Normalization and Pooling in Hierarchical Models of Natural Images

Luis G Sanchez-Giraldo, Md Nasir Uddin Laskar, and Odelia Schwartz*
Computational Neuroscience Lab, Dept. of Computer Science, University of Miami, FL 33146

Divisive normalization and subunit pooling are two canonical classes of computation that have become widely used in descriptive (what) models of visual cortical processing. Normative (why) models from natural image statistics can help constrain the form and parameters of such classes of models. We focus on recent advances in two particular directions, namely deriving richer forms of divisive normalization, and advances in learning pooling from image statistics. We discuss the incorporation of such components into hierarchical models. We consider both hierarchical unsupervised learning from image statistics, and discriminative supervised learning in deep convolutional neural networks (CNNs). We further discuss studies on the utility and extensions of the convolutional architecture, which has also been adopted by recent descriptive models. We review the recent literature and discuss the current promises and gaps of using such approaches to gain a better understanding of how cortical neurons represent and process complex visual stimuli.

Highlights:

- Subunit pooling and normalization are building blocks of hierarchical cortical models.
- Image statistics models predict when normalization is recruited in primary cortex.
- Hierarchical models can capture cortical data in secondary and higher cortex.
- There is potential for progress on when normalization is recruited in higher cortex.
- Convolutional subunit structure yields a key representation property of equivariance.

Introduction

There has been long standing interest in understanding the relation between the statistics of natural stimuli and sensory processing [1], [2], [3], [4], [5], [6]. This can provide an interpretive or normative perspective, which complements descriptive and mechanistic modeling approaches [7]. Recent advances in image statistics and in machine learning rooted in deep learning, provide an opportunity for a richer understanding of how visual cortical neurons represent and make inferences about complex images.

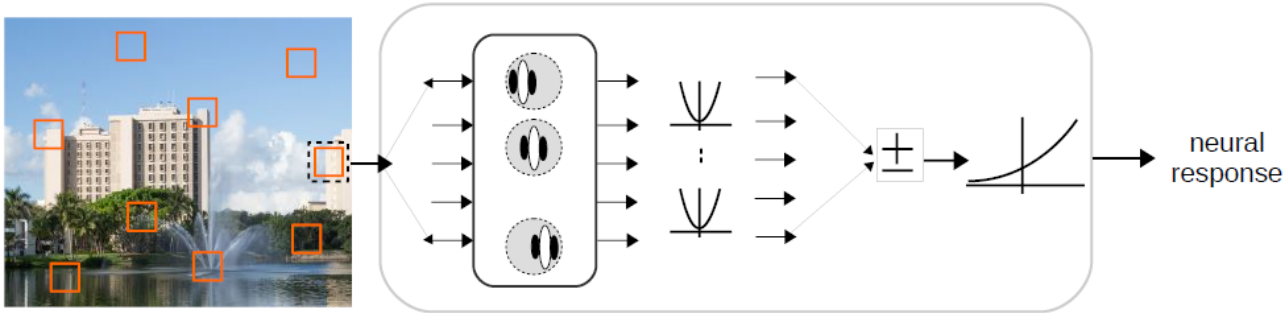

Figure 1. Subunit pooling model. The input signal goes through a bank of linear-nonlinear subunits (i.e., linear filter activation followed by a point nonlinearity such as squaring), which are then spatially pooled via a weighted sum. A rectified non-linearity is applied to obtain the final response of a higher level neural unit (schematic is after [8], [9••], [10••], [11]).

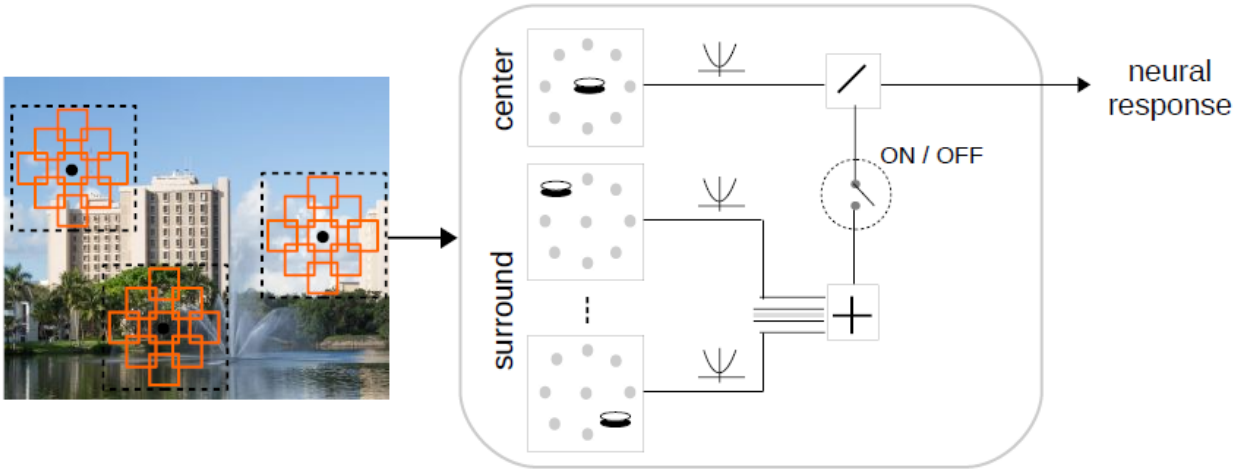

Figure 2. Flexible normalization model of center-surround activations. The response of a neural unit is given by its linear-nonlinear output, divided by a weighted sum of other linear-nonlinear units in surrounding spatial locations (extending beyond the classical receptive field of the given neural unit in the center location). The on/off switch constitutes a gating determined from natural image statistics considerations, such that normalization is only present to the degree that center and surround activations are deemed statistically dependent for the given image patch and filters (schematic is after [12], [13••], for normalization, see [14]).

