

Empirical test of ideal observer models of individual and ensemble spatial perception

Yanina E. Tena Garcia (yanina.tena-garcia@psychol.uni-giessen.de)

Bianca R. Baltaretu (bianca.baltaretu@psychol.uni-giessen.de)

Katja Fiehler (katja.fiehler@psychol.uni-giessen.de)

Justus-Liebig-Universität Gießen

Dominik M. Endres (dominik.endres@uni-marburg.de)

Marburg University

Abstract

Individual and ensemble perception are crucial for interacting with objects in our environment. Individual perception processes single objects, while ensemble perception extracts summary information from object groups. To investigate how these two modes of perception work with different set sizes (3, 6, 10) in naturalistic settings, we compare two bayesian models on our data. The first model, a variant of the summation model, is the 'Individual Encoding Model'. The second model is the 'Ensemble Encoding Model', which is related to the automatic averaging model. We conducted an experiment in which participants encoded the position of an individual object or an ensemble position that summarized multiple objects in a 3D rendered scene and indicated its remembered position by mouse click on the screen. The 'Individual Encoding Model' assumes that each object's position is encoded in memory, the ensemble position is only evaluated on demand. In the 'Ensemble Encoding Model', the ensemble position is part of the process that generates the scene and is inferred from the observable object locations. We found that the accuracy of reproducing individual object positions increased as set size increased, while the estimation of the ensemble position (arithmetic mean) only differed between the 6- and 10-object set size conditions, with smaller deviations observed for scenes with 6 objects. The Ensemble Encoding Model generally explains the human behavioral data better. The subject-specific bayes factors in its favor increase with set size. We conclude that in naturalistic scenes the choice between individual versus ensemble encoding is likely driven by the more compact scene representation of the ensemble model.

Keywords: ensemble perception; spatial perception; scene perception; cognitive model; bayesian model

Introduction

Extracting object locations is a critical task in daily life. It requires determining the precise position of a specific object (e.g., picking up a coffee mug), referred to as individual perception, and identifying the average position of a group of objects (e.g., locating where mugs are placed in a kitchen), known as ensemble perception.

Individual perception involves isolating a target object from its surroundings by combining features like color, shape, and location into a unified representation (Kahneman et al., 1992; Perry & Fallah, 2014; Xu & Chun, 2009). This sequential and resource-intensive process requires specific brain regions (for a review see Perry & Fallah (2014)) and takes approximately 100 ms to 2 s to form an accurate object representation (Haberman & Whitney, 2007; Lauer et al., 2018; Melcher et al., 2021). Encoding efficiency is influenced by multiple factors such as the number of objects (Alvarez & Cavanagh, 2004; Melcher et al., 2021; Neumann et al., 2018; Robitaille & Harris, 2011) and scene context (Bar & Aminoff, 2003; Draschkow et al., 2014; Lauer et al., 2018; Ringer et al., 2021). Generally, a lower number of objects results in more precise representations (Melcher et al., 2021; Neumann et al., 2018), while a contextually appropriate scene (e.g., a pot in a kitchen rather than a bathroom) can enhance processing speed and recall accuracy (Bar & Aminoff, 2003; Draschkow et al., 2014; Lauer et al., 2018). However, naturalistic scenes may also hinder individual object perception, e.g. when objects are obscured by other elements in the scene (Ringer et al., 2021).

Ensemble perception, in contrast, is considered to be a rapid and automatic process (Oriet & Hozempa, 2016; Yildirim et al., 2018) that encodes summary information, such as the mean or variance of group features (Whitney & Yamanashi Leib, 2018). Ensemble processing occurs early in encoding and precedes detailed individual object processing (Haberman & Whitney, 2007; Melcher et al., 2021), and provides detailed ensemble information within 50–500 ms (Chong & Treisman, 2003; Haberman & Whitney, 2007; Melcher et al., 2021; Whiting & Oriet, 2011). Unlike individual perception, ensemble processing is either unaffected by the number of objects or benefits from larger sets (Chong & Treisman, 2003; Melcher et al. 2021; Robitaille & Harris, 2011; for a review see Corbett et al., 2023). Although ensemble perception is thought to aid orientation in complex environments, previous research primarily relied on the presentation of stimuli against simplistic backgrounds, e.g. (Melcher et al., 2021; Neumann et al., 2018). Consequently, the influence of scene context on ensemble perception remains unexplored.

These behavioral findings suggest that individual and ensemble perception are influenced differently by set size and have distinct temporal resolutions, indicating two distinct underlying processing mechanisms (Ariely, 2001; Chong &

Treisman, 2005; Ward et al., 2016). Ensemble information has been proposed to be extracted automatically (Oriet & Hozempa, 2016; Yildirim et al., 2018), bypassing the need for individual object representation (Ward et al., 2016). However, especially recent computational modeling studies challenge this view, suggesting that ensemble perception relies on pooling pre-encoded individual information (Harrison et al., 2021; Robinson & Brady, 2023; Utochkin et al., 2024). For instance, Robinson and Brady (2023) showed that their Perceptual Summation model, which posits that ensemble representations are formed by summing activations from individual targets during early encoding, outperformed the Automatic Averaging model across multiple tasks. The latter model implies that the ensemble representations are derived directly as single probability distribution without representing each individual object, just like for physically present targets.

Ensemble perception is essential for efficient processing of complex visual environments, as it reduces redundancies and eases the load on the capacity-limited visual system. While computational models of ensemble perception have been developed for various features (e.g. color, orientation; Harrison et al., 2021, Robinson & Brady, 2023), how spatial perception of both individual and ensemble information are processed has been hardly considered (Lew & Vul, 2015). Moreover, to the authors' knowledge, no previous work has examined individual and ensemble spatial perception within naturalistic scenes, an important aspect of characterizing visual perception processes, as it can both enhance and hinder visual processing, as already shown for individual perception (Bar & Aminoff, 2003; Draschkow et al., 2014; Lauer et al., 2018; Ringer et al., 2021).

