Cognition 245 (2024) 105723

Contents lists available at ScienceDirect

Cognition

journal homepage: www.elsevier.com/locate/cognit

Short communication

# Individual differences in internal models explain idiosyncrasies in scene perception

Gongting Wang [a,b,1], Matthew J. Foxwell [c,1], Radoslaw M. Cichy [a], David Pitcher [c], Daniel Kaiser [b,d,*]

[a] Department of Education and Psychology, Freie Universität Berlin, Germany
[b] Department of Mathematics and Computer Science, Physics, Geography, Justus-Liebig-Universität Gießen, Germany
[c] Department of Psychology, University of York, UK
[d] Center for Mind, Brain and Behavior (CMBB), Philipps-Universität Marburg and Justus-Liebig-Universität Gießen, Germany

ARTICLE INFO

*Keywords:*
Scene perception
Individual differences
Drawing
Predictive processing
Internal models

ABSTRACT

According to predictive processing theories, vision is facilitated by predictions derived from our internal models of what the world *should* look like. However, the contents of these models and how they vary across people remains unclear. Here, we use drawing as a behavioral readout of the contents of the internal models in individual participants. Participants were first asked to draw typical versions of scene categories, as descriptors of their internal models. These drawings were converted into standardized 3d renders, which we used as stimuli in subsequent scene categorization experiments. Across two experiments, participants' scene categorization was more accurate for renders tailored to their own drawings compared to renders based on others' drawings or copies of scene photographs, suggesting that scene perception is determined by a match with idiosyncratic internal models. Using a deep neural network to computationally evaluate similarities between scene renders, we further demonstrate that graded similarity to the render based on participants' own typical drawings (and thus to their internal model) predicts categorization performance across a range of candidate scenes. Together, our results showcase the potential of a new method for understanding individual differences – starting from participants' personal expectations about the structure of real-world scenes.

## 1. Introduction

Scene perception is not only achieved through a passive analysis of sensory input. Instead, the brain actively creates predictions about the world that are compared against current inputs (Clark, 2013; Friston, 2005, 2010). In cognitive science, this idea was first highlighted by schema theory, which postulated that inputs are referenced against internal models (schemata) stored in memory, which reflect the structure of the world (Bartlett, 1932; Minsky, 1974; Rumelhart, 1980; Wagoner, 2013). Schema theory was influential in early research on human memory (Brewer & Treyens, 1981; Mandler & Parker, 1976) and perception (Biederman, 1972; Biederman, Mezzanotte, & Rabinowitz, 1982). More recently, the importance of internal models has been highlighted by theories of Bayesian inference (Kayser, Körding, & König, 2004; Yuille & Kersten, 2006) and predictive processing (Clark, 2013; Keller & Mrsic-Flogel, 2018). These theories assume that during visual processing, inputs are constantly matched against internally generated predictions of the world. Such predictions are derived from our own internal models of what we think the world *should* look like. How can we characterize the contents and individual differences of these internal models?

In the context of scene perception, internal models can be conceptualized as a collection of typical features of a scene (or scene category) that are learned from extensive real-life experience and guide the analysis of matching visual inputs. The contents of internal models are mainly inferred from carefully manipulating the structure of the visual input and observing the resulting changes in perceptual performance and neural representation. Using this approach, researchers could successfully infer key features of internal scene models, such as the typical spatial distributions of objects (Bar, 2004; Biederman et al., 1982; Kaiser, Quek, Cichy, & Peelen, 2019), semantic relationships between objects and scenes (Davenport & Potter, 2004; Oliva & Torralba, 2007; Vo,

* Corresponding author at: Department of Mathematics and Computer Science, Physics, Geography, Justus-Liebig-Universität Gießen, Germany.
*E-mail address:* danielkaiser.net@gmail.com (D. Kaiser).
[1] These authors contributed equally.

https://doi.org/10.1016/j.cognition.2024.105723
Received 14 July 2023; Received in revised form 12 January 2024; Accepted 14 January 2024

Boettcher, & Draschkow, 2019; Wolfe, Võ, Evans, & Greene, 2011), or the spatial layout of whole scenes (Biederman, 1972; Kaiser & Cichy, 2021; Kaiser, Häberle, & Cichy, 2020).

However, this approach only reveals the contents of internal models that are shared across people – although there is mounting evidence for individual variability in visual perception and neural representation (Charest, Kievit, Schmitz, Deca, & Kriegeskorte, 2014; De Haas, Iakovidis, Schwarzkopf, & Gegenfurtner, 2019; Gauthier, 2018; Mollon, Bosten, Peterzell, & Webster, 2017; Tulver, Aru, Rutiku, & Bachmann, 2019; Wang, Li, Fang, Tian, & Liu, 2012). Given that we all differ in our visual experience with scenes across the lifetime (Coutrot et al., 2022; Hartley, 2022) and in our neural architecture for visual analysis (Kanai & Rees, 2011; Llera, Wolfers, Mulders, & Beckmann, 2019; Moutsiana et al., 2016), it is likely that internal models for scenes are sculpted in different ways across people. If we could harness this individual variability, we would be able to predict and explain characteristic differences in the way each of us perceives the world.

Here, we thus developed a novel approach that focuses on distilling out key properties of internal models in individual participants. We achieved this through drawing, enabling participants to provide unconstrained descriptions of typical scenes both quickly and without prior training (Fan, Bainbridge, Chamberlain, & Wammes, 2023). Using these drawings as descriptors for internal scene models, we then tested whether individual participants' scene perception can be explained through similarities with their personal internal models.

Our participants first drew typical exemplars of natural scenes categories, as well as copies of photographs of the same categories (which served as a control for familiarity acquired during drawing). They then performed a scene categorization task, in which they viewed carefully constructed scene renders that were created based on the drawings. Across two experiments, participants were more accurate in categorizing renders based on their own drawings, compared to renders based on other people's drawings and renders based specific scenes they copied. Our results provide evidence that individual differences in internal models explain individual differences in scene categorization.

## 2. Materials and methods

### 2.1. Experiment 1

#### 2.1.1. Participants

43 participants completed the drawing session, 39 returned for the categorization task. 4 were excluded for performing at chance level (binomial test), leaving a sample of 35 participants (22.6 ± 4.3 years ± SD, 6/29 male/female). Procedures were approved by the ethics committee of the Department of Psychology, University of York, and adhered to the Declaration of Helsinki. Experiment 1 was conducted online and participants provided informed consent through an online form. Sample size was based on convenience sampling, with the target to exceed 80% statistical power for a hypothesized medium effect of d = 0.5 in a two-sided *t*-test. For this target, at least 34 participants are required.

