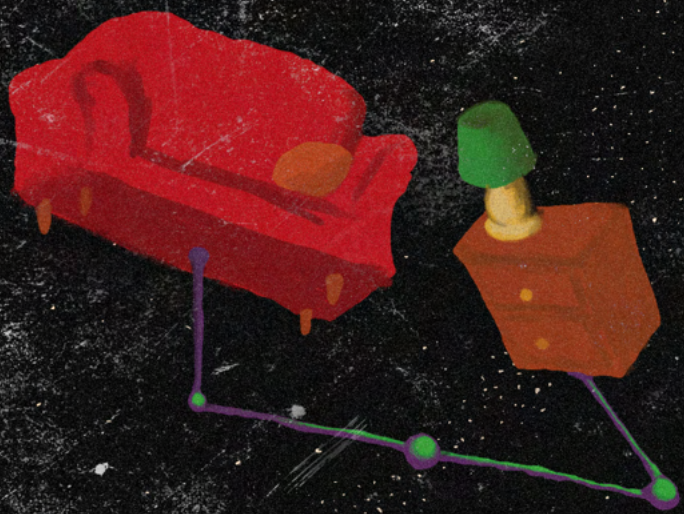
One Little World

35¢

Investigating context-driven object transformations in brain and behavior



DONDERS SERIES

Giacomo Aldegheri

One little world

Investigating context-driven object transformations in brain and behavior

Giacomo Aldegheri

Colofon

Printing:	Gildeprint www.gildeprint.nl
Layout:	Giacomo Aldegheri
Cover illustration:	Giacomo Aldegheri
Cover layout:	Giacomo Aldegheri, Qiu Han
Illustration on page 5:	from René Magritte, *Les Mots et les Images*

The work described in this thesis was carried out at the Radboud University, Donders Institute for Brain, Cognition and Behaviour, Nijmegen, The Netherlands.

One little world

Investigating context-driven object transformations in brain and behavior

Proefschrift ter verkrijging van de graad van doctor
aan de Radboud Universiteit Nijmegen
op gezag van de rector magnificus prof. dr. J.H.J.M. van Krieken,
volgens besluit van het college voor promoties
in het openbaar te verdedigen op

woensdag 13 september 2023
om 16.30 uur precies

door

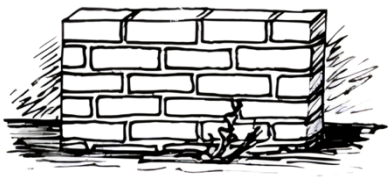
Giacomo Aldegheri
geboren op 21 maart 1992
te Venetië, Italië

Promotor:
Prof. dr. W.P. Medendorp

Copromotoren:
Dr. M.V. Peelen
Dr. S. Gayet

Manuscriptcommissie:
Prof. dr. F.P. de Lange
Prof. dr. C.M. Press (Birkbeck, University of London, Verenigd Koninkrijk)
Dr. I.I.A. Groen (Universiteit van Amsterdam)

Un objet fait supposer qu'il y en a d'autres derrière lui :



This thesis is dedicated to the memory of
Roger Shepard (1929-2022)

Tra le molte virtù di Chuang-Tzu c'era l'abilità nel disegno. Il re gli chiese il disegno d'un granchio, Chuang-Tzu disse che aveva bisogno di cinque anni di tempo e d'una villa con dodici servitori. Dopo cinque anni il disegno non era ancora cominciato. "Ho bisogno di altri cinque anni" disse Chuang-Tzu. Il re glieli accordò. Allo scadere dei dieci anni Chuang-Tzu prese il pennello e in un istante, con un solo gesto, disegnò un granchio, il più perfetto granchio che si fosse mai visto.

Italo Calvino, *Lezioni Americane, Cap. 2 - Rapidità*

Table of Contents

Chapter 1
General Introduction

1.1 Scope of this thesis

This thesis investigates how our internal representations of objects are driven by contextual information. In particular, we investigate the kind of contextual information we regularly encounter in the real world: naturalistic scenes. We explore how scene context can drive the prediction of object transformations: namely, how an object will look like from a new viewpoint. To do this, we use a novel experimental paradigm and explore behavioral and fMRI effects. We build upon an ample literature showing that expectations interact with visual processing through top-down connections to visual cortex (e.g. De Lange et al., 2018; Rao & Ballard, 1999), leading to behavioral changes in perceptual tasks (e.g. Summerfield & De Lange, 2014; Wyart et al., 2012). We find that object expectations driven by scene context share behavioral and neural hallmarks with other previously reported expectation-related effects, suggesting that common mechanisms might underlie these naturalistic expectations and more controlled ones found using synthetic stimuli. More specifically, the contributions of the experimental chapters are as follows:

In **Chapter 2**, we investigate scene-driven representations of objects from novel viewpoints by examining their effect on a perceptual task. We use fast presentation times, an orthogonal task and no explicit instructions in order to determine whether the object representation is updated automatically, as a result of the context rather than of a voluntary prediction. We also show the object from a variety of possible viewpoints and amounts of rotation, to determine the flexibility of the representation's transformation. We find that participants' performance is affected by whether an object is rotated congruently with the surrounding scene or not, suggesting that they flexibly and automatically update the object's representation together with the scene. Moreover, we find that this behavioral effect is not affected by the probability of the object respecting the scene constraint during the experiment, indicating that it results from real-world expectations rather than contingencies observed during the experiment.

In **Chapter 3**, we use fMRI to investigate the role of visual cortex in representing the updated object's view. We use multivariate pattern analysis (MVPA) to decode, from activity patterns in visual cortex, the object's proximal shape (whether it projects a wide or narrow retinal image), after the viewpoint change. We find that whether the object shape shown on the screen matches the scene-driven expectation or not is reflected in a change in decoding performance in early visual cortex, but not in higher-level object selective cortex. These results suggest that the viewpoint update ultimately results in a representation of the object's retinal shape, and that this is reflected in early visual cortex responses, consistently with ideas about early visual areas being recruited in computations that require retinotopically organized representations (Roelfsema & de Lange, 2016). Moreover, we find evidence that these scene-driven expectations share

neural mechanisms with other, previously reported expectation effects in visual cortex (e.g. Kok et al., 2012; Yon et al., 2018): an increase in multivariate decodability for expected object shapes, accompanied by an overall lower univariate response. This provides further indication that the updating of the object representation occurs automatically, and a stimulus that matches the correct updated view is perceived as the 'default', while the expectation's violation constitutes a disruptive event.

In **Chapter 4**, while the previous chapters always investigated the scene-driven expectation indirectly, by examining the effects of its violations, we directly investigate the expectation signal itself. Again using fMRI and MVPA, we design an experiment in which the rotated object is not shown after the occlusion on a majority of trials. In this way, we are able to decode a purely internally generated expectation of the updated object shape. We find that this can be successfully decoded in visual cortex, suggesting that scene-driven expectations can evoke a sensory-like signal in the absence of any visual input, again consistently with work investigating expectations in simpler domains (Kok et al., 2014, 2017).

In **Chapter 5**, as the previous studies were focused on the case of rotation out of the picture plane, we investigate a different spatial transformation. We adapt the paradigm of **Chapter 2** to study the case of object scaling, resulting from a translation in depth. We largely replicate the behavioral effects found in **Chapter 2**, suggesting that the interactions between scene and object representations we report are not idiosyncratic to rotation, and might be a general process in predicting object transformations. These results also provide a link to existing work investigating the 'rescaling' of object representations driven by scene context (Gayet & Peelen, 2022; Murray et al., 2006; Yildiz et al., 2021), suggesting a common process.

In the remainder of this Introduction, I will outline a theoretical proposal starting from the broad question of 'what constitutes a good perceptual representation'. This will largely be built around the idea that internal representations of objects should behave like those objects, making it possible to predict the outcome of transformations such as rotations or translations. There will be some notable absences in the topics I will treat. For one, the focus will be primarily on the psychological, computational and theoretical literature, as my interest here is to outline a series of desiderata for mental representations, regardless of their specific neural implementation. I will mention neuroscientific data wherever I deem it relevant, mainly for the purpose of exemplifying that some of the desiderata for representations that I list can indeed be found in the brain. Another major absence, for an Introduction addressing the topic of how we predict the behavior of external objects, is the predictive coding framework (Rao & Ballard, 1999; Bastos et al., 2012; Clark, 2015). This is a deliberate choice, in part for reasons of space: the literature on that framework is vast, and it would have been

impossible to properly cover it here. I have thus preferred to focus on literature that offers different angles on the problem of how we internally represent the world, angles that are far less explored. Another reason for this omission is the fact that much of the focus of research on predictive coding seems orthogonal to the goals of the investigation presented here. Predictive coding is largely concerned with the mechanisms of interaction between top-down expectations and bottom-up visual input (De Lange et al., 2018; Press et al., 2020). Here, on the other hand, the main focus is on the structure of internal representations themselves, rather than how they lead to changes in perceptual processing. I will primarily be interested in the ways in which internal representations behave like the things they represent: something that, to the best of my knowledge, is not a major concern of predictive coding research. I will indeed address the topic of interaction between internal representations and external stimuli, but with a different focus: I will be concerned with how this interaction reflects how external objects interact, similarly to individual representations reflecting the behavior of the corresponding objects. In my treatment, internal representations and external stimuli will be maintained as separate entities, rather than treated as two sources of evidence to be integrated. Another main focus of predictive coding research has been to develop formalisms for the top-down/bottom-up interaction in probabilistic terms (e.g. Friston 2010). I see this as a possible specific implementation of some of the ideas I discuss here, one that makes particular assumptions about the probabilistic nature of representations, and their neurophysiological implementation. Here, I will remain agnostic both on their probabilistic interpretation, and their realization at the neural level.

In the chapters, on the other hand, as I will investigate the neural bases of these representations (as summarized above), I will also address the issue of how top-down predictions interact with visual input more in depth, although without committing too strongly on any specific mechanistic or theoretical framework. The content of this Introduction, on the other hand, is mainly meant to outline a broad picture around the specific studies we have conducted, and why I hope they will be an initial step towards clarifying what ‘good representations’ of the world should look like. With these clarifications out of the way, let us begin our journey.

1.2 The little world in our head

The ordinary man asserts that he sees an external world containing various objects. It is only the philosopher who insists that he is conscious of sense-data, brown patches, and so forth.

- Kenneth Craik[1]

Do we have a little world in our head? Subjectively, we have the impression of seeing objects, people, events and places, even though of course, all that reaches our retina are rays of light. Does this mean that the world is 'reconstructed' in our brain? What does this internal world look like, and what is its relationship with the external one? If it is inside our head, how can it look 'out there'? I have no pretense to answer any of these long-standing questions within the limited scope of a thesis. Instead, I would like to focus on a specific subset of them, and in this Introduction, try to formulate them more precisely, along the way reviewing work from several disparate domains that can prove useful in understanding them. The main arguments I will try to make are summarized as follows:

a. In order for us to successfully behave in the world, our internal representations should mirror entities in the world in particular ways. I will review several related notions of what this should mean precisely: disentangled representations, analog representations, dynamic representations and linear transformations.

b. Internally generated representations and external stimuli should closely interact in order to solve several important tasks in the real world. I will review empirical evidence of this interaction and connect this idea with the concept of an *object file*.

c. In order for the interaction between internally generated representations and the external world to be successful, there need to be rules specifying how representations interact. These rules should mirror the interactions between entities in the external world, similarly to the way individual representations mirror the behavior of individual entities.

d. One particularly ubiquitous example are spatial relations (relative positions and orientations), and particularly hierarchical *part-whole* relations. I will describe an efficient way of representing these relations - the *scene graph* - and review some of the ways in which it has been implemented in artificial intelligence (AI) research.

e. Finally, I will bring these ideas together, suggesting that the task of tracking dynamic changes in structured scenes provides a motivation for:

[1] Craik (1943), p. 25

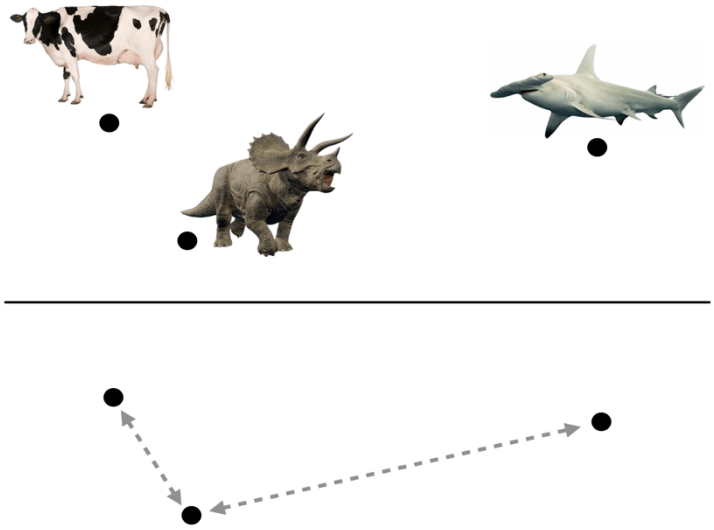


Figure 1.1. Illustration of second-order isomorphisms. The internal representations (bottom) of stimuli (top) should preserve the distances between them. After Edelman (1998).

representations that mirror objects' behavior; representations that interact with the current visual scene; representations of relations; and hierarchical spatial relations in particular.

After exposing these arguments, I will describe a simple 'model system': an experimental paradigm designed to explore some of the questions I describe. In the following Chapters, I will describe several empirical results obtained using this paradigm. My hope is that the empirical work and theoretical ideas presented here, although limited in scope, will suggest fruitful avenues for future research, as well as a new outlook on existing research.

1.3 What is a 'good' representation?

1.3.1 Representation of similarities

Our mental states can represent things in the world. I will take this here as an empirical fact rather than a thesis to be defended (for similar arguments, see John, 2021; Poldrack, 2021; Thomson & Piccinini, 2018), starting from the simple

observation that the 'units' of our perception and thought are things in the world. As I look around my kitchen, I see a table, a cupboard, and an oven. In guiding actions, I similarly refer to and target things at this scale (I reach the cupboard to get the plate). I will leave aside here the thorny questions of whether these representations exist as cohesive entities in the brain (Piccinini, 2008; Hipólito, 2022) and how they can be inferred from recordings of neural activity (Baker et al., 2022; Burnston, 2021; Ritchie et al., 2019). Assuming that, at some level of organization, mental representations exist, I will focus on perceptual representations here, and ask the following question: what constitutes a 'good' perceptual representation? As has been previously pointed out (Shepard & Chipman, 1970; Edelman, 1998) it doesn't make much sense to think that anything about a brain state that corresponds to a square object is actually square, or that a brain state representing an object that smells bad actually smells bad itself. First-order similarity (isomorphism) of an internal state with an external entity, then, seems like a hopeless criterion for a successful representation. A more promising alternative is that our representations should respect a second-order isomorphism (Figure 1.1): put simply, similarities among internal representations should reflect similarities among external entities (Shepard & Chipman, 1970). For example, the representation of a square should "have a closer functional relation to the internal representation for a rectangle than to that, say, for a green flash or the taste of a persimmon" (Shepard & Chipman, 1970, p. 2). As highlighted by Edelman (1998), this idea can account for empirical findings showing that subjects can be remarkably sensitive when comparing different stimuli (Cortese & Dyre, 1996; Cutzu & Edelman, 1996), while being poor at judging the properties of a single stimulus (Koenderink et al., 1996; Phillips & Todd, 1996; Todd & Norman, 2003). The idea that inter-stimulus similarity is central to understanding internal representations also underlies representational similarity analysis (Kriegeskorte et al., 2008), a common technique in cognitive and systems neuroscience, consisting of comparing conditions, measurement techniques or computational models by comparing the distances between stimulus representations (such as brain activation patterns), rather than the representations themselves.

Let's take this idea at face value: suppose that what representations should preserve are the distances between different entities in the world. That's great, but we are still missing a crucial ingredient: to know what distance means, we need to define an appropriate space. What is a useful notion of distance to represent objects?

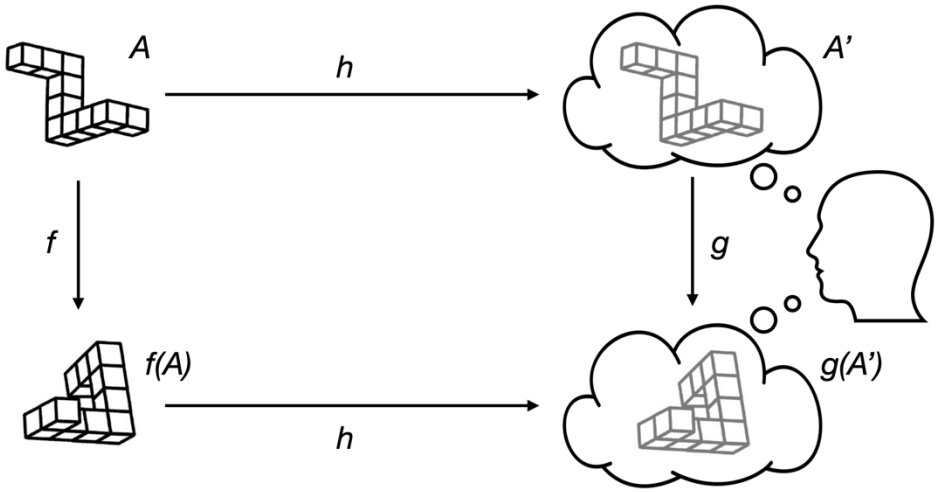


Figure 1.2. Equivariant representations: *h* is an 'encoding' function that maps stimuli to their internal representations. In an equivariant representation, transformations of external stimuli should be preserved by *h*, such that an internal operation, *g*, can be used to transform an internal representation similarly to how *f* transforms the corresponding external object. In this case, the transformation is a rotation.

1.3.2 Invariance vs. equivariance

One possibility is to exclusively represent the distances between object categories. In such a representational space, called an *invariant* representation, different views of a car would all be collapsed onto the same point: all that matters is that a car is a car. The notion that the goal of perception is to extract invariant representations, that retain information about object category while discarding all 'nuisance' information, has been the basis for much work in computer vision (e.g. Lowe, 1999; Kadir et al., 2004; Gray & Tao, 2008; Wu et al., 2008; Krizhevsky et al., 2012; Achille & Soatto, 2018) and computational neuroscience (Fukushima, 1980; Tanaka, 1996; Poggio & Bizzi, 2004; Deco & Rolls, 2004; Serre et al., 2007). Findings of neurons at the higher levels of visual cortex that respond to object identity independently of location, orientation or size (e.g. Ito et al., 1995; Tovee et al., 1994; Fujita, 2002; Booth & Rolls, 1998; Lueschow et al., 1994; Li & DiCarlo, 2008, 2010) were classically interpreted as evidence for invariant representations being the ultimate outcome of visual processing. While this kind of representation would be ideal for recognizing objects (although some might disagree: DiCarlo & Cox, 2007; Hong et al., 2016; Patel et al., 2015), it would fail at most other visual tasks in the real world. Object transformations, such as viewpoint changes, are behaviorally relevant: if I encountered a tiger in the wild, I would want to know

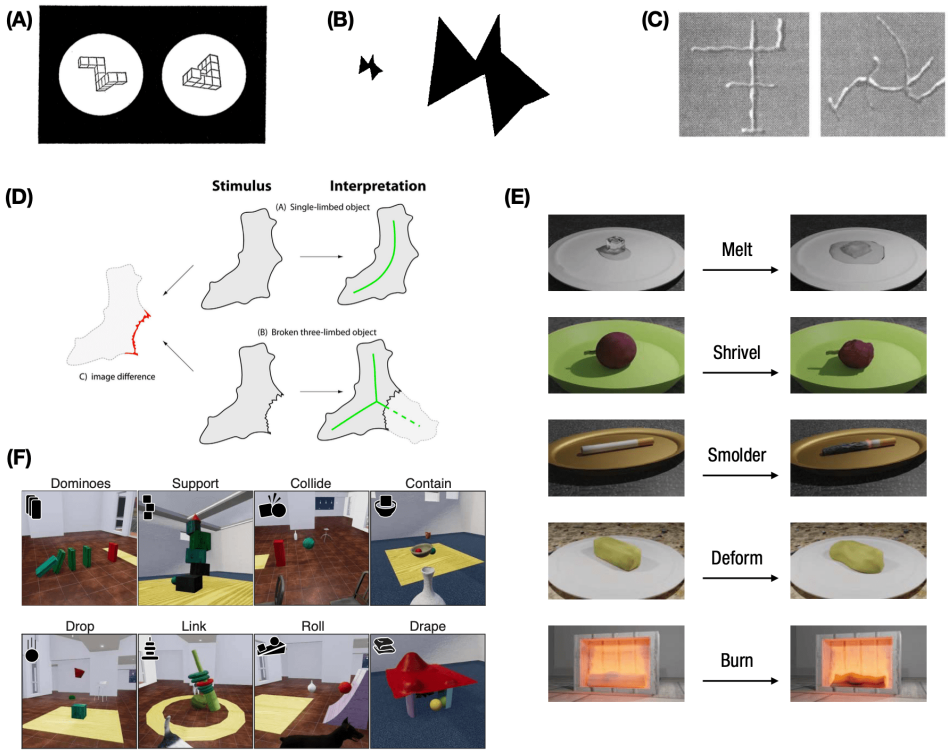


Figure 1.3. Examples of the transformations that can be predicted by human participants. Please note that these have been investigated through very different paradigms, refer to the original papers for more details. **(A)** Mental rotation (Shepard & Metzler, 1971); **(B)** Mental scaling (Bundesen & Larsen, 1975); **(C)** Simultaneous rotation and deformation of a flexible object (Kourtzi & Shiffrar, 2001); **(D)** Object breaking: by adding tear marks, an object can be perceived as a broken part of a larger object (Spröte et al., 2016); **(E)** Changes in physical state: humans tend to extrapolate these transformations forward in time (Hafri et al. 2022); **(F)** The various scenarios used by Bear et al. (2021) to test humans' and computational models' physical prediction abilities.

whether I was facing its front or its back, and it would be useful to predict how long it would take for it to turn around and see me. Our internal representations, then, should explicitly include information about these transformations of objects (e.g. their orientation, position, size, state - is the tiger awake or asleep?). Moreover, these transformations should *generalize* to new, unseen objects. If I encounter a predator that I have never encountered before, say a lion, I should still be able to predict how it will look from different viewpoints. This could be achieved through *equivariant* representations (Bengio et al., 2013; Hinton et al., 2012). The

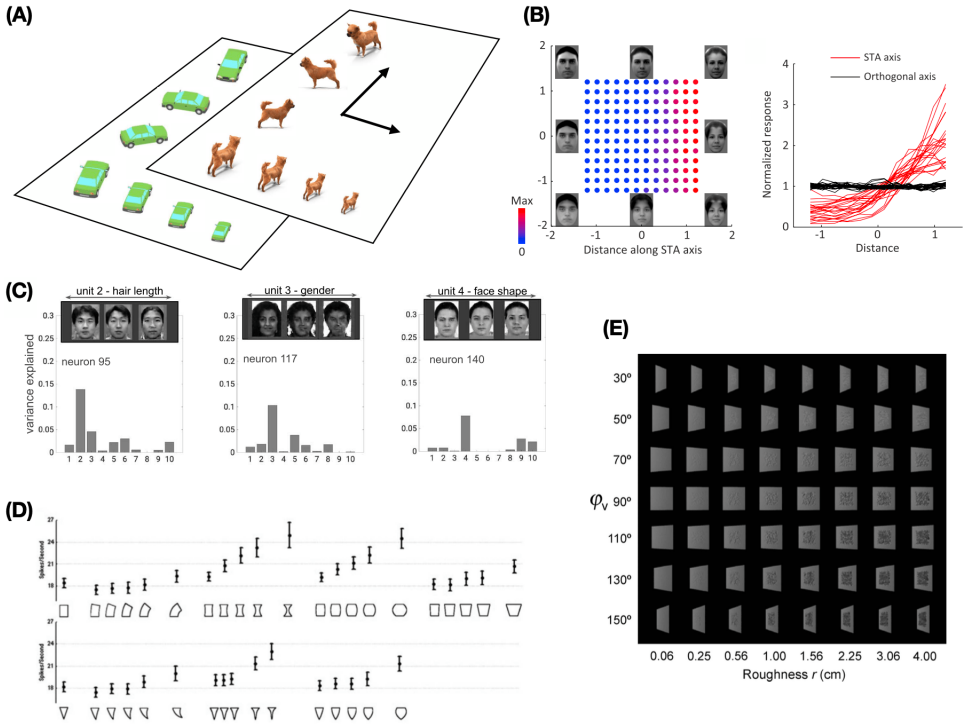


Figure 1.4. Disentangled representations. **(A)** Illustration of a disentangled representation: individual axes of the representational space correspond to meaningful factors of variations in the data. In this case, size and orientation, with category (dog, car) being an orthogonal dimension. After Higgins et al. (2022). **(B)** Disentangled representations in macaque face-selective cortex (Chang & Tsao 2017). Faces were decomposed into their principal components (left), and single neurons were found to respond according to a face's position along one specific component, while being completely unresponsive to the component orthogonal to it (right). **(C)** Single neurons responding to meaningful factors of variations of faces (Higgins et al., 2021). **(D)** Neurons in macaque IT tuned to specific shape dimensions (Kayaert et al., 2005). **(E)** An example of a failure to disentangle in human vision (Ho et al., 2007): the roughness of a surface is perceived differently depending on its viewpoint, meaning they are not represented independently.

idea of equivariant representations is illustrated in **Figure 1.2**: essentially, our internal representations of objects should transform similarly to those objects. More formally, if h is the function that maps an external object A to its corresponding internal representation, $A' = h(A)$, and f is the rotation of the object to a different viewpoint, it should be possible to apply a transformation g to the

internal representation *A'* such that *g(A') = h(f(A))*. In other words, after applying the operation *g* to the internal representation, we obtain the same representation that we would have obtained by observing the object from the new viewpoint. In this example, the internal operation *g* would be a *mental rotation* (Shepard & Metzler, 1971). The idea that our internal representations of the world behave like the external objects they represent (i.e. they can be transformed in similar ways) has a long history in cognitive science (e.g. Shepard, 1984, 2001), and is supported by a vast number of empirical findings. Without the pretense of being exhaustive, humans have been found to be able to infer and predict a variety of object transformations (**Figure 1.3**). Beyond the most notorious example of mental rotation (Shepard & Metzler, 1971; Cooper & Shepard, 1973; Shepard & Cooper, 1982), we are able to represent other *rigid* transformations (that leave object shape unaltered), such as translation and scaling (Bennett, 2002; Bundesen et al., 1983; Bundesen & Larsen, 1975; Larsen & Bundesen, 1978, 1998; Schmidt et al., 2016; Sekuler & Nash, 1972). *Non-rigid* transformations (which alter object shape) can also be represented, such as deformations (Kourtzi & Shiffrar, 2001; Hahn et al., 2009; Spröte & Fleming, 2016; Schmidt et al., 2019), tearing (Chen & Scholl, 2016; Spröte et al., 2016), or changes in physical state, like burning or melting (Hafri et al., 2022). Moreover, humans are able to predict how the physical dynamics of a scene will unfold (Battaglia et al., 2013; Bates et al., 2019; Ullman et al., 2017, 2018) in a way that surpasses current artificial systems (Bear et al., 2021). Together, these findings suggest that internal object representations can closely mirror the behavior of external objects, undergoing a wide variety of transformations. We will return to some of these 'mental transformations' later.

1.3.3 Disentangled representations

But first, what would a representation with these characteristics look like in a high-dimensional space, such as a population of neurons? A common way to think about this problem is the notion of *disentanglement* (Bengio et al., 2013; Schmidhuber, 1992; Higgins et al., 2018, 2022), which can be defined as follows: a given object transformation should only affect a subset of the dimensions in the representational space. This idea reflects the simple fact that any single transformation of an object leaves many of its features unaltered. For example, changing an object's orientation does not alter its color or shape. **Figure 1.4A** illustrates two dimensions of variation in a higher-dimensional representational space: one axis corresponds to the object's orientation, while a different axis corresponds to its size. Contrary to the *invariant* representations we encountered earlier, different views of an object are not collapsed onto a single point. Instead, they can be retrieved by following a particular direction in the representational space. In the most intuitive case, these directions correspond to single axes. That would mean that the dimensions of the representational space reflect meaningful dimensions of variation: for example, one dimension corresponds to the object's orientation, another to its size. Intriguingly, neuroscientific evidence has found

evidence for precisely this kind of code in the neural representation of high-level object features. For example, Chang & Tsao (2017) found individual neurons in macaque 'face patch' (part of high-level inferotemporal cortex, IT) tuned to single factors of variation in a low-dimensional embedding of face stimuli (**Figure 1.4B**). Higgins et al. (2021) later found that these single neurons in macaque IT had a one-to-one correspondence with the units of an artificial neural network trained to extract disentangled face representations. These single units corresponded to meaningful factors of variation, such as a face's hair length or hair style (**Figure 1.4C**). Similar results have been obtained with fMRI measurements in humans (Soulos & Isik, 2020) and with different classes of stimuli, such as shapes (Op de Beeck et al., 2001; Kayaert et al., 2005, **Figure 1.4D**). For a more exhaustive review, see Higgins et al. (2022). Behavioral findings have shown that participants independently encode particular features of 2D (Arguin & Saumier, 2000) and 3D shapes (Stankiewicz, 2002), such as aspect ratio and curvature. Other features, such as viewpoint and surface 'roughness', seem instead to be entangled in behavioral judgments (Ho et al., 2007; **Figure 1.4E**). Moreover, training participants to arrange stimuli in a disentangled space can improve their ability to flexibly learn different tasks on them (Flesch et al., 2018). Finally, one additional source of evidence for the plausibility of disentangled representations is the recent finding that they can emerge in a computational model with minimal biological constraints (Whittington et al., 2022). In summary, the intuitive idea that our internal representations of objects should mirror their transformations has found a fruitful formalization in the notion of *disentanglement*. In a disentangled representation, different meaningful object transformations are represented separately from each other, reflecting their separation in the external world. Several empirical findings show that disentangled representations closely match biological brains' strategy for representing the world, at least in some domains.

1.3.4 Linear transformations

An idea closely related to disentangled representations is that of *linear transformations* in the representational space. According to this idea, meaningful object transformations, such as rotations, which cause highly nonlinear changes in the space of images, should become linear in representational space. This is based on the notion that representing a given property of the world should entail making it explicitly available to simple readout processes (Fekete, 2010). Trivially, any property of a visual scene can be said to be represented in primary visual cortex (V1), since visual input needs to go through this area before reaching the rest of the brain. However, we do not think this is the case because that information is not in an *explicit* format. The fact that some property should be extractible via a linear readout is a reasonable criterion for qualifying as an explicit representation, a criterion that underlies analyses based on the linear decoding of brain activation patterns (Kriegeskorte & Diedrichsen, 2019). Similarly, for a particular transformation to be predictable in the representational space, it should

correspond to a linear transformation in that space.

In machine learning, several approaches based on explicitly enforcing that learned representations should transform linearly have proved successful at parsing the structure of simple scenes (e.g. Paccanaro & Hinton, 2001; Hinton et al., 2011; Cohen & Welling, 2015; Goroshin et al., 2015) and solving reinforcement learning tasks (Saanum & Schulz, 2022). In computational neuroscience, a model that learned to represent spatial locations, with the constraint that its representations should transform linearly, was shown to spontaneously give rise to a representation resembling grid cells in medial entorhinal cortex (Dorrell et al., 2022). This suggests that certain representations found in the brain might result from making useful transformations (such as shifts of one's position in space) explicitly available as linear transformations. Relatedly, work in neuroimaging (Ward et al., 2018) has shown that object transformations, such as changes in size or location, correspond to linear transformations of the representations in higher-level visual cortex. These transformations were shown to generalize across different objects, providing further evidence that they are explicitly disentangled from object identity. Interestingly, not only transformations that are actually linear in the real world, such as translation and scaling, but also nonlinear ones, such as image blurring, were shown to be similarly 'linearized' throughout visual cortex (Mocz et al., 2021). It is possible, then, that the purpose of this representational format is not to faithfully reproduce the true generating factors of the scene, but to render any potentially useful transformation explicitly available for further computations. Finally, a recent study (Hénaff et al., 2019) found that naturalistic video sequences, corresponding to highly curved trajectories in pixel space, were 'straightened' in participants' representational space, as inferred from their psychophysical discrimination performance. Linearizing, or straightening, meaningful transformations of a scene, then, might be a strategy employed by the brain to render these transformations explicit in its representational space.

In the last sections, we have described several criteria for a 'good' representation, that all boil down to the idea that internal representations of objects should 'behave like' those objects in some meaningful way. Importantly, however, our argument does not hinge on any of these criteria in particular. For example, the requirement that each different transformation should act on a subset of the representation's dimensions exclusively (the formal definition of disentanglement mentioned above) might not be necessary, and might actually lead to failure to represent certain transformations (Bouchacourt et al., 2021). But for our purposes, all that matters is to convince ourselves that 'good' perceptual representations should *in some useful way* transform like the objects they aim to represent. In the next section, we turn to an idea which has been developed separately from that of disentangled representations, but which shares several important aspects with it. It is the idea of *analog representations*.

1.3.5 Analog representations

Several definitions of what an analog representation is (as distinct from a *digital* one) have been proposed, starting from the seminal work of Goodman (1976). For the sake of brevity, we will only discuss a specific definition here, the one proposed by Maley (2011). Maley (2011) separates the distinction between analog and digital representations from that between continuous and discrete, which it had previously been equated with. He borrows a definition used in cognitive psychology (Shepard, 1978), and defines an analog representation of some quantity *Q* as a representation which has a property *P*. As the quantity *Q* increases or decreases, the property *P* similarly increases or decreases. A simple example would be a mercury thermometer: the height of the mercury column (*P*) can be said to *represent* temperature (*Q*). As the temperature increases by an amount *d*, so does the height, by an amount linearly dependent on *d*. In such a representational format, there exist intermediate representations between the representations of quantities *Q* and *Q* + *d* that will correspond to the intermediate quantities, *Q* + *1/3d*, *Q* + *1/2d*, etc. Note that this does not require the representation to be continuous: a column that, instead of being made of mercury, is drawn on a display has finite precision (the number of pixels in the display). It represents temperature in discrete steps, but still the intermediate steps in the representational space have a 1-1 correspondence with intermediate quantities, and that is what matters. Hence Maley's (2011) distinction between analog and continuous. This kind of representation can be contrasted with a *digital* representation, such as the binary representation of a number. The binary representation of 44, for example, is 101100. If we add some jitter to this binary number, by randomly flipping digits, we can equally likely get the representation of 60 (111100), 12 (001100) or 46 (101110). The 'proximity' of states in this representational space does not have any meaningful correspondence to that of the quantities they represent.

In discussing analog representations, we can notice quite a few similarities with some of the representational schemes discussed in the previous sections. For example, the straightened representations of video sequences found by Hénaff et al. (2019) are clear examples of analog representations: the representation of an intermediate frame corresponds to the point between the representations of the preceding and following frames along the straight trajectory. Disentangled representations, similarly, are clearly analog: a so-called 'latent traversal' (varying the value of one disentangled dimension, while keeping others constant) will correspond to a smooth variation in some object property, for example a face's orientation or hair length. Note that, however, positing analog representations does not require to assume any specific representational scheme, linear or non-linear, disentangled or not. Again, all that matters is what the representation *can do*.

1.3.6 Evidence for analog representations in cognition

Analog representations are thought to be involved in several cognitive processes, most notably many of the mental transformations that we listed above. Shepard & Metzler (1971) famously found that participants' reaction times for comparing two 3D shapes in different orientations increased proportionally with the angular difference between them. To confirm that participants traversed intermediate orientations in their minds during this process, Cooper (1976) determined each participant's speed of mental rotation, and then presented objects at either the correct or incorrect intermediate angles during the rotation. She found that reaction times for identifying the shape were faster when it was presented at the correct intermediate orientation. These results suggest, then, that close-by angles of rotations are represented closer together, providing strong evidence for analog representations being involved in mental rotation. A similar paradigm was used by Kourtzi & Shiffrar (2001) to show that when subjects perceived an apparent rotation between two flashed views of deformable objects, objects with an intermediate orientation and shape were processed faster. This suggests that analog representations are involved in a variety of different mental transformations, not limited to rigid ones.

A separate line of evidence for the role of analog representations comes from computational models. Recently, Rajalingham et al. (2022) showed that in predicting the final position of a bouncing ball after occlusion, recurrent neural networks that were explicitly trained to track the dynamics (intermediate positions) of the ball behind the occluder better matched human and monkey behavior, than models only trained to predict the final outcome. The traversal of intermediate internal representations, then, seems to also be involved in predicting the physical dynamics of objects. Dulberg & Cohen (2020) trained artificial neural networks to predict simple transformations (translations and rotations) of 2D shapes. They found that networks that learned to apply small transformations iteratively were better able to generalize to unseen amounts of translation or rotation than networks learning to apply larger transformations at once. This suggests that the need to flexibly generalize to unseen amount of transformation might have a role in the emergence of analog representations. These studies show that enforcing a form of analog representations in computational models both increases their similarity to human behavior, and provides computational benefits, such as better generalization.

We have seen a series of distinct, but related, ideas about what constitutes a 'good' perceptual representation, such as *disentangled* and *analog* representations. We now turn to the question of what purpose these kinds of representations might serve.

1.3.7 Representations in here

In the previous sections, we have suggested that representations which transform like the things they represent can be useful: disentangled, linear and analog representations are different ways to think about this principle. Moreover, we have reviewed substantial evidence that these types of representation exist in human perception. But what are they *good for*? One clear advantage of a representation that supports the prediction of object transformations is that it allows to explicitly imagine how things will unfold. Most people can evoke vivid image-like experiences from memory or imagination (*mental imagery* - Dijkstra et al., 2019; Pearson, 2019). This ability is believed to help in tasks such as planning (Hamrick, 2019), reasoning (Hegarty, 2004) or memory recall (Mullally & Maguire, 2014; Schacter et al., 2012). These tasks all involve slow, deliberate predictions, that happen separately from the scene currently in front of our eyes, on a sort of 'mental canvas'.

Many of the mental transformation tasks we mentioned also involve a similar detachment from external stimuli: in the classic Shepard & Metzler (1971) mental rotation study, for example, participants' task was to compare two shapes at different orientations. They solved it by bringing one into alignment with the other, rotating it in their mind and then comparing the rotated mental image with the stimulus. Subsequent research found additional evidence that the representations being manipulated in this task are image-like (Cooper & Shepard, 1973; Koriat & Norman, 1984, 1988; Shepard & Cooper, 1982; Stewart et al., 2022). Moreover, studies that broke mental rotation into separate cognitive sub-processes, using eye fixations or computational modeling (Just & Carpenter, 1976; Xue et al., 2017; Larsen, 2014; Hamrick & Griffiths, 2014), found that this task involves a complex series of computations, such as determining the correct amount and direction of rotation, and comparing the two shapes. It would seem, then, that the role of internal representations is to support these cognitively effortful mental simulations, that are initiated deliberately and unfold on a timescale of several seconds.

This seems in conflict with the example that initially motivated our foray into representations: it all started with me encountering a tiger in the wild, needing to predict how long it would take it to turn around and see me. This does not look like an operation that could be done over several seconds. It needs to be faster. Additionally, much of the evidence for disentangled representations in the brain suggests that they are involved in *any* form of perception, not just imagination- or reasoning-related processes. Of course, our perceptual representations might come in different levels of detail depending on the task (e.g. Thoma et al., 2004; Li et al., 2022). Still, what could be the purpose of representations that allow detailed predictions being involved in general perception? The answer to this question might bring us closer to our tiger example, as we will see in the next section.

1.4 Representations out there

1.4.1 Dynamic representations

The 'mental canvas' described in the previous section, completely detached from the external environment, does not seem to be of much use in real-time perception. Often, our predictive abilities need to *interact* with what is in front of our eyes rather than being siloed from it. It is true that, even in mental rotation studies, ultimately participants have to compare their internal representation to an externally presented object in order to complete the task. However, the internal and external objects are still maintained as separate entities. But consider the case of the tiger I have encountered, with its back facing me. Suppose I form a prediction, in my mind, of what it will look like from different viewpoints, so that I know which way to run away without it seeing me. Should this prediction be similar to those formed by participants in mental rotation studies? Should it occur on a mental canvas, and be compared with visual input at each moment? This does not seem to match our subjective experience. In such a situation, I don't have the experience of creating and maintaining such an image in my head, and continuously comparing it with the tiger out there. Of course, this might be happening outside of my conscious awareness. But the fast timescale needed here just does not match the empirical findings of slow, deliberate operations being involved in mental rotation (Just & Carpenter, 1976; Xue et al., 2017; Larsen, 2014; Hamrick & Griffiths, 2014). Instead, it seems like the prediction of the tiger's appearance from different viewpoints is represented *implicitly* in some way. What could be the nature of this representation? One helpful suggestion comes from some of the empirical findings we reviewed previously. Take Hénaff et al. (2019)'s finding of natural video sequences being straightened in perception. If visual input is represented in this way, there is no need for a separate 'mental canvas' to predict what an object will look like from viewpoints that are just a few movements away. The visual scene in front of my eyes, in my representational space, lies in a position that neighbors that of other visual scenes that are likely to follow it. The representation of the tiger from different viewpoints, then, is already implicitly activated as I look at it.

A similar idea was proposed by Freyd (1987), who argued that our perceptual representations are intrinsically *dynamic*. Because we live in a dynamic world, and our visual system intrinsically needs to be sensitive to temporal changes to ensure our survival, we represent things as snapshots of a sequence of events. Several experimental findings point to this idea being true: for example, we perceive hand-drawn characters to be similar, if they are distorted in a way consistent with the sequence of marks involved in drawing them (Freyd, 1983; **Figure 1.5A**). When recalling the final position of a moving stimulus, we tend to displace it forward, suggesting that we naturally tend to extrapolate its motion

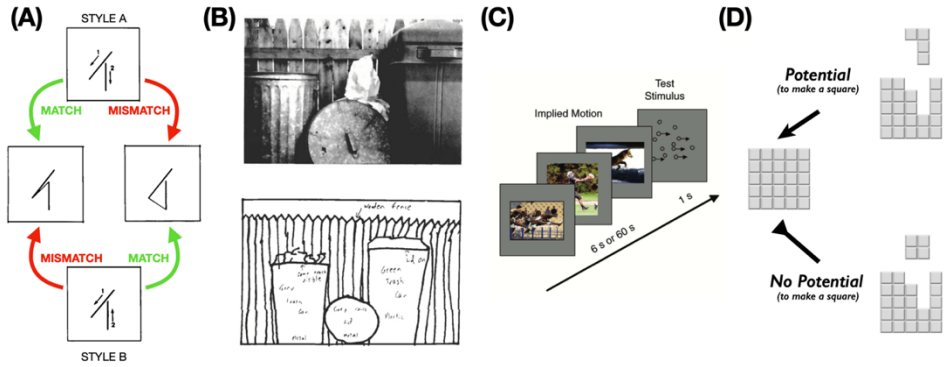


Figure 1.5. Dynamic representations. **(A)** Participants' judged similarity between handwritten character depends on whether a deformation is compatible (match) or not (mismatch) with the gestures used to draw them (Freyd, 1983). **(B)** Participants tend to extrapolate beyond the visible context in recalling (e.g. drawing) scenes (Intraub & Richardson, 1989). **(C)** After seeing static photographs implying motion, participants show a motion adaptation aftereffect for moving stimuli (Winawer et al., 2008). **(D)** When making rapid responses to the presence of a stimulus, participants confuse a display that has the potential to form a whole (Potential, above) with the corresponding whole (left). They do not confuse displays without the potential to form a whole (No Potential, below) (Guan & Firestone, 2020).

(*representational momentum*: Freyd & Finke, 1984; Hubbard, 2005). We similarly extrapolate beyond the visible information in remembering and perceiving scenes, by automatically widening the scene view in our mind (*boundary extension*: Intraub & Richardson, 1989; Intraub et al., 1996; **Figure 1.5B**). Seeing static photographs depicting motion shares some of the mechanisms involved in perceiving physical motion, eliciting a motion aftereffect (Winawer et al., 2008; **Figure 1.5C**). And in rapidly detecting stimuli comprising detached, compatible Tetris pieces, participants tend to incorrectly respond to the 'whole' that would result from joining them (Guan & Firestone, 2020; **Figure 1.5D**), suggesting that they automatically represent the plausible outcome of the scene in front of them. These disparate findings all point to perceptual representations being organized by their likelihood of co-occurring, or following one another. This might bring back to mind two ideas discussed above: second-order isomorphisms, and analog representations. When we discussed second-order isomorphisms, the idea that a 'good' representation should preserve relevant similarities among things in the world, we realized we were missing a notion of what constitutes a *relevant* similarity. In an analog representation, the proximity of representations corresponds to that of the quantities they represent (Kulvicki, 2004, 2015), such as in the case of a thermometer representing temperature by the height of its mercury column. Analog representations, then, are naturally compatible with the

idea that things should be represented as points along a dynamic trajectory. The principle of dynamic representations, then, suggests that analog representations might emerge from the need to deal with a constantly changing world. Interestingly, this idea also seems to have close ties with disentanglement. Klindt et al. (2021) showed, mathematically and empirically, that enforcing the transitions between the representations of neighboring video frames to be small leads to the emergence of representations that disentangle the factors of variation in the data. Representations that preserve the temporal contiguity structure of the world, then, naturally tend to preserve other structural regularities as well. We have now seen that disentangled and analog representations are closely related, and that beyond their potential to support offline imagination, they can have a central role in dynamic perception as well. In the next section, I will discuss some ways in which these two roles might interact.

1.4.2 The occlusion problem

Another reason to believe that the same representations might have a role in both 'regular' perceptual processing and imagination, is the fact that in the real world, the distinction between these two tasks is often blurred. A clear example of this blurring is the fact that we constantly need to deal with *occlusion*. Objects are often partially or completely occluded by other objects, yet we are able to successfully recognize them, and have the subjective impression of a seamless visual scene. In the case of partial occlusion, our visual system is known to have a series of mechanisms to 'fill in' the invisible part of the stimulus based on the surrounding context (Pessoa et al., 1998; Thielen et al., 2019), using cues such as the continuity of edges. These mechanisms are distinct from simply inferring object identity from the visible part, and there is substantial evidence that we actually represent what the object looks like behind the occluder (Ringach & Shapley, 1996; Carrigan et al., 2016; Gold et al., 2000; Lande, 2021). An occluded segment of a contour, for example, is represented as having a specific orientation and location. This process, then, already involves a form of prediction closely intermixed with regular perception. However, it does not seem to require the full machinery of disentangled or analog representations described above. It can be solved by relatively simple computations, based on low-level visual cues (e.g. Fantoni & Gerbino, 2003), while above we discussed representations that can faithfully mirror the transformations of real-world objects. A case that brings us closer to that level of processing is that of fully occluded objects. When an object is fully hidden behind an occluder, it is impossible to use contextual information from the surround to infer its appearance. Instead, what is required is some form of memory process, either long- or short-term, that uses prior information that is not currently present in the visual field. An example of long-term memory, in this case, would be visiting Piazza San Marco in Venice, but finding that a tall barrier is hiding the Basilica. Having seen the unoccluded view of the Piazza before, either in real life or in photographs, we can still represent the church behind the occluder.