The general goal of our study was to investigate how humans encode multiple object locations embedded in naturalistic scenes. In particular, we aimed to identify the computational encoding model that best accounts for the observed behavioral data, thereby gaining further insight into the possible computations underlying individual and ensemble perception. To this end, we compared two Bayesian perception models, each emphasizing different aspects of encoding individual and ensemble information. The first model, the Individual Encoding model (IEM), is related to the Perceptual Summation model (Robinson & Brady, 2023). It assumes that objects are initially encoded individually, and when required, the ensemble mean is computed as a separate sequential step. The second model, the Ensemble Encoding model (EEM), postulates that a scene is encoded by an ensemble location and object vectors relative to that location, see (Khoozani et al., 2019) for an earlier version of this model. The ensemble location is inferred during encoding, making the EEM conceptually similar to the Automatic Averaging model (Robinson & Brady, 2023). However, unlike that model, EEM does not assume the ensemble is perceived as a physically present target. Instead, it relies on (weighted) averaging of the total elicited activation without explicitly representing individual targets (Šetić et al., 2007).

In the following, we will first describe the methods of the behavioral experiment, followed by the development and evaluation of the encoding models. We then present the results of the behavioral experiment and assess the model fit to the behavioral data. Finally, we discuss the implications of our findings in the context of ensemble perception and individual encoding.

Behavioral Experiment

Participants

Data from 29 healthy students of Justus Liebig University Giessen (19 females) were collected. Participants were between 18 and 35 years of age (mean age = 23.58 ± 3.26 years), had normal or corrected vision, intact color vision (verified by Ishihara charts), no neurological or motor disorders, and were right-handed (EHI: $M = 79.34 \pm 20.81$; Oldfield, 1971). They gave their informed consent and received course credits or 8€/hour.

Stimuli and Apparatus

A total of 18 rendered pictures were created using Blender (v2.9; <https://www.blender.org/>) - six pictures for each of the three set size conditions (3, 6, and 10 objects). These pictures showed a kitchen environment in which the respective number of target objects (all kitchen-related) were randomly placed, without occluding one another (see Fig. 1). The target objects were either self-created in Blender or chosen from a repository that contained Blender-made objects (<https://www.turbosquid.com/de>).

Participants were seated in front of a 25" monitor (60 Hz, 1920 × 1080) in a dark room. A chinrest maintained a constant distance (60 cm) between their eyes and the monitor. To ensure central fixation at the start of each trial, a video-based Desktop Mount EyeLink 1000 (SR Research Ltd., Ontario, Canada; 1000 Hz sampling rate) eye-tracker was positioned 45 cm from the chinrest and below the line of sight.

Procedure

A trial started with a central fixation, followed by a three-second countdown, which participants were instructed to fixate (see Fig. 1). After the countdown, the Encoding phase started in which a test scene with either 3, 6 or 10 objects was presented for one of the three encoding times (50, 100 or 800 ms). Participants' task in the encoding phase was to encode the scene as accurately as possible. The encoding phase was followed by a mask of 50 ms and the reproduction task instruction. For the individual reproduction task, a picture of one of the previously scene target objects instructed the participants to recall the respective object position. In the ensemble reproduction task, a symbol presented at the beginning of the experiment, instructed participants to recall the average position around which all objects were arranged. During the final response phase, the empty kitchen scene was displayed, and participants clicked on the respective target position as accurately as possible, with no time limit. To prevent preemptive

positioning of the mouse, the mouse started in a random corner of the screen when the empty scene appeared. The next trial began after participant's response. The experiment was conducted using PsychoPy (v2021.2.0; Peirce et al., 2019).

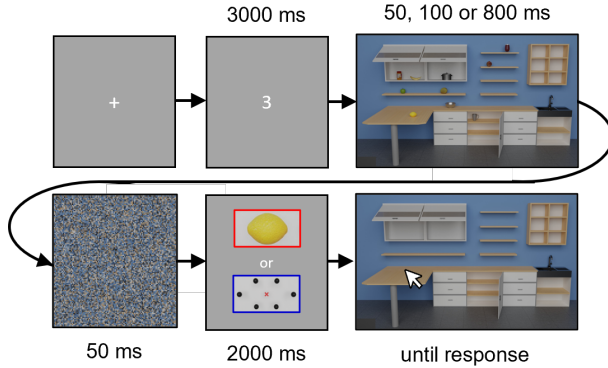


Figure 1: Example trial sequence of the behavioral experiment. Individual (red) and ensemble (blue) reproduction trials were blocked. Click endpoints were measured.

All combinations of reproduction task (individual, ensemble), set size (3, 6, 10) and encoding time (50, 100 and 800 ms) were tested for each of the six test scenes, resulting in 216 trials per participant. The trials were blocked, according to the reproduction task, creating a total of four blocks: two for individual reproduction and two for ensemble reproduction.

Preprocessing and Analysis

First, we analyzed behavioral, locating data to understand human performance in this task. These data were preprocessed using Jupyter Notebook (v6.2.0; <https://jupyter.org/>), where statistical analyses involved using repeated measures ANOVAs (RM-ANOVAs), implemented in jamovi (v2.3.28; <https://www.r-project.org/>).

To assess the accuracy of reproducing the respective target positions, the magnitude of the difference vector between the position where participants clicked relative to the actual (veridical) 2D object position in the scene was calculated. For individual perception, the 2D position of each object in each scene arrangement was defined relative to its center of mass, which was extracted from the respective Blender environment. To determine the reference position for the ensemble reproduction task, we based our assessment on participants' stated strategies after the experiment. Most reported using either the object group's centroid (arithmetic mean of target positions) or the bounding box center (midpoint of min and max coordinates). Since both yielded similar results, but the centroid aligned better with participants' ensemble perception, especially in the 10-object condition, we report deviations from the centroid.

Experimental trials were excluded based on the following criteria: (1) participants blinked during the Encoding phase of the 50 and 100 ms condition ($\approx 0.21\%$ affected), (2) participants

did not look at the computer screen for more than 30 % of the preview phase ($\approx 0.29\%$ affected) and (3) participants made high deviations from the target position (deviation > 3 standard deviations from the mean calculated separately for the reproduction task and encoding time conditions) during short reaction times (reaction time $< 1s$; $\approx 0.04\%$ affected). These outlier criteria were not mutually exclusive, and some trials met multiple criteria. This resulted in a total of 27 excluded trials, making it 0.43% of the 29 data sets.

