A Critical Test of Deep Convolutional Neural Networks' Ability to Capture Recurrent Processing in the Brain Using Visual Masking

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

Abstract

■ Recurrent processing is a crucial feature in human visual processing supporting perceptual grouping, figure-ground segmentation, and recognition under challenging conditions. There is a clear need to incorporate recurrent processing in deep convolutional neural networks, but the computations underlying recurrent processing remain unclear. In this article, we tested a form of recurrence in deep residual networks (ResNets) to capture recurrent processing signals in the human brain. Although ResNets are feedforward networks, they approximate an excitatory additive form of recurrence. Essentially, this form of recurrence consists of repeating excitatory activations in response to a static stimulus. Here, we used ResNets of varying depths (reflecting varying levels of recurrent processing) to explain EEG activity within a visual masking paradigm. Sixty-two humans and 50 artificial agents (10 ResNet models of depths −4, 6, 10, 18, and 34) completed an object categorization task. We show that deeper networks explained more variance in brain activity compared with shallower networks. Furthermore, all ResNets captured differences in brain activity between unmasked and masked trials, with differences starting at ~98 msec (from stimulus onset). These early differences indicated that EEG activity reflected "pure" feedforward signals only briefly (up to ~98 msec). After ~98 msec, deeper networks showed a significant increase in explained variance, which peaks at ~200 msec, but only within unmasked trials, not masked trials. In summary, we provided clear evidence that excitatory additive recurrent processing in ResNets captures some of the recurrent processing in humans. ■

INTRODUCTION

Deep convolutional neural networks (DCNNs) are currently the best mechanistic models of object recognition and best at predicting human neural visual processing dynamics (Kietzmann, McClure, & Kriegeskorte, 2019; Yamins & DiCarlo, 2016; Kriegeskorte, 2015). However, DCNNs have also been critiqued for lacking crucial biological features, such as recurrent processing (van Bergen & Kriegeskorte, 2020; Kietzmann, Spoerer, et al., 2019). Feedforward and recurrent processing are two modes of visual processing in biological brains (Lamme & Roelfsema, 2000). Although there is no strict separation between both modes, the distinction of both modes has been made spatio-temporally from electrophysiology, lesion studies, and neuropharmacological interventions (Lamme, Supèr, & Spekreijse, 1998). Feedforward processing is a rapid set of computations evoked by sensory information, also known as bottom–up processing (Serre, Oliva, & Poggio, 2007; Thorpe, Fize, & Marlot, 1996). Recurrent processing sets in right after the initial feedforward computations and is known to include both lateral and top–down processing. In the context of object recognition, recurrent processing is believed to be essential for recognizing noisy, occluded objects (Rajaei, Mohsenzadeh, Ebrahimpour, & Khaligh-Razavi, 2019; Spoerer, McClure, & Kriegeskorte, 2017; Tang & Kreiman, 2017), perceptual-grouping (Roelfsema, 2006), and figure-ground segmentation (Scholte, Jolij, Fahrenfort, & Lamme, 2008; Fahrenfort, Scholte, & Lamme, 2007). Its importance in human visual processing has led researchers to assert the lack of recurrent processes in DCNNs as a crucial limitation (Kreiman & Serre, 2020; Kietzmann, Spoerer, et al., 2019).

Consequently, researchers have attempted to incorporate recurrent processes in the architecture of DCNNs. This evoked discussions on different ways to model recurrent processes as researchers from separate disciplines have modeled recurrent processing differently. Physiologists who are primarily concerned with biological realism have modeled recurrent processing as an interaction between feedforward and feedback signals, influenced by neurons' feature preferences and asymmetry of feedforward and feedback signals (Mély, Linsley, & Serre, 2018; Roelfsema, Lamme, & Spekreijse, 2002). Others, mainly from the field of computer vision, modeled recurrent processing as a summation between feedforward and feedback signals (Kietzmann, Spoerer, et al., 2019;

University of Amsterdam, the Netherlands
*Contributed equally.

Kubilius, Schrimpf, Kar, Hong, & Majaj, 2019; Tang et al., 2018). Regardless of approach, these different attempts at incorporating recurrent signals in models made clear that recurrent processing improves model performance—especially when the task at hand is challenging. Thus, there is a consensus on a necessity to incorporate recurrent processing in DCNNs; however, the appropriate level of complexity in approximating recurrent signals during the task of object recognition remains unresolved. One of the simplest models for recurrent processing (as proposed by the graduate student of the founder of computational neuroscience, Tomaso Poggio, and himself) is a *deep residual network* (referred to as "ResNets" here; He, Zhang, Ren, & Sun, 2016). The proposition of using ResNets as a model for recurrent processing follows from the observation that improvement in DCNNs performance over the years came mainly from incorporating additional layers in the network architecture, and these additional layers mimicked recurrent processes in primate brains (Liao & Poggio, 2016). Furthermore, Liao and Poggio have shown that ResNets' computations are equivalent to unrolled time steps of recurrent computations in recurrent neural networks (RNNs), leading the authors to the conclusion that "moderately deep RNNs are a biologically-plausible model of the ventral stream in visual cortex." Therefore, in our study, we used a family of ResNet models of different depths (ResNet-4, 6, 10, 18, 34) as proxies for varying levels of recurrent processing to model recurrent signals in the human visual system. For each ResNet, we initialized and trained 10 seeds of the network, resulting in 50 artificial agents that we treat akin to animal models (Scholte, 2018).

Earlier studies have shown that task difficulty is an important factor for the occurrence of recurrent processes (Kar, Kubilius, Schmidt, Issa, & DiCarlo, 2019; Rajaei et al., 2019; Groen et al., 2018; Tang et al., 2018; Spoerer et al., 2017). In an earlier study, Seijdel et al. (2021) has similarly shown that image complexity predicts the occurrence of recurrent signals. Therefore, in the context of an object recognition task, image complexity could also be a factor of task difficulty—the higher the image complexity, the more difficult the task becomes. The current study complements the results from Seijdel et al. (2021) by testing the ability of deep residual networks to explain the different amounts of recurrent processing in neural data evoked through varying image complexity. We embedded the same target object in backgrounds with varying amounts of complexity and used visual masking to disrupt the occurrence of recurrent processes (Fahrenfort et al., 2007; Lamme, Zipser, & Spekreijse, 2002). We posit that the combination of target objects from natural images on artificial backgrounds gives this experiment a good balance between naturalistic image qualities and experimental control. The expectation was that, all networks, regardless of depth, would explain the same amount of variance in the brain for masked trials (when we disrupt recurrent processing), whereas deeper networks would be able to explain more variance of brain activity for unmasked trials (when visual processing is left unaffected), because of the approximation of recurrent processes in deeper layers. We found that deeper networks indeed explained more variance of brain activity than shallower networks. Furthermore, all ResNets explained more brain activity in unmasked than masked trials with differences between masking conditions starting as early as ~98 msec. In favor of replicability of model performance and variance estimation in ResNets (Mehrer, Spoerer, Kriegeskorte, & Kietzmann, 2020), we trained 10 different seeds of each ResNet model for fitting neural data. These 50 ResNet models are made available at https://osf.io/hcj27/.

We have chosen to test ResNets' ability to capture recurrent processes under the framework of Representational Similarity Analysis (RSA; see Methods section). Hence, in our experiment, ResNets' predictive ability depends on the match of its representations with representations from brain activity. The representations of ResNets are derived from its layers' activations whereas representations of brain activity are derived from time-resolved activity of EEG sensors. We assume that similar representations indicate similar structures of information in both measurements. However, this does not mean that information present is used by the brain for the task at hand (Roskies, 2021; Ritchie, Kaplan, & Klein, 2019). Nonetheless, limiting our search space to similar representations can assist us in locating the functions and processes used by the brain to perform the experimental task.

METHODS

The human participants, stimuli, and experimental paradigm are identical to the ones in Seijdel et al. (2021). However, for the ease of reading, we will also briefly describe them here. The human behavioral data have also been presented in Seijdel et al. (2021). In our Results section, we summarize the human behavioral data and present it against new networks behavioral data.

Human Participants

The experiment in Seijdel et al. (2021) had 62 participants (45 women, 18–35 years old). Its sample size was selected based on Groen et al. (2018) which used a similar paradigm, had a similar research question, and had sufficient signal-to-noise ratio. The sample size is double the sample size of Groen et al. (2018) because the data were split into exploratory and confirmatory sets. Data from one participant were excluded because of the wrong placement of electrodes I1 and I2, and two other participants were excluded because of technical errors causing missing trials (because only one trial was obtained per stimulus, missing trials cannot be used for a RSA, see Analysis: Representational Similarity Analysis section).

Stimuli

For the categorization task, we used a total of 120 unique images from five categories (i.e., 24 unique objects per category). These categories are bird, cat, fire hydrant, frisbee, and suitcase. The images were curated from several online databases—SUN (Xiao, Hays, Ehinger, Oliva, & Torralba, 2010), Microsoft COCO (Lin et al., 2014), Caltech-256 (Griffin, Holub, & Perona, 2007), Open Images V4 (Kuznetsova et al., 2018), and LabelMe (Russell, Torralba, Murphy, & Freeman, 2008). The selected images underwent several preprocessing steps. First, we cropped the images into 512 × 512 pixels and converted them to grayscale. Second, we manually extracted the target objects in the images and then repasted the target objects onto one of four backgrounds. The first type of background is a uniform gray color, referred to as the segmented condition. The second type of backgrounds was generated by phase scrambling the background of the original images (after target object removal). The backgrounds differed in complexity as determined by two values—spatial coherence (SC) and contrast energy (CE; Scholte, Ghebreab, Waldorp, Smeulders, & Lamme, 2009). Forty images with the lowest SC and CE values were labeled as "low complexity" images; 40 images with the highest SC and CE values were labeled as "high complexity"; and the remaining 40 images in the middle range were labeled as "middle complexity" images. A previous study in our laboratory (Groen et al., 2018) has shown the validity of SC and CE values as an index of image complexity. Altogether, with 120 objects embedded in four different backgrounds, we have 480 unique stimuli.

