

**Master Thesis**

*Research Master Cognitive Neuroscience,*

*Donders Graduate School,*

*Radboud University*

Searching Near and Far:

*Investigating Depth-dependent Adaptation of  
Search Template Size in Naturalistic Visual Search*

---

Maëlle Lerebourg

Advisors: Dr. Marius Peelen

Dr. Surya Gayet

2<sup>nd</sup> reader : Prof. Dr. Rob van Lier

Submitted: 19<sup>th</sup> August 2020

Defence : 26<sup>th</sup> August 2020

## Abstract

---

Current theories of visual search assume we create a template representing the target object by pre-activating neurons tuned to target features. When searching in naturalistic scenes, visual features of the target may however change drastically depending on its location in the scene, e.g. its retinal size depends on its distance. To account for this, the template may be rescaled based on depth. In a first experiment, we used breaking continuous flash suppression (b-CFS) we aimed to probe the template formed in a search task requiring participants to take into account depth-dependent size changes and test whether size-matching probes were detected faster. Suppression times to probes were however generally not modulated as a function of their match with target features. Using fMRI and MVPA we investigated the neural basis of the search template, testing whether the expected retinal size of objects participants prepared to search for could be decoded from LOC and whether depth-information from scene-selective areas modulated template size. In line with our hypotheses, we found overlapping voxel activation patterns for seeing objects of varying retinal and preparing to search for these objects near or far within LOC. This effect was however not specific to the search task. While distance-information based on low-level features may have contributed to size or depth-processing in LOC no evidence for a contribution of depth-information from scene-selective areas was found. While further research is needed to understand what specific mechanisms our findings in LOC reflect, these likely still contribute to our ability to account for changes in visual features during search.

## Introduction

---

Throughout most of our waking lives, our visual environment is highly complex and consists of a multitude of objects. We are seldomly passive observers, but instead have specific goals (e.g. wanting to go home) or questions regarding our environment (e.g. “where is my bike?”), that dynamically shape the immediate relevance of the different objects surrounding us. To locate relevant objects (our bike in this case) among the abundance present in our visual field and competing for processing resources, we engage in visual search. This entails selectively processing only parts of the momentary visual input, on the basis of their match with top-down goals.

Searching our environment requires a description of what to look out for (e.g. the bike with red flowers on the handlebar). All influential theories of visual search therefore state that search preparation involves creating a representation of the search target, commonly referred to as search- or attentional

template. Processing of incoming visual input is subsequently biased in favour of objects matching the template over non-matching input. Competition between different objects present within our visual field can thereby be resolved and processing resources efficiently allocated to likely search targets (Duncan & Humphreys, 1989; Eimer, 2014; Kastner & Ungerleider, 2001; Wolfe, 1994; Wolfe & Horowitz, 2004). When searching for our bike with the red flowers, we may e.g. be distracted by another bike with a red bell but fail to notice a friend wearing a blue pullover waiving at us.

On a neural level, the search template is likely instantiated by selective pre-activation (increase in baseline firing rate) of those neurons in visual cortex tuned to the target’s features as e.g. its shape, colour or size. Once visual input, potentially representing the search target, is presented within their receptive fields, target-matching input becomes more effective in driving the neuron’s response, at the expense of non-matching input (Desimone & Duncan, 1995). Such preparatory activation has been found both in

monkey inferotemporal cortex (IT) (Chelazzi, Duncan, Miller, & Desimone, 1998; Chelazzi, Miller, Duncan, & Desimone, 1993; Desimone, 1998), as well across the human ventral visual stream and features of varying complexity (see Battistoni, Stein, & Peelen, 2017 for a review). These findings provide important empirical evidence for template-based theories of visual search.

#### *Visual search in naturalistic scenes*

To date, much of the research and evidence on visual search is however still based on rather simplistic stimuli. In many experiments, targets and distractors are defined by relatively low-level features and the objects bear no meaningful spatial or contextual relationship with each other. One example of such a laboratory task would be searching for a red horizontal bar presented in a search array, among green horizontal and red vertical bars. This is clearly different from most real-life situations, where we are searching for more complex and visually ill-defined targets with apparent ease. Despite this, searches for objects in naturalistic scenes are often more efficient than in these artificial settings (Wolfe, Alvarez, Rosenholtz, Kuzmova, & Sherman, 2011).

The efficiency and relevance of real-world search has led to growing interest in search within naturalistic scenes. In contrast to simple search arrays, the relation between objects and the real-world scenes in which they are placed is meaningful and the visual system can make use of these regularities for object identification (Bar, 2004). A growing number of studies have investigated how scene-based information can constrain search. The general *gist* of a scene (coarse information about e.g. scene category such as indoor vs. beach and its basic spatial layout) can be extracted within around 100 ms and used to inform search, e.g. by guiding attention and eye movements towards likely target locations (e.g. Castelhamo & Henderson, 2007; Eckstein, Drescher, &

Shimozaki, 2006; Ehinger, Hidalgo-Sotelo, Torralba, & Oliva, 2009; Neider & Zelinsky, 2006; Wolfe, Võ, Evans, & Greene, 2011)

Another line of research has focused on the content of the search template for natural category-level search (e.g. for people or cars) in complex natural scenes (Reeder & Peelen, 2013). Template-related preparatory activity for such targets, less well defined in terms of simple features, was found in object-selective areas as LOC (Peelen & Kastner, 2011).

#### *Depth-dependent size changes in naturalistic scenes*

There are however still many open questions regarding search in naturalistic settings and especially how constraints imposed on the target by the scene interact with the search template. An often-overlooked issue is that the target's appearance may change drastically depending on where it is located within and across scenes, due to various factors as lighting or viewing angle. In that sense, scenes do not only determine an object's likely location, but also likely appearance and specific visual features given that it is placed at a particular location. One important factor in that regard is that for any object, its retinal size is inversely proportional to its distance from the observer. The very same object placed nearby will thus have a larger retinal size compared to when it is far away, leading to vastly different retinal images. On the other hand, objects with different physical sizes may have the same retinal size when placed at different distances. This creates an obvious problem for the visual system during search, as it needs to account for the specific scene context and position an object is placed in to determine whether it is the target (Gayet & Peelen, 2019).

Current theories of visual search do not generally address how the visual system either creates a template or matches it to such varying visual input. Essentially, whether or how these template-based theories can indeed

reflect an ecologically plausible account of visual search thus remains an important empirical question.

When looking at human search behaviour, it seems clear that the relation between an objects size and location is generally taken into account. Intuitively, we would not be strongly distracted by a nearby bonsai when searching for a full-grown tree. Indeed, human observers may sometimes fail to detect even giant targets if their size is incongruent with their surroundings, such as a toothbrush spanning the entire width of the bathroom sink (Eckstein, Koehler, Welbourne, & Akbas, 2017). Note also that, in many real-world scenarios, accounting for scene context, and thereby constraining search to objects of realistic size, will improve search performance. Since most objects have canonical sizes and the relation between an objects distance and retinal size is entirely predictable, attending only to congruently-sized objects while ignoring others represents less of a bug than a feature, likely contributing to search efficiency in naturalistic scenes (Wolfe, 2017).

#### *Aim of the current thesis*

Given this behavioural evidence, how may template-related mechanisms account for depth-dependent size changes during real-world search? One potential solution is to rescale the search template based on where in depth we are currently searching (Gayet & Peelen, 2019).

Therefore, the overarching aim of this thesis is to investigate (1) whether we are indeed searching for larger objects (in terms of retinal size) when searching nearby compared to far-away and, if that is the case, also (2) which neural mechanisms are underlying integration of scene context with the template during search and subsequent template-rescaling.

We conducted two different experiments to answer these questions. In both we used a cued visual search task asking participants to

search for target objects (melons and boxes, embedded in naturalistic scenes) at two different distances/depths (near or far).

Before investigating any neural representation of the search template, we first designed a behavioural paradigm requiring participants to take into account depth-related changes in retinal size and aimed to probe the content of the search template created using breaking continuous flash-suppression (b-CFS, Experiment 1). In a b-CFS paradigm, one eye is presented with a high-contrast dynamic mask while the other eye is presented with a probe image of increasing intensity (Jiang, Costello, & He, 2007). The time it takes for this probe image to be released from interocular suppression by the mask and become reportable can then be compared for different stimulus classes or on their basis of their match with other consciously accessible information (see Gayet, Van Der Stigchel, & Paffen, 2014; Stein, 2019 for reviews). If the probe image matches the content of visual working memory, suppression times are generally shorter (Gayet, Paffen, & Van der Stigchel, 2013; Gayet, van Maanen, Heilbron, Paffen, & Van der Stigchel, 2016; Liu, Wang, Wang, & Jiang, 2016; Pan, Lin, Zhao, & Soto, 2014). This method has not yet been applied to study search templates in naturalistic scenes. Holding online a search template however presumably relies strongly on visual working memory (Carlisle, Arita, Pardo, & Woodman, 2011; Desimone, 1995; Gunseli, Meeter, & Olivers, 2014). There is also further evidence that these b-CFS working memory effects reflect selective increase in baseline firing rate of neurons tuned to the memorandum, similar to the mechanisms thought to underlie template-based visual search (Gayet et al., 2016). We therefore reasoned that template-matching probes would similarly show reduced suppression times compared to non-matching probes. Specifically, we hypothesized that if search template is indeed rescaled based on

depth, size-matching probes should be released from interocular suppression earlier and detected faster than probes of incongruent size.

In a second experiment, we used fMRI and multivariate pattern analysis (MVPA) to investigate the neural basis of the search template created in this search task and test whether retinal size information is used during search preparation.

We focused our analyses of template-activity on the object-selective lateral occipital complex (LOC), an area previously found to encode search templates for complex shapes (Soon, Namburi, & Chee, 2013) and category-level search within naturalistic scenes (Peelen & Kastner, 2011). While object representations in LOC typically encode real-world rather than retinal size (Konkle & Oliva, 2012; Sawamura, Georgieva, Vogels, Vanduffel, & Orban, 2005), TMS or lesions to LOC can also hinder processing of retinal size (Chiou & Ralph, 2016) and impair integration of retinal size and depth information to correctly estimate physical size (Cohen, Gray, Meyrignac, Dehaene, & Degos, 1994; Zeng, Fink, & Weidner, 2020). This all suggests neurons in LOC can process an object's retinal size and that this information may be usable during search.

In addition to LOC, we also tested for template-related activity in early visual cortex (EVC), comprising V1 and V2. Neural representations in these early sensory areas are generally less invariant and clearly modulated by factors as retinal size. In terms of efficient processing, it is also evidently beneficial to identify and select likely targets and the earliest possible stage. However, given the variability of target appearance in real-world search, a template held in such an early sensory regions may not be well suited to distinguish between targets and non-targets and effectively even hinder search in naturalistic scenes as used in our task (Peelen & Kastner, 2011).

Rescaling the template based on depth would also require rapidly acquiring information about the scene's depth layout. Behavioural evidence suggests this can be done remarkably fast and efficiently, with less than 50 ms of stimulus exposure needed for humans to reliably extract information about global distance-related properties such as a scene's mean depth or presence of bounding elements in the scene (Greene, Michelle & Oliva, Aude, 2009).

Such distance-related information is likely processed in visual areas devoted to the processing of scenes, as the parahippocampal place area (PPA, R. Epstein & Kanwisher, 1998) and occipital place area (OPA, also called transverse occipital sulcus (TOS), (Grill-Spector, 2003)). These areas are sensitive to various distance-related aspects of a scene's spatial layout, such as perceived distance to objects (Amit, Mehoudar, Trope, & Yovel, 2012), arrangement and presence of spatial boundaries (Ferrara & Park, 2016; Henriksson, Mur, & Kriegeskorte, 2019; Kamps, Julian, Kubilius, Kanwisher, & Dilks, 2017; Kornblith, Cheng, Ohayon, & Tsao, 2013) and in case of the PPA also by objects' retinal size (Konkle & Oliva, 2012). The timecourse of depth-processing in these areas is not very well investigated, but recent work suggests a representation of scenic layout emerges relatively early (after around 100 ms) and is based on the feedforward sweep of visual information, especially in OPA (Bonner & Epstein, 2017a; Henriksson et al., 2019).

