Layer-specific spatiotemporal dynamics of feedforward and feedback in human visual object perception

Tony Carricarte[1,2,3*], Siying Xie[1+], Johannes Singer[1+], Robert Trampel[4], Laurentius Huber[5], Nikolaus Weiskopf[4,6,7] Radoslaw M. Cichy[1,2,3,8]

[1]Department of Education and Psychology, Freie Universität Berlin, 14195 Berlin, Germany
[2]Einstein Center for Neurosciences Berlin, Charité - Universitätsmedizin Berlin, 10117 Berlin, Germany
[3]Bernstein Center for Computational Neuroscience Berlin, Humboldt-Universität zu Berlin, 10117 Berlin, Germany
[4]Department of Neurophysics, Max Planck Institute for Human Cognitive and Brain Sciences, 04103 Leipzig, Germany
[6]Felix Bloch Institute for Solid State Physics, Faculty of Physics and Earth Sciences, Universität Leipzig, 04103 Leipzig, Germany
[7]Wellcome Centre for Human Neuroimaging, UCL Queen Square Institute of Neurology, University College London, WC1N 3AR London, United Kingdom
[8]Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, 10117 Berlin, Germany

+ These authors contributed equally
* Correspondence: tcarricarte@gmail.com

Abstract

Visual object perception is mediated by information flow between regions of the ventral visual stream along feedforward and feedback anatomical connections. However, feedforward and feedback signals during naturalistic vision are rapid and overlapping, complicating their identification and precise functional specification. Here we recorded human layer-specific fMRI responses to naturalistic object images in early visual cortex (EVC) and lateral occipital complex (LOC) to isolate feedforward and feedback information signals spatially by their cortical layer specific termination pattern. We combined these layer-specific fMRI responses with electroencephalography (EEG) responses for the same images to segregate feedforward and feedback signals in both time and space. Feedforward signals emerge early in the middle layers of EVC and LOC, followed by feedback signals in the superficial layer of both regions, and the deep layer of EVC. Comparing the identified dynamics in LOC to a visual deep neural network (DNN), revealed that early feedforward signals in LOC encode medium complexity features, whereas later feedback signals increase the representational format to high complexity features. Together this specifies the spatiotemporal dynamics and functional role of feedforward and feedback information flow mediating visual object perception.

Introduction

Human object vision relies on anatomical bidirectional connections along the ventral visual stream[1,2], spanning the visual hierarchy from early visual cortex (EVC) to the lateral occipital complex (LOC)[3]. These connections mediate visual computations via feedforward and feedback information flows, with complex overlapping spatiotemporal dynamics[4–6]. While the feedforward flow carries sensory information up the visual processing hierarchy[7,8], downstream feedback concurrently carries information down the hierarchy, refining and shaping feedforward neural dynamics[9]. Identifying the distinct contributions of these information flows to visual computation in space and time is therefore crucial for understanding the mechanism of human object vision.

Previous efforts to disentangle the specific signatures of feedforward from feedback have typically used one of two main approaches. The first involves invasive manipulations on the anatomically-[8] and functionally-defined[6,10,11] neural circuits in non-human primates, selectively targeting bottom-up vs top-down processes. While these interventions offer precise circuit-level control, their invasive nature makes them largely impractical for human research. The second approach comprises contrasting experimental conditions where the feedforward and feedback contributions vary, such as perception vs mental imagery[12], attention vs non-attention[13], and early vs late backward masking[5]. However, this method does not directly assess information flow during naturalistic vision and has conceptual limitations, as it relies on indirect comparisons of experimental conditions that may be confounded by incompletely controlled differences between them[14–16].

Instead, here we capitalized on the layer-specific anatomical connectivity found in the primate visual cortex[7,8,17], to dissect feedforward from feedback signals: while feedforward connections terminate primarily in the middle layer[18], feedback connections target superficial and deep layers[19,20].

Based on this three-compartment model of cortical depth, we characterized the layer-specific spatiotemporal dynamics of feedforward and feedback signals in object perception. Using sub-millimeter resolution fMRI at 7T, we recorded layer-specific brain activity in human EVC and LOC while participants viewed naturalistic object images. We then combined these layer-specific fMRI responses with time-resolved EEG responses for the same images within the framework of representational similarity analysis (RSA)[21–23] to resolve feedforward and feedback processing in millisecond resolution. Based on this, we then characterized the representational format in terms of visual feature complexity of feedforward and feedback processing using artificial deep neural networks.

Results

Using sub-millimeter resolution 7T MRI, we recorded gradient-echo blood oxygenation level-dependent (GE-BOLD) signals in human EVC and LOC in response to 24 different naturalistic object images (**Fig. 1A**). During the acquisition participants viewed naturalistic object images on real-world backgrounds in random order (**Fig 1B**). We estimated the neural response to each condition, i.e., object image, by fitting a general linear model.

Robust object information in both EVC and LOC is a precondition for further dissecting feedback and feedforward-related aspects. To ascertain this, we extracted voxel values from EVC and LOC separately to form condition-specific pattern vectors (**Fig. 1C**). Based on these pattern vectors we trained and tested a support vector machine (SVM) to perform pairwise-object classification of objects. We found robust decoding accuracy in both EVC (71,16 %, *P* = 0.0039, one-sample permutation test) and LOC (60,69 %, *P* = 0.0092), ascertaining reliable object information in those regions (**Fig. 1D**).

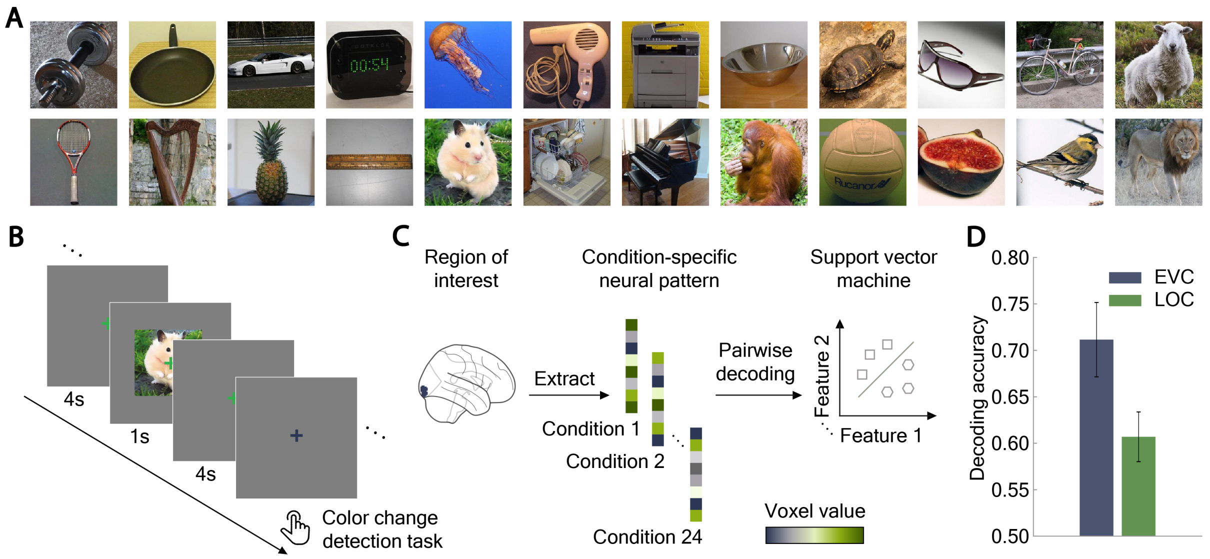


Figure 1. Stimuli, experimental design and multivariate pattern classification. (A) Stimulus set. The stimuli consisted of 24 different naturalistic object images. **(B)** fMRI experimental design. On each trial, participants viewed images for 1 s followed by a 4-s baseline interval. Participants were required to perform a color-change detection task on the fixation cross that occurred randomly throughout the experiment. (**C**) Extraction of voxel values to form condition-specific pattern vectors from region of interest (example here: EVC) and pairwise-object classification using a support vector machine. ROIs are depicted for visualization purposes only (**D**) Object-pairwise multivariate decoding output. Robust object-specific information was reliably decoded from EVC (71,16 %, *P* = 0.0039) and LOC (60,69 %, *P* = 0.0092) using one-sample permutation tests. Error bars indicate the standard error of the mean across participants.

We then proceeded to identify and examine the spatiotemporal neural dynamics of visual representations in two steps that build upon each other. First, we assessed the macroscale of cortical regions (i.e., EVC and LOC) to establish and thus validate representational EEG-fMRI fusion at 7T, and to characterize the representational format of the identified dynamics. Based on this validation, we dissect feedforward from feedback neural processing at the finer mesoscale level of cortical layers.

Spatiotemporal neural dynamics of object representations in EVC and LOC at the macroscale of brain regions

