nature human behaviour

Article

https://doi.org/10.1038/s41562-024-01870-3

Images with harder-to-reconstruct visual representations leave stronger memory traces

Received: 20 February 2023

Accepted: 19 March 2024

Published online: 13 May 2024

Check for updates

Qi Lin[1,2,5] ✉, Zifan Li[3,5], John Lafferty[3,4,6] ✉ & Ilker Yildirim[1,3,4,6] ✉

Much of what we remember is not because of intentional selection, but simply a by-product of perceiving. This raises a foundational question about the architecture of the mind: how does perception interface with and influence memory? Here, inspired by a classic proposal relating perceptual processing to memory durability, the level-of-processing theory, we present a sparse coding model for compressing feature embeddings of images, and show that the reconstruction residuals from this model predict how well images are encoded into memory. In an open memorability dataset of scene images, we show that reconstruction error not only explains memory accuracy, but also response latencies during retrieval, subsuming, in the latter case, all of the variance explained by powerful vision-only models. We also confirm a prediction of this account with 'model-driven psychophysics'. This work establishes reconstruction error as an important signal interfacing perception and memory, possibly through adaptive modulation of perceptual processing.

So much of what we remember is not the result of intentional selection, but rather the result of simply perceiving. How are perceptual experiences cast into memory? And how does perceiving exert control over remembering? These are fundamental questions in the study of the mind with multiple lines of empirical and theoretical studies designed to uncover the interface between perception and memory[1–11]. A striking illustration of the extent to which perception influences memory is the recent demonstration of 'memorability', the finding that some images are systematically more memorable than others across observers[12,13]. The formation of new visual memory traces must recruit both visual and memory-related functions, but the computational basis of how they interact to produce memory traces remains poorly understood.

Existing computational accounts, inspired by the demonstration of image memorability, largely consider models that involve vision-only computations, such as deep convolutional neural networks (DCNN) trained for image classification[14–17]. These studies have established a quantitative relationship between the summary statistics derived from the later stages of these networks (for example, the magnitude of activations of a given layer) and memorability scores of images. Interestingly, this effect is also observed neurally: the population response magnitude of the inferior temporal cortex neurons tracks the memorability scores of the presented images. However, these vision-only models do not attempt to formalize processes responsible for transforming percepts into memories and thus remain incomplete as computational accounts of how perceptual processing relates to memory traces.

A classic psychological account, the 'level-of-processing' theory of Craik and Lockart[3], has attempted to directly address this interface, proposing that a memory trace is a by-product of the perceptual analysis of incoming sensory signals and that a 'deeper' analysis is associated with better retention in memory. However, this framework, including a specification of what determines the depth (or level) of perceptual processing[18–20], remains largely qualitative and non-computational. To date, all empirical demonstrations of the level-of-processing effect

[1]Department of Psychology, Yale University, New Haven, CT, USA. [2]Center for Brain Science, RIKEN, Wako, Japan. [3]Department of Statistics & Data Science, Yale University, New Haven, CT, USA. [4]Wu-Tsai Institute, Yale University, New Haven, CT, USA. [5]These authors contributed equally: Qi Lin, Zifan Li. [6]These authors jointly supervised this work: John Lafferty, Ilker Yildirim. ✉e-mail: qi.lin@riken.jp; john.lafferty@yale.edu; ilker.yildirim@yale.edu

https://doi.org/10.1038/s41562-024-01870-3

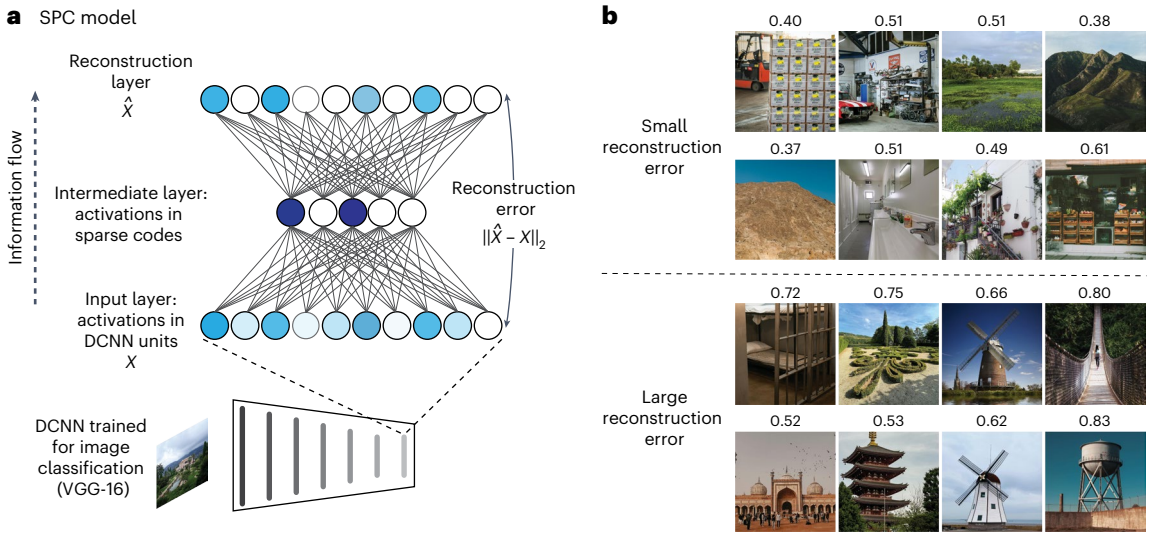

Fig. 1 | Model architecture and example images. a, Schematic representation of how reconstruction errors are quantified using a SPC model. The SPC model example shown here is based on layer 7 of a DCNN trained for image classification—the second fully connected layer from the VGG-16 network[34]. **b**, Example images that are easy and hard for the SPC to reconstruct. The numbers above the original images are the memorability scores (hit rates − false alarm rates) as measured in ref. 12. For this figure, all images were replaced by images of similar appearance from the public domain. Some images were cropped. Images are from Pexels and Pixabay.

rely on the use of an orienting task (as reviewed in ref. 21), completely missing the automatic (that is, spontaneous and stimulus-driven) nature of perceptual processing. In fact, work investigating how orientating tasks interact with image memorability has demonstrated that the effect of such orienting tasks is independent of memorability[22]. A computational form of the level-of-processing theory must take into account perceptual processing more rigorously and address how its depth can be modulated on an image-by-image basis.

By addressing the elemental computations thought to underlie memory—compression[23] and reconstruction[24,25]—we present a new computational model that yields a stimulus-driven, quantitative measure of how perceptual processing can impact memory formation. This combines the vision-only models mentioned above with a sparse coding (SPC) framework, a classic architecture used for compressing information in both computational neuroscience[26–29] and machine learning[30–33]. Our model operates on the activations evoked by natural scene images in a DCNN trained to categorize scenes and objects[34,35] and learns how to reconstruct these evoked activations. When considered over the entire space of signals to be compressed, reconstruction error, measured as the difference between the signals recovered from compressed codes and the uncompressed signal, provides a benchmark for evaluating different codes for lossy compression[36–39].

We hypothesize that reconstruction error provides the necessary computational substrate for gauging and modulating the level of perceptual analysis[3], and thus impacts memory strengths. In particular, reconstruction error resulting from the SPC model provides a principled signal to determine how much more processing might be warranted on an image-by-image basis. Ideally, a valid computational account addressing the perception-to-memory interface should capture different aspects of memory behaviour, targeting not just the accuracy of retrieval, but also systematic variance in terms of latencies during retrieval (that is, response time), in addition to predicting measurable new phenomena.

Across three studies, this work aspires to this ideal by testing our SPC model's ability to capture how sensory signals are transformed into memory, above and beyond what can already be explained by vision-only models. In Studies 1 and 2, we focus on two well-established and complimentary measures of memory strength: memory accuracy and response times during retrieval[40]. To this end, we relate the reconstruction errors of images (obtained using the SPC model) to their memorability scores and response times measured in a large-scene memorability dataset[12] while taking into account what can be explained by a standard DCNN trained for image classification (VGG-16)[34]. We find that reconstruction error explains additional variance in both memory accuracy (Study 1) and response times (Study 2), subsuming all of the variance explained by other models in the latter case. In Study 3, we turn to 'model-driven psychophysics' and predict that the architectural differences between the SPC model and DCNNs would be paralleled in the brain as temporally and functionally distinct processes. In a preregistered experiment, we manipulate encoding times in a rapid serial visual presentation (RSVP) paradigm and observe that images with large reconstruction error benefit more from longer encoding times, while controlling for DCNN-driven memorability effects. Together, these results establish compression-based reconstruction error as a previously unrecognized driver of memorability and suggest a mechanism in which such reconstruction error modulates the depth of encoding of the incoming visual inputs.

Results

Efficient compression by minimizing reconstruction error

To model the process of compressing sensory activations underlying our visual experiences, we built a SPC model to reconstruct the activations in a given layer of a DCNN pretrained to classify images of scenes and objects[34] based on the commonly used VGG-16 architecture[35]. The SPC model (see Fig. 1a for a schematic representation) consists of three layers: an input layer of 1,000 units, an intermediate layer of 500 units for recoding the inputs, and a reconstruction layer of 1,000 units for reproducing the input. We sampled seven DCNN layers (ranging from early to late) from the VGG architecture and trained a separate SPC model for each. Because the dimensionality of DCNN layers is typically very high, we designed the input layer of the SPC (and thus, the reconstruction layer) to consist of 1,000 units randomly sampled from the corresponding DCNN layer. This design also ensured that the number of parameters to be trained in the SPC model remains constant despite the fact that the dimensionality of layers varies across the DCNN architecture. Our SPC model is trained on natural scene images (10,263 images in total) from a publicly available dataset[12], allowing the model to learn an efficient code for reconstructing DCNN activations for complex

visual inputs. Critically, unlike the training of DCNN where the goal is to maximize classification accuracy, the objective of SPC training is to minimize the reconstruction error (the Euclidean distance between the reproduced and original activations, subject to a sparsity term on the intermediate recoding layer) (Fig. 1a and Methods).

