The representational nature of spatio-temporal recurrent processing in visual object recognition

Siying Xie [1*], Johannes Singer [1], Bati Yilmaz [1], Daniel Kaiser [2,3+], Radoslaw M. Cichy [1,4,5+]

1 Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany
2 Mathematical Institute, Department of Mathematics and Computer Science, Physics, Geography, Justus Liebig University Gießen, Germany
3 Center for Mind, Brain and Behavior (CMBB), Philipps-University Marburg, Justus Liebig University Gießen and Technical University Darmstadt, Germany
4 Berlin School of Mind and Brain, Faculty of Philosophy, Humboldt-Universität zu Berlin, Berlin, Germany
5 Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany

* Correspondence to siying.xie@outlook.com

+ These authors contributed equally

KEYWORDS
Backward masking, human visual system, representational dynamics, naturalistic object images, deep neural networks, brain rhythms, electroencephalography, functional magnetic resonance imaging

Summary

The human brain orchestrates object vision through an interplay of feedforward processing in concert with recurrent processing. However, where, when and how recurrent processing contributes to visual processing is incompletely understood due to the difficulties in teasing apart feedforward and recurrent processing. We combined a backward masking paradigm with multivariate analysis on EEG and fMRI data to isolate and characterize the nature of recurrent processing. We find that recurrent processing substantially shapes visual representations across the ventral visual stream, starting early on at around 100ms in early visual cortex (EVC) and in two later phases of around 175 and 300ms in lateral occipital cortex (LOC), adding persistent rather than transient neural dynamics to visual processing. Using deep neural network models for comparison with the brain, we show that recurrence changes the feature format in LOC from predominantly mid-level to more high-level features. Finally, we show that recurrence is mediated by four distinct spectro-temporal neural components in EVC and LOC, which span the theta to beta frequency range. Together, our results reveal the nature and mechanisms of the effects of recurrent processing on the visual representations in the human brain.

Introduction

Human visual object recognition is orchestrated by the interplay of feedforward and recurrent computations. Anatomically, this is evident by the number of feedforward and feedback connections complementing each other in visual circuits [1–3]. Functionally, the feedforward sweep brings in information from the retina, enabling core object recognition through basic visual analysis [4,5]. Then, the recurrent computations begin right after the first influx of feedforward information into the cortex. Recurrent activity contributes to object recognition not only when the viewing conditions are challenging [6–12], but also when objects are in plain view [13–15].

While the existence and importance of both feedforward and recurrent computations in object recognition is undoubted, their exact nature, i.e., where, when and how they affect visual processing remains incompletely understood [16–19]. This is partly because their empirical dissection is challenging: shortly after the first feedforward sweep, feedforward and recurrent activity overlap in space and time [20–22], hindering their unique characterization.

Here, we used the classical experimental protocol of backward masking [23–26] to isolate the role of recurrent from feedforward activity [27–31]. In backward masking a salient visual mask is shown shortly after a target image, impacting recurrent activity related to the target while leaving feedforward activity unaffected [28,32–34]. Thus, the comparison of brain activity when participants view masked versus unmasked target images isolates the contribution of recurrent activity.

We recorded human brain activity with EEG and fMRI to resolve visual responses in time and space when a set of naturalistic object stimuli were either backward masked or not. We then used multivariate pattern analysis [35–37] to recover the neural representations of the image contents under the different masking regimes across time and space.

Comparing the neural activity related to the target images in the masked and unmasked conditions, we determined where, when and how recurrent activity contributes to human object vision. We first identified and characterized the spatio-temporal dynamics of visual recurrent activity. We then determined its respective spectral bases using time-frequency decomposition [38–41], and finally resolved its resulting visual feature format by relating neural representations to artificial neural network [42–44].

Results

We presented 24 images of everyday objects on real-world backgrounds (Fig. 1A) to human participants while recording their brain activity with EEG (N = 31) and fMRI (N = 27) in separate sessions. On each trial, the target image was backward masked in one of two masking conditions: early mask or late mask (Fig. 1B). In the early mask condition, a dynamic mask rapidly followed the target after 17ms. The rapid succession of target and mask yields effective backward masking that disrupts recurrent processing [28,32–34]. In contrast, in the late mask condition, the mask appeared after a delay of 600ms, leaving recurrent processing unaffected across an extended time window while otherwise keeping the stimulation across the whole trial the same.

We used a multivariate pattern analysis framework to assess visual object representations captured by EEG and fMRI [36,45] to classify the objects in the target images from brain data. Because target images and masks were statistically independent by design across trials,

classifying objects isolated the neural activity that related to the target from neural activity related to the mask.

We then characterized and compared object representations across the early mask and late mask conditions, revealing the temporal, spatial, and spectral characteristics as well as the representational format of the recurrent aspects of visual processing.

The temporal dynamics of recurrent visual activity

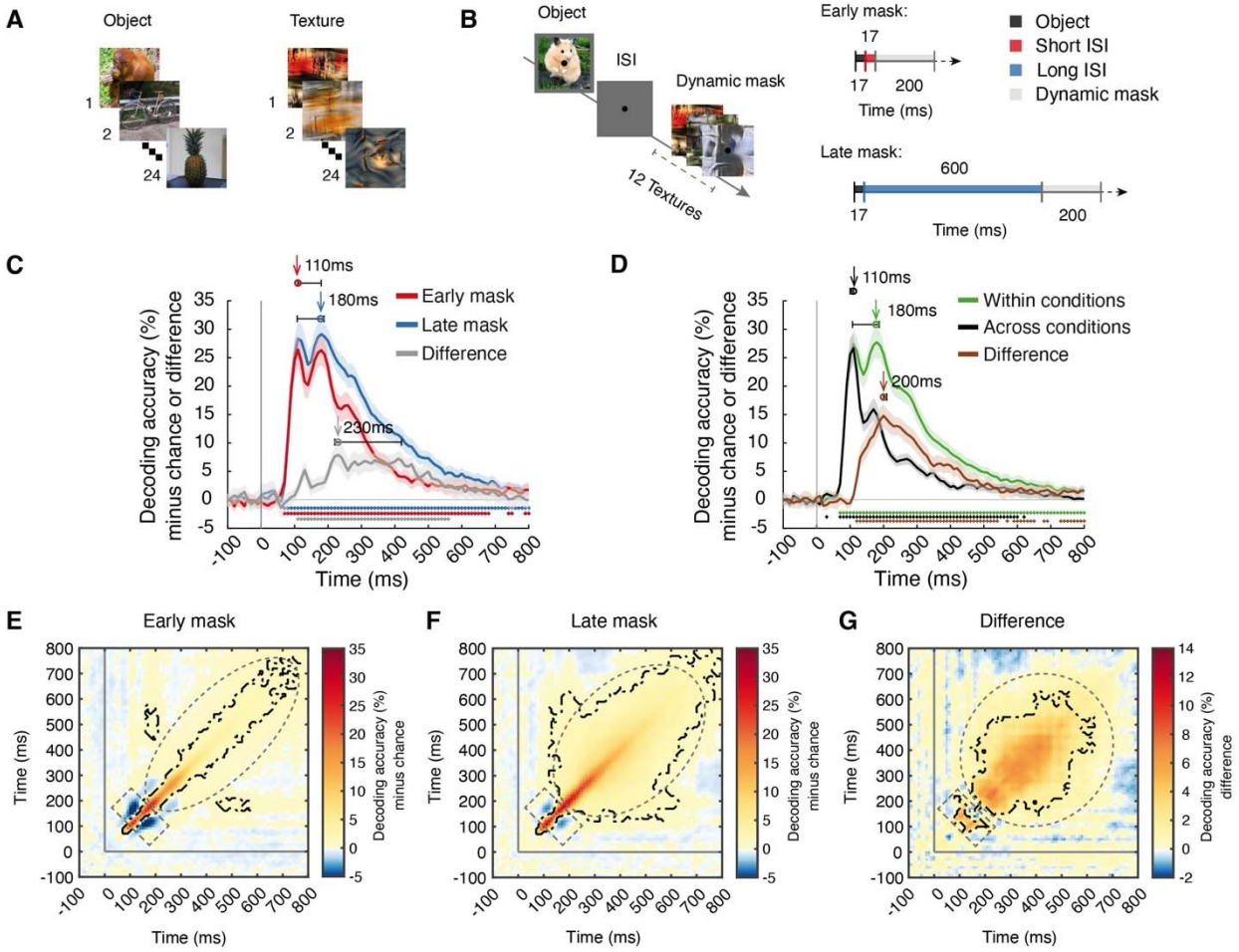


Figure 1 Experimental design and temporal dynamics of visual object representations. **(A)** Stimulus set. We used 24 real-world object images on natural backgrounds as target stimuli and 24 synthesized image textures created from an additional set of real-world object images for dynamic masks. **(B)** Experimental paradigm and timing parameters. On each trial, a briefly shown target object image was backward masked by a dynamic mask (i.e., a sequence of image textures) in one of two conditions: the early mask condition (short 17ms ISI) and the late mask condition (long 600ms ISI). **(C)** Results of object identity decoding in the early mask (red) and late mask (blue) conditions and their difference (gray). **(D)** Results of object identity decoding within (green) and across (black) masking conditions and their difference (brown). For **(C, D)**, chance level is 50%; significant above-chance level decoding is denoted by colored asterisks at the corresponding time points (N = 31, $p < 0.05$, right-tailed permutation tests, cluster definition threshold $p < 0.005$, cluster-threshold $p < 0.05$, 10,000 permutations); vertical gray line at 0ms indicates stimulus onset; shaded margins of time courses indicate 95% confidence intervals of the

decoding performance determined by bootstrapping (1,000 iterations); horizontal error bars indicate 95% confidence intervals for peak latencies. **(E-G**) Results of time-generalized decoding object identity in the **(E)** early mask condition, **(F)** late mask condition, and **(G)** the difference. For **(E-G)**, chance level is 50%; time-point combinations with significantly above-chance level decoding are outlined in black dashed lines (N = 31, right-tailed permutation tests, cluster definition threshold $p < 0.005$, cluster-threshold $p < 0.05$, 10,000 permutations); vertical and horizontal gray lines indicate stimulus onset.

To reveal the temporal dynamics of object representations in the early mask and late mask conditions, we conducted time-resolved multivariate pattern classification of object identity using EEG data. Classifying between all pairs of the 24 object images and averaging across pairs yielded a grand average object decoding time course for both masking conditions (Fig. 1C, for statistical details, see Supplementary Table 1). We assessed statistical significance using cluster-based inference (N = 31, right-tailed permutation tests, cluster definition threshold $p < 0.005$, cluster-threshold $p < 0.05$, 10,000 permutations), and report peak latencies as time points at which objects are best discriminated by neural representations with 95% confidence intervals derived by bootstrapping (1,000 samples) in brackets.

We observed a qualitatively similar and typical results pattern [45,46] in both masking conditions. Decoding accuracies fluctuated around baseline until 70ms after image onset, when they steeply rose to two peaks at ~100ms and ~200ms. The peak latencies for the objects in the early mask condition (110ms [110 – 180ms]) and the late mask condition (180ms [110 – 190ms]) coincided with the first and second peak respectively, without being significantly different ($p > 0.05$, Supplementary Table 1). This result demonstrates the presence of robust visual information in both masking conditions, warranting further analysis.

Comparing the decoding performance between the two masking conditions, we observed higher decoding in the late mask condition emerging after the first decoding peak (Fig. 1C, gray curve, cluster 110 – 560ms, peak latency 230ms [220 – 420ms]). This pattern was also present when decoding objects across the categorical boundary defined by naturalness or animacy (Supplementary Fig. 1A, B, Supplementary Table 2). Together, this provides a first characterization of the timing of recurrent activity.

The modest difference in the time-resolved decoding result patterns between the early and the late mask conditions might be interpreted as indicating a relatively minor role of recurrent processing in visual object processing. However, this conclusion is premature: similar overall time courses might hide qualitatively different visual representations across the two masking conditions.

To investigate whether the representations are strongly affected by recurrent processing, we decoded object identity across the two masking conditions [47,48]. The rationale is that if visual representations are only weakly affected by recurrent processing, decoding results should be similar for the decoding within- and across-masking conditions. However, if recurrent processing affects visual representations more strongly, the across-condition decoding accuracy should be lower than the within condition accuracy. We found that cross-decoding was strongly reduced after the first peak (110ms [100 – 110ms], Fig. 1D, black curve) when compared to decoding within each masking condition (Fig. 1D, green curve, corresponds to the average of the blue and red curve in Fig. 1C). The difference between within- and across-conditions was significant after the first within-condition decoding peak (Fig. 1D, brown curve, clusters between 120ms and

800ms), with a peak at 200ms (200 – 210ms) (for statistical details, see Supplementary Table 1). This result pattern was also obtained when comparing within- and across-conditions decoding for training the classifiers on either the early or the late mask condition (Supplementary Fig. 1C, D, Supplementary Table 2). This indicates that recurrent processing strongly affects visual object representations after the first feedforward sweep from 120ms onward, thus detailing the temporal dynamics of recurrent processing.

If recurrent processing strongly affects visual object representations, the dynamics with which those representations emerge should also differ depending on the amount of recurrent activity involved. To assess this, we used temporal generalization analysis (TGA) [49] by decoding object identity across all time-point combinations in the EEG epoch. This resulted in time-time matrices for each masking condition (Fig. 1E, F) and their difference (Fig. 1G).

We observed similarities and differences for the two masking conditions. Concerning the similarities, in both masking conditions, significant effects were present from ~70ms onwards, and decoding accuracies were highest close to the diagonal (i.e., similar time-points for training and testing), indicating that fast-evolving, transient representations dominate the neural dynamics. Further, we also observed significant off-diagonal generalization from 150ms on in both masking conditions, indicating the additional presence of stable and persistent representations. This shows that in both masking conditions, visual processing depends on both transient and persistent representations.

However, we also observed two key differences. For one, there was more widespread temporal generalization in the late mask than in the early mask condition (Fig. 1E, F, indicated by the length of the minor axis of the striped ellipses), and this difference was significant (Fig. 1G, striped ellipse). This suggests a stronger presence of persistent representations due to recurrent processing in the late mask condition. Second, we observed that below-chance decoding accuracies at the time-point combinations, i.e., ~100ms and ~200ms (Fig. 1E-G, striped rectangle), were lower in the early mask condition than in the late mask condition, emerging as a positive difference in their comparison (Fig. 1G, striped rectangle).

Negative decoding accuracies in off-diagonal regions of the TGA can be interpreted as an amplitude reversal of a time-locked (i.e., to the onset of the image) oscillatory neural component with a half-cycle length corresponding to the time difference between the time-points of the combination at which the negative decoding occurs [50,51] (for graphical illustration, see Supplementary Fig. 2). Assuming that in the early mask condition recurrent processing is reduced while feedforward processing is unaffected. This links feedforward activity to time-locked oscillatory components that are covered by time-varying recurrent activity in the late mask condition. In turn, in the early mask condition recurrent activity is reduced, and the time-locked feedforward-related oscillatory activity is uncovered. This result pattern was confirmed when comparing the decoding between within-condition of the late mask and the cross-decoding (Supplementary Fig. 3A-C), and it was reversed when comparing the decoding between the within-condition of the early mask and the cross-decoding (Supplementary Fig. 3D-F), supporting our interpretation.

Together, our results provide three key insights into the temporal dynamics of recurrent visual processing: firstly, recurrent processing affects visual object representations from ~100ms onward, after the first feed-forward sweep, and most strongly around 200ms; secondly, it

contributes specifically to the emergence of persistent representations; thirdly, it is less phase-locked to the onset of the stimulus than feed-forward activity.