We focus on advances in learning joint statistical properties of natural images, with emphasis on two main classes of nonlinear models that have been prominent in recent literature: (i) Neural subunit pooling models (Figure 1) have become popular for describing and fitting cortical neural data, including primary visual cortex (area V1; [8], [11]), secondary visual cortex (area V2; [10••]), and higher cortical areas [9••]. In such models, a weighted sum of the rectified or squared responses of linear filters in a lower level of the model (essentially making up the subunits of the given neural unit) are pooled together to give rise to the neural unit response at a higher level (Figure 1). The spatial extent of the pooling is typically within the classical receptive field of the (higher level) neural unit. Considerations of natural image statistics can complement descriptive models by constraining what subunits may pool together to form more complex representations; (ii) Divisive normalization is a ubiquitous nonlinearity in descriptive models of neural processing [15], [14]. In normalization models, the response of a given neural unit is divided by a (pooled) weighted sum of the rectified or squared responses of other neural units with receptive fields in spatially overlapping and surrounding spatial locations. Normalization models have therefore addressed not only phenomena within the classical receptive field of a neuron, but have also typically spanned a larger spatial extent than subunit pooling models. The contextual surround nonlinearly modulates the responses of cortical neurons, and can lead to striking perceptual effects such as illusions [16], [17]. For instance, in the tilt illusion, the perceived orientation of a bar in a center location appears tilted away from the orientation of a contextual surround stimulus. In recent years, image statistics approaches have resulted in richer models of divisive normalization that go beyond descriptive accounts, suggesting when normalization is recruited in area V1 (Figure 2).

The main components of the above models, namely pooling and normalization, are common building blocks of hierarchical cortical models. We discuss recent progress in the context of both unsupervised learning models based on the statistics of natural images, and supervised discriminative convolutional neural networks in which the parameters are learned based on the goal of discriminating objects [18], [19].

V1 models of pooling and normalization from image statistics

V1-like filters derived from Independent Component Analysis (ICA) or Sparse Coding have striking nonlinear dependencies [20], [21], [22], [23], [24]. These statistics have been utilized in two main ways, as detailed below. On one hand, the joint statistics can be utilized to learn what subunits are statistically coordinated in pooling models, resulting for instance in complex cell models. On the other hand, from an efficient coding perspective, maximizing information transfer requires making responses of neural unit activations statistically independent. Reducing the statistical dependencies has been approached with normalization models.

Pooling models can be learned using extensions of ICA and Sparse Coding that relax the independence assumption. In Independent Subspace Analysis (ISA), subspaces are assumed to be independent, but subunits within a subspace can be statistically dependent ([5] and references therein). This results in learning V1 “complex cells”, that are invariant to spatial transformations of the input. In particular, subspaces resembling translation invariant complex cells are made up of the sum of squares of shifted replicas of oriented V1-like filters. Allowing freedom in the combination of low level units that give rise to the variance of high level units, results in richer patterns of statistical coordination of orientations [25], [26], [27]. Other approaches have extended linear combination rules to address nonlinearities of occlusion [28], [29], [30].

Divisive normalization models derived from consideration of the joint statistics have particularly focused on understanding nonlinear contextual phenomena in V1 extending outside the classical receptive field [31], [22], [32], [33], [34], [12]. Divisive normalization models can be motivated from an efficient coding perspective. As noted earlier, V1-like filter outputs to images (including those obtained from ICA or Sparse Coding) have striking joint statistical variance dependencies even for spatially non overlapping receptive fields. A nonlinear transformation such as division can reduce these dependencies [21], [22], [35].

Related ideas have cast this as inference about the local receptive field properties given the statistically dependent (multiplicative) surround [12]. An example of a multiplicative model is lightness (or color) constancy, in which the reflectance in a given spatial location is multiplied by a global illuminant covering a larger spatial area [36], and so estimating the local reflectance can be obtained via division. More generally, global properties such as contrast or orientation structure can provide spatial coordination across receptive fields. Division can reduce this coordination and so highlight the local receptive field property that is invariant to properties such as global contrast. These ideas can be formalized by a class of statistical model known as the Gaussian Scale Mixture [37], [27]. In this model, the statistical coordination between receptive fields arises via the multiplication of local Gaussian variables by a global shared mixer variable. Removing the statistical dependencies therefore amounts to the reverse operation, namely dividing by an estimate of the shared mixer variable. Normalization has more broadly been motivated from the normative perspective of probabilistic marginalization that is invariant to nuisance parameters [38]. Another recent normative model suggests that each neural unit response is divided by other neural units responses prior to pooling, as optimal cue combination in the face of signal-dependent noise [39••].

Recent normalization models have utilized richer statistical properties of images, by taking account of the non homogeneity of images, i.e. that at different locations in the image, dependencies between the center and surround locations can be different, as viewed through the lens of V1-like neural units [12], [40] (see also other approaches for adjusting to the context across space and/or time; [32], [41], [42]). This has resulted in more sophisticated models of normalization (which we denote flexible normalization), whereby neural unit activations are divided by a weighted sum of surrounding unit activations only to the degree that center and surround are inferred to be statistically dependent (Figure 2; [12]). This also relates to visual salience as a breakdown of statistical homogeneity in area V1 [43]. Predictions of the flexible normalization model were tested neurophysiologically in area V1 with patches of images extending beyond the classical receptive field [13••], suggesting that normalization is gated by inference about statistical dependencies in images. The advantage of this approach is that the combinatorial search over a huge space of descriptive forms of normalization is avoided since the image statistics framework provides additional constraints (from consideration of the properties of images) that result in a better fit to the neural data.

