Behavioral/Cognitive

On the Necessity of Recurrent Processing during Object Recognition: It Depends on the Need for Scene Segmentation

Noor Seijdel,[1,2*] Jessica Loke,[1,2*] Ron van de Klundert,[1] Matthew van der Meer,[1] Eva Quispel,[1] Simon van Gaal,[1,2] Edward H.F. de Haan,[1,2] and H. Steven Scholte[1,2]

[1]Department of Psychology, University of Amsterdam, 1018 WS Amsterdam, The Netherlands, and [2]Amsterdam Brain and Cognition Center, University of Amsterdam, 1018 WS Amsterdam, The Netherlands

Although feedforward activity may suffice for recognizing objects in isolation, additional visual operations that aid object recognition might be needed for real-world scenes. One such additional operation is figure-ground segmentation, extracting the relevant features and locations of the target object while ignoring irrelevant features. In this study of 60 human participants (female and male), we show objects on backgrounds of increasing complexity to investigate whether recurrent computations are increasingly important for segmenting objects from more complex backgrounds. Three lines of evidence show that recurrent processing is critical for recognition of objects embedded in complex scenes. First, behavioral results indicated a greater reduction in performance after masking objects presented on more complex backgrounds, with the degree of impairment increasing with increasing background complexity. Second, electroencephalography (EEG) measurements showed clear differences in the evoked response potentials between conditions around time points beyond feedforward activity, and exploratory object decoding analyses based on the EEG signal indicated later decoding onsets for objects embedded in more complex backgrounds. Third, deep convolutional neural network performance confirmed this interpretation. Feedforward and less deep networks showed a higher degree of impairment in recognition for objects in complex backgrounds compared with recurrent and deeper networks. Together, these results support the notion that recurrent computations drive figure-ground segmentation of objects in complex scenes.

Key words: deep convolutional neural network; natural scene statistics; object recognition; scene segmentation; visual categorization; visual perception

Significance Statement

The incredible speed of object recognition suggests that it relies purely on a fast feedforward buildup of perceptual activity. However, this view is contradicted by studies showing that disruption of recurrent processing leads to decreased object recognition performance. Here, we resolve this issue by showing that how object recognition is resolved and whether recurrent processing is crucial depends on the context in which it is presented. For objects presented in isolation or in simple environments, feedforward activity could be sufficient for successful object recognition. However, when the environment is more complex, additional processing seems necessary to select the elements that belong to the object and by that segregate them from the background.

Received Nov. 9, 2020; revised Apr. 11, 2021; accepted May 13, 2021.

Author contributions: N.S., J.L., R.v.d.K., M.v.d.M., E.Q., S.v.G., E.H.F.d.H., and H.S.S. designed research; N.S., J.L., R.v.d.K., M.v.d.M., and E.Q. performed research; N.S., J.L., R.v.d.K., M.v.d.M., and E.Q. analyzed data; and N.S. and J.L. wrote the paper

This work was supported by Advanced Investigator Grant (No. 339374) from the European Research Council to E.F.H.d.H. and an Interdisciplinary Doctorate Agreement from the University of Amsterdam to H.S.S. and J.L.

*N.S. and J.L. share first authorship.

The authors declare no competing financial interests.

Correspondence should be addressed to Noor Seijdel at noorseijdel@gmail.com.

https://doi.org/10.1523/JNEUROSCI.2851-20.2021

Introduction

The efficiency and speed of the human visual system during object categorization suggests that a feedforward sweep of visual information processing is sufficient for successful recognition (VanRullen and Thorpe, 2002). For example, when presented with objects in a rapid serial visual presentation task (RSVP; Potter and Levy, 1969) or during rapid visual categorization (Thorpe et al., 1996), human subjects could still successfully recognize these objects, with EEG measurements showing robust object-selective activity within 150 ms after object presentation (VanRullen and Thorpe, 2001). Given that there is substantial

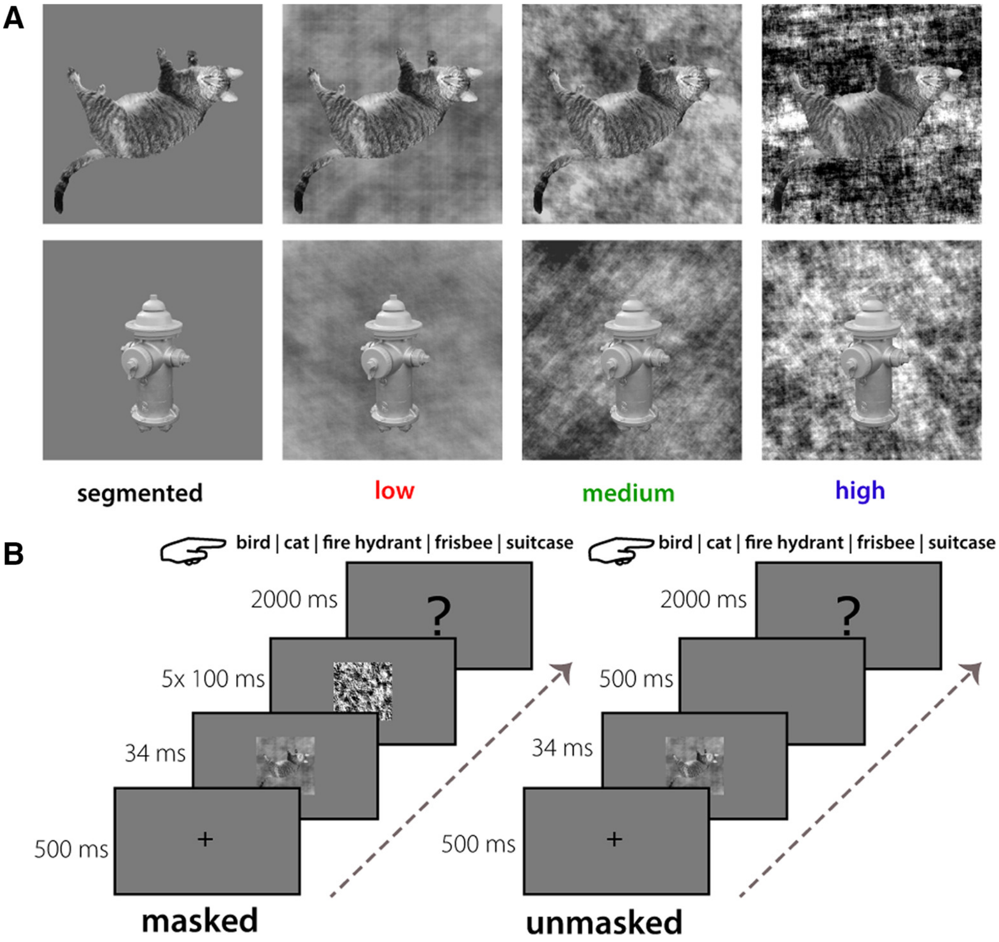

Figure 1. Stimuli and experimental paradigm. ***A***, Exemplars of two categories (cat, fire hydrant) from each stimulus complexity condition. Backgrounds were either uniform (segmented, black), or had low (red), medium (green), or high (blue) CE and SC values. ***B***, Experimental design. On masked trials, the stimulus was followed by a dynamic mask (5 × 100 ms); on unmasked trials this was replaced by a blank screen (500 ms). Participants were asked to categorize the target object by pressing the corresponding button on the keyboard.

evidence for the involvement of recurrent processing in figure-ground segmentation (Lamme and Roelfsema, 2000; Scholte et al., 2008; Wokke et al., 2012), this seems inconsistent with recognition processes that rely on explicit encoding of spatial relationships among parts and suggests instead that rapid recognition may rely on the detection of an unbound collection of image features (Crouzet and Serre, 2011).

