Human Visual Cortex and Deep Convolutional Neural Network Care Deeply about Object Background

Jessica Loke*, Noor Seijdel*, Lukas Snoek, Lynn K. A. Sörensen, Ron van de Klundert, Matthew van der Meer, Eva Quispel, Natalie Cappaert, and H. Steven Scholte

Abstract

■ Deep convolutional neural networks (DCNNs) are able to partially predict brain activity during object categorization tasks, but factors contributing to this predictive power are not fully understood. Our study aimed to investigate the factors contributing to the predictive power of DCNNs in object categorization tasks. We compared the activity of four DCNN architectures with EEG recordings obtained from 62 human participants during an object categorization task. Previous physiological studies on object categorization have highlighted the importance of figure-ground segregation—the ability to distinguish objects from their backgrounds. Therefore, we investigated whether figure-ground segregation could explain the predictive power of DCNNs. Using a stimulus set consisting of identical target objects embedded in different backgrounds, we examined the influence of object background versus object category within both EEG and DCNN activity. Crucially, the recombination of naturalistic objects and experimentally controlled backgrounds creates a challenging and naturalistic task, while retaining experimental control. Our results showed that early EEG activity (< 100 msec) and early DCNN layers represent object background rather than object category. We also found that the ability of DCNNs to predict EEG activity is primarily influenced by how both systems process object backgrounds, rather than object categories. We demonstrated the role of figure-ground segregation as a potential prerequisite for recognition of object features, by contrasting the activations of trained and untrained (i.e., random weights) DCNNs. These findings suggest that both human visual cortex and DCNNs prioritize the segregation of object backgrounds and target objects to perform object categorization. Altogether, our study provides new insights into the mechanisms underlying object categorization as we demonstrated that both human visual cortex and DCNNs care deeply about object background. ■

INTRODUCTION

Deep convolutional neural networks (DCNNs) have entered the computational modeling scene with high predictive performance of both object category and brain dynamics during object categorization tasks (Schrimpf et al., 2018; Cadieu et al., 2014; Khaligh-Razavi & Kriegeskorte, 2014; Yamins et al., 2014). These predictions on brain dynamics are not limited to low-level image statistics but also include high-level features such as animacy, object category, and semantics (Doerig et al., 2022; Takagi & Nishimoto, 2022; Dwivedi, Bonner, Cichy, & Roig, 2021; Ritchie et al., 2021; Eickenberg, Gramfort, Varoquaux, & Thirion, 2017). In fact, DCNNs' predictive performance on visual processes surpassed hand-engineered, biologically inspired models (e.g., Gabor wavelet filtered, HMAX) because DCNNs are able to achieve high performance on visual tasks (Cichy & Kaiser, 2019; Yamins & DiCarlo, 2016). Traditional mechanistic models generally include few parameters and are tested on simplistic, artificial stimuli such as bar gratings and white noise; in contrast, DCNNs generally include hundreds of thousands to millions of parameters and are tested on complex and naturalistic stimuli such as photographs of real objects or scenes. However, this acclaim is not without criticism; DCNNs have been labeled as "black boxes" (Kay, 2018; Marcus, 2018) as researchers struggled to understand how millions of parameters work together to perform tasks such as object categorization (Scholte, 2018), and also predict brain activity without being trained with brain data (Lillicrap & Kording, 2019).

The criticism toward DCNNs sharpens as studies revealed divergences in categorization strategies between humans and DCNNs—humans and DCNNs make mistakes on different images (Geirhos, Meding, & Wichmann, 2020; Rajalingham et al., 2018; Geirhos et al., 2017), DCNNs have an inherent texture bias whereas humans have an inherent shape bias (Tartaglini, Vong, & Lake, 2022; Baker, Lu, Erlikhman, & Kellman, 2018; Geirhos et al., 2018; Ritter, Barrett, Santoro, & Botvinick, 2017), and DCNNs are susceptible to adversarial attacks imperceptible to humans (Akhtar & Mian, 2018; Goodfellow, Shlens, & Szegedy, 2014). Although these studies point to differences in

University of Amsterdam
*These authors contributed equally/shared first author.

categorization strategies, they do not negate the fact that DCNNs can still produce representations that align with human visual processing (Cao & Yamins, 2021a), as reflected in its high predictive performance of brain dynamics. In other words, even if certain DCNN's categorization outputs from DCNNs are incorrect, we can probe their processing stages to find shared representations with human visual processing, thereby understanding crucial steps for the task (Dwivedi et al., 2021; Truzzi & Cusack, 2020). The right question would then be, "Which representations are useful and robust for solving the task?"

In this study, we investigated the factors leading to DCNNs' high predictive power on human visual processing within an object categorization task, focusing on essential representations for solving the task. Prior research has shown the importance of figure-ground segregation (Roelfsema, 2006; Roelfsema, Lamme, & Spekreijse, 2002)—the ability to distinguish an image's foreground and background (i.e., object and background). This ability is especially crucial when the object and its background share similar features such as line orientations, curvatures, and colors. Both humans and DCNNs showed enhanced performance when presented with presegmented objects compared with objects embedded in backgrounds (Loke et al., 2022; Borji, 2021; Seijdel, Tsakmakidis, de Haan, Bohte, & Scholte, 2020). To investigate this further, we used images with identical target objects embedded in varying background complexities, allowing us to isolate human EEG recordings and DCNN activity related to target object categorical features versus object background. This approach provides a challenging and naturalistic task while still maintaining experimental control and enables us to identify potentially useful representations in object categorization. We opted for EEG recordings over fMRI because of EEG's higher time resolution as previous studies have shown that the effects of backgrounds occur at a limited time window where feedback signals are dominant (Seijdel et al., 2021; Groen et al., 2018; Zipser, Lamme, & Schiller, 1996). Surprisingly, we discovered that both early and late activity in human EEG recordings is largely dedicated to processing object backgrounds. This pattern was mirrored in the activity of DCNNs, which also prioritized object backgrounds over object categories. Our findings suggest that the ability to distinguish between the target object and its background is essential for object categorization.

METHODS

Data

The electrophysiological data, sourced from Seijdel and colleagues (2021), consist of EEG recordings from human participants (n = 62, 18–35 years old). The sample size was determined based on a similar experimental paradigm (Groen et al., 2018). For a brief description of the experimental paradigm and example of stimuli, please see Figure 1.

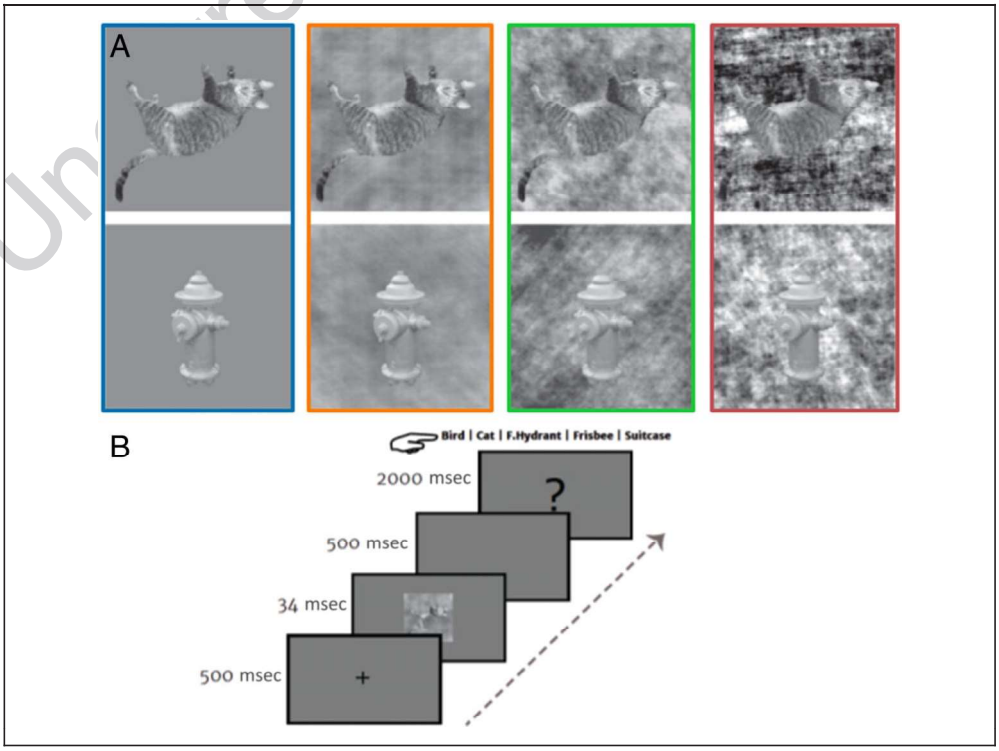

Figure 1. Stimuli sample and experimental paradigm. (A) Two object exemplars (cat and fire hydrant) are displayed across four background types. The first (highlighted in blue) is a uniform gray background, referred to as the "segmented" condition. The second (highlighted in orange), third (highlighted in green), and fourth (highlighted in red) are a low, medium, and high complexity background, respectively. The increasing levels of background complexity makes it increasingly difficult to differentiate the target object from its background. (B) The experimental paradigm had human participants perform an object categorization task. Each trial starts with a fixation cross of 500 msec, followed by a stimulus presentation of 34 msec. The stimulus presentation is followed by a blank screen for 500 msec. Finally, there is a response screen displaying the five object category options for 2000 msec. Participants completed 480 trials—120 trials per image condition. Figure adapted with permission from Seijdel and Loke and colleagues (2021).

Stimuli

The stimuli used consisted of 120 unique target objects (24 per category) from five categories (bird, cat, fire hydrant, frisbee, and suitcase), embedded within four background types (uniform gray background, low complexity, medium complexity, and high complexity), resulting in 480 unique stimuli. The backgrounds were created by phase-scrambling the original image backgrounds to remove information aiding recognition of the target object. The complexity of these phase-scrambled backgrounds varied based on spatial coherence and contrast energy (Scholte, Ghebreab, Waldorp, Smeulders, & Lamme, 2009). The segmented condition does not have phase-scrambled backgrounds but a uniform gray one. The stimuli were presented at a resolution of 512 × 512 pixels.

