Spatial perception and memory have distinct activation profiles in human visual cortex

Serra E. Favila[1,*], **Brice A. Kuhl**[2], **and Jonathan Winawer**[1]

[1]Department of Psychology, New York University, New York, NY, 10003
[2]Department of Psychology, University of Oregon, Eugene, Oregon, 97402
[*]Contact: serra.favila@nyu.edu

ABSTRACT

Neural activity evoked during perception is thought to be reactivated during later memory retrieval. While many studies have found evidence for memory reactivation in visual cortex, few have characterized differences between stimulus-driven and mnemonic activation. Here, we leveraged population receptive field (pRF) models to quantify spatial activity in the human visual system during perception and long-term memory retrieval. We found that mnemonic activity, like perceptual activity, was precisely retinotopically mapped. However, we also observed large, systematic differences between perceptual and mnemonic activity in measures of amplitude and precision. The magnitude of these differences varied according to position in the visual hierarchy, with the largest differences observed in early areas. These differences could not be accounted for by the fact that memory data had reduced signal-to-noise or the possibility that it contained a mixture of successful and failed retrieval trials. Finally, we explore predictions from several models, showing that a simple hierarchical model with reciprocal feedforward and feedback connections accounts for many of our observations. Our results reveal novel distinctions between perceptual and mnemonic activity in visual cortex and provide insight into the computational constraints governing memory reactivation.

Introduction

Episodic memory retrieval allows humans to bring to mind the details of a previous experience. This process is hypothesized to involve reactivating sensory activity that was evoked during the initial event (James, 1890; Hebb, 1968; Damasio, 1989; McClelland et al., 1995). For example, remembering a friend's face is thought to involve reactivating neural activity that was present when seeing that face. There is considerable empirical evidence showing that visual cortical areas active during perception are also active during imagery and long-term memory retrieval (Kosslyn et al., 1995; O'Craven & Kanwisher, 2000; Wheeler et al., 2000; Slotnick et al., 2005; Polyn et al., 2005; Kuhl et al., 2011; Bosch et al., 2014; Waldhauser et al., 2016; Lee et al., 2018; Bone et al., 2018). These studies have found that activity in early visual areas like V1 reflects the low-level visual features of remembered stimuli, including spatial location and orientation (Kosslyn et al., 1995; Thirion et al., 2006; Bosch et al., 2014; Naselaris et al., 2015; Sutterer et al., 2019). Likewise, category-selective activity in high-level visual areas like FFA and PPA is observed when subjects remember or imagine faces and houses (O'Craven & Kanwisher, 2000). The strength and pattern of visual cortex activity has been associated with retrieval success in memory tasks (Kuhl et al., 2011, 2013; Gordon et al., 2014), suggesting that cortical reactivation is relevant for behavior.

These studies, and many others, have focused on identifying similarities between the neural substrates of visual perception and visual memory, and have done so successfully. However, relatively little attention has been paid to identifying and explaining differences between perception and memory. In the present work, we asked the following question: which features of stimulus-driven activity are reproduced in visual cortex during memory retrieval and which are not? The extreme possibility–that all neurons in the visual system produce identical responses during visual perception and memory retrieval of the same stimulus–can likely be rejected. Early studies of sensory cortex activation during memory retrieval showed differences between perception and memory (Wheeler et al., 2000), and perception and memory give rise to distinct subjective experiences. A more plausible proposal is that visual memory functions as a "weak" version of feedforward perception (Pearson et al., 2015; Pearson, 2019), with memory activity organized in the same fundamental way as perceptual activity, but with reduced signal-to-noise. This hypothesis is consistent with informal comparisons between perception and memory BOLD amplitudes and data suggesting that visual imagery produces similar behavioral effects to weak physical stimuli in many tasks (Ishai & Sagi, 1995; Pearson et al., 2008; Tartaglia et al., 2009; Winawer et al., 2010). A third possibility is that memory reactivation differs from stimulus-driven activation in predictable and systematic ways beyond signal-to-noise. Such differences could arise due to a change in the neural populations recruited, a change in those populations' response properties, or loss of information during sensory encoding or post-sensory processing.

One way to adjudicate between these possibilities is to make use of models from visual neuroscience that quantitatively define stimulus-triggered responses in cortex. In the spatial domain, population receptive field models (pRF) specify a simple 2D Gaussian transformation between stimulus position on the retina and the BOLD response (Dumoulin & Wandell, 2008; Wandell & Winawer, 2015). These models account for a large amount of the variance in the BOLD signal observed in human

visual cortex during perception (Kay et al., 2013b). Using these models to quantify memory-triggered activity in the visual system offers the opportunity to precisely model reactivation in visual cortex and its relationship to visual perception. In particular, the fact that pRF models are stimulus-referred may aid in interpreting differences between perception and memory activation patterns by projecting these differences onto a small number of physical dimensions.

Here, we used pRF models to characterize the properties of mnemonic activity in human visual cortex. We first trained human subjects to associate spatially localized stimuli with colored fixation cues. We then measured stimulus-triggered and memory-triggered activity in visual cortex using fMRI. Separately, we fit pRF models to independent fMRI data, which allowed us to estimate receptive field location and size within multiple visual field maps for each subject. Using pRF-based analyses, we quantified the location, amplitude, and precision of activity throughout these visual field maps during perception and memory retrieval. Finally, we explored what kind of cortical computations could account for our observations by simulating responses using a stimulus-referred pRF model and a simple hierarchical model of neocortex.

Results

Behavior

Prior to being scanned, subjects participated in a behavioral training session. During this session, subjects learned to associate four colored fixation dot cues with four stimuli. The four stimuli were unique radial frequency patterns presented at 45, 135, 225, or 315 degrees of polar angle and 2 degrees of eccentricity (Fig. 1a,b). Subjects alternated between study and test blocks (Fig. 1c). During study blocks, subjects were presented with the associations. During test blocks, subjects were presented with the cues and had to detect the associated stimulus pattern and polar angle location among similar lures (Fig. 1a,c). All subjects completed a minimum of 4 test blocks (mean = 4.33, range = 4–5), and continued the task until they reached 95% accuracy. Subjects' overall performance improved over the course of training session (Fig. 1d). In particular, subjects showed large improvements in the ability to reject similar lures from the first to the last test block (Fig. 1e).

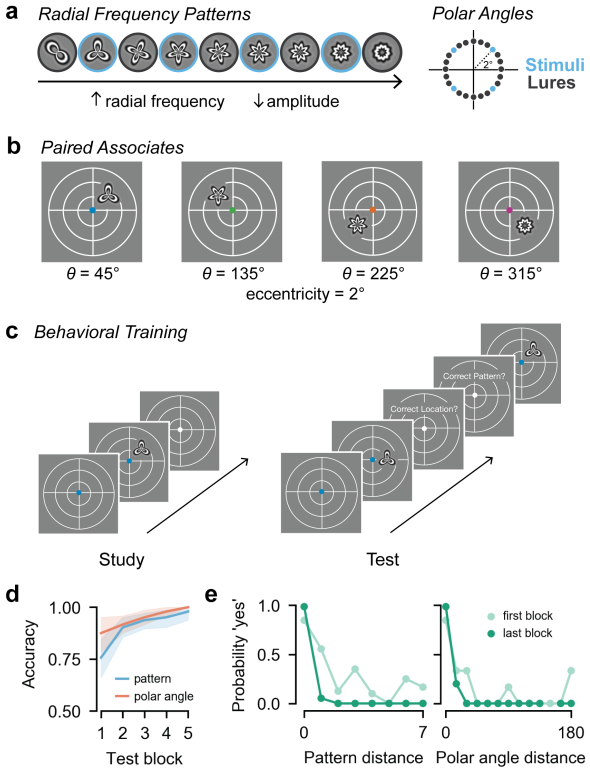

Figure 1. Stimuli and behavioral training. **(a)** The four radial frequency patterns and polar angle locations used in the fMRI experiment are outlined in blue. The intervening patterns and locations were used as lures during the behavioral training session. **(b)** Immediately prior to the scan, subjects learned that each of four colored fixation dot cues was associated with a unique radial frequency pattern that appeared at a specific location in the visual field. **(c)** Subjects alternated between study and test blocks. During study blocks, subjects were presented with the associations while maintaining central fixation. During test blocks, subjects were presented with the cues followed by test probes while maintaining central fixation. Subjects gave yes/no responses to whether the test probe was presented at the target polar angle and whether it was the target pattern. **(d)** Accuracy of pattern and polar angle responses improved over the course of the training session. Lines indicate average accuracy across subjects. Shaded region indicates 95% confidence interval. **(e)** Memory performance became more precise from the first to the last test block. During the first block, false alarms were high for stimuli similar to the target. These instances decreased by the last test block. Dots indicates probability of a 'yes' response for all trials and subjects in the first or last block. The x axis is organized such that zero corresponds to the target and increasing values correspond to lures more dissimilar to the target.

After subjects completed the behavioral training session, we collected fMRI data while subjects viewed and recalled the stimuli (Fig. 2a). During fMRI perception runs, subjects fixated on the central fixation dot cues and viewed the four stimuli in their learned spatial locations. Subjects performed a one-back task to encourage covert attention to the stimuli. Subjects were highly accurate at detecting repeated stimuli (mean = 86.8%, range = 79.4%–93.2%). During fMRI memory runs, subjects fixated on the central fixation dot cues and recalled the associated stimuli in their spatial positions. On each trial, subjects made a judgment about the subjective vividness of their memory. Subjects reported that they experienced vivid memory on an average of 89.8% of trials (range: 72.4%–99.5%), weak memory on 8.9% of trials (0.5%–25.0%), and no memory on 1.2% of trials (0.5%–2.6%).

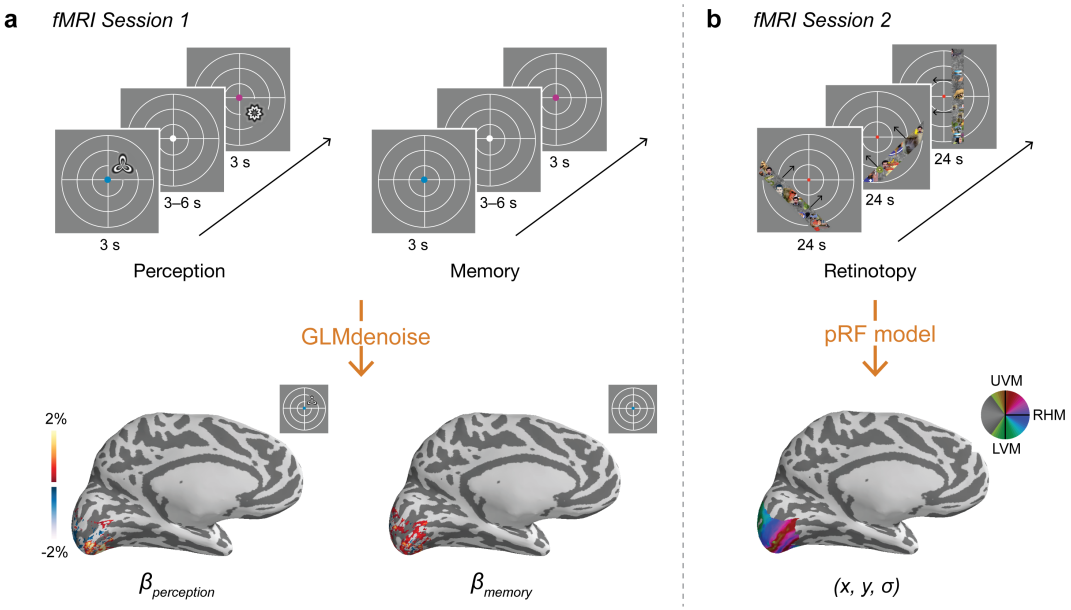

Figure 2. fMRI task design and measurements. **(a)** Following training, subjects participated in two tasks while being scanned. During perception runs, subjects viewed the colored fixation dot cues and associated stimuli while maintaining central fixation. Subjects performed a one-back task on the stimuli to enforce covert attention to each stimulus. During memory runs, subjects viewed only the cues and recalled the associated stimuli while maintaining central fixation. Subjects made a judgment about the vividness of their memory (vivid, weak, no memory) on each trial. We used the perception and memory fMRI timeseries to perform a GLM analysis that estimated the response evoked by perceiving and remembering each stimulus for each vertex on the cortical surface. Responses in visual cortex for an example subject and stimulus are shown at bottom. **(b)** In a separate fMRI session on a different day, subjects participated in a retinotopic mapping session. During retinotopy runs, subjects viewed bar apertures embedded with faces, scenes, and objects drifting across the visual field while they maintained central fixation. Subjects performed a color change detection task on the fixation dot. We used the retinotopy fMRI timeseries to solve a pRF model that estimated the receptive field parameters for each vertex on the cortical surface. A polar angle map is plotted for an example subject at bottom.

