“From photos to sketches - how humans and deep neural networks process objects across different levels of visual abstraction”

Authors:

Johannes J.D. Singer[1,2*], Katja Seeliger[1], Tim C. Kietzmann[3], Martin N. Hebart[1*]

Affiliations:

[1]Vision and Computational Cognition Group, Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

[2]Department of Psychology, Ludwig Maximilian University, Munich, Germany

[3]Donders Institute for Brain, Cognition and Behavior, Nijmegen, The Netherlands

* Corresponding author

Email: johannes.singer@arcor.de or hebart@cbs.mpg.de

Funding information:

This work was supported by a Max Planck Research Group grant of the Max Planck Society awarded to MNH.

Commercial relationships disclosures:

The authors declare no conflicts of interest.

Abstract

Line drawings convey meaning with just a few strokes. Despite strong simplifications, humans can recognize objects depicted in such abstracted images without effort. To what degree do deep convolutional neural networks (CNNs) mirror this human ability to generalize to abstracted object images? While CNNs trained on natural images have been shown to exhibit poor classification performance on drawings, other work has demonstrated highly similar latent representations in the networks for abstracted and natural images. Here, we address these seemingly conflicting findings by analyzing the activation patterns of a CNN trained on natural images across a set of photos, drawings and sketches of the same objects and comparing them to human behavior. We find a highly similar representational structure across levels of visual abstraction in early and intermediate layers of the network. This similarity, however, does not translate to later stages in the network, resulting in low classification performance for drawings and sketches. We identified that texture bias in CNNs contributes to the dissimilar representational structure in late layers and the poor performance on drawings. Finally, by fine-tuning late network layers with object drawings, we show that performance can be largely restored, demonstrating the general utility of features learned on natural images in early and intermediate layers for the recognition of drawings. In conclusion, generalization to abstracted images such as drawings seems to be an emergent property of CNNs trained on natural images, which is, however, suppressed by domain-related biases that arise during later processing stages in the network.

Keywords:

deep convolutional neural networks, generalization, drawings, representational similarity analysis

Introduction

Humans have the remarkable ability to robustly recognize objects across a wide range of visual abstractions. One striking example for this feat is that we can identify objects in simple and abstract line drawings with similar speed and accuracy as natural object images (Biederman & Ju, 1988; Eitz et al., 2012). Neuroimaging studies have shown that this behavior is supported by a similar neural representation for natural images and line drawings across a number of visually responsive brain regions (Ishai et al., 1999; Walther et al., 2011). This suggests that both the recognition of natural images and line drawings of objects rely on a general purpose neural architecture that supports general object recognition. Recent developments in computer vision and computational neuroscience have opened up new avenues for studying the computational mechanisms of object recognition *in silico* (Kietzmann, McClure, et al., 2019), with deep convolutional neural networks (CNNs) as the most successful and most popular model class in approximating the activity elicited during object recognition in the inferior temporal cortex of non-human primates and humans (Cichy et al., 2016; Khaligh-Razavi & Kriegeskorte, 2014; Storrs et al., 2020; Yamins et al., 2014). While CNNs have yielded important insights into the mechanistic underpinnings of visual object recognition and show impressive performance in object classification benchmarks (Krizhevsky et al., 2012; Simonyan & Zisserman, 2015), there is a limited understanding of how well the recognition abilities of CNNs generalize to abstracted visual stimuli such as line drawings. Identifying the extent to which CNNs generalize to drawings could therefore not only deepen our understanding of the inner workings of CNNs themselves, but also create an important benchmark case for comparing CNNs and biological vision.

Recent studies investigating human-like generalization to drawings in CNNs have yielded conflicting results. On the one hand, there is evidence that CNNs trained on natural images cannot recognize abstract drawings and achieve poor classification performance with such images (Ballester & Araujo, 2016; Evans et al., 2021; Wang et al., 2019). This might be attributed to texture bias in CNNs, referring to the recent demonstration that CNNs tend to rely more strongly on texture rather than shape information for object recognition, which contrasts with humans who rely more strongly on shape (Geirhos et al., 2019; Hermann et al., 2020). Since the texture of drawings

is altered strongly in respect to photographs, CNNs trained on the latter may no longer recognize them without this diagnostic information. On the other hand, notable performance on colored and highly detailed drawings has been shown (Kubilius et al., 2016), and others have argued for a common representational format for drawings and photos in CNNs trained on natural images (Fan et al., 2018). This can be explained with findings showing that CNNs, too, accurately capture abstract shape information in their representational spaces (Kalfas et al., 2018; Kubilius et al., 2016). Since the overall shape of an object is largely preserved in drawings, based on these findings it might be expected that the representations and performance for drawings are similar to those found for natural photographs. Together, these seemingly disparate findings and their underlying explanations point to a gap in our understanding of the emergence of generalization to drawings in CNNs.

Here we address this gap by assessing the extent to which CNNs trained on natural images process object images similarly across different levels of visual abstraction. To this end, we composed a set of images of the same objects across three levels of visual abstraction: natural photographs, line drawings and sketches. In addition to analyzing the classification performance of the networks on these images, we compare the internal representations in the networks to each other and to human similarity judgments by using representational similarity analysis (Kriegeskorte et al., 2008).

Methods

Code

We made all code used for the present analyses publicly available (https://github.com/Singerjohannes/object_drawing_DNN). The analysis of the network's performance as well as feature extraction were implemented using Python 3.6.9 and the PyTorch 1.6.0 API (Paszke et al., 2019). Further analyses based upon the network activations and the human behavioral data were implemented using MATLAB R2017a (www.mathworks.com).

Stimuli

Our stimulus set comprised 42 object categories (21 manmade, 21 natural), each depicted across three different types of depiction: “photos”, “drawings”, and “sketches” (126 stimuli total; Fig 1). Drawings and sketches were drawn by a professional artist for the purpose of this study. For the first type of depiction (“photos”), object-images were natural photographs of objects, cropped from their background. For the second type of depiction (“drawings”), we gathered line drawings of the same objects. For these line drawings, color and texture information was strongly altered, while contour-level information was highly similar to the photos. For the third type of depiction (“sketches”), the images lacked even more detail as compared to the photos, with distortions in both the contours and the size of some features, rendering them physically even more dissimilar to the photos.

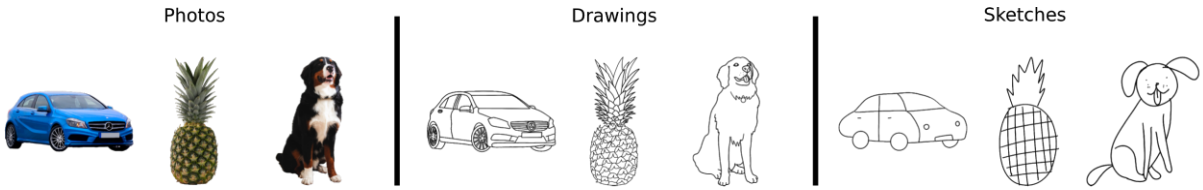


Figure 1. **Examples from the stimulus set used in all experiments.** We created a stimulus set of 126 stimuli originating from 42 object categories (21 manmade, 21 natural) depicted across three levels of visual abstraction. The three types contained the same objects as natural photographs cropped from their background (“photos”), line drawings (“drawings”), and sketch-like drawings (“sketches”).

Models

For all experiments, we used the VGG-16 architecture (Simonyan & Zisserman, 2015) which won the ILSVRC-challenge in the year 2014 with a top-5 accuracy of 92.7% on the test set. The network was chosen for its relative simplicity compared to deeper and more complex network architectures and for its documented similarities to visual processing in the human brain (Güçlü & Gerven, 2015; Jozwik et al., 2017; Kalfas et al., 2018; Nonaka et al., 2020; Schrimpf et al., 2020; Storrs et al., 2020). VGG-16 is a 16-layer network which contains five convolutional blocks, each followed by a max-pooling layer. After the convolutional stage of the network, there are three fully connected layers, with the last layer being the softmax layer which outputs the probabilities for the 1,000-way classification task. For Experiment 1, we used the

standard implementation of VGG-16 pretrained on the ILSVRC2012 dataset (Russakovsky et al., 2015) without batch normalization. For simplicity, in the following we will refer to networks trained with the ILSVRC2012 dataset as “ImageNet-trained”. For Experiment 2, we obtained the version of VGG-16 which was trained on stylized ImageNet - a modified version of the ImageNet dataset (https://github.com/rgeirhos/Stylized-ImageNet). Stylized ImageNet was created using image style transfer (Gatys et al., 2016), with the goal of making shape a more diagnostic feature than texture in the training data (Geirhos et al., 2019). For Experiment 3, we used the same standard VGG-16 implementation as in Experiment 1 but carried out fine-tuning on the weights of the layers (for details, see below).

Image preprocessing

Before passing the images through the network, they were scaled and placed on a square grey background. Further the images were resized to the input size of 224 × 224 and normalized (zero mean, unit variance).

Performance evaluation

In order to quantify the performance of the networks in our experiments, we computed the top-1 accuracy based on the outputs of the softmax layer of the models. We mapped the categories in our stimulus set to the ImageNet categories by using the WordNet 3.0 hierarchy (Miller, 1995). A response was counted as correct if the highest scoring label matched the given category label or any of its hyponyms in the WordNet 3.0 hierarchy (Miller, 1995). For example, for the category “dog”, each response was counted as correct that referred to any of the labels in the ImageNet classes that shared the hypernym “domestic dog” in the WordNet 3.0 hierarchy (Miller, 1995). While this approach may slightly overestimate the networks’ overall classification performance, it was held constant for all relevant comparisons and thus should not affect the overall interpretation of results.