Experimental Design

Participants performed a five-choice categorization task. The trial sequence is illustrated in Figure 1. The complete task consisted of 960 randomized trials (480 unique stimuli presented unmasked and masked), equally divided between visual masking, object category, and background complexity conditions. The trials were grouped into eight blocks of 120 trials with a 1-min break between each block (although participants were allowed a longer break if necessary).

The task was programmed in Presentation (Version 18.0, Neurobehavioral Systems, Inc., www.neurobs.com) and presented on a 23-in. ASUS TFT-LCD display with a spatial resolution of 1920 × 1080 pixels and refresh rate of 60 Hz. Participants were seated approximately 70 cm from the screen. The lights in the room were dimmed and kept constant between participants.

Deep Residual Neural Networks

A family of ResNets was selected: ResNet-4, 6, 10, 18, 34. The numbers in the model names indicate the models' total number of convolution and pooling layers. For each model, 10 different initializations were used as each initialization provides a certain amount of variance in its internal representations (Mehrer et al., 2020). Each of these initialized networks was trained with the ImageNet Large Scale Visual Recognition Challenge 2012 data set and fine-tuned to the five object categories with a separate data set from Microsoft COCO. For the initial training, we used a learning rate of 0.1, with a learning rate decay of 0.1 every 30 epochs. We used a stochastic gradient descent optimizer with a momentum of 0.9. All networks were trained for 150 epochs. For fine-tuning, we replaced the final fully connected layer and retrained the weights for all layers. We used 13,648 training images and 584 test images from five categories for fine-tuning the network. In regard to fine-tuning hyperparameters, we used a learning rate of 0.001, with a learning rate decay of 0.1 every 7 epochs. We also used a stochastic gradient descent optimizer with a momentum of 0.9. By 40 epochs, the fine-tuning validation performance reached a plateau. All ResNets training, fine-tuning, and feature extraction was performed in PyTorch (Paszke et al., 2019).

EEG Data Acquisition and Preprocessing

EEG recordings were made with a Biosemi 64-channel Active Two EEG system (Biosemi Instrumentation, www.biosemi.com) at a sample rate of 1024 Hz. A standard 10–10 electrode placement was used. As we were more interested in visual processing, electrodes F5 and F6 were moved to the occipital region and used as electrodes I1 and I2. Four external electrodes were used to record eye-movement artifacts. Preprocessing was performed in MNE-Python (Gramfort et al., 2013) using the following steps: (1) raw data were rereferenced to the average of left and right electrodes on the mastoids; (2) high-pass (0.1 Hz) and low-pass (30 Hz) filters were applied; (3) independent component analysis (Vigario, Sarela, Jousmiki, Hamalainen, & Oja, 2000) was performed to identify and remove remainder artifact components, specifically eye-movements and eye-blinks; (4) data were segmented into epochs from −100 to 600 msec relative to stimulus onset; (5) baseline correction was applied to the 100 msec before stimulus onset; (6) data were transformed to current source density responses (Perrin, Pernier, Bertrand, & Echallier, 1989) to emphasize local signals as done in previous studies (Seijdel et al., 2021; Groen et al., 2018); (7) multivariate noise normalization was applied (Guggenmos, Sterzer, & Cichy, 2018).

Analysis: Representational Similarity Analysis

We examined both brain activity and internal representations of the ResNets using RSA (Kriegeskorte, Mur, & Bandettini, 2008). RSA transforms both EEG activity and ResNets activations to a common representational space, allowing us to compare both modalities. Essentially, we used RSA to transform the high-dimensional activity space of EEG measurements (of 22 EEG sensors) and

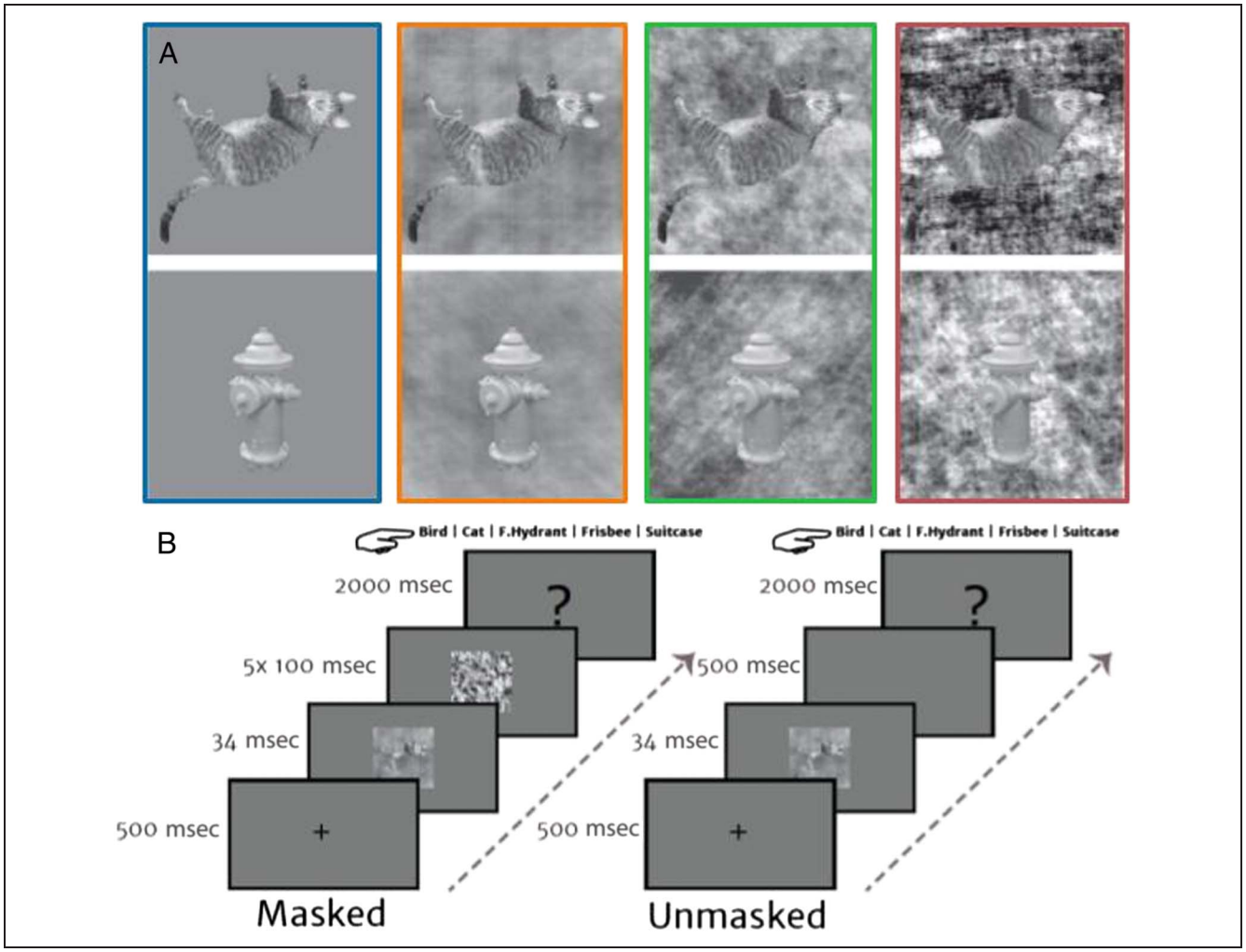

Figure 1. Stimuli and experimental paradigm. (A) Exemplars of two categories (cat and fire hydrant) from each complexity condition. Backgrounds were either uniform (segmented; blue) or had low (orange), medium (green), or high (red) complexity values. (B) Experimental design. On masked trials, the stimulus was followed by a dynamic mask (5 × 100 msec); on unmasked trials, the stimulus was followed by a blank screen (500 msec). Subsequently, participants were asked to categorize the target object. Figure was taken from Seijdel et al. (2021).

DCNN activations (within the range of thousands of units) into a lower-dimensional space represented by patterns of activity in response to our experimental stimuli. Within this lower-dimensional space, we compared the pairwise distances of activity patterns in EEG and DCNNs toward our experimental stimuli. The distances were computed as (1 − *Pearson correlation*) of the pattern responses, and were stored in representational dissimilarity matrices (RDMs). Hence, entries in the RDMs state the distances of activation patterns between all stimuli pairs. We chose the Pearson correlation as a distance metric as we are interested in the relative differences between stimuli instead of the absolute differences. This matters because the distances (and subsequent representations) are limited by the breadth of experimental stimuli. All of our analyses comparing EEG activity and DCNN activations were performed using RDMs. In the analyses involving the RDMs, only the upper triangle is used (excluding the diagonal) as the RDMs are symmetrical. The RDMs are computed separately for EEG measurements and DCNN activations.

From the EEG measurements, we obtained a RDM for every time sample from −100 msec to 600 msec relative to stimulus onset. With 180 EEG time samples per trial, this amounts to 180 RDMs. The RDMs were computed based on activity from 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). The electrodes were selected based on a previous study on the relationship between recurrent processing and image complexity (Groen et al., 2018). The electrodes selection is meant to emphasize visual processing.

From the DCNNs, we obtained RDMs from all convolutional, pooling, and fully connected layers. For example, with ResNet-4, there would be three convolutional layers, one pooling layer, and one fully connected layer. In total, there would be five layers, creating five RDMs. With the exception of fully connected layers, the DCNNs RDMs were computed based on 100 principal components of each layer's activations. The fully connected layer is exempted as its dimension is <100. We chose to use

principal components instead of raw activations as the number of units varied widely between layers; limiting our analysis to 100 components allowed us to constrain the influence of different layers to be equivalent and focus on the effects of depth. Storrs, Kietzmann, Walther, Mehrer, and Kriegeskorte (2021) had similarly used a PCA transformation to prevent the model from overfitting. Similar to the analysis in Storrs et al. (2021), we transformed the activations using 100 principal components obtained from a separate data set (n = 2986) of natural images to prevent overfitting of PCA components to our experimental data set. The natural images were selected from Microsoft COCO and retain similar image statistics as our experimental stimuli. We chose to use 100 PCA components because Storrs et al. (2021) reported the number to be a good balance between variance preserved and availability of computing ability.

