Recurrent processing strengthens feature representations throughout visual cortex

Lea-Maria Schmitt (lea-maria.schmitt@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Kapittelweg 29
Nijmegen, The Netherlands

Floris P. de Lange (floris.delange@donders.ru.nl)
Donders Institute for Brain, Cognition and Behaviour, Radboud University, Kapittelweg 29
Nijmegen, The Netherlands

Abstract

Visual processing begins with a feedforward sweep that creates an initial perceptual representation, which is then refined through recurrent (lateral and feedback) processing. While recurrent signals in visual regions are abundant, their functional role in perceptual inference remains largely unclear. Using functional magnetic resonance imaging (fMRI) and artificial neural network (ANN) modeling, we aimed to examine whether recurrence modulates how images are represented in early and late visual regions. Participants were briefly presented with novel ambiguous images that were challenging to categorize. A visual mask followed these images either immediately or after a delay, thereby blocking or allowing for recurrent processing. We found that activity in early visual regions was best encoded by early layers of a convolutional neural network. These representations could no longer be observed when images were immediately masked. Conversely, activity in later ventral and dorsal visual regions was best encoded by later layers, and remained robust in the immediate masking condition. Comparing the brain alignment of ANNs with different recurrent dynamics revealed that activity in later ventral regions under delayed masking was best explained by a model with both lateral and feedback recurrence, suggestive of a role for recurrence in perceptual inference. The general strengthening of feature-specific representations with delayed masking likely reflects an interplay between early and late visual cortex, where lateral recurrence may help denoise low-level features to form more accurate high-level interpretations, which in turn may help disambiguate low-level features through feedback.

Keywords: perceptual inference; lateral and feedback recurrence; visual feature hierarchy

Introduction

We do not always immediately understand what we see. When objects are obscured by noise (e.g., Johnson & Olshausen, 2005) or otherwise deviate from what we commonly experience (e.g., Spaak et al., 2022), their categorization may require extended processing time. This behavioral phenomenon is mirrored in the brain: While object-selective neural responses to non-challenging visual input can be decoded as early as 100 ms after stimulus onset (Liu et al., 2009), decoding of more challenging input emerges substantially later (Tang et al., 2014). One interpretation of this extended processing time is that the initial feedforward sweep of information provides an immediate first interpretation of the input (VanRullen, 2008). However, this initial interpretation may be insufficient for challenging input, necessitating multiple iterations of recurrent processing to refine the interpretation over time (van Bergen & Kriegeskorte, 2020; Thorat et al., 2021).

Two distinct lines of research have investigated different computational aspects of recurrent processing in visual perception.

The first line demonstrates that for challenging images, high-level category information emerges only later in time (Tang et al., 2018; Kar et al., 2019; Rajaei et al., 2019). These studies found that behavioral and neural responses to challenging images are better predicted by ANNs incorporating lateral recurrence, a mechanism that enhances a signal within a brain region or network layer through integration across noisy neighbouring neurons (Lindsay, 2021). However, these studies primarily focused on high-level features at higher processing stages, leaving open questions about the role of early visual regions and their interactions with higher visual regions during perceptual inference. Understanding this interplay is crucial, as lateral recurrence in early visual areas has been hypothesized to enhance signal quality for subsequent higher-level processing, while higher-level representations might simultaneously shape early visual processing through top-down influences.

The second line demonstrates that representations in earlier visual regions can become tuned to higher-level features during later stages of processing (Schwiedrzik & Freiwald, 2017; Uran et al., 2022; Richter et al., 2024). This 'feature inheritance' could potentially be the result of feedback recurrence from higher to lower visual regions, thus biasing lower-level processing based on higher-level knowledge—-effectively testing hypotheses about object categories against incoming sensory input (Lee & Mumford, 2003).

Bringing together these two lines raises the question of whether and how recurrent computations within and between the different stages of visual processing may jointly solve object recognition for challenging input.

To address this question, we conducted an fMRI study examining how early and later visual regions represented chal-

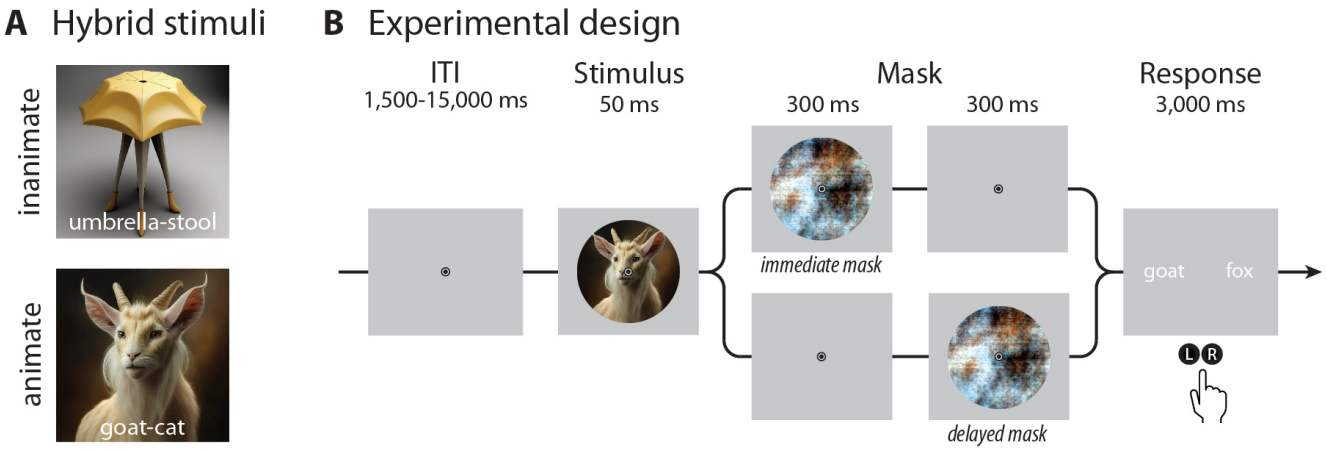

Figure 1: Stimuli and trial structure. **A)** Example images of hybrid objects created using Midjourney, combining either two inanimate or animate object categories. **B)** A 50 ms hybrid stimulus was followed by a 300 ms phase-scrambled mask, presented either directly (immediate mask) or after a 300 ms grey screen (delayed mask). On probe trials, participants indicated via button press which of two presented object categories matched the hybrid stimulus.

lenging visual input that could be processed either briefly (precluding recurrent interactions) or more extensively (allowing recurrent interactions), using a backward masking approach.

