# Spatial and Colour Opponency in Anatomically Constrained Deep Networks

**Ethan Harris**[*†] **Daniela Mihai**[*†] **Jonathon Hare**[*†]

## Abstract

Colour vision has long fascinated scientists, who have sought to understand both the physiology of the mechanics of colour vision and the psychophysics of colour perception. We consider representations of colour in anatomically constrained convolutional deep neural networks. Following ideas from neuroscience, we classify cells in early layers into groups relating to their spectral and spatial functionality. We show the emergence of single and double opponent cells in our networks and characterise how the distribution of these cells changes under the constraint of a retinal bottleneck. Our experiments not only open up a new understanding of how deep networks process spatial and colour information, but also provide new tools to help understand the black box of deep learning. The code for all experiments is available at `https://github.com/ecs-vlc/opponency`.

## 1 Introduction

Opponent colour theory, which considers how combinations of chromatic stimuli are encoded in the visual system, proposed by Hering [11], initiated nearly a century earlier by Goethe [7], was observed and formulated at a cellular level only in the 1950s by De Valois et al. [4] and others [30, 31, 23, 3]. Combined, the theories of colour opponency, trichromacy [32, 10, 22] and feature extraction in the visual cortex [17, 12, 13, 29] constitute a deep understanding of early visual processing in nature. Furthermore, the notional elegance of these theories has served to motivate much of the progress made in computer vision, most notably including the development of Convolutional Neural Networks (CNNs) [18, 2, 19] that are now so focal in our collective interests. Despite the sheer volume of these experimental discoveries, they still represent only a sparse view of the broad spectra of the natural world. This limits our ability to consider precisely which physiological differences lead to the subtle variations in visual processing between species. For this reason, deep learning offers a unique platform through which one can study the emergence of distinct visual phenomena, across the full gamut of constraints and conditions of interest.

Lindsey et al. [20] use a multi-layer CNN to explore how the emergence of centre-surround and oriented edge receptive fields changes under biologically motivated constraints. Primarily, the authors find that the introduction of a bottleneck on the number of neurons in the second layer (the 'retina' output) of a CNN (trained to classify greyscale images) induces centre-surround receptive fields in the early layers and oriented edges in the later layers. Furthermore, the authors demonstrate that as this bottleneck is decreased, the complexity of early filters increases and they tend towards orientation selectivity. The nature of colour in CNNs has also been explored [6, 8]. Specifically, Engilberge et al. [6] find that spectral sensitivity is highest in the early layers and traded for class sensitivity in deeper layers. Gomez-Villa et al. [8] demonstrate that CNNs are susceptible to the same visual illusions as those that fool human observers. This lends weight to the notion that the specifics of colour processing result from our experience of visual stimuli in the natural world [27].

[*]Authors contributed equally

[†]Vision, Learning and Control Group, Electronics and Computer Science, University of Southampton, {ewah1g13, adm1g15, jsh2}@ecs.soton.ac.uk

1st Workshop on Shared Visual Representations in Human and Machine Intelligence (SVRHM), NeurIPS 2019.

In this paper we find evidence for spectral and spatial opponency in a deep CNN with a retinal bottleneck (following Lindsey et al. [20]) and characterise the distribution of these cells as a function of bottleneck width. In doing so, we introduce a series of experimental tools, inspired by experiments performed in neurophysiology, that can be used to shed light on the functional nature of units within deep CNNs. Furthermore, we show similarities between the specific excitatory and inhibitory responses learned by our network and those observable in nature. Across all experiments our key finding is that structure (the separation of functional properties into different layers) emerges naturally in models that feature a bottleneck. Code, implemented using PyTorch [24] and Torchbearer [9] is available at `https://github.com/ecs-vlc/opponency`.

