Stimulus- and goal-oriented frameworks for understanding natural vision

Maxwell H. Turner[1,2†], Luis Gonzalo Sanchez Giraldo[3†], Odelia Schwartz[3‡], Fred Rieke[1‡*]

1. Department of Physiology and Biophysics, University of Washington, Seattle, Washington 98195, USA
2. Graduate Program in Neuroscience, University of Washington, Seattle, Washington 98195, USA
3. Department of Computer Science, University of Miami, Coral Gables, FL, 33146, USA

*Correspondence:rieke@uw.edu

†Co-first authors

‡Co-senior authors

Abstract

Our knowledge of sensory processing has advanced dramatically in the last few decades, but this understanding remains far from complete, especially for stimuli with the large dynamic range and strong temporal and spatial correlations characteristic of natural visual inputs. Here we describe some of the issues that make understanding the encoding of natural image stimuli challenging. We highlight two broad strategies for approaching this problem: a stimulus-oriented framework and a goal-oriented one. Different contexts can call for one or the other framework. Looking forward, recent advances, particularly those based in machine learning, show promise in borrowing key strengths of both frameworks and by doing so illuminating a path to a more comprehensive understanding of the encoding of natural stimuli.

Introduction

The neural circuits that process sensory inputs are shaped by the properties of the stimuli they encounter as well as the behavioral demands of the animal. Because of this, a deep understanding of sensory circuits and the computations they support requires connecting what we know about sensory systems to properties of natural stimuli. In this review, we discuss some of the progress and the challenges in describing the neural encoding of complex stimuli such as those encountered in the real world; related issues extend to many areas beyond neurophysiology. We refer to the encoding of visual scenes as a paradigmatic example, but many of the same issues arise in other sensory modalities.

Studies of sensory coding have traditionally relied on parameterized, artificial stimuli designed to isolate and characterize specific circuit mechanisms, such as nonlinearities in the integration of signals across space (reviewed by [1,2]) or adaptation to changes in particular stimulus properties such as intensity, contrast, or orientation (reviewed by [3–6]). These approaches have revealed the mechanistic basis of many important circuit computations. There is also a long history of studying the encoding of natural scenes in neurophysiology experiments (e.g. [7–12]), and recent years have seen this interest expand (e.g., [13–15] and references therein). However, the encoding of natural stimuli is incompletely understood.

Two issues make studying the encoding of natural stimuli challenging compared to typical artificial stimuli. First, complex stimuli, such as natural visual inputs, engage a host of interacting circuit mechanisms rather than individual mechanisms in isolation. These interactions can be difficult to capture with computational models. As a result, many models do not generalize well to predict responses to stimuli other than those to which they were fit[16]. For example, many predictive neurobiological models for stimulus-response transformations in the early visual system are based on a common architecture: linear filtering over space and time, followed by a nonlinear step. Such models tend to suffer from an inability to generalize to novel stimuli, especially natural ones[17–19]. Alternative model architectures may generalize better, for example those that stack multiple linear-nonlinear layers on top of one-another[20,21], or those that use multiple linear filters in parallel to capture diverse feature sensitivities[22–27].

A second challenge inherent in the study of natural stimulus encoding is the complex statistics of natural scenes (reviewed by [28–31]). For example, across different visual scenes and even within a single scene, image statistics (e.g. mean intensity, spatial contrast, and other, higher-order statistics) can vary widely but (fortunately) not randomly[32–36]. Within a single visual scene, different image features are often strongly correlated, which makes it difficult to relate a neural response to a particular feature of a scene (see [13] for a computational approach to this issue). One approach to managing this complexity is to develop generative models of natural images that enable a low dimensional representation. Parametric models exist for naturalistic textures[37] -- i.e. semi-regular, repeating patterns (see Figure 1) -- and recent advances in machine learning show promise in generating not only textures[38] but non-homogeneous naturalistic images (see [39] and references therein); for applications of these approaches see [40–42].

Stimulus- and goal-oriented approaches to natural stimulus encoding

We will focus on two theoretical frameworks that are often appealed to in the study of natural stimulus encoding: (i) A *stimulus-oriented* framework, A common approach is to identify transformations of sensory input signals that optimize statistical and information theoretic metrics, such as reducing statistical redundancies present in natural stimuli. Complementary approaches based on generative modeling seek to capture the statistical dependencies of natural scenes, and by doing so also reveal how they can be reduced. Stimulus-oriented approaches are closely related to unsupervised machine learning, for which learning is based only on properties of the input and does not require a specific task goal such as object recognition.

(ii) A *goal-oriented* framework, which appeals to the computational or behavioral goal of the circuit or animal. Unlike stimulus-oriented approaches, goal-oriented approaches explicitly treat some features of the stimulus differently than others, and which features are encoded depends on the desired behavioral output or goal. These approaches include recent advances in Deep Convolutional Neural Networks, particularly those based on supervised, discriminative learning from large databases of images with identified and labeled objects.

These two frameworks may appear to be at odds. For instance, a model focused solely on a high-level goal like object recognition will not necessarily reduce redundancies or capture general statistical properties of the stimulus. Conversely, models focusing on reducing redundancies are not likely to explain, at least not explicitly, complex tasks such as object recognition. Historically, stimulus-oriented frameworks have largely been applied to early visual areas and goal-oriented objectives to later cortical areas. But these boundaries are beginning to blur. Indeed, in some cases the two approaches can be seen as complementary. For instance, even well-established visual computations like lateral inhibition can be seen through both lenses: as a mechanism to suppress responses to low spatial frequencies and eliminate some of the redundancies present in natural images[43,44], or as a way to facilitate the detection of specific features of a scene, namely edges[45]. In addition, stimulus-oriented approaches can be relevant for pre-attentive selection and segmentation tasks, for instance by creating a saliency map in primary visual cortex[46]. We will discuss some modern computational approaches that may facilitate the merger of stimulus- and goal-oriented frameworks, allowing one to inform the other and vice-versa. In particular, deep neural networks provide a promising route for exploring how stimulus- and goal-oriented constraints together shape sensory processing.

Stimulus-oriented approaches to natural vision

An influential hypothesis that undergirds much of the study of natural scene processing is the “efficient coding hypothesis,” first proposed by Barlow[47] (see also [48]), and influenced by Shannon's earlier work on information theory[49]. Barlow proposed that an efficient coding scheme should reduce the redundancy of natural inputs, but without loss of the information that is encoded[47]. Redundancy as defined by Barlow is the fraction of the total information carrying capacity of a neuron or neural population that is not used to transmit information about the stimulus. Approaches based on producing sparse representations of natural inputs also take advantage of the redundancy in images[50].

Redundancy reduction predicts that a single noiseless neuron should distribute its responses uniformly (e.g., subject to a constraint on the maximal firing rate), such that each possible response occurs with equal frequency; to do otherwise would mean that the neuron is not making full use of its dynamic range. Examples of approximately uniformly-distributed sensory representations can be found in a variety of sensory systems[51,52]. Consideration of neural noise can substantially alter predictions of efficient coding because in that case efficiency involves both using a cell's full response range and mitigating the effect of noise[53–55].

Redundancy reduction in a population of neurons (i.e., multiple channels) relies on removing statistical dependencies among their responses[47]. Reducing redundancy for natural stimuli is particularly challenging because natural visual inputs contain strong (nonlinear) statistical regularities across time and space (for a review, see[30]). We start by describing the application of these ideas in early sensory areas (mainly the retina) and then turn to efficient coding in visual cortex.

Efficient coding and second order statistics

Second-order spatial correlations in natural scenes have been a particular focus of efficient coding approaches. Such correlations, on average, obey a power law scaling: the power spectrum of spatial frequencies falls as the inverse of the square of the spatial frequency (Figure 2b)[56]. This is the result of the scale invariance of natural images -- i.e. many statistical properties are unchanged by magnifying or demagnifying an image[36]. Scale invariance has been suggested to result from the fact that objects can appear at any distance from an observer[57].

The prevalence of low spatial frequencies in natural images produces correlated responses in nearby cells, leading to a redundant population code. Receptive field surrounds of neurons in retina and LGN decorrelate responses of nearby neurons by suppressing responses to low spatial frequencies[43,58] (but see [59–61]). The transformation that flattens the power spectrum is sometimes referred to as "whitening." Whitening, however, will increase high spatial frequency noise such as that in photoreceptor signals; consideration of noise predicts that the suppressive surround should be minimal or absent when noise is high (for a review, see [31,62]). Similar principles of whitening without amplifying noise have also been proposed in other domains, such as stereo coding in cortex[63].

