Vision Research

A FAILURE TO LEARN OBJECT SHAPE GEOMETRY: IMPLICATIONS FOR CONVOLUTIONAL NEURAL NETWORKS AS PLAUSIBLE MODELS OF BIOLOGICAL VISION

--Manuscript Draft--

Manuscript Number:	VR-20-214R1
Article Type:	VSI: Deep Neural Networks
Keywords:	Visual processing; convolutional neural networks; Shape processing
Corresponding Author:	Dietmar Heinke Birmingham, UNITED KINGDOM
First Author:	Dietmar Heinke
Order of Authors:	Dietmar Heinke
	Peter Wachman
	Wieske VanZoest
	E. Charles Leek
Manuscript Region of Origin:	UNITED KINGDOM
Abstract:	Here we examine the plausibility of deep convolutional neural networks (CNNs) as a theoretical framework for understanding biological vision in the context of image classification. Recent work on object recognition in human vision has shown that both global, and local, shape information is computed, and integrated, early during perceptual processing. Our goal was to compare the similarity in how object shape information is processed by CNNs and human observers. We tested the hypothesis that, unlike the human system, CNNs do not compute representations of global and local object geometry during image classification. To do so, we trained and tested six CNNs (AlexNet, VGG-11, VGG-16, ResNet-18, ResNet-50, GoogLeNet), and human observers, to discriminate geometrically possible and impossible objects. The ability to complete this task requires computation of a representational structure of shape that encodes both global and local object geometry because the detection of impossibility derives from an incongruity between well-formed local feature conjunctions and their integration into a geometrically well-formed 3D global shape. Unlike human observers, none of the tested CNNs could reliably discriminate between possible and impossible objects. Detailed analyses using gradient-weighted class activation mapping (GradCam) of CNN image feature processing showed that network classification performance was not constrained by object geometry. We argue that these findings reflect fundamental differences between CNNs and human vision in terms of underlying image processing structure. Notably, unlike human vision, CNNs do not compute representations of object geometry. The results challenge the plausibility of CNNs as a framework for understanding image classification in biological vision systems.

February 22nd, 2021

Dear Dr Oruç,

Re: Ms. No.: VR-20-214: A failure to learn object shape geometry: Implications for convolutional neural networks as plausible models of biological vision

Thank you for sending us these very helpful reviews and comments. To address the issues raised we have undertaken a substantial rewrite, and reorganisation of the manuscript, and included additional analyses and statistical comparisons between network and human performance. Several of the key issues noted in the reviews arose because of a lack of clarity in the writing and organisation. These have been fully addressed. The substantive conclusions of the work remain the same – but now reinforced, and strengthened, with the additional analyses. We believe that the work will make a significant novel contribution to the field.

In what follows, we have detailed the changes, and our responses to the reviews, in blue text.

Detailed Response to Reviews

AE = Action Editor, R1 = Reviewer 1; R2 = Reviewer 2

Action Editor

AE1 Both reviewers have raised concern over the modest size of your dataset with several dozen images available to train the models, compared to the typical training/development sets consisting of millions of images. This issue is especially critical here in the context of models unable to successfully classify possible vs. impossible images.

Response: Thank you for raising this issue. The augmentation approach we have taken to increase the size of the dataset is commonly used in machine learning and generates a relatively large training set and test image set. However, a key aspect of the rationale underlying our human-network performance comparison is its qualitative validity. It is important to note that we evaluate performance in both pre-trained and un-trained networks to test a specific hypothesis about the internal representational structures of object geometry that are generated by architectures that – when trained, are highly successful in image classification. Critically, we aim to test this hypothesis in a manner that is comparable to the qualitative experience of a human observer. Human observers reliably discriminate between possible and impossible forms without any prior training or experience with these specific forms of stimuli – because, we argue, the biological system computes internal representations of 3D object geometry (and CNNs do not). Thus, extensive training of networks on the classification task using larger datasets of possible and impossible forms would fundamentally undermine the validity of the human-network performance comparison that we aim to achieve. We have clarified this key point about the rationale in the revised manuscript.

AE2 Reviewer 2 has suggested that creating an additional small dataset in which the impossible/possible distinction does not depend on global shape is necessary to confirm that the dataset size is not the bottleneck in the present study, and I agree.

Response: The difficulty here is that the possible/impossible distinction relies on the mismatch, or incongruency, between well-formed local image features and global shape geometry. Thus, it is not possible to construct a dataset in which geometric impossibility does not depend on global shape.

AE3 Both reviewers have also commented on the reporting of Study 2 and raising some questions as to the usefulness of including it as a central piece of the overall work. Reviewer 1 has suggested significantly shortening and moving to supplementary materials.

AE4 Reviewer 1 is not convinced by your conclusion that the models have indeed based their decision predominantly on the image backgrounds. They have asked for additional analyses and discussion to bolster this point.

Response: The motivation for inclusion of the analyses reported as Study 2 in the original manuscript was not clear – and we thank the reviewers for highlighting this. The initial analyses of network performance showed that all of the tested networks performed worse than human observers on the classification task. This point is now strengthened by the inclusion of robust statistical contrasts using a modified t-test (Crawford, Garthwaite & Porter, 2010; Crawford & Garthwaite, 2002). We highlighted a potential confound in the original dataset which we further explored and reported as Study 3 in the original manuscript. We now report those additional analyses as 'Supplementary Analyses of Network Performance' in which we ran an additional simulation using a background size normalised dataset. This is also important to show the robustness of the original results – that is, that network performance is replicated in a different dataset. The findings confirm that the networks fail to learn the classification task.

Reviewer 1

R1.1 Although study 2 should be mentioned as an example of best practice in identifying an error and removing it (so keep in a paragraph about it to educate PhD students) giving a study with errors the largest section of the paper is very odd and highly misleading. Why have you sent in a paper where there are more results shown for the erroneous study than the fixed repeat?

Response: Please see AE4 above.

R1.2 No one is going to believe your results in the 'fixed' study when you claim that the impossible objects are assigned based on background, the exact same finding as you got from the erroneous study, and that result is not explained in any way. We only have your word that you've managed to fix the error. Have you? Are you sure?

Response: We now report statistical analyses comparing the original and normalised images.

R1.3 This paper is not strong enough to make the claims that CNNs are not plausible models for human vision, as made in the title.

Response: It is important to note that the substantive point of our paper is to test a specific hypothesis about the nature of the representational structure/s that are computed by CNNs to achieve image classification – namely, that the networks (unlike humans) rely on localist structure that does not consider global object geometry. To do so we investigated whether different CNNs are able (like human observers) to reliably discriminate between possible and impossible objects – as this task requires computing local and global object geometry. The results show that the networks fail at this task.

R1.4 Rewrite with more analysis of study 3 (but keep in the error from study 2 for pedagogical reasons - shorten it or move results to supp. info.).

Response: Please see AE4 above.

R1.5 Give me IoU numbers if you want me to believe that the CNN is attending to the background.

Response: We have now included quantitative measures for how networks attend to the background and the impossible part (heatmap analysis).

R1.6 You could look at the Grad-CAM for different layers of the CNN to try to understand your results.

Response: We have now included the results from Alex net for multiple layers. These heatmaps are consistent with our interpretation of the results.

R1.7 If Grad-CAM gives such weird results, try using other visualization methods to get more information and elucidate what is going on.

Response: The Grad-CAM analysis has provided a valuable tool for shedding light on how the CNNs are attempting to resolve the classification task. We have included further detailed analyses of the Grad-Cam data (heatmap analysis) in the revised manuscript.

R1.8 Discuss human vision with regards to perspective line drawings.

Response: We have added further discussion regarding our choice to use line drawing stimuli – which are widely used in studies of human vision. One reason is that line drawings allow us to test hypotheses about the recovery of 3D object shape from geometric cues alone (i.e., without texture, shading etc.).

R1.9 I don't believe your results, but if you fix it and still get around 50% on the correct data, you should make the point that with two outputs, a random choice would be 50%, then do the stats to show if your results re significantly different from 50%, and if not, then you can conclude that the NN has not be able to do the task.

Response: We have now included rigorous statistical comparisons between all test networks and human performance as requested.

R1.10 From my memory of my reading about human visual processing and perspective, I read that humans had to learn to 'see' (and understand) perspective images (I think these are fixed-point perspective images as they have a single vanishing point), and that the discovery of this method of drawing was the major breakthrough of the Renaissance. I think it would improve the manuscript to mention this history and add in a (short) discussion of the effects of the discovery of the perspective on human visual perception. The book Art and Visual Perspective by Rudolf Arnhelm and references within is a good place to start.

Response: See R1.8 above. We have included further discussion about the use of line drawing stimuli, and their importance to the rationale. Though interesting, we have not extended the discussion to include reference to the development of perspective as an artistic technique (which does not seem directly relevant to the current study). Note that human observers are readily able to identify objects from line drawings.

R1.11 The authors mention Grad-CAM as a method to understand how the NN is making its decision, there are other such methods, and it would improve the manuscript to mention some of these and explain why they were not chosen in this work (1-2 sentences).

Response: We have included a discussion along these lines in the introduction.

R1.12 The authors mention that they applied rotations to the objects, but not how big the rotations were. Given the way the objects are drawn (perspective projection) I suspect that a large rotation would look odd to a human being and might well involve different processing pathways. I think the authors should expand on this if they used large rotations and show some images with large rotations so the reader can see if those images appear valid. (If the rotations are small, as is standard in CNN data augmentation, this is not necessary although the rotation angle range should be added to the paper).

Response: We have now included the parameters for our augmentations in the method section.

R1.13 I notice that 5 students were excluded due to low accuracy. Was this the case that they were not doing the trail properly or that there are some humans that have difficulty with perspective projection type images?

Response: For completeness we have now included all participants in the analysis, apart one who showed an accuracy close to 50%.

R1.14 Page 16, the authors state that the task was doable but not easy, is this due to the difficulty of understanding perspective images? (c.f. my suggestions on perspective for the intro).

Response: Please see our earlier responses to this point above.

R1.15 The authors state 'There was no significant difference between impossible and possible shape (88.8% vs. 83.8%; t (19) = 1.76; p = 0.094), but the confusion matrix (Table 1) indicates that participants had a small bias towards responding impossible shape. Measured in the framework of signal detection theory (SDT) the sensitivity (d':2.43) indicated that participants signal for possible vs impossible was fairly strong and we were not able to detect any bias either way (c: 0).'

Is confusing. As there is no sig. diff. between the no correct for impossible vs impossible shapes, how can there be a small bias towards impossible shapes? Is this stating that there is a sig. diff. in the errors in the table? Also, I am familiar with signal detection theory but I don't understand how the authors have talked about it here. What is d'? what is c? can you define it please.

Response: We have now clarified this in the revised manuscript.

R1.16 Study 2. I know the authors put in the accuracy for the CNNs earlier in the document, but it would be useful to have a table of the difference between training on IM and training on these objects, this would back up the statement 'none of the networks achieved a high-level of classification accuracy.' Also regarding 'none of the networks achieved a high-level of classification accuracy.' is this true? AlexNet for example has a relatively low top-1 accuracy (from memory I think it might be as low as 56%, do check) so the AlexNet results don't look that bad to me. I know that the other cNNS have much higher accuracies (although do check that you are using top-1 accuracy as a comparison, as I do not think top-5 is comparable to this task).