Feature extraction

We extracted the activation patterns for all stimuli only from the five pooling layers and the first two fully connected layers but excluded the softmax layer from our analyses. For comparing activations in pooling layers, we flattened the extracted feature map activations.

Fine-tuning

The VGG-16 model was fine-tuned using the ImageNet-Sketch data set (Wang et al., 2019). ImageNet-Sketch consists of approximately 50 training examples of drawings for each of the 1,000 ImageNet classes, resulting in approximately 51,000 training images of drawings. It is important to note that ImageNet-Sketch is orders of magnitude smaller than the ILSVRC2012 data set (1.4 million examples from ImageNet (Deng et al., 2009)) that has been used for training the original VGG-16 network. For the fine-tuning procedure, layers conv1-1 to conv4-3 were frozen, and only the last convolutional layers conv5-1 to conv5-3 and the fully-connected layers were adjusted. The model weights of those layers were then trained using the stochastic gradient descent (SGD) optimizer. A hyperparameter search for learning rate and momentum was performed. With the optimized learning rate of 0.001 and momentum of 0.7, a top-1 accuracy of 66.35% was reached on a validation set of 2,000 images.

Behavioral tasks

All human behavioral data was acquired in online experiments using the Amazon Mechanical Turk Platform (Mturk). All participants were located in the USA. In the labelling task, 480 individuals (280 female, 196 male, 4 other) took part, with a mean age of 37.96 years (*SD*=12.45). In the triplet task 742 (404 female, 334 male, 4 other) participants took part, with a mean age of 38.29 years (*SD*=12.54). The experimental protocols were approved by the local ethics committee (012/20-ek) in accordance with the declaration of Helsinki, and all participants provided informed consent.

Labelling task

To measure the accuracy of humans in recognizing our stimuli, we conducted an image labelling experiment. In this task, workers on Mturk were instructed to give a label for a given image consisting of one word, as if they were trying to find the corresponding entry for the given image in a dictionary. Labels were regarded correct if the label matched the category-label of the object exactly or if the label was a synonym for the given category label. In order to avoid carry-over effects between types of depiction, all trials in a given task were limited to one depiction type, and a given participant could only participate in one task.

Triplet Task

To measure the perceived similarity of objects in our image sets, we conducted a triplet odd-one-out similarity task (Hebart et al., 2020). For each given trial, participants were presented with three different object images side by side. Participants were instructed to indicate which of the three object images in the trial they thought is the odd-one out by clicking on the respective image using their computer mouse or touchpad. There were no instructions as to what strategy they should use to find the odd-one out. Additionally, if they did not recognize the object, participants were instructed to base their judgment on their best guess. Similar to the labelling task, images in one task were limited to one type of depiction, and participants could only work on one task. We subsequently computed the object similarity between a given object pair for one type of depiction in our stimulus set as the probability of participants choosing objects x and y to belong together, irrespective of the context imposed by the third object z in the triplet.

Quantifying representational similarity across types of depiction

We used representational similarity analysis (RSA, (Kriegeskorte et al., 2008)) to quantitatively describe the representational object space in the CNNs in our experiments as well as in human behavioral measurements. RSA characterises representations by estimating the geometry of stimulus responses in high-dimensional population space. For this, the similarity of all pairwise comparisons of conditions in the experimental design is computed by comparing responses of a given system (e.g. activation patterns extracted from a CNN) for every pair of conditions and arranging them in a matrix format. The matrix of dissimilarities is commonly referred to as a representational dissimilarity matrix (RDM). The power of RSA lies in the fact that RDMs can be computed and subsequently compared across different systems (e.g. computational models and humans) but also across different conditions and processing stages in a neural network (Mehrer et al., 2020). As our stimulus set contained the same object categories for each type of depiction, we could directly compare the RDMs across different types of depiction by correlating the RDMs with each other. We computed RDMs based on the activation patterns that we extracted from the networks separately for each type of depiction and for each selected layer. We then calculated the Spearman rank correlation between the RDM's lower triangular

parts to acquire a measure of similarity across types of depiction in terms of the network's internal representation (Fig 2).

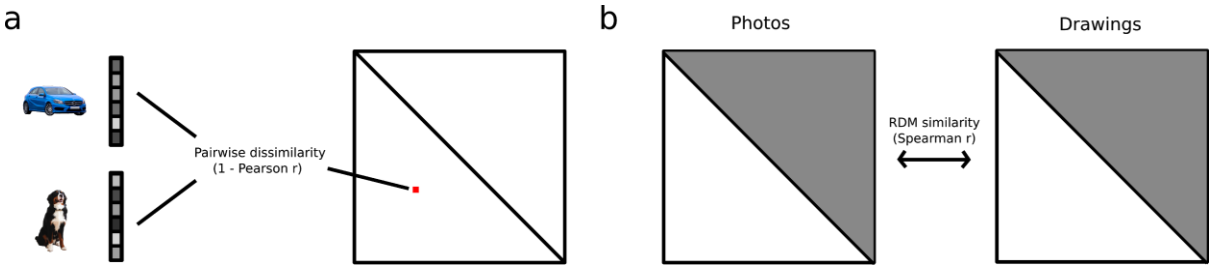


Figure 2. Schematic of the RSA approach used in all experiments. a) For analyzing the internal representations of the networks, we extracted the activation patterns to the stimuli in our set from a selection of layers. Subsequently, we computed RDMs based on the correlation distances (1- Pearson *r*) between all pairs of stimuli in one type of depiction and for each given layer separately. This yielded three RDMs (one for each type of depiction) in every selected layer. **b)** In order to obtain a measure of similarity across types of depiction regarding the internal representation in the networks, we computed the Spearman rank correlation between the lower triangular values of RDMs of the different types of depiction in each layer separately. For the RDMs comparing representations of photos and drawings, this yields a "photo-to-drawing" similarity.

Super-RDMs and multidimensional scaling

For comparisons of the representational geometries of all three types of depiction combined, we constructed super-RDMs based on the activation patterns of the responses of the network to all stimuli in the set following the RSA procedure described above. For visualization purposes, we projected the RDMs into a two-dimensional space using multidimensional scaling (metric stress).

Classification analysis

Intra-domain classification

To quantify the information contained in the activation patterns of the networks in a given layer, we performed a classification analysis on the extracted activation patterns. To this end, we trained a linear Support Vector Machine classifier (Chang & Lin, 2011) on the activation patterns to distinguish between patterns of manmade and natural objects (c=1). We followed a leave-*N*-out cross-validation framework (*N*=6), meaning

that we trained the classifier on all but a left-out sample of six activation patterns and evaluated the accuracy of the classifier's prediction on the left-out sample. We repeated this procedure 1,000 times for a given layer while randomly drawing training and test examples from all activation patterns according to the framework described above. Finally, we averaged the accuracies over all 1,000 iterations, yielding mean classification accuracies and standard errors for each layer and each type of depiction in the network.

Cross-domain classification

In an additional analysis step, we trained the classifier following the same procedure as described above but used the activation patterns of one type of depiction (e.g. photos) and tested the classifier on the activations patterns of another type of depiction (e.g. drawings).

Statistical analyses

To statistically evaluate differences in classification performance between networks or between types of depiction in one network, we used the McNemar test of homogeneity (McNemar, 1947). Furthermore, to test for differences between levels of human performance, we used a two-sided independent *t*-test. For testing the equivalence of the means in human behavioral accuracies, we followed a two one-sided test procedure as described in (Lakens, 2017). For all other comparisons, we used a one-sided randomization testing procedure. This procedure entailed computing a null distribution by repeated randomization and subsequently obtaining the *p*-value by finding the percentile of values in the null distribution that reaches or exceeds the empirical value. For comparing human performance with CNN performance, we obtained the null distribution using a sign permutation procedure. To do this, we permuted the signs of human accuracies for single objects for a given type of depiction, averaged these permuted labels, and repeated this procedure 1,000 times. To test for the statistical significance of a given RDM-correlation, we randomly shuffled object labels for one RDM, computed the correlation with the reference RDM, and repeated this procedure 1,000 times to obtain a null distribution (Mantel test (Mantel, 1967)). To test for changes in representational similarity across layers (e.g. for "photos" vs. "drawings"), we randomly shuffled object labels, using the same shuffling across layers but a different shuffling for different types of depiction. Then we created permuted

RDMs and subsequently computed the sum of squared differences between the overall mean in RDM correlations across layers and the individual RDM correlations. We then repeated this procedure 1,000 times to obtain a null distribution of sum of squared differences. Finally, in order to test if a given classification accuracy significantly exceeded chance level, we obtained the null distribution by repeating the classification analysis 1,000 times with randomly shuffled labels.

Correction for multiple testing

For all statistical tests reported here, we corrected the *p*-values for multiple comparisons by using the Benjamini-Hochberg false discovery rate correction (Benjamini & Hochberg, 1995).

Results

Experiment 1: Generalization to drawings in an ImageNet-trained CNN

The overall aim of this study was to quantify how similarly CNNs process natural images and drawings of objects. This was accomplished by comparing their performance, their internal representations, and by comparing them to human behavior.