Once coarse information about distance in the scene is extracted and processed in scene-selective regions, this may shape the size of the template held in object selective areas. In the current experiment, we therefore also analysed encoding of depth in PPA and OPA and related it to search template size in LOC. If depth information encoded in scene-selective areas modulates the size of the template, held

in object-selective regions, we would expect both to correlate on a trial-by-trial basis.

Together, the two experiments investigate a potential mechanism for template-based

## Experiment 1: breaking continuous flash suppression (b-CFS) task

---

In the first experiment, participants performed a difficult cued search task, searching for objects of two different categories (boxes and melons) either far away or nearby in scene photographs. The task required them to take into account depth-related changes in retinal size, as distractor objects with the same shape, but different size as the targets, could also appear in the same depth plane. To probe the template formed by participants, we combined the search task with a b-CFS paradigm. We introduced a delay period of unpredictable duration, during which participants had to hold in mind the category and expected retinal size of the target, while dynamic masks were presented and after which the search scene would appear briefly in half of the trials. In the other half of the trials, a probe image of a small or large melon or box was shown in the delay period (thus matching or mismatching the current search target in shape, size or both), intraocularly suppressed by the masks and to which participants had to respond as fast as possible. Importantly, both trial types were intermixed, such that participants could only determine which task they had to perform after either perceiving the search scene or the b-CFS probe and had to prepare for the search task in all trials. We hypothesized that, if participants adjusted their template in size based on where they were searching, size-matching probes and especially those for which

search in the real world, taking into account the demands naturalistic scenes impose on our visual system

both shape and size matched the target, would overcome interocular suppression and be detected faster than non-matching ones.

## Methods

### Participants

35 participants (23 females, mean age: 24 (sd = 3.68)) constitute the final sample for this experiment. Based on a power analysis, we had chosen to test until reaching a total of 34 participants with above-chance performance (allowing to find a medium-sized effect with 80 % power) and exceeded this number by one participant. A total of 56 participants (39 females, mean age: 24 (sd = 5.43)) took part in the first experiment, but only those performing above chance in the search task (as determined by a one-sided binomial test with an alpha-level of 0.05) were included, as their performance necessarily indicates successful search preparation.

Participants were recruited from the Radboud University participant pool (SONA Systems). They participated for either monetary reward (10 €) or course credit and provided written informed consent prior to the experiment. All had normal or corrected to normal vision and reported themselves free of epilepsy.

### Procedure

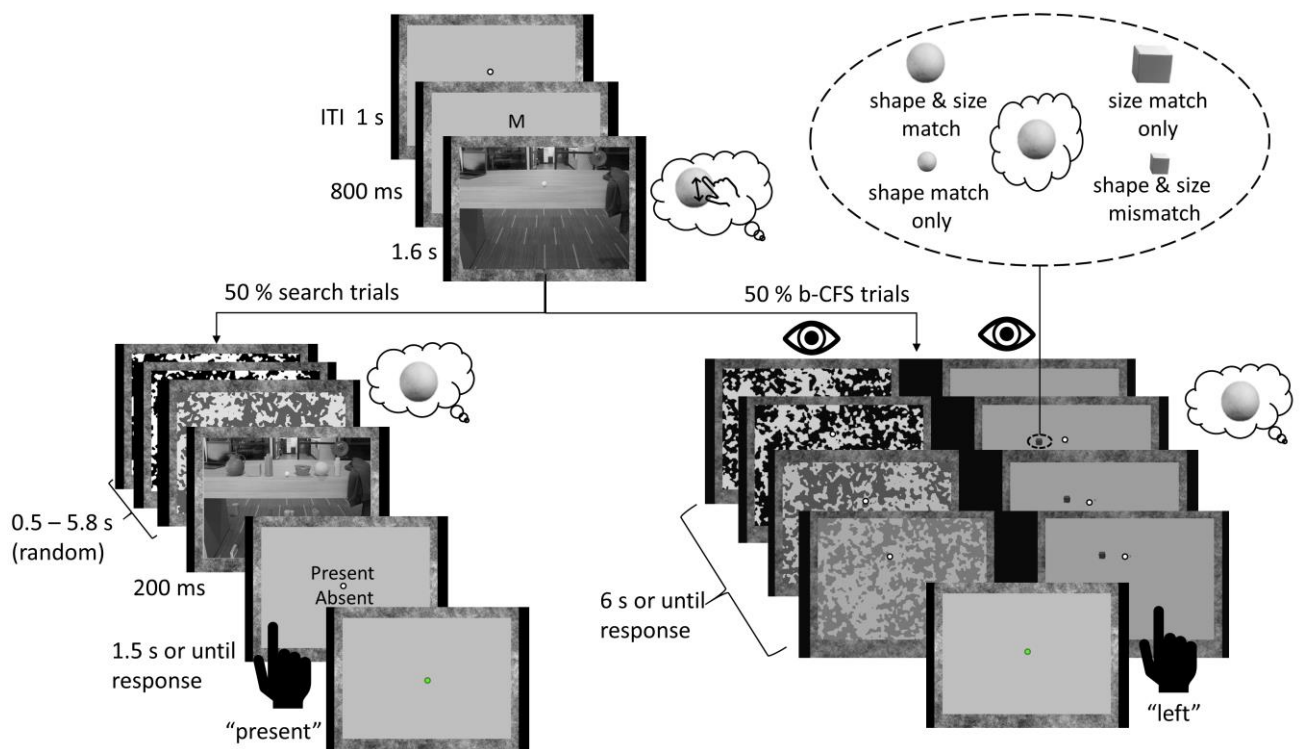
When coming into the lab, participants were first familiarized with the different conditions of the task using a click-through demo and practice trials. During this training, they were also shown the specific box depicted in the search scenes, to have an estimate of its real-world size. Thereafter, participants completed a total of 256 trials, broken up into 16 blocks of

16 trials, lasting 2-3 minutes each. Participants could take a break after every block and the entire experimental session lasted around one hour.

Each trial began with the presentation of a letter-cue (M or B for melon or box), indicating the current search target. This letter was presented either above or below the centre of the screen, requiring participants to make an up - or downward eye movement to fixate it. This was followed by the presentation of an empty search scene for 1.6 s, providing depth information. Due to the previous eye movement, participants were now fixating the centre of the currently relevant depth plane (either nearby or far away). They were

explicitly told they could use this information and prepare to search for a larger object when searching nearby compared to far away, as distractor objects in the search scenes may share shape or size with the search target but never both. The near depth plane could be either in the lower or upper visual field (see “near below” and “near above” scene types in Stimuli & Setup). Therefore, the scene preview, but not the direction of the earlier eye movement, were predicting the retinal size of the target object (with 100% validity).

The scene preview was then replaced by dynamic visual masks, initiating a delay period during which participants had to keep actively preparing to search for the current target by



**Figure 1:** Timeline of a trial in the b-CFS experiment. Each trial began with the presentation of a category cue, indicating whether participants had to search for either a melon or box. Thereafter, an empty scene provided information about the target’s expected retinal size and participants fixated within one of two depth-planes. Both size and shape information needed to be maintained over a delay period of random duration during which dynamic masks of gradually decreasing opacity were presented and after which a search scene would appear in 50% of trials. In other trials (b-CFS trials), which were randomly intermixed with search trials, a probe image matching size, shape, both or neither of the current target was presented to one eye, interocularly suppressed by the mask. Participants had to report which side of the fixation dot this probe image appeared as soon as they detected it. Reaction time to these probes were analysed as a function of their match with the target to probe the search template created by participants. Both eyes were always stimulated individually using a stereoscope but individual input to each eye only shown here when differing between the two eyes.

holding in mind both its retinal size and shape. Subsequent visual stimulation and task differed between the two possible trial types (b-CFS and search, see Figure 1). Trial type, search target category and depth plane were counterbalanced across groups of two blocks. Scene type (“near below” or “near above”) was counterbalanced within the whole experiment and the prevalence of specific scenes (taken from 1 out of 8 scene families per scene type) equated as much as possible, but not fully counterbalanced.

#### *Search trials*

In the search trials, masks of gradually decreasing opacity were presented to both eyes. The duration of the delay period was randomly drawn from a uniform distribution between 0.5 and 5.8 s, requiring participants to be prepared for search at all times. After the mask stimuli were switched off, the search scene, containing the target in half of trials, was briefly presented for 200 ms. Participants were then asked to report whether the target had been present or absent by pressing the up/down arrow keys. Instructions stressed accuracy rather than speed, but participants were made aware that they had to provide an answer within a timeframe of 2 s. Participants then received feedback on their performance in this trial through a colour change of the fixation dot and the next trial began after an intertrial interval (ITI) of 1s. Feedback on general search accuracy within a block was presented at the end of each. Search accuracy was staircased to an upper bound of 65 % correct using an Accelerated Stochastic Approximation (ASA) staircase algorithm. If performance exceeded this threshold, grayscale pixels of pink (1/f) noise were blended into the search scene with an alpha-value controlled by the staircase. Together with the short presentation time of the search scene and unpredictable time of scene onset, this ensured active search preparation was required to succeed.

Target presence and distractor type (sharing either shape or retinal size of the target) were counterbalanced across groups of two blocks. The side on which the target appeared was counterbalanced within the experiment.

#### *B-CFS trials*

B-CFS trials were identical to search trials up until the delay period. During this period, only one eye was presented with the visual masks (refreshed at 10Hz), while the other eye was presented with a probe picture either matching or mismatching the current search target in size and/or shape. More specifically, when searching for e.g. a small box, the probe could also be a small box (shape & size match), a large box (shape match only), a small melon (size match only) or a large melon (shape & size mismatch).

The probe could appear on either the left or right of the fixation dot, at the same locations at which target objects would appear in the search scene. Participants had to report the side of fixation on which the probe appeared as fast as possible using the left and right arrow keys. The probe image was presented from mask onset and gradually increased in intensity over a period of 1 second, in order to minimize immediate release from interocular suppression by abrupt stimulus onset. Mask opacity was ramped down starting after 1 s and over a period of 4 seconds. One b-CFS trial had a maximal duration of 6 s, with the probe being presented without any masking during the last second. B-CFS stimulation ended after the subject either responded to the probe or this maximal delay was reached. Feedback was again given after each trial and no search scene appeared after the delay period.

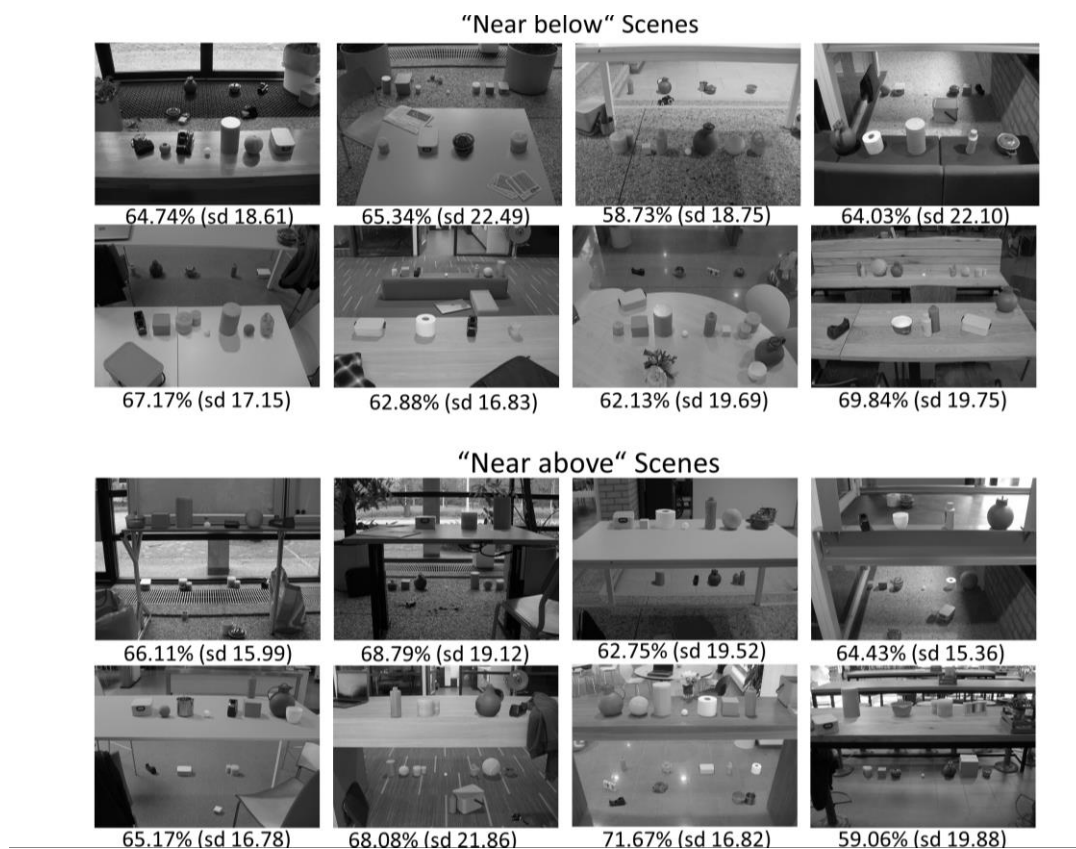
Probe object category and size were counterbalanced across groups of two blocks. Eye and hemifield (left or right side of fixation) for probe presentation were counterbalanced within the experiment. For each combination of object category and size, three different

probe images were used (cut out from different scene images). Their prevalence was equated but not fully counterbalanced within the experiment.