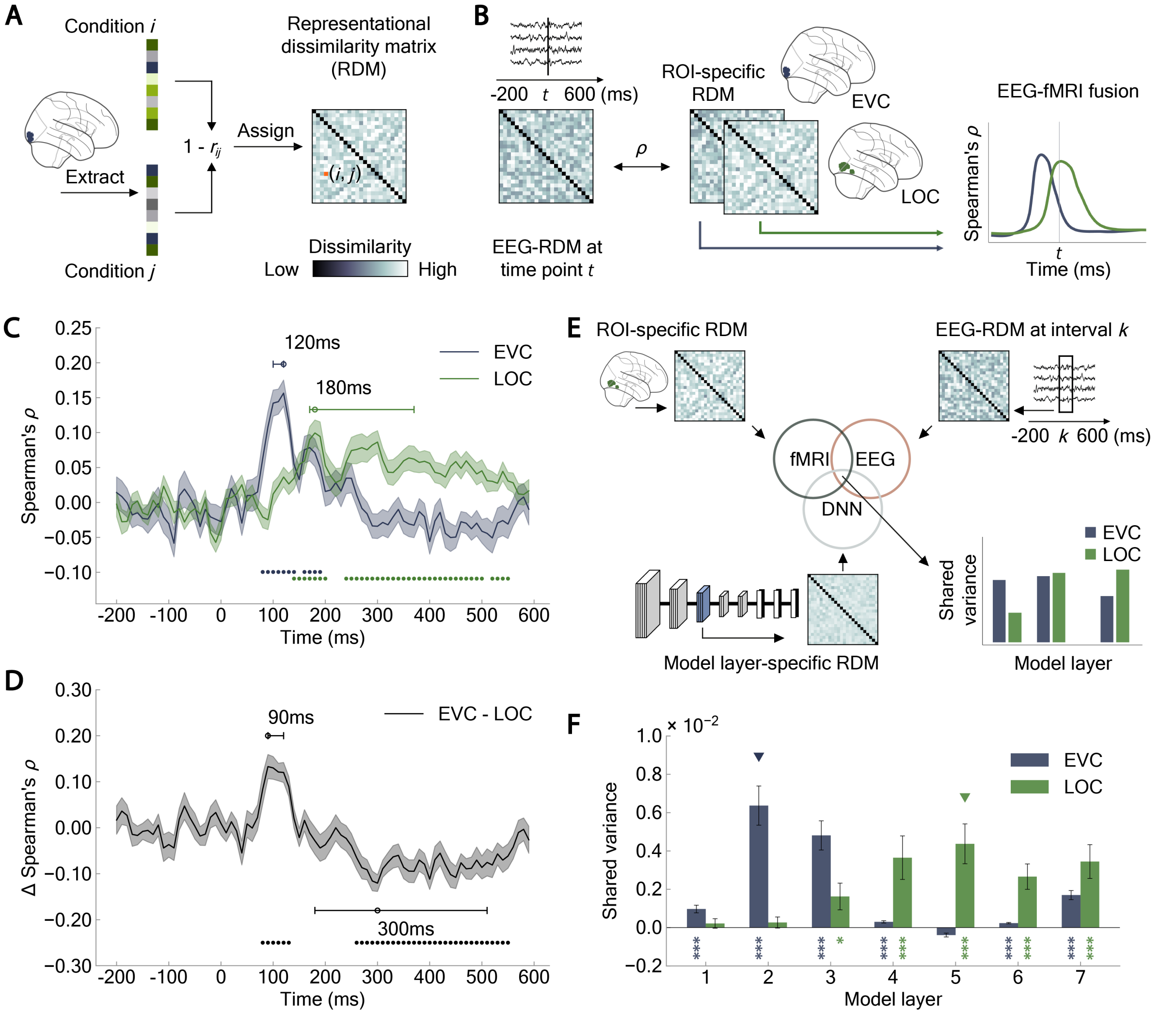


Figure 2. Representational EEG-fMRI fusion at the macroscale. (**A**) Representational similarity analysis. For each condition, we extracted the neural pattern from the region of interest (example here: EVC). We assessed the extent of pattern dissimilarity by calculating 1 – Pearson's correlation for all combinations of experimental conditions (*i*, *j*) and assigned the dissimilarity values to an fMRI representational dissimilarity matrix (RDM) indexed by the conditions in rows and columns, at entry (*i*, *j*). ROIs are depicted for visualization purposes only (**B**) Representational EEG-fMRI fusion. For each time point *t*, we correlated the EEG-RDM to the fMRI-RDMs of EVC and LOC using Spearman's rank order correlation. (**C**) Spatiotemporal neural dynamics at the macroscale level. Time course in EVC peaked earlier than in LOC. (**D**) Difference between EVC and LOC curves in **C**. EEG signals correlated first more with EVC than with LOC and later more with LOC than with EVC. Shaded area indicates the standard error of the mean across participants; colored circles indicate significant time points ($N = 32$, cluster-defining threshold $P < 0.05$, cluster threshold $P < 0.05$); uncolored circles and horizontal lines indicate peak latency means and 95% confidence intervals, respectively. (**E**) Commonality analysis. For each AlexNet layer, we correlated its RDM to each ROI-specific fMRI-RDM and the mean EEG-RDM at the time interval with significant ROI-specific temporal dynamics. (**F**) Format of representation (≈ feature complexity) in EVC and LOC. Visual representations of low-complexity emerge early in EVC, while mid-to-high-level object representations emerge later in LOC. Error bars indicate the standard error of the mean across participants; colored asterisks indicate significant correlations ($N = 32$, right-tailed permutation tests, FDR-corrected; $*P < 0.05$; $**P < 0.01$; $***P < 0.001$); colored triangles represent model layers with the highest occurrence proportion, determined through 1,000-iteration bootstraps.

To establish the time course of object processing in EVC and LOC at the macroscale, we employed representational EEG-fMRI fusion[22–24], which integrates time-resolved EEG with spatially resolved fMRI measurements for a combined spatiotemporally resolved view. For MRI, we used the dataset collected in this study, complemented with EEG responses from an existing dataset to the same set of 24 images[5] as in the MRI experiment. We assessed the representational geometry of EVC and LOC with region-specific representational dissimilarity matrices (RDMs; **Fig. 2A**), and the representational geometry of EEG signals with RDMs in a time-resolved way from -200 to 600 ms with respect to image onset (**Fig. 2B**). We related EVC and LOC RDMs to the time-resolved EEG-RDMs, yielding a time course of representational similarity for EVC and LOC each, for which we report peak latency and peak latency differences with 95% confidence intervals in parenthesis as assessed by bootstrapping (1,000 iterations).

As expected from the well-established hierarchical organization of the visual system[7,8,25], object processing in EVC preceded that in LOC (**Fig. 2C**): The EVC time course peaked at 120 ms (100 – 120 ms), whereas the LOC time course peaked later at 180 ms (170 – 370 ms), with a significant difference between peak times of 60 ms (50 – 250 ms; $P < 0.001$). Furthermore, in direct comparison by subtraction of the correlation curves of LOC from EVC, early representations correlated stronger with EVC than with LOC at 90 ms (90 – 120 ms; **Fig. 2D**), whereas late representations correlated stronger with LOC than with EVC at 300 ms (180 – 510 ms). A supplementary EEG-fMRI fusion analysis accounting for similarities in representational geometry between LOC and EVC by partial correlations confirmed these observations (**Suppl. Fig. 1A and B**). Our results extend EEG-fMRI fusion[5,22,24,26] from 3T to 7T fMRI and provide additional support to the procedure, thus warranting a spatially finer investigation of the spatiotemporal neural dynamics at the mesoscale level of cortical layers.

The format of object representations in EVC and LOC at the macroscale of brain regions

The visual cortex represents objects in formats of increasing feature complexity from low to high along the ventral visual stream[27–30]. To confirm this progression here, we related the neural dynamics of EVC and LOC to the AlexNet deep neural network (DNN) trained on object categorization[31]. The underlying rationale is that the feature complexity of visual representations progressively increases from lower to higher layers across the network's layers, so that assessing the fit of the model layers to brain data reveals feature complexity of the neural representations[30,32]. We conducted commonality analysis, linking each of the 7 layer-specific DNN-RDMs to the ROI-specific fMRI-RDMs from EVC and LOC, and to the mean EEG-RDM at corresponding time intervals, defined by the significant time points of the raw curves in Fig. 2C. We report peak layers with 95% confidence intervals in parenthesis (1,000 bootstraps).

We found commonality with predominantly low model layers, with a peak at model layer 2 (2 – 2) in EVC. and with predominantly middle to high model layers with a peak at model layer 5 (4 – 5) in LOC (**Fig. 2F**). This indicates representations of primarily low complexity in EVC and mid- to high-complexity in LOC. This pattern of results remained consistent when using an alternative experimental choice based on significant time points from the difference between EVC and LOC curves in Fig. 2D (**Suppl. Fig. 2**). Our results thus confirm a transition from representations of low complexity processed early in EVC into representations of higher complexity processed later in LOC[28], further providing additional support to the analytic approach at 7T and warranting a finer investigation of representational format at the mesoscale level.

Spatiotemporal neural dynamics of object representations in EVC and LOC at the mesoscale of cortical layers

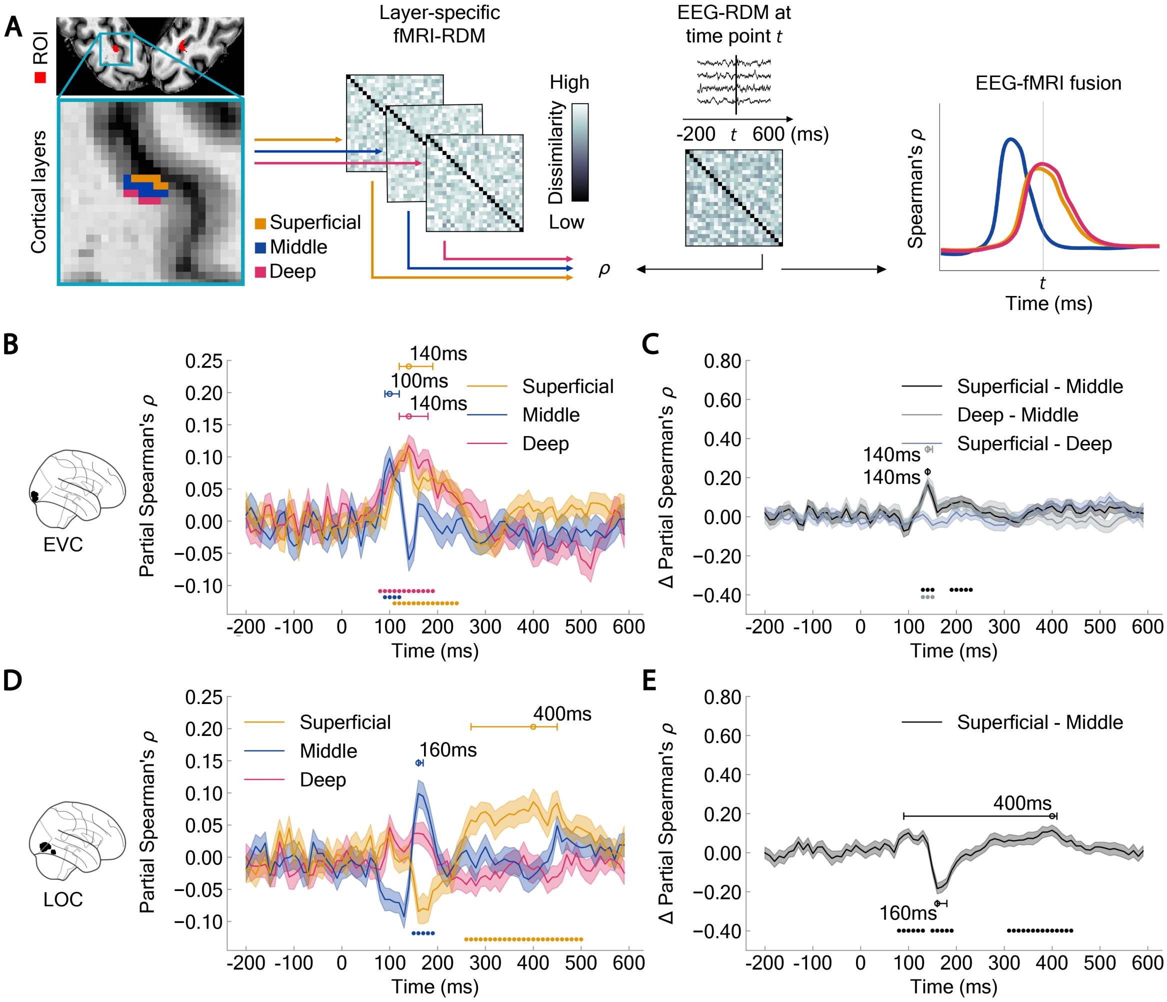


Figure 3. Representational EEG-fMRI fusion at the mesoscale. (**A**). For each cortical layer — deep, middle, and superficial — (example here: for EVC) we computed the partial Spearman's rank-order correlation between its layer-specific fMRI-RDM and the EEG-RDM at each time point *t*. (**B**) Layer-specific spatiotemporal neural dynamics in EVC. EEG signals correlated early across layers, with the time course in the middle layer peaking earlier than in the deep and superficial layers. (**C**) Difference between EVC layer curves in **B**. EEG signals correlated lately more with deep and superficial layers than with the middle layer (**D**) Layer-specific spatiotemporal neural dynamics in LOC. EEG signals correlated early in the middle layer and later in the superficial layer. (**E**) Difference between LOC layer curves in **D**. EEG signals correlated early more with the middle layer than with the superficial layer, and later more with the superficial layer than with the middle layer. Shaded area indicates the standard error of the mean across participants; colored circles indicate significant time points (N = 32, cluster-defining threshold $P < 0.05$, cluster threshold $P < 0.05$); uncolored circles and horizontal lines indicate peak latency means and 95% confidence intervals, respectively.