Similarity in terms of classification performance

As a first step, we sought to test how an ImageNet-trained CNN, specifically VGG-16 (Simonyan & Zisserman, 2015), generalizes to drawings and sketches in terms of classification performance. For this purpose, we passed the images through the network and compared predicted class labels with the actual object categories. We found that VGG-16 exhibited excellent classification accuracy for photos but very low accuracies for drawings and sketches (Fig 3). Statistical comparisons of the accuracies revealed that drawing performance was significantly lower than photo performance (M(Photos)=0.79, M(Drawings)=0.14, $\chi^2(1)=33.44$, $p<0.001$, FDR-corrected). Sketch performance was also significantly lower than photo performance (M(Sketches)=0.02, $\chi^2(1)=47.25$, $p<0.001$, FDR-corrected) but there was no significant difference between the network's performance on drawings and sketches ($\chi^2(1)=0.96$, $p=0.33$, FDR-corrected).

To compare the network's performance to humans, we measured human classification accuracy on the same images in a separate online labelling experiment. Unlike VGG-16, humans performed very well and virtually identical on photos and drawings (*M*(Photos)=0.97, *M*(Drawings)=0.96, *t*(41)=-1.93 , *p*=0.03, two one-sided test of equivalence) and showed only a slight but significant decrease in performance for sketches (*M*(Sketches)=0.9, Photos vs. Sketches - *t*(82)=3.12, *p*=0.01, FDR-corrected; Drawings vs. Sketches - *t*(82)=2.67, *p*=0.01, FDR-corrected). The network performed significantly worse across all types of depiction as compared to humans (all *p*<0.001, FDR-corrected, one-sided randomization test). These results demonstrate that VGG-16 pretrained on ImageNet fails to replicate human generalization to drawings and sketches in terms of classification performance.

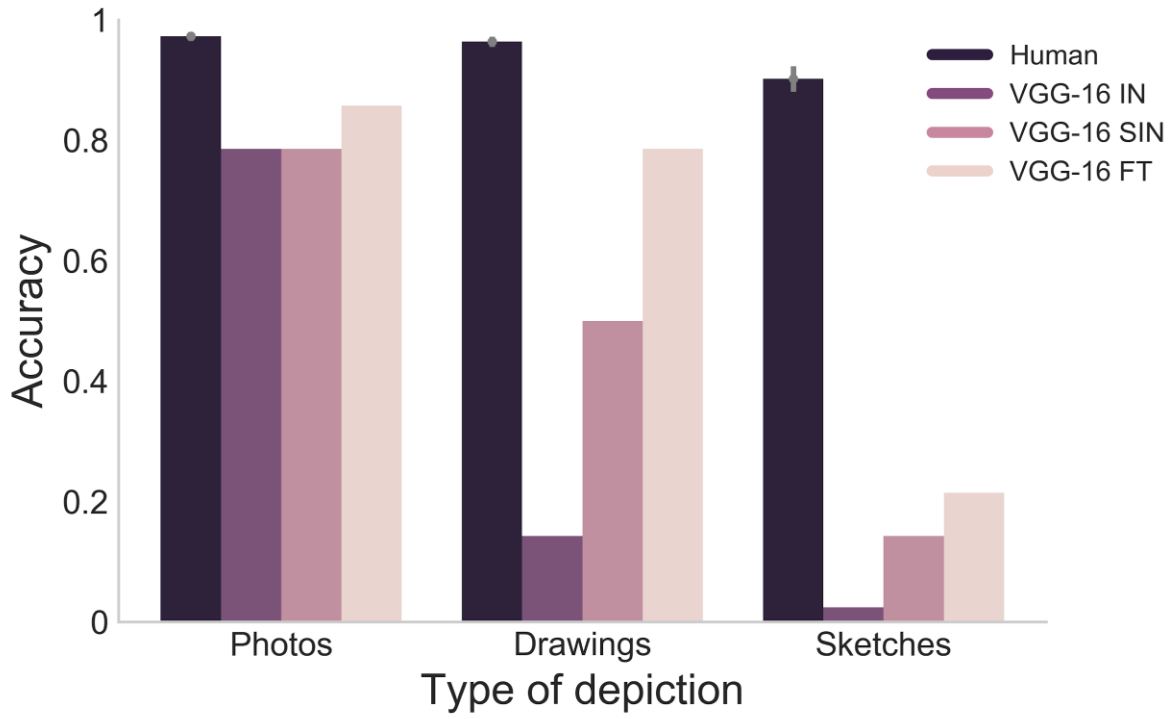


Figure 3. Classification performance in variants of VGG-16 in comparison to humans. Human accuracies measured in a separate online experiment revealed that humans performed very well and virtually identical on photos and drawings and slightly worse on sketches. The ImageNet-trained VGG-16 (VGG-16 IN) showed high performance on photos but a sharp drop in performance for drawings and sketches. VGG-16 trained on stylized ImageNet (VGG-16 SIN) performed better on drawings than VGG-16 IN but not on sketches or photos. VGG-16 fine-tuned on ImageNet-

Sketch (VGG-16 FT) performed equally well on photos and drawings but still performed poorly on sketches.

Similarity in terms of internal representations

To compare VGG-16 not only in terms of performance but also regarding its internal representations of photos, drawings and sketches, we used RSA (Kriegeskorte et al., 2008). For a given type of depiction, we calculated the pairwise dissimilarities (1 - Pearson *r*) for all pairs of object activation patterns and stored them in a representational dissimilarity matrix (RDM). This procedure yielded three RDMs (one for each type of depiction) for each of the layers in the network. Finally, we computed the Spearman rank correlation between the lower triangular part of RDMs of different types of depiction. This yielded three sets of comparisons: photo-to-drawing similarity, photo-to-sketch similarity, and drawing-to-sketch similarity. We assumed that if the network processed objects images at different levels of visual abstraction similarly, this would be reflected in a similar representational format across types of depiction.

Comparing the degree of representational similarity between photos, drawings, and sketches and how it changes across layers, we found overall differences in RDM correlations across layers for all three comparisons (all $p<0.001$, one-sided randomization test). We followed up on these overall differences with tests between pairs of layers for one type of depiction. For the photo-to-drawing similarity, we found a significant increase in correlation from pooling layer 1 to pooling layer 4 ($p=0.002$, FDR-corrected, one-sided randomization test) and after that a significant drop in similarity from pooling layer 4 to the fully connected layer 2 ($p=0.002$, FDR-corrected, one-sided randomization test) (Fig 4a). While there were overall relatively lower correlation values for the photo-to-sketch correlation (all $p<0.003$, FDR-corrected, one-sided randomization test), the shape of the results was similar, with an increase in similarity from pooling layer 1 to pooling layer 4 ($p=0.005$, FDR-corrected, one-sided randomization test) and a drop in correlation for higher layers which approached zero in the penultimate layer ($p=0.003$, FDR-corrected, one-sided randomization test). These results demonstrate a representational format which is shared between photos and drawings at intermediate layers and a weaker but stable similarity to sketches.

Interestingly, a different pattern was observed for the drawing-to-sketch similarity. Here, we found an increase in correlation from pooling layer 1 to pooling layer 4 (p=0.003, FDR-corrected, one-sided randomization test) and no significant difference in correlation between pooling layer 4 and the fully connected layer 2 (p=0.3, FDR-corrected, one-sided-randomization test).

Overall, the results reveal an increase in representational similarity across levels of visual abstraction that peaks in intermediate layers. This pattern of results diverged after pooling layer 4, with photo representations becoming less similar to drawings and sketches, while drawing and sketch representations stay at a highly similar level.

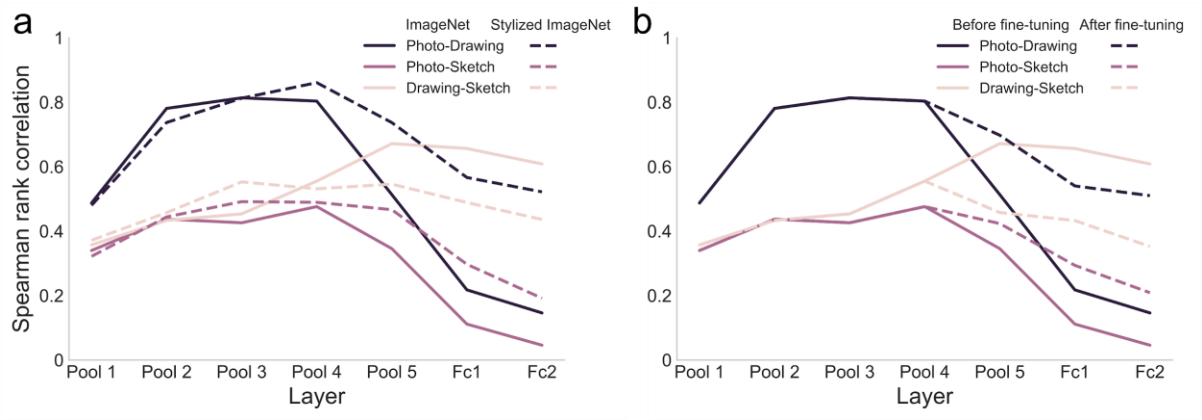


Figure 4. Similarity in representational structure between types of depiction in variants of VGG-16. a) Spearman rank correlations between RDMs for the different types of depiction in the ImageNet-trained VGG-16 (VGG-16 IN) and in the VGG-16 trained on stylized ImageNet (VGG-16 SIN). In both networks, we observed a high degree of representational similarity between photos and drawings and to a lesser extent also between photos and sketches in early and intermediate layers. In the later layers, these similarities between photos and abstracted types of depiction dropped sharply in VGG-16 IN while in VGG-16 SIN this drop was attenuated. **b)** Spearman rank correlations between RDMs for the different types of depiction in the ImageNet-trained VGG-16 before and after fine-tuning. After fine-tuning, the similarity in the representational format was increased for the photo-to-drawing and photo-to-sketch similarity, but reduced for the drawing-to-sketch-similarity, indicating increased similarity in processing between photos and abstracted types of depiction in the network after fine-tuning.