## 2 Spatial and Colour Opponency in the Brain

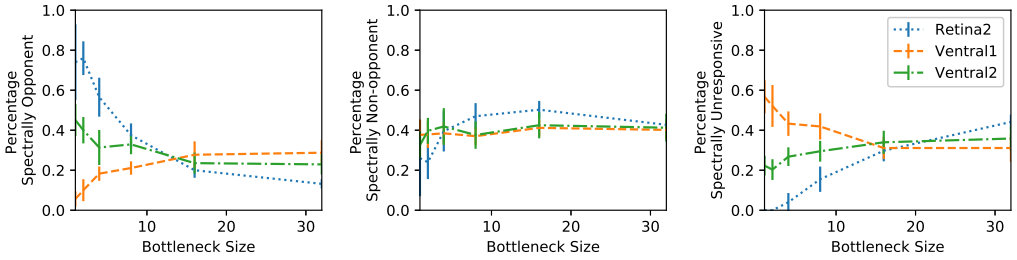
Experiments using micro-electrode recording have been used to explore how single cells respond to different stimuli. Consequently, a number of different observations and subsequent classifications of cells regarding behavioural characteristics have been made. The first key observation is the existence of two types of cell that respond to colour; spectrally opponent and spectrally non-opponent. Cells with opponent spectral sensitivity [5] are excited by particular colours,[3] and inhibited by others. For a cell to be inhibited its response must fall below its response to an empty stimulus (the 'background rate'). For excitation to occur, the response must be at some point above the background rate. Additionally, De Valois et al. [5] discovered that broadly speaking the cells could be grouped into those that were excited by red and inhibited by green (and vice-versa), and cells that were excited by blue and inhibited by yellow (and vice-versa). Cells that are spectrally non-opponent are not sensitive to specific wavelengths (or colours of equal intensity) and respond to all wavelengths in the same way . A second key observation is the existence of cells with spatial receptive fields that are opponent to each other; that is, in some spatial area, they are excited above the background rate by certain stimuli, and in other areas they are inhibited by certain stimuli [4]. Cells responsive to colour can be further grouped into 'single opponent' and 'double opponent' cells. These cells respond strongly to colour patterns but are only weakly responsive to full-field colour stimuli (e.g. solid colour across the receptive field, slow gradients or low frequency changes in colour) [28].

## 3 Experiments

In this section we detail our experimental procedures and results, characterising the emergence of spectrally, spatially and double opponent cells in deep CNNs. We focus here on the whole population of cells, for a depiction of the characterisation of a single cell see Appendix C. To preserve similarity with Lindsey et al. [20], we adopt the same deep convolutional model of the visual system. This model consists of two parts: a model of the retina, built from a pair of convolutional network layers with ReLU nonlinearities, and termed 'retina-net'; and, a ventral stream network (VVS-net) built from a stack of convolutional layers (again with ReLU) followed by a two layer MLP (with 1024 ReLU neurons in the hidden layer, and a 10-way softmax on the output layer). All convolutions are 9x9 with same padding, and each has 32 channels, with the exception of the second retinal layer whose number of channels is the retinal bottleneck. The number of convolutional layers in the ventral stream is also a parameter of the model. Our visual system model is trained with the same range of parameters (varying retinal bottlenecks and ventral system depths), with the same optimisation hyperparameters as Lindsey et al. [20], differing only in that it takes 3-channel colour inputs. As with Lindsey et al.'s work, the networks are trained to perform classification on the CIFAR-10 dataset [16]. Error bars throughout our experiments denote the standard deviation in result across all 10 models trained for each set of hyper-parameters. For further details see Appendix A.

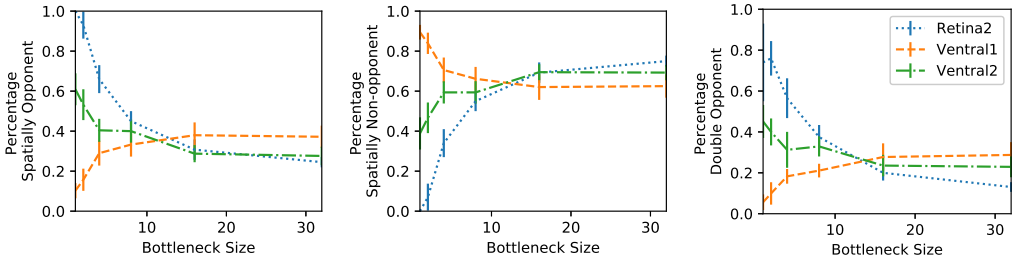
**Spectral opponency** To classify cells according to their spectral opponency, we can simulate the experimental procedure of De Valois et al. [5]. Specifically, we first present the network with uniform coloured images and measure the response of the target cell. By sampling colour patches according to hue we can show the network a range of stimuli and construct a response curve. We then classify each cell as either 'spectrally opponent' or 'spectrally non-opponent' by considering this curve relative to a background rate, defined as the response of the cell to a zero image. A spectrally non-opponent

[3] Technically, the original experiments by De Valois et al. [5] used energy-normalised single-wavelength stimuli rather than a more general notion of colour created from a mixture of wavelengths.