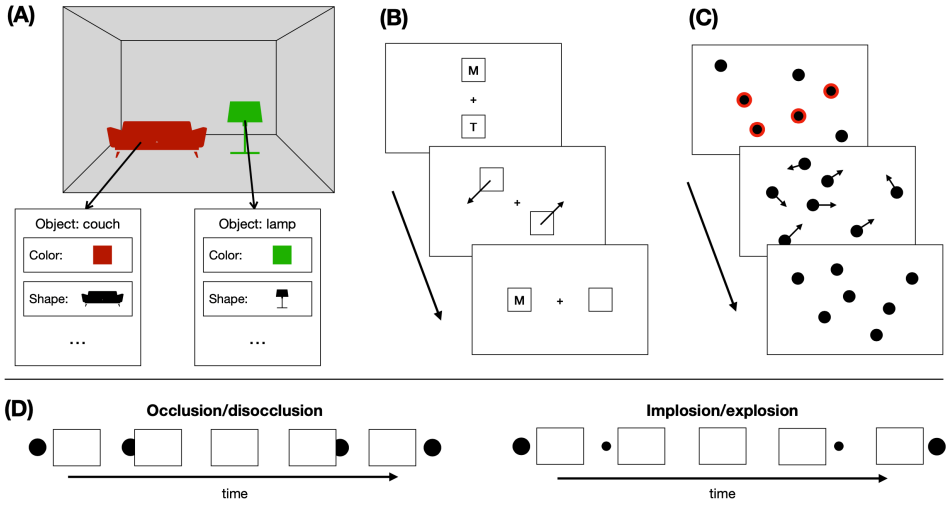


Figure 1.6. Object files. **(A)** Illustration of the concept of an object file. Each file stores the information (features) of a specific visual object, and it is bound to that object in the external scene. **(B)** Object reviewing (Kahneman et al., 1992). See text for explanation. **(C)** Multiple object tracking (Pylyshyn & Storm, 1988). See text for explanation. **(D)** Illustration of the 'occlusion/disocclusion' and 'implosion/explosion' conditions used by Flombaum & Scholl (2006).

Here, we will be more concerned with short-term processes involving experiencing an object which is later occluded, such that the representations of the object before and during the occlusion period need to be integrated.

1.4.3 Object Files

Representing objects across periods of occlusion is believed to involve a specially devoted mechanism, called an *object file* (Kahneman et al., 1992; Scholl & Flombaum, 2010; Green & Quilty-Dunn, 2020). Object files are akin to memory slots which *refer* to a particular object in the environment, such that the object's different features (e.g. color, shape) are bound together as a unit (**Figure 1.6A**). Originating as an explanation of findings from cognitive psychology, the principle of object files has been computationally implemented in a variety of ways (see Greff et al., 2020; Peters & Kriegeskorte, 2021 for reviews). It provides an elegant solution for the problem of representing objects as cohesive units, which many believe to be a fundamental precursor of our ability to interact with the world (e.g. Spelke, 1990; Lake et al., 2017; Shanahan et al., 2020).

Experimental evidence for object files comes from a variety of classic paradigms, most notably the o*bject reviewing* (Kahneman et al., 1992) and

multiple object tracking (MOT; Pylyshyn & Storm, 1988) paradigms. In object reviewing (**Figure 1.6B**), two stimuli (such as letters) are initially shown on two objects (squares). The letters then disappear, and the objects are moved to different locations. After that, a test letter appears, and the participant has to indicate whether it is the same as one of the two initial letters. When the letter appears on the same object it initially appeared on, participants are faster at correctly reporting it was one of the initial letters. Despite being in a new location, then, the letter has a processing advantage by virtue of remaining bound to the same object (*Object-specific preview benefit*, OSPB). Kahneman et al. (1992) explained this effect in terms of object files: seeing the initial display, the participant 'opens' two files associated with the objects in their short-term memory. If a letter reappears on the same object as in the initial display, it can be compared with information already present within the same object file, facilitating the response. The object file, then, aligns information related to the same object across changes in position. In MOT (**Figure 1.6C**), on the other hand, several objects are shown on the screen, and a subset of them (targets) are briefly highlighted, to instruct the participant to track them. The objects then start to move randomly for several seconds, after which the participant either has to indicate whether a particular object was a target or not, or to report all of the targets. Pylyshyn (2000, 2007) explained participants' ability to perform this task by positing the existence of *visual indexes*, which bind objects on the screen with individual object files in short-term memory.

An object file, then, is clearly an internal representation, but one that *refers* to a specific item in the external world. Most relevantly for us, it is a representation which can *track* specific object properties across periods of occlusion: in MOT, for example, participants are still able to track the targets' positions when they are temporarily occluded (Scholl & Pylyshyn, 1999). Or when a moving object reappears from behind an occluder, provided that its timing and position is consistent with continuous motion, it is perceived as the same object, despite changes in features such as color or shape (Burke, 1952; Flombaum et al., 2004). This perceived continuity is associated with increased change detection performance, which interestingly, does not occur when the object is shown to implode and then explode on the sides of the occluder, instead of going behind it (Flombaum & Scholl, 2006; **Figure 1.6D**). When contextual information suggests that two different objects have gone in and out of existence, then, it seems like the existing object file is closed and a new one is opened, eliminating the processing advantage for features within the same object file. During these occlusion periods, information about the object is not only being maintained, but also updated (to successfully predict the location and time of its reappearance). Neuroimaging evidence has confirmed that information about an object's location can be decoded in visual cortex during the occlusion period (Erlikhman & Caplovitz, 2017; Teichmann et al., 2022). Moreover, while tracking in space has been investigated most extensively, humans are also able to track objects changing in different features, such as orientation or spatial frequency (Blaser et

al., 2000), including across occlusions (Makin & Bertamini, 2014; Makin & Chauhan, 2014). They can also track location and other object features simultaneously (De Freitas et al., 2016). Information about multiple features of objects, then, can be maintained and updated across occlusions, suggesting a possible role for representations that can support continuous predictions in processes that are tightly interwoven with perception.

The representation of an object when it is visible and when it is occluded interact closely, leading some authors (e.g. Munton 2022) to question the usefulness of the distinction between perception and memory in such contexts. Munton (2022) questions what she calls the 'conveyor belt' model according to which representations are transported into memory as soon as an object becomes occluded, and back into perception when it reappears. Instead, she proposes that the relationship between perception and memory might be better described as a 'luggage carousel'. Information can be added to the carousel as we register new sensory stimuli, or removed when we cease to track an object. But our perceptual experience depends on all the information present on the carousel, whether it is currently visible or not. In this view, internal representations and external stimuli interact continuously. We have seen that representations that support predicting object transformations, such as disentangled and analog representations, can play an important role both in perceiving the world, and in imagining things on a 'mental canvas'. The tight link between 'perception' and 'memory' (for lack of better terms), exemplified in occlusion, might actually suggest an answer to the question of how these two roles are related. As a dynamic object goes out of sight behind an occluder, we use our predictive abilities to update its properties in our mind, but this prediction does not occur on a mental canvas shut off from the external world, instead being bound to a specific location in the visual scene. Through a visual index-like mechanism, we maintain this binding between our internal representation and the outside world. Exciting as this theoretical picture might sound, the examples of tracking we have mentioned thus far involved quite simple stimulus properties. In tracking across occlusions, we are able to update an object's retinotopic location, its orientation, its spatial frequency… None of this seems to scale to the complexity of real-world scenes. When we slowly, deliberately imagine things in our head, on the other hand, we seem to be able to simulate how 3D objects transform, how they physically interact, and many other incredible things. Is there, then, a fundamental difference in the complexity of predictions made during dynamic perception and deliberate imagination? This might well be the case, but before making this conclusion, it is worth considering some experimental results we have thus far neglected.

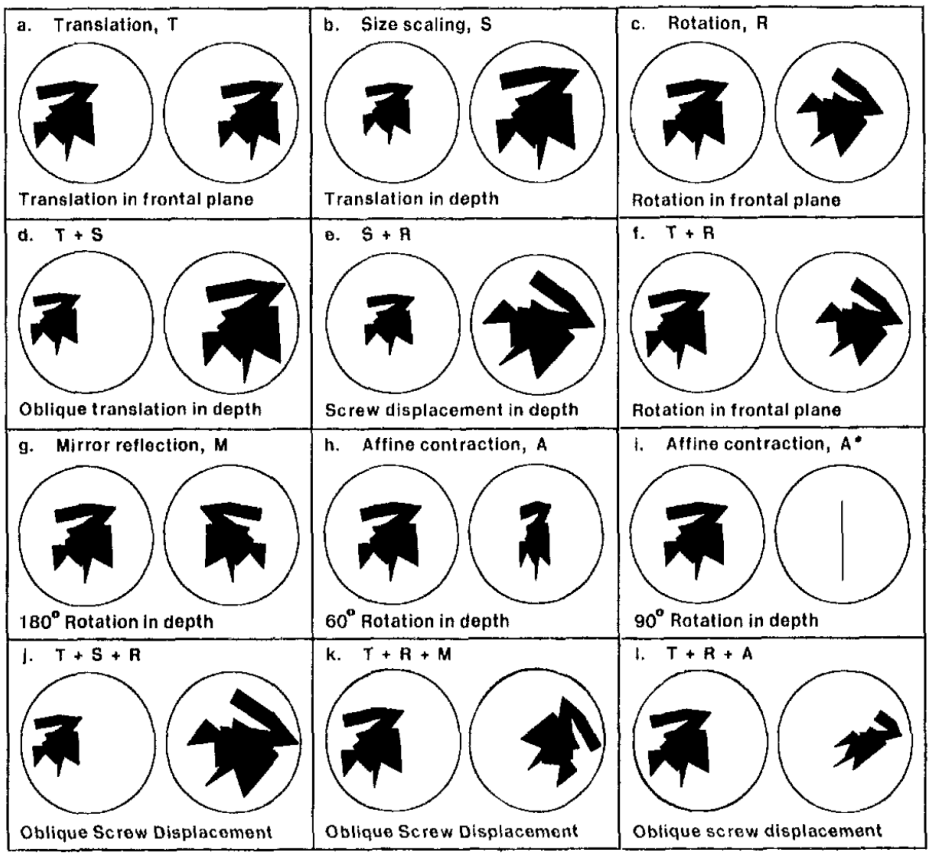


Figure 1.7. Examples of the spatial transformations that can be perceived in apparent motion (from Shepard, 1984).

1.5 Context and relations

1.5.1 Representations in context

Earlier, when describing some of the evidence for our internal representations reflecting the transformations of external objects, I focused on certain tasks. In particular, I have mentioned tasks such as mental rotation, in which a participant deliberately creates and manipulates an image in their mind. From this limited view, it would seem that predicting complex transformations, such as 3D object rotations, can be only done within one's own mind, shutting off the external world. However, this picture is far from complete. The previously mentioned (in the

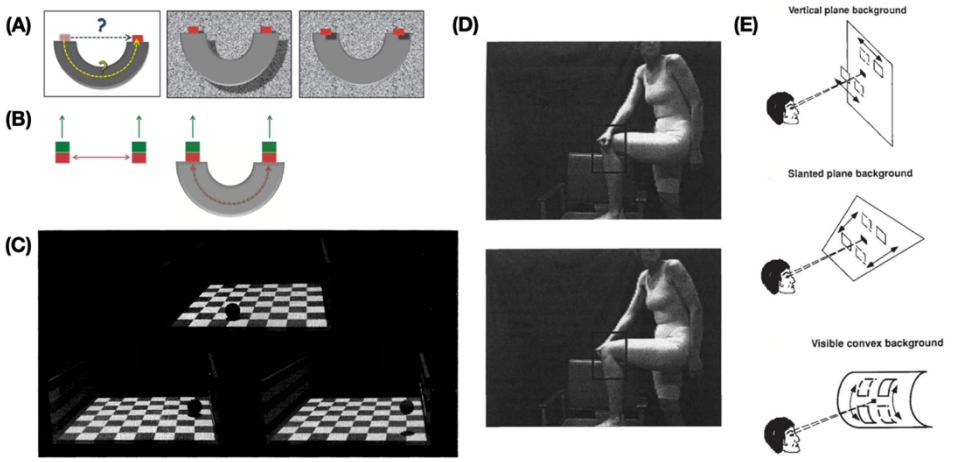


Figure 1.8. Examples of interaction of an internally generated representation (apparent motion) with contextual stimuli. **(A)** The red square, quickly flashed in the left and right position, can be seen as going through the curved 'tunnel'. This effect occurs when the tunnel is shown stereoscopically in front of the square (middle), but not behind (right) (Kim et al., 2012). **(B)** Showing the green stimuli, which move up consistently with being 'launched' by the red ones, leads to the perception of curved motion through the tunnel (Kim et al. 2013). **(C)** Flashing the stimulus at the top followed by the one at the bottom left leads to a perception of translation in depth, while the one at bottom right leads to perceived 'lifting off' from the ground, depending on the position of the cast shadow (Kersten et al. 1997). **(D)** Apparent motion of the arm, when these two views are displayed quickly, is seeing as 'jumping over' the leg (Heptulla Chatterjee et al. 1996). **(E)** The path of apparent motion follows the shape of a stereoscopically shown surface (He & Nakayama (1994).

context of analog representations) study by Cooper (1976) already showed that mental rotation itself can be sensitive to external stimuli. At any point during the mental rotation process, an object could be shown that either matched the current orientation in the participant's mind or not. Participants were more efficient at responding to the matching objects, indicating that the internal representation they were manipulating could interact with externally presented stimuli. Jolicoeur & Cavanagh (1992), and Jolicoeur et al. (1998) found that physically presented motion can influence the speed of mental rotation, which they interpreted as evidence for a common neural substrate. These results, however, could still be interpreted in terms of a separate mental canvas on which participants rotated the object. In Cooper (1976)'s study, they could have compared the presented stimulus with the object they were mentally rotating, while still maintaining them as two separate representations. In Jolicoeur's studies, on the other hand, they could have applied the physical motion they observed to the representation in

their mental canvas, rather than integrating the physical and mental rotations. A different paradigm might provide us with some more convicing evidence of complex mental transformations being tightly integrated with external stimuli: apparent motion.

Apparent motion and mental transformation paradigms have a deeply connected history: they were extensively investigated around the same time, largely by the same researchers, and with similar goals. Shepard (1984, 2001) saw them both as evidence of geometrical constraints of the world being 'internalized' in our visual system. Moreover, more recent evidence suggests that they might indeed rely on common cognitive and neural mechanisms (Larsen & Bundesen, 2009). Here, it might be worth taking a step back, and first explain what apparent motion is. It is the perception of a continuous path between two stimuli that are flashed in rapid succession in two locations (Exner, 1876; Wertheimer, 1912). Beyond simple displacement, a variety of transformations can be perceived in apparent motion, including scaling and rotations in and out of the picture plane (e.g. Kolers & Pomerantz, 1971; Foster, 1975; Shepard & Judd, 1976; Carlton & Shepard, 1990a, 1990b; see **Figure 1.7**). Apparent motion can even follow the biomechanical constraints of the human body (Shiffrar & Freyd, 1990), suggesting that it relies on the full power of our ability to predict real-world transformations. Most relevantly for our purposes, apparent motion does not seem to take place on a mental canvas, but 'out there': we see it as the motion of *that* stimulus in front of us. In this sense, it is quite close to tracking under occlusion, or even to filling-in of partially occluded objects. Moreover, several studies have shown that it can be strongly influenced by contextual stimuli. Shepard & Zare (1983) showed participants either a straight or a curved path between two rapidly flashed dots, and found that the dots could be perceived as following the displayed path. Kim et al. (2012) showed a curved occluder in between two tokens (**Figure 1.8A**), and found that with a long enough time interval between them, they could be perceived as following a curved path behind the occluder. Interestingly, when they manipulated perceived depth using a stereoscopic display, they found that the curved motion was only seen when the occluder was in front of the tokens, and not vice versa. In a subsequent study (Kim et al., 2013), when they showed two additional tokens moving consistently with being 'launched' by the token exiting the tunnel (**Figure 1.8B**), they found this to induce a percept of curved motion as well. Hubbard & Bharucha (1988) found that in the presence of an obstacle, apparent motion could be seen as 'bouncing off' the obstacle. Kersten et al., (1997) flashed a ball at two different positions in a 3D scene, and manipulated the position of its cast shadow (**Figure 1.8C**). They found that the ball was perceived as either translating in depth or lifting off from the ground, depending on the shadow's position. Heptulla Chatterjee et al. (1996) rapidly showed two pictures of a human with their hand on either side of their knee (**Figure 1.8D**). Apparent motion was perceived to follow the path around the knee, incorporating the basic physical constraint that an object cannot pass through another. He & Nakayama (1994) stereoscopically showed 3D surfaces of different shapes (e.g. vertical

plane, slanted plane, convex surface - **Figure 1.8E**) and found that the path of apparent motion could be perceived coherently with the surface. In summary, the phenomenon of apparent motion provides particularly striking evidence that internally generated predictions can occur 'out there' in the visual scene. Moreover, it can interact with a wide variety of contextual information, including 2D and 3D scene structure and basic physics. Some of the reviewed effects involve particularly sophisticated inferences about the scene, suggesting a role for rich predictive representations in a process tightly linked with ordinary perception. The main takeaway, here, is that so-called 'internal' representations, usually thought of as unfolding on a mental canvas, can be tightly integrated with a visual scene. The interaction between the visible and invisible parts of the scene happens on the basis of rules that mirror those governing the interaction of objects in the world, an idea that we will explore in the next section.

1.5.2 Relating representations, representing relations

In the previous section, we have seen how internal representations and external stimuli continuously interact, in situations such as tracking objects under occlusion and perceiving apparent motion. What is common between these situations is the need to fill-in an incomplete visual scene (because of partial occlusion and temporal discontinuity, respectively). Given that 'perception' and 'prediction' are so seamlessly integrated, it makes sense that representations which mirror the transformations of external objects are involved in both. One fact that I have not mentioned explicitly, but that was implied in the examples listed above, is that not all internal representations can interact with all external stimuli in any way. There are rules governing how they interact, and similar to how individual representations mirror the transformations of the objects they represent, so do the rules of their interactions mirror how objects interact in the world. To be sure, this is not specific to interactions between internal representations and external stimuli. All perceptual representations obey rules that constrain how they can combine with other representations, similar to the role of syntax in language (Lande, n.d.). Perceptual representations are strongly context-dependent: the same edge segments in **Figure 1.9A** can be perceived as being part of a meaningful whole (a contour) or not, depending on the context surrounding them (Geisler & Super, 2000). The orientation of the two triangles in **Figure 1.9B** is perceived to be different due to the orientation of their background, despite them being physically identical. As Lande (n.d.) notes, while in language syntax and semantics are clearly distinct (a sentence can be well-formed while being meaningless, as in "colorless green ideas sleep furiously"; Chomsky, 1957), in perception whether a representation is 'well-formed' generally depends on its probability of occurring in the real world. For example, whether a set of edges perceptually combine to form a contour depends on the probability of their orientations and locations forming a contour in the environment (Geisler et al., 2001). Contextual effects are also ubiquitous in high-level vision: for example,

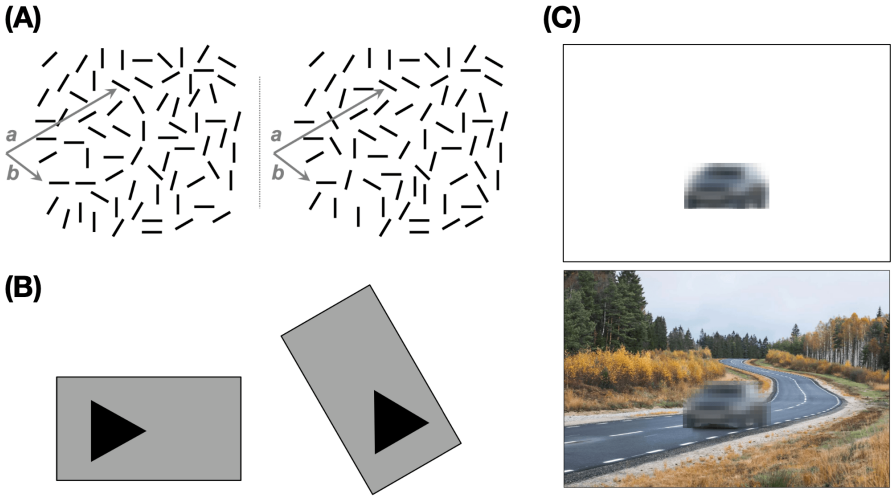


Figure 1.9. Contextual effects in visual perception. **(A)** The same two line segments are represented as being part of a contour or not, depending on the surrounding context (Geisler & Super 2000; as illustrated in Lande, n. d.). **(B)** The same triangle is seen as pointing right or top-left depending on the contextual rectangle (from Lande, n. d.). **(C)** A degraded object can be disambiguated by a background scene.

showing objects within an appropriate scene can facilitate their detection (Biederman et al., 1982), disambiguate their identity (Bar, 2004; Oliva & Torralba, 2007; Brandman & Peelen, 2017; **Figure 1.9C**), alter their perceived size (Leibowitz et al., 1969; Murray et al., 2006; Yildiz et al., 2021) or sharpness (Rossel et al., 2022). Objects shown in pairs that match likely co-occurrences in the real world elicit lower activation in visual cortex than the sum of their parts, and are ignored more effectively in visual search (Kaiser et al., 2014), suggesting that they are represented as integrated units. The presence of rules governing the interaction of perceptual representations, and the fact that these rules mirror the co-occurrences between objects in the world, is then a general fact about perception. But there is one reason why it is especially relevant in the case of incomplete (e.g. partially occluded) scenes. The problem of reconstructing the missing part of a scene involves inferring what might be hiding behind the occluder. If there were no constraints on how things in the world interact, this would be a hopeless task, as the possibilities would be limitless. Internalizing the rules of how objects interact with each other and with their context allows to predict the missing component of such interaction. The apparent motion phenomena listed above are a good instance of this. For example, the rule that objects cannot pass through

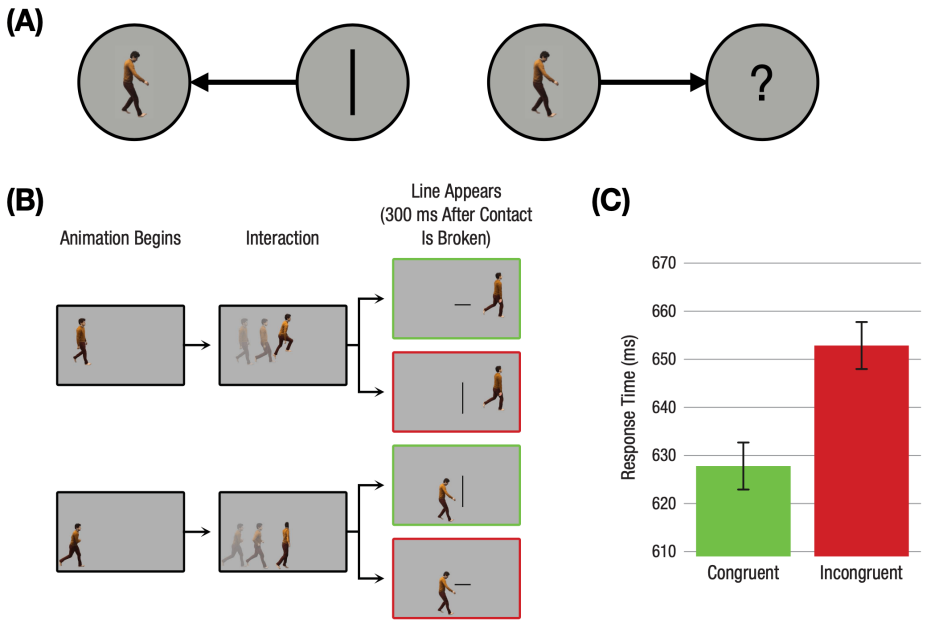


Figure 1.10. Representing relations between objects can aid the filling-in of missing objects. **(A)** Knowing how a surface and a person interact (a horizontal surface causes the person to bump into it), we can infer the shape of the surface by observing the person's behavior. Little & Firestone (2021) show precisely this kind of filling-in: a stimulus that matches the inferred invisible shape of the surface **(B)** is identified more quickly **(C)**.

other objects allows an observer to interpolate the likely movement of the hand around the leg in **Figure 1.8D**. Another clear example, this time involving filling-in an invisible static object, was recently found by Little & Firestone (2021). They showed human characters (**Figure 1.10A**) whose behavior implied an interaction with an invisible object of a particular shape: they could either step on the object (implying a horizontal shape) or bump into it (implying a vertical shape). When participants had to identify a stimulus presented subsequently, they were faster when it was congruent with the shape implied by the interaction (**Figure 1.10B**). These results, then, provide further evidence that internalizing the mechanisms by which objects interact can support filling-in of the missing components in an interaction. How might we represent these relations? Hafri & Firestone (2021) review a wide variety of findings, suggesting that physical, social and eventive relations between objects are represented as self-standing entities, and they are an integral part of visual perception. Similarly to how we can recognize objects invariantly to their lighting or viewpoint, then, we can represent a relationship such as 'containment' regardless of which object contains which other object (Hafri et

al., 2020; Ullman et al., 2019). A natural way to represent scenes in terms of objects and the relationships between them is a graph, with nodes corresponding to objects and edges to relations. Recently, artificial neural networks specifically designed to parse graph structures (Battaglia et al., 2018) have been proven to be extremely successful in a variety of tasks, including visual tasks such as object segmentation and scene captioning (see Chen et al., 2022 for a review). Is it possible that our perceptual representations are similarly organized by graph-like structures relating them? And how does the predictive power of individual object representations, described above, relate to that of the relations between them? In the next section, we will explore these questions, narrowing down our scope to a kind of relation particularly relevant for perceptual representations: spatial relations.

1.6 Spatial relations in scenes

1.6.1 Spatial relations

Spatial relations, here, are defined as the relative positions and orientations of objects within a scene. The spatial relations that are possible in real-world scenes are inevitably constrained by the physics of the world, meaning that one object's position and orientation limits the possible positions and orientations of others. A particularly clear and ubiquitous example is gravity: all objects are subject to gravity, so that for example, if a table is in a given position, other objects will be most likely to be directly on top of it, rather than floating above it. Moreover, given the orientation of the table's plane, the objects on it will most likely point in a direction parallel or perpendicular to it. Given the importance of stability as a constraint, it is considered a primary driver of our representations of spatial relations (see Kasturirangan, 2004 for an extensive treatment). Indeed, both humans (Richards et al., 1996) and computational models (Du et al., 2018) can exploit stability to infer the likely configuration of a visual scene. Moreover, objects' and scenes' alignment to gravity is represented by subpopulations in macaque object-selective cortex (Vaziri & Connor, 2016; Emonds et al., 2022). These mutual constraints, similarly to some of those mentioned above, could allow to fill-in a missing object given another. I am not aware of any work explicitly investigating if we can fill-in whole invisible objects by exploiting physical stability. Given the strength and importance of this constraint in the real world, however, it would be reasonable to think that it is incorporated into our predictions of objects' appearance.

While stability constrains the likely range of relative positions and orientations of objects, it still leaves several degrees of freedom: the vertical position of an object is constrained to be on the top of a table, for example, but it can be located anywhere on its surface (**Figure 1.11B**). Similarly, while its angles of pitch and roll are constrained to be orthogonal to the table's surface, its yaw is

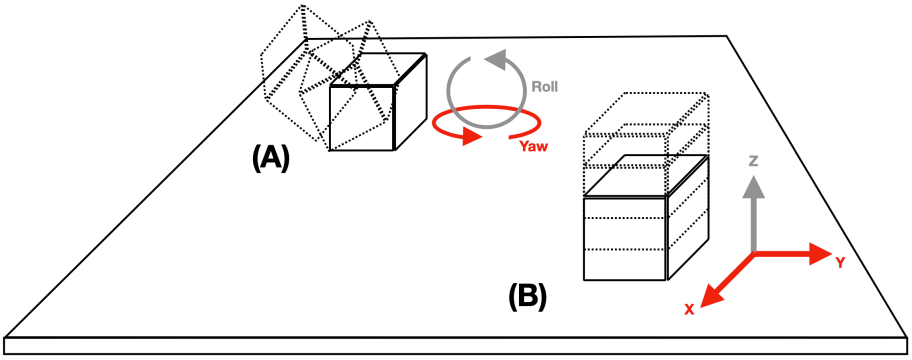


Figure 1.11. Stability provides a ubiquituous constraint for objects' orientations and positions. **(A)** Given the orientation of the table surface, two degrees of freedom of the object's orientation are constrained: pitch and roll (only roll is shown for clarity). Yaw, on the other hand, is free to vary. **(B)** Similarly for position, an object is highly unlikely to float above the table, so its *z* position is constrained. Its *x* and *y* positions, on the other hand, can vary freely.

free to vary (**Figure 1.11A**). In this case, what constrains these degrees of freedom is the fact that they tend to remain fixed across time and the observer's viewpoint. By representing the position and orientation of objects relative to each other, it is possible to exploit this constraint. A great wealth of empirical evidence indicates that humans indeed code the position of objects relative to several reference points in the environment, such as walls (Julian et al., 2016; S. A. Lee, 2017), landmarks such as buildings (Galati et al., 2010), other objects (Rieser, 1989) and the overall layout of a scene (Mou & McNamara, 2002). These relative location representations can be used, for example, to reorient ourselves in space (Julian et al., 2018), or to detect changes in a scene across viewpoints (Mou & McNamara, 2002). The case of relative *orientation* is less clear, although there is evidence that participants exploit contextual viewpoint information in object recognition: Hinton & Parsons (1988) found that when determining whether two physical objects, presented in different orientations, were the same, they rotated them to have the same orientation relative to the scene (a table), rather than to their own view (**Figure 1.12A**). Humphrey & Jolicoeur (1993) found that showing a background with a consistent slant in depth facilitated recognition for objects shown in unorthodox orientations (**Figure 1.12B**). And Christou et al. (2003) found that synthetic 3D objects in virtual reality could be recognized better when shown within a scene context, indicating that the scene could be used as a reference frame (**Figure 1.12C**). In summary, evidence points to humans using relative coding of spatial location and orientation in a variety of tasks, including spatial

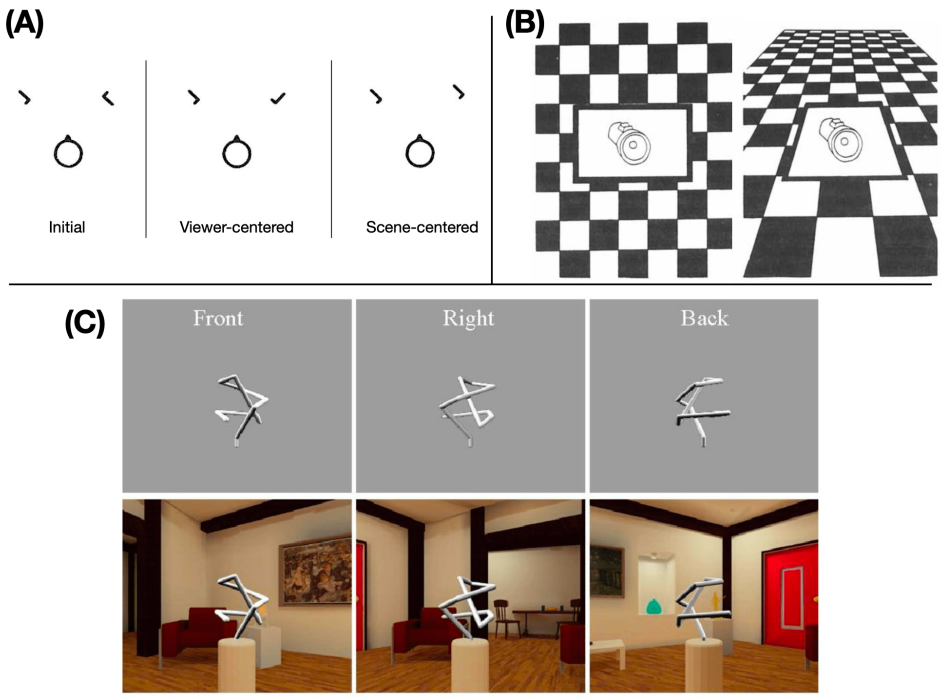


Figure 1.12. Scene context can provide a useful reference frame for object orientations. **(A)** In Hinton & Parsons (1988), participants rotated physical objects to have the same scene-centered, rather than viewer-centered, orientation in order to compare them. **(B)** Showing a background with a coherent slant can facilitate the recognition of objects from unusual angles (Humphrey & Jolicoeur, 1993). **(C)** The context of a scene can facilitate the recognition of novel objects, even when participants can freely explore them in VR (Christou et al. 2003).

reorientation and object recognition. Whether it can be used to fill-in an object, for example by providing cues to its viewpoint, is a question that will be addressed in the next chapters of this thesis. For the moment, I will turn to the question of what constitutes a 'good representation' (a question we have encountered multiple times already) in the specific case of spatial relations: enter the scene graph.

1.6.2 Scene graphs

In the previous section, I have mentioned that objects' positions and orientations can be encoded relative to other objects. While pairwise relations between two objects, or one object and its context, can already provide enough information to

support spatial reorientation, recognition, and possibly predicting object viewpoint, it might not be the most efficient representation for real-world scenes. Most scenes we encounter in daily life are structured hierarchically, with larger objects supporting smaller objects, which in turn are made of multiple parts. Generally, parts causally depend on the wholes they belong to, and smaller objects depend on the larger objects that they lie on. For example, moving a table with multiple objects on it causes all of those objects to move, while the opposite is usually not true. While interactions between other objects may occur occasionally, such interactions between wholes and their parts are default in most scenes. For this reason, it would make sense to use specialized representations to efficiently process these kinds of structures. In the field of computer graphics, *scene graphs* have been designed explicitly for this purpose. A scene graph[2] (Bar-Zeev, 2007; Cunningham & Bailey, 2001; Sowizral, 2000; **Figure 1.13A**) is a directed hierarchical graph, in which nodes correspond to objects, and edges to "part-of" relationships. A given node's children are its parts. Each object or part, beyond appearance information such as its texture and shape, stores its spatial transformations (translation, rotation, scaling) relative to its parent, usually in the form of a matrix. This structure can be traversed hierarchically, so that each individual part's transformation can be computed efficiently from its ancestors. For example, if an object, such as a robot, is updated (e.g. moved to a different location), its parts, such as its arms and legs, will be updated with it. The usefulness of this structure in modeling real world scenes has led researchers in AI to address the problem of how similar structures can be represented in distributed neural patterns and learned (e.g. Hinton, 1981, 1990, 2021). One particularly fruitful class of models has been that of *capsule networks* (e.g. Hinton et al., 2011; Sabour et al., 2017; Hinton et al., 2018; Kosiorek et al., 2019). Similarly to scene graphs, capsule networks store spatial transformations together with their corresponding objects: each 'neuron', beyond responding specifically to a particular feature, is also equipped with a representation of its relative pose (**Figure 1.13B**). The greatest challenge in capsule networks is to learn how to correctly assign parts to objects based purely on the input data, a problem which has been tackled using a variety of approaches (see Ribeiro et al., 2022 for a review). Capsule networks have shown success on tasks ranging from image classification to segmentation of 3D data such as point clouds. Granskog et al. (2021) proposed a similar approach to combine the advantages of structured scene graphs and representations that can be learned from data. They also used distributed representations of object properties, such as geometry and texture, related to each other by linear transformations (**Figure 1.13C**). In spirit, this approach is similar to enforcing linear transformations (in the representational space) of single objects, as we mentioned above. In this case, the linear

[2] This term is sometimes also used in AI (for example, in image captioning) to indicate a more general representation of all kinds of relations among objects in a scene (e.g. Johnson et al., 2015, 2018; Chang et al., 2021). To avoid confusion, please note that we are not using that meaning of the term here.

transformations are between each object and its children. They find this approach to be helpful in learning scene representations that can be manipulated piecewise, for example by only modifying the texture of a single object without affecting others. Ost et al. (2021) similarly enforced linear transformations of objects relative to each other, learning scene representations that can generate novel views and object arrangements.

Other researchers have attempted to tackle the problem of learning part-whole representations by incorporating some biologically plausible constraints: Bear et al. (2020) developed a model (**Figure 1.13D**) that learns to construct hierarchical scene representations using Gestalt-like visual cues, such as common motion, which are believed to underlie the development of object representations in early life in humans (Spelke, 1990). Hawkins and colleagues (e.g. Hawkins et al., 2017; Lewis et al., 2019) proposed a comprehensive framework for learning relative spatial representations, based on several biologically-inspired computational motifs. Gklezakos & Rao (2022) implemented part-whole representations through the use of saccade-like active sampling. Despite these early attempts at biologically plausible scene graph representations, the question of whether scenes are represented in such a hierarchical way in human perception has not been addressed thoroughly. While researchers noted early on the importance of hierarchical structure in perception in general (e.g. Palmer, 1977; Marr & Nishihara, 1978), most of the work on hierarchical representations in human vision has been in the domain of single object perception. In particular, an influential class of models (*part-based models*; Biederman, 1987) has recognized the crucial importance of parts and their arrangements in object recognition. In these models, objects are represented as collections of parts and spatial relationships between them (e.g. above, below), rather than precise metric coordinates. They can account for the finding that we often perceive changes in the part structure of objects to be much more salient than changes in their exact metric arrangement (Stankiewicz & Hummel, 1996). Moreover, since relative spatial relations are invariant of the observer's viewpoint (unlike absolute coordinates), they afford the ability to recognize objects across different orientations (Biederman & Gerhardstein, 1993). Part-based models were originally contrasted with *view-based* models (Hummel, 2000), in which objects are represented in terms of image-like templates of specific views (Bülthoff & Edelman, 1992; Tarr & Pinker, 1989; S. Ullman, 1998). In fact, experimental evidence suggests that observers might use both part-based and view-based processes in parallel during object recognition (Foster & Gilson, 2002), or switch between them through the allocation of attention (Thoma et al., 2004). Interestingly, a mixture of image-based and structural representations also underlies some recent models that learn to extract hierarchical structures from scenes (e.g. Bear et al., 2020), suggesting a promising avenue for cognitively plausible computational models. Regardless of the specific representations involved, however, the evidence that object perception involves hierarchical representations is quite abundant.

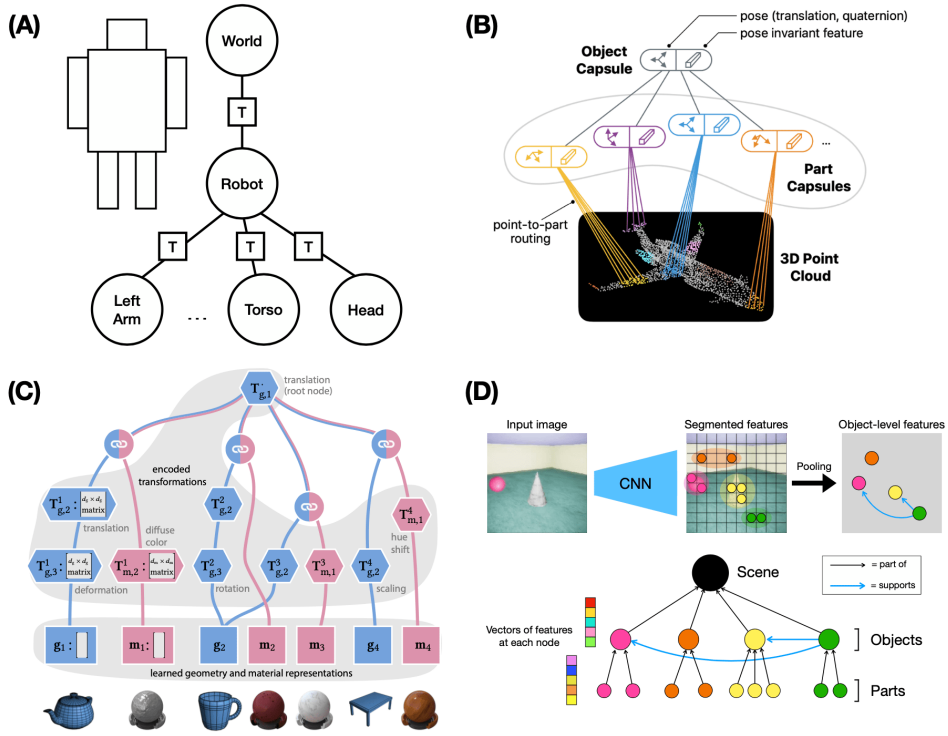


Figure 1.13. Scene graphs. **(A)** Illustration of a simple scene graph for a hierarchical object (a robot). The squares marked with 'T' stand for spatial transformations relative to the parent node. **(B)** Example of a capsule network: the network has units that respond to a specific object part, and represent its position and orientation relative to the whole (from Ribeiro et al. 2022). **(C)** Neural scene graph model from Granskog et al. (2021). **(D)** Schematic of the model from Bear et al. (2020).

Whether whole scenes are also processed in terms of hierarchical structures, similar to scene graphs in computer graphics and AI, is less clear. Such a representation would have the advantage that scenes, objects and object parts would be represented as parts of a single, deep hierarchy (see Feldman, 2003 for a view on what constitutes an 'object' in a similar framework). Processing of complex scenes, comprising multiple objects and interactions between them, seems to share some mechanisms with part-based processing of single objects. For example, in judging similarities between scenes, the features of single parts and the overall structure of the scene appear to be used in parallel: scenes are brought into 'alignment' with each other such that the features of corresponding components are compared (Markman & Gentner, 1993; Goldstone, 1996). In

attentional guidance as well, humans seem to exploit the hierarchical structure of scenes, using larger objects as 'anchors' to search for smaller ones (Võ et al., 2019). While these results could be seen as tentative evidence that hierarchical representations analogous to those for objects underlie the processing of multi-object scenes, further research should clarify whether this is really the case. Further studies should also investigate whether any of the recently proposed methods for learning scene graph-like representations from data is a viable model of scene parsing in humans. For now, we move to the final section of this Introduction, and bring together several of the topics discussed so far.

1.6.3 Scene dynamics

In the previous section, we have introduced the idea of scene graphs, which originates in computer graphics, but we have not properly motivated what its role could be in human perception. In fact, the purpose it was originally designed for in graphics might also be where its usefulness for perception lies. As we have already seen several times in this chapter, it is all about dynamics. In real-time graphics applications, where the scene needs to be rendered many times per second, speed is crucial. Structures such as scene graphs were devised to efficiently update the scene when something needs to be changed (e.g. an object is moved). As for representations of single objects, hierarchical scene representations map updates that occur frequently to small 'jumps' in representational space. Consistent with the dynamic representations account outlined above, then, efficiently representing changes, rather than static scenes, might be the goal of scene graph representations, and their main advantage. Is there any evidence that hierarchical structure is exploited by humans in dynamic tasks? Actually, there is. In motion tracking, participants are more accurate when a motion display is hierarchically structured (H. Xu et al., 2017; Bill et al., 2020, 2021), consistently with the idea that they exploit that structure to constrain their tracking. This previous research used simple, synthetic stimuli comprising multiple dots at different spatial locations. An intriguing question for future research would be whether this hierarchical structure is also exploited when tracking in complex, naturalistic scenes.

Often in the real world, objects do not simply change their retinotopic position, but can undergo several of the transformations listed above: 3D translations and rotations, deformations, changes in physical state, etc. As discussed earlier, we are able to represent and predict several of these transformations, seemingly relying on mechanisms shared with those dedicated to motion in space. At least some of these transformations also tend to follow a scene's hierarchical structure. For example in the case of rotation, if a table with several objects lying on top of it rotates, generally so do those objects. If hierarchical dependencies are exploited even when tracking these transformations, what might be the mechanisms making this possible? I argue

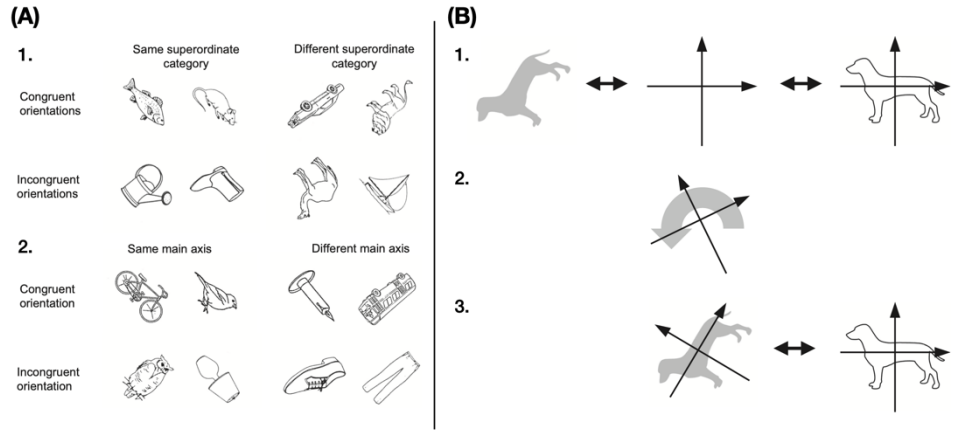


Figure 1.14. Illustration of spatial reference frames in visual perception, from Graf (2006). **(A)** An object shown with a particular orientation can facilitate the recognition of a subsequent (different) object with the same orientation (Graf et al. 2005). **(B)** This facilitation is believed to happen through the alignment of the visual scene with a canonical orientation, or in other words, by the establishment of a contextual reference frame.

that a series of empirical findings from different fields might suggest an answer. First, in motion tracking under occlusion, substantial evidence indicates that the representation of tracked objects behind the occluder is limited to certain properties of the objects. In particular, only object location seems to be represented (Scholl & Pylyshyn, 1999; Teichmann et al., 2022). In light of the 'object files' theory described above, this finding has been interpreted in terms of maintaining an index to a particular object file bound to a location in space. Spatial position, then, might play a unique role in maintaining persistent object representations (Flombaum et al., 2009; Mitroff & Alvarez, 2007), which can facilitate the detection of changes in other properties, such as color or shape (Bahrami, 2003; Flombaum & Scholl, 2006). However, some have questioned whether this distinction between space and other features is fixed (Quilty-Dunn & Green, 2021). Other features, such as color (Hollingworth & Franconeri, 2009), shape (Zhou et al., 2010) or orientation (Gordon et al., 2008), can be used to bind changing stimuli to an object file. If features different from spatial position can be used to track objects, it might be possible to infer hierarchical dependencies in a scene on the basis of these features as well. Is there any empirical evidence that inter-object dependencies can be extracted for transformations different than translation? Graf (2006) reviews a series of findings consistent with the presence of *reference frames* in object perception: the position, size or orientation of an

object can be pre-cued by other objects (Graf et al., 2005; **Figure 1.14A**) or by a coherent scene context (such as examples we have seen before, **Figure 1.12B-C**). These reference frames appear to be transformed in a manner similar to mental images during (explicit) mental transformations, by applying analog coordinate transformations (**Figure 1.14B**). However, they do not involve explicit mental images of a specific object, but rather a coordinate system that can facilitate processing of any object. In this way, they are quite similar to the abstract nature of motion tracking: one particular feature (position in the case of motion, orientation or size in the case of coordinate transformations) acts as a reference frame for an object representation. Detection of changes in other features could then be facilitated when they occur within a coherent reference frame. Clearly, further work will be required to clarify whether motion tracking and the setting of coordinate frames indeed rely on a common mechanism. It is an intriguing possibility, however, that flexible use of different features to 'guide' stable object representations could be a mechanism to exploit hierarchical dependencies in real-world scenes. While I do not commit to any specific computational implementation here, it is worth noting that disentangled representations easily lend themselves to representing links between specific features. If meaningful dimensions of objects (position, size, orientation) are associated with single dimensions of a representation, it is possible to constrain those dimensions exclusively (for example by making them dependent on a parent object) while letting others free to vary. Whittington et al. (2021) implemented a similar strategy in a model that learned to represent scenes comprising multiple objects: only the disentangled dimensions corresponding to the objects' positions were dependent on each other, while the rest was unconstrained.

In summary, scene graph-like representations would seem to be most useful in situations that require tracking the dynamic changes of a hierarchical scene in real time. Just like in the case of single objects' transformations, representing the dependencies between objects constrains which scene transformations are possible or likely given the current scene. Going through several disparate topics, I have offered some ideas of what might constitute a good representation for dynamic real-world tasks. Now, I will relate the big picture that I have just laid out with the content of the following experimental chapters.