Clustering of drawing and sketch representations

The sharp drop in representational similarity for the photo-to-drawing and photo-to-sketch comparison suggests a strong shift in the representational structure after pooling layer 4. To illustrate the source of this drop, we visualized all representational similarities using multidimensional scaling. To this end, for a given layer, we first constructed a super-RDM composed of the pairwise distances between all pairs of stimuli in the set (see Appendix A1). Subsequently, we used metric multidimensional scaling to project the representational dissimilarities into a two-dimensional space. We then repeated this entire procedure for each layer of the selected layers of VGG-16.

In these 2-D projections, we found that the activations for all types of depiction appeared evenly distributed in the first two pooling layers. However, starting already at pooling layer 3, activations for drawings and sketches started to cluster into one combined cluster, and the degree of clustering showed an increase up to the penultimate layer (Fig 5).

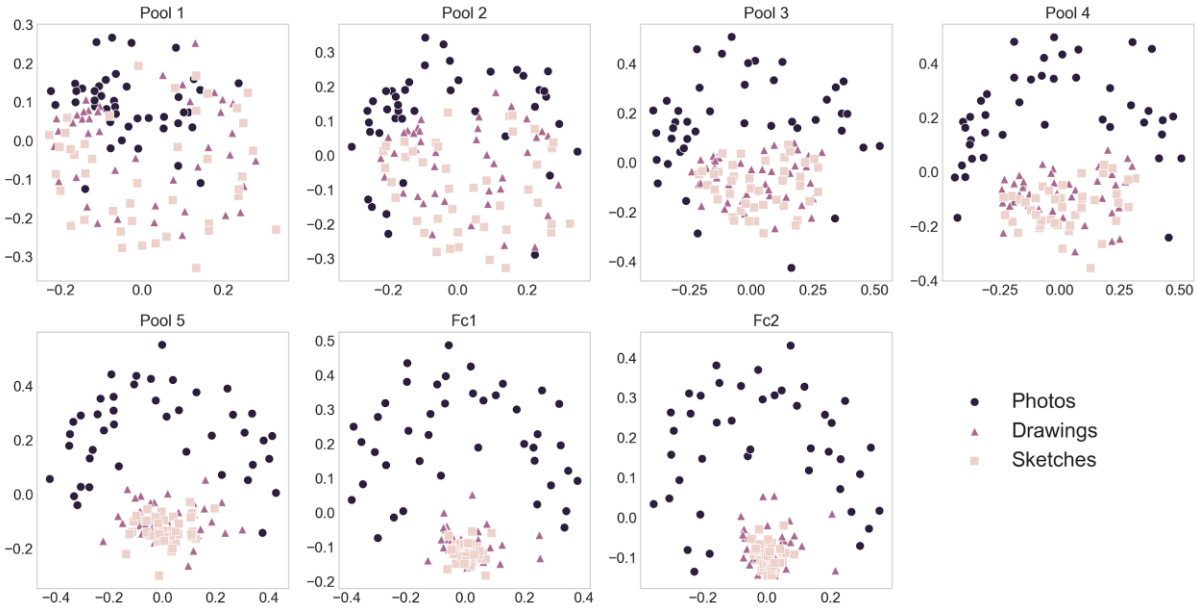


Figure 5. Clustering of abstracted object representations in the ImageNet-trained VGG-16. Multidimensional scaling plots of the super-RDMs reflecting all pairwise dissimilarities between stimulus representations in the stimulus set. Across layers, drawing and sketch representations increasingly clustered together, while photo representations remained well separated.

No collapse of representations for drawings and sketches

At first, this strong clustering might indicate a “representational collapse” of drawing and sketch representations towards specific points in the representational space. This might be expected, given the dramatic difference in image statistics between natural images the network had been trained on and drawings and sketches. A strong interpretation of this account would predict that all images of drawings and sketches are treated as largely equal by the network, possibly all leading to the same incorrect predicted category, and with differences in representations between images only reflecting meaningless residual noise. However, the fact that representational similarities between drawings and sketches remain high indicates that there is still some representational structure shared between these types of depiction. What is left open is to what extent these representations are idiosyncratic but similar, due to the similar appearance of drawings and sketches, or whether the representation still carries meaningful information, for example about the high-level category of an object.

To test for the representational content of the drawing and sketch representations, we used a Support Vector Machine classification approach to predict an object’s high-level category (manmade/natural) from the network’s activation pattern, separately for each layer and depiction. For photos, we found significant above-chance classification accuracies from pooling layer 3 onwards (all $p<0.01$, FDR-corrected, one-sided randomization test). A similar effect was observed for drawings from pooling layer 4 (all $p<0.01$, FDR-corrected, one-sided randomization test), and for sketches starting from pooling layer 2 (all $p<0.05$, FDR-corrected, one-sided randomization test) (Fig 6a). These results indicate that despite the strong clustering observed in the RDMs as visualized by the MDS plots, category-relevant information was retained in the activation patterns of drawings and sketches in the later layers. These results argue against a strong representational collapse (as described above) for drawings and sketches. Instead, they suggest a representational shift, indicating that the relative representational structure of objects is not destroyed and still allows meaningful readout of category information.

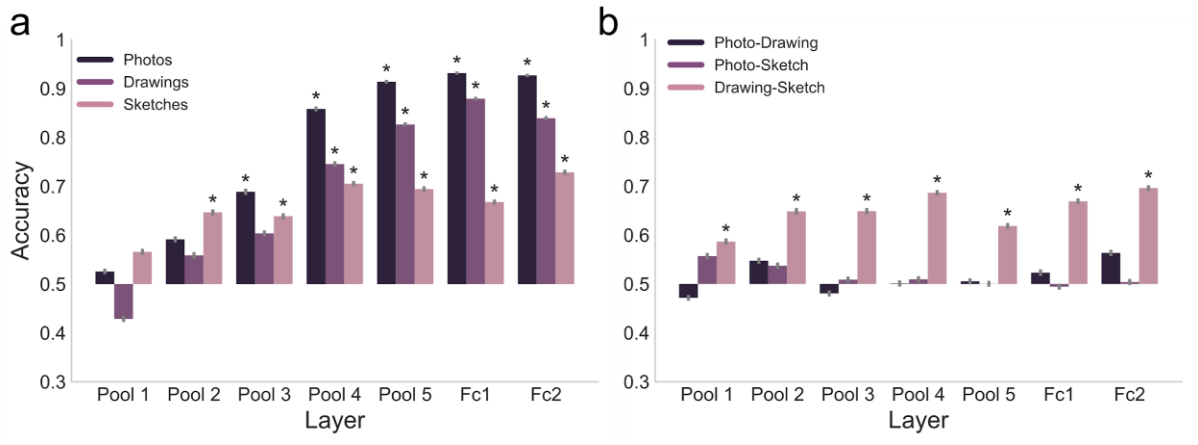


Figure 6. Classification of high-level category information based on feature representations across layers in VGG-16. a) Classification accuracies in VGG-16 IN. Based on the features extracted from single layers in VGG-16 IN high-level category information (manmade/natural classification) could be decoded with above-chance accuracy even for drawings and sketches, particularly in the later layers. **b)** Cross-classification accuracies in VGG-16 IN. Generalization of the classifier trained on one depiction (e.g. photos) to another depiction (e.g. drawings) showed above-chance accuracy only for the case where we trained on drawings and tested on sketches.

Generalizable high-level information for drawings and sketches

The classification results indicate that category information is still present in drawing and sketch representations but leaves open to what degree this information generalizes between types of depiction. To test this, we carried out cross-classification, training a Support Vector Machine classifier on the manmade/natural distinction using activation patterns from one type of depiction, and testing it on another type. We found significant above-chance cross-classification accuracies only for the case in which we trained on drawings and tested on sketches (all $p<0.04$, FDR-corrected, one-sided randomization test), however, not for the cases where we trained on photos and tested on drawings or tested on sketches (all $p>0.05$, FDR-corrected, one-sided randomization test) (Fig 6b) This suggests that the information contained in the representation only generalizes between drawings and sketches but not between photos and any of the other types of depiction. This is consistent with the idea that the representations of drawings and sketches do not collapse but are shifted and thus

retain some meaningful information, which however exhibits a different format than the information for photos.