(a) Spectrally Opponent (b) Non-opponent (c) Unresponsive

Figure 1: Distribution of spectrally opponent, non-opponent and unresponsive cells in different layers of our model as a function of bottleneck size.

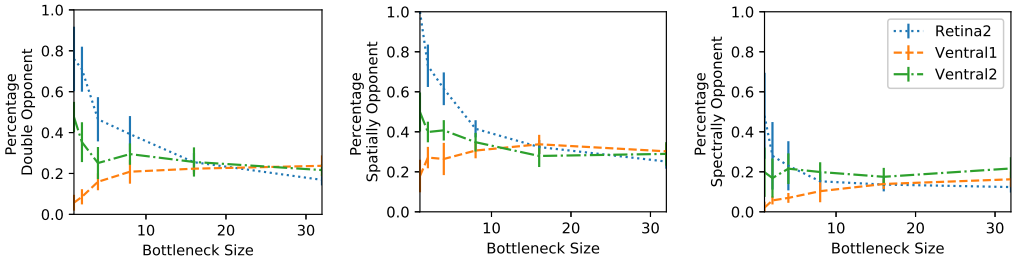



(a) Spatially Opponent (b) Non-opponent (c) Double Opponent

Figure 2: Distribution of spatially opponent, non-opponent and double opponent cells in different layers of our model as a function of bottleneck size.

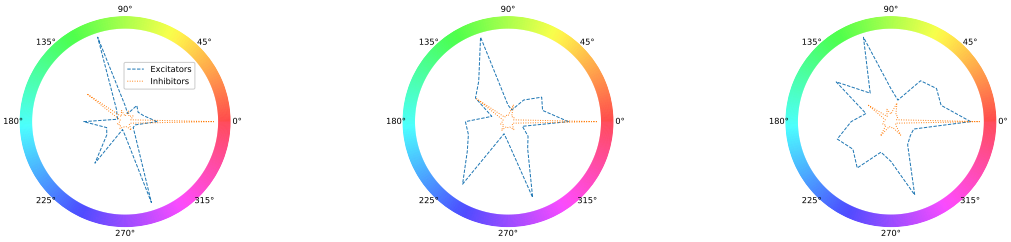
cell is one for which all responses are either above or below the baseline. A spectrally opponent cell is one for which the response is above the baseline for some colours and below the baseline for others. We further define an additional class, spectrally unresponsive, for cells which respond the same regardless of the hue of the input. Curves showing how the distributions of the spectral classes change for the second retinal and first two ventral layers as the bottleneck is increased are given in Figure 1. As the bottleneck decreases, the second retina layer exhibits a strong increase in spectral opponency, nearing $100\%$ for a bottleneck of one. Conversely, cells in the first ventral layer show a decrease in spectral opponency over the same region. Interestingly, for all but the tightest bottlenecks, up to half of the cells are spectrally non-opponent. Spectrally unresponsive cells show almost the exact opposite pattern to spectrally opponent cells.

**Spatial opponency** To explore spatial opponency, we can use a similar set-up to our experiments with spectral opponency, measuring cell response to a series of high contrast greyscale gratings produced from a sinusoidal function for a range of rotations, frequencies and phases following Johnson et al. [14] (see Appendix A.2 for an example). We can subsequently classify a cell as spatially opponent, non-opponent or unresponsive by comparing the maximum and minimum responses against the baseline in the same way as before. Further, we can characterise whether a cell is orientation tuned by isolating the grating frequency which gives the largest response for all orientations and phases, then computing the average response per orientation for that frequency across all phases. Automating the classification of cells, we can measure how spatial opponency manifests itself across the network, obtaining the results depicted in Figures 2a and 2b. We omit spatially unresponsive cells as the percentages found in each layer was always on or very near zero. For a small bottleneck, the vast majority of cells in the second retinal layer are spatially opponent. Conversely, cells in the first ventral layer are predominantly spatially non-opponent. For less constrained bottlenecks these distributions converge to be approximately equal in each of the layers. Surprisingly, and contrasting with spectral opponency, almost all cells respond to some configuration of the grating stimulus, with only a small fraction of the population being spatially unresponsive. What is again clear in these experiments is the emergent structure that arises from the introduction of a bottleneck into the model.




