Coordinated transformation of object representations across human visual cortex

Rui Xu[1]*†, Xin Xu[3], Zhipei Ling[3], Wenjing Zhou[2], Hesheng Liu[4,6], Bo Hong[1,5]*

[1]Department of Biomedical Engineering, School of Medicine, Tsinghua University, Beijing, China
[2]Department of Epilepsy Center, Tsinghua University Yuquan Hospital, Beijing, China
[3]Department of Neurosurgery, Chinese PLA General Hospital, Beijing, China
[4]Changping Laboratory, Beijing, China
[5]McGovern Institute for Brain Research, Tsinghua University, Beijing, China
[6]Biomedical Pioneering Innovation Center, Peking University, Beijing, China
*Corresponding author. Email: hongbo@tsinghua.edu.cn (B.H.); ruix@mit.edu (R.X.)
†Present address: McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA

ABSTRACT

Our brain builds increasingly sophisticated representations along the ventral visual pathway to support object recognition, but how these representations unfold over time is poorly understood. Here we characterized time-varying representations of faces, places, and objects throughout the pathway using human intracranial electroencephalography. For ~100 ms after an initial feedforward sweep, representations at all stages evolved to be less driven by low-order features and more categorical. Low-level areas like V1 showed unexpected, late-emerging tolerance to image size, and late but not early responses of high-level occipitotemporal areas best matched their fMRI responses. Besides aligned, simultaneous representational changes, we found a trial-by-trial association between concurrent response patterns across stages. Our results suggest fast, multi-areal recurrent processing builds upon initial feedforward processing to generate more sophisticated object representations.

MAIN TEXT

Humans can recognize objects, i.e., extract abstract information (e.g., category and identity) from their visual forms within a fraction of a second[1,2]. The ventral visual pathway[3] in the human brain supports object recognition, which extends from area V1 in the occipital pole rostrally to the occipitotemporal cortex (OTC)[4,5]. Areas along the pathway form a spatial hierarchy of function, in which the responses represent features of increasingly higher order[6,7], from bar orientation in V1 to category and identity in OTC. It remains elusive when these representations emerge in each area and whether and how they evolve[8], which would offer important insights into the underlying mechanisms. Standard tools in human neuroscience, such as fMRI[9–11] and MEG/EEG[12–16], track either the spatial pattern or the time course of object representations but not both. Intracranial EEG (iEEG)[17–19] and neuronal spiking[20,21] signals have a superior spatiotemporal resolution. Still, they are invasive and only available under neurosurgical settings[22], which have yet to provide a comprehensive account of the ventral pathway. Here we collected iEEG and fMRI data covering the entire pathway from V1 through OTC and found processing order in all areas substantially increased over time, for ~100 ms after the initial object-selective responses in OTC.

Results

Spatiotemporally resolved iEEG signal

We recorded cortical activity from 21 subjects implanted with iEEG electrodes to localize epileptic seizure foci and extracted log-normalized power spectrum density of high gamma activity[23] at 10-ms resolution (see Methods), which correlates with neuronal spiking[24,25]. Subjects performed a one-back task viewing serially presented images of clear, isolated objects for 300 ms every 900 to 1050 ms. The stimuli were 120 *instances* of objects from three *categories*: faces, places, and other objects (Fig. 1A), which correspond to some of the most studied category-selective regions in human OTC[10]. We divided each category into two subcategories or classes to help separate inter-instance dissimilarity structures based on low-order features vs. object categories. We

ensured that same-class instances shared some visual features, e.g., orientation and silhouette (Fig. 1A).

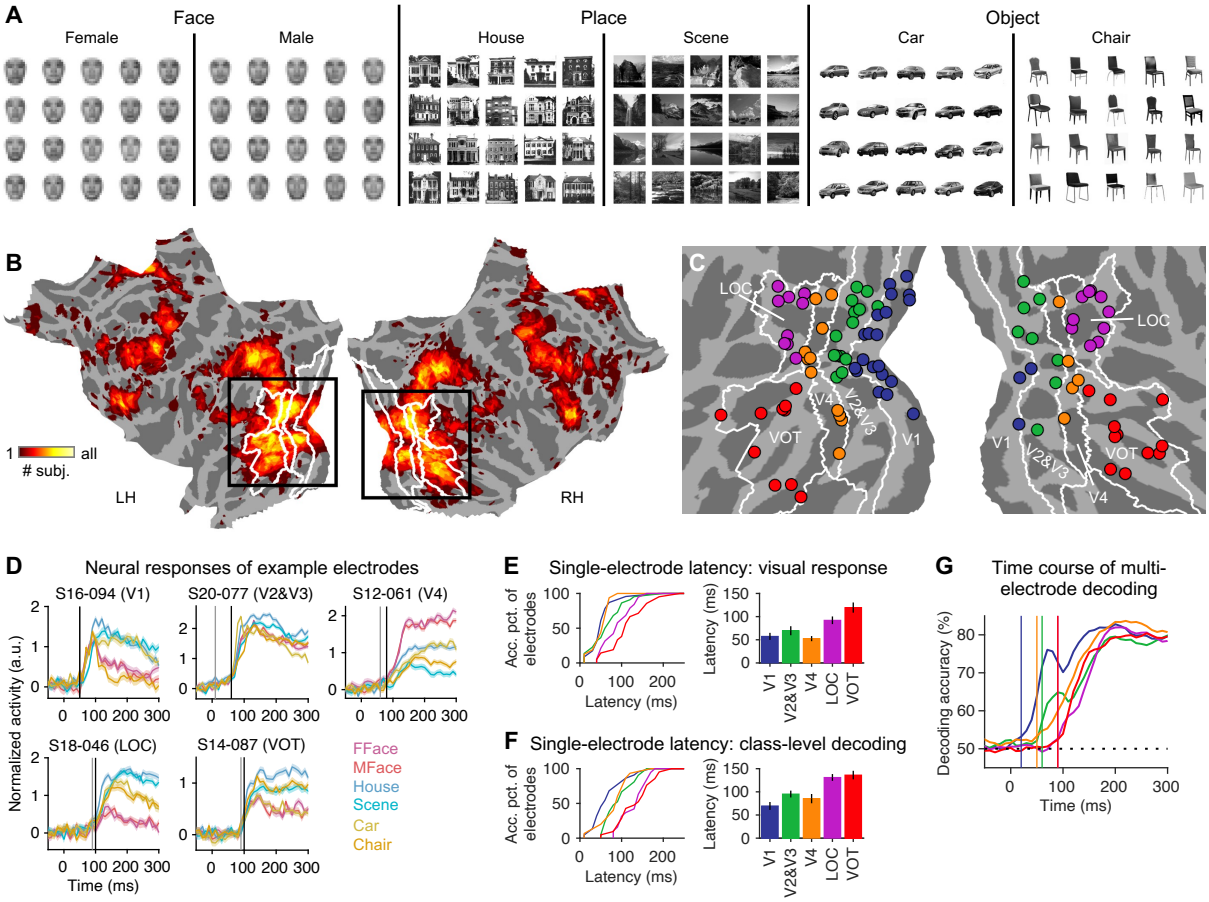

Fig. 1. Stimuli and iEEG recordings. (**A**) The stimuli are 120 images of object instances belonging to 6 classes and 3 categories. We pixelated the face images for privacy. (**B**) The number of subjects in which a vertex showed significantly positive fMRI responses to the stimuli ($P < 10^{-4}$), shown as an overlay on flattened surfaces of fsaverage. White contours are boundaries of ventral visual areas. Black rectangles match the field of view in (C). (**C**) Electrodes of interests on fsaverage colored by area. White contours are the same as in (B). (**D**) Neural responses by class at -50 to 300 ms relative to stimulus onset in example electrodes. Texts on top indicate subject ID, electrode ID, and area. Solid lines and shadings denote the mean and standard error of mean over same-class trials. Gray and black vertical lines indicate the latency of visual responses and class-level decoding, respectively. (**E**) Single-electrode latency of visual responses. Left, cumulative distribution by area (indicated by color). Y-axis indicates the accumulative percentage of electrodes. Right, areal mean. Error bar indicates standard error of mean. Hierarchically lower areas generally had significantly shorter latency (paired T-test, $P < 0.01$ for V1 vs. LOC, V2&V3 vs. VOT; $P < 0.001$ for V4 vs. LOC; $P < 10^{-4}$ for V1/V4 vs. VOT). (**F**) As (E), for class-level decoding. $P < 0.05$ for V1 vs. V2; $P < 0.01$ for V2&V3 vs. LOC/VOT; $P < 0.001$ for V4 vs. LOC/VOT; $P < 10^{-5}$ for V1 vs. LOC/VOT. (**G**) Time course of multi-electrode decoding by area (indicated by color). Vertical lines indicate latency. The dashed line indicates the chance level.

We searched for neural responses to objects at 50 to 300 ms that were larger than fixation, as observed in spiking data, and supported above-chance class-level decoding (see Methods), possibly using low-order and/or category information. Of 1227 electrodes covering much of the cerebral cortex, 122 showed such responses. Among the latter, 101 were in ventral visual areas

and co-localized with functional MRI (fMRI) responses to the same stimuli collected pre-implant from 13 subjects (Fig. 1B and figs. S1A-S1B), and we focused on these electrodes in further analyses. They covered consecutive stages of the ventral pathway (Fig. 1C and fig. S1C) according to a population atlas[26], including V1 (n = 23), V2&V3 (n = 22), V4 (n = 16), and two parts of OTC[4,5,27], lateral occipital cortex (LOC; n = 19) and ventral occipitotemporal cortex (VOT; n = 21).

To validate the spatiotemporal resolvability of iEEG, we examined the initial responses of the electrodes of interest (see Fig. 1D for examples). We quantified single-electrode latency of visual responses and class-level decoding (Figs. 1E-1F, respectively; see Methods), which were significantly longer in LOC and VOT than in lower areas, consistent with a feedforward sweep ascending the ventral pathway. We found the same trend in a separate multi-electrode decoding analysis (Fig. 1G). Thus, OTC areas began to differentiate objects at ~100 ms, as previously reported[17], whereas faster decoding in lower areas may reflect systematic variation in low-order features present in the stimuli. The neural codes of these areas showed similar fidelity, as decoding accuracy was comparable across areas (Fig. 1G and fig. S2). The clear inter-areal delay supports using iEEG to study the spatiotemporal dynamics of object processing.

A spatiotemporal visual hierarchy

We characterized the structure of neural representations[28,29] by inter-instance pairwise dissimilarity of multi-electrode response patterns, i.e., 120 × 120 representational dissimilarity matrices (RDMs; fig. S3; see Methods). We first visually inspected the neural RDMs using multidimensional scaling (MDS), which approximates dissimilarity with Euclidean distance (Fig. 2A). Non-random structure of instances by class emerged in lower areas earlier than LOC and VOT (80 vs. 130 ms; also see fig. S3E), consistent with a longer decoding latency of the latter. At 180 ms, LOC and VOT began to form an expected tripartite structure by object category. Surprisingly, representations of lower areas also became tripartite at 230 ms or even sooner. To ensure a priori category labels did not bias these observations, we performed an unsupervised clustering analysis (see Methods), which found the tripartite structure fully formed in V1 and V2&V3 in comparable fidelity yet later than LOC, VOT, and V4 (fig. S4).

