Depth in convolutional neural networks solves scene segmentation

Seijdel, N.; Tsakmakidis, N.; De Haan, E.H.F.; Bohte, S.M.; Scholte, H.S.

DOI
10.1101/2019.12.16.877753

Publication date
2019

Document Version
Submitted manuscript

Depth in convolutional neural networks solves scene segmentation

Seijdel, N [1,2], Tsakmakidis, N[3], De Haan, EHF [1,2], Bohte, SM[3], Scholte, HS [1,2]

[1] Department of Psychology, University of Amsterdam, Amsterdam, The Netherlands
[2] Amsterdam Brain & Cognition (ABC) Center, University of Amsterdam, Amsterdam, The Netherlands
[3] Machine Learning Group, Centrum Wiskunde & Informatica, Amsterdam, the Netherlands

correspondence: n.seijdel@uva.nl

Abstract

Feedforward deep convolutional neural networks (DCNNs) are, under specific conditions, matching and even surpassing human performance in object recognition in natural scenes. This performance suggests that the analysis of a loose collection of image features could support the recognition of natural object categories, without dedicated systems to solve specific visual subtasks. Research in humans however suggests that while feedforward activity may suffice for sparse scenes with isolated objects, additional visual operations ('routines') that aid the recognition process (e.g. segmentation or grouping) are needed for more complex scenes. Linking human visual processing to performance of DCNNs with increasing depth, we here explored if, how, and when object information is differentiated from the backgrounds they appear on. To this end, we controlled the information in both objects and backgrounds, as well as the relationship between them by adding noise, manipulating background congruence and systematically occluding parts of the image. Results indicate that with an increase in network depth, there is an increase in the distinction between object- and background information. For more shallow networks, results indicated a benefit of training on segmented objects. Overall, these results indicate that, de facto, scene segmentation can be performed by a network of sufficient depth. We conclude that the human brain could perform scene segmentation in the context of object identification without an explicit mechanism, by selecting or “binding” features that belong to the object and ignoring other features, in a manner similar to a very deep convolutional neural network.

Introduction

Visual object recognition is so swift and efficient that it has been suggested that a fast feed-forward sweep of perceptual activity is sufficient to perform the task [1–3]. Disruption of visual processing beyond feed-forward stages (e.g. >220 ms after stimulus onset, or after activation of higher order areas) can however lead to decreased object recognition performance [4,5], and a multitude of recent findings suggest that while feed-forward activity may suffice to recognize isolated objects that are easy to discern, the brain employs increasing feedback or recurrent processing for object recognition under more 'challenging' natural conditions [6–10]. When performing a visual object recognition task, the visual input (stimulus) elicits a feed-forward drive that rapidly extracts basic image features through feedforward connections [11]. For sparse scenes with isolated objects, this set of features appears to be enough for successful recognition. For more complex scenes, however, the jumble of visual information ('clutter') may be so great that object recognition cannot rely on having access to a conclusive set of features. For those images, extra visual operations ('visual routines'), such as scene segmentation and perceptual grouping, requiring several iterations of modulations and refinement of the feedforward activity in the same and higher visual areas, might be necessary [11–13].

While this view emphasises that object recognition relies on the integration of features that belong to the object, many studies have shown that features from the background can also influence the recognition process [14–21]. For example, objects appearing in a familiar context are detected more accurately and quickly than objects in an unfamiliar environment, and many computational models of object recognition (in both human and computer vision), use features both from within the object and from the background [22–24]. This shows that when subjects recognise an object, figure-ground segmentation has not always occurred completely.

One way to understand how the human visual system processes information involves building computational models that account for human-level performance under different conditions. Here we investigate Deep Convolutional Neural Networks (DCNNs). DCNNs are being studied often because they show remarkable performance on both object and scene recognition, rivaling human performance. Recent evidence shows that the depth of DCNNs is of crucial importance for this recognition performance [25]. In addition to better performance, deeper networks have also been shown to be more human-like (making errors similar to human subjects; [26]). More layers seem especially important when scenes are

more difficult or challenging, e.g. because of occlusion, variation, or blurring, where elaborate processing is required [8,10]. The very deep residual networks used in current object recognition tasks are nearly equivalent to a recurrent neural network unfolding over time, when the weights between their hidden layers are clamped [27]. This has led to the hypothesis that the additional layers function in a way that is similar to recurrent processing in the human visual system, and that these additional layers are solving the challenges that are resolved by recurrent computations in the brain.

In the current study, we explore how the number of layers (depth) in a DCNN relates to human vision and how depth influences to what degree object segmentation occurs. While we certainly do not aim to claim that DCNNs are identical to the human brain, we argue that they can be studied in a way similar to the way in which we use animal models (DNimals; [28]). We use Residual Networks (ResNets; [29]) as a method to systematically manipulate network depth because this type of network consists of a limited number of fixed components that can be up-scaled without altering the architecture in another way. First, we focused on the question to what extent DCNNs exhibit the same sensitivity to scene properties (object context) as human participants. To this end, we presented seven DCNNs with an increasing number of layers and 40 human participants with images of objects that were either presented on a uniform background (segmented), or on top of congruent or incongruent scenes and evaluated their performance. Additionally, for the DCNNs, we controlled the amount of information in the objects and backgrounds, as well as the relationship between them by adding noise or systematically occluding parts of the image. Next, we investigated the role of segmentation on learning ('training'), by training the DCNNs on either segmented or unsegmented objects.

A convergence of results indicated a lower degree of segregation between object- and background features in more shallow networks, compared to deeper networks. This was confirmed by the observation that more shallow networks benefit more from training on pre-segmented objects than deeper networks. Overall, deeper networks seem to perform implicit 'segmentation' of the objects from their background, by improved selection of relevant features.

Results

Experiment 1: scene segmentation + background consistency effect

Human performance

Accuracy (percentage correct) was computed for each participant. A repeated-measures ANOVA, with factor background (segmented, congruent, incongruent) differentiated accuracy across the three conditions, $F(2,74) = 366.2$, $p < .001$, $\eta^{2par} = .91$ (Figure 1D). Participants made fewer errors for segmented objects, than the congruent, $t(37) = 15.655$, $p < .001$, and incongruent condition, $t(37) = 27.6$, $p < .001$. Additionally, participants made fewer errors for congruent than incongruent, $t(37) = 9.376$, $p < .001$. Bonferroni correction at $\alpha = 0.05$ was used for all comparisons, p-values reported are the adjusted p-values. Overall, results indicate that when a scene is glanced briefly (32 ms, followed by a mask), the objects are not completely segregated from their background and semantic consistency information influences object perception.

