Object-scene conceptual regularities reveal fundamental differences between biological and artificial object vision.

Stefania Bracci[12]*, Jakob Mraz[2], Astrid Zeman[2], Gaëlle Leys[2] and Hans Op de Beeck[2].

[1] Center for Mind/Brain Sciences - CIMeC, University of Trento, Rovereto (TN), Italy.
[2] KU Leuven, Leuven Brain Institute, Brain & Cognition Research Unit, Leuven, 3000, Belgium.

Corresponding author:
Stefania Bracci
Email: stefania.bracci@unitn.it

Keywords:
Object recognition, scenes, context, CCNs, visual cortex.

Abstract

Human vision is still largely unexplained. Computer vision made impressive progress on this front, but it is unclear to what extent artificial neural networks approximate human brain strategies. Here, we confirm this gap by testing how biological and artificial systems encode object-scene contextual regularities in natural images. Both systems represent these regularities, but the underlying information processing is markedly different. In human vision, objects and backgrounds are represented separately, with rich domain-specific representations characterizing human visual cortex. Interaction between these components occurs downstream in frontoparietal areas. Conversely, neural networks represent image components in a single entangled representation revealing reduced object-segregation abilities and impoverished domain-specific object spaces. These results show the uniqueness of human vision that allows understanding that images are not just a collection of features and points to the need for developing neural network models with a similar richness of representational content.

Introduction

We live in a structured world. As a consequence, sensory input is not a random collection of lines and patterns but can be organized into meaningful and identifiable wholes such as objects. These objects show particular relationships to their environment. Some objects are most likely to appear in specific contexts, such as penguins in ice landscapes or lions in the savannah. This set of rules, similarly to the grammar of our language, provides a structure to guide our behavior (1). Here we investigate how this structure is represented in human and artificial neural systems.

Despite visual information is fully intertwined at the retina level, the human visual cortex makes an explicit distinction between objects and backgrounds (i.e., figure-background segregation). The ability to separate foreground objects from background information is a prerequisite to object recognition, for which the VTC plays a critical role (31), and largely separated brain channels process object and scene information, respectively (4, 5, 32). Further, human neuroimaging has shown that rich domain-specific object spaces characterise the different processing channels. As an example, behaviorally relevant object dimensions are encoded in object and scene areas: animate features such as the eyes/mouth useful for identifying living entities in the former (6-8), and spatial layout informative for scene navigation in the latter (9, 10, 89). At the same time, statistical regularities of the world influence perception (11, 12, 13, 16), and interaction between object and scene processing has been shown at the neural level (13-15, 33, 34, 87). While such studies indicate that the representation of objects might be influenced by whether they share similar context statistics, to our knowledge, none of them included a multivariate test pinpointing the representational similarity of neural patterns across objects and scenes. Is the frequent occurrence of a penguin in ice landscapes sufficient to modulate how penguins and ice landscapes are coded, despite the clearly separate processing channels?

We know even less about artificial vision models. In the past decade, artificial computer vision models have been developed that are able to classify visual patterns with human-level performance, such as deep convolutional neural networks (DCNNs) and recent adaptations including residual, recurrent, and transformer networks (84). These artificial models appear to develop representations similar to those in human visual cortex (24,90). Various avenues have been suggested to further improve this correspondence, such as the use of recurrent processing and changes in training regimes (85,86). However, the crucial object-background distinction has been largely neglected but it becomes apparent when we consider DCNNs training modalities. Many DCNNs employed in object recognition are trained on large image sets in which each image is attached to a single label. In such scenario, DCNNs trained in object recognition do not need to explicitly separate objects from background; short cuts are available and sufficient to solve the problem (46). In other words, no matter whether visual information pertains to foreground or background, if it is useful to succeed at the task at hand, it becomes part of the representation. It is, therefore, far from clear to which extent such models develop explicit knowledge about image structure.

What is lacking in the literature, both at the neural and computational level, is an explicit comparison of responses to objects and backgrounds. We present the first study of that kind. We compare "neural" representations in the human brain and in artificial DCNNs using an original stimulus set that includes object images and background images with a variety of object/scene

properties, and a manipulation of object-background co-occurrence in real-world images. The results show that DCNNs appear to be capable of mimic conceptual-like human knowledge of the world, but the computational strategy implemented diverges greatly from biological vision and is largely dependent on DCNNs' training. In particular, DCNNs are not optimized for object-background segregation and do not develop the rich domain-specific object space present in human visual cortex.

Results

To test how the biological or artificial brain represent image regularities, we created a stimulus set of contextually related pairs of images (Fig. 1). Each pair includes one animal (clownfish, ladybug, passerine bird, seagull, polar bear, and gorilla) and its associated background (anemone, leaves, tree branches, seashores, ice landscapes, and jungle forests).

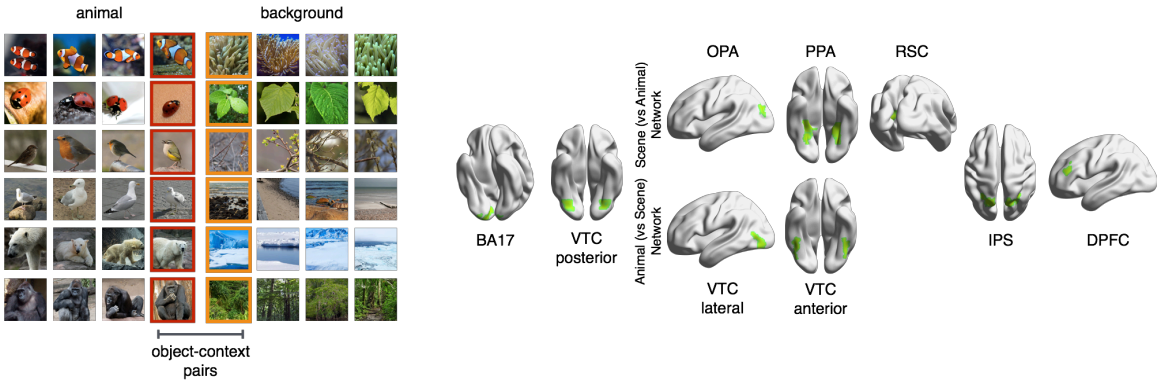

Fig. 1 – Experimental design. *The stimulus set (left) includes 2 conditions: animals and backgrounds. The pictures of animals were carefully selected to avoid that background information could be informative for object identification (e.g., the polar bear and the gorilla have a very similar neutral background). To control for shape, we further divided the animal categories in three subsets along the animacy continuum (2 mammals, 2 birds, and 2 small rounded animals). Within each subset, animals are matched for body shape (e.g., gorilla and polar bear) but each animal is paired with a different scene. As an example, the passerine bird and the seagull have similar body shape and size but are associated with two different backgrounds. As for the pictures of scenes, 3 of the backgrounds are characterized by rich navigational properties where there is no object in focus in the middle of the image: seashores, ice landscapes, and jungle forests. The other 3 backgrounds are object-like scenes with little navigational layout properties: anemones, leaves, and tree branches. The ROIs (right), included in the brain RSA analysis, are shown in light green. The areas included visual areas (for their relevance in object recognition) and frontoparietal areas (for their relevance in goal-directed behavior): BA17, ventrotemporal (VTC) posterior, VTC anterior, VTC lateral, occipital place area (OPA), parahippocampal area (PPA), retrosplenial cortex (RSC), intraparietal sulcus (IPS), and dorsal prefrontal cortex (DPFC). See Methods for details on the localization procedure.*

This unique stimulus set allows us to differentiate between at least 4 hypotheses about how objects and scenes might be represented (Fig. 2). The *object/background model* predicts a separation between animals and backgrounds. On the contrary, the *co-occurrence model* predicts contextually related associations between each animal and its associated background. These two models are orthogonal to each other (r = -0.05). Two additional control models rule out the role of visual factors: (1) The *GIST model* (29), a good descriptor of scene statistics (16), well describes responses in lower visual areas (30); (2) the *basic-level category model* captures within-category similarities for each animal and background condition. Representational similarity analysis (RSA; 28) allows to test each model's ability to capture the object space in the brain and DCNNs.

Objects and backgrounds are processed in separate pathways in visual cortex, with contextual effects emerging in prefrontal areas.

How does the human brain encode both object-scene separation and interaction? To test this, we analysed fMRI scans from participants (N=19; two long scan sessions per participant) while they watched the stimuli from Fig. 1 in an event-related design. Participants were asked to indicate to what extent each image would normally co-occur with the previous image. We investigated the representational structure along the hierarchy of visual regions from the primary visual cortex over regions involved in object and scene perception up to regions in the frontoparietal cortex involved in goal-directed behavior (35) that are known to be connected to visual domain-specific regions (36).