1.7 Our experimental paradigm

In the following chapters, I will describe an experimental paradigm that aims to investigate internal object representations in humans, integrating several of the properties I have listed. In this paradigm, participants see an object within the context of a 3D scene. This object is then temporarily occluded while the scene's viewpoint, still visible, changes. We investigate, using behavior and fMRI, whether participants update the representation of the invisible object coherently with the surrounding scene's viewpoint. Thus, we investigate whether their representation

of the object: (a) behaves like the 3D object it's representing, rotating to novel viewpoints; (b) is bound to an external object, similar to object tracking under occlusion, rather than happening on a mental canvas; (c) interacts with the surrounding visual scene by respecting real-world constraints (the simple fact that an object and the surrounding scene move coherently). This paradigm, then, combines the first three of the points listed above. As our scenes were relatively simple, we did not investigate point (d) (scene graphs, specialized representations for part-whole hierarchies). Point (e) is even more speculative, and pertains to the role of hierarchical structures in dynamic vision more broadly. These points are meant to be an inspiration for future research (I will return to this in the General Discussion).

Chapter 2
Scene context automatically drives predictions of object transformations

Abstract

Humans are able to mentally transform objects in accordance with their transformations in the external world. For example, we are able to predict how a 3D object will look from a novel viewpoint. In real-world environments, objects are not generally transformed in isolation, but in accordance with their context. As we change our viewpoint, for example, we see the whole visual scene rotate coherently. Exploiting these structural regularities would enable us to predict object transformations in complex real-world scenes. Here, in a series of online behavioral experiments (N = 152), we investigate whether scene context can automatically drive predictions of objects from novel viewpoints. We find that participants' responses in an orthogonal task are strongly influenced by whether objects appear rotated coherently with the surrounding scene after a period of occlusion. This behavioral effect holds true across a variety of possible object orientations, and different amounts of scene rotation, suggesting that it reflects a flexible transformation mechanism rather than associative learning of specific views. It also persists, and does not reverse, when the scene-driven viewpoint expectation is violated on a large proportion of trials, showing that short-term contingencies cannot easily overrule it. Altogether, these findings point to a possible mechanism for tracking and predicting objects in real-world contexts.

This chapter is based on:
Aldegheri, G., Gayet, S., and Peelen, M.V. (2023). Scene context automatically drives predictions of object transformations. *Cognition 238*, 105521.

2.1 Introduction

The appearance of objects in our everyday environments is continually transforming, for example due to their motion or to changes in our viewpoint. Predicting how objects change is crucial for survival in a dynamic world, but in most cases these transformations cannot be easily inferred from visual input, making future prediction challenging. An important function of perceptual systems, then, is to extract representations that behave similarly to objects in the external world (Shepard, 1984, 2001; Higgins et al., 2022): for example, our internal representations of objects can be *mentally rotated* similarly to how external objects rotate (Shepard & Metzler, 1971); they are extrapolated forward in time, such that a moving object will be remembered as being displaced further along its motion direction (Hubbard, 2005); and they support predictions of physical dynamics, consistent with an internal simulation of real-world physics (Battaglia et al., 2013; Bear et al., 2021).

This prior research investigating the human ability to predict object transformations has focused on objects shown in isolation, driven by properties intrinsic to the objects themselves, such as their motion, or by cognitive operations determined by the participant, such as the amount of mental rotation required to match two object views. In the real world, however, objects are strongly constrained by their context (Bar, 2004; Oliva & Torralba, 2007), so that prediction often amounts to a task of *completing* the partial information present in a scene. For example, a visually degraded object's identity can be disambiguated based on its background scene, such that a blurred shape appearing on a road will be seen as a car (Brandman & Peelen, 2017) and even be perceived as visually sharper (Rossel et al., 2022). Contextual information can also be provided by other objects or agents: for example, Little & Firestone (2021) found that human subjects mentally 'fill in' the shape of an invisible object based on how an actor interacts with it, perceiving it as horizontal if the actor steps on it, or vertical if he bumps into it.

The interaction between objects and their context is not limited to fixed object properties, like their shape or size: scenes also constrain how objects transform. A particularly clear example can be seen as we navigate an environment: assuming the objects around us remain still, all of their orientations relative to us will change jointly as we move, together with the layout of the scene (e.g., the orientations of walls). Rather than predicting how objects look from a new viewpoint based entirely on internally driven mental rotation operations, then, it is possible to use contextual scene information to fill them in. The ability to use such relational information to predict an object's appearance from a new viewpoint would be highly relevant in many real-world settings. We constantly need to track and update the representations of objects in the external environment, even when they're temporarily invisible (Munton, 2022; Scholl & Flombaum, 2010; Scholl & Pylyshyn, 1999), rather than evoke and manipulate a mental image of the object separate from the current visual input. It would also

provide a computational advantage, by alleviating the burden on several cognitive processes, such as determining the correct amount of mental rotation (Hamrick & Griffiths, 2014) or finding correspondences between rotated versions of the same object (Just & Carpenter, 1976; Larsen, 2014; Xue et al., 2017). Instead, the amount of rotation of the object can be determined from the rotation of its surrounding scene, and its updated representation can be filled-in and compared with incoming visual input at the relevant scene location. Similar interactions between internally driven processes and scene context have been found in visual search, in which the visual cortical representation of the object that subjects are searching for in a scene is 'rescaled' based on the distance at which they search (Gayet & Peelen, 2022).

Here, in three online behavioral experiments, we investigated whether human participants automatically predict an object's appearance from a new viewpoint, based exclusively on changes in the surrounding scene's viewpoint. We designed an experimental paradigm in which an object (a bed or couch) was shown in the context of a realistic indoor scene, which changed in viewpoint (**Figure 2.1**). During the viewpoint change, the object was hidden by an occluder. Afterwards, the occluder would disappear and the object would be revealed: it could either be oriented consistently with the scene's new viewpoint (*Expected* condition) or inconsistently (*Unexpected* condition). Crucially, the total viewpoint change of the scene was varied across trials, so that participants' prediction could not be driven by simply extrapolating the object's rotation by a constant amount. To process an object's view as Expected or Unexpected, they needed to take into account how much the surrounding scene had rotated. The (Expected or Unexpected) object was displayed briefly twice, and participants had to indicate whether these two appearances had the same orientation or not (same/different). We compared their performance, in terms of both sensitivity (d') and response bias (criterion) between Expected and Unexpected trials. Any performance difference would indicate that they internally represented the updated object view. The task was orthogonal to the expectancy manipulation, and participants were not explicitly instructed to use scene viewpoint or try to predict the upcoming object view. We were interested in gauging whether they would automatically extract the dependency between scene and object.

In Experiment 1, we found that both participants' response bias and sensitivity were affected by whether the object matched the scene-driven expectation. On Unexpected trials, participants tended to give more 'different' responses and had lower sensitivity compared to Expected trials. Since in this experiment, the object appeared more frequently in the Expected than the Unexpected view (75% of trials), we next asked whether the effect of scene-driven expectations could still be found if the two views appeared with equal probability (50% of trials). In Experiment 2, we found this to be the case, finding the exact same profile of results as in Exp. 1, in both response bias and sensitivity. Finally, in Experiment 3, we found that the effect of expectancy persists in the same direction even when the Expected view is shown only on a minority of trials (25%).

Together, these results indicate that scene viewpoint influences object expectations in an automatic fashion. Moreover, this influence is not attenuated or reversed when violated frequently during the experiment, suggesting that it derives primarily from real-world regularities and cannot be easily overruled. Automatically predicting object transformations on the basis of scene context might be a mechanism to overcome the complexity of the real world by exploiting its regularities.

2.2 Methods

2.2.1 Participants

All experiments were run online, hosted on Pavlovia and programmed in Javascript using JsPsych 6.3.0 (De Leeuw, 2015) and the jspsych-psychophysics library (Kuroki, 2021).
Online participants were recruited on Prolific (Palan & Schitter, 2018), and had to satisfy the following criteria: reside in Europe or the UK, to ensure their timezone was the same as ours and they were participating during day hours; be between 18 and 35 years old; have normal or corrected-to-normal vision; have participated in at least 10 previous studies on Prolific; and have a Prolific approval rate of at least 95%. Participants provided informed consent before the study and received monetary compensation for their participation. The study was approved by the Radboud University Faculty of Social Sciences Ethics Committee (ECSW2017-2306-517). Participants were included in the analysis if a one-sided binomial test comparing their hit rate in our same/different task with 50% was significant (at alpha = 0.05), meaning that they were performing better than chance. We continued data collection until the number of included participants reached 50 for each experiment. In Experiment 1, we excluded 30 participants. Of the included 50 participants, 25 were females, and mean age was 26.7 ± 5.1. In Experiment 2, we excluded 33 participants. Of the included 50 participants, 20 were female, and mean age was 24.5 ± 4.3. In Experiment 3, we excluded 56 participants. Of the included 52 participants, 25 were female, 26 male and one participant's demographic information was missing. Mean age was 24.7 ± 4.4.

The high exclusion rate was likely due to several reasons: we kept a very short presentation time (50 ms) for the two probes, in order to reduce the influence of deliberate judgment and find evidence of a perceptual representation of the object's updated appearance, also making the task more challenging; we limited the maximum orientation difference between the two probes to 20°, to avoid exceeding 1/3 of the difference between expected and unexpected views (60°). This meant that the staircase was limited in its ability to adjust to participants with a higher discrimination threshold.

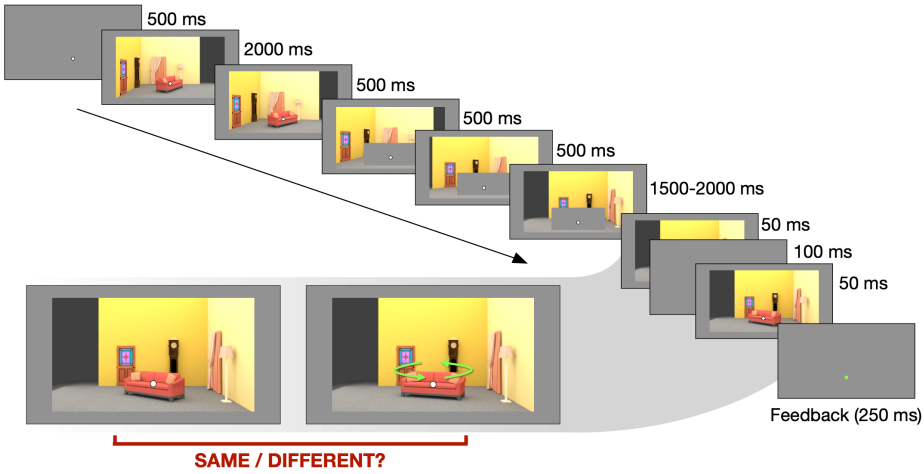


Figure 2.1. Example trial - in this case, corresponding to a "Large" total rotation (90°), an "Expected" view, and a "Different" trial: the second probe is slightly rotated relative to the first (arrows added for illustration).

2.2.2 Stimuli

The stimuli were based on 8 different indoor scenes modeled in Blender 2.92 and rendered using the Cycles rendering engine for realistic lighting. The scenes all had the same layout (floor, two walls at a right angle and a main object in the center) but contained various other objects, adjacent to the walls, and different textures on the walls and floor. The central object was a couch on half of the scenes, and a bed on the other half. The central object's size was the same across scenes. For each scene, a sequence of different viewpoints was rendered, by rotating the scene around the vertical axis between 0° to 90° in steps of 5°. The two walls were oriented such that the scene was fully visible from all these viewpoints. All scene images were presented at a resolution of 960 x 540 pixels.

2.2.3 Procedure

Each trial (**Figure 2.1**) began with a fixation dot (which was always present during the trial, radius 5 pixels) for 500 ms, followed by the first view of the scene for 2000 ms, the 3 intermediate views for 500 ms each, and the final view for a randomly jittered duration between 1500 and 2000 ms.
The central object (couch or bed) was fully visible for the first and second view, and was occluded by a grey rectangle during the third, fourth and final view. The occluder had the height and width of the largest possible view of the object, plus

a margin (horizontal: 110 pixels, vertical: 40 pixels) to ensure the object was fully covered and its shadow was not visible, which would have provided a cue to its orientation.

After the final view of the scene was shown, with the object still fully occluded, the object was briefly flashed twice (within the scene) for 50 ms, with a 100 ms inter-stimulus interval in between. We refer to these two brief presentations of the object as the *probes*. Participants' task was to report whether the second probe was the 'same' as, or 'different' from, the first, by pressing the F or J key, respectively. After responding, they would receive feedback: the fixation dot would turn green following a correct answer and red following an incorrect one for 250 ms. They had a maximum of 2500 ms to respond, after which the fixation dot would turn black, the experiment would skip to the next trial and the current trial would be counted as missed.

Participants were explicitly told that their task would be on the final two views of the objects exclusively, but that they should also pay attention to the preceding sequence of images, to ensure they wouldn't completely disengage during the seconds preceding the probes.

The first probe was randomly sampled from a normal distribution centered around the Expected or Unexpected object viewpoint, with a standard deviation of 1°, to add a small amount of jitter, and then rounded to the nearest integer.

The second probe, on half of trials ('same' trials), was exactly the same as the first probe. On the other half of trials ('different' trials), it was rotated around the vertical axis relative to the first (see **Figure 2.1**, bottom left), clockwise or counterclockwise with equal probability.

The orientation difference on the 'different' trials was titrated using a 2-down 1-up staircase, to keep the task difficulty constant across participants and across experiments. Specifically, a single staircase was used across both Expectancy conditions to ensure average performance around 70% correct (Wetherill & Levitt, 1965) across conditions, while still allowing for accuracy differences between the Expected and Unexpected conditions. Stimulus intensity (orientation difference between probes) was adjusted after both 'same' and 'different' trials. The starting value for the staircase was 10°, step size was 1° (lowered to 0.5° after 3 staircase reversals) and the minimum and maximum possible orientation differences shown were 0.5° and 20°, respectively. The means and standard deviations of the angle differences reached by the staircase in the second half of trials, in each experiment, were 12.76° ± 4.64, 11.96° ± 5.18, and 14.11° ± 4.90 respectively.

Each experiment lasted about 30 minutes in total, divided in blocks, and participants were encouraged to take a short break after the end of each block. Before the experiment began, participants read instructions, accompanied by demonstration images, at their own pace. Then they completed a short practice run. During the practice run, the presentation time of the two target probes gradually decreased across trials, from 300 ms to their presentation time in the main experiment, 50 ms. This allowed participants to familiarize with the task with

an initially less challenging presentation time.

2.2.4 Experimental design

Trials varied along three different factors (**Figure 2.2**): Expectancy (Expected, Unexpected), Object Orientation relative to the scene (6 angles: 0°, 60°, 120°, 180°, 240°, 300°), Scene Rotation (Small, Large) and Scene (1 of 4 different scene exemplars, one of two subsets of the 8 total views, selected randomly for each participant).
The overall proportion of Expected and Unexpected trials varied depending on the experiment (75% of total trials in Exp. 1, 50% in Exp. 2, and 25% in Exp. 3). All the other factors were fully balanced within the Expected and Unexpected trials, meaning that each of four partitions of the trials (variably assigned to either Expected or Unexpected depending on the experiment) were equally divided among each combination of Object Orientation, Scene Rotation and Scene (6 x 2 x 4 = 48 trials for each partition, resulting in 192 trials in total). All these trials were presented in random order throughout the experiment.

The Unexpected view corresponded, on Small rotation trials, to a view that was rotated 60° more than expected, and on Large rotation trials, to a view rotated 60° less than expected. The 6 initial object orientations were chosen to be 60° apart, so that the Unexpected view for one orientation corresponded to the Expected one for another. This way the exact same images could be presented as Expected in the context of one trial, and Unexpected in another, avoiding any possible confounds due to physical differences (**Figure 2.2**).

2.2.5 Data analysis

In order to distinguish the effects of scene-driven expectations on observers' perceptual sensitivity and response bias, we computed d' and criterion for each of the two conditions of interest (Expected and Unexpected trials). We consider 'Same' trials as noise, and 'Different' as signal, meaning that criterion measures the tendency to respond 'same'. We used the log-linear method (Hautus, 1995) to correct for the rare cases of 100% accuracy in a particular condition.

All analyses were conducted in Python using Pandas 1.2.5 (McKinney, 2011), Numpy 1.20.2 (Harris et al., 2020), Pingouin 0.3.4 (Vallat, 2018), and Scipy 1.6.2 (Virtanen et al., 2020), and results were visualized using Matplotlib 3.3.4 (Hunter, 2007), and Seaborn 0.11.1 (Waskom, 2021).

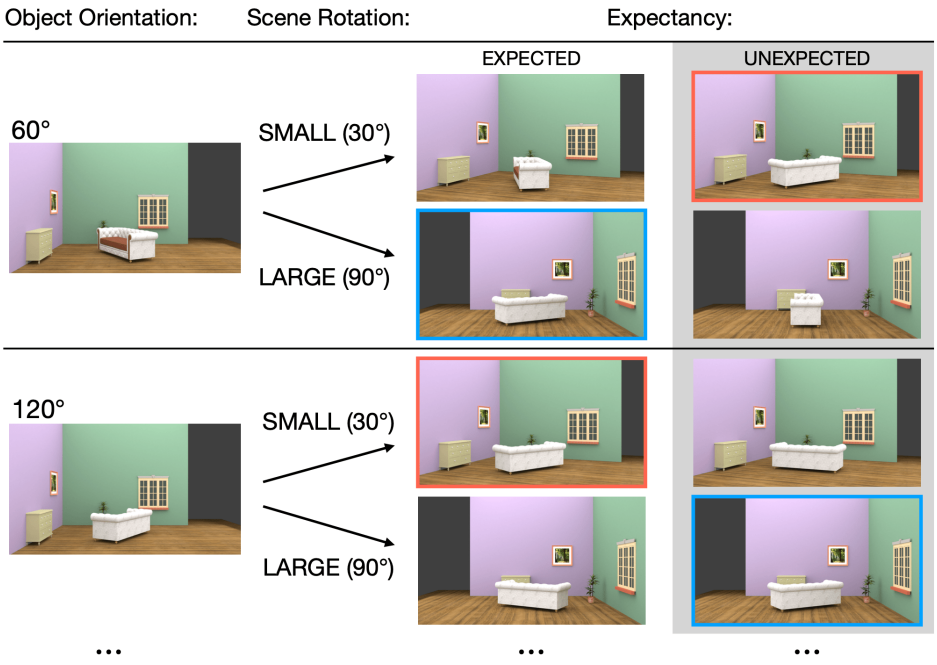


Figure 2.2. Illustration of the experimental design, showing the initial orientation of the object relative to the scene, and the final images (after the whole view sequence and the occlusion period) resulting from a Small or Large rotation on Expected or Unexpected trials. The images highlighted by the colored frames are examples of the same images appearing as either Expected or Unexpected on different trials.

2.2.6 Post-experiment survey

After completing the experiment, participants were asked three questions for us to gauge their awareness of the expectation manipulation.
The questions were:

- "Your task was only on the final image, when the object changed or not. Did you also pay attention to the sequence of images before the task image?" - the response had to be indicated on a Likert scale from 1 (Not at all) to 7 (All the time).
- "When the scene rotated, did you anticipate seeing the object in the correct viewpoint after it reappeared?" - the response also had to be indicated on a 1-7 Likert scale.
- "What percentage of objects were in line with your expectation? (They reappeared with the correct viewpoint)" - the response had to be a value in percentage, from 0 to 100%.

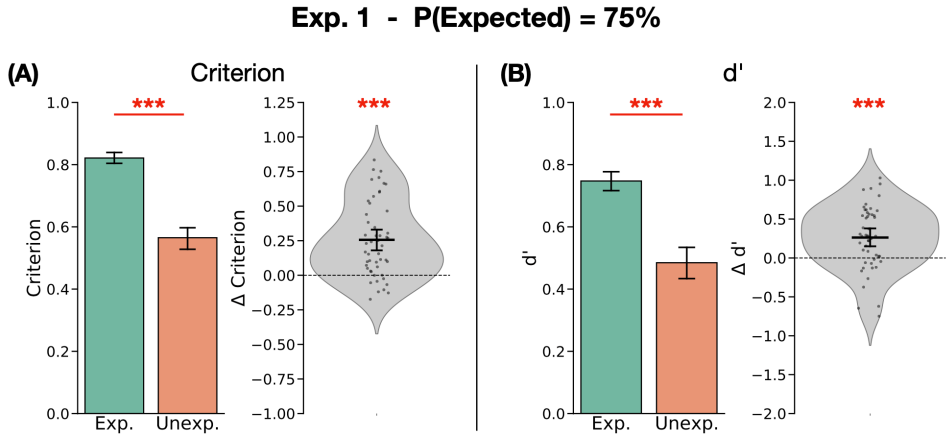


Figure 2.3. Results of Experiment 1. **(A)** Left – mean (and SEM) criterion for the Expected and Unexpected trials. Right – distribution of the differences between conditions (Expected – Unexpected) for each participant. **(B)** Same as in **A**, for d'. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

2.3 Results

2.3.1 Experiment 1: 75% Probability

In the first experiment, participants had a 75% probability of seeing the object in the expected view (given the scene viewpoint). Across conditions, their mean accuracy (and SEM) was 0.68 ± 0.01, indicating that they were fully able to do the task, and that the staircase successfully converged to the desired accuracy of 70%. Criterion overall was significantly above zero (mean: 0.69, $t(49) = 37.3$, $p < 0.001$, $d = 5.27$, 95% CI = [0.66, 0.73]), indicating a strong general bias towards responding 'same', possibly due to the small perceptual differences between the probes.

In our central analysis, we compared d' and criterion between the Expected and Unexpected trials. We found both measures to significantly differ. Criterion was higher on Expected than Unexpected trials (means: 0.82 vs. 0.56; $t(49) = 6.63$, $p < 0.001$, $d = 1.35$, 95% CI = [0.18, 0.33], **Figure 2.3A**). Participants, then, had a tendency to respond 'different' more often on Unexpected trials, reducing their overall bias. This result indicates that participants were sensitive to the object's congruence with the scene viewpoint, and that this influenced their responses. Interestingly, despite having less response bias on Unexpected trials, their sensitivity was also lower. Comparing d', we found it to be higher on Expected than Unexpected trials (means: 0.75 vs. 0.48; $t(49) = 4.53$, $p < 0.001$, $d = 0.89$, 95% CI = [0.15, 0.38], **Figure 2.3B**). The object's congruence, then,

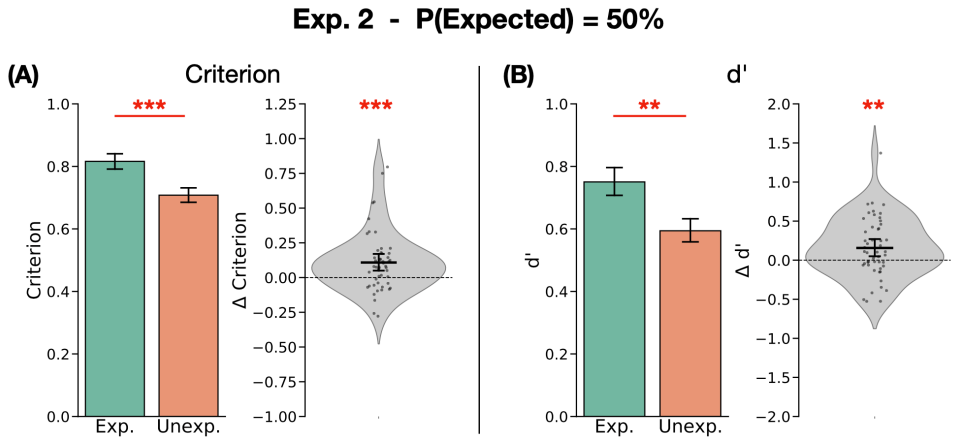


Figure 2.4. Results of Experiment 2. **(A)** Left – mean (and SEM) criterion for the Expected and Unexpected trials. Right – distribution of the differences between conditions (Expected – Unexpected) for each participant. **(B)** Same as in **A**, for d'. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

beyond influencing participants' responses, also had an effect on their perceptual sensitivity. Together, these effects on bias and sensitivity suggest that participants formed an expectation of the object's view given the scene context, and that whether this expectation was respected or violated influenced their performance in our orthogonal task.

In this experiment, the object matched participants' scene-driven expectations on a majority of trials. Real-world regularities (the coherence of an object's rotation with the surrounding scene), then, matched the short-term regularities observed during the experiment. In the next experiment, we investigated whether more frequent violations of real-world regularities would reduce this behavioral effect.

2.3.2 Experiment 2: 50% Probability

In Experiment 2, the object, after the occlusion, reappeared in an expected or unexpected view with equal probability. Besides this probability manipulation, stimuli and experimental paradigm were the same as in Experiment 1. Like in the previous experiment, participants were solidly above chance in performing the task (mean accuracy and SEM: 0.69 ± 0.01). Their overall criterion was also consistent with the previous experiment, being significantly higher than zero (mean: 0.76, $t(49) = 41.2$, $p < 0.001$, $d = 5.82$, 95% CI = [0.72, 0.80]). They were then still prone to respond 'same' (no difference between the two probes) on a majority of trials.

Again, in our main analysis, we found both criterion and d' to significantly

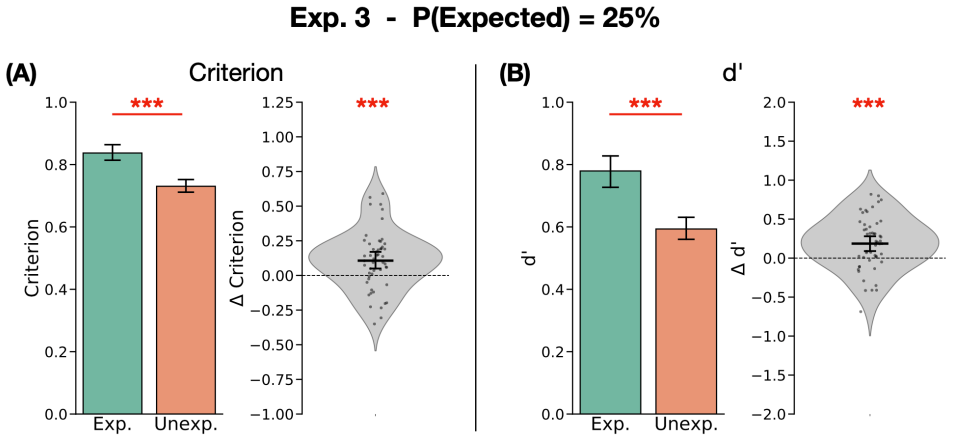


Figure 2.5. Results of Experiment 3. **(A)** Left – mean (and SEM) criterion for the Expected and Unexpected trials. Right – distribution of the differences between conditions (Expected – Unexpected) for each participant. **(B)** Same as in **A**, for d'.
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

differ between Expected and Unexpected trials. Criterion was higher in the Expected than the Unexpected condition (means: 0.82 vs. 0.71, $t(49) = 3.51$, $p < 0.001$, $d = 0.63$, 95% CI = [0.05, 0.17]; **Figure 2.4A**), and so was d' (means: 0.75 vs. 0.59, $t(49) = 2.89$, $p < 0.01$, $d = 0.54$, 95% CI = [0.05, 0.27]; **Figure 2.4B**). The behavioral effect of the object's expectancy on both bias and sensitivity was thus consistent with the previous experiment. This suggests that even when the long-term expectation of scene and object rotating coherently was not informative of the stimuli that would be shown in the experiment, it still affected participants' behavioral performance. It could thus not be easily overruled by short-term experiment regularities. In Experiment 3, we asked whether presenting unexpected object views on a *majority* of trials could overrule, and possibly even reverse, these behavioral effects.

2.3.3 Experiment 3: 25% Probability

In this experiment, the object would reappear with an expected view only on 25% of trials. Aside from this, stimuli and paradigm were the same as in the previous two experiments. Here, again, participants were well above chance (mean accuracy and SEM: 0.69 ± 0.01). Their overall bias was also consistent with the previous two experiments, with a majority of 'same' responses, leading to a significantly positive criterion (mean: 0.78, $t(51) = 46.4$, $p < 0.001$, $d = 6.44$, 95% CI = [0.75, 0.82]).

In our central comparison of criterion and d' between Expected and Unexpected trials, we again found a significant difference in both measures.

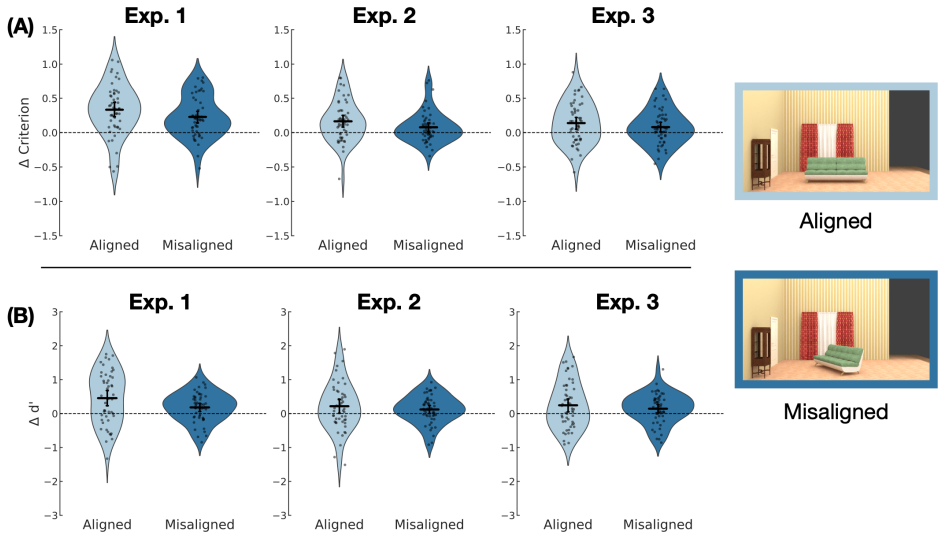


Figure 2.6. Distributions of Expected-Unexpected differences in criterion (**A**) and d' (**B**) for aligned and misaligned object orientations, for each of the three experiments.

Criterion was higher in the Expected than the Unexpected condition (means: 0.84 vs. 0.73, $t(51) = 3.59$, $p < 0.001$, $d = 0.66$, 95% CI = [0.05, 0.17]; **Figure 2.5A**), consistently with the previous experiments.
The higher d' in the Expected vs. Unexpected condition (means: 0.78 vs. 0.59, $t(51) = 3.92$, $p < 0.001$, $d = 0.58$, 95% CI = [0.09, 0.28]; **Figure 2.5B**) was also replicated in this experiment. Interestingly, then, the behavioral influence of expectancy did not reverse when the short-term experimental regularities ran counter to it. This result provides further confirmation that the real-world constraint of coherent scene-object rotation cannot be easily overruled by inconsistent evidence.

2.3.4 Role of alignment to the scene

Our object orientations could be divided into those that were aligned with one of the scene's main axes (0°, 180°) and those that were not (60°, 120°, 240°, 300°). We conducted an exploratory analysis to determine whether the behavioral effects we observed were influenced by the object's alignment to the scene's axes. If the effect were only found with aligned object orientations, this would suggest that participants' expectations of the object's view were primarily based on scene-centered cues. For each of the three experiments, we compared the magnitude of the behavioral effect (difference in d' or criterion between Expected and Unexpected trials) between aligned and misaligned object orientations. We did

	Experiment 1 **Probability = 75%**	**Experiment 2** **Probability = 50%**	**Experiment 3** **Probability = 25%**
Attention to Sequence 1-7 Likert scale	4.34 ± 0.21	3.90 ± 0.22	4.00 ± 0.23
Correlation with criterion	r = 0.12, p = 0.40, BF_{01} = 4.03	r = 0.26, p = 0.07, BF_{01} = 1.17	r = -0.07, p = 0.62, BF_{01} = 5.13
Correlation with d'	r = 0.16, p = 0.25, BF_{01} = 2.99	r = 0.11, p = 0.42, BF_{01} = 4.15	r = -0.06, p = 0.69, BF_{01} = 5.38
Object Prediction 1-7 Likert scale	3.90 ± 0.19	4.02 ± 0.20	3.88 ± 0.17
Correlation with criterion	r = 0.00, p = 0.98, BF_{01} = 5.68	r = -0.12, p = 0.39, BF_{01} = 3.97	r = -0.02, p = 0.90, BF_{01} = 5.75
Correlation with d'	r = 0.11, p = 0.45, BF_{01} = 4.29	r = -0.15, p = 0.29, BF_{01} = 3.27	r = 0.00, p = 0.98, BF_{01} = 5.78
Probability Estimate Percentage	60.48 ± 2.46	54.80 ± 2.48	50.85 ± 2.28
Correlation with criterion	r = -0.20, p = 0.17, BF_{01} = 2.32	r = -0.10, p = 0.48, BF_{01} = 4.44	r = -0.22, p = 0.12, BF_{01} = 1.74
Correlation with d'	r = 0.00, p = 0.99, BF_{01} = 5.68	r = -0.23, p = 0.11, BF_{01} = 1.66	r = 0.09, p = 0.53, BF_{01} = 4.81

Table 2.1. Mean responses (and SEM) to our final survey questions, and Pearson's r correlation with the behavioral effects (Expected – Unexpected trials) for both criterion and d'.

not find any significant effect of alignment on the magnitude of the behavioral differences, in any of the three experiments, in either criterion ($t(49) = 2.08, 2.16, t(51) = 1.22$; $p_{bonf} = 0.127, 0.108, 0.687$ for the three tests; **Figure 2.6A**) or d' ($t(49) = 2.31, 0.84, t(51) = 0.81$; $p_{bonf} = 0.074, 0.812, 0.812$ for the three tests; **Figure 2.6B**). While this analysis was purely exploratory, and based on a small number of trials (as few as 12 in some conditions), we did not find any evidence of the distinction between expected and unexpected object views being based on scene-centered cues. We discuss the possible interpretation of this in the **Discussion**.

2.3.5 Final survey data

The purpose of the final survey questions was to gauge the extent to which participants were aware of the experimental manipulation: how much they paid attention to the sequence of scene viewpoints before the target object appeared,

how much they actively tried to predict the final object viewpoint, and their estimate of the probability of the object appearing in the expected orientation. **Table 2.1** reports participants' mean responses for each of the questions, together with their correlation (Pearson's r) with the magnitude of the criterion and d' difference in our task (Expected-Unexpected) across participants. We found that none of the survey questions correlated with our behavioral effect size, suggesting that the accuracy difference was not driven by an explicit prediction strategy, by the amount of attention deliberately paid to either the scene sequence or the frequency of the stimulus matching their expectations.

Comparing the Likert ratings across the three experiments, we found that neither *Attention to Sequence* nor *Object Prediction* changed significantly depending on the probability of the object appearing in the Expected view (respectively: $F(2, 99.2) = 1.169$, $p = 0.315$, $\eta^2 = 0.014$; $F(2, 98.7) = 0.151$, $p = 0.860$, $\eta^2 = 0.002$; Welch ANOVAs). Their *Probability Estimate*, on the other hand, significantly differed across experiments ($F(2, 99.0) = 4.117$, $p = 0.019$, $\eta^2 = 0.052$), showing that participants could detect a difference in the probability of the object matching their expectation. While they were somewhat aware of the probability manipulation, then, according to their self-reports they did not seem to adopt a different strategy (e.g. paying more attention to the scene, or more actively trying to predict the object) based on the probability of the object matching their expectations. More importantly, our behavioral effects correlated with none of these self-report measures, suggesting that they did not depend on conscious adoption of one of these strategies.

2.4 Discussion

In the real world, objects and their context are strongly interdependent, meaning that it is possible to predict how an object's viewpoint will change, based on changes in the viewpoint of the surrounding scene. In this study, we manipulated whether objects respected this constraint, and measured how this affected participants' performance in an orthogonal perceptual task. Across three experiments, we found that expectancy affected both participants' sensitivity and response bias, suggesting that they formed an expectation of the object's view. This was the case even though the task did not explicitly require them any explicit prediction, or even to take scene information into account. Strikingly, the effect was still present even when the real-world constraint was not predictive during the experiment (in Experiment 2) or counter-predictive (Experiment 3). The effect we reported, then, likely did not arise from a form of associative learning between arbitrary object views occurring during the experiment, unlike in paradigms investigating the effect of probabilistic expectations on perception (Kok, Jehee, et al., 2012; Kok & Turk-Browne, 2018). Moreover, both the amount of overall scene rotation and the orientation of the object relative to the scene were varied across trials, suggesting a flexible

mechanism allowing to predict an object from novel viewpoints given any initial orientation, and adjusting this prediction to the amount of rotation in the scene.

These results provide a potential bridge between the known human ability to transform internal representations of objects, such as rotating them to novel viewpoints, and the requirements of real-world perception. Prediction in the real world generally involves filling-in missing information from a context, rather than imagining something on a 'mental canvas'. A clear example is the case of occlusion: even as objects go in and out of sight, we need to track and update their representations (Munton, 2022; Scholl & Pylyshyn, 1999; Teichmann et al., 2021). Mental transformations such as rotation, by contrast, have generally been studied in isolation, using tasks that required to manipulate a separate mental image of an object, and compare it with a target stimulus in a series of slow, deliberate cognitive steps (Shepard & Metzler, 1971; Just & Carpenter, 1976; Larsen, 2014; Xue et al., 2017). A question left open by the present study is whether the representations involved in our paradigm are the same as in classic mental transformation paradigms. Reviewing several related findings, Graf (2006) highlights the importance of spatial object transformations, such as rotation and scaling, in a variety of tasks beyond those requiring explicit imagery. However, he proposes that these tasks might fall into two distinct classes. On the one hand, there is the deliberate manipulation of mental images, a process that happens slowly, and results in a prediction of *a specific object* in a specific orientation or size (Cooper & Shepard, 1973; Koriat & Norman, 1984, 1988; Stewart et al., 2022). On the other, there is a faster process of establishing a spatial reference frame for object perception. For example, a particular orientation or size can be 'activated' and *any* object that matches it can be recognized more efficiently (Graf et al., 2005; Larsen & Bundesen, 1978). These reference frames can also be established by scene context (Humphrey & Jolicoeur, 1993; Christou et al., 2003), suggesting that the effect we observed here might also involve setting up an abstract reference frame, rather than forming a mental image of a specific object from a novel viewpoint. Interestingly, studies investigating the tracking of moving objects under occlusion have found that only the object's position, and not its surface properties such as shape or color, are represented behind the occluder (Flombaum et al., 2009; Pylyshyn, 2004; Scholl & Pylyshyn, 1999; Teichmann et al., 2022). This suggests an interesting parallel with our paradigm, raising the possibility that abstracted representations of spatial object transformations – beyond position, orientation as well – might support tracking in naturalistic environments. Further research should clarify whether the effect we report results from an image-like representation of the expected object, or an abstract reference frame. For example, by comparing expected and unexpected views for different object exemplars from the one seen at the start of the trial, it should be possible to determine whether participants form an object-specific prediction.

A related, but distinct, question for future research is how the structure of the scene is represented to support the prediction of the updated object view. The field of object perception has traditionally contrasted *structure-based* and *view-*

based models. According to structure-based models, objects are represented in terms of parts and their spatial relations (Biederman, 1987; Marr & Nishihara, 1978; Hummel, 2000; Erdogan & Jacobs, 2017; Ayzenberg & Behrmann, 2022), allowing to recognize objects across different viewpoints (Biederman & Gerhardstein, 1993). In view-based models, on the other hand, objects are represented as collection of image-like templates, with operations such as mental rotation or view interpolation allowing generalization to novel views (Bülthoff & Edelman, 1992; Tarr & Pinker, 1989; S. Ullman, 1998). In spatial cognition, at the level of whole scenes, a similar distinction has been drawn. Some evidence suggests that subjects, in tasks that require spatial reorientation or detection of changes across viewpoints, primarily rely on viewpoint-invariant cues, by representing objects' positions relative to each other (Rieser, 1989), to landmarks or boundaries in the environment (Galati et al., 2010; Julian et al., 2016; S. A. Lee, 2017), or to scene layout (Mou & McNamara, 2002). On the other hand, there is also substantial evidence for subjects primarily relying on self-centered views to orient themselves in space (Franz et al., 1998; Gillner & Mallot, 1998; Gootjes-Dreesbach et al., 2017; Vuong et al., 2019). Both in object recognition and spatial navigation, then, models based on structural descriptions, that remain invariant across viewpoints, have traditionally been contrasted with models based on image-like representations and mappings between them. Which kind of representation might underlie the scene-driven predictions reported here? To provide a tentative answer, we have compared our behavioral effects between trials in which the object was aligned with one of the cardinal axes in the scene with those in which it wasn't. This is a classic manipulation used to distinguish structure-based from view-based scene representations (e.g. Marchette et al., 2011; Marchette & Shelton, 2010; Mou & McNamara, 2002), since salient axes of the environment provide a stable reference frame that can be used across viewpoints. We did not find any difference in the magnitude of the effect, suggesting that participants might have relied more on a view-based representation. Clearly, this does not necessarily mean that they represent scenes exclusively as holistic, image-like representations. In both object perception (Edelman & Intrator, 2001; Foster & Gilson, 2002; Hayward, 2003; Hummel & Stankiewicz, 1998) and spatial navigation (Burgess, 2006; Burgess et al., 2004; Heywood-Everett et al., 2022), views and structural relations seem to be used in parallel. It is still possible, then, that scene-driven predictions of objects rely on structured representations of scenes in terms of different objects and their relations, even if those representations still rely on viewpoint-specific image features. Interestingly, a recent computational model (Bear et al., 2020) has provided a proof of concept that complex hierarchical scene representations can be constructed while remaining bound to image-centered features and spatial locations. Future research should clarify the nature of the scene representation driving object predictions in our paradigm, for example by examining systematic distortions in how distance relations are represented (e.g. Svarverud et al., 2012).

In conclusion, we have shown that participants create expectations of

objects from novel viewpoints automatically, driven exclusively by scene context. These expectations affect both their sensitivity and response bias in an orthogonal perceptual task. Moreover, they cannot easily be overruled by frequent violations, further confirming their automaticity. These results suggest that humans' mental transformation abilities might support perception in real-world scenes by automatically interacting with contextual information.

Chapter 3
Scene context drives object expectations across viewpoints in visual cortex

Abstract

As we change our viewpoint in a scene, the objects around us change coherently with each other, and with the layout of the scene. Scene context thus provides powerful cues to predict dynamic changes in object appearance. Known contextual effects on object perception, however, are limited to the disambiguation of fixed object properties, such as category. Here, we used a behavioral task and fMRI to assess whether participants formed expectations of 3D objects' appearance after a viewpoint change. Importantly, the viewpoint change could only be determined from the surrounding scene, allowing us to measure how object predictions can be driven by scene context. We found that participants' performance in an orthogonal visual task and object representations in early visual cortex are both enhanced when the object is rotated consistently with the scene. These results provide evidence that scene context, beyond disambiguating objects, can also drive predictions of object transformations.

3.1 Introduction

Human visual perception is able to handle the complexity of the real world by exploiting statistical regularities: an example is the way that objects are constrained by the surrounding scene. For instance, particular object categories are more likely to occur in some contexts, such as a boat on the sea or a car on a road: our visual system exploits this by disambiguating an object's identity based on its background (Bar, 2004; Oliva & Torralba, 2007; Brandman & Peelen, 2017; Rossel et al., 2022). Or the same retinal size can correspond to a large object far away or a small object up close, which is reflected in the object's perceived size being affected by visual depth cues (Leibowitz et al., 1969; Murray et al., 2006; Yildiz et al., 2021).

The effect of scene context on inferring these fixed object properties can be seen as a form of convergence towards a single most likely explanation. However, we live in a highly dynamic world, in which relevant properties of objects are continuously changing. For example as we move, objects' appearance keeps changing with our viewpoint. These dynamic changes are also strongly constrained by context: objects will tend to move together, coherently with the overall layout of a scene. Whether scene context automatically informs dynamic perceptual predictions is still unknown.

Here, we set out to investigate whether scene context can drive the prediction of objects from novel viewpoints. We reasoned that if these effects of scene context happen automatically, they should share some of the cognitive and neural mechanisms involved in other forms of perceptual expectations. Expectations deriving from regularities in the environment are believed to modulate perception independently of voluntary or attentional processes (Summerfield & Egner, 2009; De Lange et al., 2018). In behavior, this leads to expected stimuli being discriminated better, regardless of their relevance to the task (Wyart et al., 2012; Cheadle et al., 2015). Neuroimaging studies, on the other hand, have found representations of expected stimuli to be *sharpened* in visual cortex: decoding of stimulus information is enhanced, while the overall amount of cortical activation is reduced (e.g. Kok, Jehee, et al., 2012; Yon et al., 2018), suggesting a more efficient code.

We used fMRI and behavioral measurements to investigate (1) whether contextual information in realistic 3D scenes can drive expectations of an object's appearance from a new viewpoint, and (2) whether these expectations can lead to the sharpening of stimulus information in visual cortex. We designed a paradigm (**Figure 3.1**) in which an object was shown in the context of a realistic scene, which changed in viewpoint (rotated). During the viewpoint change, the object was temporarily occluded, and when it reappeared, it could either be oriented consistently with the rotation of the surrounding scene (Expected trials) or inconsistently (i.e., more or less rotation than the scene; Unexpected trials). Importantly, because expectations of object appearance depended on the correspondence between the amount of rotation of the object relative to the

amount of rotation of the scene, Expected and Unexpected trials could not be differentiated based on the final scene alone. Both the initial orientation of the object and the amount of scene rotation had to be taken into account. The object orientations at the start of a trial and the rotation angles were chosen such that the objects always reappeared in a 'wide' (i.e., front or back) view or 'narrow' (i.e., side) view, thus generating differentiable visually evoked responses. Participants were not given any instruction about the viewpoint changes or the scenes; they were only instructed to perform an orthogonal perceptual task on the object upon its reappearance (see **Figure 3.1**, bottom). To test whether observers generate predictions of object appearance across viewpoint changes, we compared the Expected and Unexpected conditions on (1) participants' performance on the perceptual task, and on (2) the amount of information about the shape of the reappearing object (i.e., wide versus narrow) in visually evoked activity. We focused on two regions of interest (ROIs) in visual cortex: early visual cortex (EVC) – corresponding to areas V1 and V2, and object selective cortex (OSC) – corresponding to the lateral occipital complex. We found that Expected trials were associated with (1) an increase in participants' accuracy on the task; (2) a sharper representation of object shape in EVC (but not OSC); and (3) a smaller whole-brain univariate response. Together, these results provide evidence that scene context can inform predictions of object appearance across viewpoints, that these predictions seem to happen automatically rather than as the result of deliberate cognitive operations, and that they lead to increased precision in behavioral stimulus discrimination and to sharpened representations in visual cortex.

3.2 Methods

3.2.1 Participants

Participants were recruited through the Radboud University participant pool (SONA systems) and received a monetary reimbursement for their participation. They provided informed consent before the experimental session. The study was in accordance with the institutional guidelines of the local ethical committee (CMO region Arnhem-Nijmegen, The Netherlands, Protocol CMO2014/288).
A total of 35 participants took part in the study, and one was excluded due to chance-level performance in the main behavioral task, leaving a sample of 34 participants to be analyzed (20 females, mean age = 24.2, SD = 4.4). The predetermined sample size of 34 was chosen to achieve 80% power for detecting a medium-sized effect.

