RECURRENT ISSUES WITH DEEP NEURAL NETWORK MODELS OF VISUAL RECOGNITION

A PREPRINT

Tim Maniquet
KU Leuven
Leuven, Belgium
tim.maniquet@kuleuven.be *

Hans Op de Beeck
KU Leuven
Leuven, Belgium hans.opdebeeck@kuleuven.be

Andrea Ivan Costantino
KU Leuven
Leuven, Belgium
andreaivan.costantino@kuleuven.be

April 10, 2024

ABSTRACT

Object recognition requires flexible and robust information processing, especially in view of the challenges posed by naturalistic visual settings. The ventral stream in visual cortex is provided with this robustness by its recurrent connectivity. Recurrent deep neural networks (DNNs) have recently emerged as promising models of the ventral stream. In this study, we asked whether DNNs could be used to explore the role of different recurrent computations during challenging visual recognition. We assembled a stimulus set that included manipulations that are often associated with recurrent processing in the literature, like occlusion, partial viewing, clutter, and spatial phase scrambling. We obtained a benchmark dataset from human participants performing a categorisation task on this stimulus set. By applying a wide range of model architectures to the same task, we uncovered a nuanced relationship between recurrence, model size, and performance. While recurrent models reach higher performance than their feedforward counterpart, we could not dissociate this improvement from that obtained by increasing model size. We found consistency between humans and models patterns of difficulty across the visual manipulations, but this was not modulated in an obvious way by the specific type of recurrence or size added to the model. Finally, depth/size rather than recurrence makes model confusion patterns more human-like. Contrary to previous assumptions, our findings challenge the notion that recurrent models are better models of human recognition behaviour than feedforward models, and emphasise the complexity of incorporating recurrence into computational models.

Author summary

Deep neural networks (DNNs) are considered the best current models of visual recognition. This is mainly due to the correspondence between their structure and that of the ventral stream in the primate visual system, as well as a double match between their representations and behaviour with human neural representations and error patterns. Recently, it has been suggested that adding recurrence to usually feedforward-only DNNs improved this match, while simultaneously making their architecture more brain-like. But how much of human behaviour do these models actually replicate, and does recurrence really make things better? We conducted an in-depth investigation of this question by putting DNNs to the test. In our work, we ask: do models still resemble humans when the task becomes complicated, and: are they making use of similar strategies to operate object recognition? Bringing different architectures together, we show that recurrence tends to increase model performance and consistency with humans. However, we cannot dissociate this improvement from that brought by parameter size alone. Additionally, we find a striking worsened

match with human patterns of errors in models with recurrence, as compared to purely feedforward models. Contrary to previous assumptions, our findings challenge the notion that recurrent models are better models of human recognition behaviour than feedforward models, and emphasise the complexity of incorporating recurrence into computational models.

Introduction

The seemingly effortless nature of human object recognition hides the computational difficulty of real-world object recognition. The visual complexity of scenes from everyday life calls for flexible and robust processing in order to decode the identity of objects in the environment [Kravitz et al., 2013, Grill-Spector and Weiner, 2014, Bracci and Op de Beeck, 2023]. This feat is accomplished seamlessly by the ventral stream in visual cortex, whose connectivity allows for such extraordinary abilities. While it is often modelled as a bottom-up network, every feedforward connection it contains is paralleled by one or several non-feedforward connections ([Van Essen et al., 1986, Felleman and Van Essen, 1991, Lamme et al., 1998, Ungerleider et al., 2008, Baizer et al., 1991]). It is hence better understood as a highly **recurrent network**.

Recurrence in a system allows for the dynamic processing of information. Visual recognition, consequently, is a dynamic process. Although most objects are instantaneously recognised through a fast, feedforward sweep of processing ([DiCarlo et al., 2012, Riesenhuber and Poggio, 2000, Riesenhuber, 2005]), visual information undergoes progressive transformations across time. During these, sensory inputs are modulated and integrated. This extra-feedforward processing occurs through cycles of recurrent activity considered essential for many perceptual phenomena, e.g. stimulus-context modulation, predictive processing and figure-ground segmentation [Pennartz et al., 2019, Wyatte et al., 2014, Lamme and Roelfsema, 2000, Kreiman and Serre, 2020]. This general implication of recurrence in non-canonical, non-trivial contexts has experimental evidence. Given the temporal unfolding that is inherently linked with recurrent processing, it is possible to selectively impact it while leaving initial feedforward processing intact, for example through the procedure of backward masking. Several studies made use of backward masking to manipulate recurrent processing, and found that it selectively impaired visual recognition under challenging conditions, but not under non-challenging ones ([Seijdel et al., 2021, Rajaei et al., 2019, Tang et al., 2018]).

Strong evidence in favour of the recurrent nature of the visual cortex comes from studies using *deep neural networks* (DNNs). Optimised on recognition tasks, DNNs are considered good models of the ventral stream, mimicking its hierarchical structure and operating classification with human-like accuracy ([Cadieu et al., 2014]). DNNs have been found to display representational similarity with the ventral stream ([Yamins et al., 2014, Khaligh-Razavi and Kriegeskorte, 2014]). Strikingly, this similarity increases with recurrent DNNs ([Kar et al., 2019, Kietzmann et al., 2019]), indicating that the representational dynamics of the visual cortex fit better with these of a recurrent network, compared to feedforward-only. The aforementioned studies of backward masking also included DNNs to show that models equipped with recurrent-like features show more resilience to challenging conditions than feedforward-only models.

While feedforward processing is well understood, there is no universal account of the exact mechanisms served by recurrence in visual recognition. There are two factors that complicate this matter. First, there are virtually unlimited ways an image might be challenging to process. Natural environments have infinite potential variations. To reproduce those variations, objects could be rendered cluttered, blurred, noisy, and so on. Second, recurrence is an umbrella term for a variety of anatomical connections. There are lateral and feedback connections within and between brain regions, respectively, and there is no guarantee that one type of connection in, for example, primary visual cortex, serves the same function as a lateral connection in prefrontal cortex. Similarly, a top-down connection between two nearby regions in the visual system might serve a different purpose compared to a top-down connection all the way from prefrontal cortex to V4.

Taken together, these two factors provide a high number of possible combinations of how specific recurrent connections might underlie the resolution of well-defined image challenges. Across the visual cortex, various specific forms of recurrence have been studied and linked to perceptual phenomena. For instance: uncertainty computation ([Ladret et al., 2023]) and pattern completion ([Shin et al., 2023]) by lateral connections within V1, edge detection in cluttered images through lateral processing in V1 ([Self et al., 2013]), prediction about occluded shapes by feedback from the prefrontal cortex (PFC) to V4 ([Choi et al., 2018]), figure-ground modulation by feedback from V4 to V1 ([Klink et al., 2017, Angelucci et al., 2017]), and interactions with spatial frequency transformations by feedback from frontal areas ([Bar et al., 2006, Goddard et al., 2016]). Overall, the literature points to a wide range of roles played by recurrent processing to resolve a variety of visual challenges, with each recurrent cortical connection potentially operating a different computation on incoming sensory information.

The wealth of perceptual phenomena explained by discrete recurrent connectivity patterns, and the diversity of visual challenges found to be solved by recurrence through backward masking, suggests that visual recognition is served by a wide range of recurrent computations in charge of solving visual complexity. Identifying the various roles of recurrence and merging them in a larger picture of ventral stream function would prove significant in the understanding of human visual cognition. The present study aimed at taking a step in this direction with an experiment made to dissociate between different types of recurrence. We combined a variety of image level manipulations known to involve recurrence in humans in one large experiment ([Seijdel et al., 2021, O'Reilly et al., 2013, Tang et al., 2018, Rajaei et al., 2019, Goddard et al., 2016]). We implemented different types of recurrence in a number of DNN models, and compared their behaviour on a classification task with that of humans. Taking advantage of the architectural diversity of our models, we asked whether we could use DNNs to *dissociate between different recurrent processes* responsible for the resolution of *different challenges* during visual recognition.

We found that recurrence in DNNs helped them match with human performance, but not more than added depth. This feeds into the observation that recurrent neural networks can be considered time-wrapped equivalents of size-matched feedforward neural networks ([van Bergen and Kriegeskorte, 2020]). However, our results go a step further, as we also found a decrease in models match with human confusion matrices when equipped with recurrence. Additionally, we were not able to find dissociable patterns of performance across our recurrent models, which indicates that DNNs, as tools to model the visual cortex, are unable to reproduce the variability of phenomena that occur in brain recurrence. Overall, we interpret our results, in light of current developments in the field, as an argument for the development of more elaborate, biologically plausible ways of implementing recurrence in models of visual recognition.

Materials and methods

Participants

A total of 231 subjects participated in the study (195 females, age (mean±SEM): 18.7 ± 1.7), of which 13 were excluded due to low performance (average accuracy below 0.7). All participants gave their informed consent before taking part. The protocol was approved by the ethics committee at KU Leuven. The task took place online and lasted approximately 30 minutes.

Stimuli

Images of objects were selected from real-world scenes of the publicly available online databases MS COCO ([Lin et al., 2014]) and ADE20K ([Zhou et al., 2018]). Eight object categories were included: *person*, *cat*, *bird*, *tree*, *fire hydrant*, *building*, *bus*, and *banana* (see Fig 2). The categories were chosen to cover a diverse range of levels of animacy, real world size and aspect ratio. Stimuli were segmented out from their backgrounds, transformed to a 700×700 pixel size, and equalised on their contrast and average luminance levels. 10 exemplar objects were picked for each category, resulting in a stimulus set of 80 images.

Challenging manipulations

Object recognition was rendered challenging by applying several levels of visual manipulation on the stimulus set (see Fig 1). Each one of the manipulation was selected based on literature demonstrating that it is linked to recurrent processing, as evidenced by psychophysical, neuroimaging, or DNN data. We focused on three of them: *occlusion*, *clutter*, and *phase scrambling*. In total, the full stimulus set was rendered in one control condition (segmented object on a gray background) and 16 challenging conditions.

- *Phase scrambling*: phase scrambled images were generated by randomising the phase spectrum of objects. Images were Fourier transformed, and their phase spectrum was replaced by random noise, on either side of a spatial frequency threshold of 1.5 cycles per degree (with images assumed to subtend 10 degrees of visual angle on screen). The origin of the phase spectrum was left intact. Images were then transformed back into image space.

  This resulted in two phase scrambled versions of each image: a **low-pass** phase scramble, and a **high-pass** phase scramble.

