Functional Characterizations of Cortical Visual Processing with Deep Predictive Models

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät
der Eberhard-Karls-Universität Tübingen
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
(Dr. rer. nat.)

vorgelegt von
Santiago A. Cadena Ceron
aus Bogota/Kolumbien

Tübingen
2024

Gedruckt mit Genehmigung der Mathematisch-Naturwissenschaftlichen Fakultät der Eberhard Karls Universität Tübingen.

Tag der mündlichen Qualifikation:	26.02.2026
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Alexander Ecker
2. Berichterstatter:	Prof. Dr. Philipp Berens
3. Berichterstatter:	Prof. Dr. Tim C. Kietzmann

To my mother.

ABSTRACT

How do the billions of neurons spread across cortical areas in the brain process visual information to facilitate a wide range of complex computations and visually-guided behaviors? Although electrophysiologists have produced great insights over the last six decades about visual processing in the brain, their resulting models fall short of describing the responses of visual neurons under naturalistic stimulation. In this dissertation, I leveraged recordings from mice and non-human primates to build deep learning models that predict single-cell responses to novel stimuli. These models provide higher accuracy than classical alternatives, which makes them the current best model of the visual cortex. However, these approaches are recurrently challenged because neural networks are seen as black boxes as they fail to provide compact or mechanistic interpretations that satisfy neuroscientists. Here, we used three main approaches to extract knowledge and generate testable hypothesis from deep predictive models. 1) We used pre-trained networks separately trained on a range of tasks to explain multiple areas of the primate and mouse visual cortices, 2) we baked hypotheses into the architecture of these end-to-end learned models and derived interpretations from their resulting parameters, and 3) we developed methods for the analysis and interpretation of nonlinear computations of deep network units. Our work provided insights into the nonlinear nature of monkey V1; the hierarchical organization of monkey and mouse visual cortices; the functional specialization of primate area V4 towards semantic tasks; the nature of normalization in monkey V1; and the types of invariances learned by intermediate units in convolutional neural networks and neurons in the brain. Overall, our results show that these deep learning-based approaches are valuable tools for continued scientific progress in understanding visual processing in the brain.

ZUSAMMENFASSUNG

Wie verarbeiten die Milliarden von Neuronen, die über die kortikalen Areale des Gehirns verteilt sind, visuelle Informationen, um eine Vielzahl komplexer Berechnungen und visuell gesteuerter Verhaltensweisen zu ermöglichen? Obwohl Elektrophysiologen in den letzten sechs Jahrzehnten bedeutende Erkenntnisse über die visuelle Verarbeitung im Gehirn gewonnen haben, reichen ihre Modelle nicht aus, um die Reaktionen visueller Neuronen unter natürlichen Bedingungen vollständig zu beschreiben. In dieser Dissertation nutzte ich Aufzeichnungen von Mäusen und nichtmenschlichen Primaten, um Deep-Learning-Modelle zu entwickeln, die einzelne Zellantworten auf neue Reize präzise vorhersagen. Diese Modelle übertreffen klassische Alternativen in ihrer Genauigkeit und stellen damit den aktuell besten Ansatz zur Modellierung des visuellen Kortex dar. Dennoch werden diese Ansätze häufig kritisiert, da neuronale Netzwerke als „Black Boxes“ gelten, die keine kompakten oder mechanistischen Erklärungen liefern, die Neurowissenschaftler zufriedenstellen. Um diese Herausforderung zu bewältigen, haben wir drei Hauptstrategien entwickelt, um Wissen aus Deep-Learning-Modellen zu extrahieren und testbare Hypothesen zu generieren: 1) Wir verwendeten vortrainierte Netzwerke, die auf verschiedene Aufgaben trainiert wurden, um mehrere Areale des visuellen Kortex von Primaten und Mäusen zu erklären. 2) Wir integrierten Hypothesen direkt in die Architektur dieser end-to-end gelernten Modelle und leiteten Interpretationen aus ihren Parametern ab. 3) Wir entwickelten Methoden zur Analyse und Interpretation der nichtlinearen Berechnungen in den Gewichte neuronaler Netzwerke. Unsere Arbeit lieferte wichtige Erkenntnisse, darunter die nichtlineare Natur des Cortexareals V1 bei Affen, die hierarchische Organisation des visuellen Kortex bei Affen und Mäusen, die funktionelle Spezialisierung des Cortexareals V4 bei Primaten auf semantische Aufgaben, die Normalisierung im Affen-V1 sowie die Invarianzen, die von mittleren Parameter in konvolutionellen neuronalen Netzwerken und Neuronen im Gehirn gelernt werden. Insgesamt zeigen unsere Ergebnisse, dass Deep-Learning-Modelle wertvolle Werkzeuge für den fortlaufenden wissenschaftlichen Fortschritt im Verständnis der visuellen Verarbeitung im Gehirn sind.

ACKNOWLEDGEMENTS

I have been very fortunate to be surrounded by remarkable people throughout this PhD.

First and foremost, I am deeply grateful to my supervisor, Alexander Ecker. Alex played a formative role in my development as a scientist. His guidance, trust, rigor, and steady support shaped not only this dissertation, but also the way I think about research. I have learned enormously from his clarity of thought, his standards for good science, and his encouragement through both the exciting and the difficult moments of this journey.

I am also very grateful to my colleagues and friends in the Bethge lab, for creating such a stimulating, generous, and enjoyable scientific environment. Many ideas in this thesis were sharpened through discussions, collaboration, and the everyday exchange that makes research feel like a shared endeavor.

My sincere thanks also go to the Tolias lab and the Sinz lab, and to the many people in both groups who contributed through collaboration, advice, and support. Working across these groups has been one of the great privileges of my PhD. I would also like to acknowledge Charles and Tarzan, the monkeys involved in part of this work in the Tolias lab.

I am grateful as well to my collaborators, mentors, and committee members, whose feedback, insight, and generosity improved this work in many ways.

On a personal note, I want to thank my mother, for her constant love, support, and belief in me. Finally, my deepest thanks go to Melanie, now my wife, for her patience, encouragement, and unwavering support throughout this entire journey. She carried me through more than she knows, and this dissertation would not have been possible in the same way without her.

CONTENTS

1 INTRODUCTION

1.1 MOTIVATION

How do the billions of neurons spread across cortical areas in the brain process visual information, enabling a wide range of visually-guided behaviors? Although for the reader it may seem effortless to recognize these words or any objects around, visual processing –the act of making sense of the external world from the light hitting the eyes– is a very complex problem solved by the brain. Once photons hit the light-sensitive cells in the retina in the back of the eye, their electrical responses trigger a processing cascade of other retina cells that results in a representation map of the visual scene in the form of action potentials fired. This low-level information is carried via the optic nerve to the thalamus and then relayed to the visual cortex in the cerebrum. How do the neural networks in the visual cortex make sense of the relevant properties of our three-dimensional world from that two-dimensional flickering map? For example, how does the brain compute representations that facilitate invariant object recognition despite the vast amount of possible scene variations (e.g. changes in pose, perspective, light source, illumination, etc.) that lead to radically different input patterns?

Neuroscientists have made tremendous progress toward answering these questions over the last century. This progress has been forged at the intersection of experimental recordings of neural activity, and functional descriptions of neural processing (i.e. computational models) linked to them. Early models of the visual system derived from hypothesis-driven experiments employing parametric stimuli (e.g. bars or gratings), have proven profoundly insightful, laying the foundation of our understanding of sensory processing. For example, the seminal work of Hubel and Wiesel (1959), unveiled the concept of localized orientation-tuned cells within the primary visual cortex (V1). This work gave rise to computational descriptions of tuning and invariance of individual neurons (e.g.simple and complex cells) to specific stimulus parameters (Adelson & Bergen, 1985b).

Simple and Complex cells: *Cells in primary visual cortex described by Hubel and Wiesel (1959) that respond when stimulated in localized areas of the visual field favouring a preferred orientation of stimulus bars or gratings. In contrast to simple cells, complex cells exhibit tolerance to local phase shifts of their preferred stimuli*

"What I cannot build, I do not understand"

— Richard Feynman

A growing number of neuroscientists consider that accurately predicting the firing patterns of a population of neurons in response to arbitrary stimuli is an essential step toward understanding the encoding of sensory information in the brain (Carandini et al., 2005). This idea drives the **neural system identification** approach (Butts, 2019; Wu et al., 2006) to study sensory systems and is at the heart of the work presented in this dissertation. Why should we go beyond compact – and in many ways *beautiful* – theories of sensory processing and instead build data-driven

models that accurately approximate complex functions between stimuli and (noisy) measurements of neural activity? Many simple models derived from our initially conceived theories fail to accurately predict neural responses to natural stimuli (Olshausen & Field, 2005). This suggests that there likely are fundamental determinants at play underlying sensory processing beyond those described by existing theories. Moreover, we may lack the tools or imagination to design experiments that reveal these unknown mechanisms. A data-driven approach can remove biases imposed by our incomplete theories and accelerate progress towards our goal.

Two recent technological strides facilitate the neural system identification approach to system neuroscience. On one side, electrophysiology and imaging advances enable now simultaneous recordings of thousands of neurons across the brain at finer spatial and temporal resolutions, providing ample experimental data (Jun et al., 2017; Steinmetz et al., 2021; Stevenson & Kording, 2011; Stringer et al., 2019). On the other side, machine learning breakthroughs have enhanced models' accuracy in capturing intricate input-output functions from observational data.

Deep neural networks (DNNs): *cascades of linearly connected nonlinear units (artificial neurons) whose parameters are tuned to optimize an objective function.*

Convolutional neural networks (CNNs) *A class of DNNs with convolutional weight sharing between processing layers*

At the center of these advancements in modeling lie deep neural networks (DNNs) (LeCun et al., 2015), whose unprecedented performance spans a wide range of tasks and applications, consistently setting the state-of-the-art in neural system identification benchmarks. In particular, convolutional neural networks (CNNs) have been a clear and successful candidate to model visual representations due to their tight historical relationship to the very same system we intend to model [1].

> *"... the cartographers' guilds struck a map of the empire whose size was that of the empire, and which coincided point for point with it. The following generations [...] saw that that vast map was useless ..."*
>
> — Jorge Luis Borges *On exactitude in science*

Building complex data-driven (deep) models of neural responses can bring us closer to having a functional *in-silico* replica of neurons. However, this exercise falls short in the eyes of neuroscientists at providing a comprehensive understanding of computations in the brain. There is a general wide-spread skepticism about using deep learning in a scientific setting. Criticisms are in general two-fold. First, DNN approximations

1 The ideas behind CNNs are rooted in neuroscience. After Hubel and Wiesel identified simple and complex cells in V1 (Hubel & Wiesel, 1962), they concluded that complex cells could achieve phase invariance by pooling from several simple cells with similar orientation tuning but shifted preferred locations. Their findings inspired Fukushima (1980) to create the *Neocognitron* – a hierarchical model stacking repeated layers of simple cell-like (dot products with feature templates) and complex cell-like (rectified pooling) operations. Similar models like HMAX (Riesenhuber & Poggio, 1999) followed the Neocognitron and were compared to human visual categorization (Serre et al., 2007). The first proof-of-work of CNNs on a real-world challenge came in the late eighties when LeCun et al. (1998) trained a CNN with backpropagation to recognize handwritten digits. However, CNNs gained widespread attention in 2012 when AlexNet (Krizhevsky et al., 2012), an eight-layer CNN trained using graphics processing units (GPUs), significantly surpassed traditional computer vision methods on ImageNet, a dataset of 1.2 million labeled images (Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, et al., 2015b)

may employ a different strategy to fit the data at hand than the system we intend to understand – if two chess players make the same move, it does not necessarily mean that their thought process behind the move was identical. For example, CNNs have been shown to learn shortcuts to fit the data well Geirhos et al., 2020, revealing a lack of sufficient constraints in our data and training, or exhibit unexpected behaviours that are not present in the target system we want to understand (e.g. susceptibility to adversarial examples (Goodfellow et al., 2014)) Second, DNNs are usually referred as *black boxes* because their large number of parameters does not admit compact understanding – the complexity is beyond the boundaries of any human's intellectual boundary. Even though we can indeed write a (long) equation for any intermediate neuron in a deep network, many scientist find this unsatisfactory. [2] For example we cannot predict very well how the output of the network would change if we modified the stimulus in some way, nor how much and in what ways the features extracted by any given neuron would contribute to the overall output. Interestingly, many questions that are difficult to answer about DNN (or CNN) computations – reflecting our lack of understanding even though we know their entire wire diagram (connectome)– apply to sensory neuroscience. For instance, we can ask the following questions abut both units in CNNs trained on object recognition and neurons in the ventral visual stream: what type of invariances are learned and where do they emerge? what stimulus features are driving the responses of a neuron?

1.2 THE GOAL OF THIS DISSERTATION

How can we then make use of deep learning methods to make progress in understanding visual processing in the brain? This is the main goal of this dissertation.

Through the lens of neural system identification, our collaborators and I were aware that even though all models are wrong, some are indeed useful. We embraced the predictive power of CNNs, but also researched multiple directions in which we could turn these models into neuroscientific knowledge – or at least testable hypotheses. We aimed to characterize different aspects of cortical visual processing at the interplay between i) improving predictions of recorded neural responses (neural system identification) and ii) generating hypotheses of biological function from them.

2 As an example consider this simple two-layer network acting on two binary inputs x_1 and x_2 where $\sigma(x) = 1/(1+\exp(-x))$:

$$f(x_1, x_2) = \sigma(20\sigma(20x_1 + 20x_2 - 10) + 20\sigma(-20x_1 - 20x_2 + 30) - 30)$$

Humans would consider *understanding* what is going on in this equation once they figure out that it is checking whether the two inputs are different (an XOR operation).

Figure 1: Overview of the contents of the dissertation and their relationship

We approached this goal with three complementary deep-learning-based approaches, leveraging electrophysiology recordings from macaques' areas V1 and V4 and 2-photon calcium trace recordings from multiple mouse visual areas in response to natural stimuli. These (RAs) were: **RA1: goal-driven modeling of neural responses**, which enabled understanding of target visual areas in the context of high-level goals (e.g. object recognition). **RA2: embedding knowledge into models' architectures** to constrain them to better match known priors and the data. **RA3: analysing predictive models** by probing them to replicate known phenomena (e.g. tuning curves to parametric stimuli) and applying interpretability methods (e.g. visualization of preferred stimuli).

With these approaches, we answered multiple research questions (RQs) in a body of work consisting of five research papers (Fig. 1) that are summarized in Chapter 2. Chapter 3 discusses our results in light of concurrent and novel work and Chapter 4 provides an outlook for potential next steps and open questions.

1.3 BACKGROUND

Neural system identification

To develop quantitative models that describe how sensory neurons encode information about the stimulus we require statistical models. This is because neural responses randomly vary in response to repeated presentations of the same stimuli. Most statistical models for neural system identification can be viewed as instances of *maximum a posteriori* (MAP) (Butts, 2019; Doya, 2007; Paninski et al., 2007; Wu et al., 2006) within a Bayesian framework where the goal is to find optimal parameters Θ of a model f that maps a stimulus $\mathbf{s}$ to the neural responses $\mathbf{r}$ assuming a noise distribution of the responses given the model predictions (or likelihood) $p(\mathbf{r}|f_\Theta(\mathbf{s}))$, and a prior distribution of the parameters $p(\Theta)$. The usual optimization loss for every recorded neuron, assuming independently recorded trials (stimulus presentations), is the log of the MAP objective (Eq. 1), which has two main terms: the data fitting loss, and a regularization term on the parameters.

$$\Theta_{best} = \underset{\Theta}{\operatorname{argmin}} \left[-\sum_{i=1}^{N} \log p(\mathbf{r}_i|f_\Theta(\mathbf{s})_i)) - \log p(\Theta) \right] \tag{1}$$

Assumptions about the likelihood function include the following distributions: Bernoulli (Butts et al., 2016) (to model spikes at high temporal resolutions), Poisson (Paninski, 2004) and Negative binomial (Pillow & Scott, 2012) (to model spike counts), or Gaussian (to model continuous response signals like fluorescence calcium traces).

Like in the broader machine learning, researchers don't usually make direct assumptions on the prior distribution, but rather add regularization terms that are aligned with their assumptions of model parameters that bias learning towards better generalization (David & Gallant, 2005; Ringach et al., 2002; Sahani & Linden, 2002a; Smyth et al., 2003;

Stevenson et al., 2008; Theunissen et al., 2001). Common assumptions on (linear) kernel weights include sparsity (Calabrese & Paninski, 2011), smoothness (McFarland et al., 2013), spatial locality (Park & Pillow, 2011), and space-time separability (Maheswaranathan et al., 2018; Park & Pillow, 2013; Shi, Gupta, et al., 2019)

The choices of the mapping function f (i.e. the model class), have grown in complexity as more data and computational methods become available. The first models were *linear-nonlinear* (LN) models (Marmarelis, 2004) composed by a linear operation on the stimulus (i.e. a filter) followed by a pointwise nonlinear function that guarantee a positive firing rate (Chichilnisky, 2001; Paninski, 2004; Simoncelli et al., 2004). Visualizing the learned filter(s) allowed for easy interpretation of the learned model computations and facilitated their diagnosis (e.g. noisy filters suggest overfitting). When using an invertible pointwise nonlinearity, and the *featurization* function Φ in Eq. 2 is the identity, LN models could be viewed as *generalized linear models* (GLMs) (Nelder & Wedderburn, 1972) (Eq. 2), which had wide-spread use in statistics (Hastie & Tibshirani, 1985) and quickly dominated neural system identification in the early visual system (Baccus & Meister, 2002; Carandini et al., 2005; Chichilnisky, 2001; Hubel & Wiesel, 1968; Movshon et al., 1978).

$$f_\Theta(\mathbf{s}) = g_\alpha(\mathbf{w}^T \Phi_\theta(\mathbf{s})); \Theta = \{\alpha, \mathbf{w}, \theta\} \tag{2}$$

Extensions of LN-based GLMs accounted for spike history (Paninski, 2004; Truccolo et al., 2005) and couplings between neurons (Butts et al., 2016; Okun et al., 2015; Pillow et al., 2008), but more importantly, recent advances came from learning (complex) featurization functions Φ. Earlier work focused on quadratic features (Rust et al., 2005b; Touryan et al., 2002; 2005), but later –at the cost of straight-forward interpretability– LNLN cascades (Antolík et al., 2016; Cadena, Denfield, et al., 2019a; Kindel et al., 2017; Vintch et al., 2015b) as they could in-principle approximate any function (Hornik, 1991). Importantly, there were early attempts to use multi-layer neural networks to fit neural responses with promising, but still sub-optimal performance compared to GLM-based attempts (Lau et al., 2002; Lehky et al., 1992; Prenger et al., 2004; Zipser & Andersen, 1988).

CNNs for neural system identification

Some quadratic and LNLN cascade models using convolutional filters fitted a separate model for every individual neuron (Rust et al., 2005a; Vintch et al., 2015a). This constrained the complexity of the models to be learned due to the limited experimental data. The pioneering work of Antolík et al. (2016) and Klindt et al. (2017) suggested to share the featurization (Φ_θ in Eq. 2) for all neurons from the same visual area and learned multi-layer CNN models of neural responses directly from the data. Many neurons compute similar nonlinear functions but only at different locations of the visual field (also known as *receptive fields*). Learning a translation equivariant representation (via convolutional weight

sharing) facilitates leveraging data from similarly tuned neurons with varying receptive fields –effectively reducing the amount of data needed to train the network. Moreover, many neurons may compute the same nonlinearity but only at different locations of the visual field This means that we can view the shared features extracted by the CNN as a new basis where all neurons can be viewed as (rectified) linear combinations of them, capturing a general model of the brain area beyond individual neurons. Advancements in computational methods, such as backpropagation (Rumelhart et al., 1986), alongside the emergence of automatic differentiation tools like PyTorch (Paszke et al., 2019) and TensorFlow (Abadi et al., 2016), in tandem with cutting-edge hardware, facilitated the optimization of CNNs, mirroring the strides made in computer vision applications (LeCun et al., 2015).

Although the *core* CNN extracts a representation of the stimulus shared across neurons, the *readout* weights $\mathbf{w}$ are separate for every neuron. Importantly, naively learning dense readouts would likely require an unfeasible large amount of data due to the high dimensionality of the core's output – spatial dimensions times the number of feature maps. Recent advances –largely coming from the work of collaborators and I– have helped improve the accuracy and interpretability of these readouts. Cadena, Denfield, et al. (2019a) introduced a dense readout with priors for sparsity, smoothness, and group sparsity (few feature maps should be used). Klindt et al. (2017) introduced reduced the number of learnable parameters by factorizing the readout tensor into a spatial mask and feature vector component. Further improvements on this idea tried to further restricted the spatial mask to a single activation point to push all nonlinear computations into the core (Lurz et al., 2021; Sinz et al., 2018). To overcome the challenge of maintaining gradient flow during training to progressively improve the estimate of the receptive field location, Sinz et al. (2018) used a spatial transformer network (Jaderberg et al., 2015) that involves a multi-scale spatial representation of the input tensor. Lurz et al. (2021) proposed instead to learn the parameters of a 2-dimensional Gaussian function and sample spatial points during training, while using the mean of the distribution at test time. Overall, these readout approaches facilitated assigning to every neuron a *bar code* (the feature vector factor of the readouts capturing all nonlinear computations) regardless of the neurons preferred location for improved interpretability and functional characterization (e.g. (Ustyuzhaninov et al., 2022)).

These and similar approaches based on CNN shared feature spaces learned end-to-end are a rapidly growing field (Batty et al., 2016; Burg et al., 2021; Cadena, Sinz, et al., 2019; Cadena, Denfield, et al., 2019a; Cadena et al., 2024; Cowley & Pillow, 2020; Ding et al., 2023; Ecker et al., 2018; Höfling et al., 2022; Kindel et al., 2017; Klindt et al., 2017; Lurz et al., 2021; Maheswaranathan et al., 2023; McIntosh et al., 2016; Safarani et al., 2021; Ustyuzhaninov et al., 2019; 2022; Walker et al., 2019; Willeke et al., 2023)

Evaluation of predictive models

Quantifying the similarity between a model and observations carrying trial-to-trial variability is fundamental to measure progress toward better understanding of neural processing. However, there is little consensus in previous neural system identification work on how models should be evaluated. A straightforward candidate is the evaluation of the **log-likelihood loss function** (first term in Eq. 1) on a held-out test set, but common issues associated to it are the lack of a finite range (e.g. 0 to 1) that is sensible to the scale of measurements and thus less robust to idiosyncrasies of separate studies using data from different modalities.

Although a typical choice is to compute the Pearson's correlation coefficient (Eq. 3) between observations $y_{i,j}$ to stimulus x_i on trial j and model predictions z_i (here dubbed **single-trial correlation**) has the downside that low values of this estimator could be due to high variability (low signal to noise ratio) in addition to poor model performance.

$$\rho_{\text{single-trial}} = \frac{\sum_{i,j}(y_{i,j} - \bar{y})(z_{i,j} - \bar{z})}{\sqrt{\sum_{i,j}(y_{i,j} - \bar{y})^2 \sum_{i,j}(z_{i,j} - \bar{z})^2}} \tag{3}$$

A common approach is to compute the **correlation coefficient to the average over repeats** (Eq. 4) using the sample mean $r_i = (1/m_i)\sum_j y_{i,j}$. Although studies (Pospisil & Bair, 2021) have shown that this estimator is biased and not consistent (does not converge to the correlation coefficient that uses the *true* average response to repeated trials μ_i when increasing the amount of stimuli), it behaves well when many repetitions are available (Cadena et al., 2024; Pospisil & Bair, 2021).

$$\rho_{\text{avg-trials}} = \frac{\sum_i(r_i - \bar{r})(z_i - \bar{z})}{\sqrt{\sum_i(r_i - \bar{r})^2 \sum_i(z_i - \bar{z})^2}} \tag{4}$$

To better account for the noise, multiple approaches have been proposed to evaluate the model's ability to capture the purely stimulus-driven variability (i.e. estimating some goodness of fit between predictions and the *true* average response to repeated trials μ_i) (Cadena, Denfield, et al., 2019b; David & Gallant, 2005; Haefner & Cumming, 2008; Hsu et al., 2004; Kindel et al., 2017; Pasupathy & Connor, 2001; Pospisil & Bair, 2021; Roddey et al., 2000; Sahani & Linden, 2002b; Schoppe et al., 2016; Sinz et al., 2018; Yamins et al., 2014b). For example, Yamins et al. (2014b) computed an average split-trial correlation (after randomly sampling multiple subsets of trials), and uses it to normalize $\rho^2_{\text{avg-trials}}$. Similarly, other approaches estimate this upper bound (or *oracle*) by averaging all the correlations of leave-one-out split-trials (Cadena, Sinz, et al., 2019; Sinz et al., 2018).

With a similar motivation and largely guided by intuition and good sound performance on synthetic scenarios, Cadena, Denfield, et al. (2019b) proposed the **fraction of explainable variance explained (FEVE)** as the complement of the ratio between the mean squared prediction error and the total variance of the responses, but subtracting off an estimate of trial-to-trial variability ($\sigma^2_n = E_i[\text{Var}_j[y_{i,j}|x_i]]$) from both numerator and

denominator (Eq. 5). More recent work has further proposed better unbiased and consistent estimators of the correlation to the average response over trials (Pospisil & Bair, 2021).

$$\text{FEVE} = 1 - \frac{(1/N)\sum_{i,j}(y_{i,j} - z_i)^2 - \hat{\sigma}_n^2}{\text{Var}[y] - \hat{\sigma}_n^2} \tag{5}$$

Beyond trying to evaluate the ability of a model to capture mean responses, recent work (Lurz et al., 2022) has also proposed metrics to evaluate the ability of a model to approximate the entire distribution around the mean by obtaining oracle bounds of the **normalized information gain** metric (Kümmerer et al., 2015; Theis et al., 2013). These approaches open the door to study *noise* distributions and how they may be meaningfully used for neural processing. In the work presented in this dissertation, we focused on capturing the mean stimulus-driven responses.

1.4 RESEARCH APPROACHES AND STATE OF THE ART

1.4.1 *RA1: Goal-driven modeling of neural responses*

As we've seen, a valuable consideration when building predictive models of neural responses is their ability to shed light on facets of visual computation. Akin to one of Marr's proposed levels of understanding (Marr & Poggio, 1976) we may want to interpret the computation of neurons in terms of the goals of the visual system as a whole. (e.g. the ability to recognize objects achieved by the visual ventral stream (DiCarlo & Cox, 2007; Felleman & Van Essen, 1991; Goodale & Milner, 1992)).

Marr's levels of understanding:
*Three levels of understanding that answer 1) the computational goal of the system or organism – in other words, **why** are a neuron's computations needed, 2) **what** computations are performed to solve that goal, and 3) **how** are these mechanisms implemented by the underlying cells and circuits.*

With the advent of CNNs capable to accurately recognize objects in images (He et al., 2016; Krizhevsky et al., 2012; Simonyan & Zisserman, 2014), there was a natural enthusiasm (Cadena, Denfield, et al., 2019b; Yamins et al., 2014a; Yamins & DiCarlo, 2016) to use them as candidate models of areas in the visual ventral stream – a series of visual areas that are believed to progressively transform stimulus information to disentangle object identities (DiCarlo & Cox, 2007; Felleman & Van Essen, 1991; Goodale & Milner, 1992).

Task-trained CNNs have been primarily turned into predictive models of neural responses via transfer learning (Donahue et al., 2014; Yamins et al., 2014b) where, instead of learning a featurization of the stimulus (Φ_θ from Eq. 2) end to end, neurons (linearly) read from the intermediate outputs of a pretrained CNN. This approach goes in the opposite direction of traditional decoding studies predicting behaviorally-relevant signals from populations of neurons by reading out sensory neural responses from functionally useful representations across the visual hierarchy (Glaser et al., 2020). As shown in work discussed in this dissertation (Cadena, Denfield, et al., 2019b; Cadena et al., 2024) and others' (Cadieu et al., 2014a; Khaligh-Razavi & Kriegeskorte, 2014a; Nayebi et al., 2018; Yamins et al., 2014a), this approach yielded unprecedented predictive performance of early and high visual areas of the visual system.

Why would this research direction address the problem of gaining meaningful biological insights if we can't fully explain how pretrained

CNNs solve their tasks in the first place? The main reason is that task-trained CNNs provide –up to a level of abstraction– an interpretation of the role of visual parts by establishing a functional and structural mapping between parts of the network (e.g. layers or feature maps) to specific structures in the brain (e.g. visual areas or cell types) (Cao & Yamins, 2021). Although we have incomplete interpretations of the learned representations of CNNs trained on a visual task, the set of governing principles that led to them are well-defined: the training data, the model architecture, the learning algorithm, and the loss function (Richards et al., 2019). In consequence, comparing the predictive abilities of candidate CNNs enables a systematic way to study factors that contribute to better matches to the brain (Conwell et al., 2022; Schrimpf et al., 2018; 2020; Sinz et al., 2019). Moreover CNNs can reveal hierarchical functional organizations of the visual system from the data without making strong prior assumptions about neural circuitry (Cadena et al., 2024; Yamins et al., 2014a).

1.4.2 *RA2: Embedding knowledge into models' architectures*

Stimulus-response pairs alone may underconstrain the space of functions required to fit the data in a manner that resembles biology [], so it is unclear if the series of stimulus transformations that the model learns are necessarily biologically meaningful. A complementary approach to gain insights from data-driven (deep) models is to constrain their architecture to represent prior knowledge or candidate hypotheses. Three popular ways to bake knowledge into the architecture are 1) implementing a weight-sharing scheme that leverages relevant symmetries, 2) building modules (layers) that correspond to specific biological components, and 3) including explicit nonlinearities akin to candidate canonical computations.

SYMMETRIES: Convolutional weight sharing in CNNs already carry a critical inductive bias that enable accurate models of the visual cortex in a data-efficient way (Klindt et al., 2017). The translation equivariance approximated by CNNs facilitates mirroring concepts in the visual system like receptive fields, retinotopy, and the separation of *what* versus *where* nonlinearities are computed in cortex. Other recent efforts have tried to learn rotation equivariant representations by other weight-sharing mechanisms (Ecker et al., 2018; Ustyuzhaninov et al., 2019; 2022). This follows the early discoveries by Hubel and Wiesel (1962) showing that neurons in V1 exhibit orientation selectivity, suggesting that the same nonlinear function (e.g. simple or complex cell-like functions) can be extracted not only at different locations, but also different orientations. Viewing orientation as a nuisance parameter, a rotation equivariant representation, followed by some rotation-invariant pooling (Ustyuzhaninov et al., 2019), facilitates the differentiation of cell types. Recent work of ours (not discussed in this dissertation) showed how this added symmetry in a data-driven CNN enabled a better characterization of the landscape of feature preferences in mouse V1 (Ustyuzhaninov et al., 2022).

ARCHITECTURAL CORRESPONDENCE TO BIOLOGICAL STRUCTURES: Over the last century, we have accumulated rich anatomical and neurophysiology knowledge that can guide the design of model architectures for improved performance and further insights of the visual system. Numerous studies following this line of research have focused on the retina as modeling target because it is the most comprehensively understood brain region in terms of neural circuitry and physiological characterizations (Baden et al., 2016; Gollisch & Meister, 2010; Masland, 2012). For example, Maheswaranathan et al. (2023) fitted three-layer CNNs to retinal ganglion cells (the output neurons of the retina) and showed that the learned intermediate representations capture the behavior of other interneurons (e.g. bipolar and amacrine cells) under natural and synthetic stimulation. Similar work has built-in more biophysical realism into deep networks to capture details of the temporal processing in the inner retina, including feedback circuits and synaptic release mechanisms (Schröder et al., 2020), reproducing *in-silico* the results of known pharmacology experiments. On a higher level, other work has imposed biologically inspired constraints on networks not necessarily fit to neural responses, but to a an ethological relevant task like object recognition that mirror known aspects of biological vision and provide mechanistic explanations for observed phenomena. For example, neural networks that capture antagonistic center-surround unit preferences in layers intended to model the retina after introducing a constraint in the number outputs consistent with the anatomy of the optic nerve as a stringent bottleneck (Lindsey et al., 2019). Similar work included two processing pathways akin to the dorsal and ventral stream (Bakhtiari et al., 2021), recurrent networks with feedback connections (Kietzmann et al., 2019; Kubilius et al., 2018; 2019; Nayebi et al., 2018; Spoerer et al., 2020), lateral connections (Linsley et al., 2018), V1-like variability and selectivity (Dapello et al., 2020), and Dale's law (Eccles, 1976) with exclusively excitatory and inhibitory neurons (Cornford et al., 2020; Li, Cornford, et al., 2023; Minni et al., 2019).

INCORPORATING PREDETERMINED NONLINEARITIES: Due to the rather homogeneous circuitry across cortical areas, neuroscientists hypothesise that there are fundamental or *canonical* computations performed by cortical circuits (Carandini & Heeger, 2012; Douglas et al., 1989) that may generally account for properties of mammalian intelligence (Miller, 2016). This idea is further supported by the success of architectures with standard computational motifs like CNNs and Transformers (Vaswani et al., 2017) in a a wide range of domains beyond vision. Therefore, a plausible strategy to gain insights from data-driven models is to hard-code nonlinearities derived from canonical computation hypotheses and interpret any learned parameters associated to them. A common computational candidate derived from observed neurophysiology phenomena is gain modulation (i.e. suppression or facilitation) of neural responses (Fu et al., 2014; Lee et al., 2013; Pi et al., 2013; Polack et al., 2013; Zhang et al., 2014), which is evoked by additional stimuli within the receptive field of neurons (Busse et al., 2009; Carandini & Heeger, 2012), context or surrounding stimuli outside the receptive field (Cavanaugh et al., 2002a),

or top-down influences such as attention (Reynolds & Heeger, 2009). Normalization – as in dividing the activity of neurons by the activity of others – has been proposed as a canonical computation that could account for these phenomena (Carandini & Heeger, 2012), but many of the details of how it could work under natural image stimulation are unknown. Our work in Burg et al. (2021) discussed in Chapter 2 shows how explicitly implementing normalization in an end-to-end trained CNN-like model sheds light on the nature of normalization in V1. Follow-up work has included these type of nonlinearities on image classification tasks showing robustness benefits (Cirincione et al., 2022).

1.4.3 *RA3: Analysing predictive models*

While deep models are more accurate than classical LN (or GLM) models, they lack a straightforward interpretation of the neurons' feature selectivity. Although they are usually viewed as *black-boxes*, a clear advantage over studying the representations in the brain is that we we can run virtually unlimited experiments on these models and propagate gradients through them. In an effort to improve human understanding of the learned representations of data-driven models and generate testable hypothesis, computational neuroscientists have relied on probing models *in-silico* with artificial stimuli to replicate known (nonlinear) phenomena, and visualizing training set or gradient-based synthesized inputs.

In-silico experiment using parametric stimuli produce tuning curves that unveil discrepancies between trained models and recorded neurons, and can potentially shed light into the overall functional organization of the visual system. Differences between the model's behavior and known nonlinear electrophysiology phenomena helps us *fix* models – a hallmark of what it would mean to understand a model. Moreover, these differences facilitate discriminating different, but equally performing, models on our limited experimental data. Work discussed in this dissertation Burg et al. (2021) showed how stimuli consisting with overlapping gratings with varying contrast unlocking cross-orientation inhibition nonlinearities in the brain (Carandini & Heeger, 2012) can be capture by one class of models, but not by others, although these models' performance was similar. In a similar way, Maheswaranathan et al. (2023) used showed how the layers of a CNN trained on retinal ganglion cell outputs revealed that many nonlinear retinal computations are engaged in natural scenes including fast contrast adaptation, latency coding, motion reversal, motion anticipation, omitted stimulus response and object motion sensitivity. Beyond validating models and producing testable hypotheses, tuning curves from *in-silico* experiments provide a way to characterize the functional organization of a sufficiently large sample of recorded neurons. For example, Ustyuzhaninov et al. (2022) replicated battery of classical electrophysiology experiments probing for selectivity to orientation and phase, and nonlinear response properties like cross-orientation inhibition, size-contrast tuning, and surround suppression to show how these properties are distributed across different candidate cell types.

Feature visualizations have been a popular way to analyze the representations of trained CNNs on object recognition (Cadena et al., 2018; Erhan et al., 2009; Nguyen et al., 2016; 2017; Olah et al., 2017) and used to characterize the preferred feature stimulus of neurons in different cortical visual areas (Bashivan et al., 2019; Cadena et al., 2018; Ding et al., 2023; Pierzchlewicz et al., 2023; Ponce et al., 2019; Tong et al., 2023; Walker et al., 2019; Willeke et al., 2023). In addition to making qualitative judgements about the resulting synthesized images, these images were validated *in-vivo* by displaying them back to the neurons (Bashivan et al., 2019; Ding et al., 2023; Walker et al., 2019; Willeke et al., 2023). These studies showed that model-based synthesized stimuli indeed excite real neurons maximally, proving the ability to *control* aspects of the system we aim to understand. One meaningful clear conclusion from this like of work is that unlike macaque V1, mouse V1 exhibit preference for complex features that occur frequently in natural scenes but deviated strikingly from Gabor-like functions (oriented edge detectors) (Walker et al., 2019). Beyond characterizing neurons with a single image that drives them most, other work (Cadena et al., 2018; Ding et al., 2023; Willeke et al., 2023) has focused on identifying what aspects of the stimulus the neurons become invariant to. We elaborate more on this line of work in Chapter 2.

1.5 RESEARCH QUESTIONS

We identified three research approaches (RA1,RA2, RA3) to use deep learning methods to gain insights about aspects of visual cortical processing. Here, we enumerate research questions associated to these approaches that stem from gaps in knowledge at the moment we started addressing them. Notably, there were other questions answered in our work as part of solving these main questions. Other important research questions associated to these approaches were addressed in collaboration with colleagues, but not discussed in this dissertation. Fig. 1 provides an overview of these questions relate to both research approaches, and the publications discussed in this dissertation.

Goal-driven modeling of neural responses (RA1)

The recent success of task-driven modeling (Yamins & DiCarlo, 2016) at predicting neural responses in high visual areas of the visual ventral stream like V4 and IT (Yamins et al., 2014a) begged the question of their effectiveness at predicting other sensory areas. Classical models on primary visual cortex – arguably the most widely studied cortical visual area – still left a large fraction of the variance unexplained under natural image stimulation (Olshausen & Field, 2005). Our first research question followed naturally from these:

Research question No. 1: Do the features learned by CNNs trained on object classification predict macaque V1 responses to natural stimuli?

An strong observation of goal-driven models in the primate ventral stream (Cadena, Denfield, et al., 2019b; Cadieu et al., 2014b; Kriegesko-

rte, 2015; Yamins et al., 2014a), and other sensory areas like auditory cortex (Kell et al., 2018) was the hierarchical correspondence between layers of the network and cortical areas in the brain exhibiting a gradient of complexity. Existing neurophysiology and anatomical work on the rodent brain has determined that there are clear visual cortical areas, but has not converged to a clear hierarchical ordering of them that correspond to the primate ventral stream and its role on object recognition. The unclear functional organization of mouse visual areas motivated the following question:

Research question No. 2: Can we unveil a hierarchical organization of mouse visual areas using CNNs trained on object recognition?

Object recognition has been the predominant task driving transfer learning models of visual areas (Schrimpf et al., 2018). However, there may be *other* useful visual tasks that the brain may need to solve to support various downstream behaviours. At the same time, the functional role of mid-level areas of the visual ventral stream like area V4 are not as well understood as primary visual cortex, but there is evidence of its involvement on 2D shape and texture boundary (Pasupathy & Connor, 2001; Pasupathy et al., 2020) and 3D processing (Srinath et al., 2021). In order to better characterize the functional role of area V4 from a normative point, and taking advantage of multi-task labeled datasets and models from the computer vision community (Zamir et al., 2018), we asked the following question:

Research question No. 3: How well do tasks beyond object classification (e.g. 3D, Geometric, 2D) predict macaque V1 and V4 responses to natural stimuli?

Once we were able to characterize V4 (and V1 for comparison) with a pattern of predictive performances from multiple models trained on various computer vision tasks, we found that some seemingly different tasks yielded comparable performances. This observations led to the following question:

Research question No. 4: Do different task representations explain the same aspects of V4 stimulus-driven variability?

Embedding knowledge into models' architectures (RA2)

Primary visual cortex is the most studied cortical visual area so it is a great place to start investigating how to incorporate prior knowledge into deep predictive models to gain meaningful insights. A wide range of nonlinear neurophysiology phenomena associated to local gain control has been explained with divisive normalization, suggesting that this computation may be a fundamental, canonical computation in sensory cortex (Carandini & Heeger, 2012; Miller, 2016). However, it is still unclear how normalization acts in settings beyond simple parametric stimuli like gratings or grids with varying contrasts (Carandini et al., 1997). We then asked ourselves the following question:

Research question No. 5: How can we characterize divisive normalization in V1 under natural image stimulation with data-driven models?

Analyzing predictive models (RA3)

To validate insights that we extract from data-driven models in terms of their biological relevance, we can verify that they reproduce known phenomena. Cross-orientation inhibition is an prominent nonlinear phenomenon observed in primary visual cortex where the response of a neuron to a preferred grating stimulus within its receptive field is suppressed by overlapping it with another grating at would not elicit responses on its own (DeAngelis et al., 1992; Freeman et al., 2002; Heeger, 1992a; Morrone et al., 1982). Given that we had trained multiple models based on Gabor filter banks, CNNs, and CNN networks with divisive normalization nonlinearities on V1, we asked the following question:

Research question No. 6: Do data-driven models of V1 responses to natural images capture cross-orientation inhibition?

Feature visualization approaches have been focused on maximally driving neurons, and making mostly qualitative assessments on what individual units in CNNs prefer (Cadena et al., 2018; Erhan et al., 2009; Nguyen et al., 2016; 2017; Olah et al., 2017). Beyond selectivity, *invariance* to stimulus features is critical in a lossy, hierarchical system supporting complex tasks like object recognition. Like in the visual ventral stream, our understanding of what invariances are learned throughout convolutional networks trained on object recognition is poor. To develop methods that reveal invariances in deep predictive models of neural responses (and thus build hypotheses of those in the brain), we first asked ourselves:

Research question No. 7: How can we interpret the invariances learned in early-to-mid layers of CNNs trained on object recognition?

1.6 PUBLICATIONS

Our collaborators and I addressed these research questions in multiple papers that I list bellow.

Publications included in this dissertation

The five publications below – three journal articles, one conference paper, and one conference workshop paper– form the body of this dissertation. The reader can find these publications in full in the Appendix of this dissertation together with a statement of author contributions. A summary of the motivation, results, and discussion sections of each paper are included in Chapter 2.

- **Cadena, Santiago A**, Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897

- **S. A. Cadena**, Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., Reimer, J., Bethge, M., Tolias, A., & Ecker, A. S. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? *Advances in Neural Information Processing (NeurIPS) Neuro-AI Workshop*

- **Cadena, Santiago A**, Willeke, K. F., Restivo, K., Denfield, G., Sinz, F. H., Bethge, M., Tolias, A. S., & Ecker, A. S. (2024). Diverse task-driven modeling of macaque v4 reveals functional specialization towards semantic tasks. *PLOS Computational Biology*, *20*(5), e1012056

- Burg, M. F., **Cadena, Santiago A**, Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., & Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, *17*(6), e1009028

- **Cadena, Santiago A**, Weis, M. A., Gatys, L. A., Bethge, M., & Ecker, A. S. (2018). Diverse feature visualizations reveal invariances in early layers of deep neural networks. *Proceedings of the European Conference on Computer Vision (ECCV)*, 217–232

*Related work **not** included in this dissertation*

The following publications are not formally included in this dissertation, but follow closely the research approaches enumerated before, address related research questions, and may be mentioned in Chapter 3 and 4. They comprise four conference papers, three submitted manuscripts, and a conference competition workshop.

- Ecker, A. S., Sinz, F. H., Froudarakis, E., Fahey, P. G., **Cadena, SA**, Walker, E. Y., Cobos, E., Reimer, J., Tolias, A. S., & Bethge, M. (2019). A rotation-equivariant convolutional neural network model of primary visual cortex. *International Conference on Learning Representations (ICLR)*

- Lurz, K., Bashiri, M., Willeke, K., Jagadish, A., Wang, E., Walker, E., **Cadena, SA**, Muhammad, T., Cobos, E., Tolias, A., & Sinz, F. H. (2021). Generalization in data-driven models of primary visual cortex. *Ninth International Conference on Learning Representations (ICLR 2021)*

- Ustyuzhaninov, I., **Cadena, SA**, Froudarakis, E., Fahey, P. G., Walker, E. Y., Cobos, E., Reimer, J., Sinz, F. H., Tolias, A. S., Bethge, M., et al. (2019). Rotation-invariant clustering of neuronal responses in primary visual cortex. *International Conference on Learning Representations*

- Ustyuzhaninov, I., Burg, M. F., **Cadena, Santiago A**, Fu, J., Muhammad, T., Ponder, K., Froudarakis, E., Ding, Z., Bethge, M., Tolias, A. S., et al. (2022). Digital twin reveals combinatorial code of non-linear computations in the mouse primary visual cortex. *bioRxiv*, 2022–02

- Safarani, S., Nix, A., Willeke, K., **Cadena, SA**, Restivo, K., Denfield, G., Tolias, A., & Sinz, F. (2021). Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, *34*, 739–751

- Ding, Z., Tran, D. T., Ponder, K., Cobos, E., Ding, Z., Fahey, P. G., Wang, E., Muhammad, T., Fu, J., **Cadena, Santiago A**, et al. (2023). Bipartite invariance in mouse primary visual cortex. *bioRxiv*

- Willeke, K. F., Restivo, K., Franke, K., Nix, A. F., **Cadena, Santiago A**, Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Ecker, A. S., et al. (2023). Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. *bioRxiv*, 2023–05

- Willeke, K. F., Fahey, P. G., Bashiri, M., Pede, L., Burg, M. F., Blessing, C., **Cadena, Santiago A**, Ding, Z., Lurz, K.-K., Ponder, K., et al. (2022). The sensorium competition on predicting large-scale mouse primary visual cortex activity. *arXiv preprint arXiv:2206.08666*

2 RESULTS

2.1 TASK-DRIVEN MODELING OF MACAQUE V1

> **Cadena, Santiago A**, Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897

Motivation

Classical approaches to model primary visual cortex (V1) based on LN and energy models incorporating insights from Hubel & Wiesel's work (like simple or complex cells) predicted well responses to simple, parametric stimuli like Gabor patches (Adelson & Bergen, 1985a; Heeger, 1991; Jones & Palmer, 1987), but they failed to account for neural responses to natural images (Olshausen & Field, 2005; Talebi & Baker, 2012). Existing attempt based on LN-LN cascades with (convolutional) subunits (Rust et al., 2005a; Vintch et al., 2015a) or handcrafted wavelet representations like Gabor filter banks (GFB) (Willmore et al., 2008) outperformed LN models, but still fall short at capturing nonlinearities unlocked with natural stimulation and are largely limited by the amount of available experimental data.

In this paper, our goal was to build models that capture these nonlinearities and improve predictions of V1 responses to natural stimuli using convolutional neural networks, following their success in a wide range of tasks (LeCun et al., 2015). Following the success of similar work applied to different brain areas or species, we investigated how well fully end-to-end data-driven approaches (Antolík et al., 2016; Batty et al., 2016; Klindt et al., 2017; McIntosh et al., 2016), and task-driven approaches based on CNNs trained on image classification (Yamins & DiCarlo, 2016) predict single-cell recordings from macaque V1. Moreover, we wanted to compare these and classical approaches on the same benchmark composed of stimuli capturing varying levels of higher-order statistics of natural images, and determine the point in the hierarchy within object recognition CNNs where V1 has a stronger match.

Results

We recorded 166 well-isolated V1 neurons from two awake, fixating macaques in response to thousands of natural images that were flashed one after the other for 60ms. The set of stimuli contained gray-scale samples from the ImageNet dataset (Russakovsky, Deng, Su, Krause, Satheesh, Ma, Huang, Karpathy, Khosla, Bernstein, et al., 2015a) along

with textures synthesized from these images using a texture synthesis algorithm (Gatys et al., 2015) that retained up to different levels of high-order statistics. We linked images to neural responses by counting post-stimulus spikes, using these pairs to train and evaluate our models.

Our goal-driven model, utilizing VGG19 features (Simonyan & Zisserman, 2014), featured a GLM readout with crucial readout regularizations for improved performance. Peak predictive performance emerged at an intermediate layer, about a fourth into the network. We also fitted an end-to-end CNN with a factorized readout (Klindt et al., 2017), including a flexible pointwise nonlinearity for improved performance. Performance plateaued after three convolutional layers as measured by FEVE.

Both data and task-driven models performed similarly, explaining around half of the explainable variance. These CNN models outperformed LN and GFB-based GLM models. However, the GFB reached 89% of the VGG-based model's FEVE. We found that individual neuron performance lacked correlation with classical tuning properties obtained *in-silico*. Moreover, both deep learning models effectively generalized across diverse stimulus domains (when using stimuli with varying degrees of natural image statistics). Furthermore, the data-driven model proved more sensitive to training data volume, as its performance drops compared to the goal-driven model when trained with only a fifth of the data.

Discussion

In line with work on higher areas of the ventral stream like V4 or IT (Cadieu et al., 2014a; Khaligh-Razavi & Kriegeskorte, 2014b; Yamins et al., 2014a), task-driven (and also data-driven) models set the state-of-the-art at predicting V1 responses. While object classification may not sufficiently constrain the representations to replicate the brain at a mechanistic level, we found that this goal leads to useful features for V1, despite V1's usefulness in other downstream tasks.

Human fMRI studies established hierarchical correspondence between pretrained CNN layers and the visual cortex, linking V1 to the earliest convolutional layers (Güçlü & van Gerven, 2015; Kriegeskorte, 2015). Contrary to this, our electrophysiology recordings with a deeper network (VGG19) revealed that V1 is better explained by features from layers multiple nonlinearities away from pixels (around a fourth into the network). This challenges the traditional approach of using one or two nonlinear transformations for V1 modeling and suggests the potential match between the earliest CNN layers and computations performed by the retina or LGN, consistent with findings that shallow models effectively capture their responses (McIntosh et al., 2016).

Our work still leaves unanswered the question about what specific nonlinearities are captured by the deep learning models over those present in the GFB energy model. We investigated whether the delta in performance for individual neurons was correlated to orientation tuning or phase invariance, but found no significant relationship. Based on classical electrophysiologists' work, clear candidates include divisive normalization and over-complete sparse coding-induced nonlinearities. How-

ever, there are not yet models applicable to natural stimuli that incorporate such nonlinearities. We leave for future work building these models and comparing them with our current state-of-the-art.

Our models leave roughly half of the explainable variance unexplained. Collecting datasets with larger sets of stimuli and neurons, along with incorporating inductive biases like recurrent connections, divisive normalization, or features learned for other tasks beyond object classification are promising directions to improve alignment with V1 representations.

2.2 TASK-DRIVEN MODELING OF MOUSE VISUAL AREAS

> **S. A. Cadena**, Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., Reimer, J., Bethge, M., Tolias, A., & Ecker, A. S. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? *Advances in Neural Information Processing (NeurIPS) Neuro-AI Workshop*

Motivation

What is the function of the multiple visual areas of the mouse brain? Over the last 50 years, researchers have identified that like in the primate brain, there are several retinotopic visual areas across the rodent brain (Dräger, 1975; Garrett et al., 2014; Wang & Burkhalter, 2007; Zhuang et al., 2017). Traditional efforts to bring functional characterization to these areas has been limited to the tuning of many of their neurons to variables of parametric stimuli (Andermann et al., 2011; Ayzenshtat et al., 2016; Glickfeld et al., 2013; Marshel et al., 2011; Roth et al., 2012; Tohmi et al., 2014). However, we lack a functional description of what these areas do in the context of a high-level visual task nor how they are organized to facilitate complex visually-guided behaviors. In contrast to the primate visual system, it is still unclear if there are object-processing pathways like the ventral stream across multiple areas in the mouse cortex nor the existence of parallel processing pathways.

Following the success of CNNs at predicting responses in primate visual areas (Güçlü & van Gerven, 2015; Kriegeskorte, 2015; Yamins et al., 2014a), we wanted to evaluate their functional match to mouse visual areas. Besides, these networks could shed light into the functional organization of mouse visual areas as they have successfully identified complexity gradients across areas in the ventral stream and human auditory cortex (Kell et al., 2018) – a signature of hierarchical organization.

Results

We recorded thousands of excitatory neurons in mouse visual areas V1, LM, AL, and RL in response to thousands of natural images using a large-field-of-view, two-photon mesoscope. Simultaneously, we also recorded the pupil's position and dilation, and the running speed of the animal. We build models for each area by reading out from VGG16 lay-

ers (Simonyan & Zisserman, 2014) using the spatial transformer readout (Sinz et al., 2018). In contrast to monkeys, mice do not fixate, so we corrected for eye-movements by learning a nonlinear transformation of the measured pupil's location. Finally, we modulated the output of the readout with a learned nonlinear mapping of the measured pupil dilation and running speed of the animal (Sinz et al., 2018).

We found that optimizing the input resolution of the images (the number of pixels covering a degree of visual angle) was critical to find the best predictive layers for each area. Importantly, we found no clear hierarchical correspondence between layers and areas in the mouse brain – V1, AL, and RL matched best with the same layer, while LM with one layer before that. Nevertheless, the VGG16-based model outperformed classical (shallower) models consisting of subunit energy models with Gabor quadrature pairs – same as the GFB from Cadena, Denfield, et al. (2019b)–, reaching 70-78% of the oracle correlation (the noise ceiling). Surprisingly, we found that replacing the original weights of VGG16 with randomly initialized weights, and then only learning the readout weights yielded similar performance to VGG16 features. We found that the number of linear-nonlinear (LN) layers –rather than the training objective– was critical to best match neural activity across all visual areas.

Discussion

Our work stresses that to draw conclusions about end goals in neural system identification studies with task-driven approaches, it is important to include random nonlinear features as baselines. Our results also caution against strong conclusions about object categorization as a primary driver of representations in the mouse visual cortex, despite high predictive performance. In our analysis, the critical factor influencing well-predictive representations was the number of LN transformations.

When optimizing over input resolutions we found no strong hierarchical correspondence between CNN layers and visual areas. Although the functional organization of the mouse visual areas might be different to that of the primate –or we simply did not image all of the relevant object-processing visual areas in the mouse visual cortex– our findings do spark concerns whether previously identified hierarchical correspondence would be present once input scale is taken into account.

Overall, our findings do not fully exclude the existence of object processing pathways across the mouse visual cortex, nor a hierarchical organization of the visual areas. Nevertheless, they reveal that goal-driven approaches based on object classification do not provide the right approximations to describe representations in the mouse visual cortex. We hypothesize that relevant ethological visually guided tasks involving dynamic stimuli (e.g. pray capture tasks (Hoy et al., 2016)) could be more fruitful.

2.3 DIVERSE TASK-DRIVEN MODELS OF MACAQUE V4 RESPONSES

> **Cadena, Santiago A**, Willeke, K. F., Restivo, K., Denfield, G., Sinz, F. H., Bethge, M., Tolias, A. S., & Ecker, A. S. (2024). Diverse task-driven modeling of macaque v4 reveals functional specialization towards semantic tasks. *PLOS Computational Biology*, *20*(5), e1012056

Motivation

The functional role of area V4 in primate visual information processing is unclear. Over the last decades, electrophysiologists have recorded its responses to parametric stimuli to identify what features of the stimulus V4 neurons are tuned for. This line of research has been fruitful as it unveiled V4's tuning in a high-dimensional space that enables the encoding of shapes and surface characteristics of objects in the world (Kim et al., 2019; Pasupathy & Connor, 2001; Pasupathy et al., 2020). However, this approach is limited by the available experimental time to test all the different stimulus features that could be relevant for V4. Moreover, there may be nonlinear aspects unlocked by natural stimuli that may be missed with these approaches.

A tending approach to rather characterize V4 function from a normative point of view by predicting neural responses to natural images with CNNs trained on object classification (Bashivan et al., 2019; Cadieu et al., 2014b; Pospisil et al., 2018; Willeke et al., 2023; Yamins et al., 2014a), or self-supervised training objectives (Zhuang et al., 2021). The partial success of these approaches have been taken as evidence of V4's role in object processing, but the extend to which accurate predictions are due to this training objective is still an open question – do other computational goals beyond object classification explain V4 responses well? If so, do they explain aspects of V4 function different than those explained by object recognition driven features?

Results

We used the representations learned by CNNs trained on 23 different visual tasks from the taskonomy project (Zamir et al., 2018) to predict hundreds of single-cell responses to thousands of natural images from macaque areas V1 and V4. These networks shared the same architecture and training data so any predictive performance differences we observe across task networks can be attributed to the training objectives alone. We built models using a point readout (Lurz et al., 2021) for each visual area, taskonomy network, and layer, and optimized the input scale to identify the best matching layer.

We found that indeed V4 was better explained by a higher layer than V1 across models. Moreover, the two best predictive models of V4 were the two semantic classification tasks: scene and object classification. In contrast to V4, we found that V1's top predictive models were more diverse and not specifically tied to semantic-related tasks coming also

from 3D and 2D related task groups. The features of task-trained models predicted V4 responses much more poorly than V1, although the noise ceilings of both datasets were comparable. Nevertheless, most models in V4 and all models in V1 significantly outperformed a random untrained baseline. We found that the choice of task in V1 was less detrimental to performance than in V4 evidenced by the latter's higher variance in task performances compared to the former's. Similarly, at a coarser level (i.e. grouping tasks into semantic, 3D, 2D, and geometric) we found that the specialized role of V4 towards semantic tasks was also prevalent.

By building models that jointly read out from pairs of tasks, we found that non-semantic tasks contribute useful nonlinearities beyond those captured by individual semantic classification tasks, as they improved performance over any individual task baseline. 2D keypoints, and any semantic task were consistently among the top performing pairs for V1 and V4, respectively, suggesting that these tasks capture critical portions of the variance in these areas.

Finally, we evaluated imagenet-trained models with various architectures, data augmentation strategies, and supervised and self-supervised training objectives. We found that data-rich, and robust models are the most critical factors driving better models of V1 and V4.

Discussion

Our predictive modeling approach provides another lens on how we can better understand V4 function based on the patterns of predictive performance across diverse tasks. Our results support prior indications that semantic objectives, such as object classification, strongly influence representations aligning with ventral stream responses. We also unveiled a distinct task specialization in V4, emphasizing its affinity for semantic tasks compared to the more versatile representations in V1. However, our work also highlights valuable contribution of non-semantic tasks at predicting V4 responses, suggesting that they can capture useful nonlinearities beyond those in individual semantic tasks.

In agreement with novel work focusing on single-neuron's solid shape coding (Srinath et al., 2021), we found a strong association between V4 and 3D visual processing. This carries significance as it goes beyond classical descriptions of V4 focusing on flat shape processing (Pasupathy & Connor, 2001; Pasupathy et al., 2020), and suggest a promising research avenue to understand the relationship between normative accounts of V4 at the population level and individual cell properties. At the same time, V1's strong match to 2D task features (e.g. Edge detection, 2D Keypoints) aligns with classical views of V1 function. However, other semantic objectives and ImageNet-trained CNNs showed strong performances too. Inline with recent work (Cadena, Denfield, et al., 2019b; Dapello et al., 2020), these results emphasize the ability of semantic objectives to induce orientation selectivity and nonlinear phenomena like cross-orientation inhibition (Burg et al., 2021) in its early layers.

We hypothesize that the low V4 performance values (compared to V1 at least) are due to 1) our experimental design maintaining image sequences that do not remove history effects that are likely stronger in

V4, and 2) the distribution discrepancies between stimulating images and the taskonomy training set of images. Future models accounting for dynamic effects are promising way to improve performance (Kietzmann et al., 2019). We also noticed that the performance differences across tasks were small, but likely carry consequential conclusions because 1) we found a structured pattern of tasks performance consistent with prior, 2) a substantial fraction of neurons were better explained by one task group vs another across group comparisons, and 3) there is evidence in V1 showing that small differences can carry meaningful implications about the ability of models to capture phenomena (Burg et al., 2021).

In agreement with work on V1 (Dapello et al., 2020), we found that robustness to adversarial perturbations of models trained on large-scale labeled datasets yield the top predictive performance on V1 and V4. This is also aligned with work supporting the converse: task-trained CNNs that are simultaneously trained to match neural responses tend to be more robust to input perturbations (Safarani et al., 2021).

Overall, diverse normative approaches provide another way to functionally characterize brain areas. Promising future work lies at the intersection with *in-silico* experiments of single cells to facilitate the generation of hypotheses about tuning directions. Clear candidates to do this involve maximally exciting images (MEIs) (Bashivan et al., 2019; Walker et al., 2019; Willeke et al., 2023), diverse exciting inputs (Cadena et al., 2018; Ding et al., 2023), and controversial stimuli (Golan et al., 2020).

2.4 LEARNING DIVISIVE NORMALIZATION IN V1

> Burg, M. F., **Cadena, Santiago A**, Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., & Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, *17*(6), e1009028

Motivation

We have seen that CNNs significantly improved predictions of V1 responses to natural stimuli over classical and subunit energy models (Cadena, Denfield, et al., 2019b). However, it has been difficult to gain biological conclusions about the nature of the nonlinearities approximated by these CNNs beyond those captured by the other, more interpretable, models. We also don't know if there are first principles that explain these nonlinearities nor if they can be described in a compact way in the first place.

Divisive normalization (DN) is a promising candidate for those nonlinearities (Carandini & Heeger, 2012) where a neuron's driving input is normalized divisively by a weighted sum over nearby neurons' responses Carandini and Heeger, 2012; Heeger, 1992a. At the same time, DN can be derived from a redundancy reduction objective (Schwartz & Simoncelli, 2001; Sinz & Bethge, 2008), and has explained a wide variety of neurophysiological observations during stimulation with parametric stimuli (Sawada & Petrov, 2017) both within the receptive field (Busse

et al., 2009; Heeger, 1992a; Morrone et al., 1982) and with its surround (Cavanaugh et al., 2002a; 2002b)

Here, our goal was to understand via data-driven models how divisive normalization (DN) operates under natural image stimulation in V1, especially regarding interactions between neurons with overlapping receptive fields.

Results

We developed an image-computable predictive model with DN whose parameters are learned to optimally predict V1 responses to natural stimuli (data from Cadena, Denfield, et al. (2019b)). The model has a core that takes input images and computes a shared feature space implementing a DN nonlinearity, and a readout that maps this shared space individually to each neuron. In a nutshell, input images were convolved with forward filters and then rectified with an exponential nonlinearity to produce channel outputs. Each of these outputs was then divided by a weighted sum of all other channel outputs.

We found that our DN model outperforms the subunit energy model using Gabor quadrature pairs (GFB). The DN model's improvement over this baseline DN model reached half of the gap to the SOTA data-driven CNN model with three convolutional layers. However, the DN model achieved this with significantly less parameters than the CNN.

Do the DN and CNN models trained on natural images reproduce nonlinear behaviors found by electrophysiologists? We identified the optimal Gabor stimulus for every neuron and overlapped it with a similar Gabor but orthogonal orientation and varying contrast. Such stimuli yield nonlinear phenomena in real neural responses, namely cross-orientation inhibition (Heeger, 1992a). We found that the DN and the CNN models indeed exhibit this phenomenon. Although the performance improvement of our DN model over the GFB model was roughly 3% of the fraction of explainable variance explained (FEVE), we found that the GFB model did not capture cross-orientation inhibition.

We then investigated the learned parameters of the DN core and found that normalization within the receptive field is feature-specific: the forward features showing orientation tuning are normalized more strongly by other forward features with similar orientation tuning. To test generality, we measured the cosine similarity between all pairs of forward filters and found that more similar features contribute preferentially (with higher weights) to normalization. To determine the importance of orientation-specific normalization, we learned a similar DN model where the normalization weights were shared across features. This non-specific DN model yielded poorer accuracy, favoring feature-specific DN normalization.

Finally, given that surround suppression is known to be orientation-specific as well (Cavanaugh et al., 2002a; Coen-Cagli et al., 2015), a potential concern is that we are capturing contributions from the extra-classical surround of a unit's receptive field. We ruled this out using *in-silico* experiments where the stimulus for a unit was its optimal Gabor grating masked with a circular aperture that was progressively in-

creased in diameter until it fully extended into the surround. We found for almost all neurons that there was no suppression by stimulating the surround.

Discussion

Our results showed that CNNs do not simply approximate biological heterogeneities on top of energy-based model feature spaces, but that they capture nonlinear functions of biological relevance like DN. This was revealed by the cross-orientation inhibition phenomena observed in DN and CNN, but not GFB models. Therefore, small performance differences may carry meaningful discrepancies about capturing biological phenomena, suggesting that candidate models should be verified *in-silico* with existing knowledge beyond simply testing for accuracy. Consequently, the higher performance of the CNN with respect to the DN model suggests that there could still be important nonlinearities approximated by the CNN that we have yet to account for with compact descriptions – if there are any.

By inspecting the learned parameters of the DN model, we found evidence for feature-specific normalization, which somewhat contradicts some of the existing work that uses simpler stimuli like combinations of driving and masked gratings (DeAngelis et al., 1992; Morrone et al., 1982). Some studies found phenomena that are predominantly explained by nonspecific normalization Heeger (1992b), some encountered only weak orientation-specific phenomena in only relatively few cells (Bonds, 1989; DeAngelis et al., 1992), and some proposed effective quantitative models assuming nonspecific normalization, but did not compare with an orientation-specific alternatives (Busse et al., 2009). These apparent discrepancies could be attributed to the use of natural stimuli – our work is the first to develop a quantitative analysis of orientation-specificity on a dataset of spiking neurons containing natural images. However, our conclusions do agree with models derived from classical psychophysics data that required orientation-specific divisive normalization to capture data on contrast detection and discrimination, and oblique masking (Itti et al., 2000; Schütt & Wichmann, 2017). Additionally, seminal work derived DN-like mechanisms from first principles of efficient coding (Barlow, 1961) and suggested that normalization should be stronger for neurons with higher dependencies. Orientation-specific normalization has been also supported by recurrent sparse coding models that reproduce phenomena like cross-orientation inhibition (Zhu & Rozell, 2013).

Promising research directions following our work include investigating interactions between the receptive field and its surround, and how orientation-specific normalization is implemented by the connectivity of neurons in V1.

2.5 UNVEILING INVARIANCES WITH DIVERSE FEATURE VISUALIZATIONS

> **Cadena, Santiago A**, Weis, M. A., Gatys, L. A., Bethge, M., & Ecker, A. S. (2018). Diverse feature visualizations reveal invariances in early layers of deep neural networks. *Proceedings of the European Conference on Computer Vision (ECCV)*, 217–232

Motivation

CNNs are now widely used in scientific disciplines so there is growing interest in understanding the representations they learn. A popular approach relies on visualizing input features that activate the CNN units maximally. While this gives intuitions about a unit's selectivity, it is still a very incomplete picture of its computations. Beyond selectivity, an important computational is invariance – the unit's tolerance for specific feature transformations. For example, the differentiating trait between simple and complex cells is not their selectivity (which could be the same), but invariance to phase shifts.

Recent approaches applied to units in the highest layers of a CNN have either visualized the top activating natural image patches (Erhan et al., 2009; Zeiler & Fergus, 2014) or synthesized and visualized maximally activating stimuli from different initializations (Mahendran & Vedaldi, 2016; Nguyen et al., 2016; 2017) to get an idea of invariance. We believe and show in this work, that these approaches may underestimate the true diversity of selectivity, even for early layer units. Other work (Goodfellow et al., 2009) studied invariances in deep neural networks by quantifying a unit's response to parametric changes in the stimulus (e.g. translation, rotation, or scaling), not allowing the discovery of novel transformations. Here, our goal was to visualize, characterize, and quantify the types of invariances learned in the early layers of CNNs trained on object classification by developing a method that effectively maps the manifold of highly activating inputs.

Results

We proposed a loss function with three terms that we maximize to simultaneously synthesize a batch of images that all highly activate a CNN's hidden unit. The first term is the sum of the unit's responses to the images, the second is a diversity term consisting of the minimum euclidean distance between any pair of images (or their representations), and the third is a natural image prior that acts as regularization. The later was Pixel CNN++ (Salimans et al., 2016), an autoregressive model trained to model the distribution of natural images that ultimately improved the perceptual quality of the images we synthesized.

We tested our approach on a toy simulation of a complex cell (with known phase invariance) and successfully recovered Gabor functions with equal tuning but different, well-spaced, phases. These results would have not been possible if we removed the diversity term and relied solely

on random initializations. We then applied our method to VGG-19 (Simonyan & Zisserman, 2014) and found that units in its early convolutional layers exhibit response invariances: there were diversity penalties that led to batches of images that all highly activated the unit with significant diversity. These invariances appear to be a learned property of the network – a network with matching architecture and random weights showed little activation tolerance for increases in diversity.

We identified two types of units based on visualized invariances: texture detectors and shape detectors with tolerance to localized shifts. Texture type of units showed invariance to global shifts of the same pattern. We quantified this invariance by synthesizing a larger image where overlapping patches optimized the unit's activity. We computed the ratio between the average activity of crops of this larger image and the average activity of the images synthesized with our first method (the templates). A higher ratio of this texture index implied tolerance for global shift changes. On the other hand, shape detectors showed invariance to local diffeomorphic transformations of the stimulus. We quantified this invariance by using a set of the convex hull of the template images. If a unit is tolerant to local changes, it shows higher tolerance for linear combinations of highly exciting images –even if the resulting images are very different at the pixel level with respect to the templates. We computed the ratio between convex-hull images and the templates. A high ratio of this shape index implied high selectivity for a shape, but tolerance for localized shifts. We found that these two metrics were anticorrelated across layers in VGG19. We further tested if these invariances generalize to other networks like Resnet50 (He et al., 2016) and found that it has a much fewer texture units compared to VGG, while most units exhibit higher tolerance to linear combinations of template images.

Finally we applied our method to a three-layer CNN from Cadena, Denfield, et al. (2019b) trained end-to-end to predict macaque V1 responses to natural images. Our method unveils known simple and complex cell types.

Discussion

Overall, we found that early layers of VGG19 exhibit invariance to global texture-preserving transformations and invariance to local shape-preserving transformations and quantified them. In contrast, Resnet50 did not show shift invariance. We hypothesized that this may be a potential reason why most texture synthesis (Gatys et al., 2015) and neural style transfer approaches (Gatys et al., 2016) rely mostly on VGG-like features instead of Resnet.

Our method is a promising tool for describing representations in biological neurons. Although complex cells can be identified using classic, specifically designed stimuli or analysis methods relying on quadratic features like spike-triggered covariance (Rust et al., 2005b), our non-parametric method could uncover other types of invariances beyond them. Since we don't observe such additional invariances, our results don't suggest other major features that macaque V1 could be invariant to.

3 DISCUSSION

Deep learning is an undeniably powerful tool for science at large. However, scientists are still finding out ways to derive meaningful insights from deep learning models trained on input-output pairs. In this dissertation, we presented five projects that address questions about neural processing in the visual cortex. While there are many different lenses through which our work can be viewed, we framed these projects and the questions they answer in three main research approaches: goal-driven modelling (RA1), embedding knowledge into model architectures (RA2), and analyzing trained models (RA3).

First, we addressed RQ1 and showed that models trained on object classification facilitate unprecedented predictions of neural responses in V1, with comparable results to models trained end-to-end to fit the data (Section 2.1). Since our results outperformed classical models of V1 including energy models based on Gabor filters, we wondered about the nature of the nonlinearities that deep predictive models learn beyond them. We hypothesized that nonlinearities associated to divisive normalization were a good candidate and built end-to-end data-driven models that explicitly implement this nonlinearity, addressing RQ5. We found that DN models did not outperform CNNs, but they did outperform classical models, halving the gap between them and CNNs with far less parameters (Section 2.4). Moreover, with the help of *in-silico* experiments, we addressed RQ6 and found that both DN and CNN models (but not classical GFB models) captured cross-orientation inhibition, suggesting that our natural image stimuli elicit divisive normalization nonlinearities.

We also addressed RQ2 following similar goal-driven methods and found no correspondence between CNN layers and areas in the mouse brain in contrast to results in primates (Section 2.2). While we could not rule out a potential hierarchical organization across mouse visual areas, our results did suggest that object categorization may not be the right ethological task. This was also supported by the comparably good performance of nonlinear random features.

Then, we addressed RQ3 by characterizing the functional role of macaque area V4 from a normative point of view, involving relevant visual tasks beyond object classification (Section 2.3). We found that V4 is indeed mainly explained by semantic tasks, but that it still bears significant similarity with networks trained on other tasks. To address RQ4, we trained models that jointly use features from pairs of tasks and found that non-semantic tasks do capture additional nonlinearities beyond those captured by individual semantic tasks. However, pairs of semantic tasks alone perform comparably well as the best pairs involving other tasks. These results strengthen the semantic role of V4, but also explain V4 affinities to 2D and 3D stimulus features that have been characterized before. Our results motivate future work relating individual neuron tuning properties and high-level functional goals.

Finally, we addressed RQ7 and developed a method to visualize what CNN units become be invariant to (Section 2.5). We used this method on popular CNNs and unveiled invariances to global pattern shifts (texture-like selectors), and local shape-preserving shifts. These methods revealed simple and complex cells in monkey V1, and carried significant implications for ongoing work characterizing more complex invariances existing in mouse visual areas, and macaque area V4 (Section 3.3).

In this chapter, we focus on three overarching topics that have rapidly evolved since the publication of our work and are shaping the ongoing research agenda of computational neuroscientists. We discuss major criticisms to task-driven approaches (Section 3.1), the current conclusions about task-driven modeling of mouse visual areas for their functional characterization (Section 3.2), and how (diverse) feature visualizations are helping us characterize CNN units and neurons in the brain (Section 3.3).