Results in visual cortex reveal a clear separation between animal and background representations. The object/background model captures most variance in many areas from posterior to anterior temporal cortex but, importantly, not in early visual cortex BA17 where only the GIST model reached significance ($z=0.14$; $p=0.0001$, Fig. 2B). The separation between animals and backgrounds is particularly evident in animal-selective areas (object/background model: $z > 0.84$; for all ROIs, $p < 0.0001$, Fig. 2B) with an additional minor contribution of the GIST model in posterior VTC ($z=0.17$; $p < 0.0001$, Fig. 2B). This separation is also very obvious when visually inspecting the dissimilarity matrices displayed in Fig. 2A. In scene-selective areas, the object/background model also best captures the representational structure in 2 out of 3 ROIs (PPA: $z= 0.46$; OPA: $z= 0.29$; both $p < 0.0001$, Fig. 2B). The remaining models did not explain any additional variance in these areas. The very same results were observed using single-subject data (Fig. 2A).

The information comes together again when moving downstream, with a small but significant effect of co-occurrence model in frontoparietal areas (IPS: $z= 0.18$; DPFC: $z = 0.16$, both $p < 0.0001$, Fig. 2B), thus shifting the representational content in these ROIs from object-background separation to object-background association. When using individual subject data, the effect for the co-occurrence model in these ROIs is relatively small but highly significant (DPFC and IPS: $z < 0.04$, both $p < 0.005$, Fig 2A). It also reaches reliability boundaries (between 0.04 – 0.09), which indicates the highest expected correlation after accounting for signal noise (Ritchie et al., 2017). In DPFC, the remaining models did not explain any significant variance in the patterns, while in IPS there were significant correlations for the object/background model (object/background model: $z= 0.18$; $p < 0.0001$, Fig. 2B). Together, these results show that the overall representational content in visual cortex largely distinguishes between the object and background, with interaction of these components emerging at a later processing stage in regions known to support goal-directed behaviour (Duncan, 2010).

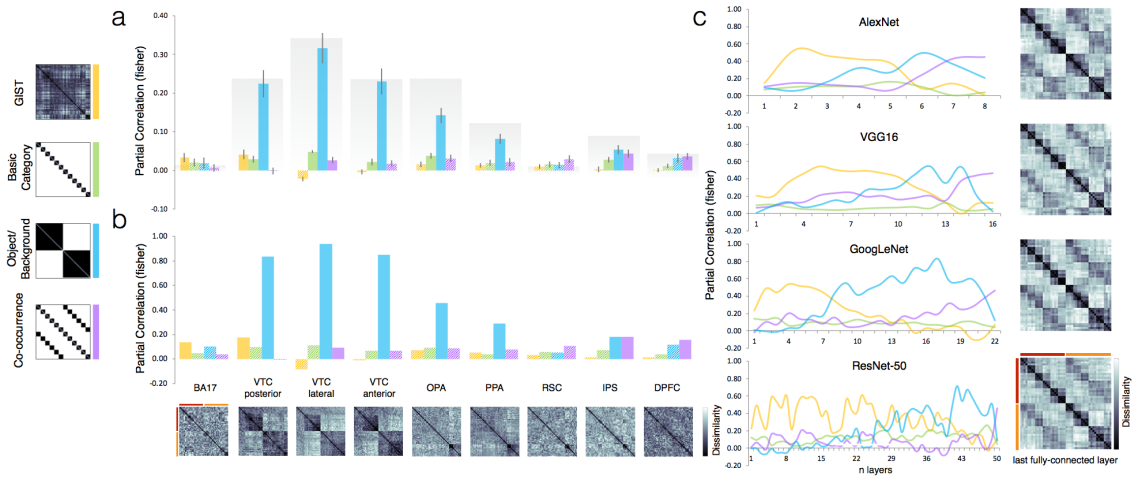

Fig. 2. Object-background separation in visual cortex and object-background integration in DCNNs highest layers. *(A-B) The brain RSA results reveal a strong separation between background and animal representations in most ventral regions. The effect of animal-background co-occurrence emerges in frontoparietal areas. The RSA results for the 4 models (GIST, basic category, object/background, co-occurrence) are shown for group-averaged neural data (B) and individual subject data (A). For group-averaged results (B), filled bars indicate significant values (p < 0.0001) computed with permutation tests (10,000 randomizations of stimulus labels). For each ROI, the neural dissimilarity matrix (1 - r) is shown. For individual subject results (A), reliability boundaries (in gray), indicate the highest expected correlation considering signal noise (see Methods) and error bars indicate SEM. Filled bars indicate significant values (p <0.005) calculated with pairwise t-tests across subjects (n=19). (C) The DCNNs RSA results for the 4 models (GIST, basic category, object/background, co-occurrence) are shown for 4 DCNNs (AlexNet, VGG16, GoogLeNet, ResNet-50). The network's depth is shown on the x axis. For each DCNN, the representational dissimilarity (1 - r) is shown for the last fully-connected layer. Red represents animal conditions and orange represents background conditions in the lines next to the dissimilarity matrix.*

The object code in DCNNs reflects object-background statistical regularities

DCNNs widely employed in image classification (23), are now considered a promising venue to explain biological object vision at the neural level proving to match human object recognition abilities in several benchmarks (38-40). Yet, it is unclear the extent to what machine vision implements computational strategies that mimic those used by the human brain (21). To test this, we use RSA to evaluate the predictive power of our models in four DCNNs: Alexnet, VGG16, GoogLeNet, and ResNet-50. All models are trained in object recognition (ImageNet) and are the most frequent benchmarks used in the literature that suggests similarities in representational structure between human and computational vision (42, 43).

The results show that the representational space at the latest processing stage of all DCNNs shows a salient object-background association (Fig. 2C). For all networks, the co-occurrence model peaks at the final processing stages (i.e., fully-connected layers) and, relative to the remaining models, explains significantly better the DCNNs' representational space ($p < 0.001$ relative to all models; for summary statistics, see Fig. 3). This structure is also visible in the secondary diagonal emerging in the lower left and upper right quadrant of the networks' representational dissimilarity matrices (Fig. 2C). The entangled animal-background representation emerges later through the network's hierarchy, ruling out the contribution of low-level image

properties (e.g., visual similarities between each animal category and its associated environment). Confirming this, the GIST model best predicts representations at an early processing stage in all neural networks (Fig. 2C). When the GIST model drops in performance, the object/background model increases and reaches significance relative to the remaining models (Fig. 2C, and Fig. 3).

Despite differences, the RSA results do also reveal some degree of similarity between biological and artificial systems in the way the object-space transforms through their computational hierarchy (Fig 3). Initially, the object space reflects image low-level visual properties (GIST model, yellow color-coded), it then shifts towards a general division between objects and backgrounds (object/background model, light-blue color-coded), to finally reveal object-background co-occurrence effects (co-occurrence model, purple color-coded). These similarities are intriguing yet might be the result of largely diverging information processing strategies. We address this point in the following sections.

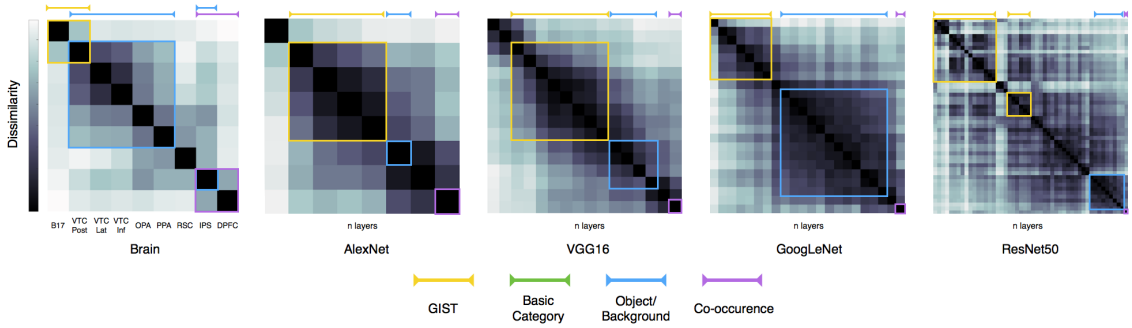

Fig. 3. Representational transformations in the brain and DCNNs' computational hierarchy. *Correlational matrices show second-order relationships among representational patterns in the brain's ROIs and individual DCNNs' layers. Color-coded line boxes highlight the ROIs and DCNNs' layers where each model reaches significance. For brain areas, significance for each model is shown relative to baseline (p <0.0001), calculated with permutation tests (10,000 randomizations of stimulus labels). For DCNNs' layers, significance for each model is shown relative to all remaining models (p < 0.001), calculated with permutation tests (1000 randomizations of stimulus labels).*

DCNNs capture humans' conceptual knowledge but lack fundamental object vision computations.

An object code that incorporates statistical regularities of an object in its background might be relevant to mimic human-like object recognition behavior, which indeed is influenced by the object-scene interaction (*11, 12*); context facilitates object recognition and vice versa (*13, 16*). Given that DCNNs representations at their final layers appear to capture contextual regularities between an object and its recurring background, we might expect similarities with human behavior when humans judge image similarities in terms of co-occurrence (see Methods).

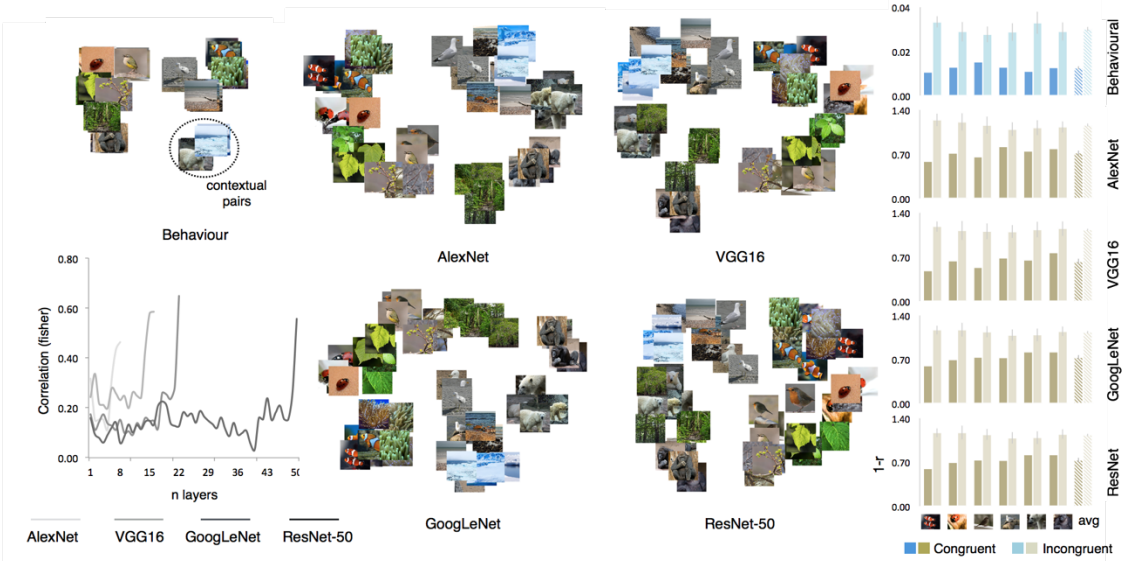

Fig. 4. DCNNs acquire human-like conceptual biases. *The correlational plot (bottom left) shows the degree of similarity between behavioral judgments and each of the four DCNN architectures. MDS spaces (metric stress) show consistent object-scene clusters in human behavioral judgments as well as DCNNs (last fully-connected layer). The object-scene cluster analysis (right) for behavioral and DCNNs (last layer) data, show a consistent significant effect of congruency (lower distance) for object-scene stimulus pairs. Dashed bars show the averaged data for the 6 stimulus pairs.*

Indeed, the correlation between human co-occurrence judgments and DNNs' object space (Fig. 4, bottom left) increases throughout the DCNNs' processing hierarchy to reach its peak at the final processing stages for all DCNNs (AlexNet: $z = 0.46$; VGG16: $z = 0.59$; GoogLeNet: $z = 0.65$; ResNet-50: $z = 0.56$). The DCNNs' ability to capture human behavior is also visible when inspecting the MDS's space (Fig. 4). To confirm this observation statistically, we tested similarities for congruent pairs (polar bear and iceberg) versus the average of all remaining incongruent pairs (polar bear and jungle). We observed significantly higher similarity for congruent relative to incongruent pairs in behavioral ($p < 0.0001$; Fig. 4, top right) as well as DCNNs data (for all models, $p < 0.001$). Thus, DCNNs' sensitivity to image statistical regularities of objects occurring in typical environments results in the acquisition of a "representation of the world" that closely resembles humans' conceptual judgments.

Despite these striking results, the similarity between DCNNs and human conceptual knowledge might be an accidental result. In other words, DCNNs might still be oblivious of the foreground-background distinction. Object-background segregation might be less optimized in DCNNs architectures trained with natural images (e.g., ImageNet) where regularities between objects and background are common. For instance, a whale lives in open water, not on the street. As such, any image feature (e.g., not necessarily object-specific) might become a useful piece of information to recognize the object (41). Here, we test this hypothesis by systematically manipulating object/background co-occurrence during DCNNs training (see Methods): from 0% co-occurrence (objects within the same category were no more likely to occur on the same background type than objects from different categories) to 100% co-occurrence (objects within the same category were always presented on the same background). Confirming the role of image regularities in facilitating object recognition (Fig 5A), the model's validation test revealed

significantly higher model performance for the 100% co-occurrence condition (accuracy: 0.85) relative to remaining conditions (accuracy < 0.72; t(4) 3.95, p < 0.02, for all pairwise t-tests). Furthermore, we predicted an increasing representational bias towards human-like conceptual object-background associations in accordance with increasing object-background co-occurrence. We run an RSA analysis with each of our predicting models and tested correlations for fully connected layer 7 in a 4x4 ANOVA within Co-occurrence (0%, 58%, 83%, 100%) and Model (GIST, basic category, object/background, co-occurrence) as within-subject factors. Results revealed a significant Co-occurrence x Model interaction (F(9,36) 11.82, p < 0.0001; Fig. 5B), showing that different levels of object-background regularities in the training set result in differences in DCNN's representational content (fc7) as captured by our models.

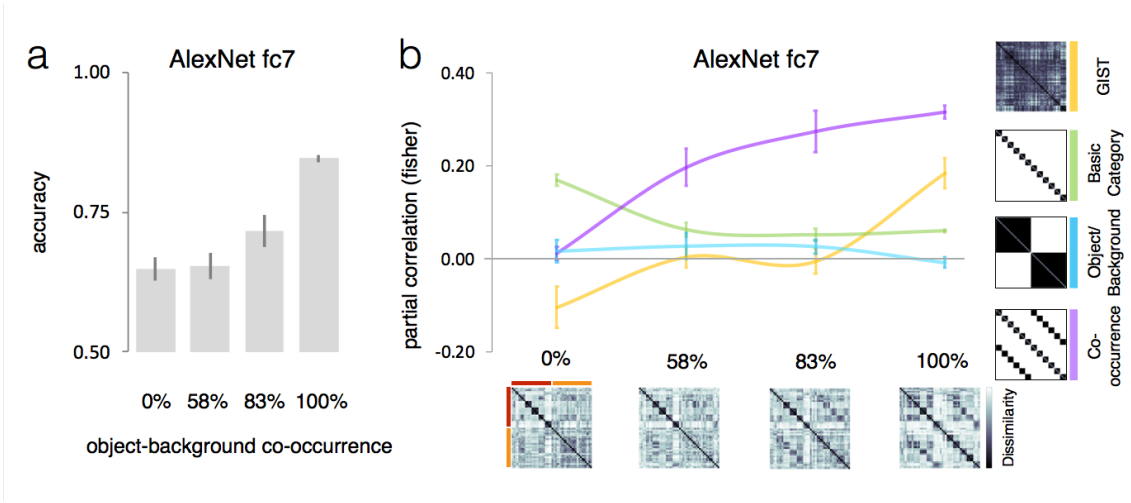

Fig. 5. The effect of images' object-background co-occurrence in the DCNN object space. *We trained a DCNN multiple times (N=5) in 4 training conditions with increasing levels of object-background regularity from 0% to 100%. (a) Accuracy results for the model's validation test. (b) The RSA results for the 4 models (GIST, basic category, object/background, co-occurrence) are shown for AlexNet's fc7 that underwent different training regimes with increasing co-occurrence (0%, 58%, 87%, 100%) between objects and backgrounds (n=6). Error bars indicate SEM. Red represents animal conditions and orange represents background conditions in the lines next to the dissimilarity matrix.*

Follow-up analyses revealed that the contextual-related effect is absent when the training is based on random object-background associations (0% co-occurrence: z = 0.01) but emerges when regularities in the training set increase (58%: z = 0.20; 83%: z = 0.27; 100%: z = 0.32). This effect is already significantly higher for the 58% condition relative to 0% condition (t(4) 3.68, p= 0.02) and it increases for higher levels of object-background co-occurrence (83%: t(4) 6.51, p= 0.003; 100%: t(4) 20.22, p < 0.0001; both relative to 0%). Interestingly, the effect of the basic category model is significantly high when the network cannot rely on background information (0%: z = 0.17), but strongly decreases when regularity increases (58%-100%: z < 0.06; t(4), > 6.24, p < 0.004, for all 3 comparisons). In the 100% co-occurrence condition, the correlation with the GIST model reaches significance (z = 0.18; t(4), > 5.78, p = 0.004), probably because in this condition there is an increase of lower-level background features that can be relied upon.

Together, these results point to a fundamental difference between biological and computational vision, namely current DCNNs employed in object vision seem to be not optimized to performed object background segregation. At the same time, in Fig 2C we observed some degree of object/background separation in DCNNs mid-layers. In the next section, we further investigate the nature of this object-background division.

Multiple object (domain-specific) spaces in the visual cortex but not in DCNNs

In visual cortex, the object/background model captures most of the variance in animal as well as in scene areas, but the dissimilarity matrices (Fig. 2A-B) reveal a marked difference between the two sets of regions that likely reflect differential domain-specific object spaces (9, 10, 37). DCNN's mid-layers do also show some object/background effects (Fig. 2C), but does this division embed domain-specific object spaces like what we observed in the human visual cortex? To investigate this, we tested two object dimensions that well characterize the object space in animal and scene areas. The *animacy continuum model* (37) captures the degree to which an animal is perceived as animate (within the animate/inanimate division some animals are perceived more animate than others: mammals > birds, birds > fish/bugs). Conversely, the *navigational layout model* (9, 10, 89) captures the degree to which scene layout provides relevant information for navigating a scene. To control for low-level image properties, we also include the GIST model. As above, these models were tested with RSA in brain and DCNNs data.

As expected, the brain RSA (Fig 6C) confirms differential representational spaces reflecting domain-specific object spaces in visual cortex: animal-selective ROIs show a significant preference for the animacy continuum model (VTC lat: $z = 0.87$; VTC inf: $z = 0.73$; both $p < 0.001$, relative to the remaining models); scene-selective ROIs show a significant preference for the navigational layout model (OPA: $z=0.48$; PPA: $z= 0.39$; both $p < 0.001$, relative to the remaining models). These results highlight the functional specialization of the human brain for multiple domain-specific computations and it highlights a representational richness that is exactly what one expects in a system with separate streams to process object and scene information. Conversely, the RSA for DCNN data only partially replicates results observed in visual cortex (Fig 6A) revealing a significant effect for the animacy model relative to the other models in those layers that revealed an object/background effect in the previous RSA analysis (Fig 6A, gray-shaded areas). Overall, this result reveals higher sensitivity for the dimension DCNNs were trained on (the object-specific dimension) and points to a potentially critical role that training tasks play in developing domain-specific object spaces in DCNNs. In other words, does training on a different object domain such as scene allow the emergence of an object space that closer resemble what we observed in scene-selective areas? To investigate this, we tested a DCNN (GoogLeNet; see Methods) trained in scene recognition (Scene365, image set).

When comparing the same architecture (GoogLeNet) on two different training tasks (object recogniton and scene recogniton) we note that only the object recognion task revealed some effects similar to those obseved in visual cortex (Fig 6A, GoogLeNet). On the contrary, the DCNN trained in scene recognition, throughout its layers, did not reveal an object space that resembles that observed in scene-selective areas (Fig 6B): none of the models differ from each other in the mid/high-level layers. This result is not totally unexpected given that scene recognition tasks requires information than might be substantially different from the information necessary to support computations regarding scene navigation (57).

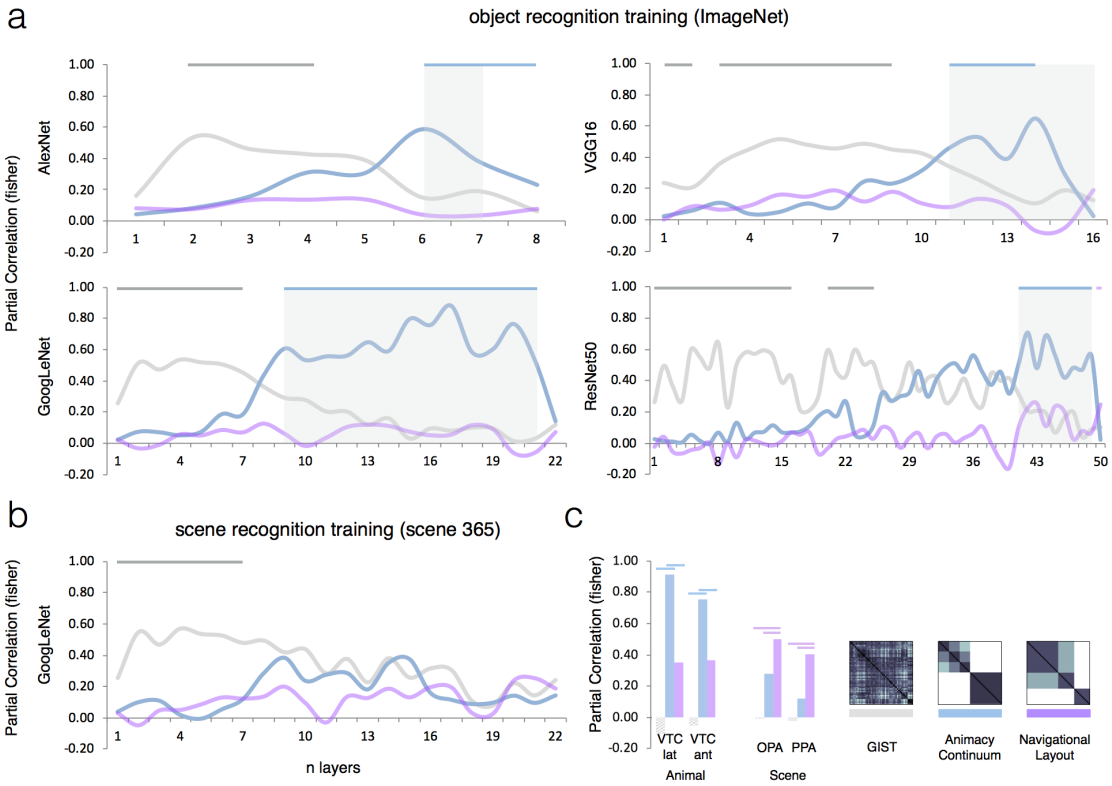

Fig. 6. Multiple object spaces in the visual cortex but not in DCNNs. *The RSA results for the 3 models (GIST, animacy continuum, navigational layout) are shown for (a, b) DCNNs and (c) group-averaged brain data. DCNNs were trained on (a) object recognition (ImageNet), or (b) scene recognition (Scene 365). In each graph, color-coded lines indicate the network's layers where each model significantly outperformed the remaining models. For comparison with the first RSA analysis, gray shaded areas indicate the network's layers in which the object/background model significantly outperformed the remaining models (see Fig. 2 and 3). For the neural data, only domain-specific ROIs (animal and scene selective) where the object/background model reached significance were included (see Fig. 2A and 3). Color-coded lines on top of bar/graphs indicate significant effects ($p < 0.001$) computed with pairwise permutations tests (1000 randomizations of stimulus labels).*

Discussion

We investigated how human and machine object vision deal with real-world object-background regularities and to what extent the latter mimics human visual computations. We measured the representational similarity within-category domains (objects and scenes) as well as across domains, in a stimulus set that includes object-scene pairs that often co-occur in the visual environment. Results revealed that artificial systems are able to capture the conceptual representations humans have about the world but the information processing leading up to this similar endpoint is very different. In the visual cortex, object and background information is processed in separated pathways, which reveal rich domain-specific representational contents for animal (animacy continuum) and scene (navigational layout) processing. Furthermore, interaction between object and scene components was observed at a later processing stage in areas that contribute to goal-directed behavior. Conversely, DCNNs, at their higher processing stage, which typically show the highest correspondence with VTC representations (44), lack a fundamental object vision computation, namely the ability to disentangle objects from the background. An additional study modulating object-scene co-occurrence during training revealed how this entangled representation depends upon training history. Finally, even those DCNNs' layers that showed a degree of object/background separation were not able to mimic the human-like richness of representational content observed in visual cortex, not even when DCCNs were trained on domain-specific recognition tasks (i.e., scene recognition).

The primate brain is, at least in part, characterized by separated modules for domain-specific processing (47). These modules are relatively independent such that confined lesions disrupt computations for one domain (e.g., animate entities) leaving unaffected other domains (e.g., inanimate objects; 48). In visual cortex, domain-specific areas encode object dimensions tailored to support specific computations (36, 49-52). As an example, view-invariant features represented in face- (53) and hand-selective regions (54) reflect domain-specific computations: the former to support identity recognition (50), the latter to support action understanding (55). In agreement, our results show that in addition to a large division between animal and scene representations, within each domain, representational content reflects the type of computations these networks support: animacy features in animal-selective areas (6, 7, 56) and layout navigational properties in scene-selective areas (9, 10, 57). We can show this representational diversity because in our study we included separate behavioral-relevant dimensions for objects (i.e., animacy continuum) and background scenes (i.e., navigational properties), while this was typically not done in previous studies. Furthermore, by including objects and backgrounds that co-occur in the environment, we also demonstrate that at the level of domain-specific representations there is not yet a strong effect of such statistical regularities.

This is not to say that there is no interaction between object and scene representations in human object vision, which in fact, has been reported in previous studies (13, 15, 57, 58). Nevertheless, the nature of such an interaction is still debated and it might occur on top of domain modularity for objects and scenes (32, 59). Our study points in this direction, showing that, ultimately, all such interaction effects ride on top of a fundamental division of labor for object versus background properties.

The information processing strategy is very different in DCNNs. On the positive side, they pick up typical object-background regularities reflecting conceptual knowledge we have about the world. Indeed, when we manipulated object-background systematically in a smaller-scale training regime, even arbitrary object-background associations resulted in strong effects of such associations upon the representations that these networks develop (Fig. 5). This result highlights the critical role played by training on the resulting object space. In the context of a one-image, one-label approach, any pixel in the image might contain features useful to recognize objects. When such regularities are prevalent in a dataset, as it is for natural image databases, shortcuts can be taken (46). In these situations, background information becomes as relevant (or even more relevant) as the object to be recognized (88).

The effect of background on the object space in DCNN fully-connected layers raises the question how much of the previously reported domain-specific effects (e.g., animacy division) observed in these same layers are due to the sensitivity of these layers to represent the scenes in addition to the objects (44, 60, 61). Such questions arise because apparently DCNNs end up with representations in which object and scene information is much more entangled than what we observe in the human brain. One possibility is that DCNNs represent animals separated from inanimate objects because the former is typically seen in specific backgrounds. Taking such shortcuts (46) can, sometimes, be advantageous such as the case shown here, where DCNNs acquired human-like "knowledge" for which they were not trained. However, in general, studies show that relying too much on background information can result in remarkable non-human-like errors. Examples have been reported where DCNNs accuracy for animal recognition drops in unfamiliar backgrounds (62), large animals go undetected in unlikely environments such as a living room (63), or more simply, object classification might be primarily driven by object-irrelevant information present in the image, which happen to correlate with categorization-relevant features (41).

The failure of DCNNs to separate objects from their background does not mean that they cannot develop interesting representations. Already in our study, the object space in mid-layers, revealed evidence for a degree of animacy continuum. Nevertheless, what these representations lack is the diversity and division of labor that we see in the human brain (Fig 6). They can be trained in separate domains though. Training a network on a specific object domain (e.g., faces and objects) leads to diverging object spaces in late DCNNs' layers (66), which however, do not generalize well to other domains (67, 68). Together, this highlights the advantages of domain-specialized modules evolved by biological vision (47). Furthermore, it shows that while these networks can learn domain-specific representations, they have not suddenly received more insight about the structure of visual images, but rather a more restricted training space that was easier to solve. To account for understanding of figure-ground segregation networks can be specifically trained to do such segregation or helped by combining an image classification network with image segmentation tools (67). Yet, in all these cases of networks trained on specific problems, we might end up with a network that only 'understands' a small part of the visual world.

This problem is not one that is simply solved by changing network architecture. Recent comparisons between human and machine vision have drawn attention to the need of including recurrence (26, 27). While this is an important architectural feature, it can be questioned how

fundamental it is for the type of representations that an information processing system can develop. In particular, a shallower recurrent network can be unfolded over time into a deeper feed-forward network (68). Seijdel and colleagues (2020)(69) has empirically demonstrated that scene segmentation, which is resolved by recurrent processing in the primate brain, can be resolved by deeper neural network architectures. From that perspective we believe that the difference between human and machine vision that we have demonstrated in the present study might be even more fundamental. As a specific example, Seijdel and colleagues (2020) noted quite good performance in scene segmentation for ResNet models with 34 and more layers. In our brain-DCNN comparisons, we found that even a ResNet50 model with 50 layers did not show more diverse intermediate representations compared to the shallower models. Also, the final fully connected ResNet50 layers that were found to show categorical representations in a previous study that dissociated category information from visual shape features (44) are the same layers that show strong object-background co-occurrence effects in the current study. Thus, also the deeper DCNNs that can, to some degree, make up for absent recurrent connections fail to capture the richness of information processing that underlies human understanding of natural images that include objects embedded in scenes.

In sum, our study demonstrates the importance of a unique aspect of human information processing relative to machine vision: human information processing represents an untangling of object and background representations. Most likely, this is related to the fact that the human brain has evolved and has been trained to implement a wide variety of tasks. As a consequence, humans understand that an object is not just a collection of features, which could erroneously lead to the assumption that green leaves are inherently linked to ladybugs or to adversarial errors such as classifying a set of yellow and black stripes as a school bus (70). While the human brain is definitely also prone to the use of shortcuts and so-called heuristics (71), the presence of multiple streams of processing and of a multitude of modular systems has the potential of limiting the impact of such shortcuts. It is a major challenge for the future to develop neural network models with a similar richness of representational content.

Methods and Materials

Participants

We recruited 21 participants (7 females, mean age 29). All participants took part in the behavioral test and in 2 functional neuroimaging sessions. Due to technical problems, 1 session from 2 participants and 2 sessions from 1 participant went lost. Additionally, due to excessive head motion, all data from 1 participant and 4 runs from an additional participant were excluded from the data analyses. The head motion exclusion criterion was set to 2 mm (equal to 1 voxel size) and defined before data collection. All participants signed the informed consent approved from the ethic committee at KU Leuven.

Stimuli

We constructed a stimulus set with contextual-related pairs of images (Fig. 1, left). For each contextual pair (1 background and 1 animal), we selected 6 conditions, each containing 4 examples. The animal conditions were depicted on a neutral background to avoid low-level associations between animals and their typical backgrounds. To balance the animals' shape information, we included 2 similarly colorful (red) rounded shaped and small size animals (clownfish and ladybug), 2 birds (passerine and seagull) and 2 mammals with similar body shape (polar bear and gorilla). Thus, conditions within each animal pair are matched in terms of taxonomic and visual information (the two birds have a similar shape, visual features, and both fly), but are specifically associated with a different background (passerine birds and trees; seagulls and coastal landscapes): one was associated with a watery background (polar bear and ice landscapes) and one with a greenery background (gorilla and jungle). Finally, each stimulus domain can be characterized by a functional-relevant dimension: the animacy continuum for the animal conditions (fish/bug, birds, mammals; 37) and the degree of navigational properties (9, 10) for the background conditions (seashores, ice landscapes, forest jungles versus anemones, leaves, tree branches).

fMRI data

The fMRI data was acquired in two separate sessions within 2 weeks. Each session comprised 8 runs for the main experiment and 2 runs for the functional localizer. For the main experiment, each run lasted 6 min and 52 sec (206 volumes). For each run, the 48 image trials and 16 fixation trials were presented twice in a random order. Each trail was presented for 1500 ms and followed by a fixation screen for 1500 ms. Each run started and ended with 14 s of fixation. During scanning participants performed a one-back task judging on a scale from 1 to 4 the degree of co-occurrence of each pair of stimuli ("How often do you see these images occurring together?"). The response order was counterbalanced across runs. Before the first scanner session, participants familiarized with the stimuli and performed a similarity judgments task (74) arranging the 48 images (Figure 1, left) according to "the degree to which you see these images occurring together". The resulting behavioral dissimilarity matrices were averaged across participants.

Acquisition parameters. The fMRI scans were acquired on a 3T Philips scanner with a 32-channel coil at the Department of Radiology of the University Hospitals Leuven. MRI volumes were collected using echo planar (EPI) T2*-weighted scans. Acquisition parameters were as follows: repetition time (TR) of 2 s, echo time (TE) of 30 ms, flip angle (FA) of 90°, field of view (FoV) of 224 mm, and matrix size of 112 x 109 and voxel size of 2 x 2 x 2 mm. Each volume comprised 60 axial slices (0.2 mm gap) acquired with a multi-band factor of 2, covering the whole brain. The T1-weighted anatomical images were acquired with an MP-RAGE sequence, with 1 x 1 x 1 mm resolution.

Preprocessing. Before statistical analysis, functional images underwent a standard preprocessing procedure (SPM 12, Welcome Department of Cognitive Neurology) including three-dimensional head-motion corrected (2th degree spine interpolation), coregistration to the individual anatomical images and normalization to an MNI (Montreal Neurological Institute) template. Spatial smoothing by convolution of a Gaussian kernel of 4 mm full width at half-maximum (75) was applied to functional images. For each participant, a general linear model (GLM) was created to model the 48 conditions and the six motion correction parameters (x, y, z for translation and for rotation). Each predictor's time course was modeled for 3 s (stimulus presentation + fixation) by a boxcar function convolved with the canonical hemodynamic response function in SPM.

Regions of interest (ROIs). All ROIs were defined at the group level based on anatomical masks defined with the Neuromorphometrics SPM toolbox, and functional masks defined with an independent localizer. All visually active voxels (all 48 conditions vs baseline) exceeding the uncorrected threshold $p < 0.0001$ within the functional/anatomical masks were included (Fig. 1). The anatomical-based ROIs included 2 visual areas (V1, posterior VTC), and 2 frontoparietal areas, intraparietal sulcus (IPS) and dorsal prefrontal cortex (DPFC). The functionally-defined ROIs included animal- and scene-selective areas defined with the contrast animals versus scenes and scenes versus animals, respectively by means of an independent localizer. Animal-selective areas were defined in lateral and ventral VTC. Scene-selective areas were defined in parahippocampal gyrus (PPA), retrosplenial cortex (RSC), and transverse occipital sulcus (TOS).

DCCN data

We extracted the vector spaces for images in our stimulus set from 4 recent DCNNs with different depth: Alexnet (76), VGG-19 (77), GoogLeNet (78) and ResNet 50 (79). DCNNs consist of various feedforward processing stages including (1) convolutional layers, (2) rectified linear unit activation, (3) max pooling layers, and (4) fully connected layers (resembling a multilayer perceptron). For each DCNNs and each convolutional layer, we extracted the vector space for all 48 stimuli by means of the Deep Learning Toolbox (https://uk.mathworks.com/products/deep-learning.html) in MATLAB. All DCNNs were pre-trained on 1.2 million natural images belonging to 1000 classes for the ImageNet dataset (http://www.image-net.org/). In addition, GoogLeNet was also pre-trained on the Place365 dataset which contains 10 million images comprising more than 400 scene classes. Each scene contains between 5000 to 30,000 training images.

Statistical Analysis

We used representational similarity analysis (RSA; 28) to compare the neural activity patterns from the brain's ROIs and DCNNs' layers to our experimental models. CoSMoMVPA (80) was used to create the representational dissimilarity matrices (RDMs) for brain and DCNNs data. For neural data, for each voxel within a given ROI, GLM parameter estimates for each stimulus condition (relative to baseline) were extracted for each participant and each run and normalized by subtracting the mean response across all conditions. Similarly, for each DCNNs, the vector features for each stimulus condition within each layer were normalized by subtracting the mean feature activations across all conditions. Next, we computed Pearson's correlation across voxels (fMRI) or units (DCNNs) between all condition pairs and converted the resulting (fisher transformed) correlation matrix (0.5*log[(1r)/(1r)]) into a representational dissimilarity matrix (RDM; 1 minus Pearson's r) and used it as input for the RSA analysis. We performed two main analyses. To test our main experimental hypothesis, we run two separated RSAs (brain and DCNNs) including the following models: the object/background model which assumed high similarity for stimuli within each domain (animals and backgrounds) and the co-occurrence model which assumed high similarity for each animal-background contextual-related pair. These two models were derived from our two alternative hypotheses and were orthogonal to each other (r=-0.05). Two additional control models were included: the GIST model (29) to capture low-level image information and the basic-category model to account for within condition visual similarities. For the neural data we performed two RSAs: one with group averaged data and one with single subject data. For the latter one we calculated reliability of RDMs which indicates the highest expected correlation in a region given its signal-to-noise ratio. For each subject and each ROI, the RDM was correlated with the averaged RDM of the remaining participants. The resulting correlation values (averaged across participants) capture noise inherent to a single subject as well as noise caused by inter-subject variability. For the two additional follow-up RSAs (brain and CNNs) the following models were included: the animacy model which captures the well-known animacy continuum (37) described in VTC and the navigational layout model which instead captures the degree to which a scene contains features useful for navigation. Before statistical analyses, results from the RSAs were Fisher transformed and tested with ANOVAs and pairwise t tests when individual-subject data were available, or permutation tests for DCNNs and fMRI group averaged data (1000 and 10000 iterations of permuting stimulus labels for DCNNs and brain data, respectively).

DCNN experiment with congruency manipulation

Construction of the training datasets: Custom datasets were constructed by joining masked images of objects with inverse-masked images of scenes obtained from two publicly available datasets. Images of 6 object categories ('bus', 'clock', 'dog', 'giraffe', 'television', 'umbrella') were obtained from the Microsoft COCO dataset that consists of over 300.000 images labelled with more than 2 million object instances (81). Each object instance is associated with a mask which allows extraction of object-representing pixels. Objects of small size (ratio between object size versus total image size smaller than 3%) were enlarged and objects of object-image ratio smaller than 0.15% omitted in order to avoid unsatisfactory classification performance on images with hardly any object-associated pixels (82). Scene images of 6 categories ('church indoor', 'sand desert, 'glacier', 'ocean underwater', 'broadleaf forest', 'landfill') were obtained from the Places dataset that consists of more than 2 million images of 205 scene categories, most of them containing several thousand unique images (83). This allowed generation of a total of 10500

training images (1750 per object category) of 256 x 256 x 3 pixels per dataset, in which each image showed an object of one of the 6 object categories embedded in a scene of one of the 6 scene categories.

Four different datasets of 0%, 58.3%, 83.3% and 100% of object-scene co-occurrence were constructed by manipulating the likelihood of objects being joined with their adhering background type. Each dataset was constructed out of the same set of object and background images that were, however, joined in a different fashion. To illustrate this on the example of 83.3% object-scene co-occurrence dataset, backgrounds of every 1st, 2nd, 3rd and 4th image were object-specific (80%) and background of every 5th image was belonging to either one of the six background categories (20%). Thus, there is an additional 1/6 chance that randomly chosen backgrounds will come from the object-specific category (adding additional 3.3% to the total object-scene specificity of the dataset). In the 0% object-scene co-occurrence, none of the 6 scene categories is more likely to be associated with a particular object category than the other scene categories (all have a 16.7% chance).

Training of the models: Five individual DCNN models were trained on each of the constructed datasets. The models followed the standard AlexNet architecture with five convolutional and three fully-connected layers with one minor difference in input layer accepting images of size 256 x 256 x 3 instead of 224 x 224 x 3 pixels, as originally defined by Krizhevsky and colleagues (2012) (76) to better fit the size of our image stimuli. All models were trained for 150 epochs in batches of 75 training images, with the Adam optimizer and categorical cross-entropy loss function. A dropout regularization rate of 0.5 was added to the first two fully-connected layers. Additionally, image augmentation was employed to improve generalization of the trained models. At the end of the training, only the weights of the best performing epoch (lowest validation loss calculated on a total 750 validation images, 125 per object category) were stored. All models were implemented in Python utilizing functions of the Keras library.

Representational similarity analysis with congruency manipulation: Additionally, representational similarity analysis was performed to compare patterns of inner neural representations of the trained networks with our experimental models. In order to do so, neural activations of each networks' 7th layer (fully connected) were extracted by presenting them with 48 image stimuli, each sprouting 4096 neural activations. This set of stimuli consisted of 24 images of objects on neutralized backgrounds and 24 images of scenes without the object entities (4 for each object category and 4 for each scene category). Neutralized backgrounds were obtained by averaging pixels of 6 scene images randomly chosen from Places dataset (83), each of them belonging to one of the above specified scene categories. This was done in order to provide objects with a background that possessed no relevant scene information yet were not too atypical/homogeneous and thus too different from training examples (usage of monotonous backgrounds was found to be problematic as it significantly impacted internal representations of the networks). Next, correlations between internal activations for each stimulus were obtained by calculation of Pearson correlation coefficients. This sprouted a single RSA matrix for each model which was in the end transformed into an RDM matrix by subtracting it from a unit matrix (with all values equal one) of a matching dimension.

Data and Code Availability

All types of brain images and statistics data are available from the authors upon request. The stimulus set and data (RDMs) used for final statistics will be made available through the Open Science Framework at xxx.

Acknowledgments

S.B. was funded by FWO (Fonds Wetenschappelijk Onderzoek) through a postdoctoral fellowship (12S1317N) and a Research Grant (1505518N). J.M. was supported by the Ad futura Scholarship of the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia. H.O.B. was supported by the KU Leuven Research Council (C14/16/031) and FWO/FNRS EOS (Excellence of Science) grant HumVisCat (nr. 30991544).

References

1. M. L. Vo, S. E. Boettcher, D. Draschkow, Reading scenes: how scene grammar guides attention and aids perception in real-world environments. Curr Opin Psychol 29, 205-210 (2019). https://doi.org/10.1016/j.copsyc.2019.03.009
2. D. Marr, Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. San Francesco. W H Freeman (1982). DOI:10.7551/mitpress/9780262514620.001.0001
3. J. Wagemans et al., A century of Gestalt psychology in visual perception: I. Perceptual grouping and figure-ground organization. Psychol Bull 138, 1172-1217 (2012). DOI: 10.1037/a0029333
4. R. Epstein, N. Kanwisher, A cortical representation of the local visual environment. Nature 392, 598-601 (1998). https://doi.org/10.1038/33402
5. R. Malach et al., Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. Proceedings of the National Academy of Sciences of the United States of America 92, 8135-8139 (1995). doi: 10.1073/pnas.92.18.8135
6. S. Bracci, J. B. Ritchie, I. Kalfas, H. P. Op de Beeck, The Ventral Visual Pathway Represents Animal Appearance over Animacy, Unlike Human Behavior and Deep Neural Networks. The Journal of neuroscience : the official journal of the Society for Neuroscience 39, 6513-6525 (2019). DOI: https://doi.org/10.1523/JNEUROSCI.1714-18.2019
7. J. B. Ritchie et al., Untangling the animacy organization of occipitotemporal cortex. BioRxiv 2020.07.17.206896 (2020). DOI: https://doi.org/10.1523/JNEUROSCI.2628-20.2021
8. S. G. Wardle, J. Taubert, L. Teichmann, C. I. Baker, Rapid and dynamic processing of face pareidolia in the human brain. Nature communications 11, 4518 (2020). https://doi.org/10.1038/s41467-020-18325-8
9. D. J. Kravitz, C. S. Peng, C. I. Baker, Real-world scene representations in high-level visual cortex: it's the spaces more than the places. The Journal of neuroscience : the official journal of the Society for Neuroscience 31, 7322-7333 (2011). DOI: https://doi.org/10.1523/JNEUROSCI.4588-10.2011
10. S. Park, T. F. Brady, M. R. Greene, A. Oliva, Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. The Journal of neuroscience : the official journal of the Society for Neuroscience 31, 1333-1340 (2011). DOI: https://doi.org/10.1523/JNEUROSCI.3885-10.2011
11. I. Biederman, Perceiving real-world scenes. Science 177, 77-80 (1972). DOI: 10.1126/science.177.4043.77
12. T. E. Palmer, The effects of contextual scenes on the identification of objects. Mem Cognit 3, 519-526 (1975). DOIhttps://doi.org/10.3758/BF03197524
13. M. Bar, Visual objects in context. Nature reviews. Neuroscience 5, 617-629 (2004). DOIhttps://doi.org/10.1038/nrn1476
14. M. Bar, E. Aminoff, Cortical analysis of visual context. Neuron 38, 347-358 (2003). DOI: 10.1016/s0896-6273(03)00167-3
15. T. Brandman, M. V. Peelen, Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding. The Journal of neuroscience : the official journal of the Society for Neuroscience 37, 7700-7710 (2017). DOI: https://doi.org/10.1523/JNEUROSCI.0582-17.2017
16. A. Oliva, A. Torralba, The role of context in object recognition. Trends in cognitive sciences 11, 520-527 (2007). https://doi.org/10.1016/j.tics.2007.09.009
17. Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE conference on computer vision and pattern recognition 86, 2278 - 2324 (1998). DOI: 10.1109/5.726791
18. M. Riesenhuber, T. Poggio, Hierarchical models of object recognition in cortex. Nature neuroscience 2, 1019-1025 (1999). DOIhttps://doi.org/10.1038/14819
19. T. Serre, A. Oliva, T. Poggio, A feedforward architecture accounts for rapid categorization. Proceedings of the National Academy of Sciences of the United States of America 104, 6424-6429 (2007). https://doi.org/10.1073/pnas.0700622104
20. Y. LeCun, Y. Bengio, G. Hinton, Deep learning. Nature 521, 436-444 (2015). DOIhttps://doi.org/10.1038/nature14539
21. T. Serre, Deep Learning: The Good, the Bad, and the Ugly. Annual review of vision science 5, 399-426 (2019). https://doi.org/10.1146/annurev-vision-091718-014951
22. D. Silver et al., A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. Science 362, 1140-1144 (2018). DOI: 10.1126/science.aar6404
23. K. He, X. Zhang, S. Ren, J. Sun, Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. arXiv 1502.01852v1 (2015).
24. S. M. Khaligh-Razavi, N. Kriegeskorte, Deep supervised, but not unsupervised, models may explain IT cortical representation. PLoS computational biology 10, e1003915 (2014). https://doi.org/10.1371/journal.pcbi.1003915
25. D. L. Yamins et al., 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, 8619-8624 (2014). https://doi.org/10.1073/pnas.1403112111
26. G. Kreiman, T. Serre, Beyond the feedforward sweep: feedback computations in the visual cortex. Ann N Y Acad Sci 1464, 222-241 (2020).
27. T. C. Kietzmann et al., Recurrence is required to capture the representational dynamics of the human visual system. Proceedings of the National Academy of Sciences of the United States of America 116, 21854-21863 (2019).

28. N. Kriegeskorte, M. Mur, P. Bandettini, Representational similarity analysis - connecting the branches of systems neuroscience. Frontiers in systems neuroscience 2, 4 (2008).
29. A. Oliva, A. Torralba, Modeling the shape of the scene: A holistic representation of the spatial envelope. Int. J. Comput. Vis. 42, 145–175 (2001).
30. Groen, II, S. Ghebreab, H. Prins, V. A. Lamme, H. S. Scholte, From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. The Journal of neuroscience : the official journal of the Society for Neuroscience 33, 18814-18824 (2013).
31. T. W. James, J. Culham, G. K. Humphrey, A. D. Milner, M. A. Goodale, Ventral occipital lesions impair object recognition but not object-directed grasping: an fMRI study. Brain : a journal of neurology 126, 2463-2475 (2003).
32. D. D. Dilks, J. B. Julian, A. M. Paunov, N. Kanwisher, The occipital place area is causally and selectively involved in scene perception. The Journal of neuroscience : the official journal of the Society for Neuroscience 33, 1331-1336a (2013).
33. M. Bar, E. Aminoff, D. L. Schacter, Scenes unseen: the parahippocampal cortex intrinsically subserves contextual associations, not scenes or places per se. The Journal of neuroscience : the official journal of the Society for Neuroscience 28, 8539-8544 (2008).
34. F. Oppermann, U. Hassler, J. D. Jescheniak, T. Gruber, The rapid extraction of gist-early neural correlates of high-level visual processing. Journal of cognitive neuroscience 24, 521-529 (2012).
35. J. Duncan, The multiple-demand (MD) system of the primate brain: mental programs for intelligent behaviour. Trends in cognitive sciences 14, 172-179 (2010).
36. R. A. Epstein, C. I. Baker, Scene Perception in the Human Brain. Annual review of vision science 5, 373-397 (2019).
37. A. C. Connolly et al., The representation of biological classes in the human brain. The Journal of neuroscience : the official journal of the Society for Neuroscience 32, 2608-2618 (2012).
38. R. M. Cichy, D. Kaiser, Deep Neural Networks as Scientific Models. Trends in cognitive sciences 23, 305-317 (2019).
39. G. W. Lindsay, Convolutional Neural Networks as a Model of the Visual System: Past, Present, and Future. Journal of cognitive neuroscience 10.1162/jocn_a_01544, 1-15 (2020).
40. M. Schrimpf et al., Artificial neural networks accurately predict language processing in the brain. BioRxiv 10.1101/2020.06.26.174482 (2020).
41. J. R. Zech et al., Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. PLoS Med 15, e1002683 (2018).
42. R. M. Cichy, A. Khosla, D. Pantazis, A. Torralba, A. Oliva, Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific reports 6, 27755 (2016).
43. H. Hong, D. L. Yamins, N. J. Majaj, J. J. DiCarlo, Explicit information for category-orthogonal object properties increases along the ventral stream. Nature neuroscience 19, 613-622 (2016).
44. A. A. Zeman, J. B. Ritchie, S. Bracci, H. Op de Beeck, Orthogonal Representations of Object Shape and Category in Deep Convolutional Neural Networks and Human Visual Cortex. Scientific reports 10, 2453 (2020).
45. I. Kalfas, S. Kumar, R. Vogels, Shape Selectivity of Middle Superior Temporal Sulcus Body Patch Neurons. eNeuro 4 (2017).
46. R. Geirhos et al., Shortcut learning in deep neural networks. Nat Mach Intell 2, 665–673 (2020).
47. N. Kanwisher, Functional specificity in the human brain: a window into the functional architecture of the mind. Proceedings of the National Academy of Sciences of the United States of America 107, 11163-11170 (2010).
48. A. Caramazza, J. R. Shelton, Domain-specific knowledge systems in the brain the animate-inanimate distinction. Journal of cognitive neuroscience 10, 1-34 (1998).
49. S. Anzellotti, S. L. Fairhall, A. Caramazza, Decoding representations of face identity that are tolerant to rotation. Cerebral cortex 24, 1988-1995 (2014).
50. L. Chang, D. Y. Tsao, The Code for Facial Identity in the Primate Brain. Cell 169, 1013-1028 e1014 (2017).
51. Groen, II, E. H. Silson, C. I. Baker, Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 372 (2017).
52. H. P. Op de Beeck, J. Haushofer, N. G. Kanwisher, Interpreting fMRI data: maps, modules and dimensions. Nature reviews. Neuroscience 9, 123-135 (2008).
53. N. Kanwisher, J. McDermott, M. M. Chun, The fusiform face area: a module in human extrastriate cortex specialized for face perception. The Journal of neuroscience : the official journal of the Society for Neuroscience 17, 4302-4311 (1997).
54. S. Bracci, M. Ietswaart, M. V. Peelen, C. Cavina-Pratesi, Dissociable neural responses to hands and non-hand body parts in human left extrastriate visual cortex. Journal of neurophysiology 103, 3389-3397 (2010).
55. S. Bracci, A. Caramazza, M. V. Peelen, View-invariant representation of hand postures in the human lateral occipitotemporal cortex. NeuroImage 181, 446-452 (2018).
56. D. Proklova, M. A. Goodale, The role of animal faces in the animate-inanimate distinction in the ventral temporal cortex. bioRxiv 2020.10.08.330639 (2020).
57. K. Dwivedi, R. M. Cichy, G. Roig, Unraveling Representations in Scene-selective Brain Regions Using Scene-Parsing Deep Neural Networks. Journal of cognitive neuroscience 10.1162/jocn_a_01624, 1-12 (2020).
58. G. Ganis, M. Kutas, An electrophysiological study of scene effects on object identification. Brain Res Cogn Brain Res 16, 123-144 (2003).

59. N. Seijdel, S. Jahfari, I. I. A. Groen, H. S. Scholte, Low-level image statistics in natural scenes influence perceptual decision-making. Scientific reports 10, 10573 (2020).
60. M. Wischnewski, M. V. Peelen, Causal Evidence for a Double Dissociation between Object- and Scene-Selective Regions of Visual Cortex: A Preregistered TMS Replication Study. The Journal of neuroscience : the official journal of the Society for Neuroscience 41, 751-756 (2021).
61. K. M. Jozwik, N. Kriegeskorte, K. R. Storrs, M. Mur, Deep Convolutional Neural Networks Outperform Feature-Based But Not Categorical Models in Explaining Object Similarity Judgments. Frontiers in psychology 8, 1726 (2017).
62. J. Kubilius, S. Bracci, H. P. Op de Beeck, Deep Neural Networks as a Computational Model for Human Shape Sensitivity. PLoS computational biology 12, e1004896 (2016).
63. S. Beery, G. Van Horn, P. Perona, Recognition in terra incognita. . In Proceedings of the European Conference on Computer Vision, 456–473 (2018).
64. A. Rosenfeld, R. Zemel, J. K. Tsotsos, The elephant in the room. . arXiv 1808.03305 (2018).
65. S. Ullman, High-Level Vision: Object Recognition and Visual Cognition. . Cambridge, MA: MIT Press (1996).
66. A. Kallmayer, J. Prince, T. Konkle, Comparing representations that support object, scene, and face recognition using representational trajectory analysis. Journal of Vision (2020).
67. K. Dobs, A. Kell, I. Palmer, M. A. Cohen, N. Kanwisher, Why Are Face and Object Processing Segregated in the Human Brain? Testing Computational Hypotheses with Deep Convolutional Neural Networks. Cognitive Computational Neuroscience Conference, Berlin, Germany (2019).
68. N. M. Blauch; et al., Assessing the similarity of cortical object and scene representations through cross-validated voxel encoding models. Journal of Vision Vol.19, (2019).
69. M. Zhang, C. Tseng, G. Kreiman, P. v. o. r. i. context., Putting visual object recognition in context. CVPR, 12985–12994 (2020).
70. Q. Liao, T. Poggio, Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. arXiv 1604.03640 (2016).
71. N. Seijdel, N. Tsakmakidis, E. H. F. de Haan, S. M. Bohte, H. S. Scholte, Depth in convolutional neural networks solves scene segmentation. PLoS computational biology 16, e1008022 (2020).
72. A. Nguyen, J. Yosinski, J. D. Clune, Deep Neural Networks are Easily Fooled: High Confidence Predictions for Unrecognizable Images Computer Vision and Pattern Recognition (CVPR), IEEE. (2015).
73. A. Tversky, D. Kahneman, Judgment under Uncertainty: Heuristics and Biases. Science 185, 1124-1131 (1974).
74. N. Kriegeskorte, M. Mur, Inverse MDS: Inferring Dissimilarity Structure from Multiple Item Arrangements. Frontiers in psychology 3, 245 (2012).
75. H. P. Op de Beeck, Against hyperacuity in brain reading: spatial smoothing does not hurt multivariate fMRI analyses? NeuroImage 49, 1943-1948 (2010).
76. A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Information Processing Systems 25 (2012).
77. K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition. arXiv 1409.1556 (2014).
78. C. Szegedy et al., Going Deeper With Convolutions. arXiv 1409.4842v1 (2015).
79. K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognitio. IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 770-778 (2016).
80. N. N. Oosterhof, A. C. Connolly, J. V. Haxby, CoSMoMVPA: Multi-Modal Multivariate Pattern Analysis of Neuroimaging Data in Matlab/GNU Octave. Frontiers in neuroinformatics 10, 27 (2016).
81. T.-Y. Lin et al., Microsoft COCO: Common Objects in Context. European Conference on Computer Vision (ECCV) (2014).
82. C. Chen, M. Y. Liu, O. Tuzel, J. Xiao, R-CNN for Small Object Detection. Asian Conference on Computer Vision (2017).
83. B. Zhou, B. Lapedriza, A. Khosla, A. Oliva, T. A., Places: A 10 million image database for scene recognition. IEEE TPAM (2017).
84. R. Geirhos, K. Narayanappa, B. Mitzkus, T. Thieringer, M. Bethge, F.A. Wichmann, W. Brendel, Partial success in closing the gap between human and machine vision. arXiv:2106.07411 (2021).
85. C. Zhuang, S. Yan, A. Nayebi, M. Schrimpf, M. C. Frank, J. J. DiCarlo, D.L.K. Yamins, Unsupervised neural network models of the ventral visual stream. PNAS, 118, 3 (2021).
86. S. Bakhtiari, P. Mineault, T. Lillicrap, C.C. Pack, B.A. Richards, The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning. bioRxiv; https://doi.org/10.1101/2021.06.18.448989 (2021).
87. M.F. Bonner, R. A. Epstein. Object representations in the human brain reflect the co-occurrence statistics of vision and language. Nature Communications 12:4081 (2021).
88. Z. Zhu, L. Xie, A. Yuille, Object Recognition with and without Objects. Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence (IJCAI, 2017).
89. M. F. Bonner and R. A. Epstein, Coding of navigational affordances in the human visual system. PNAS, (2017).
90. D.L. Yamins et al., Performance-optimized hierarchical models predict neural responses in higher visual cortex. Proc. Natl. Acad. Sci. U.S.A. 111, 8619–8624 (2014).

Author Contributions:
Conceptualization: S.B. and H.O.B;
Methodology: S.B., J.M., AZ, and G.L.;
Investigation: S.B., and J.M.;
Visualization: S.B.;
Writing—original draft: S.B., J.M. and H.O.B;
Writing—review & editing: S.B., J.M. and H.O.B.

Competing Interest Statement:
All other authors declare they have no competing interests.