- *Clutter*: cluttered images were obtained by replacing plain backgrounds with cluttered backgrounds. These were obtained by phase scrambling randomly selected natural scenes from the MS COCO dataset, and measuring an index of clutter on the outcome. The clutter index was calculated following the method of [Groen et al., 2013], whereby measures of spatial coherence (SC) and contrast energy (CE) are computed for each image, the combination of which gives a proxy of how cluttered an image is. More precisely,

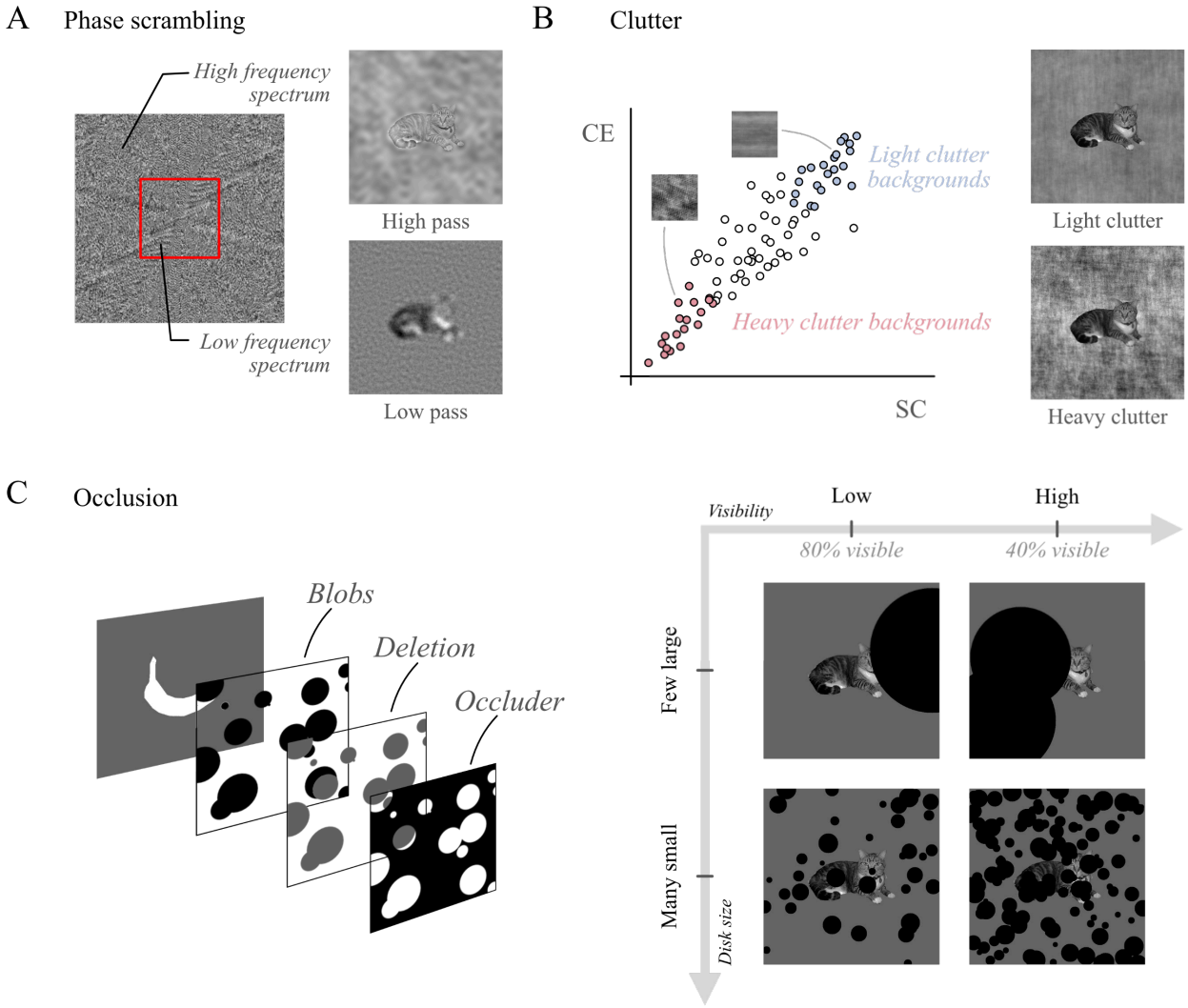

Figure 1: **Stimulus manipulations**. The stimulus set included eight object categories of 10 objects each, passed through a total of 16 visual manipulations. Objects in the images were rendered challenging by applying either *phase scrambling*, *clutter*, or *occlusion*. (A) Phase scrambling was applied by replacing the phase spectrum of images with random noise, on either sides of a 1.5 cpd threshold (red square not actual size). (B) Clutter was created following the method described in [Groen et al., 2013], whereby phase-scrambled versions of natural scenes taken from the MS COCO dataset were ranked on their *spatial coherence* and *contrast energy* values. Subsets of the most and less cluttered of a large number of such images were taken as light- and heavy-cluttered backgrounds for the stimuli. (C) Images were lightly or heavily occluded (percentage of object left visible: 80% and 40%, respectively), using many small or a few large disks. This was done in one of three possible fashions: adding black blobs, adding blobs the colour of the background (deletion) adding a full black occluder with disk-like apertures.

CE and SC values are proxys of LGN population response derived from the activation of contrast filters ([Ghebreab et al., 2009], [Scholte et al., 2009]).

Taken together, CE and SC distribute images on a two-dimensional space, with high object segmentability at the origin (high SC, low CE), and where objects become complex to segment from their background at the other end (low SC, high CE). We defined two levels of clutter by selecting 80 backgrounds at each end of the spectrum. Objects were hence rendered once on an lightly cluttered background (**light clutter**), and once on a heavily cluttered background (**heavy clutter**).

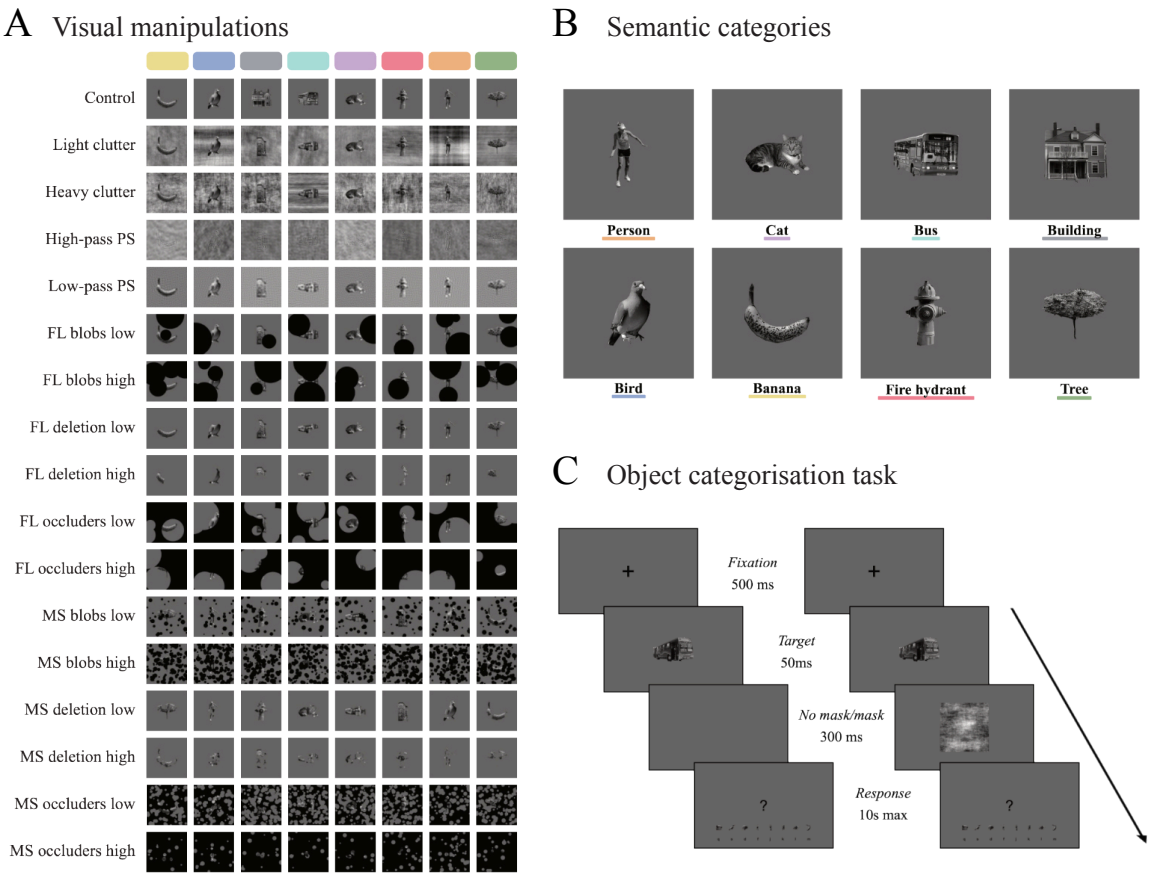

Figure 2: **Stimulus set & task**. (A) Full set of stimuli manipulations, represented by one selected exemplar per category. Each row represents a particular manipulation. Colours above the top row indicate the eight categories (there were 10 exemplar images per category, one example is selected here). PS: *phase scrambling*, FL: *few large*, MS: *many small*. (B) The eight categories included in the stimuli. (C) Experimental paradigm used for human participants.

- *Occlusion*: images were occluded by concealing them so that so that only a given proportion of the initial object would remain visible, either 80% (easy condition) or 40% (hard condition). Objects were concealed in three possible ways: by adding a full **occluder** with disk-like apertures, by adding occluding, black **blobs**, or by **deletion** of parts of the objects. Within each occluding technique, two levels of size were implemented: either **many small** disks or a **few large** disks were used.

  We combined several types of occlusion based on the observation that they are not equally challenging ([Tang et al., 2018, Johnson and Olshausen, 2005]), and therefore increase the richness of our stimulus set. This combination of difficulty levels (2), occluding techniques (3) and sizes (2) resulted in 12 different occlusion manipulations.

Experimental task

Participants were given a categorisation task, where each image was to be classified in one of the eight possible categories of the stimulus set. All trials had the same timeline (see Fig 2): after a fixation cross, a target image appeared for 50ms. Targets were either followed by an empty screen or by a mask. The instructions asked for an accurate answer that could be registered as soon as the target was presented. All trials ended with the presentation of a reminder of the eight categories and their associated response keys. The latter appeared at the bottom of the screen 200ms after the mask (or empty screen) disappeared.

Masks were phase-scrambled versions of natural scenes taken from MS COCO. Images were randomly selected from the dataset, resized to match target images (700×700), and taken to the Fourier space where their phase spectrum was fully replaced by random noise (except for the origin, left intact) before reconstruction into the image space.

In total, 16 experimental conditions and one control made 17 conditions, all of which to be presented with and without masking. The resulting 34 conditions were divided into sub-experiments: occlusion made up 2 sub-experiments (few large occluder disks and many small occluder disks, 6 conditions each). Phase scrambling and clutter were joined together into one other sub-experiment (4 conditions). These three sub-experiments were ran with and without masking, resulting in 6 experiments in total. An extra unmasked control condition was added to all three of the masked experiments to serve as a comparable baseline. All of them were conducted online over the same period, and no participant took part more than once.

Within tasks, each condition was presented once per image, i.e. 80 trials per condition. As a result, tasks consisted of 400 (without masking, phase scrambling and clutter conditions) to 640 trials (with masking, occlusion conditions).

Deep neural network architectures

We investigated thirteen distinct DNN models, each varying in architecture and recurrent connections (see Fig 3). These models were specifically chosen to represent a range of complexities, from basic feedforward structures to more sophisticated recurrent networks. We categorised these models into three groups: CORnet models, B models, and VGG models (see Table 1 for more details).