Response: We agree with the reviewer that the Top-1 accuracy is a benchmark. We have now include these accuracies in the manuscript.

R1.17 Where is the data to show if the correct possible vs incorrect possible difference is significant? You had this for the human data, I think it should be included for the CNN data.

Response: We have now included rigorous statistical comparisons between all test networks and human performance. The results show that the networks' performance is inferior to human performance. Therefore, we think the inclusion of such a comparison is not meaningful. However, we included here the results for the benefit of the reviewer.

Original dataset:

	Training			Validation		
Network	t-value	p-value	d-value	t-value	p-value	d-value
AlexNet	-0.77	0.453	-0.34	-0.73	0.474	-0.323
VGG11	-1.38	0.186	-0.61	-1.37	0.186	-0.613
VGG16	-0.89	0.385	-0.398	-0.89	0.385	-0.398
ResNet18	-12.86	0	-4.591	-11.96	0	-4.627
ResNet50	-19.29	0	-6.937	-20.17	0	-7.505
GoogLeNet	-4.09	0	-1.459	-2.11	0.049	-0.7904
AlexNet (pretrained)	-10.34	0	-4.03	-17.25	0	-6.199
VGG11 (pretrained)	-10.48	0	-3.845	-9.42	0	-3.877
VGG16 (pretrained)	-4.11	0	-1.645	-1.66	0.114	-0.720
ResNet18 (pretrained)	-18.23	0	-6.757	-11	0	-4.242
ResNet50 (pretrained)	-8.85	0	-3.616	-3.07	0.006	-1.264
GoogLeNet (pretrained)	1.55	0.139	0.626	6.05	0	1.995

Normalized dataset:

	Training			Validation		
Network	t-value	p-value	d-value	t-value	p-value	d-value
AlexNet	-1.63	0.120	-0.729	-1.791	0.089	-0.799
VGG11	-1.37	0.186	-0.613	-1.37	0.186	-0.613
VGG16	-0.438	0.666	-0.195	-0.44	0.666	-0.196
ResNet18	-6.78	0	-2.877	-7.467	0	-3.211
ResNet50	-3.79	0.001	-1.665	-2.214	0.039	-0.919
GoogLeNet	-1.00	0.331	-0.431	-0.28	0.782	-0.112
AlexNet (pretrained)	-0.876	0.392	-0.389	-0.65	0.523	-0.285
VGG11 (pretrained)	-2.29	0.033	-1.005	-2.04	0.055	-0.900
VGG16 (pretrained)	-2.00	0.061	-0.842	0.400	0.694	0.176
ResNet18 (pretrained)	-12.91	0	-5.211	-0.16	0.878	-0.055
ResNet50 (pretrained)	-4.85	0.000	-2.087	-1.04	0.312	-0.428
GoogLeNet (pretrained)	1.77	0.092	0.754	0	1	0

R1.18 Table 4: GoogLeNet results do not show that the CNN is looking at the background. Please discuss.

Response: We have now included GradCam/Heatmap analysis for GoogLeNet too.

R1.19 These findings raised the possibility of a systematic confound between the stimulus sets. In fact, we closely inspected the impossible objects and found that they are slightly smaller than possible objects. This confound should be able to explain our results, as the area size of the background is diagnostic for impossible objects. This means that all the data from study 2 is meaningless! As such I do not know why it is reported in this paper. It could be added to a supplementary information as an example of good practice fro drilling down into odd results to find an error, but this section does not show anything about the task! Why is it in this paper? It should be removed. A single paragraph explaining the error and how it was found is sufficient.

Response: See our earlier response above (AE4).

R1.20 Study 3. It is interesting that the CNN uses the background to identify impossible objects, but odd. Given study 2 this raises the question of whether the authors have properly removed the issue to do with size of the objects. This needs to be answered satisfactorily and some attempt needs to be made to check this. Also, why not add some stats, something like intersection over union values for attention (what I am calling the hot bits of the heat map) over a. the amount of the pixel space covered by the object and b. the part of the object that is impossible. These values are required to support this statement 'Grad-CAM results for this network (Table 7) seem to suggest that it attempted to use the background again to separate the two classes' (and are easy to get). This result is so odd that the authors need to be more convincing that it is true and try to understand why it is true.

Response: See our earlier responses above (AE4; R1.2).

R1.21. I want to see the results for AlexNet somewhere in this paper, as being a smaller network (easier to understand) and the claims that it learned gabor filters and is more like human vision, any paper purporting that CNNs are not like human vision (which I agree with incidentally) needs to address AlexNet.

Response: AlexNet is one of the networks included in our study (un-trained and pre-trained).

R1.22 You only had 64 images for training, ImageNet uses 1.3 million. Discuss the effects of this. Also a CNN can easily memorize this dataset. Are you sure that your results are not due to the CNN having memorized the dataset and thus it is looking at the parts of the image that cause that image to differ from the others in the set, and not the part that is impossible? Check this, it could explain the results.

Response: We have addressed this point above (AE1). Note that our augmentation led to 6400 training images. This is still smaller than commonly used, but as we explained earlier it is a reasonable size for the purpose of our study. The augmentation (0-360 rotation, horizontal flips and 0.9-1 zooms) also produces very different pixel patterns in the input images. It is not clear to us how the CNNs could have memorized such a dataset. This is also supported by fact that the heatmap analysis indicates that the networks pay much attention to the background.

Reviewer 2

R2.1 In the introduction, I would have been curious to see a cited source for the sentence "Recent work has also shown that the (human) biological system computes shape information in parallel across both global and local spatial scales, and that it integrates this information during perceptual processing to generate representations of structured scene content and object geometry" (p. 4).

Response: We have addressed this point in the revised text including supporting references.

R2.2 One question I had related to this work was whether the augmented dataset the authors used was large enough to for the networks to learn to classify between possible and impossible objects. These objects are handcrafted with important controls between the possible and impossible stimuli, so I understand it would be hard to generate thousands of different training stimuli. One way the authors might address this is by augmenting another set of 40 image pairs that do not depend on global shape and showing that in that case the network does have enough training examples to accurately classify images in the validation set.

Response: We have addressed this point above (AE2).

R2.3 Another point I would be interested to see discussed more is what the findings on Experiment 2 mean about deep networks. As the authors have currently written the paper, Experiment 2 lacks a control that, when corrected, supports the idea that DCNNs do not perceive global shape. I would recommend that if the authors think the network's success based on small size differences means something interesting about how DCNNs classify objects, they should add a little more discussion about that. Otherwise, it might make more sense to only report Experiment 3 with the size controlled.

Response: See our earlier response above (AE4; R1.22).

R2.4 One other very minor point about the size control: the authors re-tested the network after controlling for the size of the objects, but did they re-test humans? It seems extremely unlikely that humans' accurate performance in the behavioral experiment comes from an unconsciously perceived size difference, but the authors might draw a clearer distinction between DCNNs and humans if the behavioral experiment was done on size-controlled stimuli.

Response: We have addressed this point by re-analyzing the existing data. In this re-analysize we removed the bias substantially and still found no significant effect (see footnote in manuscript)

Typos corrected.

A FAILURE TO LEARN OBJECT SHAPE GEOMETRY: IMPLICATIONS FOR DEEP CONVOLUTIONAL NEURAL NETWORKS AS PLAUSIBLE MODELS OF BIOLOGICAL VISION

Dietmar Heinke[1,] Peter Wachman[1], Wieske van Zoest[1], E. Charles Leek[2]

School of Psychology, University of Birmingham[1], and Department of Psychology, University of Liverpool[2]

RUNNING HEADER: CNNs AND OBJECT SHAPE GEOMETRY

Address for Correspondence

Dietmar Heinke
School of Psychology
University of Birmingham
Birmingham B15 2TT
d.g.heinke@bham.ac.uk

Acknowledgement

D.H.‘s work was supported by a grant from the UK-ESRC ES/T002409/1.

ABSTRACT

Here we examine the plausibility of deep convolutional neural networks (CNNs) as a theoretical framework for understanding biological vision in the context of image classification. Recent work on object recognition in human vision has shown that both global, and local, shape information is computed, and integrated, early during perceptual processing. Our goal was to compare the similarity in how object shape information is processed by CNNs and human observers. We tested the hypothesis that, unlike the human system, CNNs do not compute representations of global and local object geometry during image classification. To do so, we trained and tested six CNNs (AlexNet, VGG-11, VGG-16, ResNet-18, ResNet-50, GoogLeNet), and human observers, to discriminate geometrically possible and impossible objects. The ability to complete this task requires computation of a representational structure of shape that encodes both global and local object geometry because the detection of impossibility derives from an incongruity between well-formed local feature conjunctions and their integration into a geometrically well-formed 3D global shape. Unlike human observers, none of the tested CNNs could reliably discriminate between possible and impossible objects. Detailed analyses using gradient-weighted class activation mapping (GradCam) of CNN image feature processing showed that network classification performance was not constrained by object geometry. We argue that these findings reflect fundamental differences between CNNs and human vision in terms of underlying image processing structure. Notably, unlike human vision, CNNs do not compute representations of object geometry. The results challenge the plausibility of CNNs as a framework for understanding image classification in biological vision systems.

Word count = 252

Few machine learning methods have received as much interest in recent years as deep (multi-layer) feedforward convolutional neural networks (CNNs) - the performance of which is unparalleled across a range of image processing tasks (Guo et al., 2016; LeCun, Bengio, & Hinton, 2015; Voulodimos, Doulamis, Doulamis, & Protopapadakis, 2018). CNNs are also increasingly attracting attention in vision science due to their high levels of performance in image classification (and other) tasks that matches (and sometimes exceeds) that of human observers. They also superficially share certain similarities to other properties of biological vision systems including: a hierarchical structure, convolutional sampling across increasingly large 'receptive fields', and their capacity to support category generalisation (e.g., Cox & Dean, 2014; Güçlü & van Gerven, 2015; Kuzovkin et al., 2018). Recent work has also highlighted similarities between patterns of activity within specific layers of trained networks and neural properties at intermediate and higher-levels of cortical representation using techniques such as representational similarity analysis (e.g., Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016; Khaligh-Razavi & Kriegeskorte, 2014; Yamins, Hong, Cadieu, et al, 2014).

At the same time, the suitability of CNNs as a theoretical framework for understanding biological vision remains unclear. Here we examine this issue in the context of object classification. Human observers are remarkably adept at object recognition. We can rapidly classify objects despite changes in sensory information brought about by variation in viewpoint, lighting, and other factors (e.g., Harris, Dux, Benito & Leek, 2008; Leek, Atherton & Thierry, 2007). This ability is supported by a processing system that can compute structured, hierarchical, representations of 3D object shape geometry from 2D retinal sensory input (e.g., Bar, 2003; Davitt, Cristino, Wong & Leek, 2014; Leek, Reppa, Rodriguez & Arguin, M, 2009; Leek, Reppa & Arguin, 2005; Leek, Roberts, Dundon & Pegna, 2018; Reppa, Greville & Leek, 2015; Schyns & Oliver, 1994).