#### 2.1.2. Drawing sessions

The drawing session took part online, using Skype. Participants were tasked with drawing scenes from two categories: living rooms and kitchens (Fig. 1a). Critically, they were instructed to draw what they conceived as the most typical example of the scene category. The definition of typical was given as the most generic and ordinary example they could think of. They were instructed not to draw a scene that they thought looked particularly interesting or attractive. They were further instructed to not simply draw an exact copy of a scene they knew from real life such as their own kitchen or living room (though they were reassured that similarities with known scenes did not need to be deliberate avoided). Participants were given 1 min to plan and think about what their most typical scene should look like and 3:30 min to draw the scene using a pencil, eraser, and ruler. Scene sketches were drawn into a standardized perspective grid, to allow participants to draw in 3d more easily as well as to match viewpoints across scenes. Grids were either drawn or printed by the participant on A4 paper and consisted of a large central rectangle (7.1 cm by 16.5 cm) and 4 diagonal lines going from each corner of the rectangle to the corners of the page. The rectangle was

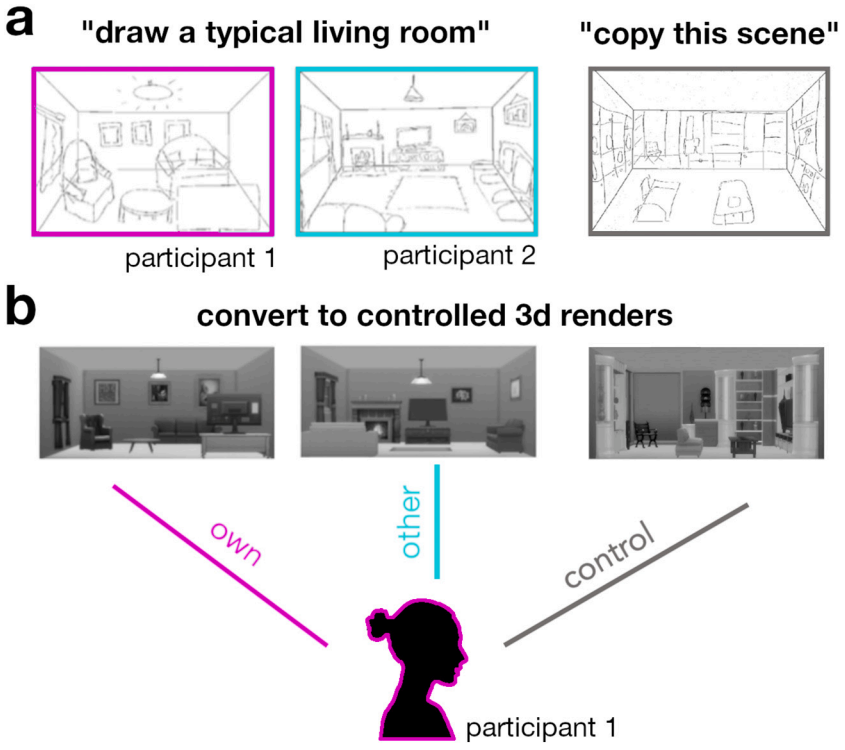


**Fig. 1.** Drawing session and stimuli. a) Participants drew typical versions of natural scenes and copies of scene photographs. b) Drawings were converted into 3d renders, based on each participant's own typical drawings, other participants' typical drawings, or scene copies.

placed 8.5 cm from the bottom of the page and 5.4 cm from the top of the page.

While drawing, participants were reminded how much time they had left at the halfway point, and when they had a minute remaining. They first drew a practice scene of a bedroom, to get them used to the timings and drawing on the perspective grid. In addition to the living rooms and kitchens, they also drew garden scenes, which were collected for another experiment. The order in which they drew the scene categories was balanced across participants. After completing each drawing, participants also drew a coarse birds-eye view of the scene, in which they labelled all objects in the scene. This was done to help clarify the room's intended 3d layout and to confirm the identity of any ambiguously drawn objects, providing additional information for generating accurate 3d renders later.

Typical kitchen drawings consisted of an average of 8.32 objects (SD = 2.03, range = [5, 12]), with the three most frequent objects cupboard (100% of drawings), hob (100%), and sink (81.40%). Living rooms consisted of an average of 8.03 objects (SD = 1.92, range = [5, 12]), with the three most frequent objects sofa (100%), table (100%), television (88.37%).

After drawing their most typical versions of the scenes, participants drew copies of given photographs of the same scene categories (Fig. 1a). Photographs were chosen to be clear exemplars of the scene category, while not being very prototypical in their visual appearance. This was done to reduce incidental similarities between the copied scenes and participants' typical drawings. All participants copied the same photographs. The copies were drawn under the same time constraints as the typical drawings, and participants were instructed to capture a similar amount of detail as they used in their drawings of typical scenes. Participants had access to the photograph throughout their drawing time. These copies acted as a control for familiarity effects in the subsequent categorization experiment: Participants will have seen and drawn these scenes, just like their typical versions, but they will not adhere to their internal models of what these scenes typically look like.

#### 2.1.3. Scene renders

We created 3d renders from the drawings by placing suitable candidate objects into an empty room. Scene renders were constructed using The SIMS4. The game includes a comprehensive design kit that allows the user to create a range of 3d environments by placing walls and objects onto a grid-like system (known in the game as "Build Mode"). The use of The SIMS4 allowed us access to a large library of thousands of 3d-modelled candidate objects for building the renders.

To create the renders, first an empty room was built to replicate the view and approximate dimensions of the perspective grid. This room was approximately 6 × 6 m in size and used wall pieces approximately 3 m high, with the outward facing wall removed. The scenes were then manually populated with objects by one of the authors, referencing both the scene sketch and birds-eye view plans the participants assembled in the drawing session. The closest matching 3d object was chosen to represent each object in the render. When objects were drawn in very little detail, a relatively generic version of the object was used at the author's discretion. Screenshots of the scenes were taken using the X box live app for Windows. Screenshots were taken from the same distance and angle for every scene render, cropped so that only the room was visible, and resized to 820 by 390 pixels. To control for low-level visual differences between the resulting images, all images were grayscaled and luminance-matched using the SHINE toolbox for MATLAB (Willenbockel et al., 2010).

In total, 88 renders were created: 86 renders were based on the typical drawings of 43 individual participants and 2 renders were based on the 2 control images (the control renders were based on the original images and thus identical for all participants).

