On the synthesis of visual illusions using deep generative models

Alex Gomez-Villa Computer Vision Center, Universitat Autónoma de Barcelona, Barcelona, Spain

Adrián Martín Department of Information and Communications Technologies, Universitat Pompeu Fabra, Barcelona, Spain

Javier Vazquez-Corral Computer Science Department, Universitat Autónoma de Barcelona and Computer Vision Center, Barcelona, Spain

Marcelo Bertalmío Instituto de Óptica, CSIC, Madrid, Spain

Jesús Malo Image Processing Lab, Faculty of Physics, Universitat de Valéncia, Spain

Visual illusions expand our understanding of the visual system by imposing constraints in the models in two different ways: i) visual illusions for humans should induce equivalent illusions in the model, and ii) illusions synthesized from the model should be compelling for human viewers too. These constraints are alternative strategies to find good vision models. Following the first research strategy, recent studies have shown that artificial neural network architectures also have human-like illusory percepts when stimulated with classical hand-crafted stimuli designed to fool humans. In this work we focus on the second (less explored) strategy: we propose a framework to synthesize new visual illusions using the optimization abilities of current automatic differentiation techniques. The proposed framework can be used with classical vision models as well as with more recent artificial neural network architectures. This framework, validated by psychophysical experiments, can be used to study the difference between a vision model and the actual human perception and to optimize the vision model to decrease this difference.

Introduction

The direct study of visual perception is an extremely challenging open problem, and for this reason most psychophysical research is performed on the study of perceptual limits, thresholds, and *errors* that may constrain the models of the system. A visual illusion is a stimulus that induces a visual percept that is not consistent with the physical description of the scene, as given by linear sensors such as spectroradiometers, rulers, protractors, and so on. An example can be seen in Figure 1, which shows a canonical contrast illusion: the gray squares have the exact same luminance (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. Visual illusions can be understood as scenes whose statistics do not correspond to the ones that are typically found in natural images, so we *misinterpret* them because of an (otherwise optimal) codification strategy. In fact, many illusions have been explained as by-products of optimal information transmission or error minimization in statistically unusual scenarios (Barlow, 1990; Laparra & Malo, 2015). Thus, visual illusions allow vision scientists to devise and test new models in their search for a better understanding of the rules that govern visual perception.

Since 2018, a handful of works have observed that artificial neural networks (ANN) trained in natural images can also be "fooled" by visual illusions, in the sense that their response to an input that induces an illusion in humans is (qualitatively) the same as that of humans, and therefore inconsistent with the actual physical values of the stimulus. This has been shown for

Citation: Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Bertalmío, M., & Malo, J. (2022). On the synthesis of visual illusions using deep generative models. *Journal of Vision*, *22*(8):2, 1–18, https://doi.org/10.1167/jov.22.8.2.

https://doi.org/10.1167/jov.22.8.2 Received March 21, 2021; published July 14, 2022 ISSN 1534-7362

Figure 1. A canonical visual illusion (brightness contrast). The squares have the same luminance, but they are perceived as having different brightnesses.

illusions of very different type: motion (Watanabe et al., 2018), brightness and color (Gomez-Villa et al., 2019), and completion (Kim et al., 2019). However, despite the fact that the inspiration for ANNs came from classical models of biological visual neurons (Haykin, 2009) and that ANNs can achieve a remarkable performance on a variety of visual tasks, they fail to emulate basic human perception abilities (Goodfellow et al., 2018; Geirhos et al., 2019; Jacob et al., 2021; Bertalmío et al., 2020; Gomez-Villa et al., 2020; Funke et al., 2021) and this may cause some of their well-known and most relevant problems.

Visual illusions constrain our knowledge of the visual system in two different ways. First, given the stimuli known to elicit illusory percepts in humans, vision models should reproduce the polarity and intensity of the illusions. Second, given different vision models, they could be used to synthesize stimuli that lead to new visual illusions.

The use of visual illusions to study vision models according to the first approach, for instance as in (Gomez-Villa et al., 2020), is inherently limited by the way illusions are created; visual illusions are a really scarce resource, they are handcrafted by psychologists and artists in a one-by-one manner, and they need to be tested using psychophysical studies before they are accepted by the community (Shapiro & Todorovic, 2016). And this approach provides only one-half of the picture: it does not tell us if stimuli designed to trick a vision model can also produce an illusion on human observers, which in terms of illusions would be the other, complementary requisite for a good vision model. Please notice that this is a different question to the well-known problem of adversarial examples in machine learning, where in very small perturbations in the input images lead to the misclasification of those images that does not occur to human observers (Goodfellow et al., 2018).

The motivations of the present work are to address the two previous points. We propose a framework to create novel visual illusions by using generative adversarial networks (GANs) that are optimized to produce illusions with a maximum effect on a given vision model (or ANN). These synthesized stimuli are indeed able to fool human observers, as corroborated by psychophysical tests. The flexibility of the proposed approach allows for the incorporation of different generative models, in particular here we explore the use of GANs in two ways: i) pretrained GANs, where we find a vision illusion by means of optimizing the latent vector, and ii) a GAN that is trained during the process of generation of illusions and, therefore, incorporates the measurement of the illusion effect into the training loss of the GAN. Moreover, the whole framework can be transformed into a GAN itself if the vision model is also trained during the process of generating illusions to not to be tricked by the stimuli produced by the generator. This strategy introduces a new way of optimizing classic vision models, or to train ANNs to better emulate human vision, that should be explored in future work.

In conclusion, the contributions of this work are the following:

- A novel method to generate visual illusions on humans in which the images are synthesized to produce a visual illusion on vision models. This method is validated by psychophysical experiments.
- A new way to study the distance between a given vision model and human perception by means of the visual illusions generated to trick the vision model.
- An extensive collection of synthetic visual illusions produced by different choices of the elements of the framework.

Our goal is not the production of the most compelling illusion, which, of course, would depend on having the perfect human vision model. Instead, we aim at showing that the proposed method works, discussing its implications, and showing how it may be used to improve vision models.

Methods: A framework to generate visual illusions

We pose the problem of synthesizing visual illusions as the problem of optimizing a generative model to create images that maximize a specific visual illusion effect. There is a large variety of known types of visual illusion (Shapiro & Todorovic, 2016); however, for the sake of illustration in this work we focus on a specific color/brightness illusion, one in which the image contains two regions of exactly the same size and radiance (from now on, the targets) that, influenced by their respective surroundings (the inducers), are perceived to be different by the observers. See Figure 1 for an example of a classical brightness illusion of this

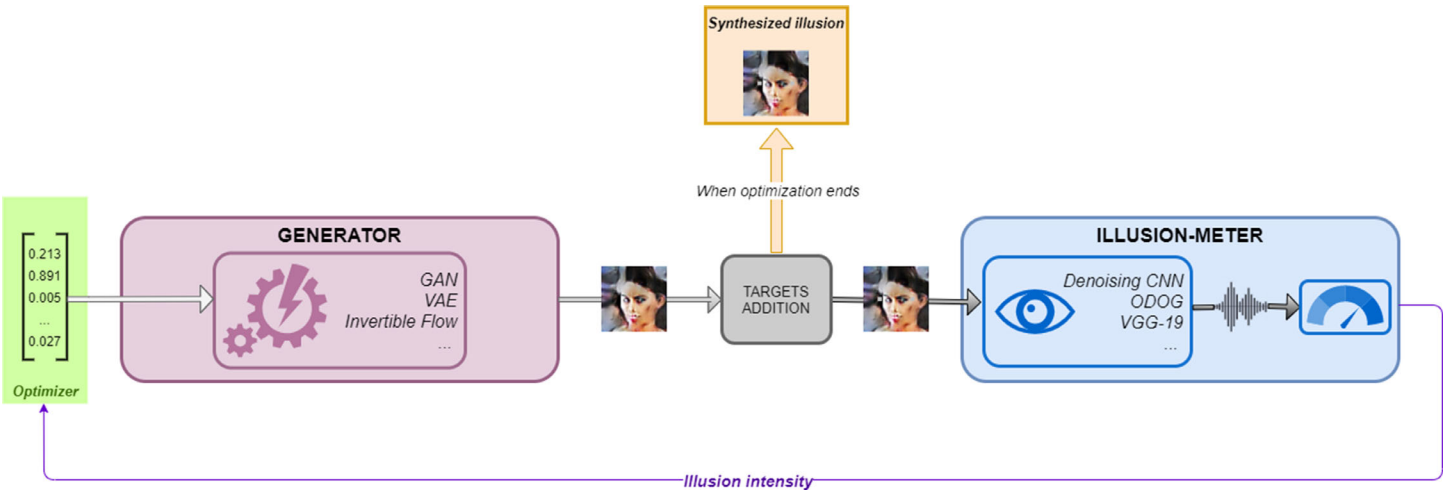

Figure 2. Overview of an instance of the proposed framework in which better illusions are generated by optimizing the latent vector. In this mode, only one visual illusion is synthesized once the optimization is finished.

