SMART SEARCH

Investigations into human visual search in structured environments

Sushrut Thorat

DONDERS SERIES

Smart Search

Investigations into human visual search
in structured environments

Sushrut Thorat

Colofon

Printing: GildePrint | www.gildeprint.nl
Layout: Sushrut Thorat | sushrutthorat.com
Cover: Ilse Modder | www.ilsemodder.nl
ISBN: 978-94-6284-400-1
Online: Donders thesis series #574 | repository.ubn.ru.nl

The image on the cover was generated using OpenAI's DALL-E 2 (openai.com/dall-e-2/) using the prompt "a human searching smartly for a fruit in a chaotic kitchen, cover art".

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

Smart Search

Investigations into human visual search in structured environments

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

woensdag 9 november 2022
om 16.30 uur precies

door

Sushrut Ramesh Thorat

geboren op 15 februari 1994
te Kolhapur (India)

Promotoren:
Prof. dr. W.P. Medendorp
Prof. dr. M.A.J. van Gerven

Copromotor:
Dr. M.V. Peelen

Manuscriptcommissie:
Prof. dr. U. Noppeney
Prof. dr. T.C. Kietzmann (Universität Osnabrück, Duitsland)
Dr. V.S. Stoermer (Dartmouth College, Verenigde Staten)

To,

Mom, Dad, & Sis

In Brief

To us, visual search for objects in the environment feels effortless as compared to other tasks such as multiplying large numbers. However, our efforts at building artificial systems have revealed that the former is computationally more challenging than the latter. That makes us wonder how our brain efficiently carries out visual searches. Decades of research indicate that the efficiency of human visual search relies on a plethora of processes, primary of which are: one, processing the hierarchical construction of the visual world (simple features such as orientations of lines constituting complex features such as shapes), two, selectively processing information relevant to the search task (e.g., suppress processing from parts of the image that contain non-target features), and three, learning the relationships between the constituent elements of the world that can guide the information selection process (e.g., knowing where an object occurs in a scene helps us constrain the search to those locations). Furthering our understanding of the processes underlying efficient search, I present new evidence using artificial neural networks, neuroimaging experiments (fMRI and EEG), and large-scale behavioral experiments. The main contributions are as follows: one, the search for body shapes can occur parallelly across our field of view; two, where selective attention needs to be deployed in a hierarchical visual system depends on the representational capacity of that visual system; three, the knowledge about the co-occurrences amongst the distractors can be learned and utilized to increase our search efficiency. I conclude the thesis by discussing the questions raised through our investigations and the research directions aimed at furthering our understanding of our seemingly effortless, but smart, visual search capabilities.

Contents

List of Figures

Chapter 1

General Introduction

As soon as we wake up in the morning, we engage in a string of searches for relevant objects - toothbrushes, coffee machines, towels, and so on. Usually, we know the locations of these objects and it becomes a matter of orienting our eyes, heads, and bodies to those locations. However, misplace an object (which happens often to me, courtesy of my flatmate) and we systematically (mostly; sometimes we get lazy and become frantic) look around our surroundings until we find the object or give up after some time. Such visual object search is ubiquitous in our lives. Most of these searches feel effortless. Unless you are someone who studies search systematically (e.g., a behavioral scientist), you won't think searching for objects is as or more complex than other abilities such as multiplying numbers. What we don't realize daily is that search is more fundamental to survival than the ability to multiply numbers, and through the millions of years leading to the rise of humans, natural selection has fine-tuned the visual system such that critical abilities such as visual search feel automatic, not requiring volition, in most situations. To demonstrate the actual complexity of the visual search for objects, I will outline the components required to build a visual search engine from scratch. The goal is not to build a replica of the human visual system but to demonstrate the various component processes that are essential for visual search (inspired by Richard Feynman's famous quote "What I cannot create, I do not understand" and Valentino Braitenberg's experiments in synthetic psychology (Braitenberg, 1986)). The goal is to add the components gradually to satisfy increasingly complex requirements of visual search. Insights underlying these components, from visual neuroscience on how the human visual system works, and computer vision on building a working visual search engine, are discussed in parallel. This exercise is followed by a deep historical dive into the visual search research that is relevant to my contributions to this field of research.

1.1 Building a visual search engine

Consider the following task (see Fig. 1.1): in a given scene (kitchen), where is the target object (faucet)? This is a typical search problem addressed in naturalistic visual search tasks in humans (Peelen and Kastner, 2014; Wolfe, 2021) and machines (Zhang et al., 2018). What could be the simplest solution to this problem?

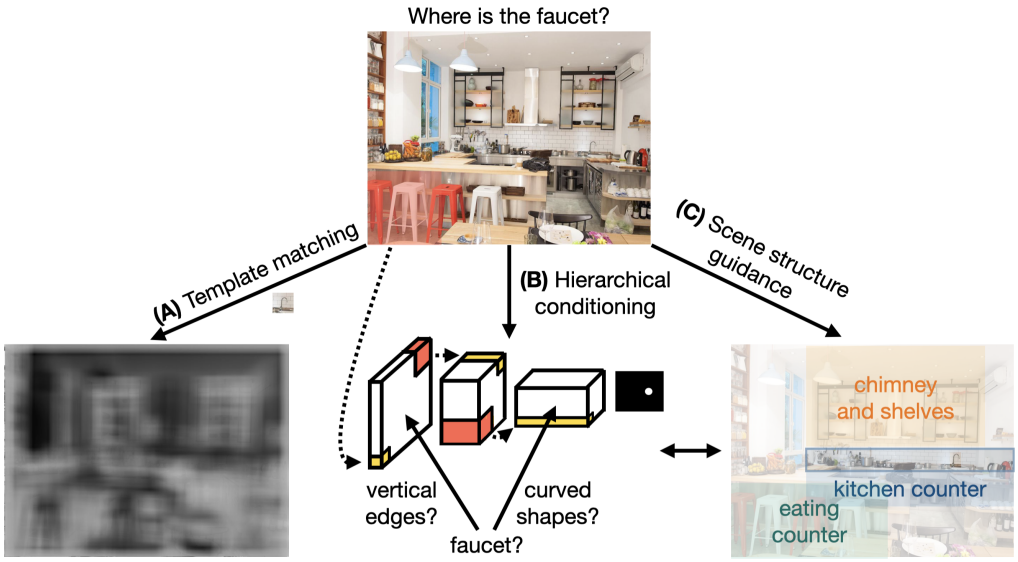

Figure 1.1: Building a visual search engine. In this image, the task is to locate the faucet in the scene. (A) Given an exact image of the faucet (with its surroundings), with template matching a direct comparison of the template image with patches of the scene can be made (here, a darker color represents more agreement). In the case of a natural scene, this procedure fails to identify the location of the template image in the scene (it will succeed in simpler displays where there is minimal clutter). (B) View-invariant (independent of the surrounding clutter and the pose of the object) object recognition can be implemented using convolutional neural networks. The information flow in such a network can be hierarchically conditioned by top-down signals to only pass information corresponding to the target (faucet) to eventually identify its location. (C) Knowledge about the occurrence of objects in certain locations in the scene (e.g., faucets typically appear on kitchen counters) can be used to restrict the search to the relevant regions in the scene, reducing the competition from the rest of the scene, making search more efficient.

1.1.1 Template matching

Let's assume we have access to the image of the faucet as it appears in the scene (i.e. a cutout - the template). The template could be slid across the entire visual field, and similarity scores between the template and the region being tested could be computed across the entire field. The similarity score could be as simple as computing the average absolute difference between the two images - the template and the region being assessed. The regions with the lowest difference scores could be the output candidate regions where the faucet is present (Fig. 1.1A).

Any occlusions or variations in the view (where we are in the kitchen scene would dictate how the faucet looks) - conditions that occur regularly in the real world - would make inference harder (Brunelli, 2009) as the template might not match the image in the actual region where the faucet is present well enough. Additionally, what if the identity of the faucet is unknown i.e. no template image is provided? Both these issues - view and exemplar variance - can be mitigated by resorting to certain transformations of the image (DiCarlo and Cox, 2007). The class of architectures that can implement these transformations is discussed below.

1.1.2 View-invariant object recognition

Objects are made of parts and a compositional hierarchy can be established amongst the parts - small parts to big parts, and simple parts (e.g., curvature of edges) to complex parts (e.g., wheels of a car). Using computing systems called convolutional neural networks (CNNs), simple features can be first extracted at a smaller spatial scale and then composed into complex features at a larger spatial scale, mirroring the natural compositional hierarchy of objects (Fukushima and Miyake, 1982; LeCun et al., 2015). These CNNs constitute state-of-art computer vision solutions to the problem of view and exemplar-invariant object recognition. In such CNNs, trained to recognize objects, the early layers extract low-level features (e.g., orientations, color) from small patches of the image. In subsequent layers, these features are combined across increasingly larger patch sizes to extract more complex features (e.g., curvature). This process is continued hierarchically through the layers of the network, to gradually obtain high-level features (e.g., shape) in the later layers of the network. These high-level features are invariant to changes in the lower-level features they are composed of (Zeiler and Fergus, 2014). These properties allow the network's inferences about the objects to be more resilient to changes in views or occlusions. Incidentally, this feature hierarchy in the CNNs matches the feature hierarchy found in the human visual system (Güçlü and van Gerven, 2015; Cichy et al., 2016). However, this is not surprising as the basic architecture of the CNN is similar to that of the human visual system - the study of which seems to have motivated the development of CNNs (Hubel and Wiesel, 1962; Lindsay, 2021).

How can such an architecture be used to identify the location of the target?

1.1.3 Target-driven hierarchical conditioning of the information flow

When searching for the faucet with no information about how it would look in the scene, in terms of its shape, pose, and any occlusions, what could be the nature of the search template? Ideally, there would be discriminative evidence for the faucet across space in the later stages of a CNN. The template could be based on the expected pattern of responses in those stages. We could perform template matching at a later stage of the CNN, using a view-invariant representation of the faucet found at that stage, rather than at the input using an image-based template of the faucet (which is prone to view-variance as described earlier). Low-level features might also be indicative of the faucet (e.g., vertical edges, silver color fill) and template matching could also be performed given the responses from the early stages.

Alternatively, instead of reading out from each stage and performing template matching separately, information about the discriminative features at every stage for the faucet could be used to condition the flow of information through the engine. For example, only the information from the locations where there is evidence for vertical edges (indicative of the existence of a faucet) might be processed further, followed by further constraints driven by other features of the faucet. Such conditioning might enhance the inference of the object present across the visual field in the later stages of the engine and template matching there might yield inferences on where the target is (Fig. 1.1B).

Such target-driven conditioning of the information flow in the engine, useful for the task of enhancing discriminative signals in the later stages of the engine, has been demonstrated in CNNs (Lindsay and Miller, 2018). It has also been posited to occur in the human visual system (Kastner and Pinsk, 2004). More generally, the ability to condition visual processing of object features based on the features of the target of the search (not about where the target is but what the target is), at various stages of the visual system, is termed feature-based attention in neuroscience (Maunsell and Treue, 2006). In the subsequent sections, I will discuss our current knowledge about the nature and neural correlates of feature-based attention, relevant to my

investigations described in Chapter 2.

What is the correct level for such target-driven conditioning? Early conditioning might be useful if any modulation of the low-level features could lead to better discriminative responses in the later stages - a function of two aspects: one, if the responses at the later stages aren't already maximally discriminative (i.e. the engine does not have sufficient representational capacity), and two, how category-discriminative and functionally connected to the high-level features the early features are. If the engine does not have sufficient capacity, early conditioning might help pass information that would help the later stages make better inferences, given the conditions described in the second aspect are met. In the study of the human visual system, capacity limits of the visual system are cited as a reason why conditioning (referred to as attention) in the early stages of visual processing might be important (Lavie and Tsal, 1994). In the subsequent sections, I will discuss our current knowledge about the influence of capacity limits of the human visual system on the deployment of information conditioning (feature-based attention) at various stages of the visual system, relevant to our investigations described in Chapters 3 and 4.

In addition to the information about the target, other sources of knowledge can be used to further constrain the information flow in the engine - the spatial and semantic relationships between the target and the distractors, and amongst the distractors.

1.1.4 Using the regularities in the scene to constrain information flow

Currently, our visual search engine has no idea about the relationships between the target and the distractors in terms of their spatial relationships. However, non-target objects could provide substantial information about where the target could be located. For example, typically faucets are found on the kitchen counter, and given this knowledge, information from only the regions above the kitchen counter could be propagated through the engine to get rid of other distractors leading to a better inference about the faucet's location (Fig. 1.1C). Such knowledge about target-distractor co-occurrence has been posited to be useful in human visual search (Wolfe et al., 2011b; Võ et al., 2019). Additionally, knowledge about the semantic relationships between the objects in the scene could also be used to disambiguate small, blurry, or occluded objects, leading to better inferences about where the target might be (Bar, 2004; Zhang et al., 2020). For example, blurry hairdryers and drills look similar but they co-occur with different objects, so knowledge about the current surroundings of the blurry object could be used to tag it as a hairdryer or a drill (see Box 1 in Bar (2004)).

To constrain the information flow based on spatial constraints driven by the non-target objects, the engine needs to extract scene structure information in parallel to extracting the features discriminative for the target object. Spatial relationships amongst the non-targets could be exploited to compress the information content about the scene structure. For example, as tables and chairs co-occur in the kitchen and other scenes, the part of space occupied by the table and the chairs in the scene could be represented by a single "dining" entity in the engine - a region unlikely to contain the faucet. In artificial neural networks, it has been shown that object co-occurrences can be exploited to compress the representation of the scene (Plaut and Vande Velde, 2017). In human visual search, it has been proposed that such co-occurrences amongst distractors could be used to compress the scene leading to a reduction in the complexity of the search. In the subsequent sections, I will discuss our current knowledge about the influence of the regularities in scenes on human visual search, specifically focusing on the influence of the regularities amongst distractors on search complexity, relevant to our investigations described in Chapters 5 and 6.

In summary, in terms of searching for a target object in a static visual scene, the search engine needs to extract hierarchical features from the image in the service of view-invariant identification

of the target. The target of the search needs to dictate the flow of information processing in the engine by modulating the responses at each relevant stage. Additional information about the location of the target could be inferred given knowledge about the spatial occurrence of the target in the scene, and be used to further constrain the information flow in the engine. This exercise sheds light on the complexity of the visual search for objects. However, human visual search has additional layers of complexity, and I will discuss them next.

1.1.5 Leaping to a model of human visual search

Although the visual search engine we built resembles the basic architecture of a state-of-art model of human visual search - Guided Search 6.0 (GS6; Wolfe (2021)) - there are many points of divergence. One of the major points of divergence stems from the fact that our engine is designed to perform one-shot target localization whereas human visual search involves multiple runs of the processes in our engine, corresponding to re-assessments of the accrued evidence, with processes such as spatial attention and eye movements. Human vision has a high resolution only in the central part of the visual field (the fovea-periphery organization of the visual system; Aubert and Foerster (1857); Curcio et al. (1990); Strasburger et al. (2011)), and eye movements are essential so the high-resolution spotlight can lay upon a prospective location for optimal inference about the presence of the target therein. Most of the other points of divergence between the search engine we built and the models of the human visual system primarily result from the requirement of iterative search.

Additionally, humans live in an evolving world, and the history of their behaviors and other associations (e.g., some objects being more important than others, for example, in terms of their impact on our survival) can also influence their search behavior. It has been indicated that simple models like the visual search engine we built or GS6 (as mentioned in Wolfe (2021) - "GS6 remains a model of a specific class of laboratory search tasks") do not exhaustively capture the richness of human visual search. For example, similar to the leap from one-shot target localization in the search engine to iterative search in GS6 (and usually studied in visual cognitive neuroscience), there is another leap from such iterative search lasting for barely a second to searches lasting minutes in detecting cancerous regions on mammograms or hours while searching for lost sailors at sea. Nonetheless, the thousands of laboratory experiments with simple visual search tasks and modeling efforts spanning decades have provided us with an understanding of the, seemingly easy but procedurally complex, human visual search behavior and how neurons in the brain process visual information leading to those behaviours (Carrasco, 2011; Eckstein, 2011).

I will present our contributions to our understanding of human visual search in the subsequent chapters. To set the stage for those contributions, I will outline the developments in our understanding of two relevant aspects of visual search: the characterization of feature-based attention in terms of its neural correlates and functional importance, and the influence of the regularities in scenes on visual search and the neural underpinnings of that influence.

1.2 Feature-based attention in parallel visual search

1.2.1 Behavioural evidence for parallel search capabilities

Consider the search task from Egeth et al. (1984): Participants had to search for a red O among black Os and red Ns (see Fig. 1.2A). Half of the participants were asked to try and restrict their search to all the red items and the other half to all the Os. In one condition (termed 'unconfounded'), for the participants told to restrict their search to red items, there were two red Ns and

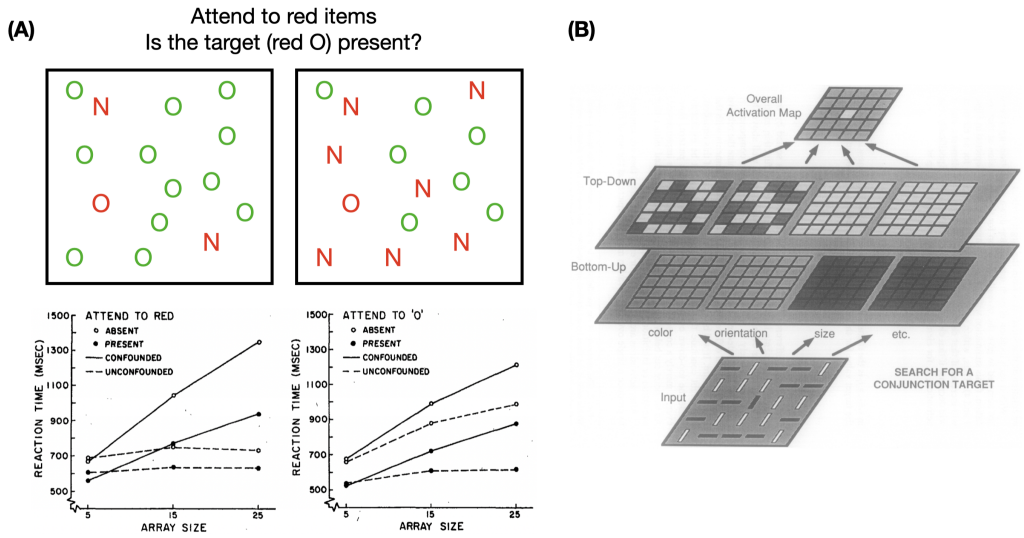

Figure 1.2: Parallel search capabilities. (A) Participants can restrict their search for items with a relevant color or shape across the visual field in parallel, thereby making their search time independent of the number of distractors that do not share the relevant feature. Figure adapted from Egeth et al. (1984). (B) Top-down information about the features of the target interacts with the bottom-up feature maps to generate a prediction for the location of the target accumulated across all target-relevant feature maps. Figure reproduced from Cave and Wolfe (1990).

a variable number of black Os, and for the participants told to restrict their search to red items, there were two black Os and a variable number of red Ns. In the other condition (termed 'confounded'), there were a variable number of distractors - half of which were red Ns and the other half black Os. The time it took participants to find the target (reaction time) as a function of the number of distractors was assessed in both conditions.

In the 'confounded' condition, the reaction time increased with an increasing number of distractors. An increasing number of distractors necessitated more examinations of the items before the participants could stumble upon the target (as no information about the target's location was available). However, in the unconfounded condition, the reaction time barely varied with the number of distractors. So, participants were indeed able to restrict their search only to the Os or the red items, as instructed, such that, in the unconfounded condition, the number of effective distractors stayed the same (equal to 2). This suggested that a process allowed the participants to examine only these 3 items (the target and the two distractors) while skipping the examination of the other distractors. Additionally, unlike the sequential examinations used in the confounded condition to ascertain which item was the target, in the unconfounded condition the same sequential examinations were not used to identify the items containing the relevant feature (either the color 'red' or the letter 'O'). Instead, we can conclude that the selection process operated over the entire display, in parallel.

Given a feature that identifies the target, the ability to increase search efficiency by reducing, in parallel, the number of items to be examined to identify the target was demonstrated in many experiments involving simple features such as colors, letters, and orientations (Green et al., 1953; Green and Anderson, 1956; Smith, 1962; Treisman and Gelade, 1980; Treisman and Gormican, 1988; Wolfe et al., 1989). Guided search (GS; Wolfe et al. (1989); Cave and Wolfe (1990)) was

put forth as a model for visual search incorporating these behavioral results. According to GS, all items in the image are represented, in parallel, in the visual system based on their features (e.g., color, orientation; Zeki (1976)). Top-down information about the features of the target interacts with these feature maps and activates the locations corresponding to the items who share the features with the target on the corresponding feature maps. The activations are summed across the feature maps and the locations are ranked by the evidence for the presence of the target (see Fig. 1.2B). The locations are then examined according to their ranks until the target is found or the search is terminated. This core aspect - the interaction between the representation of items based on their features and the top-down feature expectation based on the target - is also known as feature-based attention, and has stood the test of time as it is present in the most recent version of the GS model of visual search (GS6; Wolfe (2021)).

GS also explained previous observations that the search for targets differing from the distractors along more than one basic feature dimension depended on the number of distractors (Treisman and Gelade (1980); conjunction search; e.g., finding a red O among green Os and red Ns): if the representations of the items on the feature maps are noisy, the combined activation scores, after interaction with the top-down expectations, are noisier and the number of items needing examination before the target is found scales with the number of distractors. One way to minimize the noise would be to create feature maps for the conjunction of the basic features (e.g., representing both letter and color). Then the interaction between the representations of the items and the top-down expectation could happen on one joint feature space instead of two independent feature spaces. The efficiency of such a conjunction search could also be barely dependent on the number of distractors. Nakayama and Silverman (1986) provided evidence along these lines.

In macaques, in the middle temporal visual area (MT), cells tuned to both the direction of motion and binocular disparity of stimuli were found. In Nakayama and Silverman (1986), participants were presented with a stereoscopic image with random dot patterns (RDP). On the front plane, the distractor RDPs moved up and on the back plane, the distractor RDPs move down. If the target RDP was in the front plane, it moved down, and if it was in the back plane, it moved up. Participants could indicate the odd RDP (the target) with the same efficiency independent of the number of distractors. This independence was not found in another experiment where the target instead could differ from the distractors in terms of stereo depth and color. Correspondingly, no cells were found in the macaque visual cortex which were tuned both to binocular disparity and color. These observations support the idea that if the visual cortex contains joint tuning for multiple simple features, searching for targets along those feature dimensions could also be done in parallel. What all features along the hierarchy - simple features to complex features made out of conjunctions of the simple features - are useful in directing such parallel searches by humans?

There is ample evidence that simple features such as color, motion, orientation, and size can be used to direct parallel searches (Wolfe and Horowitz, 2004, 2017). Features of intermediate complexity, such as curvature (Treisman and Gormican (1988)) and simple shape parts (Wolfe and Bennett (1997)), can be also be used to direct parallel searches (see Wolfe and Horowitz (2017) for the complete list). On the other end of the feature spectrum, it is unclear if higher-level features, diagnostic of object categories, immune to the variations across exemplars, could be used to direct parallel searches. For example, could the search for a face or a car (where knowledge about the exemplar is unknown) amongst other distractor objects be independent of the number of distractors?

In Hershler and Hochstein (2005), participants searched for faces, cars, or houses in a grid of line drawings of objects. The reaction times for face detection increased by 3 ms/item, and for cars and houses, increased by more than 15 ms/item. In Treisman and Souther (1985), it was suggested that parallel search yields slopes of around 5-6 ms/item, while serial search commonly

yields slopes of 20 ms/item and above. Using this guideline, they concluded that a search for faces, but not cars or houses, can be done in parallel, similar to the simple search for a red O within blue Os. This specificity of the parallel search capabilities for faces was attributed to the existence of cells tuned specifically to faces in the human visual cortex (Kanwisher et al. (1997); no such specifically tuned cells were found for cars or houses). This specificity of the parallel search for faces (and relatedly, bodies) has been reported in subsequent studies (Ro et al., 2007; Reeder and Peelen, 2013; Reeder et al., 2015). Can other object categories not avail of such parallel search capabilities?

In Golan et al. (2014), participants searched for faces and cars (among other categories) in a grid of images of objects. Crucially, some participants were cars experts. In both, the car experts and non-experts, the search for faces did not depend strongly on the number of distractors (slopes 5 ms/item in agreement with the previous studies). However, the car experts had much lower search slopes for cars (< 15 ms/item) as compared to the car non-experts (> 30 ms/item). Expertise for car discrimination was related to increased search efficiency for cars (see also: Reeder et al. (2016)). So, expertise with a particular category of objects could lead to an ability to search for those objects with an efficiency rivaling that of the search for faces.

Curiously, in Golan et al. (2014), the search slopes for airplanes were of the same magnitude as that for faces. The authors argued that low-level features diagnostic of the airplane examples used might have led to flatter search slopes. This argument could also be made to explain the low search slopes for faces, in the previously discussed Hershler and Hochstein (2005) study (VanRullen (2006); but see: Hershler and Hochstein (2006)). Incidentally, in all the studies showing flatter search slopes for one category (mostly faces) as compared to other categories, it is hard to rule out that the differences manifested due to the existence of low-level features diagnostic for the categories showing flatter slopes (also see: Treisman (2006)).

In summary, decades of behavioral experiments in visual search have revealed that humans can utilize the knowledge about the features of the search target to select the items sharing those features, in parallel across the visual field, to reduce the time required to find the target. This is possibly accomplished by the top-down feature expectations interacting with the input-driven activations of the feature maps, leading to a map of candidate locations where the target could be present. This feature-based interaction (or 'attention') has been demonstrated, quantified by search slopes, for simple features such as color, orientation, and size, but also high-level features diagnostic of object categories such as human faces and bodies. It is unclear if other object categories can provide us with parallel search capabilities.

While the behavioral experiments helped reveal the human capability for parallel search, single-cell recordings and neuroimaging studies shed further light on what neural processes this capability relies on.

1.2.2 The neural correlates of spatially-global feature-based attention

In Martinez-Trujillo and Treue (2004), neurons from the macaque middle temporal area (MT) were recorded as they performed visual tasks. These MT neurons were tuned to motion direction and were implicated in motion processing (Newsome and Pare, 1988). Two random dot patterns (RDPs) were presented. Only one of the RDPs, always task-irrelevant, was present in a given neuron's receptive field (RF). The macaques pressed a button to indicate a change in the target RDP's motion direction or speed. The RDP in the neuron's RF either had the neuron's preferred motion direction or the opposite, anti-preferred direction. The target RDP's motion direction either matched that direction or was the opposite direction (see Fig. 1.3A). The neural response (average spike frequency between 200ms and 1200ms after stimulus onset) was assessed as a

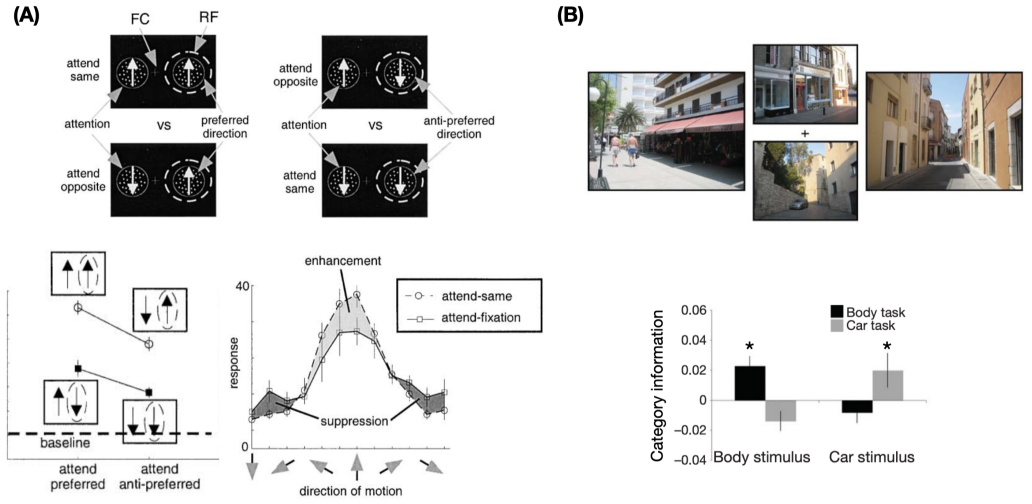

Figure 1.3: The neural correlates of spatially-global feature-based attention. (A) The feature-similarity gain model: the responses of neurons that maximally prefer the attended motion direction are maximally enhanced and the responses of the neurons which minimally prefer the attended motion direction are maximally suppressed, irrespective of the motion direction present in those neurons' receptive fields. Figure adapted from Martinez-Trujillo and Treue (2004). (B) Searching for bodies (or cars) selectively makes the representations of bodies (or cars), in task-irrelevant locations, more similar to prototypical responses for bodies (or cars) in the object-selective visual cortex, providing evidence for spatially-global feature-based attention where the features are diagnostic of categories. Figure adapted from Peelen et al. (2009).

function of the RDPs' motion directions, and the neuron's preferred motion direction.

Irrespective of the motion direction of the RDP within its RF, the neuron's response was higher when the target RDP had the neuron's preferred motion direction. This result suggests that the attention to a motion direction in one part of the visual field could lead to modulations of the responses of neurons with RFs covering any other part of the visual field. These modulations would depend on the neurons' preferences to the attended motion direction - neurons with higher activity for that direction as opposed to the other directions would have increased activity as compared to the neurons which have higher activity for the unattended directions - a feature-similarity-driven modulation.

In another experiment, they compared the activity of the neurons while attending to a motion direction not in the neurons' RFs while the same motion direction was presented in the neurons' RFs (the 'attend-same' condition) as opposed to monitoring a fixation cross in the center of the display for change of color (the 'attend-fixation' condition). They found that the responses of neurons that maximally prefer the attended direction are maximally enhanced and the responses of the neurons which minimally prefer the attended direction are maximally suppressed, irrespective of the motion direction in their RFs (see Fig. 1.3A). These modulation patterns was termed the feature-similarity gain model (FSGM; Treue and Trujillo (1999)).

Such feature-similarity-driven gain modulation had been studied before with experiments where the spatial attention of the participants was present in the RF of the recorded neuron (Haenny and Schiller, 1988; Maunsell et al., 1991; O'Craven et al., 1997). Treue and Trujillo (1999) was the first study where spatial attention was pulled away from the recorded neuron's RF, there-

fore obtaining a measure of feature-similarity-driven gain modulation in the neurons with RFs in the task-irrelevant and unattended locations - a signature of the spatially-global nature of such attention-driven gain modulation. In a subsequent study, in a visual search task, this spatially-global modulation was shown to be related to the prioritization of items that share the attended feature with the target, eventually leading to a shift in spatial attention and an eye movement to one of the prioritized stimuli (Bichot et al., 2005). Increased synchronization with the local field potential was observed in addition to increased spike rates for the neurons that preferred the attended feature and whose RFs included the prioritized items. The authors proposed that "this strong signal is the one that ultimately triggers spatial attention to the candidate target, and, in most cases, an eye movement toward it". Additionally, in Zhang and Luck (2009), using electroencephalography (EEG), it was observed that such spatially-global modulation can affect the processing of the stimulus rapidly, around 100ms after stimulus onset. Given these observations, it has been proposed (Maunsell and Treue, 2006; Carrasco, 2011; Wolfe, 2021) that such spatially-global feature-similarity gain modulation is the feature-based interaction (or 'attention') that was proposed based on the results from the behavioral experiments (Cave and Wolfe, 1990) discussed in the previous section.

Such spatially-global feature-similarity-driven modulation was subsequently observed for other low-level features other than motion direction - colour (Saenz et al., 2002; Bichot et al., 2005; Zhang and Luck, 2009; Andersen et al., 2013), orientation (McAdams and Maunsell, 2000; Jehee et al., 2011), and simple shapes (squares, crosses, etc.; Bichot et al. (2005)). What about higher-level features diagnostic of object categories? As discussed in the previous section, behavioral experiments provided evidence that high-level features diagnostic of faces and bodies can avail of parallel search capabilities. What about the evidence from experiments recording neural activity? Only two studies have provided evidence for spatially-global gain modulation for high-level features. In Peelen et al. (2009), as discussed below, functional magnetic resonance imaging (fMRI) based neural signatures of spatially-global gain modulation for human bodies (with faces) and also cars were found. In Störmer et al. (2019), an electroencephalogram (EEG) based neural signature of spatially-global gain modulation for faces (but not houses) was found.

In Peelen et al. (2009), participants were shown four images of natural scenes - two arranged horizontally and the other two arranged vertically. In a given block, they had to either search for cars or human bodies in the cued locations - horizontally or vertically aligned images - and report the presence of the target. In both the task-relevant and irrelevant locations, only one of the scenes contained cars or bodies (see Fig. 1.3B). In the category localizer runs, images of bodies and cars were shown in isolation to obtain the prototypical response patterns for these categories. In the object-selective cortex, the fMRI response patterns evoked by the stimuli presented in the task-irrelevant locations were assessed as a function of the category of the objects present in the stimuli and the search target. The responses to bodies resembled prototypical body representations more than prototypical car representations only when bodies were the search targets. Correspondingly, the responses to cars resembled prototypical car representations more than prototypical body representations only when cars were the search targets. These results suggested that spatially-global modulation of high-level features diagnostic for bodies and cars was deployed during the search for the corresponding categories in selected regions of the display.

However, in both Peelen et al. (2009) and Störmer et al. (2019), the search involved only two categories. Distinctions between examples of these categories could be made based on lower-level visual features: houses have more rectilinear edges than faces, and cars have more horizontal edges than bodies. There are also mid-level texture-related differences between these categories. It is unclear where in the hierarchy the features diagnostic of these categories lay and which of these features were used for spatially-global modulation. To constrain the features that could be

used to distinguish between categories and therefore the features that could be used for search and spatially-global modulation, we designed a new experiment. The design was along the lines of Peelen et al. (2009), however, silhouettes of objects were used to avoid texture-related distinctions and constrain the category-diagnostic features to shape space. To mitigate the possibility of low-level differences, 50 examples from 6 categories were included: beds, bottles, cars, chairs, lamps, and human bodies (without faces). The focus was on bodies - to assess if spatially-global modulation could be found for bodies in this more controlled setting. Additionally, we were curious whether spatially-global modulation could be found for the other categories (although there is little to no behavioral evidence for parallel search for these categories as discussed previously). This experiment and its results are detailed in Chapter 2.

In this section, we discussed how feature-based attention can result in spatially-global modulation of information processing, resulting in efficient parallel search. The feature-similarity gain model (FSGM) was proposed to explain how each neuron is modulated given the target of search and the response profile of that neuron. Presumably, the goal of such a modulation scheme is to generate response profiles across neurons that can optimally discriminate between the targets and the distractors in a format readable by downstream decision networks. What are the characteristics of such an optimal modulation? Is FSGM the optimal model?

1.2.3 What features are modulated for optimal search?

Suppose we have to find a tilted line (55° tilt) among other tilted lines with orientations slightly different from the target (50° tilt). FSGM states that the neurons maximally selective to the target orientation should be enhanced maximally. If the goal of a modulatory framework is to generate maximally differing responses to the target and the distractors, then maximally enhancing neurons selective to the target orientation is not the optimal solution. Instead, enhancing the neurons which are tuned away from the target, in the direction opposite to the distractor (60° tilt for the example here) provides a higher signal difference between the target and the distractors (Navalpakkam and Itti (2007); see Fig. 1.4A).

In Navalpakkam and Itti (2007), participants searched for a target line oriented at 55° among distractors oriented at 50°. After a few of these search trials, a trial with lines oriented at 30°, 50°, 55°, 60°, and 80° was shown and participants had to indicate where the target (55°) was. In all 4 participants they tested, the line oriented at 60° was mostly reported to be the target, suggesting that the participants were searching for a line oriented at 60° which helped better distinguish between the 55° and 50° orientations with a higher signal for the actual target - the 55° orientation (see Fig. 1.4B). Such a shift in the target template (characterized by orientation or colour) has been observed during discrimination tasks (Regan and Beverley, 1985; Lee et al., 1999), and in other search experiments (Scolari and Serences, 2009; Becker et al., 2013; Geng et al., 2017). With neuroimaging, Scolari et al. (2012) observed that neural populations selective for the off-target orientations were indeed modulated when faced with distractors that were similar to the target, as was the case in the experiments mentioned above. In summary, when faced with a hard search task - where the targets are very similar to the distractors - a modulation scheme, not resembling FSGM, can be used to optimally discriminate between the target and the distractors.

Moving beyond the search for simple features, what features need to be modulated while searching at the level of categories (e.g., looking for a car)? As discussed previously, there might be features at all levels of the visual hierarchy diagnostic for categories. Is FSGM a useful and optimal framework for the modulation of features along the visual hierarchy?

In Lindsay and Miller (2018), feature-based attention was deployed at various stages of a convolutional neural network (CNN) to assess if the modulation of features (according to FSGM or

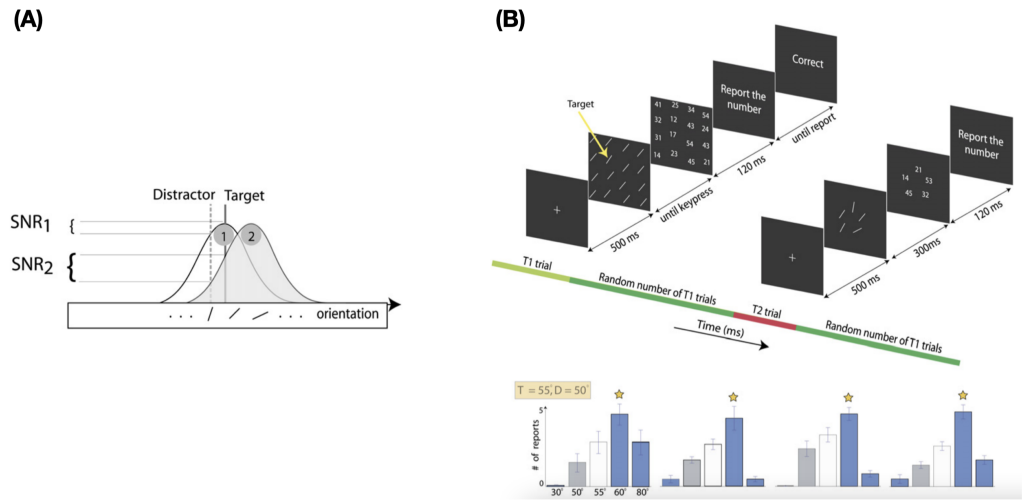

Figure 1.4: The modulation of features for optimal search. (A) Moving beyond the feature-similarity gain model: enhancing the neurons which are tuned away from the target, in the direction opposite to the distractor provides a higher signal difference between the target and the distractors. (B) In agreement with the aforementioned model, when searching for a line oriented at 55 degrees among lines oriented at 50 degrees, participants tune their search template towards a search for a line oriented at 60 degrees. Figures reproduced from Navalpakkam and Itti (2007).

otherwise) could lead to better detection of the target. The CNN was initially trained for 1000-way classification of object categories on millions of images (Simonyan and Zisserman, 2014). As discussed earlier, akin to the human visual system, such a CNN has been shown to develop a compositional hierarchy of features - from orientations and colors in the early layers, to features that are view-invariant and characterize categories in the later layers (Zeiler and Fergus, 2014). Binary classifiers for each target category (to indicate if the target was present or absent) were trained on the representations of the final layer of the CNN. Feature-based attention was then deployed across the CNN to assess if it improved the performance of those classifiers. An increment in the performance would indicate that the modulation led to more discriminative signals at the final layers.

FSGM based modulation was shown to be useful at all layers of the CNN with its impact increasing with the depth of the layer. However, a gradient-based modulation scheme was shown to outperform FSGM, maximally in the mid-level layers. This scheme computed what the change in neural response should be at a given layer to maximize the difference in representations between the target and the distractors in the final layers. These results indicate that, although useful, FSGM might not be the optimal modulation scheme when feature-based attention is deployed in the earlier stages of the network while the target of the search is at the level of categories. This finding aligns with other studies suggesting it is not straightforward to assign causal importance to any category-diagnostic lower-level features as they might not be able to impact downstream performance (Morcos et al., 2018; Zhou et al., 2018).

The gradient-based modulation scheme in Lindsay and Miller (2018) was an approximation to the solution for how neural activity upstream should be changed to result in increased discriminability in the activity downstream. Instead, we trained the modulation scheme with backpropagation iteratively, while freezing the rest of the network's weights, to reach a steady-state solution

for the optimal modulation. We then compared the performance gains due to that solution to the gains from an FSGM based modulation scheme. This experiment and its results are detailed in Chapter 3.

In this section, we discussed how during the modulation of neurons with feature-based attention, FSGM might not be the optimal modulation scheme. Instead, a modulation scheme that takes the network's information transformations into account could lead to better performance. While searching for objects such as a car, beyond how early neural activity should be optimally modulated, under what circumstances should early visual activity be modulated?