We analyzed stimulus features at different levels of abstraction, using the ANN AlexNet trained on ecologically relevant objects (Mehrer et al., 2021). Additionally, we analyzed stimulus features emerging from networks with different recurrent dynamics, using variants of the ANN BLT (Thorat et al., 2023). As challenging stimuli, we created ambiguous (animate or inanimate) images that combined features from two distinct categories (e.g., a goat-cat with slanted cat eyes and goat horns; Fig. 1A). These novel hybrid stimuli were designed to prevent rapid recognition based on prior knowledge. To manipulate the degree of recurrent processing, we employed backward masking using phase-scrambled images. These masks are thought to broadly activate visual cortex, thereby interfering with ongoing stimulus-specific processing (Macknik & Martinez-Conde, 2008). We reasoned that when the mask was presented shortly after image onset, recurrent processing is largely prevented, while delayed mask onset allows for recurrent processing to occur.

Results

Participants performed an object categorization task while undergoing fMRI. On each trial, they viewed an image containing either an animate or inanimate hybrid object for 50 ms. A phase-scrambled mask of 300 ms followed the hybrid stimulus either directly (immediate mask) or after a 300 ms grey screen (delayed mask). The temporal constraints of the visual system's neurophysiology and anatomy (Thorpe, 1990), combined with evidence from pharmacological perturbation (Kar & DiCarlo, 2021), suggest that feedback modulates stimulus processing in inferior temporal cortex after 150 ms. Since the immediate mask appears 50 ms after stimulus onset and requires 100 ms to reach high-level regions via the feedforward sweep, it arrives at these regions at the critical moment to disrupt emergent recurrent processing of the stimulus, whereas the delayed mask arrives only after some feedback processing has already occurred. Participants were occasionally prompted to indicate which of two object categories was present in the hybrid stimulus (Fig. 1B).

Delayed mask onset facilitates behavioural task performance

The categorization accuracy was well above chance level in both masking conditions (immediate mask: 70.2% accuracy, t_{30} = 18.47, $p < 0.001$, $BF_{10} > 100$; delayed mask: 81.5% accuracy, t_{30} = 21.12, $p < 0.001$, $BF_{10} > 100$), indicating that participants not only engaged with the task but also reliably perceived and categorized the hybrid stimuli. Delaying the mask onset by 300 ms improved both accuracy (11.2% increase, t_{30} = 7.23, $p < 0.001$, $BF_{10} > 100$; Fig. 2A) and response time (81 ms speed up, t_{30} = 3.89, $p < 0.001$, BF_{10} = 56.52; Fig. 2B). Importantly, since hybrid stimuli were presented for the same duration in both conditions, the performance benefits observed in the delayed mask condition suggest that the additional processing time before mask onset may have led to improved category representations.

Delayed mask onset increases visual cortical activity

Using participant-level general linear models, we estimated condition-specific cortical activity during immediate and delayed mask trials. A group-level ANOVA revealed stronger activation for delayed versus immediate masks across two broad clusters in bilateral visual regions (Fig. 3A), extending from V2 along both the ventral and dorsal streams. This widespread increase of visual activity aligns with the behavioral benefits of delayed masking, suggesting that the immediate mask successfully interrupted visual processing while the delayed mask

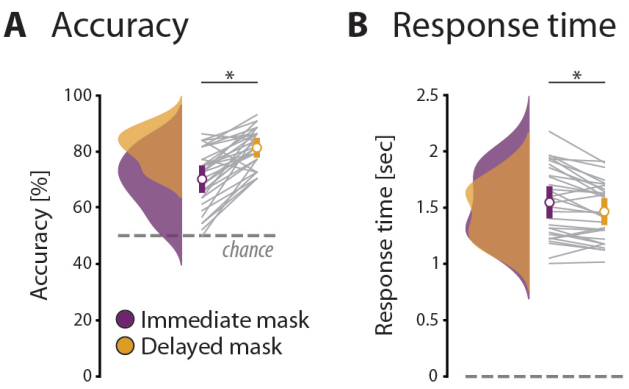

Figure 2: Object recognition performance for hybrid images. **A)** Accuracy and **B)** response time for immediate (purple) and delayed mask (orange) conditions. Density plots: participant distributions, grey lines: within-participant means, circles: means, error bars: confidence intervals, $*p < 0.001$.

permitted continued processing.

To quantify effects across the visual processing hierarchy, we selected five regions of interest (ROI; Fig. 3B) in line with the visual system as defined in the multimodal parcellation by Glasser et al. (2016): primary visual cortex (V1), early visual cortex (EVC; V2-V4), and ventral stream (V8 to ventromedial visual cortex), which process increasingly abstract visual features. Additionally, we included MT+ (V3CD to medial superior temporal area) and the dorsal stream (V3A to intraparietal sulcus) as control regions, which are less commonly implicated in processing of static images. From each ROI, we selected the 1,000 most image-responsive vertices per hemisphere, identified using an independent localizer experiment.