Harder-to-reconstruct images are more memorable

After training, sparse coding yields substantial variability in how well individual images can be reconstructed (Supplementary Fig. 1). In Fig. 1b, we show example images that are easy and hard for a representative SPC model (trained on the layer 7 activations in the DCNN) to reconstruct. We observe that the hard-to-reconstruct images tend to be more object-centred or contain humans, relative to easy-to-reconstruct images. Interestingly, such image attributes have been related to memorability behaviour in earlier work[12,41]. In fact, even from this subset of images selected solely based on their reconstruction errors, we observe that hard-to-reconstruct images are often more memorable (memorability scores indicated above each image). (The memorability score of a given image is calculated as the hit rate minus the false alarm rate, as measured in Isola et al.[12]. Hit rate is calculated as the proportion of observers who correctly indicated that the image was repeated when the image was shown to them a second time. False alarm rate is calculated as the proportion of observers who incorrectly indicated that the image was repeated when the image was shown to them for the first time. In other words, a higher score means an image is more memorable after accounting for general familiarity.) Based on these observations, we hypothesize that images that are harder to reconstruct are also more memorable.

To establish this prediction quantitatively, we focused on the 2,221 target images with available memorability scores measured in Isola et al.[12]. For each of the 2,221 target images, we obtained a reconstruction error from each of the 7 trained SPC models. For each SPC model, we then correlated the resulting reconstruction errors and memorability scores of the corresponding images. As shown in Fig. 2a (left), reconstruction errors from all sampled layers in SPC were significantly related to memorability: images with a larger reconstruction error are more memorable. Layers 5 (conv5) and 7 (fc2) showed numerically the strongest effects (layer 1: $r = 0.15$ [0.11, 0.19], $P < 0.001$ where r is Pearson's r, [l,u] are the upper and lower 95% confidence intervals (CI) and the P value was generated using two-tailed direct bootstrap hypothesis testing with 1,000 iterations and corrected for multiple comparisons; layer 2: $r = 0.16$ [0.12, 0.20], $P < 0.001$; layer 3: $r = 0.24$ [0.20, 0.28], $P < 0.001$; layer 4: $r = 0.28$ [0.25, 0.33], $P < 0.001$; layer 5: $r = 0.33$ [0.30, 0.37], $P < 0.001$; layer 6: $r = 0.25$ [0.22, 0.29], $P < 0.001$; layer 7: $r = 0.30$ [0.26, 0.33], $P < 0.001$).

Because the SPC models were trained based on the activations from a feedforward VGG-16 DCNN pretrained to classify scenes and objects[34], we next wanted to understand the extent to which the resulting reconstruction errors are just capturing the same variance in memorability as previous measures derived from purely feedforward architectures[14,16,17]. To this end, we derived a predictor from the DCNN network by sampling the same seven layers used to train the SPC models. Following Lin et al.[16], for each layer and target image, we calculated the distinctiveness as the Euclidean distance between each target image and its nearest neighbour with respect to the DCNN's feature space at this layer. Replicating similar results[14], we found that distinctiveness in the DCNN across all sampled layers was significantly related to memorability, with the later layers (layers 4–7) showing the numerically strongest effects (Fig. 2b, left; layer 1: $r = 0.14$ [0.10, 0.18], $P < 0.001$; layer 2: $r = 0.14$ [0.11, 0.19], $P < 0.001$; layer 3: $r = 0.22$ [0.18, 0.26], $P < 0.001$; layer 4: $r = 0.29$ [0.25, 0.33], $P < 0.001$; layer 5: $r = 0.36$ [0.32, 0.40], $P < 0.001$; layer 6: $r = 0.30$ [0.26, 0.34], $P < 0.001$; layer 7: $r = 0.28$ [0.24, 0.32], $P < 0.001$).

After establishing that both reconstruction error and distinctiveness are significantly correlated with memorability, we next asked: does reconstruction error capture additional variance in memorability, above and beyond what was already explainable by the DCNN's feature hierarchy optimized for image classification? To address this question, we chose distinctiveness from layer 5 as our primary measure of distinctiveness because it showed the highest correlation with memorability (Fig. 2b, left panel). We then compared distinctiveness at that layer with reconstruction error as well as another standard measure also derived from bottom-up feature hierarchy that shows correlation with memorability: the L2-Norm of activations (square-root of the sum of squares of elements in the activation vector) in a DCNN layer given an image[14]. However, we found that distinctiveness was always better correlated with memorability scores and these correlations were less sensitive to the specific choice of layer, relative to L2-Norm (Supplementary Fig. 2). Therefore, we chose to focus on distinctiveness instead of L2-Norm. We also considered a combination of both DCNN-based measures—distinctiveness and L2-Norm—by normalizing the reconstruction error with L2-Norm before submitting to partial correlation analysis with distinctiveness. As shown in Supplementary Figs. 4 and 5, our results remained qualitatively similar. First, we observed that although distinctiveness and reconstruction error were correlated ($r = 0.82$), this correlation was not perfectly co-linear and was significantly less than the correlation between the two DCNN-derived measures, distinctiveness and L2-Norm ($r = 0.99$; two-tailed Williams' t-test on the difference between the two correlations: $t(2218) = 73.67$, $P < 0.001$, Cohen's q (the difference between two Fisher-transformed correlation values) = 1.49, 95% CI = [0.15, 0.18]; Supplementary Fig. 3). These results suggest that the SPC model and the DCNN might capture different aspects of the computations underlying the memorability behaviour.

Second, we performed partial regression analysis to directly test for the unique contribution of reconstruction error to explaining memorability. Specifically, we residualized layer 5 distinctiveness (the DCNN layer that is most predictive of behaviour under distinctiveness) from both memorability and reconstruction error. We found that reconstruction error residuals continue to explain significant variance in memorability (Fig. 2a, right): reconstruction error from layers 5 and 7 in SPC was still significantly correlated with memorability, after accounting for what can be explained solely by distinctiveness (layer 1: $r = -0.02$ [−0.05, 0.01], $P > 0.999$; layer 2: $r = -0.02$ [−0.05, 0.01], $P > 0.999$; layer 3: $r = 0.00$ [−0.03, 0.04], $P > 0.999$; layer 4: $r = 0.04$ [0.00, 0.07], $P = 0.420$; layer 5: $r = 0.08$ [0.04, 0.11], $P < 0.001$; layer 6: $r = 0.03$ [−0.01, 0.06], $P > 0.999$; layer 7: $r = 0.16$ [0.13, 0.19], $P < 0.001$).

Even though we are primarily interested in evaluating whether reconstruction error can explain additional variance, for completeness, we also ran a similar analysis by regressing out the layer 5 reconstruction error (which is the most predictive reconstruction error measure of memorability of all the layers) from both memorability and distinctiveness. As shown in Fig. 2b (right), distinctiveness from layers 3–7 also remained predictive of memorability after controlling for reconstruction error (layer 1: $r = 0.01$ [−0.02, 0.04], $P > 0.999$; layer 2: $r = -0.01$ [−0.02, 0.04], $P > 0.999$; layer 3: $r = 0.06$ [0.03, 0.09], $P < 0.001$; layer 4: $r = 0.10$ [0.06, 0.13], $P < 0.001$; layer 5: $r = 0.16$ [0.13, 0.20], $P < 0.001$; layer 6: $r = 0.12$ [0.08, 0.15], $P < 0.001$; layer 7: $r = 0.20$ [0.17, 0.23], $P < 0.001$), again suggesting that distinctiveness and reconstruction error are capturing separable aspects of the variance in memorability.

Together, these results demonstrate that images with harder-to-reconstruct DCNN activations are more memorable and that reconstruction error makes an additional contribution to image memorability, above and beyond distinctiveness.

Harder-to-reconstruct images are retrieved faster

Previous work that explicitly manipulates the depth of encoding with different orienting tasks found that a deeper level of encoding is associated with faster reaction times during retrieval[22,42,43]. Thus, if our results from Study 1 have to do with a mechanism in which