CORnet Models.

From the CORnet family, we included the foundational models CORnet Z, CORnet RT, and CORnet S ([Kubilius et al., 2018, Kubilius et al., 2019]). While CORnet Z uses a simple feed-forward architecture, CORnet RT enhances this design by integrating biologically-inspired additive recurrent mechanisms within each block. CORnet S further advances this architecture by adding additional convolutional layers in a bottleneck ResNet-like block structure, thereby enriching its internal dynamics and significantly increasing its parameter count relative to its simpler counterparts. For better readability, CORnet is hereafter abbreviated *C*, and models are branded after the recurrent element of connectivity they contain (*L* for *lateral* connections, *T* for *top-down* connections). We therefore selected *C* (CORnet Z), *CL* (CORnet RT) and *CS* (CORnet S). We custom-extended the C series for this research, by adding the following: C V1-V1, featuring within-layer connections in its initial convolutional layer; C IT-IT, implementing similar within-layer connections but in its terminal convolutional layer; CT, designed with top-down connections extending from higher to lower layers; and CLT, a specialized version of CL augmented with top-down connections analogous to those in CT.

B Models.

B models ([Spoerer et al., 2017]) include B, BL, BT , and BLT models. BL is characterized by lateral connections, BT by top-down connections, while the BLT model combines both lateral and top-down mechanisms. These models were used in their original implementations, see ([Spoerer et al., 2017]) for a more detailed description.

VGG Models.

The VGG models, specifically VGG11 and VGG16 [Simonyan and Zisserman, 2015], served as the feedforward control group. Their architecture, lacking feedback mechanisms, is deeper, providing a baseline comparison against recurrent models. VGG11 is a custom implementation of VGG16 featuring only 11 trainable layers and a smaller decoding head with a single linear layer. This modification was made to create a VGG16-like very deep feedforward network, but with a number of learnable parameters more comparable to the recurrent models.

Training Regimen.

Our model training process was divided into two distinct phases: a training phase on the ImageNet dataset, followed by a phase of fine-tuning on the specific 8 categories contained in our dataset.

ImageNet Training.

Models lacking pre-existing ImageNet-trained weights or featuring custom recurrent mechanisms (i.e. C V1-V1, C IT-IT, CT, CLT, VGG11) were trained in-house. This phase involved a single cycle of 25 epochs, using a fixed seed of 42 and a batch size of 256. A one-cycle learning rate policy [Smith and Topin, 2017], as implemented in the *fastai* [Howard et al., 2018] library, was employed. The maximum learning rate for the `fit_one_cycle` algorithm was set by multiplying the optimal rate, determined using *fastai*'s implementation of the learning rate finder

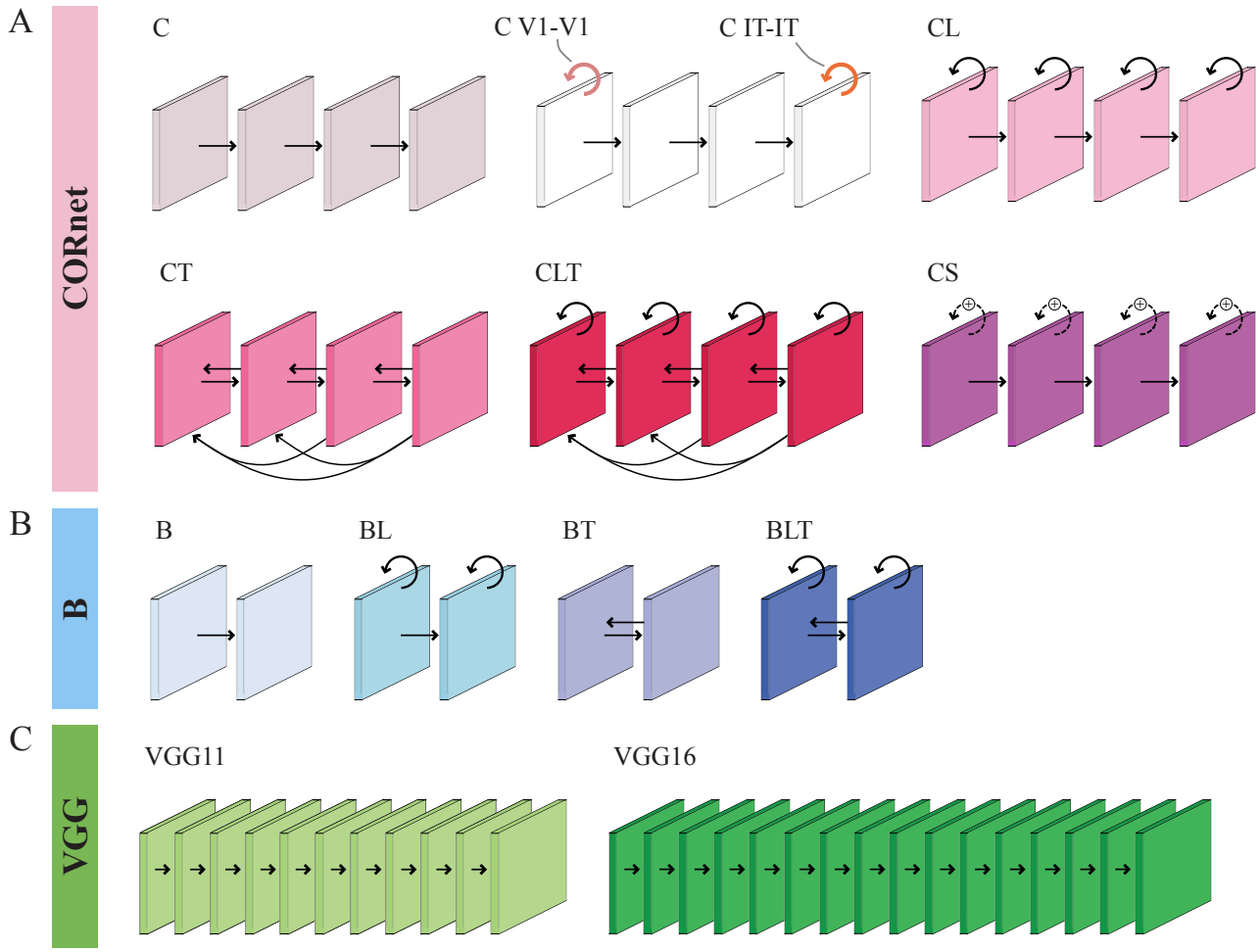

Figure 3: **Model architectures**. Rightward arrows represent feedforward connections. Circular, backward arrows above layers represent lateral connections. Leftward arrows represent feedback connections. The dashed arrows of CS represent the ResNet-inspired residual block structures. (A) Eight models were used from the *CORnet* family, including three readily available and five custom-made models. From [Kubilius et al., 2018], we selected the base FF model *CORnet Z* (C), the lateral connections-equipped *CORnet RT* (CL) and the highest-performing *CORnet S* (CS). Custom-made versions of these architectures were built, including full top-down connections (CT), full top-down and lateral connections (CLT), specific lateral connections in layers V1 (C V1-V1) and IT (C IT-IT). (B) The four available models of the *B* family were used. These models share a baseline 2-layer architecture (B), with the addition of either lateral connections (BL), top-down connections (BT), or both (BLT). (C) *VGG* models were used as feedforward controls to the recurrent models. We used the state-of-the-art VGG16 model, as well as a custom-made, smaller VGG11 to better match the size of our recurrent models.

[Smith, 2015], by a factor of 10. The training was performed on the ILSVRC2012 challenge subset of the ImageNet dataset [Russakovsky et al., 2015], which includes 1,000 classes and approximately 1.2 million images for training and an additional 50,000 for validation purposes. During the training phase, each image was normalized using the mean and standard deviation statistics from ImageNet, followed by a random crop with a minimum scale of 0.35 which was then resized to 224 x 224 pixels. In the validation phase, images were directly normalized and resized to 224 x 224 pixels, omitting the cropping process.

Fine-Tuning. For fine-tuning, we conducted 20 independent training runs on the same ImageNet-trained model, each with a distinct fixed seed. This approach ensured varied initializations for the new classifier head (adapted from 1000 output units to the 8 units needed for our classification task), for the dataloaders and different seed settings for essential libraries (e.g., *NumPy*, *PyTorch*, *fastai*, and CUDA). Each run comprised 3 cycles, each with 9 epochs, amounting to 27 epochs in total. The initial 6 epochs of each cycle focused on training the model's classifier and the last convolutional

layer before the classifier. This strategy was adopted to preserve low-level features learned from ImageNet, aligning with our goal to model human-like visual processing by retaining features from a dataset rich in natural visual variance. In the last 3 epochs of each cycle, the full model was trained, allowing for a comprehensive adaptation to our 8-way classification task. The learning rate for each cycle was dynamically determined using the learning rate finder algorithm from *fastai* at the start of each cycle. The `fit_one_cycle` method parameters were set with `lr_max` as the optimal learning rate multiplied by 10.

The fine-tuning dataset included a total of 32,000 images taken from the same datasets as these used to build the human dataset (see Stimuli), with an equal distribution of 4,000 images per category. Each category was divided to include half the images with their original background (split equally between colored and black and white) and half against a neutral gray background (also divided equally between colored and black and white). This approach allowed us to train the network to accurately categorize images irrespective of background variations and color cues.

This dual-phase training approach produced 20 uniquely fine-tuned versions of each model, ensuring robust adaptation to our classification task while preserving foundational ImageNet-acquired knowledge. The Adam optimizer with a Cross Entropy loss function was employed throughout both phases of training. Training was executed in parallel on 3 NVIDIA GeForce RTX 2080 GPUs. The models were implemented using *PyTorch* version 2.1.2 [Paszke et al., 2019], *TorchVision* version 0.16.2 [TorchVision maintainers and contributors, 2016], and *fastai* version 2.7.13 [Howard et al., 2018].

Model Sizes and Connections

A comprehensive overview of the model architectures is provided in Fig. 3. Models that implemented at least one type of recurrent connection (i.e., top-down or within-layer recurrent mechanism) have an added time dimension due to the iterative nature of recurrent dynamic. Here, we set the total time-steps to 5 during both training and testing phases. A time-step here represents an end-to-end (from input to output) processing of the input image.

Top-Down Connections.

Top-down mechanisms were implemented in three of the models used here: CT, CLT and BLT. The implementation of top-down mechanisms in the CT and CTL architectures, inspired by the CORnet model, and the BLT model, reveals distinct approaches to incorporating feedback connections in deep neural networks.

The BLT architecture incorporates top-down connections through transposed convolutional layers between consecutive layers, enabling the model to integrate abstract, higher-level information back into earlier stages of processing. The de-convolution and reintegration process could be seen as part of a predictive coding strategy, where the model minimizes prediction errors between higher level predictions and actual sensory input – akin to generative models that seek to enhance feature selectivity and reconstruct low-level sensory inputs from higher-level abstractions. The additive nature of the de-convolved signal in the implementation used here could lead to a direct influence of higher-level representations on lower-level processing, but offers less capacity for non-linear integration of the bottom-up and top-down signals compared to our other networks.

