Visual recognition memory of scenes is driven by categorical, not sensory, visual representations.

https://doi.org/10.1523/JNEUROSCI.1479-23.2024

Received: 28 July 2023
Revised: 7 February 2024
Accepted: 14 February 2024

Copyright © 2024 the authors

This Early Release article has been peer reviewed and accepted, but has not been through the composition and copyediting processes. The final version may differ slightly in style or formatting and will contain links to any extended data.

Alerts: Sign up at www.jneurosci.org/alerts to receive customized email alerts when the fully formatted version of this article is published.
Visual recognition memory of scenes is driven by categorical, not sensory, visual representations.

Ricardo Morales-Torres¹, Erik A. Wing², Lifu Deng¹, Simon W. Davis¹,³, Roberto Cabeza¹

¹Department of Psychology & Neuroscience, Duke University, Durham, NC 27708
²Rotman Research Institute, Baycrest Health Sciences, Toronto, ON M6A 2E1, Canada
³Department of Neurology, Duke University School of Medicine, Durham, NC 27708

*Corresponding Author:
Roberto Cabeza
cabeza@duke.edu
Duke University
Durham, NC 27708

The authors declare no competing financial interests

Acknowledgements
This study was supported by the National Institute of Health, R01-AG066901 and R01-AG075417.
Abstract

When we perceive a scene, our brain processes various types of visual information simultaneously, ranging from sensory features, such as line orientations and colors, to categorical features, such as objects and their arrangements. Whereas the role of sensory and categorical visual representations in predicting subsequent memory has been studied using isolated objects, their impact on memory for complex scenes remains largely unknown. To address this gap, we conducted an fMRI study in which female and male participants encoded pictures of familiar scenes (e.g., an airport picture) and later recalled them, while rating the vividness of their visual recall. Outside the scanner, participants had to distinguish each seen scene (e.g., an airport picture) from three similar lures (e.g., 3 airport pictures). We modeled the sensory and categorical visual features of multiple scenes using both early and late layers of a deep convolutional neural network. Then, we applied representational similarity analysis to determine which brain regions represented stimuli in accordance with the sensory and categorical models. We found that categorical, but not sensory, representations predicted subsequent memory. In line with the previous result, only for the categorical model, the average recognition performance of each scene exhibited a positive correlation with the average visual dissimilarity between the item in question and its respective lures. These results strongly suggest that even in memory tests that ostensibly rely solely on visual cues (such as force-choice visual recognition with similar distractors), memory decisions for scenes may be primarily influenced by categorical rather than sensory representations.

Key words: episodic memory; scene memory; representational similarity analysis; recognition memory
Significance Statement

Our memory for real-world scenes often comprises a tableau of complex visual features, but recent findings challenge the view that our memories of such stimuli rely on purely visual information. Instead, it appears that our memory for scenes is heavily influenced by higher-level categorical information. Analyzing cortical representations in regions responsive to both categorical and sensory features, we discovered that only the former can reliably predict memory outcomes. Moreover, the distinctiveness of scenes in terms of their categoric features among similar examples is positively associated with our ability to accurately recognize previously encountered scenes. In essence, this study sheds light on how our brains rely on categorical information to recognize natural scenes.
Introduction

Humans have an astonishing capacity to remember scenes (Standing, 1973), which are multi-element organizations of visual features (Epstein & Baker, 2019; Henderson & Hollingworth, 1999). Even after exposure to numerous scenes, individuals can retain sufficient details to later distinguish between seen and similar unseen exemplars of the same scene category, such as two photographs of similar airports (Konkle et al., 2010). However, it is unclear whether this capacity is driven by neural representations of sensory information (the color, size, or orientation of the planes), or by the neural representations of categorical information (the knowledge structure that contains information about the more likely objects and their spatial distribution to be encountered within a layout (Epstein & Baker, 2019; Henderson & Hollingworth, 1999), e.g. seeing an airplane next to an air traffic control tower allows an observer to categorize a scene as an airport instead of an aircraft museum). While the impact of different features on scene perception has been carefully investigated (for reviews, see Castelhano & Krzyś, 2020; Epstein & Baker, 2019), their specific influence on subsequent scene memory remains largely unknown, and it is the focus of the current study.

To identify the neural correlates of different visual features, functional Magnetic Resonance Imaging (fMRI) researchers are increasingly relying on a combination of voxel-based representational similarity analyses (RSA), and stimulus models based on deep convolutional neural networks (DNNs). RSA allows for the quantification of the representational structure of cortical activation patterns by relating the similarity in voxel activation and model values across items (Kriegeskorte & Diedrichsen, 2019). DNNs, in turn, offer ready resources for such model parameters, such that early DNN layers quantify simple visual features while late layers describe more complex categorical features. This hierarchical organization mirrors the properties of the human visual system, making DNNs an important tool in this research domain (Kriegeskorte, 2015; Peters & Kriegeskorte, 2021; Yamins et al., 2014). By employing this combined approach, researchers have discovered that during scene processing the neural representations of early DNN layers (henceforward sensory visual features) emerge rapidly across the visual cortex (Greene & Hansen, 2020) and influence the subsequent categorical processing (Dima et al., 2018; Kaiser et al., 2020). In turn, only the neural representations of late DNN layers (henceforward categorical visual features), have been shown to correlate with participants behavior when categorizing or rating the similarity between different scenes (Greene & Hansen, 2020; Groen et al., 2018; King et al., 2019).

