



Doctoral thesis

On the Color Vision of Deep Neural Networks: parallels with Humans.

Author: Alban Flachot

1st supervisor: Karl R. Gegenfurtner

2nd supervisor: Roland W. Fleming

*Synopsis zur kumulativen Dissertation zur Erlangung des Doktorgrades der
Naturwissenschaften vorgelegt von Alban Flachot*

July 28, 2022

“Color is a mean of exercising direct influence upon the soul”
from *Concerning the spiritual in art*, by Wassily Kandinsky

Abstract

Seeing in color is a primordial aspect of our visual experience. Despite its importance, it is still misunderstood what exact purpose our color vision serves. A common belief is that object recognition, crucial to our survival, is a core driving force in the development of our visual system, and our color perception by extension. And indeed, color is known for improving our ability to recognize objects.

In this thesis, I explored the limits to which Deep Neural Networks, optimized for object recognition or color constancy, can explain and help us understand our color vision. Using advanced feature visualization, stimuli generation and representational analysis methods, I carefully examined the *color vision* of these trained models, also comparing their artificial responses to biological *visual systems*.

I find that both artificial and biological systems exhibit some striking differences, but these are outweighed by the sheer number of similarities. These similarities include (1) large computing power for the processing of color, (2) single and double opponent units in their early processing stages, (3) more sensitivity to variations in hue than saturation, and (4) color representations that follow similar perceptual dimensions. Despite their limitations, Deep Neural Networks can thus astonishingly explain many color properties of our visual system. This thesis hence provides evidence that our color vision is largely shaped for and motivated by a feedforward recognition of natural objects and their surface colors.

Acknowledgements

There are many people that, in one way or another, played a part in the completion of this thesis. These next few lines are but a clumsy attempt at thanking each and everyone one of them for the help and support throughout these years.

I would first like to thank my PhD supervisor, Karl Gegenfurtner, without whom this thesis would have stayed in a state of thoughts and abstractions. He gave me the means to conduct my doctoral research as best as possible, but also to develop as a better researcher. He allowed me the freedom I wished for, and offered the guidance I needed. I am forever grateful.

Great opportunities are often the results of many fortuitous encounters. The chance I have had to join Karl's group would not have been if not for J. Kevin O'Regan's recommendation. His mentoring also greatly contributed to my enthusiasm for research, and I profoundly thank him for these reasons.

Christoph Witzel also played a role in having me come to Giessen. He also provided help and support in my early years as a doctoral students, both as a friend, a colleague and a flatmate. I heartily thank him for this.

Guido Maiello is somebody who has been a dear friend in life and a big brother in science. He has helped me countless times with his advice and his time. He taught me to be bold. I thank him very much.

I would like to thank my collaborators: Edoardo Provenzi, Roland Fleming, Felix Wichmann, Arash Akbarinia, Romain Bachy, Heiko Schütt and Jelmer de Vries.

I have been blessed with so many wonderful colleagues over the years, many of whom I now happily call friends. I would like to thank my officemates Arash Akbarinia and Florian Bayer for all the help and discussions we have had on a day to day basis - both scientific and otherwise - that have made my time in the lab so fun. I would like to thank Robert Ennis and Matteo Valsecchi for

their continuous work-help and the many great times we have shared together enjoying what Giessen-city has to offer. I also thank Doris Braun for the many chats about science and art around the kitchen, and Matteo Toscani & Anna Metzger for the great dinners. Florian schiller for the (too) many cigarette breaks and Anouk Vingerling for the amazing and very needed administrative support. I would also like to thank all other fellow raccoons for the football afternoons, and all lab members for the fruitful scientific discussion, lunches and generally for making this department so great.

I would like to further thank Arash Akbarinia and Yaniv Morgerstern for their help in making this dissertation clearer.

Other than my scientific comrades, I also need to thank my friends for the joy they bring me every day and which carries me forward. It is my privilege to know and call friend each and every one of them. Those here in Germany: Laura, Thilo and Sophie for every warm moments shared over yummy dinners, fancy week-ends and sneak previews; The K1 people Mila, Leni, Ivan, Hannah, Raphi, Lily, Damian, Schmusi, Lukas, Cille, Annika, Jakob, Manu, John, Katha, Burak, Markus and Dobby for every fond memories we shared behind these 4 decrepit walls and this garden. Sara for the fun times in Stephanstrasse. The friends back home: Sel, Agapi, Badadone, Anto, Pauline, Mounem, Stéphane, Bruno and Aleysia, whom I miss so dearly everyday but still feel at home with each time I see them; Finally, Kevin and Antoine, fellow master students and also doctoral students, with whom I shared so many doubts, ideas and hopes, and a general thirst for understanding human vision: They have been a source of inspiration and enthusiasm, both for science and in life.

I would like to thank my family for the continuous support and love, in Italy and France. And my parents, most of all, for their unconditional love. I strive to make them proud.

Finally, my partner in life, Maria. Her love and care has carried me throughout these years. She has been the air I breathed. I am thankful everyday.

Contents

1	Synopsis	1
1.1	Introduction	1
1.1.1	Introduction to Deep Learning and Deep Neural Networks	3
1.1.2	Deep Neural Networks as models of the ventral stream	4
1.1.3	Biological color vision is tuned for object recognition	7
1.1.4	Related works	8
1.1.4.1	Color in DNNs for object recognition	8
1.1.4.2	Models for color constancy	10
1.2	Summary of studies	11
1.2.1	Study 1: Processing of chromatic information in a deep convolutional neural network	11
1.2.2	Study 2: Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks	14
1.2.3	Study 3: Deep Neural Models for color classification and color constancy	16
1.3	Discussion	19
1.3.1	DNNs can explain human color vision...	19
1.3.1.1	Physiological and hierarchical correspondence	19
1.3.1.2	Human-like color representations and color categories	21
1.3.1.3	A special role of Hue	21
1.3.2	... with some limitations	22
1.3.2.1	Inverse progression of hue tuning	23
1.3.2.2	DNNs hypersensitivity to color deprivation	23
1.4	Conclusion	24
	References	25

CONTENTS

2	Publications	41
2.0.1	Study 1: Processing of chromatic information in a deep convolutional neural network	42
2.0.2	Study 2: Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks	56
2.0.3	Study 3: Deep Neural Models for color classification and color constancy	69
2.0.4	Complete list of publications	94
2.0.5	Selbstständigkeitserklärung	95

1

Synopsis

1.1 Introduction

Color is a primary feature of our visual experience. In the very first pages of his *Phenomenology of Perception*, Merleau-Ponty defines sensations as follows *sensations* [1] : "to see is to have colors or lights, to hear is to have sounds". According to him, colors are to our visual experience what sounds are to our hearing: they define it. They are so salient that the revolutionary painter Kandinsky even attributes them the power of "exercising direct influence upon the soul" [2], and even scientists use them as striking examples to illustrate perceptual phenomena [3].

While it remains unclear what exact purpose our color vision serves, there is much evidence that color enhances our ability to recognize objects in a natural environment [4, 5, 6]. Object recognition, crucial for our survival, is traditionally considered a core factor in the development of our visual system [7, 8], and, by extension, our color vision. If we can quantify the extent that state-of-the-art object classification models and color classification models explain the properties of our color vision, we would be a step closer to understanding how these emerged in biological vision. This is what this thesis attempts to do, with the help of some recent and exciting advances made in artificial intelligence algorithms – Deep Neural Networks.

Deep Neural Networks, commonly called DNNs, are a family of learning algorithms that has been dominating the field of artificial intelligence for a little less than a decade now. They have notably expanded boundaries in computer vision, surpassing even humans on object and face recognition [9]. They also reached state of the art performance in gaze prediction [10], video colorization [11] and many other complex visual tasks. DNNs are very complex non linear learning algorithms, typically consisting of millions of parameters

1. SYNOPSIS

in the form of interconnections between hundred of thousands of artificial “neurons”. Given the many ways these network learn to form the same successful interconnections, it is a real challenge to understand what features each neuron has learned in order for the model to perform its task.

Developing more efficient and more accurate DNNs for visual tasks serves the field of computer vision [12] and artificial intelligence at large. But, more generally, it is also of tremendous interest for the field of vision science, as: (1) DNNs are better predictors of neural activity [13, 14, 15, 16] and neural organisation [17] than previous models. (2) They facilitate both neurophysiological [18] and psychophysical [19, 20] studies via image generation or selection. (3) Under certain conditions they even predict human behaviour [19, 21]. Understanding the computations underlying DNNs thus has the potential to bring us a step towards understanding how we perform the same tasks - by revealing similar neural features and the factors involved in their emergence. Finally, (4) DNNs are also useful for developing tools for understanding representation in cognitive systems: if we cannot find the tools and concepts to understand artificial neural networks like DNNs, how can we hope to understand biological brains which are a million times more complex, highly dynamic and hardly accessible in vivo?

This thesis is thus concerned with characterizing, describing and understanding the processing and representations of colors in DNNs trained for the recognition of natural objects and colors in complex environments. Doing so, we can test to what extent feedforward models trained on these tasks explain how we perceive colors. Here lies the principal motivation of this thesis. More precisely, we aim at solving the three following questions:

- What are the color properties of DNNs trained for object and color recognition?
- How do these properties compare to color vision in primate?
- And what can we learn from their similarities and differences?

Before diving into the three studies that constitute the core of this thesis, I will spend the next few pages developing the themes relevant for understanding their scope. I will start by giving a short introduction to deep neural networks and their impact on the field of computer vision. I will then discuss the evidence that DNNs are useful models for understanding our visual system, and the ventral stream in particular. I will follow with what evidence we have that our color vision is tuned for object recognition. I will finish the introduction with a description of some of the recent studies that are most relevant for this thesis.

1.1.1 Introduction to Deep Learning and Deep Neural Networks

For many decades, artificial neural networks sparked a relatively marginal interest in cognitive and computer science, with none but a few scientists [22, 23, 24, 25] actively working on developing such algorithms. In 2012, Alex Krizhevski and colleagues developed AlexNet [26], an artificial neural networks that surpassed every other algorithm on the benchmark ILSVRC competition [27] ¹ by an error almost half that of the runner up.

AlexNet is a combination of several innovations [28], each taking place in a short period preceding it: a fast and efficient training procedure [25, 29], sufficiently large datasets [30, 31] and the development of software libraries for high-performance operations on Graphical Processing Units (GPUs) ². Indeed, GPUs are particularly well suited for computing many simple operations simultaneously [28]. And essentially, this is what DNNs are: hundreds of thousands of very simple and interconnected operators, or "neurons", each coding for unique, sometimes simplistic features whose combined responses are able to solve very complex tasks.

Since the introduction of AlexNet, DNNs have revolutionized the field of computer vision, exhibiting supra-human performances on object [9, 32], scene and face recognition. They also showed state of the art performance in other complex visual tasks such as saliency [10], video colorization [11], face detection [33], reverse rendering [34], medical imagery analysis [35] and 3D reconstruction [36] to name a few.

Throughout this thesis, I use several successful DNN architectures and their derivatives. These architectures are AlexNet [26], the VGG nets [37], ResNet-50 [9], MobileNet [9] and some custom derivatives of ResNet [9] and classic convolutional architectures [24]. For the sake of conciseness, let us simply address here some key concepts. For a more a detailed, technical description of these models, including their similarities and differences, please refer to their respective papers or this thesis publications.

The architectures used here are classic, feedforward convolutional DNNs which non-linearly transform their input by decomposing it into increasingly complex features, as illustrated in Figure 1.1. This hierarchical encoding takes place via linear kernels, followed by a non-linear activation function, that filter their input to enhance specific features. A whole model is divided into layers, where kernels in one layer take their input from kernels in previous layers and transmit their output to the next layers, thus building ever more large and complex features as the layer depth increases [39]. Kernels in the first layers of

¹see <https://www.image-net.org/challenges/LSVRC/index.php>

²see <https://code.google.com/archive/p/cuda-convnet/>

1. SYNOPSIS

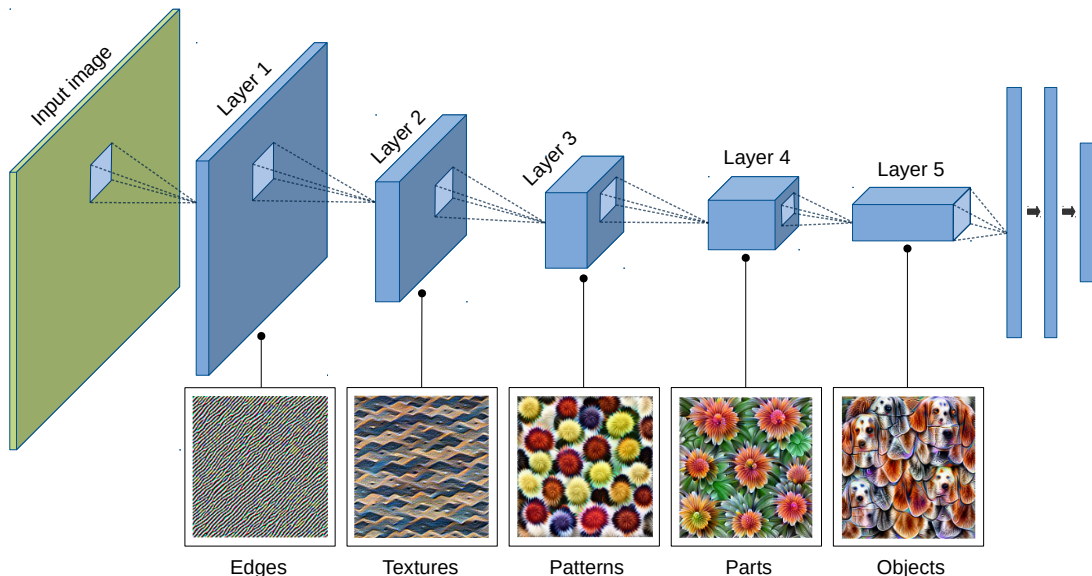


Figure 1.1: Schematic of a Deep Neural Networks and learned representations. Illustration of an AlexNet-like architecture with 5 convolutional layers and 3 fully-connected layers [26]. The visual information from the input image is processed in a feedforward way, layer per layer. As the layer number increases, so does the receptive fields of its units and the complexity of the representations learned (visualizations of representations taken from [38]).

DNNs have receptive fields typically around 1.5% to 2.5% the size of the model’s input (5 pixel-wide receptive fields for a 224 pixel-wide input), while kernels in the last layers have receptive fields of similar size as the whole input image. DNNs are Deep Learning algorithms, meaning that every kernel learns to encode features through optimization procedures, typically gradient descent [40], coupled with an optimization function, typically cross entropy for recognition tasks [26]. Thus, DNNs trained for performing object recognition are hard to interpret. To understand a model fully, one needs to discover thousands of complex features where each feature is the result of a cascade of the several non-linear operators that precede them.

1.1.2 Deep Neural Networks as models of the ventral stream

In the previous section, we have seen that DNNs revolutionized computer vision, going as far as solving complex tasks like object recognition and illumination estimation. Since biological vision resolves the same tasks, can DNNs teach us anything about the underlying computations in our own visual system?

Two different - and often opposed - conceptions of vision have concurrently driven re-

search in biological and artificial visions [41]. One, the empiricist or discriminative perspective [3, 8, 42], considers perception as a largely feedforward processing flow, driven by bottom-up mechanisms that filters and transforms the incoming retinal information into ecologically relevant constant features, such as color [3] and shape [7]. As such, vision is driven by sensory data and perception is direct. The other perspective, the generative and rationalist conception of vision [43, 44], hypothesizes that the sensory input is constantly evaluated and challenged by high-level representations that captures prior knowledge about the world, which in turn makes inferences through a top-down signal flow. As such, vision is driven by internal, complex beliefs that stabilize our sensory input.

As always, the reality is surely more refined than each perspective, and likely combines both approaches [41]. The visual systems of primates indeed exhibit the hierarchical neural organization and velocity associated with a feedforward processing of the visual input [13], but also the recurrent connections [45] and robustness [44] associated with a top-down flow of visual processing. We have yet to come up with a framework that could satisfyingly unite both approaches. Until then, we can push each approach and explore the extent in which they are sufficient to explain our visual system. With this in mind, the best current models for testing the discriminative hypothesis are DNNs [8, 41], since they indeed reach or surpass human performance on object recognition [9, 26, 46].

Interestingly, supervised DNNs trained for object recognition correlate with the ventral visual pathway - known to be involved in our object recognition [8, 47] (Cf. Figure 1.2 *A*). Notably, DNNs (1) are the best predictors for primates neural activity in the ventral pathway [13, 14, 16], meaning that they capture the computational properties of the primate brain to a larger extent than any other current models; (2) exhibit a similar organization as biology [13, 15], with a hierarchical correspondence between layers of DNNs and visual cortical areas: early cortical areas correlate with early layers of DNNs, and later cortical areas with later layers, as shown in Figure 1.2 *B*. This suggests that similarly to DNNs, the primate brain decomposes visual information into increasingly complex features; (3) have a functional neural organization alike what is found in the human visual system [48, 49] when the network is trained on the dual task of object recognition and face recognition [17]; (4) have a deep latent representation that predict human perceptual judgements [21, 50]. Together, these evidences suggests DNNs encode certain aspects of the world in the same way that we perceive them.

Supervised DNNs trained for object recognition, however, also exhibit some fundamental differences with human observers. The most well-known example is the kind of adversarial attack first proposed by Goodfellow et al. [52] where adding small amounts of noise to an

1. SYNOPSIS

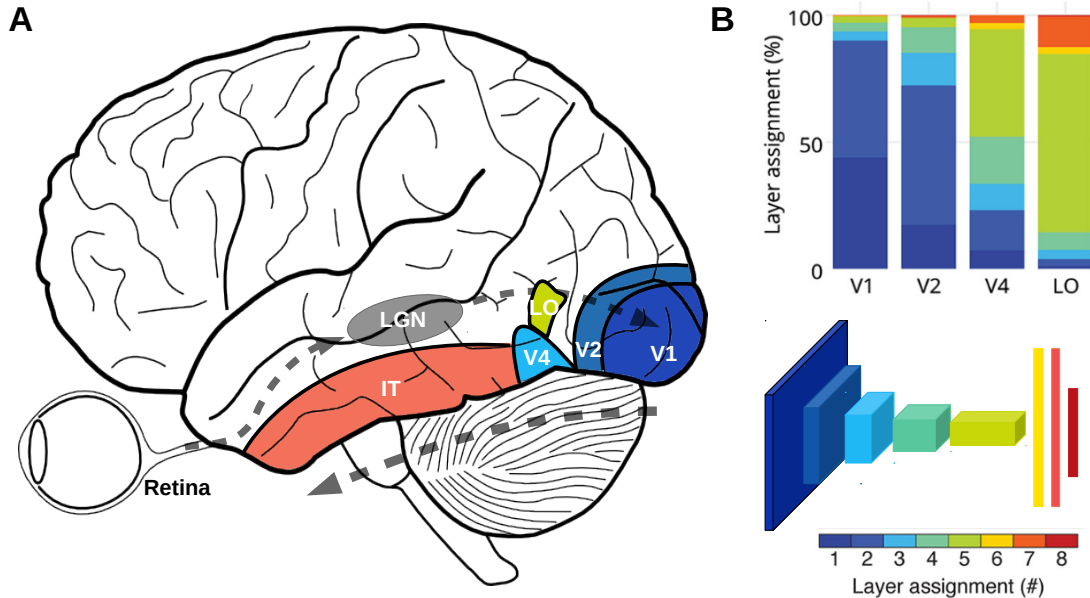


Figure 1.2: Hierarchical correspondence between layers of DNNs and visual cortical areas. *Panel A:* Schematic of the ventral stream in the human brain. It shows the approximate location of each cortical area of the ventral stream and the flow of visual information from the retina. *Panel B:* Correspondence between layers in a DNN trained for object recognition and cortical areas. The top plot is taken from [15] and shows the distribution of voxels across several cortical areas in one subject, divided according to the DNN layers that best predict their activity. Proportions are relative to the overall number of voxels significantly correlated to DNN activations (30% of total voxels). The bottom plot shows a schematic of the model’s architecture used in the study [51] and the color coding used in the layer assignment. Note that the same color coding was used in panel A: applied colors are representative of the average values found in the layer assignment [15] as well as the high correlations found between IT and top layers of DNNs [16].

image, that is imperceptible to human, lead to erroneous network responses [52]. Generally speaking, DNNs suffer from a hypersensitivity - or low robustness - to noise [53] and lack generalizability to out-of-distribution stimuli - stimuli that are outside the stimulus distribution included in the training [54]. Additionally, the error patterns of supervised DNNs often differ from human observers [21, 55], meaning that when these DNNs are wrong, they are for different reasons than us. In such cases, extreme stimuli manipulations [55] or changes in training procedure [21] are required for correcting these differences. An optimism is allowed, however, as new generations of deep neural models are ever closer to solving these issues [56, 57]. Keeping these limitations in mind, DNNs remain useful models for testing the extent to which a hierarchical, feedforward processing of the visual

information can reproduce our vision, notably the properties of cortical areas involved in object recognition.

1.1.3 Biological color vision is tuned for object recognition

DNNs seem worth pursuing as models of the visual system, and, in particular, the ventral stream. But what about color? Despite the many mysteries behind color perception [4, 5, 58], a common conception is that one crucial function of color vision is to facilitate the recognition of objects and scenes [4, 5, 58, 59].

Psychophysical studies in humans have indeed shown that color images of natural objects and scenes are recognized faster, classified more accurately and remembered better than their grayscale version [60, 61, 62, 63, 64]. This is particularly true for the so-called "color diagnostic objects" or objects which consistently appear in one specific color (e.g., lemons, tennis balls) [60, 63, 64]. For these objects, the lack of color significantly impedes our capacity to recognize them. While in our technologically advanced world, the man-made objects that we manipulate rarely come in a single color, natural objects like fruits and vegetables do [65], highlighting the importance of color for the recognition of edible objects, and our survival. Some even argue that chromatic information is more indicative of objects than luminance information, the latter rather carrying information about shadows and shadings [66, 67]. In natural settings, distinguishing between two colors can be particularly helpful for separating an object from its background e.g a fruit from its foliage [68]. Several studies have shown how color improves object segmentation in natural scenes (see [69] for a review) and how chromatic contrast drives our perception of object-contours [70]. Our visual system seems particularly responsive to chromatic contrast: color-sensitive cells with center-surround or double-opponent receptive fields appear in large proportions at the earliest stages of our visual system (e.g., LGN [71, 72] and V1 [73, 74]).

These cells are presumably also involved in color constancy: our ability to perceive surface colors consistently across various illumination conditions despite the ambiguous sensory input reaching our eyes [75, 76]. Color constancy is yet another mechanism that facilitates object recognition [5, 58]: it ensures that a same object keeps a consistent color appearance throughout the day, and under natural and artificial lights. Although color constancy has been thoroughly studied over the years (see [4, 58, 59, 75] for reviews) it has yet to be fully understood. Some theorists argue that color constancy is an ill-posed problem that cannot be perfectly solved [77, 78], and behavioral studies disagree on how color constant human observers are [58]. It also remains unclear which neural mechanisms are responsible for color constancy. Adaptation and the double opponency, Low level

1. SYNOPSIS

properties of cells in early stages of the visual system, have been shown to contribute to color constancy [76]. But higher-level and even cognitive mechanism such as memory have also been identified as being useful for color constancy. For example, humans observers have been found more color constant for familiar objects than for unknown ones [79, 80]. Thus, we are still lacking a complete neural model of color constancy, which encompasses physiological similarities to the primate’s visual system, and at the same time exhibits similar behaviour to humans on color constancy relevant tasks.

Our color vision is thus tuned for object recognition. This suggests that *object recognition* - and its derivatives like color constancy - in itself may have helped shape the way we see colors. If we want to understand how our color vision came about, it seems crucial to understand what color properties emerge in models like DNNs trained for object recognition, color recognition and color constancy, and if these networks resemble ours. Surprisingly, however, there is to my knowledge fairly little work addressing these points, although the interest seems to be growing. In the next section I give an extensive list of the few relevant studies I am aware of.

1.1.4 Related works

The studies in this thesis can be subdivided into two. The first 2 studies focus on the color properties of DNNs trained for object recognition, and how these abilities contribute to the models performance. The third study implements several architectures on the task of color recognition under varying illumination, simultaneously learning color constancy. For clarity, I divided the next short review following a similar subdivision.

1.1.4.1 Color in DNNs for object recognition

Engilberge and colleagues were first to characterize the color properties of DNNs trained for object recognition [81] using 2 popular architectures: AlexNet [26] and VGG-19[37]. They trained the models on natural images from the PASCAL dataset [82] and a subset of ImageNet[30]. They found fairly low proportions of color-sensitive units throughout the models, less than 15% in all layers. While Engilberge et al. do not show the distribution of hues for color sensitive units, they show that the color tuning of these units are more complex the deeper the layer. However, their approach has a few limitations: 1) their testing datasets only include non-color diagnostic object classes like plane and sheep, possibly explaining the low proportion of color sensitive units found; 2) they evaluated hue specificity using monochromatic images, thus removing chromatic contrast. Any chromatic

edge detector would hardly respond to these stimuli, let alone kernels with complex spatial tunings like those found in deep layers of the VGG nets [39]. In our first 2 studies, we avoided these issues by 1) using the richer ImageNet dataset [30] as a training set. While ImageNet has its limitations, it includes many natural object classes like lemon, banana, zucchini and cardoon flower as well as several breeds of dogs and birds that differ with one another in color also; 2) our test stimuli included chromatic contrasts, either through a colored disk on a grey background (Study (1)) or natural images (Study (2)).

A year later and in the same year as the first study of this thesis, Rafegas and colleagues [83] extracted some of the color properties of VGG-M net [37] trained for object recognition on the ImageNet dataset [30]. They did so using a visualization method inspired by [12, 37]. They found that an average of 32% of the network’s units were color sensitive: a much larger proportion than reported by [81]. This difference likely comes to their use of the full ImageNet for training and a representative subset for testing. They also found a prevalence of color opponency in the early layers, while kernels in higher layers tended to respond mainly to individual hues. Like in Engilberge and colleagues [81], however, their work includes a few limitations: 1) it is based on the assumption that the color properties of kernels equal the color properties of the mean image patches responsible for their max activation. As a consequence, color biases within the dataset might bias the results; 2) averaging across many images to obtain mean image patches might blur complex color and spatial tunings, particularly important in late layers [84]; 3) their study is limited to one architecture only. To address limitations (1) and (2), in our second study, we evaluated a neural unit’s maximally responsive stimulus on image patches defined in an appropriate color space, while using a segmentation algorithm to preserve the complex spatial tuning properties of deeper units. To address limitation (3) we used 3 architectures.

More recently, Harris et al. [85, 86] examined color properties in an anatomically constrained object recognition neural network [87]. The architecture mimics the biological constraints of our early visual system with the introduction of a bottleneck after the very first convolutional layers of the model [87]. Harris et al. reported simple and double opponent [72] kernels in the layers before and after the bottleneck emerged in varying proportions, depending on the size of the bottleneck. Like in [81], Harris et al. did not include chromatic contrast as their computation of a kernel’s spatial opponency was based on achromatic gratings only (Cf. Figures 6 and 7 in [85]). Despite this, their results emphasize that color sensitive edge detectors like double-opponent kernels can emerge in DNNs trained for object recognition. In the second study, we investigated this in all convolutional layers of various DNN architectures.

1. SYNOPSIS

Again more recently, Taylor Xu [84] found that color and shape are jointly represented in the higher layers of DNNs. Indeed, while colors and shape appeared separated at the first stages of the processing, the geometry of the internal representations of color varied and depended on object shape. Thus shape can influence the color tuning for kernels in the deep layers of DNNs trained for object recognition. This finding, however, should not have too many consequences for this thesis: in study (1) we used simple stimuli suited for systematically studying kernels in early layers, and, in study (2), we used stimuli with complex shapes tailored for each higher layer kernels.

In a recent work by De Vries et al [88], we show that DNNs trained for object recognition develop a categorical representation of color. We repeatedly retrained the object recognition neural networks on a color classification task across several training colors, and found systematic categorical borders. The same borders were also consistently found by a set of genetic algorithm searching for the optimal border placement. Additionally, these borders and the categories they define correlate with the basic color categories found in many populations, across cultures and languages [89]. Overall, this robust result suggests that a categorical perception of color can emerge with the development of object recognition.

1.1.4.2 Models for color constancy

There are many proposed algorithms for color constancy (e.g., from computer vision and image processing). In those fields, color constancy is typically approached by explicit estimation of the scene’s illumination [90, 91, 92, 93], followed by an image correction via the von Kries assumption [94]. This also applies to DNNs approaches [92, 93, 95, 96, 97], although all these approach outperformed previous models. Their goal is thus to correct images independently of their content, rather than model human perception. In biological vision, however, color constancy is rather tested as the ability to extract color information about the object and materials in the scene consistently across varying illuminations [58, 75, 77, 80, 98]. Following this definition, color constancy is related to object recognition, and implicitly assumes some form of color comprehension.

One reason why previous studies failed to relate DNNs trained for color constancy to object colors partially lay with the impracticability to train DNNs for color constancy on natural images, let alone object recognition. Indeed, it would require a hundreds of thousands of calibrated natural images, where the ground truth illumination would also be known. Such dataset currently does not exist. Typical natural images datasets for

color constancy are on the order of thousands of images ¹. Previous studies involving DNNs attempted at circumventing this issue through various data augmentation techniques including the application of additional color distortions and cropping [95, 96]. In study (3), instead, we rendered realistic scenes to create a large training set of naturalistic images [99].

Using this approach, we have the veridical knowledge over all relevant real-world causes such as surface colors and illuminations. Similar approaches have been used for depth and optical flow estimation tasks [100, 101, 102], and surface material inference, such as gloss [21, 103], but has to our knowledge not been applied to training DNNs for color constancy.

1.2 Summary of studies

1.2.1 Study 1: Processing of chromatic information in a deep convolutional neural network

The goal in study (1) was to compare the color properties that emerge from DNNs trained for object recognition to properties of neurons in primates. As the first work of this thesis, it thoroughly investigates one architecture - AlexNet [26]. It also mainly focuses on units in the early layers, with smaller receptive fields and fewer non-linearities than deeper units.

We trained the model on the ImageNet [30] dataset, comprised of over 1.2M hand-labeled natural images of objects. Images were divided into 1000 different categories of object, 1 category per image. To improve our statistical relevance, 34 instances of AlexNet were trained in addition to the pretrained model [104]. Randomization steps included in the training procedure [26, 29] to ensure that the training instances were slightly different from one another.

To examine color coding, we defined a color space that preserves the relationship and relative distances between colors. The cardinal dimensions of this color space, which we call RGB_{PCA} , are the principal components of the pixel distribution of the training dataset. This principal components were also found in previous work [105] for natural RGB images. The rationale behind choosing a PCA-based space is decorrelation of inputs: if DNNs trained for object recognition are similar to biological visual systems, then their internal representation should decorrelate their input along dimensions that are stastically sensible, like the principal components of the input [106, 107].

First, we looked at the color direction kernels in the first layer of all 35 instances of AlexNet are selective for. Figure 1.3 panel A shows the color direction sensitivities of the

¹see <https://colorconstancy.com/evaluation/datasets/> for a review

1. SYNOPSIS

96 kernels in the first layer of each AlexNet instance (35 total). Azimuth represents the Hue angle in RGB_{PCA} , while elevation represents the degree to which a kernel is sensitive to color: At 0° elevation, kernels are only responsive to color while at 90° they are only responsive to achromatic changes. At 45° , kernels are responsive to chromatic and achromatic information equally. In approximately equal numbers, AlexNet’s first layer kernels cluster into 2 major groups (Figure 1.3 panel A right histogram): those mostly responsive to color (*color kernels*) and those mostly sensitive to achromatic contrasts (*luminance kernels*). Additionally, kernels within the *color kernels* category tend to fall along 2 cardinal axes: the $0-180^\circ$ axis and the $90-270^\circ$ (Figure 1.3 panel A upper histogram). Thus AlexNet models seem to decorrelate their RGB input into statistically efficient dimensions in a similar fashion as the human early visual system decorrelates its LMS input [106, 107]. The distribution, however, is broad, and the 2 cardinal axes appear visible when all training instances are considered together.

Our second experiment examined the consistency underlying AlexNet’s parallel processing stream. AlexNet presents the peculiarity of having 2 parallel processing streams in its early layers, one for each of the 2 GPUs it was originally trained on, and this may have consequences on the model’s functional organisation. Alexnet’s developers [26] indeed succinctly reported that both streams each developed a specialization during training, one rather processing the chromatic information and the other the achromatic information. We asked how consistently this segregation occurred in our 35 instances and whether it correlates with the model’s performance. While the degree of segregation varied among our 35 training instances, 32 out of 35 models showed a degree of segregation significantly higher than what could be expected by chance; many instances were near perfect segregation. This consistent segregation in the early stage of visual processing is also found for the human visual system [108], suggesting that a segregation in the processing of chromatic and achromatic information facilitates the process of learning to differentiate between objects. The precise reason, however, remains unclear. Indeed, while accuracy positively correlated with the degree of segregation, it only varied by less than 0.5% across the 35 instances.

Our third experiment examined how color information is processed throughout the models. The strategy applied, inspired from physiological approaches [73, 109, 110], consisted of recording the response a single units to simple, highly controlled stimuli: colored circles on a gray background for chromatic contrast. Confronted to these stimuli the last convolutional layer of AlexNet showed a remarkable increase in responsivity to colored stimuli relative to grayscale stimuli, as is also found in the most anterior visual areas of the occipital lobe, V4 and VO [111, 112]. However, because late kernels learned complex features,

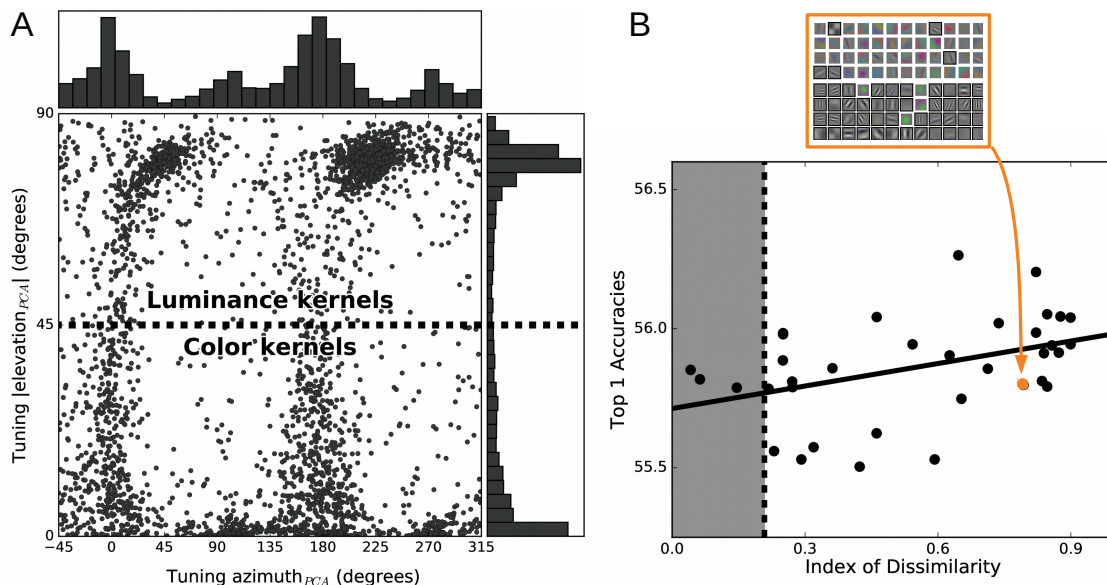


Figure 1.3: Summary of Study (1). *Panel A:* Preferred tuning directions of first layer kernels ($n=96$) in RGB_{PCA} coordinates. In the central scatter plot, individual dots are preferred elevation $_{PCA}$ angles (in absolute value) plotted against preferred azimuth $_{PCA}$ angles for each kernel in the first layer of all 35 training instances of AlexNet. Dotted lines represent the 45° threshold elevation value employed to classify kernels as either color ($\phi \leq 45^\circ$) or luminance ($\phi \geq 45^\circ$) units. The right panel shows the histogram of preferred elevation $_{PCA}$ values across all azimuths $_{PCA}$. The top panel shows the histogram of preferred azimuth $_{PCA}$ values across color kernels only. *Panel B:* Functional segregation between the processing of chromatic and achromatic information in the first layer of Alexnet. The top panel displays a visualization of the first layer kernels in one training instance. The bottom graph shows the performance of the model as a function of this functional segregation ($r = 0.41$, $p < 0.05$)

their overall response to our simple stimuli was much lower than the overall response of kernels in early layers. This suggests that stimuli with more complex spatial features are required for studying the color tuning of very late kernels accurately.

Taken together, these results show that feedforward DNNs trained on object recognition exhibit several properties similar to the primates visual system: (1) They decorrelate their input into statistically meaningful color dimensions early on; (2) They process chromatic and achromatic information separately when segregated; (3) Later stages of the visual processing exhibit a remarkable increase in color sensitivity. While the employed methods worked well for characterizing units in AlexNet’s early layers, the observed color responses in higher layers were harder to explain, presumably because of the simplicity of our stimuli. Kernels from midlayers onward are known for being responsive to complex features, where

1. SYNOPSIS

shape and colors are entangled [36]. In order to study the color properties of these kernels with higher degree of precision, in Study (2), we used more complex stimuli.

1.2.2 Study 2: Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks

While the method used in the Study (1) is well suited for examining early units of Deep Neural Networks, later units require the use of stimuli with more complex spatial features [113]. In Study (2) these limitations were solved by devising complex stimuli tailored for each one of the network’s kernels. We found each kernel’s preferred stimulus, using a method similar to [83], consisting in extracting the image patch among the 1.2M training images responsible for the model’s maximum activation. We then segmented each kernel’s preferred stimulus into meaningful segment based on their color distribution [114]. The color of each segment was then separately modified while keeping the achromatic structure of the whole stimulus the same (e.g., Figure 1.4 *left*). Finally, like in Study (1), each kernel’s response to the color changes was recorded. The main advantage of this method is that each kernel are exposed to their preferred shape, kept intact across color changes. Additionally, because the stimuli are extracted from the ImageNet training dataset and semantically meaningful to the models, we can now use them as input for the recognition task. This method thus allowed us to quantify both the color properties of the deep units and their consequences on object recognition in Deep Neural Networks.

Building up on the previous study, the pool of Deep Models included 2 additional convolutional architectures: VGG-16 and VGG-19 [37]. Very consistent results were found across all 3 models, suggesting that the color properties reported here are shared by all standard feedforward convolutional architectures, independently of their size. Additionally, many of the results confirmed the previous findings. For instance, a dichotomy was again observed in the first layers, where kernels are to a large degree either color responsive or color agnostic. A peak of color responsivity was also found in the last convolutional layers of all three models. This qualitative increase does not depend on the total number of convolutional layers the model has, be it 5 or 13 or 16.

Other than these confirmations, we were also able to describe in more detail the color properties of our 3 networks. We found a large proportion of double-opponent [72] kernels in the early layers of the models we studied - 1st layer of AlexNet and 4th layer of the VGG-nets - the determining factor appearing to be the kernel size, in each case of around 10 pixels width. Below this size - such as in the very first layers of the VGG nets - kernels are mostly similar to simple cells [72, 73] i.e selective for a single hue across their whole

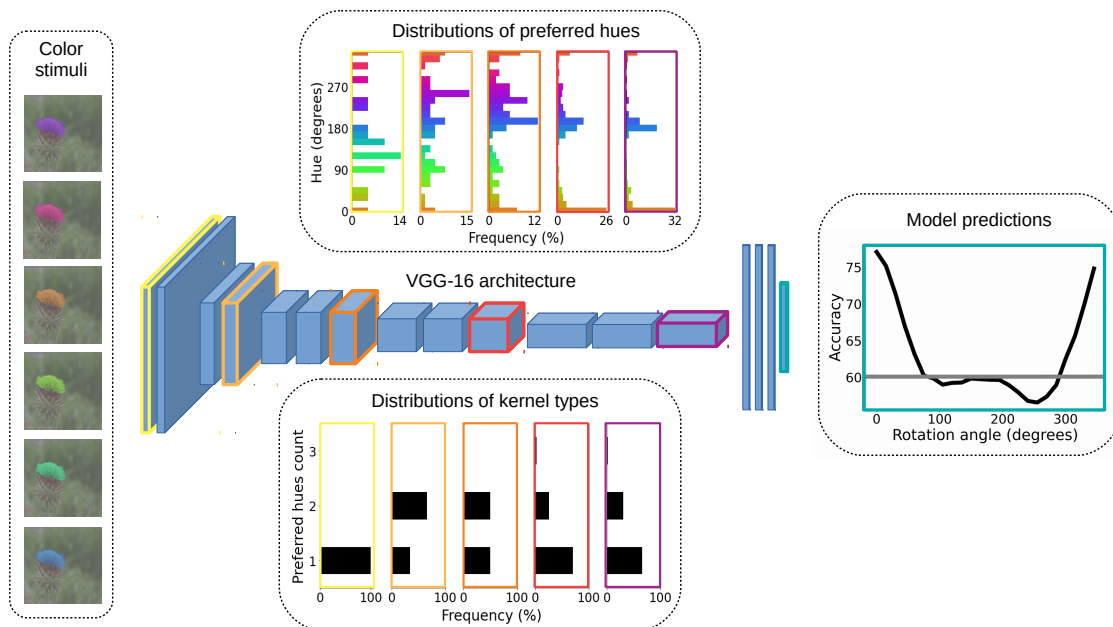


Figure 1.4: Summary of Study (2). *Center:* Schematic of the VGG-16 architecture. *Left:* Examples of colored stimuli used as network input. *Top:* Histograms of hue directions that kernels are selective for in the 1st, 4th, 7th, 10th and 13th convolutional layers (color coded). *Bottom:* Histograms of the number of hues color sensitive kernels are selective for in the same convolutional layers (as previously color coded). In the first layer, kernels are mostly single opponent i.e. selective for one hue across the whole receptive field. In the fourth layer, most color sensitive kernels are double opponent. *Right:* Accuracy of the models following hue changes. The grey line is the accuracy of the model for the same grayscale stimuli. A wrong hue can elicit a stronger drop in performance than non color, meaning that the model is more sensitive to hue than color saturation.

receptive field. Above this size - in the deep layers of all 3 networks - tunings were more varied and complex. We found kernels selective for up to 3 different hues, each in a specific spatial segment (Cf. Figure 1.4 *bottom*). Some color sensitive kernels were even found selective to 2 distinct hues within the same segment - one primary hue to which the kernel is most responsive, and one secondary. This secondary hue was the opponent color of the primary hue most of the time (around a 180° hue difference).

The distribution of hues for which kernels were most selective for also varied from layer to layer. In each individual models, early layer kernels were found selective for a wide range of hues¹. In deeper layers, however, the distributions of primary hues progressively narrows down to 1 axis: the axis blue-orange of the RGB_{PCA} coordinates, as shown in Figure

¹Despite this broad distribution, we showed in Study (1) that cardinal directions actually appear when pooling kernels of 35 training instances together.