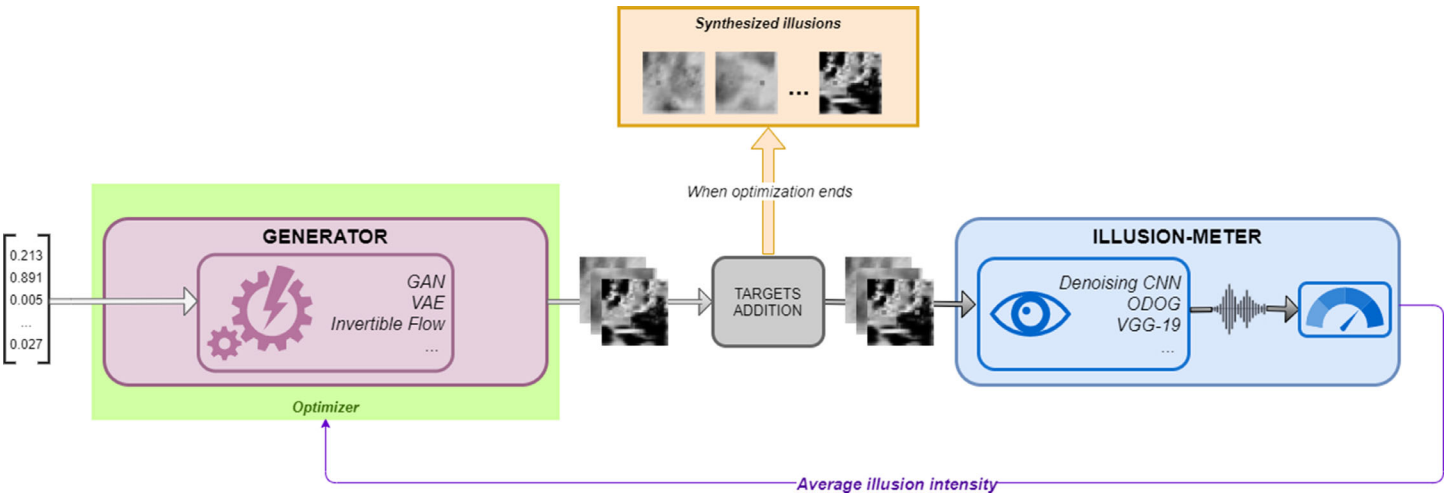

Figure 3. Overview of an instance of the proposed framework in which the generator itself is optimized to produce better visual illusions. In this mode, once that the optimization of the generator is finished, any number of visual illusions can be generated.

sort. In the Discussion, we elaborate on how to use the framework to generate other illusions (e.g., involving textures and contours). In any case, our proposal is based on measuring low-level subjective distances. Therefore, the generation of illusions that rely on high-level scene interpretation is out of the scope of this work.

The framework that we propose is composed of two main elements: one in charge of producing new images (generator), and one that first processes these images and then measures the visual illusion effect (illusion meter). After the images are generated and before they are processed by the illusion meter, there is an intermediate step in which the targets are added to the image. The intensity of the illusion as measured by the illusion meter is then fed to the generator to optimize the images produced, which are the inducers of the illusion, for the sake of increasing the provoked effect. Schemes of the proposed framework are depicted in Figures 2 and 3, in where the difference relies on the use of a fixed generator (Figure 2) or a trainable one (Figure 3), these details are further discussed in the Generator subsection.

The generator can be any generative model able to synthesize new images; in this work we have only used different GANs (Goodfellow et al., 2014), but we could also use variational autoencoders (Blei et al., 2017) or invertible flows (Kobyzev et al., 2020), for instance. Depending on the chosen specific generative model, there are different ways in which the generation of

images can operate, a these details are further discussed in the Generator subsection.

Illusion meter

The illusion meter is the most important element of the proposed framework because it is in charge of measuring the effect that the generated visual illusion produces and therefore needs to approximate the response of human observers to that illusion. It is composed of two distinct parts, the first one is a vision module that processes the image (so acting as a proxy or replicant for human vision) and a second module to measure the effect in the processed signal, as a distance measure. These two modules are intimately related; the choice for the vision module determines the type of processed signal and hence the function used to measure the vision illusion effect.

In what follows, we present the different choices of vision module used in this work (that act as a proxy of humans) and their corresponding meter functions used to determine the intensity of the illusion effect.

Using classical vision models as the vision module

A natural choice for this module is the use of classical vision models (CVM). The parameters of CVMs are carefully tuned to reproduce human responses under very specific conditions. An example of them is ODOG (Blakeslee & McCourt, 1999), a model proposed to reproduce brightness/lightness perception phenomena, which is the one that we use in the experiments.

Whether a CVM reproduces a visual illusion or not is decided in a qualitative way: After processing the visual stimulus with the CVM, the signal in the targets is usually plotted to determine if the visual illusion has taken effect. Because we need to produce a metric, we define as meter function the average of the differences of intensity between the right and the left target. In the ODOG case, given u, an $N_1 \times N_2$ grayscale image processed by the CVM and posed as a $(N_1 N_2) \times 1$ real vector ($u \in \mathbb{R}^{N_1 N_2}$), and given I_L and I_R the set of indexes corresponding to the locations of the left and right target respectively, the meter function is defined as:

$$f_{gray} = \frac{1}{M} \sum_{i \in I_R, j \in I_L} (u(i) - u(j)) \quad (1)$$

for M, the total number of pixels of each target.

Of course, ODOG is just one possible choice for this module. The inducers (surrounds) may be explicitly included in the signal as in the ODOG or in other image-computable models (Otazu et al., 2010; Schütt & Wichmann, 2017; Martinez-Garcia et al., 2018). Alternatively, the surround may be parametrically described as is in CIE color appearance models (Li et al., 2000), where the change in the surround will lead to different perceptions or corresponding color pairs (Capilla et al., 2004; Fairchild, 2013). Finally, the operation regime (or adaptation state) can be nonparametrically determined as in statistical models that try to equalize the set of responses (Twer & MacLeod, 2001; Laparra et al., 2012; Laparra & Malo, 2015). In each of these options, the intensity of the illusion would be computed according to the corresponding way to estimate differences between the responses or color descriptions.

Using convolutional neural networks trained for imaging tasks as the vision module

Following the empirical theory of vision (Kingdom, 2011), recent works (Watanabe et al., 2018; Gomez-Villa et al., 2019) studied how convolutional neural networks (CNNs) trained on databases of natural images to perform low-level processing tasks can reproduce the behavior of human perception for specific visual illusions. Despite not being designed as human vision models, CNNs offer the advantages over CVMs of the continuous development of efficient libraries for fast training through automatic differentiation techniques. Automatic differentiation uses the sequence of elementary arithmetic operations used by a computer program to execute a computer program and apply the chain rule repeatedly to these operations (to automatically compute derivatives of arbitrary order).

We distinguish two main classes of CNNs among the vast number of different types of CNNs trained for imaging tasks. The first class comprises all the CNNs that receive as an input an image and output a processed version of that same image. Examples of CNNs of this class are CNNs trained to perform denoising, deblurring, restoration, inpainting, or color correction, among others. The second class includes those CNNs whose output is not a processed version of the input image, but a signal of a different type, shape or size. This includes CNNs trained for doing image classification, object detection, segmentation, and so on.

In this work, we have selected a CNN that performs image restoration as an example for CNNs of the first class and an image classification CNN in the case of the second class. Each of them requires specific choices on the meter function depending on their type of output signal, which are detailed elsewhere in this article.

Using a CNN trained for image restoration as the vision module

CNNs trained for image restoration are able to reproduce human responses to some classical brightness

visual illusions as shown in (Gomez-Villa et al., 2019). This human-like behavior of CNNs trained for image restoration makes sense because the enhancement of the retinal signal may be one of the goals of the lateral geniculate nucleus (Martinez-Otero et al., 2014). Moreover, other human-like features (e.g., contrast sensitivity) may emerge from this error minimization goal (Gomez-Villa et al., 2020; Li et al., 2022). Analogous to the CVM case, we can measure the effect of the illusion by comparing the left and right targets in the image processed by the CNN. In the case of a grayscale $N_1 \times N_2$ image u we can use the same meter function f_{gray} presented in Equation 1 for the CVM.

In the case of RGB images, we can define different meter functions depending on the channel where we want to measure the difference between targets, or more complex color metrics that combine the different channels. In particular, in this work we have focus only in metrics based on producing differences in a single color for each specific test performed. Let u be the $N_1 \times N_2 \times 3$ image processed by the CNN. We denote as u^r, u^g, and u^b the three vectorized images corresponding with the red, blue, and green channels. Hence u^b, u^g, $u^b \in \mathbb{R}^{N_1 N_2}$ and given I_L and I_R the set of indexes corresponding with the locations of the left and right target, respectively, M the number of pixels of each target, the channel-wise meter functions are defined as:

$$f_{red} = \frac{1}{M} \sum_{i \in I_R, j \in I_L} \left(u^r(i) - u^r(j)\right), \quad (2)$$

$$f_{green} = \frac{1}{M} \sum_{i \in I_R, j \in I_L} \left(u^g(i) - u^g(j)\right), \quad (3)$$

$$f_{blue} = \frac{1}{M} \sum_{i \in I_R, j \in I_L} \left(u^b(i) - u^b(j)\right). \quad (4)$$

Additionally, we define a meter function to produce perceptual differences in the yellow color, as combination of the red and green channels:

$$f_{yellow} = \frac{1}{M} \sum_{i \in I_R, j \in I_L} \left(u^r(i) - u^r(j)\right) + \left(u^g(i) - u^g(j)\right). \quad (5)$$

Using a CNN trained for image classification as the vision module

Using CNNs trained for image classification poses the question of how to measure the illusion effect when the signal processed is not an image anymore but a vector of characteristics. Instead of measuring the effect in the final output of the CNN, we work with its internal representations (Bengio, 2009), whose perceptual properties have already been shown (Gatys et al., 2015; Zhang et al., 2018).