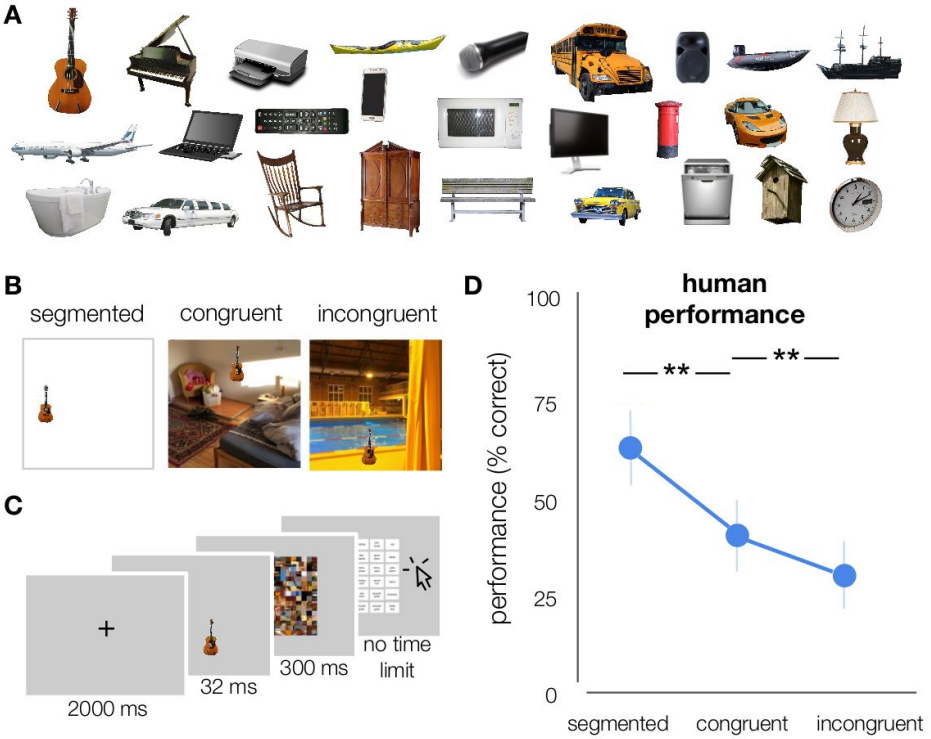


Figure 1. Stimuli and experimental design. A) Examples of the different object categories (examples shown here were not part of the actual stimulus set). 27 object categories were used in this experiment (subordinate level, based on ImageNet categories). In total, each category contained 10 exemplars. **B)** Stimuli were generated by placing the objects onto white, congruent and incongruent backgrounds (512*512 pixels, full-color). Backgrounds were sampled from the SUN2012 database [4]. For human participants, objects were downsized and placed in one of nine possible locations (3x3 grid). For DCNNs, objects were bigger and placed centrally. **C)** Participants performed on an object recognition task. At the beginning of each trial, a fixation-cross was presented in the center of the screen for 2000 ms, followed by an image. Images were presented in randomized sequence, for a duration of 32 ms, followed by a mask, presented for 300 ms. After the mask, participants had to indicate which object they saw, by clicking on one of 27 options on screen using the mouse.

After 81 (⅓) and 162 (⅔) trials, there was a short break. **D)** Human performance (% correct) on the object recognition task.

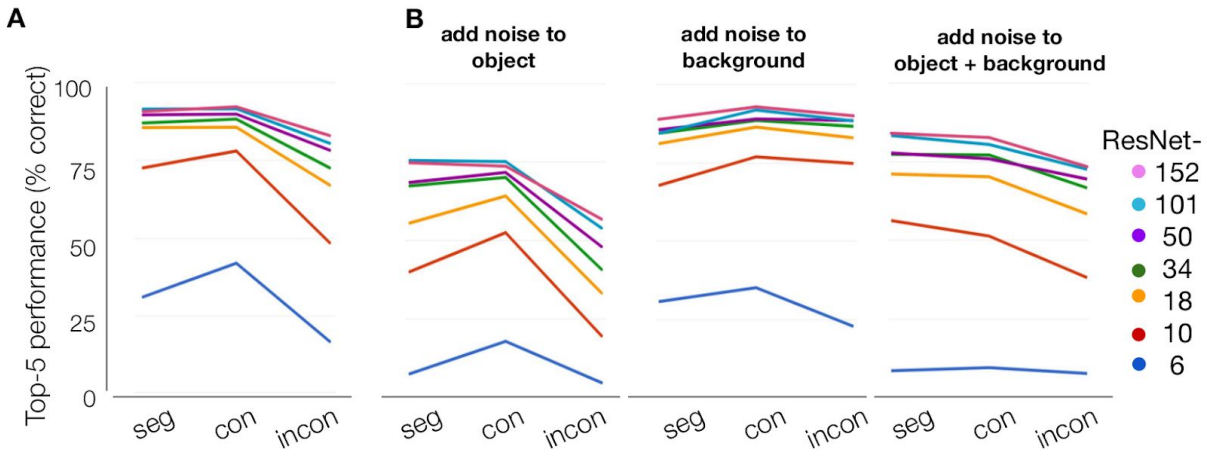


Figure 2. DCNN performance on the object recognition task. A) DCNN performance on the object recognition task. For shallow networks, performance is better for the congruent than for the incongruent condition. This decrease is most prominent for ResNet-10, and gets smaller as the models get deeper. For 'ultra-deep' networks it matters little if the background is congruent, incongruent or even present. **B)** DCNN performance on the object recognition task after adding noise to the object, the background, or both.

Model performance

For human participants, results indicated that (at a first glance) features from the background influenced object perception. Do DCNNs show a similar pattern and how is this influenced by network depth? To answer these questions, we presented the ResNets with images of objects from 27 object categories, presented segmented or on top of a (for that category) congruent or incongruent scene. Results indicated both a substantial overlap and difference in performance between human participants and DCNNs. Both were better in recognizing an object on a congruent versus an incongruent background. However, whereas human participants performed best in the segmented condition, DCNNs performed equally well (or better) for the congruent condition. Performance for the incongruent condition was lowest. This effect was particularly strong for more shallow networks (ResNet-6[1], ResNet-10), and got smaller as the networks get deeper. For 'ultra-deep' networks it mattered little if the background was congruent, incongruent or even present, behavior that humans also exhibit when these images are shown unmasked. Remarkably, performance of the most shallow networks was better for the congruent condition compared to the segmented condition. These results suggest that parts of covarying backgrounds or surroundings influence the categorization of the objects. In other words, there is 'leakage' of the natural (congruent) background in the features for classification, predominantly for the more shallow networks. For object recognition in a

[1] Technically, this network is a "Net6", because we removed the residual connection to degrade the networks' performance.

congruent scene this is not necessarily a problem, and can even increase or facilitate performance (as compared to the segmented condition). For objects on an incongruent background, however, this impairs classification performance. These results suggest that one of the ways in which network depth improves object classification, is by learning how to select the features that belong to the object, and thereby implicitly segregating the object features from the other parts of the scene.