Fit to human behavior

The classification of a single high-level object category might be seen as insufficient evidence of intact representations. For that reason, we obtained human behavioral similarity measurements to quantify how the objects in our stimulus set are represented in humans (see Appendix A2 for visualization) and to compare these human representations to the network's representations. Assuming that there was a collapse in the representations for drawings and sketches that increases across layers, we would expect to see a drop in representational similarity between VGG-16 and human behavior, both for drawings and sketches, specifically in the late layers. Alternatively, in the case of a shift of representations, we would expect smaller representational similarities for drawings and sketches than for photos but still significant representational similarities for these types of depiction in later layers. This would in turn indicate that there is some meaningful structure preserved in the representation for drawings and sketches, which is similar to human behavior. To examine this, we correlated the RDMs that we obtained for the layers of VGG-16 separately for every layer and type of depiction with the corresponding RDM that we obtained from human similarity judgments.

Perhaps not surprisingly, we found the highest fit of network representations to human behavior for photos (Fig 7), showing an increase in RDM correlation up to pooling layer 5 after which the correlation dropped slightly in the penultimate layer of the network. We observed a similar pattern for the drawings and sketches with an increase in RDM correlation up to pooling layer 5, after which the correlation stabilized in the fully connected layers. Statistical tests confirmed that there were significant correlations in all layers for photos and sketches (all $p<0.02$, FDR-corrected, one-sided Mantel test) and starting at pooling layer 2 for drawings (all $p<0.05$, FDR-corrected, one-sided Mantel test).

Comparing the RDM correlations against each other revealed that the RDM correlation for photos was significantly higher than for drawings in all layers (all $p<0.003$, FDR-corrected, one-sided randomization test) and for sketches (all $p<0.04$) in all layers but

pooling layer 1 (p=0.06) and pooling layer 2 (p=0.12). The correlations for drawings and sketches were only significantly different in the fully connected layers (both p<0.05) but not in all the other layers (all p>0.05).

In conclusion, we found the best fit of network representations to human behavior for representations of photographs. Importantly, however, we found a significant fit to human behavior for drawings and sketches, which increased across layers and remained high until the last layer. This is in line with the idea of a shift in representations of abstracted object images. While the representations are shifted, they are not completely distorted and retain category-relevant information of the objects and show similarity to human representations.

Taken together, the results of Experiment 1 suggest fairly general object representations in intermediate layers, which show generalization across different levels of visual abstraction. Yet, in the process of linking these representations to object categories, they are presumably biased towards the object features in natural images, leading to dissimilar representations in later layers and low network classification accuracies for drawings and sketches. This would indicate that the issue of incorrect classification is not an issue of representation as such, but more an issue of read-out into object classification.

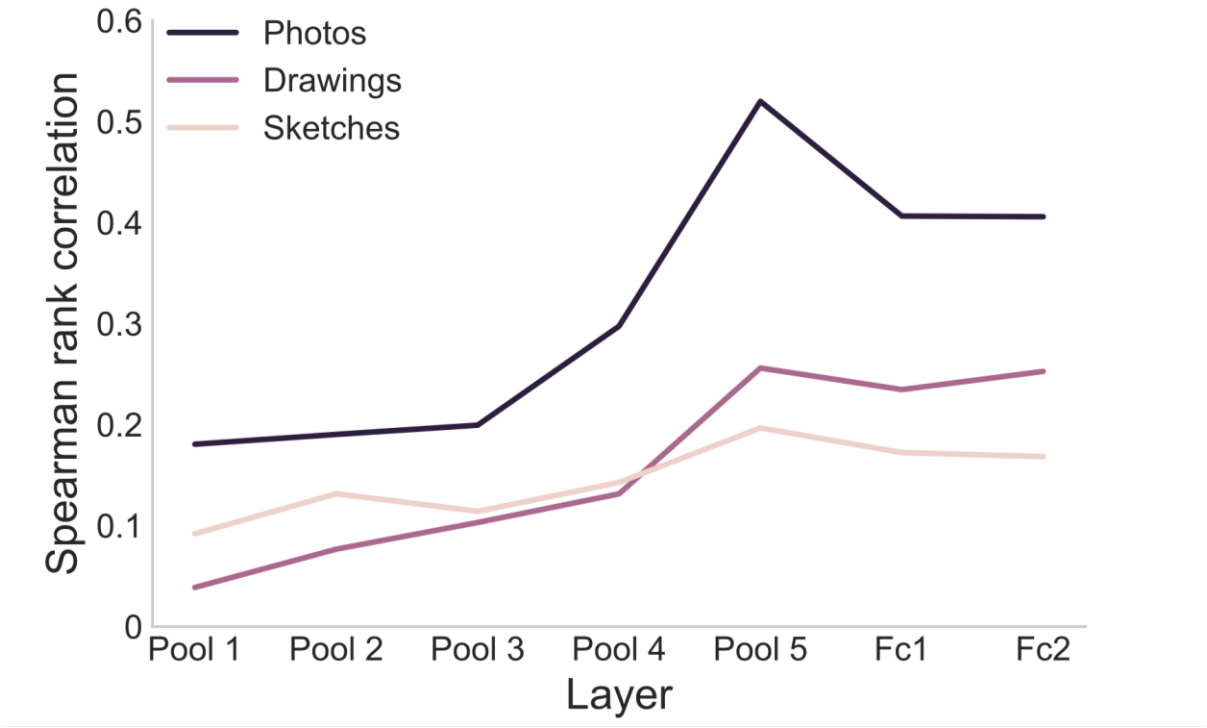


Figure 7. Fit to human behavior in VGG-16. a) Spearman rank correlations between the RDMs obtained from the ImageNet-trained VGG-16 (VGG-16 IN) for every layer in the network and the RDMs obtained in an online behavioral experiment, for each type of depiction separately. We observed the best fit to human behavior for the representations of photos. Although lower, we found a significant fit to human behavior for the drawing and sketch representations.

Experiment 2: Generalization to drawings in a shape-biased CNN

In Experiment 1 we found that the ImageNet-trained VGG-16 categorizes drawings and sketches poorly. However, in terms of its internal representations there seems to be considerable similarity between photos and drawings, and to some extent between photos and sketches. Despite this similarity in intermediate layers, as image processing approached the classification stage, the network represented drawings and sketches very differently to photos. Yet, the CNN representations for drawings and sketches retained category-relevant information about the objects and showed similar representations to those derived from human similarity judgements. While these results highlight the representational nature of object drawings and sketches in VGG-16, they leave open the critical question of why the representations shift in later layers, which seems to cause misclassifications of the network.

One reason for the dissimilar representation of natural images and abstracted images in the late stages of processing and the poor performance for abstracted images might be that ImageNet-trained CNNs rely heavily on local texture information for their classifications (Baker et al., 2018; Geirhos et al., 2019). Since texture information is strongly altered in both drawings and sketches, a reliance on texture information for the classification might favor a dissimilar representation of photos and abstracted images. In order to reduce the bias for texture in CNNs, Geirhos et al. (2019) introduced a new way of training CNNs, forcing them to rely more on shape information than texture. By training CNNs on a stylized version of the ImageNet dataset (Gatys et al., 2016), they revealed a preference of shape over texture in the classification behavior of the network. Assuming that, in Experiment 1, texture bias indeed contributed to the mismatch between the representation of photos and abstracted object images in later layers, this leads to two predictions. First, we would expect that for a CNN trained on stylized images, the drop in representational similarity between photos and abstracted types of depiction in later layers is attenuated, since they now share more similar overall shape information in the network's representation. Second, we would expect improved classification performance on abstracted object images. To test the role of the texture bias in the representation of drawings in CNNs, we analyzed the processing for VGG-16 which was trained on stylized ImageNet analogous to Experiment 1.

Better performance on drawings in a shape-biased CNN

Again, as a first step we focussed on the performance of the network in terms of predicting the object category from the images, separately for photos, drawings, and sketches. We found that the network trained on stylized ImageNet (VGG-16 SIN) compared to the ImageNet-trained variant (VGG-16 IN), performed similarly well on photos (M(Photos)=0.79, $\chi^2(1)=0.01$, p=0.91, FDR-corrected), however, showed much improved performance on drawings (M(Drawings)=0.5, $\chi^2(1)=10.01$, p=0.005, FDR-corrected), with performance on sketches remaining poor (M(Sketches)=0.14, $\chi^2(1)=0.96$, p=0.49, FDR-corrected) (Fig. 3). Despite these improvements, performance in VGG-16 SIN on drawings was still worse than on photos ($\chi^2(1)=6.30$, p=0.12, FDR-corrected) and significantly worse on sketches than on drawings ($\chi^2(1)=10.01$, p=0.002, FDR-corrected). Similar to Experiment 1 the direct comparison

of VGG-16 SIN performance to humans revealed that the network performed significantly worse in all types of depiction than humans (all $p<0.001$, FDR-corrected, one-sided randomization test).

In conclusion, VGG-16 SIN performed better on drawings than VGG-16 IN, however, the network still performed worse on drawings than on photos and did not perform better on sketches than VGG-16 IN.

Attenuated drop in representational similarity across levels of visual abstraction

Next, we quantified the similarity in processing between types of depiction in VGG-16 SIN in terms of the network's internal representations analogous to Experiment 1. We hypothesized that we would see a particular attenuation of the drop in representational similarity towards the late layers compared to what we observed in Experiment 1.