Nevertheless, how and where to measure these internal representations is not a trivial choice. In this work, we adapt the style loss proposed by (Gatys et al., 2015) in the context of style transfer that has been used later in several perceptual based approaches in CNNs (Gatys et al., 2017). Defining Gram matrices as in (Gatys et al., 2015), we denote as G^l_{left} and G^l_{right} the Gram matrices corresponding with the locations of the left and the right targets, respectively, in the l^{th} layer. The meter function is then defined as:

$$f_{style} = \sum_{l=0}^{L-1} w_l \| G^l_{left} - G^l_{right} \|^2, \quad (6)$$

where $\| \cdot \|$ is the Frobenius norm, L is the number of layers used, and w_l are weighting factors of the contribution of each layer to the meter function.

Generator

The generator in our framework is a generative model able to create new images where after the two equal targets are superimposed they are perceived to be different. For this purpose we have used different instances of GANs (Goodfellow et al., 2014), which in their most standard setting are composed of two competing CNNs, namely, the generative and discriminative network. This is not the only possible choice, any other generative model such as variational autoencoders (Blei et al., 2017) or invertible flows (Kobyzev et al., 2020), could be used for instance. Using GANs allows us to synthesize images in two operating modes: a first one in which the latent space of the trained GAN is explored to find the image that maximizes the illusion effect, and a second one in which a pretrained GAN is trained itself during an optimization process to learn to produce better illusions. In the following we detail the particulars of these two approaches.

Fixed generative model: Optimization of the latent vector

In this operation mode, our goal is to find a latent vector z (see Figure 2) that generates an image in the generator space that maximizes the illusion effect as measured by the illusion meter. The GAN has to be trained before outside the framework in some determined dataset, thus, allowing us to use any pretrained GAN in the literature.

To optimize the vector z, a gradient-based optimizer is run on the illusion meter output till convergence (green block in Figure 2). The final output in this approach is one visual illusion produced by the image generated by the found latent vector z (orange block Figure 2). This implies that, for every new visual illusion, we have to run the optimization process from a new starting point.

Trainable generative model: Learning to multi synthesize multiple visual illusions

Instead of working with a fixed GAN, a second possibility we explore is to start from a pretrained GAN that is optimized to produce images that provoke stronger visual illusions. This operating mode is depicted in Figure 3, where the green block indicates that it is the whole generator the one being optimized.The loss of the GAN (ℓ_{GAN}) is modified to incorporate the output from the illusion meter (ℓ_{IM}), a scalar value that is the average intensity of the illusion produced by the batch of images generated by the GAN, resulting on a final loss function (ℓ_{ffinal}) that is defined as

$$\ell_{\text{FINAL}} = \alpha \ell_{\text{GAN}} + (1 - \alpha)\, \ell_{\text{IM}}, \quad (7)$$

where α is an hyperparameter to balance the influence of the two terms. Once the optimization of the whole framework converges, the output from this approach are in fact as many visual illusions as required, because the generator has learned how to generate images that produce visual illusions in the two targets that are superposed in them.

Experiments

In this section we detail the different experiments we have defined, to address the ability of our framework to synthesize new visual illusions. Experiments 1 and 2 explore the two different approaches presented to generate visual illusions. The first experiment involves the case of finding visual illusions via optimization of the latent vector from a fixed generator. The second experiment focuses on training the generator to produce better visual illusions. Finally, we chose this second approach to perform a psychophysical experiment to study whether the proposed framework is able to produce visual illusions that fool human observers. Moreover, in this experiment we compare the choice of a classical vision model or a restoration CNN as the vision modules of the illusion meter. The different image datasets used in the experiments are described in Table 1.

Experiment 1: Optimizing the latent vector in a fixed generative model

In this experiment, we synthesize illusions by optimizing the latent vector following the scheme shown in Figure 2. Our choice for the generator is a DCGAN (deep convolutional generative adversarial network) (Radford et al., 2015) trained in the Celeb-faces dataset (Experiment 1.a) or the Cats faces 64 × 64 dataset (Experiment 1.b) (one DCGAN for each dataset, see Table 1 for their details). The DCGAN architecture is composed of two different CNNs modules, namely, the generative and the discriminative ones. The generative network has four modules of transposed convolution layers, batch normalization and ReLu activation with a final transposed convolution before the output *tanh* activation (with an output size of 64 × 64 pixels). The discriminative network has five modules of convolutional layers, batch normalization and Leaky ReLu activation, except for the first and last module that do not have a batch normalization layer and in which the activation of the last layer is a sigmoid. We trained the DCGAN for 80 epochs using an Adam optimizer with learning rate of 0.0002. The chosen vision module is a VGG16 network (Simonyan & Zisserman, 2014) trained for classification on Imagenet (Deng et al., 2009). The first four layers of this network are fed to the meter function that calculates the style loss (Gatys et al., 2015) in the target areas (see Equation 6). Then, the latent vector is optimized using the Adam optimizer with a learning rate of 0.0002 for 100 epochs (see Figure 2 for a scheme of this operating mode).

Experiment 2: Optimizing the generative model

In this setup, as shown in Figure 3, we optimize the generator module itself to synthesize images

Dataset	# of images	Size	Type of images
Celeb-faces (Liu et al., 2015)	200K	178 × 218	Human faces
Cat-faces cat (0000)	15K	64 × 64	Cat faces
Places 2 (Krizhevsky, 2009)	10M	128 × 128	434 categories for scene recognition
DTD (Cimpoi et al., 2014)	5640	300 × 640	Describable textures

Table 1. Summary of the datasets used in the experiments.

that produce visual illusions. This generator is a DCGAN trained in two different datasets, the Airfield category of the Places 2 dataset (Experiment 2.a) and DTD dataset (Experiment 2.b) (see Table 1 for their details). In Experiment 2.b, the generative network is composed of two fully connected layers followed by two convolutional layers. The fully connected layers have 2048 and 256 $\times$ 8 $\times$ 8 hidden nodes respectively. Both convolutional layers use 5 $\times$ 5 filter size with 128 and 1 channels, respectively. Before each convolutional layer the input is 2 $\times$ 2 upscaled. ReLu activation functions are used after each layer but the output one in which a sigmoid activation function is applied. The discriminator network is composed of three convolutional layers and two fully connected layers. The first convolutional layer uses 5 $\times$ 5 filter size and the other two use 3 $\times$ 3 filter size with 128, 256, and 512 channels, respectively. The two fully connected layers have 1,024 and 1 hidden nodes. After every layer the activation function is a Leaky ReLu ($\alpha = 0.2$) except for the output layer that uses a sigmoid function. A max pooling operation is applied after each convolutional layer. This DCGAN from Experiment 2.b was modified to perform experiments in higher resolutions in Experiment 2.a (128 $\times$ 128). The generative CNN now has additional convolutional layers (with same dimensions as before) and 2 $\times$ 2 upscaling layers while a convolution and max-pooling layers are added to the discriminative CNN. For the vision module in Experiment 2.a we use the deep denoising CNN proposed by Zhang et al. (2017) for brightness and color illusions while in Experiment 2b we use RestoreNet, a shallow image restoration CNN used in (Gomez-Villa et al., 2020), that were shown to accurate reproduce human response in several brightness illusions. We optimize the whole framework using an ADAM optimizer on the DCGAN loss modified to include the output from the chosen meter function. In the case of grayscale illusions the meter function used is the one defined in Equation (1), while for color illusions we choose between Equations (2), (4), or (5) depending on the color in which we want to perceive the differences (respectively Equation red, blue, and yellow).

Experiment 3: Psychophysical experiment

The last experiment consists in performing psychophysical tests to assess the ability of the proposed framework to synthesize lightness visual illusions that actually fool human observers. Additionally, we compare in this experiment two different choices in the vision module of the illusion meter, ODOG (Blakeslee & McCourt, 1999), a classical vision model for brightness perception, and the above explained CNN RestoreNet (Gomez-Villa et al., 2020). These two choices for the vision module are used with the same meter function defined in Equation (1), hence allowing us to compare both approaches fairly.

Observers were shown the lightness visual illusions created by our approach—using squares as targets—and they were asked to select the lighter square, having three different options to choose from left, right, or center (in case they were not able to perceive any difference). The experiment was performed on a calibrated AOC I2781FH LCD monitor. Observers were sitting at 50 cm of the screen so that the target grey squares subtended 1.5° of visual angle. The trials were untimed. Ten observers took part in the experiment. All of them presented normal color vision and emmetropic (or corrected to emmetropic) vision. None of them was an author on this paper. Each of the observers viewed each of the illusion once. We have considered the DTD dataset (see Equation 1) for training the generator module. For each choice of the vision module we have obtained 50 output images (totalling 100 images) by randomly selecting images that were considered to be an illusion by the perception quantifier module from batches of different iterations during the framework training, to ensure the diversity of the candidates generated. These images were randomized both in terms of the methods and in terms of the side in which the lighter square was expected to appear.