The spatial profile of recurrent visual activity

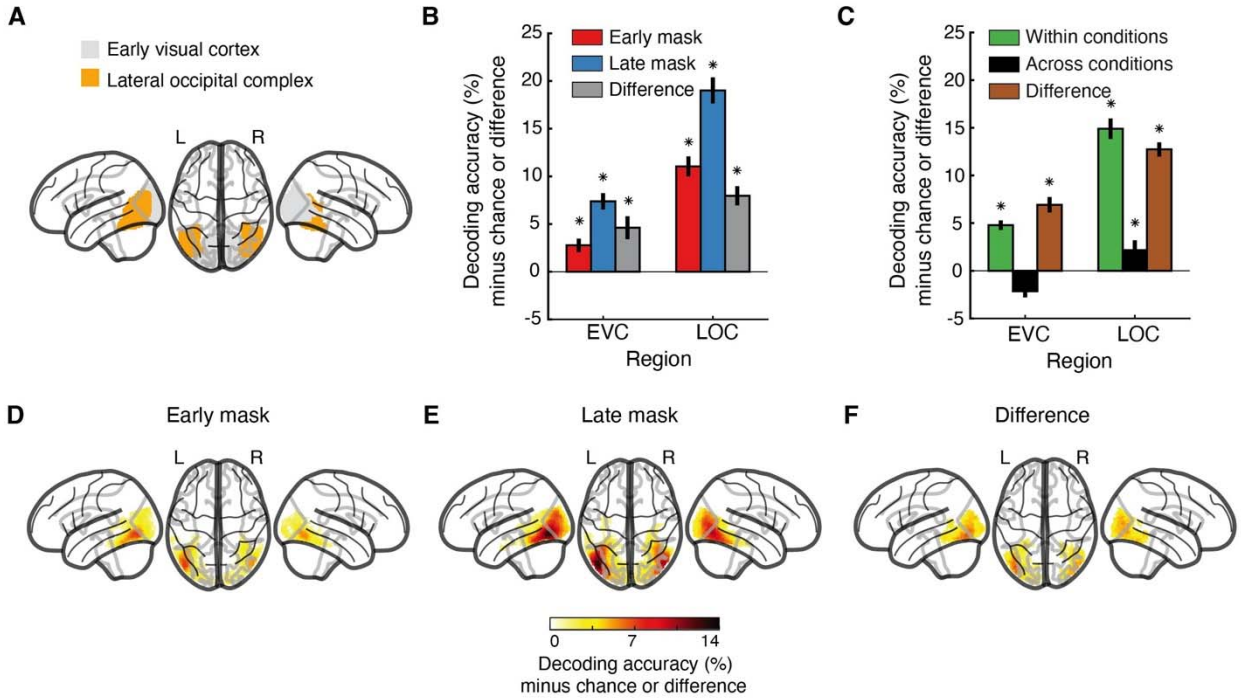


Figure 2 Cortical locus of visual object representations.
(A) Visualization of the early visual cortex (EVC, i.e., V1, V2, and V3 combined) and the lateral occipital complex (LOC) ROIs. **(B)** Results of object identity decoding in the early mask condition, the late mask condition, and their difference. **(C)** Results of object identity decoding within and across masking conditions and their difference. For **(B, C)**, chance level is 50%; significant above-chance level decoding is denoted by black asterisks above the bars (N = 27, $p < 0.05$, right-tailed permutation tests, FDR-corrected); error bars indicate standard errors of the mean. **(D-F)** Results of the spatially unbiased searchlight decoding in the **(D)** early mask condition, **(E)** late mask condition, and **(F)** the difference. For **(D-F)**, chance level is 50%; only voxels with significant above-chance level decoding are shown (N = 27, right-tailed permutation tests, cluster definition threshold $p < 0.005$, cluster-threshold $p < 0.05$).

Next, we determined the spatial profile of recurrent processing across the visual brain. For this, we used an equivalent multivariate pattern analysis scheme and comparison strategy between masking conditions as for the temporal dynamics but applied in a spatially resolved way to fMRI data.

We focused on two regions of interest (ROI) in the visual ventral stream: the early visual cortex (i.e., V1, V2, and V3 combined) as the entry point of visual information in the cortex [52,53] and the lateral occipital complex (LOC) (Fig. 2A) as a central high-level hub for object representations [54–56]. We decoded object identity in both masking conditions (Fig. 2B) as well as across masking conditions (Fig. 2C) and compared the results (N = 27, sign-permutation tests, FDR-corrected, $p < 0.05$).

In line with the EEG results, there was above-chance decoding of object identity in both ROIs in both masking conditions (Fig. 2B, blue and red bars, all ROI-results FDR-corrected). Further comparing masking conditions, we found higher decoding accuracies for the late mask condition in EVC and LOC (Fig. 2B, gray bars), indicating that recurrent processing affects representations in both regions.

Akin to the EEG analysis, we next determined the degree to which recurrent activity alters visual representations. For this, we compared the within-condition decoding results (Fig. 2C, black bars) to the across-conditions results (Fig. 1D, green bars), noting their difference (Fig. 2C, brown bars). In both ROIs, the decoding accuracy was strongly reduced when decoding across masking conditions. In LOC, but not EVC, there was low but significant cross-decoding accuracy. An equivalent results pattern emerged when comparing within- and across-conditions decoding for training the classifiers on either the early or the late mask condition (Supplementary Fig. 4A, B). This indicates that recurrent activity strongly impacts visual representations in both EVC and LOC.

To explore the differences between the two masking conditions across the whole brain, we used a spatially unbiased fMRI searchlight analysis [57,58]. Consistent with the ROI results, we found object identity information across the ventral visual stream in both masking conditions (Fig. 2D, E, right-tailed permutation tests, cluster definition threshold $p < 0.005$, cluster-threshold $p < 0.05$, 5,000 permutations). Comparing decoding in the early mask versus the late mask conditions revealed widespread effects in the ventral stream with a maximum in the high-level ventral cortex (Fig. 2F). This reinforces the view that recurrent activity strongly affects visual representations across the ventral stream.

Recurrent processing affects the format of visual representations

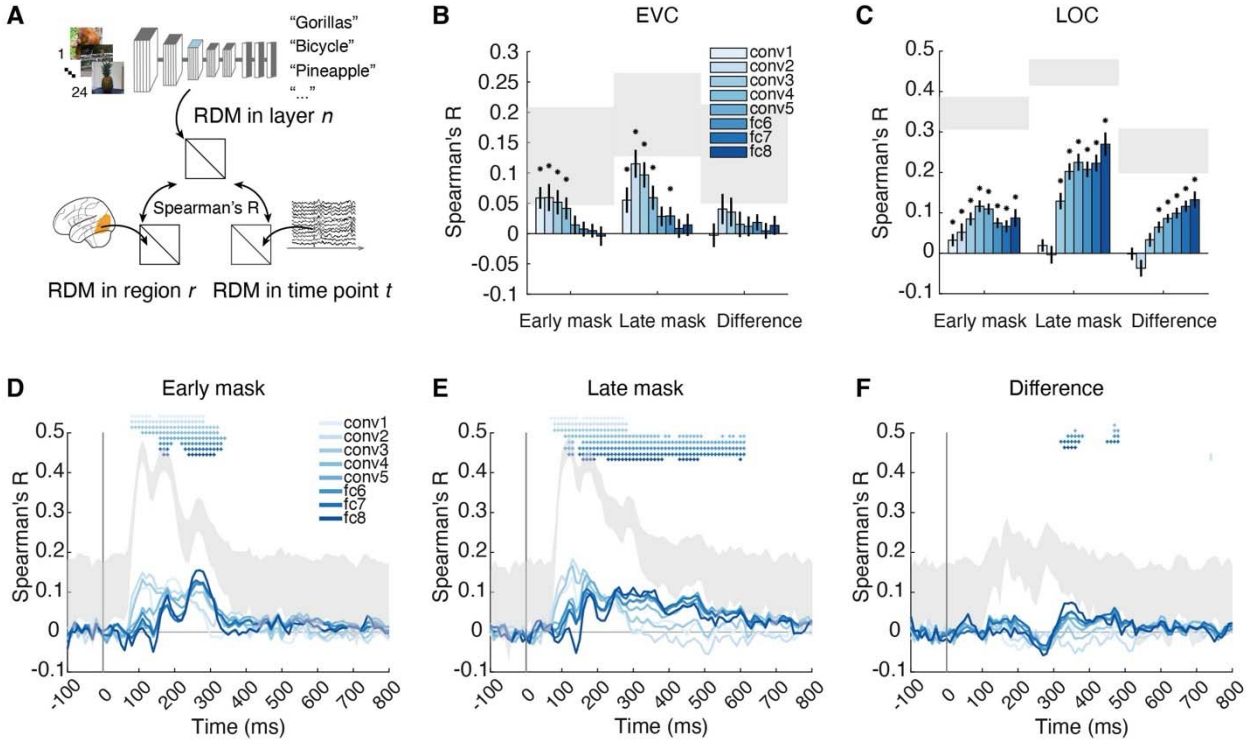


Figure 3 The representational format of visual representations resolved in space or time. **(A)** RSA linking brain responses to layer-wise activation patterns in a DNN model (AlexNet trained on object categorization). We obtained RDMs for each layer of the DNN, each ROI in fMRI, and each time point in EEG. We then compared (Spearman's R) the DNN RDMs with the EEG and fMRI RDMs, respectively. **(B, C)** RSA results linking **(B)** EVC and **(C)** LOC to DNN layers. For **(B, C)**, significant correlations are marked by black asterisks above bars (N = 27, $p<0.05$, right-tailed permutation tests, FDR corrected); error bars depict standard errors of the mean; shaded gray areas indicate the noise ceiling. **(D-F)** RSA results linking DNNs to EEG in the **(D)** early mask condition, **(E)** late mask condition, and **(F)** difference therein. For **(D-F)**, significant correlations at time points are denoted by asterisks colored by layer (N = 31, right-tailed permutation tests, cluster definition threshold $p < 0.005$, cluster-threshold $p < 0.05$, 10,000 permutations); horizontal error bars indicate 95% confidence intervals for peak latencies, shaded gray areas represented the noise ceiling.

We next investigated how recurrent processing affects the format of visual representations. For this, we used representational similarity analysis (RSA) [37,61] to compare representations in the brain and in the layers of an 8-layer AlexNet deep neural network (DNN) model trained on object categorization [59,60] (Fig. 3A). The rationale is that correspondence to layers along the DNN hierarchy reveals the complexity of the representational format in the brain, from low-complexity features in the early layers to high-complexity features in the late layers [42–44].

We began the investigation of the format of visual representations as localized in EVC and LOC using fMRI. In EVC, we identified the strongest correspondences with the early to middle DNN layers in both masking conditions (Fig. 3B). The differences between masking conditions were numerically most pronounced in the early layers, albeit not significantly. This suggests that feedforward and recurrent processing in EVC primarily involve the processing of low-level features. A supplementary analysis that compared the visual representations as revealed by the within- and across-conditions decoding to the DNN model showed an equivalent result pattern (Supplementary Fig. 5A), further strengthening this view.

In contrast, in LOC, we made two key observations. First, while there were correspondences with the middle to deep layers in both masking conditions (Fig. 3C), there was a shift in the peak correspondence from the highest layer (i.e., layer 8) in the late mask condition to a middle layer (i.e., layer 4) in the early mask condition. Correspondingly, the comparison of results between masking conditions revealed differences in mid-to-late layers, with a peak in the latest layer. Second, there was correspondence in LOC to early layers (i.e., layers 1-2) in the early mask condition but not in the late mask condition. Both observations were also present when comparing within- and across-conditions decoding results (Supplementary Fig. 5B). This suggests that in LOC the representational format shifts from lower to higher complexity through recurrent activity.

Next, we assessed the change in the representational format of visual representations across time using EEG. We observed correspondence to all layers of the DNN in both masking conditions (Fig. 3D, E) with a temporal progression in peak correspondence from lower layers early in time to higher layers later in time [62,63] (for statistical details, see Supplementary Table 3). This shows that in both masking conditions, visual representations emerge along a cascaded processing hierarchy characterized by increasing feature complexity [5,17,64,65]. To assess the feature complexity and timing of recurrent processing directly, we determined the difference in correspondence between the masking conditions (Fig. 3F). We found that the difference was highest in the middle and high layers between ~300ms and 500ms. This indicates that recurrent

activity changes the representational format to one of higher complexity features. Consistent with this conclusion, equivalent results patterns were observed in a supplementary analysis comparing the visual representations revealed by the within- to across-conditions decoding to the DNN model (Supplementary Fig. 5D).

Finally, for both EEG- and fMRI-based analyses, we confirmed the main results pattern using another DNN architecture (i.e., ResNet50 [66], Supplementary Fig. 6), indicating the generalizability of the conclusions across models.

Together, this shows that recurrent processing leaves the format of EVC representations unaffected in terms of visual feature complexity. In contrast, recurrent processing changes the format of LOC and late representations from lower to higher complexity, revealing the nature of its effect on the representational format.

The spatiotemporal dynamics of changes in representational format through recurrence

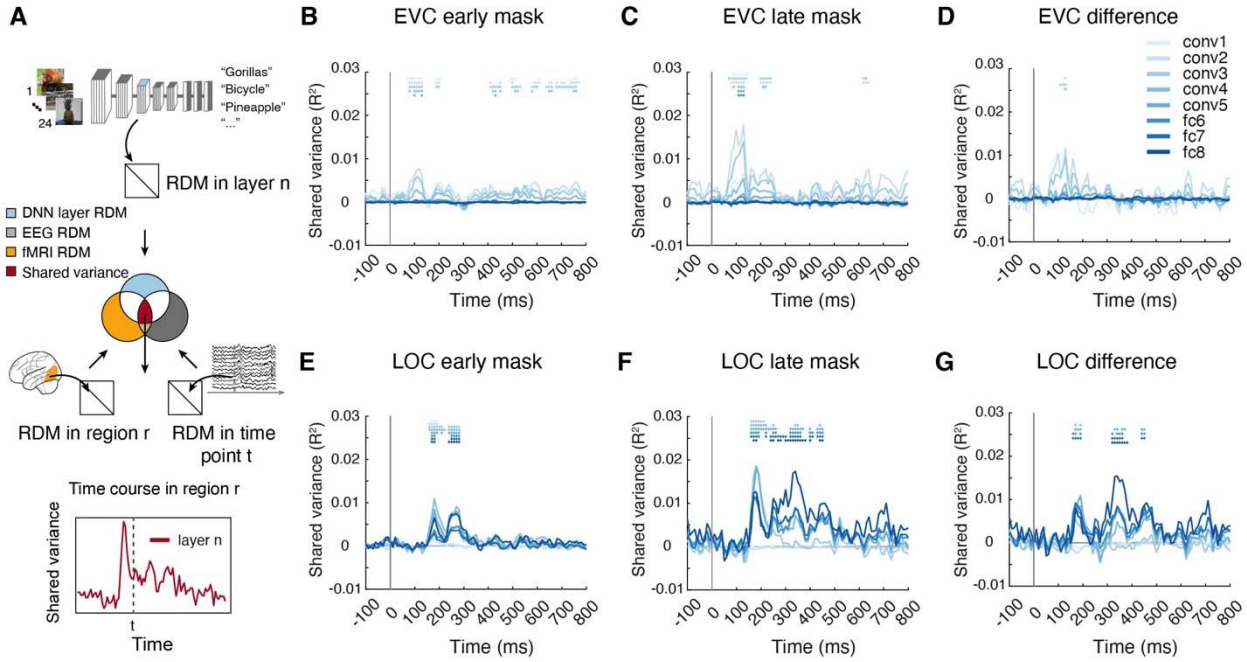


Figure 4 The format of spatiotemporally resolved visual representations.
(A) Commonality analysis based on RSA, linking temporal dynamics (EEG), cortical locus (fMRI) and feature complexity (DNN layers of AlexNet). This yielded time courses of shared variance for each DNN layer in EVC and LOC respectively (here: layer 3 in LOC). **(B-G)** Time courses of shared variance with DNN features in the **(B, E)** early mask condition, **(C, F)** late mask condition, and **(D, G)** difference between them, in EVC **(B-D)** and LOC **(E-G)** respectively. For **(B-G)**, significant effects at time points are denoted by asterisks color-coded by DNN layer (N = 31, right-tailed permutation tests, cluster definition threshold $p < 0.005$, cluster-threshold $p < 0.05$, 10,000 permutations).

Visual processing evolves dynamically across spatial locations in the brain and across time simultaneously, necessitating a spatiotemporally resolved view [35,67]. However, the analyses so far assessed visual representations and their format separately in space and time. For a fully spatio-

temporally resolved view, we used RSA-based commonality analysis [68,69] (Fig. 4A), providing time courses of shared variance with each DNN layer in EVC and LOC (for statistical details, see Supplementary Table 5).

In EVC, we observed an emergence of visual representations of low- to mid-complexity with peaks early in time, predominantly at 120 – 130ms, both in the early mask condition and the late mask condition (DNN layers 1-6, Fig. 4B, C). The difference between masking conditions emerged early (peaks at ~90 – 130ms) and was in low-to-middle complexity, too (DNN layers 1-5, Fig. 4D). This shows that recurrent activity impacts visual representations in EVC early in time and in a low-to-mid-complexity format.

In LOC, we observed the emergence of visual representations of all complexity levels at a later stage than in EVC, with two peaks at ~ 200ms and 300ms in both masking conditions (Fig. 4E, F). The difference between masking conditions was in features of middle-to-high complexity (DNN layers 4-8, Fig. 4G). This shows that recurrent activity impacts visual representations in LOC later in time and in a mid-to-high complexity format.