Eye movements are another factor that can make important contributions to the statistics of visual inputs and hence to efficient coding predictions. Human eye movements are characterized by small fixational movements and occasional discrete and rapid saccades (Figure 2a,c). The spatial frequency spectrum of natural images, subject to fixational eye movements, is roughly flat (i.e., whitened) at low spatial frequencies[64] (Figure 2b). Natural inputs that simulate fixational eye movements indeed appear to decorrelate responses in populations of salamander retinal ganglion cells[65]. This whitening effect does not hold for large and rapid eye movements like saccades[66] (see Figure 2b). Thus, Rucci & colleagues (especially [66]) suggest that a single cell may use different decorrelation strategies throughout the course of natural stimulation: classical surround-mediated decorrelation or decorrelation via nonlinearities in spike generation[60] immediately following a saccade and eye-movement generated whitening during the later parts of the fixational periods between saccades. Understanding the effects of such self-generated motion on the encoding of natural scenes will require further experiments (e.g., manipulating the statistics of synthetic eye movements in experiments on primate retina).

Efficient coding beyond second order statistics

Much of the classical work on efficient coding considers only second-order statistics and their removal by decorrelation. There is, however, much more to natural images than their spatial frequency spectra. This is evident when viewing artificial stimuli with a "natural" distribution of energy across spatial frequencies but no other statistical constraints; such images look highly unnatural (e.g Figure 3). This raises a concern that coding algorithms focusing on decorrelation may miss essential features of what early visual neurons do.

Statistical independence provides a stronger constraint on efficient coding between channels (i.e. neurons or neuron-like receptive fields) than decorrelation (for a comprehensive review, see book by [67]). Although achieving independence in general is a difficult problem, it can be simplified by considering only linear transformations followed by a point nonlinearity (i.e. a linear-nonlinear approach). Two such approaches applied to natural images (Independent Component Analysis and Sparse Coding) yield filters that qualitatively resemble the oriented and localized structure of receptive fields in primary visual cortex[68,69]; for a review, see [30]. More recent work shows that optimizing for a form of hard sparseness in which only a limited number of neurons are active can yield a better match to the full variety of cortical receptive fields in macaque[70].

Different channels can also exhibit nonlinear statistical dependencies that cannot be fully removed by linear or linear-nonlinear approaches (see [71–73] and references therein). This has prompted work on reducing statistical dependencies via nonlinear transformations. These approaches have led to more direct comparisons between models derived from scene statistics and nonlinear neural behaviors. One focus in primary visual cortex has been on modeling nonlinear contextual phenomena, whereby the responses of neurons to a target stimulus are influenced by stimuli that spatially surround the target, or by stimuli that have been observed in the past. Such effects can be modeled by reducing statistical dependencies between filter responses across space or time via a nonlinear computation known as divisive normalization or by other complementary approaches[32,73–78]. The statistical dependencies between filter responses can also be exploited to build models of complex cells that pool together filters, resulting in invariances to translation and other properties (for a review see book by [67] and references therein; see also [79,80]). Models of secondary visual cortex have been derived by stacking multiple layers of linear-nonlinear transforms to achieve statistical independence, sparseness, or other related stimulus-driven goals[81–84]. One can in principle stack many unsupervised layers, but it is not clear if efficient coding remains relevant for capturing the computations characteristic of higher cortical areas and hence provides a good fit criterion. It is often assumed instead that goal-oriented approaches become more appropriate as computations become more specialized.

Generative models that capture image statistics can complement efficient coding approaches[85,86]. Efficient coding approaches seek to transform and manipulate inputs so as to maximize the transfer of information, which can result in statistical independence of the transformed inputs. But learning to generate the statistical dependencies prevalent in natural scenes also shows how to reduce them. To make this more concrete, consider an example in which efficient coding and generative models are complementary. Multiplicative generative models for the nonlinear dependencies in filter responses to images lead immediately to approaches to reduce such dependencies via division[87]. Building on this simple example, generative approaches allow formulation of rich models of the statistical dependencies in images, based on the observation that different parts of an image could have different statistical dependencies. This leads to models in which divisive normalization (and therefore redundancy reduction) only occurs for image inputs in which center and surround locations are statistically dependent according to the model[32,88] (see also [89]).

Goal-oriented approaches to natural vision

Efficient coding predicts that neural processing will maximize the information transmitted about a stimulus without explicitly considering behavioral demands such as the specific tasks required for survival. These behavioral considerations are central to goal-oriented approaches, which view the importance of stimulus structure and circuit mechanisms on coding through the lens of specific behavioral demands. Because many behaviorally-relevant tasks require rich stimuli, goal-oriented approaches are often used to investigate the coding of natural inputs. We first illustrate these issues from studies of the retina and insect behavior, and then turn to their application in cortex.

Retinal ganglion cells support specific behavioral goals

A common observation that supports goal-oriented approaches is high neural selectivity to specific stimulus features to the exclusion of other (equally probable) features. In an early study of retinal feature selectivity, Lettvin and colleagues interpreted retinal ganglion cell (RGC) types in explicitly ethological terms, famously going so far as to speculate that one class of ganglion cell in the frog retina may be a "bug perceiver"[90]. But the idea that the earliest neurons in the visual system are tuned to highly specific features of the visual world was ahead of its time. Instead, the dominant view of retinal processing for several decades thereafter focused on basic processing, including lateral inhibition (via a center-surround spatial receptive field) and simple forms of luminance adaptation[91]. In this view, the computational heavy lifting to support specific behavioral goals is done in visual areas downstream of the retina and LGN.

A great deal of evidence has now accumulated that retinal computation is more complex (for a review see [1]). A wide variety of "non-standard" RGC computations have been discovered and often explained at the circuit and synaptic level. These include: direction-selectivity, orientation selectivity[92], an omitted stimulus response[93], and image recurrence sensitivity[94]. Of specific relevance here, recent work emphasizes intricate specializations of direction-selective circuits for extracting information about the direction of motion, often to the detriment of encoding other visual features[95,96].

The degree to which retinal neurons are specialized to guide a particular behavior or to perform general-purpose computations predicted by efficient coding may depend on species and on location within the retina. The "complex" computations discussed above (like direction selectivity) have not been observed in primate retina, although many primate RGC types remain unexplored. Further, the fovea and peripheral retina differ dramatically in circuitry (reviewed by[97]) and in functional properties[98–100]; these differences could indicate a difference in the division of computational labor between retinal and cortical circuits across retinal eccentricity.

Differences like these - across cell types, species, or retinal eccentricity - suggest one way to reconcile stimulus- and goal-oriented frameworks in the retina. Retinal neurons that support a variety of behavioral goals or project to image-forming downstream thalamocortical circuits may show more general purpose computational features consistent with efficient coding since these cells act as a common front-end for many downstream feature extractions. Other retinal neurons may violate predictions from efficient coding because they project to areas of the brain that underlie more specialized visually-guided behaviors -- for example, direction selective neurons[101] that project to superior colliculus or the accessory optic system to guide eye movements, or RGCs that control circadian rhythms (for review see [102]).

Lessons from insect vision: behavioral goals shape and constrain visual processing

Goal-oriented approaches have yielded particularly satisfying explanations for complex visual processing in insects. The insect vision community has a long history of examining visual processing as it relates to behaviors like flying[103]. Motion processing pathways in several different insects appear tuned to each species' particular flight behaviors[104]. Some visual neurons in the fly encode visual features directly relevant for flight control, such as optic flow elicited by rotations or translations around and along specific body axes[105,106] (see Figure 4). These neurons act as "matched filters" for specific types of optic flow[107,108]. Optic flow encoding may seem obvious in hindsight, but the local motion receptive fields of these cells would appear quite mysterious if not for the careful consideration of the impact of the fly's own motion on visual inputs.

Recent work on mouse directionally-selective RGCs has similarly recast their function in terms of self-generated motion while navigating the environment[109] (see Figure 4). A long-standing view of directionally-selective RGCs held that they consist of four subtypes, each preferring a cardinal axis of motion (up, down, left, right, each separated by ~90 degrees) and in alignment with the axes of eye movements produced by the four rectus muscles of the eye[101]. These RGCs project to the superior colliculus[110], which further suggests that they are involved in controlling eye movements. While this distribution of preferred directions holds in the mouse central retina, in other regions of the retina the preferred axes of directionally-selective RGCs are not perpendicular and thus do not neatly align with the rectus muscles of the eye. Sabbah & colleagues mapped retinotopic differences in direction-selectivity in relation to extrapersonal visual space and motion by the animal (Figure 4). They found that directionally-selective cells are in fact better thought of as encoding the animal's own "advance/retreat" and "rise/fall" movements than the movement of some external object.

Goal-directed Approaches in Cortex

Goal-directed approaches have also been applied to visual cortex. Geisler and colleagues have promoted the importance of understanding how particular tasks may exploit different properties of natural scenes[111,112]. They have focused on the representations learned by tasks such as patch identification, foreground identification, retinal speed estimation and binocular disparity. For instance, filters learned for a foreground identification task were oriented either parallel or perpendicular to surface boundaries[112], while filters from an image patch identification task had less discrete orientation preferences and more closely resembled primary visual cortex filters. Thus, the representations learned can depend on the visual processing goals imposed on the system.