### Stimuli & Setup

Stimuli were presented using a BenQ XL24040Z monitor with a native resolution of 1920 x 1080 pixels and refresh rate of 120 Hz. Individual stimulation of each eye was achieved using a mirror stereoscope. A chinrest kept the effective viewing distance at 61 cm. To facilitate binocular fusion, all stimuli were surrounded by a frame of Brownian ( $1/f^2$ ) noise. Masks and scene images subtended  $20 \times 14.434^\circ$  visual angle. Experimental scripts for stimulus presentation and data acquisition were coded using Matlab and Psychtoolbox (Brainard, 1997).

The search scenes were grayscale photographs taken on Radboud university campus, depicting everyday objects (including the melon and box target objects) arranged on two depth planes, at equal distance from the image centre. To avoid confounding our depth-related effects of interest with any general effects of stimulus presentation in either upper or lower visual field, the near depth plane was in the upper half of the image in 50% of the scenes (“near above” scenes) and in the lower half for the other scenes (“near below scenes”) (see Figure 2). “Near above” scenes were created by placing near objects on an elevated surface (e.g. a table), far objects below and further away as well as varying the camera angle. For each scene type, photographs were taken at eight different locations, resulting in a total of 16 distinct scene families.



**Figure 2:** Example search scenes and search accuracy for each of the 16 scene families. A small white ping -pong ball indicates the centre of the currently relevant depth plane. Accuracy values are averaged separately for near and far within participants.

In target present scenes, the search target was placed at one of four different locations (either left or right in the near or far plane). Retinal size of target objects (1.5 and 3° respectively) and the two possible locations at which they could appear within each depth plane (3,27° eccentricity) were kept equal across scenes. Near targets were always twice as large in terms of retinal size compared to far ones and the retinal size of box and melon targets equal. If the target was placed on the left side, either a size or shape distractor object was placed on the right at the eccentricity and vice versa. Shape distractors had the same shape as the search target, but the wrong retinal size given this depth plane (e.g. a large basketball far away with the same retinal size as the melon in placed in the near plane or vice versa). Size distractors shared the retinal size of the target object but had the shape of the other target object (e.g. a small ball when searching for a small box). Distractor and target objects were both flanked by other objects on each side to induce visual crowding.

Target absent scenes were identical to target present scenes except that they contained both a shape and size distractor in order to fill the two potential target locations within a depth plane. Scene previews showed the exact same scenes without objects in either depth plane.

B-CFS probe stimuli were created by cutting out the target objects from different scenes. These isolated objects were placed on a grey background, at the same position at which targets could appear in both depth planes. To avoid any effects of pixelwise correlations between mask and probes, different probe images were used for every combination of target object and size. A total of 12 different probe images was used (3 for both objects of each retinal size).

Black and white noise masks were created by smoothing pink (1/f) noise with a Gaussian filter ( $\sigma = 3$ ) and binarizing the resulting image.

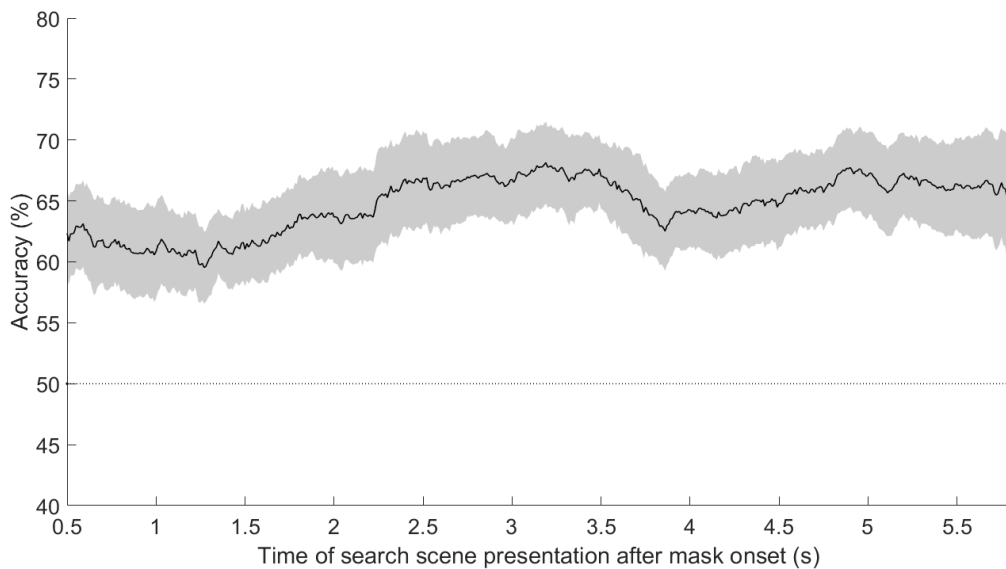
## Results

### Search Task Trials

Participants overall accuracy in the search task was 64.54% (sd 4.85), reflecting the threshold set by the staircase. To analyse whether this accuracy differed between conditions, we conducted a 2x2x2 repeated measures ANOVA with factors Depth Plane (near, far), Scene Type (“near below”, “near upright”) and Target Object (melon, box). There was a significant effect of depth plane ( $F(1,34) = 9.37, p = 0.004$ ), reflecting that participants were better at searching nearby (i.e. for larger objects) compared to far away (near: 66.81% (sd 5.53), far: 62.25% (sd 7.40)). Accuracy did not differ between scene types ( $p = 0.08$ ) or target objects ( $p = 0.20$ ) and there were no significant interactions between factors (all  $p$ 's > 0.41).

Accuracy did also not differ between individual scenes, as indicated by a second repeated measures ANOVA with single factor Scene ( $F(1,34) = 1.28, p = 0.21$ ) (see Figure 2 for accuracy for individual scenes), suggesting participants were able to successfully extract depth-information from each of them.

Given the long delay period during which the search scene could possibly appear, we also asked whether participants were able to sustain active search preparation across this entire time range. We computed accuracy as a function of search scene onset using a rolling time window of 1s within each participant (see Figure 4). We then fitted a linear model to test for changes in accuracy over time, which indicated a small increase (slope: 1.02,  $p < 0.00001$ ). Importantly however, accuracy was above 50% at every timepoint (all  $p$ 's < 0.0001 after threshold free cluster enhancement (tfce) for multiple comparison correction).

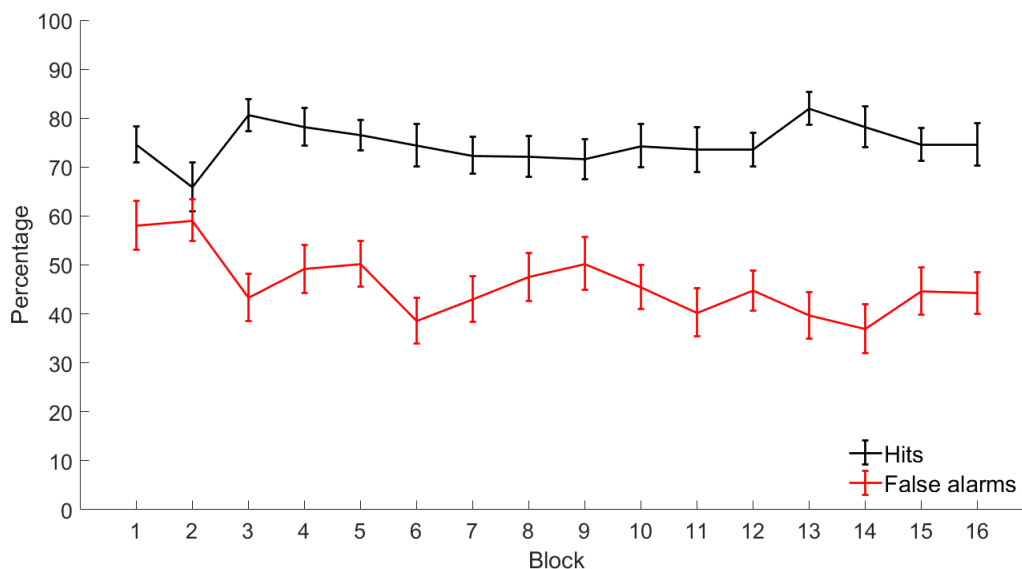


**Figure 4:** Search accuracy across the delay period as function of mask-search scene onset asynchrony. Average accuracy over time was computed using a 1s rolling window within participants. Participant average is shown. Shaded area reflects 95% CIs.

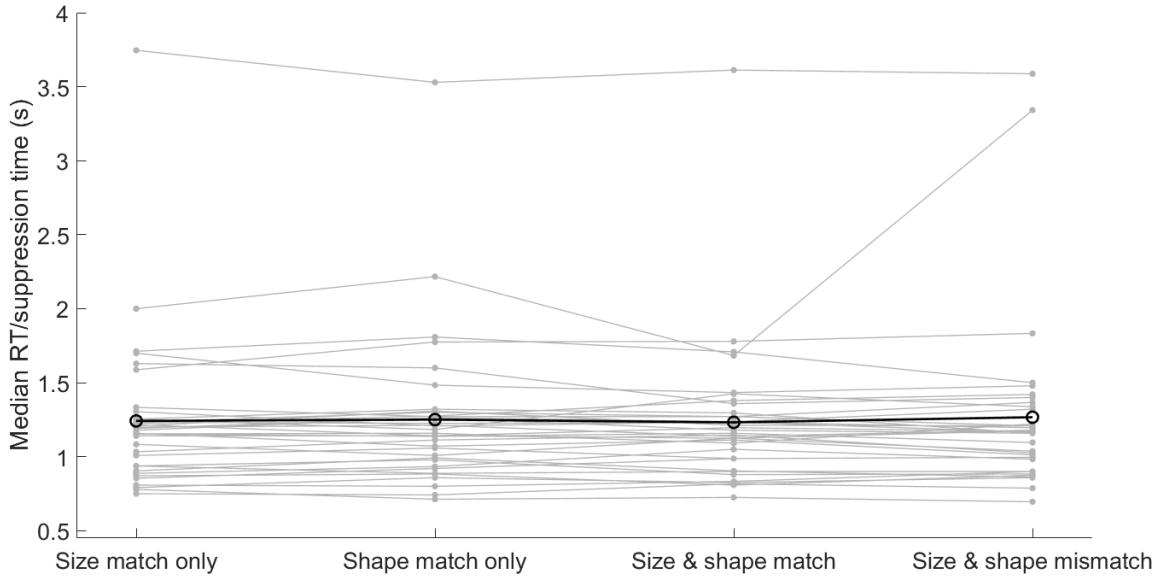
False alarms were common mistakes, with a false alarm rate of 45.79% (sd 8.71) and hit rate of 74.89% (sd 9.68). **Fehler! Verweisquelle konnte nicht gefunden werden.** shows the development of both false alarm and hit rate across the individual experimental blocks. Fitting linear models for both false alarms and hit rates revealed a slight decrease of false alarms over blocks, suggesting learning (slope:

-0.84,  $p = 0.008$ ), while hit rates remained stable (slope: -0.15,  $p = 0.48$ ).

Since target-absent scenes contained both a shape- and size-based distractor and the size-based distractor objects in some (but not all) scenes were search targets in other trials, hits and false alarms cannot meaningfully be computed separately for distractor types.