Memory reactivation is spatially organized

We used a GLM to estimate the BOLD response evoked by seeing and remembering each of the four spatially localized stimuli (Fig. 2a; see Methods). Separately, each subject participated in a retinotopic mapping session. We fit pRF models to these data to estimate pRF locations (x, y) and sizes (σ) in multiple visual areas (Fig 2b). To more easily compare evoked responses across visual areas, we transformed these responses from cortical surface coordinates into visual field coordinates using the pRF parameters. For each subject, ROI, and stimulus, we plotted the amplitude of the evoked response at the visual field position (x, y) estimated by the pRF model (Fig. 3a). We then interpolated these values over 2D space, z-scored the values, rotated all stimuli to the same polar angle, and averaged across stimuli and subjects (see Methods). These representations are useful for comparison across regions because they show the organization of the BOLD response in a common space that is undistorted by the size and magnification differences present in cortex.

We generated these visual field representations for V1, V2, and V3, separately for perception and memory measurements. Readily apparent is the fact that stimulus-evoked responses during perception were robust and spatially-specific (Fig. 3b, top). The spatial spread of perceptual responses increased from V1 to V3, consistent with estimates of increasing receptive field size in these regions (Wandell & Winawer, 2015; Kay et al., 2013b), While the memory responses were weaker and more diffuse, they were also spatially organized, with peak activity in the same location as the perception responses (Fig. 3b, bottom).

We next aimed to quantify these initial observations. Because our stimulus locations were isoeccentric, we reduced our responses to variance along one spatial dimension: polar angle. To do this, we restricted our ROIs to surface vertices with pRF locations near the stimulus eccentricity (2°), rotated stimuli to a common polar angle, normalized the responses, and averaged across stimuli and subjects. We then plotted the group average BOLD response in bins of polar angle distance from the stimulus

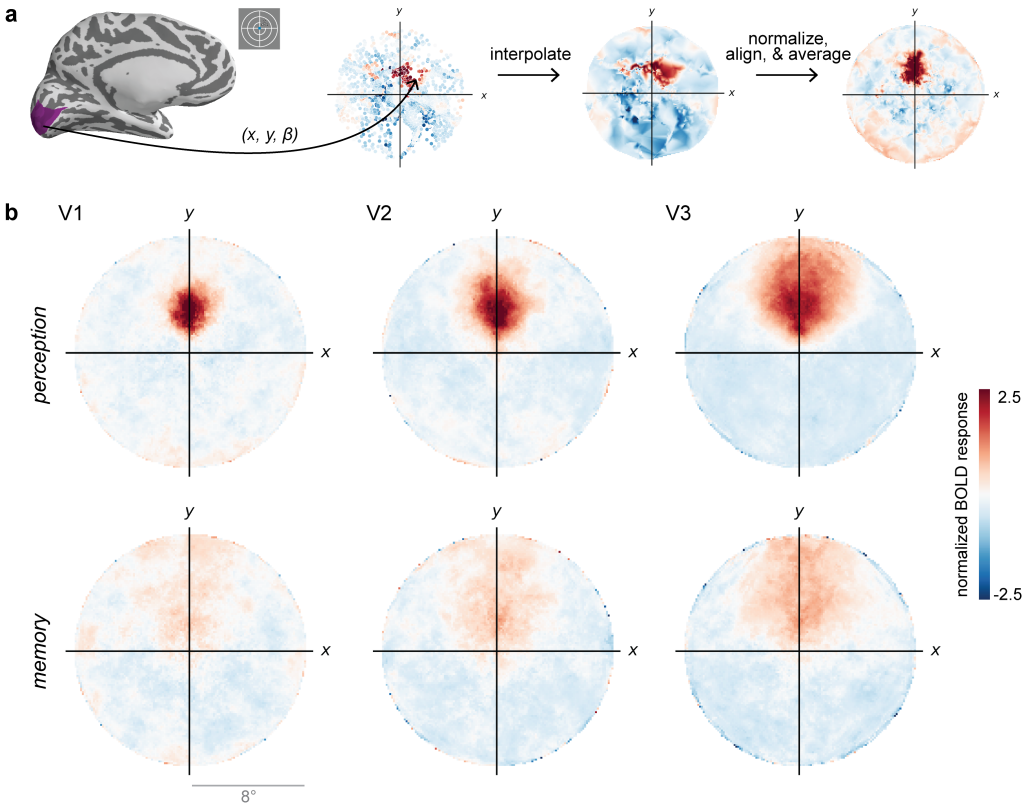

Figure 3. Perception and memory activity in visual field coordinates. (a) For a given subject, ROI, and stimulus, we plotted the perception or memory evoked response (β) in the visual field position estimated by the pRF model (x, y). We then interpolated over 2D space and z-scored the responses. We rotated these representations by the polar angle location of the stimulus so that they aligned on the upper vertical meridian, and then averaged over stimuli and subjects. This procedure produces an average activation map in visual field coordinates for each ROI. **(b)** Plots of perception-evoked and memory-evoked activity, averaged across all subjects and stimuli, from V1, V2, and V3. These plots suggest that perceptual activity is not perfectly reproduced during memory retrieval but that some retinotopic organization is preserved.

(Fig. 4a; see Methods). We generated these polar angle response functions for V1–V3 and for three mid-level visual areas: hV4, LO, and V3ab (Fig. 4b). To capture the pattern of positive and negative responses we observed, we fit the average data with a difference of two von Mises distributions, where both the positive and the negative von Mises were centered at the same location. Visualizing the data and the von Mises fits (Fig. 4b), it's clear that both perception and memory fits show a peak at 0°, or the true location of the stimulus, in every region.

To formalize this pattern, we calculated bootstrapped 95% confidence intervals for the location parameter of the von Mises distributions by resampling subjects with replacement (see Methods). We then compared the accuracy and reliability of location parameters between perception and memory (Fig. 4c, left). As expected, location parameters derived from perception data were highly accurate. Confidence intervals for perception location parameters overlapped 0°of polar angle, or the true stimulus location, for all but two ROIs. Even in those two ROIs (V3 and LO), the confidence intervals reached within a degree of 0. Perception location parameters were not only accurate (close to 0) but also extremely reliable (small error bars). Perception 95% confidence intervals spanned only 5.7°on average (range: 4.1°–6.9°;), demonstrating that there was little variability in this measure across subjects. Critically, memory parameters and confidence intervals looked similar to perception in almost all respects. Memory parameters were also highly accurate, with confidence intervals overlapping 0°in every ROI (Fig. 4c, left). Thus, in every visual area measured, the spatial locations of the remembered stimuli could be accurately estimated from mnemonic activity. Memory confidence intervals were somewhat less reliable than perception confidence intervals (mean = 13.7°, range = 7.9°–19.3°), suggesting lower signal-to-noise. However, even the widest memory confidence interval spanned only 19.3°. This is far less than the 90°separating each stimulus location, indicating no confusability between stimuli in memory activity. Because both perception and memory location parameters were highly accurate, and because differences in reliability were relatively small, perception and memory confidence intervals were highly overlapping in every visual area (Fig. 4c, left), indicating a high level of agreement in the estimated location of peak activity during perception and memory. These results provide strong evidence that memory-triggered activity in human visual cortex is spatially organized within known visual field maps, as it is during visual perception. These findings support prior reports of retinotopic activity during memory and imagery (Kosslyn et al., 1995; Slotnick et al., 2005; Thirion et al., 2006), but provide more quantitative estimates of this effect.

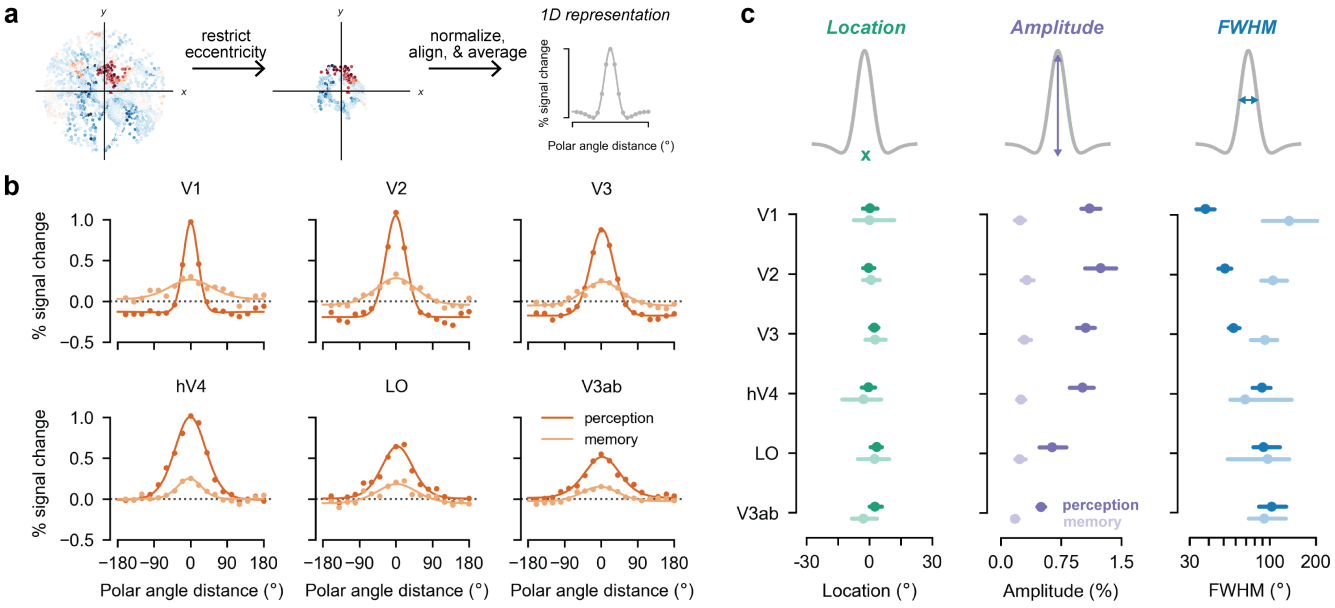

Figure 4. Perception and memory have shared and distinct activation features. (a) We created 1D polar angle response functions by restricting data to eccentricities near the stimulus, aligning stimuli to a common polar angle, and averaging responses into polar angle distance bins. A difference of two von Mises distributions was fit to the group average response. Responses in cortical locations that represent the stimulus position are plotted at x = 0. **(b)** Polar angle response functions, averaged across all subjects and stimuli, are plotted separately for perception and memory. Dots represent average data across all stimuli and subjects and lines represent the fit of the difference of two von Mises distributions to the average data. The peak response during both perception and memory retrieval occurs at cortical locations that represent the part of the visual field near the stimulus. However, there are clear differences in the amplitude and width of the responses across ROIs and between perception and memory. **(c)** Bootstrapped 95% confidence intervals for the location, amplitude and FWHM of the difference of von Mises fits are plotted to quantify these trends. In all ROIs, the peak location of the response is equivalent during perception and memory (at 0°, the stimulus location), while the amplitude of the response is reliably lower during memory than during perception. The FWHM of the response increases across ROIs during perception but not during memory, resulting in highly divergent FWHM for perception and memory in early visual areas.

