Place in perception: The role of structure and individual differences in scene processing

Matthew John Foxwell

PhD

University of York

Psychology

July 2024

Abstract

Our ability to rapidly process scenes despite the complex and varied visual information they contain represents a fundamental contradiction for cognitive science to explain. However, despite their great variation, natural scenes contain visual regularities that are highly diagnostic of their identity and function. Past research has found the visual system takes advantage of these regularities in order to facilitate this efficient processing through the use of internal models, which contain representations of typical scene information and act as a referential template for incoming visual information. In chapter 2, we use a jumbling paradigm to investigate how global scene structure is extracted from scenes, and under which conditions structure impacts categorising accuracy. We demonstrate that whilst disruptions to coherent global structure impacts processing, potential vertical biases observed in previous studies may be better explained by regularities in low and mid-level visual features. In chapter 3, we developed a novel drawing paradigm to describe the contents of internal models of the visual world, in order to investigate how individual differences in conceptions of typicality may drive efficient scene processing. Here, we found that drawings could be used to produce approximations of internal scene models, and that the strength of the match to these internal models was predictive of behavioural categorisation performance. In chapter 4 we conducted 2 further experiments to explore the contents of these internal models, by manipulating the structure and content of renders based on these drawings. However, we failed to find clear evidence for object identity and location being key features of internal models. Taken together, this thesis demonstrates how the visual system utilises structural regularities to facilitate scene processing and provides evidence for a possible influence of individual differences. We further show the ability of drawing paradigms to successfully represent internal models and investigate their application.

Acknowledgements

بِسْمِ اللهِ الرَّحْمٰنِ الرَّحِيْمِ

Thank you to my supervisors, Dr Daniel Kaiser and Dr David Pitcher, for all of your time, wisdom and patience. You've both been absolutely incredible in more ways than I can recount, and I am beyond grateful for everything you have done for me. I am incredibly lucky to have been able to work with two amazing minds, and you have both taught me so much. Daniel- thank you for all your wisdom and kindness, and for giving me this amazing opportunity. David- you helped me through some incredibly difficult times, and I will forever be grateful, and will pay it forward. Thank you both for sticking with me and seeing me to the end.

Thank you to my TAP committee, Dr Karla Evans, Prof Tim Andrews and for the help of Dr Fiona McNab, for all your great advice and feedback, and for helping me on my academic journey.

Thank you to Dr Sally Quinn and Professor Paul Bishop for the amazing opportunity to work as a GTA and tutor in the department. It was a pleasure and privilege to be able to teach, and I don't think I've ever enjoyed any job more. It was amazing to work with so many great lecturers, and to be a part of the students' time at York.

Thank you to all the administrative staff in the department, especially Vicki Hensel, for helping me navigate my somewhat strange contract. I also thank the university staff more broadly- the technicians, security, cleaners, catering staff and grounds keepers- that help keep the University of York such a lovely place to work.

Thank you to all the friends and colleagues working on their PhDs alongside me. To Alex Mepham for his wit, sense of humour, amazing cooking and equally amazing poetry. To Cátia Ferreira de Oliveira for being an amazing office mate and teaching me how to use doors. To my York Dungeons & Dragons party; Nick Souter, Emma Raat, Emma Jackson, Amanda Olsson, and Federico Segal for all the dice rolled and great laughs we had. I know you will all do incredible things, and I wish you all the best of luck for all the adventures ahead.

I thank my amazing family, for their strength and support. You believed in me when I couldn't and have always inspired me to be my best. To my mum and dad for their endless love and support. To Jack, my best friend, and the best big brother anyone could ever want. For everything you have done for me and the family, thank you.

To my incredible wife Manal, I thank you for everything. Words cannot express all you have done for me and all I owe you. You have been my strength when I had none, my light in every darkness, and

the smile in all my days. You have dreamt my dreams and been with me throughout, and I love you with all my heart. My noor, always.

And finally, to my cat, Dubba, for sitting with me throughout this write up, and only attempting to escape once!

It’s been a long road, and I thank you all for walking it with me.

Contents

Abstract.....2

Acknowledgments.....3

List of figures.....7

Authors declaration.....9

Chapter 1: Literature Review 10

1.1 Introduction.....10

1.2 How do regularities in scene structure facilitate natural vision?.....13

1.2.1 Tuning to typical object statistics.....13

1.2.2 Predictive processing 14

1.2.3 Internal scene models.....15

1.3 Regularities in low and mid-level scene properties 17

1.4 Regularities in object positioning 22

1.5 Regularities in global scene structure 26

1.6 Uncovering individual differences in internal models and why we need a new approach for studying internal models.....29

1.6.1 Challenges of the classical approach.....29

1.6.2 Line Drawings in psychological research.....30

1.6.3 Advancements in line drawing methods for cognitive research.....31

1.6.4 Implications for studying Individual differences in scene perception 35

1.7 Goals of the current thesis 36

Chapter 2: It's in the mix – how the composition of outdoor scene elements impact perception 37

2.1 Introduction.....37

2.2 Experiment 1: 180-degree rotation.....43

2.2.1 Methods 43

2.2.2 Results 47

2.2.3 Summary 51

2.3 Experiment2: 90-degree rotation.....53

2.3.1 Methods 53

2.3.2 Results 55

2.3.3 Summary 59

2.4 Comparison between inversion (180°) and rotation (90°) 59

2.4.1 Rationale 59

2.4.2 Results 60

2.4.3 Summary 61

2.5 General Discussion 62

Chapter 3: Individual differences in internal models explain idiosyncrasies in scene perception 69

3.1 Introduction 69

3.2 Methods 70

3.3 Results 77

3.4 Discussion 80

Chapter 4: Investigating the object and spatial content of internal scene models 84

4.1 Introduction 84

4.2 Experiment 1 89

4.2.1 Methods 89

4.2.2 Results 95

4.2.3 Summary 98

4.3 Experiment 2 99

4.3.1 Methods 99

4.3.2 Results 102

4.3.3 Summary 103

4.4 General Discussion 104

4.4.1 Challenges in Detecting Effects of Semantic and Syntactic Manipulations 108

4.4.2 The role of visual clutter and familiarity on scene categorisation 111

4.4.3 Conclusion 115

Chapter 5: Discussion 117

5.1 Results Summary 117

5.2 Parsing structural regularities in scene perception: vertical bias effects may be attributable to low- and mid-level visual features 118

5.3 Internal models, familiarity, and scene processing efficiency 121

5.4 The representation of object information in internal scene models 127

5.5 Drawing as a flexible method for assessing individual differences in internal scenes models 133

5.6 Conclusion 138

Appendices 121

Appendix A: Further examples of the stimuli used in chapter 1 experiment 1 121

Appendix B: Further examples of stimuli used in chapter 1 experiments 2 125

Appendix C: Perspective grids used in the drawing task utilised in chapter 3 experiment 1 and chapter 4 experiment 1 and 2……127

Appendix D: Further examples of scene drawings produced in chapter 3 and chapter 4……128

Appendix E: Further examples of the stimuli used in chapter 3……133

Appendix F: Further examples of the stimuli used in chapter 4, experiment 1 and experiment 2……136

References……140

List of Figures

Figure 1.1. Low level feature averages of different objects and scene categories, produced by averaging the low-level visual properties of 100 exemplars of each type. Originally presented in Torralba & Oliva (2003).....18

Figure 1.2. Schematic hierarchy of a bathroom scene with three phrases consisting of one anchor each (e.g. a shower, a toilet and a sink) that predict the locations of other objects (Vo, Boettcher & Draschkow, 2019). 24

Figure 1.3. Exemplars of the scene drawings produced in Bainbridge and Baker (2019).. 34

Figure 2.1. Examples of stimuli used in the classical scene jumbling paradigm, taken from Biederman (1972)…. 38

Figure 2.2. Examples of design and stimulus, and trial structure.....45

Figure 2.3. Mean reaction times (in milliseconds) across all conditions in experiment 1.....51

Figure 2.4. Mean reaction times (in milliseconds) across all condition in experiment 2... 58

Figure 3.1. Examples of the scene drawings and 3D renders.....73

Figure 3.2. The trial structure for experiment 1 session 2.....74

Figure 3.3. Visualisation of the DNN analysis conducted in chapter 3.....76

Figure 3.4. a) Mean reaction times for renders based on either participants' own, other and control scenes. b) Graded similarity analysis for DNN trained on scenes. c) Graded similarity analysis for DNN trained on objects.. 78

Figure 4.1. Example of how the impact of manipulating a scenes' typical object content is moduldated by how closely the scene resembles a persons' internal model 86

Figure 4.2. Experimental conditions included in experiment 1. 90

Figure 4.3. Examples of the manipulations applied to scenes in experiment 2. 92

Figure 4.4. a). Mean accuracies for the intact own, other and control condition in the scene categorisation task. b) The mean cateogrisation accuracies for the less severe manipulations (where only a single object was either swapped or replaced) compared to the intact conditions between drawers (own, other and control). c) The mean cateogrisation accuracies for the more severe manipulations (where 2 objects were either swapped or manipulated) compared to the intact conditions between drawers (own, other and control)..... 97

Figure 4.5. Experimental conditions included in experiment 2. 100

Figure 4.6. The control condition used in experiment 1 compared with the control condition used in experiment 2 101

Figure 4.7. a). The mean accuracies for the intact own, other and control condition in the scene categorisation task. b) The mean cateogrisation accuracies for the intact, swapped or replaced conditions between drawers (own, other and control)..... 103

Authors Declaration

I declare that this thesis is a presentation of original work, and I am the sole author. This work has not previously been presented for a degree or other qualification at this University or elsewhere. The work was carried out under the supervision of Dr Daniel Kaiser and Dr David Pitcher. This research was supported by a Deutsche Forschungsgemeinschaft (DFG) research grant and funding from the Department of Psychology, University of York, UK. All sources are acknowledged as references.

Chapter 2

Parts of chapter 2 have been published in Cognition: Wang, G.*, Foxwell, M. J.*, Cichy, R. M., Pitcher, D., & Kaiser, D. (2024). Individual differences in internal models explain idiosyncrasies in scene perception. Cognition, 245, 105723. doi.org/10.1016/j.cognition.2024.105723. * denotes joint first authors. This paper contains a second experiment, not included in the current thesis.

Matthew Foxwell contributed significantly to conceptualization, methodology, software, investigation, formal analysis, validation, visualization, project administration, drafting, writing and editing of the manuscript. Gongting Wang contributed to conceptualization, methodology, software, investigation, formal analysis, validation, visualization, project administration, and writing and editing of the manuscript. Dr Radoslaw M. Cichy contributed to the supervision, resources, project administration, funding acquisition, review and editing. Dr David Pitcher contributed to the validation, supervision, project administration, review and editing, Dr Daniel Kaiser contributed to the drafting, writing, visualization, validation, supervision, software, resources, project administration, methodology, funding acquisition, formal analysis, data curation, conceptualization, review and editing. Research assistant Daniela Marinova assisted in the drawing sessions.

Chapter 1: Literature Review

1.1 Introduction

Scenes are the environments that individuals encounter in their daily lives, encompassing a diverse array of settings and object arrangements. Whilst scenes are characterised by their inherent complexity, including diverse elements like clutter, occlusion, objects, textures and colours, they are also highly structured, with global structure stable across categories, and objects frequently placed in typical locations (Bar, 2004; Kaiser et al., 2019a; Kaiser & Cichy, 2021; Oliva & Torralba, 2007; Vo et al., 2019, Vo, 2021). As such scenes represent one of the most complex and varied sources of visual information we experience, and one central to our ability to operate and function within the world around us. Despite this complexity, the visual system is incredibly efficient at processing scenes, able to accurately gauge category information within as little as 200ms (Dima et al., 2018; Kaiser, Turini, et al., 2019; Kaiser et al., 2020a; Lowe et al., 2018).

This contradiction seemingly defies classical work on visual processing, which suggests that the visual system should struggle to process such information rich stimuli. This contrast is well illustrated in the domain of visual search, where it has long been established that distractors greatly inhibit our ability to find targets in artificial scenes, made up of abstract shapes, letters, or isolated objects(Pelli et al., 2009; Pelli & Tillman, 2008; Treisman & Gelade, 1980; Wolfe, 1994). However, research has consistently shown that this is not the case in natural scenes. Thorpe (et al., 1996) conducted a go/no-go task in which participants were required to identify whether an animal was present in a photograph of a natural scene from just 20ms exposure. Despite having no prior knowledge of the target (i.e. what the animal was or where it would be in the scene), participants were incredibly efficient at identifying the animal, with an average accuracy of 95% and median reaction times of 445ms on correct trials. Analysis of event related potential (ERP) trials measured via electroencephalogram (EEG) found that the go/ no-go decision could be decoded after only 150ms from the stimulus onset. This efficiency was achieved despite the photograph containing a large amount of irrelevant visual information, that may have acted as distractors. Functional magnetic resonance imaging (fMRI) research supports this conclusion, indicating little effect of distractors when locating objects within natural scenes and a need for only very brief exposure to be able to extract relevant content (Peelen et al., 2009; Peelen & Kastner, 2011; Seidl et al., 2012). Unlike artificial scenes, visual search in natural scenes is not modulated by set size (Wolfe, Alvarez, et al., 2011), nor attentional allocation (Li et al., 2002). This suggests that not only are we able to accurately search scenes despite their complexity, but that the process is remarkedly quick, indicating that visual search within scenes is an incredibly efficient process.

What is it about natural scenes that makes them different from artificial scenes, and helps facilitate this efficient processing despite their complexity? One possibility is rather than being overwhelmed by this complexity, the visual system could exploit statistical regularities within scene information to rapidly extract meaning. Whilst scenes can vary greatly, they are organised by predictable rules, often governed by physics, context and social norms (Kanan et al., 2009; Torralba et al., 2006). For example, gravity dictates that objects must rise from a base or platform, whilst phototropism results in plants growing upwards towards light sources. Likewise, certain scene elements are much more likely to be found grouped together based on context and meaning, for example we would expect to see mountains in a rural environment as opposed to a busy urban centre. Research increasingly seems to support this idea, with many studies indicating that typical arrangements of a scenes' features such as structure (Biederman, 1972; Kaiser et al., 2020a), layout (Kaiser et al., 2020a; Mannion et al., 2015), object content (Davenport & Potter, 2004), and surface properties (Epstein & Baker, 2019) all help facilitate efficient scene processing (Võ et al., 2019). These, and other regularities may provide scene-based guidance, allowing the visual system to quickly locate and extract the features encapsulating the most prescient scene information (Ehinger et al., 2009; Neider & Zelinsky, 2006; Torralba et al., 2006).

One way that the visual system might achieve this is through the use of predictive processing (A. Clark, 2013; Feldman & Friston, 2010; Keller & Mrsic-Flogel, 2018; Rao & Ballard, 1999). Within predictive processing models, visual inputs are compared against internal models, which represent our expectations of what a stimulus should look like. For scenes, these models might contain stored knowledge based on the regularities we experience in the our everyday visual lives, helping to explain the benefit of typical scene information (Kayser et al., 2004). Such an explanation would align with research evidencing the use of predictive processing in visual perception (Bar, 2009; Peelen et al., 2024) and neural processing of scenes (Kaiser et al., 2019b; Muckli et al., 2015), as well as the control of gaze (Henderson, 2017). Internal models might anticipate sensory inputs, transmitting only the discrepancies between these predictions and actual inputs. This mechanism would reduce the processing load by focusing on unexpected information. Such a mechanism may be particularly beneficial in natural scenes where predictability is high, whilst also containing massive amount of visual information that cannot be individually processed without compromising on the development of a rapid understanding of a person's environment.

If the visual system does make use of internal models to facilitate scene processing, one important question is precisely what information is being predicted? While it seems likely that internal models encode regularities in scene information, such as structure and object identity, we currently lack direct empirical evidence on their nature. Much of the existing research assumes a single, stable

generically “typical” scene across all observers, comparing these against atypical scenes to infer properties of internal models. However, internal models are thought to develop based on prior experiences (Friston, 2010; Köster et al., 2020), reflecting an individual’s personal exposure to the environments they encounter and the regularities within them. Consequently, what one person considers a typical kitchen—its objects, layout, and organisation—may differ significantly from another’s. Evidence for such individual differences can be found in studies suggesting that factors such as culture, socioeconomic background and vocation influence scene perception (Barrett, 2020; Hartley, 2022; Masuda & Nisbett, 2001; Rooney et al., 2017).

Furthermore, there is considerable difficulty in directly accessing the contents of internal scene models. To fully understand how internal models guide scene perception, we need a method that captures these subjective variations without imposing prior assumptions. A more flexible approach would involve directly eliciting descriptions from participants, allowing us to characterise their unique scene models. This would reveal both shared and idiosyncratic elements of internal models, enabling more targeted experiments and refined predictions about scene processing efficiency. If we could harness this variability, rather than overlooking it, we may uncover systematic individual differences in how internal models shape perception, offering a more nuanced understanding of how personal experience influences visual processing. An aim of this thesis is to develop this new flexible approach in order to investigate the content of internal scene models, and to investigate how regularities in scene structure (such as object placements and global scene structure) help facilitate scene processing.

In the following literature review, we will first discuss how the visual system might make use of regularities in scene structure to facilitate scene processing. Next, we review the existing research on how regularities in low and mid-level scene properties, object information and global layout help facilitate efficient scene perception, and how these regularities might be contained within internal scene models. Finally, we outline the need to study individual differences within internal scene models, and discuss the use of line drawings as a flexible method of creating descriptors of these models.

1.2 How do regularities in scene structure facilitate natural vision?

How might the visual system exploit regularities in scene structure to facilitate vision in real-world conditions? In the following section, we will discuss two possible theories outlining how this may occur. The first posits that the visual brain has a rigid tuning to visual regularities more broadly, and regularities are thus considered during bottom-up processing. The second account, inspired by predictive processing theories, posits a more flexible bidirectional mechanism that compares sensory inputs to top-down predictions of likely stimulus properties (including the structure of scenes).

1.2.1 Tuning to typical object statistics

Given the visual system's neural tuning for typical scenes (Davenport & Potter, 2004; Kaiser & Cichy, 2018), the brain may respond differently to scenes that feature typical compositions of constituent objects. This tuning to typical object statistics, in conjunction with tuning to typical distributions of low-level visual features could render the processing of typically structured scenes more efficient in an entirely bottom-up mediated process. Evidence for such a perspective is primarily derived from work showing that object location can be extracted at low levels of visual processing and is apparent in early scene representations (Boettcher et al., 2018; Kaiser & Peelen, 2018; Võ et al., 2019). This is well demonstrated in detection tasks using CFS, where rapid flashing causes a stimulus to appear invisible, until the visual system is able to break this suppression, and the target identified. Detection tasks only require the participant to indicate whether they saw a stimulus or where it appeared, and they do not need to conduct any higher-level tasks such as categorisation. This allows the earliest representations of scenes and objects to be investigated. In such experiments, when objects are placed in typically occurring visual field locations or are organised in meaningful groups, they are detected quicker than when they are placed in atypical locations or non-meaningful groups (Kaiser & Cichy, 2018; Stein & Peelen, 2015). Such results suggest that object regularities are extracted during basic stages of visual processing, indicative of differences in bottom-up visual processing (see Kaiser et al., 2019). The representation of typical object positioning at early, presumably feedforward-related stages of visual processing is also suggested by EEG work (Kaiser et al., 2018). Here, typical absolute locations of objects in the visual field (such as rug in the lower visual field, and a painting in the higher) facilitate object representation within 150ms of processing, again indicating a difference in bottom-up stimulus analysis.

These findings are further supported by research exploring the organisation of receptive fields (RF) in response to the typical organisation of visual features. Research has shown that receptive fields in face and place selective cortices show a similar organisation to visual field biases, with face-selective regions showing RFs closer to the centre of gaze, mirroring the foveal focus used in face processing

(Grill-Spector, 2004; Malach et al., 1995), whilst receptive field organisation in place selective cortices show a greater peripheral layout, indicative of the broader spatial processing required for natural scenes (Levy et al., 2001; Silson et al., 2016). These studies suggest that receptive fields in high-level visual cortex are organised according to stimulus distributions found under real-world processing demands (faces often fall in the fovea, scenes often fall into the periphery). Though somewhat speculative, similar receptive field tuning may facilitate the processing of objects in their typical real-world locations (Kaiser, Quek, et al., 2019).

However, in order for spatial information to be utilised in a strictly bottom-up manner as described, the visual system would need to develop a vast number of tunings for the many valid organisations of objects and their placement within scenes. Given the huge variety of different objects, and the variety of possible typical spatial arrangement between them, this represents no easy task for the brain. Furthermore, a bottom-up explanation also struggles to explain how the visual system differentiates between objects that share similar low-level visual properties and typical visual field locations but provide very different scene category information. In such scenarios, it is difficult to imagine how the brain could determine whether the placement of an object is typical for a scene, and to use this information to facilitate rapid scene processing, without first identifying what the object is.

1.2.2 Predictive processing

The brain may instead rely upon a process in which visual inputs are compared to predictions based on likely stimulus properties. Here, the brain may use stored knowledge about the structure of natural environments to actively generate predictions about likely input properties that can be adaptively compared to incoming stimuli. Predictive processing theories provide a framework for how such a mechanism might work (Clark, 2013a; Friston, 2005; Walsh et al., 2020). Predictive processing posits that in order to function efficiently and avoid disorder, the brain must minimise the level of surprise it experiences. Thus, in order to avoid surprise, the brain must maximise its ability to make accurate predictions. In the case of visual perception, predictive processing models suggest that the brain makes use of Bayesian inferences, where sensory information is stored as probability distributions, representing the likelihood of certain information being apparent. These probabilistic predictions are shaped by prior experiences with a given input (Friston, 2010), thus reflecting the regularities they might contain. In the case of scenes, a probability distribution could contain preconceived knowledge of a scene's typical spatial layout and object content, which it could then compare to incoming visual information, with the strength of the match facilitating how efficiently a given scene is processed. In this way, predictive processing models are able to explain why typicality is so important to scene processing- as it dictates the strength of the match between probability distributions and incoming visual information. In the predictive processing framework, such prior information about likely

stimulus features is often referred to as an “internal model”: an internal representation of probable world statistics that can guide the generation of adaptive predictions. The concept of an internal model conceptually resembles the more classical term of a mental “schema” (Mandler, 1984; Minsky, 1974), which was used to refer to a memory representation that houses typical properties of stimuli that are used to facilitate the perception or memory storage of visual contents.

Predictive coding could also help explain the rapid categorisation characteristic of scene processing. Neural models of predictive processing suggest that the predictions derived from our internal models of what a scene should typically look like, are passed down the processing hierarchy in the top-down direction, which then acts to suppress incoming sensory input that matches those predicted by the internal model, so that only incongruent sensory information is passed on to higher levels of processing (Feldman & Friston, 2010; Rao & Ballard, 1999). This would minimise the overall amount of sensory information being processed, only requiring incongruent sensory information to be passed on to higher stages of processing. In the case of scenes, with their high degree regularity, predictive processing would greatly reduce the amount of visual information needing to be processed to only those features that deviate from what is expected by the internal models.

1.2.3 Internal scene models

Although relatively few studies have directly investigated the role of predicative processing in scene perception, those that have provide promising support. Early research by Rao and Ballard (1999) produced a model that demonstrated how predictive coding might process the structural information found in natural scenes. First, a multi-layered neural network was trained on pictures of natural scenes. Each layer of the network was then tasked with predicting the activity of the next lowest layer, utilising a feedforward mechanism, which made predictions based on error signals from prior predictions. When a prediction was incorrect, the model received an error signal from the layer below as a form of feedback, allowing the model to adjust itself accordingly and generate more accurate predictions with each iteration. After training, the model organised itself into a hierarchy resembling the processing stages seen in the human visual system, with lower levels of the model developing simple cell like receptive fields sensitive to low level visual features such as orientations, whilst higher levels of the model were more sensitive to larger more complex spatial structures, which are more likely to deviate from standard scene statistics and thus be harder to predict. The model shows that feed-forward processing from higher levels is able to predict the simple orientations found at the lower levels, so that higher levels only need to process less predictable elements of scene structure. By demonstrating that predictive coding can explain how structure is extracted from scenes within a framework matching our hierarchical understanding of visual processing, Rao and Ballard (1999) show

that top-down mechanisms, similar to those suggested by predictive processing, are theoretically capable of explaining scene processing.

More recent evidence comes from a series of fMRI studies that investigated the interpolation of occluded information in scenes, which suggests missing information is "filled in" via predictive top-down projections (Morgan et al., 2019; Muckli et al., 2015). These studies exploited the retinotopic organisation of the visual cortex to isolate feedback from internal scene models, by comparing the neural responses to whole and occluded scenes. To achieve this, participants viewed both whole and occluded scenes during fMRI, in which one quarter corner of the scene is occluded. By occluding a section of the scene, neurons processing scene information for that section received no visual input, meaning that any information encoded during viewing must be derived from feedback from later cortical areas. Retinotopic mapping was then used to isolate voxels that responded only to the occluded portions of the scenes, and multivoxel pattern analysis (MVPA) conducted to investigate what visual information had been encoded. MVPA is a neuroimaging technique in which the activity of groups of voxels (typically based on criteria such as proximity, similarity of activity patterns) are analysed together, allowing researchers to better decode information related to a specific stimuli or task-states (Norman et al., 2006). They found that despite the lack of visual input, superficial layers of the primary visual cortex (V1) mapping to the occluded scene sections still contained contextual scene information, indicating that V1 receives some level of feedback information from cortical regions further along the processing hierarchy (Muckli et al., 2015). This conclusion is based on the established mapping between cortical depth and feedforward and feedback projections: neurons in superficial cortical layers of V1 (in contrast to neurons in the middle layers) are primarily involved in collecting feedback from higher-order cortical regions. In a follow-up study, Morgan, Petro and Muckli (2019) used line drawings to investigate what features of the occluded (and thus interpolated) scene segments are fed back during this process. To access approximations of the information stored in internal models they conducted a drawing task in which participants filled in the occluded scene section by drawing what they expected to be behind the white square. All participants drawings were then averaged to create an image representative of the average expectation for all participants, which were then rated by an independent raters to have a high degree of similarity to the actual scene information present within the occluded section. Two key conclusions can be drawn from these data: First, cortical feedback that projecting to early visual cortex "fills in" probable global scene statistics that are inferred from context. Second, drawings can be used as feasible approximations of the content of predictions in the visual system (we shall exploit this strength of drawings later in this thesis, as discussed in section 1.6).

If the brain relies on internal scene models to enhance scene processing, as suggested by the research reviewed above, understanding the nature of these models becomes essential for a deeper insight into scene perception. Perhaps most pressingly is to understand more precisely what the content of these models are, and how they facilitate the match to our internal models. Furthermore, if these internal models reflect our own visual experiences (Friston, 2010), are these differences meaningfully represented within internal scene models? Whilst many studies have demonstrated the visual systems sensitivity to the regularities found within natural scenes, suggesting such features may likely be represented within these models, we currently lack research investigating the contents of internal scene models directly. One reason for this lack of evidence may be the absence of a suitable method for assessing the contents of internal models directly. If we developed a method to read out the contents of internal models, we could make great progress in understanding these questions, and may be able to quantify the extent to which internal models are idiosyncratic, and how such idiosyncrasies impinge on scene perception. In the current thesis, we will attempt to develop a drawing method to help characterise internal scene models (chapter 3), and to explore their content (chapter 4).

In the next section we will explore literature investigating how the visual system utilises regularities in low-level visual properties to facilitate scene perception.

1.3 Regularities in low and mid-level scene properties

Low-level vision refers to the initial stages of visual processing, where the visual system detects and analyses basic features of visual input, such as colour, luminance and contrast. This processing occurs in early visual areas, including the retina and primary visual cortex, and provides the essential building blocks for more complex visual tasks (Groen et al., 2017). Research suggests that regularities in low-level visual features, such as texture, colour, and spatial distributions, play a crucial role in scene processing (Kauffmann et al., 2015; Oliva & Schyns, 2000; Rajimehr et al., 2011). Despite the high inter and intra-category variability of scenes, many low-level features remain surprisingly stable across scene categories (Geisler, 2008), providing reliable and distinctive regularities that the visual system can utilise to interpret and categorise scenes effectively. Much of this evidence comes from neuroimaging studies investigating the sensitivity of scene selective areas to low-level visual features (Kornblith et al., 2013; Lowe et al., 2017; J. Park & Park, 2017; Watson et al., 2014). These areas include the parahippocampal place area (PPA), a region of the posterior parahippocampal gyrus, which has been found to respond strongly to images of places or scenes (Aguirre et al., 1998; Epstein & Kanwisher, 1998), and is believed to play a vital role in scene processing. Additionally, two other cortical areas are strongly associated with scene perception; the retrosplenial complex (RSC) and the occipital place area (OPA), found on the lateral surface of the occipital lobe (Dilks et al., 2013).

Collectively, we will refer to these areas as scene selective cortical regions. In this section, I will discuss examples of how regularities in low-level visual properties, such as spatial frequency, colour and texture, aid scene processing, and what mechanisms might facilitate this advantage.

The stability of low-level visual properties across scenes is well visualised in a study by Torralba and Oliva (2003). They averaged the low-level statistics of 100 different objects and scene types (see Figure 1.1) in order to create protype images, where spatial similarities between local features within the stimuli category are represented by how sharply they are displayed. For objects, that have relatively stable representations, low level averages produce legible exemplars, with colour, form and contrast all being well preserved and communicating clear exemplars of their object type. For scenes, whilst perhaps less clear than in objects, it is apparent even from these simple exemplars that some important scene characteristics are being preserved. Averaging can still produce recognisable exemplars, particularly for the beach and street examples, despite the absence of any recognisable high-level information, such as objects, that we might expect to define them.

Figure 1.1. Low level feature averages of different objects and scene categories, produced by averaging the low-level visual properties of 100 exemplars of each type. Originally presented in Torralba & Oliva (2003).

Regularities in the distribution of colour across scenes have been found to communicate important scene information. Research has shown that we are both better at remembering and categorising images in colour (Goffaux et al., 2005; Homa & Viera, 1988; Spence et al., 2006), indicating that colour communicates some level of diagnostic scene information. However, colour alone may not be sufficient at facilitating this advantage, with research suggesting this advantage occurs only when colours are present in their typical real-world arrangements. Using a go/no-go task in combination

with EEG, Goffaux et al. (2005) found that presenting natural scenes in their typical colours enhances rapid categorisation, as indicated by faster reaction times and higher accuracy, with the effect being less pronounced for greyscale or non-diagnostic colours. Analysis of ERP data also revealed that colour cues influenced neural responses as early as 150ms after stimulus onset, with delayed and weaker signals for greyscale and non-diagnostically coloured scenes. Furthermore, similar beneficial effects of colour have also been observed in studies utilising classifiers to group scenes based on category, where in combination with edge-detection based features, colour could be used to group scenes into city vs landscape categories, and further distinguish between different categories of landscape scenes (such as fields vs forests) (Vailaya et al., 1998). Taken together, these findings suggest the role of colour in scene perception is contingent upon its natural, real-world distribution.

Regularities in colour have been extensively observed to enhance scene memory, with colour found to aid in directing attention and consolidating memory encoding (Dzulkifli & Mustafar, 2013). For instance, studies have demonstrated that participants exhibit reduced recall when scenes are presented in colour during learning and later tested in black and white, or vice versa, indicating that colour is integral to memory representation (Nijboer et al., 2008; Wichmann et al., 2002). Notably, this memory advantage is more pronounced when the colour is consistent with natural scenes; scenes with irregular colour representations, such as purple skies and blue fields, do not show this benefit, suggesting that the visual system utilises regular colour distributions rather than arbitrary colours. While these memory effects help highlight the significance of colour regularities, these effects may be influenced by encoding and retrieval processes, and thus while they do not necessarily indicate a perceptual advantage, they do highlight further the importance of regularities in colour for extracting information from scenes.

Regularities in texture have similarly been found to benefit scene processing. Within scenes, texture often behaves similarly to colour; both are distributed across a broad envelope, able to be broken down into distinctive texture patches, and both can be extracted rapidly (Beck, 1972; Bergen & Julesz, 1983). For example, a beach could be broken down into large patches of sand, water, and rock textures, which may be able to invoke a sense of a scene's identity. However, neuroimaging research suggests that texture alone are not enough to evoke responses in scene selective neural areas. Kornblith et al (2013) found activity in scene selective cortex of macaque monkeys was only sensitive to texture when it was combined with other important scene characteristics, such as depth, viewpoint, and object identity. Similarly, fMRI has also identified specific areas sensitive to texture patterns in the PPA, (Lowe et al., 2017; Park & Park, 2017) providing evidence for this texture sensitivity in human scene selective areas.

However, these regions may be particularly sensitive to texture when it is presented in typically occurring locations, matching the regularities seen in real world scenes. Park and Park (2017) conducted fMRI experiments to explore how the PPA represents texture within scenes. Participants viewed computer-generated room outlines with varying wall, ceiling, and floor textures. In the first experiment, MVPA indicated that the PPA's representation of texture was consistent regardless of its location within the scene, suggesting that the PPA encodes texture information independently of spatial context. However, a follow-up experiment using a repetition suppression paradigm revealed that scenes with identical textures in different locations elicited distinct neural responses compared to those with the same texture and location. This discrepancy implies that while the PPA broadly represents texture across neural populations, it also encodes specific combinations of texture and location at a more localised level. Taken together, these studies suggest that scene selective areas are most sensitive to texture when it is presented in canonical locations, reflecting the regularities we are exposed to in our visual diets.

The importance of colour and texture is further exemplified in a case study of a clinical patient, known as D.F, who had a profound from of visual agnosia (Steeves et al., 2004), meaning she could not identify objects by shape. However, she retained intact colour and texture perception, and was able to identify scenes and objects based on this information (due to representations stored in her long-term memory). D.F had lesions in the lateral occipital cortex (LOC) and the medial occipitoparietal regions which are associated with object processing (Milner et al., 1991), whilst her primary visual cortex and the fusiform gyrus were undamaged (James et al., 2003). Researchers tested D.F.'s ability to classify natural and man-made scenes using only colour and texture across five formats: regular, colour-inverted, greyscale, black-and-white, and 180-degree rotated. Behaviourally, D.F. successfully categorised scenes despite her object recognition deficit, showing the fastest reaction times for regularly coloured images and more errors for black-and-white, greyscale, or inverted colours—closely mirroring neurologically typical subjects (Nijboer et al., 2008). D.F.'s case highlights the ability of low-level visual features to drive scene perception to a degree that allowed D.F to navigate daily life effectively, without the need to identify a scenes constituent objects.

While regularities in colour and texture provide essential cues for scene processing, they do not act in isolation. As discussed, these low-level features alone are insufficient to fully support scene perception. Instead, the visual system integrates multiple low-level cues, allowing for more reliable and efficient scene categorisation (Groen et al., 2017). A central framework for understanding this integration is the concept of scene gist, which describes the summation of global spatial structure using regularities in spatial frequency, orientation, and contrast (Oliva, 2005). Gist can communicate an impression of the scene's category despite lacking any specific high-level details, such as object

content. For example, a city street might be characterised by high verticality and linear perspective, with strong parallel structures suggesting roads and buildings rather than specific objects. Research has shown that even when a scene is blurred or filtered to preserve only its spatial envelope, observers can still accurately categorise it (Oliva & Torralba, 2001). This suggests that the brain combines low-level regularities, including colour, texture, contrasts and spatial frequencies into a structured intermediate, or mid-level representation, which can facilitate the rapid and efficient extraction of scene information (Brady et al., 2017; Castelhano & Henderson, 2008; Nijboer et al., 2008; Oliva & Torralba, 2006). Categorisation can be achieved with this spatial envelop alone, with participants able to accurately categorise scene images that have been filtered so only basic spatial frequency and orientation information is present, even when display times are very short (Greene & Oliva, 2009b; Oliva & Torralba, 2001, 2002; Renninger & Malik, 2004; Torralba & Oliva, 2003). Not only can categorical information be extracted, but also more complex scene properties such as naturalness and openness can be recognised (Greene & Oliva, 2009b), indicating that specific information about a scene can also be extracted from its gist.

Gist may be particularly important for rapid scene judgements, where the need for speed outweighs accuracy. Research exploring memory for scenes by Schyns and Oliva (1994) found that when asked to match a sample scene to a target that had either been filtered with a high pass (where the scenes outline and thus object identities are preserved), low pass (where only the spatial envelop is preserved), or two variations of hybrid scenes (which combined both filtered images but empathised either high pass or low pass information), participants were more accurate at matching samples with low frequency images when exposure times were brief (30ms), and more accurate matching high frequency filtered images when exposure times were longer (150ms). However, whilst this research demonstrates the importance of low-level visual information in informing rapid recall, it is important to note that this effect was observed for a task relying on visual memory, as opposed to representing purely perceptual processes. As such, whilst it is unclear whether the same reliance on low-level visual information would hold for tasks that involve immediate scene perception rather than memory-based matching, it indicates differences in high and low-level visual information might be utilised by the visual system.

It may be that the spatial envelope of a scene is specifically useful in instances where such rapid judgements are needed, whilst in visual experiences where a longer processing time is available, or necessitated by task demands, high level object representations are utilised. This notion is somewhat supported by research showing that attentional allocation is guided more by scene content than by scene gist (Koehler & Eckstein, 2017). However, this study did use a somewhat unconventional definition of gist, where gist was characterised as scene characteristics that implicitly communicated

the scenes category, such as clutter, crowding and object/ background saliency, rather than focusing on the spatial envelope being discussed here. Despairingly, research claiming the opposite is equally inconclusive, either using a similar definition of gist as Koehler and Eckstein (2017) or utilising display times that exceed the extraction of gist, meaning the influence of other higher level scene features cannot be excluded (Hillstrom et al., 2012). As such, whilst research exploring how scene gist interacts with higher-level scene features to aid scene perception is inconclusive, research does suggest the visual system has evolved to take advantage of the broad regularities found in low-level visual features, highlighting again its sensitivity to typicality.

Despite lacking perhaps, the most typifying high-level features of a scene, such as objects or structural arrangements, regularities in low-level features provide the visual system with sufficient information for rapid scene judgements. Whilst the combination of low-level features is critical for forming these quick perceptions of our environments, as evidenced by the rapid categorisation evoked by scene gist, the visual system may also rely on object-based information—regularities in the types and spatial arrangements of objects—to refine and enhance scene perception. In the next section we discuss how regularities in object information contribute to our ability to process scenes.

1.4 Regularities in object positioning

Objects are often implicitly connected to a scene's meaning, either through its function or as a defining feature. This creates an intrinsic part-whole relationship between them, where the meaning of a scene can be derived from the configuration of its constituent objects, but also the relationship between the individual objects can be derived from the function of the scene. For example, a bedroom can be defined by its inclusion of a bed, but the bed is also necessary for the room's function, and so its presence in the scene becomes more predictable. Whilst connected, it is important to distinguish objects from scenes, as they represent distinct visual phenomena. Here, we distinguish objects from scenes based on their visual properties and function: whilst scenes consist of large-scale global environments that we act within, objects are smaller scale local entities that are acted upon (Peelen et al., 2024; Simoncelli & Olshausen, 2001). Research has found that identities and arrangement of objects help the visual system refine scene categorisation (Kaiser, Quek, et al., 2019; Koehler & Eckstein, 2017; Lowe et al., 2017; Võ et al., 2019), working in tandem with low and mid-level properties to construct a coherent percept. This section examines previous research on the role of object regularities in scene processing and how they might inform our internal models.

Võ et al (2019) suggest that a scene's constituent objects, defined as the individual objects that are found within a particular scene, contain two main sources of information: semantic and syntactic. Semantic refers to the object's identity, and how in-fitting it is with the room's category, whilst

syntactic refers to the spatial properties of objects found within the scene. Referencing our previous example, we can illustrate the presence of semantic and syntactic scene information: a bed is a highly typical object to find in a bedroom, and so would have a high semantic consistency with the scene category, whilst a bathtub would be a very unusual object to find and thus exhibit low semantic consistency. Syntactic information is also highly predictable, for example we would expect to see the bed placed on the floor, the right way up, and perhaps with one side placed against a wall (typically the headboard). Other scene objects, such as side tables or lamps could be placed around it, without blocking or interrupting the function of each constituent object. Such an arrangement would be highly typical syntactically. In this way, a scene's object content becomes analogous to the rules of written grammar, where words both have individual meaning and rules that dictate where they should be placed in a sentence. As with previously discussed scene properties, both semantic and syntactic object information often adhere to strict regularities, which our visual system can take advantage of. In this section we will focus on reviewing literature discussing the scenes syntactical rules, the regularities in object positioning and location, both within a space and in relation to one another.

The arrangement of objects within a scene are both highly structured and predictable, depending on the purpose or nature of a given scene (Epstein, 2005; Epstein & Baker, 2019). When objects are found in typically occurring locations there is improved performance in numerous scene-based tasks, such as object memory (Coco et al., 2016; Konkle et al., 2010; Mandler & Johnson, 1976) and visual search (Brockmole & Henderson, 2006; Peelen & Kastner, 2014; Torralba et al., 2006; Wolfe, Alvarez, et al., 2011). This advantage is present even when objects appear in isolation, without any contextual scene information. Research using continuous flash suppression (CFS) was used to mask objects presented at different visual field locations when they either adhered to, or violated, the spatial positions they occupy in typical scenes (e.g., placing a lamp in the upper visual field would be atypical, as they usually rise from a base in the lower visual field). Objects presented at typical locations would break suppression faster than those that were not (Kaiser & Cichy, 2018), indicating that the visual system utilises regularities in where an object is typically positioned to aid object processing. These findings are supported by neuroimaging studies that have found neural representations of objects are more efficiently decoded in object-selective lateral occipital cortex (LOC) when they appear in retinotopic locations corresponding with their typical locations found in scenes, with this encoding occurring within 140ms (Kaiser et al., 2018; Kaiser & Cichy, 2018a). This rapid encoding suggests that this location-based sensitivity occurs at the early stages of scene processing before feedback from other higher level cortical areas can be implemented. Together, this work suggests that the visual system is strongly attuned to the absolute retinotopic locations of objects matching those of their canonical placement in real world scenes. This sensitivity occurs at early stages of visual processing,

indicating that such structural regularities could be utilised by later stages of visual processing, highlighting the potential utility of these regularities to the visual system.

As well as being viewed in typical positions within scenes, objects are also frequently found in predictable positions relative to each other. For example, a computer monitor is almost always found on top of a desk, with a keyboard in front of it and perhaps speakers either side. As with single objects, the visual system becomes accustomed to this typical spatial arrangement and can take advantage of this. This has been shown in a number of behavioural tasks, such as in object detection (Stein & Peelen, 2015), identification (Biederman et al., 1982) and in improving the accuracy of visual memory (Draschkow & Võ, 2017; Gronau & Shachar, 2015; Kaiser et al., 2015). Research investigating this effect on visual perception specifically has found that viewing typical object pairs relevant to a specific scene category in isolation invokes similar patterns of activation in the LOC as viewing that scene directly (MacEvoy & Epstein, 2011).

However, within these object pairs, some individual objects may be more diagnostic then others. In a novel virtual reality paradigm, where participants constructed representations of different scenes, participants consistently placed larger objects (such a tables, counters and baths) first, before placing smaller objects around them (Draschkow & Võ, 2017). These larger objects may act as anchors, key

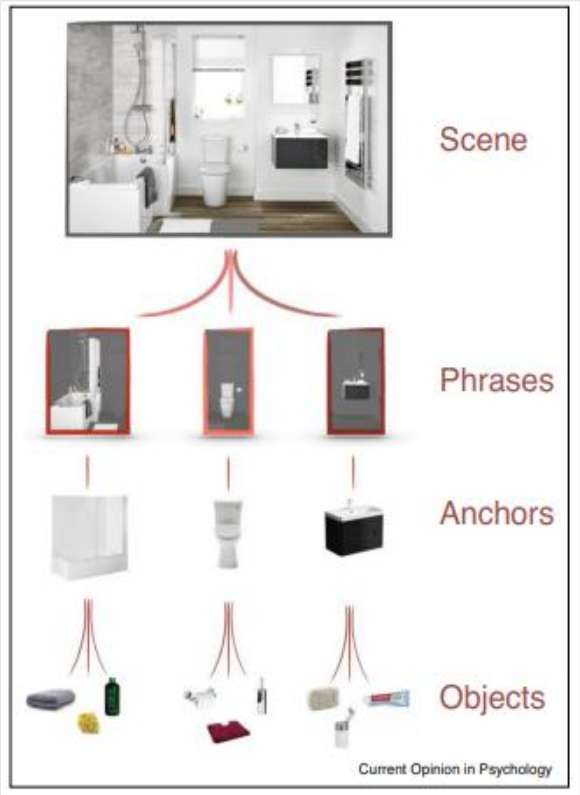


Figure 1.2. Schematic hierarchy of a bathroom scene with three phrases consisting of one anchor each (e.g. a shower, a toilet and a sink) that predict the locations of other objects (Võ, Boettcher & Draschkow, 2019).

points of reference for other objects within a scene. Võ, Boettcher and Draschkow (2019) suggested a framework for this relationship, in which anchors act as key frames of references for other objects within a scene (see Figure 1.2). As such, smaller separate clusters of objects may be connected to different anchors (such as soap, flannels and toothbrushes around a sink) in order to create phrases, which in turn are combined to create scenes.

Whilst there has been little exploration of the role of anchor type objects in guiding scene processing specifically, previous experiments have used large anchor like objects (such as cabinets, sinks, and shower stalls) as stimuli when demonstrating the effect on typical object placement in facilitating efficient scene perception (Kaiser & Peelen, 2018; Linsley & MacEvoy, 2014). If a combination of object pairs and anchors can be utilised to characterise a scene, these object relationships might be represented within our internal models. If this is true, we might expect that manipulating anchor objects within scenes more closely matching our internal models to negatively impact scene processing.

An alternative explanation for the benefit of object pairs occurring in typical relative location is the occurrence of reduced inter-object competition within the visual system. Inter-object competition occurs when different visual elements of a scene compete for the limited processing resources available to the visual system. Kaiser, Stein and Peelen (2014) demonstrated this in an fMRI experiment where they measured the activity of the PPA to images of houses when they were presented alongside objects placed in either typical or irregular relative positions to one another (for example an egg above an egg cup vs an egg cup above an egg). Their rationale posited that the PPA would show higher levels of activation when inter-object competition was low, as more resources could be allocated to the processing of the preferred element (i.e. the house). They found that PPA activation was indeed higher when objects were typically arranged, suggesting the visual system exploits the knowledge of these regularly occurring positions to group objects and process them together, lessening the perceptual load of a scene and contributing towards efficient scene processing.

Given the strong representation of the regularities in object positioning, it seems plausible that such information is represented within our internal scene models (Bar, 2004; Biederman et al., 1982; Kaiser, Quek, et al., 2019). If regularities in object structure do help drive efficient scene processing, the visual system would require a reference for this typicality, which could be facilitated by matching content to internal models. Scenes featuring more typical object arrangements could be more easily indexed against these internal models and thus be perceived more efficiently. How these regularities in object arrangements and identities are represented within internal models is a central question that this thesis seeks to explore. This will be investigated more directly in chapter 4, where we manipulate

scene content based on proxies of participants' internal models to better understand how object-level information is stored and utilised within these models.

1.5 Regularities in global scene structure

Other studies have taken a different, complementary approach to investigating the spatial regularities present in scenes. Instead of manipulating the arrangement of constituent objects, they have disrupted the congruence across a whole scene in order to disrupt the global scene structure. Global scene structure refers to the overarching organisation and arrangement of elements within a scene, encompassing both the spatial configuration of objects and the structural norms. Scene selective regions show a distinct sensitivity to many regularities in global scene structure, such as size (Park et al., 2015), openness (Henderson et al., 2011), and geometric dimensions (Dillon et al., 2018; Ferrara & Park, 2016; Henderson et al., 2008). The structural information contained within the overall geometric structure of a scene seems to be particularly diagnostic, with studies showing that even empty scenes, devoid of any constituent objects, can evoke neural activity in scene selective regions, (Epstein et al., 1999; Epstein & Kanwisher, 1998; Kamps et al., 2016; Wolbers et al., 2011), suggesting a strong sensitivity to geometric structure.

Classical work by Biederman (1972) investigated the effect of global scene structure using jumbling paradigms, in which scenes are divided into segments and then shuffled to disrupt their global spatial structure (creating a similar effect as a slide puzzle). The result is an image where segments from different parts of a scene are recombined so portions that might typically be found at the top of the image could instead be found at the bottom, or from the left to the right etc, with the degree of jumbling being modulated by how many segments the original image is divided into. Behavioural studies have shown that disruptions to the global scene structure caused by jumbling impair immediate scene perception. This impact is evident in tasks requiring rapid scene categorisation (Biederman et al., 1974) and object recognition (Biederman, 1972; Biederman et al., 1973), where structural inconsistencies interfere with real-time processing.

