

Comparing deep neural networks against humans: object recognition when the signal gets weaker*

Robert Geirhos^{1,2}, David H. J. Janssen^{1,3}, Heiko H. Schütt^{1,4,5},
 Jonas Rauber^{2,3}, Matthias Bethge^{2,6,7}, and Felix A. Wichmann^{1,2,6,8}

¹*Neural Information Processing Group, University of Tübingen, Germany*

²*Centre for Integrative Neuroscience, University of Tübingen, Germany*

³*Graduate School of Neural Information Processing, University of Tübingen, Germany*

⁴*Graduate School of Neural and Behavioural Sciences, University of Tübingen, Germany*

⁵*Department of Psychology, University of Potsdam, Germany*

⁶*Bernstein Center for Computational Neuroscience, Tübingen, Germany*

⁷*Max Planck Institute for Biological Cybernetics, Tübingen, Germany*

⁸*Max Planck Institute for Intelligent Systems, Tübingen, Germany*

Abstract

Human visual object recognition is typically rapid and seemingly effortless, as well as largely independent of viewpoint and object orientation. Until very recently, animate visual systems were the only ones capable of this remarkable computational feat. This has changed with the rise of a class of computer vision algorithms called deep neural networks (DNNs) that achieve human-level classification performance on object recognition tasks. Furthermore, a growing number of studies report similarities in the way DNNs and the human visual system process objects, suggesting that current DNNs may be good models of human visual object recognition. Yet there clearly exist important architectural and processing differences between state-of-the-art DNNs and the primate visual system. The potential behavioural consequences of these differences are not well understood. We aim to address this issue by comparing human and DNN generalisation abilities towards image degradations. We find the human visual system to be more robust to image manipulations like contrast reduction, additive noise or novel eidolon-distortions. In addition, we find progressively diverging classification error-patterns between humans and DNNs when the signal gets weaker, indicating that there may still be marked differences in the way humans and current DNNs perform visual object recognition. We envision that our findings as well as our carefully measured and freely available behavioural datasets¹ provide a new useful benchmark for the computer vision community to improve the robustness of DNNs and a motivation for neuroscientists to search for mechanisms in the brain that could facilitate this robustness.

*Please note that this paper has been greatly extended to an article published at NeurIPS 2018. Unless referring to specific results contained here, we recommend reading and citing the new paper (Geirhos et al., 2018).

¹Data and materials available at <https://github.com/rgeirhos/object-recognition>

1 Introduction

The visual recognition of objects by humans in everyday life is typically rapid and effortless, as well as largely independent of viewpoint and object orientation (e.g. Biederman, 1987). This ability of the primate visual system has been termed *core object recognition*, and much research has been devoted to understanding this process (see DiCarlo, Zoccolan, & Rust, 2012, for a review). We know, for example, that it is possible to reliably identify objects in the central visual field within a single fixation in less than 200 ms when viewing “standard” images (DiCarlo et al., 2012; Potter, 1976; Thorpe, Fize, & Marlot, 1996). Based on the rapidness of object recognition, core object recognition is often hypothesized to be achieved with mainly feedforward processing although feedback connections are ubiquitous in the primate brain (but see, e.g. Gerstner, 2005, for a critical assessment of this argument). Object recognition is believed to be realized by the ventral visual pathway, a hierarchical structure consisting of the areas V1-V2-V4-IT, with information from the retina reaching the cortex in V1 (e.g. Goodale & Milner, 1992). Although aspects of this process are known, others remain unclear.

Until very recently, animate visual systems were the only known systems capable of visual object recognition. This has changed, however, with the advent of brain-inspired deep neural networks (DNNs) which, after having been trained on millions of labeled images, achieve human-level performance when classifying objects in images of natural scenes (Krizhevsky, Sutskever, & Hinton, 2012). DNNs are now employed on a variety of tasks and set the new state-of-the-art, sometimes even surpassing human performance on tasks which were a few years ago thought to be beyond an algorithmic solution for decades to come (He, Zhang, Ren, & Sun, 2015; Silver et al., 2016). For an excellent introduction to DNNs see e.g. LeCun, Bengio, and Hinton (2015).

Although being in the first place an engineering discipline, the field of computer vision (interested in designing algorithms and building machines that can see) has always been interested in human vision: As in object recognition, our visual system is often remarkably successful, acting as de facto performance benchmark for many tasks. It is thus not surprising that there has always been an exchange between researchers in computer vision and human vision, such as the design of low-level image representations (Simoncelli, Freeman, Adelson, & Heeger, 1992; Simoncelli & Freeman, 1995) and the investigation of underlying coding principles such as redundancy reduction (Atick, 1992; Barlow, 1961; Olshausen & Field, 1996). With the advent of DNNs over the course of the last few years, this exchange has even deepened. It is thus not surprising that some studies have started investigating similarities between DNNs and human vision, drawing parallels between network and biological units or network layers and visual areas in the primate brain. Clearly, describing network units as biological neurons is an enormous simplification given the sophisticated nature and diversity of neurons in the brain (Douglas & Martin, 1991). Still, often the strength of a model lies not in replicating the original system but rather in its ability to capture the important aspects while abstracting from details of the implementation (e.g. Kriegeskorte, 2015).

1.1 Behavioural comparison between humans and DNNs

Thorough comparisons of human and DNN behaviour have been relatively rare. Behaviour goes well beyond overall performance: It comprises all performance changes as a function of certain stimulus properties, e.g. how classification accuracy depends on image background and contrast or the type and distribution of errors. Ideally, computational models of behaviour should not only be able to predict the overall accuracy of humans, but be able to describe behaviour on a more fine-grained level, e.g. in the current experiment on a category-by-category level. The ultimate goal should be the prediction of behaviour on a trial-by-trial basis, termed *molecular psychophysics* (Green, 1964; Schönfelder & Wichmann, 2012). An important early step into comparing human and DNN behaviour was the work of Lake, Zaremba, Fergus, and Gureckis (2015) reporting that DNNs are able to predict human category typicality ratings for images. Another study by Kheradpisheh, Ghodrati, Ganjtabesh, and Masquelier (2016) found largely similar performance on view-invariant, background-controlled object recognition and, for some DNNs, highly similar error distributions. On the other hand, so-called *adversarial examples* have cast some doubt on the idea of broad-ranging manlike DNN behaviour. For any given image it is possible to perturb it minimally in a principled way such that DNNs mis-classify it as belonging to an arbitrary other category (Szegedy et al., 2014). This slightly modified image is then called an adversarial example, and the manipulation is imperceptible to human observers (Szegedy et al., 2014).

The ease at which DNNs can be fooled speaks to the need of a careful, psychophysical comparison of human and DNN behaviour. As the possibility to systematically search for adversarial examples is very limited in humans, it is not known how to quantitatively compare the robustness of humans and machines against adversarial attacks. However, other behavioural measurements are known to have contributed much to our current understanding of the human visual system: Psychophysical investigations of human behaviour on object recognition tasks, measuring accuracies depending on image colour (grayscale vs. colour), image contrast and the amount of additive visual noise have been powerful means of exploring the human visual system, revealing much about the internal computations and mechanisms at work (e.g. Nachmias & Sansbury, 1974; Pelli & Farell, 1999; Wichmann, 1999; Henning, Bird, & Wichmann, 2002; Carandini & Heeger, 2012; Carandini, Heeger, & Movshon, 1997; Delorme, Richard, & Fabre-Thorpe, 2000). As a consequence, similar experiments might yield equally interesting insights into the functioning of DNNs, especially as a comparison to human behaviour. In this study, we obtain and analyse human and DNN classification data for the three above-mentioned, well-known image degradations. In addition, we employ a novel image manipulation method. The stimuli generated by the so-called eidolon-factory (Koenderink, Valsecchi, van Doorn, Wagemans, & Gegenfurtner, 2017) are parametrically controlled distortion of an image. Eidolons aim to evoke similar visual awareness as objects perceived in the periphery, giving them some biological justification. To our knowledge, we are among the first to measure DNN performance on these tasks and compare their behaviour to carefully measured human data, in particular using a controlled lab environment (instead of Amazon Mechanical Turk without sufficient control about presentation times, display calibration, viewing angles, and sustained attention of participants).

In this study, we employ a paradigm² aimed at comparing human observers and DNNs as fair as possible using an image categorization task with short presentation times (200 ms) along with backwards masking by a high-contrast $1/f$ noise mask, known to minimize, as much as psychophysically possible, feedback influence in the brain. This is important since all investigated networks rely on purely feed-forward computations. We perform psychophysical experiments on both human observers and DNNs to assess how robust the three currently well-known DNNs AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015) and VGG-16 (Simonyan & Zisserman, 2015) are towards image degradations in comparison to human participants.

DNNs provide exciting new opportunities for computational modelling of vision—and we envisage DNNs to have a major impact on our understanding of human vision in the future, essentially agreeing with assessments voiced by Kriegeskorte (2015), Kietzmann, McClure, and Kriegeskorte (2017) and VanRullen (2017). With this study, we aim to shed light on the behavioural consequences of the *currently existing* architectural, processing and training differences between the tested DNNs and the primate brain. We envision that our analyses as well as our carefully measured and freely available behavioural datasets (<https://github.com/rgeirhos/object-recognition>) may provide a new useful benchmark for the computer vision community to improve the robustness of DNNs and a motivation for neuroscientists to search for mechanisms in the brain that could facilitate human robustness.

2 Methods

2.1 General

We tested four ways of degrading images: conversion to grayscale, reducing image contrast, adding uniform white noise, and increasing the strength of a novel image distortion from the eidolon toolbox (Koenderink et al., 2017). Here we give an overview about the experimental procedure and about the observers and deep neural networks that performed these experiments. In the Appendix we provide details on the categories and image database used (Section A.1), as well as information about image preprocessing (Section A.2), including plots of example stimuli at different levels of signal strength. In Section A.3 of the Appendix we list the specifics of our experimental setup; for now it might be enough to know that images in the psychophysical experiments were always displayed at the center of the screen at a size of 3×3 degrees of visual angle.

2.2 Procedure

In each trial a fixation square was shown for 300 ms, followed by an image shown for only 200 ms, in turn immediately followed by a full-contrast pink noise mask ($1/f$ spectral shape) of the same size and duration. Participants had to choose one of 16 entry-level categories (see Section A.1 for details on these categories) by clicking on a response screen shown for 1500 ms³. During the whole experiment, the screen

²This is the same paradigm as reported by Wichmann et al. (2017).

³During practice trials the response screen was visible for another 300 ms in case an incorrect category was selected, and along with a short low beep sound the correct category was highlighted by setting its

background was set to a grey value of 0.454 in the [0, 1] range, corresponding to the mean grayscale value of all images in the dataset (41.17 cd/m^2). Figure 1 shows a schematic of a typical trial.

Prior to starting the experiment, all participants were shown the response screen and asked to name all categories to ensure that the task was fully clear. They were instructed to click on the category that they thought resembles the image best, and to guess if they were unsure. They were allowed to change their choice within the 1500 ms response interval; the last click on a category icon of the response screen was counted as the answer. The experiment was not self-paced, i.e. the response screen was always visible for 1500 ms and thus, each experimental trial lasted exactly 2200 ms (300 ms + 200 ms + 200 ms + 1500 ms).