Although our current knowledge regarding the specific impact of neural representations of scene features on memory is limited, evidence emphasizes the importance of categorical visual features in the formation of memories. Unlike sensory features, categorical visual features have been shown to be predictive of scene memorability (Isola et al., 2014), defined as the likelihood of a viewer recognizing a previously seen scene. Additionally, mnemonic interference between scenes is primarily influenced by similarities in categorical rather than sensory visual features (Konkle et al., 2010; Mikhailova et al., 2023). Given the prominent role of categorical features in both scene memory and the neural representations of scenes, we hypothesize that a stronger encoding of categorical visual features will be associated with enhanced memory performance.

To investigate the role of visual features in scene memory, we employed a widely used DNN (VGG-16) to quantify both sensory and categorical properties of complex scenes. By modelling both types
of features we aimed to isolate the influence of each feature type and examine its relationship to memory. To assess whether a stronger encoding of categorical scene features was associated with memory performance, we utilized RSA to identify brain regions representing both types of visual features and examined how these feature representations were linked to scene memory. Thus, our approach seeks to characterize the influence of multiple levels of information in processing scene information, and address how that information contributes to the formation of lasting memories.

Methods

Participants

A total of 22 adults (12 women; Mean Age = 23.5, SD = 3.0) took part in the experiment, all of whom were healthy, right-handed, English speakers with normal or corrected-to-normal vision. Prior to their participation, informed consent was obtained from each participant, and the consent process was approved by the Duke University institutional review board. One participant’s data was excluded from analysis due to a technical error during the encoding phase.

Paradigm and Procedure

As illustrated by Figure 1, the experiment consisted of three phases: encoding (3 fMRI runs), recall vividness (3 fMRI runs), and Force choice recognition (outside the scanner). There was also a short practice session prior to scanning. During each encoding run, participants viewed 32 images of scenes (96 scenes in total), each presented for 4 seconds, accompanied by a label indicating its general category (such as “Airport” or “Movie Set”). Participants were asked to rate the degree to which each image represented its label (e.g., “Is this a good picture of an airport?”). An 8-second interval separated each image presentation, during which participants made even/odd judgments to a series of digits ranging from 1 to 9. During each recall vividness run, participants read the labels of encoded scenes, and tried to recall the corresponding scenes as vividly as possible, rating vividness from 1 = least amount of detail, to 4 = highly detailed memory. Immediately following the scan session, participants completed a four-alternative forced-choice test of all 96 encoded scenes. In each trial, the target image was presented together with three scene lures for the same scene category as the target (e.g., 3 airport scenes) in the four quadrants of the screen. Participants selected the scene they believed they saw during the encoding phase, and then rated their confidence in the choice (1 = guess, 4 = very confident).

The results of some analyses on the current fMRI dataset were previously reported by Wing et al. (2015), who performed encoding-retrieval similarity (ERS) analyses to investigate the reactivation of encoding information during the recall vividness test. In contrast, the current study uses RSA to examine if and how sensory and categorical representations predict performance in recall vividness and forced-choice recognition tests.

Figure 1. Experimental Design. During the encoding phase participants saw 96 pictures of scenes, with a descriptive label. During the recall vividness test, descriptive labels for the previously encoded scenes were presented and participants were asked to first recall the encoded scene and
then rate how vivid/detailed their mental image of the scene was. During the 4-alternative forcedchoice test, participants were asked to choose the specific scene they believed was presented during encoding, and then rate their confidence about the decision.

fMRI acquisition and pre-processing

The fMRI acquisition has been previously described in Wing (2015). Briefly, data was collected with a 3T GE scanner, using a SENSE spiral-in sequence (repetition time = 2 seconds, echo time = 30 milliseconds, field of view = 24 centimeters). The anatomical image comprised 96 axial slices parallel to the AC-PC plane with voxel dimensions of 0.9x0.9x1.9 millimeters. The data was pre-processed in SPM-12 (London, United Kingdom; http://www.fil.ion.ucl.ac.uk/spm/) using several steps, including discarding the first six functional images (to allow the scanner to reach equilibrium), slice time correction to the first slice, realignment to the first scan, motion correction, and unwarping the images. The functional images were then co-registered to the skull-stripped anatomical image and normalized into the MNI space using DARTEL (Ashburner, 2007). The DRIFTER toolbox was used to denoise the images (Särkkä et al., 2012). The pre-processed functional images were non-smoothed, and each voxel had a size of 3.75 × 3.75 × 3.8 mm³.