Jumbling is a coarse manipulation of scene structure, which results in the simultaneous disruption of several aspects of scene information. When segments are moved, both the relative object positioning across segments (Kaiser, Quek, et al., 2019; Kaiser & Peelen, 2018), and the scene's overall spatial geometry (Dillon et al., 2018; Spelke & Lee, 2012) are disrupted. This high degree of spatial disruption is achieved while maintaining categorical and local object information, which remains contextualised within the individual segments, with only their absolute locations being altered. These properties allow jumbling paradigms to isolate the effect that global spatial structure

has on scene perception regardless of contextual and local object information. By preserving these scene qualities, the jumbling paradigm provides strong evidence that global spatial structure is utilised separately from object placements and relative positioning, representing a distinct source of visual information utilised during scene processing.

Jumbling was used in a series of recent studies investigating the role of intact global structure in extracting categorical information from scenes. Kaiser et al (2020b), found that when a scenes spatial structure was jumbled, extraction of categorical information was greatly decreased, and that intact spatial structure could facilitate the extraction of categorical information as early as 200ms. In order to distinguish whether the effect of jumbling on categorisation arose primarily from genuine disruptions to the global scene structure, or from disruptions to categorical-level information, Kaiser et al (2020a) compared neural activity in participants passively viewing scenes that had either their spatial or categorical information disrupted through jumbling. Categorical information was jumbled by replacing segments of the target scene image with segments taken from scenes belong to a different scene category, but arranged in a spatially consistent manner to preserve their global scene structure.

Using MVPA they found that whilst jumbled scene structure impacted cortical processing in scene selective regions, there was no effect for categorical jumbling, indicating that the effects of jumbling are not a result of disruptions to categorical information within the scene. Supporting their previous findings, they found the difference between spatially intact and jumbled scenes emerged rapidly, within 255ms. Crucially, they also included an inverted condition for both stimulus types, showing that the results could not be explained by the formidable disruptions to low level visual features present in jumbled stimuli (such as the displacement of large texture and colour patches present in typical scene structures representing portions such as sky).

In a final fMRI experiment Kaiser et al (2021) found that the jumbling scene structure impacted scene perception regardless of the task demands. They conducted MVPA on imaging data collected while participants completed an object or scene categorisation task, in which targets were displayed in either intact or jumbled scenes. Whilst they found stronger neural representations for scenes when spatial structure was intact regardless of the task, jumbled structure only weakened representation for objects during the object classification task. Taken together, these studies suggest that the use of intact global structure in scene perception occurs rapidly, and regardless of task demands. The authors suggest that as this rapid extraction of scene structure occurs later than typical object content aids scene processing (Draschkow et al., 2018; Ganis & Kutas, 2003; Mudrik et

al., 2010; Võ & Wolfe, 2013), that these results could indicate a separate stage of scene analysis dedicated to the structural arrangement of scenes holistically.

If regularities in global scene structure are important for scene perception, then we would expect this effect to be modulated by its adherence to the regularities we experience in real world scenes. For example, the global scene structure of real-world scenes tends to follow a consistent vertical organisation due to gravitational and structural constraints, with almost all natural scenes consisting of a base from which objects arise (Vaziri & Connor, 2016). Conversely, horizontal structure can vary significantly more whilst still evoking structurally typical scenes. These structural regularities in natural scene structure could be stored within internal models, and might result in a "vertical bias" for visual information arranged along the vertical axis of a scene. Regularities in global scene structure might also reflect prominent differences between structural areas of a scene, such as the difference between earth and sky found within outdoor scenes. Previous research has found that the horizon, which acts as the boundary between these two structural components, can be identified very rapidly based only on a scenes low-level visual information (Herdtweck et al., 2010), suggesting that the visual system might utilise the horizon to facilitate swift scene processing. In chapter 2 we explore this question, investigating whether the advantages of intact global scene structure are reflective of real-world structural norms. We do this by comparing the effects of jumbling across different scene axes, in order to establish whether a vertical bias exists when extracting global scene structure, reflecting the bias for vertically arranged structural information found in real-world scenes.

Taken together the research reviewed in this section highlights the critical role of global scene structure in scene processing, demonstrating that disruptions to this structure—such as through jumbling—impair both rapid scene categorisation and object recognition. The ability to rapidly extract global scene structure appears to function as a foundational stage of scene processing, distinct from object-based recognition mechanisms. Whilst we posit that the advantages of intact global scene structure may reflect the regularities observed within real-world scenes, such as a stable vertical organisation, we aim to explore this hypothesis more directly in chapter 2.

Thus far, we have reviewed how regularities in low and mid-level properties, object positioning, and global layout aid scene perception, and how this might be reflective of information stored within our internal models. However, if internal scene models are shaped by our prior experiences, could the information they contain vary on an individual level, reflective of our own individual visual experiences? In the next section, we will discuss the importance of investigating these individual differences, and methodologies that may help us do so.

1.6 Uncovering individual differences in internal models and why we need a new approach for studying internal models

Individual differences are widely neglected in the scene perception literature. Although there is consensus that many structural regularities found in natural scenes are utilised during scene processing , it is not yet well understood how differences in conceptions of these regularities might modulate this effect. This has stemmed from researchers investigating these effects assuming a shared concept for what is typical, relying upon either their own intuition (Davenport & Potter, 2004; Kaiser, Quek, et al., 2019; Võ et al., 2019; Wolfe, Alvarez, et al., 2011) or panels of raters (Torralbo et al., 2013) to determine what represents a typical scene. This results in generically typical scenes, assumed to be equally so for all participants, that then acts as a baseline to compare any manipulations against (such as changes in object content or structure).

However, if internal models are shaped by our own individual prior experiences, this classical approach (as we will refer to it here on) would miss out on these differences. A typically generic scene shared across participants may be more or less typical for each individual participant, depending on their prior experiences and expectations. This could lead to an underestimation of the effects of typical scene statistics on perception, as the manipulations applied to violate the generic scene's typicality would not affect each participant equally. For example, if an individual participant's internal model of a scene happened to differ greatly with that of the generically typical scene, the effects of such a manipulation would be weaker than expected for that participant, if exhibited at all. If instead we could create stimuli that reflects each individual participant's idea of what the most typical instance of a scene should look like (and thus reflecting their internal model of the scene), we could construct experiments that not only honour individual differences in visual experiences but also allow us to more fully explain the variance in scene perception across the population.

1.6.1 Challenges of the classical approach

Whilst the assumption of a shared conception of typicality represents a major barrier in identifying individual differences within scene processing, several other limitations hold back the classical approach. For example, the classical approach only allows for more rigid manipulations to a scene, in order to isolate which manipulations drive any observed effects. A test of all possible combinations of a scenes features, without insights on which of these statistics most prominently feature in internal scene models, is hardly feasible within a single experiment. This makes it hard to quantify which factors are more or less important to the internal model: For example, in the internal model of a living room, is the position of the sofa more important than the size of the table? By only manipulating a single feature at a time, it would be difficult to capture this relationship between scene features,

without designing an experiment with an impractical amount of stimulus conditions and permutations.

Another problem of constructing typical and atypical scenes at the discretion of the experimenter is that such experiments often expose participants to highly improbable and hence artificial scenes. This problem is especially apparent in studies of object-scene consistencies (Chen et al., 2022; Davenport & Potter, 2004; Munneke et al., 2013; Võ & Wolfe, 2013), where highly improbable objects are shown within a scene (e.g., a priest in a football field) or when using jumbling paradigms (Biederman, 1972; Kaiser et al., 2020b), where the jumbled condition is highly artificial and unlike anything people encounter in the real world. This may result in observed effects being influenced by the novelty of these strange stimuli, instead of the manipulations being applied.

1.6.2 Line Drawings in psychological research

Given the limitations of the classical approach it becomes apparent that in order to investigate internal models, a new, complimentary approach is needed. Ideally, such a new approach would focus on the contents of internal models more directly, rather than inferring them indirectly. Such an approach needs to be flexible enough to characterise an individual's internal model in an unconstrained manner, allowing them to express the contents and layout of a scene as closely as possible to their own expectations. It also needs to be practical for experimental use, in the sense the quantifications of internal models should be obtainable in a short amount of time and in a way that is intuitive for participants.

In order to achieve this, the participant, and not the experimenter, would have to lead in the construction of stimuli representative of their own idea of scene typicality, and thus approximating their internal model. One promising method that allows for this flexibility is line drawing. Line drawings allow participants to communicate scene content with a high degree of freedom, being able to include any objects, their properties and layout. Line drawings have changed little over the course of human history and may have developed as a stable way for humans to communicate important conceptual visual information, including the content and layout of different environments (Cavanagh, 2005; Sayim & Cavanagh, 2011). They can be seen as functional abstractions of the way in which an individual perceives the world, able to "exploit the underlying neural codes of vision" (Sayim & Cavanagh, 2011). As such, line drawings may provide a direct and intuitive way of conveying the contents of internal models, equipping us with a novel method of accessing the information encoded within them.

Historically the use of line drawings in psychology has primarily been in clinical settings, used as a diagnostic tool for neuropsychological conditions. The complexity of drawing, and the range of

cognitive skills required to translate mental representations into a sequence of motor commands, makes it a useful indicator for dysfunctions in a range of cognitive processes, such as memory, motor control and mental imagery (Smith, 2009). Various drawing tasks have been used to help diagnose memory disorders (Pinto & Peters, 2009), dementia (Herrmann et al., 1998), apraxia (Warrington et al., 1966), spatial neglect (Agrell & Dehlin, 1998), and in the study of lesions in the parietal lobe (Makuuchi et al., 2003). Projective drawings are also used in a number of examinations used to explore the psychological state of children and individuals with communication difficulties, in order to assess thoughts and emotions they may find difficult to describe or vocalise (Bekhit et al., 2005; Thomas & Jolley, 1998; Woolford et al., 2015),.

Within cognitive research, drawing has historically been used sparingly, with its use primarily in studies exploring memory and recall (Bainbridge et al., 2021; Intraub & Bodamer, 1993; Rubin & Kontis, 1983). Of particular methodological relevance to the current thesis, Rubin and Kontis (1983) used a drawing paradigm to explore memory schemas for coins, in order to understand what information is held in mental representations of familiar objects. They asked participants to draw different U.S coins from memory, and compared these against the drawings produced by other participants and those of real-world coins. They found that instead of representing the different types of coins separately, key features (such as inscriptions and which president was depicted on the heads side) were shared across coins, suggesting that schemas for familiar objects might exist for broad categories, rather than being object specific. Whilst this early work shares a similarity in experimental approach with our current aims, it is limited by its inability to objectively compare drawing output, instead relying on experiment judged scoring systems. Another key difference is that Rubin and Kontis (1983) used drawings to describe participants visual memory of an object, while the work in the current thesis aims to explore perceptual effects. Whilst overlapping, their underlying neural mechanisms differ, limiting how findings in drawing studies on memory can be applied to visual processing.

1.6.3 Advancements in line drawing methods for cognitive research

These early drawing experiments relied greatly on experimenter judgement in order to assess drawing content, a more subjective and complex measure compared to alternative behavioural measures like accuracy and reaction times (Fan et al., 2023). This may have deterred interest in the methodology by researchers for some time, stunting the development of the drawing paradigm further. This limitation is exemplified in the second experiment of aforementioned Rubin and Kontis (1983). Here participants were asked to draw a speculative new denomination of coin (such as a 7 pence coin) from their own imagination, in order to investigate what features they would include on such coins, without the influence of memory. In order to analyse the drawings, they used their own ratings to compare the speculative coin drawings from those produced from memory. Finding little difference between them,

they concluded that participants use a generalised coin schema, rather than schemas for specific coin denominations. However, the experimenters only compared the coin drawings on a number of pre-defined characteristics, such as the location and inclusion of key text, dates and denominations, and not on more granular details such as the figure included on the heads side, or more intricate design features. If the experimenters had been able to access a more objective measure, capable of comparing visual features in a less defined way, more nuanced comparisons might have revealed differences not immediately obvious to the researchers, yielding a more comprehensive and objective data set.

However, recently there has been a renewed interest in the use of line drawings as a method for studying cognition, in part spurred by technological advances that make it easier to collect and analyse drawing content. Deep neural networks (DNNs), trainable computer models based on the neural architecture of the visual system, allow for images to be analysed on visual characteristics in a manner analogous to the ventral visual pathway of primates, allowing for more objective comparisons to be made against line drawings (Fan et al., 2018; Jongejan et al., 2016; Yamins et al., 2014). This allows experimenters access to a more objective metric to analyse the content of line drawings and compare this with other drawings or real-world images, helping remove the requirement for experimenter judgement. Further, methodological advances in utilising drawings have also been made in recent years, with efforts to help establish more standardised procedures for line drawing experiments. Much of this work has been done by the Bainbridge lab at the University of Chicago, who developed a framework for capturing mental representations using line drawing and crowd-sourced scoring (Bainbridge, 2022), with the aim of highlighting the capabilities of drawings as a research method and standardising experimental practises, in order to develop a reliable methodological consensus on best practise. This framework not only outlines how drawing experiments can be successfully conducted in a typical laboratory setting, but also how drawing tablets can be used to collect additional information about the drawing process, such as the order in which features are drawn, allowing for a greater range of hypotheses to be tested.

By using drawings, a small number of modern studies have successfully investigated participant's mental representations without any prior assumptions about their properties. In addition to the previously described fMRI work conducted by Morgan, Petro and Muckli (2019), Bainbridge and Baker (2019, 2020) have published a series of papers using drawing to investigate scene memory, that exemplify the strengths of the methodology. In their first study, Bainbridge and Baker (2019) used drawing based visual recall tasks to investigate scene memory. In these experiments participants were asked to study photographs of real-world scenes and then draw them from memory after a distraction period. They conducted four different versions of this task, in order to investigate whether the recall

period (either delayed, immediate or drawn directly from each image) or the category information alone, would modulate the nature of the mental representation and subsequent recall. Drawings provided a measure for the scene content participants could recall, allowing the experimenters to utilise a free-recall technique with complex visual stimuli, in which participants also had complete control over which scene elements were included without the influence of the experimenter or other visual prompts (see Figure 1.3). The contents of these scene drawings were then quantified by large panels of crowd sourced raters to assess the drawings on a number of criteria, including the number, type and size of constituent objects and spatial details, so that they could be compared against the original scene images, to ascertain which image metrics could predict efficient recall. They found that drawings, across all 4 recall conditions were highly diagnostic of the original image. The drawings produced after recall contained a greater degree of diagnostic visual information, suggesting that the drawings represented a specific recall of the original scene images, and not just a categorical representation. Participants were able to construct visually detailed scenes, with few mistakes when recalling scene content, suggesting strong recall for the constituent scene objects.

Of methodological interest, participants were able to produce these detailed images regardless of no prior selection criteria for artistic ability, suggesting the method's suitability is not limited by drawing ability. Furthermore, the experimenters noted little difference in mistaken objects between pictures drawn from category level descriptions and those from memory, suggesting that participants did not take more artistic liberty in one condition over the other. As well as the scene content, they also observed the spatial layout of the scenes were accurately represented in images, suggesting that the mental representations of these scenes were not simply a visual representation of remembered constituent objects, but that the object-by-object spatial relationships were also represented. Taken together, this suggests that scene drawings are able to capture both the spatial and object information contained within scenes, indicating that these fundamental properties of scenes can be communicated accurately through the use of drawing tasks.

Figure 1.3. Exemplars of the scene drawings produced in Bainbridge and Baker (2019). Participants were instructed to draw scene drawings from a number of different categories across 4 recall conditions; delayed recall in which participants drew the images from memory following a distraction task, immediate recall where drawings were produced immediately after study, image drawing where they could draw the pictures with direct access to the original image and a category drawing where they received only a category name for each image type. Figure from Bainbridge and Baker (2019).

In a subsequent experiment, Bainbridge and Baker (2020) were able to reuse the drawings produced in their original study to investigate a new hypothesis. In many of the recall drawings produced, they noticed that the participants would often contract the boundaries of the original image. This behaviour is at odds with a widely reported scene memory phenomena observed in many experiments known as boundary extension, in which observers recall visual information about a scene extending outside of the boundaries of the originally observed view (Hubbard et al., 2010). Subsequently, they reanalysed the original drawings using online raters and additional experiments to assess boundary contraction within the drawings. They found no difference between the proportion of images judged as contracting or expanding, prompting them to compare the qualities of their drawing stimuli with a scene set used for most boundary extension experiments. They found that compared to participant's drawings, the scenes used in the original boundary extension literature contained objects displayed at a low angle, and suggested that the observed effect was instead a result of participants attempting

to realign these objects to a more canonical angle, challenging the assumptions of boundary extension. This illustrates the utility of the drawing paradigm and the richness of drawings as a source of data to challenge previously held assumptions.

Drawings have the potential to act as both a rich, explorative source of data, and to be used as reliable stimuli in more traditional experiments, allowing experimenters to utilise image sets constructed without their bias or influence. This is further complimented through technological advances, such as the advent of new image processing techniques, that allow for complex visual information to be analysed without the use of panels of raters (Fan et al., 2018), When used, crowd sourcing websites allow massive panels of raters to be recruited in order to analyse image content, that would be time demanding and impractical through traditional recruitment avenues, helping produce more reliable measures and reducing rater bias.

1.6.4 Implications for studying Individual differences in scene perception

While line drawings have been successfully used to study scene memory, their application to scene perception remains largely unexplored. Despite scene perception and memory engaging some overlapping neural mechanisms (Dalton & Maguire, 2017; Steel et al., 2021), they remain distinct neural mechanisms dissociable from each other (Bartolomeo, 2002; Bartolomeo et al., 1998; Behrmann et al., 1994; Epstein & Baker, 2019). As such it is unclear whether drawings can capture perceptual processes in the same way they reflect scene memory. However, the research reviewed above suggests that drawings provide a functional abstraction of scene content, effectively communicating spatial layouts and key visual features (Bainbridge et al., 2019; Bainbridge & Baker, 2020; Fan et al., 2023; Sayim & Cavanagh, 2011). This suggests that despite the differences between perception and memory, drawing techniques may still offer valuable insights into scene perception. Drawings could allow experimenters to objectively quantify key properties of participants' internal models and then probe these properties in targeted investigations, in order to explore the representation of different facets of scene information within them. As such, we conclude that the potential for drawings to reveal perceptual representations for scenes warrants further investigation, which will be explored in the current thesis. By doing so, we hope to open a new avenue in the study of scene perception, in which we are able to consider the individual differences that may shape our internal scene models.

1.7 Goals of the current thesis

The current thesis will aim to address two complementary questions about how scene structure impacts perception. First, we will aim to explore how regularities in global scene structure are extracted from scenes, and under which conditions it impacts scene perception. To do so, we will adopt a more traditional approach to manipulating scene structure, utilising a similar scene jumbling method as used in classical work by Biederman (1972). Specifically, we will use the jumbling paradigm to investigate how the global organisation of scenes along the horizontal and vertical axes impacts perception and how such effects are dependent on the canonical (upright) orientation of scenes.

In the later chapters, we focus on developing and utilising line drawings to describe the contents of internal scene models. In our experiments, we will ask participants to draw what they consider a typical instance of a scene (e.g., drawing a typical kitchen). These drawings are then used as descriptors of individual participants' internal scene models, allowing us to construct stimuli that are typical or atypical for individual participants. In a series of experiments, we test whether individual participants are indeed more efficient in categorising scenes that more strongly resemble their own internal scene models. We further investigate which properties of the scenes determine such performance benefits, and specifically focus on whether the presence of certain diagnostic objects or their positioning across the scene impinge more strongly on behavioural performance.

Chapter 2: It's in the mix – how the composition of outdoor scene elements impact perception

2.1 Introduction

The visual system is incredibly efficient at extracting information about our environment, able to process many attributes of visual scenes within 200ms of them being displayed, despite the rich amount of information scenes contain (Dima et al., 2018; Kaiser, Turini, et al., 2019; Kaiser et al., 2020a; Lowe et al., 2018). This efficiency is unsurprising when considering the relevance of scene processing to everyday tasks, many of which require differing conceptualisations and understandings of a scene (Malcolm et al., 2016). For example, when trying to find a new building on campus, you might need to recognise the building, identify where it is, and then understand how to navigate the environment to reach it. The information contained within scenes, whilst varied, often adheres to intrinsic rules and regularities (Geisler, 2008), which can convey meaningful information that the visual system can take advantage of. One of the main sources of this information is a scene's spatial structure, which refers to a scene's spatial layout and the relative positions of objects and scene elements (Kaiser, Quek, et al., 2019; Oliva & Torralba, 2007; Võ et al., 2019; Wolfe, Alvarez, et al., 2011).

The spatial regularities we extensively experience in our everyday lives are mirrored in cortical sensitivity to scene structure. The PPA, OPA and RSC have all demonstrated sensitivity to a scenes overall geometry, responding strongly to images of empty scenes, even when no objects were present (Epstein et al., 1999; Epstein & Kanwisher, 1998; Kamps et al., 2016; Wolbers et al., 2011). FMRI studies have shown the PPA may be particularly sensitive to global structural properties , such as size (Park et al., 2015), openness (Henderson et al., 2011), and it's geometric dimensions (Dillon et al., 2018; Ferrara & Park, 2016; Henderson et al., 2008). However similar preferences have been found for both OPA and RSC, suggesting that all scene selective regions may be sensitive to the fundamental structural properties of scenes (Bonner & Epstein, 2017; Dillon et al., 2018; Henriksson et al., 2019).

This cortical sensitivity to scene structure is also reflected in perceptual efficiency, with classical behavioural work by Biederman (et al., 1974) illustrating our reliance on scene structure through the use of jumbling paradigms. Here, scenes are divided into segments, and rearranged so that the pieces are no longer presented in their typical spatial arrangements (see Figure 2.1). Disrupting scene structure in this way has been shown to impair scene categorisation (Biederman et al., 1974), in-scene object recognition (Biederman, 1972; Biederman et al., 1973) and detection of subtle visual changes within a scene (Alexander Varakin & Levin, 2008; Zimmermann et al., 2010). Neuroimaging research by Kaiser et al (Kaiser et al., 2020a; Kaiser et al., 2020b) employed a similar jumbling paradigm to

investigate the effect of jumbling on the cortical processing of a scene's categorical information. In a series of related experiments, participants passively viewed both upright and inverted intact and jumbled scene images whilst undergoing fMRI and EGG. Using multivariate decoding they found that jumbled scenes were represented differently from intact scenes in both the PPA and OPA, and that these areas demonstrated a sensitivity towards intact scene structure. Crucially, they found a reliable inversion effect, meaning that when the scene images were shown at inversion, decoding was worse than when they were shown upright, indicating that this sensitivity was induced by the jumbling of the scenes' structural information, as opposed to manipulations to their low- and mid-level visual features.

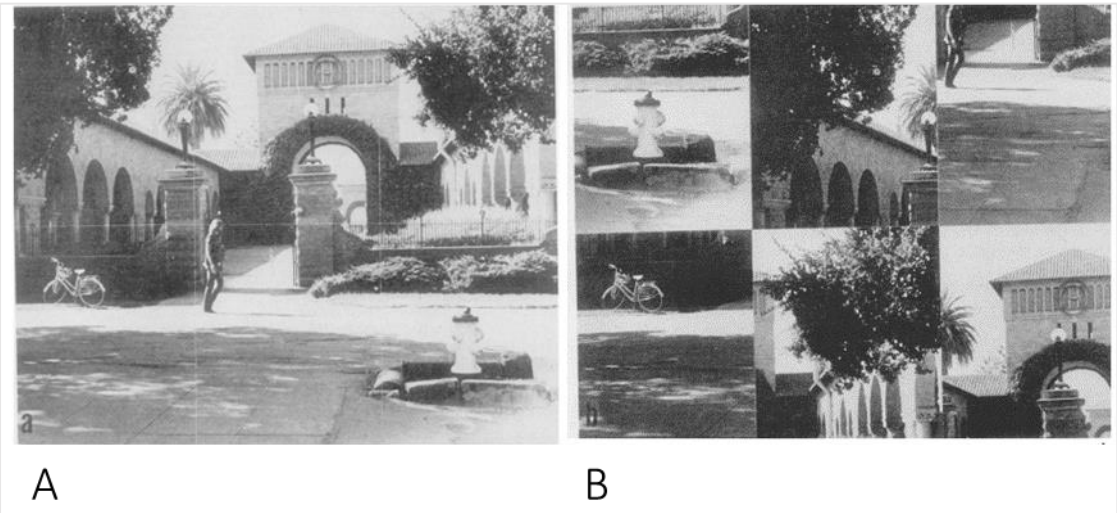


Figure 2.1. Examples of stimuli used in the classical scene jumbling paradigm, taken from Biederman (1972). The original image (A) was divided into 6 segments and rearranged so that the scene structure is disrupted (B).

If coherent scene structure is important for the processing of scenes, as previous research suggests, one fundamental question arising is whether intact scene structure is equally important across the different spatial dimensions of a scene. Chiefly, a scene's structure can be divided across two axis: its horizontal and vertical axis. These axes can present distinctive information about a scene, for example along its horizontal axis a clear horizon might be established, whilst across the vertical axis you might experience a shift from the ground plane to sky. Comparatively, horizontal organization is often more arbitrary, and can vary greatly, while vertical structure tends to be more rigid (see Adams et al., 2016). For example, scene elements tend to follow a consistent vertical organisation due to gravitational and structural constraints. Almost all natural scenes will compose of a base (be it floor or ground), from which objects arise, and a skyline. This vertical rigidity is true for many individual scene elements as well. For instance, buildings tend to have roofs which are always above walls, and levels that are

stacked vertically in a predictable sequence. If this vertical structure is violated, it can produce highly unusual or even naturally impossible scenes (such as those where the horizon is above the skyline, or where objects would float instead of standing on a base). By contrast, horizontal structure can vary significantly whilst still producing highly typical scenes. For example, buildings in an urban scene can vary significantly but still be easily identifiable; whether a church is placed to the left or right of a bakery it still produces an equally valid exemplar of an urban scene. As a result, the amount and variety of objects that can be found along the horizonal axis is far less constrained, and more dependent on the scenes semantic identity rather than structural constraints, further increasing the variability of the information it contains.

Given the evidence suggesting our visual system has evolved to take advantage of spatial regularities, the stability of vertical scene structure may make it particularly beneficial to predictive processing strategies. Predictive processing theories suggest visual inputs are compared against internal models, based on our expectations of what the world should look like (Kaiser, Turini, et al., 2019; Keller & Mrsic-Flogel, 2018; Muckli et al., 2015). This comparison allows for information that matches the internal model to be quickly processed, so more typical information can be downweighed and the system can allocate more resources to detecting novel visual features that may be more category defining, indicative of a scenes identify or behaviourally relevant. Given the high levels of standardisation and predictability of vertical structure, this information may be strongly represented within these internal models. As such, the visual system may exhibit a "vertical bias" in which scene information along the vertical axis is more easily processed, and where likewise disruptions to vertical structure may cause greater disruptions to predictive processing strategies. Conversely less predictable horizontal structure may be less easily utilised by predictive processing models, and so the visual system may be more tolerant to disruptions along this axis.

This proposed vertical bias is further supported by evidence from neurophysiological and psychophysical research, which demonstrates how the visual system has adapted to the vertical spatial regularities present in the natural world. Neurophysiological recordings from macaque monkeys provide compelling example of this adaptation. Vaziri and Connor (2016) found that scene-selective neurons in the ventral visual pathway exhibited a preference for shape arrangements that were aligned with gravity in an egocentric reference frame. Subsequent studies in humans have likewise found preferences for object arrangements obeying gravitational limitations (Tucciarelli et al., 2023), with fMRI studies also having identified cortical regions exhibiting a selective activation for scenes where physical restrictions are adhered to (Fischer et al., 2016). Taken together, these studies suggest that the visual system is optimised for interpreting the gravitational and physical, reflecting the stable and predictable nature of vertical scene elements like the ground and sky.

Psychophysical studies exploring the role of anisotropy in natural scenes provide indirect support for a potential vertical bias mirroring the content of natural scenes. The horizontal effect, as described by Essock (et al., 2003), refers to a reduced sensitivity for detecting horizontally orientated scene content compared to vertically and oblique orientated content. They define many of these horizontally orientated features as elements typically located on a scene's vertical axis, such as horizons and tree lines. They suggest that this mechanism may have evolved to tune down the prevalence of horizontal content in scenes, thus serving to discount the perceptual salience of the horizon and other predominant horizontal content (Hansen & Essock, 2004). This allows visual information with oblique and vertical orientations to become more saliant. Such elements are often scene defining or behaviourally relevant objects (such as individual trees aligned vertically, or walls of a building), and so by highlighting these more novel scene elements, they allow the visual system to prioritize and process them more efficiently, potentially enabling more efficient scene processing. Similar to the predictive processing theories discussed previously, the horizontal effect highlights how the visual system aims to downplay the perceptual salience of predictable scene information typically existing on a scenes vertical axis. However, it is important to note that whilst some horizontally orientated scene elements correspond with features found along the vertical scene axis (such as the aforementioned horizons), due to the wide array of elements found in natural scenes, many will not. Work exploring the horizontal effect has not concentrated on the former features uniquely, and instead typically employ broad range of orientation information found in scenes. As such, whilst the horizontal effect may be indicative of the role of vertical organisation in scene processing more generally, it cannot be said to explicitly support a bias in vertical scene structure explicitly.

If the visual system prioritizes vertical information during scene processing, this bias could influence how scenes are encoded, stored, and retrieved from memory. Early work exploring scene memory provides some evidence for a potential vertical bias, revealing that individuals are more adept at recalling the vertical arrangement of elements in a scene compared to their horizontal placement. Mandler and Parker (1976) presented participants with pictures of objects arranged within a scene and later asked them to reconstruct the objects' locations. Their findings showed that correlations between the original and reconstructed locations were stronger for the vertical dimension than for the horizontal dimension, suggesting memory for vertical locations was more accurate than for horizontal locations. Likewise, Previc and Intraub (1997) found that when asked to draw scenes from memory, participants typically shifted the perspective upward along the vertical axis, demonstrating a vertical bias in boundary extension. Whilst these findings suggest a preference for scene information arranged along the vertical axis, it is important to acknowledge that memory processes do not necessarily reflect perceptual processes. Memory-based biases may emerge from post-perceptual

mechanisms, such as strategies employed during encoding or retrieval, rather than from inherent properties of visual processing. Thus, while these studies provide valuable insights, they cannot definitively establish whether vertical structure is uniquely important to scene perception itself.

More direct evidence comes from neuroimaging research by Kaiser et al (2019). Kaiser et al (2019) investigated the cortical sorting of scene structure by measuring the neural activity of participants using both fMRI and EGG, whilst they performed a scene categorisation task on individual scene fragments taken from different structural positions of a scene (e.g. scenes were divided into 6 pieces, once along the vertical and 3 times along the horizontal axis to create individual scene fragments). Using representational similarity analysis to reconstruct the cortical representations of these scene fragments, they found a scene fragment's vertical, but not horizontal, location predicted its representation in the OPA. Furthermore, they found that a fragments vertical location was not predicted neural representations in V1 and PPA, suggest that the highlights coding was not the result of analysis of simple low-level visual features. These results were supported by a subsequent experiment, where participants conducted the same scene categorisation task whilst undergoing EEG. The sorting was found to occur at the early stages of scene processing, within the first 200ms, suggesting that downstream, higher level cognitive and motor systems would have access to this structural information for use in relevant real-world tasks (such as navigation).

However, whilst Kaiser et al (2019) findings contribute valuable insights into a potential vertical bias, certain methodological limitations warrant consideration. Notably, the study utilised a limited set of only six scenes, each representing a distinct category. This narrow selection may have impacted the generalisability of the results, as a broader range of scenes might yield different outcomes. Additionally, the jumbling manipulation was applied more extensively along the vertical axis than the horizontal axis, potentially introducing an imbalance that could influence the observed vertical bias. The study also relied on neuroimaging techniques, leaving a gap in understanding as to whether these vertical biases can be detected through behavioural measures; if a vertical bias in scene processing does occur rapidly and at early stages of processing, we might also expect to observe it in experiment scene categorisation.

Given the evidence discussed, there is a prominent indication that vertical scene structure plays a crucial role in scene perception. However, many of the studies discussed focus on broader scene processing mechanisms or memory effects, without specifically isolating the role of vertical structure in perceptual processing. The present experiment aims to address this gap by directly investigating whether a vertical bias exists in scene perception, and whether disruptions to vertical structure cause greater perceptual disturbances than horizontal disruptions. By exploring this question, we can

deepen our understanding of how the visual system adapts to the predictable structure of natural scenes, ultimately informing models of scene perception and prediction.

In the current study we aimed to investigate how scene structure across different axes influences scene processing by systematically manipulating spatial arrangement using a jumbling paradigm. Specifically, we had three key objectives. First, we aimed to replicate previous findings demonstrating that scene jumbling impairs scene categorisation (Biederman et al., 1974; Kaiser et al., 2020a; 2020b), in order to examine whether this impairment occurs due to disrupted spatial organisation or a more general difficulty in processing the individual components of a scene when their typical spatial relationships are altered. Secondly, to explore whether a vertical bias exists in scene processing—where disrupting vertical structure has a greater impact on categorisation than disrupting horizontal structure (Kaiser et al., 2019). This would allow us to both understand whether vertical scene structure plays a unique role in scene processing and provide further insights into how the visual system is adapted to real-world norms (as vertical structure tends to be more stable and predictable compared to horizontal structure). Thirdly, to assess whether the effect of scene structure on categorisation differs between upright and rotated scenes (180° and 90°), as observed in Kaiser et al (2020a; 2020b), in order to deduce whether this effect is the result of disrupting the scene's structural organisation, rather than simply shuffling low- and mid-level visual features.

To address these aims, we conducted two experiments using a scene jumbling paradigm Scenes were manipulated across four conditions: (1) intact, (2) vertically intact but horizontally jumbled, (3) horizontally intact but vertically jumbled, and (4) fully jumbled. Performance on a scene categorisation task was compared across these conditions to test for evidence of a vertical bias. We utilise a scene categorisation task in order to investigate how structure influences the perception of global scene information. If a vertical bias exists, we expected categorisation performance to be more impaired in the vertically jumbled condition than in the horizontally jumbled condition, consistent with prior research (Kaiser et al., 2019a). To assess whether these effects are driven by structural information rather than low-level visual features, we tested scene categorisation at different orientations (180° in Experiment 1 and 90° in Experiment 2). Additionally, we carried out analysis comparing whether the degree of rotation effected scene categorisation, in order to explore whether our results could be explained by mental rotation.

Accompanying these aims we had 3 hypotheses. First, we hypothesised that scene categorisation accuracy would be lower in the jumbled conditions compared to intact scenes, in line with previous research suggesting that structure is important for scene processing (Biederman et al., 1974). Second, that this categorisation impairment would be less pronounced for upright scenes than for inverted

scenes. Thirdly, that categorisation accuracy would be significantly worse for vertically jumbled scenes compared to horizontally scrambled scenes, indicating a vertical bias in scene processing as suggested by Kaiser et al (2019).

By investigating these key questions, we hope to clarify the role of vertical and horizontal scene structure in visual processing and contribute to our understanding of the brain's predictive strategies when interpreting natural scenes.

2.2 Experiment 1: 180-degree rotation

2.2.1 Methods

Participants

Experiment 1 was approved by the University of York ethics committee. Participants were recruited primarily through the use of an online recruitment website (https://www.prolific.com), in addition to the University of York's recruitment pool. All participants reported having normal or corrected to normal vision. They were paid for their participation and provided informed consent before taking part in the study. In experiment 1 we recruited 47 participants (18 female, mean age 25.27 years, SD 8.14, range 18–62 years).

We aimed for a sample of between 40-50 participants to reflect the sample sizes utilised in previous studies (Biederman, 1972; Biederman et al., 1973; Kaiser et al 2019b; 2020c) whilst accounting for any potential loss of experimental power due to conducting the experiment online. Whilst similar studies typically utilised sample sizes of approximately 30 participants, as our experiment was hosted online, and in an environment outside of the experimenter's control, we utilised a slightly larger sample size. This would allow us to compensate for any experimental power that may have been lost due to participants' hardware/software choices, and external distractors that may ad random noise into the results. that can largely be mitigated by larger sample sizes.

Stimuli

Stimuli consisted of photographs of outdoor scenes from 4 categories: mountains, deserts, beaches and fields. Images were taken from the "Massive memory" scene category stimulus set on the Konkle Lab website (Konkle et al., 2012), with additional scenes being sourced through online image searches. We chose to use images of outdoor scenes as these categories would be familiar to most people, and feature relatively rigid structural norms. Outdoor scenes typically have strong global features, such as a clear horizon and large regions of uniform textures (such as large areas of grass or water), whilst indoor scenes may include more visual clutter, and objects with ambiguous spatial locations. Additionally, as the experiment was conducted online, and could thus be completed by participants

globally, we excluded images of indoor scenes as these may contain more cultural and socio-economic norms that might influence participants ability to categorise scenes more effectively, thus effecting their baseline categorisation. 25 photographs were selected for each scene category, leading to a total of 100 unique photographs. In addition to the full scenes, we also presented scenes where the structure had been jumbled in 3 ways: fully jumbled, vertically jumbled and horizontally jumbled.

In order to achieve this, each photograph was split along its horizontal and vertical axis, yielding 4 position specific fragments of equal size. These fragments were rearranged to produce the scene images jumbled along the previously mentioned dimensions. The fully jumbled scenes were achieved by rearranging the scenes in 3 ways, either switching the top left piece with the bottom right piece, switching the top right piece with the bottom left piece, or by switching both top pieces with their diagonal opposites.

Horizontally jumbled scenes were produced by either switching both of the top two fragments with the bottom two fragments (switching the top half of the image with the bottom half) or by switching the top left and right fragments. Similarly, the vertically jumbled condition was produced by switching both the top left and bottom left fragments with their right-side equivalents, or by switching the top left and top right scene fragments. Participants were not shown the full scene images prior to the experiment. All conditions were also shown at 180-degree inversion (here on referred to as inversion). This resulted in a total of 16 different versions of each scene image being used in each experiment (see Figure 2.2, Appendix A for further examples).

As the experiment was conducted by participants on their own personal computers, viewing conditions varied between participants. However, they were instructed to complete the experiment on a desktop computer, in a quiet environment away from distractions. In experiment 1, we collected information about the participants display size and instructed them to sit 60cm away from the screen, so that we could work out the display size of the stimulus in visual degrees. The quality of information provided on display size varied, and the answers of 3 participants were removed due to providing spurious estimates. The average display size used in experiment 1 was 28.6cm by 41 cm (height range 37.8-5.57, width range 45.3-2.61). Stimuli were displayed at 480 by 480 pixels, however due to differing display sizes the actual size of the stimulus the participants viewed varied.

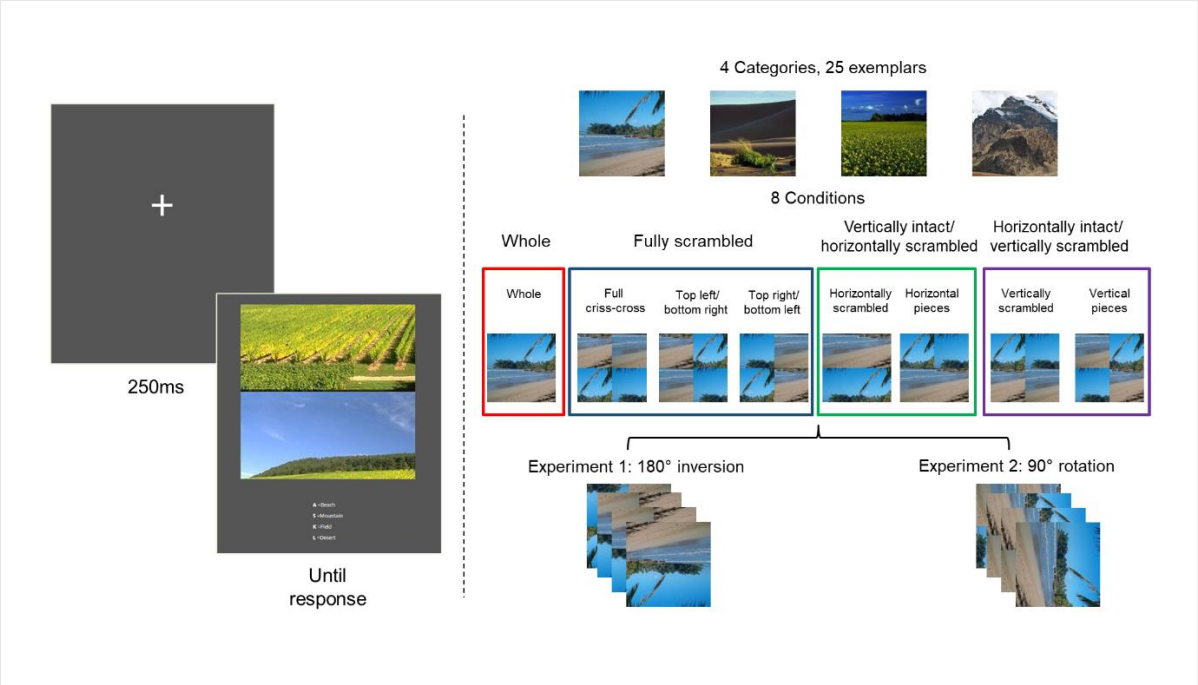


Figure 2.2. Design and Stimulus examples. Example of the trial structure. Participants first saw a white fixation cross for 250ms followed by the stimulus. The stimulus was displayed until participants gave a response. Examples of the conditions and manipulations to the scenes carried out in experiment 1 and 2. The experiment had four conditions, intact, fully jumbled, horizontally intact and vertically jumbled and vertically jumbled and horizontally intact. The intact condition consisted of the full un-manipulated scene image. The fully jumbled condition consisted of scenes which were manipulated in 3 ways, either switching the top left piece with the bottom right piece, switching the top right piece with the bottom left piece, or by switching both top pieces with their diagonal opposites. The vertically intact and horizontally jumbled condition consisted of scenes that either switched both of the top two fragments with the bottom two fragments (switching the top half of the image with the bottom half) or by the top left and right fragments. Similarly, the horizontally intact and vertically jumbled condition was produced by switching both the top left and bottom left fragments with their right-side equivalents, or by switching the top left and top right scene fragments.

Experimental Paradigm

The experiment was conducted online and was built and hosted on the online experiment building platform Gorilla (Anwyl-Irvine et al., 2020).

Participants were tasked with classifying scenes into four categories (fields, mountains, deserts and beaches) by pressing a corresponding key on their keyboard. They were instructed to respond as

quickly as possible. Each trial began with a fixation cross displayed for 250ms, followed by the stimulus scene image, which was displayed until the participant provided an answer.

Participants first completed a practise block of 64 trials, which used a unique set of 8 scene photographs (2 for each scene category) that had been manipulated in the same way as the images used in the rest of the experiment. Participants then completed 4 experimental blocks. Each block comprised of 280 trials, with the displayed stimulus drawn from a random selection of the stimulus images. Each stimulus was shown only once during the experiment, in addition to being shown at a 180-degree rotation. Key prompts were displayed underneath the stimulus, so that participants had a constant reminder of the corresponding keys for each scene (in order to minimise errors).

Data Cleaning

Trials in which incorrect answers were given, or reaction times were faster than 200ms or slower than 5000ms were excluded from the analysis. Trials faster than 200ms were excluded to best ensure that responses were to the visual stimuli, instead of anticipatory responses or instances where participants might attempt to complete the experiment as quickly as possible (this was especially important as the experiment was conducted online, without experimenter supervision). Trails slower than 5000ms were also excluded for similar reasons: as the experiment was conducted online and without experimenter supervision, we aimed to exclude trials where distractions or technical issues may have caused non-task related delays to answers. We chose 5000ms as a maximum cut-off point as the experimenters felt this gave ample time to complete each trial, whilst providing a conservative cut off point for excluding any trials where participants may have been influenced by external distractions.

Of the 52684 trials collected, 1887 were removed for being faster than 200ms, and 1084 were removed for being over 5000ms. Of the remaining 49713, 3180 were incorrect and subsequently removed. The remaining 46533 trials were used in the analysis. Approximately equal numbers of each condition were included in these trials, with the largest discrepancy being between upright intact and the inverted vertically jumbled condition, with a difference of 134 trials. Likewise, all scene categories were approximately equally represented, with the largest discrepancy being between mountains and beaches, with a difference of 205 trials, See Appendix G for tables of trials included in the final analysis by condition and scene category.

2.2.2 Results

The focus of our analysis was to investigate the effect of jumbling scene structure on scene categorisation, as well as exploring the role of inversion. Further, we also wanted to compare the difference between horizontal and vertical jumbling, in order to understand whether a bias exists for either axes. ,

A two-way repeated measures ANOVA was run to determine the effect of jumbling over orientation on reaction times in the scene categorisation task. This ANOVA had two factors, orientation with 2 levels (upright, inverted) and jumbling with 4 levels (intact, horizontally jumbled, vertically jumbled and fully jumbled.) Mauchly's test of sphericity indicated that the assumption of sphericity was met for the two-way interaction between orientation and jumbling, χ^2 (5) = 5.45, p = .360.

There was a statistically significant two-way interaction between orientation and jumbling on reaction times, F (3, 138) = 2.83, p = .041, ηp^2 = .06 (indicating a medium effect size). This suggests that the impact of jumbling is dependent on the orientation of the scene. Mean reaction times are shown in figure 2.3. To explore the simple main effects, we conducted 2 separate one-way repeated measures ANOVA to examine jumbling within each orientation, and 4 repeated measures t-tests to examine the effect of jumbling between orientation (as described below).

First, we examined the effect of jumbling within each orientation. To achieve this, we conducted two one-way repeated measures ANOVA examining the effect of jumbling within both the upright and inverted scenes. Both ANOVA used a single factor, jumbling, that consisted of all of the jumbling conditions from the upright and inverted scenes. This meant that the levels for the upright ANOVA were upright intact, upright horizontally jumbled, upright horizontally jumbled and upright fully jumbled, whilst the levels for the inverted ANOVA were inverted intact, inverted horizontally jumbled, inverted vertically jumbled and inverted fully jumbled. To account for multiple comparisons, alpha values were adjusted using a Bonferroni correction, resulting in an adjusted significance threshold of $p < .025$ (.05/2).

The one-way repeated measures ANOVA examining the effect of jumbling within the upright orientation violated the assumption of sphericity, as indicated by Mauchley's test, $\chi^2(5) = 13.44$, $p = .02$. Therefore, degrees of freedom were corrected using the Greenhouse-Geisser estimate ($\varepsilon = 8.27$). The analysis revealed a significant simple main effect of jumbling in the upright scenes, $F(2.48, 114.08) = 12.99$, $p < .001$, $\eta^2_p = .22$ (indicating a large effect size).

Bonferroni-adjusted pairwise comparisons (with an adjusted significance threshold of $p < 0.008$ (0.05/6 comparisons) found that reaction times for intact scenes were significantly lower than for

vertically scrambled scenes (*M* = -55.17, *SE* = 14.54, p < .001, 95% CI [-95.25, -15.10]) with Cohen's d of d = .54 indicating a medium effect size, and fully jumbled scenes (M = -76.63, SE = 14.30, p <.001, 95% CI [-116.05, -37.21]) with Cohen's d of d = 0.78 indicating a medium effect size, but not for horizontally scrambled scenes (*M* = -35.56, *SE* = 10.76, *p* = .011, 95% CI [-65.22, -5.90]), which had Cohen's d = .48. This suggests that whilst vertically and fully jumbled scenes were more difficult to categorise, there was no effect of horizontal scrambling. This contradicts our first hypothesis that jumbled scenes would be more difficult to categorise as a whole, and instead suggests that only when vertical structure is absent (as in the vertically and fully jumbled conditions) does a disruption occur.

However, contrary to our third hypothesis, we found no significant difference between horizontally and vertically scrambled scenes (*M* = -19.61, *SE* = 12.35, *p* = .715, 95% CI [-53.66, 14.43]) with a small effect size of Cohen's d = 0.23, which suggested that there was no preference for intact vertical scene structure over horizontal structure. We also found that reaction times for the fully jumbled condition were not significantly higher compared to the horizontally scrambled (*M* = 41.07, *SE* = 13.46, *p* = .023, 95% CI [-53.66, 14.43]) with a small effect size of Cohen's d = .44, or vertically scrambled scenes (*M* = 21.4, *SE* = 10.61, *p* = .295, 95% CI [-7.81, 50.71]) with a small effect size of Cohen's d = .29. This may indicate that whilst fully scrambling scene structure causes a greater disruption to the scene categorisation, it may be comparable to the disruption caused by removing intact vertical scene structure alone.

Next, we used a one-way repeated measures ANOVA to examine the effect of jumbling within the inverted orientations. Again, the assumption of sphericity was violated, as indicated by Mauchley's test, $\chi^2(5) = 13.83$, $p = .017$, and degrees of freedom were corrected using the Greenhouse-Geisser estimate ($\varepsilon = 8.34$). As in the upright scenes, we found a significant effect of jumbling in the inverted scenes, $F(2.50, 115.08) = 16.73$, $p < .001$, $\eta^2_p = .27$.