The main analysis involved predicting EEG RDMs using DCNNs RDMs using nonnegative least squares (Kaniuth & Hebart, 2021; Khaligh-Razavi & Kriegeskorte, 2014). In this regression, DCNNs/ResNets RDMs were the predictors (X), and EEG RDMs were the targets (y). As there are 180 EEG RDMs per trial, the regression for each ResNet model was repeated for each time sample, thus giving us a time-resolved view of ResNets' ability to capture processing of brain activity.

The EEG and DCNNs RDMs used in our regression models were computed by trials averaged within the same object categories to increase the signal-to-noise ratio and place a larger focus on object categorization. After averaging trials within object categories, we obtained RDMs of the shape 20 × 20–5 object categories × 4 background conditions. Both the EEG and DCNNs RDMs have the same 20 × 20 shape.

The regression between each ResNet model and EEG is performed with a different number of ResNet layers (see Figure 2). For example, with ResNet-4 (of five layers), we would build five different regression models. For the first model, we regressed only the first layer onto the EEG RDMs; for the second model, we regressed the first and second layers; this process continued until we included all layers in ResNet-4. The regression models are cross-validated 50 times. In every resampling, we reserved 15 (randomly chosen) test participants and 240 (randomly chosen) test stimuli, while fitting the regression model on 44 participants and 240 train stimuli. For each model, we computed a R^2 value based on the test participants and test stimuli. We also computed the upper and lower bounds of the noise ceiling by taking the averaged correlation of each test participant's RDM with the RDM averaged across all test participants (upper bound), and taking the averaged correlation of each test participant's RDM with the RDM averaged across all train participants (lower bound). Subsequently, we squared the averaged correlation values to obtain the R^2 values for the upper and lower bounds. We determined the unique R^2 of each ResNet layer based on the increase in R^2 value when that layer was included in the model.

Statistical Analysis: Behavioral Data

For all ResNets (n = 50) and human participants (n = 59), categorization accuracy was computed as the average proportion of correct trials within each condition. Differences between conditions for ResNets were tested with a two-factor ANOVA (Network Type and Background Complexity), followed by t tests between the condition pairs. The p values obtained from the ANOVA were corrected for multiple comparisons with a false discovery rate (FDR; α = .01). Significance was determined based on a p value that was less than .01. Behavioral analysis was performed and visualized in Python using the following

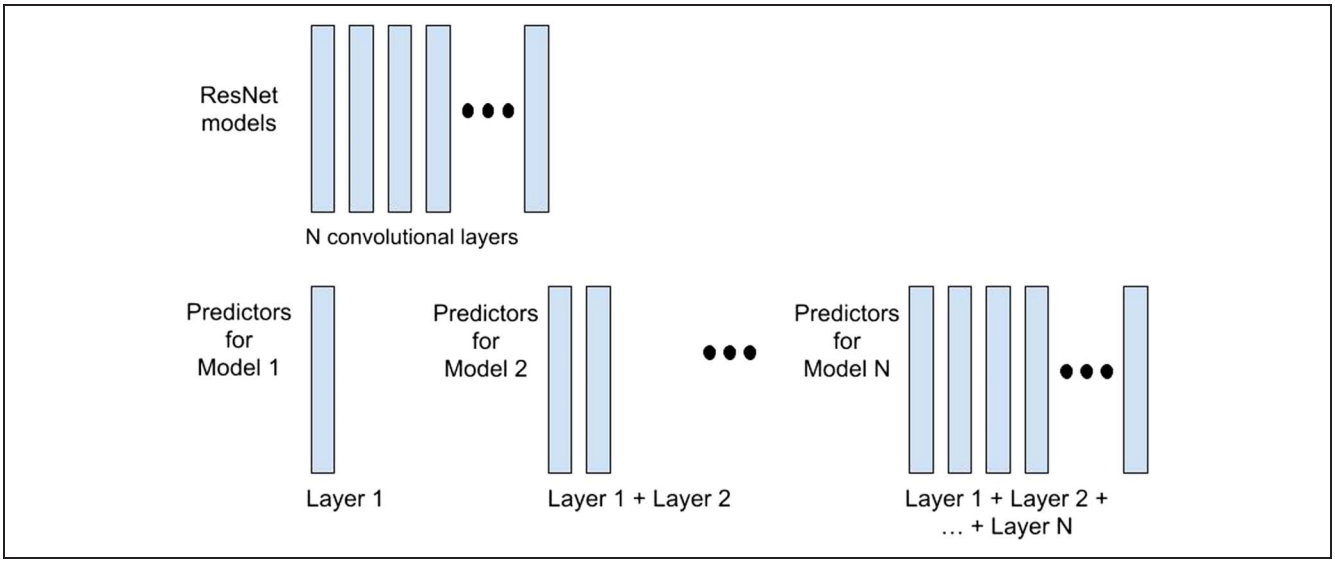

Figure 2. Regression models with varying number of layers as predictors. To observe the contributions of additional layers in predicting neural data, we fitted regression models with an incremental number of layers included as predictors.

packages: NumPy, SciPy, Statsmodels, Pandas, Seaborn (Waskom, 2021; Harris et al., 2020; Virtanen et al., 2020; McKinney, 2010; Seabold & Perktold, 2010).

Statistical Analysis: ResNets' Explained Variance on Brain Activity

We used a Mann–Whitney U test to test for pairwise differences in R^2 between ResNets. The p values obtained from the Mann–Whitney U test were corrected for multiple comparisons with an FDR ($\alpha = .01$). Similarly, the Mann–Whitney U test was used to test for differences in R^2 between unmasked and masked trials for each ResNet model. Significance was determined based on a p value less than .01.

RESULTS

We investigated the ability of DCNNs to capture recurrent processing in the human brain within an object categorization task. Human participants performed the task under both visually unmasked and masked conditions. ResNets performed the recognition task with identical stimuli. We compared both the object categorization performance of human participants with the categorization performance of ResNets, and also brain activity from human participants with unit activations from ResNets.

Visual Masking Changes Human Object Recognition Performance from One Alike a Deeper Network into a Shallower One

Human performance under visually unmasked conditions was close to performance ceiling (i.e., 100% accuracy) regardless of object background complexity (see Figure 3A). However, under visually masked conditions, human performance deteriorated with increasing background complexity. Results from repeated-measures ANOVA revealed that both factors of background complexity and masking interacted—specifically, masking impaired performance to a greater degree for more complex backgrounds (Seijdel et al., 2021).

For ResNets (see Figure 3B), the deepest network (i.e., ResNet-34) performed close to ceiling for all background complexity conditions; whereas shallower networks such as ResNet-10 suffered in performance as background complexity increased. The two most shallow networks—ResNet-4 and 6—performed poorly regardless of background complexity. A two-factor ANOVA was performed with Network Type (i.e., number of layers) and Background Complexity as independent factors; its results showed significant main effects of Network Type, $F(4, 180) = 3867.61$, $p < .001$; significant main effects of Background Complexity conditions, $F(3, 180) = 15.93$, $p < .001$; and also significant interaction effects between Network Type and Background Complexity conditions, $F(12, 180) = 7.60$, $p < .001$.

To examine if differences of each network's performance across conditions were significant, pairwise comparisons t tests were performed. For ResNet-4, significant differences were reported between segmented and medium complexity conditions, $t(9) = 5.01$, p(FDR-corrected) = .002; low and medium complexity conditions, $t(9) = 6.00$, p(FDR-corrected) = .001; between medium and high complexity conditions, $t(9) = -4.19$, p(FDR-corrected) = .005; but no significance were reported between segmented and low complexity conditions, $t(9) = -.49$, p(FDR-corrected) = .64; nor between segmented and high complexity conditions, $t(9) = 1.17$, p(FDR-corrected) = .33; nor between low and high complexity conditions, $t(9) = 2.18$, p(FDR-corrected) = .09.

For ResNet-6, significant differences were reported between segmented and low complexity conditions, $t(9) = -4.83$, p(FDR-corrected) = .005; but no significant differences were reported between segmented and medium complexity conditions, $t(9) = -3.16$, p(FDR-corrected) = .03; between segmented and high complexity conditions, $t(9) = -0.88$, p(FDR-corrected) = .46; between low and medium complexity conditions, $t(9) = -0.76$, p(FDR-corrected) = .46; between low and high complexity conditions, $t(9) = 1.30$, p(FDR-corrected) = .34; nor between medium and high complexity conditions, $t(9) = 2.41$, p(FDR-corrected) = .08.

For ResNet-10, there were significant differences between low and medium complexity conditions, $t(9) = 9.39$, p(FDR-corrected) < .001, and between low and high complexity conditions, $t(9) = 7.61$, p(FDR-corrected) < .001. However, no other significant differences were found—between segmented and low complexity conditions, $t(9) = -3.08$, p(FDR-corrected) = .02; between segmented and medium complexity conditions, $t(9) = 2.04$, p(FDR-corrected) = .07; between segmented and high complexity conditions, $t(9) = 3.29$, p(FDR-corrected) = .02; nor between medium and high complexity conditions, $t(9) = 2.48$, p(FDR-corrected) = .04.

For ResNet-18, no significant differences were found between all complexity conditions—between segmented and low complexity conditions, $t(9) = -1.06$, p(FDR-corrected) = .38; between segmented and medium complexity conditions, $t(9) = .61$, p(FDR-corrected) = .56; between segmented and high complexity conditions, $t(9) = 1.27$, p(FDR-corrected) = .38; between low and medium complexity conditions, $t(9) = 2.83$, p(FDR-corrected) = .06; between low and high complexity conditions, $t(9) = 3.71$, p(FDR-corrected) = .03; and between medium and high complexity conditions, $t(9) = 1.22$, p(FDR-corrected) = .38.