On separate days we conducted four different experiments with 1,280 trials per participant each (eidolon-experiment: three sessions of 1,280 trials each). In the colour-experiment, we used two distinct conditions (colour vs. grayscale), whereas in the contrast-experiment and in the noise-experiment eight conditions were explored (corresponding to eight different contrast values or noise power densities, respectively). In the eidolon-experiment, 24 distinct conditions were employed. For each experiment, we randomly chose 80 images per category from the pool of images without replacement (i.e., no observer ever saw an image more than once throughout the entire experiment). Within each category, all conditions were counterbalanced. Stimulus selection was done individually for each participant to reduce the probability of an accidental bias in the image selection. Images within the experiments were presented in randomized order. After 256 trials (colour-experiment, noise-experiment and eidolon-experiment) and 128 trials (contrast-experiment), the mean performance of the last block was displayed on the screen, and observers were free to take a short break. The total time necessary to complete all trials was 47 minutes per session, not including breaks and practice trials. In total, the results reported in this article are based on 39,680 psychophysical trials. Ahead of each experiment, all observers conducted approximately 10 minutes of practice trials to gain familiarity with the task and the position of the categories on the response screen.

2.3 Observers and deep neural networks

Three observers participated in the colour-experiment (all male; 22 to 28 years; mean: 25 years). In each of the other experiments, five observers took part (contrast-experiment and noise-experiment: one female, four male; 20 to 28 years; mean: 23 years. Eidolon-experiment: three female, two male; 19 to 28 years; mean: 22 years). Subject-01 is an author and participated in all but the eidolon-experiment. All other participants were either paid €10 per hour for their participation or gained course credit. All observers were students of the University of Tübingen and reported normal or corrected-to-normal vision.

We used three DNNs for our analysis: AlexNet (Krizhevsky et al., 2012), GoogLeNet (Szegedy et al., 2015) and VGG-16 (Simonyan & Zisserman, 2015). All three networks were specified within the Caffe framework (Jia et al., 2014) and acquired as a pre-trained model. VGG-16 was obtained from the Visual Geometry Group's website (<http://www.robots.ox.ac.uk/~vgg/>); AlexNet and GoogLeNet

background to white.

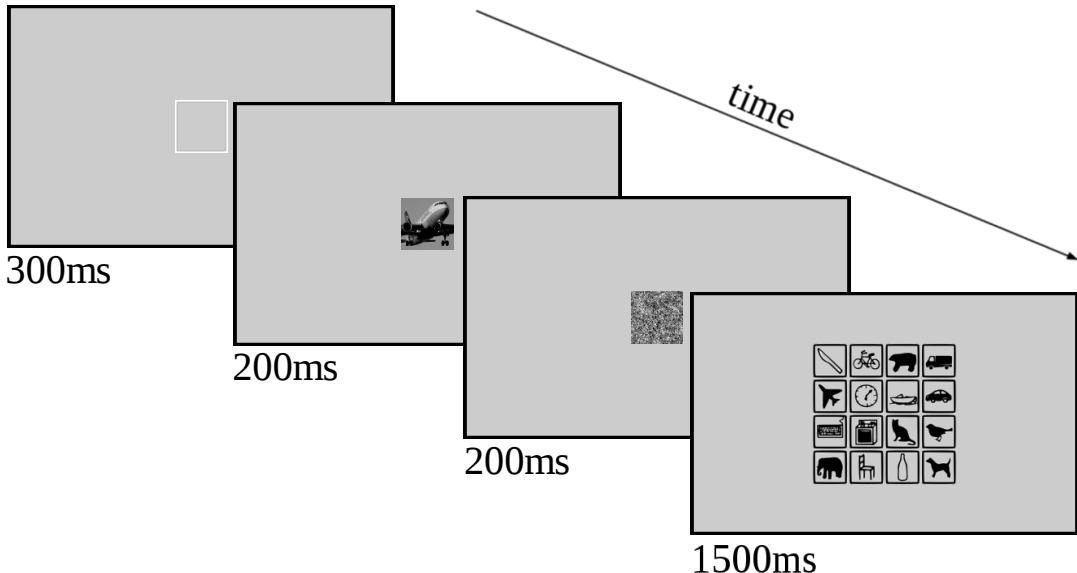


Figure 1. Schematic of a trial. After the presentation of a central fixation square (300 ms), the image was visible for 200 ms, followed immediately by a noise-mask with $1/f$ spectrum (200 ms). Then, a response screen appeared for 1500 ms, during which the observer clicked on a category. Note that we increased the contrast of the noise-mask in this figure for better visibility when printed. Categories row-wise from top to bottom: knife, bicycle, bear, truck, airplane, clock, boat, car, keyboard, oven, cat, bird, elephant, chair, bottle, dog. The icons are a modified version of the ones from the MS COCO website (<http://mscoco.org/explore/>).

from the BLVC model zoo website (<https://github.com/BVLC/caffe/wiki/Model-Zoo>). We reproduced the respective specified accuracies on the ILSVRC 2012 validation dataset in our setting.

All DNNs require images to be specified using RGB planes; to evaluate the performance using grayscale images we stacked a grayscale image three times in order to obtain the desired form specified by the caffe.io module (<https://github.com/BVLC/caffe/blob/master/python/caffe/io.py>). Images were fed through the networks using a single feedforward pass of the 224×224 pixels center crop.

3 Results

Trials in which human observers failed to click on any category were recorded as an incorrect answer in the data analysis, and are shown as a separate category (top row) in the confusion matrices (DNNs, obviously, never fail to respond). Such a failure to respond occurred in only 1.2% of all trials, and did not differ meaningfully between the different experiments. The terms 'accuracy' and 'performance' are used interchangeably. All data, if not stated otherwise, were analyzed using R version 3.2.3 (R Core Team, 2016).

3.1 Accuracy and response distribution entropy

When showing *accuracy* in any of the plots, the error bars provided have two distinct meanings: First, for DNNs they indicate the range of DNN accuracies resulting from seven⁴ runs on different images, with each run consisting of the same number of images per category and condition that a single human observer was exposed to. This serves as an estimate of the variability of DNN accuracies as a function of the random choice of images. Second, the error bars for human participants likewise correspond to the range of their accuracies (not the often shown S.E. of the means, which would be much smaller).

In addition we assessed the *response distribution entropy* of humans and DNNs as a function of image degradation strength. Entropy is a measure quantifying how close a distribution is to the uniform distribution (the higher the entropy, the closer it is). The distribution obtained by throwing a fair die many times should therefore have higher entropy than the distribution obtained from repeatedly throwing a rigged die. In the context of our experiments, it is used to measure whether observers or DNNs exhibit a bias towards certain categories: if so, the response distribution entropy will be lower than the maximal value of 4 bits (given 16 categories). We calculated the Shannon entropy H of response distribution \mathcal{X} as follows:

$$H(\mathcal{X}) = -\sum_{i=1}^{16} p(x_i) \log_2(p(x_i)), \text{ with } p(x_i) \text{ being the fraction of responses for category } i \text{ (e.g. } p(x_{car}) = 0.25 \text{ if an observer responds } \texttt{car} \text{ every fourth trial on average).}$$

3.1.1 Colour-experiment

We conducted a paired-samples *t*-test to assess the difference in accuracy between coloured and grayscale images for each network and observer (Table 2 in the Appendix). In order to account for multiple comparisons, the critical significance level of .05 was adjusted to $\frac{.05}{6} = .008\bar{3}$ by applying Bonferroni correction. As shown in Figure 2(a) all three networks performed significantly worse for grayscale images compared to coloured images (4.81% drop in performance on average: significant, but not dramatic in terms of effect size). Human observers, on the other hand, did not show on average a significant reduction in accuracy (only 1.88% accuracy drop for grayscale images). As can be seen from the range of human grayscale results, observers differed in their ability to cope with grayscale images.

The response distribution entropy shown in Figure 2(b) is innocuous: The DNNs distributed their responses perfectly among the 16 categories, and human observers are only marginally worse.

3.1.2 Contrast-experiment

As shown in Figure 3(a), accuracies for the contrast-experiment ranged from approximately 91–94% (VGG-16, GoogLeNet and human average) and 84% (AlexNet) for full contrast to chance level ($\frac{1}{16} = 6.25\%$) for 1% of contrast, except for VGG-16 which still achieves 17.5% correct responses. AlexNet’s and GoogLeNet’s performance dropped more rapidly than human and VGG-16’s performance for lower contrast levels.

⁴Seven runs are the maximum possible number of runs without ever showing an image to a DNN more than once per experiment.

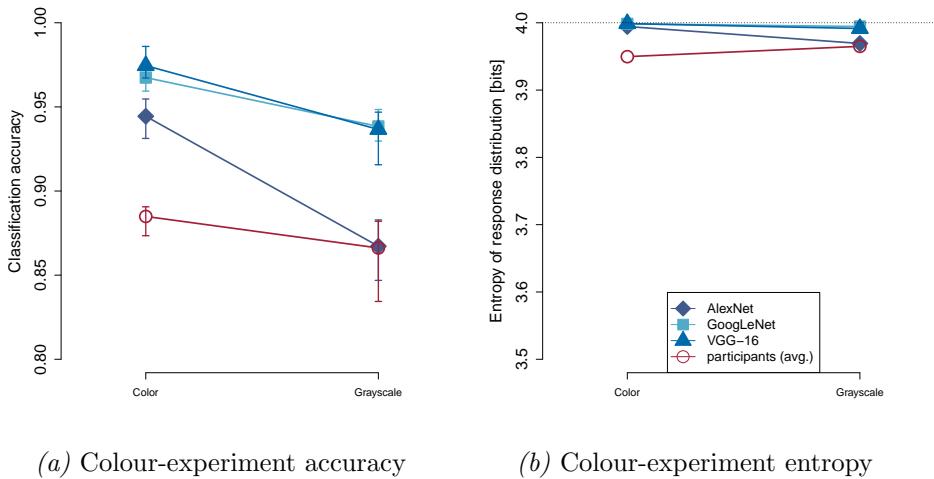


Figure 2. Results for the colour-experiment ($n=3$). **(a)** *Accuracy*. DNNs are shown in shades of blue, human data in red; diamonds correspond to AlexNet, squares to GoogLeNet, triangles to VGG-16, and circles to human observers; error bars as described in section 3.1. **(b)** *Response distribution entropy*. Plotting conventions as in (a).

The response distribution entropy shown in Figure 3(b) reveals, however, that all three DNNs showed an increasing bias towards few categories (in other words, they did no longer distribute their responses evenly among the 16 categories if the contrast was lowered). Human observers, on the other hand, still largely distributed their responses sensibly across the 16 categories.