Amplitude and precision differ between perceptual and mnemonic activity

Aspects of perception and memory responses other than the peak location differed considerably. First, memory responses were lower in amplitude than perception responses, with the largest difference in early visual areas (Fig. 4b). To quantify these observations, we derived a measure of amplitude from the difference of von Mises functions fit to our data (see Methods). We also computed 95% bootstrapped confidence intervals for this amplitude metric, following the prior analysis. We then compared these estimates between perception and memory. First, response amplitudes for memory data were lower than the perception amplitudes in all visual areas (Fig. 4c, middle). 95% confidence intervals for perception and memory did not overlap in any ROI, indicating that these differences were highly significant. The average ratio between perception and memory amplitudes ranged from 2.7 in LO and V3ab to 4.6 in V1, suggesting that the magnitude of this effect varied according to position in the visual hierarchy (Fig. 4c, middle). Critically, this does not imply that memory responses were at baseline: amplitude confidence intervals for memory data did not overlap with zero in any region (Fig. 4c, middle), demonstrating that responses were significantly above baseline in all areas measured. These results demonstrate that the amplitude of spatially-organized activity in visual cortex is largely attenuated (but present) during memory, with the strongest attenuation in early visual areas.

Second, memory responses were wider than perception responses, again, with the largest differences in early visual areas (Fig. 4b). We quantified the precision of perception and memory responses by computing the full width at half maximum (FWHM) of the difference of von Mises fits to our data and by generating 95% confidence intervals for this measure. Perception FHWM tended to increase (decreased precision) moving up the visual hierarchy (Fig. 4c, right). During perception, V1 had the narrowest (most precise) responses: 38.0°(95% CI: 33.0°–43.5°). These responses grew steadily wider through V3ab, which had the widest responses during perception at 103.0°(95% CI: 85.0°–128.0°). This increasing pattern follows previously described increases in population receptive field size in these regions (Wandell & Winawer, 2015; Kay et al., 2013b). Strikingly, this pattern was not preserved during memory. Memory confidence intervals for all regions overlapped with one another, meaning that every ROI showed equally wide memory responses within the precision of our measurements. Numerically, the trend we observed during perception was reversed during memory, with the widest responses observed in the earliest areas (FWHM: V1>V2>V3>V4). As a consequence of these trends, perception and memory FWHM were highly divergent in V1, V2, and V3 but equivalent in later areas (Fig. 4c, right). Compared to to the narrow responses we observed in V1 during perception, V1 responses during memory were significantly wider: 134.0°(95% CI: 90.0°–204.0°). In fact, V1 responses during memory were statistically equivalent to V3ab responses during perception (95% CI: 85.0°–128.0°). In V2 and V3, memory

FWHM exceeded perception FWHM by an average of 54.0°and 35.0°, respectively, with no overlap between perception and memory confidence intervals in either region. In hV4, LO, and V3ab, perception and memory confidence intervals were highly overlapping and thus statistically equivalent. In summary, we observed reliable and striking differences in the precision of perception and memory responses. These results push back on the idea that memory is a noisy copy of perception, and show that there are reliable and systematic differences in spatial tuning between perception and memory.

Differences between perception and memory are not explained by data quality

An important consideration is whether differences in data quality could manifest as apparent differences in precision. For example, is it possible that memory and perception were actually equivalent, but due to greater trial-to-trial variability, memory fits were wider? We addressed this in two ways. First, we used simulations to test the possibility that lower signal-to-noise automatically yields the pattern of FWHM that we observed during memory. Using bootstrapped parameter estimates, we confirmed that the estimated signal-to-noise ratio (SNR) for perception parameter estimates was in fact higher than for memory parameter estimates in every region. Perception SNR was between 1.2 and 1.6 times higher than memory SNR across ROIs. We then simulated new parameter estimates with the median signal we observed during perception, but with noise levels that result in *lower* SNR than what we observed during memory (see Methods). We then analyzed 100 of these simulated datasets with the identical procedure used to analyze the actual perception and memory data and examined the von Mises fits. As expected, simulating parameter estimates with added noise produced variance in the location, amplitude, and FWHM of the von Mises fits (Fig. 5a). However, it is clear that for V1, V2, and V3, adding noise did not sufficiently widen the response to match the memory data. To quantify the similarity between the low SNR simulation and the memory data, we counted the number of simulations that generated FWHM parameters within the FWHM confidence interval of the actual memory data. In V1, V2, and V3–the regions where we observed *different* FWHM for perception and memory— 0/100 low SNR simulations met this criteria (Fig. 5c). For the higher visual areas–hV4, LO, and V3ab– most of the simulations (greater than 75%) produced FWHM values within the memory confidence intervals (Fig. 5c). However, these are the regions where we observed equivalent FWHM during perception and memory in the first place, so this is not surprising. Thus, while it is certainly true that memory data had lower SNR than perception data, it is unlikely that low SNR by itself yields the pattern of memory FWHM we observed in early visual areas.

Next, we addressed the possibility that differences between perception and memory were due to the occurrence of failed retrieval trials. One might hypothesize that memory data reflect the combination of two trial types: successful retrieval, where responses are equivalent to perception, and failed retrieval, where responses do not reliably exceed baseline. Under this hypothesis, there is no true underlying difference between perception and memory, and any observed differences (such as decreased precision) are attributable to the inclusion of failed retrieval trials. To address this possibility, we simulated new parameter estimates where 25%-75% of trials had a mean response of zero (failed retrieval), and the remaining trials had a mean response equivalent to perception. We examined difference of von Mises fits for 100 different simulated data sets, for each of 25%, 50%, and 75% failed retrieval conditions (Fig. 5b). Under these assumptions, failed retrieval must occur in approximately 75% of trials to approximate our data qualitatively. This level of failed retrieval is inconsistent with subjects' vividness reports. More importantly, the 75% failed retrieval simulation fails to fully account for the data when examined quantitatively. In >75% simulations in V1, V2, and V3, the FWHM was outside the confidence interval of the actual memory data. Moreover, even in those simulations in which the estimated FWHM was within the confidence interval of the memory data (V1: 17/100, V2: 16/100, V3: 21/100; Fig. 5c), these same simulations failed to produce similarly plausible amplitude and location parameters. In all visual areas except LO, 0/100 simulations fell within the memory confidence intervals for all three parameters (Fig. 5c). In LO, only 2/100 simulations fell within the confidence intervals for all three parameters. Thus, the presence of failed retrieval trials cannot by itself account for our full set of observations.

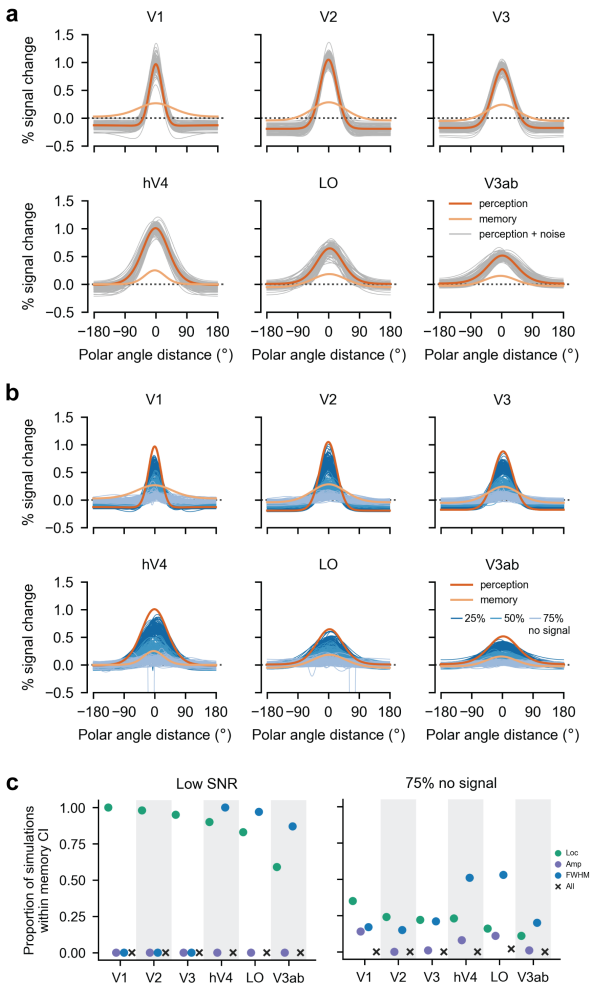

Figure 5. Differences between perception and memory are not explained by low signal-to-noise or failed retrieval (a) Gray lines represent 100 simulated perception data sets with noise levels that yield lower SNR than observed in our memory data. Orange lines represent the fits to perception and memory data, reproduced from Fig. 4b. This simulation does not produce sufficiently wide responses in V1, V2, and V3 to account for our memory results. **(b)** Blue lines represent 100 simulated data sets where data are a combination of perception signals and 25%, 50%, or 75% failed retrieval (no signal) trials. Even a very large number of failed retrieval trials (75%) does not reproduce memory data in all regions. **(c)** Proportion of SNR simulations and 75% no signal simulations that produce parameters within the confidence intervals of the memory data. Colored dots represent the proportion of 100 simulations that meet this criteria for a given parameter, and black X's represent the proportion of 100 simulations that meet this criteria for all parameters. The failure of these simulations to produce reasonable parameters in all ROIs suggests that SNR and failed retrieval are unlikely explanations for our pattern of results.

pRF models accurately predict perception but not memory responses

What cortical computations can explain our empirical results? pRF models incorporate much of what's known about the spatial computations employed in visual cortex, including center-surround receptive fields and nonlinear summation. Thus, we first compared pRF model predictions derived from independent data to the stimulus-driven and memory-driven responses we measured during the main experiment. To do this, we used the pRF parameters solved with a drifting bar stimulus (Fig. 2b) to generate predicted responses to the experimental stimuli for each subject (Fig. 6a). The pRF model we used is a novel variant of existing pRF models: we used a Difference of Gaussian pRF shape (Zuiderbaan et al., 2012) combined with a compressive spatial summation output nonlinearity, or CSS model (Kay et al., 2013b). The predictions from the model were analyzed with the same procedure as the data, yielding von Mises fits to the predicted data (Fig. 6b). Model predictions from simpler pRF models are shown in Supplemental Figure 1.