(a) CIELAB: Double (b) Shuffled: Spatial (c) Shuffled: Spectral

Figure 3: (a) Distribution of double opponent cells in different layers of our model trained on CIELAB images as a function of bottleneck size and the effect of shuffling the colour channels has on spatial opponency (b) and spectral opponency (c).




(a) Retina 2 (b) Ventral 1 (c) Ventral 2

Figure 4: Distribution of excitatory and inhibitory hues for spectrally opponent cells.

**Double opponency** Following Shapley and Hawken [28], we can automatically classify a cell as being double opponent if it is both spectrally and spatially opponent. Figure 2c shows the distribution of double opponent cells as a function of bottleneck size, giving a similar picture to the spectral and spatial opponency plots. This finding is in alignment with biological observations that most spectrally opponent cells are orientation selective for both achromatic and chromatic stimuli [15].

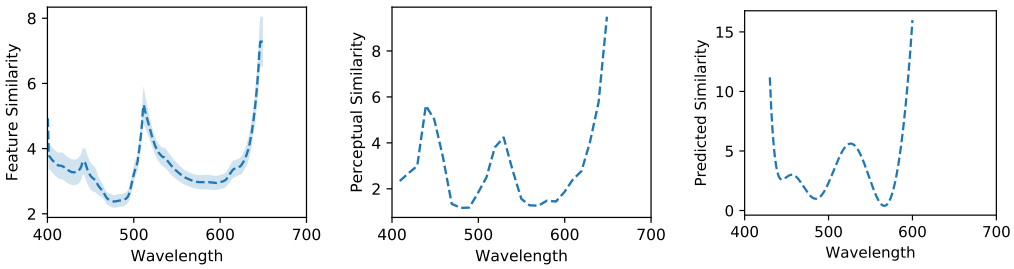
**Generalisation to CIELAB space** We performed additional experiments to validate whether double opponency is still a feature in networks trained on images in CIELAB space. Figure 3a shows the distribution of double opponent cells in this setting. As a strong validation of our findings, the distribution is nearly identical to that of networks trained on images in RGB space. This again supports the suggestion that double opponent characteristics arise from the statistics of natural images.

**Ablation: ventral depth** In order to build a greater understanding of the conditions required for opponency, we plot the distribution of spectrally and spatially opponent cells as a function of ventral depth in Figures 10 and 11 from the Appendix. With each added layer, the same pattern is found, shifted one layer to the right. This suggests that the functional organisation depends more on the distance from the output layer than from the input.

**Ablation: spectral consistency** For a final controlled demonstration of the conditions required for double opponency, we remove colour information by randomly shuffling the colour channels of inputs to the network. The resultant distribution plots show that this alteration completely removes spectral opponency (Figure 3c), whilst spatial opponency remains (Figure 3b). This again strengthens the observation that opponent characteristics arise as a result of the statistics of natural images.

## 4 Connection to Neurophysiology and Psychophysics

In this section we discuss how the learned representations from our experiments relate to the physiology and psychophysics of vision. To make comparisons between spectral processing in nature and in the trained models, we first consider the distributions of excitatory and inhibitory colours.




(a) Ours (b) Human Observer [1] (c) Natural Scenes [21]

Figure 5: (a) Similarity between successive wavelength inputs in the feature space of a retina-net with ventral depth 2 (shaded region indicates the standard error across the trained models). (b) Wavelength change needed to elicit a just-noticeable difference in hue for a Human observer from Bedford and Wyszecki [1]. (c) Predicted similarity derived from images of natural scenes from Long et al. [21].

The plots in Figure 4 show the distribution over the hue wheel of the most excitatory and most inhibitory colours in spectrally opponent cells for our models. Strikingly, the most common form of spectral opponency in the second retinal layer is predominantly red-green. The presence of cyan and magenta corresponds well with the observable colour opponents in biological vision [26], potentially in support of so-called complementary colour theory [25].