Analyzing mean activation over these image-sensitive vertices corroborated the whole-cortex findings, showing stronger activation for the delayed versus immediate mask condition across all ROIs ($p_{FDR} < 0.001$; Fig. 3C). V1 showed a small but significant modulation despite not being detected in the whole-cortex analysis, likely because cluster-based correction is less sensitive to the weaker and more diffuse activity difference in V1.

Delayed mask onset enhances feature tuning in visual cortex

To investigate what kind of visual information is present in the brain under different masking conditions, we extracted image representations at different levels of abstraction from AlexNet, a convolutional neural network that hierarchically processes low-level image features in early layers and high-level image features in late layers Cichy et al. (2016). For top image-responsive vertices of each ROI, we trained cross-validated ridge regression models to predict single-trial BOLD estimates from layer-specific AlexNet activations to hybrid images. Model performance was assessed by means of the neural predictivity of ANN features—referred to as brain score—by correlating predicted with observed BOLD estimates in held-out test data, averaged across top image-responsive vertices per ROI.

Across all layers, the delayed mask yielded higher brain scores than the immediate mask in all ROIs but MT+ (main effect of mask: $p_{FDR} < 0.05$; Fig. 4A), aligning with our univariate findings of enhanced visual processing for the delayed mask. This suggests that extended processing time enhanced the representations of visual features in both lower-order and higher-order visual regions.

For V1 and EVC, the immediate mask condition showed near-zero encoding accuracy across all layers, indicating an almost complete lack of stimulus information in early visual regions. In contrast, ventral and dorsal stream ROIs still contained significant stimulus information and tuning to high-level image features when the stimulus was masked immediately.

When the stimulus was masked after a delay, there was strong stimulus information in both early (V1 and EVC) and late (ventral and dorsal) visual regions. Activity in early visual regions could best be encoded based on early AlexNet layers, whereas the ventral and dorsal stream regions had peak performance for the final AlexNet layer.

Delayed mask onset improves ventral stream alignment with recurrent networks

To investigate whether the increase in feature tuning with the delayed mask can be attributed to recurrent processing, we compared brain scores across networks exhibiting different recurrent dynamics. For this purpose, we utilized variants of the object classification network BLT, developed and pretrained on ecologically relevant objects (Mehrer et al., 2021) by Thorat et al. (2023). BLT with its four hidden layers was chosen because networks with fewer layers have been shown to better capture hierarchical processing along the visual pathways (e.g., Nonaka et al., 2021). We included a network with only feedforward connections, a network with additional lateral (within-region) recurrence, a network with additional feedback (higher-to-lower-region) recurrence, and a network with both lateral and feedback recurrence. Activations for feedforward networks were extracted from the initial and only timestep, as these networks do not exhibit temporal dynamics. In contrast, activations for recurrent networks were extracted from the final 10th timestep to capture the influence of recurrence over extended processing time. Brain scores were evaluated on the networks' dedicated V1 layer for the V1 ROI, V2 layer for the EVC ROI, and IT layer for the ventral stream, MT+ and dorsal stream ROIs.

Focusing on the general benefit of no recurrence versus (lateral and feedback) recurrence, we found significant main effects of masking across all ROIs but MT+ ($p < 0.05$) and a significant main effect of network for the ventral stream ($F(1, 33) = 5.1$, $p = 0.031$). Specifically, the increase in brain scores for the recurrent network compared to the feedforward network under delayed masking, as shown in Fig. 4B, suggests that lateral and feedback recurrence together better account for activity in this higher visual region than a purely feedforward network.