Qualitatively, the model predictions agree well with the perception data but not the memory data across multiple ROIs (Fig. 6b. Several specific features of the perception data are well captured by the model. First, the model predicts negative responses in the surround locations of V1-V3 but not higher visual areas. This is particularly interesting given that all pRFs had a negative surround, suggesting that voxel-level parameters and population-level responses can diverge. Second, the model predicts increasingly wide response profiles from the early to late visual areas. And third, it predicts higher amplitudes in early compared to late areas. All of these patterns were observed in the perception data and none in the memory data. These patterns are especially clear when comparing the amplitudes and FWHMs of fitted von Mises distributions (Fig. 6c). While

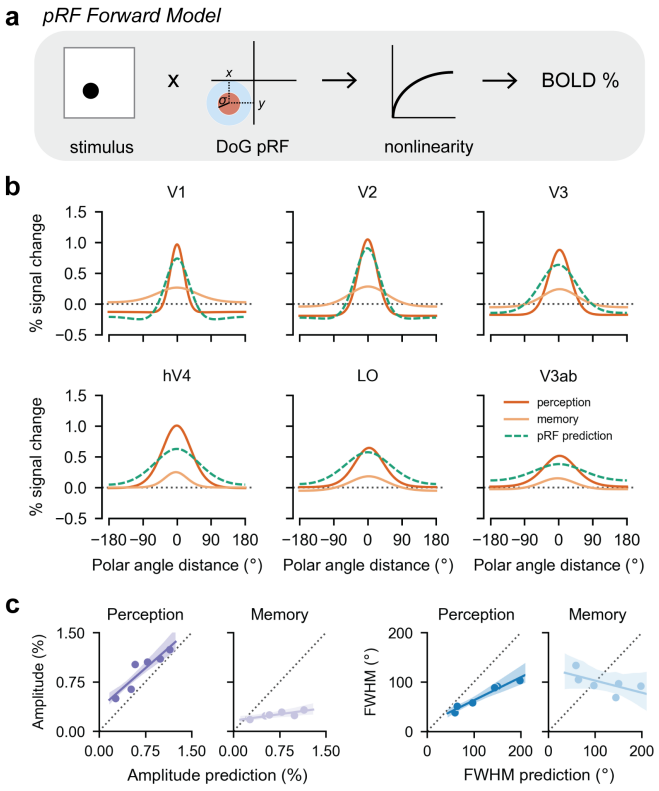

Figure 6. pRF forward model captures perception but not memory responses. (a) We used our pRF model to generate the predicted BOLD response to each of our experimental stimuli. The model assumes a Difference of Gaussians pRF shape, with a fixed positive to negative Gaussian size ratio (1:2) and amplitude (2:1). The model also incorporates a compressive nonlinearity. **(b)** Predicted polar angle response functions are plotted for the pRF model (green dashed lines), as well as the perception and memory data (dark and light orange, reproduced from Fig. 4b). The model predictions are closer to the perception data than the memory data in all visual areas. **(c)** Predicted versus observed amplitude (left) and FWHM (right), plotted separately for perception and memory. Each dot represents an ROI. The shaded region is the 95% CI from bootstrapping linear fits across participants. For both the amplitude and FWHM, the perception data lie relatively close to the pRF model predictions (dashed black lines), whereas the memory data do not.

there is a positive correlation between the model amplitude and both the perception and memory amplitudes, the slopes are very different: the perception amplitudes have a slope close to 1, indicating good agreement with the model predictions, while the memory data have a slope close to 0. Similarly, the FWHM of the model response functions are positively correlated with the perception FWHM, but only weakly and negatively correlated with the memory FWHM. Overall, the model predictions are accurate but not perfect for the perception data. The model predicts slightly lower amplitudes and larger FWHM than is observed in the perception data. These discrepancies may be due to differences between the stimuli used in the main experiment and those used in the pRF experiment, or to differences in the task (attending fixation during the pRF experiment vs attending the stimulus during the main experiment). However, the model predictions are highly inaccurate for memory data. Because pRF models incorporate many of the known properties of spatial processing in visual cortex, these analyses suggest that a different computational model is needed to account for memory.

Perception and memory responses can be simulated with a bidirectional hierarchical model

Why do the memory responses we observed differ systematically from stimulus-driven responses during perception? Responses during perception are likely to arise from a hierarchical, feedforward process with additional spatial pooling at each stage, resulting in increasingly large receptive fields. Memory reinstatement is hypothesized to begin with the hippocampus, a region bidirectionally connected to high-level visual areas in ventral temporal cortex (Felleman & Essen, 1991). Reinstated activity is then thought to propagate backwards through visual cortex (Linde-Domingo et al., 2019; Dijkstra et al., 2019). Here, we explored whether a simple hierarchical model could be adapted to account for both our perception and memory results.

We first constructed a simple feedforward hierarchical model of spatial processing in neocortex. In this model, the activity in each layer was created by convolving the activity from the previous layer with a fixed Gaussian kernel (Fig. 7a; see Methods). Beginning with a simple square wave stimulus, we cascaded this convolutional operation to simulate 8 layers of the network (Fig. 7b). The pattern of feedforward responses qualitatively matches our fMRI observations during perception. The location of the peak response is unchanged across layers, but response functions become wider and lower in amplitude in higher layers (Fig. 7c).

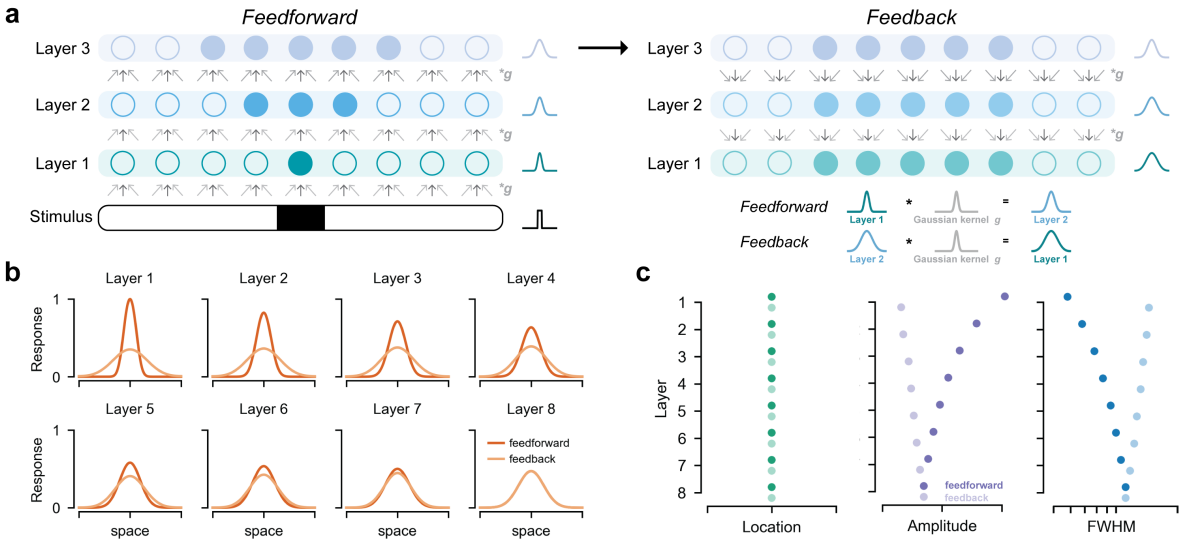

Figure 7. Perception and memory responses can be simulated with a bidirectional hierarchical model. **(a)** Illustration of stimulus-driven activity propagating through a hierarchical network in the feedforward direction (left) and mnemonic activity propagating through the network in the feedback direction (right). In both cases, a given layer's activity is generated by convolving the previously active layer's activity with a fixed Gaussian kernel. Feedforward simulation began with a square wave stimulus. Feedback simulation began with duplication of the feedforward activity from the final layer. **(b)** Results from feedforward and feedback simulations in an 8 layer network, plotted in the conventions of Figure 4b. The feedforward simulation parallels our observations during perception, and the feedback simulation parallels our observations during memory. **(c)** Location, amplitude, and FWHM parameters for each layer, plotted separately for feedforward and feedback simulations. Location is preserved across layers in the feedforward and feedback direction. Amplitude and FWHM become progressively smaller and wider, respectively, in higher layers in the feedforward direction and in lower layers in the feedback direction. This results in large differences between feedforward and feedback activity in early layers. These trends closely follow our observations in Figure 4c.

We then explored whether backwards propagation of reinstated activity in our hierarchical model could account for our memory data. To do this, we assumed that feedforward and feedback connections in the model were symmetrical, meaning that the convolutional kernel was the same in feedforward and feedback direction. We assumed perfect initial reinstatement, and thus began the feedback simulation by duplicating the feedforward activity from the final layer. Starting with this final layer activity, we convolved each layer's activity with the same Gaussian kernel to generate earlier layers' activity (Fig. 7a). The simulated responses (Fig. 7b) capture several critical aspects of our observed memory data. First, simulated feedback responses were wider and lower amplitude than feedforward responses overall (Fig. 7c). Second, the change in FWHM and amplitude across layers was much smaller in the feedback direction than in the feedforward direction. Third, the difference between feedforward and feedback responses was maximal in the earliest layers. This simulation suggests that the distinct spatial profile of mnemonic responses in visual cortex may be a straightforward consequence of reciprocal connectivity in the visual system, and that spatial pooling accumulated during feedforward processing may not be inverted during reinstatement. More broadly, these results demonstrate that models of the visual system may be useful for probing the mechanisms that support visual memory.

Discussion

In the current work, we combined empirical and modeling approaches to explore how long-term memories are represented in the human visual system. By using computational models of spatial encoding to characterize memory responses, we provide strong evidence that visual memories produce retinotopically-mapped activation in visual cortex. Just as important, we also identified several systematic differences between perceptual and mnemonic responses: mnemonic responses were lower in amplitude, less spatially precise, and varied little between areas. Differences in spatial tuning between perception and memory were most pronounced in the earliest visual areas. These results are at odds with the view that memory is best understood as a weak version of perception. Although we did confirm a lower signal-to-noise ratio during memory, simulations showed that neither reduced SNR nor failure to retrieve on a subset of trials could account for the observed differences. We speculate, instead, that these differences arise from reciprocal connectivity and hierarchical organization within visual cortex. To support this, we show that top-down activation in a simple hierarchical model elicits a systematically different pattern of responses than bottom-up activation. These simulations reproduce the properties we observe during both perception and memory. Together, these results reveal novel properties of mnemonic responses in visual cortex and advance our understanding of the computational processes that constrain memory reactivation.

Using spatial encoding models to quantify memory representations

Much work in neuroscience has been dedicated to the question of how internally generated stimulus representations are encoded in the brain. Early neuroimaging work established that sensory cortices are recruited during imagery and memory tasks (Kosslyn et al., 1995; O'Craven & Kanwisher, 2000; Wheeler et al., 2000), moving the field away from purely symbolic accounts of memory (e.g. Pylyshyn, 2002). More recently, memory researchers have favored decoding and pattern similarity approaches over univariate activation analyses to examine the content of retrieved memories (Polyn et al., 2005; Kuhl et al., 2011; Favila et al., 2018). While these approaches are powerful, they do not explicitly specify the form mnemonic activity should take, and many activation schemes can lead to successful decoding. In the present work, we leveraged models of spatial encoding from visual neuroscience, including stimulus-referred pRF models and hierarchical models, to examine and account for memory-triggered activity in visual cortex. Our results confirm earlier observations that memory and imagery produce retinotopic activation (Kosslyn et al., 1995; Slotnick et al., 2005; Thirion et al., 2006), while more precisely quantifying this effect. Moreover, our approach revealed novel properties of memory responses in visual cortex that decoding approaches have missed. For example, we found that memory activity was characterized by a different pattern of spatial pooling across regions than perceptual activity. Because spatial pooling is explicitly modeled in pRF and hierarchical models, we were able to quantify these differences and explore the computations that may underlie these observations.

Stimulus-referred and network models of neuronal responses

Population receptive field (pRF) models formalize known properties of spatial processing in human visual cortex, including center-surround receptive field structure, change in receptive field size across eccentricity and between visual areas, and nonlinear spatial summation. The pRF models used in this paper accurately predict the BOLD response we measured during perception, but not during memory. What accounts for this difference? One approach is to fit a separate set of pRF parameters for memory by having subjects imagine spatially-varying stimuli, as has been done previously for imagery pRFs (Breedlove et al., 2018) and imagery retinotopic maps (Slotnick et al., 2005). The advantage of such methods is that they allow for comparison of the same parameters (e.g. retinotopic position or pRF size) in different conditions, such as memory vs perception. But when differences are found, these models do not provide an explanation as to why they differ. For this reason, we turned to a network model.