Results

In this section, we present the results of the different experiments laid out in the previous section. We follow the same structure as above, starting with the images produced in Experiments 1 and 2, and finally introducing the results of the psychophysical experiment. A summary of the experiments performed is displayed in Table 2.

Results for Experiment 1: Optimizing the latent vector in a fixed generative model

Figure 4 shows the results of our approach when we either train a DCGAN generator with human (top row) or (bottom row) face images. The first column shows a nonoptimized image (a sample from the generator), from columns two to five an optimized version from different seeds are presented. For each image, we have also included the value of the loss function. This value indicates how different are the gray targets according to the illusion meter (bigger numbers stands for greater differences). Let us note that, in this case, as we are using the style losses, we are not allowed to specify a polarity in the visual illusion (for instance, that the left patch is perceived brighter than the right one). In

Label	Figure	Generator	Dataset	Vision module	Meter function
Experiment 1a	Figure 4, first row	Optimization of the latent vector	Celeb	VGG16	Equation 6
Experiment 1b	Figure 4, second row	Optimization of the latent vector	Cat-faces 64x64	VGG16	Equation 6
Experiment 2a	Figure 5, first row	Optimization of a DCGAN	Places2: Airplanes	Denoising CNN	Equations (1), (2), (3), (5)
Experiment 2b	Figure 5, second row	Optimization of a DCGAN	DTD	RestoreNet	Equations (1), (2), (3), (5)
Experiment 3	Figure 6, left column	Optimization of a DCGAN	DTD	ODOG	Equation (1)
Experiment 3	Figure 6, right column	Optimization of a DCGAN	DTD	RestoreNet	Equation (1)

Table 2. Table summary of the experiments performed. The vision modules used are ODOG (Blakeslee & McCourt, 1999), RestoreNet (Gomez-Villa et al., 2020), VGG16 (Simonyan & Zisserman, 2014), and the denoising CNN from Zhang et al. (2017).

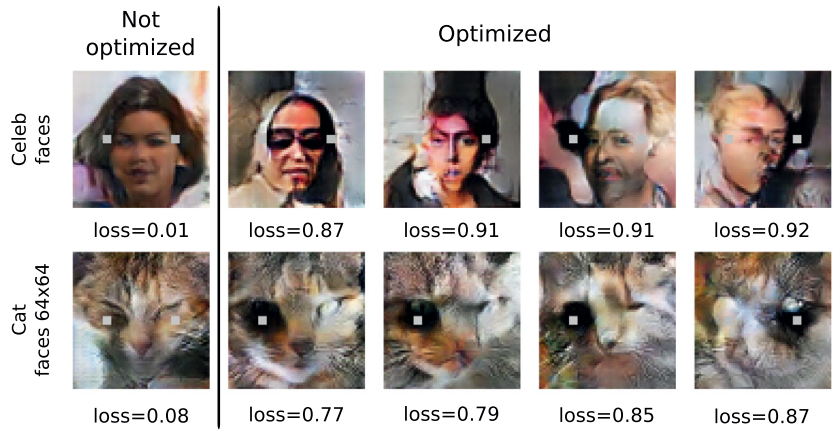

Figure 4. Experimenta 1a and 1b. Example images for visual illusions obtained at size 64x64 using the latent vector optimization approach. The first row is trained with the Celeb-faces dataset (human faces) and the second one is trained with the Cats faces 64 × 64 dataset.

this case, our condition is that a different perception should be perceived in each target (so maximizing Equation (6)). This difference in perception between the patches is clearly present in all the examples of the figure (columns two to five).

By looking at the previous figure, we can see that the generator has the clear tendency to surround one target with a dark region and the other one with a lighter region (therefore tending toward the brightness contrast illusion). This said, as we are constraining the generator to obtain images resembling either human or cat faces, those surrounding regions will be objects that do exist in the faces (e.g., hair in human faces or a black hair region in cat faces).

Results for Experiment 2: Optimizing the generative model

Results for Experiments 2.a and 2.b are shown on the first and second rows, respectively, of Figure 5. In the first row, we synthesize images of higher resolution (128 × 128) considering the airplane category of the Places dataset. The second row show images generated from the DTD dataset (32 × 32). The first column presents an intensity illusion and from the second to the fourth columns are color contrast illusions obtained by starting either in a red, blue, or yellow target color. We can see that, in all the cases, we are able to synthesize visual illusions, showing that our framework is not tied to any particular region of the color space. Going into greater detail, in the second column we specified in the meter function than the right target should be perceived redder than the left one (maximizing Equation (2)). Similarly, in the third column, we specified that the right target should be perceived bluer (maximizing Equation (4)), and in the fourth column that the right target should be perceived yellower (maximizing Equation (5)).

Results for Experiment 3: Psychophysical experiment

Six example images that were shown to the observers are shown in Figure 6; the three first images were produced using ODOG as the vision module and the last three images were synthesized with RestoreNet as the vision module. A first interpretation of the results is presented in Table 3, where the average selection of

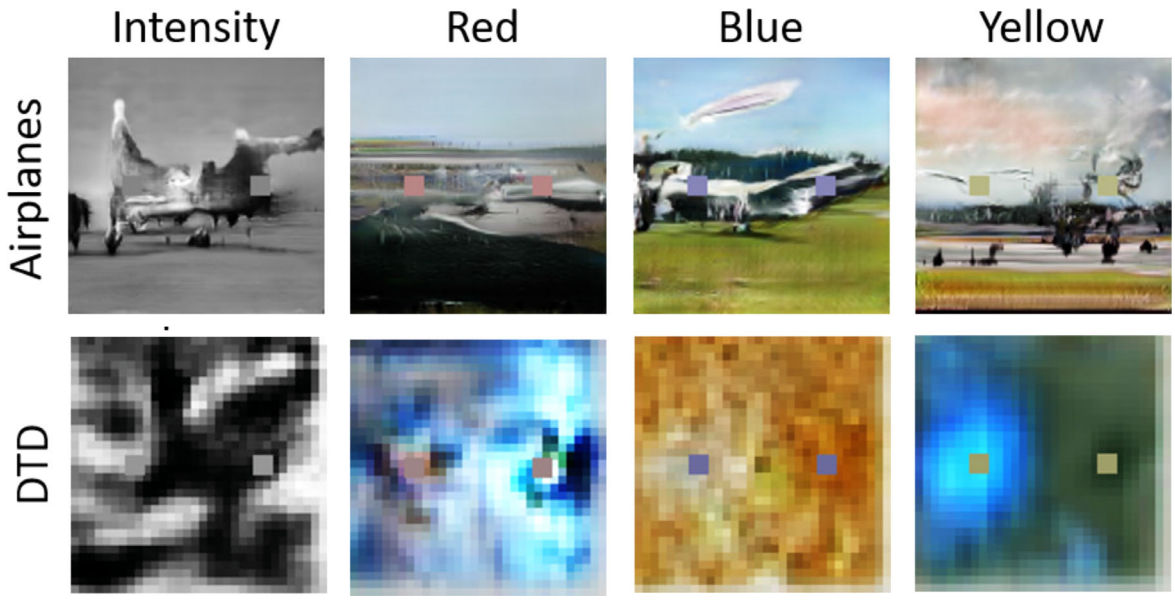

Figure 5. Experiments 2a and 2b. Results for lightness illusions using different training datasets and target shapes, and considering RestoreNet as the vision model. In all the cases the target on the right was selected to be brighter than the one on the left. Results for the lightness and color illusions using Airplanes (first row) and DTD (second row) as the training datasets for the background images and different target colors. From left to right, our cost imposed the perception of the right target to be lighter, redder, bluer, or yellower.

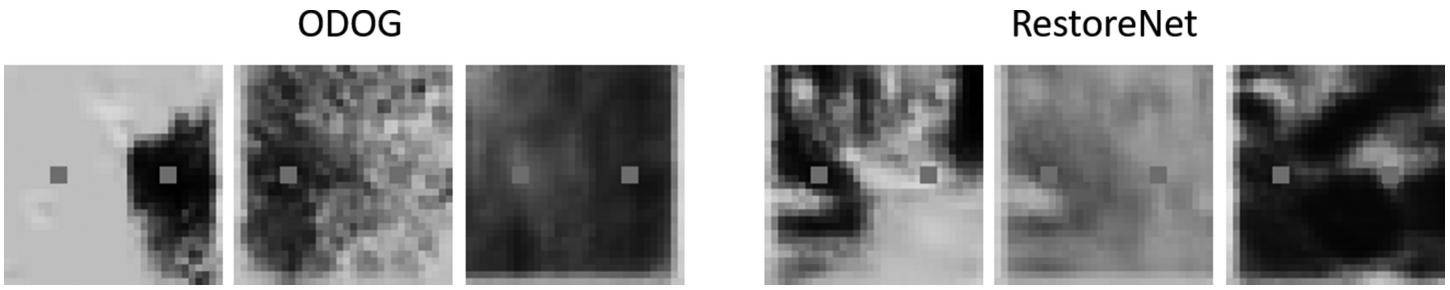

Figure 6. Experiment 3. Examples of lightness illusions generated using a classical visual model (ODOG) (left) and an image to image restoration CNN (RestoreNet) (right), with square targets and the DTD dataset. The square selected to be the one that should be perceived brighter was randomized in these images.