Deep Neural Networks

Recent years have seen tremendous advances in an area of machine learning known as deep neural networks (DNNs[113,114]); these advances have driven progress in computer vision and a host of other fields. In deep neural networks, stimuli such as natural images are represented and processed hierarchically, loosely matched to the hierarchical structure of the brain. These networks come in many different flavors, including those that are trained in an unsupervised manner -- i.e. the network learns to identify and encode statistical structure in the inputs without a specific goal. Here we focus on supervised discriminative networks, which are tasked with identifying or categorizing inputs and learn to do so by observing many examples of each category in a labeled training data set. For example, a commonly used labeled training data set is ImageNet, which is a collection of images of objects and their associated classifications (e.g. "German shepherd", "birdhouse", or "eggnog"). DNNs have many potential applications; we emphasize their potential to help understand and make predictions about the neural processing of natural images, particularly how the nervous system could achieve invariant object recognition (e.g. to pose, background clutter, and other within class variations).

Architecture and neural circuitry

Deep neural networks consist of a series of connected layers, each of which implements a set of basic computations (Figure 5). The computations in a single layer include linear filtering (convolution), rectification, pooling, and sometimes local response normalization. DNNs can be considered as a hierarchical extension of the linear-nonlinear models often used to empirically describe visual responses. By design of the network, the dimensionality (number of elements) is reduced between successive layers, and effective receptive fields become larger as one progresses along the hierarchy. Thus, individual layers implement computations like those found in descriptive models of neural circuits, and the hierarchical arrangement of layers resembles the organization of visual (and other sensory) pathways.

The parameters governing DNN behavior are not determined by specific low-level computational principles (e.g. reducing statistical dependencies as in efficient coding). Instead these parameters emerge by learning to minimize the difference between the DNN output and a desired response corresponding to the DNN goal - such as classifying images according to objects they contain. DNNs can also be used in a descriptive (and therefore not goal-oriented) manner by fitting them directly to neural data, rather than training them on a high-level task. One such model, when fit to retinal ganglion cell responses to natural movies, reproduced several of the "complex" retinal computations discussed above. The model did not reproduce these behaviors when fit to white noise stimulation[115].

While neural networks have been around for decades, recent years have seen dramatic improvements in performance due to increases in computer speed and the availability of large data sets (e.g. images with labeled objects) that together make it possible to efficiently train networks with many layers.

Learning from successes and failures of DNNs

DNNs trained on object classification show an intriguing ability to predict the responses of cortical neurons to natural images (for recent reviews, see [116,117]; for other recent work, see [118–120]). This approach has been applied with particular success to processing in the ventral visual pathway, which culminates in neurons in inferotemporal (IT) cortex. Many IT neurons exhibit high feature selectivity -- responding to specific objects and (famously) faces[121].

The flow of signals from the retina to IT is characterized by the loss of a veridical representation of the retinal image: receptive fields become progressively larger and more complex, invariances to properties like object size and position emerge, and the appropriate space to specify inputs (e.g. inputs that produce similar responses of a given neuron) becomes increasingly difficult to identify. These transformations are challenging to describe using stimulus-based models. DNNs, however, have been more successful. Interrogation of the architecture of DNNs trained on object classification suggests that invariances may arise from the pooling stages of the networks[122,123]. DNNs show an ability to generalize in two important ways: (1) they are able to classify images of objects not in the original training set, including adjusting their representation of inputs for different tasks through transfer learning[124]; and, (2) they capture several aspects of neural responses even though neural data is not used in training.

But DNNs are, of course, imperfect. For example, current DNN models fail to capture some aspects of human perception such as insensitivity to perturbations to an image[125,126]. This behavior may arise from current DNN architectures operating in rather linear regimes[127], and more biologically realistic saturating nonlinearities may improve performance[128] (although see [129]). DNNs capture some but not all aspects of responses of neurons in mid-cortical layers[120]. Interpreting DNNs can also be difficult. Unlike more principled efficient coding approaches in which the form of the computation itself (e.g., divisive normalization or gain control) can be motivated by the computational goal, it is often not clear what feature of a supervised, discriminative DNN leads to a given level of performance. This sort of insight is more readily gleaned from shallower models that share many architectural features with DNNs (see, e.g., [26,130]).

Any insights that DNNs trained on high-level tasks like classification provide about how the visual system computes comes from identifying, through learning, key statistical structure in the inputs that is important for performing the specific task used in training. Motivation for such an approach comes from convergent evolution of computations like motion detection in insect and vertebrate visual systems (see above). Given that DNNs are only loosely modeled after visual circuits, a realistic expectation is that they identify the computational capabilities and limitations of specific architectures rather than provide a literal model of how the visual system works. If statistical structure of the inputs, rather than specific hardware constraints, dominates which computational strategies are effective for a given task, we might expect DNNs and neural systems to converge on similar computational algorithms even if the implementations of these algorithms differ due to differences in hardware.

Future directions

Understanding neural computation and coding in the context of naturalistic visual stimuli is a difficult problem. But the wealth of neurophysiological data about the visual system and the emergence of new computational tools for building and fitting models put us in a good position to make progress. Below we highlight a few emerging directions that we believe will help advance understanding. Many of these approaches merge techniques and ideas from the stimulus- and goal-oriented frameworks discussed above.

Identify key circuit mechanisms and integrate into models

A complete understanding of natural visual encoding entails building models that can accurately predict neural responses to natural scenes. We believe that a major reason for the shortcomings of current models is that they lack key architectural and computational features present in biological circuits, and that these features substantially shape neural responses. Certain model abstractions (for example, linearity of the receptive field) may be appropriate under some stimulus conditions but not others. At the same time, simply building models using realistic components is not likely to explain complex computations such as object recognition. Merging DNN techniques with more realistic biological circuitry offers one path forward.

DNNs components and connectivity are typically chosen largely based on the computational efficiency of learning using current optimization tools (e.g. gradient descent). This can lead to architectures that lack key components of neural circuits. Identifying and incorporating biologically-inspired computational motifs will help identify which motifs are important for specific computations -- e.g. the computations characteristic of different stages of the visual hierarchy -- and which motifs can be simplified without loss of performance. This in turn could lead to direct predictions about the mechanisms operating in the relevant neural circuits.

One indication of the potential benefits of such an approach comes from comparing physiologically-based models of early visual areas (linear-nonlinear models with two forms of local normalization) and layers of the VGG network (which lack normalization): physiological models captured human sensitivity to image perturbations considerably better than DNNs[131]. A challenge is our current inability to identify which biological mechanisms are essential for specific computations and which can be abstracted as in linear-nonlinear models. Progress will also require probing the interactions between coactive mechanisms that are likely engaged strongly for complex stimuli such as natural images. A partial list of computational features prominent in neural circuits but under-represented in DNNs applied to neuroscience, includes normalization by stimulus context and recurrent connections. Sophisticated forms of normalization in DNNs have thus far been applied to computer vision[132,133] but offer potential for neuroscience directions[134]. Recurrent connections can improve object recognition[135] and have the potential to capture neural phenomena such as adaptation[21].

Combine the merits of stimulus- and goal-oriented approaches

DNNs are designed to perform well on the discriminative recognition task at the top level of the network, but this constraint does not uniquely specify the architecture of the other layers. On the other hand, stimulus-oriented approaches provide a principled way to capture more detailed computations and nonlinearities in early stages of visual processing, including retina and primary visual cortex. But it is not clear if such approaches can capture computations in later stages of the cortical hierarchy.

An important future task is therefore finding better ways to reconcile and integrate the merits of both approaches. For instance, most of the early processing that takes place before primary visual cortex is neglected in current DNNs (an exception is [21]). Incorporating this early processing into networks could become a merger point between goal-directed objectives shaping the top levels of the network and stimulus-driven constraints shaping the initial stages of the architecture. Another direction is to incorporate computational motifs derived from stimulus-driven normative approaches (such as the normalization discussed above) into DNNs.

New theoretical and practical approaches that balance stimulus- and goal-oriented approaches provide promising directions. For instance, an approach known as the information bottleneck formalizes the idea of capturing relevant information rather than all information (for recent application to deep learning, see [136]). Another recent approach unifies several definitions of efficient coding and considers the impact of incorporating only stimuli that are predictive about the future on coding[137,138]. Other recent work connects generative (stimulus-oriented) and discriminative (goal-oriented) components in a single model through a shared representation[139]. This combination has been exploited in 'semi-supervised' machine learning, which makes use of scarce labeled data along with unlabeled data, and therefore is a hybrid between supervised and unsupervised approaches. However, this combined stimulus and goal-oriented representation has not been applied to neuroscience and understanding natural vision. Recent theoretical work has also expanded the notion of efficient coding by recasting it as a specific case of Bayesian inference[140]. By using a broader definition of optimality, Bayesian efficient coding allows one to evaluate the efficiency of neural representations in terms of encoding goals beyond simple information maximization.

There is also a need for progress with stimulus-oriented unsupervised learning approaches that exploit the power of DNNs without specialization for a specific goal. Unsupervised learning is considered by many the "holy grail" of learning (for recent examples, see [141] which incorporates multiple levels of divisive normalization; and [142] which incorporates pooling). It is still unclear whether deep network architectures with unsupervised learning can predict responses of neurons to natural scenes or capture the invariances that characterize higher visual processing.