#### 2.1.4. Categorization task

The categorization task was conducted online, using Gorilla (Anwyl-Irvine, Massonnié, Flitton, Kirkham, & Evershed, 2020). Participants were instructed to full-screen the application and sit approximately 60 cm from the screen. During the experiment, participants were asked to categorize briefly presented scene renders into kitchens versus living rooms. On each trial, a scene render was flashed for 83 ms, followed by a mask presented for 150 ms. Masks consisted of a random arrangement of squares, diamonds, and circles. A blank screen was then displayed until the participants responded by either pressing "K" or "L" on their keyboard (to indicate whether the scene was a kitchen or living room). There was no response time limit, but both accuracy and response time were stressed. Trials were separated by a 1-s inter-trial interval.

Participants viewed renders based on their own drawing of a typical scene ("own" condition), based on other participant's drawings of typical scenes ("other" condition), and based on their copied scenes ("control" condition; the control renders were identical for all participants). In total, 88 renders were shown in the experiment, 2 of which corresponded to each participant's own drawings, 2 of which corresponded to the copied scenes, and 84 that corresponded to the other participants' drawings (based on the 43 participants who initially completed the drawing session). Each render was repeated 10 times, for a total of 880 trials. Trial order was randomized. The experiment was split into four blocks. After each block, participants were given a 1:30 min break.

#### 2.1.5. Behavioral data analysis

Responses slower than 5 s were discarded. Accuracies and response times were compared using one-way ANOVAs and *t*-tests. For the response times, only trials with correct responses were analyzed. Statistical tests were conducted in Jamovi (www.jamovi.org).

#### 2.1.6. Reliability across participants

To assess the variation in categorization performance across participants, we correlated (*Spearman* correlations) the categorization accuracies across all scenes between participants. Split-half reliability of categorization accuracies (assessed by splitting our participant group into random halves 10,000 times and averaging correlations between halves for each split) was moderate, $r = 0.72$. Data from individual participants was only relatively weakly predicted by the average of all other participants, $r = 0.35$.

#### 2.1.7. Graded similarity analysis

To investigate whether graded similarity to the internal model predicts processing efficiency across scenes, we employed a deep neural network (DNN) analysis. In this analysis, we quantified how similar each scene render in the *other* condition (hereinafter: candidate scenes) is to the renders in the *own* condition (hereinafter: reference scenes), and then correlated the resulting graded similarity score with participants' categorization accuracy across scenes. To test whether graded similarity specifically to the *own* scene (and thus to participants' internal models) predicts behavioral accuracy, we alternatively used each *other* scene as a reference (and subsequently averaged over all different other scenes) or used the *control* scene as a reference. For all candidate and reference scenes, we first extracted activation vectors from a googlenet DNN (Szegedy et al., 2015). We used two variants of this network, either trained on object categorization using the ImageNet dataset (Deng et al., 2009) or on scene categorization using the Places365 dataset (Zhou, Lapedriza, Khosla, Oliva, & Torralba, 2017). The output of the final inception module (5b) of the DNN was used as an approximation for complex visual feature processing (Kriegeskorte, 2015).

By systematically correlating the extracted activation vectors (*Spearman* correlations), we obtained two similarity relations: (i) within-category similarities, capturing how similar each candidate scene is to the reference scene of the same category, and (ii) between-category similarities, capturing how similar each candidate scene is to the reference scene of the competing category. By subtracting the between-correlations from the within-correlations, we created a graded similarity

score, which captured how similar each candidate scene render was to the reference render of the same category, relative to the reference render of the competing category. Similarity scores were collapsed across the two scene categories. To determine how well similarity scores in the early and late DNN layers predict categorization, we then correlated them with the categorization accuracies for all candidate scenes (*Spearman* correlations), separately for the *own*, *other*, and *control* scenes as the reference. Notably, each time one of the *other* scenes was the reference, one scene less was available for computing the correlations. For the analyses in which the *own* or *control* scenes were the reference, we thus iteratively removed one of the *other* scenes before computing the results and then averaged across iterations. Finally, correlations were Fisher-transformed, and statistically assessed across participants using one-way ANOVAs and paired *t*-tests.

### 2.2. Experiment 2

#### 2.2.1. Participants

36 participants completed the drawing sessions and 35 participants (23.9 ± 2.6 years ± SD, 8/27 male/female) returned for the categorization task. Procedures were approved by the ethics committee of the Department of Education and Psychology, Freie Universität Berlin, and adhered to the Declaration of Helsinki. Experiment 2 was conducted in the lab, and participants provided written informed consent. The sample size was not increased compared to Experiment 1, as we expected the lab-based experiment to yield stronger effects than the online experiment.

#### 2.2.2. Drawing sessions

The drawing sessions were similar to Experiment 1 but were conducted in the lab. Participants provided their drawings on an Apple iPad Pro using an Apple Pencil. Drawings were created using the Sketchbook app. A standardized perspective grid similar to the one used in Experiment 1 was provided for each drawing. Specifically, the full drawing display (19.5 cm by 26.1 cm) consisted of a large central rectangle (8.5 cm by 13.5 cm) and 4 diagonal lines from each corner of the rectangle to the corners of the page. The rectangle was set with the bottom length 8.5 cm from the bottom of the page and top length 2.7 cm from the top of the page. Instructions and timings were identical to Experiment 1. Here, however, participants drew typical versions and copies of six scene categories (bathroom, bedroom, café, kitchen, living room, and office). Before making their drawings, participants drew a typical classroom to practice drawing under the experimental constraints. As the bird's eye view drawings were not critical for generating the renders in Experiment 1, we did not ask participants to draw the bird's eye views in Experiment 2.

Typical bedroom drawings consisted of an average of 8.1 objects (SD = 1.7, range = [4, 11]), with the three most frequent objects bed (100% of drawings), window (89%), and carpet (81%). Kitchens consisted of an average of 8.6 objects (SD = 2.1, range = [4, 13]), with the three most frequent objects cupboard (100%), hob (100%), and sink (89%). Living rooms consisted of an average of 8.0 objects (SD = 2.0, range = [4, 12]), with the three most frequent objects sofa (100%), table (100%), and television (92%). Bathrooms consisted of an average of 7.7 objects (SD = 1.6, range = [4, 12]), with the three most frequent objects sink (100%), mirror (94%), and carpet (89%). Offices consisted of an average of 7.9 objects (SD = 1.9, range = [4, 12]), with the three most frequent objects desk (100%), chair (100%), and computer (78%). Cafés consisted of an average of 8.0 objects (SD = 2.34, range = [3, 19]), with the three most frequent objects table (100%), chair (100%), and coffee machine (72%).