1.2.4 The usefulness of feature-based attention in early visual processing

When might the modulation of early visual processing be important? In a capacity-limited neural network, all the information required for optimally performing the given task, say object discrimination, might not reach the final stages of the network. In that case, task-driven modulations of early visual processing might be essential to discard irrelevant information and make the relevant information reach the final stages. The notion of capacity limits in visual processing, and the need for attention to overcome these limits have been long considered in cognitive science (Broadbent, 1958; Lavie and Tsal, 1994; Lavie, 1995; Serences and Kastner, 2014; Bruckmaier et al., 2020). Such capacity limits are both a function of task difficulty and the neural resources of the network. Task difficulty, conceptualized as perceptual load (Lavie and Tsal, 1994), has been shown to influence the deployment of attention to early visual processing (Lavie, 1995; Schwartz et al., 2005). The perceptual load can be manipulated by changing the density and nature of the stimuli (e.g., how similar the targets and the distractors are, and how blurred the images are). However, controlled manipulation of the neural resources of the network is not possible in-vivo. We have to turn to artificial neural networks which provide us with the opportunity to run synthetic experiments with full control over both aspects of the network's capacity.

In our experiments, artificial neural networks had to perform an object detection task: given an image containing one or many objects, and given a target object, the networks had to output if the target was present in the image. Network capacity was a function of both task difficulty, indexed by the stimulus complexity or the number of objects the network had to discriminate between, and the neural resources, indexed by the number of neurons in the network. The influence of these two aspects of network capacity was assessed on the usefulness of deploying feature-based attention (to increase the discriminability between the target object and the other objects) in the earlier stage of neural processing as compared to a later stage of neural processing. These experiments and their results are detailed in Chapters 3 and 4.

In the previous sections, we discussed our ability to affect the processing of incoming information across the visual field, based on the target of our search, to efficiently discriminate between the targets and the distractors, and identify the target quickly. We discussed the neural correlates of the spatially-global aspect of this mechanism, termed feature-based attention. We also discussed, given the target of the search, what neurons are modulated across the visual stream, and when modulation at the early stages of visual processing might be essential. This concludes my introduction to the research on feature-based attention - the mechanism underlying our ability to search in parallel throughout the visual field.

1.3 The influence of the regularities in scenes on visual search

The process of parallel search for the target, discussed in the previous sections, does not take the regularities of the scene into account. Elements of the scene could constrain the location of

the target. For example, in the introductory example (see Fig. 1.1), in addition to the knowledge of how the faucet looks different from the other objects, we have the knowledge that the faucet usually occurs on the kitchen counter. This knowledge can be used to constrain the search for the faucet to a limited area of the scene, making the search faster. I will now discuss the studies investigating the influence of the knowledge about the interactions, between the elements of the scene and the target of the search, on the process of visual search.

1.3.1 The influence of the target-distractor co-occurrences

Consider the search task from Biederman et al. (1973): Participants were shown an image (cutout) of an object and they had to indicate whether that object was present or absent in the subsequent image. This subsequent image was either a coherent scene from which the object was taken or a jumbled version of that scene (see Fig. 1.5A). When the target was absent, either the scene from which it was taken was shown (the 'Possible No' case) or another scene in which the target does not typically appear (the 'Impossible No' case) was shown. Participants were faster in reporting that the target object was not present when searching in the coherent scenes as opposed to searching in the jumbled scenes. They were also faster in declaring the target was absent in the scenes where the target did not typically appear. If the search for the target object proceeded as discussed in the previous sections, such a difference in reaction times would not be expected as the distractors stayed the same in the two scenes. This suggested that the knowledge about the relative positions of the distractors and object-scene correspondence (as they were faster in rejecting the non-corresponding scene) influenced the search. Related studies showed that object recognition was impaired when the expected appearance of the object in a scene was violated (e.g., a telephone floating in the sky; Biederman (1972, 1976)). These results constitute evidence that the knowledge about the structure of natural scenes (the spatial arrangement of objects) is used by humans while searching for and recognizing the target object (for an updated account see: Bar (2004); Zhang et al. (2020)).

How does the structure of the scene guide search? In Chun and Jiang (1998), it was proposed that in visual search tasks, this guidance, termed contextual cueing, could direct spatial attention to the location of the target predicted by the distractors. To test this account, participants' search performance was gauged as they detected a rotated T ($90°$ or $270°$) amongst rotated Ls (varying orientations: $0°$, $90°$, $180°$, $270°$) and indicated which way the T was oriented (see Fig. 1.5B). Crucially, unknown to the participants, throughout the blocks of the experiment, some of the distractor arrangements were repeated. Participants were faster in reporting the target within the repeated arrangements than within the non-repeated arrangements. With subsequent experimentation, it was shown that this contextual cueing effect was independent of distractor identity and was reliant on the arrangement of distractors predicting the target location. Additionally, when asked explicitly, participants could not discriminate between the arrangements that repeated and those that did not. Subsequent studies have also suggested that these repeated arrangements influence the initial guidance of search (around $200 - 400$ ms) than later processes such as the motor response (Johnson et al., 2007; Schankin and Schubö, 2009). These results suggest that the visual system stores information about the distractor arrangements which can be used to expedite the search for the target (Chun, 2000; Sisk et al., 2019).

How can such structure in natural scenes be used to guide attention to possible target locations as indicated above? Many models have been put forth to show how scene gist could be used to make predictions about the location of the target object and direct spatial attention and eye movements towards that location (Torralba et al., 2006; Ehinger et al., 2009; Wolfe et al., 2011b). In all these models the scene is processed/segmented using a 'non-selective visual pathway' (Wolfe

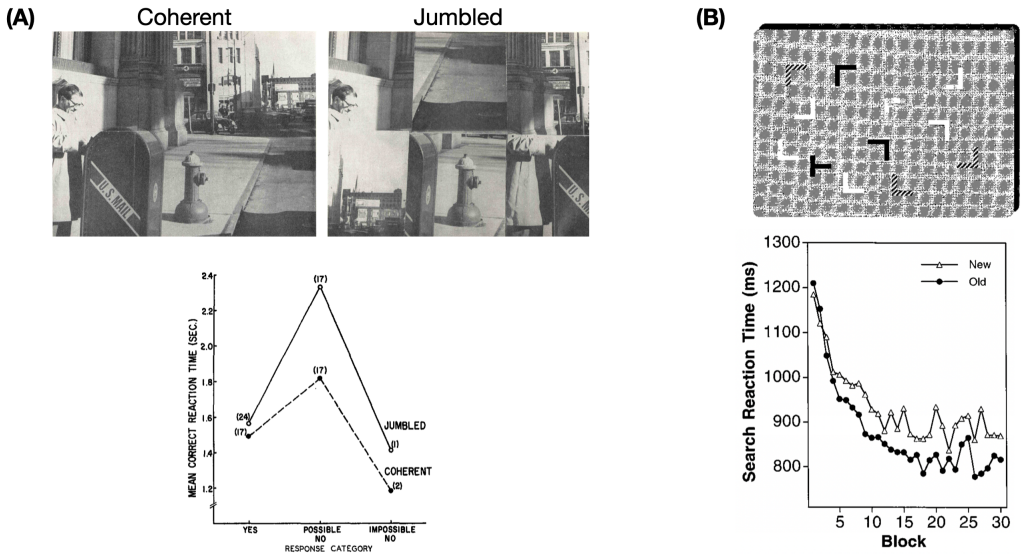

Figure 1.5: The influence of target-distractor co-occurrences on visual search. (A) Participants were faster in rejecting saying the target was absent in the coherent scenes as compared to the judgment of absence in the jumbled scenes. The spatial arrangement of objects and knowledge about the occurrence of objects in scenes influence human visual search. Figure adapted from Biederman et al. (1973). (B) Contextual cueing: participants were faster in searching for a rotated T among rotated Ls in displays that were repeated, suggesting that the memorized arrangement of distractors could guide search to the target's location faster. Figure adapted from Chun and Jiang (1998).

et al., 2011b) and the knowledge about which objects occur in what locations is used to modulate the saliency/priority maps that drive the attentional shifts. For example, regions of space such as roads or pavements, as opposed to the sky, are selectively processed further when searching for cars or people. These expectations are not restricted to regions of space but can also be driven by single objects in scenes. In Boettcher et al. (2018), it was suggested that anchor objects - objects that contain a high amount of information about local objects - could be used to restrict the coverage of the search for a target object. For example, while searching for small objects (e.g., phone, plates) in a living room, attention could be focused on top of a table as that is where such objects are usually present, as opposed to the lamp or the chair. In summary, knowledge about the co-occurrences between objects and regions of space or other objects in scenes can be used to constrain the coverage of spatial attention and make search more efficient (Wolfe et al., 2011a).

We discussed how the knowledge about target-distractor co-occurrences can lead to the efficient selection of the target. There is another aspect that can lead to efficient target selection - the predictability of distractor locations. If we think in terms of the generation of priority maps given knowledge about where the target could appear and where the distractors could appear, in addition to enhancing the priority of the locations where the target could appear, the locations where distractors could appear could also be suppressed (Leber et al., 2016; Chelazzi et al., 2019; van Moorselaar et al., 2021). This suppression ensures that spatial attention or eye movements refrain from shifting to the predictable distractor locations, effectively shifting their focus to the possible target locations. How the predictability of target and distractor locations could make search more

efficient has been well-studied. However, beyond the distractors predicting the targets (object-scene correspondence) and the target locations (contextual cueing), and the distractor locations themselves being predictable, regularities amongst distractors could also aid in visual search. I will now outline the nascent research into the influence of these distractor-distractor regularities on visual search.

1.3.2 The influence of the distractor-distractor co-occurrences

It has been well-known that distinct entities can get grouped and be perceived as single units due to the similarities between those entities, which could be due to factors such as them looking similar and being spatially proximal (Gestalt principles; Wertheimer (1923); Palmer and Rock (1994); McMains and Kastner (2010); Wagemans (2018)). For example, a display where the left side contains red dots and the right side contains green dots is perceived as containing two distinct regions, without substantial processing of the individual dots. Such grouping, due to visual similarity or connectedness, can lead to reports of participants underestimating the number of individual entities in the display (Frith and Frith, 1972; Ginsburg and Goldstein, 1987; Franconeri et al., 2009; He et al., 2009). Although the information is being lost, such grouping underlies highly efficient compression of the inputs: instead of coding the color and location of every dot (the number of bits scales with the number of dots), we can code that there are two regions with an estimate of the number of dots (the number of bits scales with the number of groups of dots). As discussed below, such grouping is not limited to items that are similar according to Gestalt principles but is generally applicable to any co-occurring items.

In Brady et al. (2009), it was shown that co-varying random colors in a display can be grouped, compressing the inputs, leading to the participants being able to memorize more information from the inputs. Notably, the covariance between colors was learned during the experiment. More generally, it has been proposed that such statistical learning of regularities could lead to the grouping of items (Fiser and Aslin, 2001; Fiser and Lengyel, 2019). For example, it has been shown that the presence of spatially co-occurring colored dots can lead to an underestimation of the number of dots in the display (Zhao and Yu, 2016). It has also been shown that the statistical learning of spatially co-occurring shapes can lead to two random shapes being grouped (Lengyel et al., 2021), leading to 'object-like' effects such as the spreading of attention from a shape to its co-occurring partner (termed object-based attention, Egly et al. (1994)). As a general principle, it has been suggested that co-occurrence-based grouping could lead to compressed, efficient, representations in the visual system (Brady et al., 2011; Brady and Tenenbaum, 2013; Kaiser et al., 2019). Such grouping-based efficient coding could occur at the level of natural objects and other elements of natural scenes such as surfaces. This efficient coding might underlie our ability to capture the gist of scenes rapidly (Cohen et al., 2016; Haun et al., 2017), which might aid in conditioning further visual processing of relevant objects or regions of the scene.

How could such efficient coding of co-occurring distractor elements in a scene lead to efficient search? In Kaiser et al. (2014), participants searched for an object (cued with the corresponding word; e.g., seahorse) in transient displays (200 ms display time) containing pairs of natural objects that co-occur in fixed arrangements (the 'regular' condition; e.g., egg on an egg cup, lamp on top of a table) or in displays containing the same pairs with the relative position of the objects within the pairs being swapped (the 'irregular' condition; e.g., egg cup on top of an egg). Participants were more accurate in searching for the targets in the regular displays than in the irregular displays (see Fig. 1.6). These displays were also shown to participants in an fMRI experiment. Instead of the target and the foil, two houses were presented in the displays. Activity specifically in the place-selective parahippocampal place area was higher when the regular displays were presented

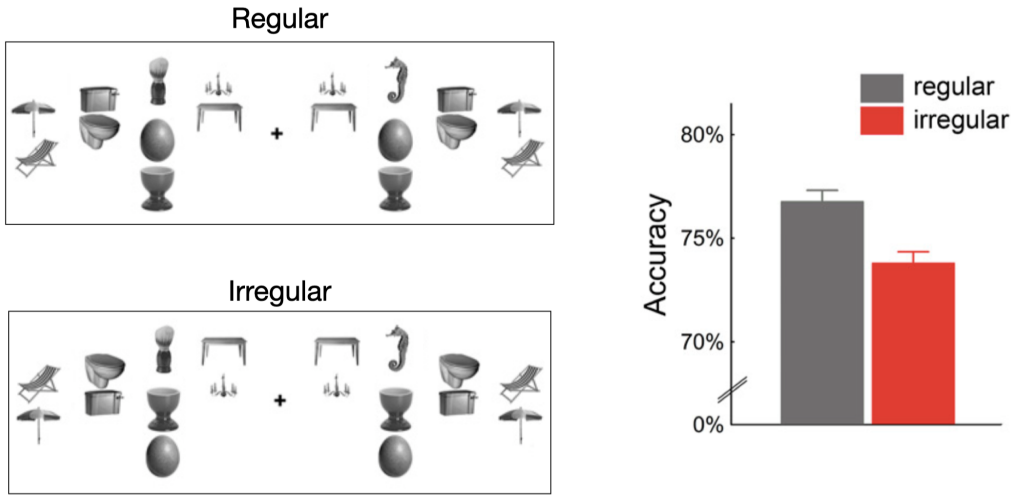

Figure 1.6: The influence of co-occurring distractors on visual search. In transiently-presented displays where the distractor object pairs are shown in their regular arrangements (e.g., egg on an egg cup), the search for a cued target (e.g., seahorse) is more accurate than a search in the displays where the distractor object pairs are shown in their irregular arrangements (e.g., egg cup on an egg). Knowledge about distractor co-occurrences influences visual search. Figure adapted from Kaiser et al. (2014).

than when the irregular displays were presented, signaling lower competition from the distractors to the house representations in the regular condition. It was concluded that co-occurrences amongst the distractors could lead to them being grouped, effectively reducing the number of distractors, reducing their competition with the target, leading to better detection of the target. This is the only study that explicitly assessed the influence of the regularities in the distractor - that could not predict the location or identity of the target but could reduce the competition faced by the target by effectively reducing the number of distractors - on the efficiency of search.

There are a couple of caveats in linking the results in Kaiser et al. (2014) to the conclusions being made about the usefulness of the distractor regularities. The search accuracy might have been worse in the irregular displays because participants' attention might have initially been grabbed by the irregular pairs which violate the expectations from a lifetime of experience. In using natural object pairs, it is hard to create a control condition where there are no expectations about the co-occurrence of objects. Even in the 'Shuffled' condition in that study, where the partners were swapped amongst the object pairs, but the relative positions were kept constant (e.g., egg on a table, and lamp on egg cup), expectations about which object co-occurs with what other object (semantic co-occurrences) were being violated. To avoid the potential influence of such violations on the results of such an experiment, novel objects which participants are not familiar with, need to be used (to avoid semantic associations) and the co-occurrences of interest for different groups of objects need to be learned anew by the participants (to avoid violating expectations from a lifetime of experience). In Chapter 5, I outline the results from a set of experiments aimed at assessing if scenes with abstract distractor shapes, co-occurring in pairs, as the participants perform a search for cued targets, facilitate a more efficient search than scenes with abstract distractor shapes that do not co-occur. In addition to behaviourally relating distractor regularities to enhanced search, in Chapter 6, I outline observations from EEG recordings, as participants per-

formed the same search task, that help us assess the time-course of the influence of the distractor regularities on visual search, informing us about the underlying processes.

In the previous sections, we discussed our ability to register the regularities in our environment - how objects co-occur with other objects and certain places in scenes - and how those regularities can be used not only to predict where targets of search might occur but also to efficiently represent the scene to reduce the competition posed by the distractors. This concludes my introduction to the research on the influence of the regularities in the environment on visual search.

1.4 Outline of the thesis

The subsequent chapters contain details about our research contributions to the field, which are followed by a general discussion of all the results in Chapter 7. Now I state the highlights of the next five chapters.

In Chapter 2, I present our assessment of the neural signatures of spatially-global feature-based modulation at the level of object categories. We used a controlled experimental design and stimulus set that overcame potential pitfalls in previous studies. Using fMRI, we observed the modulation of body representations in the object-selective visual cortex, depending on the target of the search, even though those bodies were presented in task-irrelevant locations. This observation supports the idea that human bodies can be treated as visual features that could avail of the visual system's parallel search capabilities.

In Chapters 3 and 4, I present our investigations into the influence of cue-driven feature-based modulations on network performance and its dependence on the capacity limits of the network. Using artificial neural networks and stimulus sets to manipulate the representational capacity of the network, we showed that the cue-driven modulations in the early stages of processing provide additional performance only when the network is capacity-limited. In Chapter 3, we also found that the optimal modulations do not resemble the feature-similarity gain model (FSGM), an influential model of feature-based modulations in the primate visual system. These observations confirmed the dependence of early selection of information on the capacity of the network.

In Chapter 5, I present our assessment of the usefulness of co-occurring distractors in optimizing visual search. We used a controlled experimental design and stimulus set that overcame potential pitfalls in a previous study. Using large-sample online behavioral experiments, we observed that the search for shapes in scenes containing co-occurring distractor shapes was more efficient than through scenes that did not contain co-occurring distractor shapes. This observation supports the idea that regularities amongst distractors can be used to reduce the complexity of the scene thereby making the search easier.

In Chapter 6, building over the results in Chapter 5, I present our investigations into the neural correlates - the influence on attentional orienting and representations of the target and distractors - of the increased search efficiency associated with the exposure to the co-occurring distractors during the search. Using EEG, we observed that the attention orienting component of the event-related response, the N2pc, was higher in participants who had higher search efficiency in the scenes with the co-occurring shapes. This observation indicates that the scene complexity reduction due to distractor co-occurrences (observed in Chapter 5) might happen rapidly in the visual system and influence the earliest cue-driven voluntary shift of attention during visual search.

Chapter 2

Body shape as a visual feature: evidence from spatially-global attentional modulation in human visual cortex

Feature-based attention supports the selection of goal-relevant stimuli by enhancing the visual processing of attended features. A defining property of feature-based attention is that it modulates visual processing beyond the focus of spatial attention. Previous work has reported such spatially-global effects for low-level features such as color and orientation, as well as for faces. Here, using fMRI, we provide evidence for spatially-global attentional modulation for human bodies. Participants were cued to search for one of six object categories in two vertically-aligned images. Two additional, horizontally-aligned, images were simultaneously presented but were never task-relevant across three experimental sessions. Analyses time-locked to the objects presented in these task-irrelevant images revealed that responses evoked by body silhouettes were modulated by the participants' top-down attentional set, becoming more body-selective when participants searched for bodies in the task-relevant images. These effects were observed both in univariate analyses of the body-selective cortex and in multivariate analyses of the object-selective visual cortex. Additional analyses showed that this modulation reflected response gain rather than a bias induced by the cues and that it reflected enhancement of body responses rather than suppression of non-body responses. Finally, the features of early layers of a convolutional neural network trained for object recognition could not be used to accurately categorize body silhouettes, indicating that the fMRI results were unlikely to reflect selection based on low-level features. These findings provide the first evidence for spatially-global feature-based attention for human bodies, linking this modulation to body representations in high-level visual cortex[1].

[1]This chapter has been adapted from - Thorat, S., & Peelen, M. V. (2022). *Body shape as a visual feature: evidence from spatially-global attentional modulation in human visual cortex.* NeuroImage, 255, 119207.

2.1 Introduction

The capacity limits of the human visual system require selecting visual input for further processing and conscious access (Carrasco, 2011; Chun et al., 2011). One way to do this is to select specific locations of the visual field through spatial attention and eye movements. However, when searching for task-relevant objects in our environment, the location of these objects is typically not yet known. In this case, selection may operate at the level of visual features, using a selection mechanism termed feature-based attention (Maunsell and Treue, 2006). To be an effective selection mechanism, feature-based attention would need to operate in parallel across the whole or part of the visual field, to then guide spatial attention to the location of the target object (Wolfe, 1994). While this could be a plausible mechanism of attentional selection, it raises a core question: what are the features of feature-based attention?

At a neural level, it has been proposed that feature-based attention may be restricted to features to which sensory neurons are systematically tuned (Maunsell and Treue, 2006). Accordingly, the neural mechanisms of feature-based attention have been studied extensively with experiments involving low-level features for which such tuning has been established, such as the orientations of Gabor patches (Kamitani and Tong, 2005; Liu et al., 2007a; Jehee et al., 2011) and the movement direction of random dot patterns (Treue and Trujillo, 1999; Saenz et al., 2002; Serences and Boynton, 2007). These experiments assessed how making one feature task-relevant influenced the responses of neurons that were selective or non-selective to that feature. A common finding was that attending to a low-level feature increased the responses of neurons selective to that feature and decreased the responses of neurons non-selective to that feature (Maunsell and Treue, 2006). Crucially, such modulations were shown to occur for stimuli presented in spatially-unattended and task-irrelevant locations (Treue and Trujillo, 1999; Saenz et al., 2002; Serences and Boynton, 2007; Zhang and Luck, 2009), providing evidence for a spatially-global mechanism of feature-based attention that can be distinguished from the effects of spatial attention.

In the present study, we tested whether global attentional modulation can similarly be observed for the shape of the human body, a category of high social and biological significance that is selectively represented in high-level visual cortex (Downing et al., 2001; Peelen and Downing, 2005). Behavioral studies have shown that bodies gain preferential access to awareness (Stein et al., 2012) and automatically attract attention (Downing et al., 2004; Ro et al., 2007). There is also behavioral evidence for spatially-global attention effects for bodies: in a series of studies, spatial attention was captured by body silhouettes when participants searched for people in scenes presented in different parts of the visual field (Reeder and Peelen, 2013; Reeder et al., 2015). Finally, an fMRI study reported spatially-global modulation of multivoxel activity patterns distinguishing natural scenes with people from natural scenes with cars (Peelen et al., 2009). However, in that study, the relative contributions of scene context and body, face, and car features could not be distinguished, such that it remains unknown whether feature-based attention effects exist for human bodies.

Here, we used fMRI to provide the first direct test of spatially-global attentional modulation of body processing in the visual cortex. Participants detected the presence of bodies or one of five other categories (beds, bottles, cars, chairs, lamps) in task-relevant vertically-aligned images, thereby manipulating the top-down attentional set. To test for spatially-global attentional modulation, all analyses focused on responses evoked by objects that were concurrently presented at locations that were never relevant for the object detection task across three experimental sessions (Fig. 2.1A). The inclusion of five non-body categories reduced the possibility that participants could use a low-level feature to detect the presence of bodies, for example by looking for vertical (bodies) vs horizontal (e.g., cars) stimuli: lamps and bottles shared the vertical orientation with

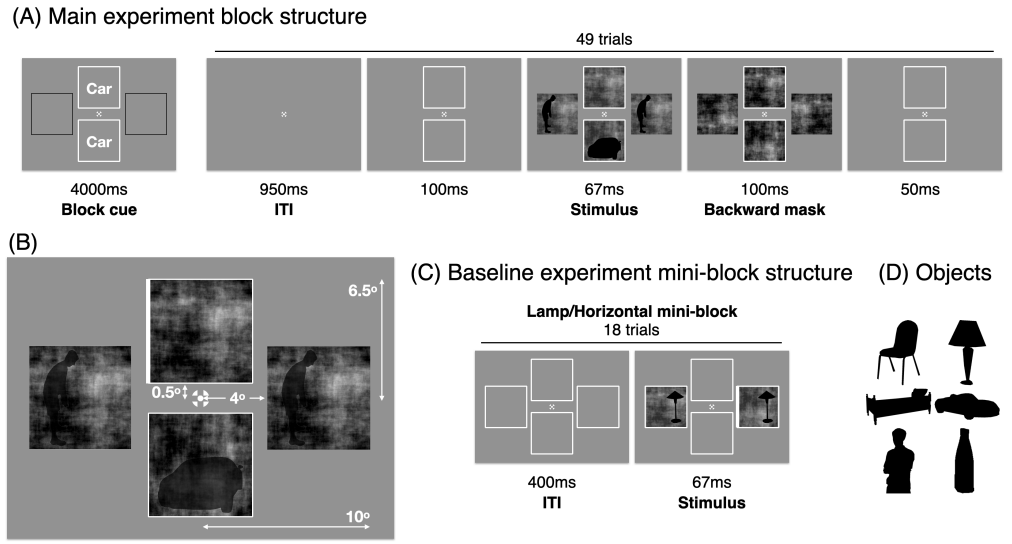

Figure 2.1: Experimental design. (A) The main experiment was designed to reveal the modulatory influence of feature-based attention on object responses evoked by stimuli presented at task-irrelevant locations (horizontal boxes). In each block (49 trials), participants had to search for the cued object category (e.g., car) in the vertical boxes, while objects were simultaneously presented in the horizontal boxes. (B) The spatial layout of the search display. (C) The baseline experiment was designed to obtain prototypical object category responses. Responses evoked by task-irrelevant objects in the main experiment were compared to these responses. Participants had to indicate if one of the edges of the two boxes thickened. The object category and location (horizontal or vertical boxes) varied across the mini-blocks. Unlike in the main experiment, the stimuli were not backward masked to increase visibility. (D) Exemplars of the six categories: chairs, lamps, beds, cars, human bodies, bottles. Fifty exemplars were used for each category.

bodies (Fig. 2.1D). To further reduce this possibility, each category was represented by a large and diverse set of exemplars cropped out of scene photographs. Finally, the use of silhouettes avoided possible low-level differences between categories in texture and color, and also excluded the possibility that attention was guided by facial features (Störmer et al., 2019) instead of the shape of bodies.

2.2 Methods and Materials

2.2.1 Participants

Twenty-three healthy adult volunteers with normal or corrected-to-normal vision gave written informed consent and participated in the experiment. All participants took part in three experimental sessions, on different days. One participant was excluded because of low performance on the visual search task (the difference between the proportion of false alarms and hits was lower than two standard deviations from the average difference). Twenty-two participants (mean age: 25.36 years; age range: $20 - 32$ years; 11 female) were included in the reported analyses. The

study was approved by the local ethics committee (CMO Arnhem-Nijmegen).

2.2.2 Experimental paradigm

In the main experiment, on each trial, the display contained two boxes in the horizontal and vertical locations (Fig. 2.1). The vertical boxes had a white bounding frame, signifying their relevance. Each of the four boxes contained a random image containing the average power spectrum of the objects from the six categories with random phases. Objects were mixed with these random images. On each trial, an exemplar from one of the six categories could be presented in one of the two vertical boxes ($1/7$ probability each) or no object would be presented ($1/7$ probability). Simultaneously, an exemplar from one of the six categories could be presented in both the horizontal boxes ($1/7$ probability each) or no object would be presented ($1/7$ probability). Each block consisted of 49 trials to fill the co-occurrence matrix of the horizontal and vertical object conditions, such that the conditions presented in the horizontal and vertical boxes were orthogonal to each other.

In each block of the main experiment, participants would either search for one of the six categories in the vertical boxes or would detect a thickening of the frames of the bounding boxes in the vertical location. Participants pressed the response button when the cued object category was shown in one of the vertical locations, which occurred on $7/49$ trials. In the thickening condition, participants had to indicate, by pressing the response button, when one of the sides of the two bounding boxes became thicker than the others (thickening occurred on $7/49$ trials in all blocks). Data from these thickening task blocks in the main experiment were not further analyzed. The simultaneously presented objects in the horizontal boxes were always task-irrelevant. Each run contained four blocks, all containing a different search condition, such that across the seven search runs in each fMRI session each search block occurred four times. Feedback about search performance was provided at the end of each block.

In the baseline experiment, in different blocks, exemplars of one of the six categories or scrambled exemplars of one of the six categories were presented in both the boxes in either the horizontal or vertical locations (the other location left empty). These objects were mixed with a random image containing the average power spectrum of the objects from the six categories with random phases. The seven object conditions (six object categories and a scrambled objects condition containing a mix of scrambled objects from the six categories) and two presentation locations were blocked into mini-blocks containing 18 trials each. In each mini-block, participants had to search for thickening of the frames of the boxes where objects were being presented ($1/7$ probability of presence; each pair of thickening events had at least two non-thickening trials between them). Each block contained seven mini-blocks, with distinct object-location pairing, such that across the four blocks in each baseline experiment run, each type of block occurred twice. At the end of each block, performance feedback was provided.

Each participant attended three experimental sessions. The first behavioral session required each participant to get exposed to the entire set of objects followed by the completion of one run of the baseline experiment and two runs of the main experiment. The second and the third sessions involved fMRI. In each of those sessions, the participant first browsed through the entire set of objects at their own pace and then performed one run of the main experiment during the anatomical scan. This was followed by the functional recordings as the participants performed one run of the baseline experiment followed by four runs of the main experiment followed by one run of the baseline experiment followed by three runs of the main experiment.

2.2.3 Stimuli

The stimulus presentation dimensions are shown in Fig. 2.1B. We acquired 50 exemplar silhouettes in real-world poses for each of the six categories of interest (beds, bottles, cars, chairs, lamps, and human bodies; shown in Fig. 2.1D). We obtained scenes containing the relevant objects from the SUN2012 database (Xiao et al., 2010) and Google images which were "Labelled for non-commercial reuse with modifications", cropped out the objects, scaled them such than on one of the axes of the objects extended throughout the image, and converted them to silhouettes.

On each trial, the chosen exemplars were shown in the boxes, embedded in noise as mentioned above. The location of the objects within the boxes was jittered to increase variability. Objects that extended throughout the image horizontally were presented in one of three places within the box: touching the upper side, centered, or touching the lower side of the box. Similarly, objects that extended throughout the image vertically could be placed touching the left side, centered, or the right side of the box. The horizontally-placed boxes in the display contained the same stimulus (Fig. 2.1C).

2.2.4 fMRI data acquisition and preprocessing

Functional (echo-planar imaging (EPI) sequence; 66 slices per volume; resolution: $2 \times 2 \times 2$ mm; repetition time (TR): 1 s; time to echo (TE): 35.2 ms; flip angle: $60°$) and anatomical (MPRAGE sequence; 192 sagittal slices; TR: 2.3 s; TE: 3.03 ms; flip angle: $8°$; $1 \times 1 \times 1$ mm resolution) images were acquired with a 3 T MAGNETOM Skyra MR scanner (Siemens AG, Healthcare Sector, Erlangen, Germany) using a 32-channel head coil.

The functional data were analyzed using MATLAB (2017a) and SPM12. During preprocessing, within each session, the functional volumes were realigned, co-registered to the structural image, re-sampled to a $2 \times 2 \times 2$ mm grid, and spatially normalized to the Montreal Neurological Institute 305 template included in SPM12. A Gaussian filter (FWHM 3 mm) was applied to smooth the images.

2.2.5 Statistical analysis

For each participant, general linear models (GLMs) were created to model the conditions in the experiment. All trials were included in the analysis. Regressors of no interest were also included to account for differences in the mean MR signal across scans and for head motion within scans. In the main experiment, the GLM included regressors for the 49 conditions of interest: 7 attention blocks x 7 stimulus conditions presented in the task-irrelevant (horizontal) location. In the baseline experiment, the GLM included regressors for the 14 conditions of interest: 7 stimulus conditions x 2 locations.

In the univariate analysis, the regression weights (betas) from the GLM were compared between conditions after averaging across the voxels of a region of interest (ROI). In the multivariate analysis, the pattern of betas from the GLM across the voxels of an ROI was compared between conditions using Kendall's tau correlation coefficient (τ) as a metric for similarity. Before comparing the betas between the main and baseline experiments, the mean across all main experiment condition betas was subtracted from those condition betas (separately for each voxel), and the mean across all baseline experiment condition betas was subtracted from those condition betas.

2.2.6 Regions of interest

In the multivariate analysis, we focused on two ROIs, the lateral occipital cortex (LOC) and the early visual cortex (EVC). The LOC ROI was defined using a group-constrained subject-specific method (Fedorenko et al., 2010). The group-level ROI was defined by first contrasting the average response to the 6 object categories with the response to the scrambled objects in the baseline experiment. Threshold-free cluster enhancement (TFCE; Smith and Nichols (2009)) with a permutation test was used to correct for multiple comparisons (at $p < 0.05$) across the whole brain. The resulting voxels were intersected with the lateral occipital cortex ROI from Julian et al. (2012) to obtain the group-level LOC ROI. Then, for each participant, the 1000 most object-selective voxels (average object response - scrambled stimulus response, in the baseline experiment horizontal conditions) within the group-level LOC ROI were selected for further analysis. The EVC ROI was defined at the individual participant level as the 1000 most responsive voxels (average object response > 0, in the baseline experiment horizontal conditions) in Brodmann area 17 (corresponding to V1; Wohlschläger et al. (2005)). Brodmann area 17 was taken from the Brodmann atlas available in SPM12.

In the univariate analysis we focused on two body-selective ROIs, the extrastriate body area (EBA; Downing et al. (2001)) and the fusiform body area (FBA; Peelen and Downing (2005)). The ROIs were defined using the method described above for LOC. The group-level ROI was defined by first contrasting the response to bodies with the average response to the other 5 categories in the baseline experiment. TFCE was used to correct for multiple comparisons (at $p < 0.05$) across the whole brain. The resulting voxels were intersected with the extrastriate body area ROI from Julian et al. (2012) to obtain the group-level EBA ROI and the fusiform face area (FFA) ROI from Julian et al. (2012) to obtain the group-level FBA ROI (FBA ROI is not provided, but the FFA and FBA closely overlap at the group-level; Peelen and Downing (2005)). Then, for each participant, the 20 most body-selective voxels (body response - average response to other objects, in the baseline experiment horizontal conditions) within the group-level ROIs were selected for further analysis.

2.2.7 Multivariate analysis approach

In the multivariate analyses, we correlated multivoxel activity patterns evoked by the task irrelevant objects in the main experiment with multivoxel activity patterns evoked by the clearly visible objects in the baseline experiment, using Kendall rank-ordered correlation (τ). We expect to find stronger correlations between corresponding object categories (e.g., between bodies in the main experiment and bodies in the baseline experiment), than between non-corresponding categories (e.g., between bodies in the main experiment and beds in the baseline experiment). As such, the difference between corresponding and non-corresponding category correlations is a measure of category processing (Peelen et al., 2009), analogous to decoding accuracy. Here, we computed proximity to the categories in the baseline experiment as the correlation with that category minus the correlation with the other categories in the baseline experiment. For example, for bodies, the proximity to bodies (in the baseline experiment) is the correlation between bodies in the main experiment and bodies in the baseline experiment minus the average correlation between bodies in the main experiment and the other five categories in the baseline experiment.

2.2.8 Image-based discriminability approach

Representations of the exemplars in the layers of a convolutional neural network (trained for object recognition in natural images; CNN; AlexNet: Krizhevsky et al. (2012)) were used to test

for image-based categorizability differences across the categories. Output activations at each layer corresponding to 50 exemplars of each of the six categories, embedded in noise as in the fMRI experiment, in the three possible locations defined by the shapes (see the subsection on Stimuli), were extracted. Balanced linear support vector machines (SVM) were trained to classify between the images of one category (150 images each) as opposed to the other categories. 10-fold cross-validated classification accuracies were reported for each category for each layer of the CNN.

2.3 Results

In the main experiment, participants (N= 22) detected the presence of object silhouettes belonging to one of six categories (Fig. 2.1D), in different blocks. Throughout the experiment, only the vertically-aligned locations were relevant for the detection task (Fig. 2.1A). Each block started with a category cue (e.g., "Car") indicating the target category for that block (Fig. 2.1A), followed by 49 object detection trials. In 42 trials (6/7th), one of the two task-relevant locations contained a briefly-presented object (67 ms) within phase-scrambled noise (Fig. 2.1B), with each category presented equally often (7 trials each). In the other 7 trials (1/7th) no object was presented.

Crucially, in 6/7th of the trials, two objects were simultaneously presented in the horizontally-aligned locations (Fig. 2.1A). These objects were briefly presented (67 ms), embedded in noise, and backward masked. Objects at these locations were never relevant for the participants and could thus be completely ignored. The occurrence probabilities of the categories were the same as for the task-relevant locations. The 7 vertical and 7 horizontal conditions were fully crossed within each block, resulting in 49 trials, which were presented in random order. Trials were coded according to the categories presented in the horizontally aligned (task-irrelevant) locations, as these were the focus of our analyses.

2.3.1 Task performance

Averaged across the two fMRI sessions and object search blocks, participants had a hit rate of 78.3% and a false alarm rate of 5.6%, resulting in an average d' of 2.7 (beds: 2.0; cars: 2.4; bottles: 2.6; bodies: 2.9; chairs: 2.9; lamps: 3.3).

2.3.2 Univariate results in EBA and FBA

Previous research has shown that bodies evoke a selective univariate response in two focal regions of the high-level visual cortex: the extrastriate body area (EBA; Downing et al. (2001)) and the fusiform body area (FBA; Peelen and Downing (2005)). Here, EBA and FBA were defined based on responses in the baseline experiment (see Material and Methods). We tested for spatially-global attention effects for bodies in these ROIs by comparing body-selective responses in EBA and FBA evoked by task-irrelevant bodies across target detection blocks in the main experiment. Betas were averaged across the voxels of each ROI to acquire one beta per condition for each ROI. For each category, the beta corresponding to within-block trials in which no objects were presented was subtracted to account for block effects. Responses to non-body objects and non-body detection blocks were averaged, such that we had 4 values for each ROI: body and non-body stimuli, presented in the body and non-body detection blocks. The difference between body and non-body stimuli within each block is a measure of body selectivity.

A 2 (ROI) $\times$2 (attention: body, other categories) ANOVA on body selectivity (response to bodies minus average response to other categories) revealed a main effect of attention ($F_{1,21} =$

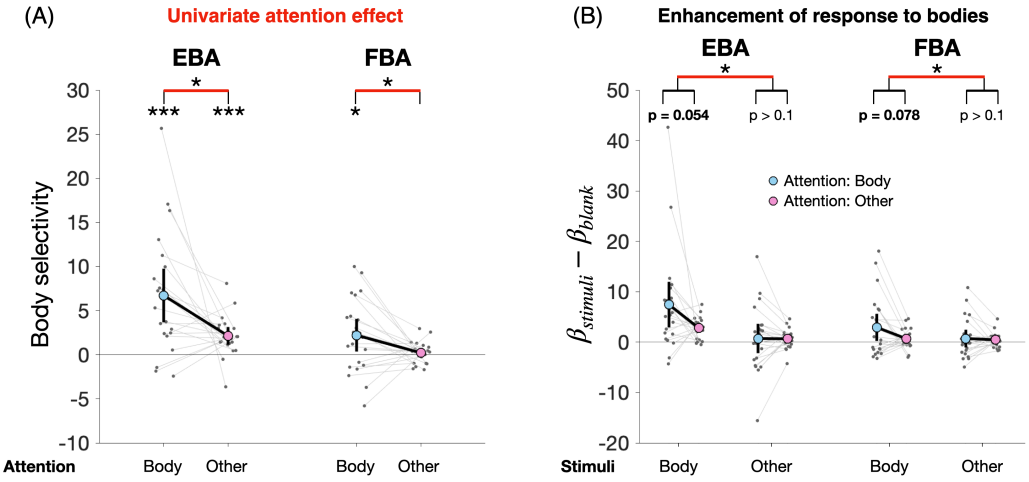

Figure 2.2: Univariate attention effect in body-selective ROIs. (A) Body-selectivity (response to body - average response to other objects) was higher when bodies were attended, in both ROIs. This provides further evidence for spatially-global attentional modulation for body silhouettes. (B) Across ROIs, the response to bodies (corrected for block-wise differences by subtracting the corresponding blank responses) was enhanced while the responses to other categories remained unchanged. Error bars indicate 95% confidence intervals for the measures indicated on the y-axes. The asterisks indicate p-values for the t-tests of the corresponding comparisons (*p$<$ 0.05, **p$<$ 0.01, ***p$<$ 0.001).