To confirm this hypothesis, and to gain more insight into the importance of the features in the object vs. the background, gaussian noise was added to either the object, the background, or both (Figure 2). When noise was added to the complete image (object included), performance decreased for all conditions and all networks. When noise was added to the object only, classification performance also decreased for all conditions Crucially, this decrease was modest for the congruent and particularly severe for the incongruent condition. This indicates that for the congruent condition, also in the no noise manipulation, performance is heavily depended on the background for classification. The other side of this conclusion, that in the incongruent condition the features in the background interfere with object classification, is confirmed by the observation that this condition improves when noise is added to the background.

To further investigate the degree to which the networks are using features from the object and/or background for classification, we systematically occluded different parts of the input image by sliding a gray patch (of either 64*64, 128*128 or 256*256 pixels) across the image in 30 pixel steps. We then evaluated the changes in activation of the correct class after occlusion of the different image parts, before the softmax activation function (compared to activation for the 'original' unoccluded image). We reasoned that, if the activity in the feature map changed after occluding a patch of the image, that those pixels were important for classification. For this analysis, positive values indicate that pixels are helping classification, with higher values indicating a higher importance. This reveals the features to be far from random, uninterpretable patterns. For example, in Figure 3, results clearly show that the network is localizing the object within the scene, as the activity in the feature map drops significantly when the object (china cabinet in this example) is occluded. To evaluate whether deeper networks are better at localizing the objects in the scene, while ignoring irrelevant background information, we quantified the importance of features in the object vs. background by averaging the change in the feature map across pixels belonging to either the object or the background. Because performance of ResNet-6 for the ‘original’

unoccluded images was already exceptionally low, the averaged interference was hard to interpret and remained low, due to many near-zero values in the data. Apart from ResNet-6, results indicated a larger influence of background pixels on classification for more shallow networks, for all conditions. For those models, pixels from the object had a larger impact as well, for the segmented and congruent condition.

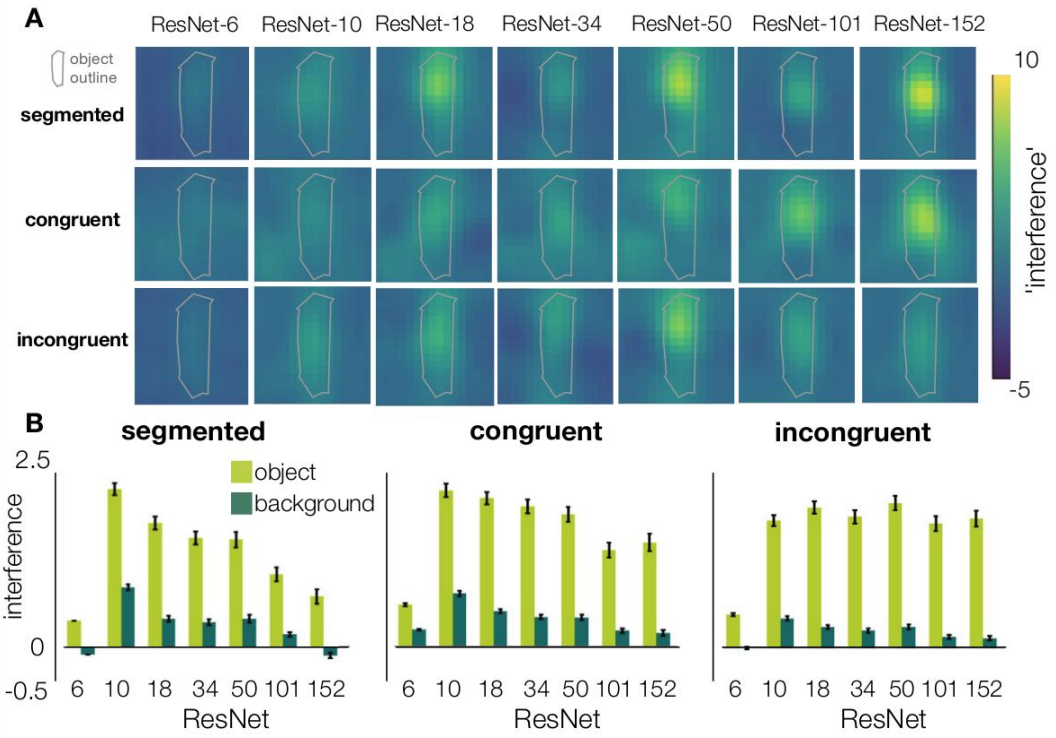


Figure 3. Systematic occlusion of parts of the image. A) Examples where we occluded different portions of the scene, and visualize how the classifier output for the correct class changed (before the softmax activation function). Images were occluded by a gray patch of 128x128 pixels, sliding across the image in 30 pixel steps. **B)** The relative change in activation (compared to the original image), after occluding pixels of either the object or the background, for the different conditions (segmented, congruent, incongruent). Error bars represent 1 SEM.