3.1 CHALLENGES AND LIMITATIONS OF GOAL-DRIVEN APPROACHES

Three of the research papers presented in this dissertations (Sections **??**) relied on a task-driven approach (Section 1.4.1) to model visual representations in the brain. Although many of these models proved useful by providing unprecedented accurate predictions of neural responses to natural stimuli (Yamins & DiCarlo, 2016), and enabling the synthesis of stimuli that maximally drives target neurons (Bashivan et al., 2019; Ponce et al., 2019; Willeke et al., 2023), there are several limitations shared by studies that follow this line of work.

The reliance on ever-evolving ML methods and resources

A limitation of studies using task-trained deep neural networks for neural system identification lies in the reliance on the contemporary state of machine learning (ML) resources and technologies. The conclusions drawn from these studies are intricately tied to the specific architectures and datasets available during the investigation, raising the question of what might have transpired if more advanced or tailored resources were employed (although there are a few ongoing attempts to alleviate this (Kubilius et al., 2018; Mehrer et al., 2021; Qiu et al., 2021)). The dynamic nature of ML research suggests that unanswered questions persist regarding potential outcomes with improved methodologies. Acknowledging this temporal dependency is crucial, as future advancements in ML could provide new insights, prompting a reevaluation of conclusions in light of evolving technologies. For example, recent studies suggest that performance-optimized object recognition neural networks are evolving into worse models of IT cortex (Linsley et al., 2024), suggesting that there are factors at play of the design of DNN models used in earlier studies that are critical beyond the training objective.

The combinatorial maze of factors giving rise to DNNs

Task-driven DNNs are characterized by a multitude of factors, including the specific dataset used for training, architectural design, training objectives, initialization methods, and learning algorithms, among others. The combinatorial explosion of these factors results in an expensive and nuanced space of possible configurations, making it impractical to comprehensively explore all permutations. Many studies, driven by the availability of existing machine learning models within the community, often focus on specific facets of DNNs, such as training objectives, while letting other factors vary. The selective exploration of the parameter space can therefore raise concerns about the generalizability of conclusions as they may be confined to specific combinations of factors, limiting the broader applicability of findings. For example, in the work presented in Section 2.3, we made an effort to keep multiple variables constant when comparing a single factor (i.e. training objective), but it is still unknown how our conclusions would generalize to configurations with other values (e.g. architecture changes). While insights into individual components of DNNs are valuable, the challenge lies in disentangling the joint effects of various factors, emphasizing the need for future research to address the complex interplay within the diverse landscape of DNN configurations.

Task-driven models behave differently than human perception

A prominent criticism of task-trained DNNs as models of the visual system is the growing body of evidence highlighting differences in decision-making strategies. These differences span the sensitivity of DNNs to image manipulations like noise (Hendrycks & Dietterich, 2019) or adversarial perturbations (Goodfellow et al., 2014), the DNNs bias towards textures vs. shapes cues (Geirhos et al., 2019), and the recurrent learning of shortcuts due to under-specification of computational tasks (Geirhos et al., 2020). Additionally, DNNs fail to replicate the majority of psychophysical measurements of human visual perception, as noted by Bowers et al., 2023. For example, these models struggle with processing objectness –a fundamental aspect of visual perception where objects are perceived as distinct entities from their background. These disconnects not only show the lack of generalizability of task-trained models but also raise concerns about their ability to develop a causal understanding of the visual scenes.

However, there is growing interest in reconciling the behavioral pessimism of DNNs with the neural predictivity enthusiasm they have generated. Despite their limitations, (task-trained) DNNs are our current single best tool for modeling visual sensory processing. Several researchers argue that (deep) image-computable models are rather a practical framework for expressing computational hypothesis and for generating testable predictions that can continue to improve (de Beeck & Bracci, 2023; DiCarlo et al., 2023; Golan et al., 2023; Wichmann et al., 2023; Yovel & Abudarham, 2023). In general, most model deficiencies are due to their training and evaluation in environments that are unrepresentative

of the real-world complexities encountered by biological agents. Highlighting this deficiencies serve rather as opportunities for refining our benchmarks and approaches to continue to make progress. Example research directions involve training on more human-like tasks (Hermann et al., 2023), improving training data and learning algorithms (Linsley & Serre, 2023), constraining models for targeted scientific understanding (Bernáez Timón et al., 2023), and including actions and interactions with the environment for perception (Rothkopf et al., 2023).

3.2 THE FUNCTIONAL ORGANIZATION OF MOUSE VISUAL AREAS

Following current success on primate data, we tried to identify a functional organization of mouse visual areas (V1, LM, AL, RL) with goal-driven models trained on image classification (RQ2, Section 2.2) and found no strong signature for hierarchical organization. However, studies based on circuit tracing and response delay methods suggest that mouse visual areas may indeed be organized in a hierarchical manner (D'Souza et al., 2022; Siegle et al., 2021; Wang et al., 2012), although in a much shallower and interconnected way than the primate visual system. In our work, we discussed that our findings did not exclude the existence of a hierarchy, but rather pointed to the failure of the model we used (VGG16 trained on ImageNet) to learn relevant features for the mouse visual system in contrast to primates. This conclusion was also corroborated by our results showing that VGG16 features exhibited similar or only marginally superior performance compared to a randomly initialized counterpart.

Concurrent papers using VGG16 to model mouse visual areas (de Vries et al., 2020; Shi, Shea-Brown, & Buice, 2019) and more recent studies (Conwell et al., 2021; Nayebi et al., 2023; Shi & Malik, 2000; Vinken & Op de Beeck, 2021) have largely found similar conclusions to ours: there is a much weaker (if any) signature of hierarchy along areas V1, LM, AL, RL (Vinken and Op de Beeck (2021) did found a *mild* one in rats but along areas V1, LM, LI, TO); the best match to any of these areas is with mid-level layers of pretrained CNNs instead of high layers like in primate IT, and deep ImageNet-trained models only show weak improvements over random baselines in stark contrast to primates.

Why do ImageNet-trained deep models fail to be a good model of the mouse visual system in contrast to primates? Follow-up work has shown that the main reasons are 1) model complexity, 2) high input resolution data streams, and 3) sub-optimal training objective (object classification with thousands of classes foreign to a rodent). Nayebi et al. (2023) and Shi and Malik (2000) found that training shallower models that may enjoy architectural similarities to the target visual system (e.g. recurrence) are a better match than deep CNNs. Moreover, Vinken and Op de Beeck (2021) show that early to mid-level layers (and not high layers like in primates) of pretrained CNNs are the best at explaining both neural data and observed behavioral experiments (e.g. Djurdjevic et al. (2018), Vinken et al. (2014), and Zoccolan et al. (2009)), suggesting that shallower models may be more adequate. Leveraging behavioral findings of the comparably low rodent visual acuity of ~ 0.5 cycles per degree (Prusky

et al., 2000), Nayebi et al. (2023) found that reducing the image resolution used for pretraining CNNs yielded strong performance improvements. Finally, Conwell et al. (2021) and Nayebi et al. (2023) identified that performance on ImageNet and Taskonomy (Zamir et al., 2018) datasets did not correlate with mouse neural predictivity. Nayebi et al. (2023) found that contrastive self-supervised training objectives far outperform object categorization and offer state-of-the-art predictions of mouse visual areas.

Overall, goal-driven approaches only weakly revealed hierarchical complexity gradients in the areas we considered. Self-supervised objectives are the most promising direction to both improve the predictive performance of neural data, and reveal functional differences across mouse visual areas. Given the evident success of using mouse-derived knowledge to improve task-driven models, the likely next big bet would be to train models on data streams that match much better the statistics of actual natural stimuli experienced by rodents. For example, Qiu et al. (2021) identified upper and lower visual field specializations in the mouse retina using recordings of mouse habitat with cameras capturing dichromatic images (UV and green channels). These stimuli unlock nonlinearities that have already facilitated novel functional characterizations of mouse V1 (Franke et al., 2022).

3.3 FEATURE VISUALIZATIONS AND INVARIANCES IN CNNS AND VISUAL CORTEX

In an effort to understand the learned invariances of trained CNNs, we developed a method to efficiently sample the stimulus manifold that maximally excites CNN units by simultaneously optimizing a batch of images with a diversity term (Section 2.5). We applied this method to early to mid-level layers of popular ImageNet-trained CNNs and unveiled invariances to various global pattern shifts (textures), and local shape-preserving shifts. Although our worked focused on further characterizing these two invariances, our visualizations qualitatively revealed a rich set of feature selectivity that more often than not carried some low-level semantic preferences like color, orientation, curvature, patterns, *eye-like* features, and more.

The remarkable concurrent work by Olah et al. (2017) revealed vastly similar-looking features and –with the help of interactive posts https://distill.pub/2017/feature-visualization/ and tools like https://microscope.openai.com/– kick-started efforts in the community to improve the interpretability of CNNs. Olah et al. (2017) also simultaneously optimized multiple *facets* of units (or more often entire channels) with a different diversity term than ours involves the similarity between layer Gram matrices of input pairs (resembling methods for texture synthesis (Gatys et al., 2015)). From these diverse visualizations they concluded that individual facet may not be helpful enough to understand selectivity of a unit at a semantic level (a single image may suggest preference for a *"dogs head"*, but multiple visualizations suggest selectivity for *"dog fur"*). Moreover, they found that many units have mixed selectivity between seemingly unrelated stimulus features (e.g. a unit that responds to both

cats and *cars*), and thus concluded that neurons are not necessarily the right semantic units for understanding neural nets (Olah et al., 2017). These observations align with our visualizations of random directions (linear combinations of neurons) yield similar-looking kinds of feature visualizations as individual CNN units.

Olah, Cammarata, Schubert, et al. (2020b) further proposed that although the field of interpretability and visualizations is largely qualitatively, we can rigorously study and understand three main facts about neural networks: features (any direction as a fundamental unit), circuits (weighted connections), and universality (analogous features across models and tasks). Universality bears special significance for studying visual representations in the brain as it allows us to identify features present both in ANNs and the visual system. A clear example is oriented Gabor-like feature detectors that emerge under a wide range of tasks that have also been identified in the brain.

What other features may be universal to support vision? Across ImageNet-trained CNNs, our work and Cammarata, Carter, et al. (2020) points to the prevalence of at least curve detectors (Cammarata, Goh, et al., 2020) and high-low frequency detectors – a family of early-vision neurons reacting to directional transitions from high to low spatial frequency (Schubert et al., 2021). In Ding et al. (2023), we applied similar methods to visualize diverse facets of mouse V1 neurons and found striking similarities with feature characterizations of CNNs. In particular, we found that the neurons' receptive fields can in general be characterized as having bipartite invariance: one portion of the receptive field tolerates phase shifts in a texture-like pattern, the other one prefers a fixed pattern. While high-low frequency in CNNs do show tolerances to shifts in both high and low parts of their receptive field, the qualitative similarity of mouse V1 features and high-low frequency detectors in CNNs suggest that these features carry universal computational importance. Future directions involve understanding how these invariances are achieved at the circuit level (do these match between CNNs and the mouse brain?), and the high-level computational goal they serve (e.g. foreground-background segmentation of naturally occurring videos?).

In addition to feature visualizations of mouse V1 neurons (Ding et al., 2023; Walker et al., 2019), feature visualizations have been applied to monkey V4 neurons (Bashivan et al., 2019; Willeke et al., 2023). In Willeke et al. (2023), we showed that many of the features found by Olah, Cammarata, Schubert, et al. (2020b) are detected by neurons in V4 (e.g. oriented fur, eyes, circles, curves, center-surround textures, boundaries). We also used diverse feature visualizations as a way to learn meaningful embedding space of highly exciting stimuli for all neurons. This embedding space was learned with a contrastive loss (Böhm et al., 2022) that tried to bring together the embeddings of diverse highly exciting stimuli for the same neuron, and pushed away embeddings from other neurons. This representation space facilitated downstream clustering and characterization of neuron types in V4 and suggested a potential topographic organization of feature preference based on columns.

Although feature visualization have proven useful to synthesize brain-relevant stimuli that can drive responses maximally, they have faced

criticism as a tool for interpretability (Borowski et al., 2020; Geirhos et al., 2023; Zimmermann et al., 2021). Using psychophysical experiments, Borowski et al. (2020) and Zimmermann et al. (2021) have shown that although they provide some information, humans often struggle to make sense of visualizations and that highly (or minimally) exciting dataset samples can be equally or more informative. Moreover, Geirhos et al. (2023) evaluated whether feature visualizations are reliable (i.e. trustworthy) and found that feature visualizations can be arbitrarily fooled, can be processed differently from natural images (i.e. follow different paths in the network), and can only be guaranteed to be reliable if we know already a lot about the network. An important research agenda moving forward is to develop methods for reliable feature visualizations (not necessarily tied to activation maximization) that facilitate interpretations of computations in artificial and real neurons.

4 OUTLOOK

4.1 HOW CAN WE BUILD BETTER MODELS OF PRIMATE V1?

Olshausen and Field (2005) argued that roughly 85% of the variance in V1 responses to natural stimulation was still left unexplained by the best models at the time. Our best CNN models improved over that baseline and explained up to 54% of the explainable variance of spike-sorted extracellular responses of macaque V1 neurons to gray scale natural images. Predictive models on other monkey V1 dataset benchmarks like the Freeman et al. (2013) on Brainscore (Schrimpf et al., 2018) showed similar results with top models achieving 0.59 scores. This suggest that the $40-50\%$ of the variance that is left unexplained by our current best goal-driven models transcend idiosyncrasies of particular experimental setups, and may involve meaningful stimulus-driven nonlinear functions that these models fail to capture.

In order to continue making progress on better V1 models, the field requires established benchmarks that help us measure progress. Examples like Brainscore (Schrimpf et al., 2018) and the Sensorium competition (Willeke et al., 2022) bring to system neuroscience a benchmark-driven mindset that has enabled fast progress in the machine learning community. Moreover, prediction benchmarks can be appended with the ability of models to capture known V1 properties (e.g. Marques et al. (2021)) as it is currently done in Brainscore. The multi-axis view of an ability of a model to predict neural responses and capture known properties could amount evidence for or against a model as a better match for V1.

In this section, I give a high-level overview of promising research directions for better generalizing models of primate V1. These attempts can fall into the known goal and data-driven perspectives and suggest ideas for better data streams, model architectures, and objectives.

Goal-driven perspective

There are three main dials that underpin the search space of models: training objective, data stream, and model architecture. However, current evidence suggest that data stream is likely the most promising avenue for continued improvements. First, several diverse computer vision tasks (conditioning on data diet and model architecture) (Cadena et al., 2024), and self-supervised losses (Zhuang et al., 2021) yield competitive V1 performance. Second, architectures that are more “brain-like" including recurrence connections (e.g. (Kubilius et al., 2018)) do not outperform purely feed-forward counterparts like Renset or VGG networks (Cadena et al., 2024; Schrimpf et al., 2018).

On the other hand, there is a strong trend that favors data-rich models –Imagenet models widely outperform CIFAR or Taskonomy-trained models– that additionally include relevant data-augmentations tolerated

by V1 responses, like certain adversarial perturbations (Cadena et al., 2024; Dapello et al., 2020; Kong et al., 2022). The power of scale of data-rich models is also supported by studies that compare CNNs and the brain at the behavioral level, concluding that models trained on web-scale datasets partially close the gap between artificial neural networks and the brain on relevant behavioral benchmarks (Geirhos et al., 2022).

Data-driven perspective

LARGER DATASETS We found that our best end-to-end trained models do not outperform task-driven models – even when architectures match. This suggests that our combination of data, training methods, and objective fail to find better optima despite our regularization schemes. A straightforward way to improve data-driven models is to simply continue to collect more data – both in the number of neurons and stimuli. For example, novel technologies like Neuropixels (Steinmetz et al., 2021) can facilitate the recording of many more neurons at once. However, we can continue to include neurons from many sessions and animals, as it has been shown –at least in mice (Lurz et al., 2021)– that involving multiple sessions and animals increase a model's ability to generalize to unseen neurons and sessions. Even more, there should be efforts to train models on multiple datasets that are collected across experimental labs – provided that any "batch effects" are carefully accounted for (e.g. (Gonschorek et al., 2021)). Going beyond the benchmark and evaluation of pretrained models like Brainscore, a community effort that also organizes and open source data from multiple sources are likely effective ways to converge to better models of V1.

CAPTURE CENTER-SURROUND INTERACTIONS In Burg et al. (2021), we normalized the output channels of the *core* with a weighted sum of all channels. We proposed an extension of this approach where a channel's normalization given by a (dilated) convolution between a filter and the activations of other channels. This approach can potentially capture spatial relationships that alter normalization strengths that can facilitate replicating center-surround effects (Heeger, 1992b) in addition to interactions within the classical receptive field. A limiting factor that prevented us from learning these interactions was the visual field coverage of our stimuli. Future datasets can focus on the sufficient stimulation of the surround to elicit center-surround effects that may be better learned with divisive normalization mechanisms.

UNIFIED MODELS THAT SIMULTANEOUSLY CARRY INDUCTIVE BIASES Architectures that reflect inductive biases can make learning more efficient (i.e. require less training data) as nonlinearities are explicitly implemented and the number of required parameters is lower than more generic models. Until now, multiple efforts tend to rather implement a type of equivariance or nonlinearity in isolation (beyond the standard translation equivariance inherit to CNNs) and showed their efficiency in learning good representations. For example, in Burg et al. (2021) implemented divisive normalization nonlinearity, Ecker et al. (2018) and

Ustyuzhaninov et al. (2019) implemented rotation equivariance, Wang et al. (2023) implemented perspective transformations (image warping) and recurrent connections. Bring together these architectural constraints into single unifying architectures could add their independent benefits and unlock potential constructive interactions between them. That said, finding optima of constrained architectures can be challenging for any random initialization –as shown by the discrepancy between distilling smaller networks from a larger network and training an equivalent small network from scratch, also referred to as the lottery ticket hypothesis (Frankle & Carbin, 2018). This issue can be potentially addressed by training separate models, each carrying a certain bias, and then initializing the weights of a unified model with them. Alternatively, teach-student networks techniques where trained large, over-parameterized network guide the training of smaller, constrained models via additional losses that include intermediate layer outputs (Hinton et al., 2015).

EXPLOIT THE FLEXIBILITY OF TRANSFORMER ARCHITECTURES The ultimate goal of constructing models of V1 should extend beyond merely beating benchmarks or focusing solely on engineering. Ideally, these models should also offer meaningful explanations (Section 1.4). However, it could be valuable to investigate whether novel deep learning architectures—despite being mechanistically more distant from the brain than CNNs—might enhance performance compared to existing models. Transformer architectures (Vaswani et al., 2017) have proven to be a versatile architecture capable of approximating functions across modalities including vision, natural language, and audio signals. With increasing dataset sizes, training transformers to predict V1 responses could help us make significant strides and potentially uncover unknown V1 mechanisms via post-training dissection methods. Recent approaches have already trained transformer architectures to predict mouse V1 data (Li, Cornacchia, et al., 2023) with successful results in competition benchmarks (Willeke et al., 2022). Other work applied to typical rodent experimental settings, trained multi-modal, multitask, generative pretrained transformers (GPT, (Radford et al., 2019)) using simultaneously recorded signals like locomotion and eye tracking to predict future signal *tokens* in an autoregressive matter (Antoniades et al., 2023). Similarly, the inclusion of multiple data modalities, paired with relatively large dataset sizes have paved the way for so-called *foundation models* of spike responses in motor cortex too (Azabou et al., 2024; Ye et al., 2024). These flexible, generic architectures applied to primate visual can potentially verify or falsify the effectiveness or current inductive biases that are ubiquitously used in the field. Finally, another argument supporting this research direction is the improved match between transformer networks trained on vast datasets and brains on behavioral benchmarks (Dehghani et al., 2023; Geirhos et al., 2021)

FIT MULTIPLE BRAIN AREAS SIMULTANEOUSLY Brainscore evaluates pretrained models' match to the brain with a series of benchmarks comprising multiple areas (Schrimpf et al., 2018). Going beyond evaluation, we could use training data from different visual areas to learn

a single model that simultaneously fits them well. Intuitively, the space of possible model representations required to fit visual responses will be constrained if the same representations need to be useful to ultimately fit multiple sets of responses. Moreover, extensive knowledge from anatomical studies (e.g. Felleman and Van Essen (1991)) can inform the architecture of models and the relationship between areas. For example, top-down connections between area V4 and area V1 could be captured with recurrent connections like in Kubilius et al. (2018) or with U-net architectures (Ronneberger et al., 2015). Furthermore, readouts that solve high-level goals likeobject classification or self-supervised objectives can be placed at the top of the network to enforce representations that are functionally useful. This promising idea has been partially pursued in Safarani et al. (2021) where we showed that a network trained to classify images and simultaneously predict macaque V1 responses from its intermediate layers results in better robustness on image classification.

Diversify stimuli to unlock more V1 nonlinearities

We've discussed opportunities from goal and data-driven perspectives for better models of V1 in our in-domain natural image test set. Importantly, while we've made great progress moving away from simple, parametric stimuli to natural images, it is unlikely that our limited, static, gray-scale, natural image set constrained to a few degrees of visual angle will unlock all nonlinearities present in V1. Promising novel stimuli that can elicit stronger responses in neurons involve video, a wider dynamic range of stimulus contrast, full field stimulation, and colored natural images (video). Experimental sessions could mix varied sets of stimuli that also include standard gray-scale images and classical gratings/plaids. Doing so, could also facilitate the evaluation of out-of-domain generalization of V1 models and further constrain the representations they learn.

4.2 HOW CAN WE BETTER DISCRIMINATE MODELS?

A prevalent observation made by collaborators and I is that starkly different models can make similar predictions, yielding small differences in test-set performance. For example, we found in Cadena et al. (2024) that models trained on very different Taskonomy tasks can lead to comparable test performance values when predicting both V1 and V4. How robust is the similarity of these representations to the brain and to each other? It may be the case that for the domain – or even the reduced set of test stimuli – these models behave similarly, but they could employ different strategies (functions) that behave fundamentally different under different circumstances. Here, I discuss a few ideas that have the potential to adjudicate among equally performing models.

Out-of-distribution evaluations

To adjudicate among equally performing models, we require stimuli for which they make distinct predictions (Golan et al., 2020). One way to

improve chances of finding these stimuli is to draw samples from a different distribution not used for model training. For example, if the training set domain involves gray-scale natural images, part of the test set could include textures, parametric stimuli, common corruptions and image perturbations (Hendrycks & Dietterich, 2019), white noise, etc. Shifting the distribution for model evaluation will likely reveal mismatches between models and the brain that can discriminate between models and further reveal failure points that a next iteration of models should address. When comparing image classification of CNNs and human behavior, out-of-domain (OOD) evaluations have revealed important differences (Geirhos et al., 2019; Hendrycks & Dietterich, 2019) that have triggered wide range of follow up work to make models better (Geirhos et al., 2021). While some data-driven models trained on natural video have been shown to reproduce *in-silico* tuning properties to parametric stimuli (Sinz et al., 2018), other work has shown that many models fail to generalize to OOD settings. For example, Ren and Bashivan (2023) found that the OOD abilities of image classification CNNs does not necessarily translate to better OOD neural predictivity properties, triggering questions about the nature of future task and data-driven models with strong generalization abilities. Further work in this direction and novel neural system identification competitions like the Dynamic Sensorium competition (Turishcheva et al., 2023) will strongly narrowing down the space of candidate models and help adjudicate among them.

Close loop experiments paired with controversial stimuli

Another way to adjudicate between models is to synthesize maximally activating inputs (MEIs) for any given neuron and then showing them back to the animal in a closed-loop fashion (Bashivan et al., 2019; Walker et al., 2019). Model-derived MEIs for which neurons respond more strongly can favor a model among others (Willeke et al., 2023). However, MEIs from different models could be qualitatively similar and yield responses that are hard to differentiate due to trail variability. An alternative approach that also uses gradient-based optimization is to synthesize *controversial* image sets for which different models make maximally different predictions (Golan et al., 2020). Golan et al. (2020) compared pairs of models trained to recognize hand-written digits by optimizing inputs that maximize a controversiality score that compels one model to classify the input as a digit different from the other model. This method facilitated the discrimination between equally accurate models based on how well they matched human perception on controversial stimuli. In a similar way, controversial stimuli can be synthesized to discriminate neural predictive models by maximizing the predictions of one model while suppressing predictions of the other model for the same neuron. These stimuli can be ideally tested for the exact same recorded neurons in close loop, but if there are experimental constraints that prevent the immediate evaluation of such stimuli, controversial stimuli for *cell-types* can be created and evaluated in future sessions instead. Recent successful results byBurg et al. (2024) relying on most discriminative stimuli for cell-types (not models) support the promising potential of synthe-

sized controversial stimuli to discriminate competitive neural predictive models.

In-silico experiments targeting known properties

A less direct way to adjudicate between models offering competitive predictions is to evaluate how well they replicate known phenomena that has been previously characterized with neurophysiology experiments. As described in Section 2.4, we were able to adjudicate between the DN and GFB models, which had a small performance gap, with *in-silico* experiments that evaluated their ability to reproduce cross-orientation inhibition (Carandini & Heeger, 2012). Despite small performance differences, we could confidently favor the DN model over the GFB, because it captured this important V1 nonlinearity. An encouraging approach to differentiate between competing models involves the development of benchmarks that assess the models' capacity to accurately reproduce established phenomena. Marques et al. (2021) has moved in this direction, establishing a set of properties that good models of V1 should replicate and it is now available in Brainscore (Schrimpf et al., 2018). For example, in V4 there is a wealth of single-cell tuning characterizations to different 2D boundary, texture, and solid shape properties (Pasupathy et al., 2020; Pospisil et al., 2018; Srinath et al., 2021) that can be turned into measurable benchmarks for models to reproduce in addition to test-set predictions. Moreover, these *in-silico* characterizations can be turned into hypotheses for the characterizations of cell types in the brain that could be later validated *in-vivo* (Ustyuzhaninov et al., 2022).

Model interventions

A nascent research direction in neural system identification that will improve singling out better models of the visual cortex is to model perturbations that causally link neural responses to behavior. In this setting, behavioral reports (e.g. face-related judgements, orientation estimates, glossiness, and object category reports) are measured after a group of neurons' activity is directly perturbed by methods that include micro-stimulation, optogenetic suppression, or muscimol suppression (Schrimpf et al., 2024). Schrimpf et al. (2024) developed perturbation modules capturing these different perturbation methods, and derived from existing experimental results a set of benchmarks that candidate models should be able to replicate. They found that a candidate topographic artificial neural network was able to qualitatively predict the effects found experimentally (e.g. Afraz et al. (2015), Afraz et al. (2006), Moeller et al. (2017), and Rajalingham and DiCarlo (2019)) but that further ideas are needed to quantitatively predict all results. These benchmarks, along with comparable research efforts, offer an intriguing and innovative approach to assess and identify possibilities for enhancing candidate models of the visual cortex.

4.3 HOW CAN WE BETTER INTERPRET DATA-DRIVEN REPRESENTATIONS?

Characterize neuron embeddings

Although the computations approximated by neural predictive models may be hard to interpret, they can be used to extract neuron embeddings – bar codes that summarize the complex input-output functions of a neuron. In contrast to traditional functional characterizations based on tuning properties to various parameters of simpler stimuli, neuron embeddings provide a low-dimensional, holistic representation of the function of a neuron in response to natural stimuli. With a rich set of neurons, these embeddings facilitate describing the functional landscape of visual neural responses via clustering and visualizations of 2D projections of all neurons (Tong et al., 2023; Ustyuzhaninov et al., 2022).

A critical step for learning meaningful neuron embeddings is to remove nuisance parameters that would otherwise dominate the feature landscape and complicate the identification of potential cell-types. For example, simple and complex cells that share the same orientation preference could be clustered together as the same type in a learned embedding space if orientation preference is not disassociated from these bar codes. In Ustyuzhaninov et al. (2022), we learned a feature space with equivariance to translation (Klindt et al., 2017) and rotation (Ecker et al., 2018), and fitted readouts to predict individual neuron responses. These factorized readouts were then aligned to a single orientation to achieve rotation invariance, yielding a feature vector description of each neuron (Ustyuzhaninov et al., 2019). These neuron embeddings facilitated the characterization of mouse V1 in the cluster vs. continuum spectrum and revealed a *combinatorial coding* of various classical tuning properties across different cell-types (or distribution modes) with the help of *in-silico* experiments (Ustyuzhaninov et al., 2022).

A similar study going beyond mouse V1 to other visual areas, synthesized maximally activating images (MEIs) of each neuron and learned an embedding space of these images (Tong et al., 2023). Although this work focuses on the stimulus mode for feature landscape characterization, the contrastive loss (Chen et al., 2020) used to learn the embedding space of neurons forces similar mappings for affine transformations of the MEIs, forcing the embedding space to account for underlying symmetries like rotation. The authors then used these embeddings to compute a matrix of local overlap between stimulus manifolds corresponding to each visual area, and derived a network structure that captures the functional relationship between visual areas (Tong et al., 2023).

Neuron embeddings can thus be powerful ways to get a holistic view of the organizational map of visual cortex. Work on two main research questions are likely going to continue to be influential in this line of work. First, how can we better characterize embeddings? *In-silico* experiments replicating known phenomena (Ustyuzhaninov et al., 2022), and (diverse) highly exciting inputs are prominently used already (Ding et al., 2023; Tong et al., 2023; Willeke et al., 2023). Novel characterizations can gain inspiration from work in the retina where other data modali-

ties like gene expression and neuron mythologies have been linked to function (Baden et al., 2016). Recent large-scale datasets collecting functional and morphology data of the same neurons (e.g. Consortium et al. (2021)) and advances in learning relevant morphology embeddings (Weis et al., 2023) will likely facilitate in the near future learning multimodal embedding representations (similar to CLIP (Radford et al., 2021)) that combine structure and function. Second, what other symmetries exist in the visual system and how can we embed them in our models? Learning symmetries from data is an active area of research (Alet et al., 2021; Dehmamy et al., 2021; Neyshabur, 2020; Zhou et al., 2020) that can potentially reduced the amount of data needed to learn better models and improve the quality of neuron embeddings. Feature visualizations and circuit analyses of trained CNNs on image classification have revealed various naturally occurring equivariances like rotation, scale, hue, hue-rotation, and reflections among others (Olah, Cammarata, Voss, et al., 2020). Baking these equivariances into the architecture of predictive models and assessing their resulting accuracy could generate testable hypotheses of symmetries captured by the brain and yield better neuron embeddings.

Generative modeling of neural responses

The work presented in this dissertation follows a *discriminative* perspective of vision where the computational goal of the visual system is to quickly and directly extract ethological relevant latent dimensions (e.g. object category, orientation) from sensory input. An alternative perspective of vision is the *generative* framework that is motivated by the fact that sensory information is ambiguous and argues that, as a result, perception is not only driven by observations but also a model of the world (Von Helmholtz, 1867). More formally, the brain encodes a probabilistic generative model of the world given by the joint distribution of sensory input and latents that represent meaningful features or objects in the world that cause sensory input. In this framework, neural activity and behavior are viewed as mappings from these latent factors.

Recent perspectives argue that the primate visual system may exploit advantages of both frameworks (e.g. speed of discriminative and stimulus disambiguation of generative) with solutions that lie on a spectrum from the purely discriminative to the purely generative (Peters et al., 2024). Given the evidence and usefulness of generative approaches that go beyond any dichotomies, the value of meaningful latent directions learned by generative models is a very promising avenue of research that is currently under-explored in the field of neural system identification.

Early work in this direction modeled the conditional distribution of stimulus given spike vs. no spike in a short time window as a mixture of Gaussians (Theis et al., 2013). More modern generative models using deep learning like variational autoencoders (VAEs, (Kingma & Welling, 2013)) with distinct regularization schemes encouraging disentangled latent representations (Higgins et al., 2016) demonstrated a strong correspondence between generative factors and neuron responses in infer-

otemporal (IT) cortex (Higgins et al., 2021). Beyond making better sense of latent factors, generative models facilitate the systematic generation of stimuli for any given perturbation of latent factors that can be used as counterfactual for future *in-vivo* experiments. These visualizations could provide more interpretable explanations of latent factors and – provided a mapping – of neural responses, that go beyond the maximally activating stimulus. Recent years have seen an explosion of generative approaches in the domain of vision with impressive results based on diffusion models (Sohl-Dickstein et al., 2015), autoregressive models (Van Den Oord et al., 2016), or normalizing flows (Rezende & Mohamed, 2015) are promising ways to generate testable stimuli that could be further confirmed experimentally.

Strip down neuron models into simpler circuits

A promising approach to understanding the computations buried in millions of parameters of deep neural predictive models is to derive *compact* models with much fewer parameters that retain accuracy (Cowley et al., 2023). These simpler models can facilitate the identification of circuit or circuit motifs that are responsible for particular nonlinear phenomena (Maheswaranathan et al., 2023). Moreover, reduced circuits paired with feature visualizations can enable a compositional understanding of how complex feature detectors emerge from simpler parts. For example circular feature detectors emerge from oriented curve detectors that are themselves the combination of various organized edge detectors (Cammarata, Carter, et al., 2020; Olah, Cammarata, Schubert, et al., 2020a). Circuits can be more rigorously studied and understood and can help bridge the gap between traditional mechanistic explanation of neural phenomena and generic deep predictive models that account for a diverse set of input-output pairs.

BIBLIOGRAPHY

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al. (2016). Tensorflow: A system for large-scale machine learning. *OSDI*, *16*, 265–283 (cit. on p. 7).

Adelson, E. H., & Bergen, J. R. (1985a). Spatiotemporal energy models for the perception of motion. *Journal of the Optical Society of America A*, *2*, 284–299. https://doi.org/10.1364/JOSAA.2.000284 (cit. on p. 19).

Adelson, E. H., & Bergen, J. R. (1985b). Spatiotemporal energy models for the perception of motion. *JOSA A*, *2*(2), 284–299 (cit. on p. 1).

Afraz, A., Boyden, E. S., & DiCarlo, J. J. (2015). Optogenetic and pharmacological suppression of spatial clusters of face neurons reveal their causal role in face gender discrimination. *Proceedings of the National Academy of Sciences*, *112*(21), 6730–6735 (cit. on p. 44).

Afraz, S.-R., Kiani, R., & Esteky, H. (2006). Microstimulation of inferotemporal cortex influences face categorization. *Nature*, *442*(7103), 692–695 (cit. on p. 44).

Alet, F., Doblar, D., Zhou, A., Tenenbaum, J., Kawaguchi, K., & Finn, C. (2021). Noether networks: Meta-learning useful conserved quantities. *Advances in Neural Information Processing Systems*, *34*, 16384–16397 (cit. on p. 46).

Andermann, M. L., Kerlin, A. M., Roumis, D. K., Glickfeld, L. L., & Reid, R. C. (2011). Functional specialization of mouse higher visual cortical areas. *Neuron*, *72*(6), 1025–1039 (cit. on p. 21).

Antolík, J., Hofer, S. B., Bednar, J. A., & Mrsic-Flogel, T. D. (2016). Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. *PLOS Comput Biol*, *12*(6), e1004927 (cit. on pp. 6, 19).

Antoniades, A., Yu, Y., Canzano, J., Wang, W., & Smith, S. L. (2023). Neuroformer: Multimodal and multitask generative pretraining for brain data. *arXiv preprint arXiv:2311.00136* (cit. on p. 41).

Ayzenshtat, I., Jackson, J., & Yuste, R. (2016). Orientation tuning depends on spatial frequency in mouse visual cortex. *eneuro*, *3*(5) (cit. on p. 21).

Azabou, M., Arora, V., Ganesh, V., Mao, X., Nachimuthu, S., Mendelson, M., Richards, B., Perich, M., Lajoie, G., & Dyer, E. (2024). A unified, scalable framework for neural population decoding. *Advances in Neural Information Processing Systems*, *36* (cit. on p. 41).

Baccus, S. A., & Meister, M. (2002). Fast and slow contrast adaptation in retinal circuitry. *Neuron*, *36*(5), 909–919 (cit. on p. 6).

Baden, T., Berens, P., Franke, K., Rosón, M. R., Bethge, M., & Euler, T. (2016). The functional diversity of retinal ganglion cells in the mouse. *Nature*, *529*(7586), 345–350 (cit. on pp. 11, 46).

Bakhtiari, S., Mineault, P., Lillicrap, T., Pack, C., & Richards, B. (2021). The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. *Advances in Neural Information Processing Systems*, *34*, 25164–25178 (cit. on p. 11).

Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. In W. A. Rosenblith (Ed.), *Sensory Communication* (pp. 217–234). MIT Press. (Cit. on p. 27).

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436 (cit. on pp. 13, 23, 25, 32, 36, 43).

Batty, E., Merel, J., Brackbill, N., Heitman, A., Sher, A., Litke, A., Chichilnisky, E., & Paninski, L. (2016). Multilayer recurrent network models of primate retinal ganglion cell responses (cit. on pp. 7, 19).

Bernáez Timón, L., Ekelmans, P., Kraynyukova, N., Rose, T., Busse, L., & Tchumatchenko, T. (2023). How to incorporate biological insights into network models and why it matters. *The Journal of Physiology*, *601*(15), 3037–3053 (cit. on p. 34).

Böhm, J. N., Berens, P., & Kobak, D. (2022). Unsupervised visualization of image datasets using contrastive learning. *arXiv preprint arXiv:2210.09879* (cit. on p. 36).

Bonds, A. B. (1989). Role of inhibition in the specification of orientation selectivity of cells in the cat striate cortex. *Visual Neuroscience*, *2*, 41–55. https://doi.org/10.1017/S0952523800004314 (cit. on p. 27).

Borowski, J., Zimmermann, R. S., Schepers, J., Geirhos, R., Wallis, T. S., Bethge, M., & Brendel, W. (2020). Exemplary natural images explain cnn activations better than state-of-the-art feature visualization. *arXiv preprint arXiv:2010.12606* (cit. on p. 37).

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., et al. (2023). Deep problems with neural network models of human vision. *Behavioral and Brain Sciences*, *46*, e385 (cit. on p. 33).

Burg, M. F., Cadena, S. A., Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., & Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, *17*(6), e1009028 (cit. on pp. 7, 12, 24, 25, 40).

Burg, M. F., Zenkel, T., Vystrčilová, M., Oesterle, J., Höfling, L., Willeke, K. F., Lause, J., Müller, S., Fahey, P. G., Ding, Z., Restivo, K., Sridhar, S., Gollisch, T., Berens, P., Tolias, A. S., Euler, T., Bethge, M., & Ecker, A. S. (2024). Most discriminative stimuli for functional cell type identification. *The Twelfth International Conference on Learning Representations* (cit. on p. 43).

Busse, L., Wade, A. R., & Carandini, M. (2009). Representation of concurrent stimuli by population activity in visual cortex. *Neuron*, *64*, 931–942. https://doi.org/10.1016/j.neuron.2009.11.004 (cit. on pp. 11, 25, 27).

Butts, D. A. (2019). Data-driven approaches to understanding visual neuron activity. *Annual review of vision science*, *5*, 451–477 (cit. on pp. 1, 5).

Butts, D. A., Cui, Y., & Casti, A. R. (2016). Nonlinear computations shaping temporal processing of precortical vision. *Journal of Neurophysiology*, *116*(3), 1344–1357 (cit. on pp. 5, 6).

Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., Reimer, J., Bethge, M., Tolias, A., & Ecker, A. S. (2019). How

well do deep neural networks trained on object recognition characterize the mouse visual system? *Advances in Neural Information Processing (NeurIPS) Neuro-AI Workshop* (cit. on pp. 7, 8).

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019a). Deep convolutional models improve predictions of macaque V1 responses to natural images. *PLoS Computational Biology*, *15*(4), e1006897 (cit. on pp. 6, 7).

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019b). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897 (cit. on pp. 8, 9, 13, 22, 24–26, 29).

Cadena, S. A., Weis, M. A., Gatys, L. A., Bethge, M., & Ecker, A. S. (2018). Diverse feature visualizations reveal invariances in early layers of deep neural networks. *Proceedings of the European Conference on Computer Vision (ECCV)*, 217–232 (cit. on pp. 13, 15, 25).

Cadena, S. A., Willeke, K. F., Restivo, K., Denfield, G., Sinz, F. H., Bethge, M., Tolias, A. S., & Ecker, A. S. (2024). Diverse task-driven modeling of macaque v4 reveals functional specialization towards semantic tasks. *PLOS Computational Biology*, *20*(5), e1012056 (cit. on pp. 7–10, 39, 40, 42).

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014a). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition (M. Bethge, Ed.). *PLoS Computational Biology*, *10*, e1003963. https://doi.org/10.1371/journal.pcbi.1003963 (cit. on pp. 9, 20).

Cadieu, C. F., Hong, H., Yamins, D. L., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., & DiCarlo, J. J. (2014b). Deep neural networks rival the representation of primate it cortex for core visual object recognition. *PLoS Comput Biol*, *10*(12), e1003963 (cit. on pp. 13, 23).

Calabrese, A., & Paninski, L. (2011). Kalman filter mixture model for spike sorting of non-stationary data. *Journal of neuroscience methods*, *196*(1), 159–169 (cit. on p. 6).

Cammarata, N., Carter, S., Goh, G., Olah, C., Petrov, M., Schubert, L., Voss, C., Egan, B., & Lim, S. K. (2020). Thread: Circuits. *Distill*, *5*(3), e24 (cit. on pp. 36, 47).

Cammarata, N., Goh, G., Carter, S., Schubert, L., Petrov, M., & Olah, C. (2020). Curve detectors. *Distill*, *5*(6), e00024–003 (cit. on p. 36).

Cao, R., & Yamins, D. (2021). Explanatory models in neuroscience: Part 1–taking mechanistic abstraction seriously. *arXiv preprint arXiv:2104.01490* (cit. on p. 10).

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., & Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience*, *25*, 10577–10597. https://doi.org/10.1523/JNEUROSCI.3726-05.2005 (cit. on pp. 1, 6).

Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, *13*, 51–62. https://doi.org/10.1038/nrn3136 (cit. on pp. 11, 12, 14, 25, 44).

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *Journal of Neuroscience*, *17*, 8621–8644. https://doi.org/10.1523/JNEUROSCI.17-21-08621.1997 (cit. on p. 14).

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002a). Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. *Journal of Neurophysiology*, *88*, 2530–2546 (cit. on pp. 11, 26).

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002b). Selectivity and spatial distribution of signals from the receptive field surround in macaque v1 neurons. *Journal of neurophysiology*, *88*(5), 2547–2556 (cit. on p. 26).

Chen, T., Kornblith, S., Norouzi, M., & Hinton, G. (2020). A simple framework for contrastive learning of visual representations. *International conference on machine learning*, 1597–1607 (cit. on p. 45).

Chichilnisky, E. (2001). A simple white noise analysis of neuronal light responses. *Network: computation in neural systems*, *12*(2), 199 (cit. on p. 6).

Cirincione, A., Verrier, R., Bic, A., Olaiya, S., DiCarlo, J. J., Udeigwe, L., & Marques, T. (2022). Implementing divisive normalization in cnns improves robustness to common image corruptions. *SVRHM 2022 Workshop@ NeurIPS* (cit. on p. 12).

Coen-Cagli, R., Kohn, A., & Schwartz, O. (2015). Flexible gating of contextual influences in natural vision. *Nature Neuroscience*, *18*, 1648–1655. https://doi.org/10.1038/nn.4128 (cit. on p. 26).

Consortium, M., Bae, J. A., Baptiste, M., Bishop, C. A., Bodor, A. L., Brittain, D., Buchanan, J., Bumbarger, D. J., Castro, M. A., Celii, B., et al. (2021). Functional connectomics spanning multiple areas of mouse visual cortex. *BioRxiv*, 2021–07 (cit. on p. 46).

Conwell, C., Mayo, D., Barbu, A., Buice, M., Alvarez, G., & Katz, B. (2021). Neural regression, representational similarity, model zoology & neural taskonomy at scale in rodent visual cortex. *Advances in Neural Information Processing Systems*, *34*, 5590–5607 (cit. on pp. 34, 35).

Conwell, C., Prince, J. S., Kay, K. N., Alvarez, G. A., & Konkle, T. (2022). What can 1.8 billion regressions tell us about the pressures shaping high-level visual representation in brains and machines? *BioRxiv*, 2022–03 (cit. on p. 10).

Cornford, J., Kalajdzievski, D., Leite, M., Lamarquette, A., Kullmann, D. M., & Richards, B. (2020). Learning to live with dale's principle: Anns with separate excitatory and inhibitory units. *bioRxiv*, 2020–11 (cit. on p. 11).

Cowley, B., & Pillow, J. W. (2020). High-contrast "gaudy" images improve the training of deep neural network models of visual cortex. *Advances in Neural Information Processing Systems*, *33*, 21591–21603 (cit. on p. 7).

Cowley, B. R., Stan, P. L., Pillow, J. W., & Smith, M. A. (2023). Compact deep neural network models of visual cortex. *bioRxiv* (cit. on p. 47).

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D., & DiCarlo, J. J. (2020). Simulating a primary visual cortex at the front of cnns improves robustness to image perturbations. *Advances in Neural Information Processing Systems*, *33*, 13073–13087 (cit. on pp. 11, 24, 25, 40).

David, S. V., & Gallant, J. L. (2005). Predicting neuronal responses during natural vision. *Network: Computation in Neural Systems*, *16*(2-3), 239–260 (cit. on pp. 5, 8).

de Beeck, H. O., & Bracci, S. (2023). Going after the bigger picture: Using high-capacity models to understand mind and brain. *Behavioral and Brain Sciences*, *46*, e404 (cit. on p. 33).

de Vries, S. E., Lecoq, J. A., Buice, M. A., Groblewski, P. A., Ocker, G. K., Oliver, M., Feng, D., Cain, N., Ledochowitsch, P., Millman, D., et al. (2020). A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature neuroscience*, *23*(1), 138–151 (cit. on p. 34).

DeAngelis, G. C., Robson, J. G., Ohzawa, I., & Freeman, R. D. (1992). Organization of suppression in receptive fields of neurons in cat visual cortex. *Journal of Neurophysiology*, *68*, 144–163. https://doi.org/10.1152/jn.1992.68.1.144 (cit. on pp. 15, 27).

Dehghani, M., Djolonga, J., Mustafa, B., Padlewski, P., Heek, J., Gilmer, J., Steiner, A. P., Caron, M., Geirhos, R., Alabdulmohsin, I., et al. (2023). Scaling vision transformers to 22 billion parameters. *International Conference on Machine Learning*, 7480–7512 (cit. on p. 41).

Dehmamy, N., Walters, R., Liu, Y., Wang, D., & Yu, R. (2021). Automatic symmetry discovery with lie algebra convolutional network. *Advances in Neural Information Processing Systems*, *34*, 2503–2515 (cit. on p. 46).

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in cognitive sciences*, *11*(8), 333–341 (cit. on p. 9).

DiCarlo, J. J., Yamins, D. L., Ferguson, M. E., Fedorenko, E., Bethge, M., Bonnen, T., & Schrimpf, M. (2023). Let's move forward: Image-computable models and a common model evaluation scheme are prerequisites for a scientific understanding of human vision. *Behavioral and Brain Sciences*, *46*, e390 (cit. on p. 33).

Ding, Z., Tran, D. T., Ponder, K., Cobos, E., Ding, Z., Fahey, P. G., Wang, E., Muhammad, T., Fu, J., Cadena, S. A., et al. (2023). Bipartite invariance in mouse primary visual cortex. *bioRxiv* (cit. on pp. 7, 13, 25, 36, 45).

Djurdjevic, V., Ansuini, A., Bertolini, D., Macke, J. H., & Zoccolan, D. (2018). Accuracy of rats in discriminating visual objects is explained by the complexity of their perceptual strategy. *Current biology*, *28*(7), 1005–1015 (cit. on p. 34).

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., & Darrell, T. (2014). Decaf: A deep convolutional activation feature for generic visual recognition. *International conference on machine learning*, 647–655 (cit. on p. 9).

Douglas, R. J., Martin, K. A., & Whitteridge, D. (1989). A canonical microcircuit for neocortex. *Neural computation*, *1*(4), 480–488 (cit. on p. 11).

Doya, K. (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press. (Cit. on p. 5).

Dräger, U. C. (1975). Receptive fields of single cells and topography in mouse visual cortex. *Journal of Comparative Neurology*, *160*(3), 269–289 (cit. on p. 21).

D'Souza, R. D., Wang, Q., Ji, W., Meier, A. M., Kennedy, H., Knoblauch, K., & Burkhalter, A. (2022). Hierarchical and nonhierarchical features of the mouse visual cortical network. *Nature communications*, *13*(1), 503 (cit. on p. 34).

Eccles, J. C. (1976). From electrical to chemical transmission in the central nervous system: The closing address of the sir henry dale centennial symposium cambridge, 19 september 1975. *Notes and records of the Royal Society of London*, *30*(2), 219–230 (cit. on p. 11).

Ecker, A. S., Sinz, F. H., Froudarakis, E., Fahey, P. G., Cadena, S. A., Walker, E. Y., Cobos, E., Reimer, J., Tolias, A. S., & Bethge, M. (2018). A rotation-equivariant convolutional neural network model of primary visual cortex. *arXiv preprint arXiv:1809.10504* (cit. on pp. 7, 10, 40, 45).

Burg, M. F., **Cadena, Santiago A**, Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., & Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, *17*(6), e1009028 (cit. on pp. 16, 25, 136).

Ding, Z., Tran, D. T., Ponder, K., Cobos, E., Ding, Z., Fahey, P. G., Wang, E., Muhammad, T., Fu, J., **Cadena, Santiago A**, et al. (2023). Bipartite invariance in mouse primary visual cortex. *bioRxiv* (cit. on p. 17).

Ecker, A. S., Sinz, F. H., Froudarakis, E., Fahey, P. G., **Cadena, SA**, Walker, E. Y., Cobos, E., Reimer, J., Tolias, A. S., & Bethge, M. (2019). A rotation-equivariant convolutional neural network model of primary visual cortex. *International Conference on Learning Representations (ICLR)* (cit. on p. 16).

Lurz, K., Bashiri, M., Willeke, K., Jagadish, A., Wang, E., Walker, E., **Cadena, SA**, Muhammad, T., Cobos, E., Tolias, A., & Sinz, F. H. (2021). Generalization in data-driven models of primary visual cortex. *Ninth International Conference on Learning Representations (ICLR 2021)* (cit. on p. 16).

Safarani, S., Nix, A., Willeke, K., **Cadena, SA**, Restivo, K., Denfield, G., Tolias, A., & Sinz, F. (2021). Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, *34*, 739–751 (cit. on p. 17).

Ustyuzhaninov, I., Burg, M. F., **Cadena, Santiago A**, Fu, J., Muhammad, T., Ponder, K., Froudarakis, E., Ding, Z., Bethge, M., Tolias, A. S., et al. (2022). Digital twin reveals combinatorial code of non-linear computations in the mouse primary visual cortex. *bioRxiv*, 2022–02 (cit. on p. 17).

Ustyuzhaninov, I., **Cadena, SA**, Froudarakis, E., Fahey, P. G., Walker, E. Y., Cobos, E., Reimer, J., Sinz, F. H., Tolias, A. S., Bethge, M., et al. (2019). Rotation-invariant clustering of neuronal responses in primary visual cortex. *International Conference on Learning Representations* (cit. on p. 16).

Willeke, K. F., Fahey, P. G., Bashiri, M., Pede, L., Burg, M. F., Blessing, C., **Cadena, Santiago A**, Ding, Z., Lurz, K.-K., Ponder, K., et al. (2022). The sensorium competition on predicting large-scale mouse primary visual cortex activity. *arXiv preprint arXiv:2206.08666* (cit. on pp. 17, 39, 41).

Willeke, K. F., Restivo, K., Franke, K., Nix, A. F., **Cadena, Santiago A**, Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Ecker, A. S., et al. (2023). Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. *bioRxiv*, 2023–05 (cit. on p. 17).

Erhan, D., Bengio, Y., Courville, A., & Vincent, P. (2009). Visualizing higher-layer features of a deep network. *University of Montreal*, *1341*(3), 1 (cit. on pp. 13, 15, 28).

Felleman, D. J., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral cortex*, *1*(1), 1–47 (cit. on pp. 9, 42).

Franke, K., Willeke, K. F., Ponder, K., Galdamez, M., Zhou, N., Muhammad, T., Patel, S., Froudarakis, E., Reimer, J., Sinz, F. H., et al. (2022). State-dependent pupil dilation rapidly shifts visual feature selectivity. *Nature*, *610*(7930), 128–134 (cit. on p. 35).

Frankle, J., & Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635* (cit. on p. 41).

Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P., & Movshon, J. A. (2013). A functional and perceptual signature of the second visual area in primates. *Nature neuroscience*, *16*(7), 974–981 (cit. on p. 39).

Freeman, T. C., Durand, S., Kiper, D. C., & Carandini, M. (2002). Suppression without inhibition in visual cortex. *Neuron*, *35*, 759–771. https://doi.org/10.1016/S0896-6273(02)00819-X (cit. on p. 15).

Fu, Y., Tucciarone, J. M., Espinosa, J. S., Sheng, N., Darcy, D. P., Nicoll, R. A., Huang, Z. J., & Stryker, M. P. (2014). A cortical circuit for gain control by behavioral state. *Cell*, *156*(6), 1139–1152 (cit. on p. 11).

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, *36*(4), 193–202 (cit. on p. 2).

Garrett, M. E., Nauhaus, I., Marshel, J. H., & Callaway, E. M. (2014). Topography and areal organization of mouse visual cortex. *Journal of Neuroscience*, *34*(37), 12587–12600 (cit. on p. 21).

Gatys, L., Ecker, A. S., & Bethge, M. (2015). Texture synthesis using convolutional neural networks. *Advances in Neural Information Processing Systems*, 262–270 (cit. on pp. 20, 29, 35).

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image style transfer using convolutional neural networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423 (cit. on p. 29).

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., & Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, *2*(11), 665–673 (cit. on pp. 3, 33).

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2021). Partial success in closing the gap between human and machine vision. *Advances in Neural Information Processing Systems*, *34*, 23885–23899 (cit. on pp. 41, 43).

Geirhos, R., Narayanappa, K., Mitzkus, B., Thieringer, T., Bethge, M., Wichmann, F. A., & Brendel, W. (2022). The bittersweet lesson: Data-rich models narrow the behavioural gap to human vision. *Journal of Vision*, *22*(14), 3273–3273 (cit. on p. 40).

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. *International Conference on Learning Representations* (cit. on pp. 33, 43).

Geirhos, R., Zimmermann, R. S., Bilodeau, B., Brendel, W., & Kim, B. (2023). Don't trust your eyes: On the (un) reliability of feature visualizations. *arXiv preprint arXiv:2306.04719* (cit. on p. 37).

Glaser, J. I., Benjamin, A. S., Chowdhury, R. H., Perich, M. G., Miller, L. E., & Kording, K. P. (2020). Machine learning for neural decoding. *Eneuro*, *7*(4) (cit. on p. 9).

Glickfeld, L. L., Histed, M. H., & Maunsell, J. H. (2013). Mouse primary visual cortex is used to detect both orientation and contrast changes. *Journal of Neuroscience*, *33*(50), 19416–19422 (cit. on p. 21).

Golan, T., Raju, P. C., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences*, *117*(47), 29330–29337 (cit. on pp. 25, 42, 43).

Golan, T., Taylor, J., Schütt, H., Peters, B., Sommers, R. P., Seeliger, K., Doerig, A., Linton, P., Konkle, T., Van Gerven, M., et al. (2023). Deep neural networks are not a single hypothesis but a language for expressing computational hypotheses. *Behavioral and Brain Sciences*, *46* (cit. on p. 33).

Gollisch, T., & Meister, M. (2010). Eye smarter than scientists believed: Neural computations in circuits of the retina. *Neuron*, *65*(2), 150–164 (cit. on p. 11).

Gonschorek, D., Höfling, L., Szatko, K. P., Franke, K., Schubert, T., Dunn, B., Berens, P., Klindt, D., & Euler, T. (2021). Removing inter-experimental variability from functional data in systems neuroscience. *Advances in Neural Information Processing Systems*, *34*, 3706–3719 (cit. on p. 40).

Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in neurosciences*, *15*(1), 20–25 (cit. on p. 9).

Goodfellow, I., Lee, H., Le, Q. V., Saxe, A., & Ng, A. Y. (2009). Measuring invariances in deep networks. *Advances in neural information processing systems*, 646–654 (cit. on p. 28).

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572* (cit. on pp. 3, 33).

Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, *35*(27), 10005–10014 (cit. on pp. 20, 21).

Haefner, R., & Cumming, B. (2008). An improved estimator of variance explained in the presence of noise. *Advances in neural information processing systems*, *21* (cit. on p. 8).

Hastie, T., & Tibshirani, R. (1985). Generalized additive models; some applications. *Generalized Linear Models: Proceedings of the GLIM 85 Conference held at Lancaster, UK, Sept. 16–19, 1985*, 66–81 (cit. on p. 6).

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (cit. on pp. 9, 29).

Heeger, D. J. (1992a). Normalization of cell responses in cat striate cortex. *Visual neuroscience*, *9*(2), 181–197 (cit. on pp. 15, 25, 26).

Heeger, D. J. (1992b). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, *9*, 181–197. https://doi.org/10.1017/S0952523800009640 (cit. on pp. 27, 40).

Heeger, D. (1991). Computational model of cat striate physiology. *Computational models of visual perception*, 119–133 (cit. on p. 19).

Hendrycks, D., & Dietterich, T. (2019). Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261* (cit. on pp. 33, 43).

Hermann, K., Nayebi, A., van Steenkiste, S., & Jones, M. (2023). For human-like models, train on human-like tasks. *Behavioral and Brain Sciences*, *46* (cit. on p. 34).

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021). Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature communications*, *12*(1), 6456 (cit. on p. 47).

Higgins, I., Matthey, L., Pal, A., Burgess, C., Glorot, X., Botvinick, M., Mohamed, S., & Lerchner, A. (2016). Beta-vae: Learning basic visual concepts with a constrained variational framework. *International conference on learning representations* (cit. on p. 46).

Hinton, G., Vinyals, O., & Dean, J. (2015). Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531* (cit. on p. 41).

Höfling, L., Szatko, K. P., Behrens, C., Qiu, Y., Klindt, D. A., Jessen, Z., Schwartz, G. W., Bethge, M., Berens, P., Franke, K., et al. (2022). A chromatic feature detector in the retina signals visual context changes. *bioRxiv*, 2022–11 (cit. on p. 7).

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural networks*, *4*(2), 251–257 (cit. on p. 6).

Hoy, J. L., Yavorska, I., Wehr, M., & Niell, C. M. (2016). Vision drives accurate approach behavior during prey capture in laboratory mice. *Current Biology*, *26*(22), 3046–3052 (cit. on p. 22).

Hsu, A., Borst, A., & Theunissen, F. E. (2004). Quantifying variability in neural responses and its application for the validation of model predictions. *Network: Computation in Neural Systems*, *15*(2), 91 (cit. on p. 8).

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, *148*(3), 574–591 (cit. on p. 1).

Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, *160*(1), 106–154 (cit. on pp. 2, 10).

Hubel, D. H., & Wiesel, T. N. (1968). Receptive fields and functional architecture of monkey striate cortex. *The Journal of physiology*, *195*(1), 215–243 (cit. on p. 6).

Itti, L., Koch, C., & Braun, J. (2000). Revisiting spatial vision: Toward a unifying model. *Journal of the Optical Society of America A*, *17*(11), 1899–1917 (cit. on p. 27).

Jaderberg, M., Simonyan, K., Zisserman, A., et al. (2015). Spatial transformer networks. *Advances in neural information processing systems*, *28* (cit. on p. 7).

Jones, J. P., & Palmer, L. A. (1987). An evaluation of the two-dimensional gabor filter model of simple receptive fields in cat striate cortex. *Journal of neurophysiology*, *58*(6), 1233–1258 (cit. on p. 19).

Jun, J. J., Steinmetz, N. A., Siegle, J. H., Denman, D. J., Bauza, M., Barbarits, B., Lee, A. K., Anastassiou, C. A., Andrei, A., Aydın, Ç., et al. (2017). Fully integrated silicon probes for high-density recording of neural activity. *Nature*, *551*(7679), 232–236 (cit. on p. 2).

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 (cit. on pp. 14, 21).

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014a). Deep supervised, but not unsupervised, models may explain IT cortical representation (J. Diedrichsen, Ed.). *PLoS Computational Biology*, *10*, e1003915. https://doi.org/10.1371/journal.pcbi.1003915 (cit. on p. 9).

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014b). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Comput Biol*, *10*(11), e1003915 (cit. on p. 20).

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 (cit. on pp. 11, 25).

Kim, T., Bair, W., & Pasupathy, A. (2019). Neural coding for shape and texture in macaque area v4. *Journal of Neuroscience*, *39*(24), 4760–4774 (cit. on p. 23).

Kindel, W. F., Christensen, E. D., & Zylberberg, J. (2017). Using deep learning to reveal the neural code for images in primary visual cortex. *arXiv preprint arXiv:1706.06208* (cit. on pp. 6–8).

Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114* (cit. on p. 46).

Klindt, D., Ecker, A. S., Euler, T., & Bethge, M. (2017). Neural system identification for large populations separating " what" and " where". In *Advances in Neural Information Processing Systems 30* (pp. 3506–3516). (Cit. on pp. 6, 7, 10, 19, 20, 45).

Kong, N. C., Margalit, E., Gardner, J. L., & Norcia, A. M. (2022). Increasing neural network robustness improves match to macaque v1 eigenspectrum, spatial frequency preference and predictivity. *PLOS Computational Biology*, *18*(1), e1009739 (cit. on p. 40).

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446 (cit. on pp. 13, 20, 21).

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 1097–1105 (cit. on pp. 2, 9).

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., et al. (2019). Brain-like object recognition with high-performing shallow recurrent anns. *Advances in neural information processing systems*, *32* (cit. on p. 11).

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *BioRxiv*, 408385 (cit. on pp. 11, 32, 39, 42).

Kümmerer, M., Wallis, T. S., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, *112*(52), 16054–16059 (cit. on p. 9).

Lau, B., Stanley, G. B., & Dan, Y. (2002). Computational subunits of visual cortical neurons revealed by artificial neural networks. *Proceedings of the National Academy of Sciences*, *99*, 8974–8979. https://doi.org/10.1073/pnas.122173799 (cit. on p. 6).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444 (cit. on pp. 2, 7, 19).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278–2324 (cit. on p. 2).

Lee, S., Kruglikov, I., Huang, Z. J., Fishell, G., & Rudy, B. (2013). A disinhibitory circuit mediates motor integration in the somatosensory cortex. *Nature neuroscience*, *16*(11), 1662–1670 (cit. on p. 11).

Lehky, S., Sejnowski, T., & Desimone, R. (1992). Predicting responses of nonlinear neurons in monkey striate cortex to complex patterns. *The Journal of Neuroscience*, *12*, 3568–3581. https://doi.org/10.1523/JNEUROSCI.12-09-03568.1992 (cit. on p. 6).

Li, B. M., Cornacchia, I. M., Rochefort, N. L., & Onken, A. (2023). V1t: Large-scale mouse v1 response prediction using a vision transformer. *arXiv preprint arXiv:2302.03023* (cit. on p. 41).

Li, P., Cornford, J., Ghosh, A., & Richards, B. (2023). Learning better with dale's law: A spectral perspective. *bioRxiv*, 2023–06 (cit. on p. 11).

Lindsey, J., Ocko, S. A., Ganguli, S., & Deny, S. (2019). A Unified Theory of Early Visual Representations from Retina to Cortex through Anatomically Constrained Deep CNNs. *arXiv:1901.00945 [cs, q-bio]* (cit. on p. 11).

Linsley, D., Kim, J., Veerabadran, V., Windolf, C., & Serre, T. (2018). Learning long-range spatial dependencies with horizontal gated recurrent units. *Advances in neural information processing systems*, *31* (cit. on p. 11).

Linsley, D., Rodriguez Rodriguez, I. F., Fel, T., Arcaro, M., Sharma, S., Livingstone, M., & Serre, T. (2024). Performance-optimized deep neural networks are evolving into worse models of inferotemporal visual cortex. *Advances in Neural Information Processing Systems*, *36* (cit. on p. 32).

Linsley, D., & Serre, T. (2023). Fixing the problems of deep neural networks will require better training data and learning algorithms. *arXiv preprint arXiv:2311.12819* (cit. on p. 34).

Lurz, K., Bashiri, M., Willeke, K., Jagadish, A., Wang, E., Walker, E., Cadena, S., Muhammad, T., Cobos, E., Tolias, A., et al. (2021). Generalization in data-driven models of primary visual cortex. *Ninth International Conference on Learning Representations (ICLR 2021)* (cit. on pp. 7, 23, 40).

Lurz, K.-K., Bashiri, M., Walker, E. Y., & Sinz, F. H. (2022). Bayesian oracle for bounding information gain in neural encoding models. *The Eleventh International Conference on Learning Representations* (cit. on p. 9).

Mahendran, A., & Vedaldi, A. (2016). Visualizing deep convolutional neural networks using natural pre-images. *International Journal of Computer Vision*, *120*(3), 233–255 (cit. on p. 28).

Maheswaranathan, N., Kastner, D. B., Baccus, S. A., & Ganguli, S. (2018). Inferring hidden structure in multilayered neural circuits. *PLoS computational biology*, *14*(8), e1006291 (cit. on p. 6).

Maheswaranathan, N., McIntosh, L. T., Tanaka, H., Grant, S., Kastner, D. B., Melander, J. B., Nayebi, A., Brezovec, L. E., Wang, J. H., Ganguli, S., et al. (2023). Interpreting the retinal neural code for natural scenes: From computations to neurons. *Neuron*, *111*(17), 2742–2755 (cit. on pp. 7, 11, 12, 47).

Marmarelis, V. Z. (2004). *Nonlinear dynamic modeling of physiological systems* (Vol. 10). John Wiley & Sons. (Cit. on p. 6).

Marques, T., Schrimpf, M., & DiCarlo, J. J. (2021). Multi-scale hierarchical neural network models that bridge from single neurons in the primate primary visual cortex to object recognition behavior. *bioRxiv*, 2021–03 (cit. on pp. 39, 44).

Marr, D., & Poggio, T. (1976). From understanding computation to understanding neural circuitry (cit. on p. 9).

Marshel, J. H., Garrett, M. E., Nauhaus, I., & Callaway, E. M. (2011). Functional specialization of seven mouse visual cortical areas. *Neuron*, *72*(6), 1040–1054 (cit. on p. 21).

Masland, R. H. (2012). The neuronal organization of the retina. *Neuron*, *76*(2), 266–280 (cit. on p. 11).

McFarland, J. M., Cui, Y., & Butts, D. A. (2013). Inferring nonlinear neuronal computation based on physiologically plausible inputs. *PLoS Computational Biology*, *9*(7), e1003143 (cit. on p. 6).

McIntosh, L., Maheswaranathan, N., Nayebi, A., Ganguli, S., & Baccus, S. (2016). Deep learning models of the retinal response to natural scenes. *Advances in Neural Information Processing Systems*, 1369–1377 (cit. on pp. 7, 19, 20).

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118 (cit. on p. 32).

Miller, K. D. (2016). Canonical computations of cerebral cortex. *Current opinion in neurobiology*, *37*, 75–84 (cit. on pp. 11, 14).

Minni, S., Ji-An, L., Moskovitz, T., Lindsay, G., Miller, K., Dipoppa, M., & Yang, G. R. (2019). Understanding the functional and structural differences across excitatory and inhibitory neurons. *bioRxiv*, 680439 (cit. on p. 11).

Moeller, S., Crapse, T., Chang, L., & Tsao, D. Y. (2017). The effect of face patch microstimulation on perception of faces and objects. *Nature neuroscience*, *20*(5), 743–752 (cit. on p. 44).

Morrone, M. C., Burr, D. C., & Maffei, L. (1982). Functional implications of cross-orientation inhibition of cortical visual cells. I. Neurophysiological evidence. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *216*, 335–354. https://doi.org/10.1098/rspb.1982.0078 (cit. on pp. 15, 26, 27).

Movshon, J. A., Thompson, I., & Tolhurst, D. (1978). Receptive field organization of complex cells in the cat's striate cortex. *The Journal of physiology*, *283*, 79 (cit. on p. 6).

Nayebi, A., Bear, D., Kubilius, J., Kar, K., Ganguli, S., Sussillo, D., DiCarlo, J. J., & Yamins, D. L. K. (2018). Task-Driven Convolutional Recurrent Models of the Visual System. *arXiv:1807.00053 [cs, q-bio]* (cit. on pp. 9, 11).

Nayebi, A., Kong, N. C., Zhuang, C., Gardner, J. L., Norcia, A. M., & Yamins, D. L. (2023). Mouse visual cortex as a limited resource system that self-learns an ecologically-general representation. *PLOS Computational Biology*, *19*(10), e1011506 (cit. on pp. 34, 35).

Nelder, J. A., & Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, *135*(3), 370–384 (cit. on p. 6).

Neyshabur, B. (2020). Towards learning convolutions from scratch. *Advances in Neural Information Processing Systems*, *33*, 8078–8088 (cit. on p. 46).

Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., & Yosinski, J. (2017). Plug & play generative networks: Conditional iterative generation of images in latent space. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 4467–4477 (cit. on pp. 13, 15, 28).

Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., & Clune, J. (2016). Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. *Advances in neural information processing systems*, *29* (cit. on pp. 13, 15, 28).

Okun, M., Steinmetz, N. A., Cossell, L., Iacaruso, M. F., Ko, H., Barthó, P., Moore, T., Hofer, S. B., Mrsic-Flogel, T. D., Carandini, M., et al. (2015). Diverse coupling of neurons to populations in sensory cortex. *Nature*, *521*(7553), 511–515 (cit. on p. 6).

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020a). An overview of early vision in inceptionv1. *Distill*, *5*(4), e00024–002 (cit. on p. 47).

Olah, C., Cammarata, N., Schubert, L., Goh, G., Petrov, M., & Carter, S. (2020b). Zoom in: An introduction to circuits. *Distill*, *5*(3), e00024–001 (cit. on p. 36).

Olah, C., Cammarata, N., Voss, C., Schubert, L., & Goh, G. (2020). Naturally occurring equivariance in neural networks. *Distill*, *5*(12), e00024–004 (cit. on p. 46).

Olah, C., Mordvintsev, A., & Schubert, L. (2017). Feature visualization. *Distill*, *2*(11), e7 (cit. on pp. 13, 15, 35, 36).

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding V1? *Neural Computation*, *17*, 1665–1699. https://doi.org/10.1162/0899766054026639 (cit. on pp. 2, 13, 19, 39).

Paninski, L. (2004). Maximum likelihood estimation of cascade point-process neural encoding models. *Network: Computation in Neural Systems*, *15*(4), 243–262 (cit. on pp. 5, 6).

Paninski, L., Pillow, J., & Lewi, J. (2007). Statistical models for neural encoding, decoding, and optimal stimulus design. *Progress in brain research*, *165*, 493–507 (cit. on p. 5).

Park, M., & Pillow, J. W. (2011). Receptive field inference with localized priors. *PLoS Comput Biol*, *7*(10), e1002219 (cit. on p. 6).

Park, M., & Pillow, J. W. (2013). Bayesian inference for low rank spatiotemporal neural receptive fields. *Advances in Neural Information Processing Systems*, *26* (cit. on p. 6).

Pasupathy, A., & Connor, C. E. (2001). Shape representation in area v4: Position-specific tuning for boundary conformation. *Journal of neurophysiology*, *86*(5), 2505–2519 (cit. on pp. 8, 14, 23, 24).

Pasupathy, A., Popovkina, D. V., & Kim, T. (2020). Visual functions of primate area v4. *Annual Review of Vision Science*, *6*, 363–385 (cit. on pp. 14, 23, 24, 44).

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. (Cit. on p. 7).

Peters, B., DiCarlo, J. J., Gureckis, T., Haefner, R., Isik, L., Tenenbaum, J., Konkle, T., Naselaris, T., Stachenfeld, K., Tavares, Z., et al. (2024). How does the primate brain combine generative and discriminative computations in vision? *arXiv preprint arXiv:2401.06005* (cit. on p. 46).

Pi, H.-J., Hangya, B., Kvitsiani, D., Sanders, J. I., Huang, Z. J., & Kepecs, A. (2013). Cortical interneurons that specialize in disinhibitory control. *Nature*, *503*(7477), 521–524 (cit. on p. 11).

Pierzchlewicz, P. A., Willeke, K. F., Nix, A. F., Elumalai, P., Restivo, K., Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Franke, K., et al. (2023). Energy guided diffusion for generating neurally exciting images. *bioRxiv* (cit. on p. 13).

Pillow, J., & Scott, J. (2012). Fully bayesian inference for neural models with negative-binomial spiking. *Advances in neural information processing systems*, *25* (cit. on p. 5).

Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E., & Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature*, *454*(7207), 995–999 (cit. on p. 6).

Polack, P.-O., Friedman, J., & Golshani, P. (2013). Cellular mechanisms of brain state–dependent gain modulation in visual cortex. *Nature neuroscience*, *16*(9), 1331–1339 (cit. on p. 11).

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving images for visual neurons using a deep

generative network reveals coding principles and neuronal preferences. *Cell*, *177*(4), 999–1009 (cit. on pp. 13, 32).

Pospisil, D. A., & Bair, W. (2021). The unbiased estimation of the fraction of variance explained by a model. *PLoS computational biology*, *17*(8), e1009212 (cit. on pp. 8, 9).

Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). 'artiphysiology'reveals v4-like shape tuning in a deep network trained for image classification. *Elife*, *7*, e38242 (cit. on pp. 23, 44).

Prenger, R., Wu, M. C. .-., David, S. V., & Gallant, J. L. (2004). Nonlinear V1 responses to natural scenes revealed by neural network analysis. *Neural Networks*, *17*, 663–679. https://doi.org/10.1016/j.neunet.2004.03.008 (cit. on p. 6).