To infer processing order, we calculated representational similarity (RS) between neural RDMs and two references (Spearman's ρ) derived from HMAX[6], a classic model of low-level vision, and ideal categorical (CAT) representation (Fig. 2B and figs. S3A-S3C; see Methods). Assuming high RS with HMAX and CAT indicates low- and high-order processing, we used their difference ($RS_{CAT} - RS_{HMAX}$) as an index of processing order. The index began to increase in all areas between ~100 and 200 ms, coincidental with the onset and peak of multi-channel decoding accuracy in OTC areas (Fig. 1G), respectively. It turned from negative ($RS_{HMAX} > RS_{CAT}$, $p < 0.05$) to around zero In V1, negative to positive In V2&V3 and V4, and zero to positive in LOC and VOT. The results suggest processing order continues increasing in all areas after the feedforward sweep, while the relative order from V1 (low) through VOT (high) holds, i.e., there is a spatiotemporal hierarchy of object processing.

Because the correlation to both references was initially low in LOC and VOT, the small difference could be due to the floor effect. The enlarged difference later may reflect improved fidelity (Fig. 1G) of the same neural code rather than altered representational structure. To test this possibility, we added Gaussian noise to late OTC responses (see Methods) to see if noise-injected RDMs (NI) could match actual ones at earlier time points in both RS_{CAT} and RS_{HMAX} (Fig. 2C). The noise level at which RS_{HMAX} of NI matched actual RDMs didn't reduce RS_{CAT} of NI sufficiently, which was significantly higher than actual ($p < 0.05$). When we increased the noise level to match RS_{CAT} instead, the noise seemed too strong as RS_{HMAX} of NI was lower than actual, albeit not always significantly. In both cases, $RS_{CAT} - RS_{HMAX}$ of NI (late) was higher than actual (early). Varying

the choice of early and late time points (110-150 and 190 ms, respectively, in Fig. 2C) didn't alter the observed pattern (data not shown). The results suggest representational structures of OTC areas change over time.

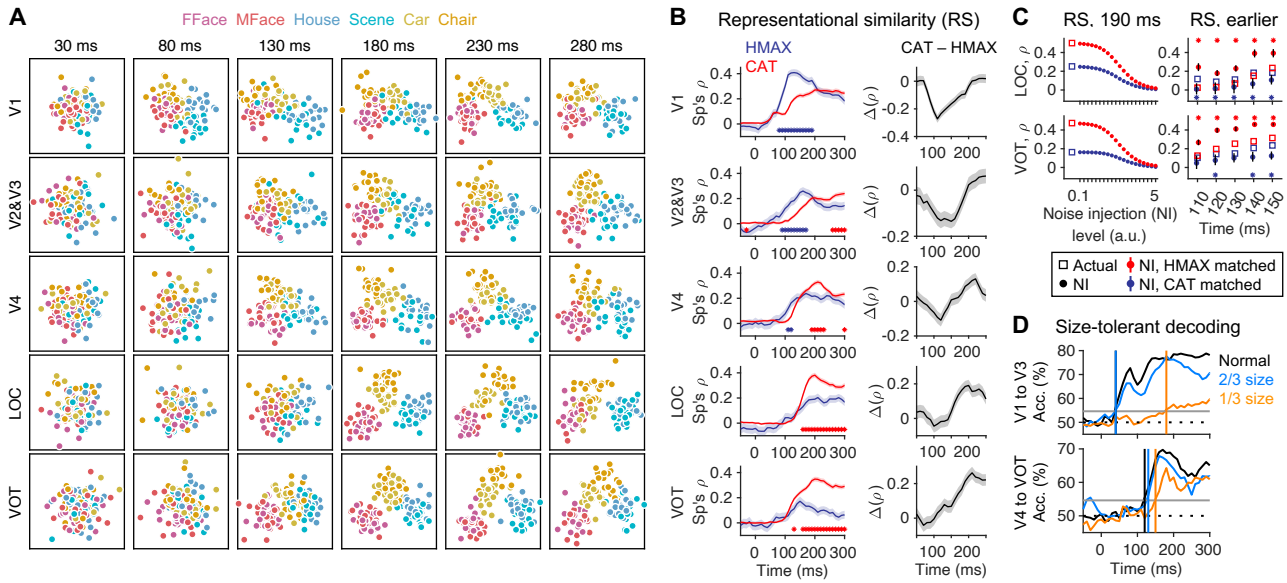

Fig. 2. Spatiotemporal dynamics of object representations. (**A**) MDS visualization of neural RDMs derived from iEEG responses by area (rows) and time (columns). Instances are shown as dots, colored by class. (**B**) Left, RS (Spearman's ρ) of iEEG vs. references (HMAX/CAT). Right, $RS_{CAT} - RS_{HMAX}$. Solid lines and shadings denote mean and standard error of bootstrap distribution. Blue/red asterisks indicate when RS_{HMAX}/RS_{CAT} is significantly higher ($p<0.05$). (**C**) RS of actual and noise injected (NI) iEEG vs. references. Left, RS of actual and NI iEEG at 190 ms (late). Squares, actual; circles, NI at various noise levels (X-axis, logarithmic scale). Right, RS of actual iEEG at 110 to 150 ms (early; squares) and NI iEEG at 190 ms (late; circles). The latter matched the former in RS vs. the other reference (red/blue, HMAX/CAT matched RS_{CAT}/RS_{HMAX}; slightly shifted along X-axis for better visibility). Error bar covers 2.5th to 97.5th quantiles of RS derived from repetitions of NI. Blue/red asterisks indicate when RS_{CAT}/RS_{HMAX} of actual iEEG is significantly lower/higher than HMAX/CAT matched NI iEEG. In some cases, RS_{HMAX} of late iEEG was lower than early iEEG (140/150 ms in VOT) so that injecting noise would never match the two, and thus RS_{CAT} of NI at the lowest noise level is shown (an HMAX-matched late iEEG at a noise level lower than actual would have even higher RS_{CAT}). (**D**) Size-tolerance decoding accuracy of V1 to V3 (top) and V4 to VOT (bottom). The color indicates the image size of the test data. The dashed and solid gray lines indicate the chance level and permuted threshold ($P < 0.01$), respectively. Vertical lines indicate latency by size.

To better understand the nature of time-varying responses across areas, we tested their tolerance to image transformations[7,17,30], a hallmark of high-order object processing. We trained class-level decoders using responses to normal stimuli and tested them using responses to either normal or size-reduced stimuli. We collected data for this analysis from one subject and divided the electrodes into two groups, V1 to V3 and V4 to VOT (n = 11/3; Fig. 2D). Decoding accuracy of 1/3-sized stimuli reached significance ($p < 0.05$) slightly later than normal (150 vs. 120 ms) in V4 to VOT. It also reached significance in V1 to V3, which was unexpected, but much later than normal (180 vs. 40 ms). The performance drop of 1/3-sized vs. normal was bigger in V1 to V3, indicating weaker tolerance. The decoding accuracy of 2/3-sized stimuli was close to normal (latency = 40/130 ms in V1-to-V3/V4-to-VOT), suggesting the transformation was short of breaking stimulus-driven class separability. Consistent with RDM-based analysis, the results

suggest late processing in lower areas is more sophisticated than earlier but not at the same level as in OTC areas.

We could also derive the spatiotemporal hierarchy by relating neural RDMs across areas and time without specifying any references. Instance-level RDMs can be noisy, especially the early ones, and we exacerbated this by calculating split-half RS to avoid bias. To address this, we used class-level RDMs (6 × 6) for this analysis (see Methods). Diagonal patterns of cross-time RS were observed both within and across areas (fig. S5A): RDMs were more similar if closer in time, whether or not from the same area. In some cases (fig. S5B), we even found cross-area, close-time RS higher than same-area, distant-time RS. We then performed MDS on the neural RDMs to infer the major directions along which they differ across areas and time. It turned out that a large portion of variation (44.7%) was along the first MDS dimension (fig. S6A), and its coordinate increases both from V1 through VOT and over time (fig. S6B), which can therefore be considered as an alternative index or processing order. Thus, the spatiotemporal hierarchy corresponded to the most prominent pattern of the data.

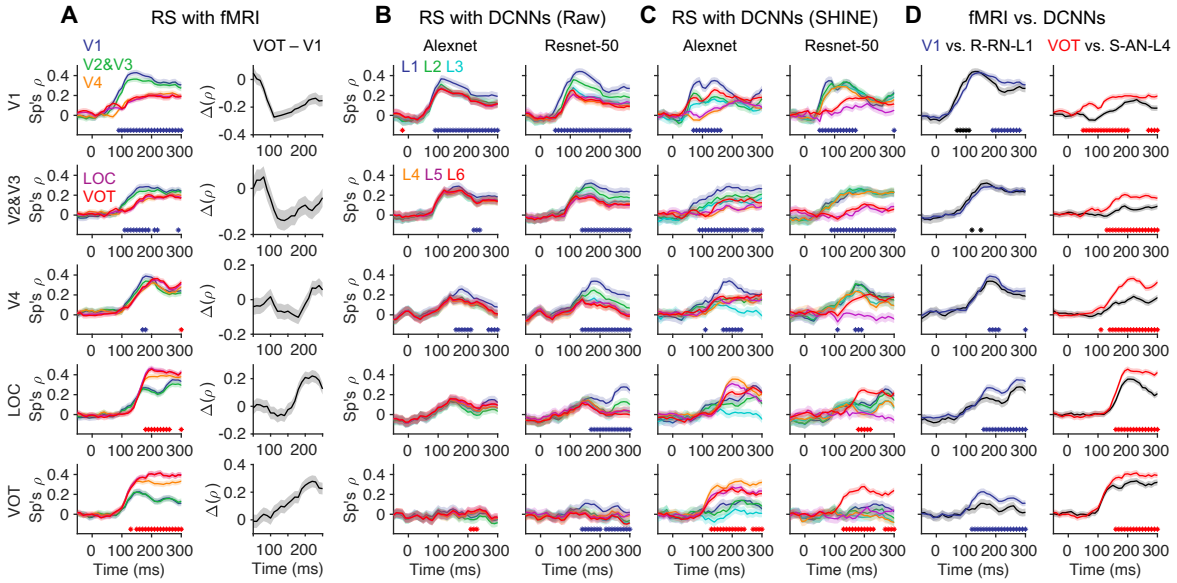

Fig. 3. Relating representational dynamics to fMRI and DCNNs. (**A**) Left, RS of iEEG vs. fMRI. Right, $RS_{fMRI\text{-}VOT} - RS_{fMRI\text{-}V1}$. Solid lines and shadings denote mean and standard error of bootstrap distribution. Blue/red asterisks indicate when $RS_{fMRI\text{-}V1}/RS_{fMRI\text{-}VOT}$ is significantly higher ($p<0.05$). (**B**) RS of iEEG vs. layers L1 to L6 of Alexnet (left) and Resnet-50 (right) taking raw stimuli as input, shown as in (A), left. Blue/red asterisks indicate when RS_{L1}/RS_{L6} is significantly higher. (**C**) As (B), taking SHINED stimuli as input of DCNNs. (**D**) Comparison of iEEG-fMRI RS vs. iEEG-DCNN RS. Left, $RS_{fMRI\text{-}V1}$ (blue) vs. $RS_{Raw\text{-}Resnet50\text{-}L1}$ (black). Right, $RS_{fMRI\text{-}VOT}$ (red) vs. $RS_{SHINE\text{-}Alexnet\text{-}L4}$ (black). Colored/black asterisks indicate when RS_{fMRI}/RS_{DCNN} is significantly higher.