	Opposite illusion	No illusion	Correct illusion
All	0.08	0.23	**0.70**
ODOG	0.07	0.22	**0.71**
RestoreNet	0.08	0.24	**0.68**

Table 3. Results of the psychophysical experiment (Experiment 3) as average of the selected square.

each option for the full set, as well as for each vision module, is shown. The average is performed on the 10 different observers of the experiment. As we can see, in a large majority of the cases (approximately 70%) the human observers perceive the illusion that was generated by the vision module. To obtain a more statistically significant result, we have also recast the experiment in terms of the Thurstone Case V Law of Comparative Judgment (Thurstone, 1927). To this end, we have partitioned the answers into two classes, as having the generated illusion or not having it. The latter category considers both the cases where an observer has not seen any illusion and where an observer has selected the opposite direction for the illusion. Results for this analysis (Experiment 3) are shown in Figures 9 and 3. As we can see, the generated illusion is perceived with statistical significance in all the cases.

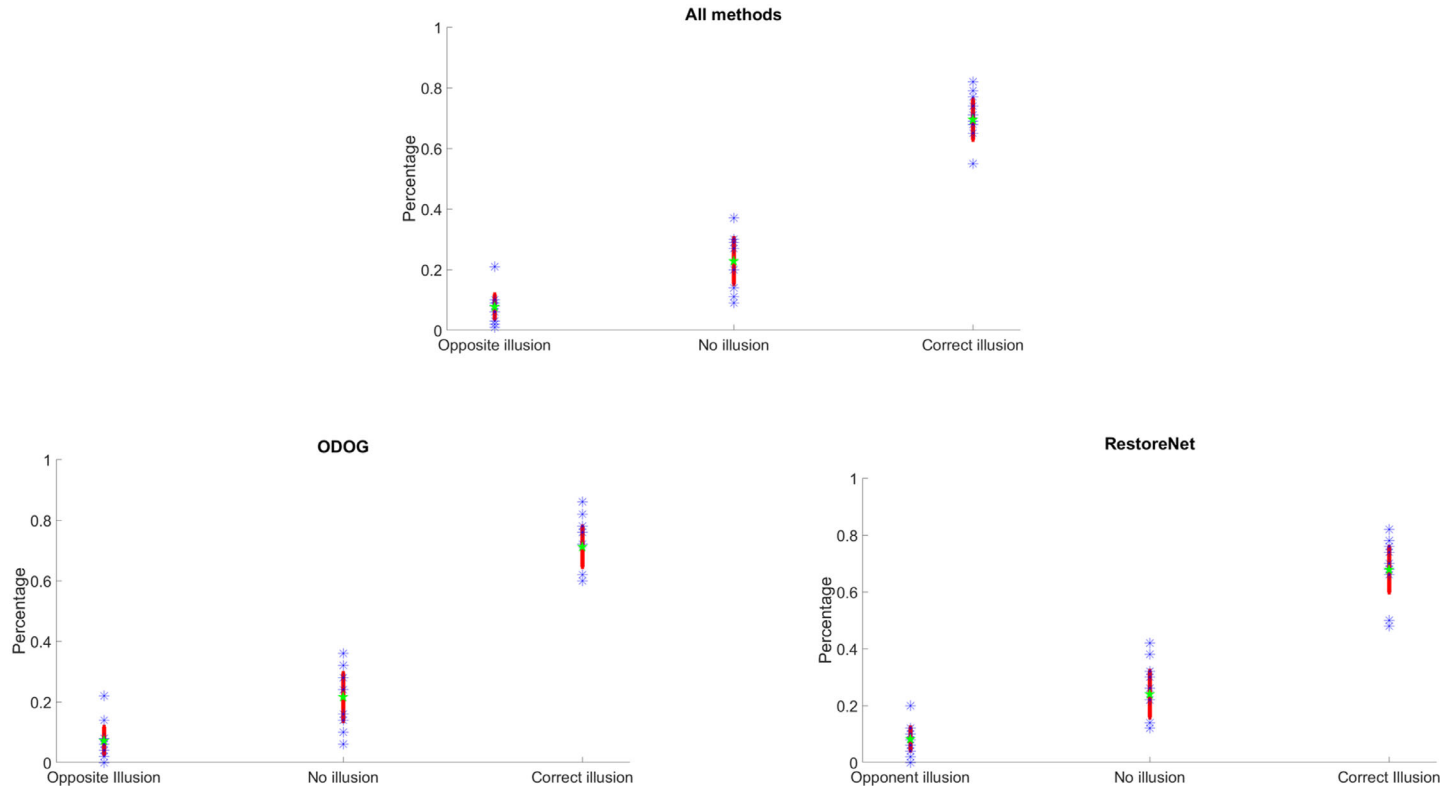

Figure 7. Statistical analysis for the results of the psychophysical experiment (Table 3).

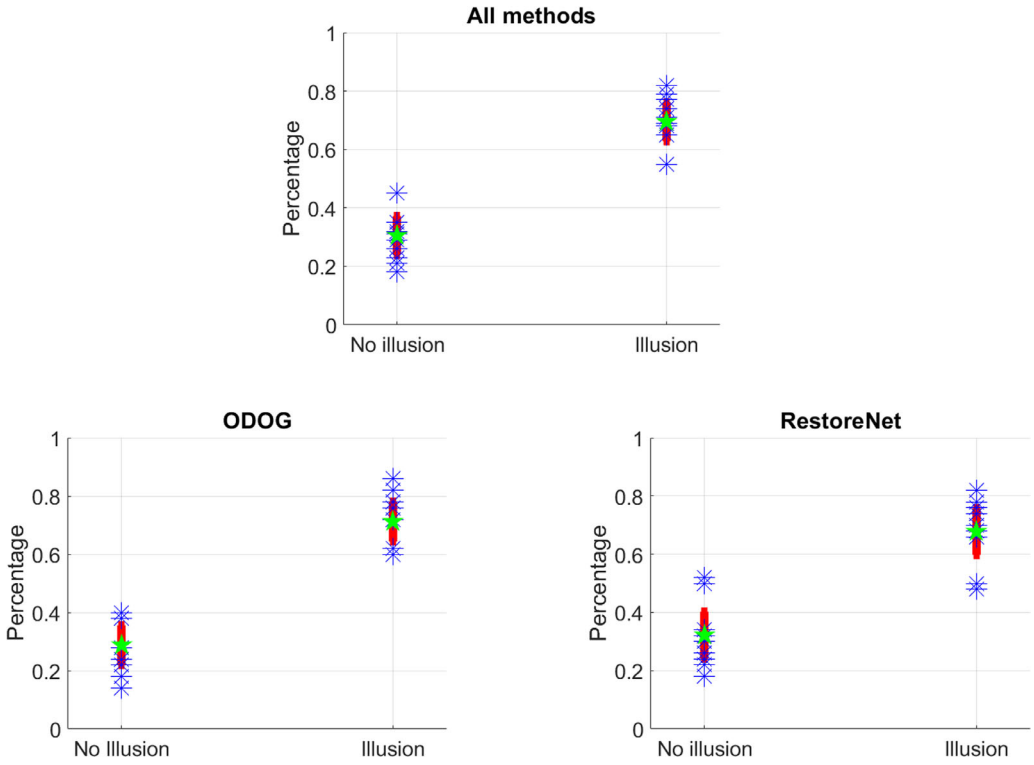

Figure 8. Statistical analysis for the results of the psychophysical experiment when results are partitioned in just two classes either obtaining or not obtaining the illusion.

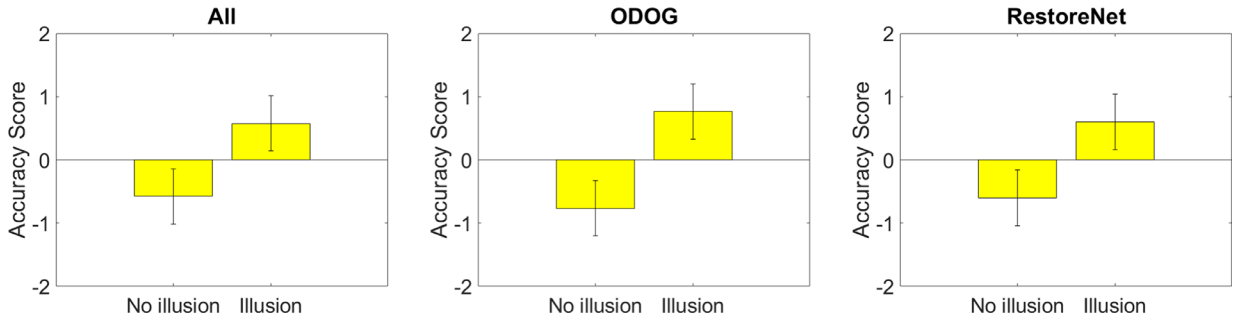

Figure 9. Thurstone Case V Results for Experiment 3. We can see that the illusions are seen with statistical significance in all the cases.