To investigate the spatiotemporal neural dynamics at the mesoscale level, we applied the research procedure validated above to the finer level of cortical layers (**Fig. 3A**). To this end, we segmented the cortical ribbon into three equidistant layers[33]: deep, middle and superficial. We then applied representational EEG-fMRI fusion as established above, but now at the level of layers, to yield time-resolved and layer-specific visual object processing time courses in EVC and LOC. To control for non-specific macrovascular responses[34] that affect layer specificity[35], we conducted EEG-fMRI fusion analysis partialing out for each layer the effect of the layers beneath.

In EVC we observed a correlation pattern suggesting two processing stages (**Fig. 3B, C**) indexed by different profiles across layers and in timing. The first stage is marked by a peak in the middle layer at 100 ms (90 – 120 ms; **Fig. 3B**). The second stage is characterized by peaks at 140 ms in both the deep (120 –180 ms) and superficial (120 – 190 ms) layers, with significant latency differences of 40 ms (10 – 50 ms; *P* = 0.004) between the middle layer and the deep layer and 40 ms (20 – 80 ms; *P* = 0.002) between the middle layer and superficial layer, but not between the deep layer and the superficial layer 0 ms (-20 – 40 ms; *P* = 0.814). This pattern was further substantiated by subtracting the layer-specific time courses which showed significant effects (**Fig. 3C**). We observed stronger correlations of late representations at 140 ms with the deep (140–150 ms) and the superficial (140–140 ms) layers than with the middle layer. This suggests initial feedforward processing in the middle layer and a later emerging feedback processing with a distinctive layer profile in the deep and superficial layers in EVC.

In LOC, the results pattern also indicated two stages with a distinctive layer and temporal profile (**Fig. 3D, E**). We observe earlier processing in the middle layer, followed by processing in the superficial layer later (**Fig. 3D**). In detail, time courses of object processing peaked first in the middle layer at 160 ms (160 – 170 ms), and later in the superficial layer at 400 ms (270 – 450 ms), with a significant difference in peak latency of 240 ms (110 – 290 ms; *P* < 0.012). A direct comparison of time courses by subtraction (**Fig. 3E**) substantiated the observation in that representations correlated stronger with the middle layer early than the superficial layer at 160 ms (160 – 180 ms), and late representations correlated stronger with the superficial layer than the middle layer later, at 400 ms (90 – 410 ms). This suggests an initial feedforward processing in the middle layer and a later emerging feedback processing with a distinctive layer profile in the superficial layer in LOC.

An analogous results pattern in EVC (**Suppl. Fig. 3A, B**) and LOC (**Suppl. Fig. 3C, D**) was observed when assessing layers directly without partialing out the effect of deeper layers, confirming the robustness of the results to particular analysis choices.

Together, this resolves the spatiotemporal dynamics of feedforward and feedback information flow during visual object processing across EVC and LOC through cortical layer-specific and temporally distinct response profiles.

The format of object representations in EVC and LOC at the mesoscale of cortical layers

One interpretation of the observed layer-specific spatiotemporal dynamics, being guided by the functional pattern of cortical layer connectivity[36,37], is that they indicate interareal communication via feedforward and feedback connections[38,39]. However, an alternative explanation is that they arise intra-areal communication via lateral connections[40] or from superficial bias in the fMRI measurements[41].

To disambiguate between these options, we characterized the format of representations by assessing feature complexity as present in DNN layers, from low to high, applying the research approach as used at the macroscale but now at the level of cortical layers. If the observed layer-specific spatiotemporal dynamics reflect an interplay of feedforward and feedback information flow, we would expect the flow to carry information of varying complexity[42,43] across different levels of the visual hierarchy[44,45], resulting in distinct representational formats across cortical layers. In contrast, if the observed layer-specific spatiotemporal dynamics reflect lateral connections modulating the neural gain[46–50] or a superficial bias, we would expect uniform feature complexity across layers[51].

We conducted commonality analysis, linking each model layer-specific DNN-RDM to the deep, middle and superficial layer-specific fMRI-RDM and mean EEG-RDM at corresponding time intervals, defined by the significant time points of the raw curves in Fig. 3B and D indicating periods of layer-specific neural dynamics (**Fig. 4A**).

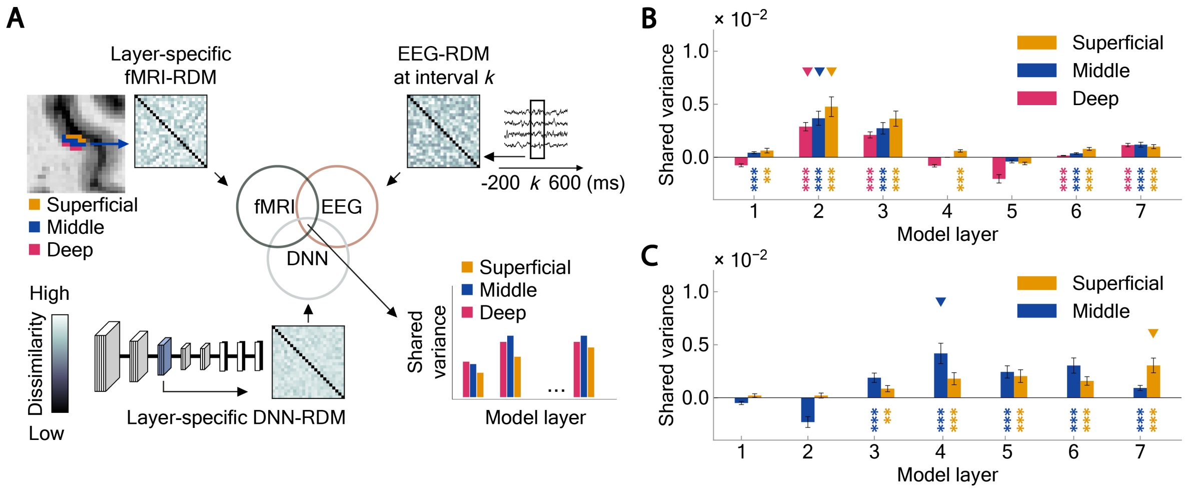


Figure 4. Commonality analysis between fMRI, EEG and AlexNet at the mesoscale. (**A**) Procedure. For each AlexNet layer, we correlated its RDM to each layer-specific fMRI-RDM in EVC and LOC and the mean EEG-RDM at the time interval with significant layer-specific temporal dynamics. (**B**) Format of representation (≈ feature complexity) across cortical layers in EVC. Low model layers correlated strongly across layers in EVC. (**C**) Format of representation (≈ feature complexity) across cortical layers in LOC. Middle model layers correlated strongly with the middle layer in LOC, while high model layers

correlated primarily with the superficial layer. Colored asterisks indicate significant correlations (N = 32, right-tailed permutation tests, FDR-corrected; *P < 0.05; **P < 0.01; ***P < 0.001); colored triangles represent model layers with the highest occurrence proportion, determined through 1,000-iteration bootstraps.

In EVC (**Fig. 4B**) we observed a uniform results pattern across cortical layers: object representations shared variance with all DNN layers, but most strongly with model layer 2 (2 – 3), indicating mainly representations of low-to-mid level complexity. This suggests that the observed layer-specific spatiotemporal dynamics in EVC reflect lateral connections modulating the neural gain[46–50] or superficial bias[41] rather than feedback.

In contrast, in LOC (**Fig. 4C**) we observed a shift from mid-to high feature complexity across cortical layers. The early emerging representations in the middle cortical layer (Fig. 3D) were of mid-complexity, as indicated by a peak at model layer 4 (4 – 6). In contrast, the late emerging representations in the superficial layers were of high complexity, with a peak at model layer 7 (4 – 7). This results pattern was robust: making alternative experimental choices based on the average of the significant time intervals from the difference between layer curves in Fig. 3C and E (**Suppl. Fig. 4 and 5**) and individual time points spanning the full-time course (**Suppl. Fig. 6**) yielded equivalent results. Together, these findings suggest a dynamic shift in representational format in LOC, transitioning from early representations of mid-level feature complexity in the feedforward flow to later representations of high feature complexity through interareal feedback.

Discussion

We leveraged layer-resolved fMRI, time-resolved EEG data and DNNs to identify and characterize the spatiotemporal neural dynamics of feedforward and feedback information flow underlying object perception. We validated our methods using 7T fMRI at the macroscale of cortical regions by replicating the temporal dynamics and representational format in EVC and LOC observed in 3T fMRI studies[5,22,26,30], allowing us to proceed to the mesoscale of cortical layers. There we made two key observations. First, we observed distinct layer-specific temporal profiles for EVC and LOC. Visual representations in the middle layers emerged earlier than in the deep and superficial layers of EVC and the superficial layers of LOC. Second, the identified layer-specific dynamics in LOC had distinctive visual feature complexity profiles: the early emerging middle layer representations were of mid-complexity and the later emerging superficial layer representations were of high complexity.

Layer-specific EEG-fMRI fusion reveals sequential feedforward and feedback processing in visual object perception

Visual object perception unfolds through intricate spatiotemporal neural dynamics[4–6,52,53], mediating feedforward and feedback information flow[22,54] rapidly and through temporally overlapping responses. Here, we disentangle the temporal dynamics of feedforward from feedback signals by leveraging the anatomical canonical microcircuit of cortex[7,8] at the input and output stage of cortical object processing[3,55,56] – EVC and LOC. We find that feedforward signals emerge early in the middle layer of both EVC at 100 ms and in LOC at 160 ms, while feedback appears relatively later in deep and superficial layers of EVC at 140 ms and superficial layer of LOC at 400 ms, supporting the idea of sequential processing of visual information first through feedforward, then through feedback processing[37,57]. Our results thus provide a functional temporal characterization for the anatomical canonical cortical microcircuit model of feedforward and feedback connectivity in the human ventral visual stream.

Our findings have theoretical implications. For example, they specify in spatiotemporal terms the dynamical communication model in predictive coding theory[58] which posits distinct neural channels for the transmission of sensory and predictive information[59]. Our results indicate that whereas sensory signals are convened early in middle layers, subsequent predictive signals are transmitted later as feedback in deep and superficial layers. This specification opens a new path for further testing predictions of predictive coding by investigating the content and integration of predictive feedback and sensory feedforward signals[60–62]. For example, our results invite investigating layer-specific neural dynamics at distinct frequency bands for feedforward and feedback processing as predicted from human[63,64] and non-human invasive studies[59,65–67].