3.2.2 Apparatus

Participants viewed the stimuli through a mirror mounted on the head coil of the scanner. Stimuli were presented on a 32-inch BOLDscreen monitor (Cambridge

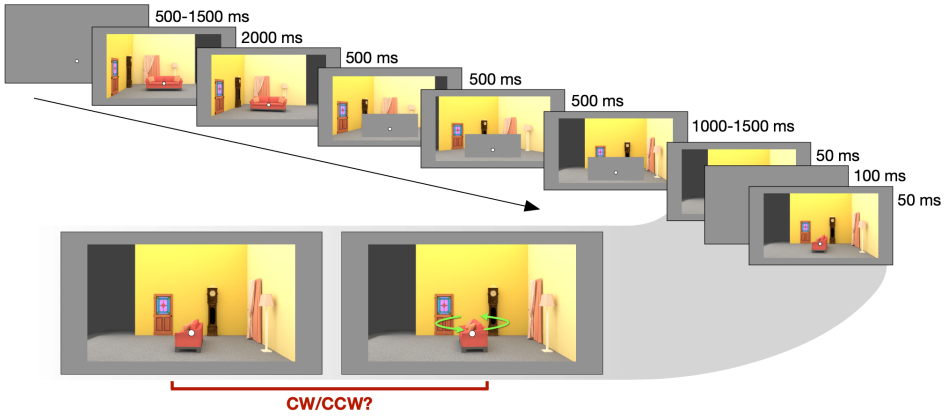


Figure 3.1. Example of the stimulus sequence within a trial. When the target object reappeared after occlusion (here in the Expected orientation), it was briefly flashed twice, and participants had to report whether the second presentation was rotated clockwise or counterclockwise relative to the first (in this example, the correct answer is counterclockwise).

Research) with 1920x1080 px resolution and 120 Hz refresh rate. The total viewing distance (eyes from mirror + mirror from screen) was 1206 mm. Stimuli were presented using Psychtoolbox (Brainard, 1997) in MATLAB R2017b. Participants provided responses on a HHSC-2x4-C button box.

3.2.3 Procedure

Before the scanning session, participants performed a short training session (40 trials, around 10 minutes duration) to familiarize with the main experimental task. During this session, they received feedback on every trial, as well as seeing their overall accuracy at the end of the session. After the training, they were also instructed about the other task they would have to perform in the scanner (1-back task in the Training and Functional Localizer runs). During the five-minute anatomical scan, they practiced the main task again, also with trial-by-trial feedback. Participants were in the scanner for a total of 12 functional runs. Each functional run began and ended with 15 seconds of fixation.

3.2.4 Experimental design & stimuli: Main task runs

In the main task (**Figure 3.1**), participants saw realistic scenes featuring a central object. The scenes underwent a change in viewpoint, during which the central object was occluded. The object then reappeared, and participants had to perform

a perceptual task on it. On each trial, a fixation dot was shown for 500 ms, followed by the initial view of the scene for 2000 ms. The scene then started rotating, in 3 intermediate views, each shown for 500 ms. The object was fully occluded starting from the second of these intermediate views. The final view of the scene, with the object still occluded, was displayed for a randomly jittered time between 1500 and 2000 ms. The object then reappeared, and was briefly flashed twice (with the scene background always present) for 50 ms each, with a 100 ms inter-stimulus interval in between.

We refer to these two brief presentations of the object as the *probes*. On a given trial, the second probe was rotated clockwise or counterclockwise, with equal probability, relative to the first, and participants' task was to report 'clockwise' or 'counterclockwise' using the index or middle finger of their right hand, respectively. Participants had a maximum of 1500 ms to respond, after which the experiment would skip to the next trial and the current trial would be counted as missed. The duration of the initial fixation for the next trial was adjusted to compensate for participants' response time on the current trial, to ensure that the overall duration of each run was constant.

Our central experimental manipulation was that, on 75% of trials, the object reappeared in the orientation that was expected if it had rotated consistently with the scene, and in an unexpected orientation (more or less rotation than the scene) on the other 25%. We call these the *Expected* and *Unexpected* trials respectively. Crucially, participants' task was completely orthogonal to this contextual manipulation: they did not have to explicitly judge whether the object remained in the same orientation relative to the beginning of the trial, or to predict its upcoming view after the occlusion period. They were instructed that their task would be on the final probes exclusively, but to remain attentive during the whole sequence. Additionally, the total amount of viewpoint change of the scene was varied on a trial-by-trial basis: it could be a small (30°) or a large (90°) rotation with equal probability. The purpose of this was to ensure that whether the reappeared object's view was expected or unexpected depended on the orientation of the background on the final frame, and not just on the initial orientation of the object, which would have been the case had the amount of rotation remained constant across trials. The central object in each scene could be oriented in two possible ways, A or B, aligning with the main axes of the scene: in A the object's long axis was orthogonal to the observer's line of view, while in B it was parallel to it (Figure **3.2B**, left). For each of the two initial orientations, then, the object could appear, in the final frame, with either a wide shape (A - small rotation or B - large rotation) or a narrow shape (B - small rotation or A - large rotation). This meant that on half of trials, the object's proximal shape was wide, and narrow on the other half, enabling us to use decoding of object shape as a measure of information about the stimulus' visual appearance in the brain. When the object was initially shown in orientation A, on Unexpected trials the corresponding rotation from orientation B was shown, and vice versa (Figure **3.2B**, right). This meant that the same images could be presented as either

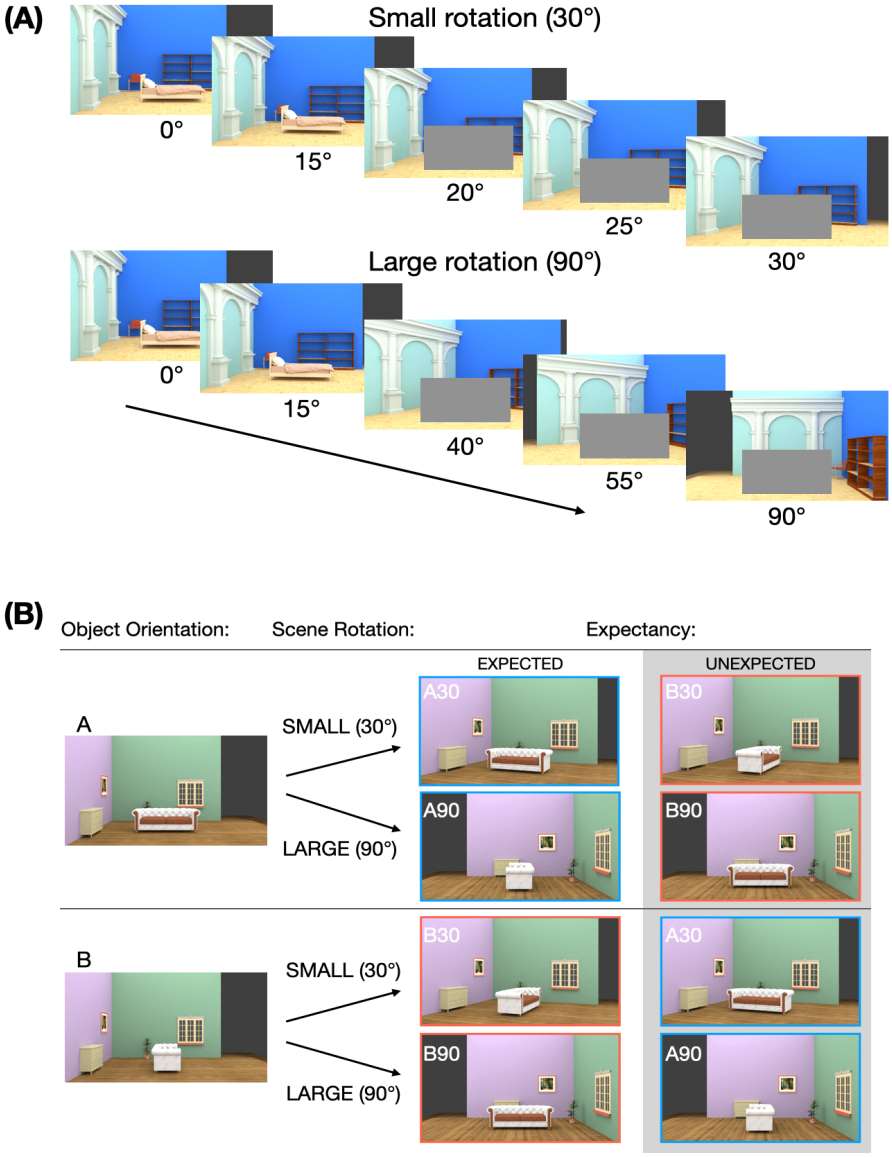


Figure 3.2. (A) Examples of 30° and 90° rotation sequences, in this case for an object in the orientation A (orthogonal to the line of view). **(B)** Overview of the 4 possible views that the object could reappear in, corresponding to combinations of initial viewpoint (A/B) and amount of rotation (30/90). On Unexpected trials, the object view corresponding to the same amount of rotation but the other object orientation was shown (e.g. A30 was shown when B30 was expected).

Expected or Unexpected depending on the context of the trial, avoiding the possibility that any of our results could be driven by visual stimulus differences.

In total, each participant completed 7 runs of the main task, each consisting of 48 trials (336 trials total). Within each run, 36 trials were Expected and 12 were Unexpected. Both Expected and Unexpected trials were equally divided among the 4 possible initial orientation/amount of rotation combinations (A30, A90, B30, B90).

The stimuli for the main task and training runs were 20 different indoor scenes modeled in Blender 2.80 and rendered using the Cycles rendering engine for realistic lighting. The scenes all had the same layout (floor, two walls at a right angle and a main object in the center) but contained various other objects, adjacent to the walls, and different textures on the walls and floor. The central object was a couch on half of the scenes, and a bed on the other half. This object's size was the same across scenes. For each scene, a range of viewpoints was rendered, by rotating the entire scene around the vertical axis between 0° and 90°, in steps of 5°. A subset of these viewpoints was presented on each trial. The two walls were oriented such that the scene was fully visible from all the viewpoints. The scenes were presented at the center of the screen at a size of 20.53 x 11.64 degrees of visual angle (dva). The occluder was a grey rectangle which had the height and width of the largest possible view of the object on that particular scene (average size: 5.50 x 2.86 dva), plus a margin (horizontal: 1.08 dva, vertical: 0.43 dva) to ensure the object was fully covered and its shadow was not visible, which would have provided a cue to its orientation. The fixation dot (size 0.2 dva, shown at the location of the central object, 3.24 dva below the center of the screen) was present throughout the whole image sequence, and participants were instructed to maintain fixation.

For the discrimination task, the first probe's orientation was randomly sampled from a normal distribution centered around the Expected or Unexpected orientation, with a standard deviation of 1°, to add a small amount of jitter, and then rounded to the nearest integer. The second probe was rotated, clockwise or counterclockwise, relative to the first by an angle that was titrated using a 2-down 1-up staircase, to keep the task difficulty constant across participants. To ensure that the visual stimuli in the Expected and Unexpected trials did not differ, a single staircase was used across both Expectancy conditions, thus allowing for accuracy differences between conditions. Unlike in the training session, participants did not receive feedback on every trial, to avoid possible effects of differing feedback between conditions (Expected and Unexpected trials) on the fMRI activity. Instead, their overall accuracy within a run was displayed at the end of the run.

3.2.5 Experimental design & stimuli: Training runs

The training runs had the purpose of estimating prototypical response patterns to the stimuli in our main task, without the context of the whole rotation sequence. The images displayed in the training runs were the final frames of the sequences

in the main task, for the 4 possible object orientation/scene rotation combinations (A30, A90, B30, B90 - see **Figure 3.3**). They were presented in mini-blocks, each consisting of 18 stimuli (different scene exemplars, all in the same orientation/rotation combination), with each stimulus presented for 350 ms and followed by a 400 ms blank interval (13.5 s in total). After a series of 4 different mini-blocks (54 s in total) a longer blank interval was shown for 6.75 s. The fixation dot was present throughout, and all stimuli were presented at the exact same location and size as in the main task runs. Participants' task was to press any button whenever the exact same image was repeated twice in a row (1-back task). Each run included 20 mini-blocks, and 3 training runs were presented in total.

3.2.6 Experimental design & stimuli: Functional localizer runs

The purpose of the functional localizer was to identify object-selective ROIs for each participant. The stimuli were images from 4 different image categories: Objects, Scrambled objects, Faces and Scenes (houses or landscapes). The stimuli used were the same as in Epstein & Kanwisher (1998), presented against a uniform gray background at a size of 12 x 12 dva. Stimuli were divided in mini-blocks, each lasting 15 s and comprising 20 unique images from a particular category, presented for 450 ms followed by a 300 ms blank. Each localizer run included 16 mini-blocks (4 repetitions of 4 stimulus types) and the experiment comprised 2 localizer runs in total.

3.2.7 fMRI data acquisition and preprocessing

Images were acquired on a 3T MAGNETOM Skyra MR scanner (Siemens AG, Healthcare Sector, Erlangen, Germany) using a 32-channel head coil. Functional data was acquired using a T2*-weighted gradient EPI sequence, with 6x multiband acceleration factor (TR 1s, TE 35.2 ms, flip angle 60°, 2x2x2 mm isotropic voxels, 66 slices). For the main task runs, 404 images were acquired per run, 333 and 318 images for the training and functional localizer runs, respectively. At the start of the scanning session, a high-resolution T1-weighted anatomical scan was acquired using an MPRAGE sequence (TR 2.3 s, TE 3.03 ms, flip angle 8°, 1x1x1 mm isotropic voxels, 192 sagittal slices, FOV 256 mm). The data was preprocessed using SPM12 (Penny et al., 2011) functions in Nipype 1.6.0 (Gorgolewski et al., 2011). The functional volumes were fieldmap-corrected, spatially realigned, co-registered with the anatomical image, normalized to MNI 152 space using the template provided in SPM, and smoothed with a 3x3x3 mm FWHM Gaussian filter.

3.2.8 GLM analysis

The responses evoked by each of the stimulus types relevant to our analysis were modelled using general linear models (GLMs) in SPM12, using Nipype 1.6.0 as

interface.

In the main task, the onsets of the final object views were modelled as impulse functions and the time series was convolved with the canonical HRF provided in SPM12. For the MVPA analysis, we included regressors for each combination of object viewpoint and final scene rotation (A30, A90, B30, B90), separately for the Expected and Unexpected trials. Since the Expected condition included 3 times as many trials as the Unexpected one, to control for the possible benefit of the larger amount of data in multivariate decoding, we randomly split the Expected trials within each run into three subsets of 12 trials each (the same number as the Unexpected trials). Considering each split of the Expected trials as a separate condition, then, we obtained a single beta weight map per condition per run. For the univariate analysis, we only included regressors for Expected and Unexpected trials, obtaining two beta weight maps per run.

In the training runs, individual mini-blocks were modeled as boxcars and the time series was convolved with the canonical HRF. Regressors were included for each of the initial object orientation/scene rotation combinations, yielding one beta weight map per condition per miniblock per run.

For functional localizer runs, we also used a block-based design, but estimated a single beta map per condition per run. Miniblocks were modeled as boxcars and convolved with the canonical HRF. To estimate the beta weights used to define the OSC ROI, we included regressors for each of the stimulus types (Faces, Objects, Scenes, and Scrambled), and to estimate the weights for the EVC ROI, we included regressors for stimulus (of any type) and baseline (fixation) only.

All GLMs included six motion parameters and one run-based regressor as nuisance regressors. As participants were performing a 1-back task in the training and localizer runs, these runs also included a nuisance regressor synchronized to participants' button presses.

3.2.9 Regions of interest

To select voxels to include in our visual cortex ROIs, we used subject-level t-contrast maps estimated using data from the functional localizers, contrasting stimulus (both objects and scrambled images) vs. fixation baseline for EVC, and intact objects vs. scrambled images for OSC. These maps were intersected with an anatomical mask corresponding to Brodmann areas 17 and 18 (corresponding to areas V1 and V2; Wohlschläger et al., 2005) for EVC, and a population-level functionally defined lateral occipital cortex mask, retrieved from Julian et al. (2012), for OSC. Each participant's map was then thresholded to only include voxels that significantly responded ($p_{uncorrected} < 0.05$) to the relevant stimuli for each ROI: general visual stimulation for EVC, and intact object pictures for OSC. To assess the robustness of our results to specific voxel inclusion criteria, the most active voxels were selected from the thresholded images, in a range from 100 to 3000 (in EVC) or 100 to 2000 (in OSC) selected voxels in steps of 100, creating 30 and 20 sub-ROIs for EVC and OSC respectively, with an increasingly liberal voxel

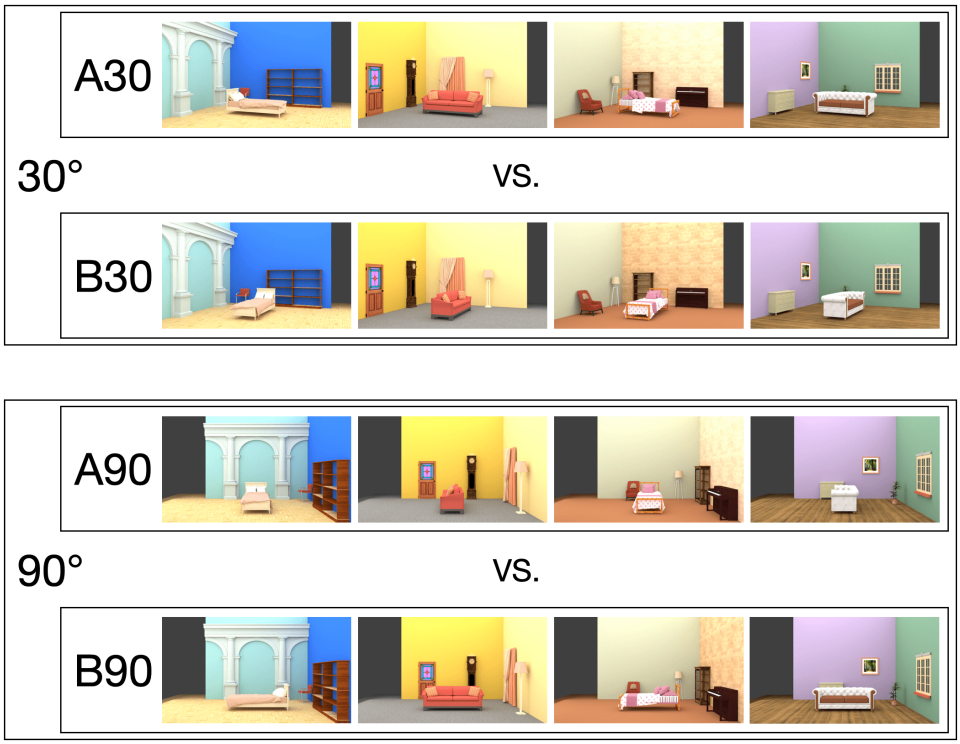


Figure 3.3. The conditions that the classifiers were trained to discriminate in the main task and training runs.

inclusion criterion.

3.2.10 Multivariate pattern analysis

The multivariate pattern analysis (MVPA) was conducted using linear support vector machines (SVMs) implemented in Scikit-learn (Pedregosa et al., 2011) and PyMVPA (Hanke et al., 2009). Our cross-classification analysis consisted of training the SVM classifiers on the miniblock-based beta weights from the training runs, and testing them on the run-based beta weights from the main task runs, and vice versa. Both directions of cross-classification were run, and the results were averaged, in order to increase robustness to task- or stimulus-unrelated factors that can lead to asymmetries between classification directions, such as different signal-to-noise ratios (van den Hurk & de Beeck, 2019). The training and testing datasets were separately z-scored before decoding. Each classifier was trained and tested on voxels within a single ROI and hemisphere, in a single subject.

To decode the stimulus feature of interest – proximal object shape (wide vs. narrow), we separately trained classifiers to discriminate between the A and B object orientations embedded in scenes rotated by 30 or 90 degrees (**Figure 3.3**), corresponding to classifying conditions A30 vs. B30, and A90 vs. B90, in such a way as to classify the object's shape against a matched background. The accuracies of classifiers trained on the two backgrounds were then averaged. Additionally, given that the number of Expected trials was 3 times larger than the number of Unexpected trials, in order to avoid a difference in decoding accuracy being driven by the number of trials (training set size), we modeled three random splits of Expected trials as separate regressors (see **GLM analysis**) and we tested (or trained, depending on decoding direction) separate linear classifiers on each of these three subsets. The accuracies from the three splits were also averaged. Importantly, the labels of the beta weights corresponding to Unexpected trials in the main task runs corresponded to the object orientation that was actually presented on the screen at the end of the trial, not the one expected given the context, as our goal was to assess how the same visual stimuli are processed differently depending on the context.

Discrete classification outcomes have been shown to be less reliable than continuous measures in estimating the distance between stimulus representations in the brain (Walther et al., 2016). For this reason, we used each sample's distance from the classification hyperplane estimated by the SVM (distance from bound), rather than classification accuracy, as our main measure of decoding performance. For each classifier, performance (classifier information) was measured as follows:

$$\text{Classifier Information} = \frac{1}{n}\sum_{i=1}^{n} d_i l_i$$

Where d_i's are the z-scored distances from bound, l_i's are the true labels (either -1 or 1), for each sample, and n is the number of test samples. Intuitively, this measure corresponds to the average match between each distance from bound and the corresponding ground-truth label, i.e. the degree to which the distance is positive when the target is positive, and negative when the target is negative. Classification is above chance when this measure is higher than 0. Averaging this measure across samples allows to compare classification performance with different numbers of test samples, enabling us to combine data across decoding directions. The classifier information was computed for each sub-ROI within EVC and OSC, and each subject. Importantly, our results were consistent, albeit noisier, when using classification accuracy instead of classifier information.

3.2.11 Significance testing

To compare the amount of classifier information between conditions (Expected and Unexpected), we used two approaches. (1) To assess the statistical

significance of differences between conditions across numbers of selected voxels, we applied threshold-free cluster enhancement (TFCE, Smith & Nichols, 2009) to the classifier information differences between conditions, across sub-ROIs. TFCE boosts the magnitude of a statistic based on its extent across neighboring samples (in this case, sub-ROIs with similar numbers of voxels), reflecting the assumption that any signal in the data should be smooth across consecutive datapoints. This measure is then compared with a null distribution generated by randomly shifting the signs of each participant's 1D map (classifier information difference across sub-ROIs). This null distribution has the same variance and autocorrelation as the original signal. The shuffling procedure was applied 10,000 times. TFCE was computed using the MNE toolbox (Gramfort et al., 2013). Statistical significance of the TFCE scores is shown in the figures. (2) To summarize results with a single statistic, we averaged the classifier information across sub-ROIs for each condition (Expected and Unexpected) and each subject, and ran a two sided paired-samples t-test between the two conditions. This is the main statistic reported in the text of the rest of the paper. These statistical tests were run using Pingouin (Vallat, 2018).

3.2.12 Univariate analysis

The purpose of the univariate analyses was to estimate the difference in overall response elicited by Expected and Unexpected trials. This was done within the main visual ROIs as well as across the whole brain. For the within-ROI analysis in visual cortex, we averaged the beta weights across voxels for each participant and each sub-ROI (number of selected voxels) in EVC and OSC. The averages across sub-ROIs in the Expected and Unexpected conditions were then compared using a two-sided paired t-test. For whole-brain analyses, we conducted a nonparametric, cluster-based analysis at the group level (one-sample t-test against 0 for the single subjects' T contrast maps) using SnPM13 (nisox.org/Software/SnPM13/), with a $p < 0.05$ I-corrected threshold and $p < 0.001$ cluster-forming threshold, for 5,000 permutations.

3.3 Results

3.3.1 Behavioral results

We first set out to test whether participants' accuracy differed between trials in which the object reappeared oriented consistently with the rotation of the background scene (Expected trials) and trials in which it was oriented inconsistently (Unexpected trials). We found that participants were more accurate on Expected than Unexpected trials: mean hit rates were 0.642 and 0.605 respectively, $t(33) = 2.83$, $p = 0.007$, $d = 0.69$, 95% CI = [0.01, 0.06] (**Figure 3.4**). This indicates that participants formed predictions of the updated object view, and

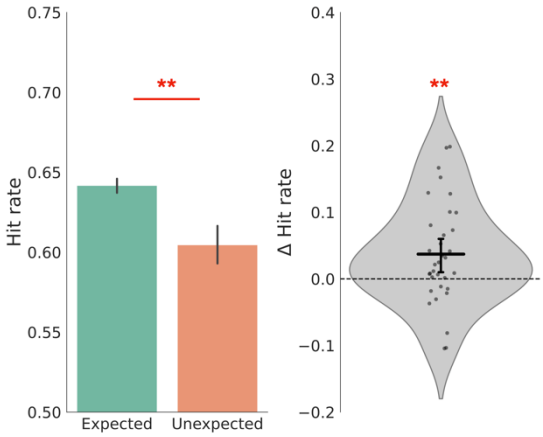


Figure 3.4. (left) Mean hit rate on Expected and Unexpected trials in our behavioral task. Error bars indicate SEM. (right) Hit rate differences between Expected and Unexpected trials. Points indicate individual subjects, horizontal bar indicates mean difference, and error bar 95% CI. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

were more accurate in performing an orthogonal task when the target matched this predicted object view. Given that the amount of scene (and thus object) rotation was varied from trial to trial, these predictions were necessarily derived from scene information, and could not have been generated by mentally rotating the object alone.

3.3.2 MVPA results

We next addressed the question of how the target object's shape was represented in multivariate activity patterns in visual cortex, and how these representations differed between Expected and Unexpected trials. For each sub-ROI (number of included voxels) in EVC and OSC, we assessed the cross-decoding performance of a classifier trained to distinguish between wide and narrow object shapes, separately for the Expected and Unexpected trials.

We found that decoding of object shape in EVC was significantly higher for Expected compared to Unexpected trials, mean classifier information: 0.344 and 0.272 respectively, $t(33) = 2.94$, $p = 0.006$, $d = 0.51$, 95% CI = [0.02, 0.12] (**Figure 3.5A**). This difference was significant in both hemispheres (Left: mean classifier information 0.367 vs. 0.306, $t(33) = 2.16$, $p = 0.038$, $d = 0.40$, 95% CI = [0.0, 0.12], Right: mean classifier information 0.321 vs. 0.239, $t(33) = 2.70$, $p = 0.011$, $d = 0.53$, 95% CI = [0.02, 0.14]). However, no such difference was present in OSC, mean classifier information: 0.129 and 0.149 respectively, $t(33) = -1.20$, $p = 0.237$, $d = 0.20$, 95% CI = [-0.05, 0.01] (**Figure 3.5B**). This result was also

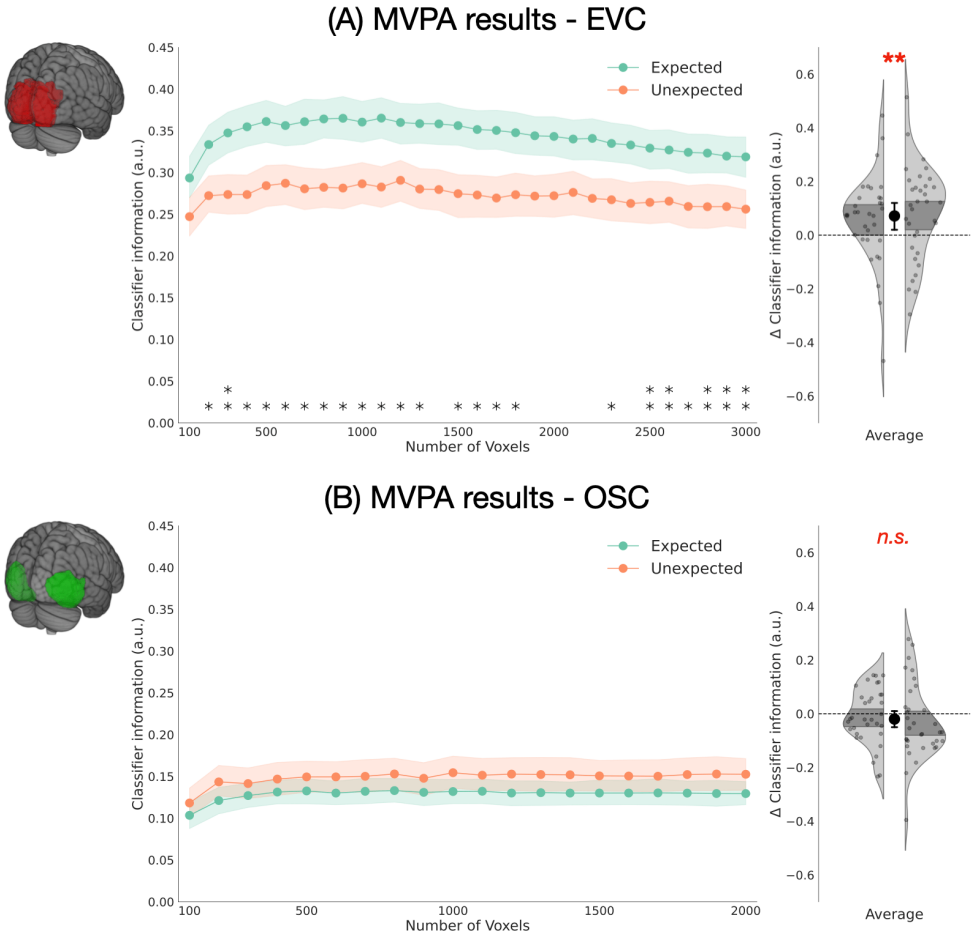


Figure 3.5. Results of multivariate classification in EVC (A) and OSC (B). (left) Average classifier information across sub-ROIs. Shaded areas represent SEM, asterisks represent TFCE significance. (right) distribution of differences across participants (averaged across sub-ROIs) for each hemisphere separately. Shaded areas represent within-hemisphere 95% CIs. For both left (TFCE) and right plots: * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

consistent across hemispheres (Left: mean classifier information 0.128 vs. 0.137, $t(33) = -0.52$, $p = 0.609$, $d = 0.09$, 95% CI = [-0.05, 0.03], Right: mean classifier information 0.129 vs. 0.160, $t(33) = -1.25$, $p = 0.219$, $d = 0.22$, 95% CI = [-0.08, 0.02]).

Given that our EVC ROI spanned both V1 and V2, we further investigated whether

Region	*p*(cluster)	*t* statistic	*k* voxels	Coordinates (mm) – x, y, z		
Left precuneus	0.003	6.08	1360	-6	-68	56
Right angular gyrus	< 0.001	5.70	2630	36	-72	42
Left inferior parietal lobule	0.003	5.65	1380	-24	-66	44
Right precuneus	0.003	5.51	1360	8	-70	54
Right precentral gyrus	0.005	4.67	1020	26	-4	50
Left inferior parietal lobule	0.005	4.66	1090	-34	-54	46

Table 3.1: Results of the whole-brain contrast for Unexpected > Expected.

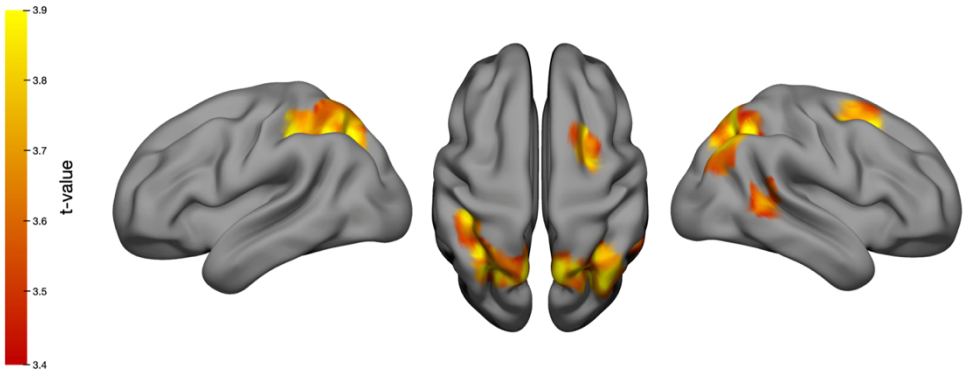


Figure 3.6. Significantly activated clusters ($p_{cFWE} < 0.05$, cluster-defining threshold $p_{uncorrected} < 0.001$) for the whole-brain univariate contrast, Unexpected > Expected.

our results were primarily driven by one or both visual areas. We found the difference between shape decoding on Expected and Unexpected trials to be significant in both ROIs. V1: mean classifier information 0.270 and 0.190 respectively, t(33) = 3.40, p = 0.002, d = 0.57, 95% CI = [0.03, 0.13]; V2: mean

classifier information 0.328 and 0.268 respectively, t(33) = 2.72, p = 0.010, d = 0.46, 95% CI = [0.02, 0.11]. A repeated measures ANOVA, ROI (V1, V2) x Expectancy (Expected, Unexpected) on mean classifier information revealed a significant main effect of Expectancy (F(1, 33) = 11.09, p = 0.002, $\eta_p^2 = 0.252$) and ROI (F(1,33) = 54.036, p < 0.001, $\eta_p^2 = 0.621$), but no significant interaction between ROI and Expectancy (F(1, 33) = 1.13, p = 0.295, $\eta_p^2 = 0.033$) indicating that the effect of Expectancy did not differ between V1 and V2.

In summary, we found information about the shape of objects otated consistently with the scene to be sharpened in EVC, but not in LOC. This effect was present in both V1 and V2, across both hemispheres. These results provide evidence for scene-driven expectations of object appearance across viewpoints in early visual areas.

3.3.3 Univariate results

The higher decoding accuracy we observed on Expected compared to Unexpected trials in EVC might be driven by a higher signal-to-noise ratio, deriving from higher univariate activation on Expected trials. This could occur, for example, if participants paid more attention to a stimulus matching their expectations. To determine whether this could have been the case, we compared average beta activations across all of the EVC sub-ROIs included in the multivariate analysis. We found no significant difference in overall activation between Expected and Unexpected conditions (mean activation: -3.20 and -3.12 for Expected and Unexpected respectively, t(33) = -0.75, p = 0.457, d = 0.02, 95% CI = [-0.28, 0.13]). In fact, numerically, the activation was slightly lower for Expected trials. This indicates that the higher multivariate decoding accuracy on Expected trials, described above, was not driven by an overall larger response on those trials.

We next ran a whole-brain univariate contrast to assess whether an overall increased response to either the Expected or Unexpected condition was present anywhere in the brain. We found no suprathreshold clusters responding more to the Expected trials. Instead, 6 clusters exhibited a significantly higher response to Unexpected trials, located in the posterior parietal cortex and right precentral gyrus (**Figure 3.6** and **Table 3.1**).

Overall, these results are in line with the view of expectation effects in visual cortex as sharpening: the increase in multivariate decoding performance for Expected trials was not associated with an overall increase in response in EVC. Moreover, across the whole brain, Unexpected trials elicited a stronger response than Expected, rather than the other way around, a pattern of results that has been proposed to distinguish expectation from attention-related effects (Kok, Rahnev, et al., 2012).

3.4 Discussion

The present results indicate that scene context, beyond influencing object perception in static settings, can also drive predictions of object appearance across viewpoint changes. These predictions seem to happen in an automatic manner, in the absence of an explicit task, and to sharpen the representation of object features in EVC.

In everyday life, we need to track objects across continuous changes in our viewpoint, and the highly structured nature of real-world scenes can aid us in doing so. By exploiting information from the surrounding scene, we can predict how objects will look from new viewpoints, while avoiding computationally expensive mental rotation operations (Hamrick & Griffiths, 2014). Using behavioral measurements and fMRI, we have found that objects rotated consistently with the surrounding scene show signatures of enhanced processing. These signatures were previously reported for other forms of perceptual expectations (Kok, Jehee, et al., 2012): (1) higher behavioral performance on orthogonal tasks; (2) enhanced multivariate decoding in the visual cortex; (3) absence of a corresponding increase in univariate activity. These shared characteristics suggest that similarly to those previously reported effects, the effect of scene context in our study reflects statistical regularities rather than explicit attentional guidance. In particular, attention is known to increase the overall response to a stimulus in visual cortex (Corbetta et al., 1990; Kastner et al., 1998). A study attempting to disentangle the effect of attention from that of expectations found that while expected stimuli elicit a reduced BOLD response in visual cortex, relative to unexpected ones, attending to the stimulus reverses this effect (Kok, Rahnev, et al., 2012). If participants in our study had adopted a deliberate attentional strategy, such as “predict the shape of the rotated object, and attend to the corresponding region of space”, we should have similarly observed a higher univariate response to stimuli matching their expectations. This was not the case in visual ROIs, and the whole-brain contrast revealed several clusters with a higher response to unexpected stimuli, but none with a higher response to expected ones.

The overlap of our results with previous reports also suggests that different forms of expectations (driven by arbitrary statistical associations, or by real world contextual regularities) rely on common mechanisms. This is consistent with the idea of common computations underlying different forms of perceptual expectations (De Lange et al., 2018; Keller & Mrsic-Flogel, 2018). For example, Kok et al. (2020) found a consistent involvement of visual cortex in expectations of different stimuli (complex shapes or oriented gratings), despite high-level areas such as the hippocampus undertaking different roles. And Ortiz-Tudela et al. (2021) found EVC to be the target of feedback conveying both contextual and mnemonic information in naturalistic scenes.

The effect of expectations we measured was specific to EVC, corresponding to visual areas V1 and V2. Several prior reports of contextual scene effects on object perception, instead, have reported effects in high-level object-

selective areas (Brandman & Peelen, 2017; Gayet & Peelen, 2022; Kaiser et al., 2021). This different pattern of results could be due to particular choices of stimulus or experimental design (Coutanche et al., 2016): for example, decoding coarse object shape, a relatively low-level feature, instead of higher-level properties such as object category. However, it is also possible that the dynamic nature of our task was fundamentally different from the static images used in those previous studies. Prior fMRI work investigating the tracking of objects across periods of occlusion (Erlikhman & Caplovitz, 2017) found that information about the feature that was being updated, object position, could be decoded from the occluded region in EVC. Object shape, a feature that remained constant throughout the trial, could instead be decoded from higher-level visual areas, consistently with previous studies of object maintenance during occlusion (C. Baker et al., 2001; Hulme & Zeki, 2007). EVC seems, then, to play a specific role in tracking object position across time. An intriguing question is whether this might generalize to tracking other object features as well. In our study, we observed involvement of EVC despite the object's location remaining constant. Instead, what changed was the object's orientation relative to the viewer, and its resulting proximal shape. Besides position, humans are also able to track objects based on other visual features (Blaser et al., 2000) but whether this tracking also relies on EVC is still unknown. Evidence coming from mental rotation paradigms (Albers et al., 2013; Christophel et al., 2015; Iamshchinina et al., 2021) suggests that this might be the case, but those studies used simpler synthetic stimuli. Whether EVC also plays a general role in dynamically tracking the properties of naturalistic objects, like those in our study, remains an open question. Roelfsema & de Lange (2016) have hypothesized that EVC might act as a 'cognitive blackboard' supporting any computation that requires a high-resolution spatial buffer. Generating expectations of upcoming object views might be one such operation.

While the present work focuses on investigating the *outcome* of expectations based on scene context, future work should clarify the *format* of the representations that make these expectations possible. One possibility is that the scene is represented as a structural description in 3D coordinates, and then translated back to retinotopic coordinates, leading to the 2D shape representations we observed in EVC. This kind of explicit coordinate transformation has been proposed to underlie spatial navigation and mental imagery (Byrne et al., 2007). The fact that we did not observe expectation effects on 2D shape decoding in higher-level OSC, in that case, could be due to expectations in high-level visual areas being represented in a 3D format. Alternatively, the scene might not be converted to an explicit structural description at all, and expectations might be based on a collection of views, perhaps related by transition probabilities (Franz et al., 1997; Glennerster, 2016). View-based representations have been shown to account for behavior in both object recognition (Bülthoff & Edelman, 1992; Tarr & Pinker, 1989; S. Ullman, 1998) and spatial orienting tasks (Gillner & Mallot, 1998; Gootjes-Dreesbach et al., 2017). Adopting similar methods to these prior works, future experiments could clarify

whether scene-based expectations exhibit inconsistencies that would not be expected if participants faithfully reconstructed the scene's 3D structure in their minds. For example, scenes with an altered 3D layout, but a similar appearance from the subject's viewpoint, might not be perceived as unexpected. Testing this will likely require designing more complex scenes that place a greater burden on participants' capacity to represent spatial relations.

Regardless of whether the representations that participants relied upon in our study are based on egocentric views or 3D structure, our results suggest that humans can represent scene-object relations in a rich enough way to support predictions across changes in viewpoint. This extends a long line of empirical and theoretical work investigating how our internal representations of objects reflect their properties in the external world (e.g. Craik, 1943; Shepard, 1984, 2001). This includes the ability to mentally rotate objects (Shepard & Metzler 1971) or to simulate their physical dynamics (Battaglia et al., 2013). It is possible that these internal representations also incorporate models of how objects interact with their context, for example the way they rotate concurrently with the surrounding scene. One way to efficiently process these kinds of spatial relations in complex scenes is to represent them in a hierarchical manner, linking scenes to the objects they contain, and objects to their parts. These kinds of hierarchical representations are extensively used in computer graphics (Cunningham & Bailey, 2001; Sowizral, 2000), and artificial intelligence research has addressed the problem of how they can be extracted from unstructured visual input (Sabour et al., 2017; Bear et al., 2020; Deng et al., 2020; Gklezakos & Rao, 2022; Hinton, 2021). Whether humans also internally represent scenes in a similar manner is still unknown, although some evidence exists that we process scenes hierarchically (Võ et al., 2019). Additionally, models based on graph-structured representations provide the best fit to human behavior on tasks such as predicting physical dynamics (Bear et al., 2021). Predicting novel object views based on scene context, as in the paradigm used here, might be another cognitive ability relying on structured scene representations.

In summary, we have found evidence for predictions of objects from new viewpoints, driven by scene context, affecting object representations in visual cortex. These results suggest that common mechanisms might underlie simple expectations learned in the lab and those resulting from the complex structure of the real world.

Chapter 4
Scene viewpoint drives the prediction of rotated objects under occlusion

Abstract

Humans have the ability to track and predict changes in external objects while they are temporarily occluded. How this ability can generalize to complex, real-world environments is still an open question. In the real world, scenes are highly structured, meaning that objects and their context tend to change coherently. A clear example can be seen when we navigate an environment: as we vary our viewpoint, we see objects rotating jointly with the scene's overall layout. Previous research has shown that our internal representations of scenes contain rich structural information about objects' relative positions and orientations. Here, we ask whether the prediction of occluded objects capitalizes on this information. We present objects within realistic 3D scenes that change in viewpoint. The object is shown in an initial view, and then occluded during the viewpoint change, while the scene remains visible. Using fMRI and multivariate pattern analysis (MVPA), we find that the object's updated appearance, consistent with the scene's new viewpoint, is represented in the visual cortex, despite the object being fully invisible. These predictions emerge in the absence of an explicit prediction or mental rotation task. Capitalizing on spatial relations might be a way in which the visual system is able to complete missing information in partially occluded real-world scenes.

4.1 Introduction

In daily life, objects continuously go in and out of sight. They are often occluded by other objects for extended periods of time, yet we are able to perceive a coherent, seamless visual scene. We are able to do this by maintaining persistent object representations (Kahneman et al., 1992; Scholl & Flombaum, 2010; Green & Quilty-Dunn, 2020; Peters & Kriegeskorte, 2021) that remain active while objects are present but not visible (C. Baker et al., 2001; Hulme & Zeki, 2007; Puneeth & Arun, 2016). Beyond being stable, these representations can also be updated during periods of occlusion. We routinely track objects' positions while they are in motion, even when they go out of sight (Scholl & Pylyshyn, 1999; Teichmann et al., 2021), and we are also able to track and extrapolate changes along other, more abstract feature dimensions (Blaser et al., 2000; Blaser & Sperling, 2008; Makin & Bertamini, 2014; Makin & Chauhan, 2014). Maintaining and updating object representations during occlusion can be seen as signatures of basic 'internal models' of the world. Respectively, they reflect the fact that objects don't suddenly cease to exist from one moment to the next, and that their dynamic changes tend to continue smoothly while we don't see them.

Previous work has shown that visual cortex seems to play a specific role in representing invisible objects during occlusion: in the absence of any visual stimulation, information about occluded objects can be decoded from visual cortex (Erlikhman & Caplovitz, 2017; Teichmann et al., 2022). This is consistent with a more general role of visual cortex in representing top-down expectations in a template-like format (Kok et al., 2014, 2017). Interestingly, these representations appear to be reflected in patterns of activity similar to those evoked during perception, suggesting that they are sensory-like in nature. A common format between these internally generated predictions and perception can support the seamless integration required in many real-world situations, such as when objects continuously disappear and reappear (Munton, 2022).

Can the 'internal models' we use in tracking occluded objects scale up to the complexity of real-world environments? In daily life, objects do not usually appear in isolation, but embedded within the context of a scene. Previous research has shown that we can use stable elements of a scene, such as its layout or environmental boundaries, as reference frames to represent objects' locations and orientations (Hinton & Parsons, 1988; Rieser, 1989; Mou & McNamara, 2002; Galati et al., 2010; Julian et al., 2016, 2018; Lee, 2017). More in general, objects might always be perceived within a reference frame established by contextual information (Graf, 2006). This contextual information, in principle, is ideally suited for driving the tracking of unseen objects. For example, as we change our viewpoint in navigating a scene, the appearance of objects changes jointly with the environment's overall layout. While objects are occluded, then, we could exploit information from the scene's viewpoint changes to predict how objects' appearance will change.

Here, we investigate whether information from scene context can

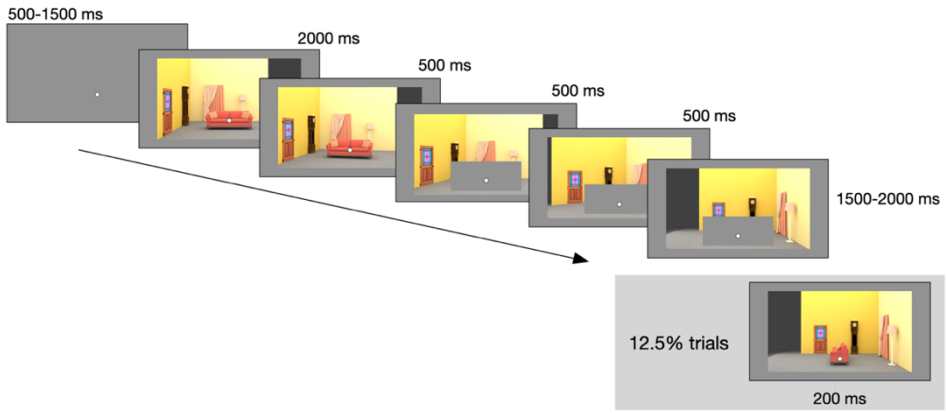


Figure 4.1. Trial outline. The object was shown in an initial view within a scene, which underwent a viewpoint change. The object was occluded during this viewpoint change, and after a series of intermediate views, the scene reached its final view (in this case, rotated by 90° relative to the beginning of the trial). On 12.5% of trials, the object was briefly shown in its updated orientation, coherent with the scene's final view.

automatically drive predictions of occluded objects from novel viewpoints in visual cortex. We show participants objects within realistic scenes, which undergo a change in viewpoint (**Figure 4.1**). During this viewpoint change, the object is completely occluded, while the surrounding scene is still visible. On a majority of trials, the object does not reappear, enabling us to gauge participants' internal representation in the absence of visual input using multivariate pattern analysis (MVPA) in fMRI. Is the representation of the invisible object updated concurrently with the scene's viewpoint? We train linear classifiers on participants' brain activity while they observe images of objects from different viewpoints, and find that they can successfully cross-decode the expected shape of the fully occluded object in the visual system. Participants were not explicitly instructed to predict the object's appearance from the new viewpoint, meaning that the updating of the object representation occurs automatically. These results indicate that representations of fully occluded objects in visual cortex are automatically driven by contextual information from the surrounding scene's viewpoint, providing a bridge towards object tracking and prediction in more naturalistic conditions.