In comparison to VGG-16 IN, photo-to-drawing correlations in the early layers in VGG-16 SIN were either not statistically different from the ones in VGG-16 IN (pooling layer 1: $p=1$, FDR-corrected, one-sided randomization test; pooling layer 3: $p=1$, FDR-corrected, one-sided randomization test) or higher in VGG-16 IN (pooling layer 2: $p=0.004$, FDR-corrected, one-sided randomization test). In contrast, starting in pooling layer 4 all photo-to-drawing correlations were higher in VGG-16 SIN (all $p<0.004$, FDR-corrected, one-sided randomization test), reflecting indeed an attenuated drop in representational similarity between photos and drawings in VGG-16 SIN as compared to VGG-16 IN. A very similar pattern was found for the photo-to-sketch correlation. For early and intermediate layers up to pooling layer 4, the only difference in similarity was found in pooling layer 3, with higher values in VGG-16 SIN ($p=0.009$, FDR-corrected, one-sided randomization test; all other $p>0.05$, FDR-corrected, one-sided randomization test). For the later layers starting from pooling layer 5, we found consistently higher values for VGG-16 SIN (all $p<0.01$, FDR-corrected, one-sided randomization test). Finally, for the drawing-to-sketch comparison, we observed higher values in VGG-16 SIN only in pooling layer 3 (both $p<0.01$, FDR-corrected, one-sided randomization test) but no significant differences in any of the other layers (all $p>0.05$, FDR-corrected, one-sided randomization test).

To summarize, while representational similarities between natural images and abstracted types of depiction in early and intermediate layers were mostly similar or lower in VGG-16 SIN compared to VGG-16 IN, in the later layers of the network we found consistently higher values in VGG-16 SIN. For the drawing-to-sketch similarity we found overall mostly similar correlations in VGG-16 SIN and VGG-16 IN. Taken together, these results suggest a specific role of texture bias in the similarity in processing across levels of visual abstraction in late layers of VGG-16 but less so in the early and intermediate layers.

Experiment 3: Generalization through transfer learning

Reducing texture bias in VGG-16 improved the similarity in processing of object images across levels of visual abstraction particularly regarding the representational structure in late layers but mostly did not affect intermediate and early layers. This implies that the features in early and intermediate layers in the ImageNet-trained VGG-16 might already be general enough to support recognition of abstracted drawings. To directly test this, we used transfer learning, from the task of classifying natural object images to the task of classifying drawings and sketches. For this purpose, we fine-tuned VGG-16 pretrained on ImageNet on the ImageNet-Sketch dataset (Wang et al., 2019). We used representational similarities between the types of depiction as an index for the generality of representations in a given layer and based our selection of layers for transfer learning on that (Dwivedi & Roig, 2019). Hence, for training the network, we froze the connection weights of the network up to pooling layer 4 and only adapted the weights of the layers after pooling layer 4 which includes the last convolutional block, the last pooling layer, and all fully connected layers.

<u>Restored performance on drawings after fine-tuning</u>

After fine-tuning, we found high and comparable performance of the network (VGG-16 FT) for photos and drawings (M(Photos)=0.86, M(Drawings)=0.79, $\chi^2(1)=0.30$, p=0.59, FDR-corrected) but still lower and relatively poor performance for sketches (M(Sketches)=0.21, photos vs. sketches - $\chi^2(1)=33.44$, $p<0.001$, FDR-corrected; drawings vs. sketches - $\chi^2(1)=26.30$, $p<0.001$, FDR-corrected) (Fig 3). When comparing VGG-16 before and after fine-tuning, we found no statistically significant difference between networks on photos ($\chi^2(1)=0.30$, p=0.59, FDR-corrected). Further,

drawing performance was significantly higher in VGG-16 FT than in VGG-16 IN ($\chi^2(1)=33.44$, $p<0.001$, FDR-corrected) but sketch performance was not ($\chi^2(1)=2.68$, $p=0.15$, FDR-corrected). In direct comparison with human performance, we found that VGG-16 FT performed worse than humans on every type of depiction (all $p<0.001$, FDR-corrected, one-sided randomization test), mirroring the results of Experiments 1 and 2. Taken together, the fine-tuned network showed largely restored performance for drawings, yet, performance on sketches remained poor.

Increased representational similarity across levels of visual abstraction after fine-tuning

Since we explicitly fine-tuned the later layers on abstracted object images, we expected to see an attenuated decrease in representational similarity between photos and abstracted types of depiction towards the output layer as compared to VGG-16 IN in Experiment 1, similar to what we observed for VGG-16 SIN in Experiment 2. Indeed, after fine-tuning the network showed higher photo-to-drawing and photo-to-sketch similarity compared to VGG-16 IN across all fine-tuned layers (all $p<0.04$, FDR-corrected, one-sided randomization test) while drawing-to-sketch similarity was decreased (all $p<0.002$, FDR-corrected, one-sided randomization test).

We conclude that the fine-tuning procedure improved the generalization to drawings regarding the classification performance of the network and the representational similarity across levels of visual abstraction. This suggests that given the right mapping between early and intermediate features learned on natural images and the classification layer, the network is able to similarly process photos and drawings. The performance for sketches, however, remained poor indicating involvement of different recognition strategies for high levels of visual abstraction.

Discussion

The main objective of this study was to investigate to what extent generalization to abstracted object images is an emergent property of CNNs trained on natural images and how these generalization capabilities compare to those found in humans. To this end, we analyzed both the classification performance and the representational similarities across different levels of visual abstraction in VGG-16 and compared them

to humans. In Experiment 1, we showed that in the ImageNet-trained VGG-16, natural and abstracted object images evoke highly similar representations, with representational similarities peaking in intermediate layers. These similarities, however, dropped off sharply in later layers of the network, culminating in poor performance on drawings and sketches. In contrast, human behavioral performance was highly accurate and comparable across types of depiction. Moreover, we observed that representations of drawings and sketches clustered together with increasing degrees across layers, while photo representations remained well separated, suggesting either a collapse or a shift in drawing and sketch representations. Two findings supported the idea of a representational shift: the presence of high-level category information in the representation for drawings and sketches, and their representational similarities to human behavior. Together, the results in Experiment 1 suggest that photos, drawings and sketches share strong similarities in processing up to intermediate layers in the network, which disappear with increasing proximity to the output layer, ultimately resulting in poor generalization to drawings and sketches. In Experiment 2, we identified the degree to which a proposed texture bias in ImageNet-trained CNNs (Geirhos et al., 2019; Hermann et al., 2020) contributed to the drop in representational similarity and the poor generalization to drawings and sketches. We demonstrate that reducing texture bias in the network attenuated the dropoff in representational similarity between types of depiction specifically in later layers. This, in turn, suggests that representations in early and intermediate layers of the ImageNet-trained VGG-16 are already general enough to serve as the foundation for the recognition of drawings. In Experiment 3, we provide direct evidence for this notion by showing that increased performance for drawings was achieved by fine-tuning only the later layers in the network on a set of drawings while keeping the features in early and intermediate layers learned on ImageNet intact. This demonstrates that large parts of the ImageNet-trained VGG-16 are general enough to support recognition of drawings, yet, not for more abstract sketches. In the later stages of the network, however, the processing is presumably biased towards the image features and texture the network was trained on. This apparent bias favours a dissimilar representation for natural and abstracted object images and prevents the network from accurately classifying drawings and sketches.

Our findings reconcile the demonstrations of poor performance of CNNs on drawings (Ballester & Araujo, 2016; Evans et al., 2021; Wang et al., 2019) with evidence for generalizable representations for natural images and drawings (Fan et al., 2018). Our results furthermore suggest that processing of object images across levels of visual abstraction remains domain-general up to the penultimate pooling layer, after which a distinct representational format emerges that is more biased towards the domain of natural images and leads to incorrect classifications on drawings and sketches. Our observations in the different parts of processing in the networks are in line with distinct bodies of research on the processing in CNNs. While several studies find evidence for abstract shape representations and brain-like shape sensitivity in CNNs (Kalfas et al., 2018; Kubilius et al., 2016; Pospisil et al., 2018) others point out the specificity of these networks for the domain of their training data, in particular regarding their classification decisions (Geirhos et al., 2019; Geirhos, Temme, et al., 2018; Goodfellow et al., 2015; Hermann et al., 2020). We propose that both lines of observations might be explained with a dynamically changing representational format that is differentially influenced by shape and texture information depending on the processing stage in the network. A recent study provides further evidence for this notion by showing that the number of shape and texture encoding units varies across processing stages in CNNs (Islam et al., 2021). However, shape and texture are clearly not the only features that influence representations in CNNs. Hence, further research investigating similarities and differences in the representational structure for different types of stimuli across processing stages is needed to provide a better understanding of the nature of representations in CNNs.

Another possible explanation for the limited similarities between natural and abstracted object images is related to the specifics of the chosen neural network architecture. One such choice is the type of layer in the network. While the decline in representational similarity was found already in the final pooling layer, the presence of fully-connected layers might have contributed to the drop in performance. However, results of a similar analysis using a network consisting only of convolutional layers (vNet, (Mehrer et al., 2021)) continued to yield a drop in classification performance for drawings and sketches, paired with decreases in representational similarities across levels of visual abstraction, albeit weaker than those found for VGG-16 (Appendix A3). This demonstrates that limitations in generalization to drawings in Experiment 1 cannot

be explained solely by the presence of fully-connected layers. Future studies need to identify the degree to which the generalization performance can be maintained by changes in network architecture alone or whether alternative training regimes become necessary.