Models

For modeling purposes, we have thus far focused on the encoding time of 800 ms, as it provides participants with the longest time, and therefore, probably the highest accuracy of encoding the scene in this experimental setup.

We created and explored two bayesian perception models for the behavioral data. The models differ in their assumptions about how individual and ensemble percepts are computed. The first model, here referred to as the 'Individual Encoding model' (IEM), assumes that the ensemble location is only evaluated on demand. During the encoding phase, individual target positions are encoded up to sensory noise \vec{V} . During recall of individual object locations, the encoded latent positions are reproduced. If presented with an ensemble query, this model summarizes the encoded information as the average location around which all target objects were arranged (Harrison et al., 2021; Robinson & Brady, 2023). In contrast, the second model posits a generative process for the visual scene that first chooses a (latent) ensemble location (Chong & Treisman, 2005; Ward et al., 2016), relative to which the objects are arranged (Lew & Vul, 2015). The sensory signals resulting from this process may also be corrupted by sensory noise. We refer to this process as the 'Ensemble Encoding model' (EEM). Percepts arise in both models through bayesian inference about the latent variables.

The coordinate system's origin was set to the center of the screen, as participants were instructed to start the exploration of the scene at this position during encoding. Owing to the inherent limitations in human vision, spatial encoding and retrieval accuracy, we incorporated zero-mean visual uncertainty with a standard deviation of 2° in both models, (Petrov et al., 2015). Our aim was to determine which of these models better explains the observed locating behavior in the two different reproduction tasks and three set size conditions, thereby gaining further insight into the possible computations underlying individual and ensemble perception.

Individual Encoding Model (IEM)

In the IEM the ensemble location \vec{E} is computed by averaging individual object representations \vec{X}_i (Harrison et al., 2021; Robinson & Brady, 2023). For K objects in a scene which are presented at locations \vec{O}_i , that differ from the internal representations \vec{X}_i by independently drawn visual noise/uncertainty \vec{V}_i , the model assumptions are

$$\vec{X}_i \sim \mathcal{N}(\vec{0}, \Sigma_X) \quad (1)$$

$$\vec{V}_i \sim \mathcal{N}(\vec{0}, 4 \cdot \mathbb{I}_2) \quad (2)$$

$$\vec{O}_i = \vec{X}_i + \vec{V}_i \quad (3)$$

$$\vec{E} = \frac{1}{K} \sum_{i=1}^K \vec{X}_i \quad (4)$$

$$\Sigma_X \sim \mathcal{W}(4, 12/\sqrt{4} \cdot \mathbb{I}_2) \quad (5)$$

where \mathcal{N} refers to a multivariate normal distribution and \mathbb{I}_2 is the 2×2 identity matrix. $\mathcal{W}(4, 12/\sqrt{4} \cdot \mathbb{I}_2)$ is a wide Wishart prior on the covariance matrix Σ_X with an expectation of 144 for the diagonal elements (parametrization via `scale_tril` in PyTorch), which reflects the experimental design: a standard deviation of 12° was used for object placement around the screen center, across all test scenes used in the behavioral experiment.

Ensemble Encoding Model (EEM)

The EEM describes a generative process for scenes that independently draws the ensemble position \vec{E} and the object positions \vec{X}_i relative to \vec{E} (Lew & Vul, 2015). Visual uncertainty is the same as for the IEM (Pertsov et al., 2015). The model is therefore specified as

$$\vec{E} \sim \mathcal{N}(\vec{0}, \Sigma_E) \quad (6)$$

$$\vec{X}_i \sim \mathcal{N}(\vec{0}, \Sigma_X) \quad (7)$$

$$\vec{V}_i \sim \mathcal{N}(\vec{0}, 4 \cdot \mathbb{I}_2) \quad (8)$$

$$\vec{O}_i = \vec{E} + \vec{X}_i + \vec{V}_i \quad (9)$$

$$\Sigma_X \sim \mathcal{W}(4, 9/\sqrt{4} \cdot \mathbb{I}_2) \quad (10)$$

$$\Sigma_E \sim \mathcal{W}(4, 9/\sqrt{4} \cdot \mathbb{I}_2) \quad (11)$$

The prior diagonal expected covariances of 81 reflect the experimental design here, too. Both the standard deviation of the average object position from the screen center, as well as the average deviation of the individual object from the average position was 9° .

In both models, the ensemble percept is given by the posterior distribution of (the latent) \vec{E} after encoding, i.e. we evaluate $P(\vec{E} | \vec{O}_{1,\dots,K}, \Sigma_X)$ for the IEM, and $P(\vec{E} | \vec{O}_{1,\dots,K}, \Sigma_X, \Sigma_E)$ for the EEM. Both posteriors can be computed analytically, details can be found in the model scripts, see link below.

Model Fitting

We fit the models to simulated data (for recovery tests) and real data (for model evaluation) by maximizing the posterior probability of the model's predictions and parameters with respect to Σ_X (and Σ_E). To this end, we implemented both models in Python (version 3.12; Foundation, 2023) using the machine learning framework PyTorch (version 2.5; Contributors, 2023) for automatic gradient computation and optimization. We performed 4000 steps with the Adam optimizer (Kingma & Ba, 2017), using a learning rate of 0.01. To compare the

models, we used a Laplace approximation to the model evidence (Bishop, 2006). All scripts are available here: *the link will be made available once the double-blind review process is over. Alternatively, we are happy to send an archive to the reviewers through the handling meta-reviewer.*

Position Reproduction Recovery

Our first test of the models was aimed at recovering the latent variables of a matching generator model with identical covariance parameters, i.e. the \vec{X}_i and \vec{E} for both models. We computed the expected position and standard deviation of both individual objects and the ensemble from the posterior distributions. The ratio of the position reproduction error (difference between actual and expected position) and the standard deviation should be ≈ 1 for ensemble position estimation. For the individual object position estimation, this ratio should be < 1 since these positions were used during encoding, thus the main source of posterior deviation stems from the (small) visual uncertainty. We computed the average ratio across 72 trials (reflecting the behavioral experimental design) and repeated this simulation 1000 times. The results are shown in Table 1.