3.1.3 Noise-experiment

The data for the noise-experiment were analyzed in the same way as the contrast-experiment data. Overall, we found drastic differences in classification accuracy, with human observers clearly outperforming all three networks. As can be seen in Figure 3(c), by increasing the noise width from 0.0 (no noise) to 0.1, VGG-16's performance drops from an accuracy of 89.91% to 44.02%; GoogLeNet's drops from 81.70% to 34.02% and AlexNet's from 70.00% to 19.29%. Human observers, on the other hand, only drop from 80.50% to 75.13%.

The response distribution entropy shown in Figure 3(d) shows again that all of the investigated DNNs exhibit a strong bias towards few categories if the images contained additive noise. For AlexNet and GoogLeNet, the response distribution entropy is close to 0 bits for a noise width of 0.6 or more, which means that they responded with a single category for these images (category `bottle` for both). Interestingly, these preferred categories are usually not the same across experiments or networks (Figures 6 and 9), and they do not simply match the probabilities of the categories in the ImageNet training database. The network responses therefore are not converging to their prior distribution, which would be a sensible way to behave in the absence of a signal. Human observers, as with low contrast, largely distributed their responses evenly across the 16 categories.

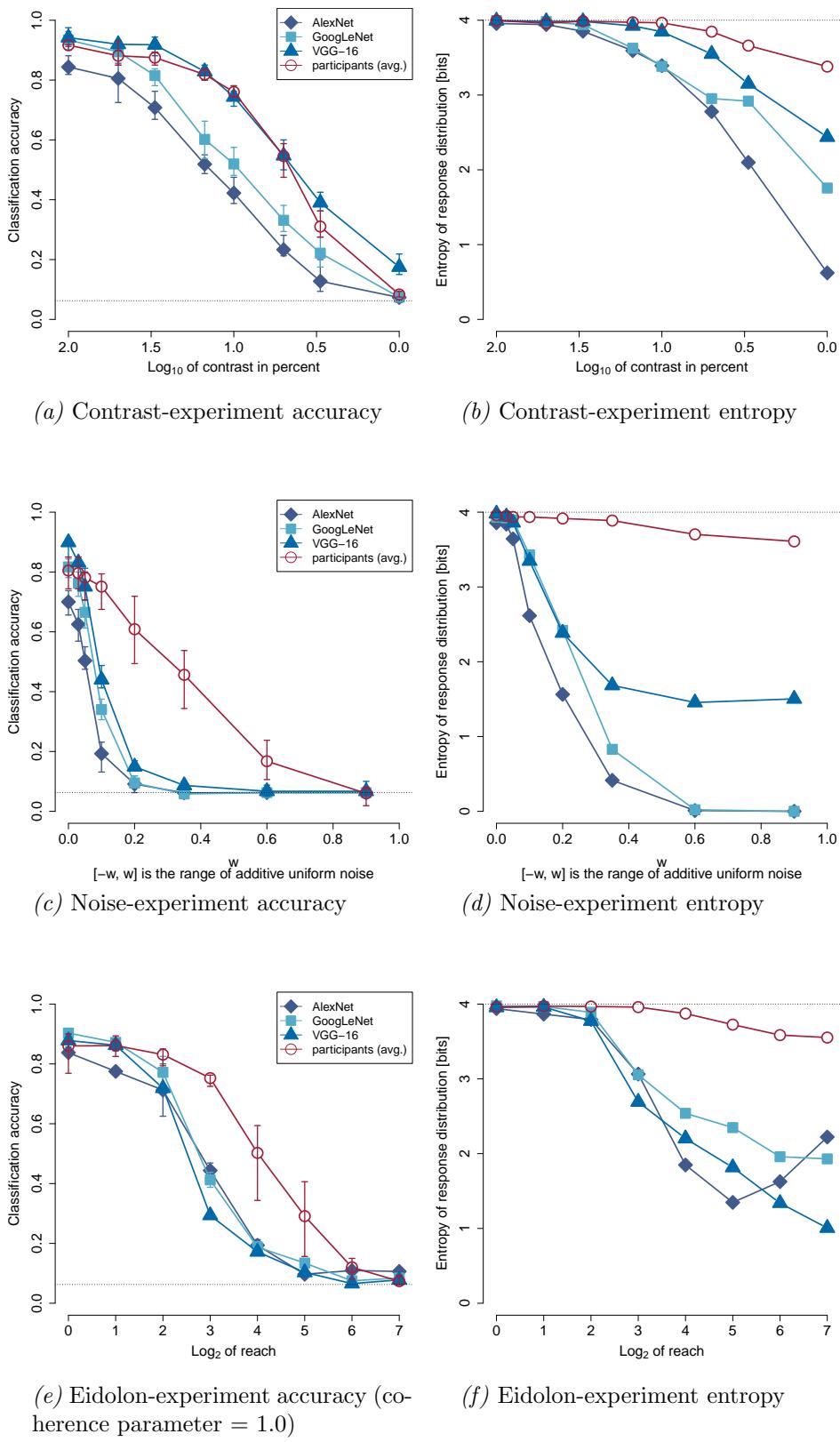


Figure 3. Results for contrast-, noise- and eidolon-experiment ($n=5$ each).

(a, c, e)(Left) *Accuracy*. Plotting conventions as in Figure 2.

(b, d, f)(Right) *Response distribution entropy*.

3.1.4 Eidolon-experiment

Results for the eidolon-experiment with maximal coherence of 1.0 are shown in Figure 3(e) and (f). The complete results of the eidolon-experiment for all coherence settings are provided in the Appendix, Figure 11. In terms of accuracy, network and human performance naturally were approximately equal for very low values of reach (no distortion, therefore high accuracies) and for very high values of reach (heavy distortion, accuracy at chance level). In the range between these extremes, their accuracies followed the typically observed s-shaped pattern known from most psychophysical experiments varying a single parameter. However, human observers clearly achieved higher accuracies than all three networks for intermediate distortions. In the full coherence case, the largest difference between network and human performance was observed for a reach value of $2^3 = 8$ (38.3% network accuracy vs. 75.3% human accuracy, averaged across networks and observers). The coherence-parameter, albeit having a considerable effect on the perceptual appearance of the stimuli, did not qualitatively change accuracies. Quantitatively, the performance was generally higher for high coherence values (see Figure 11 for details). Unlike in the case of contrast, the three networks showed only minor inter-network accuracy differences.

As for the contrast-experiment and the noise-experiment, we find all three networks to be strongly biased towards a few categories as shown by their low response distribution entropy (Figure 3f).

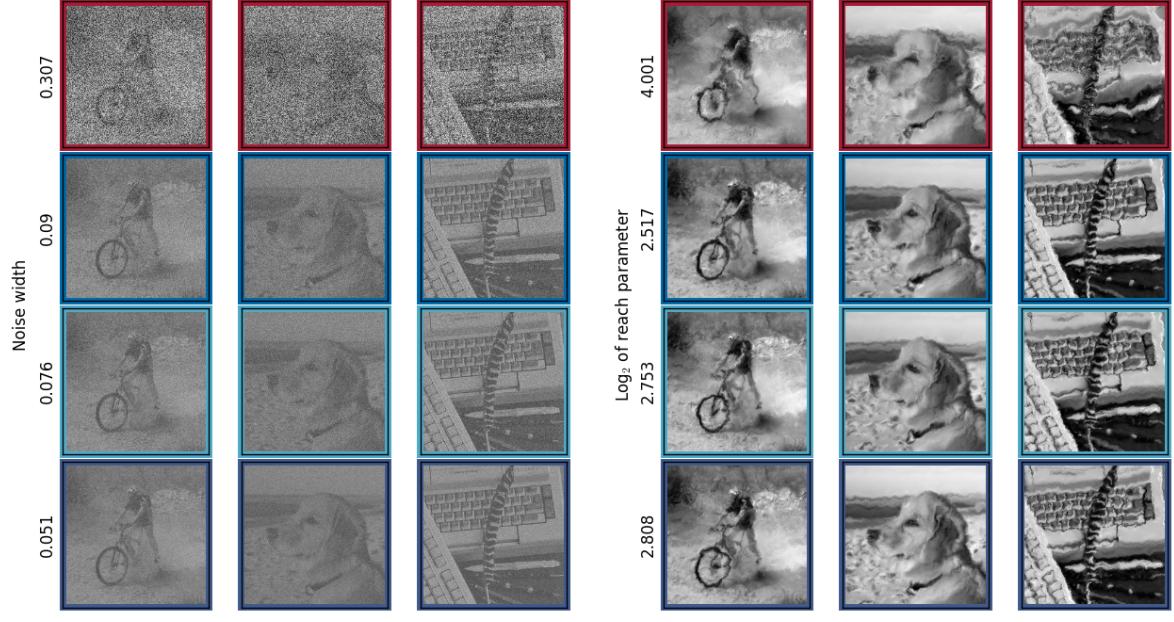
3.1.5 Performance visualization

Here we provide a visualization of the performance differences between the studied DNNs and human observers in terms of their generalisation ability (or robustness against image degradations). For all degradation-types—contrast, noise, eidolons with different coherence parameters—we estimated the stimulus levels corresponding to 50% classification accuracy. The stimulus levels were calculated assuming a linear relationship between the two closest data points measured in the experiments and shown in the left column of Figure 3.

Figure 4(a) shows the 50% accuracies for the noise-experiment, Figure 4(b) for the eidolon-experiment with maximal coherence (as in Figures 3(e) and (f)); the three illustration images of categories `bicycle`, `dog` and `keyboard` were drawn randomly from the pool of images used in the experiments. In both panels the top row shows the stimuli corresponding to 50% accuracy for the average human observer. The bottom three rows show the corresponding stimuli for VGG-16 (second row), GoogLeNet (third row) and AlexNet (bottom row). On a typical computer screen the more robust performance of human observers over DNNs should be readily appreciable. 50% accuracy stimulus plots for the contrast-experiment and the other conditions of the eidolon-experiment can be found in the Appendix, Figure 12.