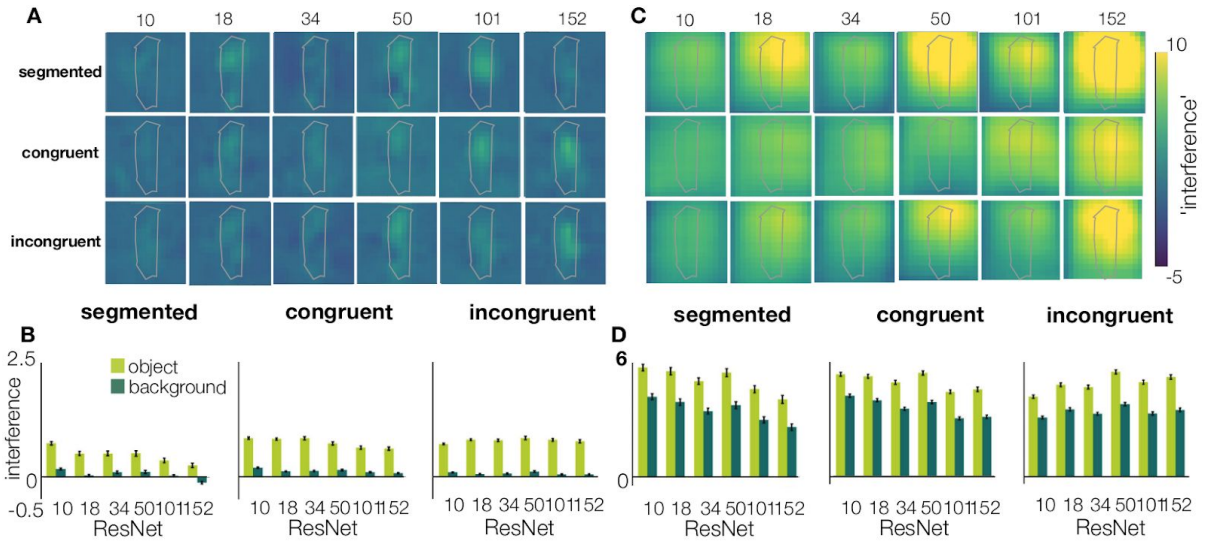


Figure 4. Analysis repeated with a smaller (64x64) and larger (256x256) patch. A) visualization of the change in classifier output for the correct class, before the softmax activation function after occlusion by a 64x64 patch, sliding across the image in 30 pixel steps. **B**) The relative change in activation (compared to the original image), after occluding pixels of either the object or the background, for the different conditions (segmented, congruent, incongruent). Error bars represent 1 SEM. **C,D)** Repeated for a large patch (256x256 pixels).

Next, we tested how training was influenced by network depth. If deeper networks indeed implicitly learn to segment object from background, we expect them to show a smaller difference in learning speed, when trained with segmented vs. unsegmented stimuli (as compared to shallow networks).

Experiment 2: training on unsegmented/segmented objects

Experiment 1 indicated that, when trained on ImageNet, the networks are influenced by visual information from both the object and the background region. In experiment 2, we investigated the influence of background on classification performance when the networks are trained on visual information from the object region only. To do so, we trained four networks (ResNet6, ResNet-10, ResNet-18, ResNet-34) on a dataset with objects that were already segmented, and on a dataset in which they were unsegmented (i.e. objects embedded in the scene). We used more shallow networks and fewer object classes to reduce computation time.

Accuracy of the ResNets was evaluated after each epoch (100 in total) on the validation sets. Results indicated a higher classification accuracy in the early stages of training for the networks trained on segmented objects compared to the networks trained on unsegmented objects (Figure 5). Additionally, these models converged in less epochs. In the later stages, accuracy between the two types of models (trained on unsegmented vs. segmented) was

similar. Results also indicated a difference between the more shallow networks (ResNet-6), where there is a significant difference in accuracy between segmented and unsegmented objects for all training epochs, and the deeper networks. For the deeper networks, the difference in accuracy quickly diminishes and finally disappears. Shallow networks trained on segmented stimuli also converged (stabilized) earlier than when they were trained on unsegmented images. The difference in epochs until convergence (between segmented and unsegmented) decreased with an increase in network depth. Deeper networks thus seem to learn to 'segment' the objects from their background during training.

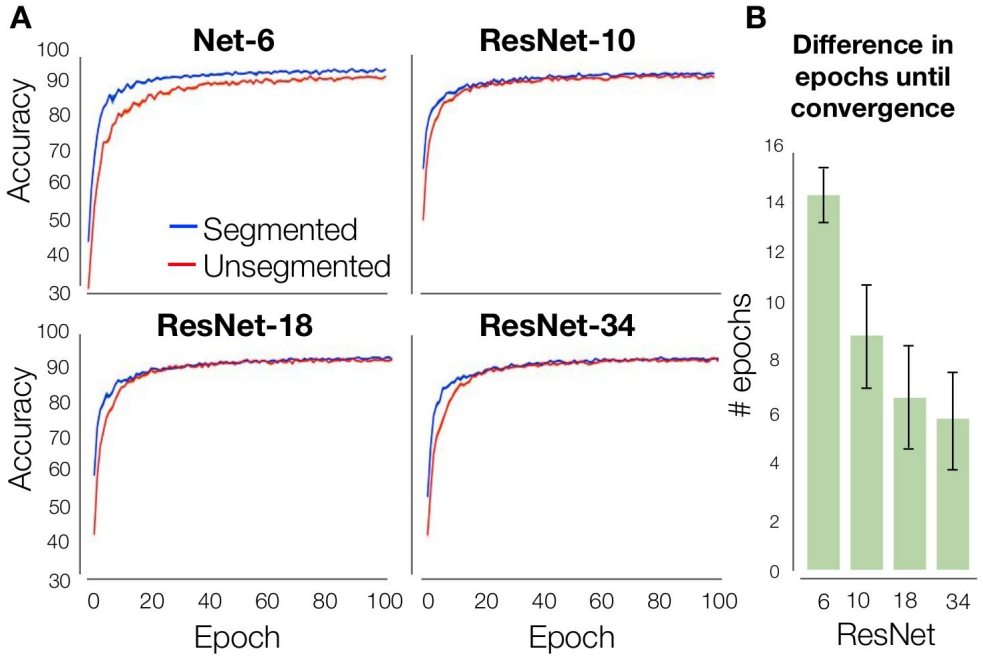


Figure 5. Accuracy during training on segmented vs. unsegmented stimuli. A) Models trained on segmented objects achieve better classification accuracy in the early stages of training than the networks trained on unsegmented objects, and they converge in less epochs. In the later stages, accuracy between the two types of models is similar. B) Shallow networks trained on segmented stimuli converge earlier than when they are trained on unsegmented stimuli. The difference in epochs until convergence decreases as the depth of the network increases.

Discussion

We investigated the extent to which object and context information is represented and used for object recognition in trained deep convolutional neural networks. Experiment 1 showed both a substantial overlap, and a difference in performance between human participants and DCNNs. Both humans and DCNNs are better in recognizing an object on a congruent versus an incongruent background. However, whereas human participants performed best in the segmented condition, DCNNs performed equally well (or better) for the congruent condition. Performance for the incongruent condition was lowest. This effect was

particularly strong for more shallow networks. Further analyses, investigating which parts of the image were most important for recognition, showed that the influence of the background features on the response outcome was relatively strong for shallow networks and almost absent for deeper networks. For shallow networks, the results of experiment 2 indicated a benefit of training on segmented objects (as compared to unsegmented objects). For deeper networks, this benefit was much less prominent. Training on segmented images thus reduced the difference in performance between shallow and deeper networks.