7.2, p $=$ 0.014), reflecting stronger body selectivity in body attention blocks than non-body attention blocks (Fig. 2.2A). This attention effect interacted with ROI ($F_{1,21} = 4.6$, p $=$ 0.043), being stronger for EBA than FBA. When analyzed separately, both EBA and FBA showed an attention effect, such that body selectivity was higher in body detection blocks than in other category detection blocks (EBA: $F_{1,21} = 7.4$, p $=$ 0.013; FBA: $F_{1,21} = 4.4$, p $=$ 0.049; Fig. 2.2A). In EBA, bodies evoked a selective response in both the body detection blocks ($t_{21} = 4.6$, p $= 2E - 4$) and the other category detection blocks ($t_{21} = 4.4$, p $= 2E - 4$), while in FBA body selectivity was only positive in the body detection blocks ($t_{21} = 2.5$, p $=$ 0.02; other category detection blocks: $t_{21} = 0.8$, p $=$ 0.42).

The attention effect for bodies in EBA and FBA could reflect enhanced responses to bodies presented in body detection blocks, but may also (or additionally) reflect reduced responses (suppression) to the other categories presented in body detection blocks. To test these alternatives, we compared body and object-evoked responses across the body and object-detection blocks (after subtracting the response to blanks within each block). Averaged across ROIs, there was a higher response to bodies in body detection blocks than in other category detection blocks (paired t-test, $t_{21} = 2.1$, p $=$ 0.05; Fig. 2.2B). There was no evidence that the response to the other objects was suppressed, with equally strong responses in both blocks (paired t-test, $t_{21} = 0.19$, p $=$ 0.85). These effects were also observed, though weaker, in each ROI separately (statistics provided in Fig. 2.2B).

These results provide the first evidence for spatially-global attentional modulation for body silhouettes, show that these effects are strongest in EBA, and link these effects to the enhancement of body responses rather than suppression of non-body responses.

2.3.3 Multivariate results in LOC

Previous studies have shown that multivoxel activity patterns in object-selective cortex distinguish between object shapes (Haushofer et al., 2008; De Beeck et al., 2008; Eger et al., 2008). This gave us another opportunity to test for spatially-global effects of attention, including for non-body categories. Here, instead of body selectivity, we used proximity (Pr) as the dependent measure. Proximity was based on correlations between response patterns in the main experiment and response patterns in the baseline experiment, following previous work (Peelen et al., 2009). Proximity reflects how similar a category's response pattern in the main experiment is to a category's response pattern in the baseline experiment, relative to the other categories in the baseline experiment (Materials and Methods). For example, for bodies, the proximity to bodies (in the baseline experiment) is the correlation between bodies in the main experiment and bodies in the baseline experiment minus the average correlation between bodies in the main experiment and the other five categories in the baseline experiment.

2.3.4 Attentional modulation for bodies in LOC

The proximity to bodies is shown in Fig. 2.3A. A 2 (attention: body, other categories) $\times 2$ (stimulus presented: body, other categories) ANOVA revealed a significant interaction ($F_{1,21} = 30.4$, $p = 2E-5$), reflecting a stronger difference between the proximities for body and non-body categories when participants attended to bodies ($t_{21} = 9.9$, $p = 2E-9$) than when they attended to the other categories ($t_{21} = 8.1$, $p = 5E-8$). These results provide further evidence for spatially-global modulation for bodies.

The attention effect for bodies in LOC could reflect enhanced proximity to bodies for the bodies presented in body detection blocks, but may also (or additionally) reflect reduced proximity to bodies (suppression) for the other categories presented in body detection blocks. To test for body-selective enhancement, we compared the proximity (to bodies in the baseline experiment) for bodies with the corresponding proximity of other objects in the body detection blocks. To account for overall differences between blocks (e.g., related to the cue or block-based attentional bias), we subtracted the proximity to bodies for the within-block trials in which no objects were presented. Results showed that proximity to bodies was significantly enhanced for bodies presented in the body detection blocks as compared with bodies presented in the other detection blocks ($t_{21} = 3.5$, $p = 0.002$; blue comparison in Fig. 2.3B). There was no evidence for suppression: proximity to bodies was not different for objects presented in the body detection blocks as compared with objects presented in the other detection blocks ($t_{21} = 0.3$, $p = 0.78$; green comparison in Fig. 2.3B). The difference between these effects (red comparison in Fig. 2.3B) corresponds to the same multivariate attention effect as shown in Fig. 2.3A. These results show that the multivariate attention effect was primarily driven by the enhancement of body-selective response patterns, in line with the univariate results (Fig. 2.2).

2.3.5 The relationship between attentional modulation and univariate body selectivity of LOC voxels

Next, we tested whether the multivariate attention effect observed for bodies in LOC depended on the (univariate) body-selectivity of voxels included in LOC. To this end, we computed the multivariate attention effect for bodies in an ROI that consisted of LOC voxels that responded less strongly to bodies than to other categories in the baseline experiment (on average 330.8 out of the original 1000 voxels satisfied this criterion). Results were compared with a size-matched ROI consisting of randomly-sampled LOC voxels (size-matching done within each participant;

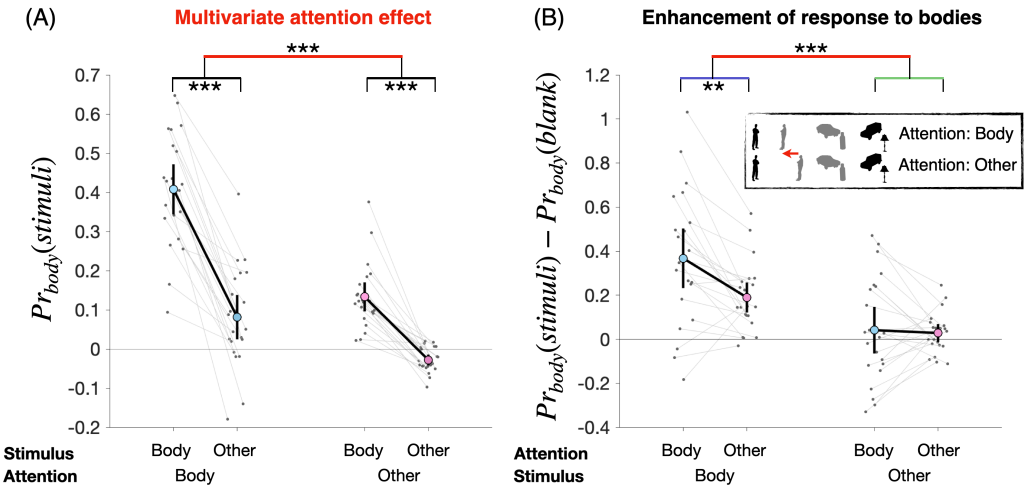

Figure 2.3: Probing the multivariate attention effect for bodies in LOC. (A) The selective proximity for bodies (proximity to bodies for Body vs Other) is higher when bodies are attended, which is evidence for a multivariate attention effect in LOC (comparison highlighted in red), reflecting response gain. (B) Proximity (to bodies) of bodies and other categories were compared between the body attention blocks and the other category attention blocks, corrected for block-wise differences by subtracting the proximity (to bodies) to blank responses within blocks. When bodies were attended, the proximity of bodies was enhanced, whereas the proximity of the other categories was not affected (inset: gray objects correspond to attention-dependent representations and black to benchmark representations). This indicated that the multivariate attention effect for bodies in LOC (the comparison corresponding to the red bar) was driven primarily by enhancement of body-selective response patterns when bodies were attended. 95% confidence intervals for the measures indicated on the y-axes are shown. The asterisks indicate the p-values for the t-tests of the corresponding comparisons (*p< 0.05, **p< 0.01, ***p< 0.001). Blue: attentional modulation for bodies; green: attentional modulation for other categories.

sampled 100 times). Attentional modulation was computed in the same way as for the whole LOC in the original analysis (red comparison in Fig. 2.3). Attentional modulation was stronger for the size-matched ROI than the non-selective ROI ($t_{21} = 3.1$, $p = 0.006$). However, attentional modulation was significant even in the non-selective ROI ($t_{21} = 2.1$, $p = 0.047$). These results suggest that the attentional modulation in LOC was partly but not exclusively driven by body-selective voxels.

2.3.6 Attentional modulation for non-body categories in LOC

Using the multivariate analysis framework outlined above for bodies, we can similarly test for spatially-global attentional modulation for the other categories. For each non-body category, we computed the multivariate attention effect as was done for bodies, now using the proximity to that category in the baseline experiment. To reduce the complexity of the ANOVA and the corresponding visualization of the data, we used selective proximity as the dependent measure. Selective proximity is the proximity difference between the corresponding and non-corresponding categories (e.g., the difference between the two left-most data points in Fig. 2.3A). As an intuition for what this new measure represents, note that in the case of bodies, selective proximity is

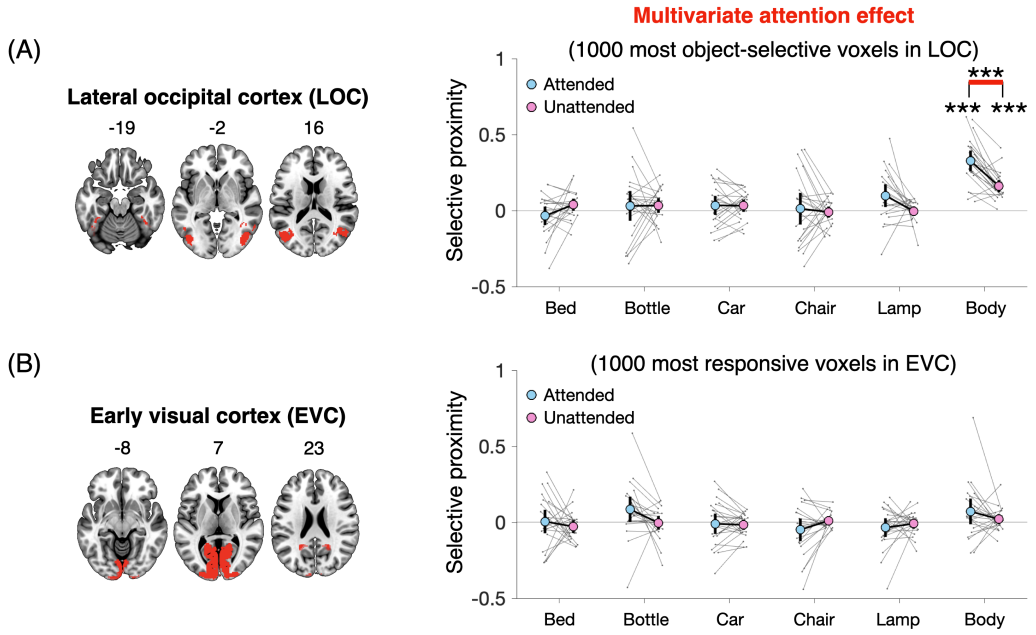

Figure 2.4: Multivariate attention effect. The selective proximities, for the attended and unattended conditions, are shown for the six categories in the two ROIs. The multivariate attention effect is the difference between attended and unattended selective proximity (comparison highlighted in red). A) In LOC, we find evidence for attentional modulation of the selective proximities of bodies. B) No attentional modulation was found in EVC. Error bars indicate 95% confidence intervals for the selective proximities. The asterisks denote Bonferroni corrected p-values for the t-tests of the twelve comparisons related to selective proximities, and Bonferroni corrected p-values for the t-tests of the six comparisons related to selective proximity modulations (*p$<$ 0.05, **p$<$ 0.01, ***p$<$ 0.001).

analogous to the body selectivity measure in the univariate analysis.

In LOC, a 6 (category of interest) $\times$2 (category attended/unattended) ANOVA on these selective proximities revealed a significant interaction ($F_{5,105} = 3.9$, $p = 0.003$; Fig. 2.4A), indicating that attention differentially affected the selective proximity of the six categories. Six paired-sample t-tests showed that attentional modulation was significant for bodies ($t_{21} = 5.5$, $p_{bonf} = 1E-4$; red comparison in Fig. 2.4A), as already shown in the previous analyses (Fig. 2.3). No significant multivariate attention effect was observed for the other categories ($t_{21} < 2.4$, $p_{bonf} > 0.1$; for all tests; Fig. 2.4A).

2.3.7 Attentional modulation in EVC

The same analysis was conducted in early visual cortex (EVC; see Materials and Methods). A 6 (category of interest) $\times$2 (category attended/unattended) ANOVA on selective proximities revealed no significant interaction ($F_{5,105} = 2.2$, $p = 0.06$; Fig. 2.4B), no significant main effect of attention ($F_{1,21} = 0.6$, $p = 0.4$), and no significant main effect of category ($F_{5,105} = 2.2$, $p = 0.06$). Paired-sample t-tests showed no significant attentional modulation for any of the categories ($|t_{21}| < 2.2$, $p_{bonf} > 0.1$; for all tests). Finally, attentional modulation for bodies was

significantly stronger in LOC than in EVC ($t_{21} = 2.9$, $p = 0.01$).

2.3.8 The relationship between attentional modulation and behavioral responses

In both multivariate and univariate analyses, we found that the body-selective response elicited by body silhouettes in task-irrelevant locations was enhanced in body detection blocks compared with other category detection blocks. This raises the question of whether this attentional modulation affected behavior in the detection task. Particularly, did participants disproportionally false alarm to the bodies at task-irrelevant locations when detecting bodies at task-relevant locations? Because of the orthogonal design, each category (+blank stimulus) in the irrelevant location appeared equally often with each category (+blank stimulus) in the relevant location. Therefore, when the target category (e.g., bodies) appeared at the task-irrelevant location no target was presented at the task-relevant locations in most trials ($6/7$th), and participants had to withhold their response. For these trials, we tested whether responses (i.e., false alarms) depended on the combination of the category presented and the category that was the target in that block. To this end, for each category, we computed the difference between the false alarm rate (FA) to that category and the average FA to the other categories, separately for each block. We then compared this ΔFA for trials in which the object matched the target category (e.g., bodies presented in body blocks) and trials in which the object mismatched the target category (e.g., bodies presented in bed blocks).

A 2 (matching, non-matching) $\times 6$ (target category) ANOVA on ΔFA revealed a significant interaction ($F_{5,105} = 3.3$, $p = 0.008$; Fig. 2.5). Six paired-sample t-tests showed that ΔFA was stronger when the object matched the target category for all categories ($t_{21} > 2.9$, $p_{bonf} < 0.05$, for all non-body categories, biggest difference of 6.5% for cars; bodies: $t_{21} = 2.79$, $p_{bonf} = 0.066$, difference of 3.7%). These results show that participants disproportionally false alarmed when the target category was shown at the task-irrelevant location. Contrary to the fMRI results, however, this effect was relatively weak for bodies.

2.3.9 Image-based discriminability

In all fMRI analyses, we found that bodies were more strongly represented and more strongly modulated by attention than the other categories. This could reflect an interesting property of bodies, for example, related to the lifetime relevance of detecting conspecifics or to the increased familiarity with body shapes. However, it could potentially also reflect uncontrolled image-based differences: perhaps the body silhouettes included in the study stood out from the other objects in terms of low-level features. To exclude this possibility, we decoded object categories from the object exemplar representations in the layers of a convolutional neural network trained for object recognition (Materials and Methods). For each of the 6 categories, in each layer of the CNN, one-vs-all linear discriminant classifiers were trained to discriminate each category from the other categories using the 50 exemplars of each category presented in the fMRI experiment. 10-fold cross-validation accuracies were analyzed across the objects.

As shown in Fig. 2.5, bodies were less discriminable than most other categories in the early layers of the CNN. It is only in the mid to final layers - where overall classification is almost at the ceiling - that the classification accuracy for bodies is similar to the average accuracies for the other categories. This result shows that the image-based discriminability was, if anything, lower for bodies than for the other objects.

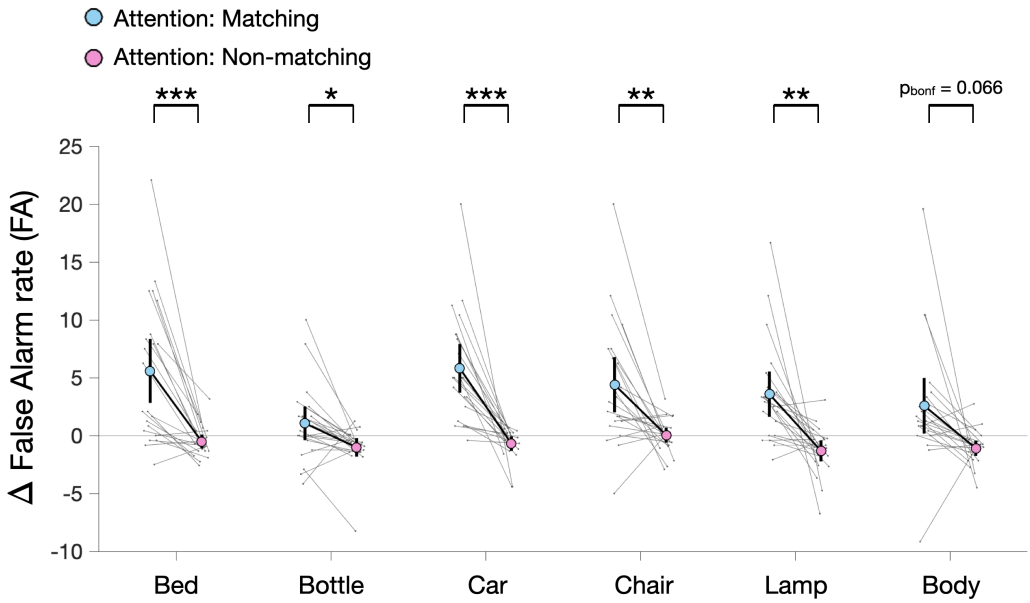

Figure 2.5: The relationship between attentional modulation and behavioral responses. Participants disproportionally false alarmed when the target category was shown at the task-irrelevant location (matching¿non-matching) but this effect was relatively weak for bodies. Error bars indicate 95% confidence intervals for ΔFA. The asterisks denote instances where t-tests returned $p_{bonf} < 0.05$ for the corresponding comparisons (*p< 0.05, **p< 0.01, ***p< 0.001).

2.4 Discussion

Across multiple analyses, we found convincing evidence that attention to human bodies enhanced visual cortex responses selective to bodies presented at task-irrelevant locations. This modulation reflected response gain rather than a generic bias, and could not be explained by low-level feature similarity of bodies. These results indicate that spatially-global attentional modulation – a hallmark of feature-based attention – can be found for features diagnostic of the presence of the human body.

The attentional effects observed here for body silhouettes are unlikely to reflect attention to low-level features such as orientation or color, for several reasons. First, we included a relatively large number of object categories in the experiment to ensure that participants could not detect objects based on low-level features, as these were shared with other categories (e.g., bottles were vertical, similar to bodies). Second, we presented object silhouettes instead of photographs to avoid possible low-level differences between categories in texture or color. Third, the image-based discriminability for each category, established using a convolutional neural network (CNN), indicated that bodies were difficult to discriminate from other categories based on low-level features encoded in the early layers of the CNN. Finally, the fMRI results showed attentional modulation in the object-selective cortex (LOC) and body-selective EBA/FBA, but not early visual cortex (EVC), indicating an attentional modulation at a higher level of visual processing.

Our results are in line with the feature similarity gain modulation model (FSGM; Maunsell and Treue (2006)) by showing that feature-based attention enhanced the response to the voxels' preferred stimuli. Specifically, attention to bodies made the response pattern evoked by task-

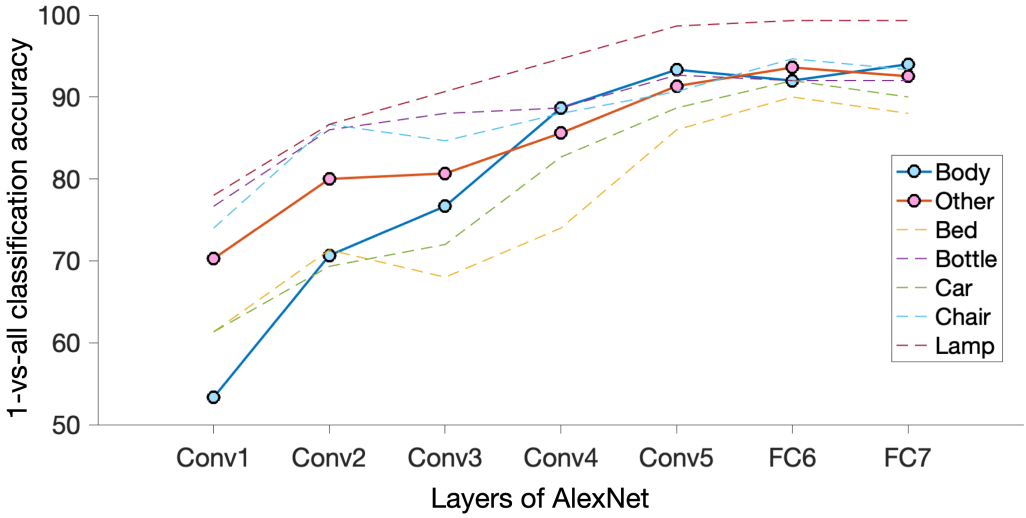

Figure 2.6: Hierarchical image-based discriminability of the exemplars used in the fMRI experiment. One-vs-all classifiers were trained for each of the six categories, on the output activations of each layer of a convolutional neural network trained for object recognition (AlexNet). 10-fold cross-validation accuracies are shown for all the objects in addition to the average accuracies for the non-body objects (termed 'Other'). Discriminability based on low-level features (corresponding to the early layers of the AlexNet) was, if anything, lower for the human bodies than for the other objects. Therefore, it is unlikely that the body-selective fMRI results reflect a distinct low-level property of bodies. 'Conv' refers to the convolutional layers of AlexNet and 'FC' refers to the fully-connected layers.

irrelevant bodies more similar to prototypical body response patterns established in a separate baseline experiment. Furthermore, these attention effects were strongest in body-selective voxels of LOC. Finally, reliable univariate attention effects were observed in independently-defined body-selective regions (EBA/FBA). It should be noted that we did not find evidence that responses to the other categories were suppressed, as proposed by FSGM. However, the response to other categories was low and any suppression (posited to be smaller in magnitude than enhancement by FSGM) might not be observable in this case.

The finding of spatially-global modulation for human bodies adds to previous evidence for global modulation for faces. Specifically, in one study, peripherally presented and task-irrelevant faces evoked a stronger face-selective N170 electroencephalography (EEG) response when participants attended to faces than to houses (Störmer et al., 2019). Furthermore, in fMRI, responses to peripheral faces in the face-selective fusiform face area (FFA) were more strongly modulated by the task-set of the participants (i.e., whether or not they focused on faces) than by spatial attention (Reddy et al., 2007). Together with the current findings, these results provide evidence for spatially-global attentional modulation for bodies and faces, two socially relevant categories that are selectively represented in the visual cortex (Downing et al., 2006; Kanwisher, 2010).

While these results suggest that bodies and faces may be special – reflecting their unique social and biological significance – we do not rule out that spatially-global attentional modulation may also exist for other highly-familiar object categories. For example, behavioral studies showed that animals and vehicles could be detected in the near-absence of spatial attention (Li et al. (2002);

but see Cohen et al. (2011)), with category-based attention facilitating object detection independently of spatial attention (Stein and Peelen, 2017). Indeed, based on the overlap in human and animal features in detection tasks (Evans and Treisman, 2005), it is plausible that our results would generalize to other animals, particularly those that activate body-selective regions (Downing et al., 2006). Similarly, extensive experience with particular objects may drive selective neural tuning (Gauthier and Logothetis, 2000; McGugin et al., 2012; Frank et al., 2014a) and give rise to similar behavioral advantages as those observed for bodies (Hershler and Hochstein, 2009; Golan et al., 2014; Reeder et al., 2016; Stein et al., 2016).

Taking everything together, the evidence suggests that features that are diagnostic of bodies meet many of the previously proposed criteria for basic features: showing spatially-global attentional modulation (Maunsell and Treue, 2006), being processed "early, automatically, and in parallel across the visual field" (Treisman and Gelade, 1980), and being represented selectively in the visual system (Treisman, 2006). Indeed, Treisman (2006) proposed that the feature detectors of the feature integration theory are not necessarily limited to low-level features such as orientation and color. Raising the possibility that there may be animal feature detectors, Treisman noted that animal features may not necessarily be more complex for the visual system than colors, line orientations, or direction of motion. By providing evidence for spatially-global attentional modulation for human bodies, our results support this proposal.

Our findings raise the question of what features are attended to when attention is directed to bodies. Addressing this question for animals, Treisman suggested that: "participants may be set to sense, in parallel, a highly overlearned vocabulary of features that characterize a particular semantic category." One possibility is thus that attention to bodies is mediated by attention to a set of mid-level features that are diagnostic of human bodies (Ullman et al., 2002; Reeder and Peelen, 2013). Alternatively, attention may be directed to holistic representations of body shape (Reed et al., 2003; Stein et al., 2012). Future studies may test these alternatives by measuring global attentional modulation for various body-related features, body parts, and inverted bodies at the task-irrelevant location while participants attend to bodies at the task-relevant locations (Reeder and Peelen, 2013).

To conclude, the current results provide the first evidence for spatially-global attentional modulation for human bodies in the high-level visual cortex, linking this modulation to body-selective representations in univariate and multivariate analyses. Combining these results with previous behavioral and neuroimaging studies, we propose that bodies may be processed as basic features, supporting the rapid and parallel detection of conspecifics in our environment even outside the focus of spatial attention.

Chapter 3

The functional role of cue-driven feature-based feedback in object recognition

Visual object recognition is not a trivial task, especially when the objects are degraded or surrounded by clutter or presented briefly. External cues (such as verbal cues or visual context) can boost recognition performance in such conditions. In this work, we build an artificial neural network to model the interaction between the object processing stream (OPS) and the cue. We study the effects of varying neural and representational capacities of the OPS on the performance boost provided by cue-driven feature-based feedback in the OPS. We observe that the feedback provides performance boosts only if the category-specific features about the objects cannot be fully represented in the OPS. This representational limit is more dependent on task demands than neural capacity. We also observe that the feedback scheme trained to maximize recognition performance boost is not the same as tuning-based feedback, and performs better than tuning-based feedback[1].

3.1 Introduction

Visual object recognition is a non-trivial task, especially when the objects are degraded, surrounded by clutter, or presented briefly. The introduction of external cues (such as verbal cues or visual context) can constrain the space of the possible object features and/or categories and improve recognition performance (Carrasco, 2011; Bar, 2004).

External cues can interact with the object processing stream in two ways. They can either modulate the information transformations in the object processing stream (through feedback) and/or get combined with the object evidence present at the end of the stream, to improve the overall decision about the category of the object. Intuitively, the interaction involving feedback would help with object recognition especially when the feature information required to recognize the object cannot be extracted by the object processing stream. This can happen either due to a capacity limit or due to a lack of information present in the input.

[1]This chapter has been adapted from - Thorat, S., van Gerven, M. A. J., & Peelen, M. V. (2018). *The functional role of cue-driven feature-based feedback in object recognition.* Proceedings of the 2018 Conference on Cognitive Computational Neuroscience, p. 1–4.

Such a capacity limit can arise due to two reasons. One, due to a limit on the number of neurons available in the object processing stream, which would reduce the object information that can be extracted from the image. We term this the *neural* capacity limit. Two, due to the limits imposed by the task for which the stream is trained. For example, if the stream is trained to represent one object, it will not perform well if two objects are presented in the same image unless feature selection is employed in the early stages of the network. We term this the *representational* capacity limit. In this work, we aim to understand how the feedback-driven performance gain due to the cue depends on the capacity limits of the visual processing stream.

Cue-driven feedback can affect the object processing stream in a feature-specific and/or location specific manner. These interactions also account for feature-based and spatial attention (Carrasco, 2011). In this work, we focus on the feature-based feedback interaction.

We developed an artificial neural network (ANN) to model the object processing stream, probing the stream output for an object's presence in the image, and the feedback-based interaction between an external cue and the stream. The parameters of the ANN let us manipulate the neural and representational capacities of the object processing stream. We train the ANN to maximize the difference (termed as the recognition performance) between correct (true positives) and incorrect (false positives) identification of the objects in the image. The external cue and the probe contain category-level information. For example, the external cue could correspond to 'Look for a Shoe' and the probe could correspond to 'Was there a Shoe?'. Then the category-level information would be 'Shoe'. We show that the external cue substantially boosts recognition performance when the object processing stream cannot represent (low representational capacity) the information required for categorizing the target object (given by the probe). We then comment on the nature of the feedback that maximizes these performance boosts.

3.2 Methods

3.2.1 Stimuli

We use the dataset Fashion-MNIST (Xiao et al., 2017), which contains 28×28 images of $70,000$ fashion products from 10 categories[2].

We want to assess the effects of cue-driven feature-based feedback on two object-feature manipulations. One, reducing the feature information through blurring. Two, introducing feature competition by adding more objects to the image. To do so, we construct 2×2 grid ($40\,\text{px} \times 40\,\text{px}$) images, in which we can place 1 to 4 objects (category overlaps are allowed) and blur them with a Gaussian kernel with standard deviations varying uniformly from 0 to 4 pixels. Example images are shown in Fig. 3.1.

3.2.2 Network architecture

The artificial neural network (ANN) accepts three inputs - the image, the cue, and the probe, and outputs whether the probe category is present in the image, as shown in Fig. 3.2. We will now describe the sub-networks corresponding to the individual inputs, and their interactions.

[2]We split the dataset into 55k, 5k, and 10k images as train, validation, and test sets (equally split over the 10 categories). For testing, we generate 10k grid images with the test image set.

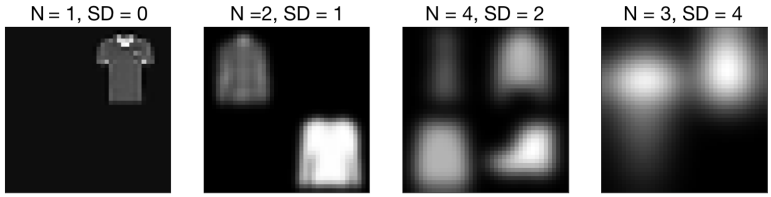

Figure 3.1: Examples of the stimuli used (with the number of objects, N, and the Gaussian blur standard deviations, SD).

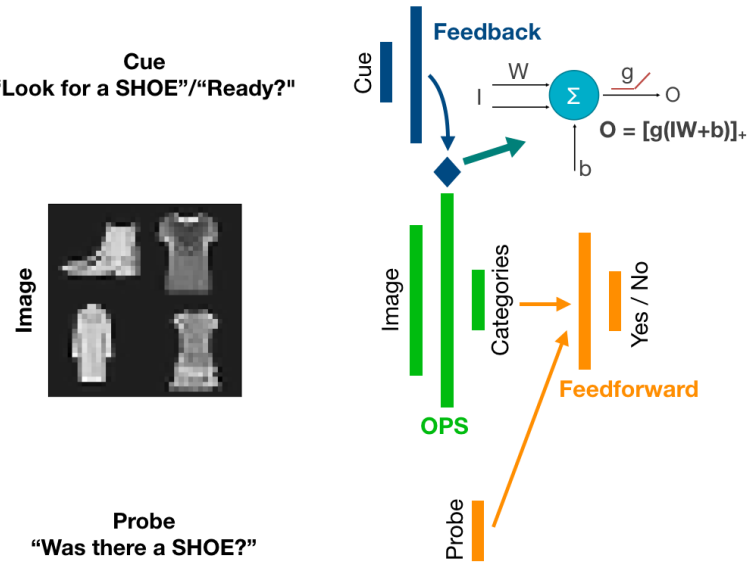

Figure 3.2: The artificial neural network (ANN) designed to gauge the dependence of feedback-driven performance boosts due to the cue on the capacity limits of the object processing stream (OPS). The ANN outputs whether the probe category is present in the input image or not. The cue interacts with the OPS through bias (b) and/or gain (g) modulation of the hidden units.

3.2.3 Nature of the object processing stream

The object processing stream (OPS, in green in Fig. 3.2) is a fully-connected ANN with one hidden layer. The input layer consists of 1600 units, representing the 40 px x 40 px images. The output layer consists of 11 units, 10 of which give the probabilities of the categories present in the image. The 11 th unit is the out-of-sample detector which gives the probability that the input image does not belong to the space of images from the training set. This is done to prevent the OPS from making high-confidence errors on out-of-sample images.

The hidden layer contains either 8, 32, or 3072 rectified linear units (ReLU). An increase in the number of hidden layer units corresponds to an increase in neural capacity.

3.2.4 Nature of the probe

The probe is a one-hot encoding[3] (10 units) of the category of interest. It is fed into another ANN with the output of the OPS (the category probabilities). This feedforward query ANN (fully connected, in orange in Fig. 3.2) has 200 rectified linear units (ReLU) in its hidden layer. It has 2 outputs (Yes/No) which give the probability of the presence of the probe category in the image.

3.2.5 Nature of the cue

The cue consists of 11 units, 10 of which correspond to the 'informative' cue as they correspond to the object categories. The 11 th unit corresponds to the 'uninformative' cue ("Ready?" as seen in Fig. 3.2). The informative cue is the same as the probe. This cue, after being transformed into 'feedback templates', interacts with the OPS through bias and gain modulation, as explained next. This cue network is shown in blue in Fig. 3.2.

3.2.6 Cue-OPS interaction

The responses $\mathbf{O}$ of the units in the hidden layer of the OPS are given by $\mathbf{O_h} = [\mathbf{g_h}(\mathbf{I}\mathbf{W_h} + \mathbf{b_h})]_+$, where $\mathbf{I}$ are the inputs, $\mathbf{W_h}$ are the input weights, and $\mathbf{b_h}$ and $\mathbf{g_h}$ are the biases and gains of the units. $[x]_+ = x$ if $x > 0$, & $[x]_+ = 0$ if $x \leq 0$. The feedback templates $\mathbf{b_c}$ and $\mathbf{g_c}$ are linear transformations of the cue, given by $\mathbf{b_c} = \mathbf{z_c}\mathbf{W_b}$ and $\mathbf{g_c} = \mathbf{z_c}\mathbf{W_g}$, where $\mathbf{z_c}$ is the one-hot encoding of the cue category. These templates are added to $\mathbf{b_h}$ and $\mathbf{g_h}$ respectively, causing either an additive or multiplicative boost in the units' responses. This interaction between the cue and the OPS was adapted from Lindsay and Miller (2017).

3.2.7 Network training

We now describe the input-output maps used in training the ANN shown in Fig. 3.2. In each case, we learn the maps by minimizing the cross-entropy between the network output and target probability distributions. We do so by using stochastic gradient descent (SGD) with Dropout regularisation Srivastava et al. (2014). The training is done in three steps.

Training the OPS

First, we train the parameters of the network marked in green in Fig. 3.2. The inputs are the images mentioned in the Stimuli section. For each image, the target distribution, at the end of the OPS, is an n-hot encoding normalized to a unit vector, representing the n unique object categories in the image. The 11 th unit of the OPS output is associated with random images[4]. To manipulate the representational capacity of the OPS, we train the OPS either with images containing a single object which is not blurred or with the full range of feature manipulations as shown in the Stimuli section. The representational capacity for the full range of feature manipulations is lower in the former case, which we refer to as low representational capacity here.

Training the probe

Second, we train the parameters of the network marked in orange in Fig. 3.2. The OPS parameters are frozen. The inputs are the images with the full range of feature manipulations. The images

[3] Given 5 categories, and $1, 2, 5$ being the categories of interest, an n-hot encoding (n= 3 here) would be $[1, 1, 0, 0, 1]$

[4] Uniformly random intensities are generated for all pixels. Then the image is scaled, blurred, and occluded to cover subspaces of interest better.

are paired with correct or incorrect probe categories equally. The output of the ANN is a one-hot encoding of the correctness of the probe.

Training the cue

Third, we train the parameters of the network marked in blue in Fig. 3.2. All other ANN parameters are frozen. In the case of informative cues, where the cued category is the same as the probe category, the maps used in training the probe are paired with the respective cues. In the case of the uninformative cue, all the maps used in training the probe are paired with the cue. We train bias and gain modulation together, allowing for interactions between them.

3.2.8 Evaluation metric

Recognition performance is defined as the difference between the proportion of correct and incorrect assessment that the probe category exists in the input image. We assess the effects of imposing the two capacity limits on the recognition performance (True positives (TP) - False positives (FP)) boosts provided by the cues. If category information in the informative cue adds any functional (in terms of object categorization) value, it should boost performance beyond the performance given by the uninformative cue (this boost is denoted by Δ).

3.3 Results and discussion

We evaluate the recognition performance on the full range of feature manipulations (number of objects and the strength of blurring), and in the case of joint training of the gain and bias modulation (in the cue ANN). The accuracies for recognizing single objects in the image when the object processing stream (OPS) is trained on single objects in the image (number of hidden units mentioned in brackets) are $82.4\%\,(3072)$, $75.7\%\,(32)$, and $54.8\%\,(8)$. So, the representational capacity for single objects reduces with a reduction in neural capacity.

The recognition performance of the probe-only, the uninformative cue, and the informative cue cases are mentioned in Fig. 3.3. As seen in the figure, the informative cue provides higher recognition performance than the uninformative cue and the probe-only case when the representational capacity is reduced. The uninformative cue boosts performance over the probe-only case when the representational capacity is low. This performance boost could be a result of boosting the overall activity of the hidden units (through bias/gain) that provide reliable differences in the activity for the object categories, in the case of the images with feature manipulations.

Trends observed in Fig. 3.3 are preserved if we vary only the number of objects (given 3072 OPS hidden units, $\Delta_{avg,RC\uparrow} = 1.5\%$, $\Delta_{avg,RC\downarrow} = 11.7\%$; if $n_{obj} = 4$, $\Delta_{RC\uparrow} = 2.3\%$, $\Delta_{RC\downarrow} = 15.5\%$) or the strength of blurring in the test images (given 3072 OPS hidden units, $\Delta_{avg,RC\uparrow} = 1.2\%$, $\Delta_{avg,RC\downarrow} = 5.4\%$; if blur SD $= 4$ px, $\Delta_{RC\uparrow} = 2.3\%$, $\Delta_{RC\downarrow} = 21.4\%$).

So, cue-driven feature-based feedback seems to be useful for recognizing objects subject to feature blurring and/or competition, only when the features required to classify those objects cannot be fully represented in the object processing stream. Intuitively, if the OPS can represent all the category-specific information about an object given the implicit representational limits imposed by the neural capacity, such feedback should not be able to add to the performance.

3.3.1 The influence of the trained feedback

How does the external cue influence the representation of the relevant object in the hidden layer?

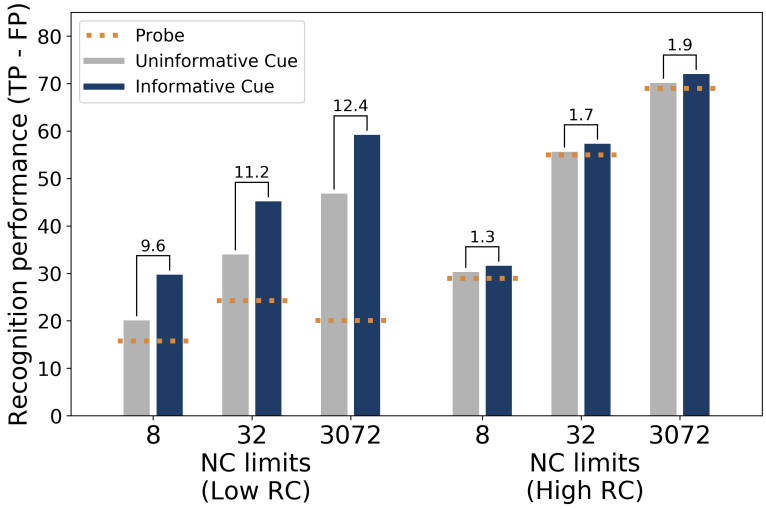

Figure 3.3: Cue-driven recognition performance boosts as a function of the neural capacity (NC) and representational capacity (RC). The values of the boosts (Δ_{avg}) given by the informative cue beyond the uninformative cue are mentioned for NC/RC pair. As the neural capacity is reduced, the informative cue provides lower performance boosts. Given a neural capacity, the informative cue provides higher performance boosts when the representational capacity is reduced.

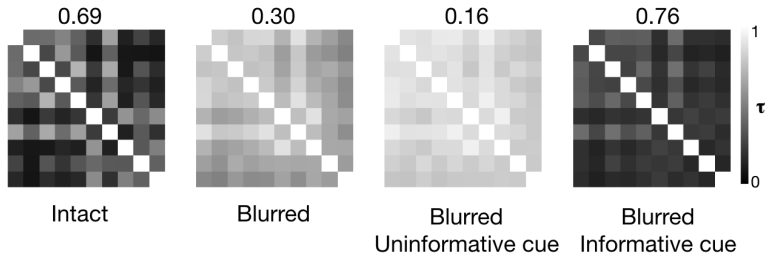

Figure 3.4: The effect of cueing on the representation of the relevant object in the OPS hidden layer (3072 units and low RC). The standard deviation of blurring is 4 px. The category-level RSMs for the activities in the hidden layer for the mentioned cases are shown. The categoricality indices for each RSM are shown. The informative cue makes the representations in the hidden layer selectively more distinct for the relevant object.