To obtain beta coefficients for each scene, a single trial model was conducted using the least squares-separate approach developed by Mumford et al. (2012). The model consisted of a regressor modeling the activity of the trial of interest and a regressor modeling the activity of all other trials within that run. The events were modeled using a stick function convolved with a standard hemodynamic response function. Additionally, the six raw motion regressors, a composite motion parameter generated by the Artifact Detection Tools, outlier time points (scan-to-scan motion > 2.0 mm or degrees, scan-to-scan global signal change > 9.0 z score; derived from the composite motion parameter), white matter, and cerebrospinal fluid time series were included in each first-level model. A high-pass temporal filter with a 128-second cutoff was applied to the data.

DNN Model

As a model of visual scene perception, we used the Keras implementation of the VGG-16 model (https://github.com/GKalliatakis/Keras-VGG16-places365). This implementation was pretrained using the Places365-Standard dataset, a subset of the Places365 dataset (Zhou et al., 2018), which contains approximately 1.8 million images from 365 scene categories. The architecture of the network consists of 13 convolutional layers, 5 max-pooling layers and 3 fully connected layers. Each layer is composed of multiple units, with the units of the first layer corresponding to every pixel of the input image (244x244 pixels), and progressively decrease the number of units until reaching the last, which is a vector of 365 units that represents the categories of the scenes.

To obtain a model of the visual features of scenes we fed the network the 96 images (resized to 244x244 pixels) used in our experiment. To model sensory and categorical representations, we followed O’Connell & Chun (2018) and Xu & Vaziri-Pashkam (2021) and extracted the values from the first and last max-pooling layers. The max-pooling layers pool information processed by the convolutional layers and pass it to the next section of the network, marking the end of a processing stage. Since the first set of convolutional layers are sensitive to low-level visual features, such as frequency, boundary, and color information (Eickenberg et al., 2017; Kriegeskorte, 2015; Peters & Kriegeskorte, 2021) we used the first max-pooling layer as a model of the sensory features of
scenes. The last block of convolutional layers contains information defined by the visual features of scenes, but with enough abstraction that feeding the activity of this layer to the last block of the network (the fully connected layers) allows the model to categorize the images. Therefore, we used the max-pooling layer at the end of the last convolutional block to model categorical visual features (See Figure 2B for a representation of how these models group images).

Regions of interest

To examine the influence of sensory and categorical features of images on brain representations, we defined bilateral ROIs using all bilateral cortical (excluding the cerebellum) and subcortical regions from the AAL3 brain atlas (Rolls et al., 2020).

Representational Similarity Analysis

In order to obtain the representations of the categorical and sensory features of the images we employed a representational similarity analysis approach (Kriegeskorte & Diedrichsen, 2019; Kriegeskorte & Kievit, 2013).

As a first step, we generated representational dissimilarity matrices (RDM) of the sensory and categorical feature of scenes by feeding to the DNN the 96 scenes. To generate the sensory RDM we vectorized the values of the first max-pooling layer for each scene and then we calculated the pairwise dissimilarity between all scenes, computed as 1 minus the Pearson correlation coefficient. Using this procedure, we generate a 96x96 matrix, each row of the matrix representing how dissimilar each image is to all other images according to the sensory features of the scenes. To obtain the categorical RDM we used the same procedure, but vectorizing the activity of the last max-pooling layer.

We then created an activity pattern matrix (APM) for all bilateral ROIs. To construct an APM, we first vectorized all voxels within an ROI, repeating this process across the 96 trials. This resulted in 96 vectors representing the voxel patterns for a given brain region during the presentation of the 96 scenes. We calculated the dissimilarity between these 96 vectors, defined as 1 minus the Pearson correlation coefficient, obtaining a 96x96 matrix. The dimensions of the APM matrix match those of our sensory and categorical RDMs.

In our last step, the objective was to identify which ROIs represented stimuli in a manner consistent with how the model represented the sensory and/or categorical features of scenes. To achieve this, we applied an item-wise procedure developed by Davis et al. (2021). This procedure yielded a value that serves as an index of the fit between the neural and the DNN representations of a specific scene, obtained as a second-order correlation between the rows of the two matrices. We computed the Spearman correlation between each row of the APM with each corresponding row of the sensory RDM and normalized the correlation coefficient using the Fisher transformation. Each row of the APM and RDMs represents how dissimilar a specific scene is from all other scenes (according to a DNN or according to an ROI). By correlating the APM row, which shows the dissimilarity in voxel activity estimates, with the RDM row, reflecting the dissimilarity in DNN layer activation values, we measure how similarly a brain region represents a particular scene compared to how the model does. This process yields 96 values, with higher values indicating a greater consistency between a specific ROI’s stimulus representation and the type of feature format that the model is
representing (Figure 2A presents a graphical example of this approach). We used the same
procedure to obtain the representational strength of the categorical features of scenes.

As in Davis et al. (2021), we interpret the item-wise Brain-Model Fit values as an index of the
sensitivity of specific regions to the particular type of feature being modeled by a DNN layer. While
this approach differs from classical RSA, where the two similarity matrices are correlated
(Dimdsdale-Zucker & Ranganath, 2018; Kriegeskorte & Kievit, 2013; Popal et al., 2019), by
calculating the brain-model fit for each scene individually, we can directly relate the sensitivity of
an ROI to a specific feature and its relationship with the memory for an individual scene.