4.2 Methods

4.2.1 Participants

Participants were recruited through the Radboud University participant pool

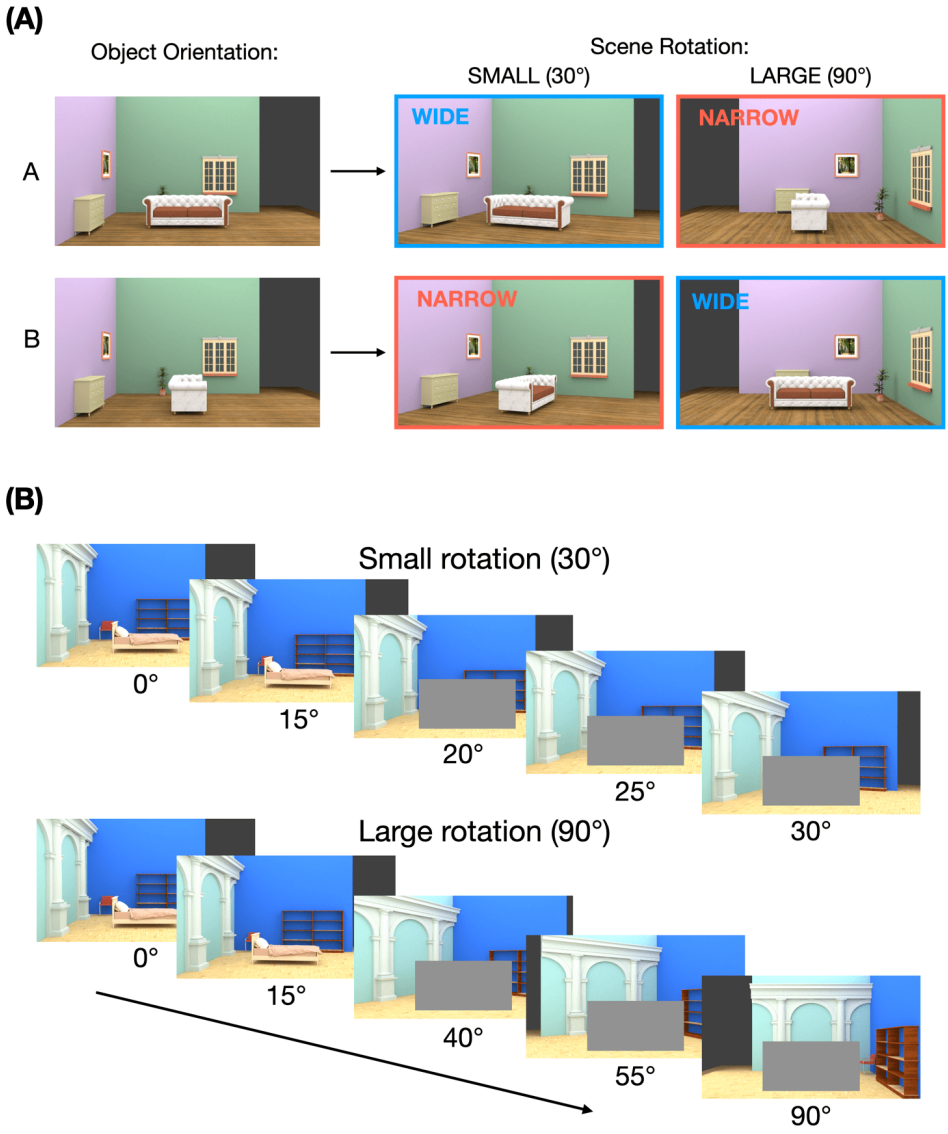


Figure 4.2. (A) Outline of the experimental design. Each initial object orientation could result in either a wide or narrow object view, depending on the amount of scene rotation.
(B) Examples of the two possible rotation amounts, small (30°) and large (90°), with the corresponding sequences of views shown on each trial.

(SONA systems) and received a monetary reimbursement for their participation. They provided informed consent before the experimental session. The study was

in accordance with the institutional guidelines of the local ethical committee (CMO region Arnhem-Nijmegen, The Netherlands, Protocol CMO2014/288). A total of 34 participants took part in the study: this predetermined sample size was chosen to achieve 80% power for detecting a medium-sized effect. Four participants were excluded due to poor performance on the main task (see **Results**), leaving a sample of 30 partipants (16 females, mean age = 25.2, SD = 8.5).

4.2.2 Apparatus

Participants viewed the stimuli through a mirror mounted on the head coil of the scanner. The stimuli were presented on a 1024x768 EIKI LC-XL100 projector (60 Hz refresh rate), back-projected onto a projection screen (Macada DAP diffuse KBA) attached to the back of the scanner bore. The effective viewing distance (eyes from mirror + mirror from screen) was approximately 1440 mm. Stimuli were presented using Psychtoolbox (Brainard, 1997) in MATLAB R2017b (Mathworks Inc.). Participants provided responses using a HHSC-2x4-C button box.

4.2.3 Procedure

Before entering the scanner, participants were instructed about the main task they were going to perform and were shown example stimuli. They were also told that on some runs they would have to detect repeated images (1-back task in the Training and Functional Localizer runs). During the five-minute anatomical scan, they practiced the main task, receiving feedback. Participants were in the scanner for a total of 13 functional runs. Each functional run began and ended with 15 seconds of fixation.

4.2.4 Experimental design & stimuli: Main task runs

In the main task (**Figure 4.1**) participants saw realistic scenes featuring a prominent object in the center (a bed or couch). The object was then completely occluded, while the scene underwent a change in viewpoint. The object reappeared only on a minority of trials, allowing us to assess the content of participants' predictions in the absence of visual input.

On each trial, a fixation dot was shown for a randomly jittered duration between 1500 and 2000 ms, followed by the initial view of the scene for 2000 ms. The scene then started rotating, in 3 intermediate views, each shown for 500 ms. The object was fully occluded starting from the second of these intermediate views. The final view of the scene was then displayed for a duration randomly jittered between 1500 and 2000 ms. On a subset of trials (40/320, or 12.5% in total - between 2 and 10 per run) the occluder disappeared, revealing the object behind it from the new viewpoint for 200 ms. To encourage participants to pay attention to the stimulus sequence, at the end of each run they were asked to report on how many trials the object reappeared after occlusion, and then shown

the correct number as feedback.

On exactly half of the trials, the final viewpoint of the object would project a *wide* shape on the screen, and on the other half, a *narrow* shape. These two conditions were both equally split into two different initial viewpoints, and two overall amounts of rotation of the scene, as illustrated in **Figure 4.2A**. The two initial viewpoints ensured that a given outcome (*wide* or *narrow*) was not uniquely associated with a particular viewpoint at the beginning of the sequence, avoiding memory-related confounds on decoding. Two different possible amounts of rotation (**Figure 4.2B**) were included to ensure that participants could not predict the final view of the object by mentally rotating it by a constant amount on every trial, ignoring the background. Instead, they needed to observe the rotation of the background scene.

The stimuli for the main task and training runs were 20 different indoor scenes modeled in Blender 2.80 (The Blender Foundation) and rendered using the Cycles rendering engine for realistic lighting. The scenes all had the same layout (floor, two walls at a right angle and a main object in the center) but contained various other objects, adjacent to the walls, and different textures on the walls and floor. The central object was a couch in half of the scenes, and a bed in the other half. This object's size was the same across scenes. For each scene, a range of viewpoints was rendered, by rotating the entire scene around the vertical axis (out of the image plane) between 0° and 90°, in steps of 5°. A subset of these viewpoints was presented on each trial. The two walls were oriented such that the scene was fully visible from all the viewpoints. The scenes were presented at the center of the screen at a size of 20.53 x 11.64 degrees of visual angle (dva). The occluder was a grey rectangle which had the height and width of the largest possible view of the object on that particular scene (average size: 5.50 x 2.86 dva), plus a margin (horizontal: 1.08 dva, vertical: 0.43 dva) to ensure the object was fully covered and its shadow was not visible, which would have provided a cue to its orientation. The fixation dot (size 0.2 dva, shown at the location of the central object, 3.24 dva below the center of the screen) was present throughout the whole image sequence, and participants were instructed to maintain fixation. Each main task run comprised 40 trials, and each participant underwent 8 main task runs, for a total of 320 trials.

4.2.5 Experimental design & stimuli: Training runs

The purpose of the training runs was to obtain prototypical response patterns to the stimuli (objects and scenes) in the main task, in order to compare the BOLD response elicited by expecting a given stimulus with that elicited by seeing it on the screen. The images shown in the training runs (**Figure 3**) were the objects as they would have appeared at the end of the roation sequence, also rotated by 30° or 90°, presented in isolation without the surrounding scene. Different object exemplars were grouped together by their proximal shape, so that a given miniblock contained exclusively wide or narrow objects, comprising different initial

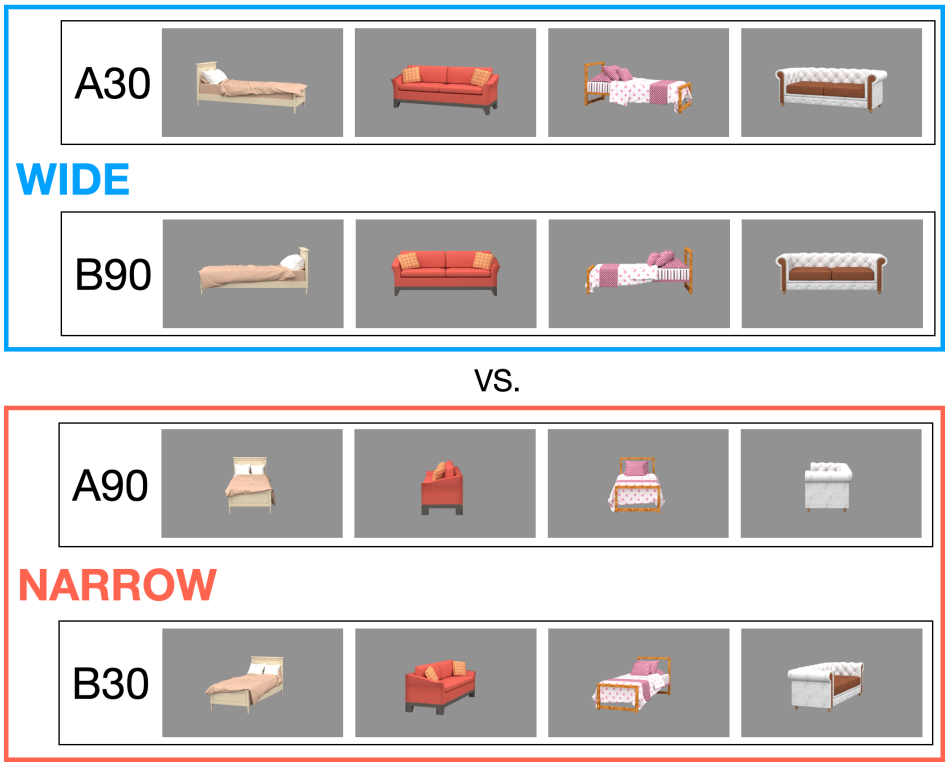


Figure 4.3. The two stimulus categories used to train the classifiers: Wide and Narrow objects. Both categories included two different views (one orientation of the object rotated by a small amount, the other by a large amount) and multiple object exemplars (couches and bed). Objects were always shown in isolation during the training runs, at the same size and position as in the main task runs.

orientations and final rotations. Images of the scenes as they appeared at the end of the rotations in the main task (either rotated by 30° or 90°), with the occluder present, were also shown during the training runs. However, this data was not used in our final analysis. Each miniblock consisted of 9 images (6.75 s in total), each image being presented for 350 ms and followed by a 400 ms blank interval. After a series of 8 different miniblocks (54 s in total) a longer blank interval was shown for 6.75 s. The fixation dot was present throughout, and all stimuli were shown at the exact same location and size as in the main task runs. Participants were instructed to press any button whenever the exact same image was repeated twice in a row (1-back task). Each training run included 40 mini-blocks, and 3 training runs were presented in total.

4.2.6 fMRI data acquisition and preprocessing

fMRI images were acquired on a 3T Magnetom PrismaFit MR scanner (Siemens AG, Healthcare Sector, Erlangen, Germany) using a 32-channel head coil. For acquisition of functional data, a T2*-weighted gradient echo EPI sequence with 6x multiband acceleration factor was used (TR 1 s, TE 34 ms, flip angle 60°, 2x2x2 mm isotropic voxels, 66 slices). For the main task runs, 315 images per run were acquired, and 333 and 318 images for the training and functional localizer runs respectively. A high-resolution T1-weighted anatomical scan was acquired at the start of the experimental session, using an MPRAGE sequence (TR 2.3 s, TE 3.03 ms, flip angle: 8°, 1x1x1 mm isotropic voxels, 192 sagittal slices, FOV 256 mm). The data was preprocessed using SPM12 (Penny et al., 2011) through the Nipype 1.6.0 interface (Gorgolewski et al., 2011). The functional volumes were fieldmap-corrected, spatially realigned, co-registered with the anatomical image, normalized to MNI 152 space using the template provided in SPM, and smoothed with a 3x3x3 mm FWHM Gaussian filter.

4.2.7 GLM analysis

The responses evoked by each of the stimulus types relevant to our analysis were modelled using general linear models (GLMs) in SPM12, using the Nipype 1.6.0 interface.

In the main task runs, the entire period from the onset of the final scene view to its offset was modeled as a boxcar and the time series was convolved with the canonical HRF provided in SPM12. We included regressors for Wide and Narrow expected final object shapes, and excluded trials in which the object reappeared after the occlusion. We thus obtained a single beta map per condition (Wide/Narrow) per run.

In the training runs, each individual miniblock was modeled as a boxcar and the time series was convolved with the canonical HRF. Regressors were included for each scene viewpoint (rotated by 30° or 90° relative to the initial viewpoint) and for each proximal object shape (Wide and Narrow), yielding one beta weight map per condition per miniblock per run.

For functional localizer runs, we also used a block-based design, but estimated a single beta map per condition per run. Miniblocks were modeled as boxcars and convolved with the canonical HRF, yielding a beta weight map for each condition (Objects, Scrambled objects, Faces and Scenes) per run.

All GLMs included six motion parameters and one run-based regressor as nuisance regressors. As participants were performing a 1-back task in the training and localizer runs, these runs also included a nuisance regressor synchronized to participants' button presses.

4.2.8 Regions of interest

Anatomical regions of interest were defined based on an anatomical atlas. For the initial analysis, in order to broadly cover the visual system, we used a large ROI including Brodmann areas (BAs) 17, 18 (corresponding to retinotopic areas V1 and V2 respectively, Wohlschläger et al., 2005), 19 (including visual areas V3, V4, V5/MT and V6) and 37 (corresponding to occipitotemporal cortex and including the posterior fusiform gyrus and the posterior inferior temporal gyrus). We subsequently split this ROI into two sub-ROIs, roughly corresponding to early (BAs 17/18) and late (BAs 19/37) visual cortex. The latter was then further split into BAs 19 and 37.

4.2.9 Sliding window analysis

The purpose of this exploratory analysis was to assess whether above-chance decoding could be reliably found at intermediate levels of the visual hierarchy. The large ROI comprising BAs 17, 18, 19 and 37 was used again, divided by hemisphere and sliced by y-coordinate, with a sliding window of 13 voxels. The sliding window proceeded from posterior to anterior, within the bounds of the macro-ROI (from MNI y-coordinate -102 to -24).

4.2.10 Robustness to voxel inclusion

To assess the robustness of our results within BA 19, we created multiple sub-ROIs, by including increasing numbers of voxels responding significantly ($p_{uncorrected} < 0.05$ for a bidirectional t-test) to either of the main conditions of interest in the main task runs: expectancy of a wide vs. narrow or narrow vs. wide shape. Voxels were sorted according to the absolute value of their t-statistic in this univariate contrast, and the top N voxels were selected for each sub-ROI, with N ranging from 100 to 1000, yielding an increasingly liberal voxel inclusion threshold. All analyses were conducted within each hemisphere separately, and results were averaged across hemispheres.

4.2.11 Multivariate pattern analysis

The multivariate pattern analysis (MVPA) was conducted using linear support vector machines (SVMs) implemented in Scikit-learn (Pedregosa et al., 2011) and PyMVPA (Hanke et al., 2009). The cross-classification analysis consisted of training SVM classifiers on the miniblock-based beta weights from the training runs, and testing them on the run-based beta weights from the main task runs, and vice versa. Both directions of cross-classification were run, and the results averaged, in order to increase the robustness to to task- or stimulus-unrelated factors that can lead to asymmetries between classification directions, such as different signal-to-noise ratios (van den Hurk & de Beeck, 2019). The training and

testing datasets were separately z-scored before decoding. Each classifier was trained and tested on voxels within a single ROI and hemisphere, in a single subject. To decode the stimulus feature of interest - proximal object shape (wide vs. narrow), we trained classifiers on the isolated objects presented in the training runs, and tested them on the expected object shapes in the test runs (or vice versa, depending on decoding direction).

We used a continuous measure of classifier performance, as discrete classification outcomes have been shown to be less reliable in estimating the distance between stimulus representations in the brain (Walther et al., 2016). Our measure, which we call 'classifier information', was based on a test sample's distance from the classification hyperplane estimated by the SVM (distance from bound). Classifier information was defined as follows:

$$\text{Classifier Information} = \frac{1}{n}\sum_{i=1}^{n} d_i l_i$$

Where d_i's are the z-scored distances from bound, l_i's are the true binary labels (either -1 or 1), for each sample, and n is the number of test samples. Intuitively, this measure corresponds to the average match between each distance from bound and the corresponding ground-truth label, i.e. the degree to which the distance is positive when the target is positive, and negative when the target is negative. Classification is above chance when this measure is higher than zero. Averaging this measure across samples allows to compare classification performance with different numbers of test samples, enabling us to combine data across decoding directions. The classifier information was computed for each ROI (or sub-ROI in the sliding window and voxel selection analyses) and each subject. Importantly, our results were consistent, albeit noisier, when using classification accuracy instead of classifier information.

4.2.12 Significance testing

To assess whether classifier information was reliably above zero for the decoding of the occluded object's shape, we used two approaches. (1) At the whole-ROI level, we used simple one-sample t-tests against a null of zero for the average classifier information of each participant, for a classifier trained and tested on all voxels. We started with a broad ROI including much of the visual stream, and then progressively split the ROIs that showed above-chance decoding into sub-regions. To correct for multiple comparisons, we used one-step Bonferroni correction. (2) For the sliding window and voxel selection analyses, we tested for the presence of above-chance decoding across y-coordinates and numbers of voxels, respectively, using threshold-free cluster enhancement (TFCE, Smith & Nichols, 2009). TFCE boosts the magnitude of a statistic based on its extent across neighboring samples (in this case, spatial windows or sub-ROIs with similar numbers of voxels), reflecting the assumption that any signal in the data should

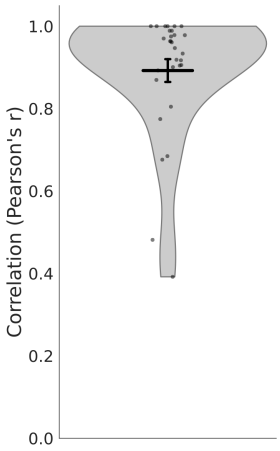


Figure 4.4. Pearson's correlation between true and estimated numbers of object reappearances for each included participant.

be smooth across consecutive datapoints. This measure is then compared with a null distribution generated by randomly shifting the signs of each participant's 1D map (classifier information across spatial windows/sub-ROIs). This null distribution has the same variance and autocorrelation as the original signal. This shuffling procedure was applied 10,000 times. TFCE was computed using the MNE toolbox (Gramfort et al., 2013).

4.3 Results

4.3.1 Behavioral data

To ensure that participants were paying attention to the stimulus sequences, we instructed them to pay attention to when the object would reappear after occlusion, and report the number of reappearances at the end of each block. As a measure of participants' attention to the task, we used the Pearson's correlation between their estimates and the correct number of reappearances for each run. This led to two participants being excluded outright due to their responses being negatively correlated with the true values ($r = -0.59$ and -0.46). Two further participants were then excluded due to their correlation being >2 standard deviations away from the sample mean. These participants' correlation coefficients were close to zero: $r = 0.07$ and 0.15. The remaining ($n = 30$) participants' responses were positively correlated with the true values (mean $r = 0.89$, minimum $= 0.39$), with a majority (25/30) having correlations higher than 0.80, as shown in **Figure 4.4**.

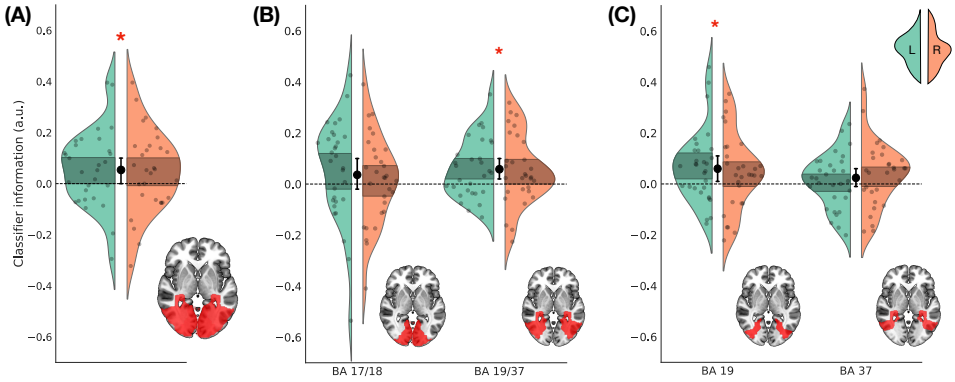


Figure 4.5. Results of the MVPA analysis on whole anatomical ROIs. Color and position indicate the left and right hemisphere, as illustrated at the top right.

4.3.2 MVPA results - whole ROIs

In our main analysis, we addressed the question of whether the expected object shape could be decoded from multivariate activity patterns in visual cortex, in the absence of visual stimulation. To be agnostic to the precise anatomical locus of this information, we first ran linear decoders on a wide ROI comprising much of the visual stream (Brodmann areas 17, 18, 19 and 37 - see **Methods** for details). We found that object shape could be decoded above chance in this macro-area ($t(29) = 2.23$, $p = 0.0339$, $d = 0.41$, 95% CI = [0.0, 0.1], **Figure 4.5A**), indicating that information about the expected shape was present in the visual system. In order to assess the robustness of our results to voxel inclusion, as well as to clarify their anatomical origin, we then split this ROI into two sub-ROIs, corresponding to low-level (BA 17/18) and high-level (BA 19/37) visual cortex. Two separate t-tests against chance revealed that BA 19/37 contained information about the expected object shape ($t(29) = 2.90$, $p_{bonf} = 0.014$, $d = 0.53$, 95% CI = [0.02, 0.1]), while BA 17/18 did not ($t(29) = 1.24$, $p_{bonf} = 0.451$, $d = 0.23$, 95% CI = [-0.02, 0.1], **Figure 4.5B**). We then further broke down the region showing significant above-chance decoding, BA 19/37, into its two constituent Brodmann areas. This analysis showed that the expected object shape could be decoded in BA 19 ($t(29) = 2.69$, $p_{bonf} = 0.023$, $d = 0.49$, 95% CI = [0.01, 0.11]) and not in BA 37 ($t(29) = 1.42$, $p_{bonf} = 0.333$, $d = 0.26$, 95% CI = [-0.01, 0.06], **Figure 4.5C**). These results indicate that a representation of the occluded object's proximal shape, updated consistently with the surrounding scene viewpoint, was present in the visual system. Moreover, it was primarily localized within the intermediate visual area BA 19.

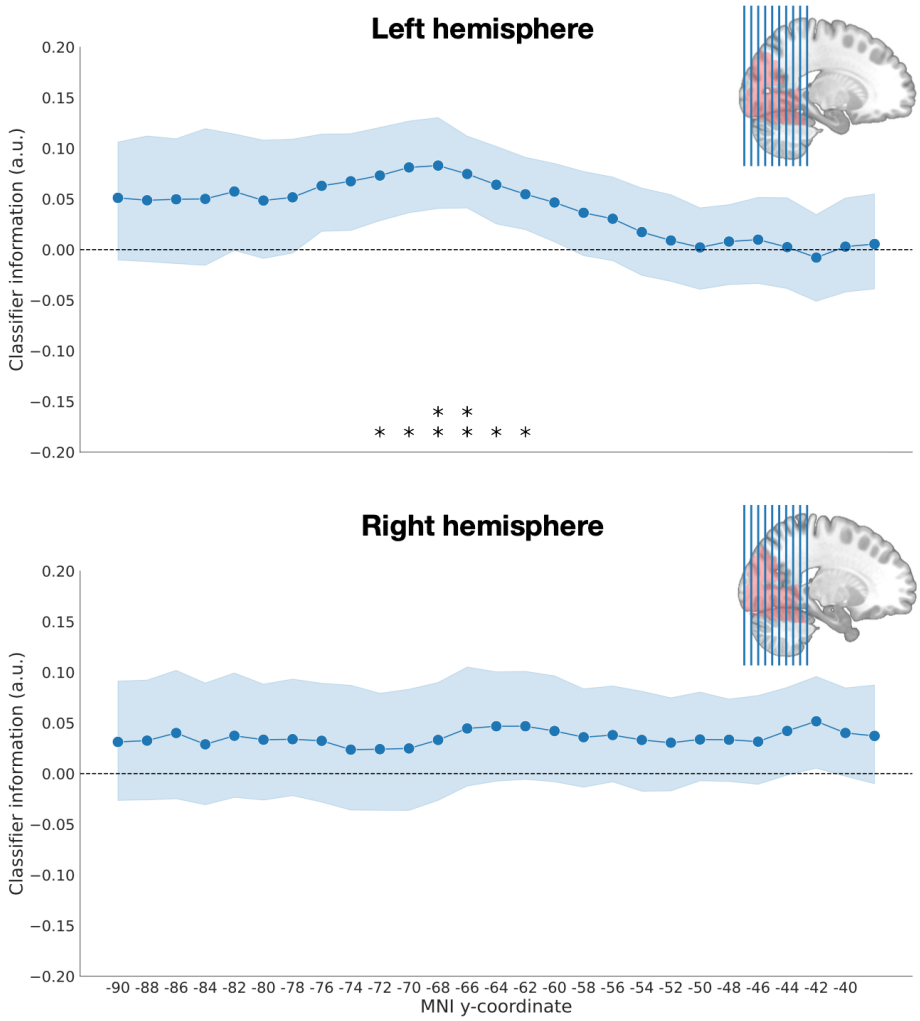


Figure 4.6. Classifier information within a sliding window along the y-axis, for each hemisphere. MNI coordinates are the centers of the sliding windows. Asterisks indicate significance as measured with TFCE. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

4.3.3 MVPA results - sliding window analysis

Our result indicated that information about the updated object shape was primarily found in mid-level visual area BA 19. As a confirmation that our results were primarily driven by intermediate levels of the visual stream, we conducted an

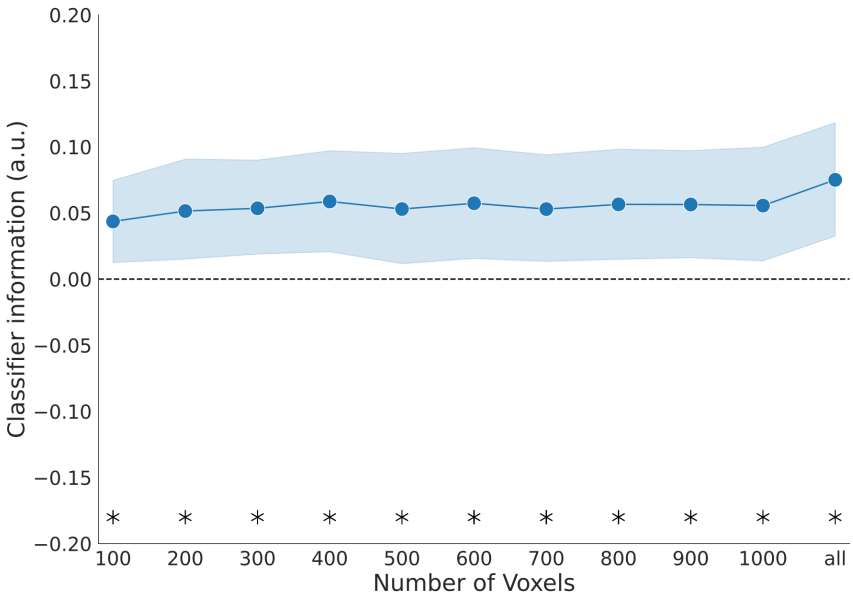


Figure 4.7. Robustness of the results within BA19 to the inclusion of different numbers of voxels.

exploratory analysis, in which object shape was decoded from a sliding window along the posterior to anterior axis of the brain (MNI y-coordinate). Inspecting the resulting plot (**Figure 4.6**), it can be seen how above-chance decoding emerges at intermediate coordinates within the left hemisphere (peak mean classifier information y = -68, p_{TFCE} = 0.007) while no above-chance decoding could be found within the right hemisphere (peak mean classifier information y = -42, p_{TFCE} = 0.140).

4.3.4 MVPA results - voxel selection

As a further check of the robustness of our results in BA 19, we cross-decoded object shape on sub-ROIs with an increasingly liberal threshold for voxel inclusion (see **Methods** for details). This analysis showed that the decoding of object shape in this ROI is robust to the inclusion of different numbers of voxels ($p_{TFCE} < 0.05$ for all sub-ROIs, **Figure 4.7**).

4.4 Discussion

The present results provide evidence for predictions of a fully occluded object's appearance from a new viewpoint, driven by contextual scene information. These predictions occur in the absence of an explicit task, and are reflected in sensory-like representations in visual cortex. These findings suggest that internal object representations are automatically updated with the surrounding scene context, possibly providing a mechanism for predicting and tracking temporarily invisible objects in the real world.

In this study, we showed that the representations of occluded objects in visual cortex resemble those of visible objects, by using a multivariate cross-decoding scheme. Several cognitive processes, such as attention (Peelen & Kastner, 2011; Battistoni et al., 2017), expectations (Kok et al., 2014; Hindy et al., 2016), mental imagery (Dijkstra et al., 2019; Pearson, 2019) and working memory (Albers et al., 2013; Christophel et al., 2015; Gayet et al., 2018) are known to elicit visual-like cortical representations in the absence of perceptual input. These representations are believed to result from feedback connections to visual cortex, and to serve the purpose of disambiguating sensory information by comparing it with top-down signals (De Lange et al., 2018). The present study occupies an intermediate position between prior work investigating representations in the complete absence of visual input, and other studies looking at contextual modulation of visible stimuli (e.g. Murray et al., 2006; Heilbron et al., 2020). Our results show that, even when an object is fully invisible, it can still be influenced by contextual information, in accordance with the way objects and the scenes surrounding them change coherently in the real world. This seamless integration of visible and invisible information can be extremely useful in tracking objects across periods of invisibility, as often happens in daily life (Munton, 2022).

In our own previous work (**Chapter 3**), using a similar paradigm to the one used here, we investigated the role of scene-driven predictions in modulating the representations of visible objects. Specifically, we compared fMRI activity patterns, and behavioral performance in a perceptual task, between objects that were or were not rotated congruently with the surrounding scene after occlusion. We found that congruently rotated objects yielded higher behavioral accuracy and enhanced multivariate decodability in visual cortex. In the present work, we show that the expectation's content can be directly decoded in the absence of visual input. This suggests that the previously reported behavioral and neural advantage might be the result of a comparison between this template-like representation and incoming sensory input. Interestingly, this comparison operation seemed to primarily occur in early visual cortex (EVC; corresponding to V1 and V2). While a precise anatomical localization of the expectation signal is beyond the scope of this study, here above-chance decoding of this signal was primarily found in mid-level BA 19. One possibility is that this discrepancy is due to recurrent interactions between neighboring visual areas (Hochstein & Ahissar, 2002; Lee & Mumford, 2003; Dijkstra et al., 2020), whereby the expectation signal represented in BA 19

is fed back to EVC to be compared with incoming visual input. Roelfsema & de Lange (2016) proposed that early retinotopic areas in the visual system might act as a 'cognitive blackboard' where spatially organized information is stored and manipulated during disparate cognitive operations. Areas with receptive field sizes that match the relevant spatial scale of the task at hand are flexibly recruited, with earlier areas being involved in finer-grained computations. In both our studies (**Chapter 3** and the present one), we explicitly designed our stimuli to maximize global shape differences between the two possible outcomes of the rotation sequence (wide or narrow). Participants might have then generated a coarse-scale representation of the occluded object's shape, involving an intermediate visual area (BA 19) with large receptive fields. In **Chapter 3**, since subjects were additionally engaged in a fine-grained visual discrimination task on the reappearing objects, this representation might have been fed back to EVC to facilitate visual processing.

While the present results indicate that information about the occluded object's appearance is 'filled in' in the visual system, an open question is what the specific nature of this information is. Prior research in perceptual expectations has often used synthetic stimuli, consisting of a few basic visual features such as orientation and spatial frequency (e.g. Kok et al., 2012, 2014). In such an impoverished scenario, top-down visual expectations could provide a full description of the predicted stimulus. In the case of complex real-world scenes, this might not be the case. While some previous studies have reported that the neural completion of occluded scenes resembles a faithful description of the missing scene section (Smith & Muckli, 2010; Svanera et al., 2021), others have suggested that it might be more schematic (Morgan et al., 2019). Particularly in dynamic settings, in which objects need to be tracked in real time, it might not be viable to predict them in every aspect. A recent study on object tracking under occlusion found that only the object's position could be decoded above chance in visual cortex during occlusion (Teichmann et al., 2022). In the training runs of our study, different object stimuli were grouped in miniblocks by their proximal shape (wide or narrow). We thus obtained prototypical object representations that collapsed across specific object exemplars, making it impossible to distinguish between a purely abstract representation of proximal shape and a richer pictorial representation. A future study using exemplar-level decoding (M. R. Johnson & Johnson, 2014; Wurm & Lingnau, 2015) or reconstruction approaches (Horikawa & Kamitani, 2017; Senden et al., 2019; Shen et al., 2019; Dado et al., 2022) could help adjudicate between these two possibilities.

Contrary to prior work investigating object tracking in space, in our study both the object's position and identity (as revealed in the few reappearance trials) remained fixed behind the occluder. What varied was instead the object's 3D orientation, and consequently its projected shape on the image plane. Is it still possible that an abstract representation of shape exclusively was being updated in visual cortex, rather than a detailed prediction of the object's appearance? Mental rotation studies, in which participants are explicitly instructed to imagine

an object's appearance from a new viewpoint (Shepard & Metzler, 1971) have found evidence that they manipulate an image-like representation in their minds (Cooper & Shepard, 1973; Shepard & Cooper, 1982; Koriat & Norman, 1984, 1988). On the other hand, Graf (2006) proposed that spatial transformations, such as rotation and scaling, might not pertain exclusively to explicit mental rotation processes, but might be a general computation in object perception. Particularly, he highlighted that object recognition can be facilitated when the observer has set up an appropriate spatial reference frame for the object, e.g. with a background that matches its viewpoint (Humphrey & Jolicoeur, 1993). Setting up these reference frames differs from mental rotation in its being object-aspecific: rather than a picture-like prediction of a specific object in a specific orientation, it resembles an abstract representation of that orientation, which can have a facilitatory effect across different objects (Graf et al., 2005). It is possible that scene viewpoint information in our study was establishing a similar reference frame, eliciting a general prediction of a particular object orientation rather than a specific object exemplar. Behaviorally, this would predict the facilitatory effect we reported in **Chapters 2 & 3** to hold across different objects. Graf (2006) also proposed that coordinate frames might be established in the visual system through feedback mechanisms conventionally associated with attentional modulation, compatible with our finding of the object shape being represented in visual cortex. Interestingly, a recent study (Gayet & Peelen, 2022) manipulated the distance at which participants were searching for objects in real-world scenes, and found, using fMRI, that their internal representation of the objects was "scaled" according to the object's expected size. Similar to our study, the size-specific object representation could be decoded in the visual system, despite the object not being visible. Our findings then raise the possibility that scene context might provide a reference frame for disparate object transformations, from scaling to rotation, and support object predictions automatically, even without an explicit task. Future research should investigate the question of whether this occurs through the generation of object-specific predictions, or through general transformation operations that generalize across objects (Ward et al., 2018).

To conclude, we found evidence for a representation of object shape in visual cortex that reflects the changed viewpoint of a surrounding 3D scene. This representation was reflected in multivariate activity patterns, and emerged in the absence of an explicit prediction task. These results indicate that predicting incomplete visual input in everyday scenes might capitalize on structured spatial relations, bridging our known ability to track occluded objects with the complexity of real-world environments.

Chapter 5
Automatic size scaling of object representations driven by scene context

Abstract

As our viewpoint varies in the real world, we see scenes changing in a coherent way. For example, as we move forward in a scene, the retinal size of all the objects in the scene will increase in accordance with the amount of forward motion. This regularity can be exploited to predict how an object will transform, allowing us to efficiently track objects in naturalistic contexts. While previous research has found that object orientation is automatically predicted based on the rotation of the surrounding scene, it is unknown whether this can generalize to other transformations beyond rotation. Here, in a series of online behavioral experiments (N = 151), we investigate whether participants automatically predict the retinal size of an object, based on the changing distance of the surrounding scene. We compare their responses in an orthogonal perceptual task, on an object that either matches the expected size or not. We find that scene-driven size expectations strongly influence task performance. The directionality of this effect remains consistent even when size expectations are violated on a majority of trials, suggesting that scene context elicits automatic predictions in line with real-world regularities, that cannot easily be overruled by short-term evidence. Together, these findings point to a general role of scenes in driving predictions of object transformations.

5.1 Introduction

The appearance of objects constantly changes with our viewpoint. As we walk around an object, we see it rotating in depth. As we get closer or farther, we see its size increase or decrease. Predicting how these transformations affect the appearance of objects is crucial for navigating the world effectively. For this reason, it is believed that we generate internal representations of external objects which behave in correspondence to those objects (Shepard, 1984, 2001; Higgins et al., 2022). For example, we can mentally rotate the representation of an object, such that the rotated representation will correspond to the rotated object in the external world (Shepard & Metzler, 1971).

In most real-world situations, when objects undergo transformations, they do so coherently with their context. For example, as we walk through a room and our viewpoint changes, so does the orientation of the objects in the room, and that of its walls. This means that real-world scenes are highly redundant: all parts of a scene tend to simultaneously undergo similar transformations. The task of predicting object transformations is computationally challenging: it requires several inferences to be made. In the case of mental rotation, for example, the direction and amount of rotation to apply to the object must be determined (Hamrick & Griffiths, 2014). It would be advantageous, then, if we could exploit the redundancy of real-world scenes by mentally transforming objects coherently with their context. For at least some mental transformations, this seems to be the case; in the previous chapters, we have shown that internal object representations are rotated concurrently with the viewpoint of the surrounding scene, as reflected in both behavior (**Chapter 2**) and activation patterns in visual cortex (**Chapters 3-4**). An important question left unanswered by our previous work is to what extent the influence of scene context on mental transformations is a general phenomenon, or one that applies strictly to rotation. Prior research has shown that our minds are able to predict a wide variety of transformations beyond rotation. These include other *rigid* transformations, that leave objects' shapes unaltered, such as translation and scaling (Bennett, 2002; Bundesen et al., 1983; Bundesen & Larsen, 1975; Larsen & Bundesen, 1978, 1998; Schmidt et al., 2016; Sekuler & Nash, 1972), and even *non-rigid* ones, that modify an object's shape, such as deforming or tearing it apart (Hafri et al., 2022; Hahn et al., 2009; Kourtzi & Shiffrar, 2001; Spröte et al., 2016; Spröte & Fleming, 2016). This suggests that predicting object transformations might be a general cognitive capacity, encompassing several qualitatively different ways in which objects can change in the world. Can other transformations also be driven by scene context, in a way similar to rotation?

Here, to investigate the role of scene context in driving mental transformations beyond rotation, we focus on *scaling*, the predictable change in an object's retinal size as a function of viewing distance. As objects rarely physically shrink or expand in the real world, their retinal size mostly varies with our distance from them: accordingly, behavioral evidence suggests that we generally perceive size changes as translations in depth (Bundesen et al., 1983;

Larsen & Bundesen, 2009). Because retinal size depends on distance, scene context should play a crucial role in influencing our representations of object size, as real-world scenes contain a rich variety of depth cues (Landy et al., 1995). Indeed, the perceived size of an object has long been known to be altered by pictorial depth cues in a scene, such that objects farther away are perceived as larger, reflecting their inferred real-world size (Leibowitz et al., 1969; Yildiz et al., 2021). The influence of scene context on object size has been shown to generalize beyond perceived objects, affecting internally generated representations as well. A recent fMRI study (Gayet & Peelen, 2022) found that as observers prepared to search for objects at a given distance in a scene, they generated internal representations of those objects that were scaled consistently with the search distance. This shows that observers can adjust the size of internal object representations based on the scene context. Their task, however, explicitly informed participants about the upcoming distance, and generating corresponding preparatory object representations was beneficial for the task at hand (reporting the presence/absence of the object). This leaves open the question of whether the rescaling of object representations occurs during day-to-day vision, when not imposed by task demands. Moreover, this study made use of static scene images, so it remains unknown whether object representations can be scaled dynamically, coherently with the changing viewpoint of a scene. Here, we address the questions of whether scene-driven scaling can occur automatically (without being imposed by task demands), and whether it can be elicited by the observed transformations of the scene. Answering these questions would provide insight into the way in which mental transformations, such as scaling, can guide perception in the real world, in the presence of rich contextual cues.

We ran a series of online behavioral experiments using a paradigm similar to the one described in **Chapter 2**, to determine whether scaling of internal object representations shares cognitive mechanisms with rotation, and is similarly influenced by scene viewpoint. On each trial, we showed participants an object placed within a realistic 3D scene (**Figure 5.1**). The viewpoint on the scene translated in depth, with the camera 'zooming in', and during this viewpoint shift, the object was concealed by an occluder. Then it reappeared, either with a size consistent with the new distance of the surrounding scene (Expected trials) or with an inconsistent size (Unexpected trials). Importantly, we varied the amount of scene viewpoint change from trial to trial, so that whether the object had been rescaled by the expected amount or not could only be determined by observing how the scene changed. Participants had to perform an orthogonal visual discrimination task on the reappeared object, a task that in principle did not require them to form an expectation of the object size or to take the scene's viewpoint change into account. We compare their performance on this task between Expected and Unexpected trials, and find that expectations of object size, driven by the scene, substantially influence their responses. This suggests that scene context drives internal predictions of object size even in the absence

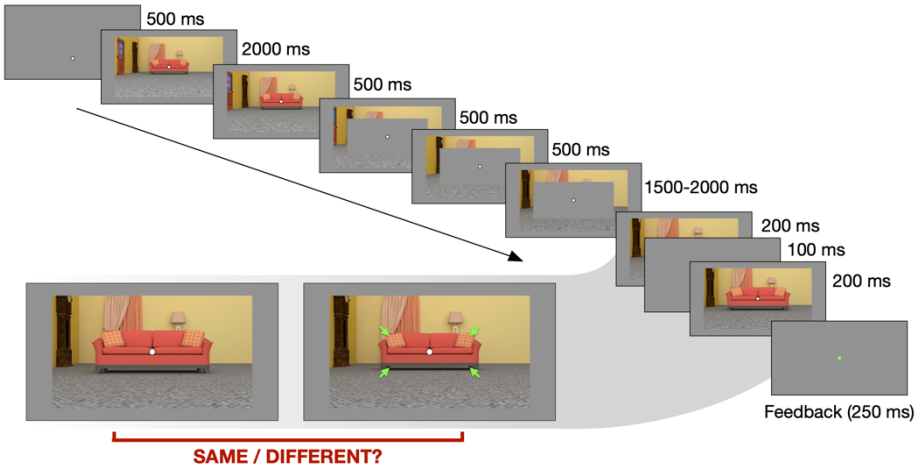


Figure 5.1. Example of a trial - in this case, corresponding to a "Large" total translation of the scene, an "Expected" view of the object, and a "Different" trial: the second, rapidly shown probe is slightly smaller than the first (arrows added for illustration).

of any explicit instruction. Moreover, across three experiments, we manipulate the probability of the object appearing with a size matching the scene's viewpoint. We find that even when scene-driven expectations are violated on a large proportion of trials, they still influence behavioral responses in a similar manner, showcasing the obligatory nature of the influence of scene context on object transformations. Altogether, these results indicate that mental scaling can be driven by scene context in an automatic way, analogously to rotation, pointing to a general role of scene viewpoint in driving transformations of object representations.

5.2 Methods

5.2.1 Participants

All experiments were run online, hosted on Pavlovia and programmed in Javascript using JsPsych 6.3.0 (De Leeuw, 2015) and the jspsych-psychophysics library (Kuroki, 2021). Participants were recruited on Prolific (Palan & Schitter, 2018) and had to satisfy the following criteria: reside in Europe or the UK, to ensure their timezone was the same as ours and they were participating during day hours; have participated in at least 10 previous studies on Prolific; and have a Prolific approval rate of at least 95%. Participants provided informed consent before the study and received monetary compensation for their participation. The study was

approved by the Radboud University Faculty of Social Sciences Ethics Committee (ECSW2017-2306-517). Participants were included in the analysis if a one-sided binomial test comparing their hit rate in our same/different task with 50% was significant (at alpha = 0.05), meaning that they were performing better than chance across all conditions. We continued data collection until the number of included participants reached 50 for each experiment. In Experiment 1, we excluded 47 participants. Of the included 50 participants, 24 were female, 25 were male, and one participant's demographic information was missing. and mean age was 27.1 ± 4.1. In Experiment 2, we excluded 37 participants. Of the included 50 participants, 25 were female, 24 were male, and 1 participant's information was missing. Mean age was 25.8 ± 4.8. In Experiment 3, we excluded 42 participants. Of the included 51 participants, 21 were female, and mean age was 26.7 ± 4.5.

The high exclusion rate was likely due to the difficulty of the task. The difference between probe stimuli was defined in 3D space (object position in depth), limiting the range of possible stimulus differences we could show. We wanted to make the difference between different object positions (and thus Expected and Unexpected positions) noticeable, so probe objects could appear at either very near or very far distances. We could thus not use the full range of distances available in the scene, and for far object positions, depth differences were very hard to notice. Moreover, we kept the presentation time short (200 ms) for each of the two target stimuli, in order to reduce the influence of deliberate judgment and investigate how scene-driven expectations influenced a primarily perceptual task. Importantly, however, all results reported here remained consistent with no participant exclusions.

5.2.2 Stimuli

The stimuli were based on 4 different indoor scenes modeled in Blender 2.92 and rendered using the Cycles rendering engine for realistic lighting. The scenes all had the same layout (floor, two walls at a right angle and a main object in the center) but the main object varied, as well as the objects present in the background (adjacent to the walls), and the textures on the walls and floor. The central object could be a couch or a bed: we chose large, immovable object categories that are generally expected to remain in a fixed position within a scene. For each scene, a sequence of different viewpoints was rendered by translating the camera gradually closer to the scene (zooming in). The main object was always fully included in the frame, while other background objects could go out of the frame. The main object was always presented with its longer side (front for the couch, side for the bed) facing the viewer. The height and pitch of the camera were chosen so that the main object would always remain at the center of the scene. All scene images had a resolution of 960 x 540 pixels.

5.2.3 Procedure

Each trial (**Figure 5.1**) began with a fixation dot (which was always present during the trial, radius 5 pixels) for 500 ms, followed by the first view of the scene for 2000 ms, the 3 intermediate views for 500 ms each, and the final view for a randomly jittered duration between 1500 and 2000 ms. The central object (couch or bed) was fully visible for the first and second view, and was occluded by a grey rectangle during the third, fourth and final view. The occluder had the height and width of the largest possible view of the object in a specific scene, plus a margin (horizontal margin: 110 pixels, vertical: 40 pixels) to ensure the object was fully covered and its shadow was not visible, which would have provided a cue to its size behind the occluder.