#### 2.2.3. Scene renders

Scene renders were created in the same way as for Experiment 1.

#### 2.2.4. Categorization task

Here, the categorization task was conducted in the lab, using the Psychtoolbox for Matlab (Brainard, 1997). During the experiment, participants categorized the scenes into the six categories. Renders were presented at central fixation with 7° horizontal visual angle. Trial design was identical to Experiment 1. To accommodate the six response options, participants saw a response screen after the mask, on which they indicated which of the six categories they had just seen, using the "S", "D", "F", "J", "K", and "L" keys on the keyboard.

In Experiment 2, we sought to make the design more efficient by not showing all renders to every participant. We instead grouped participants into groups of 4, and each participant only saw renders based on their own drawings, renders based on the other 3 participants' drawings, and the control renders created from the scene copies. Groups of 4 were chosen so that there was still sufficient variability in visual stimuli across the experiment. Each participant thus saw 5 renders for each of the 6 categories. Each of these 30 stimuli was repeated 40 times across the experiment, for a total of 1200 trials. Trial order was randomized. The experiment was split into 4 blocks. After each block, participants could take a break for as long as they needed.

For some participants, we also recorded EEG during the categorization task, in order to obtain preliminary data for another study.

#### 2.2.5. Behavioral data analysis

Accuracies and response times were analyzed in the same way as for Experiment 1.

#### 2.2.6. Post-experiment questionnaire

After the categorization task, participants completed a brief post-experiment questionnaire on paper, featuring three questions. First, we asked participants whether they felt familiar with any of the stimuli they had just seen. Here, only 3 out of the 35 participants reported that some of the scenes felt familiar to them, and analyzing the data without these three participants did not change the pattern of results. Second, we showed participants 24 scene renders (4 renders per category, with 1 corresponding to participants' own drawing and 3 corresponding to drawings from the other participants in the same group). The four renders for each category appeared on a separate questionnaire page. For each category, we asked them to circle the scene that they felt was most typical. Here, 86% of participants picked the render corresponding to their own drawing as the most typical, suggesting that the renders captured the typicality of their drawings. Finally, we showed participants the same set of 24 images again, but this time specifically asked them to circle the image that is most similar to the typical drawing they had produced in the drawing session. Here, 83% of participants correctly identified their drawing. This is not surprising, given that the categorization task was typically only conducted a week from the drawing experiment and the unlimited viewing time.

#### 2.2.7. Typicality rating experiment

As the control renders were chosen at the experimenter's discretion, we assessed whether these renders were indeed typical for the scene category. To this end, an online survey featuring typicality ratings for all scene renders was conducted via Limesurvey (http://www.limesurvey.org). Specifically, 216 scene renders (6 categories by 35 participants, plus the control renders) were presented in random order, together with their category label (e.g., "kitchen"), and an independent group of online participants ($n$ = 22, 21.3 ± 2.8 years±SD, 2/20 male/female) was asked to rate the typicality of the renders from 1 to 5 (1: not at all typical for the category, 5: very typical for the category). The resulting data showed that the control scenes were of intermediate-to-high typicality. Specifically, the typicality of the control renders ranked at the following positions out of the 36 renders: 17 (bedroom), 14 (kitchen), 15 (living room), 10 (bathroom), 8 (office), and 14 (café). Any differences between the *own*, *other*, and *control* conditions thus cannot be readily attributed to the copied scenes being less typical.

### 2.3. Open practices statement

Data, code, and materials for both experiments are accessible on the Open Science Framework (OSF), at https://osf.io/hpw9v/.

## 3. Results

In our experiments, participants first completed a drawing session, in which they drew typical versions of scene categories, such as a typical living room (Fig. 1a; see Materials and Methods). These drawings were used as descriptors of their internal models of scenes. They also copied photographs of scenes from the same categories, which later served as a control for familiarity acquired during drawing. For the subsequent experiments, we transformed these drawings into standardized 3d renders (Fig. 1b), thereby leveling out individual differences in drawing skill and style.

We then tested whether scenes designed to mimic individual participants' internal models (described by their own typical drawings) are more accurately perceived than scenes that were designed to mimic other participants' internal models. To this end, we devised a scene categorization task that required participants to accurately categorize the briefly presented and backward-masked scene renders.

In Experiment 1, participants completed an online drawing session where they drew typical versions and copies of living rooms and kitchens. They then performed an online categorization task (Fig. 2a), during which they categorized the scene renders created from participants' drawings into kitchens versus living rooms.

To investigate whether scenes that were specifically tailored to participants' personal internal models are more accurately categorized, we compared accuracies between renders based on each participant's own drawing ("own" condition), other participants' drawings ("other" condition), or the copied scenes ("control" condition). Accuracy was significantly different between conditions, $F(2,68) = 4.15$, $p = .020$, partial $\eta^2 = 0.11$ (Fig. 2b), with two significant pairwise comparisons: First, renders in the *own* condition were more accurately categorized than in the *other* condition, $t(34) = 2.18$, $p = .036$, $d = 0.37$, indicating that idiosyncrasies in categorization are indeed related to individual differences in internal models. Second, renders in the *own* condition were also more accurately categorized than in the *control* condition, $t(34) = 2.26$, $p = .031$, $d = 0.38$, indicating that the categorization advantage for renders based on typical drawings cannot be explained with participants acquiring familiarity with their drawings during the drawing session. No difference was found between the *other* and *control* conditions, $t(34) = 1.11$, $p = .28$, $d = 0.19$. Response times were also significantly different between the conditions, $F(2,68) = 3.62$, $p = .032$, partial $\eta^2 = 0.10$. Specifically, the *control* condition ($M = 828$ ms, $SE = 37$) yielded greater response times than the *own* condition ($M = 784$ ms, $SE = 30$), at the trend level, $t(34) = 1.87$, $p = .070$, $d = 0.32$, and the *other* condition ($M = 788$ ms, $SE = 31$), $t(34) = 3.18$, $p = .003$, $d = 0.54$. No difference was found between the own and other conditions, $t(34) = 0.26$, $p = .79$, $d = 0.04$.

In Experiment 2, we aimed to replicate the key result from Experiment 1 (enhanced categorization performance for *own* compared to *other* and *control* scenes) in a lab-based setup and for a wider range of scene categories. Here, participants first drew typical versions and copies for six scene categories: bathroom, bedroom, café, kitchen, living room, and office. In the subsequent categorization task, they again categorized renders based on their own, as well as other participants' drawings into the six categories (Fig. 3a).