Our approach facilitates resolving the interplay of feedforward and feedback processing in a variety of visual research contexts. It may allow dissecting the distinctive roles of feedforward and feedback information flow crucial to cognitive functions such as attention, expectation and memory by identifying and characterizing their distinct spatiotemporal dynamics. Similarly, using multi- rather than univariate methods[12,59,68] to assess layer-specific fMRI extends content-sensitive analysis[69–71] from the macro- to the mesoscale of cortical organization, allowing for the assessment of the representational contents of feedforward and feedback information flow underlying human cognition.

Layer-specific EEG-fMRI fusion clarifies neural dynamics of high-level visual cortex regions

Neural dynamics in response to faces[22], scenes[72] and objects[73,74] assessed at the macroscale of cortical regions commonly display a double peak pattern (see also **Fig. 2B**), with a sharp, early peak around 100-130 ms followed by a wide second peak around 200-450 ms, suggesting distinct contributing neural circuits and kinds of processing. Our layer-specific results clarify the neural circuits[22] and functional nature of the components underlying this pattern: it results from mixing the early, narrow-peaked and neural dynamics at the middle layer ~160 ms of feedforward information flow with the later, wide-peaked and thus more persistent dynamics at the superficial layer ~400 ms of feedback information flow. Whereas the early peak latency matches that of sensory feedforward signals[75], the late peak dynamics match that of and attention-[76], consciousness-[77] or task-related[78] feedback, originating from higher-order brain regions, such as the frontal eye fields[10] or the prefrontal cortex[79], outside the visual cortex. Thus, our results provide circuit-level mechanistic as well as functional interpretative guidance for standard 3T human neuroimaging studies that typically cannot resolve visual information flow at this fine spatiotemporal level.

Differential representational format across layer-specific dynamics in LOC implies interareal feedback mediating high-complexity visual features

The analysis of the representational format of the neural dynamics in LOC revealed that early feedforward signals primarily convey mid-complexity features, while feedback signals convey high-complexity features. This indicates that feedback does not merely modulate the gain of pre-existing features but is actively involved in the emergence of additional, more complex features that are absent in feedforward processing[5]. This finding supports the notion that interareal feedback is integral to core object recognition[80–82]. A potential source of this feedback might be the dorsolateral prefrontal cortex (DLPFC)[83], whose silencing decreases feature complexity in monkey inferior temporal cortex[84], the homologue of human LOC. Based on our results we predict that disrupting processing in human DLPFC, e.g. through transcranial magnetic stimulation[85,86], will yield analogous effects specifically on superficial, late cortical layer responses in LOC.

Limitations of the study

We based our analyses on the GE-BOLD signal that is affected by locally nonspecific responses from macrovasculature[41], compromising the estimation of the laminar response. To mitigate this effect, we used Pearson's correlation as a scale-invariant measure and partialed out effects of lower cortical layers when assessing the layers above. However, some residual bias may persist, particularly if the GE-BOLD response across layers follows a point spread function[87] that a linear model cannot fully account for. Human fMRI studies using contrasts with higher spatial specificity[88–90], optimized for condition-rich experimental designs[21], are needed to confirm our findings.

Conclusion

Understanding how the abundant feedforward and feedback connections in visual cortex mediate object vision requires specifying their functional role. Here we provided two key advances in this regard. First, leveraging layer-specific fMRI with EEG-fMRI fusion we dissociated temporally overlapping feedforward and feedback processing in LOC and EVC, specifying their unique temporal profile. Second, by assessing feature complexity, we showed that feedback in LOC actively increases the complexity of LOC's feature format.

Acknowledgments

This work was funded by the German Research Foundation (DFG, CI241/1-1, CI241/3-1, INST 272/297-1 to R.M.C.), by the European Research Council (ERC) starting grant (ERC-2018-STG 803370 to R.M.C.), supported by the NIH Intramural Program of NIMH/NINDS (#ZIC MH002884 to R. H.) and by the Einstein Center for Neuroscience (to T.C.). Computing resources were provided by the high-performance computing facilities at ZEDAT, Freie Universität Berlin, Germany.

Materials and Methods

fMRI participants

10 adult volunteers (mean age 29.4 years; age range 20-37 years; 5 female) participated in the study and provided written informed consent. The sample size was based on previous conventional layer fMRI studies conducted at 7T[68,88,91]. All participants had normal or corrected-to-normal vision and no history of neurological disorders. All participants received a monetary reward at the end of the study. The study was approved by the Ethics Committee of the Faculty of Medicine at Leipzig University, Germany, and conducted in accordance with the ethical principles of the Declaration of Helsinki, except for study preregistration, which was not performed.

Stimulus set

The stimulus set consisted of 24 naturalistic color images of everyday objects on real-world backgrounds, each of a different category from the ImageNet image database (i.e. animals, tools, vehicles, foods and others)[30].

fMRI experimental design

The study was composed of a main experimental part and a localizer run.

The main experimental part consisted of 6 to 14 runs, each lasting 621 s. Each run started and ended with a 10.5-s baseline period and included all 24 images, each repeated five times, for a total of 120 trials presented in random order. Each trial began with a 1-s stimulus-on interval, during which a centrally displayed object image (5° visual angle) appeared on a grey background with a pink fixation cross. This was followed by a 4-s stimulus-off interval, where only the fixation cross was shown. Participants were instructed to maintain their gaze on the fixation cross throughout the experiment and to perform a color-change detection task, pressing a button as soon as the fixation cross turned blue for 300 ms.

The functional localizer run was intended to define regions of interest (ROIs) EVC and LOC. Participants completed it at the beginning of the recording session. The localizer consisted of 15-s blocks displaying objects (not included in the main experiment) and scrambled objects overlayed with a fixation cross, interleaved with 7.5-s baseline blocks showing only a fixation cross on a grey background. In each block images were centrally presented (12° visual angle) for 400 ms, followed by a 350-ms display of the fixation cross. Participants were instructed to maintain their gaze on the fixation cross and to press a button if the same image appeared in consecutive trials. The localizer run included 12 blocks of each image type, resulting in a total duration of 465 s. Block order was pseudo-randomized to avoid immediate repetition of the same block type.

MRI Procedure

MRI data were acquired at the Max Planck Institute for Human Cognitive and Brain Sciences in Leipzig, Germany. Four participants completed the experiment in one scanning session. Six participants completed the experiment in two scanning sessions on two separate days. During the first scanning session, we acquired a T1-weighted anatomical image, the functional localizer and 5 to 8 runs of the main experiment. During the second scanning session, participants completed 6 to 8 runs of the main experiment. Additionally, to enable distortion correction, five volumes with reversed phase-encoding polarity were acquired following the first run of each main experimental session. To ensure that participants were familiar with the experimental tasks, we provided them with verbal and written instructions prior to the scanning, and the participants completed a 2-min training for both the localizer and the main tasks.

MRI acquisition parameters

We acquired MR images on a Siemens Magnetom Terra 7T whole-body system (Siemens Healthineers, Erlangen, Germany) with a single-channel-transmit and a 32-channel radio-frequency (RF) receive head coil (Nova Medical Inc, Wilmington, USA). We acquired the functional data using a 2D Gradient-echo (GE) echo planar imaging (EPI) sequence[92] (voxel size = 0.9 mm isotropic resolution, TE/TR = 26.2/3500 ms, in-plane field of view (FoV) 192 × 192 mm^2, 48 axial slices, flip angle = 75°, echo spacing = 1.0 ms, GRAPPA factor = 3, partial Fourier = 6/8, phase encoding direction anterior-posterior). We recorded anatomical data using an MP2RAGE sequence[93] (voxel size = 0.7 mm isotropic resolution, TE/TR = 2.01/5590 ms, in-plane FoV 224 × 224 mm, GRAPPA factor = 2) yielding two inversion contrasts (TI1 = 900 ms, flip angle 1 = 5°; TI2 = 2750 ms, flip angle 2 = 3°). The two inversion contrasts were combined to produce T1-weighted MP2RAGE uniform (UNI) images with high contrast to noise ratio.

MRI preprocessing

For each recording session, we spatially realigned the functional volumes to their mean volume using SPM12 (http://www.fil.ion.ucl.ac.uk/spm). To correct for geometric distortions in the phase encoding direction, we calculated a deformation field based on reverse gradient estimation, using the Advanced Normalization Tools (ANTs) software package (http://stnava.github.io/ANTs/). In detail, we combined the mean functional volume (forward image) with the mean volume acquired with opposing phase encoding direction to generate a distortion-corrected template. Next, we estimated the deformation map by registering the forward image to the corrected template reference using non-linear (SyN) transformations. Finally, the deformation map was used to produce a distortion-corrected mean functional volume.

To co-register the anatomical and the distortion-corrected mean functional volumes, we initially referred to the Glasser's atlas[94] and the Kanwisher's atlas[95] to identify the approximate location of EVC and LOC, respectively. Then we outlined a volume containing these regions in the occipito-temporal cortex of both hemispheres on the individual participant's native space (from here on referred to as manual mask) using ITK-SNAP with the 3D paintbrush tool (v.3.8)[96]. We then estimated a second deformation map in ANTs by registering the distortion-corrected mean volume to the T1-weighted volume, applying nonlinear (SyN) transformations within the manual mask. We visually inspected the fixed and registered volumes in ITK-SNAP for each participant. If the volume was not correctly registered within the region of the manual mask, we repeated the registration, adding linear transformations (rigid and affine) with stricter convergence criteria and increased iterations, until an accurate alignment was achieved. To minimize spatial resolution loss during resampling, we combined both deformation maps and resampled the functional images to the anatomical reference in a single interpolation step using a fifth-order spline function.

Finally, we spatially smoothed the functional localizer images using a 6-mm full width at half maximum (FWHM) Gaussian kernel. Functional images of the main runs were not smoothed to preserve spatial specificity.

Before segmenting the T1-weighted UNI volume into grey matter, white matter and cerebrospinal fluid following the procedure in ITK-SNAP outlined here (https://www.youtube.com/watch?v=tSA77mFTwcg&t=1042s), we corrected for bias field effects with a customized script[97]. Next, we divided the cortical ribbon into laminar and columnar compartments using LAYNII (v2.2.1)[98]. In detail, applying the equi-distant model[33], we segmented the gray matter into three cortical depths: deep, middle and superficial (**Fig. 3A**). Here, we used the term cortical layers to refer to the depth-dependent compartments along the cortical ribbon, distinct from the actual anatomical layers found in cortex. Additionally, we segmented the gray matter into columnar compartments within the manual masks.

fMRI univariate analysis

To estimate neural responses, we ran separate General Linear Model (GLM) analyses in SPM12 for each pre-processed functional run, i.e., all main experimental runs and the localizer run. All analyses were conducted in each participant's native anatomical space. Specifically, we modelled 25 regressors (i.e., 24 object images + baseline) for each main experimental run, and 3 regressors (i.e., objects, scrambled objects, and baseline) for the localizer run. We created the regressors by convolving a boxcar function representing the onsets and durations of the corresponding condition with the canonical (2 Gamma) hemodynamic response function (HRF). We incorporated the motion estimates into the model as nuisance regressors. By fitting a GLM, we obtained beta weight estimates for each condition (i.e., regressor) for each run, which were subsequently used in further analyses.