3.2 Confusion and confusion difference matrices

Confusion matrices are a widely used tool for visualizing error patterns in multi-class classification data, providing insight into classification behavior (e.g.: does VGG-16 frequently confuse `dogs` with `cats`?). Figure 5(a) shows a standard confusion matrix of the colour condition in the colour-experiment (Section 3.1.1) for



(a) Noise-experiment

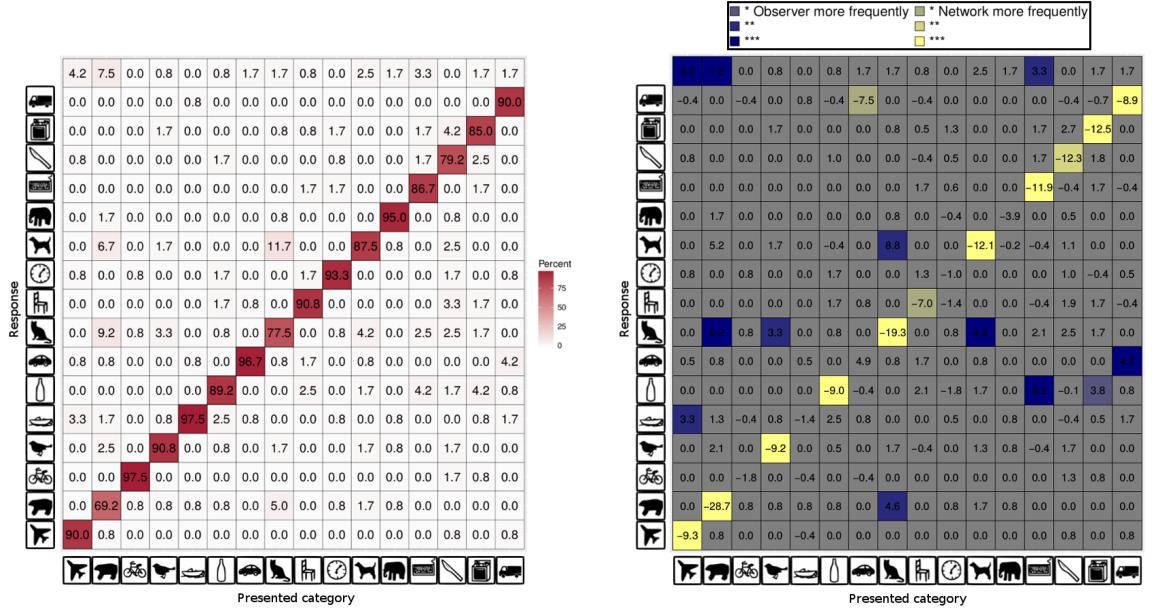
(b) Eidolon-experiment (coherence = 1.0)

Figure 4. Estimated stimuli corresponding to 50% classification accuracy. **(a)** *Noise-experiment*. **(b)** *Eidolon-experiment*. Coherence parameter = 1.0. Top row: stimuli corresponding to threshold for the [average human observer](#). Bottom three rows: stimuli corresponding to the same accuracy for [VGG-16](#) (second row), [GoogLeNet](#) (third row) and [AlexNet](#) (bottom row).

our human observers. Entries on the diagonal indicate correct classification, off-diagonal entries indicate errors, e.g. when a **cat** was presented on the screen (8^{th} column from left), human observers in 77.5% of all cases correctly clicked **cat** (8^{th} row from bottom in 8^{th} column), but in 11.7% clicked **dog** instead (11^{th} row from bottom in 8^{th} column). Participants failed to respond in 1.7% of cat trials in the colour condition of the colour-experiment (1st row from top in 8th column). Human observers typically confused physically and semantically closely related categories with each other, most notably some animal categories such as **bear**, **cat** and **dog**. Importantly, the same occurred for DNNs, albeit for different categories (confusions often between **car** and **truck**).

For the purpose of our analyses, however, we are mainly interested in comparisons between error patterns, e.g., do human observers more frequently confuse **dogs** with **cats** than VGG-16, and if so, significantly more? In order to being able to answer such questions, we developed a novel analysis and visualization technique, which we term *confusion difference matrix*.

A confusion difference matrix serves the purpose of showing the difference between two confusion matrices (e.g. human observers and VGG-16) and highlighting which differences are significant at the indicated, Bonferroni-corrected α -level. A confusion difference matrix is obtained in two steps: First, one calculates the difference between two confusion matrices' entries. In this newly obtained matrix, values close to zero indicate similar classification behavior, whereas differences point at diverging classification behavior. In the next step, we calculate whether a difference



(a) Colour-experiment, colour-condition, human participants.

(b) Colour-experiment, colour-condition, difference between human participants and VGG-16.

Figure 5. Confusion and confusion difference matrices for colour-experiment (colour-condition only). A failure to respond is shown as a separate category in the top row here. **(a)** *Standard confusion matrix.* Entries on the diagonal indicate correct classification, off-diagonal entries indicate errors. **(b)** *Confusion difference matrix.* All values indicate the signed difference of human observers' and VGG-16's confusion matrix entries. A positive sign indicates that human observers responded more frequently than VGG-16 to a certain category-response pair, for a negative sign vice-versa. The colour here indicates whether a difference for a certain cell is significant at $\alpha = \frac{5\%}{16 \cdot 17} (*)$, $\frac{1\%}{16 \cdot 17} (**)$ and $\frac{0.1\%}{16 \cdot 17} (***)$; these α -levels are Bonferroni corrected for multiple comparisons (16 categories \cdot 17 possible responses); see text in Section 3.2 for details.

for a certain cell is significant, and repeat this calculation for all cells. We calculate significance using a standard test of the probability of success in a Binomial experiment: If one thinks of the 120 colour-experiment trials in which human observers were exposed to a coloured `cat` image, of which they clicked on `cat` in 93 trials, as of a Binomial experiment with 93 successes out of 120 trials, is “93 out of 120” significantly higher or lower than we would expect under the null hypothesis of success probability $p = 96.8\%$ (VGG-16’s fraction of responses in this cell⁵)? The Binomial tests were performed with R, using the `binom.test` function of package `stats` which calculates the conservative Clopper-Pearson confidence interval⁶. The significance of a certain difference, in our experiments, is not used for traditional hypothesis testing but rather as a means of distinguishing between important and unimportant—perhaps only coincidental—*behavioural* differences between humans and DNNs even if their accuracies were equal. Confusion difference matrices thus visualize systematic category-dependent error pattern differences between human observers and DNNs—and they do this at a much more fine-grained, category-specific level than the response distribution entropy analyses shown in Section 3.1.

Figure 5(b) shows one confusion difference matrix for the colour-experiment (colour-condition only); all values indicate the signed difference of human observers’ and VGG-16’s confusion matrix entries. A positive sign indicates that human observers responded more frequently than VGG-16 to a certain category-response pair, for a negative sign vice-versa. VGG-16 is significantly better for many categories on the diagonal (correct classification) because—in the non-degraded colour condition—human observers make more errors, see Figure 2(a). Overall, however, most cells of the confusion difference matrix are grey, indicating very similar classification behaviour of human observers and VGG-16, not only in terms of overall accuracy and response entropy, but on a fine-grained category-by-category level.

In Figure 6 we show a confusion difference matrix grid for the noise-experiment (Section 3.1.3): nine confusion difference matrices for all three DNNs at three matched performance levels. Confusion difference matrices shown here are calculated as described above, however, with the important difference that we here show difference matrices for which human observers and networks have similar overall performance (accuracy difference $< 5\%$): we compare confusion matrices for different stimulus levels, but matched in performance⁷. The left column shows high performance (no noise for human observers, very little noise for DNNs; performance p-high = 80.5% which corresponds, in this order, to $w = 0.0, 0.0, 0.0$ and 0.03 for human observers, AlexNet, GoogLeNet and VGG-16). On the right, data for low performance (16.8%) are shown (high noise for human observers, moderate-to-low noise for DNNs, $w = 0.60, 0.10, 0.15, 0.19$) and in the middle results for medium performance: 45.6%, the condition for which human observers’ accuracy was approximately equal to $\frac{1}{2}(p\text{-high} + p\text{-low})$ (medium noise for human observers, low

⁵It would also be possible compare VGG-16’s number of successes to human observers’ fraction of responses. We always compared the network/observer/group with less trials to the one with more trials as null hypothesis—in the example above (colour-experiment, colour-condition, `cat` images): a total of 120 trials for human observers vs. 280 trials for VGG-16 (or any other network).

⁶If the network’s fraction of responses in a certain cell was 0.0% (not a single response in this cell), we set $p = 0.1\%$ and if it was 100.0% (every time a certain category was presented, the response lied in this cell), we set $p = 99.9\%$ instead.

⁷If DNN-human accuracy deviance was more than 5% for all conditions, we ran additional experiments to determine a suitable condition.

noise for DNNs; $w = 0.35, 0.06, 0.08, 0.10$).

Showing confusion difference matrices at matched performance levels—rather than at the same stimulus level—has the advantage that the sum over all entries of the to-be-compared confusion matrices is the same, i.e. for equally behaving classifiers the expectation is to obtain mainly grey (non-significant) cells. However, inspection of Figure 6 shows this only to be the case for the easy, low-noise, condition (left column). With increasing task difficulty (more noise), network and human behavior diverges substantially. As the noise level increases, all networks show a rapidly increasing bias for a few categories. For a noise level of $w = 0.35$, AlexNet and GoogLeNet almost exclusively respond **bottle** (92.32% and 85.71%), whereas VGG-16 homes in on category **dog** for 62.50% of all images. Note that this bias for certain categories is neither consistent across networks nor across the image manipulations.

A similar pattern emerged for the contrast-experiment: Classification behavior on a stimulus-by-stimulus basis for all three DNNs is close to that of human observers for a high accuracy (nominal contrast level). However, as task difficulty increases, the classification behavior of all three DNNs differs significantly from human behavior, despite being matched in overall accuracy (see the Appendix, Figure 9).

4 Discussion

We psychophysically examined to what extend currently well-known DNNs (AlexNet, GoogLeNet and VGG-16) could be a good model for human feedforward visual object recognition. So far thorough comparisons of DNNs and human observers on behavioural grounds have been rare. Here we proposed a fair and psychophysically accurate way of comparing network and human performance on a number of object recognition tasks: measuring categorization accuracy for single-fixation, briefly presented (200 ms) and backward-masked images as a function of colour, contrast, uniform noise, and eidolon-type distortions.

We find that DNNs outperform human observers by a significant margin for non-distorted, coloured images—the images the DNNs were specifically trained on. We speculate that this may in part be due to some images in the ImageNet database containing images with small animals in the background, making it tough to decide whether it is a **cat**, **dog** or even a **bear**. Given that the images were labelled by human observers—who thus are the ultimate guardians of what counts as right or wrong—it is clear that for unlimited inspection time and sufficient training human observers will equal DNN performance, as shown by the benchmark results of Russakovsky et al. (2015), obtained using expert annotators. What we established, however, is that under conditions minimizing feedback, current DNNs already outperform human observers on the type of images found on ImageNet..

Our first experiment also shows that human observers’ accuracy suffers only marginally when images are converted to grayscale in comparison to coloured images, consistent with previous studies (Delorme et al., 2000; Kubilius, Bracci, & Op de Beeck, 2016; Wichmann, Braun, & Gegenfurtner, 2006)⁸. For all three tested

⁸Consistent, also, with the popularity of black-and-white movies and photography: If we had a hard time recognizing objects and scenes in black-and-white we doubt that they’d ever have been a mass medium in the early and mid 20th century.