In CT and CLT models, on the other hand, each block is uniquely structured to integrate top-down inputs from all higher blocks, not just the adjacent one, through a specialized top-down pathway before being combined with the feedforward input through an additional convolutional layer. Such a design facilitates a more sophisticated non-linear integration of hierarchical information, allowing for the integration of bottom-up data with top-down information from all higher layers. This method fosters richer representation at each processing stage at the cost of increased computational complexity and a partial information loss due to channel averaging.

Specifically, in the CT and CLT models, a block processes a bottom-up input x (the output of the previous block at time-step $t-1$) and combines it with top-down feedback M, derived from the aggregated outputs of higher-level blocks at time-step $t-1$. In instances where top-down input is absent – such as during the initial time-step when no higher-level info is available – M is set to zero.

The feedforward path FF and the top-down path TD each involve convolution (*), Group Normalization (N), and ReLU activation (R):

$$FF(x) = R\left(N\left(x * \theta_{ff}\right)\right), \quad TD(M) = R\left(N\left(M * \theta_{td}\right)\right) \tag{1}$$

The weights θ_{ff} and θ_{td} represent distinct sets of learned weights for the feedforward and top-down convolutions, respectively.

The integrated top-down input M is formed by processing the outputs from higher layers $m_1, m_2, \ldots, m_n$:

$$M = \text{Concat}\left(\text{ProcessTopDown}(m_1, x), \text{ProcessTopDown}(m_2, x), \ldots, \text{ProcessTopDown}(m_n, x)\right) \tag{2}$$

Given a top-down input tensor m_i from layer i and a bottom-up input tensor x, the function ProcessLayer(m_i, x) is defined as follows:

$$\hat{m}_i = \text{AdaptiveAvgPool2d}(H_x)(m_i), \quad \text{ProcessTopDown}(m_i, x) = \frac{1}{C}\sum_{c=1}^{C} \hat{m}_i^{(c)} \tag{3}$$

where H_x is the height of the bottom-up input tensor x, $\hat{m}_i$ is the tensor obtained after applying adaptive average pooling to m_i with a target spatial dimension of H_x, and the function computes the mean across the channel dimension of $\hat{m}_i$, resulting in a tensor with a single channel and same spatial dimensions as the bottom-up input.

The final output y of the block is generated by concatenating the feedforward and top-down outputs and applying convolution, normalization, and ReLU:

$$y = O\left(\text{Concat}\left(FF(x), TD(M)\right)\right) \tag{4}$$

with $z = \text{Concat}(FF(x), TD(M))$, where concatenation occurs along the channel dimension. The output operation O applies convolution, normalization, and ReLU activation on the concatenated tensor z:

$$O(z) = R\left(N\left(z * \theta_o\right)\right) \tag{5}$$

In this setup, the feedforward path (FF) processes input from the preceding block, while the top-down path (TD) integrates inputs from multiple higher layers, creating a comprehensive feedback mechanism. This design allows for a dynamic interplay of bottom-up perceptual data and high-level predictions. Top-down connections facilitate a backward flow of information, refining feedforward outputs with knowledge acquired from more abstract representations. This approach mirrors the concept of biologically plausible models, where visual information processing involves not only a feedforward pathway but also feedback loops that dynamically adjust and refine their understanding of visual inputs [van Bergen and Kriegeskorte, 2020].

Within-layer Connections.

The CORblock RT block, used in the original implementation of CL [Kubilius et al., 2018]), was used here to model within-layer connections in models with within-layer connections only (i.e., CL, C V1-V1 and C IT-IT) to keep our custom implementations closer to the original. In these models, lateral connections are established within the block, where the output of a block at one timestep is used as an additional input in the next timestep (see [Kubilius et al., 2018]) for more details). On the other hand, within-layer connections are implemented through temporal depth (i.e., through several iterations over the same block) in CORnet S, with skip connections that facilitates the preservation and integration of information across multiple passes of the same block, effectively approximating a very deep network architecture with shared weights. In contrast, lateral connections are implemented in BLT through specific convolutional layers that are dedicated to processing inputs from the same layer at previous timesteps. These layers are specifically tasked with processing the output of the same block from the previous timestep, effectively allowing the block to integrate its past state with the current input at the "cost" of additional parameters.

A different approach was used to model within-layer connections for C LT, due to the more complex shapes and different natures of the top-down, bottom-up and lateral inputs. For each CLT block, during the forward pass for an input x at a given timestep t, a previous state s_{t-1}, and top-down input td, the output y_t and the updated state s_t can be described as follows:

$$\begin{aligned} ff &= FF\left(x\right), \\ td &= TD\left(M\right), \\ rr &= R\left(N\left(s_{t-1} * \theta_{\text{rr}}\right)\right), \\ x' &= \text{Concat}(ff, rr, td), \\ y_t &= R\left(N\left(x' * \theta_{\text{out}}\right)\right), \\ s_t &= y_t, \end{aligned} \tag{6}$$

In this representation, ff corresponds to the output of the feedforward pass, computed by convolving the input x with the weights θ_{ff}, followed by normalization N and ReLU activation R. The processed top-down input, td, is obtained by applying a sequence of operations defined by ProcessTopDown on the concatenated top-down input M (see sub-section *Top-Down Connections.* for more details). The term rr represents the within-layer processing of the previous state s_{t-1}, convolved with θ_{rr} with kernel size and stride of 1 to preserve spatial dimensions, and then normalized and ReLU-activated. The concatenation operation Concat merges the outputs ff, rr, and td into a single tensor x' along the channel dimension. This concatenation preserves the spatial dimensions of the feature maps, and effectively pools

the distinct features extracted from each individual path into a unified representation for subsequent layers. Finally, the output y_t of the block at timestep t and the updated state s_t are obtained by convolving x' with the output weights θ_{out}, followed by normalization and ReLU activation. This output serves both as the response of the block at the current timestep and the state for the next timestep.

This recurrent structure enables the layer to integrate both current and previous activations, thereby enriching the artificial system with temporal dynamics and mimicking the recurrent mechanisms observed in biological visual systems ([Kietzmann et al., 2019, Kar et al., 2019]).

Model Evaluation and Statistical Analysis

After training and fine-tuning, each model was tested against our test dataset. The test dataset included 800 black and white images (100 images for each of the same 8 categories) on a gray background. The dataset was presented to the network in a control or challenging condition (see *Challenging manipulations* for more details).

To evaluate the performance of each DNN model we employed a Top-1 Error Rate metric, which assesses classification accuracy based on the top prediction. Classification accuracy for each item in the test dataset was recorded for each model and each training seed, resulting in one confusion matrix per model across all the seeds.

Results

In this study, we compared humans and DNN models on a categorisation task in order to explore the respective advantages offered by different types of recurrence. To capture these potential differences we developed a rich stimulus set that would be sensitive enough to allow for a fine-grained test of model fit to human data. The stimulus set quality was twofold. First, it implemented a variety of visual challenges (occlusion, phase scrambling, clutter), aimed at reproducing the diversity of natural scene complexity, while maintaining control over the degree of challenge. Second, it included a high number of object categories (eight categories: people, cats, birds, trees, bananas, buses, buildings and fire hydrants, see Fig 2), thereby covering an array of human recognition behaviours.

During the experiment, images were presented either manipulated (challenging conditions, 16 conditions) or non-manipulated (control condition, segmented objects on a plain background), for a total of 17 tasks to solve. Models and humans were asked to label the category the image belonged to on every trial, with human participants instructed to do so as fast and accurately as possible. We measured model accuracy, and human accuracy and reaction time (RT) to the task, to quantify the difficulty posed by the visual manipulations.

For humans only, half of the trials contained **backward masking**. In masked trials, a phase scrambled noise pattern was presented shortly after the stimulus. This allowed for the selective impairing of recurrent processing, and served as further confirmation that the visual manipulations we implemented were linked to recurrence.

Different DNN models were trained and tested. We ensured our models covered a wide range of architectures and sizes. The architectures we focused on were either purely ***feedforward*** or contained ***recurrence***. Within our recurrent models, different types of recurrent connections were implemented. We used existing models from three families: *C*, *B*, and *VGG* ([Kubilius et al., 2018, Spoerer et al., 2017, Simonyan and Zisserman, 2015]). While the *VGG* family is purely feedforward and aims at maximising performance through depth, both the *C* and *B* families involve recurrence and are built as biologically plausible architectures. Conversely to the B models where all cases of feedback and lateral connectivity are explored, the C models included only a base, feedforward model (*C Z*, referred to hereafter as *C*), a fully laterally connected model (*C RT*, referred to hereafter as *CL*), and a high-performance, enhanced version of the latter (*C S*). To increase the spectrum of recurrent architectures in this family, we built ***four extra C architectures***: *C V1-V1*, *C IT-IT*, *CT* and *CLT*. On top of C and CL, these new models allowed us to separately look at the role of lateral and feedback connections in the C family too. On top of architecture, we had large model size differences across our DNNs, with smaller models (e.g. C, 4.5m parameters) more than twenty times smaller than our bigger ones (e.g. VGG 16, 110m parameters). Overall, we used a range of models capable of dissociating between the effects of depth, size, and recurrent connectivities on performance. All our models underwent a similar regimen of training: first on ImageNet, then fine-tuned on a custom dataset of images from the eight selected categories.

Visual manipulations trigger recurrent processing

Data from all six experiments were pooled together, after confirmation that results on the unmasked control, which was similar for all participants, did not differ significantly across experiments (one-way ANOVA on accuracy with experiment as factor, $F = 0.68, p = 0.64$). Average accuracy was then computed for each participant. A first two-way ANOVA was ran on average accuracy per manipulation (control, phase scrambling, occlusion and clutter) and per level

of masking (mask, no mask). Results showed a significant interaction effect ($F = 13.06, p = 3.89e-08$) typical of masking, with performance varying differently across levels of task difficulty (see Fig 4). Post-hoc analyses showed significant differences between *control* and both *phase scrambling* (all Bonferroni corrected - $p = 2.8e-41$) and *occlusion* ($p = 3.59e-25$), but not *clutter* ($p > 0.05$), as well as between *clutter* and both *occlusion* ($p = 1.37e-19$) and *phase scrambling* ($p = 4.16e-35$). Finally, *occlusion* and *phase scrambling* are significantly different as well ($p = 2e-16$).