https://doi.org/10.1038/s41562-024-01870-3

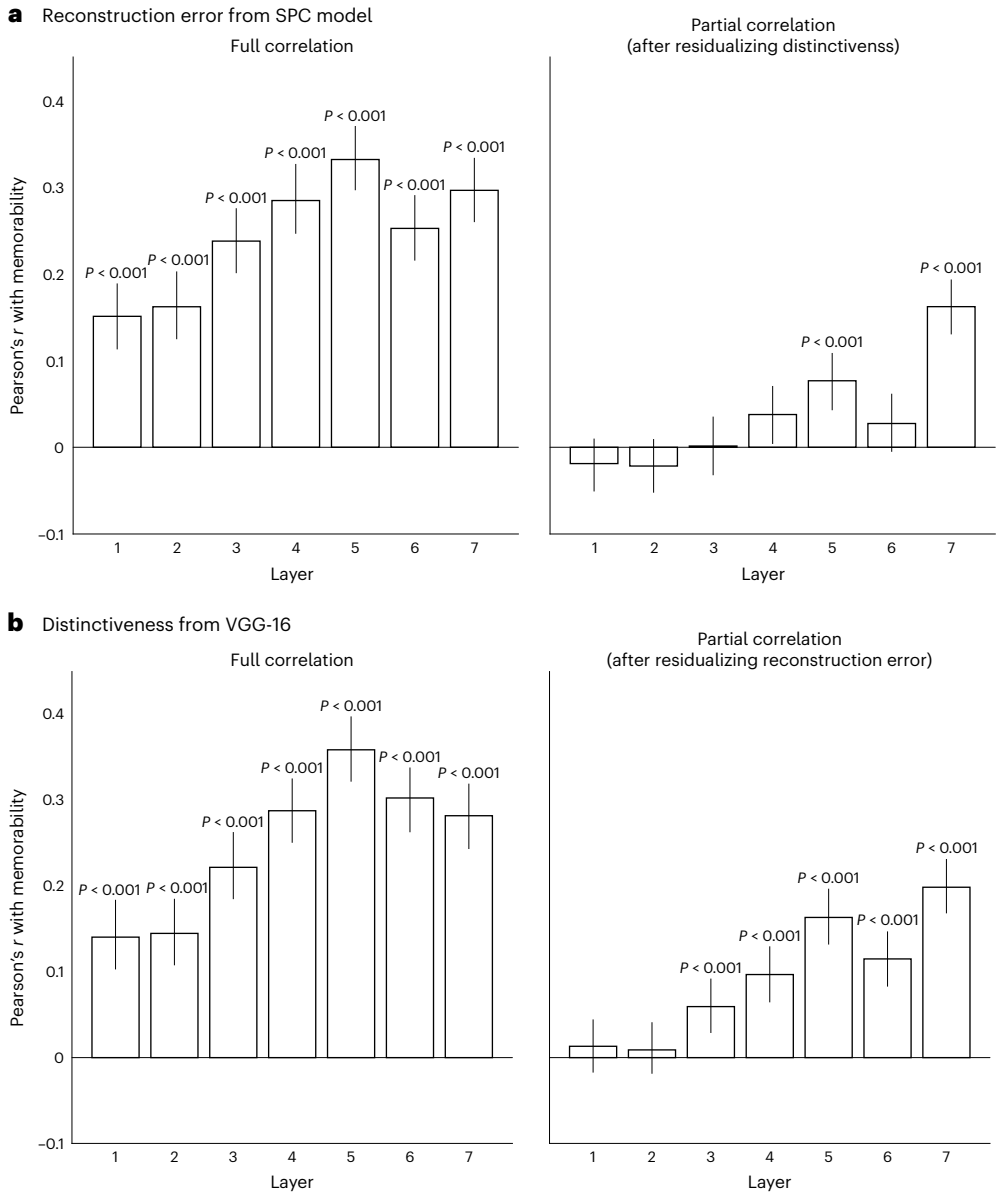

Fig. 2 | Images with a large reconstruction error are more memorable (N_{images} = 2,221). a, Pearson's *r* between memorability and reconstruction error (layer 1: *r* = 0.15 [0.11, 0.19], *P* < 0.001; layer 2: *r* = 0.16 [0.12, 0.20], *P* < 0.001; layer 3: *r* = 0.24 [0.20, 0.28], *P* < 0.001; layer 4: *r* = 0.28 [0.25, 0.33], *P* < 0.001; layer 5: *r* = 0.33 [0.30, 0.37], *P* < 0.001; layer 6: *r* = 0.25 [0.22, 0.29], *P* < 0.001; layer 7: *r* = 0.30 [0.26, 0.33], *P* < 0.001) and partial Pearson's *r* after accounting for distinctiveness (layer 1: *r* = −0.02 [−0.05, 0.01], *P* > 0.999; layer 2: *r* = −0.02 [−0.05, 0.01], *P* > 0.999; layer 3: *r* = 0.00 [−0.03, 0.04], *P* > 0.999; layer 4: *r* = 0.04 [0.00, 0.07], *P* = 0.420; layer 5: *r* = 0.08 [0.04, 0.11], *P* < 0. 001; layer 6: *r* = 0.03 [−0.01, 0.06], *P* > 0.999; layer 7: *r* = 0.16 [0.13, 0.19], *P* < 0.001). **b**, Pearson's *r* between memorability and distinctiveness (layer 1: *r* = 0.14 [0.10, 0.18], *P* < 0.001; layer 2: *r* = 0.14 [0.11, 0.19], *P* < 0.001; layer 3: *r* = 0.22 [0.18, 0.26], *P* < 0.001; layer 4: *r* = 0.29 [0.25, 0.33], *P* < 0.001; layer 5: *r* = 0.36 [0.32, 0.40], *P* < 0.001; layer 6: *r* = 0.30 [0.26, 0.34], *P* < 0.001; layer 7: *r* = 0.28 [0.24, 0.32], *P* < 0.001) and partial Pearson's *r* after accounting for reconstruction error (layer 1: *r* = 0.01 [−0.02, 0.04], *P* > 0.999; layer 2: *r* = −0.01 [−0.02, 0.04], *P* > 0.999; layer 3: *r* = 0.06 [0.03, 0.09], *P* < 0.001; layer 4: *r* = 0.10 [0.06, 0.13], *P* < 0.001; layer 5: *r* = 0.16 [0.13, 0.20], *P* < 0.001; layer 6: *r* = 0.12 [0.08, 0.15], *P* < 0.001; layer 7: *r* = 0.20 [0.17, 0.23], *P* < 0.001). Error bars represent 95% CI from 1,000 bootstrapping iterations. Statistical comparisons are made using direct bootstrap hypothesis testing based on a two-tailed test threshold. *P* values have been corrected for multiple comparisons with Bonferroni correction (α = 0.05/14).

reconstruction error modulates the depth of encoding, then we predict that harder-to-reconstruct images will be retrieved more quickly.

To this end, we analysed the response time data from the correct recognition trials in Isola et al.[12] (see Methods for details on data inclusion criteria). Indeed, we found that reconstruction errors from all seven layers in our SPC model were negatively correlated with response times during retrieval (Fig. 3a, left), such that harder-to-reconstruct images are faster to retrieve (layer 1: *r* = −0.08 [−0.12, −0.04], *P* < 0.001; layer 2: *r* = −0.08 [−0.12, −0.04], *P* < 0.001; layer 3: *r* = −0.11 [−0.15, −0.06], *P* < 0.001; layer 4: *r* = −0.13 [−0.17, −0.08], *P* < 0.001; layer 5: *r* = −0.22 [−0.26, −0.17], *P* < 0.001; layer 6: *r* = −0.33 [−0.37, −0.29], *P* < 0.001; layer 7: *r* = −0.36 [−0.40, −0.32], *P* < 0.001). We observed a similar pattern, albeit to a lesser degree and only significant for layers 3–6, when we tested distinctiveness (Fig. 3b, left; layer 1: *r* = −0.06 [−0.10, −0.02], *P* = 0.066; layer 2: *r* = −0.05 [−0.09, 0.00], *P* = 0.462; layer 3: *r* = −0.07 [−0.11, −0.02], *P* < 0.001; layer 4: *r* = −0.10 [−0.14, −0.05], *P* < 0.001; layer 5: *r* = −0.18 [−0.23, −0.14], *P* < 0.001; layer 6: *r* = −0.22 [−0.26, −0.18], *P* < 0.001; layer 7: *r* = −0.06 [−0.10, −0.01], *P* = 0.154). To dissociate the contributions of reconstruction error and distinctiveness in explaining the variance in response times during