For ResNet-34, no significance were reported between all complexity conditions—between segmented and low complexity conditions, $t(9) = 1.66$, p(FDR-corrected) = .44; between segmented and medium complexity conditions, $t(9) = 1.21$, p(FDR-corrected) = .44; between segmented and high complexity conditions, $t(9) = 1.12$, p(FDR-corrected) = .44; between low and medium complexity conditions, $t(9) = -0.55$, p(FDR-corrected) = .61;

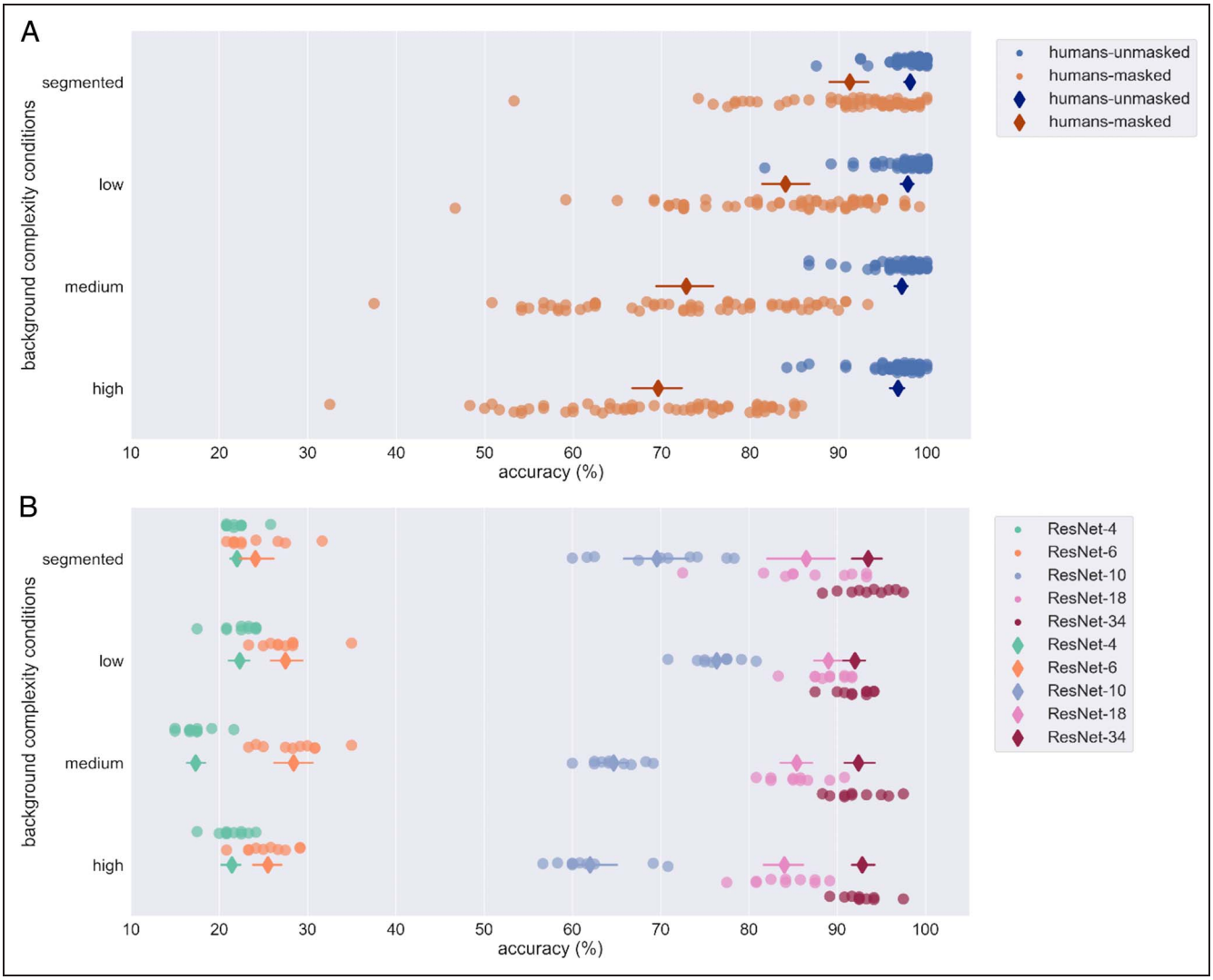

Figure 3. Human and ResNets' performance on the object categorization task. Both humans and ResNets performed an object categorization task (chance performance: 20%). The different rows indicate performance in different background complexity conditions. The diamond-shaped marker and line indicate average score and standard error. (A) Human performance was consistent across background complexity conditions within unmasked trials but differed across masked trials. (B) The deepest network (i.e., ResNet-34) mimics human performance in unmasked trials—its performance did not vary with background complexity. However, a shallower network such as ResNet-10 mimics human performance in masked trials—its performance deteriorated with medium and high complexity conditions.

between low and high complexity conditions, $t(9) = -1.13$, p(FDR-corrected) = .44; and between medium and high complexity conditions, $t(9) = -0.52$, p(FDR-corrected) = .61.

In general, each ResNet model performed consistently across different complexity conditions with the exception of ResNet-10, which performed poorer when background complexity increased from low to medium complexity.

In other words, human participants performed optimally under visually unmasked conditions, much like a deep network (i.e., ResNet-34). However, when stimuli were masked, human participants performed more like a shallower network (i.e., ResNet-10). Here, we can conclude that visual masking changes human object recognition performance from one alike deep networks to one alike shallower networks.

Deeper ResNets Explained More Variance in Brain Activity Compared with Shallower ResNets

All analyses of brain activity and ResNets activity are performed using RSA (see Methods section). With RSA, we obtained stimuli pairwise dissimilarity matrix (known as RDM) as it reveals the dissimilarity distance between stimuli condition pairs. RDMs were generated based on EEG activity and network layer activations.

We tested the ability of ResNets to capture brain activity by fitting regression models for each network. In these regression models, we used ResNets RDMs as predictors for EEG RDMs. These models were fitted separately for unmasked and masked trials as we are interested to observe the difference in explained variance (R^2) when recurrent processing is undisrupted (in unmasked trials) and disrupted (in masked

trials). The plotted R^2 for each ResNet is averaged across 10 initializations of the network (see Figure 4).

To assess if ResNet models performed differently in predicting brain activity, pairwise comparisons between ResNets were performed on samples at ~90–250 msec. Within unmasked trials, pairwise comparisons revealed no significant differences between the R^2 of ResNet-4 and ResNet-6. However, significant differences were found between the R^2 of ResNet-4 and ResNet-10 at ~94–110 msec and ~137–246 msec (α = .01, FDR corrected for 420