We demonstrated that similarity in processing of objects across visual abstractions is in part an emergent property in CNNs trained on natural images. Yet, despite attenuating texture bias and fine-tuning on drawings, the performance on very abstract sketch images remained poor. Hence, the recognition of highly abstracted and simplified object images might involve different mechanisms than the recognition of natural images and drawings, even though it is possible that fine-tuning with even more abstract images would improve performance on sketches. A possible mechanism for such robust recognition abilities might include some form of global shape processing. This ability is limited in predominant CNN architectures (Baker et al., 2018, 2020) and may involve computations including recurrent segmentation and grouping which are not performed by canonical CNNs (Doerig et al., 2020). Investigating how this challenge is solved by the human brain and networks with different architectures and processing mechanisms than current CNNs might elucidate the mechanisms that enable human generalization abilities, including those for abstract sketches.

While the generalization capabilities in human recognition are still out of reach for networks with a feedforward architecture and supervised learning regime (Geirhos et al., 2019; Geirhos, Janssen, et al., 2018; Geirhos, Temme, et al., 2018), developing models that more closely match the human brain in terms of architecture (Evans et al., 2021; Kietzmann, Spoerer, et al., 2019) and learning rules (Zhuang et al., 2021) offer new perspectives for meeting this goal. Such models have yielded important insight in how the brain solves the general problem of object recognition and show improved generalization in some cases (Geirhos, Narayanappa, et al., 2020; Spoerer et al., 2017) but are still outmatched by humans (Geirhos, Meding, et al., 2020; Geirhos, Narayanappa, et al., 2020). Only recently, a new class of models based on the concept of natural language supervision has been introduced, which show intriguingly robust performance across domains and across visual abstractions (Radford et al., 2020). It is an open and exciting question to what extent these models and their training

approach resemble principles of natural intelligence, which might ultimately provide new explanations for the remarkable generalization in human vision.

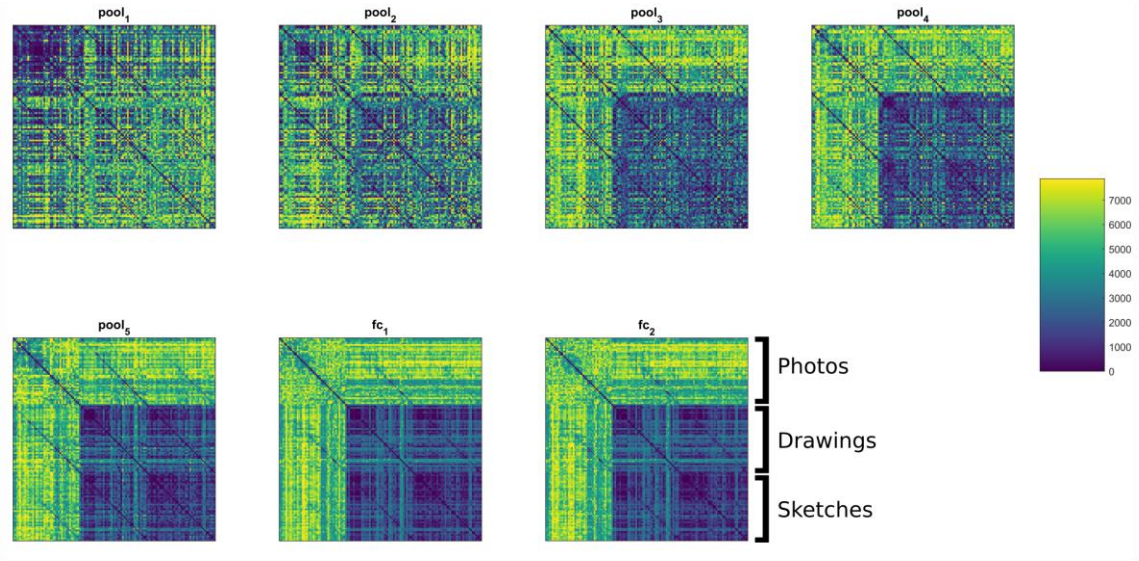
Conclusions

Our results show that a CNN trained on natural images represents photos and drawings of objects in intermediate layers highly similarly while simultaneously exhibiting low similarities between types of depiction in late layers and performing poorly on drawings and sketches. The texture bias present in CNNs (Geirhos et al., 2019; Hermann et al., 2020) contributes to the low representational similarities between types of depiction in later layers and the network's reduced performance for drawings. After fine-tuning these late layers on a set of object drawings, the network showed high performance on photos and drawings and increased representational similarities between types of depiction in late layers. In conclusion, these results contribute to the understanding of generalization to drawings in CNNs, by revealing high similarities in processing of natural images and abstracted object images in the network and providing explanations for the link between these similarities and the limitations in performance for abstracted object images.

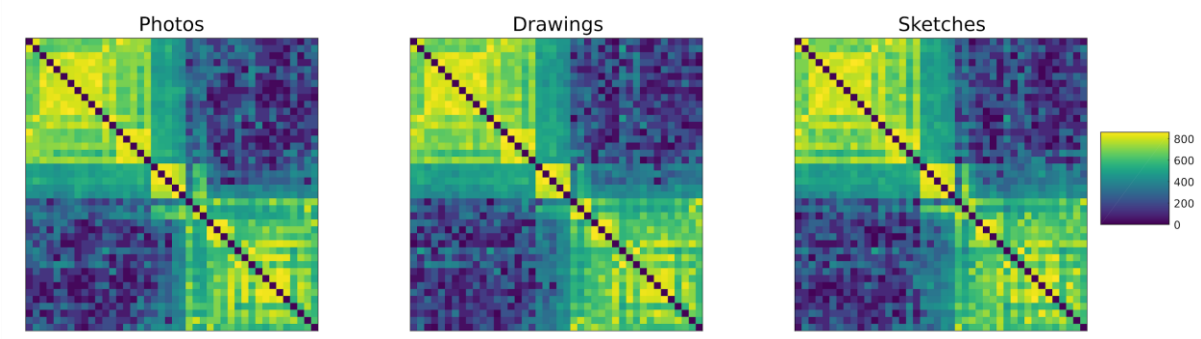
Acknowledgements

We thank D.B. Walther and Kate Storrs for early comments and discussions on the project and Juliana Sarasty Velez, who created the drawings and sketches used in this study and provided all rights to the images.

Appendix



A1. Representational dissimilarity matrices containing all pairwise distances between all stimuli in the set across types of depiction (super-RDMs) for VGG-16. Similar to the MDS visualization starting already in pooling layer 3 the distances for drawings and sketches within and between the two types of depiction became visibly smaller than the distances for within photos and between photos and the abstracted types of depiction, which increased up to the last fully connected layer. For visualization purposes all distances in one RDM were ranked and sorted according to their superordinate category (manmade/natural) within one type of depiction.



A2. Human behavioral RDMs for the three types of depiction. Based on the behavioral similarity judgments in the odd one out similarity task we calculated pairwise distances between objects as the probability of choosing object x and object y as belonging together regardless of the context imposed by object z in the triplet.

For visualization purposes all distances in one RDM were ranked and sorted according to their superordinate category (manmade/natural).

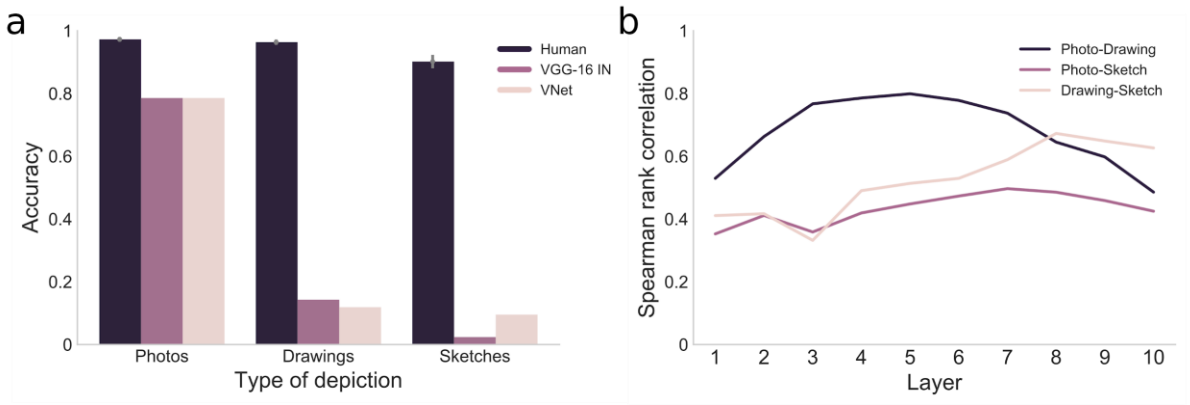
Representational similarities between type of depiction in a fully convolutional neural network

In order to test the role of fully connected layers in representational similarities of object images across levels of abstraction, we analyzed the fully convolutional neural network vNet (Mehrer et al., 2021). The architecture of vNet includes 10 layers with increasing kernel size mirroring the increase of average receptive field sizes along areas of the human ventral stream followed by a classification layer. Each of the layers in the network entails a convolution operation, dropout, max-pooling, group-norm and a ReLU nonlinearity (except layers 1,2,5,6 which do not include max-pooling). Importantly, in contrast to VGG-16, vNet does not have fully connected layers apart from the final classification layer. The network was trained on the ILSVRC2012 (Russakovsky et al., 2015) dataset such as the network in experiment 1.