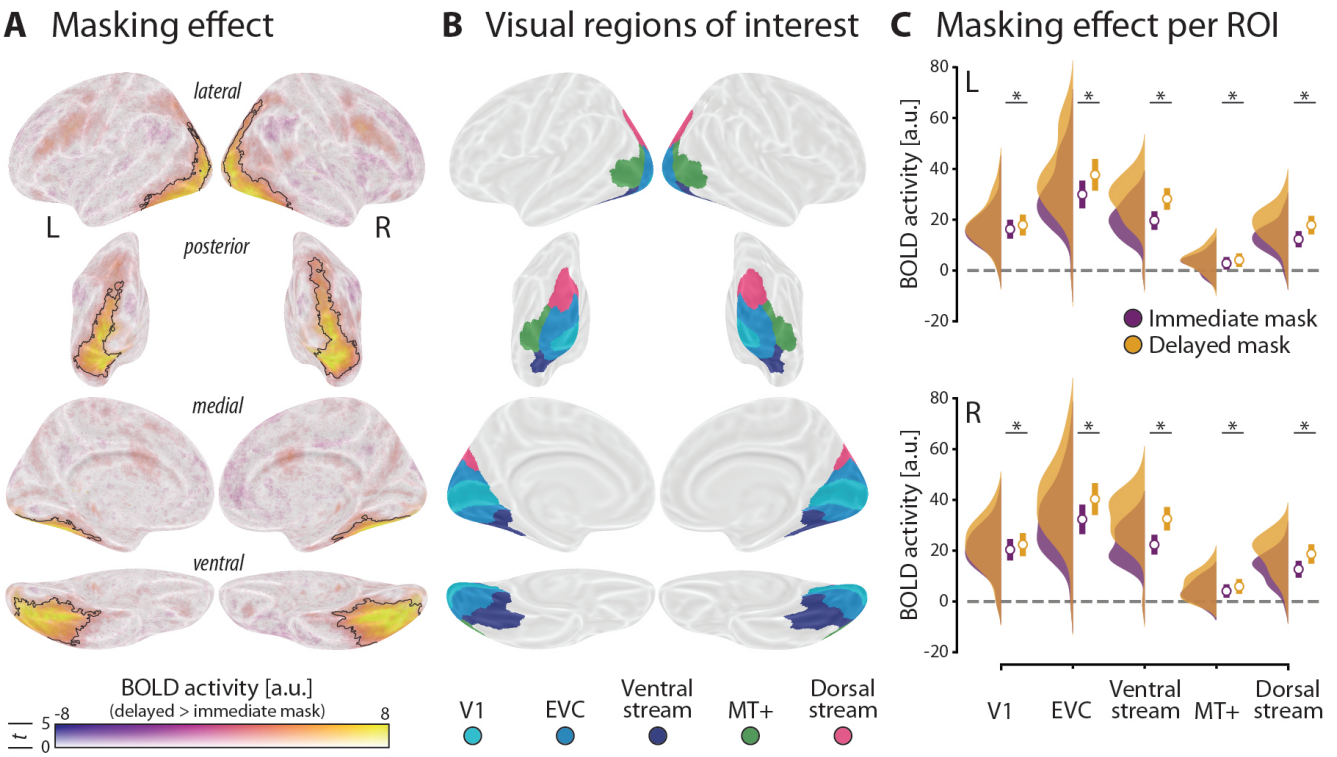

Figure 3: BOLD responses to immediate vs. delayed mask conditions. **A)** Whole-cortex contrast. Black outline: cluster at $p < 0.05$. **B)** Five visual cortex ROIs were defined from the parcellation by Glasser et al. (2016). **C)** Mean BOLD response in immediate (purple) and delayed (orange) mask conditions over top image-responsive vertices per ROI. Density plots: participant distributions; circles: means; error bars: $\pm SEM$; $^*p_{FDR} < 0.001$ between conditions.

Discussion

Visual recognition of objects under challenging conditions (e.g., occlusion) has been found to rely on recurrent processing in visual cortex (Tang et al., 2014). Previous studies have shown that visual cortex is modulated by both lateral connections within a region (Gilbert & Wiesel, 1989; Self et al., 2014) and feedback connections from later regions (Gilbert & Li, 2013; Roelfsema & de Lange, 2016), but the contribution of these connections to perceptual inference is not well understood. Here, we set out to elucidate whether and how recurrent processing transforms feature representations across the visual hierarchy.

Increased visual activity for delayed masking as a marker of recurrent processing

When stimuli were masked after a delay, compared to immediately, this led to a small activity increase in early visual and a large activity increase in later visual areas. As the only difference between immediate and delayed mask conditions pertained to the stimulus-onset asynchrony between stimulus and mask, with sensory input being otherwise identical across conditions (50 ms stimulus, 300 ms mask), we interpret this activity difference as reflecting sustained processing of stimulus information—a hallmark of recurrent processing. This interpretation is supported by our finding that networks with recurrent dynamics reflect activity in higher visual regions better than purely feedforward networks under delayed masking. Additionally, this aligns with previous electrophysiological evidence of sustained activity between image offset and mask onset (Bacon-Macé et al., 2005), neuroimaging studies showing increased visual activity with delayed mask onset (Green et al., 2005), and computational models demonstrating recurrent involvement in visual processing from 100 ms after image onset (Loke et al., 2022).

One concern with the mask design might be that an immediate mask disrupts the excitatory stimulus offset response typically seen for images, thereby interfering with bottom-up processing. Our design mitigated this concern by implementing a 300 ms interval between image and mask offset, shown to leave the offset response intact (Macknik & Livingstone, 1998). Another concern might be that the predictability of the delayed mask (inferred from the absence of an immediate mask) reduces its effectiveness by suppressing it as a distractor. However, our findings showed no significant differences in brain scores between immediate and delayed mask conditions when encoding AlexNet features of the mask instead of the stimulus ($p_{FDR} > 0.05$ for all ROIs), suggesting similar processing of the mask in both conditions with minimal differences in attention. The exclusive involvement of visual regions in the immediate versus delayed mask contrast further