[FIGURE 2]

Figure 2. (A) Overview of the RSA approach. We used the DNN VGG-16 as model of the visual
processing of scenes. We selected the first and last max-pooling layers as models of the sensory and
categorical features of scenes (note that we are depicting a simplified version of VGG-16 and not the
actual architecture of it). We generated RDMs from our sensory and categorical models, which we
 correlated with the multivariate activity of several regions of the brain. By correlating the rows
from the sensory and categorical model RDMs with the brain APMs, we obtained a measure of the
sensory and categorical representations of the scenes. (B) Multidimensional scaling plots (MDS) of
the sensory and categorical models. The MDS plots qualitatively show that the sensory model
groups images according to their color (orange scenes on the left and blue scenes on the right),
while the categorical model groups images according to their category (indoor scenes on the left
and outdoor scenes on the right). These visualizations provide insights into how the model
represents scenes at different levels of abstraction.

Behavioral Analysis

To evaluate the relationship between the two memory tests, we utilized Yule’s Q, a statistical test
that measures the level of association or dependence between the outcomes of both tests. Yule’s Q
provides a value that can be interpreted in a manner similar to a correlation coefficient (Kahana,
2000), offering insights into the degree of dependency between the recall vividness and forced-
choice recognition tasks.

For the recall-vividness test, as in previous studies, we grouped vividness ratings of 1 and 2 as 'low
vividness,' and ratings of 3 and 4 as 'high vividness' (Kuhl & Chun, 2014; Lee et al., 2019). In the
recognition memory test, as we are only interested in memory performance driven by recollection,
rather than memory performance driven by familiarity, we considered only those trials recognized
with a high-confidence rating as accurate (Kim & Cabeza, 2009; Lee et al., 2019).

Statistical Approach

To test the influence of sensory and categorical scene representations on subsequent memory we
generated 2 different types of statistical models. Our first model was generated to answer if our
ROIs represented the sensory and categorical features that we derived from the DNN. We employed
an item-wise, hierarchical mixed-effects model with the brain-model fit values for each scene in
each region as the outcome variable. To account for potential autocorrelation between different
layers of the DNN (Bone et al., 2020), and to isolate the influence of a specific type of feature, we
included the type of visual representation that was not used as the outcome variable as a control variable. Subject and scene were included as random effects. We included each scene as a random effect, such that our findings can be generalized to other scenes, thus aiming to address the stimulus as a fixed effect fallacy (Raaijmakers, 2003). Consequently, the statistical models were formulated as follows:

(1) Brain-Model Fit\textsubscript{Sensory} \sim ROI + Brain-Model Fit\textsubscript{Categorical} + (1\mid Subject) + (1\mid Scene)

(2) Brain-Model Fit\textsubscript{Categorical} \sim ROI + Brain-Model Fit\textsubscript{Sensory} + (1\mid Subject) + (1\mid Scene)

We then constructed a second type of model to examine the impact of sensory and categorical representations on subsequent memory. The objective of this model was to determine which type of representation could predict whether a scene would be remembered. We included the brain-model fit values of both features as predictors. By incorporating both features in the same model, we allowed them to 'compete' with each other, each attempting to demonstrate its unique contribution to explaining variations in the dependent variable, while controlling for the influence of the other representation type. As the dependent variable is binary, this second model took the form of a mixed-effect logistic regression, with subject and scene as random effects. Consequently, the statistical model was formulated as follows:

(3) Memory (1/0) \sim Brain-Model Fit\textsubscript{Sensory} + Brain-Model Fit\textsubscript{Categorical} + (1\mid Subject) + (1\mid Scene)

To test the significance of the logistic regression coefficients, we employed a permutation approach, which offers the advantage of better controlling for false positives and making fewer assumptions compared to standard statistical approaches (Winkler et al., 2014). In the first step, we computed the logistic regression coefficients of the mixed-effect logistic regression. Following previous approaches to randomization in mixed-effect models (Anderson & Legendre, 1999; Manly, 1986), our second step involved, for each subject, randomly changing the values of the dependent variable (which scenes were remembered or forgotten) and then computing the coefficients. We repeated this step 5000 times and compared how extreme the real coefficients obtained from the first step were compared to the distribution of the 5000 coefficients obtained during the randomization process. A p-value of less than .05 means that the real coefficient has a more extreme value than at least 95% of the coefficients obtained during the randomization approach, which represents the distribution of values if the null hypothesis were true.

Statistical analysis was done in RStudio (RStudio Team, 2020) and graphs were generated using the ggplot2 package (Wickham, 2011). Mixed-effect models were performed using the lme4 package (Bates et al., 2015). The p-values for the brain-model fit coefficients of each AAL3 bilateral ROI were obtained using the lmerTest package. (Kuznetsova et al., 2017).