After the final view of the scene was shown, the occluder disappeared, briefly revealing the object (within the scene) twice, for 200 ms each, with a 100 ms inter-stimulus interval in between. We will refer to these two brief presentations of the object as the *probes*. Participants' task was to report whether the second probe was the 'same' as, or 'different' from, the first, by pressing the F or J key, respectively. After responding, they would receive feedback: the fixation dot would turn green following a correct answer, and red following an incorrect one, for 250 ms. They had a maximum of 2500 ms to respond, after which the fixation dot would turn black, the experiment would skip to the next trial and the current trial would be counted as missed.

Participants were explicitly told that their task would be on the final two views of the objects exclusively, but that they should also pay attention to the preceding sequence of images, to ensure they wouldn't completely disengage during the seconds preceding the probes. The first probe was randomly sampled from a normal distribution centered around the Expected or Unexpected object viewpoint, to add a small amount of jitter. The second probe, on half of trials ('same' trials), was exactly the same as the first probe. On the other half of trials ('different' trials), it was translated in depth relative to the first (see **Figure 5.1**, bottom left), forward or backward with equal probability.

The depth difference between the two probes was defined in terms of distance in the virtual scene, using the default Blender unit. We henceforth refer to this measurement unit as *arbitrary unit* (a.u.). The depth difference was shown only on the 'different' trials, and was titrated using a 2-down 1-up staircase, to keep the task difficulty constant across participants and across experiments. Specifically, a single staircase was used across both Expected and Unexpected trials to ensure overall performance was around 70% correct (Wetherill & Levitt, 1965) across conditions, while still allowing for accuracy differences between the Expected and Unexpected conditions. The depth difference was adjusted after both 'same' and 'different' trials. The starting value for the staircase was 1 a.u., step size was 0.05 a.u. (halved after 3 staircase reversals) and the minimum and maximum possible depth differences shown were 0.025 and 1 a.u., respectively. The means and standard deviations of the depth differences reached by the

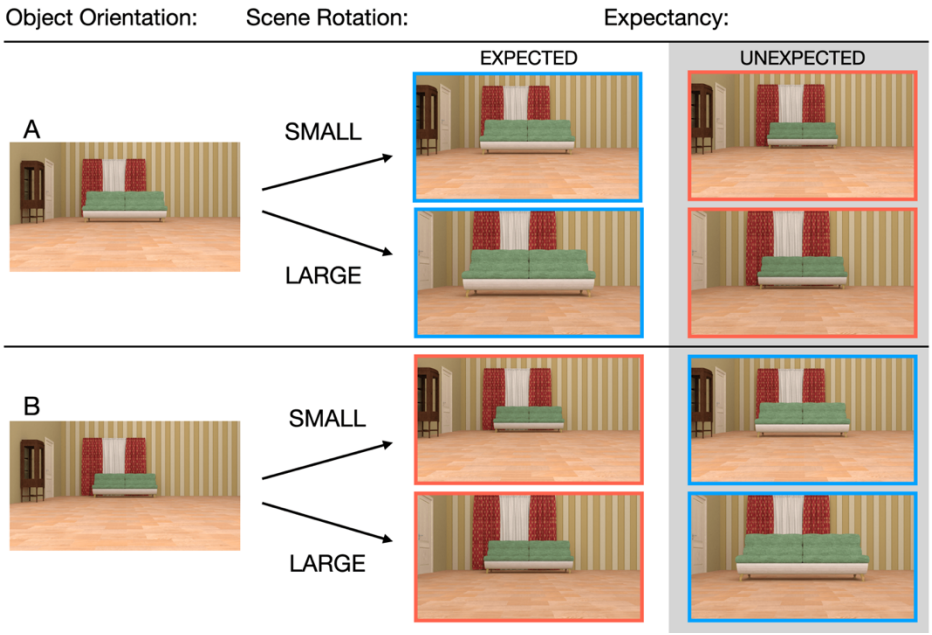


Figure 5.2. Illustration of the experimental design, showing the initial position of the object relative to the scene (either Near or Far), and the final images (after the whole sequence and the occlusion period) resulting from a Small or Large translation on Expected or Unexpected trials. As highlighted by the color frames, the same image could appear as either Expected or Unexpected on different trials.

staircase in the second half of trials, in each of the three experiments, were 0.68 ± 0.23, 0.72 ± 0.19, and 0.7 ± 0.20 a.u. respectively.

Each experiment lasted about 30 minutes in total, divided in N blocks, and participants were encouraged to take a short break after the end of each block. Before the experiment began, participants read the on-screen instructions, accompanied by demonstration images, at their own pace. Then they completed a short practice run. During the practice run, the presentation time of the two probes gradually decreased across trials, from 300 ms to the eventual presentation time that was used in the main experiment (200 ms). This allowed participants to familiarize with the task with an initially less challenging presentation time.

5.2.4 Experimental design

Trials varied along three different factors (**Figure 5.2**): Expectancy (Expected, Unexpected), Object Position relative to the scene (two possible distances from the observer, Near or Far), amount of Scene Translation (Small or Large), and

Scene (1 of 4 different exemplars). The overall proportion of Expected and Unexpected trials varied depending on the experiment (75% of total trials in Exp. 1, 50% in Exp. 2, and 25% in Exp. 3). All the other factors were fully balanced within Expected and Unexpected trials, meaning that each of four partitions of the trials (variably assigned to either Expected or Unexpected depending on the experiment) were equally divided among each combination of Object Position, Scene Translation and Scene (2 x 2 x 4 = 16 conditions, each repeated 3 times, for each partition, resulting in 192 trials in total). All these trials were presented in randomized order throughout the experiment.

On Unexpected trials, at the end of the scene translation sequence, we showed the object in a position inconsistent with the one shown at the start of the trial: on Near trials, the object appeared in the Far position, and vice versa (**Figure 5.2**). This way, the exact same image could be presented as Expected in the context of one trial, and Unexpected in another, avoiding any possible confounds due to visual differences between conditions.

5.2.5 Data analysis

In order to determine whether scene-driven expectations mostly affected observers' sensitivity or bias, we computed d' and criterion for each condition of interest (Expected and Unexpected trials). We consider 'Same' trials as noise, and 'Different' as signal, meaning that criterion is a measure of bias towards responding 'same'. We used the log-linear method (Hautus, 1995) to correct for the rare cases of 100% accuracy in a particular condition.

All analyses were conducted in Python using Pandas 1.2.5 (McKinney, 2011), Numpy 1.20.2 (Harris et al., 2020), Pingouin 0.3.4 (Vallat, 2018), and Scipy 1.6.2 (Virtanen et al., 2020), and results were visualized using Matplotlib 3.3.4 (Hunter, 2007), and Seaborn 0.11.1 (Waskom, 2021).

5.2.6 Post-experiment survey

After completing the experiment, participants were asked three questions that would help us gauge their awareness of the expectation manipulation. The questions were:

- "Your task was only on the final image, when the object changed or not. Did you also pay attention to the sequence of images before the task image?" - the response had to be indicated on a Likert scale from 1 (Not at all) to 7 (All the time).
- "When the scene shifted, did you anticipate seeing the object in the correct viewpoint after it reappeared?" - the response also had to be indicated on a 1-7 Likert scale.
- "What percentage of objects were in line with your expectation? (They reappeared with the correct viewpoint)" - the response had to be a value in percentage, from 0 to 100%.

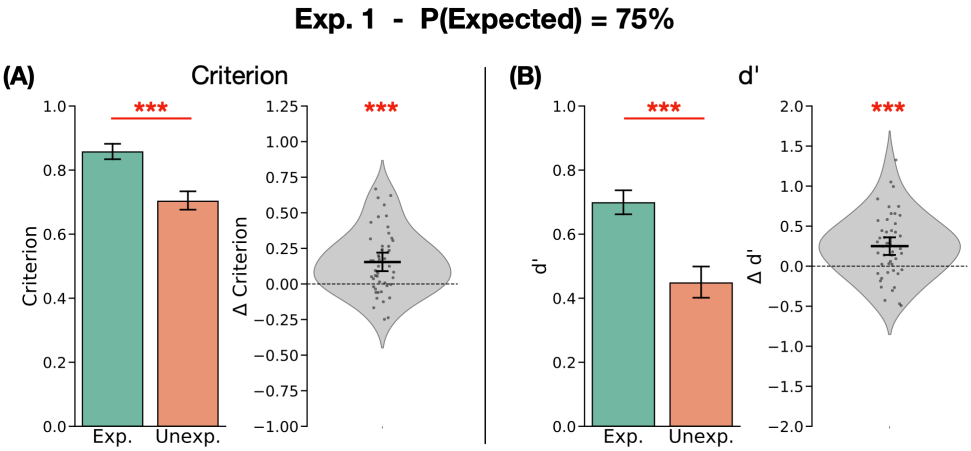


Figure 5.3. Results of Experiment 1. **(A)** Left – mean (and SEM) criterion for the Expected and Unexpected trials. Right – distribution of the differences between conditions (Expected – Unexpected) for each participant. **(B)** Same as in **A**, for d'. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

5.3 Results

5.3.1 Experiment 1: 75% Probability

In the first experiment, the object appeared in the expected view (given the scene context) on a majority (75%) of trials. Across conditions, participants' mean accuracy (and SEM) was 0.69 ± 0.01, indicating that they were able to perform the task, and that the staircase successfully converged to the desired accuracy of ~70%. Analyzing overall criterion, we found that it was significantly above zero (mean: 0.78, $t(49) = 36.34$, $p < 0.001$, $d = 5.14$, 95% CI = [0.74, 0.82]), indicating a strong general bias towards responding 'same', possibly due to the relatively small perceptual differences between the probes.

In comparing our central conditions, we found criterion to be significantly higher on Expected than Unexpected trials (means: 0.86 vs. 0.70; $t(49) = 4.96$, $p < 0.001$, $d = 0.82$, 95% CI = [0.09, 0.22]), as shown in **Figure 5.3A**. This indicates that when the object was shown with a size inconsistent with the surrounding scene's transformation, participants had a tendency to respond 'different' more often, counteracting their overall bias. This result suggests that the scene-driven size expectation affected participants' responses in the task, despite no explicit

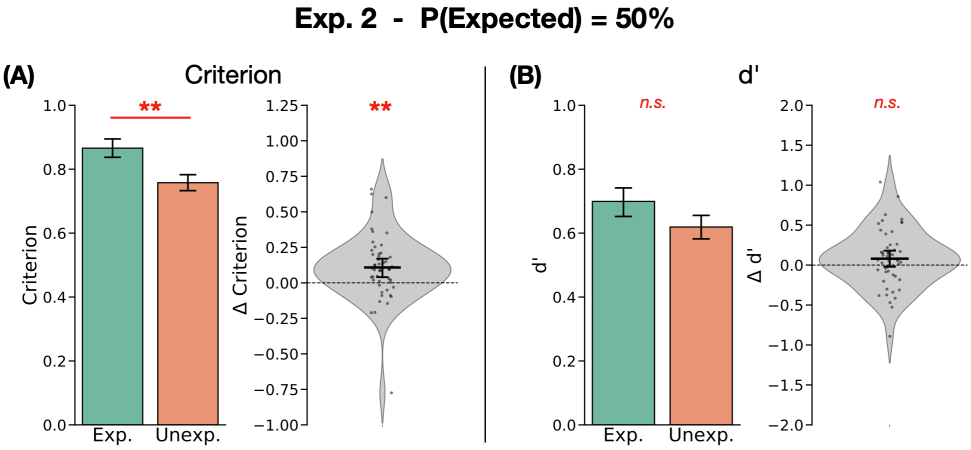


Figure 5.4. Results of Experiment 2. **(A)** Left – mean (and SEM) criterion for the Expected and Unexpected trials. Right – distribution of the differences between conditions (Expected – Unexpected) for each participant. **(B)** Same as in **A**, for d'. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

requirement to predict the object's size.

Moreover, we found d' to be significantly higher on Expected than Unexpected trials (means: 0.70 vs. 0.45; $t(49) = 4.48$, $p < 0.001$, $d = 0.80$, 95% CI = [0.14, 0.36]), as shown in **Figure 5.3B**. This indicates that scene-driven expectations, beyond influencing participants' responses in the task, affected their perceptual sensitivity as well, possibly by enabling them to set up a visual 'template' of the expected object size.

In this experiment, on a majority of trials, the object reappeared with a size that matched participants' scene-driven expectations. Thus, the short-term expectations established during the experiment matched the long-term expectations derived from real-world regularities (the fact that objects are transformed coherently with the surrounding scene). In Experiment 2, we investigated whether the interference of long-term expectations in our task would be reduced when expected and unexpected object sizes appear with the same probability during the experiment.

5.3.2 Experiment 2: 50% Probability

In the second experiment, the object would appear in an expected or unexpected size (given the background scene's viewpoint) with equal probability. Besides this probability manipulation, stimuli and experimental paradigm were the same as in Experiment 1. As in Experiment 1, participants were able to perform the task well above chance level (mean accuracy and SEM: 0.68 ± 0.01). Also consistently with the previous experiment, their criterion was significantly above zero (mean: 0.81,

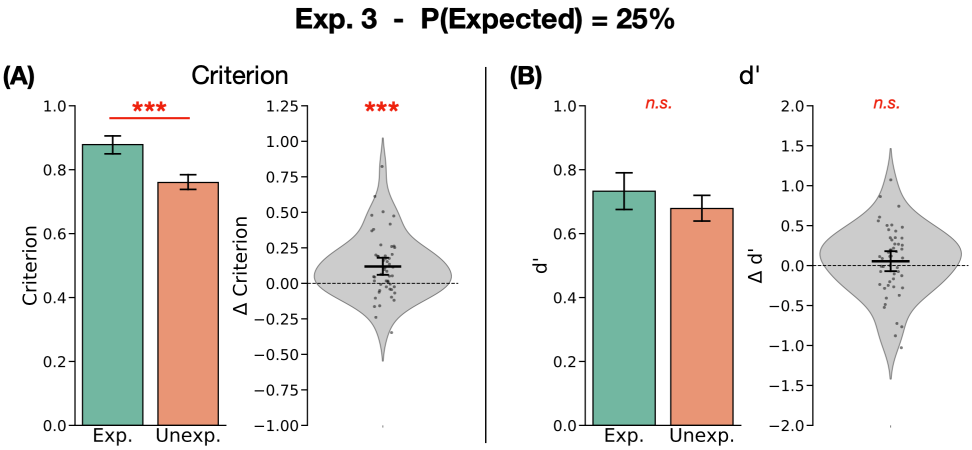


Figure 5.5. Results of Experiment 3. **(A)** Left – mean (and SEM) criterion for the Expected and Unexpected trials. Right – distribution of the differences between conditions (Expected – Unexpected) for each participant. **(B)** Same as in **A**, for d'. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$.

$t(49) = 38.09$, $p < 0.001$, $d = 5.39$, 95% CI = [0.77, 0.85]), meaning that they had a similarly strong bias towards reporting 'same' (i.e., no change between the two probes).

Comparing criterion between our main conditions of interest, we again found that it was higher on Expected than Unexpected trials (means: 0.87 vs. 0.76, $t(49) = 3.28$, $p = 0.002$, $d = 0.57$, 95% CI = [0.04, 0.17]), as shown in **Figure 5.4A**. This indicates that participants were still forming an expectation of the object size implied by the scene, and that this expectation still interfered with the context, despite the fact that it was not predictive of the actual stimuli that would be shown in the experiment.

Interestingly, contrary to the previous experiment we found no significant difference in d' between the Expected and Unexpected conditions (**Figure 5.4B**; means: 0.70 vs. 0.62, $t(49) = 1.56$, $p = 0.126$, $d = 0.27$, 95% CI = [-0.02, 0.18]). Thus, while the scene-driven expectation still influenced participants' answers, its violation did not result in a decrease in perceptual sensitivity. We address why this might have been the case in the Discussion.

The results of this experiment indicate that even when the scene-driven expectation of object size was only predictive on 50% of trials, it still interfered with participants' responses. Thus violating this expectation on a large proportion of experimental trials was still not sufficient to suppress it, providing evidence for its automaticity. In the following experiment, we asked whether further increasing the proportion of expectation violations, thereby making the scene context *counter*predictive, would suppress (or reverse) participants' predictions of object size.

	Experiment 1 Probability = 75%	**Experiment 2 Probability = 50%**	**Experiment 3 Probability = 25%**
Attention to Sequence 1-7 Likert scale	4.42 ± 0.19	4.30 ± 0.21	4.02 ± 0.23
Correlation with criterion	r = 0.07, p = 0.62, BF_{01} = 5.05	r = 0.06, p = 0.68, BF_{01} = 5.23	r = 0.08, p = 0.58, BF_{01} = 4.95
Correlation with d'	r = 0.10, p = 0.47, BF_{01} = 4.40	r = 0.04, p = 0.80, BF_{01} = 5.49	r = -0.06, p = 0.68, BF_{01} = 5.26
Object Prediction 1-7 Likert scale	3.53 ± 0.24	3.76 ± 0.22	3.32 ± 0.24
Correlation with criterion	r = -0.01, p = 0.93, BF_{01} = 5.59	r = 0.17, p = 0.23, BF_{01} = 2.79	r = -0.12, p = 0.39, BF_{01} = 4.03
Correlation with d'	r = 0.04, p = 0.76, BF_{01} = 5.35	r = 0.17, p = 0.23, BF_{01} = 2.86	r = -0.07, p = 0.63, BF_{01} = 5.10
Probability Estimate Percentage	56.96 ± 2.45	54.50 ± 2.39	55.84 ± 2.94
Correlation with criterion	r = -0.07, p = 0.63, BF_{01} = 5.08	r = 0.13, p = 0.36, BF_{01} = 3.79	r = -0.11, p = 0.43, BF_{01} = 4.22
Correlation with d'	r = 0.00, p = 0.99, BF_{01} = 5.68	r = -0.23, p = 0.11, BF_{01} = 1.66	r = 0.09, p = 0.53, BF_{01} = 4.81

Table 5.1: Mean responses (and SEM) to our final survey questions, and Pearson's r correlation with the behavioral effects (Expected – Unexpected trials) for both criterion and d'.

5.3.3 Experiment 3: 25% Probability

In this experiment, the object would reappear with its expected size only on 25% of trials. Apart from this, stimuli and paradigm were the same as in the previous two experiments. As in the previous experiments, participants performed the task well above chance level (mean accuracy and SEM: 0.69 ± 0.01). Also consistently with the previous experiments, they showed a strong overall bias towards responding 'same', leading to a significantly positive criterion (mean: 0.82, $t(50) = 37.57$, $p < 0.001$, $d = 5.26$, 95% CI = [0.78, 0.86]).

In comparing our central conditions, we again found a significant difference in criterion between Expected and Unexpected trials (**Figure 5.5A**; means: 0.88 vs. 0.76, $t(50) = 3.84$, $p < 0.001$, $d = 0.62$, 95% CI = [0.06, 0.18]). This indicates that even when a majority of trials contained inconsistent object sizes (relative to the scene background), participants still processed these inconsistently sized objects as

unexpected, leading to a response bias that is qualitatively similar to that of Experiments 1 and 2.

Comparing d' between Expected and Unexpected trials, we found no significant difference (means: 0.73 vs. 0.68, $t(50) = 0.90$, $p = 0.374$, $d = 0.15$, 95% CI = [-0.07, 0.18]), unlike Experiment 1 but consistently with Experiment 2. While scene-driven expectations still influenced par'icipants' responses, then, they did not lead to a decrease in perceptual sensitivity, confirming that this was unique to the condition in which they appeared on a minority of trials (Experiment 1).

5.3.4 Final survey data

The purpose of the final survey questions was to assess to what extent participants were aware of the experimental manipulation: how much they paid attention to the sequence of scene viewpoints before the target object appeared, how much they actively tried to predict the final object viewpoint, and their estimate of the probability of the object appearing with the expected size (see Methods for the actual questions).

Table 5.1 reports participants' mean responses for each of the questions, together with their correlations with the difference between Expected and Unexpected trials (in criterion and d') across subjects. Between-subject Welch ANOVAs revealed that none of the two Likert survey items significantly differed across experiments (Attention to Sequence: $F(2, 98.18) = 0.91$, $p = 0.406$, $\eta^2 = 0.013$; Object Prediction: $F(2, 97.67) = 0.96$, $p = 0.388$, $\eta^2 = 0.012$). This suggests that participants did not adopt a deliberate strategy of paying more or less attention to the scene, or actively trying to predict the object, depending on the probability of the prediction being accurate. Interestingly, their estimates of the probability of the object appearing with the expected size did not differ across experiments either ($F(2, 98.11) = 0.26$, $p = 0.774$, $\eta^2 = 0.003$), indicating that they were not tracking how often the contextual expectation was respected or violated, despite this expectation's impact on their responses.

Moreover, the responses of none of the survey questions correlated with the behavioral difference in either criterion or d' (**Table 5.1**). The behavioral effects we found, then, did not seem to depend on participants' awareness of the experimental manipulation or on their adoption of a specific behavioral strategy.

5.4 Discussion

Objects' retinal size in the real world depends on the distance from which they are viewed. Scene context provides a reference frame to represent that distance, and thus has the potential to guide our predictions of object size transformations. In the present work, we found that participants' responded differently in an orthogonal perceptual task, depending on whether the stimulus appeared with the size that was expected given the scene context. The influence of scene context

remained consistent even when the unexpected object size was shown on a majority of trials. This suggests that the effect was automatic, and primarily driven by expectations deriving from regularities of the real world, which overruled those observed during the experiment.

The effect of scene-driven size expectations, in our task, was reflected in a shift of criterion on trials in which the object did not match those expectations. While participants had a general bias towards reporting that the two probe views were the 'same' (possibly due to the generally small difference between them), this bias was reduced on Unexpected trials. Participants' expectations of object size, then, influenced their responses in this orthogonal task. Crucially, participants were robustly above chance in all experiments, as this was a precondition for inclusion in the analysis. This indicates that they were doing the task they were instructed to do, and not actively trying to predict the upcoming object size. Nevertheless, this prediction seems to have occurred automatically, influencing their responses. The fact that this effect was still present in Experiment 3 (25% Probability), when on most trials the scene-driven expectation was violated, provides further evidence that the formation of these expectations could not be overruled, even when the experimental setting rendered them counter predictive. In Experiment 1 (75% Probability), we additionally found that expectations affected perceptual sensitivity, with a significantly reduced d' on Unexpected trials. This suggests that when unexpected object sizes were presented infrequently, they had a more detrimental effect on perceptual sensitivity. This was possibly due to their lower frequency rendering them more salient, thus capturing attention away from the task, consistently with prior reports of involuntary attention capture being dependent on task context (Folk et al., 1992; S. W. Han & Marois, 2014). Overall, these results speak to automatic predictions of object size being driven by scene context, exerting an influence on an orthogonal task.

In the previous chapters, we found evidence that the appearance of an object from a novel viewpoint is automatically predicted based on scene cues. In those studies, similarly to the present one, the existence of scene-driven expectations was inferred from their influence on an orthogonal perceptual task. In both cases, the dimension along which the scene varied during the trial sequence was the same as the one along which the two probe views could vary in the task. In the previous studies, it was orientation around the vertical axis; in the present one, distance in depth. An interesting question for future research will be to clarify whether task interference only occurs when the overall change in the scene and the behavioral comparison happen along the same dimension. In explicit mental transformation studies, evidence suggests that participants manipulate an image-like representation in their minds (Cooper & Shepard, 1973; Koriat & Norman, 1984, 1988; Shepard & Cooper, 1982). If in our studies participants were actually creating a 'mental image' of the object at the expected size, we might expect task interference to generalize to other features of the object as well. On the other hand, it is possible that their predictions were more abstract,

only extrapolating along the changing feature of the scene and object. Research investigating object tracking under occlusion has found that only the location of the object is represented behind the occluder (Scholl & Pylyshyn, 1999; Teichmann et al., 2022). Given that tracking and prediction in motion and other feature spaces seem to share mechanisms (Blaser et al., 2000; Blaser & Sperling, 2008; Makin & Bertamini, 2014; Makin & Chauhan, 2014), it would be interesting to determine whether tracking features such as object orientation and scale can similarly involve an abstract prediction of the changing feature exclusively. In the real world, the way objects change with our viewpoint is not generally reducible to simple changes in retinotopic location: instead, it leads to complex transformations of the retinal image. It is therefore not trivial that these transformations could be 'abstracted away' from the representation of a particular object, similarly to changes in retinotopic location. A representation of visual transformations that abstracts away from specific object exemplars would allow to efficiently track and predict objects, while reducing the load on working memory, which is limited in the number of features it can track (Y. Xu & Franconeri, 2015). While there is evidence that we represent spatial transformations in a way that generalizes across specific objects (Ward et al., 2018; Mocz et al., 2021), whether these object-agnostic transformations underlie object tracking in naturalistic scenes will be an intriguing question for future research.

A theoretical idea closely related to that of object-agnostic spatial transformations is that of coordinate transformations as a ubiquitous process in visual cognition (Graf, 2006). According to Graf (2006), mental rotation, translation and scaling, beyond being involved in mental imagery-like tasks, are also involved in setting up *reference frames* for object perception across different orientations, positions and sizes. Once an observer has established an appropriate reference frame, for example of a specific orientation or size, recognition is facilitated for any object appearing with that orientation (Graf et al., 2005) or size (Larsen & Bundesen, 1978). Most relevantly to our findings, these spatial reference frames can be established by scene context (G. E. Hinton & Parsons, 1988; Humphrey & Jolicoeur, 1993; Christou et al., 2003). Our findings of scene-driven predictions of both orientation and size, then, could be explained by similar reference frame transformations. In the present study, only in Experiment 1 did we find a difference in perceptual sensitivity across Expected and Unexpected object sizes. Instead, scene-driven predictions consistently resulted in involuntary interference with task performance, with expectation violations influencing participants' response bias. A recent study (Gayet & Peelen, 2019) also found that perceived object size, driven by the depth cues in a scene, could capture participants' attention, providing further evidence that scene context can establish reference frames in an automatic way. Further research would be needed to determine how this automatic computation interacts with the structure of the task, leading to involuntary interference in some cases, and changes in perceptual sensitivity in others.

In conclusion, the present results show that beyond rotation, scene

viewpoint information can automatically drive predictions of object size, or translation in depth. This suggests that scene context might play a general role in providing a reference frame for different mental transformations of objects. Given the highly structured nature of our everyday environments, this might be an important mechanism supporting tracking and predicting objects in naturalistic vision.

Chapter 6
General Discussion

In the previous four chapters, I have described an experimental paradigm designed to investigate the nature of humans' internal representations of objects, and their interaction with external context. While this paradigm is clearly a very simplified 'little world' relative to the complexity of real-world environments and tasks, it exemplifies several interesting requirements for object representations to support everyday perception. I have discussed these in detail in the **Introduction**, but it is perhaps worth repeating them here, in light of our experimental findings:

a. Representations should behave like the objects they represent: it should be possible to predict external objects' transformations by transforming their internal representations. Here, we used 3D rotation as an example of a common spatial transformation that leads to non-trivial changes in an object's retinal image, and has been studied extensively in prior literature. In **Chapter 5**, to test the generality of our findings, we also looked at translation in depth (which leads to scaling of the retinal image).
b. Representations should be 'out there' in the world, seamlessly integrated with the visual scene rather than being projected on a mental canvas. As dealing with occluded objects is an ubiquituous example of this integration in the real world, we used occluded objects in our paradigm.
c. Representations should interact with the surrounding context in a way that mirrors the interactions between objects in the external world. We chose the ubiquitous constraint that objects tend to rotate or translate coherently with the scene that surrounds them.

In the Introduction, I mentioned two additional, more speculative points as possible directions for future research. I will return to them later. But first, I will examine some of the implications of our results, and the questions they leave unanswered.

6.1 Structure vs. Views

The distinction between *structure-based* and *view-based models* is a long-standing one in both visual object perception and spatial cognition. I have already discussed it in the **Introduction** and in **Chapters 2 & 3**, so I will not repeat it in detail here. Broadly, structure-based models posit that humans represent objects (and scenes) in terms of parts and the relationships between them, more or less independently of the observer's viewpoint (Ayzenberg & Behrmann, 2022; Biederman, 1987; Biederman & Gerhardstein, 1993; Hummel, 2000; Mou & McNamara, 2002; Rieser, 1989). According to view-based models, on the other hand, our representations are more akin to collections of views (Bülthoff & Edelman, 1992; Gillner & Mallot, 1998; Gootjes-Dreesbach et al., 2017; Tarr & Pinker, 1989; S. Ullman, 1998).

How is this debate related to our findings? These two classes of accounts could both underlie the influence of scene context on object transformations that we described. According to a structure-based account, invariant structural

relations, such as the relative orientation of the object with the scene's main axes could be extracted by observers, and the perceived distinction between Expected and Unexpected trials would be a violation of these relations, in a viewpoint-invariant fashion. In a view-based account, instead, participants would represent different viewpoints of the scene fundamentally as a collection of images, and would extrapolate to unseen views using a mental rotation-like operation (Tarr & Pinker, 1989). While we did not explicitly design our studies to investigate this distinction, a comparison, in **Chapter 2**, between object orientations that were aligned with the scene or not, revealed no difference in the magnitude of the expectancy effect on behavior. If participants were primarily relying on the orientation of the object relative to one of the scene's main axes (Mou & McNamara, 2002), objects that were aligned with those axes should show a stronger expectancy effect. Moreover, in the fMRI studies (**Chapters 3 & 4**), we found that the scene-driven expectation was reflected in decoding accuracy for the object's *proximal* shape. This provides substantial evidence that a viewer-centered representation was involved in participants' scene-driven predictions (although it might not have been the *only* representation involved).

Does this mean, then, that participants represent scenes as holistic, unstructured views? In fact, computational work has shown that it is at least possible, in principle, to perform a variety of tasks such as predicting novel views of a scene by representing scenes as single 'blocks', without any notion of individual objects or their relationships (Eslami et al., 2018; Muryy et al., 2020; Rosenbaum et al., 2018). However, given what we know about how humans parse the world into objects (Kahneman et al., 1992; Peters & Kriegeskorte, 2021), this 'fully unstructured' account seems unrealistic. If studies on the tracking of moving objects under occlusion are any indication, we likely maintain a representation of the object behind the occluder as a separate, stable unit (Scholl & Pylyshyn, 1999; Scholl & Flombaum, 2010; Green & Quilty-Dunn, 2020), albeit one that interacts with its context.

How can the scene representation be simultaneously structured and image-like? As for the original debates in object perception and spatial cognition, the answer here might be a combination of the two accounts. Several hybrid models have been proposed in which viewer-centered features are organized into structured, graph-like representations (G. Hinton, 1979; Hummel & Stankiewicz, 1998; Edelman & Intrator, 2001; D. Bear et al., 2020). These representations combine the best of both worlds: they can be directly extracted from visual input, while supporting compositional generalization (e.g. the notion that a circle is still a circle whether it is above or below a square, and that the relationship 'below' is the same whether it involves a circle below a square or vice versa). Moreover, while they contain information about distal relationships between objects in the scene, these representations remain bound to retinotopic locations in the visual field (something I will return to in the next section).

Our studies were designed to test for the *existence* of scene-driven object predictions, rather than their precise format or the mechanisms by which they

arise. This will be an interesting question for future research: how our representations balance the need to extract abstract, compositional scene structure and the need to infer this structure from visual cues, and bind it to locations in the visual field. We hope that our paradigm will be helpful in that, but also that more clever ones will be designed, and that more will be learned by comparing the predictions of different computational models. For example, different mechanisms for inferring scene structure from image cues might lead to unique distortions of the visual space representation (e.g. Svarverud et al., 2012) or parse the scene in systematically different ways depending on particular cues (such as cast shadows in humans, Yonas et al., 1978; Mamassian et al., 1998). More recent models specifically designed to parse visually rich real-world scenes (e.g. Bear et al., 2020) are of particular interest here.

6.2 Images vs. Reference frames

It is worth noting that the distinction between structure- and view-based models is distinct from the question of the *content* of scene-driven predictions. As mentioned in several of the previous chapters, our experimental designs do not allow to precisely determine what information about the object is predicted. Several studies investigating the tracking of occluded objects, for example, have found that only the object's location is represented (Scholl & Pylyshyn, 1999; Teichmann et al., 2022) leading some to believe that spatiotemporal information plays a unique role as a reference frame to bind other features to specific objects (Flombaum et al., 2009; Mitroff & Alvarez, 2007). While the special role of location has been called into question (Gordon et al., 2008; Quilty-Dunn & Green, 2021; Zhou et al., 2010), the consensus is that not all of an object's features are tracked during occlusion.

In a separate relevant line of research, Graf (2006) distinguished between the use of spatial transformations (such as rotation) in manipulating a specific object image (as in classical mental rotation studies, Shepard & Metzler, 1971; Shepard & Cooper, 1982; Koriat & Norman, 1984, 1988) and in establishing an abstract reference frame, that can facilitate processing of *any* object that is aligned with it (e.g. Graf et al., 2005; Jolicoeur, 1990; Larsen & Bundesen, 1978). He remained agnostic as to whether these processes happened in object- or viewer-centric reference frames. While, as explained above, tentative evidence points towards viewer-centered representations being involved in our studies, this does not necessarily equal a rich, picture-like representation encompassing all object features.

Our studies cannot distinguish between these two accounts: in the behavioral studies (**Chapters 2 & 5**), we did not investigate whether the behavioral difference between expected and unexpected object views generalizes to different object exemplars or categories. And in the training runs of the fMRI studies (**Chapters 3 & 4**) we grouped together objects by their proximal shape, to

maximize the power for detecting an expectation. This means that we could not test whether the prediction was object-specific, or an abstract reference frame (predicting *any object with that shape*). A recent study that, similarly, investigated the role of scene context in providing a spatial reference frame for internally generated object representations (Gayet & Peelen, 2022) found evidence for both the object's category and its size being represented. In that study, however, participants were engaged in an active visual search task, which might have led them to deliberately create an imagery-like template. In our studies, on the other hand, participants were not explicitly instructed to predict the object's appearance, and had to perform a task that, at least in principle, did not depend on doing so.

An intriguing question for future research will be to distinguish an image-like or reference-frame-like account of our findings, and more in general clarify how these two representations are used in making predictions based on scene structure in realistic scenarios. A particularly interesting possibility, already described in the Introduction, is that specific object features (those that can be 'explained away' by tracking the context, such as position and orientation) could be constrained to facilitate detection of scene changes along other dimensions. Something similar is believed to occur, for example, in object tracking (Bahrami, 2003; Flombaum & Scholl, 2006): when location can be successfully tracked, changes in features such as color can be more successfully detected.

6.3 Cognitive blackboards and 'out-there-ness'

In **Chapters 3 & 4**, we found that scene-driven object predictions were primarily reflected in early, retinotopic regions of visual cortex (early visual cortex, EVC). Is it possible that these regions play a specific role in representing the kind of predictions we observed? And what might this 'kind' be? While several areas in EVC, and particularly V1, have been implicated in several expectation- and context-related effects (e.g. Albers et al., 2013; Bosch et al., 2014; Christophel et al., 2015; Heilbron et al., 2020; Hindy et al., 2016; Kok et al., 2012; Murray et al., 2006; Smith & Muckli, 2010), others have reported effects primarily in higher-level object-selective cortex (e.g. Brandman & Peelen, 2017; Gayet & Peelen, 2022; Kaiser et al., 2021; Stokes et al., 2009). What might be the difference between these two groups of studies? Clearly, the dissociation might be related to idiosyncratic design choices due to the use of fMRI (Coutanche et al., 2016) or classes of stimuli particularly suited to studying one or the other level of processing (e.g. gratings vs. objects). However, another possibility is that these different tasks require different representational formats.

EVC is topographically organized in a way that mirrors the spatial organization of the retina. While this organization is somewhat maintained at higher levels of processing as well, and might in fact constitute a general organizing principle in the brain (Groen et al., 2022), EVC provides a particularly

detailed map of visual space. This had led some researchers to propose the idea of a 'cognitive blackboard' (Roelfsema & de Lange, 2016; Van Der Velde & De Kamps, 2002, 2006): EVC might store the intermediate results of cognitive operations that can benefit from spatially-organized representations. Locations in the visual field, then, could be 'tagged' with additional information resulting from downstream operations, such as figure-ground segmentation (e.g. Roelfsema et al., 2002). This extra information could be maintained separately from retinal bottom-up information, for example, by residing in different cortical layers (Self et al., 2013; Lawrence et al., 2018; Iamshchinina et al., 2021).

There is one aspect of this cognitive blackboard model that is of particular interest to the topics discussed here, and that to the best of my knowledge has not been previously discussed. A common spatial organization for external stimuli and internally generated signals seems ideal for the 'out-there-ness' I discussed in the Introduction. It provides a way for them to be seamlessly integrated, such that 'internal' representations can actually be 'out there' in the environment. Indeed, some of the tasks that I have highlighted as examples of this out-there-ness, such as object tracking under occlusion, have been found to involve retinotopically specific signals (e.g. Erlikhman & Caplovitz, 2017). Whether the predictions we observed here, deriving from high-level scene structure, also involve retinotopically specific representations, will be an interesting question for future studies, for example using retinotopic mapping in fMRI (Warnking et al., 2002; Wandell & Winawer, 2011). Moreover, techniques such as topographic connectivity (Knapen, 2021) could be used to determine whether retinotopic organization is a fundamental organizing principle in the interactions between brain areas that give rise to scene-driven predictions. This would provide insights into the mechanisms by which internal representations can be placed out there in the world, and interact with contextual information.

6.4 Linear and non-rigid transformations

While in our fMRI studies, through a decoding approach, we have analyzed the *outcome* of a predictive process (the expected object shape), we have not investigated the transformations that lead to that outcome. One important idea that I mentioned in the introduction was the conversion of behaviorally relevant transformations in the world into linear transformations in representational space. For example, neuroimaging work has found that object representations in visual cortex can be translated or scaled through a linear transformation (Ward et al., 2018; Mocz et al., 2021). This is not trivial, since these transformations lead to highly non-linear changes in pixel space. Moreover, transformations were found to generalize across objects, providing evidence for disentanglement of object identity from spatial transformations. By using similar techniques, future work could investigate whether object transformations are also represented as linear in our paradigm. Encoding scenes into a representation that transforms linearly

bears some resemblance to the structure-based models described for object perception: the reference frame is fundamentally scene-centric, and is thus abstracted away from viewer-centered space. An interesting question, then, is whether and how the advantages of representations that are bound to locations in retinotopic space can be combined with the ease of prediction and generalization afforded by linear transformations. This could be done by analyzing the representations learned by recent models that maintain binding to image-centered locations (such as Bear et al., 2020) to determine whether they are able to (approximately) represent complex scene transformations, such as rotations in 3D, as linear. This would provide a proof of principle that it is at least possible to combine the strengths of these two kinds of representations, while empirical work should investigate whether they are combined in the brain, or one or the other is preferably used depending on the task.

A separate question, also concerning transformations, is how interactions between scenes and objects can generalize to transformations beyond the simple, rigid ones studied here. In **Chapter 5**, we have found that translation in depth, or scaling, can be driven by scene context in a manner similar to rotation. In the case of non-rigid transformations, however, the relationship between objects and their context is less straightforward. While objects generally do rotate and translate coherently with the surrounding scene, the same can hardly be said about deformations, or changes in physical state, such as melting. It is worth considering, then, what realistic interactions between objects and context might look like for those transformations in the real world. Take the example of an object breaking into pieces. This kind of transformation is generally the result of local interventions, applied to a single object at a time, such as when someone pushes a vase off the table, causing it to break. These interactions (between the hand and the vase, and the vase and the floor) are localized, rather than involving the entirety of the scene. Moreover, breaking causes the object's parts to separate (while other non-rigid transformations, such as deformation, cause them to change their relative positions).

This calls for a fundamentally different representation from rigid transformations, in which objects are not represented as single units, but separated into parts. In fact, similar representations of objects in terms of connected parts have been found to be the best models of human observers' ability to predict physical dynamics (Bear et al., 2021; Han et al., 2022). An intriguing question for future research, then, is how representations of objects and their interactions switch according to the transformations that need to be predicted in a given task or context. Previous research has found that observers can flexibly switch between different representations of objects, for example according to whether an object is perceived as deformable or not (Kourtzi & Shiffrar, 2001) or whether the task involves predicting its physical dynamics or recognizing its shape (Y. Li et al., 2022). In general, evidence suggests that in predicting the dynamics of the world, humans have a remarkable ability to use the most efficient representation to only predict what is necessary (Ullman et al.,

2017). Understanding what the mechanisms underlying this representational switching are, and how it can go beyond single objects to the level of whole scenes, will be a crucial step towards generalizing the results presented here to more complex, realistic scenarios.

6.5 Coda: scene graphs

We are now close to the end of this Discussion, and of this thesis. But we still haven't addressed the last two points mentioned in the Introduction:

d. A representation known in computer graphics as the *scene graph* is particularly well-suited for a ubiquitous structure in real-world scenes: hierarchical part-whole relations. Much work in AI has addressed the problem of how this kind of representation can be learned from data. It remains unknown, however, whether it provides a model of how humans represent complex scenes, although some tentative evidence goes in that direction.
e. What are these representations for? Scene graphs were developed for the goal of accurately, and efficiently, simulating complex scenes. But our brains do not serve the purpose of generating a perfectly accurate simulation of the world. Internal representations should ultimately support behavior in real-world tasks. I have argued that the purpose of scene graph-like representations is the real-time tracking of hierarchically structured scenes.

How do the results presented in the previous chapters fit into this picture? Let's start with point (d). Our studies stripped down the real world to a simple world that only comprises a main object and a background scene. It is thus not possible to tell whether participants in our studies represented scenes as part-whole hierarchies, with the object as part of the scene, or as 'flat' relations, for example representing that the walls or background objects should rotate together with the object. Relatedly, we do not know whether these contextual effects on predictions are contingent on showing extended scenes, or whether any contextual information can similarly drive predictions. The long list of contextual influences on predictions reviewed in the Introduction (e.g. Shepard & Zare, 1983; Kim et al., 2012; Little & Firestone, 2021; Heptulla Chatterjee et al., 1996; He & Nakayama, 1994) shows that what constitutes a 'context' can be extremely flexible. One advantage of hierarchical graph representations is that they can accommodate such flexibility. What constitutes an 'object' in such a representation, for example, depends on the way that the scene is parsed hierarchically (Feldman, 2003). In whichever way participants perceptually organized the scenes in our studies, then, this representation might from that of scenes with a deeper hierarchical structure. Whether there exists a unified representational format that can accommodate these different situations will be a key question for future experimental, computational and theoretical work.

Regarding point (e), again, the little world we used in our paradigm was too simplified to investigate how the context-driven prediction of object transformations can aid real-world behavior. In the Introduction, I have speculated that the main role of hierarchical scene representations is to support dynamic tracking in complex, structured scenes. So far, the main line of evidence for the use of structured representations in dynamic scenes comes from studies of tracking in hierarchical motion displays (H. Xu et al., 2017; Bill et al., 2020, 2021). An exciting avenue for future studies could be to devise paradigms that combine realistic scenes, like the ones we have used here, with similar dynamic tasks. By leveraging tools such as game engines, for example, it would be possible to dynamically vary different properties of hierarchical scenes to clarify the content of humans' internal representations of these scenes. In parallel, computational work would enable us to make specific predictions as to how specific implementations of the general principles outlined here affect behavior in these dynamic tasks. Overall, I hope that the little world presented in this thesis was one little step towards clarifying how humans represent complex, real-world scenes.

References

Achille, A., & Soatto, S. (2018). Emergence of invariance and disentanglement in deep representations. *The Journal of Machine Learning Research*, *19*(1), 1947–1980.

Albers, A. M., Kok, P., Toni, I., Dijkerman, H. C., & de Lange, F. P. (2013). Shared representations for working memory and mental imagery in early visual cortex. *Current Biology: CB*, *23*(15), 1427–1431. https://doi.org/10.1016/j.cub.2013.05.065

Aldegheri, G., Gayet, S., & Peelen, M. V. (2023). Scene context automatically drives predictions of object transformations. *Cognition*, *238*, 105521.

Arguin, M., & Saumier, D. (2000). Conjunction and linear non-separability effects in visual shape encoding. *Vision Research*, *40*(22), 3099–3115. https://doi.org/10.1016/s0042-6989(00)00155-3

Ayzenberg, V., & Behrmann, M. (2022). The dorsal visual pathway represents object-centered spatial relations for object recognition. *Journal of Neuroscience*, *42*(23), 4693–4710.

Bahrami, B. (2003). Object property encoding and change blindness in multiple object tracking. *Visual Cognition*, *10*(8), 949–963.

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. *Trends in Cognitive Sciences*.

Baker, C., Keysers, C., Jellema, T., Wicker, B., & Perrett, D. (2001). Neuronal representation of disappearing and hidden objects in temporal cortex of the macaque. *Experimental Brain Research*, *140*(3), 375–381.

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

Bar-Zeev, A. (2007). *Scenegraphs: Past, Present, and Future | Reality Prime*. https://web.archive.org/web/20190207215718/http://www.realityprime.com/blog/2007/06/scenegraphs-past-present-and-future/

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695–711.

Bates, C. J., Yildirim, I., Tenenbaum, J. B., & Battaglia, P. (2019). Modeling human intuitions about liquid flow with particle-based simulation. *PLoS Computational Biology*, *15*(7), e1007210.

Battaglia, P. W., Hamrick, J. B., Bapst, V., Sanchez-Gonzalez, A., Zambaldi, V., Malinowski, M., Tacchetti, A., Raposo, D., Santoro, A., Faulkner, R., Gulcehre, C., Song, F., Ballard, A., Gilmer, J., Dahl, G., Vaswani, A., Allen, K., Nash, C., Langston, V., … Pascanu, R. (2018). *Relational inductive biases, deep learning, and graph networks* (arXiv:1806.01261). arXiv. https://doi.org/10.48550/arXiv.1806.01261

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332.

Battistoni, E., Stein, T., & Peelen, M. V. (2017). Preparatory attention in visual cortex. *Annals of the New York Academy of Sciences*, *1396*(1), 92–107. https://doi.org/10.1111/nyas.13320

Bear, D., Fan, C., Mrowca, D., Li, Y., Alter, S., Nayebi, A., Schwartz, J., Fei-Fei, L., Wu, J., Tenenbaum, J., & Yamins, Daniel LK. (2020). Learning physical graph representations from visual scenes. *Advances in Neural Information Processing Systems*, *33*, 6027–6039.

Bear, D. M., Wang, E., Mrowca, D., Binder, F. J., Tung, H.-Y. F., Pramod, R. T., Holdaway, C., Tao, S., Smith, K., & Sun, F.-Y. (2021). Physion: Evaluating physical prediction from vision in humans and machines. *ArXiv Preprint ArXiv:2106.08261*.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation Learning: A Review and New Perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798–1828. https://doi.org/10.1109/TPAMI.2013.50

Bennett, D. (2002). Evidence for a pre-match 'mental translation' on a form-matching task. *Journal of Vision*, *2*(7), 50. https://doi.org/10.1167/2.7.50

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115.

Biederman, I., & Gerhardstein, P. C. (1993). Recognizing depth-rotated objects: Evidence and conditions for three-dimensional viewpoint invariance. *Journal of Experimental Psychology: Human Perception and Performance*, *19*(6), 1162.

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

Bill, J., Gershman, S. J., & Drugowitsch, J. (2021). Structure in motion: Visual motion perception as online hierarchical inference. *BioRxiv*.

Bill, J., Pailian, H., Gershman, S. J., & Drugowitsch, J. (2020). Hierarchical structure is employed by humans during visual motion perception. *Proceedings of the National Academy of Sciences*, *117*(39), 24581–24589.

Blaser, E., Pylyshyn, Z. W., & Holcombe, A. O. (2000). Tracking an object through feature space. *Nature*, *408*(6809), 196–199.

Blaser, E., & Sperling, G. (2008). When is Motion 'Motion'? *Perception*, *37*(4), 624–627. https://doi.org/10.1068/p5812

Booth, M. C., & Rolls, E. T. (1998). View-invariant representations of familiar objects by neurons in the inferior temporal visual cortex. *Cerebral Cortex (New York, NY: 1991)*, *8*(6), 510–523.