1. SYNOPSIS

1.4 *Top*. This progression, interestingly, is almost in perfect opposition with what has been observed in the macaque’s ventral stream: while Derrington, Krauskopf and Lennie [71] found that the color responsive cells in the LGN are selective for a narrow range of hues - restricted to 2 chromatic axes, which later defined the DKL color space [115] - color responsive cells in V1 already show a higher diversity in the range of hues they are selective for [73], and cells in V2 show no preferred color axis whatsoever [109].

Finally, and perhaps more importantly, we were also able to determine what consequences some of the kernels properties have on the models performance (Cf. Figure 1.4 *right*). For instance, we found that a segment displayed with a different hue than the one it is selective for is likely to induce a lower kernel response than the same segment displayed in grey. In other words, kernels in DNNs tend to be mainly sensitive to change in hues rather than changes in chroma. This points to a special role for hue, as opposed to chroma or saturation, as has also been observed in humans [116, 117, 118]. This property propagates to the models’ recognition rate: accuracy was lowest for images with a hue around 90° apart from the original hue, lower than for the same images in grayscale. Particularly for images that color sensitive kernels are very responsive to. A similar effect was also reported in humans, both for object and scene recognition: observers took a longer time to recognize images of scenes with wrong colors than achromatic images of the same scenes [60, 119].

Overall, these results show that while supervised feedforward DNNs trained for object recognition account for a surprisingly high number of color properties of the ventral stream, they can also exhibit striking differences as illustrated by the almost opposite progression in hue tuning between both systems.

1.2.3 Study 3: Deep Neural Models for color classification and color constancy

So far, we gained insights on the color processing of state-of-the-art architectures trained on the complex task of object recognition. These networks are, however, unsuited for examining one important aspect of our color vision: color constancy, or our ability to recognize colors consistently across varying illuminations [75]. Training DNNs for color constancy requires a large dataset of calibrated images taken under varying illuminations, practically impossible for real images [95], let alone labeled for object recognition.

As an alternative, we exploited the recent advances in computer graphics [99] to generate over 450K photo-realistic images. These images were generated thanks to the real spectra

of 1600 surface reflectances [120, 121], 279 natural lights [122, 123, 124] and 2115 different realistic 3D object shapes ¹ (see Figure 1.5 *panel A* for an example).

We trained several state-of-the-art models in computer vision for color constancy: MobileNet [32], ResNet50 [9] and VGG11 [37]. We also trained two custom architectures. One, DeepCC, was a standard convolutional architecture [24]. The other, ResCC, was a Bottleneck ResNet architecture [9]. Both were similar in size and much smaller than any of the state-of-the-art models. The task of the models was to classify the images based on the object’s surface color, independently from the illumination - thus effectively learning color differences and color constancy simultaneously.

We measured the degree of color constancy achieved by the models using the Color Constancy Index (CCI) [75]. A CCI of 0 indicates no color constancy, and a CCI of 1 indicates perfect color constancy. We found that all models performed extremely well on their task, as shown in Figure 1.5 *panel B*. DeepCC, despite being the least accurate, nevertheless showed an average CCI across color classes of 0.75, around the highest values found for human observers [58], other architectures exhibiting average CCI values of 1. We also found that each one of the networks used global contextual cues surrounding the object to solve color constancy, more precisely to estimate the illumination, as indeed humans would. Like human observers, they relied on the background color to account for the illuminations [125, 126]. They also relied on scene complexity to perform color constancy, although to a lesser degree than global contextual cues, again similarly to humans [127]. Finally, the networks also exhibited lower color constancy under illuminations orthogonal to the daylight locus, which correlates with the difficulty that humans have to adapt to these unnatural illuminations [128].

Convinced that our models exhibited global behaviors qualitatively similar to humans when it comes to color constancy, we also visualized how these models perceive colors i.e. what sort of color representations they built during training. In other words, while they all differentiate between red, orange and green, do they also “perceive” orange as more similar to red than to green? This question is not trivial, as these relative distances have never been explicitly given to the models during training. And indeed, we found that not every model developed human-like representations of colors – quite far from it.

We implemented a decoding approach inspired from standard methods in brain imaging studies [129, 130] to extract matrices of distances from our models activations. These distance matrices tell us what are the relative representational distances between the surface

¹Meshes ranging from man-made objects to natural objects, issued by evermotion <https://evermotion.org/shop>

1. SYNOPSIS

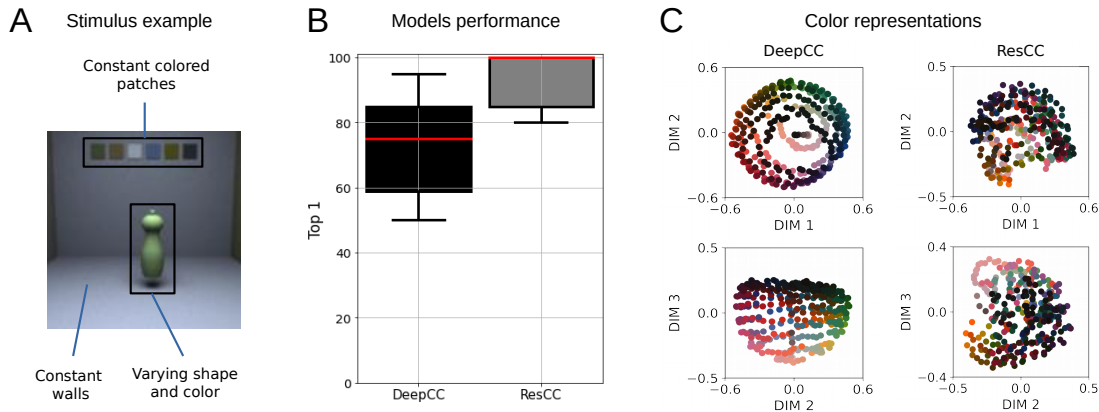


Figure 1.5: Summary of Study (3). *Panel A:* One example out of our 450K generated stimuli. The task of the models was to identify the color of the central object among 1600 different colors under varying illuminations. *Panel B:* Models performance. All models performed the task extremely well: DeepCC, the simplest convolutional architecture trained, was our poorest performer with an average Color Constancy Index of 0.75 across color classes, which is at the upper limit of color constancy in humans observers [58] *Panel C:* Internal representations of colors. Although ResCC, a custom ResNet architecture, performed better than DeepCC at the task, it represented colors in a different way than humans. Contrary to DeepCC.

colors in all different layers of our models. We can then fit the dimensions that can best explain these distances [131] and visualize the resulting representational spaces, such as in Figure 1.5 *panel C* for DeepCC and ResCC.

We found that DeepCC, our simplest convolutional architecture, was the network with the highest similarity to both perceptual spaces by a large margin. In fact, it was the only network for which the similarity grew over depth, meaning that the network progressively transformed its input into representations similar to human color perception. The other networks, like ResCC, although more accurate at object color recognition, transformed their input into color representations progressively dissimilar to human perceptual judgments.

Overall, these results show that computer graphics can be a mean to teach DNNs to distinguish between colors under varying illuminations, thus effectively also teaching them color constancy. Doing so, they consistently exhibit behaviours also previously observed in humans observers when confronted with deprivation in contextual cues across network architectures. Although the simplest architecture additionally exhibited human-like representations of colors, the other architectures did not. The latter suggests that a trade-off between an architectural simplicity and performance are necessary for developing human-

like color representations.

1.3 Discussion

In this thesis, I uncovered many of the color properties of DNNs trained for object recognition and color constancy, often comparing their emergent properties to human observers and monkeys. Here, I review the resemblance of these networks with biology, and connect my results to previous works. Doing so brings us closer to understanding the factors involved in the development of our human color vision and its properties.

1.3.1 DNNs can explain human color vision...

The DNNs studied here share many similarities with the primate visual system, either in physiological, behavioral and representational terms. Each will be discussed in detail in the next pages.

1.3.1.1 Physiological and hierarchical correspondence

The chromatic properties of DNNs trained for object recognition correlate with those in monkeys and humans. The most important similarity is perhaps that both systems devote a significant part of their resources to processing color information [4]. But the findings reported here go further and confirm a hierarchical correspondence between DNNs and biological brains ([13, 15]): At the early layers, the properties of kernels are matched to the properties of cells in the LGN and V1. At mid to late layers, kernels are matched to the properties of cells in higher areas of the visual cortex.

Indeed, in Studies (1) and (2) of this thesis as well as in [83, 85], single and double opponent kernels were found in large proportions in the early layers of DNNs trained for object recognition: single opponent kernels for receptive fields below 10 pixels in size and double opponent kernels for receptive fields around 10 pixels in size and above. These types of processing units are commonly found at the very early stages of the visual system. While single opponent cells can be found among the post-retinal ganglion cells [72] and color responsive cells in the LGN [71], double opponent cells are predominant in the subsequent stages: early areas of the visual cortex, such as V1 [72, 73]. These cells have been found to be a major early contributor of our color vision [4, 75], emphasizing the power of these models for explaining how we see colors.

We also found that DNNs tended to decorrelate their input following cardinal directions. Early layer kernels in Studies (1) and (2) were found to be either color agnostic - thus

1. SYNOPSIS

sensitive to achromatic contrasts exclusively - or strongly responsive to color only. Additionally, although individual instances showed a broad and noisy distribution of hues for which early layer kernels were most selective (Cf. Figure 1.4), we found in Study (1) that clear cardinal color directions appear when pooling several instances together (Cf. Figure 1.3). These cardinal directions match almost perfectly the principal components of the input's color distribution. Similarly, the visual system of primates is known to decorrelate its visual input into statistically optimal directions in color space [106, 107, 132, 133].

Similarities between artificial and biological systems go beyond the very early processing stages. Just like kernels in mid and late layers of our networks, cells from extra-striate cortical areas show complex color tuning and can be responsive to both achromatic and chromatic stimuli [109, 110, 134, 135, 136, 137]. Additionally, in Studies (1) and (2), as well as in Rafegas et al., [83], the highest global color sensitivities for the first and last convolutional layers. Interestingly, studies in functional imaging also show that the overall color sensitivity varies considerably between different visual cortical areas [138], seemingly without any monotonous progression. Like DNNs, both early visual areas such as the LGN and V1 and the late occipital areas, such as V4 and VO, show an overall higher color selectivity than intermediary cortical areas [112, 139]. It is precisely around these late occipital regions that neural activity is best predicted by the last convolutional layer of a DNN trained for object recognition [15, 140] (See Figure 1.2). Of course, these studies and our results do not demonstrate any strict equivalence between cortical areas and layers of AlexNet, but the tantalizing correspondence evokes that there may be a functional advantage to emphasize color information once the processing of the visual information reaches a certain complexity and scale.

Most interestingly, like the functional segregation in the LGN and to a lesser extent in V1 (see [108, 141, 142] for reviews), AlexNet shows a significant segregation in the first two layers of processing between the chromatic and achromatic information (Cf. Study 1). The exact cause for this functional segregation remains unclear: despite a significant and positive correlation of the model's performance with the degree of segregation (the more segregated the higher the accuracy) the accuracy difference remains small - below the 0.5% range. Performance may thus not be the only cause for this functional partitioning. Still, our finding suggests that the division of labor between luminance-tuned and color-tuned units is an optimal strategy in segregated neural networks such as AlexNet and the human visual system, thus a natural consequence of optimization and physical constraints. It also agrees with the hypothesis that color and luminance information serve different functions [66, 67] in both biological and artificial systems.

Overall, the neural organization of color sensitive units in DNNs, and their properties, closely match those of color sensitive cells in the primate brains. This astonishing result emphasize the usefulness of DNNs as tools for understanding and explaining the development of color vision.

1.3.1.2 Human-like color representations and color categories

DNNs can also develop color representations which correlate with human perceptual judgments. In the third study, one architecture naturally learned to differentiate between surface colors according to *lightness*, *hue* and *chroma* - the 3 color dimensions which define the most established perceptual color spaces [120, 143]. This architecture did so consistently across 10 training instances by progressively transforming its ambiguous and clustered input - an approximation of retinal transmission - into decorrelated and almost homogeneous coordinates. Not only, but the relative representational distances also closely matched the perceptual distances predicted by these spaces, indicating that the model developed human-like similarity judgments of colors.

However, no evidence of a categorical representation of colors was found in any of the models trained in study (3). Additionally, prototypical color surfaces - color surfaces representative of a color category - did not lead to higher color constancy. This refutes one theoretical idea that the perceptual singularity of prototypical colors is a consequence of their singular reflective properties, resulting in a perceptual stability under varying natural illuminations [144, 145]. This suggests that the origin of color categories does not directly come from learning to differentiate colors under a wide range of different lightning conditions. Instead, as we show in the recent work of de Vries et al. [88], color categories emerge in the latent layers of DNNs trained for object recognition. Additionally, these emergent color categories closely match the most basic color categories reported across many languages and cultures [88, 89]. How color categories help object recognition is still unclear. Nevertheless, the combination of both study (3) and de Vries et al. [88] suggests that color categories form an optimal sparse representation of color space for classifying objects visually, independently of their color stability under varying illuminations.

1.3.1.3 A special role of Hue

DNNs resemble humans in their use of color, and *hue* in particular, for recognizing objects. Geirhos et al. [146] have shown that the performance of DNNs is significantly higher when tested on color images than when tested on their grayscale version. We confirmed their

1. SYNOPSIS

observations in study (2), where the models performance decreased for grayscale images, even more so than what is reported in their study (20% relative decrease compared to 5%). This quantitative discrepancy is likely due to the higher number of color diagnostic object classes included in the ImageNet dataset [30] used here. Indeed, color was most beneficial to DNNs for recognizing images of natural objects, and generally for correctly classifying images responsible for high responses in color sensitive kernels. This benefit of color, and for recognizing natural objects in particular, is known to be important in humans as well [60, 61, 62, 63, 64]. Study (2) goes further, however, and also points towards a special role of hue for the contribution of color in DNNs. A gamut rotation of the whole image, or displaying specific segments with the wrong hue lead to an even bigger drop in the models performance than for the same images in grayscale. Again, this larger sensitivity to hue than saturation was also reported in humans. Human observers are much more sensitive to hue changes than saturation changes [117], and they have a harder time processing pictures of natural scenes displayed with a wrong hue compared to the same images in grayscale [60, 119].

Additionally, these results point towards a deep interaction between color and shape information, compatible with the evidence later provided by Taylor et al. [84] that shape and color are represented jointly in DNNs. The measured drop in performance for incorrect colors cannot be explained by a reliance of the models on chromatic contrast only for detecting object edges. If it was, the models accuracy would stay the same when colors were modified but chromatic contrasts conserved. Instead, the model's accuracy dropped sharply with a gamut rotation on the input image. This leads to the conclusion that mismatched colors act as an interference to the reliable information relayed by the stimulus luminance. Similarly, for human observers, mismatched colors interfere with the recognition of natural scenes, even more so than the lack of it [60, 119].

1.3.2 ... with some limitations

Astonishingly, many properties of our color vision emerge in DNNs, from the physiological properties of neurons in the visual cortex to the use of contextual cues and hue for solving color constancy and object recognition. These similarities have their limitations, however, which raise further questions.

1.3.2.1 Inverse progression of hue tuning

In Study (2), we found the striking difference between humans and machines that the distributions of hues for which both systems are most responsive for follow an almost exact inverse progression throughout the processing stage. Rafegas and colleagues [83] first found that kernels are on average primarily responsive to the blue and orange colors, colors most abundantly represented in the training dataset [30, 83]. In Study (2), we refined this observation showing that this mostly applies to kernels in late layers of DNNs. Kernels in early layers, however, are selective for a rather broad range of hues. The transition from a broad distribution of preferred hue to a narrow distribution representative of the input takes place progressively throughout the processing stage. In contrast, the macaque visual system exhibits the opposite transition. Cells in the LGN preferentially respond to two "cardinal directions" of color space. In the primary visual cortex, color sensitive cells are selective for a much broader range of hues [73]. Cells in V2 and later areas do not show as a whole any preference for particular hue directions, although each individual cell might be highly hue specific [109, 110, 111, 137].

One possible reason for this prominent discrepancy is the supervised nature of the training procedure and its implementation used in our networks. The gradient descent algorithm [40]] consists of an optimization procedure where the model's weights are updated in a cascade fashion, from top to bottom. Thus, kernel weights in the last layer are first updated to fit the desired output, after which weights of the penultimate layer, and so on. As a consequence, kernels of the last layers will be more specialized, more narrowly matching the dataset's color distribution than the noisier and more universal kernels of the first layer.

1.3.2.2 DNNs hypersensitivity to color deprivation

Color is beneficial to both DNNs and humans for recognizing objects, and correlates with many properties of the visual system. However, even when the experimental paradigms to test participants were designed to limit neural feedback, the accuracy loss is marginal compared to that observed for DNNs tested on grayscale images in [60, 119, 146, 146, 147]. Instead, the cost of removing color for object recognition is in humans generally rather associated with longer processing times [60, 68, 119].

Hence, despite numerous shared color properties, DNNs trained on natural colored images are more sensitive to color deprivation than humans. This hypersensitivity is likely related to the joint representation of color and shape hinted by some of our results and

1. SYNOPSIS

recently showed by Taylor et al. [84]. It suggests that except at the early stages of processing, DNNs do not evaluate both chromatic and achromatic information separately. Instead, when the color information is erroneous or missing, it interferes with the achromatic information and leads to wrong predictions. It also suggests that, in contrast, the human visual system maintain some parallelism between the flow of color and achromatic information throughout its processing of the visual information. This would agree with some recent evidence of a functional organization, including discrete cell regions sensitive to colors, in higher areas like V4 [111, 148] and IT [137, 149].

The many similarities found in the color properties of both primates and artificial brains suggest that, like DNNs, we learned to heavily rely on color for recognizing objects when color information is *available and unambiguous*. When these conditions are not met, however, then our visual system ignores the conflicting color information and relies on achromatic information - through inhibition or feedback - with the drawback of slower reaction times.

1.4 Conclusion

Despite some striking differences and limitations, DNNs surprisingly explain many of the color properties of our visual system. Like neurophysiological properties of biological vision, the networks show the emergence of simple and double opponent units in the early stage of the visual processing, and a functional separation of color and luminance. At the representational level also, with the emergence of human-like color similarities in some DNNs trained for color recognition and color constancy, and the development of color categories in DNNs trained for object recognition. At the behavioral level, finally, by their reliance on contextual cues for solving color constancy and on color information - hue especially - for solving object recognition. These numerous similarities give credit to the notion that our color vision is largely a feedforward process, motivated by and shaped for the recognition of objects and their properties.

References

- [1] MAURICE MERLEAU-PONTY. *Phenomenology of perception*. Routledge, 2013. 1
- [2] WASSILY KANDINSKY. *Concerning the spiritual in art*. Courier Corporation, 2012. 1
- [3] J KEVIN O'REGAN. *Why red doesn't sound like a bell: Understanding the feel of consciousness*. Oxford University Press, 2011. 1, 5
- [4] KARL R GEGENFURTNER AND DANIEL C KIPER. **Color vision**. *Annual review of neuroscience*, **26**(1):181–206, 2003. 1, 7, 19
- [5] ANYA HURLBERT. **Colour constancy**. *Current Biology*, **17**(21):R906–R907, 2007. 1, 7
- [6] CHRISTOPH WITZEL AND KARL R GEGENFURTNER. **Categorical sensitivity to color differences**. *Journal of vision*, **13**(7):1–1, 2013. 1
- [7] DAVID MARR. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Co, 1982. 1, 5
- [8] JAMES J DICARLO, DAVIDE ZOCCOLAN, AND NICOLE C RUST. **How does the brain solve visual object recognition?** *Neuron*, **73**(3):415–434, 2012. 1, 5
- [9] KAIMING HE, XIANGYU ZHANG, SHAOQING REN, AND JIAN SUN. **Deep residual learning for image recognition**. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 1, 3, 5, 17
- [10] MATTHIAS KÜMMERER, THOMAS SA WALLIS, AND MATTHIAS BETHGE. **DeepGaze II: Reading fixations from deep features trained on object recognition**. *arXiv preprint arXiv:1610.01563*, 2016. 1, 3

REFERENCES

- [11] BO ZHANG, MINGMING HE, JING LIAO, PEDRO V SANDER, LU YUAN, AMINE BERMAK, AND DONG CHEN. **Deep exemplar-based video colorization**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8052–8061, 2019. 1, 3
- [12] JASON YOSINSKI, JEFF CLUNE, ANH NGUYEN, THOMAS FUCHS, AND HOD LIPSON. **Understanding neural networks through deep visualization**. *arXiv preprint arXiv:1506.06579*, 2015. 2, 9
- [13] DANIEL LK YAMINS, HA HONG, CHARLES F CADIEU, ETHAN A SOLOMON, DARREN SEIBERT, AND JAMES J DICARLO. **Performance-optimized hierarchical models predict neural responses in higher visual cortex**. *Proceedings of the national academy of sciences*, **111**(23):8619–8624, 2014. 2, 5, 19
- [14] CHARLES F CADIEU, HA HONG, DANIEL LK YAMINS, NICOLAS PINTO, DIEGO ARDILA, ETHAN A SOLOMON, NAJIB J MAJAJ, AND JAMES J DICARLO. **Deep neural networks rival the representation of primate IT cortex for core visual object recognition**. *PLoS computational biology*, **10**(12):e1003963, 2014. 2, 5
- [15] UMUT GÜÇLÜ AND MARCEL AJ VAN GERVEN. **Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream**. *Journal of Neuroscience*, **35**(27):10005–10014, 2015. 2, 5, 6, 19, 20
- [16] SEYED-MAHDI KHALIGH-RAZAVI AND NIKOLAUS KRIEGESKORTE. **Deep supervised, but not unsupervised, models may explain IT cortical representation**. *PLoS computational biology*, **10**(11):e1003915, 2014. 2, 5, 6
- [17] KATHARINA DOBS, JULIO MARTINEZ, ALEXANDER JE KELL, AND NANCY KANWISHER. **Brain-like functional specialization emerges spontaneously in deep neural networks**. *bioRxiv*, 2021. 2, 5
- [18] POUYA BASHIVAN, KOHITIJ KAR, AND JAMES J DICARLO. **Neural population control via deep image synthesis**. *Science*, **364**(6439), 2019. 2
- [19] KAIYU YANG, JACQUELINE YAU, LI FEI-FEI, JIA DENG, AND OLGA RUSAKOVSKY. **A study of face obfuscation in imagenet**. *arXiv preprint arXiv:2103.06191*, 2021. 2

REFERENCES

- [20] YANIV MORGENSTERN, FRIEDER HARTMANN, FILIPP SCHMIDT, HENNING TIEDEMANN, EUGEN PROKOTT, GUIDO MAIELLO, AND ROLAND W FLEMING. **An image-computable model of human visual shape similarity.** *PLoS computational biology*, **17**(6):e1008981, 2021. 2
- [21] KATHERINE R STORRS, BARTON L ANDERSON, AND ROLAND W FLEMING. **Unsupervised learning predicts human perception and misperception of gloss.** *Nature Human Behaviour*, pages 1–16, 2021. 2, 5, 6, 11
- [22] FRANK ROSENBLATT. **The perceptron: a probabilistic model for information storage and organization in the brain.** *Psychological review*, **65**(6):386, 1958. 3
- [23] DAVID E RUMELHART, GEOFFREY E HINTON, AND RONALD J WILLIAMS. **Learning representations by back-propagating errors.** *nature*, **323**(6088):533–536, 1986. 3
- [24] YANN LECUN AND YOSHUA BENGIO. **Convolutional networks for images, speech, and time series.** *The handbook of brain theory and neural networks*, **3361**(10):1995, 1995. 3, 17
- [25] GEOFFREY E HINTON, SIMON OSINDERO, AND YEE-WHYE TEH. **A fast learning algorithm for deep belief nets.** *Neural computation*, **18**(7):1527–1554, 2006. 3
- [26] A. KRIZHEVSKY, I. SUTSKEVER, AND G. E HINTON. **Imagenet classification with deep convolutional neural networks.** In *Advances in neural information processing systems*, pages 1097–1105, 2012. 3, 4, 5, 8, 11, 12
- [27] OLGA RUSSAKOVSKY, JIA DENG, HAO SU, JONATHAN KRAUSE, SANJEEV SATHEESH, SEAN MA, ZHIHENG HUANG, ANDREJ KARPATHY, ADITYA KHOSLA, MICHAEL BERNSTEIN, ALEXANDER C. BERG, AND LI FEI-FEI. **Imagenet large scale visual recognition challenge.** *International Journal of Computer Vision*, **115**(3):211–252, 2015. 3
- [28] IAN GOODFELLOW, YOSHUA BENGIO, AND AARON COURVILLE. *Deep learning.* MIT press, 2016. 3
- [29] GEOFFREY E HINTON, NITISH SRIVASTAVA, ALEX KRIZHEVSKY, ILYA SUTSKEVER, AND RUSLAN R SALAKHUTDINOV. **Improving neural networks by preventing co-adaptation of feature detectors.** *arXiv preprint arXiv:1207.0580*, 2012. 3, 11

REFERENCES

- [30] JIA DENG, WEI DONG, RICHARD SOCHER, LI-JIA LI, KAI LI, AND LI FEI-FEI. **Imagenet: A large-scale hierarchical image database**. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009. 3, 8, 9, 11, 22, 23
- [31] ALEX KRIZHEVSKY, VINOD NAIR, AND GEOFFREY HINTON. **CIFAR-10 (canadian institute for advanced research).(2009)**. URL <http://www.cs.toronto.edu/kriz/cifar.html>, 5, 2009. 3
- [32] ANDREW G HOWARD, MENGLONG ZHU, BO CHEN, DMITRY KALENICHENKO, WEIJUN WANG, TOBIAS WEYAND, MARCO ANDREETTO, AND HARTWIG ADAM. **Mobilenets: Efficient convolutional neural networks for mobile vision applications**. *arXiv preprint arXiv:1704.04861*, 2017. 3, 17
- [33] XUDONG SUN, PENGCHENG WU, AND STEVEN CH HOI. **Face detection using deep learning: An improved faster RCNN approach**. *Neurocomputing*, 299:42–50, 2018. 3
- [34] YE YU AND WILLIAM AP SMITH. **InverseRenderNet: Learning single image inverse rendering**. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3155–3164, 2019. 3
- [35] DIMITRIOS KOLLIAS, ATHANASIOS TAGARIS, ANDREAS STAFYLOPATIS, STEFANOS KOLLIAS, AND GEORGIOS TAGARIS. **Deep neural architectures for prediction in healthcare**. *Complex & Intelligent Systems*, 4(2):119–131, 2018. 3
- [36] QIANGENG XU, WEIYUE WANG, DUYGU CEYLAN, RADOMIR MECH, AND ULRICH NEUMANN. **Disn: Deep implicit surface network for high-quality single-view 3d reconstruction**. *arXiv preprint arXiv:1905.10711*, 2019. 3, 14
- [37] KAREN SIMONYAN AND ANDREW ZISSERMAN. **Very deep convolutional networks for large-scale image recognition**. *arXiv preprint arXiv:1409.1556*, 2014. 3, 8, 9, 14, 17
- [38] CHRIS OLAH, ALEXANDER MORDVINTSEV, AND LUDWIG SCHUBERT. **Feature Visualization**. *Distill*, 2017. <https://distill.pub/2017/feature-visualization>. 4
- [39] MATTHEW D ZEILER AND ROB FERGUS. **Visualizing and understanding convolutional networks**. In *European conference on computer vision*, pages 818–833. Springer, 2014. 3, 9

-
- [40] YANN LECUN, LÉON BOTTOU, YOSHUA BENGIO, AND PATRICK HAFFNER. **Gradient-based learning applied to document recognition.** *Proceedings of the IEEE*, **86**(11):2278–2324, 1998. 4, 23
- [41] JAMES J DICARLO, RALF HAEFNER, LEYLA ISIK, TALIA KONKLE, NIKOLAUS KRIEGESKORTE, BENJAMIN PETERS, NICOLE RUST, KIM STACHENFELD, JOSHUA B TENENBAUM, DORIS TSAO, ET AL. **How does the brain combine generative models and direct discriminative computations in high-level vision?** 2021. 5
- [42] JAMES J GIBSON. *The ecological approach to visual perception: classic edition.* Psychology Press, 2014. 5
- [43] HERMANN VON HELMHOLTZ. *Handbuch der physiologischen Optik: mit 213 in den Text eingedruckten Holzschnitten und 11 Tafeln*, **9**. Voss, 1867. 5
- [44] ALAN YUILLE AND DANIEL KERSTEN. **Vision as Bayesian inference: analysis by synthesis?** *Trends in cognitive sciences*, **10**(7):301–308, 2006. 5
- [45] DAVID J HEEGER. **Theory of cortical function.** *Proceedings of the National Academy of Sciences*, **114**(8):1773–1782, 2017. 5
- [46] JIE HU, LI SHEN, AND GANG SUN. **Squeeze-and-excitation networks.** In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7132–7141, 2018. 5
- [47] JAMES J DICARLO AND DAVID D COX. **Untangling invariant object recognition.** *Trends in cognitive sciences*, **11**(8):333–341, 2007. 5
- [48] NANCY KANWISHER, JOSH MCDERMOTT, AND MARVIN M CHUN. **The fusiform face area: a module in human extrastriate cortex specialized for face perception.** *Journal of neuroscience*, **17**(11):4302–4311, 1997. 5
- [49] NANCY KANWISHER. **Functional specificity in the human brain: a window into the functional architecture of the mind.** *Proceedings of the National Academy of Sciences*, **107**(25):11163–11170, 2010. 5
- [50] JOSHUA C PETERSON, JOSHUA T ABBOTT, AND THOMAS L GRIFFITHS. **Evaluating (and improving) the correspondence between deep neural networks and human representations.** *Cognitive science*, **42**(8):2648–2669, 2018. 5

REFERENCES

- [51] K. CHATFIELD, K. SIMONYAN, A. VEDALDI, AND A. ZISSERMAN. **Return of the devil in the details: Delving deep into convolutional nets.** *arXiv preprint arXiv:1405.3531*, 2014. 6
- [52] IAN J GOODFELLOW, JONATHON SHLENS, AND CHRISTIAN SZEGEDY. **Explaining and harnessing adversarial examples.** *arXiv preprint arXiv:1412.6572*, 2014. 5, 6
- [53] ROBERT GEIRHOS, CARLOS R MEDINA TEMME, JONAS RAUBER, HEIKO H SCHÜTT, MATTHIAS BETHGE, AND FELIX A WICHMANN. **Generalisation in humans and deep neural networks.** *arXiv preprint arXiv:1808.08750*, 2018. 6
- [54] ROBERT GEIRHOS, JÖRN-HENRIK JACOBSEN, CLAUDIO MICHAELIS, RICHARD ZEMEL, WIELAND BRENDEL, MATTHIAS BETHGE, AND FELIX A. WICHMANN. **Shortcut learning in deep neural networks.** *Nature Machine Intelligence*, 2(11):665–673, 2020. 6
- [55] ROBERT GEIRHOS, PATRICIA RUBISCH, CLAUDIO MICHAELIS, MATTHIAS BETHGE, FELIX A WICHMANN, AND WIELAND BRENDEL. **ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness.** *arXiv preprint arXiv:1811.12231*, 2018. 6
- [56] EVGENIA RUSAK, LUKAS SCHOTT, ROLAND S ZIMMERMANN, JULIAN BITTERWOLF, OLIVER BRINGMANN, MATTHIAS BETHGE, AND WIELAND BRENDEL. **A simple way to make neural networks robust against diverse image corruptions.** In *European Conference on Computer Vision*, pages 53–69. Springer, 2020. 6
- [57] ROBERT GEIRHOS, KANTHARAJU NARAYANAPPA, BENJAMIN MITZKUS, TIZIAN THIERINGER, MATTHIAS BETHGE, FELIX A WICHMANN, AND WIELAND BRENDEL. **Partial success in closing the gap between human and machine vision.** *arXiv preprint arXiv:2106.07411*, 2021. 6
- [58] CHRISTOPH WITZEL AND KARL R GEGENFURTNER. **Color perception: Objects, constancy, and categories.** *Annual Review of Vision Science*, 4:475–499, 2018. 7, 10, 17, 18
- [59] ANYA HURLBERT. **Challenges to color constancy in a contemporary light.** *Current Opinion in Behavioral Sciences*, 30:186–193, 2019. 7

REFERENCES

- [60] JAMES W TANAKA AND LYNN M PRESNELL. **Color diagnosticity in object recognition.** *Perception & Psychophysics*, **61**(6):1140–1153, 1999. 7, 16, 22, 23
- [61] JAMES TANAKA, DANIEL WEISKOPF, AND PEPPER WILLIAMS. **The role of color in high-level vision.** *Trends in cognitive sciences*, **5**(5):211–215, 2001. 7, 22
- [62] FELIX A WICHMANN, LINDSAY T SHARPE, AND KARL R GEGENFURTNER. **The contributions of color to recognition memory for natural scenes.** *Journal of Experimental Psychology: Learning, Memory, and Cognition*, **28**(3):509, 2002. 7, 22
- [63] G KEITH HUMPHREY, MELVYN A GOODALE, LORNA S JAKOBSON, AND PHILIP SERVOS. **The role of surface information in object recognition: Studies of a visual form agnostic and normal subjects.** *Perception*, **23**(12):1457–1481, 1994. 7, 22
- [64] LEE H WURM, GORDON E LEGGE, LISA M ISENBERG, AND ANDREW LUEBKER. **Color improves object recognition in normal and low vision.** *Journal of Experimental Psychology: Human perception and performance*, **19**(4):899, 1993. 7, 22
- [65] ROBERT ENNIS, M TOSCANI, F. SCHILLER, AND K. R GEGENFURTNER. **Hyper-spectral database of fruits and vegetables.** 2018. 7
- [66] FREDERICK AA KINGDOM, CATHERINE BEAUCE, AND LYNDISAY HUNTER. **Colour vision brings clarity to shadows.** *Perception*, **33**(8):907–914, 2004. 7, 20
- [67] ADRIANA OLMOS AND FREDERICK AA KINGDOM. **A biologically inspired algorithm for the recovery of shading and reflectance images.** *Perception*, **33**(12):1463–1473, 2004. 7, 20
- [68] KARL R GEGENFURTNER AND JOCHEM RIEGER. **Sensory and cognitive contributions of color to the recognition of natural scenes.** *Current Biology*, **10**(13):805–808, 2000. 7, 23
- [69] FARID GARCIA-LAMONT, JAIR CERVANTES, ASDRÚBAL LÓPEZ, AND LISBETH RODRIGUEZ. **Segmentation of images by color features: A survey.** *Neurocomputing*, **292**:1–27, 2018. 7
- [70] THORSTEN HANSEN AND KARL R GEGENFURTNER. **Color contributes to object-contour perception in natural scenes.** *Journal of Vision*, **17**(3):14–14, 2017. 7

REFERENCES

- [71] ANDREW M DERRINGTON, JOHN KRAUSKOPF, AND PETER LENNIE. **Chromatic mechanisms in lateral geniculate nucleus of macaque.** *The Journal of physiology*, **357**(1):241–265, 1984. 7, 16, 19
- [72] ROBERT SHAPLEY AND MICHAEL J HAWKEN. **Color in the cortex: single-and double-opponent cells.** *Vision research*, **51**(7):701–717, 2011. 7, 9, 14, 19
- [73] PETER LENNIE, JOHN KRAUSKOPF, AND GARY SCLAR. **Chromatic mechanisms in striate cortex of macaque.** *Journal of Neuroscience*, **10**(2):649–669, 1990. 7, 12, 14, 16, 19, 23
- [74] BEVIL R CONWAY. **Spatial structure of cone inputs to color cells in alert macaque primary visual cortex (V-1).** *Journal of Neuroscience*, **21**(8):2768–2783, 2001. 7
- [75] D.H. FOSTER. **Color constancy.** *Vision Research*, **51**:674–700, 2011. 7, 10, 16, 17, 19
- [76] SHAO-BING GAO, KAI-FU YANG, CHAO-YI LI, AND YONG-JIE LI. **Color constancy using double-opponency.** *IEEE transactions on pattern analysis and machine intelligence*, **37**(10):1973–1985, 2015. 7, 8
- [77] LAURENCE T MALONEY AND BRIAN A WANDELL. **Color constancy: a method for recovering surface spectral reflectance.** *JOSA A*, **3**(1):29–33, 1986. 7, 10
- [78] AD LOGVINENKO, B. FUNT, H. MIRZAEI, AND R. TOKUNAGA. **Rethinking colour constancy.** *PloS one*, **10**(9):e0135029, 2015. 7
- [79] JEROEN JM GRANZIER AND KARL R GEGENFURTNER. **Effects of memory colour on colour constancy for unknown coloured objects.** *i-Perception*, **3**(3):190–215, 2012. 8
- [80] MARIA OLKKONEN, THORSTEN HANSEN, AND KARL R GEGENFURTNER. **Color appearance of familiar objects: Effects of object shape, texture, and illumination changes.** *Journal of vision*, **8**(5):13–17, 2008. 8, 10
- [81] M. ENGIJBERGE, E. COLLINS, AND S SÜSTRUNK. **Color representation in deep neural networks.** In *International Conference on Image Processing (ICIP), 2017*. IEEE, 2017. 8, 9

REFERENCES

- [82] M. EVERINGHAM, L. VAN GOOL, C. K. I. WILLIAMS, J. WINN, AND A. ZISSERMAN. **The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results.** <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>. 8
- [83] IVET RAFEGAS AND MARIA VANRELL. **Color encoding in biologically-inspired convolutional neural networks.** *Vision research*, 2018. 9, 14, 19, 20, 23
- [84] JOHNMARK TAYLOR AND YAODA XU. **Joint representation of color and form in convolutional neural networks: A stimulus-rich network perspective.** *Plos one*, **16**(6):e0253442, 2021. 9, 10, 22, 24
- [85] ETHAN HARRIS, DANIELA MIHAI, AND JONATHON HARE. **Spatial and colour opponency in anatomically constrained deep networks.** *arXiv preprint arXiv:1910.11086*, 2019. 9, 19
- [86] ETHAN WILLIAM ALBERT HARRIS, ANDREEA DANIELA MIHAI, AND JONATHON HARE. **Anatomically constrained ResNets exhibit opponent receptive fields; so what?** 2020. 9
- [87] 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. 9
- [88] JELMER P. DE VRIES, ARASH AKBARINIA, ALBAN FLACHOT, AND KARL R. GEGENFURTNER. **Emergent Color Categorization in a Neural Network trained for Object Recognition.** *bioRxiv*, 2021. 10, 21
- [89] B. BERLIN AND P. KAY. *Basic color terms: Their universality and evolution.* Univ of California Press, 1991. 10, 21
- [90] E. H LAND. **The retinex.** *American Scientist*, **52**(2):247–264, 1964. 10
- [91] ARASH AKBARINIA AND C ALEJANDRO PARRAGA. **Colour constancy beyond the classical receptive field.** *IEEE transactions on pattern analysis and machine intelligence*, **40**(9):2081–2094, 2017. 10
- [92] MAHMOUD AFIFI AND MICHAEL S BROWN. **Sensor-independent illumination estimation for DNN models.** *arXiv preprint arXiv:1912.06888*, 2019. 10

REFERENCES

- [93] YUANMING HU, BAORYUAN WANG, AND STEPHEN LIN. **Fc4: Fully convolutional color constancy with confidence-weighted pooling**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4085–4094, 2017. 10
- [94] J. VON KRIES. **Chromatic adaptation**. *Festschrift der Albrecht-Ludwigs-Universität*, **135**:145–158, 1902. 10
- [95] ZHONGYU LOU, THEO GEVERS, NINGHANG HU, MARCEL P LUCASSEN, ET AL. **Color Constancy by Deep Learning**. In *Proceedings of BMVC*, pages 76–1, 2015. 10, 11, 16
- [96] SIMONE BIANCO, CLAUDIO CUSANO, AND RAIMONDO SCHETTINI. **Color constancy using CNNs**. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 81–89, 2015. 10, 11
- [97] WU SHI, CHEN CHANGE LOY, AND XIAOOU TANG. **Deep specialized network for illuminant estimation**. In *European conference on computer vision*, pages 371–387. Springer, 2016. 10
- [98] DAVID WEISS, CHRISTOPH WITZEL, AND KARL GEGENFURTNER. **Determinants of colour constancy and the blue bias**. *i-Perception*, **8**(6):2041669517739635, 2017. 10
- [99] WENZEL JAKOB. **Mitsuba renderer**, 2010. 11, 16
- [100] D. J. BUTLER, J. WULFF, G. B. STANLEY, AND M. J. BLACK. **A naturalistic open source movie for optical flow evaluation**. In A. FITZGIBBON ET AL. (EDS.), editor, *European Conf. on Computer Vision (ECCV)*, Part IV, LNCS 7577, pages 611–625. Springer-Verlag, October 2012. 11
- [101] A. DOSOVITSKIY, P. FISCHER, E. ILG, P. HÄUSSER, C. HAZIRBAŞ, V. GOLKOV, P. V.D. SMAGT, D. CREMERS, AND T. BROX. **FlowNet: Learning Optical Flow with Convolutional Networks**. In *IEEE International Conference on Computer Vision (ICCV)*, 2015. 11
- [102] E. ILG, T. SAIKIA, M. KEUPER, AND T. BROX. **Occlusions, Motion and Depth Boundaries with a Generic Network for Disparity, Optical Flow or Scene Flow Estimation**. In *European Conference on Computer Vision (ECCV)*, 2018. 11

REFERENCES

- [103] KE PROKOTT, H TAMURA, AND RW FLEMING. **Gloss perception: Searching for a deep neural network that behaves like humans.** *Journal of Vision*, in press. 11
- [104] YANGQING JIA, EVAN SHELHAMER, JEFF DONAHUE, SERGEY KARAYEV, JONATHAN LONG, ROSS GIRSHICK, SERGIO GUADARRAMA, AND TREVOR DARRELL. **Caffe: Convolutional Architecture for Fast Feature Embedding.** *arXiv preprint arXiv:1408.5093*, 2014. 11
- [105] YU-ICHI OHTA, TAKEO KANADE, AND TOSHIYUKI SAKAI. **Color information for region segmentation.** *Computer graphics and image processing*, **13**(3):222–241, 1980. 11
- [106] GERSHON BUCHSBAUM AND A GOTTSCHALK. **Trichromacy, opponent colours coding and optimum colour information transmission in the retina.** *Proceedings of the Royal Society of London B: Biological Sciences*, **220**(1218):89–113, 1983. 11, 12, 20
- [107] DANIEL L RUDERMAN, THOMAS W CRONIN, AND CHUAN-CHIN CHIAO. **Statistics of cone responses to natural images: Implications for visual coding.** *JOSA A*, **15**(8):2036–2045, 1998. 11, 12, 20
- [108] EDWARD M CALLAWAY. **Structure and function of parallel pathways in the primate early visual system.** *The Journal of physiology*, **566**(1):13–19, 2005. 12, 20
- [109] KARL R GEGENFURTNER, DANIEL C KIPER, JACK MH BEUSMANS, MATTEO CARANDINI, QASIM ZAIDI, AND J ANTHONY MOVSHON. **Chromatic properties of neurons in macaque MT.** *Visual neuroscience*, **11**(3):455–466, 1994. 12, 16, 20, 23
- [110] KARL R GEGENFURTNER, DANIEL C KIPER, AND SUZANNE B FENSTEMAKER. **Processing of color, form, and motion in macaque area V2.** *Visual neuroscience*, **13**(01):161–172, 1996. 12, 20, 23
- [111] BEVIL R CONWAY, SEBASTIAN MOELLER, AND DORIS Y TSAO. **Specialized color modules in macaque extrastriate cortex.** *Neuron*, **56**(3):560–573, 2007. 12, 23, 24