Relating iEEG dynamics to fMRI and DCNNs

To relate our findings to the literature of human fMRI studies, we compared neural RDMs derived from iEEG vs. fMRI responses of ventral visual areas (fig. S3D; see Methods). RDMs of iEEG, in general, showed a similar level of RS with fMRI-V1 and fMRI-V2&V3, and with fMRI-V4, fMRI-LOC, and fMRI-VOT, respectively (Fig. 3A). RDMs of iEEG-V1/V2&V3 and iEEG-LOC/VOT were close to fMRI-V1/V2&V3 and fMRI-V4/LOC/VOT, respectively, and RDMs of iEEG-V4 were similarly close to fMRI RDMs across areas. RS of iEEG RDMs with fMRI-V1/V2&V3 peaked earlier than that with fMRI-LOC/VOT, and the difference (e.g., $RS_{fMRI\text{-}VOT} - RS_{fMRI\text{-}V1}$ in Fig. 3A, right)

increased over time between ~100 and 200 ms, following a similar trajectory as $RS_{CAT} - RS_{HMAX}$ (Fig. 2B, right). We then performed MDS on both iEEG and fMRI RDMs (fig. S6C) and found a similar pattern to the previous analysis on iEEG RDMs alone (fig. S6A). Dimension 1 coordinate of fMRI RDMs increased from V1 through VOT and fell within the range of same-area iEEG RDMs except for V4 (fig. S6D). Notably, fMRI responses of V4/LOC/VOT best matched late iEEG response after ~200 ms of the same area (figs. S6C-S6D; also see Fig. 3A). The results suggest the spatial hierarchy inferred from fMRI data aligns with the spatiotemporal hierarchy inferred from iEEG data, although fMRI underestimates how much the representational structure varies across areas and time.

Previous research reported functional correspondence[27,31–33] between high-level ventral visual areas in primates and top layers of deep convolutional neural networks (DCNNs), a family of purely feedforward models. Therefore, we compared iEEG RDMs with those derived from selected layers (L1 to L6 from bottom to top; fig. S3F; see Methods) of Alexnet and Resnet-50, two DCNNs commonly used in the field. RS with DCNNs was high for iEEG RDMs of V1 through V4 but very low for those of LOC and VOT, and RS with L1 was higher than L6 in most cases (Fig. 3B). Interestingly, performing the SHINE (spectrum, histogram, and intensity normalization and equalization) operation[34] on the stimuli before feeding them into the DCNNs made RDMs of top layers more distant from bottom layers (fig. S3F) and much closer to neural RDMs of late OTC responses (Fig. 3C). Nonetheless, the best-matching DCNN layer (L4 of Alexnet) after the SHINE operation didn't match late OTC responses as well as same-area fMRI responses (bottom two rows in Fig. 3D, right). In contrast, the best-matching layer (L1 of Resnet-50 on raw stimuli) matched early responses of V1 to V4 equally well as fMRI or slightly better (top three rows in Fig. 3D, left). The results suggest DCNNs don't fully explain high-order neural processing in the spatiotemporal hierarchy. We could also assess brain-DCNN correspondence by predicting single-electrode, instance-level responses with linearly mixed DCNN output (see Methods)[31,35]. However, this approach was unsuitable for our data because the predictive models failed to generalize across classes (fig. S7), a strong test of model validity proposed and passed in previous work[31] predicting spiking activity in monkey inferotemporal cortex (IT, homolog of human OTC).

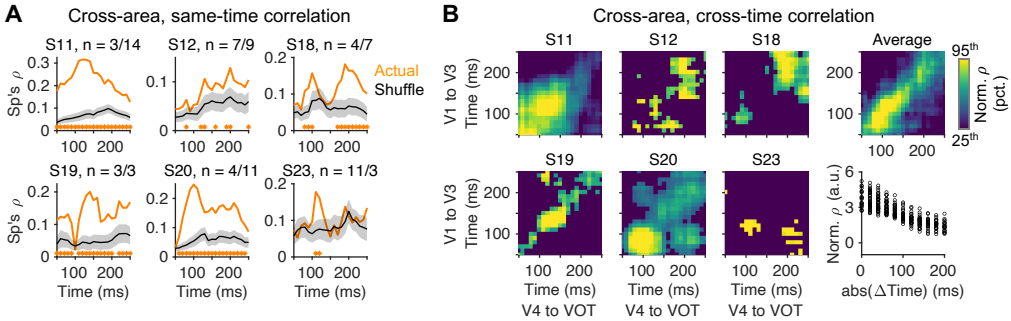

Fig. 4. Trial-by-trial association across areas. (**A**) Cross-area correlation (Spearman's ρ) of concurrent multi-electrode patterns projected to common axes. Orange/black, correlation derived from actual/shuffled data. Shadings indicate standard error of null distribution (shuffle). Orange asterisks indicate when the correlation is significantly higher than shuffle ($P < 0.05$). n indicates # electrodes in V1 to V3 and V4 to VOT per subject, respectively. (**B**) Cross-area, cross-time correlation matrices of individual subjects (left) and population average (right-top). X-axis and Y-axis indicate the time of V4 to VOT and V1 to V3 when calculating correlation, respectively. Normalized ρ is shown (actual divided by shuffle). We set insignificant correlation values to 0 before averaging. Right-bottom, population-average correlation vs. inter-areal delay (absolute value). Circles indicate entries of the correlation matrix.

Trial-by-trial association across areas

The aligned, simultaneous representational changes in OTC and lower areas imply recurrence between them[36,37] and/or a common top-down input[38]. In either case, we'd expect shared trial-by-trial neural variability between areas[39]. We quantified this using within-subject, cross-area correlation (Spearman's ρ) of multi-electrode patterns projected to common axes derived by class-level decoders (see Methods). To address the confounding effect of common visual input, we computed ρ using same-class trials only and constructed a baseline by shuffling same-instance trials. In six subjects with electrodes covering both V1 to V3 and V4 to VOT, the cross-area correlation of same-time responses (Fig. 4A) was significantly ($p < 0.05$) higher than baseline at multiple time points between 100 and 200 ms. We further examined cross-area, cross-time correlation (Fig. 4B) and found a trend of higher correlation between adjacent time points ($r = -0.87$, $p = 1.8 \times 10^{-139}$). We didn't find clear inter-areal lead/lag in this analysis. Thus, there was shared trial-by-trial variability between areas, which was strongest for concurrent responses.

Discussion

By characterizing the structure of object representations derived from spatiotemporally resolved iEEG responses in humans, we have shown that processing order in all ventral visual areas increases from ~100 to 200 ms. However, at any point, it remains higher in OTC than in lower areas. During the period, early and late responses of the same area exhibit properties expected for areas lower and higher in the ventral pathway, respectively: late responses of V1 to V3 are modulated by object category and show tolerance to image size, and early responses of LOC and VOT are more driven by low-order features and less alike same-area fMRI responses than late responses. Taken together, our findings reveal a parallel in the time domain of the much-studied spatial hierarchy of object processing.

A prominent view holds that object processing in the primate brain is largely feedforward[2], supported by the observation that the first ~50 ms of neuronal responses in monkey IT already support reliable object decoding[30]. Previous work suggests recurrent processing, beyond a cascade of intra-areal and feedforward inter-areal processing, is conditionally invoked, e.g., when recognition is challenging due to blur, clutter, occlusion, etc., at the cost of extra delay[14,40–43]. The present study goes further and supports a universal role of recurrent processing by showing that even when presented with clear, isolated objects, early responses (~100 to 150 ms) in OTC areas are less sophisticated than responses tens of milliseconds later. An important corollary is that fMRI responses in high-level ventral visual areas[9,10,44] result from recurrent processing at least partially. Coordinated representational changes and trial-by-trial association of concurrent response patterns imply that recurrent interactions between OTC and lower areas may contribute to late responses. Future causal[45–47] experiments are needed to test whether this is true, or rather, albeit less likely, intra-areal recurrent processes are sufficient. A common top-down input[38,41] may also explain our findings, but no area outside the ventral pathway has known monosynaptic projections to all stages[48,49].

Although DCNNs remain the best model of primate object recognition[35,50], a growing number of studies show they don't fully account for the behavioral and neural responses[27,43,51]. In line with these studies, we found DCNNs didn't mimic the representational structure of late OTC responses without equalizing low-order features first and still underperformed compared to same-area fMRI responses even after such operation. Nonetheless, if the computational goals were the same, feedforward DCNNs might rely on extracting the same high-order features as recurrent brain circuits. DCNNs and the brain may use different readout weights, which could explain distinct representational structures. Thus, it will be interesting to test if linearly mixed output of DCNNs predicts late OTC responses worse than early responses and to what extent[43], and if the results depend on using clear, isolated stimuli and/or eye movements. To run such tests properly, DCNN-

based predictive models ought to be accurate and capable of generalization[31], for which our analysis (fig. S7) provides a cautionary tale.

Primates often take quick actions that utilize object information, e.g., moving their eyes rapidly to look for a target object[52]. In this scenario, it can be ecologically beneficial to rely on feedforward object processing for its speed. When more time is allowed, recurrent processing might enable more robust recognition: feedback signals from downstream neurons with larger receptive fields may help an upstream neuron disambiguate[36,53] the part of the object it sees, and updated output of the latter may, in turn, refine downstream representations.

ACKNOWLEDGEMENTS

We thank Chen Song and Dan Zhang for assistance in data collection. This work was supported by the National Key R&D Program of China (2017YFA0205904 to B.H.) and National Natural Science Foundation of China (NSFC 62061136001 to B.H.).

METHODS

Participants and paradigm

Twenty-one patients (7 female, 12-47 years old) with pharmacologically intractable epilepsy participated in this study, intracranially implanted with iEEG electrodes to localize seizure foci. They were admitted to Tsinghua University Yuquan Hospital (TYH; n = 13) or Chinese PLA General Hospital (PGH). The Institutional Review Boards at both hospitals and Tsinghua University approved the study. We obtained informed consent in writing from all subjects before participation. All subjects had a normal or corrected-to-normal vision.