Prusky, G. T., West, P. W., & Douglas, R. M. (2000). Behavioral assessment of visual acuity in mice and rats. *Vision research*, *40*(16), 2201–2209 (cit. on p. 34).

Qiu, Y., Zhao, Z., Klindt, D., Kautzky, M., Szatko, K. P., Schaeffel, F., Rifai, K., Franke, K., Busse, L., & Euler, T. (2021). Natural environment statistics in the upper and lower visual field are reflected in mouse retinal specializations. *Current Biology*, *31*(15), 3233–3247 (cit. on pp. 32, 35).

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. (2021). Learning transferable visual models from natural language supervision. *International conference on machine learning*, 8748–8763 (cit. on p. 46).

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9 (cit. on p. 41).

Rajalingham, R., & DiCarlo, J. J. (2019). Reversible inactivation of different millimeter-scale regions of primate it results in different patterns of core object recognition deficits. *Neuron*, *102*(2), 493–505 (cit. on p. 44).

Ren, Y., & Bashivan, P. (2023). How well do models of visual cortex generalize to out of distribution samples? *bioRxiv*, 2023–05 (cit. on p. 43).

Reynolds, J. H., & Heeger, D. J. (2009). The normalization model of attention. *Neuron*, *61*(2), 168–185 (cit. on p. 12).

Rezende, D., & Mohamed, S. (2015). Variational inference with normalizing flows. *International conference on machine learning*, 1530–1538 (cit. on p. 47).

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., et al. (2019). A deep learning framework for neuroscience. *Nature neuroscience*, *22*(11), 1761–1770 (cit. on p. 10).

Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, *2*(11), 1019–1025 (cit. on p. 2).

Ringach, D. L., Hawken, M. J., & Shapley, R. (2002). Receptive field structure of neurons in monkey primary visual cortex revealed by stimulation with natural image sequences. *Journal of vision*, *2*(1), 2–2 (cit. on p. 5).

Roddey, J. C., Girish, B., & Miller, J. P. (2000). Assessing the performance of neural encoding models in the presence of noise. *Journal of computational neuroscience*, *8*, 95–112 (cit. on p. 8).

Ronneberger, O., Fischer, P., & Brox, T. (2015). U-net: Convolutional networks for biomedical image segmentation. *Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18*, 234–241 (cit. on p. 42).

Roth, M. M., Helmchen, F., & Kampa, B. M. (2012). Distinct functional properties of primary and posteromedial visual area of mouse neocortex. *Journal of Neuroscience*, *32*(28), 9716–9726 (cit. on p. 21).

Rothkopf, C., Bremmer, F., Fiehler, K., Dobs, K., & Triesch, J. (2023). Models of vision need some action. *The Behavioral and brain sciences*, *46* (cit. on p. 34).

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, *323*(6088), 533–536 (cit. on p. 7).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., et al. (2015a). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252 (cit. on p. 19).

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015b). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*, 211–252. https://doi.org/10.1007/s11263-015-0816-y (cit. on p. 2).

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2005a). Spatiotemporal elements of macaque V1 receptive fields. *Neuron*, *46*, 945–956. https://doi.org/10.1016/j.neuron.2005.05.021 (cit. on pp. 6, 19).

Rust, N. C., Schwartz, O., Movshon, J. A., & Simoncelli, E. P. (2005b). Spatiotemporal elements of macaque v1 receptive fields. *Neuron*, *46*(6), 945–956 (cit. on pp. 6, 29).

Safarani, S., Nix, A., Willeke, K., Cadena, S., Restivo, K., Denfield, G., Tolias, A., & Sinz, F. (2021). Towards robust vision by multi-task learning on monkey visual cortex. *Advances in Neural Information Processing Systems*, *34*, 739–751 (cit. on pp. 7, 25, 42).

Sahani, M., & Linden, J. (2002a). Evidence optimization techniques for estimating stimulus-response functions. *Advances in neural information processing systems*, *15* (cit. on p. 5).

Sahani, M., & Linden, J. (2002b). How linear are auditory cortical responses? *Advances in neural information processing systems*, *15* (cit. on p. 8).

Salimans, T., Karpathy, A., Chen, X., Kingma, D. P., & Bulatov, Y. (2016). Pixelcnn++: A pixelcnn implementation with discretized logistic mixture likelihood and other modifications. *Submitted to ICLR 2017* (cit. on p. 28).

Sawada, T., & Petrov, A. A. (2017). The divisive normalization model of V1 neurons: A comprehensive comparison of physiological data and model

predictions. *Journal of Neurophysiology*, *118*, 3051–3091. https://doi.org/10.1152/jn.00821.2016 (cit. on p. 25).

Schoppe, O., Harper, N. S., Willmore, B. D., King, A. J., & Schnupp, J. W. (2016). Measuring the performance of neural models. *Frontiers in computational neuroscience*, *10*, 10 (cit. on p. 8).

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv preprint* (cit. on pp. 10, 14, 39, 41, 44).

Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* (cit. on p. 10).

Schrimpf, M., McGrath, P., Margalit, E., & DiCarlo, J. J. (2024). Do topographic deep ann models of the primate ventral stream predict the perceptual effects of direct it cortical interventions? *bioRxiv*, 2024–01 (cit. on p. 44).

Schröder, C., Klindt, D., Strauss, S., Franke, K., Bethge, M., Euler, T., & Berens, P. (2020). System identification with biophysical constraints: A circuit model of the inner retina. *Advances in Neural Information Processing Systems*, *33*, 15439–15450 (cit. on p. 11).

Schubert, L., Voss, C., Cammarata, N., Goh, G., & Olah, C. (2021). High-low frequency detectors. *Distill*, *6*(1), e00024–005 (cit. on p. 36).

Schütt, H. H., & Wichmann, F. A. (2017). An image-computable psychophysical spatial vision model. *Journal of Vision*, *17*. https://doi.org/10.1167/17.12.12 (cit. on p. 27).

Schwartz, O., & Simoncelli, E. P. (2001). Natural signal statistics and sensory gain control. *Nature neuroscience*, *4*(8), 819–825 (cit. on p. 25).

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE transactions on pattern analysis and machine intelligence*, *29*(3), 411–426 (cit. on p. 2).

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on pattern analysis and machine intelligence*, *22*(8), 888–905 (cit. on p. 34).

Shi, J., Shea-Brown, E., & Buice, M. (2019). Comparison against task driven artificial neural networks reveals functional properties in mouse visual cortex. *Advances in Neural Information Processing Systems*, *32* (cit. on p. 34).

Shi, Q., Gupta, P., Boukhvalova, A. K., Singer, J. H., & Butts, D. A. (2019). Functional characterization of retinal ganglion cells using tailored nonlinear modeling. *Scientific reports*, *9*(1), 8713 (cit. on p. 6).

Siegle, J. H., Jia, X., Durand, S., Gale, S., Bennett, C., Graddis, N., Heller, G., Ramirez, T. K., Choi, H., Luviano, J. A., et al. (2021). Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, *592*(7852), 86–92 (cit. on p. 34).

Simoncelli, E. P., Paninski, L., Pillow, J., & Schwartz, O. (2004). Characterization of neural responses with stochastic stimuli. *The cognitive neurosciences*, *3*, 327–338 (cit. on p. 6).

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (cit. on pp. 9, 20, 22, 29).

Sinz, F., & Bethge, M. (2008). The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction. *Advances in Neural Information Processing Systems 21*, 1521–1528 (cit. on p. 25).

Sinz, F. H., Ecker, A. S., Fahey, P. G., Walker, E. Y., Cobos, E., Froudarakis, E., Yatsenko, D., Pitkow, X., Reimer, J., & Tolias, A. S. (2018). Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. *BioRxiv*, 452672 (cit. on pp. 7, 8, 22, 43).

Sinz, F. H., Pitkow, X., Reimer, J., Bethge, M., & Tolias, A. S. (2019). Engineering a less artificial intelligence. *Neuron*, *103*(6), 967–979 (cit. on p. 10).

Smyth, D., Willmore, B., Baker, G. E., Thompson, I. D., & Tolhurst, D. J. (2003). The receptive-field organization of simple cells in primary visual cortex of ferrets under natural scene stimulation. *Journal of Neuroscience*, *23*(11), 4746–4759 (cit. on p. 5).

Sohl-Dickstein, J., Weiss, E., Maheswaranathan, N., & Ganguli, S. (2015). Deep unsupervised learning using nonequilibrium thermodynamics. *International conference on machine learning*, 2256–2265 (cit. on p. 47).

Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS computational biology*, *16*(10), e1008215 (cit. on p. 11).

Srinath, R., Emonds, A., Wang, Q., Lempel, A. A., Dunn-Weiss, E., Connor, C. E., & Nielsen, K. J. (2021). Early emergence of solid shape coding in natural and deep network vision. *Current Biology*, *31*(1), 51–65 (cit. on pp. 14, 24, 44).

Steinmetz, N. A., Aydin, C., Lebedeva, A., Okun, M., Pachitariu, M., Bauza, M., Beau, M., Bhagat, J., Böhm, C., Broux, M., et al. (2021). Neuropixels 2.0: A miniaturized high-density probe for stable, long-term brain recordings. *Science*, *372*(6539), eabf4588 (cit. on pp. 2, 40).

Stevenson, I. H., & Kording, K. P. (2011). How advances in neural recording affect data analysis. *Nature neuroscience*, *14*(2), 139–142 (cit. on p. 2).

Stevenson, I. H., Rebesco, J. M., Hatsopoulos, N. G., Haga, Z., Miller, L. E., & Kording, K. P. (2008). Bayesian inference of functional connectivity and network structure from spikes. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, *17*(3), 203–213 (cit. on p. 5).

Stringer, C., Pachitariu, M., Steinmetz, N., Reddy, C. B., Carandini, M., & Harris, K. D. (2019). Spontaneous behaviors drive multidimensional, brainwide activity. *Science*, *364*(6437), eaav7893 (cit. on p. 2).

Talebi, V., & Baker, C. L. (2012). Natural versus synthetic stimuli for estimating receptive field models: A comparison of predictive robustness. *The Journal of Neuroscience*, *32*(5), 1560–1576 (cit. on p. 19).

Cadena, Santiago A, Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897 (cit. on pp. 16, 19, 72).

Cadena, Santiago A, Weis, M. A., Gatys, L. A., Bethge, M., & Ecker, A. S. (2018). Diverse feature visualizations reveal invariances in early layers of deep neural networks. *Proceedings of the European Conference on Computer Vision (ECCV)*, 217–232 (cit. on pp. 16, 28, 168).

Cadena, Santiago A, Willeke, K. F., Restivo, K., Denfield, G., Sinz, F. H., Bethge, M., Tolias, A. S., & Ecker, A. S. (2024). Diverse task-driven modeling of macaque v4 reveals functional specialization towards semantic tasks. *PLOS Computational Biology*, *20*(5), e1012056 (cit. on pp. 16, 23, 106).

S. A. Cadena, Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., Reimer, J., Bethge, M., Tolias, A., & Ecker, A. S. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? *Advances in Neural Information Processing (NeurIPS) Neuro-AI Workshop* (cit. on pp. 16, 21, 100).

Theis, L., Chagas, A. M., Arnstein, D., Schwarz, C., & Bethge, M. (2013). Beyond glms: A generative mixture modeling approach to neural system identification. *PLoS Comput Biol*, *9*(11), e1003356 (cit. on pp. 9, 46).

Theunissen, F. E., David, S. V., Singh, N. C., Hsu, A., Vinje, W. E., & Gallant, J. L. (2001). Estimating spatio-temporal receptive fields of auditory and visual neurons from their responses to natural stimuli. *Network: Computation in Neural Systems*, *12*(3), 289 (cit. on p. 6).

Tohmi, M., Meguro, R., Tsukano, H., Hishida, R., & Shibuki, K. (2014). The extrageniculate visual pathway generates distinct response properties in the higher visual areas of mice. *Current Biology*, *24*(6), 587–597 (cit. on p. 21).

Tong, R., da Silva, R., Lin, D., Ghosh, A., Wilsenach, J., Cianfarano, E., Bashivan, P., Richards, B., & Trenholm, S. (2023). The feature landscape of visual cortex. *bioRxiv*, 2023–11 (cit. on pp. 13, 45).

Touryan, J., Felsen, G., & Dan, Y. (2005). Spatial structure of complex cell receptive fields measured with natural images. *Neuron*, *45*(5), 781–791 (cit. on p. 6).

Touryan, J., Lau, B., & Dan, Y. (2002). Isolation of relevant visual features from random stimuli for cortical complex cells. *The Journal of neuroscience*, *22*(24), 10811–10818 (cit. on p. 6).

Truccolo, W., Eden, U. T., Fellows, M. R., Donoghue, J. P., & Brown, E. N. (2005). A point process framework for relating neural spiking activity to spiking history, neural ensemble, and extrinsic covariate effects. *Journal of neurophysiology*, *93*(2), 1074–1089 (cit. on p. 6).

Turishcheva, P., Fahey, P. G., Hansel, L., Froebe, R., Ponder, K., Vystrčilová, M., Willeke, K. F., Bashiri, M., Wang, E., Ding, Z., et al. (2023). The dynamic sensorium competition for predicting large-scale mouse visual cortex activity from videos. *arXiv preprint arXiv:2305.19654* (cit. on p. 43).

Ustyuzhaninov, I., Burg, M. F., Cadena, S. A., Fu, J., Muhammad, T., Ponder, K., Froudarakis, E., Ding, Z., Bethge, M., Tolias, A. S., et al. (2022). Digital twin reveals combinatorial code of non-linear computations in the mouse primary visual cortex. *bioRxiv*, 2022–02 (cit. on pp. 7, 10, 12, 44, 45).

Ustyuzhaninov, I., Cadena, S. A., Froudarakis, E., Fahey, P. G., Walker, E. Y., Cobos, E., Reimer, J., Sinz, F. H., Tolias, A. S., Bethge, M., et al. (2019). Rotation-invariant clustering of neuronal responses in primary visual

cortex. *International Conference on Learning Representations* (cit. on pp. 7, 10, 40, 45).

Van Den Oord, A., Kalchbrenner, N., & Kavukcuoglu, K. (2016). Pixel recurrent neural networks. *International conference on machine learning*, 1747–1756 (cit. on p. 47).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30* (cit. on pp. 11, 41).

Vinken, K., & Op de Beeck, H. (2021). Using deep neural networks to evaluate object vision tasks in rats. *PLoS computational biology*, *17*(3), e1008714 (cit. on p. 34).

Vinken, K., Vermaercke, B., & de Beeck, H. P. O. (2014). Visual categorization of natural movies by rats. *Journal of Neuroscience*, *34*(32), 10645–10658 (cit. on p. 34).

Vintch, B., Movshon, J. A., & Simoncelli, E. P. (2015a). A convolutional subunit model for neuronal responses in macaque V1. *Journal of Neuroscience*, *35*, 14829–14841. https://doi.org/10.1523/JNEUROSCI.2815-13.2015 (cit. on pp. 6, 19).

Vintch, B., Movshon, J. A., & Simoncelli, E. P. (2015b). A convolutional subunit model for neuronal responses in macaque v1. *The Journal of Neuroscience*, *35*(44), 14829–14841 (cit. on p. 6).

Von Helmholtz, H. (1867). *Handbuch der physiologischen optik: Mit 213 in den text eingedruckten holzschnitten und 11 tafeln* (Vol. 9). Voss. (Cit. on p. 46).

Walker, E. Y., Sinz, F. H., Cobos, E., Muhammad, T., Froudarakis, E., Fahey, P. G., Ecker, A. S., Reimer, J., Pitkow, X., & Tolias, A. S. (2019). Inception loops discover what excites neurons most using deep predictive models. *Nature neuroscience*, *22*(12), 2060–2065 (cit. on pp. 7, 13, 25, 36, 43).

Wang, E. Y., Fahey, P. G., Ponder, K., Ding, Z., Chang, A., Muhammad, T., Patel, S., Ding, Z., Tran, D., Fu, J., et al. (2023). Towards a foundation model of the mouse visual cortex. *bioRxiv* (cit. on p. 41).

Wang, Q., & Burkhalter, A. (2007). Area map of mouse visual cortex. *Journal of Comparative Neurology*, *502*(3), 339–357 (cit. on p. 21).

Wang, Q., Sporns, O., & Burkhalter, A. (2012). Network analysis of corticocortical connections reveals ventral and dorsal processing streams in mouse visual cortex. *Journal of Neuroscience*, *32*(13), 4386–4399 (cit. on p. 34).

Weis, M. A., Pede, L., Lüddecke, T., & Ecker, A. S. (2023). Self-supervised graph representation learning for neuronal morphologies. *Transactions on Machine Learning Research* (cit. on p. 46).

Wichmann, F. A., Kornblith, S., & Geirhos, R. (2023). Neither hype nor gloom do dnns justice. *The Behavioral and brain sciences*, *46* (cit. on p. 33).

Willeke, K. F., Restivo, K., Franke, K., Nix, A. F., Cadena, S. A., Shinn, T., Nealley, C., Rodriguez, G., Patel, S., Ecker, A. S., & Tolias, A. (2023). Deep learning-driven characterization of single cell tuning in primate visual area v4 unveils topological organization. *bioRxiv*, 2023–05 (cit. on pp. 7, 13, 23, 25, 32, 36, 43, 45).

Willmore, B., Prenger, R. J., Wu, M. C.-K., & Gallant, J. L. (2008). The berkeley wavelet transform: A biologically inspired orthogonal wavelet transform. *Neural computation*, *20*(6), 1537–1564 (cit. on p. 19).

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annual Review of Neuroscience*, *29*, 477–505. https://doi.org/10.1146/annurev.neuro.29.051605.113024 (cit. on pp. 1, 5).

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014a). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111 (cit. on pp. 9, 10, 13, 14, 20, 21, 23).

Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, *19*(3), 356–365 (cit. on pp. 9, 13, 19, 32).

Yamins, D. L., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014b). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624 (cit. on pp. 8, 9).

Ye, J., Collinger, J., Wehbe, L., & Gaunt, R. (2024). Neural data transformer 2: Multi-context pretraining for neural spiking activity. *Advances in Neural Information Processing Systems*, *36* (cit. on p. 41).

Yovel, G., & Abudarham, N. (2023). Why psychologists should embrace rather than abandon dnns. *The Behavioral and brain sciences*, *46*, e414 (cit. on p. 33).

Zamir, A. R., Sax, A., Shen, W., Guibas, L. J., Malik, J., & Savarese, S. (2018). Taskonomy: Disentangling task transfer learning. *Proceedings of the IEEE conference on computer vision and pattern recognition*, 3712–3722 (cit. on pp. 14, 23, 35).

Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. *European conference on computer vision*, 818–833 (cit. on p. 28).

Zhang, S., Xu, M., Kamigaki, T., Hoang Do, J. P., Chang, W.-C., Jenvay, S., Miyamichi, K., Luo, L., & Dan, Y. (2014). Long-range and local circuits for top-down modulation of visual cortex processing. *Science*, *345*(6197), 660–665 (cit. on p. 11).

Zhou, A., Knowles, T., & Finn, C. (2020). Meta-learning symmetries by reparameterization. *arXiv preprint arXiv:2007.02933* (cit. on p. 46).

Zhu, M., & Rozell, C. J. (2013). Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. *PLoS Computational Biology*, *9*(8), e1003191 (cit. on p. 27).

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3) (cit. on pp. 23, 39).

Zhuang, J., Ng, L., Williams, D., Valley, M., Li, Y., Garrett, M., & Waters, J. (2017). An extended retinotopic map of mouse cortex. *elife*, *6*, e18372 (cit. on p. 21).

Zimmermann, R. S., Borowski, J., Geirhos, R., Bethge, M., Wallis, T., & Brendel, W. (2021). How well do feature visualizations support causal understanding of cnn activations? *Advances in Neural Information Processing Systems*, *34*, 11730–11744 (cit. on p. 37).

Zipser, D., & Andersen, R. A. (1988). A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. *Nature*, *331*(6158), 679 (cit. on p. 6).

Zoccolan, D., Oertelt, N., DiCarlo, J. J., & Cox, D. D. (2009). A rodent model for the study of invariant visual object recognition. *Proceedings of the National Academy of Sciences*, *106*(21), 8748–8753 (cit. on p. 34).

APPENDIX

A.1 Deep convolutional models improve predictions of macaque V1 responses to natural images

> **Cadena, Santiago A**, Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, *15*(4), e1006897

Abstract

Despite great efforts over several decades, our best models of primary visual cortex (V1) still predict spiking activity quite poorly when probed with natural stimuli, highlighting our limited understanding of the nonlinear computations in V1. Recently, two approaches based on deep learning have emerged for modeling these nonlinear computations: transfer learning from artificial neural networks trained on object recognition and data-driven convolutional neural network models trained end-to-end on large populations of neurons. Here, we test the ability of both approaches to predict spiking activity in response to natural images in V1 of awake monkeys. We found that the transfer learning approach performed similarly well to the data-driven approach and both outperformed classical linear-nonlinear and wavelet-based feature representations that build on existing theories of V1. Notably, transfer learning using a pretrained feature space required substantially less experimental time to achieve the same performance. In conclusion, multi-layer convolutional neural networks (CNNs) set the new state of the art for predicting neural responses to natural images in primate V1 and deep features learned for object recognition are better explanations for V1 computation than all previous filter bank theories. This finding strengthens the necessity of V1 models that are multiple nonlinearities away from the image domain and it supports the idea of explaining early visual cortex based on high-level functional goals.

Author contributions

Conceptualization: **SC**, AE, AT, MB. Data Curation: **SC** Formal Analysis: **SC**, GD, EW, AT, AE. Funding Acquisition: AT, AE, MB. Investigation: **SC**, GD, EW, LG, AT, MB, AE. Methodology: **SC**, GD, EW, LG, AT, MB, AE. Project Administration: **SC**, AE, AT. Resources: EW, AT, MB, AE. Software: **SC**, EW, LG, AE. Supervision: AE, AT, MB. Validation: **SC**, MB, AE. Visualization: **SC**. Writing – Original Draft Preparation: **SC**, AT, MB, AE. Writing – Review and Editing: **SC**, GD, LG, AT, MB, AE.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Deep convolutional models improve predictions of macaque V1 responses to natural images

Santiago A. Cadena[1,2,3]*, **George H. Denfield**[3,4], **Edgar Y. Walker**[3,4], **Leon A. Gatys**[1,2], **Andreas S. Tolias**[2,3,4,5☯], **Matthias Bethge**[1,2,3,6☯], **Alexander S. Ecker**[1,2,3☯]

1 Centre for Integrative Neuroscience and Institute for Theoretical Physics, University of Tübingen, Tübingen, Germany, **2** Bernstein Center for Computational Neuroscience, Tübingen, Germany, **3** Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, Texas, United States of America, **4** Department of Neuroscience, Baylor College of Medicine, Houston, Houston, Texas, United States of America, **5** Department of Electrical and Computer Engineering, Rice University, Houston, Houston, Texas, United States of America, **6** Max Planck Institute for Biological Cybernetics, Tübingen, Germany

☯ These authors contributed equally to this work.
* santiago.cadena@uni-tuebingen.de

Citation: Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, et al. (2019) Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Comput Biol 15(4): e1006897. https://doi.org/10.1371/journal.pcbi.1006897

Editor: Wolfgang Einhäuser, Technische Universitat Chemnitz, GERMANY

Received: February 19, 2018

Accepted: February 21, 2019

Published: April 23, 2019

Data Availability Statement: The data used for this study is available in a GIN repository https://doid.gin.g-node.org/2e31e304e03d6357c98ac735a1fe5788/ (DOI: 10.12751/g-node.2e31e3). The code used to train all models in this study is available in the following repository: https://github.com/sacadena/Cadena2019PlosCB.

Funding: Research reported in this publication was supported by the German Research Foundation (DFG) grant EC 479/1-1 to A.S.E [http://www.dfg.de/] and the Collaborative Research Center (SFB

Abstract

Despite great efforts over several decades, our best models of primary visual cortex (V1) still predict spiking activity quite poorly when probed with natural stimuli, highlighting our limited understanding of the nonlinear computations in V1. Recently, two approaches based on deep learning have emerged for modeling these nonlinear computations: transfer learning from artificial neural networks trained on object recognition and data-driven convolutional neural network models trained end-to-end on large populations of neurons. Here, we test the ability of both approaches to predict spiking activity in response to natural images in V1 of awake monkeys. We found that the transfer learning approach performed similarly well to the data-driven approach and both outperformed classical linear-nonlinear and wavelet-based feature representations that build on existing theories of V1. Notably, transfer learning using a pre-trained feature space required substantially less experimental time to achieve the same performance. In conclusion, multi-layer convolutional neural networks (CNNs) set the new state of the art for predicting neural responses to natural images in primate V1 and deep features learned for object recognition are better explanations for V1 computation than all previous filter bank theories. This finding strengthens the necessity of V1 models that are multiple nonlinearities away from the image domain and it supports the idea of explaining early visual cortex based on high-level functional goals.

Author summary

Predicting the responses of sensory neurons to arbitrary natural stimuli is of major importance for understanding their function. Arguably the most studied cortical area is primary visual cortex (V1), where many models have been developed to explain its function. However, the most successful models built on neurophysiologists' intuitions still fail to account

1233, Robust Vision [https://uni-tuebingen.de/en/research/core-research/collaborative-research-centers/sfb-1233.html]); the Bernstein Center for Computational Neuroscience (FKZ 01GQ1002 [https://www.bccn-tuebingen.de/]); the German Excellency Initiative through the Centre for Integrative Neuroscience Tübingen (EXC307 [https://www.cin.uni-tuebingen.de/]); the National Eye Institute of the National Institutes of Health under Award Numbers R01EY026927 [https://nei.nih.gov/](A.S.T.), DP1 EY023176 (A.S.T.), and NIH-Pioneer award DP1-OD008301 [https://commonfund.nih.gov/pioneer/](A.S.T). This research was also supported by NEI/NIH Core Grant for Vision Research (EY-002520-37 [https://nei.nih.gov/]), NEI training grant T32EY00700140 [https://nei.nih.gov/](G.H.D) and F30EY025510 (E. Y.W.). This research was also supported by Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003 [www.iarpa.gov]. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared no competing interests.

for spiking responses to natural images. Here, we model spiking activity in primary visual cortex (V1) of monkeys using deep convolutional neural networks (CNNs), which have been successful in computer vision. We both trained CNNs directly to fit the data, and used CNNs trained to solve a high-level task (object categorization). With these approaches, we are able to outperform previous models and improve the state of the art in predicting the responses of early visual neurons to natural images. Our results have two important implications. First, since V1 is the result of several nonlinear stages, it should be modeled as such. Second, functional models of entire visual pathways, of which V1 is an early stage, do not only account for higher areas of such pathways, but also provide useful representations for V1 predictions.

Introduction

An essential step towards understanding visual processing in the brain is building models that accurately predict neural responses to arbitrary stimuli [1]. Primary visual cortex (V1) has been a strong focus of sensory neuroscience ever since Hubel and Wiesel's seminal studies demonstrated that neurons in primary visual cortex (V1) respond selectively to distinct image features like local orientation and contrast [2, 3]. Our current standard model of V1 is based on linear-nonlinear models (LN) [4, 5] and energy models [6] to explain simple and complex cells, respectively. While these models work reasonably well to model responses to simple stimuli such as gratings, they fail to account for neural responses to more complex patterns [7] and natural images [8, 9]. Moreover, the computational advantage of orientation-selective LN neurons over simple center-surround filters found in the retina would be unclear [10].

There are a number of hypotheses about nonlinear computations in V1, including normative models like overcomplete sparse coding [11, 12] or canonical computations like divisive normalization [13, 14]. The latter has been used to explain specific phenomena such as center-surround interactions with carefully designed stimuli [15–18]. However, to date, these ideas have not been turned into predictive models of spiking responses that generalize beyond simple stimuli—especially to natural images.

To go beyond simple LN models for natural stimuli, LN-LN cascade models have been proposed, which either learn (convolutional) subunits [19–21] or use handcrafted wavelet representations [22]. These cascade models outperform simple LN models, but they currently do not capture the full range of nonlinearities observed in V1, like gain control mechanisms and potentially other not-yet-understood nonlinear response properties. Because experimental time is limited, LN-LN models have to be designed very carefully to keep the number of parameters tractable, which currently limits their expressiveness, essentially, to energy models for direction-selective and complex cells.

Thus, to make progress in a quantitative sense, recent advances in machine learning and computer vision using deep neural networks ('deep learning') have opened a new door by allowing us to learn much more complex nonlinear models of neural responses. There are two main approaches, which we refer to as *goal-driven* and *data-driven*.

The idea behind the goal-driven approach is to train a deep neural network on a high-level task and use the resulting intermediate representations to model neural responses [23, 24]. In the machine learning community, this concept is known as transfer learning and has been very successful in deep learning [25, 26]. Deep convolutional neural networks (CNNs) have reached human-level performance on visual tasks like object classification by training on over one million images [27–30]. These CNNs have proven extremely useful as nonlinear feature

spaces for tasks where less labeled data is available [25, 31]. This transfer to a new task can be achieved by (linearly) reading out the network's internal representations of the input. Yamins, DiCarlo and colleagues showed recently that using deep networks trained on large-scale object recognition as nonlinear feature spaces for neural system identification works remarkably well in *higher areas* of the ventral stream, such as V4 and IT [32, 33]. Other groups have used similar approaches for early cortical areas using fMRI [34–36]. However, this approach has not yet been used to model spiking activity of early stages such as V1.

The deep data-driven approach, on the other hand, is based on fitting all model parameters directly to neural data [37–41]. The most critical advance of these models in neural system identification is that they can have many more parameters than the classical LN cascade models discussed above, because they exploit computational similarities between different neurons [38, 40]. While previous approaches treated each neuron as an individual multivariate regression problem, modern CNN-based approaches learn one model for an entire population of neurons, thereby exploiting two key properties of local neural circuits: (1) they share the same presynaptic circuitry (for V1: retina and LGN) [38] and (2) many neurons perform essentially the same computation, but at different locations (topographic organization, implemented by convolutional weight sharing) [39–41].

While both the goal-driven and the data-driven approach have been shown to outperform LN models in some settings, neither approach has been evaluated on spiking activity in monkey V1 (see [42, 43] for concurrent work). In this paper, we fill this gap and evaluate both approaches in monkey V1. We found that deep neural networks lead to substantial performance improvements over older models. In our natural image dataset, goal-driven and data-driven models performed similarly well. The goal-driven approach reached this performance with as little as 20% of the dataset and its performance saturated thereafter. In contrast, the data-driven approach required the full dataset for maximum performance, suggesting that it could benefit from a larger dataset and reach even better performance. Our key finding is that the best models required at least four nonlinear processing steps, suggesting that we need to revise our view of V1 as a Gabor filter bank and appreciate the nonlinear nature of its computations. We conclude that deep networks are not just one among many approaches that can be used, but are—despite their limitations—currently the single most accurate model of V1 computation.

Results

We measured the spiking activity of populations of neurons in V1 of two awake, fixating rhesus macaques using a linear 32-channel array spanning all cortical layers. Monkeys were viewing stimuli that consisted of 1450 natural images and four sets of textures synthesized to keep different levels of higher-order correlations present in these images (Fig 1, see Methods). Each trial consisted of a sequence of images shown for 60 ms each, with no blanks in between. In each session, we centered the stimuli on the population receptive field of the neurons.

We isolated 262 neurons in 17 sessions. The neurons responded well to the fast succession of natural images with a typical response latency of 40ms (Fig 2B). Therefore, we extracted the spike counts in the window 40–100 ms after image onset (Fig 2B). The recorded neurons were diverse in their temporal response properties (e.g. see autocorrelogram Fig 2A), average firing rates in response to stimulus (21.1 ± 20.8 spikes/s, mean ± S.D.), cortical depth (55% of cells in granular, 18% in supragranular, and 27% infragranular layers), and response-triggered average (RTA) structure (Fig 2C), but neurons recorded on the same array generally had their receptive fields at similar locations approximately centered on the stimulus (Fig 2C). Prior to analysis, we selected neurons based on how reliable their responses were from trial to trial and included only neurons for which at least 15% of their total variance could be attributed to the

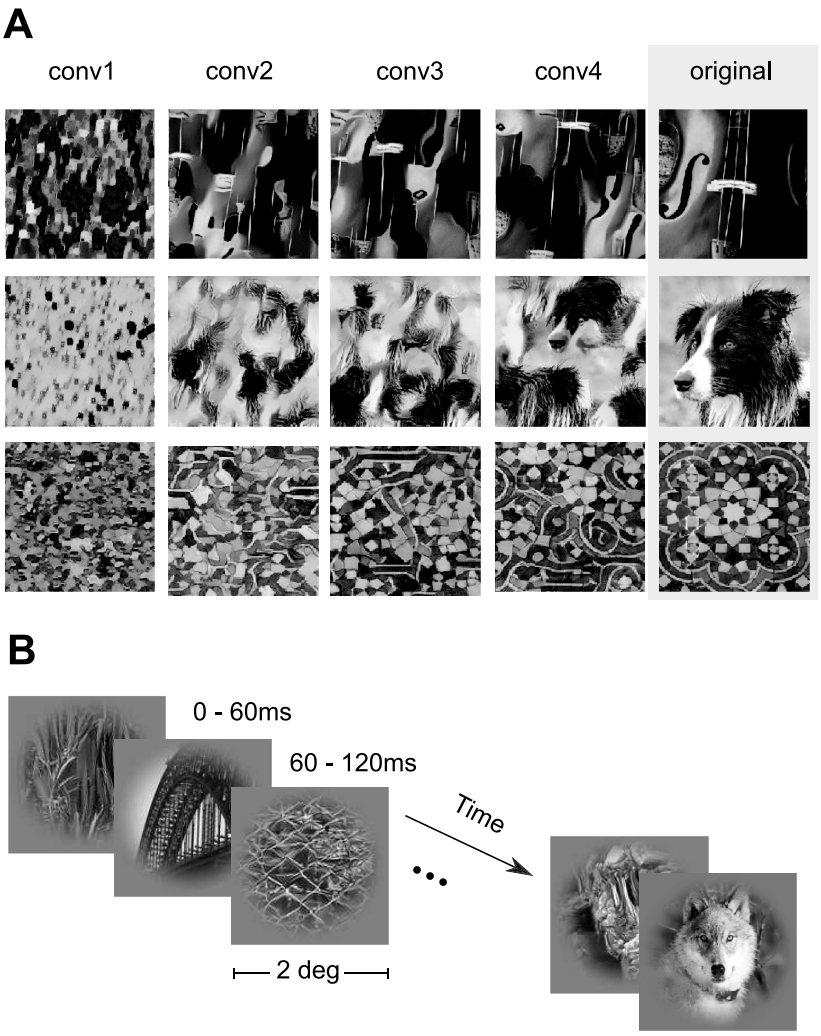


Fig 1. Stimulus paradigm. A: Classes of images shown in the experiment. We used grayscale natural images (labeled 'original') from the ImageNet dataset [44] along with textures synthesized from these images using the texture synthesis algorithm described by [45]. Each row shows four synthesized versions of three example original images using different convolutional layers (see Materials and Methods for details). Lower convolutional layers capture more local statistics compared to higher ones. B: Stimulus sequence. In each trial, we showed a randomized sequence of images (each displayed for 60 ms covering 2 degrees of visual angle) centered on the receptive fields of the recorded neurons while the monkey sustained fixation on a target. The images were masked with a circular mask with cosine fadeout.

https://doi.org/10.1371/journal.pcbi.1006897.g001

stimulus (see Methods). This selection resulted in 166 neurons, which form the basis of the models we describe in the following.

Generalized linear model with pre-trained CNN features

We start by investigating the goal-driven approach [23, 24]. Here, the idea is to use a high-performing neural network trained on a specific goal—object recognition in this case—as a non-

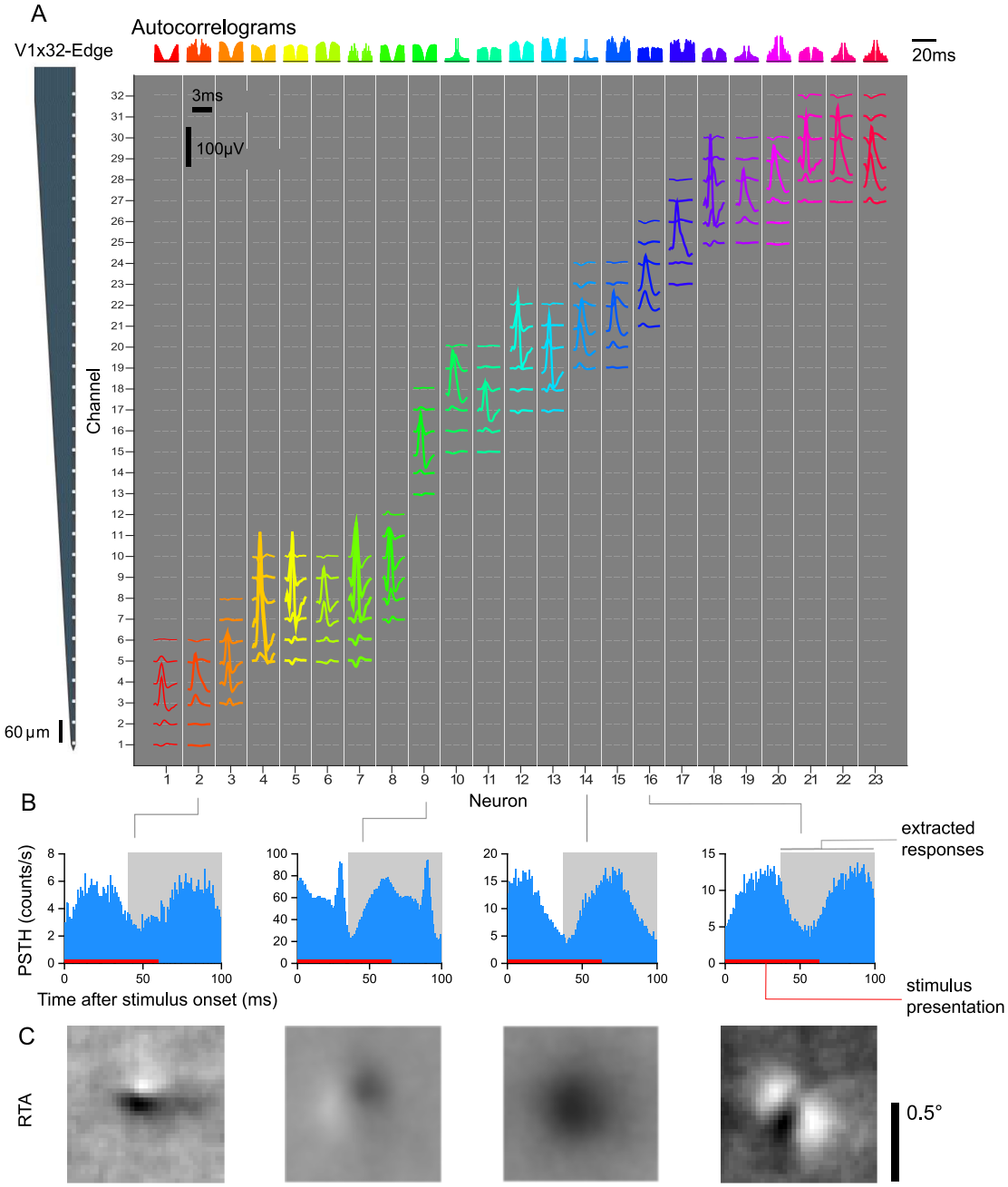


Fig 2. V1 electrophysiological responses. A: Isolated single-unit activity. We performed acute recordings with a 32-channel, linear array (NeuroNexus V1x32-Edge-10mm-60-177, layout shown in the left) to record in primary visual cortex of two awake, fixating macaques. The channel mean-waveform footprints of the spiking activity of 23 well-isolated neurons in one example session are shown in the central larger panel. The upper panel shows color-matched autocorrelograms. B: Peri-stimulus time histograms (PSTH) of four example neurons from A. Spike counts where binned with t = 1 ms, aligned to the onset of each stimulus image, and averaged over trials. The 60 ms interval where the image was displayed is shown in red. We ignored the temporal profile of the response and extracted spike counts for each image on the 40–100 ms interval after image onset (shown in light gray). C: The Response Triggered Average (RTA) calculated by reverse correlation of the extracted responses.

https://doi.org/10.1371/journal.pcbi.1006897.g002

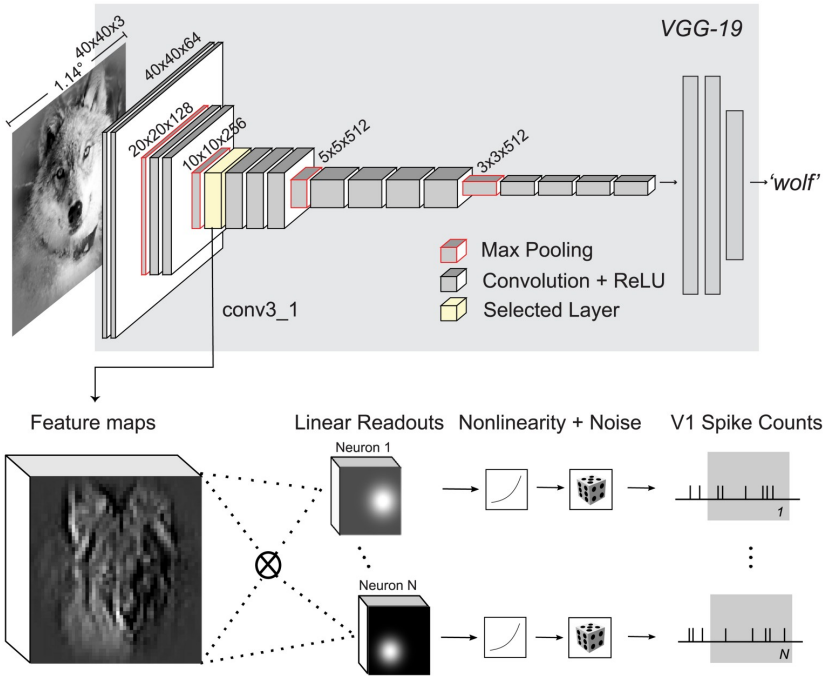


Fig 3. Our proposed model based on VGG-19 features. VGG-19 [28] (gray background) is a trained CNN that takes an input image and produces a class label. For each of the 16 convolutional layers of VGG-19, we extract the feature representations (feature maps) of the images shown to the monkey. We then train for each recorded neuron and convolutional layer, a Generalized Linear Model (GLM) using the feature maps as input to predict the observed spike counts. The GLM is formed by a linear projection (dot product) of the feature maps, a pointwise nonlinearity, and an assumed noise distribution (Poisson) that determines the optimization loss for training. We additionally imposed strong regularization constraints on the readout weights (see text).

https://doi.org/10.1371/journal.pcbi.1006897.g003

linear feature space and train only a simple linear-nonlinear readout. We chose VGG-19 [28] over other neural networks, because it has a simple architecture (described below), a fine increase in receptive field size along its hierarchy and reasonably high classification accuracy.

VGG-19 is a CNN trained on the large image classification task ImageNet (ILSVRC2012) that takes an RGB image as input and infers the class of the dominant object in the image (among 1000 possible classes). The architecture of VGG-19 consists of a hierarchy of linear-nonlinear transformations (layers), where the input is spatially convolved with a set of filters and then passed through a rectifying nonlinearity (Fig 3). The output of this operation is again an image with multiple channels. However, these channels do not represent color—as the three channels in the input image—but learned features. They are therefore also called feature maps. Each feature map can be viewed as a filtered version of its input. The collection of such feature maps serves as input for the next layer. Additionally, the network has five pooling layers, where the feature maps are downsampled by a factor of two by taking the local maximum value of four neighboring pixels. There are 16 convolutional layers that can be grouped into five groups named conv1 to conv5 with 2, 2, 4, 4, 4 convolutional layers and 64, 128, 256, 512, 512 output feature maps, respectively, and a pooling layer after each group.

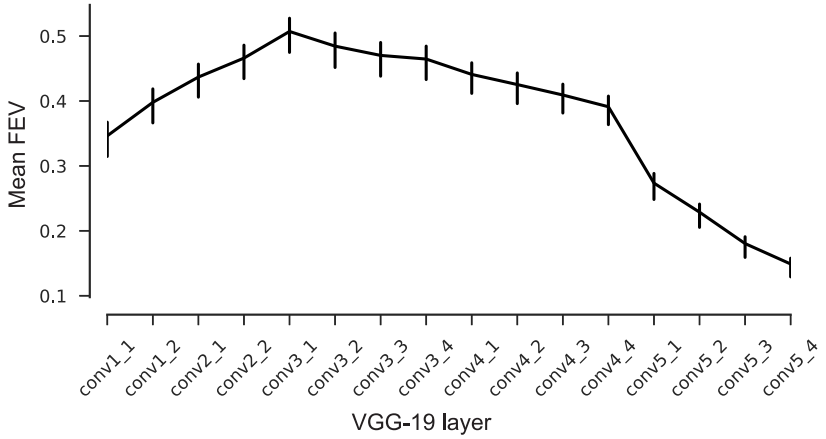


Fig 4. Model performance on test set. Average fraction of explainable variance explained (*FEV*) for models using different VGG layers as nonlinear feature spaces for a GLM. Error bars show 95% confidence intervals of the population means. The model based on layer conv3_1 shows on average the highest predictive performance.

https://doi.org/10.1371/journal.pcbi.1006897.g004

We used VGG-19 as a feature space in the following way: We selected the output of a convolutional layer as input features for a Generalized Linear Model (GLM) that predicts the recorded spike counts (Fig 3). Specifically, we fed each image x in our stimulus set through VGG-19 to extract the resulting feature maps $\Phi(x)$ of a certain layer. These feature maps were then linearly weighted with a set of learned readout weights w. This procedure resulted in a single scalar value for each image that was then passed through a (static) output nonlinearity to produce a prediction for the firing rate:

$$r(x) = \exp\left[w^T \Phi(x) + b\right] \quad (1)$$

We assumed this prediction to be the mean rate of a Poisson process (see Methods for details). In addition, we applied a number of regularization terms on the readout weights that we explain later.

Intermediate layers of VGG best predict V1 responses

We first asked which convolutional layer of VGG-19 provides the best feature space for V1. To answer this question, we fitted a readout for each layer and compared the performance. We measured performance by computing the fraction of explainable variance explained (*FEV*). This metric, which ranges from zero to one, measures what fraction of the stimulus-driven response is explained by the model, ignoring the unexplainable trial-to-trial variability in the neurons' responses (for details see Methods).

We found that the fifth (out of sixteen) layers' features (called 'conv3_1', Fig 3) best predicted neuronal responses to novel images not seen during training (Fig 4, solid line). This model predicted on average 51.6% of the explainable variance. In contrast, performance for the very first layer was poor (31% FEV), but increased monotonically up to conv3_1. Afterwards, the performance again decreased continually up the hierarchy (Fig 4). These results followed our intuition that early to intermediate processing stages in a hierarchical model should match primary visual cortex, given that V1 is the third processing stage in the visual hierarchy after the retina and the lateral geniculate nucleus (LGN) of the thalamus.

Control for input resolution and receptive field sizes

An important issue to be aware of is that the receptive field sizes of VGG units grow along the hierarchy—just like those of visual neurons in the brain. Incidentally, the receptive fields of units in the best-performing layer conv3_1 subtended approximately 0.68 degrees of visual angle, roughly matching the expected receptive sizes of our V1 neurons given their eccentricities between 1 and 3 degrees. Because receptive fields in VGG are defined in terms of image pixels, their size in degrees of visual angle depends on the resolution at which we present images to VGG, which is a free parameter whose choice will affect the results.

VGG-19 was trained on images of 224 × 224 px. Given the image resolution we used for the analyses presented above, an entire image would subtend ∼6.4 degrees of visual angle (the crops shown to the monkey were 2 degrees; see Methods for details). Although this choice appears to be reasonable and consistent with earlier work [33], it is to some extent arbitrary. If we had presented the images at lower resolution, the receptive fields sizes of all VGG units would have been larger. As a consequence, the receptive fields of units in earlier layers would match those of V1 and these layers may perform better. If this was indeed the case, there would be nothing special about layer conv3_1 with respect to V1.

To ensure that the choice of input resolution did not affect our results, we performed a control experiment, which substantiated our claim that conv3_1 provides the best features for V1. We repeated the model comparison presented above with different input resolutions, rescaling the image crops by a factor of 0.67 and 1.5. These resolutions correspond to 9.55 and 4.25 degrees of full visual field for VGG-19, respectively. While changing the input resolution did shift the optimal layer towards that with matching receptive field sizes (Fig 5, first and third row), the resolution we had picked for our main experiment yielded the best overall performance (Fig 5, second row, third column). Thus, over a range of input resolutions and layers, conv3_1 performed best, although conv2_2 at lower resolution yielded only slightly lower performance.

Careful regularization is necessary

The number of predictors given by the convolutional feature space of a large pre-trained network is much larger than the number of pixels in the image. Most of these predictors will likely be irrelevant for most recorded neurons—for example, network units at spatial positions that are not aligned with the neuron's receptive field or feature maps that compute nonlinearities

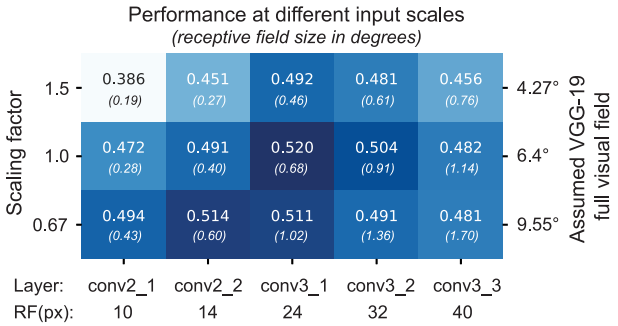


Fig 5. VGG-19 based model performance at different input scales. The performance on test set of cross-validated models that use as feature spaces layers conv2_1 to conv3_3 for different input resolutions. With the original scale used in Fig 4, we assumed that VGG-19 was trained with 6.4 degrees field of view. Scaling this resolution by a factor of 0.67 and 1.5 justifies the original choice of resolution for further analysis. At the bottom, the receptive field sizes in pixels of the different layers are shown.

https://doi.org/10.1371/journal.pcbi.1006897.g005

Table 1. Ablation experiments for VGG-based model, removing regularization terms (rows 2–5) and using factorized readout weights (row 6, [40]).

Model	FEV
Full model	**0.52**
No smoothness	0.51
No sparsity	0.49
No group sparsity	0.51
No regularization	0.33
Factorized readout [40]	0.45

https://doi.org/10.1371/journal.pcbi.1006897.t001

unrelated to those of the cells. Naïvely including many unimportant predictors would prevent us from learning a good mapping, because they lead to overfitting. We therefore used a regularization scheme with the following three terms for the readout weights: (1) sparsity, to encourage the selection of a few units; (2) smoothness, for a regular spatial continuity of the predictors' receptive fields; and (3) group sparsity, to encourage the model to pool from a small number of feature maps (see Methods for details).

We found that regularization was key to obtaining good performance (Table 1). The full model with all three terms had the best performance on the test set and vastly outperformed a model with no regularization. Eliminating one of the three terms while keeping the other two hurt performance only marginally. Among the three regularizers, sparsity appeared to be the most important one quantitatively, whereas smoothness and group sparsity could be dropped without hurting overall performance.

To understand the effect of the different regularizers qualitatively, we visualized the readout weights of each feature map of our conv3_1-based model, ordered by their spatial energy for each cell, for each of the regularization schemes (see Fig 6A for five sample neurons). Without the sparsity constraint, we obtained smooth but spread-out weights that were not well localized. Dropping the smoothness term—despite performing equally in a quantitative sense—produced sparse activations that were less localized and not smooth. Without any regularization, the weights appeared noisy and one could not get any insights about the locality of the neuron. On the other hand, the full model—in addition to having the best performance—also

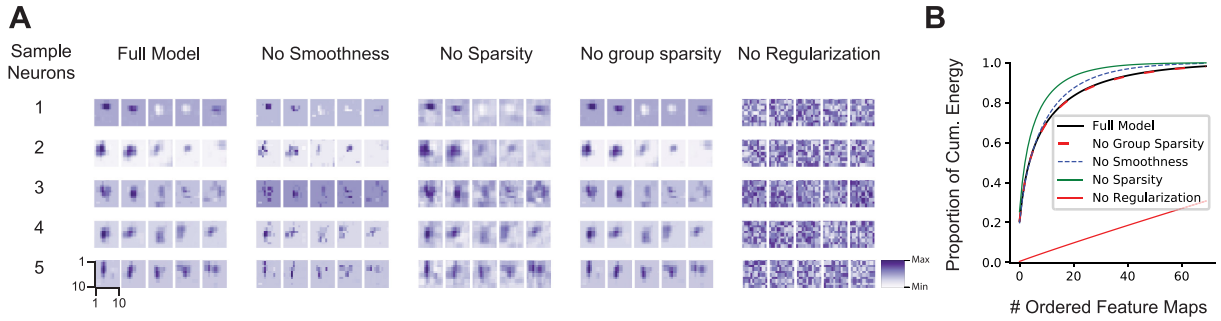


Fig 6. Learned readout weights with different regularization modes. A. For five example neurons (rows), the five highest-energy spatial readouts out of 256 feature maps of conv3_1 for each regularization scheme we explored. Feature map weights are 10 × 10 for a 40 × 40 input (∼ 1.1˚). The full model exhibits the most localized and smooth receptive fields. The scale bar is shared by all features maps of a each model and neuron from their minimum (white) to the maximum (dark). **B**. The highest normalized spatial energy of the learned readouts as a function of ordered feature maps (first 70 out of 256 of conv3_1 shown) averaged for all cells. With regularization, only a few feature maps are used for prediction, quickly asymptoting at 1.

https://doi.org/10.1371/journal.pcbi.1006897.g006

provides localized and smooth filters that provide information about the neurons' receptive field and the set of useful feature maps for prediction.

Finally, we also observed that only a small number of feature maps was used for each neuron: the weights decayed exponentially and only 20 feature maps out of 256 contained on average 82% of the readout energy (Fig 6B).

An alternative form of regularization or inductive bias would be to constrain the readout weights to be factorized in space and features [40], which reduces the number of parameters substantially. However, the best model with this factorized readout achieved only 45.5% FEV (Table 1), presumably because the feature space has not been optimized for such a constrained readout.

Goal-driven and data driven CNNs set the state of the art

Multi-layer feedforward networks have been fitted successfully to neural data on natural image datasets in mouse V1 [38, 40]. Thus, we inquired how our goal-driven model compares to a model belonging to the same functional class, but directly fitted to the neural data. Following the methods proposed by Klindt et. al [40], we fitted CNNs with one to five convolutional layers (Fig 7A; see Methods for details).

The data-driven CNNs with three or more convolutional layers yielded the best performance, outperforming their competitors with fewer (one or two) layers (Fig 7B). We therefore decided to use the CNN with three layers for model comparison, as it is the simplest model with highest predictive power on the validation set.

We then asked how the predictive performance of both data-driven and goal driven models compares to previous models of V1. As a baseline, we fitted a regularized version of the classical linear-nonlinear Poisson model (LNP; [46]). The LNP is a very popular model used to estimate the receptive field of neurons and offers interpretability and convexity for its optimization. This model gave us a good idea of the nonlinearity of the cells' responses. Additionally, we fit a model based on a handcrafted nonlinear feature space consisting of a set of Gabor wavelets [4, 47–49] and energy terms of each quadrature pair [6]. We refer to this model as the 'Gabor filter bank' (GFB). It builds upon existing knowledge about V1 function and is able to model simple and complex cells as well as linear combinations thereof. Moreover, this model is the current state of the art in the neural prediction challenge for monkey V1 responses to natural images [50] and therefore a strong baseline for a quantitative evaluation.

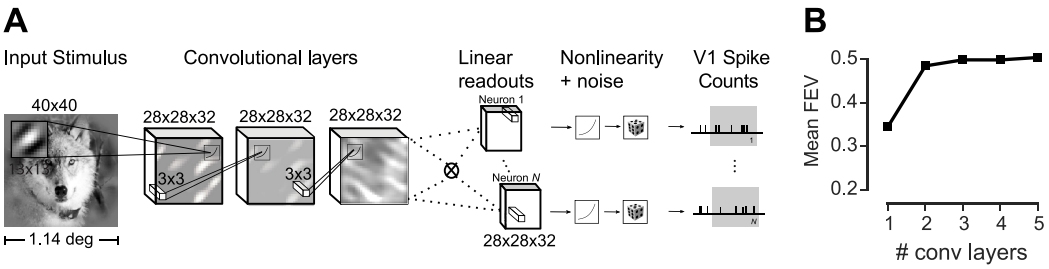


Fig 7. Data-driven convolutional network model. We trained a convolutional neural network to produce a feature space fed to a GLM-like model. In contrast to the VGG-based model, both feature space and readout weights are trained only on the neural data. **A**. Three-layer architecture with a factorized readout [40] used for comparison with other models. **B**. Performance of the data driven approach as a function of the number of convolutional layers on held-out data. Three convolutional layers provided the best performance on the validation set. See Methods for details.

https://doi.org/10.1371/journal.pcbi.1006897.g007

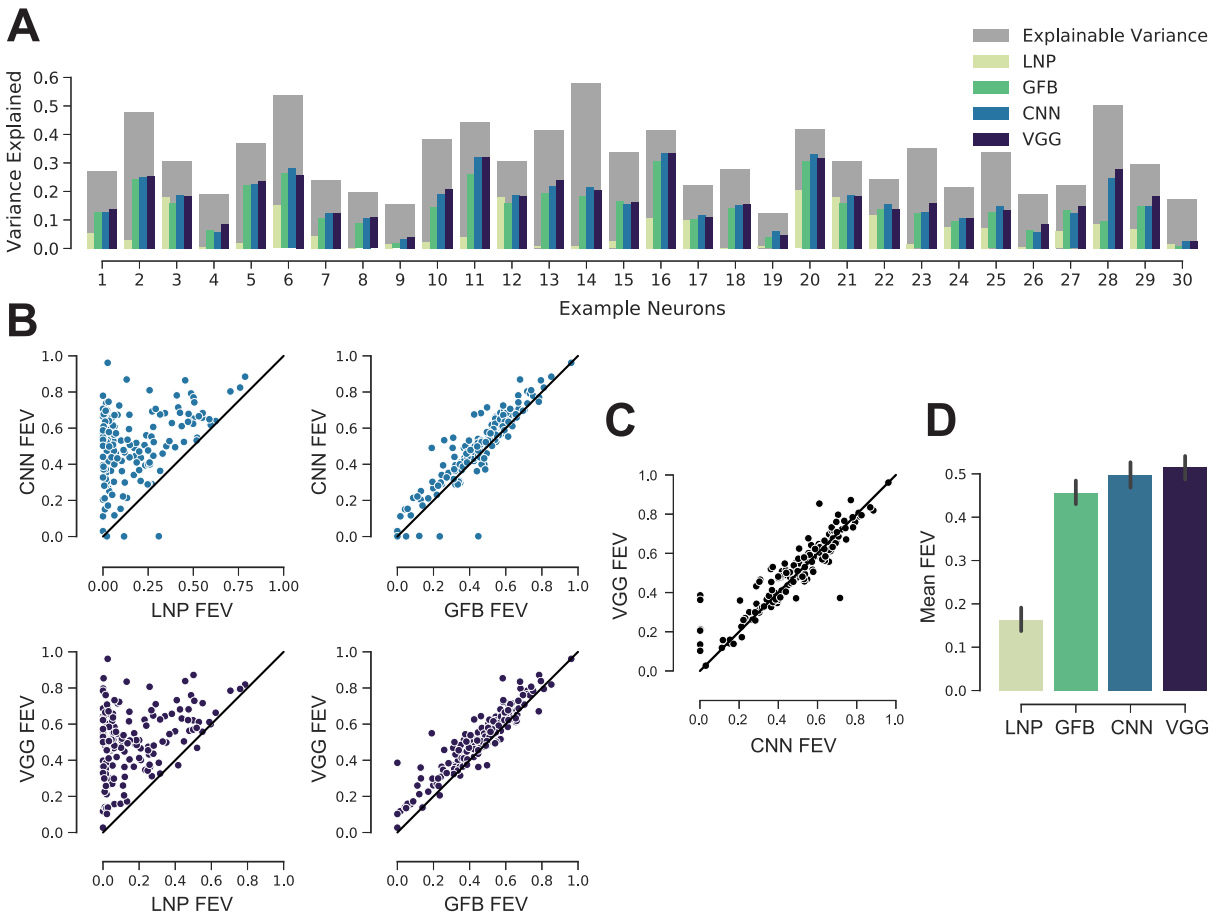


Fig 8. Deep models are the new state of the art. A: Randomly selected cells. The normalized explainable variance (oracle) per cell is shown in gray. For each cell from left to right, the variance explained of: regularized LNP [46], GFB [22, 47, 48], three-layer CNN trained on neural responses, and VGG conv3_1 model (ours). B. CNN and VGG conv3_1 models outperform for most cells LNP and GFB. Black line denotes the identity. The performance is given in *FEV*. C: VGG conv3_1 features perform slightly better than the three-layer CNN. D: Performance of the four models in fraction of explainable variance explained (*FEV*) averaged across neurons. Error bars show 95% confidence intervals of the population means. All models perform significantly different from each other (Wilcoxon signed rank test, $n = 166$; family-wise error rate $\alpha = 0.05$ using Holm-Bonferroni method to account for multiple comparisons).

https://doi.org/10.1371/journal.pcbi.1006897.g008

We compared the models for a number of cells selected randomly (Fig 8A). There was a diversity of cells, both in terms of how much variance could be explained in principle (dark gray bars) and how well the individual models performed (colored bars). Overall, the deep learning models consistently outperformed the two simpler models of V1. This trend was consistent across the entire dataset (Fig 8B and 8D). The LNP model achieved 16.3% *FEV*, the GFB model 45.6% *FEV*. The performance of the CNN trained directly on the data was comparable to that of the VGG-based model (Fig 8C and 8D); they predicted 49.8% and 51.6% *FEV*, respectively, on average. The differences in performance between all four models were statistically significant (Wilcoxon signed rank test, $n = 166$; family-wise error rate $\alpha = 0.05$ using Holm-Bonferroni method to account for multiple comparisons). Note that the one-layer CNN (mean 34.5% *FEV*, Fig 7) structurally resembles the convolutional subunit model proposed by

Vintch and colleagues [21]. Thus, deeper CNNs also outperform learned LN-LN cascade models.

Improvement of model predictions is not linked to neurons' tuning properties

We next asked whether the improvement in predictive performance afforded by our deep neural network models was related in any way to known tuning properties of V1 neurons such as the shape of their orientation tuning curve or their classification along the simple-complex axis. To investigate this question, we performed an in-silico experiment: we showed Gabor patches of the same size as our image stimulus with various orientations, spatial frequencies and phases (Fig 9A) to our CNN model of each cell. Based on the model output, we computed tuning curves for orientation (Fig 9B) and spatial phase (Fig 9D) by using the set of Gabors with the optimal spatial frequency for each neuron.

Based on the phase tuning curves we compute a linearity index (see Methods), which locates each cell on the axis from simple (linearity index close to one) to complex (index close to zero). We then asked whether there are systematic differences in model performance as a function of this simple-complex characterization. As expected, we found that more complex cells are explained better by the Gabor filter bank model than an LNP model (Fig 9C). The same was true for both the data-driven CNN and the VGG-based model. However, the simple-complex axis did not predict whether and how much the CNN models outperformed the Gabor filter bank model. Thus, whatever aspect of V1 computation was additionally explained by the CNN models, it was shared by both simple and complex cells.

Next, we asked whether there is a relationship between orientation selectivity (tuning width) and the performance of any of our models. We found that for cells with sharper orientation tuning, the performance gain afforded by the Gabor filter bank model (and both CNN-based models) over an LNP was larger than for less sharply tuned cells (Fig 9E). This result is not unexpected given that cells in layer 2/3 tend to have narrower tuning curves and also tend to be more complex [51, 52]. However, as for the simple-complex axis, tuning width was not predictive of the performance gain afforded by a CNN-based model over the Gabor filter bank (Fig 9E). Therefore, any additional nonlinearity in V1 computation captured by the CNN models is not specific to sharply or broadly tuned neurons.

Models generalize across stimulus statistics

Our stimulus set contains both natural images as well as four sets of textures generated from those images. These textures differ in how accurately and over what spatial extent they reproduce the local image statistics (see Fig 1). On the one end of the spectrum, samples from the conv1 model reproduce relatively linear statistics over small regions of a few minutes of arc. On the other end of the spectrum, samples from the conv4 model almost perfectly reproduce the statistics of natural images over larger regions of 1–2 degrees of visual angle, covering the entire classical and at least part of the extra-classical receptive field of V1 neurons.

We asked to what extent including these different image statistics helps or hurts building a predictive model. To answer this question, we additionally fit both the data-driven CNN model and the VGG-based model to subsets of the data containing only images from a single image type (originals or one of four texture classes). We then evaluated each of these models on all image types (Fig 10). Perhaps surprisingly, we found that using any of the four texture statistics or the original images for training lead to approximately equal performance, independent of which images were used for testing the model (Fig 10). This result held for both the VGG-based (Fig 10A) and the data-driven CNN model (Fig 10B). Thus, using the very

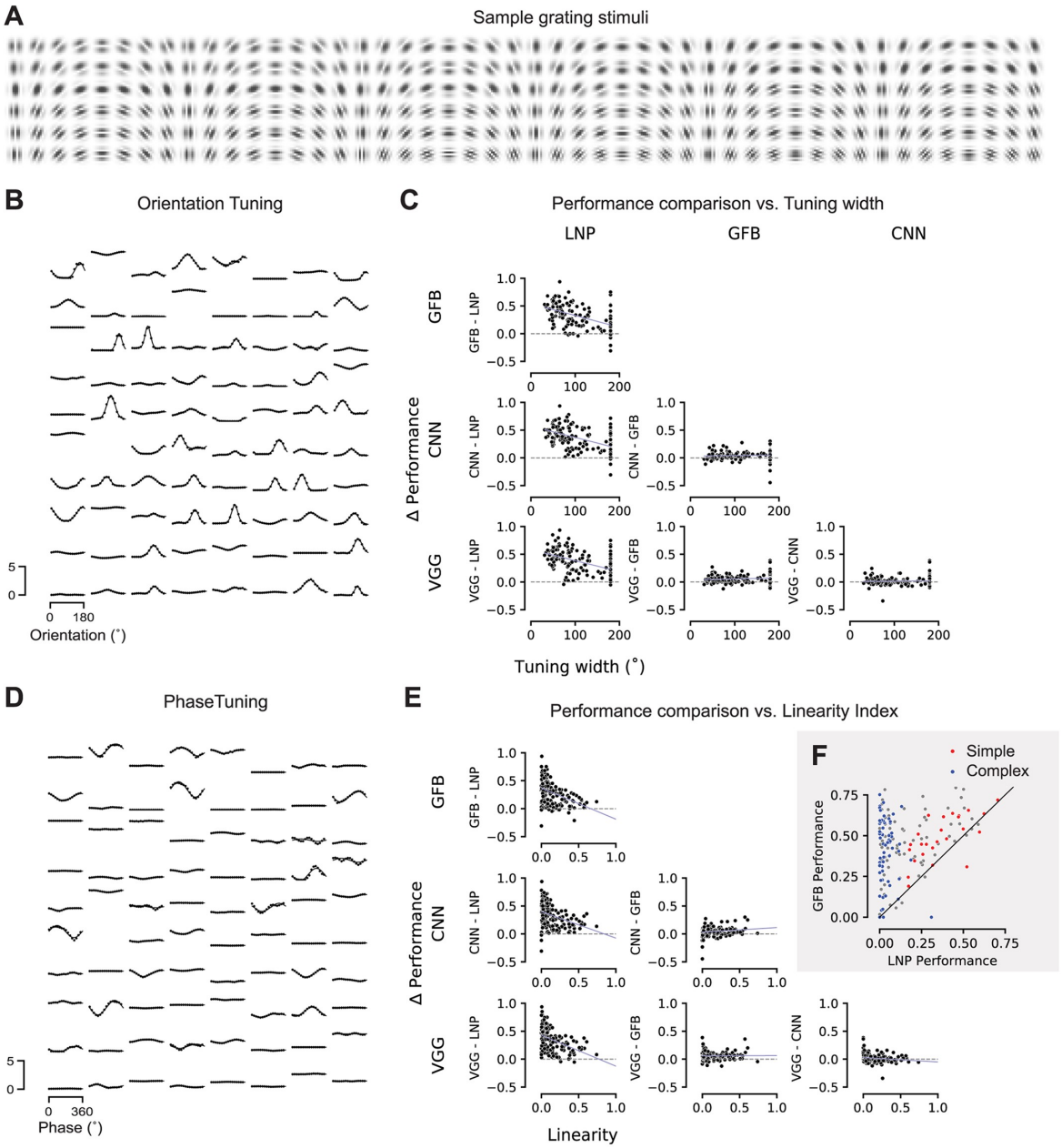


Fig 9. Relationship between model performance and neurons' tuning properties. A. A sample subset of the Gabor stimuli with a rich diversity of frequencies, orientations, and phases.**B**. Dots: Orientation tuning curves of 80 sample neurons predicted by our CNN model. Tuning curves computed at the optimal spatial frequency and phase for each neuron. Lines: von Mises fits.**C**. Difference in performance between pairs of the four models as a function of tuning width. Tuning width was defined as the full width at half maximum of the fitted tuning curve.**D**. Dots: Phase tuning curves of the same 80 sample neurons as in B, predicted by our CNN model. Tuning curves computed at the optimal spatial frequency and orientation for each neuron. Lines: Cosine tuning curve with fitted amplitude and offset (see Methods).**E**. Like C, the difference in performance between pairs of models as a function of the neurons' linearity index. Linearity index: ratio of amplitude of cosine over offset (0: complex; 1: simple).**F**. Performance comparison between GFB and LNP model. Red: simple cells (top 16% linearity, linearity > 0.3); blue: complex cells (bottom 28% linearity, linearity < 0.04).

https://doi.org/10.1371/journal.pcbi.1006897.g009

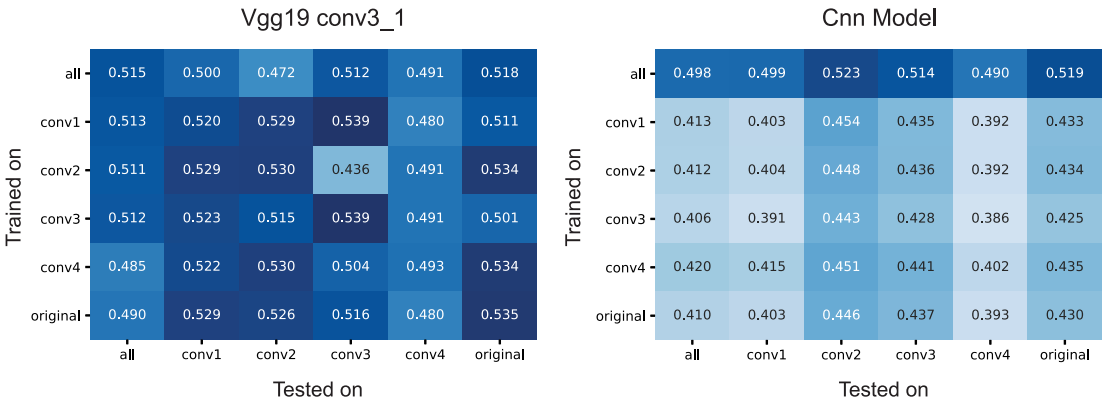


Fig 10. Training and evaluation on the different stimulus types. For both conv3_1 features of VGG-19 (left) and CNN-based models, we trained with all and every individual stimulus type (rows) (see Fig 1) and tested on all and every individual type. The VGG model showed good domain transfer in general. The same was true for the data-driven CNN model, although it performed worse overall when trained on only one set of images due to the smaller training sample. There were no substantial differences in performance across image statistics.

https://doi.org/10.1371/journal.pcbi.1006897.g010

localized conv1 textures worked just as well for predicting the responses to natural images as did training directly on natural images—or any other combination of training and test set. This result is somewhat surprising to us, as the conv1 textures match only very simple and local statistics on spatial scales smaller than individidual neurons' receptive fields and perceptually are much closer to noise than to natural images.

VGG-based model needs less training data

An interesting corollary of the analysis above is the difference in absolute performance between the VGG-based and the data-driven CNN model when using only a subset of images for training: while the performance of the VGG-based model remains equally high when using only a fifth of the data for training (Fig 10A), the data-driven CNN takes a substantial hit (Fig 10B, second and following rows). Thus, while the two models perform similarly when using our entire dataset, the VGG-based model works better when less training data is available. This result indicates that, for our current experimental paradigm, training the readout weights is not the bottleneck—despite the readout containing a large number of parameters in the VGG-based model (Table 2). Because we know that only a small number of non-zero weights are necessary, the L1 regularizer works very well in this case. In contrast, the data-driven model takes a substantial hit when using only a subset of the data, suggesting that learning the shared feature space is the bottleneck for this model. Thus collecting a larger dataset could help the data-driven model but is unlikely to improve performance of the VGG-based one.

Table 2. Number of learned parameters for the different models. 'Core' refers to the part shared among all neurons. 'Readout' refers to the parameters required for each neuron.