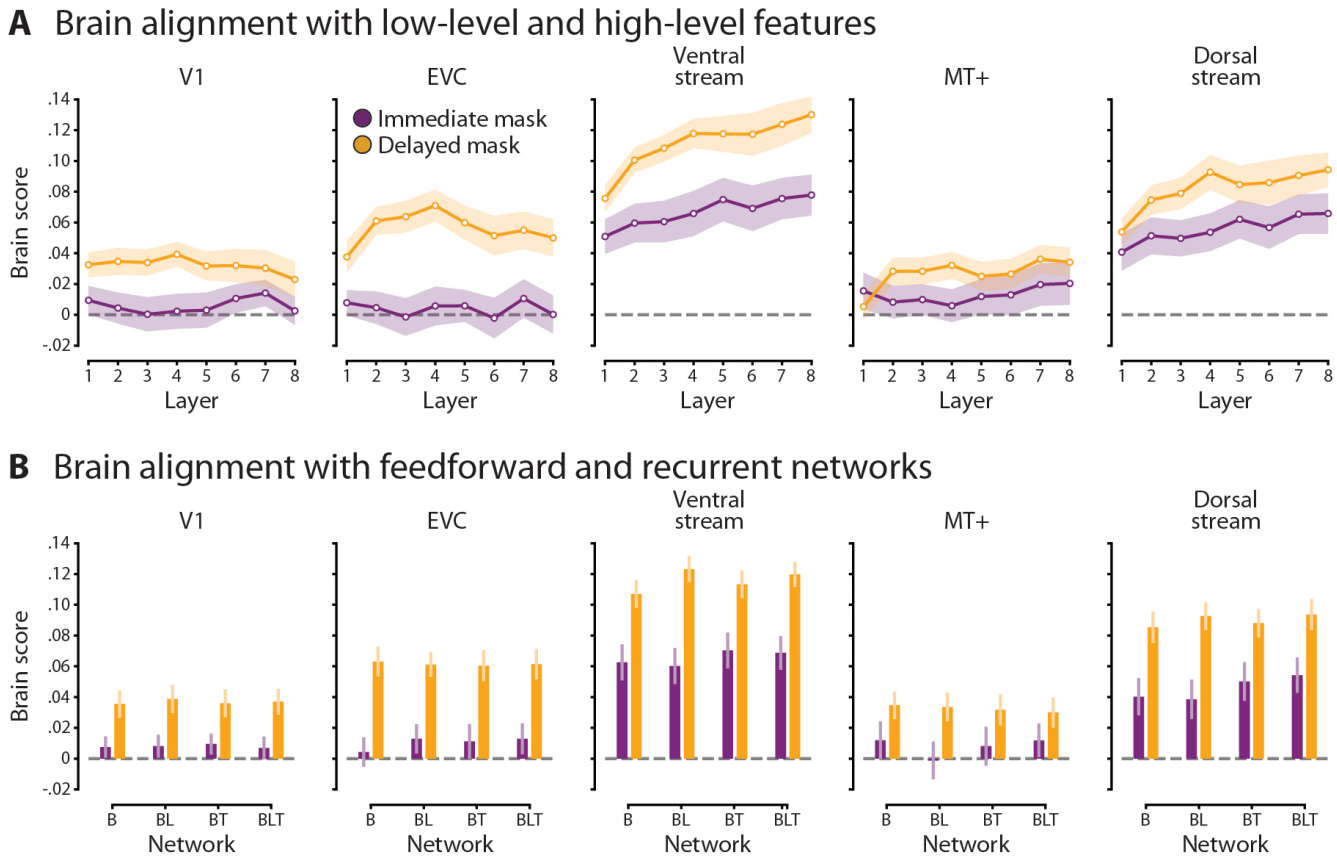

Figure 4: Cross-validated brain scores averaged across top image-responsive vertices using features from **A)** different AlexNet layers and **B)** BLT variants with different recurrent dynamics (B: feedforward connections, BL: feedforward and lateral recurrent connections, BT: feedforward and feedback recurrent connections, BLT: feedforward, lateral and feedback recurrent connections), shown for immediate (purple) and delayed (orange) mask conditions over ROIs. Error bands and bars: ±*SEM*.

indicates that our paradigm primarily engaged recurrent processes throughout the visual cortex rather than higher-order cognitive control.

Recurrent processing in early and later visual cortex

Unlike later visual regions, early visual regions showed no sensitivity to image-specific features when recurrent processing was interrupted immediately. This suggests that the mask primarily interfered with processing in early regions, which could be explained by our use of phase-scrambled masks that preserve low-level but not higher-level image statistics, potentially making them more disruptive to early visual regions. However, the robust brain scores in later visual regions and above-chance behavioral recognition even under immediate masking indicate that early visual regions must have initially represented low-level features, as feature and response processing in later regions build on this input. The brief and weak initial presence of features in early regions is likely obscured in the summed nature of the BOLD signal by the prolonged mask input.

The smaller, yet evident impact of the immediate mask on later visual regions suggests that these regions also benefited from continued recurrent processing. There are three potential mechanisms that could account for this benefit. First, lateral recurrence in these later regions could denoise higher-level representations, as proposed by Tang et al. (2018). Second, lateral recurrence in early regions may produce a denoised signal that is easier for later regions to process. Third, feedback recurrence may allow later regions to test their interpretations against early-region representations, enabling continuous refinement of perceptual inference.

One potential implication of the immediate mask mainly interfering with early regions could be that it also drives the behavioral drop in performance. This explanation is in line with recent research showing a link between masking effects in mouse V1 and behaviour (Gale et al., 2024), yet further research is needed to probe the importance of early-region recurrence for perceptual behaviour more explicitly.

Recurrent processing strengthens hierarchical feature encoding

When masking with a delay, early regions showed increased sensitivity to lower-level features and later brain regions to

higher-level features. This pattern mirrors the hierarchical feature encoding typically found along the visual pathway (Kravitz et al., 2013) and commonly observed when modelling brain responses with ANN representations (Mehrer et al., 2021). Therefore, the primary role of lateral and feedback recurrence appears to be improving feature representations over time, ultimately enabling correct object categorization.

In V1, we may not have observed the low-level feature encoding typically seen in fMRI and CNN layer alignment studies because our hybrid images were presented more briefly while being more challenging to process than the familiar objects used in previous research. For our stimuli, stronger feature tuning in V1 could require extended processing, involving sustained bottom-up drive and feedback from high-level brain regions, which may only reach V1 later.

Our results align with feedback recurrence models where high-level interpretations are translated to low-level features to test compatibility with sensory input, supported by studies showing an enhanced representation of low-level features that are predicted by the observer (Kok et al., 2012) or facilitated by semantic knowledge (Doerig et al., 2022). Other studies have found that early brain regions signal the error about a prediction made based on higher-level information (Schwiedrzik & Freiwald, 2017; Uran et al., 2022; Richter et al., 2024), showing feature inheritance. The apparent absence of such feature inheritance in our study might reflect differences in the underlying inference process. In earlier work, participants could typically develop strong expectations about upcoming stimuli, leading to robust top-down predictions of image content. In contrast, our experiment used ambiguous stimuli where participants could only iteratively build and revise expectations over time, potentially resulting in weaker and less precise top-down predictions.

Conclusions

We find that extended processing in the absence of input, resulting from recurrent processing, plays a crucial role in strengthening feature representations across the visual hierarchy, in both early and later visual regions, leading to improved perceptual performance. These findings suggest that the brain may achieve robust object recognition through dynamic interactions between different levels of visual processing and highlight the relevance of recurrence in visual cortex for behavior.