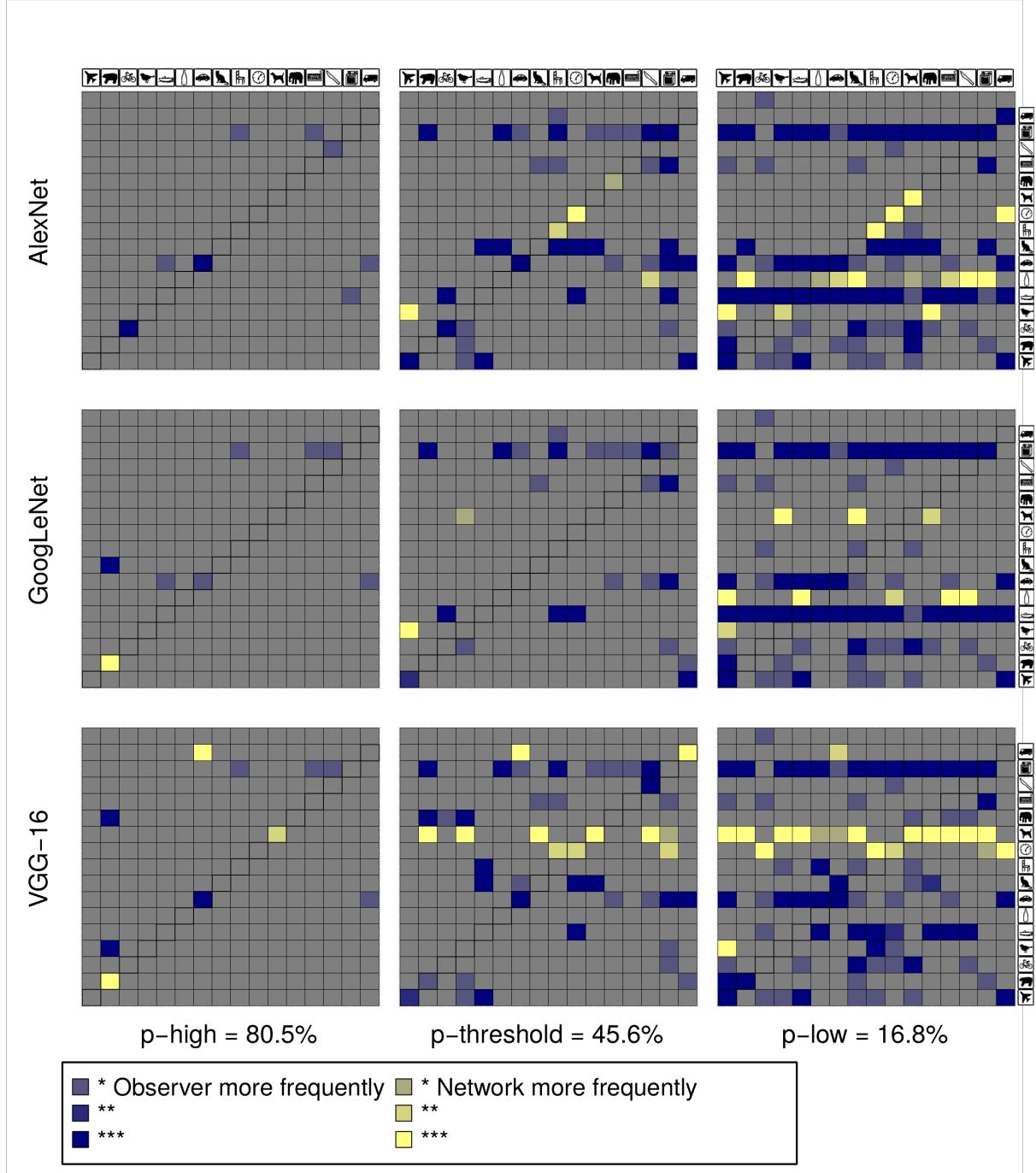


Figure 6. Confusion difference grid for the noise-experiment. Every confusion difference matrix shown here is calculated as described in the text. We here show difference matrices for which human observers and networks have similar overall performance. The left column shows for human observers the condition in which they performed best (no noise, performance $p\text{-high} = 80.5\%$) and for all three networks the condition for which they achieved the same accuracy. On the right, data for low performance (16.8%) are shown and in the middle results for medium performance: 45.6% . Colour indicates whether a difference for a certain cell is significant at $\alpha = \frac{5\%}{16 \cdot 17 \cdot 9} (*)$, $\frac{1\%}{16 \cdot 17 \cdot 9} (**)$ and $\frac{0.1\%}{16 \cdot 17 \cdot 9} (***)$; these α -levels are Bonferroni corrected for multiple comparisons (16 categories \cdot 17 possible responses \cdot 9 confusion difference matrices).

DNNs the performance decrement is significant, however. Particularly AlexNet shows a largish drop in performance ($> 7\%$), which is not human-like. VGG-16 and GoogLeNet rely less on colour information, but still somewhat more than the average human observer.

Our second experiment examined accuracy as a function of image contrast. Human participants outperform AlexNet and GoogLeNet (but not VGG-16) in the low contrast regime, where all DNNs display an increasing bias for certain categories (Figure 3(b) as well as Figure 9). Almost all images on which the networks were originally trained had full contrast. Apparently, training on ImageNet in itself only leads to a suboptimal contrast invariance. There are several solutions to overcome this deficiency: One option would be to include an explicit image pre-processing stage or have the first layer of the networks normalise the contrast. Another option would be to augment the training data with images of various contrast levels, which in itself might be a worthwhile data augmentation technique even if one does not expect low contrast images at test time. In the human visual system, probably as a response to the requirement of increasing stimulus identification accuracy (Geisler & Albrecht, 1995), a mechanism called *contrast gain control* evolved, serving the human visual system as a contrast normalization technique by taking into account the average local contrast rather than the absolute, global contrast (e.g. Carandini et al., 1997; Heeger, 1992; Sinz & Bethge, 2009, 2013). This has the (side-) effect that human observers can easily perform object recognition across a variety of contrast levels. Thus yet another, though clearly more labour-intensive way of improving contrast invariance in DNNs would be to incorporate a mechanism of contrast gain control directly in the network architecture. Early vision models could serve as a role model (e.g., Goris, Putzeys, Wagemans, & Wichmann, 2013; Schütt & Wichmann, 2017).

Our third experiment, adding uniform white noise to images, shows very clear discrepancies between the performance of DNNs and human observers. Note that, if anything, we might have underestimated human performance: randomly shuffling all conditions of an experiment instead of using blocks of a certain stimulus level are likely to yield accuracies that are lower than those possible in a blocked constant stimulus setting (Blackwell, 1953; Jäkel & Wichmann, 2006). Already at a moderate noise level, however, network accuracies drop sharply, whereas human observers are only slightly affected (visualized in Figure 4 showing stimuli corresponding to 50% accuracy for human observers and the three networks). Consistent with recent results by Dodge and Karam (2017), our data clearly show that the human visual system is currently much more robust to noise than any of the investigated DNNs.

Another noteworthy finding is that the three DNNs exhibit considerable inter-model differences; their ability to cope with grayscale and different levels of contrast and noise differs substantially. In combination with other studies finding moderate to striking differences (e.g. Cadieu et al., 2014; Kheradpisheh et al., 2016; Lake et al., 2015), this speaks to the need of carefully distinguishing between models rather than treating DNNs as a single model type as it is perhaps sometimes done in vision science.

Recent studies on so-called *adversarial examples* in DNNs have demonstrated that, for a given image, it is possible to construct a minimally perturbed version of this image which DNNs will misclassify as belonging to an arbitrary different category (Szegedy et al., 2014). Here we show that comparatively large but purely random distortions such as additive uniform noise also lead to poor network per-

formance. Our detailed analyses of the network decisions offer some clues on what could contribute to robustness against these distortions, as the measurement of confusion matrices for different signal-to-noise ratios is a powerful tool to reveal important algorithmic differences of visual decision making in humans and DNNs. All three DNNs show an escalating bias towards few categories as noise power density increases (Figures 3 and 6), indicating that there might be something inherent to noisy images that causes the networks to select a single category. The networks might perceive the noise as being part of the object and its texture while human observers perceive the noise like a layer in front of the image (you may judge this yourself by looking at the stimuli in Figure 7). This might be the achievement of a mechanism for depth-layering of surface representations implemented by mid-level vision, which is thought to help the human brain to encode spatial relations and to order surfaces in space (Kubilius, Wagemans, & Op de Beeck, 2014). Incorporating such a depth-layering mechanism may lead to an improvement of current DNNs, enabling them to robustly classify objects even when they are distorted in a way that the network was not exposed to during training. It remains subject to future investigations to determine whether such a mechanism will emerge from augmenting the training regime with different kinds of noise, or whether changes in the network architecture, potentially inspired by knowledge about mid-level vision, are necessary to achieve this feat.

One might argue that human observers, through experience and evolution, were exposed to some image distortions (e.g. fog or snow) and therefore have an advantage over current DNNs. However, an extensive exposure to eidolon-type distortions seems exceedingly unlikely. And yet, human observers were considerably better at recognising eidolon-distorted objects, largely unaffected by the different perceptual appearance for different eidolon parameter combinations (reach, coherence). This indicates that the representations learned by the human visual system go beyond being trained on certain distortions as they generalise towards previously unseen distortions. We believe that achieving such robust representations that generalise towards novel distortions are the key to achieve robust deep neural network performance, as the number of possible distortions is literally unlimited.

4.1 Conclusion

We conducted a behavioural, psychophysical comparison of human and DNN object recognition robustness against image degradations. While it has long been noticed that DNNs are extremely fragile against adversarial attacks, our results show that they are also more prone to random perturbations than humans. In comparison to human observers, we find the classification performance of three currently well-known DNNs trained on ImageNet—AlexNet, GoogLeNet and VGG-16—to decline rapidly with decreasing signal-to-noise ratio under image degradations like additive noise or eidolon-type distortions. Additionally, by measuring and comparing confusion matrices we find progressively diverging patterns of classification errors between humans and DNNs with weaker signals, and considerable inter-model differences. Our results demonstrate that there are still marked differences in the way humans and current DNNs process object information. We envision that our findings and the freely available behavioural datasets may provide a new useful benchmark for improving DNN robustness and a motivation for neuroscientists to search for mechanisms in the brain that could facilitate this robustness.

Author contributions

R.G., H.H.S. and F.A.W. designed the study; R.G. performed the network experiments with input from D.H.J.J. and H.H.S.; R.G. acquired the behavioural data with input from F.A.W.; R.G., H.H.S., J.R. M.B. and F.A.W. analysed and interpreted the data. R.G. and F.A.W. wrote the paper with significant input from H.H.S., J.R., and M.B.

Acknowledgement

This work has been funded, in part, by the German Federal Ministry of Education and Research (BMBF) through the Bernstein Computational Neuroscience Program Tübingen (FKZ: 01GQ1002) as well as the German Research Foundation (DFG; Sachbeihilfe Wi 2103/4-1 and SFB 1233 on “Robust Vision”). M.B. acknowledges support by the Centre for Integrative Neuroscience Tübingen (EXC 307) and by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior/Interior Business Center (DoI/IBC) contract number D16PC00003. J.R. is funded by the BOSCH Forschungsstiftung.

We would like to thank Tom Wallis for providing the MATLAB source code of one of his experiments, and for allowing us to use and modify it; Silke Gramer for administrative and Uli Wannek for technical support, as well as Britta Lewke for the method of creating response icons and Patricia Rubisch for help with testing human observers.