Train DNNs using multiple, behaviorally-inspired tasks

A DNN trained to perform a particular task can recapitulate some aspects of sensory circuits; for example, the middle layers of an image classification DNN resemble in some respects neurons in intermediate stages of the ventral stream[120] (reviewed by [117]). Presumably these correspondences arise from similarities in both network architecture and task. A real sensory system, however, supports a wide array of tasks or behavioral goals simultaneously. The result is that, especially in early sensory areas, neurons have to process sensory input in a way that supports multiple parallel feature extractions or behavioral goals. Neurons that make up this common biological front end (e.g. photoreceptors or some types of retinal ganglion cells) may therefore align their encoding strategies with efficient coding to support a wide variety of downstream goals. Downstream circuits performing more specialized computations, on the other hand, may not behave according to classical efficient coding principles. This agrees with our intuition that efficient coding somehow applies more neatly to peripheral sensory systems. Formalizing this intuition requires grappling with several difficult questions: Are there general rules that govern when a stimulus- or goal-oriented perspective is more appropriate? At what point does a sensory pathway stop simply efficiently packaging information and start "doing" something with that information?

Multi-task DNNs offer one approach for exploring how shared circuitry could support multiple tasks[143]. Indeed, such networks trained for speech and music classification naturally divide into separate pathways, and the level at which that split occurs can affect the performance of the network on these two tasks[144]. An interesting question is whether constraining networks by multiple mid-level tasks (as in [145]) can provide a more general-purpose representation resembling that predicted by efficient encoding. A major impediment to developing multi-task DNNs is the limited availability of datasets that could be used to train such networks (e.g. ImageNet, which consists of a collection of labeled objects, is the dominant dataset used for vision-related applications).

Acknowledgments

We thank Holger Krapp, Dean Pospisil, and Jonathon Shlens for helpful feedback on an earlier version of this review. Holger Krapp very generously provided the data and schematic shown in Fig. 4a,b. This work was supported by NIH grants F31-EY026288 (M.H.T.), EY028542 (F.R.), and a National Science Foundation Grant 1715475 (O.S.).

References cited

1. Gollisch, T. & Meister, M. Eye Smarter than Scientists Believed: Neural Computations in Circuits of the Retina. *Neuron* **65,** 150–164 (2010).
2. Schwartz, G. W. & Rieke, F. Perspectives on: Information and coding in mammalian sensory physiology: Nonlinear spatial encoding by retinal ganglion cells: when 1 + 1 != 2. *J. Gen. Physiol.* **138,** 283–290 (2011).
3. Demb, J. B. & Singer, J. H. Functional Circuitry of the Retina. *Annu. Rev. Vis. Sci.* **1,** 263–289 (2015).
4. Graham, N. V. Beyond multiple pattern analyzers modeled as linear filters (as classical V1 simple cells): Useful additions of the last 25 years. *Vision Res.* **51,** 1397–1430 (2011).
5. Rieke, F. & Rudd, M. E. The Challenges Natural Images Pose for Visual Adaptation. *Neuron* **64,** 605–616 (2009).
6. Solomon, S. G. & Kohn, A. Moving sensory adaptation beyond suppressive effects in single neurons. *Curr. Biol.* **24,** R1012–R1022 (2014).
7. Baddeley, R. *et al.* Responses of neurons in primary and inferior temporal visual cortices to natural scenes. *Proc. R. Soc. London B* **264,** 1775–1783 (1997).
8. Creutzfeldt, O. D. & Nothdurft, H. C. Representation of complex visual stimuli in the brain. *Naturwissenschaften* **65,** 307–18 (1978).
9. Smyth, D., Willmore, B., Baker, G. E., Thompson, I. D. & Tolhurst, D. J. The Receptive-Field Organization of Simple Cells in Primary Visual Cortex of Ferrets under Natural Scene Stimulation. *J. Neurosci.* **23,** 4746–4759 (2003).
10. Stanley, G. B., Li, F. F. & Dan, Y. Reconstruction of natural scenes from ensemble responses in the lateral geniculate nucleus. *J. Neurosci.* **19,** 8036–8042 (1999).
11. Vickers, N. J., Christensen, T. A., Baker, T. C. & Hildebrand, J. G. Odour-plume dynamics influence file brain's olfactory code. *Nature* **410,** 466–470 (2001).
12. Vinje, W. E. & Gallant, J. L. Sparse Coding and Decorrelation in Primary Visual Cortex During Natural Vision. *Science (80-.).* **287,** 1273–1276 (2000).
13. Sharpee, T. O. *et al.* Adaptive filtering enhances information transmission in visual cortex. *Nature* **439,** 936–942 (2006).
14. Theunissen, F. E. & Elie, J. E. Neural processing of natural sounds. *Nat. Rev. Neurosci.* **15,** 355–366 (2014).
15. Zwicker, D., Murugan, A. & Brenner, M. P. Receptor arrays optimized for natural odor statistics. *Proc. Natl. Acad. Sci.* (2016). doi:10.1073/pnas.1600357113
16. Carandini, M. *et al.* Do We Know What the Early Visual System Does? *J. Neurosci.* **25,** 10577–10597 (2005).
17. David, S. V. & Gallant, J. L. Predicting neuronal responses during natural vision. *Netw. Comput. Neural Syst.* **16,** 239–260 (2005).
18. Turner, M. H. & Rieke, F. Synaptic Rectification Controls Nonlinear Spatial Integration of Natural Visual Inputs. *Neuron* **90,** 1257–1271 (2016).
19. Heitman, A. *et al.* Testing pseudo-linear models of responses to natural scenes in primate retina. *bioRxiv* (2016). doi:10.1101/045336
20. Maheswaranathan, N., Baccus, S. A. & Ganguli, S. Inferring hidden structure in multilayered neural circuits. *PLoS Comput. Biol.* **14,** e1006291 (2018).
21. Mcintosh, L. T., Maheswaranathan, N., Nayebi, A., Ganguli, S. & Baccus, S. A. Deep Learning Models of the Retinal Response to Natural Scenes. *Adv. Neural Inf. Process. Syst.* **30,** 1–9 (2016).
22. Felsen, G., Touryan, J., Han, F. & Dan, Y. Cortical sensitivity to visual features in natural scenes. *PLoS Biol.* **3,** (2005).
23. Rust, N. C., Schwartz, O., Movshon, J. A. & Simoncelli, E. P. Spatiotemporal elements of macaque V1 receptive fields. *Neuron* **46,** 945–956 (2005).
24. Eickenberg, M., Rowekamp, R. J., Kouh, M. & Sharpee, T. O. Characterizing responses of translation-

invariant neurons to natural stimuli: Maximally informative invariant dimensions. *Neural Comput.* **24,** 2384–2421 (2012).

25. Vintch, B., Movshon, J. A. & Simoncelli, E. P. A Convolutional Subunit Model for Neuronal Responses in Macaque V1. *J. Neurosci.* **35,** 14829–41 (2015).
26. Rowekamp, R. J. & Sharpee, T. O. Cross-orientation suppression in visual area V2. *Nat. Commun.* **8,** 1–9 (2017).
27. Pagan, M., Simoncelli, E. P. & Rust, N. C. Neural Quadratic Discriminant Analysis: Nonlinear Decoding with V1-Like Computation. *Neural Comput.* **28,** 2291–2319 (2016).
28. Hyvärinen, A. Statistical models of natural images and cortical visual representation. *Top. Cogn. Sci.* **2,** 251–264 (2010).
29. Lewicki, M. S., Olshausen, B. A., Surlykke, A. & Moss, C. F. Scene analysis in the natural environment. *Front. Psychol.* **5,** 1–21 (2014).
30. Simoncelli, E. P. & Olshausen, B. A. Natural image statistics and neural representation. *Annu. Rev. Neurosci.* **24,** 1193–1216 (2001).
31. Zhaoping, L. *Theoretical understanding of the early visual processes by data compression and data selection. Network (Bristol, England)* **17,** (2006).
32. Coen-Cagli, R., Dayan, P. & Schwartz, O. Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Comput. Biol.* **8,** (2012).
33. Frazor, R. A. & Geisler, W. S. Local luminance and contrast in natural images. *Vision Res.* **46,** 1585–1598 (2006).
34. Karklin, Y. & Lewicki, M. S. A Hierarchical Bayesian Model for Learning Nonlinear Statistical Regularities in Nonstationary Natural Signals. *Neural Comput.* **17,** 397–423 (2005).
35. Parra, L., Spence, C. & Sajda, P. Higher-order statistical properties arising from the non-stationarity of natural signals. *Adv. Neural Inf. Process. Syst.* 786–792 (2001).
36. Ruderman, D. L. & Bialek, W. Statistics of natural images: Scaling in the woods. *Adv. Neural Inf. Process. Syst.* **73,** 551–558 (1994).
37. Portilla, J. & Simoncelli, E. P. Parametric texture model based on joint statistics of complex wavelet coefficients. *Int. J. Comput. Vis.* **40,** 49–71 (2000).
38. Gatys, L. A., Ecker, A. S. & Bethge, M. Texture Synthesis Using Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 262–270 (2015).
39. Karras, T., Aila, T., Laine, S. & Lehtinen, J. Progressive Growing of GANs for Improved Quality, Stability, and Variation. *ICLR 2018* (2018). doi:10.1002/joe.20070
40. Freeman, J., Ziemba, C. M., Heeger, D. J., Simoncelli, E. P. & Movshon, J. A. A functional and perceptual signature of the second visual area in primates. *Nat. Neurosci.* **16,** 974–81 (2013).
41. Okazawa, G., Tajima, S. & Komatsu, H. Image statistics underlying natural texture selectivity of neurons in macaque V4. *Proc. Natl. Acad. Sci.* **112,** E351–E360 (2015).
42. Rust, N. C. & DiCarlo, J. J. Selectivity and Tolerance (‘Invariance’) Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *J. Neurosci.* **30,** 12978–12995 (2010).
43. Atick, J. J. & Redlich, A. N. What Does the Retina Know About Natural Scenes? *Neural Comput.* **210,** 196–210 (1992).
44. Srinivasan, M. V., Laughlin, S. B. & Dubs, A. Predictive coding: a fresh view of inhibition in the retina. *Proc. R. Soc. London B* **216,** 427–59 (1982).
45. Marr, D. & Hildreth, E. Theory of edge detection. *Proc. R. Soc. London B* **207,** 187–217 (1980).
46. Zhaoping, L. Understanding Vision: Theory, Models, and Data. Oxford University Press, Oxford, UK (2014).
47. Barlow, H. B. Possible principles underlying the transformations of sensory messages. *Sens. Commun.* **6,** 217–234 (1961).
48. Attneave, F. Some informational apsects of visual perception. *Psychol. Rev.* **3,** (1954).
49. Shannon, C. E. A Mathematical Theory of Communication. *Bell Syst. Tech. J.* **27,** 379–423 (1948).
50. Field, D. J. What Is the Goal of Sensory Coding? *Neural Comput.* **6,** 559–601 (1994).
51. Bhandawat, V., Olsen, S. R., Gouwens, N. W., Schlief, M. L. & Wilson, R. I. Sensory processing in the Drosophila antennal lobe increases reliability and separability of ensemble odor representations. *Nat. Neurosci.* **10,** 1474–1482 (2007).
52. Laughlin, S. A simple coding procedure enhances a neuron’s information capacity. *Zeitschrift fur Naturforsch. - Sect. C J. Biosci.* **36,** 910–912 (1981).