Recently, a multitude of studies have reconciled these seemingly inconsistent findings by indicating that recurrent processes might be used adaptively, depending on the visual input or task. Although feedforward activity might suffice for simple scenes with isolated objects, more complex scenes or more challenging conditions (e.g., objects that are occluded or degraded) may need additional visual operations (routines) requiring recurrent computations (Seijdel et al., 2020; Groen et al., 2018; Tang et al., 2018; Spoerer et al., 2018; Kar et al., 2019; Rajaei et al., 2019; Kreiman and Serre, 2020). For objects in isolation or in very simple scenes, although recurrent computations might still aid efficient recognition during more natural viewing (Kietzmann et al., 2019), rapid recognition may thus rely on a coarse and unsegmented feedforward representation (Crouzet and Serre, 2011). For cluttered images, recurrent computations become more important or even necessary as extra visual operations; grouping parts of the object and segmenting it from its background might be needed. Similarly, even for simple scenes, recurrent processing might be used adaptively when the experimental task requires explicit encoding of spatial relationships among parts (e.g., contour detection, curve tracing, or image completion (Roelfsema et al., 1999; Bennett et al., 2016; Linsley et al., 2020).

Several studies have already shown that the segmentability of a natural scene might influence the degree of recurrent processing. For example, Koivisto et al. (2014) reported that masking, a technique shown to affect mainly recurrent but not feedforward processing (Fahrenfort et al., 2007), was more effective for objects that were difficult to segregate. Also in a more recent study, we showed that natural scene complexity, providing information about the segmentability of a scene, modulates the degree of feedback activity (Groen et al., 2018). However, both studies did not test for effects of segmentation explicitly and used natural scenes that were uncontrolled and in which complexity could correlate with other contextual factors in the scene. Therefore, we here systematically investigated whether scene complexity influenced the extent of recurrent processing during object recognition. To this end, participants performed an object recognition task with objects embedded in backgrounds of different complexity (Fig. 1), indexed by two biologically plausible measures: spatial coherence (SC) and contrast energy (CE; Ghebreab et al., 2009; Groen et al., 2013; Scholte et al., 2009). Using these hybrid stimuli, we combined relevant features of objects in natural scenes with well-controlled backgrounds of different complexity.

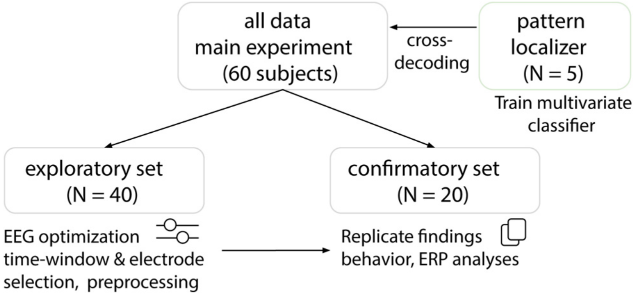

Figure 2. Experimental procedure. Sixty-two participants took part in the EEG experiment. Data from 40 participants were used to perform exploratory analyses. The resulting data (20 participants) were used to confirm our results. For the decoding analyses, five new participants took part in a separate experiment to characterize multivariate EEG activity patterns for the different object categories.

In half the trials, we impaired feedback activity with visual masking. In addition to behavior, we measured electroencephalography (EEG) responses to examine the time course of visually evoked activity. In addition to human participants, we also investigated recognition performance in deep convolutional neural networks (DCNNs), which received visual stimuli identical to that of our human participants, and performed a five-choice recognition task.

A convergence of results from behavior, EEG and DCNNs indicated that recurrent computations were critical for recognition of objects that were more difficult to segment from their background. Together, the results support the notion that recurrent computations drive figure-ground segmentation of objects in complex scenes.

Materials and Methods

Subjects main experiment. Forty-two participants (32 females, 18–35 years old) took part in a first EEG experiment. Data from two participants were excluded from further analysis because of technical problems. We used this first dataset to perform exploratory analyses and optimize our analysis pipeline (Fig. 2). Based on this dataset, we defined the time windows for further evoked response potential (ERP) analyses, electrode selection, and preprocessing steps. To confirm our results on an independent dataset, another 20 participants (13 females, 18–35 years old) were measured. Sample sizes were chosen so that the confirmatory dataset was comparable to earlier work using similar paradigms (Groen et al., 2018; Rajaei et al., 2019). Data from one participant were excluded from ERP analyses because of wrong placement of electrodes I1 and I2.

Stimuli. Images of real-world scenes containing birds, cats, fire hydrants, Frisbees, or suitcases were selected from several online databases, including MS COCO (Lin et al., 2014), the SUN database (Xiao et al., 2010), Caltech-256 (Griffin et al., 2007), Open Images V4 (Kuznetsova et al., 2020), and LabelMe (Russell et al., 2008). These five categories were selected because a large selection of images was available in which the target object was clearly visible and not occluded. For each image, one CE and one SC value was computed using a simple visual model that simulates neuronal responses in one of the earliest stages of visual processing. Specifically, they are derived by averaging the simulated population response of contrast filters at the lateral geniculate nucleus across the visual scene (Ghebreab et al., 2009; Scholte et al., 2009; Groen et al., 2013. Computing these statistics for a large set of scenes results in a two-dimensional space in which sparse scenes with just a few scene elements are separate from complex scenes with a lot of clutter and a high degree of fragmentation.

Together, CE and SC appear to provide information about the segmentability of a scene (Groen et al., 2013, 2018). High CE/SC values correspond with images that contain many edges that are distributed in an uncorrelated manner, resulting in an inherently low figure-ground segmentation. Relatively low CE/SC values, on the other hand, correspond with a homogenous image containing few edges, resulting in an inherently high figure-ground segmentation (Fig. 1). Each object was segmented from a real-world scene background and superimposed on three categories of phase-scrambled versions of the real-world scenes. This corresponded with low, medium, and high complexity scenes. Additionally, the segmented object was also presented on a uniform gray background as the segmented condition (Fig. 1). For each object category, 8 low CE/SC, 8 medium CE/SC, and 8 high CE/SC images were selected, using the cutoff values from Groen et al. (2018), resulting in 24 images for each object category and 120 images in total. Importantly, each object was presented in all conditions, allowing us to attribute the effect to the complexity (i.e., segmentability) of each trial, and rule out any object-specific effects.