Model	Core	Readout/neuron	Total
LNP	-	1,601	265,766
GFB	-	5,545	920,470
CNN	23,936	867	167,858
VGG	512	25,601	4,250,278

https://doi.org/10.1371/journal.pcbi.1006897.t002

Discussion

Our goal was to find which model among various alternatives is best for one of the most studied systems in modern systems neuroscience: primary visual cortex. We fit two models based on convolutional neural networks to V1 responses to natural stimuli in awake, fixating monkeys: a goal-driven model, which uses the representations learned by a CNN trained on object recognition (VGG-19), and a data-driven model, which learns both the convolutional and readout parameters using stimulus-response pairs with multiple neurons simultaneously. Both approaches yielded comparable performance and substantially outperformed the widely used LNP [46] and a rich Gabor filter bank (GFB), which held the previous state of the art in prediction of V1 responses to natural images. This finding is of great importance because it suggests that deep neural networks can be used to model not only higher cortex, but also lower cortical areas. In fact, deep networks are not just one among many approaches that can be used, but the only class of models that has been shown to provide the multiple nonlinearities necessary to accurately describe V1 responses to natural stimuli.

Our work contributes to a growing body of research where goal-driven deep learning models [23, 24] have shown unprecedented predictive performance in higher areas of the visual stream [32, 33], and a hierarchical correspondence between deep networks and the ventral stream [35, 53]. Studies based on fMRI have established a correspondence between early layers of CNNs trained on object recognition and V1 [35, 54]. Here, with electrophysiological data and a deeper network (VGG-19), we found that V1 is better explained by feature spaces multiple nonlinearities away from the pixels. We found that it takes five layers (a quarter of the way) into the computational stack of the object categorization network to explain V1 best, which is in contrast to the many models that treat V1 as only one or two nonlinearities away from pixels (i.e. GLMs, energy models). Earlier layers of our CNNs might explain subcortical areas better (i.e. retina and LGN), as they are known to be modeled best with multiple, but fewer, nonlinearities already [41].

What are, then, the additional nonlinearities captured by our deep convolutional models beyond those in LNP or GFB? Our first attempts to answer this question via an in-silico analysis revealed that whatever the CNNs capture beyond the Gabor filter bank model is not specific to the cells' tuning properties, such as width of the orientation tuning curve and their characterization along the simple-complex spectrum. This result suggests that the missing nonlinearity may be relatively generic and applicable to most cells. There are a few clear candidates for such nonlinear computations, including divisive normalization [55] and overcomplete sparse coding [12]. Unfortunately, quantifying whether these theories provide an equally good account of the data is not straightforward: so far they have not been turned into predictive models for V1 neurons that are applicable to natural images. In the case of divisive normalization, the main challenge is learning the normalization pool. There is evidence for multiple normalization pools, both tuned and untuned and operating in the receptive field center and surround [56]. However, previous work investigating these normalization pools has employed simple stimuli such as gratings [18] and we are not aware of any work learning the entire normalization function from neural responses to natural stimuli. Similarly, sparse coding has so far been evaluated only qualitatively by showing that the learned basis functions resemble Gabor filters [12]. Solving a convolutional sparse coding problem [57] and using the resulting representation as a feature space would be a promising direction for future work, but we consider re-implementing and thoroughly evaluating this approach to be beyond the scope of the current paper.

To move forward in understanding such nonlinearities may require developing more interpretable neural networks or methods that provide interpretability of networks, which are an

active area of research in the machine learning community. Alternatively, we could build predictive models constrained with specific hard-coded nonlinearities (such as normalization) that express our knowledge about important computations.

It is also possible that the mechanistic level of circuit components remains underconstrained by function and thus allows only for explanations up to some degree of degeneracy, requiring knowledge of the objective function the system optimizes (e.g. sparse coding, predictive coding). Our results show that object categorization—despite being a relatively impoverished visual task—is a very useful learning objective not only for high-level areas in the ventral stream, but also for a more low-level and general-purpose area like V1, despite the fact that V1 clearly serves a large number of tasks beyond object categorization. This finding resonates well with results from computer vision, where object categorization has also been found to be an extremely useful objective to learn features applicable to numerous other visual tasks [25].

Our current best models still leave almost half of the explainable variance unexplained, raising the question of how to make further progress. Our finding that the VGG-based model performed equally well with only 20% of the images in the training set suggests that its performance was not limited by the amount of data available to learn the readout weights, which make for the bulk of the parameters in this model (Table 2). Instead, the VGG-based model appears to be limited by a remaining mismatch between VGG features and V1 computation. This mismatch could potentially be reduced by using features from neural networks trained simultaneously on multiple ethologically relevant tasks beyond object categorization. The data-driven model reached its full performance only with the full training set, suggesting that learning the nonlinear feature space is the bottleneck. In this case, pooling over a larger number of neurons or recording longer from the same neurons should improve performance because most of the parameters are in the shared feature space (Table 2) and this number is independent of the number of neurons being modeled.

We conclude that previous attempts to describe the basic computations that different types of neurons in primary visual cortex perform (e.g. "edge detection") do not account for the complexity of multi-layer nonlinear computations that are necessary for the performance boost achieved with CNNs. Although these models, which so far best describe these computations, are complex and lack a concise intuitive description, they can be obtained by a simple principle: optimize a network to solve an ecologically relevant task (object categorization) and use the hidden representations of such a network. For future work, combining data- and goal-driven models and incorporating the recurrent lateral and feedback connections of the neocortex promise to provide a framework for incrementally unravelling the nonlinear computations of V1 neurons.

Methods

Ethics statement

All behavioral and electrophysiological data were obtained from two healthy, male rhesus macaque (Macaca mulatta) monkeys aged 12 and 9 years and weighing 12 and 10 kg, respectively, during the time of study. All experimental procedures complied with guidelines of the NIH and were approved by the Baylor College of Medicine Institutional Animal Care and Use Committee (permit number: AN-4367). Animals were housed individually in a large room located adjacent to the training facility, along with around ten other monkeys permitting rich visual, olfactory and auditory interactions, on a 12h light/dark cycle. Regular veterinary care and monitoring, balanced nutrition and environmental enrichment were provided by the Center for Comparative Medicine of Baylor College of Medicine. Surgical procedures on monkeys were conducted under general anesthesia following standard aseptic techniques.

To ameliorate pain after surgery, analgesics were given for 7 days. Animals were not sacrificed after the experiments.

Electrophysiological recordings

We performed non-chronic recordings using a 32-channel linear silicon probe (NeuroNexus V1x32-Edge-10mm-60-177). The surgical methods and recording protocol were described previously [58]. Briefly, form-specific titanium recording chambers and headposts were implanted under full anesthesia and aseptic conditions. The bone was originally left intact and only prior to recordings, small trephinations (2 mm) were made over medial primary visual cortex at eccentricities ranging from 1.4 to 3.0 degrees of visual angle. Recordings were done within two weeks of each trephination. Probes were lowered using a Narishige Microdrive (MO-97) and a guide tube to penetrate the dura. Care was taken to lower the probe slowly, not to penetrate the cortex with the guide tube and to minimize tissue compression (for a detailed description of the procedure, see [58]).

Data acquisition and spike sorting

Electrophysiological data were collected continuously as broadband signal (0.5Hz–16kHz) digitized at 24 bits as described previously [59]. Our spike sorting methods are based on [60], code available at https://github.com/aecker/moksm, but with adaptations to the novel type of silicon probe as described previously [58]. Briefly, we split the linear array of 32 channels into 14 groups of 6 adjacent channels (with a stride of two), which we treated as virtual electrodes for spike detection and sorting. Spikes were detected when channel signals crossed a threshold of five times the standard deviation of the noise. After spike alignment, we extracted the first three principal components of each channel, resulting in an 18-dimensional feature space used for spike sorting. We fitted a Kalman filter mixture model [61, 62] to track waveform drift typical for non-chronic recordings. The shape of each cluster was modeled with a multivariate *t*-distribution ($df = 5$) with a ridge-regularized covariance matrix. The number of clusters was determined based on a penalized average likelihood with a constant cost per additional cluster [60]. Subsequently, we used a custom graphical user interface to manually verify single-unit isolation by assessing the stability of the units (based on drifts and health of the cells throughout the session), identifying a refractory period, and inspecting the scatter plots of the pairs of channel principal components.

Visual stimulation and eye tracking

Visual stimuli were rendered by a dedicated graphics workstation and displayed on a 19" CRT monitor (40 × 30 cm) with a refresh rate of 100 Hz at a resolution of 1600 × 1200 pixels and a viewing distance of 100 cm (resulting in ∼70 px/deg). The monitors were gamma-corrected to have a linear luminance response profile. A camera-based, custom-built eye tracking system verified that monkeys maintained fixation within ∼ 0.42 degrees around the target. Offline analysis showed that monkeys typically fixated much more accurately. The monkeys were trained to fixate on a red target of ∼ 0.15 degrees in the middle of the screen. After they maintained fixation for 300 ms, a visual stimulus appeared. If the monkeys fixated throughout the entire stimulus period, they received a drop of juice at the end of the trial.

Receptive field mapping

At the beginning of each session, we first mapped receptive fields. We used a sparse random dot stimulus for receptive field mapping. A single dot of size 0.12 degrees of visual field was

presented on a uniform gray background, changing location and color (black or white) randomly every 30 ms. Each trial lasted for two seconds. We obtained multi-unit receptive field profiles for every channel using reverse correlation. We then estimated the population receptive field location by fitting a 2D Gaussian to the spike-triggered average across channels at the time lag that maximizes the signal-to-noise-ratio. We subsequently placed our natural image stimulus at this location.

Natural image stimulus

We used a set of 1450 grayscale images as well as four texturized versions of each image. We used grayscale images to avoid the complexity of dealing with color and focus on spatial image statistics. The texturized stimuli allowed us to vary the degree of naturalness, ranging from relatively simple, local statistics to very realistic textures capturing image statistics over spatial scales covering both classical and at least parts of the extra-classical receptive field of neurons. The images were taken from ImageNet [44], converted to grayscale and rescaled to 256 × 256 pixels. We generated textures with different degrees of naturalness by capturing different levels of higher-order correlations from a local to a global scale by using a parametric model for texture synthesis [45]. This texture model uses summary statistics of feature activations in different layers of the VGG-19 network [28] as parameters for the texture. The lowest-level model uses only the statistics of layer conv1_1. We refer to it as the "conv1" model. The next one uses statistics of conv1_1 and conv2_1 (referred to as conv2), and so on for conv3 and conv4. Due to the increasing level of nonlinearity of the VGG-19 features and their increasing receptive field sizes with depth, the textures synthesized from these models become increasingly more natural (see Fig 1 and [45] for more examples)

To synthesize the textures, we start with a random white noise image and iteratively refine pixels via gradient descent such that the resulting image matches the feature statistics of the original image [45]. For displaying and further analyses, we cropped the central 140 pixels of each image, which corresponds to 2 degrees of visual angle.

The entire data set contains 1450 × 5 = 7250 images (original plus synthesized). During each trial, 29 images were displayed, each for 60 ms, with no blanks in between (Fig 1B). We chose this fast succession of images to maximize the number of images we can get through in a single experiment, resulting in a large training set for model fitting. The short presentation times also mean that the responses we observe are mainly feedforward, since feedback processes take some time to be engaged. Each image was masked by a circular mask with a diameter of 2 degrees (140 px) and a soft fade-out starting at a diameter of 1 degree:

$$m(r) = \begin{cases} 1 & \text{if } 0 < r < 0.5 \\ 0.5 \cos\left(\pi(2r-1)\right) + 0.5 & \text{if } 0.5 \leq r < 1 \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Images were randomized such that consecutive images were not of the same type or synthesized from the same image. A full pass through the dataset took 250 successful trials, after which it was traversed again in a new random order. Images were repeated between one and four times, depending on how many trials the monkeys completed in each session.

Dataset and inclusion criteria

We recorded a total of 307 neurons in 23 recording sessions. We did not consider six of these sessions, for which we did not obtain enough trials to have at least two repetitions for each image. In the remaining 17 sessions, we quantified the fraction of total variance of each neuron

attributable to the stimulus by computing the ratio of explainable and total variance (grey bars in Fig 8):

$$\frac{\mathrm{Var}[y] - \sigma^2_{\mathrm{noise}}}{\mathrm{Var}[y]} \tag{3}$$

The explainable variance is the total variance minus the variance of the observation noise. We estimated the variance of the observation noise, $\sigma^2_{\mathrm{noise}}$, by averaging (across images) the variance (across repetitions) of responses to the same stimulus:

$$\sigma^2_{\mathrm{noise}} = \mathrm{E}_j[\mathrm{Var}_i[y_i|x_j]], \tag{4}$$

where x_j is the j^{th} image and y_i the response to the i^{th} repetition. We discarded neurons with a ratio of explainable-to-total variance (see Eq 3) smaller than 0.15, yielding 166 isolated neurons (monkey A: 51, monkey B: 115) recorded in 17 sessions with an average explainable variance of 0.285. Monkey A had only sessions with two repetitions while Monkey B had four repetitions per image.

Image preprocessing

All images were contrast-matched before displaying them on the screen. To do so, we rescaled the pixel intensities of all images such that the central, unmasked 1˚ (70 pixels) of each image had the same mean and standard deviation. We set the mean to 128 (same as the gray background) and the standard deviation to the average standard deviation across images. Any pixels falling outside the range of [0, 255] after this procedure were cropped to this range.

Prior to model fitting, we additionally cropped the central 80 pixels (1.1˚) of the 140-pixel (2˚) images shown to the monkey. For most of the analyses presented in this paper, we subsampled these crops to half their size (40 × 40) and z-scored them. For the input resolution control (Fig 5), we resampled with bicubic interpolation the original 80 × 80 crops to 60 × 60, 40 × 40, and 27 × 27 for scales 1.5, 1, 0.67, respectively.

GLM with pre-trained CNN features

Our proposed model consists of two parts: feature extraction and a generalized linear model (GLM; Fig 3). The features are the output maps $\Phi(x)$ of convolutional layers of VGG-19 [28] to a stimulus image x, followed by a batch normalization layer. We perform this normalization to ensure that the activations of each feature map have zero mean and unit variance (before ReLU), which is important because the readout weights are regularized by an L_1 penalty and having input features with different variances would implicitly apply different penalties on their corresponding readout weights.

We fit a separate GLM for each convolutional layer of VGG-19. The GLM consists of linear fully connected weights w_{ijk} for each neuron that compute a dot product with the input feature maps $\Phi_{ijk}(x)$, a static output nonlinearity f (also known as the inverse of the link function), and a Poisson noise model used for training. Here, i and j index space, while k indexes feature maps (denoted as depth in Fig 3). The spiking rate of a given neuron r will follow:

$$r(x) = f\left(\sum \Phi_{ijk}(x) w_{ijk} + b\right) \tag{5}$$

Additionally, three regularization terms were applied to the weights:

1. **Sparsity**: Most weights need to be zero since we expect the spatial pooling to be localized. We use the L_1 norm of the weights:

$$\mathcal{L}_{\text{sparse}} = \lambda_{\text{sparse}} \sum |w_{ijk}| \tag{6}$$

2. **Spatial smoothness**: Together with sparseness, spatial smoothness encourages spatial locality by imposing continual regular changes in space. We computed this by an L_2 penalty on the Laplacian of the weights:

$$\mathcal{L}_{\text{Laplace}} = \lambda_{\text{Laplace}} \sqrt{\sum_{ijk} (w_{:,:,k} * L)_{ij}^2}, \quad L = \begin{bmatrix} 0 & -1 & 0 \\ -1 & 4 & -1 \\ 0 & -1 & 0 \end{bmatrix} \tag{7}$$

3. **Group sparsity** encourages our model to pool from a small set of feature maps to explain each neuron's responses:

$$\mathcal{L}_{\text{group}} = \lambda_{\text{group}} \sum_k \sqrt{\sum_{i,j} w_{ijk}^2} \tag{8}$$

Considering the recorded image-response pair (x, y) for one neuron, the resulting loss function is given by:

$$\mathcal{L} = -\sum y \log r(x) + r(x) + \mathcal{L}_{\text{sparse}} + \mathcal{L}_{\text{Laplace}} + \mathcal{L}_{\text{group}} \tag{9}$$

where the sum runs over samples (image, response pairs).

We fit the model by minimizing the loss using the Adam optimizer [63] on a training set consisting of 80% of the data, and reported performance on the remaining 20%. We cross-validated the hyperparameters λ_{sparse}, λ_{Laplace}, λ_{group} for each neuron independently by performing a grid search over four logarithmically spaced values for each hyperparameter. The validation was done on 20% of the training data. The optimal hyperparameter values obtained on the validation set where $\lambda_{\text{Laplace}} = 0.1$, $\lambda_{\text{sparse}} = 0.01$, $\lambda_{\text{group}} = 0.001$. When fitting models, we used the same split of data for training, validation, and testing across all models.

Data-driven convolutional neural network model

We followed the results of [40] and use their best-performing architecture that obtained state-of-the-art performance on a public dataset [38]. Like our VGG-based model, this model also consisted of convolutional feature extraction followed by a GLM, the difference being that here the convolutional feature space was learned from neural data instead of having been trained on object recognition. The feature extraction architecture consisted of convolutional layers with filters of size 13 × 13 px for the first layer and 3 × 3 px for the subsequent layers. Each layer had 32 feature maps (Fig 7A) and exponential linear units (ELU [64])

$$ELU(x) = \begin{cases} x & \text{if } x > 0 \\ exp(x) - 1 & \text{if } x \leq 0 \end{cases} \tag{10}$$

as nonlinearities with batch normalization [65] to facilitate training in between the layers. As in the original publication [40], we regularized the convolutional filters by imposing smoothness constraints on the first layer and group sparseness on the subsequent layers. A notable difference to our VGG-based GLM is that here the readout weights are factorized in space and

feature maps:

$$w_{ijk} = u_{ij} v_k, \tag{11}$$

where u_{ij} is a spatial mask and v_k a set of feature pooling weights. We fitted models with increasing number of convolutional layers (one to five). We found that optimizing the final nonlinearity, $f(x)$, of each neuron was important for optimal performance of the data-driven CNN. To do so, we took the following approach: we split $f(x)$ into two components:

$$f(x) = h(x)g(x) \tag{12}$$

where $g(x)$ is ELU shifted to the right and up by one unit (to make it non-negative—firing rates are non-negative):

$$g(x) = ELU(x - 1) + 1 \tag{13}$$

and h is a non-negative, piecewise linear function:

$$h(x) = exp\left(\sum_{i=1}^{n} \alpha_i t_i\right) \tag{14}$$

Here, α_i are parameters learned jointly with the remaining weights of the network and the t_i are a set of 'tent' basis functions to create a piecewise linear function with interpolation points $x_i = -3, -2.82, \ldots, 6$ (i.e. $\Delta x = 0.18$):

$$t_i = \min\left(\max\left(0, \frac{x - x_{i-1}}{\Delta x}\right), \max\left(\frac{x_{i+1} - x}{\Delta x}\right)\right) \tag{15}$$

We regularize the output nonlinearity by penalizing the L_2 norm of the first and second discrete finite differences of α_i to encourage h to be close to 1 and smooth:

$$\mathcal{L}_{out} = \lambda_{out}\left(\sum_{i=2}^{n} + (\alpha_i - \alpha_{i-1})^2 + \sum_{i=2}^{n-1}(2\alpha_i - \alpha_{i-1} - \alpha_{i+1})^2\right) \tag{16}$$

Note that we applied this optimization of the output nonlinearity only to the data-driven model, as doing the same for the VGG-based model did not improve performance. One potential reason for this difference is that the VGG-based model has a much larger number of feature maps (256 for layer conv3_1) that each neuron can pool from.

Linear nonlinear poisson model (LNP)

We implemented a simple regularized LNP Model [46]. This model is fitted for each neuron separately and consists of two simple stages: The first one is a linear filter **w** with the same dimensions as the input images. The second is a pointwise exponential function as nonlinearity that converts the filter output into a non-negative spike rate. The LNP assumes spike count generation through a Poisson process, so we minimize a Poisson loss (negative log-likelihood) to obtain the kernels of each neuron (see first term of Eq 17 below). Additionally, we imposed two regularization constraints that we cross-validated: smoothness (Eq 7) and sparsity (Eq 6). With the same M image-response pairs ($\mathbf{x}$, y) of the training set that we used for all other

models, we optimized the following loss function:

$$\mathcal{L}_{LNP} = \sum_{i=1}^{M} [\mathbf{w}^T \mathbf{x}_i - y_i \log (\mathbf{w}^T \mathbf{x}_i)] + \mathcal{L}_{sparse}(\mathbf{w}) + \mathcal{L}_{laplace}(\mathbf{w}) \tag{17}$$

Gabor filter bank model (GFB)

Varying versions of the Gabor filter bank model (GFB) have been used in classical work on system identification [22, 47, 48, 66]. This model convolves the image with quadrature pairs of Gabor filters with varying scales, frequencies, aspect ratios, and orientations. Each quadrature pair consists of an 'even' (cosine/symmetric) and an 'odd' (sine/antisymmetric) Gabor filter and produces three feature maps: the results of the convolution with the two filters ('even' and 'odd' features) and an 'energy' feature, which is the spectral power of each pair (i.e. sum of the squares). Thus, this model allows for modeling simple and complex cells and linear combinations thereof.

The Gabor filters obeyed the following equations with x and y representing spatial dimensions:

$$g_{\sigma,f,\gamma,\theta,\varphi}(x', y') = \exp\left[-\frac{x'^2 + \gamma^2 y'^2}{2\sigma^2}\right] \cos(2\pi f x' + \varphi), \tag{18}$$

where

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \tag{19}$$

The standard deviation (σ) represents the scale of the Gaussian envelope, the aspect ratio (γ) specifies the ellipticity of the envelope, the spatial frequency (f) quantifies the number of sinusoidal cycles divided by the width of the Gaussian aperture ($\sim 4\sigma$), The sinusoidal grating is determined by an orientation (θ) and phase (φ). To form quadrature pairs we set φ to 0 and $\pi/2$ for even and odd filters, respectively. We set the kernel size of every Gabor filter to the minimum of the input image size and the closest integer to $4\sigma/\gamma$ for both spatial dimensions.

To compute the even and odd feature maps, we convolved the input image with each Gabor filter using strided convolutions ($\otimes$) with stride s:

$$E_{\sigma,f,\gamma,\theta} = I \otimes g_{\sigma,f,\gamma,\theta,\varphi=0} \tag{20}$$

$$O_{\sigma,f,\gamma,\theta} = I \otimes g_{\sigma,f,\gamma,\theta,\varphi=\pi/2} \tag{21}$$

We then computed the energy features as follows:

$$A_{\sigma,f,\gamma,\theta} = \sqrt{(E_{\sigma,f,\gamma,\theta})^2 + (O_{\sigma,f,\gamma,\theta})^2} \tag{22}$$

The full feature space of this model $\Phi_{Gabor}(x)$ is a concatenation of triplets of even, odd, and energy features of every Gabor filter. The filter bank consisted of N_s filter sizes, N_f spatial frequencies per size, N_θ orientations and N_γ aspect ratios. The spatial frequencies depended on the size of the envelope: $f_n = n/4\sigma$ with $n = 1, \ldots, N_f$, which means for $n = 3$ we used Gabors with 1, 2 and 3 cycles. Aspect ratios ranged from 0.5 to 1 with equal spacing, except for $N_\gamma = 1$ where we used an aspect ratio of $\gamma = 1$. As for the GLM with VGG features, we fit a linear readout on top of this feature space (Eq 5), followed by a shifted ELU output nonlinearity (Eq 13).

To train the model, we minimized a Poisson loss and regularized the readout weights to be sparse (i.e. first two terms in Eq 9); we did not enforce smoothness or group sparsity here, as they mainly improve interpretability but do not affect performance. We determined the hyperparameters of the Gabor filter bank by running a search over a number of parameter combinations and evaluating the performance of each model on the validation set. We converged to the following values: three different sizes ($N_s = 3$: 6 px/0.17˚, 11 px/0.31˚ and 21 px/0.60˚), three different spatial frequencies ($N_f = 3$: $f = 1, 2, 3$) per size, one aspect ratio ($N_\gamma = \gamma = 1$), eight orientations ($N_\theta = 8$), convolutional stride of 6 for all filters, and L_1 regularization parameter $\alpha = 0.05$. Notably, including multiple different aspect ratios did not improve performance, presumably because it increased the dimensionality of the feature space which harmed generalization.

Number of parameters to be learned

The parameters we fit for each of the models belong either to a shared set for all neurons (the core), or are specific to each neuron (the readout). Table 2 shows the number of parameters for each of the models and how many belong to either core or readout. For both the LNP and GFB models, we learn only a readout from a fixed feature space for each neuron plus a bias. For the LNP we learn one channel of pixel intensities ($40 \times 40 + 1$). For the GFB model, we have for each size $N_f N_\theta$ channels ($3 \times 8 = 24$), and each filter produced a feature output of size $\lfloor 1 + (40 - size)/stride \rfloor$. With $N_s = 3$ we got sizes 6, 11, and 21 so the output features have size 6, 5, and 4, respectively. Since each Gabor filter quadrature pair produces odd, even, and energy feature spaces, the total dimensionality is $3 \times 24 \times (6^2 + 5^2 + 4^2) + 1 = 5545$. For the three-layer CNN, we have 32 channels in all layers (32×3 biases) and filters with sizes $13 \times 13 \times 32$, $3 \times 3 \times 32 \times 32$, and $3 \times 3 \times 32 \times 32$, resulting in 23, 963 core parameters. The output feature space for an image is $28 \times 28 \times 32$ (reduced from 40×40 due to the padding of the convolutions: no padding in first layer, zero padding in second and third). With a factorized readout and a bias, the readout per neuron is then $28 \times 28 + 32$ plus a bias. In addition, our point-wise output nonlinearity has 50 parameters. Thus, overall we have 867 readout parameters per neuron for this CNN model.

For the VGG-based model, although we do not learn the feature space, we do learn batch normalization parameters at the output of the last convolutional layer. For the model that used conv3_1 (256 feature channels) this means learning scale and bias parameters common to all neurons: $2 \times 256 = 512$ for the core. For a 40×40 input, the output of the feature space is $10 \times 10 \times 256$ (due to downsampling twice via max pooling). Here, we learn a dense readout and a bias, so the readout per neuron has $10 \times 10 \times 256 + 1 = 25, 601$ parameters.

Performance evaluation

We measured the performance of all models with the fraction of explainable variance explained (*FEV*). That is, the ratio between the variance accounted for by the model (variance explained) and the explainable variance (numerator in Eq 3). The explainable variance is lower than the total variance, because observation noise prevents even a perfect model from accounting for all variance. We compute *FEV* as

$$FEV = 1 - \frac{\frac{1}{N}\sum (y - \hat{y})^2 - \sigma^2_{\text{noise}}}{Var[y] - \sigma^2_{\text{noise}}}, \tag{23}$$

where $\hat{y}$ represents the model predictions, y the observed spike counts, and the level of observation noise, σ^2_{noise} is defined in Eq 4 above.

Implementation details

We implemented all models in TensorFlow [67]. We optimized them with Adam [63] using mini-batches of size 256, and early stopping: we evaluated performance on the validation set every 100 training steps, and after ten iterations of no improvement, we decayed the learning rate by a factor of three and repeated this three times. The learning rate at the beginning of the optimization was cross-validated for the goal-driven models and set to 1e-4 for the others as this value always worked best.

Tools. We managed our data and kept track of models, parameters, and performance using DataJoint [68]. In addition, we used Numpy/Scipy [69], Matplotlib [70], Seaborn [71], Jupyter [72], Tensorflow [67], and Docker [73]. The code to fit all models is available in this repository: \url{https://github.com/sacadena/Cadena2019PlosCB}.

Acknowledgments

We thank Philipp Berens and James Cotton for valuable discussions and help during the early stages of this project. We thank Tori Shinn for help with animal training and neural recordings.

Author Contributions

Conceptualization: Santiago A. Cadena, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Data curation: Santiago A. Cadena.

Formal analysis: Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Andreas S. Tolias, Alexander S. Ecker.

Funding acquisition: Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Investigation: Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Methodology: Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Project administration: Santiago A. Cadena, Andreas S. Tolias, Alexander S. Ecker.

Resources: Edgar Y. Walker, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Software: Santiago A. Cadena, Edgar Y. Walker, Leon A. Gatys, Alexander S. Ecker.

Supervision: Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Validation: Santiago A. Cadena, Matthias Bethge, Alexander S. Ecker.

Visualization: Santiago A. Cadena.

Writing – original draft: Santiago A. Cadena, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Writing – review & editing: Santiago A. Cadena, George H. Denfield, Leon A. Gatys, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

References

1. Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, et al. Do we know what the early visual system does? The Journal of neuroscience. 2005; 25(46):10577–10597. https://doi.org/10.1523/JNEUROSCI.3726-05.2005 PMID: 16291931

2. Hubel DH, Wiesel TN. Receptive fields of single neurones in the cat's striate cortex. The Journal of physiology. 1959; 148(3):574–591. https://doi.org/10.1113/jphysiol.1959.sp006308 PMID: 14403679
3. Hubel DH, Wiesel TN. Receptive fields and functional architecture of monkey striate cortex. The Journal of physiology. 1968; 195(1):215–243. https://doi.org/10.1113/jphysiol.1968.sp008455 PMID: 4966457
4. Jones JP, Palmer LA. An evaluation of the two-dimensional Gabor filter model of simple receptive fields in cat striate cortex. Journal of neurophysiology. 1987; 58(6):1233–1258. https://doi.org/10.1152/jn.1987.58.6.1233 PMID: 3437332
5. Heeger DJ. Half-squaring in responses of cat striate cells. Visual neuroscience. 1992; 9(05):427–443. https://doi.org/10.1017/S095252380001124X PMID: 1450099
6. Adelson EH, Bergen JR. Spatiotemporal energy models for the perception of motion. JOSA A. 1985; 2 (2):284–299. https://doi.org/10.1364/JOSAA.2.000284
7. Tang S, Lee TS, Li M, Zhang Y, Xu Y, Liu F, et al. Complex Pattern Selectivity in Macaque Primary Visual Cortex Revealed by Large-Scale Two-Photon Imaging. Current Biology. 2018; 28(1):38–48. https://doi.org/10.1016/j.cub.2017.11.039 PMID: 29249660
8. Olshausen BA, Field DJ. How close are we to understanding V1? Neural computation. 2005; 17 (8):1665–1699. https://doi.org/10.1162/0899766054026639 PMID: 15969914
9. Talebi V, Baker CL. Natural versus synthetic stimuli for estimating receptive field models: a comparison of predictive robustness. The Journal of Neuroscience. 2012; 32(5):1560–1576. https://doi.org/10.1523/JNEUROSCI.4661-12.2012 PMID: 22302799
10. Eichhorn J, Sinz F, Bethge M. Natural image coding in V1: how much use is orientation selectivity? PLoS computational biology. 2009; 5(4):e1000336. https://doi.org/10.1371/journal.pcbi.1000336 PMID: 19343216
11. Field DJ. What Is the Goal of Sensory Coding? Neural Computation. 1994; 6(4):559–601. https://doi.org/10.1162/neco.1994.6.4.559
12. Olshausen BA, et al. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. Nature. 1996; 381(6583):607–609. https://doi.org/10.1038/381607a0 PMID: 8637596
13. Heeger DJ. Normalization of Cell Responses in Cat Striate Cortex. Visual neuroscience. 1992; 9 (2):181–197. https://doi.org/10.1017/S0952523800009640 PMID: 1504027
14. Bethge M, Simoncelli EP, Sinz FH. Hierarchical Modeling of Local Image Features through *L_p*-Nested Symmetric Distributions. In: Advances in neural information processing systems; 2009. p. 1696–1704.
15. Cavanaugh JR, Bair W, Movshon JA. Nature and interaction of signals from the receptive field center and surround in macaque V1 neurons. Journal of neurophysiology. 2002; 88(5):2530–2546. https://doi.org/10.1152/jn.00692.2001 PMID: 12424292
16. Cavanaugh JR, Bair W, Movshon JA. Selectivity and spatial distribution of signals from the receptive field surround in macaque V1 neurons. Journal of neurophysiology. 2002; 88(5):2547–2556. https://doi.org/10.1152/jn.00693.2001 PMID: 12424293
17. Bair W, Cavanaugh JR, Movshon JA. Time course and time-distance relationships for surround suppression in macaque V1 neurons. Journal of Neuroscience. 2003; 23(20):7690–7701. https://doi.org/10.1523/JNEUROSCI.23-20-07690.2003 PMID: 12930809
18. Carandini M, Heeger DJ. Normalization as a canonical neural computation. Nature Reviews Neuroscience. 2012; 13(1):51. https://doi.org/10.1038/nrn3136
19. Rust NC, Schwartz O, Movshon JA, Simoncelli EP. Spatiotemporal elements of macaque v1 receptive fields. Neuron. 2005; 46(6):945–956. PMID: 15953422
20. Touryan J, Felsen G, Dan Y. Spatial structure of complex cell receptive fields measured with natural images. Neuron. 2005; 45(5):781–791. https://doi.org/10.1016/j.neuron.2005.01.029 PMID: 15748852
21. Vintch B, Movshon JA, Simoncelli EP. A convolutional subunit model for neuronal responses in macaque V1. The Journal of Neuroscience. 2015; 35(44):14829–14841. https://doi.org/10.1523/JNEUROSCI.2815-13.2015 PMID: 26538653
22. Willmore B, Prenger RJ, Wu MCK, Gallant JL. The berkeley wavelet transform: a biologically inspired orthogonal wavelet transform. Neural computation. 2008; 20(6):1537–1564. https://doi.org/10.1162/neco.2007.05-07-513 PMID: 18194102
23. Zipser D, Andersen RA. A back-propagation programmed network that simulates response properties of a subset of posterior parietal neurons. Nature. 1988; 331(6158):679. https://doi.org/10.1038/331679a0 PMID: 3344044
24. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience. 2016; 19(3):356–365. https://doi.org/10.1038/nn.4244 PMID: 26906502

25. Donahue J, Jia Y, Vinyals O, Hoffman J, Zhang N, Tzeng E, et al. Decaf: A deep convolutional activation feature for generic visual recognition. In: International conference on machine learning; 2014. p. 647–655.

26. Oquab M, Bottou L, Laptev I, Sivic J. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2014. p. 1717–1724.

27. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.

28. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations; 2015.

29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.

30. Huang G, Liu Z, Weinberger KQ, van der Maaten L. Densely connected convolutional networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 4700–4708.

31. Kümmerer M, Theis L, Bethge M. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. In: ICLR Workshop; 2015.

32. Cadieu CF, Hong H, Yamins DL, Pinto N, Ardila D, Solomon EA, et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol. 2014; 10(12): e1003963. https://doi.org/10.1371/journal.pcbi.1003963 PMID: 25521294

33. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the National Academy of Sciences. 2014; 111(23):8619–8624. https://doi.org/10.1073/pnas.1403112111

34. Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Comput Biol. 2014; 10(11):e1003915. https://doi.org/10.1371/journal.pcbi.1003915 PMID: 25375136

35. Güçlü U, van Gerven MA. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. The Journal of Neuroscience. 2015; 35(27):10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015 PMID: 26157000

36. Seibert D, Yamins DL, Ardila D, Hong H, DiCarlo JJ, Gardner JL. A performance-optimized model of neural responses across the ventral visual stream. bioRxiv. 2016; p. 036475.

37. Prenger R, Wu MCK, David SV, Gallant JL. Nonlinear V1 responses to natural scenes revealed by neural network analysis. Neural Networks. 2004; 17(5):663–679. https://doi.org/10.1016/j.neunet.2004.03.008 PMID: 15288891

38. Antolík J, Hofer SB, Bednar JA, Mrsic-Flogel TD. Model Constrained by Visual Hierarchy Improves Prediction of Neural Responses to Natural Scenes. PLOS Comput Biol. 2016; 12(6):e1004927. https://doi.org/10.1371/journal.pcbi.1004927 PMID: 27348548

39. Batty E, Merel J, Brackbill N, Heitman A, Sher A, Litke A, et al. Multilayer Recurrent Network Models of Primate Retinal Ganglion Cell Responses. 2016.

40. Klindt D, Ecker AS, Euler T, Bethge M. Neural system identification for large populations separating "what" and "where". In Advances in Neural Information Processing Systems; 2017. p. 3506–3516

41. McIntosh L, Maheswaranathan N, Nayebi A, Ganguli S, Baccus S. Deep learning models of the retinal response to natural scenes. In: Advances in Neural Information Processing Systems; 2016. p. 1369–1377.

42. Kindel WF, Christensen ED, Zylberberg J. Using deep learning to reveal the neural code for images in primary visual cortex. arXiv preprint arXiv:170606208. 2017.

43. Zhang Y, Lee TS, Li M, Liu F, Tang S. Convolutional Neural Network Models of V1 Responses to Complex Patterns. Journal of computational neuroscience. 2018; p. 1–22.

44. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International Journal of Computer Vision. 2015; 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y

45. Gatys L, Ecker AS, Bethge M. Texture synthesis using convolutional neural networks. In: Advances in Neural Information Processing Systems; 2015. p. 262–270.

46. Simoncelli EP, Paninski L, Pillow J, Schwartz O. Characterization of neural responses with stochastic stimuli. The cognitive neurosciences. 2004; 3:327–338.

47. Daugman JG. Two-dimensional spectral analysis of cortical receptive field profiles. Vision research. 1980; 20(10):847–856. https://doi.org/10.1016/0042-6989(80)90065-6 PMID: 7467139

48. WATSON A, et al. The cortex transform- Rapid computation of simulated neural images. Computer vision, graphics, and image processing. 1987; 39(3):311–327. https://doi.org/10.1016/S0734-189X(87)80184-6

49. Movshon JA, Thompson I, Tolhurst D. Receptive field organization of complex cells in the cat's striate cortex. The Journal of physiology. 1978; 283:79. https://doi.org/10.1113/jphysiol.1978.sp012488 PMID: 722592

50. Gallant J, David S. The Neural Prediction Challenge;. http://neuralprediction.berkeley.edu/, last accessed on 10/02/2018.

51. Ringach DL, Hawken MJ, Shapley R. Dynamics of Orientation Tuning in Macaque Primary Visual Cortex. Nature. 1997; 387(6630):281–284. https://doi.org/10.1038/387281a0 PMID: 9153392

52. Ringach DL, Shapley RM, Hawken MJ. Orientation Selectivity in Macaque V1: Diversity and Laminar Dependence. The Journal of Neuroscience. 2002; 22(13):5639–5651. PMID: 12097515

53. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific reports. 2016; 6. https://doi.org/10.1038/srep27755

54. Kriegeskorte N. Deep neural networks: a new framework for modeling biological vision and brain information processing. Annual Review of Vision Science. 2015; 1:417–446. https://doi.org/10.1146/annurev-vision-082114-035447 PMID: 28532370

55. Heeger D. Computational model of cat striate physiology. Computational models of visual perception. 1991; p. 119–133.

56. Spillmann L, Dresp-Langley B, Tseng Ch. Beyond the Classical Receptive Field: The Effect of Contextual Stimuli. Journal of Vision. 2015; 15(9):7–7. https://doi.org/10.1167/15.9.7 PMID: 26200888

57. Zeiler MD, Krishnan D, Taylor GW, Fergus R. Deconvolutional Networks. In: Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference On. IEEE; 2010. p. 2528–2535.

58. Denfield GH, Ecker AS, Shinn TJ, Bethge M, Tolias AS. Attentional fluctuations induce shared variability in macaque primary visual cortex. Nature communications; 2018, vol. 9, no 1, p. 2654.

59. Ecker AS, Berens P, Keliris GA, Bethge M, Logothetis NK, Tolias AS. Decorrelated neuronal firing in cortical microcircuits. science. 2010; 327(5965):584–587. https://doi.org/10.1126/science.1179867 PMID: 20110506

60. Ecker AS, Berens P, Cotton RJ, Subramaniyan M, Denfield GH, Cadwell CR, et al. State dependence of noise correlations in macaque primary visual cortex. Neuron. 2014; 82(1):235–248. https://doi.org/10.1016/j.neuron.2014.02.006 PMID: 24698278

61. Calabrese A, Paninski L. Kalman filter mixture model for spike sorting of non-stationary data. Journal of neuroscience methods. 2011; 196(1):159–169. https://doi.org/10.1016/j.jneumeth.2010.12.002 PMID: 21182868

62. Shan KQ, Lubenov EV, Siapas AG. Model-based spike sorting with a mixture of drifting t-distributions. Journal of neuroscience methods, 2017, vol. 288, p. 82–98.

63. Kingma D, Ba J. Adam: A method for stochastic optimization. In: International Conference on Learning Representations; 2015.

64. Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:151107289. 2015.

65. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning; 2015. p. 448–456.

66. Willmore BD, Prenger RJ, Gallant JL. Neural representation of natural images in visual area V2. The Journal of neuroscience. 2010; 30(6):2102–2114. https://doi.org/10.1523/JNEUROSCI.4099-09.2010 PMID: 20147538

67. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, et al. TensorFlow: A System for Large-Scale Machine Learning. In: OSDI. vol. 16; 2016. p. 265–283.

68. Yatsenko D, Reimer J, Ecker AS, Walker EY, Sinz F, Berens P, et al. DataJoint: managing big scientific data using MATLAB or Python; 2015. Available from: http://biorxiv.org/lookup/doi/10.1101/031658.

69. Walt Svd, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering. 2011; 13(2):22–30. https://doi.org/10.1109/MCSE.2011.37

70. Hunter JD. Matplotlib: A 2D graphics environment. Computing in science & engineering. 2007; 9(3):90–95. https://doi.org/10.1109/MCSE.2007.55

71. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. mwaskom/seaborn: v0.8.1 (September 2017); 2017. Available from: https://doi.org/10.5281/zenodo.883859.

72. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press; 2016. p. 87—90.

73. Merkel D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. Linux J. 2014; 2014(239).

A.2 How well do deep neural networks trained on object recognition characterize the mouse visual system?

> **S. A. Cadena**, Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., Reimer, J., Bethge, M., Tolias, A., & Ecker, A. S. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? *Advances in Neural Information Processing (NeurIPS) Neuro-AI Workshop*

Abstract

Recent work on modeling neural responses in the primate visual system has benefited from deep neural networks trained on large-scale object recognition, and found a hierarchical correspondence between layers of the artificial neural network and brain areas along the ventral visual stream. However, we neither know whether such task-optimized networks enable equally good models of the rodent visual system, nor if a similar hierarchical correspondence exists. Here, we address these questions in the mouse visual system by extracting features at several layers of a convolutional neural network (CNN) trained on ImageNet to predict the responses of thousands of neurons in four visual areas (V1, LM, AL, RL) to natural images. We found that the CNN features outperform classical subunit energy models, but found no evidence for an order of the areas we recorded via a correspondence to the hierarchy of CNN layers. Moreover, the same CNN but with random weights provided an equivalently useful feature space for predicting neural responses. Our results suggest that object recognition as a high-level task does not provide more discriminative features to characterize the mouse visual system than a random network. Unlike in the primate, training on ethologically relevant visually guided behaviors – beyond static object recognition – may be needed to unveil the functional organization of the mouse visual cortex.

Author contributions

Conceptualization: **SC**, FS, AE, AT. Data Curation: **SC**, TM, EF, EC, EW, JR. Formal Analysis: **SC**. Funding Acquisition: AT, AE, MB. Investigation: **SC**, FS, AE. Methodology: **SC**, FS, AT, AE. Project Administration: **SC**, AE. Resources: **SC**, FS, EW, EC, AT, AE. Software: **SC**, FS, EW, EC. Supervision: AE, AT, MB. Validation: **SC**. Visualization: **SC**. Writing – Original Draft Preparation: **SC**, AE. Writing – Review and Editing: **SC**, EF, FS, AT, MB, AE.

How well do deep neural networks trained on object recognition characterize the mouse visual system?

Santiago A. Cadena,[1,2,5,*] Fabian H. Sinz,[5,6] Taliah Muhammad,[3]
Emmanouil Froudarakis, [3,4] Erick Cobos, [3,4] Edgar Y. Walker, [5,6] Jake Reimer, [3,4]
Matthias Bethge, [1,2,5,†] Andreas S. Tolias, [3,4,†] Alexander S. Ecker [1,2,5,†,‡]

[1] Centre for Integrative Neuroscience, University of Tübingen, Germany
[2] Institute for Theoretical Physics, University of Tübingen, Germany
[3] Department of Neuroscience, Baylor College of Medicine, Houston, TX, USA
[4] Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, USA
[5] Bernstein Center for Computational Neuroscience, University of Tübingen, Germany
[6] Institute Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Germany
† *Authors contributed equally*
‡ *Present address: Department of Computer Science, University of Göttingen, Germany*

* santiago.cadena@uni-tuebingen.de

Abstract

Recent work on modeling neural responses in the primate visual system has benefited from deep neural networks trained on large-scale object recognition, and found a hierarchical correspondence between layers of the artificial neural network and brain areas along the ventral visual stream. However, we neither know whether such task-optimized networks enable equally good models of the *rodent* visual system, nor if a similar hierarchical correspondence exists. Here, we address these questions in the mouse visual system by extracting features at several layers of a convolutional neural network (CNN) trained on ImageNet to predict the responses of thousands of neurons in four visual areas (V1, LM, AL, RL) to natural images. We found that the CNN features outperform classical subunit energy models, but found no evidence for an order of the areas we recorded via a correspondence to the hierarchy of CNN layers. Moreover, the same CNN but with random weights provided an equivalently useful feature space for predicting neural responses. Our results suggest that object recognition as a high-level task does not provide more discriminative features to characterize the mouse visual system than a random network. Unlike in the primate, training on ethologically relevant visually guided behaviors – beyond static object recognition – may be needed to unveil the functional organization of the mouse visual cortex.

1 Introduction

Visual object recognition is a fundamental and difficult task performed by the primate brain via a hierarchy of visual areas (the ventral stream) that progressively untangles object identity information, gaining invariance to a wide range of object-preserving visual transformations [1, 2]. Fueled by the advances of deep learning, recent work on modeling neural responses in sensory brain areas builds upon hierarchical convolutional neural networks (CNNs) trained to solve complex tasks like object recognition [3]. Interestingly, these models have not only achieved unprecedented performance

33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. Neuro↔AI Workshop

in predicting neural responses in several brain areas of macaques and humans [4–7], but they also revealed a hierarchical correspondence between the layers of the CNNs and areas of the ventral stream [4, 6]: the higher the area in the ventral stream, the higher the CNN layer that explained it best. The same approach also provided a quantitative signature of a previously unclear hierarchical organization of A1 and A2 in the human auditory cortex [7].

These discoveries about the primate have sparked a still unresolved question: to what extent is visual object processing also hierarchically organized in the mouse visual cortex and how well can the mouse visual system be modeled using goal-driven deep neural networks trained on static object classification? This question is important since mice are increasingly used to study vision due to the plethora of available experimental techniques such as the ability to genetically identify and manipulate neural circuits that are not easily available in primates. Recent work suggests that rats are capable of complex visual discrimination tasks [8] and recordings from extrastriate areas show a gradual increase in the ability of neurons in higher visual areas to support discrimination of visual objects [9, 10].

Here, we set out to study how well the mouse visual system can be characterized by goal-driven deep neural networks. We extracted features from the hidden layers of a standard CNN (VGG16, [11]) trained on object categorization, to predict responses of thousands of neurons in four mouse visual areas (V1, LM, AL, RL) to static natural images. We found that VGG16 yields powerful features for predicting neural activity, outperforming a Gabor filter bank energy model in these four visual areas. However, VGG16 does not significantly outperform a feature space produced by a network with an identical architecture but random weights. In contrast to previous work in primates, our data provide no evidence so far for a hierarchical correspondence between the deep network layers and the visual areas we recorded.

2 Model Architecture

Our network (Fig.1) builds upon earlier work [5, 12]. It consist of four main network components: a *core* that provides nonlinear features of input images, a *readout* that maps those features to each neuron's responses, a *shifter* that predicts receptive field shifts from pupil position, and a *modulator* that provides a gain factor for each neuron based on running speed and pupil dilation of the mouse.

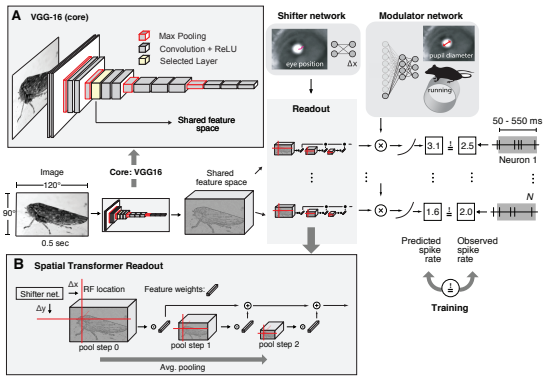


Figure 1: VGG16 based model. Input images are forwarded trough the *core* (**A**) network (first n layers of VGG16) to produce a feature space shared by all neurons. Then, the spatial transformer *readout* (**B**) finds a mapping between these features and the neural responses for each neuron separately. The *shifter* network (an MLP with one hidden layer) corrects for eye movements. The output of the readout is multiplied by a gain predicted by the *modulator* network (an MLP with one hidden layer) that uses running speed and pupil dilation. A static nonlinearity converts the result into the predicted spike rate. All components of the model are trained jointly end-to-end to minimize the difference between predicted and observed neural responses.

For the core we use VGG16 [11] up to one of the first eight convolutional layers. We chose VGG16 due to its simple feed-forward architecture, competitive object classification performance, and increasing popularity to characterize rodent visual areas [10, 13]. The collection of output feature maps of a VGG16 layer – the shared feature space – was then fed into a spatial transformer readout for each neuron (Fig.1B, see [12] for details). This readout learns one (x, y) location for each neuron (its receptive field location, RF) [12] and extracts a feature vector at this location from multiple downsampled versions (scales) of the feature maps. The output of the readout is a linear combination of the concatenated feature vectors. We regularized the feature weights with an L_1 penalty to encourage sparsity.

Shifter and modulator are multi-layer perceptrons (MLP) with one hidden layer. The shifter takes the tracked pupil position in camera coordinates and predicts a global receptive field shift $(\Delta x, \Delta y)$

in monitor coordinates. The modulator uses the mouse's running speed, its pupil diameter, and the derivative to predict a gain for each neuron by which the neuron's predicted response is multiplied. A soft-thresholding nonlinearity turns the result into a non-negative spike rate prediction (Fig.1). All components of the model (excluding the core, which is pre-trained on ImageNet) are trained jointly end-to-end to minimize the difference between predicted and observed neural responses using Adam with a learning rate of 10^{-4}, a batch size of 125 and early stopping.

3 Experiments

Neural data. We recorded responses of excitatory neurons in areas V1, LM, AL, and RL (layer 2/3) from two scans from one mouse and a third scan from a second mouse with a large-field-of-view two-photon mesoscope (see [14] for details) at a frame rate of 6.7 Hz. We selected cells based on a classifier for somata on the segmented cell masks and deconvolved their fluorescence traces, yielding 7393, 4674, 4680, 5797 neurons from areas V1, LM, AL, and RL, respectively. We further monitored pupil position, pupil dilation, and absolute running speed of the animal.

Visual stimuli. Stimuli consisted of 5100 images taken from ImageNet, cropped to 16:9 and converted to gray-scale. The screen was 55×31 cm at a distance of 15 cm, covering roughly $120° \times 90°$. In each scan, we showed 5000 of these images once (training and validation set) and the remaining 100 images 10 times each (test set). Each image was presented for 500 ms followed by a blank screen lasting between 300 ms and 500 ms. For each neuron, we extract the accumulated activity between 50 ms and 550 ms after stimulus onset using a Hamming window.

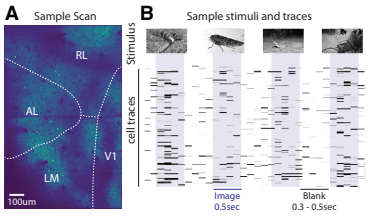


Figure 2: Neural data. **A**. Example large-field-of-view scan. **B**. Visual paradigm with sample cell traces

4 Results

We fitted one model (that of Fig.1; see [12] for training details) for each combination of scan, brain area, VGG16 layer (out of the first eight), random initialization (out of three seeds), and input resolution. We considered several resolutions of the input images because the right scale at which VGG16 layers extract relevant features that best match the representation in the brain is unknown. Optimizing the scale for each layer was critical, since the correspondence between a single layer and a brain area (in terms of best correlation performance) strongly depends on the input resolution (e.g. see Fig 3A for V1 data). For further analyses (Fig 3B & 4), we pick for each case the best performing input scale in the validation set.

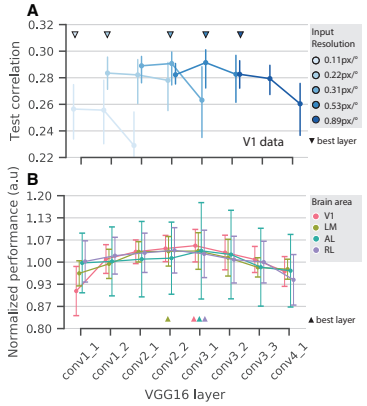


Figure 3: No clear hierarchical correspondence. **A** Test correlation in V1 data for different input resolutions as a function of VGG16 layer. **B** Normalized performance for all four brain areas. Triangles show the best predictive layer in each case

No hierarchical correspondence. Previous results in primates [4] show that a brain area higher in the hierarchy is better matched (i.e has a peak in prediction performance) by a higher network layer. In contrast, when comparing the average performance across cells and scans for each convolutional layer and brain area, we find no clear evidence for a hierarchy (Fig.3B) since there is no clear ordering of the brain areas.

VGG16 outperforms classical models. We then investigate whether the lack of an evident hierarchy was due to an overall poor performance of our model. Thus, we first revise how much of the explainable stimulus-driven variability the VGG16-based model captures. To this end we calculate the oracle correlation (the conditional mean of $n-1$ responses without the model) [12] obtaining an upper bound of the achievable performance. Then we evaluate the test correlation of our model

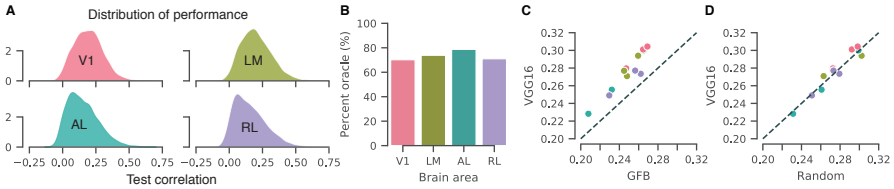


Figure 4: Performance comparison. **A**. Density of the distribution of test correlation across all neurons. **B**. Percentage of oracle performance. In **C** and **D**, each point is the average test correlation performance for one scan across all neurons. Brain areas are color coded as in **A** and the dotted line represents the identity. **B**: VGG16 vs. GFB. **C**: VGG16 vs. Random network core with the same architecture of VGG16

restricted to visual input information (no shifter and no modulator), against the oracle (Fig.4B) and find that VGG16 features explain a substantial fraction of the oracle for the for areas (70–78%)

Second, we consider a subunit energy model with Gabor quadrature pairs as a baseline due to its competitive predictive performance of macaque V1 responses [5]. We replace the core from Fig. 1 with a Gabor filter bank (GFB) consisting of a large number of Gabor filters with different orientations, sizes and spatial frequencies arranged in quadrature pairs, and followed by a squaring nonlinearity [5]. We find that for all areas and scans, the VGG16 core outperformed the GFB (Fig.4C).

Core with random weights performs similarly. The results so far show that VGG16 provides a powerful feature space to predict responses, which may suggest that static object recognition could be a useful high-level goal to describe the function of the mouse visual system. However, we were surprised that most VGG layers led to similar performance. To understand this result better, we also evaluated a core with identical architecture but random weights. This random core performed similarly well as its pre-trained counterpart (Fig.4D), suggesting that training on static object recognition as a high-level goal is not necessary to achieve state-of-the-art performance in predicting neural responses in those four visual areas. Instead, a sufficiently large collection of random features followed by rectification provides a similarly powerful feature space.

The number of LN layers is critical to best match neural activity. Since random features produced by a linear-nonlinear (LN) hierarchy closely match the performance of the pretrained VGG16, we then asked if the number of LN steps – when accounting for multiple input resolutions – was the key common aspect of these networks that yielded the best predictions. Effectively, similar to the case of the pretrained VGG16 core, we found that the fourth and fifth rectified convolutional layers of the random core are the best predictive layers for the four areas we studied (Fig. 5). However, it is important to note that in both cases the increase in performance after the second convolutional layer is only marginal. Overall, we conclude that the nonlinear degree – number of LN stages – rather than the static object recognition training goal dictates how close the representations are to the neural activity.

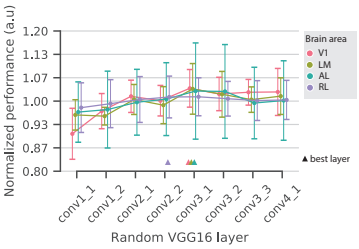


Figure 5: Normalized performance of the random core with VGG16 architecture for all four brain areas. Triangles show the best predictive layers in each case

5 Discussion

In contrast to similar work in the primate, we find no match between the hierarchy of mouse visual cortical areas and the layers of CNNs trained on object categorization. Although VGG16 achieves state-of-the-art performance, it is matched by random weights. There are three implications of our results: First, our work is in line with previous work in machine learning that shows the power of random features [15]. Therefore, we argue that models based on random features should always be reported as baselines in studies on neural system identification. Second, which VGG layer best predicted any given brain area depended strongly on the image resolution we used to feed into VGG16.

We observed a similar effect in our earlier work on primate V1 [5]. Thus, the studies reporting a hierarchical correspondence between goal-driven deep neural networks and the primate ventral stream should be taken with a grain of salt, as they – to the best of our knowledge – do not include this control. Third, optimizing the network for static object recognition alone as a high-level goal does not appear to be the right approximation to describe representations and the visual hierarchy in the mouse cortex. Although our results do not exclude a potential object processing hierarchy in the mouse visual system, they suggest that training with more ethologically relevant visually guided tasks for the mouse could be a more fruitful goal-driven approach to characterize the mouse visual system [16]. For instance, an approach with dynamic stimuli such as those found during prey capture tasks [17] could yield more meaningful features to unveil the functional organization of the mouse visual system.

Acknowledgments S.A.C was supported by the International Max Planck Research School for Intelligent Systems (IMPRS-IS). The research was supported by the German Federal Ministry of Education and Research (BMBF) via the Competence Center for Machine Learning (FKZ 01IS18039A); the German Research Foundation (DFG) grant EC 479/1-1 (A.S.E.), the Collaborative Research Center (SFB 1233, Robust Vision) and the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1, project number 390727645); the Bernstein Center for Computational Neuroscience (FKZ 01GQ1002); the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003. F.S. is supported by the Institutional Strategy of the University of Tübingen (Deutsche Forschungsgemeinschaft, ZUK 63), the Carl-Zeiss-Stiftung, and Amazon AWS through a Machine Learning Research Award. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/IBC, or the U.S. Government.

References

[1] James J DiCarlo and David D Cox. Untangling invariant object recognition. *Trends in cognitive sciences*, 11(8):333–341, 2007.

[2] James J DiCarlo, Davide Zoccolan, and Nicole C Rust. How does the brain solve visual object recognition? *Neuron*, 73(3):415–434, 2012.

[3] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19 (3):356, 2016.

[4] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[5] Santiago A Cadena, George H Denfield, Edgar Y Walker, Leon A Gatys, Andreas S Tolias, Matthias Bethge, and Alexander S Ecker. Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS computational biology*, 15(4):e1006897, 2019.

[6] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[7] Alexander JE Kell, Daniel LK Yamins, Erica N Shook, Sam V Norman-Haignere, and Josh H McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98(3):630–644, 2018.

[8] Davide Zoccolan, Nadja Oertelt, James J DiCarlo, and David D Cox. A rodent model for the study of invariant visual object recognition. *Proceedings of the National Academy of Sciences*, 106(21):8748–8753, 2009.

[9] Sina Tafazoli, Houman Safaai, Gioia De Franceschi, Federica Bianca Rosselli, Walter Vanzella, Margherita Riggi, Federica Buffolo, Stefano Panzeri, and Davide Zoccolan. Emergence of transformation-tolerant representations of visual objects in rat lateral extrastriate cortex. *Elife*, 6:e22794, 2017.

[10] Giulio Matteucci, Rosilari Bellacosa Marotti, Margherita Riggi, Federica B Rosselli, and Davide Zoccolan. Nonlinear processing of shape information in rat lateral extrastriate cortex. *Journal of Neuroscience*, 39(9):1649–1670, 2019.

[11] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[12] Fabian Sinz, Alexander S Ecker, Paul Fahey, Edgar Walker, Erick Cobos, Emmanouil Froudarakis, Dimitri Yatsenko, Zachary Pitkow, Jacob Reimer, and Andreas Tolias. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. In *Advances in Neural Information Processing Systems*, pages 7199–7210, 2018.

[13] Saskia EJ de Vries, Jerome Lecoq, Michael A Buice, Peter A Groblewski, Gabriel K Ocker, Michael Oliver, David Feng, Nicholas Cain, Peter Ledochowitsch, Daniel Millman, et al. A large-scale, standardized physiological survey reveals higher order coding throughout the mouse visual cortex. *bioRxiv*, page 359513, 2018.

[14] Edgar Y Walker, Fabian H Sinz, Emmanouil Froudarakis, Paul G Fahey, Taliah Muhammad, Alexander S Ecker, Erick Cobos, Jacob Reimer, Xaq Pitkow, and Andreas S Tolias. Inception in visual cortex: in vivo-silico loops reveal most exciting images. *bioRxiv*, page 506956, 2018.

[15] Ali Rahimi and Benjamin Recht. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pages 1177–1184, 2008.

[16] Emmanouil Froudarakis, Paul G. Fahey, Jacob Reimer, Stelios M. Smirnakis, Edward J. Tehovnik, and Andreas S. Tolias. The visual cortex in context. *Annual Review of Vision Science*, in press 2019.

[17] Jennifer L Hoy, Iryna Yavorska, Michael Wehr, and Cristopher M Niell. Vision drives accurate approach behavior during prey capture in laboratory mice. *Current Biology*, 26(22):3046–3052, 2016.

A.3 DIVERSE TASK-DRIVEN MODELING OF MACAQUE V4 REVEALS FUNCTIONAL SPECIALIZATION TOWARDS SEMANTIC TASKS

Cadena, Santiago A, Willeke, K. F., Restivo, K., Denfield, G., Sinz, F. H., Bethge, M., Tolias, A. S., & Ecker, A. S. (2024). Diverse task-driven modeling of macaque v4 reveals functional specialization towards semantic tasks. *PLOS Computational Biology*, *20*(5), e1012056

Abstract

Responses to natural stimuli in area V4 – a mid-level area of the visual ventral stream – are well predicted by features from convolutional neural networks (CNNs) trained on image classification. This result has been taken as evidence for the functional role of V4 in object classification. However, we currently do not know if and to what extent V4 plays a role in solving other computational objectives. Here, we investigated normative accounts of V4 (and V1 for comparison) by predicting macaque single-neuron responses to natural images from the representations extracted by 23 CNNs trained on different computer vision tasks including semantic, geometric, 2D, and 3D types of tasks. We found that V4 was best predicted by semantic classification features and exhibited high task selectivity, while the choice of task was less consequential to V1 performance. Consistent with traditional characterizations of V4 function that show its high-dimensional tuning to various 2D and 3D stimulus directions, we found that diverse non-semantic tasks explained aspects of V4 function beyond those captured by individual semantic tasks. Nevertheless, jointly considering the features of a pair of semantic classification tasks was sufficient to yield one of our top V4 models, solidifying V4's main functional role in semantic processing and suggesting that V4's affinity to 2D or 3D stimulus properties found by electrophysiologists can result from semantic functional goals.

Author contributions

Conceptualization: **SC**, AE, FS, AT. Data Curation: **SC**, KW, KR, GD. Formal Analysis: **SC**. Funding Acquisition: AT, AE, FS, MB. Investigation: **SC**, KR, GD. Methodology: **SC**, KW, KR, AE, AT, FB. Project Administration: **SC**, AE, AT. Resources: AT. Software: **SC**, KW, FS. Supervision: AE, AT, FS, MB. Validation: **SC**, KW. Visualization: **SC**. Writing – Original Draft Preparation: **SC**, AE. Writing – Review and Editing: **SC**, AE, FS, AT, KW, KR.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks

Santiago A. Cadena[1,2,3,4]*, **Konstantin F. Willeke**[3,4,5], **Kelli Restivo**[6,7], **George Denfield**[6,7], **Fabian H. Sinz**[1,3,4,5‡], **Matthias Bethge**[2,3‡], **Andreas S. Tolias**[6,7,8‡], **Alexander S. Ecker**[1,9‡]

1 Institute of Computer Science and Campus Institute Data Science, University of Göttingen, Göttingen, Germany, **2** Institute for Theoretical Physics and Centre for Integrative Neuroscience, University of Tübingen, Tübingen, Germany, **3** Bernstein Center for Computational Neuroscience, Tübingen, Germany, **4** International Max Planck Research School for Intelligent Systems, Tübingen, Germany, **5** Institute for Bioinformatics and Medical Informatics, University Tübingen, Tübingen, Germany, **6** Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, Texas, United States of America, **7** Department of Neuroscience, Baylor College of Medicine, Houston, Texas, United States of America, **8** Department of Electrical and Computer Engineering, Rice University, Houston, Texas, United States of America, **9** Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany

‡These authors are joint senior authors on this work.
* santiago.cadena@uni-tuebingen.de

Citation: Cadena SA, Willeke KF, Restivo K, Denfield G, Sinz FH, Bethge M, et al. (2024) Diverse task-driven modeling of macaque V4 reveals functional specialization towards semantic tasks. PLoS Comput Biol 20(5): e1012056. https://doi.org/10.1371/journal.pcbi.1012056

Editor: Wolfgang Einhäuser, Technische Universitat Chemnitz, GERMANY

Received: July 2, 2022

Accepted: April 8, 2024

Published: May 23, 2024

Data Availability Statement: The datasets of V1 and V4 macaque responses to natural images are available at Figshare: https://figshare.com/collections/Monkey_V1_and_V4_single-cell_responses_to_natural_images_ephys_Data_from_Cadena_et_al_2024_/6658331/2. Example code for training models on these data can be found in our Github repository: https://github.com/sacadena/neurovisfit.

Funding: The research was supported by the German Federal Ministry of Education and

Abstract

Responses to natural stimuli in area V4—a mid-level area of the visual ventral stream—are well predicted by features from convolutional neural networks (CNNs) trained on image classification. This result has been taken as evidence for the functional role of V4 in object classification. However, we currently do not know if and to what extent V4 plays a role in solving *other* computational objectives. Here, we investigated normative accounts of V4 (and V1 for comparison) by predicting macaque single-neuron responses to natural images from the representations extracted by 23 CNNs trained on different computer vision tasks including semantic, geometric, 2D, and 3D types of tasks. We found that V4 was best predicted by semantic classification features and exhibited high task selectivity, while the choice of task was less consequential to V1 performance. Consistent with traditional characterizations of V4 function that show its high-dimensional tuning to various 2D and 3D stimulus directions, we found that diverse non-semantic tasks explained aspects of V4 function that are not captured by individual semantic tasks. Nevertheless, jointly considering the features of a pair of semantic classification tasks was sufficient to yield one of our top V4 models, solidifying V4's main functional role in semantic processing and suggesting that V4's selectivity to 2D or 3D stimulus properties found by electrophysiologists can result from semantic functional goals.

Author summary

The functional role of area V4 in the primate visual cortex has been traditionally studied by measuring tuning properties to simple, parametric stimuli, potentially overlooking

Research (BMBF) via the Competence Center for Machine Learning (FKZ 01IS18039A to MB) [https://tuebingen.ai]; the Collaborative Research in Computational Neuroscience (CRCNS) (FKZ 01GQ2107 to FHS and NSF IIS-2113173 to AST) [https://new.nsf.gov/funding/opportunities/collaborative-research-computational-neuroscience]; the German Research Foundation (DFG) (EC 479/1-1 to ASE) [http://www.dfg.de/]; the Collaborative Research Center (SFB 1233 to MB, Robust Vision) [https://uni-tuebingen.de/en/research/core-research/collaborative-research-centers/crc-1233/]; the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1 to MB, project number 390727645)[http://www.ml-in-science.uni-tuebingen.de/]; the Bernstein Center for Computational Neuroscience (FKZ 01GQ1002 to ASE); the National Eye Institute of the National Institutes of Health (R01EY026927 to AST and DP1 EY023176 to AST) [https://nei.nih.gov/]; the NIH-Pioneer award (DP1-OD008301 to AST) [https://commonfund.nih.gov/pioneer/]; the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003 to AST [https://www.iarpa.gov]. This work was also supported by the National Institute of Mental Health (T32EY00252037 to AST) [https://www.nimh.nih.gov/]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: I have read the journal's policy and the authors of this manuscript have the following competing interests: A.S.T. holds equity ownership in Vathes LLC, which provides development and consulting for the framework (DataJoint) used to develop and operate the data analysis pipeline for this publication.

other important aspects that could be revealed with richer, natural stimuli. Here, we combine single-cell recordings of macaque V1 and V4 responses to natural images, and deep learning models trained on multiple computer vision tasks. We found that V4 responses are best predicted by representations that are critical to solve semantic tasks like object and scene classification. Moreover, our results suggest that that V4's affinity to different 2D and 3D stimulus properties likely stems from its involvement in semantic processing. Overall, our diverse task-driven modeling approach enriches our understanding of the functional role of visual areas in the brain.

Introduction

What is the functional role of area V4 in primate visual information processing? One line of evidence suggests that V4 is tuned in a high-dimensional space that facilitates the joint encoding of shape and surface characteristics of object parts [1, 2] (e.g. sensitivity to luminance [3], texture [4] and chromatic contrasts [5], blurry boundaries [6], and luminance gradients [7]). Although very insightful, these experiments are constrained to a relatively small number of stimulus feature directions that potentially miss other important aspects of V4 function that could be unlocked with richer natural stimuli. Recent work used a *transfer learning* approach to infer the functional role of different brain areas. The features extracted by convolutional neural networks (CNNs) pre-trained on object classification *transfer* well to the task of predicting V4 responses to natural stimuli [8–11]. This result has been interpreted as evidence that object recognition is one of the major goals of V4 processing. However, a natural question arises: Do other computational goals beyond object classification explain V4 responses equally well or even better? Recent work using fMRI in humans has attempted to assign different functional goals to different regions of interest in the brain [12–14], but it remains unclear whether single neurons express the same patterns of selectivity as the highly aggregated, indirect fMRI signal.

Inferring the functional role of a brain area using transfer learning is a promising avenue, but it is complicated by the fact that transfer performance is not only determined by the pre-training task itself, but also by the size of the dataset used for pre-training, the network architecture and other factors. A recent development in the computer vision community could be very promising for the neuroscience community, because it mitigates some of these problems: The *taskonomy* project [15] released a dataset that consists of 4.5 million images, ground truth labels for 23 different visual tasks for each image, and pre-trained convolutional neural networks with the same architecture (ResNet50) on each task.