Subjects viewed 120 grayscale square images of clear, isolated objects from three categories: faces, places, and other objects. Each category was further divided into two subcategories or classes (six in total), i.e., female faces, male faces, houses, scenes, cars, and chairs, and there were 20 instances per class. We ensured that same-class instances shared some visual features such as orientation and silhouette. The stimuli were used in a previous fMRI study[54]. In each trial of the main iEEG experiment, an image pseudorandomly drawn from the full set was shown for 300 ms in the center of a black screen at ~6.5 visual degrees and then disappeared for 600-750 ms, while a small white fixation point was continuously present in the center of the screen. For four subjects (2 electrodes of interest) recruited at the beginning of the study, images were shown for 500 ms, followed by blanks of 750-1000 ms. In a control experiment to assess tolerance to image size, images were presented at full, 2/3, or 1/3 of the regular size for 200 ms, followed by blanks of 700-850 ms. We instructed subjects to maintain their fixation and press a key if they saw the same image (regardless of size) shown in the previous trial, i.e., to perform a 1-back task ($P_{1\text{-back}}$ = 12%). We put 45 to 60 trials in a block, and subjects could rest between blocks. We collected 348 to 1566 trials per subject for the main experiment and 870 trials in one subject (S23; 14 electrodes of interest) for the control experiment.

Thirteen subjects also participated in a modified version of the main experiment in an fMRI session consisting of 3 to 6 runs. We used a blocked design for optimal contrast-to-noise ratio. Each run had six 32-sec stimulus blocks, one for each class, interleaved with seven 16-second fixation blocks. Each stimulus block consisted of 20 or 40 trials showing same-class images in a pseudorandom order, for 300 or 600 ms, followed by 500-ms or 1000-ms banks. We used the version with fewer, longer trials at the beginning of the study (n = 6), and the fMRI response patterns didn't differ systematically between versions. Images were presented in the center of a white screen at ~10 visual degrees with small jittering of position. Subjects performed the same 1-back task. We conducted both iEEG and fMRI experiments using Psychtoolbox-3[55].

Data acquisition and preprocessing: iEEG

Subdural surface grids/strips and/or depth electrodes were implanted based on clinical considerations only. The iEEG signals were sampled at 1200 Hz and 512 to 1024 Hz using g.Tec and Nicolet amplifiers at TYH and PGH, respectively. We applied a notch filter at 50 Hz on the data. Electrodes showing epileptic discharge were manually identified and removed from further analyses. Electrodes and trials showing other artifacts were removed semi-automatically using *ft_rejectvisual* in FieldTrip. The first two trials of a block and trials in which the image repeated the previous trial or subjects pressed the key were also discarded. We ended up with 211 to 1150 valid trials per subject in the main experiment and 548 in the control experiment (used in Fig. 2D only).

In this study, we focused on high gamma or broadband activity, which is a non-oscillatory component with no particular frequency peak and correlates with neuronal spiking. We set 10-ms spaced time points from -200 to 800 ms (relative to the stimulus onset) as time points of interest (TOIs), and for each TOI, we built ten four-cycle Morlet wavelets and had their center frequency logarithmically spaced from 110 to 140 Hz. The spectral bandwidth of the wavelets together

covered 82.5-185 Hz, and the wavelet duration was between 9.1 and 11.6 ms. We computed power spectral density (PSD) for each wavelet. We took the logarithm to make it more Gaussian, which was then normalized (subtraction before division) for each electrode separately, based on the mean and standard error of log(PSD) between -100 and 0 ms over trials and experiments. Finally, we averaged log-normalized PSD over wavelets and used it as the metric of iEEG responses (see Fig. 1D for examples) in subsequent analyses.

Data acquisition and preprocessing: MRI

We collected structural (MPRAGE; TR/TE = 2530/3.45 ms; 1 mm isotropic) and, in some cases, blood-oxygenation level-dependent (BOLD) functional (EPI; TR/TE = 2000/30 ms; voxel size = $3.125 \times 3.125 \times 4$ mm^3; 25 slices; FA = 90 degrees; no gap) MRI images in a Philips Achieva 3.0 T TX scanner before subjects were implanted. Using FreeSurfer[56], individual brain surfaces were reconstructed based on structural MRI images, nonlinearly transformed to a template brain (fsaverage), and parcellated according to a population atlas[26]. FreeSurfer also generated volume-based parcellation (e.g., of cortical areas, subcortical structures, and white matter) of individual brains. Using FreeSurfer functional analysis stream (FS-FAST), volume-based fMRI data processing steps included motion correction, slice timing correction, and quadratic linear detrending. The volume-based data were then projected onto individual surfaces and smoothed (FWHM = 5 mm). A general linear model computed each class's regression coefficient (beta). We also computed an omnibus contrast (stimulus vs. fixation). We transformed the statistics to fsaverage for further analyses.

Electrode localization and selection

We followed a semi-automatic pipeline[57] with reported precision within 3 mm to localize electrodes in the brain. We manually labeled the location of electrodes based on postoperative computer tomography (CT) images. We aligned the CT images to same-subject structural MRI images via rigid transformation using a mutual information-based algorithm and manual adjustment if needed. To address brain deformation due to the insertion of electrodes, a constrained energy-minimization algorithm was used to pull surface electrodes onto smoothed pial surface (pial-outer-smoothed) while minimizing the displacement of electrodes and deformation of grids/strips. For depth electrodes, we considered them cortical if they were closer to cortical than subcortical voxels and the distance to the nearest cortical voxel was no larger than 3.5 mm. Finally, for all surface and depth cortical electrodes, we found the nearest vertex on the individual pial surface and the corresponding vertex on fsaverage surfaces.

We narrowed the cortical electrodes down to the electrodes of interest for further analyses in three successive steps (fig. S1), using data from the main experiment. First, we looked for visually responsive electrodes by comparing mean responses over TOIs of 50 to 300 ms vs. -100 to 0 ms (fixation) for trials of each class separately. We kept electrodes showing significantly stronger-than-fixation responses (paired T-test, $P < 10^{-4}$) to at least one class. All P values reported in this study are two-sided. Second, we performed single-electrode, class-level decoding analysis and kept electrodes showing significantly above-chance-level accuracy (permutation test, $P < 0.05$). Lastly, we discarded electrodes not in ventral visual areas according to the population atlas or in a vertex in which at least one subject showed significantly positive fMRI responses (omnibus contrast, uncorrected $P < 10^{-4}$).

Single- and multi-electrode decoding analysis

For single-electrode decoding analysis, we used linear SVM classifiers (LIBSVM; C = 1) to decode the object class of single trials, taking responses from 50 to 300 ms (26 TOIs) as features. We built 15 sets of binary classifiers, one for each pair of classes. For all class pairs and electrodes, we chose the same number of trials (30 per class, 60 in total) pseudorandomly from all valid trials in each cross-validation (CV) fold, of which we used 2/3 and 1/3 to train/test the classifier. 30 was

the minimal number of total valid trials for any subject and class. We normalized training and test data in each CV fold by subtracting the mean and dividing by the standard error of training data for each TOI separately. We did 50 CV folds per classifier for each electrode and reported decoding accuracy as the mean over 15 class pairs (chance level = 50%). We performed a permutation test to assess statistical significance by shuffling the class label of all valid trials and then running decoding analysis as normal (15 classifiers, 50 CV folds per classifier). We did 2000 permutations per electrode to construct a robust null distribution. We considered decoding accuracy significant ($P < 0.05$) if it was larger than the 97.5^{th} percentile of the null distribution.

We assessed single-electrode latency of class-level decoding using the decoding analysis described above but taking single TOI (between -180 and 780 ms) as the feature. We used the mean response over 5 TOIs at -20 to 20 ms relative to each TOI in this and all other analyses that required temporal resolution. For each electrode, we did 20 permutations and merged the results across TOIs into a single null distribution (1940 permutations) as they were similar. The earliest TOI between 10 and 300 ms was taken as decoding latency, starting from which five consecutive TOIs showed significant decoding accuracy (permutation test, $P < 0.05$). Because we used few data than the previous analysis, there was no valid latency for several electrodes of interest (n = 3). Similarly, we set single-electrode latency of visual responses as the earliest TOI, starting from which five consecutive TOIs showed significantly (paired T-test, $P < 10^{-4}$) stronger-than-fixation (-100 to 0 ms) responses to at least one class, using all valid trials.

We performed multi-electrode decoding analysis for individual TOIs and areas, respectively, taking multi-electrode response pattern at a given TOI and area as features. We determined the significance and latency of multi-electrode decoding similarly to single-electrode decoding, with a stricter threshold ($P < 0.01$). We conducted Size-tolerant multi-electrode decoding analysis similar to the non-tolerant version. In each CV fold, we drew 45/23 trials per class and size (2/3 of the minimal number of total valid trials for any class and size) from main/control experiments as training/test data. We shuffled the joint class/size label of trials in the main and control experiments separately for the permutation test.

Representational similarity analysis

Because each instance was presented to a varying number of electrodes, we used a modified version of standardized Euclidean distance to quantify the inter-instance dissimilarity of multi-electrode response patterns. First, we converted responses to up to 120 instances (averaged over same-instance trials) to the Z-score for each electrode separately. Second, we computed the dissimilarity between the two instances as the root mean square of the difference between Z-score vectors consisting of electrodes to which both instances were presented. We built one 120 $\times$ 120 representational dissimilarity matrix (RDM) of iEEG for every area and TOI (fig. S3E). To visualize the RDMs, we applied non-classic MDS using *mdscale* in MATLAB with ten dimensions to get low-stress solutions. We showed the first two MDS dimensions and used Procrustes transformation to stretch/stress and rotate the 120-instance clouds to align them across areas and time (Fig. 2A).

The 120 $\times$ 120 RDM of HMAX (fig. S3B) was computed as standardized Euclidean distance across rows of 120 $\times$ # units matrix consisting of the output of all units at the C2 layer of HMAX model, using *pdist* in MATLAB (*Distance = 'seuclidean'*). The 120 $\times$ 120 RDMs of DCNNs (fig. S3F) were computed in similar ways based on the output of units in six selected layers (each) of two representative DCNNs, which were, from bottom to top (L1 to L6), 'pool1', 'pool2', 'pool5', 'fc6', 'fc7', and 'fc8' in Alexnet, and 'max_pooling2d_1', 'activation_10_relu', 'activation_22_relu', 'activation_40_relu', 'avg_pool', and 'fc1000' in Resnet-50, respectively, following a recent study[27]. We created two sets of RDMs per DCNN, one using the images presented to subjects as input ('Raw'), the other using the same images but with some low-order features equalized with the

SHINE (spectrum, histogram, and intensity normalization and equalization) operation[34] as input ('SHINE'). The 120 × 120 RDM of ideal categorical (CAT) representation (fig. S3C) was built by setting all same-category entries as zero (similar) and the rest as one (dissimilar).

For fMRI, we started with 6 × 6 inter-class RDMs and then converted them to 120 × 120 inter-instance RDMs. For each area, we pooled vertices from both hemispheres of a subject that showed significantly positive responses (omnibus contrast, $P < 10^{-4}$) to build a 6 × # vertices matrix of beta value. We then computed standardized Euclidean distance across rows of the beta matrix. The results were averaged over subjects with at least 200 significant vertices in the area to generate the inter-class RDM, in which all diagonal entries were zero. Before averaging, we normalized individual subjects' RDMs with a division by the square root of the number of significant vertices. Finally, we built the corresponding inter-distance RDM (fig. S3D) by setting dissimilarity between two instances as that between respective classes of the instances.