https://doi.org/10.1038/s41562-024-01870-3

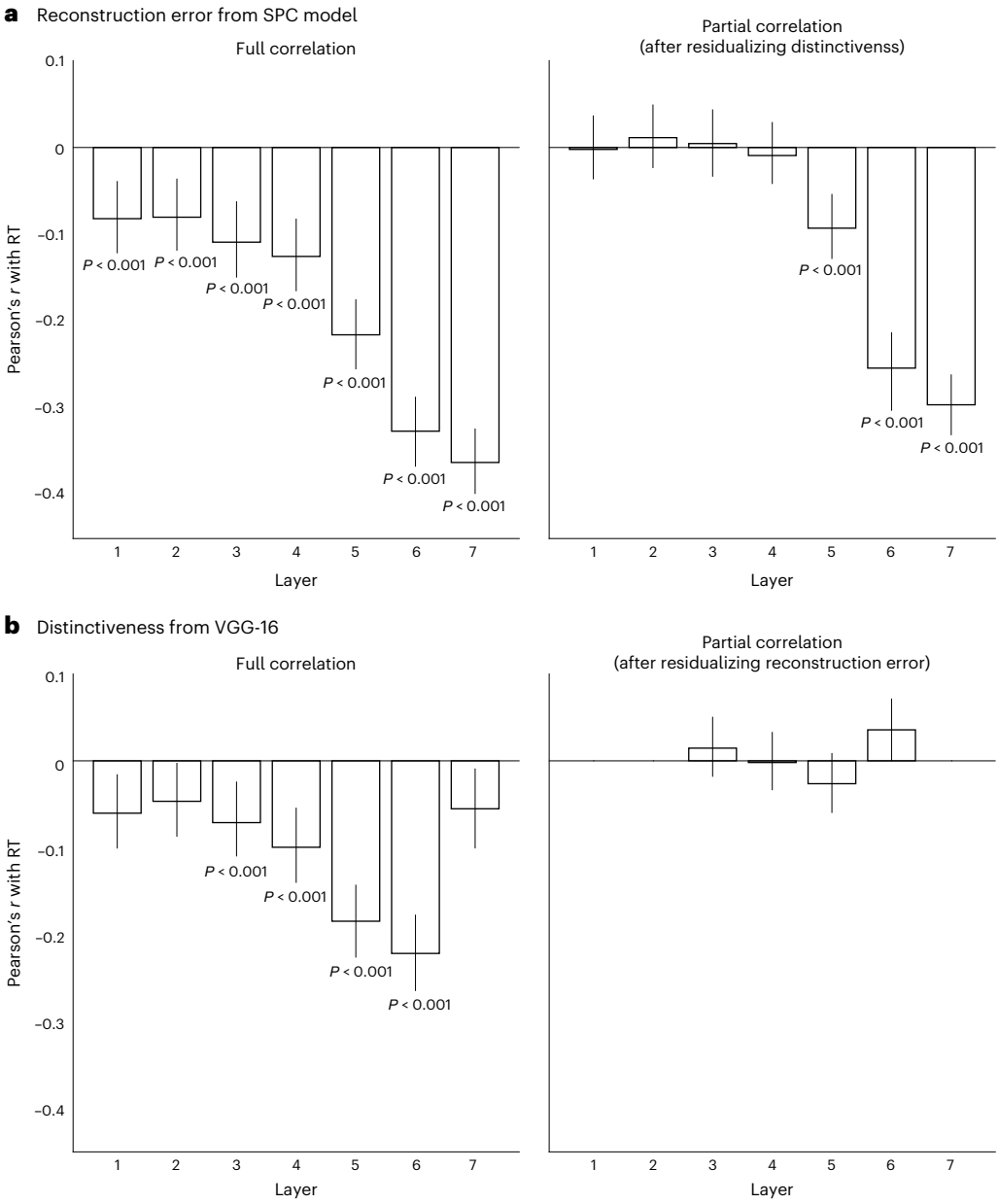

Fig. 3 | Images with a large reconstruction error are recognized faster during retrieval (N_{images} = 2,221). a, Pearson's *r* between response times during retrieval and reconstruction error (layer 1: $r = -0.08$ [−0.12, −0.04], $P < 0.001$; layer 2: $r = -0.08$ [−0.12, −0.04], $P < 0.001$; layer 3: $r = -0.11$ [−0.15, −0.06], $P < 0.001$; layer 4: $r = -0.13$ [−0.17, −0.08], $P < 0.001$; layer 5: $r = -0.22$ [−0.26, −0.17], $P < 0.001$; layer 6: $r = -0.33$ [−0.37, −0.29], $P < 0.001$; layer 7: $r = -0.36$ [−0.40, −0.32], $P < 0.001$) and partial Pearson's *r* after accounting for distinctiveness (layer 1: $r = 0.00$ [−0.04, 0.04], $P > 0.999$; layer 2: $r = 0.01$ [−0.02, 0.05], $P > 0.999$; layer 3: $r = 0.00$ [−0.03, 0.04], $P > 0.999$; layer 4: $r = -0.01$ [−0.04, 0.03], $P > 0.999$; layer 5: $r = -0.09$ [−0.13, −0.05], $P < 0.001$; layer 6: $r = -0.25$ [−0.30, −0.21], $P < 0.001$; layer 7: $r = -0.30$ [−0.33, −0.26], $P < 0.001$). **b**, Pearson's *r* between response times during retrieval and distinctiveness (layer 1: $r = -0.06$ [−0.10, −0.02], $P = 0.066$; layer 2: $r = -0.05$ [−0.09, 0.00], $P = 0.462$; layer 3: $r = -0.07$ [−0.11, −0.02], $P < 0.001$; layer 4: $r = -0.10$ [−0.14, −0.05], $P < 0.001$; layer 5: $r = -0.18$ [−0.23, −0.14], $P < 0.001$; layer 6: $r = -0.22$ [−0.26, −0.18], $P < 0.001$; layer 7: $r = -0.06$ [−0.10, −0.01], $P = 0.154$) and partial Pearson's *r* after accounting for reconstruction error (layer 3: $r = 0.01$ [−0.02, 0.05], $P > 0.999$; layer 4: $r = 0.00$ [−0.03, 0.03], $P > 0.999$; layer 5: $r = -0.03$ [−0.06, 0.01], $P > 0.999$; layer 6: $r = 0.04$ [0.00, 0.07], $P = 0.594$). Note that partial correlation was only performed if the full correlation turned out to be statistically significant. Therefore, in the bottom right plot, there were no partial correlation results for layers 1, 2 and 7 distinctiveness. Error bars represent 95% CI from 1,000 bootstrapping iterations. Statistical comparisons are made using direct bootstrap hypothesis testing based on a two-tailed test threshold. *P* values have been corrected for multiple comparisons with Bonferroni correction (reconstruction error: $\alpha = 0.05/14$; distinctiveness: $\alpha = 0.05/11$).

retrieval, we again used partial regressions as in Study 1. We focused on model layers in which we observed a significant correlation in the full correlation analysis and therefore did not perform partial correlation for layers 1, 2 and 7 distinctiveness. In the SPC model, the negative relationship between response times and reconstruction error on layers 5–7 remained significant (layer 1: $r = 0.00$ [−0.04, 0.04], $P > 0.999$; layer 2: $r = 0.01$ [−0.02, 0.05], $P > 0.999$; layer 3: $r = 0.00$ [−0.03, 0.04], $P > 0.999$; layer 4: $r = -0.01$ [−0.04, 0.03], $P > 0.999$; layer 5: $r = -0.09$ [−0.13, −0.05], $P < 0.001$; layer 6: $r = -0.25$ [−0.30, −0.21], $P < 0.001$; layer 7: $r = -0.30$ [−0.33, −0.26], $P < 0.001$) after we regressed out layer 6 distinctiveness (the most predictive measure of response times in the DCNN) from both response times and reconstruction errors. By contrast, distinctiveness decoupled from behaviour (layers 3–6) after we regressed out layer 7 reconstruction errors (the most predictive

https://doi.org/10.1038/s41562-024-01870-3

Fig. 4 | Example images from each of the four groups with different distinctiveness–reconstruction error profiles. For this figure, all images were replaced by images of similar appearance from the public domain. Some images were cropped. Image by frimufilms on Freepik (red steps). Image by Freepik (people eating and orange washing machines). Other images are from Pexels and Pixabay.

measure of response times in the SPC model) from both response times and distinctiveness (layer 3: $r = 0.01$ [−0.02, 0.05], $P > 0.999$; layer 4: $r = 0.00$ [−0.03, 0.03], $P > 0.999$; layer 5: $r = -0.03$ [−0.06, 0.01], $P > 0.999$; layer 6: $r = 0.04$ [0.00, 0.07], $P = 0.594$).

These results demonstrate the specificity of the relationship between reconstruction error and response time during retrieval. That is, although both larger distinctiveness and larger reconstruction error predict higher recognition accuracy, only reconstruction error predicts retrieval efficiency. Beyond a mere measure of visual processing, this finding is consistent with our hypothesis that the magnitude of reconstruction error directly modulates the process, in particular the depth, of encoding percepts into memory.

Harder-to-reconstruct images benefit more from longer exposure

Studies 1 and 2 revealed reconstruction error as a previously unrecognized source of image memorability and retrieval efficiency using a publicly available dataset. These results suggest that reconstruction error might be modulating the process of encoding itself, in terms of how deeply an image should be encoded. To further establish that reconstruction error is associated with a functionally distinct set of computations in the mind and brain, beyond the vision-only computations implemented in DCNN, we turn to a prediction made by our modelling framework and test it in a 'model-driven psychophysics' experiment in Study 3. Our logic is that images whose activations are harder to reconstruct and thus require a deeper level of processing will fetch additional mental resources, manifested as longer encoding times and increased memory accuracy[3]. Therefore, we predict that the memory for images with higher reconstruction errors will benefit more from longer encoding times.

To test this possibility, we started by sampling images with divergent profiles of distinctiveness and reconstruction error. Following Study 1, which focuses on memory accuracy, we used layer 5 distinctiveness as our primary measure of distinctiveness and then selected layer 7 reconstruction error from the SPC model as our primary measure of reconstruction error because this measure captures the largest amount of additional variance in memorability after accounting for distinctiveness (partial Pearson's $r = 0.16$ [0.13, 0.19], $P < 0.001$; Fig. 2a right). (To make sure that the layer 7 reconstruction error is not just capturing the memorability driven by layer 7 distinctiveness, we also tested including both layer 5 and layer 7 distinctiveness measures in the partial regression model. Even under this stringent way of controlling for distinctiveness, layer 7 reconstruction error continued to significantly capture additional variance in memorability (partial Pearson's $r = 0.11$ [0.06, 0.14], $P < 0.001$).) After settling on the primary measures of distinctiveness and reconstruction error, we then sampled four different groups of 48 images each: images with (1) large distinctiveness and large reconstruction error, (2) large distinctiveness and small reconstruction error, (3) small distinctiveness and large reconstruction error, and (4) small distinctiveness and small reconstruction error. 'Large' is defined as falling within top 30 percentile of the target measure (distinctiveness or reconstruction error) and 'small' is defined as bottom 30 percentile. Figure 4 gives example images in each of the four groups.

We adopted a within-participant design with 2 distinctiveness levels (large versus small) × 2 reconstruction error levels (large versus small) × 3 encoding durations (34, 84 or 167 ms). Each of the resulting 12 conditions was presented as a separate block for each participant. During each block, half of the trials were target-present trials (the test image was presented in the RSVP stream), whereas the other half were target-absent trials (the test image was not presented in the RSVP stream). On each trial, participants first saw an RSVP stream of images (Fig. 5a; following the trial structure of Broers et al.[44]). They were then shown a test image and had to indicate whether the test image was presented in the RSVP stream or not. The experiment design and sample size were preregistered (https://aspredicted.org/MFM_R22). Forty-five participants completed the experiment online via Prolific. We calculated hit rate for each of the 12 conditions (Fig. 5b). (We note that the target images were never used as foils in this RSVP experiment so we could not calculate false alarm rates for them).

In Fig. 5b, we can see that although memory accuracy for all images increased with longer encoding times, images with larger reconstruction error benefited more from the longer encoding times (as indicated by the steeper slopes). A preregistered three-way repeated measures analysis of variance (ANOVA) confirmed these qualitative observations. In addition to the main effects of reconstruction error, distinctiveness and encoding duration, there was a significant interaction between reconstruction error and encoding duration. That is, images with large reconstruction error benefited significantly more from longer encoding duration, relative to images with small reconstruction error (Supplementary Table 1).