**Target-Lure Dissimilarity Analysis**

Our last analysis focused on comparing which type of scene feature is used when subjects need to recognize an image from visually similar lures. The assumption of this analysis is that forced-choice recognition accuracy is better when the target is different from the lures. Following that rationale we generated a metric, named Mean Target-Lure Dissimilarity (MTLD), that summarizes the relationship between a target scene and a series of lure scenes for that target, focusing on a specific type of scene feature.
We computed MTLD values for each scene following five steps: (1) We fed the target scene and its three lures of each trial to the DNN. (2) We extracted the activation values from the first and last max-pooling layers of the DNN. (3) For both the sensory and categorical features, we calculated the dissimilarity value (1 minus the Pearson correlation) between each target and its three lures. (4) We then averaged these three dissimilarity values separately for sensory and categorical features to obtain each trial’s MTLD value. (5) We computed the Z-score of the MTLD value, using the mean and standard deviation of the dissimilarity between the target scene and the other 95 scenes.

Results

Behavioral results

The analysis of the behavior of participants aimed at testing to what degree the sequentially administered recall vividness and forced-choice recognition tests were interrelated, and thus to what extent their mnemonic processes draw upon common information (Tulving & Wiseman, 1975). The recall vividness test has as outcomes High Vividness (1) or Low Vividness (0) trials, while the forced-choice recognition test has as outcomes trials Recognized (1) or Forgotten (0). Our contingency analyses considered the number of scenes rated as high vividness and later recognized (1/1), scenes rated as high vividness but not recognized (1/0), scenes rated as low vividness but then recognized (0/1), and scenes rated as low vividness and not recognized (0/0).

We calculated Yule’s Q for each participant, averaging this value across subjects and then compared it against 0 using a t-test. Our results showed that the recall vividness and forced-choice recognition had a moderate dependency between each other (Yule’s Q = 0.44, SD = 0.26; t(20) = 7.62, p < .001, Cohen’s d = 1.66). A moderate Yule’s Q suggest that the recall vividness and forced-choice recognition tests are sensitive to similar information, therefore it is reasonable to expect that visual representations could have a similar impact on both tests.

Sensory and categorical representations regardless of subsequent memory

Functional MRI data first revealed which regions showed evidence of representing sensory and categorical features of scenes (see Figure 3). We estimated the brain-model fit for each ROI and for each feature model using a mixed-effects regression. We then contrasted these coefficients against 0, and corrected the p-values for all comparisons using the False Discovery Rate (FDR) correction. We found several regions within occipital, ventral and parietal regions that represented both types of scene features. This regions were the dorsolateral prefrontal cortex (categorical: t(125.8) = 3.69, p < .001; Sensory: t(64.46) = 3.35, p < .001), parahippocampal cortex (categorical: t(125.8) = 3.16, p = .004; Sensory: t(64.46) = 4.40, p < .001), inferior temporal cortex (categorical: t(125.8) = 4.19, p < .001; Sensory: t(64.46) = 3.31, p = .006), precuneus (categorical: t(125.8) = 3.97, p < .001; Sensory: t(64.46) = 3.46, p = .005), fusiform cortex (categorical: t(125.9) = 5.44, p < .001; Sensory: t(64.46) = 5.59, p < .001), superior occipital cortex (categorical: t(125.8) = 6.92, p < .001; Sensory: t(64.46) =
3.91, \( p = .001 \), middle occipital cortex (categorical: \( t(125.98) = 8.26, p < .001 \); Sensory: \( t(64.51) = 7.81, p < .001 \), inferior occipital cortex (categorical: \( t(125.98) = 7.52, p < .001 \); Sensory: \( t(64.49) = 6.00, p < .001 \), lingual gyrus (categorical: \( t(125.83) = 6.77, p < .001 \); Sensory: \( t(64.46) = 4.12, p < .001 \), and the calcarine fissure and surrounding cortex (categorical: \( t(125.92) = 10.05, p < .001 \); Sensory: \( t(64.54) = 6.1, p < .001 \)) ROI.

For each of the 10 ROIs showing consistent brain-model fit with both sensory and categorical scene features, we compared the coefficients of these feature types. This involved calculating the differences between the z-score-transformed coefficients to determine if any region exhibited a stronger fit for one feature type over the other. Following the FDR correction for multiple comparisons, we observed that the superior occipital cortex (\( Z_{\text{diff}} = 3.00, p = .007 \)), lingual gyrus (\( Z_{\text{diff}} = 2.65, p = .01 \), and the calcarine fissure and surrounding cortex (\( Z_{\text{diff}} = 3.95, p < .001 \)) ROIs more strongly represented the categorical features of scenes. No other significant differences were found across ROIs (all \( p > .16 \)).

Figure 3. (A) Brain-model fit values across ROIs: We calculated the Brain-model fit values for all bilateral ROIs in the AAL3 atlas. For both types of features, the brain-model fit showed higher values in occipital and temporal regions. (B) Regions representing both types of scene features: 10 ROIs across the cortex represent both types of scene features. Asterisks denote differences in brain-model fit between feature types in a specific ROI. *** \( p < .001 \), ** \( p < .01 \) * \( p < .05 \) (fdr corrected p-values).