We employed the *taskonomy* project to investigate how well the representations learned by training on each of these visual tasks predict single-cell responses to natural images recorded in macaque areas V4 and V1 (Fig 1). Using this approach, we can isolate the contribution of different pre-training tasks on how well the learned representations match those of areas V1 and V4 without the results being confounded by different network architectures or datasets across tasks. We found that a diverse set of tasks explained V1 responses almost equally well, while scene and object classification tasks provided better accounts for V4 responses than all the alternative tasks tested. We further built models that jointly read from pairs of task representations and found that 2D, 3D, and geometric types of tasks capture additional nonlinearities beyond those captured by individual semantic tasks, consistent with descriptions of V4's heterogeneous tuning observed by electrophysiologists [1, 16]. However, combining the features of both object and scene classification was sufficient to obtain peak V4 performance,

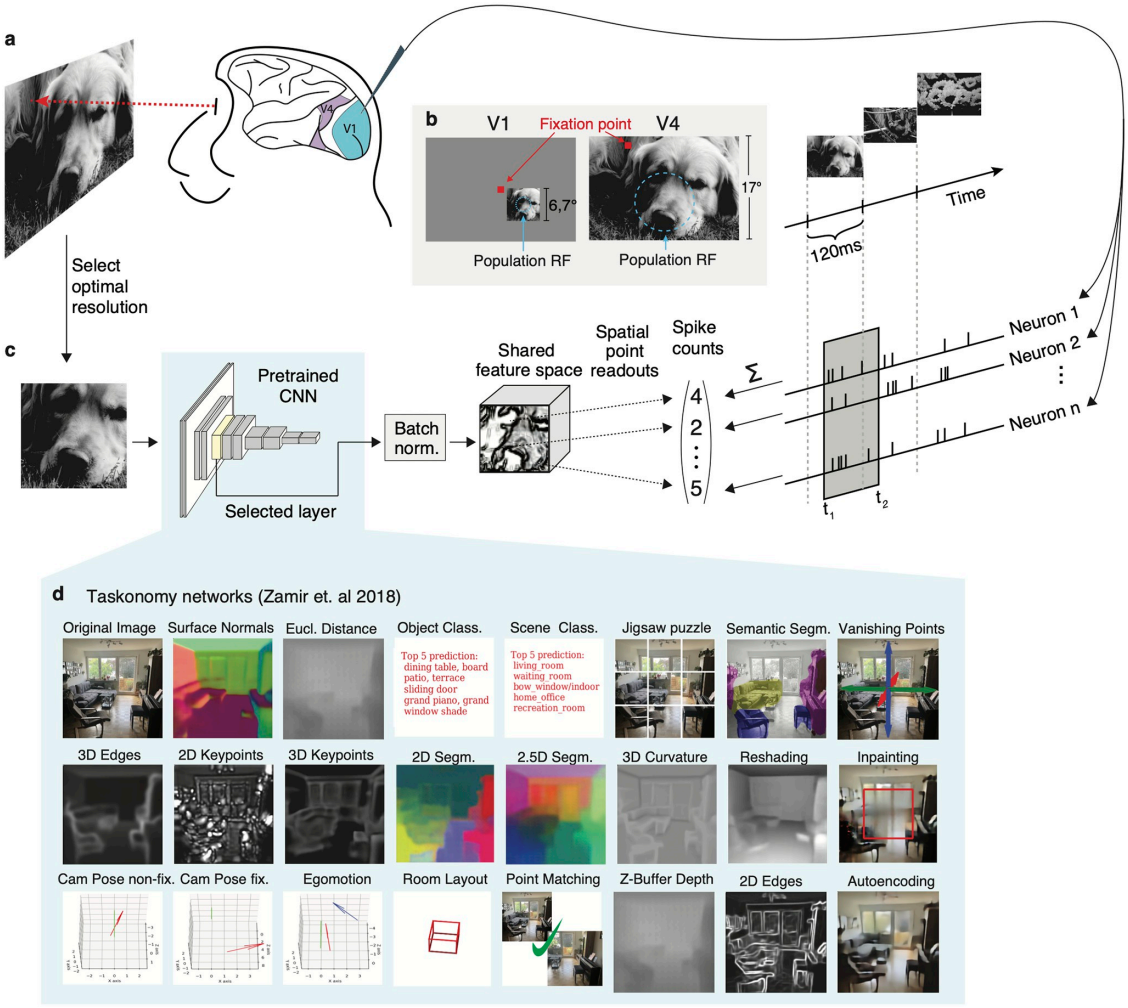


Fig 1. Experimental paradigm and diverse task modeling of neural responses. A, natural images were shown in sequence to two fixating rhesus macaques for 120ms while neural activity was recorded with a laminar silicon probe. Following careful spike-sorting, spike counts were extracted in time windows 40–160ms (V1) and 70–160ms (V4) after image onset. The screen covered ∼ 17˚ × 30˚ of visual field with a resolution of ∼63px/˚. For each area, we showed approx. 10k unique images once (*train set*). A random set of 75 images, identical for both areas, was repeated 40–55 times (*test set*). **B**, For V1 recordings (left), the fixation spot was centered on the screen. Square, gray-scale, 420px ImageNet images (6.7˚) were placed at the center of each session's population receptive field (size: ∼ 2˚, eccentricities: 2˚ − 3˚). In the V4 recordings sessions (right), the fixation spot was accommodated to bring the population receptive field as close to the center of the screen as possible (size: ∼8˚, eccentricities: 8˚ − 12˚). All images were up-sampled and cropped to cover the whole screen. We isolated 458 (V1) and 255 (V4) neurons from 32 sessions of each area. **C**, Predictive model. Cropped input images covering 2.7˚ (V1) and 12˚ (V4) were resized and forwarded through the first *l* layers of a pretrained convolutional neural network (CNN) to produce features that are then batch-normalized and shared by all neurons. The input scale factor was a hyper-parameter, cross-validated on a held-out subset of the train set (*validation set*). The point readout [22] extracts features at a single spatial location and computes a regularized linear mapping to the neural responses for each neuron separately (see Methods). The readout and batch-normalization parameters were jointly learned to minimize the Poisson loss between predicted and observed response rates. **D**, *taskonomy* networks used for feature extraction. We used the pretrained encoder CNNs of these tasks, which share a Resnet50 architecture [20], to build our models and compare their predictive abilities on V1 and V4.

https://doi.org/10.1371/journal.pcbi.1012056.g001

indicating that multiple semantic goals can induce complementary intermediate representations that are predictive of the additional nonlinearities contributed by non-semantic tasks. Overall, our results solidify V4's semantic functional role and explain that V4's affinity to other non-semantic tasks can result from semantic computational goals.

Results

We collected datasets of well-isolated single-cell responses from V4 and primary visual cortex (V1) for comparison. We measured the spiking activity of individual neurons from two awake, fixating rhesus macaques (M1, M2) using a 32-channel linear array spanning multiple cortical layers [17, 18], in response to tens of thousands of grayscale natural images presented in sequence over many trials (Fig 1A). These images were sampled uniformly from the ImageNet Large-Scale Visual Recognition Challenge (ILSVRC-2012) dataset [19] and displayed for 120 ms each without interleaving blanks (see Methods). Most of these images were shown only once (*train-set*) while a selection of 75 images was repeated multiple times (*test-set*). We isolated 458 V1 neurons from 15 (M1) and 17 (M2) sessions at eccentricities 2–3˚; and 255 V4 neurons from 11 (M1) and 21 (M2) sessions at eccentricities 8–12˚. For the V1 sessions, we centered the stimuli on the population receptive field of the neurons. For the V4 sessions, the stimuli covered the entire screen. We obtained image–response pairs by extracting spike counts in the windows 40–160 ms (V1) and 70–160 ms (V4) after image onset (Fig 1A, 1B and 1C), which corresponded to the typical response latency of the neurons in the respective brain area. We computed the stimulus-driven variability of spike counts using the repeated trial presentations on the test-set (see Methods) and found that the mean [± s.d.] fraction of explainable variance of these two areas was not significantly different (0.31 ± 0.18 (V1) and 0.32 ± 0.19 (V4); two-sided *t*-test, $t(711) = -1.152$, $p = 0.2$). Following previous work [18], we excluded unreliable neurons from the performance evaluations where the fraction of explainable variance was lower than 0.15, yielding 202 (V4) and 342 (V1) neurons (S2(A) Fig).

Task-driven modelling of neural responses

We built upon the *taskonomy* project [15], a recent large-scale effort of the computer vision community, in which CNN architectures consisting of *encoder-decoder* parts were trained to solve various visual tasks. The *encoder* provides a low-dimensional representation of the input images from which each task can be (nonlinearly) read-out by the *decoder*. We considered the encoder network of 23 of these tasks, which have been previously categorized into *semantic*, *geometric*, *2D*, and *3D* groups (listed in Fig 1D) based on hierarchical clustering of their encoder representations [15]. We chose these networks because of two key features: 1) all of them were trained on the same set of images, and 2) all encoder networks have the same architecture (ResNet-50 [20]). Any differences we observe across the learned representations are thus caused by the training objective targeted to solve a specific task.

To quantify the match between the representations extracted by intermediate layers of the *taskonomy* networks and V4 representations, we used these networks to built task-driven models [18, 21] of single-neuron recordings in response to natural stimuli: We presented the images that were shown to the monkey to each pretrained network, then extracted the resulting output feature maps from several intermediate layers and fed these to a regularized linear-nonlinear (LN) readout that was specific to each recorded neuron (Fig 1C). This readout acted on the features at a single spatial location [22], preventing any additional nonlinear spatial integration beyond what has been computed by the task-trained network. For each *taskonomy* network, we built one model for each readout layer and optimized hyperparameters (e.g. regularization penalty) for each model. To ensure that the resulting correspondence between

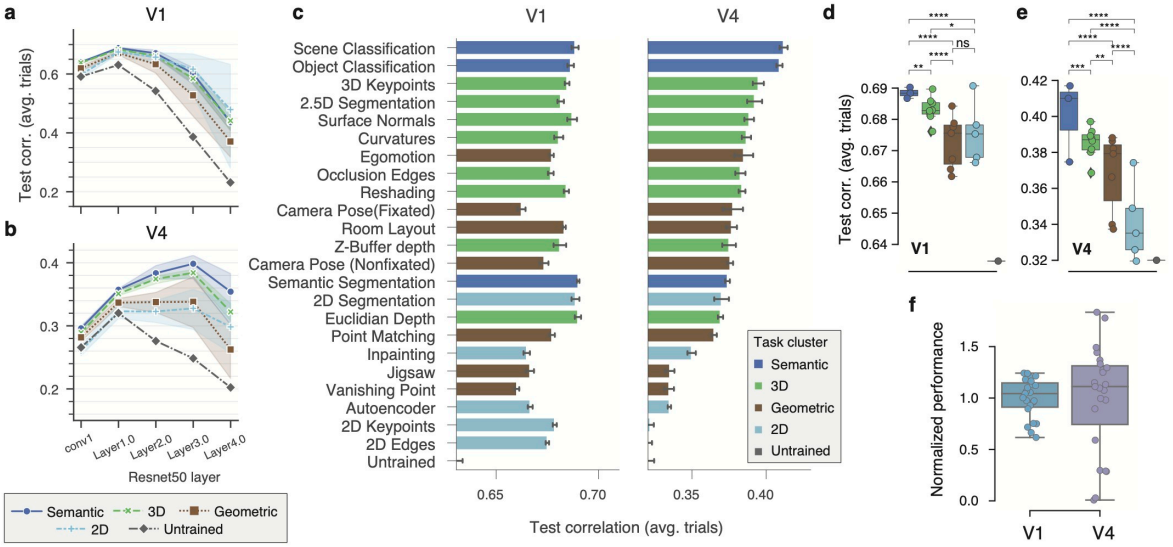


Fig 2. Comparison of diverse task-driven models on V1 and V4. a,b, Average performance of task clusters (see Fig 1F) on V1 (**a**) and V4 (**b**) as a function of network layer. Error bands represent 95% confidence intervals across individual tasks (S3 Fig). Performance is measured as the average test-set correlation across neurons between model predictions and mean spike counts over repeated presentations. All tasks outperform an untrained, random baseline that shares the same architecture (dark gray). The best predictive features for V1 can be found at an early layer (Layer1.0) for all models, while deeper layers yielded peak performance for V4 with semantic and 3D task-clusters outperforming the others. **c**, The individual task-model performance on V1 (left) and V4 (right), maximized over layers and other hyper-parameters in the validation set. Error bars show 1 s.e. of the mean for five random initializations of the same model configuration. Task-models are sorted by performance on V4. Hues represent cluster types (inset). The baseline is the average performance of an untrained network. **d,e**, Comparisons between pairs of task-clusters. Individual task-models as circles, and box-plots depict their distribution for each task-cluster. The number of stars represent the *p-value* upper thresholds (0.05, 10^{-2}, 10^{-3}, 10^{-4}) of each comparison test between pairs of clusters. For each pair, we ran a pairwise Wilcoxon signed rank test between the group contrast performance of individual units $n = 342$ (V1), and $n = 202$ (V4), and applied Holm-Bonferroni correction to account for the six multiple comparisons. **f**, The normalized performance (increment over untrained baseline divided by the mean) for all individual tasks on V1 and V4. The variance across tasks in V4 is significantly different than in V1, rendering it more functionally specialized ($P = .00135$, $n = 23$, Levene's test for equality of variances).

https://doi.org/10.1371/journal.pcbi.1012056.g002

network layers and neural data is not merely driven by the confound between growing receptive field sizes and feature complexity along the network's depth, we optimized the resolution (scale) of the input images on held-out data from the training set (S1 Fig). This prevented us from assigning V4 responses to a layer simply because of the matching receptive field coverage that could result from an arbitrary input resolution, but instead allowed us to find for each model the layer with the best aligned nonlinearities to the data.

V1 is better predicted by task representations than V4

We evaluated the predictive performance of our fitted task-driven models with the correlation between model predictions and the average response across trials. When accounting for input scale (S1 Fig), we found that performance on V4 peaks at a higher layer than V1 (Fig 2A and 2B, and S3 Fig), a signature of hierarchical correspondence and increased complexity of V4 over V1 found in anatomical and latency studies [23]. This functional hierarchy was present in most cases, but could not be explained by the architecture alone: V1 and V4 were assigned to the same layer for networks trained on 2D edges, jigsaw, vanishing points, and the untrained baseline (a network with matching architecture and random weights). However, for several 2D

tasks, the hierarchical assignment was not so strict as higher layers in the network retained the high predictive power from lower layers(S3 Fig).

After determining the optimal layer, scale, and regularization parameters for each brain area and task network on a validation set, we found that V1 responses were better predicted than V4's (Fig 2C). The top *taskonomy*-based model performances (test correlation to average over trails [± standard error of the mean across seeds]) were 0.690 ± 0.0003 (V1) and 0.412 ± 0.0014 (V4). Equivalent results in terms of fraction of explainable variance explained (Methods, [18]) were 0.496 ± 0.0005 (V1) and 0.125 ± 0.0016 (V4) (S5 Fig). This performance discrepancy between V1 and V4 could be explained only partly by differences in selectivity—measured with the selectivity index (SI) [24]—between V1 and V4 responses (mean SI ± s.d. of V1: 0.40 ± 0.23, and V4: 0.56 ± 0.26; two-sided *t*-test: $t(542) = -6.728$, $p = 4 \cdot 10^{-11}$; S2(B) and S2(C) Fig), because the SI of individual neurons was negatively correlated with the performance yielded by our top model ($\rho = -0.35$, $p < 10^{-8}$ on V4 neurons; S2(E) Fig).

Pre-training was very effective: Most models widely outperformed the untrained model baseline with random weights—unlike earlier work in the mouse visual cortex [25]. The two exceptions on V4 were 2D edges and 2D keypoints (Fig 2C), which did not improve over the untrained baseline. These results are in line with previous work [26] where models using pre-trained representations significantly outperformed untrained representations fitted to human fMRI responses in inferior temporal cortex (IT).

Semantic classification tasks predict V4 best, while V1 is well-predicted by diverse tasks

The best predictive task-models on V4 were the two semantic classification tasks: scene classification (0.4117 ± 0.0013), and—consistent with prior work [8, 9]—object classification (0.4089 ± 0.0010). In contrast, we found that in V1 the top models with comparable performance were diverse and not specifically tied to semantic-related tasks—they came from semantic, 3D, and 2D task groups: semantic segmentation (0.6900 ± 0.0003), 2D segmentation (0.6886 ± 0.0010), euclidean depth (0.6898 ± 0.0007) (Fig 2C, left). Interestingly, the performance of semantic segmentation in V4 (a pixel-to-pixel task), did not yield a high performance in comparison (0.3739 ± 0.0008), while it was among the best-performing tasks on V1 (see Discussion).

To unveil trends that apply beyond individual tasks, but that are consistent for functionally related task-groups (i.e. semantic, geometric, 3D, and 2D task clusters), we compared the average performances of each task cluster. Consistent with our previous results at the individual task level, we found evidence at this coarser level for the specialized role of V4 towards semantic tasks: there were significant differences between all pairs of clusters with the semantic group on top. In contrast to observations at the individual task level, and in line with previous work [18, 27], we found that semantic representations were significantly better at predicting V1 responses than other groups (Fig 2D) and identified significant differences between all pairs of groups (except between 2D and geometric). Nevertheless, when looking at the group median performances, we found that the group type was less critical in V1 than V4: all groups in V1 were above 74% of the gap between the untrained baseline and top median group (semantic), while the lowest group median in V4 (2D) reached only 16% of that gap (Fig 2D and 2E).

V4 is more specialized than V1

How important is the specific task objective over the untrained baseline to obtain better predictive performances? We found that the choice of task (and task cluster as shown before) did

not affect performance as much in V1 as it did in V4 (Fig 2D, 2E and 2F). We quantified this functional specialization by computing the variance of the performance of all task-models normalized to their mean (excluding the untrained network). The variance for V4 was higher than for V1 ($p < 0.01$, $n = 23$, Bartlett's test after Fisher's z-transformation; Fig 2F), rendering it more specialized. These results support the traditional notion of generality in the features extracted by V1, as they support multiple downstream tasks, while they also highlight the more specialized role of V4 in visual processing.

Building models that jointly read out from pairs of tasks

Our comparisons at the individual and cluster levels suggest that semantic tasks drive representations that best match ventral visual areas in the brain, especially area V4. However, semantic tasks outperformed the next-best predictive tasks only by a relatively small margin (Fig 2C, right), making it difficult to dismiss these other tasks altogether as computational goals of V4 function.

Moreover, the task-models that directly followed object classification—reaching 81% of the gap between the untrained and scene classification model—extract features relevant for 3D understanding (3D keypoints, 2.5D segmentation, and surface normals estimation) which could be aligned with V4 functions—as recent studies suggest that many cells in V4 are tuned to solid-shape (3D) properties [16]. For example, the 3D keypoints task aims to find points of interest that could be reliably detected even if an object in the scene is observed from different perspectives, and then to extract local surface features at these points. Detecting 3D borders is essential to solve this task because these keypoints tend to be around object corners [28], likely capturing useful information for downstream invariant object recognition.

Do computational goals (like these 3D-related objectives) provide representations that explain aspects of V4 computation beyond those explained by semantic tasks alone? We addressed this question by comparing the individual task-model performances to the performance of a single model that jointly reads out from pairs of pretrained feature spaces (Fig 3A). If we find that a larger response variance is explained by adding a second set of features extracted by a different task-network to the first one, we could claim that the nonlinearities of second network are likely not implemented by the first one.

In practice, however, it is generally harder to find global optima with an increased set of input features when the amount of training data is kept the same. Naively training the joint readout of a pair of tasks from scratch could lead to worse performance than the individual task-model performances. Therefore, we validated different training strategies together with an L_1 regularization penalty (Fig 3B): (1) The readout weights were initialized randomly, and we learned them from scratch, (2) The readout weights corresponding to the first core were initialized with the optimal ones found previously (as per validation set), while those corresponding to the second core were initialized with zeroes. Then we fine-tuned the readout weights of the second core only. This is approach was repeated by swapping the order of the cores. (3) We repeated the second approach, but instead of fine-tuning only the second core readout weights, we fine-tuned all.

We built and trained models for pairs of tasks following these strategies. In addition to scene and object classification, we considered the top three tasks on V4 of each group (Fig 2C, right), and the untrained baseline; and built models with all possible pairwise combinations (Fig 3C and 3D). This means 12 individual tasks and 78 pair task-models per brain area (Fig 3D).

We observed that initializing a portion of the readouts with the already optimized ones consistently outperformed training from scratch, as illustrated in an example involving object

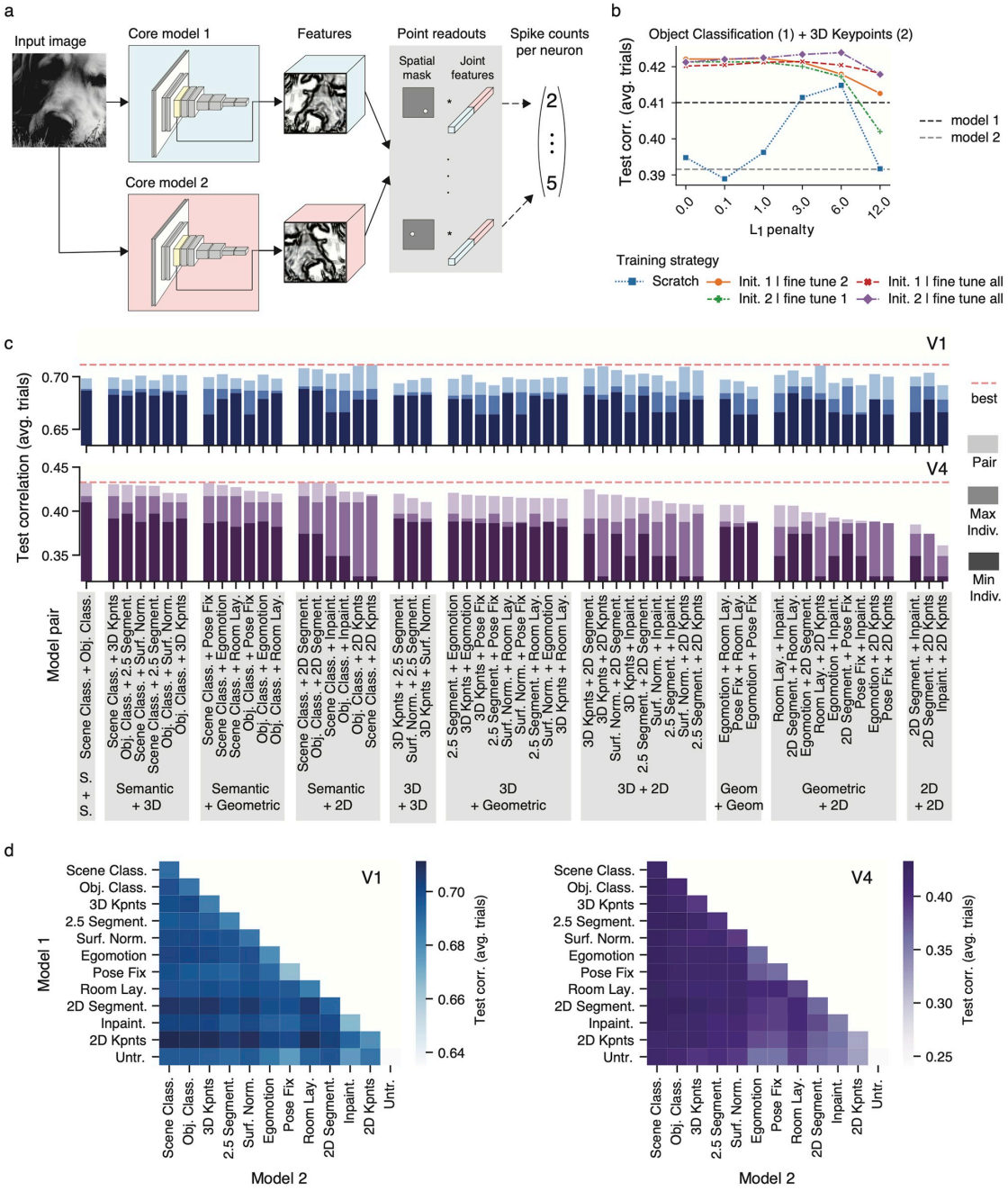


Fig 3. Jointly reading from pairs of tasks. a, Modeling approach. We modified the methods described in Fig 1 to simultaneously read out neural responses from two feature spaces. We forwarded the same input image, at the same scale, to two pretrained *taskonomy* networks, extracted the features at the same layer and concatenated them. We then learned a point readout [22] for each neuron and always keep the cores' weights frozen during training. **b**, Comparison of training strategies when jointly reading from object classification (core 1) and 3D keypoints (core 2). The resulting increased dimensionality makes it harder to find global optima, so we compared—under several regularization strengths (L_1 penalty) –five strategies that leverage

the individual task-models' optimum readout weights as per validation set from the models we trained earlier. We trained from scratch (i.e. initialized the readout weights at random), initialized the readout of core 1 (2) and finetuned the readout of core 2 (1), and initialized the readout of core 1 (2) and finetuned all readout weights. The individual task-model performances are in dotted lines for comparison. **c** Comparison of pair task models in V1 (top) and V4 (bottom). Pairs are grouped based on the task-group identities of pair members. In dark (middle-dark) color saturation, we show the worse (best) individual task model from the pair. In lightest color saturation, we show the performance of the pair model. The highest performing pair model is shown as a red dotted line to facilitate comparisons. Performances are shown in terms of correlation between model predictions and average over trials. Baseline on all bar plots is the performance of the untrained network. **d** Heat map representation of pair task models' performances in V1 (let) and V4 (right). Here, in addition to the pairs in **c**, the pair models with identical members (diagonal), and pairs built between tasks and the untrained network (bottom row) are included.

https://doi.org/10.1371/journal.pcbi.1012056.g003

classification and 3D keypoints (Fig 3B). Moreover, in most cases, fine-tuning all readout weights yielded better results than updating only some of them.

Non-semantic tasks contribute useful nonlinearities beyond those provided by individual semantic classification tasks

We found that the performance of pair task-models was always better than any of their individual constituents (Fig 3C). The average performance improvements over the highest-performing individual task in each pair were 3.9% (V4) and 2.2% (V1). The largest increase in performance in V4 came from a pair comprising 3D keypoints and 2D segmentation, which resulted in a 8.5% improvement over 3D keypoints alone. Similarly, the largest boost in V1 was obtained from a pair consisting of 2D keypoints and 3D keypoints, leading to a 4.0% increase over 2D keypoints alone.

We conducted two control experiments. First, to determine whether the observed performance gains may simply be due to the additional nonlinear capacity of the models, we trained a set of pair task-models by pairing each individual task with an untrained model. In every instance, the performance was found to be superior when pairing with another task instead of the untrained baseline (Fig 3D; last row). Second, training models with higher-dimensional feature spaces required adapting the hyperparameters of the non-convex optimization procedure (regularization, number of iterations, etc.). To ensure that these changes did not trivially lead to a better model, we trained pair task-models consisting of one task's features simply duplicated (Fig 3D; diagonals). We found that incorporating the nonlinearities of a different task always resulted in better performance than a pair of duplicate task features (Fig 3D). These results suggest that the improved performance observed in V1 and V4 is attributed to the additional nonlinear computations provided by the supplementary tasks rather than just the enhanced flexibility resulting from a greater feature dimensionality.

Semantic features are critical to explain the largest fraction of variance in V4

We found that 2D tasks—in particular 2D keypoints—were consistently among the highest performing pairs in V1: 2D keypoints + scene classification (0.711), room layout + 2D keypoints (0.710), object classification + 2D keypoints (0.710), 3D keypoints + 2D keypoints (0.709), surface normals + 2D keypoints (0.709), scene classification + 2D segmentation (0.708). Interestingly, the individual 2D keypoints model was not among the top ten tasks in V1, but pairing it with members of other task groups yields the best performing models, suggesting that 2D keypoints is the most non-redundant task with other top-performing V1 tasks.

On the other hand, we found that the semantic classification tasks were consistently among the top pairs in V4 (Fig 3C): scene classification + object classification (0.432), scene classification + 2D segmentation (0.433), scene classification + fixed camera pose (0.432), scene

classification + inpainting (0.431), object classification + 2D segmentation (0.432), scene classification + 3D keypoints (0.431).

To investigate whether the nonlinearities captured by scene and object classification are equivalent to those captured by other pairs of tasks exhibiting similar performance, we extended our previous methods (Fig 3A) to jointly read from triplets of tasks. Following our previous approach, we fine-tuned all readout weights after initializing the readouts corresponding to two task cores with the optimal weights obtained from pair task models (Fig 3) and the weights corresponding to the third core with zeroes. We focused on studying improvements provided by the features of a third task core over the object + scene classification pair. We observed only minimal improvements over this pair (0.4 ± 0.1%) ranging from 0 to 1.1%. Controls involving an untrained, scene, and object classification tasks as a third core yielded no improvements. These improvements are within the variability range obtained across seeds and are much smaller than than improvements provided by pairs of tasks over individual ones. Therefore, most nonlinearities captured by non-semantic tasks that are helpful to predict V4 responses can emerge from purely semantic training objectives.

Data-rich and robust models predict V1 and V4 better

Taskonomy networks helped us addressed a prevalent issue in most goal-driven system identification studies where multiple sources of variability across models are simultaneously at play (e.g. architecture, training dataset, training strategies, computational goal). By sharing architecture, training dataset, and optimization methods, these networks facilitated a fair comparison of the computational objective. To make progress in understanding the effects of other sources of variation, we built V1 and V4 models that use the representations of widely used, state-of-the-art networks that were largely trained on ImageNet [19] and had varying architectures and different training strategies (Fig 4): AlexNet [29], VGG19 [30], Cornet-S [31], Resnet50 [20], SimCLR [32], Resent50 with adversarial training (Robust L_2, $\epsilon = 0.1$) [33], and Resnet50 trained on ImageNet and Stylized ImageNet (StyleImNet) [34].

We found that—beyond the semantic computational objective—the training dataset is critical in driving representations that best match V4 responses: all ImageNet models yielded higher performance than the top semantic *taskonomy* networks (Fig 4; right). Our results also suggests that a Resnet50 architecture is generally better than others (VGG19, AlexNet, Cornet-S). Interestingly, a shape-biased network (StyleImNet) does not lead to better predictions than a texture-biased counterpart (Resnet50), even though, at the object recognition behavioural level, the visual system exhibits bias to shapes [34]. Furthermore, in line with previous work [35], we found that self-supervised representations from SimCLR had comparable performance to the supervised-trained Resnet50. Although not explicitly trained on object classification, SimCLR provides useful features for this task: a shallow multi-linear-perceptron readout with little supervision achieves competitive accuracy on ImageNet [32]. Finally, we obtained the highest V4 performance (0.4786 ± 0.0011) with the adversarially robust Resnet50 that was trained to produce stable classification outputs under small perturbations of at most 0.1 L_2 size. We observed that increasing the size of this robustness ball led to poorer performance in V4—as it also negatively affects ImageNet top1 accuracy [33]. Overall, these results indicate that the match to V4 neural data can be strongly influenced by the training dataset, architecture, training objective, and adversarial robustness.

Our V1 results revealed trends that partially differ from those observed in V4. First, we found that some ImageNet-trained models (AlexNet and Cornet-S) yielded comparable performance to the top semantic *taskonomy*-based models. However, when using the same Resnet50 architecture, we obtained better performance with ImageNet models. As in V4, self-

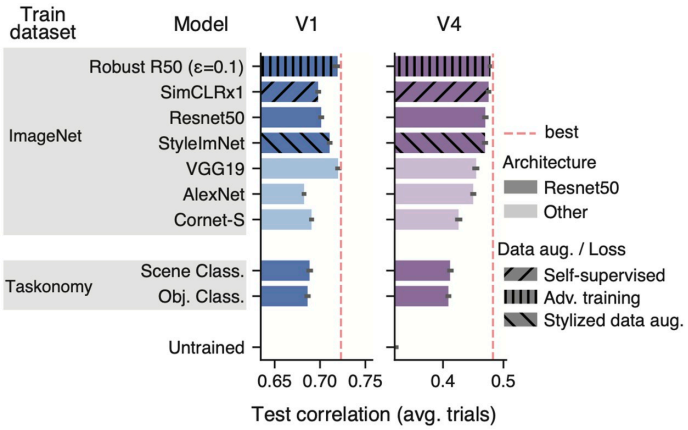


Fig 4. Comparisons with ImageNet-trained networks. We considered CNNs trained on the ImageNet dataset with varying architectures, data augmentation strategies, and losses (supervised vs. self-supervised) and compared them with the *taskonomy* semantic classification networks on V1 (left) and V4 (right). Red dotted line indicates the performance of the best model on each visual area. We found that ImageNet networks were better than *taskonomy* on V4 for any architecture. Self-supervised (SimCLRx1, [32]) and shape-biased (StyleImNet Resnet50, [34]) networks yielded comparable performance in V4 to the supervised, texture-biased Resnet50. The highest performing network on V4 was the adversarially robust Resnet50 [33]. In V1, our top *taskonomy* networks were comparable to AlexNet and Cornet-S but performed worse than the ImageNet counterpart with matching architecture (Resnet50). VGG19 [30] and the robust Resnet50 were the best match to V1.

https://doi.org/10.1371/journal.pcbi.1012056.g004

supervised learning and shape-biased networks performed similarly to their supervised, texture-biased counterpart. However, our best V1 models came from VGG19 (0.7199 ± 0.0008)—in line with [18] –, and the robust Resnet50 (0.7184 ± 0.0027), which is consistent with findings reported in [27] and [36]. In a similar way to V4, increasing the robustness size of the ResNet50 model led to poorer performance. Overall, our results suggest that data-rich and robust models are more effective in capturing V1 and V4 representations.

Discussion

A common challenge when comparing models in task-driven system identification studies is that multiple axes of variability get confounded, making it difficult to draw conclusions about individual factors like architecture or training objectives. We explored multiple normative accounts beyond object classification of V4 and V1 single neuron responses to natural stimuli using the representations of *taskonomy* networks as they have matching architectures and were trained on the same dataset but different computer vision tasks. Beyond solidifying existing evidence suggesting that semantic goals (e.g. object classification) drive representations that best match ventral stream responses [8, 18, 21, 37, 38], our results revealed a high task specialization of V4 function towards semantic tasks in contrast to the more general representations in V1 that can support multiple downstream tasks. Moreover, when predicting V4 responses from pairs of different tasks features, we found that non-semantic tasks contribute useful nonlinearities over individual semantic ones. However, a model that jointly reads from a pair of semantic-driven features alone was among the best of our *taskonomy*-based models of V4, suggesting that semantic objectives can sufficiently capture the nonlinearities explained by other models.

V4 functional role and 3D processing

Over the last decades, the functional role of V4 has been characterized by identifying neuronal tuning directions using creatively synthesized, parametric stimuli (see [1] for a review). This powerful approach has led to multiple insights about V4's preferences, but does require generating hypotheses of stimulus features (e.g. tuning for blurry vs. hard edges) and painstaking experiments to test each of them. Based on the patterns of predictive performance across tasks (and task-pairs) our systems identification approach provides another lens on how we can better understand V4 function. Importantly, our results suggest that these two approaches can inform each other. Specifically, our work predicts a strong relationship between 3D visual processing and V4 function, shifting the long-standing focus of V4's functional role in flat shape processing: 3D representations ranked high after semantic ones (Fig 2), and they captured novel nonlinearities over those of individual semantic tasks (Fig 3). This supports recent discoveries showing that some V4 cells encode solid shape (3D) features [16]. A promising future research direction is to explore the relationship between normative accounts of V4 at the population level and individual cell properties, specially now that we can first probe multiple stimulus features *in-silico* using our best CNNs before moving to the expensive *in-vivo* alternatives.

V1's affinity to 2D representations

Our results revealed the affinity of V1 to 2D processing, consistent with classical views of V1 function [39]. We found that 2D segmentation was among the top *taskonomy* networks (Fig 2), while it also maintained high predictive performance on the deepest layers that explicitly solve the task (S3 Fig). Moreover, we found that the 2D keypoints tasks was consistently among the highest performing pairs (Fig 3). This task approximates the output of a classical computer vision algorithm (SURF [40]) that was engineered to identify points of interests in an image (using an approximation of multi-scale difference-of-Gaussians [DoG]) and to extract features around them (e.g. orientation tuning) that facilitate matching these points across affine transformations of the image (rotation, translation, scaling, shear mapping). This sequence of operations (multi-scale DoGs, and identifying orientation preference with rotated wavelets) bears similarities with classical descriptions of V1 function [39]. Finally, we found that ImageNet-trained CNNs (Fig 4) yielded our best performing V1 models. This observation is consistent with recent observations [18, 27] claiming that object classification goals induce useful V1 nonlinearities beyond energy models using multi-scale Gabor features [41], and that these CNNs can learn known nonlinear phenomena like cross-orientation inhibition [42] that those models miss [43].

Low performance of semantic segmentation

We found that in contrast to object and scene classification tasks, semantic segmentation was not as effective at predicting V4 responses, while it was among the best in V1 (Fig 2). One likely explanation has to do with the architecture of the *taskonomy* semantic segmentation network: unlike modern networks like U-Net [44] or Mask R-CNN (with an FPN backbone [45]) [46] that include lateral connections between encoding and decoding streams at different scales, the decoder of the *taskonomy* network has only access to the high-level output of the encoder to infer a segmentation mask. To keep the architecture of the core consistent across tasks, we considered the encoder's representations. While the early encoder layers capture rich enough representations to predict V1, the later layers underperform on V4, possibly because 1) they miss task-relevant nonlinearities that are only implemented by the decoder network, and 2) they must maintain details about object boundaries at the pixel level in their representations due to the lack of multi-scale lateral connections provided to the decoder. This is

particularly relevant because it contrasts with V4 units, which have been shown to exhibit a degree of translation invariance [9, 47].

Low V4 performance compared to V1

We found that V1 was better predicted by task representations than V4 (Fig 2 and S5 Fig), even though the stimulus-driven variability was comparable between these two areas. We do not believe that the seemingly low explained variance by our V4 models lessens the conclusions drawn in this paper: How well a model performs in absolute terms depends on many experimental factors including the amount of data available to fit the model, the signal-to-noise ratio of the model, the nature of the pre-training dataset etc. In our work, we believe that the reasons for the comparably low values in V4 are two-fold. First, these results are likely due to our experimental design that maintained the sequence of images (without interleaving blanks) on the repeated test-trials. Although this choice makes history effects in V4 visible so future (dynamic) models can account for them, it prevents averaging out the variability introduced by the previous image, because the static image-based models cannot account for it. The dynamic effects are likely stronger in V4 than in V1 because the latter—an earlier visual area—has lower response latency and the image display times (120 ms) may be sufficient to reach stable dynamics (see Fig 2B from [18]). In contrast, display times of 120 ms in V4 have been used to model "core object recognition"—a term coined for the first feed-forward wave of visual processing [8]. Second, the models we compare were trained on the *taskonomy* dataset, whose image statistics deviate from those of the ImageNet images shown in the experiments. Hence, it is expected that these models do not perform on par with a model that was trained in-domain. We do not think that this weakens our conclusions because all *taskonomy* models have the same systematic disadvantage of having to generalize somewhat beyond their training set in terms of image statistics. Because we do not have diverse task labels for ImageNet, we make the assumption that the variable of interest (training objective) does not interact with the image statistics. To verify that the lower performance of our models is indeed caused by those two factors discussed above and not related to highly suboptimal modeling (e.g. readouts, hyperparameter optimization, etc.), we conducted additional control experiments. When using the same data-rich cores as current state-of-the-art models, we found that our models perform similarly to those displayed on the BrainScore leaderboard [37]. We used ImageNet-trained models (Fig 4) and found that our best V4 model was a Resnet50 trained to be robust to small adversarial perturbations [33]. The performance of this model in V4 in terms of correlation coefficient was $\sim$ 0.48, which is in the same ballpark as the top models on the Sanghavi, et al. (2021) [48] benchmark on BrainScore to date.

Apparent small performance differences across tasks

We found significant differences between task groups (Fig 2D and 2E), but the absolute value of performance differences across tasks are seemingly small (test correlation values range from 0.3 to 0.48 in V4), which could potentially raise concerns about our claims. We think these concerns can be partially alleviated for three reasons. First, we found a structured pattern of tasks performances that is largely consistent with expectations based on existing work and far from random. As discussed before, our results predict V4's affinity to semantic and 3D tasks, and highlight the important role of semantic and 2D representations in V1. Second, we found that substantial fractions of neurons were better explained by one task group vs. another across group comparisons (S7 Fig). This means that the overall population differences were not driven by the better explanation of a few neurons, but instead likely represent a systematic effect across the population. Third, we argue that small differences in test performance may

carry meaningful implications about the ability of models to capture neural phenomena. For example, in a previous study [43], we found that the difference between an energy model using a Gabor filter bank (GFB) and a one-layer CNN with a divisive normalization nonlinearity (CNN+DN) was seemingly small ($\sim$ 0.05 of FEVE) in favour of the latter. However, when running *in-silico* experiments, it was found that CNN+DN was able to capture cross-orientation inhibition—a known nonlinear phenomenon in V1—while the GFB model was not. Thus, the difference we find between task performances—although small—can be consequential and should not be quickly dismissed. Although we have not yet a similar (concise) description of V4 tuning that semantic models capture while others do not, we consider this an important question to tackle. A promising idea is to leverage *in-silico* experiments to find stimuli that bring to light the differences between models, for example via controversial stimuli [49, 50] that drive one model's predictions while not the other(s).

Other *taskonomy* approaches to brain representations

In contrast to similar efforts applied to measurements of aggregated neural activity (i.e. fMRI data) from multiple visual areas [12, 51], we evaluated the *intermediate* representations of *taskonomy* networks—not just their final layers—to evaluate their affinities to single-cell responses. We indeed found that intermediate layers provided the best match to the data—regardless of the task—(S4 and S6 Figs) and that diverse tasks offered comparable top performances in V1 and mostly semantic tasks were optimal in V4. [12] considered large, separate regions of interest (OPA, PPA, LOC, EarlyVis, and RSC) in human fMRI and found that object and scene classification were best across areas and subjects. V1 and V4 would likely fall into the same "EarlyVis" ROI, preventing separate conclusions of each of these areas. Dwivedi et al. (2021) [51] did consider more granular areas of the ventral stream including V1 and V4, and found that 2D tasks best explain V1 (consistent with our results) and—in contrast to our findings—also V4. This could be attributed the nature of fMRI where the nonlinear properties of individual cells can be averaged out, and last-layer 2D representations are more linear than other tasks (see similarity across layers in [15]). Overall, our work complements these studies nicely by distilling more insights about V1 and V4 function at the individual cell level than these other data modalities.

Effect of architecture choice

An important limitation of our work and other related studies is that the conclusions hinge on the particular choice of encoder architecture used in the original *taskonomy* work [15]: ResNet50. One may ask whether our results will generalize to other architectures. We believe that our results are likely to generalize if the candidate architectures do not unevenly harm performance across the 23 target tasks. If they were, resulting changes in the neural response predictivity that may come out of their representations could be attributed to sub-optimal task features. For example, a much shallower CNN could change the ratio between semantic and 2D task performance as semantic tasks requires sufficiently nonlinear computations while 2D tasks rely rather on low-level features. Similarly, a deep CNN without skip connections may yield good performance on semantic tasks but perform worse on 2D tasks (compared to ResNet50) because skip connections facilitate low-level features to reach the final layers. For novel architecture choices that share key attributes with our tested architecture (e.g. convolutional nature, sufficient depth, skip connections) it is likely that our our results will hold: Kornblith and colleagues [52] show that architectures from the same class of models with varying depth trained on the same task exhibit high similarity ($>$ 0.9 CKA) between layers found in corresponding fractions of network depth (see their Fig 5). This result suggests that we can

find layers in alternative networks with equivalent brain predictivity because of their representational similarity.

Effect of initialization

Although we account for architecture, training dataset, and optimization methods to be equivalent across core representations, a potential concern not addressed in our study is that there is evidence for variability in intermediate representations of identical networks trained on the same task but initialized differently [53]. This is a challenge faced by most goal-driven system identification studies and it is constrained by the unavailability of multiple instances of a task-network—as it is the case for *taskonomy* encoder networks. That said, we remark that the representational consistency across instances is high for early to mid layers, decreasing with network depth [53]. V1 and V4 were best explained by early to mid levels. Thus, the effects of initialization are not expected to be as strong as if later layers were used. For example, in the concrete case of AlexNet [29], an ImageNet-trained network considered in our work and in the study by [53], we found that the best predictive layer of V4 was *Layer 3*, which had a representational consistency above 0.95 (on a scale of 0–1), similar to that of the earliest layers (see Fig 6 from [53]). Moreover, we did consider multiple tasks for each of the task groups (semantic, 3D, geometric, and 2D), which serves as an implicit control for the variability induced by different network initializations. We found significant differences at the group level (Fig 2 and S7 Fig) that support our main claim favouring the semantic specialization of V4. We make our data and code available to facilitate further progress identifying the effects of network initialization.

Data richness driving better V1 and V4 representations

Are semantic training objectives critical to drive good representations of V1 and V4? We found that a self-supervised network (SimCLR) worked just as well as Resnet50 in line with [35]. Although these self-supervised methods contain no explicit semantic training objective, their inherent data augmentation strategies enforce similar invariances as required for object recognition [54] and their learned representations predict semantic labels well [32]. Moreover, we found that additional data augmentations, particularly those provided by adversarial training, were very effective at improving our models' performance in both areas. These results are aligned with what has been found in V1 [27, 36] that suggested that robustness to perturbations explain V1 responses better. Similar work that goes in the other direction, has shown that making a CNN more brain-like by co-training on object classification and to predict neural responses, makes CNNs more robust to small perturbations [55–57].

Conclusion

Taken together, our results provide evidence for semantic tasks as a normative account of area V4 and predict that particular tasks (e.g. 3D) have strong affinities to V4 representations. Although we are yet to explain most variance, our findings suggest that promising directions to improve the predictive power of V4 responses require using data-rich, robust models (also shown to be useful to match behavioural phenomena [58]), and architectural changes (e.g. recurrence, top-down processing streams) that account for dynamical processing in the brain [59]. Moreover, our multi-task modeling approach, together with *in-silico* experimentation, promises to facilitate the generation of hypotheses about tuning directions via maximally exciting inputs (MEIs) [10, 11, 60, 61], diverse exciting inputs (DEIs) [62, 63], or controversial stimuli [49].

Materials and methods

Ethics statement

All behavioral and electrophysiological data were obtained from two healthy, male rhesus macaque (*Macaca mulatta*) monkeys aged 15 and 16 years and weighing 16.4 and 9.5 kg, respectively, during the time of study. All experimental procedures complied with guidelines of the NIH and were approved by the Baylor College of Medicine Institutional Animal Care and Use Committee (permit number: AN-4367). Animals were housed individually in a large room located adjacent to the training facility, along with around ten other monkeys permitting rich visual, olfactory and auditory interactions, on a 12h light/dark cycle. Regular veterinary care and monitoring, balanced nutrition and environmental enrichment were provided by the Center for Comparative Medicine of Baylor College of Medicine. Surgical procedures on monkeys were conducted under general anesthesia following standard aseptic techniques. To ameliorate pain after surgery, analgesics were given for seven days.

Electrophysiological recordings

We performed non-chronic recordings using a 32-channel linear silicon probe (NeuroNexus V1x32-Edge-10mm-60–177). The surgical methods and recording protocol were described previously [17]. Briefly, form-specific titanium recording chambers and headposts were implanted under full anesthesia and aseptic conditions. The bone was originally left intact and only prior to recordings, small trephinations (2 mm) were made over medial primary visual cortex at eccentricities ranging from 1.4 to 3.0 degrees of visual angle. Recordings were done within two weeks of each trephination. Probes were lowered using a Narishige Microdrive (MO-97) and a guide tube to penetrate the dura. Care was taken to lower the probe slowly, not to penetrate the cortex with the guide tube and to minimize tissue compression).

Data acquisition and spike sorting

Electrophysiological data were collected continuously as broadband signal (0.5Hz–16kHz) digitized at 24 bits. Our spike sorting methods mirror those in [17, 18]. We split the linear array of 32 channels into 14 groups of 6 adjacent channels (with a stride of two), which we treated as virtual electrodes for spike detection and sorting. Spikes were detected when channel signals crossed a threshold of five times the standard deviation of the noise. After spike alignment, we extracted the first three principal components of each channel, resulting in an 18-dimensional feature space used for spike sorting. We fitted a Kalman filter mixture model to track waveform drift typical for non-chronic recordings [64, 65]. The shape of each cluster was modeled with a multivariate t-distribution ($df = 5$) with a ridge-regularized covariance matrix. The number of clusters was determined based on a penalized average likelihood with a constant cost per additional cluster [66]. Subsequently, we used a custom graphical user interface to manually verify single-unit isolation by assessing the stability of the units (based on drifts and health of the cells throughout the session), identifying a refractory period, and inspecting the scatter plots of the pairs of channel principal components.

Visual stimulation and eye tracking

Visual stimuli were rendered by a dedicated graphics workstation and displayed on a 16:9 HD widescreen LCD monitor (23.8") with a refresh rate of 100 Hz at a resolution of 1920 × 1080 pixels and a viewing distance of 100 cm (resulting in $\sim 63 px/^{\circ}$). The monitors were gamma-corrected to have a linear luminance response profile. A camera-based, custom-built eye tracking system verified that monkeys maintained fixation within $\sim 0.95^{\circ}$ around a $\sim 0.15^{\circ}$-sized

red fixation target. Offline analysis showed that monkeys typically fixated much more accurately. After monkeys maintained fixation for 300 ms, a visual stimulus appeared. If the monkeys fixated throughout the entire stimulus period, they received a drop of juice at the end of the trial.

Receptive field mapping and stimulus placing

We mapped receptive fields relative to a fixation target at the beginning of each session with a sparse random dot stimulus. A single dot of size 0.12˚ of visual angle was presented on a uniform gray background, changing location and color (black or white) randomly every 30 ms. Each fixation trial lasted for two seconds. We obtained multi-unit receptive field profiles for every channel using reverse correlation. We then estimated the population receptive field location by fitting a 2D Gaussian to the spike-triggered average across channels at the time lag that maximizes the signal-to-noise-ratio. During V1 recordings, we kept the fixation spot at the center of the screen and centered our natural image stimulus at the mean of our fit on the screen (Fig 1B). During V4 recordings, the natural image stimulus covered the entire screen. We accommodated the fixation spot so that the mean of the population receptive field was as close to the middle of the screen as possible. Due to the location of our recording sites in both monkeys, this equated to locating the fixation spot close to the upper border of the screen, shifted to the left (Fig 1C).

Natural image stimuli

We sampled a set of 24075 images from 964 categories ($\sim$ 25 images per category) from ImageNet [67], converted them to gray-scale (to be consistent with similar system identification studies and reduce complexity), and cropped them to keep the central 420 × 420px. All images had 8 bit intensity resolution (values in [0, 255]). We then sampled 75 as our *test-set*. From the remaining 24000 images, we sampled 20% as *validation-set*, leaving 19200 as *train-set*. We used the same sets of images for V1 and V4 recordings. During a recording session, we recorded $\sim$ 1000 successful trials, each consisting of uninterrupted fixation for 2.4 seconds including 300ms of gray screen (128 intensity) at the beginning and end of the trial, and 15 images shown consecutively for 120ms each with no blanks in between. Each trial contained either train and validation, or test images. We randomly interleaved trials throughout the session so that our test-set images were shown 40–50 times. The train and validation images were sampled without replacement throughout the session, so each train / validation image was effectively shown once or not at all. In V1 sessions, the images were shown at their original resolution and size covering 6.7˚ (screen resolution of 63 pixels per visual angle). The rest of the screen was kept gray (128 intensity). In V4 sessions, the images were upscaled preserving their aspect ratio with bicubic interpolation to match the width of the screen (1920px). We cropped out the upper and bottom 420px bands to cover the entire screen. As a result, we effectively stimulated both the classical and beyond the classical receptive fields of both areas. Once the neurons were sorted, we counted the spikes associated to each image presentation in a specific time window following the image onset. These windows were 40–160ms (V1) and 70–160ms (V4).

Explainable variance

A few isolated neurons were discarded if their stimulus-driven variability was too low [18]. The explainable variance in a dataset is smaller than the total variance because the observation noise prevents even a perfect model to account for all the variance in the data. Thus, targeting neurons that have sufficient explainable variance is necessary to train meaningful models of

visually driven responses. For a neuron's spike count r, the explainable variance $\mathrm{Var}_{\mathrm{exp}}[r]$ is the difference between the the total variance of all observed responses $\mathrm{Var}[r]$ and the variance of the observational noise $\sigma^2_{\mathrm{noise}}$,

$$\mathrm{Var}_{\mathrm{exp}}[r] = \mathrm{Var}[r] - \sigma^2_{\mathrm{noise}} \ . \tag{1}$$

We estimated the variance of the observational noise by computing the variance of a neuron's response r_t in multiple trials t in which we presented the same stimulus x_j and subsequently taking the expectation E_j over all images,

$$\sigma^2_{\mathrm{noise}} = \mathrm{E}_j[\mathrm{Var}_t[r_t|x_j]] \ . \tag{2}$$

Neurons for which the ratio between the explainable to total variance (Eq 3) was below 0.15 were removed. The resulting dataset includes spike count data for 202 (V1) and 342 (V4) isolated neurons, with an average ratio of explainable to total variance (s.d) of 0.306(0.181) and 0.323(0.187), respectively (S2 Fig). All variances were computed using the unbiased estimator and on the test-set responses due to the available repeated trial presentations.

$$\mathrm{EV} = \frac{\mathrm{Var}_{\mathrm{exp}}[r]}{\mathrm{Var}[\mathrm{r}]} \tag{3}$$

Measuring sparseness

We computed the selectivity index [24] (*SI*) as a measure of sparseness for every neuron on its test-set average responses. To do this, we first plotted the fraction of images whose responses where above a threshold, as a function of normalized thresholds. We considered 100 threshold bins ranging from the minimum to the maximum response values. The area under this curve (A) is close to zero for sparse neurons, and close to 0.5 for a uniform distribution of responses. Thus, following Quiroga et. al (2007) [24], we computed the selectivity index as $SI = 1 - 2A$. *SI* approaches 0 for a uniform distribution, and 1 the sparser the neuron is. In this study, we reported the mean and standard deviation of *SI* for both brain areas and found sparser responses in V4 than in V1 (see Results).

Image preprocessing and resizing

An important step of our modeling pipeline was to adjust the size and resolution of the input images to our computational models (S1 Fig). In V1, we effectively cropped the central 2.65˚ (167px) at its original 63px/˚ resolution and downsampled with bicubic interpolation to different target resolutions: 3.5, 7.0, 14, 21, 24.5, and 28px/˚. For practical and legacy reasons [18], in our codebase we first downsampled the images to a resolution of 35px/˚, followed by cropping and another downsampling step to obtained the target sizes and resolutions just reported. In V4, we cropped the images up to the bottom central 12˚, corresponding to 168px at the original 14px/˚ resolution (63px/˚ × 420/1920), in accordance with the neuron's RF positions. These images were similarly downsampled to multiple target resolutions: 1.4, 2.8, 5.6, 8.4, and 11.2px/˚.

Model architecture

Our models of cell responses consisted of two main parts: A pretrained core that outputs nonlinear features of input images, and a *spatial point readout* [22] that maps these features to each neuron's responses. We built separate model instances for each visual area, input image

resolution, task-dependent pretrained CNN, intermediate convolutional layer, regularization strength, and random initialization. Input images **x** were forwarded through all layers up to the chosen layer *l*, to output a tensor of feature maps $l(\mathbf{x}) \in \mathbb{R}^{w \times h \times c}$ (**w**idth, **h**eight, **c**hannels). Importantly, the parameters of the pretrained network were always kept fixed. We then applied batch-normalization [68] (Eq 4), with trainable parameters for scale (γ) and shift (β), and running statistics mean (μ) and standard deviation (σ). These parameters were held fixed at test time (i.e. when evaluating our model). Lastly, we rectified the resulting tensor to obtain the final nonlinear feature space ($\Phi(\mathbf{x})$) shared by all neurons, with same dimensions as *l*. The normalization of CNN features ensured that the activations of each feature map (channel) have zero mean and unit variance (before rectification), facilitating meaningfully regularized readout weights for all neurons with a single penalty—having input features with different variances would implicitly apply different penalties on their corresponding readout weights.

$$\mathrm{BN}(x) = \gamma \cdot \frac{x - \mu}{\sigma} + \beta \tag{4}$$

The goal of the readout was to find a linear-nonlinear mapping from $\Phi(\mathbf{x})$ to a single scalar firing rate for every neuron. Previous approaches have attempted to 1) do dimensionality reduction on this tensor and regress from this components (e.g. partial least squares) [8]; 2) learn a dense readout with multiple regularization penalties over space and features [18]; and 3) factorize the 3D readout weights into a lower-dimensional representation consisting of a spatial mask matrix and a vector of feature weights [69]. In this work we used the recently proposed spatial point readout [22]—also called *Gaussian readout* by the authors—that goes a step further and restricts the spatial mask to a single point. Per neuron, it computes a linear combination of the feature activations at a spatial position, parametrized as (*x*, *y*) relative coordinates (the middle of the feature map being (0, 0)). Training this readout poses the challenge of maintaining gradient flow when optimizing the objective function. In contrast to previous approaches that tackle this challenge by recreating multiple subsampled versions of the feature maps and learn a common relative location for all of them [70], the *Gaussian readout* learns the parameters of a 2D Gaussian distribution $\mathcal{N}(\mu_n, \Sigma_n)$ and samples a location during each training step for every n^{th} neuron. Σ_n is initialized large enough to ensure gradient flow, and is then shrunken during training to have a more reliable estimate of the mean location μ_n. At inference time (i.e. when evaluating our model), the readout is deterministic and uses position μ_n. Although this framework allows for rotated and elongated Gaussian functions, we found that for our monkey data, an isotropic formulation of the covariance—parametrized by a single scalar σ_n^2—was sufficient (i.e. offer similar performance as the fully parametrized Gaussian). Thus, the total number of parameters per neuron of the readout were $c + 4$ (channels, bivariate mean, variance, and bias). Finally, the resulting dot product between the features of $\Phi(\mathbf{x})$ at the chosen location with an L_1 regularized weight vector $\mathbf{w}_n \in \mathbb{R}^c$ was then followed by *f*, a pointwise nonlinear function ELU [71] offset by one (ELU + 1) to make responses positive (Eq 5).

$$\hat{r}_n(\mathbf{x}) = f\Big(\sum_k \Phi_{\mu_{n,x},\mu_{n,y},k}(\mathbf{x}) w_{n,k} + b_n\Big) \tag{5}$$

Beyond offering comparable performance compared to the factorized readout alternative with far less parameters, the most important motivation to use a single point readout was to make sure that all spatial nonlinear computations happen in the pretrained core feature extractor. We could draw mistaken claims about the nonlinear power of a feature space by computing new ones in the readout that combine rectified features computed at different spatial positions. For example, a readout that rectifies features produced by multiple simple cells with

similar orientation at different locations can easily approximate phase invariance (i.e. complex cells) [39].

Model training

We trained every model to minimize the summed Poisson loss across N neurons between observed spike counts r and our predicted spike rate $\hat{r}$ (Eq 6, first term) in addition to the L_1 regularization of the weights (Eq 6, second term) with respect to the batch-normalization, and readout parameters.

$$\mathcal{L} = \sum_{i=1}^{N} (\hat{r}_n - r_n \log \hat{r}_n) + \lambda \sum_{n,k} |w_{nk}| \tag{6}$$

Since neurons across session from the same visual area didn't necessarily *see* the same images (they were differently drawn over sessions), during each training step, we cycled through all sessions of the same visual area, sampling for each of them a fixed batch size of image-response pairs without replacement and kept track of the gradients of the loss with respect of our trainable parameters. Once a cycle was through, the gradients were added to execute an update of the weights of the weights based on the Adam optimizer [72]—an improved version of stochastic gradient descent. The initial learning rate was $3 \cdot 10^{-4}$ and momentum 0.1. We continued to exhaust image-response pair batches from all sessions until the longest session was exhausted to count a full epoch. Once all image-response pairs had been drawn from a session, we restarted sampling batches from all available image-response pairs.

Every epoch, we temporarily switched our model into evaluation mode (i.e. we froze the batch-normalization running statistics), and computed the Poisson loss on the entire single trial *validation-set*. We then used early stopping to decide whether to decay the learning rate: we scaled the learning rate by a factor of 0.3 once the validation loss did not improve over five consecutive epochs. Before decaying the learning rate, we restored the weights to the best ones up to that point (in terms of validation loss). We ran the optimization until four early stopping steps were completed. On average, this resulted in $\sim$ 50 training epochs (or $\sim$ 40 minutes on one of our GPUs) per model instance.

Taskonomy networks

The *taskonomy* networks [15] are encoder-decoder CNNs trained on multiple computer vision tasks. The original goal of the authors was to identify a taxonomy of tasks that would facilitate efficient transfer learning based on the encoder representations of these networks. Importantly for our study, all these networks were trained by the authors on the same set of images, which have labels for all tasks. These images consisted of 120k indoor room scenes. In this work, we used the encoder architecture of these networks, which was based on a slightly modified version of Resnet50 [20] that excluded average-pooling, and replaced the last stride 2 convolution with stride 1. However, these modifications did not change the number of output features of the intermediate layers we considered, keeping our *taskonomy*-based results fairly comparable with those of the original Resnet50.

The Resnet50 architecture—originally developed to solve ImageNet [67]—is made up of a series of hierarchical stages that include 1) an initial strided convolutional layer (`conv1`) followed by batch normalization, rectification, and max-pooling; 2) four processing layers, with 3,4,6, and 3 residual blocks, respectively; and 3) a final average pooling layer that maps features to the number of classes. Each residual block amounts to the rectified sum of two pathways:

one that simply projects the input to the end (i.e. skip connection), and a second that consists of three successive convolutional layers with sizes 1, 3, and 1. In this work we trained models on the output of the first convolutional layer (`conv1`), and the output of the first residual block of each processing layer (i.e. `layer1.0`, `layer2.0`, `layer3.0`, `layer4.0`). The corresponding number of output feature maps (channels) for these layers was 64, 256, 512, 1024, and 2048, respectively.

We used several *taskonomy* encoder networks, listed in (Fig 1C). The structure of the representations of these networks was presented by the authors via a metric of similarity across tasks: with agglomerative clustering of the tasks based on their transferring-out behavior, they built a hierarchical tree of tasks (see Figure 13 of their paper [15]). They found that the tasks can be grouped into 2D, 3D, low dimensional geometric, and semantic tasks based on how close (i.e. how similar) they are on the tree. We now briefly describe the tasks (for more details, see Supplementary Material from [15]):

2D tasks. *Autoencoding* PCA finds a low-dimensional latent representation of the data. *Edge Detection* responds to changes in texture. It is the output of a Canny edge detector without nonmax suppression to enable differentiation. *Inpainting* reconstructs missing regions in an image. *Keypoint Detection(2D)* both detects locally important regions in an image (keypoints), and extracts descriptive features of them that are invariant across multiple images. The output of SURF [40] was the ground-truth output of this task. *Unsupervised 2D Segmentation* uses as ground-truth the output of Normalized cuts [73] which tries to segment images into perceptually similar groups.

3D tasks. *Keypoint Detection (3D)* are like the 2D counterpart, but derived from 3D data, accounting for scene geometry. The output of the NARF algorithm [28] was the ground-truth output of this task. *Unsupervised 2.5D Segmentation* uses the same algorithm as 2D, but the labels are not only computed from RGB image, but also jointly from aligned depth, and surface normal images. It thus has access to ground-truth 3D information. *Surface Normal Estimation* are trained directly on the ground-truth surface normal vectors of the 3D meshes of the scene. *Curvature Estimation* extracts principal curvatures at each fix point of the mesh surface. *Edge Detection (3D)* (Occlusion Edges) are the edges where an object in the foreground obscures the background. It depends on 3D geometry and it is invariant to changes in color and lighting. In *Reshading*, the label for an RGB image is the shading function that results from having a single light point at the camera origin, multiplied by a constant albedo (amount of diffuse reflection of light radiation). *Depth Estimation, Z-Buffer*. *Depth Estimation, Euclidian* measures the distance between each pixel to the camera's optic center.

Geometric tasks. *Relative Camera Pose Estimation, Non-Fixated* predicts the relative six degrees of freedom (yaw, pitch, roll, x, y, z) of the camera pose between two different views with same optical centers. *Relative Camera Pose Estimation, Fixated* is a simpler variant of the previous one where the center pixel of the two inputs is always the same physical 3D point—yielding only five degrees of freedom. *Relative Camera Pose Estimation, Triplets (Egomotion)* matches camera poses for input triplets with a fixed center point. *Room Layout Estimation* estimates and aligns 3D bounding boxes around parts of the scene. *Point Matching* learns useful local feature descriptors that facilitate matching scene points across images. *Content Prediction (Jigsaw)* unscrambles a permuted tiling of the image. *Vanishing Point Estimation* predicts the analytically computed vanishing points corresponding to an x, y, and z axis.

Semantic tasks. *Object Classification* uses knowledge distillation from a high-performing network trained on ImageNet [67] where its activations serve as a supervised signal (within the manually selected 100 object classes appearing in the *taskonomy* dataset). *Scene Classification* follows a similar approach, but uses a network trained on MITPlaces [74] with 63 applicable indoor workplace and home classes for supervised annotation of the dataset. *Semantic*

Segmentation also follows the same supervised annotation procedure using a network trained on COCO [75] dataset with 17 applicable classes.

Finally, we included a control with matching architecture (Resnet50) but with random initialization.

Other network architectures

In addition to the task-models based on *taskonomy* pretrained networks, we also built models with CNN feature extractors pretrained on the large image classification task ImageNet (ILSVRC2012) [67]. This is a dataset of 1.2 million images belonging to 1000 classes. In addition to the original Resnet50 [20] described before, we also considered other popular architectures (S4 Fig): AlexNet [29], VGG19 [30], and Cornet-S [31].

AlexNet [29] consists of five convolutional layers with rectification, three max-pooling layers (between the first and second, second and third, and after the final convolutional layers), two fully connected layers after the last convolutional layer, and a final softmax layer. We used the output of all five convolutional layers in our study: `conv1_1`, `conv2_1`, `conv3_1`, `conv4_1`, `conv5_1`. Their number of output feature maps is 96, 256, 384, 384, and 256, respectively. We used the pretrained Pytorch implementation of this network found in the `torchvision` model zoo.

VGG19 [30] consists of 16 rectified convolutional layers that can be grouped into five groups (named `conv1` to `conv5`) with 2, 2, 4, 4, and 4 convolutional layers with 64, 128 256, 512, and 512 feature maps, respectively; and a pooling layer after every group. Finally, three fully connected and a softmax layers map the convolutional features to the 1000 predictions for each class. We used the original weights provided by [30], and not the default Pytorch version available in the `torchvision` model zoo.

Cornet-S [31] has a recurrent network architecture designed with known biological computations in the brain attributed to core object recognition. The network consist of a sequence of five main modules conveniently named `V1`, `V2`, `V4`, `IT`, and `decoder`. The `V1` module consist of two subsequent convolutional layers with batch normalization and rectification with 64 channels in total; `V2-IT` are recurrent modules with an initial convolutional layer with batch normalization, and a recurrent series of three convolutional layers with rectification and batch normalization. The number of features are 128, 256, and 512 with 2, 4, and 2 recurrent time steps for `V2`, `V4`, and `IT`, respectively. The final `decoder` consist of average pooling, and a linear mapping to the 1000 output classes. The implementation of this network, including pretrained weights can be found here: https://github.com/dicarlolab/CORnet. We used the output of the first four modules of this network and found that V4 was best predicted by the corresponding `V4` module, but our V1 data was best predicted (although only marginally better than `V1`) by the `V2` module (S4 Fig).

Model configurations

In this study, we fitted a large set of task-models ($> 10,000$) that include all viable combinations of 1) brain areas (2: V1, V4); 2) input resolutions (5); 3) pretrained CNNs (23 *taskonomy*, 1 random); 4) intermediate convolutional layers (5), 5) L_1 regularization strengths (1–3 for most models); and 6) random initialization (5 seeds). Because of the receptive field size of higher layers in all networks, only large enough input resolutions were permitted in those cases. We used only 1–3 regularization penalties for most model configurations because we found that the optimal parameters from a fine-grained search of a single model where also appropriate for the corresponding layers of the *taskonomy* networks—actual optimal penalties led to negligibly differences in validation performance (within the noise of random seeds). The

specific values we cross-validated over (after the fine-grained search) were: $\lambda_{conv1} = \{0.33, 1, 3\}$, $\lambda_{layer1.0} = \{3\}$, $\lambda_{layer2.0} = \{3, 6\}$, $\lambda_{layer3.0} = \{3, 9\}$, $\lambda_{layer4.0} = \{6, 12\}$.

Performance evaluation

We computed the Pearson correlation between a model's predictions with the average response over multiple presentations of our test-set to get comparable values to published results (e.g. [37]).

Furthermore, we also report the fraction of explainable variance explained (FEVE) (S5 Fig). The FEVE per neuron is given by Eq 7

$$\mathrm{FEVE} = 1 - \frac{\mathrm{Var}_{\mathrm{res}}[r]}{\mathrm{Var}_{\mathrm{exp}}[r]} \tag{7}$$

which utilizes the variance that is explainable in principle, $\mathrm{Var}_{\mathrm{exp}}[r]$ (Eq 1), and the variance of the residuals corrected by the observation noise,

$$\mathrm{Var}_{\mathrm{res}}[r] = \frac{1}{N}\sum_{j}^{N}(r_j - \hat{r}_j)^2 - \sigma^2_{\mathrm{noise}} \ , \tag{8}$$

where j indexes images. This measure corrects for observation noise, which variance $\sigma^2_{\mathrm{noise}}$ we estimated with Eq 2. To compute model performance we averaged the FEVE across neurons.

Computational tools and libraries

For this study, we used Pytorch [76], Numpy [77], scikit-image [78], matplotlib [79], seaborn [80], DataJoint [81], Jupyter [82], and Docker [83]. We also used the following open source libraries: `neuralpredictors` (https://github.com/sinzlab/neuralpredictors) for torch-based functions for data loading, model implementation, model training, and evaluation. `nnfabrik` (https://github.com/sinzlab/nnfabrik) for DataJoint-based pipelines, `ptrnets` (https://github.com/sacadena/ptrnets) for readily available pretrained CNNs and access to their intermediate layers.

Supporting information

S1 Fig. A case for input scale optimization. a, A single excitatory neuron from visual cortex, recorded from a head-anchored monkey sitting at a certain distance from a screen and fixating on a spot; extracts a nonlinear function of the input stimulus with a specific receptive field coverage. **b**, A pretrained deep convolutional neural network (CNN) extracts several nonlinear feature maps at each of its intermediate layers. A single output unit of a feature map computes a nonlinear function on its analytical receptive field with a fixed size in pixels. Even if the real neuron's nonlinear function was exactly matched to that of a CNN unit, we would have troubles finding it if we were to forward the input image at the wrong input resolution (in terms of pixels per visual angle). It is oftentimes difficult to predict *a priori* the optimal resolution at which a certain layer extracts the right nonlinearities that best match our responses, especially when the receptive field sizes of neurons are difficult to estimate for higher visual areas, and when recording beyond the foveal region of the visual field. We thus treated the input resolution as a hyperparameter that we cross-validate on the validation set. This facilitates removing the confound between the degree of nonlinearity and receptive field growth when trying to establish hierarchical correspondence between CNN layers and the biological visual system. (TIFF)

S2 Fig. Response properties. a, Explainable variance (Eq 3) distribution of V1 ($n = 458$) and V4 ($n = 255$) neurons. Red line shows the threshold (0.15) we chose to filter unreliable neurons from our test evaluations. **b**, Curves of the fraction of images evoking responses larger than a threshold vs. threshold value (normalized). We selected 100 evenly spaced thresholds between minimum and maximum value of the responses. Results for 342 neurons in V1 (left) and 202 neurons in V4 (right) show each neuron's curve (gray). A small sample of curves were colored for clarity. We then computed for each neuron the selectivity index (SI) [24] as $1 - 2AUC$ where the AUC is the area under the curve. **c**, Density distribution of selectivity indices in V1 and V4. Two-sided t-test shows that means are different between areas. **d**, Density distribution of kurtosis statistic computed for each neuron over the test images in V1 and V4. Two-sided t-test shows that means are different between areas, highlighting increase sparsity in V4. **e**, We evaluated how well SI correlates with the predictive performance of our best model on each area (using features of Robust Resnet50) and found that selectivity index only weakly explains V1 and V4 test performance.
(TIFF)

S3 Fig. Individual task-model performances on V1 (upper row) and V4 (bottom row) as a function of network layer organized in columns by the task-clusters [15]. The task-model labels are shared between V1 and V4, and placed to the right of each column. Each line represents the average performance over seeds of the mean performance over neurons of the best task-model configuration in the validation set. That means that these lines represent the test set performance after pooling over input scales, and hyper-parameters (i.e. regularization penalty). Bars represent 95% confidence intervals of 1 s.e. of the mean for five seeds. We measured performance as the average test score over single units ($n_{V1} = 458$, $n_{V4} = 255$) calculated as the correlation between model predictions and mean responses over repetitions.
(TIFF)

S4 Fig. Single-trial correlation performance of ImageNet-based models at multiple input resolutions on the validation set of V1 (upper row) and V4 (bottom row). We considered four popular CNNs with different architectures, pretrained on ImageNet (columns; from left to right: AlexNet [29], VGG-19 [30], Resnet50 [20], Cornet-S [31]). For each network, we built neural predictive models that use features from multiple layers (x axis) that span the depth of the network. Each dot in the plot is the average over seeds of the best model configuration pooled over regularization parameters (see Methods). Assigning a layer to a brain area depends on the input scale—the peak of the curves shifts across input resolutions. Moreover, optimizing the layer using the wrong input resolution may lead to sub-optimal performance (S1 Fig). We found that all of these models reveal a hierarchical ordering of nonlinear computations in the two areas, even when we account for input scale—V1 is predicted always by an earlier layer than V4 (dotted vertical lines represent the most predictive layer over scales).
(TIFF)

S5 Fig. Task-model performances in terms of fraction of explainable variance explained (FEVE). a, Individual task-model performances on V1 (upper row) and V4 (bottom row) as a function of network layer organized in columns by the task-clusters [15]. Equivalent to S3 Fig but performance is measured in terms of FEVE. **b**, Comparison of diverse task-driven models on V1 and V4 measured in FEVE (Fig 2A and 2B). **c**, Tasks performances in terms of FEVE after optimizing over layers and hyperparameters on the validation set ordered as in Fig 2C.
(TIFF)

S6 Fig. Optimal input scale and layers for task-models on V4. In contrast to area V1 where the optimal layer and scale was shared among all task-models (Layer1.0 and 21px/°), there was

variability of the optimal layer in the V4 task-models. In some models, including the untrained network, Layer1.0 with a low input resolution was optimal. The top performing models, including the two semantic classification, and most of 3D tasks (Fig 2C) chose an intermediate resolution ($\sim$ 5.6px/°) at Layer3.0. Interestingly, most geometric and 2D tasks yielded optimal performances at the same layer, but at a higher resolution.
(TIFF)

S7 Fig. Comparison of the average task-cluster performance on single-neurons in V1 (a) and V4 (b). The dotted line in each pairwise comparison represents the identity and the panels in the main diagonal shows the performance distribution of each task-cluster. A pairwise Wilcoxon signed rank test reveal that the differences between task-clusters were significant (Fig 2D and 2E). Insets show the percentage of neurons better explained by one task group (row) vs another (column).
(TIFF)

Acknowledgments

S.A.C. and K.F.W thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS). Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DARPA, DoI/IBC, or the U.S. Government.

Author Contributions

Conceptualization: Santiago A. Cadena, Fabian H. Sinz, Andreas S. Tolias, Alexander S. Ecker.

Data curation: Santiago A. Cadena, Konstantin F. Willeke, Kelli Restivo, George Denfield.

Formal analysis: Santiago A. Cadena.

Funding acquisition: Fabian H. Sinz, Matthias Bethge, Andreas S. Tolias, Alexander S. Ecker.

Investigation: Santiago A. Cadena, Kelli Restivo, George Denfield.

Methodology: Santiago A. Cadena, Konstantin F. Willeke, Kelli Restivo, Fabian H. Sinz, Andreas S. Tolias, Alexander S. Ecker.

Project administration: Santiago A. Cadena, Andreas S. Tolias, Alexander S. Ecker.

Resources: Andreas S. Tolias.

Software: Santiago A. Cadena, Konstantin F. Willeke, Fabian H. Sinz.

Supervision: Fabian H. Sinz, Matthias Bethge, Andreas S. Tolias, Alexander S. Ecker.

Validation: Santiago A. Cadena, Konstantin F. Willeke.