Hierarchical network models can make very similar predictions to stimulus-referred pRF models (Kay et al., 2013b), but also confer additional advantages. While pRF models express each region's activity as a function of the stimulus, a hierarchical network model expresses each region's activity as a function of the previous region's activity (Fukushima, 1980; Riesenhuber & Poggio, 1999), and can therefore incorporate both feedforward and feedback processes. This form of model may be particularly well-suited to comparing perception and memory because the difference between memory and perception responses may be a direct consequence of the direction of processing. Studies of anatomical connectivity provide evidence that the visual system is organized hierarchically (Felleman & Essen, 1991; Barone et al., 2000). These studies show that the hippocampus is at the highest stage of the hierarchy, receiving input from the highest visual regions, and that most connections within the visual system are reciprocal. These observations make the prediction that initial drive from the hippocampus during memory retrieval should propagate backwards through the visual system. Electrophysiological recordings from the macaque (Naya et al., 2001) and human (Linde-Domingo et al., 2019; Dijkstra et al., 2019), as well as computational modeling (Horikawa & Kamitani, 2017) support this idea.

We simulated such a process in a hierarchical network. While highly simplified, this simulation captured the dominant features of our data, providing a possible explanation for our observations. Critically, the early layers of our model showed a low amplitude and widespread response profile when triggered by memory but not by perception. This is because the model input to the memory simulation was the result of the perception simulation, and had thus already been blurred. Second, the change across layers during the memory simulation was small compared to the perception simulation. This is because the convolutional kernel was fixed, and as the input layer grew wider, the amount of additional blurring caused by convolution decreased. These findings suggest that some properties of memory-driven activation in visual cortex may be the straightforward consequence of reversing computations that occur in the feedforward direction. These findings may also be related to findings in which memory-guided behavior is less precise than stimulus-guided behavior (Kosslyn, 1980). However, this does not imply that there are no circumstances in which memory-triggered responses can have a high degree of precision in early visual areas. Other network properties may be able to explain such cases, such as a recurrent model with explicit representations of priors and state-dependent tuning (Heeger, 2017).

Implications for theories of memory reactivation

Our results have important implications for existing theories of memory reactivation. Most previous work has focused on identifying *similarities* between the neural substrates of visual perception and visual memory. These studies have been successful in that they have produced many positive findings of memory reactivation in human visual cortex (Kosslyn et al., 1995; O'Craven & Kanwisher, 2000; Wheeler et al., 2000; Slotnick et al., 2005; Polyn et al., 2005; Kuhl et al., 2011; Bosch et al., 2014; Waldhauser et al., 2016; Lee et al., 2018; Bone et al., 2018). However, much of this work implicitly assumes that

any mismatch between perception and memory is due to the fact that memory reactivation is either inherently low fidelity or susceptible to noise (Pearson et al., 2015), or is a subset of the perceptual response (Wheeler et al., 2000). Here, we confirm prior findings of reactivation by showing spatially precise retinotopic activation in multiple visual areas, as well as reduced SNR during memory. We also show clear evidence that perception and memory responses are most similar in higher level visual areas as proposed by Pearson et al. (2015) and paralleling prior findings (Favila et al., 2018; Breedlove et al., 2018). However, we also found systematic differences that could not be explained by lower SNR. This indicates that additional theories are needed to interpret and predict memory reactivation.

We focused on the domain of visual spatial processing for a few reasons. First, a great deal of research has focused on the role of space in memory (Eichenbaum et al., 1999). Second, there are detailed, well-validated models of spatial encoding in the human visual system (Kay et al., 2008; Wandell & Winawer, 2011, 2015; Graham, 1989). Finally, spatial location is coded in the human brain at a scale that is well-matched to the millimeter sampling resolution of fMRI (Horton & Hoyt, 1991; Engel et al., 1994; Sereno et al., 1995; Dougherty et al., 2003). Our interpretation of the systematic differences we observed between memory and perception is that they arise from the fact that top-down and bottom-up inputs to a hierarchical system produce different patterns. If this is correct, it is likely that systematic differences occur for other stimulus features and sensory modalities. Identifying such differences is likely to be highly informative for understanding how memory activation is generated.

Relation to other forms of memory and attention

Sensory reactivation during long-term memory retrieval may have parallels to sensory engagement in other forms of memory such as iconic memory or working memory. Nonetheless there may also be differences in the specific way that sensory circuits are used across these forms of memory. One critical factor may be how recently the sensory circuit was activated by a stimulus at the time of memory retrieval. In iconic memory studies, very detailed information can be retrieved if probed within a second of the sensory input (Sperling, 1960). In spatial working memory studies, sensory activity is thought to be maintained by active mechanisms from stimulus encoding through a seconds-long day. Many of these studies show that early visual areas have retinotopically specific signals throughout the delay (Sprague & Serences, 2013; Sprague et al., 2014; Rahmati et al., 2017), paralleling our findings. In imagery studies, eye-specific circuits presumed to be in V1 can be re-engaged if there is a delay of 5 minutes or less from when the subject viewed stimuli through the same eye, but not if there is a delay of 10 minutes (Ishai & Sagi, 1995). Hippocampally-dependent memory retrieval can likely engage visual cortex at much longer delays. Given that the mechanism for engaging sensory cortex may differ across these different forms of memory, the question of how similar sensory activation is across these timescales remains an important open question. For example, shorter-term forms of memory might, in principle, cause more spatially specific reactivation in early visual cortex than what we observed in long-term memory.

Our results also raise questions about whether long-term spatial memory and endogenous spatial attention share mechanisms for modulating the response of visual cortical populations. In typical endogenous spatial attention tasks, subjects are explicitly cued to the most likely location of an upcoming stimulus prior to being presented with a difficult visual judgment (Carrasco, 2011). fMRI studies have repeatedly found that spatial attention enhances visually-evoked responses in visual cortex (Somers et al., 1999; Gandhi et al., 1999; Buracas & Boynton, 2007; Li et al., 2008). Similar to our results, spatial attention has been show to elicit spatially localized activation in the absence of any visual stimulation (Luck et al., 1997; Kastner et al., 1999; Chawla et al., 1999; Ress et al., 2000). It is at least logically possible for attention and memory to dissociate. Most endogenous attention tasks have no memory component since the cue explicitly represents the attended location. In contrast, in most episodic memory tasks the association between a cue and a stimulus is intentionally arbitrary so that it must be acquired and retrieved in a hippocampally-dependent manner. However, it's possible that spatial attention and memory processes only differ in their dependency on the hippocampus to retrieve the target location. Once this target location is determined, the same mechanisms could be used to initiate enhanced processing of the target location in sensory areas. Future experiments and modeling efforts should determine whether memory-driven and attention-driven activation in visuals areas differ, and whether it's possible to develop a model top-down processing in visual cortex that can account for both sets of observations.

Conclusion

Together, our empirical and modeling results quantify similarities and differences between stimulus-triggered and memory-triggered activity in the human visual system and provide a plausible computational architecture for memory reactivation in sensory cortex. In agreement with prior work, we find that memory can cause retinotopic activation in visual cortex. These signals are noisier than perceptually driven signals, but tend to be centered at the same cortical locations as those triggered by perception. In contrast, we found that the amplitude and precision of memory-triggered responses differed systematically from stimulus-triggered responses, and that these differences were unlikely to be explained by decreased signal to noise or failed retrieval during the memory task. This work makes progress on specifying the mechanisms that underlie cortical reinstatement during memory retrieval and may shed light on broader computational principles that guide top-down processes in sensory systems.

Methods

Subjects

Nine human subjects participated in the experiment (5 males, 22–46 years old). All subjects had normal or correct-to-normal visual acuity, normal color vision, and no MRI contraindications. Subjects were recruited from the New York University community and included author S.E.F and author J.W. All subjects gave written informed consent to procedures approved by the New York University Institutional Review Board prior to participation. No subjects were excluded from data analysis.

Stimuli

Experimental stimuli included nine unique radial frequency patterns (Fig. 1a). We first generated patterns that differed along two dimensions: radial frequency and amplitude. We chose stimuli that tiled a one dimensional subspace of this two dimensional space, with radial frequency inversely proportional to amplitude. The nine chosen stimuli took radial frequency and amplitude values of: [2, .9], [3, .8], [4, .7], [5, .6], [6, .5], [7, .4], [8, .3], [9, .2], [10, .1]. We selected four of these stimuli to train subjects on in the behavioral training session and to appear in the fMRI session. For every subject, those stimuli were: [3, .8], [5, .6], [7, .4], [9, .2]; (radial frequency, amplitude). The remaining five stimuli were used as lures in the test trials of the behavioral training session. Stimuli were saved as images and cropped to the same size.

Experimental procedure

The experiment began with a behavioral training session, during which subjects learned four paired associates (Fig. 1). Specifically, subjects learned that four colored fixation dot cues were uniquely associated with four spatially localized radial frequency patterns. An fMRI session immediately followed completion of the behavioral session (Fig. 2a). During the scan, subjects participated in two types of functional runs (approximately 3.5 min each): (1) perception, where they viewed the spatially localized stimuli; and (2) memory, where they were presented with the fixation cues and recalled the associated spatial stimuli. Details for each of these phases are described below. A separate retinotopic mapping session was also performed for each subject (Fig. 2b), which is described in the next section.

Behavioral training

For each subject, the four radial frequency patterns were first randomly assigned to one of four polar angle locations in the visual field (45°, 135°, 225°, or 315°) and one of four colored cues (orange, magenta, blue, green; Fig. 1b). Immediately before scanning, subjects learned the association between the four colored cues and the four spatially localized stimuli through interleaved study and test blocks (Fig. 1c). Subjects alternated between study and test blocks, completing a minimum of four blocks of each type. Subjects were required to reach at least 95% accuracy, and performed additional rounds of study-test if they did not reach this threshold after four test blocks.

During study blocks, subjects were presented with the associations. Subjects were instructed to maintain central fixation and to learn each of the four associations in anticipation of a memory test. At the start of each study trial (Fig. 1c), a central white fixation dot (radius = 0.1 dva) switched to one of the four cue colors. After a 1 sec delay, the associated radial frequency pattern appeared at 2°eccentricity and its assigned polar angle location in the visual field. Each pattern image subtended 1.5 dva and was presented for 2 sec. The fixation dot then returned to white, and the next trial began after a 2 sec interval. No subject responses were required. Each study block contained 16 trials (4 trials per association), presented in random order.

During test blocks, subjects were presented with the color cues and tested on their memory for the associated stimulus pattern and spatial location. Subjects were instructed to maintain central fixation and to try to covertly recall each stimulus when cued. At the start of each test trial (Fig. 1c), the central white fixation dot switched to one of the four cue colors. This cue remained on the screen for 2.5 sec while subjects attempted to covertly retrieve the associated stimulus. At the end of this period, a test stimulus was presented at 2°of eccentricity for 2 sec. Then, subjects were cued to make two consecutive responses to the test stimulus: whether it was the correct radial frequency pattern (yes/no) and whether it was presented at the correct polar angle location (yes/no). Each test stimulus had a 50% probability of being the correct pattern. Incorrect patterns were drawn randomly from the three patterns associated with other cues and the five lure patterns (Fig. 1a). Each test stimulus had a 50% probability of being in the correct polar angle location, which was independent from the probability of being the correct pattern. Incorrect polar angle locations were drawn from the three locations assigned to the other patterns and 20 other evenly spaced locations around the visual field (Fig. 1a). This placed the closest spatial lure at 15°of polar angle away from the correct location. Responses were solicited from the subject with the words "Correct pattern?" or "Correct location?" displayed centrally in white text. The order of these queries was counterbalanced across test blocks. Subjects responses were recorded on a keyboard with a maximum response window of 2 sec. Immediately after a response was made or the response window closed, the color of the text turned black to indicate an incorrect response if one was made. After this occurred for both queries, subjects were presented with the colored fixation dot cue and correct spatially localized pattern for 1 sec as feedback. This feedback occurred for every trial, regardless of subject responses to the probe. Each test block contained 16 trials (4 trials per association), presented in random order.

fMRI session

During the fMRI session, subjects participated in two types of functional runs: perception and memory retrieval (Fig. 2a). Subjects completed 5–6 runs each of perception and memory in an interleaved order. This amounted to 40–48 repetitions of perceiving each stimulus and of remembering each stimulus per subject.