Definition of regions of interest

At the macroscale, we defined two regions of interest (ROIs) for each participant: EVC, comprising areas V1, V2, and V3, and LOC in a two-step procedure. First, we used anatomical masks from the above-mentioned brain atlases for EVC[94] and for LOC[95]. These masks were resampled from the MNI152 space into each participant's individual space. Second, we identified the overlap between the participant-specific anatomical masks and the corresponding functional contrast T-statistic map from the localizer experiment, retaining the top 2,000 voxels. Specifically, we ranked the voxels according to the objects + scrambled > baseline contrast T-statistic for EVC, and the objects > scrambled contrast T-statistic for LOC. Due to lack of activation in the lateral occipitotemporal cortex for one participant (3), we used the objects + scrambled > baseline contrast to define LOC. Any voxels overlapping across ROIs were excluded. This process resulted in one final EVC and LOC mask for each participant.

At the mesoscale, we defined six ROIs based on the two brain regions – EVC and LOC – and the three cortical layers – deep, middle and superficial (see above for definition of cortical layers). Here, the ROI definition was analogous, with the following deviations to address issues arising specifically at the mesoscale level. Voxels with higher spatial resolution exhibit a reduced SNR and increased susceptibility to noise sources[41], compromising the quality of the recorded signal. Additionally, signal reliability gradually decreases along the ventral visual cortex from lower to higher visual areas[99], likely due to signal loss and susceptibility-induced distortions[100]. To address this, we chose a higher number of voxels for EVC compared to LOC. This resulted in 3000 and 1500 voxels with the highest T-statistic within the cortical ribbon of EVC and LOC, respectively. We then assigned each voxel to one of the three layers and selected only those sharing full overlap of columnar compartments.

EEG data – paradigm, acquisition and analysis

We used a subset of the EEG data (*N* = 32) collected by Xie et al., 2024[5] for the same images as in the fMRI experiment. Below is a summary of the relevant experimental procedures and acquisition steps.

The EEG experiment employed and backward masking paradigm consisting of 2,544 trials. In each trial, an object image was briefly presented for 17 ms and followed by a dynamic mask under two conditions:

an early mask condition with inter-stimulus interval (ISI) of 17 ms or a late mask condition with an ISI of 600 ms. Object images and masks were randomly paired on each trial. All stimuli were centrally displayed on a gray background, with a size of 5° visual angle, and overlaid with a bull's-eye fixation symbol. Participants were instructed to maintain fixation and refrain from blinking during trials, except during designated task trials (i.e., two-alternative forced-choice task, 25% of total trials) where blinking was allowed after making a response. The inter-trial interval (ITI) ranged from 900 – 1,100 ms; following the task trials, the ITI was extended to 2,000 ms to reduce motor artifacts.

For the present analyses, we focused on data from the late mask condition (ISI = 600 ms), yielding approximately 53 trials per object image. We analyzed the time window from 200 ms pre-stimulus to 600 ms post-stimulus, before mask onset.

EEG data were recorded with a 64-electrode ActiCap system and a Brainvision actiChamp amplifier. Electrodes were positioned according to the international 10-10 system, with a ground electrode and a reference electrode placed on the scalp. The signals were sampled at 1,000 Hz and online filtered between 0.03 and 100 Hz. EEG data were preprocessed using Brainstorm-3[101]. Noisy channels (mean = 2.2, SD = 1.8) were removed, and the data were low-pass filtered at 40 Hz. Independent component analysis was applied to remove eye movement and other artifact components (mean = 2.7, SD = 0.9). Data were then segmented into epochs from -200 ms to 600 ms relative to stimulus onset, baseline-corrected, and multivariate noise normalization[102] was applied for decoding analyses.

Multivariate analysis was performed on a participant-specific basis using SVMs as implemented in the LIBSVM toolbox in MATLAB (2021a). To determine when the brain processes object information, time-resolved decoding analysis was conducted from -200 ms to 600 ms relative to target image onset in 10 ms intervals. At each time point, trial-specific EEG channel activations were extracted and arranged into 64-dimensional pattern vectors for each of the 24 object image conditions. For each condition, trials were randomly grouped into four equally sized bins and averaged to create four pseudo-trials, repeated over 100 permutations. Pseudo-trials were then divided into a training set (three pseudo-trials) and a testing set (one pseudo-trial) for pairwise object identity decoding. This process was repeated for all pairwise combinations of object conditions.

Multivariate pattern analysis of fMRI data

To decode object information from voxel activation patterns in EVC and LOC, we used SVMs as implemented in the scikit-learn library in Python. Analyses were conducted separately for each participant and for EVC and LOC at the macroscale. We aggregated run-specific voxel-wise beta estimates derived from the GLM into pattern vectors for each of the 24 conditions. To enhance the signal to noise ratio, we averaged beta weights across two randomly selected subsets of trials into two pseudo-trials for each condition. We repeated this process 300 times, each time performing pairwise object decoding, for all object condition combinations, using a two-fold cross-validation approach. We averaged the results across all iterations and all object condition combinations to yield one grand-average decoding accuracy for each ROI and participant.

Representational similarity analysis

To relate object representations across different signal spaces (i.e., voxel activation patterns in fMRI, sensor activation patterns in EEG, embeddings in DNNs), we used representational similarity analysis[21]. This approach is based on the rationale that if two images elicit similar neural representations, their corresponding fMRI, EEG or DNN signal patterns should also be similar. We used two variants of RSA: (i) representational similarity analysis-based fusion[21,23,24], relating spatially localized object representations (from fMRI) to specific temporal dynamics (from EEG) to resolve the spatiotemporal dynamics with which visual representations emerge, and ii) representational similarity analysis-based commonality analysis[103,104] to determine the visual feature complexity (from layer-specific DNN embeddings) of spatiotemporally identified dynamics. For this we calculated the common variance between spatially localized object representations (from fMRI), specific temporal dynamics (from EEG), and layer embeddings (from DNNs). We detail both approaches below after describing the specifics of summarizing representational similarity for each signal space in representational dissimilarity matrices (RDMs).

Construction of fMRI representational dissimilarity matrices

To construct the fMRI-RDMs, we first extracted and vectorized beta weight estimates for each of the 24 conditions to form fMRI neural patterns for a given ROI (area or cortical layer) and applied multivariate noise normalization in order to enhance the SNR. We then quantified the extent of pattern similarity by calculating the Pearson’s correlation for all pairwise combinations of experimental conditions. Output correlations were transformed into dissimilarity values using 1 – Pearson’s correlation, and organized into a 24×24 fMRI representational dissimilarity matrix (RDM), indexed in rows and columns by the compared conditions. For each participant, this yielded two fMRI-RDMs at the macroscale level of brain areas: the EVC RDM and LOC RDM; and six layer-specific fMRI-RDMs at the mesoscale level of cortical layers: Deep EVC RDM, Middle EVC RDM, Superficial EVC RDM, and Deep LOC RDM, Middle LOC RDM and Superficial LOC RDM. For all further analyses we averaged the MRI-RDMs across participants.

Construction of EEG representational dissimilarity matrices

A detailed description of the creation of the EEG-RDMs is provided by Xie et. al., (2024)[5]. Briefly, for each participant and each time point in the epoch from -200 to 600 ms, we pairwise decoded object information using EEG channel activation patterns, via SVMs as implemented in the LIBSVM toolbox[105] in MATLAB (2021a). The decoding accuracies for each pair of object images were assembled into 24×24 EEG-RDM for each time point (801 in total), with rows and columns indexed by the conditions. Each matrix was symmetric across the diagonal, with diagonal entries left undefined.

Construction of DNN representational dissimilarity matrices

To compute DNN-RDMs, we used the AlexNet architecture[31] trained on visual object categorization on the ImageNet dataset[106]. In detail, we fed the pre-trained network with the same set of 24 object images as input and extracted the activation pattern for each object condition from each of the five convolutional layers and the two subsequent fully-connected layers. Next, we quantified the pairwise pattern dissimilarity using 1 – Pearson's correlation for all pairs of image combinations and organized the outputs into 24×24 DNN-RDMs, with rows and columns representing Pearson's based dissimilarity measure between object conditions. This resulted in a total of 7 layer-specific DNN RDMs.

Representational similarity analysis-based EEG-fMRI fusion

We used representational similarity analysis-based fusion[21,23,24], relating fMRI-RDMs to EEG-RDMs using Spearman rank order correlation. We applied this analysis at the macroscale and at the mesoscale, for each participant separately. At the macroscale, we related region-specific RDMs (i.e., EEG, LOC) to EEG-RDMs, yielding one time course for each ROI for each participant. At the mesoscale we related region- and layers-specific ROIs to EEG-RDMs, yielding one time course for each of the six ROI-layer combinations for each participant.

Representational similarity analysis-based commonality analysis

To characterize the level of feature complexity of the neural representations identified above in space and time, we related them to DNN embeddings across model layers. Feature complexity increases progressively across DNN layers: early model layers are more sensitive to low complexity features, whereas later layers are tuned to high-complexity features[107,108], analogous to the ventral visual hierarchy in humans and non-human primates[30,32]. Thus, by linking the neural representations to those from the DNN layers, we can determine the feature complexity of spatially localized object representations to particular time points. We did this using representational similarity analysis-based commonality analysis[109]. In detail, we calculated the commonality coefficient corresponding to the shared variance among fMRI-RDMs for each brain region and cortical layer, EEG-RDMs at time points with significant effects in EEG-fMRI fusion (averaged across time), and DNN RDMs for each model layer.

Quantification and statistical analysis

To assess statistical significance, we performed non-parametric statistical analyses that do not rely on assumptions about the data distribution. The empirical estimations, including decoding accuracy as well as correlation values from the representational similarity analysis-based fusion and from the commonality analysis, were tested against a null distribution created by sign permutation (1,000 permutations). We randomly multiplied the participant-specific data by ±1 to generate permutation samples, recomputed the statistic for each sample, and derived *P* values.

We controlled the familywise error rate across time points using cluster-based statistics. First, *P*-value maps were thresholded at $P < 0.05$ to define temporally contiguous suprathreshold clusters. These clusters were then used to construct an empirical null distribution of maximum cluster weights, and a corrected threshold was determined at the 95th percentile of the right tail of this distribution. For the commonality analyses we corrected *P* values for multiple comparisons by FDR-correction.

We estimated the 95% confidence intervals for the peak latencies in the RSA-derived time courses using bootstrapping (1,000 paired bootstrap resamples with replacement). For each bootstrap we calculated the statistic, resulting in a bootstrap estimate of peak latencies from which we derived the confidence intervals.

To estimate confidence intervals for peak latency differences, we used an analogous bootstrapping approach, resampling the mean peak-to-peak latency difference for each resample. This generated a distribution of mean differences, from which we derived the 95% confidence interval. We set $P < 0.05$, i.e., if the 95% confidence interval did not include 0, we rejected the null hypothesis of no peak-to-peak latency differences.

References

1. Mishkin, M., Ungerleider, L. G. & Macko, K. A. Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences* **6**, 414–417 (1983).

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

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

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

5. Xie, S., Singer, J., Yilmaz, B., Kaiser, D. & Cichy, R. M. The representational nature of spatio-temporal recurrent processing in visual object recognition. 2024.07.30.605751 Preprint at https://doi.org/10.1101/2024.07.30.605751 (2024).