The previous analysis revealed that several brain regions demonstrated significant brain model fits for both categorical and sensory scene features. However, to preclude the potential confound of the visual model capturing objects within the scenes rather than scenes as a whole, we conducted a follow-up analysis. This analysis tested whether a model trained on object classification would perform similarly to one trained on scene classification. For this, we calculated the brain-model fit of sensory and categorical scene features using a version of VGG-16 trained for object classification. We then compared the brain-model fit estimates of the 42 bilateral ROIs previously calculated with a DNN trained on scene-classification against the brain-model fit estimates obtained with the same DNN trained on object-classification.

Our results indicated that the brain-model fits of the sensory features did not differ significantly between the model trained for scene-classification (M = 0.011, SD = 0.009) and the model trained for object-classification (M = 0.009, SD = 0.008; \( t(41) = 1.85, p = .07 \), Cohen’s d = 0.23). In contrast, for the representations of categorical features, the brain-model fits obtained using the scene-classification model (M = 0.015, SD = 0.009) were significantly higher compared to those obtained using the object-classification model (M = 0.006, SD = 0.005; \( t(41) = 8.63, p < .001 \), Cohen’s d = 1.22). These results reveal that while the early stages of models trained for scene and object classification show comparable results in capturing the representational structure of the brain, the later stages of the network trained with scenes are significantly more effective in capturing the representational structure of the brain during scene perception. This suggests that the processing of scenes cannot be reduced to the processing of the objects within them.
Only categorical representations predict subsequent scene memory

To test and compare the influence of sensory and categorical feature representations on scene memory, we conducted a mixed-effect logistic regression using the brain-model fit of both types of features as predictors and memory as the outcome variable, as well as including subject and scene as random factors. The inclusion of both types of feature representations within the same model enabled a direct comparison of their influence on scene memory, while also accounting for shared information between them. Since the Yule’s Q analysis revealed a significant relationship between both types of memory tests, we classified trials that were recalled with high vividness and subsequently recognized as 'remembered' and coded them as '1' for the mixed-effect logistic regression. All other trials were coded as '0'.

We applied the mixed-effects model to the 10 ROIs that demonstrated a significant fit for both types of feature representations. Following the adjustment for multiple comparisons using the FDR correction, we identified that, among these 10 ROIs, associations between memory and brain-model fit were significant in three ROIs. In the Precuneus, we observed a positive association with subsequent memory exclusively for categorical feature representations ($\beta = 0.15, SE = 0.05, p = .011$), while sensory feature representations ($\beta = -0.04, SE = 0.05, p = .77$) did not show a significant link to memory. A similar pattern was found in the inferior temporal cortex, where categorical feature representations ($\beta = 0.12, SE = 0.05, p = .018$) were positively associated with subsequent memory, while sensory feature representations ($\beta = 0.02, SE = 0.05, p = .69$) did not exhibit a significant relationship with memory outcomes. Lastly, within the superior occipital cortex, we found that only categorical feature representations ($\beta = 0.13, SE = 0.05, p = .017$) displayed a positive association with subsequent memory, whereas sensory feature representations ($\beta = -0.03, SE = 0.05, p = .77$) did not exhibit a significant connection to memory performance.

To validate our results in these 3 ROIs, we directly compared the influence of each type of feature representation in explaining subjects memory by employing a likelihood ratio test. This test compared the fit between different statistical models in predicting the outcome variable and directly compares the models’ ability to account for the data (Nie, 2006). In our case, for the 3 ROIs, we compared the ability of three models to explain behavior. The first model only contained the intercept and the random effects of subjects and scenes, the second contained the random effects as well as the sensory brain-model fit values for each scene and the last model contained the random effects, as well as the sensory and categorical brain-model fit values for each scene. This stepwise approach allowed us to directly compare the ability of sensory and categorical feature representations to explain subsequent memory.

The likelihood ratio test revealed that, in comparison to a model devoid of any form of feature representations, sensory feature representations in the Precuneus ($\chi^2(1) = 0.12, p = .91$), inferior temporal cortex ($\chi^2 (1) = 0.69, p = .40$) and superior occipital cortex ($\chi^2(1) = 0.04, p = .85$) did not make a significant contribution to the model’s capacity to explain the behavioral data. Conversely, a model that incorporated categorical feature representations displayed a notably improved fit to the data in the Precuneus ($\chi^2 (1) = 7.85, p = .005$), inferior temporal cortex ($\chi^2 (1) = 4.66, p = .03$) and superior occipital cortex ($\chi^2(1) = 5.65, p = .017$), when compared to a model using only sensory feature representations as predictors of memory.
In summary, from the group of ROIs that showed both sensory and categorical representations, only categorical representations in the precuneus, inferior temporal cortex and superior occipital cortex predicted scene memory. Consistent with our hypothesis, stronger encoding of categorical scene features was associated with better memory performance. To further investigate the effect of categorical representations on memory recognition, we conducted a follow-up analysis on the discrimination between forced-choice recognition targets and lures.

**Figure 4.** Stronger categorical representations predict memory performance. Model coefficients of the relationship between subsequent memory and Brain-Model Fit (± standard mean error) are displayed as a function of representation type and ROI. Across both tasks only the strength of categorical representations is associated with memory performance. * p < .05, all p-values were FDR corrected.