In sum, recurrent activity modulates EVC representations early in processing in low-to-mid complexity format, and LOC representations later in processing in mid-to-high complexity format.

The spectro-temporal basis of recurrent processing

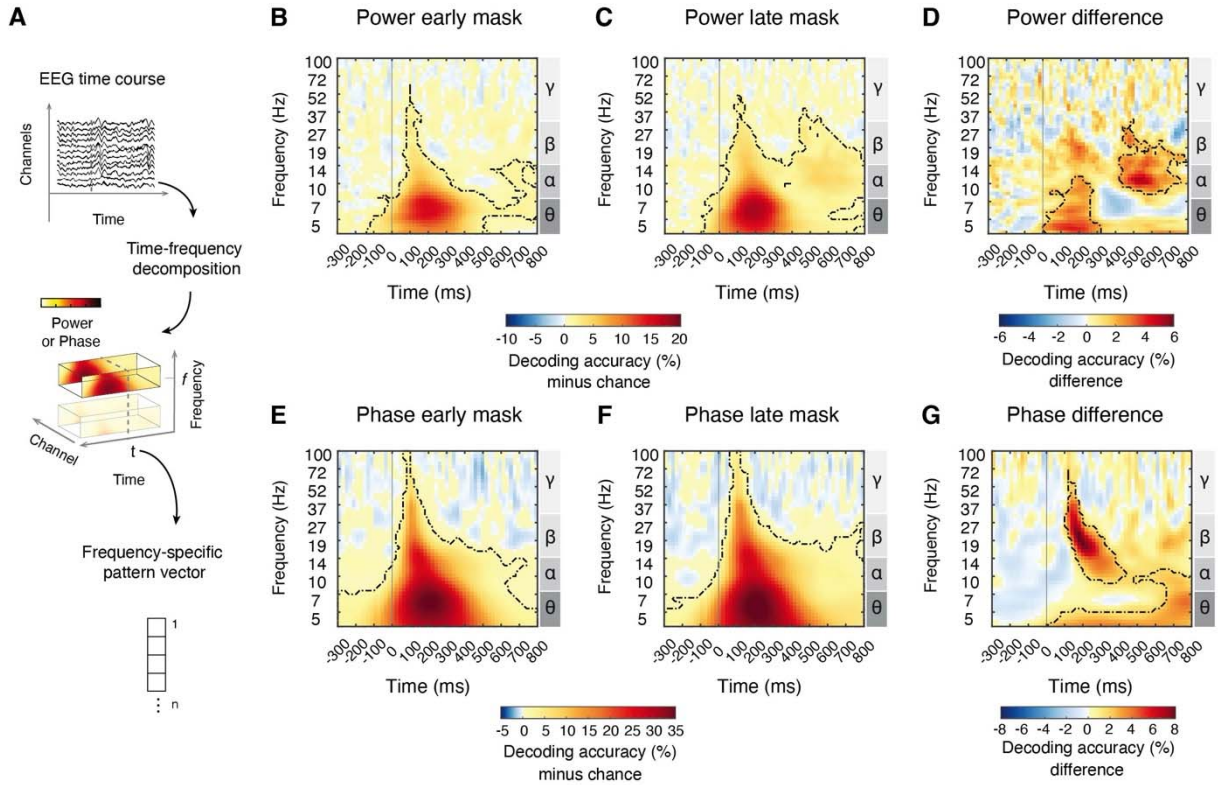


Figure 5 Spectral characteristics of visual representations.
(A) Using time-frequency decomposition we extracted frequency-specific response pattern vectors across EEG channels for power (63) and phase values (63 × 2 = 126) separately. **(B-G)** Results of time- and frequency-resolved object identity decoding in the **(B, E)** early mask condition and **(C, F)** late mask

condition, and **(D, G)** difference between them, based on power values **(B-D)** and phase values **(E-G)**. For (**B-G)**, chance level was 50%; time-frequency combinations with significant above-chance decoding are outlined by black dash lines (N = 31, right-tailed permutation tests, cluster definition $p < 0.05$, significance $p < 0.05$, 10,000 permutations); the vertical gray line indicates stimulus onset, and the right y-axis labels indicate frequency bands.

The transmission of visual information and the formation of visual representations [70,71] is fundamentally indexed by oscillatory neural activity. Based on previous work in human and non-human primates, we hypothesized that recurrent processing should be evident in the low-frequency range between the theta- and the beta-range [72–74]. Thus, in the next step, we investigated the spectral characteristics of visual processing in the early mask condition and the late mask condition. For this, we decoded object identity from EEG data resolved both in time and frequency (Fig. 5A), considering power and phase of the signals separately.

Across both masking conditions and for both power (Fig. 5B, C) and phase (Fig. 5E, F), we observed significant object decoding in a broad frequency range. The decoding peak was consistently within the theta band (~6 Hz) at ~200ms (for statistical details see Supplementary Table 6). This establishes the sensitivity of the analysis and warrants further inspection by contrasting the masking conditions.

Comparing the results of the early mask condition to the late mask condition, we observed four components with distinct temporal and spectral characteristics (Fig. 5D, G; for statistical details see Supplementary Table 6). Two clusters were in the power domain (Fig. 5D) and two in the phase domain (Fig. 5G). In detail, in the power domain, there was a cluster before 300ms in the theta-alpha frequency range (peak at 4.27 Hz, 160ms), and a later cluster after 400ms in the alpha-beta frequency range (peak at 10.72 Hz, ~540ms). In the phase domain, there was a cluster between 100ms and 400ms in the alpha-beta frequency range (peak at 19.35 Hz, 200ms) and a cluster in the theta range across the entire temporal range after stimulus onset (peak at 10.03 Hz, 560ms). A supplementary analysis comparing the within- and across-conditions decoding (Supplementary Fig. 7) revealed a more widespread effect that largely encompassed the clusters observed here.

Together, this establishes the spectro-temporal basis underlying recurrent visual processing as four distinct components with specific spectro-temporal profiles.

Feature format and cortical origin of the spectral components underlying recurrent processing

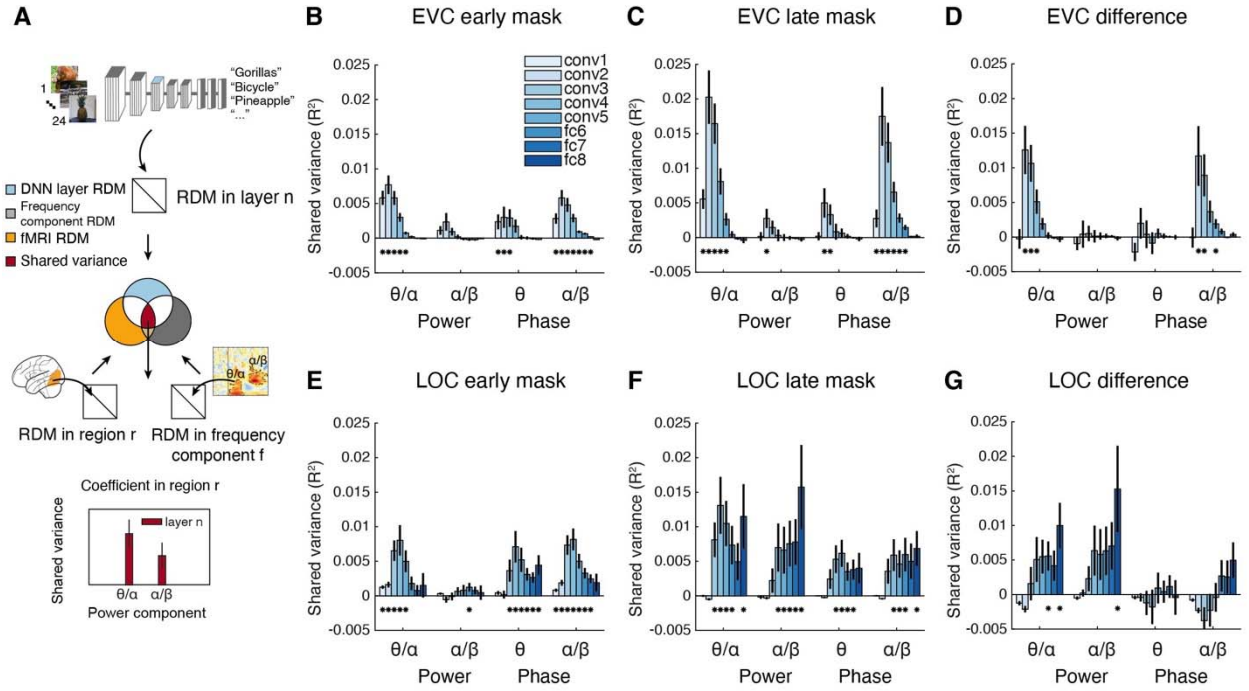


Figure 6 Spatial profile and feature format of the spectral components underlying recurrent processing.
(A) Commonality analysis based on RSA linking identified time-frequency resolved dynamics (EEG), cortical locus (fMRI) and feature complexity (DNN layers of AlexNet). This analysis yielded coefficients of shared variance for each of the four identified components and for each DNN layer in EVC and LOC respectively. **(B-G)** Shared variance in the **(B, E)** early mask condition, **(C, F)** late mask condition, and **(D, G)** difference between them, in EVC **(B-D)** and LOC **(E-G)** respectively. For **(B-G)**, significant effects at DNN layers are denoted by asterisks (N = 31, right-tailed permutation tests, cluster definition threshold $p < 0.005$, cluster-threshold $p < 0.05$, 10,000 permutations).

In a final step, we asked for each of the four spectro-temporally identified components: what is their spatial origin in the brain, and in what feature format do they represent object information? To address both questions, we again used RSA-based commonality analysis [68,69], relating frequency-based EEG signals to fMRI signals from EVC and LOC and layers of the DNN model (Fig. 6A).

We first note results common to all four components, forming the basis for further discriminative investigation. First, we observed significant relationships to DNN layers for all components, regions, and both masking conditions (Fig. 6 B, C and Fig. 6 E, F; except for EVC and the alpha-beta power component in the early mask condition, see Fig. 6B), demonstrating the analytical feasibility of the approach. Second, similar to the previous analyses, the shared variance was generally lower in the early mask condition compared to the late mask condition.

However, key differences emerged when isolating recurrent processing by considering the difference between the early and late mask conditions with respect to the feature complexity of visual representations (Fig. 6D, G). Concerning EVC (Fig. 6D), we found a relationship between

the two components emerging early in time, that is, the theta-alpha power component and the alpha-beta phase component, but not for the other two clusters. This relationship was strongest in low-and mid-level visual features. In contrast, concerning LOC (Fig. 6G), we found a relationship between the two power-defined components, one early in time and the other late in time, strongest for high-level visual features.

Together, these results reveal the spectro-temporal basis of recurrent processing in EVC and LOC by comprehensively characterizing its distinct spectrally identified components, in terms of their specific feature complexity and temporal profile.

Discussion

Summary

We combined a backward masking paradigm with multivariate analysis on EEG and fMRI data, along with computational modelling, to characterize when, where and how recurrent processing affects object representations. Harvesting the detailed structure of visual representations beyond grand-average responses to visual stimulation, we showed that recurrence substantially affects the image-specific geometry of visual representations.

First, regarding the spatiotemporal dynamics, we found that recurrence affects visual representations across the ventral visual stream, early on at ~100ms in EVC and in two later phases of ~175 and 300ms in LOC, adding persistent rather than transient neural dynamics to visual processing. Next, we determined the feature complexity and spectral basis of the effect of recurrence on visual representations. We found that recurrence changes the feature format in LOC from mid- to high-level features and is mediated by four distinct spectro-temporal components in EVC and LOC in the theta to beta frequency range.

The spatio-temporal dynamics of recurrent processing

Our separate analyses of EEG and fMRI data revealed a broad impact of recurrent processing: it affects visual representations starting at 100–120ms, with a peak at 200ms in a wide plateau, and across the ventral visual stream.

The combination of EEG and fMRI dissected these broad effects into distinct components for EVC and LOC. In EVC, recurrence affected visual representations early with a peak at 100ms. This is in the range of previously observed early effects of recurrence in non-human primate EVC [75], associated with contextual modulation and figure-ground segregation [76–78] that originate from within the ventral visual stream rather than with attentional effects [16,79,80].

In LOC, recurrence affected visual representations later, with two peaks at ~175 and 300ms. The earlier peak at 175ms is consistent with effects of masking observed in humans invasively in V4/pIT [32], potentially originating from prefrontal cortex [81,82] and modulating visual activity in monkey V4 and IT [83,84]. The later peak at 300ms might reflect pattern completion, as indicated by delayed responses in invasive studies of human IT in a similar time frame [11,85]. The origin of this late effect might be medial temporal lobe regions such as parahippocampal cortex that activates as early as 270ms after stimulus onset [86–88]. Alternatively, attentional effects might be driving the late effect, consistent with reports of human and non-human attentional modulation in high-level ventral visual cortex starting at 150ms [75,89–92].

Our results cannot ultimately determine whether non-visual regions contributed to the observed effects as here, fMRI coverage was restricted to the ventral visual stream. Future research assessing the whole brain, including frontal [84,93–95] and parietal [96–99] regions, is needed.

Temporal generalization analysis added two further insights into the temporal dynamics of recurrent processing. For one, recurrence specifically contributed to the emergence of persistent, rather than transient representations. This is consistent with the observation that masking reduces firing duration in single cells in monkey IT [100,101], and that masking reduces persistence in the visual representations of occluded objects in humans [9]. Together this supports the view that recurrence plays an active role in accruing and maintaining important information online for further processing and decision making [45,102–106].

The second insight is that recurrent activity is less phase-locked to the onset of the stimulus than feedforward activity. Notably, the spectral basis of recurrent activity in LOC is in power only (Fig. 6G), in contrast to EVC where it is in power and phase (Fig. 6D). This suggests an increase of variability in phase over the course of processing, possibly due to accruing variability as information propagates increasingly back and forth along the visual processing hierarchy.

Recurrence transforms the feature format in LOC from mid- to high-level complexity

Using deep neural networks to model the representations from EEG and fMRI data [42–44], we found that recurrence changes the feature complexity of representations in LOC, but not in EVC.

In LOC, we observed a shift of representational format from predominantly mid-level to more high-level features through recurrent processing. This has three implications. First, it adds algorithmic specificity to the observations from invasive recordings in non-human primates that feature coding in high-level ventral visual cortex is dynamic, changing the code over time from global to fine-grained [107], individual object parts to multipart configuration [108], and from a code supporting detection to one for discrimination [109]. Second, it qualifies the finding that masking affects firing rate and stimulus specificity in monkey IT [31,100], linking those observations to the lack of recurrent activity mediating high-complexity features [7,84]. Finally, it converges with visual imagery and working memory studies indicating that recurrent processing carries high-complexity features [41,110]. However, a limitation of our finding is that we cannot distinguish whether the observed effect indicates the addition of new features to LOC representations through recurrence that are absent in feedforward processing [13], or the modulation of the gain of already present features, e.g., through attention [111–114].

In contrast to LOC, we did not find evidence for a change in feature complexity in EVC from its low-level complexity format (Fig. 4D and Fig. 6D). Analogous to the case of LOC, this suggests two different mechanisms underlying recurrence in EVC. One is that recurrent activity in EVC amplifies features encoded already in the feed-forward sweep [16]. The other is that it adds new features of low-level complexity, consistent with observations of dynamical feature coding in orientation and color [115,116] and changes to receptive field structure [117]. To distinguish these potential mechanisms of recurrence in both LOC and EVC future work is needed, for example, investigating the finer-grained encoding of single features rather than feature complexity [118,119] and modulating attentional state [120–122].

Note that here we used DNNs as a tool to characterize feature complexity rather than to directly model human visual processing. Future work is needed that carefully and explicitly models how feedforward and recurrent activity [18,123,124] account for core object recognition [13], as well as visual behavior [125].

The spectral basis of recurrent processing

Our results on the spectral basis of recurrent processing go beyond previous work in several ways: by identifying distinct oscillatory components of the spectral basis of recurrent processing, by linking those components differentially and directly to stimulus properties and by clarifying their distinct relationship to EVC and LOC as well as their distinct feature format [39,126–128].

We find that a set of distinct spectro-temporal components of power and phase in the theta to beta frequency range subserve recurrent processing. Our findings refine the view that low-frequency rhythms may generally serve as a neural index for recurrent processing [73,74] by showing that recurrent processes can further be subdivided into early recurrent processes (in the phase domain) that refine the representations of basic visual features, followed by later recurrent processes (in the power domain) that sculpt the representations of complex visual features in higher levels of the visual hierarchy (for a detailed discussion of each component, see Supplementary Discussion 1).