DCNNs

We selected four established DCNN architectures, commonly used in computational modeling—AlexNet (Krizhevsky, Sutskever, & Hinton, 2012), VGG-16 (Simonyan & Zisserman, 2014), ResNet-18, and ResNet-50 (He, Zhang, Ren, & Sun, 2016). We initialized and trained five different seeds of each network using the ImageNet Large Scale Visual Recognition Challenge 2012 (ILSVRC) data set. These networks were then fine-tuned to the experimental object categories with the Microsoft COCO data set (Lin et al., 2014). We used different seeds to capture variance between different initializations and obtain reliable results (Mehrer, Spoerer, Kriegeskorte, & Kietzmann, 2020). For the initial training on ILSVRC, we used a learning rate of 0.1 (except for VGG-16, which needed a lower learning rate of 0.05) with a learning rate decay of 0.1 every 30 epochs and a weight decay of 1e-4. We also used a stochastic gradient optimizer with a momentum of 0.9. AlexNet, ResNet-18, and ResNet-50 were trained for 150 epochs whereas VGG-16 was trained for 74 epochs. All DCNNs reached similar performance accuracies reported in the original articles. For fine-tuning, we replaced the last fully connected layer and retrained weights from all layers. We fine-tuned the network with a learning rate of 1e-3 with a learning rate decay of 0.1 every seven epochs. The fine-tuning was performed for 20 epochs. We also used a stochastic gradient descent optimizer with a momentum of 0.9 for fine-tuning. In addition to trained networks, we initialized five different seeds of each architecture with no training as untrained networks. All DCNN's training and fine-tuning were done in PyTorch (Paszke et al., 2019).

Analysis: Representational Similarity Analysis

We used the framework of representational similarity analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008) to compare EEG activity with DCNN's activations. RSA is a method of analysis allowing for the comparison between different modalities by first generating a representational structure of the stimuli set as reflected in brain activity (as recorded using EEG sensors) and DCNNs (as reflected through its unit activations), and then comparing both those representational structures. This abstraction from EEG sensors and DCNNs unit activations allows us to compare the transformations performed by both modalities on the stimuli. Using RSA, we obtained time-resolved EEG activity and layer-wise DCNN activations in the form of representational dissimilarity matrices (RDMs). The RDMs consist of pairwise distances computed from multivariate responses (i.e., pattern of EEG activity or pattern of layer-wise DCNN's activations) toward every possible stimulus pair. Pairwise distances were computed as ($1 - \textit{Pearson correlation}$). An entry in the RDM between stimuli A and B would be: $1 - \textit{Pearson correlation of multivariate responses towards stimuli A and B}$; whereas, an entry in the RDM between stimuli A and A would be 0. With 480 unique stimuli (120 unique objects × 4 background types), we obtained 480 × 480 RDMs. In all analyses using RDMs, we used only the upper triangle (excluding the diagonal) because the RDMs are symmetrical.

RDMs of EEG recordings were computed using 22 posterior electrodes (Iz, I1, I2, Oz, O1, O2, POz, PO3, PO4, PO7, PO8, Pz, P1, P2, P3, P4, P5, P6, P7, P8, P9, and P10). These electrodes are chosen to focus on activity from visual processing areas and were confirmed in previous studies (Seijdel et al., 2021; Groen et al., 2018). The electrodes placement followed a 10–10 layout, modified with two additional occipital electrodes (I1 and I2) replacing two frontal electrodes (F5 and F6). RDMs were computed from every time sample from −100 msec to 600 msec relative to stimulus onset. RDMs of DCNN's activations were obtained from activity of all convolutional, pooling, and fully connected layers.

In addition to RDMs from EEG and DCNNs, we also constructed categorical RDMs to evaluate the main effects of our experimental manipulations. We built three categorical RDMs—segmentation, background complexity, and object category (see Figure 2). All three RDMs consisted of binary values: "0" representing pairs from the same group and "1" representing pairs from different groups. Segmentation distinguishes between stimuli with and without backgrounds (see Figure 2A). Background complexity distinguishes between the four background types (see Figure 2B): segmented (no background), low complexity, medium complexity, and high complexity. Object category distinguishes between the five object categories (see Figure 2C). Here, it should be noted that the categorical RDMs of segmentation and background complexity correlate substantially ($r = .45$), because the segmented stimuli all have the same complexity (i.e., 0; see Figure 2A and B). As such, to separate the variance associated with segmentation or background complexity, we performed partial correlations between the categorical RDMs and EEG RDMs.

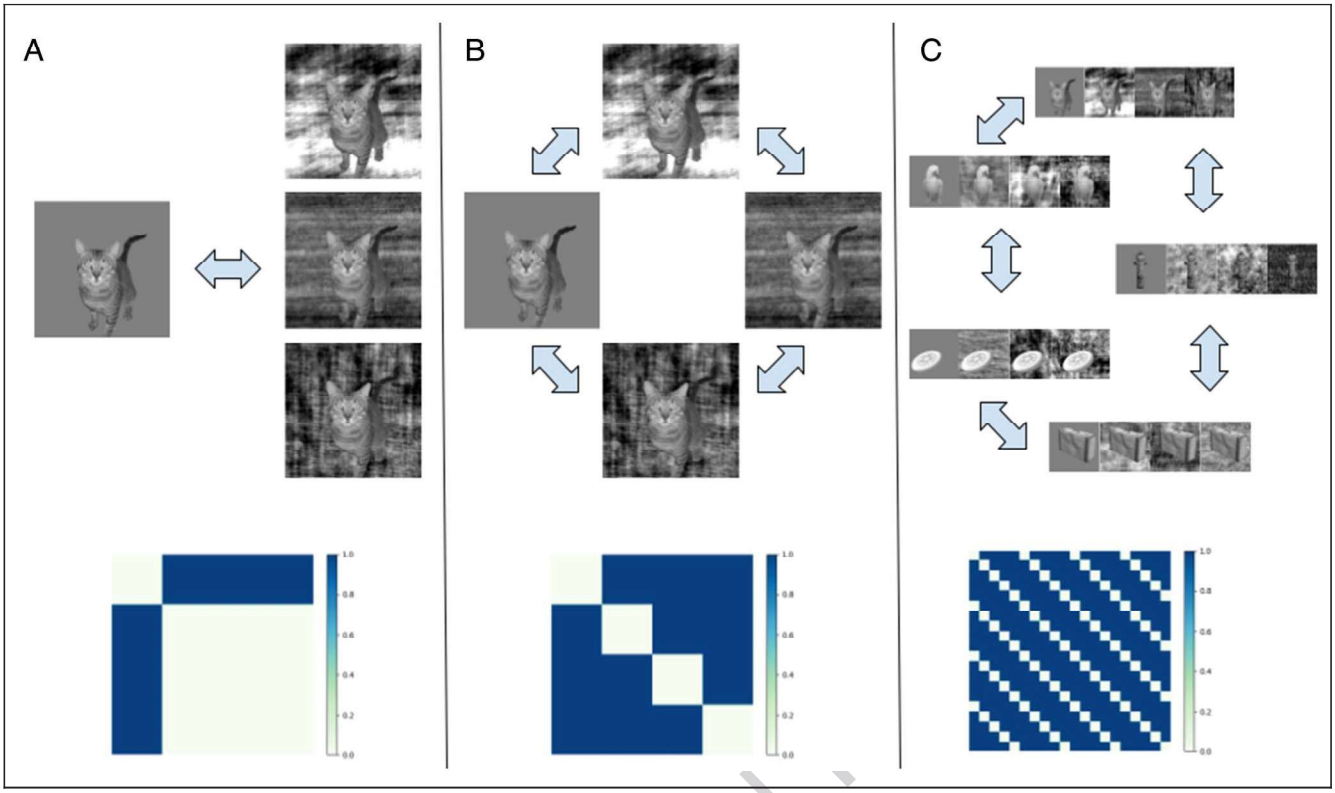

Figure 2. Categorical models of main experimental manipulations. (A) The categorical RDM of segmentation distinguishes between trials with and without backgrounds. (B) The categorical RDM of background complexity distinguishes between trials with different background complexities. (C) The categorical RDM of object category distinguishes between trials based on the target object category.

With the RDMs, first, we correlated the EEG RDMs per background condition with the categorical RDM of object category. This allowed us to assess the amount of object category information present in EEG as modulated by the image background complexity. Second, we performed partial correlations between the categorical RDMs and EEG RDMs, and between the categorical RDMs and DCNN RDMs to identify the shared representational structure. We chose to use a partial correlation instead of a regression to control for the correlation between the segmentation and background complexity categorical model. Third, to assess which DCNN layer's representation was most similar to EEG, we performed a Spearman correlation (i.e., classical RSA) between EEG RDMs (for every time sample) and DCNN RDMs (per layer). Fourth, we normalized each layer's explained variance from the Spearman correlation against the upper noise ceiling (the upper bound of EEG data) for all time samples and then plotted its median correlation against the layer's correlation with the categorical RDMs. This allowed us to summarize each layer's correlation with EEG data across all time samples. Finally, we qualitatively inspected the representations from DCNNs using t-distributed stochastic neighbor embedding (tSNE; van der Maaten & Hinton, 2008).

All statistical analysis was performed and visualized in Python using the following packages: NumPy, SciPy, Statsmodels, Pandas, Seaborn, Matplotlib (Waskom, 2021; Harris et al., 2020; Virtanen et al., 2020; McKinney, 2010; Seabold & Perktold, 2010; Hunter, 2007).

Analysis: Statistical

We used a Friedman test to determine if human participants and networks' categorization performance differed per background complexity (Figure 3). Significant Friedman test results were followed up with a post hoc Wilcoxon signed-ranks test to determine which condition pairs were significantly different from each other.

We used the Wilcoxon signed-ranks test to determine the onset of correlation significance between EEG RDMs per background complexity condition and the categorical RDM of object category (Figure 4). The p values obtained from the Wilcoxon signed-ranks test are Bonferroni corrected for multiple comparisons ($\alpha = .01$).

We used the Wilcoxon signed-ranks test to determine the onset of correlation significance between categorical RDMs and EEG RDMs, and to determine statistical significant differences in the correlation values of categorical RDMs (Figure 5). The p values obtained from the Wilcoxon signed-ranks test are Bonferroni corrected for multiple comparisons ($\alpha = .01$).