Jaini and Burge [44••] have derived pooling models that incorporate normalization, based on an optimal Bayesian decoding goal of determining which stimulus properties are most useful for a given task. Quadratic models akin to those used in descriptive subunit models [45], [9••] arise from their analysis. The assumption that neural unit responses given the hidden task variable should be Gaussian distributed lends to a simple estimation process. The model also includes a form of divisive normalization acting on the input stimuli (rather than after the filter responses are computed) that serves as a simple form of contrast normalization.

Hierarchical models targeting V2-like neural units

Area V2 contains more complex representations that combine the features captured in V1, but the exact nature of the representation has remained unclear. Nevertheless, a number of properties have been emerging for V2 from neurophysiology studies, including sensitivity to combinations of edges [46], [47]; figure ground [48], [49], [50]; sensitivity to textures [51••], [52]; cross orientation suppression formulations in V2 [10••]; and other changes that occur between V1 and V2 [53], [54], [55].

Hierarchical image statistics models can capture edge combinations, for instance, by incorporating two layers of sparse coding [56], as well as with other approaches [33], [57], [58••]. But can hierarchical models obtain a larger repertoire of V2 like units, beyond edge combinations and corner units? Similar questions have also come up in V1 studies, in which aspects such as hard as opposed to soft sparseness [59] and highly overcomplete representations [60] have been suggested as important for obtaining more diverse receptive fields. For V2, Hosoya and Hyvärinen [58••] have proposed introducing a significant dimensionality reduction after a V1 complex cell layer, followed by an additional expansive (overcomplete) sparse coding [61]. The model has been comprehensively compared to a number of properties observed in V2, accounting for neurophysiology experiments of local orientation integration [47] and length and width suppression [55]. Cagli *et al*. [57] have shown that in a two layer model, incorporating flexible divisive normalization in the first layer prior to pooling versus an equivalent model without normalization, makes more apparent the linear dependencies in the second layer and leads to a richer combination of units in the second layer. This goes beyond corner units, capturing some texture boundaries. The model with flexible normalization achieves better performance on object recognition and on a figure ground task, compared to the model without normalization.

Recent neurophysiology studies have shown that a distinguishing factor between area V1 and V2, is the sensitivity to the high order structure of textures in V2 but not V1 [51••], [52]. These studies demonstrate this sensitivity by comparing the responses of cortical neurons to naturalistic texture stimuli, versus spectrally matched noise images which lack the high order statistical structure. This presents a rich set of data that can be tested against hierarchical models. Recent work has shown that early stages of discriminative CNN models, which we will discuss in more detail below, can capture some of the texture sensitivity of the cortical V2 versus V1 data [62], [63] (see also [64]). It is worth noting that a collective testing of the V2 properties that have emerged from neurophysiology has not been tapped into by any single model discussed above.

Multi-layer hierarchical models

Sparse coding can be extended to multiple hierarchical layers [65], [66], but higher layers have not been tested against neural data. Hirayama *et al*. [67] have recently hierarchically extended ISA to multiple layers, by designing a tractable generative model for which subspaces in each layer may be dependent on others via higher layer hidden variables. This leads in the second layer to some similar properties of [58••], and allows extension to higher layers. Deep CNN models, which include pooling and sometimes local normalization, can capture aspects of cortical data across the hierarchy [18], [19]. This may be related to a gradual untangling of object manifolds [68], [69] (see also [70], [71]). Two main stages of computation take place in this untangling: a dimensionality expansion that linearizes the problem, and a second stage of dimensionality reduction. Pooling, in particular average pooling, falls in the second category.

Some hierarchical models have incorporated normalization beyond the first layer, but mostly for image processing and computer vision applications. Balle *et al.* [72••], [73] have proposed a multilayer unsupervised learning model for image compression that incorporates a joint nonlinear transformation that seeks to Gaussianize the data in each layer, amounting to a generalized divisive normalization. Spratling [74] has extended unsupervised predictive coding models of V1 to include two or more levels, with application to image recognition. CNNs have included various restricted forms of divisive normalization, typically either to help training or to improve object recognition. Local response normalization in AlexNet [75] (see also [76]) normalizes groups of spatially overlapping units, akin to cross orientation suppression [15]. In batch normalization [77], each single neural unit response is normalized by the mean and standard deviation of a given batch over time. In layer normalization [78], all units in a given layer are normalized by the mean and standard deviation. In Ren *et al*. [79], normalization by center and surround units is used to improve object recognition. Han and Vasconcelos proposed a more sophisticated V1-motivated divisive normalization model for deep neural networks [80], [81] to improve object recognition. Interestingly, recent approaches geared largely at computer vision raise some of the same questions that have been studied longer in the context of computational neuroscience and V1: How should normalization operate locally and more broadly in terms of spatial context? What units should group together in the normalization pool?

Contextual influences beyond the classical receptive field have been most widely studied for area V1. Understanding when normalization is recruited in cortical area V2 (and indeed in higher areas) is more challenging (see recent neurophysiology studies of [82]). We propose to incorporate contextual normalization models inspired by V1 image statistics modeling directions, into hierarchical models of V2 for application to neuroscience [83]. We expect this will allow us to make predictions about when normalization is recruited in V2.

