Deep Convolutional Networks do not Perceive Illusory Contours

Nicholas Baker (nbaker9@ucla.edu)
Department of Psychology, Franz Hall, 502 Portola Plaza, Los Angeles, CA 90095 USA
Gennady Erlikhman (gerlikhman@unr.edu)
Department of Psychology, 1664 N. Virginia Street, Reno, NV 89557 USA
Philip Kellman (kellman@cognet.ucla.edu)
Department of Psychology, Franz Hall, 502 Portola Plaza, Los Angeles, CA 90095 USA
Hongjing Lu (hongjing@ucla.edu)
Department of Psychology, Franz Hall, 502 Portola Plaza, Los Angeles, CA 90095 USA

Abstract

Deep learning networks have shown impressive performance in object recognition. We used the classification image method to probe whether a deep learning model employs the same features as humans in perceiving real and illusory contours. We adopted a deep learning network, pre-trained with natural images, and retrained the decision layer with laboratory stimuli to perform shape discrimination in the "fat/thin" task. We tested the network with real and illusory contour stimuli contaminated with luminance noise. We found that deep networks trained on natural images can be readily adapted to discriminate between psychophysical stimuli with an extremely high degree of accuracy. However, deep learning networks do not appear to represent illusory contours where they may aid performance in the fat/thin task, a process automatically performed in human vision. This divergence indicates an important difference between the kinds of visual representations formed by deep networks and by humans.

Keywords: Deep learning, contour interpolation, classification images

Introduction

Object recognition is among the most important and remarkable functions of biological vision. Classifying objects into categories allows us to interpret a visual scene and make inferences about objects beyond the information present in the retinal image. The task of categorizing objects is made difficult by the vast diversity of visual features among objects of the same category and by the variety of contexts under which objects are viewed. These variations include differences in viewing angle, distance from the observer, qualities of the illuminant, and possible occluders fragmenting the projection of the object.

In the past decade, computational vision researchers have made remarkable progress in overcoming the many difficulties of object recognition. Most influential has been the application of deep convolutional neural networks (DCNNs) to object recognition. DCNNs built for object recognition are trained with millions of labeled photographs of objects and animals to classify an image into one of 1000 categories. They take an image as input and perform a series of matrix operations and nonlinear transformations to output a vector of probabilities for each of their trained categories. Unlike traditional neural networks, DCNNs have convolutional layers with filters that operate on only a subset of contiguous image pixels at a time. The effect is that spatial information is preserved in the image because two pixels must fall into the same convolutional window in order for correlations between them to be considered (LeCun, Bottou, Bengio & Haffner 1998). DCNN architectures have won the ImageNet object classification competition since their first entrance in 2012 (Krizhevsky, Sutskever & Hinton 2012), now achieving accuracies even better than human recognition performance.

Similarities between DCNNs and humans, both in structure and performance, have raised questions about the extent to which the computational processes taking place in deep networks are similar to those in human vision. One obstacle to answering these questions is that most research has been restricted to comparisons of categorization performance between deep networks and humans (e.g., Dubey, Peterson, Khosla, Yang & Ghanem 2015; Peterson, Abbott & Griffiths 2016). This can be a useful metric, but it can also be misleading if humans and artificial systems reach the same classification decision through very different computational processes. For example, consider Ringach and Shapley's (1996) fat and thin Kanizsa squares (Figure 1). For humans, discrimination of fat and thin stimuli is aided by the perception of illusory contours between the inducing elements (Gold, Murray, Bennett & Sekuler 2000). If deep networks were presented with similar stimuli, usual methods of comparison could assess discrimination between fat and thin stimuli, but not how this discrimination is accomplished. It would be impossible to know if DCNNs interpolate between inducing elements as humans do, or if they make their classification based on other information, such as the orientation of the black elements. In this study, we undertake to apply classification image techniques (Gold et al. 2000) to DCNNs to study the intermediate representations that drive their ultimate classification decisions.