Analogous to the procedure in experiments 1 to 3, we first obtained the top-1 accuracies for the stimuli in each type of depiction separately. VNet showed high accuracy on photos (*M*(Photos)=0.79) and a drop in performance for drawings and sketches (*M*(Drawings)=0.12, *M*(Sketches)=0.1), similar to the results in Experiment 1. Subsequently, we extracted the activations for our stimuli from all of the layers in vNet and computed RDMs based on the activations for all the types of depiction and layers separately. Finally, we computed the Spearman rank correlation between the lower triangular values of the RDMs of the different types of depiction for each of the layers. We found that representational similarities between photos and drawings increased in the early layers and reached a peak in layer 5 after which there was a drop in representational similarity (Fig A3). For the photo-to-sketch similarity a similar pattern was observed with an increase in similarity in early layers and a drop in the later layers after a peak in layer 7. Further, for the drawing-to-sketch correlation the similarity increased steadily up until the last layer in which the similarity dropped slightly compared to the penultimate layer.

In sum, the representational similarities between photos and abstracted types of depiction followed a comparable pattern to what was observed in the ImageNet-trained

VGG-16 in experiment 1 with a drop in similarity in the later layers close to the classification layer. Taken together, this indicates that the poor performance on drawings and sketches and the drop in representational similarity in VGG-16 between types of depiction cannot be fully explained by the presence of fully connected layers.



A3. Similarity in processing across levels of visual abstraction in a fully convolutional neural network. a) Top-1 accuracies for vNet for each type of depiction separately in comparison to human and VGG-16 performance. vNet performed similarly as VGG-16 with high performance on photos and poor performance on drawings and sketches. **b)** Representational similarities between types of depiction in vNet. Akin to the experiments with VGG-16 we observed a drop in similarity between representations of photos and both drawings and sketches in the later layers of vNet.

References

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLOS Computational Biology*, *14*(12), e1006613. https://doi.org/10.1371/journal.pcbi.1006613

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2020). Local features and global shape information in object classification by deep convolutional neural networks. *Vision Research*, *172*, 46–61. https://doi.org/10.1016/j.visres.2020.04.003

Ballester, P., & Araujo, R. M. (2016). On the performance of GoogLeNet and AlexNet applied to sketches. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, 1124–1128.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. JSTOR.

Biederman, I., & Ju, G. (1988). Surface versus edge-based determinants of visual recognition. *Cognitive Psychology*, *20*(1), 38–64. https://doi.org/10.1016/0010-0285(88)90024-2

Chang, C.-C., & Lin, C.-J. (2011). LIBSVM: A library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, *2*(3), 27:1-27:27. https://doi.org/10.1145/1961189.1961199

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755–27755.

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

Doerig, A., Schmittwilken, L., Sayim, B., Manassi, M., & Herzog, M. H. (2020). Capsule networks as recurrent models of grouping and segmentation. *PLOS Computational*

Biology, *16*(7), e1008017. https://doi.org/10.1371/journal.pcbi.1008017

Dwivedi, K., & Roig, G. (2019). Representation similarity analysis for efficient task taxonomy & transfer learning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12387–12396.

Eitz, M., Hays, J., & Alexa, M. (2012). How Do Humans Sketch Objects? *ACM Transactions on Graphics - TOG*, *31*. https://doi.org/10.1145/2185520.2185540

Evans, B. D., Malhotra, G., & Bowers, J. S. (2021). Biological convolutions improve DNN robustness to noise and generalisation. *BioRxiv*, 2021.02.18.431827. https://doi.org/10.1101/2021.02.18.431827

Fan, J. E., Yamins, D. L. K., & Turk-Browne, N. B. (2018). Common Object Representations for Visual Production and Recognition. *Cognitive Science*, *42*(8), 2670–2698. https://doi.org/10.1111/cogs.12676

Gatys, L. A., Ecker, A. S., & Bethge, M. (2016). Image Style Transfer Using Convolutional Neural Networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2414–2423. https://www.cv-foundation.org/openaccess/content_cvpr_2016/html/Gatys_Image_Style_Transfer_CVPR_2016_paper.html

Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2018). Comparing deep neural networks against humans: Object recognition when the signal gets weaker. *ArXiv:1706.06969 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1706.06969

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *ArXiv:2006.16736 [Cs, q-Bio]*. http://arxiv.org/abs/2006.16736

Geirhos, R., Narayanappa, K., Mitzkus, B., Bethge, M., Wichmann, F. A., & Brendel, W. (2020). On the surprising similarities between supervised and self-supervised models. *ArXiv:2010.08377 [Cs, q-Bio]*. http://arxiv.org/abs/2010.08377

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. *ArXiv:1811.12231 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1811.12231

Geirhos, R., Temme, C. R. M., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *ArXiv:1808.08750 [Cs, q-Bio, Stat]*. http://arxiv.org/abs/1808.08750

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and Harnessing Adversarial Examples. *ArXiv:1412.6572 [Cs, Stat]*. http://arxiv.org/abs/1412.6572

Güçlü, U., & Gerven, M. A. J. van. (2015). Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *Journal of Neuroscience*, *35*(27), 10005–10014. https://doi.org/10.1523/JNEUROSCI.5023-14.2015

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(11), 1173–1185.

Hermann, K. L., Chen, T., & Kornblith, S. (2020). The Origins and Prevalence of Texture Bias in Convolutional Neural Networks. *ArXiv:1911.09071 [Cs, q-Bio]*. http://arxiv.org/abs/1911.09071

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, *96*(16), 9379. https://doi.org/10.1073/pnas.96.16.9379

Islam, M. A., Kowal, M., Esser, P., Jia, S., Ommer, B., Derpanis, K. G., & Bruce, N. (2021). Shape or Texture: Understanding Discriminative Features in CNNs. *ArXiv:2101.11604 [Cs]*. http://arxiv.org/abs/2101.11604

Jozwik, K. M., Kriegeskorte, N., Storrs, K. R., & Mur, M. (2017). Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. *Frontiers in Psychology*, *8*, 1726.

https://doi.org/10.3389/fpsyg.2017.01726

Kalfas, I., Vinken, K., & Vogels, R. (2018). Representations of regular and irregular shapes by deep Convolutional Neural Networks, monkey inferotemporal neurons and human judgments. *PLOS Computational Biology*, *14*(10), e1006557. https://doi.org/10.1371/journal.pcbi.1006557

Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLOS Computational Biology*, *10*(11), e1003915–e1003915.

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. In T. C. Kietzmann, P. McClure, & N. Kriegeskorte, *Oxford Research Encyclopedia of Neuroscience*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190264086.013.46

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, *116*(43), 21854 LP – 21863. https://doi.org/10.1073/pnas.1905544116

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4–4.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 1*, 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. https://doi.org/10.1371/journal.pcbi.1004896

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Mantel, N. (1967). The Detection of Disease Clustering and a Generalized Regression Approach. *Cancer Research*, *27*(2 Part 1), 209.

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*(2), 153–157. https://doi.org/10.1007/BF02295996

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118. https://doi.org/10.1073/pnas.2011417118

Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*(1), 5725. https://doi.org/10.1038/s41467-020-19632-w

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. https://doi.org/10.1145/219717.219748

Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2020). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *BioRxiv*, 2020.07.22.216713. https://doi.org/10.1101/2020.07.22.216713

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Advances in Neural Information Processing Systems 32*, 8026–8037. http://papers.nips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pospisil, D. A., Pasupathy, A., & Bair, W. (2018). “Artiphysiology” reveals V4-like shape tuning in a deep network trained for image classification. *ELife*, *7*, e38242. https://doi.org/10.7554/eLife.38242

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2020). *Learning Transferable*

Visual Models From Natural Language Supervision. https://cdn.openai.com/papers/Learning_Transferable_Visual_Models_From_Natural_Language_Supervision.pdf

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., & Bernstein, M. (2015). Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252.

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Geiger, F., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). Brain-Score: Which Artificial Neural Network for Object Recognition is most Brain-Like? *BioRxiv*, 407007. https://doi.org/10.1101/407007

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. *ArXiv:1409.1556 [Cs]*. http://arxiv.org/abs/1409.1556

Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. *Frontiers in Psychology*, *8*, 1551–1551.

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2020). Diverse deep neural networks all predict human IT well, after training and fitting. *BioRxiv*, 2020.05.07.082743. https://doi.org/10.1101/2020.05.07.082743

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(23), 9661–9666. https://doi.org/10.1073/pnas.1015666108

Wang, H., Ge, S., Xing, E. P., & Lipton, Z. C. (2019). Learning Robust Global Representations by Penalizing Local Predictive Power. *ArXiv:1905.13549 [Cs]*. http://arxiv.org/abs/1905.13549

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23),

8619 LP – 8624. https://doi.org/10.1073/pnas.1403112111

Zhuang, C., Yan, S., Nayebi, A., Schrimpf, M., Frank, M. C., DiCarlo, J. J., & Yamins, D. L. K. (2021). Unsupervised neural network models of the ventral visual stream. *Proceedings of the National Academy of Sciences*, *118*(3), e2014196118. https://doi.org/10.1073/pnas.2014196118