Experimental design. Participants performed a five-choice categorization task (Fig. 1), differentiating images containing cats, birds, fire hydrants, Frisbees, and suitcases as accurately as possible. Participants indicated their response using five keyboard buttons corresponding to the different categories. Images were presented in a randomized sequence for a duration of 34 ms. Stimuli were presented at eye level in the center of a 23-inch ASUS TFT LCD display, with a spatial resolution of 1920*1080 pixels, at a refresh rate of 60 Hz. Participants were seated ∼70 cm from the screen so that stimuli subtended a 6.9° visual angle. The object recognition task was programmed in and performed using Presentation software version 18.0 (Neurobehavioral Systems). The experiment consisted of 960 trials in total, of which 480 were backward masked trials, and 480 were unmasked trials, randomly divided into 8 blocks of 120 trials for each participant. After each block, participants took a short break. The beginning of each trial consisted of a 500 ms fixation period in which participants focused their gaze on a fixation cross at the center of the screen. In the unmasked trials, stimuli were followed by a blank screen for 500 ms and then a response screen for 2000 ms. To disrupt recurrent processes (Breitmeyer and Ogmen, 2000; Fahrenfort et al., 2007; Lamme et al., 2002), in the masked trials five randomly chosen phase-scrambled masks were presented sequentially for 500 ms. The first mask was presented immediately after stimulus presentation, and each mask was presented for 100 ms (Fig. 1). The ambient illumination in the room was kept constant across different participants.

Subjects pattern localizer. Five new participants took part in a separate experiment to characterize multivariate EEG activity patterns for the different object categories. For this experiment, we measured EEG activity while participants viewed the original experimental stimuli followed by a word (noun). Participants were asked to press the button only when the image and the noun did not match to ensure attention (responses were not analyzed). A classifier was trained on the EEG data from this experiment and subsequently tested on the data from the main experiment using a cross-decoding approach. All participants had normal or corrected-to-normal vision, provided written informed consent, and received monetary compensation or research credits for their participation. The ethics committee of the University of Amsterdam approved the experiment.

Deep convolutional neural networks. First, to investigate the effect of recurrent connections, we tested different architectures from the CORnet model family (Kubilius et al., 2018), CORnet-Z (feed forward), CORnet-RT (recurrent), and CORnet-S (recurrent with skip connections). In CORnet-RT and CORnet-S, recurrence is introduced only within an area (no feedback connections between areas). In CORnet-S a skip connection is included so that the result of adding the state to the input is combined with the output of the last convolution just before applying a nonlinearity. Recurrent processing in all CORnet models is different from recurrent processing in the brain. However, compared with strictly feedforward models, they include and imitate, to a limited extent, recurrent processing.

Then, to further evaluate the influence of network depth on scene segmentation, tests were conducted on three deep residual networks (ResNets; He et al., 2016) with an increasing number of layers, ResNet-10, ResNet-18, and Resnet-34. Ultra-deep residual networks are mathematically equivalent to a recurrent neural network unfolding over time, when the weights between their hidden layers are clamped (Liao and

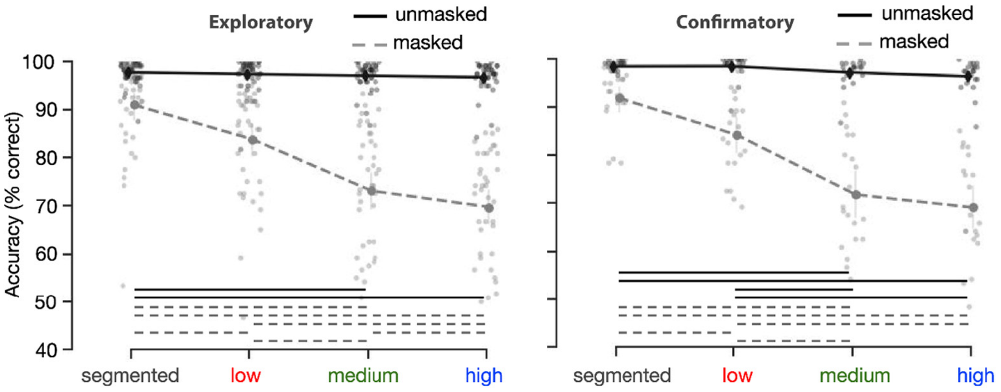

Figure 3. Human performance on the object recognition task. Performance (percentage correct) on the five-option object recognition task. For masked trials, performance decreased with an increase in background complexity. Left, Results from the exploratory set are plotted. Right, Results from the confirmatory set are plotted. Error bars represent the bootstrap 95% confidence interval; dots indicate the average performance of individual participants.

Poggio, 2016). This has led to the hypothesis that the additional layers function in a way that is similar to recurrent processing in the human visual system (Kar et al., 2019).

After initialization of the pretrained networks (on the ImageNet Large Scale Visual Recognition Challenge dataset), the model's weights were fine-tuned for our task on images from the MSCoco database (Lin et al., 2014), using PyTorch (Paszke et al., 2019), generating five probability outputs (for our five object categories). To obtain statistical results, we fine-tuned each network architecture 10 different times.

EEG data acquisition and preprocessing. EEG was recorded using a 64-channel Active Two EEG system (Biosemi) at a 1024 Hz sample rate. As in previous studies investigating early visual processing (Groen et al., 2013, 2018), we used caps with an extended 10–10 layout modified with two additional occipital electrodes (I1 and I2, which replaced F5 and F6). Eye movements were recorded with additional electro-oculograms (vEOG and hEOG). Preprocessing was done using MNE software in Python (Gramfort et al., 2014) and included the following steps for the ERP analyses: (1) After importing, data were rereferenced to the average of two external electrodes placed on the mastoids. (2) High-pass (0.1 Hz, 0.1 Hz transition band) and low-pass (30 Hz, 7.5 Hz transition band) basic finite impulse response filters were sequentially applied. (3) an independent component analysis (Vigario et al., 2000) was run to identify and remove eye-blink- and eye-movement-related noise components (mean = 1.73 of the first 25 components removed per participant). (4) Epochs were extracted from −200 ms to 500 ms from stimulus onset. (5) Trials were normalized by a 200 ms prestimulus baseline. (6) Five percent of trials with the most extreme values within each condition were removed, keeping the number of trials within each condition equal. (7) Data were transformed to current source density responses (Perrin et al., 1989).

Statistical analysis: behavioral data. For human subjects, choice accuracy was computed for each condition in the masked and unmasked trials (Fig. 3). Reaction times were not analyzed as participants were asked to perform the task as accurately as possible and were told that the speed of their reaction was not important. Differences between the conditions were tested using two-factor (scene complexity: segmented, low, med, high; and masking: masked, unmasked) repeated measures ANOVAs. Significant main effects were followed up by *post hoc* pairwise comparisons between conditions using Sidak multiple comparisons correction at $\alpha = 0.05$. For DCNNs, a nonparametric Friedman test was used to differentiate accuracy across the different conditions (segmented, low, medium, high), followed by pairwise comparisons using a Mann–Whitney *U* test. Behavioral data were analyzed in Python using the following packages: Statsmodels, SciPy, NumPy, Pandas, (Jones et al., 2001; Oliphant, 2006; Seabold and Perktold, 2010; McKinney, 2010).