To directly measure the effect of reconstruction error on memory, following the procedures stated in our preregistration, we compared regression slopes relating encoding times to hit rates between the images with large versus small reconstruction error. We bootstrapped

https://doi.org/10.1038/s41562-024-01870-3

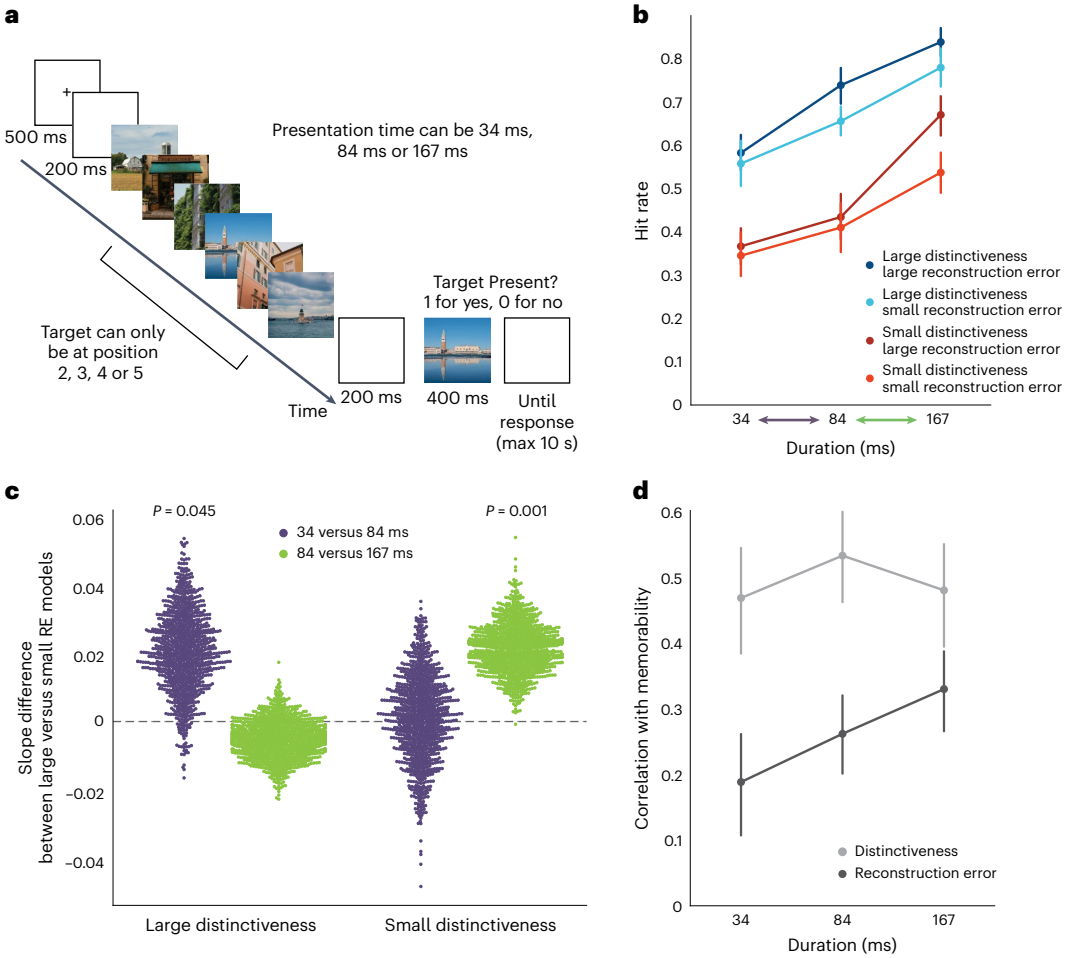

Fig. 5 | Images with harder-to-reconstruct representations benefit more from longer encoding times ($N_{participants}$ = 45). a, An example trial from the RSVP experiment. **b**, Hit rates for each of the four image groups. Error bars represent 95% CI from 1,000 bootstrapping iterations. **c**, Difference between the regression slopes of hit rate versus time across images with large versus small reconstruction error (RE), conducted separately for each sequential pair of encoding times (indicated by colours; also highlighted in **b**, *x* axis) and distinctiveness level. Distributions represent the slope differences from 1,000 bootstrapping iterations. Large distinctiveness (34 versus 84 ms): mean Δ slope = 0.020, one-tailed $P = 0.045$, 95% CI [−0.003, 0.043]; large distinctiveness (84 versus 167 ms): mean Δ slope = −0.005, one-tailed $P = 0.813$, 95% CI [−0.016, 0.006]; small distinctiveness (34 versus 84 ms): mean Δ slope = 0.001, one-tailed $P = 0.449$, 95% CI [−0.024, 0.026]; small distinctiveness (84 versus 167 ms): mean Δ slope = 0.022, one-tailed $P = 0.001$, 95% CI [0.008, 0.036]. **d**, Pearson's *r* between distinctiveness/reconstruction error and hit rates across different presentation times. Distinctiveness: mean slope = −0.050, one-tailed $P = 0.526$, 95% CI [−0.962, 0.909]; reconstruction error: mean slope = 0.940, one-tailed $P < 0.001$, 95% CI [0.702, 1.000]. Error bars represent 95% CI from 1,000 bootstrapping iterations. For this figure, all images were replaced by images of similar appearance from the public domain. Some images were cropped. Images are from Pexels.

the difference of these regression slopes between reconstruction error levels ($\beta_{LargeRE} - \beta_{SmallRE}$) separately for each distinctiveness level and sequential pair of encoding times (early: 34 ms versus 84 ms; late: 84 ms versus 167 ms) (Methods). This analysis not only confirmed that images with larger reconstruction error indeed benefit more from longer encoding times, but also revealed a finer-grained temporal picture of this effect. As shown in Fig. 5c, for images with large distinctiveness, the memory benefit of large reconstruction error was observed earlier (when encoding time was increased from 34 ms to 84 ms; mean Δ slope (the difference between the two slopes) = 0.020, one-tailed $P = 0.045$, 95% CI [−0.003, 0.043] using 1,000 iterations of bootstrapping), whereas for images with small distinctiveness, this reconstruction error benefit was only observed later (when encoding time was increased from 84 ms to 167 ms; mean Δ slope = 0.022, one-tailed $P = 0.001$, 95% CI [0.008, 0.036] using 1,000 iterations of bootstrapping). Note that here, because we had clear prediction about the direction of the effects based on results from Study 1 and 2, one-tailed *P* values were reported (also stated clearly in the preregistration).

Although Study 3 was conceived to reveal the differential effects of distinctiveness and reconstruction error in a categorical design (for example, large distinctiveness and small reconstruction error), we next turned to a finer-grained and non-preregistered analysis to ask whether the ability of these models to predict behaviour depends on time. We predict that if the memory benefit of reconstruction error is the result of deeper processing and thus requires more time to complete the computation, reconstruction error should account for more and more variance in memory as exposure time increases. To this end, for each duration, we correlated the behavioural hit rates at the level of individual images with their corresponding distinctiveness and reconstruction error. Indeed, as shown in Fig. 5d, we observed that the correlations between reconstruction error and hit rate increase monotonically as a function of exposure time (mean slope = 0.940, one-tailed $P < 0.001$, 95% CI [0.702, 1.000]; *P* value and 95% CI generated from 1,000 iterations of bootstrapping) (Methods), whereas this pattern is absent for distinctiveness (mean slope = −0.050, one-tailed $P = 0.526$, 95% CI [−0.962, 0.909]).

Together, these behavioural results further demonstrate that reconstruction error contributes to memorability by modulating encoding depth such that the memory benefit driven by reconstruction error manifests itself only when given enough encoding time. Moreover, distinctiveness-driven and reconstruction error-driven memorability effects exhibit differential temporal profiles and thus probably reflect functionally distinct processes in the mind and brain.

Discussion

We present a computational model that combines sparse coding with recent vision models based on DCNNs and gives rise to a new quantitative measure—compression-based reconstruction error—of how perception modulates the strength of memory traces. In Studies 1 and 2, we show that reconstruction error predicts both memory accuracy and response times at the level of individual images. Critically, reconstruction error explains additional variance in both aspects of memory performance, beyond what can be explained by vision-only models[14,16], including capturing all of the model-explained variance in response times. To further demonstrate that the modulation of memorability because of reconstruction error reflects a separate process relative to vision-only models, we run an RSVP experiment and show that the effects of distinctiveness and reconstruction error on memorability have distinct temporal profiles. Memory performance increases to a more substantial degree for images that are harder to reconstruct (controlling for distinctiveness). Finally, as exposure time increases, allowing a deeper analysis of the relevant scene features when necessary, reconstruction error becomes increasingly predictive of memory, whereas the effect of distinctiveness stays more or less constant across durations. Together, these results not only establish compression-based reconstruction error as a previously unrecognized driver of memorability, but also suggest a perception-to-memory interface in which such reconstruction error modulates the encoding depth of incoming visual inputs.

A signal for modulating depth of encoding

By formalizing the interface between perception and memory with the SPC model, our work places the level-of-processing theory in a new light. In the original and revised versions of the theory[3,20,45], encoding depth has always been conceptualized as a continuum. Yet, to date, all empirical demonstrations of the effect of encoding depth on memory remain qualitative, through manipulating encoding depth with different orienting tasks (for example, paying attention to the physical versus semantic property of the stimuli; for a review, see ref. 21). By demonstrating that reconstruction error, at the level of images, predicts memory accuracy, response latency during retrieval and the need for more encoding time, we suggest that reconstruction error can serve as an indication of the resource-rational depth of encoding, filling in a longstanding explanatory gap in the theory.

Multiple mechanisms are possible for how reconstruction error can affect perceptual processing, leading to deeper or shallower analysis, and thus memory strength. One possibility is suggested by the predictive coding framework[46], a version of which suggests the existence of predictive auto-encoders that could compute the compressibility and reconstruction error of incoming information[47]. Another possibility is related to the idea of 'analysis-by-synthesis', which posits the reconstruction of sensory activations via a process of 'synthesis' to be part of perceptual processing[48–51]. The current study suggests that such a synthesis process may also provide a learning signal for memory.

We believe that the adaptive deployment of these perceptual mechanisms, alongside purely feedforward computations (as implemented in standard DCNNs as we tested here), can be temporally mapped out with increasingly greater resolution. The RSVP results presented in Study 3 are broadly consistent with a behavioural report finding that the difference between memorable and forgettable images emerges at presentation times as brief as 13 ms (ref. 44). In our study, the effect of distinctiveness also emerges at the shortest duration (34 ms) we tested, whereas the effect of reconstruction error is only detectable at 84 ms and not before 34 ms, suggesting a temporal window of a few tens of milliseconds after encountering a stimulus in which additional perceptual mechanisms, beyond feedforward processing, may be initially recruited. Future work should use behavioural paradigms—ultra-fast RSVP[44] and others—and computational implementations of these adaptive perceptual mechanisms to uncover a precise temporal evolution of the encoding processes that ultimately lead to traces of remembered or forgotten memories.