One important characteristic of object processing in human vision is that both global, and local, shape information is computed, and integrated, during perceptual processing. For example, numerous studies have shown that observers can rapidly classify scenes based on coarse analyses of

global image content alone (e.g., Bullier, 2001; Peyrin, Michel, Schwartz, Thut, Seghier et al., 2010; Peyrin, Baciu, Segebarth & Marendaz, 2004; Schyns & Oliva, 1994), and that rapid analyses of low-spatial frequency global image content constrains local high spatial frequency processing of local structure during object recognition (e.g., Bar 2003; Bar et al., 2006). Other work has shown that global and local information is integrated during the perceptual processing of object shape – as shown, for example, in the context of global-to-local processing in Navon-type displays (Navon, 1977 – see also, Beaucousin, Simon, Cassotti et al., 2013; Deco & Heinke, 2007; Han, He & Woods, 2000; Proverbio, Minniti & Zani, 1998), dissociations between local and global processing in patients with unilateral brain lesions (Robertson, Lamb & Knight, 1988; Robertson & Lamb, 1991), and – more recently, deficits to global but not local eye movement scanning patterns during object recognition in patients with acquired visual agnosia (Leek, Patterson, Paul, Rafal & Cristino, 2012). Further work, using event-related potentials (ERPs), has found evidence for an early differential perceptual sensitivity to local and global 3D shape structure during image classification within the first 200ms of stimulus onset (Leek, Roberts, Oliver, et al, 2016; Oliver, Cristino, Roberts et al, 2018).

In contrast, the functional contribution of local and global shape structure to image classification in CNNs is unclear. The architecture of CNNs (increasing larger receptive fields) seems to suggest that, in principle, they could process global or higher-order image structure (e.g., Kriegeskorte, 2015; LeCun et al., 2015; Zeiler and Fergues, 2014). However, other work suggests that CNNs rely exclusively on local image information (e.g., Baker et al., 2018; 2020; Brendel & Bethge, 2019; Geirhos et al. 2019). For example, Baker et al. (2018) examined the performance of two pretrained CNN architectures (VGG19 and AlexNet) in their ability to classify images of objects with either congruent or incongruent (e.g., mixed) global shape and local textures (e.g., a camel outline shape with a zebra's texture). The results showed that network performance (unlike human observers) was perturbed by incongruency - with classification errors biased towards classification based on local but not global image properties (see also Geirhos et al., 2019).

This current study aims to further investigate this issue by testing whether CNNs compute representations of global and local 3D object geometry. To do so, we examined the ability of six (pre-trained and un-trained) CNNs (AlexNet, VGG-11, VGG-16, ResNet-18, ResNet-50, GoogLeNet) to discriminate geometrically possible and impossible novel objects (see Figure 2 and 3). These sorts of stimuli comprise a 2D depiction of a 3D form that cannot be geometrically reconstructed in 3D space – like the well-known Penrose triangle (Penrose & Penrose, 1958). This class of stimuli has also been extensively used previously to study how the human visual system computes representations of object shape (e.g., Carrasco & Seamon, 1996; Cooper, Schacter, Ballesteros, & Moore, 1992; Freud, Avidan, & Ganel, 2013; Freud et al., 2017; Freud, Hadad, Avidan, & Ganel, 2015; Schacter, Cooper, & Delaney, 1990; Schacter, Cooper, Delaney, Peterson, & Tharan, 1991). By definition, the ability to discriminate possible from impossible objects requires computation of a representational structure of shape that encodes both global and local object geometry, since the detection of impossibility derives from an incongruity between well-formed local feature conjunctions and their integration into a structured representation of global object shape. That is, object impossibility can only be detected at a level of perceptual processing in which local geometric structure is integrated into a coherent and physically possible 3D object. Thus, it follows that perceptual sensitivity to object impossibility implies a level of processing in the biological vision system that involves the integration of local and global object geometry. Our goal was to examine whether CNNs can, in principle, learn to discriminate geometrically possible from impossible objects to evaluate whether the networks – like the biological system, also compute, and integrate, local and global representations of 3D object shape geometry.

The design of the study had two further important aspects. First, human observers can readily detect object impossibility without prior training, or exposure, to this specific class of stimulus (e.g., Carrasco & Seamon, 1996; Cooper et al., 1992; Freud et al., 2013) suggesting that this ability reflects fundamental representational properties of the object processing system in human vision. For this

reason, we wanted to test the performance of both pre-trained and un-trained networks with minimal prior exposure to impossible objects to indirectly probe the internal representational structures that the networks have acquired to support image classification. The rationale is that the failure of the networks to reliably discriminate possible and impossible forms can be taken as evidence that image classification is not based on the integration of internal representational structures that make explicit local and global object geometry. Second, a further key aspect of the rationale was the use of datasets comprising line drawing depictions of novel 3D polyhedral. This class of stimulus provides a strong test of the ability to generate representations of 3D object structure from geometric cues alone – and are readily perceived by human observers (e.g., Attneave, 1954; Biederman, 1987; Pizlo (2014); see Sayim & Cavanagh, 2011, for a recent review). Here we use this stimulus class to provide a strong test of the capability of CNNs to generate representations of 3D object geometry.

METHOD

Networks

We trained 12 CNNs to perform an object discrimination task involving the classification of possible and impossible object shapes. The 12 CNNs were based on four architectures, AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGG (Simonyan & Zisserman, 2015), ResNet (He, Zhang, Ren, & Sun, 2016) and GoogleNet (Szegedy et al., 2015). For each of these architectures we also tested a pre-trained and an un-trained version. The pre-trained version was based on the ImageNet database as set-up in PyTorch (Paszke et al. 2019).

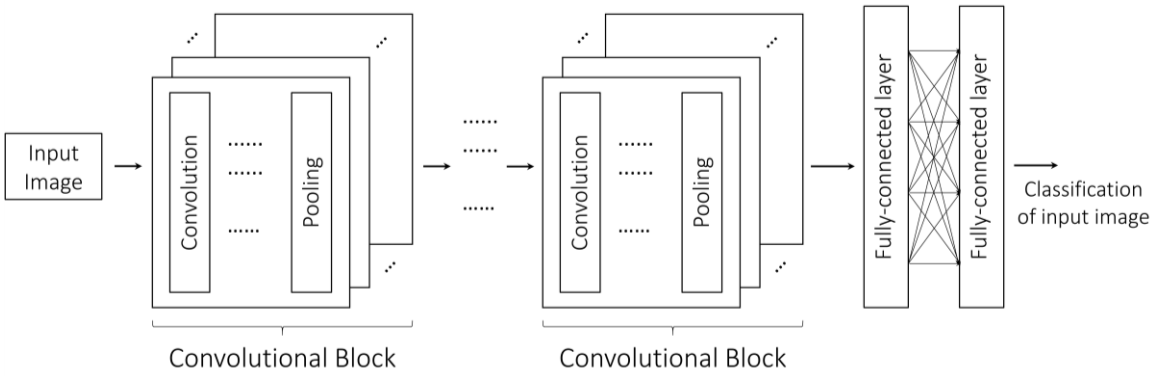


Figure 1. *This figure illustrates the structure common to all CNNs used in the papers. The CNNs categorize an input image through a pipeline of stages. The first stages consist of convolutional blocks which in turn are made of convolution layers and pooling layers. The second series of stages comprises fully connected layers. A fully connected layer calculates weighted sums across all inputs. The convolution layers convolve the input with a kernel of pre-defined size that often varies across the convolutional layers. The kernel size and the type of pooling layer, and the number of layers depend on the particulars of the architecture (see main text for details).*

As illustrated in Figure 1, the four CNN architectures categorize an input image through a pipeline of stages. Each stage consists of a set number of layers. The number of layers vary across the different architectures and their specific instantiations (see below for details). Typically, the first stage consists

of blocks made from convolutional layers and pooling layers. The second stage is a classifier comprised of fully connected layers. The convolutional layers convolve the input with a kernel of pre-defined size that often varies across the convolutional layers of a particular network. Subsequently, the output of convolutional layers may be processed with a pooling layer. There are different types of pooling mechanisms. The most common form of pooling, called max pooling, simply divides the input into patches of a predefined size, and then outputs the maximal value in each patch. The results of convolutional layers and pooling layers are then vectorised and fed into fully connected layers. A fully connected layer calculates weighted sums across all inputs. During the training process the values of the kernels and the weights of the fully connected layers are modified. Other characteristics like max pooling, kernel sizes, and number of layers are constant. We outline below the main characteristics of the tested CNN architectures together with their accuracy on ImageNet. Typically, a network's response is considered as accuracy if the correct response is among the five categories with the highest output activations (Top-5 accuracy). However, given that our benchmark is a two-category problem we report the Top-1 accuracies here.

AlexNet (Krizhevsky, Sutskever, & Hinton, 2012) has five convolutional layers and three fully connected layers (62.4 M parameters: Top-1 accuracy on ImageNet 62.5%). The first layer has a kernel size of 11x11, and Layer 2 a kernel size of 5x5. All other layers have a kernel size of 3x3.

VGG (Simonyan & Zisserman, 2015) architecture is based on five blocks with a kernel size of three across all blocks. We tested two VGG networks, **VGG-11** (113M parameters; accuracy: 69%) and **VGG-16** (138M parameters; accuracy: 74%) (Configuration D, Simonyan & Zisserman, 2015). In VGG-11 the first two blocks consist of one layer (convolutional layer plus pooling layer) each while the other three blocks are made of two layers each. A pooling layer is used only at the end of the three blocks. In VGG-16 the first two blocks consist of two layers followed by a pooling layer while the remaining three blocks consist of three layers followed by a pooling layer. There are three fully connected layers in each VGG-version.

ResNet (He, Zhang, Ren, & Sun, 2016) has a 7x7 kernel convolution layer in the first block and one fully connected layer. Importantly, ResNet contains weighted "short-cut connections" which bypasses convolutional layers, and their result is added to the output of the convolutional layers ("short-cut connections"). We tested two ResNet networks, **ResNet-18** (11M parameters; accuracy: 72.12%) and ResNet-50 (25.5M parameters; accuracy 77.15%). In ResNet-18 the short-cut connections bypass only one convolutional layer (4 blocks with two 3x3 kernel layers each) while in **ResNet-50** three layers are skipped (4 blocks with varying number of 3x3 kernel layers).

GoogLeNet (Szegedy et al., 2015; 6.4M parameters; accuracy of 69.78%) is a 22-layer network that also includes a new mechanism for each layer termed an inception module. An inception module consists of three filters with different kernel sizes (1x1, 3x3, 5x5) and a max pooling layer. The outputs of these filters are concatenated and form the input to the next layer. Prior to the 3x3 and 5x5 filters a channel pooling (1x1 convolution) takes place creating a "bottleneck" for these filters. The channels are made of several parallel convolutional layers also called feature maps. The network consists of 11 blocks. The first block is a standard layer with a 7x7 kernel. All other blocks comprise two parallel inception modules. There is only one fully connected layer.