References

- Atick, J. J. (1992). Could information theory provide an ecological theory of sensory processing? *Network: Computation in neural systems*, 3(2), 213–251.
- Barlow, H. B. (1961). Possible principles underlying the transformations of sensory messages. *Sensory Communication*, 217–234.
- Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological Review*, 94(2), 115–147.
- Blackwell, H. R. (1953). Psychophysical thresholds: experimental studies of methods of measurement. *University of Michigan Engineering Research Institute Bulletin*, No. 36, xiii–227.
- Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision*, 10, 433–436.
- Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardis, D., Solomon, E. A., ... DiCarlo, J. J. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, 10(12).
- Carandini, M., & Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1), 51–62.
- Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *The Journal of Neuroscience*, 17(21), 8621–8644.
- Delorme, A., Richard, G., & Fabre-Thorpe, M. (2000). Ultra-rapid categorisation of natural scenes does not rely on colour cues: a study in monkeys and humans. *Vision Research*, 40(16), 2187–2200.
- DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3), 415–434.
- Dodge, S., & Karam, L. (2017). *A study and comparison of human and deep learning recognition performance under visual distortions*. arXiv preprint arXiv:1705.02498.

- Douglas, R. J., & Martin, K. A. C. (1991). Opening the grey box. *Trends in Neurosciences*, 14(7), 286-293.
- Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. In *Advances in neural information processing systems* (pp. 7548–7560).
- Geisler, W. S., & Albrecht, D. G. (1995). Bayesian analysis of identification performance in monkey visual cortex: nonlinear mechanisms and stimulus certainty. *Vision Research*, 35(19), 2723-2730.
- Gerstner, W. (2005). How can the brain be so fast? In J. L. van Hemmen & T. J. Sejnowski (Eds.), *23 problems in systems neuroscience* (p. 135-142). Oxford University Press.
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1), 20–25.
- Goris, R. L., Putzeys, T., Wagemans, J., & Wichmann, F. A. (2013). A neural population model for visual pattern detection. *Psychological Review*, 120(3), 472.
- Green, D. M. (1964). Consistency of auditory judgements. *Psychological Review*, 71(5), 592-407.
- He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on ImageNet classification. In *Proceedings of the IEEE International Conference on Computer Vision* (pp. 1026–1034).
- Heeger, D. J. (1992). Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9, 181-197.
- Henning, G. B., Bird, C. M., & Wichmann, F. A. (2002). Contrast discrimination with pulse trains in pink noise. *Journal of the Optical Society of America A*, 19(7), 1259-1266.
- Jäkel, F., & Wichmann, F. A. (2006). Spatial four-alternative forced-choice method is the preferred psychophysical method for naïve observers. *Journal of Vision*, 6(11), 1307-1322.
- Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R., ... Darrell, T. (2014). Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the 22nd ACM International Conference on Multimedia* (pp. 675–678).
- Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., & Masquelier, T. (2016). Deep networks resemble human feed-forward vision in invariant object recognition. arXiv preprint arXiv:1508.03929.
- Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2017). Deep neural networks in computational neuroscience. *bioRxiv*, <http://dx.doi.org/10.1101/133504>.
- Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., & Broussard, C. (2007). What's new in Psychtoolbox-3. *Perception*, 36(14), 1.
- Koenderink, J., Valseschi, M., van Doorn, A., Wagemans, J., & Gegenfurtner, K. (2017). Eidolons: Novel stimuli for vision research. *Journal of Vision*, 17(2), 7–7.
- Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, 1(15), 417–446.
- Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems* (pp. 1097–1105).
- Kubilius, J., Bracci, S., & Op de Beeck, H. P. (2016). Deep neural networks as a computational model for human shape sensitivity. *PLoS Computational Biology*, 12(4), e1004896.
- Kubilius, J., Wagemans, J., & Op de Beeck, H. P. (2014). A conceptual framework of computations in mid-level vision. *Frontiers in Computational Neuroscience*, 8, 158.
- Lake, B. M., Zaremba, W., Fergus, R., & Gureckis, T. M. (2015). Deep Neural Networks predict category typicality ratings for images. In *Proceedings of the 37th Annual Conference of the Cognitive Science Society*.
- LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, 521(7553), 436–444.
- Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., ... Zitnick, C. L. (2015). Microsoft COCO: Common objects in context. In *European Conference on*

- Computer Vision* (pp. 740–755).
- Nachmias, J., & Sansbury, R. V. (1974). Grating contrast: Discrimination may be better than detection. *Vision Research*, 14(10), 1039-1042.
- 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.
- Pelli, D. G., & Farell, B. (1999). Why use noise? *Journal of the Optical Society of America A*, 16(3), 647-653.
- Potter, M. C. (1976). Short-term conceptual memory for pictures. *Journal of Experimental Psychology: human learning and memory*, 2(5), 509.
- R Core Team. (2016). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from <https://www.R-project.org/>
- Rosch, E. (1999). Principles of categorization. In E. Margolis & S. Laurence (Eds.), *Concepts: core readings* (pp. 189–206).
- Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., ... Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3), 211–252.
- Schütt, H. H., & Wichmann, F. A. (2017). *An image-computable psychophysical spatial vision model* (Vol. under review).
- Schönfelder, V. H., & Wichmann, F. A. (2012). Sparse regularized regression identifies behaviorally-relevant stimulus features from psychophysical data. *Journal of the Acoustical Society of America*, 131(5), 3953-3969.
- Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., ... Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, 529(7587), 484–489.
- Simoncelli, E. P., & Freeman, W. T. (1995). The steerable pyramid: a flexible architecture for multi-scale derivative computation. In *2nd ieee international conference on image processing* (Vol. III, p. 444-447). Washington, DC.
- Simoncelli, E. P., Freeman, W. T., Adelson, E. H., & Heeger, D. J. (1992). Shiftable multiscale transforms. *IEEE Transactions on Information Theory*, 38(2), 587-607.
- Simonyan, K., & Zisserman, A. (2015). *Very deep convolutional networks for large-scale image recognition*. arXiv preprint arXiv:1409.1556.
- Sinz, F., & Bethge, M. (2009). The conjoint effect of divisive normalization and orientation selectivity on redundancy reduction. In *Advances in neural information processing systems* (pp. 1521–1528).
- Sinz, F., & Bethge, M. (2013). Temporal adaptation enhances efficient contrast gain control on natural images. *PLoS Comput Biol*, 9(1), e1002889.
- Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 1–9).
- Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2014). *Intriguing properties of neural networks*. arXiv preprint arXiv:1312.6199.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582), 520–522.
- Van der Walt, S., Schönberger, J. L., Nunez-Iglesias, J., Boulogne, F., Warner, J. D., Yager, N., ... Yu, T. (2014). scikit-image: image processing in Python. *PeerJ*, 2, e453.
- VanRullen, R. (2017). Perception science in the age of deep neural networks. *Frontiers in Psychology*, 8, 142 , doi: 10.3389/fpsyg.2017.00142.
- Wichmann, F. A. (1999). *Some aspects of modelling human spatial vision: Contrast discrimination* (Unpublished doctoral dissertation). The University of Oxford.
- Wichmann, F. A., Braun, D. I., & Gegenfurtner, K. R. (2006). Phase noise and the classification of natural images. *Vision Research*, 46(8), 1520–1529.
- Wichmann, F. A., Janssen, D. H., Geirhos, R., Aguilar, G., Schütt, H. H., Maertens, M., & Bethge, M. (2017). Methods and measurements to compare men against machines. *Electronic Imaging, Human Vision and Electronic Imaging*, 2017(14), 36–45.

Appendix

A Stimuli & Apparatus

A.1 Categories and image database

The images serving as psychophysical stimuli were images extracted from the training set of the ImageNet Large Scale Visual Recognition Challenge 2012 database (Russakovsky et al., 2015). This database contains millions of labeled images grouped into 1,000 very fine-grained categories (e.g., the database contains over a hundred different dog breeds). If human observers are asked to name objects, however, they most naturally categorize them into many fewer so-called basic or entry-level categories, e.g. dog rather than **German shepherd** (Rosch, 1999). The Microsoft COCO (MS COCO) database (Lin et al., 2015) is an image database structured according to 91 such entry-level categories, making it an excellent source of categories for an object recognition task. Thus for our experiments we fused the carefully selected entry-level categories in the MS COCO database with the large quantity of images in ImageNet. Using WordNet’s *hypernym* relationship (x is a hypernym of y if y is a “kind of” x , e.g., dog is a hypernym of **German shepherd**), we mapped every ImageNet label to an entry-level category of MS COCO in case such a relationship exists, retaining 16 clearly non-ambiguous categories with sufficiently many images within each category (see Figure 1 for a iconic representation of the 16 categories; the figure shows the icons used for the observers during the experiment). A complete list of ImageNet labels used for the experiments can be found in our github repository, <https://github.com/rgeirhos/object-recognition>. Since all investigated DNNs, when shown an image, output classification predictions for all 1,000 ImageNet categories, we disregarded all predictions for categories that were not mapped to any of the 16 entry-level categories. Amongst the remaining categories, the entry-level category corresponding to the ImageNet category with the highest probability (top-1) was selected as the network’s response. This way, the DNN response selection corresponds directly to the forced-choice paradigm for our human observers.

A.2 Image preprocessing

We used Python (Version 2.7.11) for all image preprocessing and for running the DNN experiments. From the pool of ImageNet images of the 16 entry-level categories, we excluded all grayscale images (1%) as well as all images not at least 256×256 pixels in size (11% of non-grayscale images). We then cropped all images to a center patch of 256×256 pixels as follows: First, every image was cropped to the largest possible center square. This center square was then downsampled to the desired size with `PIL.Image.thumbnail((256, 256), Image.ANTIALIAS)`. Human observers get adapted to the mean luminance of the display during experiments, and thus images which are either very bright or very dark may be harder to recognize due to their very different perceived brightness. We therefore excluded all images which had a mean deviating more than two standard deviations from that of other images (5% of correct-sized colour-images excluded). In total we retained 213,555 images from ImageNet.

For the experiments using grayscale images the stimuli were converted using the `rgb2gray` method (Van der Walt et al., 2014) in Python. This was the case

for all experiments and conditions except for the colour-condition of the colour-experiment. For the contrast-experiment, we employed eight different contrast levels $c \in \{1, 3, 5, 10, 15, 30, 50, 100\}\%$. For an image in the $[0, 1]$ range, scaling the image to a new contrast level c was achieved by computing $new_value = \frac{c}{100\%} \cdot old_value + \frac{1 - \frac{c}{100\%}}{2}$ for each pixel. For the noise-experiment, we first scaled all images to a contrast level of $c = 30\%$. Subsequently, white uniform noise of range $[-w, w]$ was added pixelwise, $w \in \{0.0, 0.03, 0.05, 0.1, 0.2, 0.35, 0.6, 0.9\}$. In case this resulted in a value out of the $[0, 1]$ range, this value was clipped to either 0 or 1. By design, this never occurred for a noise range less or equal to 0.35 due to the reduced contrast (see above). For $w = 0.6$, clipping occurred in 17.2% of all pixels and for $w = 0.9$ in 44.4% of all pixels. Clearly, clipping pixels changes the spectrum of the noise and is undesirable. However, as can be seen in Section 3, specifically Figure 3, all DNNs were already at chance performance for noise with a w of 0.35 (no clipping), whereas human observers were still supra-threshold. Thus changes in the exact shape of the spectrum of the noise due to clipping have no effect on the conclusions drawn from our experiment. See Figure 7 for example contrast and noise stimuli.