We interpret these findings as indicating that with an increase in network depth there is better selection of the features that belong to the output category. This process is similar, at least in terms of its outcome, to figure-ground segmentation in humans and might be one of the ways in which scene segmentation is performed in the brain using recurrent computations. It also suggests that, with adequate deployment of attention, a deeper network is not necessary to recognise the object.

Explicit vs. implicit models of grouping and segmentation

Classic models focussing on grouping and segmentation presume an explicit process in which certain elements of an image are grouped, whilst others are segregated from each other, by a labelling process [30,31]. Several studies have established the involvement of such explicit grouping mechanisms during specific visual tasks. For example, different curve tracing paradigms require grouping of spatially separate contour segments [32], and recent findings by Doerig, Bornet, Rosenholtz, Francis, Clarke & Herzog [33], comparing a wide range of computational models, indicate that an explicit grouping step is crucial to explain different (un)crowding phenomena. Adding explicit segmentation mechanisms to DCNNs is promising to explain human behavior in tasks that require integrating and grouping of global features, or shape-level representations. Our results from behavioral experiments with segmented and unsegmented objects show that when the task is *object recognition* an explicit segmentation step is typically not necessary. We show that with an increase in network depth, there is a stronger influence of the features that belong to the object on recognition performance, showing that 'implicit' segmentation occurs. When this process becomes more efficient (with a deeper network, or recurrent processing) the result is a situation in which, just as in 'explicit' segmentation, the network (or visual system) knows which features belong together, and which ones do not.

Previous studies have already looked into DCNN performance on unsegmented images [34,35], or have even shown a decrease in classification accuracy for unsegmented,

compared to segmented objects [26]). In those images, however, objects were placed on a random background, thereby often incongruent (or coincidentally, congruent). In the current study, by manipulating the relevance and usefulness of the background information, we could disentangle whether this decrease was due to a segmentation problem, or the presence of incongruent, misleading information.

Contextual effects in object recognition

Different accounts of object recognition in scenes propose different loci for contextual effects [36,37]. It has been argued that a bottom-up visual analysis is sufficient to discriminate between basic level object categories, after which context may influence this process in a top-down manner by priming relevant semantic representations, or by constraining the search space of most likely objects (e.g. [38]). Recent studies have also indicated that low-level features of a scene (versus high-level semantic components) can modulate object processing [37] by showing that seemingly meaningless textures with preserved summary statistics contribute to the effective processing of objects in scenes. Comparably, in the current study the DCNNs were agnostic to the meaning of the backgrounds, as they were not trained to recognize, for example, kitchens or bedrooms. The current results show that visual context features may impact object recognition in a bottom-up fashion, even for objects in a spatially incongruent location.

Previous studies have indicated that explicitly augmenting DCNNs with human-derived contextual expectations (likelihood, scale and location of a target object) was able to improve detection performance, potentially indicating a difference in contextual representations in the networks and the humans [39]. In the current study, findings show that only training DCNNs on a large dataset (ImageNet), enables them to learn human-like contextual expectations as well.

Feed-forward vs. recurrent processing

Instead of being an ultra-deep feedforward network, the brain likely uses recurrent connections for object recognition in complex natural environments. There are a multitude of findings that have firmly established the involvement of feedback connections during figure-ground segmentation. For example, behavior and neural activity in V1 evoked by figure-ground stimuli are affected by backward masking [40], region-filling processes that are mediated by feedback connections lead to an enhanced neural representation for figure regions compared to backgrounds in early visual areas [41], responses by neurons showing selectivity to border ownership are modulated depending on the location of a 'figure'

relative to other edges in their receptive field [42], and the accuracy of scene segmentation seems to depend on recurrent connections to sharpen the local elements within early visual areas [43] (and there are many more). The current results do not speak to those findings, but merely indicate that a very deep feedforward architecture is capable of obtaining a 'segmented' representation of an object, without recurrent projections.

The interpretation that deeper networks are better at object recognition, because they are capable of limiting their analysis to (mostly) the object –when necessary– is consistent with the idea that deeper networks are solving the challenges that are resolved by recurrent computations in the brain [27]. Previous findings comparing human behavior or the representational geometry of neural responses to DCNNs (e.g. [34,44]) often use images that contain (mostly) frontal views of objects on uniform backgrounds. For segmented objects, on a white or uniform background, all incoming information is relevant and segmentation is not needed. For those scenes, feed-forward activity in the brain may suffice to recognize the objects [6]. In line with those findings, we also see that even very shallow networks are able to perform well on those scenes. For more complex scenes, on the other hand, the first feed-forward sweep might not be not sufficiently informative, and correctly classifying or recognizing the object might require additional processing. For those scenes, we see a decrease in classification performance, mainly for the more shallow networks. These findings are in line with the global-to-local (or coarse-to-fine) processing framework, in which a coarse visual representation is acquired by the initial feedforward sweep. If this coarse representation is not informative enough to solve the task at hand, additional, more sophisticated visual processes ('routines') can be recruited to refine this representation [6,11,12,45–48]).

Background congruency

In human natural vision, extraction of gist can lead to a set of expectations regarding the scene's composition, indicating the probability of the presence of a certain object in a scene, but also its most probable locations [19,21]. In the current study, in incongruent scenes, objects did not only violate the overall meaning of the scene category (semantic violation), but were also placed in a position that was not predicted by the local structure of the scene (syntactic violation). On top of that, objects in the human categorization task were placed in a semi-random location across trials to make the task more difficult. This spatial uncertainty, however, has the additional benefit that it makes the task more comparable to the task we ask DCNNs to perform, as DCNNs have no knowledge about the spatial location. A pilot study using stimuli with centered 3D-rendered objects indicated no

difference in performance between congruent and incongruent images. While this is contrary to published literature [49], there are several factors that might explain this difference. First of all, we used 3D-rendered, computer generated objects, placed on natural scenes (real-world pictures, Supporting Figure S1). The difference in visual quality and 'style' between the object and the background might have influenced perception, by making it easier to distinguish them from each other. A second reason might be the size of the objects. Compared to the stimuli used by Davenport and Potter [15] or Munneke et al. [49], our objects were quite large, in order to obtain good network performance.