We evaluated this large number of CNNs to explore whether specific network characteristics contribute to classification accuracy in the possible/impossible discrimination task. For instance, CNNs with the highest number of parameters (VGG-11 and VGG-18) may be better equipped to learn mappings from the objects to the two categories. On the other hand, since the test images are simple contour-based line drawings and provide only a small training set (even though we used data augmentation) these networks may be prone to overfitting. Here using the pre-trained network (where only the fully connected layer is trained) may alleviate this problem. However, given Geirhos et al.'s (2018) study, we expect that the pre-trained approach would fail as these networks are biased towards the local level, while the untrained networks can be adapted to the task. Of course, network architecture is also likely to be a critical determinant of performance. One such property is kernel

size. Larger kernels may be assumed to capture properties of global shape, and smaller kernels local elements (e.g., corners, line crossing, etc.). Hence, AlexNet with the larger kernels in the first blocks, may be superior compared to other networks. On the other hand, GoogLeNet can adapt the kernel size and, together with the bottleneck mechanism, might be predicted to have more success in the task. The short-cut connections in ResNet also provide an important mechanism for the task at hand as they may, in principle, allow ResNet to integrate global and local levels of shape information.

Datasets

We used a base set of 40 possible and 40 impossible objects (adapted from Williams and Tarr, 1997; see Figures 2). Impossible objects were created by one modification of the drawing of a possible object (see Figure 3). For each possible object there was a corresponding matched impossible object. Some stimuli were modified and redrawn to ensure that possible and impossible objects were matched for complexity in terms of contours and vertices. The complexity of objects was not significantly different: possible vertices (M=29.15, SD=5.69), impossible vertices (M=29.35, SD=5.65), t (39) = 1.275, *ns*; possible contours (M=38.98, SD=7.9), impossible contours (M=38.63, SD=7.78), t (39) = 1.617, *ns*.

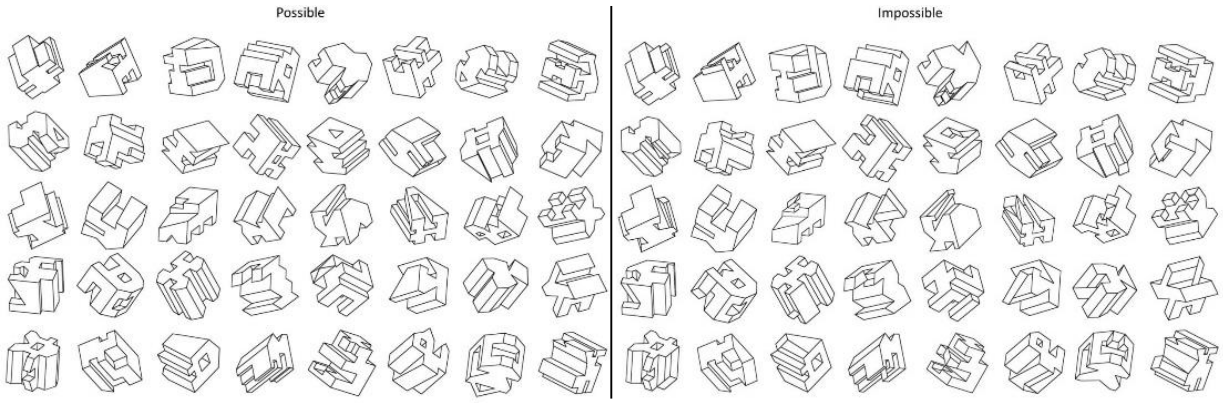


Figure 2*. This figure shows all possible and impossible shapes used to train the networks. The right panel shows the impossible shape corresponding the shapes on the left.*

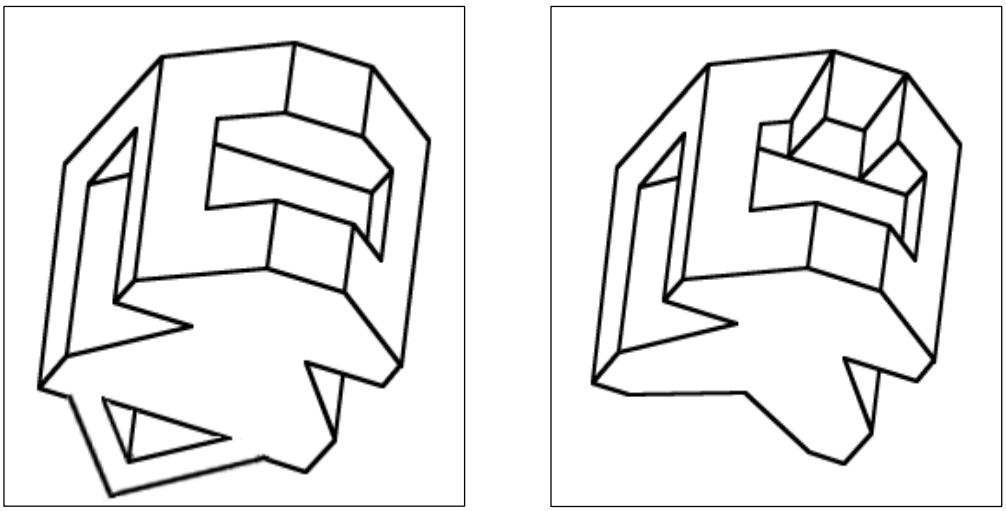


Figure 3 *Example of how a possible shape (left) was turned into an impossible shape (right).*

Network Training

Each network was trained for 100 epochs, which appeared to be approximately when the loss and accuracy scores stopped improving, based on preliminary testing. All networks were built using PyTorch 1.2.0 on a cuda-enabled NVIDIA GeForce GTX 1060 GPU. We fitted the CNNs with the *Adam optimiser* which typically shows good performance with little to no hyperparameter tuning (Kingma & Ba, 2015). We averaged the results across 20 different seeds in line with the Monte Carlo validation method or Repeated random sub-sampling validation (e.g., Picard & Cook, 1984).

We inverted the images (black pixels to white pixels, and vice versa). They were then converted to 224×224-pixel images with three colour channels and normalised in accordance with the pre-processing procedures used for ImageNet. We applied data augmentation at every training batch consisting of random rotations (0-360 degrees picture plane), horizontal flips and random zooms (0.9 – 1). We also applied data augmentation to the validation set so that we could perform two iterations of the validation at the end of each epoch, reducing the sensitivity of the validation scores to noise.

We pseudo-randomly divided the images into training and validation sets where 20% of the data (N=16) was reserved for validation and the remaining 80% (N=64) was used for training. Importantly,

since we augmented the images for each epoch, we obtain a training set comprised of 6400 images. When dividing the dataset, we ensured that each possible-impossible object pair was in the same (training or validation) dataset. This was done to facilitate the networks' ability to learn what constitutes a possible or impossible object, and to ensure that the number of possible and impossible images was balanced between each set.

The code for the project is available at https://github.com/PWman/Impossible-Shapes-Paper.

Analyses of Network Performance

1. Network accuracy

The outputs of each network tested were adapted to have two output nodes (one-hot encoding), as opposed to binary encoding with a single output node, to ensure the network was compatible with Grad-CAM. For all networks tested, we used the PyTorch Cross-Entropy Loss Function to calculate network error, since this internally applies softmax to the outputs during the calculation of loss. Mean network accuracy on the validation dataset was compared to human performance using the modified t-test (Crawford, Garthwaite & Porter, 2010; Crawford & Garthwaite, 2002) with an a priori $p < .05$ alpha criterion. We also analysed network performance and human performance using discriminability (d') and criterion shift (c) based on signal detection theory (SDT; Macmillan and Creelman, 1991).

2. Heatmap analysis

The goal of these analyses was to elucidate which region of the images contribute to network classification performance. At the present there are three types of methods to determine these regions: gradient-based methods (e.g., Simonyan, & Zisserman, 2015), perturbation-based methods (e.g., Wagner et al., 2019) and class activation mapping (CAM) methods. We used a recently developed tool from the CAM family, Grad-CAM (*Gradient-weighted Class Activation Mapping*; Selvaraju et

al., 2019). GradCAM usefully provides heatmap visualisations representing the degree by which regions in the input image contribute to the correct classification. Here we expect that Grad-CAM heat maps should highlight the single local region of each impossible objects that gives rise to the local-global shape incongruency. The heatmaps were determined for each of the 16 images in the validation dataset and then averaged across all 20 seeds.

To further analyse the spatial distributions of the activation in the heatmaps, we defined two ROIs, ROI-Background and ROI-Impossible. ROI-Background was determined with the Flood Fill algorithm from Python's scikit-image toolbox. ROI-Impossible marked the local region of shape impossibility, determined by one of the authors. Based on these two ROIs we then calculated the ratio of GradCam activation in each ROI and the total Grad-Cam activation. Note that a value of 100% would indicate perfect correspondence between regions of Grad-Cam activation and the region defined by the ROI. The ratios reported here are the averages across all heatmaps and all seeds.

Human Performance: Stimulus Validation Study

We also conducted a stimulus validation study to determine whether human observers could reliably discriminate possible and impossible objects using the stimulus set described above.

Participants The experiment tested 25 students recruited via advertisement on University of Birmingham social media pages who were reimbursed £10 for their time. Written informed consent was gained prior to participation with procedures approved by the local ethics committee.

Materials and Apparatus The stimuli were the same 80 base images used in the network dataset and another 160 images generated by randomly flipping and rotating (0-360 degrees) the base set. The stimuli were scaled to 768x 768 pixels and presented centrally on a standard 22" monitor.

Design and Procedure Participants first completed 6 practice trials for which they received feedback. These trials used a random choice of the original images. Each trial began with the presentation of a fixation cross (see Figure 4). The presentation time varied randomly between 300ms and 900ms to prevent the trials becoming too predictable. Stimulus duration was 2500ms. Each stimulus was followed with a screen asking them to indicate whether the shape was possible or impossible by pressing a key. There were two breaks: one after 74 trials and another halfway after another 80 trials. After the practical trials, the images were presented in two blocks. First the images not seen during the practice session were presented. In the second block 160 randomly generated images were shown. The order of stimuli was randomised within each block, and each block and contained equal numbers of possible and impossible shapes.

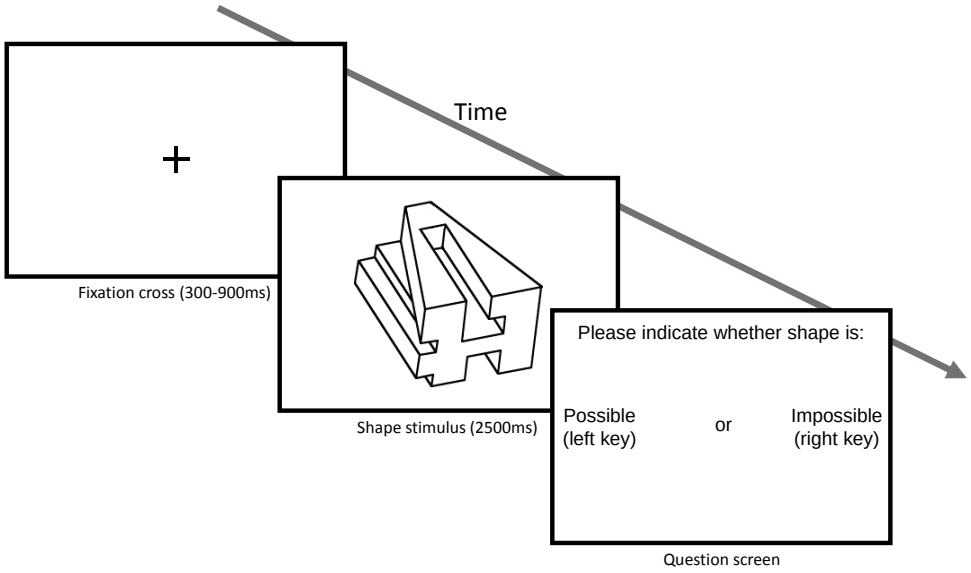


Figure 4 *Illustration of the trial structure for the behavioural study (see text for details).*

RESULTS

Human Performance: Behavioural Stimulus Validation Study

One participant was removed from the analysis as they showed an accuracy close to 50%. Participants classified images with a high degree of accuracy (M = 86.7%; SD = 4.88; 95%CI 84.6-88.7). There was no significant difference between classification accuracy for impossible and possible objects (*M*

= 88.1%; SD=7.3 vs. *M* = 85.4%; SD=9.4%; t (23) = 0.97; p = 0.341; Cohen's d = 0.32). Based on the confusion matrix (Table 1), the discriminability between possible and impossible was high (d' = 2.11), and there was no criterion shift (c: 0). These results show that human observers, without prior experience or training, can reliably discriminate the possible and impossible objects used in the network dataset.