To assess this influence, we generate category-level representational similarity matrices (RSM, based on Kendall's τ correlation) for the case of a single object in the grid, which is either presented intact, with blurring (with SD $= 4$ px), with blurring and the uninformative cue, or with blurring and the informative cue. We define the *categoricality index* of the RSM as the difference between the mean values of the diagonal and off-diagonal elements. As seen in Fig. 3.4, the informative cue makes the representation of the relevant object more distinct, making it accessible to the output of the OPS.

In Abdelhack and Kamitani (2018), it was shown that neural representations of blurred objects in the human visual cortex are more similar to the corresponding intact object representations in a

feedforward neural network than the blurred object representations in that neural network. They attributed this effect to top-down information interacting with the stimulus representations. This effect became stronger when a category cue was introduced. However, in our case, the neural representations of blurred objects (with either the informative or the uninformative cue in effect) are equally similar to the corresponding intact object representations (with no cue) and the blurred object representations (with no cue) ($\Delta\tau < 0.04$). This inconsistency will be probed in further work by using more complex and more biologically-plausible networks (such as convolutional neural networks and recurrent neural networks) for the object processing stream, which would make our network a better model of the human visual system.

3.3.2 Comparison with tuning-based feedback

A popular model to describe the effects of the cue on neuronal responses in the brain is the feature similarity gain model (FSGM) Martinez-Trujillo and Treue (2004). It claims that the neuronal response is multiplicatively scaled according to its preference to the properties of the attended (or task-relevant) stimuli. Such cue-driven 'tuning-based' feedback was shown to boost object recognition performance (with multiple objects in a grid or overlaid) in Lindsay and Miller (2017). We deployed tuning-based bias and gain modulation according to the mathematical framework outlined in Lindsay and Miller (2017)[5]. We ran a grid search to compute the parameters to maximize the recognition performance boosts provided by tuning-based feedback over the probe-only case when representational capacity is low. Across the three neural capacities, the maximum performance boost observed was 3%. This is small compared to the boosts observed with the feedback trained with SGD as seen in Fig. 3.3. This implies the trained feedback is not the same as tuning-based feedback.

Lindsay and Miller (2017) did observe a higher performance using gradient-based feedback (of which feedback trained with SGD is the natural extension) than with tuning-based feedback. As also mentioned in their paper, this is not surprising as neuronal tuning is not necessarily a measure of neuronal function. It has been shown that category-selective responses of hidden units in ANNs do not imply that those units are relatively more important to the recognition of objects of those categories Morcos et al. (2018). The greater the category-selectivity of the hidden units, the harder it is for the ANN to generalize to new data.

3.4 Conclusions

In this work, we investigated the nature and usefulness of cue-driven feature-based feedback in recognizing objects suffering from feature blurring and/or competition. We built an artificial neural network and asked how feature-based feedback can be deployed, and how its recognition performance boosts are dependent on the neural and representational capacities of the object processing stream. We found that the feedback boosts performance only if the category-specific features about the objects cannot be fully represented in the base ANN. These representational limits are not dependent on the neural capacity but on the task demands on the object processing stream. The trained feedback does not resemble (but performs better than) tuning-based feedback which is based on the feature similarity gain model Martinez-Trujillo and Treue (2004).

[5]To implement the tuning-based modulations, the following steps are taken in Lindsay and Miller (2017). Compute category-specific (averaged across multiple images) hidden layer activations. Mean- and variance-normalize the category values for each hidden unit to generate the feedback templates. Tune (multiplicative scaling only) these templates for bias (additive) or gain (multiplicative) modulation.

To gauge the robustness of our observations, in subsequent work we will run these analyses on different datasets and architectures (such as convolutional neural networks). We shall also assess these effects for location-based feedback.

Chapter 4

Modulation of early visual processing alleviates capacity limits in solving multiple tasks

In daily life situations, we have to perform multiple tasks given a visual stimulus, which requires task-relevant information to be transmitted through our visual system. When it is not possible to transmit all the possibly relevant information to higher layers, due to a bottleneck, task-based modulation of early visual processing might be necessary. In this work, we report how the effectiveness of modulating the early processing stage of an artificial neural network depends on the information bottleneck faced by the network. The bottleneck is quantified by the number of tasks the network has to perform and the neural capacity of the later stage of the network. The effectiveness is gauged by the performance on multiple object detection tasks, where the network is trained with a recent multi-task optimization scheme. By associating neural modulations with task-based *switching* of the state of the network and characterizing when such switching is helpful in early processing, our results provide a functional perspective towards understanding why task-based modulation of early neural processes might be observed in the primate visual cortex[1].

4.1 Introduction

Humans and other animals have to perform multiple tasks given a visual stimulus. For example, seeing a face, we may have to say whether it is happy or sad, or recognize its identity. For each of these tasks, a subset of all the features of the face is useful. In principle, it could be possible for a visual system to extract all of the features necessary to solve all possible tasks, and then select the relevant information from this rich representation downstream. However, as the number of tasks increases, a network with a limited capacity may not be able to extract all of the potentially relevant features (an information bottleneck is manifest), requiring the information that is extracted from the stimulus in the early processing stages to change according to the task.

[1]This chapter has been adapted from - Thorat, S., Aldegheri, G. van Gerven, M. A. J., & Peelen, M. V. (2019). *Modulation of early visual processing alleviates capacity limits in solving multiple tasks.* 2019 Conference on Cognitive Computational Neuroscience, p. 226–229.

Several studies in neuroscience have found evidence for such task-dependent modulations of sensory processing in the primate visual system, including at the early levels (Carrasco, 2011; Maunsell and Treue, 2006; Gilbert and Li, 2013). For example, human neuroimaging studies have shown that attending to a stimulus could lead to an increase in the accuracy with which its task-relevant features could be decoded by a classifier in early visual areas (Jehee et al., 2011), and neurophysiological experiments in nonhuman primates have shown that the stimulus selectivity of neurons in primary visual cortex was dependent on the task the monkeys had to perform (Gilbert and Li, 2013).

Despite the observation of such modulations of early visual processing, it is not clear whether they are causally necessary for performing better on the corresponding tasks. This question has been addressed by deploying biologically-inspired task-based modulations on computational models. Lindsay and Miller (2018) showed that task-based modulation deployed on multiple stages of a convolutional neural network improves performance on challenging object classification tasks. Another recent work (Thorat et al., 2018; Rosenfeld et al., 2018) has also shown that task-based modulation of early visual processing aids in object detection and segmentation in addition to the task-based modulation of late processing. However, the conditions under which early modulation can be beneficial in performing multiple tasks have not been systematically investigated.

In the present work, we assessed the effectiveness of task-based modulation of early visual processing as a function of an information bottleneck in a neural network, quantified by the number of tasks the network had to execute and the neural capacity of the network. To do so, we trained networks to, given an image, provide an answer conditioned on the cued task. Every task required detecting the presence of the corresponding object in the image. The networks were trained according to a recent framework proposed in the field of continual learning (Cheung et al., 2019), which helps them execute multiple tasks by switching their state given a task cue, to transmit relevant information through the network. In this work, to quantify the effectiveness of task-based modulation of early neural processing, we measured the increase in performance provided by modulating early neural processing in addition to modulating the late neural processing in the networks.

4.2 Methods

4.2.1 Task and system description

In a multi-task setting, object detection can be thought of as solving one of a set of possible binary classification (one object versus the rest) problems. Given an image and a task cue indicating the identity of the object to be detected, a network had to output if the object in the image matched the task cue.

We used MNIST (LeCun et al., 1998) digits and their permutations as objects (Kirkpatrick et al., 2017). The original MNIST dataset has $28 \times 28\,\mathrm{px}^2$ images of 10 digits. Each permuted version consists of images of those 10 digits, whose pixels undergo a given permutation, creating 10 new objects. We varied the number of permutations used (10, 25, and 50) to modulate the number of tasks the networks had to perform (which are 10 times the number of permutations).

We considered a multi-layer perceptron with rectified linear units (ReLU), which had one hidden layer between the input (image) and the binary output. The number of neurons in the hidden layer was variable (32, 64, and 128) and determined the neural capacity of the late stage of the network.

4.2.2 Task-based modulation and its function

Modelling biological neurons as perceptrons (Rosenblatt, 1957), task-based modulations have been shown to affect the effective biases and gains of the neurons (Maunsell and Treue, 2006; Boynton, 2009; Ling et al., 2009). The nature of modulation - which neurons should be modulated and how - is under debate (Boynton, 2009; Thorat et al., 2018). We adapted these findings by introducing task-based modulation into our networks via the biases of the perceptrons and the gains of their ReLU activation functions. The modulations were then trained end-to-end with the rest of the network.

Given a particular task, the task cue is a one-hot encoding of the relevant object. Task-based modulation is mediated through bias and gain modulation in the following manner.

$$x_n = [W_n(g_{n-1} \circ x_{n-1}) + b_n]_+ \tag{4.1}$$

$$g_n = G_n c, \;\; b_n = B_n c \tag{4.2}$$

where the transformation between layers $n-1$ and n ($L_{n-1} \rightarrow L_n$) is modulated by changing the slope of the ReLU activation function (gains, g_{n-1}) in L_{n-1} and the biases (b_n) to the perceptrons in L_n; x_n are the pre-gain activations of the perceptrons in L_n, W_n is the task-independent transformation matrix between L_{n-1} and L_n, G_n and B_n map the task cue c (one-hot encoding of the relevant object k) to the gain and bias modulations of the perceptrons in L_n respectively, and $\circ$ refers to element-wise multiplication.

Given a task k, modulating the gains of the pre-synaptic perceptrons (in L_{n-1}) and the biases of the post-synaptic perceptrons (in L_n) transforms the information transformation between L_{n-1} and L_n. This allows for the transmission of information required to perform task k while ignoring the information required to perform the other tasks, as formalized in Cheung et al. (2019). This transformation can also be thought of as the network *switching* its state to selectively transmit task-relevant information downstream (see Fig. 4.1). The conditions - the nature of these modulations and the neural capacity of the network - under which the network can switch between a given number of tasks are preliminarily described in Cheung et al. (2019).

For every layer L_n, W_n, B_n, and G_n were jointly learned for the given number of tasks.

4.2.3 Evaluation metric and expected trends

The effectiveness of early neural modulation was quantified by the average absolute increase in detection performance across all the tasks when modulations were implemented on both the transformations $L_1 \rightarrow L_2$ and $L_2 \rightarrow L_3$ (L_1 corresponds to the input layer and L_3 to the output layer) as opposed to when the modulations were trained on the transformation $L_2 \rightarrow L_3$ only.

We expected the effectiveness of task-based early neural modulation to be directly proportional to the number of neurons in L_2 and inversely proportional to the number of tasks (permuted MNIST sets used).

4.2.4 Neural network training details

All the networks were trained with adaptive stochastic gradient descent with backpropagation through the ADAM optimiser (Kingma and Ba, 2014) with the default settings in TensorFlow (v1.4.0) and $\alpha = 10^{-5}$. We used a batch size of 100. Half of each batch contained randomly selected images of randomly selected tasks where the cued object was present and a half where the cued object was not present. These images were taken from the MNIST training set and its corresponding permutations. The images were augmented by adding small translations and noise.

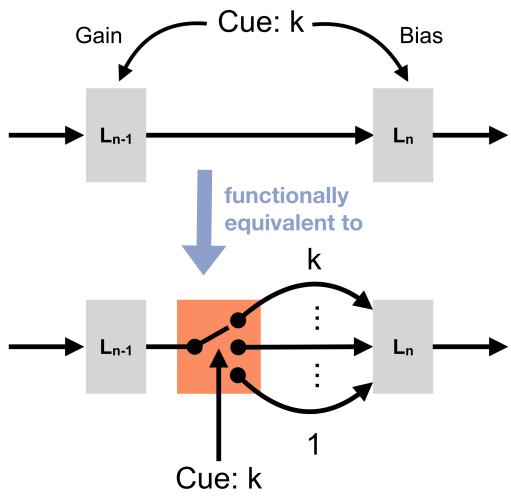

Figure 4.1: The effect of bias and gain modulation on the transformations in the network. Modulating the gains and biases is *functionally* equivalent to *switching* the transformation being performed to one suited for the relevant task. Such an example of switching is visualized in the figure. Given a task cue corresponding to object k, corresponding gain and bias modulations are applied, which results in the $L_{n-1} \rightarrow L_n$ transformation being switched into one that transmits feature information required to detect the presence or absence of object k.

We trained each network with 10^7 such batches. The relevant metrics discussed in the previous section are computed at the end of training over a batch of size 10^5 created from the MNIST test set and its corresponding permutations.

4.3 Results

We first analyzed the detection performance of the network with only $L_2 \rightarrow L_3$ modulation. The network performance as a function of the number of neurons in L_2 and the number of detection tasks the network had to perform is shown in Fig. 4.2 (red circles). The network performance increased with an increase in the number of neurons in L_2, as the neural capacity increased. The performance decreased with an increase in the number of tasks to be performed, as the representational capacity of the network for any of the tasks was reduced. A network with as little as 32 neurons in its hidden layer was able to switch between as many as 500 detection tasks, while keeping the average detection performance across all the tasks as high as 87%, thus replicating the success of the multi-task learning framework proposed by Cheung et al. (2019).

To assess the dependence of the effectiveness of task-based modulation of early neural processing ($L_1 \rightarrow L_2$) on the bottleneck in the network, we analyzed the boost in average detection performance when task-based modulation of $L_1 \rightarrow L_2$ was deployed in addition to task-based modulation of $L_2 \rightarrow L_3$, as a function of the number of neurons in L_2 and the number of detection tasks the network had to perform. The resulting boosts are shown in Fig. 4.2 ($\Delta_\uparrow$ quantification). The performance boost increased as the number of neurons in L_2 decreased, and as the number of tasks the network has to perform increased. This confirms the hypothesis that task-based modulation of early neural processing is essential when an information bottleneck exists in a subsequent

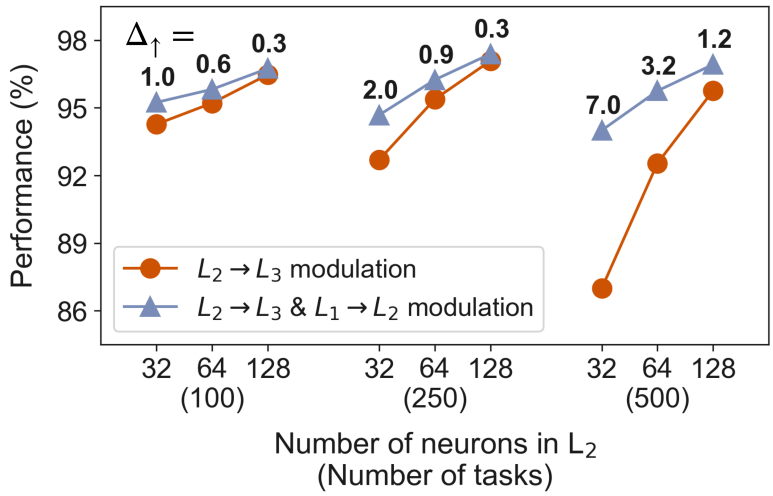

Figure 4.2: The effectiveness of task-based modulation (quantified by the performance boost, $\Delta_{\uparrow}$) of early neural processing ($L_1 \rightarrow L_2$) as a function of the number of neurons in L_2 and the number of tasks the network has to perform. The performance boost was inversely proportional to the number of neurons in L_2 and directly proportional to the number of tasks the network had to perform. The absolute performance profiles given either the modulation of $L_2 \rightarrow L_3$ only or the joint modulation of $L_2 \rightarrow L_3$ and $L_1 \rightarrow L_2$ are also shown. (See the Supplementary results for substantiation.)

processing stage (see the Supplementary results for further analyses elucidating these results).

4.3.1 The contributions of bias and gain modulation

Gain, but not so much bias, modulation of neural responses has been observed in experiments investigating feature-based attention in the monkey/human brain (Maunsell and Treue, 2006; Boynton, 2009). We assessed how the two contributed to the overall modulation of the transformations in the network.

We selectively turned off the bias or gain modulation for all the variants of the network that were trained. The average detection performance decreased by $43.0 \pm 2.0\%$ when gain modulation was turned off and by $3.9 \pm 0.9\%$ when bias modulation was turned off, suggesting that in our framework, when jointly deployed, gain modulation is more important than bias modulation in switching the state of the network to be able to perform the desired task well.

We also trained a network with 32 neurons in L_2, on 25 permutations of MNIST, with gain-only or bias-only modulations of both the $L_1 \rightarrow L_2$ and $L_2 \rightarrow L_3$ transformations. When the gain and bias modulations were jointly trained, the network performance was 94.7%. With gain-only modulation, the performance was 94.8%, and with bias-only modulation, the performance was 90.9%. As the performance when only bias modulation was deployed was much higher than chance (50%), we can conclude that bias modulation alone can also lead to efficient task-switching. When the bias and gain modulations are jointly trained, the gain might take over as it multiplicatively impacts responses, and therefore has higher gradients during training, as opposed to the additive impact of bias.

4.4 Discussion

Adding to the discussion about the functional role of task-based modulation of early neural processing, in this work, we have shown that modulating the early layer of an artificial neural network in a task-dependent manner can boost performance, beyond just modulating the late layer, in a multi-task learning scenario in which a network contains an information bottleneck, either due to a large number of tasks to be performed or to a small number of units in the late layer.

Adapting a formalism proposed by Cheung et al. (2019), we showed how bias and gain modulation, two prevalent neuronal implementations of top-down modulation in the brain, could functionally lead to switching the state of a network to perform transformations effective for the task at hand. While task-dependent computations are widespread in higher-level areas of the primate brain, such as prefrontal cortex (Mante et al., 2013), it is not clear to what extent sensory streams (which perform early visual processing) can also be seen as switching their state according to the current task (although see Gilbert and Li (2013) for a proposal), and what the functional relevance of doing so would be. Here we show how this switching could be computationally advantageous when it is not possible to send the information required for all tasks to higher layers, which might well be the case in the complex environments that humans and other animals can navigate.

To further investigate the relevance of our findings to biological visual systems, in follow-up work our modulation scheme could be deployed on architectures that bear more similarity to the primate visual hierarchy, such as deep convolutional networks (Kriegeskorte, 2015), datasets of naturalistic images such as ImageNet (Russakovsky et al., 2015), and general naturalistic tasks such as visual question answering (Agrawal et al., 2017). This will allow us to assess whether the functional advantage provided by early modulation holds in a more realistic scenario and whether the resulting modulation schemes resemble those observed in the early visual areas of the brain.

Finally, a key aspect of our approach is the fact that the network is constantly operating in a task-dependent manner. Most previous approaches to task-dependent modulation have assumed the presence of an underlying task-free representation on which the modulation operates (for example, in the case of Lindsay and Miller (2018) this corresponds to a network pre-trained on object recognition). Providing the network with task cues during the training phase, on the other hand, has been used in the field of continual learning (Cheung et al., 2019; Masse et al., 2018; Yang et al., 2019), and according to one influential theory in neuroscience, the interplay between sparse, context-specific information encoded by the hippocampus and shared structural information in the neo-cortex is crucial for learning new tasks without overwriting previous ones (Kumaran et al., 2016). To our knowledge, the question of how the task-based modulations observed in the visual cortex might be learned has not been explicitly addressed in previous literature. On the one hand, a context-free representation may be learned first, possibly through unsupervised learning, and then modulated upon. On the other, learning of representations and task modulations might interact at all stages, allowing the representations to be optimized for the type of modulations they are subject to. Whether one scheme or the other constitutes a better explanation for the modulations observed in biological visual systems is an important direction for research.

4.5 Supplementary results

Here we present the results from further analyses performed to address reviewer comments, post-acceptance into the 2019 Conference on Cognitive Computational Neuroscience.

4.5.1 Dependence of the results on parameter expansion

When early modulation is performed in addition to late modulation, the addition of parameters is constant across the different networks with a varying number of hidden neurons. However, the relative increase in the number of parameters is higher in the network with a lower number of neurons in the hidden layer. This network, with a lower number of hidden neurons, also benefits most from the addition of early modulation, as observed in Fig. 4.2. So, are the observations in Fig. 4.2 an effect of simple proportionate parameter expansion? Two additional analyses show that this is not the case.

Matching the proportionate increase in the number of parameters across networks

To match the proportionate increase in parameters, we introduced an additional layer between the task cue c and the gain modulation g_1 to the first (input) layer of the network. No bias modulation was included in these networks as we found it did not aid gain modulation. We set the number of hidden neurons in this modulation network to 300. We wanted to match the proportionate increase in parameters accompanying the addition of early gain modulation to late gain modulation, between two networks with 32 and 128 neurons in L_2 respectively. To accomplish this matching, in the case of the network with 32 neurons the hidden layer of the modulation needs to contain approximately 75 neurons. As seen in Fig. 4.3, reducing the number of hidden neurons in early modulation to 75, to match the proportionate increase in parameters between the two networks with 32 and 128 neurons in L_2, does not substantially reduce the increase in performance provided by task-based modulation of early neural processing.

Increasing the task difficulty to increase capacity constraints

In addition to manipulating the number of hidden neurons in the network and the number of tasks to be performed, the difficulty of the tasks could also contribute to capacity limits. If we increase the noise in the stimuli from 20% (which is the case in the main analysis; additive uniform random noise) to 40%, the performance boost increases as seen in Fig. 4.4. As the number of parameters stays the same across the change in the level of noise, this result also suggests that the trends observed in Fig. 4.2 are not simply a consequence of simple proportionate parameter expansion.

4.5.2 Comparison with Cheung et al. 2019

The task-based modulation scheme in our work was adapted from Cheung et al. (2019). However, there are major differences between our work and Cheung et al. (2019). Their work proposes a solution to the problem of catastrophic forgetting in neural networks when faced with a continual stream of tasks. In our work, the tasks are interleaved. They used random task-based modulations, while in our work the task-based modulations are learned to optimize detection performance. If two tasks are similar, the task-based modulations should also be similar and this relationship can be learned by the networks in our work. We wanted to assess if learning the modulations instead of fixing them randomly, as in Cheung et al. (2019), aids performance on the detection tasks here. To do so, we trained a network with 32 neurons in L_2 on 500 tasks, with or without training the bias and gain modulations (early and late modulation) which are initialized with random binary vectors as in Cheung et al. (2019). When the modulation schemes can be learned, the performance of the network was 94.0%, while when the scheme was fixed as random binary vectors, the performance was 71.1%, confirming the idea that task-based modulations can account

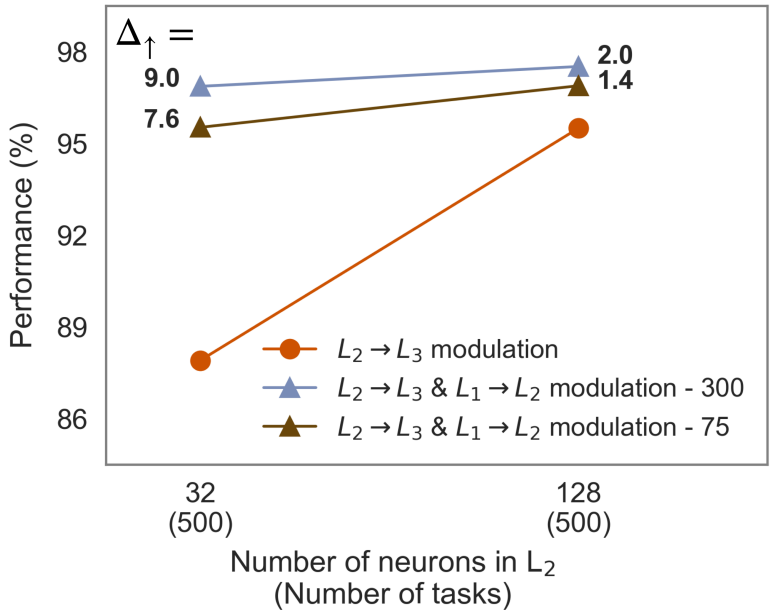

Figure 4.3: The effectiveness of task-based *gain* modulation (quantified by the performance boost, $\Delta_{\uparrow}$) of early neural processing ($L_1 \rightarrow L_2$), as a function of the number of neurons in L_2 and the number of neurons in the hidden layer of early modulation. Reducing the number of hidden neurons to 75 does not reduce the performance boost substantially. This observation suggests that the trends observed in Figure 4.2 are not simply a consequence of a disproportionate increase in the number of parameters in the different networks.

for similarities across tasks. How such task similarities could be included in learning task-based modulations in a continual learning setting is an open question and beyond the scope of our work.

4.5.3 Robustness of presented effects

To assess if the differences in the performance ($\Delta_{\uparrow}$) of the networks mentioned in Fig. 4.2 are robust, we used McNemar's test (McNemar, 1947), a statistical test used on paired nominal data. This test compares the quantities of examples where the decisions of the two networks being compared differ. If one network misclassifies examples the other network classifies correctly more often (say b examples) than the other way around (say c examples), the test statistic ($\chi^2 = (b-c)^2/(b+c)$) is higher. The test statistic is a chi-squared distribution with 1 degree of freedom.

For each network (varying on the number of hidden neurons and the number of tasks performed), we compared the outputs when only late modulation was active and when both early and late modulation (global modulation) were active. Across all the comparisons, the χ^2 values were above 15 (which corresponds to $p = 10^{-4}$). So, all the differences ($\Delta_{\uparrow}$) in Fig. 4.2 correspond to robust differences in the performance of the corresponding networks.

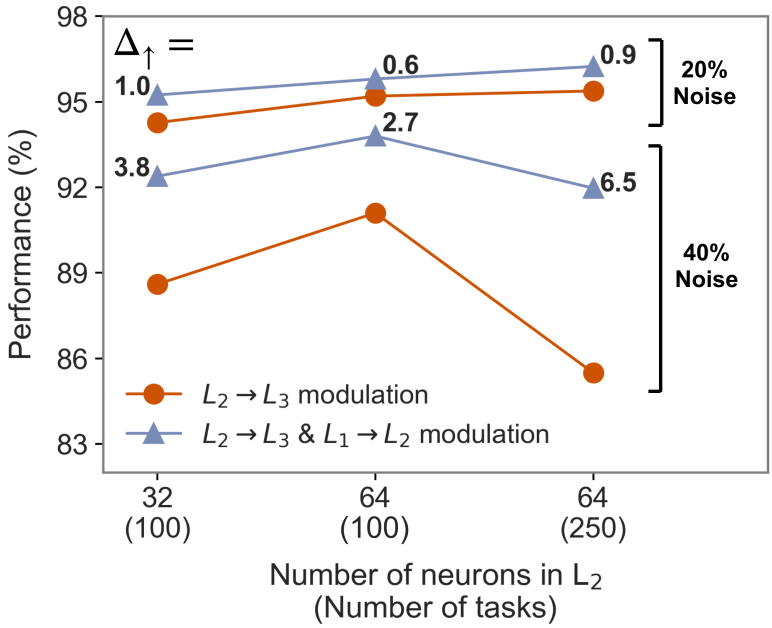

Figure 4.4: The effect of increasing task difficulty on the additive effectiveness of early modulation. The performance boost ($\Delta_{\uparrow}$) increases when the noise in the stimuli is increased making each task more difficult. The performance boost in the high noise case also decreases with an increasing number of neurons in L_2, and increases with an increasing number of tasks to be performed, echoing the trends in Fig. 4.2.

4.5.4 Additional observations about the behavior of the trained neural networks

Below we clarify the training setup and present an additional observation about the effectiveness of early modulation.

The network mainly performs permutation discrimination

In training the networks, each batch contained 50 examples corresponding to the cue (example cue: 5 in permutation 10, corresponding to task 95) and 50 examples corresponding to every other task (the invalid case, where the network outputs 'No'). As the invalid cases are drawn randomly, the probability that the same digit as the valid case (5) would be included is $1/10$. The probability that the same permutation as the valid case (permutation 10) would be included is $1/50$ (in the case of 50 permutations). In this setting, the network might thus mainly perform permutation discrimination rather than digit discrimination. To show that this was indeed the case, we probed the behavior of a network with 32 neurons in L_2 trained to perform 500 tasks (with joint modulation). When we only included permutation-matched digits as invalid test examples, the performance was 60.9%, while when we only included digit-matched permutations as invalid test examples, the performance was 94.3%. This demonstrates that the high performance of the network (Fig. 4.2) largely reflected permutation discrimination.

Modulating late neural processing in addition to modulating early neural processing does not boost performance

How well does modulating early neural processing alone perform? We trained a network with 32 neurons in L_2 to perform 250 or 500 tasks, with only the modulation of early neural processing. The performance of this network was equal to the performance of the same network trained with joint modulation. This suggests that, in this setting, only modulating early neural processing is better than only modulating late neural processing, and that late modulation does not aid performance on top of early modulation.

Early modulation might be performing so well because the stimuli used might be distinguishable at the pixel level. For example, as the network is mostly performing permutation discrimination, each permutation might have a characteristic spread of pixels that the first transformation could pick on and distinguish between the valid and invalid permutations. Switching to naturalistic stimuli might remove such low-level distinctions between stimuli, thereby not allowing the network to capitalize solely on early modulation.

In biological systems, the trade-off between wiring costs and task optimality might exist even if early modulation is always equal to or better than late modulation. This is because it might only be in the (capacity-limited) cases, where late modulation performs poorly, that early modulation is worth its wiring costs. Further work involving naturalistic stimuli would provide a better understanding of the nature of this trade-off.

Chapter 5

Statistical learning of distractor object pairs facilitates visual search

Visual search is known to depend on the relationship between the target and the distractors – i.e., how the target differs from the distractors and where the target is likely to be amongst the distractors. Whether the statistical structure amongst distractors themselves facilitates search is less well understood. Here, we assessed the benefit of distractor structure using novel shapes whose relationship to each other was learned implicitly during visual search. Participants searched for target items in arrays of shapes that comprised either four co-occurring pairs of distractor shapes (structured scenes) or eight distractor shapes randomly partitioned into four pairs on each trial (unstructured scenes). Across five online experiments (N=1140), we found that after a period of search training, participants were more efficient when searching for targets in structured vs. unstructured scenes. This structure-benefit emerged independently of whether the position of the shapes within each pair was fixed or varied, despite participants having no explicit knowledge of the structured pairs they had seen (assessed with a 2AFC task after the main experiment). These results show that learned co-occurrence statistics between distractor shapes can help increase search efficiency[1].

5.1 Introduction

The world is full of regularities amongst its constituent elements. For example, cars are usually found on roads but not on sidewalks, birds are usually found in the sky and not underwater, and forks are usually found next to plates and not clothes. The human visual system is sensitive to these regularities, with these relationships between the objects and their surroundings influencing perceptual processing of both the objects and the scenes, and the search for objects (Biederman, 1976; Bar, 2004; Bonner and Epstein, 2021). In such a structured world, visual object search capitalizes on the covariance between the target and the distractors. For example, in searching for a faucet in a kitchen, our search can be guided by how a faucet looks different from other items that can occur in the kitchen (guiding feature-based attention, Carrasco (2011)), and by the knowledge that a faucet usually exists on the kitchen counter (guiding spatial attention, Chelazzi et al. (2019)). The target location can be predicted given the arrangement of distractors (termed

[1] This chapter has been adapted from - Thorat, S., Quek, G., & Peelen, M. V. (2022). *Statistical learning of distractor co-occurrences facilitates visual search.* bioRxiv.

contextual cueing, Chun and Jiang (1998); Sisk et al. (2019)) or the presence of anchor objects - large objects that usually accompany a lot of small objects and constrain where the target object could occur in the scene (e.g., a table, Boettcher et al. (2018)).

In addition to these well-studied processes using knowledge about the distractors to predict the target location and identity, it has been proposed that regularities amongst distractors themselves could be used to make search more efficient (Kaiser et al., 2014). Recently, it has been shown that pairs of co-occurring objects in a fixed arrangement exhibit effects similar to object-based attention (Lengyel et al., 2021) conforming to the idea that the objects might get grouped into one larger object (Kaiser and Peelen, 2018). It has been proposed that such object grouping could lead to a compression of the input information from the constituent objects (Brady et al., 2009). When such grouping would occur in the distractors, this compression could effectively reduce the numerosity of distractors and therefore search complexity, enhancing search performance, similar to the numerosity reduction accounts for simpler stimuli with gestalt grouping (Zhao and Yu, 2016). For example, in our initial example, as chairs in a kitchen are usually around a table, the chairs and the table could be processed as one big object, reducing the effective number of objects amongst which the target exists, leading to a more efficient search for the faucet. The characterization of such processes related to complexity reduction of the search display, and their influence on search, is in its infancy.

In the only study directly assessing the influence of co-occurring distractor objects on visual search, Kaiser et al. (2014), participants searched for a cued object in two types of transiently-presented displays. In the regular displays, pairs of objects were presented as distractors in their regular arrangements (e.g., an egg on top of an egg cup or a lamp on top of a table). In the irregular displays, those pairs were presented in irregular arrangements (e.g., an egg cup on top of an egg or a table on top of a lamp). Participants were more accurate in indicating the location of the target in the regular displays than in the irregular displays. Separately, a different group of observers saw these same displays while in an fMRI scanner. Instead of a target, here two houses were presented amid the distractor items, with activity in the place-selective parahippocampal place area taken as an index of house-processing. Notably, PPA activation was higher when for houses embedded in regular displays compared to irregular displays, signaling lower competition from the distractors to the house representations in the regular condition. It was concluded that co-occurrences amongst the distractors could lead to them being grouped, effectively reducing the number of distractors, reducing their competition with the target, leading to better detection of the target.

The relationships between objects in the real world are learned through a lifetime of experience, both seeing and interacting with the objects. These relationships lead to strong positional and semantic associations between co-occurring objects (e.g., chimneys occur on top of stoves in the kitchen, but chimneys seldom occur with cars in any arrangement). The violation of these associations can attract attention and impair the search for another object. Such an account could also explain why participants in Kaiser et al. (2014) were better at locating the target in the regular displays. To avoid such association-violation-related influences, we consider abstract objects (simple shapes) and their co-occurrences. Numerous studies in the statistical learning literature have shown that humans can learn novel co-occurrences of shapes rapidly (Fiser and Aslin, 2001; Schapiro and Turk-Browne, 2015; Fiser and Lengyel, 2019) and possibly implicitly in the absence of task relevance and attention (Turk-Browne et al., 2009). By learning random co-occurrences between abstract shapes, any reliance on shape similarity or implied semantic similarity is avoided. Additionally, instead of violating those learned associations to assess the influence of the co-occurrences, we can just compare the co-occurring set of distractors with a set of other distractors that do not co-occur. In this study, we sought to assess whether novel

co-occurrences of abstract shapes, always appearing as distractors, could be learned during the search and be utilized to reduce the complexity of the search.

We report observations from a set of online behavioral experiments in which participants searched for pre-cued target shapes amongst scenes that consisted of co-occurring distractor pairs (structured scenes) or non-co-occurring distractors (unstructured scenes). Participants were not informed about the co-occurrences and any co-occurrence statistics they could use had to be learned during the search task. In separate experiments, the co-occurring shapes either had fixed arrangements (e.g., circle over square) or their locations within the pairs could be swapped (e.g., circle over square, or square over circle). This was done to additionally assess if fixing the arrangement of co-occurring shapes within the pairs (required by the grouping-based accounts discussed earlier) was essential for search complexity reduction. We found that participants were more efficient in searching for targets in the structured scenes than the unstructured scenes. Interestingly, this pattern was independent of whether the arrangement of co-occurring shapes within the pairs was fixed or not. Although their search performance was dependent on the co-occurrences, participants could not explicitly indicate which shapes co-occurred during the visual search experiment. In sum, we found evidence that the knowledge about distractor shape co-occurrences, gained through exposure during visual search, can lead to more efficient search.

5.2 Methods and materials

5.2.1 Stimuli

We used 20 abstract shapes (see Fig. 5.1 for examples), a subset of which overlap with those from seminal statistical learning studies (Fiser and Aslin, 2001, 2005). For each participant, we randomly assigned the shapes to three different sets that were maintained throughout the entire experiment: 8 were allocated into 4 co-occurring pairs (structured set), 4 were assigned as search targets, and the remaining 8 shapes were used to create 4 random pairs on each new trial (unstructured set). Critically, a shape assigned to the structured set only ever appeared in a vertical pairing with its nominated partner shape. There were two types of possible structures. In one experiment, the shape pairs in the structured set had fixed arrangements (e.g., circle always occurred with a square and always on top of the square). In the other experiment, the shape pairs had free arrangements (e.g., circle always occurred with a square but could be either on top of or below the square). These two types of structures occurred in separate experiments. In contrast, on any given trial, a shape assigned to the unstructured set could be paired with any other shape from the unstructured set and occupied either the top or bottom position within this random pairing.

5.2.2 Visual search task

On each trial, participants saw a letter cue indicating which of the 4 memorized target shapes they had to search for in the upcoming display. After a brief delay, a search display with 10 shapes appeared (either structured or unstructured, see below). Participants used the keyboard to indicate whether the target was present on the left or the right side of the display. The temporal details of the trial structure are mentioned in Fig. 5.1B.

The search display was $16\,\text{em} \times 28\,\text{em}$, where em is the font size on the participant's display. This size was chosen such that the display would approximately extend around 6 degrees (parafoveal) of visual angle. We did not employ any methods to explicitly control for the visual angle subtended by the search display. We reasoned that participants who use smaller screens (desktop or laptop only) have smaller font sizes and hold them relatively closer to their eyes, to

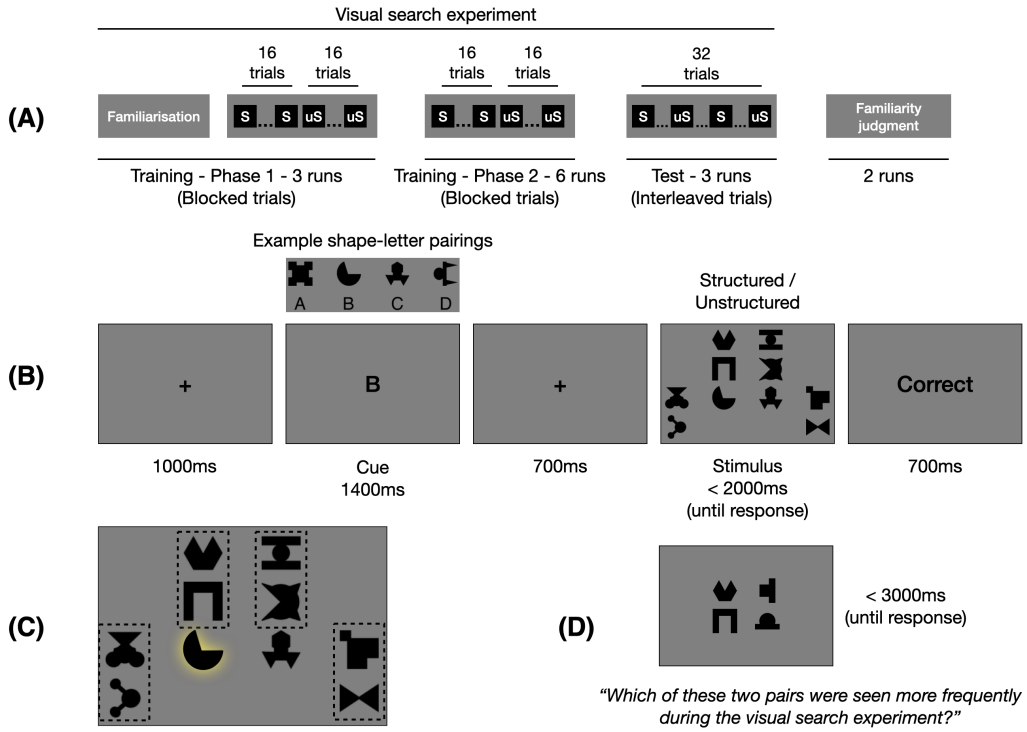

Figure 5.1: Experimental design. (A) The structure of the experiment. S refers to structured scenes and uS refers to unstructured scenes. In the training runs of the visual search experiment, the structured and unstructured scenes were blocked, while they were interleaved in the test runs. The visual search experiment was followed by a familiarity judgment task. (B) The trial structure of the visual search experiment. Participants had to search for a target shape cued by its corresponding letter in the subsequent search display and indicate if the target was present on the left or the right part of the display within 2 s. (C) The layout of the search display. 10 shapes were presented - 8 distractors, 1 target (highlighted in yellow - color not shown during the experiment), and 1 foil (which could be a target on other trials). The distractors were presented as 4 pairs. In the structured scenes, the distractors co-occurred in pairs of two(with either fixed arrangements within the pairs or not - in separate experiments). In the unstructured scenes, the distractors were randomly partitioned into four pairs on each trial. The search performance was compared between structured and unstructured displays. (D) A trial of the familiarity judgment task. Participants had to judge which of the two pairs (one taken from the structured scenes and the other from the unstructured scenes) were seen more frequently during the visual search experiment.

effectively equate the visual angle subtended by the relevant stimuli with the angle subtended by the stimuli for participants using larger screens.