Table 1: Mean ratio of the position reproduction error and the predicted object placement uncertainties for all combinations of reproduction task and set size for the Individual Encoding model (IEM) and the Ensemble Encoding model (EEM). All standard errors $< 10^{-2}$.

Set Size	IEM		EEM	
	Ind.	Ens.	Ind.	Ens.
3	0.03	1.00	0.04	1.00
6	0.03	1.00	0.04	1.00
10	0.03	1.00	0.04	1.00

All mean ratios are in the expected ranges, indicating that the models' latent variables behave as expected. An example of recovered object and ensemble positions is depicted in Figure 2. The notably small covariance ellipses of the ensemble position predicted by the IEM model are a consequence of Eqn. 4.

Parameter Recovery

The second test of our modeling pipeline targeted the models' parameters, i.e. Σ_X for both models and Σ_E for the EEM. We generated artificial datasets by varying the generating models' parameters in increments of 0.1, and trained a model of the same type as the generator on these data. Increments of 0.1 reflect behavior differences smaller than those resolvable by our experimental design, see Figure 3. Here, we show the recovered vs. generating diagonal elements of the covariance matrices, as well as the Pearson correlation coefficients (r) between these elements. While there is some compression especially in Σ_X , all model parameters can be recovered well enough to allow for meaningful statements about relative magnitude.

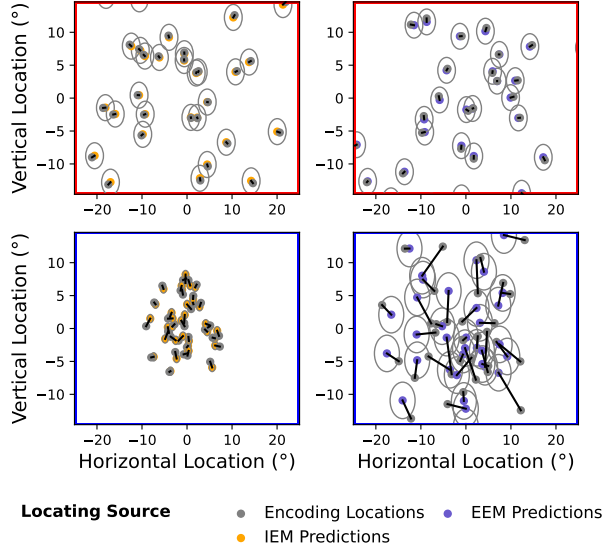


Figure 2: Examples of position reproduction recovery in the 10-object condition. Individual object trials are shown in the top row (red boxes) and ensemble estimates trials in the bottom row (blue boxes). The left column represents the Individual Encoding model (IEM, yellow dots), while the right columns corresponds to the Ensemble Encoding model (EEM, purple dots). Data were sampled from a matching model. The respective encoding locations are indicated as grey dots, while their placement variance are presented as grey circles.

Model Comparison and Recovery

To determine which model best describes the behavior of a given participant, we performed approximate bayesian model comparison with the aforementioned Laplace approximation to each model's evidence. To test if this approach recovers the correct model type with high certainty, we generated 100 datasets with 72 trials for 3, 6 and 10 objects from each model type and fitted both models to those datasets. Repeating the analysis with 1000 datasets yielded almost the same results. For the direct comparison of both models we then computed the log-bayes factor in favor of the EEM, i.e. positive values indicate evidence for EEM. The result is shown in Figure 4.

In virtually in all cases, the model comparison recovers the data generating model with near certainty. I.e. the absolute value of the log-bayes factor is well above $\log(100) \approx 4.6$, which is 'decisive' for the generating model (Kass & Raftery, 1995), as can be seen in Figure 4.

Results

Behavioral Experiment

To analyze the effects of set size on individual and ensemble reproduction deviation, we conducted two RM-ANOVAs, one for the individual and one for ensemble reproduction task. In this study, we report only the results from the 800 ms conditions to ensure comparability with the modeling approach. The

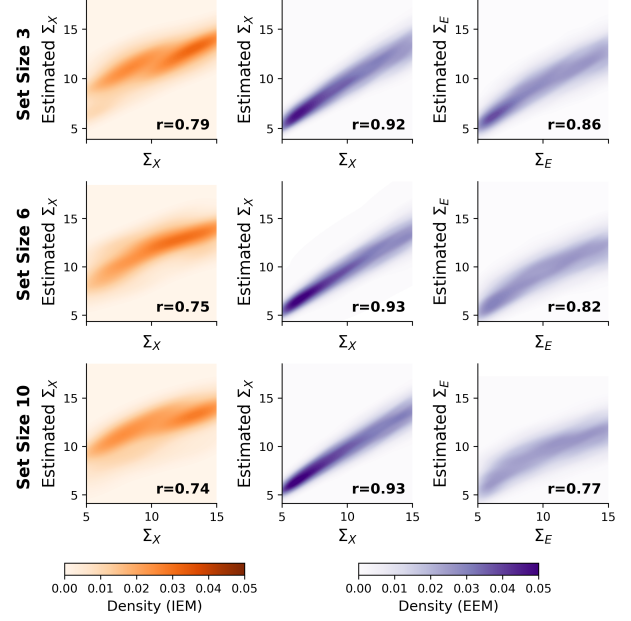


Figure 3: Parameter recovery of the IEM and the EEM for the 3, 6 and 10 object conditions of the parameters Σ_X and Σ_E . The strength of the recovered parameter relationships is indicated by Pearson correlation coefficients (r).

remaining analyses will be made available at a later stage.

In the individual reproduction condition, there was a significant main effect of set size ($F_{2,56} = 55.50$, $p < .001$, $\eta_p^2 = 0.67$). Post hoc t-tests, corrected using the Bonferroni-Holm method, indicated that reproduction error increased with each successive increase in set size (mean difference₃₋₆ = -2.59 , $t(28) = -5.65$, $p < .001$; mean difference₃₋₁₀ = -5.00 , $t(28) = -11.36$, $p < .001$; mean difference₆₋₁₀ = -2.41 , $t(28) = -4.62$, $p < .001$, all p-values are Bonferroni-Holm corrected), as shown in Figure 5.