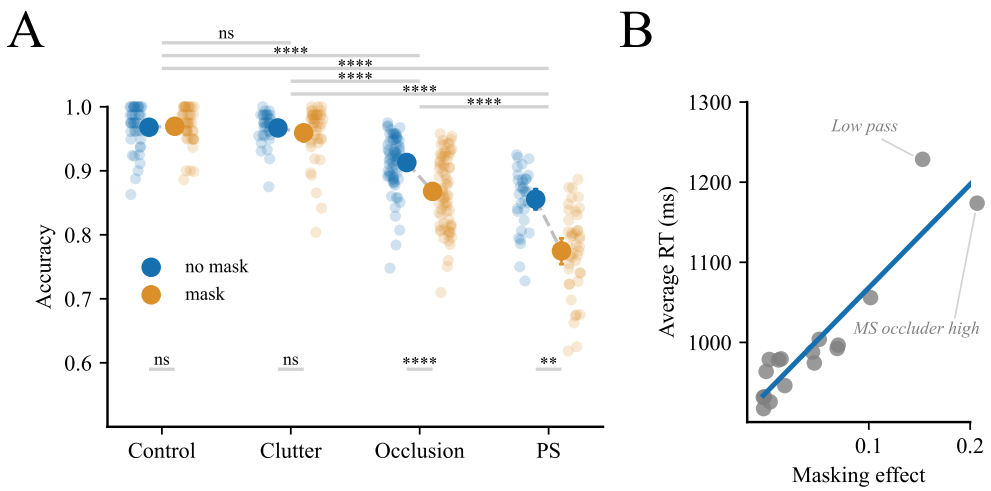

Figure 4: **Human behavioural results**. (A) Masking shows an interaction effect with manipulations, and impairs performance in occlusion and phase scrambling but not clutter. *PS: phase scrambling*. (B) A strong correlation is found between average RT and masking effect ($r = 0.93$). Their relationship is shown per task. Each dot represents a task (17 tasks in total). The line shows the best linear regression fit.

To further investigate the role of masking, three two-way ANOVAs were then performed on accuracy within each of the three main manipulations, with factors *manipulation level* and *masking*. Results vary across tasks. There is no significant effect of masking, clutter difficulty (light, heavy) or their interaction (all $p > 0.2$). There is, however, a significant interaction effect of masking and occlusion level (easy, hard - $F = 23.93, p = 1.69e-06$), on top of significant main effects of masking ($F = 43.45, p = 2.15e-10$) and occlusion level ($F = 265.29, p = 2.03e-42$). We finally found a significant interaction effect of masking and phase scrambling level (low pass, high pass - $F = 33.78, p = 3.65e-08$), as well as significant main effects of masking ($F = 45.07, p = 3.8e-10$) and phase scrambling level ($F = 670.55, p = 7.81e-57$).

To better account for the variations in task difficulty across our manipulations, we turned to RT. Correlations were computed between the **masking effect** and the **average RT** of each of our 17 tasks (see Fig 4). Masking effect is defined as the difference in average accuracy between *non-masked* and *masked* trials, and shows how performance decreases in the presence of masking. We found a strong, positive correlation between average RT and masking effect (Pearson's correlation $r = 0.92$, $p = 7.04e-08$), which remains strong even after the exclusion of the two most extremes performance values (*low-pass phase scrambling* and *MS occluder high*, Pearson's correlation $r = 0.86$, $p = 4.5e-05$).

Masking effect and RT are traditionally used to measure recurrence in visual recognition, but separately. Longer RTs are typically taken as indicators of extra, post-feedforward computations, while backward masking effects are direct evidence for the need of recurrent processing to solve the task. The concurrence of both measures of recurrence in our results shows that our tasks covary on difficulty and need for recurrent processing.

Model fit to human behaviour varies with recurrence and size

Having established the role of recurrent processing in our stimulus set, we next asked how our models would perform on it. We started by comparing model average performances, in order to see whether added recurrence or larger model size would lead to an improvement at all. Average accuracy per model was collected for each of the 20 random initiation seeds (see Fig 5). Performances were first compared across models by running a one-way ANOVA on accuracy with model as a main factor. Results showed a significant effect of models ($F = 474.63$, $p = 5.64e-163$). Post-hoc

comparisons of the models average accuracies showed significant differences across almost all pairs of models (Tukey test; all pair-wise comparisons significant, except: C - C V1-V1, C IT-IT - B, CT - BL, CLT - BT - BLT, CL - BL - BT, BT - BLT). In particular, VGG16 showed significantly higher accuracy than all models ($p < 0.001$, Tukey test). This reflects a striking difference in performance, with VGG16 on average 10% more accurate than other models.

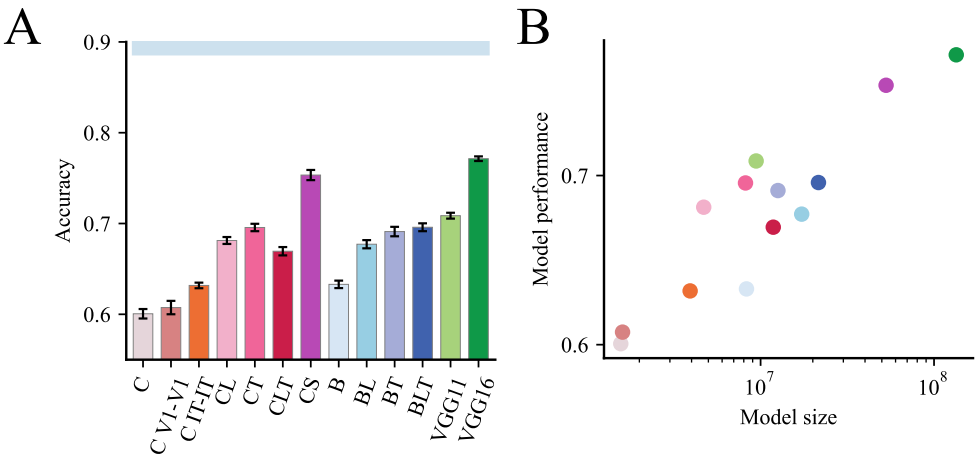

Figure 5: **Model performance results**. (A) Model performances: bar plot of average performance per model. Error bars represent SEM. The blue horizontal bar represents average human performance (0.89). (B) Scatter plot of model average accuracy per model size, in number of parameters. Note that the x-axis is in logarithmic scale.

Within our recurrent models, we observed an increase in performance with added recurrent connections. The more recurrent models especially (CL, CT, CLT, CS, BL, BT & BLT) seem to perform higher than their baseline counterparts (CL, CT, CLT and CS $>$ 95% CI of C; BT and BLT$>=$ 95% CI of B). This improvement, however, could also be explained by the increase in size underwent by models as recurrence is added to them. To check whether performance across models was driven by size, we correlated model parameters numbers with model average performance. Results showed a large correlation (Pearson's correlation $r = 0.74$, $p = 0.003$). Overall, although recurrent models perform better than their non-recurrent counterparts, this analysis seems to point to **model size** as the main driver of the performances of our models.

Better performing models are more consistent with humans performance patterns

While performance is a good indicator of how well a given model can solve the complex tasks of our stimulus set, it does not take into account the variability in task difficulty. A model could have a lower overall performance but a better task-wise fit with humans. To look beyond accuracy, we took advantage of our stimulus set variability to ask which models most consistently fit the ***pattern of performance*** of humans, and find difficulty where human participants found it. To do so, we divided the data into **17 tasks**, from which a vector of 17 average accuracy values (one for each task) was calculated per model and for humans (see Fig 6). We then correlated each model with participants to measure their consistency with human data.

Results from this analysis (see Fig 7) show an overall high correlation (> 0.65), indicating that DNNs, on average, correlate with humans and agree on which task is easier or more difficult. Models overall correlate better with humans when equipped with recurrent connectivity. With the exception of *C V1-V1*, all recurrent models achieve larger correlation scores than their baseline equivalent (all recurrent *C* and *B* models with a mean correlation above the 95% CI of both baseline *C* and *B* models). Notably, all these models also show a higher correlation than VGG11, but not VGG16. The latter surpasses all with a correlation score of 0.85. These results show that, at approximately equivalent sizes, some recurrent models fare better on matching human behaviour than feedforward-only models (for instance, CT & CLT do better than VGG11 with approximately similar parameter numbers). However, that is not always true (for instance, C compared to C V1-V1 and C IT-V1). Regardless, the highest correlation is found for VGG16, which is an indication that the computational flexibility brought by recurrence in our models is not superior to that brought by added depth. In fact, model size seems to drive the consistency of models with humans (Pearson's correlation $r = 0.83$, $p = 3.9e - 04$), in a way comparable to how it drives performance.

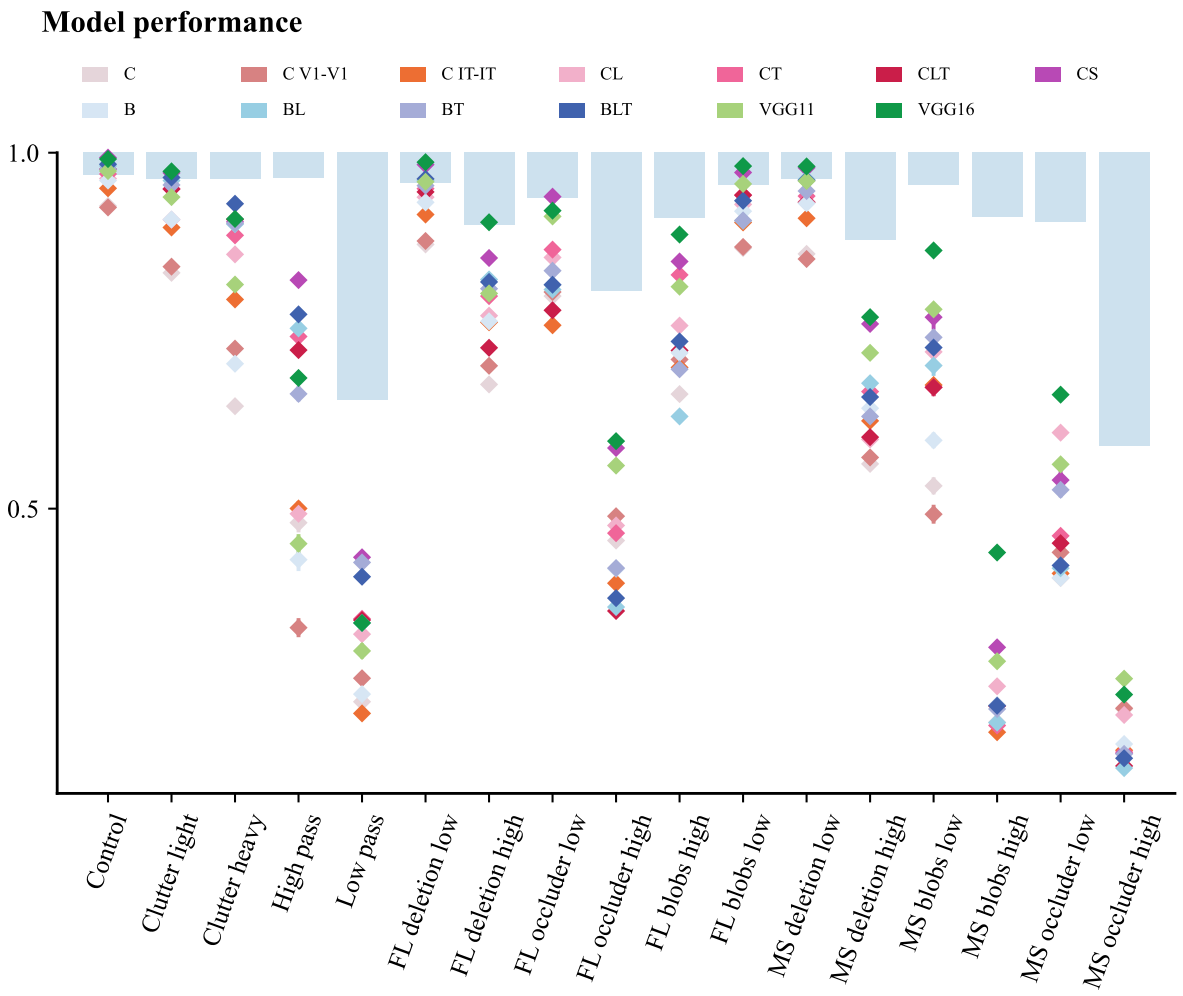

Figure 6: **Performance across task per models**. Each diamond represents the average performance of a model on a given task (bars on the diamonds indicate SEM). Human performance is shown in blue, with the bottom of the blue bars indicating average human performance. While models on average perform lower than humans, they overall follow the same pattern.