Our results further support the broad notion that theta [129], alpha [74,130] and beta [72–74,131] frequencies mediate recurrent activity and play an active role in cognition [132–135] and vision in particular [41,136–138], rather than in inhibition of irrelevant information [130,139] or cortical idling [140,141].

Backward masking as a tool to dissect recurrent processing

A key assumption on which our interpretations rest is that the difference between early and late mask conditions in neural activity isolates recurrent processing to a relevant degree. While not undoubted [142,143], this assumption is supported by a large number of studies linking backward masking to recurrent rather than feedforward processing [20,28,32,34,144], impacting the communication between and to visual regions [34,84,145].

Our results invite future backward masking studies employing multivariate analysis to further confirm and dissect the sources of recurrent activity identified here. This might in particular involve causal interventions such as TMS [146] to determine the sources of recurrent activity across cortex, and layer-specific fMRI analysis [147–149] to distinguish recurrent from feedforward processing based on cortical layers [1,2,150].

Conclusion

In sum, recurrent activity substantially affects the ventral visual stream, first in EVC and subsequently in LOC. Recurrent processing drives a shift in the feature format of LOC from mid- to high-level complexity and is linked to distinct spectro-temporal components in the theta to the beta frequency range. These findings characterize where, when, and how recurrence affects visual representations, furthering the understanding of how the recurrent information flow in the brain mediates visual object perception.

Materials and Methods

Participants in EEG and fMRI experiments

We conducted two independent experiments: an EEG and an fMRI experiment. Thirty-two participants took part in the EEG experiment, of whom one was excluded due to high-frequency noise in the recordings (N = 31, mean age 26.6 years, standard deviation 4.8 years, 20 female). Twenty-eight participants took part in the MRI experiment, of whom one was excluded due to failure of the stimulus presentation equipment (N = 27, mean age 27.7 years, standard deviation 4.6 years, 19 female). There was an overlap of four participants between the EEG and the fMRI participant sample. All participants had normal or corrected-to-normal vision. The study was conducted according to the Declaration of Helsinki and approved by the local ethics committee of the Freie Universität Berlin.

Stimulus set

The stimulus set consisted of a set of target object images and a set of image textures used to create dynamic object masks.

The set of target object images consisted of 24 object images (Fig. 1A). Each image showed an object of a different object category and was cropped quadratically to the size of the centrally presented object. The 24 object images were a subset of a larger set of 118 images [151]. The rationale for selecting the stimulus subset was as follows. Brain responses to natural images are typically highly correlated across the stages of the visual processing hierarchy. That is, two images that elicit similar responses at one stage tend to elicit similar responses at another stage, too. This makes assessing the role of different processing stages and the information they send in a forward or backward direction using multivariate analysis methods particularly difficult: due to the high correlations observed, experimental effects cannot be uniquely assigned to particular stages. To improve the chances of eliciting dissociable responses across the visual processing hierarchy in our experiment, we selected the stimulus set that yielded low correlations between the entry (early visual cortex, EVC) and the endpoint (inferior temporal cortex, IT) of the ventral visual pathway. For this, we used fMRI data in EVC and IT for the 118-image superset from a previous experiment [151]. We assessed the similarity of representations in EVC and IT for the 118 images using representational similarity analysis (RSA) [37,61]. To select 24 images that yielded uncorrelated responses, we used a genetic algorithm [152] for optimization. In detail, the optimization constraint was to minimize the absolute value of correlation between EVC and IT representational dissimilarity matrices (RDMs). The RDMs for the chosen 24-stimulus set yielded the desired low similarity between EVC and IT (R = 0.0018) on the preexisting fMRI data set. In comparison, this was lower than a random selection of 24 stimuli would have been (as assessed by 1,000 random draws, average R = 0.211, standard deviation = 0.101).

We created a set of image textures to be used for dynamic backward masks. For this, we chose a different subset of 24 object images randomly from the 118-image set and converted the images to textures that conserved the low- and mid-level image statistics of the images without portraying identifiable objects [153]. We next created 24 dynamic masks that consisted of a sequence of 12 textures each, by randomly assigning 12 of the 24 texture images in random order to a dynamic mask.

Experimental procedures

Main experiment & experimental design

We presented object images to participants in a backward masking paradigm (Fig. 1B). The general experimental design, stimulus presentation parameters, and trial structure were equivalent in both the EEG and the fMRI experiments. We describe the crucial elements common to EEG and fMRI first before detailing the modality-specific differences.

On each trial, a single object image (referred to as "target") was briefly displayed for 17ms, followed by a 200ms dynamic mask. Object images and dynamic masks were randomly paired for each trial. We manipulated the target's visibility by varying the inter-stimulus interval (ISI) between target and mask. This defined two conditions: in the early mask condition, the ISI was 17ms; in the late mask condition, the ISI was 600ms. During each trial, one of the 24 dynamic masks was presented. Stimuli were presented centrally on a gray background with a size of 5 x 5 degrees visual angle, overlaid with a bull's-eye fixation symbol with a diameter of 0.1-degree visual angle [154]. The texture images of dynamic mask were positioned and sized identically to the target object images. Participants were instructed to fixate on the fixation symbol throughout the experiment. We used Psychophysics Toolbox [155] for experimental presentation.

EEG experimental procedures

In the EEG experiment, participants completed a total of 2,544 main trials partitioned into 26 blocks of 3.5 minutes each. Throughout the experiment, each object image was presented a total of 53 times in both the early mask condition and the late mask condition.

We assessed the participants' recognition performance with additional task trials that were interspersed every 4 to 6 (average: 5) main trials. The task was to identify the object image in the previous trial from a pair of images in a two-alternative forced choice (2-AFC) task. For this, two images were presented side by side for 500ms: one of the images presented was the image from the previous trial, and the other image was randomly chosen from the remaining 23 images. Participants indicated their response with a button press.

Participants were instructed to refrain from blinking throughout the experiment except during the additional interspersed task trials, when participants were asked to blink when they gave their responses. While the inter-trial interval (ITI) between main trials was between 900ms and 1,100ms, following the 2-AFC trial, the ITI was extended to 2,000ms to prevent motor artifacts from influencing the EEG recordings of the subsequent trial.

Participants had high task performance in both masking conditions, suggesting that they attended to the stimuli even under viewing challenging conditions (for details and statistics, see Supplementary Table 7). Further, as expected, the task performance was worse for the early mask condition than for the late mask condition trials. This confirms the efficacy of the backward masking procedure in reducing object visibility.

fMRI experimental procedures

In the fMRI experiment, participants performed a total of 12 runs, each lasting 6.5 minutes. In each run, each object image was presented twice in the early mask condition and the late mask condition, resulting in 96 main trials per run. The trial-onset synchrony (TOA) was 3,000ms.

Main trials were interspersed with null trials (34 per run), during which only the background but no stimulus was shown.

Participants were instructed to attend to the object images and respond with a button press if an object image was repeated in two consecutive trials (i.e., a one-back task on the target images). Object repetitions occurred ten times per run.

As in the EEG experiment, participants had overall high task performance, with worse performance for the early mask condition than for the late mask condition trials (for details and statistics, see Supplementary Table 7).

fMRI localizer experiment

To define the regions-of-interest (ROIs) early visual cortex (EVC) and object-selective lateral occipital cortex (LOC), we performed a separate fMRI localizer run. The localizer run was conducted prior to the fMRI main experiment runs. The stimulus set comprised 40 images of objects and scrambled objects each.

The localizer run used a fMRI block design. Each block lasted 15 s. During each block, 20 stimuli were centrally presented within an area of 5 x 5 degrees visual angle at a rate of 650ms on and 100ms off. There were 6 object and scrambled object blocks each. They were presented in counterbalanced order and randomly interspersed with 7 baseline blocks during which only the background was shown.

Participants were instructed to fixate on a centrally presented fixation symbol that was presented throughout the experiment, and to respond to one-back repetitions of images with a button press. Repetitions occurred a total of 9 times over the course of the localizer experiment.

EEG data acquisition, preprocessing, and time-frequency decomposition

We recorded EEG data using an ActiCap 64 electrodes system and a Brainvision actiChamp amplifier. 64 electrodes were placed according to the 10-10 system, with an additional ground electrode and a reference electrode placed on the scalp. The signals were sampled at a rate of 1,000 Hz and online filtered between 0.03 and 100 Hz. All electrodes' impedances were kept below 10 kΩ during the recording.

We preprocessed EEG data offline using the Brainstorm-3 toolbox [156]. We removed noisy channels (average 2.2 channels per participant, standard deviation 1.8 channels) identified through visual inspection. We then filtered the data with a low-pass filter at 40 Hz. Eyeblinks and eye movement artifacts were detected using independent component analysis (ICA). We visually inspected the resulting components and removed those resembling the spatial properties of eyeblinks and eye movements (average 2.7 components per participant, standard deviation 0.9 components). We segmented the continuous data in epochs between -200ms and 800ms with respect to target image onset and baseline-corrected the segmented data by subtracting the mean of the 200ms interval before stimulus onset from the entire epoch. We finally applied multivariate noise normalization on the preprocessed data to improve the signal-to-noise ratio and reliability of the data [157]. This formed the data for the temporally resolved decoding analyses.

For time-frequency analysis, we preprocessed the data again the same way except for two differences: 1) we did not apply offline filtering, and 2) we segmented the continuous data into longer epochs (-600ms to 1,200ms) to enable better estimation of signals at lower frequencies.

Time-frequency decomposition of the EEG data

We performed time-frequency decomposition by applying complex *Morlet* wavelets. The wavelets, resembling complex sine waves modified by a Gaussian function, covered frequencies from 4 to 100 Hz in 50 logarithmically spaced increments. The Gaussian taper characteristics varied across this frequency range, with temporal full-width-half-maximum (FWHM) ranging from 20ms to 500ms as frequency decreased and spectral FWHM ranging from 1Hz to 31Hz as frequency increased.

We applied the complex *Morlet* wavelets for each channel and each trial of the EEG data at 2ms intervals (i.e., 500Hz). At each time point, this yielded 50 distinct frequency coefficients corresponding to the range of 4 to 100 Hz. At each time-frequency point, we computed two measures: the power and phase of the oscillation. To determine the absolute power values, we took the square root of the resulting time-frequency coefficients. To determine the phase values, we determined the real (sine) and imaginary (cosine) components from the time-frequency coefficients. This decomposition procedure yielded frequency-resolved EEG signals to be used for further time-frequency resolved decoding analyses. To decrease computation time and disk space usage, we downsampled the time points of frequency-resolved signals at 20ms intervals after time-frequency decomposition.

fMRI data acquisition, preprocessing and univariate analysis

We acquired T2* and T1-weighted MRI data using a 3T Siemens Tim Trio scanner with a 32-channel head coil. We acquired T2*-weighted BOLD images using a gradient-echo EPI sequence. The acquisition parameters were as follows: TR = 2,000ms, TE = 30ms, FOV = 224 x 224 mm^2, matrix size = 112 x 112, voxel size = 2 x 2 x 2 mm^3, flip angle = 70°, with 30 slices and a 20% gap. The acquisition volume covered the occipital and temporal lobes and was oriented parallel to the inferior temporal cortex. Additionally, we obtained a T1-weighted image for each participant as an anatomical reference (MPRAGE; TR = 1,900ms, TE = 2.52ms, TI = 900ms, matrix size = 256 x 256, voxel size = 1 x 1 x 1 mm^3, and 176 slices).

We performed fMRI data preprocessing using SPM12 (https://www.fil.ion.ucl.ac.uk/spm/software/spm12/). This involved realignment, slice-time correction, co-registration to the anatomical image, and normalization to MNI space. For the fMRI data of the localizer experiment, but not the main experiment, we additionally applied smoothing with a Gaussian kernel (FWHM = 5 mm). For the fMRI data from the main experiment, we additionally estimated noise components using the Tapas PhysIO toolbox [158,159] by creating tissue-probability maps from each participant's anatomical image and extracting noise components from the white matter and CSF maps combined with the fMRI time series.

We used a general linear model (GLM) to estimate responses for the 48 experimental conditions (i.e., the 24 object images presented in either the early mask condition or the late mask condition). The analysis was conducted in a participant-specific fashion. We applied the GLM estimation to the preprocessed fMRI data for each run. We entered experimental condition onsets and durations as regressors into the GLM. Nuisance regressors comprised noise components and

movement parameters. We evaluated 20 different GLMs by convolving regressors with 20 distinct hemodynamics response functions (HRFs) as derived from a large fMRI dataset [160]. For each voxel, we then identified the HRF that resulted in the lowest average residual [161] and chose the corresponding estimates for further analysis. This approach resulted in 48 beta maps (one for each experimental condition) for each run and participant.

We used a separate GLM to estimate responses for the localizer run. We included block onsets and durations as regressors for the 3 conditions (i.e., objects, scrambled objects, and baseline), along with movement parameters as nuisance regressors. We convolved the regressors with the canonical HRF. We computed two contrasts from the resulting GLM parameter estimates that were used at a later step for voxel selection in the ROI analysis. The first contrast was defined as object + scrambled objects > baseline to define EVC. The second contrast was defined as objects > scrambled objects to define LOC. This yielded two t-value maps for the localizer run per participant.

Definition of fMRI regions of interest (ROIs)

For each participant, we identified two regions of interest (ROIs) within the ventral visual stream: early visual cortex (EVC) and lateral occipital complex (LOC). To determine the boundaries of these ROIs, we used participant-specific t-value maps from the localizer run threshold at $p < 0.0001$ intersected with anatomical masks. For the EVC definition, we intersected the thresholded t-value map (object + scrambled objects > baseline) with the combined anatomical region masks of V1, V2, and V3 obtained from the Glasser Brain Atlas [162]. For the LOC definition, we intersected the thresholded t-value map (objects > scrambled objects) with a mask of LOC derived from a functional atlas [163]. We removed any voxels shared between the EVC and LOC ROIs to avoid overlap. This process resulted in the definitions of two ROIs for each participant.

Multivariate pattern analysis on EEG and fMRI data

An analytical challenge in comparing neural activity evoked by target images versus target image with a backward mask is the confounding effect introduced by the mask. Previous studies addressed this challenge by using subtraction design, for example, by including trials showing only the mask and subtracting the resulting neural activity from the neural activity evoked by the stimulus plus mask [28,164]. Here, instead, we used a content-sensitive multivariate pattern analysis on EEG and fMRI data to dissect neural activity of the target image from neural activity evoked by the mask. The rationale is that in our design, target and mask stimuli were statistically independent, so multivariate pattern analysis classifying target object images revealed neural activity related to object images rather than the mask.

We performed multivariate pattern analysis on EEG and fMRI data using linear support vector machines [165] as implemented in the LIBSVM toolbox [166] in MATLAB (2021a). We conducted all analyses on a participant-specific basis.

Temporally resolved decoding analysis from EEG data

To determine when the brain processes object information, we conducted a time-resolved decoding analysis [45,167]. We examined EEG data from -200ms to 800ms with respect to target image onset, in 10ms intervals. At each time point, we extracted trial-specific EEG channel activations and arranged them into 64-dimensional pattern vectors for each of the 24 object

image conditions for each masking condition, separately. We conducted two types of analysis: within- and across-masking conditions object decoding.

In the within-masking condition analysis, we separately decoded object conditions for the early mask and the late mask conditions. For each of the 24 image conditions, we first randomly grouped trials into four equally sized bins and averaged them to create four pseudo-trials to enhance the signal-to-noise ratio (SNR). Employing a leave-one-out cross-validation approach, we then divided these pseudo-trials into training (three pseudo trials) and testing sets (one pseudo trial) to pairwise decode object identity. We then decoded object conditions pairwise for all object condition combinations. The resulting decoding accuracies were arranged into a 24 × 24 decoding accuracy matrix, with rows and columns corresponding to the decoded object conditions. This matrix is symmetric across the diagonal, with the diagonal being undefined. We repeated this analysis 100 times, randomly assigning trials to pseudo-trials each time. Averaging results over repetitions yielded one 24 × 24 decoding accuracy matrix for each time point, separately for the early and late mask conditions.

In the across-masking conditions analysis, we proceeded accordingly but assigned pseudo-trials to the training set and testing set from different masking conditions. That is, we trained on data recorded in the early mask condition and tested on data from the late mask condition (or vice versa). We averaged the results across both training and testing directions. This yielded one 24 × 24 decoding accuracy matrix for each time point.

In both analyses, averaging across the 24 × 24 entries of decoding accuracy at each time point resulted in a grand-average decoding accuracy time course.