The RM-ANOVA for the ensemble reproduction task showed a significant main effect of set size as well ($F_{2,56} = 5.04$, $p = .010$, $\eta_p^2 = 0.15$). Bonferroni-Holm corrected post hoc t-tests indicated that with our choice of significance level, we are able to reject the hypothesis that the average deviation in the 6 object condition is equal to the average deviation in the 10 object condition. Deviations were smaller on average when six objects were presented in the scene (mean difference₃₋₆ = 0.15 , $t(28) = 1.26$, $p = .219$; mean difference₃₋₁₀ = -0.29 , $t(28) = -2.09$, $p = .092$; mean difference₆₋₁₀ = -0.43 , $t(28) = -2.71$, $p = .034$). However, this does not imply that there are no differences between any other pairs of conditions.

Models

We illustrate the fit of the models to the behavioral data for an example trial in Figure 6. For individual object location reproduction, both models make predictions close enough to the participant's estimation, i.e. within their covariance ellipses

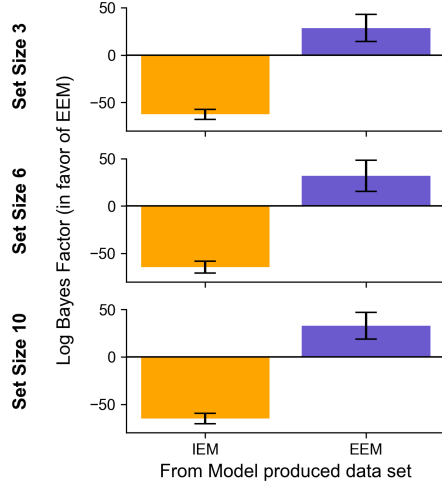


Figure 4: Model comparison test on model generated data between the IEM and EEM for the 3-, 6-, and 10-object conditions. The log-bayes factors of the comparisons for data generated by the IEM are shown in the left column, while those for data generated by the EEM are shown in the right column. Positive log-bayes factors favor the EEM, whereas negative values favor the IEM. Error bars are standard deviations.

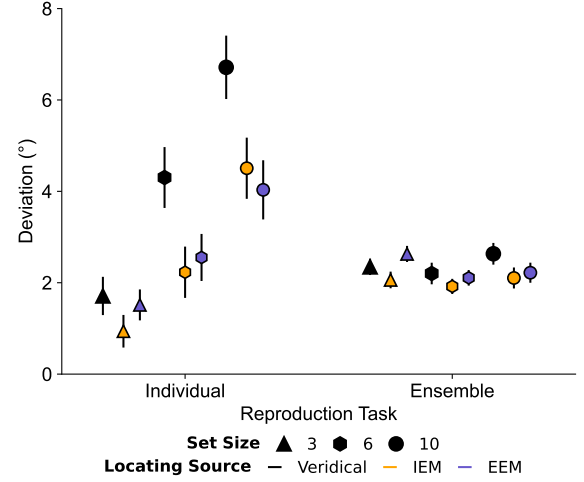


Figure 5: Mean reproduction errors (black) for individual and ensemble average reproduction task for all three set sizes with confidence intervals as error bars. Individual errors are relative to veridical positions, ensemble errors are relative to average veridical positions. Also shown are the median angular distances between model posterior means and participants reports for the IEM (orange) and EEM (purple), with confidence intervals as error bars.

(upper panel of Fig. 6). We also computed the average angular distances between model posterior means and participants reports (see Fig.5). The results indicate that the model predicts participants' reports for individual object locations better than the veridical position.

For ensemble perception, only the EEM covariance prediction contains the participant's response (lower panel of Fig. 6), the IEM is 'overconfident' as a consequence of Eqn. 4. However, as it can be seen in Figure 5, the reported ensemble locations are predicted at least as well by both models' ensemble mean as the average of the veridical object locations. This overconfidence of the IEM is probably due to the fact that the models currently account only for variance in encoding and do not incorporate other aspects of human behavior, such as memory-based uncertainty (Robinson & Brady, 2023).

In a next step, we calculated the log-bayes factors for each participant across and within each set size condition, as well as the mean log-bayes factor per set size condition across all participants (see Fig. 7). While the EEM shows a tendency to explain participants' behavior at smaller set sizes better, this advantage becomes especially pronounced in the 10-item set size condition. Although the IEM provides a better fit for some participants in the 3-item set size condition, this advantage diminishes with increasing set sizes, ultimately resulting in a clear superiority of the EEM over the IEM. The cause for this result might be the overconfidence of the IEM in estimating the ensemble position, which we illustrated in Figure 6, while the predicted ensemble means of EEM and IEM are almost identical, as illustrated in Figure 5. Hence, the EEM can assign a larger probability density to the participant's estimate

Set Size	BF_{ind}	BF_{ens}
3	-0.03 [-0.06, 0.06]	0.71 [0.1, 1.4]
6	0.02 [-0.14, 0.26]	1.29 [0.81, 3.4]
10	0.24 [0.11, 0.58]	8.0 [4.9, 10.9]

Table 2: Median and interquartile ranges of bayes factor contributions in favour of EEM, for individual BF_{ind} and ensemble BF_{ens} trials for different set sizes.

To investigate if this difference in posterior covariance holds across all participants' data and drives the model comparison results, we computed the across-participant median contribution of individual and ensemble trials to the bayes factors in favor of EEM, see table 2. Since the Bayes factors are mostly driven by the ensemble trials, we conclude that it is indeed the difference in predicted covariance that makes the EEM a better explanation for our data.