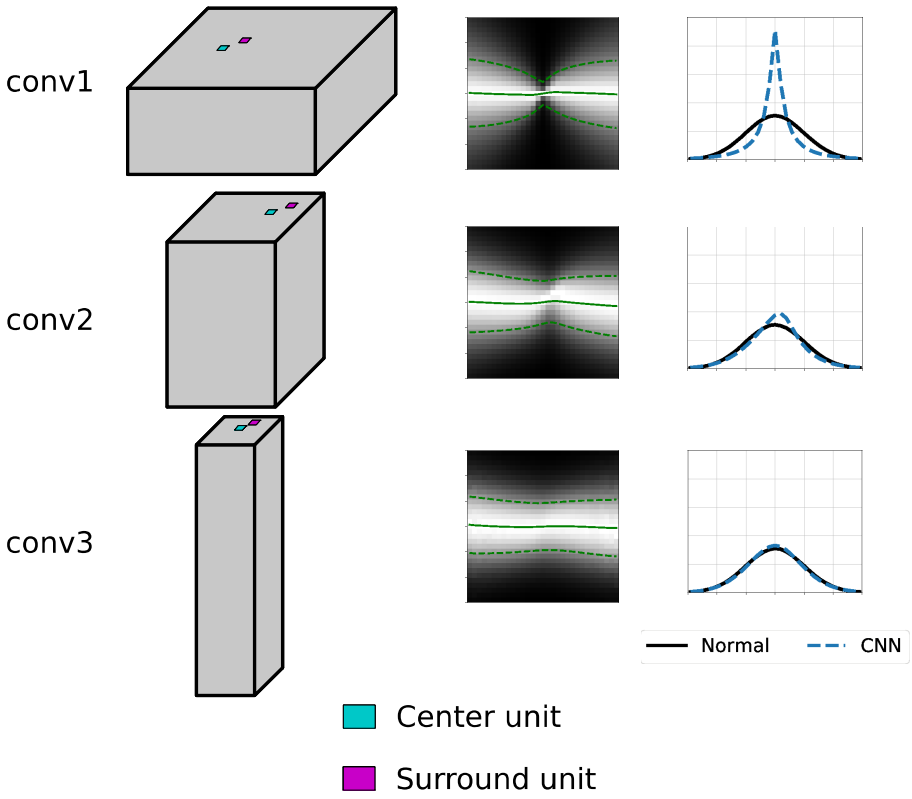

Figure 3. Spatial dependencies are reduced for higher layers of deep convolutional neural networks. The employed CNN is AlexNet. Activation maps for the first three convolutional layers (conv1, conv2, conv3) are depicted by the gray boxes on the left side of the figure. The width and depth of each box represent the spatial locations, and the height represents the number of filters. For higher layers, spatial resolution is reduced and at the same time the number of filters are increased. Here, we display (as an average over all filters in a given layer) the normalized conditional histograms (bowtie plots) between the activation values obtained by convolution at two spatial locations that overlap, but no more than 30%. We also show the average marginal distributions.

Figure 3 depicts the joint statistical dependencies across space and the marginal statistics in response to images for different layers of AlexNet [75]. The first two layers exhibit joint statistical dependencies between the responses of neural units in center and surround locations, and sparse marginal statistics, similar to what has been previously observed for V1-like filter outputs. This suggests that flexible surround normalization can be used to reduce the statistical dependencies and Gaussianize the marginals. In addition, it is interesting to note that although efficient coding is not set as a goal, even without surround normalization, the joint statistics become more independent and the marginal statistics more Gaussian on average as one proceeds from the first layer to higher layers of AlexNet. This echoes the discussion in [73], that hierarchical models without normalization may perform with similar accuracy, but requiring more layers to do so--a suggestion that needs more thorough testing.

Recent work has also focused on more theoretical aspects of CNN models, including the role of convolution. Deep CNNs that perform tasks like object recognition often include max pooling and average pooling. Intuitively, these forms of pooling applied to the convolutional architecture are suitable for translation invariance. This assumption allows to reduce the dimensionality of the parameters that need to be learned. Approaches for visualizing invariances in deep CNNs have revealed both shift invariance and other forms of invariance [84]. Theoretical approaches for understanding deep convolutional networks have pointed out that a main property of the convolutional architectures is equivariance (a generalization of invariance), i.e. that the representation changes in predictable ways to group actions [85••], [86], [87], [88]. Translation is one of the most basic examples of a group action, but other transformations include rotations of the image plane and changes in scale (see also [89]).

It has been argued that invariance alone at early levels of detecting simple features can hurt selectivity at higher levels. Equivariance, on the other hand, provides a more complete account of *what* simple features are present and *how* these are instantiated in the images. Recent extensions of CNN models known as capsule networks [90], [91] exploit the equivariance property. A main difference with conventional pooling, is that the outputs not only provide information about the likelihood of a feature being present which is closer to what max pooling provides, but also about the transformation taking place in the feature. One potential way that pooling and normalization can combine is motivated by the role of equivariance in the stimuli representation. Similar to population coding, subunits can be pooled to retrieve the instantiation parameters (such as the location) of a feature in the form of a neural unit (simple cell) response. Normalization would then guarantee that the response is robust to variations in the input.

Discussion

In recent years, there has been a convergence of similar architectural components (convolution, subunit pooling, and normalization) in both descriptive and normative visual cortical models. We have suggested that for divisive normalization models in V1, rather than combinatorially searching the space of possibilities for computing the normalization signal, normative image statistics approaches can add value to descriptive accounts by making predictions about when normalization is recruited for natural stimuli [13••]. In the context of subunit pooling models, an interesting direction for combining descriptive and normative perspectives, is through transfer learning. Subunit groups capturing properties of natural stimuli can be learned from much larger image datasets with normative approaches, and the subsequent pooling or normalization weights could be tuned to neural response predictions.