Bosch, S. E., Jehee, J. F. M., Fernández, G., & Doeller, C. F. (2014). Reinstatement of Associative Memories in Early Visual Cortex Is Signaled by the Hippocampus. *Journal of Neuroscience*, *34*(22), 7493–7500. https://doi.org/10.1523/JNEUROSCI.0805-14.2014

Bouchacourt, D., Ibrahim, M., & Deny, S. (2021). *Addressing the Topological Defects of Disentanglement via Distributed Operators* (arXiv:2102.05623). arXiv. https://doi.org/10.48550/arXiv.2102.05623

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

Brandman, T., & Peelen, M. V. (2017). Interaction between scene and object processing revealed by

human fMRI and MEG decoding. *Journal of Neuroscience*, *37*(32), 7700–7710.

Bülthoff, H. H., & Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, *89*(1), 60–64.

Bundesen, C., & Larsen, A. (1975). Visual transformation of size. *Journal of Experimental Psychology: Human Perception and Performance*, *1*, 214–220. https://doi.org/10.1037/0096-1523.1.3.214

Bundesen, C., Larsen, A., & Farrell, J. E. (1983). Visual Apparent Movement: Transformations of Size and Orientation. *Perception*, *12*(5), 549–558. https://doi.org/10.1068/p120549

Burgess, N. (2006). Spatial memory: How egocentric and allocentric combine. *Trends in Cognitive Sciences*, *10*(12), 551–557.

Burgess, N., Spiers, H. J., & Paleologou, E. (2004). Orientational manoeuvres in the dark: Dissociating allocentric and egocentric influences on spatial memory. *Cognition*, *94*(2), 149–166.

Burke, L. (1952). On the tunnel effect. *Quarterly Journal of Experimental Psychology*, *4*(3), 121–138. https://doi.org/10.1080/17470215208416611

Burnston, D. C. (2021). Contents, vehicles, and complex data analysis in neuroscience. *Synthese*, *199*(1), 1617–1639. https://doi.org/10.1007/s11229-020-02831-9

Byrne, P., Becker, S., & Burgess, N. (2007). Remembering the past and imagining the future: A neural model of spatial memory and imagery. *Psychological Review*, *114*(2), 340.

Carlton, E. H., & Shepard, R. N. (1990a). Psychologically simple motions as geodesic paths I. Asymmetric objects. *Journal of Mathematical Psychology*, *34*(2), 127–188. https://doi.org/10.1016/0022-2496(90)90001-P

Carlton, E. H., & Shepard, R. N. (1990b). Psychologically simple motions as geodesic paths II. Symmetric objects. *Journal of Mathematical Psychology*, *34*(2), 189–228. https://doi.org/10.1016/0022-2496(90)90002-Q

Carrigan, S. B., Palmer, E. M., & Kellman, P. J. (2016). Differentiating global and local contour completion using a dot localization paradigm. *Journal of Experimental Psychology: Human Perception and Performance*, *42*, 1928–1946. https://doi.org/10.1037/xhp0000233

Chang, L., & Tsao, D. Y. (2017). The Code for Facial Identity in the Primate Brain. *Cell*, *169*(6), 1013-1028.e14. https://doi.org/10.1016/j.cell.2017.05.011

Chang, X., Ren, P., Xu, P., Li, Z., Chen, X., & Hauptmann, A. (2021). Scene graphs: A survey of generations and applications. *ArXiv Preprint ArXiv:2104.01111*.

Cheadle, S., Egner, T., Wyart, V., Wu, C., & Summerfield, C. (2015). Feature expectation heightens visual sensitivity during fine orientation discrimination. *Journal of Vision*, *15*(14), 14–14.

Chen, C., Wu, Y., Dai, Q., Zhou, H.-Y., Xu, M., Yang, S., Han, X., & Yu, Y. (2022). *A Survey on Graph Neural Networks and Graph Transformers in Computer Vision: A Task-Oriented Perspective* (arXiv:2209.13232). arXiv. https://doi.org/10.48550/arXiv.2209.13232

Chen, Y.-C., & Scholl, B. J. (2016). The Perception of History: Seeing Causal History in Static Shapes

Induces Illusory Motion Perception. *Psychological Science*, *27*(6), 923–930. https://doi.org/10.1177/0956797616628525

Chomsky, N. (1957). Syntactic Structures (The Hague: Mouton, 1957). *Review of Verbal Behavior by BF Skinner, Language*, *35*, 26–58.

Christophel, T. B., Cichy, R. M., Hebart, M. N., & Haynes, J.-D. (2015). Parietal and early visual cortices encode working memory content across mental transformations. *Neuroimage*, *106*, 198–206.

Christou, C. G., Tjan, B. S., & Bülthoff, H. H. (2003). Extrinsic cues aid shape recognition from novel viewpoints. *Journal of Vision*, *3*(3), 1. https://doi.org/10.1167/3.3.1

Clark, A. (2015). Surfing uncertainty: Prediction, action, and the embodied mind. Oxford University Press.

Cohen, T. S., & Welling, M. (2015). *Transformation Properties of Learned Visual Representations* (arXiv:1412.7659). arXiv. https://doi.org/10.48550/arXiv.1412.7659

Cooper, L. A. (1976). Demonstration of a mental analog of an external rotation. *Perception & Psychophysics*, *19*(4), 296–302.

Cooper, L. A., & Shepard, R. N. (1973). CHRONOMETRIC STUDIES OF THE ROTATION OF MENTAL IMAGES. In W. G. Chase (Ed.), *Visual Information Processing* (pp. 75–176). Academic Press. https://doi.org/10.1016/B978-0-12-170150-5.50009-3

Corbetta, M., Miezin, F. M., Dobmeyer, S., Shulman, G. L., & Petersen, S. E. (1990). Attentional modulation of neural processing of shape, color, and velocity in humans. *Science*, *248*(4962), 1556–1559.

Cortese, J. M., & Dyre, B. P. (1996). Perceptual similarity of shapes generated from fourier descriptors. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 133–143. https://doi.org/10.1037/0096-1523.22.1.133

Coutanche, M. N., Solomon, S. H., & Thompson-Schill, S. L. (2016). A meta-analysis of fMRI decoding: Quantifying influences on human visual population codes. *Neuropsychologia*, *82*, 134–141. https://doi.org/10.1016/j.neuropsychologia.2016.01.018

Craik, K. J. W. (1943). *The nature of explanation* (pp. viii, 123). University Press, Macmillan.

Cunningham, S., & Bailey, M. J. (2001). Lessons from scene graphs: Using scene graphs to teach hierarchical modeling. *Computers & Graphics*, *25*(4), 703–711. https://doi.org/10.1016/S0097-8493(01)00099-1

Cutzu, F., & Edelman, S. (1996). Faithful representation of similarities among three-dimensional shapes in human vision. *Proceedings of the National Academy of Sciences*, *93*(21), 12046–12050. https://doi.org/10.1073/pnas.93.21.12046

Dado, T., Güçlütürk, Y., Ambrogioni, L., Ras, G., Bosch, S., van Gerven, M., & Güçlü, U. (2022). Hyperrealistic neural decoding for reconstructing faces from fMRI activations via the GAN latent space. *Scientific Reports*, *12*(1), 1–9.

De Freitas, J., Myers, N. E., & Nobre, A. C. (2016). Tracking the changing feature of a moving object. *Journal of Vision*, *16*(3), 22. https://doi.org/10.1167/16.3.22

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*(9), 764–779.

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12.

Deco, G., & Rolls, E. T. (2004). A Neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, *44*(6), 621–642. https://doi.org/10.1016/j.visres.2003.09.037

Deng, F., Zhi, Z., Lee, D., & Ahn, S. (2020). Generative scene graph networks. *International Conference on Learning Representations*.

DiCarlo, J. J., & Cox, D. D. (2007). Untangling invariant object recognition. *Trends in Cognitive Sciences*, *11*(8), 333–341. https://doi.org/10.1016/j.tics.2007.06.010

Dijkstra, N., Ambrogioni, L., Vidaurre, D., & van Gerven, M. (2020). Neural dynamics of perceptual inference and its reversal during imagery. *ELife*, *9*, e53588. https://doi.org/10.7554/eLife.53588

Dijkstra, N., Bosch, S. E., & van Gerven, M. A. (2019). Shared neural mechanisms of visual perception and imagery. *Trends in Cognitive Sciences*, *23*(5), 423–434.

Dorrell, W., Latham, P. E., Behrens, T. E. J., & Whittington, J. C. R. (2022). *Actionable Neural Representations: Grid Cells from Minimal Constraints* (arXiv:2209.15563). arXiv. https://doi.org/10.48550/arXiv.2209.15563

Du, Y., Liu, Z., Basevi, H., Leonardis, A., Freeman, B., Tenenbaum, J., & Wu, J. (2018). Learning to Exploit Stability for 3D Scene Parsing. *Advances in Neural Information Processing Systems*, *31*. https://proceedings.neurips.cc/paper/2018/hash/43feaeeecd7b2fe2ae2e26d917b6477d-Abstract.html

Dulberg, Z., & Cohen, J. (2020). Learning canonical transformations. *ArXiv Preprint ArXiv:2011.08822*.

Edelman, S. (1998). Representation is representation of similarities. *Behavioral and Brain Sciences*, *21*(4), 449–467. https://doi.org/10.1017/S0140525X98001253

Edelman, S., & Intrator, N. (2001). A productive, systematic framework for the representation of visual structure. *Advances in Neural Information Processing Systems*, 10–16.

Emonds, A. M., Srinath, R., Nielsen, K. J., & Connor, C. E. (2022). Object representation in a gravitational reference frame. *BioRxiv*.

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601.

Erdogan, G., & Jacobs, R. A. (2017). Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychological Review*, *124*(6), 740.

Erlikhman, G., & Caplovitz, G. P. (2017). Decoding information about dynamically occluded objects in visual cortex. *NeuroImage*, *146*, 778–788.

Eslami, S. A., Jimenez Rezende, D., Besse, F., Viola, F., Morcos, A. S., Garnelo, M., Ruderman, A., Rusu, A. A., Danihelka, I., Gregor, K., & others. (2018). Neural scene representation and

rendering. *Science*, *360*(6394), 1204–1210.

Exner, S. (1876). Uber das Sehen von Bewegungen und die Theorie des Zusammengestzen. *Auges. Sber. Akad. Wiss. Wein (Math.-Nat. Kl.)*, *72*, 157–191.

Fantoni, C., & Gerbino, W. (2003). Contour interpolation by vector-field combination. *Journal of Vision*, *3*(4), 4. https://doi.org/10.1167/3.4.4

Fekete, T. (2010). Representational Systems. *Minds and Machines*, *20*(1), 69–101. https://doi.org/10.1007/s11023-009-9166-2

Feldman, J. (2003). What is a visual object? *Trends in Cognitive Sciences*, *7*(6), 252–256. https://doi.org/10.1016/s1364-6613(03)00111-6

Flesch, T., Balaguer, J., Dekker, R., Nili, H., & Summerfield, C. (2018). Comparing continual task learning in minds and machines. *Proceedings of the National Academy of Sciences*, *115*(44), E10313–E10322. https://doi.org/10.1073/pnas.1800755115

Flombaum, J. I., Kundey, S. M., Santos, L. R., & Scholl, B. J. (2004). Dynamic object individuation in rhesus macaques: A study of the tunnel effect. *Psychological Science*, *15*(12), 795–800. https://doi.org/10.1111/j.0956-7976.2004.00758.x

Flombaum, J. I., & Scholl, B. J. (2006). A temporal same-object advantage in the tunnel effect: Facilitated change detection for persisting objects. *Journal of Experimental Psychology. Human Perception and Performance*, *32*(4), 840–853. https://doi.org/10.1037/0096-1523.32.4.840

Flombaum, J. I., Scholl, B. J., & Santos, L. R. (2009). Spatiotemporal priority as a fundamental principle of object persistence. *The Origins of Object Knowledge*, 135–164.

Folk, C. L., Remington, R. W., & Johnston, J. C. (1992). Involuntary covert orienting is contingent on attentional control settings. *Journal of Experimental Psychology. Human Perception and Performance*, *18*(4), 1030–1044.

Foster, D. H. (1975). Visual apparent motion and some preferred paths in the rotation group SO (3). *Biological Cybernetics*, *18*(2), 81–89.

Foster, D. H., & Gilson, S. J. (2002). Recognizing novel three-dimensional objects by summing signals from parts and views. *Proceedings of the Royal Society B: Biological Sciences*, *269*(1503), 1939–1947. https://doi.org/10.1098/rspb.2002.2119

Franz, M. O., Schölkopf, B., Georg, P., Mallot, H. A., & Bülthoff, H. H. (1997). Learning view graphs for robot navigation. *Proceedings of the First International Conference on Autonomous Agents*, 138–147.

Franz, M. O., Schölkopf, B., Mallot, H. A., & Bülthoff, H. H. (1998). Where did I take that snapshot? Scene-based homing by image matching. *Biological Cybernetics*, *79*(3), 191–202.

Freyd, J. J. (1983). Representing the dynamics of a static form. *Memory & Cognition*, *11*(4), 342–346.

Freyd, J. J., & Finke, R. A. (1984). Representational momentum. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 126–132. https://doi.org/10.1037/0278-7393.10.1.126

Fujita, I. (2002). The inferior temporal cortex: Architecture, computation, and representation. *Journal*

of Neurocytology, *31*(3), 359–371.

Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, *36*(4), 193–202. https://doi.org/10.1007/BF00344251

Galati, G., Pelle, G., Berthoz, A., & Committeri, G. (2010). Multiple reference frames used by the human brain for spatial perception and memory. *Experimental Brain Research*, *206*(2), 109–120. https://doi.org/10.1007/s00221-010-2168-8

Gayet, S., Paffen, C. L., & Van der Stigchel, S. (2018). Visual working memory storage recruits sensory processing areas. *Trends in Cognitive Sciences*, *22*(3), 189–190.

Gayet, S., & Peelen, M. V. (2019). Scenes modulate object processing before interacting with memory templates. *Psychological Science*, *30*(10), 1497–1509.

Gayet, S., & Peelen, M. V. (2022). Preparatory attention incorporates contextual expectations. *Current Biology*, *32*(3), 687–692.

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*(6), 711–724. https://doi.org/10.1016/s0042-6989(00)00277-7

Geisler, W. S., & Super, B. J. (2000). Perceptual organization of two-dimensional patterns. *Psychological Review*, *107*(4), 677.

Gillner, S., & Mallot, H. A. (1998). Navigation and acquisition of spatial knowledge in a virtual maze. *Journal of Cognitive Neuroscience*, *10*(4), 445–463.

Gklezakos, D. C., & Rao, R. P. (2022). Active Predictive Coding Networks: A Neural Solution to the Problem of Learning Reference Frames and Part-Whole Hierarchies. *ArXiv Preprint ArXiv:2201.08813*.

Glennerster, A. (2016). A moving observer in a three-dimensional world. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *371*(1697), 20150265. https://doi.org/10.1098/rstb.2015.0265

Gold, J. M., Murray, R. F., Bennett, P. J., & Sekuler, A. B. (2000). Deriving behavioural receptive fields for visually completed contours. *Current Biology*, *10*(11), 663–666. https://doi.org/10.1016/S0960-9822(00)00523-6

Goldstone, R. L. (1996). Alignment-based nonmonotonicities in similarity. *Journal of Experimental Psychology. Learning, Memory, and Cognition*, *22*(4), 988–1001. https://doi.org/10.1037//0278-7393.22.4.988

Goodman, N. (1976). Languages of art: An approach to a theory of symbols. Hackett publishing.

Gootjes-Dreesbach, L., Pickup, L. C., Fitzgibbon, A. W., & Glennerster, A. (2017). Comparison of view-based and reconstruction-based models of human navigational strategy. *Journal of Vision*, *17*(9), 11–11.

Gordon, R. D., Vollmer, S. D., & Frankl, M. L. (2008). Object continuity and the transsaccadic representation of form. *Perception & Psychophysics*, *70*(4), 667–679.

Gorgolewski, K., Burns, C. D., Madison, C., Clark, D., Halchenko, Y. O., Waskom, M. L., & Ghosh, S.

S. (2011). Nipype: A flexible, lightweight and extensible neuroimaging data processing framework in python. *Frontiers in Neuroinformatics*, 13.

Goroshin, R., Mathieu, M. F., & LeCun, Y. (2015). Learning to Linearize Under Uncertainty. *Advances in Neural Information Processing Systems*, *28*. https://proceedings.neurips.cc/paper/2015/hash/eefc9e10ebdc4a2333b42b2dbb8f27b6-Abstract.html

Graf, M. (2006). Coordinate transformations in object recognition. *Psychological Bulletin*, *132*, 920–945. https://doi.org/10.1037/0033-2909.132.6.920

Graf, M., Kaping, D., & Bülthoff, H. H. (2005). Orientation congruency effects for familiar objects: Coordinate transformations in object recognition. *Psychological Science*, *16*(3), 214–221. https://doi.org/10.1111/j.0956-7976.2005.00806.x

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., & Parkkonen, L. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, 267.

Granskog, J., Schnabel, T. N., Rousselle, F., & Novák, J. (2021). Neural scene graph rendering. *ACM Transactions on Graphics (TOG)*, *40*(4), 1–11.

Gray, D., & Tao, H. (2008). Viewpoint Invariant Pedestrian Recognition with an Ensemble of Localized Features. In D. Forsyth, P. Torr, & A. Zisserman (Eds.), *Computer Vision – ECCV 2008* (pp. 262–275). Springer. https://doi.org/10.1007/978-3-540-88682-2_21

Green, E. J., & Quilty-Dunn, J. (2020). What is an object file? *The British Journal for the Philosophy of Science*.

Greff, K., van Steenkiste, S., & Schmidhuber, J. (2020). *On the Binding Problem in Artificial Neural Networks* (arXiv:2012.05208). arXiv. https://doi.org/10.48550/arXiv.2012.05208

Groen, I. I. A., Dekker, T. M., Knapen, T., & Silson, E. H. (2022). Visuospatial coding as ubiquitous scaffolding for human cognition. *Trends in Cognitive Sciences*, *26*(1), 81–96. https://doi.org/10.1016/j.tics.2021.10.011

Guan, C., & Firestone, C. (2020). Seeing what's possible: Disconnected visual parts are confused for their potential wholes. *Journal of Experimental Psychology: General*, *149*, 590–598. https://doi.org/10.1037/xge0000658

Hafri, A., Boger, T., & Firestone, C. (2022). Melting ice with your mind: Representational momentum for physical states. *Psychological Science*, *33*(5), 725–735.

Hafri, A., Bonner, M. F., Landau, B., & Firestone, C. (2020). *A phone in a basket looks like a knife in a cup: The perception of abstract relations*. PsyArXiv. https://doi.org/10.31234/osf.io/jx4yg

Hafri, A., & Firestone, C. (2021). The perception of relations. *Trends in Cognitive Sciences*, *25*(6), 475–492.

Hahn, U., Close, J., & Graf, M. (2009). Transformation direction influences shape-similarity judgments. *Psychological Science*, *20*(4), 447–454. https://doi.org/10.1111/j.1467-9280.2009.02310.x

Hamrick, J. B. (2019). Analogues of mental simulation and imagination in deep learning. *Current Opinion in Behavioral Sciences*, *29*, 8–16. https://doi.org/10.1016/j.cobeha.2018.12.011

Hamrick, J. B., & Griffiths, T. (2014). What to simulate? Inferring the right direction for mental rotation. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *36*(36). https://escholarship.org/uc/item/064367d4

Han, J., Huang, W., Ma, H., Li, J., Tenenbaum, J. B., & Gan, C. (2022). Learning Physical Dynamics with Subequivariant Graph Neural Networks. *ArXiv Preprint ArXiv:2210.06876*.

Han, S. W., & Marois, R. (2014). The effects of stimulus-driven competition and task set on involuntary attention. *Journal of Vision*, *14*(7), 14. https://doi.org/10.1167/14.7.14

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37–53.

Harris, C. R., Millman, K. J., Van Der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., & Smith, N. J. (2020). Array programming with NumPy. *Nature*, *585*(7825), 357–362.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d′. *Behavior Research Methods, Instruments, & Computers*, *27*(1), 46–51.

Hawkins, J., Ahmad, S., & Cui, Y. (2017). A Theory of How Columns in the Neocortex Enable Learning the Structure of the World. *Frontiers in Neural Circuits*, *11*. https://www.frontiersin.org/articles/10.3389/fncir.2017.00081

Hayward, W. G. (2003). After the viewpoint debate: Where next in object recognition? *Trends in Cognitive Sciences*, *7*(10), 425–427. https://doi.org/10.1016/j.tics.2003.08.004

He, Z. J., & Nakayama, K. (1994). Apparent motion determined by surface layout not by disparity or three-dimensional distance. *Nature*, *367*(6459), Article 6459. https://doi.org/10.1038/367173a0

Hegarty, M. (2004). Mechanical reasoning by mental simulation. *Trends in Cognitive Sciences*, *8*(6), 280–285. https://doi.org/10.1016/j.tics.2004.04.001

Heilbron, M., Richter, D., Ekman, M., Hagoort, P., & de Lange, F. P. (2020). Word contexts enhance the neural representation of individual letters in early visual cortex. *Nature Communications*, *11*(1), Article 1. https://doi.org/10.1038/s41467-019-13996-4

Hénaff, O. J., Goris, R. L. T., & Simoncelli, E. P. (2019). Perceptual straightening of natural videos. *Nature Neuroscience*, *22*(6), Article 6. https://doi.org/10.1038/s41593-019-0377-4

Heptulla Chatterjee, S., Freyd, J. J., & Shiffrar, M. (1996). Configural processing in the perception of apparent biological motion. *Journal of Experimental Psychology: Human Perception and Performance*, *22*, 916–929. https://doi.org/10.1037/0096-1523.22.4.916

Heywood-Everett, E., Baker, D. H., & Hartley, T. (2022). Testing the precision of spatial memory representations using a change-detection task: Effects of viewpoint change. *Journal of Cognitive Psychology*, *34*(1), 127–141.

Higgins, I., Amos, D., Pfau, D., Racaniere, S., Matthey, L., Rezende, D., & Lerchner, A. (2018). *Towards a Definition of Disentangled Representations* (arXiv:1812.02230). arXiv. https://doi.org/10.48550/arXiv.1812.02230

Higgins, I., Chang, L., Langston, V., Hassabis, D., Summerfield, C., Tsao, D., & Botvinick, M. (2021).

Unsupervised deep learning identifies semantic disentanglement in single inferotemporal face patch neurons. *Nature Communications*, *12*(1), Article 1. https://doi.org/10.1038/s41467-021-26751-5

Higgins, I., Racanière, S., & Rezende, D. (2022). Symmetry-Based Representations for Artificial and Biological General Intelligence. *Frontiers in Computational Neuroscience*, *16*. https://www.frontiersin.org/articles/10.3389/fncom.2022.836498

Hindy, N. C., Ng, F. Y., & Turk-Browne, N. B. (2016). Linking pattern completion in the hippocampus to predictive coding in visual cortex. *Nature Neuroscience*, *19*(5), Article 5. https://doi.org/10.1038/nn.4284

Hinton, G. (1979). Some demonstrations of the effects of structural descriptions in mental imagery. *Cognitive Science*, *3*(3), 231–250.

Hinton, G. (2021). How to represent part-whole hierarchies in a neural network. *ArXiv Preprint ArXiv:2102.12627*.

Hinton, G. E. (1990). Mapping part-whole hierarchies into connectionist networks. *Artificial Intelligence*, *46*(1–2), 47–75.

Hinton, G. E., Krizhevsky, A., & Wang, S. D. (2011). Transforming Auto-Encoders. In T. Honkela, W. Duch, M. Girolami, & S. Kaski (Eds.), *Artificial Neural Networks and Machine Learning – ICANN 2011* (pp. 44–51). Springer. https://doi.org/10.1007/978-3-642-21735-7_6

Hinton, G. E., & Parsons, L. M. (1988). Scene-based and viewer-centered representations for comparing shapes. *Cognition*, *30*(1), 1–35.

Hinton, G. E., Sabour, S., & Frosst, N. (2018). Matrix capsules with EM routing. *International Conference on Learning Representations*.

Hinton, G. F. (1981). A parallel computation that assigns canonical object-based frames of reference. *Proceedings of the 7th International Joint Conference on Artificial Intelligence-Volume 2*, 683–685.

Hinton, G., Krizhevsky, A., Jaitly, N., Tieleman, T., & Tang, Y. (2012). Does the brain do inverse graphics. *Brain and Cognitive Sciences Fall Colloquium*, *2*.

Hipólito, I. (2022). Cognition Without Neural Representation: Dynamics of a Complex System. *Frontiers in Psychology*, 5472.

Ho, Y.-X., Maloney, L. T., & Landy, M. S. (2007). The effect of viewpoint on perceived visual roughness. *Journal of Vision*, *7*(1), 1. https://doi.org/10.1167/7.1.1

Hochstein, S., & Ahissar, M. (2002). View from the Top: Hierarchies and Reverse Hierarchies in the Visual System. *Neuron*, *36*(5), 791–804. https://doi.org/10.1016/S0896-6273(02)01091-7

Hollingworth, A., & Franconeri, S. L. (2009). Object correspondence across brief occlusion is established on the basis of both spatiotemporal and surface feature cues. *Cognition*, *113*(2), 150–166. https://doi.org/10.1016/j.cognition.2009.08.004

Hong, H., Yamins, D. L., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*(4), 613–622.

Horikawa, T., & Kamitani, Y. (2017). Generic decoding of seen and imagined objects using hierarchical visual features. *Nature Communications*, *8*(1), Article 1. https://doi.org/10.1038/ncomms15037

Hubbard, T. L. (2005). Representational momentum and related displacements in spatial memory: A review of the findings. *Psychonomic Bulletin & Review*, *12*(5), 822–851.

Hubbard, T. L., & Bharucha, J. J. (1988). Judged displacement in apparent vertical and horizontal motion. *Perception & Psychophysics*, *44*(3), 211–221. https://doi.org/10.3758/BF03206290

Hulme, O. J., & Zeki, S. (2007). The sightless view: Neural correlates of occluded objects. *Cerebral Cortex*, *17*(5), 1197–1205.

Hummel, J. E. (2000). Where view-based theories break down: The role of structure in shape perception and object recognition. *Cognitive Dynamics: Conceptual Change in Humans and Machines*, 157–185.

Hummel, J. E., & Stankiewicz, B. J. (1998). Two Roles for Attention in Shape Perception: A Structural Description Model of Visual Scrutiny. *Visual Cognition*, *5*(1–2), 49–79. https://doi.org/10.1080/713756775

Humphrey, G. K., & Jolicoeur, P. (1993). An examination of the effects of axis foreshortening, monocular depth cues, and visual field on object identification. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *46A*, 137–159. https://doi.org/10.1080/14640749308401070

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(03), 90–95.

Iamshchinina, P., Kaiser, D., Yakupov, R., Haenelt, D., Sciarra, A., Mattern, H., Luesebrink, F., Duezel, E., Speck, O., Weiskopf, N., & Cichy, R. M. (2021). Perceived and mentally rotated contents are differentially represented in cortical depth of V1. *Communications Biology*, *4*(1), Article 1. https://doi.org/10.1038/s42003-021-02582-4

Intraub, H., Gottesman, C. V., Willey, E. V., & Zuk, I. J. (1996). Boundary extension for briefly glimpsed photographs: Do common perceptual processes result in unexpected memory distortions? *Journal of Memory and Language*, *35*(2), 118–134.

Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(2), 179.

Ito, M., Tamura, H., Fujita, I., & Tanaka, K. (1995). Size and position invariance of neuronal responses in monkey inferotemporal cortex. *Journal of Neurophysiology*, *73*(1), 218–226. https://doi.org/10.1152/jn.1995.73.1.218

John, Y. (2021, August 16). 'Representing' means exactly what you think it means. *Yohan J. John, PhD*. https://yohanjohn.com/neurologism/representing-means-exactly-what-you-think-it-means/

Johnson, J., Gupta, A., & Fei-Fei, L. (2018). Image generation from scene graphs. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 1219–1228.

Johnson, J., Krishna, R., Stark, M., Li, L.-J., Shamma, D., Bernstein, M., & Fei-Fei, L. (2015). *Image Retrieval Using Scene Graphs*. 3668–3678. https://openaccess.thecvf.com/content_cvpr_2015/html/Johnson_Image_Retrieval_Using_

2015_CVPR_paper.html

Johnson, M. R., & Johnson, M. K. (2014). Decoding individual natural scene representations during perception and imagery. *Frontiers in Human Neuroscience*, *8*. https://doi.org/10.3389/fnhum.2014.00059

Jolicoeur, P. (1990). Identification of Disoriented Objects: A Dual-systems Theory. *Mind & Language*, *5*(4), 387–410. https://doi.org/10.1111/j.1468-0017.1990.tb00170.x

Jolicoeur, P., & Cavanagh, P. (1992). Mental rotation, physical rotation, and surface media. *Journal of Experimental Psychology: Human Perception and Performance*, *18*, 371–384. https://doi.org/10.1037/0096-1523.18.2.371

Jolicoeur, P., Corballis, M. C., & Lawson, R. (1998). The influence of perceived rotary motion on the recognition of rotated objects. *Psychonomic Bulletin & Review*, *5*(1), 140–146. https://doi.org/10.3758/BF03209470

Julian, J. B., Fedorenko, E., Webster, J., & Kanwisher, N. (2012). An algorithmic method for functionally defining regions of interest in the ventral visual pathway. *Neuroimage*, *60*(4), 2357–2364.

Julian, J. B., Keinath, A. T., Marchette, S. A., & Epstein, R. A. (2018). The neurocognitive basis of spatial reorientation. *Current Biology*, *28*(17), R1059–R1073.

Julian, J. B., Ryan, J., Hamilton, R. H., & Epstein, R. A. (2016). The Occipital Place Area Is Causally Involved in Representing Environmental Boundaries during Navigation. *Current Biology*, *26*(8), 1104–1109. https://doi.org/10.1016/j.cub.2016.02.066

Just, M. A., & Carpenter, P. A. (1976). Eye fixations and cognitive processes. *Cognitive Psychology*, *8*(4), 441–480. https://doi.org/10.1016/0010-0285(76)90015-3

Kadir, T., Zisserman, A., & Brady, M. (2004). An affine invariant salient region detector. *European Conference on Computer Vision*, 228–241.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The reviewing of object files: Object-specific integration of information. *Cognitive Psychology*, *24*(2), 175–219.

Kaiser, D., Häberle, G., & Cichy, R. M. (2021). Coherent natural scene structure facilitates the extraction of task-relevant object information in visual cortex. *NeuroImage*, *240*, 118365. https://doi.org/10.1016/j.neuroimage.2021.118365

Kaiser, D., Stein, T., & Peelen, M. V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences*, *111*(30), 11217–11222. https://doi.org/10.1073/pnas.1400559111

Kastner, S., De Weerd, P., Desimone, R., & Ungerleider, L. G. (1998). Mechanisms of directed attention in the human extrastriate cortex as revealed by functional MRI. *Science*, *282*(5386), 108–111.

Kasturirangan, R. (2004). *Mapping spatial relations* [PhD Thesis]. Massachusetts Institute of Technology.

Kayaert, G., Biederman, I., Op de Beeck, H. P., & Vogels, R. (2005). Tuning for shape dimensions in macaque inferior temporal cortex. *The European Journal of Neuroscience*, *22*(1), 212–224.

https://doi.org/10.1111/j.1460-9568.2005.04202.x

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

Kersten, D., Mamassian, P., & Knill, D. C. (1997). Moving Cast Shadows Induce Apparent Motion in Depth. *Perception*, *26*(2), 171–192. https://doi.org/10.1068/p260171

Kim, S.-H., Feldman, J., & Singh, M. (2012). Curved apparent motion induced by amodal completion. *Attention, Perception, & Psychophysics*, *74*(2), 350–364.

Kim, S.-H., Feldman, J., & Singh, M. (2013). Perceived Causality Can Alter the Perceived Trajectory of Apparent Motion. *Psychological Science*, *24*(4), 575–582. https://doi.org/10.1177/0956797612458529

Klindt, D., Schott, L., Sharma, Y., Ustyuzhaninov, I., Brendel, W., Bethge, M., & Paiton, D. (2021). *Towards Nonlinear Disentanglement in Natural Data with Temporal Sparse Coding* (arXiv:2007.10930). arXiv. https://doi.org/10.48550/arXiv.2007.10930

Knapen, T. (2021). Topographic connectivity reveals task-dependent retinotopic processing throughout the human brain. *Proceedings of the National Academy of Sciences*, *118*(2), e2017032118. https://doi.org/10.1073/pnas.2017032118

Koenderink, J. J., Van Doorn, A. J., & Kappers, A. M. (1996). Pictorial surface attitude and local depth comparisons. *Perception & Psychophysics*, *58*(2), 163–173.

Kok, P., Failing, M. F., & de Lange, F. P. (2014). Prior expectations evoke stimulus templates in the primary visual cortex. *Journal of Cognitive Neuroscience*, *26*(7), 1546–1554. https://doi.org/10.1162/jocn_a_00562

Kok, P., Jehee, J. F., & De Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, *75*(2), 265–270.

Kok, P., Mostert, P., & de Lange, F. P. (2017). Prior expectations induce prestimulus sensory templates. *Proceedings of the National Academy of Sciences*, *114*(39), 10473–10478. https://doi.org/10.1073/pnas.1705652114

Kok, P., Rahnev, D., Jehee, J. F., Lau, H. C., & De Lange, F. P. (2012). Attention reverses the effect of prediction in silencing sensory signals. *Cerebral Cortex*, *22*(9), 2197–2206.

Kok, P., Rait, L. I., & Turk-Browne, N. B. (2020). Content-based dissociation of hippocampal involvement in prediction. *Journal of Cognitive Neuroscience*, *32*(3), 527–545.

Kok, P., & Turk-Browne, N. B. (2018). Associative Prediction of Visual Shape in the Hippocampus. *Journal of Neuroscience*, *38*(31), 6888–6899. https://doi.org/10.1523/JNEUROSCI.0163-18.2018

Kolers, P. A., & Pomerantz, J. R. (1971). Figural change in apparent motion. *Journal of Experimental Psychology*, *87*, 99–108. https://doi.org/10.1037/h0030156

Koriat, A., & Norman, J. (1984). What is rotated in mental rotation? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*, 421–434. https://doi.org/10.1037/0278-7393.10.3.421

Koriat, A., & Norman, J. (1988). Frames and images: Sequential effects in mental rotation. *Journal of*

Experimental Psychology: Learning, Memory, and Cognition, *14*, 93–111. https://doi.org/10.1037/0278-7393.14.1.93

Kosiorek, A., Sabour, S., Teh, Y. W., & Hinton, G. E. (2019). Stacked capsule autoencoders. *Advances in Neural Information Processing Systems*, *32*.

Kourtzi, Z., & Shiffrar, M. (2001). Visual representation of malleable and rigid objects that deform as they rotate. *Journal of Experimental Psychology: Human Perception and Performance*, *27*, 335–355. https://doi.org/10.1037/0096-1523.27.2.335

Kriegeskorte, N., & Diedrichsen, J. (2019). Peeling the Onion of Brain Representations. *Annual Review of Neuroscience*, *42*(1), 407–432. https://doi.org/10.1146/annurev-neuro-080317-061906

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. https://www.frontiersin.org/articles/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in Neural Information Processing Systems*, *25*, 1106–1114.

Kulvicki, J. (2004). Isomorphism in information carrying systems. *Pacific Philosophical Quarterly*, *85*(4), 380–395.

Kulvicki, J. (2015). Analog representation and the parts principle. *Review of Philosophy and Psychology*, *6*(1), 165–180.

Kuroki, D. (2021). A new jsPsych plugin for psychophysics, providing accurate display duration and stimulus onset asynchrony. *Behavior Research Methods*, *53*(1), 301–310.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2017). Building machines that learn and think like people. *Behavioral and Brain Sciences*, *40*, e253. https://doi.org/10.1017/S0140525X16001837

Lande, K. J. (n.d.). Compositionality and Context in Perception (Draft).

Lande, K. J. (2021). Seeing and visual reference. *Philosophy and Phenomenological Research*.

Landy, M. S., Maloney, L. T., Johnston, E. B., & Young, M. (1995). Measurement and modeling of depth cue combination: In defense of weak fusion. *Vision Research*, *35*(3), 389–412.

Larsen, A. (2014). Deconstructing mental rotation. *Journal of Experimental Psychology: Human Perception and Performance*, *40*, 1072–1091. https://doi.org/10.1037/a0035648

Larsen, A., & Bundesen, C. (1978). Size scaling in visual pattern recognition. *Journal of Experimental Psychology: Human Perception and Performance*, *4*, 1–20. https://doi.org/10.1037/0096-1523.4.1.1

Larsen, A., & Bundesen, C. (1998). Effects of spatial separation in visual pattern matching: Evidence on the role of mental translation. *Journal of Experimental Psychology: Human Perception and Performance*, *24*, 719–731. https://doi.org/10.1037/0096-1523.24.3.719

Larsen, A., & Bundesen, C. (2009). Common mechanisms in apparent motion perception and visual pattern matching. *Scandinavian Journal of Psychology*, *50*(6), 526–534. https://doi.org/10.1111/j.1467-9450.2009.00782.x

Lawrence, S. J. D., van Mourik, T., Kok, P., Koopmans, P. J., Norris, D. G., & de Lange, F. P. (2018). Laminar Organization of Working Memory Signals in Human Visual Cortex. *Current Biology*, *28*(21), 3435-3440.e4. https://doi.org/10.1016/j.cub.2018.08.043

Lee, S. A. (2017). The boundary-based view of spatial cognition: A synthesis. *Current Opinion in Behavioral Sciences*, *16*, 58–65.

Lee, T. S., & Mumford, D. (2003). Hierarchical Bayesian inference in the visual cortex. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *20*(7), 1434–1448. https://doi.org/10.1364/josaa.20.001434

Leibowitz, H., Brislin, R., Perlmutrer, L., & Hennessy, R. (1969). Ponzo perspective illusion as a manifestation of space perception. *Science*, *166*(3909), 1174–1176.

Lewis, M., Purdy, S., Ahmad, S., & Hawkins, J. (2019). Locations in the neocortex: A theory of sensorimotor object recognition using cortical grid cells. *Frontiers in Neural Circuits*, *13*, 22.

Li, N., & DiCarlo, J. J. (2008). Unsupervised natural experience rapidly alters invariant object representation in visual cortex. *Science*, *321*(5895), 1502–1507.

Li, N., & DiCarlo, J. J. (2010). Unsupervised natural visual experience rapidly reshapes size-invariant object representation in inferior temporal cortex. *Neuron*, *67*(6), 1062–1075.

Li, Y., Wang, Y., Boger, T., Smith, K., Gershman, S. J., & Ullman, T. (2022). *An Approximate Representation of Objects Underlies Physical Reasoning*. PsyArXiv. https://doi.org/10.31234/osf.io/vebu5

Little, P. C., & Firestone, C. (2021). Physically implied surfaces. *Psychological Science*, *32*(5), 799–808.

Lowe, D. G. (1999). Object recognition from local scale-invariant features. *Proceedings of the Seventh IEEE International Conference on Computer Vision*, *2*, 1150–1157 vol.2. https://doi.org/10.1109/ICCV.1999.790410

Lueschow, A., Miller, E. K., & Desimone, R. (1994). Inferior temporal mechanisms for invariant object recognition. *Cerebral Cortex*, *4*(5), 523–531.

Makin, A. D. J., & Bertamini, M. (2014). Do different types of dynamic extrapolation rely on the same mechanism? *Journal of Experimental Psychology: Human Perception and Performance*, *40*(4), 1566.

Makin, A. D. J., & Chauhan, T. (2014). Memory-guided tracking through physical space and feature space. *Journal of Vision*, *14*(13), 10–10.

Maley, C. J. (2011). Analog and Digital, Continuous and Discrete. *Philosophical Studies*, *155*(1), 117–131. https://doi.org/10.1007/s11098-010-9562-8

Mamassian, P., Knill, D. C., & Kersten, D. (1998). The perception of cast shadows. *Trends in Cognitive Sciences*, *2*(8), 288–295. https://doi.org/10.1016/S1364-6613(98)01204-2

Marchette, S. A., & Shelton, A. L. (2010). Object Properties and Frame of Reference in Spatial Memory Representations. *Spatial Cognition & Computation*, *10*(1), 1–27. https://doi.org/10.1080/13875860903509406

Marchette, S. A., Yerramsetti, A., Burns, T. J., & Shelton, A. L. (2011). Spatial memory in the real world: Long-term representations of everyday environments. *Memory & Cognition*, *39*(8), 1401. https://doi.org/10.3758/s13421-011-0108-x

Markman, A. B., & Gentner, D. (1993). Structural alignment during similarity comparisons. *Cognitive Psychology*, *25*(4), 431–467.

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B. Biological Sciences*, *200*(1140), 269–294.

McKinney, W. (2011). pandas: A foundational Python library for data analysis and statistics. *Python for High Performance and Scientific Computing*, *14*(9), 1–9.

Mitroff, S. R., & Alvarez, G. A. (2007). Space and time, not surface features, guide object persistence. *Psychonomic Bulletin & Review*, *14*, 1199–1204. https://doi.org/10.3758/BF03193113

Mocz, V., Vaziri-Pashkam, M., Chun, M. M., & Xu, Y. (2021). Predicting Identity-Preserving Object Transformations across the Human Ventral Visual Stream. *Journal of Neuroscience*, *41*(35), 7403–7419. https://doi.org/10.1523/JNEUROSCI.2137-20.2021

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. https://doi.org/10.1523/JNEUROSCI.0852-19.2019

Mou, W., & McNamara, T. P. (2002). Intrinsic frames of reference in spatial memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(1), 162.

Mullally, S. L., & Maguire, E. A. (2014). Memory, Imagination, and Predicting the Future: A Common Brain Mechanism? *The Neuroscientist*, *20*(3), 220–234. https://doi.org/10.1177/1073858413495091

Munton, J. (2022). How to see invisible objects. *Noûs*, *56*(2), 343–365.

Murray, S. O., Boyaci, H., & Kersten, D. (2006). The representation of perceived angular size in human primary visual cortex. *Nature Neuroscience*, *9*(3), 429–434.

Muryy, A., Siddharth, N., Nardelli, N., Glennerster, A., & Torr, P. H. (2020). Lessons from reinforcement learning for biological representations of space. *Vision Research*, *174*, 79–93.

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527.

Op de Beeck, H., Wagemans, J., & Vogels, R. (2001). Inferotemporal neurons represent low-dimensional configurations of parameterized shapes. *Nature Neuroscience*, *4*(12), 1244–1252. https://doi.org/10.1038/nn767

Ortiz-Tudela, J., Bergmann, J., Bennett, M., Ehrlich, I., Muckli, L., & Shing, Y. L. (2021). Concurrent contextual and time-distant mnemonic information co-exist as feedback in human visual cortex. *BioRxiv*.

Ost, J., Mannan, F., Thuerey, N., Knodt, J., & Heide, F. (2021). Neural scene graphs for dynamic scenes. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2856–2865.

Paccanaro, A., & Hinton, G. E. (2001). Learning Hierarchical Structures with Linear Relational Embedding. *Advances in Neural Information Processing Systems*, *14*. https://proceedings.neurips.cc/paper/2001/hash/814a9c18f5abff398787c9cfcbf3d80c-Abstract.html

Palan, S., & Schitter, C. (2018). Prolific.ac—A subject pool for online experiments. *Journal of Behavioral and Experimental Finance*, *17*, 22–27.

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, *9*(4), 441–474.

Patel, A. B., Nguyen, T., & Baraniuk, R. G. (2015). A probabilistic theory of deep learning. *ArXiv Preprint ArXiv:1504.00641*.

Pearson, J. (2019). The human imagination: The cognitive neuroscience of visual mental imagery. *Nature Reviews Neuroscience*, *20*(10), 624–634.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., & Dubourg, V. (2011). Scikit-learn: Machine learning in Python. *The Journal of Machine Learning Research*, *12*, 2825–2830.

Peelen, M. V., & Kastner, S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex. *Proceedings of the National Academy of Sciences*, *108*(29), 12125–12130. https://doi.org/10.1073/pnas.1101042108

Penny, W. D., Friston, K. J., Ashburner, J. T., Kiebel, S. J., & Nichols, T. E. (2011). *Statistical parametric mapping: The analysis of functional brain images*. Elsevier.

Pessoa, L., Thompson, E., & Noë, A. (1998). Finding out about filling-in: A guide to perceptual completion for visual science and the philosophy of perception. *The Behavioral and Brain Sciences*, *21*(6), 723–748; discussion 748-802. https://doi.org/10.1017/s0140525x98001757

Peters, B., & Kriegeskorte, N. (2021). Capturing the objects of vision with neural networks. *Nature Human Behaviour*, *5*(9), 1127–1144.

Phillips, F., & Todd, J. T. (1996). Perception of local three-dimensional shape. *Journal of Experimental Psychology: Human Perception and Performance*, *22*(4), 930.

Piccinini, G. (2008). Computation without Representation. *Philosophical Studies*, *137*(2), 205–241. https://doi.org/10.1007/s11098-005-5385-4

Poggio, T., & Bizzi, E. (2004). Generalization in vision and motor control. *Nature*, *431*(7010), 768–774.

Poldrack, R. A. (2021). The physics of representation. *Synthese*, *199*(1), 1307–1325. https://doi.org/10.1007/s11229-020-02793-y

Press, C., Kok, P., & Yon, D. (2020). The Perceptual Prediction Paradox. *Trends in Cognitive Sciences*, *24*(1), 13–24. https://doi.org/10.1016/j.tics.2019.11.003

Puneeth, N. C., & Arun, S. P. (2016). A neural substrate for object permanence in monkey inferotemporal cortex. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep30808

Pylyshyn, Z. W. (2000). Situating vision in the world. *Trends in Cognitive Sciences*, *4*(5), 197–207. https://doi.org/10.1016/S1364-6613(00)01477-7

Pylyshyn, Z. W. (2004). Some puzzling findings in multiple object tracking: I. Tracking without keeping track of object identities. *Visual Cognition*, *11*, 801–822. https://doi.org/10.1080/13506280344000518

Pylyshyn, Z. W. (2007). Things and places: How the mind connects with the world. MIT press.

Pylyshyn, Z. W., & Storm, R. W. (1988). Tracking multiple independent targets: Evidence for a parallel tracking mechanism. *Spatial Vision*, *3*(3), 179–197. https://doi.org/10.1163/156856888x00122

Quilty-Dunn, J., & Green, E. J. (2021). Perceptual attribution and perceptual reference. *Philosophy and Phenomenological Research*.

Rajalingham, R., Piccato, A., & Jazayeri, M. (2022). Recurrent neural networks with explicit representation of dynamic latent variables can mimic behavioral patterns in a physical inference task. *Nature Communications*, *13*(1), Article 1. https://doi.org/10.1038/s41467-022-33581-6

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects. *Nature Neuroscience*, *2*(1), 79–87.

Ribeiro, F. D. S., Duarte, K., Everett, M. A., Leontidis, G., & Shah, M. (2022). *Learning with capsule: A survey*. https://doi.org/10.48550/arXiv.2206.02664

Richards, W., Jepson, A., & Feldman, J. (1996). Priors, preferences and categorical percepts. *Perception as Bayesian Inference*, 93–122.

Rieser, J. J. (1989). Access to knowledge of spatial structure at novel points of observation. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*(6), 1157.

Ringach, D. L., & Shapley, R. (1996). Spatial and Temporal Properties of Illusory Contours and Amodal Boundary Completion. *Vision Research*, *36*(19), 3037–3050. https://doi.org/10.1016/0042-6989(96)00062-4

Ritchie, J. B., Kaplan, D. M., & Klein, C. (2019). Decoding the brain: Neural representation and the limits of multivariate pattern analysis in cognitive neuroscience. *The British Journal for the Philosophy of Science*.