Bonferroni-adjusted pairwise comparisons (with an adjusted significance threshold of $p < .0083$ (.05/6 comparisons) found that as in the upright scenes, reaction times for inverted intact scenes were significantly faster than those for vertically scrambled (*M* = -81.92, *SE* = 15.21, *p* < .001, 95% CI [-123.85, -39.19]) with a strong effect size of Cohen's d = 0.78, and fully jumbled scenes (*M* = -60.82, *SE* = 13.01, *p* <.001, 95% CI [-96.69, -24.95]) with a strong effect size of Cohen's d = 0.89, and again that there was no significant difference between reaction times for the intact scenes and the horizontally scrambled scenes (*M* = -7.92, *SE* = 11.42, *p* = 1.000, 95% CI [-39.40, 23.57]) with Cohen's d = .1. This may indicate that the effects of vertical and fully jumbling observed in the upright conditions could be influenced by disruptions to the scenes low-level visual characteristics, and not uniquely to the scene's structural information, whilst inversion does little to modulate the effect of horizontal jumbling.

Unlike in the upright scenes we found that reaction times were significantly higher for the inverted fully jumbled scenes compared to the horizontally jumbled scenes (*M* = -52.91, *SE* = 10.55, *p* < .001, 95% CI [-81.91, -23.82]) with a medium effect size of Cohen's d = .73, whilst reaction times were not significantly different compared to the inverted vertically jumbled condition (*M* = 21.101, *SE* = 15.60, *p* = .116, 95% CI [-21.88, 64.09]) with Cohen's d = .19. However, unlike in the upright condition we also found that reaction times were significantly higher for the inverted vertically jumbled scenes compared to the inverted horizontally jumbled scenes (*M* = 74.01, *SE* = 16.17, *p* < .001, 95% CI [-29.4, 118.61,]) with a medium effect size of Cohen's d = .67. As both the vertically and fully jumbled conditions lack intact vertical structure, it could suggest that it is specifically manipulations to the scene content along this axis that results in the impact of inversion in jumbled scenes. As we only observed the difference between vertical and horizontal jumbling in the inverted scenes, it may suggest that it is specifically manipulations to the low-level visual properties along this axis that causes the disruption.

To explore the interaction between jumbling and orientation further, we conducted four repeated measures t-tests examining the effect of orientation within each jumbling condition. For each, we compared the jumbling type (either intact, horizontally jumbled, vertically jumbled or fully) for the upright and inverted scenes. To account for multiple comparisons, alpha values were adjusted using a Bonferroni correction, resulting in an adjusted significance threshold of p < .012 (.05/4).

For intact scenes, we found reaction times were significantly higher when inverted compared to upright (t(46) = 3.86, *p* < .001), with Cohen's d = .56 indicating a medium effect size. This result is unsurprising considering many previous studies have demonstrated that inversion causes disruptions to categorisation (Kelley et al., 2003; Lauer et al., 2020). However, into itself this finding tells us little about what factors cause this disruption.

For horizontally jumbled scenes we found no significant difference when scenes were displayed inverted or upright (t(46) = 1.49, *p* = .143), with Cohen's d = .21, suggesting that jumbling horizontal structure had a similar effect in both upright and inverted scenes. As we found that horizontal jumbling did not cause significantly higher reaction times compared to intact scenes in both upright and inverted conditions, this could suggest that neither manipulations to horizontal scene structure or low-level visual characteristics along the horizontal axis disrupt scene categorisation.

Conversely, for vertically jumbled scenes we found reaction times were significantly higher for inverted compared to the upright scenes (t(46) = 4.83, *p* < .001), with a medium effect size of Cohen's d = .71, indicating that inversion modulated the effect of vertical jumbling. This could suggest that the

disruption caused by jumbling the vertical scenes is a result of disrupting low-level visual characteristics along the scenes vertical axis, as opposed to the structure itself.

When exploring the fully jumbled condition we found no significant difference between the upright and inverted conditions, ($t(46) = 2.05$, $p = .23$), with a small effect size of Cohen's $d = .29$, indicating that inversion had no effect when all intact spatial structure was removed from a scene. We suggest two possible interpretations for this finding; either that the increase in reactions times for the fully jumbled condition is primarily the result of manipulations to the scenes structure, as opposed to disruptions in low level visual characteristics, or that the jumbling present in the upright scenes disrupts the low-level visual characteristics of the scene to such an extent that inversion no longer impacts categorisation. As we observed that reactions times were significantly greater for jumbled compared to intact conditions in both upright and inverted conditions (where the inverted scenes similarly had their low level visual properties disrupted), this could suggest that our results support the former explanation, aligning with previous work showing the impact of jumbling scene content (Biederman et al., 1973; Kaiser, Turini, et al., 2019).

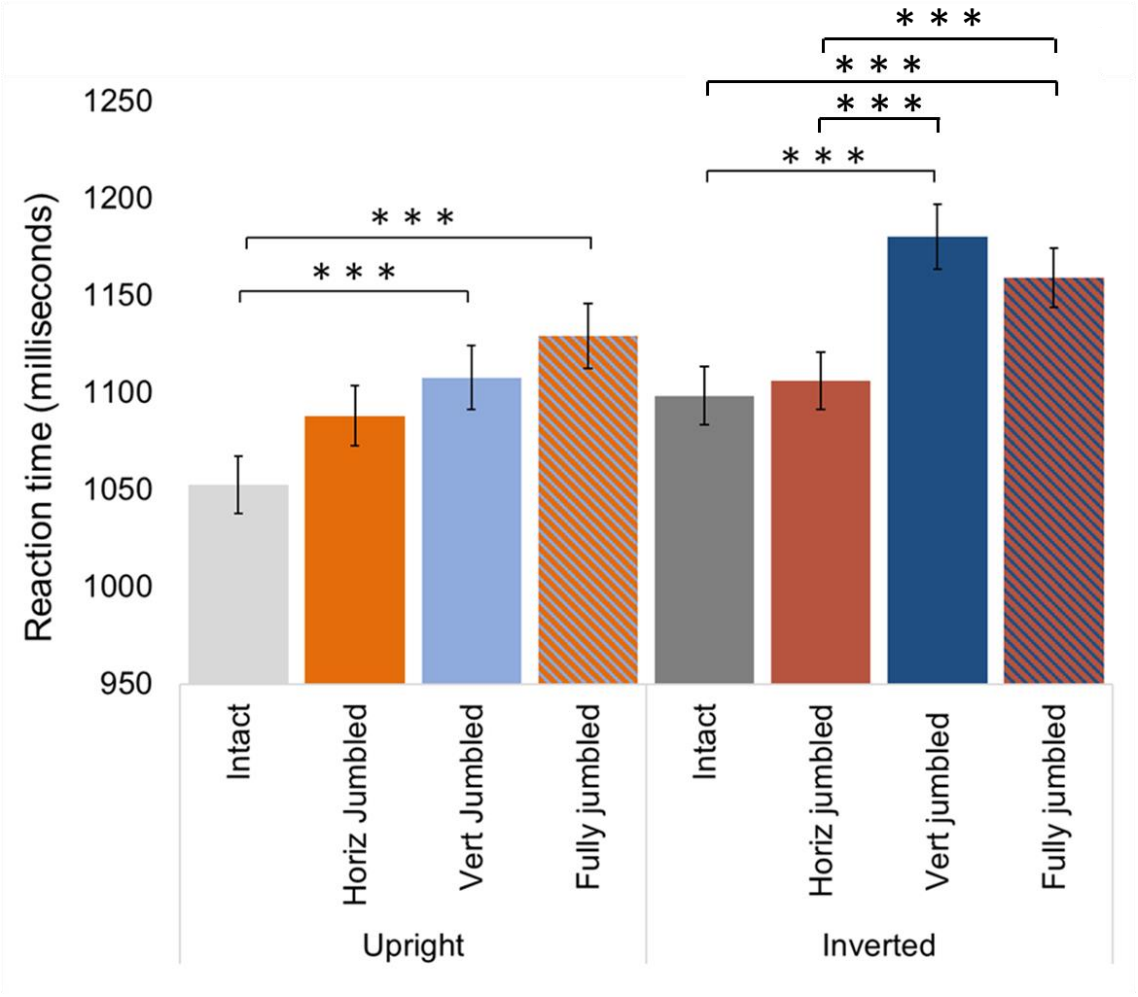


Figure 2.3. Mean reaction times (in milliseconds) across all conditions in experiment 1. Error bars represent standard errors of the mean. * Indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$.

2.2.3 Summary

In experiment 1, we found that only fully and vertically jumbled scenes significantly impacted scene categorisation for both upright and inverted scenes. This suggests that jumbling scene content only had a negative effect on scene categorisation, when vertical scene structure was absent. However, whilst we observed an effect of inversion, unlike previous research (Kaiser et al., 2020a, 2020b) we did not find that jumbling had a stronger effect on upright compared to inverted scenes, contrasting with our second hypothesis. Instead, we found inversion produced higher reaction times for intact and vertically jumbled scenes, whilst having no effect on the horizontally and fully jumbled scenes. This effect was particularly pronounced in the vertically jumbled condition, where reaction times were higher for the inverted than for upright vertically scrambled condition and comparable to the inverted fully jumbled condition.

In both upright and inverted conditions, we found that reaction times for the fully jumbled scenes were not statistically different to those of the vertically jumbled scenes. This may indicate that disrupting vertical scene structure impacts scene processing as much as removing coherent structure entirely, which may suggest that intact visual information arranged along a scene's vertical axes is uniquely important for scene processing. However, only in the inverted conditions did we find a difference between horizontal and vertical scrambling, which could indicate that it is manipulations to low-level visual characteristics arranged along this axis, as opposed to the structural information, that aids scene categorisation.

As such, whilst the results of experiment 1 do not provide direct evidence for a vertical bias resulting from scene structure, it provides tentative evidence for a bias resulting from low-level visual characteristics arranged along the vertical axis.

This interpretation relies on the assumption that inversion effects all jumbling conditions equally: that it removes meaningful coherent structural information. However, when inverting jumbled scenes, the inversion may have complex effects on the relative and absolute positions of the segments. For example, when an intact scene is inverted, it retains the relative positioning of its segments, but when a vertically jumbled scene is inverted, segments lose their relative positioning but may in return gain an intact absolute positioning (e.g., a piece of sky would be in the upper part of an inverted jumbled scene, which is where it belongs). This effect may have limited the inversion effect evoked in experiment 1, by retaining some intact vertical scene structure within the vertically jumbled condition, whilst the horizontally jumbled condition represented a true disruption to horizontal scene information. We initially explored a full 180-degree inversion in order to try to replicate the classical inversion effect witnessed in previous studies, but subsequently may have failed to adequately account for the difference in jumbling.

Whilst we did not find a statistically significant difference between vertical and horizontal jumbling in the upright condition, as we had predicted in our third hypothesis, there was some tentative emerging evidence for more severe disruptions being caused by vertical jumbling. Non-significant, reaction times for vertically jumbled scenes were nominally higher than the horizontally jumbled scenes, indicating a possible trend in the data. However, it is important to note that this nominal difference is only small (with a mean difference of only 19.6ms between upright vertically and horizontally scrambled scenes), and where significant differences were found between jumbling conditions the effect sizes were only weak. Further, whilst we found upright and inverted vertically jumbled scenes had significantly lower reaction times compared to the intact scenes, we found no effect of horizontally jumbling, , which as discussed may indicate some additional importance of vertical

structure. In conjunction with the potential differential effects of inversion on horizontal and vertical scenes, these contradictions make it difficult to draw conclusive interpretations from the current experiment, and call for further clarification.

As such, in order to clarify the results of experiment 1, we conducted a second experiment in which scenes were rotated at 90-degrees instead of 180-degrees. By displaying scenes at 90-degree rotation, it would ensure that both horizontally and vertically jumbled scenes would be positioned outside both their relative and absolute positions, ensuring that the effect of rotation would manipulate the structural content of each scene equally.

2.3 Experiment2: 90-degree rotation

2.3.1 Methods

Participants

Experiment 2 was approved by the University of York ethics committee. Participants were again recruited primary through the use of an online recruitment website (https://www.prolific.com), in addition to the University of York's recruitment pool. All participants reported having normal or corrected to normal vision. They were paid for their participation and provided informed consent before taking part in the study.

Due to an error in our online script, where the software used to host the experiment recruited more participants than we had originally intended, we collected responses from 99 participants. Although we had originally intended to collect responses from 50 participants (in line with the strategy used in experiment 1), we decided to analyse the full sample that we had collected. We had two main reasons for this decision; Firstly, as experiment 1 was conducted online and not in experimental conditions, the variability in the data was inevitably higher due to uncontrolled environmental factors. This may have resulted in the weaker effect sizes found in experiment 1, and a lack of power resulting in potential type 2 errors. This could have been particularly detrimental to our experimental design, as the small differences between reaction times may have been lost due to distractions or differences in device settings, such as internet lag, processing power or screen size. As many of the significant differences observed in experiment 1 were only very small in nominal terms (such as the mean difference between the intact and fully jumbled conditions only being 76.25ms), we felt these effects may have been undetectable without sufficient experimental power under the current experimental conditions. In experiment 1, we had attempted to address this weakness by increasing the sample size from those typically used in lab settings, but this increase may not have been enough to offset these differences. Secondly, we felt that arbitrarily removing participants post-hoc may raise concerns about

selection bias, as we had no pre-registered criteria on how to remove participants (Open Science Collaboration, 2015; Simmons et al., 2011). It is acknowledged that the ideal solution to this issue would have been to base our initial sample size on a power analysis, as opposed to the sample size of previous studies, and establish pre-registered exclusion criteria, as suggested by Nosek (et al., 2018). However, as we were not able to take these precautions after the fact, we have taken the current approach to both mitigate further sampling issues caused by attempting to rectify this mistake and based on the principle of experimental transparency. Ultimately, whilst we decided to include these participants in our final analysis, it is important to interpret the results in consideration of this sampling error, and the subsequent increase in sample size.

Thus in experiment 2 our sample consisted of 99 participants (44 female, mean age 26.34 years, SD 9.9, range 18–60 years).

Stimuli

The same scene images were used as in experiment 1, and these were manipulated in the same ways to create the 4 conditions (intact, fully jumbled, vertically intact and horizontally jumbled and horizontally intact and vertically jumbled). However, in experiment 2 scenes were also shown rotated at 90-degrees clockwise, as opposed to a full 180-degree inversion. See Appendix B for stimuli examples.

Participants were given the same viewing instructions as in experiment 1.

Experimental Paradigm

Experiment 2 used the same paradigm as experiment 1. The experiment was conducted online and was built and hosted on the online experiment building platform Gorilla (Anwyl-Irvine et al., 2020).

Participants were tasked with classifying scenes into four categories (fields, mountains, deserts and beaches) by pressing a corresponding key on their keyboard. Each trial began with a fixation cross displayed for 250ms, followed by the stimulus scene image, which was displayed until the participant provided an answer. Participants first completed a practise block of 64 trials, which used a unique set of 8 scene photographs (2 for each scene category) that had been manipulated in the same way as the images used in the rest of the experiment. Participants then completed 4 experimental blocks. Each block comprised of 280 trials, with the displayed stimulus drawn from a random selection of the stimulus images. Each stimulus was shown only once during the experiment, in addition to being shown at a 90-degree rotation. Key prompts were displayed underneath the stimulus, so that participants had a constant reminder of the corresponding keys for each scene (in order to minimise errors).

Data Cleaning

As in experiment 1, trials in which incorrect answers were given, or reaction times were faster than 200ms or slower than 5000ms were excluded from the analysis. Of the 110223 trials collected, 4358 were removed for being faster than 200ms, and 1518 were removed for being over 5000ms. Of the remaining 104345, 7027 were incorrect and subsequently removed. The remaining 97318 trials were used in the analysis.

Approximately equal numbers of each condition were included in the remaining trials, with the largest discrepancy being between upright whole and the rotated whole condition, with a difference of 392 trials. Likewise, all scene categories were approximately equally represented, with the largest discrepancy being between field and desert, with a difference of 2737 trials. See Appendix H for tables of trials included in the final analysis by condition and scene category.

2.3.2 Results

As in experiment 1, the focus of our analysis was to investigate the effect of jumbling scene structure on scene categorisation, and to explore the role of rotation. We again wanted to compare the difference between horizontal and vertical jumbling.

As in experiment 1, a two-way repeated measures ANOVA was used to determine the effect of jumbling over orientation on reaction times in the scene categorisation task. This ANOVA had two factors, orientation with 2 levels (upright, rotated) and jumbling with 4 levels (intact, horizontally jumbled, vertically jumbled and fully jumbled.) Mauchly's test of sphericity indicated that the assumption of sphericity had been violated for the two-way interaction between orientation and jumbling, $\chi^2(2) = 8.54$, $p = .128$, so degrees of freedom were corrected using the Greenhouse-Geisser estimate ($\varepsilon = .96$). .

As in experiment 1, there was a statistically significant two-way interaction between orientation and jumbling on reaction times, $F(2.88, 282.37) = 2.69$, $p = .048$, $\eta^2_p = 0.02$. This suggests that the impact of jumbling is dependent on the orientation of the scene. Mean reaction times are shown in figure 2.4. To explore the simple main effects, we conducted 2 separate one-way repeated measures ANOVA to examine jumbling within each orientation, and 4 repeated measures t-tests to examine the effect of jumbling between orientation (as described below). following the same analysis as conducted in experiment 1.

To examine the effect of jumbling within each orientation we conducted two one-way repeated measures ANOVA, with jumbling as the factor, consisting of all of the jumbling conditions (intact, horizontally jumbled, vertically jumbled and fully jumbled) for both the upright and inverted scenes.

To account for multiple comparisons, alpha values were adjusted using a Bonferroni correction, resulting in an adjusted significance threshold of $p < .025$ (.05/2).

The one-way repeated measures ANOVA examining the effect of jumbling within the upright orientation violated the assumption of sphericity, as indicated by Mauchley's test, $\chi^2(5) = 12.12$, $p = .033$. Therefore, degrees of freedom were corrected using the Greenhouse-Geisser estimate ($\varepsilon = 0.93$). The analysis revealed a significant simple main effect of jumbling in the upright scenes, $F(2.79, 272.31) = 43.21$, $p < .001$, $\eta^2_p = .31$.

Bonferroni-adjusted pairwise comparisons (with an adjusted significance threshold of $p < .0083$ (.05/6 comparisons) found that reaction times for intact scenes were significantly lower than for all other jumbling conditions: horizontally scrambled ($M = -36.04$, $SE = 9.19$, $p < .001$, 95% CI [-60.54, -11.54]) with a small effect size of Cohen's d = .39, vertically scrambled ($M = -84.68$, $SE = 9.58$, $p < .001$, 95% CI [-110.51, -58.91]) with a large effect size of Cohen's d = .88, and fully jumbled scenes (M = -98.66, SE = 10.57, $p < .001$, 95% CI [-127.12, -70.21]) with a large effect size of Cohen's d = .93. These results contrast with those of experiment 1, and indicate that all types of scene jumbling made scene's more difficult to categorise, supporting our first hypothesis.

However, unlike in experiment 1, we also found that reaction times were significantly lower for horizontally scrambled scenes compared to vertically scrambled scenes ($M = -19.61$, $SE = 12.35$, $p = 1.000$, 95% CI [-53.66, 14.43]) with a Cohen's d = 58 indicating a moderate effect size, supporting our third hypothesis that there is a preferential effect for intact vertical scene structure. However, it is important to consider the sampling errors made during experiment 2, and the resulting increased sample size utilised. Whilst this may indicate that experiment 1 lacked the experimental power to find this effect, especially considering that the study was conducted online, it is also possible that this result represents a type 1 error resulting from the sampling error. This will be discussed further in the discussion section.

Further, we also found that reaction times for the fully jumbled condition were significantly higher compared to the horizontally scrambled condition ($M = 62.62$, $SE = 9.53$, $p < .001$, 95% CI [36.94, 88.29]) with a Cohen's d = .66 indicating a moderate effect size, but unlike in experiment 1 they were not significantly so when compared to the vertically scrambled condition ($M = 13.97$, $SE = 11.15$, $p = 1.000$, 95% CI [-16.06, 44.01]), with a Cohen's d = .12. This may indicate that fully jumbling scene structure causes a comparable degree of disruption as jumbling vertical scene structure alone.

Next, we examined the effect of jumbling within the rotated orientation, again using a one-way repeated measures ANOVA. Mauchly's test of sphericity indicated that the assumption of sphericity

was not violated ($\chi^2(5) = 7.87$, $p = 0.163$), allowing for the interpretation of the standard repeated measures ANOVA results. As in the upright scenes, we found a significant effect of jumbling in the inverted scenes, $F(3, 294) = 68.56$, $p < .001$, with a medium effect size of $\eta^2_p = 0.41$.

Bonferroni-adjusted pairwise comparisons (with an adjusted significance threshold of $p < .008$ (.05/6 comparisons) found that as in the upright scenes, reaction times for rotated intact scenes were significantly lower than those for vertically scrambled ($M = -91.69$, $SE = 7.52$, $p < .001$, 95% CI [-111.96, -71.42]) with a Cohen's d= 1.24 indicating a large effect size, and fully jumbled scenes ($M = -76.31$, $SE = 8.25$, $p < .001$, 95% CI [-98.51, -54.11]) with a Cohens d = 93, but that there was no significant difference between reaction times for the horizontally scrambled scenes ($M = -21.91$, $SE = 7.48$, $p = .022$, 95% CI [-41.77, -2.05]) with a Cohen's d = .29. This suggests the effects of vertical, horizontal and fully jumbling observed in the upright conditions could be influenced by disruptions to the scenes low-level visual characteristics.

Following the pattern observed in the upright scenes we found that reaction times were significantly higher for the rotated fully jumbled scenes compared to the horizontally jumbled scenes ($M = 54.39$, $SE = 6.45$, $p < .001$, 95% CI [36.92, 71.86]) with a strong effect size indicated by Cohen's d = .83, but not when compared to the rotated vertically jumbled condition ($M = -15.38$, $SE = 7.67$, $p = .286$, 95% CI [-36.03, 5.27]) with a Cohen's d = .21. We also observed that reaction times were significantly higher for the rotated vertically jumbled scenes compared to the rotated horizontally jumbled scenes ($M = 69.78$, $SE = 7.23$, $p < .001$, 95% CI [50.31, 89.25]) with a strong effect size indicated by Cohen's d = .97, similar to the effect of inversion observed in experiment 1.

Next, we conducted four repeated measures t-tests examining the effect of orientation within each jumbling condition. For each, the jumbling type (either intact, horizontally jumbled, vertically jumbled or fully) was compared between the upright and rotated version of each scene. To account for multiple comparisons, alpha values were adjusted using a Bonferroni correction, resulting in an adjusted significance threshold of $p < .012$ (.05/4).

Reaction times for the rotated intact scene were significantly higher for those of the upright intact scene ($t(98) = 4.14$, $p < .001$) with Cohen's d = .41 indicating a small effect size, suggesting that rotation negatively impacts scene categorisation. This result was similar to that observed of inversion in experiment 1.

We found no significant difference between the upright and rotated horizontal conditions, ($t(98) = 1.84$, $p = .077$) with Cohen's d = .18, suggesting that jumbling horizontal structure had a similar effect in both upright and rotated scenes. These results suggest that manipulations to horizontal scene

structure, as opposed to jumbling low-level characteristics along the horizontal axis, result in the higher reaction times for the horizontally jumbled conditions compared to the intact condition.

As in experiment 1, we found reaction times were higher for inverted vertically jumbled scenes compared to upright, ($t(98) = 3.53$, $p < .001$) with Cohen's $d = .35$, indicating that rotation modulated the effect of vertical jumbling. Again, this finding supports those observed in experiment, further suggesting that the disruption caused by jumbling vertical scenes is a result of disrupting low-level visual characteristics along the scenes vertical axis, as opposed to the structure itself.

We found no significant difference between the upright and rotated fully jumbled conditions, ($t(98) = 6.62$, $p = .267$) with Cohen's $d = .12$, suggesting that as observed with inversion in experiment 1, rotation had no effect on reaction times when all intact structure was removed.

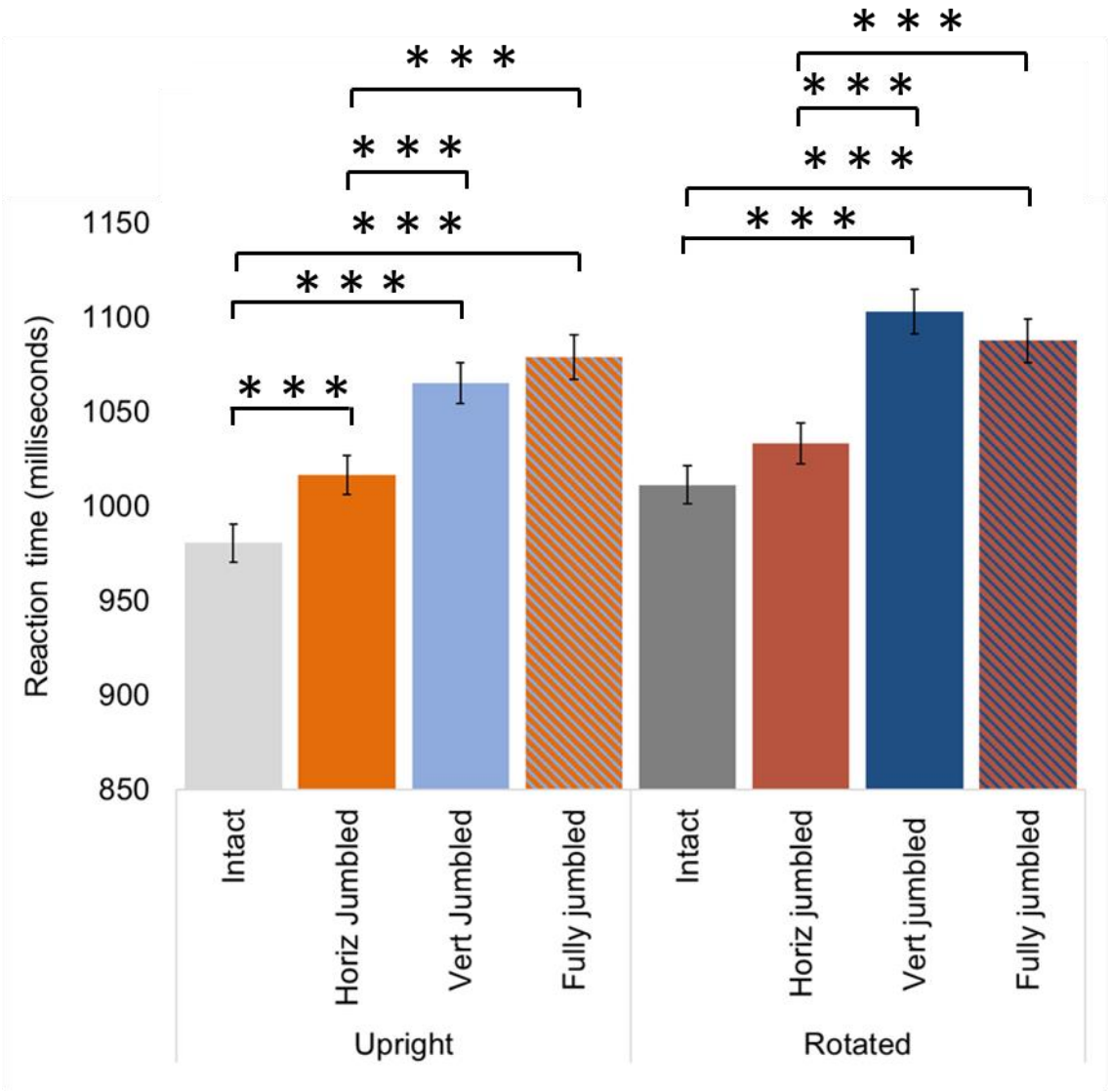


Figure 2.4. Mean reaction times (in milliseconds) across all condition in experiment 2. Error bars represent standard errors of the mean. * Indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$.

2.3.3 Summary

The results of experiment 2 suggest that reaction times were greater for all upright jumbling conditions compared to intact scenes, suggesting that disrupting scene structure reduces categorisation efficiency. This finding varied from those observed in experiment 1, that found that only vertical and fully jumbling scene content resulted in increased reaction times. As such, the results of experiment 1 support hypothesis 1; that jumbling would negatively affect scene categorisation.

Unlike in experiment 1 we also found that reaction times for the fully jumbled scenes were significantly higher for the horizontal condition, but not for the vertically jumbled scenes. This difference may be the result of the larger sample size used increasing the ability to detect this difference between the horizontally and fully jumbled conditions. Additionally, we found that reaction times were also significantly higher for upright vertically jumbled scenes compared to horizontally jumbled scenes. Taken together, these results could indicate a possible bias for vertical scene structure, in support of our third hypothesis.

However, we found that rotation had a similar effect as inversion did in experiment 1, and found higher reaction times for rotated intact and vertically jumbled scenes compared to their upright equivalents. Rotation likewise had a particularly strong effect on vertically jumbled scenes, and caused reaction times to be comparable to those of the fully jumble condition (with no statistically significant difference between the conditions). As such, whilst we did observe that disruptions caused a greater disruption to scene categorisation than horizontal structure, the prevalence of this effect in rotated scenes suggests this bias may result from a sensitivity to mid or low-level visual characteristics arranged along the vertical axis, as opposed to structural information specifically.

The effect of rotation found in experiment 2 also suggests that the inversion effect evoked in experiment 1 was not caused by the retention of intact vertical scene structure within the vertically jumbled condition, and instead that it represented a true disruption to the scenes vertical and horizontal scene structure.

2.4 Comparison between inversion (180°) and rotation (90°)

2.4.1 Rationale

Across both Experiments 1 and 2, we did not find the expected effect of inversion. That is, instead of finding a greater impact of scene jumbling in upright scenes, we found that inversion had no effect

on scene categorisation for horizontally and fully jumbled scenes, and caused even greater disruptions for vertical jumbled scenes. One possible explanation is that participants attempted to mentally rotate the scenes to their canonical upright orientation. This strategy may have been more or less effective depending on the type of jumbling: for horizontally jumbled scenes, mental rotation may have been highly successful, whereas for vertically jumbled scenes, it may have been less effective, amplifying the effects of jumbling. In fully jumbled scenes, the absence of a coherent structure may have discouraged participants from using mental rotation, resulting in minimal impact of this strategy.

To further investigate this explanation, we combined data from experiments 1 and 2 to examine whether the degree of rotation (180 degrees in Experiment 1 vs. 90 degrees in Experiment 2) affected reaction times. If mental rotation underlies our findings, we expect a greater impact of jumbling on scenes rotated 180 degrees compared to 90 degrees, as the larger degree of rotation would require more processing effort (Dalecki et al., 2012; Shepard & Metzler, 1971).

2.4.2 Results

In order to compare the difference between inversion (180°, as collected in experiment 1) and rotation (90°, as collected in experiment 2), we combined data from experiment 1 and 2 and conducted a two-way mixed ANOVA. This ANOVA had 2 factors, 1 between and 1 within. The between factor was orientation with 2 levels (inversion and rotation), whilst the within factor was jumbling with 4 levels (intact, horizontally jumbled, vertically jumbled and fully jumbled.) We did not include the upright orientation in this analysis because it was present in both Experiment 1 and Experiment 2 as a within-subjects condition. Since the 90° and 180° orientations were unique to their respective experiments (with different participants), rotation could only be analysed as a between-subjects factor. Including upright would violate the independence assumption of a between-subjects ANOVA, as all participants completed this condition, making it non-independent across experiments.

Mauchly's test of sphericity indicated that the assumption of sphericity was not violated for the two-way interaction ($\chi^2(5) = 7.87$, $p = .168$). Here we found a statistically significant interaction between orientation and jumbling on reaction times, $F(3, 432) = 33.71$, $p < .001$, $\eta^2 = 0.19$. To explore this further, we conducted four independent samples t-tests to compare the effect of orientation (180° inversion or 90° rotation) between each jumbling condition (intact, vertically jumbled, horizontally jumbled and fully jumbled). In order to account for multiple comparisons, we applied Bonferroni corrections to alpha levels, resulting in an adjusted significance threshold of $p < .012$ (.05/4).

For the vertically jumbled conditions, reaction times were significantly higher in the 180° inverted compared to the 90° rotated condition, ($t(144) = 2.39$, $p < .018$) with a moderate effect size as indicated by a Cohen's d = .66. This suggests that the degree of rotation modulated the effect of vertical scene jumbling.

However, there were no other statistically significant differences between orientation and jumbling condition (intact: $t(144) = 2.39$, $p < .001$ with Cohen's d = .42, horizontally jumbled: $t(144) = 0.76$, $p < .936$ with Cohen's d = .03, and fully jumbled: $t(144) = 1.87$f, $p < .072$ with Cohen's d = .32), indicating the degree of rotation had no effect on reaction times for these conditions.

2.4.3 Summary

From our analysis comparing the effects of inversion and rotation we found that the degree of rotation modulated the effect of vertical scene jumbling, but that it had no effect on intact, horizontally and fully jumbled scenes. As reaction times were higher when scenes were inverted, than when they were rotated, this could suggest that the more a vertically jumbled scene was rotated the more difficult categorisation becomes. This could indicate that participants attempted to utilise mental rotation when trying to categorise vertically jumbled scenes, but this strategy negatively impacted their performance. These findings may support our proposal that the increased effect of inversion and rotation for vertically jumbled scenes observed in experiment 1 and 2 are the result of participants adopting a maladaptive strategy of mental rotation when trying to categorise these scenes.

However, we only observed a difference between degrees of rotation in the vertically jumbled scenes, suggesting that this was the only condition in which participants attempted to utilise this strategy. This interpretation is challenging, as it is not clear why participants would utilise differing strategies for categorising stimuli when the task demand remains constant. A possible explanation is that the intact horizontal structure within these scenes encouraged this strategy, but if so, it is unclear why it would not also be used for intact scenes (which likewise contain intact horizontal structure). It may be that inverted and rotated intact scenes did not require mental rotation to categorise, or that the unrestricted stimulus display time made the strategy unnecessary, but there is little evidence to support this assumption. Furthermore, it remains unclear how participants would be able to selectively apply this strategy given the studies design. Conditions were randomised, and participants were required to rapidly categorise each scene. If the increased reaction times for the inverted and rotated vertically jumbled scenes resulted from participants attempting to utilise mental rotation, and failing to do so (thus increasing their reaction times) it is not clear why this

additional cognitive load would not be apparent in other conditions where they would likewise need to consider whether or not to mentally rotate the image.

An alternative explanation for the degree of rotation modulating the effect of vertical jumbling could instead be the result of greater manipulation to the scenes low-level visual characteristics. If participants primarily utilise regularities in a scenes low-level visual characteristic when categorising vertically jumbled scenes, as our results suggest, these may be disrupted further when fully inverted compared to rotated.

2.5 General Discussion

In the current study we aimed to investigate how structural regularities across different scenes axes influences scene processing by systematically manipulating spatial structure using a jumbling paradigm. To achieve this, we aimed to replicate previous findings demonstrating that jumbling impairs scene categorisation (Kaiser et al., 2020a; 2020b), explore whether a vertical bias exists in scene processing (Kaiser et al., 2019) and to assess whether the effect of scene structure on categorisation differs between upright and inverted scenes (180° and 90°). Whilst both experiments found evidence for the effect of vertical and fully jumbling scene structure on categorisation, as well as a detrimental effect of inversion and tentative evidence for a vertical bias, there were several inconsistencies between the results that raise important questions about the underlying mechanisms driving the observed effects and suggest that additional factors may have influenced our findings.

In experiment 1 we found that only vertical and fully jumbling structure impacted scene categorisation, whilst in experiment 2 we found an additional effect of horizontal jumbling. Whilst the results of both experiments suggest that vertical and fully jumbling scene structure impacts categorisation, partially supporting our first hypothesis (that jumbling would impact categorisation), only the results of experiment 2 found horizontal scrambling impacts categorisation. However, whilst we failed to demonstrate the effect of jumbling for all of our jumbling types, we did consistently find an impact of fully jumbling scene content. This manipulation most closely resembled the jumbling utilised in previous research (Biederman, 1974; Kaiser et al., 2020a; 2020b. As such, whilst the current study found that jumbling scene structure impacts categorisation, the failure to replicate the effect of horizontal structure across experiments may suggest that further research is required to clarify whether horizontal structure truly impacts categorisation. This discrepancy is particularly important when assessing whether there is a differential effect of jumbling across the two structural axes, as the absence of an effect of horizontal jumbling in conjunction with the observed impact of vertical jumbling could support the notion of a vertical bias.

Across both experiments we found tentative evidence supporting our third hypothesis; that vertical scene jumbling would be more detrimental to scene categorisation than horizontal jumbling. Whilst we did not show a statistically significant difference between vertical jumbling in experiment 1, we did in experiment 2, which used the same paradigm but with a greater number of participants. Furthermore, in both experiments we found that whilst fully jumbling scene structure caused a greater disruption to categorisation than horizontal jumbling, there was no difference between the impacts of fully and vertical jumbling. This could imply that the impact of vertical jumbling was as severe as removing all coherent structure entirely, providing tentative evidence of a potential vertical bias.

However, crucially we did not find the expected effect of inversion observed in previous studies (Kaiser et al., 2020a, 2020b). We initially predicted that the effect of jumbling would be stronger in upright scenes, where some coherent local structural information is maintained within the individual segments, indicating that the jumbling resulted in manipulations to the scenes structure, as opposed to mid and low-level visual characteristics. Contrary to these predictions, in both experiments we found that the effect of vertical jumbling was stronger when the scenes were inverted or rotated. Furthermore, we found that the differences between the jumbling conditions found in the upright scenes were also present during inversion and rotation. The only exception to this was the effect of horizontal jumbling found in experiment 2, which was only present in the upright condition. However, we did not find any statistical difference for horizontal jumbling when displayed upright or at inversion, again failing to replicate the expected inversion effect and making it difficult to determine whether this effect was a result of manipulations to the scenes structural information, or the effect of rotating other mid-level properties such as the edges, contour structures and textures (Oliva & Torralba, 2001). As such, our results contradict our first hypothesis, and instead suggest that the observed effects of vertical and fully jumbling were the result of manipulations to the scenes mid and low-level visual characteristics, as opposed to the scene structure.

What mid or low-level visual characteristics could be responsible for the increased reaction times observed for vertically jumbled scenes in the current study? As the impact of vertical jumbling is amplified in the inverted condition, where these characteristics are further disrupted, this could suggest that the effect of vertical jumbling found in the upright scenes is a result of manipulating low level visual characteristics along the vertical axis. When jumbling vertical structure, segments taken from the upper half of the scene are swapped with those from lower segments. As we used outdoor scenes, these upper segments typically contained portions of sky, whilst lower segments those of the ground. In outdoor scenes, the upper portion (sky) and lower portion (ground) often contain distinct low-level visual features. For example, the sky typically consists of low spatial frequency information, characterized by smooth gradients and large, uniform areas (Julesz, 1981), while the ground is rich in

high spatial frequency content, with sharp edges and complex textures (Thorpe et al., 1996). Swapping these segments during vertical scrambling may have disrupted the visual system's ability to utilise the regularities found in these low-level features. Conversely, when switching segments along the horizontal axis, many of the low-level visual features remain in a typical arrangement, as the left and right halves of a scene often contain similar spatial properties. For example, natural outdoor scenes tend to exhibit horizontal symmetry, where visual elements such as trees, buildings, or landscape features are often balanced on either side of the scene (Torralba & Oliva, 2003). This means that when horizontally jumbling an image, the general distribution of spatial frequencies and textures remains relatively intact, making it less disruptive to scene recognition compared to vertical scrambling. Thus, while vertical scrambling may disrupt key low-level scene statistics by swapping sky and ground components, horizontal scrambling preserves much of the scene's structural integrity, allowing the visual system to maintain familiar global relationships. This may help to explain our failure to replicate the effect of inversion observed in Kaiser et al (2020a, 2020b), which utilised a wider variety of scene types, including both urban and natural indoor and outdoor scenes. The inclusion of indoor scenes may have limited the influence of jumbling these low-level features, thus allowing for the effect of structure to emerge. As such, in order to clarify these results future studies are needed that utilise a greater variety of scenes, especially those of indoor and outdoor places. Additionally, such studies could aim to test this explanation by comparing the effect of different jumbling types between indoor and outdoor scenes.

However, such an explanation does not fully explain why we observed worse categorisation in the inverted and rotated vertically jumbled conditions compared to when they were displayed upright. If the jumbling of low-level visual features can fully explain the effect of scene jumbling, we would expect to have seen similar levels of disruption across both upright and inverted scenes. This interaction between jumbling and rotation could be the result of these conditions representing the greatest possible permutation of the intact scenes, with not only global scene properties (such as structure) disrupted by the coarse jumbling of the scene segments but the local and mid-level visual properties being disrupted by rotation. This interpretation may be supported by the highest levels of disruption occurring in scenes which were both inverted and contained no intact vertical structure. As such, an important consideration for future research is to control for these low- and mid-level visual characteristics through filtering, and the inclusion of scene images that do not have such a stark divide in visual features between the two vertical segments. The latter could easily be achieved through the inclusion of interior scene images where the horizon is higher in the scene images.

Alternatively, a potential explanation is that participants tried to partly solve the task by mentally rotating the scenes to their canonical upright orientation. This process may be less efficient for

jumbled scenes, causing the effects of jumbling to become greater in the inverted and rotated conditions. We tested this by comparing the impact of inversion and rotation, but found that the degree of rotation only impacted vertically jumbled scenes. Whilst this could suggest that participants only attempted to utilise mental rotation when categorising vertically jumbled scenes, it is difficult to explain why they would utilise differing strategies for categorising stimuli when the task demand remains constant between conditions.

The difference between the orientations could alternatively represent the effect of even greater manipulations to the scenes low-level visual characteristics. Such an interpretation would align with the idea that the inverted vertically jumbled condition represented the greatest possible permutation of the intact scene, and so reducing the degree of rotation reduces the level of disruption by bringing it closer to its canonical upright orientation. However, it is important to note that the analysis comparing the degree of rotation was conducted post-hoc, and the study was not designed with this analysis in consideration. As such, the design of our study limited the accuracy of the comparisons that we could draw, especially by necessitating the exclusion of the upright orientation, as this was used in both studies and subsequently was not comparable to the inverted and rotated conditions (which were viewed by different groups of participants). Furthermore, our analysis also compared two groups of considerably different sample sizes. Whilst mixed ANOVA are robust, the presence of a larger sample for the rotated condition may have caused it to dominate the interaction effect, making it harder to interpret the true interaction between orientation and jumbling condition (Schmider et al., 2010). As such, these results require further clarification in order to account for any possible role of mental rotation. This could be achieved in future experiments by utilising brief stimulus presentation times, where mental rotation is not an adaptive strategy for solving the task. Future studies could also aim to test scenes rotated at multiple angles within subjects, so that a more reliable comparison can be made accounting for potential individual differences, which could be especially prevalent when conducting experiments online (due to the added factors of hardware, software and environmental effects).

Our studies failure to replicate the expected vertical bias may also be the result of several key methodological differenced with previous research. Firstly, Kaiser et al (2019) detected the effect utilising a combination of EEG and fMRI. Our inability to isolate its effect utilising a behavioural measure may suggest that the influence of vertical structure could be quite subtle, and require a more controlled design to detect. Additionally, if there is a further influence of low-level characteristics located along the scene's vertical axis, as our experiment suggests, the role of structure may become further obfuscated. As such, the effects observed in previous behavioural experiments demonstrating a bias for information presented along a scene vertically axes (Mandler & Parker, 1976; Previc &

Intraub, 1997;Essock et al., 2003) may have been influenced by a combination of differing underlying mechanisms that utilise various source of information located along a scene's vertical axis, and not just structure. Secondly, the stimuli used in Kaiser et al (2019) differed considerably from those used in the current study, utilising a smaller number of scenes and a different method of dividing scene structure. Whilst in the current experiment, we divided scenes into 4 segments, Kaiser et al (2019) divided scenes once along the horizontal axis and 3 times along the vertical. By dividing the scenes in this way, vertical scene structure may have been more greatly disrupted than horizontal information, which was only split twice, leading to an unequal comparison between the manipulations. In the current study, by dividing the scene into 4 equal segments, both vertical and horizontal scene information was equally disrupted. Additionally, as only a single scene was used for 6 distinct categories, the observed effect may have reflected the organisation of spatial structure unique to the restricted stimulus set used. For example, one of the outdoor scenes used was of an alley way, where vertical structure may have been more diagnostic of scene type due to the edges of the surrounding buildings. The use of both indoor and outdoor scenes in Kaiser et al (2019) may have also caused the differences in our results; as previously discussed, it is possible that the nature of vertical structure may vary considerably between indoor and outdoor scenes, and this could have biased the results of the current study.

As such an important question remains whether these results would apply to other types of scenes, particular indoor scenes. In the current study we focused on outdoor scenes due to supposed high levels of general familiarity, their relatively rigid structural norms and strong global features. However, it is arguable that many indoor scene categories share similarly rigid structural norms, and are also highly familiar to most people, and as such it is plausible that the results of these experiments may replicate for indoor scenes. However, one important difference to consider is the lack of skyline in indoor scenes. As discussed previously, the jumbling of portions of sky, which often have drastically different low level visual properties compared to other scene segments, may explain the current results. With these sky segments being absent in indoor scenes, results may have differed. Specifically, if the observed disruption was the result of manipulations to low level visual properties caused by swapping the sky segments, we would not expect to see these same disruptions in indoor scenes.

Additionally, the differences between the typical spatial structures of indoor and outdoor scenes may also change how structural information is used, and whether an axis-based preference is utilised. Neural processing for indoor scenes elicits stronger activity in areas associated with object recognition, such as the lateral occipital cortex, due to the prevalence of discrete, identifiable objects, whilst outdoor scenes activate regions tied to navigation and environmental context, reflecting the evolutionary importance of recognizing landscapes for survival (Epstein & Kanwisher, 1998; Greene &

Oliva, 2009). As such, the extraction of vertical scene structure may be more relevant to navigational and environmental context, which may be less beneficial in extracting relevant structural information from indoor scenes, where objects provide more pertinent structural scene information. Whilst these ideas are only speculative, they highlight the challenges of applying these results to indoor scenes.

Another outstanding question is whether we would expect scene jumbling to cause similar patterns of disruption if alternative experimental tasks were used? In the current experiment, we chose to use a categorisation task, prioritising fast response times, as we predicted that the extraction of coherent scene structure would occur early within visual analysis, with previous EEG research showing that structural information is extracted from as early as 200ms (Kaiser et al 2020a; 2020b; 2020c). As such, if a bias for vertical scene structure does exist, we would have expected that it would manifest at the early stages of visual processing, and might subsequently impact other tasks that rely on information extracted at later stages of processing. This could result in any potential vertical bias being particularly detectable in tasks that rely upon understanding scene information or context. Visual search tasks, where the structure of a scene is more directly relevant to the task (it is easy to imagine how jumbling scene structure could disrupt search strategies invoked in such tasks i.e. looking for keys on a table is more difficult if the table is in a segment of a scene one would not normally expect to find it), scene categorisation has no implicit task driven reliance on structure. This could suggest that any disruption to categorisation caused by the absents of vertical structure might reflect a more general disruption to scene processing directly, which could create a more general effect observable in other tasks. Consequently, we might also expect that disrupting scene structure might affect other tasks that are similarly less implicitly reliant on structure, such as those reliant on scene memory. This idea is somewhat supported by research showing that more distinctive scenes are both more easily categorised, and also more easily remembered (Greene & Oliva, 2009b; Konkle et al., 2010). As such if intact scene structure aids with scene categorised, it may also make it more easily recalled or remembered. Whilst previous research has found that jumbling impacts memory related tasks, such as recognition (Velisavljević & Elder, 2008) and change detection (Zimmermann et al., 2010), it is unclear whether any axes based bias would also impact these tasks. If and how a potential vertical bias in scene structure would generalise to other tasks remains an open question, and one necessary to understand whether this bias exists for scene processing more generally or is unique to scene categorisation. Future research could explore this by systematically manipulating the axis of jumbling as in the current experiment, whilst utilising a range of different experimental tasks to explore how any potential bias may operate across cognitive domains.

What mechanisms could be responsible for a sensitivity to low level visual information arranged along a scene's vertical axis? As previously discussed, the observed effect of vertical jumbling in both upright

and inverted scenes may be reflective of jumbling segments from the sky and ground, disrupting the scenes typical low-level visual characteristics. One possibility as to why this effect is particularly prominent along the vertical axis, is because it may disrupt the prominent horizon found within outdoor scenes. The horizon may represent an important structural feature for outdoor scenes, as it defines the relative physical constraints of a scene, such as establishing the base from which objects should arise from. If the visual system has adapted to make use of structural regularities within scenes, as suggested by previous research (Kaiser, Quek, et al., 2019; Oliva & Torralba, 2007; Võ et al., 2019; Wolfe, Alvarez, et al., 2011), it may be particularly sensitive to disruptions to information that is used to quickly identify the horizon. Previous research has found that the horizon can be identified very rapidly based only on a scenes low-level visual information. Herdtweck et al (2010) investigated horizon estimates after a brief (150ms) masked presentation of outdoor scenes and found that participants' judgments were consistent and aligned well with annotated horizon data, even when images were blurred, suggesting that global, low-frequency information plays a key role in horizon perception. Additionally, computational modelling revealed that human performance was best predicted by a simple gradient change across the scene, further supporting the idea that the visual system relies on low-level visual cues to rapidly determine the horizon. Of particular interest, they also found that when scenes were inverted, estimates of horizon became significantly worse, suggesting that orientation may be particularly important for identifying the horizon. As such it may be possible that regularities in low-level features organised along a scenes vertical axes provide important cues to identify the horizon, and that the manipulation caused by both jumbling and inversion disrupt this process specifically.