In this review, we have discussed normalization from various normative perspectives. So what is normalization good for? Normalization is considered a canonical computation in the brain [14]. From an efficient coding standpoint, neurons have a limited response range, a limit to the range of stimuli they can respond to differentially. Normalization in the visual system to the lightness or contrast level can set the responses of neurons to more finely cover the current range. Another way to think about normalization, is that neural responses become invariant to global properties such as lightness, contrast, or more complex properties such as orientation texture. We have motivated this from the point of view of a generative statistical model. Normalization may serve to reduce predictable information, which may relate to highlighting salient aspects in the scene [6]. In addition, normalization may help in recognizing objects invariant to such global properties. Interestingly, restricted forms of normalization have been incorporated into deep convolutional neural networks for the purpose of improving object recognition. Normalization may help by reducing some invariances and equalizing the response range of the neural units. But in recent years, some very deep architectures perform well without normalization [92]–[94]. Indeed, it is possible that very deep networks can better estimate the transformations required for invariant object recognition and so bypass the need for normalization at all. The brain, in contrast, may incorporate richer nonlinearities and sacrifice on the depth. However, there is a need to study more sophisticated forms of normalization by contextual information in such networks. More importantly, we believe that normalization will be useful for generalizing to a broader range of images and tasks that were not in the original training set, a feat that humans are better at than artificial neural networks [95].

We have suggested that hierarchical models now have the potential to push forward progress in understanding cortical representation and processing in mid level cortical areas. In particular, we highlighted area V2 as a potentially tractable goal. Recent hierarchical models have started to make progress in understanding V2 processing inside the classical receptive field. In terms of contextual surround influences, most modeling work has focused on cortical area V1. However, a number of testable hierarchical models are emerging. We believe there is now the potential to incorporate learning of surround normalization at higher levels of hierarchical models, and therefore to make headway in understanding contextual influences in V2.

Acknowledgements

This work was supported by the National Science Foundation (NSF), grant 1715475.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

• of special interest

•• of outstanding interest

[1] H. B. Barlow, “Possible Principles Underlying the Transformations of Sensory Messages,” in *Sensory Communication*, MIT press, 1961, pp. 217–234.

[2] F. Attneave, “Some Informational Aspects of Visual Perception,” *Psychol. Rev.*, vol. 61, no. 3, pp. 183–193, 1954.

[3] E. P. Simoncelli and B. A. Olshausen, “Natural Image Statistics and Neural Representation,” *Annu. Rev. Neurosci.*, vol. 24, pp. 1193–1216, 2001.

[4] M. S. Lewicki, B. A. Olshausen, A. Surlykke, and C. F. Moss, “Scene analysis in the natural environment,” *Frontiers in Psychology*. 2014.

[5] A. Hyvärinen, J. Hurri, and Patrik O. Hoyer, *Natural Image Statistics: A probabilistic approach to early computational vision*. Springer, 2009.

[6] L. Zhaoping, *Understanding Vision*. Oxford University Press, 2014.

[7] P. Dayan and L. F. Abbott, *Theoretical Neuroscience: Computational and Mathematical Modeling of Neural Systems*. Cambridge, MA: MIT Press, 2001.

[8] B. Vintch, J. A. Movshon, and E. P. Simoncelli, “A Convolutional Subunit Model for Neuronal Responses in Macaque V1,” in *Neural Information Processing Systems (NIPS)*, 2015.

[9] M. Pagan, E. P. Simoncelli, and N. C. Rust, “Neural Quadratic Discriminant Analysis: Nonlinear Decoding with V1-Like Computation,” *Neural Comput.*, vol. 28, no. 11, pp. 2291–2319, 2016.

•• Develops a decoding model for higher cortical areas, and shows that the form of the computation is similar to subunit pooling encoding models typically applied in earlier visual areas.

[10] R. J. Rowekamp and T. O. Sharpee, “Cross-orientation suppression in visual area V2,” *Nat. Commun.*, vol. 8, no. 15739, 2017.

•• Highlights selective and invariant responses in V2, via a quadratic convolutional subunit pooling model that captures patterns such as textures or texture boundaries.

[11] A. Wu, I. M. Park, and J. W. Pillow, “Convolutional Spike-triggered Covariance Analysis for Neural Subunit Models,” in *Neural Information Processing Systems (NIPS)*, 2015.

[12] R. Coen-Cagli, P. Dayan, and O. Schwartz, “Cortical Surround Interactions and Perceptual Salience Via Natural Scene Statistics,” *PLoS Comput. Biol.*, vol. 8, no. 3, 2012.

[13] R. Coen-Cagli, A. Kohn, and O. Schwartz, “Flexible Gating of Contextual Modulation During Natural Vision,” *Nat. Neurosci.*, vol. 18, pp. 1648–1655, 2015.

•• Fits V1 natural image data with a normative divisive normalization model that yields selective surround suppression through gating based on inferring dependencies between center and surround responses.

[14] M. Carandini and D. J. Heeger, “Normalization as a canonical neural computation,” *Nat. Rev. Neurosci.*, vol. 13, pp. 51–62, 2012.

[15] D. J. Heeger, “Normalization of cell responses in cat striate cortex,” *Vis. Neurosci.*, vol. 9, pp. 181–197, 1992.

[16] O. Schwartz, A. Hsu, and P. Dayan, “Space and Time in Visual Context,” *Nat. Rev. Neurosci.*, vol. 8, pp. 522–535, 2007.

[17] A. Angelucci, M. Bijanzadeh, L. Nurminen, F. Federer, S. Merlin, and P. C. Bressloff, “Circuits and Mechanisms for Surround Modulation in Visual Cortex,” *Annu. Rev. Neurosci.*, vol. 40, no. 1, pp. 425–451, 2017.

[18] D. L. K. Yamins and J. J. DiCarlo, “Using goal-driven deep learning models to understand sensory cortex,” *Nat. Neurosci.*, vol. 19, no. 3, pp. 356–365, 2016.

[19] N. Kriegeskorte, “Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing,” *Annu. Rev. Vis. Sci.*, vol. 1, pp. 417–446, 2015.

[20] C. Zetzsche, B. Wegmann, and E. Barth, “Nonlinear Aspects of Primary Vision: Entropy Reduction Beyond Decorrelation,” in *Int’l Symposium, Society for Information Display*, 1993, vol. XXIV, pp. 933–936.

[21] E. P. Simoncelli, “Statistical Models for Images: Compression, Restoration and Synthesis,” in *Proc 31st Asilomar Conf on Signals, Systems and Computers*, 1997, pp. 673–678.

[22] O. Schwartz and E. P. Simoncelli, “Natural signal statistics and sensory gain control,” *Nat. Neurosci.*, vol. 4, no. 8, pp. 819–825, 2001.

[23] J. Eichhorn, F. Sinz, and M. Bethge, “Natural Image Coding in V1: How Much Use Is Orientation Selectivity?,” *PLoS Comput Biol.*, vol. 5, no. 4, 2009.

[24] J. R. Golden, K. P. Vilankar, M. C. Wu, and D. J. Field, “Conjectures regarding the nonlinear geometry of visual neurons,” *Vision Res.*, vol. 120, pp. 74–92, 2016.

[25] Y. Karklin and M. S. Lewicki, “A hierarchical Bayesian model for learning nonlinear statistical regularities in nonstationary natural signals,” *Neural Comput.*, vol. 17, pp. 397–423, 2005.

[26] Y. Karklin and M. S. Lewicki, “Emergence of complex cell properties by learning to generalize in natural scenes,” *Nature*, vol. 457, no. 1, pp. 83–87, 2009.

[27] O. Schwartz, T. J. Sejnowski, and P. Dayan, “Soft Mixer Assignment in a Hierarchichal Generative Model of Natural Scene Statistics,” *Neural Comput.*, no. 11, pp. 2680–2718, 2006.

[28] J. Lücke, R. Turner, M. Sahani, and M. Henniges, “Occlusive components analysis,” in *NIPS*, 2009, pp. 1069–1077.

[29] Z. Dai, G. Exarchakis, and J. Lücke, “What Are the Invariant Occlusive Components of Image Patches? A Probabilistic Generative Approach,” in *Neural Information Processing Systems (NIPS)*, 2013.

[30] M. Henniges, R. E. Turner, M. Sahani, J. Eggert, and J. Lücke, “Efficient Occlusive Components Analysis,” 2014.

[31] R. P. N. Rao and D. H. Ballard, “Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects,” *Nat. Neurosci.*, vol. 2, no. 1, pp. 79–87, 1999.

[32] T. Lochmann, U. A. Ernst, and S. Deneve, “Perceptual Inference Predicts Contextual Modulations of Sensory Responses,” *J. Neurosci.*, vol. 32, no. 12, pp. 4179–4195, 2012.

[33] M. W. Spratling, “Unsupervised Learning of Generative and Discriminative Weights Encoding Elementary Image Components in a Predictive Coding Model of Cortical Function,” *Neural Comput.*, vol. 24, no. 1, pp. 60–103, 2012.

[34] M. Zhu and C. J. Rozell, “Visual Nonclassical Receptive Field Effects Emerge from Sparse Coding in a Dynamical System,” *PLOS Comput. Biol.*, vol. 9, no. 8, Aug. 2013.

[35] S. Lyu and E. P. Simoncelli, “Modeling multiscale subbands of photographic images with fields of Gaussian scale mixtures,” *IEEE Trans. Patt. Anal. Mach. Intell.*, vol. 31, no. 4, pp. 693–706, 2008.

[36] E. H. Adelson, “Lightness Perception and Lightness Illusions,” in *The New Cognitive Neurosciences*, Cambridge, MA: MIT Press, 2000, pp. 339–351.

[37] M. J. Wainwright and E. P. Simoncelli, “Scale Mixtures of Gaussians and the Statistics of Natural Images,” in *Adv. Neural Information Processing Systems*, 2000, vol. 12, pp. 855–861.

[38] J. M. Beck, P. E. Latham, and A. Pouget, “Marginalization in Neural Circuits with Divisive Normalization,” *JNeurosci*, vol. 31, no. 43, pp. 15310–15319, 2011.

[39] M. Chalk, P. Masset, S. Deneve, and B. Gutkin, “Sensory noise predicts divisive reshaping of receptive fields,” *PLOS Comput. Biol.*, vol. 13, no. 6, 2017.

•• Derives normalization from the perspective of optimal cue combination in the face of noise.

[40] G.-C. Jose A, L. Mancera, and J. Portilla, “Image Restoration Using Space-Variant Gaussian Scale Mixtures in Overcomplete Pyramids,” *IEEE Trans Image Process.*, vol. 17, no. 1, pp. 27–41, 2008.

[41] M. W. Spratling, “Predictive coding as a model of response properties in cortical area V1,” *J. Neurosci.*, vol. 30, no. 9, pp. 3531–3543, 2010.

[42] Z. M. Westrick, D. J. Heeger, and M. S. Landy, “Pattern Adaptation and Normalization Reweighting,” *J. Neurosci.*, vol. 36, no. 38, pp. 9805–9816, 2016.

[43] Z. Li, “Contextual influences in V1 as a basis for pop out and asymmetry in visual search,” *Proc. Natl. Acad. Sci.*, vol. 96, no. 18, pp. 10530–10535, 1999.

[44] P. Jaini and J. Burge, “Linking normative models of natural tasks to descriptive models of neural response,” *J. Vis.*, vol. 17, no. 12, pp. 1–26, 2017.

•• Introduces a Bayesian ideal observer model that gives rise to quadratic interactions between linear subunits.

[45] I. M. Park, E. Archer, N. Priebe, and J. W. Pillow, “Spectral methods for neural characterization using generalized quadratic models,” in *NIPS*, 2013.

[46] M. Ito and H. Komatsu, “Representation of angles embedded within contour stimuli in area V2 of macaque monkeys,” *J. Neurosci.*, vol. 24, pp. 3313–3324, 2004.

[47] A. Anzai, X. Peng, and D. C. Van Essen, “Neurons in monkey visual area V2 encode combinations of orientations.,” *Nat Neurosci.*, vol. 10, no. 10, pp. 1313–1321, 2007.

[48] H. Zhou, H. S. Friedman, and R. Von der Heydt, “Coding of border ownership in monkey visual cortex,” *J. Neurosci.*, vol. 20, no. 17, pp. 6594–6611, 2000.

[49] J. R. Williford and R. von der Heydt, “Figure-Ground Organization in Visual Cortex for Natural Scenes.,” *eNeuro*, vol. 3, no. 6, 2016.

[50] L. Zhaoping, “Border ownership from intracortical interactions in visual area V2,” *Neuron*, vol. 47, no. 1, pp. 143–153, 2005.

[51] J. Freeman, C. M. Ziemba, D. J. Heeger, E. P. Simoncelli, and J. A. Movshon, “A functional and perceptual signature of the second visual area in primates,” *Nat. Neurosci.*, vol. 16, no. 7, pp. 974–981, 2013.

•• Natural texture sensitivity is a key property that emerges in V2.

[52] C. M. Ziemba, J. Freeman, A. Movshon, and E. P. Simoncelli, “Selectivity and tolerance for visual texture in macaque V2,” *Proc Natl Acad. Sci.*, vol. 113, no. 22, 2016.

[53] A. M. Hermundstad, J. J. Briguglio, M. M. Conte, J. D. Victor, V. Balasubramanian, and G. Tkačik, “Variance predicts salience in central sensory processing,” *Elife*, vol. 3, 2014.

[54] Y. Yu, A. M. Schmid, and J. D. Victor, “Visual processing of informative multipoint correlations arises primarily in V2,” *Elife*, vol. 4, 2015.

[55] A. M. Schmid, K. P. Purpura, and J. D. Victor, “Responses to Orientation Discontinuities in V1 and V2: Physiological Dissociations and Functional Implications,” *J Neurosci*, vol. 34, no. 10, pp. 3559–3578, 2014.

[56] H. Lee, C. Ekanadham, and A. Ng, “Sparse deep belief net model for visual area V2,” in *Advances in neural information processing systems*, 2008, pp. 873–880.

[57] R. Coen-Cagli and O. Schwartz, “The Impact on Mid-Level Vision of Statistically Optimal Divisive Normalization in V1,” *J. Vis.*, vol. 13, no. 8, 2013.

[58] H. Hosoya and A. Hyvärinen, “A Hierarchical Statistical Model of Natural Images. Explains Tuning Properties in V2,” *J. Neurosci.*, vol. 35, no. 29, pp. 10412–10428, 2015.

•• Presents a hierarchical unsupervised model of V2 that includes a significant dimensionality reduction after the V1 stage, followed by an overcomplete sparse coding.

[59] M. Rehn and F. T. Sommer, “A network that uses few active neurones to code visual input predicts the diverse shapes of cortical receptive fields,” *J. Comput. Neurosci.*, vol. 22, no. 2, pp. 135–146, 2007.

[60] B. A. Olshausen, “Highly overcomplete sparse coding,” in *SPIE vol 5681*, 2013.

[61] H. Shan and G. Cottrell, “Efficient Visual Coding: From Retina To V2,” A*rXiv Prepr. ArXiv: 1312.6077*, 2014.

[62] M. N. U. Laskar, L. G. Sanchez-Giraldo, and O. Schwartz, “Deep learning captures V2 selectivity for natural textures,” in *Computational and Systems Neuroscience (Cosyne), Abstract*, 2017.

[63] M. N. U. Laskar, L. G. Sanchez-Giraldo, and O. Schwartz, “Correspondence of Deep Neural Networks and the Brain for Visual Textures,” A*rXiv Prepr. ArXiv: 1806.02888*, 2018.

[64] C. Zhuang, Y. Wang, D. Yamins, and X. Hu, “Deep Learning Predicts Correlation between a Functional Signature of Higher Visual Areas and Sparse Firing of Neurons,” *Front. Comput. Neurosci.*, vol. 11, no. 100, 2017.

[65] Q. V Le *et al.*, “Building High-level Features Using Large Scale Unsupervised Learning,” in *ICML*, 2012.

[66] M. D. Zeiler, D. Krishnan, G. W. Taylor, and R. Fergus, “Deconvolutional networks,” in *Computer Vision and Pattern Recognition*, 2010, pp. 2528–2535.

[67] J. Hirayama, A. Hyvärinen, and M. Kawanabe, “SPLICE: Fully Tractable Hierarchical Extension of ICA with Pooling,” in *ICML*, 2017.

[68] J. J. Dicarlo and D. D. Cox, “Untangling invariant object recognition,” *Trends Cogn. Sci.*, vol. 11, no. 8, 2007.

[69] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, “How Does the Brain Solve Visual Object Recognition?,” *Neuron*, vol. 73, no. 3, pp. 415–434, Feb. 2012.

[70] Y. Chen, D. M. Paiton, and B. A. Olshausen, “The Sparse Manifold Transform,” NIPS, 2018.

[71] O. J. Hénaff, R. L. T. Goris, and E. P. Simoncelli, “Perceptual straightening of natural videos,” in *Cosyne abstract*, 2018.

[72] J. Balle, V. Laparra, and E. P. Simoncelli, “Density Modelling of Images using a Generalized Normalization Transformation,” in *International Conference on Learning Representations*, 2016.

•• Develops a generative model based on an invertible transformation that maps the data as close as possible to Gaussian distributed, based on a divisive normalization model with extra degrees of freedom as compared to conventional normalization models.

[73] J. Balle, V. Laparra, and E. P. Simoncelli, “End-to-end Optimized Image Compression,” in *ICLR*, 2017.

[74] M. W. Spratling, “A Hierarchical Predictive Coding Model of Object Recognition in Natural Images.,” *Cognit. Comput.*, vol. 9, no. 2, pp. 151–167, 2017.

[75] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “ImageNet Classification with Deep Convolutional Neural Networks,” in *Neural Information Processing Systems*, 2012.

[76] K. Jarrett, K. Kavukcuoglu, M. Ranzato, and Y. LeCun, “What is the best multi-stage architecture for object recognition?,” in

ICCV, 2009, pp. 2146–2153.

[77] S. Ioffe and C. Szegedy, “Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift,” in *International Conference on Machine Learning*, 2015, pp. 448–456.

[78] J. L. Ba, J. R. Kiros, and G. E. Hinton, “Layer Normalization,” A*rXiv Prepr. ArXiv: 1607.06450*, 2016.

[79] M. Ren, R. Liao, R. Urtasun, F. H. Sinz, and R. S. Zemel, “Normalizing the Normalizers: Comparing and Extending Network Normalization Schemes,” in *Proceedings of the 5th International Conference on Learning Representations*, 2017.

[80] S. Han and N. Vasconcelos, “Biologically plausible saliency mechanisms improve feedforward object recognition,” *Vision Res.*, vol. 50, no. 22, pp. 2295–2307, 2010.

[81] S. Han and N. Vasconcelos, “Object recognition with hierarchical discriminant saliency networks,” *Front. Comput. Neurosci.*, vol. 8, p. 109, 2014.

[82] C. M. Ziemba, J. Freeman, E. P. Simoncelli, and J. A. Movshon, “Contextual modulation of sensitivity to naturalistic image structure in macaque V2,” *J. Neurophysiol.*, 2018.

[83] L. G. Sanchez-Giraldo and O. Schwartz, “Integrating Flexible Normalization into Mid-Level Representations of Deep Convolutional Neural Networks,” in A*rXiv preprint ArXiv:1806.01823*, 2018.

[84] S. A. Cadena, M. A. Weis, L. A. Gatys, M. Bethge, and A. S. Ecker, “Diverse feature visualizations reveal invariances in early layers of deep neural networks,” A*rXiv Prepr. ArXiv: 1807.10589*, 2018.

[85] R. Kondor and S. Trivedi, “On the Generalization of Equivariance and Convolution in Neural Networks to the Action of Compact Groups,” in *International Conference on Machine Learning*, 2018, pp. 2747–2755.

•• Shows that convolutional structure is necessary and sufficient for equivariant representations.

[86] S. Mallat, “Understanding Deep Convolutional Networks,” A*rXiv Prepr.*, 2016.

[87] K. Lenc and A. Vedaldi, “Understanding image representations by measuring their equivariance and equivalence,” A*rXiv Prepr. ArXiv: 1411.5908*, 2014.

[88] T. Poggio and F. Anselmi, *Visual Cortex and Deep Networks: Learning Invariant Representations*. MIT Press, 2016.

[89] X. Miao and R. P. N. Rao, “Learning the Lie Groups of Visual Invariance,” *Neural Comput.*, vol. 19, no. 10, pp. 2665–2693, 2007.

[90] G. E. Hinton, A. Krizhevsky, and S. D. Wang, “Transforming Auto-Encoders,” in *International Conference on Artificial Neural Networks*, 2011, pp. 44–51.

[91] S. Sabour, N. Frosst, and G. E. Hinton, “Dynamic Routing Between Capsules,” in *Neural Information Processing Systems (NIPS)*, 2017.

[92] K. Simonyan and A. Zisserman, “Very Deep Convolutional Networks for Large-Scale Image Recognition,” in *International Conference on Learning Representations, ICLR*, 2015.

[93] C. Szegedy *et al.*, “Going deeper with convolutions,” *ArXiv:1409.4842*, 2014.

[94] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, “Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning,” *ArXiv: 1602.07261*, 2016.

[95] R. Geirhos, C. R. Medina Temme, J. Rauber, H. H. Schütt, M. Bethge, and F. A. Wichmann, “Generalisation in humans and deep neural networks,” in *NIPS*, 2018.