One aspect of recognition that these methods could clarify is recognition of partially occluded objects. DCNNs develop some robustness by training with many images with different

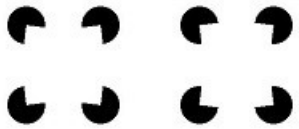

Figure 1. Fat and thin modified Kanizsa squares.

viewing contexts, for example, from non-canonical viewing angles or with partial occlusion. When occlusion is minimal, network classification remains fairly good. The VGG-19 network (Simonyan & Zisserman, 2014) correctly classifies Figure 2a as a lion despite the occluding cage bars. However, DCNN performance drops off considerably when presented with more significant partial occlusion. It cannot correctly classify Figure 2b, which is identical apart from the addition of four wider occluding black bars. The assigned probability for "lion" goes down from .777 (first choice) in Figure 2a to .002 (75^{th} choice) in the occluded image. On testing sets with multiple occluders, mean classification accuracy is between 35% and 20% among top performing DCNNs, depending on the number of occluders (Wang et al., 2017).

Figure 2. (a) Minimally occluded lion. Found online at: https://c1.staticflickr.com/3/2169/3527269138_36f6ce1988_b.jpg (b) Substantially occluded lion.

To the extent that deep networks recognize partially occluded objects, they could be doing so by completion of the object's shape or by recognition from partial information. In human perception, both strategies play a role in object classification, but there is substantial evidence that local completion is the more basic and obligatory perceptual process (Kanizsa 1979; Carrigan, Palmer & Kellman 2016). For example, it is much easier for humans to classify displays such as those in Figure 1 as fat or thin based on completion between the inducers than by looking at the orientation of individual elements.

In human perception, amodal completion (behind occluders) depends on the same visual mechanisms that give rise to illusory contour perception (Kellman, Yin & Shipley, 1998). In displays like the Kanizsa square, people see a subjective contour despite a total absence of luminance contrast between inducers. Gold, Murray, Bennett and Sekuler (2000) used classification image techniques to show that the image region between contour inducers is influential in subjects' classification of a presented Kanizsa square, or a partly occluded square, as fat or thin, even though the signal was totally absent from these regions.

Classification images are computed by first having observers make decisions about hundreds of images containing a signal (the stimulus pertinent to the perceptual decision) and random visual noise. The patterns of noise in the images are then correlated with classification decisions in order to determine which pixels (i.e. regions of the image) were important for classifying the image into one or the other category. This kind of analysis can give insight into where the behavioral receptive fields (BRFs) – areas important to observers' perceptual decisions – are in the image (see Murray (2011) for more information).

In the present study, we aimed to establish a method for conducting psychophysical experiments on DCNNs that would be informative not only about the network's final classification decisions, but would also provide insight into the stimulus information influential in the network's final output. First, we adapted a pre-trained deep network to new perceptual tasks by replacing the final layer and learning new weights between it and the preceding layer in order to allow for testing on more tightly controlled laboratory stimuli. This retraining only on the decision layer preserves all the learned features from training for object recognition, but repurposes the network's representations for a different task. We then used classification image techniques to systematically examine whether deep convolutional networks are sensitive to illusory contours between inducing elements. If classification image analyses revealed that networks formed behavioral receptive fields between inducers, that would be strong evidence of similarity between humans and such artificial systems. On the other hand, if networks did not show BRFs in the interpolating region, that would be evidence that DCNNs are not performing object completion, or at least that object completion does not involve illusory contour interpolation as it does in humans.

Experiment 1

The purpose of Experiment 1 was to develop and validate a method of using classification images to derive behavioral receptive fields in deep convolutional networks. We trained a network to classify wire frames as fat or thin, then tested the network with impoverished stimuli which had added pixel noise (Fig. 3). We then analyzed the noise fields from the testing phase to determine which image regions played a role in the network's classification decision.

Method

Training All training and testing was done using the AlexNet deep network model (Krizhevsky et al. 2012). We adopted a pre-trained network from Matconvnet (Veldadi & Lenc 2015) which was trained in the standard way to classify natural images from the ImageNet database including 1.2 million images and 1000 object categories. The decision layer of AlexNet has 1000 nodes, one for each object category. We replaced this layer with a single node layer for the binary "fat/thin" classification. The weights between the

penultimate, fully connected (fc8) layer and this final decision node layer were trained to classify wire frames as fat or thin, depending on the curvature of their vertical contour segments.

The network was trained to make the fat/thin classification from 22,000 wire frame images with the size of 227x227, half labeled fat, and half labeled thin. The curvature of the vertical segments varied from extreme (curves nearly touching in the thin stimulus) to negligible (horizontal position of the curves' midpoint only a few pixels away from the corners of the wire frame). The position of the wire frames in the image also varied, with the constraint that the whole shape must be visible. We added a small amount of Gaussian noise (SD of contrast = .16) to every pixel in the training image, as it was found through experimentation that this reduced decision bias in the training. See Figure 3 for training examples.