Statistical analysis: EEG event-related potentials. EEG analyses were conducted in Python, using the MNE software. For each participant, the difference in ERPs to scene complexity was computed within masked and unmasked conditions, pooled across occipital and perioccipital electrodes (Oz, POz, O1, O2, PO3, PO4, PO7, PO8). This was done by subtracting the signal of each complexity condition (i.e., low, medium, or high) from the segmented condition. Doing so enabled us to investigate differences among low-, medium-, and high-complex scenes regardless of masking effects. Based on the exploratory dataset, we established five time windows by performing *t* tests on every time point for each condition and selecting windows in which the amplitude differed from zero for all complexity conditions (low, med, high). Then, a repeated measures ANOVA with factor background complexity (low, medium, high) and masking (masked, unmasked) was performed on the average activity in these established time windows.

Statistical analysis: EEG—exploratory multivariate classification. The same preprocessing pipeline was used as for the ERP analyses. To evaluate how object category information in our EEG signal evolves over time, cross-decoding analyses were performed by training a support vector machine classifier on all trials from the pattern localizer experiment (performed by five different subjects) and testing it on each of the main experiment conditions. Object category classification was performed on a vector of EEG amplitudes across 22 electrodes, including occipital (I1, Iz, I2, O1, Oz, O2), perioccipital (PO3, PO7, POz, PO4, PO8), and parietal (Pz, P1-P10) electrodes. Per condition, decoding accuracy was tested against chance (20%) and against the other conditions using Wilcoxon signed rank tests. Given the large number of statistical comparisons, all *p* values were corrected for multiple comparisons across the two masking conditions, four complexity conditions and 151 time points by means of false discovery rate (FDR) correction at $\alpha = 0.01$.

Differences in onset latency were compared using Wilcoxon signed-rank tests. Following the procedure in Rajaei et al. (2019), we defined onset latency as the earliest time where performance became significantly above chance for at least three consecutive time points (~11.7 ms). Onset latencies were calculated by leave one subject out, repeated for $N = 58$ times.

Data availability. Data and code to reproduce the analyses in this article are available at https://github.com/noorseijdel/2020_EEG_figureground.

Results

Behavior

During the task, participants viewed images of objects placed on top of a gray (segmented), low-, medium-, or high-complexity background. On each trial, they indicated the object category the scene contained, using the corresponding keyboard buttons. In half of the trials, the target image was followed by a dynamic backward mask (5 × 100 ms); the other half of the trials was unmasked (Fig. 1). Accuracy (percentage correct trials) was computed for each participant. A repeated measures ANOVA on the exploratory dataset ($N = 40$), with factors background (segmented, low, medium, high) and masking (masked, unmasked) indicated, apart from main effects, an interaction effect. Results indicated that masking impaired performance for objects presented

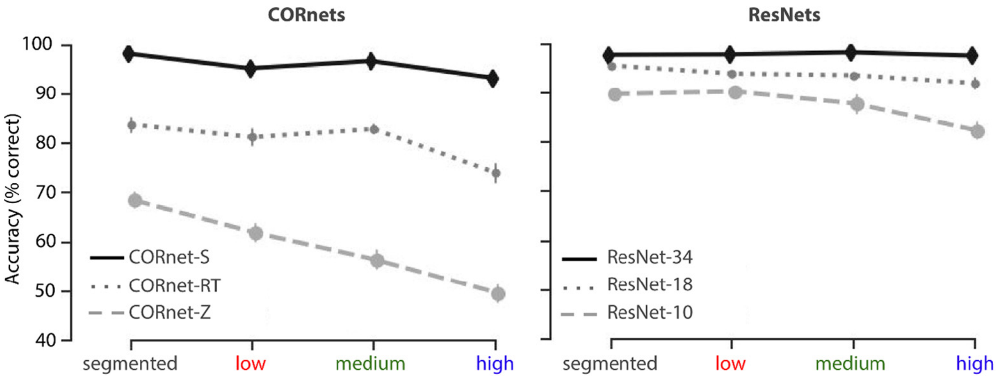

Figure 4. Deep convolutional neural network performance on the object recognition task. Performance (percentage correct) on the five-option object recognition task. Networks were fine-tuned on the five target categories; top-1 accuracy was computed. For the CORnets (left), performance of the feedforward architecture decreased with an increase in background complexity. For recurrent architectures, this decrease was less prominent. For CORnet-S, there was no difference between conditions. Error bars represent the bootstrap 95% confidence interval.

on more complex backgrounds stronger than for less complex backgrounds ($F_{(3,117)}$ = 185.6748, $p < 0.001$). *Post hoc* comparisons showed that for masked trials, accuracy decreased for both medium ($t_{(39)}$ = 2.88, p = 0.038 Sidak-corrected) and high ($t_{(39)}$ = 3.84, p = 0.003 Sidak-corrected) complexity condition compared with the low condition (all other $p > 0.203$). For unmasked trials, all conditions differed from each other, with an incremental decrease in accuracy for objects presented on more complex backgrounds. Analysis of the confirmatory dataset (N = 20) indicated similarly, apart from the main effects, an interaction between masking and background complexity. For masked trials, there was a larger decrease in performance with an increase in background complexity ($F(_{3,57}$ = 101.3338, $p < 0.001$). *Post hoc* comparisons showed that for masked trials, accuracy decreased for both medium- and high-complexity conditions compared with the segmented ($t_{(19)}$ = 3.47, p = 0.003 Sidak-corrected; ($t_{(19)}$ = 3.47, p = 0.003 Sidak corrected) and low conditions ($t_{(19)}$ = 4.23, $p < 0.001$ Sidak-corrected, ($t_{(19)}$ = 4.31, $p < 0.001$ Sidak corrected). For unmasked trials, all conditions differed from each other with the exception of medium and high, with an incremental decrease in accuracy for objects presented on more complex backgrounds.

Network performance

Next, we presented the same images to deep convolutional neural networks with different architectures. For the CORnets (Fig. 4, left), a nonparametric Friedman test differentiated accuracy across the different conditions (segmented, low, medium, high) for all architectures, Friedman's Q(3) = 27.8400, 24.7576, 26.4687 for CORnet-Z, -RT, and -S respectively, all $p < .001$. A Mann–Whitney *U test* on the difference in performance between segmented and high complexity trials indicated a smaller decrease in performance for CORnet-S compared with CORnet-Z (Mann–Whitney U = 100.0, n1 = n2 = 10, $p < 0.001$, two tailed). For the ResNets (Fig. 4, right), a nonparametric Friedman test differentiated accuracy across the different conditions for ResNet-10 and ResNet-18, Friedman's Q (3) = 23.9053, 22.9468, for ResNet-10 and ResNet-18, respectively, both $p < .001$. A Mann–Whitney *U test* on the difference in performance between segmented and high-complexity trials indicated a smaller decrease in performance for ResNet-34 compared with ResNet-10 (Mann–Whitney U = 100.0, n1 = n2 = 10, $p < 0.001$, two tailed). Overall, in line with human performance, results indicated a higher degree of impairment in recognition for objects in complex backgrounds for feedforward or more shallow networks, compared with recurrent or deeper networks.