Each search display consisted of 4 distractor shape pairs, the target shape, and a foil shape (i.e., one of the other three target shapes not currently being searched for) arrayed symmetrically on a 4×4 grid (Fig. 5.1C). All the shape pairs in a given display were taken from the same scene condition (e.g., 4 pairs from the structured set (either the fixed or free arrangement - in the different experiments) or 4 randomly generated (on each trial) pairs from the unstructured set). Thus, on each trial, participants searched for the target in either a structured or unstructured scene. The

4 shape pairs were placed at random locations, one in each column of a 4×4 grid. The locations were mirrored horizontally. The target appeared in one of the remaining locations adjoining the pairs and a foil (one of the other 3 targets) was presented on the horizontally mirrored location. The location randomization process ensured that the probability of the target's location was uniform across the entire grid. We assessed the dependence of participants' search performance, in terms of accuracy or reaction time, on the scene structure.

Each run of the visual search experiment consisted of 16 trials each from the structured and unstructured conditions. The trial types were either interleaved or completely blocked (rationale explained below). The order of blocking (structured trials first or unstructured trials first) was maintained for a participant throughout the experiment and balanced across participants.

There were three types of runs in the experiment - two phases of training (blocked scene conditions), and test (interleaved scene conditions). In phase 1 of the training runs, each run with blocked scene conditions was preceded by a familiarisation block where the letters and their associated target shapes were shown sequentially four times. Participants completed 3 phase-1 training runs. In phase 2 of the training runs and the test runs, there were no familiarisation blocks, but at the beginning of each run, all the letter-shape associations were shown with the instructions to refresh participants' memory. Participants completed 6 phase-2 training runs with blocked scene conditions, followed by 3 test runs with interleaved scene conditions. The experimental design is schematized in Fig. 5.1A.

We blocked the scene conditions for most of the training trials since there is evidence (Flesch et al., 2018) that humans learn multiple statistical associations (that some shapes were grouped and the others were not) faster when the different statistical associations were presented in a blocked than interleaved order. We included the three test runs with interleaved scene conditions at the end of training to help eliminate any strategy differences blocking might have introduced. For example, within each block, the distractor shapes stay the same across the trials which might have helped participants in making their search extremely efficient, reducing any efficiency differences across the blocks due to ceiling effects. Our assessment of the impact of the structured distractors on visual search is therefore focussed on the post-training test runs.

5.2.3 Familiarity judgement task

After the main search experiment had concluded, we used a two-alternative-forced-choice (2AFC) familiarity judgment task (Fiser and Aslin, 2001) to assess participants' knowledge about the shape pairs presented in the visual search experiment as distractors. To directly contrast the shapes from the two scene conditions, we created 4 random (forced) pairs from the shapes in the unstructured scene condition. These 4 forced pairs were objectively seen less frequently (14 times less frequent) than the 4 pairs from the structured scenes, and we assessed if participants had explicit knowledge about this fact. 16 comparisons were possible between these pairs from the structured and unstructured scenes. These comparisons were termed original comparisons in contrast to the partner-swapped and position-swapped comparisons described below.

In the case of the fixed arrangement condition for the co-occurring pairs (Experiment 2A and Experiment 3), we additionally assessed whether the two constituent regularities - shape co-occurrence and relative positioning - were also registered by the participants. To do so we asked participants to compare partner-swapped and position-swapped versions of the pairs from the structured and unstructured scenes. 4 partner-swapped pairs were constructed from the structured or the unstructured scenes, by taking the 4 fixed pairs or the 4 forced pairs respectively, and swapping the partners of the shapes while maintaining their relative positions in the pairs. 4 position-swapped pairs were constructed from the structured or unstructured scenes, by taking

the 4 fixed or forced pairs respectively and swapping the positions of the shapes within their pairs. These two additional manipulations led to 32 more comparisons between the shapes from the structured and unstructured scenes.

In the case of the free arrangement condition for the co-occurring pairs (Experiment 2B), no partner-swapped and position-swapped comparisons were assessed, as position-swapping was redundant and partner-swapping would have destroyed all co-occurrence information. To compare between the pairs from the structured and unstructured scenes, in addition to the creation of the 4 forced pairs for the unstructured set as explained above, 4 forced pairs were created for the structured set too where the locations of the shapes within the 4 pairs were fixed, thus again leading to 16 possible comparisons between the pairs from the structured and unstructured scenes.

On each trial of the familiarity judgment task, participants were shown 2 pairs corresponding to one of the three comparisons (see Fig. 5.1D) between the pairs from the structured and unstructured scenes and were asked to indicate - guess if they have to - which of the two pairs they saw more frequently during the visual search experiment, within 3 s.

5.2.4 Experiments and participants

Participants were recruited from Prolific (Prolific, 2021) and the experiment was hosted online on Pavlovia (Open Science Tools Limited, 2021). They provided informed consent before beginning the experiment. Participants from whom we obtained partial data from Pavlovia were excluded from the analysis (10% dropout rate). For any given experiment requiring a particular number of participants, we tested around that number of participants balancing the blocking order of scene structure. Then participants whose overall accuracy and reaction times were above or below 3 standard deviations (SDs) from the means were removed (outlier detection). This was done iteratively until no exclusions happened. Then more participants were added to get to the desired number and this exclusion process was repeated. In the end, we obtained the desired number of participants for each experiment whose accuracies and reaction times (for correct responses) were within 3 SDs from the means and the blocking order was balanced. This procedure resulted in a further 10% of the total participants being rejected from subsequent analysis.

Five online studies were conducted. Two pilot experiments (N = 40 each) were conducted to preliminarily assess the differences in search efficiency between the scene conditions, one experiment each with fixed (Experiment 1A) and free arrangements (Experiment 1B) for the co-occurring shapes in the pairs in the structured scenes. The pilot experiments did not contain the familiarity judgment task. These experiments were followed by two large sample experiments (N = 400 each), one each for the fixed (Experiment 2A) and free arrangement (Experiment 2B) cases. These experiments contained the familiarity judgment task for a subset of the participants (see the section Familiarity judgment task). These experiments were followed by a pre-registered replication (Experiment 3; N = 260) for the fixed arrangement case.

In the large sample experiment for the fixed arrangement case (Experiment 2A), after the exclusion process (ran two times to obtain 200 participants each corresponding to the slight differences in the familiarity judgment task), 400 participants' data were analyzed. As the 400 participants' accuracies and reaction times fell within 3 SDs of their means, they were pooled for further analysis.

In the familiarity judgment task, for the first 200 participants, in each of the two runs, the 16 original comparisons between the 4 pairs from the structure scenes and 4 forced pairs from the unstructured scenes were interleaved with the 16 partner-swapped and 16 position-swapped comparisons. Feedback on the responses was provided at the end of each run. For the remaining 200 participants and in the following pre-registered replication experiment, in each of the two

runs, the 16 original comparisons between the 4 pairs from the structured scenes and 4 forced pairs from the unstructured scenes were shown first, followed by the 16 partner-swapped and 16 position-swapped comparisons interleaved. Feedback on the responses was provided only at the end of the two runs. The task was challenging leading to non-responses on many trials. Therefore, we used the criterion for only the inclusion of participants who responded for at least one each of the original, position-swapped, and partner-swapped comparisons, in each run. The familiarity judgments of 368 of the 400 participants were further analyzed for any associations between the familiarity judgments and the structure-related efficiency differences in the visual search task.

In the large sample experiment for the free arrangement case (Experiment 2B), after the exclusion process (ran two times, once to obtain 200 participants who did not complete the familiarity judgment task and then to obtain 200 participants who did complete the familiarity judgment task), 400 participants' data were analyzed. As the 400 participants' accuracies and reaction times fell within 3 SDs of their means, they were pooled for further analysis.

The first half of the participants did not complete the familiarity judgment task. For the second half of the participants, in each run, only the 16 original comparisons between the 4 forced pairs from the structured scenes and 4 forced pairs from the unstructured scenes were shown. All 200 participants responded for at least one of the comparisons, in each run, and their data were further analyzed for any associations between the familiarity judgments and the structure-related efficiency differences in the visual search task.

Experiment 3 was designed as a replication of the interaction between the structure-related efficiency differences and familiarity judgments across the participants, for the fixed arrangement case. In the exclusion process, participants who did not respond to at least one each of the original, position-swapped, and partner-swapped comparisons, in each run, were additionally excluded before running the outlier detection analysis. 260 participants' data were analyzed as per the requirement of the replication (pre-registration form can be found on AsPredicted (Wharton Credibility Lab, 2021): `https://aspredicted.org/blind.php?x=5ne7qa`).

5.3 Results

5.3.1 Search efficiency as a function of distractor structure in the scenes

To assess if the presence of co-occurring distractors facilitated participants' search, we evaluated the difference between search performance in structured scenes and unstructured scenes in terms of both accuracy and reaction time (for the correct response trials) in the test runs (i.e., after a period of exposure during training). This difference was termed the structure-benefit (indicated by a higher search accuracy or faster reaction times in the structured scenes). We first used a two-sample t-test to assess if the structure-benefit was different across the two arrangement types for co-occurring shapes, i.e., fixed or free within their pairs. When a significant difference was identified, we conducted individual one-sample t-tests for the structure-benefits corresponding to the two types of arrangements within pairs. We used the inverse efficiency score (IES = average reaction time / average accuracy) which includes both the speed and accuracy of the search as a primary metric for the structure-benefit.

In the pilot experiments (N= 40 each; Fig. 5.2), the structure-benefit did not differ across the arrangements within pairs in the IES (two-sample t-test: $t_{78} = 1.38$, $p = 0.17$), similarly reflected in accuracy ($t_{78} = 0.71$, $p = 0.48$) and reaction times ($t_{78} = 1.36$, $p = 0.18$). Pooling across the arrangement types, there was evidence for structure-benefit in the IES (one-sample t-test: $t_{79} = 3.7$, $p = 4E-4$), reflected both in the accuracies ($t_{79} = 3.6$, $p = 5E-4$) and the reaction times ($t_{79} = 2.3$, $p = 0.03$). Thus, these pilot experiments provided evidence that

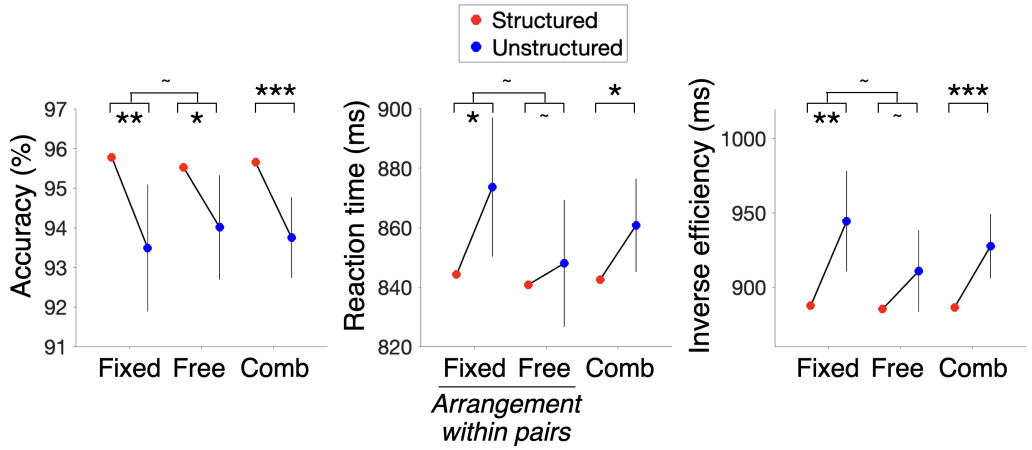

Figure 5.2: Search efficiency as a function of scene condition: Pilot experiments 1A and 1B. Structure-benefit (increased accuracy or decreased reaction time or decreased inverse efficiency in the structured scenes) was observed in both the experiments with fixed or free arrangements of the co-occurring shapes within their pairs. As no differences were observed between the experiments in either of the measures, the data from the two experiments were combined ('Comb') to accumulate the evidence for the structure-benefit. Error bars indicate 95% confidence intervals for the structure-benefit on each measure, for each experiment. The asterisks indicate p-values for the t-tests for the corresponding comparisons (*p$<$ 0.05, **p$<$ 0.01, ***p$<$ 0.001, p$>$ 0.05).

participants search for targets more efficiently in the context of structured distractor arrays than unstructured ones, irrespective of the arrangement of pairs in the structured scenes.

Next, we conducted large sample experiments for the two arrangement types with two goals in mind: First, to check if the structure-benefit observed in the pilot data was robust (i.e., replicable in a large sample), and second, to obtain an estimation of participants' explicit familiarity for the co-occurring shapes. Here we used one-sided t-tests to test for the existence of structure-benefits since we had preliminary evidence suggesting the direction of the effect from Expts. 1A and 1B. In these large sample experiments (N = 400 each; Fig. 5.3), we again observed that the structure-benefit did not differ across arrangement type within pairs in the IES (two-sample t-test: $t_{798} = 0.26$, $p = 0.8$), similarly reflected in the accuracy ($t_{798} = -1.42$, $p = 0.16$) and reaction time ($t_{798} = 1.18$, $p = 0.24$). Pooling across arrangement type, there was evidence for structure-benefit in the IES (one-sample, one-sided, t-test: $t_{799} = 2.8$, $p = 3E-3$), which was reflected both in the accuracy ($t_{799} = 2.5$, $p = 6E-3$) and the reaction time ($t_{799} = 1.8$, $p = 0.04$). Thus, we found additional, confirmatory, evidence that after a period of exposure to distractor co-occurrence in the search displays, participants performed a more efficient search in the structured scenes than the unstructured scenes. Notably, the benefit of distractor co-occurrence was evident irrespective of whether the co-occurring shapes in the structured scenes occurred in fixed or free arrangements within their pairs.

5.3.2 Explicit knowledge of the distractor structure and its relationship to structure-benefit

Did participants explicitly recognize the distractor structure, as in previous experiments where the co-occurrences were attended (Fiser and Aslin, 2001)? To assess whether this was the case,

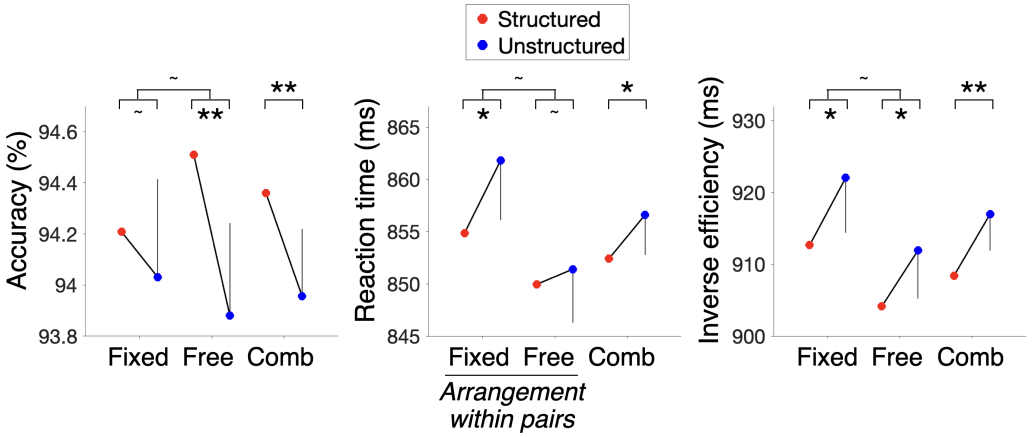

Figure 5.3: Search efficiency as a function of scene condition: Large sample experiments 2A and 2B. Structure-benefit (increased accuracy or decreased reaction time or decreased inverse efficiency in the structured scenes) was again observed in both the experiments with fixed or free arrangements of the co-occurring shapes within their pairs, replicating the effects from the pilot experiments. As no differences were observed between the experiments in either of the measures, the data from the two experiments were combined ('Comb') to accumulate the evidence for the structure-benefit. Error bars indicate 95% confidence intervals for the structure-benefit on each measure (corresponding to a one-sided t-test), for each experiment. The asterisks indicate p-values for the t-tests (*p$<$ 0.05, **p$<$ 0.01, ***p$<$ 0.001, p$>$ 0.05).

we inspected familiarity judgments for structured and unstructured pairs for a subset of the participants in the large sample experiments (see Methods and Materials for the selection details). We defined familiarity score as the proportion of responses where the pairs corresponding to the shapes from the structured scenes were indicated as more familiar than the pairs corresponding to the shapes from the unstructured scenes. The familiarity scores were averaged across the two runs of the task for each comparison type.

Since familiarity scores given by the original comparisons (the pairs from the structured set vs the pairs from the unstructured set) did not differ between Experiments 2A and 2B (two-sample t-test: $t_{566} = 0.9$, $p = 0.4$), we pooled the data across the two experiments. Here, familiarity scores for the original comparisons did not differ from 0.5 (one-sample t-test: $t_{567} = 1.7$, $p = 0.09$), suggesting that on average, observers did not have explicit knowledge of which shapes co-occurred during the search task, although these co-occurrences boosted search efficiency.

Although there was no difference in familiarity for structured vs. unstructured pairs at the group level, could there be individual differences across participants that would be reflected as an association between the structure-benefit with the familiarity judgments? It could be the case that the participants who exhibit a higher structure-benefit would be more explicitly familiar with the distractor co-occurrences as they might be using the co-occurrences to make their search more efficient. To this end, we assessed the correlation between the participants' familiarity scores and their structure-benefit reflected in IES. We observed a negative correlation in Experiment 2A ($r = -0.16$, $p = 0.001$; Fig. 5.4A), but not in Experiment 2B ($r = 0.02$, $p = 0.67$; Fig. 5.4B; across-experiment test for a difference between correlations, computed using the calculator in `https://www.psychometrica.de/correlation.html`, that follows Eid et al. (2017): $z = 2.05$, $p = 0.02$). Contrary to our expectations, in Experiment 2A, participants

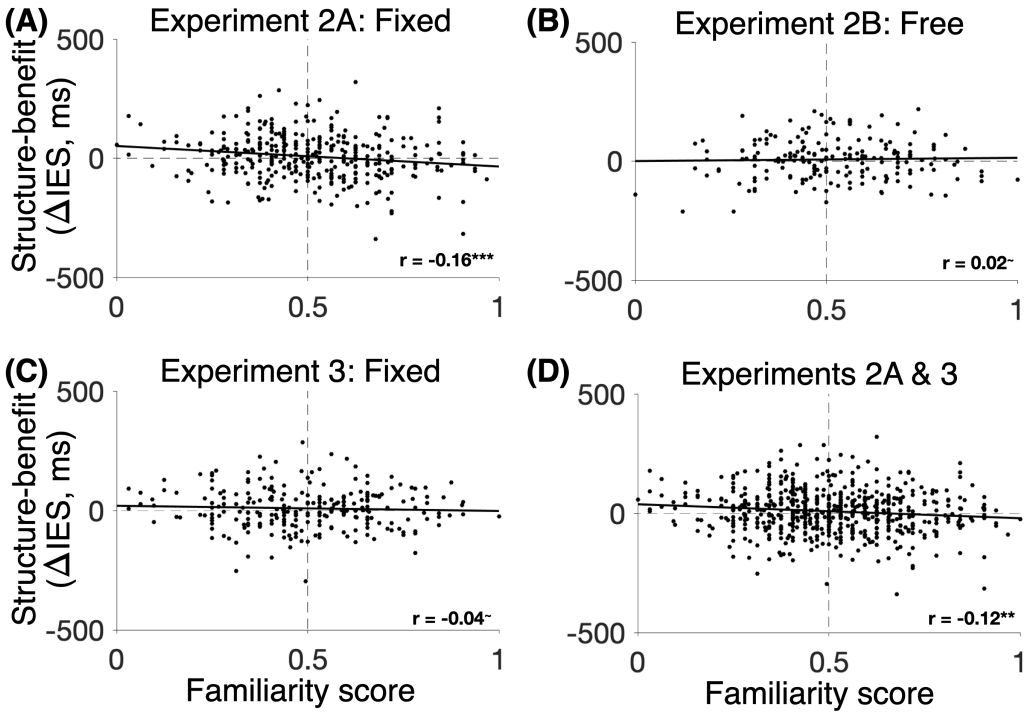

Figure 5.4: The relationship between the structure-benefit and explicit knowledge about the co-occurring distractors. (A) In Experiment 2A, with the fixed arrangement of co-occurring distractors within their pairs, the structure-benefit (in the inverse efficiency score, IES) was negatively correlated with the familiarity scores. (B) In Experiment 2B, with the free arrangement of co-occurring distractors within their pairs, no such correlation was observed. (C) Experiment 3 did not replicate the negative correlation found in Experiment 2A. (D) However, as there was no evidence for a difference in the correlations in Experiments 2A and 3 (see text for details), pooling across the two experiments, we found evidence for a negative correlation between the structure-benefit and the familiarity scores. Error bars indicate 95% confidence intervals for the structure-benefit (corresponding to a one-sided t-test). The asterisks indicate p-values for the t-tests (*p< 0.05, **p< 0.01, ***p< 0.001, p> 0.05).

who indicated that the structured pairs were more familiar had less structure-benefit.

To assess if the negative correlation found in Experiment 2A was replicable, we ran a pre-registered experiment (Experiment 3; N $= 260$) that mirrored Experiment 2A. Pleasingly, the test runs of Experiment 3 demonstrated a structure-benefit in IES (one-sample t-test: $t_{259} = 1.7$, p $= 0.04$), replicating the main findings from the pilot and large sample experiments. Mirroring the findings of Experiment 2A, the familiarity scores for the original comparisons across participants did not differ from 0.5 (one-sample t-test: $t_{567} = -0.3$, p $= 0.79$). However, unlike in Experiment 2A, we found no evidence for a correlation between the familiarity scores and the structure-benefit across participants (r $= -0.04$, p $= 0.44$; Fig. 5.4C), failing to replicate the negative correlation observed in Experiment 2A.

We wondered if some differences between the responses in Experiments 2A and 3 could explain the non-replication of the negative correlation. However, there was no difference be-

tween the two experiments in either the magnitude of the the structure-benefit in IES or the familiarity scores for the original comparisons (two-sample t-tests, structure-benefit: $t_{626} = 0.2$, $p = 0.8$; familiarity score: $t_{626} = 0.7$, $p = 0.5$). Additionally, there was no evidence that the correlations differed between these two experiments (across-experiment test for a difference between correlations, computed using the calculator in `https://www.psychometrica.de/correlation.html`, that follows Eid et al. (2017): $z = 1.5$, $p = 0.07$). Thus, we considered it justified to pool the data across Experiments 2A and 3. This global analysis confirmed the negative correlation between the structure-benefit and the familiarity scores ($r = -0.12$, $p = 0.003$). We might have overestimated the true value of the correlation given the samples from Experiment 2A, leading to a null result in Experiment 3. Further testing, with a larger number of participants (estimated including the updated correlation values), is required to assess the robustness of this finding of a negative correlation between the structure-benefit and the familiarity scores.

5.4 Discussion

In this study, we report evidence that observers exploit statistical co-occurrences between distractor shapes to increase efficiency during the search. This benefit of scene structure arose irrespective of whether the spatial arrangement of co-occurring shapes in the pairs was fixed or not. Surprisingly, the increase in search efficiency was not accompanied by an increase in participants' subjective familiarity with the underlying statistical regularities (if anything, they seem to be inversely related). These findings provide support to the idea that humans can make their search more efficient by utilizing statistical regularities in the environment.

How might reliable co-occurrences between distractor items give rise to a visual search benefit? Object grouping has been proposed as a complexity reduction mechanism supporting more efficient search (Kaiser et al., 2014, 2019). In Lengyel et al. (2021), fixed arrangements of co-occurring objects produced object attention effects, thus providing further support to the idea that co-occurring objects could be treated as one large object, which could lead to numerosity and search-complexity reduction. We found a similar effect in our data, insofar as co-occurring distractor shapes in fixed arrangements within their pairs produced more efficient search than randomly paired distractor shapes. However, a search benefit was also present (and not statistically different in magnitude) when the co-occurring shapes had no fixed arrangement, i.e., could vary freely in their spatial arrangement within the pair. The latter finding does not fit easily with an object grouping account similar to that proposed by Lengyel et al. (2021) - unless we assume that observers effectively learned two large objects, corresponding to the two configurations of the co-occurring objects, giving rise to numerosity and search-complexity reduction.

An alternate explanation is possible: while a distractor shape was rejected as the target, the nearby (above or below) co-occurring, and therefore predictable, distractor could have been rejected as a target faster and more accurately (via a mechanism such as inter-object priming, see Stein et al. (2015) - although in that case, such priming was facilitative), leading to quicker localization of the actual target. Such inter-object priming could happen serially on some trials, leading to the observed benefit for co-occurring distractor items. Alternatively, perhaps such inter-object priming happens weakly (corresponding to the non-existence of the explicit knowledge about the co-occurrences), but in parallel across multiple distractor locations, effectively creating a better priority map for attentional orientation that can facilitate target identification (Chaumon et al., 2008; Zinchenko et al., 2020). Both accounts could explain the small effect size (5 ms of structure-benefit in the reaction times) observed. Support for the second account is found in Chapter 6, where using EEG, we present preliminary evidence that the increased search efficiency due to the distractor co-occurrences is related to the increased attentional orientation towards the

target (as indexed by the N2pc component - between 200 to 400 ms after the onset of the search display; Luck (2012)).

An important aspect of our design is that it targets learning of statistical regularities outside the focus of attention. Here, co-occurrences of distractor items were irrelevant to the participant's task - not just in the sense that it did not matter if the participants explicitly paid attention to the regularities (as is the case in the seminal statistical learning studies - Fiser and Aslin (2001, 2005)), but insofar as deliberately attending to co-occurrences between distractors would likely have been detrimental to the task participants had to perform (i.e., find an abstract target in a complex array, with the target item changing trial to trial). Prior work has proposed that statistical learning is gated by visual attention (Turk-Browne et al., 2005), such that co-occurrences between items are not indicated as familiar post-experiment when the co-occurrences are task-irrelevant. Indeed, we found here that participants had no explicit familiarity with pairs of co-occurring shapes, even though these co-occurrences affected their visual search. Is this evidence for implicit learning of the statistical regularities (Turk-Browne et al., 2009)? The absence of informative judgments about the familiarity to the co-occurrences might just be due to the low sensitivity of the differences between the neural processing underlying the two types of scenes and does not necessarily reflect an absence of familiarity while affecting search which would have made the case for implicit learning (Meyen et al., 2021). Although participants could not indicate which shapes co-occurred, further neuroimaging studies could reveal the neural correlates of the co-occurrences in the visual system (Kaiser and Peelen, 2018) and the hippocampus and other regions (such as the medial temporal lobe) involved in statistical learning (Turk-Browne et al., 2009; Schapiro and Turk-Browne, 2015).

Yet another finding was the possibility of a negative relationship between how familiar participants indicated the co-occurring shapes to be and how much their search efficiency was boosted by the co-occurring shapes (see also: Spaak and de Lange (2020) who report a similar negative relationship between a contextual cueing effect and the participants' familiarity with the regularities). What neural processes could underlie such a relationship? It has been proposed (Meyer and Rust, 2018) that the neural representations of memorized objects in the inferotemporal cortex reflect both the visual differences between the objects (in terms of the neural activity patterns) and the familiarity of those objects (in terms of the overall neural activity: lower familiarity = lower activity). In our study, participants who showed higher structure-benefit could have suppressed the co-occurring distractors more efficiently. This suppression throughout the visual search experiment might have been reflected as a reduced activity associated with the distractor shapes. Subsequently, when participants had to judge the familiarity of the co-occurring pairs, those that suppressed the co-occurring pairs more might have reported lower familiarity towards those pairs. Future behavioral studies re-assessing the existence of this negative relationship and neuroimaging studies directly assessing the neural representations of the co-occurring shapes could help us understand the validity of this account.

In summary, beyond utilizing the regularities in the environments to predict (in addition to the target's identity) where the target object could appear, the regularities amongst the distractors in the environment could themselves be used to reduce the complexity of the search. The former process has been studied extensively in the contextual cueing and anchor object literature (Boettcher et al., 2018; Sisk et al., 2019). Reliance on the regularities amongst objects to compress the incoming information for efficient and rapid visual processing and memorization has been demonstrated (Bar, 2004; Brady et al., 2009; Kaiser et al., 2019). In Kaiser et al. (2014) and this study, such regularities amongst the distractors themselves, not predicting any aspect of the target, were shown to aid in visual search. Grouping of (Lengyel et al., 2021), and inter-object priming between (Stein et al., 2015), the distractors were proposed as mechanisms underlying

these regularity-driven search benefits. Going beyond the artificial stimuli and displays used here, future research with natural objects and scenes could help elucidate how impactful such complexity reduction based on the regularities amongst the distractors is during real-world search.

Chapter 6

The impact of distractor object co-occurrences on the orientation of attention in visual search

Efficient visual search capitalizes on the structure in the environment. Aspects of target-distractor co-occurrences - how the target looks as opposed to the distractors and where the target could be located - have been well studied and the neural mechanisms supporting these processes are well characterized. On the other hand, how the co-occurrences amongst distractor objects aid visual search is a nascent line of investigation. It has been proposed that co-occurring objects (e.g., mirror and sink) could be grouped and treated as one object, reducing the numerosity of distractors in the scene and allowing the search to be more efficient. Chapter 5 provided evidence for this proposal showing that the search for targets amongst co-occurring distractor shapes (structured scenes) was more efficient than the search amongst non-co-occurring distractors (unstructured scenes). In this study, using EEG recordings we observed that the increased search efficiency in the structured scenes was associated with increased attentional orientation towards the target during the search, indexed by the N2pc component. The neural representations of the co-occurring distractor pairs, obtained through separate runs during the experiment where the pairs were presented in isolation, were not found to be affected by exposure to the co-occurrences during the visual search runs of the experiment. These results indicate that the increased efficiency of search in the structured scenes found in Chapter 5 was an outcome of the increased attentional orientation towards the target, possibly driven by better distractor rejection owing to the distractor co-occurrences.

6.1 Introduction

Efficiently focussing on relevant information is crucial while searching for objects in complex real-world scenes. How the human brain achieves this feat is an ongoing investigation. Target-distractor relationships (e.g., computers look different from other objects and are mostly found on tables) are used to efficiently identify (via feature-based attention and object recognition) and locate (via spatial attention) the target in everyday search (Carrasco, 2011; Chelazzi et al., 2019). Recent studies have revealed that, in addition to the well-studied, target-distractor relationships, statistical regularities amongst distractors themselves can also be exploited by observers to in-

crease search efficiency (Kaiser et al., 2019). These statistical regularities are not just restricted to low-level features, leading to Gestalt grouping (Zhao and Yu, 2016), but can also exist at the level of co-occurrences between individual objects (Kaiser et al. (2014); see also Chapter 5). For example, in natural scenes, objects such as mirrors and sinks, or chimneys and fireplaces, commonly occur together and can be effectively grouped as a single unit, increasing the efficiency with which an observer can parse a scene and locate a target. The neural mechanisms underlying the employment of such object-level statistical regularities in visual search are unclear.

In Chapter 5, a series of online experiments with a large participant pool showed that visual search amongst abstract distractor shapes which co-occurred in pairs (i.e., structured scenes) was more efficient than when the distractor shapes did not co-occur (i.e., unstructured scenes). The co-occurring shapes could occur in fixed spatial arrangements or free spatial arrangements. That study provided evidence that observers can make use of distractor co-occurrence statistics to increase search efficiency, regardless of the spatial arrangement within pairs. In this study, we sought to reveal the stages of neural processing related to the visual search that were influenced by the presence of those co-occurring distractor shapes, by using electroencephalography (EEG). Our goal was two-fold: 1) to assess the impact of distractor co-occurrences in the scenes on the orientation of attention to the target, and 2) to assess what impact exposure to distractor co-occurrences has on the neural representations of those distractors. As a starting point, we chose to focus solely on the co-occurring pairs with fixed spatial arrangements within the pairs.

A possible explanation for the increased search efficiency in the structured scenes observed in Chapter 5 concerns improved deployment of visual attention to the target when distractor items are groupable thanks to their co-occurrence. If this is the case, then a relevant electrophysiological metric to consider will be the N2pc, which is commonly understood to index attentional deployment in space. The N2 posterior contralateral (N2pc) component in the event-related potentials (ERP) in the EEG signal is considered to be an index of the deployment of attention in the ventral visual pathway during visual search (Luck, 2012). It is observed as a negative deflection that is larger over the hemisphere contralateral to the attended location compared to the ipsilateral hemisphere, around $200 - 400$ ms after the onset of the search display. The increased search efficiency observed in Chapter 5 about the co-occurrences could have been a result of rapid distractor suppression processes that led to the deployment of top-down attention towards the target, similar to the deployment of attention in typical visual search experiments with simple shapes, where the target shape is cued on each trial (see Luck (2006)). Therefore, mirroring the increase in search efficiency observed in Chapter 5, we hypothesized that the N2pc deflection might be higher during the search in the structured scenes as compared to the unstructured scenes. In addition to the influence of the structure on the N2pc, due to the reduction in scene complexity due to grouping, we also hypothesized the decodability of the target's shape and location in the scene, from the neural representations, might also have improved.

A separate question of interest concerned whether the neural representations of the distractor shapes might change as a result of exposure to their co-occurrences during the visual search task. Co-occurrences between objects affect the neural representations of those objects (Bonner and Epstein, 2021; Turk-Browne et al., 2009; Kaiser and Peelen, 2018). It has been shown that co-occurring objects - with the co-occurrences learned during the experiment - could get grouped, effectively becoming one new object (Lengyel et al., 2021), showing object attention effects. Additionally, co-occurring objects could elicit neural responses unlike any of the component objects (Kaiser and Peelen, 2018), and/or each of the component objects' representations might get biased towards the other component object (Yu and Zhao, 2018). Exposure to the distractor co-occurrences during the visual search task in Chapter 5 might have led to changes in the distractor representations too. We hypothesized that the shape pairs from the structured scenes

might have been grouped, getting expressed as distinct entities on par with objects. This grouping could have resulted in neural representations distinct from the pairs from the unstructured scenes. Additionally, the neural representations of the grouped pairs might have become more distinct from each other due to them becoming object-like, as compared to the distinctiveness of the pairs from the unstructured scenes.

This study evaluated the above hypotheses about the neural underpinnings of the increase in search efficiency due to structure in the distractors observed in Chapter 5. After an online training session where participants performed the visual search task, we collected EEG data as participants performed 1) a visual search task containing structured or unstructured distractor pairs, and 2) a central letter discrimination task where the individual shape pairs from the structured and unstructured scenes were presented peripherally (to obtain neural representations of those pairs). We present evidence for an association between the increased search efficiency due to structured distractors and increased N2pc deflection for search amongst the structured distractors, which sheds light on the processes underlying the observations in Chapter 5. No evidence was found for changes in neural representations of either the target shape or the distractor shape pairs attributable to the structure in the scenes.

6.2 Methods and materials

6.2.1 Participants

32 English-speaking individuals (11 males; age range: $18-35$, mean = 23.9 years) completed the experiment in exchange for monetary compensation. All reported having a normal or corrected-to-normal vision and no neurological or psychiatric history. Before experimental testing, all participants gave their written informed consent. Experimental data were stored under pseudo anonymized codes per the European General Data Protection Regulation. The study was approved by the Radboud University Faculty of Social Sciences Ethics Committee (ECSW2017–2306-517). We excluded data from one participant whose testing session was terminated prematurely due to fatigue; the final sample consisted of 31 subjects.

The effects reported in Chapter 5 were small and we might require hundreds of participants to find those effects. Specifically, we would need to test 546 participants to be afforded 90% power for detecting the effect (across all the 1140 participants) observed in Chapter 5 (power computed using G*Power 3.1; Faul et al. (2009)). This requirement was not satisfied by the current study - the current sample affords us 16.8% power for detecting that effect. However, we reasoned that even if the structure-benefit was not found across participants here, the neural representational differences, proposed in the subsequent sections, between the two conditions might still be observed, or perhaps the variability between participants could still be linked to the variability in the various neural representational differences.

6.2.2 Stimuli

Stimuli were the same 20 abstract shapes as used in Chapter 5, a subset of which overlap with those from seminal statistical learning studies (Fiser and Aslin, 2001, 2005). For each participant, we randomly assigned the shapes to three different sets that were maintained throughout the entire experiment: 8 were allocated into 4 fixed pairs (structured set), 4 were assigned as search targets, and the remaining 8 shapes were used to create 4 random pairs on each new trial (unstructured set). Critically, a shape assigned to the structured set only ever appeared in a vertical pairing with its nominated partner shape, always in the same relative position (e.g., circle above square).

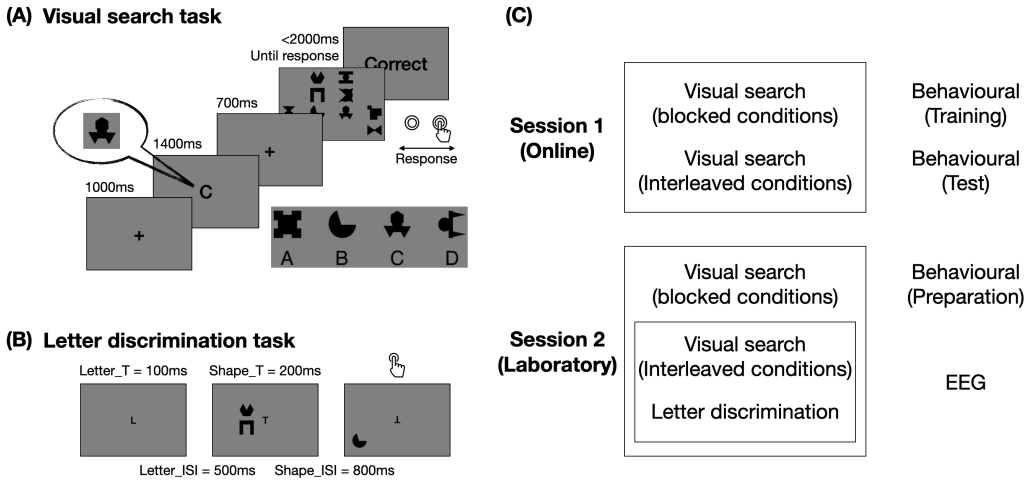

Figure 6.1: Experimental design. (A) In the visual search task, participants searched for a target cued with its associated letter in the upcoming search display. They indicated if the target was present on the left or right part of the display within 2 s. There were two types of search displays: the structured scenes contained 4 pairs of co-occurring distractors and the unstructured scenes contained 8 distractors randomly partitioned into 4 pairs on each trial. (B) In the letter discrimination task, at fixation participants monitored the letters L and T for inversion, and pressed a button if an inverted T was seen (the Letter stream). Concurrently, task-irrelevant singleton target shapes and shape pairs from the structured and unstructured scenes were presented in the locations they occupied during the visual search task (the Shape stream). (C) The structure of the experiment. Session 1 was conducted online and mirrored the experiment in Chapter 5. In Session 2, participants completed six runs of the visual search task as they were prepared for EEG, and subsequently completed six runs each (interleaved) of the visual search and letter discrimination tasks while EEG recording was active.

In this way, both the shape identities and their relative position within the pair were completely fixed. In contrast, on any given trial, a shape assigned to the unstructured set could be paired with any other shape from the unstructured set and could occupy either the top or bottom position within this random pairing.

6.2.3 Procedure

Participants attended two sessions within 48 hours: an online pre-training session and an EEG recording session in the lab. During pre-training, participants completed the same online visual search experiment described in Chapter 5. In the second session, participants continued training with an additional six runs of the visual search task, completed while the experimenter fitted the EEG cap and electrodes. After initiating the EEG recording, participants underwent alternate runs of the visual search and letter discrimination tasks (details in the following sections), totaling 6 runs per task across the full experiment The overall structure of the study is shown in Fig. 6.1C.

6.2.4 Visual search task

The trial structure in the visual search task followed that described in Chapter 5. On each trial, participants saw a letter cue indicating which of the 4 memorized target shapes they had to search for in the upcoming display. After a brief delay, a search display with 10 shapes appeared (either structured or unstructured, see below) until participants indicated via a button press whether the target was present on the left or the right side of the display. The temporal details of the trial structure are described in Fig. 6.1A.

Each search display consisted of 4 distractor shape pairs, the target shape, and a foil shape (i.e., one of the other three target shapes not currently being searched for) arrayed symmetrically on a 4×4 grid, which subtended a visual area of $10.1^\circ \times 6.8^\circ$ on the visual field (Fig. 6.1A). All shape pairs within a given display belonged to the same scene condition (e.g., 4 fixed pairs from the structured set or else 4 randomly generated (on each trial) pairs from the unstructured set). Thus, on each trial, participants searched for the target in either a structured or unstructured scene. The 4 shape pairs were placed at random locations, one in each column of the grid. The locations were mirrored horizontally. The target appeared in one of the remaining locations, always immediately adjacent to a pair (see Fig. 6.1A). The foil (one of the other 3 targets) appeared at the horizontally mirrored location to the target. The location randomization process ensured that the probability of the target's location was uniform across the entire grid.