Conclusion

With an increase in network depth there is better selection of the features that belong to the output category. This process is similar, at least in terms of its outcome, to figure-ground segmentation in humans and might be one of the ways in which scene segmentation is performed in the brain.

Materials and methods

Experiment 1 (scene segmentation and background consistency):

Ethics statement. All participants provided written informed consent and were rewarded with research credits or received a monetary compensation. The experiment was approved by the ethics committee of the University of Amsterdam.

Participants. 40 participants (9 males) aged between 18 and 30 years (M = 22.03, SD = 3.02) with normal or corrected-to-normal vision, took part in the experiment.

Networks. To investigate the effect of network depth on scene segmentation, tests were conducted on seven deep residual networks (ResNets; [29]) with increasing number of layers (6, 10, 18, 34, 50, 101, 152). We chose ResNet architectures because they can be scaled up and down in size (depth) by adding or removing their basic building blocks, and they perform very well on datasets like ImageNet. This approach allowed us to investigate the effect of network depth (adding layers) while keeping other model properties as similar as possible. The fb.resnet.torch implementation of ResNets by Sam Gross and Michael Wilber [50] was used. In this implementation, input images from the ImageNet dataset [25] were 224x224 randomly cropped from a resized image using the scale and aspect ratio augmentation of Szegedy et al. (2015) [51]. Downsampling was done by stride-2 convolutions in the 3x3 layer of the first block in each stage (instead of the first 1x1 convolution) and weight decay was applied to all weights and biases (instead of just the weights of the convolution layers). ResNet-6 and ResNet-10 were trained on ImageNet [25] with 1 GPU. The other ResNets were downloaded (pretrained).

Stimuli. Images of 27 different object categories were generated by placing cut-out objects from the ImageNet validation set onto white (segmented), congruent and incongruent backgrounds. The categories were defined at a (sub)ordinate level, based on ImageNet categories: acoustic guitar, airliner, bathtub, birdhouse, cab, canoe, cellular telephone, china cabinet, dishwasher, grand piano, laptop, limousine, loudspeaker, mailbox, microphone, microwave, park bench, pirate ship, printer, remote, rocking chair, schoolbus, screen, speedboat, sports car, table lamp, wall clock (Figure 1A). There were ten exemplars for every object category. Backgrounds were sampled from a large database of images obtained from the SUN2012 database [52] (512*512 pixels, full-color). For each category, three typical backgrounds were selected using the five most common places where this

object was found within the database (sorted by number of instances inside each scene type). Three atypical backgrounds were manually chosen (Figure 1B).

To familiarize human participants with the categories, one of the ten exemplars for each category was randomly selected and used in a practice-run. Using the remaining nine exemplars - three for each condition (segmented, congruent, incongruent) - 243 images were generated for the actual experiment. Each exemplar was only presented once for each participant. To ensure participants processed the complete image, exemplars were downsized and placed in one of 9 possible locations (3x3 grid). Importantly, to rule out any effect of 'exemplar-complexity' (e.g. one guitar being easier to recognize than another) or an interaction between the object, location and the background, all possible exemplar-background-location combinations were balanced over participants.

For DCNNs, all exemplars were placed centrally, resulting in 810 images with a congruent background, 810 with an incongruent background and 270 images with segmented objects.

Experimental procedure. Participants performed on an object recognition task (Figure 1C). At the beginning of each trial, a fixation-cross was presented for 2000 ms, followed by an image. Images were presented in randomized sequence, for a duration of 32 ms, followed by a mask. The masks consisted of scrambled patches of the images and was presented for 300 ms. After the mask, participants had to indicate which object they had seen, by clicking on one of 27 options on screen using the mouse. After 81 (⅓) and 162 (⅔) trials, there was a short break. Using this paradigm, our human object recognition task was closely analogous to the large-scale ImageNet 1000-way object categorization for which the ResNets were optimized and thus expected to perform well.
For the models, all images were used to assess object recognition performance.

Statistical analysis: human performance. Accuracy (percentage correct) was computed for each participant. Differences in accuracy between the three conditions (segmented, congruent, incongruent) were statistically evaluated using a repeated-measures ANOVA. A significant main effect was followed up by two-tailed post-hoc pairwise comparisons using a Bonferroni correction at $\alpha = 0.05$, and the p-values reported in the main text are the adjusted p-values. Data were analyzed in Python.

Statistical analysis: DCNNs. For each of the images, the DCNNs (ResNet-6, ResNet-10, ResNet-18, ResNet-34, Resnet-50, ResNet-101, ResNet-152) assigned a probability value

to each of the 1000 object categories it had been trained to classify. For each condition (segmented, congruent, incongruent) the Top-5 Error (%) was computed (classification is correct if the object is among the objects categories that received the five highest probability assignments). Then, to gain more insight in the importance of the features in the object vs the background for classification, we added Gaussian noise to either the object, background, or to both (the complete image) and evaluated performance.

Experiment 2: training on unsegmented/segmented objects

Results from experiment 1 suggested that information from the background is present in the representation of the object, predominantly for more shallow networks. What happens if we train the models on segmented objects, when all features are related to the object? To further explore the role of segmentation on learning, we trained ResNets differing in depth on a dataset with objects that were already segmented, and a dataset in which they were intact (i.e. embedded in a scene).

Models. As in experiment 1, we used deep residual network architectures (ResNets; [29]) with increasing number of layers (6, 10, 18, 34). We did not use ResNets with more than 34 layers, as the simplicity of the task leads to overfitting problems for the 'ultra-deep' networks.

Stimuli. To train the models, a subset of images from 10 different categories were selected from ImageNet. The categories were: bird 1 t/m 7, elephant, zebra, horse. Using multiple different types of birds helped us to increase task difficulty, enforcing the models to learn specific features for each class. The remaining (bigger) animals were added for diversity. From this subselection, we generated two image sets: one in which the objects were segmented, and one with the original images (objects embedded in scene). Because many images are needed to train the models, objects were segmented using a DCNN pretrained on the MS COCO dataset [53], using the Mask R-CNN method [54] (instead of manually). Images with object probability scores lower than 0.98 were discarded, to minimize the risk of selecting images with low quality or containing the wrong object. All images were resized to 128x128 pixels. In total, the image set contained ~9000 images. 80% of these images was used for training, 20% was used for validation.

Table 1. Dataset classes (categories) and the number of training and test stimuli. Multiple different types of birds increased task difficulty, enforcing the models to learn specific features for each class. The remaining (bigger) animals were added for diversity.