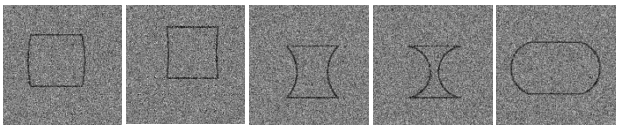

Figure 3. Sample training images from the second phase of training in Experiment 1. The training images varied the curvature of the vertical segments and location. Gaussian pixel noise were added to training images.

The network was trained for 20 epochs, after which it was tested with a validation set of an additional 2128 wire frame images that had been removed from the training set. The error rate on the validation set was .048.

Testing After the network had been trained to classify fat and thin wire frames, we conducted classification image analysis to examine which parts of the image were relevant to the network's classification decisions. To do this, we generated noise fields with a standard deviation of 0.16, then took fat and thin wire frames of intermediate curvature as signals with a contrast of 0.12, and added them atop the noise field.

In order to derive classification images, it is necessary to have both correct and incorrect responses for each target shape. To that end, we varied the contrast of the signal over several thousand trials and used the Palamedes toolbox (Prins & Kingdom, 2009) to fit a psychometric curve to the data, and find the contrast at which the network correctly classified about 75% of presented stimuli, as is standard in psychophysical classification image analysis (see Figure 4).

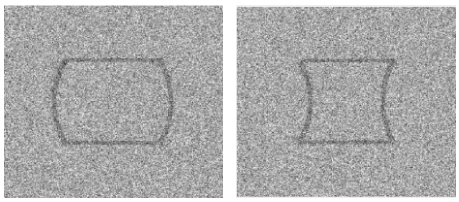

Figure 4. Sample test images from Experiment 1. Test images have fixed curvature but with adjusted signal contrast to maintain the accuracy at 75%.

We tested the network on 100,000 stimuli, recording the signal, the noise field, and the network response for each trial. Stimuli were identical in position, size and magnitude of curvature. The only stimulus features that changed from trial to trial were the convexity of curved segments (corresponding to "fat" or "thin" stimuli), and the randomly generated noise field.

Analysis We first analyzed the behavioral receptive fields from Experiment 1 using classical classification image methods, which was used in the human study by Gold et al. (2000). Trials were grouped into four categories: signal fat/response fat (S_fR_f), signal thin/response fat (S_tR_f), signal thin/response thin (S_tR_t), and signal fat/response thin (S_fR_t). We calculated the mean of the noise fields for each of these four kinds of trials, and then found the classification image by computing (1), where μ is the mean of the noise field corresponding to each classification type.

(1) $CI = (\mu.S_fR_f + \mu.S_tR_f) - (\mu.S_tR_t + \mu.S_fR_t)$

After examining the network's results, we found that it made considerably more S_tR_f misclassifications than S_fR_t misclassifications, caused by a bias term in the decision layer. Because it is important to have all response types well represented in classification image analysis, a biased pattern of response could make the derived CI less interpretable. The bias term in the network is unaffected by the presented stimulus, so to reduce its effects on the resulting classification images, we also performed a reverse correlation analysis. Rather than grouping noise fields by their four possible response types, we correlated each pixel intensity in the noise fields with the activity of the network's response nodes across 100,000 trials. This continuous measure was simply the dot product of the input to second to last network layer and the connection weights between the last and second to last layers. Conceptually, this analysis is almost identical to classification image analysis used in psychophysics, but it has the advantage of not being subject to the network's response bias in shape judgments.

Results and Discussion

Both the traditionally calculated classification image based on mean noise fields and the correlation map are shown in Figure 5. Gaussian smoothing was applied to both images to aid visualization. Darker regions correspond to areas that influence the network towards a "fat" classification, while lighter regions are areas that influence the network towards a "thin" classification.

The purpose of reverse noise image correlation techniques is to find areas that are influential to the network's ultimate classification. By analyzing the noise fields in the absence of the stimulus signal, psychophysicists can examine how random variation in the presented image can influence a subject's decision one way or the other. The results of Experiment 1 suggest that the same techniques can be applied to gauge deep networks to find what areas influence

the artificial systems' ultimate classification decision. In the wire frame experiment, there is significant correlation between the image region where a fat or thin vertical segment was overlaid and the network's final classification.