We used the Wilcoxon signed-ranks test to determine if the correlation values between the networks' layer

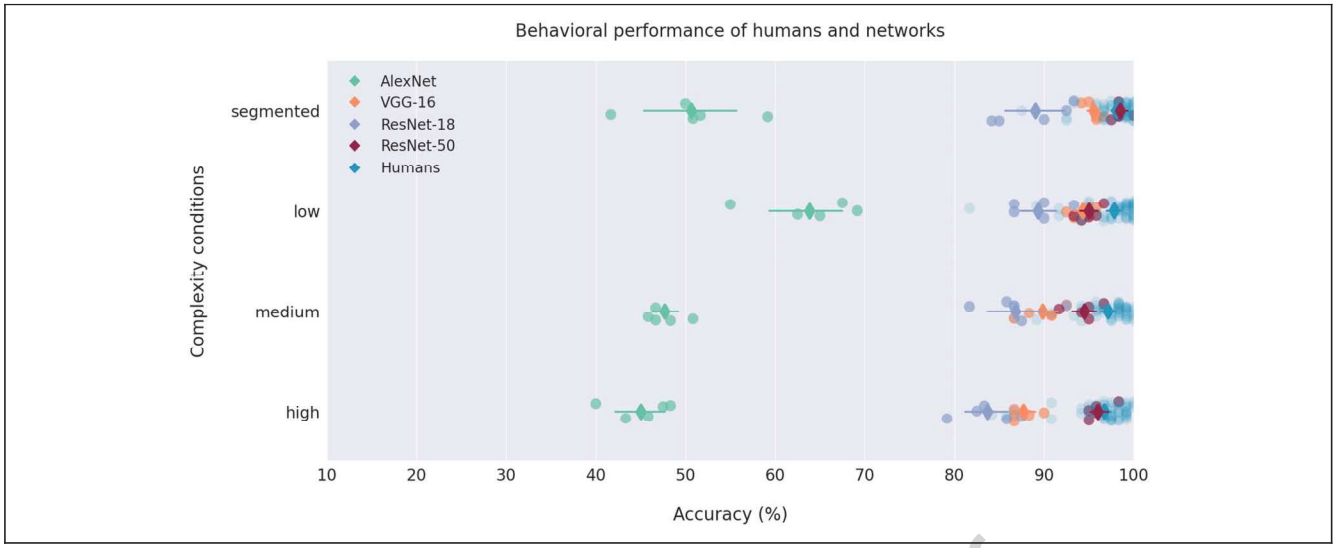

Figure 3. Behavioral performance of humans and networks. Human participants' categorization performance was near optimal across background complexity conditions. The Friedman test revealed that human participants' performance was influenced by the image background complexity. Post hoc Wilcoxon signed-ranks tests revealed significant differences between the low and medium, segmented and medium, segmented and high, and low and high complexity conditions. Among all networks, ResNet-50 most closely resembles human participants' performance. The data from human participants and ResNet-18 have previously been published in Seijdel and colleagues (2021).

activation and EEG RDM significantly differed from the correlation values of the categorical RDM of segmentation with EEG RDM (Figure 6). The p values obtained from the Wilcoxon signed-ranks test are Bonferroni corrected for multiple comparisons ($\alpha = .01$).

RESULTS

In this study, we investigated the factors contributing to the high predictive performance of DCNNs in human visual processing dynamics. First, we demonstrated that

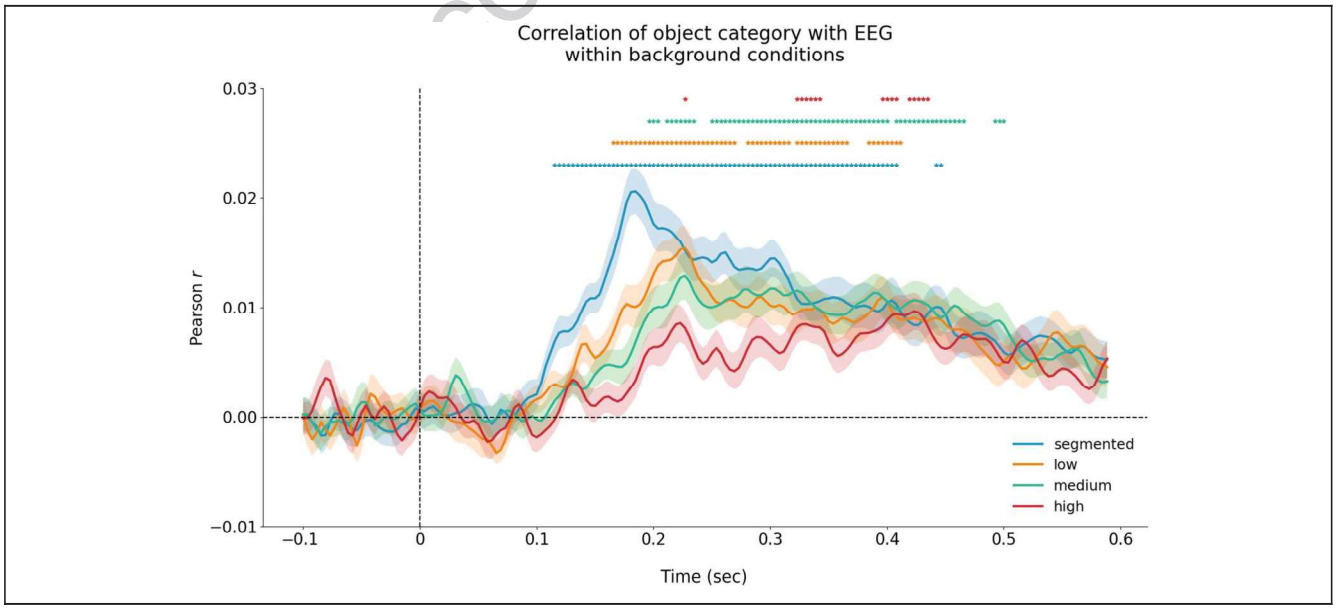

Figure 4. Neural signal for object concept. Objects without a background (segmented condition) have the earliest onset (117.78 msec) and highest degree of object concept ($r = .021$). This is followed by objects embedded in low complexity backgrounds (168.33 msec; $r = .015$), then objects in medium complexity backgrounds (199.44 msec; $r = .013$), and finally, objects in high complexity backgrounds (230.56 msec; $r = .010$). The colored asterisks indicate when values are significantly above zero.

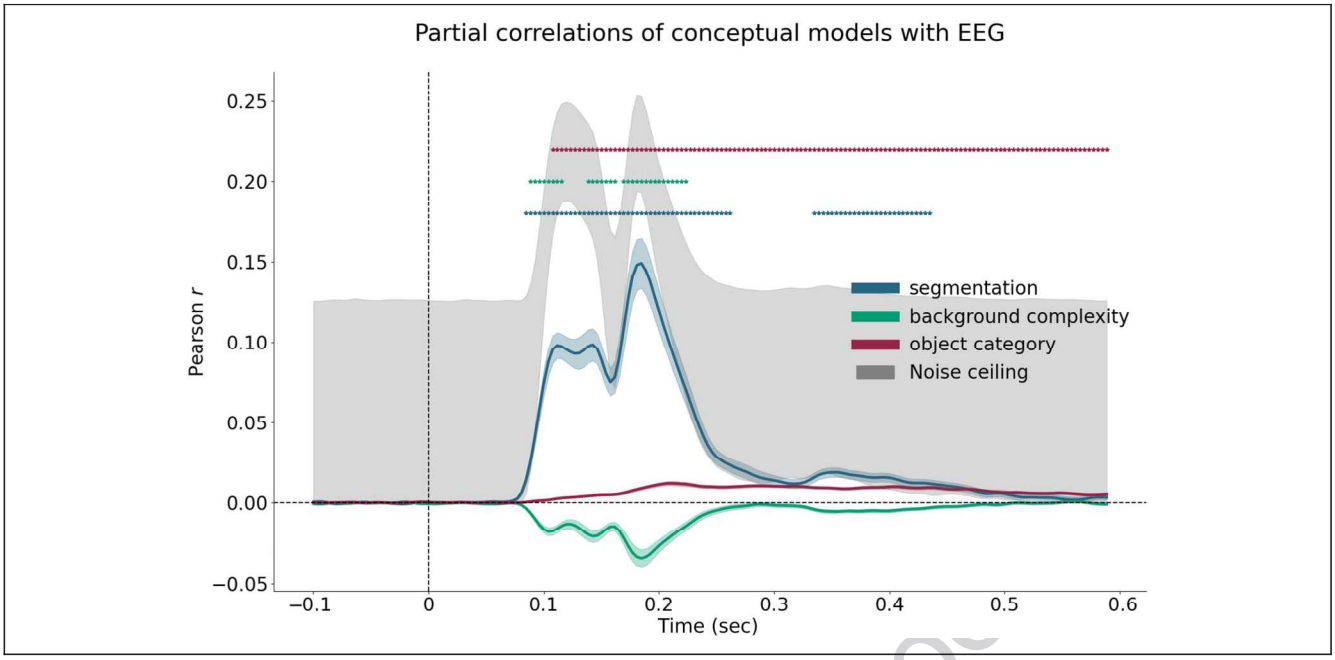

Figure 5. Partial squared correlation of conceptual models with EEG RDMs. By correlating our categorical RDMs with EEG RDMs, we find that the correlation with segmentation was the largest and earliest at 86.67 msec. This was followed by the correlation with background complexity with an onset at 90.56 msec. Finally, the correlation with object category was much smaller and later at 110 msec, compared with both factors related to object backgrounds. The colored asterisks indicate when values are significantly above zero. Wilcoxon signed-ranks tests also indicated that both background factors—segmentation and background complexity—had larger correlation values (in magnitude) as compared with object category. Altogether, this shows that EEG RDMs are largely modulated by object backgrounds, and the processing of object backgrounds precedes object category.