**Figure 3:** Search performance across experimental blocks. Error bars are 95% CIs.



**Figure 5:** Median raw RTs to probes of the different target-match conditions. Grey lines are individual participants, the black line represents mean over participants.

### B-CFS trials

Participants responded to the b-CFS probes in 99.71% of trials (sd 0.61), and correctly reported the side on which the probe appeared in 99.62% of those (sd 0.61). Trials with responses faster than 300 ms were excluded from this analysis, as these could not reflect meaningful responses to the probes. Median suppression time was 1.24s on average (sd 0.51).

To analyse suppression times as a function of their match with the template (Figure 5), we used a latency-normalization procedure as proposed by Gayet & Stein (2017), since suppression times remained non-normally distributed even after log-transform or z-scoring. This method generally tends to give both sensitive estimates, by taking into account inter-individual variability in overall suppression times, as well as normally distributed measures. Normalized RT differences were computed on the raw suppression times of every participant as follows:

$$\Delta RT_{size\ match} = \left( \frac{\text{median } RT_{shape\ match\ only} - \text{median } RT_{shape\ \&\ size\ match}}{\text{median } RT_{overall}} \right) / 2 +$$

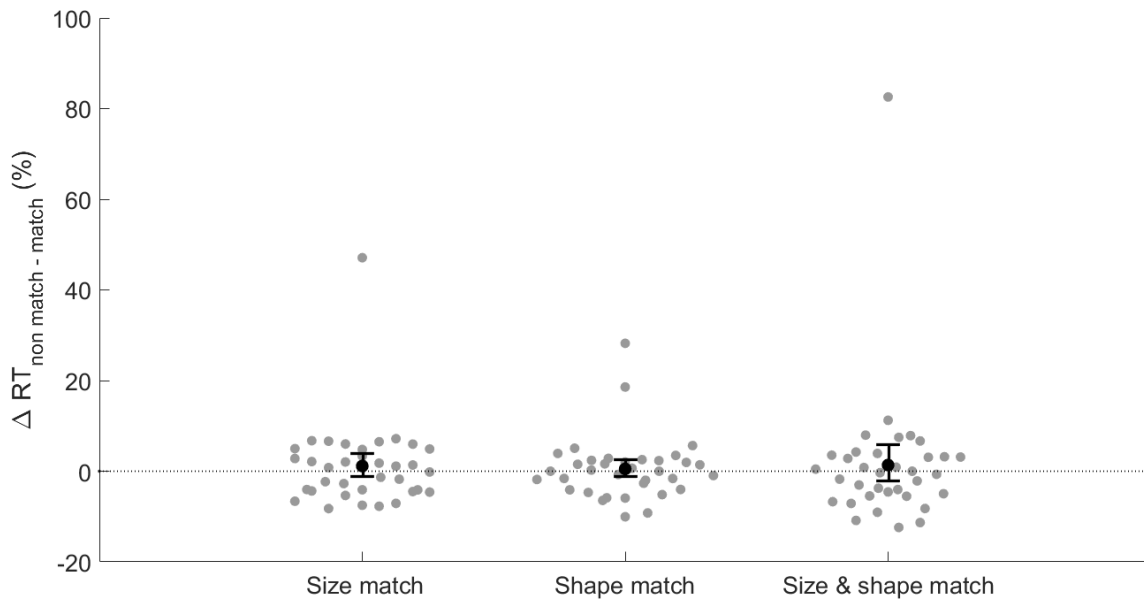
$$\left( \frac{\text{median } RT_{shape\ \&\ size\ mismatch} - \text{median } RT_{size\ match\ only}}{\text{median } RT_{overall}} \right) / 2 \times 100$$

$$\Delta RT_{shape\ match} = \left( \frac{\text{median } RT_{size\ match\ only} - \text{median } RT_{shape\ \&\ size\ match}}{\text{median } RT_{overall}} \right) / 2 + \left( \frac{\text{median } RT_{shape\ \&\ size\ mismatch} - \text{median } RT_{shape\ match\ only}}{\text{median } RT_{overall}} \right) / 2 \times 100$$

$$\Delta RT_{shape\ \&\ size\ match} = \left( \frac{\text{median } RT_{shape\ \&\ size\ mismatch} - \text{median } RT_{shape\ \&\ size\ match}}{\text{median } RT_{overall}} \right) \times 100$$

As also the distribution of those latency normalized RT differences remained non-normal, we resorted to bootstrap tests which do not rely on any assumptions regarding the underlying distribution of the metric of interest.

We first asked whether in general, probes whose size matched the current search target were detected faster than incongruently sized ones. Contrary to our hypotheses, we found no reduction of suppression times for size-matching probes (size match: bootstrapped mean: 1.14%, 90% CI: [-1.09, 3.88],  $p = 0.25$ ) (Figure 6). There was also no general effect of shape-match (0.51%, [-1.28, 2.55],  $p = 0.34$ ).



**Figure 6:** Effect of target match on suppression times as normalized latency differences between feature-matching and non-matching probes. Grey dots represent individual participants, black dot and error bars bootstrapped mean and 90% CIs. Positive values reflect shorter suppression times for template-matching compared to non-matching ones.

Arguably, the largest template-related effects are expected when comparing RTs to probes for which both shape and size match the current search target to those for which neither matches. However, also this comparison indicated no difference in suppression times (size & shape match: 1.37%, [-2.18, 6.15],  $p = 0.34$ ). Thus, whether or not the probe matched any feature of the current search template did not significantly alter suppression times.

## Discussion

We tested a novel behavioural paradigm combining a visual search task, in which participants had to take into account changes in retinal size based on depth, with breaking-continuous flash suppression to probe the search template created when searching either far away or nearby. The task was clearly challenging, as indicated by the large number of participants not significantly performing above chance. Our subgroup of participants was however able to correctly account for depth-dependant size changes and ignore similarly shaped distractor objects of

incongruent size present in the same depth plane. Aside from a general advantage in detecting larger objects, participants were equally accurate across scenes and scene types. This indicates they successfully used all scene previews as cues for the expected retinal size of objects, validating the general stimuli and search task used. Further, above chance performance across the entire delay period at least suggests participants may have maintained an active search template throughout the entire time in which the b-CFS probes were presented.

Despite this, even for those participants whose performance necessarily indicated successful search preparation and comparing probes that exactly matched the search target in both size and shape to others for which both features were mismatching, no template-related reduction in suppression times was found.

Given previous findings of reduced suppression times for probes matching the content of visual working memory and evidence the search template relies on visual working memory, this seems surprising. What may be the reason(s)

for matching effects found in previous working memory tasks but not in the current search task?

In general, working memory match effects on suppression times have been found using simple stimuli as colours (Gayet et al., 2013), but also more complex ones as shapes (Gayet et al., 2016) and human faces (Pan et al., 2014). This makes it unlikely the category of our chosen stimuli or specific visual areas in which different working memory - or search templates are maintained can explain the null effects found in the current study.

In spite of the apparent similarities between search and general working memory content, the template formed in this task could however have been qualitatively different from the ones in previous working memory studies. In comparison to those studies, we used a rather limited set of memoranda and the task did not require fine-grained within category discrimination (e.g. between different shades of the same basic colour or individual faces). Further, in previous studies specific memoranda were presented as image cues whose representation the participants could maintain online or reactivate during the delay period. In our task, the specific template needed to be generated internally by correctly estimating depth-information from the scene and combining the inferred target size with the symbolic shape cue. Both of these factors could have resulted in a more abstract and less visual template, not able to interact with the probes before they overcame interocular suppression and were consciously processed on a more conceptual level.

One previous study using a detection instead of working memory task found feature-based attention (i.e. creating an attentional template for a particular feature) based on symbolic cues was not sufficient to facilitate conscious access and reduce suppression times (Gayet, Douw, van der Burg, Van der Stigchel, & Paffen, 2018).

However, in this study target and probes were also visually distinct and probes matched only one feature (colour). Moreover, an effect of feature-based attention was found when the target was defined more broadly, potentially increasing the overlap of probes and template. The previously mentioned differences may therefore not entirely explain why even probes exactly matching the search target were not detected faster than non-matching ones.

It is also possible that the template-related activity in our search task was qualitatively similar to previous studies, generally able to interact with the probes, but still weaker and our paradigm not sensitive enough to probe it. One reason may be the task's difficulty. Even though the task was generally designed to require effortful search preparation and strong template activation, the task may simply have become too difficult, reducing any template-driven effect. The participants may e.g. have been less engaged in the search task on individual trials. If they began generally focusing more on detecting any kind of object in the much easier b-CFS task, this could have reduced the bias in favour of template-matching objects. The fact that the template needed to be internally generated could also have made template creation more difficult, and participants therefore created wrong or no templates on individual trials, reducing overall matching effects.

Besides potentially weaker neural activation, another relevant factor to consider is that suppression times in this study were generally low. Median suppression time was 1.24s (note that the probe reached its full contrast only after 1s), while response times around 1.7 - 2s or longer are common in other studies. Interocular suppression was therefore likely shallower in our study, independent of matching condition. This is of particular relevance as RT differences between conditions typically become more pronounced with longer overall RTs in the b-CFS paradigm

(Gayet & Stein, 2017). Reasons for this shallow suppression may be visual characteristics of the specific mask and probe stimuli used, the decision to present probes already from mask onset, the fact that probes were repeated more often or a combination thereof.

Overall, while (at least a subset of) participants successfully interpreted depth-information in the scene and likely engaged in sustained search preparation across the entire delay period, we were not able to probe any template due to either general differences in the preparatory mechanisms involved or insensitivity of our paradigm. Whether participants adjust their search template in size could thus not be conclusively answered with the present experiment. We reasoned that a more direct neural measure using fMRI and MVPA, previously used successfully to study search templates in naturalistic scenes, may provide more insight into the specific template created when searching at a particular depth. For this second experiment, we also adapted the search task in order to ease some of its difficulty. As an additional advantage, using fMRI allowed to probe the template without the need for a dual task and therefore more similar to visual search in naturalistic settings.

## Experiment 2: fMRI

---

In this second experiment, we aimed to probe the neural basis of the template created when searching at different depths and mechanisms to integrate distance-information with the template, focusing on object-selective LOC, scene-selective areas OPA and PPA as well as early visual cortex (EVC). The search task was similar to the previously used one. To probe the template, we introduced trials in which participants prepared to search for an object either near or far, but only the empty search scene was shown (catch – trials, randomly intermixed with search trials). These allowed us to isolate neural activity solely reflecting search

preparation, in the absence of actual targets. In addition to neural activity patterns related to *merely searching* for either large or small objects, we also extracted activity patterns related to *seeing* those large or small target objects in isolation within a different task and tested for an overlap in these patterns. This was done by training classifiers on decoding the retinal size of these isolated objects and test whether they could also correctly classify the size of objects participants were searching for (cross-decoding). Successful (i.e. above-chance) cross-decoding accuracy for size in this approach would suggest participants created differently sized-templates during search, resembling the neural activity evoked by seeing target objects. We specifically hypothesized to find such template-related activity in LOC.

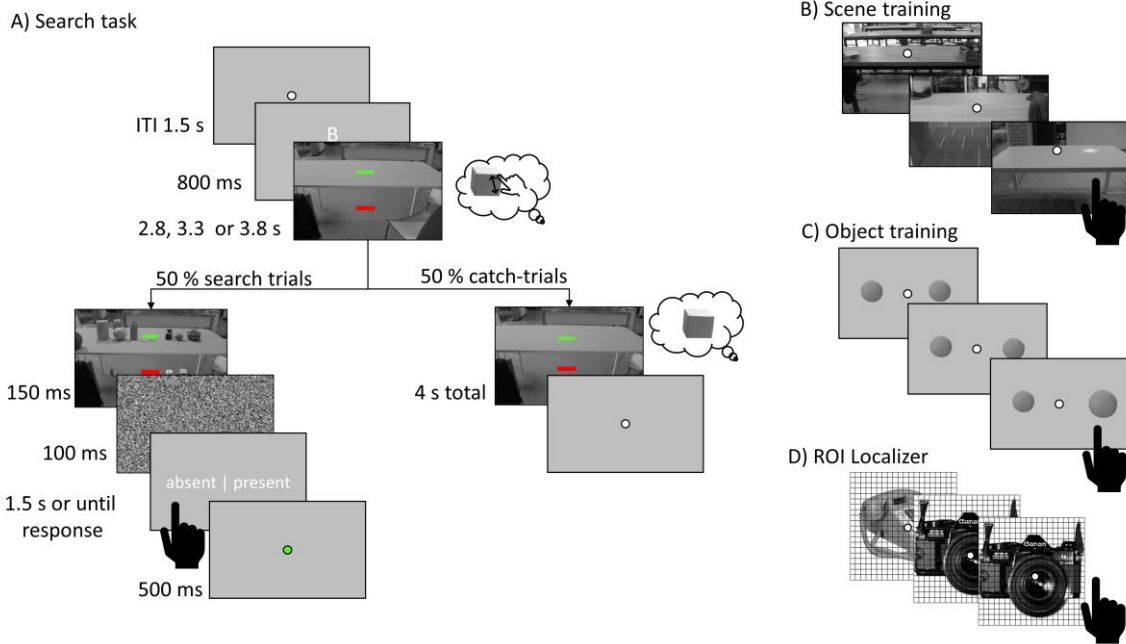
In a similar manner, we probed the depth-representation during search. To this end we extracted activity patterns related to fixating and attending near and far in the scenes to cross-decode and isolate scene-based distance information when participants were searching for target objects in these depth planes.