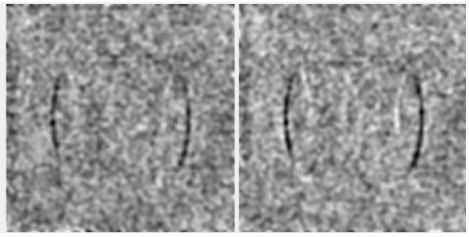

Figure 5. Left: Classification image Right: Correlation map. The pixel contrasts in the result images reflect the degrees that different locations influence the classification decision.

These findings validate the idea that deep networks trained for object recognition can be trained to do other perceptual tasks while preserving the features learned from training on natural image classification. Moreover, the correlation maps recovered from Experiment 1 confirm that classification images can be recovered from deep networks and give important insight into which stimulus regions are influential in a network's ultimate classification.

Experiment 2

In Experiment 2, we used reverse correlation analyses to test whether deep convolutional networks interpolate illusory contours between inducing elements. Classification image analysis on human perception has found that the region between inducers is influential in subjects' perceptual decision, even when the signal is only present at the inducers' locations (Gold et al. 2000). If deep networks process visual scenes as humans do, we would expect the same scene conditions that produce an illusory contour percept in humans to give rise to an illusory contour in the artificial system. We tested this by presenting to a DCNN fat and thin Kanizsa square stimuli with both real and illusory contours, and compared the classification images from the two conditions to see if the network, like humans, had a representation of the interpolating contour in the illusory condition.

Method

Training The first phase of training was identical to Experiment 1—we used AlexNet, a DCNN that had already been trained to classify natural images. In the second phase, we retrained the connection weights between the last two layers, this time to classify Kanizsa squares as fat or thin. The training set consisted of 22,000 images of sectors of circles that could define fat or thin shapes depending on the orientation of the circle inducers. In all training stimuli, a curved contour was drawn to connect between the corner inducers, so that all training was on stimuli with real contours. Sample training stimuli are shown in Figure 6. The stimuli are slightly longer vertically than horizontally. This is done because DCNNs have rotation invariance, so we needed there to be a difference between fat and thin images regardless of orientation. Training images varied in curvature, from one degree off true vertical to 44 degrees off true vertical, and in position in the image.

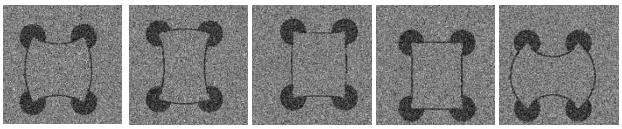

Figure 6. Training images for Experiment 2. Training images varied in curvature of the vertical segments and location. Gaussian pixel noise was added to training images.

The network was trained for 20 epochs, after which it was tested on 2128 images not included in training, for which it had an error rate of .027.

Testing Two testing conditions were carried out using the same retrained DCNN. First, we tested on stimuli with real contours connecting between inducers. We chose fat and thin signals with intermediate curvature and overlaid one or the other atop a randomly generated Gaussian noise field (standard deviation 0.16). As in Experiment 1, contrast between the signal and background was set so that the network correctly classified the image with about 75% accuracy (see Figure 7). We then ran 100,000 trials, half of which used the fat signal, and half the thin signal. The only stimulus features that varied across trials were the orientation of the inducers (angled inward for "thin" stimuli and outward for "fat stimuli"), the convexity of the segments between inducers, and the randomly generated noise field. Network response and the noise image were recorded for each trial.

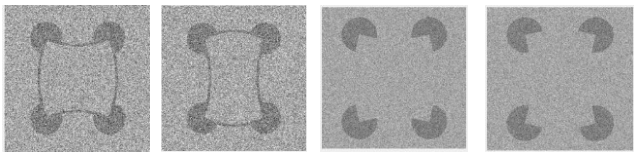

Figure 7. Testing images for the real and illusory contour condition. Test images have fixed curvature, but with adjusted signal contrast to maintain the accuracy at 75%.

We also tested the network on stimuli with no physical contour between partial circle inducers. We used inducers with the same orientation as in the real contour condition, and varied the contrast between the signal and background to find the 75% accuracy threshold (Figure 7, right). We tested the network on 100,000 illusory contour trials, recording network response and the noise image for each stimulus.

Analysis We analyzed the data by computing the classification image and correlation map for both conditions, as in Experiment 1.