In conclusion, the current study provides evidence that vertical and fully jumbled scene structures impact categorisation, with tentative support for a vertical bias. However, inconsistencies between results in experiment 1 and 2, particularly regarding horizontal jumbling, highlight the need for further research. The findings suggest that low level visual characteristics along the vertical axis may play an important role in scene processing, particularly through disruptions to sky-ground segmentation that may be reflective of important cues used to identify the horizon in outdoor scenes. In order to clarify the results of the current study, future research should explore the effect of jumbling along different axes with a broader range of scene categories, particularly including indoor scenes, and consider alternative experimental tasks to determine the generalisability of a potential vertical bias.

Chapter 3: Individual differences in internal models explain idiosyncrasies in scene perception[1]

3.1 Introduction

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

In the context of scene perception, internal models can be conceptualized as a collection of typical features of a scene (or scene category) that are learned from extensive real-life experience and guide the analysis of matching visual inputs. The contents of internal models are mainly inferred from carefully manipulating the structure of the visual input and observing the resulting changes in perceptual performance and neural representation. Using this approach, researchers could successfully infer key features of internal scene models, such as the typical spatial distributions of objects (Bar, 2004; Biederman et al., 1982; Kaiser, Quek, et al., 2019), semantic relationships between objects and scenes (Davenport & Potter, 2004; Evans & Wolfe, 2022; Oliva & Torralba, 2007; Võ et al., 2019; Wolfe, Võ, et al., 2011), or the spatial layout of whole scenes (Biederman, 1972; Kaiser et al., 2020a; Kaiser & Cichy, 2021).

However, this approach only reveals the contents of internal models that are shared across people – although there is mounting evidence for individual variability in visual perception and neural representation (Charest et al., 2014; de Haas et al., 2019; Gauthier, 2018; Mollon et al., 2017; Tulver et al., 2019; R. Wang et al., 2012). Given that we all differ in our visual experience with scenes across

[1] Parts of this chapter have been published in Cognition (Wang*, Foxwell*, et al., 2024).

our lifetime (Coutrot et al., 2022; Hartley, 2022) and in our neural architecture for visual analysis (Kanai & Rees, 2011; Llera et al., 2019; Moutsiana et al., 2016), it is likely that internal models for scenes are sculpted in different ways across people. If we could harness this individual variability, we would be able to predict and explain characteristic differences in the way each of us perceives the world.

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

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

3.2 Methods

The experiment consisted of 2 parts. In an initial drawing session, participants took part in a drawing task where they constructed scenes representative of typical exemplars of kitchens and living rooms, in order to create approximations of their internal scene representations. Their drawings were then converted into controlled 3D renders. Next, participants completed an online scene categorisation task, where we investigated whether scenes more representative of participants' personal internal models are more efficiently categorised.

Participants

The experiment was approved by the University of York ethics committee. Participants were recruited using the University of York's recruitment pool, Sona-systems (https://www.sona-systems.com). All participants reported having normal or corrected to normal vision. All participants were English speaking and with an average age of 22.6 ± 4.3 years ± SD. 6 participants identified as male, and 29 participants identified as female. They were paid for their participation and provided informed consent before taking part in the study. 43 participants took part in the drawing session. 39 of the participants returned for part 2, whilst 4 participants were excluded because their performance did not exceed guessing performance (based on binomial tests against chance level), leaving us with a

final sample of 35 participants. Sample size was based on convenience sampling, with the target to exceed 80% statistical power for a hypothesised medium effect of d = 0.5 in a two-sided t-test. For this target, at least 34 participants are required.

Drawing session

To obtain descriptors of participants' internal scene models, they first took part in a drawing task, where they were instructed to draw typical examples of different scenes. Here, participants were tasked with drawing scenes from 2 different categories: living rooms and kitchens. Critically, they were instructed to draw their interpretation of the most typical example of that scene type. The definition of typical was given as the most generic and ordinary example they could think of. They were also instructed not to draw a scene that they thought looked particularly interesting or attractive, nor an exact copy of a scene they knew from real life (such as simply producing a copy of their own kitchen or living room). They were given 1 minute to plan and think about what their most typical scene should look like. They then had 3 minutes and 30 seconds to draw the scene, using a pencil, rubber and ruler. Scene sketches were drawn into a perspective grid, to allow participants to more easily draw in 3D as well as to standardise the participants viewpoint across all scenes. See Appendix D for examples of participant drawings.

Perspective grids were either drawn or printed by the participant on A4 paper and consisted of a large central rectangle (7.1cm by 16.5cm) and 4 diagonal lines going from each corner of the rectangle to the corners of the page. The rectangle was drawn slightly raised from the vertical centre on the page, with the bottom length 8.5cm from the bottom of the page and top length 5.4cm from the top of the page. Both sides of the rectangle were drawn 5.4cm away from the sides of the page. Grids thereby created the outline of a room, with the large central rectangle acting as the back wall, the top and bottom segments the ceiling and floor, and side segments as the side walls (see Appendix C).

Participants were reminded how much time they had left at the halfway point, and when they had a minute remaining. They first drew a practise scene of a bedroom, to get them used to the timings and drawing on the perspective grid. This also allowed the experimenter to check that they understood the task instructions. The order in which they drew the other scenes was balanced across participants. After completing each drawing, participants also drew a coarse birds-eye view of the scene, in which they labelled all the objects in the scene. This was done to help clarify the room's intended 3D layout and to confirm the identity of any ambiguously drawn objects, providing additional information for generating accurate 3D renders of the drawings.

In addition to drawing their most typical versions of living rooms and kitchens, participants drew copies of a given photograph of a living room and kitchen. These copies were drawn under the same time constraints as the sketches, and participants were instructed to capture a similar amount of detail as they used in their own drawings. They were given 1 minute to study the photo, followed by 3.5 minutes to sketch it, and had access to the photograph throughout their drawing time. These copies acted as a control for memory effects in the subsequent scene categorisation experiments: Participants will have seen and drawn these scenes, just like their typical versions of living rooms and kitchens, but they will not adhere to their internal models of what living rooms and kitchens typically look like.

Stimuli

To produce stimuli specifically tailored to participants' internal models, we created a set of 3D renders that optimally captured the properties of participants' scene sketches produced in the drawing session. We used renders instead of the original line sketches as they allowed us to standardize the stimulus set in several ways: Firstly, renders allowed us to equalise differences in drawing ability, whilst accurately maintaining key aspects of the scenes' content and structure. Secondly, low-level visual features can be readily standardized in the 3D renders. Finally, the object content of the 3D renders can be manipulated in precisely controlled ways.

In order to create the 3D renders we used the video game "The Sims 4" (*The Sims4*, 2014). The Sims 4 is a social simulation game, that allows the user to create and design different characters and houses, and then to play out different social scenarios and objectives. The game includes a comprehensive, highly detailed and easy to use design software that allows the user to create a range of 3D environments by placing walls and objects onto a grid-like system (known in the game as "Build Mode"). The use of the Sims 4 allowed us access to a large library of thousands of 3D modelled candidate objects for building the renders: We could thereby choose from a comprehensive and diverse set of exemplars for any objects we required.

When constructing the 3D renders, first an empty room was built to replicate the view and approximate dimensions of the perspective grid. This room was approximately 6 x 6 cm in size and used wall pieces approximately 3m high, with the outward facing wall was removed (see Figure 3.1). This created an empty room structure that approximately resembled the perspective grids the scene sketches were drawn in, which acted as the starting point for building other 3D renders. The scenes were then populated with objects by referencing both the scene sketch and birds-eye view plans the participants constructed in the drawing session. The amount of detail a given individual object was drawn in varied greatly; the closest matching 3D object was chosen to represent it in the render, but

when objects were drawn in very little detail, a highly generic version of that object was used (at the experimenter's discretion). Once a scene render was completed, screen shots were taken using the Xbox live app for Windows (*Xbox App for Windows*, 2009). Screen shots were taken from the same distance and angle for every scene render, cropped so that only the room was visible, and resized to 820 x 390 pixels. In order to control for low-level visual differences between the resulting images, all images were grayscaled and their mean luminance and contrast were matched using the SHINE toolbox for MATLAB (Willenbockel et al., 2010). See Appendix E for further examples of the stimuli.

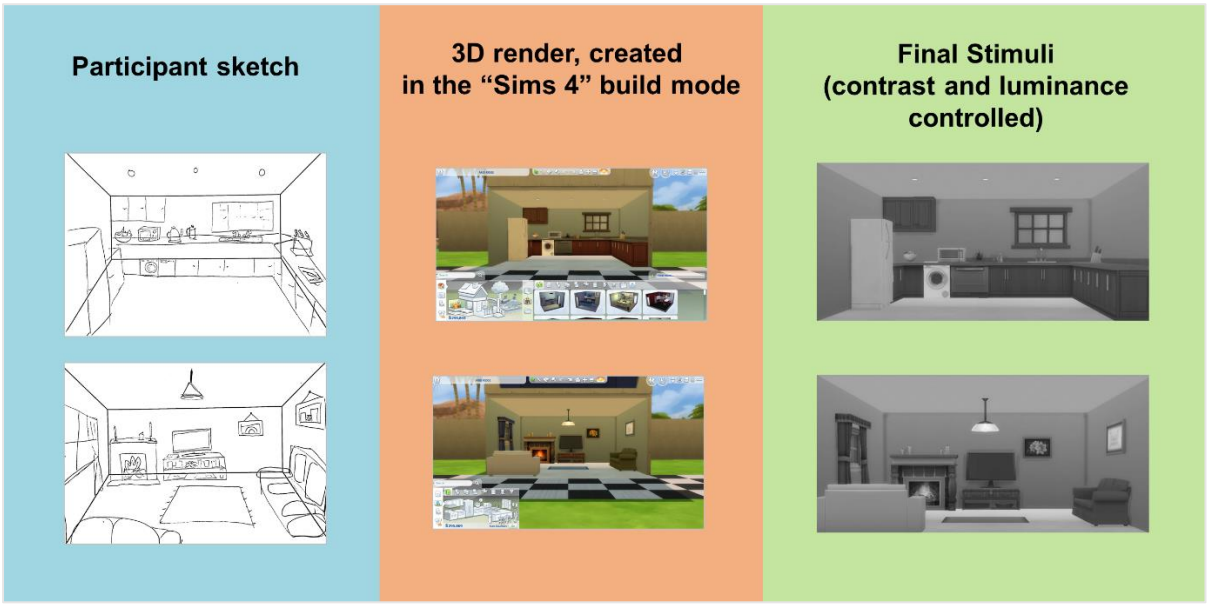


Figure 3.1. Examples of the scene drawings and 3D renders. Participants first sketched the scene inside the perspective grid. These were then converted into 3D renders using the design tool in "The Sims 4" build mode. They were then cropped, grayscaled and standardized in their low-level features.

Procedure

Participants took part in an online scene categorisation task where they were asked to indicate whether a briefly presented scene was either a living room or a kitchen as accurately as they could (see figure 3). The experiment was created and hosted using the Gorilla online experiment builder (Anwyl-Irvine et al., 2020). Before the experiment, they were instructed to maximise their browser window and sit approximately 60 cm away from the screen. After reading the instructions, participants were shown 2 examples of each scene category (these were not included in the experiment).

During the experiment, participants viewed 3D renders based on their own drawing of a typical scene ("own" condition), based on other participant's drawings of typical scenes ("other" condition), and

based on their copied scenes ("control" condition; the control renders were identical for all participants). In total, 88 renders were shown in the experiment, 2 of which corresponded to each participants' own drawings, 2 of which corresponded to the copied scenes, and the other 82 corresponded to the other participants' drawings. The stimuli were thus initially based on the drawings of 43 participants, but 5 of them did not return for the experiment after the drawing session. Each scene render was repeated 10 times, for a total of 880 trials. Trial order was randomized. The experiment was split into four blocks. After each block, participants were given a 1-and-a-half-minute break.

Stimuli were displayed on a grey screen. Trials began with a blank screen, followed by a central fixation cross for 1000ms. Next, the scene render was flashed for 83ms, followed by a mask presented for 150ms. Masks consisted of a random arrangement of squares, diamonds, and circles. On each trial one of 43 unique masks was chosen randomly. A blank screen was then displayed until the participants responded by either pressing "K" or "L" on their keyboard (to indicate whether a scene was a kitchen or living room). There was no time limit for participants to give their answer. After the participants gave their response there was a 100ms delay before the fixation cross was shown again and the next trial started (see Figure 3.2).

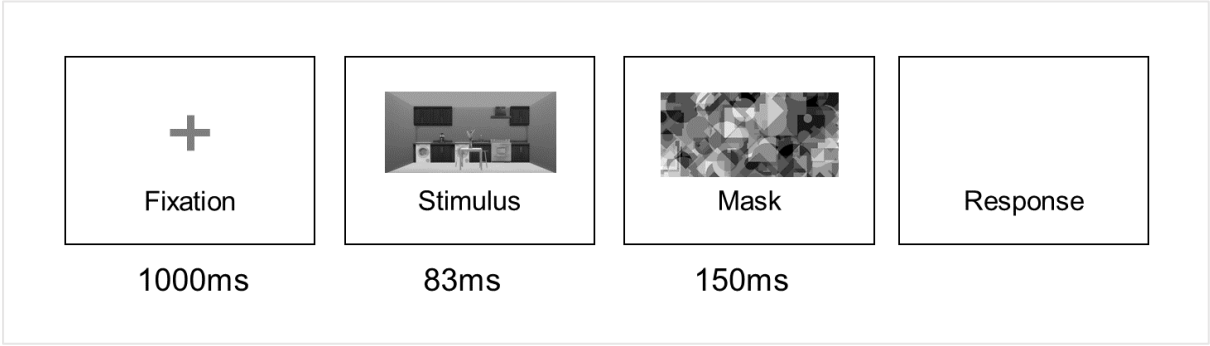


Figure 3.2. The trial structure for experiment 1 session 2. First a fixation cross was shown for 1000ms, followed by the stimulus for 83ms and a mask for 150ms. Participants than gave their answers by pressing either "K" or "L" on their keyboard (to indicate whether a scene was a kitchen or living room).

Statistical Analysis

To compare categorisation accuracies across conditions, we used one-way repeated-measured ANOVAs and paired-samples t-tests.

To investigate whether a graded similarity to the participants own scene images predicted processing efficiency, we used a deep neural network (DNN) to measure how similar each scene render based on other participants' typical drawings (hereinafter: candidate scenes) is to the renders based on the current participant's drawings (hereinafter: reference scenes), and then correlated the resulting similarity score with the participants' categorisation accuracy for these scene renders (see Figure 3.3). For all candidate and reference scenes, we first extracted activation vectors from the convolutional layers and the final fully-connected layer of GoogLenet deep convolutional neural network (Szegedy et al., 2015), which was either pre-trained on scene categorisation using 1.8 million scene images from the Places365 data set (B. Zhou et al., 2018) or objects trained on the ImageNet dataset (Deng et al., 2009). Activation in hierarchical DNN layers were used as approximations for the hierarchical stages of visual processing, with earlier layers of the DNN being more representative of lower-level processing (such as shapes, forms and colours) and higher levels representative of higher level visual processing (such as the scene's object content)(Cichy & Kaiser, 2019; Kriegeskorte, 2015).

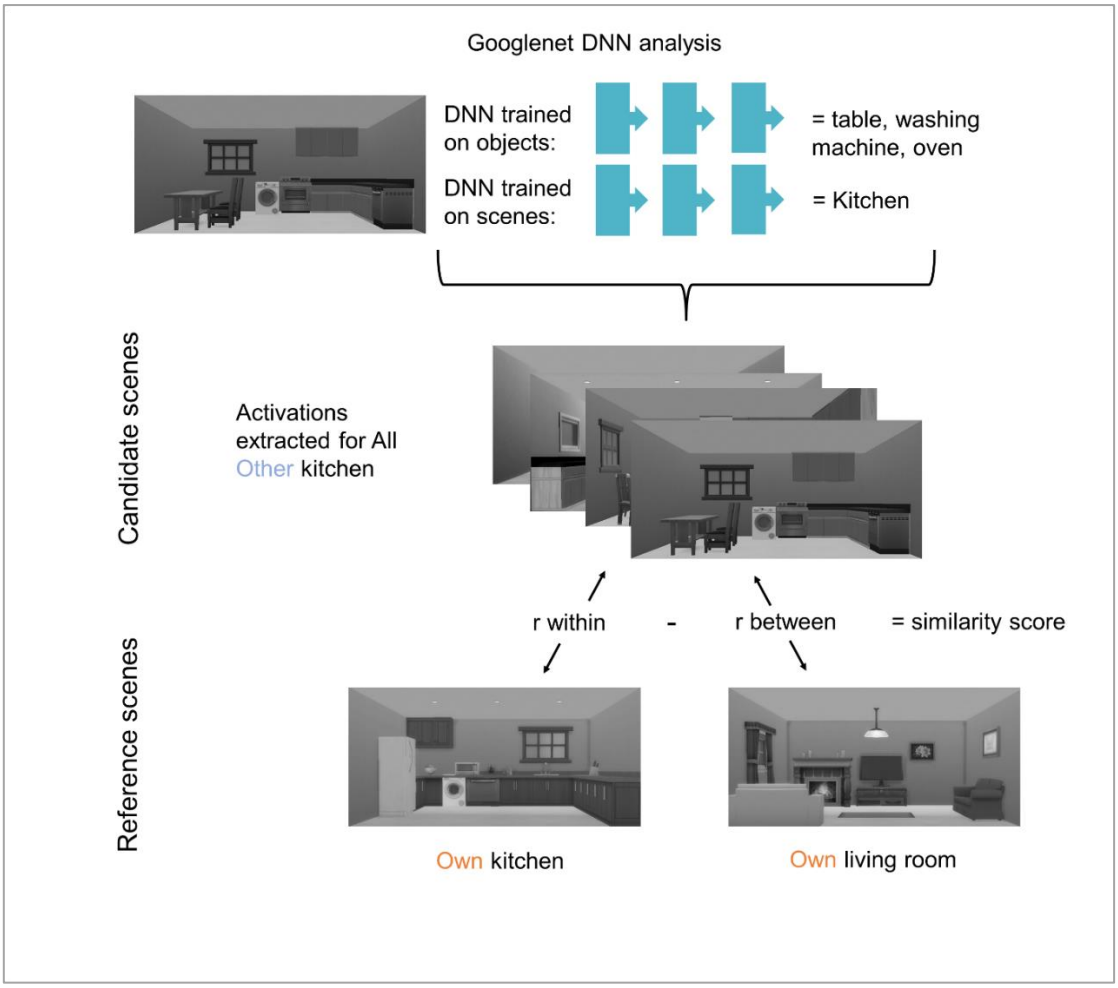


Figure 3.3. We extracted activation patters for all scene renders in Experiment 1 from GoogLenet DNNs trained on scene or object classification. To approximate the processing of complex, high-level visual features, we extracted activation patterns from the last inception module of the DNN. To quantify similarity to the internal model, we correlated the activation pattern for each other scene to the own scene of the same category (within-category correlation) and each own scene of the other category (between-category correlation), separately for each participant. By subtracting the within- and between-category correlations, we obtained a graded similarity measure, which we correlated with the behavioural categorisation accuracy across all candidate images. This analysis was repeated with all possible other scenes or the control scenes as the reference images.

By systematically correlating these activation vectors, we obtained two similarity relations: (1) within-category similarities, capturing how similar of candidate scenes is to the reference scene of the same category, and (2) between-category similarities, how similar of candidate scenes is to the reference scene of the opposite category. By subtracting the between-correlation from the within-correlation, we created "similarity score", which captured how similar each scene render was to the internal model

of the same category, relative to the internal model of the opposite category. The aggregated correlations were Fisher-transformed, and subjected one-sided t-tests across participants, where correlations greater than zero indicated that higher DNN similarity scores indeed predict higher categorisation accuracy. The resulting *p*-values were FDR-corrected across the DNN blocks. Notably, each time one of the other scenes was the reference, one scene less was available for computing the correlations. For the analyses in which the own or control scenes were the reference, we thus iteratively removed one of the other scenes before computing the results and then averaged across iterations.

3.3 Results

Scenes tailored to participants' internal models are processed more efficiently

We first analysed participants categorisation performance (kitchen versus living room) across the own, other and control conditions (see Figure 3.4.a). The three conditions yielded significantly different accuracies ($t(2,68) = 4.15$, $p = .020$): Participants were significantly more accurate for scenes that were tailored to their own drawings of typical scenes than for scene tailored to others' drawings ($t(34) = 2.18$, $p = .036$) and scenes they had copied before ($t(34) = 2.26$, $p = .031$). We found no statistically significant difference in accuracy between the other and control conditions ($t(34) = 1.11$, $p = .280$). These results suggest that scenes resembling individual participants' internal scene model are more efficiently processed than those resembling other people's internal models. The lower accuracy in the copy condition suggests that the benefit for scenes tailored to participants' own internal models could not be explained by familiarity to the scenes acquired during the drawing session. Together, this finding suggests that variations in scene perception are indeed explained by our own personal priors of what the world should look like.

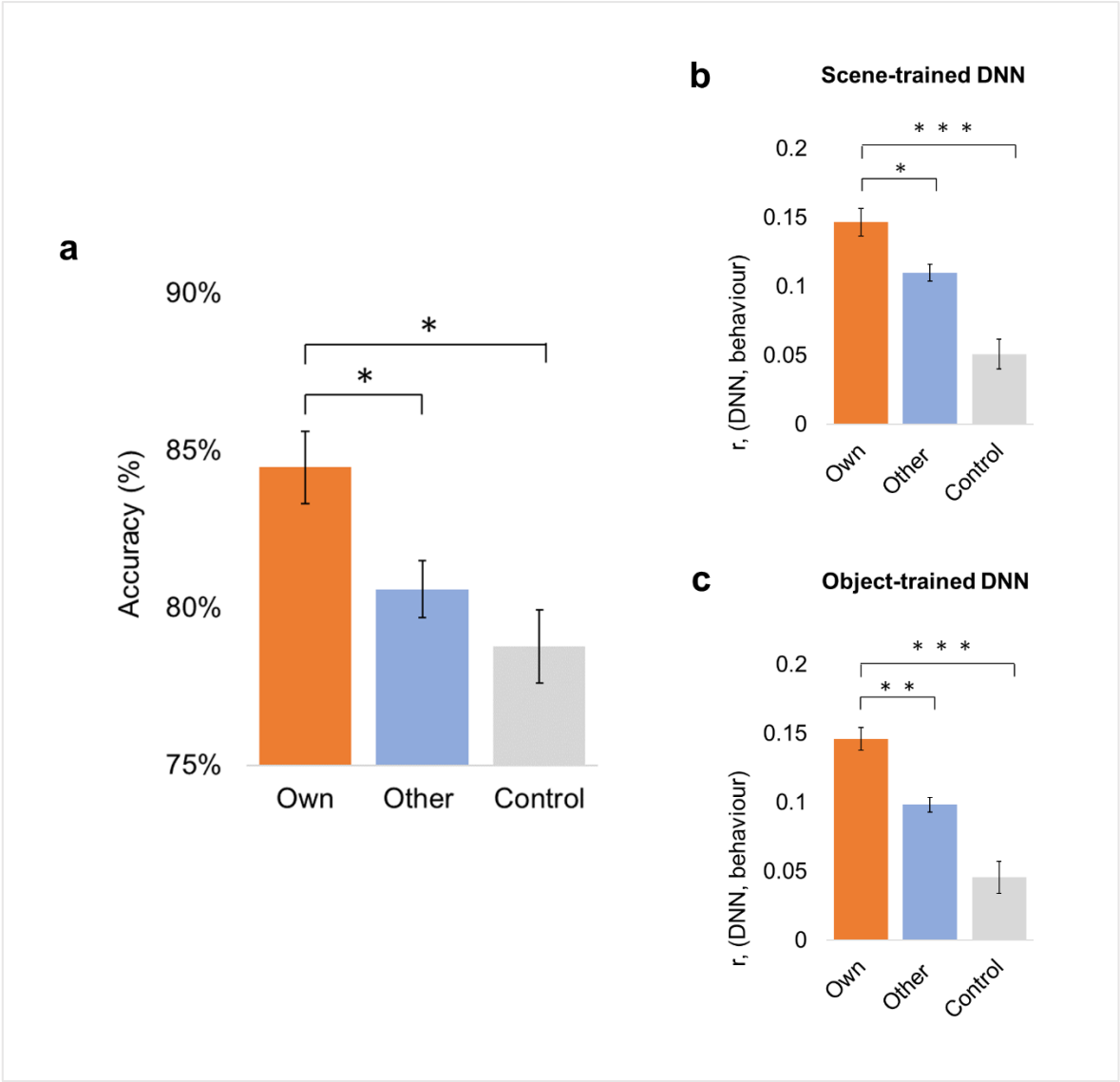


Figure 3.4. a) Mean reaction times for renders based on either participants' own, other and control scenes. We found that participants were significantly more accurate at categorising scenes based off their own drawings than those produced by other participants or a copied scene control. b) Graded similarity analysis for DNN trained on scenes. c) Graded similarity analysis for DNN trained on objects. In both DNNs, graded similarity to the own scene predicted categorisation better than graded similarity to the other or control scenes, suggesting that similarity to participants' personal internal models predicts behavioural categorisation across the range of images used in the experiment. Error bars represent standard errors of the mean. * Indicates $p < .05$, ** indicates $p < .01$, *** indicates $p < .001$.

Graded similarity to internal representations predicts categorisation accuracy

The results from Experiment 1 suggest that scenes specifically engineered to match participants internal models are preferentially categorised. However, if internal models indeed function as templates for categorisation, we should also see a graded benefit of similarity to the internal model: The more similar any scene is to participants internal model, the better categorisation performance should be. To investigate whether graded similarity to internal scene models could predict scene classification, we used a deep neural network (DNN) model to compute the similarity of all scene renders that were generated to match other participants' drawings (henceforth: candidate scenes) to the renders that were generated to match the participants' own drawings (henceforth: reference scenes). In order to quantify the similarity between these scene renders, we first processed both the candidate and reference scenes through two GoogLenet DNN models trained on scene or object categorisation and extracted activation vectors from deep layers of the DNN (see statistical analysis for details), analogous of high-level visual processing. We then correlated the activation vectors for each candidate scene with those for the two target scenes (kitchen and living room). For each candidate scenes we then subtracted the similarity to the reference scene of the same category from the similarity to the reference scene of the opposite category. This yielded a "similarity score" which was higher if the image was more similar to the participants' internal model of the corresponding category or more dissimilar to the internal model of the opposite condition.

For each participant, we then correlated the similarity scores for each candidate scene with their accuracy in the scene categorisation task, allowing us to infer how well similarity predicted categorisation. We found that in both object and scene trained DNNs, across participants, there was a positive correlation between similarity score and categorisation accuracy for both own and other scenes ($t(34) = 7.37$, $p < .001$). Graded similarity to the control scenes was a weaker predictor of categorisation, both in the object-trained ($t(34) = 1.96$, $p = .058$), and scene-trained DNNs ($t(34) = 2.34$, $p = .025$) (see Figure 3.4.b and Figure 3.4.c). Comparing predictions between the own, other, and control scenes as references, we found a significant difference between conditions in both networks (both $F(2,68) > 13.3$, $p < .001$). Critically, graded similarity to the own scenes between predicted behavioural performance better than graded similarity to the other scenes for both the object trained DNN ($t(34) = 3.31$, $p = .002$) and scene trained DNN ($t(34) = 2.26$, $p = .030$), and better than graded similarity to the control scenes (object trained DNN: $t(34) = 4.04$, $p < .001$, scene trained DNN: $t(34) = 4.65$, $p < .001$). This confirmed our prediction that graded similarity to participants' individual internal models determines categorisation performance.

3.4 Discussion

Together, our findings provide new insights on individual differences in naturalistic vision. We show that participants are better at categorizing scenes that resemble a typical drawing they had produced prior to the experiment, compared to scenes that resemble other people's typical drawings, or scenes that resemble scene copies they had produced earlier. Using a DNN as a measure of graded similarity, we further show that categorisation varies as a function of the similarity between participants' drawings and the scene that they are asked to categorize. We interpret these findings to reflect differences in participants' internal models of the world that are captured by their typical scene drawings. These differences in internal models may in turn drive idiosyncrasies in scene categorisation.

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

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

than providing a single monolithic reference point, are perhaps defined by a probability distribution in representational space.

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

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

Our study further prompts questions that provide new avenues for future research. First, we currently do not know why internal models systematically differ across participants. Future studies could relate variations in internal models to idiosyncrasies in cortical representation (Charest et al., 2014; J. Lee & Geng, 2017) and visual exploration behaviour (de Haas et al., 2019; Henderson & Luke, 2014), as well as to individual differences in brain anatomy (Kanai & Rees, 2011; Llera et al., 2019; Moutsiana et al., 2016). Second, we do not know exactly how visual inputs are matched against the internal models. There is a variety of dimensions along which this match could be computed, such as the objects included in a scene as well as their spatial distribution (Kaiser, Quek, et al., 2019; Oliva & Torralba, 2007; Võ et al., 2019; Wolfe, Võ, et al., 2011)the global geometry of the scene (Epstein & Baker, 2019; Kaiser & Cichy, 2021; Oliva & Torralba, 2006), or low- and mid-level features correlated with the content of a scene (Geisler, 2008; Groen et al., 2017; Watson et al., 2014). By object-related features, we refer to scene information that pertain to the identity, position, and spatial relationships of constituent objects. Object identity refers to recognising an item as belonging to a certain category (e.g., a chair, a tree), while object position and spatial relationships capture how objects are arranged relative to each other within a scene (Kaiser et al., 2018; Epstein, 2008). These features are critical for scene understanding and are processed across multiple brain regions, including the LOC for object recognition and the PPA for spatial configurations (Grill-Spector & Weiner, 2014; Konen & Kastner, 2008; Epstein & Baker, 2019).

Our DNN-based analysis of graded similarity indeed suggests that high-level features are important, given that graded similarity in a deep layer of a scene-trained DNN predicted categorisation performance. We define high-level visual features as the complex properties of a scene that go beyond basic low-level and mid-level visual attributes (such as edges, contrast, or luminance). High-level features include object identity, spatial layout, and global scene properties, all of which contribute to semantic scene understanding (Kravitz et al., 2011; Oliva & Torralba, 2007). These features may be processed in higher-level visual areas, such as the inferotemporal cortex for object recognition and the parahippocampal cortex for scene categorisation (Grill-Spector & Weiner, 2014; Epstein, 2008). The observation that predictions were enabled by both object- and scene-trained DNNs suggests that the features useful for prediction are not uniquely critical for either object or scene recognition. However, our scene renders were carefully matched for low-level features, and this matching may have obscured a possible contribution of low-level features that are relevant under more naturalistic conditions. To chart relevant visual features more comprehensively, future studies could systematically manipulate inputs to deviate from the internal model in targeted ways.

More generally, our study highlights the potential of drawing for quantifying internal representations (Fan et al., 2023). Drawings indeed received renewed attention recently, in studies of scene memory

(Bainbridge et al., 2019; Bainbridge & Baker, 2020) and perception (Fan et al., 2018; Matthews & Adams, 2008; Morgan et al., 2019; Ostrofsky et al., 2017; Singer et al., 2023). Our study suggests that drawings also yield the potential to advance our understanding of the internal models that guide the visual representation of objects, faces, or actions. Furthermore, our drawing method may prove useful for studying the maturation of internal models across development (see Long et al., 2024) or their alterations in disorders of prediction like autism (Pellicano & Burr, 2012).

In sum, our work provides two critical advances for studying vision on the individual level. First, our findings offer a new interpretation of individual differences in perception. They suggest that humans categorize real-world environments in different ways because we all have different internal models of the world. Second, our work provides researchers with a new drawing-based method for unveiling the contents of internal models in individual participants. This method has the potential to be widely applied to derive explicit predictions about individual differences in vision.

Chapter 4: Investigating the object and spatial content of internal scene models

4.1 Introduction

Understanding how humans perceive and interpret scenes is fundamental to vision science. Scene processing has been theorised to utilise internal scene models—mental representations that guide recognition and processing (Morgan et al., 2019; Muckli et al., 2015; Peelen et al., 2024). While internal models are thought to guide scene perception, it remains unclear what specific types of information they contain and how this may influence scene perception. Previous studies have shown that scene perception relies on semantic and syntactic object information (Võ et al., 2019; Võ & Wolfe, 2013), yet it is unknown whether this information is explicitly represented within internal models. In chapter 3, we show how drawings can be used to probe the contents of internal scene models, showing that participants categorised scenes more efficiently when they were based on their own drawings rather than those of others. In a control condition, we also found that this effect was not just a result of familiarity with the stimulus itself, as participants showed no such improved accuracy for renders based on drawings of photographed scenes produced at the same time. These findings suggest that drawings are a reliable readout of internal scene models, and that similarity to these models can predict the efficiency of scene processing. However, a critical question remains: What types of information within an internal model drive this effect?

One possible source of this information are the objects found within a scene. Whilst scenes contain a wealth of visual information, they are inherently defined by their constituent objects. Võ (et al., 2019) proposed that objects in scenes contain two primary sources of information: semantic and syntactic. Semantic information pertains to an object's identity and its alignment with the room category, while syntactic information relates to the spatial properties of objects within the scene. In a bedroom, a bed would exhibit high semantic consistency with the scene category, whereas a bathtub would be unusual and thus show low semantic consistency. Syntactic information is also highly predictable; for instance, we expect the bed to be placed on the floor, upright, and against a wall. Both syntactic and semantic scene information has been shown to be distinguishable from each other within the visual system. Võ & Wolfe (2013) conducted an EEG experiment in which participants were exposed to semantically and syntactically inconsistent scenes. They found that semantic inconsistencies produced a negative deflection in the N300/N400, whist syntactic inconsistencies elicited a late positivity resembling the P600. This finding mirrored differences in neural processing for syntactic and semantic information found outside of scene perception, with activation of N400 found for semantic violations of verbal (Holcomb, 1993; Kutas & Hillyard, 1980) and pictorial information (Kutas & Federmeier,

2011), and P600 activations associated with intact syntactic information in language processing (Hagoort et al., 1993; Kutas et al., 2006).

Both typical semantic and syntactic scene content has been shown to aid efficient scene processing, and as such may be represented within internal scene models. Davenport and Potter (2004) found that when placing objects within either semantically consistent or inconsistent scenes (such as a football on a football pitch vs a shark in a desert) and asking participants to identify the objects and background scene as quickly as possible, not only did they find that semantic consistency facilitated object recognition, but also that the presence of semantically consistent objects aided in scene recognition. This benefit has been found to be very robust, with Evans and Wolfe (2022) finding that participants were unable to separate objects from their backgrounds, even when these were detrimental to task performance or semantically inconsistent, indicating the strong role of context in object processing.

Similarly to Davenport and Potter (2004), Brandman and Peelen (2019) conducted a study investigating how objects facilitate scene processing when other scene information is obscured, rendering the scene category ambiguous. They blurred indoor and outdoor scene photographs, making the scene category challenging to discern, but this difficulty was mitigated by the inclusion of semantically consistent objects. Testing classifiers trained on un-blurred indoor and outdoor scenes, they discerned response patterns in scene-selective brain regions, finding more accurate classification in left PPA and the OPA for scenes containing semantically consistent objects compared to those without, suggesting that typical objects positioned in typical scene locations aids in scene categorisation. If typical syntactic and semantic scene information facilitates efficient scene processing, the visual system would require a way to gauge what information is indeed typical. This typicality could be inferred by referencing a scene's object content against an internal model for that scene category. Scenes featuring typical object types and arrangements may be more easily indexed against prototypical category level semantic and syntactic object information contained in the internal model and thus be perceived more efficiently.

While previous studies have demonstrated that typical semantic and syntactic scene content aids recognition (Davenport & Potter, 2004; Brandman & Peelen, 2019), these studies assume a shared, universal representation of scene structure. However, our findings from Chapter 3 suggest that internal models vary across individuals, raising the question as to whether individual differences in semantic and syntactic scene representations within internal models shape scene perception? If semantic and syntactic information are contained in individually specific internal models, then we might expect to observe a greater sensitivity to manipulations of semantic and syntactic information

in scenes that more closely resembles that individual's internal model, whilst violations to scenes further from their internal model would be less likely to cause disruptions, as the content already differs from that observed within the internal model (see Figure 4.1). Alternatively, semantic and syntactic information might be representative of broader rules of scene grammar, shared across individuals. This information could be acquired through exposure to broadly shared rules found within scene content that are almost universally shared. In the case of semantic information, this might represent a sensitivity to violations in key defining objects almost always present within specific exemplars of that scene category, such as beds in bedrooms, whilst broad syntactic rules might reflect equally universal rules of object placement, such as the physical confounds dictated by forces such as gravity or in objects interjecting within each other (Võ et al., 2019; Võ & Henderson, 2009; Võ & Wolfe, 2013). Given that many studies have found that shared conception of typical object information helps facilitate efficient scene processing, it is perhaps most likely that internal models could contain a combination of typical semantic and syntactic information derived from both personal exposure, and those more universally. However, no study has explicitly examined individual differences in the use of semantic and syntactic information during scene processing.

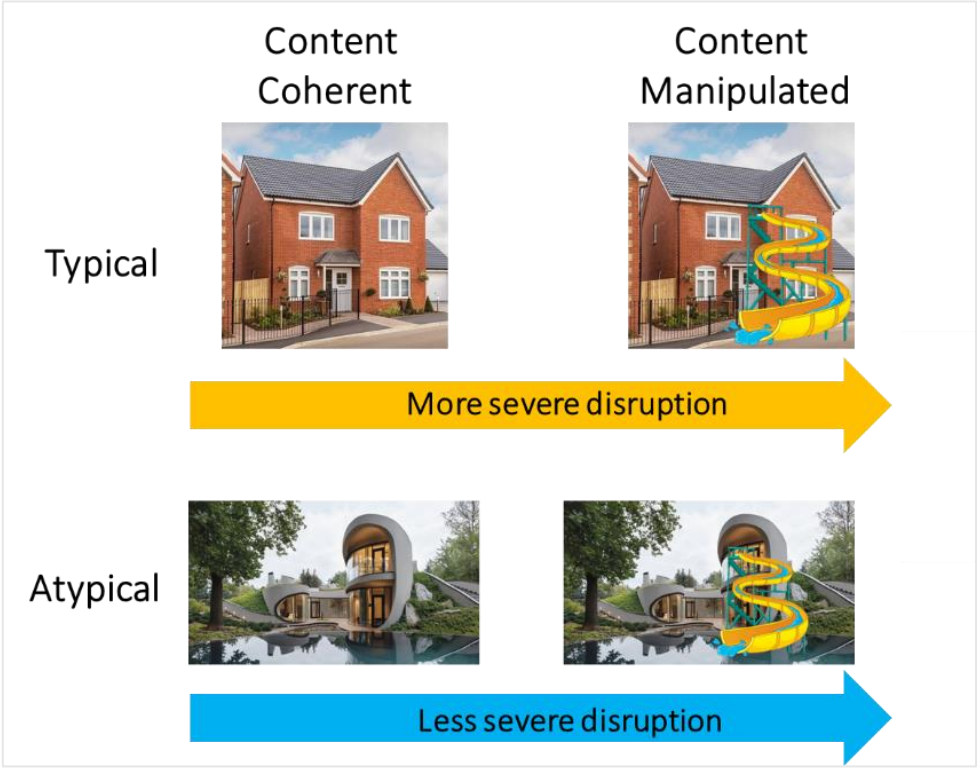


Figure 4.1. In chapter 3 we demonstrated that the benefit of scene typicality may be dictated by the strength of the match between the incoming scene information and an internal model of the scene. When scene content is manipulated (exemplified with the toy example of adding a water slide to a

house) in a scene more closely resembling the internal model, this manipulation may cause a greater relative disruption in processing compared to scenes that are already a poor match to the internal model, and thus appear more atypical. Manipulating the content within less typical scenes may only change scene information that is already a poor match to the internal model, rather than changing more diagnostic scene information found in more typical scene exemplars.

Our drawing paradigm provides an opportune method for investigating individual differences in the representation of semantic and syntactic information in internal scene models. Here, participants produce line drawings based on their own judgement of what a typical scene in a given category looks like, acting as a proxy for the content of their own internal model. These line drawings are then used to produce 3D scene renders, representing the same objects and locations as the original drawing. Crucially, the objects within the scene renders can be easily manipulated, without necessarily causing inconsistencies in the representation of the original scene image. This means, that unlike manipulating a line drawing, objects can be moved, replaced, or otherwise manipulated without causing disruptions to the scene (such as terminating lines or creating artificial object placements in a drawing) that may cause the changes to be artificially more noticeable. Furthermore, changing objects within computerised renders does not require the style of the original drawer to be replicated, meaning that changes to the scene are not made more obvious by the success of the imitation. The intensity of the manipulation can easily be modulated by increasing the number of objects within the scene that are changed, allowing us to investigate the sensitivity of the perceptual outcome to these changes. Together, these qualities allow both object identity and location to be easily changed, manipulating elements of scene grammar without disrupting the overall style or coherency of the image.

In the current study we had 3 aims; 1) to investigate whether object related information is stored within internal scene models, 2) and how robustly, and 3) to replicate the results of chapter 3, in order to provide further evidence for the ability of line drawing to access information about our internal scene models. To achieve this, we use our previously established drawing paradigm to investigate the role of semantic and syntactic object information stored within internal scene models, by manipulating both the identity and location of constituent objects within scene renders based on a participant own drawing of typical scenes, drawings produced by other participants or copies of existing scenes. If scene renders are able to represent the content of an individual's internal scene model, as the results of chapter 3 suggest, then we would expect manipulating objects within renders based on a participants own drawing would disrupt their ability to successfully categorise renders, as these manipulations reduce their semblance to a participants' own internal model. Conversely, we would not expect any impact of categorisation of renders based on drawings produced by other

participants or based on existing scenes, as there is equal chance that any changes made would make these renders an any better or worse match to a participants own internal model.

We further investigated whether any effects were modulated by the severity of the manipulation (i.e., whether one or two objects were manipulated), to test how robustly this object information is encoded within internal models. If object information does help facilitate the comparison of internal models to external scenes, we would expect scenes containing object information more closely matching that of the internal model to be categorised more accurately than those that deviate further. As such, if representations of object content are more rigid, we would expect more severe manipulations to cause a greater impact on categorisation compared to less severe manipulations. Alternatively, if the representation of object content is more flexible, then we would expect to see little difference between the severity of the manipulations applied, indicating that internal models either rely little on object information or that the matching process is robust enough for such manipulations to have little effect.

Two experiments were conducted: in experiment 1 we compared scene categorisation between renders based on either participant own drawings, those based on other people's drawings or based of copies of existing scenes when constituent objects were manipulated by either replacing (manipulating semantic information) or swapping (manipulating syntactic information but preserving semantic information), at two different levels of severity. In the second experiment, we aimed to confirm our results by only testing the more severe manipulations, due to concerns that displaying both manipulations may have biased our finding.

We hypothesised that manipulating both object identity and location would have a greater impact on categorisation performance for scenes based on a participant's own drawings, acting as proxies for their own internal models, than for scenes based on other participants' drawings or copies of existing scenes. We further hypothesised that the severity of the manipulation would modulate this effect, with more severe manipulations causing the scene render to become an even weaker match to the participants internal model. As such, we would only expect the severity of the manipulation to impact renders based on a participant's own drawing, and not those of the based of other's drawings or copies of pre-existing scenes. Finally, we also hypothesised that as in chapter 3, we would find scenes based on a participant's own drawings were more accurately categorised compared to those based on other people's drawings or copies of pre-existing scenes. By investigating these questions, we hope to further our understanding of how the brain utilises internal models to help facilitate scene categorisation.

4.2 Experiment 1

4.2.1 Methods

Participants

This study was exploratory in terms of sample size as there was little prior research with directly comparable methodologies; therefore, we did not have a strong empirical basis for conducting a formal power analysis. Instead, we aimed to approximately match the sample size used in Chapter 3, expecting similar effect sizes. This approach allowed for consistency across studies and an initial examination of potential effects. In total 38 participants took part in experiment 1, 7 participants identified as male, 30 identified as female and 1 as another gender, with an average age of 24.76 years ± SD. 6.11 . Of these 15 participants had also taken part in our previous drawing experiment (as outlined in chapter 3). This allowed us to take advantage of the already existing scene drawings and renders, meaning that the scene drawing session did not need to be repeated for these participants. Procedures were approved by the ethics committee of the Department of Psychology, University of York, and adhered to the Declaration of Helsinki. Experiment 1 was conducted online, and participants provided informed consent through an online form.

Whilst returning participants may have been more experienced with the paradigm, and already had some exposure to the renders designed to match their drawings, we did not expect that this would affect the experiment systematically. This was based on the results of chapter 3 where we included a control condition for memory effects and found prior familiarity with scene drawings could not explain the improved recognition we observed. In addition, if any undetected memory effects did exist, these would be mitigated by the design, as we compare the relative ability of participants to recognise scenes within individual participants, as opposed to between them. Returning participant also produced their scene drawings at the same time as the controls, so any memory effects would affect both stimuli equally. In addition, for new participants who had to complete the drawing task, at least 1 week would elapse between the drawing task and the scene categorisation experiment, meaning all participants would experience a significant gap between these tasks.

Participants were randomly assigned into groups of 3. Participants in each batch only viewed the renders based on their own and the two other participants drawings, as well as the renders based on the control scenes. This was done to achieve a viable length to the experiment, as participants were required to view multiple versions of each image, and including more sets would greatly increase the length of the experiment, leading to possible fatigue effects.

Stimuli

Scene renders

Stimuli consisted of intact and manipulated 3D renders of scenes. The intact scene was produced using the same method as in chapter 3. These were based on participants' own drawings of typical living rooms and kitchen scenes, other people's drawings of these scenes, and copied control scenes. These drawings always consisted of living rooms and kitchens due to participants' familiarity with these scene categories, and variability of the items associated with these categories.

Each render (own, other and control) was displayed as either intact or manipulated. Two manipulations were applied to each scene; either swapping the location of two objects within the scene or replacing one object with a similar scene-appropriate object. For both manipulations, we additionally created two levels of intensity, by either swapping one or two object pairs or replacing one or two objects, respectively. This created a further 4 conditions in which either one object was replaced (replace 1), two objects replaced (replace 2), 1 object swapped (swap 1) or two objects swapped (swap2) (see Figure 4.2).

	Intact	**Manipulated**			
		Swapped		**Replaced**	
Own	1x Own Intact	2x Swap 1	1x Swap 2	2x Replace 1	1x Replace 2
Other	1 x Other Intact	2x Swap 1	1x Swap 2	2x Replace 1	1x Replace 2
Control	1 x Control	2x Swap 1	1x Swap 2	2x Replace 1	1x Replace 2

Figure 4.2. Experimental conditions included in experiment 1. Renders were based on either participants' own drawings of typical living room and kitchen scenes, other people's drawings of these scenes, or copied control scenes, and displayed as either intact or manipulated. Two types of manipulations were applied to were scenes, either swapping or replacing their content. These manipulations were further divided into less (where 1 object was manipulated) and more (where 2 objects were manipulated) severe conditions.

Conditions with 1 manipulation (swap 1 and replace 1) were subsets of the conditions with 2 manipulations (swap 2 and replace 2). This meant that to create the 2-object manipulation image, we

took both single object manipulations and applied them to the same image, so that the 2-object manipulated image comprised of both. For example, if for one image in the first replace 1 manipulation we replaced a sofa with a coffee table, and in the second we replaced a TV with a bookshelf, the replace 2 condition for that image would have both the sofa replaced with the coffee table and the TV replaced with a book shelf (see Figure 4.3, and Appendix F for further examples).

When selecting which objects to manipulate we attempted to match the original object with an object of similar size and shape. For example, we would avoid replacing a bookcase (a large rectangular object) with a house plant (a smaller narrow object), and instead choose a more similar object, such as a large TV on a stand. We did this so the objects that were replaced would have similar coarse visual properties where possible. We also prioritised manipulating objects that appeared in a relatively central part of the scene and were clearly categorically consistent. For example, when replacing an object in a kitchen such as a stove, we would replace it with another object you might likely find in a kitchen (such as a dish washer).

We also aimed to match the orientation of the existing object as closely as possible, whilst still creating a plausible scene: avoiding placing objects in ways that would violate how they might be experienced in real life. This meant that some objects were rotated to fit into their new locations, or to maintain a viewpoint in which they could be identified. This was done to avoid producing any manipulated scenes that looked artificial, so that they still resembled a scene that could exist and would not stand out due to being overly artificial or implausible (such as placing a washing machine on a kitchen table.)