This allowed us to derive two measures for every catch-trial in the search task, reflecting template size and depth-information. We correlated these on a trial-by-trial basis to test whether depth-information in scene-selective areas PPA and OPA may modulate the size of the template in LOC.

## Methods

### Participants

30 participants (14 females, mean age: 24 years,  $sd = 3.28$ ) participated in the fMRI study. They were recruited through the Radboud university participant pool (SONA systems) and participated for monetary reward (20 or 25 €). All provided written informed consent and reported having normal (or corrected to normal) vision. Two additional participants were excluded as they stated having difficulties to see the isolated target objects presented in



**Figure 7:** Timeline of a trial in the search task and overview over all fMRI tasks. A) In the search task, each trial began with a category cue reminding participants of the target category (melon or box) in the current run. After fixating the letter cue, an empty search scene appeared and participants fixation was now in the center of one depth plane, providing information about the retinal size of the target objects. In half of the trials, objects briefly appeared in the search scene and participants reported whether the target was present. In the other half, which were randomly intermixed with search trials, participants prepared to search but the scene remained empty of objects. These catch-trials were used to probe neural activity solely related to search preparation (i.e. the search template). In addition to the search task, three types of localizer tasks were included in the experimental session. B) the scene training task was used to extract voxel activity pattern reflecting scene-based distance information and fixating in either depth plane. Participants attended to oddball targets (light patches appearing left or right in the depth plane). C) In the object training tasks pair of target objects were shown in isolation, to capture voxel activity patterns related to seeing those large and small objects. Participants attended to size-oddballs in which one of the two presented objects was slightly larger or smaller than the other. D) The functional ROI localizer was used to locate object- and scene-selective areas and included photographs of objects, scenes, faces and scrambled objects. Participants reported 1-back image repetitions.

the object localizer task, required for any meaningful cross-decoding of the search template. For these participants, the stimulation > baseline contrast in this localizer (used for defining early visual cortex) also yielded less than 6 active voxels in V1. As in the previous experiment, we aimed for a total number of 34 participants based on a power-analysis, but this number was not reached before completion of the thesis.

### Procedure

During the experimental session, participants took part in eight runs of the search task, as well as in two runs each of three different localizer tasks. Search task and localizer runs

were intermixed. All tasks were briefly explained and practiced outside the scanner. During training, participants also saw the original box used for creating the search scenes to ensure they could estimate its real-world size. The whole experimental session lasted around 2 hours. One subject only completed 6 runs of the search task and another one only 7.

### Search task

Each search task run consisted of 32 trials, yielding a total of 256 trials. The task was similar to the one in experiment 1, adapted to the fMRI paradigm (Figure 7). To decrease task difficulty, target category was now blocked per run, making retinal size (the main feature of

interest) the only relevant target feature changing on a trial-by-trial basis and decreasing working memory demands. In addition to a prompt at the beginning of every run, a letter cue (M or B) at the beginning of each trial reminded participants of the current target category. This letter was presented either above or below the centre of the screen, requiring participants to make an eye movement to fixate it. The category cue was followed by a depth cue, provided by the empty search scene. Due to the previous eye movement, participants were now fixating in the middle of the currently relevant depth plane, indicated by a green horizontal bar at fixation. A second red bar marked the centre of the other (currently irrelevant) depth plane. Again, participants were told they could use this distance information to anticipate the size of the target and about the presence of distractors sharing either shape or retinal size of the target. As previously, only the search scene itself, but not the direction of the earlier eye movement, was predicting the target's retinal size. In contrast to the b-CFS experiment, the empty scene remained on screen until objects appeared or the trial ended, to minimize evoked responses by stimulus onsets.

In half of all trials (search trials), objects briefly appeared in the scene for 150 ms, after a randomly varying cue-target onset asynchrony of 2.8, 3.3 or 3.8s. This was followed by a 100 ms backward mask. Once the search scene disappeared, participants were asked to report whether the target had been present or absent, by pressing a button with either their left or right hand within a response deadline of 1.5s. Feedback was given by a colour change of the fixation dot and the next trial began after an intertrial interval of 1.5s. Cumulative feedback about performance in the current block was provided at the end of each run.

In the other trials (catch-trials), no objects appeared and the empty scene remained on

screen until a total delay of 4.03s had passed. These were the trials of interest for our analysis, as they allowed to isolate neural activity solely related to search preparation and thus to the search template. No response was required.

Trial type (search or catch-trials), target presence, distractor type (size or shape based), depth plane and scene type ("near below" and "near above" scenes) were fully counterbalanced per run. Prevalence of scene family (scenes taken at one of 16 locations), target side (left or right of fixation) and the duration of cue-target asynchrony (2.8, 3.3, or 3.8s) was equated as far as possible within runs, but not fully counterbalanced.

#### *Localizer tasks*

For all localizer tasks, a miniblock based design was used. Within each miniblock, 20 images belonging to the same condition were presented in rapid succession (for 450 ms each). One run consisted of 16 miniblocks (4 repetitions of 4 conditions), lasting 14.7 s and interspersed with baseline fixation blocks. Participants had to either respond to oddball stimuli or 1-back repetitions, appearing twice per miniblock, by pressing a button with either hand.

#### *Scene training runs*

Scene training runs were used to isolate scene-based distance information, specifically voxel activity patterns related to fixating and attending either near or far away. The same empty scenes as used in the main search task were shown. To create a miniblock, scenes were grouped by attended depth plane and scene type. Oddballs were small oval light patches appearing either to the left or right of fixation (the centre of one depth plane).

#### *Object training runs*

With object training runs we aimed to isolate neural activity related to the retinal size of the targets in the search task. Per miniblock, images of two isolated objects of the same

category and size were presented left and right of a central fixation dot, at the same retinotopic positions at which targets would appear in the search scene. Since participants always fixated within a given depth plane, there were four different locations, but effectively only two retinotopic locations at which targets would potentially appear in the search task. For oddball stimuli, one of the two objects was 30% larger or smaller than the other.

#### Functional ROI localizer

A functional ROI localizer was used to locate both object- and scene-selective regions. Participants were shown grayscale images of faces, scenes, objects and scrambled objects. They performed a 1-back task, reporting immediate repetitions of the same image.

#### Stimuli & Setup

Stimuli were presented on a 1920 x 1080 pixel IPS LCD BOLDscreen (120 Hz refreshrate) and backprojected into the scanner bore. Participants viewed the stimuli through a mirror placed on the head coil.

Scene stimuli for the main search task and scene training runs were the same as used in Experiment 1. The scene localizer only used the empty scenes. Scenes subtended  $20 \times 14.434^\circ$ . Light patches were created by blending 2D Gaussians ( $\sigma = [12^\circ, 0^\circ, 3.96^\circ, 0^\circ]$ ) into the scene images. These did not change in size dependent on depth.

For the object training runs, isolated images of the target objects were used. Two objects of the same object category and size were presented left and right of a central fixation dot, at the same retinotopic positions ( $3,27^\circ$  eccentricity) and with the same retinal size ( $1,5$  and  $3^\circ$ ) as in the search task. For the size-oddball stimuli, size was changed to 130% or 70% of the original size, making large objects even larger and small targets even smaller. Target objects were cropped out from all 16 scenes and mean luminance of the isolated

objects equated before placing them on an equiluminant grey background, having the mean luminance of the search scenes (RGB 104, 104, 104).

For the functional ROI localizer, greyscale images for each of the four categories (faces, scenes, objects and scrambled objects) were used.

Experimental code for stimulus presentation and response collection was run using Matlab and Psychtoolbox (Brainard, 1997).

#### Data Acquisition and Preprocessing

Data were acquired on a 3T Siemens Prismafit Scanner using a 32-channel head coil. A T2-weighted gradient echo EPI sequence was used for acquisition of functional data (TR 1 s, TE 34 ms, flip angle  $60^\circ$ , 2 mm isotropic voxels, 66 slices). For the search task, 295 images were acquired per run and 318 per run for all localizer runs. A high-resolution T1-weighted anatomical scan was acquired prior to the experimental runs, using an MPRAGE sequence (TR 2.3 s, TE 3.03 ms, flip angle:  $8^\circ$ , 1 mm isotropic voxels, 192 sagittal slices, FOV 256 mm).

#### Preprocessing

Data preprocessing was performed using SPM12 ([www.fil.ion.ucl.ac.uk/spm](http://www.fil.ion.ucl.ac.uk/spm)). Preprocessing steps included spatial realignment, co-registration of functional and anatomical scans and normalization to MNI 152 space. A Gaussian filter (FWHM 3 mm) was then applied to smooth the images.

#### General Linear Model (GLM) Estimation

Subject level GLMs were estimated on the preprocessed images. For the search task, boxes and melons runs were modelled individually as separate GLMs. Regressors included the two possible sizes of the individual search targets, modelled as boxcar functions across the duration of the distance cue (empty search scene) presentation (4.03s) and convolved with the canonical SPM

hemodynamic response function. Only catch-trials, in which no objects appeared in the scene, were modelled for the search template analysis. Single-trial beta estimates were obtained by modelling a separate GLM for each trial, including one regressor for the trial of interest and another common one for all other trials to reduce collinearity, following Mumford et al. (2012).

GLMs for the scene training runs included regressors for fixations in the near and far depth plane (collapsing across scene type), modelled individually for each miniblock within a run as boxcar functions over the duration of a miniblock (14.7 s) and convolved with the canonical SPM hemodynamical response function.

The object training GLM included regressors for each of the four size and shape combinations (large and small melons and boxes). As for the scene training runs, each miniblock within a run was modelled individually.

For the functional ROI localizer, miniblock-based regressors for the four object categories (faces, scenes, objects and scrambled objects) were included.

All GLMs also included the estimated head motion parameters as nuisance regressors and GLM estimation included temporal high-pass filtering (cutoff: 128s) to remove low-frequency drift in the signal.

#### *ROI definition*

We defined functional ROI masks of different sizes for LOC, PPA, OPA and EVC (early visual cortex) for every subject by intersecting anatomical or functional masks with the  $x$  most active voxels of the relevant contrast map (thresholded at 0.05). We first determined the maximum number of active voxels within each ROI across participants, before creating smaller ROIs of varying voxel counts going up this maximum in 20 equidistant steps.

LOC was defined on the basis of the objects > scrambled objects contrast in the functional localizer runs (range of included voxels: 14 - 534, median: 189). Scene selective regions PPA and OPA were defined as being more strongly activated by scenes than other objects in the functional localizer (contrast: scenes > objects + faces; PPA: 24 - 388 voxels, median: 186; OPA: 17 - 237 voxels, median: 88). All were intersected with bilateral functionally defined masks in MNI space.

EVC was defined as voxels responsive to visual stimulation in the object training runs [objects > fixation baseline contrast], intersected with Brodmann areas (BA) 17 and 18 (253 - 2634 voxels, median: 1109). This ensured we specifically included those voxels sensitive to stimulus presentation at the retinotopic locations at which targets appeared in the search task.