53. Brinkman, B. A. W., Weber, A. I., Rieke, F. & Shea-Brown, E. How Do Efficient Coding Strategies Depend on Origins of Noise in Neural Circuits? *PLoS Comput. Biol.* **12,** 1–34 (2016).
54. Gjorgjieva, J., Sompolinsky, H. & Meister, M. Benefits of pathway splitting in sensory coding. *J. Neurosci.* **34,** 12127–44 (2014).
55. Kastner, D. B., Baccus, S. a. & Sharpee, T. O. Critical and maximally informative encoding between neural populations in the retina. *Pnas* **112,** 2533–8 (2015).
56. Field, D. J. Relations between the statistics of natural images and the response properties of cortical cells. *J. Opt. Soc. Am. A* **4,** 2379–94 (1987).
57. Ruderman, D. L. Origins of Scaling in Natural Images. *Vision Res.* **37,** 3385–3398 (1997).
58. Dan, Y., Atick, J. J. & Reid, R. C. Efficient Coding of Natural Scenes in the Lateral Geniculate Nucleus: Experimental Test of a Computational Theory. *J. Neurosci.* **16,** 3351–3362 (1996).
59. Franke, K. *et al.* Balanced excitation and inhibition decorrelates visual feature representation in the mammalian inner retina. *Nature* **542,** 439–444 (2017).
60. Pitkow, X. & Meister, M. Decorrelation and efficient coding by retinal ganglion cells. *Nat. Neurosci.* **15,** 628–635 (2012).
61. Vincent, B. T. & Baddeley, R. J. Synaptic energy efficiency in retinal processing. *Vision Res.* (2003). doi:10.1016/S0042-6989(03)00096-8
62. Atick, J. J. Could information theory provide an ecological theory of sensory processing? *Netw. Comput. Neural Syst.* **22,** 4–44 (2011).
63. Li, Z. & Atick, J. J. Efficient stereo coding in the multiscale representation. *Netw. Comput. Neural Syst.* **5,** 157–174 (1994).
64. Kuang, X., Poletti, M., Victor, J. D. & Rucci, M. Temporal encoding of spatial information during active visual fixation. *Curr. Biol.* **22,** 510–514 (2012).
65. Segal, I. Y. *et al.* Decorrelation of retinal response to natural scenes by fixational eye movements. *Proc. Natl. Acad. Sci.* **112,** 3110–5 (2015).
66. Boi, M., Poletti, M., Victor, J. D. & Rucci, M. Consequences of the Oculomotor Cycle for the Dynamics of Perception. *Curr. Biol.* **27,** 1268–1277 (2017).
67. Hyvärinen, A., Hurri, J. & Hoyer, P. O. *Natural Image Statistics-A Probabilistic Approach to Early Computational Vision. Computational Imaging and Vision* **39,** (2009).
68. Bell, A. J. & Sejnowski, T. J. The 'independent components'"of natural scenes are edge filters'. *Vision Res.* **37,** 3327–3338 (1997).
69. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381,** 607–609 (1996).
70. Rehn, M. & Sommer, F. T. A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields. *J. Comput. Neurosci.* **22,** 135–146 (2007).
71. Eichhorn, J., Sinz, F. & Bethge, M. Natural Image Coding in V1: How Much Use Is Orientation Selectivity? *PLoS Comput. Biol.* **5,** (2009).
72. Golden, J. R., Vilankar, K. P., Wu, M. C. K. & Field, D. J. Conjectures regarding the nonlinear geometry of visual neurons. *Vision Res.* **120,** 74–92 (2016).
73. Schwartz, O. & Simoncelli, E. P. Natural signal statistics and sensory gain control. *Nat. Neurosci.* **4,** 819–25 (2001).
74. Karklin, Y. & Lewicki, M. S. Emergence of complex cell properties by learning to generalize in natural scenes. *Nature* **457,** 83–86 (2009).
75. Lochmann, T., Ernst, U. A. & Deneve, S. Perceptual Inference Predicts Contextual Modulations of Sensory Responses. *J. Neurosci.* **32,** 4179–4195 (2012).
76. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2,** 79–87 (1999).
77. Spratling, M. W. Predictive Coding as a Model of Response Properties in Cortical Area V1. *J. Neurosci.* **30,** 3531–3543 (2010).
78. Zhu, M. & Rozell, C. J. Visual Nonclassical Receptive Field Effects Emerge from Sparse Coding in a Dynamical System. *PLoS Comput. Biol.* **9,** 1–15 (2013).
79. Berkes, P. & Wiskott, L. Slow feature analysis yields a rich repertoire of complex cell properties. *J. Vis.* **5,** 9–9 (2005).
80. Cadieu, C. F. & Olshausen, B. A. Learning intermediate-level representations of form and motion from natural movies. *Neural Comput.* **24,** 827–866 (2012).

