Vision Research 176 (2020) 156–174

Contents lists available at ScienceDirect

Vision Research

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

Color illusions also deceive CNNs for low-level vision tasks: Analysis and implications

A. Gomez-Villa[a,*], A. Martín[a], J. Vazquez-Corral[a], M. Bertalmío[a], J. Malo[b]

[a] Dept. Inf. Comm. Tech., Universitat Pompeu Fabra, Barcelona, Spain
[b] Image Proc., Lab, Universitat de València, València, Spain

ARTICLE INFO

Keywords:
Visual illusions
Artificial neural networks
Efficient representation
Natural image statistics

ABSTRACT

The study of visual illusions has proven to be a very useful approach in vision science. In this work we start by showing that, while convolutional neural networks (CNNs) trained for low-level visual tasks in natural images may be deceived by brightness and color illusions, some network illusions can be inconsistent with the perception of humans. Next, we analyze where these similarities and differences may come from. On one hand, the proposed linear eigenanalysis explains the overall similarities: in simple CNNs trained for tasks like denoising or deblurring, the linear version of the network has center-surround receptive fields, and global transfer functions are very similar to the human achromatic and chromatic contrast sensitivity functions in human-like opponent color spaces. These similarities are consistent with the long-standing hypothesis that considers low-level visual illusions as a by-product of the optimization to natural environments. Specifically, here human-like features emerge from error minimization. On the other hand, the observed differences must be due to the behavior of the human visual system not explained by the linear approximation. However, our study also shows that more 'flexible' network architectures, with more layers and a higher degree of nonlinearity, may actually have a *worse* capability of reproducing visual illusions. This implies, in line with other works in the vision science literature, a word of caution on using CNNs to study human vision: on top of the intrinsic limitations of the L + NL formulation of artificial networks to model vision, the nonlinear behavior of flexible architectures may easily be markedly different from that of the visual system.

1. Introduction

A visual illusion (VI) is an image stimulus that induces a visual percept that is not consistent with the visual information that can be physically measured in the scene. An example VI can be seen in Fig. 1: the center squares have the exact same gray value, and therefore send the same light intensity to our eyes (as a measurement with a photometer could attest), but we perceive the gray square over the white background as being darker than the gray square over the black background. There are many types of VIs, involving for instance the perception of brightness (White et al., 1979; McCourt et al., 1982; DeValois et al., 1990, color Kitaoka et al., 2005; Zaidi, Ennis, & Lee, 2012; Loomis et al., 1972; Hillis et al., 2005; Abrams, Hillis, & Brainard, 2007), texture (Blakemore et al., 1996; Ross et al., 1991; Foley et al., 1997; Watson et al., 1997, motion Morgan, Chubb, & Solomon, 2006; George Mather, Andrea Pavan, & Casco, 2008; Morgan & Chubb, 2011), geometry (Weintraub et al., 1971; Westheimer et al., 2008), etc.

In the context of the efficient representation view of biological vision (Attneave et al., 1954; Barlow et al., 1961), VIs are not seen as failures but as a by-product of strategies to adapt to the statistics of the images that the individuals typically encounter (Barlow, 1990; Clifford, Wenderoth, & Spehar, 2000; Clifford et al., 2002; Clifford et al., 2007; Laparra et al., 2015). This is the reason why VIs provide compelling case examples which are useful to probe theories about how our perception works. Following classical works in visual neuroscience and visual perception (Hubel et al., 1959; Campbell et al., 1968) that successfully predicted visual responses as a linear filtering operation followed by a pointwise nonlinearity, the "standard model" of vision (Olshausen et al., 2005) has become that of a filter bank or rather a cascade of linear and nonlinear (L + NL) modules (Carandini et al., 2012; Martinez-Garcia, Cyriac, Batard, Bertalmío, & Malo, 2018). The design of artificial neural networks (ANNs) has also taken neurobiological models as the source of inspiration (Haykin et al., 2009), and for this reason Convolutional Neural Networks (CNNs) can be seen as

* Corresponding author.
E-mail addresses: alexander.gomez@upf.edu (A. Gomez-Villa), adrian.martin@upf.edu (A. Martín), javier.vazquez@upf.edu (J. Vazquez-Corral), marcelo.bertalmio@upf.edu (M. Bertalmío), jesus.malo@uv.es (J. Malo).

https://doi.org/10.1016/j.visres.2020.07.010
Received 2 December 2019; Received in revised form 10 July 2020; Accepted 22 July 2020
Available online 04 September 2020

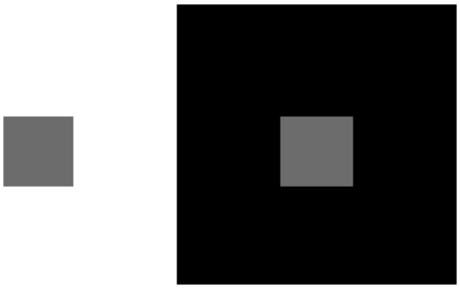

Fig. 1. An example visual illusion. The squares have the same gray value, but one is perceived as being brighter than the other.

constituted by a stack of L + NL modules as well.

Since 2018, a handful of works have found that CNNs trained on natural images can also be "fooled" by VIs, in the sense that their response to an image input that is a VI for a human is (qualitatively) the same as that of humans, and therefore inconsistent with the actual physical values of the light stimulus. In order to demonstrate the idea that visual perception is the result of image statistics learned thought goal-directed behaviour (Purves et al., 2003), Corney and Lotto (2007) trained a CNN to identify the reflectance of target surfaces in synthetic images, and saw that this CNN responded to some visual illusions in a way similar to humans. Watanabe, Kitaoka, Sakamoto, Yasugi, and Tanaka (2018) trained a CNN to predict videos and demonstrated that, as a side effect, it was able to reproduce motion illusions. Gomez-Villa, Martin, Vazquez-Corral, and Bertalmio (2019) showed that a CNN trained for low-level visual tasks in natural images is able to reproduce human perception in several instances of color and brightness illusions. Kim et al. (YYYY) showed how a CNN trained for classification exhibits the law of closure (a Gestalt principle), hence replicating visual completion illusions. Benjamin, Qiu, Zhang, Kording, and Stocker (2019) studied the orientation bias in CNNs trained for classification, and found that the early layers of these CNNs are capable of reproducing orientation visual illusions. Sun and Dekel (2019) demonstrated that a classification CNN is able to reproduce the Scintillating Grid visual illusion. Ward et al. (2019) showed that a deep CNN trained for object recognition is able to reproduce the Muller-Lyon illusion, a type of geometric visual illusion. Benjamin et al. (2019) fine-tuned the last layers of the AlexNet classification network to report perceived orientation and proved it in several orientation and geometric visual illusions. Linsley, Kim, Ashok, and Serre (2019) designed a deep recurrent neural network architecture based on the perception of the orientation-tilt illusion which was able to surpass the state-of-the-art in contour detection. Finally, Jacob, Pramod, Katti, and Arun (2019) run several experiments (including reproduction of visual illusions such as the Thatcher effect) in order to test perceptual capabilities of the CNNs.

This very recent line of research, devoted to the study of similarities and differences between the VIs suffered by human viewers and artificial neural networks, may be relevant to explore the limitations of simplified architectures and suggest better models of biological vision.

The reason for this is that CNNs still fail to emulate very basic perceptual phenomena (Martinez, Bertalmío, & Malo, 2019; Geirhos et al., 2019; Jacob et al., 2019; Geirhos et al., 2020) even when they achieve state-of-the-art results in modeling cortical activity (Cadena et al., 2019), and match human performance in vision tasks like face recognition and object classification.

This work expands our initial findings on visual illusions that deceive artificial neural networks, presented in Gomez-Villa et al. (2019). Our contributions in this paper are:

1. Providing more insight on why some CNNs trained for basic visual tasks in natural images are deceived by brightness and color illusions while others do not.
2. Performing a linear eigenanalysis in a simple CNN trained for denoising and deblurring that reproduces illusions, which shows that the network's response is qualitatively very similar to the human achromatic and chromatic contrast sensitivity functions (CSFs), and consistent with natural image statistics.
3. Performing a psychophysical-like analysis of CNNs to show that, while these artificial networks are deceived by illusions, their nature might be significantly different to that of humans.

These contributions suggest the following.

From result (1) above, low-level VIs may appear as a by-product of basic visual goals in natural environments, and simpler or more linear network architectures seem to suffer from stronger illusions.

From (2), and in line with error minimization explanations of visual function (Laparra et al., 2015; Atick, Li, & Redlich, 1993; Twer & MacLeod, 2001; MacLeod et al., 2003; Laparra, Jimenez, Camps, & Malo, 2012), simpler CNNs trained for low-level visual tasks also develop human-like achromatic and opponent chromatic channels with band-pass/low-pass spatial frequency response because the optimal removal of non-natural features (like noise or blur) leads to the identification of principal directions in natural scenes.

More interestingly, from (3), discrepancies with humans in quantitative experiments imply a word of caution on using CNNs to study human vision, in line with what's argued in some very recent works (e.g. see Jacob et al., 2019; Geirhos et al., 2020; Martinez et al., 2019 and references therein) and also as previously suggested in other contexts (with regards to L + NL formulations) in the vision science literature (Wandell et al., 1995; Carandini et al., 2005; Olshausen, 2013).

The structure of the paper is as follows. In the *Materials and Methods* section we introduce the stimuli used in the experiments, we describe the considered architectures and the visual tasks used to train them, and we discuss two alternative methods to describe the illusions that may be found in the networks. In the *Results* section we compute the shifts of the responses due to context and the corresponding pairs of the networks in asymmetric matching experiments. In the *Linear Analysis* section, eigenanalysis of the networks reveals intrinsic filters which are similar to the CSFs in opponent channels, and finally, the *Discussion* analyzes the implications of the results in terms of complexity of the networks and appropriateness to model human vision.

2. Materials and methods

2.1. The stimuli

In this work we deal with two sets of stimuli. First, experiments using the classical visual illusions shown in Fig. 2 will be used to point out that CNNs can have illusions that are *qualitatively* similar to those of human viewers. The illusions in Fig. 2 present test regions that are physically the same but are seen differently depending on their surrounds. Sometimes the context induces *assimilation* (the perception of the test shifts towards that of its surround), while others induces *contrast* (the perception of the test moves away from that of its surround). The test regions are, in the Dungeon illusion (Bressan et al., 2001), Fig. 2a, the large central squares, in Hong et al. (2004), Fig. 2b, the middle rings, in the White illusion (White et al., 1979), Fig. 2c, the small grey bars, and in the Luminance gradient (combination of Brucke, 1865; Adelson, 2000), Fig. 2d, the circles. The fact that the tests are identical can be seen in the second and fourth to sixth rows of Fig. 2, that plots the digital values along some line segments marked over the visual illusions in the first and third rows. The Chevreul illusion (Ratliff et al., 1965), Fig. 2e, presents homogeneous bands of increasing intensity, from left to right, but these bands are perceived to be inhomogeneous, with darker and brighter lines at the borders between adjacent bands.

Then, a second experiment simulating asymmetric color matching is performed with the networks to have results that are quantitatively comparable to those of human viewers. In an asymmetric matching

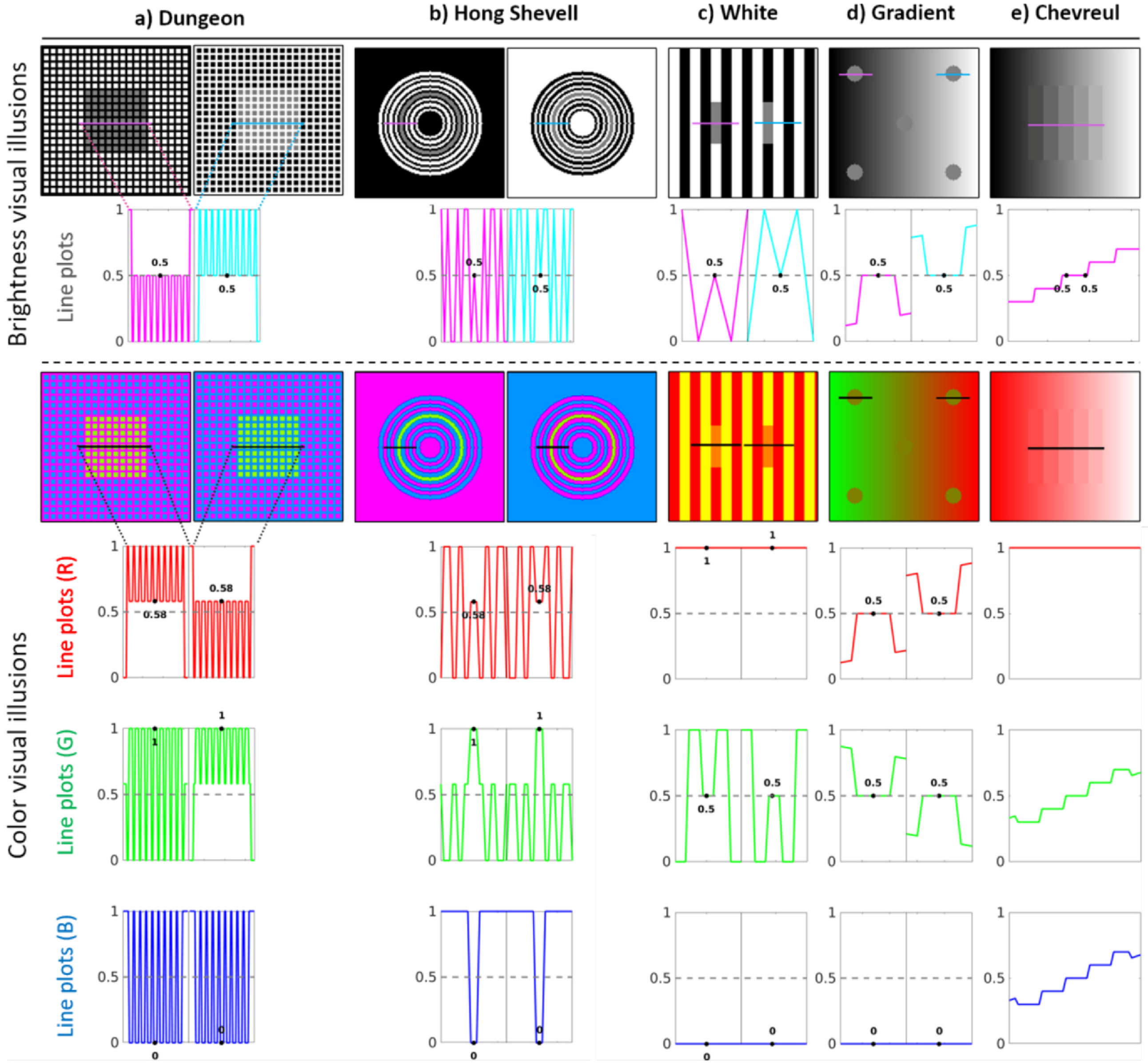

Fig. 2. Classical stimuli to check the existence of brightness and chromatic visual illusions. The first and the third rows show the selected grayscale and color stimuli respectively. The second row plots the intensity values of the images at the lines marked over in magenta and cyan. Rows 4–6 show the profiles representing the RGB digital values of the color stimuli at the lines depicted in the images. Equal digital values imply physically equal tests despite what we perceive.