Relation to subsequent memory effect and visual exploration

Decades of work in neuroimaging have clearly demonstrated that neural processing during perceptual encoding has a profound impact on what will later be remembered (for example, the subsequent memory effect)[1,52,53]. Compared with items later forgotten, remembered items elicit greater activation in higher-level visual regions, frontoparietal attention regions and the medial temporal lobe[54], and higher pattern similarity across repetitions[2,55,56]. However, the computational processing that underlies these neural signals measured at the interface between perception and memory has been elusive. The empirical success of our SPC model in explaining multiple aspects of memory behaviour suggests reconstruction error is a plausible quantitative covariate to explain aspects of the neural signals underlying the subsequent memory effect.

Previous work has linked active visual exploration (for example, eye movements) with successful memory formation and brain activity (for reviews, see refs. 57–59). Moreover, previous work demonstrated that images with larger numbers of fixations and greater 'fixation-map' consistency across individuals are more memorable[41,60]. Future work should explore the possibility that reconstruction error, in the service of effectively modulating depth of encoding of an image, might drive spontaneous visual exploration behaviour and modulate brain activity (for example, potential regions of exploration include the hippocampus, frontoparietal regions and medial temporal lobe).

These explorations and others can be executed more effectively by refining or extending the architecture presented in this work. The current work explores only one way of computing compression-based reconstruction error—using SPC to compress 1,000 randomly sampled DCNN units in a layer. As a result, what is compressed in the model only partially captures what is in the DCNN activations. Refining this architecture, simply by replacing the random sampling strategy with an additional compression block (via principal component analysis, suggested by an anonymous reviewer) appears readily effective (Supplementary Figs. 6 and 7). Our model can also be extended to represent the reconstruction error with spatially-varying gradients, and thus generate predictions of how the depth of processing can be modulated at finer levels of granularity (covertly, or via a combination of covert processes and eye movements).

Beyond image memorability

Beyond what we studied here (static images), recent work has begun to establish that memorability is a more domain-general phenomenon. This includes the stimulus-driven memorability of videos[61], events[62] and even words[63]. We believe that the presented framework—compression-based reconstruction error—should be a good starting point for characterizing what may be a shared, domain-general mechanism of memorability. Future work should build on our research to explore models that integrate compression-based reconstruction error with different forms of encoding sequential inputs, including using recent advances in neural network models for encoding of videos and events[64], and text-based natural language input[65].

More broadly, beyond its role in driving memorability, reconstruction error can provide a more general priority signal that can be useful across multiple domains of cognition. These include the broader

range of memory-related processes such as retrieval[24,25], modulation of learning during development[66] and guiding the deployment of visual attention beyond bottom-up salience. Finally, our study also showcases the utility of 'model-driven psychophysics'. Computational modelling allows us to generate quantitative predictions about behavioural performance and in turn, behavioural phenomena can help constrain and arbitrate between different models.

Methods

This study complies with all relevant ethical regulations and was approved by the Yale University Institutional Review Board (protocol no. 2000026775).

Studies 1 and 2

Dataset. The dataset we used is from Isola et al.[12] and includes 2,222 target images (with memorability scores measured from human observers; memorability score = hit rate − false alarm rate) and 8,042 filler images (without memorability scores). For the reaction time (RT) measure, we only included RTs during the correctly recognized (hit) trials. After excluding outliers (±2.5 s.d. from the mean RT of all the correct trials), we averaged the remaining RTs for each target image to obtain one single measure of RT for the target image. The dataset covers a wide range of outdoor and indoor scenes. After removing duplicate images (images with different indices but identical content), we are left with 2,221 target images and 8,038 filler images.

Quantifying distinctiveness. We used the VGG-16 network trained for classifying both objects and scenes[34]. The model is trained to make a 1,365-way classification consisting of both object and scene categories (by merging the 1,000 classes from ImageNet and the 365 classes from Places365-Standard). We used the Keras[67] implementation of this model (https://github.com/GKalliatakis/Keras-VGG16-places365; Keras v.1.0.8). We then sampled seven layers across the hierarchy of the network to capture visual features at each layer. The seven layers were maxpooling 1–5 and fully connected (fc) 1–2. For simplicity, we referred to these as layers 1–7.

To quantify distinctiveness, we first passed all the images (resized from 256 × 256 to 227 × 227 pixels) in the Isola dataset (both targets and fillers) through the VGG-16 and extracted activation patterns across the seven layers. Following our previous work[16], for each layer, we calculated distinctiveness as the Euclidean distance between each target image and its nearest neighbour (among all target and filler images) with respect to the feature space defined by the given layer. We also considered an alternative measure for quantifying DCNN network activations that was shown to be predictive of memorability scores (L2-Norm)[14]. We found that both approaches yielded highly similar results (Supplementary Fig. 1). So we report results based on the nearest neighbour approach.

Quantifying reconstruction error. *Sparse coding*. A separate SPC model was trained for each of the seven layers of the VGG-16 network[34]. SPC involves using linear combinations of a limited number of codewords (a codebook) to reconstruct inputs. To do so, we used LCA (locally competitive algorithm)[31] to implement sparse coding, which is a state-of-the-art method but computationally intensive algorithm. The main source of computational cost arises from the dimensionality of the inputs to be reconstructed. Thus, for the sake of computational tractability, we randomly sample $d = 1{,}000$ columns from the flattened feature activations for each layer. The training objective is then to reconstruct the resulting 1,000-dimensional feature vectors. In the language of SPC, this means that the dimension of each codeword is 1,000. (At this codeword length, we found that the pattern of reconstruction errors was not sensitive to the specific random sample of units.)

Training SPC also requires a regularizer λ for controlling how many codewords can be used to reconstruct a given feature vector. We choose $\lambda = 0.001$ based on a simple grid search to minimize the reconstruction error. Because this regularization term is sensitive to the scale of feature vectors, and because we wished to use the same λ for all layers, we applied a preprocessing step in which we scaled feature vectors to be reconstructed using a layer-specific constant. This preprocessing step ensured that the same value of λ worked equally well for all layers. Note that this preprocessing only affects the absolute magnitude of reconstruction errors within a layer, but does not affect relative order among the images with respect to that layer. For our study, the relevant metric is these relative differences in reconstruction errors, instead of the otherwise arbitrary absolute reconstruction error magnitudes. Finally, we choose the codebook size (the number of codewords available to use for reconstructing a given input) to be $n = 500$. This codebook size is a reasonable trade-off between the computational cost of a large codebook in LCA and the reconstruction error incurred: setting smaller values of n substantially increases reconstruction error.

As mentioned above, a separate SPC model was trained for each layer of the VGG-16 network using all images in the Isola dataset, including both target images and filler images. (We also tried training only on the filler images and the resulting reconstruction errors for the target images were largely the same (average $r = 0.94$).) We used 800 iterations to train each model, where a random batch of 50 images in each iteration are used as training inputs. In each iteration, model weights are updated at most 500 times in an inner loop, or less if the weights converged for the batch. To quantify reconstruction error, we calculated the Euclidean distance between the reconstructed 1,000-dimensional vectors and the input 1,000-dimensional activation vectors for each image.

The contribution of distinctiveness versus reconstruction error. We performed partial regression to evaluate the relative contributions of distinctiveness and reconstruction error to memorability/response times. More specifically, we first perform a simple linear regression (with intercept) that regresses the memorability scores/response times on distinctiveness and obtain residual r_1. Then, we perform another simple linear regression (with intercept) that regresses the reconstruction error on distinctiveness and obtain residual r_2. Finally, we correlate r_1 and r_2, and report statistics from this correlation. We also performed the same set of analyses to control for reconstruction error.

Study 3

Participants. We recruited 65 participants via Prolific (www.prolific.co). All participants provided informed consents before starting the experiment. As specified in our preregistration submitted on 16 June 2021 (https://aspredicted.org/MFM_R22), we excluded participants who did not complete all 12 blocks of the experiment ($n = 19$) or who did complete all 12 blocks but did not give a response on over 10% of the trials on any given block ($n = 1$). Our final sample included 45 participants. The memory performance of all 45 participants was above chance (all $d' > 0$, sensitivity measure d' = z(Hit) − z(False Alarm)). Each participant was paid US$5. We did not collect any information about the sex and age of the participants. There is no existing literature showing that these factors can impact the phenomena we are studying.

Stimuli. To investigate the roles distinctiveness and reconstruction error play in determining memorability, we sampled four groups of 48 target images each, with divergent profiles of distinctiveness and reconstruction error: (1) images with large distinctiveness and large reconstruction error; (2) images with large distinctiveness and small reconstruction error; (3) images with small distinctiveness and large reconstruction error; and (4) images with small distinctiveness and small reconstruction error.

To sample these images, we started by using layer 5 nearest − 1 neighbour distinctiveness as our distinctiveness measure because this yielded the highest correlation with the memorability scores (Fig. 2a,

right). Then based on the partial regression results, we selected layer 7 reconstruction error from the SPC model as the reconstruction measure because it explained largest proportion of variance in memorability score beyond layer 5 distinctiveness (Fig. 2b). In other words, we benchmarked distinctiveness and reconstruction error using the 'best' performing combination. Using these benchmarks, we designated images in the top and bottom 30 percentiles of all the images (including targets and fillers) as 'large' and 'small' in terms of each measure (distinctiveness/reconstruction error). For each of the four image groups, we then sampled the 48 target images with the most extreme values considering both measures (for example, for the large distinctiveness and large reconstruction error group, we sampled the 48 images with the largest sum of both percentiles; for the large distinctiveness and small reconstruction error group, we sampled the 48 images with the largest difference between the distinctiveness percentile and the reconstruction error percentile).

In addition to the 192 target images sampled in the way described above, we also randomly sampled 2,304 images from the fillers in the Isola dataset to use as fillers in our experiment.