REFERENCES

- [112] KATHY T MULLEN, SERGE O DUMOULIN, KATIE L MCMAHON, GREIG I DE ZUBICARAY, AND ROBERT F HESS. **Selectivity of human retinotopic visual cortex to S-cone-opponent, L/M-cone-opponent and achromatic stimulation.** *European Journal of Neuroscience*, **25**(2):491–502, 2007. 12, 20
- [113] QUANSHI ZHANG AND SONG-CHUN ZHU. **Visual interpretability for deep learning: a survey.** *arXiv preprint arXiv:1802.00614*, 2018. 14
- [114] SONG YUHENG AND YAN HAO. **Image segmentation algorithms overview.** *arXiv preprint arXiv:1707.02051*, 2017. 14
- [115] THORSTEN HANSEN AND KARL R GEGENFURTNER. **Higher order color mechanisms: Evidence from noise-masking experiments in cone contrast space.** *Journal of vision*, **13**(1):26–26, 2013. 16
- [116] DEANE B JUDD. **Ideal color space.** *Color Eng*, **8**(2):36–52, 1970. 16
- [117] MV DANILOVA AND JD MOLLON. **Superior discrimination for hue than for saturation and an explanation in terms of correlated neural noise.** *Proceedings of the Royal Society B: Biological Sciences*, **283**(1831):20160164, 2016. 16, 22
- [118] J KRAUSKOPF AND K. R. GEGENFURTNER. **Color discrimination and adaptation.** *Vision research*, **32**(11):2165–2175, 1992. 16
- [119] AUDE OLIVA AND PHILIPPE G SCHYNS. **Diagnostic colors mediate scene recognition.** *Cognitive psychology*, **41**(2):176–210, 2000. 16, 22, 23
- [120] A. H MUNSELL. **A pigment color system and notation.** *The American Journal of Psychology*, **23**(2):236–244, 1912. 17, 21
- [121] JONI ORAVA, TIMO JAASKELAINEN, AND JUSSI PARKKINEN. **Color differences in a spectral space.** In *IS AND TS PICS CONFERENCE*, pages 205–209. SOCIETY FOR IMAGING SCIENCE & TECHNOLOGY, 2003. 17
- [122] D. B JUDD, D. L MACADAM, G. WYSZECKI, HW BUDDE, HR CONDIT, ST HENDERSON, AND JL SIMONDS. **Spectral distribution of typical daylight as a function of correlated color temperature.** *JOSA A*, **54**(8):1031–1040, 1964. 17
- [123] JÁNOS SCHANDA. *Colorimetry: understanding the CIE system.* John Wiley & Sons, 2007. 17

REFERENCES

- [124] C-C. CHIAO, T. W. CRONIN, AND D. OSORIO. **Color signals in natural scenes: characteristics of reflectance spectra and effects of natural illuminants.** *JOSA A*, **17**(2):218–224, 2000. 17
- [125] JAMES M KRAFT AND DAVID H BRAINARD. **Mechanisms of color constancy under nearly natural viewing.** *Proceedings of the National Academy of Sciences*, **96**(1):307–312, 1999. 17
- [126] JOONG NAM YANG AND LAURENCE T MALONEY. **Illuminant cues in surface color perception: Tests of three candidate cues.** *Vision Research*, **41**(20):2581–2600, 2001. 17
- [127] JAMES M KRAFT, SHANNON I MALONEY, AND DAVID H BRAINARD. **Surface-illuminant ambiguity and color constancy: Effects of scene complexity and depth cues.** *Perception*, **31**(2):247–263, 2002. 17
- [128] STACEY ASTON, ANA RADONJIĆ, DAVID H BRAINARD, AND ANYA C HURLBERT. **Illumination discrimination for chromatically biased illuminations: Implications for color constancy.** *Journal of vision*, **19**(3):15–15, 2019. 17
- [129] NIKOLAUS KRIEGESKORTE, MARIEKE MUR, AND PETER A BANDETTINI. **Representational similarity analysis-connecting the branches of systems neuroscience.** *Frontiers in systems neuroscience*, **2**:4, 2008. 17
- [130] GEOFFREY KARL AGUIRRE. **Continuous carry-over designs for fMRI.** *Neuroimage*, **35**(4):1480–1494, 2007. 17
- [131] MICHAEL AA COX AND TREVOR F COX. **Multidimensional scaling.** In *Handbook of data visualization*, pages 315–347. Springer, 2008. 18
- [132] QASIM ZAIDI. **Decorrelation of L-and M-cone signals.** *JOSA A*, **14**(12):3430–3431, 1997. 20
- [133] MATTHEW S CAYWOOD, BENJAMIN WILLMORE, AND DAVID J TOLHURST. **Independent components of color natural scenes resemble V1 neurons in their spatial and color tuning.** *Journal of Neurophysiology*, **91**(6):2859–2873, 2004. 20
- [134] BEVIL R CONWAY. **Color vision, cones, and color-coding in the cortex.** *The neuroscientist*, **15**(3):274–290, 2009. 20

REFERENCES

- [135] ROBERT SHAPLEY AND MICHAEL HAWKEN. **Neural mechanisms for color perception in the primary visual cortex.** *Current opinion in neurobiology*, **12**(4):426–432, 2002. 20
- [136] HIDEHIKO KOMATSU. **Mechanisms of central color vision.** *Current opinion in neurobiology*, **8**(4):503–508, 1998. 20
- [137] QASIM ZAIDI AND BEVIL CONWAY. **Steps towards neural decoding of colors.** *Current Opinion in Behavioral Sciences*, **30**:169–177, 2019. 20, 23, 24
- [138] BEVIL R CONWAY AND DORIS Y TSAO. **Color architecture in alert macaque cortex revealed by fMRI.** *Cerebral Cortex*, **16**(11):1604–1613, 2006. 20
- [139] KATHY T MULLEN, DORITA HF CHANG, AND ROBERT F HESS. **The selectivity of responses to red-green colour and achromatic contrast in the human visual cortex: an fMRI adaptation study.** *European Journal of Neuroscience*, **42**(11):2923–2933, 2015. 20
- [140] RADOSLAW M CICHY, ADITYA KHOSLA, DIMITRIOS PANTAZIS, ANTONIO TORRALBA, AND AUDE OLIVA. **Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition.** *arXiv preprint arXiv:1601.02970*, 2016. 20
- [141] KARL R GEGENFURTNER. **Cortical mechanisms of colour vision.** *Nature reviews. Neuroscience*, **4**(7):563, 2003. 20
- [142] JONATHAN J NASSI AND EDWARD M CALLAWAY. **Parallel processing strategies of the primate visual system.** *Nature Reviews Neuroscience*, **10**(5):360, 2009. 20
- [143] CIE. *Recommendations on uniform color spaces, color-difference equations, psychometric color terms.* 1978. 21
- [144] D. L PHILIPONA AND J K. O’REGAN. **Color naming, unique hues, and hue cancellation predicted from singularities in reflection properties.** *Visual neuroscience*, **23**(3-4):331–339, 2006. 21
- [145] ALBAN FLACHOT, EDOARDO PROVENZI, AND J KEVIN O’REGAN. **An illuminant-independent analysis of reflectance as sensed by humans, and its applicability to computer vision.** In *Proceedings of the 6th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2016. 21

REFERENCES

- [146] ROBERT GEIRHOS, DAVID HJ JANSSEN, HEIKO H SCHÜTT, JONAS RAUBER, MATTHIAS BETHGE, AND FELIX A WICHMANN. **Comparing deep neural networks against humans: object recognition when the signal gets weaker.** *arXiv preprint arXiv:1706.06969*, 2017. 21, 23
- [147] FELIX A WICHMANN, DAVID HJ JANSSEN, ROBERT GEIRHOS, GUILLERMO AGUILAR, HEIKO H SCHÜTT, MARIANNE MAERTENS, AND MATTHIAS BETHGE. **Methods and measurements to compare men against machines.** *Electronic Imaging*, **2017**(14):36–45, 2017. 23
- [148] ANNA W ROE, LEONARDO CHELAZZI, CHARLES E CONNOR, BEVIL R CONWAY, ICHIRO FUJITA, JACK L GALLANT, HAIDONG LU, AND WIM VANDUFFEL. **Toward a unified theory of visual area V4.** *Neuron*, **74**(1):12–29, 2012. 24
- [149] ROSA LAFER-SOUSA AND BEVIL R CONWAY. **Parallel, multi-stage processing of colors, faces and shapes in macaque inferior temporal cortex.** *Nature neuroscience*, **16**(12):1870–1878, 2013. 24

REFERENCES

2

Publications

2. PUBLICATIONS

2.0.1 Study 1: Processing of chromatic information in a deep convolutional neural network



Processing of chromatic information in a deep convolutional neural network

ALBAN FLACHOT AND KARL R. GEGENFURTNER* 

Abteilung Allgemeine Psychologie, Justus-Liebig-Universität-Giessen, 35394 Giessen, Germany

*Corresponding author: gegenfurtner@uni-giessen.de

Received 2 November 2017; revised 6 March 2018; accepted 7 March 2018; posted 8 March 2018 (Doc. ID 312517); published 30 March 2018

Deep convolutional neural networks are a class of machine-learning algorithms capable of solving non-trivial tasks, such as object recognition, with human-like performance. Little is known about the exact computations that deep neural networks learn, and to what extent these computations are similar to the ones performed by the primate brain. Here, we investigate how color information is processed in the different layers of the AlexNet deep neural network, originally trained on object classification of over 1.2M images of objects in their natural contexts. We found that the color-responsive units in the first layer of AlexNet learned linear features and were broadly tuned to two directions in color space, analogously to what is known of color responsive cells in the primate thalamus. Moreover, these directions are decorrelated and lead to statistically efficient representations, similar to the cardinal directions of the second-stage color mechanisms in primates. We also found, in analogy to the early stages of the primate visual system, that chromatic and achromatic information were segregated in the early layers of the network. Units in the higher layers of AlexNet exhibit on average a lower responsivity for color than units at earlier stages. © 2018 Optical Society of America

OCIS codes: (330.4060) Vision modeling; (330.4270) Vision system neurophysiology; (330.1690) Color.

<https://doi.org/10.1364/JOSAA.35.00B334>

1. INTRODUCTION

Deep neural networks have emerged as the state-of-the-art algorithms for artificial intelligence and computer vision applications. In the challenging task of object recognition, deep convolutional neural networks (CNNs) have reached and even surpassed human performance in the ImageNet large-scale visual recognition challenge [1]. These algorithms have biologically inspired architectures mimicking the hierarchical processing structure of the primate brain [2], and have been proposed as potentially useful models of how the primate brain executes feed-forward visual processing [3–5]. However, little is currently known about the internal representations these algorithms develop during training, and to what extent the computations performed by these networks are similar to the known processing stages of the human and primate visual cortex.

A common application of CNNs is object recognition. Object recognition is considered to be one of the most important purposes of the development of biological visual systems [5,6], and one of the major questions is how different visual attributes contribute to this process. Here, we focus on color, as color is known to play a significant part in object recognition. Psychophysical studies in humans have shown that adding color information leads to faster and more accurate

performance in object and scene classification tasks, as well as improvements in visual memory for such stimuli [7–12]. In particular, color was found to highly facilitate the recognition of color-diagnostic objects, i.e., objects with which a specific color can be associated such as lemons or strawberries [7,10,11].

Color also seems to play a critical role for CNNs trained in object recognition. In a recent study [13], the authors compared the rate of successful classifications in three widely used CNNs for achromatic and colored images in 16 object classes. In parallel, they carefully designed a psychophysical experiment and tested three human subjects for comparison purposes. Despite their different architectures, all three CNNs showed a significant 3%–8% drop in accuracy on achromatic relative to colored images, while humans showed a somewhat smaller difference (0.5%–4%). All of the selected object classes in this study had weak color diagnosticity, suggesting that color plays a significant role in object recognition in CNNs, even for non-color diagnostic objects.

Several studies have worked towards a better understanding of the processing of visual information in CNNs trained for object recognition [14–17]. Yet very few, to our knowledge, focused on their processing of color information. Only recently, Rafegas and colleagues [18,19] studied the color responsivity and tuning of units in the different layers of one training

instance of the VGG-M network [20]. To do so, for each kernel learned by the VGG-M network after training, they extracted the N first stimuli responsible for their maximal activation. They then made the assumption that the mean features of these N stimuli were representative of the features of their corresponding kernel and analyzed their chromatic characteristics. We use a different approach here, with a fixed set of highly controlled stimuli similar to what has been used in numerous electrophysiological experiments. We also investigate a different CNN, AlexNet, that has been compared to the human visual system via several different methods. Khaligh-Razavi and Kriegeskorte [3] have shown that AlexNet, when trained for object recognition, can better explain inferotemporal cortex activity, measured by functional magnetic resonance imaging (fMRI), than an extensive list of other computational models, and also fairly well explains activities in mid-level visual areas. It has also been shown that early- and mid-convolutional layers of AlexNet correlate with early and anterior areas of the occipital visual cortex, while the late fully connected layers correlate with areas of the temporal cortex [4,21], thus providing a clear correspondence between layers and different stages of the processing in the human visual system. Furthermore, AlexNet has the interesting feature of often being implemented with two independent streams in the early stages of its processing. This was a purely pragmatic feature, designed to reduce computational demands during training by splitting the layers with the largest numbers of units over two independent GPUs. Curiously, although both streams are initialized randomly, Alexnet's developers reported in their original paper that each stream becomes somewhat specialized after training, in either the processing of the chromatic or achromatic information [22]. However, they did not further investigate this segregation and whether it bears any consequences for network performance.

In the first part of this study, we focus on the chromatic properties of the kernels in the first layer of a set of 35 training instances. Thanks to the high linearity of these kernels, we could directly analyze their weights and characterize their chromatic tuning. In the second part, we focus on the color processing occurring in the deeper layers of AlexNet. Because kernels in these layers can no longer be considered linear, we used our physiologically inspired approach to assess the chromatic properties of kernels up to the last convolutional layer. In the third part, we combine the two approaches and measure, in layer 1 and 2 of AlexNet, to what degree the functional segregation reported by AlexNet's developers systematically occurs throughout our set of 35 training instances. We further correlate our results with the performance of the model.

2. METHODS

A. AlexNet

Deep convolutional neural networks are layered algorithms, with each layer performing a set of processing operations. Like most other CNNs, AlexNet is a feedforward system. It takes as input a $227 \times 227 \times 3$ image and outputs the category the input image most likely belongs to. The first two input dimensions represent the spatial extent of the image (width and height), and the third input dimension represents the three RGB color channels. AlexNet consists of convolutional layers and fully connected layers. A convolutional layer consists of a set

of linear kernels (i.e., filters) with equally sized receptive fields (e.g., $11 \times 11 \times 3$ in the first layer) at equally spaced intervals, followed by half-wave rectification (ReLU) [22]. This results in a two-dimensional map encoding the response of a given filter at each spatial position. The activation maps from all filters within a layer are stacked to produce the output volume of that layer, which is the input volume of the next layer. In fully connected layers, the network units get input from all units of the previous layer. The units in fully connected layers thus have receptive fields the same size as the input image, and their activation maps can be computed through a simple multiplication of their weights with the previous unit's responses. AlexNet's architecture consists of five convolutional layers followed by three fully connected layers. The convolutional layers 1, 2, and 5 of the AlexNet architecture are followed by max pooling, a down-sampling operation which reduces the size of the input volume along its first two dimensions by taking the maximum response of 3×3 neighboring units. Following the pooling operations in layers 1 and 2 are two normalization layers. Most of these features of the AlexNet architecture are shared by several most recent and efficient architectures, in terms of the nature of the layers [14,15,20,23,24], number of layers [14,20], and nonlinearities implemented [14,15,20,23,24].

AlexNet has an interesting feature which it does not share with other network architectures: throughout its convolutional layers 1 and 2, processing occurs in two functionally separate and independent processing streams. This segregation occurs because the network was originally trained on two separate graphic processing units (GPUs), and the authors found that restraining the connectivity between the two GPUs was highly beneficial for training efficiency. Thus, two parallel, intra-layer processing streams were built into the first two convolutional layers of the AlexNet architecture. Each layer was divided into two groups of kernels; one group was trained on GPU₁ while the other group was trained on GPU₂. All the way up to the convolutional layer 3, each network unit receives input only from the units of the previous layer residing on the same GPU. The training of AlexNet and other deep convolutional neural networks includes several randomization steps [22,25]. This implies that different network instances are likely to contain different network parameters even though they might lead to similar levels of performance. We investigated the AlexNet instantiation available with the Caffe framework and trained by the Berkeley team [1,22], denoted by AlexNet_B, together with 34 novel instances trained on the same training dataset as the original network and using the same procedures [22], but with each one randomly initialized with values drawn from a standard normal distribution. Training images were taken from the 2012 ImageNet image dataset, which is composed of over 1.2 million realistic JPEG images [26]. Prior to each training run, the image dataset was randomly shuffled. The implementation and training of each network instance was done through the Caffe deep learning framework [27] via its Python interface. All analyses presented in this work were scripted in Python.

B. Chromatic Coordinates

The processing taking place in the photoreceptors and in the second-stage color opponent channels of the retinal ganglion

cells has been characterized in great detail [28–30]. Psychophysically, three cardinal mechanisms have been shown to provide independent axes, and in electrophysiological recordings these axes could be related to different layers of lateral geniculate neurons. One of these mechanisms is thought to convey luminance information by combining information from long and medium wavelength cones (L + M). The remaining two channels convey chromatic information and exhibit color opponent mechanisms by taking differences across cone activations (L – M) and S – (L + M) (where S represents short wavelength cones). Computationally, this transformation provides an efficient coding and decorrelation of the cone signals, akin to a principal component analysis [31,32]. Since one major aim of our study was to compare processing in AlexNet with known primate physiology, we wanted to test whether the CNN used similar coding mechanisms in its first layers.

Our initial approach was therefore to calculate cone excitations from the RGB input images, convert these into the cardinal direction coordinates (DKL) proposed by Derrington, Krauskopf, and Lennie in 1984 [30], and use these new coordinates to characterize chromatic processing in AlexNet. One immediate difficulty with this approach is that the AlexNet training set consists of 1.2M uncalibrated and JPEG compressed images, which makes it impossible to recover cone excitations from the camera RGBs. On average, a camera standard such as sRGB [33] might produce reasonable results. However, doing so made it obvious that the concept of “luminance,” based on flicker photometry and leading to an approximate ratio of 10:3:1 for the sRGB primaries, does not make sense for a network trained on RGB images. Therefore, we decided to use as the relevant input quantities the principal components of the complete training set. The first principal component represents intensity by summing the R, G, and B inputs. The second and third principal components represent decorrelated chromatic dimensions: R – B for the second principal component and G – (R + B) for the third. The variances explained by the three PCs are 90%, 8%, and 2%, respectively, meaning that intensity is the dimension with overwhelmingly the most discriminative power. These vectors are virtually identical to the ones found by Otha *et al.* (1980) [34] for a much smaller set of images. We will refer to the new system of chromatic coordinates defined by the principal components as the RGB_{PCA} coordinates. All of our analyses will therefore be presented solely in the RGB_{PCA} coordinates, or their spherical representation as azimuth and elevation. To understand the conversion from Cartesian to spherical coordinates, we may consider a point *P* in a given color space and call **V** the vector associated with *P*. The elevation, or ϕ , of a point in space represents the angle between the vector **V** and the chromaticity plane, with $\phi \in [-90, 90]$. Conversely, the *azimuth*, or ψ , represents the angle between the projection of **V** on the chromaticity plane and the first chromaticity axis. Note that the first chromaticity axis will be given by the second principal component in the case of the RGB_{PCA} chromatic coordinates and the L–M axis in the case of the DKL color coordinates. ψ can assume values in the interval [0, 360]. A convenient aspect of this coordinate transformation is that the length of the projection of **V** on the chromaticity plane intuitively represents the *chromatic contrast* of

P. Since the terms *azimuth* and *elevation* are mainly used in the context of DKL coordinates in color science, we will use the subscript PCA here.

C. AlexNet First Layer Kernels

The 96 kernels in the first convolution layer of AlexNet are linear filters followed by half-wave rectification [22], composed of 363 weights ($11 \times 11 \times 3$). The filters' activation results from the weighted sum of the input image pixel values coded in RGB. Hence, the 363 weights fully describe each filter, and can be directly expressed in the RGB space and then converted into RGB_{PCA} coordinates following the procedures described previously. Since we focus on the color properties of the kernels in this study, we will disregard their spatial properties. For a given kernel, the direction in color space in which the kernel weights vary the most defines the direction for which the kernel is most sensitive to. Therefore, we performed a PCA on the distribution of the kernel's weights in RGB color space expressed in RGB_{PCA} coordinates. The first principal component defines the direction in color space for which a kernel is mainly sensitive, and the variance explained by this first principal component indicates how narrow the distribution of weights is around this direction (i.e., how preferably tuned the kernel is to this direction in color space).

D. Artificial Cell Responses

The kernels in the first layer of AlexNet can be directly expressed in color space, and this provides a straightforward way of assessing how these kernels process color information. Different methods need to be used to probe the color processing of kernels in deeper layers of AlexNet. Physiologists are of course confronted with the same kind of problem (see [35–39] for reviews). A common way to investigate color processing at different stages throughout the primate visual system is to record the response of single neurons to simple, highly constrained color stimuli, thus characterizing each neuron's color tuning curves [30,40–45]. We will use the same approach here to characterize the color tuning characteristics of individual units of the AlexNet algorithm.

We opted for a fixed set of stimuli, all sharing the same spatial characteristics, which only varied in color and in achromatic contrast with the background. Stimuli were RGB images of colored and achromatic disks on a gray background. The images were 227×227 pixels in size, as required by the AlexNet architecture. The disks had a 70 pixel radius and were placed at the center of the image. Disks were chosen as the basic stimulus shape since the color responsive kernels of the first layer of AlexNet are for the most part selective for low spatial frequencies at different orientations. Hence, employing circular stimuli allowed us to roughly fit the spatial characteristics of all the different kernels across different orientations. The color of our stimuli was initially generated in the RGB_{PCA} chromatic coordinates and then converted into RGB values.

All color stimuli had the same constant chromatic-contrast_{PCA} of 0.3 and the same neutral gray background, set to an intensity_{PCA} value of 0. We designed a total of 300 color stimuli at 60 different azimuth_{PCA} values (ψ ranging from 0 to 360° in 6° steps) and at five different intensity_{PCA} values (*I* ranging from –0.6 to 0.6 in steps of 0.3).

Intensity_{P_{CA}} values were chosen such that, at -0.3 and $0.3I$, the intensity_{P_{CA}} difference between the central disk and the background and the chromatic-contrast_{P_{CA}} were of equal magnitude. The choice of a relatively low chromatic-contrast_{P_{CA}} guaranteed that our stimuli would remain within the RGB gamut. Henceforth, we will sometimes describe our stimuli in terms of their elevation_{P_{CA}} ϕ instead of their intensity_{P_{CA}} value. Since chromatic-contrast_{P_{CA}} is kept constant, the conversion from intensity_{P_{CA}} to elevation_{P_{CA}} is straightforward: $\phi = \arctan(\frac{I}{0.30})$. In addition to these color stimuli, we also designed five achromatic (gray) stimuli, one for each of the five intensity_{P_{CA}} values employed to construct the color stimuli.

The convolutional layers of deep neural networks are designed such that a set of artificial neurons applies each kernel of the layer at equally spaced intervals that tile the input volume. For example, AlexNet's first convolutional layer is composed of 96 different kernels. Each of the 96 kernels is applied throughout the input image at four pixel intervals. Therefore, 55×55 network units are needed in order to apply each kernel throughout the 227×227 input image. This means that AlexNet's first convolutional layer contains a total of $55 \times 55 \times 96$ units. However, this study focuses on the chromatic properties of the kernels of AlexNet independently of their spatial properties. Therefore, we did not record the response of each artificial neuron in each layer. Instead, for each kernel K , we selected the maximal response of its set of units to each stimulus image. If we denote N_i as the number of kernels of the convolutional layer i , we thus end up with only N_i tuning curves \mathbf{Y}_ϕ^K per layer for each elevation_{P_{CA}} level ϕ of our set of stimuli. If we denote $Y_{\phi,\psi}^K$ as the component of \mathbf{Y}_ϕ^K corresponding to the azimuth_{P_{CA}} ψ , we obtain

$$Y_{\phi,\psi}^K = \max_j R_{\phi,\psi}^{K_j}, \quad (1)$$

where $R_{\phi,\psi}^{K_j}$ is the response of the j th artificial neuron applying the K th kernel map for a stimulus whose color is defined by elevation_{P_{CA}} ϕ and azimuth_{P_{CA}} ψ .

E. Linearity

Some studies have shown that, throughout the hierarchy of the primate visual system, the tuning of color responsive neurons increases in complexity and becomes progressively more non-linear [41,46]. Because of the non-linearities present throughout AlexNet and deep neural networks in general, we similarly expect that the tuning of color-responsive kernels in the deep layers of these networks should also become increasingly non-linear. By definition, the response of a linear kernel is given by the dot product of its parameters with the input values. In our case, at iso-intensity, this means that the response of a linear kernel is given by the dot product, in color space, between the stimulus vector and the direction for which the kernel is maximally tuned. Some cells have been found on several occasions to behave in a similar fashion in the early visual system [30,40], and were thus called linear. For such cells, the response R_ϕ to a set of chromatic stimuli with a given elevation_{P_{CA}} ψ is given by the equation

$$R_\phi(\psi) = B + A|b + \cos(\psi - az)|, \quad (2)$$

where A is the maximum response amplitude, B is the response baseline and b is constant, az is the preferred azimuth_{P_{CA}} of the kernel, and ψ is the azimuth_{P_{CA}} of the stimulus. We determined tuning curves empirically, e.g., using the methods described in subsection 2.D. Then, if Eq. (2) provides a good fit to the tuning curves, we may conclude that the response of the cell or network kernel is well approximated by the linear combination of the chromatic input to the system. If the properties of AlexNet match those of the visual system, the quality of the fit should decrease in higher layers.

F. Color and Achromatic Responses

We also measured color responsivity throughout AlexNet using a metric taken from the physiology literature [44]. This measure, which we call here color responsivity (CR), can be defined as follows:

$$CR = \frac{\text{max response to color} - \text{response to gray}}{\text{max response to color} + \text{response to gray}}, \quad (3)$$

where gray and colored stimuli have the same intensity-contrast_{P_{CA}}. A color-responsivity value of -1 occurs when a unit exhibits no response to any colored stimulus but is responsive to achromatic stimuli. Conversely, a color-responsivity value of 1 occurs when a unit responds to at least one colored stimulus and is unresponsive to achromatic stimuli.

We classified kernels into color and achromatic kernels according to their chromatic properties. In the case of the linear kernels of the first layer of AlexNet, color kernels are those tuned for directions in color space with a chromatic component larger than the achromatic component, i.e., tuned for elevations_{P_{CA}} $\leq 45^\circ$. In the case of kernels in subsequent layers of AlexNet, the distinction is not as straightforward, as the tuning characteristics of these kernels cannot be directly expressed in color space. Instead, the color responsivity CR, measured for colored stimuli with chromatic and intensity contrasts_{P_{CA}} of equal magnitude, may be used to classify kernels. According to Eq. (3), a unit with a response to one chromatic stimulus twice as large as its response to an achromatic stimulus with the same intensity features has a color responsivity of $1/3$. Thus, following the definition of color kernels given previously, a color kernel has a color responsivity $\geq 1/3$, given that the chromatic contrast_{P_{CA}} of the colored stimuli is *equal* to their intensity contrast_{P_{CA}}. For colored stimuli with different achromatic and chromatic contrasts, the distinction between color and achromatic (luminance) kernels is given by different thresholds.

G. Measure of Functional Segregation

The differentiation between color and achromatic kernels we propose in this study is motivated by the segregation observed by Krizhevsky *et al.* [22]. This feature of AlexNet is potentially very interesting particularly because AlexNet exhibits a unique architecture, with two functionally separate and independent processing streams implemented on two separate GPUs (which we refer to as residing on GPU₁ and GPU₂). This two-stream architecture is reminiscent of the organization of the primate visual system into two separate visual processing streams, which also exhibits a segregation between color and achromatic processing [36,47,48]. If segregation between color and

achromatic tuning were indeed found to systematically occur across multiple AlexNet training instances, this would suggest that the segregation naturally emerges from an architecture with independent processing streams.

To measure the degree of functional segregation occurring in AlexNet we employ the *Index of Dissimilarity* (ID) [49], widely used in the field of economics, and defined as

$$ID = \frac{1}{2} \left(\left| \frac{c_1}{C} - \frac{l_1}{L} \right| + \left| \frac{c_2}{C} - \frac{l_2}{L} \right| \right), \quad (4)$$

where C and L are the total number of color and luminance kernels, respectively; c_1 , l_1 are the number of color and luminance kernels residing on GPU₁; and c_2 , l_2 are the number of color and luminance kernels residing on GPU₂. An index of dissimilarity of 0 corresponds to the absence of functional segregation, whereas an index of dissimilarity of 1 corresponds to a complete functional segregation.

3. RESULTS

A. Chromatic Properties of First-Layer Kernels

Following the procedure described in subsection 2.C, we expressed each kernel of the first layer of our training instances in terms of their tuning in elevation_{PCA} and elevation_{PCA} < 27°. We did so after performing a PCA on their weights expressed in RGB_{PCA} coordinates. The first principal component explains on average 94.3% of the variance of kernel weights. About 90% of kernels had over 84.2% of their weights' variance explained by the first principal component. This result confirms that individual kernels in the first layer of AlexNet have weights distributed primarily along one direction in color space. Figure 1 displays the tuning directions found for all kernels in the first layer of all 35 training instances of AlexNet as a scatter plot of the preferred elevation_{PCA} (in absolute value) and azimuth_{PCA}, marginal histograms of preferred elevation_{PCA} values across all azimuths_{PCA}, and of preferred azimuths_{PCA} for color kernels.

The histogram of preferred elevation_{PCA} values in Fig. 1 exhibits a bimodal distribution, with one peak at high elevations_{PCA} close to 90° and a second peak at elevations_{PCA} near 0°. The marginal histogram of preferred azimuths_{PCA} in Fig. 1 further shows that kernels tuned for low elevations_{PCA} exhibit azimuth_{PCA} values that cluster roughly around 0° and 180°, and to a lesser degree around 90° and 270°. Therefore, the tuning of kernels in the first layer of AlexNet aligns well with the axes of the RGB_{PCA} chromatic coordinates. This indicates that the kernels in the first layer basically perform a PCA on the RGB triplets. The bimodal distribution of preferred kernel elevation_{PCA} angles suggests that first-layer kernels are likely to be either strongly color responsive or strongly color agnostic. Additionally, the mean and median values of absolute elevation_{PCA} tuning were 47.6° and 58.4°, respectively. The observed median elevation_{PCA} corresponds to the angle for which the intensity component of a kernel is 1.6 times larger than its chromatic component. In other words, nearly half of the kernels in the first layer of AlexNet are at least 1.6 times more responsive to intensity contrasts than to chromatic contrasts. This imbalance is sensible, since 90% of variance in the pixel values of the training dataset is distributed

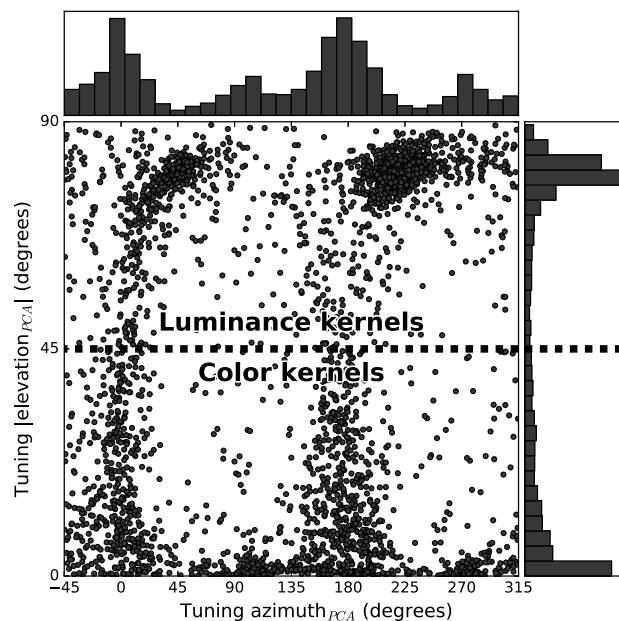


Fig. 1. Preferred tuning directions of first-layer kernels in RGB_{PCA} coordinates. Kernel tuning directions represented in RGB_{PCA} coordinates. In the left panel, individual dots are preferred elevation_{PCA} angles (in absolute value) plotted against preferred azimuth_{PCA} angles for each kernel in the first layer of all 35 training instances of AlexNet. Dotted lines represent the 45° threshold elevation value employed to classify kernels as either color ($\phi \leq 45^\circ$) or luminance ($\phi \geq 45^\circ$) kernels. The right panel shows the histogram of preferred elevation_{PCA} values across all azimuths_{PCA}. The top panel shows the histogram of preferred azimuth_{PCA} values across color kernels only.

along the intensity axis. The bimodal distribution of network kernels in Fig. 1 highlights that only a few kernels respond to both intensity and chromatic contrasts. This suggests it is indeed appropriate to classify first-layer network kernels into color and luminance kernels. Figure 2 shows all 96 kernel maps

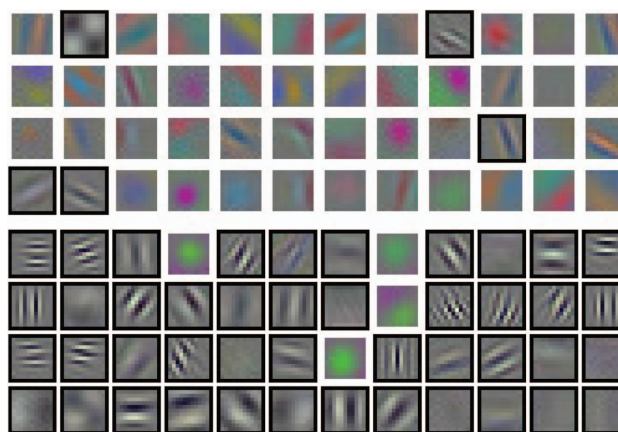


Fig. 2. Kernels of the first layer of AlexNet_B. Kernels are displayed according to their index order in the architecture. The top group of 48 kernels was trained on GPU₁. The bottom group of 48 kernels was trained on GPU₂. Outlined in black are those kernels classified as luminance kernels following the procedure described in Section 2.F. Note how the majority of GPU₁ kernels are color kernels, while the majority of GPU₂ kernels are luminance kernels.

of the first layer of AlexNet_B, where luminance kernels are outlined in black and color kernels are outlined in white. Qualitatively, we can observe that our classification method correctly identifies whether a kernel is or is not tuned to colors. This figure also allows some other qualitative remarks: color kernels have color-opponent receptive fields while achromatic kernels have achromatic-opponent receptive fields. Color kernels are either oriented or non-oriented. These features are shared by the kernels in the first layer of the 34 other training instances.

Across all 35 training instances of AlexNet we investigated, on average $54 \pm 3\%$ of first-layer kernels were classified as luminance kernels. This rough equality between the number of luminance and color kernels is a common feature of neural models. Rafegas and colleagues in 2017 [18,19] observed similar proportions in the first layer of the VGG-M net, as did studies implementing linear models [32,50–52].

The response of luminance kernels depends very little on the chromatic variations in the input. Therefore, the chromatic information contained within the input images will be primarily transmitted to upstream network layers through the output of color-selective kernels. For this reason, we now only focus on the chromatic properties of kernels which we have identified as color kernels. Figure 3(a) shows the distribution of preferred azimuth_{PCA} across all color kernels in the first layer of all 35

AlexNet training instances. The central symmetry of the histogram arises from the linear and opponent nature of the kernels. Panels B through D show the distributions of preferred azimuth_{PCA} in the first-layer kernels of three individual AlexNet training instances randomly selected out of the total set of 34. Panel E shows the same distribution of preferred azimuth_{PCA}, except for AlexNet_B.

Several things are striking about the histograms in Fig. 3. First of all, azimuth_{PCA} tuning of all kernels in all AlexNet training instances exhibits a clear bimodal distribution. Network kernels appear to be preferably tuned to two directions in the chromatic plane, one direction along the blueish–orange axis, and one direction along the green–magenta axis. Secondly, the directions defined by the two modes of the azimuth_{PCA} distribution are remarkably well aligned with the two chromatic axes of the RGB_{PCA} chromatic coordinates, with the greatest discrepancy being a small 5° misalignment of the green–magenta axis. The peak along the blueish–orange axis is larger than the peak occurring along the green–magenta axis corresponding to the third principal component of the RGB_{PCA} coordinates. The asymmetry between the two modes of the azimuth distribution thus follows the difference in discriminative power between the second and third principal components defining the chromatic axes of the RGB_{PCA} chromatic coordinates. The alignment of the network's tuning with the principal chromatic directions of RGB_{PCA} is not as obvious when inspecting the azimuth_{PCA} tuning of individual training instances. A few representative examples are given in Figs. 3(b)–3(e). The similarities between AlexNet's chromatic tuning directions and the RGB_{PCA} chromatic axes strongly suggest that kernels in the first layer of AlexNet are attempting to transform and decorrelate the input signal in a fashion which is similar to a principal component analysis.

B. Linearity and Color Responsivity in Deep Layers

The characteristics of color processing occurring in kernels beyond the first convolutional layer cannot be directly accessed and analyzed. Therefore, we employed an indirect approach, inspired from physiology, to investigate color processing in the AlexNet deep layers. Our approach is described in detail in subsections 2.D and 2.E.

Figure 4 shows some examples of tuning curves obtained in different convolutional layers of AlexNet, fitted using Eq. (2). We observe that the fits to Eq. (2) seem very accurate in the first few layers, and that accuracy drops in higher layers. This observation is quantitatively confirmed in Fig. 5(a), in which the mean accuracy of the fitted tuning curves (in terms of r^2) is shown to decrease with network depth (i.e., layer number). This was to be expected, as the responses of the artificial neurons in higher layers are preceded by several nonlinear processing steps (see subsection 2.A). Fit accuracy decreases almost linearly, from a mean r^2 above 0.9 in layer 1 down to a mean r^2 only slightly above 0.5 for layer 7. Concerning the chromatic tuning of linear kernels in layer 1, both methods give nearly identical results. For 50% of the first-layer kernels, the difference was less than 3° between the two methods, and the bimodality around the two chromatic axes of the RGB_{PCA} coordinates was present with both methods as well.

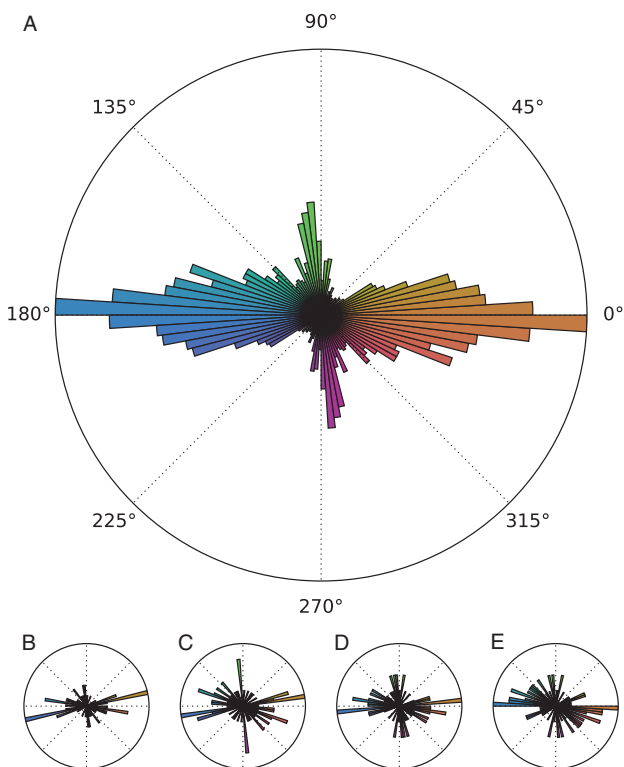


Fig. 3. Azimuth tuning distributions of color kernels in RGB_{PCA}. (a) Circular histogram displaying the distribution of preferred azimuth_{PCA} of all color kernels in the first layer of all 35 AlexNet training instances. (b–d) Distributions of preferred azimuth_{PCA} of color, first-layer kernels for three individual AlexNet training instances selected from our set of 35. (e) Same as (b–d), except for the AlexNet_B network instantiation provided within the CAFFE framework.

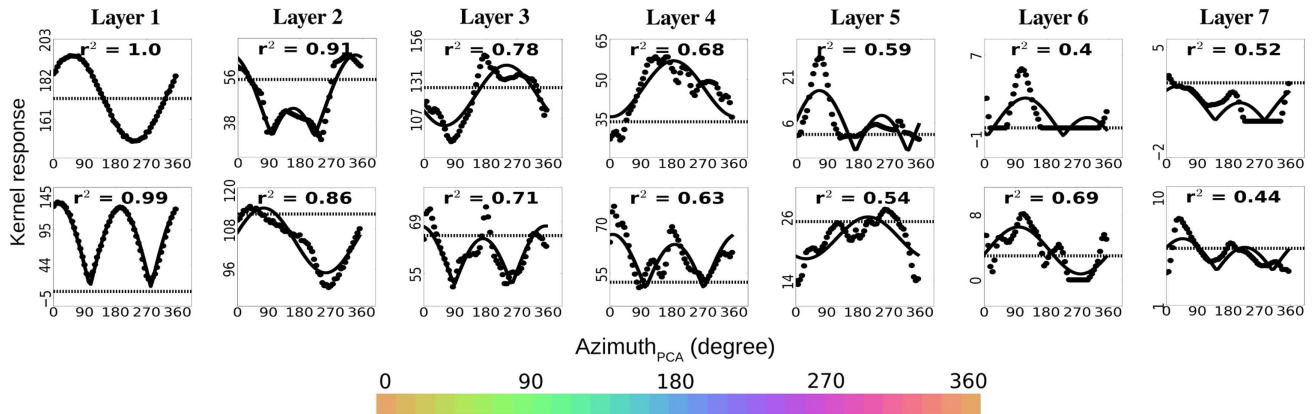


Fig. 4. Color tuning curves from representative kernels in the different AlexNet layers. In each panel, the dotted curve represents the response of a kernel, plotted as a function of the azimuth_{PCA} of the input stimulus. The continuous line is the tuning curve of filter that linearly combines chromatic input, fitted to the kernel response using the formula in Eq. (2). The horizontal dashed line is the kernel's response to an achromatic (gray) stimulus with the same spatial and intensity characteristics as the colored stimuli.

Physiological studies have shown that the response of early visual neurons, in the primate lateral geniculate nucleus (LGN) and in smaller proportion in V1 [40,46,53,54], are also well fitted by Eq. (2). Therefore, the early visual processing units of AlexNet and of the primate visual system both behave like linear filters of chromatic input. Additionally studies in primates have shown that cells in visual areas as early as V1, but mostly V2 and beyond, show more complex tunings than the one described by Eq. (2) [40,41,54–56]. Therefore, as processing progresses throughout the hierarchy of both biological and artificial visual systems, cells or units process color information in ways which are increasingly nonlinear.

To facilitate the conversion from color responsivity to elevation_{PCA} tuning in layer 1, we used exclusively in the following analysis the set of stimuli for which the intensity and

chromatic contrasts_{PCA} were of the same magnitude (cf. subsection 2.F). Figure 5(b) shows the mean proportion of units exhibiting noticing behaviors across all 35 AlexNet training instances, in every network layer up to layer 7. The full bold line gives the proportion of *unresponsive* kernels across our set of stimuli, i.e., of kernels which have no response to our stimuli. The dashed-dotted line gives the proportion of *luminance-only* kernels with sensitivity to intensity contrasts_{PCA} at least two times superior to their sensitivity to chromatic contrasts_{PCA} (elevation_{PCA} > 63° in layer 1, color responsivity < 1/5). The full fine line is the proportions of *color-only* kernels with sensitivity to chromatic contrasts_{PCA} at least two times superior to intensity contrasts_{PCA} (elevation_{PCA} < 27° in layer 1, color responsivity > 1/2). The dashed line is the proportion of *color-luminance*

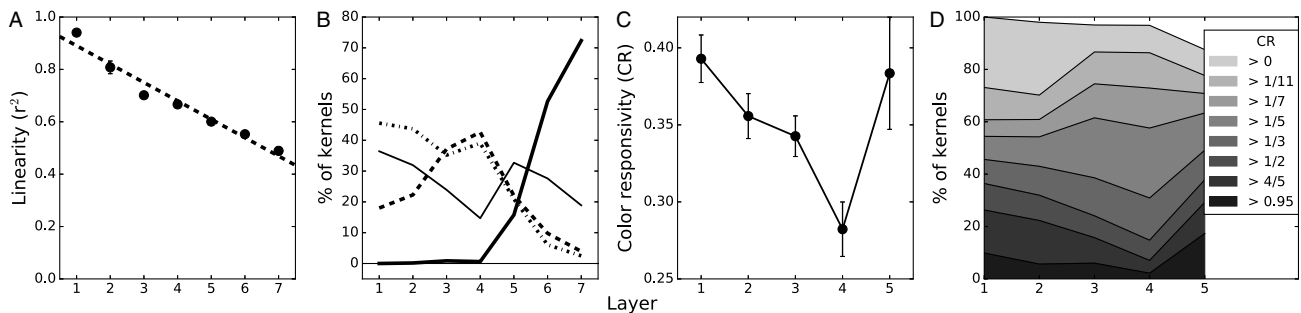


Fig. 5. Linearity, responsivity, and color responsivity of kernels across the different AlexNet layers. (a) Fit accuracy plotted as a function of network depth (i.e., layer number). Accuracy data are the r^2 score between kernel responses to colored stimuli and the model response of a linear chromatic filter [Eq. (2)] fit to the measured kernel responses. Filled dots are mean accuracy at each network layer, computed across all kernels and across all 35 AlexNet training instances. (b) Proportion of kernels with different chromatic processing characteristics as a function of layer number. The full bold line gives the proportion of kernels which have no response to our stimuli. Dashed-dotted line gives the proportion of kernels with a sensitivity to intensity contrasts_{PCA} at least two times superior to their sensitivity to chromatic contrasts_{PCA} (= elevation_{PCA} > 63° in layer 1, color responsivity > 1/5). Full fine line is the proportions of kernels with sensitivity to chromatic contrast_{PCA} at least two times superior to intensity contrasts_{PCA} (= elevation_{PCA} < 27° in layer 1, color responsivity > 1/2). Dashed line is the proportion of kernels sensitive to both intensity and chromatic contrasts_{PCA} (63° > elevation_{PCA} > 27° in layer 1, 1/2 > color responsivity > 1/5). All data are the mean proportions, computed across all 35 AlexNet training instances. (c) Color responsivity [as computed from Eq. (3)] in all five convolutional layers. Data are the mean color responsivity computed across all kernels and across all 35 AlexNet training instances. (d) Proportion of kernels with color responsivities superior to different thresholds, computed across all 35 AlexNet training instances. In panels (a) and (c), error bars represent the standard deviation across the 35 training instances.

kernels sensitive to both intensity and chromatic contrasts_{P_{CA}} ($63^\circ > \text{elevation}_{\text{PCA}} > 27^\circ$ in layer 1, $1/2 > \text{color responsiveness} > 1/5$). All data are the mean proportions, computed across all 35 AlexNet training instances. From this figure, we can see that the proportion of unresponsive kernels increases dramatically in the fully connected network layers 6 and 7. On average, the proportion of unresponsive kernels increases from 16% in layer 5 to 53% in layer 6. In layer 7, 72% of units are completely unresponsive to both chromatic and achromatic stimuli. This suggests that simple stimuli, such as the gray and colored disks we employ here, are not well suited to eliciting meaningful responses from the units of fully connected layers. Comparing the responses to colored and achromatic stimuli of units in layers 6 and 7 of AlexNet would likely result in unreliable estimates of color responsiveness in these layers. Additionally, the color responsiveness cannot be computed for unresponsive units, as Eq. (3) does not allow a null sum as its denominator. Therefore, we excluded the fully connected layers 6 and 7 of AlexNet from further analyses.

Another observation one can make from this figure is that the population of color-luminance kernels increases from layers 1 to 4, while the populations of color-only and luminance-only kernels decreases. This means that while kernels in layer 1 of AlexNet are mainly color-only or luminance-only kernels, subsequent layers, up to layer 4, build more representations where achromatic and chromatic information are of similar importance. In layer 5, however, an opposite tendency seems to occur.

To investigate this point further, we plotted in Fig. 5(c) the mean value of color responsiveness across the 35 AlexNet training instances. In Fig. 5(d), we show the proportions of color responsive units for different criterion values of color responsiveness across all AlexNet training instances. However, since color responsiveness cannot be computed for unresponsive kernels, we did not consider these units.

Figure 5(c) shows the mean color responsiveness across training instances in all five convolutional layers of AlexNet. The error bars are the standard deviations of color responsiveness across all 35 instances. We can observe a significant decrease of the mean color responsiveness in layers 1 to 4 in AlexNet, from a value of 0.39 in layer 1 to a value of 0.29 in layer 4. Surprisingly, the value increases again in layer 5, with a mean color responsiveness of 0.38, almost the same as in layer 1. To further understand this, Fig. 5(d) shows the proportion of units in their respective layers having different degrees of color responsiveness, for criterion values ranging from 0 to 0.95. We can see that compared to layer 1, layers 2 to 4 show an increasing proportion of low-color-responsive kernels but, more importantly, a decreasing proportion of high-color-responsive kernels with color-responsivity values above 1/3. This color-responsivity value of 1/3 corresponds to units with equal sensitivity to achromatic and chromatic contrasts_{P_{CA}}, or tuned to an elevation_{P_{CA}} of 45° in the case of the first layer. In layer 5, however, color responsiveness increases at all levels, hence producing a mean color responsiveness very similar to the mean color responsiveness in layer 1.

Note that for kernels in layer 1, it is possible to relate their elevation_{P_{CA}} tuning, obtained after analyzing their weights directly, with their corresponding CR values (cf. subsection 2.F), and then compute the proportions of units with different

color-responsivity values in layer 1 similar as is done in Fig. 5(d). We obtained equivalent proportions between the two methods, with a mean difference of 2.4% across all color-responsivity values displayed in Fig. 5(d), thus giving further evidence that the physiologically inspired approach yields reliable results.

C. Functional Segregation

Table 1 sums up the proportions of color and luminance kernels in layers 1 and 2 of AlexNet across training instances, according to our criteria described in subsection 2.F. In each instance and layer, the half with more color kernels was selected as the color group and the other half was selected as luminance group. We see that, on average, over 70% of kernels in the color group were color kernels, and over 80% were luminance kernels in the luminance group, in both layers. Correspondingly, we obtained an average index of dissimilarity of 0.54 ± 0.27 and of 0.56 ± 0.26 for layers 1 and 2, respectively. In comparison, the 95th percentile index of dissimilarity of a random distribution only amounted up to 0.21 and 0.13. Only one training instance had an index of dissimilarity lower than these values in both layers; two others had indices of dissimilarity in the first layer lower than 0.21 and one other had an index of dissimilarity in layer 2 lower than 0.13. Maximum indices of dissimilarity were 0.90 and 0.87 for the first and second layers. The distributions of indices of dissimilarity in layers 1 and 2 had a high correlation of 0.85.

These results mean that color and luminance kernels in layers 1 and 2 of AlexNet are functionally segregated. However, the standard deviations of 0.27 and 0.26 for the index of dissimilarity indicate a good degree of variability from one training instance to another. The question is why such a segregation tends to naturally occur during training, and what it is good for. Does it lead to better accuracy? We tried to answer this question by relating the performance of AlexNet to the degree of segregation. Figure 6 shows plots of the top 1 accuracies of the 35 training instances as a function of index of dissimilarity measured in layer 1. Top 1 accuracies were computed on the 50,000 images of the validation dataset of the ImageNet 2012 contest, and correspond to the proportion of images for which exact object classes were selected as the most probable one by the model among the 1000 classes available. The right axis shows the number of successful classifications among the validation dataset corresponding to the left accuracy values. We find a significant and positive correlation ($\rho = 0.41$, $p = 0.014$). While the effect is not dramatically big, it does lead to a correct classification for about 200 more out of the 50,000 test images. We obtained a similar result for layer 2.

4. DISCUSSION

Our main results show that AlexNet essentially performs a principal components analysis on the chromatic properties of the input signals in its first layer. In subsequent layers, units respond increasingly nonlinear to color, but color responsiveness remains high throughout the convolutional layers. The segregation into two anatomically distinct streams that AlexNet imposes leads to a functional segregation into one part being

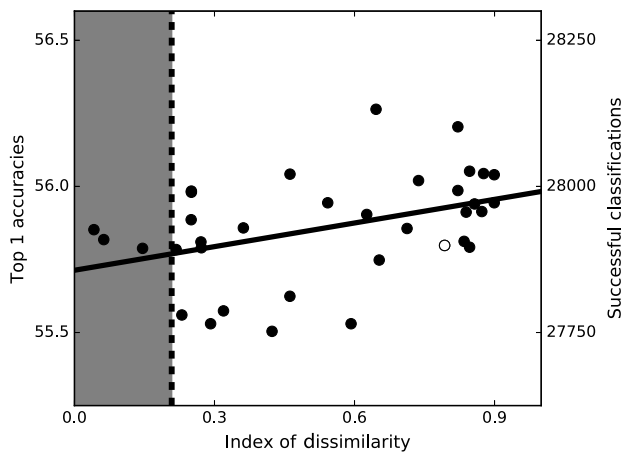


Fig. 6. Accuracy as a function of the index of dissimilarity in layer 1. Correlation of the top 1 accuracy, computed on the whole validation dataset, of the training instances of AlexNet with the degree of segregation (ID, cf. Section 2.G) in layer 1. The right axis shows the number of successful classifications among the 50,000 images of the validation dataset. The white dot stands for the AlexNet_{*β*} training instance provided with CAFFE. Kernels in the first layer of AlexNet_{*β*} are shown in Fig. 2.

mainly responsive to achromatic stimuli and the other to chromatic variations. This segregation is beneficial for object recognition performance of the network. In this section, we will discuss the similarities of these results with findings on the processing of visual information in humans and non-human primates. In particular, we will discuss how the tuning of color-responsive kernels in the first layer of AlexNet may relate to the tuning of color-responsive and linear cells of the primary visual cortex. We will compare our results with models of the primary visual cortex based on a linear analysis of the statistical properties of natural scenes. We will further discuss the possible relationship between the partial functional segregation found in the first two layers of AlexNet and the partial segregation of the chromatic and achromatic information observed in the early stages of the visual system. And we will discuss the relatively high color responsivity we found in units of the last convolutional layer of AlexNet and the high color responsivity of the most anterior visual areas of the primate occipital lobe.

A. Layer 1 and Early Visual Processing

The color space on which our networks have been trained is defined by the uncalibrated RGB images of our training dataset. Statistics of RGB images of natural scenes captured by uncalibrated cameras cannot be expected to match the statistics of natural scenes as captured by the retinal cones. As a consequence, any neural system trained on these RGB images cannot be expected to tune itself to the cardinal directions found in the LGN, but rather to statistically efficient directions in this RGB space.

The fact that the tunings of kernel maps in the first layer of AlexNet, across all training instances, preferentially fall along the axes of the RGB_{PCA} coordinates suggests that, similar to the early visual system [31,32,57], AlexNet learns to transform and decorrelate its visual inputs in the early stages of its processing to represent them more efficiently. The tuning distributions

Table 1. Proportions of Color and Luminance Kernels in the Color and Luminance Groups of the First Two Layers and Corresponding Indexes of Dissimilarity

	% Color Kernels		% Luminance Kernels		ID
	color gp	lum gp	color gp	lum gp	
Layer 1	72 ± 13	19 ± 15	28 ± 13	81 ± 15	0.54 ± 0.27
Layer 2	70 ± 13	14 ± 4	30 ± 12	86 ± 4	0.56 ± 0.26

along the axes, such as shown in Fig. 1, are remarkably similar to the distributions observed in the LGN (see Fig. 5 of [30]). For individual instances of the network [cf. Figs. 3(b) through 3(e)], the tuning resembles the more broad distributions found in V1 [40,53]. This distribution in tuning has also been observed in the first convolutional layer of the VGG-M network by Rafegas *et al.* [19], suggesting that it may be a shared characteristic of deep convolutional networks and not specific to AlexNet.

Color kernels in the first layer of AlexNet share other similarities with color-responsive linear cells in V1: they show a high degree of linearity and can be oriented or non-oriented (cf. Fig. 2) [40,43,58,59]. Several studies have reported correlates between the first layers of AlexNet and early visual processing in humans [3,4,21]. The color tuning of layer 1 kernels in AlexNet around the cardinal directions of the RGB space, analogous to the tuning of LGN and V1 cells along the cardinal directions, is further evidence that the early processing stages in both AlexNet and the primate brain share some similarities.

Another similarity of the early processing in AlexNet with the early visual cortex is the significant segregation occurring in the first two layers of AlexNet between the processing of the chromatic and achromatic information. A functional segregation between color and achromatic information is also thought to occur in the LGN and to a lesser extent in V1 (see [36,47,48] for reviews). This finding suggests that the division of labor between luminance- and color-tuned units may be an optimal strategy in segregated neural networks such as AlexNet and the human visual system, allowing for separate normalization pools for the different inputs.

B. Statistical Models of the Early Visual Cortex

Many studies have compared cells in the early visual system, particularly in V1, with basis functions resulting from statistical analyses of natural images. PCA performed on small patches extracted from natural images [32,50,51]) results in color opponency, a feature shared with V1 cells. V1 cells' spatial characteristics, however, were not matched and the chromatic tuning of the basis functions were aligned with the directions given by the principal component and did not explain the broad chromatic tuning observed in V1.

Independent component analysis (ICA) seems like a more promising approach, as several studies have reported that basis functions found via ICA performed on natural images bear a closer resemblance to V1 cells in terms of the spatial features of simple cells such as Gabor-like receptive fields, orientation tuning, and spatial frequency bandwidth [51,60–62]. All of these features are shared by kernels in the first layer of AlexNet. Phase

invariance, a property of complex cells in V1 as well as color-opponent receptive fields, could also emerge from an ICA performed on natural images [63–65]. In addition, Caywood *et al.* [52] showed that color-responsive basis functions, derived from an ICA on natural images encoded in LMS, were preferably tuned to directions close to the cardinal directions of the LGN. The basis functions obtained after performing an ICA on natural color images [51,64,65], however, did not show a color-tuning distribution as broad as that found for V1 linear cells by Lennie and colleagues, even after the addition of noise to simulate their experimental conditions [52]. More recently, Kellner *et al.* [66] showed that broad color tuning could be obtained with ICA performed on natural images after non-linear preprocessing of the images. The nonlinearities consisted of a center-surround filtering, followed by a half-wave rectification. Interestingly, this half-wave rectification is done by the same function as the nonlinearity implemented in the response of units in convolutional layers, the ReLU activation function [67]. To understand this phenomenon further and better relate it to the visual system, it would be interesting to directly compare the kernels of the first layer of AlexNet with the basis functions of an ICA performed on the training dataset, with and without this nonlinearity.

C. Mid-level Processing

AlexNet's lack of response to our color stimuli in deep layers starting after layer 5 confirms the inherent characteristic of CNNs to build more complex representations at each step of the processing. Representations in deep layers have complex spatial preferences [14,16] that simple stimuli fail to activate, thus reducing the overall responsivity of the layers. Many units in deeper layers show little response when presented with simple stimuli. Although to some degree this demonstrates a limitation of simple stimuli for understanding the processing of visual information, it also shows that biological and artificial systems share some similarities. Indeed, similar proportions of cells showing responses to color have been found in V1, V2, V3, and V4 in physiological studies [41,43,68–72], as is true for the layers 1 to 4 in AlexNet.

The mean color-responsivity curve of Fig. 5(c) shows that the color responsivity across the layers of AlexNet steadily decreases from layers 1 to 4 before suddenly increasing in layer 5. Some caution has to be taken in interpreting the magnitude of this increase, but Rafegas *et al.* [19] also found an increase, relative to layers 3 and 4, of color-responsive kernels in the fifth layer of the VGG-M network. As the VGG-M network's architecture shares with AlexNet the feature of having five convolutional layers and three fully connected ones, we cannot assess whether this increase in color responsivity is strictly related to the number of layers. One reason for this increase could be, however, the particular status of layer 5 in both cases as the input to the first fully connected layer.

It has been shown that the convolutional layers of AlexNet, trained for object recognition on the same dataset as in this study, best predict cortical activity in the occipital lobe, while fully connected layers best predict cortical activity in areas of the temporal lobe, particularly the object-responsive Inferior Temporal cortex [3,4,21]. Of course, these studies do not

demonstrate any strict equivalence between cortical areas and layers of AlexNet, but show that when comparing visual representations in AlexNet and the visual system, convolutional layers are more similar to areas in the occipital lobe than to other areas, while fully connected layers are more similar to areas in the temporal lobe. Again, these studies focus on AlexNet and thus do not allow for comparison or generalization to other CNNs' architectures.

Interestingly, studies of color processing in the human brain have shown that V4 [44,73,74] and a more anterior occipital area termed VO [74,75], exhibit an increase in color responsivity relative to previous areas. It is exactly these areas where activity can be best predicted by layer 5 of AlexNet [4,21]. This tantalizing similarity hints that there may be a functional advantage to emphasizing color information once feature analyses have reached a certain level of spatial scale or complexity, which is not present when analyzing features at lower scales or complexities.

D. Limitations

It is clear that the properties of CNNs, and thus AlexNet, depend largely on the input that is used for training. While the 1.2 million images of the 2012 ImageNet competition dataset [1,26] used for training AlexNet are useful in exploring object recognition, they have some disadvantages when studying color vision.

First of all, the images of the dataset are originally sampled from the Internet. While this makes it feasible to obtain a large number of images, there is no way to be able to control even the most basic aspects of calibration for these images. In fact, this is a feature of the data set with respect to object recognition, which should be invariant relative to any of the camera settings. For color vision, we would ideally have hyperspectral images, and at least images with calibrated camera sensors. This lack of calibration information was the major reason for us to perform all of our analyses on the RGB coordinates rather than attempting conversion to cone excitations. Therefore, the statistics of the training set of natural images are different from statistics of natural scenes as sensed by the primate cone photoreceptors [52].

A second limitation of the dataset is that all the training images were compressed using lossy JPEG compression. The JPEG algorithm compresses images by taking blocks of 8×8 pixels and then downsampling their information, both in spatial frequency and color [76]. Caywood *et al.* [52] investigated the statistical biases that JPEG compression could induce. They compared the basis functions obtained after performing an independent component analysis (ICA) on a set of natural RGB images and the same set but JPEG compressed, and found several biases induced by the JPEG compression. In particular, they found checkerboard-like blue–yellow and red–green basis functions in the JPEG condition but not in the raw images. We did not find such artifacts in our set of kernels (cf. Fig. 2). Caywood *et al.* [52] also warn that biases could come from the color encoding in JPEG compression. The JPEG downsampling of colors is performed after a conversion of the RGB values into YCbCr color-opponent space [76]. This transformation mimics the color-opponent channels in the human visual system and is actually quite similar to the

transformation from RGB to DKL coordinates [77]. We did not see any bias in the tuning of the kernels in the first layer of AlexNet towards these axes. Rather, they preferentially fell around the directions defined by the PCA performed on the color distribution of the training dataset as shown in Fig. 1. In summary, JPEG compression does not seem to substantially bias our results. Note also that color is downsampled more heavily than luminance in JPEG, and this would lead us to underestimate the role of color.

The third caveat is that the 1000 classes of objects constituting the dataset, and those which the model has been trained to differentiate, contain few color diagnostic objects. Among the 1000 classes, 352 of them are manmade instruments, 130 are dog breeds, and 67 are vehicles. In comparison, birds and food only number 59 and 27, respectively. It should also be noted that most cameras use fairly good white balance algorithms, achieving color constancy even before the stimulus enters the network and thus further reducing the demands on color processing within the network.

To escape these limitations of the image set, one would need to use a dataset of fully calibrated natural images with no JPEG compression. The ideal, of course, would be hyper-spectral natural images. However, such images are cumbersome to obtain [78] and datasets with a sufficient number of labeled object images, either in terms of classes or samples per class, do not yet exist.

Another big issue concerns the question how much our findings can be generalized to other CNNs. AlexNet was the first CNN that was highly successful in the ImageNet challenge, and for that reason it has been studied a lot in the past, including comparisons to the primate visual system [3,4,21]. Subsequent network architectures—for example VGG or Google-Net—still bear similarities to AlexNet, and in the few instances where several architectures were compared to human vision, they all behaved reasonably similar [13]. Still, detailed studies on the influence of specific network architectures on correlates of CNNs with the visual system are lacking. However, at least the VGG-M network investigated by Rafegas and colleagues [18,19] seems to behave relatively similar with respect to color. The framework and tools we have developed here will allow us to investigate these interesting issues in more detail in the future.

5. CONCLUSION

In this study, we revealed several similarities between the color processing in AlexNet and the primate visual system. First, we have shown that, similar to the early visual system, kernels in the first layer of AlexNet are linear and preferentially selective for decorrelated and statistically sensible directions in the color space of the input. Second, the distribution of tuning around these directions is comparable to the distribution of color tuning of linear and color-responsive cells in early visual processing. Third, we observed a functional segregation of achromatic and chromatic information in the early layers of AlexNet, analogous to what has been found in early visual processing. Fourth, the responsivity of the kernels to simple stimuli decreases as a function of the layer's depth, as has been reported in the visual cortex. Finally, layer 5, the last convolutional layer of AlexNet, shows a

remarkable increase in color responsivity, as is also found in the most anterior visual areas of the occipital lobe, V4 and VO.

Our results show that the electrophysiology-inspired approach of using a set of fixed stimuli with highly constrained parameters can in principle give reliable insights into the chromatic processing in a neural network, in our case an artificial one. Indeed, we found that a direct analysis of the kernel weights in the first layer of AlexNet and the physiological approach yield equivalent results when analyzing both tuning and color responsivity. This approach has clear limitations with regards to the analysis of deeper layers, where simple stimuli are less effective. Nevertheless, we find that color responsivity increases in the last convolutional layer of AlexNet. This finding is in good agreement with the previously shown increase in color responsivity found in the last convolutional layer of the VGG-M network using a direct optimization approach to discover, for each unit of the network, an optimal stimulus [19].

We also observed that there is a degree of variability in the features learned by AlexNet from one training instance to another. In particular, kernels in the first layer do not become tuned to the same directions in color space every time, and a preferential chromatic selectivity for two directions appears clearly only when considering all instances together. Furthermore, different training instances can exhibit diverse degrees of segregation between color-sensitive and non-color-sensitive kernels. This underlines the importance of taking into account several training instances of the same convolutional neural network in order to thoroughly study its general behavior.

The further exploration of CNNs and their properties presents many fascinating opportunities. It is possible to present network units with huge numbers of stimuli to explore their properties. Through training with a large set of more than 1 million natural images, it is also possible to explore exactly which combinations of features in natural images drive each network unit, even at the higher layers. And finally, we can record the activity of any subset of units at the same time. There are also shortcomings. CNNs do not yet represent any temporal dynamics of neural processing. The properties of CNNs depend largely on the input that is used for training. While the 1.2 million images used for training AlexNet are likely useful for exploring object recognition, it would be ideal to have available a tailored image set for the study of color vision.

Funding. Deutsche Forschungsgemeinschaft (DFG) (SFB TRR 135 C2).

Acknowledgment. We thank Felix Wichmann, Matthias Bethge, Guido Maiello, and Kate Storrs for insightful scientific discussion that helped to improve this study. We further thank Guido Maiello and Kate Storrs for their suggestions that helped to improve the presentation of our data.

REFERENCES

1. O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.* **115**, 211–252 (2015).

2. Y. LeCun and Y. Bengio, "Convolutional networks for images, speech, and time series," in *The Handbook of Brain Theory and Neural Networks* (MIT, 1995), p. 3361.
3. S.-M. Khaligh-Razavi and N. Kriegeskorte, "Deep supervised, but not unsupervised, models may explain it cortical representation," *PLoS Comput. Biol.* **10**, e1003915 (2014).
4. U. Güçlü and M. A. van Gerven, "Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream," *J. Neurosci.* **35**, 10005–10014 (2015).
5. J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?" *Neuron* **73**, 415–434 (2012).
6. N. K. Logothetis and D. L. Sheinberg, "Visual object recognition," *Annu. Rev. Neurosci.* **19**, 577–621 (1996).
7. J. W. Tanaka and L. M. Presnell, "Color diagnosticity in object recognition," *Percept. Psychophys.* **61**, 1140–1153 (1999).
8. J. Tanaka, D. Weiskopf, and P. Williams, "The role of color in high-level vision," *Trends Cognit. Sci.* **5**, 211–215 (2001).
9. F. A. Wichmann, L. T. Sharpe, and K. R. Gegenfurtner, "The contributions of color to recognition memory for natural scenes," *J. Exp. Psychol.* **28**, 509–520 (2002).
10. G. K. Humphrey, M. A. Goodale, L. S. Jakobson, and P. Servos, "The role of surface information in object recognition: studies of a visual form agnostic and normal subjects," *Perception* **23**, 1457–1481 (1994).
11. L. H. Wurm, G. E. Legge, L. M. Isenberg, and A. Luebker, "Color improves object recognition in normal and low vision," *J. Exp. Psychol.* **19**, 899–911 (1993).
12. K. R. Gegenfurtner and J. Rieger, "Sensory and cognitive contributions of color to the recognition of natural scenes," *Curr. Biol.* **10**, 805–808 (2000).
13. R. Geirhos, D. H. Janssen, H. H. Schütt, J. Rauber, M. Bethge, and F. A. Wichmann, "Comparing deep neural networks against humans: object recognition when the signal gets weaker," arXiv:1706.06969 (2017).
14. M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European Conference on Computer Vision* (Springer, 2014), pp. 818–833.
15. K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv:1409.1556 (2014).
16. J. Yosinski, J. Clune, A. Nguyen, T. Fuchs, and H. Lipson, "Understanding neural networks through deep visualization," arXiv:1506.06579 (2015).
17. A. Paul and S. Venkatasubramanian, "Why does deep learning work? A perspective from group theory," arXiv:1412.6621 (2014).
18. I. Rafegas, M. Vanrell, and L. A. Alexandre, "Understanding trained CNNs by indexing neuron selectivity," arXiv:1702.00382 (2017).
19. I. Rafegas and M. Vanrell, "Color representation in CNNs: parallelisms with biological vision," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017), pp. 2697–2705.
20. K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: delving deep into convolutional nets," arXiv:1405.3531 (2014).
21. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, and A. Oliva, "Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition," arXiv:1601.02970 (2016).
22. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105.
23. C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," arXiv:1312.6199 (2014).
24. K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2016), pp. 770–778.
25. G. E. Hinton, N. Srivastava, A. Krizhevsky, I. Sutskever, and R. R. Salakhutdinov, "Improving neural networks by preventing co-adaptation of feature detectors," arXiv:1207.0580 (2012).
26. J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: a large-scale hierarchical image database," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE, 2009), pp. 248–255.
27. Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: convolutional architecture for fast feature embedding," arXiv:1408.5093 (2014).
28. V. C. Smith and J. Pokorny, "Spectral sensitivity of the foveal cone photopigments between 400 and 500 nm," *Vis. Res.* **15**, 161–171 (1975).
29. J. Krauskopf, D. R. Williams, and D. W. Heeley, "Cardinal directions of color space," *Vis. Res.* **22**, 1123–1131 (1982).
30. A. M. Derrington, J. Krauskopf, and P. Lennie, "Chromatic mechanisms in lateral geniculate nucleus of macaque," *J. Physiol.* **357**, 241–265 (1984).
31. G. Buchsbaum and A. Gottschalk, "Trichromacy, opponent colours coding and optimum colour information transmission in the retina," *Proc. R. Soc. London B* **220**, 89–113 (1983).
32. D. L. Ruderman, T. W. Cronin, and C.-C. Chiao, "Statistics of cone responses to natural images: Implications for visual coding," *J. Opt. Soc. Am. A* **15**, 2036–2045 (1998).
33. K. N. Plataniotis and A. N. Venetsanopoulos, *Color Image Processing and Applications* (Springer, 2013).
34. Y.-I. Ohta, T. Kanade, and T. Sakai, "Color information for region segmentation," *Comput. Graph. Image Process.* **13**, 222–241 (1980).
35. H. Komatsu, "Mechanisms of central color vision," *Curr. Opin. Neurobiol.* **8**, 503–508 (1998).
36. K. R. Gegenfurtner, "Cortical mechanisms of colour vision," *Nat. Rev. Neurosci.* **4**, 563–572 (2003).
37. K. R. Gegenfurtner and D. C. Kiper, "Color vision," *Annu. Rev. Neurosci.* **26**, 181–206 (2003).
38. B. R. Conway, "Color vision, cones, and color-coding in the cortex," *Neuroscience* **15**, 274–290 (2009).
39. R. Shapley and M. J. Hawken, "Color in the cortex: single-and double-opponent cells," *Vis. Res.* **51**, 701–717 (2011).
40. P. Lennie, J. Krauskopf, and G. Sclar, "Chromatic mechanisms in striate cortex of macaque," *J. Neurosci.* **10**, 649–669 (1990).
41. D. C. Kiper, S. B. Fenstemaker, and K. R. Gegenfurtner, "Chromatic properties of neurons in macaque area v2," *Vis. Neurosci.* **14**, 1061–1072 (1997).
42. K. R. Gegenfurtner, D. C. Kiper, J. M. Beusmans, M. Carandini, Q. Zaidi, and J. A. Movshon, "Chromatic properties of neurons in macaque MT," *Vis. Neurosci.* **11**, 455–466 (1994).
43. E. N. Johnson, M. J. Hawken, and R. Shapley, "The spatial transformation of color in the primary visual cortex of the macaque monkey," *Nat. Neurosci.* **4**, 409–416 (2001).
44. B. R. Conway, S. Moeller, and D. Y. Tsao, "Specialized color modules in macaque extrastriate cortex," *Neuron* **56**, 560–573 (2007).
45. T. M. Sanada, T. Namima, and H. Komatsu, "Comparison of the color selectivity of macaque v4 neurons in different color spaces," *J. Neurophysiol.* **116**, 2163–2172 (2016).
46. R. L. De Valois, N. P. Cottaris, S. D. Elfar, L. E. Mahon, and J. A. Wilson, "Some transformations of color information from lateral geniculate nucleus to striate cortex," *Proc. Natl. Acad. Sci. USA* **97**, 4997–5002 (2000).
47. E. M. Callaway, "Structure and function of parallel pathways in the primate early visual system," *J. Physiol.* **566**, 13–19 (2005).
48. J. J. Nassi and E. M. Callaway, "Parallel processing strategies of the primate visual system," *Nat. Rev. Neurosci.* **10**, 360–372 (2009).
49. D. S. Massey and N. A. Denton, "The dimensions of residential segregation," *Soc. Forces* **67**, 281–315 (1988).
50. E. Provenzi, J. Delon, Y. Gousseau, and B. Mazin, "On the second order spatiochromatic structure of natural images," *Vis. Res.* **120**, 22–38 (2016).
51. T. Wachtler, T.-W. Lee, and T. J. Sejnowski, "Chromatic structure of natural scenes," *J. Opt. Soc. Am. A* **18**, 65–77 (2001).
52. M. S. Caywood, B. Willmore, and D. J. Tolhurst, "Independent components of color natural scenes resemble v1 neurons in their spatial and color tuning," *J. Neurophysiol.* **91**, 2859–2873 (2004).
53. A. Hanazawa, H. Komatsu, and I. Murakami, "Neural selectivity for hue and saturation of colour in the primary visual cortex of the monkey," *Eur. J. Neurosci.* **12**, 1753–1763 (2000).
54. G. D. Horwitz and C. A. Hass, "Nonlinear analysis of macaque v1 color tuning reveals cardinal directions for cortical color processing," *Nat. Neurosci.* **15**, 913–919 (2012).
55. D. J. Felleman and D. C. Van Essen, "Receptive field properties of neurons in area v3 of macaque monkey extrastriate cortex," *J. Neurophysiol.* **57**, 889–920 (1987).

56. M. Kusunoki, K. Moutoussis, and S. Zeki, "Effect of background colors on the tuning of color-selective cells in monkey area v4," *J. Neurophysiol.* **95**, 3047–3059 (2006).
57. Q. Zaidi, "Decorrelation of l-and m-cone signals," *J. Opt. Soc. Am. A* **14**, 3430–3431 (1997).
58. M. S. Livingstone and D. H. Hubel, "Anatomy and physiology of a color system in the primate visual cortex," *J. Neurosci.* **4**, 309–356 (1984).
59. R. Shapley and M. Hawken, "Neural mechanisms for color perception in the primary visual cortex," *Curr. Opin. Neurobiol.* **12**, 426–432 (2002).
60. J. H. Van Hateren and A. van der Schaaf, "Independent component filters of natural images compared with simple cells in primary visual cortex," *Proc. R. Soc. London B* **265**, 359–366 (1998).
61. P. O. Hoyer and A. Hyvärinen, "Independent component analysis applied to feature extraction from colour and stereo images," *Network* **11**, 191–210 (2000).
62. A. Hyvärinen and E. Oja, "Independent component analysis: algorithms and applications," *Neural Networks* **13**, 411–430 (2000).
63. A. Hyvärinen and P. Hoyer, "Emergence of phase-and shift-invariant features by decomposition of natural images into independent feature subspaces," *Neural Comput.* **12**, 1705–1720 (2000).
64. D. R. Taylor, L. H. Finkel, and G. Buchsbaum, "Color-opponent receptive fields derived from independent component analysis of natural images," *Vis. Res.* **40**, 2671–2676 (2000).
65. T.-W. Lee, T. Wachtler, and T. J. Sejnowski, "Color opponency is an efficient representation of spectral properties in natural scenes," *Vis. Res.* **42**, 2095–2103 (2002).
66. C. J. Kellner and T. Wachtler, "A distributed code for color in natural scenes derived from center-surround filtered cone signals," *Front. Psychol.* **4**, 661 (2013).
67. V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *27th International Conference on Machine Learning (ICML-10)* (2010), pp. 807–814.
68. B. M. Dow and P. Gouras, "Color and spatial specificity of single units in rhesus monkey foveal striate cortex," *J. Neurophysiol.* **36**, 79–100 (1973).
69. P. Gouras, "Opponent-colour cells in different layers of foveal striate cortex," *J. Physiol.* **238**, 583–602 (1974).
70. L. G. Thorell, R. L. de Valois, and D. G. Albrecht, "Spatial mapping of monkey VI cells with pure color and luminance stimuli," *Vis. Res.* **24**, 751–769 (1984).
71. S. Shipp and S. Zeki, "The functional organization of area V2, I: specialization across stripes and layers," *Vis. Neurosci.* **19**, 187–210 (2002).
72. H. S. Friedman, H. Zhou, and R. Heydt, "The coding of uniform colour figures in monkey visual cortex," *J. Physiol.* **548**, 593–613 (2003).
73. A. Bartels and S. Zeki, "The architecture of the colour centre in the human visual brain: new results and a review," *Eur. J. Neurosci.* **12**, 172–193 (2000).
74. K. T. Mullen, D. H. Chang, and R. F. Hess, "The selectivity of responses to red-green colour and achromatic contrast in the human visual cortex: an fMRI adaptation study," *Eur. J. Neurosci.* **42**, 2923–2933 (2015).
75. K. T. Mullen, S. O. Dumoulin, K. L. McMahon, G. I. De Zubicaray, and R. F. Hess, "Selectivity of human retinotopic visual cortex to s-cone-opponent, l/m-cone-opponent and achromatic stimulation," *Eur. J. Neurosci.* **25**, 491–502 (2007).
76. A. Skodras, C. Christopoulos, and T. Ebrahimi, "The jpeg 2000 still image compression standard," *IEEE Signal Process. Mag.* **18**(5), 36–58 (2001).
77. T. Hansen and K. R. Gegenfurtner, "Higher order color mechanisms: evidence from noise-masking experiments in cone contrast space," *J. Vis.* **13**(1):26 (2013).
78. R. Ennis, M. Toscani, F. Schiller, and K. R. Gegenfurtner, "Hyperspectral database of fruits and vegetables," *J. Opt. Soc. Am. A* **35**, B256–B266 (2018).

2. PUBLICATIONS

2.0.2 Study 2: Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks



Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks

Alban Flachot^{a,*}, Karl R. Gegenfurtner^a

^a Abteilung Allgemeine Psychologie, Giessen University, Germany

ARTICLE INFO

Keywords:

Deep learning
Object recognition
Hue selectivity
Chroma responsivity
Feature visualization

ABSTRACT

In this work, we examined the color tuning of units in the hidden layers of AlexNet, VGG-16 and VGG-19 convolutional neural networks and their relevance for the successful recognition of an object.

We first selected the patches for which the units are maximally responsive among the 1.2 M images of the ImageNet training dataset. We segmented these patches using a k-means clustering algorithm on their chromatic distribution. Then we independently varied the color of these segments, both in hue and chroma, to measure the unit's chromatic tuning.

The models exhibited properties at times similar or opposed to the known chromatic processing of biological system. We found that, similarly to the most anterior occipital visual areas in primates, the last convolutional layer exhibited high color sensitivity. We also found the gradual emergence of single to double opponent kernels. Contrary to cells in the visual system, however, these kernels were selective for hues that gradually transit from being broadly distributed in early layers, to mainly falling along the blue-orange axis in late layers. In addition, we found that the classification performance of our models varies as we change the color of our stimuli following the models' kernels properties. Performance was highest for colors the kernels maximally responded to, and images responsible for the activation of color sensitive kernels were more likely to be mis-classified as we changed their color.

These observations were shared by all three networks, thus suggesting that they are general properties of current convolutional neural networks trained for object recognition.

1. Introduction

Convolutional Neural Networks (CNNs) are the state-of-the-art for object recognition algorithms. However, little is known about their internal representations, and how these representations relate to object classification. The difficulty of the task resides in several factors, including the numerous non-linearities and the entanglement of features, such as shape and color, in hidden layers.

This study takes its place in an ongoing debate on the validity of CNNs, particularly those trained for object recognition, as models of biological neural systems. There is evidence that, similarly to CNNs, object recognition in human is mainly a feedforward process (DiCarlo, Zoccolan, & Rust, 2012), and that CNNs can be good predictors of primate brain activity (Khaligh-Razavi & Kriegeskorte, 2014; Güçlü & van Gerven, 2015; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016). However, other studies have shown differences between CNNs and the primate brain (Goodfellow, Shlens, & Szegedy, 2014; Szegedy et al., 2014; Geirhos et al., 2017), or that CNNs rely on very different cues than our

visual system (Szegedy et al., 2014; Geirhos et al., 2018). Studying how artificial neural networks learn to solve their tasks, and identifying and characterising their similarities and differences with biological brains are promising approaches. They will help us to understand *why* and *how* CNNs do solve the same tasks, and to answer the questions: what caused the two systems to behave similarly here, and different there?

The processing of visual color information, and its importance for object recognition, is a field of study that offers such an opportunity. Decades of physiological and psychophysical studies (see Komatsu, 1998; Gegenfurtner, 2003; Witzel & Gegenfurtner, 2018 for reviews) form a great basis for comparing CNNs to biological systems.

For these reasons, we focus here on the general color tuning properties of CNNs trained for object recognition. More specifically, we studied the color properties of the units constituting these CNNs, what consequences these properties may have on the classification performances of CNNs and, finally, how these properties and there consequences may relate to their counterparts in the macaque and human visual systems.

* Corresponding author.

<https://doi.org/10.1016/j.visres.2020.09.010>

Received 23 March 2020; Received in revised form 2 September 2020; Accepted 18 September 2020

Available online 18 February 2021

0042-6989/© 2021 Elsevier Ltd. All rights reserved.

To our knowledge, there are relatively few studies that address this question. In our own earlier work (Flachot & Gegenfurtner, 2018), we used a physiologically-inspired approach to study the processing of chromatic information in AlexNet (Krizhevsky, Sutskever, & Hinton, 2012). We used simple shape stimuli to analyze the chromatic tuning of kernels in a large number of training instances of AlexNet. We showed that units in early layers tended to be either color sensitive or color agnostic. Furthermore, there was a functional segregation of color sensitive and color agnostic units, probably due to the specific architecture of AlexNet, which is split into two different streams (ie graphics cards) in the early layers. Those network instances with a higher degree of segregation tended to perform better, implying that it might be advantageous to perform normalization operations separately to color and luminance components. Despite these promising results, our approach was limited to studying the early and middle layers, as the stimuli we used to probe the models were highly constrained to simple shapes, and to one CNN architecture only.

Rafegas et al. (2018) used natural images from the ImageNet (Deng et al., 2009) data set to study the color properties of VGG-M net. Using visualization methods developed in (Simonyan & Zisserman, 2014; Yosinski, Clune, Nguyen, Fuchs, & Lipson, 2015), Rafegas and colleagues examined patches for which the neural network units are maximally responsive among the 1.2 M RGB natural images of the training dataset. For each unit, they thus selected 100 patches and computed their weighted mean as an estimate of the feature that kernel would respond to best. They found that a large number of neurons were color selective in the sense that they responded much better to the colored patches than to the same patches in grayscale. An analysis of mean image patches showed a prevalence of color opponency in the early layers, while kernels in higher layers tended to respond mainly to individual hues. Their work includes a few limitations, however: (1) it is based on the assumption that the color properties of kernels equal the color properties of their corresponding mean image patches. As a consequence, color biases within the dataset might bias the results; (2) averaging across 100 images to obtain mean image patches might blur complex color tuning, particularly for late layers; (3) their study is limited to one architecture only, very similar to the one we used previously.

Engilberge, Collins, and Süsstrunk (2017) also used natural images, but they evaluated the units' responses to monochromatic images of different hues. This way, any kind of chromatic contrast was removed from the images.

Here, we try to overcome some of the limitations of the earlier work. We measure the chromatic properties of units using natural images, but we do so by varying the color of the images, either through global transformations or by modifying the color of segmented regions in these images. We not only investigate the effect of chromatic changes on the responses of individual units, but also on the recognition performance of the whole network.

2. Methods

2.1. Models and training

We used 3 networks in this study: AlexNet (Krizhevsky et al., 2012), VGG-16 and VGG-19 (Simonyan & Zisserman, 2014). We chose these networks because they are well established architectures of CNNs: more recent models are all inspired from or compared to these architectures. They also have more straightforward architectures than other networks such as Inception nets (Szegedy et al., 2015) or ResNets (He, Zhang, Ren, & Sun, 2016) making it easier to draw conclusions on their general properties.

2.1.1. Models

Deep convolutional neural networks are layered algorithms, each layer performing a set of processing operations. Like most other CNNs,

AlexNet is a feedforward system. It takes as input a $227 \times 227 \times 3$ image and outputs 1 out of 1000 category labels that the input image most likely belongs to. The first two input dimensions represent the spatial extent of the image (width and height), and the third input dimension represents the three RGB color channels. AlexNet consists of convolutional layers and fully-connected layers. A convolutional layer consists of a set of linear kernels (i.e. filters) with equally sized receptive fields (e.g. $11 \times 11 \times 3$ in the first layer) applied at equally spaced intervals, followed by half-wave rectification (ReLU) (Krizhevsky et al., 2012). This results in a two-dimensional map encoding the response of a given filter at each spatial position. The activation maps from all filters within a layer are stacked to produce the output volume of that layer, which is the input volume of the next layer. In fully-connected layers the network units get input from all units of the previous layer. The units in fully connected layers thus have receptive fields of the same size as the input image, and their activation maps can be computed through a simple multiplication of their weights with the responses of the previous units. AlexNet's architecture consists of five convolutional layers followed by three fully-connected layers. The convolutional layers 1, 2 and 5 of the AlexNet architecture are followed by max pooling, a down-sampling operation which reduces the size of the input volume along its first two dimensions by taking the maximum response of 3×3 neighboring units. Following the pooling operations, in layers 1 and 2 are two normalization layers.

We included two other networks in our study, the VGG-16 and VGG-19 networks (Simonyan & Zisserman, 2014). The main difference between AlexNet and these two is the number of convolutional layers: as their names suggest, VGG-16 and VGG-19 have 16 and 19 layers respectively. Similar to AlexNet, the last three of these layers are fully connected, and the others convolutional. As opposed to AlexNet, VGG-16 and VGG-19 do not have normalization layers. Rather the non-linearities implemented within the nets come only from the ReLU activation functions and pooling layers. In the case of VGG-16, the pooling layers are after convolutional layers 2, 4, 7, 10 and 13, while in the case of VGG-19, the pooling layers are after convolutional layers 2, 4, 8, 12 and 16. Without the normalization layers, the VGG nets have simpler architectures than AlexNet.

2.1.2. Software and dataset

All three models were pretrained by the Berkeley team and are available with the CAFFE deep learning framework (Jia et al., 2014). The models were trained on the ILSVRC 2012 dataset (Russakovsky et al., 2015). This dataset consists in over 1.2 M labeled RGB images, divided into 1000 object classes. All analyses presented in this work were scripted in python. The code used in this study is available through Github.¹

2.2. RGB_{PCA} color coordinates

Many color spaces and chromatic coordinates are commonly used in colorimetry, color science and computer graphics (Plataniotis & Venetsanopoulos, 2013). Depending on the task, some are better suited than others. The color space most suitable for our analysis is that which our CNNs are tuned towards, as well as a product of the distribution of RGB values of pixels in the training dataset. ImageNet is indeed biased in its pixels distribution mainly towards achromatic variations, but also towards bluish-orangish colors (Rafegas et al., 2018; Flachot & Gegenfurtner, 2018), which seems to be a common feature of RGB natural images (Ohta, Kanade, & Sakai, 1980). As such, a Principal Component Analysis (PCA) performed on the pixel distribution of the training dataset lead to a first Principal Component along the achromatic direction, and a second along the bluish-orangish direction. In a recent study, we showed that kernels in early layers of AlexNet also preferred

¹ https://github.com/AlbanFlachot/optimal_patch.

these directions (Flachot & Gegenfurtner, 2018). Given that these principal components are nearly identical to the optimal features found by Ohta and colleagues (Ohta et al., 1980), with a maximum relative difference of 3% per element, we used their color-axes transformation values. The resulting three color axes define a coordinate system in RGB space which we call RGB_{PCA} . The three coordinates, sorted according to the ranking of their corresponding principal components, are I_{PCA} for intensity as the achromatic dimension, $C1_{PCA}$ and $C2_{PCA}$ as the chromatic dimensions. The transformation from RGB values to RGB_{PCA} is as follows:

$$\begin{pmatrix} I_{PCA} \\ C1_{PCA} \\ C2_{PCA} \end{pmatrix} = \begin{pmatrix} 2/3 & 2/3 & 2/3 \\ 1 & 0 & -1 \\ -0.5 & 1 & -0.5 \end{pmatrix} \begin{pmatrix} R \\ G \\ B \end{pmatrix} - 0.5 \quad (1)$$

All of our analyses will be presented solely in the RGB_{PCA} coordinates, or their cylindrical representation as *Hue* and *Chroma*. To understand the conversion from cartesian to cylindrical coordinates, we may consider a point P in color space, and call V the vector associated with P . The *Hue* represents the angle between the projection of V on the chromaticity plane (orthogonal to I_{PCA}) and $C1$. The *Hue* can assume values in the interval $[0,360]$. *Chroma* is defined as the length of the projection of V onto the chromaticity plane i.e. the degree to which a color diverges from gray.

The choice of using RGB_{PCA} as the unique color coordinates for our analysis is motivated by previous studies on CNNs trained on ImageNet (Flachot & Gegenfurtner, 2018) and is as such not arbitrary. Still, since it has been used for analysis only and not training, our results should not be too dependent on this choice, as other sensible color coordinates should lead to qualitatively similar conclusions. This is particularly the case given that the color dimension most relevant for this study - *Hue* - is almost identically represented across color spaces. The main difference is that the same hue might be referenced at two different angles in two

different color spaces.

2.3. Stimulus selection

We aim at understanding the characteristics of the color properties of kernels learned in CNNs trained for object recognition, meaning that we would like to single out the dependence of a kernel’s response to the color of its input independently of any other feature. The main issue with CNNs is that the features learned individually by each of their kernels are mixtures of specific shapes and colors i.e. have specific spatial, achromatic and chromatic characteristics (Zeiler & Fergus, 2014; Simonyan & Zisserman, 2014; Yosinski et al., 2015). In particular, kernels in deep hidden layers learn features of such specific and complex spatial and achromatic properties that one needs to first match in order to study the kernels’ chromatic properties (Flachot & Gegenfurtner, 2018).

To do so, for each kernel of our 3 models, we aimed at obtaining an image patch with an “optimal” shape. By optimal, we mean that the patch should display a shape feature that we know the given kernel is highly responsive to. This was done by picking, for each kernel, the image patch within the entire training dataset for which it is most responsive, similarly to Rafegas and colleagues (Rafegas et al., 2018). Some examples are provided in Fig. 1 A.

Note that the size of the optimal patch is equal to the receptive field of the kernel it corresponds to. For example, optimal patches for the first layer kernels of the VGG-19 net are 3x3 pixels large, while optimal patches of layer 11 are 100x100 pixels large. This will matter when we will look into the impact of color changes on the classification performance of our models.

Since the selected patch is the one responsible for the maximal activation of the given kernel across over 1.2 million images, it is reasonable to assume that its shape characteristics match the kernel’s shape features.

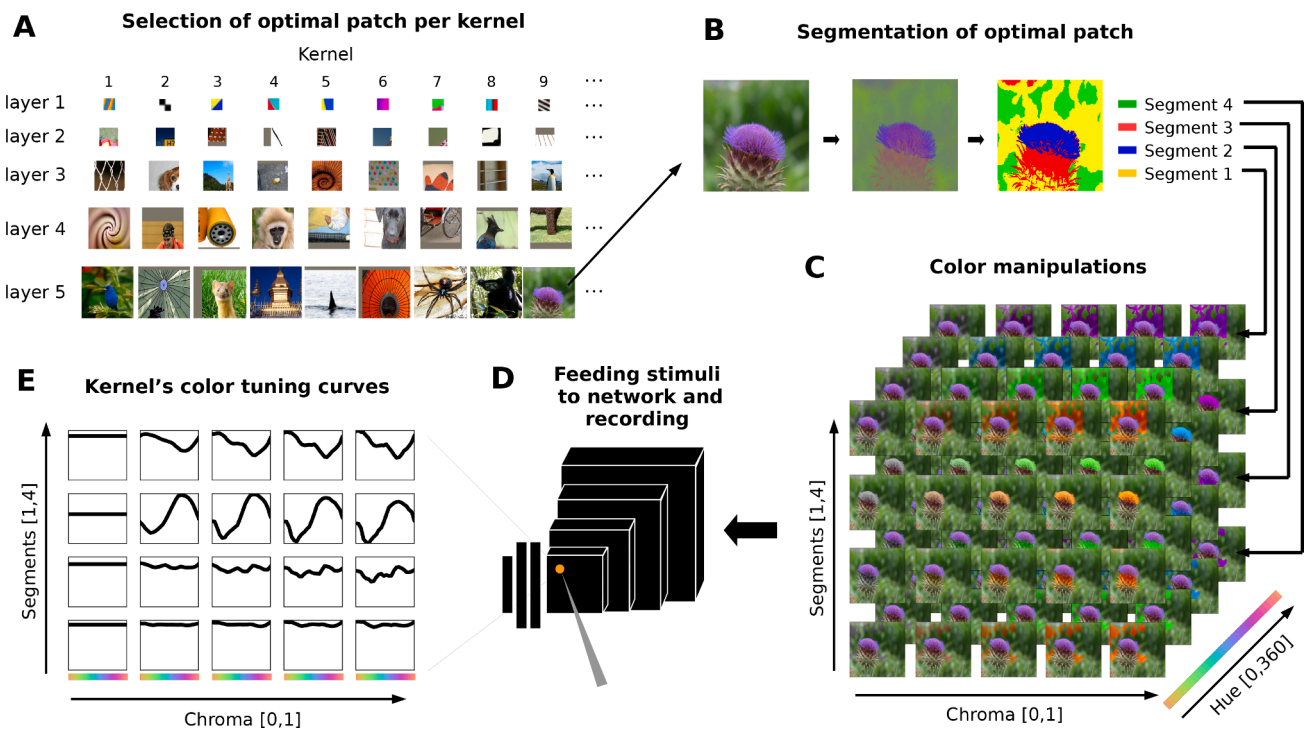


Fig. 1. Method used to extract the refined color tuning curves of kernels in the deep layers of our networks. **A:** We start by selecting, for each kernel in each layer, the *optimal patch* that results in its maximal activation across the entire training dataset; **B:** For each optimal patch, we subtract the achromatic information and apply a k-means segmentation algorithm ($K = 4$) to the chromatic distribution; **C:** We then modified independently and uniformly the color of each of the resulting segments in both hue (24 different hues, from 0 to 360°) and chroma (5 levels of C, from 0 to 1, from gray to colorful); **D:** We used each modified image as input to the model and recorded the kernel response to the modified optimal patch; **E:** We extracted the response of the kernel as a function of hue for each of the segment and values of chroma, resulting in a 4x5 tuning curves per kernel.

2.4. Color manipulation

Similar to our previous study (Flachot & Gegenfurtner, 2018), we varied the color of our stimuli in a systematic and controlled manner, and recorded model responses. As opposed to using simple stimuli like colored circles, however, here we used the more complex optimal patches as the basis.

Manipulating colors in these optimal patches lead to two issues. The first one was the RGB gamut. Since the color manipulations would be made according to the RGB_{PCA} coordinates, changes to the hue of individual pixels could lead to results outside of the RGB gamut. Note, however, that these manipulated images never get displayed on any device. These are purely virtual color coordinates and thus we do not need to be concerned with the gamut here. The "images" are simply color distributions that can take any value. The second issue was retaining the shape information within the optimal patches across our color modifications, e.g sharp color edges. We used two approaches to make sure that any change in response from our models indeed came from the color changes and their color tuning, and not their shape tuning. First, we applied a global color transformation to the whole image, by rotating all pixel colors along the intensity axis in RGB_{PCA} space, similar to (Nascimento, Albers, & Gegenfurtner, 2018). This conserves the color edges and local chromatic contrasts but modifies the hue. We applied gamut rotations for 24 angles in Hue, equally spaced by 15° . Second, we segmented the optimal image patches into different color regions and manipulated the hue of these regions separately.

In order to extract refined tuning curves from our models, we had to choose a segmentation algorithm that would segment the optimal patch in a sensible way, color wise, while retaining the shape information. This allowed us to study the tuning of kernels in different regions of the patch. We chose to use the k-means segmentation algorithm (Forsyth & Ponce, 2003) on the chromatic distribution of the pixels of the patch, after we projected the color of every pixel onto the chromaticity plane ($I = 0$). After some exploration, we fixed K at 4: the upper limit for the number of hues the kernels were selective for. We thus obtained 4 segments of our image patch based on their colors. This is illustrated in Fig. 1 B. The k-means algorithm has obvious shortcomings, such as forcing a fixed value of segments that can lead to non semantically sensible segment distinctions (see segments 1 and 4 in the example). However, it seemed to work for most image patches. We discuss this choice of segmentation algorithm in more details in the discussion section.

After identifying our 4 segments for each optimal patch, we modified the color of each segment independently. We used 24 hue values equally spaced (every 15°) within the interval $[0,360]$, for 5 values of chroma, from 0 to 1. Fig. 1 C shows an example of such manipulation for 4 hue values and all 5 chroma values. Note that at zero chroma, the segment only retains its achromatic characteristics. We then measured how the kernel responded to these changes (Cf. Fig. 1 D).

For each kernel K of our 3 networks, we thus measured 4×5 color tuning curves separately, one for each of the 4 segments at each of the 5 values of chroma (cf. Fig. 2 E). At zero chroma, the tuning curves are flat since there are no color variations as the segment was converted to grayscale.

2.5. Measures of color sensitivity

Given the richness of our set of stimuli for each kernel of all three models, we defined several measures of color sensitivity. The first, and most straightforward measure, is the normalized maximal change of a kernel's response induced by our set of color modifications. We call this measure the *overall color sensitivity* ($CS_{overall}$). A $CS_{overall}$ of 0.5 describes a kernel whose response halved, compared to its maximal response, across all tested color modifications. More formally, we define $CS_{overall}$ as:

$$CS_{overall}^K = 1 - \frac{\min(\mathbf{R}^K)}{\max(\mathbf{R}^K)}. \quad (2)$$

where K denotes a kernel and \mathbf{R} the set of measured responses.

In other words, if a usually responsive kernel was to show a null response to one of our stimuli, say for one specific gamut rotation or segment modification, then it would have the maximal value of 1 in overall color sensitivity. If its response was to stay absolutely constant across our entire set of stimuli, thus not caring about any color change, then it would have the minimum value of zero.

For each kernel K and segment S , we also applied a more restrictive measure that we called *hue selectivity* (CS_{hue}). It is defined as the normalized relative change of response induced by a hue modification, at constant chroma. More formally,

$$CS_{hue}^{K_S} = \max_C \left(1 - \frac{\min_H(\mathbf{R}_{S,C}^K)}{\max_H(\mathbf{R}_{S,C}^K)} \right). \quad (3)$$

where S denotes our set of segments, C denotes our set of chroma and H denotes our set of hues. Most often, these changes were largest at the highest levels of chroma.

We will describe a kernel as showing a major hue selectivity for segment S if its response varies by more than 50% across hues ($CS_{hue}^{K_S} > 0.5$), and a minor hue selectivity if its response varies by more than 25% ($CS_{hue}^{K_S} > 0.25$). We will call the hue eliciting the maximal activation *preferred hue* for the kernel K at segment S . We will also say that the kernel K is *hue selective* if it shows a major hue selectivity for at least one segment.

Finally, we also considered the *responsivity to chroma* (CR). Not to be confused with the minimum perceived chroma (or chroma sensitivity) used in behavioral studies (Witzel & Gegenfurtner, 2014; Bednarek & Grabowska, 2002). Here, we defined responsivity to chroma as the relative change in a kernel response to a colored segment compared to the response to the same segment in grayscale:

$$CR_S^K = 1 - \frac{R_{S,C=0}^K}{\max(\mathbf{R}_S^K)}. \quad (4)$$

As with hue selectivity, we describe a kernel chroma responsive if it showed a major chroma responsivity in at least one of its segments. In other words, if a kernel showed a response 2 times higher for a colored segment than for the grayscale version of the segment, then this kernel is chroma responsive.

3. Results

Except when explicitly stated, the results presented here are essentially shared across all three networks and thus for synthesis purposes, only the results for VGG-19 are shown. Results for the other two networks can be found in the [supplementary material](#).

3.1. Hue and chroma sensitivity

An important first step in understanding the color processing of our models is to map the degree of color sensitivity of their underlying kernels (i.e., to which degree the responses of their kernels vary with color).

Fig. 2 A shows the proportions of kernels with overall color sensitivities above various thresholds. In very early layers, overall color sensitivity is bimodally distributed. There are many kernels with little overall color sensitivity and many kernels with a high overall color sensitivity (Eq. 2). On average, across all three networks, we found that 35% of the kernels saw their response vary by less than 12.5% when we changed the color of one of the patch's segment (overall color sensitivity

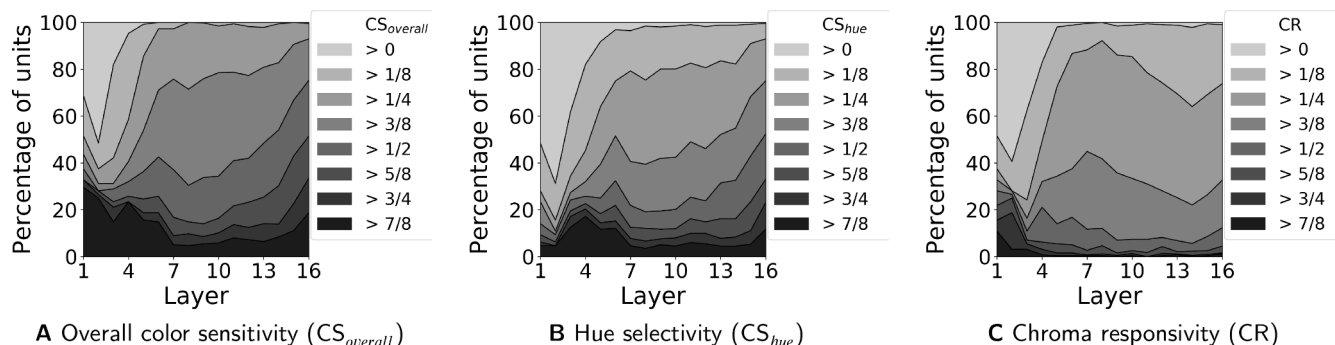


Fig. 2. Proportions of VGG-19’s kernels with different levels of A overall color sensitivity; B hue selectivity; and C chroma responsivity.

<1/8) while 37% saw their response vary by more than 87.5% (overall color sensitivity >7/8). Past the first or second layer, we found that the spectrum of hue sensitivity spreads out and kernels gradually started showing intermediate degrees of hue sensitivity. This pattern hold up to the mid convolutional layers for the VGG networks and the last convolutional layer for AlexNet. This progressive change in the distribution of color sensitivity from early to late layers is representative of the progressive entanglement of shape and color. While early kernels code almost exclusively for either the chromatic, either the achromatic information, kernels in deep layers rather code for a mixture of the two. Interestingly, we also observed an increase in overall color sensitivity for kernels in the last convolutional layers, and the highest proportion of strongly color sensitive kernels for all three networks. We found the mean color sensitivity across all three networks equal to 0.63 and the mean proportion of strongly color sensitive kernels of 63%. These results are in line with the observations previously made for individual training instances of AlexNet (Flachot & Gegenfurtner, 2018) and VGG-M (Rafegas et al., 2018). Specific to the VGG nets, we also found a secondary peak in sensitivity around the 6th convolutional layer.

The proportions of kernels with hue selectivity (cf. Eq. 3) above various thresholds are shown in Fig. 2 B. Except in very early layers, where hue selectivity was on average very low in the VGG nets, we found a similar pattern as for overall color sensitivity. In fact, we found that both measures were extremely highly correlated, with the lowest correlation being of 0.94 for the VGG-19 network.

Results for the proportions of responsivity to chroma (cf. Eq. 4) are displayed in Fig. 2 C. Similarly as for the two other measures, kernels in early layers tend to be either very responsive, either not responsive to chroma, while in later layers the spectrum of chroma responsivities is more broadly represented. Responsivity to chroma was on average, however, lower than for overall color sensitivity and hue selectivity, particularly in the deeper layers. Seemingly, kernels selective for hues tended to also be responsive to chroma. Positive correlations between the two measures were indeed found in every layers and model, the lowest correlation found being of 0.26 for the 11th layer of VGG-19, while AlexNet and early layers of the VGG-nets showed correlations greater than 0.75. On average, the correlation between the two measures is 0.62.

These results suggest the kernels in CNNs tend to be mainly sensitive to change in hues rather than changes in chroma. In other words, a segment displayed with a wrong hue is likely to induce a lower kernel response than the same segment with different saturation. This points to a special role for hue, as opposed to chroma or saturation, as has been observed in some psychophysical studies (Judd, 1970; Danilova & Mollon, 2016; Krauskopf & Gegenfurtner, 1992).

3.2. Hue tuning and color opponency

Studies in the primate visual system have also focused on the sensitivity of cells the early visual cortex towards direction in color

space (Krauskopf, Williams, & Heeley, 1982; Lennie, Krauskopf, & Sclar, 1990; Gegenfurtner, Kiper, & Fenstemaker, 1996; Gegenfurtner et al., 1994; Gegenfurtner, 2003; Komatsu, Ideura, Kaji, & Yamane, 1992; Yasuda, Banno, & Komatsu, 2009) with cells along the primate visual pathway showing different degrees of color opponency, from single opponent cells in the LGN to double opponent cells in the visual cortex (Shapley & Hawken, 2011; Conway & Livingstone, 2006). Single opponent cells decorrelate the input channels, here in terms of RGB, by combining them in a spatially uniform way. Single opponent kernels show spatially uniform color selectivity. Double opponent cells are selective for opponent colors in different spatial regions of their receptive fields. Here we define a kernel as double opponent if it is selective for opponent hues in two different segments.

Given our definition of hue sensitivity, one kernel can be selective to up to 4 hues, one for each color segment within the patch resulting from the k-means segmentation. To identify which hues the kernels are selective for each of their color segment their preferred hue (i.e the hue eliciting the kernel’s highest response), under the condition that the kernel was found to be majorly hue selective

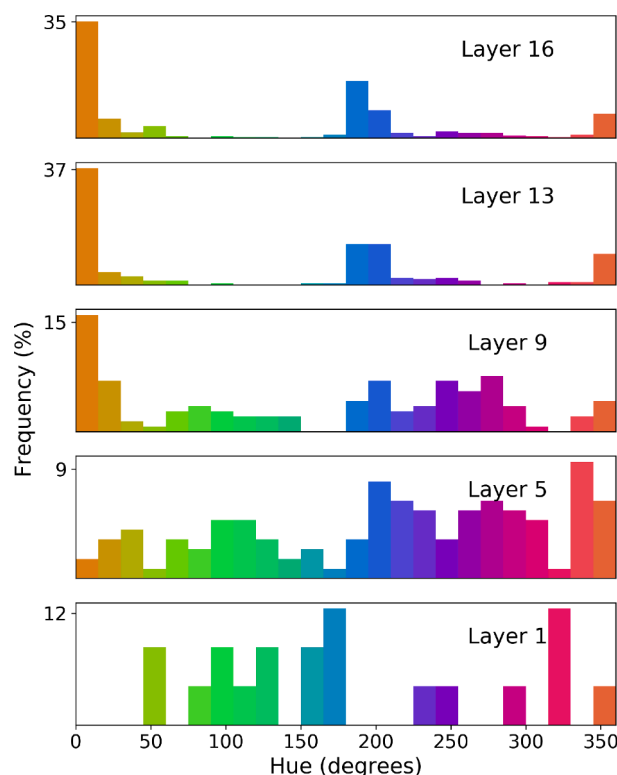


Fig. 3. Histograms of hues for which kernels are most responsive to across layers of the VGG-19 network.

($CS_{hue} > 0.5$) on that same segment. If a kernel was found to be majorly hue selective in two segments, then it could be selective for 2 hues. There is a slight chance, however, that the 2 preferred hues at 2 different segments are actually very similar. To prevent over counting based on such a bias, we considered that a kernel can be considered a selective for two hues if and only if the hues are at a minimum of 30° from one another.

Fig. 3 shows histograms of hues kernels are selective for across layers of the VGG-19 network. In the networks' early layers, the different kernels show a broad distribution of preferred hues. There are no particular color directions that are over-represented. This broad distribution becomes a bi-modal distribution in the later layers, with hue preferences falling along the blue-orange direction of 0 and 180 hue degrees. In other words, kernels in the last convolutional layers of the VGG-19 net are mostly responsive to stimuli along the C1 axis of the RGB_{PCA} coordinates. Kernels thus follow the color bias towards bluish-orangish colors of the pixels distribution of the training dataset (Rafegas et al., 2018; Flachot & Gegenfurtner, 2018). Such a bias is typically found for natural images (Nascimento, Ferreira, & Foster, 2002) due to the strong variation of natural images along the daylight locus, i.e bluish-orangish direction. It is also partially caused by the cubic nature of the RGB space (Ohta et al., 1980). Therefore, the bias is not a consequence of the particular choice of the RGB_{PCA} coordinates. Rather, it confirms that the RGB_{PCA} coordinates, because they are aligned with this preferred direction, are highly suitable to study the color processing in CNNs trained for object recognition.

This large bias, however, does not mean that VGG-19 is color-deficient (e.g green) in its last layers. While ConvNets like AlexNet and VGG nets start with a relatively low number of kernels in their first layers (Krizhevsky et al., 2012; Simonyan & Zisserman, 2014) (96 and 64 respectively), the number of kernels increases progressively to reach high values in late convolutional layers. As such, although 1.8 of kernels are only selective for the green direction in the last layer of VGG-19, for example, this small percentage of kernels still makes a significant contribution.

In order to identify single opponent and double opponent kernels, we counted the number of hues for which kernels are hue selective. If a kernel is selective for a single hue, and this hue is the preferred hue in all segments, then this kernel would be single opponent. If a kernel is selective for 2 different hues in two different segments, then it might be double opponent. Fig. 4 A shows the histograms of the number of hues for which kernels are majorly selective for (Cf 2.5). In the very early layers of the VGG nets, hue selective kernels were only selective for a

single hue. Out of these hue selective kernels, 38% of them shared the same preferred hue across all segments, showing standard deviations of less than 10° in hue angle. By definition, these kernels are thus single opponent. In deeper layers, a large proportion of hue selective kernels were also selective for only one hue, although different segments showed different preferred hues. However, it was also the case that in these layers a significant proportion of hue selective kernels was selective for 2 hues. The highest proportions were found at the 4th layer of the VGG-nets and at the 1st layer of AlexNet, layers where the receptive fields are all in the order of magnitude of 10 pixels wide. In these layers, proportions are on average of 66%. In the last layers, the proportions of kernels selective for 2 hues are on average of 28%. To figure out whether a kernel found selective for 2 hues at 2 different segments is actually double opponent, we need to compute the difference between these 2 hues. Fig. 4 B shows a histogram of these hue differences across all layers of the VGG-19 net. From this figure, it appears that in their large majority, over 73% on average, kernels selective for 2 hues are selective for hues more than 165° from one another, meaning these kernels are, indeed, double opponent.

We also describe *minor* hue selectivity as cases where the response of a kernel vary by 25%, or more, with changes in hue within a segment ($CS > 1/4$ in Fig. 2 panel B; See also methods Section 2.5). Minor hue selectivity thus *includes* hue selective segments. Most kernels in middle to late layers were found to have minor hue selectivity, with proportion superior to 60% starting from layer 5 in the VGG-nets and layer 2 in AlexNet. Out of these minor hue selective kernels, the majority were selective for at least 2 hues, with a maximum of 4 hues found for 2 kernels in each of the VGG-nets last layers. Although little, this number defines an upper boundary for the maximal number of segments in which kernels may be selective for different hues. This is also why 4 segments were set in the k-means segmentation algorithm.

So far, when a kernel was found to be hue selective within one of its segment, we focused on the hue they were maximally responsive to. But within one segment, a kernel could actually be selective to several hues, measurable by their tuning curves exhibiting auxiliary peaks. We thought interesting to quantify these auxiliary peaks using the peak detection algorithm in scipy (Virtanen et al., 2019). To be detected, a peak had a prominence of over 1/6 of the curves highest value and be above 30° hue apart from other peaks. We conducted this analysis for only hue selective kernels and segments. We considered only one tuning curve per segment (out of the 4 with non-zero chroma), the curve with the highest number of detected peaks. Fig. 5 shows some examples of tuning curves exhibiting 1, 2 or 3 peaks according to our algorithm. Fig. 6 shows the results of the analysis for hue selective kernels. We found that in early layers, hue selective kernels are exclusively selective for one hue in individual segments, with a proportion of 99% up to layer 4 in VGG-19. From layer 5 on, this proportion decreases in favour of kernels with tuning curves showing 2 peaks, 1 for their primary hue and another for their secondary hue, reaching average proportions of 39% in the last layers. If applicable, we computed the angle difference between the secondary and primary hues. In Fig. 7 we show a histogram of these angle differences. We find that over 70% of these are at $180^\circ \pm 15^\circ$ away from the preferred hue.

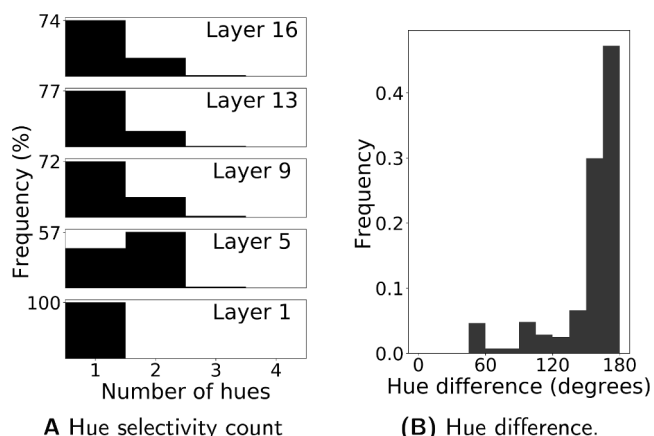


Fig. 4. A: Histograms of the number of hues for which kernel are selective (see Eq. 3). B: For kernels selective for 2 different hues at different segments: histogram of the hue difference between the 2 hues. Except for early layers of the VGG nets, a significant proportion of the hue selective kernels are selective for two hues in two different segments. For the majority of these, the two hues are approximately opponent, suggesting that these kernels are double opponent.

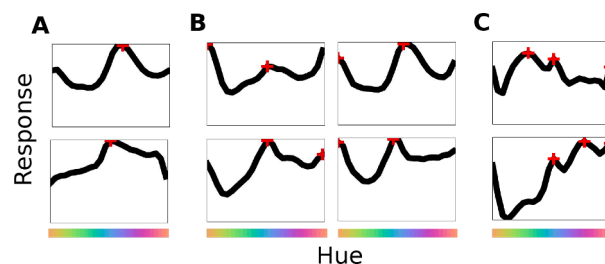


Fig. 5. Example of tuning curves displaying A 1 peak, B 2 peaks and C 3 peaks according to our algorithm (see Section 3).

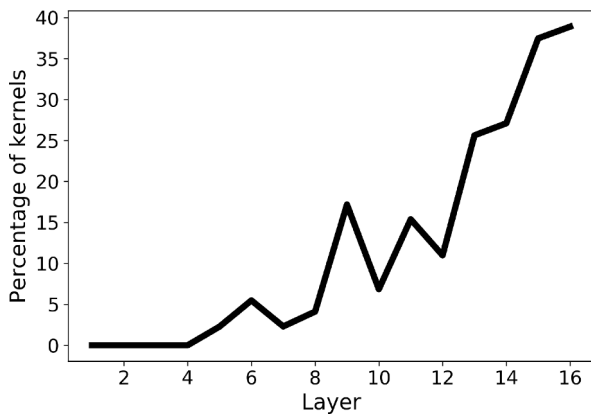


Fig. 6. Proportion of VGG-19's hue selective kernels showing a secondary peak in their tuning curves.

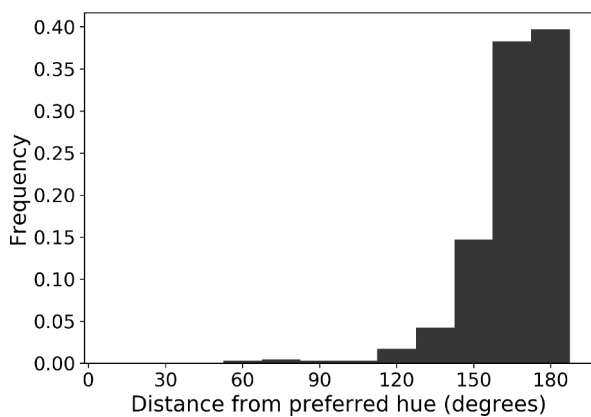


Fig. 7. Histograms of the distance, in hue, of the secondary with respect to the primary hue for the VGG-19 networks. We can see that most of the peaks fall 180° from the primary hue, meaning that kernels can be secondarily selective for hues 180° from the main hue in the same segment.

This characteristic of late kernels goes beyond simple and double opponency. In fact, it rather resembles the behaviour of complex cells found in the primary cortex of cats and macaques (Spitzer & Hochstein, 1985; Spitzer & Hochstein, 1985; Lennie et al., 1990). A complex cell response is modeled by taking in the signal of linearly summing elements distributed throughout its receptive field, performing a half-wave rectification on each of them before combining them linearly (Spitzer & Hochstein, 1985; Lennie et al., 1990). In a similar fashion, a kernel of deep layers of CNNs linearly combines the outputs of kernels in the previous layer, each output resulting from a half-wave rectified weighted sum of inputs (ReLU, cf. Section 2.1 of this manuscript) (Krizhevsky et al., 2012).

Once again, similar tendencies were found across the three examined networks, suggesting a general property of convolutional neural networks.

In summary, in the last layer of our models, we found on average that 50 % of the kernels were selective for one hue in at least one of their 4 segments. 39% of these showed, within the same segment, a secondary selectivity for another hue. This other hue was, in over 70 % of cases, around 180° from the preferred and optimal hue. This means that, in the eventuality that this segment carried semantically relevant information for object classification, the classification could still be successful if the object had the optimal hue or its opponent hue, and unsuccessful if the object had a hue in between.

3.3. Hue tuning and classification performances

The color tuning of the kernels do not say how their color characteristics impact the classification performance of the whole network. We therefore obtained color tuning curves for the whole network and compared them to the tuning curves measured in the previous section.

We thus showed the networks our color-modified set of images, and recorded the classification results of the models. We first focus on the simple case of the global transformations of the whole images colors, i.e via a rotation along the achromatic axis or turning in black and white.

We looked at how the model performs as we modify the color of the images more and more. Fig. 8 shows the performance of VGG-19 as we modify the original colors of the stimuli by applying a rotation around the achromatic axis, in color space, of the their pixel distribution (black). At zero (or 360) degrees we have thus the accuracy obtained for original images. In gray is plotted the classification performance for the same images but converted to grayscale. We found that converting the stimuli to grayscale had already a significant impact on the classification performance. We observed a drop in performance from 76.5% to 59.5% for VGG-19. Across all three networks, we found a relative decrease of 25% in performance, 33% for AlexNet. However, we found an even bigger effect of hue modifications. The models reached even lower performance for large rotation angles, between 60 and 285° off the original colors. On average, we found that models showed a relative decrease in performance up to 31.6%, and 42% for AlexNet. This means that showing the wrong color to the network can be more detrimental than showing no color at all.

After analysing the change in classification induced by the global transformation, we looked at the change in classification induced by the local transformations of the segments. First, we looked at the proportion of images which were originally classified correctly then misclassified at least once as we modified the color of the image segments (Fig. 9). The black curve (in Fig. 9) corresponds to images responsible for the activation of color sensitive kernels. The gray curve corresponds to images responsible for the activation of non color sensitive kernels. The red curve stands for both kinds combined. We found that in all three cases the proportions increased as we used images related to kernels in higher levels. This is not surprising, as the size of the color modifications, in terms of pixels, increases as well, as described in the Section 2.3. The modification size is indeed a function of the patch size, itself equal to the receptive field of the kernel considered. We can also see that starting from the mid level layers, the black line is above the gray and red lines, meaning that images including the optimal patch for overall color sensitive kernels are more likely to be misclassified as we modify their color.

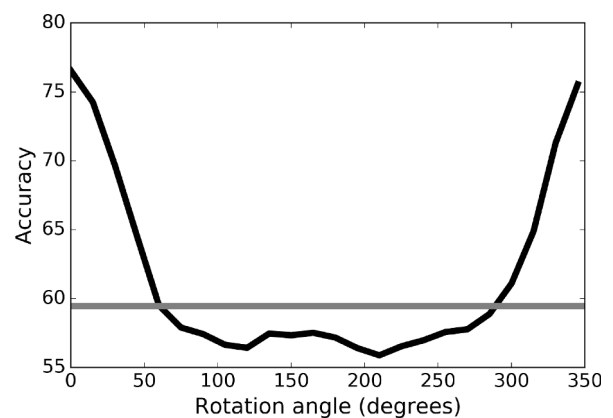


Fig. 8. VGG-19 classification performance as a function of hue angle rotation, relative to the original colors. In black: performance of VGG-19 as we modify images from the original colors by applying a rotation around the achromatic axis of color space. Gray horizontal line: performance of the model for the images converted to grayscale.

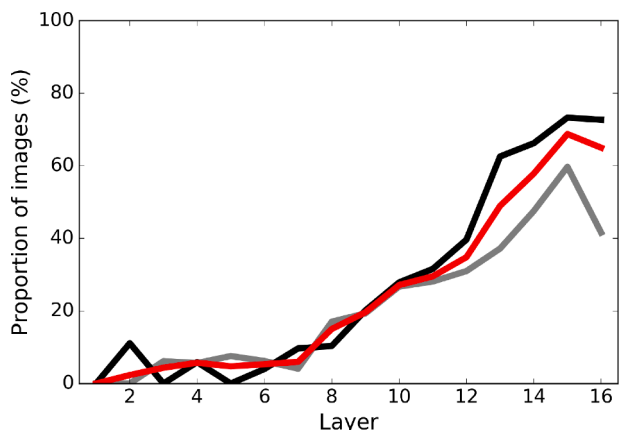


Fig. 9. The proportion of correctly classified images which are misclassified at least once when the color of a segment is modified in color (black) and non-color (grey) sensitive kernels, and in any kernel (red).

For these images, color plays a higher diagnostic role for images including the optimal patch of non color sensitive kernel. Across all three networks, we found that a proportion of 65.7% images were misclassified at least once when they included the optimal patch of kernels in the last layer, 73.1% when these kernels were color sensitive and 45.2% when these kernels were non color sensitive.

To be able to obtain a curve of classification as a function of hue, we cannot just consider the hue per se, as kernels were selective for different hues. Instead, we need to consider the degree of hue rotation with respect to the preferred hue of the corresponding kernel at this segment of the stimulus. In other words, if a kernel was mainly selective for blue at this particular segment, we started by showing to the model the corresponding image with the blue segment, then showed the images successively with hues progressively going away from the preferred blue. Since the classification is binary, we averaged across segments to obtain tuning curves.

The color manipulations with the highest impact on classification were the ones based on the optimal patches in the highest layers, as opposed to the patches found in early layers being of smaller sizes. To measure tuning curves for classification as a function of hue, we thus

considered the optimal stimuli corresponding to the kernels in the last convolutional layers: 256 original images for AlexNet and 512 for the VGG networks, one for each kernel.

Figs. 10 A and B show the result of this procedure for VGG-19. **Fig. 10 A** shows the classification accuracy of the model averaged across chroma, for different values of hue selectivity: in full gray is the classification accuracy as a function of hue angle away from the preferred hue for images corresponding to non hue selective kernels. In full black line, the equivalent but for hue selective kernels. Dotted lines correspond to the accuracy of the model for the original images. In red is the mean accuracy across all segments. **Fig. 10 B**, on the other hand, shows the classification accuracy of the model averaged across hue selectivity, for different values of chroma. Full lines are obtained for different chroma, from 0 to 1. The lightest, straight line corresponds thus to color manipulation with a chroma of 0, meaning images with achromatic segments and no variation in hue.

Several conclusions follow from **Fig. 10**. First, for all conditions, the maximal accuracies were obtained for the preferred hue, at 0° on the graph. This indicates that for a given chroma, the preferred hue was indeed the optimal hue for classification. Second, image classification, including the optimal patch of hue selective kernels, varies more with color modifications than for images including the optimal patch of non hue selective kernels. For the former, color played thus a more important role. On average across models, we observed a 27.4% relative decrease in accuracy for images including the optimal patch of hue selective kernel. In the non hue selective case, the relative decrease in accuracy is limited to 4.7% on average. Overall, as shown in red, we observed on average a relative decrease of 8.9%. Third, we see that the magnitude of the change in classification performance increased as we increased the chroma, from a relative difference of 4.3% to 14% for chromas of 0.25 and 1 respectively (**Fig. 10 B**). Note that the maximal average accuracy decreased with chroma as well, with the same order of magnitude as the variation within chroma. Lastly, and perhaps most interestingly, we observed that the accuracy dropped gradually for hue angles that are 0 to 90° apart from the preferred hues (0° in **Fig. 10**). However, the accuracy increased for angles roughly above 90° , particularly in the case of the VGG networks, to peak again around 180° . This was a robust effect across images, hue selectivities, chroma responsivities and networks.

This peculiar secondary peak cannot be a direct consequence of the opponency of kernels, in the classical definition of it. Single and double

Hue and chroma sensitivity in CNNs

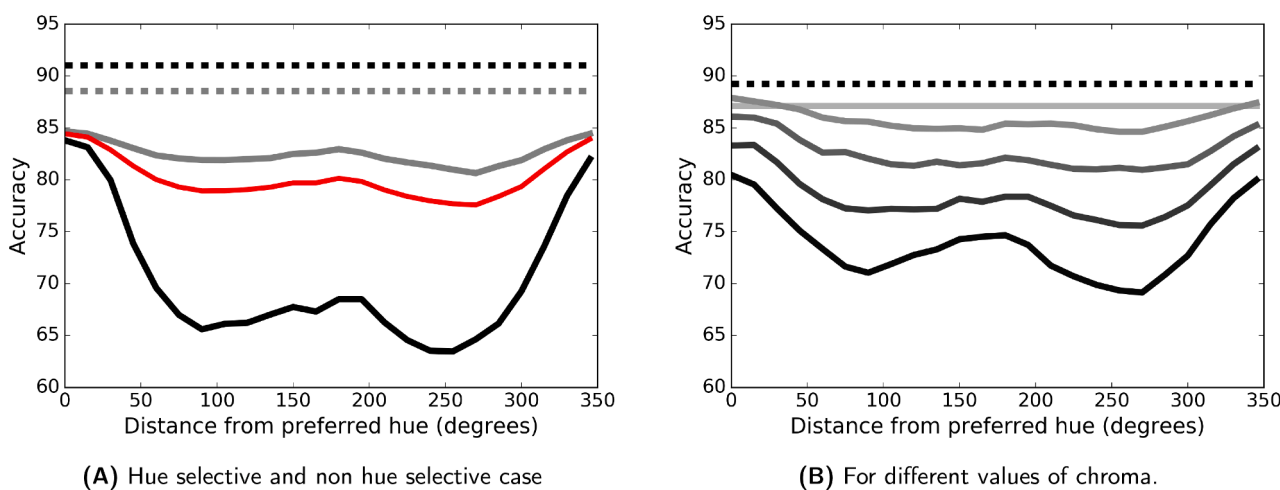


Fig. 10. VGG-19 network classification performance as a function of the distance, in hues, to the preferred hue. **A:** Results for different levels of hue selectivity, averaged across chroma. In full gray, results obtained for images including the optimal patches of non hue selective kernels. In full black line, for images including the optimal patches of hue selective kernels. Dotted lines are the classification performance for the original images in the 2 corresponding groups. In red is the mean accuracy across all kernels. **B:** Results for different chroma. Five levels of chroma, from 0 to 1, displayed from light gray to black. Dotted line correspond to the mean accuracy of the model for the original images.

opponent cells, as defined by Shapley and Hawken (2011), respond positively to specific hues with a specific spatial configuration, and respond negatively to inverse hues with the same spatial arrangement. Here, however, our nets correctly classify one image, whether one segment exhibits a hue or its inverse, but not those in between. This result is similar to the observations made in 7, where we found that hue selective kernels in late convolutional layers were often selective to two hues *within* one segment, the preferred hue and a secondary hue, that were in most cases around 180° apart from one another. The secondary peak in accuracy could possibly be directly related to the secondary peak sometimes present in the kernels tuning curves.

4. Discussion

We have made several observations in this work about the chromatic processing of kernels in 3 well established CNNs, about the color physiology of these nets so to speak. We also used a psychophysical-like approach to investigate the importance of color for their successful recognition of objects.

Understanding how the color properties of the CNNs trained here relate to what we know of the macaque's visual systems would be useful for assessing the extent to which CNNs can be accurate models of biological neural systems. It would, in turn, give us ground for extending our understanding of *how* and *why* these biological neural systems - and in particular the macaque's visual system - get to organize themselves. Color vision is particularly suitable for comparing both biological and artificial systems due to its long list of physiological and psychophysical studies performed over the last decades.

4.1. Comparison with the physiology of color processing in the primate visual system

On many occasions do both biological and artificial systems share similarities. In particular, the kernel properties in the early layers of the CNNs tend to be comparable to the properties of cells in the early visual system of the primate and human brains. Similarly to cells in the Lateral Geniculate Nucleus (LGN) and to a lower extent in V1 (Gegenfurtner, 2003; Callaway, 2005; Nassi & Callaway, 2009; Krauskopf et al., 1982), kernels in early layers show a clear separation between highly color sensitive kernels and non color sensitive kernels (cf. Fig. 2). Color sensitive kernels in these layers show a simple hue tuning (cf. Fig. 6) similarly to cells in the LGN and simple cells in the primary visual cortex (Krauskopf et al., 1982; Lennie et al., 1990).

Similarities between artificial and biological systems can also be identified in extra-striate cortical areas. Just like kernels in mid and late layers of our networks, cells from cortical areas from V2 onwards show complex color tuning and can be responsive to both achromatic and chromatic stimuli (Conway, 2009; Shapley & Hawken, 2002; Komatsu, 1998; Gegenfurtner et al., 1996; Gegenfurtner et al., 1994; Zaidi & Conway, 2019). Functional imaging shows that global color sensitivity varies considerably between different visual cortical areas (Conway & Tsao, 2006). Similar to the CNNs studied here, early visual areas such as the LGN and V1, as well as late occipital areas, such as V4 and VO, show an overall higher color selectivity compared to mid occipital areas (Mullen, Chang, & Hess, 2015; Mullen, Dumoulin, McMahon, De Zubicaray, & Hess, 2007). Neural regions of high color responsiveness have also been found in more anterior areas such as IT (Zaidi & Conway, 2019).

Another notable similarity between the CNNs studied here and biological visual systems is the emergence of different degrees of color opponency, from *single* to *double opponent* kernels, just like the *single* and *double opponent* cells found in the early visual system of the monkey (Lennie et al., 1990; Shapley & Hawken, 2011; Conway, Hubel, & Livingstone, 2002). Kernels exhibiting non-linear color response, likened to the color response of *complex cells* of the macaque visual systems (Lennie et al., 1990; Kiper, Fenstemaker, & Gegenfurtner, 1997), were also found in mid to late layers of our models.

While we found many similarities between CNNs and the macaque's visual system, massive differences can also be observed. In terms of hue tuning, indeed, striking differences can be found between biological and artificial brains in mid to late processing levels. On the one hand, we found here that CNN's kernels progressively become preferentially selective for two specific hues, along the axis of the first chromatic principal component of the input images. In the primate's visual system on the other hand, cells in the LGN preferentially respond to two "cardinal directions" of color space. Color sensitive cells in the primary visual cortex are selective for a much broader range of hues (Lennie et al., 1990). In V1 and later areas, they do not show as a whole any preference for particular hue directions, although each individual cell might be highly hue specific (Zaidi & Conway, 2019; Gegenfurtner et al., 1994; Gegenfurtner et al., 1996). Cells of the primate visual system show a transition from being selective for a narrow set of hues to a broad set, while it is just the opposite in CNNs.

4.2. Comparison to psychophysical studies in humans

There are many psychophysical studies investigating the role of color for recognition (for a review, see Bramão, Reis, Petersson, & Faisca, 2011; Witzel & Gegenfurtner, 2018). Color enhances the recognition of objects and scenes by reducing reaction times needed for recognition (Wurm, Legge, Isenberg, & Luebker, 1993; Gegenfurtner & Rieger, 2000) and increasing recognition accuracy (Gegenfurtner & Rieger, 2000). This is especially true for objects so called color diagnostic, i.e., objects with a redundant color (Tanaka & Presnell, 1999; Tanaka, Weiskopf, & Williams, 2001; Nagai & Yokosawa, 2003; Wichmann, Sharpe, & Gegenfurtner, 2002; Oliva & Schyns, 2000). Same as for humans, networks trained on colored images also use color to perform better at recognizing objects. Figs. 8 indeed show that performance is significantly higher for the original colored images than for their grey-scale counterparts.

Not only is color helpful, but previous work showed that incorrect colors also hinder humans recognition performance. Oliva and colleagues (Oliva & Schyns, 2000) had an extra condition where they modified the color of the images of natural scenes by swapping the projections of their pixels on the CIELab color axes. They found that observers took a longer time to recognize images of scenes with swapped colors than achromatic images of the same scenes. Since these results are about scene perception, they do not allow a direct comparison with the observations made in this study. They do nonetheless show interesting similarities with some of our results: that kernels show a lower response to the wrong hues than to black and white stimuli, or that the classification performance of our models are indeed lower for stimuli with the wrong colors than for black and white stimuli (see Fig. 8). It remains an open question, however, whether the secondary peak in performance at around 180° off the kernels preferred hues in CNN (Fig. 10) would reoccur for human observers.

4.3. Potential causes for similarities and differences

The reasons for these similarities and differences remain unclear. Nevertheless, some possible explanations are at hand. Some of these are related to the general similarities and differences between CNNs and biological vision, other more specific to color. These may arise from differences in the input, the computational architecture, or the task (the output). One obvious similarity is that both systems devote a significant part of their resources to processing color information. The main reason would be that both systems try to make sense of the "world" they see in order to solve their "task", and both this "world" and the "task" gives an important role to color. This is only possible because the CNNs studied here are trained on naturalistic color images for object recognition, a task for which color is highly relevant. Nevertheless, there are many important differences between the two systems' inputs and tasks which could explain the differences between both systems in the processing of

color. The inputs have different constraints. ImageNet is composed of presumably white balanced, static RGB encoded images, while humans deal with the much more ambiguous retinal images that are constantly changing. While the sole task of our models was to solve object recognition for a few image classes, humans and macaques' behaviour is dictated by constantly changing needs, from survival to reproduction, to which object recognition contributes as one of many subtasks. The hierarchical and feedforward processing of CNNs and primate visual system could at least partially account for the progressive transition from the separation of achromatic and chromatic information at the early stage of processing, to a progressive entanglement in later stages, found for both systems. Still, feedback connections, so numerous in humans and primates brains, are missing in CNNs. A feedback loop is implemented during training when updating the CNNs parameters, but it is no longer part of the recognition process after the models are trained. The supervised nature of the training procedure and its implementation is possibly one reason for the difference in hue tuning between CNNs and the primate visual system. The now classical gradient descent commonly implemented consists in training steps where the CNNs weights are updated in a cascade fashion, from top to bottom. Thus, kernel weights in the last layer are first modified to match the desired output, after which weights of the penultimate layer, and so on. As a consequence, kernels of the last layers will be more specialized, more narrowly matching the dataset's color distribution than the noisier and more universal kernels of the first layer.

4.4. Limitations

We discuss here the limitations of our method, and in particular on the use of the k-means segmentation algorithm. The purpose of using the segmentation algorithm was to modify the color of segments of the kernels' optimal patches in order to finely study the kernels' color tuning (Fig. 1B). The segments should sensibly follow the color distribution of the patch while conserving the semantic information of the patch.

Segmentation algorithms are a field of research in itself, of which we will not pretend to have an exhaustive knowledge. We looked into several kinds of algorithms, which could be divided into algorithms based on semantics or low-level features.

A very accurate semantic segmentation, capable of segmenting the object from its surrounding sounds like it should be optimal. Since 2012 and the advent of CNNs, as with many other complex visual tasks, semantic segmentation has improved considerably. To improve segmentation, previous work used complex architectures (Jégou, Drozdal, Vazquez, Romero, & Bengio, 2017; Long, Shelhamer, & Darrell, 2015), better learning strategies (Papandreou, Chen, Murphy, & Yuille, 2015) or data augmentation (Zhu et al., 2019). The main limitation with semantic segmentation, however, is that it requires *learning*, thus a dataset to learn from and with a precise ground truth to compare to the model's output. Although several of these datasets do exist (Everingham, Van Gool, Williams, Winn, & Zisserman, 2012; Brostow, Fauqueur, & Cipolla, 2008; Nathan Silberman, Derek Hoiem, & Fergus, 2012), none of them unfortunately include a number of semantic classes comparable to 1000 object classes of ILSVRC 2012, the dataset used here, and they do not necessarily coincide with the nature of ILSVRC classes. Some of them, such as CamVid (Brostow et al., 2008), are for the purpose of automatic driving and present essentially street views only. The PASCAL dataset (Everingham et al., 2012) has the interesting feature of having datasets for both object classifications and segmentation with the same object classes. These classes, however, are very few (10 to 20 classes), very broad and mainly man-made. None of these classes were classified as "color-diagnostic" by Tanaka and colleagues (Tanaka & Presnell, 1999), and thus inappropriate to study the importance of color for object recognition. Color indeed contributes very little for the recognition accuracy of these models when tested on these classes (Geirhos et al., 2017). In addition, these classes transfer poorly to the broader ILSVRC

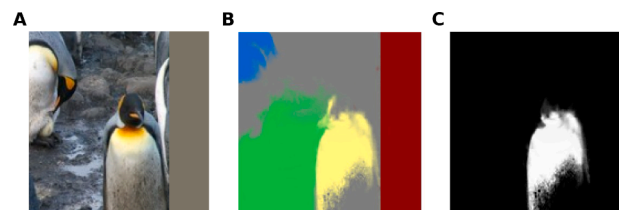


Fig. 11. Figure of an optimal image patch for which current segmentation algorithms failed to serve our purposes. In A, the optimal image patch found for kernel 98 of AlexNet's layer 5. It shows a penguin, an animal the segmentation algorithm has never seen before. As a consequence, the segmentation algorithm outputs incorrect segments as shown in B and thus an unusable object segment as shown in C.

2012 dataset and would require additional training. As an example, Fig. 11 shows a failed segmentation of one optimal image patch (a penguin) obtained with a recently developed soft-segmentation semantic algorithm (Aksoy, Oh, Paris, Pollefeys, & Matusik, 2018) and trained on the PASCAL dataset. The segmentation algorithm failed to recognise the penguin and thus gave an incorrect and unusable set of segments.

Thus, we relied on low feature based algorithms, and decided to use the k-means algorithm (Forsyth & Ponce, 2003). As our research interests were in color properties, we performed the clustering on the chromatic distribution of pixels rather than achromatic information. The main drawback of k-means algorithm, aside from the fact that it bears no semantic knowledge, is the set number of segments one needs to define a priori. We chose the number 4, as it was found to be the upper bound for the number of hues in kernels with minor hue selectivity (see Section 3.2). This number, however, is unlikely the correct number of segments for all optimal image patches. As a consequence, we might find areas of the image patch which would be unnecessarily divided, such as segments 1 and 4 in our example Fig. 1 B. Given our purposes and analysis, however, we have several reasons to believe this is not an issue. If the extra segment(s) found are so nonsensical that they bear no significance to the kernel itself, any color modification would have no consequence on the kernels response. In addition, we always considered the maximal value across segments in our measures of color sensitivity for the models' kernels. Finally, we accounted for this issue when we counted the number of preferred hues for which a kernel would be color selective. We did so by discounting a hue if its hue angle is too close (30° or less) to the preferred hue found for at another segment (Cf. Section 3 and Fig. 4). Considering all these points, the consequences of the k-means algorithm shortcomings should bear no, if not quantitative at least qualitative, significance in our results.

5. Conclusion

In this study, we looked into the color tuning of kernels in deep convolutional neural networks trained for object recognition, and its influence on the models' performance. The obscurity, non-linearity and complexity of these networks makes it a difficult task. We thus came up with a complex but complete approach, which allowed us to come up with stimuli tailored to study the color properties of each and everyone of the convolutional kernels of our three models. Thanks to this, we were able to extract the amount and nature of hues for which kernels were mainly responsive to. We show that the complexity of the color tuning of kernels in higher layers gets progressively higher, either because they are selective for several hues at distinct position, or because they show non linear tuning at the same position. We also show that most kernels are majorly responsive to the same hue directions in color space. This direction corresponds to the second principal component of the color distribution of pixels in the training dataset where the first component corresponds to the achromatic direction in color space. Finally, we were able to relate the color tuning of the models' kernels with their

performance by looking at the proportion of successful classification despite the color changes. We found that color had a significant importance for the object recognition by CNNs, and that the proportion of successful classifications is highest for the colors the kernels maximally responded to. These findings support in part the applicability of CNNs trained for object recognition as models for the primate's ventral stream. Significant discrepancies between the two systems were nevertheless made obvious, particularly with respect to the hue tuning of kernels in late convolutional layers versus the hue tuning of cells in late occipital areas. These differences can however serve as a basis for developing CNNs even further and, in doing so, lead to an expanded understanding of how biological systems get to organize themselves.

CRedit authorship contribution statement

Alban Flachot: Conceptualization, Formal analysis, Methodology, Software, Writing - original draft. **Karl R. Gegenfurtner:** Conceptualization, Formal analysis, Resources, Software, Writing - review & editing.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This work was funded by the DFG (German Research Foundation) as part of the SFB TRR 135: Cardinal Mechanisms of Perception - project number 222641018.

We thank all our friends and colleagues within our team and lab for their support and insightful scientific discussion that helped improving this study, namely Christoph Witzel, Guido Maiello, Florian Bayer, Matteo Valsecchi, Robert Ennis, Arash Akbarinia, Raquel Gil, Matteo Toscani, Thorsten Hansen, Kate Storrs, Anke-Marit Albers, Philipp Schmidt. In particular, we would like to thank Yaniv Morgenstern for helping with improving the readability of our data. We would also like to thank Felix A. Wichmann, Heiko H. Schütt, Matthias Kümmerer and Tom Wallis for their useful feedback.

Appendix A. Supplementary data

Supplementary data associated with this article can be found, in the online version, at <https://doi.org/10.1016/j.visres.2020.09.010>.

References

- Aksoy, Y., Oh, T. H., Paris, S., Pollefeys, M., & Matusik, W. (2018). Semantic segmentation. *ACM Transactions on Graphics (Proc. SIGGRAPH)*, 37, 72:1–72:13.
- Bednarek, D. B., & Grabowska, A. (2002). Luminance and chromatic contrast sensitivity in dyslexia: the magnocellular deficit hypothesis revisited. *Neuroreport*, 13, 2521–2525.
- Bramão, I., Reis, A., Petersson, K. M., & Faisca, L. (2011). The role of color information on object recognition: A review and meta-analysis. *Acta psychologica*, 138, 244–253.
- Brostow, G. J., Fauqueur, J., & Cipolla, R. (2008). Semantic object classes in video: A high-definition ground truth database. *Pattern Recognition Letters* xx.
- Callaway, E. M. (2005). Structure and function of parallel pathways in the primate early visual system. *The Journal of Physiology*, 566, 13–19.
- Cichy, R.M., Khosla, A., Pantazis, D., Torralba, A., Oliva, A., 2016. Deep neural networks predict hierarchical spatio-temporal cortical dynamics of human visual object recognition. arXiv preprint arXiv:1601.02970.
- Conway, B. R. (2009). Color vision, cones, and color-coding in the cortex. *The Neuroscientist*, 15, 274–290.
- Conway, B. R., Hubel, D. H., & Livingstone, M. S. (2002). Color contrast in macaque v1. *Cerebral Cortex*, 12, 915–925.
- Conway, B. R., & Livingstone, M. S. (2006). Spatial and temporal properties of cone signals in alert macaque primary visual cortex. *Journal of Neuroscience*, 26, 10826–10846.
- Conway, B. R., & Tsao, D. Y. (2006). Color architecture in alert macaque cortex revealed by fMRI. *Cerebral Cortex*, 16, 1604–1613.
- Danilova, M., & Mollon, J. (2016). Superior discrimination for hue than for saturation and an explanation in terms of correlated neural noise. *Proceedings of the Royal Society B: Biological Sciences*, 283, 20160164.
- Deng, J., Dong, W., Socher, R., Li, L.J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In: *Computer Vision and Pattern Recognition*, 2009. CVPR 2009. IEEE Conference on, IEEE. pp. 248–255.
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73, 415–434.
- Engilberge, M., Collins, E., & Süsstrunk, S. (2017). Color representation in deep neural networks. In *International Conference on Image Processing (ICIP) 2017*. IEEE.
- Everingham, M., Van Gool, L., Williams, C.K.I., Winn, J., & Zisserman, A., (2012). The PASCAL Visual Object Classes Challenge 2012 (VOC2012) Results. <http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html>.
- Flachot, A., & Gegenfurtner, K. R. (2018). Processing of chromatic information in a deep convolutional neural network. *JOSA A*, 35, B334–B346.
- Forsyth, D. A., & Ponce, J. (2003). A modern approach. *Computer Vision: A Modern Approach*, 88–101.
- Gegenfurtner, K. R. (2003). Cortical mechanisms of colour vision. *Nature Reviews Neuroscience*, 4, 563.
- Gegenfurtner, K. R., Kiper, D. C., Beusmans, J. M., Carandini, M., Zaidi, Q., & Movshon, J. A. (1994). Chromatic properties of neurons in macaque mt. *Visual Neuroscience*, 11, 455–466.
- Gegenfurtner, K. R., Kiper, D. C., & Fenstemaker, S. B. (1996). Processing of color, form, and motion in macaque area v2. *Visual Neuroscience*, 13, 161–172.
- Gegenfurtner, K. R., & Rieger, J. (2000). Sensory and cognitive contributions of color to the recognition of natural scenes. *Current Biology*, 10, 805–808.
- Geirhos, R., Janssen, D.H., Schütt, H.H., Rauber, J., Bethge, M., Wichmann, & F.A., (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. arXiv preprint arXiv:1706.06969.
- Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F.A., & Brendel, W. (2018). Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv preprint arXiv:1811.12231.
- Goodfellow, I.J., Shlens, J., & Szegedy, C., (2014). Explaining and harnessing adversarial examples. arXiv preprint arXiv:1412.6572.
- Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35, 10005–10014.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).
- Jégou, S., Drozdal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 11–19).
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., Guadarrama, S., & Darrell, T., (2014). Caffe: Convolutional architecture for fast feature embedding. arXiv preprint arXiv:1408.5093.
- Judd, D. B. (1970). Ideal color space. *Color Eng*, 8, 36–52.
- Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS Computational Biology*, 10, Article e1003915.
- Kiper, D. C., Fenstemaker, S. B., & Gegenfurtner, K. R. (1997). Chromatic properties of neurons in macaque area v2. *Visual Neuroscience*, 14, 1061–1072.
- Komatsu, H. (1998). Mechanisms of central color vision. *Current Opinion in Neurobiology*, 8, 503–508.
- Komatsu, H., Ideura, Y., Kaji, S., & Yamane, S. (1992). Color selectivity of neurons in the inferior temporal cortex of the awake macaque monkey. *Journal of Neuroscience*, 12, 408–424.
- Krauskopf, J., & Gegenfurtner, K. R. (1992). Color discrimination and adaptation. *Vision Research*, 32, 2165–2175.
- Krauskopf, J., Williams, D. R., & Heeley, D. W. (1982). Cardinal directions of color space. *Vision Research*, 22, 1123–1131.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, 1097–1105.
- Lennie, P., Krauskopf, J., & Sclar, G. (1990). Chromatic mechanisms in striate cortex of macaque. *Journal of Neuroscience*, 10, 649–669.
- Long, J., Shelhamer, E., & Darrell, T. (2015). Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 3431–3440).
- Mullen, K. T., Chang, D. H., & Hess, R. F. (2015). The selectivity of responses to red-green colour and achromatic contrast in the human visual cortex: an fMRI adaptation study. *European Journal of Neuroscience*, 42, 2923–2933.
- Mullen, K. T., Dumoulin, S. O., McMahon, K. L., De Zubicaray, G. I., & Hess, R. F. (2007). Selectivity of human retinotopic visual cortex to s-cone-opponent, l/m-cone-opponent and achromatic stimulation. *European Journal of Neuroscience*, 25, 491–502.
- Nagai, J.I. & Yokosawa, K., (2003). What regulates the surface color effect in object recognition: Color diagnosticity or category. Technical Report on Attention and Cognition 28, 1–4.
- Nascimento, S., Albers, A. M., & Gegenfurtner, K. (2018). Naturalness and aesthetics of colors in the human brain. *Journal of Vision*, 18, 868.
- Nascimento, S. M., Ferreira, F. P., & Foster, D. H. (2002). Statistics of spatial cone-excitation ratios in natural scenes. *JOSA A*, 19, 1484–1490.
- Nassi, J. J., & Callaway, E. M. (2009). Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience*, 10, 360.

- Ohta, Y. I., Kanade, T., & Sakai, T. (1980). Color information for region segmentation. *Computer Graphics and Image Processing*, 13, 222–241.
- Oliva, A., & Schyns, P. G. (2000). Diagnostic colors mediate scene recognition. *Cognitive Psychology*, 41, 176–210.
- Papandreou, G., Chen, L. C., Murphy, K. P., & Yuille, A. L. (2015). Weakly-and semi-supervised learning of a deep convolutional network for semantic image segmentation, in. *In Proceedings of the IEEE international conference on computer vision* (pp. 1742–1750).
- Plataniotis, K. N., & Venetsanopoulos, A. N. (2013). *Color image processing and applications*. Springer Science & Business Media.
- Rafegas, L., & Vanrell, M. (2018). Color encoding in biologically-inspired convolutional neural networks. *Vision Research*.
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115, 211–252. <https://doi.org/10.1007/s11263-015-0816-y>.
- Shapley, R., & Hawken, M. (2002). Neural mechanisms for color perception in the primary visual cortex. *Current Opinion in Neurobiology*, 12, 426–432.
- Shapley, R., & Hawken, M. J. (2011). Color in the cortex: single-and double-opponent cells. *Vision Research*, 51, 701–717.
- Nathan Silberman, Derek Hoiem, P.K., & Fergus, R. (2012). Indoor segmentation and support inference from rgb-d images, in: ECCV.
- Simonyan, K. & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556.
- Spitzer, H., & Hochstein, S. (1985). A complex-cell receptive-field model. *Journal of Neurophysiology*, 53, 1266–1286.
- Spitzer, H., & Hochstein, S. (1985). Simple-and complex-cell response dependences on stimulation parameters. *Journal of Neurophysiology*, 53, 1244–1265.
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). Intriguing properties of neural networks. arXiv preprint arXiv:1312.6199.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions, in. *In Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).
- Tanaka, J. W., & Presnell, L. M. (1999). Color diagnosticity in object recognition. *Perception & Psychophysics*, 61, 1140–1153.
- Tanaka, J., Weiskopf, D., & Williams, P. (2001). The role of color in high-level vision. *Trends in Cognitive Sciences*, 5, 211–215.
- Virtanen, P., Gommers, R., Oliphant, T.E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S.J., Brett, M., Wilson, J., Millman, K.J., Mayorov, N., Nelson, A.R.J., Jones, E., Kern, R., Larson, E., Carey, C., İlhan Polat, Feng, Y., Moore, E.W., VanderPlas, J., Laxalde, D., Perktold, J., Cimrman, R., Henriksen, I., Quintero, E.A., Harris, C.R., Archibald, A.M., Ribeiro, A. H., Pedregosa, F., van Mulbregt, P., Contributors, S., 2019. Scipy 1.0—fundamental algorithms for scientific computing in python. arXiv:1907.10121.
- Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 28, 509.
- Witzel, C., & Gegenfurtner, K. (2014). Chromatic contrast sensitivity. *Encyclopedia of Color Science and Technology*, 1–7.
- Witzel, C., & Gegenfurtner, K. R. (2018). Color perception: Objects, constancy, and categories. *Annual Review of Vision Science*, 4, 475–499.
- Wurm, L. H., Legge, G. E., Isenberg, L. M., & Luebker, A. (1993). Color improves object recognition in normal and low vision. *Journal of Experimental Psychology: Human perception and performance*, 19, 899.
- Yasuda, M., Banno, T., & Komatsu, H. (2009). Color selectivity of neurons in the posterior inferior temporal cortex of the macaque monkey. *Cerebral Cortex*, 20, 1630–1646.
- Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. arXiv preprint arXiv:1506.06579.
- Zaidi, Q., & Conway, B. (2019). Steps towards neural decoding of colors. *Current Opinion in Behavioral Sciences*, 30, 169–177.
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European conference on computer vision* (pp. 818–833). Springer.
- Zhu, Y., Sapra, K., Reda, F. A., Shih, K. J., Newsam, S., Tao, A., & Catanzaro, B. (2019). Improving semantic segmentation via video propagation and label relaxation, in. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 8856–8865).

2.0.3 Study 3: Deep Neural Models for color classification and color constancy

Deep neural models for color classification and color constancy

Alban Flachot

Abteilung Allgemeine Psychologie,
Justus Liebig University, Giessen, Germany



Arash Akbarinia

Abteilung Allgemeine Psychologie,
Justus Liebig University, Giessen, Germany



Heiko H. Schütt

Center for Neural Science, New York University,
New York, NY, USA



Roland W. Fleming

Experimental Psychology, Justus Liebig University,
Giessen, Germany



Felix A. Wichmann

Neural Information Processing Group,
University of Tübingen, Germany



Karl R. Gegenfurtner

Abteilung Allgemeine Psychologie,
Justus Liebig University, Giessen, Germany



Color constancy is our ability to perceive constant colors across varying illuminations. Here, we trained deep neural networks to be color constant and evaluated their performance with varying cues. Inputs to the networks consisted of two-dimensional images of simulated cone excitations derived from three-dimensional (3D) rendered scenes of 2,115 different 3D shapes, with spectral reflectances of 1,600 different Munsell chips, illuminated under 278 different natural illuminations. The models were trained to classify the reflectance of the objects. Testing was done with four new illuminations with equally spaced CIE L*a*b* chromaticities, two along the daylight locus and two orthogonal to it. High levels of color constancy were achieved with different deep neural networks, and constancy was higher along the daylight locus. When gradually removing cues from the scene, constancy decreased. Both ResNets and classical ConvNets of varying degrees of complexity performed well. However, DeepCC, our simplest sequential convolutional network, represented colors along the three color dimensions of human color vision, while ResNets showed a more complex representation.

Introduction

Color constancy denotes the ability to perceive constant colors, even though variations in illumination change the spectrum of the light entering the eye. Although extensively studied (see Gegenfurtner & Kiper, 2003; Witzel & Gegenfurtner, 2018; Foster, 2011, for reviews), it has yet to be fully understood. Behavioral studies disagree on the degree of color constancy exhibited by human observers (Witzel & Gegenfurtner, 2018), and color constancy is considered an ill-posed problem. It is argued from theoretical and mathematical considerations that perfect color constancy is not possible using only the available visual information (Maloney & Wandell, 1986; Logvinenko et al., 2015). Yet, observing that humans do achieve at least partial color constancy sparks the question about which cues and computations they use to do so. It also remains unclear which neural mechanisms contribute to color constancy. Low-level, feedforward processes, such as adaptation and the double opponency of cells in early stages of the visual system, have been identified as being useful for color constancy (Gao et al., 2015). Yet, other studies suggest that higher-level and even cognitive processes such as memory also contribute. For example, better color constancy has been observed

Citation: Flachot, A., Akbarinia, A., Schütt, H. H., Fleming, R. W., Wichmann, F. A., & Gegenfurtner, K. R. (2022). Deep neural models for color classification and color constancy. *Journal of Vision*, 22(4):17, 1–24, <https://doi.org/10.1167/jov.22.4.17>.

<https://doi.org/10.1167/jov.22.4.17>

Received December 29, 2020; published March 30, 2022

ISSN 1534-7362 Copyright 2022 The Authors

This work is licensed under a Creative Commons Attribution 4.0 International License.



for known objects than for unknown ones (Granzier & Gegenfurtner, 2012; Olkkonen et al., 2008). Thus, we are still lacking a complete neural model of color constancy, which encompasses physiological similarities to the primate's visual system and at the same time exhibits similar behavior to humans on color constancy relevant tasks.

In contrast to earlier computer vision approaches, deep neural networks (DNNs) may have greater potential to be models for biological color constancy and color vision. Conceptually inspired by biology (LeCun & Bengio, 1995), DNNs can solve many complex visual tasks such as face and object recognition (Zeiler & Fergus, 2014; Yosinski et al., 2015), and DNNs trained for object recognition have been shown to correlate with neuronal activity in visual cortical regions (Güçlü & van Gerven, 2015; Cichy et al., 2016). The predictions for cortical activity are not perfect, though, and DNN responses are far less robust to distortions of the input images than human observers (Goodfellow et al., 2014; Brendel et al., 2017; Geirhos et al., 2017, 2018; Akbarinia & Gil-Rodríguez, 2020). Furthermore, it has been shown that current DNNs and human observers do not agree which individual images are easy or difficult to recognize (Geirhos et al., 2020b).

For the processing of color information specifically, similarities have been observed between DNNs trained on complex tasks and the visual system (Rafegas & Vanrell, 2018; Flachot & Gegenfurtner, 2018). In addition, DNNs trained on illumination estimation from images have outperformed all previous approaches (Lou et al., 2015; Bianco et al., 2015; Hu et al., 2017; Shi et al., 2016; Afifi & Brown, 2019). This success was enabled by fine-tuning networks pretrained on other tasks (Lou et al., 2015), various data augmentation techniques including the application of additional color distortions and cropping (Lou et al., 2015; Bianco et al., 2015), and architectural innovations and adversarial training (Hu et al., 2017; Shi et al., 2016; Afifi & Brown, 2019). Notably, none of these networks were trained only on natural variation in illuminations, and most of them aimed at the task of color-correcting images, not estimating object color.

Color constancy is also a well-studied problem in computer vision and image processing, yet the extent to which the algorithms in these engineering fields can inform our understanding of human color constancy is limited. In those fields, color constancy is typically approached by explicit estimation of the scene's illumination (Land, 1964; Akbarinia & Parraga, 2017; Afifi & Brown, 2019; Bianco & Cusano, 2019; Hu et al., 2017), followed by an image correction via the von Kries assumption (von Kries, 1902). In biological vision, however, color constancy is rather tested as the ability to extract color information about the object and materials in the scene consistently across varying

illuminations (Maloney & Wandell, 1986; Foster, 2011; Witzel & Gegenfurtner, 2018; Weiss et al., 2017; Olkkonen et al., 2008), thus going one step further than illumination estimation and requiring some form of color comprehension.

Deep learning approaches to color constancy are limited by their need for large datasets. The heavy requirements for a good color constancy image dataset (calibrated cameras, pictures taken from the same angle at different times of day, or with many different controlled and measured illuminations) result in datasets rarely containing more than a thousand images.¹ One approach to generate larger training datasets for this kind of situation is to use computer graphics to render images or videos instead. This approach has successfully been used for depth and optical flow estimation tasks (Butler et al., 2012; Dosovitskiy et al., 2015; Ilg et al., 2018), as well as other aspects of surface material inference, such as gloss perception (Storrs et al., 2021; Prokott et al., in press), but has to our knowledge not been applied to color constancy yet.

The goal of this study is (1) to teach DNNs to identify color in settings that require color constancy, (2) to assess whether the trained models exhibit behaviors akin to observations made in psychophysical studies for color constancy, and (3) to test whether human-like color representations emerge with training. To do so, we proceeded as follows: We generated artificial training and validation images using three-dimensional (3D) spectral rendering with a naturalistic distribution of illuminations to overcome the limitations of previous approaches. Instead of RGB encoded inputs, we used images encoded using human cone sensitivities. Instead of training our models on illumination estimation, we trained them to extract the color of a foreground object within the scene. Specifically, the task was to classify objects floating in a room based on their surface color, under a large set of different illumination conditions. Chromaticities of colored surfaces and illuminations were such that color constancy was necessary to attain high accuracy, that is, the chromaticity shifts induced by colorful illuminations were often larger than the chromaticity difference between neighboring surfaces. We then devised an evaluation procedure of the trained models to allow comparison with human studies. Finally, instead of using only a large, complicated standard deep learning model, we trained both complex and relatively simple ones and compared their performance as well as the color representations they developed during training.

We found that all our models performed very well at recognizing objects surface colors, even for illuminations they had never seen, with a supra-human accuracy. Like humans (Kraft & Brainard, 1999), the accuracy of the models drastically degraded, however, as we manipulated the input by gradually removing

cues necessary for color constancy. Similarly, we also found a better performance for illuminations falling along the daylight axis than for illuminations falling in the orthogonal direction. This result is in line with observations made in psychophysical studies (Pearce et al., 2014; Aston et al., 2019). We found, however, that different architectures learned to represent the surface colors of objects very differently. One of them, *DeepCC*—the most straightforward convolutional architecture we implemented—seems to represent surface colors following criteria resembling the perceptual color dimensions of humans, as determined by psychophysical studies. Other architectures like ResNets, on the other hand, did not. This suggests that while perceptual color spaces may aid color constancy, they are certainly not necessary for achieving human-like robustness to changes in illumination.

This article is divided into sections following our main findings. We start by reporting the results obtained for *DeepCC*'s evaluation, with a focus on the effect of illumination on *DeepCC*'s performance. Then we analyze how *DeepCC* represents surface colors and gradually becomes color constant throughout its processing stages. We finish with a summary of the results obtained for other deep net architectures, in particular, custom ResNet architectures.

General methods

Munsell and CIEL*a*b* coordinates

Throughout this study, two-color coordinate systems are used. The first one is the Munsell color system (Munsell, 1912; Nickerson, 1940), defined by the Munsell chips themselves. Each Munsell chip is indexed according to three coordinates: *Hue*, *Value*, and *Chroma*. *Hue* is divided into 5 main hues: Red, Yellow, Green, Blue, and Purple, each one divided into 8 intermediary hues, for a total of 40 hues. *Value* is close to *lightness* as it refers to how light a Munsell chip is perceived to be. In terms of surface reflectance, it approximately corresponds to the amount of light that gets reflected by the Munsell chip, that is, the area under curve (Flachot, 2019). *Value* varies from 0 to 10, 0 being the darkest and 10 being the lightest. *Chroma* refers to the colorfulness of the chip, or its distance from gray. In terms of surface reflectance, it corresponds to the contrast in the amount of light reflected by different wavelengths. The higher the chroma, the less flat the surface reflectance spectrum (Flachot, 2019) and the more colorful the chip. *Chroma* varies from 0 to 16. Note, however, that the Munsell color system does not have perfect cylindrical shape but has a limited gamut: Certain hues and values do not allow for high chromas. Hence, the full set of Munsell chips consists

of only 1600 chips instead of $40 \times 16 \times 10 = 5,600$ chips. Because the Munsell color system is defined by the Munsell chips, it is the most appropriate space to discriminate Munsells. In addition, the Munsell chips were chosen in an attempt to be perceptually uniformly distant, and as such, the Munsell coordinate system is an approximately perceptually uniform space.

Another perceptually uniform color space is the CIEL*a*b* (Ohno, 2000) coordinate system. It was constructed such that its Euclidean distance, commonly called ΔE , is an approximate measure of perceptual difference: Two colors equidistant to another in CIEL*a*b* are *approximately* perceptually equidistant. Additionally, it is commonly considered that the average just noticeable difference (JND) between two colors is approximately $2.3 \Delta E$ (Mokrzycki & Tatol, 2011), meaning that a human observer is not able to discriminate two color patches closer than this value, even if placed side-by-side. Of the three dimensions, *L** accounts for lightness, *a** accounts for greenish-reddish variations, and *b** accounts for blueish-yellowish variations. The white point (point of highest Lightness) was computed using the spectrum of the light reflected by the Munsell chip of highest value, under the D65 illumination. This Munsell chip is also an achromatic chip.

To relate the two color coordinate systems, the median distance between two adjacent Munsell chips is equal to $7.3 \Delta E$ (i.e., significantly above the JND).

Image generation

In the present study, we generated our own images using the physically based renderer.² Mitsuba was developed for research in physics and includes accurate, physics-based approximations for the interaction of light with surfaces (Pharr et al., 2016; Bergmann et al., 2016), yielding a perceptually accurate rendering (Guarnera et al., 2018). Most important, it also allows the use and rendering of spectral data: One can use physically measured spectra of lights and surfaces as parameters. Outputs can also be multispectral images rather than simple RGB images. We exploited this multispectral characteristic of Mitsuba using the reflectance spectra of 1,600 Munsell chips (Munsell, 1912) downloaded from Joensuu University³ (Kalenova et al., 2005). As illuminations, we used the power spectra of 279 natural lights: 43 were generated from the D series of CIE standard illuminations (Judd et al., 1964; Ohno, 2000) at temperatures ranging from 4,000K to 12,000K; 236 were taken from the forest illuminations measured by (Chiao et al., 2000). Each illumination spectrum was normalized such that their highest point reaches the same, arbitrary value of a

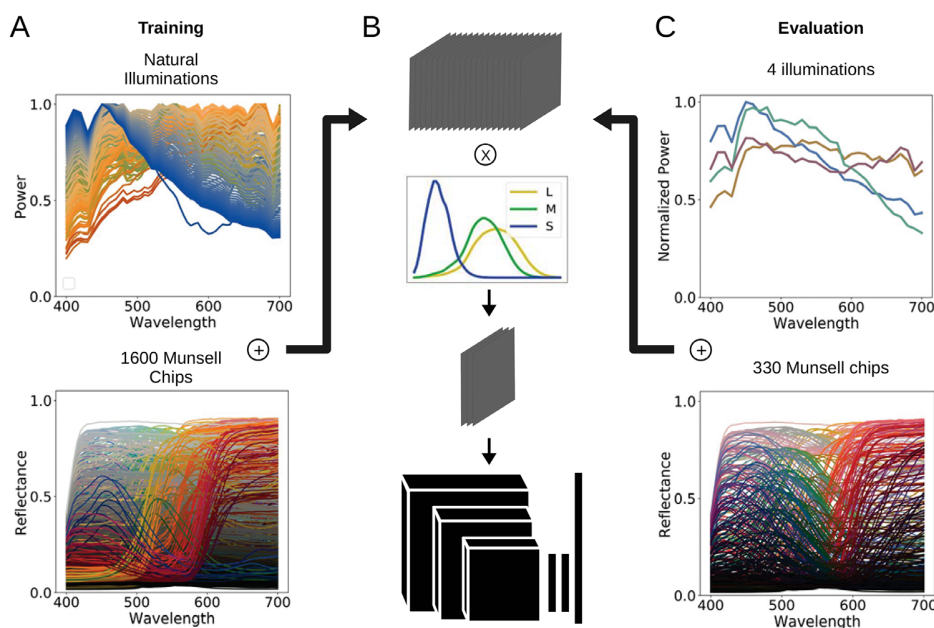


Figure 1. Figure illustrating our method, both for training and evaluation. (A) To generate the training set of images, sets of 279 spectra of natural illuminations and 1,600 spectra of Munsell reflectances were used. The resulting multispectral images (B) were then converted into three “LMS” channels using human cone sensitivity spectra and fed to the network. (C) The four illuminations R, G, Y, and B were used exclusively in the evaluation. Note that while Y and B fall on the daylight locus, R and G have chromaticities different from the illuminations of the training set. Out of 1,600, only 330 Munsell spectra were used.

100. The spectra of both Munsell reflectances and illuminations are displayed in Figure 1A.

For meshes, we used a compilation of object datasets issued by Evermotion⁴ for a total of 2,115 different meshes, ranging from human-made objects to natural objects. Each mesh was normalized such that they have the same size (equal longest dimension).

In order to approximate the input to human visual processing, we first generated our images with 20 channels, at equally spaced wavelengths ranging from 380 to 830 nm. These were then collapsed onto three “LMS” channels using measured human cone sensitivities (Stockman & Sharpe, 2000). Images were saved with floating points, thus without the need for any gamut correction or further processing. This procedure is illustrated in Figure 1B.

The 3D scene consisted of a simple “room” (see Figure 2), with three walls, a floor, and a ceiling with constant Munsell reflectances as surfaces. On the ceiling, a rectangular light source was defined. On the back wall, six colorful patches with constant Munsell reflectances were added. Their purpose was giving additional cues for the model to solve color constancy, as seems to be necessary for humans (Brainard et al., 2003; Yang & Maloney, 2001).

Finally, each LMS image consisted of a random object floating at a random position and orientation in the scene, with a given Munsell surface reflectance. The shape of the object was taken randomly among our

pool of 2,115 meshes. Although its position was also random, it was bounded so that the object would never occlude the six patches in the background and would stay fully within the field of view. We generated two datasets, the *Set-CC* and *Set-D65* datasets. Illustrations of these datasets are available in Figure 2. In the *CC* dataset, we generated 279 images per Munsell chip, one for each of the 279 natural illuminations. In the *D65* dataset, we also generated 279 images per Munsell chip value but kept the illumination constant with the power spectrum of the standard D65 illumination. Each dataset thus consisted of $1,600 \times 279 = 446,400$ images, with a resolution of 128×128 pixels and three color channels, one for each L, M, and S cone photoreceptor. Images were labeled according to the mesh type, object position, illumination, and, most important in this study, according to the Munsell chip used for the mesh’s surface reflectance. All surfaces were defined as Lambertian. This dataset is publicly available,⁵ as well as the pipeline to generate it.

Deep architecture

One network architecture has been extensively studied throughout this work. Several others were also tested, evaluated, and analyzed, for which results are described in detail in “Standard and custom

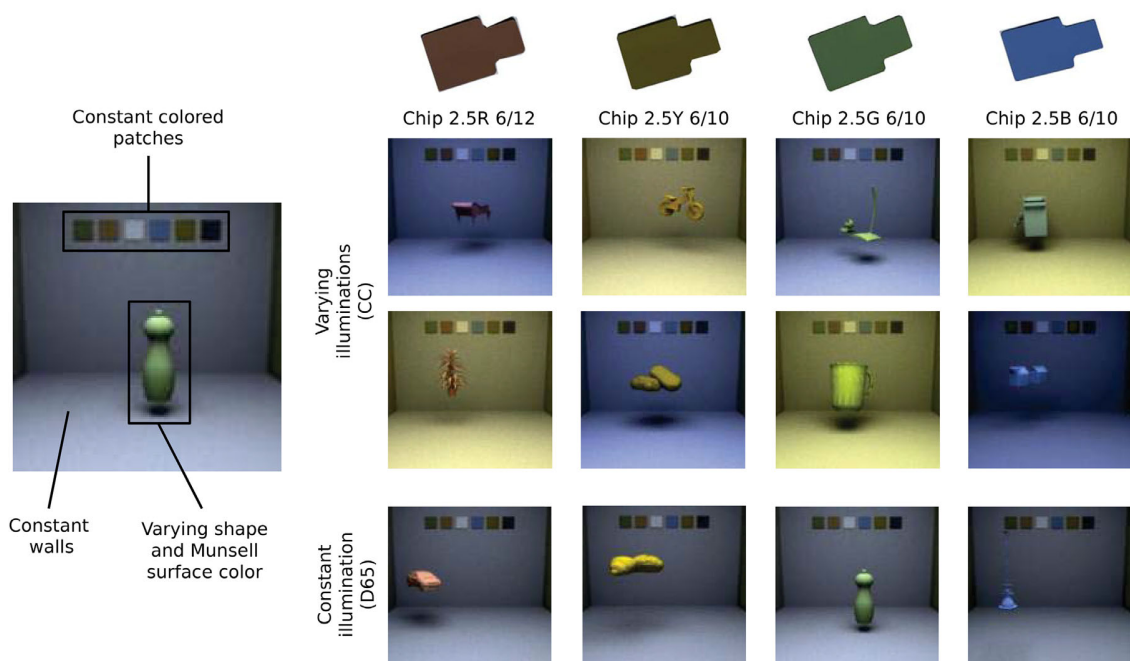


Figure 2. Illustration of the two training datasets used: one with varying illumination (CC), another with a constant illumination (D65). The classification task consisted of identifying the correct Munsell chip used as surface reflectance for a random object floating in the room. In order to be performant on the CC dataset, the network had to account for the illumination.

architectures.” For now, we limit ourselves to describing the network architecture most relevant for this study, which we refer to as *Deep*.

Deep has a convolutional architecture (LeCun et al., 1998; Krizhevsky et al., 2012) with three convolutional layers and two fully connected layers preceding a classification layer. Convolutional layers can be described as a set of linear kernels. Each kernel applies the same linear filter of limited size on different portions of the input, at regular intervals. The output of one linear filter applied on one input patch, coupled with a half-wave rectification (ReLU), is the output of one unit. Units in convolutional layers have thus limited receptive fields in the image input. Fully connected layers instead take all units of the previous layer as input, such that the units’ receptive fields cover the whole input image. The convolutional layers of the *Deep* architecture have 16, 32, and 64 kernels with kernel sizes of 5, 3, and 3, respectively. After each convolutional layer follows a 2×2 maxpooling layer. The fully connected layers have 250 units each. The classification layer is a simple fully connected layer, preceded by a 40% dropout layer for regularization.

Deep’s input consisted of the set of images we generated, thus with a dimension of 128×128 pixels and three color channels, one for each L, M, and S cone photoreceptor.

Task and training

The training was supervised with the learning objective of outputting the Munsell chip label for each image (i.e., the color of the object in each scene). Cross-entropy was used as loss. Training took place for 90 epochs. We used the Adam optimizer (Kingma & Ba, 2015), with a learning rate of 0.001, divided every 30 epochs by 10.

We trained separate models on the *CC* and *D65* datasets. Each dataset was further divided into training and validation subsets, the former consisting of 90% of the dataset’s images and the latter the remaining 10%. Training and validation subsets are quite similar: They use the same viewpoint and the same room, although the floating object was at random position and orientations. But they also have differences: They do not use the same object meshes, and in the case of *CC*, neither do they use the same illumination spectra. The validation subsets were generated with 212 object meshes and for *CC* 28 illumination spectra exclusively, selected randomly among the 2,115 meshes and 279 illuminations. The remaining meshes and spectra were used for generating the training subsets. Training subsets were only used for training our models, while the validation sets were only used for testing the model during training at regular intervals (each epoch)

to monitor its performance on images it had never seen.

We can now see how our task requires the models to become color constant: In order for the models to achieve a high recognition accuracy on the *CC* dataset, they would need to compensate for the chromatic shifts that are induced by the varying illuminations interacting with the Lambertian surfaces. By extension, this means they would need to achieve some degree of color constancy. Indeed, the standard deviation of the training illumination's distribution is equal to 8.55 ΔE , higher than the median distance between two adjacent Munsell classes of 7.3 ΔE . Out of the 279 illuminations in our training and validation sets, 202 are distant by more than 10 ΔE from the reference illumination D65.

Given that there are two datasets, *CC* and *D65*, two kinds of training instances need to be distinguished: *DeepCC* when trained on *CC* and *Deep65* when trained on *D65*. Due to several randomization procedures implemented during training, two training instances of the same architecture trained on the same dataset will give slightly different results. To allow broader claims and a statistical analysis, we trained 10 instances of *DeepCC* and *Deep65* each.

Each model was trained on one GeForce GTX 1080. Batch size varied from architecture to architecture but was maximized to fit the GPUs memory. In the case of *Deep*, the batch size was 800 images. All the code is available on Github.⁶

Other than the validation dataset, we devised other datasets to further evaluate our models. These evaluation datasets mimicked the typical experimental procedures for studying color constancy, consisting in removing or ambiguously modifying contextual cues to make the task more difficult (Witzel & Gegenfurtner, 2018; Kraft et al., 2002). They facilitated identifying the relevance of diverse cues for the task, the testing the model's robustness to scene modifications, and the comparison with previous psychophysical studies. These contextual modifications were (1) removing the colored patches in the background—if the models use the constancy information transmitted by these patches, a drop in performance should follow. (2) Swapping the colored patches in the background with patches under a different illumination—again, if the models use the constancy information transmitted by these patches, a drop in performance should follow. (3) Placing the floating object in a background illuminated with a wrong illumination—if the models follow the information within the scene to estimate the illumination's color, then the resulting incorrect estimation should lead to a misclassification of the floating object's color.

A detailed description of the evaluation datasets will follow in “Evaluation DeepCC and Deep65” and “Impoverished visual scene,” sections where the results of these evaluations are presented.

Metrics

To assess the performance of *DeepCC* and *Deep65*, we used several measures of accuracy. Given that the task is the classification of Munsell chips, two are the standard top-1 and top-5 accuracies (Krizhevsky et al., 2012): top-1 counts as hit when the correct Munsell is the one selected as most probable by the model; top-5 counts as hit when the correct Munsell is among the five selected as most probable by the model. In addition, we defined the Muns³ accuracy: A hit occurs whenever the Munsell selected as most probable by the model is 1 Munsell away from the correct one (within a cube of side 3 in Munsell space centered on the correct Munsell).

Due to their discrete nature, however, top-1, top-5, and Muns³ accuracies do not discriminate between cases when a model selected a Munsell just outside Muns³ or when it was completely off. To correct this shortcoming, we converted the model's output into chromaticity coordinates. We did so by considering the Munsell chips' chromaticities under the D65 illuminant in CIEL*a*b* space. We then defined the model's *selected chromaticity* as the chromaticity of the Munsell selected by the model. The Euclidean distance between the correct Munsell's chromaticity and the model's selected chromaticity now defines a continuous measure of the model's error. Following the literature (Ohno, 2000; Weiss et al., 2017), we call this error ΔE (with its 1976 definition).

To further compare with the color constancy literature, we considered another measure called the *Color Constancy Index* (CCI) (Foster, 2011; Arend & Reeves, 1986; Weiss et al., 2017). This measure has the benefit of taking into account the quantitative error of the model in color space (ΔE) relative to chromaticity shift induced by the illumination. Consider that we present to the model an image showing a floating object under an illumination *I* with the surface reflectance of a Munsell *M*. Consider now that the model recognizes the wrong Munsell *N*. Then the Color Constancy Index is defined as

$$\begin{aligned} CCI &= 1 - \frac{|C_I^N - C_I^M|}{|C_{D65}^M - C_I^M|}, \\ &= 1 - \frac{\Delta E}{|C_{D65}^M - C_I^M|}. \quad (1) \end{aligned}$$

where C_I^M is the chromaticity of the Munsell *M* under the illumination *I*, C_{D65}^M is the chromaticity of the same Munsell chip but under the standard illumination D65, and C_I^N is the chromaticity of Munsell *N* under the illumination *I* and recognized by the model. If the model recognizes the correct Munsell, then the ratio in the formula is neutral and CCI would be equal to 1. However, if the model does not compensate for the

illumination's shift in chromaticity and recognizes the wrong Munsell chip, CCI would be close to 0. Negative values of CCI indicate that the network chose the wrong Munsell for other reasons, beyond the chromaticity shifts induced by the illumination.

DeepCC and Deep65 evaluation

This section focuses on the evaluation of DeepCC and Deep65. Results for other architectures can be found in “Standard and custom architectures.”

We first present the results of training and validation for both DeepCC's and Deep65's instances. We then present thorough evaluations of the models using additional, custom datasets (description below).

We found that both DeepCC and Deep65 reached high top-1 accuracies on their respective validation datasets. DeepCC instances reached on average 76% accuracy on the CC validation set, while Deep65 reached on average 86% accuracy on the D65 validation set. These values clearly show that the two sets of networks learned how to solve their task and are able to differentiate between 1,600 different surface colors reasonably accurately (random performance would be 0.0625%). The higher performance of the Deep65 network also indicates, as expected, that the D65 task is inherently easier than when illumination is allowed to vary, and thus color constancy is required to perform the task.

In order to evaluate DeepCC in greater detail, as well as allowing some comparison with observations made in psychophysical studies, we generated another set of testing images, with settings closer to conditions found in typical perceptual experiments.

Methods

To facilitate our analysis, an evaluation dataset was generated using a slightly different procedure than for the training sets. First, a subset of 330 Munsell chips was used, instead of the original set of 1,600 (cf. Figure 1C). This subset was originally used for the World Color Survey and is now a standard for studies focusing on color naming (Berlin & Kay, 1969). It is also widely used in studies related to color categories (Witzel, 2019) and unique hues (Philipona & O'Regan, 2006; Flachot et al., 2016). As such, they are an excellent basis for comparing our models with human judgments.

Second, we used four illuminations (cf. Figure 1C) equidistant to the CIEL*a*b* gray point by 10 ΔE (Ohno, 2000) in the chromaticity plane. This procedure was inspired by experimental studies on illumination discrimination and estimation (Aston et al., 2019). Two, B and Y, lie on the daylight locus projected onto the

chromatic plane, and are thus within the distribution of the natural illuminations used during training. The other two, G and R, lie in the orthogonal direction, which crosses the daylight locus at the gray point, and are outside of the distribution of illuminations used during training. More precisely, G is 4.45 ΔE away from its closest illumination within the training set, while R is 7.9 ΔE away, making R then G the two illuminations DeepCC is less familiar with. Their power spectra were generated with the principal components of natural daylight spectra defined by Judd et al. (1964), which serve as the basis for the D series of the CIE standard illuminations. These illuminations were normalized such that their areas under curve were equalized, thus minimizing their difference in Lightness. For each Munsell of the 330 Munsell classes and each of the four illuminations, we generated 10 images for a total of $330 \times 4 \times 10 = 13,200$ images.

Note the fundamental difference between the validation sets employed earlier and the evaluation set defined here: While the validation datasets consisted of illuminations and 3D shapes the networks had never seen (to prevent overfitting), these illuminations and shapes were still taken randomly from the same distributions as for the training set (see General methods). The evaluation dataset, however, included illuminations that were completely outside of the illumination distribution used at training time. As such, our evaluation procedure is in accordance with the recommendations from the machine learning community and formally defined recently (Geirhos et al., 2020a): using one *independent and identically distributed* (i.i.d.) test set—our validation set—and another *out of the distribution* (o.o.d.) test set—the evaluation set described here.

Although the illumination spectra were different from the ones used during training and validation, the scene in which the floating objects were displayed was exactly the same. We therefore refer to this evaluation dataset as *normal*. Because we are evaluating *DeepCC* and *Deep65*, each trained on different datasets, we distinguish between two conditions: *CC* and *D65*.

Results

Figure 3A shows the distributions obtained for each of our five metrics under the CC and D65 conditions. For the accuracies, we considered the distributions of values found for each Munsell class and illuminations (each point of the distribution is thus computed with 10 images). For ΔE and CCI, we plot the distributions of values found for individual images. Under the CC condition, we found median top-1, top-5, and Muns³ accuracies of 80%, 100%, and 100%, respectively, across Munsell classes. The first quartiles are at 60%, 90%, and 90%, respectively. This means that for the majority of

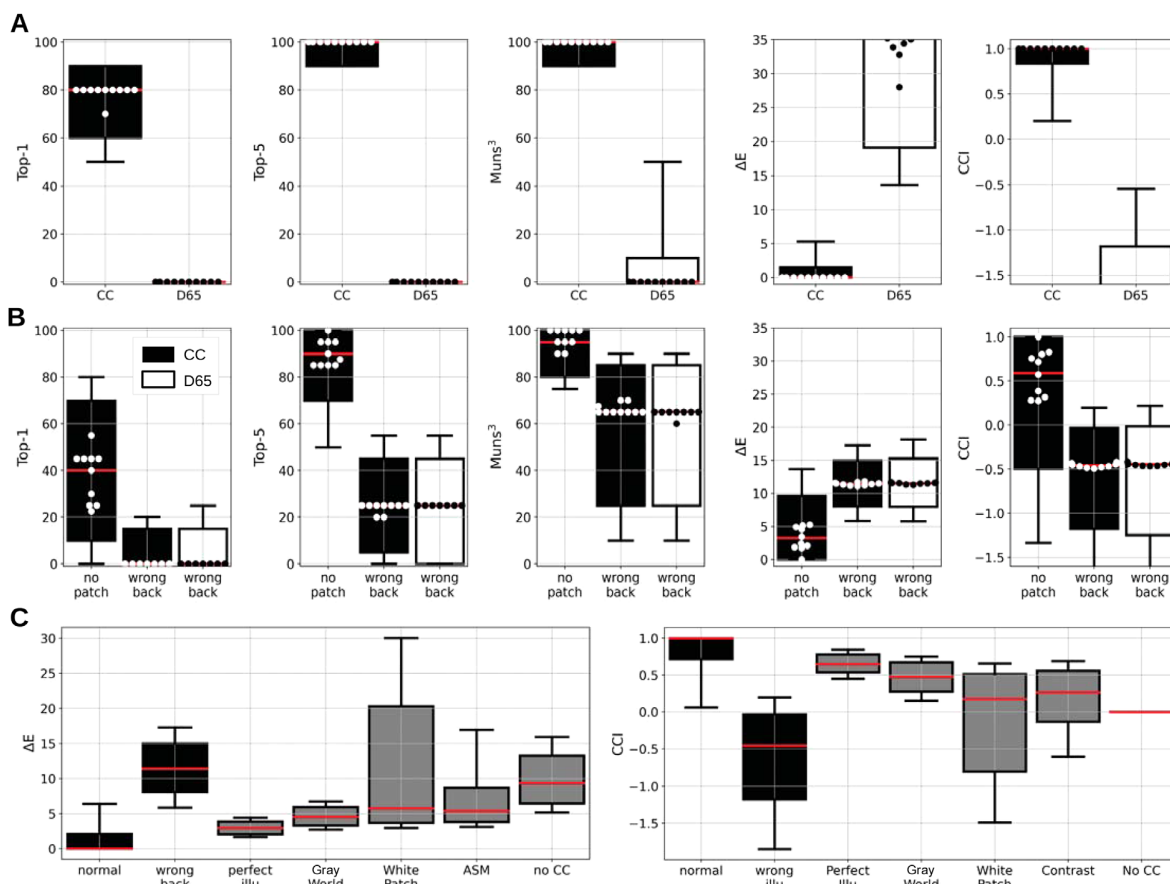


Figure 3. DeepCC's evaluation results obtained for all measures and all conditions. Each column corresponds to one measure (*Top-1*, *Top-5*, and *Muns³* accuracies, ΔE errors to ground truth, and CCI). Boxplots show distributions across Munsell chips; the swarm shows the performance of the 10 training instances. (A) Performance for models trained under varying illuminations CC or the D65 illumination only D65 ("DeepCC and Deep65 evaluation"). The models trained on the CC dataset learned to classify Munsell colors accurately even under novel illuminations, contrary to the models trained on D65 only. (B) In black, performance of DeepCC under the *no patch* and *wrong back* conditions ("Impoverished visual scenes"). In white, performance of Deep65 under the *wrong back* condition. DeepCC learned to rely on contextual cues within the scene to solve the task. When these cues are taken away or incongruously altered, the model's performance decreases. Under the wrong back condition, where the background is artificially kept constant despite various illuminants shining on the object, DeepCC performs at the level of Deep65. (C) Performance of DeepCC compared to other approaches to color constancy ("Classical approaches"), namely, perfect illumination estimation and von Kries adaptation (perfect illu), Gray World, White Patch, ASM, and no account for illumination whatsoever (no CC). Under the normal condition, DeepCC performed better than any algorithm tested, even better than a model that would perfectly estimate the illumination and then perform the standard von Kries adaptation (perfect illu condition).

Munsell classes, DeepCC selects the correct Munsell class in four out of five images, and when wrong, it still selects a neighboring chip. This is confirmed by the distributions found for ΔE and CCI, with median values of 0 and 1. Eighty-five percent of the images yielded less than 5 ΔE error as indicated by the whiskers, 93% less than 10 ΔE error, and 99% less than 19 ΔE . As a comparison, note that the median ΔE distance between adjacent chips is approximately 7.5. This means that when DeepCC instances selected the wrong chip, it tended to be a close neighbor of the

correct one. This is confirmed by the *Muns³* accuracy, according to which the model had an accuracy equal to or above 90% for 95% of the Munsell classes. Similarly, DeepCC showed a CCI higher than 0.83 in 75% of cases. This CCI value of 0.83 is among the higher end of CCI values measured in humans psychophysical experiments (cf. Foster, 2011; Witzel & Gegenfurtner, 2018, for reviews), thus indicating the supra-human performance of the model on this dataset. We also found a positive CCI value in more than 87% of cases, evidence that DeepCC not only learned to discriminate

between Munsell colors with high accuracy but also learned to account for potential color shifts induced by the illumination.

Results were, however, very different for Deep65—the network trained using only a single illuminant, D65. We found median values of 0 in all three accuracies, meaning the 10 training instances of Deep65 rarely came close to selecting the right Munsell class. This is made clear with the distributions of the ΔE and CCI measures. For the vast majority of the images, Deep65 exhibited errors of above 10 ΔE and negative CCI, meaning that Deep65's error cannot be explained by the illumination change alone. This indicates that Deep65 lacks the ability to cope with illuminant deviations from the one it has been trained on, whereas DeepCC could generalize to novel illuminants beyond the 279 different illuminants it had been trained upon.

Interim conclusion

Results so far show that DeepCC did learn to accurately classify color surfaces under varying illumination. In doing so, it also learned to discount the illumination color, reaching a high degree of color constancy, even for illuminations outside of the gamut of illumination spectra used for training. Deep65, on the other end, performed very poorly on the four illuminations used for testing.

Impoverished visual scenes

We have seen that DeepCC achieved supra-human performance under normal conditions on the devised evaluation dataset, thus achieving some degree of color constancy. A remaining question is which elements within the scene DeepCC used to compensate for illumination change: Does it consider, for example, the six constant color patches in the background? Given that there are interreflections between the floating object and the surrounding walls, is there any need for the model to use cues in the background at all?

Computer graphics allow us to manipulate the scene elements to test these questions. We thus devised new datasets to gain insights into which cues within the images DeepCC might use to achieve color constancy. Three manipulations were conducted: (1) removing the constant patches in the background, (2) modifying the colored patches in the background to have the wrong color, and (3) showing a floating object illuminated by one illumination in a scene illuminated by another illumination.

We then tested DeepCC on these three new datasets, without any additional training.

Methods

We generated three new image datasets to test DeepCC, in which some elements within the scene were removed or incongruously modified. These elements constituted cues that are known to be useful to humans for achieving color constancy. Previous experiments (Kraft et al., 2002) have shown that increasing the color cues within a scene, in their case adding a Macbeth color checker, can increase color constancy for humans. Thus, in one dataset, the *no patch* dataset, we removed the six constant patches located on the back wall. If the networks do partially rely on the information given by the background patches to solve color constancy, then the missing information should lead to a drop in model performance. Other studies (Kraft & Brainard, 1999) showed that human color constancy is neutralized when the context surrounding the object of interest is manipulated incongruously. Thus, in two other datasets, *wrong patch* and *wrong background*, we gave the network conflicting contextual cues. In *wrong patch*, we modified the chromaticities of the six colored patches, originally under one of the four test illuminations, to be replaced by their color under the D65 illumination. In *wrong background*, the floating object, illuminated by one of the four test illuminations, was cropped out and placed in the same scene but illuminated by the D65 illumination. If the networks do use the background information to solve color constancy, then the misleading information should also lead the models' performance to drop, and significantly more so than in the *no patch* condition. Note that for the last condition, human observers would be expected to be unable to solve the task. Examples of images illustrating these conditions are shown in Figure 4.

Results

Results are shown in Figure 3B. The results for DeepCC are plotted in black and the results for Deep65 under the *wrong background* condition are plotted in white. Overall, DeepCC performed significantly worse in each of the three new conditions than in the normal condition, but still better than Deep65 in the normal condition. Performance for the *no patch* condition was on average still fairly high, indicating that the networks did not rely solely on the constant patches to perform the task. The three accuracy distributions include medians of 40%, 90%, and 100%. Muns³ in particular shows a first quartile at 90% accuracy, evidence that deepCC was selecting a Munsell chip within the direct vicinity of the correct one in the vast majority of cases under this condition. ΔE and CCI measures lead to the same conclusions: Median ΔE is found at 3.3 and a third quartile at 9.40, thus showing that in the

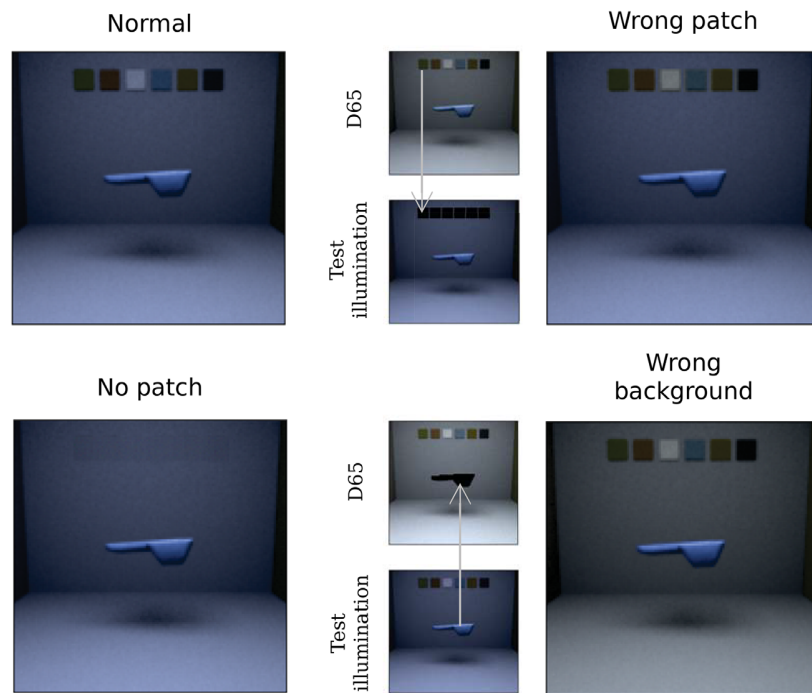


Figure 4. Example for the four types of images we used during testing: normal, no patch (the colored patches in the background are removed), wrong patch (the colored patches are cropped and replaced by the same surfaces but under the D65 illumination), and wrong background (the object is cropped and placed in the room illuminated by D65 illumination).

large majority of cases, the model showed an error of the same magnitude as the interdistance between Munsell chips in CIEL*a*b*. The analysis of the CCI distribution leads to the same conclusions: We found a median value of 0.62 but a first quartile at -0.43 . This indicates that, while for most images DeepCC performs relatively well under the no patch condition (a CCI of 0.62 remains in the upper-half of CCI values reported in humans psychophysics), it is generally more difficult for the model to solve the task, to the extent that a significant proportion of images yields a negative CCI.

Interestingly, the reliance on the back patches' presence was not equal across DeepCC's training instances. One instance saw its accuracy change by merely 20%, while another experienced a drop of 60%. Refining our scene manipulations, we also looked at how the model's instances responded when masking one colored patch in the background at a time. Some patches appeared more critical than others: Masking the red and yellow patches (second and fifth from the left) led to the largest loss in accuracy, with average losses of 9.9% and 8.9%. Masking the white and black patches (third and sixth from the left), however, had the least impact on the model's performance, accounting for losses of 0.1% and 4%, respectively, on average. Individual differences were also confirmed. When masking the red patch, for example, one instance dropped by 22% in accuracy, while another dropped

only by 2.3%. Some instances were also mainly affected by the red patch, others by the yellow patch. Nevertheless, the relatively high accuracies and CCI show that the model remained able to perform the task, albeit less successfully. The fact that different patches had different influences also tends to suggest that the decline in performance was not just a generic decline associated with deviations from the training set, but rather reflected the use of specific information from the patches to support color constancy.

These results are evidence that DeepCC indeed uses the information provided by the six constant colored patches in the back wall—particularly the chromatic ones. This is confirmed by the performance obtained for the model on the *wrong patch* dataset (data not shown). Indeed, we found that the models performed equally well or worse under this condition than under the *no patch* condition. Contrary to the latter, *wrong patch* introduces a conflicting cue rather than the absence of one, thus making the task even more difficult for the model if it partially relies on the colored patches in the background. Still, we found a median CCI value of 0.22, thus showing that despite the conflicting cues, the model retained some degree of color constancy and must rely on other cues to account for the illumination's color.

In the *wrong background* condition, however, DeepCC's performance dropped considerably, with

a median top-1 accuracy at 0, and median CCI values below 0 for all training instances. In fact, its performance dropped to the same level as our control model's Deep65 tested on the same dataset. DeepCC shows a median ΔE error of 11.4, for instance, and 11.3 for Deep65. In the *wrong background* dataset, the background was manipulated such that it appeared constant across all test illuminations, and illuminated by our reference D65. This near equality is strong evidence that DeepCC relies solely on the contextual cues surrounding the floating object to perform color constancy: When deprived of these cues, it interprets any chromaticity shifts induced by the test illuminations as being intrinsic to the object's surface and thus wrongly classifies its color, just like Deep65 would.

Interim conclusion

Thanks to the controlled manipulation of the scene surrounding the floating object, we saw in this section that all DeepCC instances solely rely on contextual cues to identify the object's Munsell surface and account for illumination change: When deprived of reliable cues surrounding the object of interest, it behaves the same as Deep65, the same architecture trained with the D65 illumination only. Similarly, humans rely on contextual cues to solve color constancy (Kraft & Brainard, 1999; Kraft et al., 2002). Individual differences were observed between training instances, however, when the colored patches in the background were removed, with some instances relying more on certain patches than others.

Standard approaches

To further evaluate DeepCC, we compared its performance to the error expected with classical approaches to illumination estimation, coupled with the von Kries correction (Von Kries, 1902), standard in computer vision (Akbarinia & Parraga, 2017; Hu et al., 2017).

Methods

For comparison purposes we also computed, on our test images of the CC normal condition, the errors expected from classical approaches to illumination estimation: *Gray World*, *White Patch* (Land, 1977), and *adaptive-surround modulation* (ASM) (Akbarinia & Parraga, 2017). All of these approaches are driven by low-level features (as opposed to learning): *Gray World* makes the assumption that the world is on average "gray" under a neutral illumination and takes the average pixel value as an estimation of the illumination's

color; *White Patch* considers the brightest pixel as an estimation of the illumination; ASM assumes that image areas with high to middle spatial frequencies (typically edges) are most informative and computes the illumination by dynamically pooling a portion of the brightest pixels according to the average image contrast. Each of these approaches delivers a single global triplet of values specifying the illuminant for a given image.

To enable a link from the global illumination estimations to our classification task of the floating object's surface color, we coupled these approaches with a global von Kries correction (von Kries, 1902). This correction consisted in dividing each image pixel by the three illumination values estimated by each approach. For each resulting image, we then segmented the floating object and estimated its chromaticity by considering the mean value of all its pixels. We then compared this estimated chromaticity with the chromaticity found for the exact same object, at the exact same position and orientation, but under a reference illumination. In this way, any difference between the estimated chromaticity and the reference chromaticity would be a consequence of the illumination estimation + correction only. As a reference, we used the computed chromaticity of the object rendered under the D65 illumination.

Of course, there are many other approaches to illumination estimation and white-balance correction than the ones tested here, some of which may be more accurate (see Akbarinia & Parraga, 2017, for a review; Shi et al., 2016; Afifi & Brown, 2019). All of them, however, deal with RGB images and rely on the global von Kries adaptation for the final correction, which in itself is an approximation. As an upper bound for any approach based on illumination estimation and von Kries adaptation, we also estimated the error of the von Kries method based on the ground truth illumination (perfect illumination estimator) using the same evaluation procedure as for estimated illuminations. This object color estimate is not perfect, because it does not take into account local variations in illumination, due to interreflections within the scene, for instance (Worthey & Brill, 1986; Flachot et al., 2016; Foster, 2011). Finally, we also computed the error obtained without compensating for the illumination at all. This would serve as an error estimate for a model that would perfectly segment the object of interest in the scene, but not compensate for the illumination (a perfect non-color constant model). By definition, such a model would thus have a CCI of 0.

Results

Figure 3C shows the distributions of ΔE errors and CCI predicted from the classical approaches

to color constancy, together with the results obtained under the *normal* and *wrong background* conditions, described previously, for comparison purposes.

We found median ΔE values for all of the aforementioned approaches to be higher than for DeepCC under the *normal* condition. Even the error merely induced by the von Kries adaptation (perfect illumination condition in the figure) leads to higher errors, with a median value of 2.9. This median value is in fact very similar to the median found for the *no patch* condition, although slightly better. This is confirmed by the corresponding median CCI of 0.65. Of the classical approaches, the Gray World hypothesis proved to be the most accurate, with median values of 4.6 ΔE and 0.48 CCI, slightly worse than for DeepCC on the *no patch* condition. This suggests that not only did the DeepCC instances accurately identify the region of interest that is the object within the image and managed to accurately estimate the illumination, but they also accounted for the object's position with respect to the illumination. It also implies that DeepCC found a better correction strategy than a global discounting of the illumination like in the von Kries approach. This is presumably thanks to the nature of the task, which tries to estimate object color rather than a global illumination, and thanks to the convolutional nature of the model's architecture, which allows local discounts of the illumination.

Although DeepCC under the *wrong background* condition exhibits errors greater than every one of the standard approaches, it is as well to note that its distribution is quite close to the distribution predicted for a perfect non-color constant model (*no CC* condition in the figure). Indeed, we find a median error of 9.4 ΔE for the *no CC* condition, similar to the 11.4 ΔE found for the *wrong background* condition. This suggests that DeepCC is indeed misled to attribute a neutral illumination on a floating object and thus behaves like a non-color constant model. Since DeepCC performs at the same level as DeepCC on the same dataset, it is likely that the discrepancy of 2 ΔE between *no CC* and *wrong background* comes from the fact that DeepCC is no perfect Munsell classifier, even with all contextual cues available.

Gray World's success compared to other approaches can be explained by the relative simplicity of the scene: a room with fairly neutral walls, with a single illumination. ASM would be expected to perform better using images of more complex scenes. The poor performance of the White Patch approach for many images can be understood by the proximity of the object of interest to the camera: When a Munsell reflectance of high value is applied to the object, the brightest pixels are likely to be found on the object itself, rather than on some other parts of the context.

Interim conclusion

Comparisons with classical approaches to color constancy show that under the *normal* condition, DeepCC learned how to compensate for the illumination better than any of the classical approaches we tested. It even performed better than a hypothetical model provided with omniscient knowledge of the true illumination and compensating through the von Kries correction, the standard procedure for discounting in a scene the illumination after its estimation (Akbarinia & Parraga, 2017). Under the *wrong background* condition, its performance lies close to the predicted performance of a model that would perfectly segment the object of interest in the scene and extract its chromaticity, but not account for the illumination color. This suggests that similarly to humans, it also relies on context to achieve color constancy (Kraft & Brainard, 1999; Kraft et al., 2002; Yang & Maloney, 2001).

Effect of illumination

To test the DeepCC models, we used the four illuminations: Yellow (Y), Blue (B), Green (G), and Red (R) (see Figure 1C). These were chosen to be equidistant to D65 in CIEL*a*b* space, with Y and B on the daylight locus and G and R in the orthogonal direction. Note, however, that even though none of these four illuminations were used during training, Y and B are expected to appear more “familiar” to the models than the other two. Indeed, the distribution of natural illuminations used for training includes several other illuminations along the daylight locus. G and R, however, were outside the distribution of the training set. More precisely, G is 4.45 ΔE away from its closest illumination within the training set, while R is 7.9 ΔE away, making R then G the two illuminations DeepCC is less familiar with.

This anisotropy in the distribution of natural illuminations had consequences on the performance of our models and their degree of color constancy. For each training instance and illumination, we computed the mean CCI per Munsell class and training instances, with each mean value computed across 10 image exemplars in the *normal* conditions. Figure 5 shows the distributions of these mean values for each of the four illuminations in the form of a boxplot. Additionally, we also plotted the average CCI value for each training instance under each illumination in the form of bee swarms. We observed a significant effect of the illumination on the CCI of our models: DeepCC models showed higher CCI for the “familiar” illuminations (Yellow and Blue) than for the “unfamiliar” illuminations (Green and Red). The highest degree of color constancy was found under the

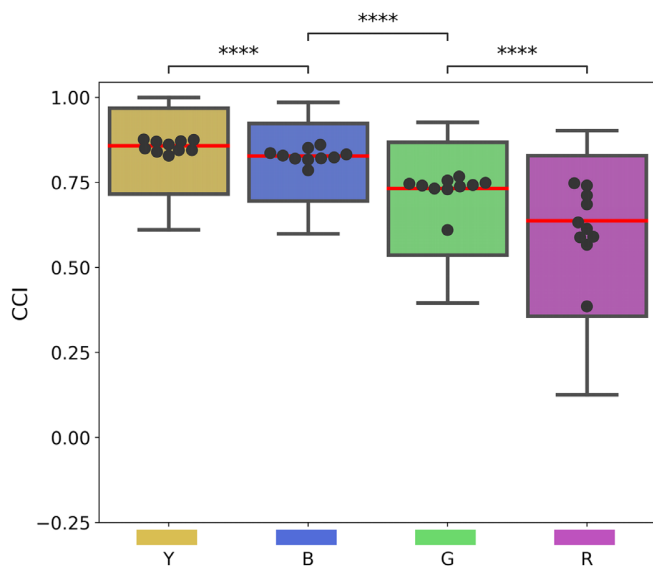


Figure 5. Effect of the illumination on color constancy: distributions of DeepCC's mean Color Constancy Index (CCI) for each Munsell class under each of the four testing illuminations. Medians are in red. Each dot of the bee swarm plots is to the average CCI found for a training instance of DeepCC. Statistical significance was computed applying pairwise *t* tests with Bonferroni corrections.

Yellow illumination, with an average CCI value of 0.86, while the lowest was found under the Red illumination, with an average CCI value of 0.64.

Results of Figure 5 are very similar to observations made regarding the capacity of humans to perceive illumination changes (Pearce et al., 2014; Aston et al., 2019). It was found that human observers were more sensitive to illumination changes happening along the green–red color direction compared to changes along the yellow–blue direction, meaning that they are less likely to perceive an illumination shift along the yellow–blue direction than along the green–red one. This suggests, the authors argue, that the human visual system compensates better for changes in the blue–yellow directions, which could have consequences for color constancy.

Interim conclusion

Results in this section show a significant effect of the illumination on DeepCC's performance. Higher color constancy indices were observed for illuminations along the yellow–blue direction in CIEL*a*b* color space compared to illuminations falling onto the orthogonal direction. This difference is presumably explained by the model being more accustomed to variations along the daylight locus, the direction along which daylight and natural illuminations, such as the ones used for training,

vary most. The parallel one can draw between our result and observations made in human psychophysics (Aston et al., 2019) implies that the higher variation along the daylight locus may be a cause of similar consequences in humans.

Color constancy throughout DeepCC

There is uncertainty regarding where the neural mechanisms for color constancy would take place in the brain. Many studies emphasize early mechanisms, such as cone adaptation (Lee et al., 1999), or cells sensitive to chromatic contrasts between object and background in V1 (Wachtler et al., 2003). Other have shown that lesions in macaque area V4 also led to impaired color constancy (Wild et al., 1985; see Foster, 2011, for a review). In contrast to biological brains, deep neural networks like DeepCC allow access to the activations of every unit. Taking advantage of this, we added linear readouts to every layer of DeepCC in order to measure at which processing step color constancy emerges.

Methods

In order to apply the Color Constancy Index at different processing stages of DeepCC, we trained readout networks for each one of its five layers (three convolutional and two fully connected). These linear probes (Alain & Bengio, 2016) consisted of very simple, fully connected linear models with 1,600 kernels, 1 per Munsell class. They take as input the ReLU-corrected output of DeepCC's layer they read out, before the maxpooling operation. For example, the readout network of DeepCC's first convolutional layer (RC1) takes as input the output of that layer after the ReLU operation and is trained on the same task as DeepCC, using the same dataset. The parameters of DeepCC's convolutional layer are not updated during this training iteration, only the weights of RC1. RC1 being fully connected and linear, no complex or nonlinear operations are added, and as such, RC1's performance is an indication of the amount of information available in the first convolutional layer of DeepCC.

Results

Figure 6 shows the average CCI obtained for DeepCC readout models. We named these readout models RC1, RC2, RC3, and RF1, RF2, corresponding to the convolutional layers 1, 2, 3, and the fully connected layers 1, 2, respectively. We trained 10 instances of each

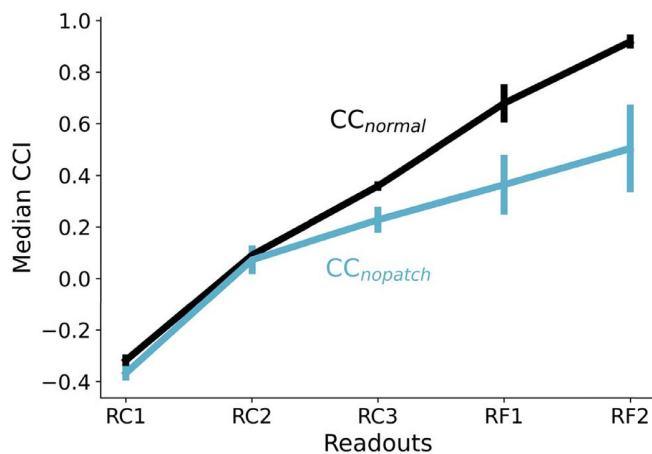


Figure 6. Color Constancy Index (CCI) for the five readout models tested with the *normal* and *no patch* image sets. Each readout takes input from all units of the designated layer: from the three convolutional layers (readouts RC1, RC2, and RC3) to the two fully connected layers (readouts RF1 and RF2). By extension, the value of CCI reflects the degree of color constancy at the different layers of DeepCC.

readout model, one for each instance of the original model. As shown in the plot, the readout models were tested under two conditions: CC_{normal} (black) and $CC_{nopatch}$ (cyan). Error bars are the standard deviation obtained across the 10 training instances. The CCI gradually increases in the normal condition in an almost linear fashion across processing stages, consistently across the 10 models. In the *no patch* condition, CCI follows the normal condition only up to RC2, at which point it continues increasing but at a much lower rate. The difference between the two conditions becomes significant from RC3 onward. Error bars are also larger for the following layers, another indication of the large individual differences between training instances and observed in “Impoverished visual scene”.

Interim conclusion

Contrary to many physiological studies emphasizing the early neural mechanisms for color constancy (Foster, 2011), we found that color constancy seemed to increase steadily throughout DeepCC, both under the normal condition and the no patch condition.

Color representations in DeepCC

We next performed a representational similarity analysis (Kriegeskorte et al., 2008) on unit activations within each layer to probe the models’ internal

representations of colors. We find that although the training objective treated each Munsell value as an entirely distinct class, the DeepCC networks nonetheless learned similarity relationships between the colors that closely resemble their true embeddings in the Munsell space.

Methods

To estimate the similarity between Munsell colors as seen by DeepCC, we computed representational dissimilarity matrices (RDMs) (Kriegeskorte et al., 2008) between the average unit activations per Munsell classes for each layer in the DeepCC networks using the correlation distance as a metric (Aguirre, 2007). Activations were recorded using the evaluation dataset under the normal condition, augmented with additional images under the D65 illumination (i.e., the 330 test Munsell classes under the D65, Y, B, G, and R illuminations). In turn, the RDMs were used as input to a classical multidimensional scaling analysis (MDS) (Cox & Cox, 2008) to compute the underlying dimensions best explaining the previously found dissimilarities. Previous work has shown that the activations of complex deep neural models were able to predict neural response in biological brains (e.g., in mice), even when untrained, that is, with random weights (Cadena et al., 2019). As a control, we thus also performed the same analysis for 10 instances of the deep architecture with random weights, denoted *DeepRand*.

Results

We performed MDS on the RDMs for each of the five layers of DeepCC. Figure 7 shows two-dimensional (2D) representations of the first three dimensions of the MDS results for each layer, tested under the *normal* condition and averaged across all 10 training instances. These three dimensions are the dimensions of maximal variance, in decreasing order. Each column corresponds to one layer. The upper row plots the first and second dimensions, the lower row the second and third. Colored dots correspond to Munsell chips and are displayed using their corresponding sRGB values.