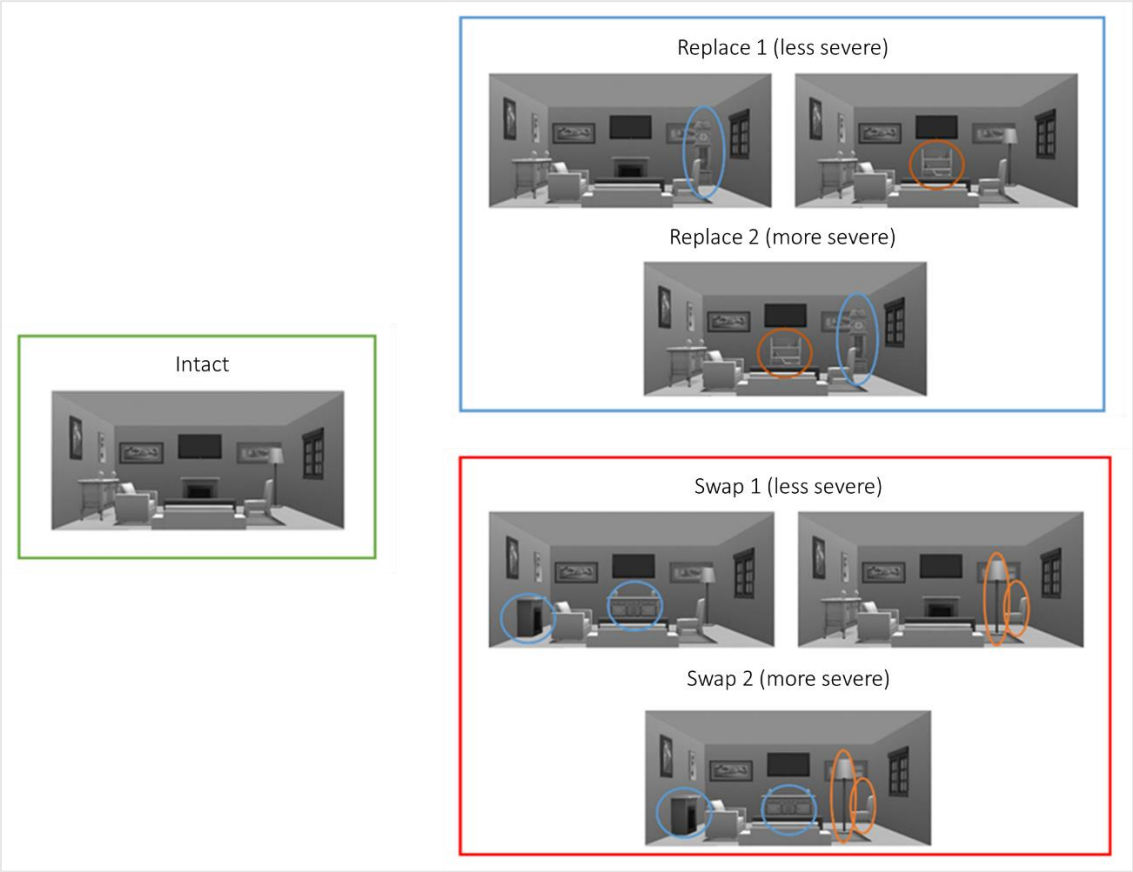


Figure 4.3. Examples of the manipulations applied to scenes in experiment 2. Participants were shown the intact scene (green box), 2 scenes where one object was replaced and 1 where 2 objects were replaced (blue box), and 2 scenes where one object was switched and one where 2 object were switched (red box).

Procedure

Scene drawing task

Participants first drew scenes from 2 categories (living rooms and kitchens) that matched their interpretation of the most typical exemplar of each scene type. The definition of typical was given as the most generic and ordinary example they could think of. They were also instructed not to draw a scene that they thought looked particularly interesting or attractive, nor an exact copy of a scene they knew from real life. They were given 1 minute to plan the scene, and then 3 minutes and 30 seconds to draw the scene. Scene sketches were drawn into a perspective grid, to allow participants to represent their scenes more easily in a 3D space, without having to employ any specific drawing techniques. It also standardised the participants viewpoint across all scenes. The perspective grids

used are described in more detail in the methods section of chapter 3 (see Appendix C for an example of the perspective grids used).

Participants were reminded how much time they had left at the halfway point, and when they had a minute remaining. They first drew a practise scene of a bedroom, to get them used to the timings and drawing on the perspective grid. The order in which they drew the other scenes was balanced across participants. After completing each drawing, participants also drew a rough birds-eye view of the scene, in which they labelled all the objects in the scene. This was done to help clarify the room's intended 3D layout and to confirm the identity of any ambiguously drawn objects, providing additional information for generating accurate 3D renders of the drawings.

In addition to drawing their most typical versions of living rooms and kitchens, participants drew copies of a given photograph of a living room and kitchen. These were the same photographs as used in chapter 3. These copies were drawn under the same time constraints as the typical drawings, and participants were instructed to capture a similar amount of detail as they used in their own drawings. They were given 1 minute to study the photo, followed by 3.5 minutes to sketch it, and had access to the photograph throughout their drawing time.

By employing the same drawing task between experiments, we were able to continue to use scene renders and participants from our previous experiment and compare the results of both with greater parity.

Scene Categorisation experiment

Participants completed a similar scene categorisation task as in chapter 3. Participants were asked to indicate whether a briefly presented scene was either a living room or a kitchen as accurately as they could. Before the experiment, they were instructed to full screen the web page and sit approximately 60 cm away from the screen. After reading the instructions, participants were shown 2 examples of each scene category (these were not included in the experiment).

During the experiment, participants viewed 3D renders based on either their own, the other 2 members of their group, or the control drawings of typical scenes. These were displayed as either intact or manipulated.

Within each group there were 8 intact scene renders; 2 based on a participants own typical living room and kitchen drawings, 4 based on those of other participants and 2 based on the control scenes. Each of these intact scenes were displayed in the 6 different manipulations (2x replace 1, 1x replace 2, 2x swap 1, and 1x swap 2), for 48 manipulated scenes total. Together, this yielded 12 conditions, comprised of 56 unique scene renders per group. Each scene was repeated 16 times, for a total of 896

trials. Trials were divided into 4 blocks of 224 trials, in which conditions were balanced, so that each stimulus was repeated 4 times. Blocks were separated by a 90 second break.

Stimuli were displayed on a grey screen. Trials began with a blank screen, followed by a central fixation cross for 1000ms. Next, the scene render was flashed for 83ms, followed by a mask presented for 150ms. Masks consisted of a random arrangement of squares, diamonds, and circles. For each trial one of 43 unique masks were randomly chosen. A blank screen was then displayed until the participants responded by either pressing "K" or "L" on their keyboard (to indicate whether a scene was a kitchen or living room). There was no time limit for participants to give their answer. After the participants gave their response there was a 100ms delay before the fixation cross was shown again and the next trial started.

We chose a brief stimulus presentation time and mask duration based on the hypothesis that rapid, coarse analysis is sufficient for individuals to activate their internal scene models, guiding categorisation decisions. This was supported by the findings of chapter 3, where we were able to find an influence of individual differences on scene perception with the same low display times. This approach further aligns with theories of prediction and prototype activation (Bar, 2004; Friston, 2005), suggesting that individuals generate predictions through a fast, initial processing of scene features. Previous research has shown that typical scene exemplars elicit more accurate detection and categorisation responses (Caddigan et al., 2017; Csathó et al., 2015; Torralbo et al., 2013), which supports the notion that brief exposures allow for the rapid engagement of internal representations. Thus, the short exposure time and mask duration were intended to investigate this rapid predictive processing mechanism whilst minimising the influence of post-perceptual factors (such as stimulus novelty or preference).

Statistical Analysis

To compare categorisation accuracies across conditions, we used repeated-measured ANOVAs and post-hoc paired-samples t-tests. We chose to conduct post hoc tests over priori contrasts despite having a specific directional hypothesis aimed at replicating previous findings. This decision was influenced by the implementation of a newly developed drawing paradigm, which in combination with the effect of manipulating scene content, introduced the potential for unforeseen interactions. Utilising post hoc tests allowed us to comprehensively explore all possible comparisons, thereby minimising the risk of selectively focusing on certain interactions and ensuring a thorough analysis of the data.

4.2.2 Results

Scenes based on participants' typical drawings are more accurately categorised than those based on other participants' drawings, but not of copied scenes

In order to test whether we could replicate the findings of the original scene drawing experiment (chapter 3), we first investigated whether scenes that were specifically tailored to participants' personal internal models are more accurately categorised. We used a one-way repeated measures ANOVA to compare categorisation accuracy between renders based on each participant's own drawing ("own" condition), other participants' drawings ("other" condition), or the copied scenes ("control" condition). The ANOVA had one factor (drawer), with 3 levels (own, other and control). The one-way repeated measures ANOVA violated the assumption of sphericity, as indicated by Mauchley's test, $\chi^2(2) = 3.611$, $p = 0.164$. Therefore, degrees of freedom were corrected using the Greenhouse-Geisser estimate ($\varepsilon = 0.91$).. As in our chapter 3, we found a significant differences in categorisation accuracy between drawers ($F(1.82,67.55) = 5.53$, $p = .007$), with a medium effect size of $\eta^2_p = 0.13$.

A Bonferroni-adjusted pairwise t-tests (with an adjusted significance threshold of $p < 0.017$ (0.05/3 comparisons) were then used to explore the results further by comparing categorisation accuracy between drawers Here we found that accuracy for the own condition was only significantly higher) than the other condition ($t(37) = 2.64$, $p = .012$, with Cohen's d of $d = 0.42$ indicating a small to medium effect size , but not the control condition ($t(37) = 0.16$, $p = .875$). We also found a significant difference between the other and control conditions ($t(37) = 3.29$, $p = .002$), with a medium effect size of $d=0.53$. This deviated from the original experiment, where accuracy was significantly greater for the own condition compared to both the other and control conditions, and there was no significant difference between the control and other conditions. As such, experiment 1 only provided a partial replication of the original scene drawing experiment.

The intensity of the manipulations to object content does not modulate scene categorisation for scenes based on participants' personal internal models

To investigate the effects of manipulating scene content on categorisation accuracy for the own, other and control scenes, we conducted two 3x3 repeated measures ANOVAs, with the factor's drawer (own, other, control) and manipulation (intact, replace and swap), separately for the less and more severe manipulations. To account for multiple comparisons, alpha values were adjusted using a Bonferroni correction, resulting in an adjusted significance threshold of $p < 0.025$ (0.05/2).

The 3x3 repeated measures ANOVA investigating the less severe manipulations violated the assumption of sphericity, as indicated by Mauchley's test, $\chi^2(9) = 13.973$, $p = 0.124$. Therefore, degrees of freedom were corrected using the Greenhouse-Geisser estimate ($\varepsilon = 0.85$).We found no significant interaction between manipulation and drawer for the less severe manipulations ($F(3.4, 126.06) = 0.89$, $p = .461$), indicating that the less severe manipulations to scene content did not disrupt scene categorisation differently regardless of which drawing a render was based on. We found a main effect of drawer ($F(1.73, 64) = 6.62$, $p = .004$), with a large effect size of $\eta^2_p = 0.15$, driven by the difference between the own and control condition against the other condition, as reported above.

For the more severe manipulations, we again failed to find a significant interaction effect between drawer and manipulation ($F(3.12, 115.74) = 2.17$, $p = .093$). Similar to the less severe manipulations, we also found a significant main effect of drawer ($F(1.65, 61.23) = 3.69$, $p = .038$), with a medium effect size of $\eta^2_p = 0.09$,.

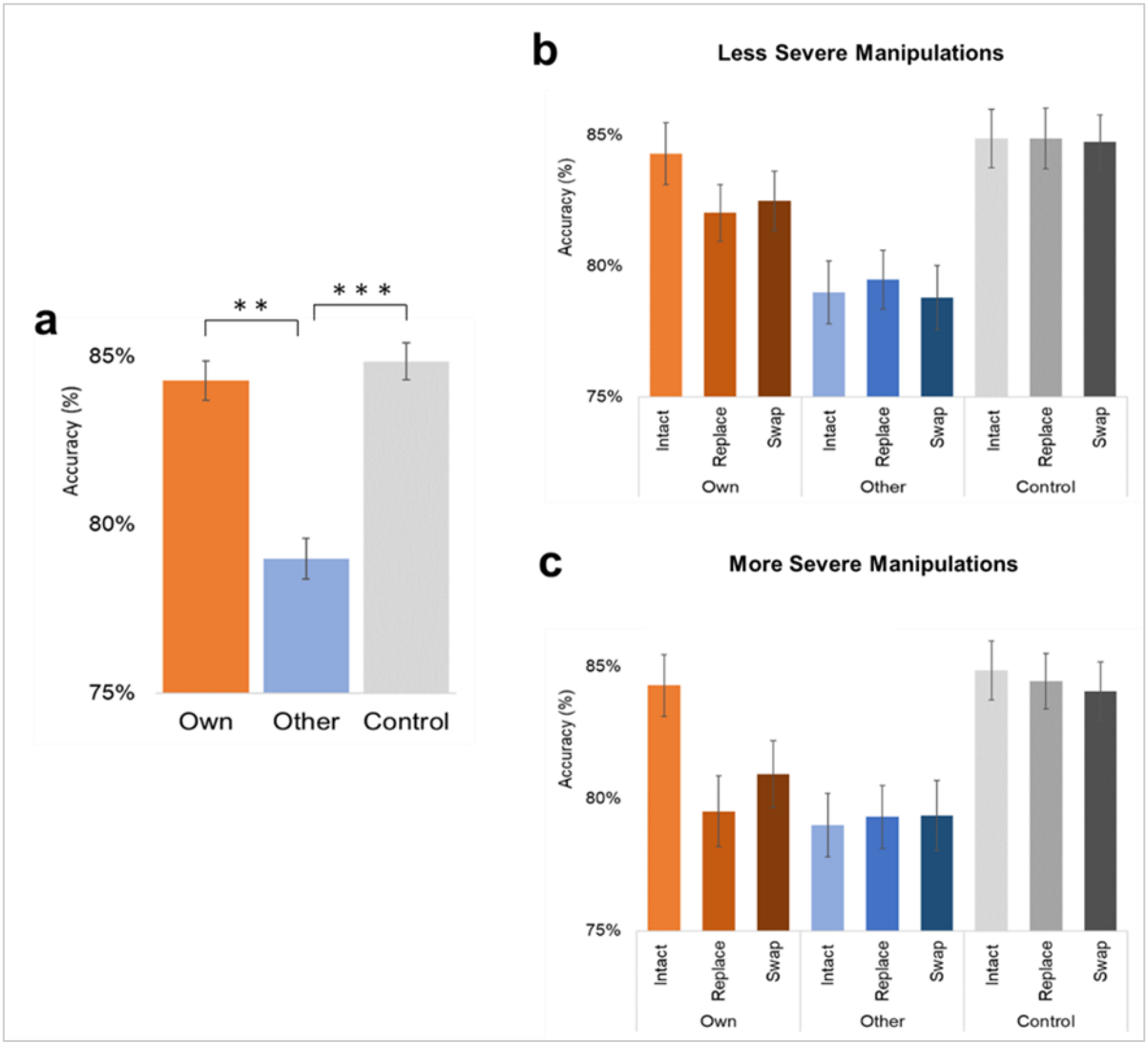


Figure 4.4. a). In order to test whether experiment 1 replicated our original drawing experiment, we compared mean accuracies for the intact own, other and control condition in the scene categorisation task. Here, we found that whilst participants were significantly more accuracte at classifying scenes based on their own renders compared to those produced by others, they were not more so than those in the control condition. b) The mean cateogrisation accuracies for the less severe manipulations (where only a single object was either swapped or replaced) compared to the intact conditions between drawers (own, other and control). Here, we found no effect of manipulating scene content. c) The mean cateogrisation accuracies for the more severe manipulations (where 2 objects were either swapped or manipulated) compared to the intact conditions between drawers (own, other and

control).Again, we found no signficnat effect of manipualting scene content. ** indicates $p < .01$, *** indicates $p < .001$.

4.2.3 Summary

In experiment 1 we found no significant interaction effect between drawer and manipulation for both the less or more severe conditions. However, observing the trends in the data, the effect of manipulation was greater where participants categorised renders based on their own scenes, compared to the other and control conditions. Additionally, the greatest impact was caused by replacing objects which could suggest that internal models contain more information about object identities typically found in a scene than these objects' locations within the scene. Whilst these differences were not statistically significantly different, and so do not be represent real effects, they may be indicative of a possible trend that the current experiment was unable to detect.

By including both less and more severe manipulations in the same task we may have limited our ability to detect a stronger effect from the more severe manipulations. The less severe manipulations were a subset of the more severe manipulations. While the less severe manipulations may not have been strong enough to be detected during the brief display time for the images, they may have acted to adjust participants ability to judge the scene, by providing a middle point between the intact and more severe manipulation conditions, limiting the overall effect. Due to the low stimulus variety, participants may have passively learnt to recognise the images throughout the experiment, mitigating some of the disruption from the manipulations. As such, in experiment 2 we proceeded with only testing the more severe manipulations, to investigate this effect more directly.

In experiment 1 we produced a partial replication of our drawing experiment in chapter 3, finding participants were more efficient at scenes based on their own renders compared to those of others. However, unlike in chapter 3, we found no difference between a participant's own renders and those of the control condition. Furthermore, participants were also better at categorising the control condition than those produced by others. This could suggest that the observed effect is a result of familiarity with the stimulus, contrary to the findings of chapter 3. In chapter 3, a far greater variety of unique scenes were shown (45 compared to 4 in the current experiment), which could have made it more difficult for the scenes to be recognised, especially considering the rapid display time. However, in the current experiment, if one was to consider the manipulated scenes as distinctive from the intact originals, this experiment would have a greater variety of unique stimuli (at 56 unique scenes). It is difficult to determine the extent to which these scenes could be considered truly unique, especially in comparison to the variety found in the original experiment.

Alternatively, the difference could be attributed to the content of the control condition. As the control condition was constructed from a copy of a real-world photograph, the level of detail produced in the images was generally greater than those produced by participants in their own sketches. This resulted in the control scene containing a greater number of objects than the drawings produced by the participants, and subsequently both the own and other scenes. This may have resulted in the scene containing more scene information that the visual system could use to categorise the image more efficiently. Furthermore, this could affect the manipulations applied to the control scene, as the changes in content may have been harder to detect within the greater scene clutter. This effect might have only become apparent in the current experiment again due to the lower variety of scenes used for each participant.

4.3 Experiment 2

The aim of experiment 2 was to confirm the results of experiment 1. Experiment 2 utilised the same methodology as experiment 1 with two changes. First, only the more severe conditions were used for the manipulated scenes (replace 2 and swap 2), as nominal trends in the data in experiment 1 suggested only the more severe manipulations may be able to disrupt scene categorisation. Further, by reducing the number of manipulations, we are able to increase the number of repetitions for each image, increasing the reliability of the categorisation data. Second, we reduced the number of objects used to design the control renders. This was done for two reasons: to make the overall number of objects within the scene more comparable to those in the own and other conditions, and to reduce visual clutter, so that the effect of manipulations might be more easily detectable (in-line with the own and other scenes).

4.3.1 Methods

Participants

As in experiment 1, we aimed to achieve a similar sample size as used in chapter 3. In total 33 participants took part in the experiment. This was 2 less than took part in chapter 3, and was a result of participants not returning to take part in the scene categorisation experiment after they had completed the drawing session. 17 participants identified as male and 16 identified as female, with an average age of 28.12 years ± SD. 9.10. Of these all participants had also taken part in experiment 1, and 13 in our original scene drawing experiment (as outlined in chapter 3). All participants reported having normal or corrected to normal vision and were English speaking.

Stimuli

The stimuli in experiment 2 comprised of scene renders constructed using the same method as experiment 1. Both intact and manipulated scene renders were used, but only the more severe manipulation conditions (replace 2 and swapped 2) were included. This resulted in 9 conditions (see Figure 4.5).

	Intact	**Manipulated**	
		Swapped	**Replaced**
Own	1x Own Intact	1x Swap 2	1x Replace 2
Other	1 x Other Intact	1x Swap 2	1x Replace 2
Control	1 x Control	1x Swap 2	1x Replace 2

Figure 4.5. Experimental conditions included in experiment 2. As in experiment 1 renders were based on either participants' own drawings of typical living room and kitchen scenes, other people's drawings of these scenes, or copied control scenes, and displayed as either intact or manipulated (with constituent objects either being swapped or replaced). However, unlike in experiment 1 only more severe manipulations were utilised.

The control condition for experiment 2 was also remade, so that it was more comparable to the scene renders based on participants' typical drawings. This was done so that manipulations to the control scene caused a similar degree of change as to those in the own and other conditions, and thus make the condition more comparable. In the new control renders we retained the main typifying objects, such as large pieces of furniture, but removed smaller objects that added a greater level of detail and clutter than in other scene renders (such as many of the decorative vases found in the control living room scene). We also changed the 3d models used to depict some furniture items (such as the sofa, chair, and coffee table) to models utilised in other scene renders. This was because most scene renders shared 3D models for these objects, but the control scene utilised unique models not found in other renders. By changing these to more commonly used model, the control scene became more analogous with the renders used in other conditions, whilst remaining a unique scene exemplar and a strong representation of the control image (see Figure 4.6).

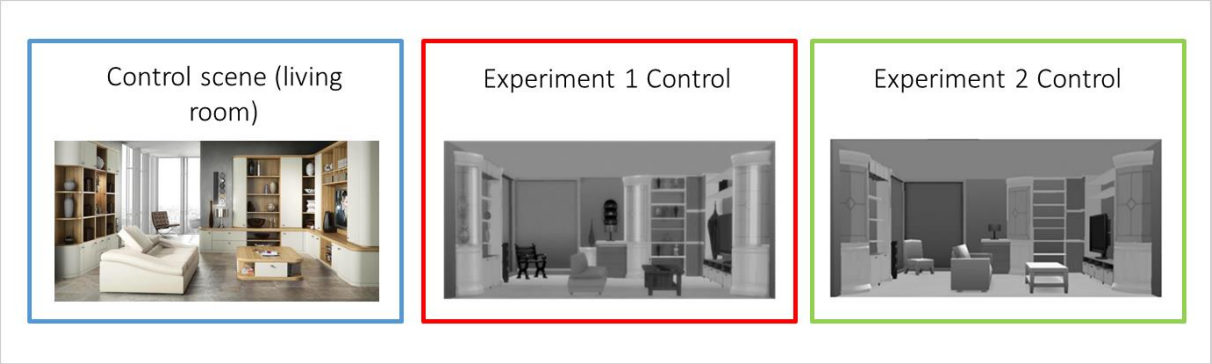


Figure 4.6. The control condition used in experiment 1 compared with the control condition used in experiment 2. Many smaller decorative objects (such as vases) have been removed to reduce the visual clutter and create a scene with a comparable amount of individual objects and visual clutter as the scene renders produced by participants.

Procedure

Drawing Task

As all participants had participated in drawing sessions before, no new drawing sessions were conducted for this experiment.

Scene categorisation experiment

The same scene categorisation experiment was used as in experiment 1, with all instructions and experimental parameters being identical. As in experiment 1 participants were grouped into batches of 3, viewing the same stimulus sets as each other. Participants remained in the same batch as they were allocated in experiment 1. The only difference in procedure for experiment 2 was that only the more severe manipulations were utilised (replace 2 and swapped 2) alongside the intact scene renders.

Within each group there were 8 intact scene renders; 2 based on a participants own typical scene drawings (1 of a living room and 1 of a kitchen scene), 4 based on those of other participants and 2 based on the control scenes. 2 manipulations (replace 2 and swap 2) were applied to each of these scenes, totalling 16 manipulated scenes. This yielded 24 unique scene renders per group. Each trial was repeated 30 times, for a total of 720 trials. These trials were divided into 4 blocks, with 8 repetitions being included in blocks 1, 2 and 3 (consisting of 192 trials) and 6 repetitions in block 4 (consisting of 144). Block 4 contained less trials as a result of dividing the total trial numbers across the four blocks (to match the previous experiment). Each block was separated by a 90 second break, in order to allow participants to rest their eyes and reduce any effects of after images.

4.3.2 Results

Scenes based on participants' personal internal models were more accurately categorised than those based on other participants' internal models, but not of copied scenes.

We conducted the same analysis as used in experiment 1 in order to investigate whether scenes that were specifically tailored to participants' personal internal models are more accurately categorised, in order to establish whether we had replicated the findings of chapter 3.

We compared categorisation accuracy between renders based on each participant's own drawing ("own" condition), other participants' drawings ("other" condition), or the copied scenes ("control" condition) using a one-way ANOVA. The ANOVA had one factor (drawer), with 3 levels (own, other and control). The one-way repeated measures ANOVA violated the assumption of sphericity, as indicated by Mauchley's test, $\chi^2(2) = 0.29$, $p = 0.865$. Therefore, degrees of freedom were corrected using the Greenhouse-Geisser estimate ($\varepsilon = 0.99$).

As in our chapter 3, we found a significant differences in categorisation accuracy between drawers ($F(1.98,63.41)= 3.40$, $p= .040$), with a medium effect size of $\eta^2_p = 0.09$. Bonferroni-adjusted pairwise t-tests (with an adjusted significance threshold of $p < 0.017$ (0.05/3 comparisons) were then used to explore the results further by comparing categorisation accuracy between drawers. We found that renders in the own condition were more accurately categorised than those in the other condition ($t(32) = 2.54$, $p = .016$) (see fig 11.a). The effect size, as measured by Cohen's d, was d = 0.44, indicating a small to medium effect size. As in experiment 1, we found no significant difference between own and control scenes ($t(32) = 1.29$, $p = .208$). However, unlike experiment 1 there was also no significant difference found between the *other* and *control* conditions ($t(32) = 1.36$, $p = .184$). Overall, the pattern qualitatively matches the pattern observed in chapter 3 better than experiment 1, although the effects did not replicate fully.

Manipulations to scene content may not disrupt scene categorisation for scenes based on participants' personal internal models

In order to investigate the effect of manipulating the scene content, we compared categorisation accuracy between the own, other and controls conditions, when they were displayed as either intact, or manipulated by either replacing 2 objects with new ones (the replace condition) or swapping the location of 2 sets of objects (the swap condition).

We conducted a 3x3 repeated measures ANOVA, with 2 factors (drawer and manipulation) each with 3 levels (drawer: own, other, control; manipulation: intact, replace, swap.) The interaction effect

between drawer and manipulation met the assumption of sphericity, as indicated by Mauchley's test, $\chi^2(9) = 19.1$, $p = 0.025$.

We found no significant interaction between drawer and manipulation ($F(3.12, 99.83) = 0.96$, $p = .416$), indicating no significant differences in the effect of the manipulation across drawers (see fig 11.b). Although the trends in the data were similar to experiment 1, this effect was not supported by a significant interaction effect, as initially hypothesized.

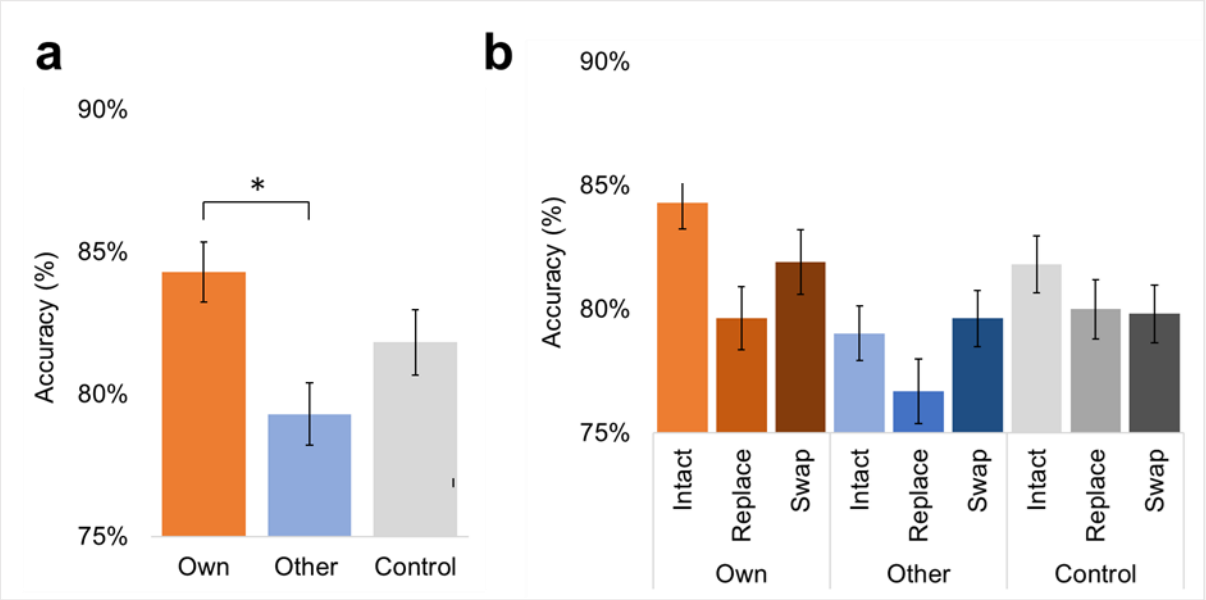


Figure 4.7. a). As in experiment 1 we tested whether our results replicated our original drawing experiment, by comparing the mean accuracies for the intact own, other and control condition in the scene categorisation task. As in experiment 1 we found that participants were significantly more accuracte at classifying scenes based on their own renders compared to those produced by others, but no longer found that accuracy was significantly higher in the control condition. b) The mean cateogrisation accuracies for the intact, swapped or replaced conditions between drawers (own, other and control).. * Indicates $p < .05$, ** indicates $p < .01$.

4.3.3 Summary

Experiment 2 confirmed the same effect of scene manipulation as observed in experiment 1. We again found no interaction effect between manipulation and drawer. When examining the trends in the data we again observed little effect of switching object locations within these same scenes. Whilst this may tentatively indicate the identity of objects within internal models are more important than their position, due to the lack of interaction effect we cannot make this assumption from the current study.

Instead, based on our results we conclude that manipulating scene content had no impact on scene categorisation, regardless of which drawing the renders were based on.

As in experiment 1 we only produced a partial replication of the findings of our original scene study (see chapter 3), with the own condition only yielding significantly higher categorisation accuracy than the other condition. However, unlike in experiment 1 the control condition did not yield significantly higher performance than the other condition. Whilst this does not allow us to disregard familiarity with the scene as an influencing factor in the increased scene categorisation accuracy observed for the scenes based of participants own drawing, it does suggest that familiarity is not the sole factor driving the increased accuracy.

Reducing the visual clutter in the control scene lowered categorisation accuracy (M = 81.83%, SE = 2.3%), compared to experiment 1 (M = 84%, SE = 2.5%). This may suggest that the increased categorisation accuracy for the control scenes observed in experiment 1 was the result of the scene containing distinctly more objects than the scenes used in the own and other conditions. However, this same control scene was used in chapter 3, but scene categorisation accuracy was lower, and importantly lower than in the own condition. This may further evidence that the difference in the experimental design used in current experiments was responsible for our inability to replicate our original findings.

In this context, however, it is worth noting that the difference in categorisation performance between the own and control conditions also emerged in another independent replication (experiment 2; Wang et al., 2024), in which the stimuli were also shown in batches. The reasons for the discrepancies observed in these experiments are currently unclear, although they may relate to a difference in online and laboratory testing (as in experiment 2, Wang et al., 2024).

4.4 General Discussion

Together, across both experiments, we found no significant difference between manipulating scene content regardless of whether the scene was based on a participant's own judgement of typicality or not. If semantic and syntactic information were represented in internal models for scenes, we would expect to have observed a greater disruption of categorisation in the scenes approximating a participant's internal model, as the strength of the match is reduced when introducing object manipulations. Scenes based on the control and other participants' scene drawings would already be less of a match to the internal model, so further manipulations would not necessarily make them less of a match to the participant's own model (the manipulations may even bring them closer to the participant's internal model). Furthermore, we were unable to fully replicate the findings of our

original drawing experiment, as although we found that accuracy was higher for scenes based on participants own renders when compared to those based on others, accuracy was not higher than our control condition in either experiment, suggesting a possible effect of familiarity. Below, we will first provide an interpretation of the current results, before moving on to discuss how methodological weaknesses and potential flaws in our earlier assumptions based on the findings of chapter 3, might provide possible explanations for the discrepancies with our expected results.

Across both experiments we were only able to partially replicate the idiosyncratic effect on categorisation accuracy observed in chapter 3. Whilst we found that drawings based on participants own scenes were more efficiently categorised compared to those produced by others, performance was not significantly greater than the control condition. In addition, in experiment 1 we also found scene categorisation was also significantly better for control scenes compared to those produced by other participants. This would suggest that the improved categorisation accuracy for renders based on participants own scene drawings was a result of familiarity with the stimulus, as both the participant's own scenes and controls were produced at the same time. Previous research has found that participants are better at categorising familiar scenes (Bainbridge, 2022; Epstein et al., 2007; Klink et al., 2023), which may have led to the improved accuracy we observed. If familiarity can explain these effects, this may suggest that the lack of object manipulation we observed could have resulted from render's being categorised based on familiarity, and this effect masking any effects of matching scenes to internal models.

This further makes it difficult to infer what the impacts of manipulating object content in the current study can tell us about their representation in internal scene models. The findings of the current study conflict with those found in chapter 3, which suggested that constituent scene objects and their placements were the main drivers of individual differences we observed. Across our drawing experiments, the differences between the renders based on participants' own internal models and those based on other participants were primarily resulted from differences in the objects included and their locations, with other factors such as viewpoint, colour and size of the scene being controlled for. The identity of individual objects was also somewhat controlled for, as participant drawings often did not contain much detail for single objects, and due to limitations of the software used to produce the renders, often the closest match to a limited set of pre-existing models was used, meaning similar commonly occurring objects (such as ovens, chairs and TVs) were used across renders. As such, our original drawing study would suggest that it was differences in the object type and location that produced the improved categorisation accuracy we observed.

Subsequently, if the lack of significant disruption caused by manipulating object content reflects a similar process within internal models, than our results would challenges the idea that idiosyncratic object and location-based information are represented within internal scene models. Instead, they would suggest that representations of typical object information is represented similarly across participants. Whilst we manipulated object content, we aimed to do this in such a way that the manipulated scenes still represented a plausible exemplar of that category. As such, no scene type would be any less typical than another for its category, meaning that shared representations of scene regularities may not have been violated in the manipulated scene conditions, which could account for why we did not observe any effect of object manipulation. Such an interpretation is coherent with previous research that has shown beneficial effects of shared conceptions of typical object identities and placements across participants (Davenport & Potter, 2004; Faivre et al., 2019; Mudrik et al., 2010, 2011; Võ & Wolfe, 2013; Kaiser et al., 2018; Kaiser & Cichy, 2018a, 2018b).

This notion would align with the idea that internal models provide coarse representations of scene information that would match their expected use in natural vision (Brandman & Peelen, 2019). For internal models to be able to facilitate rapid scene processing, they would need to be applicable to a broad range of scenes. If minor deviations in object identity or placement could meaningfully disrupt this process, then we would expect internal models to only be useful at recognising a narrow range of scenes based closely around the contents of the internal model itself. This could lead to us having very narrow and individualised conceptions of scene categories. Instead, what we see is the opposite – people have little difficulty identifying a broad range of different arrangements of object types as an example of a given scene type. For example, living rooms can vary in design, with a broad range of possible constituent objects, but they are still easily identifiable as a living room.

However, whilst the current results may suggest that internal scene models do not contain idiosyncratic representations of scene specific semantic and syntactic object-level information, it is possible that they may instead contain a relatively coarse, or broad, representation of these features. As mentioned, in the current study the manipulations we applied still produced plausible exemplars of their scene category, thus representing relatively subtle changes to the scenes semantic and syntactic information. If internal models contain a coarser representation of this information, such manipulations may not have been sufficient at disrupting the match to the internal models. In the current experiment, we only changed one or two objects at a time, and only either the semantic or syntactic information separately. Not only could these manipulations not have been severe enough to disrupt the match to the contents of the internal model, but it may be the remaining intact source of object information was sufficient to successfully facilitate the match. For example, when the syntactic information was disrupted, the intact semantic information could have been enough to facilitate the

match to the internal model, and vice versa. Previous studies have found that object information is most efficient at facilitating scene processing when both semantic and syntactic object information is intact, supporting the notion that internal models might contain a more holistic representation of object information (Kaiser et al., 2014, 2015). In order to explore this further, future studies could utilise parametric manipulations to both semantic and syntactic scene information simultaneously, in order to disrupt both concurrently and thereby disrupt any possible integration. Whilst proximity to an individual level internal model may help facilitate this match, the need for robust internal models may make it difficult to detect this benefit by only manipulating one or two objects within a scene.

Alternatively, the current study may have lacked the experimental power to properly detect the influence of manipulating object content of scene categorisation. Whilst we did not observes any significant effects of object manipulation on categorisation accuracy, across both experiments we did observe a nominal trend in the data showing a greater difference between the intact and manipulated scenes based on participants own drawings, compared to those drawn by other participants and the control. This difference is particularly apparent in experiment 1, where the difference between the intact and manipulated scenes produced by other participants and the control was less than 1% accuracy, whilst the difference between the intact and manipulated scenes for the own condition was 4% for the more serve manipulations. This could suggest that future studies employing an experimental design more sensitive to object level information may observe an effect of manipulating object content in future studies. Whilst the data of the current experiments do not provide evidence for this relationship, this inference may be a useful indicator for future experiments seeking to clarify the representation of semantic and syntactic information within internal models. We will discuss how future studies could be improved to better investigate these effects in the sections below, and highlight what can be learnt from the limitations of the current study.

Taken together, the results of experiment 1 and 2 failed to meet our expected pattern of results. Whilst nominal trends in the data partially aligned with our expectations, as discussed above, we found no evidence of the hypothesised effect of manipulating object identity and location within renders based on a participant's own drawings, nor that the severity of these manipulations modulated these effects. Further, we were unable to fully replicate the findings of chapter 3, suggesting that the current results might be better explained by the influence of familiarity with the scenes, rather than being driven by a match to a participants own internal models. Alternatively, these results may be better explained by several methodological weaknesses inherit to the current study, that may have considerably influenced our findings. Next, we will discuss these limitations, and how they might have affected our results.

4.4.1 Challenges in Detecting Effects of Semantic and Syntactic Manipulations

If internal scene models are constructed from broader-level scene information, it may be that our current measure was not fine enough to detect the effect the relatively minor disruptions to semantic and syntactic information had on scene processing. In both our current and original study, we found that scene categorisation performance was high for all scenes, indicating the relative ease of the task. The ability to categorise scenes quickly and accurately has been long established, even when scene information is disrupted or the image blurred (Joubert et al., 2007; Li et al., 2002; Trouilloud et al., 2020; Wiesmann & Võ, 2022), highlighting the relative ease of the task. Further, in chapter 3 we found that the differences by idiosyncratic scene representations were only very slight, further suggesting that these effects may be fragile, and easily masked by competing influences on scene categorisation.

As such it may be necessary to apply more severe manipulations to object content to detect an idiosyncratic effect. In the current study, our manipulations still produced plausible exemplars of scenes from their category, with much of the scene information remaining intact, and thus potentially matching participants own internal models. In previous experiments that have shown effects for semantically typical objects on scene processing, the manipulations applied often disrupt the category level information of that object (Davenport & Potter, 2004; Faivre et al., 2019; Mudrik et al., 2010, 2011; Võ & Wolfe, 2013), creating stronger disruption to the scenes semantic content. As such, it may be valuable for future research to try and establish whether similarly stark category level semantic disruptions differentially impact scenes approximating participants internal scene models. If such disruptions effect processing more for scenes based on participants own internal models, than this could suggest that semantic information is represented more holistically. Similarly, it may be that greater manipulations may also need to be applied to syntactic object information. Whilst this might be achievable by simply manipulating more objects, it may be more pertinent to consider the relationships between objects. Whilst research has shown that typical positioning for single objects can help scene processing (Kaiser et al., 2018; Kaiser & Cichy, 2018a, 2018b), there is growing evidence to suggest that the relative locations of objects to each other may have an even greater impact (Bilalić et al., 2019; Kaiser et al., 2014; Kaiser & Peelen, 2018; Stein & Peelen, 2015; Peelen et al, 2024). Recent research has demonstrated that clusters of objects forming typical arrangements provide a strong representation of scene information. If semantic scene information is more holistically represented in internal modes, it may be these relative relationships that are more important than the placement of individual objects, as explored in the current study.

However, in situations where participants produce drawings with few constituent objects, it may be difficult to manipulate syntactic object information further. Such scenarios were common in the current experiment and may represent a potential limitation of the drawing method. It may be

possible that providing participants with more time to draw their scenes, or an alternative means of constructing them (such as the use of simple computer aided design program software) would encourage participants to include more objects that could be manipulated.

A further potential limitation of our current study was the use of experimenter judgements when deciding which objects were manipulated. Objects were selected on a preestablished set of criteria, which featured manipulating large, scene-defining objects as a priority. This was based on previous research that found when constructing representations of different scene categories participants consistently placed larger objects (such a tables, counters, baths) first, before placing smaller objects around them (Boettcher et al., 2018). These larger objects are hypothesised to act as anchors, and key points of reference for other objects within a scene, and as such most likely to cause disruptions in schema matching. However, the application of these rules was based on the experimenter's assumption about what objects would be most meaningful to change. In some instances where such objects were limited, this choice would be more obvious, but when multiple candidate objects are available, it relies on more subjective experimenter judgements. Which object the experimenter chose to manipulate may have been less meaningful to the drawer, and thus failed to consider the individual differences the experimenters aimed to investigate.

Instead, the experiment could have been improved by collecting the participants judgement of which objects they judged as being the most diagnostic for their own scene drawings. This could have been achieved by asking participants to rank the importance of objects in the scene, and subsequently manipulating the objects rated the highest. However, asking for participant judgements may have introduced additional confounding variables associated with self-report methods. For example, participants may have chosen a prominent scene feature they felt would match the experimenters' expectations, rather than their own. Such demand characteristics would be particularly undesirable, as they would diminish the individual differences trying to be captured. Alternatively, drawing order could have been used to establish which objects were more prominently represented within internal models. However, whilst the order in which objects are drawn has been suggested to be diagnostic of how well they were recalled during experiments investigating scene memory (Bainbridge, 2022; Bainbridge et al., 2019), it is unclear whether this would translate to how strongly they are represented in internal scene models. Kinematic drawing data can be achieved on drawing tablets, where pen strokes can be monitored quickly and accurately, so the order of the objects was drawn in could be more easily collected. This was, unfortunately, not easily possible, as our experiments were conducted online.

Our failure to find an effect of manipulating scene content may have resulted from our choice of experimental paradigm. Here, we utilised a rapid scene categorisation task, with brief stimulus presentation times, as we hypothesised that a rapid coarse analysis would be sufficient for individuals to activate their internal scene models. This decision was based on the findings of chapter 3 that utilised a similar categorisation task, and previous studies suggesting that typical scene content helps facilitate rapid scene categorisation (Caddigan et al., 2017; Csathó et al., 2015; Torralbo et al., 2013). However, our inability to detect an effect of manipulating scene content on categorisation may not necessarily indicate that object-related information is absent from internal scene models completely. Instead, it might suggest that object information has less relevance for rapid scene categorisation, where course matching could be applied without the need to identify object information.

As such, in order to explore whether object related information is stored within internal scene models, it may be necessary to utilise a task where object level information is more directly behaviourally relevant. This could be achieved through the use of visual search tasks, where participants are required to actively scan a scene for a specific target (such as an object). Visual search has successfully been used to demonstrate the role of typical semantic and syntactic scene information for guiding search in natural scenes (Biederman et al., 1973; Malcolm & Henderson, 2010; Neider & Zelinsky, 2006). If individual differences are present within these conceptions of typicality, then we might expect that manipulating scene content to adhere to, or violate, these individual conceptions to modulate search efficiency, in a similar manner as hypothesised in the current study. Such an experiment could consist of asking participants to locate a category neutral object (such as luggage) within scenes based on their own and others renders, when the object location and identity are manipulated. This approach could help clarify the role of individual differences in object-related information in scene perception, and is a clear direction for future research.

However, combining visual search with the current drawing paradigm may make it difficult to isolate the mechanisms responsible for the preference for renders based on participants own scenes compared to those based on others' drawings. In the current study we were unable to fully replicate the findings of chapter 3, particularly the categorisation difference between participants own and the control scenes, which may suggest that familiarity with the render could drive the observed effect. Whilst this discrepancy may result from methodological differences (as discussed below), it remains an important question to understand whether line drawings truly reflect the content of internal scenes, or rather represent a closer proximity to specific scenes participants are familiar with. As visual search is theorised to rely on a combination of regularities in a scenes content and

structure, as well memory for specific scenes (Wolfe, Võ, et al., 2011), it may be difficult to attribute which mechanism drives this effect, or in what combination, and to further ascertain how much of the effect may be driven by familiarity. Whilst this does not discount the utility of combing visual search with the line drawing method, it highlights further the need to clarify the role of familiarity in driving this effect before further exploration can be conducted.

Additionally, the use of rapid presentation times may have been particularly inappropriate for a study conducted online. Due to differences in hardware and the stability of participants internet connections, this may have caused issues with displaying the stimuli that could have resulted in participants having less exposure to the stimuli, or some scenes not being displayed at all. Studies comparing the effects of these factors on stimulus display found that when using Gorilla (the software used to host our experiment) stimuli could be displayed for approximately 10ms longer or shorter than expected, even when operating at optimal conditions, and that these effects were further compounded by even greater delays caused by differences between operating systems and browsers (Bridges et al., 2020). Given our short stimulus display time of only 83ms, such variations could have altered stimulus presentation times considerably. Any impact of stimuli not being displayed may have further been exasperated by the requirement for participants to provide an answer before trials continued. This may have resulted in participants guessing in instances where stimuli were not appropriately displayed.

Even if all stimuli were displayed correctly, the online setting may have impacted participant performance. Crump et al. (2013) found that participants were more likely to lose attention, and begin utilising inefficient strategies for online experiments that were long and repetitive, which could fairly describe the scene categorisation experiments used in the current study. Such strategies may have added further extraneous variability to our data, diminishing differences between conditions. As the expected results only represented relatively small differences in accuracy, and in combination with the high base accuracy for the task (regardless of condition), this may have had a considerable effect on the current data set., As such, in order to help clarify the results of the current study, it may be necessary to conduct a replication under laboratory conditions, in order to discount these potential extraneous technical and behavioural variables caused by conducting the experiment online.

4.4.2 The role of visual clutter and familiarity on scene categorisation

The discrepancy between the current results and those of chapter 3 were surprising, as we had expected to find a similar categorisation preference for participants own scenes compared to the

control. The failure to replicate this result challenged the findings of our original drawing experiment, and suggested a greater influence of familiarity on the effect we observed.

To explore this discrepancy, we investigated if the higher levels of visual clutter found in the control scene may contribute to this discrepancy. Visual clutter has been theorised to be an integral part of scene complexity, alongside other scene characteristics such as openness and object organisation (Kyle-Davidson et al., 2022; Olivia et al., 2004). When reducing the visual clutter from the control scene, we primarily removed smaller objects that were replicated many times. For the living room scene, this primarily involved removing small decorative statues found on the bookshelves, whilst in the kitchen, this involved removing plates from cupboards and other small decorative objects. These objects were not frequently represented in drawings produced by participants and added a greater level of comparative detail to the control scenes. Other aspects associated with scene complexity were preserved; with no changes to the layout of more diagnostic constituent objects (such as chairs, tables and other furniture) preserving object organisation and openness.

Experiments using 3D virtual environments have found that after training visual clutter can improve performance on tasks such as visual search, navigation, and spatial judgements (Bacim et al., 2013; Handali et al., 2021; Ragan et al., 2015). Meijer (et al., 2009) found that navigation was improved in virtual environments that included a greater degree of detail for constituent objects, finding a preference for navigating a virtual supermarket environment when shelves contained detailed models of products. This is similar to the type of visual clutter that was removed from the control scenes in experiment 2, with both representing smaller objects incorporated into a larger scene feature. Whilst these experiments utilise different tasks and drastically different display times, it may suggest that visual clutter improves our ability to process scene information by increasing the realism of a scene and strengthening its representation. As such, the additional clutter in the control scene in experiment 1 may have aided participants in developing and learning more efficient strategies for recognising the control scenes, compared to the less detailed scenes based on participant drawings. However, comparison with studies using 3D virtual environments should only be made tentatively, as the exposure to the scene information differed from those of the current study. During the current study, participants were only briefly exposed to the scene information, whilst in the virtual reality environments they were able to the freely explore the scene for several minutes, which may modulate the role of visual clutter.

As discussed, the brief exposure to the stimuli in the current study may have only allowed participants access to the gist of the scene (Oliva & Torralba, 2001). Here, the role of visual clutter is less clear. Clutter has been shown to be an essential element in judging scene complexity (Kyle-Davidson et al.,

2023; Oliva & Torralba, 2001), however it is less clear how this would impact scene categorisation. It might be expected that additional clutter may make recognition of individual scene objects more difficult (Edquist & Johnston, 2008; Meyers et al., 2010), impacting scene recognition by decreasing the diagnostic ability of constituent objects. By decreasing the clutter, and subsequently the complexity, we may have also impacted how memorable the control scenes were. Research has shown a complicated relationship between scene complexity and memorability, with both high and low level complexity aiding in memorability, whilst medium-level complexity scenes are more forgettable (Kyle-Davidson & Evans, 2023). If the increased clutter in the controls used for experiment 1 helped make the scenes more memorable, this may have further aided participants in developing more efficient strategies in the categorisation task not available for the renders based on participant drawings.