(Heinemann et al., 1955; Ware et al., 1982), given a scene consisting of a *test* **t** seen on an inducing *surround* **s**, the observer modifies a variable stimulus seen on a neutral background to look for the *corresponding pair* **t'** that matches the perception of the test **t**. The surround induces an illusion because in general $\mathbf{t}' \neq \mathbf{t}$. The magnitude of the illusion is quantified by the displacement in chromatic coordinates required to match the color perception. Fig. 3 illustrates the layout and the colors used in our simulation of the Ware-Cowan experiments (Ware et al., 1982).

2.2. The networks: architectures and training

We trained two CNN architectures for three low-level visual tasks: *denoising* and *deblurring* (as in the work of Gomez-Villa et al. (2019)) and also *restoration* (a combination of the denoising and deblurring problems). Hence, we have 6 models. The general setting to obtain the parameters of the models is supervised learning (see Fig. 4).

For consistency with the spatial extent of the stimuli used in the experiment with humans reported by Ware et al. (1982), we assume the images subtend 1.83 deg with sampling frequency of 70 cpd (128 × 128 pixels).

The first architecture has input and output layers of size 128 × 128 × 3 pixels. The architecture has *two* hidden layers with *eight* feature maps with a kernel size of 5 × 5 and no stride, and sigmoid activation functions. The second architecture is a bit deeper and hence with substantially more free parameters. It also has input and output layers of size 128 × 128 × 3, but *four* hidden layers with 24 feature maps. Kernel sizes and non-linearities are the same as in the first architecture. The two hidden-layer architectures (shallow) are named with respect to the task they are trained for: DN-NET (denoising network), DB-NET (deblurring network), and RestoreNET (restoration network). As for the four hidden-layer architectures (deep), we added the “Deep” word to the corresponding shallow architecture name, hence: Deep DN-NET, Deep DB-NET, and Deep RestoreNET.

Mean squared error was used as loss function in all the tasks and all the models were implemented using Abadi et al. (2015). The maximum

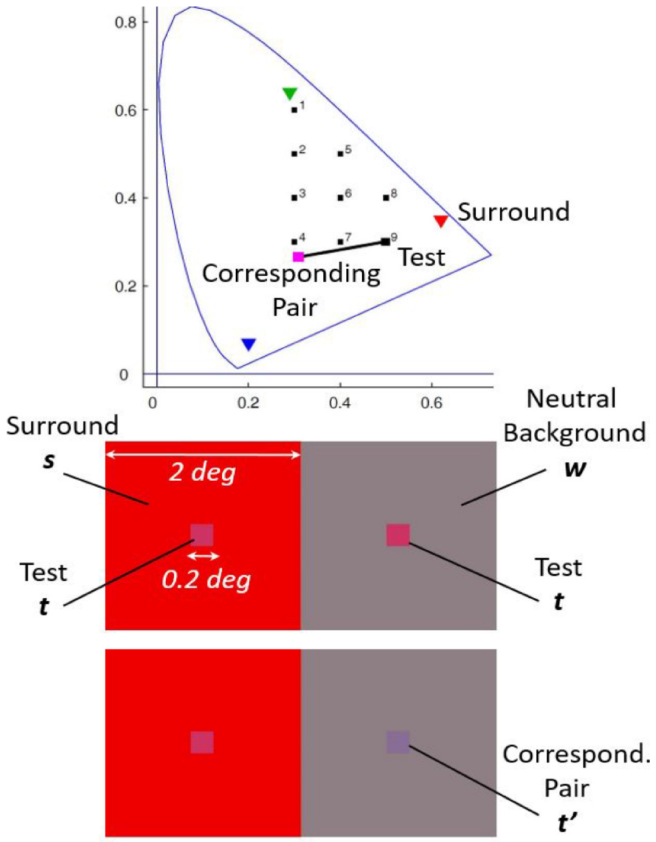

Fig. 3. (Top CIE xy diagram) test points (1–9) and inducing surrounds (in R,G, and B) used in our simulation of the Ware-Cowan corresponding pair experiment (Ware et al., 1982). Luminance of all colors was set to 30 cd/m^2. The highlighted colors illustrate one match: the effect of the *red* surround on test number 9 implies a shift of the corresponding pair, in magenta. (Middle Panel) The test seen on top of the neutral background at the right is modified by the observer till it matches the color appearance of the test-surround scene (Bottom panel).

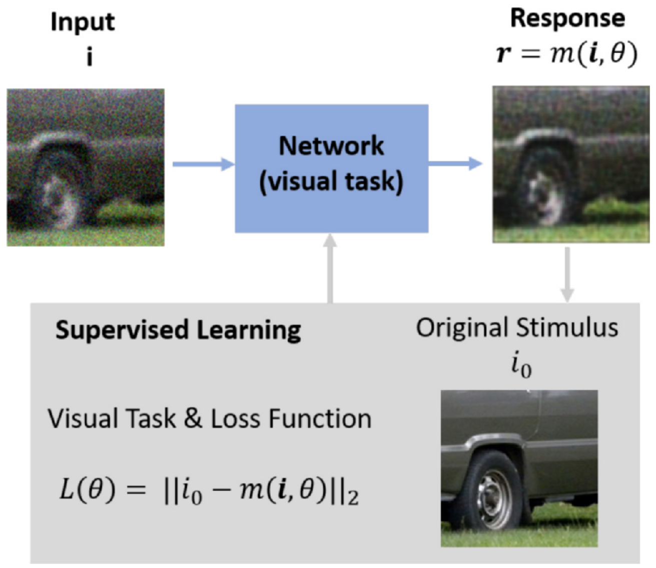

Fig. 4. In *supervised learning* the parameters, θ, of the models, $\mathbf{m}$, are obtained by minimizing a distance between the response of the model and a desired result known by the *supervisor*. The networks we train here take degraded photographic images as input stimuli, $\mathbf{i}$, and the response is compared to the known original, i_0.

number of epochs was set to 100 and we stopped the optimization early if there was no improvement in the validation set after 5 consecutive evaluations.

The dataset used for training the above architectures is the Large Scale Visual Recognition Challenge 2014 CLS-LOC validation dataset (which contains 50 k images), leaving 10 k images for validation purposes. This dataset is a subset of the whole ImageNet dataset (Russakovsky et al., 2015).

For denoising, we corrupted the images with additive Gaussian noise of $\sigma = 25$ in each RGB channel (digital counts in the range [0, 255]). In the case of deblurring, we blurred the images with a spatial Gaussian kernel of width $\sigma_x = 0.03$ deg (2 pixels). As for restoration, first we blurred the images with a Gaussian kernel of $\sigma_x = 0.03$ deg and then we corrupted the images with additive Gaussian noise of $\sigma = 25$. Note that restoration combines the other two tasks, thus being more general.

The above case-study architectures with 2 and 4 hidden layers[1] give us the opportunity to check the behavior of the networks in a systematic and comparable way. However, it is also interesting to explore the eventual illusions happening in current much deeper networks used in real image processing applications. To this end, we also explored the behavior of the 17-layer CNN with batch normalization for denoising by Zhang, Zuo, Chen, Meng, and Zhang (2017), and the 21-layer CNN with recurrent units for deblurring by Tao, Gao, Shen, Wang, and Jia (2018). These networks represent the state-of-the art in these tasks. In these *really deep* cases we used the implementations pretrained and provided by Zhang[2] and Tao[3] respectively.

Let us finally note that in this paper we selected low-level nets (i.e. nets reconstructing signals in image domain) because the "physiological" interpretation of their response is straightforward. This said, it can not be discarded that networks trained for higher level tasks – such as recognition and segmentation – may present different illusions, but they may need to be studied using other techniques to relate their outputs to observers' answers.

2.3. *Strategies to measure the illusions in CNNs*

In this work we use two alternative strategies to measure the visual illusions of the network. The first strategy is inspired from physiology while the second simulates psychophysics.

The physiological-like strategy consists of measuring the shifts in the response of the CNN for identical test values surrounded by different contexts (the classical VIs shown in Fig. 2). This approach implicitly assumes that the CNN output is akin to the human perception of the input. However, while it is useful to spot and compare trends between the CNN output and the human opinion (e.g. if the CNN is increasing the gray value of a test region this would be consistent with humans perceiving this regions as a lighter gray), it is not adequate to perform quantitative comparisons with human responses. For this, we employ the psychophysical-like strategy, using the CNN as an observer to simulate perceptual matches, as it's done in psychophysics. In this manner, the units to quantify the strength of the illusion of the network are exactly the same as in human psychophysics, so the performance of CNNs and humans can be compared quantitatively.

It is worth noting that comparisons with humans through simulated psychophysics can be applied to analyze CNNs trained to more general goals (not involving image reconstruction). Matching the responses at specific inner layers is always possible regardless of the nature of the output. This would inform about which layers have a human-like image representation. Proposition of comparison criteria at different layers is a matter of active research (Jacob et al., 2019).

[1] Source code will be made publicly available.

[2] See *https://github.com/cszn/DnCNN*.

[3] See *https://github.com/jiangsutx/SRN-Deblur*.

3. Results

We did two numerical experiments with the *shallow* and the *deep* CNNs trained according to the low-level visual tasks considered above: the first experiment, presented in Section 3.1, uses a physiological-like strategy and measures shifts in the response of the CNN, while the second experiment, presented in Section 3.2, uses a psychophysical-like strategy and matches corresponding pairs in color induction tests.

In these experiments where the networks are applied to the illusion-inducing stimuli, it is convenient to think of stimuli, the images, $\mathbf{i}$, and responses, $\mathbf{r}$, as column vectors where the subindexes t and s denote respectively the *test* and *surround* part of the stimuli that are spatially disjoint, i.e. in different rows of the corresponding column vector. This is just a formalism that we use for the equations. In practice since we work with images, what we have are masks indicating which pixels correspond to the *test* and *surround* parts of the stimuli. Schematically,

$$\mathbf{i} = \begin{bmatrix} \mathbf{i}_t \\ \mathbf{i}_s \end{bmatrix} \quad \text{and} \quad \mathbf{r} = \begin{bmatrix} \mathbf{r}_t \\ \mathbf{r}_s \end{bmatrix}, \tag{1}$$

where, considering $n \times n \times 3$ color images, $\mathbf{i}$ and $\mathbf{r}$ are $(n^2 \cdot 3) \times 1$ vectors.

3.1. Shifts in the response

This experiment consists of computing the response of the network for the physically identical tests seen in different spatio-chromatic contexts (the stimuli in Fig. 2). It reduces to applying the kth model to each stimulus, $\mathbf{i}$, to compute the corresponding response, $\mathbf{r}$:

$$\mathbf{r} = m(\mathbf{i}, \theta_k), \tag{2}$$

where m represents the model response function that depends on the parameters θ_k learnt with certain architecture and task. With 2 architectures and 3 tasks, $k = 1, \ldots, 6$. As stated above, we also consider the response of two pretrained really deep nets which are state-of-the-art in denoising (Zhang et al., 2017) and deblurring (Tao et al., 2018).

The qualitative interpretation of the behavior of the responses in different settings can be done by checking if the response shifts in certain direction. For instance, in the achromatic cases, *is the response departing from average brightness?* and if so, *in which direction, darker or brighter?* In the color cases, The kind of questions could be *are the values of the response departing from the input hue?*, and if so, *in which direction, towards orange or towards green?*

3.1.1. Achromatic case

Figs. 5 and 6 show the response profiles obtained from the different CNNs considered in this work when they are fed with the achromatic illusion-inducing images from the first row of Fig. 2. Below each panel, the words “Lighter” and “Darker” describe the direction of the shift of the CNN response for the test region, and the font color indicates if it is human-like (in black), weak (in gray), or non–human (in red).

These figures show that the simpler networks have shifts in perceived brightness in similar directions as humans in about 80% of the cases, while the deeper networks trained for similar goals have very small brightness illusions, and in most cases (about 60%), not in the human direction.

3.1.2. Chromatic case

Shifts of the chromatic responses are more difficult to assess since three curves have to be evaluated at the same time. See for instance a small fraction of the results in Fig. 7. For this reason, we are going to discuss each visual illusion at a time in the following way. First, we will refer to Fig. 2 (third row) to describe verbally the color shifts of the tests in each illusion. Second, we will focus on the numerical values of the responses at the central pixels of the tests. Reporting the original input values and the response values is a more concise way to give this information. This is systematically presented in Tables 1 and 2, where we also add symbol marks to identify how the tests shift in hue or brightness: a green “check” mark when the shift is *human-like*, a red “cross” mark when it is *non-human*, or a red “~” mark for a *very weak effect*.

Dungeon: For humans the left test shifts towards orange/yellow and the right shifts towards green. All the simpler networks have substantially bigger R response in the left test and substantially bigger G response in the right test (see Fig. 7, and the Dungeon column in Table 1). That is why all these cases are labeled as *human-like*. On the contrary, note that the variations of the responses for the networks of Zhang and Tao are substantially smaller (particularly for Zhang). That is why we labeled them as *very weak*.

Hong-Shevell: The left test shifts towards green while the right test shifts towards orange/yellow in the case of humans. Therefore, a network qualifies as human-like if it gives substantially bigger G response at the left and substantially bigger R response at the right. This is the case in all simple networks except for the deblurring nets, hence labeled as *weak effect*. Zhang network introduces totally negligible changes in the values, and Tao network leads to slightly more reddish response at the right, but the relative differences are much smaller than in simpler networks. That is why both have been labeled as *very weak*.

White: The rectangle at the right is darker and more reddish for human observers. When network responses at the right have relatively bigger R than G compared to the left, the behavior has been labeled as *human-like*. This happens in all the simpler networks except for Deep DB-Net (in which the green response at the right does not reduce much, meaning it may be perceived lighter). Again for both state-of-the-art networks the illusion is negligible.

Gradient: For humans the dots at the right are seen greener (contrast effect). Therefore, networks with bigger G component at the right are said to reproduce the illusion. This is not the case in any of the simpler networks. On the contrary, it is the case in both state-of-the-art nets (though weak in Zhang).

Chevreul: Humans perceive the left side of the regions as lighter. Therefore, in this reddish case, we qualify as non-human behavior when the R response at the left is not higher than at the right. All the simple networks have higher R response at the left, but in the denoising networks the other responses decay a bit, so we have labeled this behavior as *weak*. Both state-of-the-art networks have non–human behavior (no illusion or shift in the opposite direction).

To summarize the results on color shifts in Tables 1 and 2, we have seen that simpler architectures lead to substantial human-like illusions in 66% of the considered cases. This rate is bigger, up to 80%, for the (more general) restoration task. On the contrary, the state-of-the-art networks only develop strong human-like shifts in 20% of the cases.

These trends in the color results are consistent with the previous achromatic results. In this (qualitative) experiment based on the shift of the responses, the behavior of simpler (more rigid) networks is more similar to the human behavior than the considered much deeper architectures.