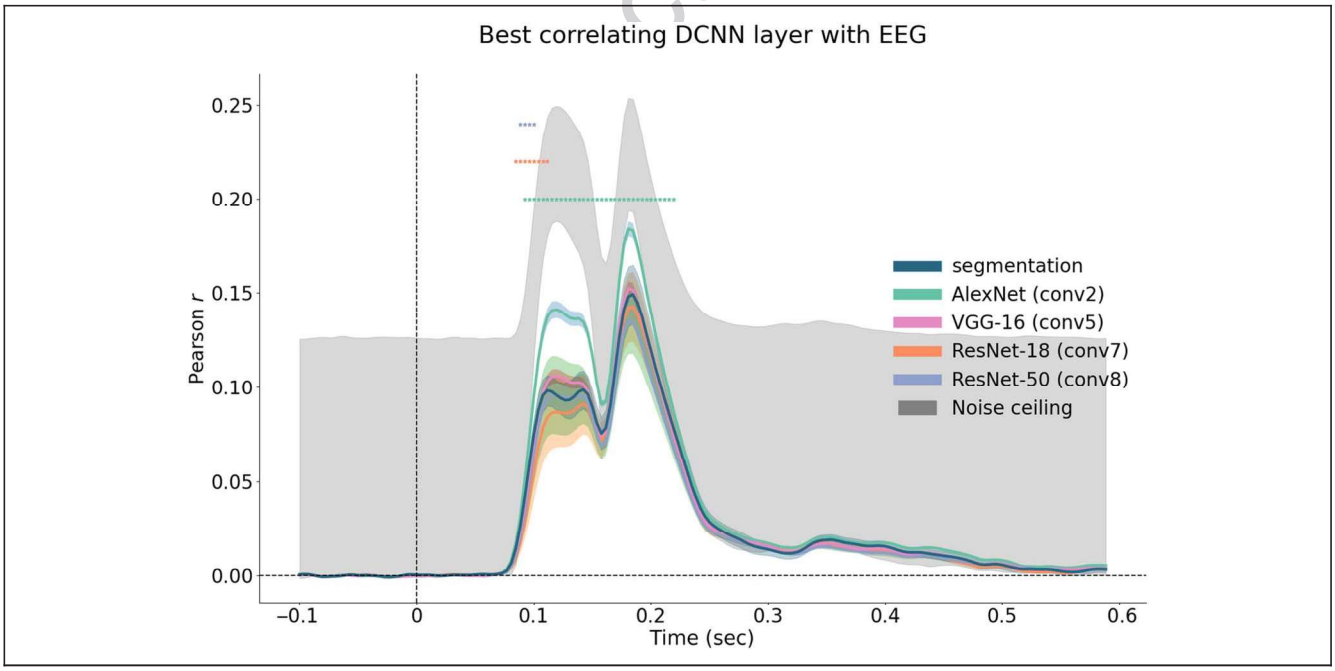

Figure 6. Best correlating DCNNs layers with EEG. We correlated DCNN RDMs (per layer) with EEG RDMs and observed that only AlexNet's second convolutional layer was close to the noise ceiling of the EEG data. AlexNet was also the only network which surpassed the explained variance of the segmentation RDM (asterisks indicate significant differences from segmentation RDM). AlexNet's second convolutional layer had significantly higher correlation values compared with the segmentation RDM between 94.44 msec and 222.78 msec. ResNet-18's seventh convolutional layer had significantly higher correlation values compared with the segmentation RDM between 86.67 msec and 113.89 msec. ResNet-50's eighth convolutional layer had significantly higher correlation values compared with the segmentation RDM between 90.56 msec and 102.22 msec. VGG-16's fifth convolutional layer was statistically equivalent with the segmentation RDM at all time points. The colored asterisks indicate when values significantly differed from the categorical RDM of segmentation.

both human participants and DCNNs could perform the object categorization task above chance. Second, we demonstrated that human participants' EEG activity differed per object background complexity. Third, using RSA (see Materials and Methods section), we examined the representations of EEG recordings using three categorical RDMs (see Materials and Methods section)—segmentation, background complexity, and object category (see Figure 2). We computed partial correlations between the categorical RDMs and EEG RDMs, and between the categorical RDMs and DCNN RDMs. Results from both sets of partial correlations revealed that EEG recordings and DCNN activations alike shared a high proportion of activity that distinguished between objects with backgrounds and those without. To investigate which processing stage (i.e., which layer) was most similar between human participants and DCNNs, we performed Spearman correlations between EEG RDMs (at every time sample) with DCNN RDMs (per layer). Finally, we showed that DCNN layers that correlate highly with EEG recordings are also layers that correlate highly with the categorical RDM of segmentation.

Deeper Networks Better Resemble Human Performance Compared with Shallower Networks

Human participants exhibited near-optimal categorization performance, despite the images being displayed for only 32 msec (see Figure 3). Results from the Friedman test showed that human participants' performance differed per background complexity condition, $\chi^2(3) = 31.45$, $p < .001$. Post hoc Wilcoxon signed-ranks tests revealed: no significant differences between the segmented and low condition, $W = 480.5$, p(uncorrected) = 0.51; significant differences between the low and medium condition, $W = 334.5$, p(uncorrected) = .003; no significant differences between the medium and high condition, $W = 346.5$, p(uncorrected) = .13; significant differences between the segmented and medium condition, $W = 176.5$, $p < .001$; significant differences between the segmented and high condition, $W = 129.0$, $p < .001$; significant differences between the low and high condition, $W = 248.0$, $p < .001$.

Among all networks, the deepest network (ResNet-50) best resembled human performance. Compared with all other networks, its mean categorization performance was the closest to human performance. Results from the Friedman test showed that ResNet-50's categorization performance differed per background complexity condition, $\chi^2(3) = 14.04$, $p = .002$. Post hoc Wilcoxon signed-ranks tests revealed: no significant differences between the segmented and low condition, $W = 0$, p(uncorrected) = .06; no significant differences between the low and medium condition, $W = 1$, p(uncorrected) = .13; no significant differences between the medium and high condition, $W = 0$, p(uncorrected) = .06; no significant differences between the segmented and medium condition, $W = 0$, p(uncorrected) = .06; no significant differences between the segmented and high condition, $W = 0$, p(uncorrected) = .06; no significant differences between the low and high condition, $W = 0$, p(uncorrected) = .06.

For all other networks (AlexNet, VGG-16, and ResNet-18), their categorization performance is similarly influenced by background complexity; (AlexNet) $\chi^2(3) = 15.0$, $p = .002$; (VGG-16) $\chi^2(3) = 13.56$, $p = .003$; (ResNet-18) $\chi^2(3) = 14.02$, $p = .003$. Post hoc Wilcoxon signed-ranks tests revealed similar results for all three networks (AlexNet, VGG-16, and ResNet-18): no significant differences between all condition pairs. However, these post hoc test results on DCNNs need to be interpreted with caution as they are based on a small sample size ($n = 5$).

Object Background Modulates Availability of Object Concept in EEG

Although human participants' categorization performance was near optimal across all background complexity conditions, the neural signals varied depending on the complexity of the object's background (see Figure 4). Here, we correlated EEG RDMs per background complexity condition with the categorical RDM of object category. This correlation informs us the degree to which object concept is present in the neural signal within each background complexity condition. Although the correlation values between object category RDM and EEG are modest, they were consistent with previous findings for a similar stimulus presentation rate (Grootswagers, Robinson, & Carlson, 2019). Overall, we observed that the degree of object concept decreases with increasing background complexity. In line with Grootswager and colleagues (2019), who noted that stimulus presentation rate influences EEG representation of object categories, our results add that background complexity also plays a significant role in modulating these neural signals.

Objects in the segmented condition has the earliest significant onset of object concept at 117.78 msec, $W = 1406$, p(Bonferonni corrected) < .001, and also the highest degree of object concept ($r = .021$), followed by the low complexity condition with significant onset of object concept at 168.33 msec, $W = 1415$, p(Bonferonni corrected) < .001, and object concept of $r = .015$, then medium complexity condition with significant onset of object concept at 199.44 msec, $W = 1397$, p(Bonferonni corrected) < .001, and object concept of $r = .013$, and finally high complexity condition with significant onset of object concept at 230.56 msec, $W = 1407$, p(Bonferonni corrected) < .001, and object concept of $r = .010$.

Object Background Modulates Early Neural Activity in Humans

To investigate which of our experimental factors best explained human participants' EEG recordings, we performed partial correlations between the categorical RDMs with EEG RDMs (see Figure 5). The EEG RDMs correlated highly with segmentation; this correlation had an onset of 86.67 msec, $W = 79$, p(Bonferonni corrected) $< .001$. This was followed by a correlation between the EEG RDMs with background complexity (onset of 90.56 msec), $W = 197$, p(Bonferonni corrected) $< .001$. Finally, there was a much smaller correlation between the EEG RDMs with object category (onset of 110 msec), $W = 222$, p(Bonferonni corrected) $< .01$. The order of onset significance started with segmentation and background complexity, both factors relating to object background, and subsequently arrived at object category.

We performed Wilcoxon signed-ranks tests between the correlation values of segmentation and background complexity, and between the correlation values of segmentation and object category. The correlation between the EEG RDMs with segmentation is significantly higher than the correlation between the EEG RDMs with background complexity at ~87–258 msec and ~339–441 msec, p(Bonferonni corrected) $< .01$, and also significantly higher than the correlation between EEG RDMs with object category at ~87–242 msec. We also performed Wilcoxon signed-ranks tests between the correlation values of background complexity and object category. The correlation between the EEG RDMs with background complexity is significantly larger (irrespective of signs) than the correlation between the EEG RDMs with object category at ~91–596 msec, p(Bonferonni corrected) $< .01$. Thus, both factors related to object backgrounds have earlier onsets and higher correlations as compared with object category. We can infer two things from these results—1. object background modulates majority of visual processing signals, not object category, and 2. visual processing of object backgrounds takes place before visual processing of object category.

Object Background Predicts Brain Activity As Well As DCNNs

To investigate which DCNNs processing stage (i.e., which layer) was most similar to human participants, we performed Spearman correlations between DCNN RDMs (per layer) with EEG RDMs (at every time sample). We plotted the best layer (i.e., the layer from each DCNN with the highest correlation with EEG), next to the best categorical RDM—segmentation.