6. Lamme, V. A. F., Zipser, K. & Spekreijse, H. Figure-ground activity in primary visual cortex is suppressed by anesthesia. *Proceedings of the National Academy of Sciences* **95**, 3263–3268 (1998).

7. Markov, N. T. *et al.* Anatomy of hierarchy: Feedforward and feedback pathways in macaque visual cortex. *Journal of Comparative Neurology* **522**, 225–259 (2014).

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

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

10. Hüer, J., Saxena, P. & Treue, S. Pathway-selective optogenetics reveals the functional anatomy of top–down attentional modulation in the macaque visual cortex. *Proc. Natl. Acad. Sci. U.S.A.* **121**, e2304511121 (2024).

11. Nurminen, L., Merlin, S., Bijanzadeh, M., Federer, F. & Angelucci, A. Top-down feedback controls spatial summation and response amplitude in primate visual cortex. *Nature Communications* **9**, 2281 (2018).

12. Bergmann, J. *et al.* Cortical depth profiles in primary visual cortex for illusory and imaginary experiences. *Nature Communications* **15**, 1002 (2024).

13. Lawrence, S. J., Norris, D. G. & de Lange, F. P. Dissociable laminar profiles of concurrent bottom-up and top-down modulation in the human visual cortex. *eLife* **8**, e44422 (2019).

14. Carricarte, T. *et al.* Laminar dissociation of feedforward and feedback in high-level ventral visual cortex during imagery and perception. *iScience* **27**, 110229 (2024).

15. Dowdle, L. *et al.* Characterizing top-down microcircuitry of complex human behavior across different levels of the visual hierarchy. Preprint at https://doi.org/10.1101/2022.12.03.518973 (2023).

16. Liu, C. *et al.* Layer-dependent multiplicative effects of spatial attention on contrast responses in human early visual cortex. *Progress in Neurobiology* **207**, 101897 (2021).

17. Rockland, K. S. & Pandya, D. N. Laminar origins and terminations of cortical connections of the occipital lobe in the rhesus monkey. *Brain Research* **179**, 3–20 (1979).

18. Rockland, K. S. What do we know about laminar connectivity? *NeuroImage* **197**, 772–784 (2019).

19. Siu, C., Balsor, J., Merlin, S., Federer, F. & Angelucci, A. A direct interareal feedback-to-feedforward circuit in primate visual cortex. *Nat Commun* **12**, 4911 (2021).

20. Federer, F., Ta'afua, S., Merlin, S., Hassanpour, M. S. & Angelucci, A. Stream-specific feedback inputs to the primate primary visual cortex. *Nature Communications* **12**, 228 (2021).

21. Kriegeskorte, N., Mur, M. & Bandettini, P. A. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience* **2**, (2008).

22. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nature Neuroscience* **17**, 455–462 (2014).

23. Cichy, R. M. & Oliva, A. A M/EEG-fMRI Fusion Primer: Resolving Human Brain Responses in Space and Time. *Neuron* **107**, 772–781 (2020).

24. 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. *Cerebral Cortex* **26**, 3563–3579 (2016).

25. Essen, D. C. V. & Maunsell, J. H. R. Hierarchical organization and functional streams in the visual cortex. *Trends in Neurosciences* **6**, 370–375 (1983).

26. Singer, J. J. D., Karapetian, A., Hebart, M. N. & Cichy, R. M. Identifying and characterizing scene representations relevant for categorization behavior. *Imaging Neuroscience* **3**, imag_a_00449 (2025).

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

28. Kourtzi, Z. & Connor, C. E. Neural Representations for Object Perception: Structure, Category, and Adaptive Coding. *Annual Review of Neuroscience* vol. 34 45–67 (2011).

29. Muukkonen, I., Ölander, K., Numminen, J. & Salmela, V. R. Spatio-temporal dynamics of face perception. *NeuroImage* **209**, 116531 (2020).

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

31. Krizhevsky, A., Sutskever, I. & Hinton, G. E. ImageNet Classification with Deep Convolutional Neural Networks. in *Advances in Neural Information Processing Systems* (eds. Pereira, F., Burges, C. J., Bottou, L. & Weinberger, K. Q.) vol. 25 (Curran Associates, Inc., 2012).

32. Güçlü, U. & van Gerven, M. A. J. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. *J. Neurosci.* **35**, 10005 (2015).

33. Waehnert, M. D. *et al.* Anatomically motivated modeling of cortical laminae. *NeuroImage* **93**, 210–220 (2014).

34. Barth, M. & Norris, D. G. Very high-resolution three-dimensional functional MRI of the human visual cortex with elimination of large venous vessels. *NMR in Biomedicine* **20**, 477–484 (2007).

35. Markuerkiaga, I., Barth, M. & Norris, D. G. A cortical vascular model for examining the specificity of the laminar BOLD signal. *NeuroImage* **132**, 491–498 (2016).

36. 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. *Nature Communications* **8**, 13804 (2017).

37. Self, M. W., van Kerkoerle, T., Supèr, H. & Roelfsema, P. R. Distinct Roles of the Cortical Layers of Area V1 in Figure-Ground Segregation. *Current Biology* **23**, 2121–2129 (2013).

38. Barzegaran, E. & Plomp, G. Four concurrent feedforward and feedback networks with different roles in the visual cortical hierarchy. *PLOS Biology* **20**, e3001534 (2022).

39. Grossberg, S. Linking the laminar circuits of visual cortex to visual perception: Development, grouping, and attention. *Neuroscience & Biobehavioral Reviews* **25**, 513–526 (2001).

40. Smith, M. A., Jia, X., Zandvakili, A. & Kohn, A. Laminar dependence of neuronal correlations in visual cortex. *Journal of Neurophysiology* **109**, 940–947 (2013).

41. Polimeni, J. R., Fischl, B., Greve, D. N. & Wald, L. L. Laminar analysis of 7T BOLD using an imposed spatial activation pattern in human V1. *NeuroImage* **52**, 1334–1346 (2010).

42. Schwiedrzik, C. M. & Freiwald, W. A. High-Level Prediction Signals in a Low-Level Area of the Macaque Face-Processing Hierarchy. *Neuron* **96**, 89-97.e4 (2017).

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

44. Morgan, A. T., Petro, L. S. & Muckli, L. Scene Representations Conveyed by Cortical Feedback to Early Visual Cortex Can Be Described by Line Drawings. *J. Neurosci.* **39**, 9410 (2019).

45. Papale, P. *et al.* The representation of occluded image regions in area V1 of monkeys and humans. *Current Biology* **33**, 3865-3871.e3 (2023).

46. Del Rosario, J. *et al.* Lateral inhibition in V1 controls neural and perceptual contrast sensitivity. *Nature Neuroscience* (2025) doi:10.1038/s41593-025-01888-4.

47. Gilbert, C. D. & Wiesel, T. N. Morphology and intracortical projections of functionally characterised neurones in the cat visual cortex. *Nature* **280**, 120–125 (1979).

48. Rockland, K. S. & Lund, J. S. Widespread Periodic Intrinsic Connections in the Tree Shrew Visual Cortex. *Science* **215**, 1532–1534 (1982).

49. Gilbert, C. & Wiesel, T. Clustered intrinsic connections in cat visual cortex. *J. Neurosci.* **3**, 1116 (1983).

50. Stettler, D. D., Das, A., Bennett, J. & Gilbert, C. D. Lateral Connectivity and Contextual Interactions in Macaque Primary Visual Cortex. *Neuron* **36**, 739–750 (2002).

51. Fujita, I. & Fujita, T. Intrinsic connections in the macaque inferior temporal cortex. *Journal of Comparative Neurology* **368**, 467–486 (1996).

52. Grootswagers, T., Robinson, A. K., Shatek, S. M. & Carlson, T. A. Mapping the dynamics of visual feature coding: Insights into perception and integration. *PLOS Computational Biology* **20**, e1011760 (2024).

53. Gallimore, C. G., Ricci, D. A. & Hamm, J. P. Spatiotemporal dynamics across visual cortical laminae support a predictive coding framework for interpreting mismatch responses. *Cerebral Cortex* **33**, 9417–9428 (2023).

54. Dobs, K., Isik, L., Pantazis, D. & Kanwisher, N. How face perception unfolds over time. *Nature Communications* **10**, 1258 (2019).

55. Grill-Spector, K. & Malach, R. The human visual cortex. *Annual Review of Neuroscience* vol. 27 649–677 (2004).

56. Grill-Spector, K. *et al.* A sequence of object-processing stages revealed by fMRI in the human occipital lobe. *Human Brain Mapping* **6**, 316–328 (1998).

57. Callaway, E. M. Feedforward, feedback and inhibitory connections in primate visual cortex. *Neural Networks* **17**, 625–632 (2004).

58. 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).

59. Scheeringa, R. & Fries, P. Cortical layers, rhythms and BOLD signals. *NeuroImage* **197**, 689–698 (2019).

60. Larkum, M. E., Zhu, J. J. & Sakmann, B. A new cellular mechanism for coupling inputs arriving at different cortical layers. *Nature* **398**, 338–341 (1999).

61. Aru, J., Suzuki, M. & Larkum, M. E. Cellular Mechanisms of Conscious Processing. *Trends in Cognitive Sciences* **24**, 814–825 (2020).

62. Schuman, B., Dellal, S., Prönneke, A., Machold, R. & Rudy, B. Neocortical Layer 1: An Elegant Solution to Top-Down and Bottom-Up Integration. *Annual Review of Neuroscience* **44**, 221–252 (2021).

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

64. Xie, S., Kaiser, D. & Cichy, R. M. Visual Imagery and Perception Share Neural Representations in the Alpha Frequency Band. *Current Biology* **30**, 2621-2627.e5 (2020).

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

66. van Kerkoerle, T. *et al.* Alpha and gamma oscillations characterize feedback and feedforward processing in monkey visual cortex. *Proceedings of the National Academy of Sciences* **111**, 14332–14341 (2014).

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

68. Muckli, L. *et al.* Contextual Feedback to Superficial Layers of V1. *Current Biology* **25**, 2690–2695 (2015).

69. Haynes, J.-D. & Rees, G. Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience* **7**, 523–534 (2006).

70. Kriegeskorte, N. & Kievit, R. A. Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences* **17**, 401–412 (2013).

71. Mur, M., Bandettini, P. A. & Kriegeskorte, N. Revealing representational content with pattern-information fMRI—an introductory guide. *Social Cognitive and Affective Neuroscience* **4**, 101–109 (2009).

72. Henriksson, L., Mur, M. & Kriegeskorte, N. Rapid Invariant Encoding of Scene Layout in Human OPA. *Neuron* **103**, 161-171.e3 (2019).

73. Mohsenzadeh, Y., Qin, S., Cichy, R. M. & Pantazis, D. Ultra-Rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway. *eLife* **7**, e36329 (2018).

74. Motlagh, S. C., Joanisse, M., Wang, B. & Mohsenzadeh, Y. Unveiling the neural dynamics of conscious perception in rapid object recognition. *NeuroImage* **296**, 120668 (2024).

75. Hung, C. P., Kreiman, G., Poggio, T. & DiCarlo, J. J. Fast readout of object identity from macaque inferior temporal cortex. *Science* **310**, 863–866 (2005).