It is important to note that the trends mentioned above are independent of two common implementation issues: (1) the initialization of the weights, and (2) the choice of the nonlinearity in each convolutional unit. These implementation points apply to the simpler networks that have been defined and trained for this work. Note that the more flexible networks have not been trained, but used with default parameters. Separate exhaustive analysis in the specific RestoreNET (results not shown) indicates that the above choices are not an issue in the observed performance of the models. On the one hand, variance over 10 random initializations of the shallow and deep architectures are very small and do not change the qualitative trends described. On the other hand, regarding the nonlinearity, we selected sigmoid functions because, in principle, it seems more biologically plausible. However experiments on RestoreNET using the ReLU nonlinearity show that the response profiles also follow the same trends, and hence this choice is

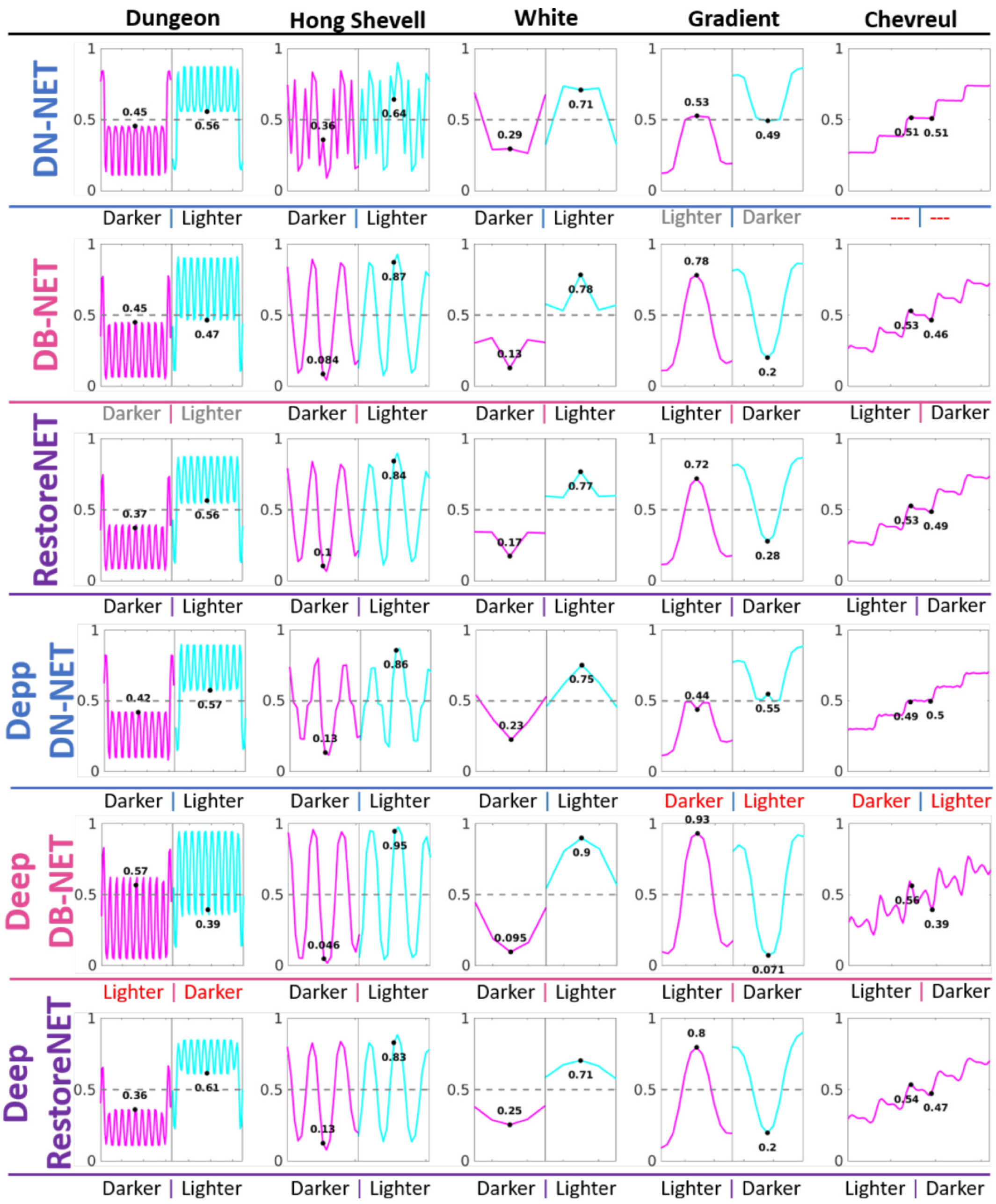

Fig. 5. Response of the simpler networks (2-layer shallow and 4-layer deep) CNNs to the stimuli inducing brightness illusions. The profiles refer to the stimuli depicted in Fig. 2 (in which tests appear distinctly lighter or darker to humans). Below the response profile in each case we describe the direction of the shift (darker or lighter). Descriptions in *black* indicate the shift is in the same direction as humans (as seen in Fig. 2). Descriptions in *gray* also mean correspondence with humans but weak effect. And those in *red* mean non–human shifts.

less relevant than the task and the complexity of the architecture.

3.2. *Corresponding pairs in color induction*

In this experiment we replicate the Ware-Cowan approach of Ware et al. (1982). Given a fixed test-surround configuration, $\mathbf{i} = [\mathbf{i}_t\ \mathbf{i}_s]^\top$, the observer looks for the *corresponding pair*, $\mathbf{t}'$, that seen on a neutral reference background, $\mathbf{w}$, matches the perception of $\mathbf{t}$ (see Fig. 4).

While human observers look for the corresponding pair by physically changing the color in the lab, we say that the network matches the perception when, given these two responses,

$$\begin{bmatrix}\mathbf{r}_t\\ \mathbf{r}_s\end{bmatrix} = m\left(\begin{bmatrix}\mathbf{i}_t\\ \mathbf{i}_s\end{bmatrix}, \theta_k\right) \quad\text{and}\quad \begin{bmatrix}\mathbf{r}_{t'}\\ \mathbf{r}_w\end{bmatrix} = m\left(\begin{bmatrix}\mathbf{i}_{t'}\\ \mathbf{i}_w\end{bmatrix}, \theta_k\right) \tag{3}$$

it holds $\mathbf{r}_t = \mathbf{r}_{t'}$. Therefore, the cost function for the *numerical* corresponding pair experiment is just:

$$\mathbf{i}_t{}' = \mathrm{argmin}_{i_t^\star} \mid \mathbf{r}_{t^\star}([\mathbf{i}_t^\star\ \mathbf{i}_w]) - \mathbf{r}_t([\mathbf{i}_t\ \mathbf{i}_s]) \mid_2. \tag{4}$$

The optimization is solved by means of an exhaustive search in a discretized RGB space between 0 and 255 with a step size of 8. Results for this experiment are shown in Figs. 8–10. In these figures, the black squares represent the test colors, while the magenta squares represent the corresponding pair required by the model to match the responses. The red, green and blue squares are the inducers. The first row in Fig. 8 presents the results obtained by the human observers (Ware et al., 1982). The rest of rows in Fig. 8 show the results for the shallow CNNs, while Fig. 9 presents the results for our deep CNNs. Finally, Fig. 10 shows the results for the recent methods of Zhang (denoising) and Tao

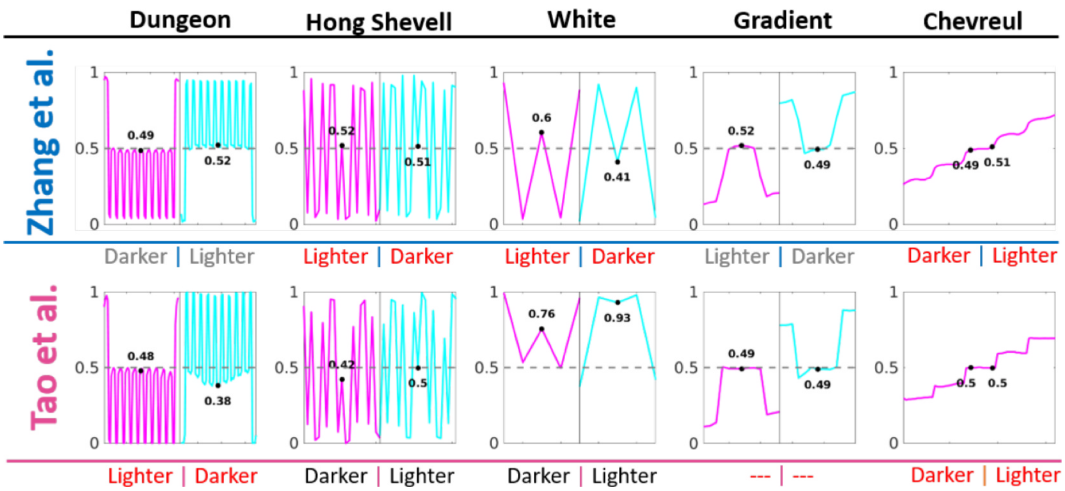

Fig. 6. Response of state-of-the-art CNNs to the stimuli inducing brightness illusions. Below the response profile in each case we describe the direction of the shift (darker or lighter). Descriptions in *black* indicate the shift is in the same direction as humans. Descriptions in *gray* also mean correspondence with humans but weak effect. And those in *red* mean non-human shifts.

(deblurring).

In Figs. 8 and 9 we see that the simpler CNNs require different **t'** to match the response of **t**. Exhaustive experiments using multiple random initalizations with shallow and deep RestoreNET (results not shown) indicate that the cluster of corresponding pairs found, **t'**, *do not* overlap with the test **t** in 93% of the cases. This means that the illusions are significant. That said, in shallow networks the magnitude of the illusions is substantially smaller than in humans. Interestingly, the 4-layer versions of these networks do have illusions as strong as humans. However, in both cases these networks develop assimilation instead of the contrast effect suffered by humans.

Fig. 10 shows that Zhang network displays virtually no illusion while Tao does (with smaller strength than our 4-layer network). Interestingly, displacements in Tao go away from the inductor, i.e. it is developing a contrast effect. This last result relates to the behavior obtained for the Gradient illusion in Section 3.1.2, where the Tao (and also weakly Zhang) were the only ones presenting the contrast effect. This connection was to be expected, as the Gradient illusion is the closest one to the Ware-Cowan experiment of all the illusions studied there.

3.3. Summary of the experimental results

On the one hand, the (qualitative) experiment based on the shift of the responses, shows that simpler (more rigid) networks are more similar to humans than the considered really deep architectures, which have negligible or non–human illusions.

On the other hand, the (quantitative) experiment simulating asymmetric matching psychophysics shows the following. Simpler networks do have significant illusions, and in particular the magnitude of the illusions in the 4-layer networks is similar in strength to human illusions. However, the direction of the illusions is the opposite: the simpler CNNs perceive assimilation while humans perceive contrast. As in the qualitative response experiment, here the Zhang network also seems to have negligible illusions. But now the Tao network does have substantial illusions, although their magnitude is smaller than those of human observers. Interestingly, the Tao network is able to reproduce the human-like contrast behavior.

4. Linear analysis of the networks

Results of the numerical experiments with the networks trained for low-level visual tasks confirm that they do have illusions, as anticipated before (Gomez-Villa et al., 2019), but they do not necessarily have the same illusions as humans do. This section analyzes why, particularly for the assimilation effect seen in Ware-Cowan experiments.

Here we show how the behavior of these CNNs can be understood through the analysis of the linear approximation of their response. Specifically, consider the following first order approximation of Eq. 2:

$$\mathbf{r} = m(\mathbf{0}, \theta_k) + \nabla_{\mathbf{i}}\, m(\mathbf{0}, \theta_k)\cdot\mathbf{i} \approx M_{\theta_k}\cdot\mathbf{i} \qquad (5)$$

where the matrix M_{θ_k} is the Jacobian of the network response w.r.t. the input at **0**, and we assumed that the response for the null stimulus is also zero. Note that in end-to-end networks with error-minimization tasks (denoising/deblurring/restoration) this assumption is reasonable: $m(\mathbf{0}, \theta_k) \approx \mathbf{0}$ should hold to keep deviation with the input small. We explicitly checked this in all the considered networks and found that the mean and variance of the response to a null image is always 2 orders of magnitude smaller than the maximum possible response. The matrix M_{θ_k} is fixed for a certain architecture/task combination.

The Jacobian for different points gives important information about the behavior of the network and could be computed analytically for any input (Martinez-Garcia et al., 2018). However, for our purposes here (we need it only at the origin, **0**), it can be estimated through plain linear regression. Once some architecture has been trained for certain task, the resulting model, characterized by the parameters θ_k, can be applied to a set of N stimuli. Then, by stacking the vectors representing the N stimuli and the N responses in the $(n^2{\cdot}3)\times N$ matrices, $\mathbf{I} = [\mathbf{i}^{(1)}\mathbf{i}^{(2)}\cdots\mathbf{i}^{(N)}]$, and $\mathbf{R}_{\theta_k} = [\mathbf{r}^{(1)}\mathbf{r}^{(2)}\cdots\mathbf{r}^{(N)}]$, respectively, we have:

$$M_{\theta_k} = \mathbf{R}_{\theta_k}\cdot\mathbf{I}^{\dagger} \qquad (6)$$

where $\mathbf{I}^{\dagger}$ is the pseudoinverse of the rectangular matrix with the input images and M_{θ_k} is then a $(n^2{\cdot}3)\times(n^2{\cdot}3)$ square matrix.

While CNNs are, in general, difficult to understand (Samek et al., 2019), if the proposed linear approximation captures a substantial fraction of the energy of the response, it can be very useful for two reasons: (1) it allows the use of well understood linear algebra tools in the analysis, and (2) it allows the comparison with classical linear descriptions of human vision.

In particular, here we perform two kinds of linear analysis, where the second is justified by the results of the first:

1. First, we make no extra assumptions (apart from linearity) and we perform an eigenvector analysis of the matrix M_{θ_k}. This analysis shows that this kind of networks are stationary (shift invariant), they are roughly spatio-chromatically separable, they implicitly operate in a color opponent space, and they have markedly different spatial bandwidth in these chromatic channels.