Time generalization decoding analysis

We used time-generalization decoding analysis to determine how visual representations relate to each other across different time points. We proceeded as for the within masking condition time-resolved decoding analysis, except that classifiers trained on data from a particular time point were tested iteratively on data from all other time points. The rationale here is that successful generalization across time points indicates the similarity of visual representations over time. This analysis yielded 24 × 24 decoding accuracy matrices for each combination of time points from -200 to +800ms. By averaging the entries of each decoding accuracy matrix across time point combinations, we obtained a temporal generalization matrix (TGM), where rows and columns are indexed by training and testing time points, respectively.

Time-frequency-resolved decoding analysis from EEG frequency power and phase

To determine the spectral properties of visual object representations in the two masking conditions, we conducted a time-frequency-resolved decoding analysis. This analysis was identical to the time-resolved analysis described above, but instead of decoding from raw activation values, we decoded object identity from patterns of power or phase value. We performed the analysis separately for 50 frequency bins spanning from 4 Hz to 100 Hz, using either power or phase values. In the power-based analysis, decoding was based on 64 power values corresponding to the 64 EEG channels. For the phase-based analysis, decoding used 128 values corresponding to the concatenation of the 64 sine and 64 cosine values. This resulted in one 24 × 24 decoding accuracy matrix for each time point and frequency bin, for the power- and phase-based analyses. Averaging across the 24 × 24 entries of decoding accuracy resulted in a

grand average time-frequency matrix, where time points and frequency bins are indexed in rows and columns, respectively.

Spatially resolved decoding analysis from fMRI data

We conducted two types of decoding analyses on the fMRI data: region-of-interest (ROI)-based and spatially unbiased volumetric searchlight-based decoding [57,58] on the fMRI data.

For the ROI-based analysis, we arranged beta values from voxels of a given ROI into pattern vectors for each of the 24 experimental conditions and each of the 12 runs of the main fMRI experiment. To enhance signal-to-noise ratio, we grouped 3 runs into 4 bins and averaged across runs, creating four pseudo-run fMRI pattern vectors [168]. Then for each ROI, we performed object decoding on these pseudo-run fMRI pattern vectors in a leave-one-pseudo-run-out manner. Averaging across iterations yielded a 24 × 24 decoding accuracy matrix for each ROI, participant, and masking condition.

For the searchlight-based analysis, for each voxel in the 3D fMRI volume, we defined spheres of voxels around it with a radius of four voxels. For each sphere, we arranged voxel values into pattern vectors. We then decoded object identity as described for the ROI-based analysis. This yielded a 24 × 24 decoding accuracy matrix for each voxel in the 3D fMRI volume for each participant, and each masking condition.

In both ROI and searchlight-based analyses, averaging across the 24 × 24 entries of decoding accuracy resulted in either a single value or a 3D map of grand average decoding accuracy, respectively.

Representational similarity analysis (RSA)

RSA is a framework to relate representations across different measurement and signal spaces, such as those defined by different brain imaging modalities (EEG and fMRI) or computational models [37,61]. The idea is to abstract from incommensurate measurement spaces into a common similarity space where representations can be directly compared.

For each masking condition, the analysis proceeded in two steps. In the first step, within each signal space of interest (e.g., fMRI responses in ROI, EEG broadband responses at particular time points, EEG spectral responses at time-frequency combinations, and activations of DNN layers), we calculated the dissimilarity between condition-specific multivariate activity patterns for all pairwise combinations of the 24 object conditions. We aggregated the results in representational dissimilarity matrices (RDMs), where rows and columns were indexed by the 24 object conditions. These RDMs summarize the representational geometry within each signal space. In the second step, we compared the RDMs across signal spaces using Spearman correlations, yielding a measure of their similarity. We provide the details for each of the two steps below.

Step 1: Construction of RDMs

For the brain data, we used the decoding accuracy matrices resulting from the decoding analyses detailed above as RDMs. This yielded RDMs a) from the temporally resolved EEG decoding analysis for each time point, b) from the time-frequency-resolved EEG decoding analysis for every time-point and frequency combination, separately for power and phase, and c) from the spatially resolved fMRI decoding analysis for each ROI.

For the computational model, we built RDMs from an AlexNet architecture trained for object categorization on the ImageNet dataset [59,60]. AlexNet is an 8-layer deep neural network (DNN) commonly used as a baseline for brain-DNN comparisons [169]. Using the MatConvNet toolbox [170], we fed our object stimuli into the pre-trained AlexNet and extracted the activation patterns for each stimulus from each of the five convolutional layers (conv1 to conv5) and the three fully connected layer (fc6, fc7, and fc8).

To test the generalizability of our conclusion across different DNN models, we also built RDMs using the ResNet50 architecture [66], pre-trained on the ImageNet dataset [59] for object categorization. ResNet50 features a distinct architecture compared to AlexNet, consisting of an initial convolutional layer followed by four residual blocks, each containing multiple convolutional layers with skip connections, and leading to a final classification layer. We fed the object stimuli into ResNet50 and extracted the activation patterns for each stimulus from the last layer of each of the four residual blocks (block1 to block4) as well as from the final classification layer (fc).

We quantified the dissimilarity of the activation patterns by calculating 1-Pearson's R for each pair of stimuli. This resulted in eight RDMs for AlexNet layers and five RDMs for ResNet50 layers.

Step 2a: Standard RSA - relating DNN RDMs to EEG and fMRI RDMs

To characterize the format of neural representations, we related DNN RDMs from each layer to EEG and fMRI RDMs (Fig. 3a). The idea is that ascending layers of a DNN capture features of increasing complexity. Thus, relating neural representations to each DNN layer informs about the feature complexity of the neural representations [42–44].

For the EEG-based analysis, we correlated the DNN RDMs with EEG RDMs across all time points obtained from temporally resolved EEG decoding analysis. This yielded a time course of correlation values for each DNN layer, participant, and masking condition. For the fMRI-based analysis, we correlated the DNN RDMs with RDMs from two ROIs (i.e., EVC and LOC), yielding a correlation value per ROI for each DNN layer, participant, and masking condition.

Step 2b: Commonality analysis - shared variance among EEG, fMRI and DNN RDMs

To investigate the temporal dynamics of specific visual features emerging in brain regions, we extended standard RSA to commonality analysis [68,69] (Fig. 4A). Specifically, we computed the coefficients of shared variance separately among EEG RDMs at each time point, fMRI RDMs in each ROI, and DNN RDMs for each layer. This resulted in a time course of shared variance (R^2) for each DNN layer, ROI, participant, and masking condition.

To investigate where in the brain the specific visual features originate and how each of the four spectro-temporally identified components carries them, we conducted commonality analysis once more (Fig. 6A). Here, we calculated coefficients of shared variance among frequency-based EEG RDMs corresponding to each spectro-temporally identified component, fMRI RDMs within each ROI, and DNN RDMs across each layer. This analysis resulted in a coefficient of shared variance (R^2) for each DNN layer, ROI, power- or phase-based component, participant, and masking condition.

Noise ceilings

We calculated an upper and lower bound for the noise ceiling [61], that is the maximal correlation in the RSA analyses that might be achieved given the noisiness of the data. This was done for the EEG data and fMRI data (i.e., ROIs) separately. To estimate the lower bound, we correlated each participant's RDM with the average RDM of all other participants. To estimate the upper bound, we correlated each participant's RDM with the average RDM of all participants. We averaged the results, thus obtaining estimates of the lower and upper noise ceilings for each EEG time point or time-point and frequency combination, as well as for all fMRI ROIs.

Statistical analyses

We used sign permutation tests [171] that do not make assumptions about the data distribution. We compared the statistic of interest (i.e., mean decoding accuracy, correlation coefficients in RSA, coefficients of shared variance in commonality analysis, or differences therein between the masking conditions) against the null hypothesis that the statistic of interest was equal to chance (i.e., 50 % decoding accuracy for pair-wise decoding, a correlation of 0, a coefficient of shared variance of 0, or a difference of 0). To obtain a null distribution, we multiplied participant-specific data randomly by either +1 or -1 and computed the statistic of interest for 10,000 permutations. Based on these null distributions, we obtained p-values by comparing the original statistic to the null distribution. We conducted one-tailed (i.e., the right-tailed) tests for all statistics of interest except for differences, for which we used two-tailed tests.

To correct for multiple comparisons with a small number of unrelated comparisons, we used FDR correction at a $p < 0.05$ [172]. In cases involving a large number of comparisons in contiguous and correlated results (i.e., time points, frequencies, or voxels), we used cluster-based inference [173]. For the cluster-size-based inference, we calculated the statistic of interest both for the empirical results and for each permutation sample under the null hypothesis. This resulted in 1-dimensional (e.g., decoding time courses, RSA-based correlation time courses, time courses of shared variance), 2-dimensional (e.g., decoding time-time matrices, decoding time-frequency matrices, RSA-based correlation matrices), or 3-dimensional (i.e., fMRI volumetric decoding results) p-value maps. We defined clusters based on temporal or spatial contiguity with a $p < 0.005$ (i.e., cluster-definition threshold) for most analyses, except for time-frequency decoding matrices, which used a threshold of $p < 0.05$. We determined the maximum cluster size for each permutation sample, yielding a distribution of the maximum cluster size statistic. We set the cluster threshold at $p < 0.05$.

We calculated 95% confidence intervals for the peak latencies in the resulted time courses (e.g., decoding time courses, RSA-based correlation time courses, time courses of shared variance). For this, we randomly sampled participants with replacements 1,000 times. For each bootstrap sample, we determined the peak latency. This yielded a distribution of peak latencies for which we report the 95 % confidence intervals.

Acknowledgments

We are grateful to Marleen Haupt and Agnessa Karapetian for their valuable comments on the manuscript, to Lixiang Chen for his assistance with data collection, and to all the participants for their time and contributions to this study. We thank the HPC service of FUB-IT, Freie Universitaet Berlin, for computing time on Curta (http://dx.doi.org/10.17169/refubium-26754).

R.M.C. is supported by the German Research Foundation (DFG; CI241/1-1, CI241/3-1, CI241/7-1) and by a European Research Council (ERC) starting grant (ERC-2018-STG 803370). D.K. is supported by the German Research Foundation (DFG; SFB/TRR135, project number 222641018; KA4683/5-1, project number 518483074), "The Adaptive Mind", funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art, and an ERC Starting Grant (PEP, ERC-2022-STG 101076057).

Author contributions

Conceptualization: S.X., J.S., D.K. and R.M.C.; Methodology: S.X., J.S., D.K. and R.M.C.; Software: S.X.; Formal analysis: S.X. and J.S.; Investigation: S.X. and B.Y.; Resources: S.X. and R.M.C.; Data curation: S.X.; Writing - original draft preparation: S.X. and R.M.C.; Writing - review and editing: S.X., J.S., D.K. and R.M.C.; Visualization: S.X.; Supervision: R.M.C. and D.K.; Project administrations: R.M.C. and D.K.; Funding acquisition: R.M.C. and D.K.

Declaration of interests

The authors declare no competing interests.

Supplementary information

Document S1. Supplementary Discussion 1, Supplementary Figures 1-7 and Supplementary Tables 1-7.

References

1. Felleman, D. J. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex N. Y. NY 1991* **1**, 1–47 (1991).

2. Markov, N. T. *et al.* Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *J. Comp. Neurol.* **522**, 225–259 (2014).

3. Sporns, O. & Zwi, J. D. The small world of the cerebral cortex. *Neuroinformatics* **2**, 145–162 (2004).

4. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How Does the Brain Solve Visual Object Recognition? *Neuron* **73**, 415–434 (2012).

5. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–1025 (1999).

6. Groen, I. I. A. *et al.* Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLOS Comput. Biol.* **14**, e1006690 (2018).

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

8. Koivisto, M., Kastrati, G. & Revonsuo, A. Recurrent Processing Enhances Visual Awareness but Is Not Necessary for Fast Categorization of Natural Scenes. *J. Cogn. Neurosci.* **26**, 223–231 (2014).

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

10. Seijdel, N. *et al.* On the Necessity of Recurrent Processing during Object Recognition: It Depends on the Need for Scene Segmentation. *J. Neurosci.* **41**, 6281–6289 (2021).

11. Tang, H. *et al.* Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci.* **115**, 8835–8840 (2018).

12. Wyatte, D., Curran, T. & O'Reilly, R. The Limits of Feedforward Vision: Recurrent Processing Promotes Robust Object Recognition when Objects Are Degraded. *J. Cogn. Neurosci.* **24**, 2248–2261 (2012).

13. Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl. Acad. Sci.* **116**, 21854–21863 (2019).

14. Koivisto, M., Railo, H., Revonsuo, A., Vanni, S. & Salminen-Vaparanta, N. Recurrent Processing in V1/V2 Contributes to Categorization of Natural Scenes. *J. Neurosci.* **31**, 2488–2492 (2011).

15. Rao, R. P. N. & Ballard, D. H. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nat. Neurosci.* **2**, 79–87 (1999).

16. Lamme, V. A. F. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).

17. Hochstein, S. & Ahissar, M. View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron* **36**, 791–804 (2002).

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

19. Pennartz, C. M. A., Dora, S., Muckli, L. & Lorteije, J. A. M. Towards a Unified View on Pathways and Functions of Neural Recurrent Processing. *Trends Neurosci.* **42**, 589–603 (2019).

20. Boehler, C. N., Schoenfeld, M. A., Heinze, H.-J. & Hopf, J.-M. Rapid recurrent processing gates awareness in primary visual cortex. *Proc. Natl. Acad. Sci.* **105**, 8742–8747 (2008).

21. Bullier, J. Integrated model of visual processing. *Brain Res. Rev.* **36**, 96–107 (2001).

22. Gilbert, C. D. & Li, W. Top-down influences on visual processing. *Nat. Rev. Neurosci.* **14**, 350–363 (2013).

23. Bachmann, T. *Psychophysiology of Visual Masking: The Fine Structure of Conscious Experience*. (Nova Science Publishers, 1994).

24. Breitmeyer, B. G. *Visual Masking: An Integrative Approach*. (Clarendon Press, 1984).

25. Breitmeyer, B. G. & Ogmen, H. Recent models and findings in visual backward masking: A comparison, review, and update. *Percept. Psychophys.* **62**, 1572–1595 (2000).

26. Kahneman, D. Method, findings, and theory in studies of visual masking. *Psychol. Bull.* **70**, 404–425 (1968).

939 27. Bacon-Macé, N., Macé, M. J.-M., Fabre-Thorpe, M. & Thorpe, S. J. The time course of
940 visual processing: Backward masking and natural scene categorisation. *Vision Res.* **45**,
941 1459–1469 (2005).
942 28. Fahrenfort, J. J., Scholte, H. S. & Lamme, V. A. F. Masking Disrupts Reentrant Processing
943 in Human Visual Cortex. *J. Cogn. Neurosci.* **19**, 1488–1497 (2007).
944 29. Grill-Spector, K., Kushnir, T., Hendler, T. & Malach, R. The dynamics of object-selective
945 activation correlate with recognition performance in humans. *Nat. Neurosci.* **3**, 837–843
946 (2000).
947 30. Kravitz, D. J., Saleem, K. S., Baker, C. I., Ungerleider, L. G. & Mishkin, M. The ventral
948 visual pathway: an expanded neural framework for the processing of object quality. *Trends*
949 *Cogn. Sci.* **17**, 26–49 (2013).
950 31. Rolls, E. T., Tovée, M. J. & Panzeri, S. The Neurophysiology of Backward Visual Masking:
951 Information Analysis. *J. Cogn. Neurosci.* **11**, 300–311 (1999).
952 32. Cauchoix, M., Crouzet, S. M., Fize, D. & Serre, T. Fast ventral stream neural activity enables
953 rapid visual categorization. *NeuroImage* **125**, 280–290 (2016).
954 33. Jiang, Y. & He, S. Cortical Responses to Invisible Faces: Dissociating Subsystems for
955 Facial-Information Processing. *Curr. Biol.* **16**, 2023–2029 (2006).
956 34. Lamme, V. A. F., Zipser, K. & Spekreijse, H. Masking Interrupts Figure-Ground Signals in
957 V1. *J. Cogn. Neurosci.* **14**, 1044–1053 (2002).
958 35. Cichy, R. M. & Oliva, A. A M/EEG-fMRI Fusion Primer: Resolving Human Brain
959 Responses in Space and Time. *Neuron* **107**, 772–781 (2020).
960 36. Haynes, J.-D. A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and
961 Perspectives. *Neuron* **87**, 257–270 (2015).