Finally, to assess the variability associated with the latent variables X and E in the two models, we calculated the average estimated Σ across and within each set size condition (see Fig. 8). In the EEM, the total variance (in the horizontal direction) of the object locations is divided between Σ_X and Σ_E , hence these (co)variances are smaller than Σ_X in the IEM. Furthermore, for three objects $\Sigma_X < \Sigma_E$ in the EEM. This results in an ensemble location that is 'pulled' towards the geometric center of the individual objects and away from the screen center, reflecting participants' behavior. In contrast, for ten objects, both Σ s can be very similar to achieve the same result, which is in line with our prior choice as described in the

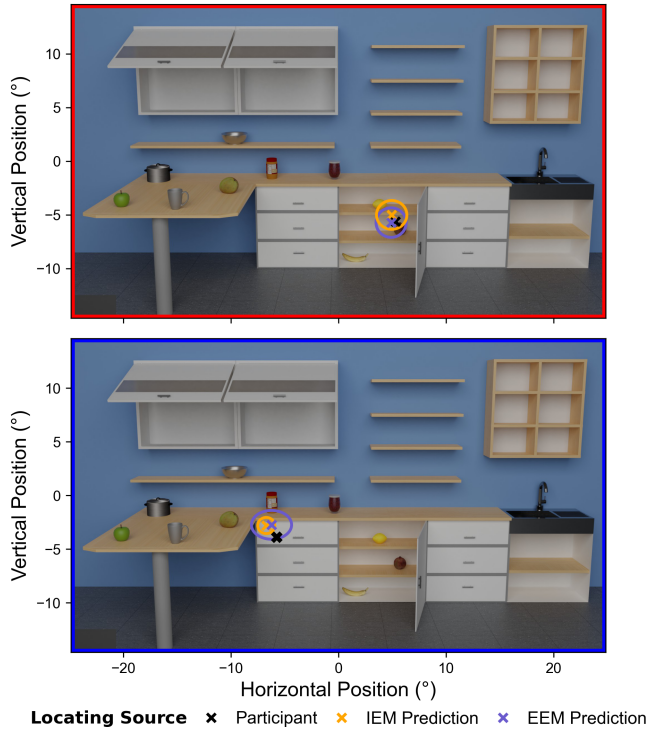


Figure 6: Example of a comparison of participants' locating behavior (black crosses) with predictions from the IEM (orange crosses) and EEM (purple crosses) in an individual reproduction trial (first row, target: pomegranate) and an ensemble reproduction trial (second row). Colored circles indicate the estimated $\sqrt{\Sigma}$ of each model for the target position.

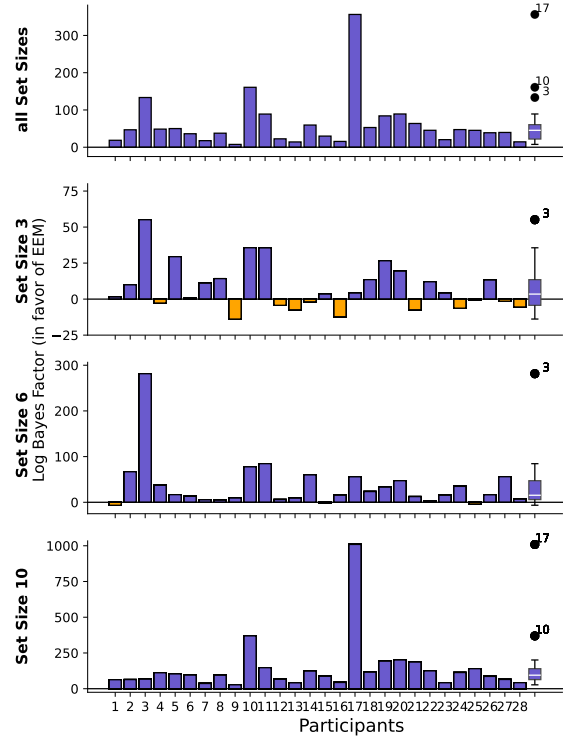


Figure 7: Log-bayes factor for each participant and the overall average for all set size conditions. The average log-bayes factor is represented as a box plot, illustrating the distribution, with confidence intervals as error bars and extreme values indicated by dots. Positive log-bayes factors favor the EEM, whereas negative values favor the IEM.

modeling section.

Discussion

Previous research has primarily examined individual and ensemble perception without accounting for the visual complexity of realistic scene contexts, which can either facilitate or hinder object perception (Draschkow et al., 2014; Ringer et al., 2021). To bridge this gap, we investigated how the perception of both individual objects and ensemble spatial information within naturalistic scenes is influenced by variations in set size. Our behavioral data suggest that, consistent with findings from studies using uniform backgrounds (Melcher et al., 2021; Neumann et al., 2018), individual object perception benefits more from smaller set sizes, whereas ensemble perception remains relatively unaffected by set size.

Furthermore, we compared two computational models — Individual Encoding model (IEM) and Ensemble Encoding model (EEM) — to address the ongoing debate about the exact processing of ensemble information (Harrison et al., 2021; Robinson & Brady, 2023; Yildirim et al., 2018; for a review see Corbett et al., 2023). The IEM suggests that ensembles are constructed only when needed, relying on the averaging of individually encoded object percepts. Meanwhile, the EEM

proposes that the ensemble's average position is directly processed and used to encode individual representations. Contrary to recent studies (Harrison et al., 2021; Robinson & Brady, 2023; Utochkin et al., 2024), the EEM demonstrated a better fit for the observed locating behavior, questioning the superior account of the IEM. Notably, the advantage of the EEM became more pronounced at higher set sizes. But the explanatory power of EEM is not limited to ensemble perception: EEM was shown to be a likely candidate for allocentric scene coding in reaching tasks, cf. (Khoozani et al., 2019).

Given three objects, the ability to form a perceptual group and facilitate ensemble perception remains debated (Whitney & Yamanashi Leib, 2018). While two objects suffice for ensemble perception (Whitney & Yamanashi Leib, 2018), a small set does not necessarily need to be processed as one. When objects are widely dispersed or divided into separate subgroups, their integration is weakened, contributing less to the overall percept (Klinghammer et al., 2017; Lew & Vul, 2015). In our 3-object set size condition, spatial dispersion across the scene likely reduced the advantage of extracting a shared position, making it harder to perceive them as a cohesive group. This is reflected in the log-bayes factor, which shows greater variability across participants in model fit. On average, model