For a small subset of participants, intersecting contrast maps with the functional masks yielded less than 6 active voxels (for LOC definition in three participants, and OPA definition in three participants). In this case, the threshold of the contrast map was lowered to 0.10 before intersecting.

#### Multivariate Analyses

Multivariate analyses were conducted using linear support vector machines (SVMs) and the single-trial (for the search task) or miniblock- (for the two training tasks) based beta estimates within an ROI. Classification was performed using the decoding toolbox (TDT) (Hebart, Gorgen, & Haynes, 2015) and the `libsvm` library (<https://www.csie.ntu.edu.tw/~cjlin/libsvm/>).

For within-run type classification, a leave-one-run-out cross-validation design was used. For the two training tasks, each miniblock was effectively considered a run.

For cross-classification, we trained SVMs on all betas of either the object or scene training task

and used either all beta estimates from the search task or from the other training task as testing set (or vice versa for reverse cross-decoding). Individual SVMs were trained for every participant, ROI and voxel count. To take into account potential classifier bias, arising using different training and testing sets, we also recomputed accuracy after creating a median decision hyperplane, in which those testing betas closest to one label (e.g. near or large, determined by their distance to the decision boundary) were relabelled as such.

Above chance-classification performance was determined by two-sided t-tests against 0. To correct for multiple comparisons when training and testing SVMs of the same general ROIs across different voxel counts we used threshold-free cluster enhancement (tfce). Decoding accuracy across voxel counts was tested against a null distribution created by randomly permuting test labels within each SVM (using 10000 bootstrap iterations).

#### Searchlight Analyses

In addition to our planned ROI analyses, we also tested whether retinal size-information was present during search preparation (reflecting a depth-dependent template) in other brain areas. We conducted a whole-brain searchlight analysis using spheres with a radius of 5 voxels (corresponding to approximately 523 voxels enclosed in the sphere). Within each sphere, we calculated accuracy for cross-decoded retinal size (training on the object training task and testing on the search task). Subject-level searchlights (in MNI space) were combined to a group map, testing for consistent above-chance decoding accuracy in individual voxels across participants and using tfce to correct for multiple comparisons.

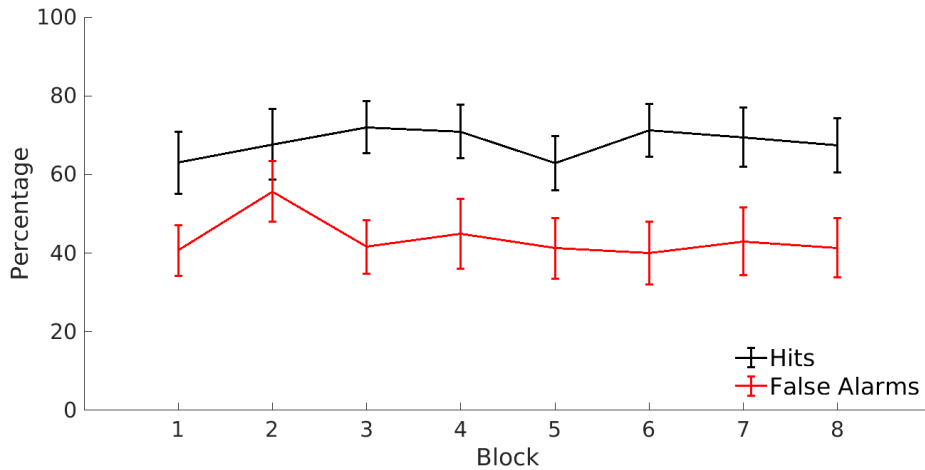
#### Univariate Analyses

In addition to the multivariate analysis, we further analysed univariate responses in our search and localizer tasks, which may influence cross-classification. We computed mean

responses to large and small objects and near and far scene-fixations by averaging all beta estimates within an ROI (including all active voxels per participant, i.e. the highest voxel count) and collapsing across the two separate shape-GLMs created for the search task.

#### Trialwise correlations of size and distance-decoding between ROIs

For every trial in the search task, our design allowed us to have both an estimate of template size and the representation of depth in this trial, by comparing how classifiers trained on either the object or scene task labelled it. To analyse whether template size in object-selective areas was informed by distance information in scene-selective regions, we extracted the distance to the decision boundary of all betas in the search task after cross-classification of size or distance for each SVM. These provided us with a continuous measure jointly reflecting the predicted label (near or far and large and small) and classifier certainty (with large distances reflecting greater certainty a particular trial belonged to one class rather than the other). Next, those distance measures were converted to ranks and averaged across unique voxel counts within an ROI. Near and far trials were correlated separately before averaging them to an estimate we named  $\tau_{\text{split}}$ . This was done to avoid finding correlations simply based on above-chance classification of both size and distance or different classifier bias. Correlations were computed within participants before calculating the mean correlation for each ROI pair. Put simply, the resulting  $\tau_{\text{split}}$  reflects whether, on a particular search trial, if this trial is classified as small (i.e. a smaller template was formed), it is also represented as further away in either the same or another ROI, independent of its true label.



**Figure 8:** Performance across experimental blocks of the fMRI search task. Error bars are 95% CIs.

## Results

### Behavioural Performance

#### Search Task

Participants performed above chance in the search task, with a mean accuracy of 62.54% (sd 9.75,  $t(29) = 7.04$ ,  $p < 0.00001$ ). As in the first experiment, a 2x2x2 repeated measures ANOVA with factors Scene Type, Depth Plane and Target Object revealed that searching nearby and therefore for larger objects was easier compared to far away ( $F(1,29) = 4.21$ ,  $p = 0.0494$ , near: 64.66% (sd 10.54), far: 60.41% (sd 12.00)). No other factors or interactions were significant (all  $p$ 's  $> 0.14$ ).

Participants detected 68.53% of targets (sd 11.98) and false alarm rate was 43.46% (sd 13.06). Both hits and false alarms remained stable across experimental blocks (hits: slope 0.004,  $p = 0.55$ , false alarms: - 0.81,  $p = 0.35$ ) (see Figure 8).

Six participants did not individually perform above chance level, as determined by a one-sided binomial test (alpha level 0.05). These were included in all analyses, but main findings also remained consistent when excluding them.

#### Localizer Tasks

In the object training task, general hitrate was 69.74% (sd 12.75) and false alarm rate 1.67% (sd 2.12). In the scene training task, participants responded to 93.49% of oddballs (sd 8.03) and incorrectly responded to non-targets in 1.67% (sd 2.12) of trials. For the ROI localizer task, hitrate was 89.74% (sd 7.47) and false alarm rate 0.60% (sd 0.53).

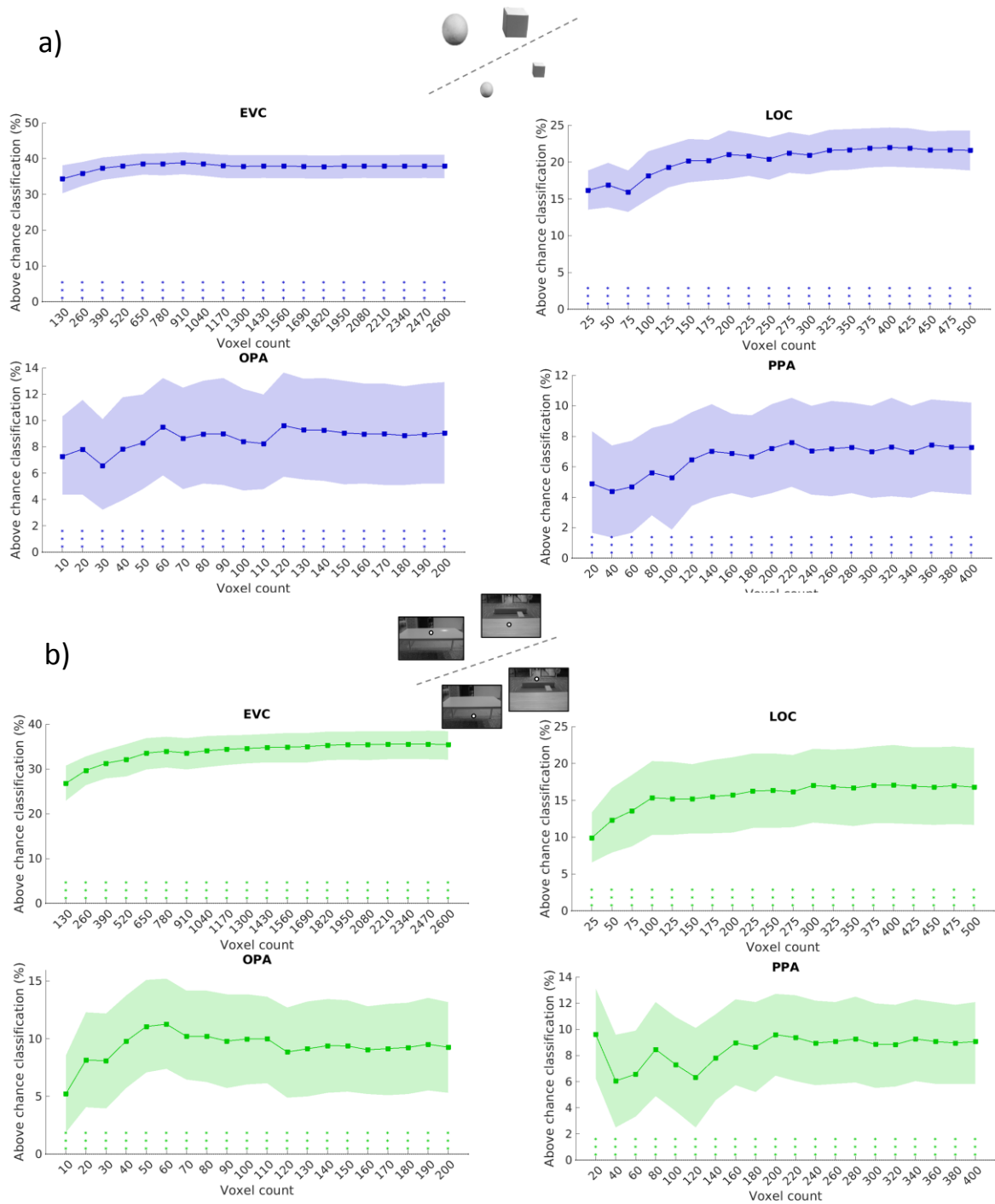
#### Decoding within localizers

Before investigating template-size and depth representation in the search task, we first ensured our ROIs generally encoded size- and distance information by decoding the retinal size of objects participants saw or depth plane in which they fixated within the two training tasks.

When large or small target objects were present on screen and in isolation, their size could be decoded above chance-level<sup>1</sup> from all chosen ROIs and across all ROI sizes (all  $p$ 's = 0.0002 after tfce) (see Figure 9 a)).

Decoding which depth plane participants fixated in the scene training showed a similar pattern of consistent decoding across ROIs,

<sup>1</sup> Chance level for all multivariate analyses was 50% and results always reported as deviation from chance.



**Figure 9:** Decoding of A) retinal size of isolated target objects within the object training task and B) distance within the scene training task across ROIs and ROI sizes. Asterisks indicate significant difference from chance-level (50%) after threshold free cluster enhancement. Shaded areas are 95% CIs.