We found that AlexNet was the only DCNN that surpassed the explained variance of the categorical RDM of segmentation (see Figure 6). The correlation values of AlexNet's second convolutional layer was significantly higher than the correlation values of the categorical RDM of segmentation, within the time window of 94.44 msec until 222.78 msec. On the other hand, the categorical RDM of segmentation captured an equivalent amount of variance in the human participant EEG recordings as layer activations of VGG-16, ResNet-18, and ResNet-50. The correlation values of the categorical RDM of segmentation were higher than ResNet-18's seventh convolutional layer within the time window of 86.67 msec until 113.89 msec, and also higher than ResNet-50's eighth convolutional layer within the time window of 90.56 msec and 102.22 msec. The correlation values of the categorical RDM of segmentation were statistically equivalent with VGG-16's fifth convolutional layer for all EEG time points. Out of the four DCNNs, AlexNet was the closest to the noise ceiling of the EEG RDMs whereas the other networks fell far from the noise ceiling.

Object Background Modulates Early Layers' Activations in DCNNs

To investigate which of our experimental factors best explained DCNNs' activity, we similarly performed the partial correlation between the categorical RDMs with DCNNs' activations (per layer; see Figure 7). First, we observed that early layers of the DCNNs have high correlation values with segmentation and background complexity—indicating that a large proportion of DCNNs' early activity is related to object background, not object category, similar to activity in human brains as shown in Figure 5. Second, we observed that correlations with object category arose in later layers. In deeper networks (with more layers), the correlations with object category became much higher toward the penultimate layer as compared with shallower networks. As a control, we performed the partial correlations between categorical RDMs and untrained DCNN RDMs. We observed that the correlation for segmentation (and not background complexity nor object category) similarly captured a large proportion of untrained DCNNs' activations. However, unlike their trained counterparts, untrained DCNNs' correlations arose more gradually and remained until the penultimate (fully connected) layer. In addition, the correlation for background complexity and object category remained close to null throughout the untrained DCNN layers. This indicates that the background differences in untrained DCNNs were not resolved or made invariant, unlike their trained counterparts. Presumably, this transformation of making backgrounds invariant allowed the networks to learn object categorically relevant features.

To further understand the network activations, we visualized its activity with tSNE (van der Maaten & Hinton, 2008)). tSNE maps high-dimensional data points to 2-D or 3-D spaces. We selected to visualize the activations of DCNNs' first and final layers, and also the layer with the highest correlation with human participants EEG recordings. The tSNE visualization showed that DCNNs' layers that correlate most with EEG RDMs

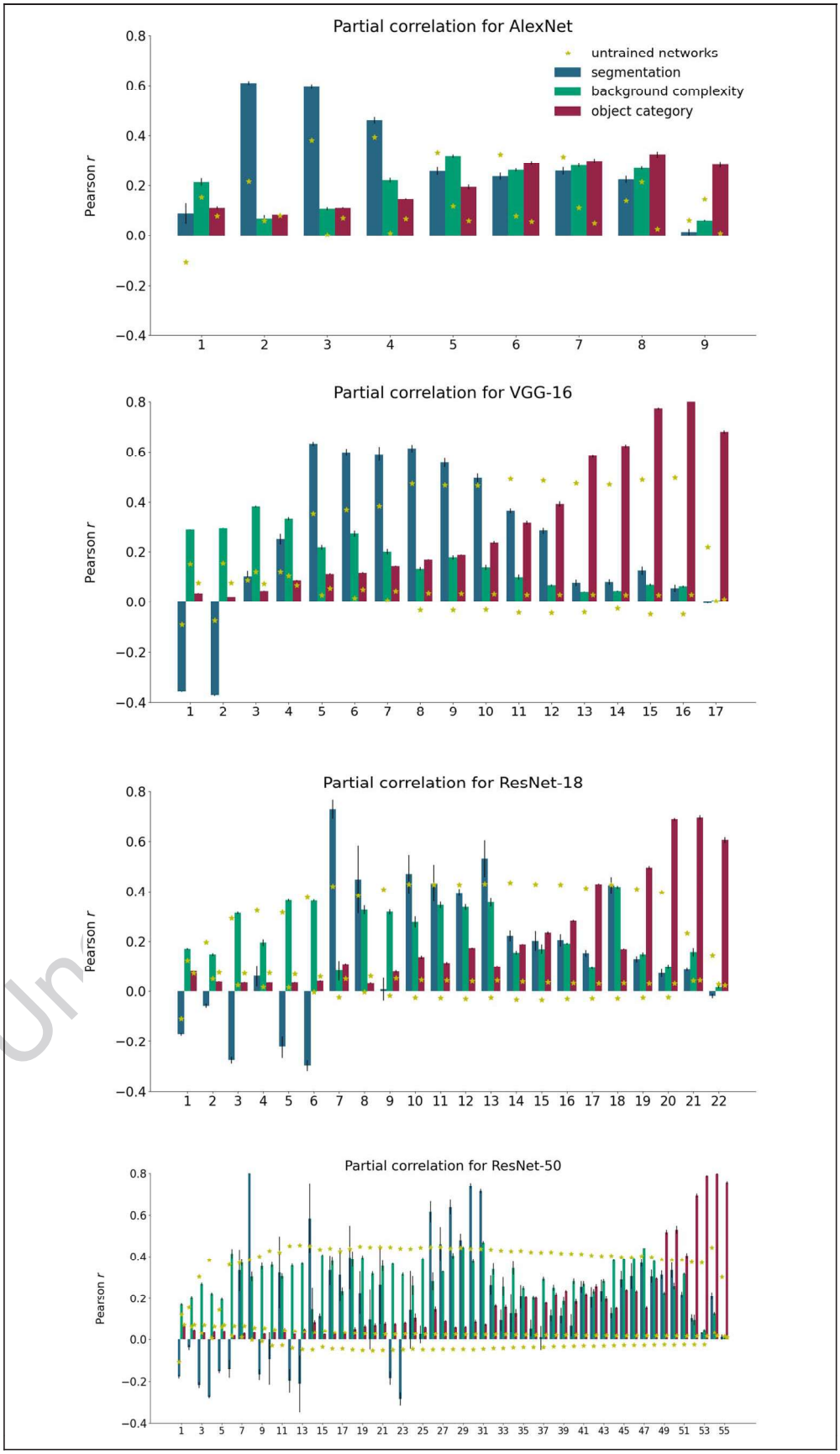

Figure 7. Partial correlation of categorical RDMs with DCNNs. The partial correlations between categorical RDMs (segmentation, background complexity, and object category) and DCNN RDMs are shown for each layer of the network. Partial correlations for untrained DCNN RDMs are marked by the yellow stars. Values on the *x* axis indicate layer number; values on the *y* axis indicate the layer's partial correlation with the categorical RDMs. We observed that the early layers of DCNNs correlate largely with both segmentation and background complexity but not with object category. The correlation with object category gradually increases in the later layers, with deeper networks showing a larger increase compared with shallower networks. This pattern of correlation is robust across all networks.

showed activation that is differentiated along object background—not object category (see Figure 8). In the first layer of all networks, we see a random initialization with no clear clustering of stimuli. In the layer that correlates most with brain activity, we see a clustering of activity according to object backgrounds. And in the final layer, we see a clustering of activity according to object category. With the tSNE visualization, we showed that DCNNs' activity differentiates first according to object background and then according to object category. One

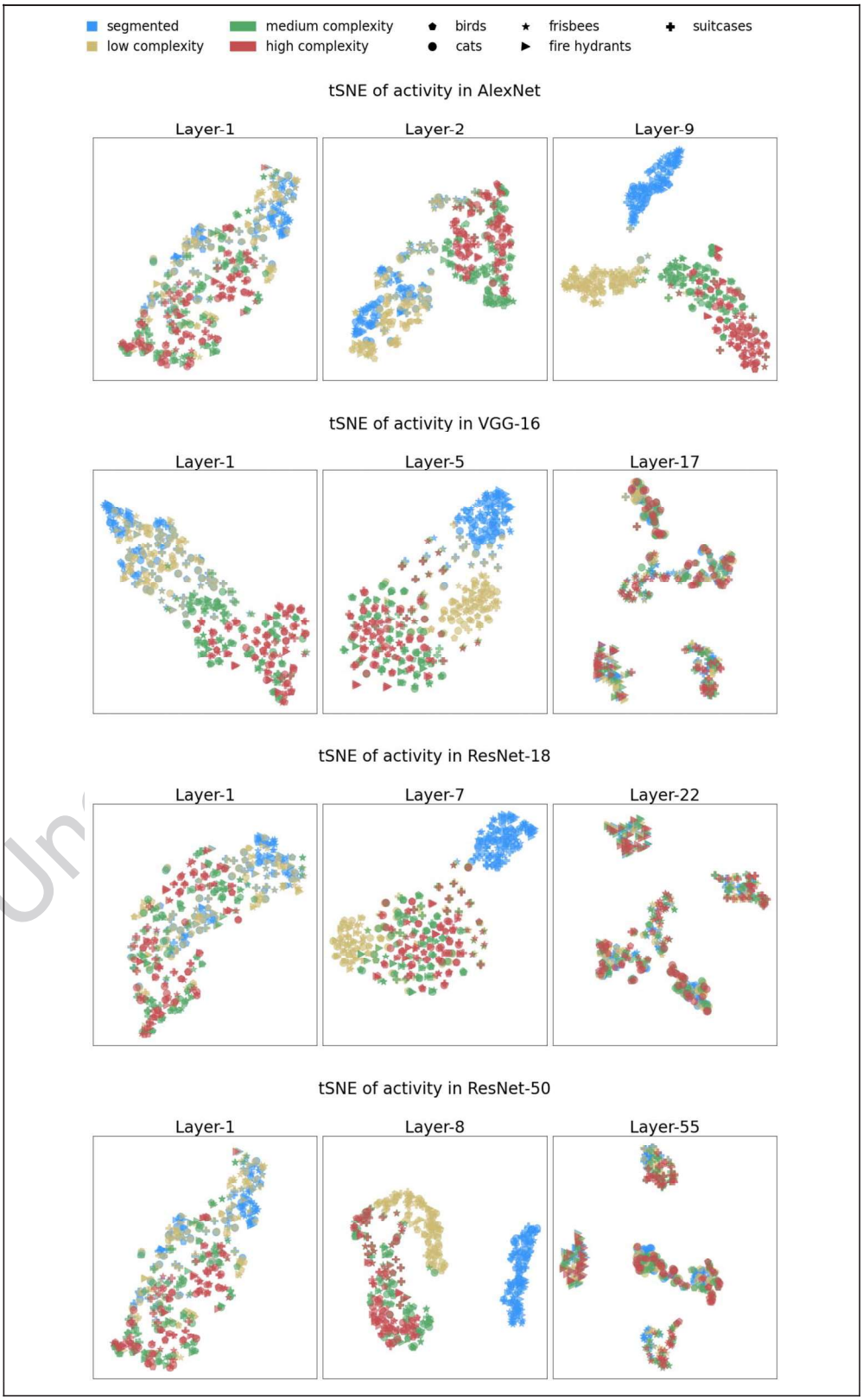

Figure 8. tSNE of DCNN's activations. We applied tSNE to DCNNs' activations in the first (leftmost) and last (rightmost) layers, and also the layer that correlated most with brain activity (middle). Colors indicate object background conditions—segmented (blue), low complexity (yellow), medium complexity (green), high complexity (red). Markers indicate object category—bird (crosses), cat (circles), frisbee (stars), fire hydrant (triangles), suitcase (plusses). We observed that DCNNs' layer that best captured human participants' EEG recordings has activity differentiated along object background—not object category. In the first layer of all networks (leftmost), we see a random initialization with no clear clustering of stimuli. In the layer that correlates most with brain activity (middle), we see a clustering of activity according to object background (in colors). In the final layer (rightmost), we see a clustering of activity according to object category (in marker shapes). Here, we show that DCNN's activity differentiates first according to object background and then according to object category.

notable exception of this pattern of results is AlexNet; in its output layer (Layer 9), its activity is still clustered along object background. One possible explanation is that AlexNet, being a shallower network compared with the other three, lacks the depth and additional processing needed to differentiate stimuli based on their categories.

We showed that DCNNs' layers that capture differences related to object background are also layers that best capture human participants' EEG recordings. As these layers with activations differentiating object background correlate with brain activity, we posit that the predictive power of DCNNs on brain activity is largely derived from its ability to differentiate object backgrounds, or, more specifically, image textures (Geirhos et al., 2018).

DCNNs Layers That Correlate Highly with EEG RDMs Also Correlate Highly with Segmentation

After observing that both EEG RDMs and DCNNs' RDMs correlate highly with the categorical RDM of segmentation (see Figures 5 and 7), we wanted to investigate the relationship between the three groups of RDMs—EEG RDMs, DCNNs' RDMs, and the categorical RDMs. Specifically, we examined if the correlation values of EEG with a categorical RDM (e.g., segmentation), and the correlation values of DCNNs with the same categorical RDM, correlated with each other. By doing so, we directly investigate if DCNNs' layers, which correlate with a categorical RDM, also correlate well with EEG. This correlation analysis gives us a bridge between EEG and DCNNs to observe if their correlation with a categorical RDM helps explain DCNNs' predictive power on EEG dynamics. Thus, we took the correlation values of DCNNs with the three categorical RDMs (one datapoint per layer, averaged across five initializations) and plotted each DCNN layer's median correlation with EEG across all time points. We observed that DCNNs' RDMs, which correlate highly with EEG RDM also correlate highly with the categorical RDM of segmentation (AlexNet, r =.98, $p < .001$; VGG-16, $r = .50$, $p = .04$; ResNet-18, $r = .53$, $p = .01$; ResNet-50, $r = .53$, $p < .001$). This indicates that DCNNs' correlation with brain activity is derived from its ability to distinguish between objects' backgrounds. DCNNs' RDMs, which correlate moderately with background complexity, have a weaker correlation with EEG RDM (AlexNet, $r = -.42$, $p = .26$; VGG-16, $r = -.33$, $p = .20$; ResNet-18, $r = -.16$, $p = .47$; ResNet-50, $r = .16$, $p = .26$). DCNNs' RDMs, which correlate moderately with the categorical RDM of object category, also have a weaker correlation with EEG RDMs (AlexNet, $r = -.58$, $p = .10$; VGG-16, $r = -.11$, $p = .68$; ResNet-18, $r = .14$, $p = .52$; ResNet-50, $r = .17$, $p = .20$). Therefore, we can conclude that much of the predictive power of DCNNs on EEG dynamics stems from the shared representations of object backgrounds between DCNNs and the human visual cortex.

DISCUSSION

We set out to investigate the factors leading to DCNNs' high predictive performance on human visual processing dynamics by studying objects and their backgrounds. Using RSA (Kriegeskorte et al., 2008), we compared the activity of four DCNN architectures with EEG recordings of human participants. We focused on three factors: segmentation, background complexity, and object category. First, we showed that object background modulates the amount of object concept in EEG signals. Second, we showed that object background largely modulates early EEG signals and early DCNNs layers. Third, we showed that both representations from EEG and DCNNs reflected the distinction between objects with and without backgrounds. Fourth, we showed that the shared distinction of object backgrounds is associated with DCNNs' high predictive performance on human visual processing dynamics. We posit that DCNNs' ability to predict EEG signals is derived from its ability to distinguish between target object and object backgrounds.

Processing of Object Backgrounds in Humans Happens Earlier and Is More Substantial Than Processing of Object Features

We found high correlations between the categorical RDMs of segmentation and background complexity with EEG—revealing that visual processing (as recorded with EEG) is largely modulated by object backgrounds instead of object category (see Figure 5). Furthermore, the correlations between segmentation and background complexity with EEG have earlier onsets compared with object category—segmentation at 86.67 msec, background complexity at 90.56 msec, and object category at 110 msec. Our result suggests that the processing of object background precedes object features, and through this process, target objects and their backgrounds become distinct. This is evident not only in the latency of significant correlation between the conceptual models and EEG, but also in the correlation between the conceptual models and DCNNs layers—where correlations with segmentation and background complexity precede object category.

Our finding agrees with previous findings showing that object background complexity influences object categorical perception, with objects embedded in more complex backgrounds reaching categorical perception later (Seijdel et al., 2021; Groen et al., 2018). The longer latency for categorical perception could be explained by time taken to distinguish between the target object and its background. In addition, our result also extends initial findings that categorical perception is fast (within 150 msec; Hung, Kreiman, Poggio, & DiCarlo, 2005; Thorpe, Fize, & Marlot, 1996). Results from earlier studies demonstrating the quickness of categorical perception holds when the presented stimuli are simple (i.e., object with a plain background); however, if the presented stimuli are more

complex (i.e., object with a complex background), longer latency incorporating additional processing steps would be required (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019). As natural scenes comprise a myriad complexities in backgrounds, we recommend a careful consideration of not only object category but also backgrounds.

DCNNs Processes on Object Backgrounds Explain EEG Activity

In our experiment, we show that DCNNs' predictive power on EEG data is derived from DCNNs' inherent ability to distinguish between objects with and without backgrounds. Crucially, the distinction of object backgrounds is orthogonal to the object categorization task. The selected DCNNs for the experimental task have been pretrained on a naturalistic data set (ImageNet) and further optimized with a separate data set (MSCOCO). Nonetheless, DCNNs' activations reflect a distinction between objects with and without backgrounds. The distinction is apparent in its partial correlation with the categorical RDMs of segmentation and background complexity (see Figure 7), especially in DCNNs' early and mid-layers. In addition, we also showed that DCNNs' layers, which correlated with segmentation, also correlated with EEG (see Figure 9), suggesting that DCNNs' predictive power on EEG data is largely derived from the shared ability of both modalities to distinguish between the target object and its background.

Our conclusion that DCNNs' predictive power on EEG data is derived from the shared ability of both modalities to distinguish between objects' backgrounds needs to be considered carefully because we have reconstructed an experimental data set with target objects embedded within artificial backgrounds. There is a high necessity to identify the target object as separate from its background because the artificial backgrounds are uninformative on the object category. In contrast, if the object category correlated with its background (e.g., frisbee with the background of a park), and if the discrimination of object categories could be performed sufficiently well based on the object backgrounds, no distinction needs to be made between target objects and their backgrounds. In reality, most naturalistic scenes will have backgrounds that are informative of its target objects' categories as these are a matter of statistical correlations (Oliva & Torralba, 2007). In our study, we constructed an object categorization task that required the distinction of target object and its background with the intention of investigating the mechanism of figure-ground segmentation; surprisingly, we found that both DCNNs and our human participants shared this ability.

Constraints of EEG Signals in Capturing Object Categories

Although the correlation values between categorical RDMs and EEG in our study are modest, it is essential to interpret these findings within the broader research context. These low, yet consistent correlation values are aligned with previous studies conducted at a comparable stimulus presentation rate (Grootswagers et al., 2019). These results underscore the inherent limitations of capturing the neural processes of object categorization through EEG. The signal-to-noise ratio in EEG recordings is generally low, but further limited by presentation rate. On top of presentation rate, our data revealed that background complexity also significantly modulates these neural signals, adding another layer of complexity to our understanding of object categorization. We believe that these insights will be valuable for future research, aiding in the design of experiments and setting of appropriate methodological expectations.

Emergence of Shared Solutions for Object Categorization

The shared ability to distinguish between target objects and their backgrounds within human visual processing and DCNNs is intriguing. It prompts us to explore a fundamental question: Why does this shared ability exist in the first place? This ability was not directly implemented in both systems yet emerged as part of the solution for categorizing objects. Within vision neuroscience, this ability to distinguish between target objects and its backgrounds has long been studied as part of processes known as perceptual grouping or figure-ground segmentation (Kirchberger et al., 2021; Self et al., 2019; Scholte, Jolij, Fahrenfort, & Lamme, 2008; Roelfsema, 2006; Roelfsema et al., 2002; Lamme, Supèr, & Spekreijse, 1998). Specifically, these processes refer to the grouping of image elements that belong to different entities. It has been shown that if these processes were interrupted in human participants, object categorization becomes impaired (Fahrenfort, Scholte, & Lamme, 2007). In our study, the emergence of a shared solution (i.e., perceptual grouping) for object categorization suggests that it is a crucial solution for the task at hand and could elucidate the evolutionary constraints on the problem (Cao & Yamins, 2021b). This helps us arbitrate which biological processes are necessary to incorporate in artificial systems depending on their contexts.

Figure-ground Segregation Assists Object Features Learning

Previous research has shown the surprising prediction performance of random weights networks (Storrs, Kietzmann, Walther, Mehrer, & Kriegeskorte, 2021; Truzzi & Cusack, 2020; Cichy, Khosla, Pantazis, Torralba, & Oliva, 2016); it is indeed impressive that random weights networks are able to explain any brain activity at all. Our experimental results similarly showed that untrained networks can explain variance in brain activity through its inherent ability to process low-level image statistics.

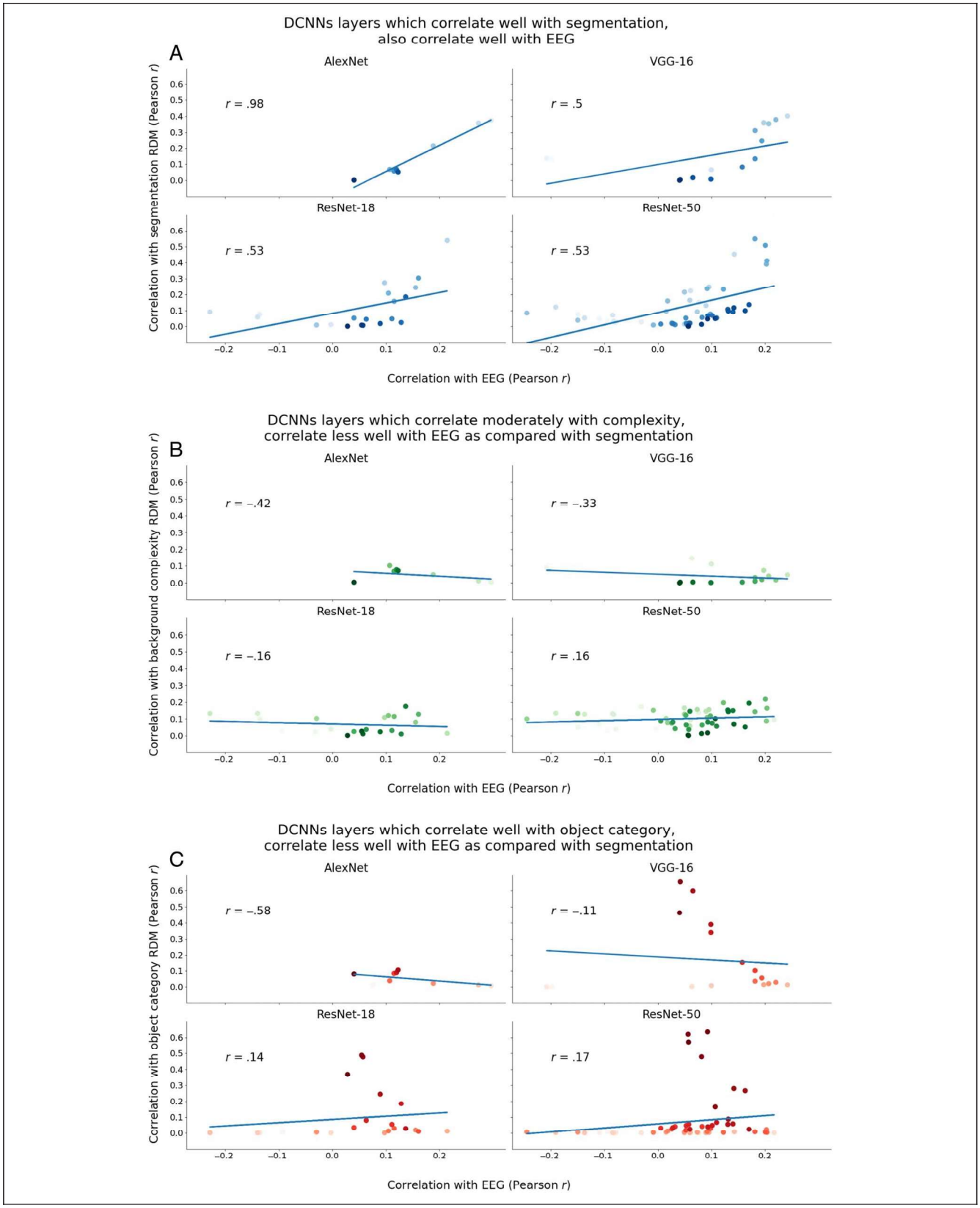

Figure 9. Relationship between DCNN's correlation with EEG and categorical RDMs. Each dot represents a DCNN layer (averaged across five initializations). The correlation value plotted is the median correlation with EEG across all time points. Darker colors indicate deeper layers within a network, and lighter colors indicate shallower layers. (A) We observed that layers that correlate highly with EEG are also layers that correlate with the categorical RDM of segmentation. (B) The relationship between DCNNs' correlation with EEG and the categorical RDM of background complexity is much weaker, and similarly, (C) the relationship between DCNNs' correlation with EEG and the categorical RDM of object category is also much weaker as compared with the categorical RDM of segmentation.

Through correlating untrained networks' RDMs with conceptual RDMs, we find that the networks' activity is modulated only by object background and not object category at all (see Figure 7). We observed a similar predictive performance of an untrained network on V1 in previous studies, where the correlation of the untrained network gradually increased in the early layers and remained until the late layers (Cichy et al., 2016). In our study, we observed that the conceptual RDMs of segmentation correlated moderately with the layers of untrained networks, whereas the conceptual RDMs of background complexity and object category did not correlate with the layers of untrained networks. This indicates that untrained networks can partially distinguish between objects with and without backgrounds. However, they are unable to distinguish between background types or categorical features. In contrast, layers of trained networks showed a correlation with segmentation up until the middle layers of the network, which then gradually decreased, matched by the gradual increase of correlation with object category. This suggests that trained networks "resolved" figure-ground segregation, allowing them to learn object categorical features.

Conclusion

In summary, we have tested the best mechanistic models of visual processing and showed that both early human visual processing and early layers of DCNNs are highly influenced by object backgrounds rather than object categories. Furthermore, this shared ability to distinguish between object backgrounds accounts for the predictive power of DCNNs on EEG activity. Interestingly, neither humans nor DCNNs were explicitly trained to make these distinctions, yet this shared solution emerged to address the experimental task of object categorization. Overall, our findings indicate that both human visual processing and DCNNs care deeply about object backgrounds.

Corresponding author: Jessica Loke, Psychology Department - Brain & Cognition, University of Amsterdam, Nieuwe Achtergracht 129b, 1018 XE Amsterdam, The Netherlands, or via e-mail: jessloke08@gmail.com.

Data Availability Statement

Data and code to reproduce the analyses in this article will be made available at https://osf.io/es34u/.

Author Contributions

Jessica Loke: Conceptualization; Formal analysis; Investigation; Methodology; Validation; Visualization; Writing—Original draft; Writing—Review & editing. Noor Seijdel: Conceptualization; Formal analysis; Investigation; Methodology; Visualization; Writing—Original draft; Writing—Review & editing. Lukas Snoek: Formal analysis; Methodology; Visualization; Writing—Review & editing. Lynn K. A. Sörensen: Formal analysis; Methodology; Visualization. Ron van de Klundert: Data curation; Project administration; Writing—Review & editing. Matthew van der Meer: Data curation; Project administration; Writing—Review & editing. Eva Quispel: Data curation; Project administration; Writing—Review & editing. Natalie Cappaert: Funding acquisition; Supervision; Writing—Review & editing. H. Steven Scholte: Conceptualization; Funding acquisition; Project administration; Supervision; Writing—Review & editing.

Funding Information

This work is supported by an Interdisciplinary Doctorate Agreement from the University of Amsterdam to H. Steven Scholte and Natalie Cappaert and an Advanced Investigator Grant from the European Research Council (ERC; https://dx.doi.org/10.13039/501100000781) to Edward de Haan, grant number: 339374.

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience (JoCN)* during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance. The authors of this paper report its proportions of citations by gender category to be: M/M = .75; W/M = .175; M/W = .075; W/W = 0.

REFERENCES

Akhtar, N., & Mian, A. (2018). Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, *6*, 14410–14430. https://doi.org/10.1109/ACCESS.2018.2807385

Baker, N., Lu, H., Erlikhman, G., & Kellman, P. J. (2018). Deep convolutional networks do not classify based on global object shape. *PLoS Computational Biology*, *14*, e1006613. https://doi.org/10.1371/journal.pcbi.1006613, PubMed: 30532273

Borji, A. (2021). Contemplating real-world object classification. *ArXiv*. https://doi.org/10.48550/arXiv.2103.05137

Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., et al. (2014). Deep neural networks rival the representation of primate IT cortex for core visual object recognition. *PLoS Computational Biology*, *10*, e1003963.