Recurrence types do not dissociate across tasks

Better performing DNNs are on average more consistent with humans. This correspondence between general performance and performance consistency suggests that solving the task better leads to matching humans patterns better. However, this does not rule out model-specific effects, or connection-specific effects, whereby some tasks could be better solved by some models than others. For instance, a task could be better solved by a model exhibiting lateral recurrence, with that model not performing best overall. To look for such effects and try to dissociate between types of recurrence, we checked whether the order of performance across models (see Fig 5) was repeated within each task, and whether there were exceptions to that order. For each task, one vector of thirteen values was created with the average performance of each DNN model. Each vector (17 vectors in total, one per task) was then correlated to the overall model order of performance.

Within-task model performance rankings seem to converge with the overall results: all tasks show a strong (Pearson's $r > 0.54$, 0.76 on average) correlation with the order of model performance, with the only exception of *many small occluder, high* (correlation of 0.04). The latter deviates from the global trend, with baseline models reaching better accuracies than expected (e.g. VGG11 surpasses VGG16, B surpasses its B counterparts, C is the second best of the C models). While it is not yet clear why this task falls out of the general trend, it is noticeably more difficulty than average, and might show a *floor effect*. We further address this point in the discussion. This convergence of within-task

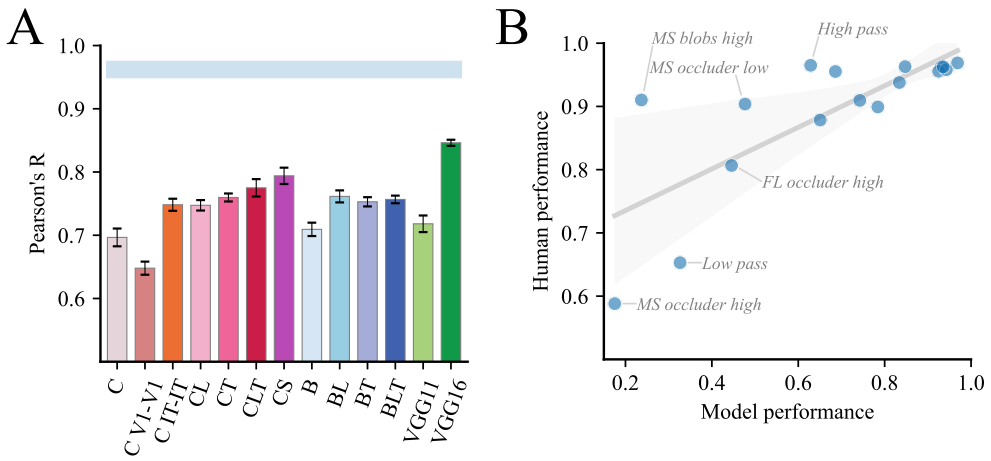

Figure 7: **Model consistency with humans**. (A) Correlation between human performance on 17 tasks and models. Error bars show a 95% CI around the mean of the correlations calculated from each of the model's 20 initiation seeds. The blue shaded area represents the noise ceiling, calculated as the average split-half reliability of the human data (over 100 iterations). (B) Scatter plot of human and average model performance per task. Each dot represents a task. The grey line shows the best linear regression fit, with the grey shaded area showing a 95% confidence interval (CI) around the regression.

with overall model performance ranking indicates that models do not display strong task-specific effects that could link architectural features to particular task challenges.

Despite the absence of model-specific effects, task-specific effects could remain: some tasks could fall out of the human-DNN consensus, and instead be better solved by humans or by models. To check for this, we looked for outliers on the human-model agreement (see Fig 7). Two noticeable pairs of outliers were found. In the phase scrambling condition, the *low pass* and *high pass* pair, surprisingly, shows models performing better than expected on the former, and worse than expected on the latter. This clashes with the generally accepted notion that DNNs rely on high-frequency information to perform classification ([Geirhos et al., 2022, Avberšek et al., 2021]). In the occlusion condition, the *MS occluder high* and *MS blobs high* pair also falls out of distribution, with models performing surprisingly well on the former and surprisingly low on the latter. This might be due to the high number of black pixels in these manipulations (see Fig 2). The *occluder* images, indeed, have on average more black pixels than the *blobs* images (301e3 v. 137e3 black pixels per image, on average). These black pixels might create a visual transient that could distract participants and interfere with performance, which could explain the unexpectedly low human accuracy.

Depth, not recurrence, makes model mistakes more human-like

In order to compare the respective advantages of recurrence and size in matching with human patterns of performance, we looked at category-level errors. We used ***confusion matrices*** to get a measure of how categories are represented with respect to each other. Confusion matrices are built by calculating the number of times each of the eight categories was given as a response when each of the eight categories was presented. The result is a 8x8 matrix where each cell counts the number of responses to a given category. A confusion matrix was built for humans, and for each DNN model. The results were then visualised with principal component analysis (PCA). We ran PCAs at the image level, with two main components fitted onto the confusion data of each individual image (see Fig S10). To quantify the agreement between models and humans, we ran a correlation analysis at the confusion matrix level. Matrices were correlated without diagonal values using Pearson's R (see Fig 8).

The results of this analysis show a pattern diverging from that seen in the performance correlation ran previously. Here, recurrence in models does not seem to give an advantage. Rather, within the C and B families, adding recurrence decreases the correlation with humans. Within family, baseline models seem to outperform recurrent ones. To confirm the lack of an advantage of recurrence, we built 95% CIs around model correlation scores. To do so, we calculated participant-level correlations: we obtained 218 correlation values for each model, from the correlations of each participant's confusion matrix with the overall model confusion matrix. This allowed us to build CIs around the average individual-level correlations (see Fig 8). This more conservative approach confirms the observed patterns. In

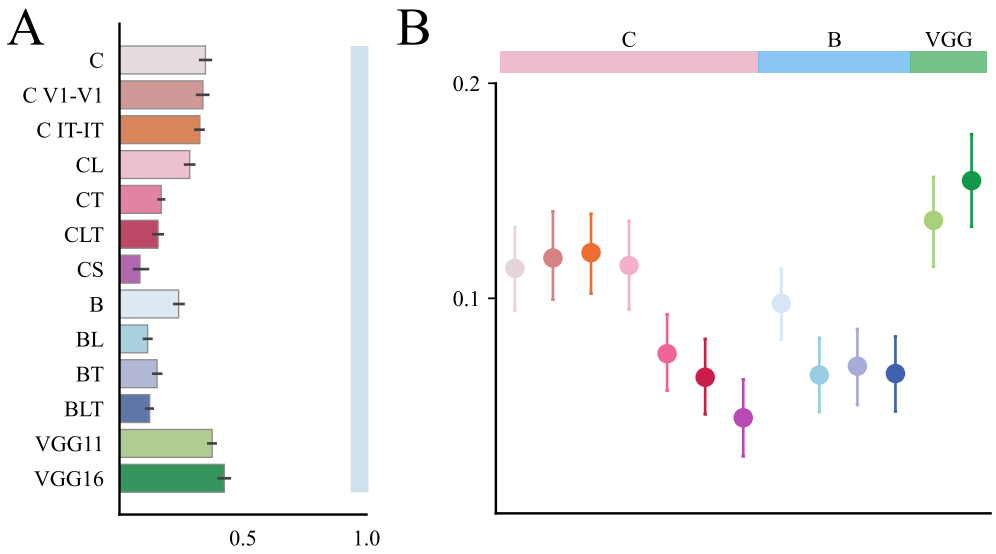

Figure 8: **Correlating confusion matrices**. (A) Bar plot of correlations between the human confusion matrix and the model confusion matrices. Bars indicate the standard error of the mean, calculated from the 20 random starting initializations of each model. Blue shaded area represents the noise ceiling, calculated as the average split-half reliability of the human group-level confusion matrix (over 100 iterations). (B) Average individual participant confusion matrix correlation with each model. Bars represent a **95% confidence interval** around the mean. VGG 16 shows a higher correlation than any other model (all models below the lower CI boundary, 0.13).

particular, CS and recurrent versions of B all have average participant-level correlation scores outside of and lower than the CI boundaries of the correlation scores of their baseline counterparts (boundaries of 0.09 and 0.08 for C and B, respectively). Additionally, all recurrent versions of C and B show correlations lower than VGG 16 (all below the CI lower boundary of 0.13). These values are striking when compared to model-model confusion matrix correlations: all models have a relatively higher average correlation with other models (Pearson's r, seed-level correlation with all other models in the range $0.24 - 0.94$, with average 0.73). Therefore, when looking at the pattern of mistakes committed by our models, we do not find an advantage of adding recurrence.

Overall, the results of our analyses all have VGG 16, the deep, large, feedforward-only model, as the best to human data. We cannot, from the tests we ran, conclude for recurrent connectivity in DNNs as a better strategy to achieve a task in a more human-like fashion, as compared to depth and overall size.

Discussion

In this study, we investigated the role of recurrence in DNNs and its implication in the resolution of visual complexity. We implemented a range of image-level manipulations known to involve recurrence in humans, and aimed at associating them with different types of recurrent connectivity in a set of DNN models. Our objective was to use DNNs to dissociate between distinct recurrent processes responsible for resolving different challenges during visual recognition. By comparing human and DNN behaviour in a classification task, we set out to find conditions where certain models more closely aligned with human behaviour. Our results on model performance across various architectures showcases the current limitations of recurrence in DNNs and their implications for modelling human visual recognition.

We found a nuanced relationship between recurrence, model size, and performance. The inclusion of recurrent connections in DNNs improved performance, however, this improvement was tied to the overall size of the network. Larger models, therefore, fared better on our tasks than smaller ones, with VGG16 performing best.

In a comparable way, when looking at performance consistency across tasks, we did not find that recurrent models aligned better with human difficulty perception. Instead, a similar pattern emerged whereby larger models aligned more consistently with humans, with model size broadly predicting how well a given model would do. Additionally, the accuracy ranking of models collapsed when task difficulty reached a certain level (notably on the *MS blobs high occlusion* task, see Recurrence types do not dissociate across tasks), which seems to match with reports of DNNs

failing to replicate human performance in highly difficult conditions ([Geirhos et al., 2018, Ghodrati et al., 2014]). Overall, we found an agreement between models and humans on task difficulty, in line with other reports showing a tendency of DNNs to generally fail when humans fail ([Kheradpisheh et al., 2016]). We also found this agreement to be performance-dependent, with larger models displaying more consistency than smaller ones, in line with other reports ([Lee and DiCarlo, 2023]).