When computing representational similarity (RS) between inter-instance RDMs, we only considered 6000 above-diagonal (for the RDMs were all symmetrical) entries corresponding to the dissimilarity between instances from different classes (fig. S3A). We didn't consider dissimilarity between same-class instances because it was undefined for fMRI RDMs and expected to be small for both low- and high-order representations due to our choice of stimuli. To assess statistical significance, we created a bootstrap distribution of RS using iEEG RDMs derived from a subset of data pseudorandomly drawn from all trials with replacement. We made 2000 rounds of bootstrap, and trials were selected once per round for all electrodes of the same subject. We considered an iEEG RDM significantly more similar to one reference RDM than another ($P < 0.05$) if the RS of the former reference was higher in more than 97.5% of bootstrap rounds.

To build noise-injected iEEG RDMs (NI), we added Gaussian noise to instance-level responses for each electrode separately ($\mu = 0$; $\sigma = k$ times the stand error of responses, where k was the noise injection level and could take 20 values spaced logarithmically from 0.1 to 5). We then took the steps of building non-bootstrap iEEG RDMs as described at the beginning of this section. We did this 2000 times per noise injection level, and we considered RS of NI significantly higher or lower than actual ($P < 0.05$) if this was the case more than 97.5% of the time.

Because iEEG RDMs were noise, especially those of early TOIs, when comparing iEEG RDMs with each other, we used 6 × 6 inter-class RDMs built the same way as inter-instance RDMs but based on responses averaged over same-class trials. To avoid bias toward a stronger correlation between same-area, close-time RDMs, we calculated RS between inter-class RDMs based on two separate halves of trials (split-half RS). In fig. S5, we did bootstrap resampling and calculated split-half RS in each bootstrap round (500 rounds, one split-half per round) to assess statistical significance. In fig. S6, where a significance test wasn't needed, we averaged RS over 50 split-halves of the full dataset for less noisy results.

Unsupervised clustering analysis

We applied the K-means algorithm on inter-instance RDMs to group instances into three clusters without being biased by *a priori* category labels. Since the algorithm doesn't guarantee equally sized clusters, for each cluster larger than 40, we found its member closest to the centroid of any cluster smaller than 40 and moved it to the latter cluster. We repeated this process until the clusters were equally sized. We calculated the percentage of instances in the "correct" cluster for each of the six possible ways of assigning category labels to clusters and reported the highest percentage as clustering accuracy. This metric is, by design, positively biased when clustering is noisy, which explains why pre-stimulus clustering accuracy was around 40% instead of 1/3. We used bootstrap resampling of trials (2000 rounds) to assess statistical significance.

Predicting single-electrode responses with DCNN output

We tested how well single-electrode, instance-level response can be modeled as linearly mixed output of units in selected layers of DCNNs taking raw or SHINED stimuli as input (see "Representational similarity analysis"). We used partial least squares (PLS) regression[31] (# components = 5) to estimate the mixing weights based on responses to some instances and used the learned weights to predict responses to other instances. Changing # components didn't alter the pattern of results (data not shown). The performance of prediction was measured as the cross-validated normalized mean square error (norm. MSE) using *goodnessOfFit* in MATLAB (*cost_func = 'NMSE'*), which is effectively 1 – goodness-of-fit (r^2). A value of 0 indicates a perfect fit, 1 indicates fit as good as a flat line through the mean response, and larger than 1 indicates a bad fit likely due to overfitting in training. We did two types of six-fold cross-validation, and in each fold, we either used 5/6 and 1/6 of instances, regardless of their class, as training and test data, respectively (object generalization) or used instances of 5 class as training data and the other class as test data (class generalization).

Reliable class- and instance-level modulation of single-electrode responses

We quantified class- and instance-level modulation of single-electrode, single-trial responses based on the partition of sums of squares (SS): the total sum of squares (SS_{Total}), i.e., the sum of squares of deviation from single-trial response to the total mean summed over all trials, equals the sum of SS_{Class} (deviation from class mean to total mean), $SS_{Instance}$ (deviation from instance mean to class mean), and SS_{Error} (deviation from single-trial response to instance mean). We only considered instances presented in at least three trials for each electrode to separate $SS_{Instance}$ from SS_{Error}. SS_{Class} and $SS_{Instance}$ divided by SS_{Error} (normalized SS) can be used as class- and instance-level modulation metrics, respectively. We didn't separate class-level modulation within and across categories. Because non-error SS terms are always overestimated with limited data due to noise, we created a control term for SS_{Class} (SSC_{Class}) by shuffling the class label of instances without changing the instance label of trials. We created another control term for $SS_{Instance}$ ($SSC_{Instance}$) by shuffling the instance label of trials within each class. We subtracted the SSC terms (mean of 500 shuffles) from respective non-error SS terms and added both to SS_{Error} when calculating normalized SS.

Trial-by-trial association of multi-electrode response patterns

To measure the within-subject, trial-by-trial association between multi-electrode response patterns, we projected them to axes perpendicular to classification hyperplanes of the 15 binary classifiers, which are comparable across areas and time. In each CV fold of decoding analysis, we used 30/15 trials per class as training/test data in five subjects with coverage of both V1 to V3 and V4 to VOT and 20/9 trials in another subject who had fewer trials available. Unlike other decoding analyses, here we applied the trainer classifier to test trials of all classes. We computed Spearman's ρ using projected responses along each of the 15 axes for each class to avoid spurious correction since classes modulated responses. To address the additional confounding effect of instance-level modulation, we constructed a baseline by shuffling the order of same-instance trials before computing ρ. We didn't analyze trials of instances presented fewer than three times to a subject. We assessed statistical significance based on the null distribution constructed from 2000 shuffles.

REFERENCES

1. Thorpe, S., Fize, D. & Marlot, C. Speed of processing in the human visual system. *Nature* **381**, 520–522 (1996).
2. DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How Does the Brain Solve Visual Object Recognition? *Neuron* **73**, 415–434 (2012).
3. 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 Cogn. Sci.* **17**, 26–49 (2013).
4. Grill-Spector, K. & Malach, R. THE HUMAN VISUAL CORTEX. *Annu. Rev. Neurosci.* **27**, 649–677 (2004).
5. Weiner, K. S. & Grill-Spector, K. Neural representations of faces and limbs neighbor in human high-level visual cortex: evidence for a new organization principle. *Psychol. Res.* **77**, 74–97 (2013).
6. Riesenhuber, M. & Poggio, T. Hierarchical models of object recognition in cortex. *Nat. Neurosci.* **2**, 1019–25 (1999).
7. Rust, N. C. & DiCarlo, J. J. Selectivity and Tolerance (“Invariance”) Both Increase as Visual Information Propagates from Cortical Area V4 to IT. *J. Neurosci.* **30**, 12978–12995 (2010).
8. Mur, M. & Kriegeskorte, N. What’s there, distinctly, when and where? *Nat. Neurosci.* **17**, 332–333 (2014).
9. Haxby, J. V *et al.* Distributed and Overlapping Representations of Faces and Objects in Ventral Temporal Cortex. *Science* **293**, 2425–2430 (2001).
10. Op de Beeck, H. P., Haushofer, J. & Kanwisher, N. G. Interpreting fMRI data: maps, modules and dimensions. *Nat. Rev. Neurosci.* **9**, 123–135 (2008).
11. Konkle, T. & Oliva, A. A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron* **74**, 1114–1124 (2012).
12. Liu, J., Harris, A. & Kanwisher, N. Stages of processing in face perception: an MEG study. *Nat. Neurosci.* **5**, 910–916 (2002).
13. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat. Neurosci.* **17**, 455–462 (2014).
14. 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**, 1–23 (2018).
15. 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).
16. Graumann, M., Ciuffi, C., Dwivedi, K., Roig, G. & Cichy, R. M. The spatiotemporal neural dynamics of object location representations in the human brain. *Nat. Hum. Behav.* **6**, 796–811 (2022).
17. Liu, H., Agam, Y., Madsen, J. R. & Kreiman, G. Timing, Timing, Timing: Fast Decoding of Object Information from Intracranial Field Potentials in Human Visual Cortex. *Neuron* **62**, 281–290 (2009).
18. Ghuman, A. S. *et al.* Dynamic encoding of face information in the human fusiform gyrus. *Nat. Commun.* **5**, 5672 (2014).
19. Jacques, C. *et al.* Corresponding ECoG and fMRI category-selective signals in human ventral temporal cortex. *Neuropsychologia* **83**, 14–28 (2016).
20. Quiroga, R. Q., Reddy, L., Kreiman, G., Koch, C. & Fried, I. Invariant visual

representation by single neurons in the human brain. *Nature* **435**, 1102–1107 (2005).

21. Cao, R., Li, X., Brandmeir, N. J. & Wang, S. Encoding of facial features by single neurons in the human amygdala and hippocampus. *Commun. Biol.* **4**, 1394 (2021).
22. Parvizi, J. & Kastner, S. Promises and limitations of human intracranial electroencephalography. *Nat. Neurosci.* **21**, 474–483 (2018).
23. Miller, K. J. *et al.* Broadband changes in the cortical surface potential track activation of functionally diverse neuronal populations. *Neuroimage* **85**, 711–720 (2014).
24. Manning, J. R., Jacobs, J., Fried, I. & Kahana, M. J. Broadband Shifts in Local Field Potential Power Spectra Are Correlated with Single-Neuron Spiking in Humans. *J. Neurosci.* **29**, 13613–13620 (2009).
25. Ray, S. & Maunsell, J. H. R. Different Origins of Gamma Rhythm and High-Gamma Activity in Macaque Visual Cortex. *PLoS Biol.* **9**, e1000610 (2011).
26. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536**, 171–178 (2016).
27. Xu, Y. & Vaziri-Pashkam, M. Limits to visual representational correspondence between convolutional neural networks and the human brain. *Nat. Commun.* **12**, 2065 (2021).
28. Edelman, S. Representation is representation of similarities. *Behav. Brain Sci.* **21**, 449–67; discussion 467-98 (1998).
29. Kriegeskorte, N. *et al.* Matching Categorical Object Representations in Inferior Temporal Cortex of Man and Monkey. *Neuron* **60**, 1126–1141 (2008).
30. 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).
31. Yamins, D. L. K. *et al.* Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl. Acad. Sci.* **111**, 8619–8624 (2014).
32. Guclu, 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**, 10005–10014 (2015).
33. Long, B., Yu, C.-P. & Konkle, T. Mid-level visual features underlie the high-level categorical organization of the ventral stream. *Proc. Natl. Acad. Sci.* **115**, E9015–E9024 (2018).
34. Willenbockel, V. *et al.* Controlling low-level image properties: The SHINE toolbox. *Behav. Res. Methods* **42**, 671–684 (2010).
35. Ratan Murty, N. A., Bashivan, P., Abate, A., DiCarlo, J. J. & Kanwisher, N. Computational models of category-selective brain regions enable high-throughput tests of selectivity. *Nat. Commun.* **12**, 5540 (2021).
36. Chen, M. *et al.* Incremental Integration of Global Contours through Interplay between Visual Cortical Areas. *Neuron* **82**, 682–694 (2014).
37. Nayebi, A. *et al.* Recurrent Connections in the Primate Ventral Visual Stream Mediate a Trade-Off Between Task Performance and Network Size During Core Object Recognition. *Neural Comput.* **34**, 1652–1675 (2022).
38. 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).
39. Panzeri, S., Moroni, M., Safaai, H. & Harvey, C. D. The structures and functions of correlations in neural population codes. *Nat. Rev. Neurosci.* **23**, 551–567 (2022).
40. Tang, H. *et al.* Spatiotemporal Dynamics Underlying Object Completion in Human Ventral

Visual Cortex. *Neuron* **83**, 736–748 (2014).