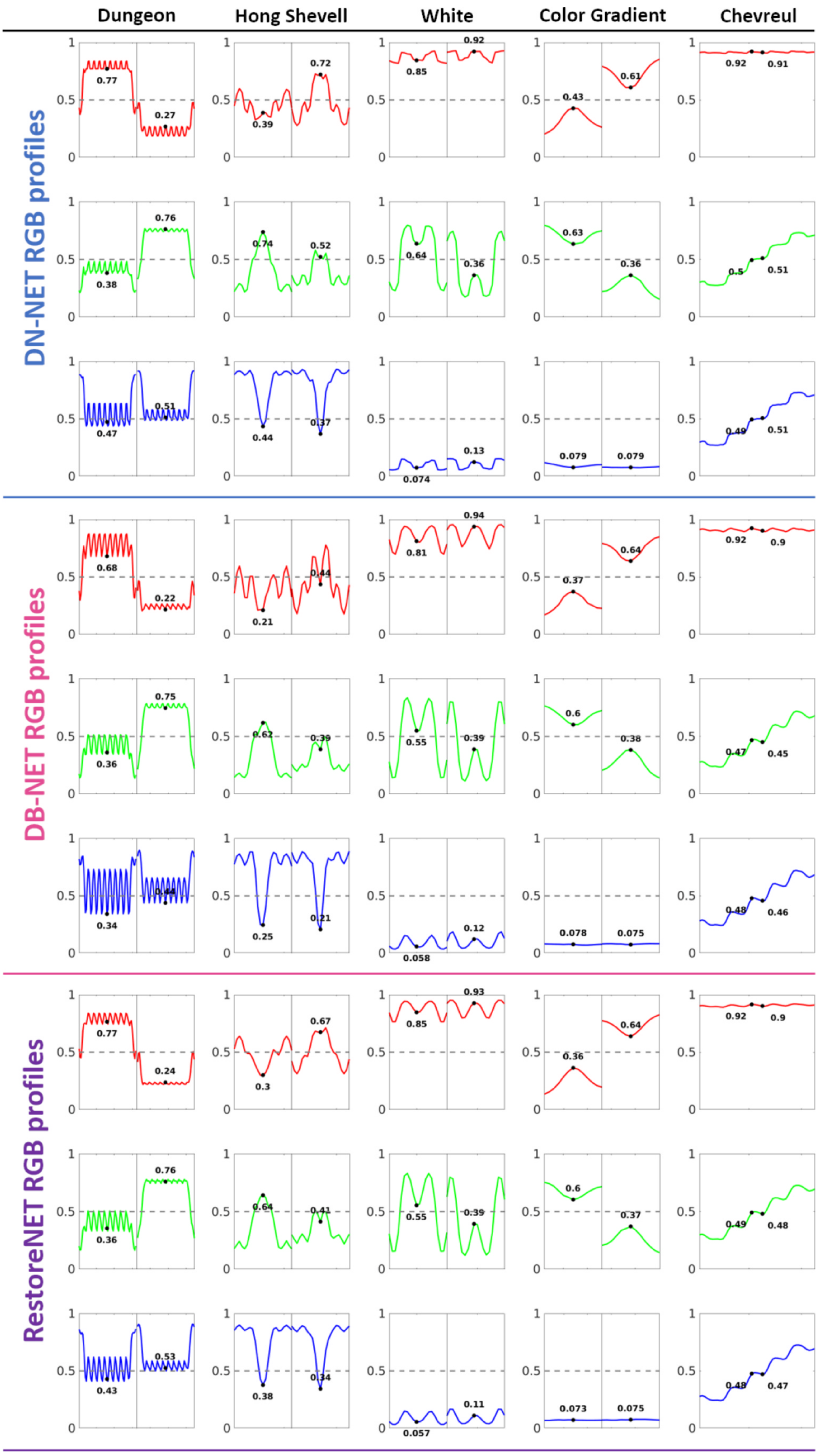

Fig. 7. Response profiles of the shallow CNNs to the color stimuli that illustrate different color illusions. The responses of the sensors tuned to RGB hues at the highlighted locations represent the *perception* of the network for the considered tests.

Table 1
Input and model responses for the simpler CNNs (2-layers shallow CNNs and deep 4-layer CNNs) for the considered color illusions.

DN-NET

DN-NET

	Dungeon ✓			Hong-Shevell ✓			White ✓			Gradient X			Chevreul ~		
	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R
R	0.58	0.77	0.27	0.58	0.39	0.72	1	0.85	0.92	0.5	0.43	0.61	1	0.92	0.91
G	1	0.38	0.76	1	0.74	0.53	0.5	0.64	0.36	0.5	0.63	0.36	0.5	0.5	0.51
B	0	0.47	0.51	0	0.44	0.37	0	0.074	0.13	0	0.079	0.079	0.5	0.49	0.51

DB-NET

	Dungeon ✓			Hong-Shevell ~			White ✓			Gradient X			Chevreul ✓		
	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R
R	0.58	0.68	0.22	0.58	0.21	0.44	1	0.81	0.94	0.5	0.37	0.64	1	0.92	0.9
G	1	0.36	0.75	1	0.62	0.39	0.5	0.55	0.39	0.5	0.6	0.38	0.5	0.47	0.45
B	0	0.34	0.44	0	0.25	0.21	0	0.058	0.12	0	0.078	0.075	0.5	0.48	0.46

RestoreNET

	Dungeon ✓			Hong-Shevell ✓			White ✓			Gradient X			Chevreul ✓		
	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R
R	0.58	0.77	0.24	0.58	0.3	0.67	1	0.85	0.93	0.5	0.36	0.64	1	0.92	0.9
G	1	0.36	0.76	1	0.64	0.41	0.5	0.55	0.39	0.5	0.6	0.37	0.5	0.49	0.48
B	0	0.43	0.53	0	0.38	0.34	0	0.057	0.11	0	0.073	0.075	0.5	0.48	0.47

Deep DN-NET

	Dungeon ✓			Hong-Shevell ✓			White ✓			Gradient X			Chevreul ~		
	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R
R	0.58	0.77	0.27	0.58	0.42	0.63	1	0.94	0.97	0.5	0.47	0.49	1	0.96	0.95
G	1	0.38	0.72	1	0.57	0.4	0.5	0.61	0.38	0.5	0.54	0.48	0.5	0.49	0.49
B	0	0.5	0.57	0	0.56	0.49	0	0.028	0.056	0	0.027	0.03	0.5	0.52	0.52

Deep DB-NET

	Dungeon ✓			Hong-Shevell ~			White X			Gradient X			Chevreul ✓		
	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R
R	0.58	0.67	0.29	0.58	0.36	0.45	1	0.89	0.97	0.5	0.5	0.45	1	0.95	0.92
G	1	0.27	0.72	1	0.51	0.25	0.5	0.48	0.47	0.5	0.56	0.41	0.5	0.52	0.45
B	0	0.41	0.57	0	0.52	0.36	0	0.024	0.094	0	0.068	0.052	0.5	0.53	0.46

Deep RestoreNET

	Dungeon ✓			Hong-Shevell ✓			White ✓			Gradient X			Chevreul ✓		
	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R
R	0.58	0.74	0.28	0.58	0.31	0.58	1	0.93	0.97	0.5	0.48	0.55	1	0.93	0.91
G	1	0.39	0.7	1	0.5	0.42	0.5	0.51	0.39	0.5	0.48	0.46	0.5	0.53	0.5
B	0	0.49	0.56	0	0.49	0.53	0	0.036	0.075	0	0.074	0.073	0.5	0.5	0.48

Table 2
Input and model responses for the (more flexible) state-of-the-art CNNs for the considered color illusions.

Zhang et al.

Zhang et al.

	Dungeon ~			Hong-Shevell ~			White ~			Gradient ✓			Chevreul X		
	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R
R	0.58	0.60	0.56	0.58	0.58	0.59	1	1	1	0.5	0.52	0.49	1	1	1
G	1	0.96	0.98	1	0.99	0.98	0.5	0.5	0.49	0.5	0.49	0.51	0.5	0.48	0.51
B	0	0.008	0.008	0	0.02	0.02	0	0.012	0.012	0	0.012	0.012	0.5	0.48	0.51

Tao et al.

	Dungeon ~			Hong-Shevell ~			White ~			Gradient ✓			Chevreul X		
	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R	In	Out-L	Out-R
R	0.58	0.72	0.55	0.58	0.56	0.69	1	0.98	0.95	0.5	0.47	0.55	1	0.99	0.99
G	1	0.83	0.89	1	0.98	1	0.5	0.47	0.47	0.5	0.43	0.58	0.5	0.49	0.5
B	0	0.17	0.2	0	0.12	0.13	0	0	0	0	0.02	0	0.5	0.5	0.5

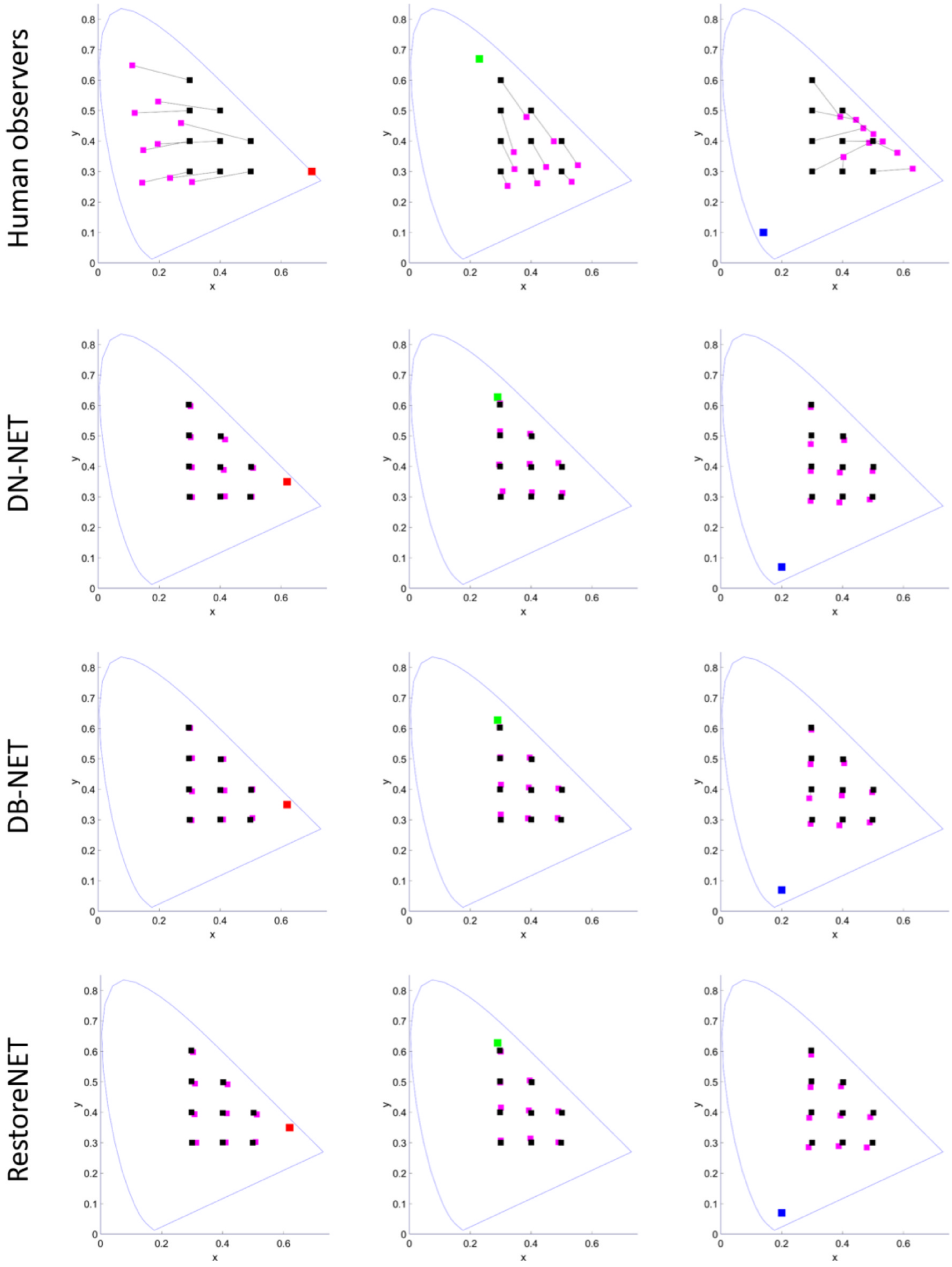

Fig. 8. First row: Results for the human observers in the Ware-Cowan corresponding pair experiment. Second to last raw: Results for the shallow CNNs studied in the Ware-Cowan corresponding pair experiment. We can see that the displacements are small, and in the opposite direction than for human observers (they suffer from *assimilation* with the inductor as opposed to *contrast* happening in human observers). Note that the inductors used by Ware & Cowan in their psychophysical experiments are slightly more saturated than those used in our *numerical psychophysics*. This is because we were using images expressed in digital counts. Nevertheless, this small difference in the inductors does not justify the differences in the corresponding pairs. Therefore, qualitative conclusions about the differences of behavior between networks and humans are valid.

2. The above results imply that extra assumptions can be done on top of linearity and hence they justify an analysis of the transfer functions of the networks in the Fourier domain.

Before going into the details of the above linear analysis, first we visually illustrate that the linear approximation done in Eq. 5 is reasonable by showing the output of a restoration network and its linearized version. Then we address points 1 and 2. Finally, we come back

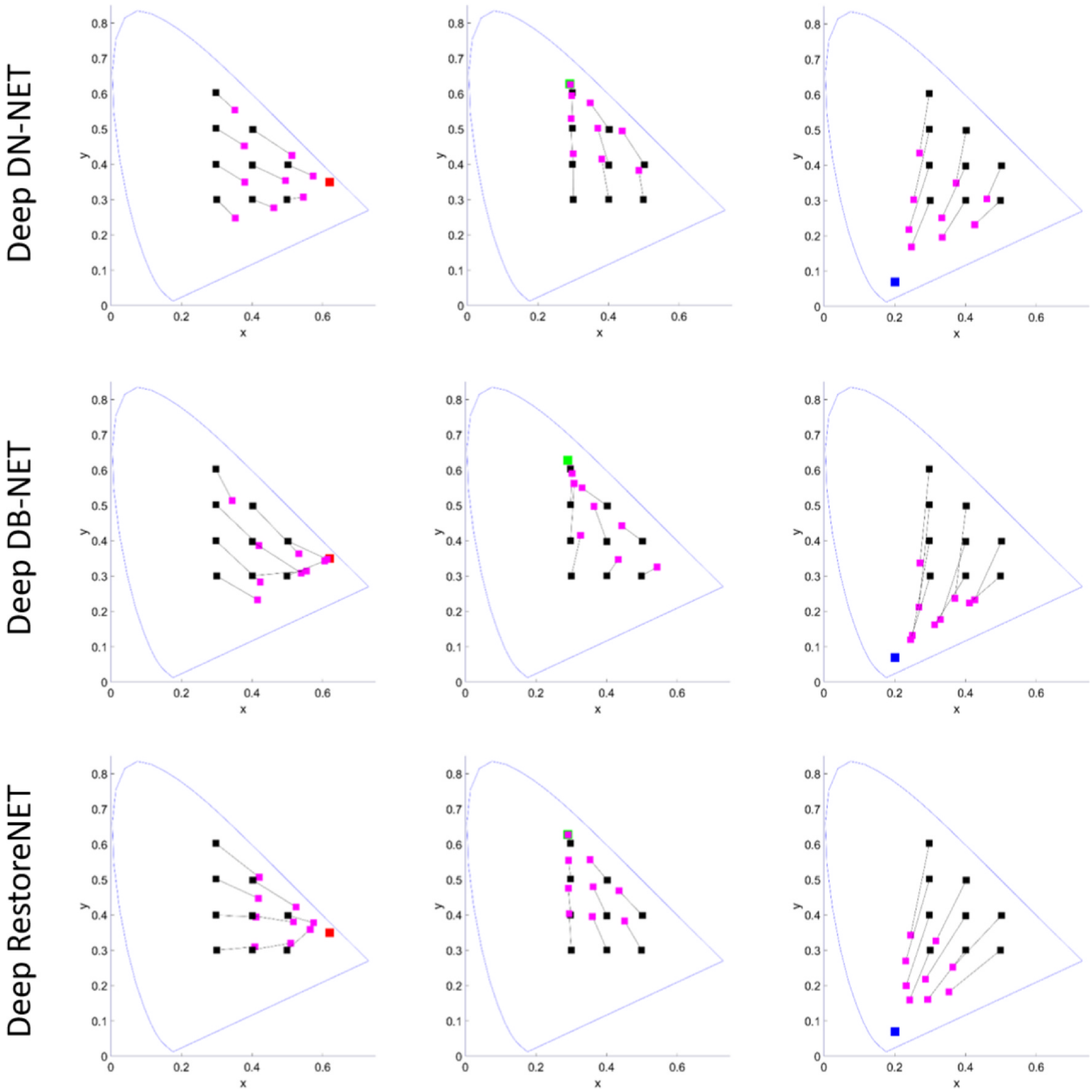

Fig. 9. Results for the deep CNNs studied in the Ware-Cowan corresponding pair experiment. We can see that the displacements of the 4 layer networks (Deep DN-NET, deep DB-NET, and deep RestoreNET) are comparable in magnitude to the human illusions, but again in the opposite direction than for human observers.

to the quantification of the nonlinear nature of the networks in discussing the amount of illusion depending on their complexity/flexibility (Tables 3 and 4).