Roelfsema, P. R., & de Lange, F. P. (2016). Early visual cortex as a multiscale cognitive blackboard. *Annual Review of Vision Science*, *2*(1), 131–151.

Roelfsema, P. R., Lamme, V. A., Spekreijse, H., & Bosch, H. (2002). Figure—Ground segregation in a recurrent network architecture. *Journal of Cognitive Neuroscience*, *14*(4), 525–537.

Rosenbaum, D., Besse, F., Viola, F., Rezende, D. J., & Eslami, S. (2018). Learning models for visual 3d localization with implicit mapping. *ArXiv Preprint ArXiv:1807.03149*.

Rossel, P., Peyrin, C., Roux-Sibilon, A., & Kauffmann, L. (2022). It makes sense, so I see it better! Contextual information about the visual environment increases its perceived sharpness. *Journal of Experimental Psychology: Human Perception and Performance*.

Saanum, T., & Schulz, E. (2022). *Learning Parsimonious Dynamics for Generalization in Reinforcement Learning* (arXiv:2209.14781). arXiv. https://doi.org/10.48550/arXiv.2209.14781

Sabour, S., Frosst, N., & Hinton, G. E. (2017). Dynamic routing between capsules. *Advances in Neural Information Processing Systems*, *30*.

Schacter, D. L., Addis, D. R., Hassabis, D., Martin, V. C., Spreng, R. N., & Szpunar, K. K. (2012). The Future of Memory: Remembering, Imagining, and the Brain. *Neuron*, *76*(4), 10.1016/j.neuron.2012.11.001. https://doi.org/10.1016/j.neuron.2012.11.001

Schmidhuber, J. (1992). Learning Factorial Codes by Predictability Minimization. *Neural Computation*, *4*(6), 863–879. https://doi.org/10.1162/neco.1992.4.6.863

Schmidt, F., Phillips, F., & Fleming, R. W. (2019). Visual perception of shape-transforming processes:'Shape scission.' *Cognition*, *189*, 167–180.

Schmidt, F., Spröte, P., & Fleming, R. W. (2016). Perception of shape and space across rigid transformations. *Vision Research*, *126*, 318–329. https://doi.org/10.1016/j.visres.2015.04.011

Scholl, B. J., & Flombaum, J. I. (2010). Object persistence. *Encyclopedia of Perception*, *2*, 653–657.

Scholl, B. J., & Pylyshyn, Z. W. (1999). Tracking Multiple Items Through Occlusion: Clues to Visual Objecthood. *Cognitive Psychology*, *38*(2), 259–290. https://doi.org/10.1006/cogp.1998.0698

Sekuler, R., & Nash, D. (1972). Speed of size scaling in human vision. *Psychonomic Science*, *27*, 93–94. https://doi.org/10.3758/BF03328898

Self, M. W., van Kerkoerle, T., Super, H., & Roelfsema, P. R. (2013). Distinct roles of the cortical layers of area V1 in figure-ground segregation. *Current Biology*, *23*(21), 2121–2129.

Senden, M., Emmerling, T. C., van Hoof, R., Frost, M. A., & Goebel, R. (2019). Reconstructing imagined letters from early visual cortex reveals tight topographic correspondence between visual mental imagery and perception. *Brain Structure and Function*, *224*(3), 1167–1183. https://doi.org/10.1007/s00429-019-01828-6

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust Object Recognition with Cortex-Like Mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426. https://doi.org/10.1109/TPAMI.2007.56

Shanahan, M., Crosby, M., Beyret, B., & Cheke, L. (2020). Artificial Intelligence and the Common Sense of Animals. *Trends in Cognitive Sciences*, *24*(11), 862–872. https://doi.org/10.1016/j.tics.2020.09.002

Shen, G., Horikawa, T., Majima, K., & Kamitani, Y. (2019). Deep image reconstruction from human brain activity. *PLOS Computational Biology*, *15*(1), e1006633. https://doi.org/10.1371/journal.pcbi.1006633

Shepard, R. N. (1978). The mental image. *American Psychologist*, *33*, 125–137. https://doi.org/10.1037/0003-066X.33.2.125

Shepard, R. N. (1984). Ecological constraints on internal representation: Resonant kinematics of perceiving, imagining, thinking, and dreaming. *Psychological Review*, *91*(4), 417.

Shepard, R. N. (2001). Perceptual-cognitive universals as reflections of the world. *Behavioral and Brain Sciences*, *24*(4), 581–601.

Shepard, R. N., & Chipman, S. (1970). Second-order isomorphism of internal representations: Shapes of states. *Cognitive Psychology*, *1*(1), 1–17. https://doi.org/10.1016/0010-0285(70)90002-2

Shepard, R. N., & Cooper, L. A. (1982). *Mental images and their transformations* (pp. viii, 364). The MIT Press.

Shepard, R. N., & Judd, S. A. (1976). Perceptual Illusion of Rotation of Three-Dimensional Objects. *Science*, *191*(4230), 952–954. https://doi.org/10.1126/science.1251207

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science*, *171*(3972), 701–703.

Shepard, R. N., & Zare, S. L. (1983). Path-Guided Apparent Motion. *Science*, *220*(4597), 632–634. https://doi.org/10.1126/science.6836307

Shiffrar, M., & Freyd, J. J. (1990). Apparent Motion of the Human Body. *Psychological Science*, *1*(4), 257–264. https://doi.org/10.1111/j.1467-9280.1990.tb00210.x

Smith, F. W., & Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences*, *107*(46), 20099–20103.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage*, *44*(1), 83–98.

Soulos, P., & Isik, L. (2020, July 6). *Disentangled Face Representations in Deep Generative Models and the Human Brain*. NeurIPS 2020 Workshop SVRHM. https://openreview.net/forum?id=ME5Uh_tyld5

Sowizral, H. (2000). Scene graphs in the new millennium. *IEEE Computer Graphics and Applications*, *20*(1), 56–57.

Spelke, E. S. (1990). Principles of object perception. *Cognitive Science*, *14*(1), 29–56. https://doi.org/10.1016/0364-0213(90)90025-R

Spröte, P., & Fleming, R. W. (2016). Bent out of shape: The visual inference of non-rigid shape transformations applied to objects. *Vision Research*, *126*, 330–346.

Spröte, P., Schmidt, F., & Fleming, R. W. (2016). Visual perception of shape altered by inferred causal history. *Scientific Reports*, *6*(1), Article 1. https://doi.org/10.1038/srep36245

Stankiewicz, B. J. (2002). Empirical evidence for independent dimensions in the visual representation of three-dimensional shape. *Journal of Experimental Psychology. Human Perception and Performance*, *28*(4), 913–932.

Stankiewicz, B. J., & Hummel, J. E. (1996). Categorical relations in shape perception. *Spatial Vision*, *10*(3), 201–236.

Stewart, E. E. M., Hartmann, F. T., Morgenstern, Y., Storrs, K. R., Maiello, G., & Fleming, R. W. (2022). Mental object rotation based on two-dimensional visual representations. *Current Biology*, *32*(21), R1224–R1225. https://doi.org/10.1016/j.cub.2022.09.036

Stokes, M., Thompson, R., Nobre, A. C., & Duncan, J. (2009). Shape-specific preparatory activity

mediates attention to targets in human visual cortex. *Proceedings of the National Academy of Sciences*, *106*(46), 19569–19574. https://doi.org/10.1073/pnas.0905306106

Summerfield, C., & De Lange, F. P. (2014). Expectation in perceptual decision making: Neural and computational mechanisms. *Nature Reviews Neuroscience*, *15*(11), 745–756.

Summerfield, C., & Egner, T. (2009). Expectation (and attention) in visual cognition. *Trends in Cognitive Sciences*, *13*(9), 403–409.

Svanera, M., Morgan, A. T., Petro, L. S., & Muckli, L. (2021). A self-supervised deep neural network for image completion resembles early visual cortex fMRI activity patterns for occluded scenes. *Journal of Vision*, *21*(7), 5. https://doi.org/10.1167/jov.21.7.5

Svarverud, E., Gilson, S., & Glennerster, A. (2012). A Demonstration of 'Broken' Visual Space. *PLOS ONE*, *7*(3), e33782. https://doi.org/10.1371/journal.pone.0033782

Tanaka, K. (1996). Inferotemporal Cortex and Object Vision. *Annual Review of Neuroscience*, *19*(1), 109–139. https://doi.org/10.1146/annurev.ne.19.030196.000545

Tarr, M. J., & Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, *21*(2), 233–282.

Teichmann, L., Edwards, G., & Baker, C. I. (2021). Resolving visual motion through perceptual gaps. *Trends in Cognitive Sciences*, *25*(11), 978–991.

Teichmann, L., Moerel, D., Rich, A. N., & Baker, C. I. (2022). The nature of neural object representations during dynamic occlusion. *Cortex*, *153*, 66–86.

Thielen, J., Bosch, S. E., van Leeuwen, T. M., van Gerven, M. A. J., & van Lier, R. (2019). Neuroimaging Findings on Amodal Completion: A Review. *I-Perception*, *10*(2), 2041669519840047. https://doi.org/10.1177/2041669519840047

Thoma, V., Hummel, J. E., & Davidoff, J. (2004). Evidence for holistic representations of ignored images and analytic representations of attended images. *Journal of Experimental Psychology: Human Perception and Performance*, *30*(2), 257.

Thomson, E., & Piccinini, G. (2018). Neural Representations Observed. *Minds and Machines*, *28*(1), 191–235. https://doi.org/10.1007/s11023-018-9459-4

Todd, J. T., & Norman, J. F. (2003). The visual perception of 3-D shape from multiple cues: Are observers capable of perceiving metric structure? *Perception & Psychophysics*, *65*(1), 31–47. https://doi.org/10.3758/BF03194781

Tovee, M. J., Rolls, E. T., & Azzopardi, P. (1994). Translation invariance in the responses to faces of single neurons in the temporal visual cortical areas of the alert macaque. *Journal of Neurophysiology*, *72*(3), 1049–1060. https://doi.org/10.1152/jn.1994.72.3.1049

Ullman, S. (1998). Three-dimensional object recognition based on the combination of views. *Cognition*, *67*(1–2), 21–44.

Ullman, S., Dorfman, N., & Harari, D. (2019). A model for discovering 'containment'relations. *Cognition*, *183*, 67–81.

Ullman, T. D., Spelke, E., Battaglia, P., & Tenenbaum, J. B. (2017). Mind games: Game engines as an architecture for intuitive physics. *Trends in Cognitive Sciences*, *21*(9), 649–665.

Ullman, T. D., Stuhlmüller, A., Goodman, N. D., & Tenenbaum, J. B. (2018). Learning physical parameters from dynamic scenes. *Cognitive Psychology*, *104*, 57–82. https://doi.org/10.1016/j.cogpsych.2017.05.006

Vallat, R. (2018). Pingouin: Statistics in Python. *Journal of Open Source Software*, *3*(31), 1026.

van den Hurk, J., & de Beeck, H. P. O. (2019). Generalization asymmetry in multivariate cross-classification: When representation A generalizes better to representation B than B to A. *BioRxiv*, 592410.

Van Der Velde, F., & De Kamps, M. (2002). Involvement of a visual blackboard architecture in imagery. *Behavioral and Brain Sciences*, *25*(2), 213–214.

Van der Velde, F., & De Kamps, M. (2006). Neural blackboard architectures of combinatorial structures in cognition. *Behavioral and Brain Sciences*, *29*(1), 37–70.

Vaziri, S., & Connor, C. E. (2016). Representation of gravity-aligned scene structure in ventral pathway visual cortex. *Current Biology*, *26*(6), 766–774.

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., & Bright, J. (2020). SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, *17*(3), 261–272.

Võ, 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.

Vuong, J., Fitzgibbon, A. W., & Glennerster, A. (2019). No single, stable 3D representation can explain pointing biases in a spatial updating task. *Scientific Reports*, *9*(1), 1–13.

Walther, A., Nili, H., Ejaz, N., Alink, A., Kriegeskorte, N., & Diedrichsen, J. (2016). Reliability of dissimilarity measures for multi-voxel pattern analysis. *Neuroimage*, *137*, 188–200.

Wandell, B. A., & Winawer, J. (2011). Imaging retinotopic maps in the human brain. *Vision Research*, *51*(7), 718–737. https://doi.org/10.1016/j.visres.2010.08.004

Ward, E. J., Isik, L., & Chun, M. M. (2018). General transformations of object representations in human visual cortex. *Journal of Neuroscience*. https://doi.org/10.1523/JNEUROSCI.2800-17.2018

Warnking, J., Dojat, M., Guérin-Dugué, A., Delon-Martin, C., Olympieff, S., Richard, N., Chéhikian, A., & Segebarth, C. (2002). fMRI Retinotopic Mapping—Step by Step. *NeuroImage*, *17*(4), 1665–1683. https://doi.org/10.1006/nimg.2002.1304

Waskom, M. L. (2021). Seaborn: Statistical data visualization. *Journal of Open Source Software*, *6*(60), 3021.

Wertheimer, M. (1912). Experimentelle studien uber das sehen von bewegung. *Zeitschrift Fur Psychologie*, *61*.

Wetherill, G., & Levitt, H. (1965). Sequential estimation of points on a psychometric function. *British Journal of Mathematical and Statistical Psychology*, *18*(1), 1–10.

Whittington, J. C. R., Dorrell, W., Ganguli, S., & Behrens, T. E. J. (2022). *Disentangling with Biological Constraints: A Theory of Functional Cell Types* (arXiv:2210.01768). arXiv.

https://doi.org/10.48550/arXiv.2210.01768

Whittington, J. C. R., Kabra, R., Matthey, L., Burgess, C. P., & Lerchner, A. (2021). *Constellation: Learning relational abstractions over objects for compositional imagination* (arXiv:2107.11153). arXiv. https://doi.org/10.48550/arXiv.2107.11153

Winawer, J., Huk, A. C., & Boroditsky, L. (2008). A Motion Aftereffect From Still Photographs Depicting Motion. *Psychological Science*, *19*(3), 276–283. https://doi.org/10.1111/j.1467-9280.2008.02080.x

Wohlschläger, A. M., Specht, K., Lie, C., Mohlberg, H., Wohlschläger, A., Bente, K., Pietrzyk, U., Stöcker, T., Zilles, K., & Amunts, K. (2005). Linking retinotopic fMRI mapping and anatomical probability maps of human occipital areas V1 and V2. *Neuroimage*, *26*(1), 73–82.

Wu, C., Clipp, B., Li, X., Frahm, J.-M., & Pollefeys, M. (2008). 3D model matching with viewpoint-invariant patches (VIP). *2008 IEEE Conference on Computer Vision and Pattern Recognition*, 1–8.

Wurm, M. F., & Lingnau, A. (2015). Decoding Actions at Different Levels of Abstraction. *Journal of Neuroscience*, *35*(20), 7727–7735. https://doi.org/10.1523/JNEUROSCI.0188-15.2015

Wyart, V., Nobre, A. C., & Summerfield, C. (2012). Dissociable prior influences of signal probability and relevance on visual contrast sensitivity. *Proceedings of the National Academy of Sciences*, *109*(9), 3593–3598.

Xu, H., Tang, N., Zhou, J., Shen, M., & Gao, T. (2017). Seeing “what” through “why”: Evidence from probing the causal structure of hierarchical motion. *Journal of Experimental Psychology: General*, *146*(6), 896.

Xu, Y., & Franconeri, S. L. (2015). Capacity for Visual Features in Mental Rotation. *Psychological Science*, *26*(8), 1241–1251. https://doi.org/10.1177/0956797615585002

Xue, J., Li, C., Quan, C., Lu, Y., Yue, J., & Zhang, C. (2017). Uncovering the cognitive processes underlying mental rotation: An eye-movement study. *Scientific Reports*, *7*(1), 1–12.

Yildiz, G. Y., Sperandio, I., Kettle, C., & Chouinard, P. A. (2021). A review on various explanations of Ponzo-like illusions. *Psychonomic Bulletin & Review*, 1–28.

Yon, D., Gilbert, S. J., de Lange, F. P., & Press, C. (2018). Action sharpens sensory representations of expected outcomes. *Nature Communications*, *9*(1), 1–8.

Yonas, A., Goldsmith, L. T., & Hallstrom, J. L. (1978). Development of Sensitivity to Information Provided by Cast Shadows in Pictures. *Perception*, *7*(3), 333–341. https://doi.org/10.1068/p070333

Zhou, K., Luo, H., Zhou, T., Zhuo, Y., & Chen, L. (2010). Topological change disturbs object continuity in attentive tracking. *Proceedings of the National Academy of Sciences*, *107*(50), 21920–21924.

Appendices

Appendix A

Nederlandse samenvatting

De wereld om ons heen bestaat niet uit geïsoleerde objecten: dingen om ons heen zijn altijd verweven in een web van relaties. Als ik om me heen kijk, zie ik een tafel met daarop een fles, ik zie dat de wind de bladeren van de boom doet schudden, ik zie mijn eigen handen deze woorden typen op een toetsenbord. Een nuttig model van deze wereld, dat ons in staat stelt te voorspellen hoe dingen zich zullen gedragen en hoe ze eruit zullen zien, moet rekening houden met deze complexe interacties.

We weten, uit eerder onderzoek, dat de manier waarop we objecten zien sterk afhangt van hun context. Bijvoorbeeld, objecten kunnen beter worden herkend als ze op een plek verschijnen waar ze worden verwacht: een auto op een weg en een boot op zee is makkelijker te herkennen dan een auto op zee en een boot op een weg. Menselijk zicht behandelt objecten dus niet als geïsoleerde entiteiten. We weten echter nog niet waar ons brein deze contextuele informatie voor kan gebruiken: helpt deze contextuele informatie alleen om objecten te herkennen, of ook om te voorspellen hoe deze objecten zullen veranderen?

In dit proefschrift heb ik geprobeerd een antwoord te geven op die vraag. Om te bestuderen hoe context voorspellingen beïnvloedt in de echte wereld, heb ik ruimtelijke transformaties bestudeerd binnen 3D omgevingen: zo heb ik gekeken naar de manier waarop het uiterlijk van objecten verandert als we ze vanuit een andere hoek bekijken, en de manier waarop de grootte van objecten toe- of afneemt naarmate we dichterbij of verder weg zijn. Een halve eeuw aan onderzoek heeft aangetoond dat het menselijke brein in staat is om dit soort transformaties (schalen, roteren) te ‘simuleren’: bijvoorbeeld, we kunnen voorspellen hoe objecten eruit zullen zien vanuit een nieuwe hoek, en deze voorspelling neemt meer tijd in beslag naarmate de hoek toeneemt, alsof we het object in onze geest roteren.

Deze mogelijkheid om objecten mentaal te transformeren wordt verondersteld een rol te spelen bij het maken van voorspellingen terwijl we naar de wereld kijken. Maar in de echte wereld, hangt de manier waarop objecten veranderen meestal ook af van hun context. Als ik me bijvoorbeeld door een kamer beweeg, zal ik zien dat het meubilair in de kamer en de muren gezamenlijk (en coherent) bewegen. Een tafel die in de tegenovergestelde richting beweegt van de rest van de kamer is vrij onwaarschijnlijk, dus het zou nuttig zijn om gebruik te maken van deze regelmatigheden bij het maken van voorspellingen. Eerdere experimenten met mentale transformaties, die deelnemers meestal vroegen om geïsoleerde objecten mentaal te transformeren, zijn in deze zin niet representatief voor waarneming in de echte wereld.

Om deze kloof te dichten, hebben we een experimenteel paradigma

ontworpen dat precies dit soort situaties modelleert: je ziet een kamer met daarin een object, die samen bewegen, zoals je zou verwachten wanneer je in die kamer zou rondlopen. Het object wordt dan plotseling afgeschermd, waardoor je alleen de omliggende kamer nog ziet bewegen. In een reeks experimenten hebben we geprobeerd te achterhalen of mensen het verborgen object in hun geest blijven 'zien', en of ze het object dan ook zien meeveranderen met de bewegende kamer. Als de kamer bijvoorbeeld roteert naar een bepaalde draaihoek, zou het object op dezelfde manier moeten roteren als de kamer; en als de kamer dichterbij komt, zou het object -in het geestesoog - groter moeten worden. Om deze voorspellingen te toetsen, hebben we twee technieken gebruikt: gedragsexperimenten en functionele MRI (fMRI), een beeldvormingstechniek waarmee indirect hersenactiviteit kan worden gemeten.

In **Hoofdstuk 2** hebben we getest of deelnemers een object mentaal roteren in samenspraak met de rotatie van de kamer, door hun gedrag te meten in twee situaties: één situatie waarin het afgeschermde object weer verscheen in een oriëntatie die overeenkomt met de rotatie van de kamer (*congruente* conditie), en een andere situatie waarin het object weer verscheen in een oriëntatie die niet overeenkomt met de rotatie van de kamer (*incongruente* conditie). Deelnemers moesten een eenvoudige en *orthogonale* visuele taak op het object uitvoeren: Het object dat weer verscheen werd tweemaal kort achter elkaar getoond, en participanten moesten aangeven of het object in die twee beelden precies hetzelfde georiënteerd was, of een nét verschillende oriëntatie had. Om deze taak te doen, hoefden participanten niet te voorspellen hoe het object eruit zou zien gegeven de nieuwe orientatie van de kamer; de taak was immers puur gebaseerd op de laatste twee beelden van het verschenen object zelf. Als participanten echter een mentale weergave van het object bijhouden, dan zou een object dat overeenkomt met die weergave nauwkeuriger moeten zijn waargenomen. We ontdekten dat dit inderdaad het geval was: deelnemers waren beter in het uitvoeren van de visuele taak wanneer het verschijnende object overeenkwam met de oriëntatie van de geroteerde kamer, dan wanneer het object verscheen in een oriëntatie die niet overeenkwam met de kamer. Een kritiek aspect is dat exact dezelfde beelden werden getoond in de congruente en incongruente condities; het enige verschil tussen deze condities betrof of de uiteindelijke oriëntatie van het object in de kamer overeenkwam met de initiële oriëntatie van het object in de kamer. De enige verklaring van dit resultaat is dan ook dat participanten automatisch de veranderende oriëntatie van de kamer gebruikten om de oriëntatie van het object te voorspellen.

Interessant genoeg kon dit effect worden waargenomen bij een verscheidenheid aan verschillende objecten en oriëntaties, wat suggereert dat we objecten flexibel kunnen roteren en niet alleen specifieke beelden met elkaar associëren. Bovendien was het effect nog steeds aanwezig wanneer contextuele verwachtingen frequent werden geschonden tijdens het experiment (bijvoorbeeld als het object in 50% of zelfs 75% van de gevallen in een incongruente oriëntatie verscheen). Dit suggereert dat de voorspellingen die mensen maken voortkomen

uit kennis van hoe objecten zich gedragen in de echte wereld (objecten bewegen doorgaans coherent met hun omgeving), en deze voorspellingen niet gemakkelijk kunnen worden overschreven. Al met al leveren deze gedragsbevindingen sterk bewijs voor het menselijk vermogen om interne objectrepresentaties automatisch coherent met de context te transformeren.

In **Hoofdstuk 3** gebruikten we fMRI om de hersenactiviteit van deelnemers te meten terwijl ze een taak uitvoerden die vergelijkbaar was met die van het vorige hoofdstuk. We maakten gebruik van zogenaamde multivariate patroonanalyse (MVPA), een techniek die het mogelijk maakt om de mentale representaties in een bepaald hersengebied te achterhalen met behulp van een classificatie-algoritme. De classifier wordt getraind op een serie voorbeelden van hersenactiviteit die is opgewekt door het tonen van verschillende stimuli, en leert zo onderscheid te maken tussen situaties waarin de participant de ene of de andere stimuluscategorie heeft gezien, puur op basis van de hersenactiviteit. In ons geval trainden we de classifier om onderscheid te maken tussen objecten (bankstel of bed) die vanuit een bepaald standpunt *smal* of *breed* leken (afhankelijk van hoe ze geroteerd waren). De classifier werd vervolgens getest op objecten die congruent of incongruent waren met de rotatie van de kamer, zoals omschreven in **Hoofdstuk 2**. Als de representatie van congruente objecten in het visuele systeem inderdaad nauwkeuriger is, zoals de gedragsresultaten lijken te suggereren, zou de classifier de oriëntatie (smal of breed) beter kunnen achterhalen door hersenactiviteit in de congruente conditie (waar de interne representatie van het object dezelfde oriëntatie heeft als het object op het scherm) dan in de incongruente conditie (waar de oriëntaties van de interne representaties en het waargenomen object niet overeenkomen).

Dit bleek inderdaad het geval. De classifier kon de oriëntatie van congruente objecten beter classificeren dan van incongruente objecten. Verrassend genoeg werd dit verschil in classificatie alleen gevonden in de vroegste niveaus van het visuele systeem, de gebieden V1 en V2 (de primaire en secundaire visuele hersenschors), maar bijvoorbeeld niet in LOC, een gebied in de laterale occipitale hersenschors dat gespecialiseerd is in het verwerken van objecten. Dit zou verband kunnen houden met de aard van de voorspellingen die werden opgeroepen door de context: omdat deelnemers een heel specifieke voorspelling konden maken van hoe het object er na de rotatie uit kwam te zien (en er geen abstracte representatie van hoefden te maken), waren vroege visuele gebieden wellicht beter geschikt om een mentale representatie van het object te maken. Omdat onze analysemethode echter specifiek was afgestemd op het achterhalen van dit soort concrete verschillen tussen representaties (smal versus breed), kunnen we niet uitsluiten dat er andere informatie over de voorspelde objecten elders in de hersenen werd gerepresenteerd.

De congruente en incongruente condities verschilden niet alleen in hoe goed we de oriëntatie van het object konden classificeren op basis van de hersenactiviteit, ze verschilden ook in de algehele hoeveelheid veroorzaakte hersenactiviteit. We vonden dat over het algemeen de hersenen sterker reageren

op incongruente objecten, met name ook in gebieden die betrokken zijn bij het verwerken van verrassende of onverwachte stimuli. Dit resultaat suggereert dat congruente objectoriëntaties werden waargenomen als 'standaard', terwijl incongruente oriëntaties verrassend waren. Dit is een verdere indicatie van de automatisering van het congruentie-effect. Bovendien is het consistent met verschillende eerdere bevindingen over de invloed van verwachtingen op waarneming. Eerdere studies hebben namelijk aangetoond dat verwachte stimuli met hogere precisie worden gerepresenteerd in het visuele systeem, terwijl ze lagere algehele niveaus van hersenactiviteit teweegbrengen; dit duidt op een *verscherpte*, efficiëntere representatie van de visuele informatie. Eerder bewijs voor deze effecten van verwachtingen op waarneming werden echter voornamelijk gevonden met sterk vereenvoudigde stimuli (bijvoorbeeld lijnpatronen): onze resultaten suggereren dat vergelijkbare mechanismen betrokken kunnen zijn bij complexe natuurgetrouwe afbeeldingen van objecten en scènes.

In de vorige hoofdstukken hebben we de context-gedreven voorspellingen van objecten telkens indirect gemeten, door de respons te meten op daadwerkelijk getoonde objecten die dan wel congruent dan wel incongruent waren met de context-gedreven voorspellingen. In **Hoofdstuk 4**, probeerden we deze verwachtingen in het visuele systeem direct te observeren. We gebruikten opnieuw fMRI en MVPA, zoals in **Hoofdstuk 3**, maar in dit geval probeerden we de oriëntatie (smal versus breed) van het voorspelde object te classificeren vóórdat het op het scherm verscheen. Deelnemers zagen opnieuw een roterende kamer waarin het centrale object was afgeschermd, maar in de meeste gevallen bleef het object tot het eind afgeschermd en verscheen het niet meer. Participanten deden hier een veel simpelere taak; ze hoefden alleen maar te tellen hoe vaak het object wel weer tevoorschijn kwam, en dit aan het einde van elk blok te melden. Wij vroegen ons af of we de voorspelde oriëntatie van het object (smal versus breed) konden reconstrueren, alleen op basis van de hersenactiviteit die werd veroorzaakt door de geroteerde kamer met het afgeschermde object. Als dit zou lukken, zou dit aantonen dat participanten een mentale representatie van het object creëren, die correspondeert met de rotatie van de kamer. Net als in het voorgaande hoofdstuk, werd de classifier getraind om onderscheid te maken tussen hersenactiviteit die werd opgeroepen door daadwerkelijk waargenomen (smalle versus brede) objecten, in een los onderdeel van het experiment.

We vonden dat de oriëntatie van het object inderdaad kon worden achterhaald aan de hand van activiteit in het visuele systeem, ondanks dat het object niet op het scherm was getoond. De context-gedreven object-representaties, die we ook al waarnamen in het vorige hoofdstuk,, komen dus niet alleen tot uiting in de waarneming van zichtbare objecten, maar bestaan ook op zichzelf, wanneer er (nog) geen object zichtbaar is. In tegenstelling tot het vorige hoofdstuk, waar de context-gedreven objectrepresentaties voornamelijk in de vroege verwerkingsgebieden van het visuele systeem werden gevonden, vonden we deze in het huidige experiment in hoger gelegen visuele gebieden; gebieden die mogelijkerwijs objecten representeren op een meer abstract niveau. Dit kan te

maken hebben met het feit dat de verwachtingen worden gerepresenteerd op een meer abstract niveau, passend bij de grotere receptieve velden die in deze gebieden worden gevonden. Deze gebieden omvatten echter ook delen van de dorsale visuele stroom, gespecialiseerd in het verwerken van beweging. Een andere mogelijkheid is daarom dat de verwachting van het objectbeeld voornamelijk wordt aangedreven door de bewegingspatronen van de scèneachtergrond. Een recente studie heeft gevonden dat visuele bewegingspatronen het gedrag van deelnemers in een mentale rotatietaak konden voorspellen, wat suggereert dat visuele gebieden die beweging verwerken een rol spelen bij verschillende ruimtelijke transformaties van objectrepresentaties. Het verduidelijken van de verbanden tussen deze verschillende processen zal een belangrijk vraag zijn voor toekomstig onderzoek.

Samengevat hebben we in dit hoofdstuk gevonden dat context-gedreven verwachtingen van objectoriëntatie zowel direct als indirect kunnen worden gedecodeerd in het menselijk visuele systeem. Bovendien lijken de verwachtingen zelf in een later stadium van het visuele systeem gepresenteerd te worden dan hun modulerend effect, wat suggereert dat het effect dat in **Hoofdstuk 3** werd gevonden het resultaat is van feedback van deze latere visuele gebieden.

In het laatste experimentele hoofdstuk, **Hoofdstuk 5**, onderzochten we of automatische voorspellingen van nieuwe objectstandpunten uit contextuele informatie gevonden kunnen worden voor andere transformaties dan rotatie. In de vorige hoofdstukken gebruikten we rotatie vanwege de alomtegenwoordigheid ervan in het dagelijks leven, de complexe manieren waarop het het uiterlijk van objecten transformeert, en de vele eerdere onderzoeken naar mentale rotatie. De mentale transformatievaardigheden van mensen zijn echter niet beperkt tot rotatie: eerder onderzoek heeft bewijs gevonden voor verschillende andere transformaties, waaronder mentale translatie en schaling. In **Hoofdstuk 5** onderzochten we daarom of we een vergelijkbaar verwachtingseffect konden vinden voor scènes en objecten die in de diepte bewogen (dichter bij de waarnemer kwamen) in plaats van roteerden.

De resultaten toonden aan dat translatie een soortgelijk verwachtingseffect oproept als rotatie, waarbij de deelnemers verschillend presteerden op objecten die congruent of incongruent waren met de omliggende scène. Dit verschil werd echter alleen weerspiegeld in visuele gevoeligheid en nauwkeurigheid wanneer het object meestal (75% van de gevallen) in de congruente positie werd getoond. Wanneer de congruente objectpositie slechts op 50% of 25% van de gevallen werd getoond, had congruentie alleen een algemeen effect op hoe deelnemers beslissingen namen. Denk terug aan de orthogonale taak die al in **Hoofdstuk 2** werd gebruikt. Na het verschijnen van het object in een congruente of incongruente grootte, moesten de deelnemers aangeven of het object veranderde. We vonden dat deelnemers eerder aangaven dat het object veranderde wanneer het object incongruent was, maar zonder dat ze nauwkeuriger werden in de taak. Wellicht dat deelnemers de neiging hadden om 'verschillend' te antwoorden wanneer hun verwachting werd geschonden in

plaats van wanneer de teststimuli daadwerkelijk verschillend waren. Een interessante vraag is hoe dit interferentie-effect gerelateerd is aan de effecten op nauwkeurigheid die we waarnamen in **Hoofdstuk 2**, en of dit verschil te wijten is aan de verschillende onderzochte transformaties (rotatie vs. translatie).

Over het algemeen tonen deze resultaten echter aan dat deelnemers nog steeds automatisch objectrepresentaties transformeren samen met de context, ook in het geval van translatie. Dit suggereert dat het cognitieve proces dat in de vorige hoofdstukken werd onderzocht, generaliseert naar verschillende objecttransformaties.

We zijn nu aan het einde gekomen van het experimentele deel van dit proefschrift. Ik zou nu een stap terug willen doen en vragen: wat betekent dit allemaal? Zijn de effecten die we hebben onderzocht beperkt tot zeer specifieke laboratoriumsituaties, of kunnen ze ons iets vertellen over perceptie in de echte wereld? In **Hoofdstukken 1 & 6** reflecteer ik op de theoretische betekenis van deze empirische bevindingen. Het werk in dit proefschrift is een eerste stap naar het karakteriseren van hoe ons vermogen om mentaal objecten te transformeren nuttig kan zijn tijdens het navigeren in de wereld. Door interactie met contextuele informatie 'daarbuiten' in de wereld, kunnen onze representaties 'hierbinnen' in ons hoofd in real-time worden bijgewerkt en direct worden vergeleken met wat we zien. De experimenten die we hebben uitgevoerd zijn een voorbeeld van dit proces: de zichtbare informatie van de achtergrondscène stuurt de mentale rotatie of translatie van het object. Deze interactie tussen externe stimuli en interne representaties volgt dezelfde regels als die de interactie tussen dingen in de externe wereld bepalen, bijvoorbeeld dat objecten over het algemeen coherent samen bewegen.

In het algemeen denk ik dat ons vermogen om de buitenwereld in onze gedachten te simuleren het nuttigst is bij het dynamisch volgen van objecten, omdat we hierdoor kunnen interacteren in een steeds veranderende omgeving. Hoe dit uitpakt in meer realistische situaties zal een vraag zijn voor toekomstig onderzoek - hier hebben we slechts één kleine wereld bestudeerd.

(Thanks to ChatGPT, Surya Gayet and Marius Peelen for the translation)

Appendix B

Research data management

This research followed the applicable laws and ethical guidelines. Research Data Management was conducted according to the FAIR principles. The paragraphs below specify in detail how this was achieved.

Ethics

This thesis is based on the results of human studies, which were conducted in accordance with the principles of the Declaration of Helsinki. The Ethical Committee of the faculty of Social Sciences (ECSS) has given a positive advice to conduct these studies to the Dean of the Faculty, who formally approved the conduct of these studies (ECSW2017-2306-517). This research was funded by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement no. 725970).

Findable, Accessible

The table below details where the data and research documentation for each chapter can be found on the Donders Repository (DR), the Open Science Framework (OSF) and Github. All data archived as a Data Sharing Collection remain available for at least 10 years after termination of the studies.

Chapter	DAC	RDC	OSF	Github
2	2020.00041_240	2020.00041_697	WNEFH	GAldegheri/scenecontext-transforms
3	2019.00114_057	2019.00114_952	-	-
4	2018.00091_503	2018.00091_612	-	-
5	2020.00041_240	2020.00041_697	-	-

DAC = Data Acquisition Collection, RDC = Research Documentation Collection, OSF = Open Science Framework

For Chapters 3 and 4, research data have also been stored on the Donders project drive (respectively, project 3018040.05 and 3018040.07). These data were accessible to all members involved in the project. The publication resulting from these chapters is still in preparation. Upon publication, the respective data will be made publicly available under the RU-DI-HD-1.0 license as it contains potentially identifiable data, and will be removed from the project drives.

Informed consent for Chapters 3 and 4 was obtained on paper, and for Chapters 2 and 5 digitally, following the Centre procedure. The forms are archived in the central archive of the Centre for 10 years after termination of the studies.

Interoperable, Reusable

The raw data are stored in the DAC in their original form. For RDC and DSC long-lived file formats (e.g. .csv, .nii, .mat) have been used ensuring that data remains usable in the future. The data of the RDC and DSC are organized according to the BIDS standards, with concomitant README files. Results are reproducible by providing a description of the experimental setup, raw data (DAC and OSF), and code for running the experiment and analyzing the data (RDC and Github). Also, the used software including version numbers is specified.

Privacy

The privacy of the participants in this thesis has been warranted using random individual subject codes. A pseudonymization key linked this random code with the personal data. This pseudonymization key was stored on a network drive that was only accessible to members of the project who needed access to it because of their role within the project. The pseudonymization key was stored separately from the research data. The pseudonymization key of Chapter 2 was destroyed within one month after finalization of the project. The keys of Chapters 3, 4 and 5 are still stored on a dedicated restricted network drive and will be destroyed within one month after finalization. Data in Chapters 2 and 5 are not identifiable and shared without restrictions. MRI data of Chapters 3 and 4, upon publication, will be defaced and shared under the restricted license RU-DI-HD-1.0, which provides extra statements for the protection of the identity of the participants.

Appendix C

Acknowledgements

I kept this part last, and procrastinated writing it for a long time. It's always so hard to figure out what to write! Anyway, I really have to do it now.

I think the word "rollercoaster" is an apt description for this PhD. It was a time of unprecedented learning for me, often in the form of clashing with a reality that didn't quite match my expectations. I think this is how it's supposed to go. There's no point in pretending it was easy or always fun and pleasant. But it taught me so many things that I would never have learned otherwise. And I wouldn't have learned so much, had as much fun, nor been able to cope with the down moments without the incredible people that accompanied me through this time, some that came from the "before times" and others that I met along the way.

First of all, I want to thank my supervisor, Marius Peelen. Marius, you have taught me so much about research: you are a terrific example of what being a researcher means, in terms of clarity of thought, scientific integrity, scope of knowledge and so much more. You have allowed me to define myself both by agreeing and disagreeing with you. Most of all, you have taught me how to express my thought in writing in the clearest way possible, and through many iterations, make that thought clearer, first and foremost to myself. Your kindness and tolerance as a supervisor were also indispensable in getting me to the end of this journey.

To my co-supervisor, Surya Gayet: you have been a fundamental figure during this time, you were always there whenever I needed you, for advice, for reassurance, for technical assistance, or simply to talk. You have also been an incredible role model for your open-mindedness, curiosity and the speed with which you can pick up new ideas, as well as a wonderful example that it's possible to be a successful researcher and an incredibly fun person at the same time. Who would have thought!

Right after my supervisors, another fundamental figure that accompanied me during the PhD was Sushrut Thorat. I would say the amount of stuff I learned from you was comparable to a supervisor. Thanks to you, I ventured into areas of research that I would have probably thought were too difficult or distant otherwise, areas that I now want to keep exploring further and further. You have taught me to set a very high standard on my arguments, both when I agreed and when I disagreed with you. Most of all, you have been and are an incredible friend, and I'm really happy to have met you.

Charlotte, since you joined the lab a year after Sush and I, you have been another fundamental companion in this PhD journey, as well as in other journeys around the world... I'm always happy to be a spectator or a co-conspirator of your ranting sessions!

Maëlle, during the last year of my PhD you have been my main source of encouragement and emotional support. If you are available as an office buddy on demand, I would be interested in hiring you! And I hope I can be as helpful to you some day.

I also want to mention many other fantastic people that have gone through the lab and the Donders during these years (in no particular order): Genevieve, Yuanfang, Chuanji, Micha, Johannes, Miles, Linlin, Elena, Nestor, Paula, Lu-Chun, Marco, Elie, Simen, Eelke, Natalia, Alexandra, Jorie, Myrthel, Nicolò. Thanks to all of you for contributing to make the lab and the Donders feel like a family.

A special mention goes to Simen, to whom I owe the title of this thesis: you probably don't remember, but after being in the scanner for my experiment, you said that after seeing these rooms for 1.5 hours, they become "your little world". I thought that sounded very cool.

Of course, one name is missing from this list. Qiu, this is hardly the place to express what I feel for you. Thanks to you, I consider writing this thesis my second-best accomplishment of 2022. I love you so much, and I'm looking forward to many more adventures with you.

Beyond the lab and institute, my life in Nijmegen has been, in large part, filled with music, again thanks to many wonderful people. Anja, what I learned from you during these years is comparable to a PhD. You are an incredible teacher and you have made me several times the singer I was before meeting you. Popkoor Switch, for close to four years, was a great outlet to let out, each Wednesday, the music I had accumulated in me during the week. And more recently, Lucas and "Bandname TBD" (Valentin and Thomas, and Henri before) have given me the fun of playing in a band again.

Speaking about life in Nijmegen, I spent the large part of it in Burghardt van den Berghstraat. There, I was lucky to have many great neighbors coming and going, and in particular my quasi-roommate Anna. Anna, you deserve a mention here and a big thank you for tolerating me during this time, even when I was singing loudly in the middle of the day during the pandemic, and being really helpful whenever I needed it, even removing dead pigeons from the roof for me.

While I met many amazing people during these years, the wonderful friends I had before were always with me. Among many bizarre situations brought about by the pandemic, I have found myself sharing an Office™ with some of those friends for a few months in Venice. Most of all the host, Gabriele (& Barbara). Thanks to you guys, I have been able to continue my work in a completely different 'scene context'. I like to think that this has allowed me to see things in a different light, and fostered new ideas.

Two central people in my life have also continued to play a major role during this time, and particularly during the tougher parts. Anna, you have been an indispensable support for me during a tough time, and continue to be my rock whenever I need someone to talk to. I hope I can do a little bit of the same for you. Andrea, since high school you have been an incredible friend, and I consider you and your parents as a second family. I'm incredibly lucky to have your friendship

and support.

Another person that was fundamental for me during these years was my therapist, Roberta. I'm really not sure if I could have gotten to the end of this PhD without a major process of self-evaluation and discovery, and you have played a central role in that process, helping me to deal with a crisis first, and then to successfully navigate the small and large difficulties my mind puts me through everyday.

Last, but absolutely not least, my family. Mom, dad and my brother Giovanni, being able to always count on your love and support is an incredible privilege that I do not take for granted. Growing up in a highly stimulating and supportive environment, thanks to you, was fundamental in making me the curious human being I am today. I realize this again every time we all gather around the dinner table. The work in this thesis is one of many things that would have never been possible without you.

Appendix D

About the author

Giacomo Aldegheri was born on March 21st, 1992 in Venice, Italy. He graduated from Liceo Classico Marco Foscarini in Venice in 2011. Towards the end of high school, he grew an interest in studying the mind, realizing that many phenomena that are commonly considered the exclusive domain of introspection, poetry, art or at best philosophy are in fact amenable to scientific investigation. Oliver Sacks' book, *The Man Who Mistook His Wife for a Hat*, was one of the sources of this realization.

Driven by this interest, in 2011, Giacomo started studying psychology and cognitive science at the University of Trento, in Rovereto. During his Bachelor's, he progressively got exposed to the various subfields of psychology, each replacing the previous one as his favorite: first thought and decision making, then memory, then language. But ultimately, it was a course in visual perception that truly sparked his flame, as a primarily visual animal with a lifelong passion for art. He got his first research experience assisting Elisa Infanti, then a PhD student in Massimo Turatto's lab, with a psychophysical study on perceptual learning, using simple artificial stimuli (Gabor patches). Wanting to explore more naturalistic stimuli, he did a second internship with Jorien van Paaschen in David Melcher's lab studying the effect of semantics on false memories in real-world scenes.

After graduating in 2014, he faced the choice of doing his Master's in London, Amsterdam or Rovereto. He chose to stay in Rovereto, primarily because of the vibrant research environment there. During his Master's, he again worked with David Melcher on several projects, one of which became his thesis on how different visual features influence recognition of 2D and 3D natural scenes. He also spent time in Pascal Mamassian's lab at the École Normale Supérieure, in Paris, working on the psychophysics of perceptual organization in stereo vision (Gabor patches again), and in Nicu Sebe's lab at the Department of Computer Science in Trento, where he experimented with deep neural networks for the first time. Advised by his supervisor David Melcher, he contacted Marius Peelen, who had just received an ERC grant, for a PhD position. Marius was then still working in Rovereto, but he would soon move to Nijmegen, in the Netherlands.

So after he finished his Master's in 2017, Giacomo started as a PhD in Marius Peelen's lab at the Donders Center for Cognition. After a long time spent looking for a research question, he finally settled on the topic of this thesis, the interaction between scene context and mental object transformations. He investigated this topic using functional magnetic resonance imaging (fMRI) and behavioral experiments, both in the laboratory and online. During his PhD, he also explored other interests, such as computational modeling and neural networks, which led to two publications.

He is now working as a postdoctoral researcher at the University of Amsterdam with Steven Scholte and Iris Groen, finally combining his interests in high-level perception of relations in natural scenes and in neural networks as a model of cognition.

List of Publications

* indicates equal contribution

Aldegheri, G., Gayet, S., & Peelen, M. V. (2023). Scene context automatically drives predictions of object transformations. *Cognition*, *238*, 105521. https://www.sciencedirect.com/science/article/pii/S0010027723001555

Thorat, S.*, Aldegheri, G.*, & Kietzmann, T. C. (2021). Category-orthogonal object features guide information processing in recurrent neural networks trained for object categorization. In *SVRHM 2021 Workshop, NeurIPS*. https://arxiv.org/abs/2111.07898

Thorat, S.*, Aldegheri G.*, van Gerven, M.A.J. & Peelen, M.V. (2019). Modulation of early visual processing alleviates capacity limits in solving multiple tasks. In *Cognitive Computational Neuroscience*, Berlin, Germany. https://arxiv.org/abs/1907.12309

Appendix E

Donders Graduate School for Cognitive Neuroscience

For a successful research Institute, it is vital to train the next generation of young scientists. To achieve this goal, the Donders Institute for Brain, Cognition and Behaviour established the Donders Graduate School for Cognitive Neuroscience (DGCN), which was officially recognised as a national graduate school in 2009. The Graduate School covers training at both Master's and PhD level and provides an excellent educational context fully aligned with the research programme of the Donders Institute.

The school successfully attracts highly talented national and international students in biology, physics, psycholinguistics, psychology, behavioral science, medicine and related disciplines. Selective admission and assessment centers guarantee the enrolment of the best and most motivated students.

The DGCN tracks the career of PhD graduates carefully. More than 50% of PhD alumni show a continuation in academia with postdoc positions at top institutes worldwide, e.g. Stanford University, University of Oxford, University of Cambridge, UCL London, MPI Leipzig, Hanyang University in South Korea, NTNU Norway, University of Illinois, North Western University, Northeastern University in Boston, ETH Zürich, University of Vienna etc. Positions outside academia spread among the following sectors: specialists in a medical environment, mainly in genetics, geriatrics, psychiatry and neurology. Specialists in a psychological environment, e.g. as specialist in neuropsychology, psychological diagnostics or therapy. Positions in higher education as coordinators or lecturers. A smaller percentage enters business as research consultants, analysts or head of research and development. Fewer graduates stay in a research environment as lab coordinators, technical support or policy advisors. Upcoming possibilities are positions in the IT sector and management position in pharmaceutical industry. In general, the PhDs graduates almost invariably continue with high-quality positions that play an important role in our knowledge economy.

For more information on the DGCN as well as past and upcoming defenses please visit: http://www.ru.nl/donders/graduate-school/phd/

DONDERS INSTITUTE

Radboud University

Radboudumc