Category	Train	Test
Bird 1	944	236
Bird 2	867	217
Bird 3	312	227
Bird 4	455	114
Bird 5	421	105
Bird 6	930	233
Bird 7	462	241
Elephant	700	175
Horse	290	72
Zebra	316	230

Experimental procedure. First, we trained the different ResNets for 100 epochs and monitored their accuracy after each epoch on the validation sets. Then, we reinitialized the networks with different seeds and repeated the process for 10 different seeds to obtain statistical results. During training, we monitored the training loss to ensure that there was no overfitting.

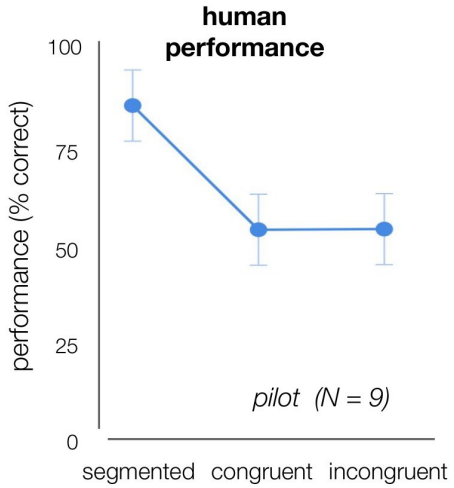
Acknowledgments

We thank Yannick Vinkesteijn for help with data collection for the human object recognition task, Sara Jahfari for helpful comments on the manuscript and all members of the Scholte lab for discussion. This research was supported by an Advanced Investigator Grant to EdH by the European Research Council (ERC grant FAB4V #339374).

Data and code availability:

Data and code to reproduce the analyses in this article are available at http://github.com/noorseijdel/2019_scenecontext.

Supporting information



S1. Human performance (% correct) on the object recognition task, using centered 3D-rendered objects on white, congruent or incongruent backgrounds. Performance was higher for the segmented condition, compared to congruent and incongruent.

References

1. VanRullen R, Thorpe SJ. Surfing a spike wave down the ventral stream. Vision Res. 2002;42: 2593–2615. doi:10.1016/S0042-6989(02)00298-5

2. DiCarlo JJ, Cox DD. Untangling invariant object recognition. Trends Cogn Sci. 2007;11: 333–341. doi:10.1016/j.tics.2007.06.010

3. Serre T, Oliva A, Poggio T. A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences. 2007;104: 6424–6429. doi:10.1073/pnas.0700622104

4. Koivisto M, Railo H, Revonsuo A, Vanni S, Salminen-Vaparanta N. Recurrent processing in V1/V2 contributes to categorization of natural scenes. J Neurosci. 2011;31: 2488–2492. doi:10.1523/JNEUROSCI.3074-10.2011

5. Camprodon JA, Zohary E, Brodbeck V, Pascual-leone A. Two Phases of V1 activity for visual recognition of natural images. 2013;18: 1199–1216. doi:10.1016/j.micinf.2011.07.011.Innate

6. Groen IIA, Jahfari S, Seijdel N, Ghebreab S, Lamme VAF, Scholte HS. Scene complexity modulates degree of feedback activity during object detection in natural scenes. PLoS Comput Biol. 2018;14: e1006690. doi:10.1371/journal.pcbi.1006690

7. Groen IIA, Ghebreab S, Prins H, Lamme VAF, Scholte HS. From Image Statistics to Scene Gist: Evoked Neural Activity Reveals Transition from Low-Level Natural Image Structure to Scene Category. Journal of Neuroscience. 2013;33: 18814–18824. doi:10.1523/JNEUROSCI.3128-13.2013

8. Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat

Neurosci. 2019;22: 974–983. doi:10.1038/s41593-019-0392-5

9. Herzog MH, Clarke AM. Why vision is not both hierarchical and feedforward. Front Comput Neurosci. 2014;8: 135. doi:10.3389/fncom.2014.00135

10. Rajaei K, Mohsenzadeh Y, Ebrahimpour R, Khaligh-Razavi S-M. Beyond core object recognition: Recurrent processes account for object recognition under occlusion. PLoS Comput Biol. 2019;15: e1007001. doi:10.1371/journal.pcbi.1007001

11. Lamme V a. F, Roelfsema PR. The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci. 2000;23: 571–579. doi:10.1016/S0166-2236(00)01657-X

12. Hochstein S, Ahissar M. View from the top: Hierarchies and reverse hierarchies in the visual system. Neuron. 2002;36: 791–804. doi:10.1016/S0896-6273(02)01091-7

13. Howe PDL. Natural scenes can be identified as rapidly as individual features. Atten Percept Psychophys. 2017;79: 1674–1681. doi:10.3758/s13414-017-1349-y

14. Bar M, Ullman S. Spatial context in recognition. Perception. 1993;25: 343–352. doi:10.1068/p250343

15. Davenport JL, Potter MC. Scene consistency in object and background perception. Psychol Sci. 2004;15: 559–564. doi:10.1111/j.0956-7976.2004.00719.x

16. Davenport JL. Consistency effects between objects in scenes. Mem Cognit. 2007;35: 393–401. doi:10.3758/BF03193280

17. Joubert OR, Rousselet G a., Fize D, Fabre-Thorpe M. Processing scene context: Fast categorization and object interference. Vision Res. 2007;47: 3286–3297. doi:10.1016/j.visres.2007.09.013

18. Joubert OR, Fize D, Rousselet GA, Fabre-Thorpe M. Early interference of context congruence on object processing in rapid visual categorization of natural scenes. J Vis. 2008;8: 11.1–18. doi:10.1167/8.13.11

19. Rémy F, Saint-Aubert L, Bacon-Macé N, Vayssière N, Barbeau E, Fabre-Thorpe M. Object recognition in congruent and incongruent natural scenes: A life-span study. Vision Res. 2013;91: 36–44. doi:10.1016/j.visres.2013.07.006

20. Sun H-M, Simon-Dack SL, Gordon RD, Teder WA. Contextual influences on rapid object categorization in natural scenes. Brain Res. 2011;1398: 40–54. doi:10.1016/j.brainres.2011.04.029

21. Greene MR, Botros AP, Beck DM, Fei-Fei L. What you see is what you expect: rapid scene understanding benefits from prior experience. Atten Percept Psychophys. 2015; 1239–1251. doi:10.3758/s13414-015-0859-8

22. Riesenhuber M, Poggio T. Hierarchical models of object recognition in cortex. Nat Neurosci. 1999;2: 1019–1025. doi:10.1038/14819

23. Fink M, Perona P. Mutual Boosting for Contextual Inference. In: Thrun S, Saul LK, Schölkopf B, editors. Advances in Neural Information Processing Systems 16. MIT Press; 2004. pp. 1515–1522.