On top of this size comparison, we set up our experiment to be able to distinguish between different types of recurrence. By choosing and building DNNs that contained different types of recurrence in different layers, we aimed at finding task-dependent effects of recurrent processing. However, we did not find any such distinctions in our results, as evidenced by the overall consensus of performance ranking across tasks (see Better performing models are more consistent with humans performance patterns). This null result is surprising considering the ample evidence for region-specific perceptual phenomena in the brain. One could have expected, for instance, C V1-V1 to behave differently in tasks requiring figure-ground modulation (e.g. clutter tasks) given the known role of recurrent processing in V1 in this phenomenon ([Lamme et al., 1998, Self et al., 2013]). This result, combined with the general performance-dependent model consistency, shows that the recurrent connections we implemented do not critically change the behaviour of our models.

When considering task performance and consistency, our results align well with the notion that recurrent neural networks can be considered time-wrapped equivalents of size-matched feedforward networks ([van Bergen and Kriegeskorte, 2020, Liao and Poggio, 2020]). Recurrent architectures, though, could be considered superior because more brain-like. Furthermore, recurrence allows to increase the number of computations without an increase in the number of neurons. Thus, if we were to express model size in terms of the number of units rather than the number of parameters, then recurrence results in an improved performance without an increase in 'size'.

However, even with this handy way out to promote recurrence as a solution with unique benefits, recurrence faces other challenges. Our examination of confusion matrices provided a finer-grained comparison of models, and showed, contrary to expectations, a drop in the alignment of recurrent models, specifically CS and BLT, with human behaviour. This corroborates reports showing that DNNs rely on different strategies to operate visual recognition ([Lonnqvist et al., 2019, Biscione and Bowers, 2023, Hosseini et al., 2017]). Importantly, we found a characteristic tradeoff between overall performance and fit with human data ([Fel et al., 2022]) within the recurrent model families C and B, but not VGG.

This challenges the observation that recurrent models better mimic the representational patterns of the brain than non-recurrent ones ([Nayebi et al., 2022, Kietzmann et al., 2019, Kar et al., 2019, Spoerer et al., 2020]). While recurrent models are reported to match better the brain's functioning, our study suggests that they could simultaneously be worse models of human behaviour. Therefore, we stress behavioural fidelity as a metric for developing better models of human visual recognition.

Strikingly, although we used a design equipped for it, we found no clear dissociations between the effect of specific types of recurrence. Lateral connections did not induce specific effects compared to top-down connections, either in general or in a layer-dependent way. This lack of dissociation could mean that the implementation of recurrence within DNNs does not inherently alter the mechanisms of information processing. Rather than fundamentally reshaping how inputs are processed, recurrent processing, in its current implementation, seems to operate similarly to added depth, as a tool adding to the computational power of a network. This is in line with our observation that despite vast differences in architecture, the approach to visual recognition, as indexed by confusion matrices, remains relatively homogeneous for a given model size (see Fig S11).

It is noticeable that each of the recurrent connection in our models can be implemented in multiple ways. Here, we adhered to the common implementation as it is used by architectures such as C and B, yet, there is a wide range of other options and frequent developments in the field ([Nayebi et al., 2022, Wang and Hu, 2022, Linsley et al., 2019, Fukui et al., 2019, Lotter et al., 2017, Lazar et al., 2009, Konkle and Alvarez, 2023]). In addition, even though the visual challenges that we created were motivated by previous research that referred to a potential specific role of lateral versus top-down connections ([Rajaei et al., 2019, Seijdel et al., 2021, Bar et al., 2006]), there might be better manipulations to be found. Overall, before we can accept the hypothesis that the type of recurrence does not matter, further studies are needed with more refined, biologically plausible implementations of recurrence and with additional visual challenges.

The need to search further in the space of visual manipulations also finds support in the wide variety we found across our 17 tasks. Interesting distinctions emerged, with seemingly similar tasks bringing surprisingly different results. The implementation of many small occluders, for instance, brought two such experimental conditions: *MS occluder, high* and *MS blobs, high*. While in the former, both humans and models see significant drops in their performance, the latter shows disagreement, with humans finding it much less difficult. A similar distinction can be found by comparing phase

scrambling in the *low pass* condition and *high pass* condition. Even in the absence of model-specific effects, our results highlight the behavioural variability that can be measured using a rich stimulus set.

In conclusion, contrary to our expectations, our findings highlight the limitations of recurrence as implemented in our models. Furthermore, they point out the overall discrepancy of visual processing in DNNs compared to humans. While we reproduce the general DNN ability to capture patterns of task difficulty, our results do not support that recurrent models outperform feedforward counterparts in capturing human visual recognition processes. As a consequence, we stress the importance of behavioural fidelity as a metric in developing new models, and emphasise the need for refined, biologically plausible implementations of recurrence.

Supporting information

Table 1: **General model information**. Extra information about models parameter size, connections (lateral/feedback) and references.

Table 2: **Custom model extra information: C V1-V1**. Layer specific information (type, input and output shape, parameters, trainability) about **C V1-V1** (CORnet V1-V1).

Table 3: **Custom model extra information: C IT-IT**. Layer specific information (type, input and output shape, parameters, trainability) about **C IT-IT** (CORnet IT-IT).

Table 4: **Custom model extra information: C IT-IT**. Layer specific information (type, input and output shape, parameters, trainability) about **CT** (CORnet T).

Table 5: **Custom model extra information: CLT**. Layer-specific information (type, input and output shape, parameters, trainability) about **CLT** (CORnet LT).

Table 6: **Custom model extra information: VGG11**. Layer-specific information (type, input and output shape, parameters, trainability) about **VGG11**.

Table 7: **Average performance per task per model**. Detailed average performance per model and per task. Acronyms: FL: *few large*, MS: *many small*.

Figure S9: **Colour-blind friendly performance pointplot**. A less colour-heavy reproduction of figure 6. The colour palette has been adapted for more contrast. Shape distinctions have been added as well.

Figure S10: **Model PCA plots**. PCA plots built from the category-wise confusion data of individual images, for each model and for human participants.

Figure S11: **Within-model confusion matrix correlations**. Correlation matrix (Pearson's R) of the pair-wise confusion matrix correlations of all models.

Author contributions

Conceptualization: Tim Maniquet, Hans Op de Beeck, Andrea Costantino

Data curation: Tim Maniquet, Andrea Costantino

Formal analysis: Tim Maniquet, Andrea Costantino

Funding acquisition: Tim Maniquet, Hans Op de Beeck

Investigation: Tim Maniquet, Andrea Costantino

Methodology: Tim Maniquet, Hans Op de Beeck, Andrea Costantino

Project administration: Hans Op de Beeck, Andrea Costantino

Resources: Hans Op de Beeck

Software: Tim Maniquet, Andrea Costantino

Supervision: Hans Op de Beeck, Andrea Costantino

Visualization: Tim Maniquet

Writing - original draft: Tim Maniquet, Hans Op de Beeck, Andrea Costantino

References

[Angelucci et al., 2017] Angelucci, A., Bijanzadeh, M., Nurminen, L., Federer, F., Merlin, S., and Bressloff, P. C. (2017). Circuits and Mechanisms for Surround Modulation in Visual Cortex. *Annual Review of Neuroscience*, 40(1):425–451.

[Avberšek et al., 2021] Avberšek, L. K., Zeman, A., and Op de Beeck, H. (2021). Training for object recognition with increasing spatial frequency: A comparison of deep learning with human vision. *Journal of Vision*, 21(10):14.

[Baizer et al., 1991] Baizer, J., Ungerleider, L., and Desimone, R. (1991). Organization of visual inputs to the inferior temporal and posterior parietal cortex in macaques. *The Journal of Neuroscience*, 11(1):168–190.

[Bar et al., 2006] Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hamalainen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., and Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, 103(2):449–454.

[Biscione and Bowers, 2023] Biscione, V. and Bowers, J. S. (2023). Mixed Evidence for Gestalt Grouping in Deep Neural Networks. arXiv:2203.07302 [cs].

[Bracci and Op de Beeck, 2023] Bracci, S. and Op de Beeck, H. P. (2023). Understanding Human Object Vision: A Picture Is Worth a Thousand Representations. *Annual Review of Psychology*, 74(1):113–135. _eprint: https://doi.org/10.1146/annurev-psych-032720-041031.

[Cadieu et al., 2014] Cadieu, C. F., Hong, H., Yamins, D. L. K., Pinto, N., Ardila, D., Solomon, E. A., Majaj, N. J., and DiCarlo, J. J. (2014). Deep Neural Networks Rival the Representation of Primate IT Cortex for Core Visual Object Recognition. *PLOS Computational Biology*, 10(12):e1003963. Publisher: Public Library of Science.

[Choi et al., 2018] Choi, H., Pasupathy, A., and Shea-Brown, E. (2018). Predictive Coding in Area V4: Dynamic Shape Discrimination under Partial Occlusion. *Neural Computation*, 30(5):1209–1257.

[DiCarlo et al., 2012] DiCarlo, J., Zoccolan, D., and Rust, N. (2012). How Does the Brain Solve Visual Object Recognition? *Neuron*, 73(3):415–434.

[Fel et al., 2022] Fel, T., Felipe, I., Linsley, D., and Serre, T. (2022). Harmonizing the object recognition strategies of deep neural networks with humans. arXiv:2211.04533 [cs].

[Felleman and Van Essen, 1991] Felleman, D. J. and Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex (New York, N.Y.: 1991)*, 1(1):1–47.

[Fukui et al., 2019] Fukui, H., Hirakawa, T., Yamashita, T., and Fujiyoshi, H. (2019). Attention Branch Network: Learning of Attention Mechanism for Visual Explanation. pages 10705–10714.

[Geirhos et al., 2018] Geirhos, R., Janssen, D. H. J., Schütt, H. H., Rauber, J., Bethge, M., and Wichmann, F. A. (2018). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv:1706.06969 [cs, q-bio, stat]*. arXiv: 1706.06969.

[Geirhos et al., 2022] Geirhos, R., Rubisch, P., Michaelis, C., Bethge, M., Wichmann, F. A., and Brendel, W. (2022). ImageNet-trained CNNs are biased towards texture; increasing shape bias improves accuracy and robustness. arXiv:1811.12231 [cs, q-bio, stat].

[Ghebreab et al., 2009] Ghebreab, S., Scholte, S., Lamme, V., and Smeulders, A. (2009). A Biologically Plausible Model for Rapid Natural Scene Identification. In *Advances in Neural Information Processing Systems*, volume 22. Curran Associates, Inc.

[Ghodrati et al., 2014] Ghodrati, M., Farzmahdi, A., Rajaei, K., Ebrahimpour, R., and Khaligh-Razavi, S.-M. (2014). Feedforward object-vision models only tolerate small image variations compared to human. *Frontiers in Computational Neuroscience*, 8.

[Goddard et al., 2016] Goddard, E., Carlson, T. A., Dermody, N., and Woolgar, A. (2016). Representational dynamics of object recognition: Feedforward and feedback information flows. *NeuroImage*, 128:385–397.

[Grill-Spector and Weiner, 2014] Grill-Spector, K. and Weiner, K. S. (2014). The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548.