EEG event-related potentials

To investigate the time course of figure-ground segmentation in the visual cortex, evoked responses to the masked and unmasked scenes were pooled across occipital and perioccipital electrodes (Oz, POz, O1, O2, PO3, PO4, PO7, PO8) for each condition. Although we certainly do not rule out frontal effects and acknowledge that (recurrent) processing through frontal regions contributes to object recognition (Kar and DiCarlo, 2021; Scholte et al., 2008), we based our pooling on previous work showing neural correlates of figure-ground segmentation in these channels (Scholte et al., 2008; Pitts et al., 2011; Wokke et al., 2012; Groen et al., 2018).

Difference waves were generated by subtracting the signal of each condition from the segmented condition (Fig. 5*B*,*E*). Doing so enabled us to eliminate the effect of masking on the EEG signal and to investigate differences between low-, medium-, and high-complex scenes. For each participant, data were averaged across five time windows based on analyses on the exploratory dataset (see above, Materials and Methods).

For every time window, a repeated measures ANOVA was performed on the average EEG amplitude of the difference waves, with complexity (low, medium, high) and masking (masked, unmasked) as within-subject factors. As the preprocessing procedure and time point selection were based on t tests on the exploratory set, we do not report subsequent repeated measures ANOVA for this dataset. Results on the confirmatory dataset (Fig. 5*D*–*F*) showed no main or interaction effects in the first time window (92–115 ms; Fig. 5*F*). Critically, differences among complexity conditions only emerged in time windows 2 and 3 (120–150 ms: $F_{(36)}$ = 22.87, η^{2par} = 0.56, $p < 0.001$; 155–217 ms: $F_{(36)}$ = 24.21, η^{2par} = 0.57, $p < 0.001$), suggesting a differential contribution of recurrent processing to object recognition in varying complexity scenes. In time window 2, there was a main effect of masking ($F_{(18)}$ = 5.38, η^{2par} = 0.576, p = .03). Only in time window 4 (221–275 ms), an interaction effect of masking and complexity ($F_{(18)}$ = 59.60, η^{2par} = 0.07, $p < 0.001$) started to emerge.

Exploratory: EEG multivariate classification

To further investigate the representational dynamics of object recognition under different complexity conditions, exploratory multivariate decoding analyses were performed on the EEG data from all participants (N = 58; Fig. 6). As compared with the confirmatory behavioral and ERP analyses, we used data from all participants to increase statistical power, and we did not a priori decide on the testing procedure. To control for response-related activity (keyboard buttons were fixed across the task), a cross-

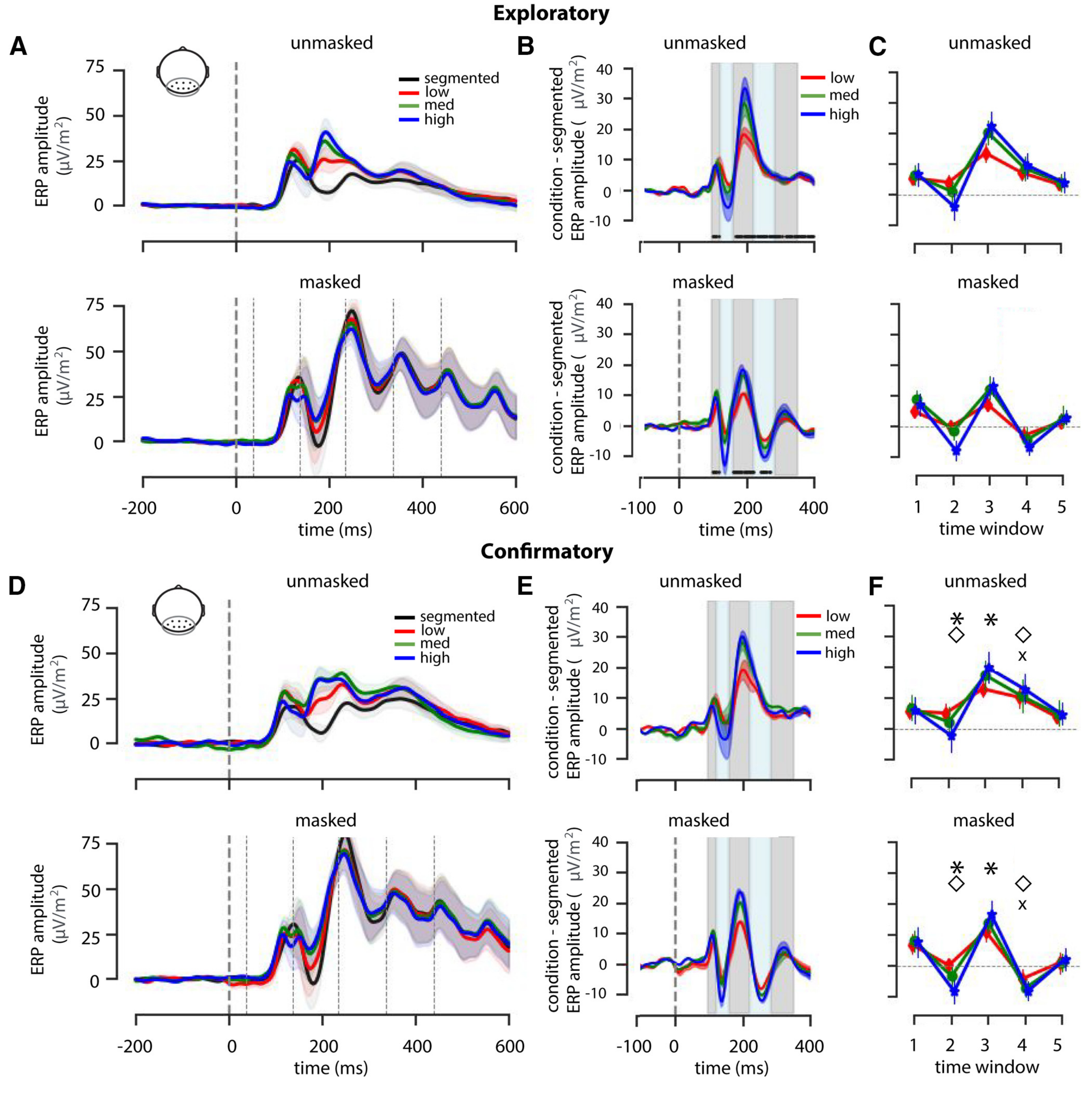

Figure 5. ERP results. ***A***, Average ERP amplitude for segmented, low, medium, and high complexity scenes for an occipital-perioccipital pooling of EEG channels (0z, P0z, 01, 02, P03, P04, P07, P08) for masked and unmasked trials. Shaded regions indicate SEM across participants. Mask onsets are indicated with thin dashed lines (bottom only). ***B***, Difference waves were generated by subtracting the signal of each condition from the segmented condition. ***C***, Based on significant time points in the exploratory dataset, five time windows were defined: 92–115 ms, 120–150 ms, 155–217 ms, 221–275 ms, 279–245 ms). Symbol markers indicate main or interaction effects: asterisk, main effect of condition; diamond, main effect of masking; × sign, interaction effect. ***D***–***F***, Analyses repeated for the confirmatory dataset.