81. Coen-Cagli, R. & Schwartz, O. The impact on midlevel vision of statistically optimal divisive normalization in V1. *J. Vis.* **13,** 1–20 (2013).
82. Hosoya, H. & Hyvarinen, A. A Hierarchical Statistical Model of Natural Images Explains Tuning Properties in V2. *J. Neurosci.* **35,** 10412–10428 (2015).
83. Lee, H., Ekanadham, C. & Ng, A. Y. Sparse deep belief net model for visual area V2. *Adv. Neural Inf. Process. Syst. 20* 873–880 (2008). doi:10.1.1.120.9887
84. Shan, H. & Cottrell, G. Efficient Visual Coding: From Retina To V2. *arXiv Prepr.* (2013).
85. Dayan, P., Sahani, M. & Deback, G. Adaptation and Unsupervised Learning. *Adv. Neural Inf. Process. Syst. 15* 237–244 (2003).
86. Hinton, G. E. & Ghahramani, Z. Generative models for discovering sparse distributed representations. *Philos. Trans. R. Soc. Lond. B. Biol. Sci.* **352,** 1177–90 (1997).
87. Wainwright, M. J. & Simoncelli, E. P. Scale mixtures of Gaussians and the statistics of natural images. *Adv. Neural Inf. Process. Syst.* **12,** 855–861 (2000).
88. Coen-Cagli, R., Kohn, A. & Schwartz, O. Flexible gating of contextual influences in natural vision. *Nat. Neurosci.* **18,** 1648–1655 (2015).
89. Li, Z. Contextual influences in V1 as a basis for pop out and asymmetry in visual search. *Proc. Natl. Acad. Sci.* **96,** 10530–10535 (1999).
90. Lettvin, J. Y., Maturana, H. R., McCulloch, W. S. & Pitts, W. H. What the Frog's Eye Tells the Frog's brain. *Proc. Natl. Acad. Sci.* 1940–1951 (1959).
91. Masland, R. H. & Martin, P. R. The unsolved mystery of vision. *Curr. Biol.* **17,** 577–582 (2007).
92. Nath, A. & Schwartz, G. W. Cardinal Orientation Selectivity Is Represented by Two Distinct Ganglion Cell Types in Mouse Retina. *J. Neurosci.* **36,** 3208–21 (2016).
93. Schwartz, G., Harris, R., Shrom, D. & Berry, M. J. Detection and prediction of periodic patterns by the retina. *Nat. Neurosci.* **10,** 552–554 (2007).
94. Krishnamoorthy, V., Weick, M. & Gollisch, T. Sensitivity to image recurrence across eye-movement-like image transitions through local serial inhibition in the retina. *Elife* e22431 (2017). doi:10.7554/eLife.22431
95. Franke, F. *et al.* Structures of Neural Correlation and How They Favor Coding. *Neuron* **89,** 409–422 (2016).
96. Zylberberg, J., Cafaro, J., Turner, M. H., Shea-Brown, E. & Rieke, F. Direction-Selective Circuits Shape Noise to Ensure a Precise Population Code. *Neuron* **89,** 369–383 (2016).
97. Rodieck, R. W. *The First Steps in Seeing*. (1998). doi:10.1001/archopht.117.4.550
98. Hecht, S. & Verrijp, C. Intermittent Stimulation By Light III. The relation between intensity and critical fusion frequency for different retinal locations. *J. Gen. Physiol.* 251 (1933).
99. Sinha, R. *et al.* Cellular and Circuit Mechanisms Shaping the Perceptual Properties of the Primate Fovea. *Cell* **168,** 413–426.e12 (2017).
100. Solomon, S. G., Martin, P. R., White, A. J. R., Rüttiger, L. & Lee, B. B. Modulation sensitivity of ganglion cells in peripheral retina of macaque. *Vision Res.* **42,** 2893–2898 (2002).
101. Oyster, C. W. & Barlow, H. B. Direction-selective units in rabbit retina: distribution of preferred directions. *Science (80-.).* **155,** 841–842 (1967).
102. Hughes, S. *et al.* Signalling by melanopsin (OPN4) expressing photosensitive retinal ganglion cells. *Eye* **30,** 247–254 (2016).
103. Hausen, K. & Egelhaaf, M. in *Facets of Vision* (eds. Stavenga, D. G. & Hardie, R. C.) 391–424 (Springer London, 1989).
104. O'Carroll, D. C., Bidwell, N. J., Laughlin, S. B. & Warrant, E. J. Insect motion detectors matched to visual ecology. *Nature* **382,** 63–66 (1996).
105. Krapp, H. G. & Hengstenberg, R. Estimation of self-motion by optic flow processing in single visual interneurons. *Nature* **384,** 463–466 (1996).
106. Longden, K. D., Wicklein, M., Hardcastle, B. J., Huston, S. J. & Krapp, H. G. Spike Burst Coding of Translatory Optic Flow and Depth from Motion in the Fly Visual System. *Curr. Biol.* **27,** 3225–3236.e3 (2017).
107. Franz, M. O. & Krapp, H. G. Wide-field, motion-sensitive neurons and matched filters for optic flow fields. *Biol. Cybern.* **83,** 185–197 (2000).
108. Kohn, J. R., Heath, S. L. & Behnia, R. Eyes Matched to the Prize : The State of Matched Filters in Insect Visual Circuits. *Front. Neural Circuits* **12,** 26 (2018).

109. Sabbah, S. *et al.* A retinal code for motion along the gravitational and body axes. *Nature* (2017). doi:10.1038/nature22818
110. Gauvain, G. & Murphy, G. J. Projection-Specific Characteristics of Retinal Input to the Brain. *J. Neurosci.* **35,** 6575–6583 (2015).
111. Burge, J. & Jaini, P. *Accuracy Maximization Analysis for Sensory-Perceptual Tasks: Computational Improvements, Filter Robustness, and Coding Advantages for Scaled Additive Noise*. *PLoS Computational Biology* **13,** (2017).
112. Geisler, W. S., Najemnik, J. & Ing, A. D. Optimal stimulus encoders for natural tasks. *J. Vis.* **9,** 17–17 (2009).
113. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. *Adv. Neural Inf. Process. Syst.* 1–9 (2012). doi:http://dx.doi.org/10.1016/j.protcy.2014.09.007
114. Lecun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521,** 436–444 (2015).
115. Maheswaranathan, N. *et al.* Deep learning models reveal internal structure and diverse computations in the retina under natural scenes. *bioRxiv* (2018). doi:10.1101/340943
116. Kriegeskorte, N. Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annu. Rev. Vis. Sci.* **1,** 417–446 (2015).
117. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19,** 356–365 (2016).
118. Cadena, S. A. *et al.* Deep convolutional models improve predictions of macaque V1 responses to natural images. *bioRxiv Prepr.* 201764 (2017). doi:10.1101/201764
119. Cichy, R. M., 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. *Sci. Rep.* **6,** (2016).
120. Pospisil, D., Pasupathy, A. & Bair, W. Comparing the brain's representation of shape to that of a deep convolutional neural network. *Proc. 9th EAI Int. Conf. Bio-inspired Inf. Commun. Technol. (formerly BIONETICS)* 516–523 (2016). doi:10.4108/eai.3-12-2015.2262486
121. Young, M. & Yamane, S. Sparse population coding of faces in the inferotemporal cortex. *Science* **256,** 1327–1331 (1992).
122. Fukushima, K. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biol. Cybern.* **36,** 193–202 (1980).
123. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2,** 1019–25 (1999).
124. Razavian, A. S., Azizpour, H., Sullivan, J. & Carlsson, S. CNN Features Off-the-Shelf: An Astounding Baseline for Recognition. *CVPR2014 Work.* 806–813 (2014).
125. Szegedy, C. *et al.* Intriguing properties of neural networks. *arXiv Prepr.* arXiv:1312.6199v4 (2013). doi:10.1021/ct2009208
126. Ullman, S., Dorfman, N. & Harari, D. Discovering 'containment': from infants to machines. *arXiv Prepr.* arXiv:1610.09625v1 (2016).
127. Goodfellow, I. J., Shlens, J. & Szegedy, C. Explaining and Harnessing Adversarial Examples. *arXiv Prepr.* arXiv:1412.6572v3 (2014).
128. Nayebi, A. & Ganguli, S. Biologically inspired protection of deep networks from adversarial attacks. *arXiv Prepr.* arXiv:1703.09202v1 (2017).
129. Brendel, W. & Bethge, M. Comment on 'Biologically inspired protection of deep networks from adversarial attacks'. *arXiv Prepr.* arXiv:1704.01547v1 (2017).
130. Nishimoto, S. & Gallant, J. L. A Three-Dimensional Spatiotemporal Receptive Field Model Explains Responses of Area MT Neurons to Naturalistic Movies. *J. Neurosci.* **31,** 14551–14564 (2011).
131. Berardino, A., Ballé, J., Laparra, V. & Simoncelli, E. P. Eigen-Distortions of Hierarchical Representations. *arXiv Prepr.* arXiv:1710.02266v3 (2017).
132. Han, S. & Vasconcelos, N. Object recognition with hierarchical discriminant saliency networks. *Front. Comput. Neurosci.* **8,** 1–20 (2014).
133. Ren, M., Liao, R., Urtasun, R., Sinz, F. H. & Zemel, R. S. Normalizing the Normalizers: Comparing and Extending Network Normalization Schemes. *ICLR 2017* (2017).
134. Giraldo, L. G. S. & Schwartz, O. Flexible normalization in deep convolutional neural networks. in *COSYNE abstracts* (2017).
135. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: A better model

of biological object recognition. *Front. Psychol.* **8,** 1–14 (2017).

136. Shwartz-Ziv, R. & Tishby, N. Opening the Black Box of Deep Neural Networks via Information. *arXiv Prepr.* arXiv:1703.00810v3 (2017).
137. Chalk, M., Marre, O. & Tkačik, G. Toward a unified theory of efficient, predictive, and sparse coding. *Proc. Natl. Acad. Sci.* **115,** 186–191 (2018).
138. Sederberg, A. J., MacLean, J. N. & Palmer, S. E. Learning to make external sensory stimulus predictions using internal correlations in populations of neurons. *Proc. Natl. Acad. Sci.* **115,** 201710779 (2018).
139. Kuleshov, V. & Ermon, S. Deep Hybrid Models: Bridging Discriminative and Generative Approaches. *Uncertain. Ai* (2017).
140. Park, I. M. & Pillow, J. W. Bayesian Efficient Coding. *bioRxiv* 178418 (2017). doi:10.1101/178418
141. Ballé, J., Laparra, V. & Simoncelli, E. P. End-to-end Optimized Image Compression. *arXiv Prepr.* arXiv:1611.01704v3 (2016). doi:10.1016/S0197-3975(03)00059-6
142. Hirayama, J., Hyvärinen, A. & Kawanabe, M. SPLICE: Fully Tractable Hierarchical Extension of ICA with Pooling. *Proc. 34th Int. Conf. Mach. Learn.* **70,** 1491–1500 (2017).
143. Scholte, H. S., Losch, M. M., Ramakrishnan, K., de Haan, E. H. F. & Bohte, S. M. Visual pathways from the perspective of cost functions and deep learning. *BioRxiv* 146472 (2017).
144. Kell, A. J. E. *et al.* A Task-Optimized Neural Network Replicates Human Auditory Behavior, Predicts Brain Responses, and Reveals a Cortical Processing Hierarchy. *Neuron* **98,** 1–15 (2018).
145. Chengxu Zhuang, D. Y. Using multiple optimization tasks to improve deep neural network models of higher ventral cortex. in *COSYNE abstracts* (2018).
146. Van Der Linde, I., Rajashekar, U., Bovik, A. C. & Cormack, L. K. DOVES: a database of visual eye movements. *Spat. Vis.* **22,** 161–77 (2009).
147. Rucci, M. & Victor, J. D. The unsteady eye: An information-processing stage, not a bug. *Trends Neurosci.* **38,** 195–206 (2015).
148. Thomsom, M. G. A. Visual coding and the phase structure of natural scenes. *Netw. Comput. Neural Syst.* **10,** 123–132 (1999).