41. Fyall, A. M., El-Shamayleh, Y., Choi, H., Shea-Brown, E. & Pasupathy, A. Dynamic representation of partially occluded objects in primate prefrontal and visual cortex. *Elife* **6**, e25784 (2017).
42. Tang, H. *et al.* Recurrent computations for visual pattern completion. *Proc. Natl. Acad. Sci.* **115**, 8835–8840 (2018).
43. 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).
44. Bao, P., She, L., McGill, M. & Tsao, D. Y. A map of object space in primate inferotemporal cortex. *Nature* **583**, 103–108 (2020).
45. Pascual-Leone, A. & Walsh, V. Fast Backprojections from the Motion to the Primary Visual Area Necessary for Visual Awareness. *Science* **292**, 510–512 (2001).
46. 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).
47. Siddiqi, S. H., Kording, K. P., Parvizi, J. & Fox, M. D. Causal mapping of human brain function. *Nat. Rev. Neurosci.* **23**, 361–375 (2022).
48. Felleman, D. J. & Van Essen, D. C. Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cereb. Cortex* **1**, 1–47 (1991).
49. Xu, R., Bichot, N. P., Takahashi, A. & Desimone, R. The cortical connectome of primate lateral prefrontal cortex. *Neuron* **110**, 312-327.e7 (2022).
50. Bashivan, P., Kar, K. & DiCarlo, J. J. Neural population control via deep image synthesis. *Science* **364**, eaav9436 (2019).
51. Rajalingham, R. *et al.* Large-Scale, High-Resolution Comparison of the Core Visual Object Recognition Behavior of Humans, Monkeys, and State-of-the-Art Deep Artificial Neural Networks. *J. Neurosci.* **38**, 7255–7269 (2018).
52. Bichot, N. P., Xu, R., Ghadooshahy, A., Williams, M. L. & Desimone, R. The role of prefrontal cortex in the control of feature attention in area V4. *Nat. Commun.* **10**, 5727 (2019).
53. Epshtein, B., Lifshitz, I. & Ullman, S. Image interpretation by a single bottom-up top-down cycle. *Proc. Natl. Acad. Sci. U. S. A.* **105**, 14298–303 (2008).
54. Xu, R., Zhen, Z. & Liu, J. Mapping Informative Clusters in a Hierarchical Framework of fMRI Multivariate Analysis. *PLoS One* **5**, e15065 (2010).
55. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
56. Fischl, B. FreeSurfer. *Neuroimage* **62**, 774–781 (2012).
57. Dykstra, A. R. *et al.* Individualized localization and cortical surface-based registration of intracranial electrodes. *Neuroimage* **59**, 3563–3570 (2012).

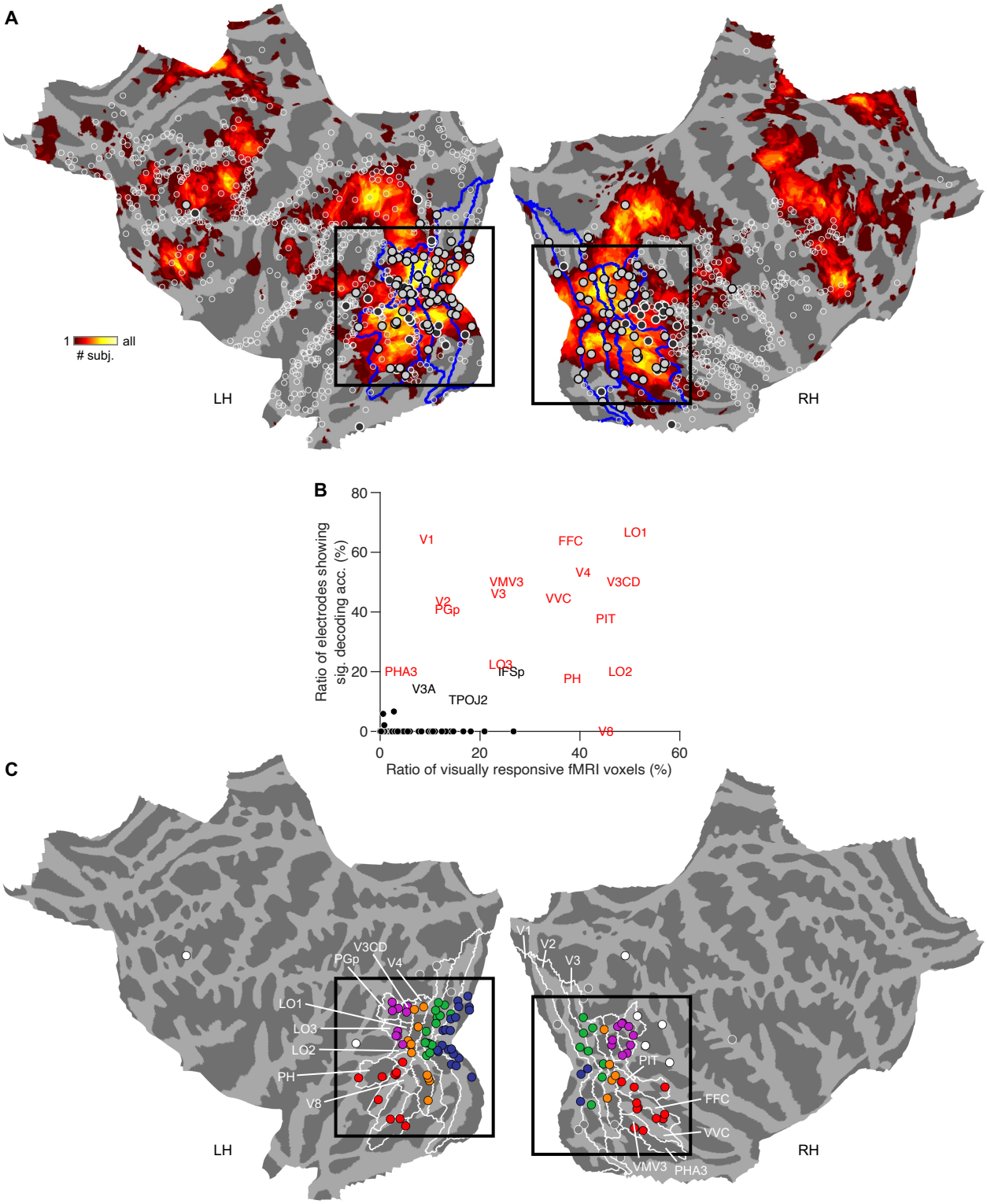

Fig. S1. Coverage and grouping of iEEG electrodes. (**A**) All cortical electrodes on flattened surfaces of fsaverage, along with fMRI responses to the same stimuli shown as in Fig. 1B. Hollow circles, electrodes not showing significantly stronger-than-fixation responses ($P < 10^{-4}$) to any object class (n = 1055, including four cortical electrodes not shown due to cuts made to flatten

the surface). Gray/black circles, visually responsive electrodes showing / not showing significantly above-chance-level decoding accuracy ($P < 0.05$; $n = 122/50$). Overlay indicates # subjects showing significantly positive fMRI responses ($P < 10^{-4}$). Blue contours are boundaries of five areas/stages along the ventral visual pathway (V1, V2&V3, V4, LOC, and VOT), which each may consist of one or more areas in a population atlas. Black rectangles match the field of view in Fig. 1C. Most decodable electrodes were found in ventral visual areas and co-localized with strong fMRI responses. (**B**) The ratio of decodable electrodes vs. visually responsive fMRI voxels in the same area (both hemispheres combined) for all cortical areas in the atlas covered by at least five electrodes. We show ventral visual areas as red text, areas outside the pathway where>10% of electrodes were decodable as black text, and the other areas as black dots. iEEG decodability and fMRI visual responsiveness were significantly correlated ($r = 0.65$, $p = 2.8 \times 10^{-10}$), suggesting good consistency between modalities. (**C**) Decodable electrodes on flattened surfaces of fsaverage. Colored and white circles indicate electrodes in one of the five ventral visual areas (electrodes of interest) and the other cortical areas, respectively ($n = 101/7$), both located in a vertex in which at least one subject showed significantly positive fMRI responses ($P < 10^{-4}$). Gray circles indicate electrodes in a vertex not showing significantly positive fMRI responses in any subject ($n = 14$). White contours are boundaries of ventral visual areas in the atlas (see labels in white text).

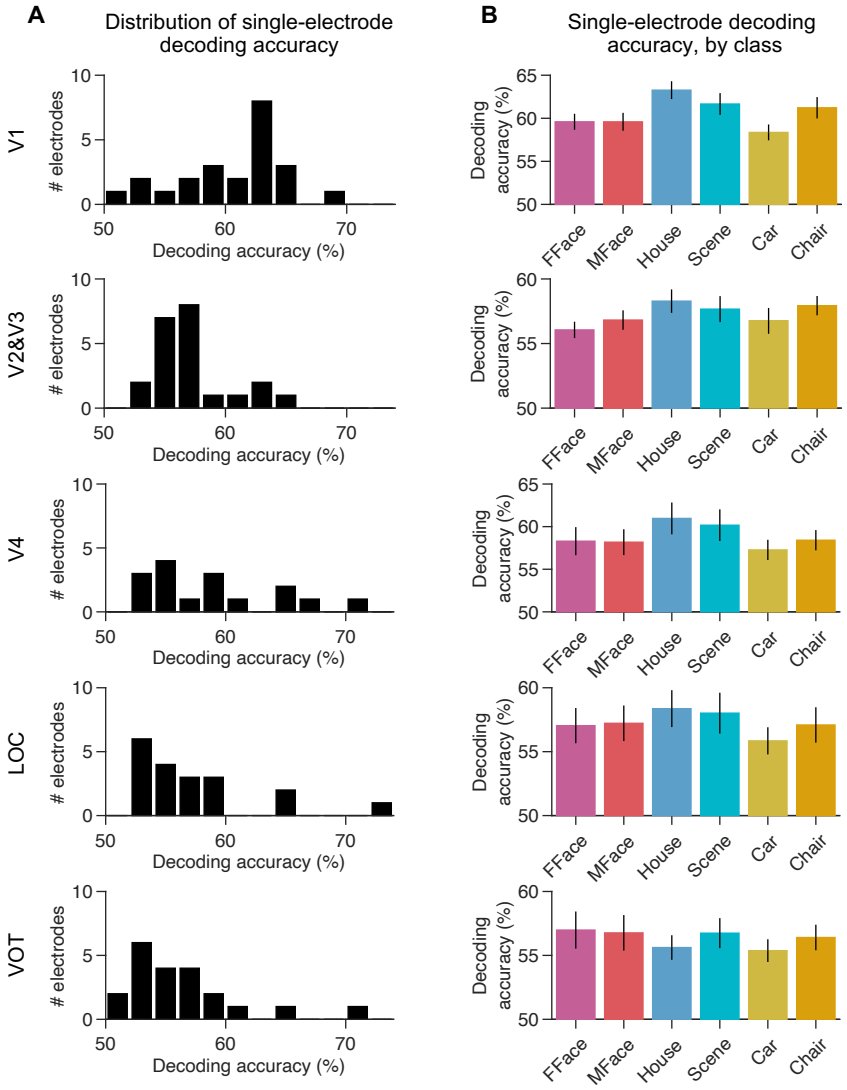

Fig. S2. Single-electrode decoding accuracy by area and class. (**A**) Distribution of single-electrode decoding accuracy by area (rows). (**B**) Mean decoding accuracy over electrodes by area and class. The accuracy of five binary classifiers was averaged for each class. The error bar indicates the standard error of mean over electrodes.