Alternatively, the higher levels of visual clutter may have changed the scenes' low-level scene characteristics, by causing them to appear as more dense texture patches in the brief exposure time. This could have aided in participants learning more efficient strategies for classifying the control scenes that did not rely on recognising its categorical information. For example, participants may have initially identified the scene in earlier trials, and through the course of the experiment learnt to classify it more quickly based on its more unique low-level features. To investigate this possibility, future experiments could compare participants' ability to categorise the more or less cluttered control scenes utilising scene inversion. Scene inversion has been shown to disrupt low level visual information whilst disrupting the semantic meaning of a scene. If the difference in scene categorisation observed between the cluttered and uncluttered control scene is the result of changes in low-level visual information, we would expect inversion to diminish this advantage, whilst if it was the result of the object information found in the scene, we would not expect inversion to impact scene categorisation to diminish this advantage.

Alternatively, the failure to replicate the results of chapter 3 may have been caused by differences in the experimental design. In chapter 3, participants were exposed to a greater variety of individual scenes, as they were shown renders based on drawings produced by all other participants (88 individual intact scenes) whilst in the current study they were only shown 8 intact scenes. Further, these stimuli were repeated more time in the current experiment, with 16 repeats in experiment 1 and 30 in experiment 2 compared to only 10 in chapter 3. The lower stimulus variety and greater number of repeats may have made it easier for participants to develop more efficient strategies to recognise the individual stimuli, and this learning effect could have diminished the effect size and weakened the experimental power. The higher accuracies in both experiment 1 (M = 82.33%) and 2 (M = 81.72%) compared to chapter 3 (M = 80.6%) may support this conclusion, although these differences are admittedly only very slight.

This issue may have been further compounded by the inclusion of the same participants in both experiment 1 and 2, as well as returning participants from our original drawing study in chapter 3. As participants would see the same stimuli for the own and other condition in both experiments, and only a slightly modified control stimuli (with elements believed to add extraneous visual clutter being removed in experiment 2), this may have resulted a greater familiarity with the stimuli, potentially amplifying any learning effects caused by the low stimulus variety. Familiarity has consistently been demonstrated to aid visual processing across various stimulus categories (Dosher & Lu, 2017; Sagi, 2011), including scenes (Epstein et al., 2007; Klink et al., 2023; Lee & Quessy, 2002), where fMRI studies have reported differential responses in scene-selective cortical regions to familiar and unfamiliar scenes (Bainbridge & Baker, 2022; Epstein, Higgins, et al., 2007; Epstein, Parker, et al., 2007). In addition to the low stimulus variety, research has shown that participants are more efficient at recognising familiar scenes when the viewpoint remains stable (Christou & Bülthoff, 1999), as in the current study. Given the combination of the relative ease of the task, and the conditions encouraging the recognition of familiar scenes, these factors may have resulted in a potential ceiling effect, where differences between the conditions could no longer be observed.

This issue may have been further amplified by many participants having also taken part in our original drawing experiment (chapter 3), with 15 taking part in experiment 1, and 11 in experiment 2. This may have resulted in increased familiarity with the participants own and control scenes, as these remained constant across the current study and our original drawing experiment. This familiarity could explain the increased accuracy for the renders based on participants own drawings and the controls found experiment 1 and 2. However, returning participants may have also had prior exposure to the stimuli in the other condition, as in our original drawing experiment participants viewed all of the renders produced by all other participants. This means that if a participant was placed into a batch with another returning participant, they would have been both previously exposed to the renders based on each other's drawing (comprising part of the other condition). Subsequently, it is difficult to ascertain how much familiarity could have affected the differences between the own, other and control conditions, as the degree of familiarity to the other condition may not be equal between all returning participants. As such, the use of non-naïve participants represents a further confound in the current study.

Could this familiarity have additionally modulated the impact of manipulating scene content, potentially diminishing its impact? Prior research investigating familiarity for objects, faces and letter arrangements found that familiarity can lead to a more holistic processing approach, potentially causing observes to rely more on broad category level visual information and to overlook specific details within familiar stimuli (Barenholtz et al., 2016; Garcia-Marques & Mackie, 2007; Huang, 2011;

Tovey & Herdman, 2014, 2014; Q. Wang et al., 1994). If such an effect is present within familiar scenes, than this may have reduced the impact of manipulating content within renders participants were more familiar with, such as the own and control scenes. It may have also contributed to our ability to detect these changes in experiment 2, where participants had already been exposed to all stimuli. However, evidence suggests that this effect may not be present within scenes, and instead that familiarity may cause even greater attentional allocation to scene content. Early work by Teitelbaum (et al., 1978) found that prior visual exposure to a scene facilitated faster detection of incongruities, suggesting that familiarity instead enhances sensitivity to scene detail. More recent research by Cohen (et al., 2024) supports this conclusion. Here they found that participants were better at detecting changes within more familiar scenes, and that familiarity may expand the bandwidth of perceptual awareness. As such, the role of familiarity within the current study may be complex, and not easily predicted with the current data. This complexity again suggests the necessity for the current experiments to be repeated with naïve participants, and for more work to be conducted exploring the role of familiarity in driving the effect observed in chapter 3.

4.4.3 Conclusion

Taken together, whilst experiments 1 and 2 suggest that semantic and syntactic object information is not represented within internal scene models, it is difficult to distinguish whether the results we observed truly reflected an investigation of the object content of internal models, or whether such effects were masked by experimental confounds such as increased stimulus familiarity. A tentative interpretation of our results suggests that whilst internal models may not contain specific object level representations, they could contain a broader representation of semantic and syntactic scene information, that is robust enough to tolerate the relatively minor object-level manipulations applied in the current experiments. This may be reflective of the role of internal models in natural vision, where in order to help facilitate the rapid processing of scene information, they need to be able to be matched to a wide range of different scene exemplars, regardless of the intra-category differences in the layout and identity of the scene's constituent objects. However, whilst our study provides a useful initial exploration of the content of internal models, weaknesses in the experimental design limit the applicability of these findings. First, the use of non-naïve participants and a small subset of stimuli may have increased participants familiarity with the stimuli, causing potential ceiling effects and limiting our ability to detect differences between conditions. This problem is further compounded by our failure to replicate the findings of chapter 3, where we instead found no difference between render's based on a participant's own drawings and those of a control scene, further suggesting that familiarity with the stimulus may have influenced our results. Secondly, which objects were manipulated relied

upon experimenter judgement, which may have meant that the most saliant objects for a given participants own internal model were not altered, limiting the effects of the manipulations applied. Finally, the choice of paradigm relied on the assumption that matching scene information to internal models would occur rapidly. However, as previously discussed rapid scene categorisation may rely on mechanisms that do not require analysis of object level information. This may have limited our ability to detect any individual differences that might occur at later processing stages, or that might be more task dependent.

As such, in order to clarify the results of the current study, future research needs to be conducted that addresses these limitations, in order to better isolate the factors contributing to the observed effects. This may be best achieved through the use of a visual search paradigm, in which participants locate category neutral objects within renders based on their own, other and control scene drawings, where the constituent objects are manipulated based on the participants ratings of which objects in their own scenes were the most defining. In addition, whilst a replication of the results of chapter 3was a secondary objective for the current study (as the results were already successfully replicated in our manuscript, see experiment 2 in Wang & Foxwell et al., 2024), the results of the current study suggest that further investigation is required, particularly to better understand the influence of familiarity. One potential approach could involve comparing participants' scene drawings to photographs of their real-world equivalents (e.g., their own living room), in order to clarify to what degree drawings, reflect internal models or are influenced by familiarity with specific locations. Although it is important to consider that the failure to replicate these findings may be due to the changes made to the experimental design in order to prioritise the current studies aims, greater clarity of the original effect would allow for future research to use the drawing paradigm with a greater understanding of the factors that drive the observed effects.

Chapter 5: Discussion

5.1 Results Summary

In the current thesis we aimed to explore how typical scene structure influences scene perception, by investigating two complimentary research questions: how is global scene structure extracted and utilised during scene processing, and how individual differences in internal scene models shape categorisation performance. Across three studies, we examined these questions, first using a more traditional scene jumbling method, and then by developing a new drawing-based method to assess internal scene models directly.

In chapter 2, we investigated how global scene structure impacts perception using a jumbling paradigm. Whilst we found that disruptions to scene structure impair categorisation, we were specifically interested in investigating whether disruption along a scenes vertical axis would cause a greater disruption to scene processing, reflecting the rigid vertical structure found in natural scenes. Whilst we observed a strong disruption when structure was manipulated along the vertical axis, our results suggest that this may be driven more by disruptions to low or mid-level visual properties rather than to a sensitivity in vertical structure. This difference may reflect the importance of low and mid-level features in communicating natural sky-ground segmentation cues, indicative of a horizon. However, inconsistencies between experiments, particularly regarding the interplay between horizontal jumbling and inversion, indicate that further research is needed to clarify the robustness of these effects. Whilst future work is needed to explore whether these findings generalise across a broader range of scene categories and tasks, particularly to better understand whether the same effects can be observed in indoor scenes (where a less prominent sky-ground segmentation occurs), it provides insights into how the visual system may be adapted to real world regularities observed in natural scenes.

In chapter 3 we explored the use of line drawings as a tool to quantify internal scene models. Participants created drawings representing a typical exemplar of a given scene category, which were then used to assess whether individual differences in internal representations influence categorisation performance. We found that renders based on participants own drawings were more preferentially categorised compared to renders based on other participants drawings and a control scene, drawn at the same time, indicating this effect was not the result of familiarity with the stimulus. The results suggest that these drawings can approximate internal scene models, and that scene categorisation performance is, to some extent, influenced by the degree to which a scene matches an individual's internal model. This supports predictive processing theories that suggest that internal models are used

to facilitate scene perception, and provides evidence for the influence of individual differences in scene processing.

However, the results of chapter 4 raise some doubts about these initial findings. Here, we aimed to further investigate the content of internal scene models, testing whether object identity and location were important features. We did not find a significant effect of manipulating object identity or location, regardless of the severity of these manipulations, which could suggest that internal models may encode scene representations in a more holistic manner, rather than being strictly dependent on object-level features. However, we failed to fully replicate the effects observed in chapter 3, finding that whilst participants were more accurate at processing scenes based on their own drawings compared to those of other participants, they were not significantly better at doing so compared to the control scenes, suggesting that the effect may in part reflect a familiarity with the stimulus. Whilst the experimental designs utilised in chapters 3 and 4 varied, reflecting differences in the studies aims, this result was surprising, as it was expected that our findings would replicate given the similar paradigms used. Several methodological differences may have contributed to these results, including the use of a small stimulus set, familiarity with the stimuli resulting from the use of non-naïve participants, and the possibility that object-level manipulations may not have been observable with a scene categorisation task, which might have better been detected using an alternative experimental paradigm that relied on participants attending to a scenes constituent objects, such as visual search.

Taken together, whilst these findings provide some evidence that scene perception is guided by internal models that mirror real-world scene regularities, the inconsistencies observed in chapter 4 highlight key limitations in our ability to fully support this claim. In the following sections, we will discuss the key findings of each chapter in relation to each other, and their implication to existing theories of scene perception, followed by a discussion of the drawing method developed in the current thesis and its implications for the study of scene perception and internal models more broadly.

5.2 Parsing structural regularities in scene perception: vertical bias effects may be attributable to low- and mid-level visual features

In chapter 2, we investigated how global scene structure facilitates scene processing, utilising a similar jumbling paradigm to the one originally used in classical work by Biederman et al (1974), in which scenes are divided into rectangular segments, and rearranged so that they are no longer presented in their typical spatial configuration. In order to explore whether there were further axes-based differences between vertical and horizontal structure, we rearranged segments so that they were either fully jumbled, vertically jumbled, or horizontally jumbled.

We found that fully jumbling scene structure caused categorisation accuracy to decrease, indicating that coherent global scene structure helps facilitate efficient scene processing. This result replicated previous experiments that utilised jumbling to demonstrate the importance of coherent global scene structure (Biederman et al., 1974; Kaiser, Häberle, et al., 2020b, 2020a; Kaiser, Inciuraite, et al., 2020). Importantly, this effect was not influenced by inversion, suggesting that it was the disruption caused to the scene's global structure, as opposed to the low or mid-level visual characteristics, that drove this effect.

However, when examining the effects of jumbling across different scene axes, whilst we were able to find a significant impact of vertical jumbling, categorisation became increasingly worse when scenes were displayed at inversion and rotation. This finding contradicts previous research that found inversion reduced the impact vertical jumbling had on scene categorisation, that suggested the effect could be attributed to disrupting the scenes' vertical global structure (Kaiser, Turini, et al., 2019), as opposed to low and mid-level scene characteristics also disrupted by the course manipulations applied during scene jumbling. Instead, our results found the opposite, suggesting that the effect of vertical jumbling we observed may be the result of disruptions caused to the scene's low and mid-level characteristics. This challenges the idea of a vertical bias within global scene structure, and instead suggests that previously observed benefits to the organisation of vertical structure within scenes (Fischer et al., 2016; Hansen & Essock, 2004; Previc & Intraub, 1997; Tucciarelli et al., 2023) may be better explained by the influence of low or mid-level visual features.

What low or mid-level visual features could benefit from intact vertical organisation? One possibility is that this effect is driven by the more rigid organisation of low-level visual features across a scene's vertical axes. An analysis of the amplitude spectrum of 1,017 natural scenes, reflective of the low-level visual information, found the greatest variety across the horizontal axis, with variation in vertical scene information primarily being derived from flora (such as the difference between the trunk and canopy of trees) (Hansen et al., 2008). This greater variation in visual information across the horizontal axis may evidence fewer universal commonalities in horizontal information across different scenes. Conversely, the more rigid low-level characteristics along the vertical axis may provide stronger visual regularities that can be utilised to help facilitate scene processing.

As discussed in chapter 2, this vertical rigidity may be the result of the distinct low-level characteristics of sky and ground sections present in many outdoor scenes, with the sky typically composed of smooth, low-frequency gradients, and the ground consisting of high-frequency textures and sharp edges (Julesz, 1981; Thorpe et al., 1996). These features are maintained within the individual segments, and may be important indicators for identifying the horizon within outdoor scenes, which

might be important for quickly assessing the spatial structure of a scene. Such an explanation may help to explain why the observed effect of vertical jumbling differed to those observed in Kaiser et al (2019) which also utilised indoor scenes where horizons are less prevalent. This difference might not strictly represent a distinction between indoor and outdoor scenes however, but instead scenes where horizons are key defining features. Whilst the outdoor scenes used in the current study all consisted of clear horizons, many outdoor scenes have far less prevalent horizons (such as forests) and subsequently might be invariant to any disruptions along the vertical axes. Such an interpretation could align with predictive processing theories, as the visual system may have adapted to make use of the horizon as it acts as a constant and predictable feature for many outdoor scenes, and an important indicator of a scenes structure and overall organisation. Low-level visual characteristics may be important in identifying this horizon, and subsequently, disrupting this information may have resulted in the impaired categorisation. This effect could have been made worse when scenes were shown at inversion, where these low-level horizon indicators were disrupted further.

The disruption caused by manipulating low and mid-level visual features along the vertical axis could also have disrupted the scene's spatial envelope, impairing gist extraction and subsequently delaying scene categorisation. Research has shown that the horizon can be identified in as little as 153ms in outdoor scenes, from the gist of the scene alone (Herdtweck & Wallraven, 2013), which could suggest it's importance in establishing a holistic representation of a scene's spatial envelope. Whilst the current study provided participants with long viewing times, meaning that the task did not explicitly rely on gist extraction, if initial gist extraction was disrupted this could have resulted in delayed reaction times.

However, it is impossible to fully understand what features could drive the observed effect of vertical jumbling from the current study alone, as although the effect of inversion suggests the influence of low and mid-level features, we cannot parse which of these features drove this effect. Regularities in various low level visual features have been found to help categorise outdoor scenes, such as colour, (Ganesan & Balasubramanian, 2019), edge alignment (Payne & Singh, 2005) and contrast (Stürzl & Zeil, 2007), all of which are disrupted during the vertical jumbling condition. In addition, previous research investigating the influence of low-level visual features on scene jumbling have found that the influence of factors such as colour and spatial frequency on the extraction of global scene structure vary on a category level (Vogel et al., 2007). This may add additional complications when trying to parse the causes of these effects, as they may vary by scene category, which was not explored in the current thesis.

Whilst we posit that disruptions to low-level features along the vertical axis may align with meaningful structural scene elements (such as the horizon), further work is required to understand whether such an effect is truly present, and which low or mid-level features might drive it. This could be achieved by systematically controlling for these low and mid-level visual features whilst jumbling scenes across different axes, to assess their individual and combined effects on scene categorisation. Additionally, in order to investigate whether the horizon might drive this effect, future studies could compare the effects of vertical jumbling in scenes where horizons are more or less prevalent. Such experiments may help us better understand how coherent visual information presented along a scenes vertical axis may help facilitate efficient scene processing.

As such, whilst the replication of previous studies showing the importance of intact global scene structure reinforces the notion that the visual system utilises regularities in global scene structure to help facilitate efficient scene processing, our results do not indicate that this extends to an axes level bias reflective of the organisation of real-world scenes. This could indicate that the benefits of intact global-scene structure represent a more holistic representation of a scene's layout. Additionally, the identification of a possible sensitivity to low and mid-level level features organised along a scene's vertical axis could suggest that local features drive a potential vertical bias for the arrangement of scene information. Whilst we suggest that such features might be important at establishing a horizon within outdoor scenes, ultimately, the methodological limitations of the chapter 2 necessitate further experimentation in order to identify what features might drive the observed effect of vertical jumbling,

5.3 Internal models, familiarity, and scene processing efficiency

A key objective of the current thesis was to explore the contribution of internal models to scene perception, in order to better understand how predictive processing mechanisms might underpin the visual system's ability to anticipate and interpret natural scenes. To achieve this chapter's 3 and 4 aimed to develop and test a new drawing paradigm that would allow us to characterise the contents of a participant's own internal model, so that we could investigate whether individual differences within these models could explain idiosyncratic differences in scene processing.

In chapter 3 we found that participants were more efficient at categorising scenes based on their own drawings of typical scenes, compared to those based on drawings produced by other participants or copied during a control condition. Crucially, utilising a DNN we found that categorisation accuracy for all scenes correlated with how similar they were to the participants own scene drawing. This improved accuracy could suggest that the generation of accurate predictions, derived from the participants own internal scene models, directs the processing of incoming visual input. The strength of this match

would in turn modulate how efficiently a scene is processed, with scenes closely adhering to the predictions being more efficiently processed, whilst those deviating require further analysis to extract their meaning (Bar, 2004; Friston, 2005). Whilst such effects have previously been observed on the group level, with participants demonstrating improved accuracy and stronger neural responses to highly typical scenes (Caddigan et al., 2017; Torralbo et al., 2013), these results suggest that individual differences between participants own conceptions of typicality may influence scene processing, supporting the notion that internal scene models reflect the different visual experiences of observers (Hartley, 2022).

However, the results of chapter 4 were less conclusive. Here, we were unable to fully replicate the results of chapter 3, finding instead that although categorisation accuracy was higher for renders based on participants' own drawings compared to those of others, it was not significantly higher than the control condition. This could suggest that the improved accuracy observed for participant own scenes may have been influenced by the increased familiarity with the stimuli. Familiarity has been found to evoke differential response in scene selective brain regions and to help facilitate more efficient scene categorisation (Epstein, Higgins, et al., 2007; Epstein, Parker, et al., 2007; Noad et al., 2024). Recent EEG work has also found that personally similar scenes evoke distinct decodable neural activity (Klink et al., 2023), indicating that individual differences found in the current study could instead stem from idiosyncratic differences in familiarity. These conflicting results challenge the assertion that the improved accuracy observed in chapter 3 relied upon the strong match of scene renders to internal scene models, and suggest that familiarity may have influenced the observed effect.

As such, a crucial question for the current thesis is to understand to what degree familiarity can explain the increased categorisation accuracy for participants own scene renders, and to whether these effects were present in the experiments conducted in both chapters. Several methodological factors of the experiments conducted in chapter 4 may have resulted in an increased familiarity with the stimuli compared to chapter 3; including the use of a smaller subset of stimuli, with each individual scene (and manipulated version) being repeated considerably more times than in chapter 3. Chapter 4 also utilised many non-naïve participants that had previously taken part in chapter 3, meaning they had previously been exposed to their own and the control scenes, which may have increased their familiarity further. These factors may suggest that the effect of familiarity could be unique to chapter 4, driven by methodological changes implemented to test a different experimental hypothesis. Alternatively, these conditions could have increased the chances of us detecting the influence of familiarity within our paradigm, not initially detected within chapter 3.

If familiarity accounts for the discrepancy observed in our results, it is important to consider the nature of that familiarity. Was it driven by familiarity with the scene content depicted in the stimuli, or by a more general familiarity with the stimuli as visual artefact? For example, were participants familiar with the scene being depicted in the renders, or simply with the specific images themselves? Whilst in chapter 3 participants had no prior exposure to the renders, only the drawings they had produced, in chapter 4 many participants had seen the specific renders used previously. This may have counter acted one of the perceived advantages of using the renders instead of the participant's drawings directly; whilst the renders would maintain the identities and layout of constituent objects, the participants would not be biased by a familiarity with their own drawing style or prior exposure to the drawings presented. This could suggest that the familiarity observed in chapter 4 was a result of a more general familiarity with the renders, and was not present in chapter 3 as participants had no prior familiarity with the stimuli themselves.

As such, the effect of familiarity in chapter 4 could represent the influence of a more general familiarity with the stimuli, as opposed to the scene information they represent. Given the relative ease of the task, and possible training effects resulting from non-naïve participants having previously completed a similar experiment with similar stimuli, this familiarity could have produced a ceiling effect, making it difficult to detect any additional advantage caused by a similarity to a participants own internal scene model. Conversely, participants in chapter 3 had no such prior exposures, and thus would not have been influenced by a more general familiarity with the stimuli. As such, the influence of familiarity represents a difficult question for the current experiment to answer, that may necessitate future research to disentangle. One way that this might have been achieved is through showing scene renders from different viewpoints; if the familiarity results from the scene information, we might expect it to persist even when scenes are shown from new viewpoints, whilst if it derives from a specific familiarity with the stimuli, it might be more dependent on a fixed recognisable perspective (Epstein, Higgins, et al., 2007).

The results of the DNN graded similarity analysis may further suggest that familiarity might have had less of an influence on the results of chapter 3. Here we found that categorisation varied systematically across renders as a function of how similar they were to the render based on their own typical drawing. As participants had no prior exposure to the renders based on other participants drawings, this similarity could not be driven by familiarity. Importantly, we did not find that a graded similarity to the control scene correlated with categorisation accuracy, which would be expected if these results were driven by a familiarity to the scenes.

However, this does not fully discount the influence of familiarity entirely. The graded similarity may have been modulated by the degree of familiarity, with participants not being familiar enough with the control scene to evoke this effect. If this familiarity derives from the scene content, then it is plausible that participants may have been more familiar with their own scene drawings than those of the control. Although instructed not to draw a specific scene they were familiar with, if internal models are derived from the regularities we experience in our personal visual experience, than it is possible these scenes would share many features with those that they are most familiar with. Recent studies utilising MVPA have found that scene selective regions differentially process scenes that are personally familiar, such as showing a preferences for a person's own office compared to a generic office, (Epstein & Morgan, 2012; Sugiura et al., 2005; Wiese et al., 2023), indicating that the degree of personal familiarity can modulate processing efficiency. In the current studies, this could have been better controlled for by asking participants to rate all of the scene renders produced based on familiarity, and seeing if a graded familiarity score could better predict categorisation, and whether this in turn correlated with how well similarity to a participants own drawing predicted categorisation accuracy.

Alternatively, the difference between the own and control condition might stem from the differing cognitive processes involved when composing an original scene, compared to copying a picture of a scene. Differences in neural activation patterns during drawing tasks suggest that composing an original scene and copying an existing picture engage distinct cognitive processes. Research has found that copying primarily involves the intraparietal sulcus, which facilitates the direct transformation of visual input into corresponding motor actions, whilst in contrast, creating a novel drawing activates the anterior cingulate cortex, reflecting the engagement of higher-order functions such as planning, decision-making, and creative integration (Ferber et al., 2007; Ogawa & Inui, 2009). This distinction implies that the cognitive demands of generating original content are more complex, requiring additional neural resources beyond those utilised in mere replication.

Could the differences within the cognitive demands for copying and constructing scenes help explain the different effects of control scene in chapter 3 and 4? One possibility is copying and drawing original scenes may have differential effects on the visual memory benefits of drawing (Fernandes et al., 2018; Peynırcıoğlu, 1989; Wammes et al., 2016; Zhou et al., 2025). Recent research by Wammes, Jonker and Fernades (2019) sought to explore what aspects of drawing account for improved memory evoked by drawing, by comparing different types of drawing tasks (tracing, viewing, imagining, and drawing without seeing the output). Whilst they found that memory effects decreased for tasks lacking the visual and motor components of natural drawing behaviours, the greatest impact was caused by elaborative component of drawing. They defined the elaborative component as the generative processes used to imagine an internal representation of visual and semantic features. Whilst they did

not include a copying task specifically, copying distinctively lacks this elaborative element, and thus may produce a weaker benefit to copying a scene.

However, whilst an increased effect of personal familiarity with one's own scene renders, or differences in memory caused by drawing vs copying, might explain why the DNN found a graded similarity for the own scenes but not the control, it is unclear why such effects would not be detectable in the categorisation accuracies measured in chapter 4 (where we found no increased accuracy between the own scene vs the control). One possibility is that this difference might be explained by a decrease in the potency of these effects over time, as several months passed between the experiments conducted in chapter 3 and 4. However, the benefits of familiarity have been found to be relatively stable over time (de Chastelaine et al., 2017; Friedman et al., 2010; Koen & Yonelinas, 2016), suggesting that it might be unlikely they fade within this time frame. Whether the potential benefits familiarity has on drawing could be impacted by time remains underexplored, as currently no studies have investigated how long these benefits last over extended periods. As such, it remains difficult to understand how these factors may have influenced our findings, necessitating further research.

Whilst the difficulty of isolating the effects of individual differences in familiarity from those of internal models reflects a broader conceptual challenge, it is possible that familiarity does not operate as a distinct influence. Instead, familiarity might interact with internal models—potentially enhancing the accessibility or precision of predictions, and in turn improving processing for scenes that are both highly typical and familiar. In their recent review paper, Servajean and Wiese (2024) outlined how familiarity might act as an important indicator of the precision of estimates derived from internal models, suggesting that familiar stimuli may help to reduce uncertainty in predictive processing mechanisms. They propose that familiarity forms an important element of a fluency heuristic, which reflects how easily a stimulus is processed compared to how easily internal models expect it to be processed. When this difference is large, it may signal a need for structural changes in internal models, thus helping to expand our predictive capacity. This implies that the brain may weigh sensory input more heavily when it aligns with both a familiar context and strong internal expectations, thereby facilitating more efficient and accurate perception. Whilst little work has explored how familiarity might modulate predictive processing mechanisms in scenes specifically, research into the learning of familiar faces has found that familiarity helps to modulate predictive processing mechanisms when leaning new facial identities (Apps & Tsakiris, 2013). Given recent research suggesting that familiarity plays a similar role in scene processing as face perception (Klink et al., 2023), this could suggest that familiarity may likewise help facilitate the predictive processing mechanisms in scene processing. Whilst the current thesis did not aim to explore how familiarity could influence predictive processing

mechanisms, and subsequently cannot attest to how the two may interact, our findings do highlight the importance of considering the interplay between these factors in future research

It remains a consideration for future research to continue to investigate how individual familiarity and typicality interact to shape the efficiency of scene processing. Disentangling these effects completely may prove challenging, as internal models are likely to be developed as the sum experience of our exposure to our environments, and subsequently resemble those that we are more familiar with. Future studies could take two possible approaches. Firstly, they could compare participants' drawings when they are instructed to either draw the most typical scene exemplars, to when they are instructed to draw the most familiar. Whilst this would allow for a more direct comparison between conceptions of typical and familiar scene content, it is unclear whether participants would be able to accurately draw a familiar scene, and how much their internal model would influence their ability to mentally reconstruct this environment (and thus one task instruction tainting the other one). Alternatively, participants' own drawings of typical scenes could be compared against photographs of their everyday environments, and categorisation could be predicted separately from each scene's similarity to participants' scene drawings, as well as photos of their current living environments, allowing the effects to be better differentiated. Another approach could be to utilise tasks that have been shown to be unaffected by familiarity. Whilst previous research indicates a beneficial effect of familiarity in scene categorisation tasks (Bainbridge, 2022; Epstein et al., 2007; Klink et al., 2023), it has been shown to provide less of a benefit when searching for objects within scenes (Võ & Wolfe, 2012). If similarity to internal models indeed facilitates efficient scene processing, we might expect visual search to be improved in scenes more closely approximating internal models, helping to further differentiate the influence of familiarity versus typicality on the individual level.

Our findings highlight the complex interplay between familiarity and typicality in shaping scene perception, raising important questions about how internal models are formed and utilised. While previous research has demonstrated group-level advantages for processing highly typical scenes (Caddigan et al., 2017; Torralbo et al., 2013), the current work suggests that there may also be individual differences in these advantages, potentially reflecting the idiosyncratic visual experiences that shape an individual's internal scene model (Hartley, 2022). This assertion does not imply that perception is fully unique, or even radically different, between observers. We still found a fair reliability of categorisation performance across observers, with a modest split half-reliability of $r = 0.72$. What our results do suggest is that on top of this coarse stability in performance, there may be interesting additional variance that is systematic across observers and can be captured by our drawing-based method. However, the current thesis also highlights the importance of disentangling effects driven by typicality from those rooted in familiarity (Epstein et al., 2007; Klink et al., 2023). This

complicates the interpretation of results and underscores the need for future research that directly compares these influences or employs tasks less susceptible to familiarity-driven effects (Võ & Wolfe, 2012). Overall, the results of chapter 3 and 4 contributes to the growing body of work on scene perception by providing tentative evidence for the individual nature of internal models and by identifying familiarity as a key factor that must be carefully controlled in future research.

5.4 The representation of object information in internal scene models

An important question we sought to answer in the current thesis was what features define internal scene models. Initially, we hypothesised that internal scene models would store information about typical objects and their placements, in line with previous research suggesting typicality in the identity (Davenport & Potter, 2004; Faivre et al., 2019; Mudrik et al., 2010, 2011; Võ & Wolfe, 2013), positioning (Kaiser et al., 2018; Kaiser & Cichy, 2018a, 2018b), and relative spatial relationships (Bilalić et al., 2019; Kaiser et al., 2014; Kaiser & Peelen, 2018; Stein & Peelen, 2015) of constituent objects helps to facilitate efficient scene processing However, whilst the results of chapter 3 provided some support for our hypothesis, this is contrasted by the findings of chapter 4, where the content of internal scene models were more directly investigated. Below we will discuss what our results could suggest about the content of internal scene models, and how methodological constraints may have limited our ability to detect object level manipulations.

In our drawing paradigm participants primarily constructed their scenes by deciding what objects they would include and where they would be placed, with other aspects of the scenes being somewhat standardised, including the viewpoint, approximate shape of the room, and the representation of specific objects (due to a limited amount of 3D models being used). This suggests that it was the participants choice of objects, and their placement, that characterised their drawings, and subsequently drove the improved categorisation accuracy we observed in chapter 3. The importance of the object content in driving this effect was supported by the findings of our DNN analysis, where models trained on both scenes and objects expressed a graded similarity in categorisation accuracy, indicating that scenes with similar object content were more accurately categorised. Whilst both models found graded similarity could predict categorisation accuracy, the models trained on objects yielded more accurate predictions. This analysis also found that graded similarity was best expressed in later convolutional layers of the Googlenet model, which approximate the later stages of visual processing and simulate the processing of high-level features such as object-related information, further evidencing the importance of object representations.

However, contrary to these results, in chapter 4 where we sought to specifically investigate the representation of object information within internal models, we found no significant effect of

manipulating the scene content of renders, regardless of whether the drawings more or less closely resembled a participant's own drawing. If high-level object-related features were represented in internal models, we would expect that changing this content would disrupt the matching process to the internal model, with further changes in object information reducing the match even more. Below we explore two possible interpretations of these null results; that either object level information may not be represented within internal scene models, or that internal models could contain a broader and more "holistic" representation of object content, resistant to the manipulations we applied. We also explore the possibility that that this discrepancy may stem from methodological limitations of chapter 4, and what these might indicate about the role of internal models in scene processing.

This first interpretation raises the possibility that individual object information may not be necessary to facilitate rapid scene categorisation as previous research implies (Davenport & Potter, 2004; Võ & Wolfe, 2013; Kaiser et al., 2018), and instead may rely more heavily upon the combination of other visual features. Subsequently, this could challenge the assumptions of predictive processing, and suggest that instead of making predictions based on the statistical regularities in the object content in a top-down process, our results could potentially be explained by a bottom-up scene-first explanation of visual processing. Scene-first theories suggest that instead of analysing specific object level information, scenes are processed on a global to local level (Hochstein & Ahissar, 2002; Oliva & Torralba, 2006; Schyns & Oliva, 1994), where low spatial frequency information can rapidly extract the gist of a scene, which in turn helps provide contextual information to facilitate object processing (Oliva & Torralba, 2006; Wu et al., 2018). Such a view aligns with the arguments put forth by Groen (et al., 2017), who argue that even in the absence of object information, low and mid-level visual characteristics can facilitate the extraction of the visual information needed for many scene related tasks, such as categorisation and navigation (Greene & Oliva, 2009a, 2009b). They further suggest that instead of necessitating hierarchical processing, low and mid-level features are instead combined with high-level object information in a more integrated framework. Supporting this idea, they cite research exploring the temporal dynamics of how these features are represented during scene processing showing that whilst low, mid and high-level features can be decoded separately (Bieniek et al., 2013; Hansen & Hess, 2007; Thorpe et al., 1996), the time course of when each feature is most strongly represented does not follow a straightforward low to high level progression, and instead overlap and intermix (Martin Cichy et al., 2017; Ramkumar et al., 2016; Wardle et al., 2016), suggesting further than scene categorisation is not reliant on object information.

It is unclear from the current results which low and mid-level characteristics could explain the individual differences we observed in chapter 3. Here we controlled for several low-level visual

features, such as colour, luminance and contrast, suggesting these features were unlikely to account for this variance. However other properties such as depth, spaciousness and navigability, could not be controlled for, as these are somewhat defined by the scenes content. One possibility is that improved gist extraction might be facilitated by participants' own scene drawings representing an optimal arrangement of the category specific scene statistics learnt by that participant. This could be evoked by the spatial statistics defined by the edge boundaries and surface properties of the overall arrangement of constituent objects represented in the renders based on participant's own drawings. Previous research has found that the presence of highly typical constituent objects placed in typical locations biased encoding of a scenes spatial properties in PPA towards the average representation for that category, with MVPA finding that this was driven by the objects modulating the perceived spaciousness of the scene (Linsley & MacEvoy, 2015). If the categorical information communicated by a scenes spatial statistics are modulated by the presence of typically occurring objects, then this may be further controlled by adherence to an individual's own conception of typicality, thus evoking the idiosyncratic categorisation effects observed in chapter 3.

Such an explanation may be challenged when considering some of the proposed properties of gist extraction. Whilst gist extraction is believed to be reliant on the learning of statistical regularities through repeated exposure to stimuli (Brady & Oliva, 2008; Groen et al., 2013; Loschky et al., 2015; Raat et al., 2022), research has found this process may already operate at peak performance for scenes (Fabre-Thorpe et al., 2001), indicating it may be unlikely to be sensitive to individually learnt regularities that would produce idiosyncratic differences. Further, the primary strengths of scene gist is its ability to accurately categorise scenes irrespective of the many inter-category differences, such as specific object representations or viewpoints (Greene & Oliva, 2009a). Conversely, evidence suggests that gist may be less efficient at detecting inter-category differences, where it has been shown that gist extraction is less efficient at identifying subordinate scene categories, such as distinguishing between a diner and a fine dining restaurant (Malcolm et al., 2012). Another issue with this explanation is that if the holistic arrangement of objects does communicate an optimal arrangement of a scene's spatial statistics, we might predict that the manipulations to the location of objects, as applied in the swap condition in chapter 4, to redefine the edge boundaries and surface properties of the overall arrangement, and subsequently impact any categorisation advantage. As such, given the strength of scene gist at effectively capturing the general statistical regularities of a scene, it seems unlikely that the idiosyncratic differences we observed could reflect similar individual differences derived from gist level descriptors. In order to explore this possibility, future research could investigate if the spatial envelope of participants own scenes can evoke similar idiosyncratic differences in categorisation accuracy as observed in chapter 3. This might be achieved by blurring the

scene renders, to obscure object identities, or by displaying them for briefer intervals so that only gist can be extracted, and investigating whether these idiosyncratic effects remain.

Alternatively, the results of chapter 4 may suggest that internal scene models represent a more holistic, broad level interpretation of a scene's object content. Instead of containing information about the identity and location of single objects, internal models might make predictions based on the sum of this information across the entire scene. As such, a more holistic representation might be best represented by the meaningful relationships between all of the constituent objects within a scene, which we did not consider and subsequently may have failed to disrupt in chapter 4. This would suggest that whilst the drawings produced by participants represented the most "typical" arrangement of objects, they may be robust to relatively specific changes, and able to facilitate a more accurate match to a greater number of scenes that share a broadly similar arrangement of objects. For example, a participant's internal model of a kitchen might contain a table near some chairs, with a fridge, oven and sink on a countertop. It may be that other scenes that do not include a fridge, or with the table placed in a different location would still produce a very close match to that internal model, as the content still closely aligns to the predictions driven by that model. This would match the relatively small idiosyncratic effects we observed, suggesting that the predictions derived from internal models are relatively robust.

This could suggest that instead of the distinctive separation of semantic and syntactic object information suggested by Draschkow & Võ (2017), object information may not be as rigidly differentiated within internal models. Instead, the relative relationships between objects might be particularly important to internal scene models. In their recent review paper Peelen et al (2024) makes a similar argument, suggesting that predictive processing mechanisms might make use of the typical spatial configuration of semantically consistent object, which they refer to as object constellations. The notion of object constellations is supported by CFS research showing that when objects are displayed together in typical configurations, they break suppression more quickly (Stein et al., 2015). Crucially, this effect is reliant at showing objects in canonical upright perspectives, suggesting that this grouping is not reliant on low or mid-level visual information. Similar benefits of object clusters have also been observed in neuroimaging studies, where the benefit of grouping objects together with meaningful spatial relationships produces a stronger representation in LOC than showing objects displayed without a coherent spatial arrangement (MacEvoy & Epstein, 2009).

Subsequently if object constellations are represented within internal scene models, and thus drove the idiosyncratic categorisation accuracy observed in chapter 3, the manipulations we applied might not have been severe enough to disrupt this more holistic representation of object information. One

possibility is that whilst we manipulated semantic and syntactic object information separately, we did not disrupt both during any trials, which may have preserved enough of the information about the meaningful relationships within the object clusters to allow for efficient matching to internal scene models.

How might predictive processing mechanisms make use of a more holistic representation of object information? Peelen et al (2024) suggest that in line with models of Bayesian inference, internal models can produce more accurate predictions by different sources of information weighed by their uncertainty. As high-level perception is less ambiguous than low level information, predictions derived from more holistic representation can help disambiguate the information derived from lower-level visual features. For example, the broad spatial envelope of a scene may be able to identify a scenes broad category (i.e. this is a restaurant), whilst the arrangement of objects could help differentiate more sub-ordinate category information (i.e. this is a fancy restaurant). This possibility could align with recent research showing that object information can help disambiguate scene category during gist extraction (Joubert et al., 2007; Malcolm et al., 2012; Wiesmann & Võ, 2022). Such a benefit is well evidenced in a recent study by Furtak (et al., 2022) found that semantically coherent foreground objects aided the categorisation of background scenes even when they were only displayed very briefly, allowing for only the gist to be extracted, suggesting coherent object content can be utilised in conjunction with low and mid-level scene information to help aid scene processing. Given range of valid meaningful arrangements of constituent object information within a single scene category, a more holistic representation of object information may provide internal scene models with a more flexible criteria from which to derive predications, thus allowing them to more accurately predict information about a broader range of different scenes within a category. For example, if internal models contained possible information about the relative positioning and identity of multiple categorically consistent objects within a scene, the absence or irregular placement of some of these objects may be compensated for by accurate predictions that do match the models' expectations.

However, future research is needed to clarify the contents of internal models and whether they might contain a more holistic scene representation of object content. In doing so, it may be pertinent to attempt to disrupt the possible relationships between groups of objects, rather than targeting the influence of individual objects. Achieving this may be challenging, as it is difficult to discern what relationships may be stored in internal models (if any). One possibility is that these holistic representations contain a combination of both semantic and syntactic object information. In chapter 4, we manipulated these factors separately, and so it is possible that the source of object information that remained intact was sufficient at facilitating an efficient match to internal scene

models. By applying more targeted manipulations, this may help future research to better understand the representation of object content within internal scene models.

Whilst the previous interpretations provide possible explanations of what the null results of chapter 4 might tell us about the content of internal models, it is important to consider that methodological limitations in the experiment may have limited our ability to detect the effects of any object level manipulations. Although we did not find any significant effects of object manipulation in chapter 4, we did observe a nominal non-significant trend in the data, suggesting a stronger impact of manipulating object identity on the categorisation of renders based on participants' own scene drawings . This trend may not have reached significance due to several methodological weaknesses present in the study. As discussed in the previous section, the combination of non-naïve participants and the greater number of repeats of a considerably smaller stimulus set could have increased the influence of familiarity on categorisation accuracy. Given the relatively small effect sizes observed, the effect of familiarity may have masked any benefits derived from matching stimuli to internal scene models. This may explain why manipulating the content of participants' own scenes had no differential effect compared to manipulating the content of other scene drawings. However, research exploring the effect of familiarity on the processing of individual features is mixed. Whereas work on object, face and letter arrangements suggests that familiarity causes individual features to be represented more holistically, potentially limiting the impact of manipulating individual elements (Barenholtz et al., 2016; Garcia-Marques & Mackie, 2007; Huang, 2011; Tovey & Herdman, 2014, 2014; Q. Wang et al., 1994), research on scenes suggests that familiarity may instead increase attentional allocation to individual objects (Teitelbaum et al., 1978; Cohen et al., 2024), from which we might have expected object manipulations to have caused a greater impact. As such, the role of familiarity within chapter 4 remains a consistent confound, and necessitates further research to fully understand its influence.

Another important methodological factor that may have influenced our results was the choice of paradigm. We used a rapid scene categorisation task as we originally hypothesised that the benefits of matching input to internal scene models would occur at early stages of visual processing, based on the findings of chapter 3 and previous studies showing that typical object content helps facilitate rapid scene categorisation (Caddigan et al., 2017; Csathó et al., 2015; Torralbo et al., 2013). However, as previous research has shown rapid scene categorisation can be achieved in the absence of object information (Greene & Oliva, 2009a), it may be that matching object information to internal scene models was unnecessary for achieving scene categorisation, and thus gone undetected. This could suggest that predictions derived from more specific object level information could emerge in tasks that require a greater adherence to the object content. To explore this possibility further, future

studies could explore the impact of manipulating the object level information in renders based on participants' own drawings on performance in visual search tasks (Biederman et al., 1973; Malcolm & Henderson, 2010; Neider & Zelinsky, 2006), potentially revealing effects that were not detectable using rapid categorisation. By contrasting the impact of object manipulation between these tasks, it may help us develop a better understanding of how individual differences in internal scene models influence scene processing.

5.5 Drawing as a flexible method for assessing individual differences in internal scenes models

Developing a flexible method to describe and investigate internal scene models was an important goal of the current thesis. The current thesis developed and applied a new method of studying the individual differences between conceptions of typical scene content. In chapters 3 and 4, we used the method to successfully study the content of internal models for everyday scene categories, demonstrating the viability of this method. Whilst scene construction and drawing experiments have been used previously to study both scene memory (Bainbridge et al., 2019; Bainbridge & Baker, 2020) and perception (Fan et al., 2018; Matthews & Adams, 2008; Morgan et al., 2019; Ostrofsky et al., 2017; Singer et al., 2023), our method has built upon the success of these techniques, by not only analysing the content of the images produced, but its variation across participants.

An important methodological advancement in our drawing method was the process of converting scene drawings into 3D renders. This was done to both control for drawing ability and to allow the scenes to be manipulated without creating visual distortions (for example it is difficult to replace an object whilst matching the drawers style). A potential limitation in the process of converting the scenes to renders was the difficulty in accurately representing each object participants included within their drawings exactly as participants drew them, and instead the closest possible 3D model to the object was chosen. However, the objects drawn within scenes were typically simple, lacking specific detail. This was likely facilitated by both the task instructions, which encouraged participants not to allocate too much of their drawing time to individual objects, and instead focus on the scenes content and layout. A practical consideration for future research utilising this technique is to decide how important the representation of individual objects is for their research question, and ensure the software used for constructing the renders has ample 3D models available to choose from. In our experiments, we utilised the popular video game "The Sims 4" to produce these renders, due to the massive representation of objects available in the game, as well as the ability to easily include additional fan made content available online. Whilst we did not encounter a scenario where we were

unable to find an object included in a participants scene drawing, future studies could utilise more flexible 3D design software, that would allow an even greater choice of objects available.

The use of renders to control for participant's drawing ability allowed for standardising the representation of scenes across participants. This meant that participants ability to categorise scenes was based on the content and layout of the scene itself, as opposed to how well they could interpret other participants drawings. This was particularly beneficial to the within-groups design of our scene categorisation experiments, as it meant that we could discount the improved categorisation of participants own scenes being a result of a greater ability to interpret the content of their own drawings, or other metacognitive judgements they might form around their own drawings. Previous research exploring the judgment of participants own drawings compared to others in adults is limited, but research in children suggest considerable differences between judgements of correctness, as well as overall quality, between participants own drawings and those produced by others (Bonoti & Metallidou, 2010), which may bias experiments that use the participants' drawings as stimuli.

However, although the use of renders limited the influence of drawing ability on the scene categorisation, it may influence how participants constructed their drawings. Although participants were instructed not to concern themselves about the overall drawing quality, their confidence and experience drawing may have influenced which objects they chose to represent in the scene and where. To try and mitigate this, participants were granted time to plan out their scenes before drawing them, so that they did not feel rushed when constructing their scenes. However, across our experiments, although drawing ability varied greatly, the drawings produced provided sufficient detail to evoke idiosyncratic differences in scene categorisation, suggesting that the impact of drawing ability was minimal. However, contrary to these observations, recent fMRI research has found that increased observational drawing ability correlates with functional changes in brain areas involved in attention, decision making, motor control, visual information processing, and working memory (Katz et al., 2021). Such functional changes could indicate the participants with increased drawing ability may be more efficient at not only representing their internal models graphically, but in recalling their contents. In particular, improvements in decisions about light sources, tonal values, line variation and linear perspective drove these changes. It may be possible that due to the simple nature of the drawing task we employed, asking participants to draw simple line sketches in a limited time window, that any effect of individual differences in these more advanced drawing skills was somewhat mitigated.

Drawing ability may also act as a potential limitation for what sources of scene information can be investigated using the current method. The focus of our research was to explore typicality in a scenes'

high-level visual content, which may have been easier for participants to represent in their drawings. However, typicality in mid-level features such as geometry (Bertamini et al., 2018; Hill & Bruce, 1993; Kanizsa & Gerbino, 1976; Mamassian & Landy, 1998) or depth (Kersten, 1997; Mamassian & Landy, 1998) have also been found to benefit scene processing as well, and investigating individual differences within these features may be more challenging for participants to represent within their drawings. Likewise, it may be more difficult to utilise this technique to explore internal models of more complex scene categories, in particular those that are less constrained by spatial boundaries. In the current drawing task, participants were provided with a template of a rectangular empty room (the perspective grid). This was in part to help them more easily construct the scene, as they did not need to think about the spatial dimensions, but also helped them represent their drawings in 3D, allowing us to measure the spatial arrangement of the objects and not just their identity. For indoor scenes, these restrictions were deemed as acceptable, as most indoor rooms consist of a rectangular space, however such restrictions may be less appropriate for outdoor scenes, such as landscapes, which are not confined to these limitations.

A potential limitation of the drawing paradigm for studying internal models is the confound of familiarity. As we have discussed, differentiating between whether the observed effects resulted from a representation of participants internal scene models or a sense of familiarity was a challenge, and limited the application of our results. Whilst our control condition allowed us to monitor the effects of stimulus familiarity based on when the images were drawn, differences between the cognitive loads of drawing vs copying a scene may have made these conditions less comparable than initially expected (Ferber et al., 2007). Further, whilst we initially expected that transforming drawings into renders would reduce the impact of familiarity, as participants were not shown their actual scene drawing or a stimulus they had previously been exposed to, the discrepancy between chapter 3 and 4 may suggest that repeated exposure to the renders may increase the influence of familiarity further. As such, a key learning from this thesis is the need to develop a better control task for this paradigm to account for familiarity and differential effects of constructing scenes. The former could be achieved through asking participants to provide photographs of scenes they are personally familiar with, and comparing these to the drawings they produce. However, this could be more difficult when investigating scene categories participants might be less personally familiar with (such as outdoor or communal spaces). Alternatively, both of these factors could be controlled for by instructing participants to also draw a scene they are most familiar with (such as their own lounge or kitchen), in addition to one that they find most typical. Whilst this would rely on participants ability to differentiate these two concepts sufficiently, and be able to represent this via drawing.