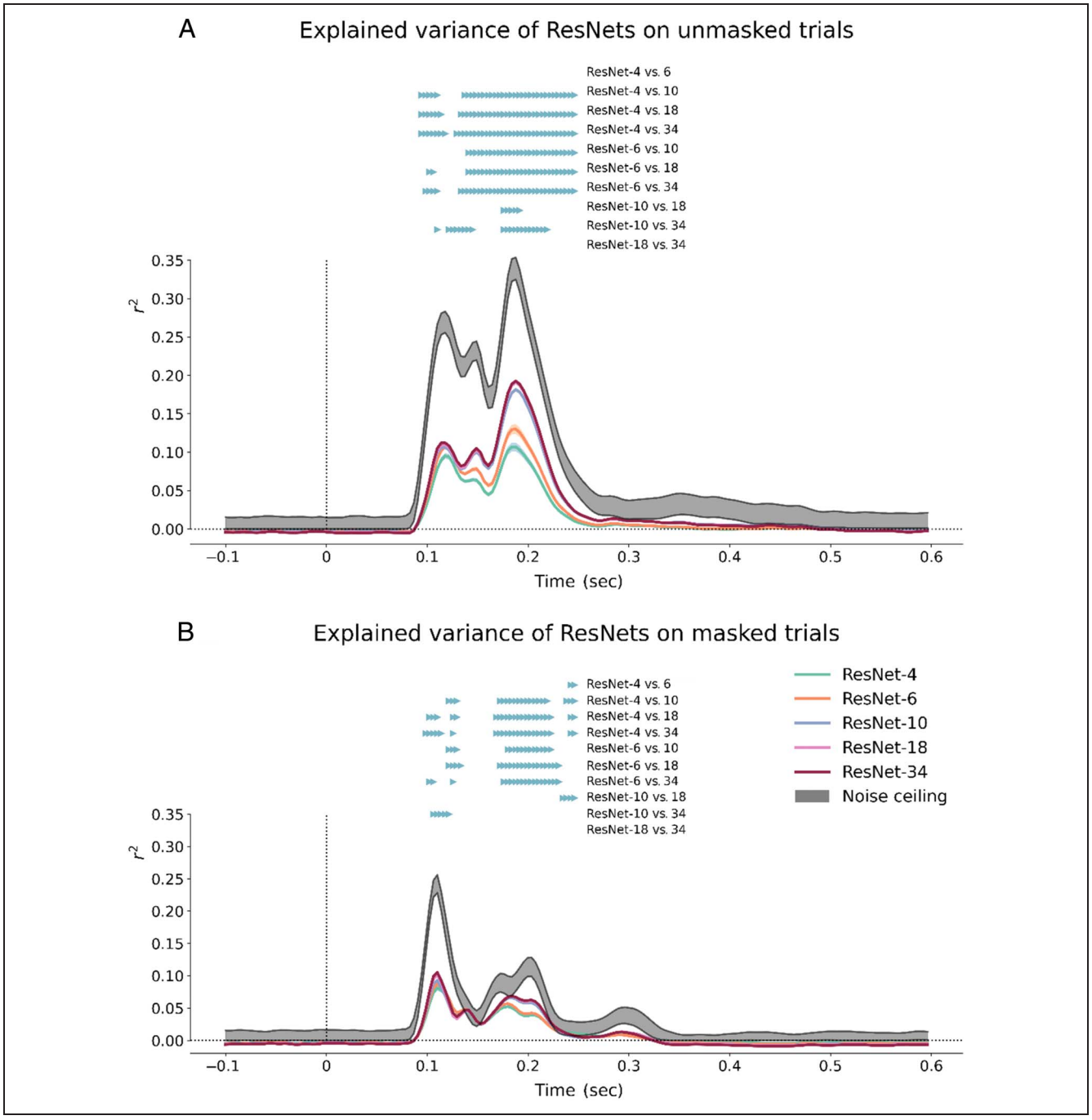

Figure 4. Regression models using ResNets' layers as predictors for unmasked (A) and masked (B) trials. Pairwise comparisons between the ResNets revealed that deeper networks have higher R^2 than shallower networks. (A) Within unmasked trials, results showed that the R^2 of ResNets are higher for deeper networks than shallower ones. The R^2 of ResNet-10, 18, and 34 is higher compared with the R^2 of ResNet-4 and 6 for all time points 141–250 msec (±12 msec). In addition, the R^2 of ResNet-18 and 34 is also higher compared with the R^2 of ResNet-10 between ~176 and 219 msec. However, no differences were found between ResNet-18 and 34. (B) Within masked trials, results similarly showed that the R^2 of deeper networks are higher than shallower networks, although the differences between networks have decreased, compared with unmasked trials. Specifically, the R^2 of ResNet-10, 18, and 34 is higher than the R^2 of ResNet-4 and 6 for all time points 173–225 msec (±11 msec). The R^2 of ResNet-34 is also higher compared with the R^2 of ResNet-10 at ~106–122 msec. However, the R^2 of ResNet-18 and ResNet-34 did not significantly differ from each other. For masked trials, all networks' R^2 did not differ significantly from each other at ~133–168 msec.

pairwise comparisons – 10 model pairs × 42 time samples); between the R^2 of ResNet-4 and ResNet-18 at ~94–114 msec and ~133–246 msec; and between the R^2 of ResNet-4 and ResNet-34 at ~94–118 msec and ~129–246 msec. Pairwise comparisons between the R^2 of ResNet-6 and ResNet-10 revealed significant differences at ~141–246 msec; between the R^2 of ResNet-6 and ResNet-18 at ~102–106 msec and ~141–246 msec; and between the R^2 of ResNet-6 and ResNet-34 at ~98–110 msec and ~133–246 msec. Pairwise comparisons between the R^2 of ResNet-10 and ResNet-18 revealed significant differences at ~176–192 msec; and between the R^2 of ResNet-10 and ResNet-34 at ~110 msec, ~122–145 msec, and ~176–219 msec. Pairwise differences between the R^2 of ResNet-18 and ResNet-34 revealed nonsignificant differences. All statistical significance were determined with $\alpha = .01$ and FDR corrected.

In summary, the R^2 of ResNet-4 and ResNet-6 did not significantly differ, and both are significantly lower than the R^2 of ResNet-10, 18, and 34. The R^2 of ResNet-10 significantly differed from the R^2 of ResNet18 at ~176–192 msec. The R^2 of ResNet-10 also significantly differed from the R^2 of ResNet-34 at ~110 msec, ~122–145 msec, and

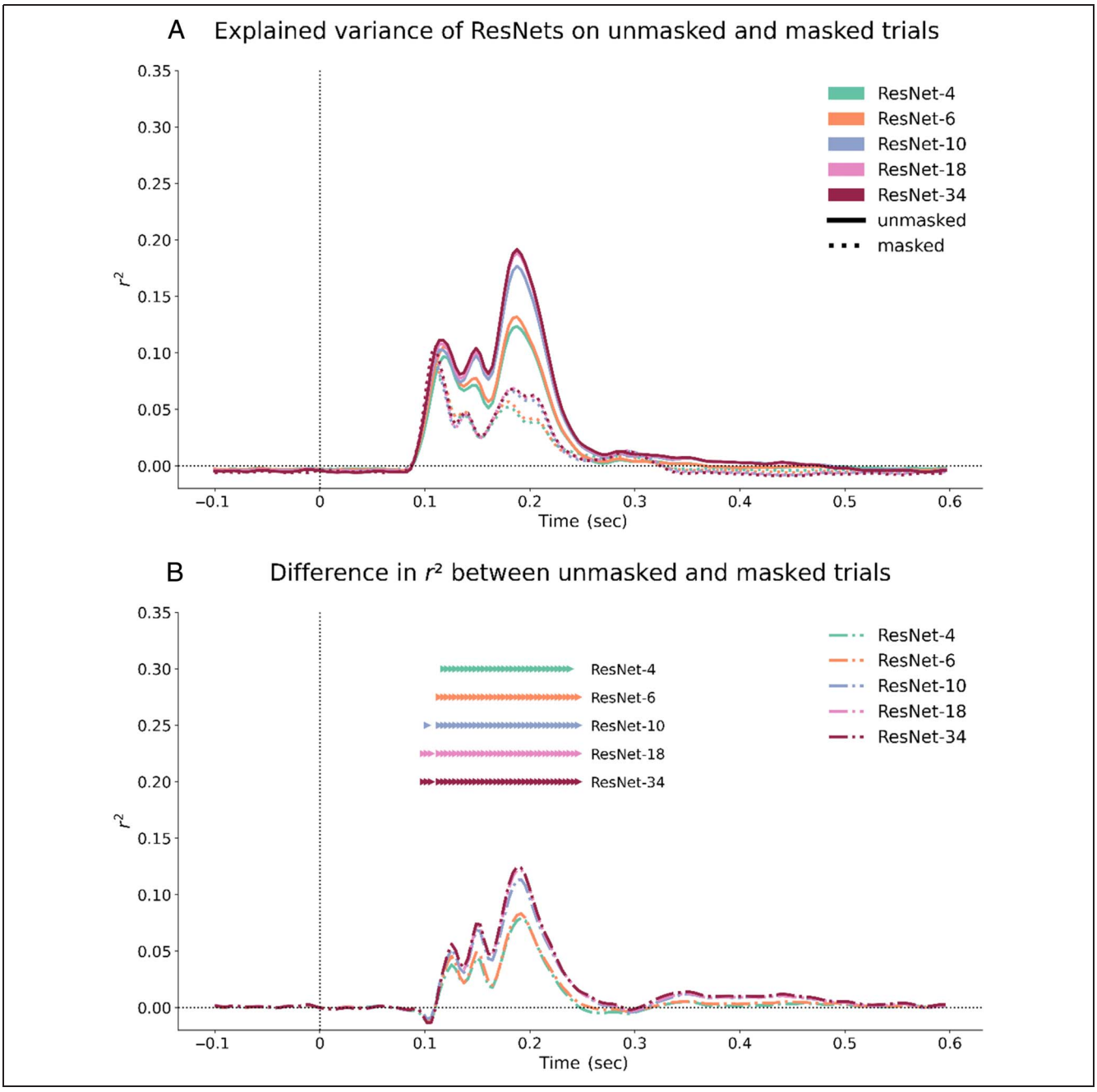

Figure 5. R^2 of models for unmasked and masked data and differences in models performance between masking conditions. (A) The R^2 for both unmasked and masked trials are plotted together. (B) The difference in R^2 between masking conditions is plotted. Colored markers above reflect significant differences between masking conditions for each network. Here, we see that the magnitude of differences are larger for deeper networks (ResNet-10, 18, and 34) compared with shallower networks (ResNet-4 and 6).

~176–219 msec. The R^2 of ResNet-18 and ResNet-34 did not significantly differ from each other. We observed that deeper networks have higher R^2 compared with shallower networks, indicating an increased ability to predict EEG data.

Subsequently, we fitted the models to masked trials. We similarly performed pairwise comparisons between the R^2 of ResNet models on samples at ~90–250 msec. Pairwise comparisons between the R^2 of ResNet-4 and ResNet-6 revealed significant differences at ~242–246 msec; between the R^2 of ResNet-4 and ResNet-10 at ~122–129 msec, ~172–219 msec, and ~238–246 msec; between the R^2 of ResNet-4 and ResNet-18 at ~102–110 msec, ~126–129 msec, ~168–223 msec, and ~242–246 msec; and between the R^2 of ResNet-4 and ResNet-34 at ~98–114 msec, ~126 msec, ~168–223 msec, and ~242–246 msec. Pairwise comparisons between the R^2 of ResNet-6 with ResNet-10 revealed significant differences at ~122–129 msec, and ~180–223 msec; between the R^2 of ResNet-6 with ResNet-18 at ~122–133 msec, and ~172–231 msec; and between the R^2 of ResNet-6 with ResNet-34 at ~102–106 msec, ~126 msec, and ~176–231 msec. Pairwise comparisons between the R^2 of ResNet-10 with ResNet-18 revealed significant differences at ~234–246 msec; between the R^2 of ResNet-10 with ResNet-34 at ~106–122 msec. Pairwise comparisons between the R^2 of ResNet-18 with ResNet-34 revealed no significant differences for all time samples. All statistical differences are determined with $\alpha = .01$ and FDR corrected.

In summary, the R^2 of ResNet-4 and ResNet-6 for masked trials showed significant differences for the time window at ~242–246 msec. However, the R^2 of ResNet-10, 18, and 34 is higher than the R^2 of ResNet-4 and 6 at ~173–225 msec (±11 msec). Between the R^2 of ResNet-10 and ResNet-18, significant differences were found at ~234–246 msec. Comparisons between R^2 of ResNet-10 and ResNet-34 showed significant differences at ~106–122 msec. Comparisons between R^2 of ResNet-18 and ResNet-34 showed no significant differences. Similar to unmasked trials, we also observed that deeper networks have higher R^2 compared with shallower networks; however, the magnitude and duration of significant differences between both deeper and shallower networks have shrunk in masked trials compared with unmasked trials. In addition, within masked trials, the R^2 of all ResNets did not significantly differ from each other at ~133–168 msec.