During perception runs, subjects viewed the colored fixation dot cues and the radial frequency patterns in their learned locations. Subjects were instructed to maintain central fixation and to perform a one-back task on the stimuli. The purpose of the one-back task was to enforce covert stimulus-directed attention on each trial. At the start of each perception trial (Fig. 2a, top), a central white fixation dot (radius = 0.1 dva) switched to one of the four cue colors. After a 0.5 sec delay, the associated radial frequency pattern appeared at 2°of eccentricity and its assigned polar angle location in the visual field. Each pattern subtended 1.5 dva and was presented for 2.5 sec. The fixation dot then returned to white and the next trial began after a variable interval. Intervals were drawn from an approximately geometric distribution sampled at 3, 4, 5, and 6 sec with probabilities of 0.5625, 0.25, 0.125, and 0.0625 respectively. Subjects indicated when a stimulus repeated from the previous trial using a button box. Responses were accepted during the stimulus presentation or during the interstimulus interval. Each perception run contained 32 trials (8 trials per stimulus). The trial order was randomized for each run, separately for every subject.

During memory runs, subjects viewed the colored fixation dot cues and recalled the associated patterns in their learned spatial locations. Subjects were instructed to maintain central fixation, to use the cues to initiate recollection, and to make a subjective judgment about the vividness of their memory on each trial. The purpose of the vividness task was to enforce attention to the remembered stimulus on each trial. At the start of each memory trial (Fig. 2a, top), the central white fixation dot switched to one of the four cue colors. This cue remained on the screen for a recollection period of 3 sec. The fixation dot then returned to white and the next trial began after a variable interval. Subjects indicated whether the stimulus associated with the cue was vividly remembered, weakly remembered, or not remembered using a button box. Responses were accepted during the cue presentation or during the interstimulus interval. Each memory run contained 32 trials (8 trials per stimulus). For a given subject, each memory run's trial order and trial onsets were exactly matched to one of the perception runs. The order of these matched memory runs was scrambled relative to the order of the perception runs.

Retinotopic mapping procedure

Each subject completed either 6 or 12 identical retinotopic mapping runs in a separate fMRI session from the main experiment (Fig. 2b, top). Stimuli and procedures for the retinotopic mapping session were based on those used by the Human Connectome Project (Benson et al., 2018) and were identical to those reported in Benson & Winawer (2018). During each functional run, bar apertures on a uniform gray background swept across the central 24 degrees of the subject's visual field (circular aperture with a radius of 12 dva). Bar apertures were a constant width (1.5 dva) at all eccentricities. Each sweep began at one of eight equally spaced positions around the edge of the circular aperture, oriented perpendicularly to the direction of the sweep. Horizontal and vertical sweeps traversed the entire diameter of the circular aperture while diagonal sweeps stopped halfway and were followed by a blank period. A full-field sweep or half-field sweep plus blank took 24 s to complete. One functional run contained 8 sweeps, taking 192 s in total. Bar apertures contained a grayscale pink noise background with randomly placed faces, scenes, objects, and words at a variety of sizes. Noise background and stimuli were updated at a frequency of 3 Hz. Each run of the task had an identical design. Subjects were instructed to maintain fixation on a central dot and to use a button box to report whenever the dot changed color. Color changes occurred on average every 3 s.

MRI acquisition

Images were acquired on a 3T Siemens Prisma MRI system at the Center for Brain Imaging at New York University. Functional images were acquired with a T2*-weighted multiband EPI sequence with whole-brain coverage (repetition time = 1 s, echo time = 37 ms, flip angle = 68°, 66 slices, 2 x 2 x 2 mm voxels, multiband acceleration factor = 6, phase-encoding = posterior-anterior) and a Siemens 64-channel head/neck coil. Spin echo images with anterior-posterior and posterior-anterior phase-encoding were collected to estimate the susceptibility-induced distortion present in the functional EPIs. Between one and three whole-brain T1-weighted MPRAGE 3D anatomical volumes (.8 x .8 x .8 mm voxels) were also acquired for seven subjects. For two subjects, previously acquired MPRAGE volumes (1 x 1 x 1 mm voxels) from a 3T Siemens Allegra head-only MRI system were used.

MRI processing

Preprocessing

Anatomical and functional images were preprocessed using FSL (Smith et al., 2004) and Freesurfer (Fischl, 2012) tools implemented in a Nipype workflow (Gorgolewski et al., 2011). To correct for head motion, each functional image acquired in a session was realigned to a single band reference image and then registered to the spin echo distortion scan acquired with the same phase encoding direction. The two spin echo images with reversed phase encoding were used to estimate the susceptibility-induced distortion present in the EPIs. For each EPI volume, this nonlinear unwarping function was concatenated with the previous spatial registrations and applied with a single interpolation. Freesurfer was used to perform segmentation and cortical surface reconstruction on each subject's average anatomical volume. Registration from the functional images to each

subject's anatomical volume was performed using boundary-based registration. Preprocessed functional timeseries were then projected onto each subject's reconstructed cortical surface.

GLM analyses

Beginning with each subject's surface-based timeseries, we used GLMdenoise (Kay et al., 2013a) to estimate the neural pattern of activity evoked by the perception and memory of every stimulus (Fig. 2a). GLMdenoise improves signal-to-noise ratios in GLM analyses by identifying a pool of noise voxels whose responses are unrelated to the task and regressing them out of the timeseries. This technique first converts all timeseries to percent signal change and determines an optimal hemodynamic response function for all vertices using an iterative linear fitting procedure. It then identifies noise vertices as vertices with negative R^2 values in the task-based model. Then, it derives noise regressors from the noise pool time series using principal components analysis and iteratively projects them out of the timeseries of all vertices, one noise regressor at a time. The optimal number of noise regressors is determined based on cross-validated R^2 improvement for the task-based model. We estimated two models using this procedure. We constructed design matrices for the perception model to have four regressors of interest (one per stimulus), with events corresponding to stimulus presentation. Design matrices for the memory model were constructed the same way, with events corresponding to the the cued retrieval period. These models returned parameter estimates reflecting the BOLD amplitude evoked by perceiving or remembering a given stimulus for every vertex on a subject's cortical surface (Fig. 2a, bottom).

Fitting pRF models

Images from the retinotopic mapping session were preprocessed as above, but omitting the final step of projecting the timeseries to the cortical surface. Using these timeseries, nonlinear symmetric 2D Gaussian population receptive field (pRF) models were estimated in Vistasoft (Fig. 2b), as described previously (Dumoulin & Wandell, 2008; Kay et al., 2013b). We refer to this nonlinear version of the pRF model as the compressive spatial summation (CSS) model, following Kay et al. (2013b). Briefly, we estimated the receptive field parameters that, when applied to the drifting bar stimulus images, minimized the difference between the observed and predicted BOLD timeseries. First, stimulus images were converted to contrast apertures and downsampled to 101 x 101 grids. Timeseries from each retinotopy run were resampled to anatomical space and restricted to gray matter voxels. Timeseries were then averaged across runs. pRF models were solved using a two stage coarse-to-fine fit on the average time series. The first stage of the model fit was a coarse grid fit, which was used to find an approximate solution robust to local minima. This stage was solved on a volume-based timeseries that was first temporally decimated, spatially blurred on the cortical surface, and spatially subsampled. The parameters obtained with this fit were interpolated and then used as a seed for subsequent nonlinear optimization, or fine fit. This procedure yielded four final parameters of interest for every voxel: eccentricity (r), polar angle (θ), sigma (σ), exponent (n). The eccentricity and polar angle parameters describe the location of the receptive field in space, the sigma parameter describes the size of the receptive field, and the exponent describes the amount of compressive spatial summation applied to responses from the receptive field. Eccentricity and polar angle parameters were converted from polar coordinates to rectangular coordinates (x, y) for some analyses. Variance explained by the pRF model with these parameters was also calculated for each voxel. All parameters were then projected from each subject's anatomical volume to the cortical surface (Fig. 2b, bottom).

ROI definitions

Regions of interest were defined by hand-drawing boundaries at polar angle reversals on each subject's cortical surface, following established practice (Wandell et al., 2007). We used this method to define six ROIs spanning early to mid-level visual cortex: V1, V2, V3, hV4, LO (LO1 and LO2), and V3ab (V3a and V3b).

We further restricted each ROI by preferred eccentricity in order to isolate vertices responsive to our stimuli. We excluded vertices with eccentricity values less than 0.5°and greater than 8°. This procedure excluded vertices responding primarily to the fixation dot and vertices near the maximal extent of visual stimulation in the scanner. We also excluded vertices whose variance explained by the pRF model (R^2) was less than 0.1, indicating poor spatial selectivity. All measures used to exclude vertices from ROIs were independent of the measurements made during the perception and memory tasks.

Analyses

Our main empirical analyses examined the evoked BOLD response to our experimental stimuli during perception and memory as a function of visual field parameters estimated from the pRF model. Our first step was to visualize evoked activity during perception and memory in visual field coordinates (Fig. 3a). Transforming the data in this way allowed us to view the activity in a common reference frame across all brain regions, rather than on the cortical surface, where comparisons are made difficult by the fact that surface area and cortical magnification differ substantially from one area to the next. To do this, we selected the (x, y) parameters for each surface vertex from the retinotopy model and the β parameters from the GLM analysis. Separately for a given ROI, subject, stimulus, and task (perception/memory), we interpolated the β values over (x, y) space. We rotated each of these representations according to the polar angle location of the stimulus so that they would be aligned at the upper

vertical meridian. We then z-scored each representation before averaging across stimuli and subjects. We used these images to gain intuition about the response profiles and to guide subsequent quantitative analyses.

Before quantifying these representations, we simplified them further. Because our stimuli were all presented at the same eccentricity, we reduced our 2D stimulus coordinate representations to 1D dimensional responses functions on the polar angle dimension (Fig. 4a). We did this by selecting surface vertices whose (x,y) coordinates were within one σ of the stimulus eccentricity (2°) for each ROI. We then binned the evoked BOLD response into 19 bins of polar angle distance from the stimulus and averaged within each bin to produce polar angle response functions for each subject. We divided each subject's response function by the norm of the response vector before averaging across subjects and then multiplying by the average vector norm. This procedure prevents a subject with a high baseline BOLD response from dominating the average response. The resulting average polar angle response functions showed clear surround suppression for polar angles near the stimulus during perception. Given this, we fit a difference of two von Mises distributions to the average data, with the location parameters (μ) of the two von Mises distributions fixed to be equal, but the spread (κ) and scale allowed to differ.

We quantitatively assessed the similarities and differences between perception and memory responses using these fits. We examined the location parameter of the two von Mises distributions, and also computed the amplitude and FWHM of the fit. We repeated the fitting procedure 500 times, drawing subjects with replacement, to create bootstrapped 95% confidence intervals of location, amplitude, and FWHM for perception and memory.

Signal-to-noise and failed retrieval simulations

We performed two simulations designed to rule out alternative explanations for our data. First, we simulated 100 new datasets where perception parameter estimates were noisier than memory parameter estimates. This tested whether lower SNR by itself could produce the pattern of memory responses we observed. We determined the amount of signal and noise actually observed for each perception and memory parameter estimate by examining bootstrapped distributions produced by GLMdenoise. We defined the median parameter estimate across bootstraps as the amount of signal and the standard error of this distribution as the amount of noise. Averaging across all data included in Figure 4, memory data had lower signal-to-noise ratios than perception data. Perception SNR was less than twice as high as memory SNR in our regions of interest. We overcompensated for these observed differences by simulating new perception data with twice the amount of noise. For every surface vertex included in our analyses, we selected a new parameter estimate from a normal distribution defined by the true signal value (median) and twice the true noise value (SE). Critically, we made the draws correlated across vertices for each dataset. We did this by selecting a scale factor from a standard normal distribution which determined how many (noisy) SEs away from the median each vertex's simulated value would lie. This scale factor was shared across all vertices in an ROI for a given simulation. This procedure overcompensates for the spatial correlation present in BOLD data by assuming that all vertices in an ROI are 100% correlated. We analyzed our simulated datasets using the same procedure we applied to the actual data. This produced von Mises fits for the simulated data (Fig. 5a). We evaluated the location, amplitude, and FWHM parameters derived from these simulated data fits by comparing them to the confidence intervals estimated from the actual data (Fig. 5c, left).

We performed a second simulation where a subset of memory trials were assumed to contain no signal while the remaining memory trials were assumed to have the same signal as perception. This simulation tested whether including failed retrieval trials in our analyses could produce apparent differences in spatial tuning properties between perception and memory. We simulated 100 datasets in each of three conditions: 25%, 50% and 75% failed retrieval. For each of these conditions, one, two, of three of the four stimuli were randomly designated as forgotten. For these forgotten stimuli, new memory parameter estimates were drawn from a distribution defined by zero signal for every vertex. For the remaining stimuli, new memory parameter estimates were drawn from a distribution defined by the true perception signal. Noise was equated for both trial types; for each vertex, we used the the amount of noise observed during perception. As in the previous analysis, simulated data were correlated across vertices in an ROI. We analyzed our simulated datasets separately for each condition, plotting the von Mises fits in Figure 5b and evaluating location, amplitude, and FWHM for the 25% failed retrieval condition.

pRF forward model

We evaluated the ability of our pRF model to account for our perception and memory measurements. To do this, we used our pRF model as a forward model. This means that we took the pRF model parameters fit to fMRI data from the retinotopy session (which used a drifting bar stimulus) and used them to generate predicted BOLD responses to our four experimental stimuli. The model takes processed stimulus images as input, and for each of these images, outputs a predicted BOLD response (in units of % signal change) for every cortical surface vertex. Before running the model, we transformed our experimental stimuli into binary contrast apertures with values of 1 where the stimulus was and values of 0 everywhere else. These images were downsampled to the same resolution as the images used to fit the pRF model (101 x 101).

The pRF forward model has two fundamental operations. In the first operation, a stimulus contrast aperture image is multiplied by a voxel's pRF. In the version of the model we fit to the retinotopy data (CSS model), this pRF is defined as a circular symmetric 2D Gaussian, paramaterized by a location in the visual field (x,y) and a size (σ). Multiplying stimulus images by the pRF is a linear operation, where smaller stimuli and larger pRFs yield smaller BOLD responses. Further, because

both the stimulus images and the pRFs are defined as positive only, this multiplication can only generate positive values. The second operation applies a power-law exponent (n) to the result of the multiplication, effectively boosting small responses. This nonlinear operation is the key component of the CSS model and improves model accuracy in high-level visual areas that are known to exhibit subadditive spatial summation (Kay et al., 2013b; Mackey et al., 2017). The values of the exponent range from 0 to 1, where a value of 1 returns the model to linear. The output of this nonlinear stage is multiplied by a final scale parameter (β), which returns the units to % signal change. Again, each voxel's parameters (x, y, σ, n, β) were fit using fMRI data from the retinotopy session (see Retinotopic mapping procedure and Fitting pRF models for details).

We analyzed the predicted responses using the same procedure as the data to generate predicted polar angle response functions. The CSS model produced sensible predictions for our perception measurements (Supplementary Fig. 1a). However, it cannot account for the surround suppression we observed in V1–V3 during perception because the model only generates positive values. Prior work has shown that difference-of-Gaussians (DoG) pRF models can account for the center-surround structure we observed (Zuiderbaan et al., 2012). We tested whether converting each 2D Gaussian pRF in the CSS model to a 2D DoG pRF would better account for our data. To do this, we took every 2D Gaussian pRF from the CSS model, and we subtracted from it a second 2D Gaussian pRF that was centered at the same location but was twice as wide and half as high. This ratio of 2σ and $.5\beta$ between the negative and positive Gaussians was fixed for all voxels. In order to prevent the resulting DoG pRF from being systematically narrower and lower in amplitude than the original pRF, we rescaled the σ and β of the original pRF before converting it to a DoG. We multiplied the original σ by $\sqrt{2}$ and the original β by 2, resulting in a DoG pRF with equivalent FWHM and amplitude as the original pRF. Thus, the DoG pRF differed from the original pRF only in the presence of a suppressive surround. After modifying the pRF in this way, we ran the model exactly as described in the prior paragraph (Fig. 6a).

We next compared how well the DOG+CSS model predictions matched our perception versus memory measurements. We correlated the predicted location, amplitude, and FWHM parameters for each ROI with the actual perception and memory parameters. We evaluated these relationships by fitting a linear model to the observations and to 500 bootstrapped datasets and their predictions (Fig. 6c).

We also compared the model accuracy of the DoG+CSS and CSS models alongside a linear model with no exponent parameter (Supplementary Fig. 1a). We calculated the coefficient of determination (R^2) for the predicted response functions in each ROI, separately for perception and memory data (Supplementary Fig. 1b). Under this measure, a model that predicts the mean response for every value of polar angle distance will have an R^2 of zero, with better models producing positive values and worse models producing negative values. In order to evaluate how much these accuracies were influenced by failure to capture mean amplitude, we re-scaled our predictions with the scale factor that best fit the data and re-computed accuracy.

Hierarchical model

While stimulus-referred pRF models have been very successful in accounting for stimulus-triggered activity during perception, it's less clear how these models can be used to account for memory-triggered activity. A more useful model for this question would allow for directly simulating top-down activation in the visual system. To this end, we assessed whether a simple instantiation of a single neural network model could account for both the perception and memory data. We implemented a a simple hierarchical model of neocortex in which the activity from each layer was created by pooling activity from the previous layer. This model shares some features of hierarchical models of object recognition (Riesenhuber & Poggio, 1999; Serre et al., 2007), but is much simpler in that it encodes only one stimulus feature: space (Kay et al., 2013b). We began with a 1D square wave stimulus, which spanned -20 to 20 degrees of polar angle. We created a fixed Gaussian convolution kernel ($\mu = 0$, $\sigma = 15$), which we convolved with the stimulus to create the activity in layer 1. This layer 1 activity was convolved with the same Gaussian kernel to create the layer 2 activity, and this process was repeated recursively (Fig. 7a, left). In order to simulate memory-triggered responses in this network, we made two assumptions. First, we assumed that the feedback phase began with the layer 10 activity from the feedforward phase. That is, we assumed no information loss or distortion between perception and memory in the last layer. Second, we assumed that all connections were reciprocal and thus that the same Gaussian kernel was applied to transform layers in the feedback direction as in the feedforward direction (Fig. 7a, right). Thus, in the feedback simulation, we convolved the layer 8 activity with the Gaussian kernel to produce the layer 7 activity and repeated this procedure recursively (Fig. 7b). Note that these computations can be performed with matrix multiplication rather than convolution by converting the convolutional kernel to a Toeplitz matrix, which is symmetric. In this case, the transpose of the Toeplitz matrix (itself) is used in the feedback direction. We plot the location, amplitude and FWHM for each layer's activation in the same convention as the data (Fig. 7c)

Acknowledgements

S.E.F. was supported by an NIH Blueprint D-SPAN Award (F99-NS105223).

References

Barone, P., Batardiere, A., Knoblauch, K., & Kennedy, H. (2000). Laminar distribution of neurons in extrastriate areas projecting to visual areas V1 and V4 correlates with the hiearchical rank and intimates the operation of a distance rule. *Journal of Neuroscience*, 20(9), 3263–3281.

Benson, N. C., Jamison, K. W., Arcaro, M. J., Vu, A. T., Glasser, M. F., Coalson, T. S., Van Essen, D. C., Yacoub, E., Ugurbil, K., Winawer, J., & Kay, K. (2018). The Human Connectome Project 7 Tesla retinotopy dataset: Description and population receptive field analysis. *Journal of Vision*, 18(13), 23.

Benson, N. C. & Winawer, J. (2018). Bayesian analysis of retinotopic maps. *eLife*, 7, 0–45.

Bone, M. B., St-Laurent, M., Dang, C., McQuiggan, D. A., Ryan, J. D., & Buchsbaum, B. R. (2018). Eye Movement Reinstatement and Neural Reactivation During Mental Imagery. *Cerebral Cortex*, 29(3), 1075–1089.

Bosch, S. E., Jehee, J. F. M., Fernandez, G., & Doeller, C. F. (2014). Reinstatement of Associative Memories in Early Visual Cortex Is Signaled by the Hippocampus. *Journal of Neuroscience*, 34(22), 7493–7500.

Breedlove, J. L., St-Yves, G., Olman, C. A., & Naselaris, T. P. (2018). Mental imagery encoding models reveal signatures of inference in a hierarchical generative model. *bioRxiv*.

Buracas, G. T. & Boynton, G. M. (2007). The Effect of Spatial Attention on Contrast Response Functions in Human Visual Cortex. *Journal of Neuroscience*, 27(1), 93–97.

Carrasco, M. (2011). Visual attention: The past 25 years. *Vision Research*, 51(13), 1484–1525.

Chawla, D., Rees, G., & Friston, K. J. (1999). The physiological basis of attentional modulation in extrastriate visual areas. *Nature Neuroscience*, 2(7), 671–676.

Damasio, A. R. (1989). Time-locked multiregional retroactivation: A systems level proposal for the neural substrates of recall and recognition. *Cognition*, 33, 25–62.

Dijkstra, N., Ambrogioni, L., & Gerven, M. A. J. V. (2019). Neural dynamics of perceptual inference and its reversal during imagery. *bioRxiv*.

Dougherty, R. F., Koch, V. M., Brewer, A. A., Fischer, B., Modersitzki, J., & Wandell, B. A. (2003). Visual field representations and locations of visual areas v1/2/3 in human visual cortex. *Journal of Vision*, 3(10), 586–598.

Dumoulin, S. O. & Wandell, B. A. (2008). Population receptive field estimates in human visual cortex. *NeuroImage*, 39(2), 647–660.

Eichenbaum, H., Dudchenko, P., Wood, E., Shapiro, M., & Tanila, H. (1999). The Hippocampus, Memory, and Place Cells: Is It Spatial Memory or a Memory Space? *Neuron*, 23(2), 209–226.

Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Chichilnisky, E.-J., & Shadlen, M. N. (1994). fMRI of human visual cortex.

Favila, S. E., Samide, R., Sweigart, S. C., & Kuhl, B. A. (2018). Parietal representations of stimulus features are amplified during memory retrieval and flexibly aligned with top-down goals. *Journal of Neuroscience*, 38(36), 0564–18.

Felleman, D. J. & Essen, D. C. V. (1991). Distributed Hierarchical Processing in the Primate Cerebral Cortex. *Cerebral Cortex*, 1, 1–47.

Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2), 774–781.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4), 193–202.

Gandhi, S. P., Heeger, D. J., & Boynton, G. M. (1999). Spatial attention affects brain activity in human primary. *Proc. Natl. Acad. Sci. USA*, 96(March), 3314–3319.

Gordon, A. M., Rissman, J., Kiani, R., & Wagner, A. D. (2014). Cortical Reinstatement Mediates the Relationship Between Content-Specific Encoding Activity and Subsequent Recollection Decisions. *Cerebral Cortex*, 24(12), 3350–3364.

Gorgolewski, K., Madison, C., Burns, C. D., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S. S. (2011). Nipype: A Flexible, Lightweight and Extensible Neuroimaging Data Processing Framework in Python. *Frontiers in Neuroinformatics*, 5(August).