An encouraging finding from our research was that drawings could be used to characterise scenes even outside of laboratory conditions. Due to Covid-19 lockdown restrictions imposed on in-person testing, our drawing sessions were conducted online over video call. Whilst efforts were made to control this environment, it was less controlled than a lab setting, with occasional distractions caused by extraneous background activity. Regardless of these non-laboratory conditions, the images produced were able to act as useful approximations of participants' internal models that could in turn predict behavioural performance in scene categorisation. Combined with the relative ease of the task and low resource costs, drawing methods are an ideal option for collecting descriptors of internal representations amongst populations that cannot easily be studied in a laboratory environment. This could include studying patients with disorders affecting their ability to accurately predict information like autism (Pellicano & Burr, 2012) or semantic dementia (Lambon Ralph & Patterson, 2008), without causing disruption to their care by removing them from their regular environments.

This may also be useful in studying how cultural or environmental differences effect the formation and content of internal models, by allowing us to study participants world-wide and across very different cultures. Previous research has provided compelling evidence for the impact of environmental norms on perceptions of typical scene content (Medin et al., 1997; Miyamoto et al., 2006; Rogers & McClelland, 2004), however, these studies have focused on studying largely urban environments in western and eastern populations. These developed urban populations are likely to have access to media that at least partially exposes them to a broad range of environments from across different cultures, potentially effecting their perception of typical scene content. By studying the formation of internal scene models in more remote communities, these extraneous variables could be avoided. Comparisons to remote communities exposed to different visual environments have been utilised to successfully investigate the environmental influence on other aspects of visual processing, such as colour perception (Roberson et al., 2006), visual attention (de Fockert et al., 2011), depth perception (Hudson, 1960) and the experience of visual illusions (de Fockert et al., 2007; Segall et al., 1963). Whilst we might reasonably assume the contents of internal models developed in remote cultures to vary greatly compared to those of more globalised cultures, differences in how sources of scene information are represented within internal models could help us understand how much our internal models reflect learnt environmental information and how much they reflect the functionality of their neural correlates.

The drawing method may also be useful in investigating how our knowledge and understanding of scenes develop over time. Previously, Öhlschläger and Võ (2020) have used scene construction techniques to study the early development of scene knowledge by instructing children to arrange the furniture in toy doll houses. Such laboratory-based experiments necessitate greater commitment from

participants, which could increase dropouts, and greater logistical demands on the lab, potentially limiting the scope of experimental designs. By using an online drawing task, it may be easier to collect data from participants as they would no longer have to visit a laboratory setting. Further, by allowing infants to take part in experiments from home, this may help reduce potential extraneous impacts caused by the novelty of the laboratory setting (Allen & Bickhard, 2013; Kominsky et al., 2022; Lamm et al., 2014). Such practical advantages could allow for longitudinal studies with larger samples investigating the development of scene knowledge to be more easily conducted.

The drawing method may also be useful at studying the development of internal scene models more specifically. If we do utilise internal scene models, one pressing question is to understand whether the representations in these models are learnt at a young age and remain relatively stable, or if they are constantly being updated to reflect our new visual experiences. Whilst longitudinal studies could be used to explore how internal models develop over time and in response to changes in our environments, it may be difficult to control for the variety of environments people experience, and impractical to conduct over a long period of time. Instead, future studies could explore the development of internal models by comparing drawings of typical scenes to those produced by people who have previously and currently shared living environments. If typical drawings were found to more closely resemble those of people that shared living spaces early in life, such as siblings, it could suggest that the content is shaped by early exposure to real world environments, whilst if they more closely resembled pictures produced by those they currently share their lived space with, such as their spouses or flat mates, it could suggest the content of internal models updates to reflect new scene information.

Beyond helping to develop our understanding of visual processes, the drawing paradigm may help designers and architects create more inclusive and functional environments, by improving our understanding of how individual differences shape our engagement and use of spaces. The concept of place attachment within the field of environmental psychology describes how individuals and groups develop affective bonds to the spaces they inhabit, which encourages people to use spaces and develop a sense of ownership and belonging towards them (Altman & Low, 2012). Whilst place attachment is evoked when built environments reflect the expectations and norms of the people and groups using those spaces, conversely when spaces feel like they are designed for other groups or are highly unfamiliar, this discourages people from using these spaces (Lewicka, 2011; Williams & Vaske, 2003). Such effects have been found in various groups, such as dissuading elderly people from using public spaces- contributing to feelings of isolation and loneliness (Phillips et al., 2011), and in discouraging students from lower socio-economic backgrounds from utilising publics spaces when attending university (Trawalter et al., 2021). These effects have also been found to help enforce the

economic and class divides caused by the gentrification of urban areas, where redevelopments often adopt modern architectural styles more familiar with wealthy, affluent groups, whilst changing the local vernacular and subsequently reducing the feeling of ownership of lower socioeconomic people (Bullock, 2017; Song & Levine, 2024; Ujang & Zakariya, 2015). Whilst the current thesis did not focus on these group differences, the drawing paradigm developed may provide a way to better understand and visualise the expectations different groups have of their environments, so that these can be better reflected and incorporated into the designs of the spaces that we share.

5.6 Conclusion

The current thesis adds to the existing literature demonstrating how regularities in scene information help to facilitate efficient processing. Whilst it provides evidence supporting the role of coherent global structure on scene processing, it suggests that potential vertical biases observed in previous studies may be better explained by regularities in low and mid-level visual features. We suggest that this effect may reflect the stark divide between colours and textures in scenes segments containing portions of the sky and ground, and might be indicative of the importance of these features in establishing a scenes horizon. It also provides an exploration of potential individual differences within internal scene models, and puts forward an updated drawing paradigm to attempt to characterise their contents. The results provide insight into the complex nature of these models, suggesting that in addition to influence of shared regularities found in previous studies, there may be some effect of additional variance between individuals. However, whilst our drawing paradigm was initially able to describe these internal scene models, we were less successful at exploring their content directly. Whilst our results suggest that semantic and syntactic information about individual objects may be less important at characterising internal scene models, methodological limitations restrict the interpretation of these results. Specifically, the use of non-naïve participants and numerous repetitions of a small stimulus set may have increased the influence of familiarity, making it difficult to discern whether the effects of manipulating object content truly reflected an impact on predictive processing mechanisms. This confound of familiarity highlights the need for future research to continue to explore the contents of internal models and what factors might shape these potential individual differences. We suggest that the absence of an effect of manipulating semantic and syntactic object information separately, could suggest the idiosyncratic differences we originally observed are reflective of internal scene models containing a more holistic representation of relative object relationships, and highlight this as a possible direction for future experiments to explore. The drawing paradigm developed in this thesis may provide a useful tool for future studies to not only explore internal scene models, but other differences in how individuals and groups understand their environments, potentially helping architects and designers to develop more inclusive spaces. Overall,

the thesis contributes to the growing body of work providing evidence for the role of regularities in scenes processing, and identifies familiarity as a key factor that must be carefully controlled in future research exploring potential individual differences.

Appendices

Appendix A: Further examples of the stimuli used in chapter 1 experiment 1

Whole upright

Whole inverted

Fully scrambled: Full criss-cross upright

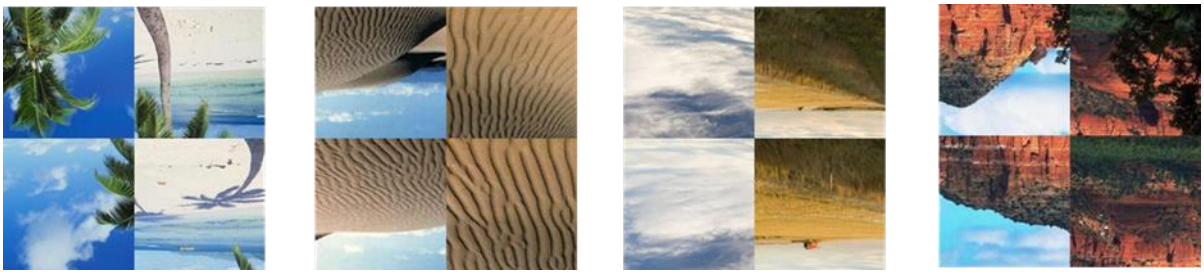
Fully scrambled: Fully criss-cross inverted

Fully scrambled: Top left/ bottom right upright

Fully scrambled: Top left/ bottom right inverted

Fully scrambled: Top right/ bottom left upright

Fully scrambled: Top right/ bottom left inverted



Vertically intact/ horizontally scrambled: Horizontally scrambled upright

Vertically intact/ horizontally scrambled: Horizontally scrambled inverted

Vertically intact/ horizontally scrambled: Horizontal pieces upright

Vertically intact/ horizontally scrambled: Horizontal pieces inverted

Horizontally intact/ vertically scrambled: Vertically scrambled upright

Horizontally intact/ vertically scrambled: Vertically scrambled inverted

Horizontally intact/ vertically scrambled: Vertical pieces upright

Horizontally intact/ vertically scrambled: Vertical pieces inverted

Appendix B: Further examples of stimuli used in chapter 1 experiments 2

Whole rotated

Fully scrambled: Fully criss-cross rotated

Fully scrambled: Top left/ bottom right rotated

Fully scrambled: Top right/ bottom left rotated

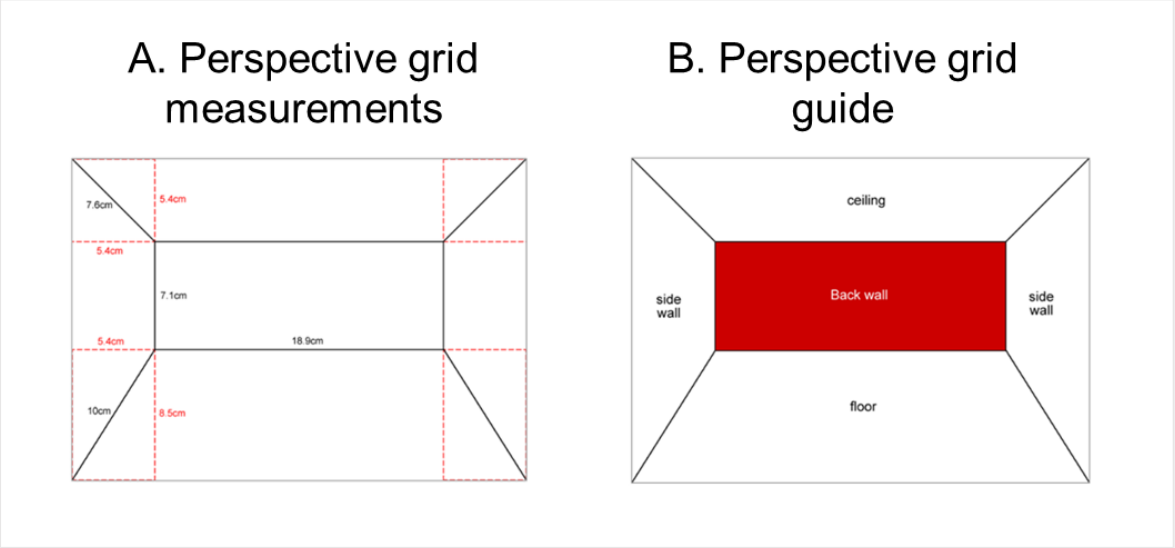
Vertically intact/ horizontally scrambled: Horizontally scrambled rotated

Vertically intact/ horizontally scrambled: Horizontal pieces rotated

Horizontally intact/ vertically scrambled: Vertically scrambled rotated

Horizontally intact/ vertically scrambled: Vertical pieces rotated

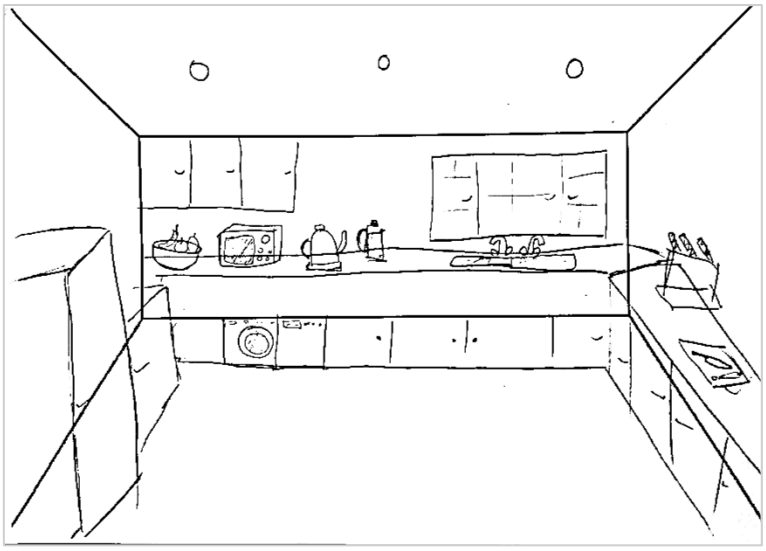
Appendix C: Perspective grids used in the drawing task utilised in chapter 3 experiment 1 and chapter 4 experiment 1 and 2

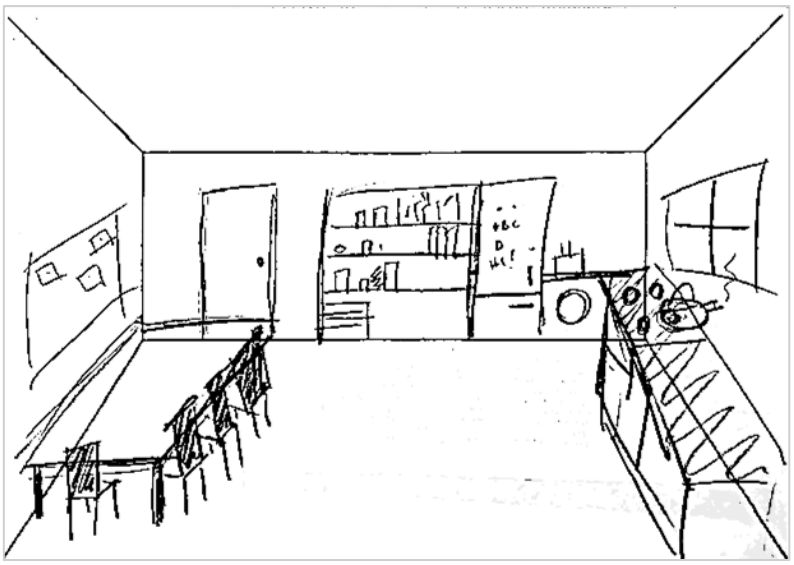


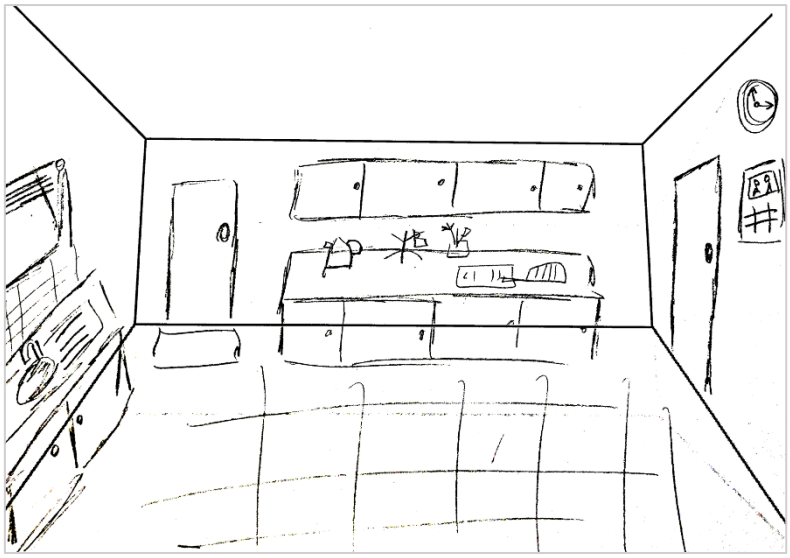
Examples of the layout of the perspective grids and the guidance on how to draw them. A. The measurements of the perspective grid used to draw indoor scenes (bedroom, kitchen and living room). B. The layout of the perspective grid. The back rectangle comprised the back wall of the room, the larger bottom segment the floor, and the side segments the left and right walls. The top section made up the rooms ceiling.

Appendix D: Further examples of scene drawings produced in chapter 3 and chapter 4

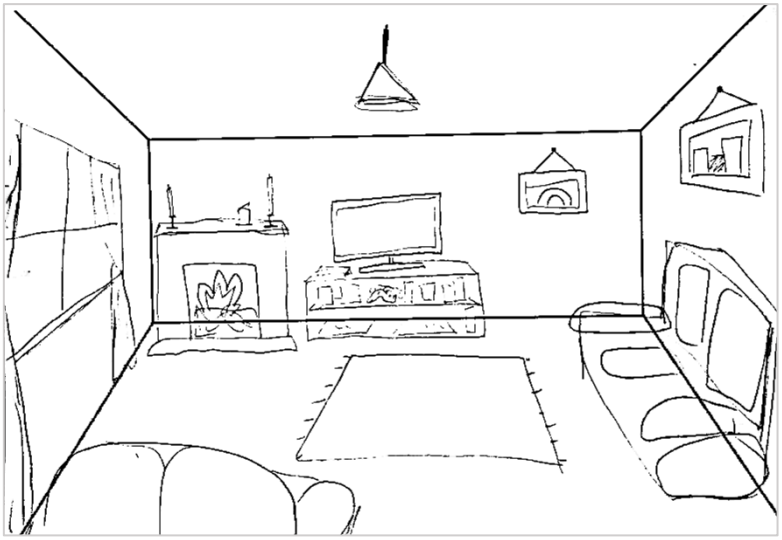
Participants own kitchen drawings



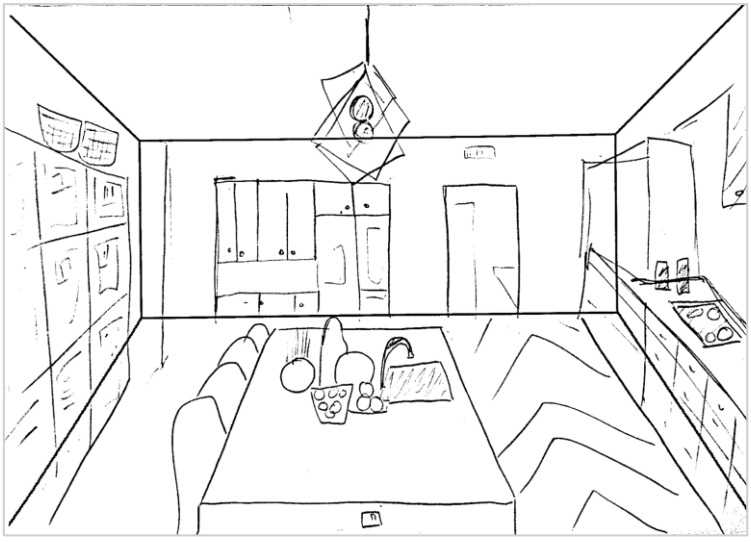


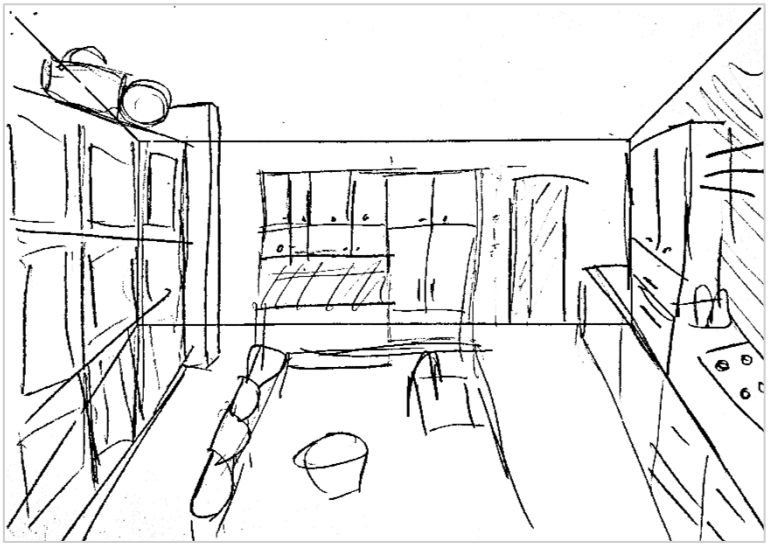


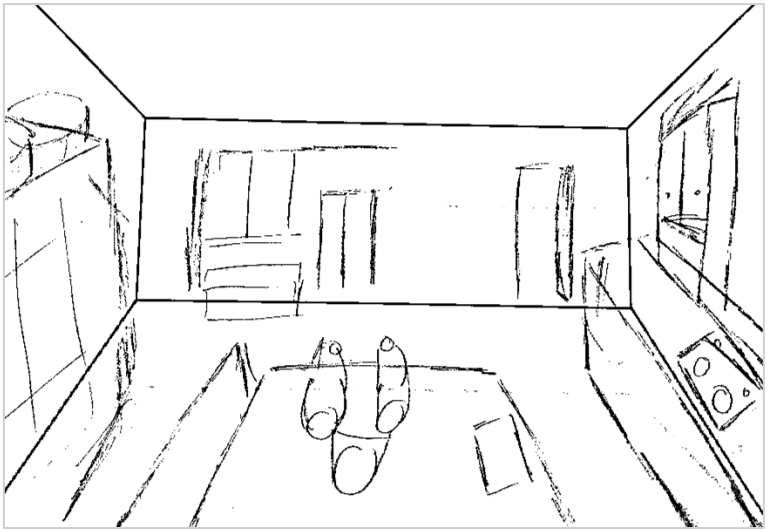
Participants own living room drawings



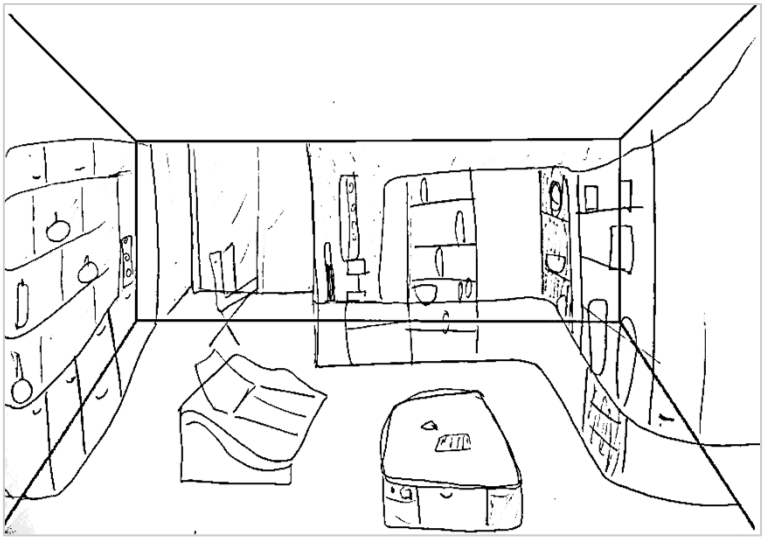
Participants copied kitchen drawings

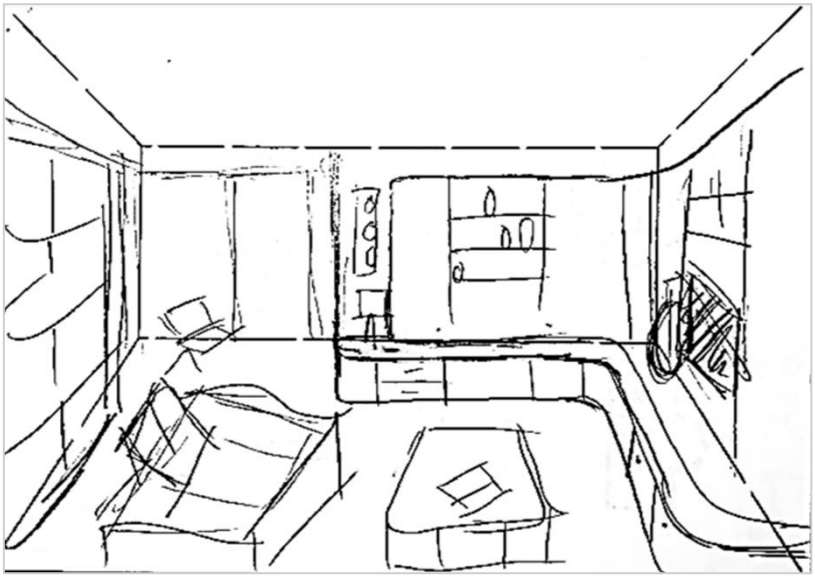


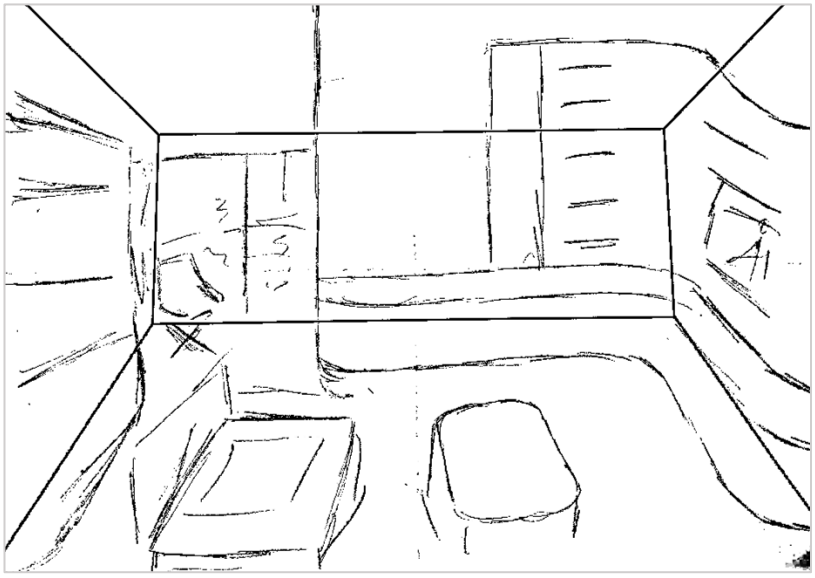




Participants copied living room drawings







Appendix E: Further examples of the stimuli used in chapter 3

Kitchen

Living room

Control kitchen

Control living room

Appendix F: Further examples of the stimuli used in chapter 4, experiment 1 and experiment 2

Kitchen: Whole


Kitchen: Swap 1




Kitchen: Swap 2


Kitchen: Replace 1

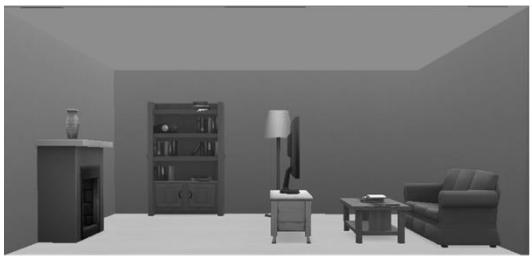



Kitchen: Replace 2

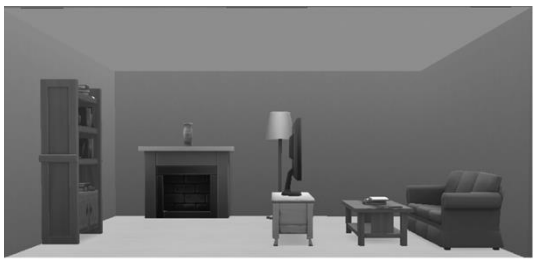

Living room: Whole


Living room: Swap 1

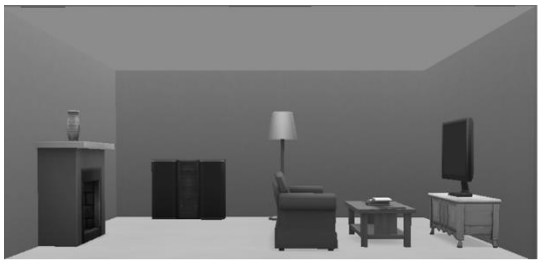


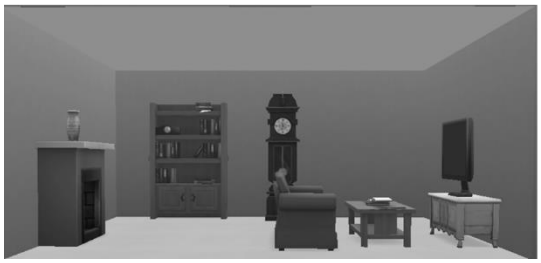



Living room: Swap 2

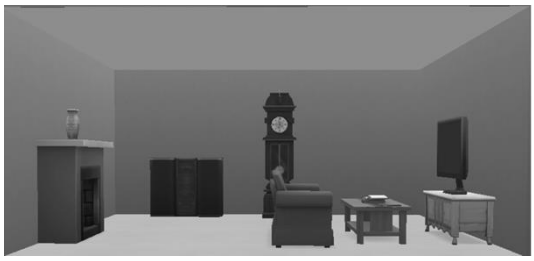


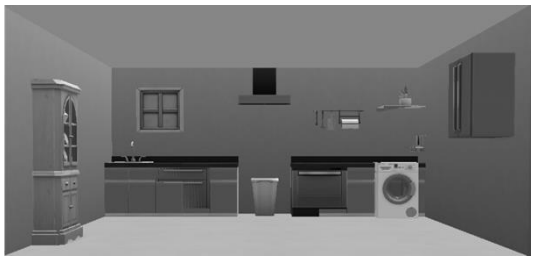

Living room: Replace 1






Living room: Replace 2





Appendix G: Table showing the conditions and scene categories of the trials included in the analysis for chapter 2 experiment 1 after exclusions

	Beach	**Desert**	**Field**	**Mountain**	**Total**
Whole, Upright	1504	1360	1500	1528	5892
Whole, Inverted	1400	1385	1478	1507	5770
Vertically Jumbled, Upright	1476	1383	1492	1507	5858
Vertically Jumbled, Inverted	1449	1329	1484	1496	5758
Horizontally Jumbled, Upright	1504	1347	1484	1504	5839
Horizontally jumbled, Inverted	1480	1354	1480	1494	5808
Fully Jumbled, Upright	1489	1332	1488	1500	5809
Fully Jumbled, Inverted	1519	1306	1484	1490	5799
Total	11821	10796	11890	12026	46533

Appendix H: Table showing the conditions and scene categories of the trials included in the analysis for chapter 2 experiment 2 after exclusions

	Beach	**Desert**	**Field**	**Mountain**	**Total**
Whole, Upright	2934	2788	3133	3138	11993
Whole, Rotated	3343	2786	3111	3145	12385
Vertically Jumbled, Upright	3132	2814	3133	3125	12204
Vertically Jumbled, Rotated	3075	2783	3144	3118	12120
Horizontally Jumbled, Upright	3170	2780	3128	3154	12232
Horizontally jumbled, Rotated	3115	2772	3116	3121	12124
Fully Jumbled, Upright	3129	2807	3139	3104	12179
Fully Jumbled, Rotated	3074	2763	3126	3118	12081
Total	24972	22293	25030	25023	97318

References

Adams, W. J., Elder, J. H., Graf, E. W., Leyland, J., Lugtigheid, A. J., & Muryy, A. (2016). The Southampton-York Natural Scenes (SYNS) dataset: Statistics of surface attitude. *Scientific Reports*, *6*, 35805. https://doi.org/10.1038/srep35805

Agrell, B., & Dehlin, O. (1998). The clock-drawing test. *Age and Ageing*, *27*(3), 399–404.

Alexander Varakin, D., & Levin, D. T. (2008). Short Article: Scene Structure Enhances Change Detection. *Quarterly Journal of Experimental Psychology*, *61*(4), 543–551. https://doi.org/10.1080/17470210701774176

Allen, J. W. P., & Bickhard, M. H. (2013). Stepping off the pendulum: Why only an action-based approach can transcend the nativist–empiricist debate. *Cognitive Development*, *28*(2), 96–133. https://doi.org/10.1016/j.cogdev.2013.01.002

Altman, I., & Low, S. M. (2012). *Place Attachment*. Springer Science & Business Media.

Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. https://doi.org/10.3758/s13428-019-01237-x

Apps, M. A. J., & Tsakiris, M. (2013). Predictive codes of familiarity and context during the perceptual learning of facial identities. *Nature Communications*, *4*(1), 2698. https://doi.org/10.1038/ncomms3698

Bacim, F., Ragan, E., Scerbo, S., Polys, N. F., Setareh, M., & Jones, B. D. (2013). *The Effects of Display Fidelity, Visual Complexity, and Task Scope on Spatial Understanding of 3D Graphs*.

Bainbridge, W. A. (2022). A tutorial on capturing mental representations through drawing and crowd-sourced scoring. *Behavior Research Methods*, *54*(2), 663–675. https://doi.org/10.3758/s13428-021-01672-9

Bainbridge, W. A., & Baker, C. I. (2020). Boundaries Extend and Contract in Scene Memory Depending on Image Properties. *Current Biology*, *30*(3), 537-543.e3. https://doi.org/10.1016/j.cub.2019.12.004

Bainbridge, W. A., & Baker, C. I. (2022). Multidimensional memory topography in the medial parietal cortex identified from neuroimaging of thousands of daily memory videos. *Nature Communications*, *13*(1), 6508. https://doi.org/10.1038/s41467-022-34075-1

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2019). Drawings of real-world scenes during free recall reveal detailed object and spatial information in memory. *Nature Communications*, *10*(1), 5. https://doi.org/10.1038/s41467-018-07830-6

Bainbridge, W. A., Hall, E. H., & Baker, C. I. (2021). Distinct Representational Structure and Localization for Visual Encoding and Recall during Visual Imagery. *Cerebral Cortex*, *31*(4), 1898–1913. https://doi.org/10.1093/cercor/bhaa329

Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. https://doi.org/10.1038/nrn1476

Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M. S., Marinkovic, K., Schacter, D. L., Rosen, B. R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the National Academy of Sciences*, *103*(2), 449–454. https://doi.org/10.1073/pnas.0507062103

Barrett, H. C. (2020). Towards a Cognitive Science of the Human: Cross-Cultural Approaches and Their Urgency. *Trends in Cognitive Sciences*, *24*(8), 620–638. https://doi.org/10.1016/j.tics.2020.05.007

Bartlett, S. F. C. (1932). *Remembering: A Study in Experimental and Social Psychology*. Cambridge University Press.

Bartolomeo, P. (2002). The relationship between visual perception and visual mental imagery: A reappraisal of the neuropsychological evidence. *Cortex; a Journal Devoted to the Study of the Nervous System and Behavior*, *38*(3), 357–378. https://doi.org/10.1016/s0010-9452(08)70665-8

Bartolomeo, P., Bachoud-Lévi, A.-C., De Gelder, B., Denes, G., Barba, G. D., Brugières, P., & Degos, J.-D. (1998). Multiple-domain dissociation between impaired visual perception and preserved

mental imagery in a patient with bilateral extrastriate lesions. *Neuropsychologia*, *36*(3), 239–249. https://doi.org/10.1016/S0028-3932(97)00103-6

Behrmann, M., Moscovitch, M., & Winocur, G. (1994). Intact visual imagery and impaired visual perception in a patient with visual agnosia. *Journal of Experimental Psychology. Human Perception and Performance*, *20*(5), 1068–1087. https://doi.org/10.1037//0096-1523.20.5.1068

Bekhit, N. S., Thomas, G. V., & Jolley, R. P. (2005). The use of drawing for psychological assessment in Britain: Survey findings. *Psychology and Psychotherapy: Theory, Research and Practice*, *78*(2), 205–217. https://doi.org/10.1348/147608305X26044

Bertamini, M., Silvanto, J., Norcia, A., Makin, A., & Wagemans, J. (2018). The neural basis of visual symmetry and its role in mid- and high-level visual processing: Neural basis of visual symmetry. *Annals of the New York Academy of Sciences*, *1426*. https://doi.org/10.1111/nyas.13667

Bertsch, S., Pesta, B. J., Wiscott, R., & McDaniel, M. A. (2007). The generation effect: A meta-analytic review. *Memory & Cognition*, *35*(2), 201–210. https://doi.org/10.3758/BF03193441

Biederman, I. (1972). Human performance in contingent information-processing tasks. *Journal of Experimental Psychology*, *93*(2), 219–238. https://doi.org/10.1037/h0032511

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, *97*(1), 22–27. https://doi.org/10.1037/h0033776

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-X

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, *103*(3), 597–600. https://doi.org/10.1037/h0037158

Bieniek, M. M., Frei, L. S., & Rousselet, G. A. (2013). Early ERPs to faces: Aging, luminance, and individual differences. *Frontiers in Psychology*, *4*, 268. https://doi.org/10.3389/fpsyg.2013.00268

Boettcher, S. E. P., Draschkow, D., Dienhart, E., & Võ, M. L.-H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision*, *18*(13), 11. https://doi.org/10.1167/18.13.11

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences*, *114*(18), 4793–4798. https://doi.org/10.1073/pnas.1618228114

Bonoti, F., & Metallidou, P. (2010). Children's judgments and feelings about their own drawings. *Psychology*, *1*(5), 329–336. https://doi.org/10.4236/psych.2010.15042

Brady, T. F., & Oliva, A. (2008). Statistical Learning Using Real-World Scenes: Extracting Categorical Regularities Without Conscious Intent. *Psychological Science*, *19*(7), 678–685. https://doi.org/10.1111/j.1467-9280.2008.02142.x

Brady, T. F., Shafer-Skelton, A., & Alvarez, G. A. (2017). Global ensemble texture representations are critical to rapid scene perception. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1160–1176. https://doi.org/10.1037/xhp0000399

Brandman, T., & Peelen, M. V. (2019). Signposts in the Fog: Objects Facilitate Scene Representations in Left Scene-selective Cortex. *Journal of Cognitive Neuroscience*, *31*(3), 390–400. https://doi.org/10.1162/jocn_a_01258

Brewer, W. F., & Treyens, J. C. (1981). Role of schemata in memory for places. *Cognitive Psychology*, *13*(2), 207–230. https://doi.org/10.1016/0010-0285(81)90008-6

Bridges, D., Pitiot, A., MacAskill, M. R., & Peirce, J. W. (2020). The timing mega-study: Comparing a range of experiment generators, both lab-based and online. *PeerJ*, *8*, e9414. https://doi.org/10.7717/peerj.9414

Brockmole, J. R., & Henderson, J. M. (2006). Using real-world scenes as contextual cues for search. *Visual Cognition*, *13*(1), 99–108. https://doi.org/10.1080/13506280500165188

Bullock, H. E. (2017). The widening economic divide: Economic disparities and classism as critical community context. In *APA handbook of community psychology: Theoretical foundations, core concepts, and emerging challenges, Vol. 1* (pp. 353–368). American Psychological Association. https://doi.org/10.1037/14953-017

Caddigan, E., Choo, H., Fei-Fei, L., & Beck, D. M. (2017). Categorization influences detection: A perceptual advantage for representative exemplars of natural scene categories. *Journal of Vision*, *17*(1), 21. https://doi.org/10.1167/17.1.21

Castelhano, M. S., & Henderson, J. M. (2008). The influence of color on the perception of scene gist. *Journal of Experimental Psychology: Human Perception and Performance*, *34*(3), 660–675. https://doi.org/10.1037/0096-1523.34.3.660

Cavanagh, P. (2005). The artist as neuroscientist. *Nature*, *434*(7031), 301–307. https://doi.org/10.1038/434301a

Charest, I., Kievit, R. A., Schmitz, T. W., Deca, D., & Kriegeskorte, N. (2014). Unique semantic space in the brain of each beholder predicts perceived similarity. *Proceedings of the National Academy of Sciences*, *111*(40), 14565–14570. https://doi.org/10.1073/pnas.1402594111

Chen, L., Cichy, R. M., & Kaiser, D. (2022). Semantic scene-object consistency modulates N300/400 EEG components, but does not automatically facilitate object representations. *Cerebral Cortex*, *32*(16), 3553–3567. https://doi.org/10.1093/cercor/bhab433

Christou, C. G., & Bülthoff, H. H. (1999). View dependence in scene recognition after active learning. *Memory & Cognition*, *27*(6), 996–1007. https://doi.org/10.3758/BF03201230

Cichy, R. M., & Kaiser, D. (2019). Deep Neural Networks as Scientific Models. *Trends in Cognitive Sciences*, *23*(4), 305–317. https://doi.org/10.1016/j.tics.2019.01.009

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*(3), 181–204. https://doi.org/10.1017/S0140525X12000477

Clark, S. E. (1995). The generation effect and the modeling of associations in memory. *Memory & Cognition*, *23*(4), 442–455. https://doi.org/10.3758/BF03197245

Coco, M. I., Keller, F., & Malcolm, G. L. (2016). Anticipation in Real-World Scenes: The Role of Visual Context and Visual Memory. *Cognitive Science*, *40*(8), 1995–2024. https://doi.org/10.1111/cogs.12313

Coutrot, A., Manley, E., Goodroe, S., Gahnstrom, C., Filomena, G., Yesiltepe, D., Dalton, R. C., Wiener, J. M., Hölscher, C., Hornberger, M., & Spiers, H. J. (2022). Entropy of city street networks linked to future spatial navigation ability. *Nature*, *604*(7904), 104–110. https://doi.org/10.1038/s41586-022-04486-7

Crump, M. J. C., McDonnell, J. V., & Gureckis, T. M. (2013). Evaluating Amazon's Mechanical Turk as a Tool for Experimental Behavioral Research. *PLOS ONE*, *8*(3), e57410. https://doi.org/10.1371/journal.pone.0057410

Csathó, Á., van der Linden, D., & Gács, B. (2015). Natural Scene Recognition with Increasing Time-On-Task: The Role of Typicality and Global Image Properties. *Quarterly Journal of Experimental Psychology*, *68*(4), 814–828. https://doi.org/10.1080/17470218.2014.968592

Dalecki, M., Hoffmann, U., & Bock, O. (2012). Mental rotation of letters, body parts and complex scenes: Separate or common mechanisms? *Human Movement Science*, *31*(5), 1151–1160. https://doi.org/10.1016/j.humov.2011.12.001

Dalton, M. A., & Maguire, E. A. (2017). The pre/parasubiculum: A hippocampal hub for scene-based cognition? *Current Opinion in Behavioral Sciences*, *17*, 34–40. https://doi.org/10.1016/j.cobeha.2017.06.001

Davenport, J. L., & Potter, M. C. (2004). Scene Consistency in Object and Background Perception. *Psychological Science*, *15*(8), 559–564. https://doi.org/10.1111/j.0956-7976.2004.00719.x

de Chastelaine, M., Mattson, J. T., Wang, T. H., Donley, B. E., & Rugg, M. D. (2017). Independent contributions of fMRI familiarity and novelty effects to recognition memory and their stability across the adult lifespan. *NeuroImage*, *156*, 340–351. https://doi.org/10.1016/j.neuroimage.2017.05.039

de Fockert, J. W. de, Caparos, S., Linnell, K. J., & Davidoff, J. (2011). Reduced Distractibility in a Remote Culture. *PLOS ONE*, *6*(10), e26337. https://doi.org/10.1371/journal.pone.0026337

de Fockert, J., Davidoff, J., Fagot, J., Parron, C., & Goldstein, J. (2007). More accurate size contrast judgments in the Ebbinghaus Illusion by a remote culture. *Journal of Experimental Psychology. Human Perception and Performance*, *33*(3), 738–742. https://doi.org/10.1037/0096-1523.33.3.738

de Haas, B., Iakovidis, A. L., Schwarzkopf, D. S., & Gegenfurtner, K. R. (2019). Individual differences in visual salience vary along semantic dimensions. *Proceedings of the National Academy of Sciences*, *116*(24), 11687–11692. https://doi.org/10.1073/pnas.1820553116

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). ImageNet: A large-scale hierarchical image database. *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 248–255. https://doi.org/10.1109/CVPR.2009.5206848

Dillon, M. R., Persichetti, A. S., Spelke, E. S., & Dilks, D. D. (2018). Places in the Brain: Bridging Layout and Object Geometry in Scene-Selective Cortex. *Cerebral Cortex*, *28*(7), 2365–2374. https://doi.org/10.1093/cercor/bhx139

Dima, D. C., Perry, G., & Singh, K. D. (2018). Spatial frequency supports the emergence of categorical representations in visual cortex during natural scene perception. *NeuroImage*, *179*, 102–116. https://doi.org/10.1016/j.neuroimage.2018.06.033

Dosher, B., & Lu, Z.-L. (2017). *Visual Perceptual Learning and Models | Annual Reviews*. https://www.annualreviews.org/content/journals/10.1146/annurev-vision-102016-061249#f1

Draschkow, D., Heikel, E., Võ, M. L.-H., Fiebach, C. J., & Sassenhagen, J. (2018). No evidence from MVPA for different processes underlying the N300 and N400 incongruity effects in object-scene processing. *Neuropsychologia*, *120*, 9–17. https://doi.org/10.1016/j.neuropsychologia.2018.09.016

Draschkow, D., & Võ, M. L.-H. (2017). Scene grammar shapes the way we interact with objects, strengthens memories, and speeds search. *Scientific Reports*, *7*(1), 16471. https://doi.org/10.1038/s41598-017-16739-x

Dzulkifli, M. A., & Mustafar, M. F. (2013). The Influence of Colour on Memory Performance: A Review. *The Malaysian Journal of Medical Sciences : MJMS*, *20*(2), 3–9.

Edquist, J., & Johnston, I. (2008). *Visual Clutter in Road Environments—What It Does, and What to Do About It*.

Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*(6–7), 945–978. https://doi.org/10.1080/13506280902834720

Epstein, R. (2005). The cortical basis of visual scene processing. *Visual Cognition*, *12*(6), 954–978. https://doi.org/10.1080/13506280444000607

Epstein, R. A., & Baker, C. I. (2019). Scene Perception in the Human Brain. *Annual Review of Vision Science*, *5*(Volume 5, 2019), 373–397. https://doi.org/10.1146/annurev-vision-091718-014809

Epstein, R. A., Higgins, J. S., Jablonski, K., & Feiler, A. M. (2007). Visual Scene Processing in Familiar and Unfamiliar Environments. *Journal of Neurophysiology*, *97*(5), 3670–3683. https://doi.org/10.1152/jn.00003.2007

Epstein, R. A., & Morgan, L. K. (2012). Neural responses to visual scenes reveals inconsistencies between fMRI adaptation and multivoxel pattern analysis. *Neuropsychologia*, *50*(4), 530–543. https://doi.org/10.1016/j.neuropsychologia.2011.09.042

Epstein, R. A., Parker, W. E., & Feiler, A. M. (2007). Where Am I Now? Distinct Roles for Parahippocampal and Retrosplenial Cortices in Place Recognition. *Journal of Neuroscience*, *27*(23), 6141–6149. https://doi.org/10.1523/JNEUROSCI.0799-07.2007

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The Parahippocampal Place Area: Recognition, Navigation, or Encoding? *Neuron*, *23*(1), 115–125. https://doi.org/10.1016/S0896-6273(00)80758-8

Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*(6676), 598–601. https://doi.org/10.1038/33402

Essock, E. A., DeFord, J. K., Hansen, B. C., & Sinai, M. J. (2003). Oblique stimuli are seen best (not worst!) in naturalistic broad-band stimuli: A horizontal effect. *Vision Research*, *43*(12), 1329–1335. https://doi.org/10.1016/S0042-6989(03)00142-1

Evans, K., & Wolfe, J. (2022). Sometimes it helps to be taken out of context: Memory for objects in scenes. *Visual Cognition*, *30*, 1–16. https://doi.org/10.1080/13506285.2021.2023245

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of Cognitive Neuroscience*, *13*(2), 171–180. https://doi.org/10.1162/089892901564234

Fan, J. E., Bainbridge, W. A., Chamberlain, R., & Wammes, J. D. (2023). Drawing as a versatile cognitive tool. *Nature Reviews Psychology*, *2*(9), 556–568. https://doi.org/10.1038/s44159-023-00212-w

Fan, J. E., Yamins, D. L. K., & Turk-Browne, N. B. (2018). Common Object Representations for Visual Production and Recognition. *Cognitive Science*, *42*(8), 2670–2698. https://doi.org/10.1111/cogs.12676

Feldman, H., & Friston, K. (2010). Attention, Uncertainty, and Free-Energy. *Frontiers in Human Neuroscience*, *4*. https://doi.org/10.3389/fnhum.2010.00215

Ferber, S., Mraz, R., Baker, N., & Graham, S. J. (2007). Shared and differential neural substrates of copying versus drawing: A functional magnetic resonance imaging study. *Neuroreport*, *18*(11), 1089–1093. https://doi.org/10.1097/WNR.0b013e3281ac2143

Fernandes, M. A., Wammes, J. D., & Meade, M. E. (2018). The Surprisingly Powerful Influence of Drawing on Memory. *Current Directions in Psychological Science*, *27*(5), 302–308. https://doi.org/10.1177/0963721418755385

Ferrara, K., & Park, S. (2016). Neural representation of scene boundaries. *Neuropsychologia*, *89*, 180–190. https://doi.org/10.1016/j.neuropsychologia.2016.05.012

Fischer, J., Mikhael, J. G., Tenenbaum, J. B., & Kanwisher, N. (2016). Functional neuroanatomy of intuitive physical inference. *Proceedings of the National Academy of Sciences*, *113*(34), E5072–E5081. https://doi.org/10.1073/pnas.1610344113

Friedman, D., de Chastelaine, M., Nessler, D., & Malcolm, B. (2010). Changes in familiarity and recollection across the lifespan: An ERP perspective. *Brain Research*, *1310*, 124–141. https://doi.org/10.1016/j.brainres.2009.11.016

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *360*(1456), 815–836. https://doi.org/10.1098/rstb.2005.1622

Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787

Furtak, M., Mudrik, L., & Bola, M. (2022). The forest, the trees, or both? Hierarchy and interactions between gist and object processing during perception of real-world scenes. *Cognition*, *221*, 104983. https://doi.org/10.1016/j.cognition.2021.104983

Ganesan, A., & Balasubramanian, A. (2019). Indoor versus outdoor scene recognition for navigation of a micro aerial vehicle using spatial color gist wavelet descriptors. *Visual Computing for Industry, Biomedicine, and Art*, *2*(1), 20. https://doi.org/10.1186/s42492-019-0030-9

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, *16*(2), 123–144. https://doi.org/10.1016/S0926-6410(02)00244-6

Gauthier, I. (2018). Domain-Specific and Domain-General Individual Differences in Visual Object Recognition. *Current Directions in Psychological Science*, *27*(2), 97–102. https://doi.org/10.1177/0963721417737151

Geisler, W. S. (2008). Visual Perception and the Statistical Properties of Natural Scenes. *Annual Review of Psychology*, *59*(Volume 59, 2008), 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632

Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P., & Rossion, B. (2005). Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual Cognition*, *12*(6), 878–892. https://doi.org/10.1080/13506280444000562

Greene, M. R., & Oliva, A. (2009a). Recognition of natural scenes from global properties: Seeing the forest without representing the trees. *Cognitive Psychology*, *58*(2), 137–176. https://doi.org/10.1016/j.cogpsych.2008.06.001

Greene, M. R., & Oliva, A. (2009b). The Briefest of Glances: The Time Course of Natural Scene Understanding. *Psychological Science*, *20*(4), 464–472.