Table 1 *Confusion matrix for human performance showing the % of responses by stimulus category (possible/impossible) and response.*

		Response [%]	
		impossible	possible
Stimulus Category	impossible	42.6	7.4
	possible	5.9	44.1
	TOTAL	48.5	51.5

Analyses of Network Performance

Table 2 shows a summary of network accuracy for all versions (pre-trained and un-trained) of the tested networks. Overall, performance was poor. In all cases, network performance was significantly below human performance as indicated by the modified t-test. The best result was achieved by ResNet-18 (pre-trained) with a mean accuracy of 67.7%. To understand better the influence of network architecture, we further analysed the results of the best network from each architecture, VGG-11 (pre-trained), AlexNet (pre-trained) and GoogLeNet (un-trained). The confusion matrix (Table 3) shows that the networks were better at identifying possible objects than impossible objects while being biased towards responding with "possible object" (apart from GoogLeNet). In other words, these results are not consistent with human behaviour.

Table 2. *Mean and standard deviation of training and validation accuracies and losses for each network from 20 seeds. The train loss and validation loss were determined with PyTorch's native cross-entropy loss function. The networks in red indicate the best validation accuracy for each network architecture. The modified t-test compares the validation accuracy with human performance.*

Network	Train Accuracy [%]	Validation Accuracy [%]	Modified t-test	Train Loss	Validation Loss
Un-trained					
AlexNet	56.2 ± 7.8	55.5 ± 6.3	t=6.26, p<0.001	0.67 ± 0.04	0.67 ± 0.03
VGG-11	50.0 ± 5.6	50.0 ± 0.0	t=7.36, p<0.001	0.69 ± 0.0	0.69 ± 0.0
VGG-16	48.8 ± 5.3	50.0 ± 0.0	t=7.36, p<0.001	0.69 ± 0.0	0.69 ± 0.0
ResNet-18	70.2 ± 5.0	63.9 ± 8.1	t=4.40, p<0.001	0.57 ± 0.05	0.73 ± 0.23
ResNet-50	65.3 ± 6.9	65.9 ± 8.9	t=4.17, p<0.001	0.64 ± 0.08	0.65 ± 0.11
GoogLeNet	66.0 ± 4.8	65.5 ± 9.5	t=4.25, p<0.001	0.99 ± 0.05	0.62 ± 0.07
Pre-trained					
AlexNet	66.1 ± 4.1	64.1 ± 7.7	t=4.53, p<0.001	0.65 ± 0.08	0.66 ± 0.11
VGG-11	67.6 ± 6.6	64.1 ± 4.4	t=4.53, p<0.001	0.7 ± 0.1	0.65 ± 0.1
VGG-16	67.3 ± 7.2	63.6 ± 6.6	t=4.53, p<0.001	0.73 ± 0.16	0.71 ± 0.13
ResNet-18	68.9 ± 5.3	67.7 ± 8.3	t=3.57, p=0.001	0.57 ± 0.04	0.59 ± 0.07
ResNet-50	69.4 ± 5.4	66.4 ± 6.5	t=4.07, p<0.001	0.57 ± 0.05	0.6 ± 0.06
GoogLeNet	67.6 ± 4.4	64.7 ± 7.3	t=4.41, p<0.001	0.59 ± 0.05	0.62 ± 0.04

Table 3 *Confusion matrix for four convolution neural networks using the original dataset. The confusion matrix shows the % of the network's response by stimulus category (possible/impossible) and response.*

		GoogLeNet (un-trained) Response [%]		AlexNet (pre-trained) Response [%]		VGG-11 (pre-trained) Response [%]		ResNet-18 (pre-trained) Response [%]	
		impossible	possible	impossible	possible	impossible	possible	impossible	possible
Stimulus Category	impossible	34	16	12	38	16	34	21	30
	possible	12	38	1	49	3	47	9	41
	Total	46	54	13	88	19	81	29	71
	SDT	d':0.94 c: 0.47		d': -1.42 c: -0.72		d': -0.97 c: -0.49		d': -0.44 c: -0.22	

Figure 5 shows the results from Grad-CAM. Inspection of the visualisations shows an inconsistent pattern across stimuli and networks. While there is some indication that ResNet18 (pre-trained) based its decisions on impossible parts of some objects, the network classification decisions frequently involve image background. This bias seems particularly striking in GoogleNet, where the distinction between possible and impossible shape is based on the background. Since AlexNet is very popular in brain imaging and its layered structure is relatively easy to understand we also included the

GradCam results for the layers in the un-trained and pre-trained in Table5. As expected, the activation from the pre-trained network tends to show an increase of the receptive field size. The untrained network shows a similar effect, but the receptive fields are generally wider possibly indicating that the network was trying to capture the global shape but failed as the architecture is too constraining.

Table 4 *Grad-CAM heatmap visualisations representing the degree to which regions in the input image contribute to the correct classification. Red indicates high contributions while blue indicates no contribution.*

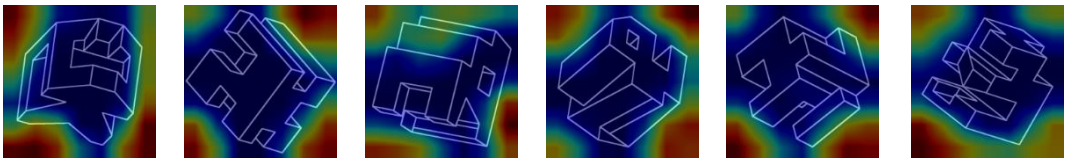
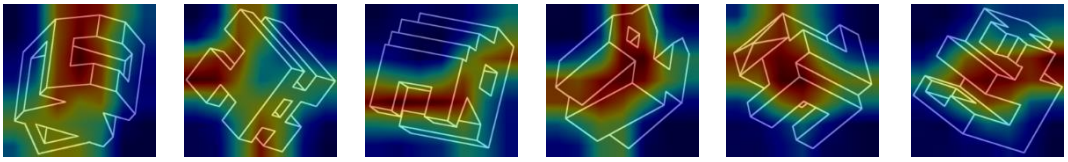
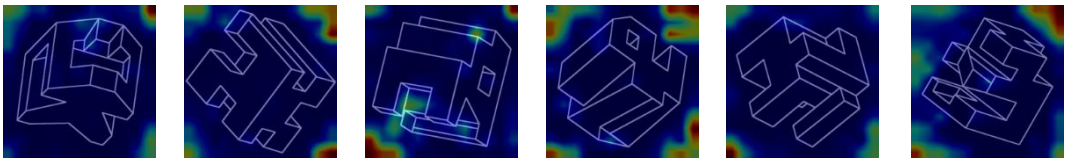
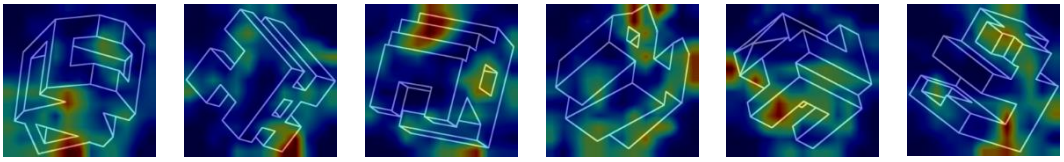
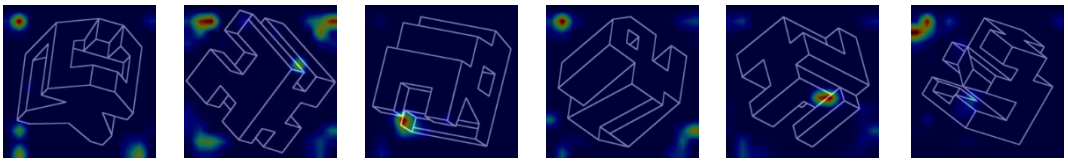
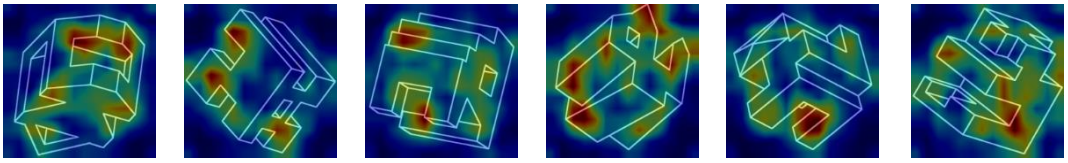
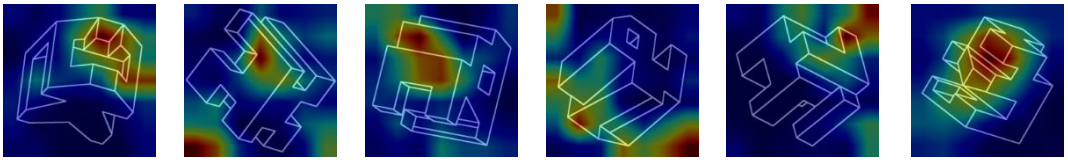
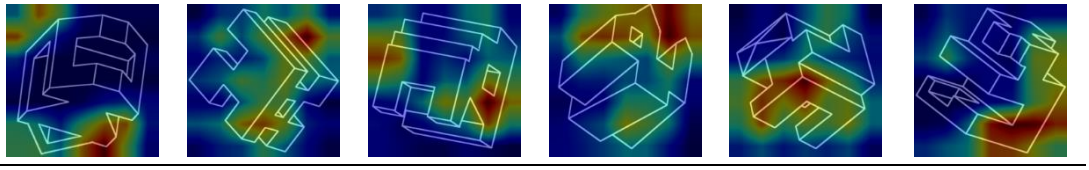
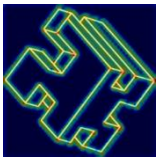
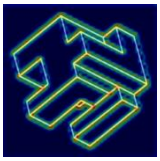
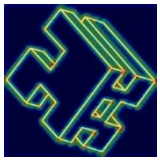
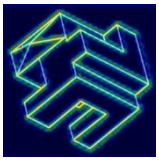
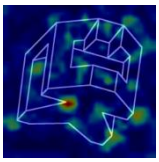
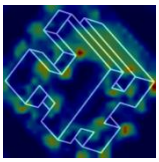
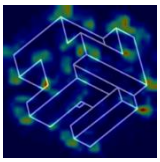
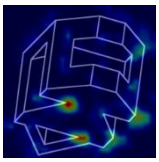
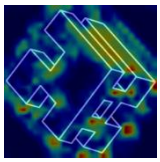
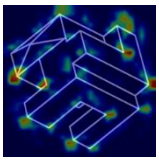
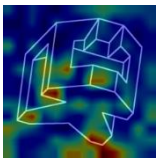
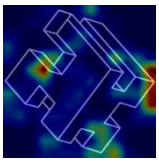
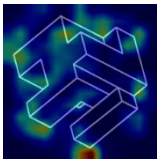
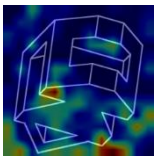
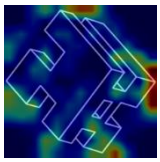
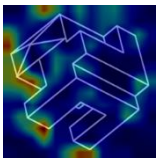
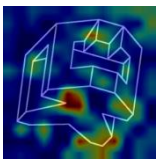
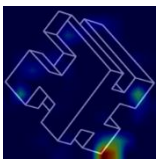
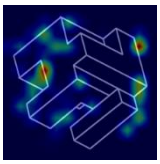
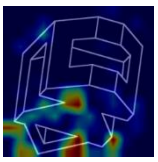
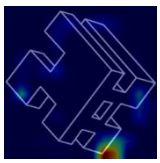
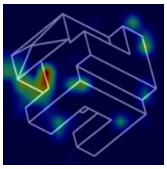
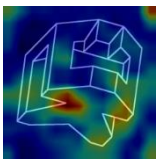
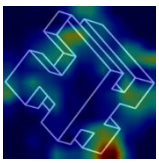
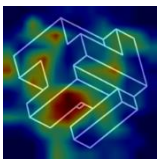
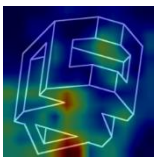
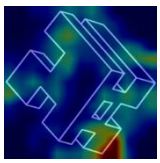
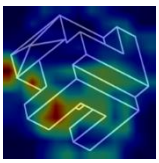
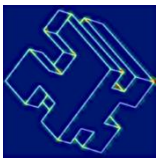
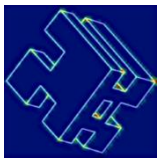
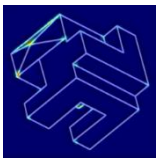
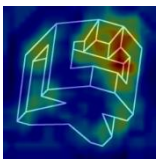
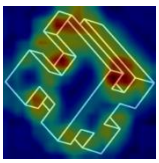
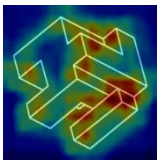
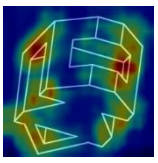
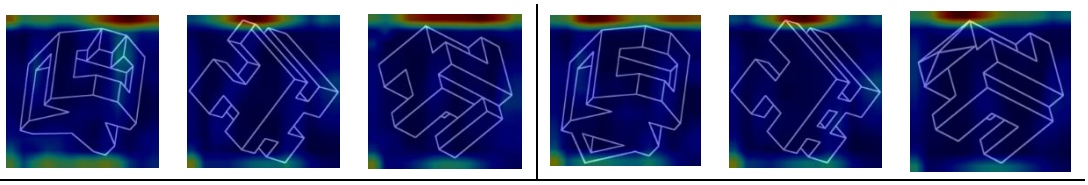
	GoogLeNet (un-trained)
Impossible	
Possible	
	AlexNet (pre-trained)
Impossible	
Possible	
	VGG-11 (pre-trained)
Impossible	
Possible	
	ResNet-18 (pre-trained)
Impossible	
Possible	