https://doi.org/10.1371/journal.pcbi.1003963, PubMed: 25521294
Cao, R., & Yamins, D. (2021a). Explanatory models in neuroscience: Part 1—Taking mechanistic abstraction seriously. *ArXiv*. https://doi.org/10.48550/arXiv.2104.01490
Cao, R., & Yamins, D. (2021b). Explanatory models in neuroscience: Part 2—Constraint-based intelligibility. *ArXiv*. https://doi.org/10.48550/arXiv.2104.01489
Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*, 305–317. https://doi.org/10.1016/j.tics.2019.01.009, PubMed: 30795896
Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 27755. https://doi.org/10.1038/srep27755, PubMed: 27282108
Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., et al. (2022). Semantic scene descriptions as an objective of human vision. *ArXiv*. https://doi.org/10.48550/arXiv.2209.11737
Dwivedi, K., Bonner, M. F., Cichy, R. M., & Roig, G. (2021). Unveiling functions of the visual cortex using task-specific deep neural networks. *PLoS Computational Biology*, *17*, e1009267. https://doi.org/10.1371/journal.pcbi.1009267, PubMed: 34388161
Eickenberg, M., Gramfort, A., Varoquaux, G., & Thirion, B. (2017). Seeing it all: Convolutional network layers map the function of the human visual system. *Neuroimage*, *152*, 184–194. https://doi.org/10.1016/j.neuroimage.2016.10.001, PubMed: 27777172
Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. F. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience*, *19*, 1488–1497. https://doi.org/10.1162/jocn.2007.19.9.1488, PubMed: 17714010
Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: Object recognition when the signal gets weaker. *ArXiv*. https://doi.org/10.48550/arXiv.1706.06969
Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: Quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in Neural Information Processing Systems*. https://proceedings.neurips.cc/paper/2020/hash/9f6992966d4c363ea0162a056cb45fe5-Abstract.html
Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., & Brendel, W. (2018). ImageNet-trained CNNs are biased towards texture; Increasing shape bias improves accuracy and robustness. *ArXiv*. https://doi.org/10.48550/arXiv.1811.12231
Goodfellow, I. J., Shlens, J., & Szegedy, C. (2014). Explaining and harnessing adversarial examples. *ArXiv*. https://doi.org/10.48550/arXiv.1412.6572
Groen, I. I. A., Jahfari, S., Seijdel, N., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2018). Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLoS Computational Biology*, *14*, e1006690. https://doi.org/10.1371/journal.pcbi.1006690, PubMed: 30596644
Grootswagers, T., Robinson, A. K., & Carlson, T. A. (2019). The representational dynamics of visual objects in rapid serial visual processing streams. *Neuroimage*, *188*, 668–679. https://doi.org/10.1016/j.neuroimage.2018.12.046, PubMed: 30593903
Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., et al. (2020). Array programming with NumPy. *Nature*, *585*, 357–362. https://doi.org/10.1038/s41586-020-2649-2, PubMed: 32939066
He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. *2016 IEEE conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 770–778). Las Vegas, NV, USA: IEEE. https://doi.org/10.1109/CVPR.2016.90
Hung, C. P., Kreiman, G., Poggio, T., & DiCarlo, J. J. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*, 863–866. https://doi.org/10.1126/science.1117593, PubMed: 16272124
Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*, 90–95. https://doi.org/10.1109/MCSE.2007.55
Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*, 974–983. https://doi.org/10.1038/s41593-019-0392-5, PubMed: 31036945
Kay, K. N. (2018). Principles for models of neural information processing. *NeuroImage*, *180*, 101–109. https://doi.org/10.1016/j.neuroimage.2017.08.016, PubMed: 28793238
Khaligh-Razavi, S.-M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS Computational Biology*, *10*, e1003915. https://doi.org/10.1371/journal.pcbi.1003915, PubMed: 25375136
Kirchberger, L., Mukherjee, S., Schnabel, U. H., van Beest, E. H., Barsegyan, A., Levelt, C. N., et al. (2021). The essential role of recurrent processing for figure-ground perception in mice. *Science Advances*, *7*, eabe1833. https://doi.org/10.1126/sciadv.abe1833, PubMed: 34193411
Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008, PubMed: 19104670
Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet classification with deep convolutional neural networks. In F. Pereira, C. J. Burges, L. Bottou, & K. Q. Weinberger (Eds.), *Advances in neural information processing systems* (Vol. 25, pp. 1–9). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2012/file/c399862d3b9d6b76c8436e924a68c45b-Paper.pdf.
Lamme, V. A., Supèr, H., & Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, *8*, 529–535. https://doi.org/10.1016/S0959-4388(98)80042-1, PubMed: 9751656
Lillicrap, T. P., & Kording, K. P. (2019). What does it mean to understand a neural network? *ArXiv*. https://doi.org/10.48550/arXiv.1907.06374
Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., et al. (2014). Microsoft COCO: Common Objects in Context. *Computer Vision – ECCV*, *8693*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48
Loke, J., Seijdel, N., Snoek, L., van der Meer, M., van de Klundert, R., Quispel, E., et al. (2022). A critical test of deep convolutional neural networks' ability to capture recurrent processing in the brain using visual masking. *Journal of Cognitive Neuroscience*, *34*, 2390–2405. https://doi.org/10.1162/jocn_a_01914, PubMed: 36122352
Marcus, G. (2018). Deep learning: A critical appraisal. *ArXiv*. https://doi.org/10.48550/arXiv.1801.00631
McKinney, W. (2010). Data structures for statistical computing in Python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56. https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf
Mehrer, J., Spoerer, C. J., Kriegeskorte, N., & Kietzmann, T. C. (2020). Individual differences among deep neural network models. *Nature Communications*, *11*, 5725. https://doi.org/10.1038/s41467-020-19632-w, PubMed: 33184286
Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*, 520–527. https://doi.org/10.1016/j.tics.2007.09.009, PubMed: 18024143

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., et al. (2019). PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

Rajalingham, R., Issa, E. B., Bashivan, P., Kar, K., Schmidt, K., & DiCarlo, J. J. (2018). Large-scale, high-resolution comparison of the core visual object recognition behavior of humans, monkeys, and state-of-the-art deep artificial neural networks. *Journal of Neuroscience*, *38*, 7255–7269. https://doi.org/10.1523/JNEUROSCI.0388-18.2018, PubMed: 30006365

Ritchie, J. B., Zeman, A. A., Bosmans, J., Sun, S., Verhaegen, K., & Op de Beeck, H. P. (2021). Untangling the animacy organization of occipitotemporal cortex. *Journal of Neuroscience*, *41*, 7103–7119. https://doi.org/10.1523/JNEUROSCI.2628-20.2021, PubMed: 34230104

Ritter, S., Barrett, D., Santoro, A., & Botvinick, M. (2017). Cognitive psychology for deep neural networks: A shape bias case study. *International Conference on Machine Learning*. https://www.semanticscholar.org/paper/39fb9fa2615620f043084a2ecbbdb1a1f8c707c9

Roelfsema, P. R. (2006). Cortical algorithms for perceptual grouping. *Annual Review of Neuroscience*, *29*, 203–227. https://doi.org/10.1146/annurev.neuro.29.051605.112939, PubMed: 16776584

Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (2002). Figure—ground segregation in a recurrent network architecture. *Journal of Cognitive Neuroscience*, *14*, 525–537. https://doi.org/10.1162/08989290260045756, PubMed: 12126495

Scholte, H. S. (2018). Fantastic DNimals and where to find them. *NeuroImage*, *180*, 112–113. https://doi.org/10.1016/j.neuroimage.2017.12.077, PubMed: 29288865

Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W. M., & Lamme, V. A. F. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. *Journal of Vision*, *9*, 29.1–29.15. https://doi.org/10.1167/9.4.29, PubMed: 19757938

Scholte, H. S., Jolij, J., Fahrenfort, J. J., & Lamme, V. A. F. (2008). Feedforward and recurrent processing in scene segmentation: Electroencephalography and functional magnetic resonance imaging. *Journal of Cognitive Neuroscience*, *20*, 2097–2109. https://doi.org/10.1162/jocn.2008.20142, PubMed: 18416684

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., et al. (2018). Brain-score: Which artificial neural network for object recognition is most brain-like? *BioRxiv*, 407007. https://doi.org/10.1101/407007

Seabold, S., & Perktold, J. (2010). Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in science conference*, *57*, 61. https://pdfs.semanticscholar.org/3a27/6417e5350e29cb6bf04ea5a4785601d5a215.pdf

Seijdel, N., Loke, J., van de Klundert, R., van der Meer, M., Quispel, E., van Gaal, S., et al. (2021). On the necessity of recurrent processing during object recognition: It depends on the need for scene segmentation. *Journal of Neuroscience*, *41*, 6281–6289. https://doi.org/10.1523/JNEUROSCI.2851-20.2021, PubMed: 34088797

Seijdel, N., Tsakmakidis, N., de Haan, E. H. F., Bohte, S. M., & Scholte, H. S. (2020). Depth in convolutional neural networks solves scene segmentation. *PLoS Computational Biology*, *16*, e1008022. https://doi.org/10.1371/journal.pcbi.1008022, PubMed: 32706770

Self, M. W., Jeurissen, D., van Ham, A. F., van Vugt, B., Poort, J., & Roelfsema, P. R. (2019). The segmentation of proto-objects in the monkey primary visual cortex. *Current Biology*, *29*, 1019–1029.e4. https://doi.org/10.1016/j.cub.2019.02.016, PubMed: 30853432

Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *ArXiv*. https://doi.org/10.48550/arXiv.1409.1556

Storrs, K. R., Kietzmann, T. C., Walther, A., Mehrer, J., & Kriegeskorte, N. (2021). Diverse deep neural networks all predict human inferior temporal cortex well, after training and fitting. *Journal of Cognitive Neuroscience*, *33*, 2044–2064. https://doi.org/10.1162/jocn_a_01755, PubMed: 34272948

Takagi, Y., & Nishimoto, S. (2022). High-resolution image reconstruction with latent diffusion models from human brain activity. *BioRxiv*, *11*, 517004. https://doi.org/10.1101/2022.11.18.517004

Tartaglini, A. R., Vong, W. K., & Lake, B. M. (2022). A developmentally-inspired examination of shape versus texture bias in machines. *ArXiv*. https://doi.org/10.48550/arXiv.2202.08340

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522. https://doi.org/10.1038/381520a0, PubMed: 8632824

Truzzi, A., & Cusack, R. (2020). Understanding CNNs as a model of the inferior temporal cortex: Using mediation analysis to unpack the contribution of perceptual and semantic features in random and trained networks. *NeurIPS 2020 Workshop SVRHM*. https://openreview.net/forum?id=r7R7VAN6t-k

van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research*, *9*, 2579–2605. https://www.jmlr.org/papers/v9/vandermaaten08a.html

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., et al. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2, PubMed: 32015543

Waskom, M. (2021). seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*, 3021. https://doi.org/10.21105/joss.03021

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*, 356–365. https://doi.org/10.1038/nn.4244, PubMed: 26906502

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *111*, 8619–8624. https://doi.org/10.1073/pnas.1403112111, PubMed: 24812127

Zipser, K., Lamme, V. A., & Schiller, P. H. (1996). Contextual modulation in primary visual cortex. *Journal of Neuroscience*, *16*, 7376–7389. https://doi.org/10.1523/JNEUROSCI.16-22-07376.1996, PubMed: 8929444