decoding analysis was performed by training the classifier on all trials from an independent pattern localizer experiment and testing it on each of the main experiment conditions (see above, Materials and Methods). Statistical comparisons (FDR corrected across all time-points, masking conditions and complexity conditions) of the decoding accuracy indicated above-chance decoding for unmasked segmented trials and low trials early in time (Fig. 6; ~106 and ~110, respectively). For objects on a medium-complex background, decoding accuracy did not diverge from chance before ~186 ms. For objects on high-complex backgrounds, there was no successful decoding. Direct comparison of the decoding accuracy across conditions indicated that decoding accuracy for the segmented and low conditions was significantly enhanced compared with to high and medium conditions (Fig. 6; blue x, high vs segmented; blue dot, high vs low, green x, medium vs segmented; green dot, medium vs low). Comparison of the onset latencies (earliest time where performance became significantly above chance for at least three consecutive time points) indicated the earliest onset latency for segmented trials, followed by low and then medium trials (segmented vs low W = 15; segmented vs medium W = 0; low vs medium W = 0; all $p < 0.001$, two-sided Wilcoxon signed rank). For masked trials,

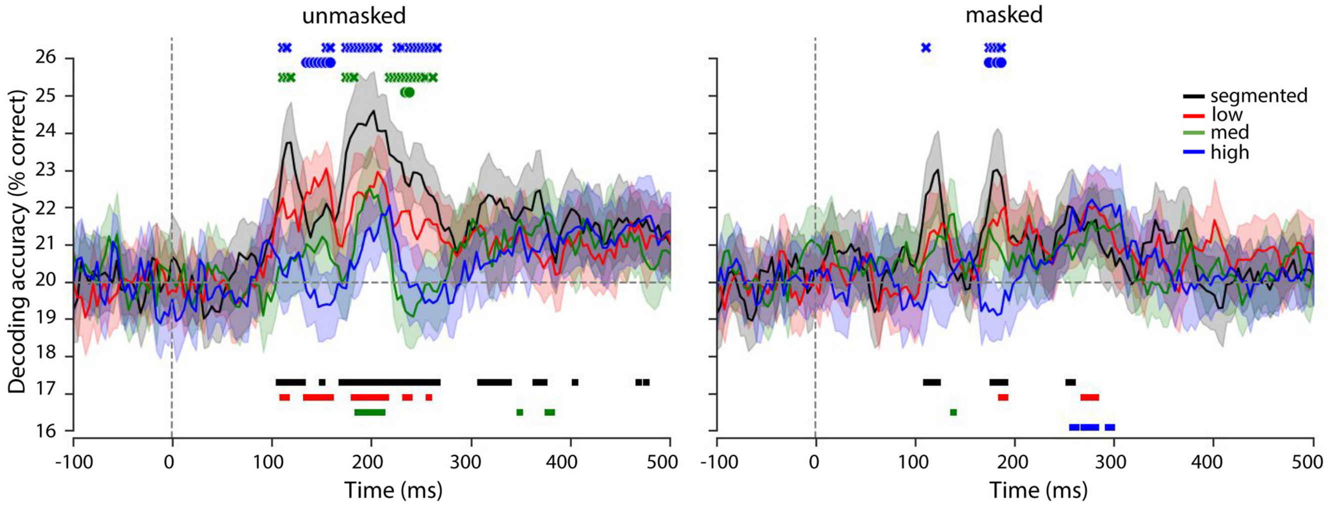

Figure 6. Cross-decoding results using the pattern localizer. Decoding object category in EEG signal for masked and unmasked trials with varying complexity. The dotted line represents the 20% chance level, shaded error bars represent the bootstrap 95% confidence interval. Results from the Wilcoxon signed rank tests across chance (20%) are indicated with thick lines at the bottom of the graphs. Symbol markers at the top of the graphs indicate significant differences in decoding accuracy between conditions: blue x, high versus segmented; blue dot, high versus low; green x, medium versus segmented; green dot, medium versus low. All *p* values are corrected for multiple comparisons using an FDR of 0.01.

successful decoding for the segmented objects started ~111 ms, followed by later additional decoding of low-complexity (186 ms) and high-complexity (257 ms) trials. Again, decoding accuracy was enhanced for the segmented and low condition compared with images with a high-complex background. Comparison of the onset latencies indicated an earlier onset for segmented trials compared with low trials (W = 0, $p < 0.001$). For high trials, the presence of a significant onset (>3 time points) varied across repetitions, depending on which subject was left out.

Overall, these findings showed that different objects evoked reliably different sensor patterns when presented in isolation or in simple environments within the first feedforward sweep of visual information processing. Additionally, results indicated decreased and later decoding for objects embedded in more complex backgrounds, suggesting that object representations for objects on complex backgrounds emerge later. Finally, these findings indicate that object category representations generalized across tasks and participants. Whether these effects are robust should emerge from future confirmatory research.

Discussion

This study systematically investigated whether recurrent processing is required for figure-ground segmentation during object recognition. A converging set of behavioral, EEG and computational modeling results indicated that recurrent computations are required for figure-ground segmentation of objects in complex scenes. These findings are consistent with previous findings showing enhanced feedback for complex scenes (Groen et al., 2018) and visual backward masking being more effective for images that were more difficult to segment (Koivisto et al., 2014). We interpret these results as showing that figure-ground segmentation, driven by recurrent processing, is not necessary for object recognition in simple scenes, but it is for more complex scenes.

Effects of scene complexity using artificial backgrounds

In an earlier study using natural scenes, we already showed that feedback was selectively enhanced for high-complexity scenes, during an animal detection task. Although there are numerous reasons for using naturalistic scenes (Felsen et al., 2005; Felsen and Dan, 2005; Talebi and Baker, 2012), it is difficult to do controlled experiments with them because they vary in many (unknown) dimensions. For example, SC and CE (measures of scene complexity) could correlate with other contextual factors in the scene (e.g., SC correlates with perception of naturalness of a scene (Groen et al., 2013) and could be used as diagnostic information for the detection of an animal. Additionally, previous research has shown that natural scenes and scene structure can facilitate object recognition (Davenport and Potter, 2004; Kaiser and Cichy, 2018; Neider and Zelinsky, 2006). Results from the current study, using artificial backgrounds of varying complexity, replicate earlier findings while allowing us to attribute the effects to SC and CE and the subsequent effect on segmentability. A limitation of any experiment with artificially generated (or artificially embedded) images is that it is not clear whether the findings will generalize to real images that have not been manipulated in any way. Together with the previous findings, however, our results corroborate the idea that more extensive processing (possibly in the form of recurrent computations) is required for object recognition in more complex, natural environments (Groen et al., 2018; Rajaei et al., 2019).

Time course of object recognition

Based on the data from the exploratory dataset (N = 40), we selected five time windows in the ERPs to test our hypotheses on the confirmatory dataset. For our occipital-perioccipital pooling, we expected the first feedforward sweep to be unaffected by scene complexity (i.e., low, medium, high). Indeed, amplitudes of the difference waves (complexity condition, segmented ERP amplitudes) averaged across the selected time windows indicated no influence of masking or scene complexity early in time (92–115 ms). The observation that all three difference waves deviated from zero, however, indicates that there was an effect of segmentation. In this early time window, background presence thus seems to be more important than the complexity of the background. This difference could be attributed to the detection of additional low-level features in the low-, medium-, and high-complexity condition, activating a larger set of neurons that participate in the first feedforward sweep (Lamme and Roelfsema, 2000). In the second and third time window (120–217 ms), differences among the complexity conditions emerged. We interpret these differences as reflecting the increasing need for recurrent processes when backgrounds are more complex.