Graham, N. (1989). *Visual Pattern Analyzers*. New York, NY: Oxford University Press.

Hebb, D. O. (1968). Concerning imagery. *Psychological Review*, 75(6), 466–77.

Heeger, D. J. (2017). Theory of cortical function. *Proceedings of the National Academy of Sciences*, 114(8), 1773–1782.

Horikawa, T. & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, 8(May), 1–15.

Horton, J. C. & Hoyt, W. F. (1991). The Representation of the Visual Field in Human Striate Cortex: A Revision of the Classic Holmes Map. *Archives of Ophthalmology*, 109(6), 816–824.

Ishai, A. & Sagi, D. (1995). Common mechanisms of visual imagery and perception. *Science*, 268(5218), 1772–1774.

James, W. (1890). *The Principles of Psychology*. New York, NY: Holt.

Kastner, S., Pinsk, M. A., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1999). Increased activity in human visual cortex during directed attention in the absence of visual stimulation. *Neuron*, 22(4), 751–61.

Kay, K. N., Naselaris, T., Prenger, R. L., & Gallant, J. L. (2008). Identifying natural images from human brain activity. *Nature*, 452(March), 352–355.

Kay, K. N., Rokem, A., Winawer, J., Dougherty, R. F., & Wandell, B. A. (2013a). GLMdenoise: A fast, automated technique for denoising task-based fMRI data. *Frontiers in Neuroscience*, 7(7 DEC), 1–15.

Kay, K. N., Winawer, J., Mezer, A., & Wandell, B. A. (2013b). Compressive spatial summation in human visual cortex. *Journal of Neurophysiology*, 110(2), 481–494.

Kosslyn, S. M. (1980). *Image and Mind*. Cambridge, MA: Harvard University Press.

Kosslyn, S. M., Thompson, W. L., Kim, I. J., & Alpert, N. M. (1995). Topographical representations of mental images in primary visual cortex. *Nature*, 378(6556), 496–8.

Kuhl, B. A., Johnson, M. K., & Chun, M. M. (2013). Dissociable neural mechanisms for goal-directed versus incidental memory reactivation. *The Journal of Neuroscience*, 33(41), 16099–109.

Kuhl, B. A., Rissman, J., Chun, M. M., & Wagner, A. D. (2011). Fidelity of neural reactivation reveals competition between memories. *Proceedings of the National Academy of Sciences*, 108(14), 5903–5908.

Lee, S.-h., Kravitz, D. J., & Baker, C. I. (2018). Differential Representations of Perceived and Retrieved Visual Information in Hippocampus and Cortex. *Cerebral Cortex*, (pp. 1–10).

Li, X., Lu, Z.-L., Tjan, B. S., Dosher, B. A., & Chu, W. (2008). Blood oxygenation level-dependent contrast response functions identify mechanisms of covert attention in early visual areas. *Proceedings of the National Academy of Sciences*, 105(16), 6202–6207.

Linde-Domingo, J., Treder, M. S., Kerrén, C., & Wimber, M. (2019). Evidence that neural information flow is reversed between object perception and object reconstruction from memory. *Nature Communications*, 10(1), 179.

Luck, S. J., Chelazzi, L., Hillyard, S. A., & Desimone, R. (1997). Neural Mechanisms of Spatial Selective Attention in Areas V1, V2, and V4 of Macaque Visual Cortex. *Journal of Neurophysiology*, 77(1), 24–42.

Mackey, W. E., Winawer, J., & Curtis, C. E. (2017). Visual field map clusters in human frontoparietal cortex. *eLife*, 6(e22974).

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: insights from the successes and failures of connectionist models of learning and memory. *Psychological review*, 102(3), 419–457.

Naselaris, T., Olman, C. A., Stansbury, D. E., Ugurbil, K., & Gallant, J. L. (2015). A voxel-wise encoding model for early visual areas decodes mental images of remembered scenes. *NeuroImage*, 105, 215–228.

Naya, Y., Yoshida, M., & Miyashita, Y. (2001). Backward Spreading of Memory-Retrieval Signal in the Primate Temporal Cortex. *Science*, 291(5504), 661–664.

O'Craven, K. M. & Kanwisher, N. (2000). Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions. *Journal of Cognitive Neuroscience*, 12(6), 1013–1023.

Pearson, J. (2019). The human imagination: the cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*.

Pearson, J., Clifford, C. W., & Tong, F. (2008). The Functional Impact of Mental Imagery on Conscious Perception. *Current Biology*, 18(13), 982–986.

Pearson, J., Naselaris, T., Holmes, E. A., & Kosslyn, S. M. (2015). Mental Imagery: Functional Mechanisms and Clinical Applications. *Trends in Cognitive Sciences*, 19(10), 590–602.

Polyn, S. M., Natu, V. S., Cohen, J. D., & Norman, K. A. (2005). Category-Specific Cortical Activity Precedes Retrieval During Memory Search. *Science*, 310(5756), 1963–6.

Pylyshyn, Z. W. (2002). Mental imagery: In search of a theory. *Behavioral and Brain Sciences*, 25(2), 157–182.

Rahmati, M., Saber, G., & Curtis, C. (2017). Population Dynamics of Early Visual Cortex During Working Memory. *Journal of Cognitive Neuroscience*.

Ress, D., Backus, B. T., & Heeger, D. J. (2000). Activity in primary visual cortex predicts performance in a visual detection task. *Nature Neuroscience*, 3(9), 940–945.

Riesenhuber, M. & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11), 1019–1025.

Sereno, M. I., Dale, A. M., Reppas, J. B., Kwong, K. K., Belliveau, J. W., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Borders of multiple visual areas in humans revealed by functional magnetic resonance imaging. *Science*, 268(5212), 889–893.

Serre, T., Oliva, A., & Poggio, T. (2007). A feedforward architecture accounts for rapid categorization. *Proceedings of the National Academy of Sciences*, 104(15), 6424–6429.

Slotnick, S. D., Thompson, W. L., & Kosslyn, S. M. (2005). Visual mental imagery induces retinotopically organized activation of early visual areas. *Cerebral Cortex*, 15(10), 1570–1583.

Smith, S. M., Jenkinson, M., Woolrich, M. W., Beckmann, C. F., Behrens, T. E., Johansen-Berg, H., Bannister, P. R., Luca, M. D., Drobnjak, I., Flitney, D. E., Niazy, R. K., Saunders, J., Vickers, J., Zhang, Y., Stefano, N. D., Brady, J. M., & Matthews, P. M. (2004). Advances in functional and structural MR image analysis and implementation as FSL. *NeuroImage*, 23, S208–S219.

Somers, D. C., Dale, A. M., Seiffert, A. E., & Tootell, R. B. H. (1999). Functional MRI reveals spatially specific attentional modulation in human primary visual cortex. *Proceedings of the National Academy of Sciences*, 96(4), 1663–1668.

Sperling, G. (1960). The information available in brief visual presentations. *Psychological Monographs: General and Applied*, 74(11), 1–29.

Sprague, T. C., Ester, E. F., & Serences, J. T. (2014). Reconstructions of information in visual spatial working memory degrade with memory load. *Current Biology*, 24(18), 2174–2180.

Sprague, T. C. & Serences, J. T. (2013). Attention modulates spatial priority maps in the human occipital, parietal and frontal cortices. *Nature neuroscience*, 16(12), 1879–1887.

Sutterer, D. W., Foster, J. J., Serences, J. T., Vogel, E. K., & Awh, E. (2019). Alpha-band oscillations track the retrieval of precise spatial representations from long-term memory. *Journal of Neurophysiology*, 122(2), 539–551.

Tartaglia, E. M., Bamert, L., Mast, F. W., & Herzog, M. H. (2009). Human Perceptual Learning by Mental Imagery. *Current Biology*, 19(24), 2081–2085.

Thirion, B., Duchesnay, E., Hubbard, E., Dubois, J., Poline, J. B., Lebihan, D., & Dehaene, S. (2006). Inverse retinotopy: Inferring the visual content of images from brain activation patterns. *NeuroImage*, 33(4), 1104–1116.

Waldhauser, G. T., Braun, V., & Hanslmayr, S. (2016). Episodic Memory Retrieval Functionally Relies on Very Rapid Reactivation of Sensory Information. *The Journal of Neuroscience*, 36(1), 251–260.

Wandell, B., Dumoulin, S. O. S., & Brewer, A. A. a. (2007). Visual Field Maps in Human Cortex. *Neuron*, 56(2), 366–383.

Wandell, B. A. & Winawer, J. (2011). Imaging retinotopic maps in the human brain. *Vision Research*, 51(7), 718–737.

Wandell, B. A. & Winawer, J. (2015). Computational neuroimaging and population receptive fields. *Trends in Cognitive Sciences*, 19(6), 349–357.

Wheeler, M. E., Petersen, S. E., & Buckner, R. L. (2000). Memory's echo: Vivid remembering reactivates sensory-specific cortex. *Proceedings of the National Academy of Sciences*, 97(20), 11125–11129.

Winawer, J., Huk, A. C., & Boroditsky, L. (2010). A motion aftereffect from visual imagery of motion. *Cognition*, 114(2), 276–284.

Zuiderbaan, W., Harvey, B. M., & Dumoulin, S. O. (2012). Modeling center-surround configurations in population receptive fields using fMRI. *Journal of Vision*, 12(3), 10–10.

Supplementary Figures

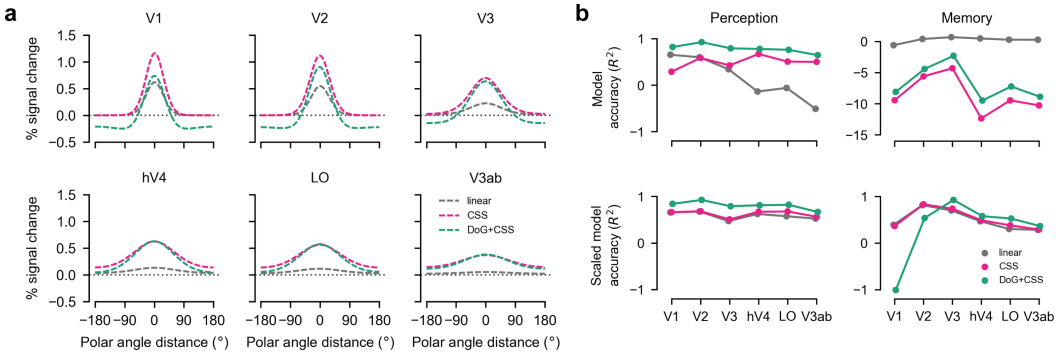

Figure 1. pRF model comparisons. (a) Predicted polar angle response functions are plotted for three pRF models: linear, CSS, and DoG+CSS. Comparing these responses to perception data plotted in 4b, the linear model did the poorest job of predicting perception responses. Linear predictions underestimated the amplitude of the observed response, particularly in higher regions. Both nonlinear models (CSS and DOG+CSS) avoided this magnitude of failure. The DoG+CSS model selectively captured negative responses in V1–V3. **(b)** Top: Model accuracy (R^2) of the predicted polar angle response functions for each pRF model, evaluated separately for perception and memory data in each ROI. Accuracy of the linear model in predicting perception data dropped steadily moving away from V1, indicating poor fit. Model accuracies for the the CSS and DoG+CSS models were higher and more stable across ROIs, with the DoG+CSS performing slightly better in every region. Accuracy of all three models was far worse for memory data than perception data in every ROI. Bottom: Because model accuracy is strongly influenced by mean amplitude, we re-scaled our predictions with the best-fitting scale factor and then recomputed model accuracy. This greatly improves the accuracy of the model predictions for memory. However, perception predictions remain superior, suggesting that mean activation differences alone do not account for differences in model accuracy for perception and memory data.