Methods

Participants

Thirty-four participants (25 female; age 19-34 years, M = 24) took part in the experiment. All participants reported no neurological or psychiatric disorders and had normal or corrected-to-normal vision. Participants gave written informed consent and received an expense allowance of €15/hour of testing. The study was conducted in accordance with the Declaration of Helsinki and approved by the METC Oost-Nederland ethics committee under the blanket approval for the protocol 'Imaging Human Cognition' (NL45659.091.14). Behavioral data from four participants were excluded due to technical issues.

Stimuli

We developed a set of 702 hybrid object images by combining two object categories into novel images (e.g., an umbrella-stool or goat-cat, Fig. 1A). Using the AI program Midjourney 5.1, we paired each object category within sets of 27 animate or 27 inanimate categories with every other one. The object categories were selected from a normative dataset of 1,200 concrete nouns (VanArsdall & Blunt, 2022). We included only object categories that were known by $>$ 95% of participants in previous norming studies and were primarily experienced through vision rather than other sensory modalities (Lynott et al., 2020). Any ambiguous, disturbing, unlimited, extinct, class-level, subordinate and synonymous object names were excluded. To ensure diverse object categories, we clustered object categories based on their semantic similarity using GloVe word embeddings (Pennington et al., 2014) and randomly selected one object category from each group of semantically similar words ($r > 0.45$). Object categories exceeding 2.5 *SD* in concreteness, familiarity, word frequency, valence, arousal (VanArsdall & Blunt, 2022), or perceptual strength (Lynott et al., 2020) were excluded.

An online validation study with 34 native English speakers (19 female; age 21-40 years, M = 32) assessed hybrid image quality. Nine participants rated each image on a 5-point Likert scale, indicating how likely each of the two constituent object categories was represented in the hybrid ('Not at all' to 'Extremely'). Images with median ratings ≥ 3 for both object categories were retained. The final stimulus set comprised 216 images, with each object category appearing in 8 different hybrids.

Mask stimuli were created by phase-scrambling hybrid images, which retains low-level image properties but renders hybrid objects unrecognizable.

Procedures

In the object recognition experiment, hybrid stimuli and masks were presented as spheres subtending 5° of visual angle while participants fixated a central bull's eye. On probe trials (5% of all trials per run), two object categories appeared on screen and participants indicated via button press with their right index finger (HHSC-2x4-C response pad, Current Designs) whether the category on the left or right matched the preceding hybrid image. The alternative category was randomly selected from the object categories of matching animacy. Participants had 3 s to respond. Inter-trial intervals followed a truncated exponential distribution ranging from 2 to 10 s (M = 3.1 s). The experiment was implemented using the Psychophysics Toolbox (v3.0.18) in MATLAB (R2022a, MathWorks).

The stimulus set comprised 216 images, evenly and randomly assigned to immediate and delayed mask conditions for every participant, with equal numbers of animate and inanimate objects in each condition. Each image appeared once

per run across the four experimental runs, yielding 864 total trials. Images were paired with a random mask. Prior to the experiment, participants were familiarized with half of the images from each masking condition over two days. Familiarity effects are not reported here.

Following the object categorization experiment, participants completed a localizer experiment. Each hybrid image from the stimulus set was presented once, with images grouped into blocks by animacy × mask condition. A block contained 19 images, with individual images displayed for 750 ms followed by a 250 ms grey inter-stimulus interval. Three blocks of phase-scrambled stimuli and three blocks of grey blank screens served as control conditions. To maintain attention, participants performed a one-back task, responding via button press to immediate image repetitions (5% of trials). The total experimental duration was 1.5 hours.

MRI data acquisition

MRI data were collected on a 3-T Siemens MAGNETOM Prisma or PrismaFit scanner using a 32-channel head coil. During the object recognition and localizer experiments, continuous whole-brain fMRI data were acquired using an echo-planar imaging sequence with a simultaneous multislice factor of 6 (repetition time (TR) = 1,000 ms, echo time (TE) = 34 ms, flip angle = 60°, voxel size = 2 mm isotropic, slice thickness = 2 mm, slice number = 66, slice orientation = transversal, slice order = interleaved, phase-encoding direction = anterior to posterior, field of view = 210 mm by 210 mm, bandwidth = 2090 Hz/Px).

Before the object recognition experiment, field maps were acquired with a gradient echo sequence (TR = 425 ms, TE1 = 2.2 ms, TE2 = 4.66 ms, phase-encoding direction = right to left, bandwidth = 843 Hz/Px, all other parameters as reported above). After the localizer experiment, anatomical images were acquired using a T1-weighted magnetization-prepared rapid gradient-echo sequence (TR = 2,300 ms, TE = 3.03 ms, flip angle = 8°, voxel size = 1 mm isotropic, slice thickness = 1 mm, slice number = 192, slice orientation = sagittal, phase-encoding direction = anterior to posterior, field of view = 256 mm by 225 mm, bandwidth = 130 Hz/Px).

MRI preprocessing

Data were preprocessed using fMRIPrep 23.2.0 (Esteban et al., 2019), which is based on Nipype 1.8.6 (Gorgolewski et al., 2011).

Anatomical data preprocessing. T1-weighted (T1w) images underwent intensity non-uniformity correction using ANTs' N4BiasFieldCorrection (Tustison et al., 2010) and skull-stripping using ANTs' antsBrainExtraction.sh workflow with OASIS30ANTs as target template. Brain tissue segmentation was performed using FSL's fast (Zhang et al., 2001). Cortical surfaces were reconstructed using FreeSurfer's recon-all (Dale et al., 1999). Brain masks were refined with a custom variation of the method to reconcile ANTs-derived and FreeSurfer-derived segmentations of the cortical gray matter of Mindboggle (Klein et al., 2017).