Table 5. *Heatmaps for pre-trained and un-trained AlexNet's convolutional layers.*

Pre-trained AlexNet						
	impossible			**possible**		
Layer 1						
Layer 2						
Layer 3						
Layer 4						
Layer 5						
Un-trained AlexNet						
	Impossible			**Possible**		
Layer 1						
Layer 2						
Layer 3						
Layer 4						

Layer 5



The heatmap analysis in Table 6 confirms our initial observation that network classification is not reliably based on impossible parts of the shapes (ROI-impossible heatmap) but is rather biased towards image content outside of the shape bounding contour (ROI-Background heatmap). Critically, this suggests that network analyses during the classification task is not constrained by any representation of object geometry.

Table *6* *Heatmap analysis. This table lists the percentage of the activation falls on the impossible part (ROI-impossible heatmap) and background of possible and impossible shapes (ROI-background heatmap).*

Network	ROI-Impossible heatmap [%]	ROI-Background heatmap	
		Impossible [%]	Possible [%]
Un-trained			
AlexNet	1	94	57
VGG-11	0	100	100
VGG-16	0	100	100
ResNet-18	0	92	29
ResNet-50	0	95	34
GoogLeNet	1	86	24
Pre-trained			
AlexNet	3	84	36
VGG-11	5	82	28
VGG-16	5	39	28
ResNet-18	12	38	38
ResNet-50	10	42	43
GoogLeNet	19	32	32

Supplementary Simulation and Analyses of Network Performance

To further elucidate the determinants of network performance, and to verify the robustness of the results, we ran an additional simulation on a modified dataset. One motivation for this was based on the observation from the heatmap analysis that for some networks there was an apparent bias towards background image properties in the impossible object set. Although this was not consistent across all the tested networks and cannot account for the near chance level of overall network performance, we wanted to rule out a potential confound in the proportion of background area in the image sets between possible and impossible objects as a contributor to network performance.

To examine this, we used the Flood Fill algorithm from the Python Scikit-Image Toolbox to identify the pixel size of the background for each image. We found that there was a significant effect of background area between impossible (M=26051 pixels; SD=2108 pixels) and possible (M=23769 pixels; SD=1960 pixels) images ($t(39)=10.98$, $p<0.001$, $d=0.98$). We then modified the images by applying zooms using the PyTorch augmentation function. This operation reduced the bias of background in impossible (M=25621 pixels; SD=1940 pixels) shapes compared to possible (M=24900 pixels; SD=2031 pixels) shapes ($t(39)=3.54$, $p=0.001$, $d=0.36$). This reduction was significant ($t(39)=5.22$, $p<0.001$, $d=1.03$)[1]. Following this we reran the network tests on this background normalised dataset. The classification results are shown in Table 7. As previously

[1] A reanalysis of the stimulus set used in the human observer validation study showed that Cohen's d = 0.684. This differs from the apparent bias in the datasets used to evaluate network performance because the stimuli were not scaled. To examine whether a bias influenced our findings in the behavioural study we removed possible/impossible shape pairings that showed a particularly large difference in terms of background area. To be more specific, we calculated the SD of background differences and removed all stimuli pairings that showed a difference >1 SD (N=102 augmented stimuli). As a result, the background bias was reduced to 0.392 - which is comparable to the network dataset bias. A reanalysis of human performance with this stimulus subset showed no significant difference between impossible (88.2% accuracy) and possible (85.2% accuracy) objects: $t(23) = 1.06$, $p = 0.299$, $d = 0.35$). This suggests that the original results were not influenced by background area differences.

observed, network performance was very near to chance – and well below the level of performance achieved by the human observers as indicated by the modified t-test. The performance of ResNet-50 (pre-trained) is perhaps notable at 54% (see Table 8 for its confusion matrix, d' and c).

Table 7 *Classification results for supplementary network simulations using the background normalised dataset.*

Network	Train Accuracy [%]	Validation Accuracy [%]	Modified t-test	Train Loss	Validation Loss
Un-trained					
AlexNet	50.1 ± 6.3	51.1 ± 3.6	t=7.14, p<0.001	0.7 ± 0.01	0.69 ± 0.0
VGG-11	50.1 ± 7.3	50.0 ± 0.0	t=7.36, p<0.001	0.69 ± 0.0	0.69 ± 0.0
VGG-16	49.3 ± 4.8	50.0 ± 0.0	t=7.36, p<0.001	0.69 ± 0.0	0.69 ± 0.0
ResNet-18	62.2 ± 4.4	49.1 ± 6.7	t=7.54, p<0.001	0.64 ± 0.03	0.86 ± 0.17
ResNet-50	57.9 ± 5.8	47.7 ± 6.8	t=-7.82, p <0.001	0.69 ± 0.04	0.74 ± 0.04
GoogLeNet	56.4 ± 4.0	52.3 ± 9.2	t=-6.90, p<0.001	1.09 ± 0.03	0.7 ± 0.03
Pre-trained					
AlexNet	56.0 ± 4.9	45.6 ± 5.3	t=824, p<0.001	0.73 ± 0.04	0.77 ± 0.06
VGG-11	57.5 ± 5.6	50.6 ± 6.0	t=7.24, p<0.001	0.8 ± 0.1	0.75 ± 0.05
VGG-16	58.5 ± 5.0	50.6 ± 7.3	t=7.24, p<0.001	0.81 ± 0.17	0.76 ± 0.06
ResNet-18	63.3 ± 6.0	50.2 ± 7.0	t=7.32, p<0.001	0.64 ± 0.04	0.74 ± 0.06
ResNet-50	64.2 ± 4.4	54.5 ± 6.0	t=6.46, p<0.001	0.63 ± 0.03	0.72 ± 0.06
GoogLeNet	59.4 ± 4.0	52.3 ± 5.8	t=6.90, p<0.001	0.66 ± 0.03	0.72 ± 0.03

Table 8 *Confusion matrix for ResNet50 (pre-trained) using the background normalised dataset.*

		ResNet50 (pre-trained) Response [%]	
		impossible	possible
Stimulus Category	impossible	31	19
	possible	16	34
	Total	46	46
	SDT	d': 0.57; c: 0.29	

Heatmap results for this network (see Table 9) suggests that it attempted to use the background again to separate the two classes of images. The network clearly failed to consistently identify local regions of impossibility. This can also be seen for all other networks as shown in Table 10.

Table 9 *Grad-CAM heatmap visualisations representing the degree to which regions in the input image contribute to the correct classification. Red indicates high contributions while blue indicates no contribution.*

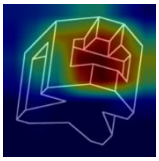
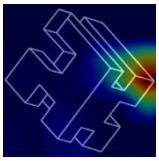
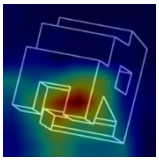
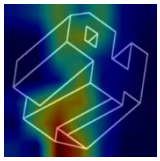
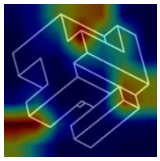
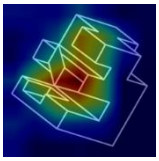
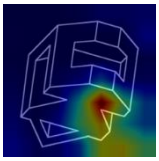
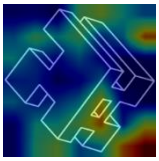
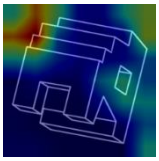
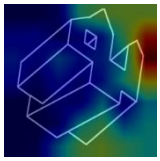
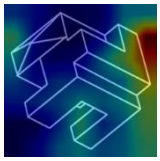
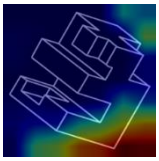
	ResNet-50 (pre-trained)					
impossible						
possible						

Table 10 *Heatmap analysis on normalised dataset. This table lists the percentage of the activation falls on the impossible part (ROI-impossible heatmap) and background of possible and impossible shapes (ROI-background heatmap).*

	ROI-Impossible heatmap [%]	**ROI-Background heatmap**	
		Impossible [%]	**Possible [%]**
Un-trained			
AlexNet	1	94	61
VGG-11	0	100	100
VGG-16	0	100	100
ResNet-18	1	92	34
ResNet-50	1	94	39
GoogLeNet	3	85	28
Pre-trained			
AlexNet	5	83	41
VGG-11	11	83	31
VGG-16	6	37	31
ResNet-18	16	36	38
ResNet-50	16	41	47
GoogLeNet	23	31	37

GENERAL DISCUSSION

We investigated the performance of a range of CNNs (AlexNet, VGG-11, VGG-16, ResNet-18, ResNet-50, GoogLeNet) in a task involving the classification of geometrically possible and impossible objects. The ability to complete this task requires computation of a representational structure of shape that encodes both global and local object geometry because the detection of impossibility derives from an incongruity between well-formed local feature conjunctions and their integration into a geometrically well-formed 3D global shape. Unlike human observers, none of the tested CNNs could reliably discriminate between possible and impossible objects. Detailed analyses using gradient-weighted class activation mapping (GradCam) of CNN image feature processing showed that network classification performance was not constrained by object geometry.