76. Hopf, J.-M. *et al.* Neural Sources of Focused Attention in Visual Search. *Cerebral Cortex* **10**, 1233–1241 (2000).

77. Cul, A. D., Baillet, S. & Dehaene, S. Brain Dynamics Underlying the Nonlinear Threshold for Access to Consciousness. *PLOS Biology* **5**, e260 (2007).

78. Duan, Y., Zhan, J., Gross, J., Ince, R. A. A. & Schyns, P. G. Pre-frontal cortex guides dimension-reducing transformations in the occipito-ventral pathway for categorization behaviors. *Current Biology* **34**, 3392-3404.e5 (2024).

79. Bar, M. A Cortical Mechanism for Triggering Top-Down Facilitation in Visual Object Recognition. *Journal of Cognitive Neuroscience* **15**, 600–609 (2003).

80. Kietzmann, T. C. *et al.* Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences* **116**, 21854–21863 (2019).

81. 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. *Nature Neuroscience* **22**, 974–983 (2019).

82. von Seth, J., Nicholls, V. I., Tyler, L. K. & Clarke, A. Recurrent connectivity supports higher-level visual and semantic object representations in the brain. *Commun Biol* **6**, 1–15 (2023).

83. Hamm, J. P., Shymkiv, Y., Han, S., Yang, W. & Yuste, R. Cortical ensembles selective for context. *Proceedings of the National Academy of Sciences* **118**, e2026179118 (2021).

84. Kar, K. & DiCarlo, J. J. Fast Recurrent Processing via Ventrolateral Prefrontal Cortex Is Needed by the Primate Ventral Stream for Robust Core Visual Object Recognition. *Neuron* **109**, 164-176.e5 (2021).

85. Hallett, M. Transcranial Magnetic Stimulation: A Primer. *Neuron* **55**, 187–199 (2007).

86. Bergmann, T. O. *et al.* Concurrent TMS-fMRI for causal network perturbation and proof of target engagement. *NeuroImage* **237**, 118093 (2021).

87. Markuerkiaga, I., Marques, J. P., Gallagher, T. E. & Norris, D. G. Estimation of laminar BOLD activation profiles using deconvolution with a physiological point spread function. *J Neurosci Methods* **353**, 109095 (2021).

88. Huber, L. *et al.* High-Resolution CBV-fMRI Allows Mapping of Laminar Activity and Connectivity of Cortical Input and Output in Human M1. *Neuron* **96**, 1253-1263.e7 (2017).

89. Huber, L. *et al.* Techniques for blood volume fMRI with VASO: From low-resolution mapping towards sub-millimeter layer-dependent applications. *NeuroImage* **164**, 131–143 (2018).

90. Pizzuti, A. *et al.* Imaging the columnar functional organization of human area MT+ to axis-of-motion stimuli using VASO at 7 Tesla. *Cerebral Cortex* **33**, 8693–8711 (2023).

91. Kok, P., Bains, L. J., van Mourik, T., Norris, D. G. & de Lange, F. P. Selective Activation of the Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback. *Current Biology* **26**, 371–376 (2016).

92. Moeller, S. *et al.* Multiband multislice GE-EPI at 7 tesla, with 16-fold acceleration using partial parallel imaging with application to high spatial and temporal whole-brain fMRI. *Magnetic Resonance in Medicine* **63**, 1144–1153 (2010).

93. Marques, J. P. *et al.* MP2RAGE, a self bias-field corrected sequence for improved segmentation and T1-mapping at high field. *NeuroImage* **49**, 1271–1281 (2010).

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

95. 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).

96. Yushkevich, P. A. *et al.* User-Guided Segmentation of Multi-modality Medical Imaging Datasets with ITK-SNAP. *Neuroinformatics* **17**, 83–102 (2019).

97. Lüsebrink, F., Sciarra, A., Mattern, H., Yakupov, R. & Speck, O. T1-weighted in vivo human whole brain MRI dataset with an ultrahigh isotropic resolution of 250 μm. *Scientific Data* **4**, 170032 (2017).

98. Huber, L. (Renzo) *et al.* LayNii: A software suite for layer-fMRI. *NeuroImage* **237**, 118091 (2021).

99. Badwal, M. W., Bergmann, J., Roth, J., Doeller, C. F. & Hebart, M. N. The scope and limits of fine-grained image and category information in the ventral visual pathway. *J. Neurosci.* **45**, e0936242024 (2024).

100. Malekian, V. *et al.* Mitigating susceptibility-induced distortions in high-resolution 3DEPI fMRI at 7T. *NeuroImage* **279**, 120294 (2023).

101. Tadel, F., Baillet, S., Mosher, J. C., Pantazis, D. & Leahy, R. M. Brainstorm: A User-Friendly Application for MEG/EEG Analysis. *Computational Intelligence and Neuroscience* **2011**, 879716 (2011).

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

103. Mood, A. M. Partitioning Variance in Multiple Regression Analyses as a Tool For Developing Learning Models. *American Educational Research Journal* **8**, 191–202 (1971).

104. Reichwein Zientek, L. & Thompson, B. Commonality Analysis: Partitioning Variance to Facilitate Better Understanding of Data. *Journal of Early Intervention* **28**, 299–307 (2006).

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

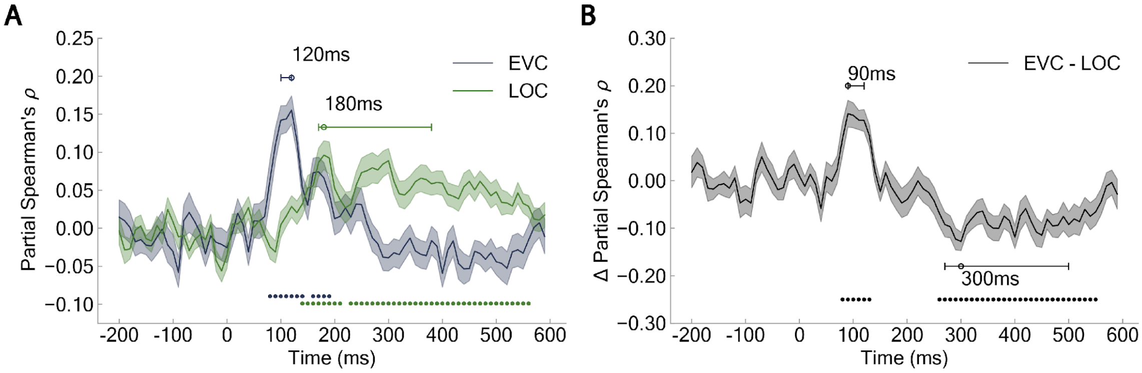
106. J. Deng *et al.* ImageNet: A large-scale hierarchical image database. in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009). doi:10.1109/CVPR.2009.5206848.

107. Zeiler, M. D. & Fergus, R. Visualizing and Understanding Convolutional Networks. Preprint at https://doi.org/10.48550/arXiv.1311.2901 (2013).

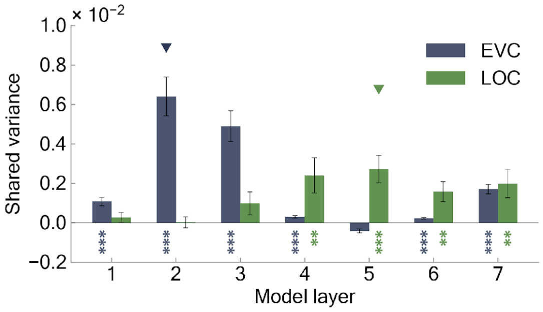
108. Zhou, B., Khosla, A., Lapedriza, A., Oliva, A. & Torralba, A. Object Detectors Emerge in Deep Scene CNNs. Preprint at https://doi.org/10.48550/arXiv.1412.6856 (2015).

109. Hebart, M. N., Bankson, B. B., Harel, A., Baker, C. I. & Cichy, R. M. The representational dynamics of task and object processing in humans. *Elife* **7**, e32816 (2018).

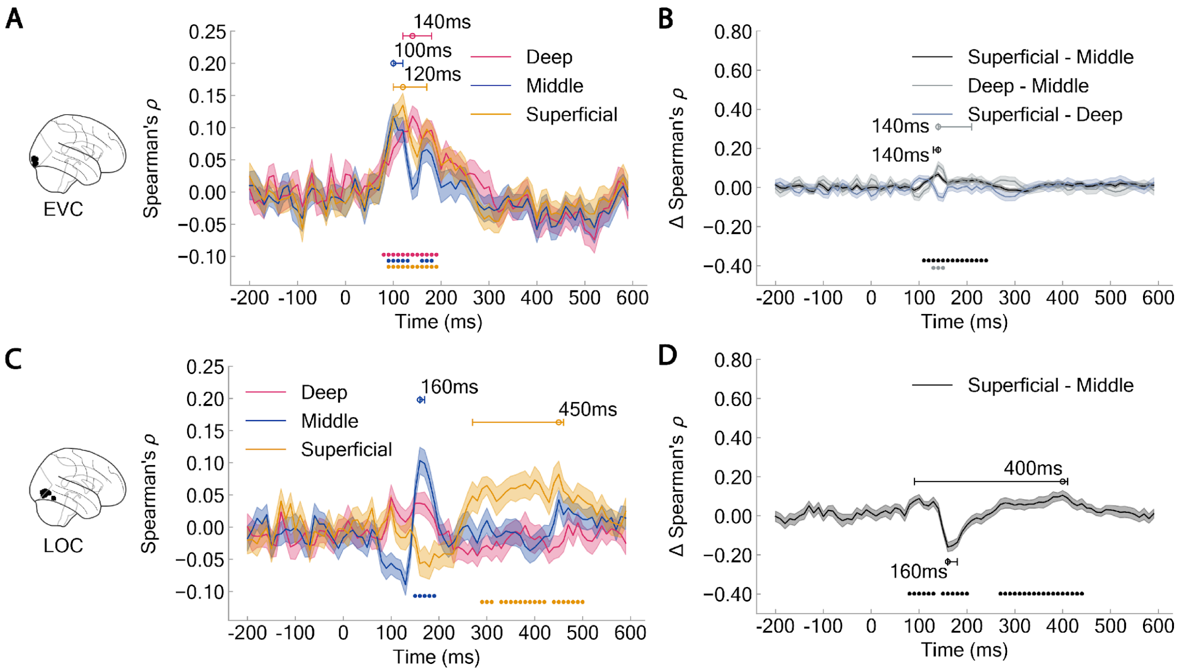
Supplementary information



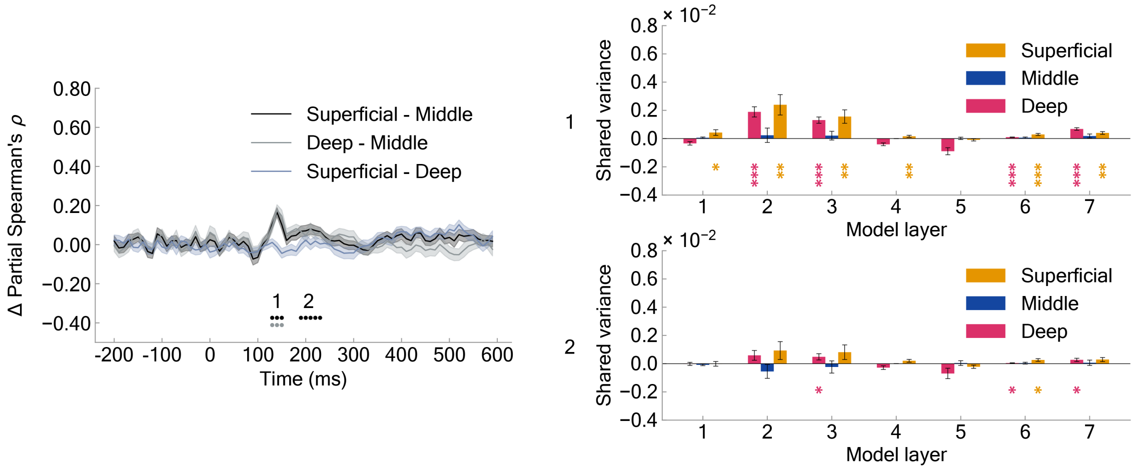
Supplementary Figure 1. Representational EEG-fMRI fusion at the macroscale using partial correlations. For EVC we partialed out the effect of LOC and for LOC we partialed out the effect of EVC. Early representations correlated stronger with EVC than with LOC and later representations stronger with LOC than with EVC. Shaded area indicates the standard error of the mean across participants; colored circles indicate significant time points ($N = 32$, cluster-defining threshold $P < 0.05$, cluster threshold $P < 0.05$); uncolored circles and horizontal lines indicate peak latency means and 95% confidence intervals, respectively.