For a next point of comparison we consider the work of Bedford and Wyszecki [1] and Long et al. [21]. In Bedford and Wyszecki [1] the authors show that the change needed to elicit a just-noticeable difference in hue is a complex function of wavelength. In Long et al. [21] the authors further suggest that the reason for this non-uniform spectral sensitivity derives from the statistics of natural scenes, showing that the curve predicted from a dataset of natural images bares a strong resemblance to that obtained for a human observer. We can reproduce this experiment for our models by considering the amount of change in response to successive stimuli across the spectrum, in this case using the mean squared difference. The results for all three experiments are given in Figure 5. Perhaps surprisingly, our results show a strong degree of similarity with Human colour sensitivity and the predicted function derived from natural scenes. This in turn suggests a strong correlation between the learned spectral representation of our networks and that found in nature. Furthermore, this experiment validates the notion that classification on natural images (such as those in CIFAR-10) is a biologically valid way to model the early layers of the visual system.

Enumerating the connections regarding spatial processing is a harder task as it is more difficult to draw direct analogues to our experimental procedure. Specifically, the experiments which have inspired our approach (such as those from Johnson et al. [15] and Zhao et al. [33]) predominantly use dynamic stimuli which change with time, such as flashing on and off or sliding across the field of view. Conversely, in our experiments we use only static stimuli since the model considered exhibits no time dependence. That said, we have included some example orientation tuning curves from Zhao et al. [33] in Figure 8 from Appendix B, demonstrating that the curves found through our procedure have a similar form to those observed in the Mouse lateral geniculate nucleus.

## 5 Discussion

Our investigation has shown that a dimensionality bottleneck has the power to do more than just induce centre-surround receptive fields in the retina and oriented receptive fields in V1. We have shown that spectral, spatial and double opponent characteristics arise from this constriction and made two key observations. First, we have shown that a retinal bottleneck induces structure in the network where all cells in each layer follow a layer dependant functional archetype. Second, we have given a strong demonstration that opponent cells emerge as a result of the statistics of the input space; in this case, natural images. Our findings also have the potential to support the development of new network architectures. Specifically, if one accepts that human superiority in visual problems (such as adversarial robustness or constructing a notion of shape) can be approached through increasing similarity between deep networks and the human visual system, then we have provided a strong mandate for future research into the use of convolutional bottlenecks.

## References


[1] R.E. Bedford and Günter W. Wyszecki. Wavelength discrimination for point sources. *JOSA*, 48 (2):129–135, 1958.

[2] Léon Bottou, Corinna Cortes, John S. Denker, Harris Drucker, Isabelle Guyon, Larry D. Jackel, Yann LeCun, Urs A. Müller, Eduard Säckinger, Patrice Y. Simard, et al. Comparison of classifier methods: a case study in handwritten digit recognition. In *International conference on pattern recognition*, pages 77–77. IEEE Computer Society Press, 1994.

[3] Nigel W. Daw. Goldfish retina: organization for simultaneous color contrast. *Science*, 158 (3803):942–944, 1967.

[4] Russell L. De Valois, C.J. Smith, Stephen T. Kitai, and A.J. Karoly. Response of single cells in monkey lateral geniculate nucleus to monochromatic light. *Science*, 1958.

[5] Russell L. De Valois, Israel Abramov, and Gerald H. Jacobs. Analysis of response patterns of lgn cells∗. *J. Opt. Soc. Am.*, 56(7):966–977, Jul 1966. doi: 10.1364/JOSA.56.000966. URL `http://www.osapublishing.org/abstract.cfm?URI=josa-56-7-966`.

[6] Martin Engilberge, Edo Collins, and Sabine Süsstrunk. Color representation in deep neural networks. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 2786–2790. IEEE, 2017.

[7] Johann Wolfgang von Goethe. *Theory of colours*, volume 3. MIT Press, 1840.

[8] Alexander Gomez-Villa, Adrian Martin, Javier Vazquez-Corral, and Marcelo Bertalmio. Convolutional neural networks can be deceived by visual illusions. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[9] Ethan Harris, Matthew Painter, and Jonathon Hare. Torchbearer: A model fitting library for pytorch. *arXiv preprint arXiv:1809.03363*, 2018.

[10] Hermann von Helmholtz. LXXXI. On the theory of compound colours. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 4(28):519–534, 1852.

[11] Ewald Hering. Outlines of a theory of the light sense. 1964.

[12] David H. Hubel and Torsten N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1):106–154, 1962.

[13] David H. Hubel and Torsten N. Wiesel. *Brain and visual perception: the story of a 25-year collaboration*. Oxford University Press, 2004.

[14] Elizabeth N. Johnson, Michael J. Hawken, and Robert Shapley. The spatial transformation of color in the primary visual cortex of the macaque monkey. *Nature neuroscience*, 4(4):409, 2001.