Before discussing the broader implications of these results, we consider some relevant methodological points concerning the dataset. First, one possible argument is that network performance is underestimated because of the relatively small size of the augmented dataset. On this point, it is relevant to note that neither the pretrained nor untrained networks were able to perform the task. Furthermore, human observers can discriminate possible and impossible objects without prior training or experience with these forms of stimuli. Thus, extensive training of networks on the classification task using larger datasets of possible and impossible forms would fundamentally undermine the validity of the human-network performance comparison that we aimed to achieve. Second, the dataset comprised contour-based line drawing objects. We have argued that this class of stimulus provides a strong test of the ability to generate representations of 3D object structure from geometric cues alone. Additionally, human observers are readily able to extract 3D object structure from such stimuli (e.g., Attneave, 1954; Biederman, 1987; Pizlo 2014; see Sayim & Cavangh, 2011, for a recent review). Thus, the use of this stimulus class provides a strong test of the capability of CNNs to generate representations of 3D object geometry.

Taken together, these results have important implications for our understanding of convolutional neural networks and their suitability as models for image classification in biological vision systems. Evidence from studies of human performance suggests that the extraction of visual information about both local and global image properties, as well as the integration of this information at intermediate levels of perceptual representation, is characteristic of human vision (e.g., Bar, 2003; Bar et al., 2006; Bullier, 2001; Leek et al., 2016; Oliver et al., 2018; Peyrin et al., 2010). For example, Bar (2003; Bar et al., 2006) has shown that image classification in human vision is constrained by parallel analyses of object shape as coarse (global) and fine (local) spatial scales which are mediated by distinct neural pathways. This is supported by other recent work using high-density ERPs showing evidence of parallel processing of local and global object structure (Leek, Roberts, Oliver, et al, 2016; Oliver, Cristino, Roberts et al, 2018). We have demonstrated, across a broad range of network architectures, that CNNs are unable to discriminate possible and impossible objects based on object geometry alone. We hypothesised that sensitivity to object impossibility necessitates a level of object shape processing in which local shape features are integrated with a representation of global 3D shape geometry. Thus, our results suggest that object processing in CNNs, unlike in human vision, is not constrained by representations of local and global object geometry.

The results add to a growing body evidence highlighting important differences between CNNs and the human visual system. Other recent work has also demonstrated that CNNs can fail to mimic human abilities in ways which suggest fundamental differences in processing architecture between the networks and the biological system. This finding is consistent with other recent studies of CNNs demonstrating their reliance on local image features in classification (e.g., Baker et al., 2020; 2018; Brendel & Bethge, 2019; Geirhos et al. 2019). For example, as noted earlier, Baker et al. (2018) examined the performance of two pretrained CNN architectures (VGG19 and AlexNet) in their ability to classify images of objects with either congruent or incongruent (e.g., mixed) global shape and local textures. The results showed that network performance was biased towards classification based on

local but not global image properties (see also Geirhos et al., 2019). Other important evidence comes from network performance under conditions of adversarial attack in which pre-trained networks can be shown to make classification decisions that human observers do not make (e.g., Nguyen, Yosinski, & Clune, 2015; Szegedy, et al., 2014; see also Zhang, Liu, Suen, 2020; for a recent review). We propose here that the failure of networks to learn possible-impossible image classification, and their (hyper)sensitivity to local feature perturbation in adversarial examples, derives from the absence of an explicit representation of 3D object geometry. A future grand challenge will be to explore whether other combinations of CNN architectures, and processing parameters, will be more successful. For instance, whether modification of filter sizes, and the incorporation of short- and long-range recurrent connections, may provide a means to capture and integrate both local and global shape structure. One promising line of development is dual pathway architectures in which processing of image content is constrained by parallel analyses across multi spatial scales – and which take some inspiration from neurobiological models of human vision (e.g., Bar, 2003; Bar et al., 2006; Mishkin & Ungerleider, 1983; Milner & Goodale, 2006). Some recent examples of such architectures include SAIM (Selective Attention for Identification model, e.g., Abadi et al., 2019; Narbutas et al., 2017; Heinke & Humphreys, 2003), and NAM (Naming and Action model; Yoon et al., 2002) and CoRLEGO (a model of reaching; Strauss, Woodgate, Sami, & Heinke, 2015).

It is worth noting that, compared to CNNs, the design of these architecture takes a very different approach. Here a theoretical framework informs the architecture's structure and mechanisms. By and large, this approach follows the traditional method commonly used in natural sciences (see Farrell & Lewandowsky, 2018; Mavritsaki et al., 2011; Heinke, 2009; for reviews). In contrast, in the CNN approach the implemented operations are determined through a combination of architectural constrains and training material. While the architecture is often loosely inspired by theories about biological structures, the processing is not informed by conceptual frameworks, but by

the training material provided by the user. Future work will need to compare these approaches and evaluate which is better at advancing our understanding of biological vision.

In summary, we tested the hypothesis that, unlike the human system, CNNs do not compute representations of global and local object geometry during image classification. To do so we trained and tested six CNNs (AlexNet, VGG-11, VGG-16, ResNet-18, ResNet-50, GoogLeNet), and human observers, to discriminate geometrically possible and impossible objects. The ability to complete this task requires computation of a representational structure of shape that encodes both global and local object geometry because the detection of impossibility derives from an incongruity between well-formed local feature conjunctions and their integration into a geometrically well-formed 3D global shape. Unlike human observers, none of the tested CNNs could reliably discriminate between possible and impossible objects. Detailed Grad-Cam analyses of CNN image feature processing showed that network classification performance was not constrained by object geometry. We argue that these findings reflect fundamental differences between CNNs and human vision in terms of underlying image processing structure. Notably, unlike human vision, CNNs do not compute representations of object geometry. The results challenge the plausibility of CNNs as a framework for understanding image classification in biological vision systems.

REFERENCES

Abadi, A. K., Yahya, K., Amini, M., Friston, K., & Heinke, D. (2019). Excitatory versus inhibitory feedback in Bayesian formulations of scene construction. Journal of The Royal Society Interface, 16(154), https://doi.org/10.1098/rsif.2018.0344

Attneave, F. (1954). Some informational aspects of visual perception. Psychological Review, 61(3), 183–193. https://doi.org/10.1037/h0054663

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

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).

Bar, M. (2003). A cortical mechanism for triggering top-down facilitation in visual object identification. Journal of Cognitive Neuroscience, 15, 600-609.

Bar, M., Kassam, K.S., Ghuman, A.S., Boshyan, J, Schmid, A.M., Dale, A.M., Hamalainen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R. & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, 103, 449-454.

Beaucousin, V., Simon, G., Cassotti, M., Pineau, A., Houdé, O. & Poirel, N (2013). Global interference during early visual processing: ERP evidence from a rapid global/local selection task. Frontier in Psychology, 4, 1-6: doi: 10.3389/fpsyg.2013.00539

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. Psychological Review, 94(2), 115–147. https://doi.org/10.1037/0033-295X.94.2.115

Bullier, J. (2001). Integrated model of visual processing. Brain Research Reviews, 36, 96-107.

Brendel, W., & Bethge, M. (2019). Approximating CNNs with Bag-of-local-Features models works surprisingly well on ImageNet. Retrieved from http://arxiv.org/abs/1904.00760.

Carrasco, M., & Seamon, J. G. (1996). Priming impossible figures in the object decision test: The critical importance of perceived stimulus complexity. Psychonomic Bulletin & Review, 3(3), 344–351. https://doi.org/10.3758/BF03210758.

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. https://doi.org/10.1038/srep27755.

Cooper, L. A., Schacter, D. L., Ballesteros, S., & Moore, C. (1992). Priming and recognition of transformed three-dimensional objects: Effects of size and reflection. Journal of Experimental Psychology: Learning, Memory, and Cognition, 18(1), 43–57. https://doi.org/10.1037/0278-7393.18.1.43

Cox, D. D., & Dean, T. (2014, September 22). Neural networks and neuroscience-inspired computer vision. Current Biology, Vol. 24, pp. R921–R929. https://doi.org/10.1016/j.cub.2014.08.026

Crawford, J.R., Garthwaite, P.H. & Porter, S. (2010). Point and interval estimates of effect sizes for the case-controls design in neuropsychology: Rationale, methods, implementations, and proposed reporting standards. Cognitive Neuropsychology, 27, 245-260.

Crawford, J.R. & Garthwaite, P.H. (2002). Investigation of the single case in neuropsychology: Confidence limits on the abnormality of test scores and test score differences. Neuropsychologia, 40, 1196-1208.

Davitt, L.I., Cristino, F., Wong, A.C.N. & Leek, E.C. (2014). Shape information mediating basic-and subordinate-level object recognition revealed by analyses of eye movements. Journal of Experimental Psychology: Human Perception and Performance 40 (2), 451-456.

Deco, G., & Heinke, D. (2007) Attention and Spatial Resolution: A theoretical and experimental study of visual search in hierarchical patterns. Perception, 36(3), 335-354

Deng J, Dong W, Socher R, Li LJ, Li K, Fei-Fei L. (2009) Imagenet: A large-scale hierarchical image database. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on 2009 Jun 20(pp. 248–255).

Farrell, S., & Lewandowsky, S. (2018). Computational Modeling of Cognition and Behavior. Cambridge University Press. https://doi.org/10.1017/cbo9781316272503

Freud, E., Avidan, G., & Ganel, T. (2013). Holistic processing of impossible objects: Evidence from Garner's speeded-classification task. Vision Research, 93, 10–18. https://doi.org/10.1016/J.VISRES.2013.10.001

Freud, E., Ganel, T., & Avidan, G. (2015). Impossible expectations: fMRI adaptation in the lateral occipital complex (LOC) is modulated by the statistical regularities of 3D structural information. NeuroImage, 122, 188–194. https://doi.org/10.1016/J.NEUROIMAGE.2015.07.085

Freud, E., Ganel, T., Shelef, I., Hammer, M. D., Avidan, G., & Behrmann, M. (2017). Three-Dimensional Representations of Objects in Dorsal Cortex are Dissociable from Those in Ventral Cortex. Cerebral Cortex, 27(1), 422–434. https://doi.org/10.1093/cercor/bhv229

Freud, E., Hadad, B.-S., Avidan, G., & Ganel, T. (2015). Evidence for similar early but not late representation of possible and impossible objects. Frontiers in Psychology, 6, 94. https://doi.org/10.3389/fpsyg.2015.00094

Heaton, J. (2017). Goodfellow, I., Bengio, Y., & Courville, A.: Deep learning. Genetic Programming and Evolvable Machines, 19(1–2), 305–307. https://doi.org/10.1007/s10710-017-9314-z

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, abs/1811.12231.