All eidolon stimuli were generated using the eidolon toolbox for Python obtained from <https://github.com/gestaltrevision/Eidolon>, more specifically its `PartiallyCoherentDisarray(image, reach, coherence, grain)` function. Using a combination of the three parameters reach, coherence and grain, one obtains a distorted version of the original image (a so-called eidolon). The parameters reach and coherence were varied in the experiment, grain was held constant with a value of 10.0 throughout the experiment (grain indicates how fine-grained the distortion is; a value of 10.0 corresponds to a medium-grainy distortion). Reach $\in \{1.0, 2.0, 4.0, 8.0, 16.0, 32.0, 64.0, 128.0\}$ is an amplitude-like parameter indicating the strength of the distortion, coherence $\in \{0.0, 0.3, 1.0\}$ defines the relationship between local and global image structure. Those two parameters were fully crossed, resulting in $8 \cdot 3 = 24$ different eidolon conditions. A high coherence value "retains the local image structure even when the global image structure is destroyed" (Koenderink et al., 2017, p. 10). A coherence value of 0.0 corresponds to 'completely incoherent', a value of 1.0 to 'fully coherent'. The third value 0.3 was chosen because it produces images that perceptually lie—as informally determined by the authors—in the middle between those two extremes. See Figure 8 for example eidolon stimuli.

All images, prior to showing them to human observers or DNNs, were saved in the JPEG format using the default settings of the `skimage.io.imsave` function. The JPEG format was chosen because the image training database for all three networks, ImageNet (Russakovsky et al., 2015), consists of JPEG images. However, one has to bear in mind that JPEG compression is lossy and introduces, under certain circumstances, unwanted artefacts. We therefore ran all DNN experiments additionally saving them in the (up to rounding issues) lossless PNG format. We did not find any noteworthy differences in DNN results for colour-, noise- and eidolon-experiment but did find some for the contrast-experiment, which is why we report data for PNG images in the case of the contrast-experiment (Figure 3). In particular, saving a low-contrast image to JPEG may result in a slightly different contrast level, which is why we refer to the contrast level of JPEG images as *nominal contrast* throughout this paper. For an in-depth overview about JPEG vs. PNG results, see Section C of this Appendix.



Figure 7. Three example stimuli for all conditions of contrast-experiment and noise-experiment. The three images (categories `bicycle`, `dog` and `keyboard`) were drawn randomly from the pool of images used in the experiments.



(a) Coherence parameter = 1.0

(b) Coherence parameter = 0.0

Figure 8. Three example stimuli (bicycle, dog, keyboard) for all except the coherence = 0.3 conditions of the eidolon-experiment, split by coherence levels.

A.3 Apparatus

All stimuli were presented on a VIEWPixed LCD monitor (VPixx Technologies, Saint-Bruno, Canada) in a dark chamber. The 22" monitor (484×302 mm) had a spatial resolution of 1920×1200 pixels at a refresh rate of 120 Hz. Stimuli were presented at the center of the screen with 256×256 pixels, corresponding, at a viewing distance of 123 cm, to 3×3 degrees of visual angle. A chin rest was used in order to keep the position of the head constant over the course of an experiment. Stimulus presentation and response recording were controlled using MATLAB (Release 2016a, The MathWorks, Inc., Natick, Massachusetts, United States) and the Psychophysics Toolbox extensions version 3.0.12 (Brainard, 1997; Kleiner et al., 2007) along with our in-house iShow library (<http://dx.doi.org/10.5281/zenodo.34217>) on a desktop computer (12 core CPU i7-3930K, AMD HD7970 graphics card “Tahiti” by AMD, Sunnyvale, California, United States) running Kubuntu 14.04 LTS. Responses were collected with a standard computer mouse.

B Fine-tuning on distortions

AlexNet, GoogLeNet and VGG-16 have not been designed for or trained on images with reduced contrast, added noise or other distortions. It is therefore natural to ask whether simple architectural modifications or fine-tuning of these networks can improve their robustness. Our preliminary experiments indicate that fine-tuning DNNs on specific test conditions can improve their performance on these conditions substantially, even surpassing human performance on noisy low-contrast images, for example. At the same time, fine-tuning on specific conditions does not seem to generalise well to other conditions (e.g. fine-tuning on uniform noise does not improve performance for salt-and-pepper noise), a finding consistent with results by Dodge and Karam (2017) who examined the impact of fine-tuning on noise and blur. This clearly indicates that it could be difficult to train a single network to reach human performance on all of the conditions tested here. A publication containing a detailed description and analysis of these experiments is in preparation. The question what kind of training would lead to robustness for arbitrary noise models remains open.

C JPEG vs. PNG

As mentioned in Section A.2, all experiments were performed using images saved in the JPEG format for compatibility with the image training database ImageNet (Russakovsky et al., 2015), which consists of JPEG images. That is, a certain image was read in, distorted as described earlier and then saved again as a JPEG image using the default settings of the `skimage.io.imsave` function. Since the lossy compression of JPEG may introduce artefacts, we here examine the difference in DNN results between saving to JPEG and to PNG, which is lossless up to rounding issues. Some results for the contrast-experiment using JPEG images were already reported by Wichmann et al. (2017).

The results of this comparison can be seen in Table 1. For all experiments but the contrast-experiment, there was hardly any difference between PNG and JPEG images. For the contrast-experiment, however, we found a systematic difference: all networks were better for PNG images. We therefore collected human data for

Table 1*Classification performance difference for PNG and JPEG images.*

Experiment	AlexNet	GoogLeNet	VGG-16	human avg.
colour-experiment	0.03% (0.03%)	-0.01% (0.01%)	0.02% (0.02%)	-
contrast-experiment	1.64% (1.64%)	3.25% (3.27%)	8.82% (8.84%)	2.68% (3.67%)
noise-experiment	0.03% (0.48%)	0.22% (0.71%)	0.45% (0.71%)	-
eidolon-experiment	-0.43% (1.08%)	0.03% (1.02%)	-0.34% (1.09%)	-

Notes. Each entry corresponds to the average performance difference for PNG minus JPEG performance for a certain network and experiment. The value in brackets indicates the average absolute difference. A value of 0.03% for AlexNet in the colour-experiment therefore indicates that AlexNet performance on PNG images was, in absolute terms, 0.03% higher compared to JPEG images (in this example: 90.58% vs. 90.61%). Human data ($n=3$) was collected for the contrast-experiment only.

this experiment employing PNG instead of JPEG images (reported in Figure 3). In this experiment, three of the original contrast-experiment’s observers participated, seeing the same images as in the first experiment⁹. The results are compared in Figure 10. Both human observers and DNNs were better for PNG images than for JPEG images, especially in the low-contrast regime. Especially VGG-16 benefits strongly from saving images to PNG (on average: 8.82 % better performance) and achieves better-than-human performance for 1% and 3% contrast stimuli. In the main paper, we therefore show the performance of humans and DNNs when the images are saved as in the PNG rather than the JPEG format to disentangle JPEG compression and low contrast.

The cause of this effect could most likely be attributed to JPEG compression artefacts for low-contrast images. Based on our JPEG vs. PNG examination, we draw the following conclusions: First of all, we recommend using a lossless image saving routine for future experiments even though networks may be trained on JPEG images, since performance, as our data indicate, will be either equal or better in both man and machine. Secondly, we showed that our results with JPEG images for the colour-, the noise- and the eidolon-experiment are not influenced by this issue, whereas the contrast-experiment’s results are to some degree.

⁹A time gap of approximately six months between both experiments should minimize memory effects; furthermore, human participants were not shown any feedback (correct / incorrect classification choice) during the experiments.

Table 2

Colour-experiment: difference between colour and grayscale conditions (paired-samples t-test).

Network / Observer	Difference (%)	95% CI (%)	t	df	p
AlexNet	7.72	[6.55, 8.90]	12.86	4479	<.001*
VGG-16	3.79	[2.97, 4.62]	8.99	4479	<.001*
GoogLeNet	2.90	[2.04, 3.76]	6.61	4479	<.001*
subject-01	0.47	[-2.91, 3.85]	0.27	639	.785
subject-02	1.25	[-2.33, 4.83]	0.69	639	.493
subject-03	3.91	[0.10, 7.72]	2.01	639	.045

Notes. * $p < .008\bar{3}$. Difference stands for colour minus grayscale performance; significant at $\alpha = \frac{5\%}{6} = .008\bar{3}$ after applying Bonferroni correction for multiple comparisons.