**Participants rely on categorical features to discriminate forced-choice recognition targets and lures**

To confirm that categorical, but not sensory, representations were associated with memory performance we conducted a follow-up analysis to examine which type of features participants rely on when discriminating targets vs. lures. To do this, we generated MTLD values for each scene (the procedure is depicted in Figure 5A) and correlated such values (excluding 3 scenes with z-scores greater than 3 or less than -3) with the average forced-choice recognition memory for each scene.

To control for the influence of the alternate type of information, we conducted two partial correlations: the first correlating sensory MTLD with memory while controlling for categorical MTLD, and the second correlating categorical MTLD with memory while controlling for sensory MTLD.

As illustrated by Figure 5B, we found that forced-choice recognition performance was positively correlated with the target-lure dissimilarity values of the categorical features of images (r(90) = 0.33, p = .002), but not with the sensory features of images (r(90) = 0.15, p = .15). These results confirm that participants rely on categorical features when recognizing previously encoded scenes.

To directly compare the influences of categorical and sensory MTLD on recognition memory, we employed a likelihood ratio test. Similar to our approach with the fMRI data, we initially created a model with only the intercept. The second model included sensory MTLD values for each scene, and the third model incorporated both sensory and categorical MTLD values. These models were then compared sequentially. The likelihood ratio test showed that sensory MTLD (χ²(1) = 1.35, p = .25) did not significantly contribute to the model's explanatory power for scene recognition, compared to a model with only the intercept. Conversely, the model with categorical MTLD values demonstrated a significant fit to the data, compared to the sensory MTLD-only model (χ²(1) = 10.4, p = .001).

In sum, our follow-up analysis on target-lure discrimination converges with the conclusion from the mixed effect model on RSA data that categorical, but not sensory, representations predict forced-choice recognition memory.
Figure 5. Only the categorical features of scenes impact forced-choice recognition memory performance. (A). We investigated the impact of sensory and categorical features of scenes on forced-choice recognition performance. Specifically, we examined the relationship between each target image (e.g., the airport on the left) and its three lures (e.g., the three airports on the center). For each target image we calculated the dissimilarity \((1 - r)\) between the target image and its three lures, according to the sensory \((ds)\) and categorical \((dc)\) models. As we can see by the different values of \(ds\) and \(dc\), the model perceived the relationship between the target and lures differently depending on whether sensory or categorical features were considered. (B) We found a significant partial-correlation between memory performance and the average dissimilarity between the target and the lures only for categorical features, while controlling for dissimilarity of sensory features. The plus sign represents the average memory for the airport as a function of the target-lure dissimilarity, for the two types of features. ** \(p < .01\).

Discussion

In the current study, we investigated the contributions of sensory and categorical visual representations to scene memory. Our research yielded two primary findings. First, we demonstrated that the encoding of categorical scene representations was linked to memory performance. Second, we identified a positive correlation between memory accuracy and the categorical distinctiveness of a scene relative to its distractors. However, no significant relationship was found for sensory features. We discuss these findings in greater detail below.

Memory performance were predicted only by categorical scene representations.

Consistent with our hypothesis, the stronger encoding of categorical features was associated with better memory performance. Previous research about scene perception has shown that the neural representation of categorical scene features explains participants behavior across scene classification tasks (Groen et al., 2018; King et al., 2019), while memory research has shown that the categorical features of scenes are related to scene memory (Isola et al., 2014; Konkle et al., 2010). Our results build upon these previous findings, demonstrating not only the significant role of categorical features in scene processing, but also illustrating how the neural representation of these categorical features supports subsequent memory for scenes.

The influence of higher-level features on subsequent memory performance has also been demonstrated in a recent study by Liu et al. (2021). In their research, they utilized intracortical electroencephalographic recordings while subjects viewed objects and subsequently underwent memory testing. By employing DNNs to model visual and semantic features, the authors found that greater representational transformations from sensory to categorical features correlated with improved memory for objects. To our knowledge, no study has directly compared the transformation of sensory to semantic representations of scene features and the impact of this transformation for memory. Nonetheless, considering that higher-level categorical information in scenes has been shown to emerge as early as the first 100 ms of visual processing (Lowe et al., 2018), one might speculate that scenes are a type of visual stimulus undergoing a rapid
transformation from sensory to categorical representations, which could explain their mnemonic advantage (Standing, 1973).

In our current study, we have demonstrated that categorical representations within regions spanning the occipital, parietal, and temporal lobes predict subsequent memory for naturalistic stimuli. Our findings are in line with those of a recent study conducted by Hebscher et al. (2023). In their research, they showed that within the same regions more similar neural representations of naturalistic stimuli (recordings of subjects performing different actions) were correlated with enhanced memory performance. While the authors used DNNs to model the sensory and categorical features of their stimuli and identified the type of information encoded in these ROIs, they did not link the brain-model fit values to subsequent memory. To gain a comprehensive understanding of how neural representations influence memory for naturalistic stimuli, future studies should strive to establish links between both types of representational information: feature representation and neural similarity between events.