Results for Experiment 3: Psychophysical Experiment

Six example images that were shown to the observers are shown in Figure 6, the three first images were produced using ODOG as the vision module while the last three images were synthesized with RestoreNet as the vision module. A first interpretation of the results is presented in Table 3, where the average selection of each option for the full set, as well as for each vision module, is shown. As we can see, in a large majority of the cases (approximately 70%), the human observers perceive the illusion that was generated by the vision module. The results of this table can be further interpreted in Figure 7, where we plot the individual results for the 10 observers (blue crosses), as well as the average (green star), and the 95% confidence intervals (orange lines). The confidence intervals have been computed in terms of the Student *t's* distribution approximation. In the top Figure, we can see the results for where both methods are considering, and in the bottom part, we present the results for both the ODOG and RestoreNet cases. Looking at the figures, we can clearly see that there is a statistical significance in our results, and that, when comparing the figures, for ODOG and RestoreNET, the former seems to surpass the latter.

To have a more detailed result looking at whether our methods obtain/does not obtain the illusion, we have also partitioned the answers into two classes, as having the generated illusion or not having it. The latter category considers both the cases where an observer has not seen any illusion and where an observer has selected the opposite direction for the illusion. Figure 8 presents the same structure from the previous Figure. We can clearly see that our results are statistically significant, and that for almost all observers in all of the cases, our method obtains the desired effect. We have also recasted this partitioned case in terms of the Thurstone Case V Law of Comparative Judgement (Thurstone, 1927), because it is a standard approach to look at pairwise comparison experiments, and in this case, our experiments is one. Results for this analysis (Experiment 3) are shown in Figure 9 and Table 3. As we can see, this analysis also confirms that the generated illusion is perceived with statistical significance in all the cases; the error bars display 95% confidence intervals.

We further analyzed whether our results can be explained by differences in the contrast level surrounding our patches. For that, we removed those images in which the contrast of one of the channels with respect to its surrounds was smaller than 2% or 5%, and we computed at two different scales, a 20×20 and a 30×30 window surrounding our patch. For all the analyzed cases, the results presented the same trend detailed elsewhere in this article.

Discussion

Here we discuss the critical elements of the framework, the opportunities for theory-driven experiments to improve vision models, and for a better understanding of the role of the statistics of the environment in visual illusions.

Limitations of the elements of the framework

Generation strategy

In Experiments 1 and 2, we have shown results obtained using the two different generating modes of the proposed framework. When we optimize a latent vector as generation strategy, we have the advantages of a fixed network (generator): there is no need to adjust, optimize or tune any parameter in the approach, just the components of z and there is no risk to fall in common optimization issues of generative models such as mode collapse (when the generator starts producing the same output or small set of outputs and the discriminator learns to reject that output). However, we face drawbacks such as the need for the generator to be trained to learn well enough the natural image manifold. This setting is restricted to the type of images from the training dataset; hence, it can not be modified in order to properly produce targets inside the image.

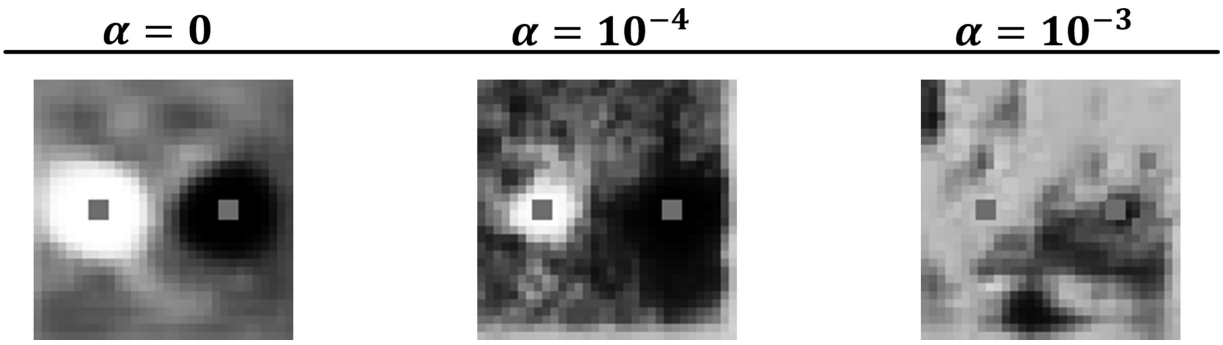

Figure 10. Effect of the influence of the hyperparameter α in the conditions of Experiment 3.

Finally, only a single visual illusion is produced for each optimization process we run.

In contrast, when we train the generator, after the optimization it will be capable of generating multiple visual illusions just sampling the created manifold. In this setting, it will be possible to create the images with targets if we impose a condition over the generator, for instance following an approach similar to Pix2Pix (Isola et al., 2017). The limitations of this approach include the usual generative models optimization problems, such as mode collapse and the tune procedure of the weights of each loss; that is, there is a fine tuning for balancing the value we give to the influence of the strength of the illusion and the generation loss that forces to produce a natural image as background (the hyperparameter α in Equation (7)). As observed in our experiments, without this fine tuning, the visual illusions generated by this approach tend to images resembling a *canonical* visual illusion such as the one showed in Figure 1. In Figure 10, we show an example of how the value of α influences the appearance of the images generated. All the other choices in the framework correspond to the ones used in Experiment 3 with RestoreNet as vision module (see Table 2). When $\alpha = 0$, the optimization of the generator is only driven by the generation of visual illusions, producing an images that does not resemble the images in the dataset. When the value of α increases, the quality of the produced images improves until reaching a point in where the effect of the illusion is not perceived anymore. We have found that $\alpha = 10^{-3}$ is a good compromise between these two extreme cases.

Finally, in terms of computational cost, this second approach requires a bigger effort in terms of time and resources due to the need to train the generator of images to produce better visual illusions. Optimizing just the latent vector is much faster and allows to use complex models that have been already trained.

Measuring the illusion

When we are synthesizing visual illusions, there is a fundamental issue that is not solved in vision science, which is how to quantitatively measure whether an image indeed is producing a visual illusions. We have posed this problem in two stages in the illusion meter: first, we simulate human perception (hence, also the distortion of the reality) with the vision module, secondly, we define meter functions that should objectively measure the processed signal (true tristimulus values). How to compute these measurements remains an issue as shown in Figure 11, where we can see three different examples of image profiles over a line crossing the targets that have been processed with different choices of the vision module (Figure 11b) and the profile of the original image (Figure 11a). In this classical visual illusion, human observers perceive the square over the black background to be lighter than the square over the white background. Where should we quantify the perceived brightness is an important question to consider. One approach could be to measure the value of the central pixel of each square, but this will not work for the outputs presented in Figures 11c and 11d, that, however, are correctly reproducing the illusion as observed in the profiles. Another possibility would be to compare the average values in the target regions, although this approach will not necessarily work again for these two models. This choice is, therefore, an important decision intimately connected with our vision module choice.

If standard distances are used in the illusion meter, the proposed technique can be seen just as a way to obtain different versions of preexisting illusions. This is a limitation imposed by the illusion meter (e.g., restricting the kind of generated illusion to the feature taken into account by the measure). However, the automatic generation of different versions of an illusion is still interesting for two reasons: 1) it gives the researcher the opportunity to check the influence of diverse parameters through the cost function, and 2) looking for the best inductors with spatial structure remains an open issue (masking by surround texture) and could be explored with this framework. See the discussion on using the framework with more general illusions elsewhere in this article.

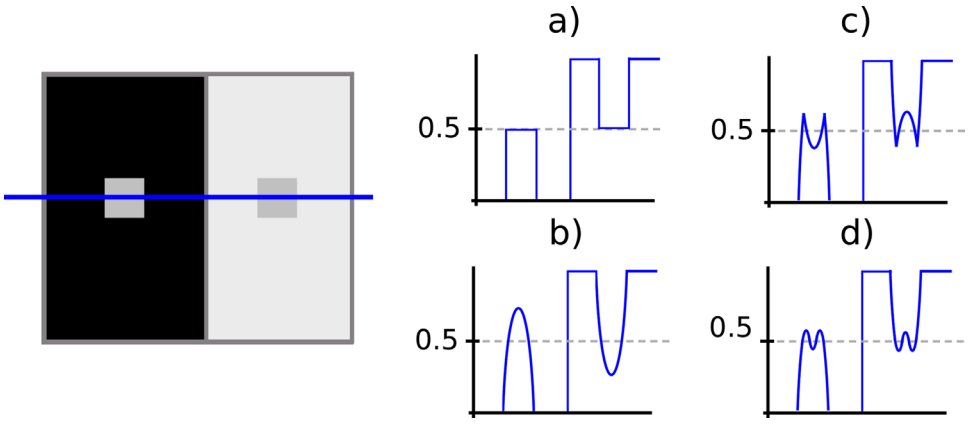

Figure 11. A classical visual illusion (brightness contrast). The squares have the same gray value, but they are perceived as being different. In (a) is depicted a plot of the values of the image over the blue line. Subfigures (b), (c) and (d) show sample output images from different vision models or same vision model using different parameters. This figure was created for visualization purposes (it was not explicitly generated from a vision model); nevertheless, any brightness vision model such as Blakeslee and McCourt (1999) present these kind of plots.

Synthesis of illusions as a tool to decide between models

The proposed framework for illusion synthesis could also be seen as a contribution to the literature on theory-driven methods to decide between competing vision models.