[15] Elizabeth N. Johnson, Michael J. Hawken, and Robert Shapley. The orientation selectivity of color-responsive neurons in macaque v1. *Journal of Neuroscience*, 28(32):8096–8106, 2008.

[16] Alex Krizhevsky. Learning multiple layers of features from tiny images. Technical report, 2009.

[17] Stephen W. Kuffler. Discharge patterns and functional organization of mammalian retina. *Journal of neurophysiology*, 16(1):37–68, 1953.

[18] Yann Le Cun, Ofer Matan, Bernhard Boser, John S. Denker, Don Henderson, Richard E. Howard, Wayne Hubbard, L.D. Jackel, and Henry S. Baird. Handwritten zip code recognition with multilayer networks. In *Proc. 10th International Conference on Pattern Recognition*, volume 2, pages 35–40, 1990.

[19] Yann Le Cun, Yoshua Bengio, et al. Convolutional networks for images, speech, and time series. *The handbook of brain theory and neural networks*, 3361(10):1995, 1995.

[20] Jack Lindsey, Samuel A Ocko, Surya Ganguli, and Stephane Deny. A unified theory of early visual representations from retina to cortex through anatomically constrained deep cnns. *arXiv preprint arXiv:1901.00945*, 2019.

[21] Fuhui Long, Zhiyong Yang, and Dale Purves. Spectral statistics in natural scenes predict hue, saturation, and brightness. *Proceedings of the National Academy of Sciences*, 103(15): 6013–6018, 2006.

[22] James Clerk Maxwell. IV. On the theory of compound colours, and the relations of the colours of the spectrum. *Philosophical Transactions of the Royal Society of London*, (150):57–84, 1860.

[23] K.I. Naka and William A.H. Rushton. S-potentials from colour units in the retina of fish (cyprinidae). *The Journal of physiology*, 185(3):536–555, 1966.

[24] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. Automatic differentiation in pytorch. 2017.

[25] Ralph W. Pridmore. Theory of corresponding colors as complementary sets. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur*, 30(5):371–381, 2005.

[26] Ralph W. Pridmore. Complementary colors theory of color vision: Physiology, color mixture, color constancy and color perception. *Color Research & Application*, 36(6):394–412, 2011.

[27] Dale Purves and R. Beau Lotto. *Why we see what we do redux: A wholly empirical theory of vision.* Sinauer Associates, 2011.

[28] Robert Shapley and Michael J. Hawken. Color in the Cortex: single- and double-opponent cells. *Vision Research*, 51(7):701 – 717, 2011. ISSN 0042-6989. doi: https://doi.org/10.1016/j.visres.2011.02.012. URL `http://www.sciencedirect.com/science/article/pii/S0042698911000526`. Vision Research 50th Anniversary Issue: Part 1.

[29] John B. Troy and T. Shou. The receptive fields of cat retinal ganglion cells in physiological and pathological states: where we are after half a century of research. *Progress in retinal and eye research*, 21(3):263–302, 2002.

[30] Henry G. Wagner, E.F. MacNichol, and Myron L. Wolbarsht. Opponent color responses in retinal ganglion cells. *Science*, 131(3409):1314–1314, 1960.

[31] Torsten N. Wiesel and David H. Hubel. Spatial and chromatic interactions in the lateral geniculate body of the rhesus monkey. *Journal of neurophysiology*, 29(6):1115–1156, 1966.

[32] Thomas Young. Ii. the bakerian lecture. on the theory of light and colours. *Philosophical transactions of the Royal Society of London*, (92):12–48, 1802.

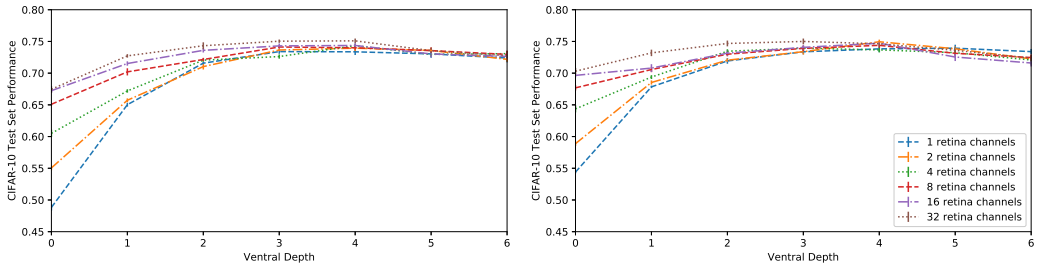
[33] Xinyu Zhao, Hui Chen, Xiaorong Liu, and Jianhua Cang. Orientation-selective responses in the mouse lateral geniculate nucleus. *Journal of Neuroscience*, 33(31):12751–12763, 2013.