[Groen et al., 2013] Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., and Scholte, H. S. (2013). From Image Statistics to Scene Gist: Evoked Neural Activity Reveals Transition from Low-Level Natural Image Structure to Scene Category. *Journal of Neuroscience*, 33(48):18814–18824.

[Hosseini et al., 2017] Hosseini, H., Xiao, B., Jaiswal, M., and Poovendran, R. (2017). On the Limitation of Convolutional Neural Networks in Recognizing Negative Images. arXiv:1703.06857 [cs, stat].

[Howard et al., 2018] Howard, J., and others (2018). fastai. GitHub. `https://github.com/fastai/fastai`

[Johnson and Olshausen, 2005] Johnson, J. S. and Olshausen, B. A. (2005). The recognition of partially visible natural objects in the presence and absence of their occluders. *Vision Research*, 45(25-26):3262–3276.

[Kar et al., 2019] Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., and DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, 22(6):974–983. Number: 6 Publisher: Nature Publishing Group.

[Khaligh-Razavi and Kriegeskorte, 2014] Khaligh-Razavi, S.-M. and Kriegeskorte, N. (2014). Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. *PLoS Computational Biology*, 10(11):e1003915.

[Kheradpisheh et al., 2016] Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. (2016). Humans and Deep Networks Largely Agree on Which Kinds of Variation Make Object Recognition Harder. *Frontiers in Computational Neuroscience*, 10.

[Kietzmann et al., 2019] Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K. A., Cichy, R. M., Hauk, O., and Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences*, 116(43):21854–21863. Publisher: National Academy of Sciences Section: PNAS Plus.

[Klink et al., 2017] Klink, P. C., Dagnino, B., Gariel-Mathis, M.-A., and Roelfsema, P. R. (2017). Distinct Feedforward and Feedback Effects of Microstimulation in Visual Cortex Reveal Neural Mechanisms of Texture Segregation. *Neuron*, 95(1):209–220.e3.

[Konkle and Alvarez, 2023] Konkle, T. and Alvarez, G. A. (2023). Cognitive Steering in Deep Neural Networks via Long-Range Modulatory Feedback Connections.

[Kravitz et al., 2013] Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., and Mishkin, M. (2013). The ventral visual pathway: an expanded neural framework for the processing of object quality. *Trends in Cognitive Sciences*, 17(1):26–49.

[Kreiman and Serre, 2020] Kreiman, G. and Serre, T. (2020). Beyond the feedforward sweep: feedback computations in the visual cortex. *Annals of the New York Academy of Sciences*, 1464(1):222–241.

[Kubilius et al., 2019] Kubilius, J., Schrimpf, M., Kar, K., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L. K., and DiCarlo, J. J. (2019). Brain-Like Object Recognition with High-Performing Shallow Recurrent ANNs. arXiv:1909.06161 [cs, eess, q-bio].

[Kubilius et al., 2018] Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., and DiCarlo, J. J. (2018). CORnet: Modeling the Neural Mechanisms of Core Object Recognition. Pages: 408385 Section: New Results.

[Ladret et al., 2023] Ladret, H. J., Cortes, N., Ikan, L., Chavane, F., Casanova, C., and Perrinet, L. U. (2023). Cortical recurrence supports resilience to sensory variance in the primary visual cortex. *Communications Biology*, 6(1):1–13. Number: 1 Publisher: Nature Publishing Group.

[Lamme and Roelfsema, 2000] Lamme, V. A. and Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579.

[Lamme et al., 1998] Lamme, V. A., Supèr, H., and Spekreijse, H. (1998). Feedforward, horizontal, and feedback processing in the visual cortex. *Current Opinion in Neurobiology*, 8(4):529–535.

[Lazar et al., 2009] Lazar, A., Pipa, G., and Triesch, J. (2009). SORN: a self-organizing recurrent neural network. *Frontiers in Computational Neuroscience*, 3.

[Lee and DiCarlo, 2023] Lee, M. J. and DiCarlo, J. J. (2023). How well do rudimentary plasticity rules predict adult visual object learning? *PLOS Computational Biology*, 19(12):e1011713.

[Liao and Poggio, 2020] Liao, Q. and Poggio, T. (2020). Bridging the Gaps Between Residual Learning, Recurrent Neural Networks and Visual Cortex. arXiv:1604.03640 [cs].

[Lin et al., 2014] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In Fleet, D., Pajdla, T., Schiele, B., and Tuytelaars, T., editors, *Computer Vision – ECCV 2014*, Lecture Notes in Computer Science, pages 740–755, Cham. Springer International Publishing.

[Linsley et al., 2019] Linsley, D., Kim, J., Veerabadran, V., and Serre, T. (2019). Learning long-range spatial dependencies with horizontal gated-recurrent units. arXiv:1805.08315 [cs].

[Lonnqvist et al., 2019] Lonnqvist, B., Clarke, A. D. F., and Chakravarthi, R. (2019). Crowding in humans is unlike that in convolutional neural networks. arXiv:1903.00258 [cs].

[Lotter et al., 2017] Lotter, W., Kreiman, G., and Cox, D. (2017). Deep Predictive Coding Networks for Video Prediction and Unsupervised Learning. arXiv:1605.08104 [cs, q-bio].

[Nayebi et al., 2022] Nayebi, A., Sagastuy-Brena, J., Bear, D. M., Kar, K., Kubilius, J., Ganguli, S., Sussillo, D., DiCarlo, J. J., and Yamins, D. L. K. (2022). Recurrent Connections in the Primate Ventral Visual Stream Mediate a Tradeoff Between Task Performance and Network Size During Core Object Recognition. Pages: 2021.02.17.431717 Section: New Results.

[O'Reilly et al., 2013] O'Reilly, R. C., Wyatte, D., Herd, S., Mingus, B., and Jilk, D. J. (2013). Recurrent Processing during Object Recognition. *Frontiers in Psychology*, 4.

[Paszke et al., 2019] Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems 32*, pages 8024–8035. Curran Associates, Inc.

[Pennartz et al., 2019] Pennartz, C. M., Dora, S., Muckli, L., and Lorteije, J. A. (2019). Towards a Unified View on Pathways and Functions of Neural Recurrent Processing. *Trends in Neurosciences*, 42(9):589–603.

[Rajaei et al., 2019] Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., and Khaligh-Razavi, S.-M. (2019). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLOS Computational Biology*, 15(5):e1007001.

[Riesenhuber, 2005] Riesenhuber, M. (2005). Object Recognition in Cortex: Neural Mechanisms, and Possible Roles for Attention. In *Neurobiology of Attention*, pages 279–287. Elsevier.

[Riesenhuber and Poggio, 2000] Riesenhuber, M. and Poggio, T. (2000). Models of object recognition. 3.

[Russakovsky et al., 2015] Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A.C., and Fei-Fei, L. (2015). ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*. doi: 10.1007/s11263-015-0816-y. Volume: 115, Number: 3, Pages: 211-252.

[Scholte et al., 2009] Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W. M., and Lamme, V. A. F. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. *Journal of Vision*, 9. Place: US Publisher: Assn for Research in Vision & Ophthalmology (ARVO).

[Seijdel et al., 2021] Seijdel, N., Loke, J., Klundert, R. v. d., Meer, M. v. d., Quispel, E., Gaal, S. v., Haan, E. H. F. d., and Scholte, H. S. (2021). On the Necessity of Recurrent Processing during Object Recognition: It Depends on the Need for Scene Segmentation. *Journal of Neuroscience*, 41(29):6281–6289. Publisher: Society for Neuroscience Section: Research Articles.

[Self et al., 2013] Self, M. W., van Kerkoerle, T., Supèr, H., and Roelfsema, P. R. (2013). Distinct Roles of the Cortical Layers of Area V1 in Figure-Ground Segregation. *Current Biology*, 23(21):2121–2129. Publisher: Elsevier.

[Shin et al., 2023] Shin, H., Ogando, M. B., Abdeladim, L., Durand, S., Belski, H., Cabasco, H., Loefler, H., Bawany, A., Hardcastle, B., Wilkes, J., Nguyen, K., Suarez, L., Johnson, T., Han, W., Ouellette, B., Grasso, C., Swapp, J., Ha, V., Young, A., Caldejon, S., Williford, A., Groblewski, P., Olsen, S., Kiselycznyk, C., Lecoq, J., and Adesnik, H. (2023). Recurrent pattern completion drives the neocortical representation of sensory inference. Pages: 2023.06.05.543698 Section: New Results.

[Simonyan and Zisserman, 2015] Simonyan, K. and Zisserman, A. (2015). Very Deep Convolutional Networks for Large-Scale Image Recognition. arXiv:1409.1556 [cs] version: 6.

[Smith, 2015] Smith, L. N. (2015). Cyclical learning rates for training neural networks. *Preprint at https://arxiv.org/abs/1506.01186*.

[Smith and Topin, 2017] Smith, L. N. and Topin, N. (2017). Super-convergence: Very fast training of neural networks using large learning rates. *arXiv preprint arXiv:1708.07120*, 6.

[Spoerer et al., 2020] Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., and Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Computational Biology*, 16(10):e1008215.

[Spoerer et al., 2017] Spoerer, C. J., McClure, P., and Kriegeskorte, N. (2017). Recurrent Convolutional Neural Networks: A Better Model of Biological Object Recognition. *Frontiers in Psychology*, 8:1551.

[Tang et al., 2018] Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Ortega Caro, J., Hardesty, W., Cox, D., and Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840.

[TorchVision maintainers and contributors, 2016] TorchVision maintainers and contributors (2016). TorchVision: PyTorch's Computer Vision library. BSD-3-Clause. `https://github.com/pytorch/vision`

[Ungerleider et al., 2008] Ungerleider, L. G., Galkin, T. W., Desimone, R., and Gattass, R. (2008). Cortical Connections of Area V4 in the Macaque. *Cerebral Cortex*, 18(3):477–499.

[van Bergen and Kriegeskorte, 2020] van Bergen, R. S. and Kriegeskorte, N. (2020). Going in circles is the way forward: the role of recurrence in visual inference. *Current Opinion in Neurobiology*, 65:176–193.

[Van Essen et al., 1986] Van Essen, D. C., Newsome, W. T., Maunsell, J. H., and Bixby, J. L. (1986). The projections from striate cortex (V1) to areas V2 and V3 in the macaque monkey: asymmetries, areal boundaries, and patchy connections. *The Journal of Comparative Neurology*, 244(4):451–480.

[Wang and Hu, 2022] Wang, J. and Hu, X. (2022). Convolutional Neural Networks With Gated Recurrent Connections. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(7):3421–3435. Conference Name: IEEE Transactions on Pattern Analysis and Machine Intelligence.

[Wyatte et al., 2014] Wyatte, D., Jilk, D. J., and O'Reilly, R. C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, 5.

[Yamins et al., 2014] Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., and DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624. Publisher: Proceedings of the National Academy of Sciences.

[Zhou et al., 2018] Zhou, B., Zhao, H., Puig Fernandez, X., Xiao, T., Fidler, S., Barriuso, A., and Torralba, A. (2018). Semantic Understanding of Scenes Through the ADE20K Dataset. *arXiv*. Accepted: 2020-06-11T20:32:21Z Publisher: Springer Nature.