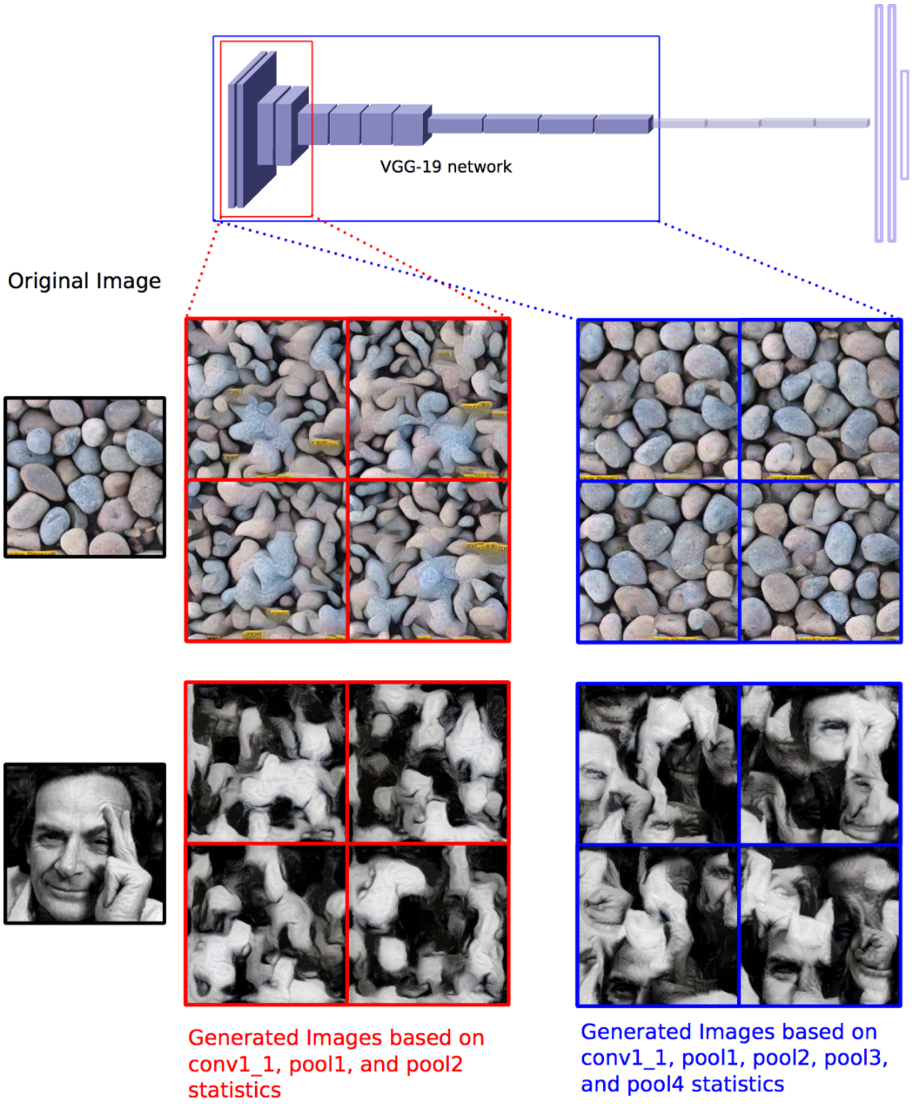
VGG-19 network
Original Image
Generated Images based on conv1_1, pool1, and pool2 statistics
Generated Images based on conv1_1, pool1, pool2, pool3, and pool4 statistics

Figure 1: Texture Synthesis based on Deep Convolutional Neural Networks. The activations of different layers of a DNN trained for object recognition can be employed to capture statistics of textures beyond second order[38]. Texture synthesis is accomplished by numerical optimization of the pixel values of an image that matches the statistics of a reference image (Original Image enclosed in black). Statistics can be obtained from activation values at different stages of the deep DNN. Images enclosed in red are synthesized by considering only activations from the first and second pooling stages of the DNN, whereas images enclosed in blue include the third and fourth pooling stages in their statistics. In the case of the inhomogeneous images (bottom row) the texture generation tiles local features in scrambled places that will match the activation statistics that have been averaged over space. Original images outlined in black (Feynman portrait and rocks) are from http://www.cns.nyu.edu/~lcv/texture/ and are used with permission from Eero Simoncelli.

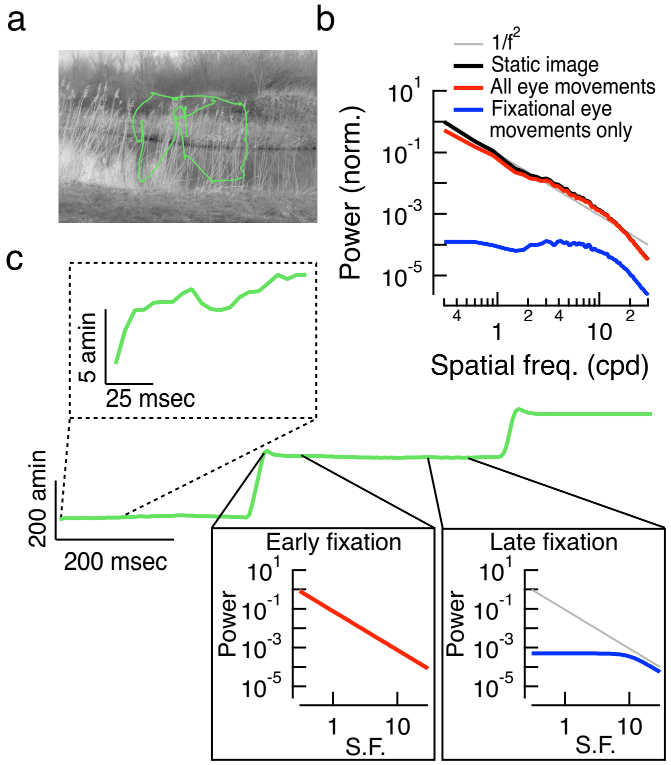

Figure 2: Efficient coding strategies rely on self-generated movement. (a) A natural image and measured human eye movement trajectory from [146]. An observer will explore a scene using large, ballistic changes in fixation called saccades. In the time between saccades, observers make much smaller, involuntary eye movements called fixational eye movements (for review, see [147]). (b) Using these eye movement data, we can reconstruct the time-varying image on the retina into a naturalistic movie stimulus. We summed the Fourier spatial power spectra of each frame of this movie, resulting in a roughly $1/f^2$ power law scaling, which is characteristic of static natural images (black trace). Following the analysis in [64], we then measured spatial power spectra for the dynamic component of the natural movie. To produce these spatial power spectra, we computed the spatiotemporal power spectrum of a movie and summed over all *non-zero temporal frequencies*. Fixational eye movements simply shift much of the power, except that at the lowest spatial frequencies, to higher temporal frequencies. The removal of the temporal DC component of the movie thus selectively removes low spatial frequency content, and the result is a whitened spatial power spectrum (Fig. 2b, blue trace). Importantly, this result relies on fixational eye movements and *not* saccades. When saccades are included in the natural movie stimulus, considerable low spatial frequency content is still present at nonzero temporal frequencies, so whitening does not occur (Fig. 2b, red trace). (c) The position (in one dimension) of the eye as a function of time is shown by the green trace. Examining the eye position at a finer time scale (dashed inset) reveals smaller fixational eye movements. Boi et al.[66] suggested that during a saccade, the dynamic spatial frequency content of natural images follows the familiar $1/f^2$ power law scaling (left inset, red trace). As the fixation proceeds, the retinal input is whitened (right inset, blue trace). Between saccades (when the image is relatively stable), any low spatial frequency content is present mostly in the temporal DC component of the input. In other words, the large-scale spatial structure isn't changing very much within a single fixation. The whitening effect of fixational eye movements will depend on how completely (and how quickly) a visual neuron adapts to the (mostly static) low spatial frequency content imposed by each new fixation.