# Appendices

## A Model Details

In this appendix we provide further exposition of the details of our model and experimentation process. In A.1 we detail our training process and results. In A.2 we give example gratings images used in our spatial experiments.

### A.1 Model Training



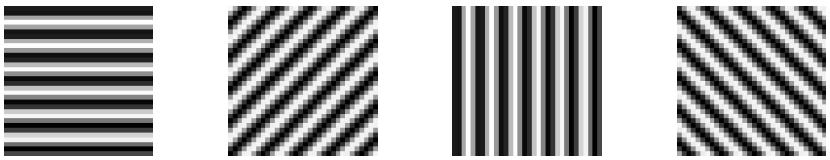

(a) Greyscale (b) Colour

Figure 6: Test accuracy for the different combinations of retinal bottleneck and ventral stream depth explored in the experiments. Data points are the average over 10 trials.

We exactly follow the model construction and training procedure defined by Lindsey et al. [20]. We note in addition that to replicate the results of the original experiments (see below). We additionally use a weight decay of $1e-6$ to provide mild regularisation of the networks weights, and data augmentation (random translations of 10% of the image width/height, and random horizontal flipping) to avoid over-fitting. Figure 6 gives the average terminal accuracy for models trained both on greyscale and colour images. The greyscale accuracy curves match those given in Lindsey et al. [20].

### A.2 Gratings

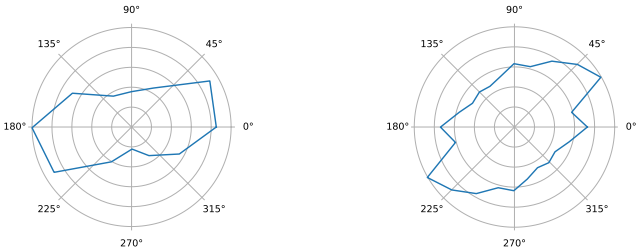
The grating patterns in Figure 7 illustrate the type of stimuli used to classify cells according to their orientation and form sensitivity. These samples have been generated with different angles ($\theta$), frequency of $\frac{8}{2\pi}$ and $0^\circ$ phase. For the experiments described in Section 4.2 of the paper, a wide range of values has been used to generate this type of stimuli.



(a) $\theta = 0^\circ$ (b) $\theta = 45^\circ$ (c) $\theta = 90^\circ$ (d) $\theta = 135^\circ$

Figure 7: Examples of grating patterns used as stimuli for the spatial opponency experiments.

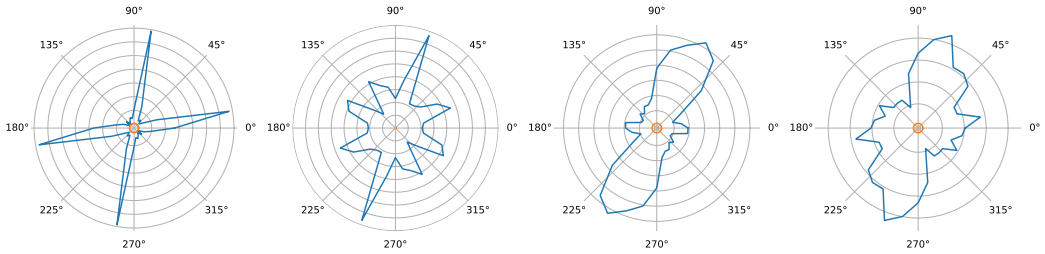
## B Mouse LGN Spatial Tuning



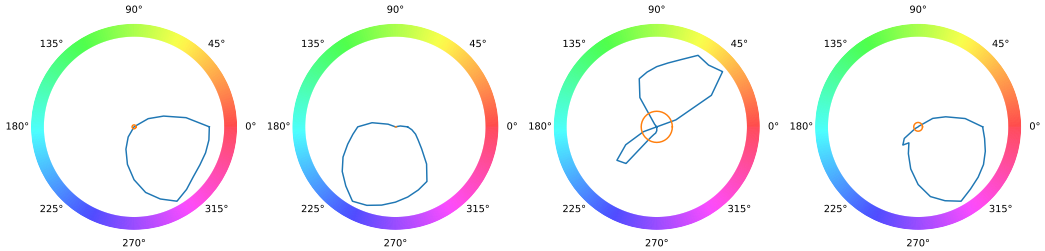

(a) Drifting Gratings (b) Flashing Bar

Figure 8: Spatial tuning curves for cells in the Mouse Lateral Geniculate Nucleus (LGN) from Zhao et al. [33].