Results and Discussion

The correlation map for the real contour and illusory condition are shown in Figure 8, along with the classification image derived from human subjects by Gold et al. (2000). The classification images were also computed, but are not shown because they look very similar to the correlation map, but with slightly less contrast between behavioral receptive fields and the background.

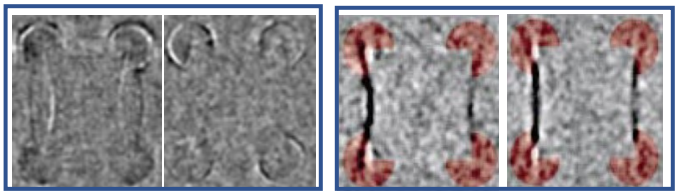

Figure 8 Left panel: Correlation map for the real contour and illusory condition from Experiment 2, respectively. Right panel: Classification image for the real contour condition and illusory contour condition (from Gold et al. (2000)).

When physical contours connect between the figure's inducing elements, both the orientation of the elements and the contours themselves appear to be influential in classification. These results are similar to Experiment 1, except that now there are two information streams that could lead to correct classification—orientation of inducer and contour curvature.

In the inducer-only contour condition, classification can be done by examining the orientation of the inducing elements, or by the curvature of an illusory contour connecting pairs of inducers. The correlation map for inducer-only stimuli looks dramatically different from the map for stimuli with real contours, and from the inducer-only condition in human subjects. The image region where inducers are present is highly influential in classification, but there appears to be no behavioral receptive field in the area between the partial circle inducers. This suggests that in the absence of real contours between inducers, the network classifies fat and thin stimuli purely based on the orientation of individual inducing elements, without perceiving interpolated contours between these elements. This is true even though real contours were present in all training images. Such a training regimen gives the network the best chance of representing illusory contours because the network will have learned to expect diagnostic information to be present between inducing elements, but correlation analysis reveals no contribution from the interpolating region. This differs from the behavioral receptive fields observed in humans for the same task (Fig. 8 right), which include the illusory contour region as well as oriented inducer region.

General Discussion

The purpose of this study was to develop a method for conducting more rigorous psychophysical tests of deep convolutional networks in order to probe the nature of their representations and computations, and to apply this method specifically to the question of contour interpolation.

Experiment 1 served as a validation for our method of using artificial stimuli and classification image techniques for probing the capabilities of DCNNs. Even though humans' visual systems did not evolve to process laboratory stimuli typically used in vision research, psychophysicists find it useful to simplify the visual input in an experiment in order to make their findings more interpretable. The same can be done in deep convolutional networks by replacing the decision layer with one more appropriate to a given perceptual task. One problem that comes with training DCNNs with millions of parameters is the risk of overfitting, as when a monkey is mistaken for a person due to its proximity to a vehicle (Wang et al., 2017). Use of laboratory stimuli can mitigate this issue by more tightly controlling what information is available to the network in classification.

The methodology we used in our experiments also provides insight into how deep convolutional networks make their classification decisions. In Experiment 1, we knew that the influential region in the wire frame images should be along the curved vertical contours, and we were able to produce classification images that confirmed this expectation. The structural complexity of DCNNs makes it very difficult to track computational processes from input to output, so a method like reverse image correlation is a promising tool for learning what information deep networks are using when they make one classification instead of another.

The usefulness of such a method becomes clear when we look at results from Experiment 2. Most research into the capabilities of DCNNs has been restricted to the performance level. In the inducer-only condition in Experiment 2, evaluation of the network based solely on performance would seem to suggest broad similarity between human and artificial perceptual processes. Like humans, the deep network was able to accurately classify oriented inducers as fat or thin configurations, even when there were no real contours connecting them. Differences between biological and artificial vision are only revealed when we look past the performance level and analyze the information that was used by each system in its ultimate perceptual decision.

Evidence that deep networks do not perceive illusory contours could support the notion that DCNNs do not do completion behind occluders, but recognize partially occluded objects from partial information. Unlike humans, the presence of illusory contour inducing edges satisfying geometric constraints of relatability is not sufficient to induce contour completion (Kellman & Shipley, 1991). An alternative explanation is that deep networks do amodal completion, but not modal completion. Under this hypothesis, there might be some scene requirements beyond the presence of tangent discontinuities and relatable edges to engage completion processes. Our current findings cannot decide between these possibilities, but it must be noted that either hypothesis represents a divergence from human perception,

where modal and amodal completion appear to depend on a common process (Kellman, Yin & Shipley, 1998).

One reason deep networks might not interpolate between relatable inducers is that they are purely feedforward systems. It is possible that a deep network with recurrent connections would be better suited to fill in spaces between tangent discontinuities on a backwards pass from higher level areas. Importantly, though, even this would constitute a difference between networks and humans, for whom interpolation is generally thought to be a feedforward process (e.g., Heitger, von der Heydt, Peterhans, Rosenthaler & Kubler, 1998).

Another reason networks might not interpolate between inducers is that the natural images on which they are trained do not have occluded target objects, so completion capabilities may be a low priority during training. It would be an interesting future direction to train networks on an image set with more occluded objects to test if more robust training would result in deep networks perceiving illusory contours.

One limitation of this study is that classification image techniques assume linearity in a system's decision-making process, but deep convolutional networks are inherently nonlinear. (We thank James Elder for bringing this issue to our attention). This is a subject of ongoing research, but preliminary findings suggest that analyses that do not assume linearity, such as regression using the general linear model, produce similar results.

Another limitation is that in Gold et al.'s (2000) study, exposure time for the stimuli was limited to 500 ms. It is possible that given unlimited time, human observers would make their classification based on the orientation of individual partial circle inducers, rather than on the features of the illusory contour. Since there is no way to limit exposure time for DCNNs, it is possible that the same regions are influential in humans and deep networks, given unlimited viewing time. We cannot rule this out, but it seems unlikely given the strength of the illusory contour percept. It does not seem probable that human observers would be more accurate in their "fat/thin" classifications by attending to individual inducer orientations, and it would certainly make the task more effortful and unpleasant.

Overall, our findings suggest that although deep convolutional networks resemble humans on many performance-based measures, there is a great deal of work to be done to evaluate how similar their intermediate computations really are to human perception. In the case of illusory contour displays like the Kanizsa square, the representations of humans and deep networks appear very different, as DNNs do not appear to interpolate between tangent discontinuities in the same way human observers do.

Acknowledgements

This research was funded by an NSF grant BSC-1655300 to HL, NEI F32EY025520 to GE.

References

Carrigan, S. B., Palmer, E. M., & Kellman, P. J. (2016). Differentiating global and local contour completion using a dot localization paradigm. *J. Experimental Psychology: Human Perception and Performance*, *42*(12), 1928-1947.

Dubey R, Peterson J, Khosla A, Yang M. H., & Ghanem B. (2015). What makes an object memorable? In Proc. of the IEEE Int. Conference on Computer Vision, 1089-1097.

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology*, *10*(11), 663-666.

Heitger, F., von der Heydt, R., Peterhans, E., Rosenthaler, L., & Kübler, O. (1998). Simulation of neural contour mechanisms: representing anomalous contours. *Image and Vision Computing*, *16*(6-7), 407-421.

Kanizsa, G. (1979). *Organization in vision: Essays on Gestalt perception*. Praeger Publishers.

Kellman, P. J., & Shipley, T. F. (1991). A theory of visual interpolation in object perception. *Cognitive psychology*, *23*(2), 141-221.

Kellman, P.J., Yin, C. & Shipley, T.F. (1998). A common mechanism for illusory and occluded object completion. *Journal of Experimental Psychology: Human Perception & Performance, Vol. 24,* No. 3, 859-869.

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).

LeCun, Y., Bottou, L., Bengio, Y., & Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, *86*(11), 2278-2324.

Murray, R. F. (2011). Classification images: A review. *Journal of Vision*, *11*(5), 2-2.

Peterson JC, Abbott JT, Griffiths TL. Adapting deep network features to capture psychological representations. arXiv preprint arXiv:1608.02164. 2016 Aug 6.

Prins, N. (2014). Kingdom, FAA (2009). Palamedes: Matlab routines for analyzing psychophysical data.

Ringach, D. L., & Shapley, R. (1996). Spatial and temporal properties of illusory contours and amodal boundary completion. *Vision research*, *36*(19), 3037-3050.

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

Vedaldi, A., & Lenc, K. (2015, October). Matconvnet: Convolutional neural networks for matlab. In *Proceedings of the 23rd ACM international conference on Multimedia* (pp. 689-692). ACM.

Wang, J., Zhang, Z., Xie, C., Zhou, Y., Premachandran, V., Zhu, J., ... & Yuille, A. (2017). Visual Concepts and Compositional Voting. *arXiv preprint arXiv:1711.04451*.