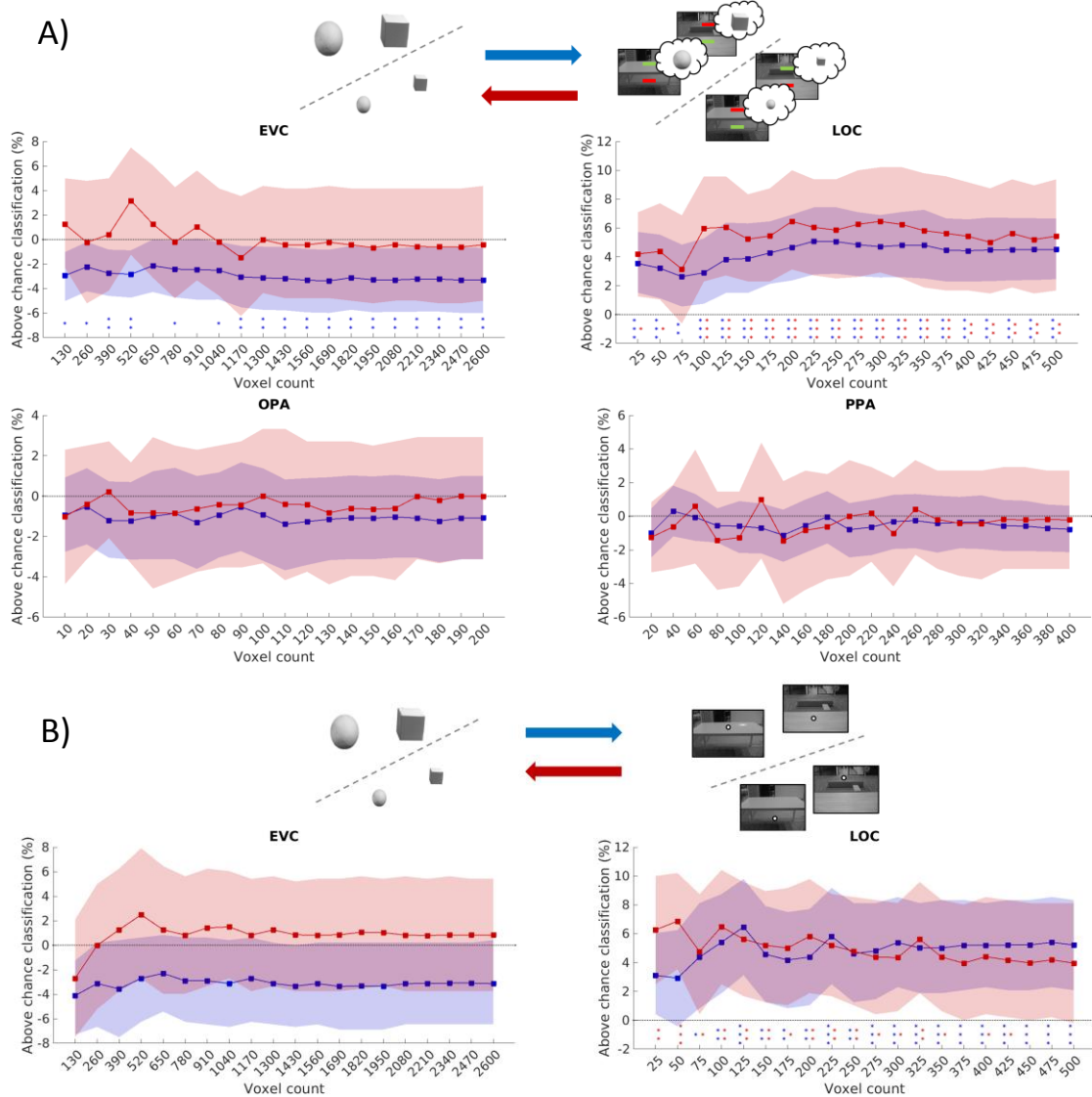
independent of their general selectivity (all  $p$ 's = 0.0002).

#### Cross-decoding of size

To test whether participants formed depth-dependent search templates in the search task, we trained classifiers to decode the size of objects participants saw in the object training

task and then applied these classifiers to decode the retinal size of objects participants prepared to search for (see Figure 10 a)).

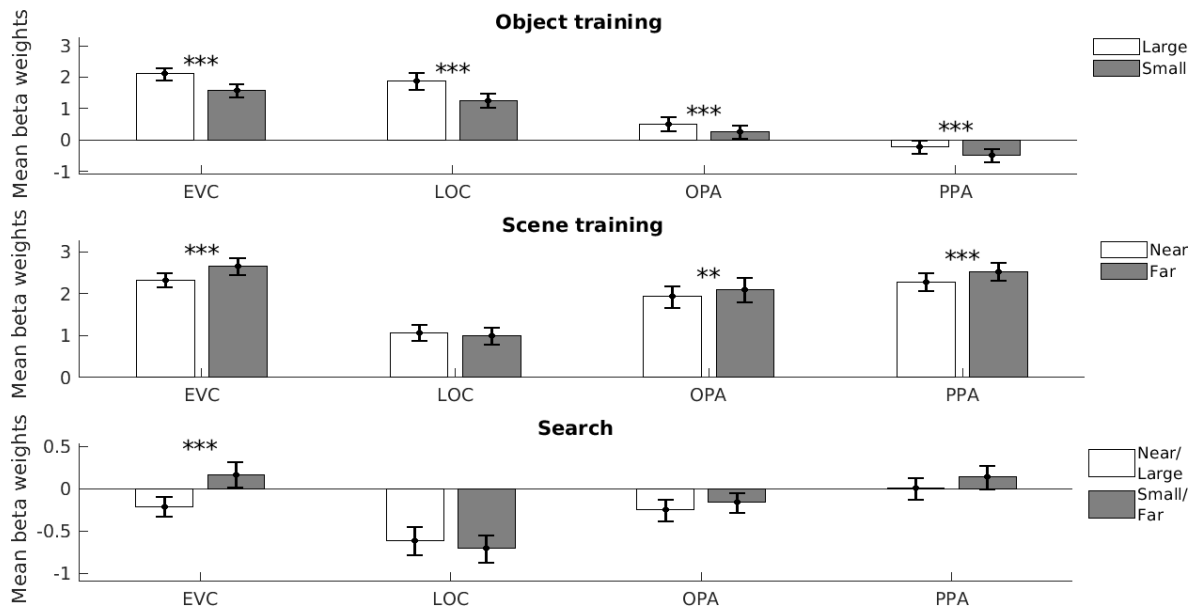
Consistent with our notion of depth-dependent templates, size information could be successfully cross-decoded in LOC across all voxel counts (all  $p$ 's < 0.001 after tfce). Within



**Figure 10:** A) Cross-decoding of retinal size within different ROIs and ROI sizes. Classifiers were trained on decoding the retinal size of target objects when these were present on screen in the object training task and tested on the search task to decode the retinal size of objects participants were searching for, when these were not present on screen (or vice versa). Asterisks indicate significant above-chance cross-decoding after tfce. Colour of the arrows and lines reflect decoding direction (blue: training on object training task and testing on search task, red: training on search task and testing on object training task). Shaded area represents 95% CIs. B) Cross-decoding of retinal size and distance across localizer task. Classifiers were trained on the object training task to decode retinal size and tested on the scene training task to decode whether participants fixated near or far (or vice versa, blue: training on object localizer, testing on scene training task; red: training on scene training task, testing on object training task).

scene-selective regions, no such information was present (all  $p$ 's > 0.99). In EVC, decoding accuracy was even significantly below chance across nearly all ROI sizes (18/20, all  $p$ 's <= 0.056). This may reflect a stronger low-level visual similarity between small objects and near parts of the scene in which participants were fixating during search (or vice-versa).

Further, this suggests successful size-decoding in LOC was not driven by a simple overlap in visual features, but related to object size on a more abstract level. Decoding in LOC did however not correlate with general search accuracy across participants ( $\tau = 0.02$ , 95 % CIs [-0.25, 0.29],  $p = 0.39$ ).



**Figure 11:** Univariate responses in the search and training tasks. Error bars represent 95% CIs

Reversing the cross-decoding direction (training on the search task and testing on the object localizer) also yielded equivalent results, with exception of the negative decoding in EVC.

To further investigate whether the effects found in LOC and EVC were specific to the search task, we also attempted to cross-classify size and distance between the two training tasks (Figure 10 b)). We trained classifiers on decoding object size and tested whether they would correctly classify whether participants were fixating near or far in the scene training task, which did not require taking into account depth-dependent size-changes (or vice versa). Surprisingly, we found above-chance decoding in LOC across most voxel counts (all  $p$ 's  $\leq 0.15$  in both cross-decoding directions)<sup>2</sup>. Decoding accuracy was not significantly below chance-level in EVC (range  $p$ 's: 0.99 – 0.13 across decoding directions) and remained at chance for scene-selective areas (all  $p$ 's  $> 0.8$ ).

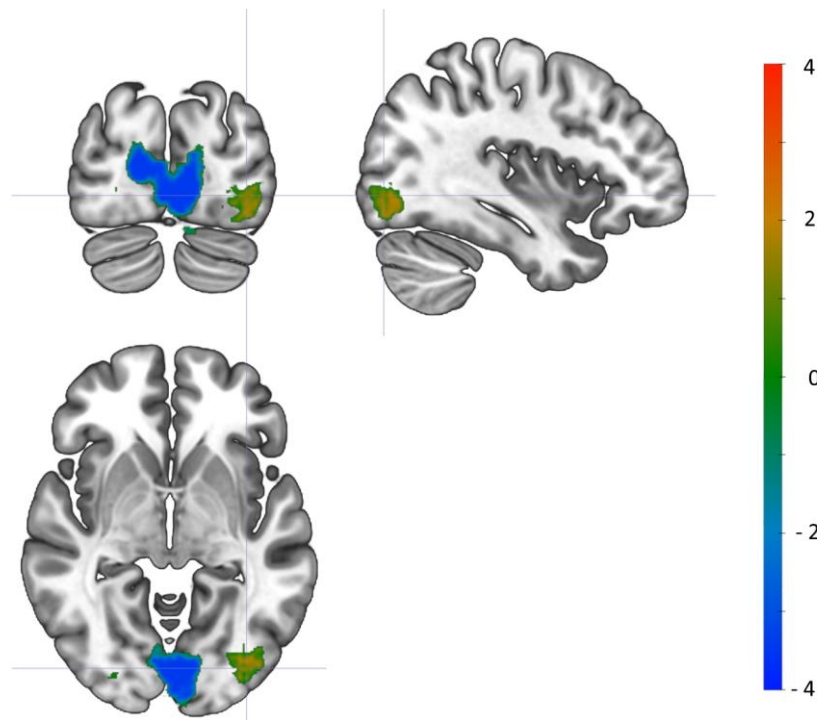
<sup>2</sup> Miniblock-based beta estimates were used in the training tasks in comparison to single-trial estimates in the search task and the number of testing betas for decoding differed (128 in the

### Univariate Analyses

Given these surprising results, we asked whether general univariate differences might explain our findings in LOC (Figure 11). Generally, a stronger univariate response to nearby objects and of larger retinal size has been previously found in LOC (Amit et al., 2012; Cate, Goodale, & Köhler, 2011). Such a preference could potentially even translate to stronger responses to nearby parts of scenes in the absence of objects, as LOC can be sensitive to general scene-based information and e.g. shows a bias towards small spaces (Park, Konkle, & Oliva, 2015). A congruent univariate response pattern for large vs. small objects and near vs far fixations may lead to successful cross-decoding, without necessarily indicating overlap in more fine-grained voxel activity patterns.

Larger objects indeed evoked stronger activation across all ROIs in the object training task (EVC:  $t(29) = 12.5$ ,  $p < 0.0001$ ; LOC:  $t(29) = 10.24$ ,  $p < 0.0001$ , OPA:  $t(29) = 5.20$ ,  $p < 0.0001$ ;

search task, 32 in the training task). A direct numerical comparison between decoding accuracies for different cross-decoding approaches is therefore not immediately meaningful.



**Figure 12:** Whole-brain searchlight for size-cross decoding. Colours indicate z-scores and map is thresholded at 0.05.