In visual psychophysics, theory-driven experiments imply i) the synthesis of stimuli according to the models under consideration, and ii) the psychophysical judgements of the stimuli done by human observers. The judgments by humans determine which model better describes visual function. In this context, the eventual complications of the synthesis are just a computational issue: the expensive part is the psychophysical test. Therefore, the (cheaper) computational effort in stimulus design is key to minimize the (more expensive) validation through measurements. These considerations led to a series of techniques for theory-driven stimulus generation (Wang & Simoncelli, 2008; DiMattina & Zhang, 2013; DiMattina, 2016). For instance, (Wang & Simoncelli, 2008) proposed that one can rule out certain vision model A (or distance A) in front of a reference model B (or reference distance B) by generating maximally distant images according to model A while having a fixed distance for model B. The observers decide which of the model-generated pairs is actually more different, thus indicating which distance A or B is more correct. This maximum differentiation procedure has been used to perceptually decide the parameters of divisive normalization models (Malo & Simoncelli, 2015). The iterative procedure to generate the stimuli in the original maximum differentiation has been simplified by second-order approximations of the distance (Martinez-Garcia et al., 2018), or by selecting pairs of stimuli from preexisting databases instead of being explicitly synthesized (Ma et al., 2020).

In this stimuli–synthesis context, the current deep learning artificial neural networks to model biological vision (Kriegeskorte, 2015; Yamins & DiCarlo, 2016; Majaj & Pelli, 2018; Kietzmann et al., 2019) has brought an interest in generating stimuli to falsify these architectures (Berardino et al., 2017; Geirhos et al., 2019; Martinez et al., 2019; Fruend, 2020). But, more important, from the experimental point of view, deep generative models excel in the synthesis of realistic natural images (Goodfellow et al., 2014): they implicitly capture the properties of the manifold of natural stimuli and one can control the properties of the manifold by restricting the images in the training set. Therefore, deep generative models seem the appropriate candidate to sample stimuli from. Recent research uses the stimulus optimization ability of deep learning to synthesize tests that allow the discrimination between competing models. For instance, given two models A and B (Golan et al., 2020) generate controversial stimuli in the sense that they belong to different classes for the classification model A, while being the same class for model B. Similar to maximum differentiation, the observers decide which of the models leads to (perceptually) more meaningful stimuli/class distribution. On the same vein Fruend (2020) uses a two-stage (synthesis + experimental decision) process. First, he uses GANs to embed targets into natural scenes and the models that fit the visibility of such targets are used to synthesize new model-specific stimuli. Then, observers decide between the models assessing between the model-specific stimuli.

In the case of the proposed framework for illusion synthesis, given two competing vision models (or subjective distances), our method will synthesize visual illusions by generating backgrounds that maximize the subjective distance between physically identical tests. In this context, the best model will be the one leading to the more compelling illusion, a decision that can be reduced to simple 2 alternative forced choice experiments. As a consequence, the psychophysical results reported above to validate the procedure (Experiment 3) can also be seen as a way to test which model explains better human behavior. The criterion is how often the human observers are fooled by the illusions generated by each model. In this specific example, the difference between the considered models (the classical vision model and the CNN) is small. However, note that the proposed framework is generic and it can be applied to any pair of models as the maximum differentiation method cited elsewhere in this article.

More generally, psychophysical decisions on illusions similar to Experiment 3 could be used to fit the free parameters of a vision model. In this case, the best parameter is the one that leads to higher alignment between the perceptual distance (illusion strength) and human opinion. Nevertheless, to make a more conclusive comparison between two models (say A and B), one should impose the selection of surrounds that maximize the distance for model A while keeping it constant for model B, as it is done in maximum differentiation (Wang & Simoncelli, 2008; Malo & Simoncelli, 2015; Ma et al., 2020) and controversial stimuli (Golan et al., 2020). We did not do this in our proof of concept (Experiment 3), and hence the use of the proposed framework to this end is left as a matter for future research.

Synthesis of illusions in different statistical environments

Another advantage of the proposed framework is that it allows systematic restrictions of the dataset that determines the surrounds. This can clarify the relation between the statistics of the environment and the strength/direction of the visual illusions. Note that mismatch between the statistics of the environment and the target has been largely considered as the origin of the illusions (Barlow, 1990; Laparra & Malo, 2015). Recently, the ideas of Purves et al. (2014) on the relative scales of brightness learned from the experience in certain environments have been used in practice to synthesize illusions via normalizing flows (Hirsch & Tal, 2020). These approaches may benefit from the proposed method by exploring the interaction between the class of illusions that can be generated and the restrictions applied to the manifold of admissible backgrounds.

Training the vision module within the framework

Although not explored in this work, the proposed framework could be modified to use the output from the meter function to optimize the vision module during the main process of generation of visual illusions. In the case of a classical vision model this opens the possibility of an automatic adjustment of its parameters, that are typically manually adjusted. When using CNNs, this training could be seen as a way to impose human perceptual properties in these architectures with the aim of solving some of the errors that could be originated by their differences with respect to humans. This topic very compelling and should be explored deeply in future work.

Using the framework with more general illusions

The consideration of spatial information in the illusion meter would lead to effects beyond brightness and color induction. For instance, it is known the perceived frequency, orientation and contrast of a spatial pattern depends on the texture of the surround and the background (Blakemore et al., 1970; Tolhurst & Thompson, 1975; Foley, 1994; Ellemberg et al., 1998). These effects are described by the interactions between the responses of wavelet-like filter banks (Watson & Solomon, 1997; Cavanaugh et al. 2002a, b). Models of these interactions have been used to generate maximally visible distortions in certain backgrounds (Martinez-Garcia et al., 2018). If these models are used in the illusion meter, an optimization procedure similar to maximum differentiation may be used to look for the background/surround that induces the greatest changes in spatial the perception of the targets.

Similar models based on oriented filters at different scales have been shown to explain contour illusions such as Kanizsa-type subjective figures, phase-induced subjective contours, the Zöllner illusion, and the Müller–Lyer illusion (Rodriguez-Sanchez et al. 1999, 2000), so it is easy to include them in the proposed framework to assess the strength of the contours when the targets are superimposed to the background.

Finally, artificial neural models have been shown to perceive illusory contours (Pang et al., 2021; Kim et al., 2021), so they could also be used and even trained in the illusion meter module of the framework.

Conclusions

We have proposed a framework for the synthesis of visual illusions through deep generative models

and the maximization of the perceptual difference between the targets embedded in the background. Psychophysical experiments validate the correctness of the framework by showing that the stimuli synthesized assuming different vision models do induce illusions in human viewers. This means that the proposed synthesis method, based on distance maximization, works. Of course, the intensity of the illusions depends on the correctness/generality of the considered models, but the optimization of the models to get more compelling illusions is a matter for future research. Here we explored specific options for the basic building blocks of the framework (the background generator and the illusion meter), and in every case we showed illustrative results for these different options. A systematic analysis of these different options is also a matter for future research.

Finally, future work should address extension of this work to other types of illusions. For example, we can consider the works of Watanabe et al. (2018); Kim et al. (2019), where they studied how CNNs replicate motion or completion illusions respectively. Both of these CNNs seem to be good candidates for vision modules.

Keywords: visual illusions, synthesis of stimuli, visual response models, image distortion metrics, deep generative models, generative adversarial networks

Acknowledgments

Commercial relationships: none.
Corresponding author: Alex Gomez-Villa.
Email: agomezvi@cvc.uab.es.
Address: Edifici O, Campus UAB, 08193 Bellaterra (Cerdanyola), Barcelona, Spain.

References

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

Bengio, Y. (2009). *Learning deep architectures for AI*. Delft, The Netherlands: Now Publishers Inc.

Berardino, A., Laparra, V., Ballé, J., & Simoncelli, E. (2017). Eigen-distortions of hierarchical representations. *Advances in Neural Information Processing Systems, 30*, 3533–3542

Bertalmío, M., Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Kane, D., & Malo, J. (2020). Evidence for the intrinsically nonlinear nature of receptive fields in vision. *Scientific Reports, 10*(1), 1–15, doi:10.1038/s41598-020-73113-0.

Blakemore, C., Nachmias, J., & Sutton, P. (1970). The perceived spatial frequency shift: Evidence for frequency-selective neurones in the human brain. *Journal of Physiology, 210*(3), 727–750.

Blakeslee, B., & McCourt, M. E. (1999). A multiscale spatial filtering account of the white effect, simultaneous brightness contrast and grating induction. *Vision Research, 39*(26), 4361–4377.

Blei, D. M., Kucukelbir, A., & McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association, 112*(518), 859–877.

Capilla, P., Diez, M., Luque, M., & Malo, J. (2004). Corresponding-pair procedure: A new approach to simulation of dichromatic color perception. *Journal of the Optical Society of America A, 21*(2), 176–186.

Cats faces 64x64. (n.d.), https://www.kaggle.com/spandan2/cats-faces-64x64-for-generative-models. Accessed March 17, 2021.

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002a). Nature and interaction of signals from the receptive field center and surround in macaque v1 neurons. *Journal of Neurophysiology, 88*(5), 2530–2546.