There were 16 structured and 16 unstructured trials in each run of the visual search experiment. These trial types were blocked during the training phases of both session 1 and session 2 and were fully interleaved during the test phases of each session. The order of blocking (structured first or unstructured first) was maintained for a participant throughout the experiment and counterbalanced across participants.

6.2.5 Letter discrimination task

One of our goals was to assess the influence of the exposure to the shape co-occurrences during the search on the representation of the shape pairs. To obtain the neural representations of the pairs from both scene conditions, we ran an orthogonal, central, letter-discrimination task as the pairs were flashed in the locations they would occur in the search displays.

In this task, observers were instructed to attend to a stream of centrally-presented letters (T or L, equal probability) that appeared one at a time for 100 ms with an ISI of 500 ms (±100ms random jitter). The letters were inverted with a $1/10$ probability. The task was to press the spacebar each time an inverted T appeared in the stream. While observers were engaged in this demanding central task, task-irrelevant shape stimuli appeared serially in the same (peripheral) grid locations of the visual search task (the Shape stream). All three stimulus types appeared with equal probability: the four targets, the four fixed pairs comprising the structured set, and four randomly generated shape pairs from the unstructured set (generated at the start of the experiment and maintained across letter discrimination task runs). The stimuli appeared for 200ms, with an ISI of 800ms (±200ms jitter, see Fig. 6.1B).

Each run of the letter discrimination task consisted of 24 trials corresponding to each of the 12 stimuli (the four targets, the four fixed pairs comprising the structured set, and four randomly generated shape pairs from the unstructured set) in the Shape stream. Each run was terminated upon the completion of the Shape stream.

6.2.6 EEG acquisition

We recorded scalp EEG with a 64-channel active electrode actiCAP system (500 Hz sample rate) with customized electrode positions adapted from the actiCAP 64Ch Standard-2 system (ground electrode placed at AFz; TP10 placed on right mastoid as a reference electrode). We verified that electrode offsets were $< 50\,k\Omega$ before recording commenced. Data were referenced online to the left mastoid and filtered between 0.016 and 125 Hz using BrainVision Recorder (BrainVision Recorder, Version 1.21.0402, Brain Products GmbH, Gilching, Germany). We recorded horizontal and vertical eye movements using external passive electrodes placed at the outer canthi of both eyes and immediately above and below the right eye. These external channels were referenced to a ground electrode placed on the nose and were visually monitored by the experimenter during recording to help provide feedback to the participant regarding eye movements/blinks. During the experiment, the experimental code sent triggers to the EEG recording to mark relevant stimulus onsets via a custom-developed BITSI button box.

6.2.7 EEG preprocessing

We preprocessed each participant's EEG trace in MatLab (2016b) using custom code adapted from the FieldTrip toolbox. For each subject individually, we applied a bandpass filter to the full EEG trace ($0.05 - 100$ Hz), and a line noise filter to remove electrical noise at 50, 100, and 150 Hz. We re-referenced all scalp channel data to the average of all scalp channels (excluding mastoids), before downsampling to 250 Hz for easier handling and storage. Finally, we segmented trial epochs from -100 ms to 600 ms around stimulus onsets of interest, producing 192 epochs corresponding to onsets of the visual search experiment stimuli, and 1728 epochs corresponding to onsets of peripheral stimuli (fixed pairs, random pairs, and singleton targets) during the letter discrimination task. Each of these epochs was baseline corrected by subtracting the average activity from -100 ms to 0 ms from the waveform.

For a subset of participants (S1-S12), a programming error led to not all triggers being correctly registered during the letter discrimination task, resulting in a degree of data loss for those participants (on average, 6.23% of the epochs for these subjects). The error did not affect the visual search epochs; there was no data loss for S13 onward.

6.2.8 ERP analysis

To assess the impact of the distractor structure on attentional orientation, we analyzed the N2pc component of the ERPs. 4 ERPs corresponding to the scene conditions (structured/unstructured) and target location (left/right) were constructed for each participant for channels PO7 and PO8 (Gaspelin and Luck, 2018). For each target location, the ERP at the ipsilateral channel was subtracted from the ERP at the contralateral channel. These difference waveforms were then averaged. The N2pc window was defined as the period between 200ms to 400ms, following Gaspelin and Luck (2018). The N2pc negativity in this period was compared between the two scene conditions.

6.2.9 Decoding analysis

To assess the existence of various factors of interest in neural representations (e.g., location and shape of the target), we ran multiple decoding analyses employing linear discriminant analysis (LDA) classifiers as implemented in MatLab (2017a) with default parameters. Details about the

classifiers used are mentioned below. For each analysis, the epochs corresponding to -100 ms to 600 ms from the stimulus onset (in the Shape stream) were considered.

First, we assessed if the target was processed better in the structured scenes, given the increased search efficiency, which could be reflected as the information about both the location and shape of the target being higher for the trials with the structured scenes. For the visual search task, we trained a classifier on trials labeled with the target location (left/right) or target shape $(1-4)$. For each participant, separately for the trials with the structured and unstructured scenes, and separately for the classification type (location or shape), we implemented 6-fold cross-validation (average accuracy considered) using all the trials across the 6 runs (all scalp channels included), at each time point.

Second, we assessed the decodability of individual shapes and the cross-decodability of target shapes in the visual search task using the singleton target presentations in the letter discrimination task. Classifiers were trained to classify the shape of the targets presented as singletons in the letter discrimination experiment. 6-fold cross-validation (average accuracy considered) was performed using all the relevant trials (target shape stimuli) across the 6 runs, for each participant, across all scalp channels, at each time point. To assess if the shape discriminant patterns in the letter discrimination experiment generalized to the visual search experiment, classifiers were trained to classify the shape of the targets presented as singletons in the letter discrimination experiment and tested on the target shapes present in the visual search experiment, separately for the scenes.

Third, we assessed if the shape pairs from the structured scenes could be distinguished from the shape pairs from the unstructured scenes, disregarding the visual feature differences between the pairs. To do so, classifiers were trained to classify which set (structured or unstructured) each of the 8 shape pairs, presented in the letter discrimination experiment, belonged to. 4-fold cross-validation (average accuracy considered) was performed. In each fold, all the trials corresponding to 3 of the pairs from each set were used to train the classifier which was then tested on the 2 left-out pairs, for each participant, across all scalp channels, at each time point. This leave-one-set-of-pairs-out approach ensured that the decoding accuracy did not reflect simple visual feature differences between the pairs but any differences specific to those pairs belonging to distinct classes of neural representations - driven by exposure to the shape co-occurrences in this case.

Lastly, we assessed if the shape pairs from the structured scenes were more distinct from each other than the shape pairs from the unstructured scenes were. Classifiers were trained to classify between the 4 pairs corresponding to a set (structured or unstructured). 6-fold cross-validation (average accuracy considered) was performed using all the trials across the 6 runs, for each participant, across all scalp channels, at each time point, separately for the sets.

6.2.10 Statistical analysis

Threshold-free cluster enhancement (TFCE; Smith and Nichols (2009)) with a permutation test was used to correct for multiple comparisons, across the time points, of the contrasts relative to the corresponding baselines (at $p < 0.05$). In each figure, the time points where the contrasts survive TFCE correction are marked with asterisks.

6.3 Results

6.3.1 Behavioral results

We began by inspecting the test phase data for session 1, which in theory should replicate the observations of Chapter 5 (greater inverse efficiency (IES = average reaction times for the cor-

rect responses / average accuracy) for search in the structured scenes compared to unstructured scenes). However, we found no evidence for such a structure-benefit in IES between the scene conditions in session 1 (paired t-test: $t_{30} = 1.0$, $p = 0.31$). Similarly, there was also no evidence for this difference in the EEG experiment (in session 2; paired t-test: $t_{30} = 0.7$, $p = 0.49$; both sessions combined (IES averaged): $t_{30} = 1.04$, $p = 0.31$). Thus, the results from Chapter 5 did not replicate in this sample of participants. However, as discussed earlier, this replication was apriori unexpected given the sample size here (although the average difference in IES was comparable between session 1 (12.8 ms) and Chapter 5 (10.97 ms)). The EEG signals might provide us with stronger signatures of the influence of the structure in the scenes on the attentional processes and the neural representations of the shapes.

6.3.2 The impact of distractor structure on attentional orienting

To assess the influence of structured distractors on the orientation of attention in search, we examined how the N2pc ERP component in posterior electrodes (PO7 and PO8) varied as a function of scene condition. Fig. 6.2A shows the contralateral and ipsilateral ERPs, averaged across target locations (left/right) and scene conditions (structured/unstructured). Notably, contralateral and ipsilateral ERPs diverged around 200 ms following the onset of the search display (which is the onset of the N2pc window - $200 - 400$ ms; Gaspelin and Luck (2018)). We then compared the difference waveforms (contra-ipsi) for structured and unstructured scenes (Fig. 6.2B), however, there was no significant difference within the N2pc window (averaged within the window; paired t-test: $t_{30} = 0.2$, $p = 0.8$) nor at any other time point. Thus, it appeared there was no difference in the orientation of attention observed between the scene conditions.

Although scene conditions did not, on average, appear to modulate either IES or N2pc magnitude, it could still be the case that these two metrics are related to each other at the individual participant level. Here we considered whether the structure-effect in search performance was correlated with our neural measure of attentional allocation (i.e., N2pc) Averaged across the scene conditions, the IES was indeed correlated with the N2pc negativity ($r = -0.56$, $p = 0.001$), such that participants with greater N2pc negativity (higher degree of attentional orientation) tended to also have higher search efficiency. The structure-benefit in IES was also significantly correlated with the difference in N2pc negativity between the scene conditions, across participants ($r = 0.36$, $p = 0.048$; Fig. 6.2C). In sum, not only did the N2pc negativity index overall search efficiency, but also the differences in search efficiencies across scene conditions. Whenever the distractor structure was accompanied by increased or decreased search efficiency, it was also accompanied by a corresponding increase or decrease in the degree of the orientation of attention (as indexed by N2pc) towards the target.

As the overall IES was related to the overall N2pc negativity, the correlation between the differences in these two measures across conditions might not be specific to the condition-specific splitting of trial data - the differences in IES or N2pc negativity in a random split of the data might also reveal a correlation. To rule out this possibility, we asked how many of these random splits would produce a correlation coefficient larger than the relevant, condition (structure) specific split described above. To this end, we generated $10,000$ random splits of the data and compared the difference across the splits for IES and N2pc negativity. Only 3.9% of the random splits yielded correlations higher than the condition-specific split. We can conclude that the observed correlation across participants between the differences between the scene conditions for IES and N2pc negativity was unlikely to be found due to it being a random split of the data (in traditional statistical terms, the control analysis resulted in $p = 0.039$ for the hypothesis that a random split could generate the observed correlation).

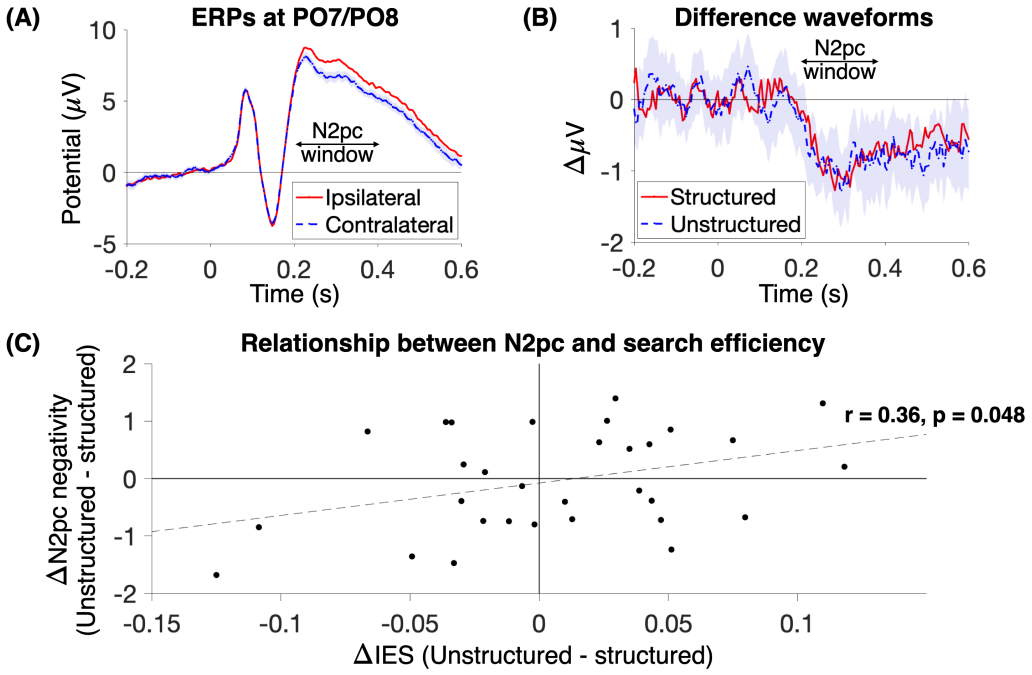

Figure 6.2: The impact of distractor co-occurrences on attentional orienting. (A) The event-related potentials (ERPs) collapsed across the visual search trials when the targets appeared in the left or right sides of the display are shown for the corresponding ipsi- and contra-lateral electrodes. The N2pc window is defined as the period between 200 and 400 ms (Gaspelin and Luck, 2018). The blue envelope shows the 95% confidence intervals for the difference between the two conditions. (B) The difference waveforms (contralateral - ipsilateral) are shown for the trials with the structured and unstructured scenes. No overall difference was found between the two conditions in the N2pc window. The blue envelope shows the 95% confidence intervals for the difference between the two conditions. (C) Although an overall structure-benefit effect was not found in the N2pc negativity or the inverse efficiency scores (IES), the two effects were correlated across participants.

6.3.3 The impact of distractor structure on the target's neural representation

To assess the impact of distractor structure on the extraction of information about the target for efficient search, we ran separate decoding analyses for target location (left/right) and specific target shape ($1-4$) for the visual search data (all scalp channels included, we report 6-fold cross-validation accuracies and TFCE-corrected statistics; see Methods and Materials). Target location could be decoded from around 300 ms onward (Fig. 3A), for both structured and unstructured scene conditions There was no difference in location decoding between the two scene conditions at any time point. No evidence was found for a correlation across participants between the condition-specific difference between location decoding (averaged across 300 ms to 600 ms - the window in which target location could be decoded) and the difference in IES ($r = 0.28$, $p = 0.13$). In contrast to the fairly robust location decoding we observed (Fig 3A), the target shape could only be decoded weakly at certain time points post-200 ms (Fig. 3B). There was no differ-

ence in shape decoding between the two scene conditions at any time point. Again, no evidence was found for a correlation across participants between the condition-specific difference between shape decoding (averaged across 200 ms to 600 ms - the window in which target shape could be decoded) and the difference in IES ($r = -0.16$, $p = 0.4$). Thus, although the visual search displays evoked neural responses that contained information about both the target's location and its identity, no evidence was found for differential neural representations of the target attributable to distractor structure.

Maybe the previous analysis of target shape decoding failed to reveal strong effects as the EEG signal was not fine-grained enough to be able to decode the abstract shapes used in the study. To assess if shape information was decodable at all, given the abstract shapes used in the study, we ran a separate analysis in which we decoded the shape of the targets when presented as singletons in the letter discrimination experiment. Target shape was indeed decodable at time points between 278 ms and 344 ms (Fig. 3C). Additionally, although the target shape was decodable in the trials of the letter discrimination experiment and weakly decodable in the trials of the visual search experiment, no evidence was found for cross-decoding (classifiers trained on the trials of the letter-discrimination experiment and tested on the trials of the visual search experiment) of target shape (Fig. 3D). Although the target shape could be decoded when the shapes were presented as singletons, the decoding accuracy was weak (although it was a more robust effect than the shape decoding during the search trials). These results suggest that the EEG signal was indeed not fine-grained enough to be able to decode the abstract shapes used in the study.

6.3.4 The impact of the exposure to distractor structure on the neural representations of the shape pairs

To assess how the representations of shape pairs belonging to the structured set might have changed due to exposure during the visual search experiments, we decoded which set - structured or unstructured - the eight pairs belonged to (with a leave-one-set-of-pairs-out approach, see Methods and Materials), and we also decoded the identity of the pairs separately within the structured and unstructured sets, using the trials from the letter discrimination task. The pairs from the structured set could not be distinguished from the pairs from the unstructured set at any time point (Fig. 6.4A). Averaging the accuracies across the two sets, the four pairs could be distinguished from each other around 232 ms (Fig. 6.4B). No evidence was found for a difference between the decoding accuracies for the two sets of pairs. Additionally, no evidence was found for a correlation across participants between the condition-specific difference between pair identity decoding (at 232 ms - the timepoint at which pair identity could be decoded) and the difference in IES ($r = -0.16$, $p = 0.37$). In sum, no evidence was found for a change in the representations of the pairs from the structured scenes as compared to those from the unstructured scenes. This could, again, be a result of the EEG signal not being fine-grained enough to reveal the differences in the representations of the abstract shapes used.

6.4 Discussion

In Chapter 5, we observed that participants could search more efficiently for target shapes in scenes containing distractor shapes that co-occurred in pairs. The neural processes underlying this observation were unclear. In this study, we found that while participants engaged in visual search amongst those co-occurring distractors, EEG recordings revealed that the increased efficiency attributed to distractor co-occurrences in the scenes (termed structure-benefit) was accompanied by a higher degree of attentional orienting towards the target, indexed by the N2pc component

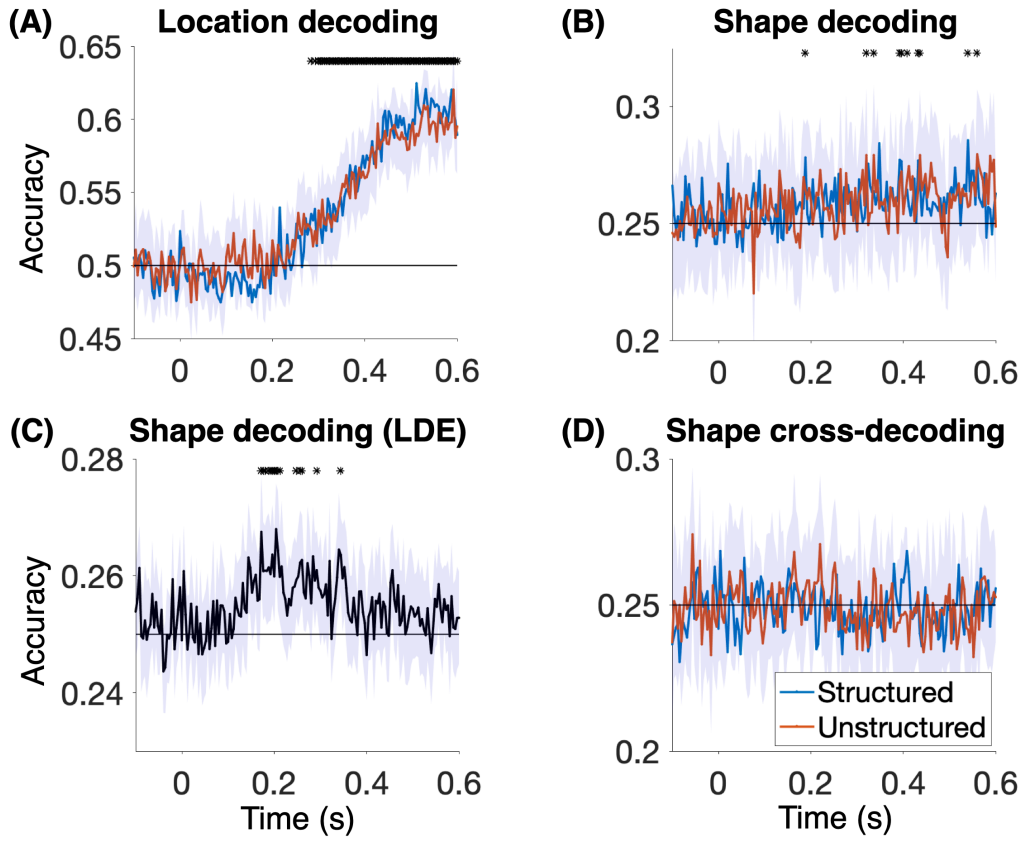

Figure 6.3: The impact of distractor co-occurrences on the neural representation of the target. (A) Target location was linearly decodable from around 300 ms onwards, with no difference in the decoding profiles across the two scene conditions. The blue envelope shows the 95% confidence intervals for the difference between the two conditions. (B) Target shape was weakly decodable at some timepoints post-200 ms, again with no difference in the decoding profiles across the two scene conditions. The blue envelope shows the 95% confidence intervals for the difference between the two conditions. (C) When presented as singletons in the letter discrimination task, the shape of the targets was decodable between 278 and 344 ms. The blue envelope shows the 95% confidence intervals for the accuracies. (D) The target shapes in the visual search trials could not be cross-decoded using the shape-elicited patterns in the letter discrimination task, irrespective of the scene conditions those trials belonged to. In summary, we found no evidence that the distractor co-occurrences were associated with differential processing of the target shape and location. The blue envelope shows the 95% confidence intervals for the difference between the two conditions. In all the panels, asterisks indicate the time points where the measures indicated on the y-axes differed from the corresponding baselines, as gauged via Threshold free cluster enhancement (TFCE).

of the ERPs. However, no evidence was found for any changes in neural representations of the distractor pairs attributable to the exposure to the co-occurrences during the search experiment.

Although no overall structure-benefit was found across participants, a weak association (re-

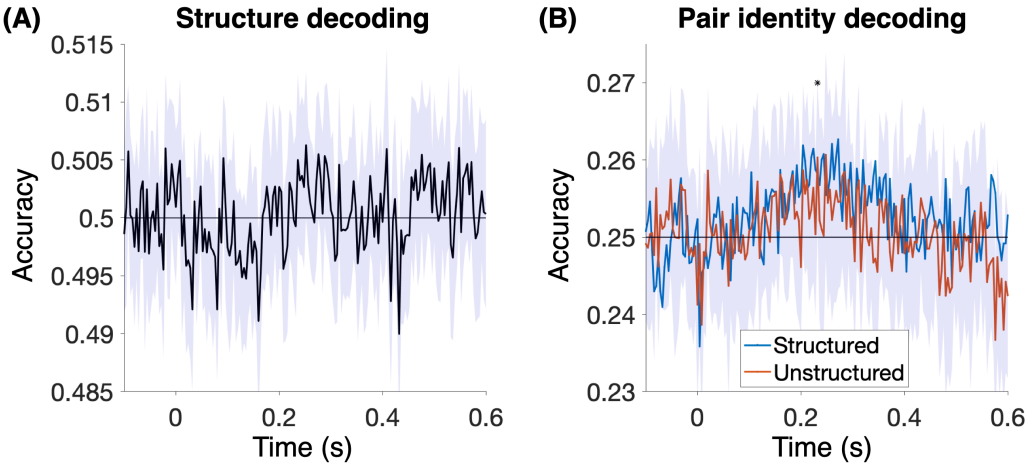

Figure 6.4: The impact of the exposure to distractor co-occurrences on the neural representations of shape pairs. (A) Given 3 pairs each from the structured and unstructured scenes, the set to which the left out pairs belonged could not be decoded at any timepoint. The two sets of shape pairs could be distinguished based on which scene condition they belonged to. (B) The identities of the 4 pairs within the structured and unstructured sets could be decoded around 232 ms, but there was no difference in the decodability across the conditions. In summary, we found no evidence that the distractor co-occurrences were associated with differential processing of the distractor shapes. In both panels, the blue envelope shows the 95% confidence intervals for the difference between the two conditions. The asterisks indicate the time points where the measures indicated on the y-axes differed from the corresponding baselines, as gauged via Threshold free cluster enhancement (TFCE).

quiring replication) was found between structure-benefit and attentional orientation. Attentional orienting was indexed by the N2pc component. A similar relationship between the benefit of target-distractor regularities to search efficiency and attentional orienting has been observed in contextual cueing. Studies in the contextual cueing literature (Johnson et al., 2007; Schankin and Schubö, 2009; Sisk et al., 2019) have used N2pc to assess if repeating the search displays (where participants can learn the target-distractor co-occurrences) lead to an increased attentional orientation towards the target which could lead to the primarily observed reduction in the reaction time for searching the target. In Schankin and Schubö (2009), participants who showed a higher contextual cueing effect also showed a higher difference in the N2pc amplitudes between the repeated and the non-repeated displays. Mirroring their findings, in this study, participants with higher structure-benefit showed higher attentional orientation towards the target in the structured scenes. It is important to note that there is a difference between contextual cueing and our setup: the distractor arrangement can predict the target location in contextual cueing, whereas, in our setup, the distractor arrangement does not predict the target location but can only lead to processes such as object grouping (Kaiser et al., 2019; Lengyel et al., 2021) or inter-object priming (Stein et al., 2015) which could reduce the complexity of the scene paving the way to efficient detection of the target.

The observed association between the attentional orientation and structure-benefit suggests that the overall structure-benefit observed in Chapter 5, across a large sample of participants, could be associated with an increased attentional orientation towards the target driven by the in-

formation about the distractor co-occurrences. The speed of attentional orientation - 200-400ms after display onset (indexed by the N2pc component) - places constraints on the possible accounts of how the co-occurrences might have been used towards a more efficient search. In Chapter 5, it was proposed that on some trials when a participant rejected a distractor shape as the possible target, knowledge about the shape co-occurrences could aid the participant in rejecting that shape's partner faster (inter-object priming, see Stein et al. (2015)), thereby arriving at the target faster. Such a serial search cannot account for the speed of attentional orientation towards the target observed here, as the N2pc component is associated with the first cue-driven voluntary spatial attention shift post display onset (Luck, 2012). An alternate account suggested that such inter-object priming could instead happen in parallel throughout the visual field, thereby affecting the priority map by suppressing the activity in the distractor locations, influencing the orientation of attention around the N2pc window (Chaumon et al., 2008; Zinchenko et al., 2020). The results described here provide support for that alternate account.

Although participants' attentional orientation was associated with distractor structure in the scenes, we did not find any evidence for structure-related differences in the representations of the distractor pairs. The overall decoding performance was weak, both in decoding the singleton targets and the pair identities within the two sets of pairs (structured or unstructured). The differences across the two sets of pairs were probably weaker compared to the overall classifiability, and therefore not measurable using this design and/or measurement technique. Previous studies have shown that neural responses elicited by multiple objects could be classified using EEG (Wang et al., 2012; Kaneshiro et al., 2015; Cichy and Pantazis, 2017). However, most of these studies used natural object exemplars as opposed to the abstract shapes used in this study which might be challenging to classify using EEG. Additionally, as the question of interest is primarily about the differences between the neural representation of shapes and not about the timecourse of the difference, other neuroimaging techniques with higher spatial resolution such as fMRI could be better suited here (e.g., Kaiser and Peelen (2018)).

To conclude, our results suggest that distractor co-occurrences could be utilized to direct top-down attention towards the target rapidly ($200 - 400$ ms after display onset), constraining the possible mechanisms underlying the usefulness of the co-occurrences in search observed in Chapter 5. Further research using high spatial-resolution techniques such as fMRI could shed light on the changes in the neural representations of shapes due to their co-occurrences.

Chapter 7

General Discussion

The intricacies of how humans search for objects in structured environments have been under investigation, through the lens of cognitive neuroscience, for half a century. Through numerous studies, much has been revealed about the fundamental aspects of search such as the neural basis of object recognition, the capability to search for visual features (that constitute objects) in parallel across the visual field, and the reliance of search on the regularities in the environment, as discussed in Chapter 1. However, there are several outstanding issues about the nature and neural basis of these fundamental aspects. In the previous chapters, using neuroimaging techniques (fMRI and EEG), large-sample behavioral experiments, and artificial neural networks, I presented observations that help us shed new light on those issues. In each of the chapters, the implications of the presented findings were discussed. These observations and implications are briefly summarized:

- In Chapter 2, we observed that the neural (fMRI) response to bodies presented in task-irrelevant locations depended on the search target - bodies or other objects - in the task-relevant locations. This observation supports the idea that feature-based attention can be deployed at higher stages of the visual hierarchy, where the features are diagnostic of object categories. This statement seems to be valid in the case of body shape, but it remains unclear if other objects (e.g., cars) can similarly avail of feature-based attention.
- In both Chapters 3 and 4, we observed that the advantage of deploying feature-based modulations in an earlier layer of an artificial neural network in addition to deploying them in the later layer manifests when the network has a low representational capacity. These observations constitute a formal demonstration of the presumed notion in cognitive science that the attentional modulation of early visual processing is essential when the visual system is capacity-limited. Representational capacity was operationalized differently in the two chapters.
  - In Chapter 3, we additionally found that the modulations trained for the earlier layer did not resemble those predicted by the feature-similarity gain model (FSGM) of feature-based attention. This observation agrees with other reports suggesting FSGM might not be optimal nor might it be used in the visual cortex for all situations.
- In Chapter 5, we observed that the search for a target shape is faster, and more accurate, amongst distractor shapes that co-occur in pairs than amongst distractor shapes that do not co-occur in pairs. This observation supports the theory that humans can exploit the regularities amongst distractors to parse the scene better and make the search easier.

- In Chapter 6, we observed that the participants showing increased search efficiency in the scenes with co-occurring distractor pairs also showed an increased attentional orienting (indexed by the N2pc component of the EEG waveforms) towards the target in those scenes. This observation suggests that the benefit derived through the co-occurring distractors could manifest in the attentional processes during the search.

In this chapter, in light of all our findings, I will discuss some of the overarching frameworks and models of the fundamental aspects of visual search.

7.1 Neural modulations due to feature-based attention

Humans can constrain their search to objects that contain a target feature, in parallel across the visual field, via a mechanism termed feature-based attention. How exactly does the information about the target feature interact with the incoming information from the eyes in the visual cortex? The feature-similarity gain model (FSGM; Treue and Trujillo (1999)) proposes that the activity of neurons is multiplicatively upregulated or downregulated depending on whether and how much the neurons prefer the target feature. This modulation has been posited to increase the relevance of locations where the target feature is most likely to be present in the priority map, which can lead to the efficient deployment of spatial attention and eye movements (Bichot et al., 2005). This modulation has been observed mostly for low and mid-level features (such as orientations and curvatures), but scarcely for higher-level features diagnostic of object category.

Conforming with FSGM, in Chapter 2, using fMRI, we observed the spatially-global modulation of human body silhouettes in the voxels in high-level visual cortex that were body-selective. Additionally, an analysis of the response patterns across voxels in the object-selective cortex revealed another signature of feature-based attention. When bodies were the targets of search, the response patterns to bodies in the task-irrelevant locations became more similar to the prototypical response patterns to bodies than the prototypical response patterns to the other objects. The prototypical response patterns were recorded from a separate fMRI run where participants were attending to the task-irrelevant locations and the images were not masked, giving us an estimate of the baseline response to the categories. Such signatures, at the level of response patterns, of feature-based attention have been reported in previous studies that used multivariate pattern analysis (Peelen et al., 2009; Jehee et al., 2011). We additionally found that this modulation of response patterns was not solely driven by body-selective voxels. Can FSGM accommodate these observations?

According to FSGM, the modulation of neurons is proportional to their selectivities to the target feature. A weak pattern elicited by bodies, which could be the result of the masking, presence of other objects, or lack of spatial attention, will get enhanced when FSGM is deployed with bodies as the target feature: voxels that are more selective to bodies will have their response boosted more thus making the response pattern more similar to that evoked by bodies when presented in isolation, without masking and reduction of spatial attention. Hence, FSGM can accommodate the attentional modulations observed at the level of response patterns. The added benefit of analyzing the response patterns, over overall activity across voxels, is that it capitalizes on the graded modulation posited by FSGM. In summary, FSGM can accommodate the observations from Chapter 2 regarding the signatures of spatially-global feature-based attention for bodies.

Despite its usefulness in predicting the influence feature-based attention has on neural activity, FSGM is not the optimal modulation scheme in every scenario. The modulation scheme leading to an optimal search for the target takes the distractors into account (Navalpakkam and Itti, 2007; Scolari et al., 2012). If the task is to find a grating oriented at 55° degrees amongst other gratings

oriented at $50°$, it is optimal to upregulate the activity of neurons supposing that the target grating is instead oriented at $60°$, to maximize the difference between the responses to the target and the distractors. Lindsay and Miller (2018) also showed that when the target is a category, FSGM is not the best feature-based attention scheme (deployed across an artificial neural network trained for object recognition). Their alternate scheme was computed using the gradient of the task performance (target detection) accounting for each network layer's activity. However, the gradient direction does not necessarily stay the same if we continue such a procedure iteratively to maximize the task performance - required to obtain the optimal modulation scheme. In Chapter 3, in an artificial neural network, we trained the top-down attentional modulation to maximize the task performance using backpropagation iteratively. In agreement with Lindsay and Miller's results, we found that the trained modulation scheme outperformed FSGM. These results suggest that while searching for objects, the optimal modulation of neurons in the early visual layers might not be FSGM.

In the case of the trained modulation scheme, it is hard to understand the nature of the modulations (unlike the initial example where the target template changed to $60°$). The mapping from categories to low and mid-level features is non-linear and hard to interpret in terms of the tuning curves for category-diagnostic features at those layers (Zeiler and Fergus, 2014). Additionally, neurons in the early or mid-level layers which might be highly selective to one category might not be essential for neurons at the output that indicate the presence of that category (Morcos et al., 2018; Zhou et al., 2018). At this stage, all we can state is the optimal scheme modulates the neurons to maximally differentiate the target from the distractors at the final layer of the network. Further research is required to understand if we can better interpret the optimal modulation scheme in terms of the activity of the neurons across the network. If we cannot, that could mean that similar attentional modulations in the primate visual system would be uninterpretable too (the same way neuronal tuning curves are mostly uninterpretable in the mid-level visual cortex - Bashivan et al. (2019); Richards et al. (2019)). In Chapter 2, the existence of attentional modulations in the visual cortex was gauged by assuming FSGM. If FSGM is not the modulation scheme in use in the early and mid-layers of the visual cortex, it is unclear what other analyses could reveal any existing attentional modulations.

In summary, FSGM is an experimentally-verified model of the modulations underlying feature-based attention when the modulations are deployed at the stage of visual processing at which the target feature is expressed (i.e. where the neural tuning curves exist for that feature; e.g., orientations in the early visual cortex, bodies in high-level visual cortex). However, if the modulations are deployed at an earlier stage when the target feature is at a higher level (e.g., categories), or when the distractors are too similar to the target, then FSGM might not be the optimal modulation scheme. How we could characterize the alternate optimal scheme in terms of the visual features being modulated remains to be seen.

7.2 The deployment of feature-based attention

Consider a multi-layered feedforward neural network as a model of visual processing, akin to the ventral visual stream. Suppose the network's task is to perform cued object detection (e.g., "is there a car in the image?"). The cue can be conveyed to the network with top-down signals (corresponding to feature-based attention or spatial attention depending on the nature of the cue; here we only consider feature-based attention). Where in the network is this interaction between bottom-up and top-down signals essential? The simplest solution is to only feed it at the end of the network, assuming that the network has enough representational capacity, to allow for all possible 1-vs-all classifications (e.g., corresponding to "car present" or "car absent") across the

objects of interest. What if it does not have sufficient capacity? Then the top-down signals need to be communicated to earlier layers to constrain the information flow from there such that the downstream layers receive more information necessary for the 1-vs-all classification required by the cue.

The dependence of where these top-down modulations need to be deployed on the capacity of the network was studied in Chapters 3 and 4. It was observed that when the representational capacity of a network is low, either due to the network not having enough neurons or there being too many tasks to solve, additionally deploying top-down modulations in the earlier layer provided performance gains over the deployment of those modulations solely in the final layer. These results provided a normative modeling demonstration of the idea that early task-based selection of information is essential in capacity-limited situations of biological visual systems (Lavie and Tsal, 1994). To the best of my knowledge, no other studies have explicitly, computationally, gauged this relationship, although some studies have shown that deploying attention in early layers of deep neural networks helps in downstream tasks (Lindsay and Miller, 2018; Rosenfeld et al., 2018; Luo et al., 2021). Our investigations also led to other questions about how the organization of the biological visual processing stream and its interaction with top-down modulations emerges during development, as discussed below.

One of the main differences between Chapters 3 and 4 was about the way the networks were trained. In Chapter 3, the object processing stream (the feedforward sweep) was first trained on object recognition and then the top-down cue-based modulations were trained on it. In Chapter 4, both the object-processing stream and the top-down modulations were jointly trained. While the training scheme in Chapter 4 is optimal, as the object-processing stream performs transformations on the input that can be optimally switched to the transformations required by the detection task dictated by the cue, it requires changes to both the object-processing stream and top-down modulations whenever new cues are to be learned. How are the object-processing stream and top-down modulations in the human brain trained?

From a developmental standpoint, it has been suggested that the object-processing stream and top-down modulation are trained jointly, akin to the suggestion in Chapter 4 (Amso and Scerif, 2015). As the object-processing stream becomes better at representing more details, the top-down modulations can get better at selecting the relevant details to be sent downstream. In adult participants, it has been suggested that repeated exposure to attention tasks can train the top-down modulations to switch the state of the object-processing stream, akin to the suggestion in Chapter 3 (Bartolucci and Smith, 2011; Gilbert and Li, 2013; Harel et al., 2014). However, it has also been suggested that the changes might be reflected in the object-processing stream itself (termed perceptual learning), with no influence on the top-down modulations (Frank et al., 2014b; Reavis et al., 2016). Further research is essential to understand in what situations changes are made to the object-processing stream as opposed to or in concert with the top-down modulations, as the primate visual system is trained on a task or during development.

The issue about where top-down modulations are deployed in the primate brain is further complicated by the considerations of lateral and feedback connections in the visual system. These connections establish recurrent information flow in the visual stream which could lead to phenomena such as figure-ground segmentation, perceptual grouping, and surface-based segmentation, that could lead to better representations of the objects themselves, presumably without any top-down guidance from other brain regions, as demonstrated by in-vivo and in-silico modeling studies (Lamme and Roelfsema, 2000; O'Reilly et al., 2013; Wyatte et al., 2014; Kar et al., 2019; Kietzmann et al., 2019; Linsley et al., 2020; Thorat et al., 2021). In such a recurrent visual stream, top-down signals from other brain areas incident on later stages can get propagated to earlier layers. Indeed, a backward progression (in terms of latency) of attentional effects has been observed

in the primate visual system (Buffalo et al., 2010). Whether this progression reflected modulations conveyed via feedback within the visual stream or long-range top-down connections to the earlier layers remains to be resolved. In any case, given the ubiquitous nature of the recurrent information flow, it is critical to account for its existence in any account of top-down modulation of the information processing in the visual system.

In summary, the deployment of cue-driven attention in earlier stages of visual processing is essential when the representational capacity of the downstream processing is limited. This limitation could arise due to there being too few neurons given the amount and difficulty of the tasks the system has to perform. Questions such as how the interaction between the visual stream and such cue-driven top-down modulation works in the primate brain, and how the interaction accounts for the other facets of the visual stream such as the existence of lateral and feedback connections, remain largely unanswered.

7.3 The influence of the regularities in scenes on visual search

In searching for a target object, both the target-distractor co-occurrences and the distractor-distractor co-occurrences could play a role. The spatial predictability - distractor arrangements and identities predicting the location of the target drive spatial attention to the target location (Biederman et al., 1973; Bar, 2004; Zhang et al., 2020). For example, we would search for a big object if told to search for a car outdoors whereas we would search for a small object if told to search for a car in the living room. The influence of target-distractor co-occurrences on visual search has been well-studied. On the other hand, the study of the influence of distractor-distractor co-occurrences on search is in its infancy.

Most of the discussion about the distractor-distractor co-occurrences focuses on the reduction of the complexity of the scene due to these co-occurrences (Wertheimer, 1923; Brady et al., 2011; Kaiser et al., 2019; Lengyel et al., 2021). For example, co-occurring objects in a fixed arrangement (e.g., egg on egg cup) can be treated as one composite object, leading to a reduction of the information needed to characterize the scene. It has been proposed that such complexity reduction is similar to a reduction of the number of distractors and could therefore make the search for the target more efficient (Kaiser et al., 2014, 2019). In Chapter 5, using a controlled behavioral experiment, we observed that participants could indeed search faster and more accurately in scenes that had co-occurring distractor shapes than in scenes that did not contain such co-occurring distractors.