24. Torralba A, Oliva A, Castelhano MS, Henderson JM. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol Rev. 2006;113: 766. doi:10.1037/0033-295X.113.4.766

25. Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, et al. ImageNet Large Scale Visual Recognition Challenge. Int J Comput Vis. 2015;115: 211–252. doi:10.1007/s11263-015-0816-y

26. Kheradpisheh SR, Ghodrati M, Ganjtabesh M, Masquelier T. Deep Networks Can Resemble Human Feed-forward Vision in Invariant Object Recognition. Sci Rep. 2016;6: 32672. doi:10.1038/srep32672

27. Liao Q, Poggio T. Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. arXiv. 2016. Available: http://arxiv.org/abs/1604.03640

28. Scholte HS. Fantastic DNimals and where to find them. NeuroImage. 2018. pp. 112–113. doi:10.1016/j.neuroimage.2017.12.077

29. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. Proceedings of the IEEE conference on computer vision and pattern recognition. 2016. pp. 770–778. doi:10.1109/CVPR.2016.90

30. Treisman A. Solutions to the binding problem: progress through controversy and convergence. Neuron. 1999;24: 105–10, 111–25. doi:10.1016/S0896-6273(00)80826-0

31. Neisser U, Becklen R. Selective looking: Attending to visually specified events. Cogn Psychol. 1975;7: 480–494. doi:10.1016/0010-0285(75)90019-5

32. Roelfsema PR, Scholte HS, Spekreijse H. Temporal constraints on the grouping of contour segments into spatially extended objects. Vision Res. 1999;39: 1509–1529. doi:10.1016/s0042-6989(98)00222-3

33. Doerig A, Bornet A, Rosenholtz R, Francis G, Clarke AM, Herzog MH. Beyond Bouma's window: How to explain global aspects of crowding? PLoS Comput Biol. 2019;15: e1006580. doi:10.1371/journal.pcbi.1006580

34. Martin Cichy R, Khosla A, Pantazis D, Oliva A. Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. Neuroimage. 2017;153: 346–358. doi:10.1016/j.neuroimage.2016.03.063

35. Cadieu CF, Hong H, Yamins DLK, Pinto N, Ardila D, Solomon EA, et al. Deep neural networks rival the representation of primate IT cortex for core visual object recognition. PLoS Comput Biol. 2014;10: e1003963. doi:10.1371/journal.pcbi.1003963

36. Oliva A, Torralba A. The role of context in object recognition. Trends Cogn Sci. 2007;11: 520–527. doi:10.1016/j.tics.2007.09.009

37. Võ ML-H, Boettcher SE, Draschkow D. Reading scenes: how scene grammar guides attention and aids perception in real-world environments. Curr Opin Psychol. 2019;29: 205–210. doi:10.1016/j.copsyc.2019.03.009

38. Bar M. A cortical mechanism for triggering top-down facilitation in visual object recognition. J Cogn Neurosci. 2003;15: 600–609. doi:10.1162/089892903321662976

39. Katti H, Peelen MV, Arun SP. Machine vision benefits from human contextual

expectations. Sci Rep. 2019;9: 2112. doi:10.1038/s41598-018-38427-0

40. Lamme VAF, Zipser K, Spekreijse H. Masking interrupts figure-ground signals in V1. J Cogn Neurosci. 2002;14: 1044–1053. doi:10.1162/089892902320474490

41. Self MW, Roelfsema PR. The Neural Mechanisms of Figure-ground Segregation. Oxford Handbooks Online. 2014. doi:10.1093/oxfordhb/9780199686858.013.036

42. von der Heydt R. Figure–ground organization and the emergence of proto-objects in the visual cortex. Front Psychol. 2015;6: 10391. doi:10.3389/fpsyg.2015.01695

43. Self MW, Jeurissen D, van Ham AF, van Vugt B, Poort J, Roelfsema PR. The Segmentation of Proto-Objects in the Monkey Primary Visual Cortex. Curr Biol. 2019;29: 1019–1029.e4. doi:10.1016/j.cub.2019.02.016

44. Khaligh-Razavi S-M, Kriegeskorte N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol. 2014;10: e1003915. doi:10.1371/journal.pcbi.1003915

45. Crouzet SM, Serre T. What are the Visual Features Underlying Rapid Object Recognition? Front Psychol. 2011;2: 326. doi:10.3389/fpsyg.2011.00326

46. Epshtein B, Lifshitz I, Ullman S. Image interpretation by a single bottom-up top-down cycle. Proc Natl Acad Sci U S A. 2008;105: 14298–14303. doi:10.1073/pnas.0800968105

47. Petro LS, Vizioli L, Muckli L. Contributions of cortical feedback to sensory processing in primary visual cortex. Front Psychol. 2014;5: 1223. doi:10.3389/fpsyg.2014.01223

48. Zheng S, Yuille A, Tu Z. Detecting object boundaries using low-, mid-, and high-level information. Comput Vis Image Underst. 2010;114: 1055–1067. doi:10.1016/j.cviu.2010.07.004

49. Munneke J, Brentari V, Peelen MV. The influence of scene context on object recognition is independent of attentional focus. Front Psychol. 2013;4: 552. doi:10.3389/fpsyg.2013.00552

50. Gross S, Wilber M. Training and investigating residual nets. Facebook AI Research, CA [Online] Avilable: http://torch ch/blog/2016/02/04/resnets html. 2016.

51. Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, et al. Going deeper with convolutions. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. pp. 1–9. doi:10.1109/CVPR.2015.7298594

52. Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A. SUN database: Large-scale scene recognition from abbey to zoo. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. 2010. pp. 3485–3492. doi:10.1109/CVPR.2010.5539970

53. Lin T-Y, Maire M, Belongie S, Hays J, Perona P, Ramanan D, et al. Microsoft COCO: Common Objects in Context. Computer Vision – ECCV 2014. Springer International Publishing; 2014. pp. 740–755. doi:10.1007/978-3-319-10602-1_48

54. He K, Gkioxari G, Dollar P, Girshick R. Mask R-CNN. IEEE Trans Pattern Anal Mach

Intell. 2018. doi:10.1109/TPAMI.2018.2844175