Our results are generally consistent with prior work investigating the time course of visual processing of objects under more or less challenging conditions (Cichy et al., 2014; Contini et al., 2017; CDiCarlo and Cox, 2007; Rajaei et al., 2019; Tang et al., 2018). In line with multiple earlier studies, masking left the early evoked neural activity (<120 ms) relatively intact, whereas the neural activity after ~150 ms was decreased (Boehler et al., 2008; Del Cul et al., 2007; Fahrenfort et al., 2007; Koivisto and Revonsuo, 2010; Lamme et al., 2002; Lamme and Roelfsema, 2000).

Exploratory decoding results corroborated these findings, showing decreased or delayed decoding onsets for objects embedded in more complex backgrounds, suggesting that object representations for those images emerge later. Additionally, when recurrent processing was impaired using backward masking, only objects presented in isolation or in simple environments evoked reliably different sensor patterns that our classifiers were able to pick up (Figs. 5, 6).

Influence of masking on behavior

Based on the strong interaction effect on behavior, it is tempting to conclude that complexity significantly increases the effect of masking on recognition accuracy. However, performance on all unmasked trials was virtually perfect (96–97%), raising concerns about ceiling effects obscuring the actual variation among these conditions (Uttl, 2005). Therefore, although masked stimuli show a decrease in performance along increases in complexity, based on the current findings we cannot conclude that this is because of masking (i.e., reducing recurrent processes). Although we do not claim that unmasked segmented, low, medium, or high images are equally difficult or processed in the same way (we actually argue for the opposite), our results show that apparently the brain is capable of arriving at the correct answer with enough time. It is hard to come up with an alternative (more difficult) task without affecting our experimental design and subsequent visual processing (e.g., stimulus degradation generally affects low-level complexity; reducing object size or varying object location creates a visual search task that could benefit from spatial layout properties). Combined functional magnetic resonance imaging and EEG results from an earlier study already showed that for complex scenes only, early visual areas were selectively engaged by means of a feedback signal (Groen et al., 2018). Here, using controlled stimuli and backward masking, we replicated and expanded on these findings. Importantly, results from both EEG and deep convolutional neural networks supported the notion that recurrent computations drive figure-ground segmentation of objects in complex scenes.

Probing cognition with deep convolutional neural networks

One way to understand how the human visual system processes visual information involves building computational models that account for human-level performance under different conditions. Here we used deep convolutional neural networks because they show remarkable performance on both object and scene recognition (He et al., 2016; Russakovsky et al., 2015). Although we certainly do not aim to claim that DCNNs are identical to the human brain, we argue that studying how performance of different architectures compares to human behavior could be informative about the type of computations that are underlying this behavior (Cichy and Kaiser, 2019). In the current study, it provides an additional test for the involvement of recurrent connections. Comparing the (behavioral) results of DCNNs with findings in humans, our study adds to a growing realization that more extensive processing, in the form of recurrent computations, is required for object recognition in more complex, natural environments (Groen et al., 2018; Tang et al., 2018; Kar et al., 2019; Rajaei et al., 2019). Here it's important to note that although multiple runs of fine-tuning do introduce some variance, the current results rely on a single DCNN instance only. As shown by Mehrer et al. (2020), different network seeds may give rise to substantial differences in the network internal representations. Additionally in the current study, using different pretrained networks, the question remains whether our behavioral findings are the result of network architecture (recurrent connections and network depth) and/or network training (optimization). However, a separate investigation comparing object decoding performance based on the final layer of trained and untrained networks (J. Loke, N. Seijdel, L. Snoek, R. van de Klundert, M. van der Meer, E. Quispel, N. Cappaert, and H.S. Scholte, unpublished observations) indicated a lack of convergence for untrained networks, suggesting that there is no reliable object representation in these layers and that above-chance performance on our experimental task requires training. Further analyses, examining the contribution of different layers within the networks, indicated that layers from trained networks are better at explaining variance in neural activity and that untrained networks do capture background complexity properties (segmentation, scene complexity), but not object category.

Conclusion

Results from the current study show that how object recognition is resolved depends on the context in which the target object appears. For objects presented in isolation or in simple environments, object recognition appears to be dependent on the object itself, resulting in a problem that can likely be solved within the first feedforward sweep of visual information processing on the basis of unbound features (Crouzet and Serre, 2011). When the environment is more complex, recurrent processing seems necessary to group the elements that belong to the object and segregate them from the background.

References

Bennett M, Petro L, Muckli L (2016) Investigating cortical feedback of objects and background scene to foveal and peripheral V1 using fMRI. J Vis 16:568–568.

Boehler CN, Schoenfeld MA, Heinze H-J, Hopf J-M (2008) Rapid recurrent processing gates awareness in primary visual cortex. Proc Natl Acad Sci U S A 105:8742–8747.

Breitmeyer BG, Ogmen H (2000) Recent models and findings in visual backward masking: a comparison, review, and update. Percept Psychophys 62:1572–1595.

Cichy RM, Kaiser D (2019) Deep neural networks as scientific models. Trends Cogn Sci 23:305–317.

Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and time. Nat Neurosci 17:455–462.

Contini EW, Wardle SG, Carlson TA (2017) Decoding the time-course of object recognition in the human brain: from visual features to categorical decisions. Neuropsychologia 105:165–176.

Crouzet SM, Serre T (2011) What are the visual features underlying rapid object recognition? Front Psychol 2:326.

Davenport JL, Potter MC (2004) Scene consistency in object and background perception. Psychol Sci 15:559–564.

Del Cul A, Baillet S, Dehaene S (2007) Brain dynamics underlying the nonlinear threshold for access to consciousness. PLoS Biol 5:e260.

DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. Trends Cogn Sci 11:333–341.

Fahrenfort JJ, Scholte HS, Lamme VA (2007) Masking disrupts reentrant processing in human visual cortex. J Cogn Neurosci 19:1488–1497.

Felsen G, Dan Y (2005) A natural approach to studying vision. Nat Neurosci 8:1643–1646.

Felsen G, Touryan J, Han F, Dan Y (2005) Cortical sensitivity to visual features in natural scenes. PLoS Biol 3:e342.

Ghebreab S, Smeulders AWM, Scholte HS, Lamme VAF (2009). A biologically plausible model for rapid natural image identification. Advances in Neural Information Processing Systems, pp 629–637.

Gramfort A, Luessi M, Larson E, Engemann DA, Strohmeier D, Brodbeck C, Parkkonen L, Hämäläinen MS (2014) MNE software for processing MEG and EEG data. Neuroimage 86:446–460.

Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset. Technical Report 7694, California Institute of Technology, Pasadena.