4.1. *Linear approximation is representative and shift invariant*

Fig. 11 shows a representative example that illustrates the accuracy of the linear approximation. We show the behavior of the shallow RestoreNET and its linearized version with the kind of images used in the training and with the kind of stimuli used in the simulations of the visual illusions. We applied Eq. 6 to $1.3{\cdot}10^5$ image patches subtending 0.23 deg (16×16 pixels, i.e. vectors $\mathbf{i}$ and $\mathbf{r}$ of dimension $16 \times 16 \times 3 = 768$). Therefore this specific illustration required the pseudoinverse of a matrix of size $768 \times 1.3{\cdot}10^5$.

In this example we see that: (1) The network carries out the visual task (i.e. it is reducing the degradation). (2) The network visually behaves as classical restoration techniques (e.g. Wiener, Tikhonov Wiener et al., 1949; Tikhonov & Arsenin, 1977), see Gutiérrez, Ferri, and Malo (2006) for visual examples. (3) The network seems to have a stationary behavior, which may not be surprising given the stationary nature of the degradation learnt (signal-independent blur and noise). (4) Given the above, a linear version of the network may be sensible. (5) The intuitive meaningfulness of the linear approximation is confirmed by the results shown at the right column. First, note the visual resemblance of the actual and linear responses, particularly in natural images (where M_{Θ_k} came from). Second, note the large fraction of energy of the response captured by the linear approximation in these particular images (for a larger dataset of natural images the figure is about 93%). Incidentally, for this specific image the linear approximation to the CNN gives a slightly better restoration result than the actual CNN, but this is not representative of the whole natural image dataset. And finally, (6) note that the effect of the network on the Ware-Cowan image (bottom row of the figure) is *spreading the surround into the test*, thus leading to the assimilation effect. The linear approximation also has this effect.

Fig. 12 explicitly shows the matrix M_{Θ_k} for the RestoreNET example considered in Fig. 11. We can see in this figure (1) The existence of well defined submatrices in M_{Θ_k} (highlighted in red in Fig. 12) that correspond to similar spatial processing in the different chromatic channels. This suggests that the behavior of the network maybe *roughly* separable in chromatic and spatial terms. (2) The Toeplitz-like structure of the submatrices, that confirms the spatially stationary (roughly convolutional-like) behavior of the network. (3) The *equivalent* convolution

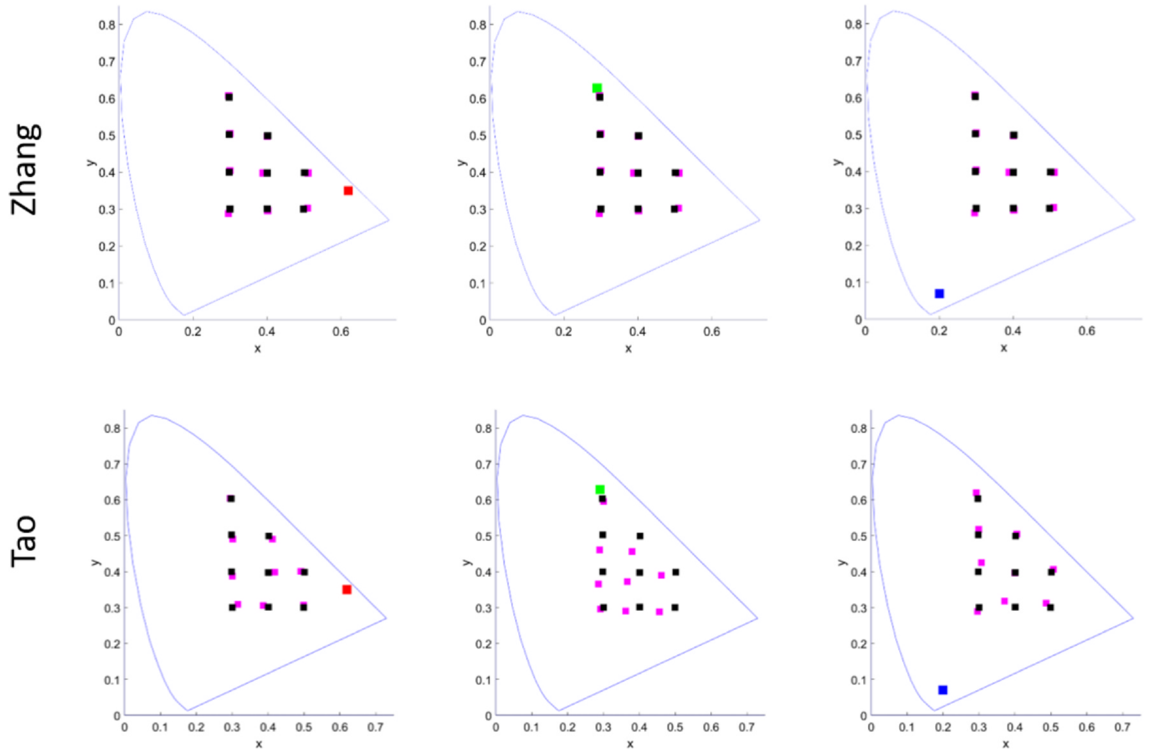

Fig. 10. Results for the state-of-art (really) deep networks of Zhang and Tao. The net of Zhang has the smallest displacements of all explored networks. However, in some cases we got contrast (departure from inductor) as in human observers, but in any case, illusions of very small magnitude. The net of Tao present largest displacements than that of Zhang, specially for the green inducer, and, it some cases it get contrast as in human observers.

Table 3
Nonlinearity, Performance & Illusion Strength (Denoising)

	Shallow	Deep	Zhang et al.
Fract. Lin. Resp.	90%	93%	84%
Error NonLin.	11.2%	10.8%	7.5%
Error Linear	12.5%	12.1%	15.9%
Illusion Strength	+ +	+ + +	-

Table 4
Nonlinearity, Performance & Illusion Strength (Deblurring)

	Shallow	Deep	Tao et al.
Fract. Lin. Resp.	88%	93%	94%
Error NonLin.	17.1%	16.3%	15.0%
Error Linear	19.0%	17.1%	15.6%
Illusion Strength	+ +	+ + +	+

kernels (*equivalent* receptive fields) of the network are a combination of a Gaussian-like blurring operator for the channel at hand, and center-surround operators at adjacent channels. And finally, (4) the width of the equivalent receptive fields explains how the surround spreads into the test.

These features qualitatively resemble the properties of LGN cells, but additional insight is definitely required. The diagonalization of the matrix M_{Θ_k} done in the next section helps to obtain extra intuition on the inner working of the network.

4.2. Eigenvector/ eigenvalue analysis

The eigendecomposition of the linear transform M_{Θ_k} identifies the stimuli that are considered by the system in a *special* way. By definition, the eigenfunctions, $\mathbf{b}^{(i)}$, are stimuli whose response is just an attenuated version of the input: $\lambda_i \mathbf{b}^{(i)} = M_{\Theta_k} \cdot \mathbf{b}^{(i)}$, and hence, $M_{\Theta_k} = B \cdot \lambda \cdot B^{-1}$, where $B = (\mathbf{b}^{(1)} \mathbf{b}^{(2)} \cdots \mathbf{b}^{(d)})$. Moreover, the eigendecomposition ranks the eigenfunctions according to the eigenvalues.

Therefore, the eigendecomposition describes the response of the network as a *linear autoencoder*:

$$\mathbf{r} = B \cdot \lambda \cdot B^{-1} \cdot \mathbf{i} \tag{7}$$

where the *rows* of B^{-1} contain the *encoding functions*, and the *columns* of B contain the *decoding functions*.

In this interpretation of the action of the network, the *encoder*, B^{-1}, transforms the input stimuli into a new representation. This is the inner eigenrepresentation of the network. In this inner representation, coefficients of the signal are dimension-wise attenuated by the diagonal matrix λ, and then the final response is synthesized by the *decoder* B.

Fig. 13 shows the eigenfunctions (columns of B) of the considered M_{Θ_k}. The most relevant stimuli for the network appear first.

The diagonalization of M_{Θ_k} shows that: (1) Eigenfunctions are oscillating stimuli of different frequencies extended over the spatial domain (stationary textures over the spatial domain). (2) Oscillations appear on the achromatic direction and in two *very specific* chromatic directions: namely *pink/green*, and *yellow-orange/blue*. (3) The most important functions are the achromatic ones and only afterwards there are functions that display chromatic variations (but also brightness oscillations of different frequency). These facts strongly suggest that the network is *implicitly* analyzing the stimuli in a *frequency-domain* representation in a color opponent space.

In order to clarify this intuition, we did the following analysis: first we computed the change of basis matrix that transforms the CIE XYZ primaries into the color basis defined by the extreme colors of the pink/green, yellow-orange/blue, and dark/light gray directions found in the eigenfunctions. This matrix allows to compute the color matching functions in the new basis. The perceptual meaningfulness of the *intrinsic* color basis of the network is demonstrated in Fig. 14.

Then, in order to estimate the spatial bandwidth of the network in these chromatic channels just found, we accumulated the spectra of the

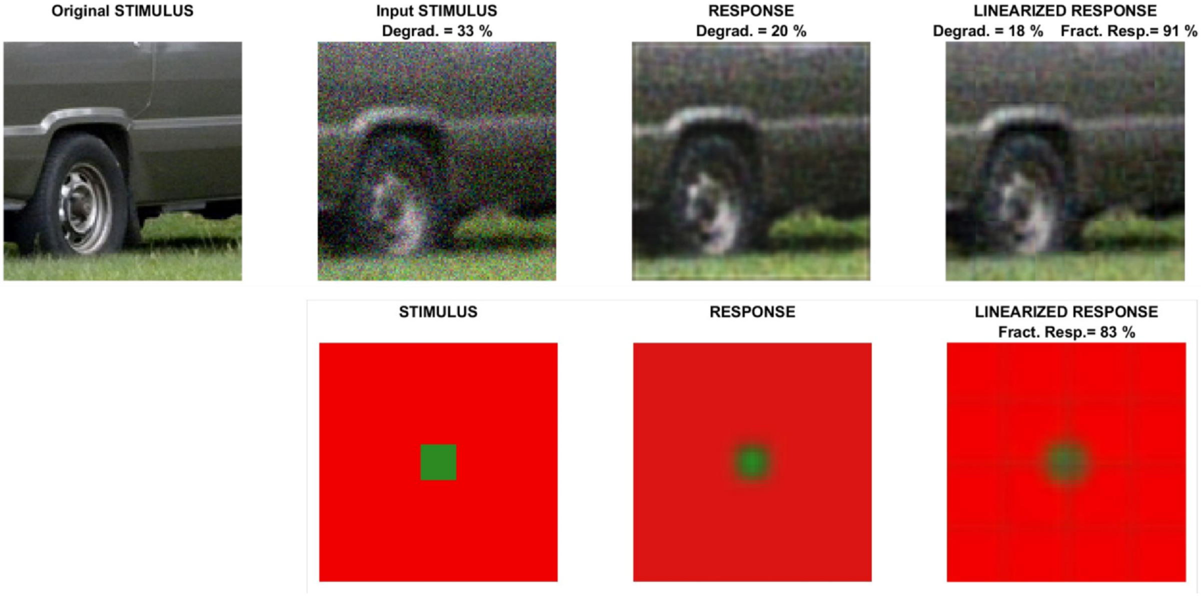

Fig. 11. Representative response of RestoreNET on a natural image and on a illusion inducing stimulus. In the restoration example (top) *Degradation* represents the RMSE of the considered signal (either input signal or response -restored signal-) referred to the square root of the average energy of the original stimulus. The *Fraction of Response* is the RMSE difference between the linear response and the actual (nonlinear) response referred to the square root of the average energy of the network response.

eigenfunctions decomposed in this color space and weighted the spectra by the corresponding eigenvalues. This is how images decomposed in this color space would be weighted when passing through the network. The result of such analysis is shown in Fig. 15. According to it, the intrinsic representation of the network can be interpreted as a color decomposition of stimuli in certain opponent color space (which is similar to human opponency), and the application of filters of markedly different bandwidth in the achromatic and the chromatic channels.

These filters could be compared with the achromatic and chromatic Contrast Sensitivity Functions (CSFs) of human viewers (Campbell et al., 1968; Mullen et al., 1985), but the frequency resolution of this eigenanalysis is limited by the size of the image patches used in computing the matrix M_{Θ_k}.

Note that the only assumption or approximation made so far is *linearity*. Fortunately, the properties of M_{Θ_k} and B found in the network allow us to make extra assumptions beyond linearity that make possible a more accurate analysis.

4.3. Spatial Fourier analysis in opponent channels

The properties found above for linear approximations using small-size image patches justify a straightforward Fourier analysis of the transfer functions of the network. After finding that the network *implicitly* operates in an opponent color space and that it is shift-invariant or stationary, and hence it has eigenfunctions which are Fourier-like, we did the following analysis. For 2046 full-size images subtending 1.83 deg (128×128 pixels) we computed the quotient of the Fourier spectra of the input stimuli and the output responses, both decomposed in a classical opponent space (Hurvich et al., 1957), the one of the color matching functions represented in Fig. 14 (right). In this way, the filters will be directly comparable to the human CSFs. These filters are shown in Fig. 16, and the actual CSFs are plotted for convenient reference in

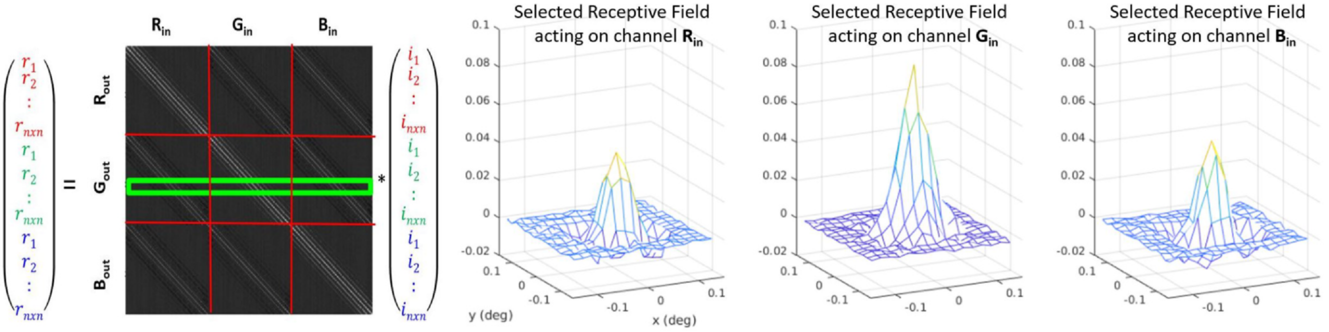

Fig. 12. Linear version of RestoreNET: the matrix M_Θ. In order to interpret the matrix (at the left) remember that, as we pointed at the beginning of the section, here we are rearranging $n \times n$ color images as column vectors of dimension $n \times n \times 3$. Specifically, the pixels of each RGB channel are organized as a column, and vertically stacked one after the other into a single vector. This is represented by the i_x elements in red, green and blue. In this linear approximation, Eq. 5, every row of the matrix (as for instance the one highlighted in green) acts on the input column vector arranged in this specific way. Therefore, as each response is the scalar product of the corresponding row times the input vector, these rows represent the *receptive fields* of the linear version of the network. Large submatrices highlighted here by red lines represent the spatial processing within each color channel, and then these responses are linearly combined to lead to the final responses. The properties of the receptive field corresponding to one row of the matrix are more evident by undoing the vector arrangement. The surfaces at the right correspond to the spatial arrangement of the R, G and B portions of the receptive field weights in the highlighted row. Then, we can see that this is a center-surround sensor tuned to the central location of the image with excitatory center mainly tuned to green (it receives more input from G) and inhibitory purple surround (with an input of the form -(R + B)).

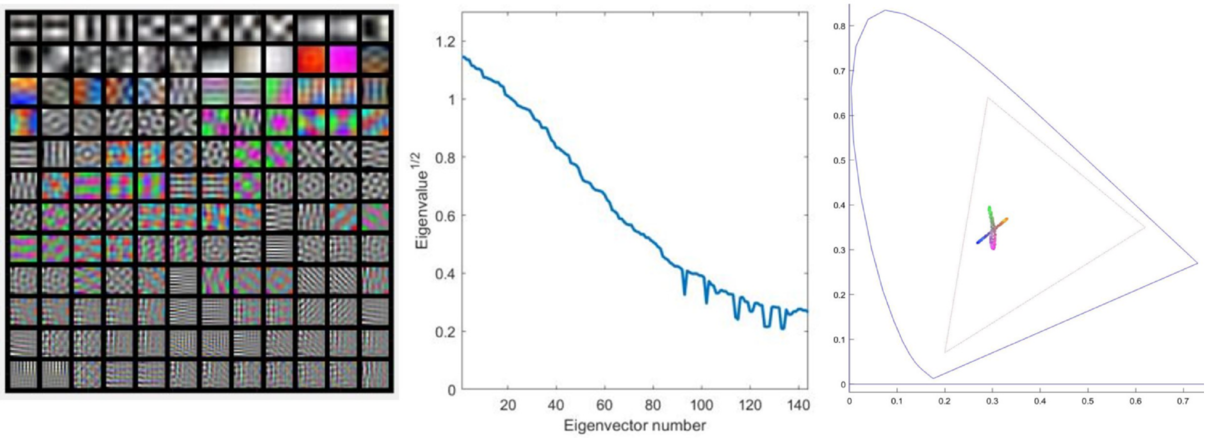

Fig. 13. Eigenfunctions (left) and eigenvalues (center) of the linear network. Eigenfunctions (columns of B) are sorted sorted (left-to-right and top-to-bottom) according to the associated eigenvalues. The CIE xy chromatic diagram (right) displays the chromatic coordinates of the colors in the 50 eigenfunctions with bigger eigenvalue. Note that as the input stimuli in the considered dataset are given in digital counts the eigenfunctions are also given in digital counts. Therefore, the representation of the colors of the eigenfuncitons in the chromatic diagram involved the assumption of a standard display calibration conversion (Malo & Luque, 2002). Note also that, given the fact that M_{ϑ_k} is almost symmetric (see Fig. 12), B is almost orthonormal and hence, the *encoding* functions (rows of B^{-1}, not shown) are very similar to the *decoding* functions represented here.

the bottom row of Fig. 17.

In this context in which models and linear approximations are obtained from large image databases an important safety check was necessary. For this specific illustration we not only trained the networks with images from the massive database CLS-LOC (Russakovsky et al., 2015) (uncalibrated images and eventually subject to uncontrolled manipulations), but we also did a separate training with images coming from two calibrated databases (images in CIE XYZ with no spatial manipulation Gutmann, Laparra, Hyvärinen, & Malo, 2014; Laparra, Jimenez, Camps, & Malo, 2012; Parraga, Vazquez-Corral, & Vanrell, 2009; Vazquez-Corral, Párraga, Baldrich, & Vanrell, 2009). Results were qualitatively the same (see Fig. 17 first row). This implies that the database CLS-LOC can be trusted with regard to the average spatio-chromatic spectra (covariance matrix) of the image samples.

A similar conclusion to that of Fig. 13 can be drawn by looking at the first few layers of other deep networks (Krizhevsky, Sutskever, & Hinton, 2012). This is a related result, but note that the eigenanalysis is different and more powerful in some ways. In Krizhevsky et al. (2012) the frequency filters in color opponent spaces explicitly emerge in the weights of the linear part of the layers, while here the linear approximation displays center-surround filters, and not narrow frequency sensors. In our case, these frequency sensors are revealed only after the eigenanalysis of the learned weights. More interestingly, the eigenanalysis shows the different gain (or sensitivities) for the different frequencies (the eigenvalues), which is not obvious from the raw receptive fields in Krizhevsky et al. (2012). The frequency sensitivities in opponent channels is what can be compared to the human CSFs. Such comparisons with specific psychophysical results (CSFs and spectral sensitivities) are not mentioned at all in Krizhevsky et al. (2012) nor computable from their results.

4.4. Linear approximation and strength of illusions

The linear approximation of the simpler networks (2-hidden layers and 4-hidden layers) reveals a number of human-like characteristics in their intrinsic image representation, namely the chromatic opponent channels and filters of bandwidths similar to the CSFs.

In relatively rigid networks (simpler architectures) the emergence of this specific frequency selectivity to fulfill the low-level visual task explains that color and luminance profiles in the stimuli are distorted in the response of the network in specific ways. The responses at certain region changes depending on the spatial context (e.g. Figs. 5–7), thus

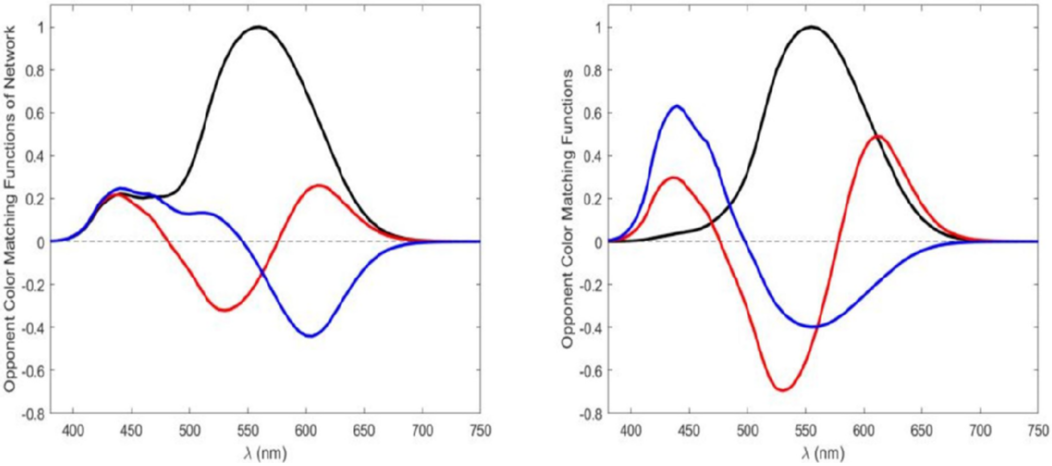

Fig. 14. Intrinsic color matching functions of the network (left) compared with the classical opponent color-matching functions of human color vision (by Hurvich et al. (1957), on the right). Both systems of primaries have an achromatic channel (all-positive color matching function in black), and two opponent chromatic channels (with positive and negative values).

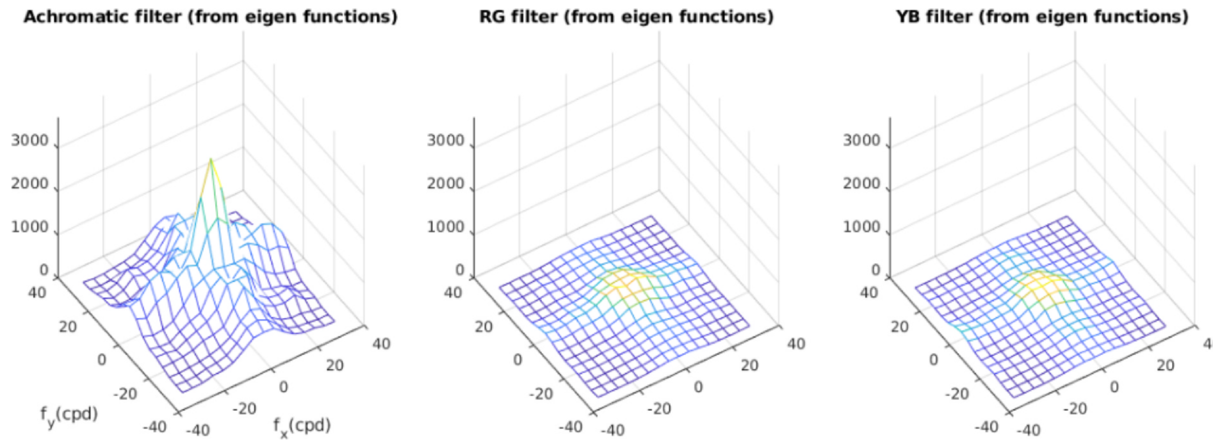

Fig. 15. Accumulated spectra of eigenfunctions decomposed in their intrinsic color space and weighted by eigenvalues. Limited frequency resolution is due to the fact that this result comes from small 16×16 image blocks. This may give rise to artifacts in the spectra.

leading to significant visual illusions in these networks.

The emergence of these properties can be understood from the relation of image restoration methods with the statistics of natural images: if the goal is removing non-natural features, the network will learn transfer functions matched to the statistics of the signal to filter out undesired features. And it is known that the covariance of natural colors (Buchsbaum et al., 1983; Lee, Wachtler, & Sejnowski, 2002) and natural color images (Wachtler, Lee, & Sejnowski, 2001) is consistent with the opponent color representation and eigenfunctions found by our analysis.

As a by-product, the effect of the context due to the width of the equivalent kernels of the linear network explains the shifts in perceived brightness in similar directions to that happening in humans. But the simplicity of this behavior, e.g. as in Fig. 11, also explains why the illusions may be markedly different from those of humans in some color cases.

For instance, when the spatial layout of the stimulus is relatively simple, as in the *Gradient color* illusion or in the center-surround setting in the Ware-Cowan experiment, the simple filtering found in Fig. 11 only leads to penetration of the surround in the region corresponding to the test thus leading to *assimilation* instead of the *contrast* found in human observers. This eventually simple behavior revealed by the linear analysis certainly applies to the six models with not that many layers (DN-Net, Deep DN-Net, etc.).

However, the situation may be different for more flexible architectures for which the linear filter view may not be that appropriate, like the 17-layer CNN of Zhang et al. (2017) or the 21-layer CNN of Tao et al. (2018). In the following we compare the quality of the simpler CNNs trained in this work with the state-of-the-art networks and their respective linear approximations.

In Tables 3 and 4, we list the following descriptors. The *Fraction of Linear Response* is the proportion of the energy of the response explained by the linear approximation. The *error* measurements (either for the nonlinear network or for its linear approximation) correspond to the fraction of the energy of the clean signal not recovered by network (either in denoising or in deblurring), and the final row qualitatively describes the magnitude of illusions found.

In these tables, the linear approximations of the networks were estimated using Eq. 6 and $2{\cdot}10^3$ input–output pairs of image patches of size 20×20 from the CLS-LOC 2014 ImageNet validation dataset (Russakovsky et al., 2015). In the denoising case the images were degraded with the same kind of distortions used in the training of the shallow and deep networks. In the deblurring case the images were degraded with Gaussian blur of the same width used in the training and with double width. The fraction of the response captured by the linear approximation and the performance measure were computed with 10^5 image patches not included in the training set of the linear approximation. The deviation over 10 realizations was found to be about 0.1% in all cases, and hence not included in the table for the sake of clarity.

In the specific case of the simpler networks (Shallow and Deep), the increased complexity does not make a big difference in terms of their nonlinear nature, which explains the similarity of their intrinsic filters and the slight improvement in performance for the deeper net. However, the state-of-the-art network of Zhang el al. is the more nonlinear. Interestingly, the state-of-the-art network for deblurring is very well described by a linear approximation.

From a pure machine learning perspective, it is obvious that the nonlinear nature of the networks and their performance in the goal have to increase when substantially increasing the number of parameters. However, regarding the magnitude of the illusions (and more in general, regarding the eventual similarity with the visual system) this does not necessarily increase with the complexity of the model.

This can be understood in the following way: increasing the complexity usually leads to systems that are too specialized in the specific goal. Therefore, it is reasonable that the networks by Zhang et al. (2017) and Tao et al. (2018) do not show visual illusions with the considered stimuli, because these stimuli lack the perturbations (noise, blur) that these highly specialized deep neural networks were trained to remove.

5. Discussion and final remarks

This work confirms and expands our original report (Gomez-Villa et al., 2019) on color and brightness illusions suffered by CNNs trained to solve low-level visual tasks. Specifically, we explored a range of five classical brightness illusions and their color counterparts (a total of 10 different illusions) to point out the existence of illusions in CNNs and assess their qualitative correspondence with human behavior. Additionally, we proposed a quantitative comparison by studying CNN illusions through asymmetric color matching experiments as done by humans (Ware et al., 1982). In those experiments we explored simple CNN architectures (with 2 or 4 hidden layers) trained for image denoising, deblurring and restoration (simultaneous denoising and deblurring). And we also studied the behavior of recent, much deeper CNNs trained for the same kind of tasks: the 17-layer architecture of Zhang et al. (2017) pretrained for denoising and the 21-layer architecture of Tao et al. (2018) pretrained for deblurring.