In Chapter 6, using EEG, we observed, across participants, that the higher this structure-benefit (increased search efficiency due to the co-occurring distractor shapes) the higher the attentional orientation response towards the target (indexed by the N2pc component of the event-related potential, $200 - 400$ ms after scene onset; Luck (2012)). Thus, the increased speed and accuracy of search in the structured scenes could be associated with the first voluntary spatial attention shift driven by the cue (Liu et al., 2007b). These observations suggest that the visual system registered these co-occurrences and used it to reduce the competition to the target leading to a stronger bias in the attentional shift towards the target location as opposed to the other side of the display, which could have led to an advantage in search efficiency in the structured scene.

How could this reduction in competition from the distractors have occurred? First, the structure-benefit was found independent of whether the relative positions of shapes within the co-occurring pairs were fixed (A and B always occur next to each other, and A could only be on top of B) or free (A and B always occur next to each other, either could be on top of the other). To explain the findings for the second case where the relative positions are free, the previously discussed - object grouping leading to the creation of a unified object leading to a reduced number of dis-

tractors - account is not ideal. Instead, inter-object priming mediated by the lateral and feedback connections in the visual system might provide the following solution: The partial recognition of a shape is communicated to the neighboring receptive fields providing a cue to disambiguate the partial evidence for the shape present in that receptive field due to the observed co-occurrences. However, instead of enhancing the response of the co-occurring shape, as is the case in inter-object priming observations (Stein et al., 2015), the response is suppressed as a signature of the shape being a distractor is found. This process could occur, in parallel, across the entire visual field, suppressing the responses in visual field locations that confirm the co-occurrence relationship with their neighboring locations, leaving the target location with a higher response, thereby biasing spatial attention towards that location.

Given the speed of influence on the attentional orientation, these co-occurrence driven, suppressive, inter-object priming effects might have occurred within the visual cortex without the guidance from the other brain regions implicated in statistical learning of co-occurrences (such as the caudate and the hippocampus; Turk-Browne et al. (2009); Covington et al. (2018)). In Chapter 5, we also found that participants could not indicate which of the distractor shapes co-occurred during the search experiment. This could be a result of the co-occurrence registration being weak which also gets reflected as a minute advantage in the efficiency due to the structure in the scenes. Further research is required to assess the validity of this account where co-occurrences between distractor shapes, relatively fixed or not, can be learned and influence the recurrent information flow in the visual cortex to suppress the locations where the co-occurring shapes could exist, thus reducing the competition to the target location.

In summary, in addition to the usefulness of the distractors predicting the identity and location of the target, co-occurrences amongst the distractors can reduce the competition from the distractors to the target, leading to an efficient search. This reduction in competition could be a result of the co-occurring distractors suppressing each other's response in the visual system. How such co-occurrences amongst distractors could be learned and exactly how they could be used to reduce the competition to the target to optimize search remains to be seen.

7.4 Conclusion

Consider the task of searching for an object in a scene. The knowledge about the relationships between the target and the distractors could help us select relevant information from the scene: spatial selection of information driven by the distractors predicting the target's location, and the selection of feature dimensions along which the target could be best discriminated from the distractors. Additionally, the knowledge about the relationships amongst the distractors could help us group the co-occurring distractors and make the scene easier to search through. These relationships exist and could be used to drive the selection processes underlying search, but are they used by humans during visual search? All the relationships mentioned above are associated with human visual search (Carrasco, 2011; Wolfe, 2021; Kaiser et al., 2019). They are the fundamental aspects of the information selection process in the state-of-art theory of human visual search (Guided Search 6.0; Wolfe (2021)), and in this thesis, I presented our evidence in agreement with this theory.

The support for this theory mostly comes from experiments where the participant is asked to sit still in a chair (or lay down in an MRI scanner), not move their head, and shown a 2D picture of a scene after being told what object to rapidly search for. It is unclear if the elements of the theory, borne out of observations under these constraints, would generalize to real-world visual search (Carrasco, 2011; Wolfe, 2021). Contrary to the rapid nature of search in our experiments, real-world search spans multiple timescales: finding the toothbrush in the morning (< 1 s), finding

misplaced keys ($\sim$ 1min), searching bags at the airport for dangerous items ($\sim$ 5 mins), and searching for lost sailors at sea (hours or days). It has been suggested that short searches might not require as much strategy or planning as longer searches could - "it seems to be faster to let covert attention bounce around in an anarchic manner than to bring it under strict control" (Wolfe et al., 2000; Wolfe, 2021). Additionally, in the real world, we search for multiple objects (e.g., airport baggage screening) and execute sequential searches as a part of navigation (e.g., walking to the supermarket, crossing roads, while monitoring traffic lights, and looking out for bicycles and cars). In these cases, it is unclear to what extent the other non-search processes, such as deciding the objects of relevance and the sequence of searches, can be dissociated from the search processes (Gottlieb and Oudeyer, 2018; Wolfe, 2021). In summary, it remains to be seen how far the insights gleaned from our laboratory search experiments generalize to real-world search.

Structured investigations into visual search began recently. Compared to the long history of the characterization of physical laws by Galileo Galilei, his successors, spanning at least half a millennia, the first reports of characterizing how humans search can be traced to the 1950s (Koopman, 1956a,b; Nakayama and Martini, 2011). In the subsequent decades, researchers shed light on many aspects of search, leading to an understanding of the true complexity of a process that seems effortless to us. While we have come a long way in our understanding of the procedural and neural underpinnings of search, bridging the gap between what humans *can* do in visual search, as revealed by our experiments, to what humans *actually* do during a real-world search, and possibly going beyond to posit what any artificial agent *should* do for optimal visual search, are the next steps to be taken.

Bibliography

[Abdelhack and Kamitani 2018] ABDELHACK, Mohamed ; KAMITANI, Yukiyasu: Sharpening of Hierarchical Visual Feature Representations of Blurred Images. In: *eNeuro* 5 (2018), Nr. 3, P. ENEURO–0443

[Agrawal et al. 2017] AGRAWAL, Aishwarya ; LU, Jiasen ; ANTOL, Stanislaw ; MITCHELL, Margaret ; ZITNICK, C L. ; PARIKH, Devi ; BATRA, Dhruv: Vqa: Visual question answering. In: *International Journal of Computer Vision* 123 (2017), Nr. 1, P. 4–31

[Amso and Scerif 2015] AMSO, Dima ; SCERIF, Gaia: The attentive brain: insights from developmental cognitive neuroscience. In: *Nature Reviews Neuroscience* 16 (2015), Nr. 10, P. 606–619

[Andersen et al. 2013] ANDERSEN, Søren K ; HILLYARD, Steven A. ; MÜLLER, Matthias M.: Global facilitation of attended features is obligatory and restricts divided attention. In: *Journal of Neuroscience* 33 (2013), Nr. 46, P. 18200–18207

[Aubert and Foerster 1857] AUBERT, H. R. ; FOERSTER, C. F. R.: Beitrage zur Kenntniss des indirecten Sehens. (I). Untersuchungen uber den Raumsinn der Retina. In: *Archiv fur Ophthalmologie* 3 (1857), P. 1–37

[Bar 2004] BAR, Moshe: Visual objects in context. In: *Nature Reviews Neuroscience* 5 (2004), Nr. 8, P. 617

[Bartolucci and Smith 2011] BARTOLUCCI, Marco ; SMITH, Andrew T.: Attentional modulation in visual cortex is modified during perceptual learning. In: *Neuropsychologia* 49 (2011), Nr. 14, P. 3898–3907

[Bashivan et al. 2019] BASHIVAN, Pouya ; KAR, Kohitij ; DICARLO, James J.: Neural population control via deep image synthesis. In: *Science* 364 (2019), Nr. 6439

[Becker et al. 2013] BECKER, Stefanie I. ; FOLK, Charles L. ; REMINGTON, Roger W.: Attentional capture does not depend on feature similarity, but on target-nontarget relations. In: *Psychological Science* 24 (2013), Nr. 5, P. 634–647

[Bichot et al. 2005] BICHOT, Narcisse P. ; ROSSI, Andrew F. ; DESIMONE, Robert: Parallel and serial neural mechanisms for visual search in macaque area V4. In: *Science* 308 (2005), Nr. 5721, P. 529–534

[Biederman 1972] BIEDERMAN, Irving: Perceiving real-world scenes. In: *Science* 177 (1972), Nr. 4043, P. 77–80

[Biederman 1976] BIEDERMAN, Irving: On processing information from a glance at a scene: Some implications for a syntax and semantics of visual processing. In: *Proceedings of the ACM/SIGGRAPH workshop on user-oriented design of interactive graphics systems*, 1976, P. 75–88

[Biederman et al. 1973] BIEDERMAN, Irving ; GLASS, Arnold L. ; STACY, E W.: Searching for objects in real-world scenes. In: *Journal of experimental psychology* 97 (1973), Nr. 1, P. 22

[Boettcher et al. 2018] BOETTCHER, Sage E. ; DRASCHKOW, Dejan ; DIENHART, Eric ; VÕ, Melissa L-H: Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. In: *Journal of vision* 18 (2018), Nr. 13, P. 11–11

[Bonner and Epstein 2021] BONNER, Michael F. ; EPSTEIN, Russell A.: Object representations in the human brain reflect the co-occurrence statistics of vision and language. In: *Nature Communications* 12 (2021), Nr. 1, P. 1–16

[Boynton 2009] BOYNTON, Geoffrey M.: A framework for describing the effects of attention on visual responses. In: *Vision Research* 49 (2009), Nr. 10, P. 1129–1143

[Brady et al. 2009] BRADY, Timothy F. ; KONKLE, Talia ; ALVAREZ, George A.: Compression in visual working memory: using statistical regularities to form more efficient memory representations. In: *Journal of Experimental Psychology: General* 138 (2009), Nr. 4, P. 487

[Brady et al. 2011] BRADY, Timothy F. ; KONKLE, Talia ; ALVAREZ, George A.: A review of visual memory capacity: Beyond individual items and toward structured representations. In: *Journal of vision* 11 (2011), Nr. 5, P. 4–4

[Brady and Tenenbaum 2013] BRADY, Timothy F. ; TENENBAUM, Joshua B.: A probabilistic model of visual working memory: Incorporating higher order regularities into working memory capacity estimates. In: *Psychological review* 120 (2013), Nr. 1, P. 85

[Braitenberg 1986] BRAITENBERG, Valentino: *Vehicles: Experiments in synthetic psychology*. MIT press, 1986

[Broadbent 1958] BROADBENT, Donald E.: *Perception and communication*. London: Pergamon, 1958

[Bruckmaier et al. 2020] BRUCKMAIER, Merit ; TACHTSIDIS, Ilias ; PHAN, Phong ; LAVIE, Nilli: Attention and capacity limits in perception: A cellular metabolism account. In: *Journal of Neuroscience* 40 (2020), Nr. 35, P. 6801–6811

[Brunelli 2009] BRUNELLI, Roberto: *Template matching techniques in computer vision: theory and practice*. John Wiley & Sons, 2009

[Buffalo et al. 2010] BUFFALO, Elizabeth A. ; FRIES, Pascal ; LANDMAN, Rogier ; LIANG, Hualou ; DESIMONE, Robert: A backward progression of attentional effects in the ventral stream. In: *Proceedings of the National Academy of Sciences* 107 (2010), Nr. 1, P. 361–365

[Carrasco 2011] CARRASCO, Marisa: Visual attention: The past 25 years. In: *Vision research* 51 (2011), Nr. 13, P. 1484–1525

[Cave and Wolfe 1990] CAVE, Kyle R. ; WOLFE, Jeremy M.: Modeling the role of parallel processing in visual search. In: *Cognitive psychology* 22 (1990), Nr. 2, P. 225–271

[Chaumon et al. 2008] CHAUMON, Maximilien ; DROUET, Valérie ; TALLON-BAUDRY, Catherine: Unconscious associative memory affects visual processing before 100 ms. In: *Journal of vision* 8 (2008), Nr. 3, P. 10–10

[Chelazzi et al. 2019] CHELAZZI, Leonardo ; MARINI, Francesco ; PASCUCCI, David ; TURATTO, Massimo: Getting rid of visual distractors: the why, when, how, and where. In: *Current opinion in psychology* 29 (2019), P. 135–147

[Cheung et al. 2019] CHEUNG, Brian ; TEREKHOV, Alex ; CHEN, Yubei ; AGRAWAL, Pulkit ; OLSHAUSEN, Bruno: Superposition of many models into one. In: *arXiv preprint arXiv:1902.05522* (2019)

[Chun 2000] CHUN, Marvin M.: Contextual cueing of visual attention. In: *Trends in cognitive sciences* 4 (2000), Nr. 5, P. 170–178

[Chun et al. 2011] CHUN, Marvin M. ; GOLOMB, Julie D. ; TURK-BROWNE, Nicholas B.: A taxonomy of external and internal attention. In: *Annual review of psychology* 62 (2011), P. 73–101

[Chun and Jiang 1998] CHUN, Marvin M. ; JIANG, Yuhong: Contextual cueing: Implicit learning and memory of visual context guides spatial attention. In: *Cognitive psychology* 36 (1998), Nr. 1, P. 28–71

[Cichy et al. 2016] CICHY, Radoslaw M. ; KHOSLA, Aditya ; PANTAZIS, Dimitrios ; TORRALBA, Antonio ; OLIVA, Aude: Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. In: *Scientific Reports* 6 (2016), P. 27755

[Cichy and Pantazis 2017] CICHY, Radoslaw M. ; PANTAZIS, Dimitrios: Multivariate pattern analysis of MEG and EEG: A comparison of representational structure in time and space. In: *NeuroImage* 158 (2017), P. 441–454

[Cohen et al. 2011] COHEN, Michael A. ; ALVAREZ, George A. ; NAKAYAMA, Ken: Natural-scene perception requires attention. In: *Psychological science* 22 (2011), Nr. 9, P. 1165–1172

[Cohen et al. 2016] COHEN, Michael A. ; DENNETT, Daniel C. ; KANWISHER, Nancy: What is the bandwidth of perceptual experience? In: *Trends in cognitive sciences* 20 (2016), Nr. 5, P. 324–335

[Covington et al. 2018] COVINGTON, Natalie V. ; BROWN-SCHMIDT, Sarah ; DUFF, Melissa C.: The necessity of the hippocampus for statistical learning. In: *Journal of cognitive neuroscience* 30 (2018), Nr. 5, P. 680–697

[Curcio et al. 1990] CURCIO, Christine A. ; SLOAN, Kenneth R. ; KALINA, Robert E. ; HENDRICKSON, Anita E.: Human photoreceptor topography. In: *Journal of comparative neurology* 292 (1990), Nr. 4, P. 497–523

[De Beeck et al. 2008] DE BEECK, Hans P O. ; HAUSHOFER, Johannes ; KANWISHER, Nancy G.: Interpreting fMRI data: maps, modules and dimensions. In: *Nature Reviews Neuroscience* 9 (2008), Nr. 2, P. 123–135

[DiCarlo and Cox 2007] DICARLO, James J. ; COX, David D.: Untangling invariant object recognition. In: *Trends in cognitive sciences* 11 (2007), Nr. 8, P. 333–341

[Downing et al. 2004] DOWNING, Paul E. ; BRAY, David ; ROGERS, Jack ; CHILDS, Claire: Bodies capture attention when nothing is expected. In: *Cognition* 93 (2004), Nr. 1, P. B27–B38

[Downing et al. 2006] DOWNING, Paul E. ; CHAN, AW-Y ; PEELEN, MV ; DODDS, CM ; KANWISHER, N: Domain specificity in visual cortex. In: *Cerebral cortex* 16 (2006), Nr. 10, P. 1453–1461

[Downing et al. 2001] DOWNING, Paul E. ; JIANG, Yuhong ; SHUMAN, Miles ; KANWISHER, Nancy: A cortical area selective for visual processing of the human body. In: *Science* 293 (2001), Nr. 5539, P. 2470–2473

[Eckstein 2011] ECKSTEIN, Miguel P.: Visual search: A retrospective. In: *Journal of vision* 11 (2011), Nr. 5, P. 14–14

[Eger et al. 2008] EGER, Evelyn ; ASHBURNER, John ; HAYNES, John-Dylan ; DOLAN, Raymond J. ; REES, Geraint: fMRI activity patterns in human LOC carry information about object exemplars within category. In: *Journal of cognitive neuroscience* 20 (2008), Nr. 2, P. 356–370

[Egeth et al. 1984] EGETH, Howard E. ; VIRZI, Robert A. ; GARBART, Hadley: Searching for conjunctively defined targets. In: *Journal of Experimental Psychology: Human Perception and Performance* 10 (1984), Nr. 1, P. 32

[Egly et al. 1994] EGLY, Robert ; DRIVER, Jon ; RAFAL, Robert D.: Shifting visual attention between objects and locations: evidence from normal and parietal lesion subjects. In: *Journal of Experimental Psychology: General* 123 (1994), Nr. 2, P. 161

[Ehinger et al. 2009] EHINGER, Krista A. ; HIDALGO-SOTELO, Barbara ; TORRALBA, Antonio ; OLIVA, Aude: Modelling search for people in 900 scenes: A combined source model of eye guidance. In: *Visual cognition* 17 (2009), Nr. 6-7, P. 945–978

[Eid et al. 2017] EID, Michael ; GOLLWITZER, Mario ; SCHMITT, Manfred: *Statistik und forschungsmethoden*. Beltz, 2017

[Evans and Treisman 2005] EVANS, Karla K. ; TREISMAN, Anne: Perception of objects in natural scenes: is it really attention free? In: *Journal of Experimental Psychology: Human Perception and Performance* 31 (2005), Nr. 6, P. 1476

[Faul et al. 2009] FAUL, Franz ; ERDFELDER, Edgar ; BUCHNER, Axel ; LANG, Albert-Georg: Statistical power analyses using G* Power 3.1: Tests for correlation and regression analyses. In: *Behavior research methods* 41 (2009), Nr. 4, P. 1149–1160

[Fedorenko et al. 2010] FEDORENKO, Evelina ; HSIEH, Po-Jang ; NIETO-CASTAÑÓN, Alfonso ; WHITFIELD-GABRIELI, Susan ; KANWISHER, Nancy: New method for fMRI investigations of language: defining ROIs functionally in individual subjects. In: *Journal of neurophysiology* 104 (2010), Nr. 2, P. 1177–1194

[Fiser and Aslin 2001] FISER, József ; ASLIN, Richard N.: Unsupervised statistical learning of higher-order spatial structures from visual scenes. In: *Psychological science* 12 (2001), Nr. 6, P. 499–504

[Fiser and Aslin 2005] FISER, József ; ASLIN, Richard N.: Encoding multielement scenes: statistical learning of visual feature hierarchies. In: *Journal of Experimental Psychology: General* 134 (2005), Nr. 4, P. 521

[Fiser and Lengyel 2019] FISER, József ; LENGYEL, Gábor: A common probabilistic framework for perceptual and statistical learning. In: *Current Opinion in Neurobiology* 58 (2019), P. 218–228

[Flesch et al. 2018] FLESCH, Timo ; BALAGUER, Jan ; DEKKER, Ronald ; NILI, Hamed ; SUMMERFIELD, Christopher: Comparing continual task learning in minds and machines. In: *Proceedings of the National Academy of Sciences* 115 (2018), Nr. 44, P. E10313–E10322

[Franconeri et al. 2009] FRANCONERI, Steven L. ; BEMIS, Douglas K. ; ALVAREZ, George A.: Number estimation relies on a set of segmented objects. In: *Cognition* 113 (2009), Nr. 1, P. 1–13

[Frank et al. 2014a] FRANK, Michael C. ; AMSO, Dima ; JOHNSON, Scott P.: Visual search and attention to faces during early infancy. In: *Journal of experimental child psychology* 118 (2014), P. 13–26

[Frank et al. 2014b] FRANK, Sebastian M. ; REAVIS, Eric A. ; TSE, Peter U. ; GREENLEE, Mark W.: Neural mechanisms of feature conjunction learning: enduring changes in occipital cortex after a week of training. In: *Human Brain Mapping* 35 (2014), Nr. 4, P. 1201–1211

[Frith and Frith 1972] FRITH, Christopher D. ; FRITH, Uta: The solitaire illusion: An illusion of numerosity. In: *Perception & Psychophysics* 11 (1972), Nr. 6, P. 409–410

[Fukushima and Miyake 1982] FUKUSHIMA, Kunihiko ; MIYAKE, Sei: Neocognitron: A self-organizing neural network model for a mechanism of visual pattern recognition. In: *Competition and cooperation in neural nets*. Springer, 1982, P. 267–285

[Gaspelin and Luck 2018] GASPELIN, Nicholas ; LUCK, Steven J.: Combined electrophysiological and behavioral evidence for the suppression of salient distractors. In: *Journal of cognitive neuroscience* 30 (2018), Nr. 9, P. 1265–1280

[Gauthier and Logothetis 2000] GAUTHIER, Isabel ; LOGOTHETIS, Nikos K.: Is face recognition not so unique after all? In: *Cognitive Neuropsychology* 17 (2000), Nr. 1-3, P. 125–142

[Geng et al. 2017] GENG, Joy J. ; DIQUATTRO, Nicholas E. ; HELM, Jonathan: Distractor probability changes the shape of the attentional template. In: *Journal of Experimental Psychology: Human Perception and Performance* 43 (2017), Nr. 12, P. 1993

[Gilbert and Li 2013] GILBERT, Charles D. ; LI, Wu: Top-down influences on visual processing. In: *Nature Reviews Neuroscience* 14 (2013), Nr. 5, P. 350–363

[Ginsburg and Goldstein 1987] GINSBURG, Norman ; GOLDSTEIN, Stephen R.: Measurement of visual cluster. In: *The American Journal of Psychology* (1987), P. 193–203

[Golan et al. 2014] GOLAN, Tal ; BENTIN, Shlomo ; DEGUTIS, Joseph M. ; ROBERTSON, Lynn C. ; HAREL, Assaf: Association and dissociation between detection and discrimination of objects of expertise: Evidence from visual search. In: *Attention, Perception, & Psychophysics* 76 (2014), Nr. 2, P. 391–406

[Gottlieb and Oudeyer 2018] GOTTLIEB, Jacqueline ; OUDEYER, Pierre-Yves: Towards a neuroscience of active sampling and curiosity. In: *Nature Reviews Neuroscience* 19 (2018), Nr. 12, P. 758–770

[Green and Anderson 1956] GREEN, Bert F. ; ANDERSON, Lois K.: Color coding in a visual search task. In: *Journal of experimental psychology* 51 (1956), Nr. 1, P. 19

[Green et al. 1953] GREEN, Bert F. ; MCGILL, William J. ; JENKINS, Herbert M.: *The time required to search for numbers on large visual displays*. MIT Lincoln Laboratory, 1953

[Güçlü and van Gerven 2015] GÜÇLÜ, Umut ; GERVEN, Marcel A. van: Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. In: *Journal of Neuroscience* 35 (2015), Nr. 27, P. 10005–10014

[Haenny and Schiller 1988] HAENNY, PE ; SCHILLER, PH: State dependent activity in monkey visual cortex. In: *Experimental brain research* 69 (1988), Nr. 2, P. 225–244

[Harel et al. 2014] HAREL, Assaf ; KRAVITZ, Dwight J. ; BAKER, Chris I.: Task context impacts visual object processing differentially across the cortex. In: *Proceedings of the National Academy of Sciences* 111 (2014), Nr. 10, P. E962–E971

[Haun et al. 2017] HAUN, Andrew M. ; TONONI, Giulio ; KOCH, Christof ; TSUCHIYA, Naotsugu: Are we underestimating the richness of visual experience? In: *Neuroscience of Consciousness* 2017 (2017), Nr. 1, P. niw023

[Haushofer et al. 2008] HAUSHOFER, Johannes ; LIVINGSTONE, Margaret S. ; KANWISHER, Nancy: Multivariate patterns in object-selective cortex dissociate perceptual and physical shape similarity. In: *PLoS biology* 6 (2008), Nr. 7, P. e187

[He et al. 2009] HE, Lixia ; ZHANG, Jun ; ZHOU, Tiangang ; CHEN, Lin: Connectedness affects dot numerosity judgment: Implications for configural processing. In: *Psychonomic Bulletin & Review* 16 (2009), Nr. 3, P. 509–517

[Hershler and Hochstein 2005] HERSHLER, Orit ; HOCHSTEIN, Shaul: At first sight: A high-level pop out effect for faces. In: *Vision research* 45 (2005), Nr. 13, P. 1707–1724

[Hershler and Hochstein 2006] HERSHLER, Orit ; HOCHSTEIN, Shaul: With a careful look: Still no low-level confound to face pop-out. In: *Vision research* 46 (2006), Nr. 18, P. 3028–3035

[Hershler and Hochstein 2009] HERSHLER, Orit ; HOCHSTEIN, Shaul: The importance of being expert: Top-down attentional control in visual search with photographs. In: *Attention, Perception, & Psychophysics* 71 (2009), Nr. 7, P. 1478–1486

[Hubel and Wiesel 1962] HUBEL, David H. ; WIESEL, Torsten N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. In: *The Journal of physiology* 160 (1962), Nr. 1, P. 106–154

[Jehee et al. 2011] JEHEE, Janneke F. ; BRADY, Devin K. ; TONG, Frank: Attention improves encoding of task-relevant features in the human visual cortex. In: *Journal of Neuroscience* 31 (2011), Nr. 22, P. 8210–8219

[Johnson et al. 2007] JOHNSON, Jeffrey S. ; WOODMAN, Geoffrey F. ; BRAUN, Elsie ; LUCK, Steven J.: Implicit memory influences the allocation of attention in visual cortex. In: *Psychonomic Bulletin & Review* 14 (2007), Nr. 5, P. 834–839

[Julian et al. 2012] JULIAN, Joshua B. ; FEDORENKO, Evelina ; WEBSTER, Jason ; KANWISHER, Nancy: An algorithmic method for functionally defining regions of interest in the ventral visual pathway. In: *Neuroimage* 60 (2012), Nr. 4, P. 2357–2364

[Kaiser and Peelen 2018] KAISER, Daniel ; PEELEN, Marius V.: Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. In: *Neuroimage* 169 (2018), P. 334–341

[Kaiser et al. 2019] KAISER, Daniel ; QUEK, Genevieve L. ; CICHY, Radoslaw M. ; PEELEN, Marius V.: Object vision in a structured world. In: *Trends in cognitive sciences* 23 (2019), Nr. 8, P. 672–685

[Kaiser et al. 2014] KAISER, Daniel ; STEIN, Timo ; PEELEN, Marius V.: Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. In: *Proceedings of the National Academy of Sciences* 111 (2014), Nr. 30, P. 11217–11222

[Kamitani and Tong 2005] KAMITANI, Yukiyasu ; TONG, Frank: Decoding the visual and subjective contents of the human brain. In: *Nature neuroscience* 8 (2005), Nr. 5, P. 679–685

[Kaneshiro et al. 2015] KANESHIRO, Blair ; PERREAU GUIMARAES, Marcos ; KIM, Hyung-Suk ; NORCIA, Anthony M. ; SUPPES, Patrick: A representational similarity analysis of the dynamics of object processing using single-trial EEG classification. In: *Plos one* 10 (2015), Nr. 8, P. e0135697

[Kanwisher 2010] KANWISHER, Nancy: Functional specificity in the human brain: a window into the functional architecture of the mind. In: *Proceedings of the National Academy of Sciences* 107 (2010), Nr. 25, P. 11163–11170

[Kanwisher et al. 1997] KANWISHER, Nancy ; MCDERMOTT, Josh ; CHUN, Marvin M.: The fusiform face area: a module in human extrastriate cortex specialized for face perception. In: *Journal of neuroscience* 17 (1997), Nr. 11, P. 4302–4311

[Kar et al. 2019] KAR, Kohitij ; KUBILIUS, Jonas ; SCHMIDT, Kailyn ; ISSA, Elias B. ; DICARLO, James J.: Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. In: *Nature neuroscience* 22 (2019), Nr. 6, P. 974–983

[Kastner and Pinsk 2004] KASTNER, Sabine ; PINSK, Mark A.: Visual attention as a multilevel selection process. In: *Cognitive, Affective, & Behavioral Neuroscience* 4 (2004), Nr. 4, P. 483–500

[Kietzmann et al. 2019] KIETZMANN, Tim C. ; SPOERER, Courtney J. ; SÖRENSEN, Lynn K. ; CICHY, Radoslaw M. ; HAUK, Olaf ; KRIEGESKORTE, Nikolaus: Recurrence is required to capture the representational dynamics of the human visual system. In: *Proceedings of the National Academy of Sciences* 116 (2019), Nr. 43, P. 21854–21863

[Kingma and Ba 2014] KINGMA, Diederik P. ; BA, Jimmy: Adam: A method for stochastic optimization. In: *arXiv preprint arXiv:1412.6980* (2014)

[Kirkpatrick et al. 2017] KIRKPATRICK, James ; PASCANU, Razvan ; RABINOWITZ, Neil ; VENESS, Joel ; DESJARDINS, Guillaume ; RUSU, Andrei A. ; MILAN, Kieran ; QUAN, John ; RAMALHO, Tiago ; GRABSKA-BARWINSKA, Agnieszka et al.: Overcoming catastrophic forgetting in neural networks. In: *Proceedings of the National Academy of Sciences* 114 (2017), Nr. 13, P. 3521–3526

[Koopman 1956b] KOOPMAN, Bernard O.: The theory of search. II. Target detection. In: *Operations research* 4 (1956b), Nr. 5, P. 503–531

[Koopman 1956a] KOOPMAN, Bernard O.: The theory of search. I. Kinematic bases. In: *Operations research* 4 (1956a), Nr. 3, P. 324–346

[Kriegeskorte 2015] KRIEGESKORTE, Nikolaus: Deep neural networks: a new framework for modeling biological vision and brain information processing. In: *Annual Review of Vision Science* 1 (2015), P. 417–446

[Krizhevsky et al. 2012] KRIZHEVSKY, Alex ; SUTSKEVER, Ilya ; HINTON, Geoffrey E.: Imagenet classification with deep convolutional neural networks. In: *Advances in neural information processing systems* 25 (2012), P. 1097–1105

[Kumaran et al. 2016] KUMARAN, Dharshan ; HASSABIS, Demis ; MCCLELLAND, James L.: What learning systems do intelligent agents need? Complementary learning systems theory updated. In: *Trends in Cognitive Sciences* 20 (2016), Nr. 7, P. 512–534

[Lamme and Roelfsema 2000] LAMME, Victor A. ; ROELFSEMA, Pieter R.: The distinct modes of vision offered by feedforward and recurrent processing. In: *Trends in neurosciences* 23 (2000), Nr. 11, P. 571–579

[Lavie 1995] LAVIE, Nilli: Perceptual load as a necessary condition for selective attention. In: *Journal of Experimental Psychology: Human perception and performance* 21 (1995), Nr. 3, P. 451

[Lavie and Tsal 1994] LAVIE, Nilli ; TSAL, Yehoshua: Perceptual load as a major determinant of the locus of selection in visual attention. In: *Perception & psychophysics* 56 (1994), Nr. 2, P. 183–197

[Leber et al. 2016] LEBER, Andrew B. ; GWINN, Rachael E. ; HONG, Yoolim ; O'TOOLE, Ryan J.: Implicitly learned suppression of irrelevant spatial locations. In: *Psychonomic Bulletin & Review* 23 (2016), Nr. 6, P. 1873–1881

[LeCun et al. 2015] LECUN, Yann ; BENGIO, Yoshua ; HINTON, Geoffrey: Deep learning. In: *nature* 521 (2015), Nr. 7553, P. 436–444

[LeCun et al. 1998] LECUN, Yann ; BOTTOU, Léon ; BENGIO, Yoshua ; HAFFNER, Patrick: Gradient-based learning applied to document recognition. In: *Proceedings of the IEEE* 86 (1998), Nr. 11, P. 2278–2324

[Lee et al. 1999] LEE, Dale K. ; ITTI, Laurent ; KOCH, Christof ; BRAUN, Jochen: Attention activates winner-take-all competition among visual filters. In: *Nature neuroscience* 2 (1999), Nr. 4, P. 375–381

[Lengyel et al. 2021] LENGYEL, Gábor ; NAGY, Márton ; FISER, József: Statistically defined visual chunks engage object-based attention. In: *Nature communications* 12 (2021), Nr. 1, P. 1–12

[Li et al. 2002] LI, Fei F. ; VANRULLEN, Rufin ; KOCH, Christof ; PERONA, Pietro: Rapid natural scene categorization in the near absence of attention. In: *Proceedings of the National Academy of Sciences* 99 (2002), Nr. 14, P. 9596–9601

[Lindsay 2021] LINDSAY, Grace: *Models of the Mind: How Physics, Engineering and Mathematics Have Shaped Our Understanding of the Brain*. Bloomsbury Publishing, 2021

[Lindsay and Miller 2017] LINDSAY, Grace W. ; MILLER, Kenneth D.: Understanding Biological Visual Attention Using Convolutional Neural Networks. In: *bioRxiv* (2017), P. 233338

[Lindsay and Miller 2018] LINDSAY, Grace W. ; MILLER, Kenneth D.: How biological attention mechanisms improve task performance in a large-scale visual system model. In: *eLife* 7 (2018), P. e38105

[Ling et al. 2009] LING, Sam ; LIU, Taosheng ; CARRASCO, Marisa: How spatial and feature-based attention affect the gain and tuning of population responses. In: *Vision Research* 49 (2009), Nr. 10, P. 1194–1204

[Linsley et al. 2020] LINSLEY, Drew ; ASHOK, Alekh K. ; GOVINDARAJAN, Lakshmi N. ; LIU, Rex ; SERRE, Thomas: Stable and expressive recurrent vision models. In: *arXiv preprint arXiv:2005.11362* (2020)

[Liu et al. 2007a] LIU, Taosheng ; LARSSON, Jonas ; CARRASCO, Marisa: Feature-based attention modulates orientation-selective responses in human visual cortex. In: *Neuron* 55 (2007), Nr. 2, P. 313–323

[Liu et al. 2007b] LIU, Taosheng ; STEVENS, Sean T. ; CARRASCO, Marisa: Comparing the time course and efficacy of spatial and feature-based attention. In: *Vision research* 47 (2007), Nr. 1, P. 108–113

[Luck 2006] LUCK, Steven J.: The Operation of Attention–Millisecond by Millisecond–Over the First Half Second. In: *The first half second: The microgenesis and temporal dynamics of unconscious and conscious visual processes*. Mit Press, 2006, P. 187–206

[Luck 2012] LUCK, Steven J.: Electrophysiological correlates of the focusing of attention within complex visual scenes: N2pc and related ERP components. In: *The Oxford Handbook of Event-Related Potential Components* (2012)

[Luo et al. 2021] LUO, Xiaoliang ; ROADS, Brett D. ; LOVE, Bradley C.: The costs and benefits of goal-directed attention in deep convolutional neural networks. In: *Computational Brain & Behavior* 4 (2021), Nr. 2, P. 213–230

[Mante et al. 2013] MANTE, Valerio ; SUSSILLO, David ; SHENOY, Krishna V. ; NEWSOME, William T.: Context-dependent computation by recurrent dynamics in prefrontal cortex. In: *Nature* 503 (2013), Nr. 7474, P. 78

[Martinez-Trujillo and Treue 2004] MARTINEZ-TRUJILLO, Julio C. ; TREUE, Stefan: Feature-based attention increases the selectivity of population responses in primate visual cortex. In: *Current Biology* 14 (2004), Nr. 9, P. 744–751

[Masse et al. 2018] MASSE, Nicolas Y. ; GRANT, Gregory D. ; FREEDMAN, David J.: Alleviating catastrophic forgetting using context-dependent gating and synaptic stabilization. In: *Proceedings of the National Academy of Sciences* 115 (2018), Nr. 44, P. E10467–E10475

[Maunsell et al. 1991] MAUNSELL, John H. ; SCLAR, Gary ; NEALEY, Tara A. ; DEPRIEST, Derryl D.: Extraretinal representations in area V4 in the macaque monkey. In: *Visual neuroscience* 7 (1991), Nr. 6, P. 561–573

[Maunsell and Treue 2006] MAUNSELL, John H. ; TREUE, Stefan: Feature-based attention in visual cortex. In: *Trends in neurosciences* 29 (2006), Nr. 6, P. 317–322

[McAdams and Maunsell 2000] McAdams, Carrie J. ; Maunsell, John H.: Attention to both space and feature modulates neuronal responses in macaque area V4. In: *Journal of neurophysiology* 83 (2000), Nr. 3, P. 1751–1755

[McGugin et al. 2012] McGugin, Rankin W. ; Gatenby, J C. ; Gore, John C. ; Gauthier, Isabel: High-resolution imaging of expertise reveals reliable object selectivity in the fusiform face area related to perceptual performance. In: *Proceedings of the National Academy of Sciences* 109 (2012), Nr. 42, P. 17063–17068

[McMains and Kastner 2010] McMains, Stephanie A. ; Kastner, Sabine: Defining the units of competition: influences of perceptual organization on competitive interactions in human visual cortex. In: *Journal of Cognitive Neuroscience* 22 (2010), Nr. 11, P. 2417–2426

[McNemar 1947] McNemar, Quinn: Note on the sampling error of the difference between correlated proportions or percentages. In: *Psychometrika* 12 (1947), Nr. 2, P. 153–157

[Meyen et al. 2021] Meyen, Sascha ; Zerweck, Iris A. ; Amado, Catarina ; Luxburg, Ulrike von ; Franz, Volker H.: Advancing research on unconscious priming: When can scientists claim an indirect task advantage? In: *Journal of Experimental Psychology: General* (2021)

[Meyer and Rust 2018] Meyer, Travis ; Rust, Nicole C.: Single-exposure visual memory judgments are reflected in inferotemporal cortex. In: *Elife* 7 (2018), P. e32259

[van Moorselaar et al. 2021] Moorselaar, Dirk van ; Daneshtalab, Nasim ; Slagter, Heleen A.: Neural mechanisms underlying distractor inhibition on the basis of feature and/or spatial expectations. In: *Cortex* 137 (2021), P. 232–250

[Morcos et al. 2018] Morcos, Ari S. ; Barrett, David G. ; Rabinowitz, Neil C. ; Botvinick, Matthew: On the importance of single directions for generalization. In: *arXiv preprint arXiv:1803.06959* (2018)

[Nakayama and Martini 2011] Nakayama, Ken ; Martini, Paolo: Situating visual search. In: *Vision research* 51 (2011), Nr. 13, P. 1526–1537

[Nakayama and Silverman 1986] Nakayama, Ken ; Silverman, Gerald H.: Serial and parallel processing of visual feature conjunctions. In: *Nature* 320 (1986), Nr. 6059, P. 264–265

[Navalpakkam and Itti 2007] Navalpakkam, Vidhya ; Itti, Laurent: Search goal tunes visual features optimally. In: *Neuron* 53 (2007), Nr. 4, P. 605–617

[Newsome and Pare 1988] Newsome, William T. ; Pare, Edmond B.: A selective impairment of motion perception following lesions of the middle temporal visual area (MT). In: *Journal of Neuroscience* 8 (1988), Nr. 6, P. 2201–2211

[O'Craven et al. 1997] O'Craven, Kathleen M. ; Rosen, Bruce R. ; Kwong, Kenneth K. ; Treisman, Anne ; Savoy, Robert L.: Voluntary attention modulates fMRI activity in human MT–MST. In: *Neuron* 18 (1997), Nr. 4, P. 591–598

[O'Reilly et al. 2013] O'Reilly, Randall C. ; Wyatte, Dean ; Herd, Seth ; Mingus, Brian ; Jilk, David J.: Recurrent processing during object recognition. In: *Frontiers in psychology* 4 (2013), P. 124

[Palmer and Rock 1994] PALMER, Stephen ; ROCK, Irvin: Rethinking perceptual organization: The role of uniform connectedness. In: *Psychonomic bulletin & review* 1 (1994), Nr. 1, P. 29–55

[Peelen and Downing 2005] PEELEN, Marius V. ; DOWNING, Paul E.: Selectivity for the human body in the fusiform gyrus. In: *Journal of neurophysiology* 93 (2005), Nr. 1, P. 603–608

[Peelen et al. 2009] PEELEN, Marius V. ; FEI-FEI, Li ; KASTNER, Sabine: Neural mechanisms of rapid natural scene categorization in human visual cortex. In: *Nature* 460 (2009), Nr. 7251, P. 94–97

[Peelen and Kastner 2014] PEELEN, Marius V. ; KASTNER, Sabine: Attention in the real world: toward understanding its neural basis. In: *Trends in cognitive sciences* 18 (2014), Nr. 5, P. 242–250

[Plaut and Vande Velde 2017] PLAUT, David C. ; VANDE VELDE, Anna K.: Statistical learning of parts and wholes: A neural network approach. In: *Journal of Experimental Psychology: General* 146 (2017), Nr. 3, P. 318

[Reavis et al. 2016] REAVIS, Eric A. ; FRANK, Sebastian M. ; GREENLEE, Mark W. ; TSE, Peter U.: Neural correlates of context-dependent feature conjunction learning in visual search tasks. In: *Human brain mapping* 37 (2016), Nr. 6, P. 2319–2330