Our current findings provide additional insights into the role of memory representations, building upon earlier research from our group. In their exploratory study, Davis et al. (2021) found that in the inferior temporal gyrus, both types of representations predicted object memory, while only sensory representations predicted subsequent memory in the precuneus. Contrasting with these earlier findings, our current study reveals a distinct pattern. We found that in both the inferior temporal gyrus and the precuneus, only the representations of categorical features are correlated with scene memory. This different pattern of results could be explained by how scenes context influences object processing (Biderman & Mudrik, 2018; Furtak et al., 2022) which, in turn, affects the neural representations of those objects, rendering these representations context-dependent (Lowe et al., 2017; Wang et al., 2018). Therefore, while further studies are necessary to confirm this hypothesis, scene context could provide additional semantic support for individual objects, making their sensory characteristics less significant for subsequent memory performance.

In direct contrast to our findings, Bone et al. (2020) reported a contradictory pattern, such that the reactivation of early visual features of images, and their corresponding cortical sensory representations were predictive of subsequent forced-choice recognition. One possible explanation for the opposite direction of these results is that the nature of the distractors used in these two experiments differed in one critical factor: their visual dissimilarity with the target image. We discuss this critical factor in the context of our second main finding in the next section.

**The effect of lure dissimilarity confirms the role of categorical representations on forced-choice recognition memory**

Our second main finding was that the categorical dissimilarity between the target image and its lures predicted forced-choice recognition memory, but sensory dissimilarity has no such influence on scene memory. Previous research about memory for scenes has shown that the encoding of more similar scenes according to their categorical, but not sensory, features negatively influences memory performance (Konkle et al., 2010). Our results expand this previous finding by showing that only the categorical (dis)similarity of the target and distractor images can predict forced-choice recognition memory.
To the best of our knowledge, this is the first study to model the relationship between a target image and its distractors using different types of visual features to investigate their unique contributions to forced-choice recognition memory. This relationship can help to explain the difference between the results of Bone et al. (2020) and our results. We hypothesized that while the distractors used in our study were more dissimilar in term of their categorical features (e.g., lures representing different models of airplanes), and thus participants relied more on these types of features to recognize them, the distractors used by Bone et al. (2020) were more dissimilar on their sensory features (e.g., a picture of the same statue with a differently colored background), thus making participants relying more on those features during the memory test.

**Limitations and further considerations**

One limitation of our study that could have influenced our results is related to the nature of the encoding task. We asked participants to rate how well the scene image represented its category, a task that relies on participants paying attention to the more categorical features of scenes, while not attending their more basic sensory features. By making participants pay more attention to the categorical features of images, they could have encoded more strongly these types of features. An important follow up experiment would be to test if making participants focus on the basic sensory aspects of the scenes, in order to test how that influences the encoding of sensory features.

Given the rapid development of neural network models of vision it is important to acknowledge a potential limitation of our study. While DNNs have been used as a model of visual processing (Jozwik et al., 2023; Peters & Kriegeskorte, 2021), there has been an increased attention on the limits between this computational model and its ability to explain human visual processing (Xu & Vaziri-Pashkam, 2021). While we recognize this as a limitation, it’s worth noting that we employed a DNN architecture widely used on memory research (Bone et al., 2020; Davis et al., 2021), trained on a database that has been shown to generate accurate model of brain activity during visual scene processing (Cichy et al., 2017; Greene & Hansen, 2020; Groen et al., 2018). Nevertheless, it would be an important step forward to test if more biologically plausible DNN, such as architectures with a recurrent structure (Kubilius et al., 2018), offers an improved model to study the influence of visual representations on memory.

Although not the central focus of our study, an intriguing avenue for future research emerges from our result where early stages of visual processing displayed higher brain-model fit values for categorical features compared to sensory features. This outcome, while unexpected, aligns with several studies that have reported higher brain-model fit values for later DNN layers as opposed to early ones in early visual regions (Davis et al., 2021; Devereux et al., 2018; Groen et al., 2018). Considering the reciprocal connections between early and late regions within the visual cortex hierarchy (Bastos et al., 2012; Muckli et al., 2015), it may not be surprising that early visual regions show sensitivity to visual features modeled by the later layers of a DNN, particularly when taking into account the integration of neural signals recorded through the slow acquisition time of fMRI.

Nonetheless, further investigation utilizing more time-sensitive measures is required to fully explore this phenomenon.

**Conclusions**
By integrating computational models of scene processing with fMRI data, our study has yielded evidence supporting the role of categorical scene representations in memory. Firstly, our findings demonstrate that robust encoding of categorical features contributes to scene memory. Secondly, we established that categorical representations also play an important role in scene recognition. Furthermore, our results suggest that, when confronted with stimuli that encompass multiple layers of visual details, the brain effectively utilizes pre-existing knowledge (captured by categorical models) to perceive and recall these complex environments. In summary, our study advances our understanding of the specific visual information that underlies our ability to recall and navigate our surrounding visual environment.
References


Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. Scientific Reports, 8(1), 10636. https://doi.org/10.1038/s41598-018-28865-1