Design. We adopted a within-participant 2 distinctiveness levels (high versus low) × 2 reconstruction error levels (high versus low) × 3 encoding durations (34, 84 or 167 ms) design. Each condition was presented as a separate block (that is, 12 blocks in total, randomized across participants). During each block, half of the trials were target-present trials (the test image was presented in the RSVP stream), whereas the other half were target-absent trials (the test image was not presented in the RSVP stream). Trial order was randomized. Each block consists of 32 trials.

Procedure. We used lab.js (https://lab.js.org/; Release v.20.2.2) to build the experiment for Study 3. Participants first completed a practice block of 20 trials with images not used in the actual experiment and all the practice images were presented for 167 ms each. At the end of each practice trial, participants received feedback on whether their response was correct or not. After the practice block, participants were instructed to press the space bar to begin each experimental block and no feedback was given during the experimental blocks. Given the importance of timing for our experiment, images were preloaded at the beginning of each block to minimize the processing times during trial presentation.

On each trial (following the trial structure of Broers et al.[44]; Fig. 5b), participants first saw a 500 ms fixation cross, followed by a 200 ms blank screen. They then saw a stream of six scene images presented back-to-back. Each image was shown for 34, 84 or 167 ms, depending on the block. If this was a target-present trial, the target image can be presented as the second, third, fourth or fifth image in the RSVP stream (fully counter-balanced). After the RSVP stream, there was another 200 ms blank screen, followed by the test image, which was presented for 400 ms. Participants had a maximum of 10 s to respond either yes (press '1') or no (press '0'). After the participant made a response, the next trial would start.

Analysis. Analyses were conducted using customized Python v.3.7 scripts and RStudio v.1.2.1335 for the ANOVA scripts (deposited in the GitHub repository and included in the 'Code availability' section). The primary measure of interest is hit rate (percentage of correctly recognized target-present trials), calculated separately for each of the 12 within-participant conditions. To assess the effect of encoding times as a function of reconstruction accuracy, we conducted a three-way repeated measures ANOVA with the dependent variable being hit rate and the three factors being distinctiveness, reconstruction error (RE) and encoding durations (time).

To further investigate how the interaction between distinctiveness, reconstruction error and encoding times affects memory, we ran bootstrapping analysis to resample the participants with replacements and only analyse trials in which the encoding times are 34 and 84 ms (early) or 84 and 167 ms (late). For each participant, we first separated the trials into large and small distinctiveness. For each distinctiveness level, we fit two different linear regression models where time was included as the predictor and hit rate as the dependent variable: one for images with large RE and the other for those with small RE. We then subtracted the fitted beta for time for images with small RE from that for images with large RE. We next averaged this beta difference for all the participants in each bootstrapped sample. If our hypothesis holds, the slope for the linear regression model for large RE should be steeper than that for small RE and therefore the mean beta difference should be >0. The bootstrapping procedure was repeated 1,000 times and the *P* value was calculated as the number of iterations in which the mean beta difference is <0 (that is, went against our hypothesis).

The preregistered analysis described above treated distinctiveness and reconstruction error as categorical variables and aggregated images into separate categories. In the following analysis, we aimed to test whether the time-dependent effect of reconstruction error on memory would also hold when we evaluated the contributions of distinctiveness and reconstruction error to memorability of images presented at different exposure times in a continuous manner across all the target images. For each duration, we calculated the correlation between distinctiveness/reconstruction error and hit rate across all target images, respectively. To test statistical significance, we again ran bootstrapping analysis to resample the participants with replacement. For each iteration, we calculated correlations between distinctiveness/reconstruction error with hit rate across the three durations. We then fit two separate linear regression models, one for distinctiveness and the other for reconstruction error, to relate duration to correlation values. If our hypothesis holds, we should expect that the beta for the reconstruction error model is positive, whereas the beta for the distinctiveness model should not be consistently positive. The bootstrapping procedure was repeated 1,000 times and the *P* value was calculated as the number of iterations in which the beta value is <0 (that is, went against our hypothesis), separately for the reconstruction error model and the distinctiveness model.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data used in Studies 1 and 2 are from a publicly available dataset from Isola et al.[12] (https://web.mit.edu/phillipi/Public/MemorabilityPAMI/index.html). De-identified data collected for Study 3 have been deposited on GitHub (https://github.com/CNCLgithub/ReconMem)[68].

Code availability

Codes have been deposited on GitHub (https://github.com/CNCLgithub/ReconMem)[68].

References

1. Wagner, A. D. et al. Building memories: remembering and forgetting of verbal experiences as predicted by brain activity. *Science* **281**, 1188–1191 (1998).
2. Xue, G. The neural representations underlying human episodic memory. *Trends Cogn. Sci.* **22**, 544–561 (2018).
3. Craik, F. I. & Lockhart, R. S. Levels of processing: a framework for memory research. *J. Verbal Learning Verbal Behav.* **11**, 671–684 (1972).
4. Schurgin, M. W., Wixted, J. T. & Brady, T. F. Psychophysical scaling reveals a unified theory of visual memory strength. *Nat. Hum. Behav.* **4**, 1156–1172 (2020).

5. Chun, M. M. & Johnson, M. K. Memory: enduring traces of perceptual and reflective attention. *Neuron* **72**, 520–535 (2011).
6. Kurby, C. A. & Zacks, J. M. Segmentation in the perception and memory of events. *Trends Cogn. Sci.* **12**, 72–79 (2008).
7. Favila, S. E., Lee, H. & Kuhl, B. A. Transforming the concept of memory reactivation. *Trends Neurosci.* **43**, 939–950 (2020).
8. Liu, J. et al. Transformative neural representations support long-term episodic memory. *Sci. Adv.* **7**, eabg9715 (2021).
9. Libby, A. & Buschman, T. J. Rotational dynamics reduce interference between sensory and memory representations. *Nat. Neurosci.* **24**, 715–726 (2021).
10. Serences, J. T. Neural mechanisms of information storage in visual short-term memory. *Vision Res.* **128**, 53–67 (2016).
11. Xu, Y. Reevaluating the sensory account of visual working memory storage. *Trends Cogn. Sci.* **21**, 794–815 (2017).
12. Isola, P., Xiao, J., Parikh, D., Torralba, A. & Oliva, A. What makes a photograph memorable? *IEEE Trans. Pattern Anal. Mach. Intell.* **7**, 1469–1482 (2014).
13. Bainbridge, W. A., Isola, P. & Oliva, A. The intrinsic memorability of face photographs. *J. Exp. Psychol. Gen.* **142**, 1323–1334 (2013).
14. Jaegle, A. et al. Population response magnitude variation in inferotemporal cortex predicts image memorability. *eLife* **8**, e47596 (2019).
15. Khosla, A., Raju, A. S., Torralba, A. & Oliva, A. Understanding and predicting image memorability at a large scale. In *Proc. IEEE International Conference on Computer Vision*, 2390–2398 (2015).
16. Lin, Q., Yousif, S. R., Scholl, B. & Chun, M. M. Image memorability is driven by visual and conceptual distinctivenes. *J. Vis.* **19**, 290c (2019).
17. Kramer, M. A., Hebart, M. N., Baker, C. I. & Bainbridge, W. A. The features underlying the memorability of objects. *Sci. Adv.* **9**, eadd2981 (2023).
18. Baddeley, A. D. The trouble with levels: a reexamination of Craik and Lockhart's framework for memory research. *Psychol. Rev.* **85**, 139–152 (1978).
19. Treisman, A. in *Levels of Processing in Human Memory* (eds Cermak, L. S. & Craik, F. I. M.) 301–330 (Psychology Press, 2014).
20. Craik, F. I. Remembering: an activity of mind and brain. *Annu. Rev. Psychol.* **71**, 1–24 (2020).
21. Cermak, L. S. & Craik, F. I. M. *Levels of Processing in Human Memory* (Psychology Press, 2014).
22. Bainbridge, W. A. The resiliency of image memorability: a predictor of memory separate from attention and priming. *Neuropsychologia* **141**, 107408 (2020).
23. Bates, C. J. & Jacobs, R. A. Efficient data compression in perception and perceptual memory. *Psychol. Rev.* **127**, 891–917 (2020).
24. Schacter, D. L. Adaptive constructive processes and the future of memory. *Am. Psychol.* **67**, 603–613 (2012).
25. Hemmer, P. & Steyvers, M. A Bayesian account of reconstructive memory. *Top. Cogn. Sci.* **1**, 189–202 (2009).
26. Olshausen, B. A. & Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* **381**, 607–609 (1996).
27. Olshausen, B. A. & Field, D. J. Sparse coding with an overcomplete basis set: a strategy employed by v1? *Vision Res.* **37**, 3311–3325 (1997).
28. Benna, M. K. & Fusi, S. Place cells may simply be memory cells: memory compression leads to spatial tuning and history dependence. *Proc. Natl Acad. Sci. USA* **118**, e2018422118 (2021).
29. Lewicki, M. S. Efficient coding of natural sounds. *Nat. Neurosci.* **5**, 356–363 (2002).
30. Zemel, R. & Hinton, G. E. Developing population codes by minimizing description length. *Adv. Neural Info. Process. Syst.* **6**, 11–18 (1993).
31. Rozell, C. J., Johnson, D. H., Baraniuk, R. G. & Olshausen, B. A. Sparse coding via thresholding and local competition in neural circuits. *Neural Comput.* **20**, 2526–2563 (2008).
32. Rumelhart, D., Hinton, G. & Williams, R. Learning representations by back-propagating errors. *Nature* **323**, 533–536 (1986).
33. Gregor, K. & LeCun, Y. Learning fast approximations of sparse coding. In *Proc. 27th International Conference on International Conference on Machine Learning*, 399–406 (2010).
34. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A. & Torralba, A. Places: a 10 million image database for scene recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **40**, 1452–1464 (2017).
35. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. In *Proc. of the 3rd International Conference on Learning Representations* 1–14 (ICLR, 2015).
36. Berger, T. *Rate Distortion Theory: A Mathematical Basis for Data Compression* (Prentice-Hall, 1971).
37. Cover, T. M. & Thomas, J. A. *Elements of Information Theory* (Wiley, 1991).
38. MacKay, D. J. *Information Theory, Inference, and Learning Algorithms* (Cambridge Univ. Press, 2003).
39. Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. The 'wake–sleep' algorithm for unsupervised neural networks. *Science* **268**, 1158–1161 (1995).
40. Kahana, M. & Loftus, G. in *The Nature of Cognition* (ed. Sternberg, R. J.) 322–384 (MIT Press, 1999).
41. Bylinskii, Z., Isola, P., Bainbridge, C., Torralba, A. & Oliva, A. Intrinsic and extrinsic effects on image memorability. *Vision Res.* **116**, 165–178 (2015).
42. Vincent, A., Craik, F. I. & Furedy, J. J. Relations among memory performance, mental workload and cardiovascular responses. *Int. J. Psychophysiol.* **23**, 181–198 (1996).
43. Ragland, J. D. et al. Levels-of-processing effect on word recognition in schizophrenia. *Biol. Psychiatry* **54**, 1154–1161 (2003).
44. Broers, N., Potter, M. C. & Nieuwenstein, M. R. Enhanced recognition of memorable pictures in ultra-fast RSVP. *Psychon. Bull. Rev.* **25**, 1080–1086 (2018).
45. Craik, F. I. Levels of processing: past, present... and future? *Memory* **10**, 305–318 (2002).
46. Friston, K. & Kiebel, S. Predictive coding under the free-energy principle. *Philos. Trans. R. Soc. B* **364**, 1211–1221 (2009).
47. Rosenbaum, R. On the relationship between predictive coding and backpropagation. *PLoS ONE* **17**, e0266102 (2022).
48. Barrow, H. G. & Tenenbaum, J. M. In *Computer Vision Systems* (eds. Hanson A. & Riseman E. M.) 3–26 (Academic Press, 1978).
49. Olshausen, B. A., Mangun, G. & Gazzaniga, M. *Perception as an Inference Problem* (MIT Press, 2014).
50. Yuille, A. & Kersten, D. Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308 (2006).
51. Mumford, D. in *Large-Scale Neuronal Theories of the Brain* (eds Koch, C & Davis, J. L.) 125–152 (MIT Press, 1994).
52. Brewer, J. B., Zhao, Z., Desmond, J. E., Glover, G. H. & Gabrieli, J. D. Making memories: brain activity that predicts how well visual experience will be remembered. *Science* **281**, 1185–1187 (1998).
53. Paller, K. A. & Wagner, A. D. Observing the transformation of experience into memory. *Trends Cogn. Sci.* **6**, 93–102 (2002).
54. Kim, H. Neural activity that predicts subsequent memory and forgetting: a meta-analysis of 74 fMRI studies. *Neuroimage* **54**, 2446–2461 (2011).
55. Xue, G. et al. Greater neural pattern similarity across repetitions is associated with better memory. *Science* **330**, 97–101 (2010).
56. Ward, E. J., Chun, M. M. & Kuhl, B. A. Repetition suppression and multi-voxel pattern similarity differentially track implicit and explicit visual memory. *J. Neurosci.* **33**, 14749–14757 (2013).