Results fully replicated the pattern observed in Experiment 1. Categorization accuracies varied significantly across conditions, $F(2,68) = 8.05$, $p < .001$, partial $\eta^2 = 0.19$ (Fig. 3b), with higher accuracy in the *own* condition, compared to both the *other*, $t(34) = 2.89$, $p = .007$, $d = 0.49$, and *control* conditions, $t(34) = 3.61$, $p < .001$, $d = 0.61$. No difference was found between the *other* and *control* conditions, $t(34) = 1.59$, $p = .12$, $d = 0.27$. This again shows that scenes are categorized more accurately when they are similar to participants' typical scene drawings, suggesting that similarity to individual participants' internal models determines categorization performance. Response times were not significantly different between the conditions, $F(2,68) = 0.69$, $p = .51$, partial $\eta^2 = 0.02$.

The results from Experiments 1 and 2 demonstrate that scenes that are specifically tailored to match participants' internal models are more accurately categorized. However, if internal models indeed serve as templates for categorization, performance should gradually increase as a scene becomes (i) more similar to the internal model of the same category and (ii) more dissimilar to the internal model of another category. The data from Experiment 1, where each participant viewed a variety of individual scenes, allowed us to test this prediction. To objectively quantify such graded similarity, we used deep neural network (DNN) models trained on object or scene categorization, from which we extracted activations from a deep layer as a proxy for high-level feature processing. From these activations, we computed how similar each scene render in the *other* condition was to the same-category scene in the *own* condition versus the other-category scene in the *own* condition (Fig. 4a). We then correlated this graded similarity measure with behavioral categorization accuracies across scenes in Experiment 1. This analysis was repeated by comparing activation patterns for each other scene to the same- and different-category activation patterns from the *other* scenes or *control* scenes. If similarity to the internal model is indeed driving categorization, then we should see a better prediction of behavioral accuracy when the *own* scenes are used as a reference than when the *other* or *control* scenes are used as a reference.

In both the object- and scene-trained DNNs, graded similarity to the

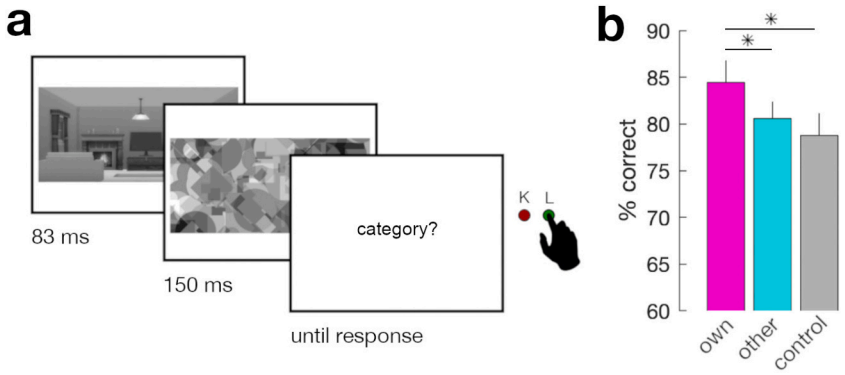


**Fig. 2.** Experiment 1 – behavioral results. a) Participants categorized briefly presented renders into kitchens versus living rooms. b) Categorization was more accurate for renders based on participants' own drawings (own condition) than for those based on other participants' drawings (other condition) or copies (control condition). Error bars represent standard errors of the mean. * indicates $p < .05$.

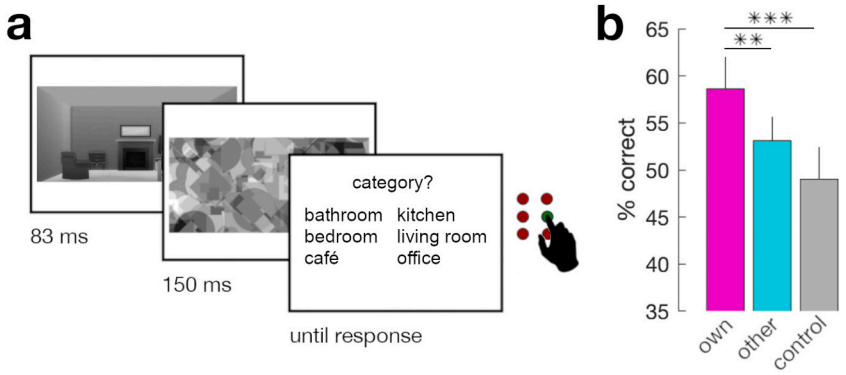


**Fig. 3.** Experiment 2 – behavioral results. a) Participants categorized briefly presented renders into six scene categories. b) Categorization was again more accurate for renders based on participants' own drawings (own condition) than for those based on other participants' drawings (other condition) or copies (control condition). Error bars represent standard errors of the mean. ** indicates $p < .01$, *** indicates $p < .001$.