Grill-Spector, K. (2004). The Functional Organization of the Ventral Visual Pathway and its Relationship to object Recognition. In N. Kanwisher & J. Duncan (Eds.), *Functional Neuroimaging of Visual Cognition* (pp. 169–193). Oxford University PressOxford. https://doi.org/10.1093/oso/9780198528456.003.0008

Groen, I. I. A., Ghebreab, S., Prins, H., Lamme, V. A. F., & Scholte, H. S. (2013). From Image Statistics to Scene Gist: Evoked Neural Activity Reveals Transition from Low-Level Natural Image Structure to Scene Category. *Journal of Neuroscience*, *33*(48), 18814–18824. https://doi.org/10.1523/JNEUROSCI.3128-13.2013

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160102. https://doi.org/10.1098/rstb.2016.0102

Gronau, N., & Shachar, M. (2015). Contextual consistency facilitates long-term memory of perceptual detail in barely seen images. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(4), 1095–1111. https://doi.org/10.1037/xhp0000071

Hagoort, P., Brown, C., & Groothusen, J. (1993). The syntactic positive shift (sps) as an erp measure of syntactic processing. *Language and Cognitive Processes*, *8*(4), 439–483. https://doi.org/10.1080/01690969308407585

Handali, J. P., Schneider, J., Gau, M., Holzwarth, V., & Brocke, J. vom. (2021). Visual Complexity and Scene Recognition: How Low Can You Go? *2021 IEEE Virtual Reality and 3D User Interfaces (VR)*, 286–295. https://doi.org/10.1109/VR50410.2021.00051

Hansen, B. C., & Essock, E. A. (2004). A horizontal bias in human visual processing of orientation and its correspondence to the structural components of natural scenes. *Journal of Vision*, *4*(12), 5. https://doi.org/10.1167/4.12.5

Hansen, B. C., & Hess, R. F. (2007). Structural sparseness and spatial phase alignment in natural scenes. *Journal of the Optical Society of America. A, Optics, Image Science, and Vision*, *24*(7), 1873–1885. https://doi.org/10.1364/josaa.24.001873

Hansen, B., Haun, A., & Essock, E. (2008). *The "Horizontal Effect": A perceptual anisotropy in visual processing of naturalistic broadband stimuli*.

Hartley, C. A. (2022). How do natural environments shape adaptive cognition across the lifespan? *Trends in Cognitive Sciences*, *26*(12), 1029–1030. https://doi.org/10.1016/j.tics.2022.10.002

Henderson, J. M. (2017). Gaze Control as Prediction. *Trends in Cognitive Sciences*, *21*(1), 15–23. https://doi.org/10.1016/j.tics.2016.11.003

Henderson, J. M., Larson, C. L., & Zhu, D. C. (2008). Full Scenes produce more activation than Close-up Scenes and Scene-Diagnostic Objects in parahippocampal and retrosplenial cortex: An fMRI study. *Brain and Cognition*, *66*(1), 40–49. https://doi.org/10.1016/j.bandc.2007.05.001

Henderson, J. M., & Luke, S. G. (2014). Stable individual differences in saccadic eye movements during reading, pseudoreading, scene viewing, and scene search. *Journal of Experimental*

Psychology: Human Perception and Performance, *40*(4), 1390–1400. https://doi.org/10.1037/a0036330

Henderson, J. M., Zhu, D. C., & Larson, C. L. (2011). Functions of parahippocampal place area and retrosplenial cortex in real-world scene analysis: An fMRI study. *Visual Cognition*, *19*(7), 910–927. https://doi.org/10.1080/13506285.2011.596852

Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid Invariant Encoding of Scene Layout in Human OPA. *Neuron*, *103*(1), 161-171.e3. https://doi.org/10.1016/j.neuron.2019.04.014

Herdtweck, C., & Wallraven, C. (2013). Estimation of the Horizon in Photographed Outdoor Scenes by Human and Machine. *PLOS ONE*, *8*(12), e81462. https://doi.org/10.1371/journal.pone.0081462

Herrmann, N., Kidron, D., Shulman, K. I., Kaplan, E., Binns, M., Leach, L., & Freedman, M. (1998). Clock Tests in Depression, Alzheimer's Disease, and Elderly Controls. *The International Journal of Psychiatry in Medicine*, *28*(4), 437–447. https://doi.org/10.2190/5QA5-PHUN-1Q9F-C0PB

Hill, H., & Bruce, V. (1993). Independent effects of lighting, orientation, and stereopsis on the hollow-face illusion. *Perception*, *22*(8), 887–897. https://doi.org/10.1068/p220887

Hillstrom, A. P., Scholey, H., Liversedge, S. P., & Benson, V. (2012). The effect of the first glimpse at a scene on eye movements during search. *Psychonomic Bulletin & Review*, *19*(2), 204–210. https://doi.org/10.3758/s13423-011-0205-7

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

Holcomb, P. J. (1993). Semantic priming and stimulus degradation: Implications for the role of the N400 in language processing. *Psychophysiology*, *30*(1), 47–61. https://doi.org/10.1111/j.1469-8986.1993.tb03204.x

Homa, D., & Viera, C. (1988). Long-term memory for pictures under conditions of thematically related foils. *Memory & Cognition*, *16*(5), 411–421. https://doi.org/10.3758/BF03214221

Hubbard, T. L., Hutchison, J. L., & Courtney, J. R. (2010). Boundary extension: Findings and theories. *Quarterly Journal of Experimental Psychology*, *63*(8), 1467–1494. https://doi.org/10.1080/17470210903511236

Hudson, W. (1960). Pictorial Depth Perception in Sub-Cultural Groups in Africa. *The Journal of Social Psychology*. https://www.tandfonline.com/doi/abs/10.1080/00224545.1960.9922077

Intraub, H., & Bodamer, J. L. (1993). Boundary extension: Fundamental aspect of pictorial representation or encoding artifact? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*(6), 1387–1397. https://doi.org/10.1037/0278-7393.19.6.1387

James, T. W., Culham, J., Humphrey, G. K., Milner, A. D., & Goodale, M. A. (2003). Ventral occipital lesions impair object recognition but not object-directed grasping: An fMRI study. *Brain*, *126*(11), 2463–2475. https://doi.org/10.1093/brain/awg248

Jongejan, R., Ranasinghe, R., Wainwright, D., Callaghan, D. P., & Reyns, J. (2016). Drawing the line on coastline recession risk. *Ocean & Coastal Management*, *122*, 87–94. https://doi.org/10.1016/j.ocecoaman.2016.01.006

Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision Research*, *47*(26), 3286–3297. https://doi.org/10.1016/j.visres.2007.09.013

Kaiser, D., & Cichy, R. M. (2018). Typical visual-field locations facilitate access to awareness for everyday objects. *Cognition*, *180*, 118–122. https://doi.org/10.1016/j.cognition.2018.07.009

Kaiser, D., & Cichy, R. M. (2021). Parts and Wholes in Scene Processing. *Journal of Cognitive Neuroscience*, *34*(1), 4–15. https://doi.org/10.1162/jocn_a_01788

Kaiser, D., Häberle, G., & Cichy, R. M. (2020a). Cortical sensitivity to natural scene structure. *Human Brain Mapping*, *41*(5), 1286–1295. https://doi.org/10.1002/hbm.24875

Kaiser, D., Häberle, G., & Cichy, R. M. (2020b). Real-world structure facilitates the rapid emergence of scene category information in visual brain signals. *Journal of Neurophysiology*, *124*(1), 145–151. https://doi.org/10.1152/jn.00164.2020

Kaiser, D., Inciuraite, G., & Cichy, R. M. (2020). Rapid contextualization of fragmented scene information in the human visual system. *NeuroImage*, *219*, 117045. https://doi.org/10.1016/j.neuroimage.2020.117045

Kaiser, D., Moeskops, M. M., & Cichy, R. M. (2018). Typical retinotopic locations impact the time course of object coding. *NeuroImage*, *176*, 372–379. https://doi.org/10.1016/j.neuroimage.2018.05.006

Kaiser, D., & Peelen, M. V. (2018). Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *NeuroImage*, *169*, 334–341. https://doi.org/10.1016/j.neuroimage.2017.12.065

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object Vision in a Structured World. *Trends in Cognitive Sciences*, *23*(8), 672–685. https://doi.org/10.1016/j.tics.2019.04.013

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

Kaiser, D., Stein, T., & Peelen, M. V. (2015). Real-world spatial regularities affect visual working memory for objects. *Psychonomic Bulletin & Review*, *22*(6), 1784–1790. https://doi.org/10.3758/s13423-015-0833-4

Kaiser, D., Turini, J., & Cichy, R. M. (2019). A neural mechanism for contextualizing fragmented inputs during naturalistic vision. *eLife*, *8*, e48182. https://doi.org/10.7554/eLife.48182

Kamps, F. S., Julian, J. B., Kubilius, J., Kanwisher, N., & Dilks, D. D. (2016). The occipital place area represents the local elements of scenes. *NeuroImage*, *132*, 417–424. https://doi.org/10.1016/j.neuroimage.2016.02.062

Kanai, R., & Rees, G. (2011). The structural basis of inter-individual differences in human behaviour and cognition. *Nature Reviews Neuroscience*, *12*(4), 231–242. https://doi.org/10.1038/nrn3000

Kanan, C., Tong, M. H., Zhang, L., & Cottrell, G. W. (2009). SUN: Top-down saliency using natural statistics. *Visual Cognition*, *17*(6–7), 979–1003. https://doi.org/10.1080/13506280902771138

Kanizsa, W., & Gerbino, W. (1976). *Convexity and Symmetry in Figure-Ground Organization*. https://www.semanticscholar.org/paper/Convexity-and-Symmetry-in-Figure-Ground-Kanizsa-Gerbino/d936cb884979471b31805368fcbb8ecbbfa8502d

Katz, J. S., Forloines, M. R., Strassberg, L. R., & Bondy, B. (2021). Observational drawing in the brain: A longitudinal exploratory fMRI study. *Neuropsychologia*, *160*, 107960. https://doi.org/10.1016/j.neuropsychologia.2021.107960

Kauffmann, L., Ramanoël, S., Guyader, N., Chauvin, A., & Peyrin, C. (2015). Spatial frequency processing in scene-selective cortical regions. *NeuroImage*, *112*, 86–95. https://doi.org/10.1016/j.neuroimage.2015.02.058

Kayser, C., Körding, K. P., & König, P. (2004). Processing of complex stimuli and natural scenes in the visual cortex. *Current Opinion in Neurobiology*, *14*(4), 468–473. https://doi.org/10.1016/j.conb.2004.06.002

Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive Processing: A Canonical Cortical Computation. *Neuron*, *100*(2), 424–435. https://doi.org/10.1016/j.neuron.2018.10.003

Kelley, T. A., Chun, M. M., & Chua, K.-P. (2003). Effects of scene inversion on change detection of targets matched for visual salience. *Journal of Vision*, *3*(1), 1. https://doi.org/10.1167/3.1.1

Kersten, D. (1997). Perceptual categories for spatial layout. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *352*(1358), 1155–1163.

Klink, H., Kaiser, D., Stecher, R., Ambrus, G. G., & Kovács, G. (2023). Your place or mine? The neural dynamics of personally familiar scene recognition suggests category independent familiarity encoding. *Cerebral Cortex*, *33*(24), 11634–11645. https://doi.org/10.1093/cercor/bhad397

Koehler, K., & Eckstein, M. P. (2017). Beyond scene gist: Objects guide search more than scene background. *Journal of Experimental Psychology: Human Perception and Performance*, *43*(6), 1177–1193. https://doi.org/10.1037/xhp0000363

Koen, J. D., & Yonelinas, A. P. (2016). Recollection, not familiarity, decreases in healthy ageing: Converging evidence from four estimation methods. *Memory (Hove, England)*, *24*(1), 75–88. https://doi.org/10.1080/09658211.2014.985590

Kominsky, J. F., Lucca, K., Thomas, A. J., Frank, M. C., & Hamlin, J. K. (2022). Simplicity and validity in infant research. *Cognitive Development*, *63*, 101213. https://doi.org/10.1016/j.cogdev.2022.101213

Konkle, T., Brady, Alvarez, & Oliva. (2012). 'Massive Memory' Scene Categories. *Psychological Science*. https://konklab.fas.harvard.edu/#

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene Memory Is More Detailed Than You Think: The Role of Categories in Visual Long-Term Memory. *Psychological Science*, *21*(11), 1551–1556. https://doi.org/10.1177/0956797610385359

Kornblith, S., Cheng, X., Ohayon, S., & Tsao, D. Y. (2013). A network for scene processing in the macaque temporal lobe. *Neuron*, *79*(4), 766–781. https://doi.org/10.1016/j.neuron.2013.06.015

Köster, M., Kayhan, E., Langeloh, M., & Hoehl, S. (2020). Making Sense of the World: Infant Learning From a Predictive Processing Perspective. *Perspectives on Psychological Science*, *15*(3), 562–571. https://doi.org/10.1177/1745691619895071

Kriegeskorte, N. (2015). Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing. *Annual Review of Vision Science*, *1*(Volume 1, 2015), 417–446. https://doi.org/10.1146/annurev-vision-082114-035447

Kutas, M., & Federmeier, K. D. (2011). Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP). *Annual Review of Psychology*, *62*(Volume 62, 2011), 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Kutas, M., & Hillyard, S. A. (1980). Event-related brain potentials to semantically inappropriate and surprisingly large words. *Biological Psychology*, *11*(2), 99–116. https://doi.org/10.1016/0301-0511(80)90046-0

Kutas, M., Van Petten, C. K., & Kluender, R. (2006). Chapter 17—Psycholinguistics Electrified II (1994–2005). In M. J. Traxler & M. A. Gernsbacher (Eds.), *Handbook of Psycholinguistics (Second Edition)* (pp. 659–724). Academic Press. https://doi.org/10.1016/B978-012369374-7/50018-3

Kyle-Davidson, C., Bors, A., & Evans, K. (2022). *Predicting Human Perception of Scene Complexity*.

Kyle-Davidson, C., & Evans, K. K. (2023). Complexity & Memorability have a Nonlinear Relationship when Remembering Scenes. *Journal of Vision*, *23*(9), 5251. https://doi.org/10.1167/jov.23.9.5251

Kyle-Davidson, C., Zhou, E. Y., Walther, D. B., Bors, A. G., & Evans, K. K. (2023). Characterising and dissecting human perception of scene complexity. *Cognition*, *231*, 105319. https://doi.org/10.1016/j.cognition.2022.105319

Lambon Ralph, M. A., & Patterson, K. (2008). Generalization and Differentiation in Semantic Memory. *Annals of the New York Academy of Sciences*, *1124*(1), 61–76. https://doi.org/10.1196/annals.1440.006

Lamm, B., Gudi, H., Freitag, C., Teubert, M., Graf, F., Fassbender, I., Schwarzer, G., Lohaus, A., Knopf, M., & Keller, H. (2014). Mother–Infant Interactions at Home and in a Laboratory Setting: A Comparative Analysis in Two Cultural Contexts. *Journal of Cross-Cultural Psychology*, *45*(6), 843–852. https://doi.org/10.1177/0022022114532357

Lauer, T., Willenbockel, V., Maffongelli, L., & Võ, M. L.-H. (2020). The influence of scene and object orientation on the scene consistency effect. *Behavioural Brain Research*, *394*, 112812. https://doi.org/10.1016/j.bbr.2020.112812

Lee, D., & Quessy, S. (2002). Scene familiarity facilitates visual search in monkeys. *Journal of Vision*, *2*(7), 531. https://doi.org/10.1167/2.7.531

Lee, J., & Geng, J. J. (2017). Idiosyncratic Patterns of Representational Similarity in Prefrontal Cortex Predict Attentional Performance. *Journal of Neuroscience*, *37*(5), 1257–1268. https://doi.org/10.1523/JNEUROSCI.1407-16.2016

Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center–periphery organization of human object areas. *Nature Neuroscience*, *4*(5), 533–539. https://doi.org/10.1038/87490

Lewicka, M. (2011). Place attachment: How far have we come in the last 40 years? *Journal of Environmental Psychology*, *31*(3), 207–230. https://doi.org/10.1016/j.jenvp.2010.10.001

Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, *99*(14), 9596–9601. https://doi.org/10.1073/pnas.092277599

Linsley, D., & MacEvoy, S. P. (2014). Evidence for participation by object-selective visual cortex in scene category judgments. *Journal of Vision*, *14*(9), 19. https://doi.org/10.1167/14.9.19

Linsley, D., & MacEvoy, S. P. (2015). Encoding-Stage Crosstalk Between Object- and Spatial Property-Based Scene Processing Pathways. *Cerebral Cortex*, *25*(8), 2267–2281. https://doi.org/10.1093/cercor/bhu034

Llera, A., Wolfers, T., Mulders, P., & Beckmann, C. F. (2019). Inter-individual differences in human brain structure and morphology link to variation in demographics and behavior. *eLife*, *8*, e44443. https://doi.org/10.7554/eLife.44443

Long, B., Fan, J. E., Huey, H., Chai, Z., & Frank, M. C. (2024). Parallel developmental changes in children's production and recognition of line drawings of visual concepts. *Nature Communications*, *15*(1), 1191. https://doi.org/10.1038/s41467-023-44529-9

Loschky, L. C., Ringer, R. V., Ellis, K., & Hansen, B. C. (2015). Comparing rapid scene categorization of aerial and terrestrial views: A new perspective on scene gist. *Journal of Vision*, *15*(6), 11. https://doi.org/10.1167/15.6.11

Lowe, M. X., Rajsic, J., Ferber, S., & Walther, D. B. (2018). Discriminating scene categories from brain activity within 100 milliseconds. *Cortex*, *106*, 275–287. https://doi.org/10.1016/j.cortex.2018.06.006

Lowe, M. X., Rajsic, J., Gallivan, J. P., Ferber, S., & Cant, J. S. (2017). Neural representation of geometry and surface properties in object and scene perception. *NeuroImage*, *157*, 586–597. https://doi.org/10.1016/j.neuroimage.2017.06.043

MacEvoy, S. P., & Epstein, R. A. (2009). Decoding the Representation of Multiple Simultaneous Objects in Human Occipitotemporal Cortex. *Current Biology*, *19*(11), 943–947. https://doi.org/10.1016/j.cub.2009.04.020

MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience*, *14*(10), 1323–1329. https://doi.org/10.1038/nn.2903

Makuuchi, M., Kaminaga, T., & Sugishita, M. (2003). Both parietal lobes are involved in drawing: A functional MRI study and implications for constructional apraxia. *Cognitive Brain Research*, *16*(3), 338–347. https://doi.org/10.1016/S0926-6410(02)00302-6

Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jiang, H., Kennedy, W. A., Ledden, P. J., Brady, T. J., Rosen, B. R., & Tootell, R. B. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135–8139. https://doi.org/10.1073/pnas.92.18.8135

Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making Sense of Real-World Scenes. *Trends in Cognitive Sciences*, *20*(11), 843–856. https://doi.org/10.1016/j.tics.2016.09.003

Malcolm, G. L., & Henderson, J. M. (2010). Combining top-down processes to guide eye movements during real-world scene search. *Journal of Vision*, *10*(2), 4. https://doi.org/10.1167/10.2.4

Malcolm, G. L., Nuthmann, A., & Schyns, P. G. (2012). Beyond Gist: Diagnostic Information Changes with Level of Scene Categorization. *Journal of Vision*, *12*(9), 800. https://doi.org/10.1167/12.9.800

Mamassian, P., & Landy, M. S. (1998). Observer biases in the 3D interpretation of line drawings. *Vision Research*, *38*(18), 2817–2832. https://doi.org/10.1016/S0042-6989(97)00438-0

Mandler, J. M. (1984). Representation and Recall in Infancy. In M. Moscovitch (Ed.), *Infant Memory: Its Relation to Normal and Pathological Memory in Humans and Other Animals* (pp. 75–101). Springer US. https://doi.org/10.1007/978-1-4615-9364-5_4

Mandler, J. M., & Johnson, N. S. (1976). Some of the thousand words a picture is worth. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(5), 529–540. https://doi.org/10.1037/0278-7393.2.5.529

Mandler, J. M., & Parker, R. E. (1976). Memory for descriptive and spatial information in complex pictures. *Journal of Experimental Psychology: Human Learning and Memory*, *2*(1), 38–48. https://doi.org/10.1037/0278-7393.2.1.38

Mannion, D. J., Kersten, D. J., & Olman, C. A. (2015). Scene coherence can affect the local response to natural images in human V1. *European Journal of Neuroscience*, *42*(11), 2895–2903. https://doi.org/10.1111/ejn.13082

Martin Cichy, R., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, *153*, 346–358. https://doi.org/10.1016/j.neuroimage.2016.03.063

Masuda, T., & Nisbett, R. E. (2001). Attending holistically versus analytically: Comparing the context sensitivity of Japanese and Americans. *Journal of Personality and Social Psychology*, *81*(5), 922–934. https://doi.org/10.1037/0022-3514.81.5.922

Matthews, W. J., & Adams, A. (2008). Another Reason Why Adults Find it Hard to Draw Accurately. *Perception*, *37*(4), 628–630. https://doi.org/10.1068/p5895

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and Reasoning among Tree Experts: Do All Roads Lead to Rome? *Cognitive Psychology*, *32*(1), 49–96. https://doi.org/10.1006/cogp.1997.0645

Meijer, F., Geudeke, B. L., & van den Broek, E. L. (2009). Navigating through Virtual Environments: Visual Realism Improves Spatial Cognition. *CyberPsychology & Behavior*, *12*(5), 517–521. https://doi.org/10.1089/cpb.2009.0053

Meyers, E., Embark, H., Freiwald, W., Serre, T., Kreiman, G., & Poggio, T. (2010). *Examining high level neural representations of cluttered scenes*. https://dspace.mit.edu/handle/1721.1/57463

MILNER, A. D., PERRETT, D. I., JOHNSTON, R. S., BENSON, P. J., JORDAN, T. R., HEELEY, D. W., BETTUCCI, D., MORTARA, F., MUTANI, R., TERAZZI, E., & DAVIDSON, D. L. W. (1991). PERCEPTION AND ACTION IN 'VISUAL FORM AGNOSIA'. *Brain*, *114*(1), 405–428. https://doi.org/10.1093/brain/114.1.405

Minsky, M. (1974). *A framework for representing knowledge*. Massachusetts Institute of Technology AI Laboratory Cambridge. https://direct.mit.edu/books/edited-volume/chapter-pdf/2301803/9780262275071_c000400.pdf

Miyamoto, Y., Nisbett, R. E., & Masuda, T. (2006). Culture and the Physical Environment: Holistic Versus Analytic Perceptual Affordances. *Psychological Science*, *17*(2), 113–119. https://doi.org/10.1111/j.1467-9280.2006.01673.x

Mollon, J. D., Bosten, J. M., Peterzell, D. H., & Webster, M. A. (2017). Individual differences in visual science: What can be learned and what is good experimental practice? *Vision Research*, *141*, 4–15. https://doi.org/10.1016/j.visres.2017.11.001

Morgan, A. T., Petro, L. S., & Muckli, L. (2019). Scene representations conveyed by cortical feedback to early visual cortex can be described by line drawings. *Journal of Neuroscience*, *39*(47), 9410–9423.

Moutsiana, C., de Haas, B., Papageorgiou, A., van Dijk, J. A., Balraj, A., Greenwood, J. A., & Schwarzkopf, D. S. (2016). Cortical idiosyncrasies predict the perception of object size. *Nature Communications*, *7*(1), 12110. https://doi.org/10.1038/ncomms12110

Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., Goebel, R., & Yacoub, E. (2015). Contextual Feedback to Superficial Layers of V1. *Current Biology*, *25*(20), 2690–2695. https://doi.org/10.1016/j.cub.2015.08.057

Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia*, *48*(2), 507–517. https://doi.org/10.1016/j.neuropsychologia.2009.10.011

Munneke, J., Brentari, V., & Peelen, M. (2013). The influence of scene context on object recognition is independent of attentional focus. *Frontiers in Psychology*, *4*. https://doi.org/10.3389/fpsyg.2013.00552

Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research*, *46*(5), 614–621. https://doi.org/10.1016/j.visres.2005.08.025

Nijboer, T. C. W., Kanai, R., de Haan, E. H. F., & van der Smagt, M. J. (2008). Recognising the forest, but not the trees: An effect of colour on scene perception and recognition. *Consciousness and Cognition*, *17*(3), 741–752. https://doi.org/10.1016/j.concog.2007.07.008

Noad, K. N., Watson, D. M., & Andrews, T. J. (2024). Familiarity enhances functional connectivity between visual and nonvisual regions of the brain during natural viewing. *Cerebral Cortex (New York, N.Y.: 1991)*, *34*(7), bhae285. https://doi.org/10.1093/cercor/bhae285

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430. https://doi.org/10.1016/j.tics.2006.07.005

Nosek, B. A., Ebersole, C. R., DeHaven, A. C., & Mellor, D. T. (2018). The preregistration revolution. *Proceedings of the National Academy of Sciences*, *115*(11), 2600–2606. https://doi.org/10.1073/pnas.1708274114

Ogawa, K., & Inui, T. (2009). The role of the posterior parietal cortex in drawing by copying. *Neuropsychologia*, *47*(4), 1013–1022. https://doi.org/10.1016/j.neuropsychologia.2008.10.022

Öhlschläger, S., & Võ, M. L.-H. (2020). Development of scene knowledge: Evidence from explicit and implicit scene knowledge measures. *Journal of Experimental Child Psychology*, *194*, 104782. https://doi.org/10.1016/j.jecp.2019.104782

Oliva, A. (2005). CHAPTER 41—Gist of the Scene. In L. Itti, G. Rees, & J. K. Tsotsos (Eds.), *Neurobiology of Attention* (pp. 251–256). Academic Press. https://doi.org/10.1016/B978-012375731-9/50045-8

Oliva, A., & Schyns, P. G. (2000). Diagnostic Colors Mediate Scene Recognition. *Cognitive Psychology*, *41*(2), 176–210. https://doi.org/10.1006/cogp.1999.0728

Oliva, A., & Torralba, A. (2001). Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope. *International Journal of Computer Vision*, *42*(3), 145–175. https://doi.org/10.1023/A:1011139631724

Oliva, A., & Torralba, A. (2002). Scene-Centered Description from Spatial Envelope Properties. In H. H. Bülthoff, C. Wallraven, S.-W. Lee, & T. A. Poggio (Eds.), *Biologically Motivated Computer Vision* (pp. 263–272). Springer. https://doi.org/10.1007/3-540-36181-2_26

Oliva, A., & Torralba, A. (2006). Chapter 2 Building the gist of a scene: The role of global image features in recognition. In S. Martinez-Conde, S. L. Macknik, L. M. Martinez, J.-M. Alonso, & P. U. Tse (Eds.), *Progress in Brain Research* (Vol. 155, pp. 23–36). Elsevier. https://doi.org/10.1016/S0079-6123(06)55002-2

Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527. https://doi.org/10.1016/j.tics.2007.09.009

Olivia, A., Mack, M. L., Shrestha, M., & Peeper, A. (2004). Identifying the Perceptual Dimensions of Visual Complexity of Scenes. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*(26). https://escholarship.org/uc/item/17s4h6w8

Open Science Collaboration. (2015). Estimating the reproducibility of psychological science. *Science*, *349*(6251), aac4716. https://doi.org/10.1126/science.aac4716

Ostrofsky, J., Nehl, H., & Mannion, K. (2017). The effect of object interpretation on the appearance of drawings of ambiguous figures. *Psychology of Aesthetics, Creativity, and the Arts*, *11*(1), 99–108. https://doi.org/10.1037/aca0000084

Park, J., & Park, S. (2017). Conjoint representation of texture ensemble and location in the parahippocampal place area. *Journal of Neurophysiology*, *117*(4), 1595–1607. https://doi.org/10.1152/jn.00338.2016

Park, S., Konkle, T., & Oliva, A. (2015). Parametric Coding of the Size and Clutter of Natural Scenes in the Human Brain. *Cerebral Cortex*, *25*(7), 1792–1805. https://doi.org/10.1093/cercor/bht418

Payne, A., & Singh, S. (2005). Indoor vs. Outdoor scene classification in digital photographs. *Pattern Recognition*, *38*(10), 1533–1545. https://doi.org/10.1016/j.patcog.2004.12.014

Peelen, M. V., Berlot, E., & de Lange, F. P. (2024). Predictive processing of scenes and objects. *Nature Reviews Psychology*, *3*(1), 13–26. https://doi.org/10.1038/s44159-023-00254-0

Peelen, M. V., Fei-Fei, L., & Kastner, S. (2009). Neural mechanisms of rapid natural scene categorization in human visual cortex. *Nature*, *460*(7251), 94–97. https://doi.org/10.1038/nature08103

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

Peelen, M. V., & Kastner, S. (2014). Attention in the real world: Toward understanding its neural basis. *Trends in Cognitive Sciences*, *18*(5), 242–250. https://doi.org/10.1016/j.tics.2014.02.004

Pelli, D. G., Majaj, N. J., Raizman, N., Christian, C. J., Kim, E., & Palomares, M. C. (2009). Grouping in object recognition: The role of a Gestalt law in letter identification. *Cognitive Neuropsychology*, *26*(1), 36–49. https://doi.org/10.1080/13546800802550134

Pelli, D. G., & Tillman, K. A. (2008). The uncrowded window of object recognition. *Nature Neuroscience*, *11*(10), 1129–1135.

Pellicano, E., & Burr, D. (2012). When the world becomes 'too real': A Bayesian explanation of autistic perception. *Trends in Cognitive Sciences*, *16*(10), 504–510. https://doi.org/10.1016/j.tics.2012.08.009

Peynırcıoğlu, Z. F. (1989). The generation effect with pictures and nonsense figures. *Acta Psychologica*, *70*(2), 153–160. https://doi.org/10.1016/0001-6918(89)90018-8

Phillips, J., Walford, N., & Hockey, A. (2011). How do unfamiliar environments convey meaning to older people? Urban dimensions of placelessness and attachment. *International Journal of Ageing and Later Life*, *6*(2), Article 2. https://doi.org/10.3384/ijal.1652-8670.116273

Pinto, E., & Peters, R. (2009). Literature Review of the Clock Drawing Test as a Tool for Cognitive Screening. *Dementia and Geriatric Cognitive Disorders*, *27*(3), 201–213. https://doi.org/10.1159/000203344

Previc, F. H., & Intraub, H. (1997). Vertical biases in scene memory. *Neuropsychologia*, *35*(12), 1513–1517. https://doi.org/10.1016/S0028-3932(97)00091-2

Raat, E. M., Farr, I., Kyle-Davidson, C., & Evans, K. K. (2022). Learning to perceive the gist of cancer through perceptual training. *Journal of Vision*, *22*(14), 3342. https://doi.org/10.1167/jov.22.14.3342

Ragan, E. D., Bowman, D. A., Kopper, R., Stinson, C., Scerbo, S., & McMahan, R. P. (2015). Effects of Field of View and Visual Complexity on Virtual Reality Training Effectiveness for a Visual Scanning Task. *IEEE Transactions on Visualization and Computer Graphics*, *21*(7), 794–807. IEEE Transactions on Visualization and Computer Graphics. https://doi.org/10.1109/TVCG.2015.2403312

Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C., & Tootell, R. B. H. (2011). The 'parahippocampal place area' responds preferentially to high spatial frequencies in humans and monkeys. *PLoS Biology*, *9*(4), e1000608. https://doi.org/10.1371/journal.pbio.1000608

Ramkumar, P., Hansen, B. C., Pannasch, S., & Loschky, L. C. (2016). Visual information representation and rapid-scene categorization are simultaneous across cortex: An MEG study. *NeuroImage*, *134*, 295–304. https://doi.org/10.1016/j.neuroimage.2016.03.027

Rao, R. P. N., & 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. https://doi.org/10.1038/4580

Renninger, L. W., & Malik, J. (2004). When is scene identification just texture recognition? *Vision Research*, *44*(19), 2301–2311. https://doi.org/10.1016/j.visres.2004.04.006

Roberson, D., Davidoff, J., Davies, I., & Shapiro, L. (2006). *Colour categories and category acquisition in Himba and English*. https://doi.org/10.1075/z.pics2.14rob

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press.

Rooney, K. K., Condia, R. J., & Loschky, L. C. (2017). Focal and Ambient Processing of Built Environments: Intellectual and Atmospheric Experiences of Architecture. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.00326

Rubin, D. C., & Kontis, T. C. (1983). A schema for common cents. *Memory & Cognition*, *11*(4), 335–341. https://doi.org/10.3758/BF03202446

Rumelhart, D. E. (1980). Schemata: The Building Blocks of Cognition. In *Theoretical Issues in Reading Comprehension*. Routledge.

Sagi, D. (2011). Perceptual learning in *Vision Research*. *Vision Research*, *51*(13), 1552–1566. https://doi.org/10.1016/j.visres.2010.10.019

Sayim, B., & Cavanagh, P. (2011). What Line Drawings Reveal About the Visual Brain. *Frontiers in Human Neuroscience*, *5*. https://doi.org/10.3389/fnhum.2011.00118

Schyns, P. G., & Oliva, A. (1994). From Blobs to Boundary Edges: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition. *Psychological Science*, *5*(4), 195–200. https://doi.org/10.1111/j.1467-9280.1994.tb00500.x

Segall, M. H., Campbell, D. T., & Herskovits, M. J. (1963). Cultural differences in the perception of geometric illusions. *Science (New York, N.Y.)*, *139*(3556), 769–771. https://doi.org/10.1126/science.139.3556.769

Seidl, K. N., Peelen, M. V., & Kastner, S. (2012). Neural Evidence for Distracter Suppression during Visual Search in Real-World Scenes. *Journal of Neuroscience*, *32*(34), 11812–11819. https://doi.org/10.1523/JNEUROSCI.1693-12.2012

Servajean, P., & Wiese, W. (2024). Processing Fluency and Predictive Processing: How the Predictive Mind Becomes Aware of its Cognitive Limitations. *Topics in Cognitive Science*, *n/a*(n/a). https://doi.org/10.1111/tops.12776

Shepard, R. N., & Metzler, J. (1971). Mental rotation of three-dimensional objects. *Science (New York, N.Y.)*, *171*(3972), 701–703. https://doi.org/10.1126/science.171.3972.701

Silson, E. H., Steel, A. D., & Baker, C. I. (2016). Scene-Selectivity and Retinotopy in Medial Parietal Cortex. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00412

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. https://doi.org/10.1177/0956797611417632

Simoncelli, E. P., & Olshausen, B. A. (2001). Natural image statistics and neural representation. *Annual Review of Neuroscience*, *24*, 1193–1216. https://doi.org/10.1146/annurev.neuro.24.1.1193

Singer, J. J. D., Cichy, R. M., & Hebart, M. N. (2023). The Spatiotemporal Neural Dynamics of Object Recognition for Natural Images and Line Drawings. *Journal of Neuroscience*, *43*(3), 484–500. https://doi.org/10.1523/JNEUROSCI.1546-22.2022

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*(6), 592–604. https://doi.org/10.1037/0278-7393.4.6.592

Smith, A. D. (2009). On the use of drawing tasks in neuropsychological assessment. *Neuropsychology*, *23*(2), 231–239. https://doi.org/10.1037/a0014184

Song, R., & Levine, C. S. (2024). Gentrification creates social class disparities in belonging. *Journal of Personality and Social Psychology*. https://doi.org/10.1037/pspi0000477

Spelke, E. S., & Lee, S. A. (2012). Core systems of geometry in animal minds. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1603), 2784–2793. https://doi.org/10.1098/rstb.2012.0210

Spence, I., Wong, P., Rusan, M., & Rastegar, N. (2006). How Color Enhances Visual Memory for Natural Scenes. *Psychological Science*, *17*(1), 1–6. https://doi.org/10.1111/j.1467-9280.2005.01656.x

Steel, A., Billings, M. M., Silson, E. H., & Robertson, C. E. (2021). A network linking scene perception and spatial memory systems in posterior cerebral cortex. *Nature Communications*, *12*(1), 2632. https://doi.org/10.1038/s41467-021-22848-z

Steeves, J. K. E., Humphrey, G. K., Culham, J. C., Menon, R. S., Milner, A. D., & Goodale, M. A. (2004). Behavioral and Neuroimaging Evidence for a Contribution of Color and Texture Information to Scene Classification in a Patient with Visual Form Agnosia. *Journal of Cognitive Neuroscience*, *16*(6), 955–965. https://doi.org/10.1162/0898929041502715

Stein, T., Kaiser, D., & Peelen, M. V. (2015). Interobject grouping facilitates visual awareness. *Journal of Vision*, *15*(8), 10. https://doi.org/10.1167/15.8.10

Stein, T., & Peelen, M. V. (2015). Content-specific expectations enhance stimulus detectability by increasing perceptual sensitivity. *Journal of Experimental Psychology: General*, *144*(6), 1089–1104. https://doi.org/10.1037/xge0000109

Stürzl, W., & Zeil, J. (2007). Depth, contrast and view-based homing in outdoor scenes. *Biological Cybernetics*, *96*(5), 519–531. https://doi.org/10.1007/s00422-007-0147-3

Sugiura, M., Shah, N. J., Zilles, K., & Fink, G. R. (2005). Cortical Representations of Personally Familiar Objects and Places: Functional Organization of the Human Posterior Cingulate Cortex.

Journal of Cognitive Neuroscience, *17*(2), 183–198. https://doi.org/10.1162/0898929053124956

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). *Going Deeper With Convolutions*. 1–9. https://www.cv-foundation.org/openaccess/content_cvpr_2015/html/Szegedy_Going_Deeper_With_2015_CVPR_paper.html

The Sims4. (2014). [Computer software]. Electronic Arts.

Thomas, G. V., & Jolley, R. P. (1998). Drawing conclusions: A re-examination of empirical and conceptual bases for psychological evaluation of children from their drawings. *British Journal of Clinical Psychology*, *37*(2), 127–139. https://doi.org/10.1111/j.2044-8260.1998.tb01289.x

Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. https://doi.org/10.1038/381520a0

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*(3), 391. https://doi.org/10.1088/0954-898X/14/3/302

Torralba, A., Oliva, A., Castelhano, M. S., & Henderson, J. M. (2006). Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, *113*(4), 766–786. https://doi.org/10.1037/0033-295X.113.4.766

Torralbo, A., Walther, D. B., Chai, B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2013). Good Exemplars of Natural Scene Categories Elicit Clearer Patterns than Bad Exemplars but Not Greater BOLD Activity. *PLOS ONE*, *8*(3), e58594. https://doi.org/10.1371/journal.pone.0058594

Trawalter, S., Hoffman, K., & Palmer, L. (2021). Out of place: Socioeconomic status, use of public space, and belonging in higher education. *Journal of Personality and Social Psychology*, *120*(1), 131–144. https://doi.org/10.1037/pspi0000248

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136.

Trouilloud, A., Kauffmann, L., Roux-Sibilon, A., Rossel, P., Boucart, M., Mermillod, M., & Peyrin, C. (2020). Rapid scene categorization: From coarse peripheral vision to fine central vision. *Vision Research*, *170*, 60–72. https://doi.org/10.1016/j.visres.2020.02.008

Tucciarelli, R., Ferrè, E. R., Amoruso, E., Azañón, E., & Longo, M. R. (2023). Gravitational and retinal reference frames shape spatial memory. *Journal of Experimental Psychology. General*, *152*(12), 3433–3439. https://doi.org/10.1037/xge0001441

Tulver, K., Aru, J., Rutiku, R., & Bachmann, T. (2019). Individual differences in the effects of priors on perception: A multi-paradigm approach. *Cognition*, *187*, 167–177. https://doi.org/10.1016/j.cognition.2019.03.008

Ujang, N., & Zakariya, K. (2015). The Notion of Place, Place Meaning and Identity in Urban Regeneration. *Procedia - Social and Behavioral Sciences*, *170*, 709–717. https://doi.org/10.1016/j.sbspro.2015.01.073

Vailaya, A., Jain, A., & Zhang, H. J. (1998). ON IMAGE CLASSIFICATION: CITY IMAGES VS. LANDSCAPES. *Pattern Recognition*, *31*(12), 1921–1935. https://doi.org/10.1016/S0031-3203(98)00079-X

Vaziri, S., & Connor, C. E. (2016). Representation of Gravity-Aligned Scene Structure in Ventral Pathway Visual Cortex. *Current Biology*, *26*(6), 766–774. https://doi.org/10.1016/j.cub.2016.01.022

Velisavljević, L., & Elder, J. H. (2008). Visual short-term memory for natural scenes: Effects of eccentricity. *Journal of Vision*, *8*(4), 28. https://doi.org/10.1167/8.4.28

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. https://doi.org/10.1016/j.copsyc.2019.03.009

Võ, M. L.-H., & Henderson, J. M. (2009). Does gravity matter? Effects of semantic and syntactic inconsistencies on the allocation of attention during scene perception. *Journal of Vision*, *9*(3), 24. https://doi.org/10.1167/9.3.24

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential Electrophysiological Signatures of Semantic and Syntactic Scene Processing. *Psychological Science*, *24*(9), 1816–1823. https://doi.org/10.1177/0956797613476955

Vogel, J., Schwaninger, A., Wallraven, C., & Bülthoff, H. H. (2007). Categorization of natural scenes: Local versus global information and the role of color. *ACM Trans. Appl. Percept.*, *4*(3), 19-es. https://doi.org/10.1145/1278387.1278393

Wagoner, B. (2013). Bartlett's concept of schema in reconstruction. *Theory & Psychology*, *23*(5), 553–575. https://doi.org/10.1177/0959354313500166

Walsh, K. S., McGovern, D. P., Clark, A., & O'Connell, R. G. (2020). Evaluating the neurophysiological evidence for predictive processing as a model of perception. *Annals of the New York Academy of Sciences*, *1464*(1), 242–268. https://doi.org/10.1111/nyas.14321

Wammes, J. D., Jonker, T. R., & Fernandes, M. A. (2019). Drawing improves memory: The importance of multimodal encoding context. *Cognition*, *191*, 103955. https://doi.org/10.1016/j.cognition.2019.04.024

Wammes, J. D., Meade, M. E., & Fernandes, M. A. (2016). The drawing effect: Evidence for reliable and robust memory benefits in free recall. *Quarterly Journal of Experimental Psychology (2006)*, *69*(9), 1752–1776. https://doi.org/10.1080/17470218.2015.1094494

Wang, G., Foxwell, M. J., Cichy, R. M., Pitcher, D., & Kaiser, D. (2024). Individual differences in internal models explain idiosyncrasies in scene perception. *Cognition*, *245*, 105723. https://doi.org/10.1016/j.cognition.2024.105723

Wang, R., Li, J., Fang, H., Tian, M., & Liu, J. (2012). Individual Differences in Holistic Processing Predict Face Recognition Ability. *Psychological Science*, *23*(2), 169–177. https://doi.org/10.1177/0956797611420575

Wardle, S. G., Kriegeskorte, N., Grootswagers, T., Khaligh-Razavi, S.-M., & Carlson, T. A. (2016). Perceptual similarity of visual patterns predicts dynamic neural activation patterns

measured with MEG. *NeuroImage*, *132*, 59–70. https://doi.org/10.1016/j.neuroimage.2016.02.019

Warrington, E., James, M., & Kinsbourne, M. (1966). Drawing disability in relation to laterality of cerebral lesion. *Brain : A Journal of Neurology*, *89*, 53–82. https://doi.org/10.1093/brain/89.1.53

Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. *NeuroImage*, *99*, 402–410. https://doi.org/10.1016/j.neuroimage.2014.05.045

Wichmann, F. A., Sharpe, L. T., & Gegenfurtner, K. R. (2002). The contributions of color to recognition memory for natural scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *28*(3), 509–520. https://doi.org/10.1037/0278-7393.28.3.509

Wiese, H., Schipper, M., Popova, T., Burton, A. M., & Young, A. W. (2023). Personal familiarity of faces, animals, objects, and scenes: Distinct perceptual and overlapping conceptual representations. *Cognition*, *241*, 105625. https://doi.org/10.1016/j.cognition.2023.105625

Wiesmann, S. L., & Võ, M. L.-H. (2022). What makes a scene? Fast scene categorization as a function of global scene information at different resolutions. *Journal of Experimental Psychology: Human Perception and Performance*, *48*(8), 871–888. https://doi.org/10.1037/xhp0001020

Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: The SHINE toolbox. *Behavior Research Methods*, *42*(3), 671–684. https://doi.org/10.3758/BRM.42.3.671

Williams, D. R., & Vaske, J. J. (2003). The Measurement of Place Attachment: Validity and Generalizability of a Psychometric Approach. *Forest Science*, *49*(6), 830–840. https://doi.org/10.1093/forestscience/49.6.830

Wolbers, T., Klatzky, R. L., Loomis, J. M., Wutte, M. G., & Giudice, N. A. (2011). Modality-Independent Coding of Spatial Layout in the Human Brain. *Current Biology*, *21*(11), 984–989. https://doi.org/10.1016/j.cub.2011.04.038

Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search. *Psychonomic Bulletin & Review*, *1*(2), 202–238. https://doi.org/10.3758/BF03200774

Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics*, *73*(6), 1650–1671. https://doi.org/10.3758/s13414-011-0153-3

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84. https://doi.org/10.1016/j.tics.2010.12.001

Woolford, J., Patterson, T., Macleod, E., Hobbs, L., & Hayne, H. (2015). Drawing helps children to talk about their presenting problems during a mental health assessment. *Clinical Child Psychology and Psychiatry*, *20*(1), 68–83. https://doi.org/10.1177/1359104513496261

Wu, K., Wu, E., & Kreiman, G. (2018). Learning scene gist with convolutional neural networks to improve object recognition. *2018 52nd Annual Conference on Information Sciences and Systems (CISS)*, 1–6. https://doi.org/10.1109/CISS.2018.8362305

Xbox app for Windows. (2009). [Computer software]. Microsoft Gaming.

Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, *111*(23), 8619–8624. https://doi.org/10.1073/pnas.1403112111

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308. https://doi.org/10.1016/j.tics.2006.05.002

Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2018). Places: A 10 Million Image Database for Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(6), 1452–1464. IEEE Transactions on Pattern Analysis and Machine Intelligence. https://doi.org/10.1109/TPAMI.2017.2723009

Zhou, S. S., Rowchan, K., Mckeown, B., Smallwood, J., & Wammes, J. D. (2025). Drawing behaviour influences ongoing thought patterns and subsequent memory. *Consciousness and Cognition*, *127*, 103791. https://doi.org/10.1016/j.concog.2024.103791

Zimmermann, E., Schnier, F., & Lappe, M. (2010). The contribution of scene context on change detection performance. *Vision Research*, *50*(20), 2062–2068. https://doi.org/10.1016/j.visres.2010.07.019