Güçlü, U., & van Gerven, M. A. J. (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

Guo, Y., Liu, Y., Oerlemans, A., Lao, S., Wu, S., & Lew, M. S. (2016). Deep learning for visual understanding: A review. Neurocomputing, 187, 27–48. https://doi.org/10.1016/J.NEUCOM.2015.09.116.

Harris, I.M., Dux, P.E., Benito, C.T. & Leek, E.C. (2008). Orientation sensitivity at different stages of object processing: Evidence from repetition priming and naming. PLoS One 3 (5)

Han, S., He, X. & Woods, D.L. (2000). Hierarchical processing and level-repetition effect as indexed by early brain potentials. Psychophysiology, 37, 817-830.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770–778. https://doi.org/10.1109/CVPR.2016.90

Heinke, D., & Humphreys, G. W. (2003). Attention, spatial representation and visual neglect: Simulating emergent attention and spatial memory in the Selective Attention for Identification Model (SAIM). Psychological Review, 110(1):29-87.

Heinke, D. (2009). Computational modelling in behavioural neuroscience: Methodologies and Approaches - Minutes of discussions at the workshop in Birmingham, UK in May 2007. In Heinke, D. & Mavritsaki, E. (Eds.) (2009) Computational Modelling in Behavioural Neuroscience: Closing the gap between neurophysiology and behaviour, London: Psychology Press.

Hinz, T., Navarro-Guerrero, N., Magg, S., & Wermter, S. (2018). Speeding up the Hyperparameter Optimization of Deep Convolutional Neural Networks. International Journal of Computational Intelligence and Applications, 17(2). https://doi.org/10.1142/S1469026818500086.

Keskar, N.S., & Socher, R. (2017). Improving Generalization Performance by Switching from Adam to SGD. ArXiv, abs/1712.07628.

Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology, 10(11).

Kriegeskorte, N. (2015) Deep neural networks: A new framework for modeling biological vision and brain information processing. Annual Review of Vision Science, 1(15):417–446, 2015.

Kingma, D. P., & Ba, J. L. (2015). Adam: A method for stochastic optimization. 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. International Conference on Learning Representations, ICLR.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. Communications of the ACM, 60(6), 84–90. https://doi.org/10.1145/3065386

Kuzovkin, I., Vicente, R., Petton, M., Lachaux, J. P., Baciu, M., Kahane, P., … Aru, J. (2018). Activations of deep convolutional neural networks are aligned with gamma band activity of human visual cortex. Communications Biology, 1(1). https://doi.org/10.1038/s42003-018-0110-y

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539.

Leek, EC., Roberts, M., Oliver, ZJ., Cristino, F. & Pegna, AJ. (2016). Early differential sensitivity of evoked-potentials to local and global shape during the perception of three-dimensional objects. Neuropsychologia, 89, 495-509.

Leek, E.C., Roberts, M., Dundon, N.M. & Pegna, A.J. (2018). Early sensitivity of evoked potentials to surface and volumetric structure during the visual perception of three-dimensional object shape. European Journal of Neuroscience, doi.org/10.1111/ejn.14270

Leek, E.C., Patterson, C., Paul. M.A., Rafal, R. & Cristino, F. (2012). Eye movement patterns during object recognition in visual agnosia. Neuropsychologia, 50 (9), 2142-2153.

Leek, E.C., Reppa, I., Rodriguez, E. & Arguin, M. (2009). Surface but not volumetric part structure mediates three-dimensional shape representation: Evidence from part-whole priming. Quarterly Journal of Experimental Psychology, 62 (4), 814-830.

Leek, E.C., Atherton, C.J. & Thierry, G. (2007). Computational mechanisms of object constancy for visual recognition revealed by event-related potentials. Vision Research 47 (5), 706-713

Leek, E.C., Reppa, I. & Arguin, M. (2005). The structure of three-dimensional object representations in human vision: Evidence from whole-part matching. Journal of Experimental Psychology: Human Perception and Performance, 31 (4), 668-684.

Macmillan, N. A. & Creelman, C. D. (1991) "Detection Theory: A User's guide" Cambridge University Press.

Masters, D., & Luschi, C. (2018). Revisiting Small Batch Training for Deep Neural Networks. ArXiv, abs/1804.07612.

Mavritsaki, E., Heinke, D., Allen H., Deco, G., & Humphreys, G. W. (2011) Bridging the gap between physiology and behavior: Evidence from the sSoTS model of human visual attention. Psychological Review, 118(1), 3-41.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. Trends in Neurosciences, 6, 414–417. https://doi.org/10.1016/0166-2236(83)90190-x

Narbutas, V., Lin, Y.-S., Kristan, M., & Heinke, D. (2017) Serial versus parallel search: A model comparison approach based on reaction time distributions. Visual Cognition, 1-3, 306-325. https://doi.org/10.1080/13506285.2017.1352055

Navon, D. (1977). Forest before trees: The precedence of global feature in visual perception. Cognitive Psychology, 9, 353-383.

Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 427-436.

Oliver, Z., Cristino, F., Roberts, M.V., Pegna, A.J. & Leek, E.C. (2018). Stereo viewing modulates three-dimensional shape processing during object recognition: A high-density ERP study. Journal of Experimental Psychology: Human Perception and Performance, 44(4), 518-534.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... & Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. arXiv preprint arXiv:1912.01703.

Penrose, L. S., & Penrose, R. (1958). Impossible Objects: A Special Type of Visual Illusion. British Journal of Psychology, 49(1), 31–33. https://doi.org/10.1111/j.2044-8295.1958.tb00634.x

Peyrin, C., Baciu, M., Segebarth, C. & Marendaz, C. (2004). Cerebral regions and hemispheric specialization for processing spatial frequencies during natural scene recognition: An event-related fMRI study. Neuroimage, 23, 698-707.

Picard, R. R. & Cook, R. D. (1984) Cross-Validation of Regression Models. J. Am. Stat. Assoc., 387, 575–583.

Pizlo, Z. (2014). Making a Machine that Sees Like Us. United Kingdom: Oxford University Press.

Proverbio, A.M., Minniti, A. & Zani, A. (1998). Electrophysiological evidence of a perceptual precedence of global vs. local visual information. Brain Research, 6, 321-334.

Ratcliff, R., & McKoon, G. (1995). Bias in the priming of object decisions. Journal of Experimental Psychology. Learning, Memory, and Cognition, 21(3), 754–767. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/7602269

Reddi, S. J., Kale, S., & Kumar, S. (2018). On the convergence of Adam and beyond. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. International Conference on Learning Representations, ICLR.

Reppa, I., Greville, W.J. & Leek, E.C. (2015). The role of surface-based representations of shape in visual object recognition. Quarterly Journal of Experimental Psychology, 68 (12), 2351-2369.

Robertson, L.C. & Lamb, M.R. (1991). Neuropsychological contributions to theories of part/whole organisation. Cognitive Psychology, 23, 299-330.

Robertson, L.C., Lamb, M.R. & Knight, R.T. (1988). Effects of lesions of temporal-parietal junction on perceptual and attentional processing in humans. Journal of Neuroscience, 8, 3757-3769.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., … Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. International Journal of Computer Vision, 115(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Schacter, D. L., Cooper, L. A., & Delaney, S. M. (1990). Implicit memory for unfamiliar objects depends on access to structural descriptions. Journal of Experimental Psychology. General, 119(1), 5–24. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2141064

Schacter, D. L., Cooper, L. A., Delaney, S. M., Peterson, M. A., & Tharan, M. (1991). Implicit memory for possible and impossible objects: constraints on the construction of structural descriptions. Journal of Experimental Psychology. Learning, Memory, and Cognition, 17(1), 3–19. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1826731.

Sayim, B., & Cavanagh, P. (2011). What Line Drawings Reveal About the Visual Brain. Frontiers in Human Neuroscience, 5. https://doi.org/10.3389/fnhum.2011.00118

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2019). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization. International Journal of Computer Vision, 128(2), 336–359. https://doi.org/10.1007/s11263-019-01228-7

Simonyan, K., & Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. CoRR, abs/1409.1556.

Sharma, A., van Rijn, J. N., Hutter, F., & Müller, A. (2019). Hyperparameter Importance for Image Classification by Residual Neural Networks. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11828 LNAI, 112–126. https://doi.org/10.1007/978-3-030-33778-0_10

Smith, S. L., & Le, Q. V. (2017). A Bayesian Perspective on Generalization and Stochastic Gradient Descent. 6th International Conference on Learning Representations, ICLR 2018 - Conference Track Proceedings. Retrieved from http://arxiv.org/abs/1710.06451

Strauss, S., Woodgate, P.J.W., Sami, S. A., & Heinke, D. (2015) Choice reaching with a LEGO arm robot (CoRLEGO): The motor system guides visual attention to movement-relevant information. Neural Networks, 72, 3-12. http://dx.doi.org/10.1016/j.neunet.2015.10.005

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 07-12-June-2015, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Vuilleumier, P. (2010). The neural substrates and timing of top-down processes during coarse-to-fine categorization of visual scenes: A combined fMRI and ERP study. Journal of Cognitive Neuroscience, 22, 2768-2780.

Voulodimos, A., Doulamis, N., Doulamis, A., & Protopapadakis, E. (2018). Deep Learning for Computer Vision: A Brief Review. Computational Intelligence and Neuroscience, 2018, 1–13. https://doi.org/10.1155/2018/7068349

Wagner, J. M. Kohler, T. Gindele, L. Hetzel, J. T. Wiedemer, and S. Behnke. Interpretable and fine-grained visual explanations for convolutional neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 9097–9107, 2019.

Wilson, A. C., Roelofs, R., Stern, M., Srebro, N., & Recht, B. (2017). The Marginal Value of Adaptive Gradient Methods in Machine Learning. Advances in Neural Information Processing Systems.

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 of the United States of America, 111(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111.

Yoon, E. Y., Heinke, D., & Humphreys, G. W. (2002). Modelling direct perceptual constraints on action selection: The Naming and Action model (NAM). Visual Cognition, 9(4/5):615-661.

Zhang, X.-Y., Liu, C.-L., & Suen, C. Y. (2020). Towards Robust Pattern Recognition: A Review. Proceedings of the IEEE, 108(6), 894–922. https://doi.org/10.1109/jproc.2020.2989782

The following authors have contributed to this manuscript:

Dietmar Heinke, Peter Wachman, Wieske van Zoest, E. Charles Leek

Dietmar Heinke drafted and revised the manuscript and conducted some of the data analysis.

Peter Wachman conducted the study with human observers, and trained and analysed the CNNs

Wieske van Zoest design the study with human observers and helped with the data analysis.

E. Charles Leek conceived the idea, supplied the stimuli and helped with the revision of the manuscript.