To compare between the explained variance in unmasked and masked trials, we performed pairwise comparisons between the models' R^2 for unmasked trials and R^2 for masked trials on time samples ~90–250 msec. For ResNet-4, significant differences were found—~118–238 msec. For ResNet-6, significant differences were found at ~113–246 msec. For ResNet-10, significant differences were found at ~102 msec, and ~114–246 msec. For ResNet-18, significant differences were found at ~98–106 msec, and ~114–246 msec. For ResNet-34, significant differences were found at ~98–106 msec, and ~114–246 msec. All statistical significance were determined with $\alpha = .01$ and FDR corrected.

In short, significant differences in R^2 between masking conditions were found as early as ~98 msec for ResNet-18 and 34, ~102 msec for ResNet-10, ~113 msec for ResNet-6, and ~118 msec for ResNet-4. Here, deeper networks also show earlier significant masking differences compared with shallower networks. There were no effects of masking nor network depth before ~98 msec, indicating EEG activity reflected purely feedforward processes.

In addition, we computed the difference between the R^2 for unmasked trials and masked trials (see Figure 5B)

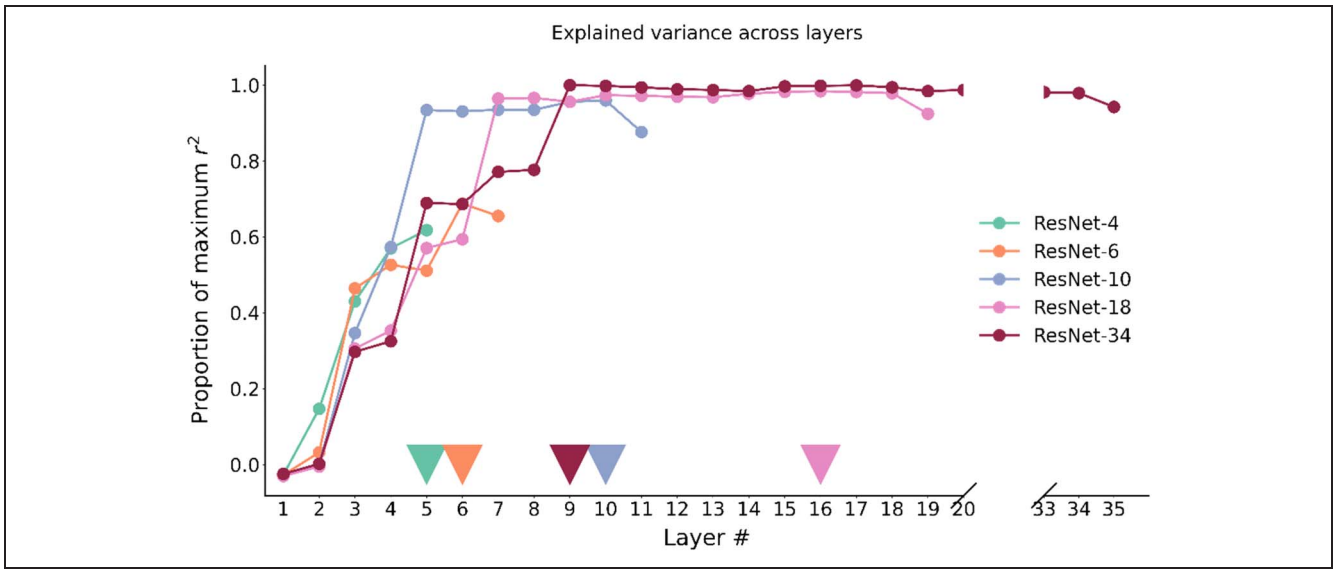

Figure 6. Models' R^2 plotted as a proportion of the maximum R^2. We fitted regression models with an increasing number of convolutional layers as predictors. With each additional layer, models increase its R^2. However, the increase in R^2 stopped at a certain number of layers. The colored triangles indicate the layer where R^2 is maximum for the model. For ResNet-10 and 34, the maximal R^2 was reached at its early layers (i.e., first half layers of the network). Notably, including the fully connected (i.e., classification) layer to the model does not improve its predictions.

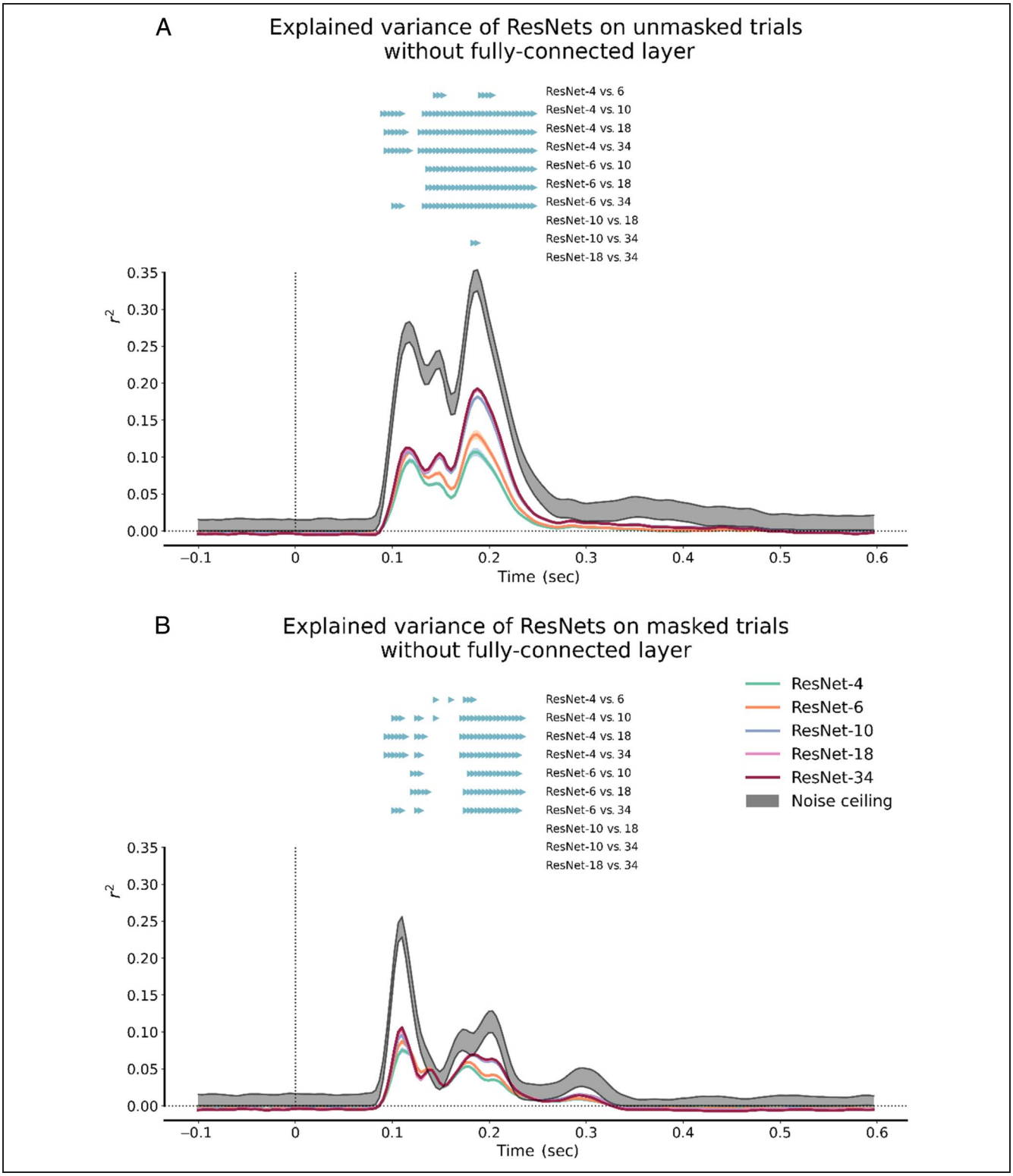

Figure 7. Regression models using ResNets' layers as predictors, excluding fully connected layers, for unmasked (A) and masked (B) trials. Without the fully connected layer, we see a larger difference between ResNet-4 and ResNet-6, but smaller differences between ResNet-10 and ResNet-18, 34. However, our conclusion still holds—deeper models still explain more variance in the brain than shallower models. Furthermore, the differences in R^2 between deep and shallow models also decreased for masked trials, supporting our conclusion that both deep and shallow models perform more similarly because of a reduction of recurrence with visual-masking.

and observed that the difference between masking conditions gradually increased to a peak at ~200 msec. The magnitude of differences between masking conditions are also larger for deeper networks (ResNet-10, 18, and 34) than shallower networks (ResNet-4 and 6), indicating that deeper networks are better at capturing recurrent processing signals at later time points.

Early DCNN Layers Are Sufficient to Explain Brain Activity

To understand the contribution of ResNets' depth in capturing brain activity, we built a series of regression models differentiated by the number of layers used as predictors (see Figure 2). As more layers are included as predictors, the model increases in R^2 (see Figure 6). However, at a certain number of layers, including more layers no longer increases the R^2 of the model. ResNet-4 reached maximal R^2 at layer 5. ResNet-6 reached maximal R^2 at layer 6. ResNet-10 reached maximal R^2 at layer 10. ResNet-18 reached maximal R^2 at layer 16. ResNet-34 reached maximal R^2 at layer 9. Surprisingly, ResNet-10 and 34 reached maximal R^2 in its early layers (i.e., first half layers of the network). Layers beyond these layers of maximal R^2 do not improve the model performance (i.e., amount of R^2).

In fact, inclusion of the fully connected layer decreases the explained variance for ResNet-6, 10, 18, and 34, revealing that the models have overfitted on the training data set. Interestingly, this overfitting phenomenon did not apply for ResNet-4, nor for deeper convolutional layers for other networks. For the sake of transparency and clarity, we also show the R^2 of models without the fully connected layer (see Figure 7). Without the fully connected layer, we see a larger difference between ResNet-4 and 6, but a much smaller difference between ResNet-10 and ResNet-18, 34. Nonetheless, both results still support our conclusion that deeper networks explain more variance in the brain than shallower networks.