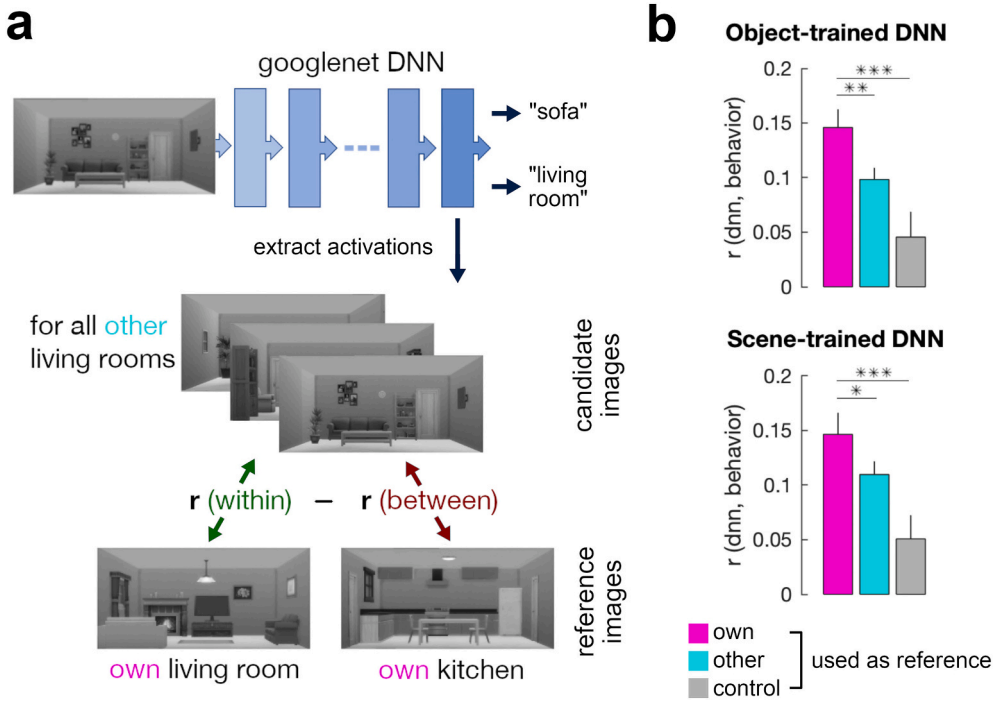


**Fig. 4.** Experiment 1 – graded similarity analysis. a) We extracted activation patters for all scene renders in Experiment 1 from googlenet DNNs trained on scene or object classification. To approximate the processing of complex, high-level visual features, we extracted activation patterns from the last inception module of the DNN. To quantify similarity to the internal model, we correlated the activation pattern for each other scene to the own scene of the same category (within-category correlation) and each own scene of the other category (between-category correlation), separately for each participant. By subtracting the within- and between-category correlations, we obtained a graded similarity measure, which we correlated with the behavioral categorization accuracy across all candidate images. This analysis was repeated with all possible other scenes or the control scenes as the reference images. b) In both DNNs, graded similarity to the own scene predicted categorization better than graded similarity to the other or control scenes, suggesting that similarity to participants' personal internal models predicts behavioral categorization across the range of images used in the experiment. Error bars represent standard errors of the mean. * indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$.

*own* and *other* scenes was positively correlated with categorization performance, all $t(34) > 7.37$, $p < .001$, $d > 1.25$ (Fig. 4b). Graded similarity to the *control* scenes was a weaker predictor of categorization, both in the object-trained, $t(34) = 1.96$, $p = .058$, $d = 0.33$, and scene-trained DNNs, $t(34) = 2.34$, $p = .025$, $d = 0.40$. Comparing predictions between the *own*, *other*, and *control* scenes as references, we found a significant difference between conditions in both networks, both $F(2,68) > 13.3$, $p < .001$, partial $\eta^2 > 0.28$. Critically, graded similarity to the *own* scenes between predicted behavioral performance better than graded similarity to the *other* scenes, $t(34) = 3.31$, $p = .002$, $d = 0.57$ (object-trained DNN), $t(34) = 2.26$, $p = .030$, $d = 0.39$ (scene-trained DNN), and better than graded similarity to the *control* scenes, $t(34) = 4.04$, $p < .001$, $d = 0.69$ (object-trained DNN), $t(34) = 4.65$, $p < .001$, $d = 0.80$ (scene-trained DNN). This confirmed our prediction that graded similarity to participants' individual internal models determines categorization performance.

## 4. Discussion

Together, our findings provide new insights on individual differences in naturalistic vision. We show that participants are better at categorizing scenes that resemble a typical drawing they had produced prior to

the experiment, compared to scenes that resemble other people's typical drawings, or scenes that resemble scene copies they had produced earlier. Using a DNN as a measure of graded similarity, we further show that categorization varies as a function of the similarity between participants' drawings and the scene that they are asked to categorize. We interpret these findings to reflect differences in participants' internal models of the world that are captured by their typical scene drawings. These differences in internal models may in turn drive idiosyncrasies in scene categorization.

The more accurate categorization of scenes that are similar to descriptions of participants' internal models can be explained by the rapid formation of accurate predictions that guide the analysis of the sensory input (Bar, 2004; Friston, 2005). It has been suggested that such predictions are generated through the activation of candidate prototypes from rapid and coarse stimulus analysis (Bar, 2004; Bar et al., 2006). This idea is consistent with previous studies reporting that participants – on the group level – show enhanced detection, categorization, and more diagnostic neural responses for more typical scene exemplars (Caddigan, Choo, Fei-Fei, & Beck, 2017; Csathó, van der Linden, & Gács, 2015; Torralbo et al., 2013). Here, we show that the activation of such categorical prototypes occurs in an idiosyncratic way, where each individual activates their own internal models of a scene. This reinforces the idea that internal representations of the world are only fully understood if we take the differential experience of individual observers with their real-world environments into account (Hartley, 2022). This assertion does not imply that perception is fully unique, or even radically different, between observers. We still found a fair reliability of categorization performance across observers, with a modest split half-reliability of $r = 0.72$ in Experiment 1 (see Materials and Methods). What our results do suggest is that on top of this coarse stability in performance, there is interesting additional variance that is systematic across observers and can be captured by our drawing-based method.

We demonstrate that a single drawing of a typical scene is able to capture essential properties of the individually specific internal model that gives rise to these predictions. This highlights the potential of our approach: A simple drawing composed in just a few minutes is enough to capture characteristic properties of the internal models in individual participants. While a single drawing thus seems sufficient to uncover individual differences, our approach is somewhat simplistic, as it assumes that (1) the internal model is a single point in the space of possible scenes and (2) the internal model is stable across time. Moving forward, it would be very interesting to see how internal models vary when probed with multiple drawings and across time. Such studies could reveal that internal models, rather than providing a single monolithic reference point, are perhaps defined by a probability distribution in representational space.

Our findings further suggest that familiarity acquired during drawing is insufficient to explain categorization benefits for stimuli that are similar to it. Renders created from the scenes that people copied, and that they also acquired familiarity with during drawing, did not yield the same performance benefit as renders that were created from drawings that reflect participants' own typical scenes. This shows that the generation of a drawing per se – the copy drawings were produced under the exact same constraints as the typical drawings – does not produce performance benefits in a subsequent task. Another concern relates to the mental construction of a scene, which is more demanding for the typical scene where the scene contents need to be thought up without a direct visual reference. Mental generation has indeed been linked to subsequent memory benefits in the memory literature, referred to as the generation effect (Clark, 1995; Slamecka & Graf, 1978). Though generation effects in memory are mostly probed on purely semantic contents and under long presentation regimes (Bertsch, Pesta, Wiscott, & McDaniel, 2007), generation may in principle lead to more pronounced familiarity in the subsequent categorization task. Our graded similarity analysis argues against our effects being driven solely by a preferential recognition of renders constructed from the typical drawings that participants had mentally generated before: Categorization also varied in a systematic way across renders based on other participants' drawings, as a function of how similar they were to the render based on their own typical drawing.