Qualitative analysis shows that the simpler networks do modify their response in the same direction as the humans in most cases, and that the more complex networks lead to negligible or non-human illusions.

On the other hand, quantitative results on asymmetric color

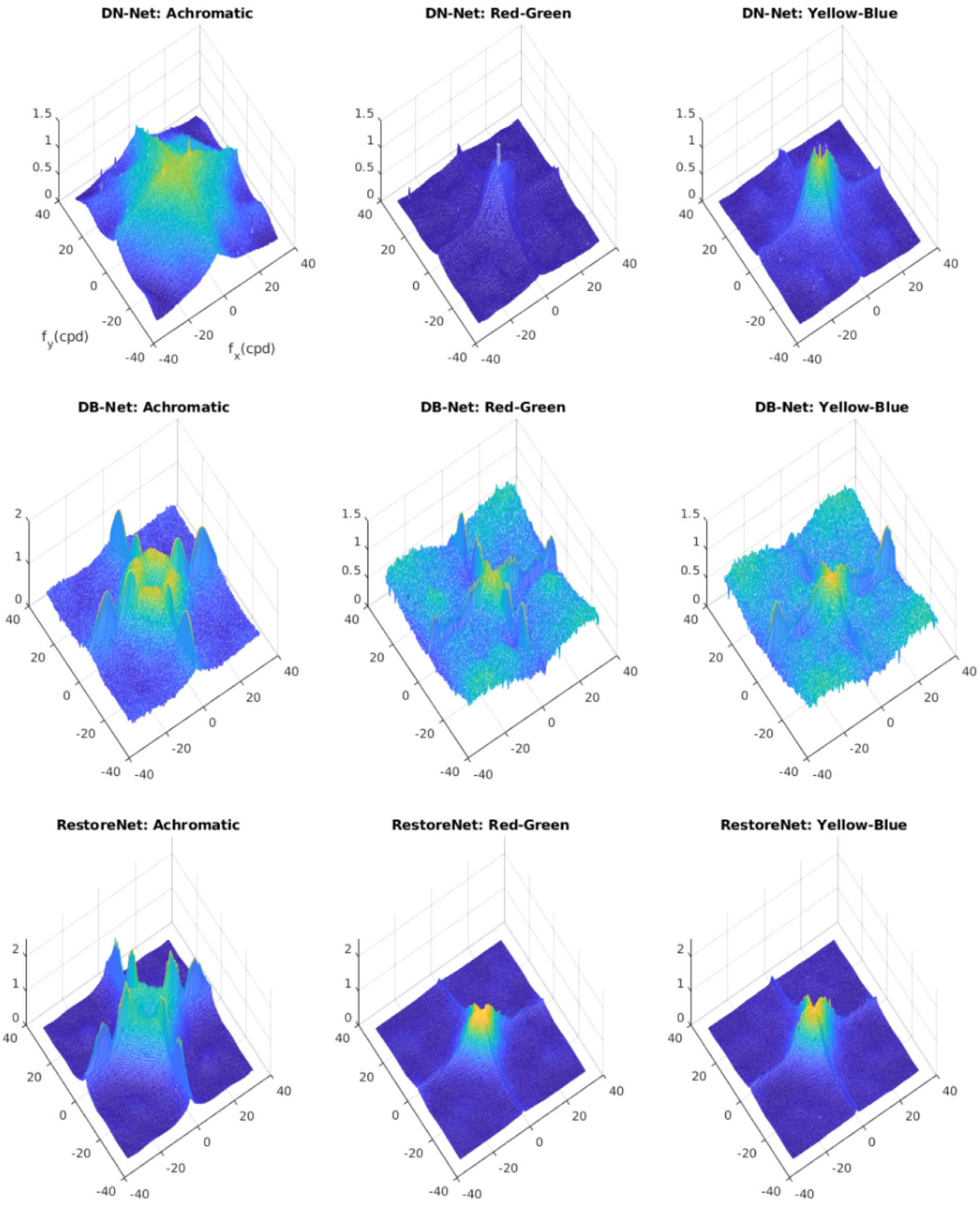

Fig. 16. Filters estimated for DB-net, DN-net and RestoreNET assuming a classical opponent space and assuming the Fourier representation. The filters for the deeper version of these networks are similar. Note that, following classical ideas from optimal filtering and regularization (Wiener et al., 1949; Tikhonov & Arsenin, 1977), the denoising filters happen to be low-pass, the deblurring filters happen to be highpass, but of course, also preserving the low-frequencies, and the restoration filters are a combination of both.

matching show that the simpler networks in center-surround settings have substantial illusions, but of opposite nature to those of humans (while a human observer perceives chromatic contrast, the CNN shows assimilation) and the much deeper networks display illusions which are either negligible or of less magnitude than humans.

The proposed eigenanalysis of simple networks reveals interesting similarities with human vision, showing that these simple networks *implicitly* operate in an opponent color space, with low-pass filtering for the chromatic channels and band-pass filtering for the luminance. This simple linear description may explain why in the color-matching experiment the CNNs suffer from assimilation, unlike humans: in center-surround settings, the low-pass nature of the filtering in the chromatic channels shifts the hue of the test towards the hue of the surround.

From the results and associated analysis, these considerations may

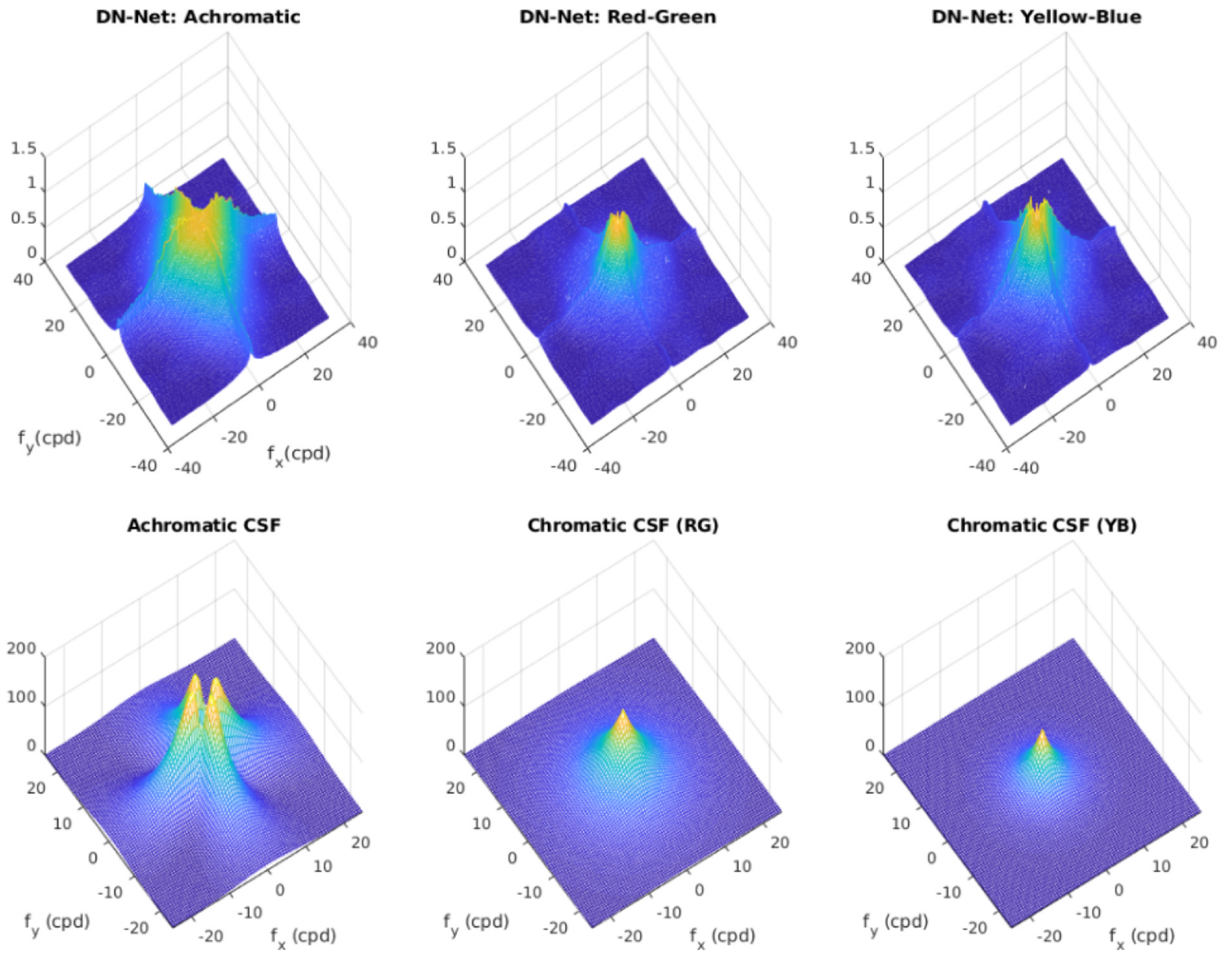

Fig. 17. Top row: Filters estimated for DB-net, trained with images coming from two calibrated databases (images in CIE XYZ with no spatial manipulation). Results are qualitatively the same that for the large uncalibrated dataset. Bottom row: Human CSFs for convenient reference (Achromatic CSF from the Standard Spatial Observer [?], and chromatic CSFs from Mullen et al. (1985)).

follow.*Visual illusions and image statistics*: The findings in this work are consistent with the long-standing tradition in vision science that considers low-level visual illusions as by-product of optimization of visual systems to perform basic tasks in natural environments (Barlow, 1990; Clifford et al., 2000; Clifford et al., 2002; Clifford et al., 2007). More specifically, the results of our linear analysis link the behavior of the networks with (a) the statistical basis of the image restoration problem and, more interestingly, with (b) optimal coding theories of human vision. First, it is interesting to note that our linear analysis of the networks leads to functions that resemble the principal directions of natural colors (Buchsbaum et al., 1983; Lee et al., 2002), and natural color images (Wachtler et al., 2001; Hyvärinen, Hurri, & Hoyer, 2009). However, note that in our analysis we are not diagonalizing the covariance matrix of natural signals. In fact, signal decorrelation or independence was not enforced in anyway. Therefore, although similar, our result should not be attributed to information maximization or sparse coding (Olshausen et al., 1996; Wachtler et al., 2001). Instead, as stated above, if the goal is removing non-natural features from stimuli, Wiener or Tikhonov ideas (Wiener et al., 1949; Tikhonov & Arsenin, 1977) naturally lead to filters matched to the signal spectrum. With this in mind, and given the relation between the average spectrum of the signals, their autocorrelation and their covariance, it makes sense that the optimal filter obtained from our linear approximation is very similar to the covariance of the natural signals. Therefore, our result to reconstruct signals with minimum error turns out to be very similar to the PCA result. Nevertheless, the reason why CNNs trained for image restoration develop opponent chromatic channels and CSF-like filters would be more in line with signal/noise explanations of visual function (Atick, Li, & Redlich, 1992). Note that error minimization and information maximization are similar, but not the same (see Lloyd et al., 1982 for the original account, and see Twer & MacLeod, 2001; MacLeod et al., 2003; Laparra et al., 2012; Laparra et al., 2015 for sequels in vision science).*Some implications on the use of artificial neural networks to study vision* The analysis of the fraction of response captured by the linear approximation would suggest that more 'rigid' (less non-linear) networks suffer from stronger illusions; a possibility is that the key for a quantitative replication of human behavior resides in the (small or even tiny) nonlinear part. But more importantly, our psychophysical-like analysis of ANNs shows that while they are deceived by illusions, their response might be significantly different to that of humans. These discrepancies with humans in quantitative experiments imply a word of caution on using ANNs to study human vision, a point that has been getting significant attention lately, e.g. see Jacob et al. (2019), Geirhos et al. (2020) and references therein. In particular, when fitting flexible nonlinearities to specific goals it is easy to miss basic psychophysical phenomena if the proper precautions are not taken (Martinez et al., 2019).

More generally, as mentioned earlier, ANNs were inspired by classical biological models of vision, and for this reason they share the L + NL formulation (Haykin et al., 2009) of the "standard" model of vision (Olshausen et al., 2005). But this model is questioned in the vision science literature.

Vision models and ANNs use L + NL modules derived from fitting some data, and in every case either the linear filters are constant or the models do not have general rules as to how the filters should be modified depending on the input (Betz, Shapley, Wichmann, & Maertens, 2015; Li, Wang, Hu, & Yang, 2019). This is an essential weakness of all these models as visual adaptation shows. Visual adaptation, an essential feature of the neural systems of all species, produces a change in the input–output relation of the system that is driven by the stimuli (Wark, Fairhall, & Rieke, 2009). Therefore, it requires that the linear and/or the nonlinear stages of a L + NL model change with the input in order to explain neural responses (Meister et al., 1999; Coen-Cagli, Dayan, & Schwartz, 2012; Jansen et al., 2018), thus resulting in a crucial weakness of these L + NL models.

L + NL models are not tests of how well the linear filter of a neuron describes its behavior, they have been obtained simply by *assuming* that the neuron performs a linear summation and then searching for the best-fitting linear model. In visual perception, experimental data contradicts in many situations the central notions of L + NL models (Wandell et al., 1995), which fail to predict image appearance in the general case (Fairchild et al., 2013). The state-of-the-art deep learning metric for perceived appearance (Zhang et al., 2018) – designed to predict perceptual image error like human observers and trained on a large scale dataset of 160 K images with close to 500 K human judgements- has been shown to correlate poorly with human observer responses (Zamir et al., YYYY; Bertalmío et al., 2019). In visual neuroscience, the standard model was able to explain in 2005 at the most a 40% of the data variance in V1 (Olshausen et al., 2005), and fifteen years later this value has increased just to ~ 50% (Cadena et al., 2019). The prevailing belief, as pointed out in Jacob et al. (2019), appears to be that ANNs can be treated as accurate models of vision and that the differences are only a matter of degree that will eventually be solved. But the limited performance of even deep ANNs suggests that a much more complex network nonlinearity is at work in the visual system than what L + NL models are capturing (Carandini et al., 2005). Even further, it has even been proposed (Olshausen, 2013) that the standard model is not just in need of revision, but it is the wrong starting point and needs to be discarded altogether.

CRediT authorship contribution statement

A. Gomez-Villa, A.Martín, J. Vazquez-corral, M. Bertalmío, J. Malo: Conceptualization, Methodology, Software, Data curation, Writing- Original draft preparation, Visualization, Investigation, Writing- Reviewing and Editing. **M. Bertalmío, J. Malo:** Supervision.

Acknowledgements

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 761544 (project HDR4EU) and under grant agreement number 780470 (project SAUCE), and by the Spanish government and FEDER Fund, grant Ref. PGC2018-099651-B-I00 (MCIU/AEI/FEDER, UE). The work of AM was supported by the Spanish government under Grant FJCI-2017–31758. JM has been supported by the Spanish government under the MINECO grant Ref. DPI2017-89867 and by the Generalitat Velanciana grant Ref. GrisoliaP-2019-035. We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan Xp GPU used for this research.