962 37. Kriegeskorte, N. Representational similarity analysis – connecting the branches of systems
963 neuroscience. *Front. Syst. Neurosci.* (2008) doi:10.3389/neuro.06.004.2008.
964 38. Engel, A. K., Fries, P. & Singer, W. Dynamic predictions: Oscillations and synchrony in
965 top–down processing. *Nat. Rev. Neurosci.* **2**, 704–716 (2001).
966 39. Reddy, L., Cichy, R. M. & VanRullen, R. Representational Content of Oscillatory Brain
967 Activity during Object Recognition: Contrasting Cortical and Deep Neural Network
968 Hierarchies. *eneuro* **8**, ENEURO.0362-20.2021 (2021).
969 40. Siegel, M., Donner, T. H. & Engel, A. K. Spectral fingerprints of large-scale neuronal
970 interactions. *Nat. Rev. Neurosci.* **13**, 121–134 (2012).
971 41. Xie, S., Kaiser, D. & Cichy, R. M. Visual Imagery and Perception Share Neural
972 Representations in the Alpha Frequency Band. *Curr. Biol.* **30**, 2621-2627.e5 (2020).
973 42. Cichy, R. M. & Kaiser, D. Deep Neural Networks as Scientific Models. *Trends Cogn. Sci.* **23**,
974 305–317 (2019).
975 43. Doerig, A. *et al.* The neuroconnectionist research programme. *Nat. Rev. Neurosci.* **24**, 431–
976 450 (2023).
977 44. Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand
978 sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
979 45. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and
980 time. *Nat. Neurosci.* **17**, 455–462 (2014).
981 46. Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J. & Turret, J. High temporal resolution
982 decoding of object position and category. *J. Vis.* **11**, 9–9 (2011).
983 47. Oosterhof, N. N., Tipper, S. P. & Downing, P. E. Crossmodal and action-specific:
984 neuroimaging the human mirror neuron system. *Trends Cogn. Sci.* **17**, 311–318 (2013).

48. Peelen, M. V. & Downing, P. E. Testing cognitive theories with multivariate pattern analysis of neuroimaging data. *Nat. Hum. Behav.* (2023) doi:10.1038/s41562-023-01680-z.

49. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210 (2014).

50. Vidaurre, D. A generative model of electrophysiological brain responses to stimulation. *eLife* **12**, RP87729 (2024).

51. Vidaurre, D., Cichy, R. M. & Woolrich, M. W. Dissociable Components of Information Encoding in Human Perception. *Cereb. Cortex* **31**, 5664–5675 (2021).

52. Grill-Spector, K. & Malach, R. THE HUMAN VISUAL CORTEX. *Annu. Rev. Neurosci.* **27**, 649–677 (2004).

53. Hubel, D. H. & Wiesel, T. N. Receptive fields and functional architecture of monkey striate cortex. *J. Physiol.* **195**, 215–243 (1968).

54. Grill-Spector, K., Kourtzi, Z. & Kanwisher, N. The lateral occipital complex and its role in object recognition. *Vision Res.* **41**, 1409–1422 (2001).

55. Malach, R. *et al.* Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc. Natl. Acad. Sci.* **92**, 8135–8139 (1995).

56. Vinberg, J. & Grill-Spector, K. Representation of Shapes, Edges, and Surfaces Across Multiple Cues in the Human Visual Cortex. *J. Neurophysiol.* **99**, 1380–1393 (2008).

57. Haynes, J.-D. & Rees, G. Predicting the orientation of invisible stimuli from activity in human primary visual cortex. *Nat. Neurosci.* **8**, 686–691 (2005).

58. Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl. Acad. Sci.* **103**, 3863–3868 (2006).

1007 59. Nili, H. *et al.* A Toolbox for Representational Similarity Analysis. *PLoS Comput. Biol.* **10**,
1008 e1003553 (2014).
1009 60. Deng, J. *et al.* Imagenet: A large-scale hierarchical image database. in 248–255 (Ieee, 2009).
1010 61. Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep
1011 convolutional neural networks. *Adv. Neural Inf. Process. Syst.* **25**, (2012).
1012 62. Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A. & Oliva, A. Comparison of deep neural
1013 networks to spatio-temporal cortical dynamics of human visual object recognition reveals
1014 hierarchical correspondence. *Sci. Rep.* **6**, 27755 (2016).
1015 63. Seeliger, K. *et al.* Convolutional neural network-based encoding and decoding of visual
1016 object recognition in space and time. *NeuroImage* **180**, 253–266 (2018).
1017 64. DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**,
1018 333–341 (2007).
1019 65. Serre, T. & Poggio, T. A neuromorphic approach to computer vision. *Commun. ACM* **53**, 54–
1020 61 (2010).
1021 66. He, K., Zhang, X., Ren, S. & Sun, J. Deep Residual Learning for Image Recognition.
1022 Preprint at https://doi.org/10.48550/arXiv.1512.03385 (2015).
1023 67. Rousselet, G. A., Thorpe, S. J. & Fabre-Thorpe, M. How parallel is visual processing in the
1024 ventral pathway? *Trends Cogn. Sci.* **8**, 363–370 (2004).
1025 68. Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I. & Cichy, R. M. The representational
1026 dynamics of task and object processing in humans. *eLife* **7**, e32816 (2018).
1027 69. Seibold, D. R. & McPhee, R. D. Commonality Analysis: A Method for Decomposing
1028 Explained Variance in Multiple Regression Analyses. *Hum. Commun. Res.* **5**, 355–365
1029 (1979).

70. Buzsaki, G. Neuronal Oscillations in Cortical Networks. *Science* **304**, 1926–1929 (2004).

71. Fries, P. Rhythms for Cognition: Communication through Coherence. *Neuron* **88**, 220–235 (2015).

72. Bastos, A. M. *et al.* Visual Areas Exert Feedforward and Feedback Influences through Distinct Frequency Channels. *Neuron* **85**, 390–401 (2015).

73. Michalareas, G. *et al.* Alpha-Beta and Gamma Rhythms Subserve Feedback and Feedforward Influences among Human Visual Cortical Areas. *Neuron* **89**, 384–397 (2016).

74. van Kerkoerle, T. *et al.* Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proc. Natl. Acad. Sci.* **111**, 14332–14341 (2014).

75. Wyatte, D., Jilk, D. J. & O'Reilly, R. C. Early recurrent feedback facilitates visual object recognition under challenging conditions. *Front. Psychol.* **5**, (2014).

76. Hupé, J. M. *et al.* Cortical feedback improves discrimination between figure and background by V1, V2 and V3 neurons. *Nature* **394**, 784–787 (1998).

77. Muckli, L. *et al.* Contextual Feedback to Superficial Layers of V1. *Curr. Biol.* **25**, 2690–2695 (2015).

78. Roelfsema, P. R., Lamme, V. A. F., Spekreijse, H. & Bosch, H. Figure—Ground Segregation in a Recurrent Network Architecture. *J. Cogn. Neurosci.* **14**, 525–537 (2002).

79. Poort, J., Self, M. W., van Vugt, B., Malkki, H. & Roelfsema, P. R. Texture Segregation Causes Early Figure Enhancement and Later Ground Suppression in Areas V1 and V4 of Visual Cortex. *Cereb. Cortex* **26**, 3964–3976 (2016).

80. Poort, J. *et al.* The Role of Attention in Figure-Ground Segregation in Areas V1 and V4 of the Visual Cortex. *Neuron* **75**, 143–156 (2012).

1053 81. Bar, M. A Cortical Mechanism for Triggering Top-Down Facilitation in Visual Object
1054 Recognition. *J. Cogn. Neurosci.* **15**, 600–609 (2003).
1055 82. Bichot, N. P., Heard, M. T., DeGennaro, E. M. & Desimone, R. A Source for Feature-Based
1056 Attention in the Prefrontal Cortex. *Neuron* **88**, 832–844 (2015).
1057 83. Fyall, A. M., El-Shamayleh, Y., Choi, H., Shea-Brown, E. & Pasupathy, A. Dynamic
1058 representation of partially occluded objects in primate prefrontal and visual cortex. *eLife* **6**,
1059 e25784 (2017).
1060 84. Kar, K. & DiCarlo, J. J. Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is
1061 Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition. *Neuron*
1062 **109**, 164-176.e5 (2021).
1063 85. Tang, H. *et al.* Spatiotemporal Dynamics Underlying Object Completion in Human Ventral
1064 Visual Cortex. *Neuron* **83**, 736–748 (2014).
1065 86. Bar, M. Visual objects in context. *Nat. Rev. Neurosci.* **5**, 617–629 (2004).
1066 87. Brandman, T. & Peelen, M. V. Interaction between Scene and Object Processing Revealed
1067 by Human fMRI and MEG Decoding. *J. Neurosci.* **37**, 7700–7710 (2017).
1068 88. Epstein, R. A. & Baker, C. I. Scene Perception in the Human Brain. *Annu. Rev. Vis. Sci.* **5**,
1069 373–397 (2019).
1070 89. Buffalo, E. A., Fries, P., Landman, R., Liang, H. & Desimone, R. A backward progression of
1071 attentional effects in the ventral stream. *Proc. Natl. Acad. Sci.* **107**, 361–365 (2010).
1072 90. Roelfsema, P. R., Lamme, V. A. F. & Spekreijse, H. Object-based attention in the primary
1073 visual cortex of the macaque monkey. *Nature* **395**, 376–381 (1998).

1074 91. Von Seth, J., Nicholls, V. I., Tyler, L. K. & Clarke, A. Recurrent connectivity supports
1075 higher-level visual and semantic object representations in the brain. *Commun. Biol.* **6**, 1207
1076 (2023).
1077 92. Kaiser, D., Oosterhof, N. N. & Peelen, M. V. The Neural Dynamics of Attentional Selection
1078 in Natural Scenes. *J. Neurosci.* **36**, 10522–10528 (2016).
1079 93. Freedman, D. J., Riesenhuber, M., Poggio, T. & Miller, E. K. Categorical Representation of
1080 Visual Stimuli in the Primate Prefrontal Cortex. *Science* **291**, 312–316 (2001).
1081 94. Kornblith, S. & Tsao, D. Y. How thoughts arise from sights: inferotemporal and prefrontal
1082 contributions to vision. *Curr. Opin. Neurobiol.* **46**, 208–218 (2017).
1083 95. Morishima, Y. *et al.* Task-specific signal transmission from prefrontal cortex in visual
1084 selective attention. *Nat. Neurosci.* **12**, 85–91 (2009).
1085 96. Freud, E., Plaut, D. C. & Behrmann, M. 'What' Is Happening in the Dorsal Visual Pathway.
1086 *Trends Cogn. Sci.* **20**, 773–784 (2016).
1087 97. Konen, C. S. & Kastner, S. Two hierarchically organized neural systems for object
1088 information in human visual cortex. *Nat. Neurosci.* **11**, 224–231 (2008).
1089 98. Sawamura, H., Georgieva, S., Vogels, R., Vanduffel, W. & Orban, G. A. Using Functional
1090 Magnetic Resonance Imaging to Assess Adaptation and Size Invariance of Shape
1091 Processing by Humans and Monkeys. *J. Neurosci.* **25**, 4294–4306 (2005).
1092 99. Sereno, A. B. & Maunsell, J. H. R. Shape selectivity in primate lateral intraparietal cortex.
1093 *Nature* **395**, 500–503 (1998).
1094 100. Kovács, G., Vogels, R. & Orban, G. A. Cortical correlate of pattern backward masking.
1095 *Proc. Natl. Acad. Sci.* **92**, 5587–5591 (1995).

101. Rolls, E. T. & Tovee, M. J. The responses of single neurons in the temporal visual cortical areas of the macaque when more than one stimulus is present in the receptive field. *Exp. Brain Res.* **103**, 409–420 (1995).

102. Christophel, T. B., Iamshchinina, P., Yan, C., Allefeld, C. & Haynes, J.-D. Cortical specialization for attended versus unattended working memory. *Nat. Neurosci.* **21**, 494–496 (2018).

103. Gwilliams, L. & King, J.-R. Recurrent processes support a cascade of hierarchical decisions. *eLife* **9**, e56603 (2020).

104. O'Connell, R. G., Dockree, P. M. & Kelly, S. P. A supramodal accumulation-to-bound signal that determines perceptual decisions in humans. *Nat. Neurosci.* **15**, 1729–1735 (2012).

105. Wang, X.-J. Synaptic reverberation underlying mnemonic persistent activity. *Trends Neurosci.* **24**, 455–463 (2001).

106. Karapetian, A. *et al.* Empirically Identifying and Computationally Modelling the Brain–Behavior Relationship for Human Scene Categorization. *J. Cogn. Neurosci.* 1–19 (2023) doi:10.1162/jocn_a_02043.

107. Sugase, Y., Yamane, S., Ueno, S. & Kawano, K. Global and fine information coded by single neurons in the temporal visual cortex. *Nature* **400**, 869–873 (1999).

108. Brincat, S. L. & Connor, C. E. Dynamic Shape Synthesis in Posterior Inferotemporal Cortex. *Neuron* **49**, 17–24 (2006).

109. Shi, Y. *et al.* Rapid, concerted switching of the neural code in inferotemporal cortex. 2023.12.06.570341 Preprint at https://doi.org/10.1101/2023.12.06.570341 (2023).

110. Liu, J. *et al.* Stable maintenance of multiple representational formats in human visual short-term memory. *Proc. Natl. Acad. Sci.* **117**, 32329–32339 (2020).

111. Çukur, T., Nishimoto, S., Huth, A. G. & Gallant, J. L. Attention during natural vision warps semantic representation across the human brain. *Nat. Neurosci.* **16**, 763–770 (2013).

112. Itthipuripat, S., Garcia, J. O., Rungratsameetaweemana, N., Sprague, T. C. & Serences, J. T. Changing the Spatial Scope of Attention Alters Patterns of Neural Gain in Human Cortex. *J. Neurosci.* **34**, 112–123 (2014).

113. Serences, J. T., Saproo, S., Scolari, M., Ho, T. & Muftuler, L. T. Estimating the influence of attention on population codes in human visual cortex using voxel-based tuning functions. *NeuroImage* **44**, 223–231 (2009).

114. Treue, S. & Trujillo, J. C. M. Feature-based attention influences motion processing gain in macaque visual cortex. *Nature* **399**, 575–579 (1999).

115. Cottaris, N. P. & De Valois, R. L. Temporal dynamics of chromatic tuning in macaque primary visual cortex. *Nature* **395**, 896–900 (1998).

116. Ringach, D. L., Hawken, M. J. & Shapley, R. Dynamics of orientation tuning in macaque primary visual cortex. *Nature* **387**, 281–284 (1997).

117. Keller, A. J., Roth, M. M. & Scanziani, M. Feedback generates a second receptive field in neurons of the visual cortex. *Nature* **582**, 545–549 (2020).

118. Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532**, 453–458 (2016).

119. Naselaris, T., Kay, K. N., Nishimoto, S. & Gallant, J. L. Encoding and decoding in fMRI. *NeuroImage* **56**, 400–410 (2011).

120. Baldauf, D. & Desimone, R. Neural Mechanisms of Object-Based Attention. *Science* **344**, 424–427 (2014).

121. Chapman, A. F. & Störmer, V. S. Representational structures as a unifying framework for attention. *Trends Cogn. Sci.* **28**, 416–427 (2024).

122. White, A. L., Kay, K. N., Tang, K. A. & Yeatman, J. D. Engaging in word recognition elicits highly specific modulations in visual cortex. *Curr. Biol.* **33**, 1308-1320.e5 (2023).

123. Maniquet, T., Beeck, H. O. de & Costantino, A. I. Recurrent issues with deep neural network models of visual recognition. 2024.04.02.587669 Preprint at https://doi.org/10.1101/2024.04.02.587669 (2024).

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

125. Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I. & Kriegeskorte, N. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLOS Comput. Biol.* **16**, e1008215 (2020).

126. Romei, V., Driver, J., Schyns, P. G. & Thut, G. Rhythmic TMS over Parietal Cortex Links Distinct Brain Frequencies to Global versus Local Visual Processing. *Curr. Biol.* **21**, 334–337 (2011).

127. Schyns, P. G., Thut, G. & Gross, J. Cracking the Code of Oscillatory Activity. *PLOS Biol.* **9**, e1001064 (2011).

128. Smith, M. A. & Kohn, A. Spatial and Temporal Scales of Neuronal Correlation in Primary Visual Cortex. *J. Neurosci.* **28**, 12591–12603 (2008).

129. Busch, N. A. & VanRullen, R. Spontaneous EEG oscillations reveal periodic sampling of visual attention. *Proc. Natl. Acad. Sci.* **107**, 16048–16053 (2010).

130. Klimesch, W., Sauseng, P. & Hanslmayr, S. EEG alpha oscillations: The inhibition–timing hypothesis. *Brain Res. Rev.* **53**, 63–88 (2007).

131. Fontolan, L., Morillon, B., Liegeois-Chauvel, C. & Giraud, A.-L. The contribution of frequency-specific activity to hierarchical information processing in the human auditory cortex. *Nat. Commun.* **5**, 4694 (2014).

132. Griffiths, B. J. *et al.* Alpha/beta power decreases track the fidelity of stimulus-specific information. *eLife* **8**, e49562 (2019).

133. Hanslmayr, S., Staudigl, T. & Fellner, M.-C. Oscillatory power decreases and long-term memory: the information via desynchronization hypothesis. *Front. Hum. Neurosci.* **6**, (2012).

134. Palva, S. & Palva, J. M. New vistas for α-frequency band oscillations. *Trends Neurosci.* **30**, 150–158 (2007).

135. Sutterer, D. W., Foster, J. J., Adam, K. C. S., Vogel, E. K. & Awh, E. Item-specific delay activity demonstrates concurrent storage of multiple active neural representations in working memory. *PLOS Biol.* **17**, e3000239 (2019).

136. Chen, L., Cichy, R. M. & Kaiser, D. Alpha-frequency feedback to early visual cortex orchestrates coherent naturalistic vision. *Sci. Adv.* **9**, eadi2321 (2023).

137. Kaiser, D. Spectral brain signatures of aesthetic natural perception in the α and β frequency bands. *J. Neurophysiol.* **128**, 1501–1505 (2022).

138. Stecher, R. & Kaiser, D. Representations of imaginary scenes and their properties in cortical alpha activity. *Sci. Rep.* **14**, 12796 (2024).

139. Jensen, O. & Mazaheri, A. Shaping Functional Architecture by Oscillatory Alpha Activity: Gating by Inhibition. *Front. Hum. Neurosci.* **4**, (2010).

140. Berger, H. Über das Elektrenkephalogramm des Menschen. *Arch. Für Psychiatr. Nervenkrankh.* **87**, 527–570 (1929).

141. Pfurtscheller, G., Stancák, A. & Neuper, Ch. Event-related synchronization (ERS) in the alpha band — an electrophysiological correlate of cortical idling: A review. *Int. J. Psychophysiol.* **24**, 39–46 (1996).

142. Breitmeyer, B. G. & Ganz, L. Implications of sustained and transient channels for theories of visual pattern masking, saccadic suppression, and information processing. *Psychol. Rev.* **83**, 1–36 (1976).

143. Macknik, S. L. & Martinez-Conde, S. The role of feedback in visual masking and visual processing. *Adv. Cogn. Psychol.* **3**, 125–152 (2007).

144. Breitmeyer, B. G. Visual masking: past accomplishments, present status, future developments. *Adv. Cogn. Psychol.* **3**, 9–20 (2007).

145. Haynes, J.-D., Driver, J. & Rees, G. Visibility Reflects Dynamic Changes of Effective Connectivity between V1 and Fusiform Cortex. *Neuron* **46**, 811–821 (2005).

146. Wagner, T., Valero-Cabre, A. & Pascual-Leone, A. Noninvasive Human Brain Stimulation. *Annu. Rev. Biomed. Eng.* **9**, 527–565 (2007).

147. Carricarte, T. *et al.* Laminar dissociation of feedforward and feedback in high-level ventral visual cortex during imagery and perception. *iScience* **0**, (2024).

148. Finn, E. S., Huber, L., Jangraw, D. C., Molfese, P. J. & Bandettini, P. A. Layer-dependent activity in human prefrontal cortex during working memory. *Nat. Neurosci.* **22**, 1687–1695 (2019).

149. van Kerkoerle, T., Self, M. W. & Roelfsema, P. R. Layer-specificity in the effects of attention and working memory on activity in primary visual cortex. *Nat. Commun.* **8**, 13804 (2017).

150. Harris, K. D. & Mrsic-Flogel, T. D. Cortical connectivity and sensory coding. *Nature* **503**, 51–58 (2013).

151. Cichy, R. M., Pantazis, D. & Oliva, A. Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. *Cereb. Cortex* **26**, 3563–3579 (2016).

152. Jozwik, K. M. *et al*. Disentangling five dimensions of animacy in human brain and behaviour. *Commun. Biol.* **5**, 1–15 (2022).

153. Portilla, J. & Simoncelli, E. P. A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *Int. J. Comput. Vis.* **40**, 49–70 (2000).

154. Thaler, L., Schütz, A. C., Goodale, M. A. & Gegenfurtner, K. R. What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Res.* **76**, 31–42 (2013).

155. Brainard, D. H. & Vision, S. The psychophysics toolbox. *Spat. Vis.* **10**, 433–436 (1997).

156. Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D. & Leahy, R. M. Brainstorm: A User-Friendly Application for MEG/EEG Analysis. *Comput. Intell. Neurosci.* **2011**, 1–13 (2011).

157. Guggenmos, M., Sterzer, P. & Cichy, R. M. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage* **173**, 434–447 (2018).

158. Behzadi, Y., Restom, K., Liau, J. & Liu, T. T. A component based noise correction method (CompCor) for BOLD and perfusion based fMRI. *NeuroImage* **37**, 90–101 (2007).

159. Kasper, L. *et al.* The PhysIO Toolbox for Modeling Physiological Noise in fMRI Data. *J. Neurosci. Methods* **276**, 56–72 (2017).

160. Allen, E. J. *et al.* A massive 7T fMRI dataset to bridge cognitive neuroscience and artificial intelligence. *Nat. Neurosci.* **25**, 116–126 (2022).

161. Prince, J. S. *et al.* Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *eLife* **11**, e77599 (2022).

162. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).

163. Julian, J. B., Fedorenko, E., Webster, J. & Kanwisher, N. An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *NeuroImage* **60**, 2357–2364 (2012).

164. Rieger, J. W., Braun, C., Bülthoff, H. H. & Gegenfurtner, K. R. The dynamics of visual pattern masking in natural scene processing: A magnetoencephalography study. *J. Vis.* **5**, 10 (2005).

165. Vapnik, V. *The Nature of Statistical Learning Theory*. (Springer Science & Business Media, 2013).

166. Chang, C.-C. & Lin, C.-J. LIBSVM: a library for support vector machines. *ACM Trans. Intell. Syst. Technol. TIST* **2**, 1–27 (2011).

167. Robinson, A. K., Quek, G. L. & Carlson, T. A. Visual Representations: Insights from Neural Decoding. *Annu. Rev. Vis. Sci.* **9**, 313–335 (2023).

168. Stehr, D. A., Garcia, J. O., Pyles, J. A. & Grossman, E. D. Optimizing multivariate pattern classification in rapid event-related designs. *J. Neurosci. Methods* **387**, 109808 (2023).

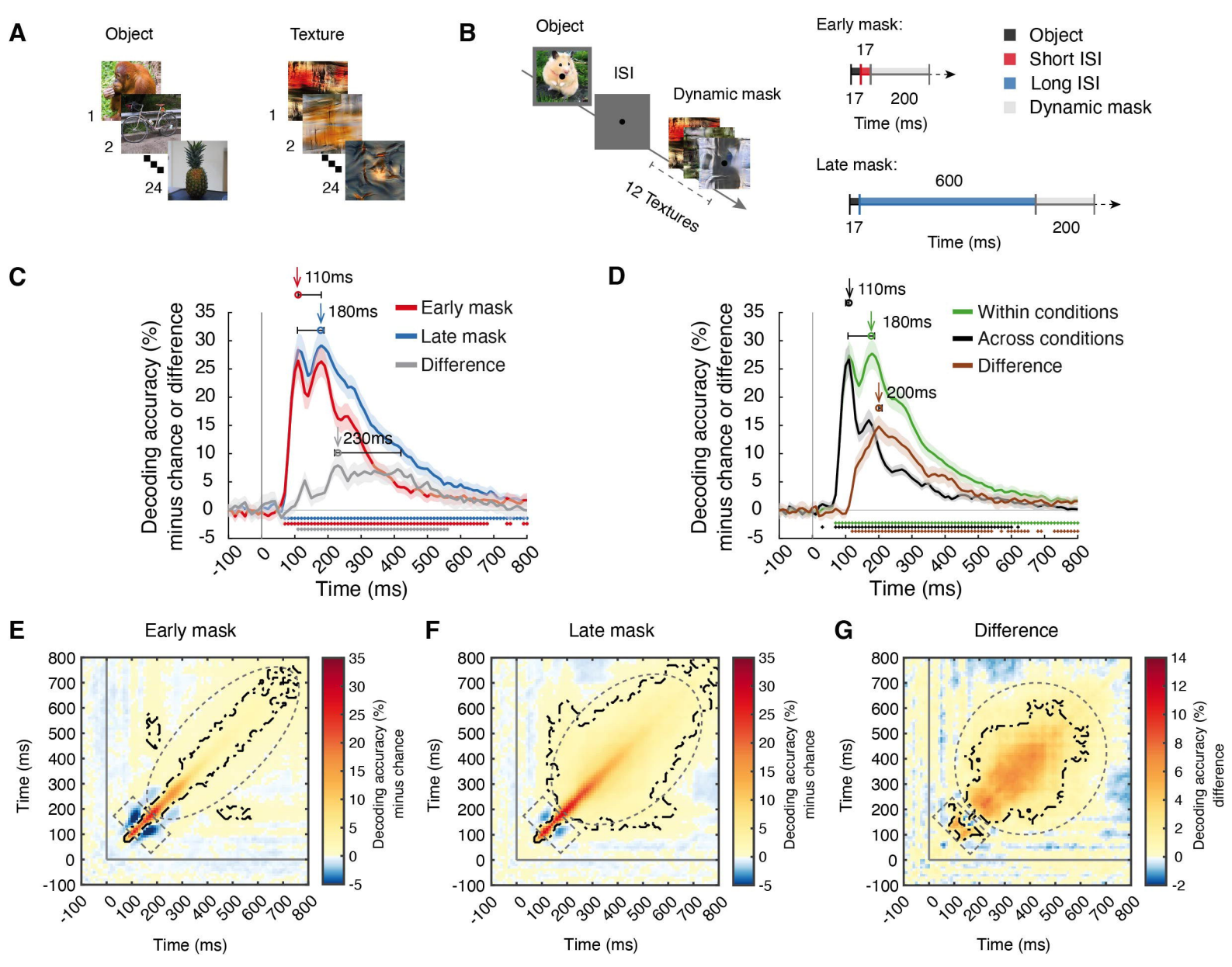
169. Schrimpf, M. *et al. Brain-Score: Which Artificial Neural Network for Object Recognition Is Most Brain-Like?* http://biorxiv.org/lookup/doi/10.1101/407007 (2018) doi:10.1101/407007.

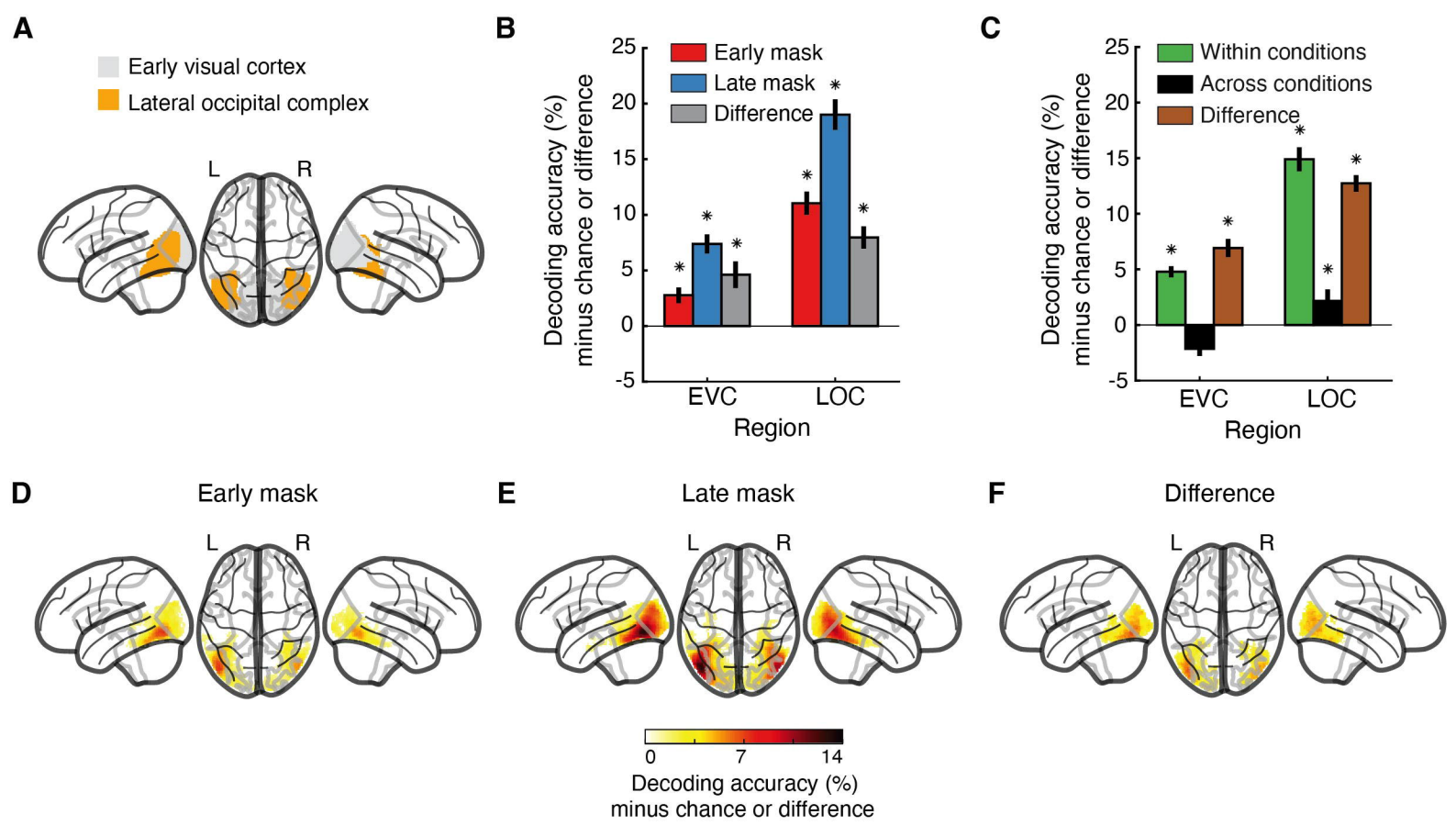
170. Vedaldi, A. & Lenc, K. Matconvnet: Convolutional neural networks for matlab. in 689–692 (2015).

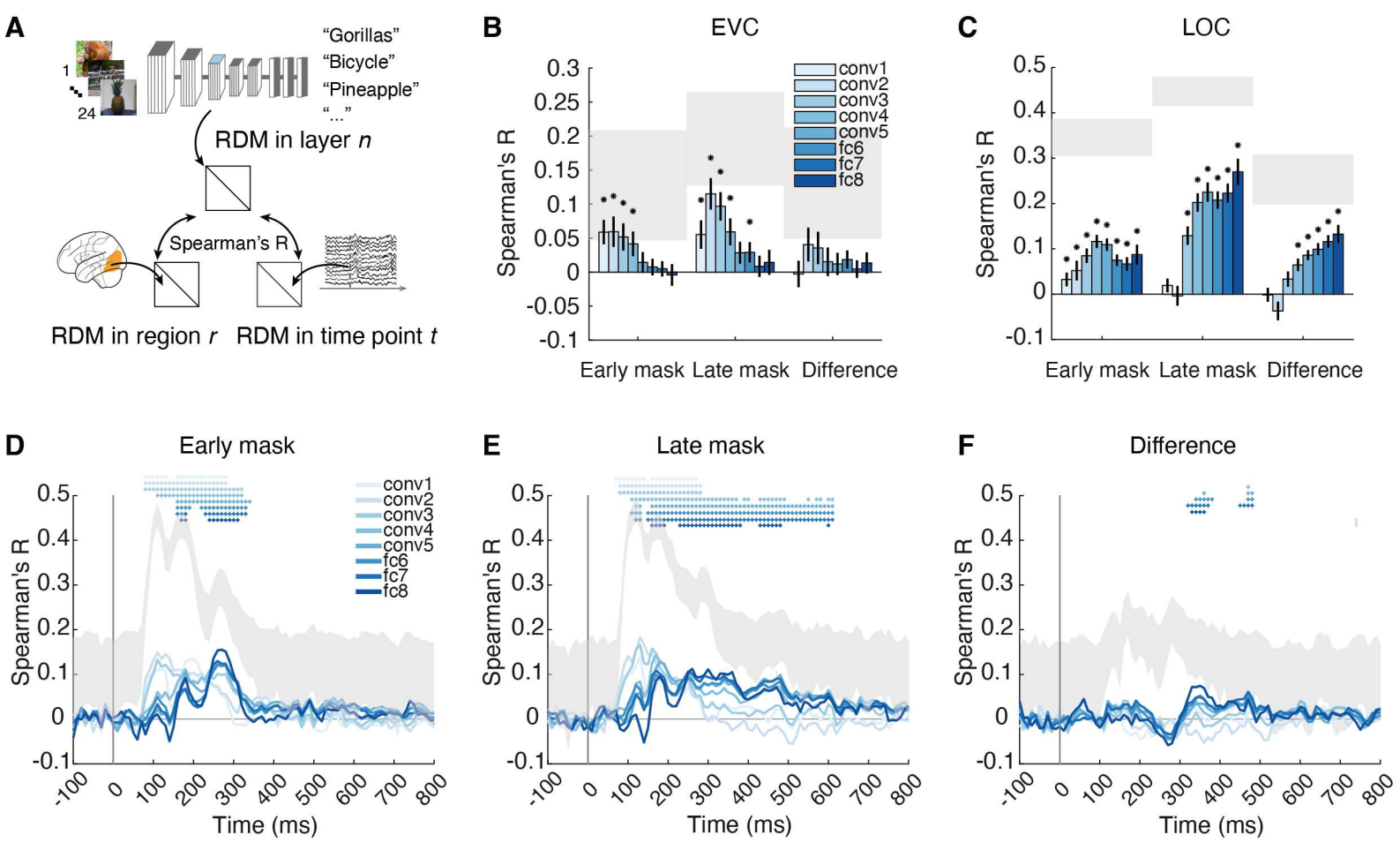
171. Nichols, T. E. & Holmes, A. P. Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Hum. Brain Mapp.* **15**, 1–25 (2002).

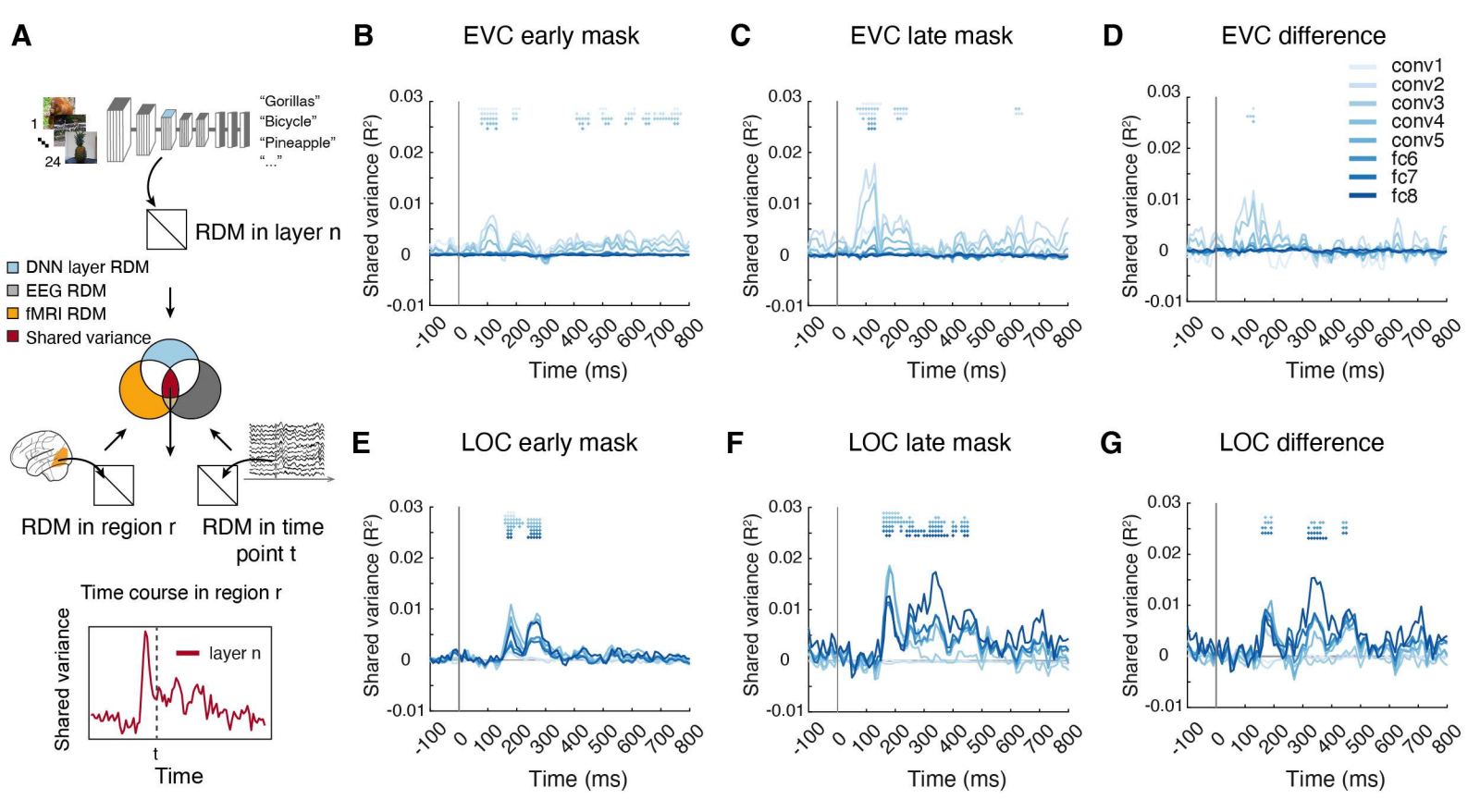
172. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B Methodol.* **57**, 289–300 (1995).

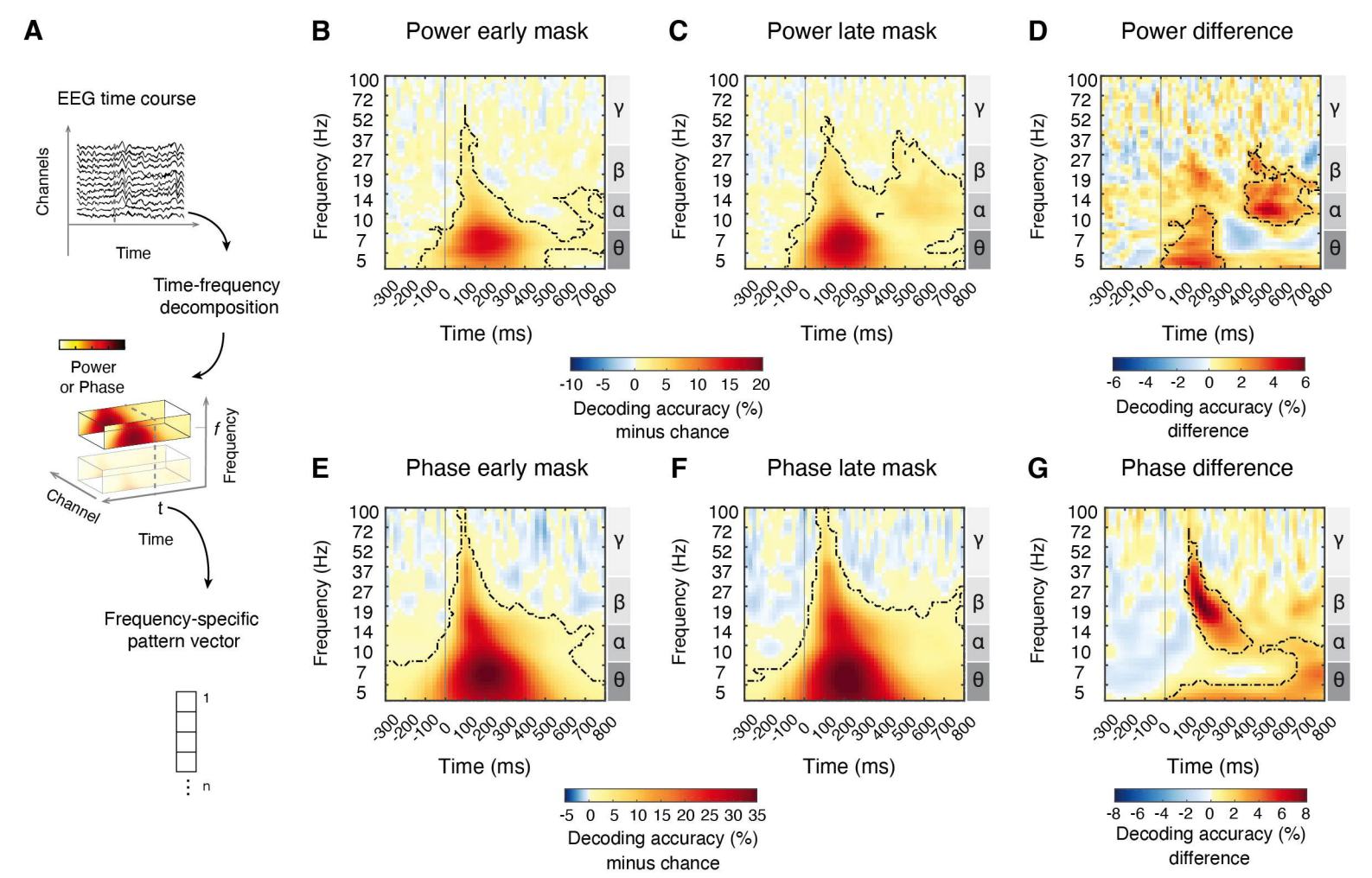
173. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG-and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).

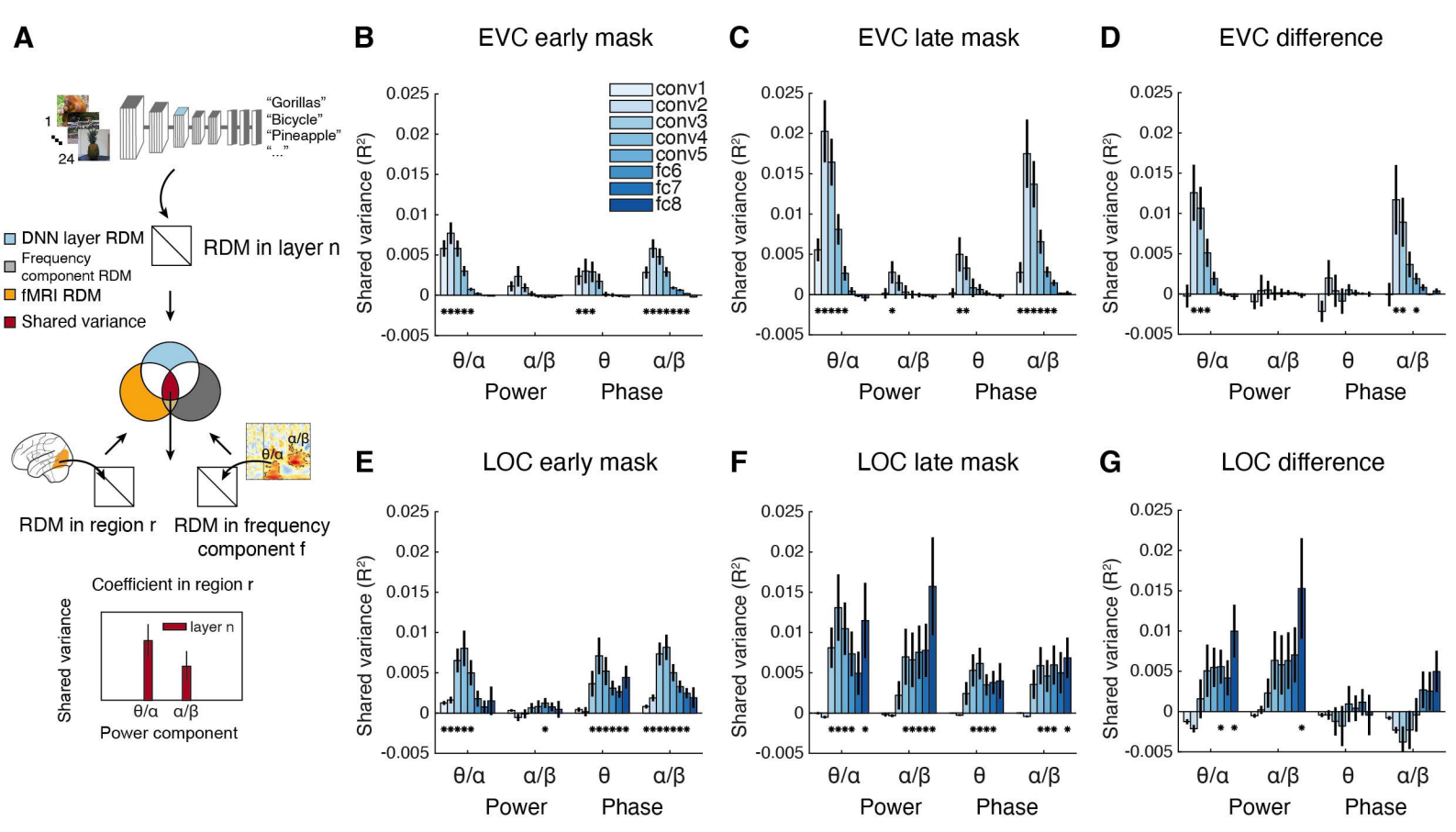

A
Object
Texture
1
2
24
B
Object
ISI
Dynamic mask
12 Textures
Early mask:
17
17
200
Time (ms)
Late mask:
600
17
200
Time (ms)
Object
Short ISI
Long ISI
Dynamic mask
C
110ms
180ms
230ms
Early mask
Late mask
Difference
Decoding accuracy (%)
minus chance or difference
Time (ms)
D
110ms
180ms
200ms
Within conditions
Across conditions
Difference
Decoding accuracy (%)
minus chance or difference
Time (ms)
E
Early mask
F
Late mask
G
Difference
Time (ms)
Decoding accuracy (%)
minus chance
Decoding accuracy (%)
difference


A
Early visual cortex
Lateral occipital complex
L
R
B
Early mask
Late mask
Difference
Decoding accuracy (%)
minus chance or difference
25
20
15
10
5
0
-5
EVC
LOC
Region
C
Within conditions
Across conditions
Difference
Decoding accuracy (%)
minus chance or difference
EVC
LOC
Region
D
Early mask
E
Late mask
F
Difference
0
7
14
Decoding accuracy (%)
minus chance or difference


A
1
24
"Gorillas"
"Bicycle"
"Pineapple"
"..."
RDM in layer n
Spearman's R
RDM in region r
RDM in time point t
B
EVC
Spearman's R
conv1
conv2
conv3
conv4
conv5
fc6
fc7
fc8
Early mask
Late mask
Difference
C
LOC
Spearman's R
Early mask
Late mask
Difference
D
Early mask
Spearman's R
Time (ms)
E
Late mask
Spearman's R
Time (ms)
F
Difference
Spearman's R
Time (ms)


A
1
24
"Gorillas"
"Bicycle"
"Pineapple"
"..."
RDM in layer n
DNN layer RDM
EEG RDM
fMRI RDM
Shared variance
RDM in region r
RDM in time point t
Time course in region r
Shared variance
layer n
t
Time
B
EVC early mask
C
EVC late mask
D
EVC difference
conv1
conv2
conv3
conv4
conv5
fc6
fc7
fc8
E
LOC early mask
F
LOC late mask
G
LOC difference
Shared variance (R²)
0.03
0.02
0.01
0
-0.01
-100
0
100
200
300
400
500
600
700
800
Time (ms)


A
EEG time course
Channels
Time
Time-frequency decomposition
Power or Phase
Frequency
f
Channel
t
Time
Frequency-specific pattern vector
1
n
B
Power early mask
C
Power late mask
D
Power difference
Frequency (Hz)
100 72 52 37 27 19 14 10 7 5
γ
β
α
θ
-300 -200 -100 0 100 200 300 400 500 600 700 800
Time (ms)
-10 -5 0 5 10 15 20
Decoding accuracy (%) minus chance
-6 -4 -2 0 2 4 6
Decoding accuracy (%) difference
E
Phase early mask
F
Phase late mask
G
Phase difference
-5 0 5 10 15 20 25 30 35
Decoding accuracy (%) minus chance
-8 -6 -4 -2 0 2 4 6 8
Decoding accuracy (%) difference


A
1
24
"Gorillas"
"Bicycle"
"Pineapple"
"..."
DNN layer RDM
Frequency component RDM
fMRI RDM
Shared variance
RDM in layer n
RDM in region r
RDM in frequency component f
θ/α
α/β
Coefficient in region r
Shared variance
layer n
θ/α α/β
Power component
B
EVC early mask
C
EVC late mask
D
EVC difference
E
LOC early mask
F
LOC late mask
G
LOC difference
conv1
conv2
conv3
conv4
conv5
fc6
fc7
fc8
Shared variance (R²)
0.025
0.02
0.015
0.01
0.005
0
-0.005
θ/α α/β θ α/β
Power Phase