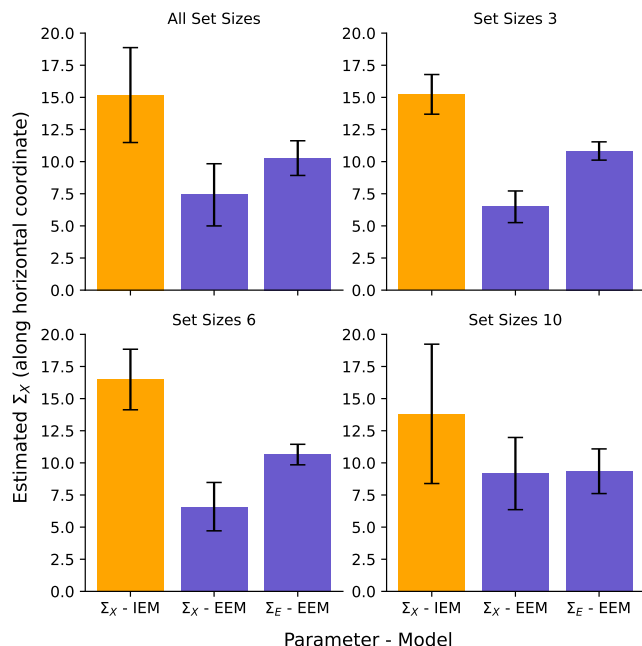


Figure 8: Estimated Σ along the horizontal coordinate of the IEM (orange) and EEM (purple) across all set sizes and for the 3, 6, and 10 object conditions, with standard deviations as error bars.

fit in this condition hovers just above zero (i.e. in favor of EEM), though some participants' data align more clearly with the IEM than others. Smaller set sizes (e.g., 3 objects) may be perceived as individual objects because the advantage of ensemble representation is minimal compared to simply encoding their positions. When more objects are present, the perceptual system may shift toward ensemble processing as a resource-efficient strategy, favoring summary representation over encoding each object separately. This aligns with theories of resource-limited rationality, which suggest that perceptual and cognitive systems optimize information processing by prioritizing strategies that reduce computational demands (cf. Binz et al., 2022). In contrast, when only three objects are present — especially if they are widely spaced — the benefit of ensemble encoding may be minimal, making an individual-object-based representation more viable.

It is important to note that our current models focus exclusively on the encoding of individual and ensemble information, accounting for uncertainty only at this initial stage of visual processing. However, human locating behavior is influenced by multiple sources of uncertainty beyond encoding, including memory-related degradation (Robinson & Brady, 2023; Pertzov et al., 2015) and motor errors (Hertzum & Hornbæk, 2010). These additional factors likely contribute to the differences observed between participant responses and model predictions, as seen in Figure 6. Further, our models in their current state assume ideal observer conditions, meaning they postulate that all objects can be accurately encoded. However,

in our experimental setup, the encoding duration was limited to 800 ms, which is particularly restrictive in the 10-object condition, where encoding every object individually would be highly demanding. Given that encoding time is a key factor in the debate on the underlying ensemble process (Melcher et al., 2021; Neumann et al., 2018), a natural next step is to integrate a time-dependent variable into our models. This addition would enable a systematic exploration of how encoding duration may influence this transition from individual to ensemble encoding strategies, helping to bridge the gap between behavioral findings and computational models.

Our findings also demonstrate that in complex, naturalistic scenes, set size effects differentially influence individual and ensemble perception. While both encoding strategies are viable at smaller set sizes, our computational modeling approach reveals that as set size increases, the EEM, which posits direct processing of ensemble averages, better explains human behavior. This suggests that ensemble perception becomes increasingly advantageous when encoding demands grow, suggesting that in rich and dynamic real-world environments, the perceptual system flexibly selects between individual and ensemble encoding based on efficiency constraints.

However, our results may be biased in favor of the EEM because participants knew when they would have to reproduce the ensemble position. An argument for why our findings might still hold if participants were not primed for the reproduction task is given by the Bayes factor contributions in table 2. The EEM explains the individual reproduction task just as well as the IEM. It is superior in explaining the ensemble reproduction data. Thus, independent of the task EEM provides the overall better explanation. Nevertheless, it would be interesting to investigate this explicitly in future experiments.

By bridging behavioral and computational perspectives, our study contributes to a more dynamic view of individual and ensemble perception in real-world contexts, by flexibly adapting the perceptual strategies to the task demands. This highlights the adaptive nature of how humans extract and utilize visual object information. Building models that fuse both modes of perception and can flexibly weight them in a continuous fashion, rather than contrasting them as we did here, will be interesting for future research.

References

- Alvarez, G. A., & Cavanagh, P. (2004). The capacity of visual short-term memory is set both by visual information load and by number of objects. *Psychological Science*, 15(2), 106–111. doi: 10.1111/j.0963-7214.2004.01502006.x
- Arieli, D. (2001). Seeing sets: Representation by statistical properties. *Psychological Science*, 12(2), 157–162. doi: 10.1111/1467-9280.00327
- Bar, M., & Aminoff, E. (2003). Cortical analysis of visual context. *Neuron*, 38(2), 347–358. doi: 10.1016/S0896-6273(03)00167-3