Visualization: Santiago A. Cadena.

Writing – original draft: Santiago A. Cadena, Alexander S. Ecker.

Writing – review & editing: Santiago A. Cadena, Konstantin F. Willeke, Kelli Restivo, Fabian H. Sinz, Andreas S. Tolias, Alexander S. Ecker.

References

1. Pasupathy A, Popovkina DV, Kim T. Visual functions of primate area V4. Annual Review of Vision Science. 2020; 6:363–385. https://doi.org/10.1146/annurev-vision-030320-041306 PMID: 32580663

2. Pasupathy A, Connor CE. Shape representation in area V4: position-specific tuning for boundary conformation. Journal of neurophysiology. 2001; 86(5):2505–2519. https://doi.org/10.1152/jn.2001.86.5.2505 PMID: 11698538
3. Bushnell BN, Harding PJ, Kosai Y, Bair W, Pasupathy A. Equiluminance cells in visual cortical area V4. Journal of Neuroscience. 2011; 31(35):12398–12412. https://doi.org/10.1523/JNEUROSCI.1890-11.2011 PMID: 21880901
4. Kim T, Bair W, Pasupathy A. Neural coding for shape and texture in macaque area V4. Journal of Neuroscience. 2019; 39(24):4760–4774. https://doi.org/10.1523/JNEUROSCI.3073-18.2019 PMID: 30948478
5. Conway BR, Moeller S, Tsao DY. Specialized color modules in macaque extrastriate cortex. Neuron. 2007; 56(3):560–573. https://doi.org/10.1016/j.neuron.2007.10.008 PMID: 17988638
6. Oleskiw TD, Nowack A, Pasupathy A. Joint coding of shape and blur in area V4. Nature communications. 2018; 9(1):1–13. https://doi.org/10.1038/s41467-017-02438-8 PMID: 29386511
7. Hanazawa A, Komatsu H. Influence of the direction of elemental luminance gradients on the responses of V4 cells to textured surfaces. Journal of Neuroscience. 2001; 21(12):4490–4497. https://doi.org/10.1523/JNEUROSCI.21-12-04490.2001 PMID: 11404436
8. Yamins DL, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proceedings of the national academy of sciences. 2014; 111(23):8619–8624. https://doi.org/10.1073/pnas.1403112111 PMID: 24812127
9. Pospisil DA, Pasupathy A, Bair W. 'Artiphysiology'reveals V4-like shape tuning in a deep network trained for image classification. Elife. 2018; 7:e38242. https://doi.org/10.7554/eLife.38242 PMID: 30570484
10. Bashivan P, Kar K, DiCarlo JJ. Neural population control via deep image synthesis. Science. 2019; 364(6439):eaav9436. https://doi.org/10.1126/science.aav9436 PMID: 31048462
11. Willeke KF, Restivo K, Franke K, Nix AF, Cadena SA, Shinn T, et al. Deep learning-driven characterization of single cell tuning in primate visual area V4 unveils topological organization. bioRxiv. 2023; p. 2023–05.
12. Wang A, Tarr M, Wehbe L. Neural taskonomy: Inferring the similarity of task-derived representations from brain activity. Advances in Neural Information Processing Systems. 2019; 32.
13. Dwivedi K, Roig G. Representation similarity analysis for efficient task taxonomy &; transfer learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition; 2019. p. 12387–12396.
14. Conwell C, Prince JS, Alvarez GA, Konkle T. What can 5.17 billion regression fits tell us about artificial models of the human visual system? In: SVRHM 2021 Workshop@ NeurIPS; 2021.
15. Zamir AR, Sax A, Shen W, Guibas LJ, Malik J, Savarese S. Taskonomy: Disentangling task transfer learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 3712–3722.
16. Srinath R, Emonds A, Wang Q, Lempel AA, Dunn-Weiss E, Connor CE, et al. Early emergence of solid shape coding in natural and deep network vision. Current Biology. 2021; 31(1):51–65. https://doi.org/10.1016/j.cub.2020.09.076 PMID: 33096039
17. Denfield GH, Ecker AS, Shinn TJ, Bethge M, Tolias AS. Attentional fluctuations induce shared variability in macaque primary visual cortex. Nature communications. 2018; 9(1):1–14. https://doi.org/10.1038/s41467-018-05123-6 PMID: 29985411
18. Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS computational biology. 2019; 15(4):e1006897. https://doi.org/10.1371/journal.pcbi.1006897 PMID: 31013278
19. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. Imagenet large scale visual recognition challenge. International journal of computer vision. 2015; 115(3):211–252. https://doi.org/10.1007/s11263-015-0816-y
20. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.
21. Yamins DL, DiCarlo JJ. Using goal-driven deep learning models to understand sensory cortex. Nature neuroscience. 2016; 19(3):356–365. https://doi.org/10.1038/nn.4244 PMID: 26906502
22. Lurz K, Bashiri M, Willeke K, Jagadish A, Wang E, Walker E, et al. Generalization in data-driven models of primary visual cortex. In: Ninth International Conference on Learning Representations (ICLR 2021); 2021.
23. Felleman DJ, Van Essen DC. Distributed hierarchical processing in the primate cerebral cortex. Cerebral cortex (New York, NY: 1991). 1991; 1(1):1–47. PMID: 1822724

24. Quiroga RQ, Reddy L, Koch C, Fried I. Decoding visual inputs from multiple neurons in the human temporal lobe. Journal of neurophysiology. 2007; 98(4):1997–2007. https://doi.org/10.1152/jn.00125.2007 PMID: 17671106

25. Cadena SA, Sinz FH, Muhammad T, Froudarakis E, Cobos E, Walker EY, et al. How well do deep neural networks trained on object recognition characterize the mouse visual system? In: Advances in Neural Information Processing (NeurIPS) Neuro-AI Workshop; 2019. Available from: https://openreview.net/forum?id=rkxcXmtUUS.

26. Storrs KR, Kietzmann TC, Walther A, Mehrer J, Kriegeskorte N. Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. Journal of Cognitive Neuroscience. 2021; 33(10):2044–2064. PMID: 34272948

27. Dapello J, Marques T, Schrimpf M, Geiger F, Cox D, DiCarlo JJ. Simulating a primary visual cortex at the front of CNNs improves robustness to image perturbations. Advances in Neural Information Processing Systems. 2020; 33:13073–13087.

28. Steder B, Rusu RB, Konolige K, Burgard W. NARF: 3D range image features for object recognition. In: Workshop on Defining and Solving Realistic Perception Problems in Personal Robotics at the IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS). vol. 44; 2010. p. 2.

29. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. Advances in neural information processing systems. 2012; 25:1097–1105.

30. Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:14091556. 2014;.

31. Kubilius J, Schrimpf M, Nayebi A, Bear D, Yamins DL, DiCarlo JJ. Cornet: Modeling the neural mechanisms of core object recognition. BioRxiv. 2018; p. 408385.

32. Chen T, Kornblith S, Norouzi M, Hinton G. A simple framework for contrastive learning of visual representations. In: International conference on machine learning. PMLR; 2020. p. 1597–1607.

33. Salman H, Ilyas A, Engstrom L, Kapoor A, Madry A. Do adversarially robust imagenet models transfer better? Advances in Neural Information Processing Systems. 2020; 33:3533–3545.

34. Geirhos R, Rubisch P, Michaelis C, Bethge M, Wichmann FA, Brendel W. ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In: International Conference on Learning Representations; 2019. Available from: https://openreview.net/forum?id=Bygh9j09KX.

35. Zhuang C, Yan S, Nayebi A, Schrimpf M, Frank MC, DiCarlo JJ, et al. Unsupervised neural network models of the ventral visual stream. Proceedings of the National Academy of Sciences. 2021; 118(3). https://doi.org/10.1073/pnas.2014196118 PMID: 33431673

36. Kong NC, Margalit E, Gardner JL, Norcia AM. Increasing neural network robustness improves match to macaque V1 eigenspectrum, spatial frequency preference and predictivity. PLOS Computational Biology. 2022; 18(1):e1009739. https://doi.org/10.1371/journal.pcbi.1009739 PMID: 34995280

37. Schrimpf M, Kubilius J, Hong H, Majaj NJ, Rajalingham R, Issa EB, et al. Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? bioRxiv preprint. 2018;.

38. Güçlü U, van Gerven MA. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience. 2015; 35(27):10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015 PMID: 26157000

39. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology. 1962; 160(1):106–154. https://doi.org/10.1113/jphysiol.1962.sp006837 PMID: 14449617

40. Bay H, Ess A, Tuytelaars T, Van Gool L. Speeded-up robust features (SURF). Computer vision and image understanding. 2008; 110(3):346–359. https://doi.org/10.1016/j.cviu.2007.09.014

41. Willmore B, Prenger RJ, Wu MCK, Gallant JL. The berkeley wavelet transform: a biologically inspired orthogonal wavelet transform. Neural computation. 2008; 20(6):1537–1564. https://doi.org/10.1162/neco.2007.05-07-513 PMID: 18194102

42. Carandini M, Heeger DJ. Normalization as a canonical neural computation. Nature Reviews Neuroscience. 2012; 13(1):51–62. https://doi.org/10.1038/nrn3136

43. Burg MF, Cadena SA, Denfield GH, Walker EY, Tolias AS, Bethge M, et al. Learning divisive normalization in primary visual cortex. PLOS Computational Biology. 2021; 17(6):e1009028. https://doi.org/10.1371/journal.pcbi.1009028 PMID: 34097695

44. Ronneberger O, Fischer P, Brox T. U-net: Convolutional networks for biomedical image segmentation. In: Medical Image Computing and Computer-Assisted Intervention–MICCAI 2015: 18th International Conference, Munich, Germany, October 5-9, 2015, Proceedings, Part III 18. Springer; 2015. p. 234–241.

45. Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S. Feature pyramid networks for object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2017. p. 2117–2125.
46. He K, Gkioxari G, Dollár P, Girshick R. Mask r-cnn. In: Proceedings of the IEEE international conference on computer vision; 2017. p. 2961–2969.
47. El-Shamayleh Y, Pasupathy A. Contour curvature as an invariant code for objects in visual area V4. Journal of Neuroscience. 2016; 36(20):5532–5543. https://doi.org/10.1523/JNEUROSCI.4139-15.2016 PMID: 27194333
48. Sanghavi S, Jozwik KM, DiCarlo JJ. SanghaviJozwik2020; 2021. Available from: osf.io/fhy36.
49. Golan T, Raju PC, Kriegeskorte N. Controversial stimuli: Pitting neural networks against each other as models of human cognition. Proceedings of the National Academy of Sciences. 2020; 117(47):29330–29337. https://doi.org/10.1073/pnas.1912334117 PMID: 33229549
50. Burg MF, Zenkel T, Vystrčilová M, Oesterle J, Höfling L, Willeke KF, et al. Most discriminative stimuli for functional cell type identification. The Twelfth International Conference on Learning Representations; 2024. Available from: https://openreview.net/forum?id=9W6KaAcYIr
51. Dwivedi K, Bonner MF, Cichy RM, Roig G. Unveiling functions of the visual cortex using task-specific deep neural networks. PLoS computational biology. 2021; 17(8):e1009267. https://doi.org/10.1371/journal.pcbi.1009267 PMID: 34388161
52. Kornblith S, Norouzi M, Lee H, Hinton G. Similarity of neural network representations revisited. In: International conference on machine learning. PMLR; 2019. p. 3519–3529.
53. Mehrer J, Spoerer CJ, Kriegeskorte N, Kietzmann TC. Individual differences among deep neural network models. Nature communications. 2020; 11(1):5725. https://doi.org/10.1038/s41467-020-19632-w PMID: 33184286
54. Geirhos R, Narayanappa K, Mitzkus B, Bethge M, Wichmann FA, Brendel W. On the surprising similarities between supervised and self-supervised models. arXiv preprint arXiv:201008377. 2020;.
55. Li Z, Brendel W, Walker E, Cobos E, Muhammad T, Reimer J, et al. Learning from brains how to regularize machines. Advances in neural information processing systems. 2019; 32.
56. Safarani S, Nix A, Willeke K, Cadena S, Restivo K, Denfield G, et al. Towards robust vision by multi-task learning on monkey visual cortex. Advances in Neural Information Processing Systems. 2021; 34:739–751.
57. Li Z, Ortega Caro J, Rusak E, Brendel W, Bethge M, Anselmi F, et al. Robust deep learning object recognition models rely on low frequency information in natural images. PLOS Computational Biology. 2023; 19(3):e1010932. https://doi.org/10.1371/journal.pcbi.1010932 PMID: 36972288
58. Geirhos R, Narayanappa K, Mitzkus B, Thieringer T, Bethge M, Wichmann FA, et al. The bittersweet lesson: data-rich models narrow the behavioural gap to human vision. Journal of Vision. 2022; 22 (14):3273–3273. https://doi.org/10.1167/jov.22.14.3273
59. Kietzmann TC, Spoerer CJ, Sörensen LK, Cichy RM, Hauk O, Kriegeskorte N. Recurrence is required to capture the representational dynamics of the human visual system. Proceedings of the National Academy of Sciences. 2019; 116(43):21854–21863. https://doi.org/10.1073/pnas.1905544116 PMID: 31591217
60. Walker EY, Sinz FH, Cobos E, Muhammad T, Froudarakis E, Fahey PG, et al. Inception loops discover what excites neurons most using deep predictive models. Nature neuroscience. 2019; 22(12):2060–2065. https://doi.org/10.1038/s41593-019-0517-x PMID: 31686023
61. Franke K, Willeke KF, Ponder K, Galdamez M, Zhou N, Muhammad T, et al. State-dependent pupil dilation rapidly shifts visual feature selectivity. Nature. 2022; 610(7930):128–134. https://doi.org/10.1038/s41586-022-05270-3 PMID: 36171291
62. Ding Z, Tran DT, Ponder K, Cobos E, Ding Z, Fahey PG, et al. Bipartite invariance in mouse primary visual cortex. bioRxiv. 2023;. https://doi.org/10.1101/2023.03.15.532836 PMID: 36993218
63. Cadena SA, Weis MA, Gatys LA, Bethge M, Ecker AS. Diverse feature visualizations reveal invariances in early layers of deep neural networks. In: Proceedings of the European Conference on Computer Vision (ECCV); 2018. p. 217–232.
64. Calabrese A, Paninski L. Kalman filter mixture model for spike sorting of non-stationary data. Journal of neuroscience methods. 2011; 196(1):159–169. https://doi.org/10.1016/j.jneumeth.2010.12.002 PMID: 21182868
65. Shan KQ, Lubenov EV, Siapas AG. Model-based spike sorting with a mixture of drifting t-distributions. Journal of neuroscience methods. 2017; 288:82–98. https://doi.org/10.1016/j.jneumeth.2017.06.017 PMID: 28652008

66. Ecker AS, Berens P, Cotton RJ, Subramaniyan M, Denfield GH, Cadwell CR, et al. State dependence of noise correlations in macaque primary visual cortex. Neuron. 2014; 82(1):235–248. https://doi.org/10.1016/j.neuron.2014.02.006 PMID: 24698278

67. Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. Imagenet: A large-scale hierarchical image database. In: 2009 IEEE conference on computer vision and pattern recognition. Ieee; 2009. p. 248–255.

68. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International conference on machine learning. PMLR; 2015. p. 448–456.

69. Klindt DA, Ecker AS, Euler T, Bethge M. Neural system identification for large populations separating "what" and "where". arXiv preprint arXiv:171102653. 2017;.

70. Sinz FH, Ecker AS, Fahey PG, Walker EY, Cobos E, Froudarakis E, et al. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. BioRxiv. 2018; p. 452672.

71. Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:151107289. 2015;.

72. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014;.

73. Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on pattern analysis and machine intelligence. 2000; 22(8):888–905. https://doi.org/10.1109/34.868688

74. Zhou B, Lapedriza A, Xiao J, Torralba A, Oliva A. Learning deep features for scene recognition using places database. Advances in neural information processing systems. 2014; 27.

75. Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft coco: Common objects in context. In: European conference on computer vision. Springer; 2014. p. 740–755.

76. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In: Wallach H, Larochelle H, Beygelzimer A, d'Alché-Buc F, Fox E, Garnett R, editors. Advances in Neural Information Processing Systems 32. Curran Associates, Inc.; 2019. p. 8024–8035. Available from: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

77. Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020; 585(7825):357–362. https://doi.org/10.1038/s41586-020-2649-2 PMID: 32939066

78. Van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ. 2014; 2:e453. https://doi.org/10.7717/peerj.453 PMID: 25024921

79. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007; 9 (3):90–95. https://doi.org/10.1109/MCSE.2007.55

80. Waskom M, Botvinnik O, Ostblom J, Gelbart M, Lukauskas S, Hobson P, et al. mwaskom/seaborn: v0. 10.1 (April 2020). zenodo. 2020;.

81. Yatsenko D, Reimer J, Ecker AS, Walker EY, Sinz F, Berens P, et al. DataJoint: managing big scientific data using MATLAB or Python. BioRxiv. 2015; p. 031658.

82. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press; 2016. p. 87–90.

83. Merkel D. Docker: lightweight linux containers for consistent development and deployment. Linux journal. 2014; 2014(239):2.

A.4 Learning divisive normalization in primary visual cortex

Burg, M. F., **Cadena, Santiago A**, Denfield, G. H., Walker, E. Y., Tolias, A. S., Bethge, M., & Ecker, A. S. (2021). Learning divisive normalization in primary visual cortex. *PLOS Computational Biology*, *17*(6), e1009028

Abstract

Divisive normalization (DN) is a prominent computational building block in the brain that has been proposed as a canonical cortical operation. Numerous experimental studies have verified its importance for capturing nonlinear neural response properties to simple, artificial stimuli, and computational studies suggest that DN is also an important component for processing natural stimuli. However, we lack quantitative models of DN that are directly informed by measurements of spiking responses in the brain and applicable to arbitrary stimuli. Here, we propose a DN model that is applicable to arbitrary input images. We test its ability to predict how neurons in macaque primary visual cortex (V1) respond to natural images, with a focus on nonlinear response properties within the classical receptive field. Our model consists of one layer of subunits followed by learned orientation-specific DN. It outperforms linear-nonlinear and wavelet-based feature representations and makes a significant step towards the performance of state-of-the-art convolutional neural network (CNN) models. Unlike deep CNNs, our compact DN model offers a direct interpretation of the nature of normalization. By inspecting the learned normalization pool of our model, we gained insights into a long-standing question about the tuning properties of DN that update the current textbook description: we found that within the receptive field oriented features were normalized preferentially by features with similar orientation rather than non-specifically as currently assumed.

Author contributions

Conceptualization: MFB, **SC**, AE, AT, MB. Data Curation: MFB, **SC**, GD, EW. Formal Analysis: MFB, **SC**, AE. Funding Acquisition: AT, AE, MB. Investigation: MFB, **SC**, AE. Methodology: MFB, **SC**, AT, MB, AE. Project Administration: MFB, AE, AT, MB. Resources: **SC**, GD, EW, AT, MB, AE. Software: MFB, **SC**, GD, EW, AE. Supervision: AE, AT, MB. Validation: MFB, **SC**. Visualization: MFB, **SC**, AE. Writing – Original Draft Preparation: MFB, **SC**. Writing – Review and Editing: MFB, **SC**, GD, EW, AT, MB, AE.

PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

Learning divisive normalization in primary visual cortex

Max F. Burg[1,2,6]*, **Santiago A. Cadena**[1,2,3], **George H. Denfield**[3,4], **Edgar Y. Walker**[3,4], **Andreas S. Tolias**[2,3,4,5☯], **Matthias Bethge**[1,2,3☯], **Alexander S. Ecker**[6,7☯]

1 Institute for Theoretical Physics and Centre for Integrative Neuroscience, University of Tübingen, Tübingen, Germany, **2** Bernstein Center for Computational Neuroscience, Tübingen, Germany, **3** Center for Neuroscience and Artificial Intelligence, Baylor College of Medicine, Houston, Texas, United States of America, **4** Department of Neuroscience, Baylor College of Medicine, Houston, Texas, United States of America, **5** Department of Electrical and Computer Engineering, Rice University, Houston, Texas, United States of America, **6** Institute of Computer Science and Campus Institute Data Science, University of Göttingen, Göttingen, Germany, **7** Max Planck Institute for Dynamics and Self-Organization, Göttingen, Germany

☯ These authors contributed equally to this work.
* max.burg@bethgelab.org

Citation: Burg MF, Cadena SA, Denfield GH, Walker EY, Tolias AS, Bethge M, et al. (2021) Learning divisive normalization in primary visual cortex. PLoS Comput Biol 17(6): e1009028. https://doi.org/10.1371/journal.pcbi.1009028

Editor: Blake A. Richards, McGill University, CANADA

Received: July 24, 2020

Accepted: April 30, 2021

Published: June 7, 2021

Data Availability Statement: The data analyzed in this study is available in a GIN repository at https://doid.gin.g-node.org/2e31e304e03d6357c98ac735a1fe5788/ (DOI: 10.12751/g-node.2e31e3). Code is available from https://eckerlab.org/code/burg2021_learning_divisive_normalization.

Funding: The research was supported by the German Federal Ministry of Education and Research (BMBF) via the Competence Center for Machine Learning (FKZ 01IS18039A) [https://tuebingen.ai]; the Deutsche

Abstract

Divisive normalization (DN) is a prominent computational building block in the brain that has been proposed as a canonical cortical operation. Numerous experimental studies have verified its importance for capturing nonlinear neural response properties to simple, artificial stimuli, and computational studies suggest that DN is also an important component for processing natural stimuli. However, we lack quantitative models of DN that are directly informed by measurements of spiking responses in the brain and applicable to arbitrary stimuli. Here, we propose a DN model that is applicable to arbitrary input images. We test its ability to predict how neurons in macaque primary visual cortex (V1) respond to natural images, with a focus on nonlinear response properties within the classical receptive field. Our model consists of one layer of subunits followed by learned orientation-specific DN. It outperforms linear-nonlinear and wavelet-based feature representations and makes a significant step towards the performance of state-of-the-art convolutional neural network (CNN) models. Unlike deep CNNs, our compact DN model offers a direct interpretation of the nature of normalization. By inspecting the learned normalization pool of our model, we gained insights into a long-standing question about the tuning properties of DN that update the current textbook description: we found that within the receptive field oriented features were normalized preferentially by features with similar orientation rather than non-specifically as currently assumed.

Author summary

Divisive normalization (DN) is a computational building block throughout sensory processing in the brain. We currently lack an understanding of what role this normalization mechanism plays when processing complex stimuli like natural images. Here, we use

Forschungsgemeinschaft (DFG, German Research Foundation) via grant EC 479/1-1 (A.S.E.) [http://www.dfg.de/], the Collaborative Research Centers SFB 1233 (Robust Vision, project number 276693517) [https://uni-tuebingen.de/en/research/core-research/collaborative-research-centers/crc-1233/] and SFB 1456 (Mathematics of Experiment, project number 432680300) [https://www.uni-goettingen.de/en/628179.html], and the Cluster of Excellence "Machine Learning – New Perspectives for Science" (EXC 2064/1, project number 390727645) [http://www.ml-in-science.uni-tuebingen.de/]; the National Eye Institute of the National Institutes of Health under Award Numbers R01EY026927 (A.S.T.) and DP1 EY023176 (A.S.T.) [https://nei.nih.gov/], and NIH-Pioneer award DP1-OD008301 (A.S.T) [https://commonfund.nih.gov/pioneer/]. This research was also supported by NEI/NIH Core Grant for Vision Research (EY-002520-37) [https://nei.nih.gov/], NEI training grant T32EY00700140 (G.H.D) and F30EY025510 (E.Y.W.) [https://nei.nih.gov/], DARPA grant N66001-17-C-4002 [https://www.darpa.mil], and the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003 [https://www.iarpa.gov]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

modern machine learning methods to build a general DN model that is directly informed by data from primary visual cortex (V1). Contrary to high-predictive deep learning models, our DN-based model's parameters offer a straightforward interpretation of the nature of normalization. Within the receptive field, we found that neurons responding strongly to a specific orientation are preferentially normalized by other neurons that are highly active for similar orientations, rather than being normalized by all neurons as it is currently assumed by textbook models.

Introduction

A crucial step towards understanding the visual system is to build models that predict neural responses to arbitrary stimuli with high accuracy [1]. The classical standard models of primary visual cortex (V1) are based on linear-nonlinear models [2], energy models [3] and subunit (LN-LN) models [4–9]. Fueled by advances in machine learning technology, recent studies have shown that multi-layer convolutional neural networks (CNNs) can significantly improve the prediction of neural responses to complex images and videos at several stages of the visual pathway, outperforming classical models [10–17]. The current state-of-the-art data-driven model of single-unit activity in monkey V1 is a three-layer CNN [14]. However, it is challenging to gain insights into V1 function from the features produced by deeper layers of such models. In particular, we do not have first principles explaining the kind of nonlinearities approximated in successive layers of CNNs, or if these nonlinearities can be described in a compact way in the first place.

A promising candidate to facilitate a more succinct description of V1 neurons is to replace the computations from the CNN's deeper layers by a divisive normalization nonlinearity [18] that is easier to understand. Divisive normalization has been proposed as a canonical neural computation present throughout the visual pathway [19] because it (1) explains a wide variety of neurophysiological phenomena using simple stimuli [19, 20] and (2) can be derived from first principles of redundancy reduction [21, 22]. The significance of DN is also supported by its recent success in computer vision where it enables state-of-the-art natural image compression with high perceptual quality [23].

Divisive normalization has been invoked to explain several nonlinear neurophysiological phenomena in V1, which can be classified into (1) phenomena that are restricted to the receptive field (RF) only or (2) phenomena that involve an interaction between receptive field and its surround. A prominent example for a mechanism restricted to the receptive field is cross-orientation inhibition [18, 24–29], whereas an example for an interaction of the RF with its spatially adjacent surround is surround suppression [30–34]. In our work, we focus on the first aspect: divisive normalization *within* the RF and cross-orientation inhibition. In this phenomenon, the response of a neuron to a driving grating stimulus in the RF is suppressed by superimposing a second grating also within the RF that does not elicit a response when presented alone: for instance, a grating with orientation orthogonal to the neuron's preferred orientation.

The basic idea of divisive normalization (Fig 1A) is that a neuron's driving input is normalized divisively by a weighted sum over nearby neurons' responses [18, 19]. While the general idea is simple, elegant and powerful, our current knowledge of DN is limited in two important ways: (1) DN has been studied mostly with simple, synthetic stimuli, but its implications under natural image stimulation are unclear. For example we do not know whether incorporating DN into system identification models predicting neural responses to natural stimuli improves

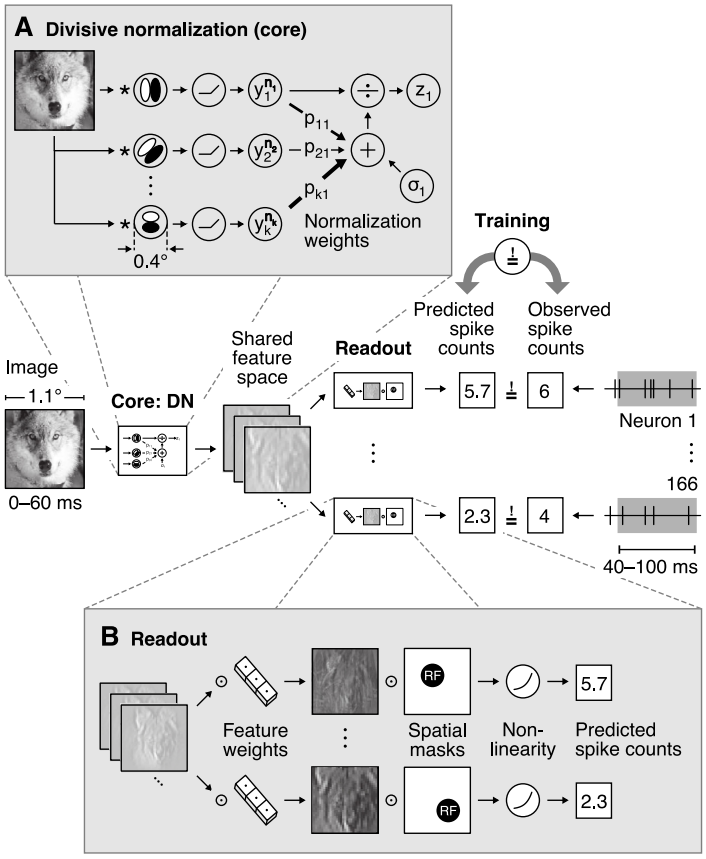


Fig 1. Overview of our divisive normalization (DN) model. The model takes as input an image covering 1.1˚ of visual field and predicts neurons' spike counts in response to this image (details in Fig 2). The model is split into two parts: a *core* that computes a shared nonlinear feature space and a *readout* that maps the shared feature space individually to each neuron's spike count. **A**. Divisive normalization mechanism (simplified). The visual input is convolved with 32 filters covering 0.4˚ of visual field and then rectified and exponentiated to produce an excitatory output. The output of each filter is then divided by a weighted sum of the excitatory outputs of all filters with normalization weights p_{kl} and a semi-saturation constant σ_l. In our general formulation, all weights and constants are learned from the data. **B**. Readout that maps the shared feature space to each neuron's spike count through an individual weighted sum over the entire shared feature space and a pointwise output nonlinearity. The readout weights are factorized into a feature vector—capturing the nonlinear feature(s) that a neuron computes—and a spatial mask—localizing each neuron's receptive field (RF).

https://doi.org/10.1371/journal.pcbi.1009028.g001

performance. (2) We currently do not know how response properties of neurons with overlapping receptive fields determine whether a neuron contributes to the normalization pool of another neuron and, if so, with which strength [27].

To explain normalization phenomena within the receptive field like cross-orientation inhibition, current models of divisive normalization assume that all nearby neurons with diverse orientation tuning preferences and with similar receptive field locations contribute equally to the normalization pool [18, 27, 29]. However, some earlier experimental studies suggest that this assumption may not be correct for some neurons [24, 26], and normative models of

normalization predict that the magnitude with which a given neuron contributes to another neuron's normalization should depend on the relationship of their response properties [21].

In this paper, we address the two main questions raised above: (1) How well does a data-driven model with divisive normalization predict V1 responses to natural images, compared to classical subunit models and deep convolutional neural networks, and (2) what stimulus features are learned for normalizing a V1 neuron's response in relation to its favoured feature selectivity within the receptive field?

We focused on effects restricted to the receptive field and on models that only account for local normalization interactions between neurons with overlapping receptive fields. We developed an end-to-end trainable model predicting V1 spike counts from natural stimuli, replacing the CNNs deeper layers by the canonical neural DN computation [19] and learning the filters of all neurons as well as their normalization weights directly from the data. We also explored how our DN model could be extended in a way that might enable it to capture surround interactions from outside of the RF, however, our control experiments demonstrate that our results are very unlikely to include the extra-classical surround.

We fitted our DN model to monkey V1 responses to natural images and found that it outperforms linear-nonlinear and subunit models, suggesting that DN mechanisms are useful for capturing nonlinear computations unlocked by natural stimuli beyond those evoked by simpler gratings. The power of our model was further strengthened as it captured cross-orientation inhibition within the receptive field, brought to light via *in silico* experiments. In contrast to the three-layer CNN, which also reproduced this phenomenon, the single divisive layer of the DN model provides direct interpretations of the normalization pool. These revealed that oriented features were preferentially normalized by other features with similar orientation, which quantitatively improved predictive performance over a model with nonspecific normalization. However, the three-layer CNN outperformed our DN models, suggesting that there are additional non-linear mechanisms beyond the ones captured by our DN model (e. g. phase invariance and divisive effects), that are required to achieve optimal performance in modeling V1 responses. Thus, our work advances our understanding of V1 function, highlighting the importance of orientation-specific gain control to predict responses to natural images, and shows how embedding theories of neural processing into the architectures of generic data-driven models can improve our understanding of sensory processing in the brain.

Results

Learning divisive normalization

The basic idea of divisive normalization (Fig 1A) is that the response of neuron l

$$z_l(x) = \frac{y_l^{n_l}(x)}{\sigma_l^{n_l} + \sum_{k \in \mathcal{K}} p_{kl} \cdot y_k^{n_k}(x)} \tag{1}$$

is given by its driving input activity $y_l(x)$ exponentiated by n_l and divisively normalized by a weighted sum over nearby neurons' exponentiated responses $y_k^{n_k}(x)$ [18, 19], where x represents the stimulus, σ_l is a semi-saturation constant, and all quantities are non-negative. Here, the set of normalizing neurons $\mathcal{K}$ and the normalization weights p_{kl} define which neurons k contribute to the normalization pool of neuron l and with what strength, respectively.

While this formulation is straightforward to write down, it is challenging to build quantitative models based on it that are applicable to arbitrary inputs. The denominator depends on a potentially large population of neurons—which is unknown in general—and the structure of the normalization weights has been studied only using very restricted sets of simple stimuli such as oriented gratings and bars. Previous modeling work on divisive normalization has

therefore made specific assumptions about the filter properties of the underlying neuronal population and either modeled only a closed set of stimuli such as gratings of different orientation [18, 27–29, 35] or evaluated models only qualitatively [21, 36, 37].

We developed a general, image-computable predictive model of divisive normalization following Eq (1), which is applicable to arbitrary images and whose parameters are learned by optimizing the accuracy of the model in predicting the spiking activity of a large number of neurons in response to natural images (see Fig 1). Our model builds on a recent innovation in predictive modeling [12, 14, 38–40], jointly modeling all recorded neurons instead of learning a predictive model for each neuron individually. Because many neurons perform similar computations—up to shifts in receptive field location—jointly modeling them makes more efficient use of the data and we can learn more complex models. The basic idea is to split the model into two parts (Fig 1): (A) a *core* that transforms the input image into nonlinear features shared among all neurons, and (B) a *readout* that linearly combines the features to produce a prediction of each neuron's response.

We use a convolutional network for the core, whose architecture lends itself very well to model divisive normalization. By construction, we have a model that contains all filters necessary to account for the recorded neurons' responses. All of these filter responses are automatically extracted at each location, providing a good approximation of the underlying population of neurons in the brain although it is only sparsely sampled during the experiment. As a consequence, we can optimize the pool of neurons providing normalizing inputs and their corresponding weights p_{kl} (Eq 1) to account for the neural responses. In this work we mostly focus on normalization effects from neurons with overlapping receptive fields, rather than investigating interactions with the RFs' surround.

In summary, our model's core (Fig 1A) consists of a set of convolutional filters w_l and biases o_l (we use 32) that provide the driving inputs y_l from an image x at all locations in the neurons' RFs,

$$y_l = \max(0,\ w_l * x + o_l)\ . \tag{2}$$

As a specific feature is extracted by its according convolutional filter at each spatial location of the image, the driving responses y_l form a map of responses for the according features, hence named feature map. This operation is followed by a DN stage (Eq 1) in which all operations are performed element-wise across space. This core is shared among all neurons and converts the image into a set of feature maps z_l containing information from the neurons' RFs. We mapped those normalized feature maps into response predictions $\hat{r}$ by a linear readout step (Fig 1B) that picks the relevant features by readout weights b and spatial locations by readout weights a for each neuron i, followed by a pointwise output nonlinearity ϕ,

$$\hat{r}_i = \phi_i\Big(\sum_{u,v,l}(a_{uv,i}\, b_{l,i})\, z_{uvl}\Big)\ , \tag{3}$$

where u, v index space (see Methods for additional details). To ensure that the readout does not model any inhibitory or complex interactions, we constrained its weights a, b to be non-negative and added a sparsity prior. The non-negativity ensures that activations can only add, preventing the readout stage from accounting for any suppressive effects. The readout can, however, account for response invariances such as phase invariance of complex cells; see Methods for an in-depth explanation. While our model reflects the general formulation of divisive normalization, mathematically it is not identical to the classical formulation of DN due to the additional linear-nonlinear readout after the model's core.

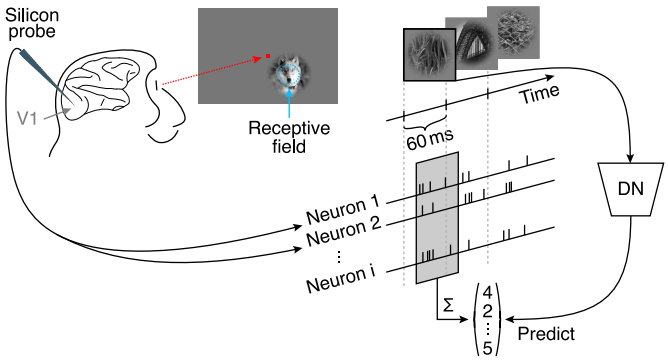


Fig 2. Experimental paradigm from Cadena and colleagues [14]. Natural images were flashed to a monkey covering 2˚ of their visual angle, and located at the center of the multi-unit receptive field. Multiple neurons were isolated from recordings with silicon probes inserted into V1 [41]. Natural images were shown in a fast sequence without blanks, each presented for 60 ms. Spike counts from all isolated neurons corresponding to each image were extracted from a window 40 ms after the image onset lasting 60 ms.

https://doi.org/10.1371/journal.pcbi.1009028.g002

DN model outperforms subunit model, half-way closing the gap to state-of-the-art performance

We fitted the model described above to a dataset of 166 neurons recorded in V1 of two awake, fixating monkeys, who viewed a fast sequence of localized natural images and textures (Fig 2; data from Cadena and colleagues [14]). The stimuli were placed at the neurons' receptive fields and primarily stimulated the center (i. e. the RF) as they covered one degree of visual field, quickly fading off reaching zero contrast at two degrees (see Discussion and Methods for additional details). Images were shown for 60 ms each, without blanks in between. Single unit activity was recorded with laminar silicon probes sampling from all cortical layers. We fit the model jointly to the responses of all neurons. As neurons were recorded in 17 recording sessions, the dataset sampled a diverse range of preferred orientations. The objective function during training was to minimize the difference between the model's prediction and the observed spike counts of the neurons in a time window 40–100 ms after image onset (to account for response latency).

To evaluate model performance, we estimated the fraction of explainable variance explained (FEV), which quantifies the fraction of the stimulus-driven response variance that is accounted for by the model, and ignores unexplainable trial-to-trial variability in the response of the neurons (see Methods). A perfect model would reach a FEV of 100%. Since model performance depends on the parameters' initial values before optimization, we initialized all weights randomly before model training and repeated this procedure multiple times. To ensure that the reported performance is not a spurious characteristic of one particular model, we picked the best ten models of a model class (assessed in terms of validation set accuracy) and report the mean performance together with its 95% confidence interval. We perform this analysis for each of the four model classes.

Subunit models are an established approach to model primary visual cortex responses [4–9]. In addition to capturing a fair portion of the explainable variance [14], their computations are easily understandable because they are rather shallow models consisting of a first stage of rectified linear filtering followed by a static nonlinearity, and then a pooling stage with a subsequent output nonlinearity. Although the final output nonlinearity is capable to learn various pointwise functions like sigmoids, in most cases (across all model types) it leads to only slight

expansive deviations of the identity function. Structurally, the subunit model is the same as our DN model, but without the normalization stage. Therefore, we considered a convolutional subunit model as a strong baseline for our model. This subunit model accounted for (46.6 ± 0.2)% FEV (mean and 95% confidence interval over the ten best models selected in terms of validation set accuracy). In comparison, a regularized linear nonlinear Poisson model (LNP) only accounted for 16.3% FEV on the same dataset [14] due to its inability to model complex cells.

As recent developments in machine learning technology have allowed us to improve predictive performance, we used the current best data-driven model as a gold standard. This model is a convolutional neural network with three convolutional layers and a linear-nonlinear readout [14]. To enable fair comparison of this model to all other models in this study, we optimized the CNN comparable to our DN model, reaching (52.3 ± 0.1)% FEV (mean and 95% confidence interval over the ten best models selected in terms of validation set accuracy; outperforming previously reported performance of 49.8% FEV [14]). However, although this CNN model outperforms the simpler subunit model, it is of higher complexity requiring a higher number of parameters.

To evaluate how well our DN model accounts for the data, we placed its performance on a scale between 0% (baseline: subunit model) and 100% (gold standard: CNN). On that scale, our DN model achieved a score of (52 ± 3)% between the baseline and gold standard (Fig 3; (49.6 ± 0.1)% FEV; mean and 95% confidence interval over the ten best models selected in terms of validation set accuracy). The performance differences between all model pairs were statistically significant (pairwise Wilcoxon signed rank test on best models in terms of validation accuracy: $p < 0.024$, $N = 166$ neurons, family-wise error rate $\alpha = 0.05$ using Holm-Bonferroni correction to account for multiple comparisons). Notably, we achieved this performance gain by simply adding the trainable DN stage to the convolutional subunit model, which

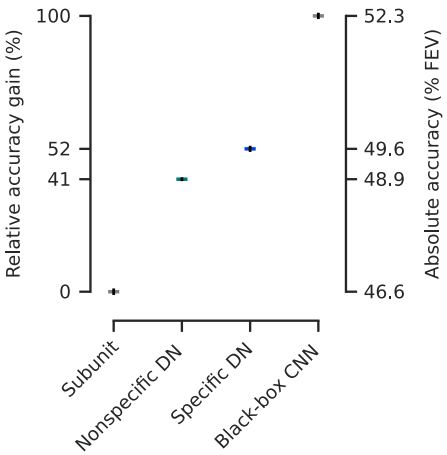


Fig 3. Performance comparison of our models fitted to the data from Cadena and colleagues [14] relative to the gap between the best shallow model—a subunit one layer convolutional neural network (CNN)—and the deeper data-driven state-of-the-art three-layer CNN [14]. Non-specific divisive normalization (DN) accounts for 41% of this gap, while specific DN improves it up to 52%. Absolute values in terms of percentage of explainable variance explained (FEV) on the right (mean over the ten best models selected in terms of validation set accuracy, see main text for details). Error-bars (black) indicate the standard error of the mean. Model performance is significantly different between each model class (pairwise Wilcoxon signed rank test on best models in terms of validation accuracy: $p < 0.024$, $N = 166$ neurons, family-wise error rate $\alpha = 0.05$ using Holm-Bonferroni correction).

https://doi.org/10.1371/journal.pcbi.1009028.g003

quantifies the importance of divisive normalization as computational mechanism to predict neural responses in V1 under stimulation with complex, natural images. At the same time, the CNN outperformed the DN model, suggesting that further non-linear dependencies in addition to phase invariance (complex cells) and divisive normalization are required to reach optimal predictive performance.

Divisive normalization and CNN models learn cross-orientation inhibition

We wanted to test if our DN model captured non-linear interactions between the neurons inside the receptive field leading to cross-orientation inhibition, a phenomenon that was explained by DN within the RF before [18, 24–29]. As the gold standard three-layer convolutional neural network explains more variance than the divisive normalization model, we asked if the CNN captured cross-orientation inhibition, too. To assess if our models learned cross-orientation inhibition, we performed the experiments done before by experimentalists *in silico*. Specifically, for each unit we determined the optimal Gabor maximizing our models' predictions. Subsequently, we combined the optimal Gabor by adding the same Gabor pattern with orthogonal orientation, obtaining a plaid stimulus (Fig 4A insets). We then presented the plaids to our models varying the contrasts of each Gabor component and measured the resulting tuning curve for all models *in silico*, averaging across the orthogonal Gabor's phase to get independent of any phase dependency. For the DN model and the CNN, we found cross-orientation inhibition in a number of cells: the response elicited by presenting the optimal Gabor was inhibited by increasing the contrast of the orthogonal Gabor component (Fig 4A). However, such behavior was not clearly visible for the subunit model.

To quantify the models' capability to learn cross-orientation inhibition, we defined a cross-orientation inhibition index measuring the percentage an orthogonal Gabor inhibits response when added to an optimal Gabor presented alone (Fig 4A left; see Methods for details). To determine how many cells are cross-orientation inhibited by at least 10%, we assess the highest cross-orientation inhibition index across all contrast combinations of the two component Gabors. To ensure we report a general property of all models, we quantified cross-orientation inhibition across the ten best models of each model-type (assessed in terms of performance on the validation set, only differing in the initialization of the model parameters before training). In the subunit model we found no significant cross-orientation inhibition (Fig 4B). Although the 3% FEV performance gain of the DN model over the subunit model might seem small, the phenomenological difference between those models was quite substantial, as in the DN model half of the cells showed cross-orientation inhibition. The gold-standard three-layer CNN captured cross-orientation inhibition in fewer cells, likely approximating divisive normalization through its deeper layers as its filters in the first convolutional layer are qualitatively similar to the DN model, showing Gabor-like and blob patterns. At the same time, the CNN seems to capture additional effects from the data that are not explained by DN in the RF center, leading to the CNN's higher performance.

Normalization within the receptive field is feature-specific

Having established that both the DN and the CNN models learned a phenomenon explained by divisive normalization within the RF, namely cross-orientation inhibition, we next asked which features and how strongly they contribute to the normalization pool. In the CNN model there is no explicit notion of a normalization pool that we could comprehensively analyze. On the other hand, our DN model offers a way to investigate the structure of the normalizing input within the receptive field, i. e. the image-averaged products $\langle p_{kl} \cdot y_k^{n_k}(x) \rangle_x$ in the denominator of Eq (1), to quantify the strength by which different features contribute to it (see

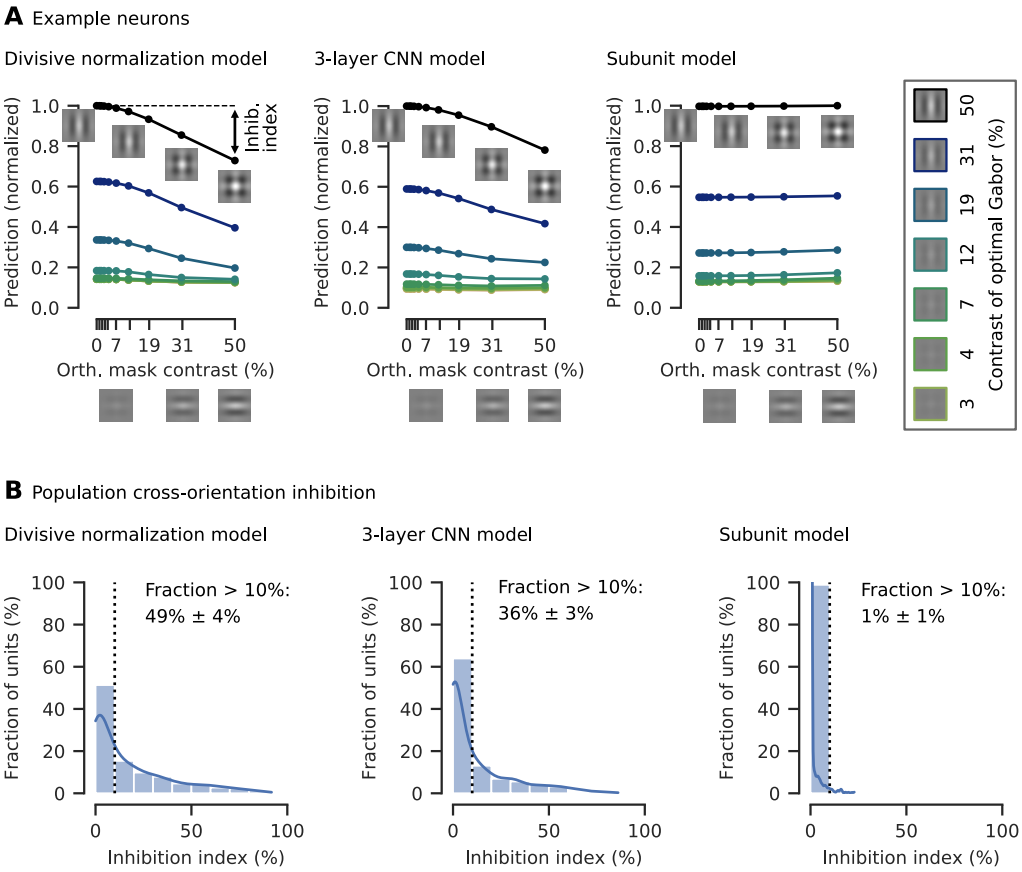


Fig 4. Cross-orientation inhibition was learned by our DN model and the three-layer CNN, but not significantly by the subunit model. **A**. Tuning curves for an example neuron of all three models and various contrast combinations of the optimal Gabor (box on the right, examples for contrasts of 0%, 1% and 2% not shown) and an orthogonal Gabor masking. As the contrast of the orthogonal mask increases, the model prediction (normalized by the maximum response) decreases. The cross-orientation inhibition (inhib.) index measures the percentage of response inhibition by adding the masking compared to the optimal Gabor presented alone, in this case approximately 20%. **A**. **insets**: Illustration of plaid stimuli, created by overlaying an optimally oriented Gabor with an orthogonal mask. **B**. Histograms of the cross-orientation inhibition indices accumulated across the best ten models (in terms of validation set accuracy) per model type, with kernel density estimate of the underlying distribution. The fraction of cells that show more than 10% cross-orientation inhibition is displayed right of the dotted line (mean and 95% confidence interval over the ten best models selected in terms of validation set accuracy). For the DN model, more cells show cross-orientation inhibition compared to the other models. The subunit model shows almost no cross-orientation inhibition.

https://doi.org/10.1371/journal.pcbi.1009028.g004

Methods for details). When visualizing the feed-forward features learned by the DN model we found that several exhibited clear orientation selectivity (i. e. features with an anisotropic structure; see Fig 5A). Thus, for these features we first wondered about the relationship between their preferred orientation and the orientation of the features by which they are preferentially normalized. To this end, we visualized the feature pair-wise normalizing input (Fig 5A) and found that oriented features are normalized preferentially by features with similar orientation preference, while orthogonal features seem to contribute less.

First, we coarsely quantified the difference with which these two groups contribute to normalization. We split the sum in Eq (1) into two parts and collected the normalizing input of

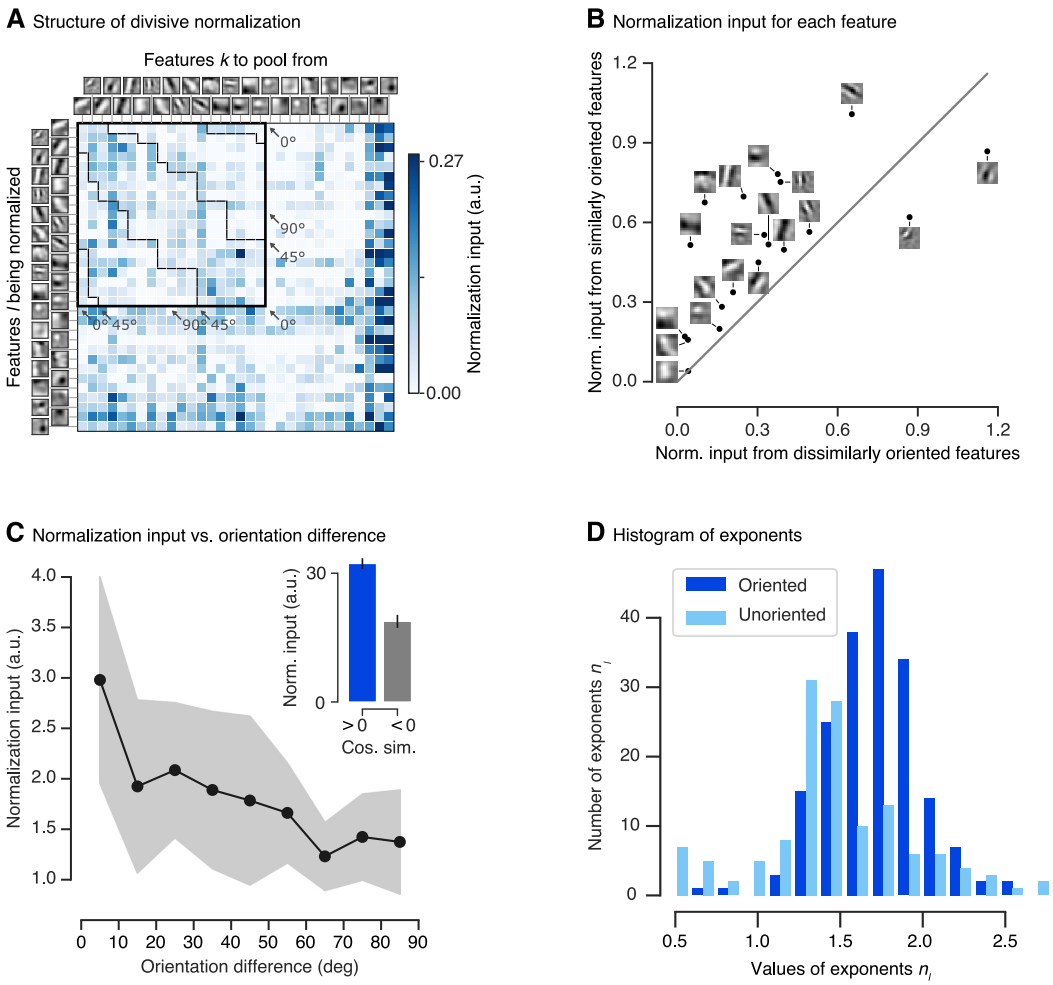


Fig 5. Structure of divisive normalization. A. The matrix shows the average strength of the normalizing inputs (products $\langle p_{kl} y_k^{n_k}(x) \rangle_x$ in denominator of Eq (1) averaged across images; see Methods) for each combination of filter response being normalized (rows) and filter response providing normalizing input (columns). Darker shades of blue indicate stronger normalization. Orientation-selective filters are grouped at the top, ordered by preferred orientation and marked by the black square. The dashed black lines within the square separate pairs of filters with similar ($< 45^\circ$) and dissimilar ($\leq 45^\circ$) orientations. Normalizing inputs are stronger for similarly tuned filters. Unoriented filters mainly accounting for orientation-unspecific contrast are sorted by total normalization input. Darkest blue color corresponds to the maximum normalization input for the group of oriented filters, higher normalization input values for the unoriented filters are clipped to that value. Data of the model with highest accuracy on the validation set is shown. **B**. Normalization input from similar orientations ($< 45^\circ$) compared to the normalization input from dissimilar orientations ($\geq 45^\circ$) for each oriented linear filter. Grey line: identity. Most features are normalized preferentially by the responses of filters with similar preferred orientations. Data of the model with highest accuracy on the validation set is shown. **C**. Normalization input, binned into orientation difference of 10°. Each bin was averaged over the top-10 models (assessed on the validation set). The shaded area depicts the standard deviation per bin. **C inset**. Normalization input (norm. input) vs. cosine similarity between linear filters (cos. sim.) averaged across the top-10 models (assessed on the validation set). A cosine similarity greater than zero corresponds to similar features. Error bars: standard error of the mean. **D**. Histogram of DN exponents (n_l in Eq 1) of the ten best performing models in terms of validation set accuracy. Darker/lighter color: exponents corresponding to driving inputs due to oriented/unoriented linear filters. Most values are larger than one, with a few exceptions mainly corresponding to unoriented filters.

https://doi.org/10.1371/journal.pcbi.1009028.g005

features with similar orientation as the driving feature ($< 45^\circ$) and that of features with dissimilar orientations ($\geq 45^\circ$). By analyzing the normalization of each oriented feature individually, we found that most oriented features are more strongly normalized by features with similar orientations (Fig 5B). To assess whether our qualitative observation above is a general property of the data or a spurious characteristic of the best model chosen, we repeated this analysis for the top-10 performing models (assessed in terms of accuracy on the validation set, only differing in the initialization of the model parameters before training) and observed a similar behavior. Summing up the normalizing input across the features, we found that, for all of these models, similar orientations contributed more strongly than dissimilar orientations. Taking the data of all top-10 models into account, we found that similarly oriented features contributed (1.5 ± 0.1)-fold more normalizing input than dissimilar features (mean and 95% confidence interval over the ten best models selected in terms of validation set accuracy; Wilcoxon signed rank test: $p < 0.006$, $N = 10$ models; Cohen's $d = 1.2$).

We further quantified the orientation-specific nature of normalization at a more fine-grained level. Instead of lumping orientation differences into two groups as before, we split up the normalizing inputs into nine bins of 10° width each and averaged those bins across the top-10 models. This analysis revealed that the strength of the normalizing inputs decreased as the difference in orientation increased (Fig 5C). Hence, the more similar a normalizing feature's orientation was to the feature to be normalized, the stronger was its contribution to normalization (linear regression analysis on all normalization input vs. orientation difference pairs of best model in terms of validation accuracy: $p < 10^{-6}$). In fact, features in the group most similar to the driving input contributed (2.3 ± 0.4)-fold more than those in the orthogonal group (mean and 95% confidence interval over the ten best models selected in terms of validation set accuracy; Wilcoxon signed rank test: $p < 0.006$, $N = 10$ models; Cohen's $d = 1.9$).

An immediate question that follows from our previous findings relates to the contribution of more general feature properties beyond orientation selectivity. Is the normalization input higher only for similar orientation, or is normalization input generally higher for similar filters, without pre-specifying a certain property like orientation? To answer that question, we split the input to normalization in terms of cosine similarity between all oriented and unoriented filters into a group of similar (cosine similarity > 0) and dissimilar (cosine similarity < 0) features. Summing up the normalizing input across the features and averaging across the best ten models, we found that in general similar features contributed (1.9 ± 0.5)-fold more normalizing input than dissimilar features (Fig 5C inset; mean and 95% confidence interval over the ten best models selected in terms of validation set accuracy; Wilcoxon signed rank test: $p < 0.006$, $N = 10$ models; Cohen's $d = 2.9$).

Analysing the exponents of our divisive normalization model, we found a mean value of 1.61 (averaged across the best ten models on the validation set) with most values being greater than one (Fig 5D), which previously has been connected to an intensified winner-take-all behavior of responses [19, 29]. In a few exceptions exponents learned by our model are smaller than one, mainly corresponding to driving inputs from unoriented filters.

Control: Nonspecific divisive normalization reduces accuracy. To determine how important orientation-specific normalization is, we performed a control experiment: For each feature l being normalized, we constrained all of its incoming normalization weights p_{kl} to be identical. This model is an instantiation of the previous models [18, 27, 29] assuming non-specific normalization from all features. Note that, since the feature orientations were non-uniformly distributed (Fig 5A), this model's normalization was weakly orientation tuned despite equal weights for all features (orientation difference $< 45^\circ$ contributed 1.1-fold stronger than $\geq 45^\circ$; Wilcoxon signed rank test; $p < 0.007$, $N = 10$ models; Cohen's $d = 0.3$). This model achieved a performance of (41 ± 2)% between the baseline and gold standard

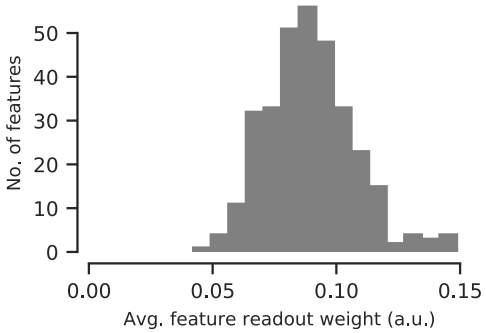


Fig 6. Histogram of feature readout weights of the ten best performing models in terms of validation set accuracy. For each model, feature weights are normalized across channels and averaged across individual neurons. All model's channels are used to predict neural activity.

https://doi.org/10.1371/journal.pcbi.1009028.g006

((48.9 ± 0.1)% FEV, mean and 95% confidence interval over the ten best models selected in terms of validation set accuracy). The unspecific DN model was included in the hypothesis test showing statistically significant performance differences between all model types (pairwise Wilcoxon signed rank test on best models in terms of validation accuracy: $p < 0.024$, $N = 166$ neurons, family-wise error rate $\alpha = 0.05$ using Holm-Bonferroni correction). While the unspecific control model does not match the performance of our more general DN model, it does outperform the subunit baseline. Thus, stronger orientation-specific normalization is necessary to achieve the full DN model performance, but a simpler form of uniform divisive normalization can account for a large fraction of the improvement over the subunit baseline.

Control: All channels contribute to our model's prediction. One potential caveat of our analyses so far is that we analyzed the orientation specificity of DN in terms of the convolutional feature maps in our model's core rather than the actual neurons we recorded. These features provide a much more compact view of the population of neurons, because they are invariant to the receptive field locations and the neural responses are simple linear combinations of those features. Moreover, the model's core offers a direct interpretation of its normalization weights. However, it is not clear a-priori whether all features are equally important for predicting the activity of the neurons in our population. Thus, considering convolutional features instead of actual neurons may lead to a skewed view of the population. To verify that this is not an issue, we quantified how much each feature contributed to the overall activity of all neurons by normalizing the feature readout weights across channels and averaging across neurons. The resulting distribution (Fig 6) containing these averaged feature readout weights for the best ten models had a coefficient of variation of 0.2. We therefore concluded that all features were read out by roughly the same number of neurons and hence were similarly important to predict neural activity. Thus, our interpretation of orientation-specific normalization is unlikely to be an artifact of analyzing the convolutional features rather than the actual neurons.

Control: Surround interactions are unlikely driving orientation-specific normalization in the RF center. We observed orientation-specific divisive normalization in the receptive field. Surround suppression is known to be orientation-specific as well [30–34], so a potential concern would be that some of the extra-classical surround of a unit's RF contributed to the results presented above. To rule out this possibility, we explicitly verified *in silico* that our model learning DN inside the RF learned no significant surround-suppression. For each

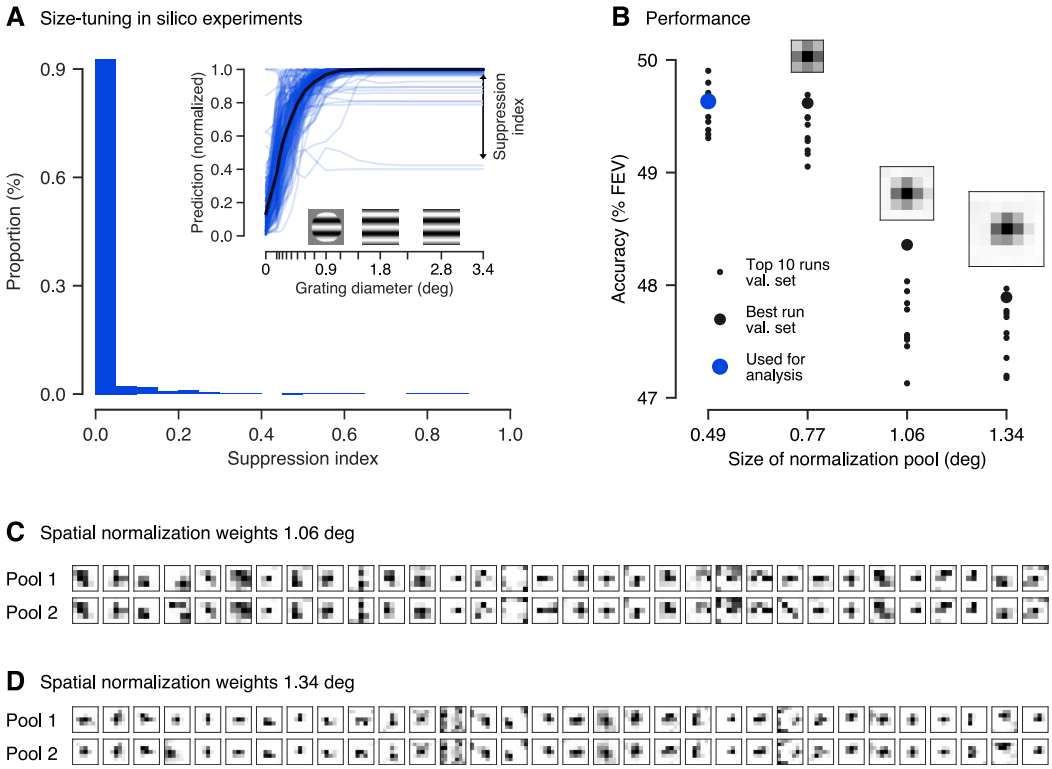


Fig 7. Size-tuning *in silico* experiments and spatially extended DN control models. A. inset: Prediction of the best DN model (chosen by validation set accuracy) for all neurons to gratings of increasing size. The gratings' properties were determined from the units' optimally stimulating Gabor pattern. As grating diameter increased, only very few neurons showed mostly weak suppression. Predictions normalized to maximum response per neuron. Suppression index measures asymptotic suppression relative to the maximum prediction **A. main panel**: Across all neurons and the ten best DN models (chosen by validation set accuracy), almost no neurons show significant surround suppression. **B**. Test set performance of the ten best performing DN models. The model's performance rapidly decreases for spatially increasing normalization pool size (in units of visual angle in degrees). The best model on the validation set is indicated by a blue dot. **C. & D**. Weights of the spatial normalization pool for the best performing model with pool size of (**C.**) 1.06˚ of visual field (5 px × 5 px) and (**D.**) 1.34˚ of visual field (7 px × 7 px; all evaluated in terms of the validation set accuracy). For each feature (columns), the two components (rows) of the in total 32 spatial normalization pools are shown. Darker color corresponds to higher weights. Both components are similar. **B. insets**: Average across features and normalization pool components. The model learned normalization from the receptive field center (on average).

https://doi.org/10.1371/journal.pcbi.1009028.g007

neuron, we generated a circular grating stimulus with optimal orientation and spatial frequency, which we presented to our model. As we increased the grating diameters, our DN model's spike count predictions were suppressed only for very few cells for very large stimuli (Fig 7A, inset), suggesting only weak surround-suppression in very few cells. To quantify this observation, we computed a surround suppression index reporting the reduction from the highest measured response to the response for a grating completely filling the whole image for each neuron (Fig 7A, inset). Thus, no suppression is indicated by a suppression index of zero, whereas a value of one would mean maximum suppression (see Methods for details). Across the best ten DN models and all neurons, we found only very few exceptions from no suppression (Fig 7A) with an average suppression index of (1.6 ± 0.8)% (mean and 95% confidence interval over the ten best models selected in terms of validation set accuracy). If our DN model would have learned surround suppression, we would have expected that the predictions of a

substantial number of neurons would significantly decrease for larger grating sizes leading to substantially higher suppression indices [32], which is not the case for our DN model. In conclusion, we expect the influence of very few neurons showing mostly very weak surround-suppression to our results to be very small making it unlikely that surround suppression could have led to the substantial orientation-specific normalization we observed. For the non-specific DN model and the three-layer CNN we found very similar behavior.

As a further control to verify that the orientation-specific normalization we observed in the RF is unlikely to be caused by surround influence, we fit a more general DN model which could additionally learn the spatial structure of the normalization pool, instead of limiting it only to neurons with overlapping receptive fields. To account for the RF's surround, we spatially expanded the model's normalization weights $p_{kl} \rightarrow p_{kluv}$ and introduced a convolution across space into the normalizing sum in Eq (1):

$$z_l(x) = \frac{y_l^{n_l}(x)}{\sigma_l^{n_l} + \sum_{k \in K} p_{kl\cdot\cdot} * y_k^{n_k}(x)} \quad . \tag{4}$$

Here, the normalization weights $p_{kl\cdot\cdot} = p_{kluv}$ encode which input drives $y_k^{n_k}$ are pooled over what spatial extent (indexed by u, v) to normalize the input drive $y_l^{n_l}$. Note that for the special choice of normalization weights p_{kl11} we perform a $u \times v = 1 \times 1$ convolution resulting in Eq (1) describing a DN model without surround. To keep the number of parameters computationally tractable, we factorized the normalization weights p_{kluv} allowing for two normalization pools (indexed by m) per output channel l, each consisting of a set of spatial integration weights c and feature weightings d,

$$p_{kluv} = \sum_{m=1}^{2} c_{luv,m} \cdot d_{kl,m}. \tag{5}$$

We constrained c and d to be non-negative to make sure the denominator in Eq (4) is non-negative (recall that $y_k^{n_k} \geq 0$ due to Eq (2)). The two normalization pools indexed by m could have different patterns of weights along the feature and spatial dimensions. Our extended model is therefore general enough to account for the standard model of DN with a nonspecific center normalization pool and orientation-specific surround suppression (see Methods for details).

In line with what we expect from our small stimuli mostly covering the RF center, spatially expanding the normalization pool to cover larger surround areas did not increase our model's accuracy; in fact, for lager normalization pools the performance even decreased (Fig 7B; for the best models on the validation set statistically significant beyond 0.77° of visual field, pairwise Wilcoxon signed rank test: $p < 0.025$, $N = 166$ neurons, family-wise error rate $\alpha = 0.05$ using Holm-Bonferroni correction). This could be due to the increased number of parameters, which might lead to over-fitting or non-optimal learning, since adding more weights to the model makes optimization harder. The best performance was achieved by models with the smallest normalization pool mostly within the units' RFs (0.25°–0.75° diameter estimated by spike triggered average, see Methods; 0.86° estimated from the smallest grating that leads to at least 95% of the maximum response in our *in silico* analysis, averaged across neurons and the best ten DN models with smallest normalization pool). The normalization weights of the extended spatial normalization pools showed no visible separation into center and surround and exhibited no or only weak contributions from the RF's surround (Fig 7). If leakage from the surround was the source of orientation-specific normalization, we would expect to observe spatial profiles that cover the surround, which is not the case (except for a few exceptions, which are mostly untuned features).

In summary, we controlled that the orientation-specific normalization learned by our spatially constrained DN model is unlikely caused by orientation-specific influence from the RF surround but mainly from the RF center. The reason for this limitation could be in our dataset, as the stimuli we used were mostly restricted to the receptive field masking out any surround influences and their short presentation time compared to the delay after which surround influences kick in (see Discussion and Methods), preventing any model to capture information from the surround.

Discussion

To improve our understanding of primary visual cortex, we asked what condensed nonlinear function state-of-the-art CNNs might implement for predicting V1 responses to localized natural stimuli. To answer this question, we developed an end-to-end learnable divisive normalization model and fit it to neural responses. Both the unspecific control model and the full model that learned the normalization pool outperformed the subunit model, with the full DN model outperforming its non-specific variant. Through the superior performance of the DN model, we quantify the relevance of divisive normalization for predicting V1 responses to natural images. With the help of *in silico* experiments, we found that both, the DN and the CNN model learned cross-orientation inhibition, showing that the deeper layers of the CNN model approximate the DN nonlinearities. Compared to the multi-layer CNN, the DN model offers a direct, easily understandable interpretation of its normalization pool. For the DN model, we found that normalization by similar orientations, and in general by similar features, is higher compared to dissimilar features within the receptive field.

Along those lines, one may ask whether the difference between the non-specific DN model and the full model that learns orientation specific normalization weights is relevant, because the full model may simply be able to better account for some insignificant biological heterogeneous imperfection due to its additional parameters. We believe that this explanation is unlikely, because oriented features are preferentially normalized by channels with similar orientation. If the model was simply picking up some biological imperfection, we would expect random deviations from uniform normalization weights rather than weights that depend systematically on preferred orientation.

Previous experimental work investigated suppressive phenomena within the receptive field only qualitatively or used simple stimuli that mainly consisted of a combination of driving and mask gratings. Morrone and colleagues [25] found suppression at all orientations, but did not investigate orientations similar to preferred orientation. Bonds [24] reported predominantly orientation-nonspecific suppression, although three of fourteen cells exhibit stronger suppression with masks oriented similarly to the neurons' preferred orientations, and a few other cells were suppressed most strongly by mask orientations orthogonal to the preferred orientation. Similarly, DeAngelis and colleagues [26] found suppression to be predominantly independent of orientation, although for some cells an increased suppression for a range of orientations near the optimal excitatory orientation was apparent. Heeger [18] explained those results by proposing an orientation-nonspecific divisive normalization model. Carandini and colleagues [27] considered the possibility of orientation-specific normalization which provides a marginal improvement in the quality of their model fits to the data. However, they concluded that their dataset was not specifically designed to provide a strong test of this question and their results were inconclusive in this respect. Busse and colleagues [29] developed a quantitative model for the response of a population of neurons to a combination of gratings. Assuming nonspecific normalization by overall contrast, their model predicted the collective action of the whole neuron population better than linear and winner-take-all baselines, but they did not test against

an orientation-specific alternative model. To summarize, these studies found phenomena that are predominantly explained by nonspecific normalization [18], some of them encountered only weak orientation-specific phenomena and only in relatively few cells. Until now, a quantitative analysis of orientation-specificity on a data set of natural images was missing.

Our findings are largely consistent with previous experimental results and quantitatively refine them using a larger dataset, place them in the context of other models of V1 and show that the same normalization mechanisms observed with simple stimuli also apply under more natural stimulus conditions. Interestingly, and somewhat in contrast to earlier work, features with preferred orientations within 10˚ of the driving feature provided twice as strong normalizing input than those with orthogonal preferred orientations.

The reason for this difference between our findings and previous studies could be that we used natural stimuli, which have different image statistics compared to simple stimuli used in earlier studies. Furthermore, most previous studies of divisive normalization were performed in cats [24–26, 29] and the results therein may not generalize to monkeys, for which preceding studies are inconclusive regarding orientation specificity [27].

Recent work modeling a large set of classical psychophysical data also suggests an orientation-specific divisive normalization: Schütt and Wichmann [42] developed an image-computable model of early vision very similar in structure to ours, and found that in order to explain classical data on contrast detection, contrast discrimination and oblique masking, their model required divisive normalization to be orientation-specific. Similar results had been reported in an earlier study by Itti and colleagues [43]. In contrast to those psychophysical studies involving experiments with human observers, in our paper we studied individual spiking neurons in primary visual cortex in response to natural stimulation, providing physiological evidence for orientation-specific divisive normalization. Moreover, the previous psychophysical models used a predefined Gabor filter bank and assumed that divisive normalization's orientation specificity is Gaussian distributed, learning only the standard deviation of the distribution. In contrast, our model is far more flexible, not making such strong assumptions, relying solely on the data to learn oriented filters and their divisive interactions.

Following a normative approach, Schwartz and Simoncelli [21] derive an ecologically justified divisive normalization model from the efficient coding hypothesis [44] that is able to qualitatively describe the orientation masking data of Bonds [24]. Reducing the statistical redundancy of responses to natural stimuli predicts that normalization should be stronger for neurons that exhibit a higher dependency in their unnormalized responses. This theoretical result implies that normalization weights should not be uniform, consistent with our empirical findings. Iyer and Burge [45] compare the output statistics of a divisive normalization model with broadband and narrowband normalization to natural images, $1/f$ stimuli and white noise stimuli. They report that feature-specific divisive normalization improves the discriminability among natural images compared to unspecific normalization.

Other research motivated by the efficient coding hypothesis [44] found that a recurrent sparse coding model reproduces several phenomena explained by DN, like cross-orientation inhibition [46]. In a recurrent excitation-inhibition network trained with Hebbian learning rules the inhibitory connections yielded stronger inhibition for similarly oriented cells, while the responses of excitatory cells were decorrelated [47]. This result of orientation specific normalization is qualitatively consistent with the findings of Schwartz and Simoncelli [21] and our results. These earlier models were trained using an unsupervised objective. Future research could focus on investigating the relationship between DN and recurrence within cortical circuits using recurrent models trained to predict neural responses in V1 to natural stimuli. Another interesting question could be to which degree efficient coding models match our findings of orientation specific normalization.

Is our discovery of divisive normalization by similar orientations actually implemented by the connectivity of neurons in primary visual cortex? The answer to this question could be reflected in the connectivity from inhibitory parvalbumin-expressing (PV) interneurons to pyramidal cells and their relation to neurons' tuning properties. Hofer and colleagues [48] find that, in the mouse, pyramidal cells and PV cells are homogeneously connected. Although a weak bias towards orientation tuning is apparent, they conclude that local inhibition in V1 is primarily non-specific. However, despite the connection probability between PV and pyramidal cells being homogeneous, connection strengths are quite heterogeneous: Individual PV cells strongly inhibit those pyramidal cells that share their visual selectivity [49]. This result is in line with our finding of orientation-specific normalization. Similarly, recent work employs single-neuron perturbations in layer 2/3 of mouse V1 and directly measures higher suppression of neighbouring excitatory neurons with a similar preferred orientation [50].

The stimuli in our dataset were specifically designed to investigate nonlinear processing in the RF center and to minimize surround suppression. The control analyses showing that surround suppression is unlikely in our dataset serve only the purpose of verifying that our conclusions about orientation-specific normalization in the receptive field are not due to leakage of surround suppression (which is known to be orientation-specific [30–34]). Without question, surround suppression is an important component of nonlinear processing in V1 and deserves further investigation and modeling using a dataset designed for that purpose.

In our stimuli the fully visible part was spatially restricted to approximately the size of the receptive field. For our dataset the RF diameters of 0.25˚—0.75˚ have been roughly estimated by spike-triggered average. RF size estimates based on the grating summation field (GSF) tend to be larger by approximately a factor of two [32]. The GSF is defined by the smallest grating diameter that leads a unit to respond with at least 95% of its maximum activity [32]. Accordingly, the average GSF size we estimated based on our *in silico* size-tuning experiments had a diameter of 0.86˚, roughly corresponding to the fully visible part of our stimuli (1˚ of visual field). For the remaining stimulus portion (1˚—2˚) outside of the GSF there could be normalizing influences from the surround (Figs 1A and 2 in ref. [32]). However, it is rather unlikely that this region strongly influenced our data, as in our stimuli a cosine mask fading to zero was applied in this area (see Methods). Rather, we would expect significant surround effects beyond 2˚ of visual field, which is not covered by our stimuli. Furthermore, in the results we have shown here, our model saw only the central 1.14˚ of the images (except for the variants extended to the surround, which did not learn significant surround interactions either). Thus, as we expected, our DN model did not exhibit significant surround suppression in our size-tuning control experiments and spatially extended DN control models suggested no surround interaction (Fig 7). Summarizing our control analysis, literature suggests that in general there are surround interactions under stimulation with natural images [34, 51], however, our spatially constrained stimuli likely restricted us to study DN within the receptive field and very likely prevented us from learning the influence of the RF's surround on normalization, making the surround unlikely to affect our result of orientation-specific normalization within the RF center.

Another property of our dataset that minimized the effect of surround suppression is that we focused on single images to predict a spike count in a relatively short time window covering the transient response, and ignored any temporal aspects or more sustained periods of the response. If there was some interaction from the partly masked surround, we would expect the onset of these effects to be delayed. Several studies report that the onset of surround suppression exhibits a 15—60 ms delay relative to the onset of center RF responses [52]. Bair and colleagues [53] reported average delays as short as 9 ms, however, surround suppression's peak effect was typically reached 40—50 ms after its onset. As responses to the center RF occur on

average 52 ms after stimulus onset [53], in the dataset we used [14] we would expect significant surround influences to set in at the end of the recording window, which was 40—100 ms after stimulus onset. Hence, if there might be any contribution from the surround, we would expect it to contribute only very weakly to the overall spike counts determined in this time window. Note, the limitations in studying surround effects are imposed by the available data—the modeling approach may well generalize to cover both the surround and the temporal structure —and thus the extra-classical receptive field should be addressed in future work with a dataset containing larger stimuli.

Compared to our divisive normalization model, the three-layer CNN performs a few percentage points FEV better, showing that it captures additional (non-linear) phenomena that divisive normalization cannot account for. Future work will be necessary to find out what these unexplained effects are. This result shows the importance of quantitative, data-driven modeling: our work suggests that a model that can account for pooling of subunits (complex cells) and divisive normalization is structurally insufficient to achieve optimal performance in modeling V1 responses.

The performance difference between subunit and DN model measured in terms of FEV is also just a few percentage points. However, the phenomenological difference between those models is quite substantial, as the DN model correctly captures cross-orientation inhibition while the subunit model does not. Therefore, one should probably not only measure model performance by considering FEV over a set of randomly sampled natural images, but instead also consider other metrics or specifically chosen test sets that highlight certain aspects of neural computation—such as cross-orientation inhibition.

Important future research directions that follow our methods include studying the role of divisive normalization in higher areas of the ventral stream, like V4 and inferotemporal cortex. Building a predictive model by stacking successive DN layers or including a final divisive nonlinearity for the shared feature space are great candidates for future consideration, especially as earlier studies already report signs of normalization in these higher areas [54–56].

In conclusion, we developed a model consisting of one layer of subunits followed by learned orientation-specific divisive normalization, which accounted remarkably well for V1 data. We hope that this quantitative approach of evaluating theories of computation in the brain by formalizing them as (components of) trainable predictive models will be used more widely in the future, so the field will (slowly) converge to an accurate general-purpose model of the visual system applicable to natural inputs.

Methods

Experimental details

We used the dataset described in detail by Cadena and colleagues [14] and provide a summary of the most important characteristics here. Electrophysiological recordings from two healthy, male rhesus macaque monkeys aged 12 and 9 years were performed with a 32-channel linear silicon probe. The monkeys were head-fixed and placed in front of a screen. They were trained to fixate on a target located at the center of the screen. The start of a trial was determined by maintained fixation on the target for 300 ms. The fixation tolerance was set to 0.42˚ around the center of the target. At the beginning of each recording session, population receptive fields were mapped with a sparse random dot stimulus. Each dot was of size 0.12˚ of visual angle and was presented over a uniform gray background, changing location and light intensity (black or white) randomly every 30 ms. The receptive field profiles per electrode channel were then obtained via reverse correlation (i. e. spike-triggered average). The center location of the population receptive field was subsequently estimated by averaging over channels and fitting a two-

dimensional Gaussian to the reverse correlation profiles. Afterwards, this location was used to place the images of the natural stimulus paradigm.

The dataset by Cadena and colleagues [14] consists of 7 250 distinct natural, greyscale images which were presented two to four times each. A fifth of these images (1 450) were taken from ImageNet [57]. Four additional texturized images were synthesized from each of them, preserving varying degrees of higher-order statistics. The images were cropped to 2 × 2 degrees of visual angle (140 px × 140 px). Before displaying the images on the screen, the images were normalized such that the central 1˚ (70 px) of each image had the same mean (111.5) and standard deviation (45) determined across the central portion of all original images. Pixels with an intensity that fell outside the display's range [0, 255] where clipped. Afterwards, all images were overlaid with a circular mask with a soft cosine fade-out fading to the screen's mean gray intensity (128) and an aperture with a diameter of 1˚.

Images were presented for 60 ms with no blanks in between. Neural responses were extracted in time windows of 40–100 ms after image onset (Fig 2), accounting for typical response latencies in primary visual cortex. The image sequence was randomized with the restriction that consecutive images do not belong to the same type (i. e. natural or one of the four texturized versions).

A few isolated neurons were discarded if their stimulus driven variability was too low [14]. The explainable variance in a dataset is smaller than the total variance because the observation noise prevents even a perfect model to account for all the variance in the data. Thus, targeting neurons that have sufficient explainable variance is necessary to train meaningful models of visually driven responses. For a neuron's spike count r, the explainable variance $\mathrm{Var}_{\mathrm{exp}}[r]$ is the difference between the the total variance $\mathrm{Var}[r]$ and the variance of the observational noise $\sigma^2_{\mathrm{noise}}$,

$$\mathrm{Var}_{\mathrm{exp}}[r] = \mathrm{Var}[r] - \sigma^2_{\mathrm{noise}} \ . \tag{6}$$

We estimated the variance of the observational noise by computing the variance of a neuron's response r_t in multiple trials t in which we presented the same stimulus x_j and subsequently taking the expectation E_j over all images,

$$\sigma^2_{\mathrm{noise}} = \mathrm{E}_j[\mathrm{Var}_t[r_t|x_j]] \ . \tag{7}$$

Neurons for which the ratio between the explainable to total variance was below 0.15 were removed. The resulting dataset includes spike count data for 166 isolated neurons, with an average ratio of explainable to total variance of 0.285. These neurons were recorded at 1˚–3˚ eccentricities and estimated receptive field size diameters were between 0.25˚ and 0.75˚. Since RF sizes were roughly estimated using spike-triggered average, it is likely that the values reported here underestimate the grating summation field defined by the smallest grating diameter that leads a unit to respond with at least 95% of its maximum activity by a factor of approximately two (see Discussion; similar to the minimum response field (MRF) underestimating the GSF as reported by Cavanaugh and colleagues [32]).

To keep the results of our models consistent and comparable to the gold standard baseline from Cadena and colleagues [14], we down-sampled the images by a factor of two to train our models. Likewise, images were cropped symmetrically, keeping the 40 × 40 central pixels (1.14˚ of visual angle). This size covers all of the recorded neurons' receptive fields, with a slight variability in their spatial location. Furthermore, the stimuli light intensities across all pixels and all images were centered around zero and normalized to have unit standard deviation. Additionally, we used the same random dataset splits of Cadena and colleagues [14] into training (64%), validation (16%) and testing (20%). We assessed our models' accuracy for a

specific architecture or set of hyper-parameters in the validation set and we report performance on the test set. We consistently used the same split throughout our study.

Divisive normalization model

Our model consists of two parts, a nonlinear core and a linear readout (see Results and Fig 1). The core (Fig 1A) processes the input stimulus x by convolving it with 32 filters w_k of size 13 px × 13 px (0.37˚ of visual angle) without padding, defining a bank of features indexed by k. Subsequently, we apply batch normalization [58] without re-scaling (BN*) which adds a bias term (o_l in Eq 2) and scales the responses to be of unit variance. This operation does not affect the overall computation and biological interpretation of our model, since scaling the driving inputs (y in Eq 1) by a factor β whilst scaling the normalization weights (p in Eq 1) and linear readout weights (Eq 3) by $1/\beta$ yields the same output. The batch normalization step is followed by a rectified linear unit (ReLU) nonlinearity

$$f(\cdot) = \max(0, \cdot) \ . \tag{8}$$

Hence, equivalent to Eq (2), the resulting 32 feature maps of size 28 px × 28 px for the excitatory drive are given by

$$y_l = f(\,\mathrm{BN} * (w_l * x)) \tag{9}$$

and contain information about the full input space covering 1.14˚ of visual field. Many neurons perform similar computations but respond at different localized areas of the visual field. Those receptive fields are represented by the kernels w_l, which we implemented convolutionally to make use of this knowledge. Furthermore, the ReLU nonlinearity (Eq 8) ensures that all feature maps are non-negative, $y_l \geq 0$ being coherent to the biological interpretation of an excitatory drive.

The feature maps y_l are then normalized divisively to produce 32 output feature maps

$$z_l = \frac{y_l^{n_l}}{\sigma_l^{n_l} + \sum_k p_{kl} \langle y_k^{n_k} \rangle} \tag{10}$$

shared by all neurons. Here, all operations are element-wise and the scalar semi-saturation constant $\sigma_l \geq 0$ is learned from the data. To include normalization by other channels k, we first exponentiate the excitatory feature maps y_k by the scalar $n_k \geq 0$ element-wise, which is learned from the data as well. Subsequently, low-pass filtering $\langle y_k^{n_k} \rangle$ is performed through average pooling in space with pool-size 5 px × 5 px (0.14˚, covering 0.49˚ in the input space taking the convolutional layer into account). We perform this pooling in order to achieve (approximate) phase invariance of the normalizing input without requiring a large number of filters with different phases. Subsequently, the results of the low-pass filtering are summed up, weighted by the normalization weights p_{kl}, and added into the denominator, resembling Eq (1). Furthermore, the normalization weights are constrained to be non-negative, $p_{kl} \geq 0$. Together with $y_k \geq 0$ and $\sigma_l \geq 0$, this ensures that the denominator in Eq (10) is non-negative, hence having a well-defined biological interpretation.

We converted the core's output feature maps z_l, shared by all neurons, to the activity of individual neurons via a readout for each of them (Fig 1B). To do so, we factorized the readout into spatial readout weights $a_{uv,i} \geq 0$ and feature readout weights $b_{l,i} \geq 0$ that pick the relevant locations and features plus a bias q_i,

$$g_i = \sum_{u,v,l} (a_{uv,i}\, b_{l,i})\, z_{uvl} + q_i \ . \tag{11}$$

Here, u, v index space and i indexes neurons. This factorization is beneficial because it reduces the number of parameters in the readout. Also, we wanted to ensure that the readout does not model any complex computations, which we achieved by this factorization and the non-negativity of the readout weights. Additionally, we limited complexity by imposing a sparseness prior on both weights, because each neuron should only respond to its receptive field which is represented by a sparse spatial readout weight and should not mix many different features which corresponds to a sparse feature readout weight. The readout can, however, model a complex cell [59] by linearly combining multiple channels of the shared feature space.

We fitted an output nonlinearity to obtain the final prediction of a neuron's activity $\hat{r}_i$ as suggested by Cadena and colleagues [14] from which we paraphrase the description here. Optimizing the output nonlinearity improves data-driven models, but has to be done carefully when simultaneously learning the shared core and the readout weights end-to-end. We therefore construct the output nonlinearity inspired by the shifted exponential linear unit (ELU*) [60],

$$\hat{r}_i = h_i(g_i) \cdot \mathrm{ELU}^*(g_i) \ , \quad \mathrm{ELU}^*(g_i) = \begin{cases} g_i & \text{if } g_i \geq 1 \\ \exp(g_i - 1) & \text{otherwise} \ , \end{cases} \tag{12}$$

keeping the resulting activity non-negative, and multiply it by a non-negative function

$$h_i(g_i) = \sum_j t_j(g_i)\, \mathrm{e}^{\alpha_{ji}} \tag{13}$$

with parameters α_{ji} which are learned for each neuron i independently. The basis functions $t_j(x)$ in the argument of h are tent functions leading to a piecewise linear function inside the exponential. The tent functions are defined as

$$t_j(x) = \min\left(\max\left(0, \frac{x - x_{j-1}}{\Delta x}\right), \max\left(0, \frac{x_{j+1} - x}{\Delta x}\right)\right) \ , \tag{14}$$

with interpolation points $x_j = -3, -2.82, \ldots, 6$ and $\Delta x = 0.18$. By regularizing the coefficients α_j with a L_2 and a smoothness penalty, the model is pushed towards using the identity function for inputs larger than one and a shifted exponential function for smaller inputs. If the data provides strong evidence in favor of a different output nonlinearity, the piecewise function h_i can be used to modify it. For example, the output nonlinearity is general enough to learn a sigmoidal shape, although for our best-fit models across model types it learned a slight expansive deviation from the identity in most cases and in a few cases a mix of slight expanding and slight compressing deviations from the identity function.

To optimize our model's parameters, we maximized the log-likelihood of the model's predictions given the data. To do so, we assumed that neurons' spikes are produced by a Poisson process. Our model predicts the average spike count $\hat{r}$ of a neuron, hence the probability of observing r spikes in the experiment is

$$P(r|\hat{r}) = \frac{\hat{r}^r}{r!} \mathrm{e}^{-\hat{r}} \ . \tag{15}$$

From that follows the Poisson log-likelihood

$$\ln P(r|\hat{r}) = \sum_{i,j} (r_i(x_j) \ln \hat{r}_i(x_j) - \ln(r_i(x_j)!) - \hat{r}_i(x_j)) \tag{16}$$

for all neurons i and all stimuli x_j. A neuron's response $r_i \equiv r_i(x_j)$ depends on the stimulus x_j, which we suppress in our further notation for better readability. For implementation reasons,

we wanted to minimize the Poisson loss function

$$\mathcal{L}_{\text{Poisson}} = \sum_{i,j} (\hat{r}_i - r_i \ln \hat{r}_i) \ , \tag{17}$$

which is the negative of the Poisson log-likelihood (Eq 16), where we omitted $\ln(r_i!)$ since this term does not depend on our model.

Furthermore, three terms regularizing the model's parameters were applied to the loss. We imposed a smoothness prior on the kernels w_k to ensure the spatial continuity of the predictors' receptive fields. The according penalty on the loss for weights being not smooth was determined with a Laplace filter L to be

$$\mathcal{L}_{\text{smooth}} = \sqrt{\sum_{u,v,k} (L * w_k)_{uv}^2} \ , \quad L = \begin{pmatrix} 0.25 & 0.5 & 0.25 \\ 0.5 & -3 & 0.5 \\ 0.25 & 0.5 & 0.25 \end{pmatrix} \ . \tag{18}$$

Due to their receptive fields, neurons only respond to a small, localized area of the visual field, which is why we imposed a sparsity prior on the spatial readout weights a_{uv}. Furthermore, neurons should only pool from a small set of feature maps to ensure that the readout does not perform complex computations. Thus we imposed a sparsity prior on the feature readout weights b_l as well. We achieved this by adding the L_1-norm of both weights

$$\mathcal{L}_{\text{sparse}} = \sum_i \sum_{u,v,l} |a_{uv,i}| \cdot |b_{l,i}| \tag{19}$$

to the loss function. As described by Cadena and colleagues [14], we regularized the output nonlinearity for all neurons by penalizing the sum of squares of the first and second finite differences of the weights α_{ji} to keep the learnable exponential function $h_i(x)$ smooth,

$$\mathcal{L}_{\text{out}} = \frac{1}{N} \sum_i \sum_{j=1}^{N} \left((\alpha_{j,i} - \alpha_{j-1,i})^2 + (2\alpha_{j,i} - \alpha_{j-1,i} - \alpha_{j+1,i})^2 \right) \ . \tag{20}$$

The final loss function to minimize with respect to our model's parameters is

$$\mathcal{L} = \mathcal{L}_{\text{Poisson}} + \lambda_{\text{smooth}} \mathcal{L}_{\text{smooth}} + \lambda_{\text{sparse}} \mathcal{L}_{\text{sparse}} + \lambda_{\text{out}} \mathcal{L}_{\text{out}} \ , \tag{21}$$

where λ_{smooth}, λ_{sparse} and λ_{out} are hyper-parameters which set the strength of the smoothness, the sparsity and the output nonlinearity penalty, respectively.

Divisive normalization model extended to normalization from surround

To extend our DN model to capture normalization from the spatial surround of a unit's RF, we replaced the weighted sum accounting for normalization (Eq 10) by a convolution that also covers space, keeping the rest of the original DN model unchanged,

$$z_l = \frac{y_l^{n_l}}{\sigma_l^{n_l} + s_l} \ , \quad s_l = \sum_k p_{kl\cdot\cdot} * \langle y_k^{n_k} \rangle \ , \tag{22}$$

equivalent to Eq (4). The new shared feature maps z_l consist of all element-wise operations where the newly introduced normalization feature maps s_l represent the strength by which the excitatory drive $y_l^{n_l}$ is normalized. The normalization feature maps are the result of a convolution between $\langle y_l^{n_l} \rangle$ and normalization pool kernels p_{kluv}.

For a larger convolutional kernel p, the feature maps s have smaller spatial dimensions than the excitatory feature maps y due to the convolution without padding. To be able to perform the element-wise division, we symmetrically cropped the excitatory feature maps y so that the resulting feature maps had the same spatial dimensions as s. Additionally, we wanted to keep the complexity (number of parameters) of the linear readout constant for all the size choices of the normalization kernel p. To this end, we symmetrically cropped the input images to a size that corresponds—after a forward pass through our model—to a shared feature space of spatial dimensions 28 px × 28 px (0.80˚, covering 1.14˚ of the input image taking the convolutional layer into account). Overall, this process enabled a fair comparison across all sizes of p.

To keep the kernel size of p computationally tractable, we used dilated convolutions which have defined gaps between the weights of the convolutional kernels. We chose to skip four pixels between each weight which is reflected by a dilation factor of five. Thereby, we were able to pool from a relatively large extra-classical RF while using few parameters. If we would compute this convolution directly on the feature maps $y_k^{n_k}$, the dilation would cause us to skip some elements in the feature map $y_k^{n_k}$ in the convolution's inner product for one specific position of the convolutional kernel, i. e. for one specific element in the suppression feature maps s_l. To consider all those elements we preceded the inner product computation of the convolution with an average pooling of 5 px × 5 px pool size (same as the dilation factor; 0.14˚, covering 0.49˚ in the input space taking the convolutional layer into account) which we performed at every spatial position in the input feature maps $y_k^{n_k}$. Then, exactly one weight of the convolutional kernel accounts for one 5 px × 5 px pool. In this view, the pools of the dilated kernel's neighbouring weights have coinciding boundaries. So in addition to implementing shift invariance, the average pooling makes sure that we do not loose information for the extended DN model. Due to this pooling, a normalization kernel p of spatial size 3 px × 3 px would spatially cover a normalization pool of size 15 px × 15 px (0.43˚, covering 0.77˚ in the input image taking the convolutional layer into account). We investigated models with normalization kernel sizes of 1 px × 1 px, 3 px × 3 px, 5 px × 5 px and 7 px × 7 px which spatially covered a five times larger normalization pool due to dilation. Those normalization pools covered visual angles of 0.49˚, 0.77˚, 1.06˚ and 1.34˚, respectively.

Baseline models

Convolutional neural network. Since the divisive normalization computation in our model was completely learned from the data, we wanted to compare to a baseline model that is purely data-driven as well. For this, the current state-of-the-art model is a convolutional neural network with three layers [14]. Its first convolutional layer consists of a kernel with spatial size of 13 px × 13 px (0.37˚ of visual field) and for the second and third layer of size 3 px × 3 px (0.09˚) each, covering 0.43˚ and 0.49˚ of the input image, respectively. All layers use 32 channels, batch normalization [58] and ELU nonlinearity [60],

$$\mathrm{ELU}(x) = \begin{cases} x & \text{if } x \geq 0 \\ \exp(x) - 1 & \text{otherwise} \end{cases}. \tag{23}$$

Similar to our model's architecture, the core part of the CNN model results in a nonlinear feature space shared by all neurons which is mapped to each neuron's activity with individual readout weights factorized in spatial and feature weightings. Sparseness of both of them is achieved by adding an L_1-penalty to the according loss function. This readout employs the same final output nonlinearity as in our DN model but differs from ours in having no non-negativity constraints on the weights.

Convolutional subunit model. Our convolutional subunit baseline model is structurally a one-layer convolutional neural network with multiple filters followed by a readout. It is exactly the same as our divisive normalization model but with the normalization function (Eq 10) replaced by the identity function

$$z_l = \mathrm{id}(y_l) = y_l \ . \tag{24}$$

Hence, the only difference to our DN model is the lack of normalization. The shared feature maps z_l consist of rectified outputs of linear filters (Eq 9) which approximate simple cells. The subsequent readout can sum up those simple cell responses with additional weightings, enabling the model to approximate complex cells [59]. We trained the model with the same loss function (Eq 21) as the divisive normalization model.

Hyper-parameter optimization

Our model's accuracy depends on several hyper-parameters. We set the initial learning-rate to 10^{-3} and used an early stopping training scheme: We evaluated the Poisson loss (Eq 17) every 100 training steps, after ten iterations of no improvement we decayed the learning-rate by a factor of three, and we repeated this four times. For the filters w_k in the first convolution, we found that a size of 13 px × 13 px (0.37˚ of visual field) was optimal, the same is true for the number of 32 channels.

The weights λ_{smooth} of the smoothness penalty (Eq 18), λ_{sparse} of the readout sparsity penalty (Eq 19) and λ_{out} of the output nonlinearity's penalty term (Eq 20) in the full loss function (Eq 21) were extensively cross-validated using the validation set of our data. We randomly sampled the smooth-weight λ_{smooth} from a logarithmic uniform distribution in the interval $[10^{-9}, 10^{-3.5}]$ and the readout sparse-weight λ_{sparse} from a logarithmic uniform distribution in the interval $[10^{-9}, 10^{-4.5}]$ for all models. For the divisive normalization model constrained to the receptive field center and the subunit model, we drew the output nonlinearity penalty weight λ_{out} from a logarithmic uniform distribution in the interval $[10^{-8}, 10^{2}]$. For those two models, we sampled 1 000 runs. For the non-specific divisive normalization model and the DN models extended to the spatial surround, we narrowed down the relevant parameter space of λ_{out} and sampled them from a logarithmic uniform distribution in the interval $[10^{-5}, 10^{0}]$. Since this halved the search space of the hyper-parameters, we halved the number of samples drawn for those models. Thus, the density of sample points in the search space was the same for all models and should lead to a fair comparison between their performance. For a fair comparison of all models in this study to the state-of-the-art three-layer CNN, we retrained this model with the same optimization procedure. After running optimization for a few 100 hyper-parameter samples, we observed that the best performing models were close to the upper bound of the readout sparsity penalty weight. Thus, we discarded those data-points and started searching again with a shifted search-interval for $\lambda_{\text{sparse}} \in [10^{-5.5}, 10^{-1}]$, sampling 1 000 model runs as before. In layer 2 and 3, we set the smooth-weights to zero and group-sparsity weights to $2.5 \cdot 10^{-4}$ since Cadena and colleagues [14] found best performance for those values and these parameters are specific to the three-layer CNN, being not applicable to the other models used in this study.

Accuracy evaluation

Average correlation. For architecture search, hyper-parameter optimization and the selection of specific models for analysis we evaluated models' accuracies on the validation set with the Pearson correlation coefficient between the measured spike counts and our models' predictions, averaged over neurons. If the prediction for one neuron is constant, the according

standard deviation is zero. Hence, for such neurons the correlation coefficient was not computable due to division by zero. For those neurons, we set the correlation coefficient to zero before averaging. This average correlation measure does not consider observational noise (Eq 7).

Fraction of explainable variance explained. For reporting accuracy values in this paper, we used the data's test set to compute the fraction of explainable variance explained (FEV) per neuron

$$\mathrm{FEV} = 1 - \frac{\mathrm{Var}_{\mathrm{res}}[r]}{\mathrm{Var}_{\mathrm{exp}}[r]} \tag{25}$$

which utilizes the variance that is explainable in principle, $\mathrm{Var}_{\mathrm{exp}}[r]$ (Eq 6), and the variance of the residuals corrected by the observation noise,

$$\mathrm{Var}_{\mathrm{res}}[r] = \frac{1}{N}\sum_{j}^{N}(r_j - \hat{r}_j)^2 - \sigma^2_{\mathrm{noise}} \ , \tag{26}$$

where j indexes images. This measure corrects for observation noise, which variance $\sigma^2_{\mathrm{noise}}$ we estimated with Eq (7). To compute model performance we averaged the FEV across neurons.

In silico experiments

To perform cross-orientation and size tuning experiments *in silico*, we first had to determine the optimal Gabor pattern that elicited maximum response for all units in the models we investigated. We specified an image containing a Gabor pattern by the Gabor's center location, size, spatial frequency, phase, orientation and contrast. For each neuron and model, we obtained the optimal Gabor image by finding the parameter combination that lead to maximum model prediction. For the center location we tested all positions in the 40 px × 40 px (1.14˚ of visual field) image using 12 different orientations with equally spaced angles $\phi/\pi = 0, 1/12, 2/12, \ldots, 11/12$ (in units of π) and 8 phases $\psi/2\pi = 0, 1/8, 2/8, \ldots, 7/8$ (in units of 2π). We searched for size, spatial frequency and contrast with equidistant values in log-space. Specifically, we sampled 8 sizes $s_i = 4\ \mathrm{px} \cdot (1.3895)^i$, $i = 0, 1, \ldots, 7$, resulting in a maximum size of 40px, where the size of the Gabor is defined as ±2 standard deviations of its Gaussian envelope. For spatial frequency, we used 10 values $f_i = (1.3)^{-1} \cdot (1.3)^i$, $i = 0, 1, 2, \ldots, 9$ measured in cycles per four standard deviations of the according Gaussian envelope. We determined the maximum contrast from the images in our training set. After normalization, the mean across all pixel values was zero, the minimum and maximum pixel values were −2.52 and 3.20, respectively. Thus, making sure that the Gabor's pixel values are within the value range seen by the models during training, we set the maximum contrast to $c_{\mathrm{max}} = 2 \cdot 2.52$ corresponding to a maximum amplitude of $a_{\mathrm{max}} = 2.52$. We chose 6 amplitude (contrast) values $a_i = c_i/2 = 0.01\ a_{\mathrm{max}} \cdot (2.51189)^i$, $i = 0, 1, 2, \ldots, 5$ resulting in $c_5 = c_{\mathrm{max}}$.

Cross-orientation inhibition. For the cross-orientation inhibition experiment, we used the parameters determined for the optimal Gabor to generate plaids (Fig 4A, insets) by linear superposition of the optimal Gabor with a 90˚ rotated version of it. We sampled the contrasts of both Gabor components from all combinations out of $\mathcal{C} = \{0, c_i\}$ with $c_i = 0.01\ c_{\mathrm{max}} \cdot (1.63069)^i$, $i = 0, 1, 2, \ldots, 8$ where the highest contrast was 50% of the maximum contrast, $c_8 = 0.5\ c_{\mathrm{max}}$, making sure the pixel values of the combined plaid stimulus stayed within the value range seen by our models during training. In addition, for the orthogonal Gabor we used 8 phases $\psi/2\pi = 0, 1/8, 2/8, \ldots, 7/8$ (in units of 2π). We obtained contrast tuning curves presenting all plaid stimuli averaging across phases to get results that are independent of relative

phase between optimally and orthogonally oriented Gabor components. To quantify cross-orientation inhibition strength, we defined a cross-orientation inhibition index (COI) measuring the relative reduction in response due to adding an orthogonal Gabor with contrast c_i to the driving Gabor of optimal orientation with contrast c_j for all contrast combinations $(c_i, c_j) \in \mathcal{C}_{\text{orthogonal}} \times \mathcal{C}_{\text{optimal}}$,

$$\text{COI}\,(c_i, c_j) = 1 - \frac{\hat{r}(c_i, c_j)}{\hat{r}(c_i = 0, c_j)} \quad . \tag{27}$$

We define a cell to be cross-orientation inhibited if the cross-orientation inhibition index COI is at least 10% for any combination of optimal and orthogonal Gabor contrasts, i.e. the presence of the orthogonal Gabor reduces response by at least 10%.

Size tuning and surround-suppression control analysis. To obtain size tuning curves, we generated gratings using the parameters of the optimal Gabor for each cell. Then, instead of presenting a Gabor with a Gaussian envelope, we present a sinusoidal grating with full contrast and a sharp cut-off to zero intensity (50% grey) outside of a circular area with predefined diameter (Fig 7). We set the maximum grating diameter to d_{max} = 120px (3.43˚ of visual field), making sure to cover the whole 40 px × 40 px (1.14˚ of visual field) image even if the grating's center would be in one of the image corners. Then, we chose diameters $d_i = 0.05\, d_{\text{max}} \cdot (1.23859)^i$, i = 0, 1, 2, . . ., 14. As almost all optimal Gabors had maximum contrast, we presented the gratings in this experiment with maximum contrast, too. To quantify any surround suppression, we used the suppression index (SI) proposed by Cavanaugh and colleagues [32], measuring the reduction from the highest model prediction $\hat{r}_{\text{max}}$ to the prediction $\hat{r}_{\text{supp}}$ for a grating completely filling the whole image for each neuron,

$$\text{SI} = 1 - \frac{\hat{r}_{\text{supp}}}{\hat{r}_{\text{max}}} \quad . \tag{28}$$

Hence, a SI of zero corresponds to no suppression, whereas a fully suppressed cell would lead to a SI of one.

Evaluation of orientation-specific normalization

To analyse how the preferred orientation of the features being normalized depend on that of the features providing normalizing inputs (Fig 5), we determined for each feature map whether it extracts oriented features and—if so—its preferred orientation. To do so, we windowed each convolutional kernel of size 13 px × 13 px (0.37˚ of visual field) with a Gaussian window (standard deviation: 3 px, corresponding to 0.09˚), normalized it and then computed its 2D power spectrum (using the discrete Fourier transform with 64 × 64 samples). We then quantified how power spectral density is distributed across orientations by computing a mean resultant vector m given by:

$$m = \sum_{u,v \in R} F_{uv} e^{2i\phi}, \tag{29}$$

where F_{uv} is the Fourier transformed kernel, $R = \{(u, v) : 0.3 < \sqrt{u^2 + v^2} < 0.7\}$ contains all frequencies between 0.3 and 0.7 (with 1.0 being the Nyquist frequency), ϕ = atan2(v, u) is the orientation, i the imaginary unit and the factor 2 in the complex exponential accounts for the fact that we are interested in orientation, which is periodic in 180˚ or π. If all power in a kernel is concentrated in one orientation, the mean resultant vector will be long, whereas an unoriented kernel will have a mean resultant vector near zero. Based on visual inspection of the

kernels in model fits of the top 11 to 20 model runs (in terms of validation set accuracy), we found $|m| = 0.125$ to be a reasonable threshold for separating oriented from unoriented features and used it as a heuristic for further analyses of the best ten model runs, avoiding issues with post-hoc statistical testing.

To quantify how strong a feature l is normalized by other features k, we computed the average normalizing input, which is given as the expected value (over images) of the product $p_{kl} \cdot y_{kuv}(x)$ in Eq (1). Since this normalization input depends on the stimulus, we computed its expected value of all images in the validation set. We removed the dependence on space by averaging over all locations within the feature map.

Control: All channels contribute to our model's prediction

To verify that all features contribute to normalization, we analyzed the readout feature weights for the best ten models (assessed in terms of performance on the validation set). However, there are two issues that prevent a direct comparison across models and neurons of the feature weightings. First, the factorization of the readout into spatial and feature weightings is not unique: scaling the spatial weights (a in Eq (11)) by a factor β whilst scaling the feature weights (b in Eq (11)) by $1/\beta$ yields the same output limiting comparisons across neurons. Second, a similar exercise between the normalization weights p and the semi-saturation constant σ (Eq 10) impedes comparison across models. To solve these issues, we normalized the feature readout weights of each neuron for this control analysis so that the resulting vectors for each neuron and model convey how much a certain feature map contributes to predict a neuron's response compared to the other feature maps, making the feature readout weights comparable across neurons. Next, we averaged these weights across neurons to assess the importance of the features in a model. Since these normalized feature readout weights were comparable across both neurons and models, we calculated a collective distribution of the averaged feature readout weights from the best ten models. To make sense of this distribution's absolute values, we evaluated its width in terms of the coefficient of variation, which is the standard deviation in units of the mean.

Implementation details

We used Tensorflow [61] to implement models as well as Python, which we additionally used for data analysis. We optimized models with the Adam optimizer [62] using mini-batches of size 256. In addition, we used the Python packages Numpy/Scipy [63], Pandas [64], Matplotlib [65], Seaborn [66], Datajoint [67, 68] and the tools Jupyter [69] and Docker [70].

Acknowledgments

We thank Fabian H. Sinz, Felix A. Wichmann, David A. Klindt, Alexander Böttcher, Matthias Kümmerer, Dylan M. Paiton, Robert Geirhos, Claudio Michaelis and Ivan Ustyuzhaninov for valuable discussions. M.F.B. and S.A.C. thank the International Max Planck Research School for Intelligent Systems (IMPRS-IS). Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DARPA, DoI/IBC, or the U. S. Government.

Author Contributions

Conceptualization: Max F. Burg, Santiago A. Cadena, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Data curation: Max F. Burg, Santiago A. Cadena, George H. Denfield, Edgar Y. Walker.

Formal analysis: Max F. Burg, Santiago A. Cadena, Alexander S. Ecker.

Funding acquisition: Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Investigation: Max F. Burg, Santiago A. Cadena, George H. Denfield.

Methodology: Max F. Burg, Santiago A. Cadena, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Project administration: Max F. Burg, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Resources: Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Software: Max F. Burg, Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Alexander S. Ecker.

Supervision: Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

Validation: Max F. Burg, Santiago A. Cadena.

Visualization: Max F. Burg, Santiago A. Cadena, Alexander S. Ecker.

Writing – original draft: Max F. Burg, Santiago A. Cadena.

Writing – review & editing: Max F. Burg, Santiago A. Cadena, George H. Denfield, Edgar Y. Walker, Andreas S. Tolias, Matthias Bethge, Alexander S. Ecker.

References

1. Carandini M, Demb JB, Mante V, Tolhurst DJ, Dan Y, Olshausen BA, et al. Do we know what the early visual system does? Journal of Neuroscience. 2005; 25:10577–10597. https://doi.org/10.1523/JNEUROSCI.3726-05.2005 PMID: 16291931
2. Simoncelli EP, Paninski L, Pillow J, Schwartz O. Characterization of neural responses with stochastic stimuli. The Cognitive Neurosciences. 2004; 3:327–338.
3. Adelson EH, Bergen JR. Spatiotemporal Energy Models for the Perception of Motion. Journal of the Optical Society of America A. 1985; 2:284–299. https://doi.org/10.1364/JOSAA.2.000284
4. Rust NC, Schwartz O, Movshon JA, Simoncelli EP. Spatiotemporal Elements of Macaque V1 Receptive Fields. Neuron. 2005; 46:945–956. https://doi.org/10.1016/j.neuron.2005.05.021
5. Touryan J, Felsen G, Dan Y. Spatial structure of complex cell receptive fields measured with natural images. Neuron. 2005; 45(5):781–791. https://doi.org/10.1016/j.neuron.2005.01.029
6. Willmore B, Prenger RJ, Wu MCK, Gallant JL. The berkeley wavelet transform: a biologically inspired orthogonal wavelet transform. Neural Computation. 2008; 20(6):1537–1564. https://doi.org/10.1162/neco.2007.05-07-513
7. Butts DA, Weng C, Jin J, Alonso JM, Paninski L. Temporal precision in the visual pathway through the interplay of excitation and stimulus-driven suppression. Journal of Neuroscience. 2011; 31(31):11313–11327. https://doi.org/10.1523/JNEUROSCI.0434-11.2011
8. McFarland JM, Cui Y, Butts DA. Inferring nonlinear neuronal computation based on physiologically plausible inputs. PLoS Computational Biology. 2013; 9(7):e1003143. https://doi.org/10.1371/journal.pcbi.1003143
9. Vintch B, Movshon JA, Simoncelli EP. A Convolutional Subunit Model for Neuronal Responses in Macaque V1. Journal of Neuroscience. 2015; 35:14829–14841. https://doi.org/10.1523/JNEUROSCI.2815-13.2015
10. Yamins DLK, Hong H, Cadieu CF, Solomon EA, Seibert D, DiCarlo JJ. Performance-Optimized Hierarchical Models Predict Neural Responses in Higher Visual Cortex. Proceedings of the National Academy of Sciences. 2014; 111:8619–8624. https://doi.org/10.1073/pnas.1403112111
11. Khaligh-Razavi SM, Kriegeskorte N. Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS Computational Biology. 2014; 10:e1003915. https://doi.org/10.1371/journal.pcbi.1003915

12. McIntosh L, Maheswaranathan N, Nayebi A, Ganguli S, Baccus S. Deep learning models of the retinal response to natural scenes. In: Advances in Neural Information Processing Systems 29; 2016. p. 1369–1377.
13. Zhang Y, Lee TS, Li M, Liu F, Tang S. Convolutional neural network models of V1 responses to complex patterns. Journal of Computational Neuroscience. 2019; 46(1):33–54. https://doi.org/10.1007/s10827-018-0687-7
14. Cadena SA, Denfield GH, Walker EY, Gatys LA, Tolias AS, Bethge M, et al. Deep convolutional models improve predictions of macaque V1 responses to natural images. PLoS Computational Biology. 2019; 15(4):e1006897. https://doi.org/10.1371/journal.pcbi.1006897 PMID: 31013278
15. Kindel WF, Christensen ED, Zylberberg J. Using deep learning to probe the neural code for images in primary visual cortex. Journal of Vision. 2019; 19(4):29–29. https://doi.org/10.1167/19.4.29
16. Walker EY, Sinz FH, Cobos E, Muhammad T, Froudarakis E, Fahey PG, et al. Inception loops discover what excites neurons most using deep predictive models. Nature Neuroscience. 2019; 22(12):2060–2065. https://doi.org/10.1038/s41593-019-0517-x PMID: 31686023
17. Sinz F, Ecker AS, Fahey P, Walker E, Cobos E, Froudarakis E, et al. Stimulus domain transfer in recurrent models for large scale cortical population prediction on video. In: Advances in Neural Information Processing Systems 31; 2018. p. 7199–7210.
18. Heeger DJ. Normalization of Cell Responses in Cat Striate Cortex. Visual Neuroscience. 1992; 9:181–197. https://doi.org/10.1017/S0952523800009640
19. Carandini M, Heeger DJ. Normalization as a Canonical Neural Computation. Nature Reviews Neuroscience. 2012; 13:51–62. https://doi.org/10.1038/nrn3136
20. Sawada T, Petrov AA. The Divisive Normalization Model of V1 Neurons: A Comprehensive Comparison of Physiological Data and Model Predictions. Journal of Neurophysiology. 2017; 118:3051–3091. https://doi.org/10.1152/jn.00821.2016
21. Schwartz O, Simoncelli EP. Natural Signal Statistics and Sensory Gain Control. Nature Neuroscience. 2001; 4:819–825. https://doi.org/10.1038/90526
22. Sinz F, Bethge M. The Conjoint Effect of Divisive Normalization and Orientation Selectivity on Redundancy Reduction. In: Advances in Neural Information Processing Systems 21; 2008. p. 1521–1528.
23. Ballé J, Laparra V, Simoncelli EP. End-to-end optimized image compression. In: International Conference on Learning Representations; 2017.
24. Bonds AB. Role of Inhibition in the Specification of Orientation Selectivity of Cells in the Cat Striate Cortex. Visual Neuroscience. 1989; 2:41–55. https://doi.org/10.1017/S0952523800004314
25. Morrone MC, Burr DC, Maffei L. Functional Implications of Cross-Orientation Inhibition of Cortical Visual Cells. I. Neurophysiological Evidence. Proceedings of the Royal Society of London Series B Biological Sciences. 1982; 216:335–354.
26. DeAngelis GC, Robson JG, Ohzawa I, Freeman RD. Organization of Suppression in Receptive Fields of Neurons in Cat Visual Cortex. Journal of Neurophysiology. 1992; 68:144–163. https://doi.org/10.1152/jn.1992.68.1.144
27. Carandini M, Heeger DJ, Movshon JA. Linearity and Normalization in Simple Cells of the Macaque Primary Visual Cortex. Journal of Neuroscience. 1997; 17:8621–8644. https://doi.org/10.1523/JNEUROSCI.17-21-08621.1997
28. Freeman TCB, Durand S, Kiper DC, Carandini M. Suppression without Inhibition in Visual Cortex. Neuron. 2002; 35:759–771. https://doi.org/10.1016/S0896-6273(02)00819-X
29. Busse L, Wade AR, Carandini M. Representation of Concurrent Stimuli by Population Activity in Visual Cortex. Neuron. 2009; 64:931–942. https://doi.org/10.1016/j.neuron.2009.11.004
30. Blakemore C, Tobin EA. Lateral Inhibition between Orientation Detectors in the Cat's Visual Cortex. Experimental Brain Research. 1972; 15:439–440.
31. DeAngelis GC, Freeman RD, Ohzawa I. Length and Width Tuning of Neurons in the Cat's Primary Visual Cortex. Journal of Neurophysiology. 1994; 71:347–374. https://doi.org/10.1152/jn.1994.71.1.347
32. Cavanaugh JR, Bair W, Movshon JA. Nature and Interaction of Signals From the Receptive Field Center and Surround in Macaque V1 Neurons. Journal of Neurophysiology. 2002; 88:2530–2546. https://doi.org/10.1152/jn.00692.2001
33. Cavanaugh JR, Bair W, Movshon JA. Selectivity and Spatial Distribution of Signals From the Receptive Field Surround in Macaque V1 Neurons. Journal of Neurophysiology. 2002; 88:2547–2556. https://doi.org/10.1152/jn.00693.2001
34. Coen-Cagli R, Kohn A, Schwartz O. Flexible Gating of Contextual Influences in Natural Vision. Nature Neuroscience. 2015; 18:1648–1655. https://doi.org/10.1038/nn.4128

35. Heuer HW, Britten KH. Contrast dependence of response normalization in area MT of the rhesus macaque. Journal of Neurophysiology. 2002; 88(6):3398–3408. https://doi.org/10.1152/jn.00255.2002

36. Wainwright MJ, Schwartz O, Simoncelli EP. Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In: Rao R, Olshausen B, Lewicki M, editors. Probabilistic Models of the Brain: Perception and Neural Function. MIT Press; 2002. p. 203–222.

37. Froudarakis E, Berens P, Ecker AS, Cotton RJ, Sinz FH, Yatsenko D, et al. Population code in mouse V1 facilitates readout of natural scenes through increased sparseness. Nature Neuroscience. 2014; 17 (6):851. https://doi.org/10.1038/nn.3707 PMID: 24747577

38. Antolík J, Hofer SB, Bednar JA, Mrsic-Flogel TD. Model constrained by visual hierarchy improves prediction of neural responses to natural scenes. PLoS Computational Biology. 2016; 12(6):e1004927. https://doi.org/10.1371/journal.pcbi.1004927

39. Klindt D, Ecker AS, Euler T, Bethge M. Neural System Identification for Large Populations Separating "What" and "Where". In: Advances in Neural Information Processing Systems 30; 2017. p. 3506–3516.

40. Batty E, Merel J, Brackbill N, Heitman A, Sher A, Litke A, et al. Multilayer Recurrent Network Models of Primate Retinal Ganglion Cell Responses. In: International Conference on Learning Representations; 2017.

41. Denfield GH, Ecker AS, Shinn TJ, Bethge M, Tolias AS. Attentional fluctuations induce shared variability in macaque primary visual cortex. Nature Communications. 2018; 9(1):2654. https://doi.org/10.1038/s41467-018-05123-6

42. Schütt HH, Wichmann FA. An Image-Computable Psychophysical Spatial Vision Model. Journal of Vision. 2017; 17.

43. Itti L, Koch C, Braun J. Revisiting spatial vision: Toward a unifying model. Journal of the Optical Society of America A. 2000; 17(11):1899–1917. https://doi.org/10.1364/JOSAA.17.001899

44. Barlow HB. Possible Principles Underlying the Transformations of Sensory Messages. In: Rosenblith WA, editor. Sensory Communication. Cambridge, Massachusetts: MIT Press; 1961. p. 217–234.

45. Iyer A, Burge J. The statistics of how natural images drive the responses of neurons. Journal of Vision. 2019; 19(13):4–4. https://doi.org/10.1167/19.13.4

46. Zhu M, Rozell CJ. Visual nonclassical receptive field effects emerge from sparse coding in a dynamical system. PLoS Computational Biology. 2013; 9(8):e1003191. https://doi.org/10.1371/journal.pcbi.1003191

47. King PD, Zylberberg J, DeWeese MR. Inhibitory interneurons decorrelate excitatory cells to drive sparse code formation in a spiking model of V1. Journal of Neuroscience. 2013; 33(13):5475–5485. https://doi.org/10.1523/JNEUROSCI.4188-12.2013

48. Hofer SB, Ko H, Pichler B, Vogelstein J, Ros H, Zeng H, et al. Differential Connectivity and Response Dynamics of Excitatory and Inhibitory Neurons in Visual Cortex. Nature Neuroscience. 2011; 14:1045–1052. https://doi.org/10.1038/nn.2876 PMID: 21765421

49. Znamenskiy P, Kim MH, Muir DR, Iacaruso MF, Hofer SB, Mrsic-Flogel TD. Functional Selectivity and Specific Connectivity of Inhibitory Neurons in Primary Visual Cortex. bioRxiv. 2018; p. 294835.

50. Chettih SN, Harvey CD. Single-neuron perturbations reveal feature-specific competition in V1. Nature. 2019; 567(7748):334–340. https://doi.org/10.1038/s41586-019-0997-6

51. Vinje WE, Gallant JL. Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. Science. 2000; 287:1273–1276. https://doi.org/10.1126/science.287.5456.1273

52. Angelucci A, Bressloff PC. Contribution of feedforward, lateral and feedback connections to the classical receptive field center and extra-classical receptive field surround of primate V1 neurons. Progress in Brain Research. 2006; 154:93–120. https://doi.org/10.1016/S0079-6123(06)54005-1

53. Bair W, Cavanaugh JR, Movshon JA. Time course and time-distance relationships for surround suppression in macaque V1 neurons. Journal of Neuroscience. 2003; 23(20):7690–7701. https://doi.org/10.1523/JNEUROSCI.23-20-07690.2003

54. Zoccolan D, Cox DD, DiCarlo JJ. Multiple object response normalization in monkey inferotemporal cortex. Journal of Neuroscience. 2005; 25(36):8150–8164. https://doi.org/10.1523/JNEUROSCI.2058-05.2005

55. Kaliukhovich DA, Vogels R. Divisive normalization predicts adaptation-induced response changes in macaque inferior temporal cortex. Journal of Neuroscience. 2016; 36(22):6116–6128. https://doi.org/10.1523/JNEUROSCI.2011-15.2016

56. Reynolds JH, Desimone R. Interacting roles of attention and visual salience in V4. Neuron. 2003; 37 (5):853–863. https://doi.org/10.1016/S0896-6273(03)00097-7

57. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision. 2015; 115:211–252. https://doi.org/10.1007/s11263-015-0816-y

58. Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: International Conference on Machine Learning; 2015. p. 448–456.

59. Hubel DH, Wiesel TN. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology. 1962; 160:106–154. https://doi.org/10.1113/jphysiol.1962.sp006837

60. Clevert DA, Unterthiner T, Hochreiter S. Fast and accurate deep network learning by exponential linear units (elus). arXiv preprint arXiv:151107289. 2015.

61. Abadi M, Agarwal A, Barham P, Brevdo E, Chen Z, Citro C, et al. TensorFlow: Large-Scale Machine Learning on Heterogeneous Distributed Systems; 2015.

62. Kingma D, Ba J. Adam: A method for stochastic optimization. In: International Conference on Learning Representations; 2015.

63. Walt Svd, Colbert SC, Varoquaux G. The NumPy array: a structure for efficient numerical computation. Computing in Science & Engineering. 2011; 13(2):22–30. https://doi.org/10.1109/MCSE.2011.37

64. McKinney W. Data Structures for Statistical Computing in Python. In: van der Walt S, Millman J, editors. Proceedings of the 9th Python in Science Conference; 2010. p. 51–56.

65. Hunter JD. Matplotlib: A 2D graphics environment. Computing in Science & Engineering. 2007; 9 (3):90–95. https://doi.org/10.1109/MCSE.2007.55

66. Waskom M, Botvinnik O, O'Kane D, Hobson P, Lukauskas S, Gemperline DC, et al. mwaskom/seaborn: v0.8.1 (September 2017); 2017.

67. Yatsenko D, Reimer J, Ecker AS, Walker EY, Sinz F, Berens P, et al. DataJoint: managing big scientific data using MATLAB or Python. BioRxiv. 2015; p. 031658.

68. Yatsenko D, Walker EY, Tolias AS. DataJoint: a simpler relational data model. arXiv preprint arXiv:180711104. 2018.

69. Kluyver T, Ragan-Kelley B, Pérez F, Granger B, Bussonnier M, Frederic J, et al. Jupyter Notebooks—a publishing format for reproducible computational workflows. In: Loizides F, Schmidt B, editors. Positioning and Power in Academic Publishing: Players, Agents and Agendas. IOS Press; 2016. p. 87–90.

70. Merkel D. Docker: Lightweight Linux Containers for Consistent Development and Deployment. Linux Journal. 2014; 2014(239).

A.5 Diverse feature visualizations reveal invariances in early layers of deep neural networks

Cadena, Santiago A, Weis, M. A., Gatys, L. A., Bethge, M., & Ecker, A. S. (2018). Diverse feature visualizations reveal invariances in early layers of deep neural networks. *Proceedings of the European Conference on Computer Vision (ECCV)*, 217–232

Abstract

Visualizing features in deep neural networks (DNNs) can help understanding their computations. Many previous studies aimed to visualize the selectivity of individual units by finding meaningful images that maximize their activation. However, comparably little attention has been paid to visualizing to what image transformations units in DNNs are invariant. Here we propose a method to discover invariances in the responses of hidden layer units of deep neural networks. Our approach is based on simultaneously searching for a batch of images that strongly activate a unit while at the same time being as distinct from each other as possible. We find that even early convolutional layers in VGG-19 exhibit various forms of response invariance: near-perfect phase invariance in some units and invariance to local diffeomorphic transformations in others. At the same time, we uncover representational differences with ResNet-50 in its corresponding layers. We conclude that invariance transformations are a major computational component learned by DNNs and we provide a systematic method to study them.

Author contributions

Conceptualization: **SC**, MB, AE. Data Curation: **SC**, MW. Formal Analysis: **SC**, MW, AE. Funding Acquisition: AE, MB. Investigation: **SC**, MW, AE. Methodology: **SC**, MW, AE. Project Administration: **SC**, AE. Resources: **SC**, MW, LG, AE, MB. Software: **SC**, MW, AE. Supervision: AE, MB. Validation: **SC**, MW. Visualization: **SC**, MW. Writing – Original Draft Preparation: **SC**, AE. Writing – Review and Editing: **SC**, MW, LG, MB, AE.

Diverse feature visualizations reveal invariances in early layers of deep neural networks

Santiago A. Cadena, Marissa A. Weis, Leon A. Gatys,
Matthias Bethge, and Alexander S. Ecker

Centre for Integrative Neuroscience and Institute for Theoretical Physics
Bernstein Center for Computational Neuroscience
University of Tübingen, Germany
{first.last}@bethgelab.org

Abstract. Visualizing features in deep neural networks (DNNs) can help understanding their computations. Many previous studies aimed to visualize the selectivity of individual units by finding meaningful images that maximize their activation. However, comparably little attention has been paid to visualizing to what image transformations units in DNNs are invariant. Here we propose a method to discover invariances in the responses of hidden layer units of deep neural networks. Our approach is based on simultaneously searching for a batch of images that strongly activate a unit while at the same time being as distinct from each other as possible. We find that even early convolutional layers in VGG-19 exhibit various forms of response invariance: near-perfect phase invariance in some units and invariance to local diffeomorphic transformations in others. At the same time, we uncover representational differences with ResNet-50 in its corresponding layers. We conclude that invariance transformations are a major computational component learned by DNNs and we provide a systematic method to study them.

Keywords: Feature visualization, invariance, phase invariance, deep neural networks, early visual system.

1 Introduction

As deep neural networks have gained popularity in many scientific disciplines and technological applications, there is a growing interest in understanding the representations they learn and the computations they perform. One approach towards achieving such understanding is to visualize the features that activate the neurons in a network. There is a growing body of work that seeks to visualize features by synthesizing images which maximally drive hidden layer units. While this approach can give us a rough intuition about a unit's selectivity, it provides only a very incomplete picture of its computation. In addition to characterizing feature detectors by the stimulus that elicits the largest response, it is important to identify the nuisance parameters to which the neuron is invariant. As hidden layers build up response invariances gradually with depth, it is not the *image*

that most strongly drives a unit that is the most telling about this unit's function, but instead the *set of images* that elicit a strong response. While some previous work has visualized multiple 'facets' of neurons' selectivity, these efforts focused mostly on the highest layers of the network and relied on initialization or random sampling strategies to create multiple images for each unit. However, as we show in the present paper, these approaches underestimate the true diversity of the selectivity of even relatively low-level units. Additionally, these approaches have not offered insights about how the representations of different networks trained on the same task compare. Our contributions are the following:

1. Motivated by the phase invariance of complex cells in the early visual system of the brain, we show why visualizing invariance is as important as visualizing selectivity for understanding the computations of even low-level units.
2. We develop a non-parametric approach to map the manifold of highly-activating inputs as exhaustively as possible.
3. We show that even relatively low-level units exhibit a remarkable degree of invariance in VGG-19 [28], which is not revealed by finding highly activating stimuli from multiple optimization runs with random initializations.
4. We find that in low to intermediate layers of VGG-19, at least two types of invariances emerge: tolerance to local diffeomorphic transformations tuned to specific features, and phase invariance, where units respond well to periodic texture patterns and are insensitive to their phase. We additionally offer a way to quantify these invariances.
5. In contrast, we find that low to intermediate layers of a network with skip connections (ResNet-50 [11]) that was trained on the same task as VGG-19 exhibit far less phase invariance, revealing representational differences between these two networks.
6. We showcase our visualization approach on a CNN trained to predict responses to natural images in primary visual cortex of the primate brain.

We provide the code to replicate our results. [1]

2 Related work

One way to identify selectivity of hidden units is to look for image patches in the dataset that drive them maximally [6,33]. These image patches can sometimes hint at a unit's selectivity, but it can be difficult to identify their common features. Optimization-based techniques have proven more useful for feature visualization: a common approach is to search for pre-images that drive individual neurons maximally via gradient ascent [6]. Most previous work focused on deep layers, where finding natural-looking pre-images is challenging. For example, the activation objective leads to adversarial-like patterns [20,29]. As a consequence, much of the follow-up work focused on developing regularization techniques to obtain more natural pre-images, including penalties on high-frequency

[1] https://github.com/sacadena/diverse_feature_vis

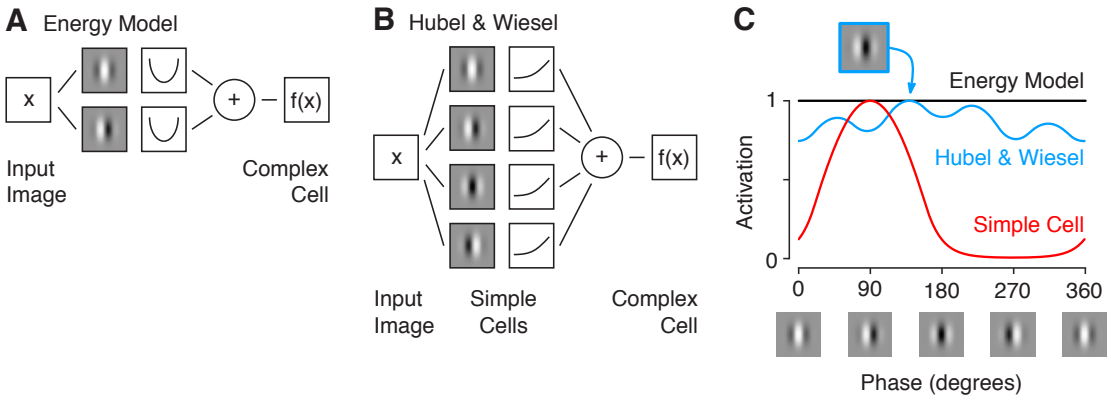


Fig. 1. Simple and complex cells, phase invariance. **A.** Energy model of complex cell. **B.** Hubel & Wiesel model of complex cell. **C.** Neural response as a function of phase of Gabor stimulus with optimal orientation and spatial frequency.

noise [16,20] or the distance between the generated visualizations and natural images patches [32], or performing gradient descent in the feature space of a deep generator network [19].

Goodfellow et al. [9] were the first (to our knowledge) to study invariances in deep networks. Their approach allows to quantify how invariant a unit is to known transformations such as translation, (3D-) rotation or scaling, but it does not allow to discover these transformations if they are unknown in advance.

Recent work proposes visualizing multiple 'facets' of the neuron's selectivity by obtaining multiple images from different random initializations [17], using a diverse set of highly activating images as initializations [21], or using a generative image model to sample highly-activating images [18].

These methods do not explicitly specify an objective to produce a diverse set of images. In contrast, we optimize a batch of images to drive the neuron of interest strongly while simultaneously being as distinct from each other as possible. Recent concurrent work [22] introduces a similar idea, albeit with a different loss function based on texture representations [7,8].

3 Discovering invariances

3.1 Motivation: simple and complex cells

We illustrate our point by considering a toy example well known from early vision in the brain (Fig. 1): simple and complex cells [12], which are found in the primary visual cortex, an early stage of visual processing in the mammalian brain. Simple cells can be approximated well by a linear filter followed by a thresholding nonlinearity (e.g. ReLU). The linear filter usually resembles a Gabor filter. Complex cells are, like simple cells, selective for a specific orientation and spatial frequency. However, unlike simple cells they respond to Gabor patches of arbitrary phases – they are phase-invariant. The standard model for this phase invariance is the so-called energy model (Fig. 1A, [1]), which sums over

the squared responses of two Gabor filters phase-shifted by 90° (Fig. 1C, black). This energy model has also been used to study rotation, scaling and more general invariances in the context of unsupervised representation learning [2,3,15]

An alternative formulation was originally proposed by Hubel & Wiesel, who discovered complex cells in the 1960ies in the primary visual cortex of cats [12]. Their model suggests that complex cells are the result of pooling over multiple simple cells with a range of phase preferences (Fig. 1B). If the learned weights and phase preferences exhibit some variability, the resulting phase invariance is only approximate (Fig. 1C, blue).

Now, consider what happens when we study simple and complex cells using activity maximization. For a simple cell, we will recover its selectivity. For a complex cell, however, all Gabor patches of optimal orientation and spatial frequency will elicit a high response, irrespective of their phase. In the case of the Energy Model, which is perfectly phase-invariant, we may obtain this set of optimal images by starting with random initializations. However, for an imperfect model more likely to occur in reality (e. g. Hubel & Wiesel model, blue in Fig. 1C), there is a unique maximum, which we will find despite the fact that activations are consistently above 80% of the maximum for all phases. Thus, activity maximization will produce the same result for both simple and complex cells (a single Gabor patch), but this result will miss the key aspect of the complex cell's computation: its phase invariance.

3.2 Mapping invariances

Objective. The idea behind our approach is to find a batch of images in which each image maximally drives a specific unit while the images are maximally different from one another. Starting with a batch of n images $\{x_1, \cdots, x_n\}$, initialized as white noise, we *maximize* the following objective using gradient ascent:

$$L = \sum_{i=1}^{n} y_{ik}^{(l)} + \alpha \sum_{i=1}^{n} \log P(x_i) + \lambda \min_{i,j} d(x_i, x_j). \tag{1}$$

Here, $y_{ik}^{(l)}$ is the output activation of unit k in layer l for the i^{th} image in the batch, $P(x_i)$ is the likelihood of the image under a generative model of natural images and $d(x_i, x_j)$ is a distance between two images, The likelihood and distance measures are specified below. Note that we set the image size to the receptive field size of units in the layer to be visualized, such that the outputs $y_{ik}^{(l)}$ are 1×1 spatially and we can omit the indices over space. We constrain the norm of the synthesized images to be equal to half the average norm of natural images patches of the same size taken from the ImageNet dataset[2], where we assume that zero in each color channel corresponds to the average value of this channel

[2] Using half the average norm is a heuristic that we use because the synthesized images tend to be localized to the center of the patch.

across the ImageNet training set. For visualization, we add this mean and clip the values between 0 and 255. Very few pixels fall outside this range.

The first and the second term in the objective are similar to previous work, encouraging the optimization to find natural images that strongly activate the unit. The third term forces all images in the batch to be as distinct as possible from all other images, since we penalize the minimum distance between any pair of images. This objective presents a trade-off: we allow for some degree of non-maximal responses if this allows us to increase the set of strongly activating pre-images substantially.

It is important to use the minimum distance in the objective rather than the average. Maximizing the average distance does not necessarily lead to coverage of the invariant subspace. Consider the Energy Model: assuming we generate an even number of n images, the optimal solution maximizing the average L_2 distance is to place all images at either of two distinct phases separated by 180°. Now we fail to generate a diverse set of images but the average distance is high (90°). In contrast, the desired solution of images evenly separated by $360°/n$ will give a smaller average distance for $n > 4$ and can be obtained when maximizing the minimum distance.

It has also some advantages to consider a single unit within a feature map compared to considering the entire feature map. When maximizing the activation of the entire feature map, the resulting image will be shift-invariant by construction and properties such as phase invariance of individual units cannot be detected.

Natural image prior. We use PixelCNN++ [27] as a natural image prior, as it allows directly evaluating and optimizing the likelihood of an image patch of arbitrary size. In a nutshell, PixelCNN++ improves upon PixelCNN [23] and earlier autoregressive models [24,30,31] that attempt to capture the distribution of natural images by expressing the joint distribution of all pixels as the product of the distributions of individual pixels conditioned on a causal neighborhood. We use the model pre-trained on Cifar-100 provided by OpenAI[3] which is state-of-the-art in terms of likelihood on natural images.

Distance metric. To evaluate the distance between two images, we use a feature space given by the neural network to encourage diversity on perceptually interesting image properties. For an output unit y_k in layer l, we compute the Euclidean distance in the feature space of the preceding convolutional layer:

$$d(x_i, x_j) = \|\mathbf{y}_i^{(l-1)} - \mathbf{y}_j^{(l-1)}\|_2; \quad i \neq j \tag{2}$$

where $\mathbf{y}_i^{(l-1)}$ and $\mathbf{y}_j^{(l-1)}$ are vectors of activations in the preceding layer flattened over space and channels.

[3] https://github.com/openai/pixel-cnn

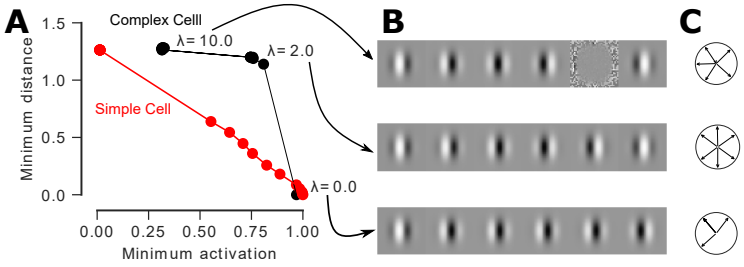


Fig. 2. Mapping invariances as a trade-off between diversity and maximizing activation. **A.** Trade-off between activation and image diversity. For a complex cell, images can be made quite diverse while keeping the activation level high. When λ gets too large ($\lambda > 2$), there is a qualitative change. **B.** Set of images for three different λ. **C.** Distribution of phases of synthesized Gabor patches, showing that with the optimal $\lambda = 2$ we get equally spaced images, i.e. cover the invariant subspace well.

Optimization. We optimize the objective defined in Eq. (1) using the Adam optimizer [13] with a learning rate of 0.1 until the objective converges (maximum of 1000 steps). Similar to Olah et al. [22], we precondition the gradient to reduce the effect of high frequencies by dividing each frequency component by $\sqrt{f}$.

We manually set the hyperparameter α, which controls the strength of the natural image prior, based on qualitative inspection of the resulting images in an exploratory experiment. We used $\alpha = 0.0005$ for all experiments.

We sweep a range of values for λ (0.02, 0.04, 0.08, ... 20.48) and for each unit pick the largest such λ that the average activation level remains above a threshold. This threshold is 80% of the maximum for the complex cell model and 90% for VGG-19 and ResNet-50. See Fig. 2A and Fig. 4 for a qualitative justification of these thresholds.

3.3 Application to complex cell models

Before applying our approach to a deep neural network, we verify that it works when the units are only approximately invariant to some transformation. To this end, we use the Hubel & Wiesel model of a complex cell outlined above (Fig. 1B), which does not produce perfect phase invariance, but still responds strongly to Gabor patches of all phases.

Indeed, our approach can visualize the entire invariant subspace spanning the full range of phases (Fig. 2). Without the diversity term ($\lambda = 0$), the optimization tends to converge to the same pre-image (Fig. 2B). Four out of six solutions correspond to the globally most strongly driving image (see also Fig. 1C, top). In contrast, with an appropriate choice of λ, the images distribute uniformly (Fig. 2B, C). If we increase λ too much, however, the diversity penalty becomes too large and the optimization will converge to solutions including non-optimal images. Thus, to visualize the invariant subspace, we should pick the largest λ that leads to only a small decrease in activation level. This point depends on how 'clean' the invariance of the cell is. For the Hubel & Wiesel model considered

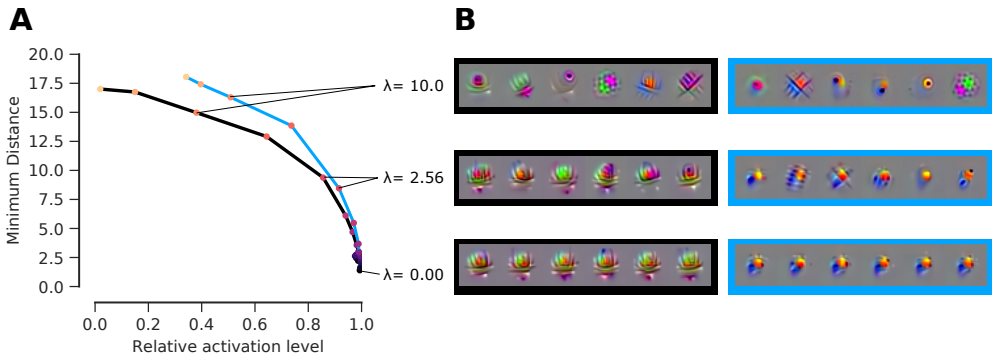


Fig. 3. Invariant subspace of two example units in conv3_2 (feature maps 9 and 26). **A.** Invariance/activation trade-off **B.** Pre-images obtained for different values of λ.

here, this drop in activation occurs when the average activation falls below 80% of the maximum, which corresponds to the response range for images within the approximately invariant subspace (see Fig. 1C, blue line).

Note that for the simple cell, which does not exhibit any such response invariance, the curve looks qualitatively different (Fig. 2A, red line). Thus, we can quantify response invariance of units in a DNN by computing the minimum distance between any two images in the batch at the optimal λ.

4 Invariances in VGG-19

We asked to what extent deep neural networks trained on large-scale object recognition (ImageNet [25]) exhibit response invariances in their convolutional layers. Previous work focused mostly on higher layers and did not find much invariance in low and intermediate layers. However, in neuroscience it is well-known that low- and mid-level neurons in the brain – like complex cells – can exhibit a substantial degree of response invariance. Moreover, there is evidence for a considerable degree of similarity between neural representations in DNNs trained on object recognition and the primate visual system [14,10,4,5]. In particular, we have shown [4] that the convolutional layers of VGG-19 [28] around layer conv3_1 best predict neural activity in primary visual cortex, including that of many complex cells. Therefore we would expect that these layers in the VGG-19 network should also exhibit some degree of invariance to phase and potentially other transformations.

4.1 Convolutional layers of VGG-19 exhibit response invariances

We start by considering two example units from layer conv3_2 (Fig. 3) of VGG-19. As in the complex cell example, we can increase the diversity of generated images quite substantially while maintaining a high activation level (Fig. 3A). Only when we increase λ too much, the activation level drops substantially and the images start deteriorating (Fig. 3B, top row). Overall, the trade-off between

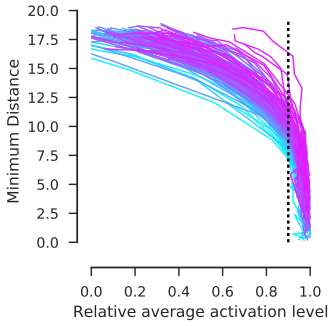


Fig. 4. Invariance/activity maximization trade-off for all units in layer conv3_1. Based on visual inspection we deem 90% an appropriate threshold and use the largest λ such that the average activation remains above 90% of the maximum activity.

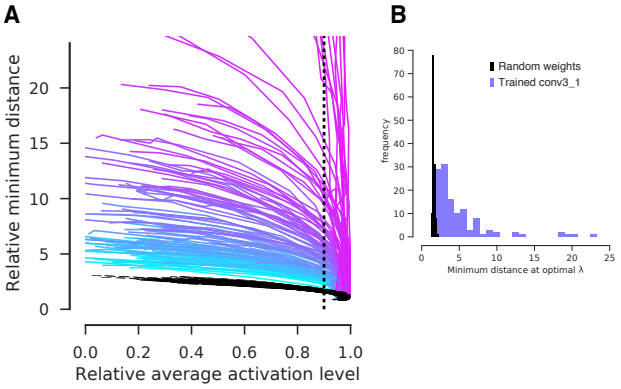


Fig. 5. VGG units are more invariant than expected from random weights. **A.** All 256 units in conv3_1 (colored lines) are more invariant than units in a network with the same architecture but random weights (black lines). **B.** Histogram of the diversity terms for the optimal λ relative to their value for $\lambda = 0$ (black: random weights; purple: trained conv3_1). This means that for the least invariant units we can increase the diversity of the images two-fold while maintaining the average activation above 90% of the maximum obtained with $\lambda = 0$.

image diversity and activation level looks qualitatively similar to the complex cell example above.

Moreover, the images generated with the optimal λ look significantly more diverse than those obtained by random initialization at $\lambda = 0$ (Fig. 3B, middle and bottom rows). Indeed, most units showed quite some degree of invariance: we can increase the image diversity considerably while maintaining activation levels above 90% of the maximum (Fig. 4 for conv3_1; see Sect. 1 in the Supp. for additional convolutional layers). Below, we therefore use the largest such λ that maintains the average activation level above 90% of the maximum.

4.2 Response invariances are a learned property of the network

Is this invariance a learned property of the network or does it arise trivially from the network architecture? We repeated the analysis on a network with the same architecture as VGG-19 but random weights. To keep the two networks comparable, we normalized both the activations and the distances between images such that they are equal to one for $\lambda = 0$. We found that units in the random network are substantially less invariant than those of VGG-19 (Fig. 5A), suggesting that the neurons' response invariance is indeed a learned property. Remarkably, by introducing the diversity term into the pre-image search, we could increase the minimum distance between any two images in a batch by a factor of at least two and up to 100-fold without 'sacrificing' more than 10% of the unit's activation level (Fig. 5B), a property that the random network does not exhibit.

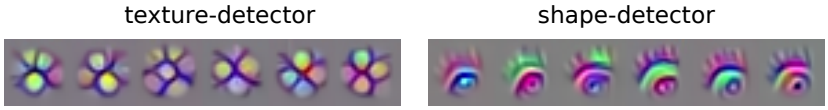


Fig. 6. Examples of invariant subspaces of texture-like and shape-like detectors of feature maps 13 (left) and 22 (right) in conv3_1.

4.3 Types of invariance: texture vs. shape detectors

We now investigate the types of invariance learned by different units in the network. We start by considering two example units from layer conv3_1 (Fig. 6). The first unit responds to a dark grid on brighter background of arbitrary color. In addition to this selectivity, it appears to be entirely phase- and rotation-invariant: the location of the grid lines and their orientation is irrelevant for the unit's activation, but their general spatial scale and the foreground color are important. We refer to units that exhibit this property as *texture* detectors.

The second unit, in contrast, detects a circular feature in the lower half of its receptive field. While it is sensitive to the location of this pattern within its receptive field, it exhibits a substantial degree of color and scale invariance: the contours have a sinusoidal cross-section whose local phase varies across images, such that by using linear combinations of multiple of these images one can obtain the circular pattern in various different sizes and color combinations. We refer to such units as *shape* detectors: they are sensitive to location but allow for some degree of local diffeomorphic transformation.

The two units shown here are representative of a larger number of units in various layers of VGG-19 (see Fig. 7 and Sect. 2 from Supp. for more examples). As we will quantitatively show below, they lie on two extremes of a spectrum along which we can characterize low- and intermediate-level units.

4.4 Quantification of phase invariance (textures)

So far, we have described texture and shape units only qualitatively. We therefore developed metrics to quantify these properties more systematically. We start by quantifying phase invariance, the property that characterizes texture detectors.

While shift *equivariance* is built into CNNs, phase *invariance* of individual units has to our knowledge not been reported. A perfectly phase-invariant unit would maintain a high activation when presented with shifted versions of its preferred texture. Therefore, to quantify phase invariance, we optimize an image twice as large as the unit's receptive field such that the average activation of all possible windowed crops from this image is maximized (Fig. 8A, 1–4). Indeed, for a decent number of units we had qualitatively labeled as 'texture detectors,' the crops generated in this way (Fig. 8A, 3) resemble the templates we synthesized earlier (Fig. 8A, 4) and elicit similarly high activations (Fig. 8C). On the other hand, 'shape-selective' units expect certain structures in specific locations within their receptive field. Generating a texture where arbitrary crops are highly activating is not possible for these units (Fig. 8B).

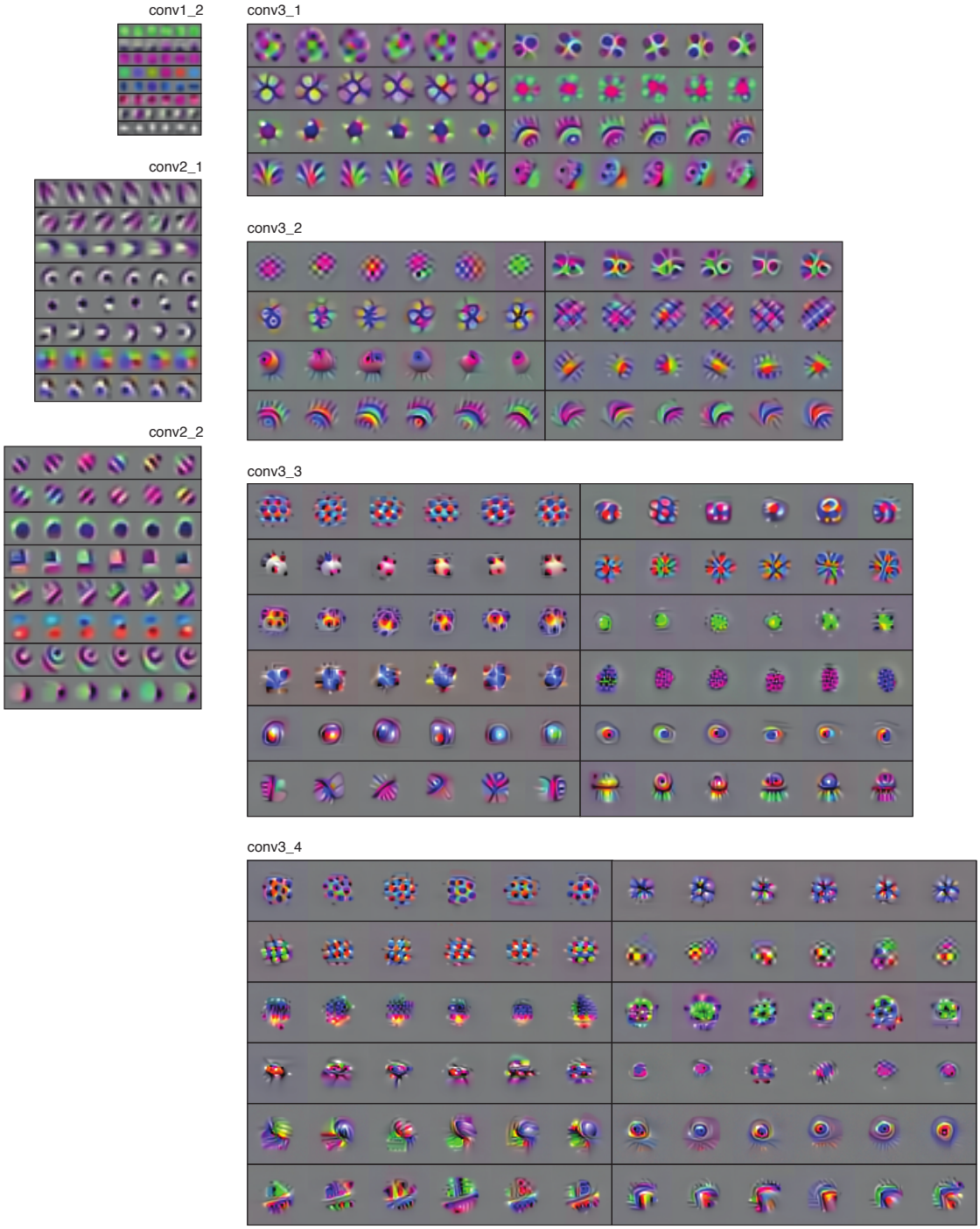


Fig. 7. Invariant subspaces of a selection of units in convolutional layers conv1_2 to conv3_4 of VGG-19. Each horizontal block of six images represents one unit. It contains the six maximally diverse images resulting in an activation of the unit above 90% of its maximum. Images for higher layers are scaled up slightly to improve visibility, but the pixel sizes are not matched across layers (lower layers have comparably larger pixels).

To quantify this intuitive argument, we defined shift invariance as the ratio between the average activation of all crops from the larger texture and the aver-

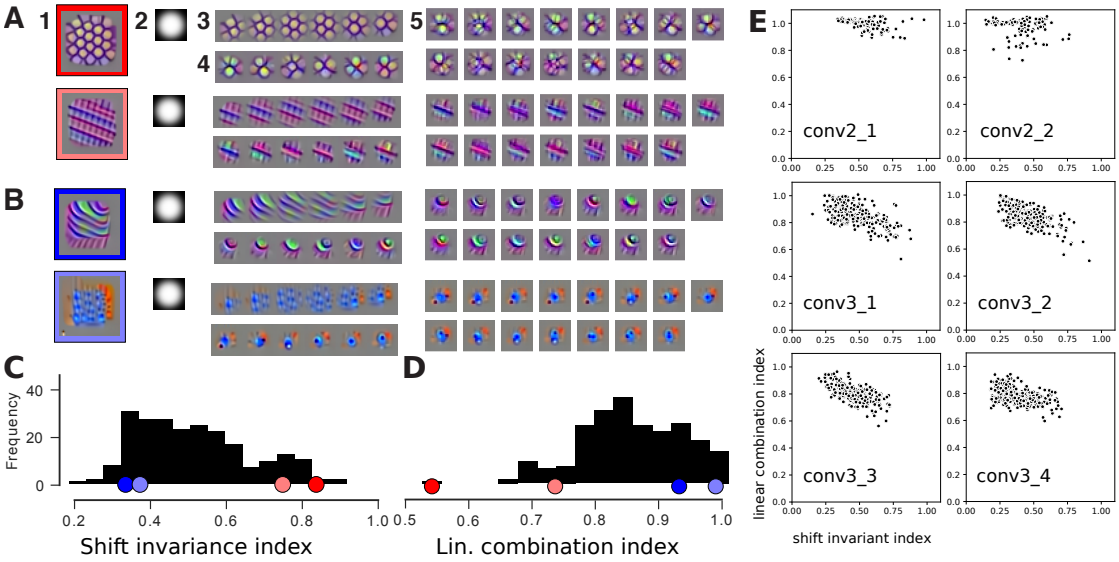


Fig. 8. Quantification of invariances in VGG19. For layer conv3_1, examples of texture (**A**) and shape (**B**) units. **Left:** Phase invariance. We optimize a texture (**1**) to maximize the average activation of all windowed (**2**) crops (**3**). The mask has the form $\exp\left(-\left(r/\sigma\right)^4\right)$ where $r = \sqrt{x^2 + y^2}$. We picked σ so that the ratio between the unit's receptive field and σ is ~ 2.5. (**4**): individual images maximizing the unit's activation. **Right:** Invariance to local deformations is supported by features that locally form quadrature pairs. Linear combinations (**5**) of templates (**4**) produce images with high activations. **C.** Histogram of the phase invariance (examples from A+B labeled). **D.** Histogram of metric measuring invariance to local deformations. **E.** Scatter plot of the two metrics (shift invariant index and linear combination index) for all units at each convolutional layer of VGG19.

age activation of the diverse templates produced earlier (see example histogram in Fig. 8C, for conv3_1). Indeed, the units labeled as phase-invariant (Fig. 8A), maintain a high activations despite arbitrary phase shifts, while the activation of the shape-selective units (Fig. 8B) drops substantially (Fig. 8C).

Note that synthesizing a larger image by maximizing all crops is similar to maximizing an entire channel's activity (i. e. feature map) for a sufficiently large input image, an approach other authors have taken for feature visualization [22]. Although insightful in many occasions, the drawback is that this procedure often occludes shape selectivity. For instance, the first unit in Fig. 8B is selective to a circular pattern in the top-right with rays pointing towards the bottom-left when maximized individually. However, the resulting texture looks like a field of oriented edges, thus missing the crucial pattern that drives this unit.

4.5 Tolerance to local deformations (shapes)

The second invariance we identify is tolerance to local deformations. A closer look at some examples (e. g. Fig 6, right; Fig. 8B, top) reveals that some of the units have local tolerance for phase changes. The patterns these units are tuned for

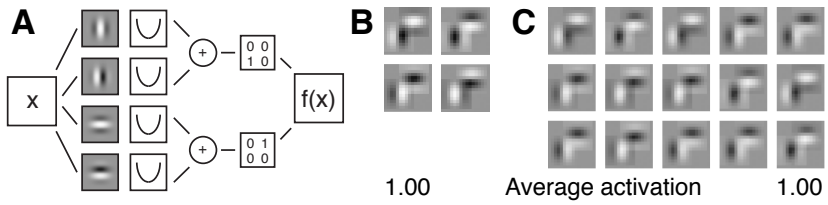


Fig. 9. Toy example (**A**) where linear combinations (**C**) of highly activating images (**B**) are also highly activating. It detects a top-left corner by combining two complex cells.

can be locally built by spatially arranging multiple complex-cell-like quadrature pairs. This would suggest, that – although mapped into a nonlinear feature space – linear combinations of the 'template' images spanning the invariant subspace should highly activate these units as well. We illustrate this seemingly counter-intuitive hypothesis with a toy example and then show how it applies to CNNs.

Consider the following example comprised of two complex cells arranged such that they detect a top-left corner (Fig. 9). The unit allows for individually shifting up or down the horizontal edge, and left or right the vertical edge. Each of the two edges is detected by an energy model of a complex cell (Fig. 9A), each at a defined location within the receptive field. Accordingly, the highly activating template images are made up of combinations of odd and even Gabors (Fig. 9B) and any linear combination of them is again a highly activating image (Fig. 9C).

To quantify whether the same property holds for VGG units, we computed the average activation level of linear combinations of the maximally activating images. Specifically, we took the averages (in pixel space) of all 15 pairs of templates (Fig. 8A.5), renormalized them to the same norm as the templates and compared their average activation to that of the templates. For 'texture-selective' units this procedure deteriorates the clear texture patterns revealed by the templates (see for instance Fig. 8A.5). Accordingly, the unit's activation level to these images drops substantially (Fig. 8D, red+orange). We quantify this drop by computing a linear combination index, defined as the ratio between the average activation of average-image pairs and the average activation of the diverse templates. Units tuned to shape patterns that are tolerant to local transformations give average-pairs that are fairly similar to the original templates, producing a high linear combination index.

4.6 Characterization of invariances across layers

We have identified two metrics that quantify two different forms of invariance in VGG units. Our examples from Fig. 8 suggest that these two types of invariance are anticorrelated. As this does not have to be the case a priori – a complex cell would score high on both metrics – we asked whether this was just due to our selection of examples or whether it holds more generally across layers. Indeed, shift invariance and tolerance to local deformations appear to be anticorrelated across a wide range of layers (Fig. 8E; conv3 in particular). We also observe that

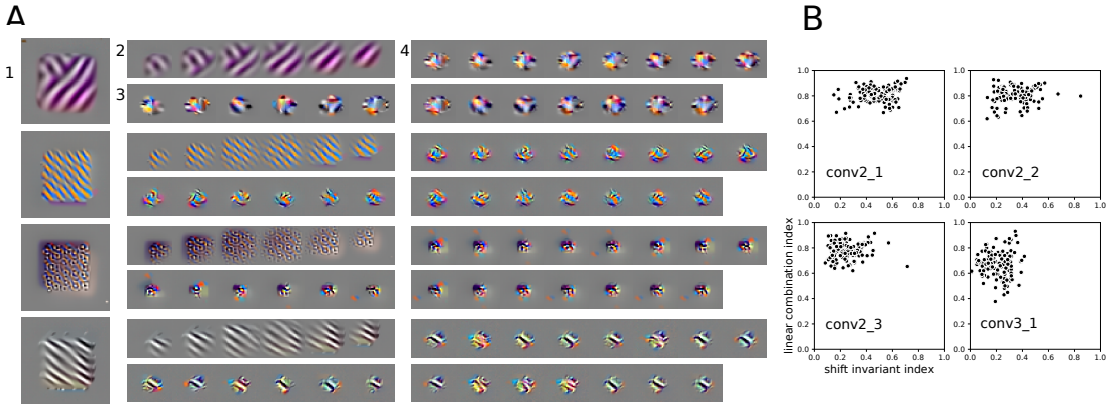


Fig. 10. ResNet-50 results. **A**, Example units of block conv2_3 (compare to Fig.8). We noticed that maximizing windowed crops (2) of a big texture (1) are largely different from the maximizing templates (3). 15 template-pair averages (4) are on the other hand highly activating and similar to the templates. **B**, Scatter plot of the two metrics proposed for said layers of ResNet-50.

higher layers tend to be less shift-invariant than lower ones (e. g. compare within conv3 in Fig. 8E).

5 Diverse visualizations of early layers of ResNet-50

To test whether our results so far are properties of VGG-19 or apply more generally to CNNs trained on ImageNet, we also applied our methods to ResNet-50 [11]. We considered its early layers up to conv3_1 (fourth block), which have receptive field sizes comparable to the layers we studied in VGG-19. We first synthesized diverse image batches with different diversity penalties and found a similar trade-off between activation and diversity as found before (see Sect. 3 in Suppl. Material). However, for the λ that evoked at least 90% of the maximal responses we observed on average a smaller diversity compared to that of VGG-19 units. We then ran our analysis to identify both phase and shape invariance and surprisingly found a much reduced number of phase-invariant units compared to VGG-19 (Fig. 10): there are basically no ResNet-50 units for which the crops from the optimal texture look like the optimized templates (e.g. Fig. 10A,2+3). On the other hand, template-pair averages do not appear to qualitatively deviate from the synthesized templates (Fig. 10A,4) indicating a strong presence of tolerance to local changes. The two metrics introduced above confirm this observation quantitatively: the distribution of shift invariance indices is shifted towards zero in ResNet-50 layers (Fig. 10B) with respect to those in VGG-19.

This is a very interesting finding, because it shows that the different architectures learn quite different features in their early layers despite both being trained on ImageNet and achieving comparable classification accuracy. Thus, our novel approach to feature visualization helped us identify strong representational dif-

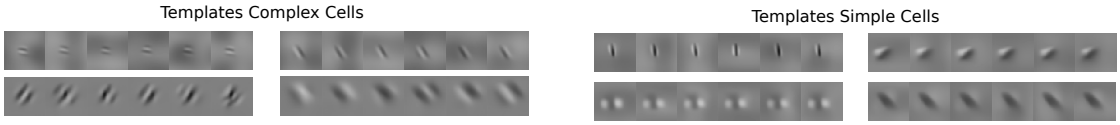


Fig. 11. Subspaces of V1 cells. Complex (left) and simple (right) cells

ferences in the canonical directions between two architectures that would not have been observed with conventional activity maximization

6 Phase invariance in Primary visual cortex (V1)

As a final practical use case, we applied our method to a three-layer CNN that has been trained to predict neural responses in V1 when monkeys are shown natural images (data from [4]; see also their Fig. 3). Our method unveils the known cell types – simple: phase-selective and complex: phase-invariant (Fig 11). Although complex cells can also be identified using specifically designed stimuli or analysis methods relying on quadratic features (e. g. spike-triggered covariance [26]), our non-parametric approach could in principle also uncover other types of invariance that are not captured by quadratic features. Given that we see no such additional invariances, there are likely no other major features V1 cells are invariant to – a conclusion that could not be drawn using parametric approaches.

7 Conclusion

Motivated by early vision in the brain, we investigated the response invariances in the early to intermediate convolutional layers of DNNs. We found that units in early layers of VGG-19 show invariance to global texture-preserving transformations and invariance to local shape-preserving transformations. In contrast, ResNet-50 does not exhibit the same degree of shift invariance. This difference could explain why practitioners working on texture synthesis and style transfer observe that the features of VGG work substantially better than those of more modern architectures such as residual networks.

We conclude that these methods not only give new insights into the computations performed by DNNs and how they compare with other architectures, but also constitutes an important step towards a unified language for describing neural representations in both biological and computer vision.

Acknowledgements. We thank Jonas Rauber and Andreas Tolias for useful discussions. This work was supported by the German Research Foundation (DFG) grant EC 479/1-1 to A.S.E. The International Max Planck Research School for Intelligent Systems (IMPRS-IS) supported S.A.C. The work was also supported by IARPA via Department of Interior (DoI) contract D16PC00003.

References

1. Adelson, E.H., Bergen, J.R.: Spatiotemporal energy models for the perception of motion. J. Opt. Soc. Am. A **2**(2), 284–299 (1985). https://doi.org/10.1364/JOSAA.2.000284
2. Berkes, P., Wiskott, L.: Slow feature analysis yields a rich repertoire of complex cell properties. Journal of vision **5**(6), 9–9 (2005)
3. Bethge, M., Gerwinn, S., Macke, J.H.: Unsupervised learning of a steerable basis for invariant image representations. In: Human Vision and Electronic Imaging XII. vol. 6492, p. 64920C. International Society for Optics and Photonics (2007)
4. Cadena, S.A., Denfield, G.H., Walker, E.Y., Gatys, L.A., Tolias, A.S., Bethge, M., Ecker, A.S.: Deep convolutional models improve predictions of macaque v1 responses to natural images. bioRxiv (2017). https://doi.org/10.1101/201764
5. Cadieu, C.F., Hong, H., Yamins, D.L., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J.: Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS computational biology **10**(12), e1003963 (2014), 00152
6. Erhan, D., Bengio, Y., Courville, A., Vincent, P.: Visualizing higher-layer features of a deep network. Tech. Rep. 1341, University of Montreal (Jun 2009), also presented at the ICML 2009 Workshop on Learning Feature Hierarchies, Montréal, Canada.
7. Gatys, L., Ecker, A.S., Bethge, M.: Texture synthesis using convolutional neural networks. In: Advances in Neural Information Processing Systems. pp. 262–270 (2015)
8. Gatys, L.A., Ecker, A.S., Bethge, M.: Image style transfer using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pp. 2414–2423 (2016)
9. Goodfellow, I., Lee, H., Le, Q.V., Saxe, A., Ng, A.Y.: Measuring invariances in deep networks. In: Advances in neural information processing systems. pp. 646–654 (2009)
10. Güçlü, U., van Gerven, M.A.J.: Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. Journal of Neuroscience **35**(27), 10005–10014 (2015). https://doi.org/10.1523/JNEUROSCI.5023-14.2015
11. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 770–778 (2016)
12. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of physiology **160**(1), 106 (1962), 09139
13. Kingma, D.P., Ba, J.: Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
14. Kriegeskorte, N.: Deep neural networks: A new framework for modeling biological vision and brain information processing. Annual Review of Vision Science **1**(1), 417–446 (2015). https://doi.org/10.1146/annurev-vision-082114-035447
15. Lies, J.P., Häfner, R.M., Bethge, M.: Slowness and sparseness have diverging effects on complex cell learning. PLoS computational biology **10**(3), e1003468 (2014)
16. Mahendran, A., Vedaldi, A.: Understanding deep image representations by inverting them. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 5188–5196 (2015)

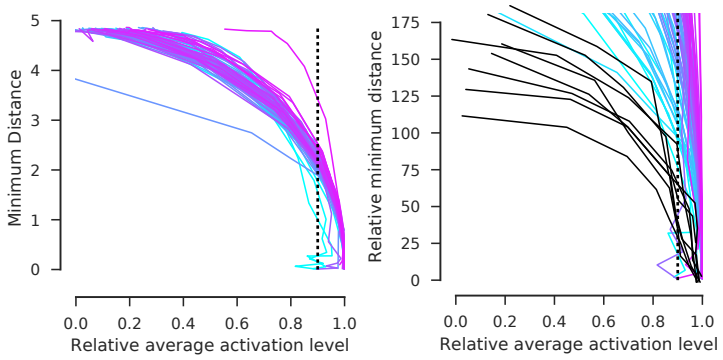
17. Mahendran, A., Vedaldi, A.: Visualizing deep convolutional neural networks using natural pre-images. International Journal of Computer Vision **120**(3), 233–255 (2016)
18. Nguyen, A., Clune, J., Bengio, Y., Dosovitskiy, A., Yosinski, J.: Plug & play generative networks: Conditional iterative generation of images in latent space. In: CVPR. vol. 2, p. 7 (2017)
19. Nguyen, A., Dosovitskiy, A., Yosinski, J., Brox, T., Clune, J.: Synthesizing the preferred inputs for neurons in neural networks via deep generator networks. In: Advances in Neural Information Processing Systems. pp. 3387–3395 (2016)
20. Nguyen, A., Yosinski, J., Clune, J.: Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In: The IEEE Conference on Computer Vision and Pattern Recognition (June 2015)
21. Nguyen, A.M., Yosinski, J., Clune, J.: Multifaceted feature visualization: Uncovering the different types of features learned by each neuron in deep neural networks. Visualization for Deep Learning workshop, ICML (2016)
22. Olah, C., Mordvintsev, A., Schubert, L.: Feature visualization. Distill (2017). https://doi.org/10.23915/distill.00007
23. van den Oord, A., Kalchbrenner, N., Espeholt, L., Vinyals, O., Graves, A., et al.: Conditional image generation with pixelcnn decoders. In: Advances in Neural Information Processing Systems. pp. 4790–4798 (2016)
24. Oord, A.v.d., Kalchbrenner, N., Kavukcuoglu, K.: Pixel recurrent neural networks. arXiv preprint arXiv:1601.06759 (2016)
25. Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., Fei-Fei, L.: ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision (IJCV) **115**(3), 211–252 (2015). https://doi.org/10.1007/s11263-015-0816-y
26. Rust, N.C., Schwartz, O., Movshon, J.A., Simoncelli, E.P.: Spatiotemporal elements of macaque v1 receptive fields. Neuron **46**(6), 945–956 (2005)
27. Salimans, T., Karpathy, A., Chen, X., Kingma, D.P., Bulatov, Y.: Pixelcnn++: A pixelcnn implementation with discretized logistic mixture likelihood and other modifications. In: Submitted to ICLR 2017 (2016)
28. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014), `http://arxiv.org/abs/1409.1556`
29. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., Fergus, R.: Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199 (2013)
30. Theis, L., Bethge, M.: Generative image modeling using spatial lstms. In: Advances in Neural Information Processing Systems. pp. 1927–1935 (2015)
31. Theis, L., Hosseini, R., Bethge, M.: Mixtures of conditional gaussian scale mixtures applied to multiscale image representations. PloS one **7**(7), e39857 (2012)
32. Wei, D., Zhou, B., Torrabla, A., Freeman, W.: Understanding intra-class knowledge inside cnn. arXiv preprint arXiv:1507.02379 (2015)
33. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: European conference on computer vision. pp. 818–833. Springer (2014)

8 Supplementary

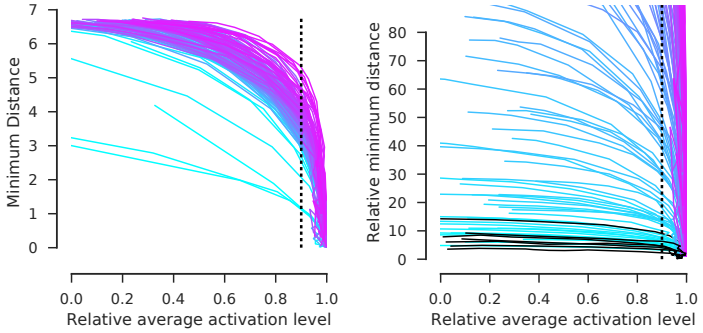
8.1 Diversity/activation maximization trade-off VGG19

As in Figures 4 and 5, we show here the trade-off between diversity and activity maximization for all layers including the natural image prior. Diversity is measured as the minimum L_2 distance in feature space between all pairs of synthesized templates. Each curve represents a unit (feature map) of the corresponding layer. The curves connect the average of three optimization runs for a choice of λ from Equation 1. The penalty for the natural image prior α was set to 0.0005 after visual inspection. The curves were normalized to the maximum sum of activations (relative average activation level). On the left: The trade-off between minimum distance and relative average activation. On the right: The same curves normalized to have a unit minimum distance. This facilitated comparison with the network with random weights (black). Here, we show in black a sample of units from a random network with the same architecture as VGG-19. Note that the VGG units exhibit more invariance at each layer than expected from random weights for all studied layers.

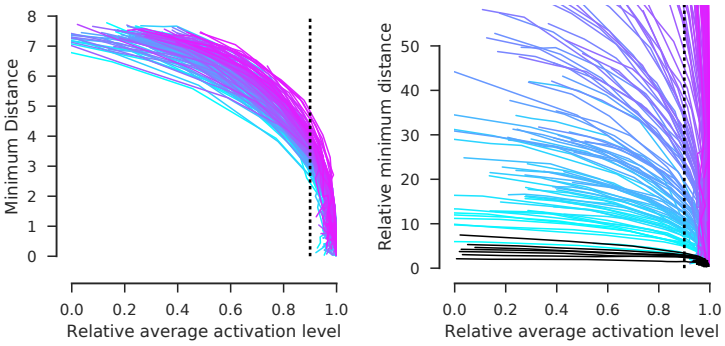
conv1_2 .



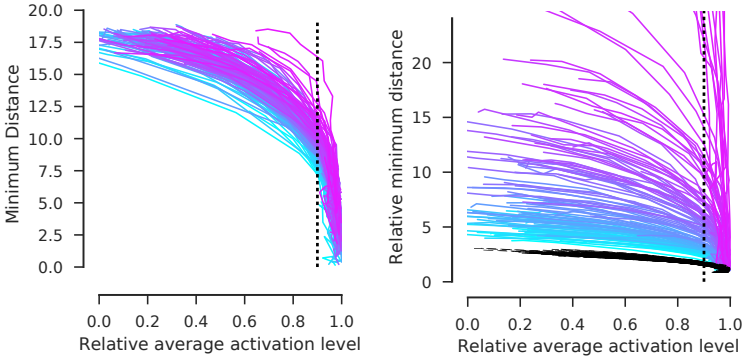
conv2_1 .



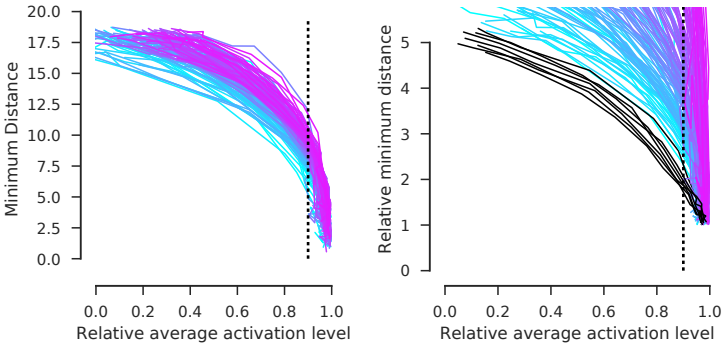
conv2_2 .



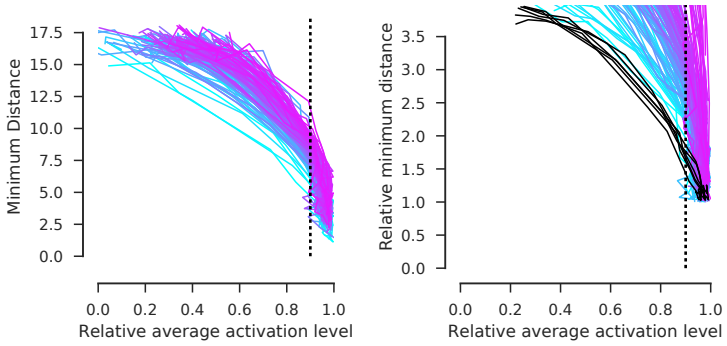
conv3_1 .



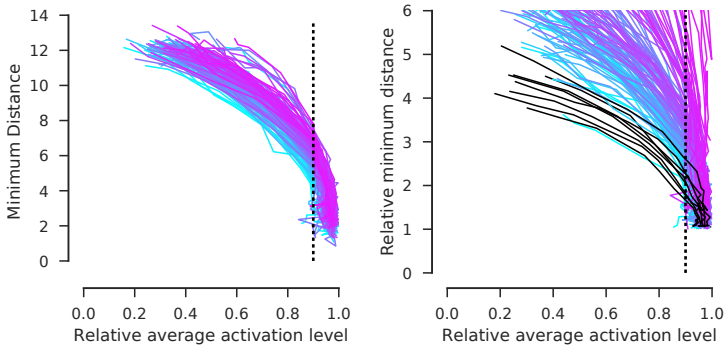
conv3_2 .



conv3_3 .

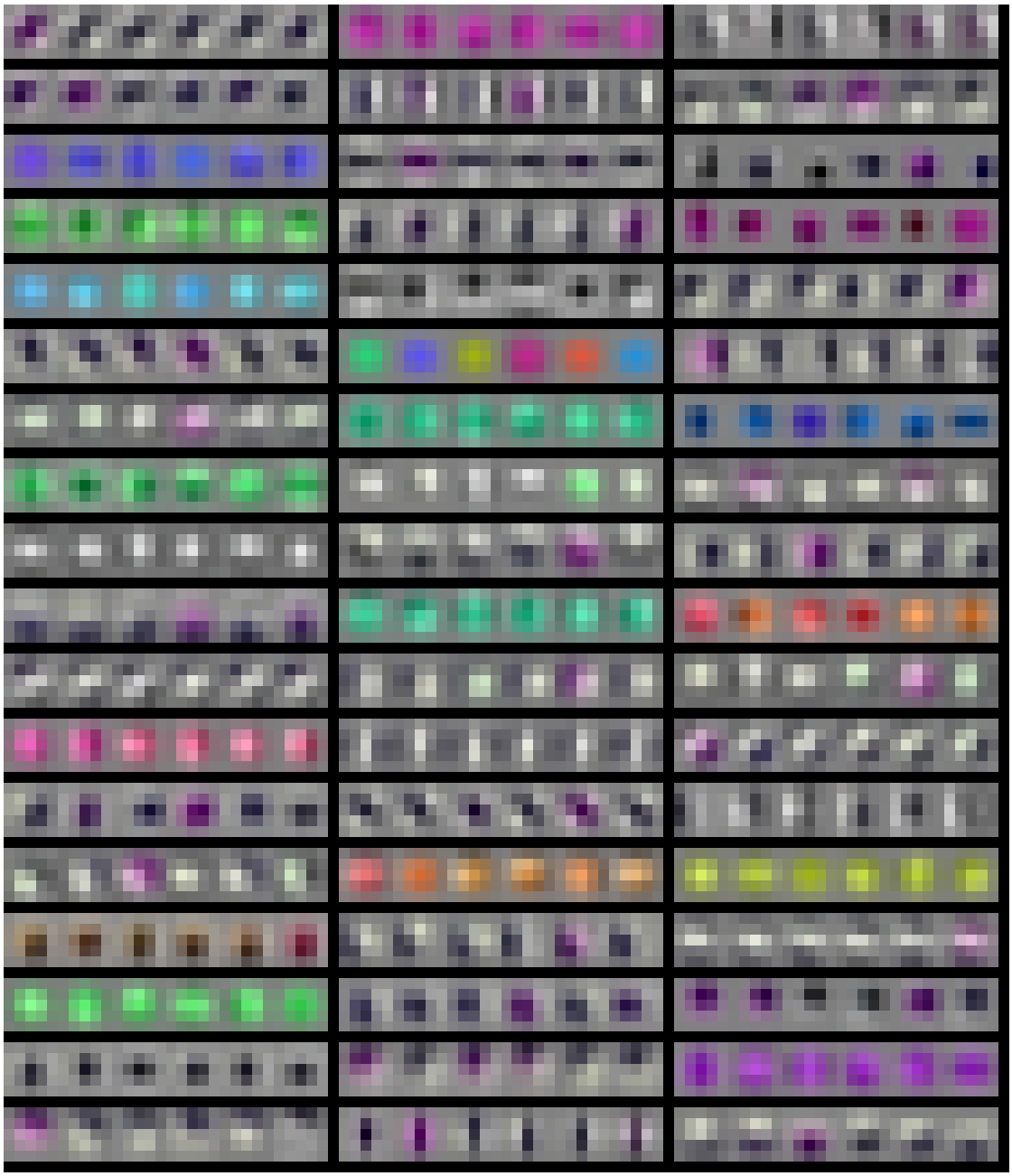


conv3_4 .

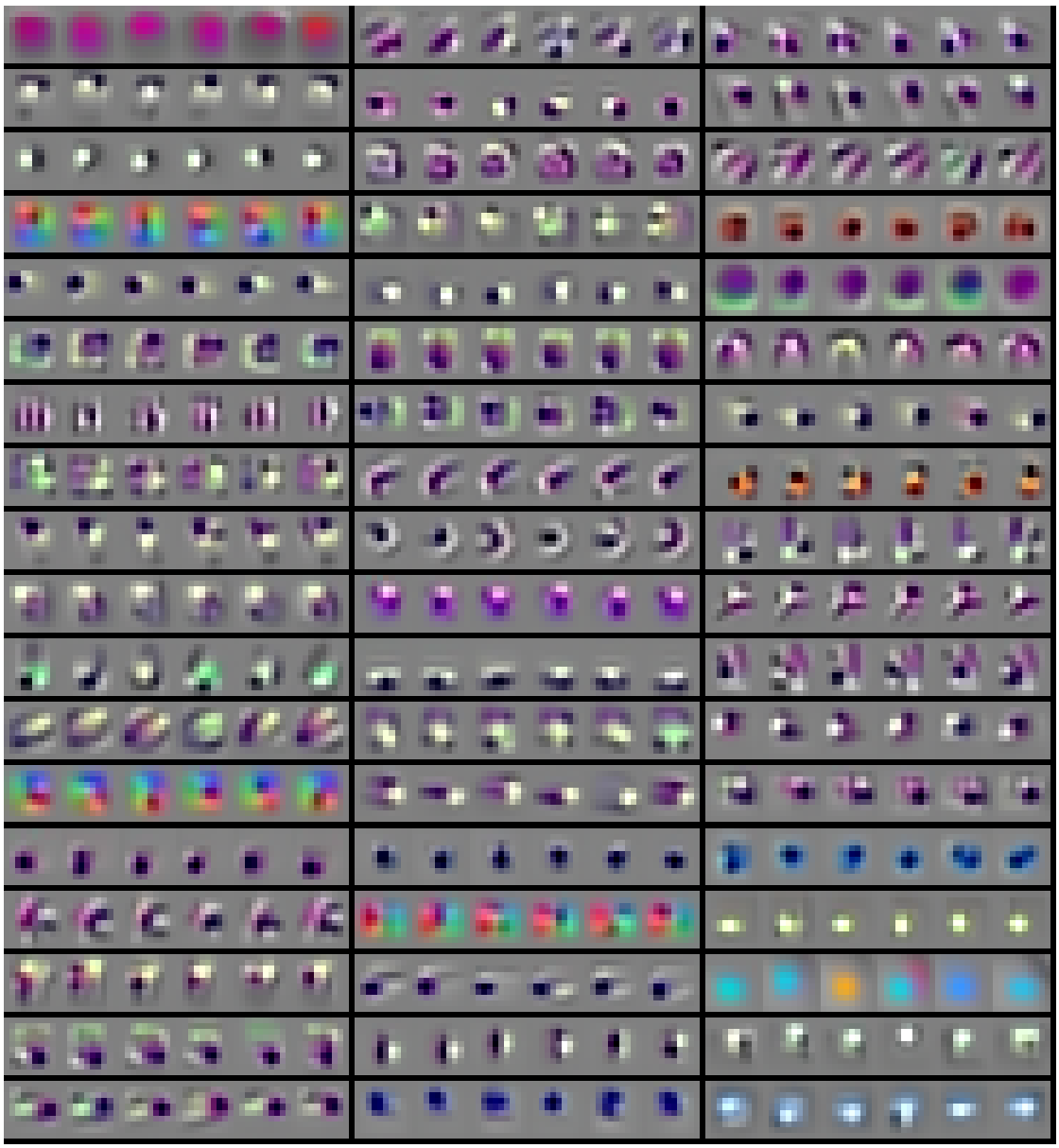


8.2 Example invariant subspaces at optimal λ for early convolutional layers of VGG-19

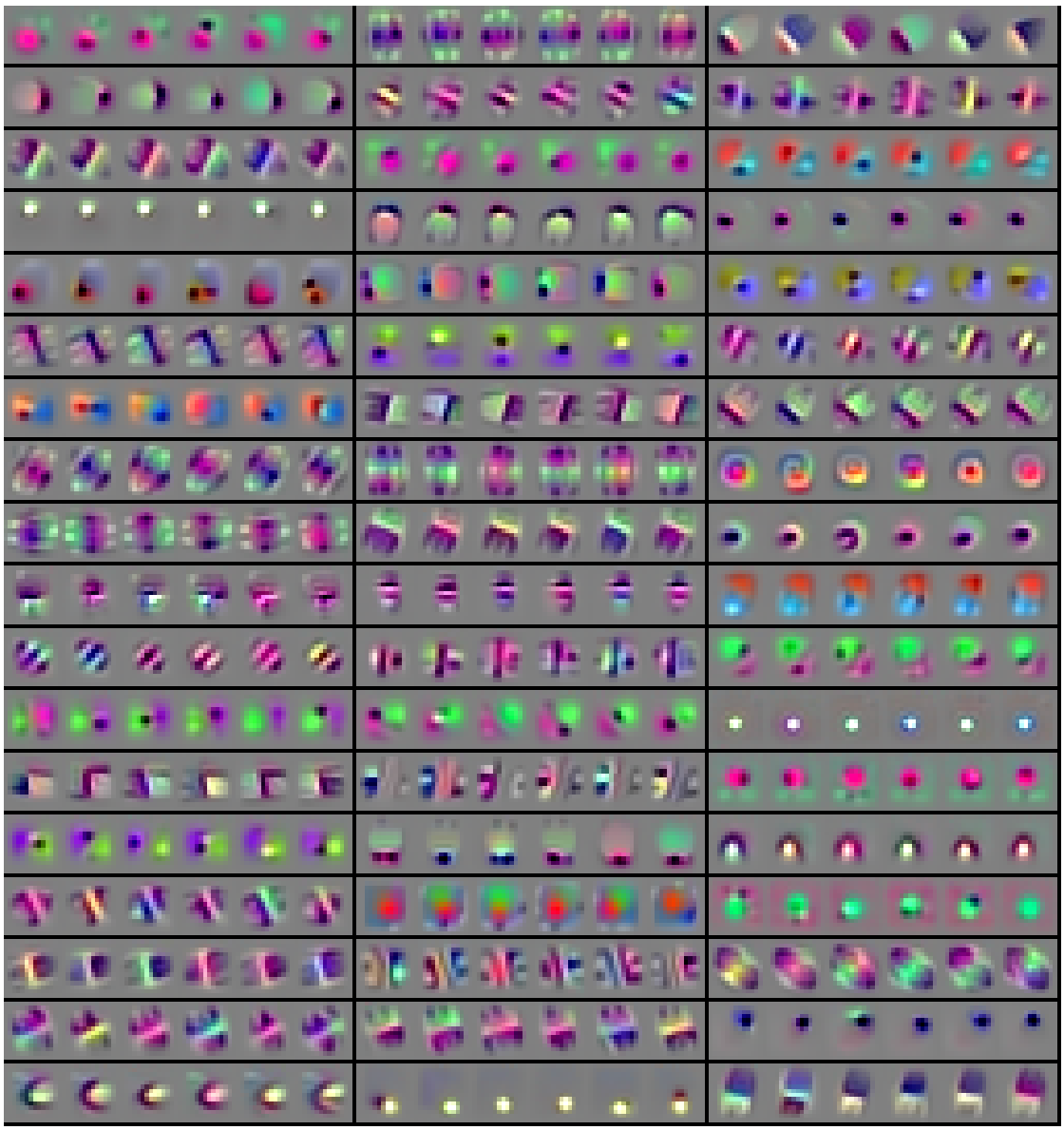
conv1_2 .



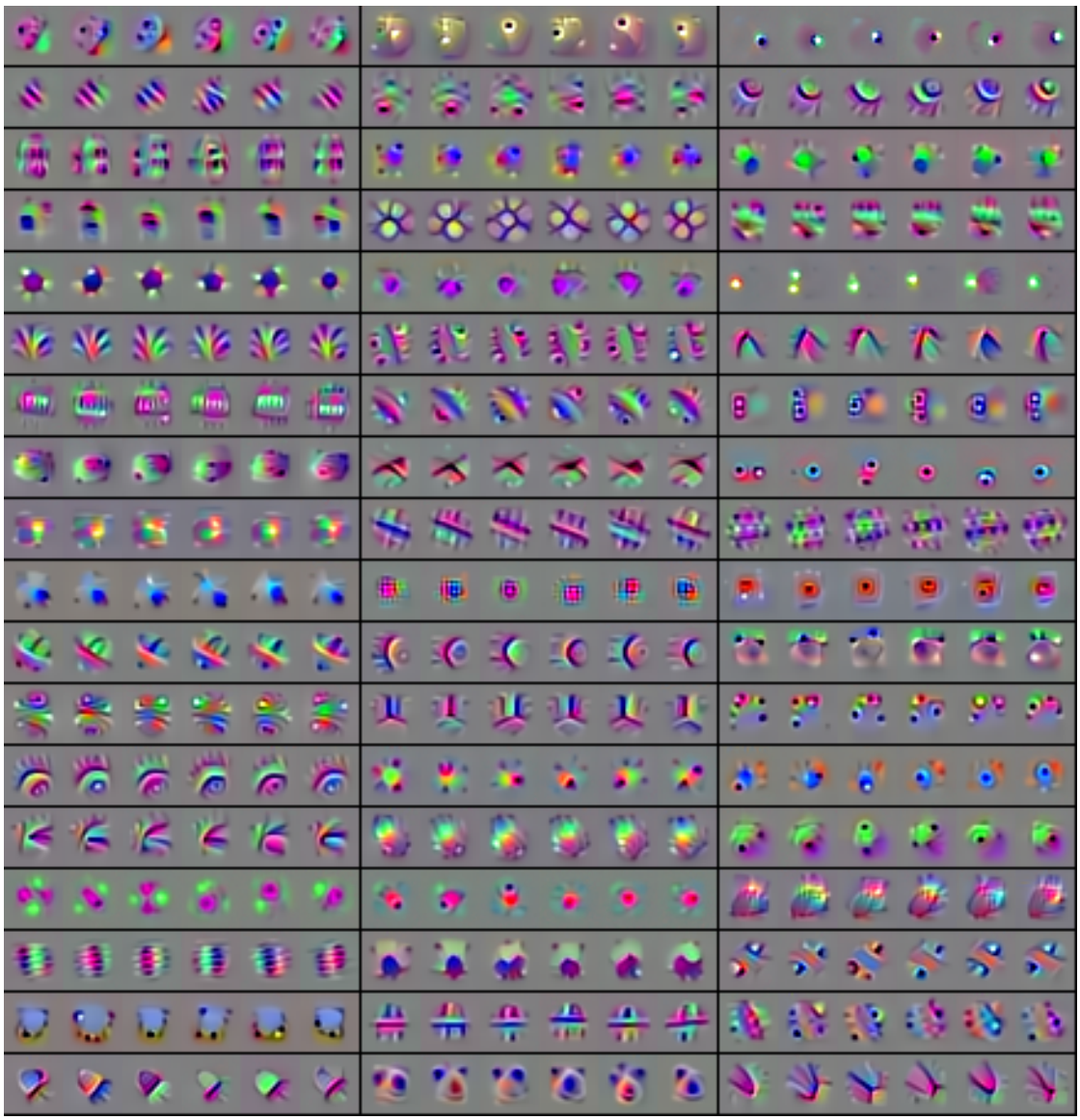
conv2_1 .



conv2_2 .

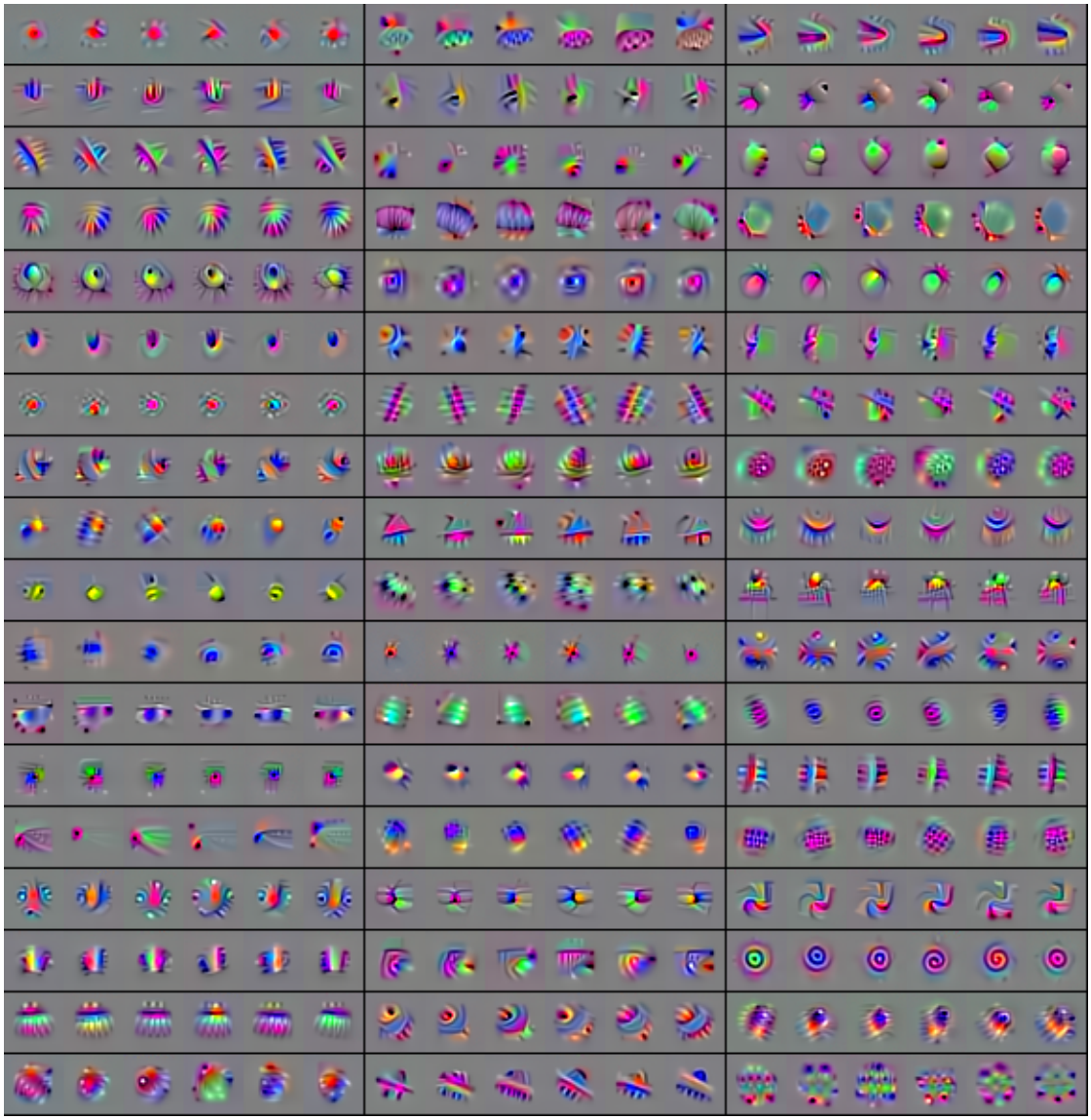


conv3_1 .

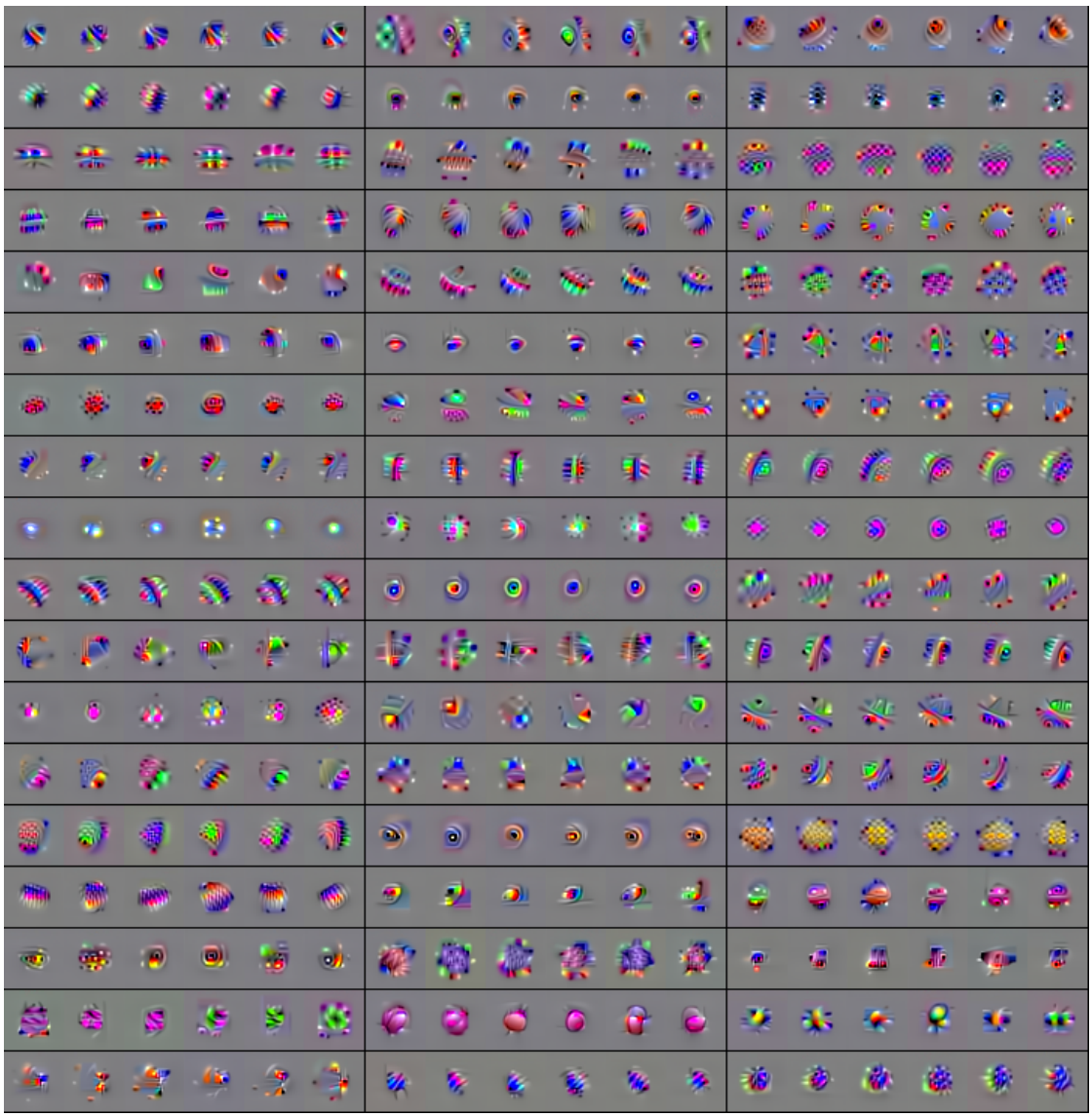


conv3_2

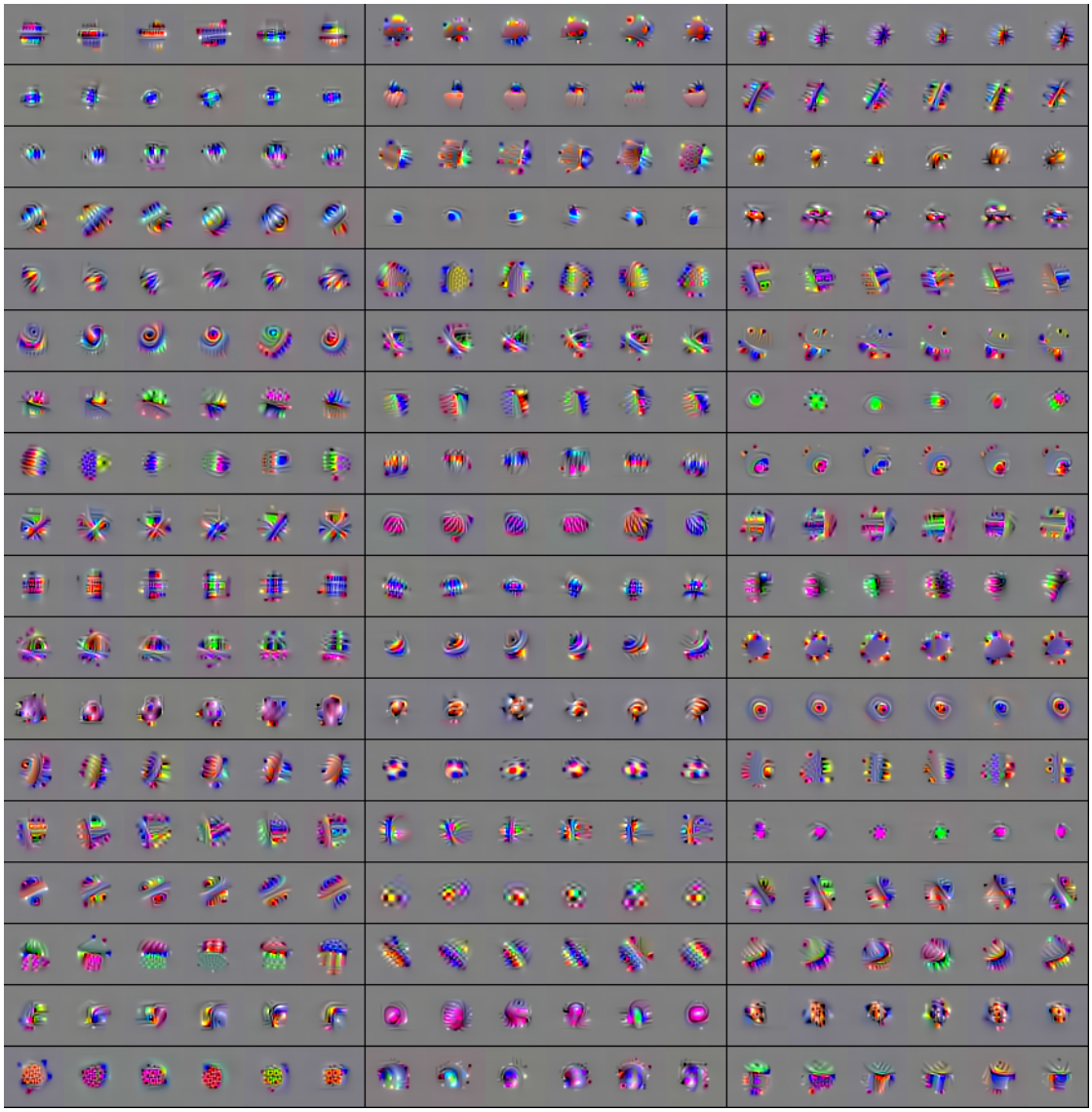
.



conv3_3 .

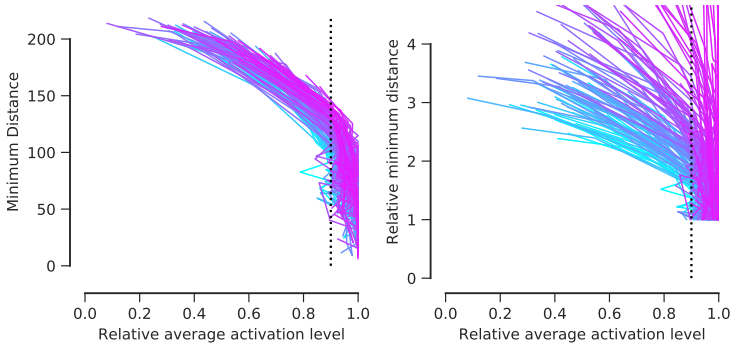


conv3_4 .

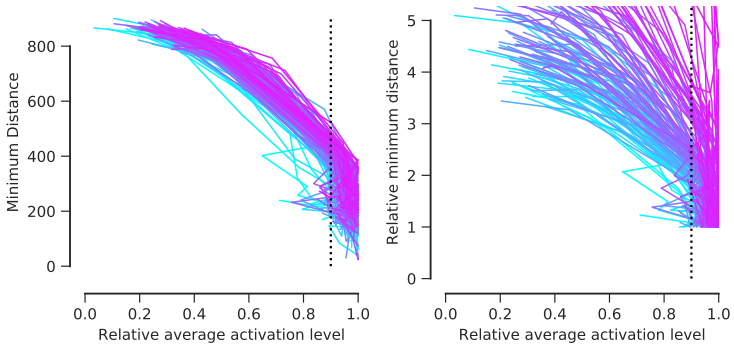


8.3 Diversity/activation maximization trade-off ResNet-50

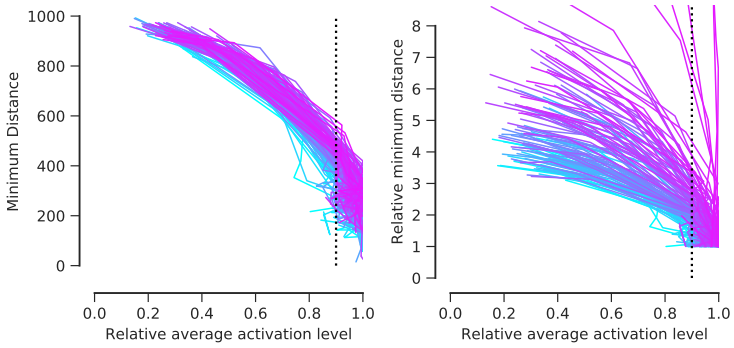
conv2_1 .



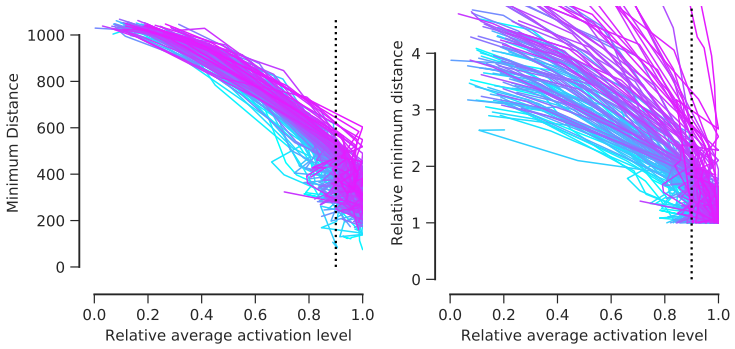
conv2_2 .



conv2_3 .

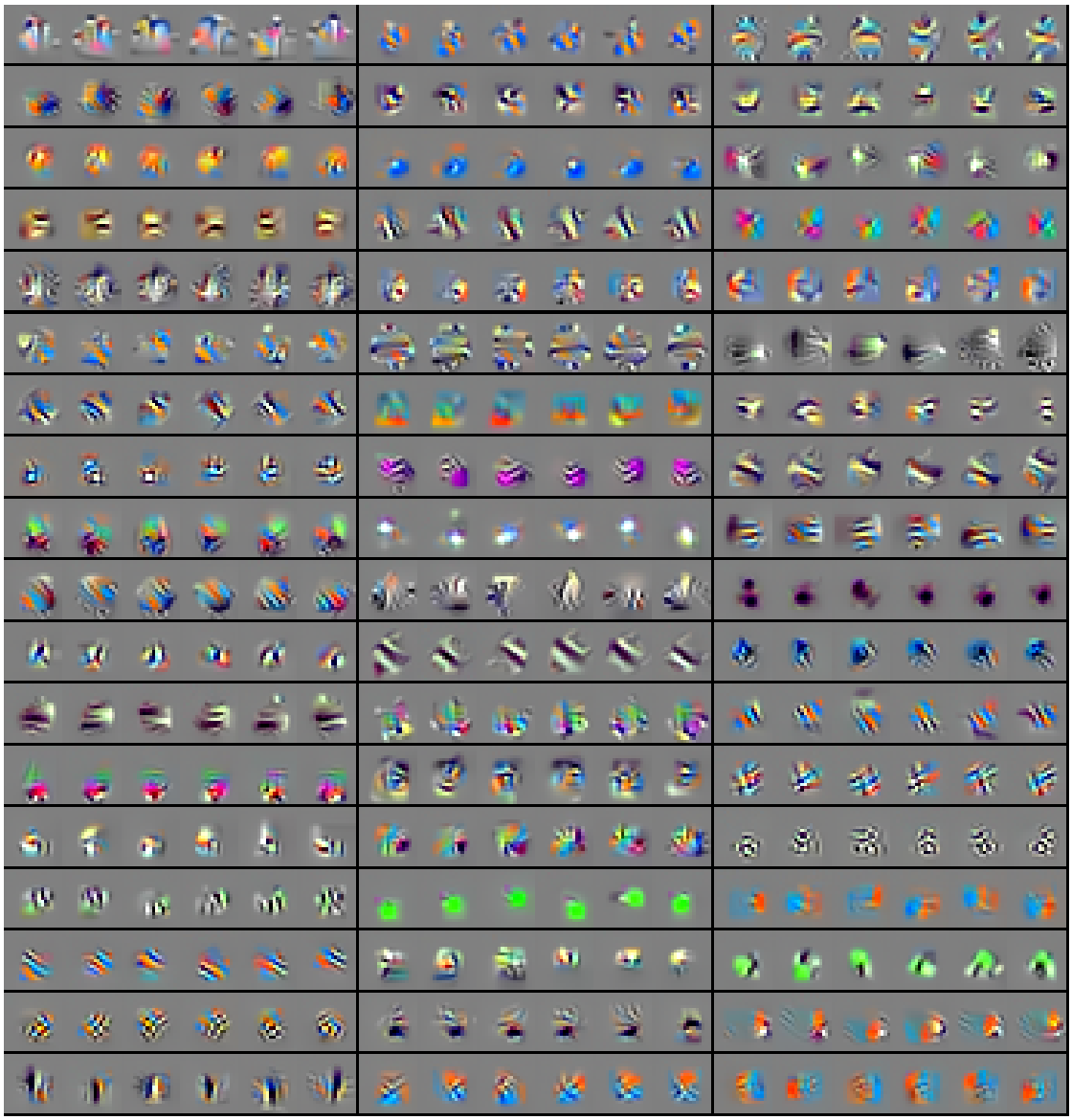


conv3_1 .

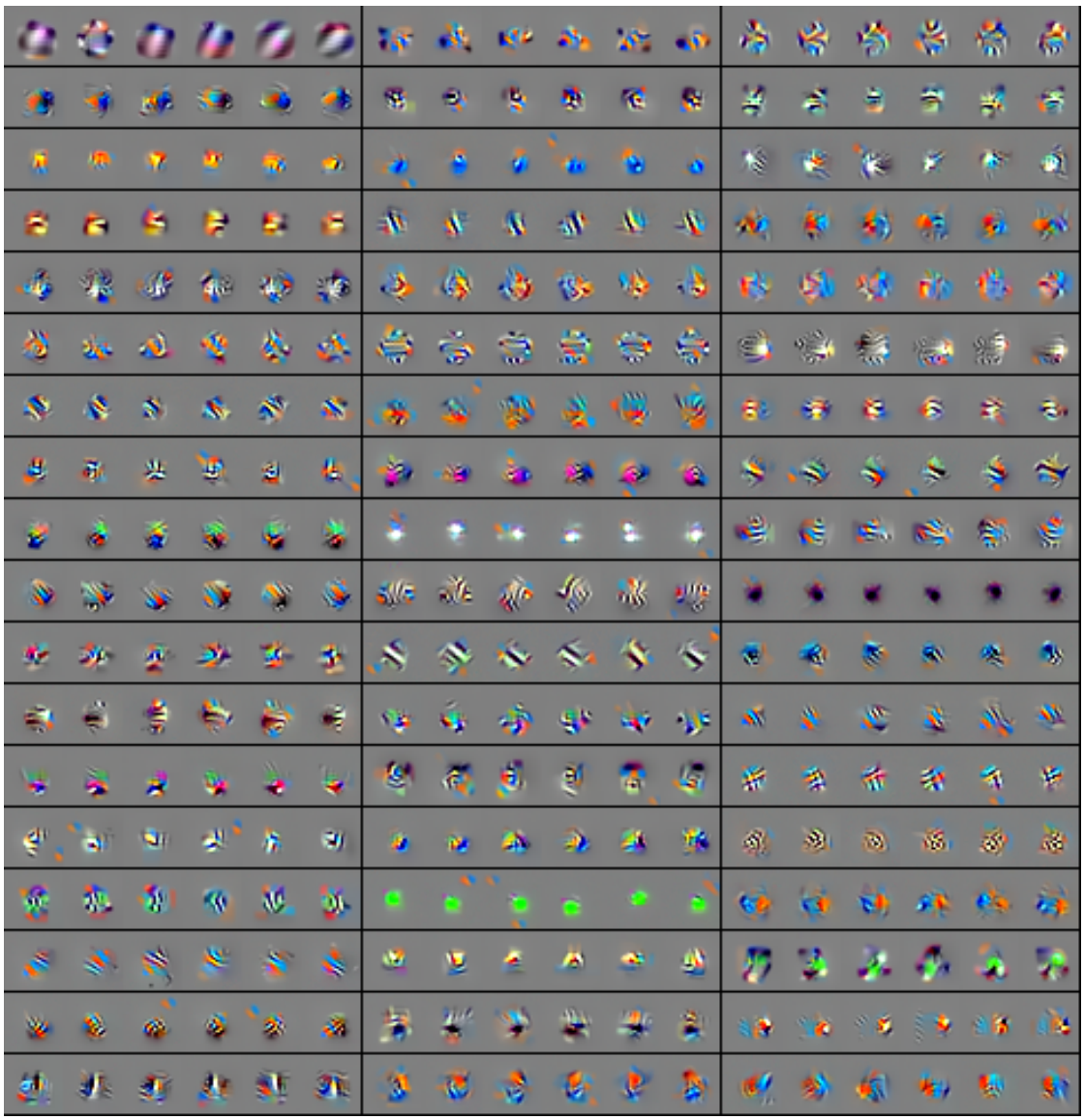

Minimum Distance
1000
800
600
400
200
0
0.0
0.2
0.4
0.6
0.8
1.0
Relative average activation level
Relative minimum distance
4
3
2
1
0
0.0
0.2
0.4
0.6
0.8
1.0
Relative average activation level

8.4 Example invariant subspaces at optimal λ for early convolutional layers of ResNet-50

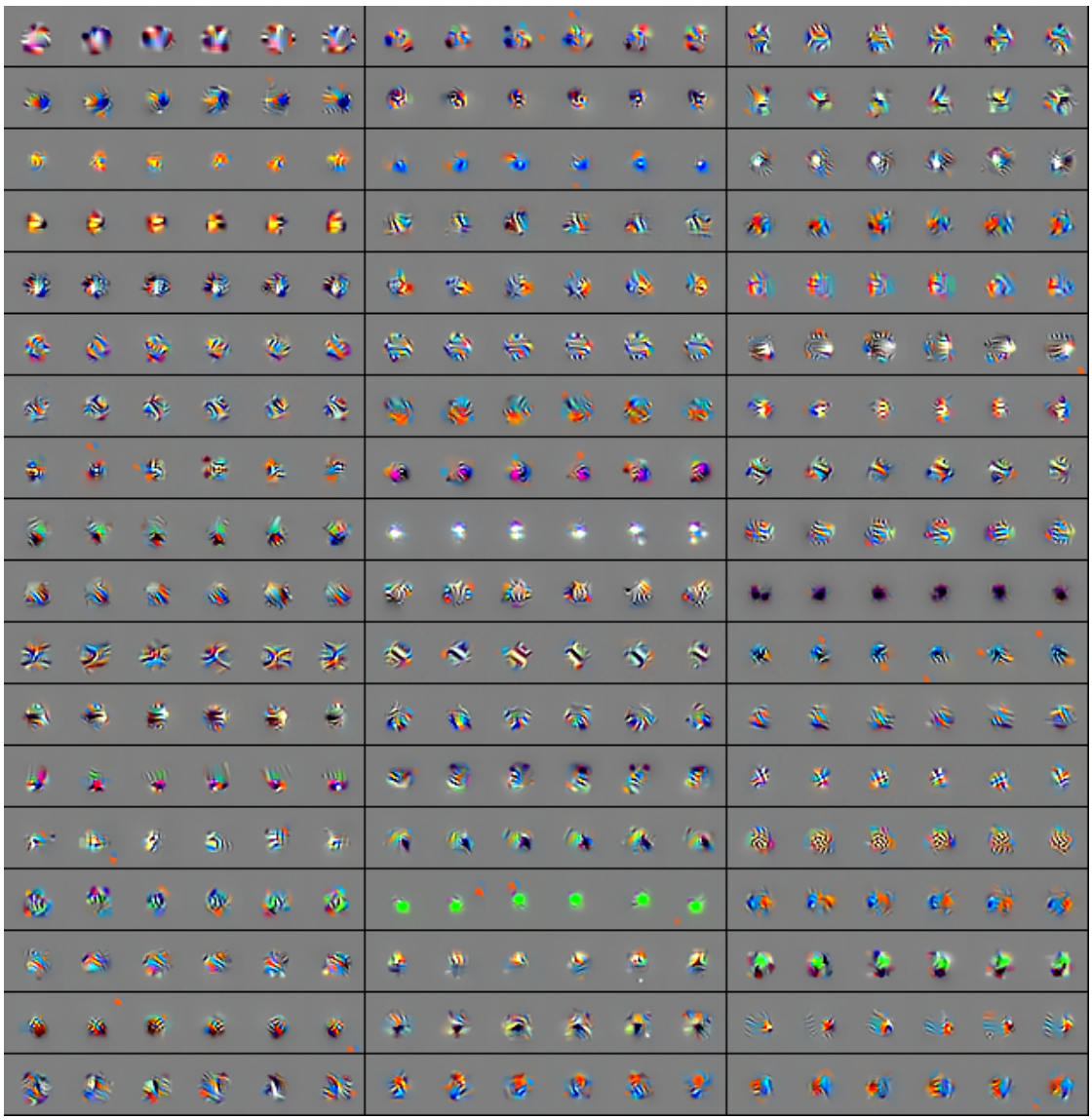
conv2_1 .



conv2_2 .



conv2_3 .



conv3_1 .