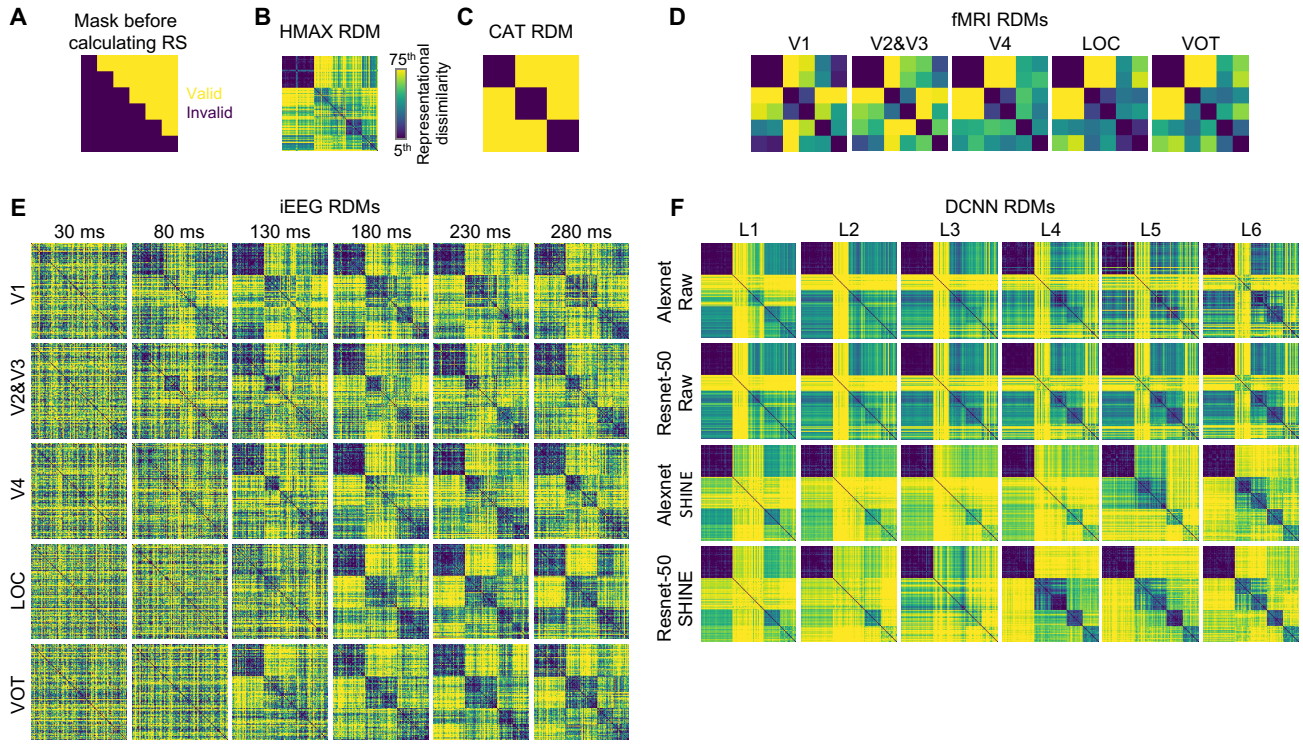

Fig. S3. Representational dissimilarity matrices of all modalities. (**A**) RDM entries used to calculate representational similarity (RS). Valid (used) and invalid (discarded) entries are indicated by color. (**B**) RDM derived from the output of HMAX-C2 units. Colormap covered the 5th to 75th percentile of dissimilarity values in the RDM. (**C**) As (B), of ideal categorical representation (CAT). (**D**) As (B), derived from fMRI responses in five ventral visual areas (columns). (**E**) As (B), derived from iEEG responses across areas (rows) and time (columns), using all trials (used to generate MDS visualization in Fig. 2A). (**F**) As (B), derived from out of layers L1 to L6 (columns) in Alexnet and Resnet-50 using raw or SHINED stimuli as input (rows). RDMs of L1 through L6 were very similar using raw stimuli, but not so much using SHINED stimuli: RS of L1 vs. L6 of Alexnet/Resnet-50 was 0.79/0.58 and 0.26/-0.02, using raw and SHINED stimuli as input, respectively. L6-SHINE was also distant from L1-raw (RS of Alexnet/Resnet-50 = 0.12/0.03) but much closer to fMRI-VOT (RS = 0.51/0.59) than was L6-SHINE (RS = 0.00/-0.15).

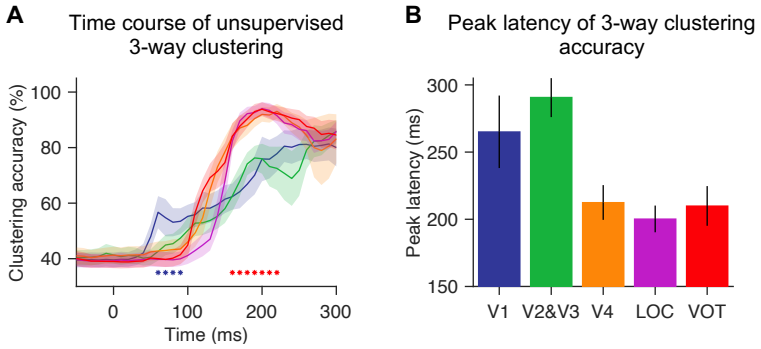

Fig. S4. Unsupervised clustering of instance-level response patterns. (**A**) Unsupervised 3-way clustering accuracy across areas and time. Higher accuracy indicates that more instances of the same category were clustered together. Solid lines and shadings indicate the mean and standard error of bootstrap distribution. Blue/red asterisks indicate when the accuracy of VOT/V1 is significantly higher ($p<0.05$). Early clustering accuracy in V1 (before 100 ms) was weak yet significantly higher than VOT (at the same level as pre-stimulus baseline), which could be explained by the stimulus-driven similarity of same-category classes (e.g., female and male faces; see fig. S3E). Cluster accuracy of LOC, VOT, and V4 increased quickly after ~100 ms, and that of V1 and V2&V3 increased slowly and caught up with higher areas after ~200 ms, as indicated by the insignificant difference between V1 and VOT. (**B**) Peak latency of clustering accuracy by area. For each round of bootstrap resampling, we determined peak latency based on the time course of clustering accuracy smoothed using a 50-ms-span moving average and then averaged the latency over bootstrap rounds. The error bar indicates standard error. Hierarchically lower areas generally had significantly longer latency ($P < 0.05$ for V1 vs. LOC, V2&V3 vs. V4/LOC/VOT).

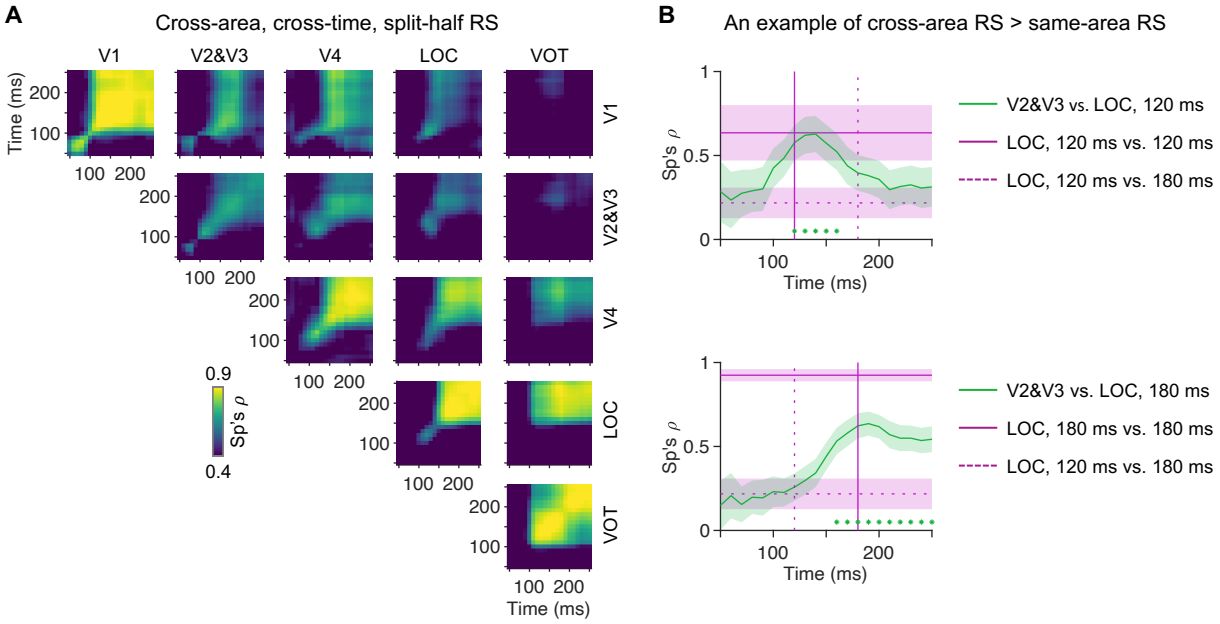

Fig. S5. Relating neural representations across areas and time. (**A**) Cross-area, cross-time RS of inter-class RDMs built using separate trials (bootstrap resampling and split-half; see Methods). X- and Y-axes indicate the time of RDMs to be compared, which could be of the same or different areas. The mean value over bootstrap distribution is shown. (**B**) An example of higher cross-area, close-time RS than same-area, distant-time RS. Top, RS of V2&V3 at 50 to 250 ms vs. LOC at 120 ms (green), compared to RS of LOC at 120 ms vs. itself (purple solid) and LOC at 180 ms (purple dotted). The latter two are drawn as flat lines as they only correspond to one time point each. Lines and shadings indicate the mean and standard error of bootstrap distribution. Green asterisks indicate when RS of V2&V3 vs. LOC at 120 ms is significantly higher than RS of LOC at 120 ms vs. LOC at 180 ms ($P < 0.05$). Vertical solid/dotted line indicates 120/180 ms, i.e., corresponding to the respective flat line. Bottom, as top, RS of V2&V3 vs. LOC at 180 ms (green), compared to RS of LOC at 180 ms vs. itself (purple solid) and LOC at 120 ms (purple dotted).