To further assess the networks' ability to predict brain activity, we computed the unique R^2 from each network layer to observe each layer's contribution to the model performance. The unique R^2 of a layer is computed as the increase in R^2 of the model with the additional layer as a predictor (see equation below).

$$\textit{unique } R^2 \textit{ at layer } n = \textit{model's } R^2 \textit{ at layer } n - \textit{model's } R^2 \textit{ at layer } n-1 \quad (1)$$

We observed that the early layers of deeper networks contributed disproportionately to the models' R^2 (see Figure 8). Layers 3, 5, and 7 contributed most to deeper networks' (ResNet-10, 18, and 34) R^2. Paradoxically, although deeper networks have significantly higher R^2 than shallower networks, deeper networks' performance can be attributed to activity in its early layers; deeper layers do not further contribute to the models' R^2.

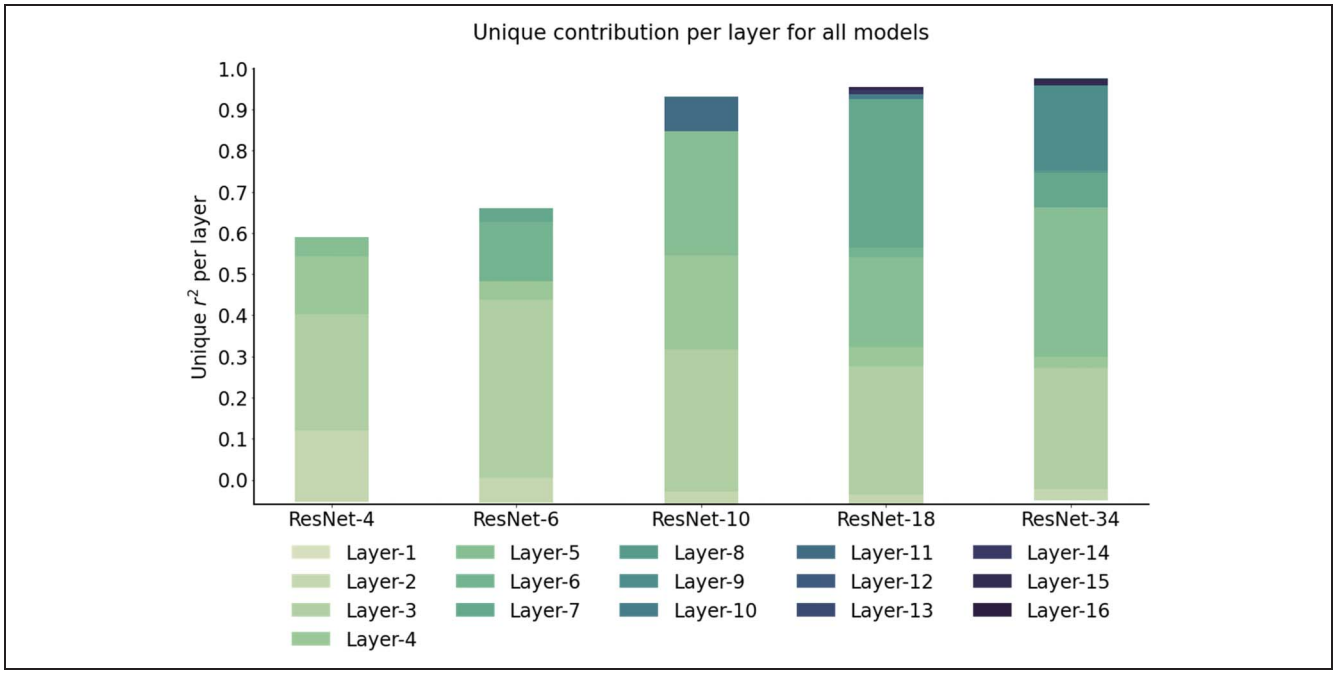

Figure 8. Unique R^2 of each ResNet layer. We computed the unique variance for each layer in the regression model. For shallower networks (ResNet-4 and 6), Layers 2, 3, and 6 contributed most to the model. For deeper networks (ResNet-10, 18, and 34), Layers 3, 5, and 7 contributed at least half of the models' R^2. Layer 1 has negative R^2 values that happen when the model is arbitrarily worse than chance. This is also the reason why the plot begins below zero. Layers beyond 16 are not plotted as there are no further increases of R^2 values.

DISCUSSION

Summary

In this study, we investigated the ability of DCNNs to capture recurrent processing in the human brain. Specifically, we tested ResNets as they approximate an additive form of recurrent processing consisting of repeating excitatory activations on static inputs. We used ResNets of varying depths as proxies for varying amounts of recurrent processing. We expected deeper networks (i.e., networks with more recurrent processing) to explain more variance in brain activity (i.e., have higher R^2) than shallower neworks in unmasked trials when recurrent processing was not disrupted, but for all networks to have similar R^2 in masked trials when recurrent processing was disrupted. Our expectations were partially met as deeper networks (ResNet-10, 18, and 34) indeed have higher R^2 than shallower networks (ResNet-4 and 6), but for both unmasked and masked trials. We also find that all ResNet models (ResNet-4, 6, 10, 18, 34) have higher R^2 under unmasked trials than masked trials, with differences in R^2 starting as early as ~98 msec. These differences in R^2 gradually increased to a peak at ~200 msec, with deeper networks (ResNet-10, 18, 34) showing larger magnitudes of differences compared with shallower networks (ResNet-4 and 6). By building regression models with increasing numbers of layers as predictors, we see that only early layers (i.e., first half layers of the model) contributed to the R^2 of the model.

Deeper Networks Capture Behavioral Performance But Not Recurrent Processes in Early Visual Cortex

We found that humans performed similarly as a deep network (i.e., ResNet-34) under visually unmasked conditions, but visual masking deteriorated human performance to become more like a shallower network (i.e., ResNet-10). Based on categorization performance, we had expected a deep model like ResNet-34 to have higher R^2 as compared with a shallower model like ResNet-10. This expectation stems from previous studies—where DCNNs that perform better in categorization accuracy also better predict brain data (Yamins et al., 2014). However, although ResNet-34 significantly outperforms ResNet-10 at categorizing objects, the R^2 ResNets-34 only significantly differed from the R^2 of ResNet-10 for a short time window at ~102–126 msec and ~184 msec. Nonetheless, our finding agrees that model depth improves the model's object categorization performance—similar to the findings of the original ResNet creators (He et al., 2016). However, the mechanisms in ResNets' deeper layers do not seem to match the underlying mechanisms in humans' early visual cortex as measured with EEG. This can be observed in Figure 6 where inclusion of ResNets' deeper layers in the regression model does not improve its predictions. We speculate that this is caused by the fact that EEG signals reflect only part of the activity in the early visual cortex, namely, the activity at the cortical gyri (Nunez & Srinivasan, 2009). Furthermore, a discrepancy between behavioral outcomes and brain activity prediction can also be observed on the Brain-Score platform, where we see that the unmodified version of ResNets tend to score well on explaining behavior but score less well on explaining activity in V1 (Schrimpf et al., 2018).

Simple Recurrence Captures Recurrent Processing But Still Insufficient

ResNets are made up of residual blocks. Each residual block has multiple convolutional layers with the same filter size and same number of filters. Thus, the repetition of identical convolutional operations could be perceived as a simple, additive form of recurrence—recurrence by repeating excitatory activations in response to a static stimulus, similar to recurrence in current RNNs and DCNNs like CorNet. In this experiment, we had approximated this additive form of recurrence as we did not tether the weights of the convolutional operations. Nonetheless, even our approximation through network depth showed differences. However, this form of recurrence only works to a certain extent. With deeper networks, our results revealed that the R^2 of ResNets-10, 18, and 34 only significantly differed on a few time points or none at all (between ResNet-18 and 34), whereas there was still a large gap between ResNet-34's R^2 with the noise ceiling. Thus, it appears that although additive recurrence captures recurrent processing partially, by itself, it is insufficient to fully capture recurrent processing in humans. This finding is supported by studies in animals demonstrating the importance of inhibition in feedback processes (Klink, Dagnino, Gariel-Mathis, & Roelfsema, 2017), and also recent studies modeling recurrent signals, where researchers have shown that both lateral (i.e., repeating excitatory activations) and feedback (i.e., top–down) connections are necessary to improve the model's performance beyond feedforward processes (Kietzmann, Spoerer, et al., 2019; Spoerer et al., 2017). An alternative consideration would be a more complex form of recurrence, which can be found in predictive hierarchical models (Friston, 2010) where top–down recurrent signals are able to explain away variance in processes in lower levels, leading to reduced activity in lower levels. Last but not the least, although an improvement in predictions suggest a match in representations between a ResNet model and EEG activity, we do not believe that a ResNet-like architecture underlies brain connectivity. Other architectures could similarly give rise to similar representations (Diedrichsen & Kriegeskorte 2017; Kriegeskorte et al., 2008)—the architecture is merely one form of constraints on the function. Nonetheless, matches in representations are important criteria for locating functions that could approximate processes in the brain.

Recurrence Disrupted But Still Present within Masked Trials

Earlier, we had hypothesized no differences between the R^2 of all ResNets for masked trials. However, we observed that differences are still present, especially between ResNet-4, 6 and ResNet-10, 18, 34 (see Figure 4). We presume that visual-masking disrupts or reduces the occurrence of recurrent processes, but does not completely eliminate the occurrence of recurrent processes. Presumably, these traces of recurrence could explain the differences of R^2 between ResNets models within masked trials, although the magnitude of differences have become much smaller, supporting the presumption of reduction in recurrent processing.