[Reddy et al. 2007] REDDY, Leila ; MORADI, Farshad ; KOCH, Christof: Top–down biases win against focal attention in the fusiform face area. In: *Neuroimage* 38 (2007), Nr. 4, P. 730–739

[Reed et al. 2003] REED, Catherine L. ; STONE, Valerie E. ; BOZOVA, Senia ; TANAKA, James: The body-inversion effect. In: *Psychological science* 14 (2003), Nr. 4, P. 302–308

[Reeder and Peelen 2013] REEDER, Reshanne R. ; PEELEN, Marius V.: The contents of the search template for category-level search in natural scenes. In: *Journal of Vision* 13 (2013), Nr. 3, P. 13–13

[Reeder et al. 2016] REEDER, Reshanne R. ; STEIN, Timo ; PEELEN, Marius V.: Perceptual expertise improves category detection in natural scenes. In: *Psychonomic bulletin & review* 23 (2016), Nr. 1, P. 172–179

[Reeder et al. 2015] REEDER, Reshanne R. ; ZOEST, Wieske van ; PEELEN, Marius V.: Involuntary attentional capture by task-irrelevant objects that match the search template for category detection in natural scenes. In: *Attention, Perception, & Psychophysics* 77 (2015), Nr. 4, P. 1070–1080

[Regan and Beverley 1985] REGAN, D ; BEVERLEY, KI: Postadaptation orientation discrimination. In: *JOSA A* 2 (1985), Nr. 2, P. 147–155

[Richards et al. 2019] RICHARDS, Blake A. ; LILLICRAP, Timothy P. ; BEAUDOIN, Philippe ; BENGIO, Yoshua ; BOGACZ, Rafal ; CHRISTENSEN, Amelia ; CLOPATH, Claudia ; COSTA, Rui P. ; BERKER, Archy de ; GANGULI, Surya et al.: A deep learning framework for neuroscience. In: *Nature neuroscience* 22 (2019), Nr. 11, P. 1761–1770

[Ro et al. 2007] RO, Tony ; FRIGGEL, Ashley ; LAVIE, Nilli: Attentional biases for faces and body parts. In: *Visual Cognition* 15 (2007), Nr. 3, P. 322–348

[Rosenblatt 1957] ROSENBLATT, Frank: *The perceptron, a perceiving and recognizing automaton Project Para*. Cornell Aeronautical Laboratory, 1957

[Rosenfeld et al. 2018] ROSENFELD, Amir ; BIPARVA, Mahdi ; TSOTSOS, John K.: Priming Neural Networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, P. 2011–2020

[Russakovsky et al. 2015] RUSSAKOVSKY, Olga ; DENG, Jia ; SU, Hao ; KRAUSE, Jonathan ; SATHEESH, Sanjeev ; MA, Sean ; HUANG, Zhiheng ; KARPATHY, Andrej ; KHOSLA, Aditya ; BERNSTEIN, Michael et al.: Imagenet large scale visual recognition challenge. In: *International Journal of Computer Vision* 115 (2015), Nr. 3, P. 211–252

[Saenz et al. 2002] SAENZ, Melissa ; BURACAS, Giedrius T. ; BOYNTON, Geoffrey M.: Global effects of feature-based attention in human visual cortex. In: *Nature neuroscience* 5 (2002), Nr. 7, P. 631–632

[Schankin and Schubö 2009] SCHANKIN, Andrea ; SCHUBÖ, Anna: Cognitive processes facilitated by contextual cueing: Evidence from event-related brain potentials. In: *Psychophysiology* 46 (2009), Nr. 3, P. 668–679

[Schapiro and Turk-Browne 2015] SCHAPIRO, A ; TURK-BROWNE, Nicholas: Statistical learning. In: *Brain mapping* 3 (2015), P. 501–506

[Schwartz et al. 2005] SCHWARTZ, Sophie ; VUILLEUMIER, Patrik ; HUTTON, Chloe ; MARAVITA, Angelo ; DOLAN, Raymond J. ; DRIVER, Jon: Attentional load and sensory competition in human vision: modulation of fMRI responses by load at fixation during task-irrelevant stimulation in the peripheral visual field. In: *Cerebral cortex* 15 (2005), Nr. 6, P. 770–786

[Scolari et al. 2012] SCOLARI, Miranda ; BYERS, Anna ; SERENCES, John T.: Optimal deployment of attentional gain during fine discriminations. In: *Journal of Neuroscience* 32 (2012), Nr. 22, P. 7723–7733

[Scolari and Serences 2009] SCOLARI, Miranda ; SERENCES, John T.: Adaptive allocation of attentional gain. In: *Journal of Neuroscience* 29 (2009), Nr. 38, P. 11933–11942

[Serences and Boynton 2007] SERENCES, John T. ; BOYNTON, Geoffrey M.: Feature-based attentional modulations in the absence of direct visual stimulation. In: *Neuron* 55 (2007), Nr. 2, P. 301–312

[Serences and Kastner 2014] SERENCES, John T. ; KASTNER, Sabine: A multi-level account of selective attention. In: *The Oxford Handbook of Attention* (2014)

[Simonyan and Zisserman 2014] SIMONYAN, Karen ; ZISSERMAN, Andrew: Very deep convolutional networks for large-scale image recognition. In: *arXiv preprint arXiv:1409.1556* (2014)

[Sisk et al. 2019] SISK, Caitlin A. ; REMINGTON, Roger W. ; JIANG, Yuhong V.: Mechanisms of contextual cueing: A tutorial review. In: *Attention, Perception, & Psychophysics* 81 (2019), Nr. 8, P. 2571–2589

[Smith 1962] SMITH, Sidney L.: Color coding and visual search. In: *Journal of Experimental Psychology* 64 (1962), Nr. 5, P. 434

[Smith and Nichols 2009] SMITH, Stephen M. ; NICHOLS, Thomas E.: Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. In: *Neuroimage* 44 (2009), Nr. 1, P. 83–98

[Spaak and de Lange 2020] SPAAK, Eelke ; LANGE, Floris P. de: Hippocampal and prefrontal theta-band mechanisms underpin implicit spatial context learning. In: *Journal of Neuroscience* 40 (2020), Nr. 1, P. 191–202

[Srivastava et al. 2014] SRIVASTAVA, Nitish ; HINTON, Geoffrey ; KRIZHEVSKY, Alex ; SUTSKEVER, Ilya ; SALAKHUTDINOV, Ruslan: Dropout: A simple way to prevent neural networks from overfitting. In: *The Journal of Machine Learning Research* 15 (2014), Nr. 1, P. 1929–1958

[Stein et al. 2015] STEIN, Timo ; KAISER, Daniel ; PEELEN, Marius V.: Interobject grouping facilitates visual awareness. In: *Journal of Vision* 15 (2015), Nr. 8, P. 10–10

[Stein and Peelen 2017] STEIN, Timo ; PEELEN, Marius V.: Object detection in natural scenes: Independent effects of spatial and category-based attention. In: *Attention, Perception, & Psychophysics* 79 (2017), Nr. 3, P. 738–752

[Stein et al. 2016] STEIN, Timo ; REEDER, Reshanne R. ; PEELEN, Marius V.: Privileged access to awareness for faces and objects of expertise. In: *Journal of Experimental Psychology: Human Perception and Performance* 42 (2016), Nr. 6, P. 788

[Stein et al. 2012] STEIN, Timo ; STERZER, Philipp ; PEELEN, Marius V.: Privileged detection of conspecifics: Evidence from inversion effects during continuous flash suppression. In: *Cognition* 125 (2012), Nr. 1, P. 64–79

[Störmer et al. 2019] STÖRMER, Viola S. ; COHEN, Michael A. ; ALVAREZ, George A.: Tuning attention to object categories: Spatially global effects of attention to faces in visual processing. In: *Journal of cognitive neuroscience* 31 (2019), Nr. 7, P. 937–947

[Strasburger et al. 2011] STRASBURGER, Hans ; RENTSCHLER, Ingo ; JÜTTNER, Martin: Peripheral vision and pattern recognition: A review. In: *Journal of vision* 11 (2011), Nr. 5, P. 13–13

[Thorat et al. 2019] THORAT, SR ; ALDEGHERI, G ; GERVEN, MAJ van ; PEELEN, MV: Modulation of early visual processing alleviates capacity limits in solving multiple tasks. In: *2019 Conference on Cognitive Computational Neuroscience*, 2019, P. 226–229

[Thorat et al. 2018] THORAT, SR ; GERVEN, MAJ van ; PEELEN, MV: The functional role of cue-driven feature-based feedback in object recognition. In: *Proceedings 2018 Conference on Cognitive Computational Neuroscience*, 2018, P. 1–4

[Thorat et al. 2021] THORAT, Sushrut ; ALDEGHERI, Giacomo ; KIETZMANN, Tim C.: Category-orthogonal object features guide information processing in recurrent neural networks trained for object categorization. In: *SVRHM 2021 Workshop@ NeurIPS*, 2021, P. 1–12

[Thorat and Peelen 2022] THORAT, Sushrut ; PEELEN, Marius V.: Body shape as a visual feature: evidence from spatially-global attentional modulation in human visual cortex. In: *NeuroImage* 255 (2022), P. 119207

[Thorat et al. 2022] THORAT, Sushrut ; QUEK, Genevieve ; PEELEN, Marius V.: Statistical learning of distractor co-occurrences facilitates visual search. In: *bioRxiv* (2022)

[Torralba et al. 2006] TORRALBA, Antonio ; OLIVA, Aude ; CASTELHANO, Monica S. ; HENDERSON, John M.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. In: *Psychological review* 113 (2006), Nr. 4, P. 766

[Treisman 2006] TREISMAN, Anne: How the deployment of attention determines what we see. In: *Visual cognition* 14 (2006), Nr. 4-8, P. 411–443

[Treisman and Gormican 1988] TREISMAN, Anne ; GORMICAN, Stephen: Feature analysis in early vision: evidence from search asymmetries. In: *Psychological review* 95 (1988), Nr. 1, P. 15

[Treisman and Souther 1985] TREISMAN, Anne ; SOUTHER, Janet: Search asymmetry: a diagnostic for preattentive processing of separable features. In: *Journal of Experimental Psychology: General* 114 (1985), Nr. 3, P. 285

[Treisman and Gelade 1980] TREISMAN, Anne M. ; GELADE, Garry: A feature-integration theory of attention. In: *Cognitive psychology* 12 (1980), Nr. 1, P. 97–136

[Treue and Trujillo 1999] TREUE, Stefan ; TRUJILLO, Julio C M.: Feature-based attention influences motion processing gain in macaque visual cortex. In: *Nature* 399 (1999), Nr. 6736, P. 575–579

[Turk-Browne et al. 2005] TURK-BROWNE, Nicholas B. ; JUNGÉ, Justin A. ; SCHOLL, Brian J.: The automaticity of visual statistical learning. In: *Journal of Experimental Psychology: General* 134 (2005), Nr. 4, P. 552

[Turk-Browne et al. 2009] TURK-BROWNE, Nicholas B. ; SCHOLL, Brian J. ; CHUN, Marvin M. ; JOHNSON, Marcia K.: Neural evidence of statistical learning: Efficient detection of visual regularities without awareness. In: *Journal of cognitive neuroscience* 21 (2009), Nr. 10, P. 1934–1945

[Ullman et al. 2002] ULLMAN, Shimon ; VIDAL-NAQUET, Michel ; SALI, Erez: Visual features of intermediate complexity and their use in classification. In: *Nature neuroscience* 5 (2002), Nr. 7, P. 682–687

[VanRullen 2006] VANRULLEN, Rufin: On second glance: Still no high-level pop-out effect for faces. In: *Vision research* 46 (2006), Nr. 18, P. 3017–3027

[Võ et al. 2019] VÕ, Melissa Le-Hoa ; BOETTCHER, Sage E. ; DRASCHKOW, Dejan: Reading scenes: how scene grammar guides attention and aids perception in real-world environments. In: *Current opinion in psychology* 29 (2019), P. 205–210

[Wagemans 2018] WAGEMANS, Johan: Perceptual organization. In: *Stevens' handbook of experimental psychology and cognitive neuroscience* 2 (2018), P. 1–70

[Wang et al. 2012] WANG, Changming ; XIONG, Shi ; HU, Xiaoping ; YAO, Li ; ZHANG, Jiacai: Combining features from ERP components in single-trial EEG for discriminating four-category visual objects. In: *Journal of neural engineering* 9 (2012), Nr. 5, P. 056013

[Wertheimer 1923] WERTHEIMER, Max: Laws of organization in perceptual forms. In: *A source book of Gestalt Psychology* 1 (1923)

[Wohlschläger et al. 2005] WOHLSCHLÄGER, Afra M. ; SPECHT, Karsten ; LIE, C ; MOHLBERG, Hartmut ; WOHLSCHLÄGER, Andreas ; BENTE, Kay ; PIETRZYK, Uwe ; STÖCKER, Tony ; ZILLES, Karl ; AMUNTS, Katrin et al.: Linking retinotopic fMRI mapping and anatomical probability maps of human occipital areas V1 and V2. In: *Neuroimage* 26 (2005), Nr. 1, P. 73–82

[Wolfe 1994] WOLFE, Jeremy M.: Guided search 2.0 a revised model of visual search. In: *Psychonomic bulletin & review* 1 (1994), Nr. 2, P. 202–238

[Wolfe 2021] WOLFE, Jeremy M.: Guided Search 6.0: An updated model of visual search. In: *Psychonomic Bulletin & Review* (2021), P. 1–33

[Wolfe et al. 2000] WOLFE, Jeremy M. ; ALVAREZ, George A. ; HOROWITZ, Todd S.: Attention is fast but volition is slow. In: *Nature* 406 (2000), Nr. 6797, P. 691–691

[Wolfe et al. 2011a] WOLFE, Jeremy M. ; ALVAREZ, George A. ; ROSENHOLTZ, Ruth ; KUZMOVA, Yoana I. ; SHERMAN, Ashley M.: Visual search for arbitrary objects in real scenes. In: *Attention, Perception, & Psychophysics* 73 (2011), Nr. 6, P. 1650–1671

[Wolfe and Bennett 1997] WOLFE, Jeremy M. ; BENNETT, Sara C.: Preattentive object files: Shapeless bundles of basic features. In: *Vision research* 37 (1997), Nr. 1, P. 25–43

[Wolfe et al. 1989] WOLFE, Jeremy M. ; CAVE, Kyle R. ; FRANZEL, Susan L.: Guided search: an alternative to the feature integration model for visual search. In: *Journal of Experimental Psychology: Human perception and performance* 15 (1989), Nr. 3, P. 419

[Wolfe and Horowitz 2004] WOLFE, Jeremy M. ; HOROWITZ, Todd S.: What attributes guide the deployment of visual attention and how do they do it? In: *Nature reviews neuroscience* 5 (2004), Nr. 6, P. 495–501

[Wolfe and Horowitz 2017] WOLFE, Jeremy M. ; HOROWITZ, Todd S.: Five factors that guide attention in visual search. In: *Nature Human Behaviour* 1 (2017), Nr. 3, P. 1–8

[Wolfe et al. 2011b] WOLFE, Jeremy M. ; VÕ, Melissa L-H ; EVANS, Karla K. ; GREENE, Michelle R.: Visual search in scenes involves selective and nonselective pathways. In: *Trends in cognitive sciences* 15 (2011), Nr. 2, P. 77–84

[Wyatte et al. 2014] WYATTE, Dean ; JILK, David J. ; O'REILLY, Randall C.: Early recurrent feedback facilitates visual object recognition under challenging conditions. In: *Frontiers in psychology* 5 (2014), P. 674

[Xiao et al. 2017] XIAO, Han ; RASUL, Kashif ; VOLLGRAF, Roland: Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. In: *arXiv preprint arXiv:1708.07747* (2017)

[Xiao et al. 2010] XIAO, Jianxiong ; HAYS, James ; EHINGER, Krista A. ; OLIVA, Aude ; TORRALBA, Antonio: Sun database: Large-scale scene recognition from abbey to zoo. In: *2010 IEEE computer society conference on computer vision and pattern recognition* IEEE (event), 2010, P. 3485–3492

[Yang et al. 2019] YANG, Guangyu R. ; JOGLEKAR, Madhura R. ; SONG, H F. ; NEWSOME, William T. ; WANG, Xiao-Jing: Task representations in neural networks trained to perform many cognitive tasks. In: *Nature Neuroscience* 22 (2019), Nr. 2, P. 297

[Yu and Zhao 2018] YU, Ru Q. ; ZHAO, Jiaying: Implicit updating of object representation via temporal associations. In: *Cognition* 181 (2018), P. 127–134

[Zeiler and Fergus 2014] ZEILER, Matthew D. ; FERGUS, Rob: Visualizing and understanding convolutional networks. In: *European conference on computer vision* Springer (event), 2014, P. 818–833

[Zeki 1976] ZEKI, SM: The functional organization of projections from striate to prestriate visual cortex in the rhesus monkey. In: *Cold Spring Harbor Symposia on Quantitative Biology* Vol. 40 Cold Spring Harbor Laboratory Press (event), 1976, P. 591–600

[Zhang et al. 2018] ZHANG, Mengmi ; FENG, Jiashi ; MA, Keng T. ; LIM, Joo H. ; ZHAO, Qi ; KREIMAN, Gabriel: Finding any Waldo with zero-shot invariant and efficient visual search. In: *Nature communications* 9 (2018), Nr. 1, P. 1–15

[Zhang et al. 2020] ZHANG, Mengmi ; TSENG, Claire ; KREIMAN, Gabriel: Putting visual object recognition in context. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, P. 12985–12994

[Zhang and Luck 2009] ZHANG, Weiwei ; LUCK, Steven J.: Feature-based attention modulates feedforward visual processing. In: *Nature neuroscience* 12 (2009), Nr. 1, P. 24–25

[Zhao and Yu 2016] ZHAO, Jiaying ; YU, Ru Q.: Statistical regularities reduce perceived numerosity. In: *Cognition* 146 (2016), P. 217–222

[Zhou et al. 2018] ZHOU, Bolei ; SUN, Yiyou ; BAU, David ; TORRALBA, Antonio: Revisiting the importance of individual units in cnns via ablation. In: *arXiv preprint arXiv:1806.02891* (2018)

[Zinchenko et al. 2020] ZINCHENKO, Artyom ; CONCI, Markus ; TÖLLNER, Thomas ; MÜLLER, Hermann J. ; GEYER, Thomas: Automatic guidance (and misguidance) of visuospatial attention by acquired scene memory: evidence from an N1pc polarity reversal. In: *Psychological Science* 31 (2020), Nr. 12, P. 1531–1543

Appendices

Appendix A

Samenvatting

Neem het visueel zoeken naar een interessant object. Als we niet weten waar dat object is, moeten we rondkijken totdat we dat object zien. We kijken echter niet lukraak om ons heen. Denk na over de procedure die we gebruiken bij het rondkijken. Ook al weten we niet waar het object is, we kunnen wel weten waar het object zou kunnen zijn. Als u bijvoorbeeld thuis naar een boek zoekt, ligt het hoogstwaarschijnlijk op tafel en niet op de vloer of het plafond. Deze kennis kan ons helpen veel zoeklocaties te vermijden, waardoor we sneller kunnen zoeken. Een andere procedures betreft hoe het object eruitziet. Het is gemakkelijker om weg te kijken van een locatie die een object bevat dat er heel anders uitziet (verschillende kenmerken) dan het object van belang.

Door gedragsexperimenten ontdekten onderzoekers dat dergelijke op kenmerken gebaseerde vergelijkingen tegelijkertijd in ons hele gezichtsveld kunnen plaatsvinden, zonder de expliciete noodzaak om onze ogen of onze "aandachtsschijnwerper" naar de verschillende objecten te verplaatsen. De snelheid van het zoeken naar een blauwe bal tussen rode ballen is bijvoorbeeld onafhankelijk van het aantal rode ballen in zicht. Het zoeken naar een blauwe bal tussen rode ballen en blauwe dozen is echter afhankelijk van het aantal ballen en dozen. Dit komt doordat we ons hele gezichtsveld kunnen vergelijken op basis van de kleur of de vorm, maar niet beide tegelijk. Dit proces van parallelle vergelijking wordt kenmerk-gebaseerde aandacht genoemd.

Eén van de vragen die in dit proefschrift aan de orde komt gaat over welke kenmerken van een dergelijke parallelle vergelijking kunnen profiteren. Hoewel kleur en vorm als kenmerk niet kunnen worden gecombineerd tot één kenmerk dat kenmerk-gebaseerde aandacht benut, ontdekten onderzoekers dat andere combinaties van kenmerken, zoals bewegingsrichting en kleur, kunnen worden gecombineerd tot een kenmerk dat kenmerk-gebaseerde aandacht kan benutten. Momenteel wordt gedacht dat combinaties van kenmerken die specifieke neuronen in de hersenen activeren kenmerk-gebaseerde aandacht kunnen benutten. Er zijn bijvoorbeeld geen neuronen die specifiek reageren op de aanwezigheid van een blauwe bal, maar er zijn neuronen die specifiek reageren op blauwe stippen die naar rechts bewegen. De vraag die we in hoofdstuk 2 van het proefschrift verkennen was of een andere combinatie van kenmerken die selectief neuronen activeren, zoals lichaamsvormen, ook kenmerk-gebaseerde aandacht benutten. We hebben ook beoordeeld of andere vormen, zoals auto's en lampen, eveneens kenmerk-gebaseerde aandacht gebruiken.

Voor deze beoordelingen gebruikten we functionele Magnetic Resonance Imaging (fMRI), een methode die wordt gebruikt om neurale activiteit af te leiden door de bloedstroom naar neuronen te registreren als reactie op een stimulus. Proefpersonen zochten naar de vorm van een lichaam en andere vormen in twee verticaal uitgelijnde dozen. We analyseerden de activiteit van neuronen die specifiek reageren op lichaamsvormen die werden gepresenteerd in twee horizontaal

uitgelijnde dozen. Toen deelnemers zochten naar lichamen, was de neurale reactie op lichamen hoger dan de neurale reactie op de andere vormen. Dit verschil in respons was groter tijdens het zoeken naar lichamen dan bij het zoeken naar de andere vormen. Deze resultaten suggereren dat zelfs op locaties waar lichamen niet gedetecteerd zouden moeten worden, lichamen gemakkelijker worden gedetecteerd indien lichamen ergens anders het zoekdoel zijn. Dit is een kenmerk van kenmerk-gebaseerde aandacht, waarbij lichamen de kenmerken zijn. Dergelijke eigenschappen werden niet gevonden voor de andere vormen. In Hoofdstuk 2 hebben we dus bewijs gevonden dat eerder onderzoek ondersteunt dat lichaamsvormen gebruik kunnen maken van kenmerk-gebaseerde aandacht bij het visueel zoeken van mensen. Of lichamen, naast gezichten, speciale vormen zijn die zeer ecologisch relevant zijn en daarom meer op kenmerken gebaseerde aandacht krijgen dan andere vormen zoals auto's of lampen, is een onderwerp voor verder onderzoek.

Naast op kenmerken gebaseerde aandacht die moduleert hoe de kenmerk-selectieve neuronen reageren op basis van het doel van visueel zoeken, is ook voorgesteld dat dergelijke modulatie kan optreden tijdens visuele verwerking. Objecten kunnen meerdere verschillende kenmerken hebben, op verschillende niveaus van visuele verwerking. Hoewel bedden en auto's verschillende vormen hebben, kunnen ze ook worden onderscheiden op basis van de algemene oriëntatie van hun randen - auto's hebben meer horizontale randen dan bedden. Dergelijke oriëntaties activeren specifiek neuronen in de vroege stadia van visuele verwerking de hersenen. Er wordt echter aangenomen dat de modulatie van neuronen in de vroege delen hogere metabolische kosten met zich meebrengt dan de modulatie van neuronen in de latere delen. Dit komt omdat het beginpunt van doelwit-gestuurde modulaties anatomisch (en conceptueel) dichter bij de latere delen ligt. Er rijst echter een vraag: wanneer wordt het essentieel dat de doelwit-gestuurde modulatie naast de latere delen ook op de eerdere delen van de visuele verwerking wordt gericht?

In Hoofdstukken 3 en 4 hebben we deze vraag behandeld met behulp van computationele modellering met kunstmatige neurale netwerken (ANN's) - een verzameling van algoritmen die zijn geïnspireerd op de netwerken van de hersenen. We redeneerden dat het filteren van informatie door modulatie in de eerdere delen niets zou toevoegen als de informatie die relevant is voor de huidige taak in de latere delen van de hiërarchie beschikbaar is. Aan de andere kant, als die informatie niet beschikbaar is, bijvoorbeeld omdat het netwerk te klein is voor de taak, d.w.z. een lagere capaciteit heeft, dan kan modulatie van de eerdere delen de relevante informatie naar de latere delen leiden, wat leidt tot betere taakuitvoering. Om deze hypothese te testen, hebben we een ANN getraind om aanwijzingsgebaseerde taken uit te voeren - het netwerk moest aangeven of een aangewezen object in een gegeven afbeelding aanwezig was. De aanwijzing werd ofwel geleverd als een modulatie van de late laag van het ANN of zowel van de vroege als de late laag van het ANN. De capaciteit van het ANN was een functie van het aantal neuronen en het aantal categorieën dat kon worden gecued. We ontdekten dat de modulatie van de vroege laag alleen nuttig was wanneer de capaciteit van de ANN laag was, dat wil zeggen, het netwerk had een laag aantal neuronen gezien het aantal categorieën. Deze modelleringsaanpak riep ook een aantal vragen op. Hoe worden bijvoorbeeld het basisnetwerk en de modulatie in de hersenen getraind: wordt de modulatie getraind nadat het basisnetwerk is getraind of worden ze samen getraind tijdens de ontwikkeling?

Een derde procedure die we gebruiken bij het zoekend rondkijken heeft betrekking op de relaties tussen objecten in de wereld. Objecten komen tegelijk voor in een ruimte en hebben semantische relaties. Een tv-meubel verschijnt bijvoorbeeld vaker bij een tv, en met de tv erop, dan bij een auto. Deze relaties beïnvloeden onze zoekprocedures. Bij het zoeken naar een tv in een wazige scène, geeft de aanwezigheid van een tv-standaard ons meer vertrouwen. In verband hiermee is voorgesteld dat gelijktijdig voorkomende objecten worden gegroepeerd, waardoor ze in feite één groot object in ons gezichtsveld worden. Een interessant gevolg van dit fenomeen zou

kunnen zijn dat omgevingen met gelijktijdig voorkomende objecten gemakkelijker te doorzoeken zouden kunnen zijn, aangezien het effectieve aantal objecten dat moet worden beoordeeld om het doelwit te vinden zou worden verlaagd als gevolg van groepering.

In Hoofdstuk 5 hebben we de mogelijkheid onderzocht om de efficiëntie van visueel zoeken te verbeteren bij de aanwezigheid van gelijktijdig voorkomende objecten. Deelnemers zochten naar aangewezen vormen in displays die ofwel gelijktijdig voorkomende (gestructureerde scènes) of niet-gelijktijdige afleidende vormen (ongestructureerde scènes) bevatten. De gelijktijdig voorkomende afleiders kwamen voor in paren van twee, waarbij de relatieve posities van de vormen binnen de paren vast of vrij waren (in afzonderlijke experimenten). Na enkele trainingssessies merkten we dat deelnemers zowel sneller als nauwkeuriger waren in het aangeven van de locatie van de cued-vorm in de gestructureerde scènes. We ontdekten ook dat dit voordeel bij het zoeken niet afhing van het feit of de relatieve posities van de vormen in de paren vast of vrij waren. Een verklaring op basis van groepering kan het voordeel bij het zoeken in de vrij geordende vormen niet rechtstreeks verklaren, tenzij we aannemen dat twee keer zoveel groepen werden geregistreerd zonder dat dit ten koste ging van het voordeel bij het zoeken. We hebben een alternatieve verklaring voorgesteld waarbij de relaties tussen gelijktijdig voorkomende vormen snel kunnen helpen elkanders vorm te verwerpen wanneer één van de vormen werd geïdentificeerd. Of een dergelijk proces parallel in het gezichtsveld kan plaatsvinden of sequentieel door de afleiders moet gaan, en hoe relevant een dergelijk proces kan zijn tijdens visueel zoeken in de echte wereld, zijn onderwerpen voor verder onderzoek.

In Hoofdstuk 6 hebben we de neurale processen onderzocht die ten grondslag liggen aan het waargenomen voordeel bij het zoeken, met behulp van elektro-encefalografie (EEG), een methode die wordt gebruikt om neurale activiteit af te leiden door elektrische potentialen op de hoofdhuid te registreren. We beoordeelden de verschillen tussen de gemiddelde elektrische potentialen vanaf het begin van gestructureerde en ongestructureerde scènes. Voor beide scènes zagen we ongeveer 200-400 ms na het begin van de weergave een grotere afbuiging over de hemisfeer tegenover de kant waar de vooraf aangewezen vorm aanwezig was dan de hemisfeer aan dezelfde kant als het doelwit. Dit verschil, de N2 posterieure contralaterale (N2pc) component genoemd, is een kenmerk van verhoogde aandachtsoriëntatie naar het doelwit. We ontdekten dat het N2pc-verschil tussen de gestructureerde en ongestructureerde scènes groter bleek te zijn bij deelnemers die een groter voordeel hadden bij het zoeken onder de gelijktijdig voorkomende afleiders. Dit effect suggereerde dat het waargenomen voordeel bij het zoeken kon worden toegeschreven aan snelle aandachtsoriëntatie, en niet aan latere besluitvormingsprocessen. Dit effect was echter zwak en moet met zorg worden geïnterpreteerd. Verdere experimenten, die gebruik maken van andere methoden (zoals fMRI), zijn nodig om de mechanismen te begrijpen die ten grondslag liggen aan het voordeel van gelijktijdig voorkomende afleiders bij het zoekproces.

In dit proefschrift hebben we, in de geest van computationele cognitieve neurowetenschap, een kijkje genomen onder de motorkap van de procedures die we gebruiken bij visueel zoeken, om beter te begrijpen hoe het menselijk brein diens machinerie gebruikt om die procedures uit te voeren. Deze onderzoeken omvatten het gebruik van sterk gecontroleerde, kunstmatige stimuli om de effecten te isoleren die van belang zijn bij het beantwoorden van de gestelde vragen. Zoals hierboven uiteengezet, zijn verdere experimenten echter essentieel om te weten of de waargenomen/gepostuleerde processen inderdaad aan het werk zijn tijdens onze zoektocht in de echte wereld. Dit proefschrift dient als een nieuwe kleine stap in de richting van het ontrafelen van de werking van het menselijk brein dat ten grondslag ligt aan al het verbijsterende gedrag dat wordt geproduceerd op deze aardbol.

(Translated from the English version by Charlotte de Blecourt)

Appendix B

Research Data Management

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

Ethics

This thesis is based on the results of human studies, which were conducted in accordance with the principles of the Declaration of Helsinki. The Ethical Committee of the faculty of Social Sciences (ECSS) has given a positive advice to conduct these studies to the Dean of the Faculty, who formally approved the conduct of these studies.

Findable, Accessible

Table B.1 details where the data and research documentation for each chapter can be found on the Donders Repository (DR), the Open Science Framework (OSF), and GitHub. All data archived as a Data Sharing Collection remain available for at least 10 years after termination of the studies. The data in the DSC have been shared with a CC0-1.0 Universal license, the data in OSF have been shared with a CC-By Attribution 4.0 International license, and the data in GitHub have been shared with a MIT license. Informed consent was obtained on paper following the Centre procedure. The forms are archived in the central archive of the Centre for 10 years after termination of the studies.

Chapter	**DAC**	**DSC**	**OSF**	**GitHub**
2	2019.00052_614	2019.00052_491	HJ5VC	-
3	-	-	-	novelmartis/cue-feedback-ccn18
4	-	-	-	novelmartis/early-vs-late-multi-task
5	2020.00049_924	-	EM2XF	-
6	2020.00111_120	-	-	-

Table B.1: **Data collections for each chapter.** DAC = data acquisition collection. DSC = data sharing collection. OSF = Open Science Framework.

Interoperable, Reusable

The raw data are stored in the DAC in their original form. For the DSC long-lived file formats have been used ensuring that data remains usable in the future. The data of the DSC are organized according to the BIDS standards, with concomitant readme files. Results are made reproducible by providing a description of the experimental setup, raw data (DAC and DSC), and analysis scripts/pipelines (DSC, OSF, and GitHub). The used software with their version numbers are specified in the mentioned repositories.

Privacy

The privacy of the participants in this thesis has been warranted using random individual subject codes. A pseudonymization key linked this random code with the personal data. This pseudonymization key was stored on a network drive that was only accessible to members of the project who needed access to it because of their role within the project. The pseudonymization key was stored separately from the research data. The pseudonymization keys were destroyed within one month after finalization of these projects.

Appendix C

Acknowledgement

I haven't made up my mind on how the overall experience of working through the Ph.D. was. There were always a few days of intense thinking about new questions, reframing old questions, designing experiments, and coming up with appropriate ways of analyzing the data, followed by a multitude of days just doing the grind work and wondering what it all was leading to anyway. Only at the end, when I wrote a summary for the possible thesis, did the amount of work and the unification of the findings feel like it led to something of substance - a thesis wandering through the rugged landscape of research into human visual search. Looking back at my journey, I find this landscape untraversable without the transcendent guiding figures along the way. Here I reminisce about them.

The MVP among those figures was, of course, my advisor Marius Peelen. He has played the biggest role in shaping my ability to science systematically - probably the most important aspect of a Ph.D. I also have been constantly amazed and inspired by his ability to look through the fluff and understand experimental designs and conceptual details rapidly. I also am grateful for the amount of patience he has had dealing with my musings.

In all phases of my academic journey, I have had a companion who I have bothered with all the tangents my thoughts could take. I cannot imagine how I could come up with all my research ideas without those tangents. This companion, possibly majorly troubled by my constant pinging, is extremely important to my journey. During the Ph.D. this place was occupied by Giacomo Aldegheri. From consciousness to mind uploading to the theory of everything to artificial general intelligence to finally coming up with scientifically approachable questions that led to two publications during the Ph.D., Giacomo has been an indispensable colleague and friend.

All the members of my lab - Surya, Genevieve, Lu-Chun, Jochem, Sjoerd, Charlotte, Yuanfang, Jorie, Alexandra, Maelle, Marco, Simen, Qiu, Chuanji, Miles, Myrthel, and Linlin - and other members at the Donders Institute - Johannes, Gabi, Erdi, Micha, Nadine, Jordy, Nasir, Nestor, and Sara - thank you for making the workplace an intellectually-stimulating and caring place. In the same vein, I would also like to thank the support staff - Jolanda, Vanessa, and Karin.

The final year and half of my Ph.D. were immersed in the pandemic. Beyond academia, life would be impossible to sustain without my family in Nijmegen. It is funny that the family being large followed me from India to Nijmegen. I won't even attempt to name everyone for fear of getting kicked later for forgetting about someone. However, I have to highlight that the journey through those pandemic years would have been very hard without my flatmate, Elena Mainetto. Saturday morning rituals, going shopping around Nijmegen through the markets and having lunch with the huge gang, which sometimes included random friends picked off the street and even Tugay who sold us Poffertjes, were all I could've asked for to end the tiring weeks.

I never thought my academic journey was ever hard, all the way from school to the end of my Ph.D. This is thanks to my support structures - surreal friendships and my parents. “Bhiu Nakos Mi Tuzya Pathishi Aahe” (Don’t fear, I’ve got your back) is what I hear whenever I need it, thanks to these folks. Stephanie, Tanvi, Lu-Chun, Abhijit, Giacomo, Charlotte, Digvijay, David, Shriya, Elena, and my dearest Zhazha - thanks for checking on me from time to time and making me realize how lucky I am to have amazing people like you care for me. Mom, Dad, and Shruti - I know you will always be there and that’s more of a relief than you might realize.

Lastly, I would like to express my gratitude to all the support structures in place at the Donders Institute and Radboud University at large - the supervision team, the Donders graduate school, and the thesis review team - for helping me through this worthwhile journey. Thank you all!

Appendix D

About the Author

Sushrut Thorat was born in Kolhapur on 15 February 1994. He completed his schooling at KCT's Krishna School in Karad. After a brief time in Kota, he completed a B.Tech. in engineering physics at the Indian Institute of Technology (IIT-B) in Mumbai, and an M.Sc. in cognitive neuroscience at the Centre for Mind/Brain Sciences (CIMeC) in Rovereto.

During his B.Tech., Sushrut's interests steadily pivoted away from fundamental physics to the consideration of the nervous systems' dynamics from the viewpoint of complexity theory. Initially, he dabbled at the level of small neural circuits: one, in his thesis project involving spiking neural networks for quadcopter flight control, which led to a publication; two, at the Computational Approaches to Memory and Plasticity (CAMP) summer school at the National Centre for Biological Sciences (NCBS) in Bengaluru. Exposed to the complexity generated by the simplest neural networks, he decided he needed to study that complexity from the other, top-down, side - through the cognitive sciences. He, therefore, enrolled in the M.Sc. program in cognitive neuroscience at CIMeC, also in a small part because their brochure showed a picture of Lake Garda which, a bit sadly, turned out to be 30 minutes away by car.

At CIMeC, in addition to exploring the field of cognitive neuroscience, Sushrut explored the emerging field of deep learning, which required reasoning at all levels, from neural circuits to emergent behavior. He was drawn into both visual information processing and language understanding and production. During his holidays preceding the M.Sc., he had started a project with a school friend on building simple algorithms for relating phrases to single-word summaries - a reverse dictionary - which eventually led to a publication. During his M.Sc., after dabbling in various projects, he started working with Marius Peelen on using deep convolutional neural networks (CNNs) for studies requiring expressive models of visual processing as alternatives to higher-order models such as semantic processing (in this case, information about the animacy of objects). This research was conducted amidst the very first wave of papers comparing CNNs and the primate visual system. It led to a thesis that led to a cum laude graduation and a publication.

In 2017, Sushrut started his Ph.D. with Marius at the Donders Centre for Cognition, aiming at furthering our understanding of human visual search processes through neuroimaging and computational modeling. With the use of fMRI, EEG, large-sample online behavioral experiments, and artificial neural networks, his research corroborated previous theories and findings and led to new insights on how the visual system deploys feature-based attention and how it uses the regularities in the environment to optimize its search operations, leading to four publications. Additionally, he performed several side projects, with colleagues from the lab, such as studying the influence of scene context on object recognition, predictive and task-dependent visual processing, and understanding the internal operations in recurrent neural networks, many of which are being written

up for publication or are already published. He has presented his findings at several local and international conferences (his work has won awards at three).

Aside from his research activities, Sushrut was involved in teaching during his Ph.D. He acted as a teaching assistant for various undergraduate and graduate courses such as 'Advanced Academic & Professional Skills', 'Neural Networks', and 'Brain for AI'. He supervised the research activities of many bachelor's students and one master's student, during their thesis projects. These activities reflect his passion for educating the upcoming generation of researchers in neuroscience and artificial intelligence.

After his Ph.D., Sushrut will continue his academic journey as a postdoctoral researcher at the Institute of Cognitive Science at the University of Osnabruck. In terms of his research focus, he will pivot to the field of machine learning and focus on building artificial agents that can learn continually and function flexibly in the world as biological creatures do, in close collaboration with Tim Kietzmann.

List of Publications

Thorat, S., & Rajendran, B. (2015). *Arithmetic computing via rate coding in neural circuits with spike-triggered adaptive synapses.* In 2015 International Joint Conference on Neural Networks (IJCNN) (pp. 1-8). IEEE.

Thorat, S., & Choudhari, V. (2016). *Implementing a Reverse Dictionary, based on word definitions, using a Node-Graph Architecture.* In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers (pp. 2797-2806).

Thorat, S., van Gerven, M. A. J., & Peelen, M. V. (2018). *The functional role of cue-driven feature-based feedback in object recognition.* In Proceedings 2018 Conference on Cognitive Computational Neuroscience (pp. 1-4).

Thorat, S., Aldegheri, G., van Gerven, M. A. J., & Peelen, M. V. (2019). *Modulation of early visual processing alleviates capacity limits in solving multiple tasks.* In 2019 Conference on Cognitive Computational Neuroscience (pp. 226-229).

Thorat, S., Proklova, D., & Peelen, M. V. (2019). *The nature of the animacy organization in human ventral temporal cortex.* Elife, 8, e47142.

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

Thorat, S., & Peelen, M. V. (2022). *Body shape as a visual feature: evidence from spatially-global attentional modulation in human visual cortex.* NeuroImage, 255, 119207.

Thorat, S., Quek, G., & Peelen, M. V. (2022). *Statistical learning of distractor co-occurrences facilitates visual search.* bioRxiv. *in print* at the Journal of Vision.

Appendix E

Donders Graduate School for Cognitive Neuroscience

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

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

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

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

DONDERS INSTITUTE

Max Planck Institute for Psycholinguistics

Radboud University

Radboudumc