Functional data preprocessing. For each BOLD run, head-motion parameters were estimated with respect to the BOLD reference before any spatiotemporal filtering using FSL's mcflirt (Jenkinson et al., 2002). The estimated fieldmap was aligned to the BOLD reference with rigid-registration and field coefficients were then mapped onto the BOLD reference using the transform. The BOLD reference was then co-registered to the T1w reference using boundary-based registration with six degrees of freedom (Greve & Fischl, 2009) as implemented in FreeSurfer's bbregister. The nuisance time series derived from head motion estimates were expanded with their temporal derivatives and quadratic terms (Satterthwaite et al., 2013) as well as principal components from a thin band (or crown) of voxels around the edge of the brain (Patriat et al., 2017). BOLD time-series were resampled onto fsnative surfaces using FreeSurfer's mri_vol2surf and smoothed to 6 mm FWHM using AFNI's 3dBlurToFWHM.

Data analysis

fMRI univariate analyses. For the object recognition experiment, single-participant BOLD responses were modeled using vertex-wise general linear models (GLMs) for each run (AFNI's 3dREMLfit). Experimental conditions (familiarity × mask) were modeled as 650 ms events using canonical hemodynamic response functions. The model included nuisance regressors for presentation of task screens, button responses, head motion parameters with their quadratic terms and temporal derivatives, as well as edge components.

After transformation to surface standard space (*fsaverage*), participant-level coefficients were submitted to a group-level ANOVA (familiarity × mask × run; AFNI's 3dANOVA3). The main effect of the mask was cluster-corrected (formation threshold $p < 0.001$, significance threshold $p < 0.05$; AFNI's SurfClust).

For the localizer experiment, each experimental condition (animacy × mask conditions, and phase-scrambled condition) was modeled by a separate regressor. Within each regressor, individual image blocks were modeled as boxcar functions spanning the duration of the respective block. Apart from these details, the participant-level GLM analysis followed the specifications of the object recognition experiment. For each participant and ROI, we determined the 1,000 vertices showing the strongest image responsiveness, defined as the highest absolute mean beta coefficients across all animacy × mask conditions.

fMRI single-trial response estimation. Vertex-wise single-trial responses to images in the object recognition experiment were estimated using least-squares-sum regression (AFNI's 3dLSS). Each trial was modeled separately while all other trials were combined into a single regressor, requiring separate model estimation per trial. Apart from these details, the participant-level single-trial models followed the specifications of the multi-condition univariate analysis. Single-trial es-

timates were averaged across their four repetitions to reduce noise.

fMRI encoding models. To assess feature processing across visual ROIs, we used ANN unit activations evoked by hybrid images to model single-trial response estimates. For activations from each ANN layer of both AlexNet (trained on ecoset; Mehrer et al., 2021) and variants of BLT (Thorat et al., 2023), encoding models were estimated for the 1,000 most image-responsive vertices per ROI and hemisphere. Encoding models were fitted using ridge regression with four-fold cross-validation. Regularization parameters were optimized within each training fold using generalized cross-validation across 100 logarithmically spaced values (10^{-5} to 10^{8}) as implemented in scikit-learn's RidgeCV function (Pedregosa et al., 2011). The performance of encoding models, or brain score, was evaluated by correlating predicted with actual single-trial BOLD estimates in held-out test data, averaging correlation coefficients across folds and top image-responsive vertices per ROI.

Statistics. Group-level comparisons against zero or chance level were performed using one-sample t-tests, while condition differences were assessed using dependent-sample t-tests. Wilcoxon signed-rank tests were used when parametric assumptions were violated. Evidence for alternative hypotheses was quantified using Bayes factors (BF_{10}), with $BF_{10} > 100$ indicating extreme evidence.

Acknowledgments

This work was supported by a Consolidator Grant from the European Research Council (101000942) and a Vici Grant from the Dutch Research Council (VI.C.231.043) awarded to FPdL, as well as a Marie Skłodowska-Curie Actions Postdoctoral Fellowship from the European Commission awarded to LMS (101111402).

References

Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision Research*, *45*(11), 1459–1469.

Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*(1), 27755.

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *Neuroimage*, *9*(2), 179–194.

Doerig, A., Kietzmann, T. C., Allen, E., Wu, Y., Naselaris, T., Kay, K., & Charest, I. (2022). Semantic scene descriptions as an objective of human vision. *arXiv:2209.11737*, *10*.

Esteban, O., Markiewicz, C. J., Blair, R. W., Moodie, C. A., Isik, A. I., Erramuzpe, A., ... others (2019). fmriprep: a robust preprocessing pipeline for functional mri. *Nature Methods*, *16*(1), 111–116.

Gale, S. D., Strawder, C., Bennett, C., Mihalas, S., Koch, C., & Olsen, S. R. (2024). Backward masking in mice requires visual cortex. *Nature Neuroscience*, *27*(1), 129–136.

Gilbert, C. D., & Li, W. (2013). Top-down influences on visual processing. *Nature Reviews Neuroscience*, *14*(5), 350–363.

Gilbert, C. D., & Wiesel, T. N. (1989). Columnar specificity of intrinsic horizontal and corticocortical connections in cat visual cortex. *Journal of Neuroscience*, *9*(7), 2432–2442.

Glasser, M. F., Coalson, T. S., Robinson, E. C., Hacker, C. D., Harwell, J., Yacoub, E., ... others (2016). A multi-modal parcellation of human cerebral cortex. *Nature*, *536*(7615), 171–178.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: a flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, *5*, 13.

Green, M. F., Glahn, D., Engel, S. A., Nuechterlein, K. H., Sabb, F., Strojwas, M., & Cohen, M. S. (2005). Regional brain activity associated with visual backward masking. *Journal of Cognitive Neuroscience*, *17*(1), 13–23.