Cavanaugh, J. R., Bair, W., & Movshon, J. A. (2002b). Selectivity and spatial distribution of signals from the receptive field surround in macaque v1 neurons. *Journal of Neurophysiology, 88*(5), 2547–2556.

Cimpoi, M., Maji, S., Kokkinos, I., Mohamed, S., & Vedaldi, A. (2014). Describing textures in the wild. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3606–3613), doi:10.1109/CVPR.2014.461.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. In *Procedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 248–255), doi:10.1109/CVPR.2009.5206848.

DiMattina, C. (2016). Comparing models of contrast gain using psychophysical experiments. *Journal of Vision, 16*(9), 1.

DiMattina, C., & Zhang, K. (2013). Adaptive stimulus optimization for sensory systems neuroscience. *Frontiers in Neural Circuits, 7*, 101.

Ellemberg, D., Wilkinson, F., Wilson, H., & Arsenault, A. (1998). Apparent contrast and spatial frequency of local texture elements. *Journal of the Optical Society of America A, 15*(7), 1733–1739.

Fairchild, M. D. (2013). *Color appearance models*. John Wiley & Sons, doi:10.1002/9781118653128.

Foley, J. M. (1994). Human luminance pattern-vision mechanisms: Masking experiments require a new model. *Journal of the Optical Society of America A, 11*(6), 1710–1719.

Fruend, I. (2020). Constrained sampling from deep generative image models reveals mechanisms of human target detection. *Journal of Vision, 20*(7), 32.

Funke, C. M., Borowski, J., Stosio, K., Brendel, W., Wallis, T. S., & Bethge, M. (2021). Five points to check when comparing visual perception in humans and machines. *Journal of Vision, 21*(3), 16–16.

Gatys, L. A., Ecker, A. S., & Bethge, M. (2015). A neural algorithm of artistic style. *arXiv preprint arXiv:1508.06576*.

Gatys, L. A., Ecker, A. S., Bethge, M., Hertzmann, A., & Shechtman, E. (2017). Controlling perceptual factors in neural style transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3985–3993), doi:10.1109/CVPR.2017.397.

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 (ICLR)*, https://openreview.net/forum?id=Bygh9j09KX.

Golan, T., Raju, P., & Kriegeskorte, N. (2020). Controversial stimuli: Pitting neural networks against each other as models of human cognition. *Proceedings of the National Academy of Sciences of the United States of America, 117*(47), 29330–29337.

Gomez-Villa, A., Martin, A., Vazquez-Corral, J., & Bertalmío, M. (2019). Convolutional neural networks can be deceived by visual illusions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 12309–12317), doi:10.1109/CVPR.2019.01259.

Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Bertalmío, M., & Malo, J. (2020). Color illusions also deceive cnns for low-level vision tasks: Analysis and implications. *Vision Research, 176*, 156–174.

Goodfellow, I., McDaniel, P., & Papernot, N. (2018). Making machine learning robust against adversarial inputs. *Communications of the ACM, 61*(7), 56–66, doi:10.1145/3134599.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., . . . Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2* (pp. 2672–2680). Cambridge, MA: MIT Press. Available: http://dl.acm.org/citation.cfm?id=2969033.2969125. Accessed July 5, 2022.

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

Hirsch, E., & Tal, A. (2020). Color visual illusions: A statistics-based computational model. In *Advances in neural information processing systems* (Vol. *33*, pp. 9447–9458). Red Hook, NY: Curran Associates, Inc.

Isola, P., Zhu, J. Y., Zhou, T., & Efros, A. A. (2017). Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1125–1134), doi:10.1109/CVPR.2017.632.

Jacob, G., Pramod, R. T., Katti, H., & Arun, S. P. (2021). Qualitative similarities and differences in visual object representations between brains and deep networks. *Nature Communications, 12*(1), 1–14, doi:10.1038/s41467-021-22078-3.

Kietzmann, T., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience.* Retrieved from https://oxfordre.com/neuroscience/view/10.1093/acrefore/9780190264086.001.0001/acrefore-9780190264086-e-46. Accessed July 5, 2022.

Kim, B., Reif, E., Wattenberg, M., & Bengio, S. (2019). Do neural networks show gestalt phenomena? an exploration of the law of closure. *arXiv preprint arXiv:1903.01069*.

Kim, B., Reif, E., Wattenberg, M., Bengio, S., & Mozer, M. (2021). Neural networks trained on natural scenes exhibit gestalt closure. *Computational Brain Behavior, 4*, 251–263.

Kingdom, F. A. (2011). Lightness, brightness and transparency: A quarter century of new ideas, captivating demonstrations and unrelenting controversy. *Vision Research, 51*(7), 652–673.

Kobyzev, I., Prince, S. J., & Brubaker, M. A. (2020). Normalizing flows: An introduction and review of current methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 43*(11), 3964–3979, doi:10.1109/TPAMI.2020.2992934.

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*(1), 417–446.

Krizhevsky, A. (2009). Learning multiple layers of features from tiny images. Retrieved from https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf. Accessed July 5, 2022.

Laparra, V., Jiménez, 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.

Li, C. J., Luo, M. R., & Hunt, R. W. G. (2000). A revision of the CIECAM97s model. *Color Research & Application: Endorsed by Inter-Society Color Council, The Colour Group (Great Britain), Canadian Society for Color, Color Science Association of Japan, Dutch Society for the Study of Color, The Swedish Colour Centre Foundation, Colour Society of Australia, Centre Français de la Couleur, 25*(4), 260–266, doi:10.1002/1520-6378(200008)25:4⟨260::AID-COL6⟩3.0.CO;2-9.

Li, Q., Gomez-Villa, A., Bertalmio, M., & Malo, J. (2022). Contrast sensitivity functions in autoencoders. *Journal of Vision, 22*(6), 8–8, doi:10.1167/jov.22.6.8.

Liu, Z., Luo, P., Wang, X., & Tang, X. (2015). Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 3730–3738), doi:10.1109/ICCV.2015.425.

Ma, K., Duanmu, Z., Wang, Z., Wu, Q., Liu, W., Yong, H., … Zhang, L. (2020). Group maximum differentiation competition: Model comparison with few samples. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 42*(4), 851–864.

Majaj, N., & Pelli, D. (2018). Deep learning—using machine learning to study biological vision. *Journal of Vision, 18*(13), 2.

Malo, J., & Simoncelli, E. P. (2015). Geometrical and statistical properties of vision models obtained via maximum differentiation. *Proc. SPIE 9394, Human Vision and Electronic Imaging XX, 93940L*; https://doi.org/10.1117/12.2085653.

Martinez, M., Bertalmío, M., & Malo, J. (2019). In praise of artifice reloaded: Caution with natural image databases in modeling vision. *Frontiers in Neuroscience, 13*, 8.

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.

Martinez-Otero, L., Molano, M., Wang, X., Sommer, F., & Hirsch, J. (2014). Statistical wiring of thalamic receptive fields optimizes spatial sampling of the retinal image. *Neuron, 81*(4), 943–956.

Otazu, X., Parraga, C. A., & Vanrell, M. (2010). Toward a unified chromatic induction model. *Journal of Vision, 10*(12), 5–5.

Pang, Z., O'May, C. B., Choksi, B., & VanRullen, R. (2021). Predictive coding feedback results in perceived illusory contours in a recurrent neural network. *Neural Networks, 144*, 164–175, https://doi.org/10.1016/j.neunet.2021.08.02.

Purves, D., Monson, B. B., Sundararajan, J., & Wojtach, W. T. (2014). How biological vision succeeds in the physical world. *Proceedings of the National Academy of Science of the United States of America, 111*(13), 4750–4755.

Radford, A., Metz, L., & Chintala, S. (2015). Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*.

Rodriguez-Sanchez, R., Garcia, J., Fdez-Valdivia, J., & Fdez-Vidal, X. (1999). The rgff representational model: A system for the automatically learned partitioning of "visual patterns" in digital images. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 21*(10), 1044–1073.

Rodriguez-Sanchez, R., Garcia, J., Fdez-Valdivia, J., & Fdez-Vidal, X. (2000). Origins of illusory percepts in digital images. *Pattern Recognition, 33*(12), 2007–2017.

Schütt, H., & Wichmann, F. (2017). An image-computable psychophysical spatial vision model. *Journal of Vision, 17*(12), 12–12.

Shapiro, A. G., & Todorovic, D. (Eds.). *The oxford compendium of visual illusions* (New York, 2017; online edn, Oxford Scholarship Online, June 2017), http://dx.doi.org/10.1093/acprof:oso/9780199794607.001.0001. Accessed July 05, 2022.

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological Review, 34*(4), 273.

Tolhurst, D., & Thompson, P. (1975). Orientation illusions and after-effects: Inhibition between channels. *Vision Research, 15*(8), 967–972.

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

Wang, Z., & Simoncelli, E. (2008). Maximum differentiation (MAD) competition: A methodology for comparing computational models of perceptual quantities. *Journal of Vision, 8*(12), 1–13.

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. B., & Solomon, J. A. (1997). Model of visual contrast gain control and pattern masking.

Journal of the Optical Society of America A, 14(9), 2379–2391.

Yamins, D., & DiCarlo, J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience, 19*, 356–365.

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.

Zhang, R., Isola, P., Efros, A. A., Shechtman, E., & Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 586–595), doi:10.1109/CVPR.2018.00068.