Our results may still be related to familiarity with scenes acquired throughout our lifetimes: The scenes we encounter during everyday experience ultimately eventually led to the formation of our internal models for scene categories. Previous studies indeed suggest that familiarity modulates scene processing (Bainbridge & Baker, 2022; Epstein, Higgins, Jablonski, & Feiler, 2007; Epstein, Parker, & Feiler, 2007; Klink, Kaiser, Stecher, Ambrus, & Kovács, 2023). In our study, we explicitly instructed our participants to not draw individual scenes from their immediate real-life experience but to draw the most typical scenes they could think of (with the idea that typical scenes reflect a weighted mix of features encountered in scenes across life). Thoroughly disentangling effects of typicality and familiarity in creating the reported effects will nonetheless require further studies. To comprehensively address this issue, studies need to either (i) track participants longitudinally, monitoring how their internal models change as they learn about new types of environments, or (ii) construct detailed descriptors of participants' visual experience, for instance by collecting descriptions and images from their everyday environments.

Our study further prompts interesting questions that provide new avenues for future research. First, we currently do not know *why* internal models systematically differ across participants. Future studies could relate variations in internal models to idiosyncrasies in cortical representation (Charest et al., 2014; Lee & Geng, 2017) and visual exploration behavior (De Haas et al., 2019; Henderson & Luke, 2014), as well as to individual differences in brain anatomy (Kanai & Rees, 2011; Llera et al., 2019; Moutsiana et al., 2016). Second, we do not know exactly *how* visual inputs are matched against the internal models. There is a variety of dimensions along which this match could be computed, such as the objects included in a scene as well as their spatial distribution (Kaiser et al., 2019; Oliva & Torralba, 2007; Vo et al., 2019; Wolfe et al., 2011), the global geometry of the scene (Epstein & Baker, 2019; Kaiser & Cichy, 2021; Oliva & Torralba, 2006), or low- and mid-level features correlated with the content of a scene (Geisler, 2008; Groen, Silson, & Baker, 2017; Watson, Hartley, & Andrews, 2014). Our DNN-based analysis of graded similarity indeed suggests that high-level features are important, given that graded similarity in a deep layer of a scene-trained DNN predicted categorization performance. The observation that predictions were enabled by both object- and scene-trained DNNs suggests that the features useful for prediction are not uniquely critical for either object or scene recognition. However, our scene renders were carefully matched for low-level features, and this matching may have obscured a possible contribution of low-level features that are relevant under more naturalistic conditions. To chart relevant visual features more comprehensively, future studies could systematically manipulate inputs to deviate from the internal model in targeted ways.

More generally, our study highlights the potential of drawing for quantifying internal representations (Fan et al., 2023). Drawings indeed received renewed attention recently, in studies of scene memory (Bainbridge & Baker, 2020; Bainbridge, Hall, & Baker, 2019) and perception (Fan, Yamins, & Turk-Browne, 2018; Matthews & Adams, 2008; Morgan, Petro, & Muckli, 2019; Ostrofsky, Nehl, & Mannion, 2017; Singer, Cichy, & Hebart, 2023). Our study suggests that drawings also yield the potential to advance our understanding of the internal models that guide the visual representation of objects, faces, or actions. Furthermore, our drawing method may prove useful for studying the maturation of internal models across development (see Long, Fan, Chai, & Frank, 2021) or their alterations in disorders of prediction like autism (Pellicano & Burr, 2012).

In sum, our work provides two critical advances for studying vision on the individual level. First, our findings offer a new interpretation of individual differences in perception. They suggest that humans categorize real-world environments in different ways because we all have

different internal models of the world. Second, our work provides researchers with a new drawing-based method for unveiling the contents of internal models in individual participants. This method has the potential to be widely applied to derive explicit predictions about individual differences in vision.

## CRediT authorship contribution statement

**Gongting Wang:** Writing – review & editing, Validation, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Matthew J. Foxwell:** Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Radoslaw M. Cichy:** Writing – review & editing, Supervision, Resources, Project administration, Funding acquisition. **David Pitcher:** Writing – review & editing, Validation, Supervision, Project administration. **Daniel Kaiser:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Formal analysis, Data curation, Conceptualization.

## Declaration of competing interest

The authors declare that there are no competing interests.

## Data availability

Link to data in the manuscript.

## Acknowledgements

G.W. is supported by a PhD stipend from the China Scholarship Council (CSC). R.M.C. is supported by the Deutsche Forschungsgemeinschaft (DFG; CI241/1-1, CI241/3-1, CI241/7-1) and by a European Research Council (ERC) starting grant (ERC-2018-STG 803370). D.K. is supported by the DFG (SFB/TRR135, project number 222641018; KA4683/5-1, project number 518483074), "The Adaptive Mind", funded by the Excellence Program of the Hessian Ministry of Higher Education, Science, Research and Art, and an ERC Starting Grant (PEP, ERC-2022-STG 101076057). Views and opinions expressed are those of the authors only and do not necessarily reflect those of the European Union or the European Research Council. Neither the European Union nor the granting authority can be held responsible for them. We thank Daniela Marinova for assisting in the drawing sessions for Experiment 1. The authors declare that there are no competing interests.

## References

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods, 52*, 388–407.

Bainbridge, W. A., & Baker, C. I. (2020). Boundaries extend and contract in scene memory depending on image properties. *Current Biology, 30*(3), 537–543.

Bainbridge, W. A., & Baker, C. I. (2022). Multidimensional memory topography in the medial parietal cortex identified from neuroimaging of thousands of daily memory videos. *Nature Communications, 13*(1), 6508.

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications, 10*(1), 1–13.

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience, 5*(8), 617–629.

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., … Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences, 103*(2), 449–454.

Bartlett, F. C. (1932). *Remembering: A study in experimental and social psychology*. Cambridge, UK: Cambridge University Press.

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition, 35*, 201–210.

Biederman, I. (1972). Perceiving real-world scenes. *Science, 177*(4043), 77–80.

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology, 14*(2), 143–177.

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial Vision, 10*(4), 433–436.

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology, 13*(2), 207–230.

Caddigan, E., Choo, H., Fei-Fei, L., & Beck, D. M. (2017). Categorization influences detection: A perceptual advantage for representative exemplars of natural scene categories. *Journal of Vision, 17*(1), 21.

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences, 111*(40), 14565–14570.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181–204.