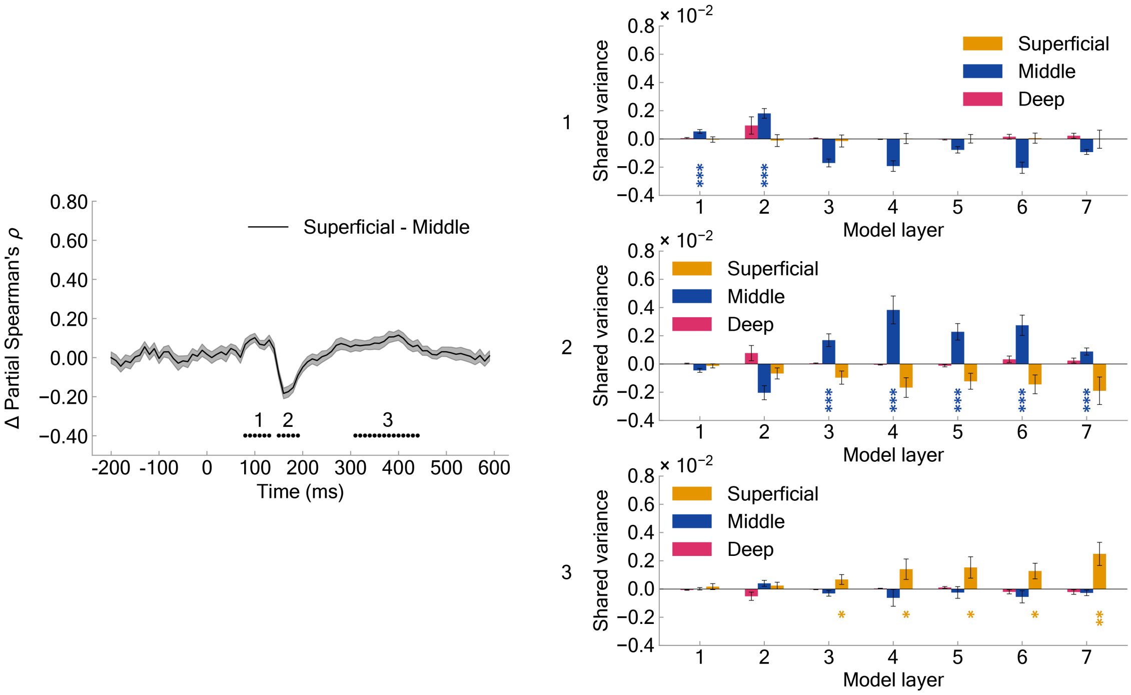
Supplementary Figure 2. Commonality analysis at the macroscale based on significant time intervals of curve EVC – LOC differences. Visual representations of low-complexity emerge primarily in EVC, while mid-to-high-level object representations emerge in LOC. Colored asterisks indicate significant correlations (*N* = 32, right-tailed permutation tests, FDR-corrected; **P* < 0.05; ***P* < 0.01; ****P* < 0.001); colored triangles represent model layers with the highest occurrence proportion, determined through 1,000-iteration bootstraps.



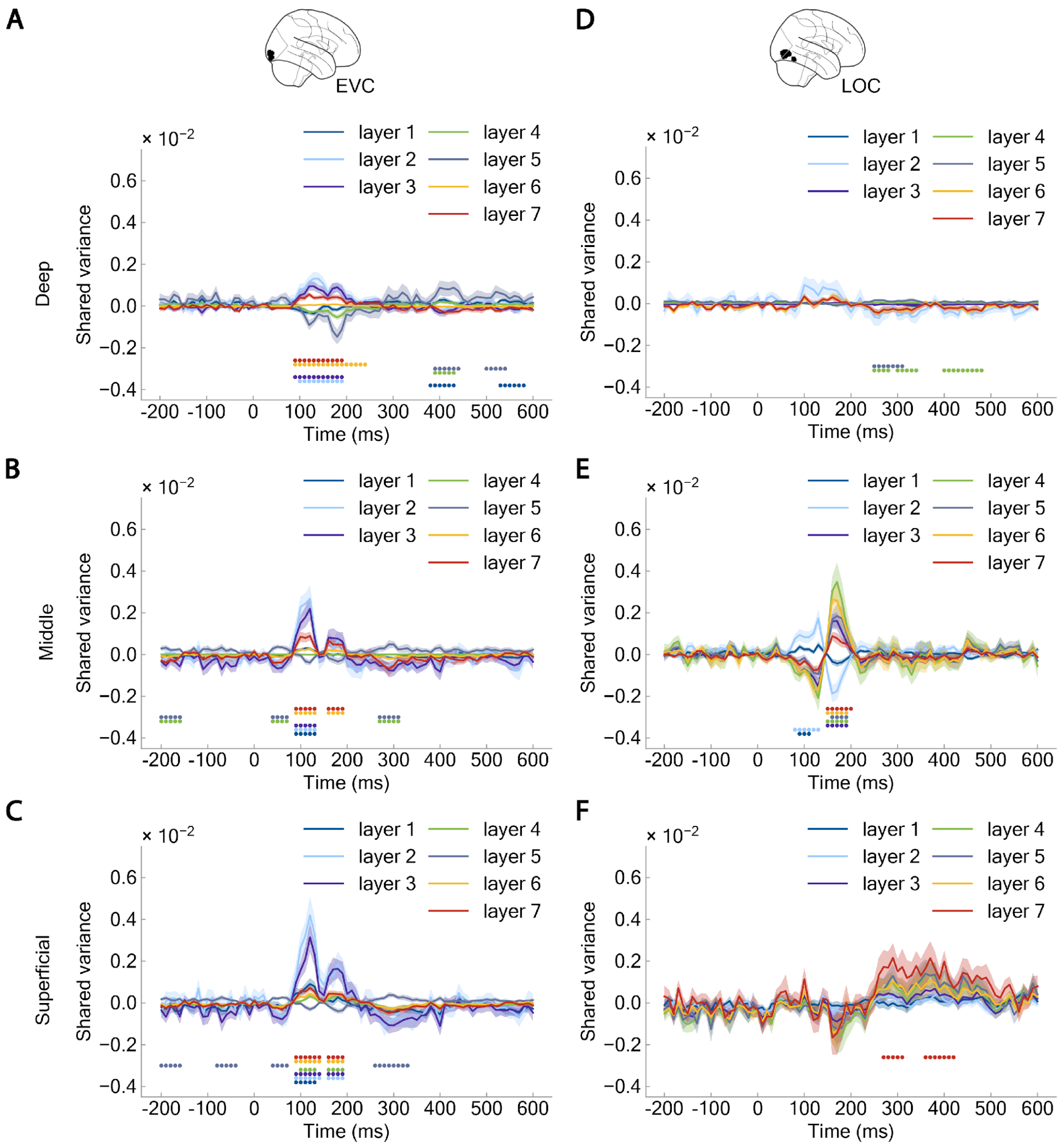
Supplementary Figure 3. Representational EEG-fMRI fusion at the mesoscale without partial correlations. Object representations derived from GE-BOLD signals are strongly influenced by non-specific macrovascular responses, which can compromise layer-specificity. To address this, we applied partial rank-order Spearman correlation, reducing the laminar influence of the layers beneath. Specifically, for the superficial layer, we partialed out the effect of the middle layer. Similarly, for the middle layer, we partialed out the effect of the deep layer. The deep layer remained unaffected by this approach. (**A**) Early representations emerged across layers in EVC, while late representations appeared in middle and superficial layers. (**B**) Early representations emerged in middle layers of LOC, whereas late representations appeared in superficial layers. Shaded area indicates the standard error of the mean across participants; colored circles indicate significant time points ($N = 32$, cluster-defining threshold $P < 0.05$, cluster threshold $P < 0.05$); uncolored circles and horizontal lines indicate peak latency means and 95% confidence intervals, respectively.



Supplementary Figure 4. Commonality analysis in EVC at the mesoscale based on significant time intervals of curve differences across cortical layers. We defined two clusters (labeled 1 and 2) from the significant time intervals. For each cluster, we performed commonality analysis by linking each model layer-specific DNN-RDM to each layer-specific fMRI-RDM in EVC and LOC and the mean EEG-RDM at the time interval with significant layer-specific temporal dynamics. Error bars indicate the standard error of the mean across participants; colored asterisks indicate significant correlations (*N* = 32, right-tailed permutation tests, FDR-corrected; **P* < 0.05; ***P* < 0.01; ****P* < 0.001).



Supplementary Figure 5. Commonality analysis in LOC at the mesoscale based on significant time intervals of curve differences across cortical layers. We defined three clusters (labeled 1, 2 and 3) from the significant time intervals. For each cluster, we performed commonality analysis by linking each model layer-specific DNN-RDM to each layer-specific fMRI-RDM in EVC and LOC and the mean EEG-RDM at the time interval with significant layer-specific temporal dynamics. Error bars indicate the standard error of the mean across participants; colored asterisks indicate significant correlations (*N* = 32, right-tailed permutation tests, FDR-corrected; *$P < 0.05$; **$P < 0.01$; ***$P < 0.001$).



Supplementary Figure 6. Commonality analysis at the mesoscale based on all individual time points. (**A, D**) Deep layer, (**B, E**) Middle layer and (**C, F**) Superficial layer in (**A-C**) EVC and (**D-F**) LOC. Shaded area indicates the standard error of the mean across participants; colored circles indicate significant time points ($N = 32$, cluster-defining threshold $P < 0.05$, cluster threshold $P < 0.05$).