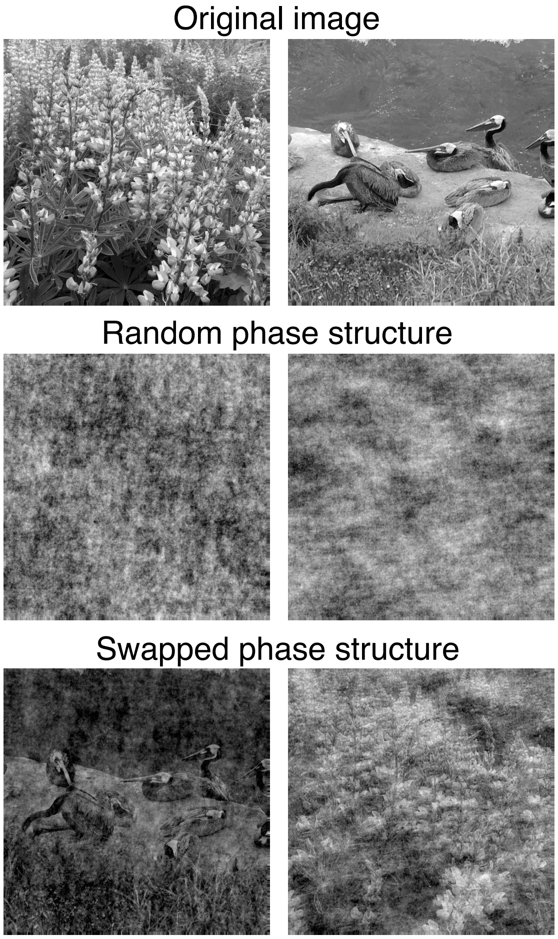

Figure 3: Beyond-pairwise statistics contribute to complex structure in natural images. Top row: Two grayscale natural images. Middle row: The natural images above with randomized phase spectra. Both of these images have the roughly $1/f^2$ spatial power spectrum characteristic of natural images, yet appear quite unnatural. Bottom row: The natural images with their phase spectra swapped, such that the image on the left now has the phase spectrum of the original image on the right, and vice-versa. See [30,148]. Original photographs were taken by the authors.

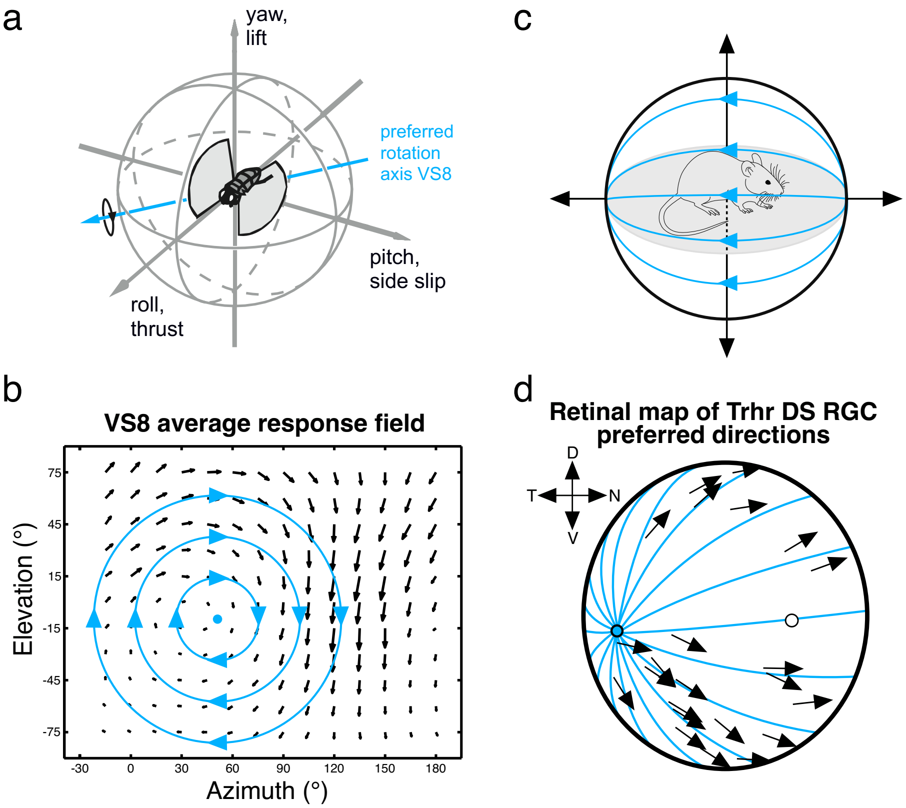

Figure 4: Motion sensitive neurons encode self-movement across the animal kingdom. (a) Schematic showing a fly in flight. (b) Local motion receptive field of the VS8 neuron in the blowfly Calliphora. The direction of each arrow indicates the local preferred direction, and the length of each arrow indicates the cell's motion sensitivity. This local motion receptive field corresponds to the optic flow pattern that would result from a rotation of the animal. The rotation axis around which the fly would need to turn to maximally activate this neuron is indicated in (a). Data & schematic provided by Holger Krapp. (c) Schematic showing a mouse ambulating in a forward direction. The resulting visual input is an optic flow pattern emanating from a singularity directly ahead of the animal (blue lines). (d) Direction preferences of a population of DS RGCs in mouse retina are overlaid on the retinal surface. Forward motion optic flow moves outward from a point in the retina (blue lines). The direction preferences of this cell type roughly align with the optic flow lines that result from forward motion. Other DS RGC types similarly respond to optic flow resulting from other directions of motion of the animal. Data redrawn from [109].

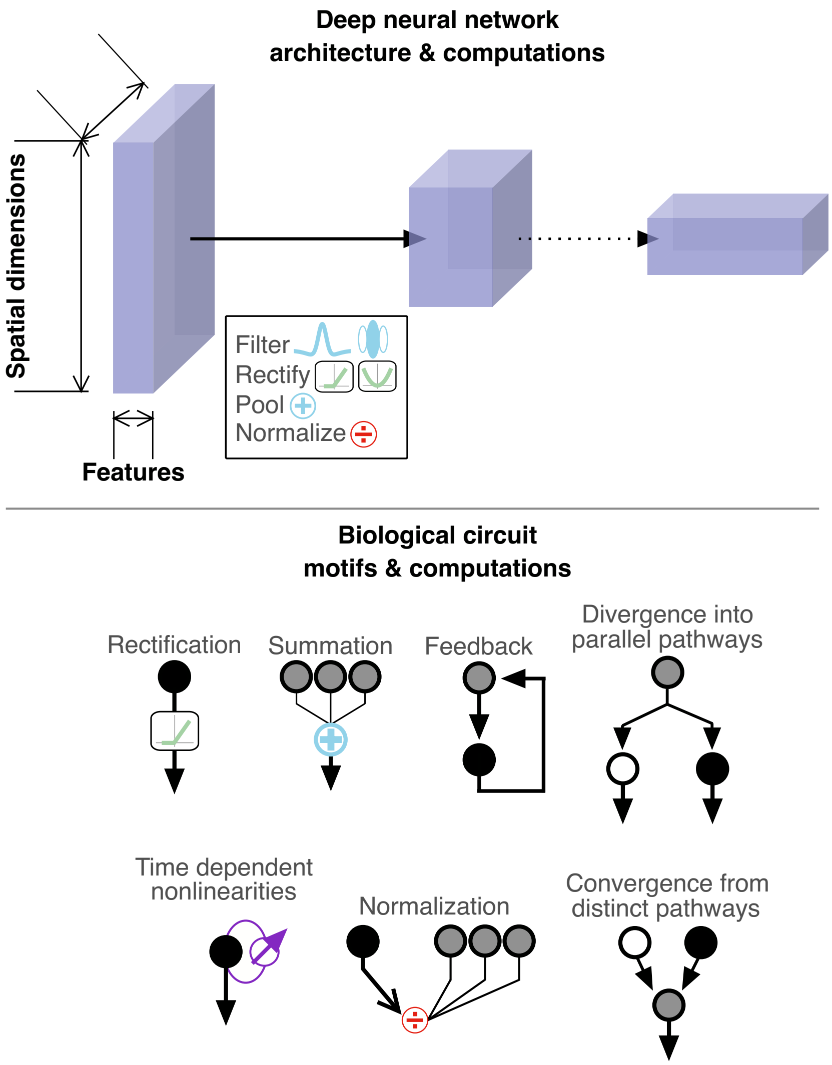

Figure 5: Deep neural networks reflect some, but not all, architectural and computational motifs found in neural circuits. Top: Deep neural networks are composed of multiple, connected layers. Several basic computations are performed within each layer. Bottom: examples of common circuit motifs and computations observed in neural circuits. Some of these examples are well-represented by many DNNs (e.g. pooling / filtering), others can be included in DNNs but their precise nature & location are not necessarily well reflected (e.g. rectification or normalization), and still others are excluded from most DNNs (e.g. time-dependent nonlinearities).