References

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., et al. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems, software available from tensorflow.org. URL:http://tensorflow.org/.

Abrams, A., Hillis, J., & Brainard, D. (2007). The relation between color discrimination and color constancy: When is optimal adaptation task dependent? *Neural Computation, 19*(10), 2610–2637.

Adelson, E. H. 2000. Lightness perception and lightness illusions. New Cognitive Neurosciences 339.

Atick, J., Li, Z., & Redlich, A. (1992). Understanding retinal color coding from first principles. *Neural Computation, 4*(4), 559–572.

Atick, J., Li, Z., & Redlich, A. (1993). What does post-adaptation color appearance reveal about cortical color representation? *Vision Research, 33*(1), 123–129.

Attneave, F. (1954). Some informational aspects of visual perception. *Psychological Review, 61*(3), 183.

Barlow, H. (1961). Possible principles underlying the transformation of sensory messages. *Sensory Communication, 1*, 217–234.

Barlow, H. (1990). Vision: Coding and efficiency, Cambridge, UK: Cambridge Univ. Press. Ch. A theory about the functional role and synaptic mechanism of visual aftereffects.

Benjamin, A.S., Qiu, C., Zhang, L. -Q., Kording, K. P. & Stocker, A. A. (2019). Shared visual illusions between humans and artificial neural networks. Proceedings of conference on cognitive computational neuroscience.

Bertalmío, M. (2019). *Vision models for high dynamic range and wide colour gamut imaging: Techniques and applications.* Academic Press.

Betz, T., Shapley, R., Wichmann, F. A., & Maertens, M. (2015). Testing the role of luminance edges in white's illusion with contour adaptation. *Journal of Vision, 15*(11), 14.

Blakemore, C., & Campbell, F. W. (1996). On the existence of neurones in the human visual system selectively sensitive to the orientation and size of retinal images. *Journal of Physiology, 203*, 237–260.

Bressan, P. (2001). Explaining lightness illusions. *Perception, 30*(9), 1031–1046.

Bruke, E. (1865). uber erganzungs und contrasfarben. Wiener Sitzungsber, 51.

Buchsbaum, G., & Gottschalk, A. (1983). Trichromacy, opponent colours coding and optimum colour information transmission in the retina. *Proceedings of the Royal Society B, 220*(1218), 89–113.

Cadena, S. A., Denfield, G. H., Walker, E. Y., Gatys, L. A., Tolias, A. S., Bethge, M., & Ecker, A. S. (2019). Deep convolutional models improve predictions of macaque v1 responses to natural images. *PLoS Computational Biology, 15*(4), Article e1006897.

Campbell, F. W., & Robson, J. (1968). Application of fourier analysis to the visibility of gratings. *The Journal of Physiology, 197*(3), 551–566.

Campbell, F. W., & Robson, J. G. (1968). Application of fourier analysis to the visibility of gratings. *The Journal of Physiology, 197*(3), 551.

Carandini, M., Demb, J. B., Mante, V., Tolhurst, D. J., Dan, Y., Olshausen, B. A., Gallant, J. L., & Rust, N. C. (2005). Do we know what the early visual system does? *Journal of Neuroscience, 25*(46), 10577–10597.

Carandini, M., & Heeger, D. (2012). Normalization as a canonical neural computation. *Nature Reviews. Neuroscience, 13*(1), 51–62.

Clifford, C. W. (2002). Perceptual adaptation: Motion parallels orientation. *Trends in Cognitive Sciences, 6*, 136–143.

Clifford, C., Webster, M., Stanley, G., Stocker, A., Kohn, A., Sharpee, T., & Schwartz, O. (2007). Visual adaptation: Neural, psychological and computational aspects. *Vision Research, 47*, 3125–3131.

Clifford, C. W., Wenderoth, P., & Spehar, B. (2000). A functional angle on some aftereffects in cortical vision. *Proceedings of the Royal Society B, 267*, 1705–1710.

Coen-Cagli, R., Dayan, P., & Schwartz, O. (2012). Cortical surround interactions and perceptual salience via natural scene statistics. *PLoS Computational Biology, 8*(3), Article e1002405.

Corney, D., & Lotto, R. B. (2007). What are lightness illusions and why do we see them? *PLoS Computational Biology, 3*(9), 1790–1800.

DeValois, R. L., & DeValois, K. K. (1990). *Spatial vision.* Oxford University Press.

Fairchild, M. D. & Heckaman, R. L. (2013). Metameric observers: a monte carlo approach, in: Color and imaging conference (Vol. 2013, pp. 185–190). Society for Imaging Science and Technology.

Foley, J., & Chen, C. (1997). Analysis of the effect of pattern adaptation on pattern pedestal effects: A two-process model. *Vision Research, 37*(19), 2779–2788.

Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A. & Brendel, W. (2019). Imagenet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. In International conference on learning representations. URL:https://openreview.net/forum?id=Bygh9j09KX.

Geirhos, R., Jacobsen, J.-H., Michaelis, C., Zemel, R., Brendel, W., Bethge, M. & Wichmann, F. (2020). A. Shortcut learning in deep neural networks, arXiv preprint arXiv:2004.07780.

George Mather, G. C., Pavan, A., & Casco, C. (2008). The motion aftereffect reloaded. *Trends in Cognitive Sciences, 12*, 482–487.

Gomez-Villa, A., Martin, A., Vazquez-Corral, J., & Bertalmio, M. (2019). Convolutional neural networks can be deceived by visual illusions, in. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 12309–12317). .

Gutiérrez, J., Ferri, F. J., & Malo, J. (2006). Regularization operators fornatural images based on nonlinear perception models. *IEEE Transactions on Image Processing, 15*(1), 189–200.

Gutmann, Laparra, Hyvärinen, & Malo (2014). Spatiochromatic adaptation via higher-order canonical correlation analysis of natural images. *PloS ONE.*

Haykin, S. (2009). *Neural networks and learning machines.* New York: Prentice Hall.

Heinemann, E. G. (1955). Simultaneous brightness induction as a function of inducing- and test-field luminances. *Journal of Experimental Psychology, 50*(2), 89.

Hillis, J. M., & Brainard, D. (2005). Do common mechanisms of adaptation mediate color discrimination and appearance? *JOSA A, 22*(10), 2090–2106.

Hong, S. W., & Shevell, S. K. (2004). Brightness contrast and assimilation from patterned inducing backgrounds. *Vision Research, 44*(1), 35–43.

Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology, 148*(3), 574–591.

Hurvich, L., & Jameson, D. (1957). An opponent-process theory of color vision. *Psychology Review, 64*(6), 384–404.

Hyvärinen, A., Hurri, J., & Hoyer, P. (2009). *Natural image statistics: A probabilistic approach to early computational vision.* Heidelberg, Germany: Springer-Verlag.

Jacob, G., Pramod, R. T., Katti, H. & Arun, S. P. 2019. Do deep neural networks see the way we do?, bioRxiv arXiv:https://www.biorxiv.org/content/early/2020/03/05/860759.full.pdf, doi:10.1101/860759. URL:https://www.biorxiv.org/content/early/2020/03/05/860759.

Jansen, M., Jin, J., Li, X., Lashgari, R., Kremkow, J., Bereshpolova, Y., Swadlow, H. A., Zaidi, Q., & Alonso, J.-M. (2018). Cortical balance between on and off visual responses is modulated by the spatial properties of the visual stimulus. *Cerebral Cortex, 29*(1), 336–355.

Kim, B., Reif, E., Wattenberg, M. & Bengio, S. Do neural networks show gestalt phenomena? an exploration of the law of closure, arXiv preprint arXiv:1903.01069.

Kitaoka, A. (2005). Illusion and color perception 29, 150–151.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. In F. Pereira, C. J. C. Burges, L. Bottou, & K. Q. Weinberger (Eds.). *Advances in neural information processing systems* (pp. 1097–1105). Curran Associates Inc 25.

Laparra, V., Jimenez, S., Camps, G., & Malo, J. (2012). Nonlinearities and adaptation of color vision from sequential principal curves analysis. *Neural Computation, 24*(10), 2751–2788.

Laparra, V., & Malo, J. (2015). Visual aftereffects and sensory nonlinearities from a single statistical framework. *Frontiers in Human Neuroscience, 9*, 557. https://doi.org/10.3389/fnhum.2015.00557.

Lee, T.-W., Wachtler, T., & Sejnowski, T. J. (2002). Color opponency is an efficient representation of spectral properties in natural scenes. *Vision Research, 42*(17), 2095–2103.

Linsley, D., Kim, J., Ashok, A. & Serre, T. (2019). Recurrent neural circuits for contour detection. In International conference on learning representations.

Li, X., Wang, W., Hu, X., & Yang, J. (2019). Selective kernel networks. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 510–519). .

Lloyd, S. P. (1982). Least squares quantization in PCM. *IEEE Transactions on Information Theory, 28*, 129–137.

Loomis, J. M. (1972). The photopigment bleaching hypothesis of complementary after-images: A psychophysical test. *Vision Research, 12*, 1587–1594.

MacLeod, D. A. (2003). Colour discrimination, colour constancy, and natural scene statistics. In J. Mollon, J. Pokorny, & K. Knoblauch (Eds.). *Normal and defective colour vision* (pp. 189–218). Oxford University Press.

Martinez, M., Bertalmío, M., & Malo, J. (2019). In paraise of artifice reloaded: Caution with natural image databases in modeling vision. *Frontiers in Neuroscience, 13*(8), https://doi.org/10.3389/fnins.2019.00008.

Malo, J. & Luque, M. 2002. Colorlab: A color processing toolbox for matlab, Internet site: http://www.uv.es/vista/vistavalencia/software.html.

Martinez-Garcia, M., Cyriac, P., Batard, T., Bertalmío, M., & Malo, J. (2018). Derivatives and inverse of cascaded linear + nonlinear neural models. *PLOS One, 13*(10), 1–49. https://doi.org/10.1371/journal.pone.0201326 URL:https://doi.org/10.1371/journal.pone.0201326.

McCourt, M. E. (1982). A spatial frequency dependent grating-induction effect. *Vision Research, 22*(1), 119–134.

Meister, M., & Berry, M. J. (1999). The neural code of the retina. *Neuron, 22*(3), 435–450.

Morgan, J. A. S. M., & Chubb, C. (2011). Evidence for a subtractive component in motion adaptation. *Vision Research, 51*, 2312–2316.

Morgan, M., Chubb, C., & Solomon, J. (2006). Predicting the motion after-effect from sensitivity loss. *Vision Research, 46*, 2412–2420.

Mullen, K. T. (1985). The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings. *The Journal of Physiology, 359*(1), 381–400.

Olshausen, B. A. (2013). 20 years of learning about vision: Questions answered, questions unanswered, and questions not yet asked, in: 20 Years of computational neuroscience. Springer. pp. 243–270.

Olshausen, B. A., & Field, D. J. (1996). Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature, 381*(6583), 607.

Olshausen, B. A., & Field, D. J. (2005). How close are we to understanding v1? *Neural Computation, 17*(8), 1665–1699.

Parraga, C. A., Vazquez-Corral, J., & Vanrell, M. (2009). A new cone activation-based natural images dataset. *Perception, 36*(Suppl), 180.

Purves, D., & Lotto, R. B. (2003). Why we see what we do: An empirical theory of vision. *Sinauer Associates*.

Ratliff, F. (1965). *Mach bands: Quantitative studies on neural networks.* San Francisco London Amsterdam: Holden-Day.

Ross, J., & Speed, H. (1991). Contrast adaptation and contrast masking in human vision. *Proceedings of the Royal Society of London, 246*, 61–69.

Russakovsky, et al. (2015). Imagenet large scale visual recognition challenge. International Journal of Computer Vision 115 (3), 211–252.

Samek, W., & Explainable, A. I. (2019). *Interpreting, explaining and visualizing deep learning.* Springer Nature.

Sun, E. D. & Dekel, R. 2019. Imagenet-trained deep neural network exhibits illusion-like response to the scintillating grid, arXiv preprint arXiv:1907.09019.

Tao, X., Gao, H., Shen, X., Wang, J., & Jia, J. (2018). Scale-recurrent network for deep image deblurring. *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 8174–8182). .

Tikhonov, A. N., & Arsenin, V. I. (1977). *Solutions of ill-posed problems, Vol. 14.* Washington, DC: Winston.

Twer, T., & MacLeod, D. I. (2001). Optimal nonlinear codes for the perception of natural colours. *Network: Computation in Neural Systems, 12*(3), 395–407.

Vazquez-Corral, J., Párraga, C., Baldrich, R. & Vanrell, M. (2009). Color constancy algorithms: Psychophysical evaluation on a new dataset. Journal of Imaging Science and Technology 53 (3), 31105-1–31105-9.

Wachtler, T., Lee, T.-W., & Sejnowski, T. J. (2001). Chromatic structure of natural scenes. *JOSA A, 18*(1), 65–77.

Wandell, B. A. (1995). *Foundations of vision, Vol. 8.* MA: Sinauer Associates Sunderland.

Ward, E. J. (2019). Exploring perceptual illusions in deep neural networks. In https://www.biorxiv.org/content/10.1101/687905v1, 2019.

Ware, C., & Cowan, W. B. (1982). Changes in perceived color due to chromatic interactions. *Vision Research, 22*(11), 1353–1362. https://doi.org/10.1016/0042-6989(82)90225-5.

Wark, B., Fairhall, A., & Rieke, F. (2009). Timescales of inference in visual adaptation. *Neuron, 61*(5), 750–761.

Watanabe, E., Kitaoka, A., Sakamoto, K., Yasugi, M., & Tanaka, K. (2018). Illusory motion reproduced by deep neural networks trained for prediction. *Frontiers in Psychology, 9*, 345.

Watson, A., & Solomon, J. (1997). A model of visual contrast gain control and pattern masking. *JOSA A, 14*, 2379–2391.

Weintraub, D. J., & Krantz, D. H. (1971). The Poggendorff illusion: Amputations, rotations, and other perturbations. *Attention, Perception, & Psychophysics, 10*(4), 257–264.

Westheimer, G. (2008). Illusions in the spatial sense of the eye: geometrical-optical illusions and the neural representation of space. *Vision Research, 48*(20), 212–2142.

White, M. (1979). A new effect of pattern on perceived lightness. *Perception, 8*(4), 413–416.

Wiener, N. (1949). *Extrapolation, interpolation, and smoothing of stationary time series, Vol. 2.*

Zaidi, D. C. Q., Ennis, R., & Lee, B. (2012). Neural locus of color afterimages. *Current Biology, 22*, 220–224.

Zamir, S. W., Vazquez-Corral, J. & Bertalmio, M. Vision models for wide color gamut imaging in cinema, IEEE Transactions on Pattern Analysis and Machine Intelligence.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. *CVPR*.

Zhang, K., Zuo, W., Chen, Y., Meng, D., & Zhang, L. (2017). Beyond a gaussian denoiser: Residual learning of deep cnn for image denoising. *IEEE Transactions on Image Processing, 26*(7), 3142–3155.