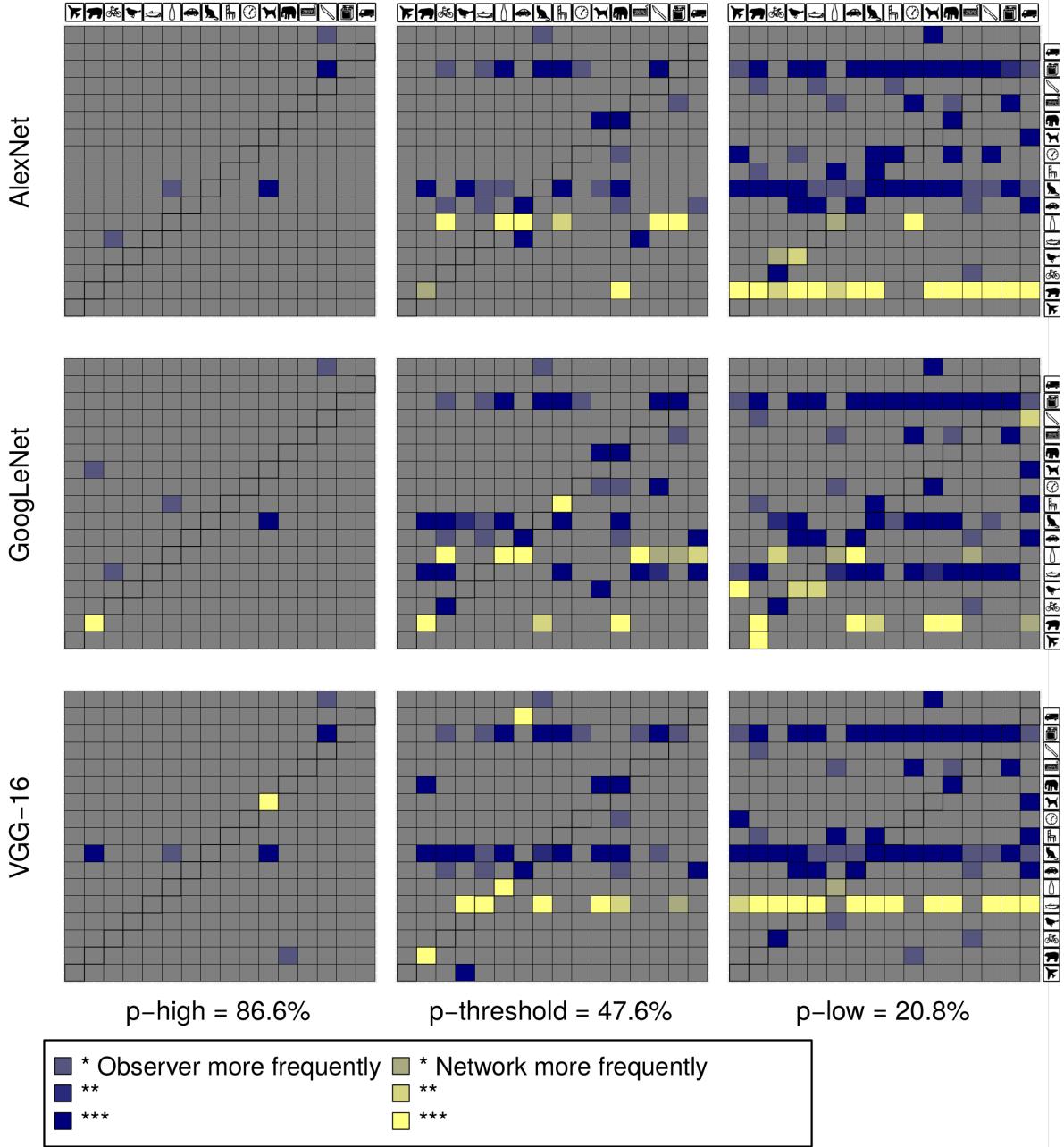


Figure 9. Confusion difference grid for contrast-experiment (as in Figure 6). Left column: for human observers the condition in which they performed best (100% contrast, performance $p\text{-high} = 86.6\%$) and for all networks the condition for which they achieved the same accuracy. Right column: data for low performance (20.8%); middle column: results for medium performance: 47.6%. The colour here shows whether a difference for a certain cell is significant at $\alpha = \frac{5\%}{16 \cdot 17 \cdot 9} (*)$, $\frac{1\%}{16 \cdot 17 \cdot 9} (**)$ and $\frac{0.1\%}{16 \cdot 17 \cdot 9} (***)$; these α -levels are Bonferroni corrected for multiple comparisons (16 categories \cdot 17 possible responses \cdot 9 confusion difference matrices). Data obtained with JPEG images.

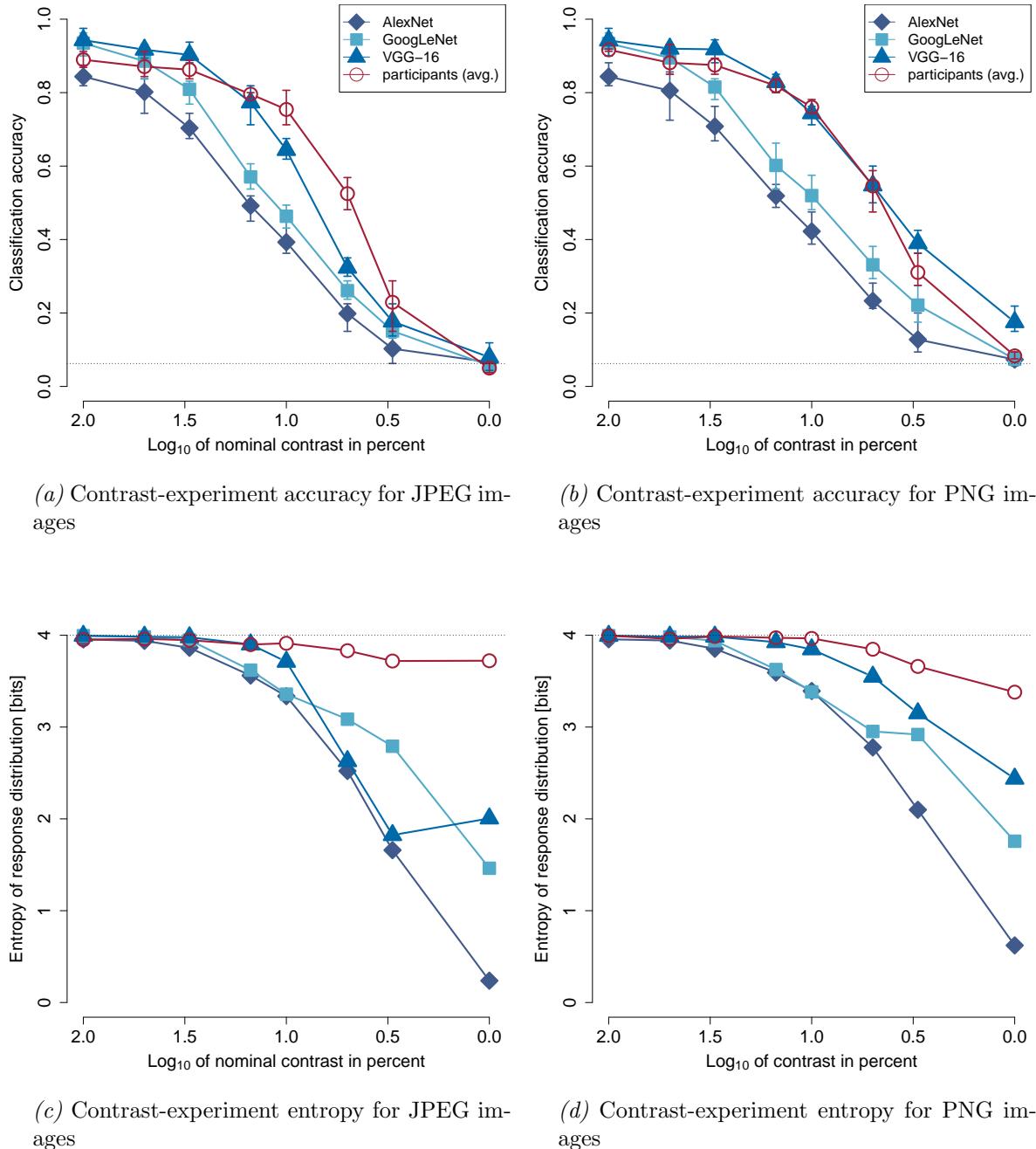
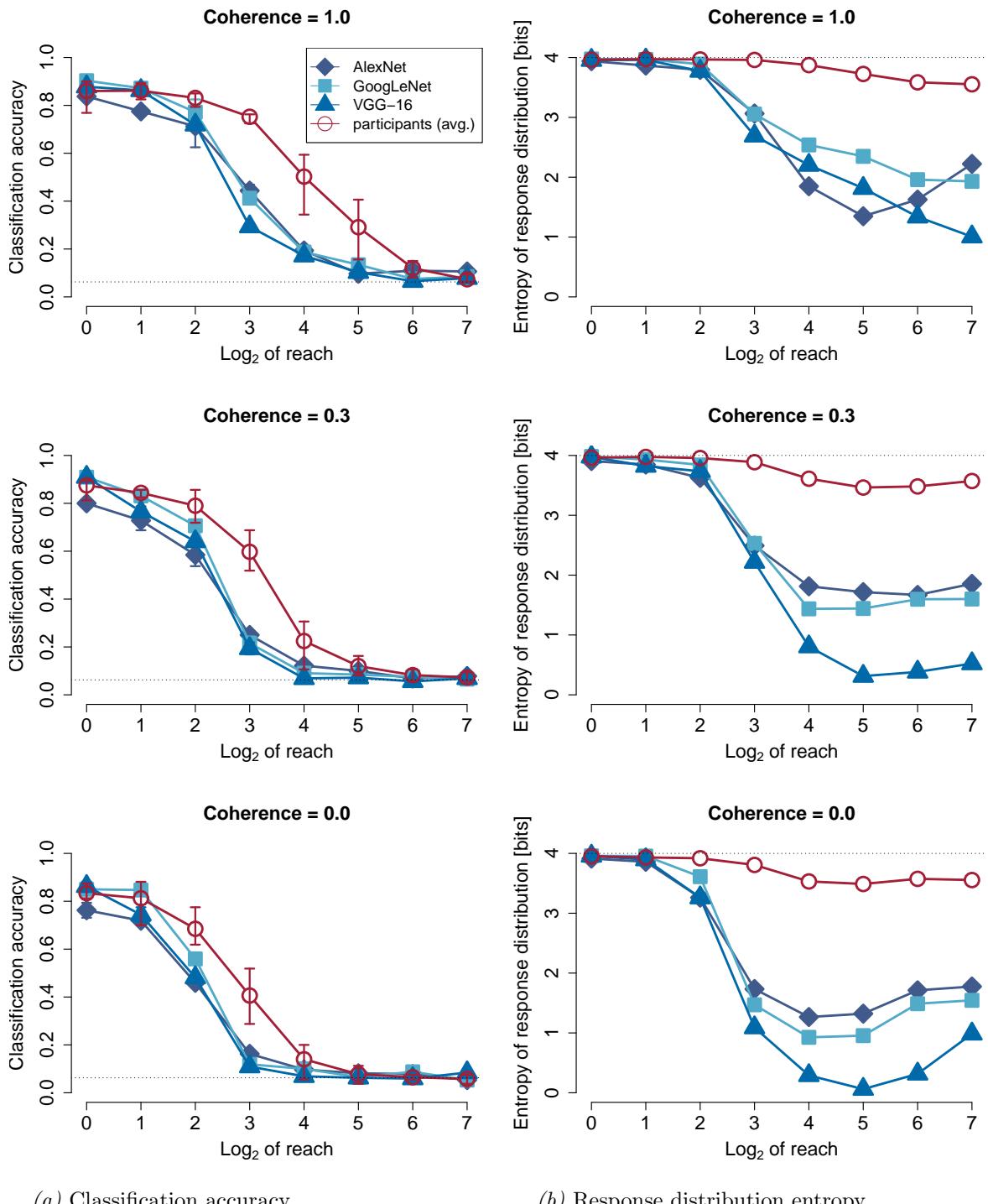


Figure 10. Accuracy and response distribution entropy for contrast-experiment, split by JPEG and PNG images. (a)(c) Results for JPEG images. N=3, the same three observers as for the PNG experiment. (b)(d) Results for PNG images.



(a) Classification accuracy

(b) Response distribution entropy

Figure 11. Complete eidolon-experiment results ($n=5$). **(a)** Accuracy including range. This range was obtained as for the other experiments. **(b)** Response distribution entropy. Note that the data for coherence = 1.0 are already visualized in Figure 3 (e) and (f), but are here shown again for better comparison to the results obtained for different coherence values.



(a) Contrast-experiment, JPEG images



(b) Contrast-experiment, PNG images



(c) Eidolon-experiment, coherence = 0.3



(d) Eidolon-experiment, coherence = 0.0

Figure 12. Estimated stimuli corresponding to 50% accuracy for contrast-experiments and eidolon-experiment (coherence parameter = 0.3 and 0.0). Top row: stimuli corresponding to threshold for the **average human observer**. Bottom three rows: stimuli corresponding to the same accuracy for **VGG-16** (second row), **GoogLeNet** (third row) and **AlexNet** (bottom row). The corresponding stimulus levels were calculated by assuming a linear relationship between the two closest data points measured in the experiments.