Greve, D. N., & Fischl, B. (2009). Accurate and robust brain image alignment using boundary-based registration. *Neuroimage*, *48*(1), 63–72.

Jenkinson, M., Bannister, P., Brady, M., & Smith, S. (2002). Improved optimization for the robust and accurate linear registration and motion correction of brain images. *Neuroimage*, *17*(2), 825–841.

Johnson, J. S., & 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, K., & DiCarlo, J. J. (2021). Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, *109*(1), 164–176.

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*(6), 974–983.

Klein, A., Ghosh, S. S., Bao, F. S., Giard, J., Häme, Y., Stavsky, E., ... others (2017). Mindboggling morphometry of human brains. *PLoS Computational Biology*, *13*(2), e1005350.

Kok, P., Jehee, J. F., & De Lange, F. P. (2012). Less is more: expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270.

Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G., & 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.

Lee, T. S., & Mumford, D. (2003). Hierarchical bayesian inference in the visual cortex. *Journal of the Optical Society of America A*, *20*(7), 1434–1448.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 2017–2031.

Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, *62*(2), 281–290.

Loke, J., Seijdel, N., Snoek, L., Van der Meer, M., Van de Klundert, R., Quispel, E., ... Scholte, H. S. (2022). A critical test of deep convolutional neural networks' ability to capture recurrent processing in the brain using visual masking. *Journal of Cognitive Neuroscience*, *34*(12), 2390–2405.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, *52*, 1271–1291.

Macknik, S. L., & Livingstone, M. S. (1998). Neuronal correlates of visibility and invisibility in the primate visual system. *Nature Neuroscience*, *1*(2), 144–149.

Macknik, S. L., & Martinez-Conde, S. (2008). The role of feedback in visual masking and visual processing. *Advances in Cognitive Psychology*, *3*(1-2), 125.

Mehrer, J., Spoerer, C. J., Jones, E. C., Kriegeskorte, N., & Kietzmann, T. C. (2021). An ecologically motivated image dataset for deep learning yields better models of human vision. *Proceedings of the National Academy of Sciences*, *118*(8), e2011417118.

Nonaka, S., Majima, K., Aoki, S. C., & Kamitani, Y. (2021). Brain hierarchy score: Which deep neural networks are hierarchically brain-like? *IScience*, *24*(9).

Patriat, R., Reynolds, R. C., & Birn, R. M. (2017). An improved model of motion-related signal changes in fmri. *NeuroImage*, *144*, 74–82.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... others (2011). Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

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*(5), e1007001.

Richter, D., Kietzmann, T. C., & de Lange, F. P. (2024). High-level visual prediction errors in early visual cortex. *PLoS Biology*, *22*(11), e3002829.

Roelfsema, P. R., & de Lange, F. P. (2016). Early visual cortex as a multiscale cognitive blackboard. *Annual Review of Vision Science*, *2*(1), 131–151.

Satterthwaite, T. D., Elliott, M. A., Gerraty, R. T., Ruparel, K., Loughead, J., Calkins, M. E., ... others (2013). An improved framework for confound regression and filtering for control of motion artifact in the preprocessing of resting-state functional connectivity data. *Neuroimage*, *64*, 240–256.

Schwiedrzik, C. M., & Freiwald, W. A. (2017). High-level prediction signals in a low-level area of the macaque face-processing hierarchy. *Neuron*, *96*(1), 89–97.

Self, M. W., Lorteije, J. A., Vangeneugden, J., van Beest, E. H., Grigore, M. E., Levelt, C. N., ... Roelfsema, P. R. (2014). Orientation-tuned surround suppression in mouse visual cortex. *Journal of Neuroscience*, *34*(28), 9290–9304.

Spaak, E., Peelen, M. V., & de Lange, F. P. (2022). Scene context impairs perception of semantically congruent objects. *Psychological Science*, *33*(2), 299–313.

Tang, H., Buia, C., Madhavan, R., Crone, N. E., Madsen, J. R., Anderson, W. S., & Kreiman, G. (2014). Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron*, *83*(3), 736–748.

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

Thorat, S., Aldegheri, G., & Kietzmann, T. C. (2021). Category-orthogonal object features guide information processing in recurrent neural networks trained for object categorization. *arXiv:2111.07898*.

Thorat, S., Doerig, A., & Kietzmann, T. C. (2023). Characterising representation dynamics in recurrent neural networks for object recognition. *arXiv preprint arXiv:2308.12435*.

Thorpe, S. J. (1990). Spike arrival times: A highly efficient coding scheme for neural networks. *Parallel Processing in Neural Systems*, 91–94.

Tustison, N. J., Avants, B. B., Cook, P. A., Zheng, Y., Egan, A., Yushkevich, P. A., & Gee, J. C. (2010). N4itk: improved n3 bias correction. *IEEE Transactions on Medical Imaging*, *29*(6), 1310–1320.

Uran, C., Peter, A., Lazar, A., Barnes, W., Klon-Lipok, J., Shapcott, K. A., ... Vinck, M. (2022). Predictive coding of natural images by v1 firing rates and rhythmic synchronization. *Neuron*, *110*(7), 1240–1257.

VanArsdall, J. E., & Blunt, J. R. (2022). Analyzing the structure of animacy: Exploring relationships among six new animacy and 15 existing normative dimensions for 1,200 concrete nouns. *Memory & Cognition*, *50*(5), 997–1012.

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

VanRullen, R. (2008). The power of the feed-forward sweep. *Advances in Cognitive Psychology*, *3*(1-2), 167.

Zhang, Y., Brady, M., & Smith, S. (2001). Segmentation of brain mr images through a hidden markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, *20*(1), 45–57.