57. Voss, J. L., Bridge, D. J., Cohen, N. J. & Walker, J. A. A closer look at the hippocampus and memory. *Trends Cogn. Sci.* **21**, 577–588 (2017).
58. Ryan, J. D., Shen, K. & Liu, Z.-X. The intersection between the oculomotor and hippocampal memory systems: empirical developments and clinical implications. *Ann. N Y Acad. Sci.* **1464**, 115–141 (2020).
59. Kragel, J. E. & Voss, J. L. Looking for the neural basis of memory. *Trends Cogn. Sci.* **26**, 53–65 (2022).
60. Lyu, M. et al. Overt attentional correlates of memorability of scene images and their relationships to scene semantics. *J. Vis.* **20**, 1–17 (2020).
61. Cohendet, R., Demarty, C.-H., Duong, N. Q. & Engilberge, M. Videomem: constructing, analyzing, predicting short-term and long-term video memorability. In *Proc. IEEE/CVF International Conference on Computer Vision*, 2531–2540 (2019).
62. Xu, Q., Fang, F., Molino, A., Subbaraju, V. & Lim, J.-H. Predicting event memorability from contextual visual semantics. *Adv. Neural Info. Process. Syst.* **34**, 22431–22442 (2021).
63. Lau, M. C., Goh, W. D. & Yap, M. J. An item-level analysis of lexical-semantic effects in free recall and recognition memory using the megastudy approach. *Q. J. Exp. Psychol. (Hove)* **71**, 2207–2222 (2018).
64. Majumdar, A. et al. Where are we in the search for an artificial visual cortex for embodied intelligence? *Adv. Neural Info. Process. Syst.* **36**, 1–23 (2024).
65. Radford, A. et al. Language models are unsupervised multitask learners. *OpenAI Blog* **1**, 9 (2019).
66. Stahl, A. E. & Feigenson, L. Observing the unexpected enhances infants' learning and exploration. *Science* **348**, 91–94 (2015).
67. Chollet, F. et al. Keras. https://keras.io (2015).
68. Lin, Q., Li, Z., Lafferty, J. & Yildirim, I. From seeing to remembering: Images with harder-to-reconstruct representations leave stronger memory traces. *GitHub* https://github.com/CNCLgithub/ReconMem (2023).

Acknowledgements

This project was funded by an Air Force Office of Scientific Research (AFOSR) award #FA9550-22-1-0041 (to I.Y.). The funder had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript. We thank the Yale Center for Research Computing for maintaining HPC resources for computation. We also thank R. Jacobs, B. Scholl and members of the Yale Cognitive & Neural Computation Lab for comments on an earlier version of this manuscript.

Author contributions

Q.L., J.L. and I.Y. conceived the study. Q.L., Z.L., J.L. and I.Y. developed the methodology. Q.L. and Z.L. developed the software. Q.L. collected the data. Q.L. and Z.L. formally analysed the data. Q.L., Z.L. and I.Y. wrote the original draft. Q.L., Z.L., J.L. and I.Y. wrote and edited the manuscript. Q.L. visualized the data. J.L. and I.Y. supervised the study. I.Y. acquired the funding.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-024-01870-3.

Correspondence and requests for materials should be addressed to Qi Lin, John Lafferty or Ilker Yildirim.

Peer review information *Nature Human Behaviour* thanks the anonymous reviewers for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): Ilker Yildirim, Qi Lin, and John Lafferty

Last updated by author(s): Feb 1, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Only Study 3 includes newly collected data in our manuscript. The data collection for Study 3 was conducted online and we used lab.js (https://lab.js.org/; Release 20.2.2) to build our experiment.
Data analysis	The analysis codes for data analysis have also been deposited to Github repository for the current project (https://github.com/CNCLgithub/ReconMem). The analysis scripts were written in Python 3.7. To extract features from the deep neural network (VGG16-Places365: https://github.com/GKalliatakis/Keras-VGG16-places365), we used Keras (Version: 1.0.8). For the ANOVA results reported in Study 3, we used RStudio (1.2.1335).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

The data used in Studies 1 &2 are publicly available (https://web.mit.edu/phillipi/Public/MemorabilityPAMI/index.html). The de-identified data collected for Study 3 are deposited in our Github repository (https://github.com/CNCLgithub/ReconMem).

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	For Study 3 (for which new data were collected), we did not think there would be difference in memory performance across sex/gender. Therefore, we did not collect or report any information on sex and gender.
Population characteristics	We did not collect or report any information on the demographic characteristics of our sample.
Recruitment	We recruited participants through Prolific (https://www.prolific.co/) in a randomized manner (as offered by the platform). We required the participants to be using a computer, rather than a tablet. We acknowledge that this may introduces a selection bias, including those with access to computers and the Internet. However, there is no existing literature showing that this bias can significantly impact the phenomena we are studying.
Ethics oversight	The study has been approved by and overseen by the Yale University Institutional Review Board.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	As stated above, only Study 3 includes new quantitative data collected in a psychophysics experiment and is a within-participant design.
Research sample	Studies 1&2 used a publicly available dataset with memory performance from 665 participants for 2221 unique images in an online experiment (https://web.mit.edu/phillipi/Public/MemorabilityPAMI/index.html). Study 3 was conducted online through Prolific. We did not implement any restriction (except that they had to use a computer rather than a tablet to access our study to ensure that our stimuli were displayed properly) on Prolific. Therefore, we consider our sample to be representative of the Prolific participant pool.
Sampling strategy	For Studies 1&2, we chose the memorability dataset released by Isola et al. (2013) because of the representativeness of its stimuli selection and its sample size (2221 unique scene images with behavioral responses from 665 participants). For Study 3, we adopted a random sampling approach. The sample size (N=45) was decided based on a separate pilot study and was pre-registered before data collection (Pre-registration report here: https://aspredicted.org/blind.php?x=MFM_R22).
Data collection	The experiment was conducted online and participants completed the study on their own computers without researcher supervision.
Timing	Data was collected online from 2021 June 28 to 2021 July 1.
Data exclusions	As stated in our pre-registration (https://aspredicted.org/blind.php?x=MFM_R22), we excluded participants who met any of the following criteria: 1) did not complete all 12 blocks of the experiment (N=19); 2) failed to provide a response on over 10% of the trials on any given block (N=1); 3) performed below chance overall (overall d' [z(hit) – z(false alarm)] < 0, collapsing across all conditions). (N=0)

Non-participation	Based on the recorded responses transferred to our server, 19 participants did not complete all 12 blocks of the experiment and 1 of them made no response on more than 10% on at least 1 block. Of the 19 participants with incomplete data, only 2 reached out to us through Prolific and reported that the images did not load. We suspect that this arose from the large number images loaded for our experiment (2688 images in total) and therefore, for these participants with limited bandwidth, they were only able to complete part of the experiment.
Randomization	Given that the study was a within-participant design, there was no random assignment to conditions (i.e., all participants saw the same set of conditions). The only randomization was done on the order of the experiment conditions and the order of trials.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging