Vision Research 182 (2021) 89–100

Contents lists available at ScienceDirect

Vision Research

journal homepage: www.elsevier.com/locate/visres

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 selectitivity
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

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 normalizations 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:

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

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.

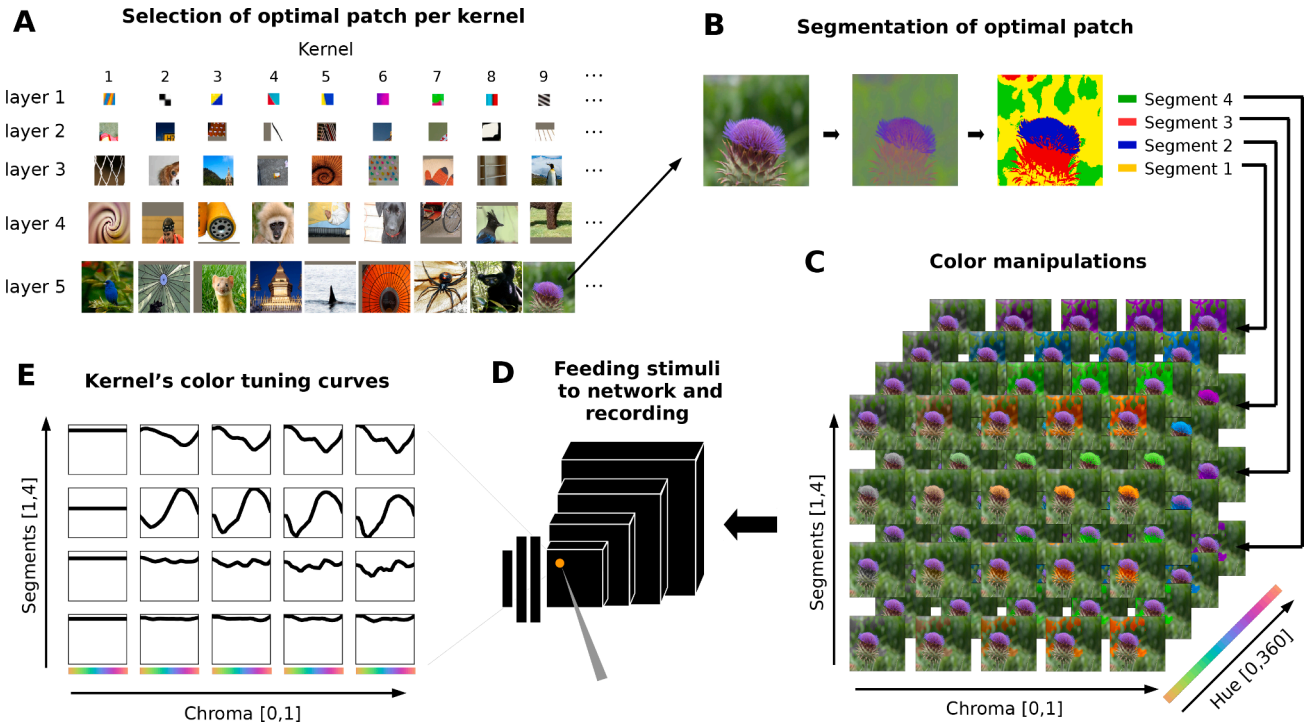

Fig. 1. Method used to extract the refined color tuning curves of kernels in the deep layers of our networks. **A**: We start by selecting, for each kernel in each layer, the *optimal patch* that results in its maximal activation across the entire training dataset; **B**: For each optimal patch, we subtract 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 l ike 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 4x5 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:

$$CS^{K}_{overall} = 1 - \frac{min(\mathbf{R}^{K})}{max(\mathbf{R}^{K})}. \tag{2}$$

where K denotes a kernel and **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,

$$CS^{K_S}_{hue} = max_{\mathbf{C}}\left(1 - \frac{min_{\mathbf{H}}\left(\mathbf{R}^{K}_{S,C}\right)}{max_{\mathbf{H}}\left(\mathbf{R}^{K}_{S,C}\right)}\right). \tag{3}$$

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 ($CS^{K_S}_{hue} > 0.5$), and a minor hue selectivity if its response varies by more than 25% ($CS^{K_S}_{hue} > 0.25$). 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:

$$CR^{K}_{S} = 1 - \frac{R^{K}_{S,C=0}}{max(\mathbf{R}^{K}_{S})}. \tag{4}$$

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

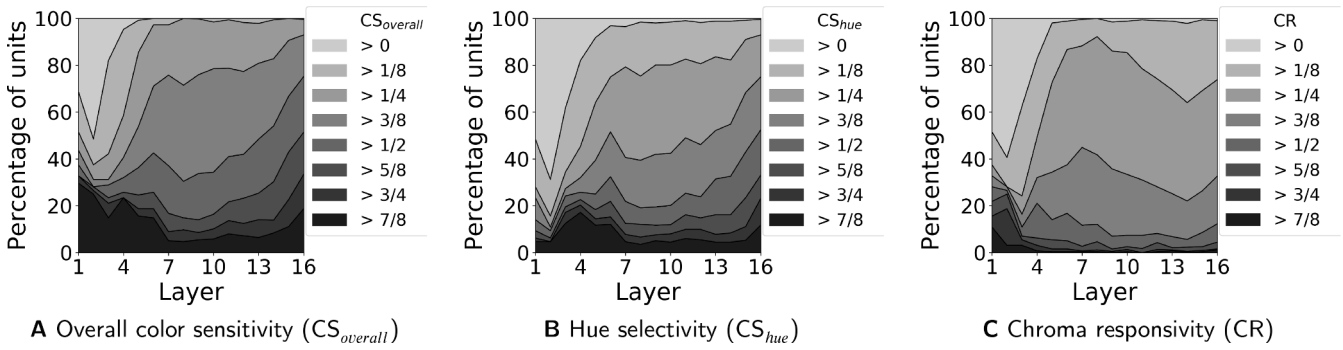

Fig. 2. Proportions of VGG-19's kernels with different levels of **A** overall color sensitivity; **B** hue selectivity; and **C** chroma responsivity.

<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, in Fig. 3, we selected 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

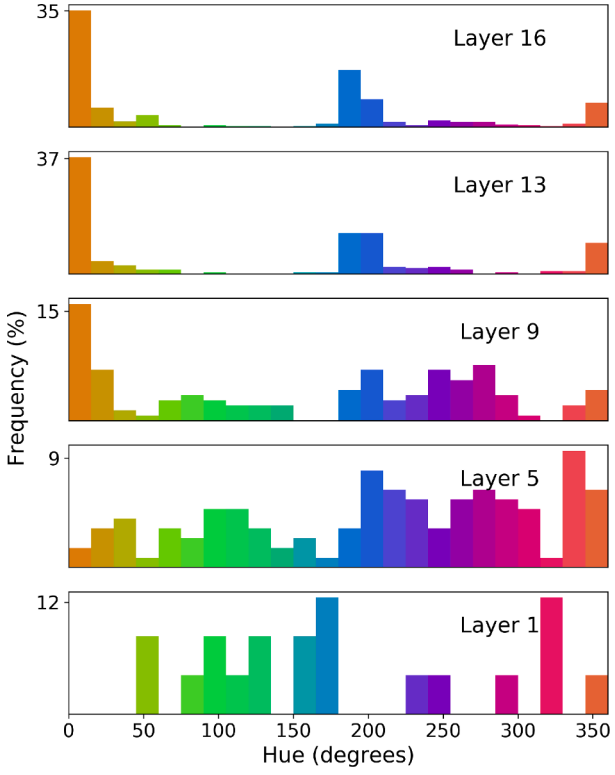

Fig. 3. Histograms of hues for which kernels are most responsive to across layers of the VGG-19 network.

($CS_{hue} > 0.5$) 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.

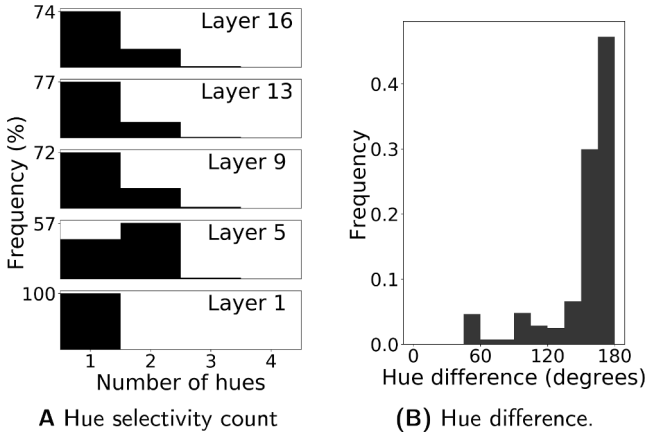

Fig. 4. **A**: Histograms of the number of hues for which kernel are selective (see Eq. 3). **B**: For kernels selective for 2 different hues at different segments: histogram of the hue difference between the 2 hues. Except for early layers of the VGG nets, a significant proportion of the hue selective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

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° ±15° away from the preferred hue.

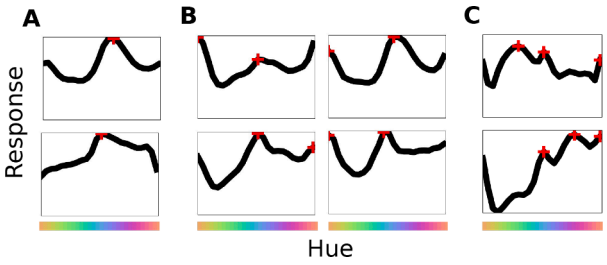

Fig. 5. Example of tuning curves displaying **A** 1 peak, **B** 2 peaks and **C** 3 peaks according to our algorithm (see Section 3).

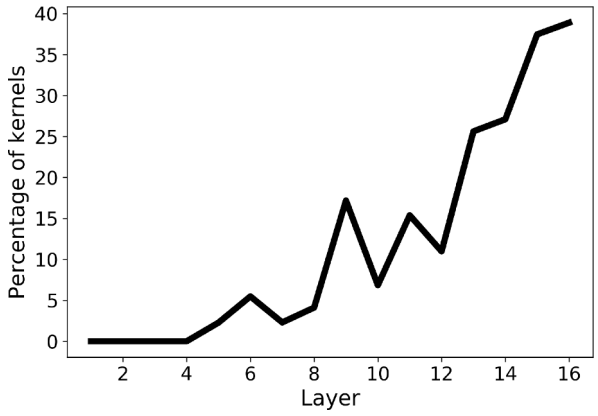

Fig. 6. Proportion of VGG-19's hue selective kernels showing a secondary peak in their tuning curves.

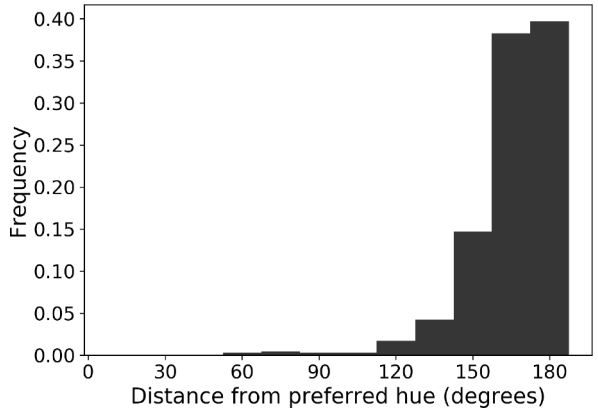

Fig. 7. Histograms of the distance, in hue, of the secondary with respect to the primary hue for the VGG-19 networks. We can see that most of the peaks fall 180° from the primary hue, meaning that kernels can be secondarily selective for hues 180° from the main hue in the same segment.

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.

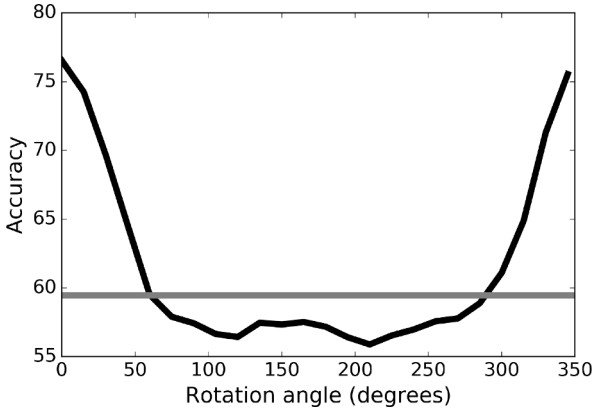

Fig. 8. VGG-19 classification performance as a function of hue angle rotation, relative to the original colors. In black: performance of VGG-19 as we modify images from the original colors by applying a rotation around the achromatic axis of color space. Gray horizontal line: performance of the model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

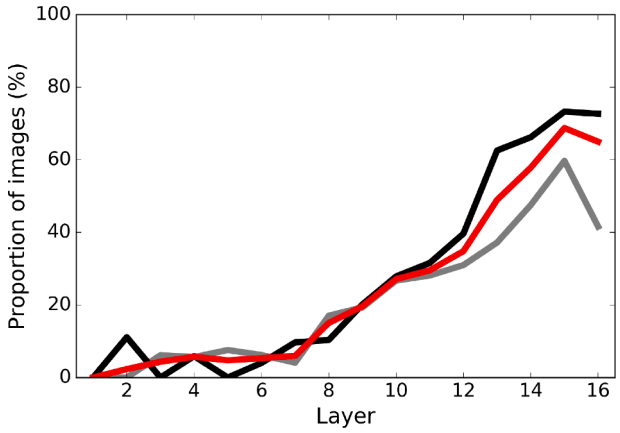

Fig. 9. The proportion of correctly classified images which are misclassified at least once when the color of a segment is modified in color (black) and non-color (grey) sensitive kernels, and in any kernel (red).

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

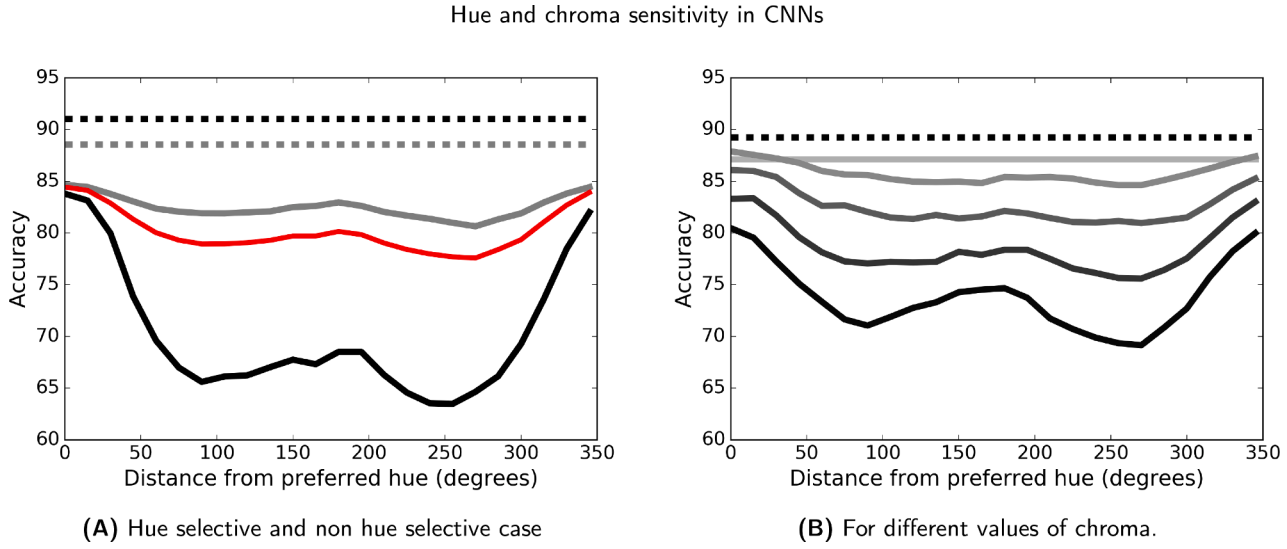

Fig. 10. VGG-19 network classification performance as a function of the distance, in hues, to the preferred hue. **A**: Results for different levels of hue selectivity, averaged across chroma. In full gray, results obtained for images including the optimal patches of non hue selective kernels. In full 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 responsivity 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, & Faísca, 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 greyscale 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, Drozdzal, 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 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.

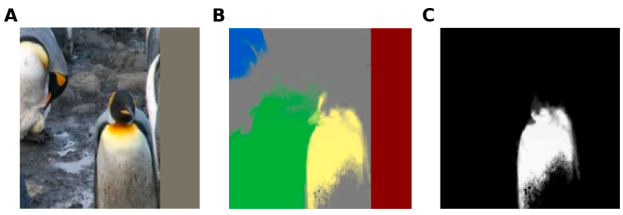

Fig. 11. Figure of an optimal image patch for which current segmentation algorithms failed to serve our purposes. In **A**, the optimal image patch found for kernel 98 of AlexNet's layer 5. It shows a penguin, an animal the segmentation algorithm has never seen before. As a consequence, the segmentation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

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, Filipp 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, athttps://doi.org/10.1016/j.visres.2020.09.010.

References

Aksoy, Y., Oh, T. H., Paris, S., Pollefeys, M., & Matusik, W. (2018). Semantic soft 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., & Faísca, 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 neual networks, in. 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. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Jégou, S., Drozdzal, M., Vazquez, D., Romero, A., & Bengio, Y. (2017). The one hundred layers tiramisu: Fully convolutional densenets for semantic segmentation, in. 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. 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, I., & 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 rgbd 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).