We find that increasingly human-like color dimensions emerge in all layers: Munsells are separated according to their lightness, sometimes also their hue. There is a progression in the way DeepCC represents Munsells: In early layers, many colors are clustered together, especially in the dark regions, rendering them less easily discriminable from one another. This changes in the last two layers, in which colors are more clearly separated from one another. Additionally, the dimensions are easy to interpret. In the first fully

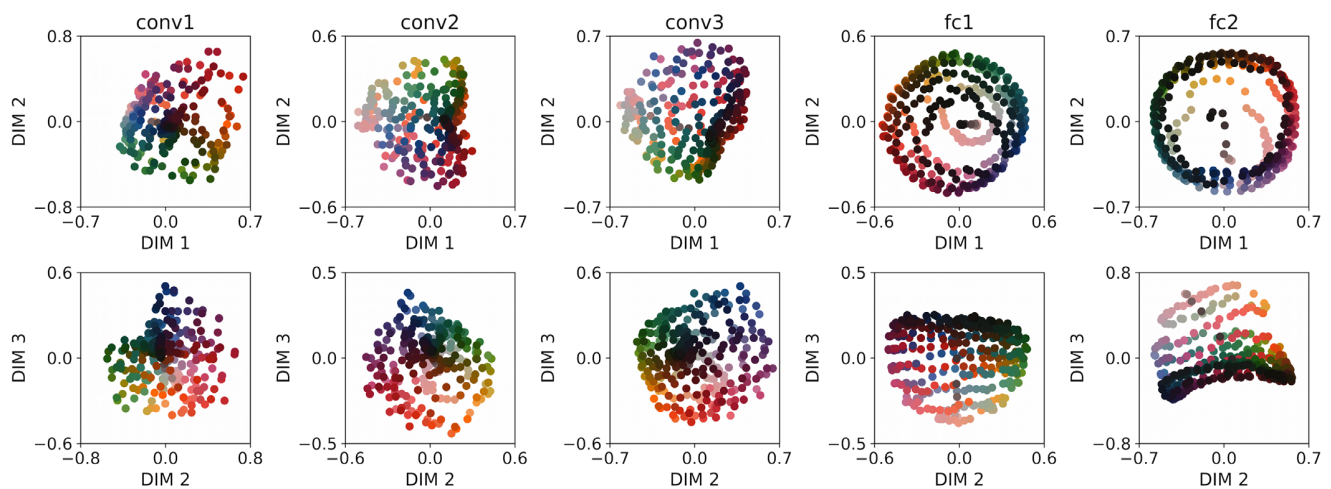


Figure 7. Results of a multidimensional scaling performed on the correlation of Munsell representations for different layers of DeepCC, from convolutional layer 1 (Conv1) to fully connected layer 2 (Fc2). Each column corresponds to one layer, each row to the different dimensions resulting from the MDS: first row, Dimensions 1 and 2 of maximal variance (decreasing order); second row, Dimensions 2 and 3 of maximal variance (decreasing order). Each dot corresponds to one Munsell surface, displayed with its color under the D65 illumination. While Munsell surfaces appear clustered in the early layers, particularly with respect to lightness, a progressive disentanglement in terms of chromaticity and lightness takes place throughout the network.

connected layer, for example, each dimension seems to code for a standard color dimension: Dimensions 1 and 2 for “yellow–blue” and “red–green,” with an almost perfect hue color circle and a radius correlated with saturation, and dimension 3 for lightness.

At each layer, we also computed the cumulative percentage of activation’s variance explained by the three first dimensions given by the MDS, both for DeepCC and DeepRand, the latter consisting of deep instances with random weights. We interestingly found that, although the MDS could potentially yield a much larger number of dimensions, the first three dimensions are enough to explain more than 85% of the variance in most of the layers, for both model types. The highest percentage of explained variance in DeepCC is found for fc1, with 91%. This means that the representations of Munsell are mostly 3D. This result is particularly surprising because fc1 contains the highest number of kernels (250, same as fc2) and thus is more likely to lead to a higher-dimensional latent space. And indeed, the explained variance is lowest at fc1 layers for DeepRand, with 68%.

We next sought to quantify the similarity—or difference—between Munsell representation in our models and their coordinates in a perceptual color space. To do this, we performed a Procrustes analysis (Gower, 1975) to identify the rigid transformation that best mapped the coordinates obtained from the first three MDS dimensions, performed on each layer, to the corresponding coordinates in the Munsell color space. The percentage of explained variance is an indication of the goodness of the mapping: The closer to 100%, the better. As shown in Figure 8, we find that in all

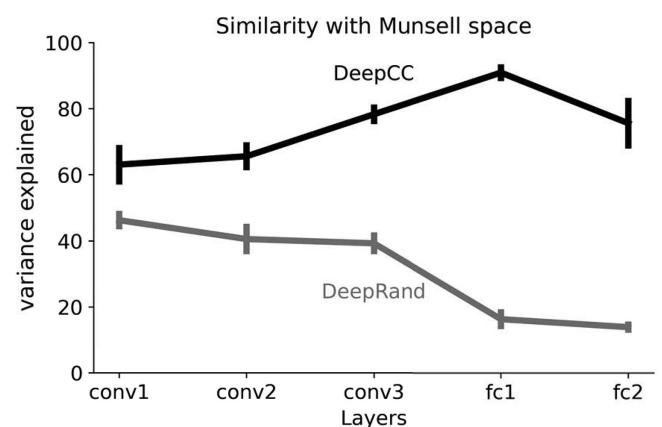


Figure 8. Result of the similarity analysis for all layers of the deep architecture trained on the CC dataset (DeepCC) and with random weights (DeepRand), from convolutional layer 1 (conv1) to fully connected layer 2 (fc2). The figure shows that the highest similarity with Munsell coordinates was found for DeepCC at the first fully connected layer fc1. Additionally, DeepCC always rates higher than DeepRand.

layers, the variance explained by DeepCC progressively increases from 63% in convolutional layer 1 to 91% in fc1. Fc2’s subsequent drop likely reflects the demands of the objective function to deliver categorical outputs. Additionally, DeepCC significantly explains more of the variance than the same architecture with random weights (DeepRand) with a maximal difference in fc1. Indeed, while the variance explained progressively increases for DeepCC, it progressively decreases for

DeepRand. Note the relatively high explained variance for both DeepCC and DeepRand models in the first layer conv1. It is likely a consequence of the input space: Performing the Procrustes analysis from the Munsell chromaticities in LMS space (input) to Munsell coordinates yields a percentage of accounted variance of 66%, very close to the 63% found in DeepCC's conv1.

It is important to note that this finding is nontrivial and cannot be explained solely by the loss function we used. During training, the networks were never taught similarity relationships between Munsell color values. Rather, the error signal was the same whether the models wrongly selected a Munsell close to or far from the correct one in color space. Theoretically, a model could reach a high accuracy and not learn human-like similarities between the Munsell colors. And indeed, as reported below, other architectures trained and tested following the same procedures represent colors in a different manner.

Qualitatively similar results were also obtained when using a L2 norm instead of the correlation metric. Additionally, we also performed this analysis using the CIEL*a*b* coordinates as a reference for the Procrustes analysis and found extremely similar results as with the Munsell coordinates. We excluded these results from the figures to avoid redundancy.

Interim conclusion

Similarly to the increasing CCI observed throughout the network in the previous section, the representational analysis also uncovered a progression in the way Munsell colors are represented within the model's layers. Visually, we could observe a progressive disentanglement of Munsell colors with increasing layer depth. More important, the representation of color also progressively increased their resemblance with human perception, peaking at FC1, where there was a very high correspondence to the Munsell perceptual color space. This was quantitatively confirmed using a similarity analysis, where it was found that the representational distances and dimensions between Munsell values, in the penultimate layer in particular, matched very well the human perceptual distances and dimensions found empirically in previous psychophysical studies. The subsequent drop found in the last layer likely reflects the demands of the objective function to deliver categorical outputs.

Standard and custom architectures

We observed in the previous section that DeepCC represents Munsell colors following color dimensions

found empirically to be perceptually relevant for humans. Is this a special feature of this architecture (i.e., would different architectures learn different representations)? If yes, it would be strong evidence that there is not one globally optimal system of representations to solve color classification. To answer this question, we trained and evaluated several other standard deep learning architectures.

Methods

Architectures

For the sake of comparison, we also trained three standard, high-performance deep learning models on the CC dataset: VGG-11 (Simonyan & Zisserman, 2014), MobileNet (Howard et al., 2017), and ResNet-50 (He et al., 2016). All of these architectures have specific features that make them significantly different from one another. These standard architectures, however, are relatively large and complex compared to the DeepCC architecture. While DeepCC only has 676 kernels (outside of the classification layer's 1,600) and 3.6 million interconnections between units, all three others have more than 13,000 kernels, the highest being ResNet-50 with almost 54,000. In order to allow some comparison with networks of a size more similar to DeepCC, we additionally devised another, shallower model. It consisted of a custom ResNet architecture, generated thanks to a ResNet bottleneck architecture generator (available in [github](#)⁷). To distinguish it from ResNet-50, we will call this architecture *ResCC*. It has three layers, each with three, one, and two bottleneck blocks, respectively. The first layer starts with 16 kernels, layer 2 with 32, and layer 3 with 64. Including the kernels within the bottleneck layers, it reaches 3,424 kernels and 0.6 million interconnections. Similarly to DeepCC, where we trained 10 instances, 6 independent training instances of ResCC were trained for further analysis.

Results

We evaluated each one of the DNN architectures on the *normal*, *no patch*, *wrong patch*, and *wrong back* conditions. Here, for the sake of simplicity, we show only a summary of the results through a table with the distributions' medians. Table 1 shows the median measurements of performance obtained for all architectures under those conditions, with the results obtained for DeepCC as a reminder on the last column. MobileNet, VGG-net, ResNet-50, and ResCC all showed higher performance than DeepCC in all conditions. Interestingly, there was almost no difference in performance for every model other than

Model Nb param	MobileNet 4.3 M		VGG-11 135.3 M		ResNet-50 29.8 M		ResCC 0.6 M		DeepCC (ref ConvNet) 3.6 M	
	<i>normal</i>	<i>no patch</i>	<i>normal</i>	<i>no patch</i>	<i>normal</i>	<i>no patch</i>	<i>normal</i>	<i>no patch</i>	<i>normal</i>	<i>no patch</i>
Top-1	95	95	100	100	100	95	85	80	75	40
Top-5	100	100	100	100	100	100	100	100	100	90
Muns ³	100	100	100	100	100	100	100	100	100	95
ΔE	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	3.3
CCI	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	0.6

Condition	<i>wrg patch</i>	<i>wrg back</i>	<i>wrg patch</i>	<i>wrg back</i>	<i>wrg patch</i>	<i>wrg back</i>	<i>wrg patch</i>	<i>wrg back</i>	<i>wrg patch</i>	<i>wrg back</i>
Top-1	85	0.0	92.5	0.0	87.5	0.0	80	0.0	25	0.0
Top-5	100	25	100	25	100	25	100	25	65	25
Muns ³	100	70	100	75	100	70	100	65	85	65
ΔE	0.0	10.6	0.0	10.0	0.0	10.34	0.0	11.2	5.6	11.4
CCI	1	-0.34	1.0	-0.28	1	-0.32	1	-0.4	0.23	-0.46

Table 1. Median values found for all measures and all models under the normal, no patch, wrong patch, and wrong back conditions. All models show higher performances than DeepCC in all test sets. Interestingly, except DeepCC, none of the models are sensitive to the absence (*no patch*) or incongruence (*wrong patch*) of the colored patches in the background. This suggests that in contrast to DeepCC, these other models barely rely on the constant colored patches in the background to perform color constancy. The sharp drop in performance for the *wrong back* condition, however, suggests that like DeepCC, all other models also rely on the contextual cues surrounding the floating object to perform color constancy.

DeepCC between the *normal*, *no patch*, and *wrong patch* conditions. All models, however, have shown a significant loss in accuracy for the *wrong back* condition, suggesting that all tested models rely heavily on cues in the background to perform their task.

Up to now, standard networks and ResCC essentially shared the same characteristics as DeepCC: While they outperformed the classical approaches to color constancy, such as Gray World (cf. “Comparison with classical approaches”) under the *normal* condition, they failed to account for the illumination color under the *wrong back* condition (cf. Figure 3), as indeed essentially any observer would. Additionally, we found they also show a significant effect of the illumination on the Color Constancy Index, with higher performance for the Yellow and Blue illuminations than for the Green and Red illuminations (not shown).

However, when it came to the analysis of Munsell representations within the latent layers, they all exhibited a very different picture from DeepCC: Munsell chips did not appear to be differentiated following human-like color dimensions. As in the previous section, we performed multidimensional scaling on the RDMs for each layer of each architecture, followed by a Procrustes analysis using Munsell coordinates as a reference space. Across all architectures, the highest percentage of explained variance resulting from the Procrustes analysis was 53%. It was obtained for the VGG-11 architecture’s fourth layer and stands substantially below the 91% explained variance of DeepCC’s penultimate layer.

As an example, we show in Figure 9 the results of the MDS analysis averaged over the ResCC instances. We can observe that none of the three layers visibly separate Munsell colors along human-like perceptual dimensions like Hue, Lightness, or Chroma. This is particularly true for Layer 3. For this last layer, the first three dimensions of the MDS account for only 54% of the dissimilarity between Munsell representations, meaning that Munsell discrimination took place in a space with more than three dimensions.

This observation is further confirmed by Figure 10. The variance explained by the best fit for mapping Munsell representations in ResCC layers onto the Munsell coordinates was always lower than for DeepCC, meaning that ResCC distinguished Munsell values using color dimensions different from the ones defined by human color perception, contrary to DeepCC. Additionally, the low percentage of variance explained by the same architecture but with random weights (ResRand) suggests that the architecture is the major factor for this difference. Interestingly, this result correlates with a recent observation (Zhou et al., 2018) that ResNet architectures lead to much less interpretable representations than more conventional convolutional architectures like AlexNet and VGG Nets.

Interim conclusion

The results of our comparisons with other architectures show that if performance was our only

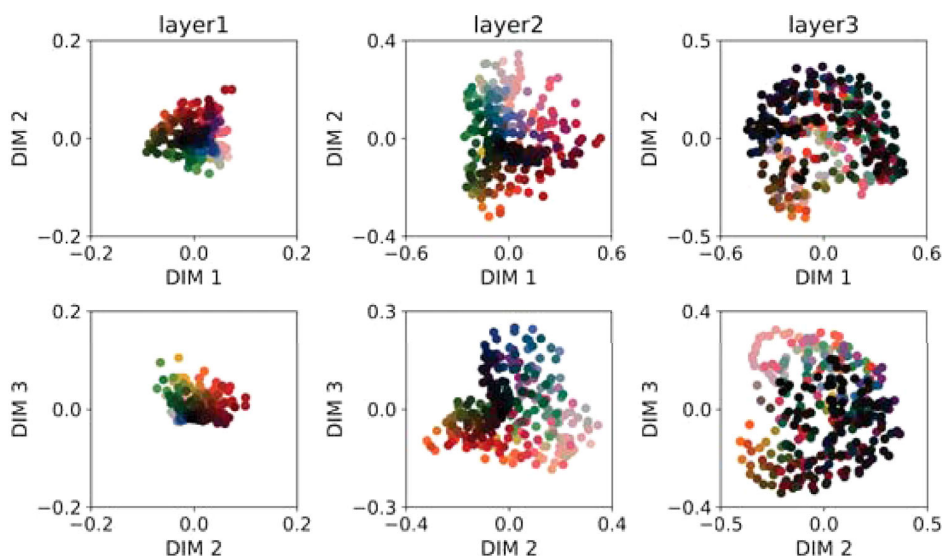


Figure 9. Results of a multidimensional scaling performed on the correlation distance of Munsell representations for different layers of ResCC. Compared to DeepCC (cf. Figure 7), ResCC does not seem to classify Munsells following the same dimensions as those defined by human perception, particularly in Layer 3.

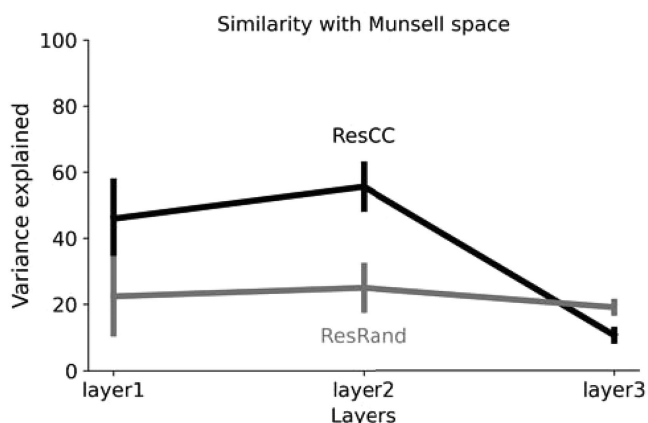


Figure 10. Results of the Procrustes analysis for the Res architecture trained on the CC dataset (black) and with random weights (gray). The analysis was performed on the outcomes of the multidimensional scaling at different layers using Munsell space as reference coordinates. The variance explained for ResCC was consistently lower than for DeepCC throughout its layers, meaning that ResCC discriminate Munsells following color dimensions dissimilar to those defined by human color perception. The fact that the Res architectures systematically rate lower both when trained and with random weights suggests that the major factor for this difference is the architecture.

goal, many architectures other than *Deep* could have been used to solve the Munsell classification task and indeed achieved superior performance. The similarity analysis we used, however, showed that other architectures, such as ResCC, seemingly differentiate between Munsell colors according to color dimensions

very different from those empirically found for human perception, contrary to DeepCC.

This last observation is thus evidence that there is not one globally optimal system of representations to which all networks tend to converge. Rather, multiple possible systems of representations deliver good performance at the task, the one shared between humans and DeepCC being one of them. This result also emphasizes the need for careful examination when it comes to selecting a DNN architecture for a given task. While at first sight, ResCC might have seemed a better choice for our tasks (highest performance and few parameters), the analysis of the Munsell representations shows that DeepCC presents characteristics more similar to human color discrimination. This last point suggests that DeepCC is thus potentially a better candidate for modeling human discrimination of Munsell color surfaces. It also emphasizes the need to develop further methods and strategies to analyze and understand the features learned by different architectures.

Discussion

We have trained deep neural models for the classification of Munsell chips under varying natural illuminations using 3D spectral renderings. We found that our models did learn to discount the illumination’s contribution to the color appearance of the surface, hence learning color constancy. When manipulating the contextual cues within the scene, in such the way that these cues no longer gave information about the illumination shining on the object, our models were no

longer color constant, performing exactly at the same level as our control network Deep65, trained under our reference illumination D65 only. Additionally, we found that despite using the same training procedure, different architectures led to very different color representations of Munsell chips within their layers: One network, DeepCC, developed color representations very similar to the Munsell chips coordinates, while the other models did not.

In the following, we discuss how these findings relate to human color constancy and color vision. We also discuss the opportunities offered by the combination of deep learning and computer graphics for studying properties of human vision such as color constancy.

Deep neural networks for biological color vision

We find that as a result of training, the deep neural network models became similar to humans in several respects: They classified Munsell colors largely independently of changes in illumination, thus learning color constancy. They used contextual information to do so: When we manipulate the scene elements to provide incorrect information about the illuminant, the models perform at the same level as a non-color constant model, meaning that they are no longer able to discount the illuminant. Likewise, numerous previous studies have shown that humans also rely on context to achieve color constancy (Kraft & Brainard, 1999; Kraft et al., 2002; Yang & Maloney, 2001). One model, DeepCC, was also sensitive to the cues provided by the constant color patches in the background. Additionally, the models showed higher degrees of color constancy for illuminations along the daylight locus than for illuminations along the orthogonal color direction. This also correlates with the lower sensitivity to illuminant change along the daylight locus observed in humans (Aston et al., 2019).

In addition, our analysis of the networks' inner representations revealed that DeepCC represented surface colors using dimensions similar to the Munsell and CIELab spaces, which are based on human perception. This similarity seems to be the exception rather than the rule, as other architectures like ResCC, represented color in a different way, despite achieving similar or superior performance on the objective. The observation that one architecture learned human-like features and not the other hints at architectural influences shaping human color perception. Better understanding these architectural influences—and how they relate to the architecture of primate visual systems—may help us understand human color vision in the future.

It remains unclear what exact mechanisms within the networks are responsible for achieving color constancy, and to what extent these are comparable to neural mechanisms found in biological visual systems. Some

possibilities, however, are more likely than others. One mechanism thought to significantly contribute to primate color constancy is *adaptation* (Foster, 2011) present as early as at the retinal level (Lee et al., 1999). Adaptation, however, is commonly accepted to require either neural feedback from recurrent interactions within the network (del Mar Quiroga et al., 2016), or an intrinsic suppression mechanism in the neuron itself (Whitmire & Stanley, 2016), neither of which are explicitly implemented in the architectures used here: They are feedforward networks with simple ReLU activation functions. Recently, Vinken et al. have implemented an exponentially decaying intrinsic adaptation state within each unit of a feedforward CNN architecture (Vinken et al., 2020). They were successfully able to reproduce neurophysiological and perceptual properties of adaptation. Their proposed architecture could thus have the potential to learn the adaptation mechanism for color constancy if trained on our task. Nevertheless, the fact that networks can achieve color constancy without such adaptation mechanisms suggests that in humans, the primary role of adaptation may be in controlling sensitivity given limited dynamic range and noise, rather than surface reflectance estimation per se. Another mechanism thought to contribute to color constancy in biological brains is *cell response invariance*, or the tendency of certain cells to be largely sensitive to chromatic contrasts between target and background (Foster, 2011), both at the early stages of the visual system (Wachtler et al., 2003) and the later stages (Kusunoki et al., 2006). Recent studies have shown that kernels sensitive to chromatic contrasts can be found in the early and late convolutional layers of feedforward CNNs trained for object recognition (Flachot & Gegenfurtner, 2018, 2021; Harris et al., 2019).

3D-rendered dataset for color constancy

Unfortunately, large datasets consisting of numerous photographs of real, complex scenes with controlled conditions suitable for training deep neural networks from scratch on color constancy tasks do not yet exist. The popular ImageNet (Deng et al., 2009), for instance, consists of millions of natural images but taken from noncalibrated cameras, presumably white-balanced. The ColorChecker dataset (Gehler et al., 2008) has the opposite characteristic: It presents precise and well-calibrated complex images, but less than 1,000 of them. Large hyperspectral datasets of natural scenes at different times of the day would be optimal, of course, but the difficulty of controlled hyperspectral captures is such that most datasets count a few hundreds of images at most (Vazquez-Corral et al., 2009; Nascimento et al., 2016).

Some challenges remain, however, such as the efficient creation of convincing outdoor scenes. It

is possible that reproducing the statistics of more complex, naturalistic scenes would contribute toward greater robustness of DNNs to scene changes and perhaps allow the emergence of higher features of color vision, such as color categories (Witzel & Gegenfurtner, 2018; Parraga & Akbarinia, 2016).

Implications for color constancy in general

Our results have several implications for color constancy in general, independent of whether we believe that DNNs are a good model of human color constancy. First, we trained networks to extract the surface color more accurately than a perfect global von Kries correction. This implies that a global illumination correction is not the optimal solution to the color constancy problem, even in a situation with a single illumination color. This may guide future computer vision and image-processing work that aims to extract object properties rather than color-correcting images. Second, we confirm earlier suspicions that the prior distribution over illuminations causes the better performance of humans along the daylight axis, as employing a naturalistic range of illuminations was sufficient to cause our networks to have this bias as well. Third, our finding that network architectures like ResCC can achieve outstanding color constancy performance despite not reproducing human perceptual color similarity representations suggests that these representations are not necessary for color constancy. Although perceptual color spaces presumably have many advantages for human color vision, our findings do not support the notion that they are specifically optimized for color constancy—at least in the class of images we investigated. An interesting direction for future research would be to train networks explicitly on perceptual color representations and test how this improves performance at other tasks. This would potentially provide answers to the teleological question of why human color space is shaped as it is (DiCarlo et al., 2012).

Conclusion

In this study, we approached color constancy as a surface reflectance classification task under varying illumination using deep neural networks. This methodology closely mimics what humans do on a daily basis and differs from the common approach to computational modeling of color constancy that mainly focuses on the illumination estimation and image correction. We then devised a set of testing conditions to thoroughly evaluate our models and compare them to previous human behavioural studies. We found that similarly to humans, all models heavily relied on

contextual cues to solve color constancy and show the same bias toward illuminations along the daylight locus as humans. However, a similarity analysis on the activation patterns within the deep latent layers of the trained models showed significant differences in the way they represented color surfaces. Only one convolutional network, DeepCC, learned to discriminate colored surfaces following similar dimensions to those used by humans. This suggests that in computational models of human color constancy, the highest performance alone might not be the best metric to measure fidelity of a model to human color representations. This is in line with reports in object classification, where lower performance networks may better correlate with human brain recordings and behavioral measurements (Kubilius et al., 2019; Geirhos et al., 2020b).

Keywords: color constancy, deep learning, spectral renderings, color classification

Acknowledgments

The authors thank their colleagues in Giessen for their support and useful discussions. In particular, we thank Guido Maiello for his statistical advice, Philipp Schmidt for making the meshes dataset accessible, and Robert Ennis and Florian Bayer for their discussions and software support. We also thank Raquel Gil for her mathematical expertise and Christoph Witzel for his color expertise. In Tübingen, we are grateful to Bernhard Lang for helping us to get started with spectral renderings and Uli Wannek for technical support.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 222641018-SFB/TRR 135 TPs C1 and C2 and by “The Adaptive Mind,” funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art. HHS was funded, in part, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) through research fellowship SCHU 3351/1-1. FAW was funded, in part, by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany’s Excellence Strategy—EXC number 2064/1—Project number 390727645. K.R.G. received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement no. 884116).

Commercial relationships: none.

Corresponding author: Alban Flachot.

Email: flachot.alban@gmail.com.

Address: Abteilung Allgemeine Psychologie, Otto-Behagel-Str. 10F, 35394, Giessen, Germany.

Footnotes

- ¹See <https://colorconstancy.com/evaluation/datasets/> for a review.
²<http://www.mitsuba-renderer.org/>.
³http://www.cs.joensuu.fi/~spectral/databases/download/munsell_spec_glossy_all.htm.
⁴<https://evermotion.org/shop>.
⁵<https://www.dropbox.com/sh/gz52alcoue9ew6w/AADYg3EJZD9bRLb04aifByNJa?dl=0>.
⁶https://github.com/AlbanFlachot/color_constancy.
⁷<https://github.com/ArashAkbarinia/kernelphysiology>.

References

- Affi, M., & Brown, M. S. (2019). Sensor-independent illumination estimation for DNN models. *arXiv preprint arXiv:1912.06888*.
- Aguirre, G. K. (2007). Continuous carry-over designs for fMRI. *Neuroimage*, 35(4), 1480–1494.
- Akbarinia, A., & Gil-Rodríguez, R. (2020). Deciphering image contrast in object classification deep networks. *Vision Research*, 173, 61–76.
- Akbarinia, A., & Parraga, C. A. (2017). Colour constancy beyond the classical receptive field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(9), 2081–2094.
- Alain, G., & Bengio, Y. (2016). Understanding intermediate layers using linear classifier probes. *arXiv preprint arXiv:1610.01644*.
- Arend, L., & Reeves, A. (1986). Simultaneous color constancy. *Journal of the Optical Society of America A*, 3(10), 1743–1751.
- Aston, S., Radonjić, A., Brainard, D. H., & Hurlbert, A. C. (2019). Illumination discrimination for chromatically biased illuminations: Implications for color constancy. *Journal of Vision*, 19(3), 15, <https://doi.org/10.1167/19.3.15>.
- Bergmann, S., Mohammadikaji, M., Irgenfried, S., Wörn, H., Beyerer, J., & Dachsbacher, C. (2016). A phenomenological approach to integrating Gaussian beam properties and speckle into a physically-based renderer. In *Proceedings of the Conference on Vision, Modeling and Visualization* (pp. 179–186).
- Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Berkeley, California: University of California Press.
- Bianco, S., & Cusano, C. (2019). Quasi-unsupervised color constancy. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12212–12221). IEEE.
- Bianco, S., Cusano, C., & Schettini, R. (2015). Color constancy using CNNs. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops* (pp. 81–89). IEEE.
- Brainard, D. H., Kraft, J. M., & Longere, P. (2003). Color constancy: Developing empirical tests of computational models. In R. Mausfeld, & D. Heyer (Eds.), *Colour Perception: Mind and the physical world* (pp. 307–334). Oxford University Press.
- Brendel, W., Rauber, J., & Bethge, M. (2017). Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. *arXiv preprint arXiv:1712.04248*.
- Butler, D. J., Wulff, J., Stanley, G. B., & Black, M. J. (2012). A naturalistic open source movie for optical flow evaluation. In *European Conference on Computer Vision (ECCV)* (pp. 611–625). Berlin, Heidelberg: Springer.
- Cadena, S. A., Sinz, F. H., Muhammad, T., Froudarakis, E., Cobos, E., Walker, E. Y., . . . Ecker, A. S. (2019). How well do deep neural networks trained on object recognition characterize the mouse visual system? *Real Neurons & Hidden Units NeurIPS Workshop*.
- Chiao, C.-C., Cronin, T. W., & Osorio, D. (2000). Color signals in natural scenes: Characteristics of reflectance spectra and effects of natural illuminants. *Journal of the Optical Society of America A*, 17(2), 218–224.
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Deep neural networks predict hierarchical spatiotemporal dynamics of human visual object recognition. *arXiv preprint arXiv:1601.02970*.
- Cox, M. A., & Cox, T. F. (2008). Multidimensional scaling. In *Handbook of data visualization* (pp. 315–347). Berlin, Heidelberg: Springer.
- Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255). IEEE.
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–434.
- Dosovitskiy, A., Fischer, P., Ilg, E., Häusser, P., Hazırbaş, C., Golkov, V., . . . Brox, T. (2015). FlowNet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 2758–2766).
- Flachot, A. (2019). Extensions of a linear model of surface reflectance. *PsyArXiv*.
- Flachot, A., & Gegenfurtner, K. R. (2018). Processing of chromatic information in a deep convolutional neural network. *Journal of the Optical Society of America A*, 35(4), B334–B346.

- Flachot, A., & Gegenfurtner, K. R. (2021). Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks. *Vision Research*, 182, 89–100.
- Flachot, A., Provenzi, E., & O'Regan, J. K. (2016). An illuminant-independent analysis of reflectance as sensed by humans, and its applicability to computer vision. In *Proceedings of the 6th European Workshop on Visual Information Processing (EUVIP)*, Marseille, France (pp. 1–6).
- Foster, D. (2011). Color constancy. *Vision Research*, 51, 674–700.
- Gao, S.-B., Yang, K.-F., Li, C.-Y., & Li, Y.-J. (2015). Color constancy using double-opponency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(10), 1973–1985.
- Gegenfurtner, K. R., & Kiper, D. C. (2003). Color vision. *Annual Review of Neuroscience*, 26(1), 181–206.
- Gehler, P. V., Rother, C., Blake, A., Minka, T., & Sharp, T. (2008). Bayesian color constancy revisited. In *2008 IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–8). IEEE.
- Geirhos, R., Jacobsen, J. H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M., . . . Wichmann, F. A. (2020). Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11), 665–673.
- Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: Object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.
- Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*, 33, 13890–13902.
- Geirhos, R., Temme, C. R. M., Rauber, J., Schtt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 31, 7549–7561.
- Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*.
- Gower, J. C. (1975). Generalized Procrustes analysis. *Psychometrika*, 40(1), 33–51.
- Granzier, J. J., & Gegenfurtner, K. R. (2012). Effects of memory colour on colour constancy for unknown coloured objects. *i-Perception*, 3(3), 190–215.
- Guarnera, D. Y., Guarnera, G. C., Toscani, M., Glencross, M., Li, B., Hardeberg, J. Y., . . . Gegenfurtner, K. R. (2018). Perceptually validated cross-renderer analytical BRDF parameter remapping. *IEEE Transactions on Visualization and Computer Graphics*, 26(6), 2258–2272.
- Güçlü, U., & van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27), 10005–10014.
- Harris, E., Mihai, D., & Hare, J. (2019). Spatial and colour opponency in anatomically constrained deep networks. *arXiv preprint arXiv:1910.11086*.
- He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 770–778). IEEE.
- Howard, A. G., Zhu, M., Chen, B., Kalenichenko, D., Wang, W., Weyand, T., . . . Adam, H. (2017). Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*.
- Hu, Y., Wang, B., & Lin, S. (2017). Fc4: Fully convolutional color constancy with confidence-weighted pooling. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4085–4094). IEEE.
- Ilg, E., Saikia, T., Keuper, M., & Brox, T. (2018). Occlusions, motion and depth boundaries with a generic network for disparity, optical flow or scene flow estimation. In *Proceedings of the European Conference on Computer Vision (ECCV)* (pp. 614–630).
- Judd, D. B., MacAdam, D. L., Wyszecki, G., Budde, H. W., Condit, H. R., Henderson, S. T., . . . Simonds, J. L. (1964). Spectral distribution of typical daylight as a function of correlated color temperature. *Journal of the Optical Society of America A*, 54(8), 1031–1040.
- Kalenova, D., Toivanen, P., & Bochko, V. (2005). Color differences in a spectral space. *Journal of Imaging Science and Technology*, 49(4), 404–409.
- Kingma, P., & Ba, J. (2015). Adam: A method for stochastic optimization, arxiv (2014). *arXiv preprint arXiv:1412.6980*, 106.
- Kraft, J. M., & Brainard, D. H. (1999). Mechanisms of color constancy under nearly natural viewing. *Proceedings of the National Academy of Sciences*, 96(1), 307–312.
- Kraft, J. M., Maloney, S. I., & Brainard, D. H. (2002). Surfaceilluminant ambiguity and color constancy: Effects of scene complexity and depth cues. *Perception*, 31(2), 247–263.
- Kriegeskorte, N., Mur, M., & Bandettini, P. A. (2008). Representational similarity analysis-connecting

- the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems* (Vol. 25, pp. 1097–1105).
- Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N. J., . . . DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems* (pp. 12805–12816).
- Kusunoki, M., Moutoussis, K., & Zeki, S. (2006). Effect of background colors on the tuning of color-selective cells in monkey area v4. *Journal of Neurophysiology*, 95(5), 3047–3059.
- Land, E. H. (1964). The retinex. *American Scientist*, 52(2), 247–264.
- Land, E. H. (1977). The retinex theory of color vision. *Scientific American*, 237(6), 108–129.
- LeCun, Y., & Bengio, Y. (1995). Convolutional networks for images, speech, and time series. *The Handbook of Brain Theory and Neural Networks*, 3361(10), 1995.
- LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11), 2278–2324.
- Lee, B. B., Dacey, D. M., Smith, V. C., & Pokorny, J. (1999). Horizontal cells reveal cone type-specific adaptation in primate retina. *Proceedings of the National Academy of Sciences*, 96(25), 14611–14616.
- Logvinenko, A., Funt, B., Mirzaei, H., & Tokunaga, R. (2015). Rethinking colour constancy. *PLoS One*, 10(9), e0135029.
- Lou, Z., Gevers, T., Hu, N., & Lucassen, M. P. et al. (2015). Color constancy by deep learning. In *Proceedings of BMVC* (pp. 76–1).
- Maloney, L. T., & Wandell, B. A. (1986). Color constancy: A method for recovering surface spectral reflectance. *Journal of the Optical Society of America A*, 3(1), 29–33.
- Mar Quiroga, M. del, Morris, A. P., & Krekelberg, B. (2016). Adaptation without plasticity. *Cell Reports*, 17(1), 58–68.
- Mokrzycki, W., & Tatol, M. (2011). Colour difference_e-a survey. *Machine Graphics and Vision*, 20(4), 383–411.
- Munsell, A. H. (1912). A pigment color system and notation. *The American Journal of Psychology*, 23(2), 236–244.
- Nascimento, S. M., Amano, K., & Foster, D. H. (2016). Spatial distributions of local illumination color in natural scenes. *Vision Research*, 120, 39–44.
- Nickerson, D. (1940). History of the Munsell color system and its scientific application. *Journal of the Optical Society of America*, 30(12), 575–586.
- Ohno, Y. (2000). CIE fundamentals for color measurements. In *NIP & Digital Fabrication Conference* (Vol. 2000, pp. 540–545). Society for Imaging Science and Technology.
- Olkkonen, M., Hansen, T., & Gegenfurtner, K. R. (2008). Color appearance of familiar objects: Effects of object shape, texture, and illumination changes. *Journal of Vision*, 8(5), 13–17, <https://doi.org/10.1167/8.5.13>.
- Parraga, C. A., & Akbarinia, A. (2016). Nice: A computational solution to close the gap from colour perception to colour categorization. *PLoS One*, 11(3), e0149538.
- Pearce, B., Crichton, S., Mackiewicz, M., Finlayson, G. D., & Hurlbert, A. (2014). Chromatic illumination discrimination ability reveals that human colour constancy is optimised for blue daylight illuminations. *PLoS One*, 9(2), e87989.
- Pharr, M., Jakob, W., & Humphreys, G. (2016). *Physically based rendering: From theory to implementation*. Morgan Kaufmann, https://books.google.de/books?hl=en&lr=&id=iNMVBQAAQBAJ&oi=fnd&pg=PP1&dq=Physically+based+rendering:+From+theory+to+implementation&ots=iwxiBhWiOK&sig=ghbSKaqb5t_pqru7D7nFN9RFJgs#v=onepage&q=Physically%20based%20rendering%3A%20From%20theory%20to%20implementation&f=false.
- Philipona, D. L., & O'Regan, J. K. (2006). Color naming, unique hues, and hue cancellation predicted from singularities in reflection properties. *Visual Neuroscience*, 23(3–4), 331–339.
- Prokott, K. E., Tamura, H., & Fleming, R. W. (2021). Gloss perception: Searching for a deep neural network that behaves like humans. *Journal of Vision*, 21(12), 14–14, <https://doi.org/10.1167/jov.21.12.14>.
- Rafegas, I., & Vanrell, M. (2018). Color encoding in biologically-inspired convolutional neural networks. *Vision Research*, 151, 7–17.
- Shi, W., Loy, C. C., & Tang, X. (2016). Deep specialized network for illuminant estimation. In *European Conference on Computer Vision* (pp. 371–387). Springer, Cham.
- Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.
- Stockman, A., & Sharpe, L. T. (2000). The spectral sensitivities of the middle-and long-wavelength-

- sensitive cones derived from measurements in observers of known genotype. *Vision Research*, 40(13), 1711–1737.
- Storrs, K. R., Anderson, B. L., & Fleming, R. W. (2021). Unsupervised learning predicts human perception and misperception of gloss. *Nature Human Behaviour*, 5(10), 1402–1417.
- Vazquez, J., Párraga, C. A., Vanrell, M., & Baldrich, R. (2009). Color constancy algorithms: Psychophysical evaluation on a new dataset. *Journal of Imaging Science and Technology*, 1(3), 1.
- Vinken, K., Boix, X., & Kreiman, G. (2020). Incorporating intrinsic suppression in deep neural networks captures dynamics of adaptation in neurophysiology and perception. *Science Advances*, 6(42), eabd4205.
- von Kries, J. (1902). Chromatic adaptation. *Festschrift der Albrecht-Ludwigs-Universität*, 135, 145–158.
- Wachtler, T., Sejnowski, T. J., & Albright, T. D. (2003). Representation of color stimuli in awake macaque primary visual cortex. *Neuron*, 37(4), 681–691.
- Weiss, D., Witzel, C., & Gegenfurtner, K. (2017). Determinants of colour constancy and the blue bias. *i-Perception*, 8(6), 2041669517739635.
- Whitmire, C. J., & Stanley, G. B. (2016). Rapid sensory adaptation redux: A circuit perspective. *Neuron*, 92(2), 298–315.
- Wild, H., Butler, S., Carden, D., & Kulikowski, J. (1985). Primate cortical area v4 important for colour constancy but not wavelength discrimination. *Nature*, 313(5998), 133–135.
- Witzel, C. (2019). Misconceptions about colour categories. *Review of Philosophy and Psychology*, 10(3), 499–540.
- Witzel, C., & Gegenfurtner, K. R. (2018). Color perception: Objects, constancy, and categories. *Annual Review of Vision Science*, 4, 475–499.
- Worthey, J. A., & Brill, M. H. (1986). Heuristic analysis of von Kries color constancy. *Journal of the Optical Society of America A*, 3(10), 1708–1712.
- Yang, J. N., & Maloney, L. T. (2001). Illuminant cues in surface color perception: Tests of three candidate cues. *Vision Research*, 41(20), 2581–2600.
- Yosinski, J., Clune, J., Nguyen, A., Fuchs, T., & Lipson, H. (2015). Understanding neural networks through deep visualization. *arXiv preprint arXiv:1506.06579*.
- Zeiler, M. D., & Fergus, R. (2014). Visualizing and understanding convolutional networks. In *European Conference on Computer Vision* (pp. 818–833). IEEE.
- Zhou, B., Bau, D., Oliva, A., & Torralba, A. (2018). Interpreting deep visual representations via network dissection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(9), 2131–2145.

2. PUBLICATIONS

2.0.4 Complete list of publications

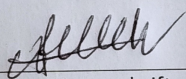
- **Flachot, A.** and Gegenfurtner, K.R., 2018. Processing of chromatic information in a deep convolutional neural network. *JOSA A*, 35(4), pp.B334-B346.
- **Flachot, A.** and Gegenfurtner, K.R., 2021. Color for object recognition: Hue and chroma sensitivity in the deep features of convolutional neural networks. *Vision Research*, 182, pp.89-100.
- **Flachot, A.**, Akbarinia, A., Schütt, H.H., Fleming, R.W., Wichmann, F.A. and Gegenfurtner, K.R., 2020. Deep Neural Models for color discrimination and color constancy. Accepted, *Journal of Vision*
- Akbarinia, A., Gil-Rodríguez, R., **Flachot, A.** and Toscani, M., 2020. The Utility of Decorrelating Colour Spaces in Vector Quantised Variational Autoencoders. In prep. arXiv preprint arXiv:2009.14487.
- de Vries, J.P., Akbarinia, A., **Flachot, A.** and Gegenfurtner, K.R., 2021. Emergent Color Categorization in a Neural Network trained for Object Recognition. Under review, *eLife*.

2.0.5 Selbstständigkeitserklärung

Selbstständigkeitserklärung

Hiermit versichere ich, die vorgelegte Thesis selbstständig und ohne unerlaubte fremde Hilfe und nur mit den Hilfen angefertigt zu haben, die ich in der Thesis angegeben habe. Alle Textstellen, die wörtlich oder sinngemäß aus veröffentlichten Schriften entnommen sind, und alle Angaben die auf mündlichen Auskünften beruhen, sind als solche kenntlich gemacht. Bei den von mir durchgeführten und in der Thesis erwähnten Untersuchungen habe ich die Grundsätze guter wissenschaftlicher Praxis, wie sie in der ‚Satzung der Justus-Liebig-Universität zur Sicherung guter wissenschaftlicher Praxis‘ niedergelegt sind, eingehalten. Gemäß § 25 Abs. 6 der Allgemeinen Bestimmungen für modularisierte Studiengänge dulde ich eine Überprüfung der Thesis mittels Anti-Plagiatssoftware.

20.02.2022
Datum


Unterschrift