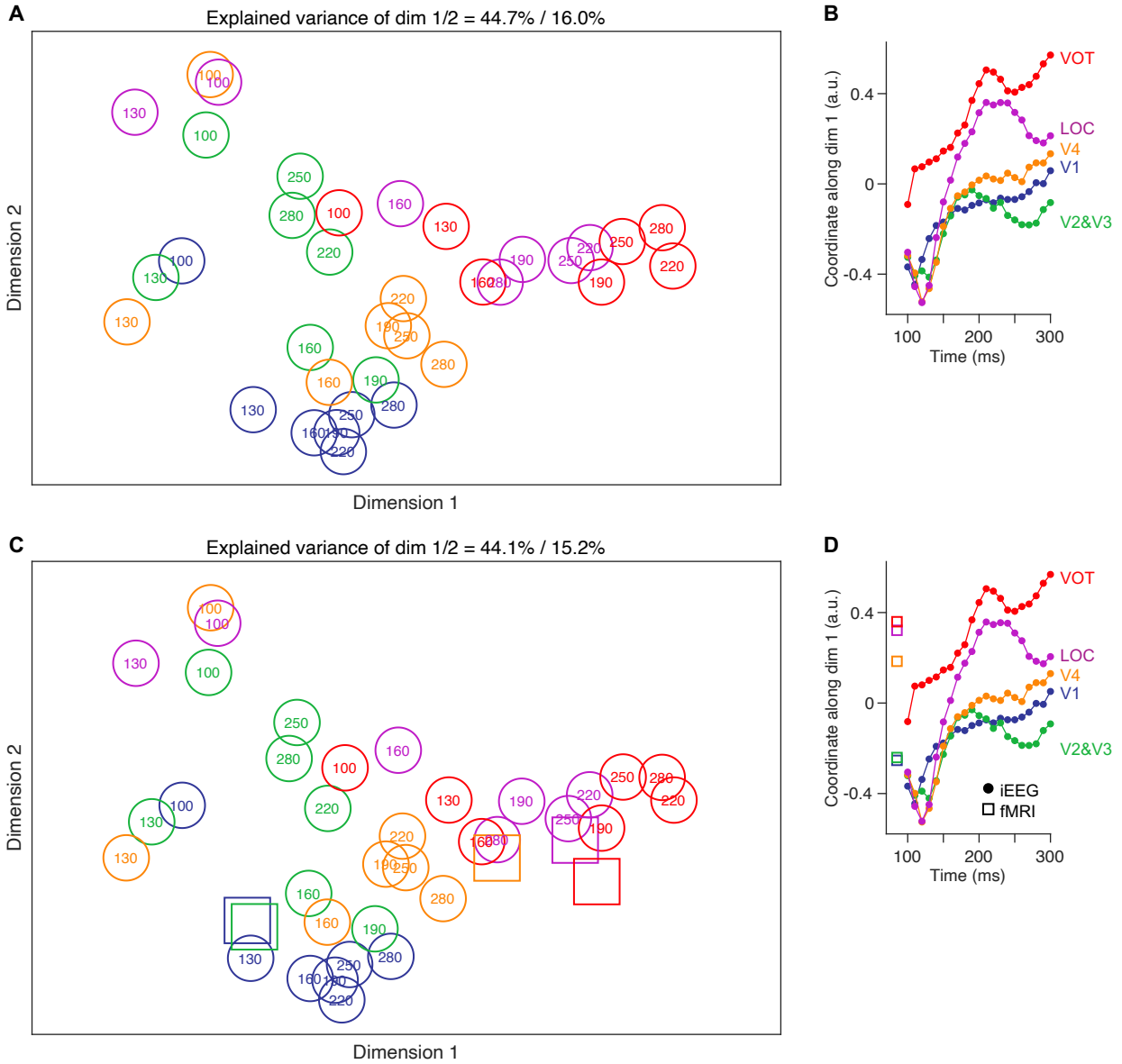

Fig. S6. Multidimensional scaling of neural representations across areas and time. (**A**) MDS visualization of iEEG RDMs across areas and time. We applied classical MDS (*cmdscale* in MATLAB) to split-half RS between inter-class RDMs in five areas at 100 to 300 ms because this algorithm reports explained variation. Coordinate along the first two MDS dimensions of RDMs at one of every three TOIs is shown to avoid clutter. The color indicates the area, and the text indicates time (ms). (**B**) Coordinate along the first MDS dimension of RDMs at all TOIs from 100 to 300 ms is shown by area. The first dimension accounted for a much larger portion (44.7%) of variation than the others. The coordinate tended to be more positive for RDMs in higher areas and at later time points. Nonetheless, distance in MDS space would better approximate the full matrix of RS (see fig. S5A) by adding more dimensions. The results suggest representational changes across OTC and lower areas were largely aligned, although they were not strictly one-dimensional. (**C**) and (**D**), as (**A**) and (**B**), results of MDS applied to both iEEG and fMRI RDMs. The pattern didn't change much as there were fewer fMRI RDMs (squares), which typically had their coordinate along the first dimension within the range of same-area iEEG RDMs, except V4.

fMRI RDM of V4 was close to same-area iEEG RDMs after ~200 ms in the MDS space, like LOC and VOC. The latency suggests recurrent processing contributes to fMRI responses in the higher ventral visual area.

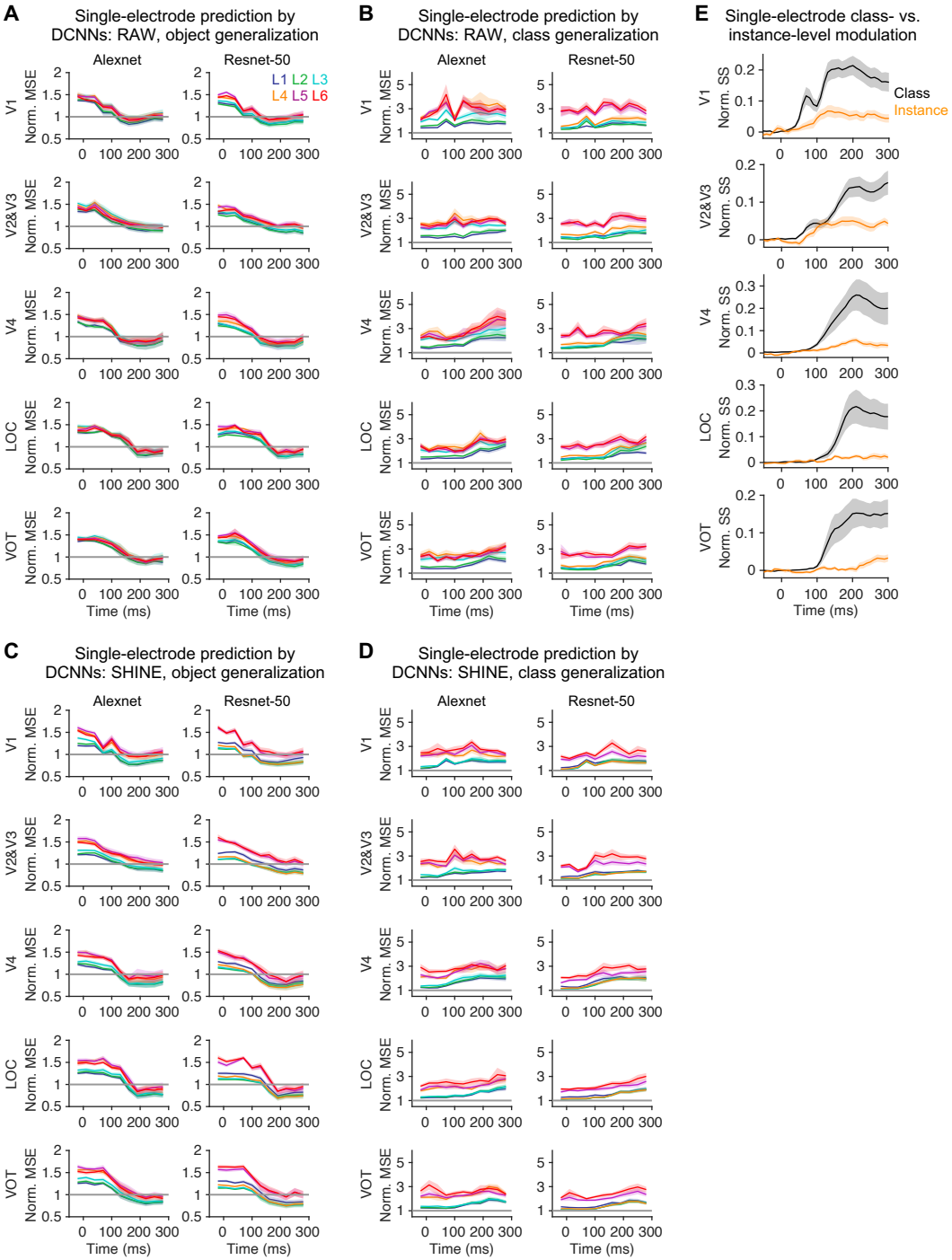

Fig. S7. Predicting single-electrode responses with DCNN output. (**A**) Performance (cross-validated normalized mean square error) of predicting single-electrode, instance-level responses with linearly mixed output of layers L1 to L6 of Alexnet (left) and Restnet-50 (right) using raw stimuli as input. We did his analysis for one of every three TOIs as it was computationally intensive. Lines and shadings indicate mean and standard error of mean over same-area electrodes. Training and test data were separate instances regardless of their class (object generalization). (**B**) As (A), the predictive performance of class generalization, in which training and test data were instances from separate classes. (**C** and **D**) As (A) and (B), respectively, predictive performance

when SHINED stimuli served as input to DCNNs. DCNN-based models passed the object generalization test, which predicted single-electrode responses better than baseline (norm. MSE < 1), i.e., a flat line through the mean, taking either raw or SHINED stimuli as input of DCNNs. However, the models failed the class generalization test (norm. MSE > 1), a stronger test of model validity proposed in previous work. (**E**) Class- and instance-level modulation of single-electrode, single-trial responses, quantified as the bias-corrected sum of squares (SS) at the class and instance levels divided by SS of trial-by-trial errors, respectively (see Methods). Classes rather than instances (of the same classes) explained single-electrode modulation much better, especially in V4, LOC, and VOT, which suggests the inherent dimensionality of 120-dim instance-level responses might be no larger than 6 (# classes). Considering the predictive models for multiple electrodes in matrix form, i.e., Y (responses; # electrodes × # instances) = beta (mixing weights; # electrodes × # units plus a constant) × X (DCNN output; # units plus a constant × # instances), the rank of both response and weight matrices (X and beta) should also be six or fewer. Finding the weights then became trivial because the DCNN output (X) should be close to full rank (120) due to high # units (1k to 800k) for all layers and both models, which implies the predictive performance may not be a proper metric of brain-DCNN correspondence. Indeed, the relative performance (object generalization) of L1 through L6 was largely the same regardless of which area's response to predict and at which time for the same model/input. The weights (beta) learned from instances of a subset of classes (class generalization) should have a lower rank than that learned from instances of all classes (object generalization), which may explain the poor predictive performance of the former that was indicative of severe overfitting during training. In a recent study[35] that reported good DCNN-based predictive performance for fMRI responses, more repetitions per image and averaging over voxels and subjects may have increased the fidelity of instance-level responses. Such a strategy could, in the future, improve the feasibility/validity of assessing brain-DCNN correspondence from the response prediction approach.