Clark, S. E. (1995). The generation effect and the modeling of associations in memory. *Memory & Cognition, 23*(4), 442–455.

Coutrot, A., Manley, E., Goodroe, S., Gahnstrom, C., Filomena, G., Yesiltepe, D., … Spiers, H. J. (2022). Entropy of city street networks linked to future spatial navigation ability. *Nature, 604*(7904), 104–110.

Csathó, Á., van der Linden, D., & Gács, B. (2015). Natural scene recognition with increasing time-on-task: The role of typicality and global image properties. *Quarterly Journal of Experimental Psychology, 68*(4), 814–828.

Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological Science, 15*(8), 559–564.

De Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., & Gegenfurtner, K. R. (2019). Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences, 116*(24), 11687–11692.

Deng, J., Dong, W., Socher, R., Li, L. J., Li, K., & Fei-Fei, L. (2009, June). Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science, 5*, 373–397.

Epstein, R. A., Higgins, J. S., Jablonski, K., & Feiler, A. M. (2007). Visual scene processing in familiar and unfamiliar environments. *Journal of Neurophysiology, 97*(5), 3670–3683.

Epstein, R. A., Parker, W. E., & Feiler, A. M. (2007). Where am I now? Distinct roles for parahippocampal and retrosplenial cortices in place recognition. *Journal of Neuroscience, 27*(23), 6141–6149.

Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology, 2*(9), 556–568.

Fan, J. E., Yamins, D. L., & Turk-Browne, N. B. (2018). Common object representations for visual production and recognition. *Cognitive Science, 42*(8), 2670–2698.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 360*(1456), 815–836.

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127–138.

Gauthier, I. (2018). Domain-specific and domain-general individual differences in visual object recognition. *Current Directions in Psychological Science, 27*(2), 97–102.

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology, 59*, 167–192.

Groen, I. I., Silson, E. H., & Baker, C. I. (2017). Contributions of low-and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society, B: Biological Sciences, 372*(1714), 20160102.

Hartley, C. A. (2022). How do natural environments shape adaptive cognition across the lifespan? *Trends in Cognitive Sciences, 26*(12), 1029–1030.

Henderson, J. M., & Luke, S. G. (2014). Stable individual differences in saccadic eye movements during reading, pseudoreading, scene viewing, and scene search. *Journal of Experimental Psychology: Human Perception and Performance, 40*(4), 1390.

Kaiser, D., & Cichy, R. M. (2021). Parts and wholes in scene processing. *Journal of Cognitive Neuroscience, 34*(1), 4–15.

Kaiser, D., Häberle, G., & Cichy, R. M. (2020). Cortical sensitivity to natural scene structure. *Human Brain Mapping, 41*(5), 1286–1295.

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences, 23*(8), 672–685.

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience, 12*(4), 231–242.

Kayser, C., Körding, K. P., & König, P. (2004). Processing of complex stimuli and natural scenes in the visual cortex. *Current Opinion in Neurobiology, 14*(4), 468–473.

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. *Neuron, 100*(2), 424–435.

Klink, H., Kaiser, D., Stecher, R., Ambrus, G. G., & Kovács, G. (2023). Your place or mine? The neural dynamics of personally familiar scene recognition suggests category independent familiarity encoding. *Cerebral Cortex*. https://doi.org/10.1093/cercor/bhad397 (in press).

Kriegeskorte, N. (2015). Deep neural networks: A new framework for modeling biological vision and brain information processing. *Annual Review of Vision Science, 1*, 417–446.

Lee, J., & Geng, J. J. (2017). Idiosyncratic patterns of representational similarity in prefrontal cortex predict attentional performance. *Journal of Neuroscience, 37*(5), 1257–1268.

Llera, A., Wolfers, T., Mulders, P., & Beckmann, C. F. (2019). Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. *Elife, 8*, Article e44443.

Long, B., Fan, J., Chai, Z., & Frank, M. C. (2021). Parallel developmental changes in children's drawing and recognition of visual concepts. *PsyArXiv*. https://doi.org/10.31234/osf.io/5yv7x

Mandler, J. M., & Parker, R. E. (1976). Memory for descriptive and spatial information in complex pictures. *Journal of Experimental Psychology: Human Learning and Memory, 2*(1), 38.

Matthews, W. J., & Adams, A. (2008). Another reason why adults find it hard to draw accurately. *Perception, 37*(4), 628–630.
Minsky, M. (1974). A framework for representing knowledge. In P. Winston (Ed.), *The psychology of computer vision*. McGraw-Hill.
Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice? *Vision Research, 141*, 4–15.
Morgan, A. T., Petro, L. S., & Muckli, L. (2019). Scene representations conveyed by cortical feedback to early visual cortex can be described by line drawings. *Journal of Neuroscience, 39*(47), 9410–9423.
Moutsiana, C., De Haas, B., Papageorgiou, A., Van Dijk, J. A., Balraj, A., Greenwood, J. A., & Schwarzkopf, D. S. (2016). Cortical idiosyncrasies predict the perception of object size. *Nature Communications, 7*(1), 12110.
Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research, 155*, 23–36.
Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences, 11*(12), 520–527.
Ostrofsky, J., Nehl, H., & Mannion, K. (2017). The effect of object interpretation on the appearance of drawings of ambiguous figures. *Psychology of Aesthetics, Creativity, and the Arts, 11*(1), 99.
Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences, 16*(10), 504–510.
Rumelhart, D. E. (1980). Schemata: The building blocks of cognition. In J. R. Spiro (Ed.), *Theoretical issues in Reading comprehension*. CRC Press.
Singer, J. J., Cichy, R. M., & Hebart, M. N. (2023). The spatiotemporal neural dynamics of object recognition for natural images and line drawings. *Journal of Neuroscience, 43*(3), 484–500.
Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592.
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., … Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1–9).
Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater BOLD activity. *PLoS One, 8*(3), Article e58594.
Tulver, K., Aru, J., Rutiku, R., & Bachmann, T. (2019). Individual differences in the effects of priors on perception: A multi-paradigm approach. *Cognition, 187*, 167–177.
Vo, M. L. H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology, 29*, 205–210.
Wagoner, B. (2013). Bartlett's concept of schema in reconstruction. *Theory & Psychology, 23*(5), 553–575.
Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual differences in holistic processing predict face recognition ability. *Psychological Science, 23*, 169–177.
Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. *NeuroImage, 99*, 402–410.
Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods, 42*, 671–684.
Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences, 15*(2), 77–84.
Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences, 10*(7), 301–308.
Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 40*(6), 1452–1464.