Groen IIA, Ghebreab S, Prins H, Lamme VAF, Scholte HS (2013) From image statistics to scene gist: evoked neural activity reveals transition from low-level natural image structure to scene category. J Neurosci 33:18814–18824.

Groen IIA, Jahfari S, Seijdel N, Ghebreab S, Lamme VAF, Scholte HS (2018) Scene complexity modulates degree of feedback activity during object detection in natural scenes. PLoS Comput Biol 14:e1006690.

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. Paper presented at IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, June.

Jones E, Oliphant T, Peterson P (2001) SciPy: open source scientific tools for Python. Austin, TX: SciPy.

Kaiser D, Cichy RM (2018) Typical visual-field locations facilitate access to awareness for everyday objects. Cognition 180:118–122.

Kar K, DiCarlo JJ (2021) Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. Neuron 109:164–176.e5.

Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ (2019) Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. Nat Neurosci 22:974–983.

Kietzmann TC, Spoerer CJ, Sörensen LKA, Cichy RM, Hauk O, Kriegeskorte N (2019) Recurrence is required to capture the representational dynamics of the human visual system. Proc Natl Acad Sci U S A 116:21854–21863.

Koivisto M, Revonsuo A (2010) Event-related brain potential correlates of visual awareness. Neurosci Behav Rev 34: 922–934.

Koivisto M, Kastrati G, Revonsuo A (2014) Recurrent processing enhances visual awareness but is not necessary for fast categorization of natural scenes. J Cogn Neurosci 26:223–231.

Kreiman G, Serre T (2020) Beyond the feedforward sweep: feedback computations in the visual cortex. Ann N Y Acad Sci 1464:222–241.

Kubilius J, Schrimpf M, Nayebi A, Bear D, Yamins DLK, DiCarlo JJ (2018) CORnet: modeling the neural mechanisms of core object recognition. BioRxiv. doi: 10.1101/408385.

Kuznetsova A, Rom H, Alldrin N, Uijlings J, Krasin I, Pont-Tuset J, Kamali S, Popov S, Malloci M, Kolesnikov A, Duerig T, Ferrari V (2020) The open images dataset v4. Int J Comput Vis 128:1956–1981.

Lamme VA, Roelfsema PR (2000) The distinct modes of vision offered by feedforward and recurrent processing. Trends Neurosci 23:571–579.

Lamme VAF, Zipser K, Spekreijse H (2002) Masking interrupts figure-ground signals in V1. J Cogn Neurosci 14:1044–1053.

Liao Q, Poggio T (2016) Bridging the gaps between residual learning, recurrent neural networks and visual cortex. arXiv 1604.03640.

Lin TY, Maire M, Belongie S, Hays J, Perona P, Ramanan D, Dollár P, and Zitnick CL (2014) Microsoft coco: Common objects in context. In European conference on computer vision, pp 740–755, Springer, Cham.

Linsley D, Kim J, Ashok A, Serre T (2020) Recurrent neural circuits for contour detection. arXiv 2010.15314.

McKinney W (2010) Data structures for statistical computing in python. In: Proceedings of the 9th Python in Science Conference, pp 51–56, Austin.

Mehrer J, Spoerer CJ, Kriegeskorte N, and Kietzmann TC (2020) Individual differences among deep neural network models. bioRxiv. doi: 10.1101/2020.01.08.898288. .

Neider MB, Zelinsky GJ (2006) Scene context guides eye movements during visual search. Vision Res 46:614–621.

Oliphant TE (2006) A guide to NumPy, Vol 1, p 85. USA: Trelgol Publishing.

Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, Desmaison A (2019). Pytorch: An imperative style, high-performance deep learning library. arXiv 1912.01703.

Perrin F, Pernier J, Bertrand O, Echallier JF (1989) Spherical splines for scalp potential and current density mapping. Electroencephalogr Clin Neurophysiol 72:184–187.

Pitts MA, Martínez A, Brewer JB, Hillyard SA (2011) Early stages of figure-ground segregation during perception of the face-vase. J Cogn Neurosci 23:880–895.

Potter MC, Levy EI (1969) Recognition memory for a rapid sequence of pictures. J Exp Psychol 81:10–15.

Rajaei K, Mohsenzadeh Y, Ebrahimpour R, Khaligh-Razavi S-M (2019) Beyond core object recognition: recurrent processes account for object recognition under occlusion. PLoS Comput Biol 15:e1007001.

Roelfsema PR, Scholte HS, Spekreijse H (1999) Temporal constraints on the grouping of contour segments into spatially extended objects. Vision Res 39:1509–1529.

Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, Berg AC, Fei-Fei L (2015) ImageNet large scale visual recognition challenge. Int J Comput Vis 115:211–252.

Russell BC, Torralba A, Murphy KP, Freeman WT (2008) LabelMe: a database and web-based tool for image annotation. Int J Comput Vis 77:157–173.

Scholte HS, Jolij J, Fahrenfort JJ, Lamme VAF (2008) Feedforward and recurrent processing in scene segmentation: electroencephalography and functional magnetic resonance imaging. J Cogn Neurosci 20:2097–2109.

Scholte HS, Ghebreab S, Waldorp L, Smeulders AWM, Lamme VAF (2009) Brain responses strongly correlate with Weibull image statistics when processing natural images. J Vis 9(4):29 1–15.

Seabold S, Perktold J (2010) Statsmodels: econometric and statistical modeling with python. Paper presented at the 9th Python in Science Conference, Austin, TX, June–July. .

Seijdel N, Tsakmakidis N, de Haan EHF, Bohte SM, Scholte HS (2020) Depth in convolutional neural networks solves scene segmentation. PLoS Comput Biol 16:e1008022.

Spoerer CJ, McClure P, Kriegeskorte N (2018) Corrigendum: recurrent convolutional neural networks: a better model of biological object recognition. Front Psychol 9:1695.

Talebi V, Baker CL Jr (2012) Natural versus synthetic stimuli for estimating receptive field models: a comparison of predictive robustness. J Neurosci 32:1560–1576.

Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, Ortega Caro J, Hardesty W, Cox D, Kreiman G (2018) Recurrent computations for visual pattern completion. Proc Natl Acad Sci U S A 115:8835–8840.

Thorpe S, Fize D, Marlot C (1996) Speed of processing in the human visual system. Nature 381:520–522.

Uttl B (2005) Measurement of individual differences: lessons from memory assessment in research and clinical practice. Psychol Sci 16:460–467.

VanRullen R, Thorpe SJ (2001) The time course of visual processing: from early perception to decision-making. J Cogn Neurosci 13:454–461.

VanRullen R, Thorpe SJ (2002) Surfing a spike wave down the ventral stream. Vision Res 42:2593–2615.

Vigario R, Sarela J, Jousmiki V, Hamalainen M, Oja E (2000) Independent component approach to the analysis of EEG and MEG recordings. IEEE Trans Biomed Eng 47: 589–593.

Wokke ME, Sligte IG, Steven Scholte H, Lamme VAF (2012) Two critical periods in early visual cortex during figure-ground segregation. Brain Behav 2:763–777.

Xiao J, Hays J, Ehinger KA, Oliva A, Torralba A (2010) SUN database: large-scale scene recognition from abbey to zoo. Paper presented at the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, San Francisco, CA, June.