- Binz, M., Gershman, S. J., Schulz, E., & Endres, D. (2022).
Heuristics from bounded meta-learned inference. *Psychological Review*. doi: 10.1037/rev0000330
- Bishop, C. M. (2006). *Pattern recognition and machine learning*. Springer. Retrieved from /bib/bishop/Bishop2006/Pattern-Recognition-and-Machine-Learning-Christophe-M-Bishop.pdf, /bib/bishop/Bishop2006/978-0-387-31073-2_sm.pdf, https://www.microsoft.com/en-us/research/people/cmbishop/#!prml-book
- Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, 43(4), 393–404. doi: 10.1016/S0042-6989(02)00596-5
- Chong, S. C., & Treisman, A. (2005). Statistical processing Computing the average size in perceptual groups. *Vision Research*, 45(7), 891–900. doi: 10.1016/j.visres.2004.10.004
- Contributors, P. (2023). *Pytorch (version 2.5)*. Retrieved from https://pytorch.org/ (Machine learning library)
- Corbett, J. E., Utochkin, I., & Hochstein, S. (2023). *The pervasiveness of ensemble perception: Not just your average review*. Cambridge, UK: Cambridge University Press.
- Draschkow, D., Wolfe, J. M., & Vö, M. L.-H. (2014). Seek and you shall remember: Scene semantics interact with visual search to build better memories. *Journal of Vision*, 14(2), 10. doi: 10.1167/14.2.10
- Foundation, P. S. (2023). *Python (version 3.12)*. Retrieved from https://www.python.org/ (Programming language)
- Haberman, J., & Whitney, D. (2007). Rapid extraction of mean emotion and gender from sets of faces. *Current Biology*, 17(17), R751–R753. doi: 10.1016/j.cub.2007.06.039
- Harrison, W. J., McMaster, J. M., & Bays, P. M. (2021). Limited memory for ensemble statistics in visual change detection. *Cognition*, 214, 104763. doi: 10.1016/j.cognition.2021.104763
- Hertzum, M., & Hornbæk, K. (2010). How age affects pointing with mouse and touchpad: A comparison of young, adult, and elderly users. *International Journal of Human-Computer Interaction*, 26(7), 703–734. doi: 10.1080/10447318.2010.487198
- Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, 24(2), 175–219. doi: 10.1016/0010-0285(92)90007-O
- Kass, R. E., & Raftery, A. E. (1995). Bayes factors. *Journal of the American Statistical Association*, 90(430), 773–795. Retrieved from http://www.stat.washington.edu/people/raftery/Research/PDF/kass1995.pdf doi: 10.1080/01621459.1995.10476572
- Khoozani, P., Schrater, P., Endres, D., Fiehler, K., & Blohm, G. (2019). Models of allocentric coding for reaching in naturalistic visual scenes. In *Proceedings of ccn*. Retrieved from https://2019.ccneuro.org/proceedings/0000161.pdf
- Kingma, D. P., & Ba, J. (2017). *Adam: A method for stochastic optimization*. Retrieved from https://arxiv.org/abs/1412.6980
- Klinghammer, M., Blohm, G., & Fiehler, K. (2017). Scene configuration and object reliability affect the use of allocentric information for memory-guided reaching. *Frontiers in Neuroscience*, 11, 204. doi: 10.3389/fnins.2017.00204
- Lauer, T., Cornelissen, T. H. W., Draschkow, D., Willenbockel, V., & Vö, M. L.-H. (2018). The role of scene summary statistics in object recognition. *Scientific Reports*, 8(1), 14666. doi: 10.1038/s41598-018-32985-6
- Lew, T. F., & Vul, E. (2015). Ensemble clustering in visual working memory biases location memories and reduces the weber noise of relative positions. *Journal of Vision*, 15(4), 10. doi: 10.1167/15.4.10
- Melcher, D., Huber-Huber, C., & Wutz, A. (2021). Enumerating the forest before the trees: The time courses of estimation-based and individuation-based numerical processing. *Attention, Perception, & Psychophysics*, 83(3), 1215–1229. doi: 10.3758/s13414-020-02188-0
- Neumann, M. F., Ng, R., Rhodes, G., & Palermo, R. (2018). Ensemble coding of face identity is not independent of the coding of individual identity. *Quarterly Journal of Experimental Psychology*, 71(6), 1357–1366. doi: 10.1080/17470218.2017.1327988
- Oldfield, R. (1971). The assessment and analysis of handedness: The edinburgh inventory. *Neuropsychologia*, 9(1), 97–113. doi: 10.1016/0028-3932(71)90067-4
- Oriet, C., & Hozempa, K. (2016). Incidental statistical summary representation over time. *Journal of Vision*, 16(15), 3. doi: 10.1167/16.15.3
- Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., ... Lindeløv, J. (2019). Psychopy2: Experiments in behavior made easy. *Behavior Research Methods*, 51, 195–203. doi: 10.3758/s13428-018-01193-y
- Perry, C. J., & Fallah, M. (2014). Feature integration and object representations along the dorsal stream visual hierarchy. *Frontiers in Computational Neuroscience*, 8, 84. doi: 10.3389/fncom.2014.00084
- Pertsov, Y., Heider, M., Liang, Y., & Husain, M. (2015). Effects of healthy ageing on precision and binding of object location in visual short term memory. *Psychology and Aging*, 30(1), 26–35. doi: 10.1037/pag0000029
- Ringer, R. V., Coy, A. M., Larson, A. M., & Loschky, L. C. (2021). Investigating visual crowding of objects in complex real-world scenes. *i-Perception*, 12(1), 1–23. doi: 10.1177/20416695211006345
- Robinson, M. M., & Brady, T. F. (2023). A quantitative model of ensemble perception as summed activation in feature space. *Nature Human Behaviour*, 7, 1638–1651. doi: 10.1038/s41562-023-01634-0

- 691 Robitaille, N., & Harris, I. M. (2011). When more is less:
692 Extraction of summary statistics benefits from larger sets.
693 *Journal of Vision*, 11(12), 18. doi: 10.1167/11.12.18
- 694 Utochkin, I. S., Choi, J., & Chong, S. C. (2024). A population
695 response model of ensemble perception. *Psychological Re-*
696 *view*, 131, 36–57. doi: 10.1037/rev0000431
- 697 Ward, E. J., Bear, A., & Scholl, B. J. (2016). Can you per-
698 ceive ensembles without perceiving individuals? the role
699 of statistical perception in determining whether awareness
700 overflows access. *Cognition*, 152, 78–86. doi: 10.1016/
701 j.cognition.2016.03.010
- 702 Whiting, B. F., & Oriet, C. (2011). Rapid averaging? not so
703 fast! *Psychonomic Bulletin & Review*, 18(3), 484–489. doi:
704 10.3758/s13423-011-0079-8
- 705 Whitney, D., & Yamanashi Leib, A. (2018). Ensemble per-
706 ception. *Annual Review of Psychology*, 69, 105–129. doi:
707 10.1146/annurev-psych-010416-044232
- 708 Xu, Y., & Chun, M. M. (2009). Selecting and perceiving mul-
709 tiple visual objects. *Trends in Cognitive Sciences*, 13(4),
710 167–174. doi: 10.1016/j.tics.2009.01.008
- 711 Yildirim, I., Öğreden, O., & Boduroglu, A. (2018). Impact of
712 spatial grouping on mean size estimation. *Attention, Per-*
713 *ception, & Psychophysics*, 80(7), 1847–1862. doi: 10.3758/
714 s13414-018-1570-0
- 715 Šetić, M., Švegar, D., & Domijan, D. (2007). Modelling the sta-
716 tistical processing of visual information. *Neurocomputing*,
717 70(10), 1808–1812. doi: 10.1016/j.neucom.2006.11.017