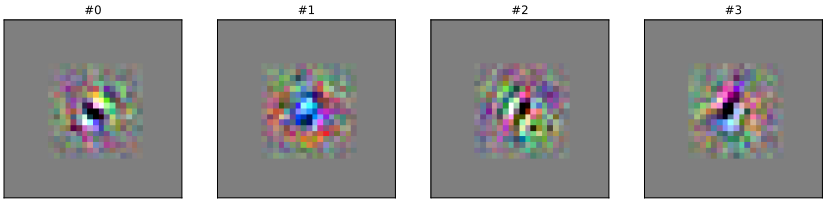
## C Characterising a Single Cell




(a) Orientation Tuning Curve




(b) Spectral Tuning Curve

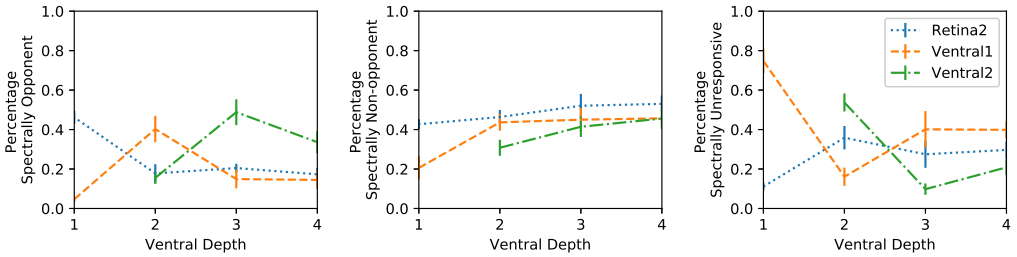



(c) First Order RF Approximation

Figure 9: Characterisation of the 4 cells in the second retinal layer of a network with bottleneck of 4 and ventral depth of 2, based on (a) orientation and form sensitivity to a range of grating patterns, (b) colour sensitivity to the colour stimuli shown on the hue wheel, and (c), the receptive field approximation via 1-step gradient ascent towards a blank image.

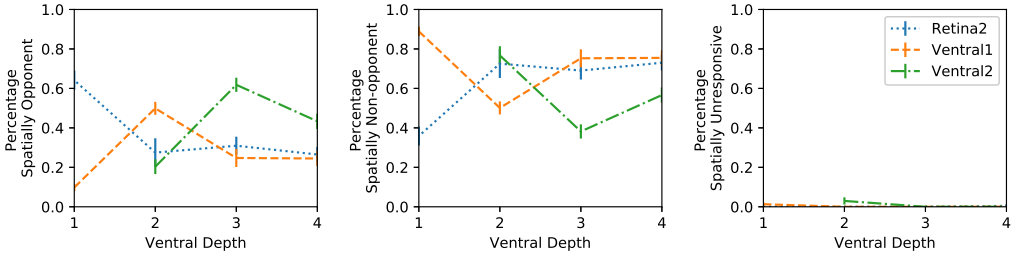
# D Ventral Depth

In this appendix we include ablation studies for the ventral depth of the model. Specifically, ventral depth corresponds to the number of layers of the convolutional network which follows the retinal bottleneck in the model from Lindsey et al. [20]. Networks were trained for a range of ventral depths following Lindsey et al. [20]. Plots show the representation of each cell type for networks with ventral depth from one to four.




(a) Spectrally Opponent (b) Non-opponent (c) Unresponsive

Figure 10: Distribution of spectrally opponent, non-opponent and unresponsive cells in different layers of our model following the definitions given by De Valois et al. [4] as a function of ventral depth.




(a) Spatially Opponent (b) Non-opponent (c) Unresponsive

Figure 11: Distribution of spatially opponent, non-opponent and unresponsive cells in different layers of our model following the definitions given by Johnson et al. [14] as a function of ventral depth.