Recurrent Signals Set in as Early as ~98 msec

When we observed the difference in R^2 between masking conditions, we see significant differences as early as ~98 msec for ResNet-18 and 34, ~102 msec for ResNet-10, ~113 msec for ResNet-6, and ~118 msec for ResNet-4—with deeper networks showing earlier significant differences compared with shallower networks. These early differences indicate that recurrent signals set in as early as ~98 msec. Before ~98 msec, there were no effects of masking nor effects of network depth, suggesting that EEG activity before ~98 msec are feedforward processes. The differences in R^2 between masking conditions gradually increased until a peak difference at ~200 msec, indicating that recurrent signals build up across time. As such, early EEG activity is a mixture of feedforward and recurrent processes, whereas late EEG activity is mainly dominated by recurrent processes. Consequently, our results suggest that models including interactions between feedforward and feedback streams across time steps could better capture recurrent processes in humans.

Conclusion

In this article, we tested the ability of DCNNs to capture recurrent processes in human brains. Specifically, we tested an additive form of recurrence in ResNets to predict recurrent signals in human visual systems. We found that deeper ResNets explained more variance in brain activity than shallower ResNets. Furthermore, all ResNets explained more variance in brain activity during unmasked trials than masked trials. Differences in explained variance between masking conditions set in as early as ~98 msec and gradually increased to a peak at ~200 msec, indicating that early brain activity consists of both feedforward and recurrent processes but gradually becomes dominated by recurrent processes. Accordingly, deeper networks showed larger differences in explained variance between masking conditions than shallower networks, providing further evidence that deeper networks capture larger proportions of recurrent processing signals. However, given the substantial distance between the models' explained variance and data's noise ceiling, we posit that other types of recurrent processes (inhibition, multiplicative), which are not present in current regular deep neural networks (alexnet, cornet, resnet), are of paramount importance toward better visual models.

Acknowledgments

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 to Edward de Haan (#339374). The authors also express their gratitude to Kate Storrs and Martin Hebart for guidance on parts of the analysis.

Reprint request should be sent to Jessica Loke, Psychology Department - Brain & Cognition University of Amsterdam, Nieuwe Achtergracht 129b, 1018 XE Amsterdam, The Netherlands, or via email: jessica_2020@hotmail.com.

Data Availability Statement

The data, experimental stimuli, and trained models are made available at https://osf.io/hcj27/.

Author Contributions

Jessica Loke: Conceptualization; Data curation; Formal analysis; Visualization; Writing–Original draft; Writing–Review & editing. Noor Seijdel: Conceptualization; Data curation; Formal analysis; Visualization; Writing–Original draft; Writing–Review & editing. Lukas Snoek: Formal analysis; Writing–Review & editing. Matthew van der Meer: Conceptualization; Data curation; Writing–Review & editing. Ron van de Klundert: Conceptualization; Data curation. Eva Quispel: Conceptualization; Data curation. Natalie Cappaert: Supervision; Visualization; Writing–Review & editing. H. Steven Scholte: Conceptualization; Formal analysis; Supervision; Visualization; Writing–Original draft; Writing–Review & editing.

Diversity in Citation Practices

Retrospective analysis of the citations in every article published in this journal from 2010 to 2021 reveals 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 = .407, W(oman)/M = .32, M/W = .115, and W/W = .159, the comparable proportions for the articles that these authorship teams cited were M/M = .549, W/M = .257, M/W = .109, and W/W = .085 (Postle and Fulvio, *JoCN*, 34:1, pp. 1–3). 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 article report its proportions of citations by gender category to be as follows: M/M = .865; W/M = .108; M/W = .027; W/W = 0.

REFERENCES

Diedrichsen, J., & Kriegeskorte, N. (2017). Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLoS Computational Biology*, *13*, e1005508. https://doi.org/10.1371/journal.pcbi.1005508, PubMed: 28437426

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

Friston, K. (2010). The free-energy principle: A unified brain theory? Nature reviews. *Neuroscience*, *11*, 127–138. https://doi.org/10.1038/nrn2787, PubMed: 20068583

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., et al. (2013). MEG and EEG data analysis with MNE-python. *Frontiers in Neuroscience*, *7*, 267. https://doi.org/10.3389/fnins.2013.00267, PubMed: 24431986

Griffin, G., Holub, A., & Perona, P. (2007). *Caltech-256 object category dataset* (p. 20). California Institute of Technology https://authors.library.caltech.edu/7694.

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

Guggenmos, M., Sterzer, P., & Cichy, R. M. (2018). Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *Neuroimage*, *173*, 434–447. https://doi.org/10.1016/j.neuroimage.2018.02.044, PubMed: 29499313

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. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778). https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html.

Kaniuth, P., & Hebart, M. N. (2021). Feature-reweighted RSA: A method for improving the fit between computational models, brains, and behavior. *bioRxiv*. https://www.biorxiv.org/content/10.1101/2021.09.27.462005.abstract

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

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

Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep neural networks in computational neuroscience. *Oxford Research Encyclopedia of Neuroscience*. https://doi.org/10.1093/acrefore/9780190264086.013.46

Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences, U.S.A.*, *116*, 21854–21863. https://doi.org/10.1073/pnas.1905544116, PubMed: 31591217

Klink, P. C., Dagnino, B., Gariel-Mathis, M.-A., & Roelfsema, P. R. (2017). Distinct feedforward and feedback effects of microstimulation in visual cortex reveal neural mechanisms of texture segregation. *Neuron*, *95*, 209–220.e3. https://doi.org/10.1016/j.neuron.2017.05.033, PubMed: 28625487

Kreiman, G., & Serre, T. (2020). Beyond the feedforward sweep: Feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, *1464*, 222–241. https://doi.org/10.1111/nyas.14320, PubMed: 32112444

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science*, *1*, 417–446. https://doi.org/10.1146/annurev-vision-082114-035447, PubMed: 28532370

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

Kubilius, J., Schrimpf, M., Kar, K., Hong, H., & Majaj, N. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. *arXiv Preprint*. https://arxiv.org/abs/1909.06161

Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., et al. (2018). The open images dataset V4: Unified image classification, object detection, and visual relationship detection at scale. *arXiv [cs.CV]*. https://arxiv.org/abs/1811.00982

Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, *23*, 571–579. https://doi.org/10.1016/s0166-2236(00)01657-x, PubMed: 11074267

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

Lamme, V. A. F., Zipser, K., & Spekreijse, H. (2002). Masking interrupts figure-ground signals in V1. *Journal of Cognitive Neuroscience*, *14*, 1044–1053. https://doi.org/10.1162/089892902320474490, PubMed: 12419127

Liao, Q., & Poggio, T. (2016). Bridging the gaps between residual learning, recurrent neural networks and visual cortex. *arXiv:1604.03640*. https://arxiv.org/abs/1604.03640

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*, *2014*, 740–755. https://doi.org/10.1007/978-3-319-10602-1_48

McKinney, W. (2010). Data structures for statistical computing in python. *Proceedings of the 9th python in science conference* (Vol. 445, pp. 51–56). https://doi.org/10.25080/Majora-92bf1922-00a

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

Mély, D. A., Linsley, D., & Serre, T. (2018). Complementary surrounds explain diverse contextual phenomena across visual modalities. *Psychological Review*, *125*, 769–784. https://doi.org/10.1037/rev0000109, PubMed: 30234321

Nunez, P. L., & Srinivasan, R. (2009). *Electric fields of the brain: The neurophysics of EEG*. (2nd ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195050387.001.0001

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..
Perrin, F., Pernier, J., Bertrand, O., & Echallier, J. F. (1989). Spherical splines for scalp potential and current density mapping. *Electroencephalography and Clinical Neurophysiology*, *72*, 184–187. https://doi.org/10.1016/0013-4694(89)90180-6, PubMed: 2464490
Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2019). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLoS Computational Biology*, *15*, e1007001. https://doi.org/10.1371/journal.pcbi.1007001, PubMed: 31091234
Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *British Journal for the Philosophy of Science*, *70*, 581–607. https://doi.org/10.1093/bjps/axx023, PubMed: 31086423
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
Roskies, A. L. (2021). Representational Similarity Analysis in neuroimaging: Proxy vehicles and provisional representations. *Synthese*, *199*, 5917–5935. https://doi.org/10.1007/s11229-021-03052-4
Russell, B. C., Torralba, A., Murphy, K. P., & Freeman, W. T. (2008). LabelMe: A database and web-based tool for image annotation. *International Journal of Computer Vision*, *77*, 157–173. https://doi.org/10.1007/s11263-007-0090-8
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: 19757938
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* (Vol. 57, p. 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
Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences, U.S.A.*, *104*, 6424–6429. https://doi.org/10.1073/pnas.0700622104, PubMed: 17404214
Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*, 1551. https://doi.org/10.3389/fpsyg.2017.01551, PubMed: 28955272
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
Tang, H., & Kreiman, G. (2017). Recognition of occluded objects. In Q. Zhao (Ed.), *Computational and cognitive neuroscience of vision* (pp. 41–58). Springer Singapore. https://doi.org/10.1007/978-981-10-0213-7_3
Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., et al. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences, U.S.A.*, *115*, 8835–8840. https://doi.org/10.1073/pnas.1719397115, PubMed: 30104363
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
van Bergen, R. S., & Kriegeskorte, N. (2020). Going in circles is the way forward: The role of recurrence in visual inference. *arXiv [q-bio.NC]*. https://arxiv.org/abs/2003.12128
Vigario, R., Sarela, J., Jousmiki, V., Hamalainen, M., & Oja, E. (2000). Independent component approach to the analysis of EEG and MEG recordings. *IEEE Transactions on Biomedical Engineering*, *47*, 589–593. https://doi.org/10.1109/10.841330
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
Xiao, J., Hays, J., Ehinger, K. A., Oliva, A., & Torralba, A. (2010). SUN database: Large-scale scene recognition from abbey to zoo. *2010 IEEE computer society conference on computer vision and pattern recognition* (pp. 3485–3492). https://doi.org/10.1109/CVPR.2010.5539970
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