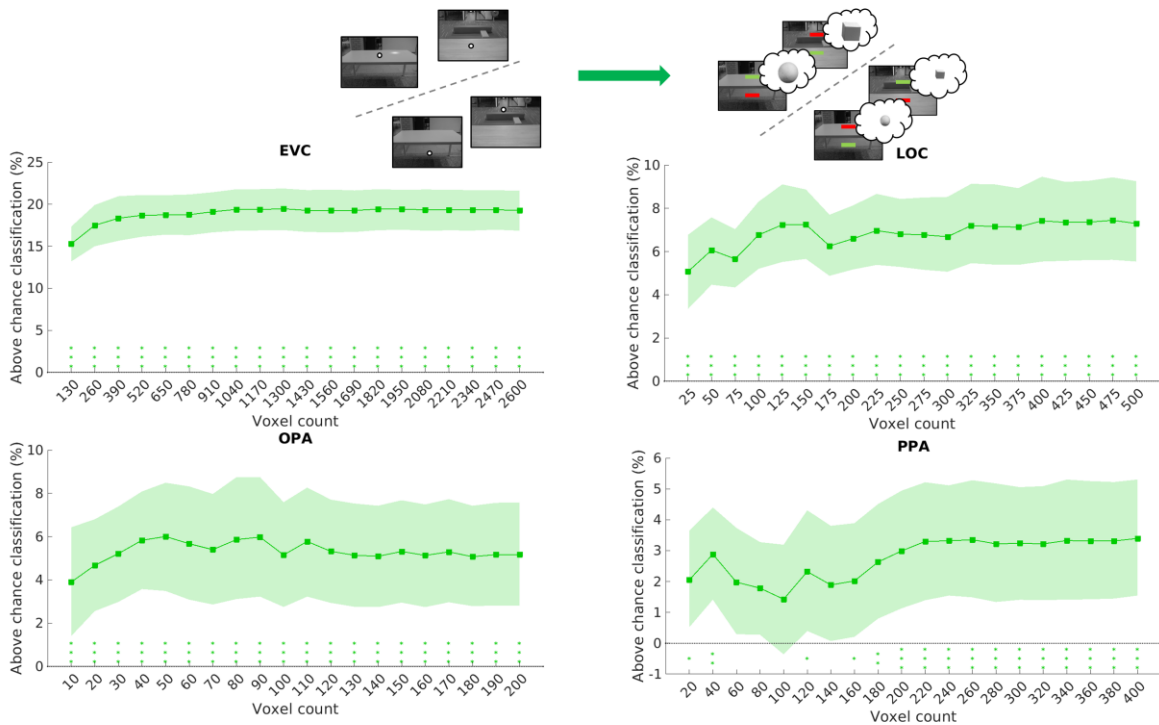
PPA:  $t(29) = 5.93$ ,  $p < 0.0001$ ). This could e.g. be partly driven by the larger extent of cortex they activated in retinotopically organized areas. A corresponding difference between near and far fixations was however not significant in LOC for either scene training or search task (scene training:  $t(29) = 1.76$ ,  $p = 0.09$ ; search:  $t(29) = 1.88$ ,  $p = 0.07$ ). Although general univariate differences related to object size and distance tended to go in the same direction in LOC, they are thus unlikely to explain successful cross-decoding of size.

For EVC however, an opposed pattern of univariate responses for object size and scene locations was found, which could account for the negative cross-decoding from object training to search task. Fixating far away in empty scenes yielded larger responses compared to near fixations (scene training:  $t(29) = -6.02$ ,  $p < 0.0001$ ; search:  $-6.00$ ,  $p < 0.0001$ ). After inspecting the scenes, this higher activation may be reflecting higher local spatial frequency content in far-away parts of the

scenes, which seem less visually uniform than the near-parts which are often blank surfaces.

#### Searchlight analyses for size

Using a searchlight approach, we tested whether further brain regions may encode size-information during search (and potentially also more exclusive to the search task) (Figure 12). Our whole-brain searchlight for cross-decoding of size revealed a large occipital cluster showing negative decoding accuracy (overlapping with EVC, 203828 mm<sup>3</sup>, centre of mass MNI coordinates : 1.39, -88.5, 6.94) and smaller bilateral clusters showing above-chance decoding within the occipito-temporal lobe, overlapping with LOC (left: 33496 mm<sup>3</sup>, centre of mass -37, -81.5, -6.03; right: 160 mm<sup>3</sup>, centre of mass: 26.4, -86.7, -6.6). No further areas were found to represent retinal size during search, confirming our a priori ROI selection.



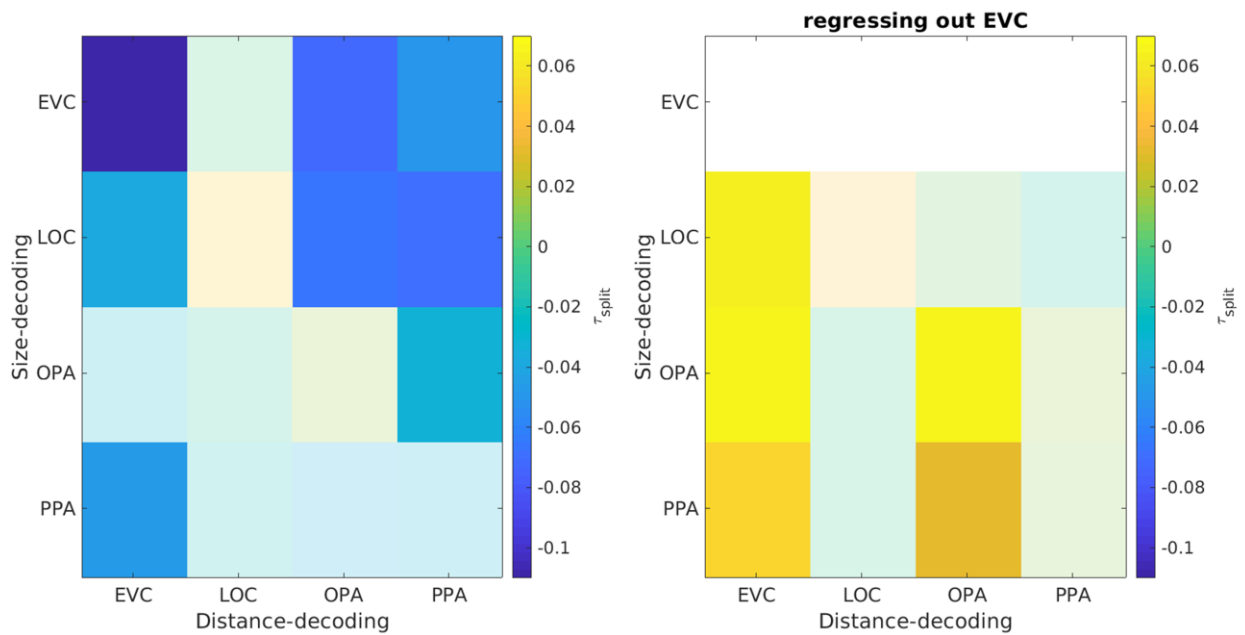
**Figure 13:** Cross-decoding of distance. Classifiers were trained on decoding whether participants fixated in the near or far depth plane in the scene training task and tested on the search tasks in which participants were also fixating in these depth planes to search for objects of varying size. Asterisks indicate above-chance cross-decoding after tfce. Shaded areas are 95% CIs.

### Cross-decoding of distance and trial-wise correlations

Since no simple visual or univariate differences would easily explain size decoding in LOC, it instead seemed to reflect a genuine link between the representation of an object's size and distance, albeit unspecific to the search task itself. We therefore still asked whether distance information from scene-selective areas contributed to it. To isolate distance-information from all ongoing brain activity within the search task, we trained classifiers on distinguishing whether participants were fixating near or far within the empty scenes in the scene task and tested these on the search task. The distance at which participants were searching could be successfully cross-decoded within all ROIs. For EVC, LOC and OPA this was also consistent across all ROI sizes (all  $p$ 's = 0.0002) and across most for PPA ( $p$  (100 voxels) = 0.33, all other  $p$ 's  $\leq$  0.058).

Next, we correlated cross-decoding of size and distance for every trial in the search task (Figure

14). Intriguingly, we found that size decoding in LOC correlated negatively with distance decoding in all other areas (EVC:  $\tau_{\text{split}} = -0.04$ , [-0.07, -0.01],  $p = 0.027$ ; OPA:  $\tau_{\text{split}} = -0.07$ , [-0.10, -0.04],  $p = 0.001$ ) and was also not correlated with distance information in LOC itself ( $\tau_{\text{split}} = 0.05$ ,  $p = 0.2$ ). Negative-going correlations between size and distance decoding were a general pattern observed across almost all ROIs, strongest when correlating both within EVC ( $\tau_{\text{split}} = -0.11$ , [-0.17, -0.05],  $p = 0.002$ ). This seemed in line with generally opposed univariate responses to isolated objects and their corresponding positions within scenes across different visual areas and a tendency for below-chance decoding when training on the object localizer in EVC. As this may therefore reflect low-level differences potentially still influencing size-decoding on individual trials, we attempted to account for them by regressing out the ranks of EVC distance to boundary measures from those of other ROIs in size-decoding before re-computing  $\tau_{\text{split}}$ . This



**Figure 14:** Trialwise correlations between size- and distance cross-decoding in the search task. Correlations between distance-to-bound measures were calculated within participants and separately for searching near and far before calculating the average correlation for each ROI pair. Correlations in nontransparent squares are significant at an alpha-level of 0.05.

revealed a positive correlation of size-decoding in LOC with the representation of distance in EVC ( $\tau_{split} = 0.06$ , [0.1, 0.11],  $p = 0.018$ ), but no association with scene-selective areas (OPA:  $\tau_{split} = 0.01$ , [-0.01, 0.03],  $p = 0.5$ , PPA:  $\tau_{split} = -0.01$ , [-0.01, 0.04],  $p = 0.298$ ) and also not with distance-decoding within LOC itself ( $\tau_{split} = 0.04$ , [-0.01, 0.10],  $p = 0.14$ ).

## Discussion

We investigated neural mechanisms allowing the visual system to account for depth-dependent changes in retinal size, testing whether participants created differently-sized templates based on where they were searching and modulation of template-size by distance-information processed in scene-selective areas.

Generally in line with our hypotheses, we found robust cross-decoding between voxel activity patterns elicited by searching for objects of varying retinal size and viewing those objects, exclusively within LOC and consistent across ROI definitions and decoding directions.

Unexpectedly, this was however not specific to the search task but also found in the scene training task, in which participants similarly fixated near or far in the scenes but there were no explicit instructions to search for objects of varying size.

Besides putative size-information in LOC, we also found negative or negative-going decoding of size within the search task. This was however not consistent across decoding-directions and presumably explained by univariate and visual differences that classifiers trained on the object localizer task capitalized on.

Even though not specific to the search task, the effects found in LOC were unlikely to be due to overlap in low-level features or univariate responses. Instead, they seemed to reflect a more genuine link between locations in the scene and object size.

Given that this link was also found without explicit instructions and need to search for objects of varying size, it is possible our findings

do not reflect template-related activity but instead a process unrelated to search per se. In the object localizer, we presented objects of varying retinal size, which were closely tied to a particular depth plane and of known physical size. This may have incited a representation of distance in the absence of specific depth cues. This depth-representation may then have been successfully cross-decoded across all three tasks. Given the explicit instructions in the search task, it cannot be entirely ruled out this represents an artifact of the strong contextual link between object size and respective scene-locations within the task context. It may however also reflect more automatic mechanisms, estimating an object's size and distance based on object knowledge and/or depth cues. In that case, smaller objects may truly have been represented as further away and LOC likely played an active role in this process. To create object representations reflecting real-world size and independent of retinal size changes (Grill-Spector & Malach, 2001; Konkle & Oliva, 2012), LOC may be involved in estimating the distance of isolated objects and scene parts and integrating this distance information with the retinal input. It has generally been implicated in size-constancy, allowing us to perceive an object as having the same size independent of where it is placed (see Sperandio & Chouinard, 2015 for a review) and lesions to LOC or its monkey homologue IT lead to impairments to correctly judge the physical as well as retinal size of objects (Chiou & Ralph, 2016; Cohen et al., 1994; Humphrey & Weiskrantz, 1969; Ungerleider, Ganz, & Pribram, 1977; Zeng et al., 2020). Distance processing as such is not well investigated in LOC, but it has e.g. also been found to be sensitive to scene-based information in the absence of objects, as the overall size of space (Park et al., 2015). This

combined evidence suggests an important role for LOC in processing retinal size information and integrating size and depth cues to estimate an objects' true size and distance. By rescaling object representations and making them independent of retinal size, such mechanisms may clearly also help to account for scene context and changing visual features during object recognition or search.

Another potential interpretation is however that we indeed found depth-dependent templates in both the search and scene training task. Given that the search task was undoubtedly challenging, participants may have tried to practice in the scene localizer task by extracting depth-information from the scenes and explicitly preparing to search for objects of a particular size, even though they never appeared as targets within that task. It is also important to note that the scene task still required participants to actively detect targets within either depth plane (even though these were not changing in size based on depth). Even more interesting from the point of view of naturalistic visual search, this may potentially reflect a more general preparatory mechanism expecting larger targets when searching nearby.

Dissociating between these accounts based on the current data is unfortunately not easily possible. Further studies may test whether similar effects are found without any explicit instructions to detect targets and consider depth-dependent size changes. This may reveal whether these effects reflect automatic mechanisms tuned to regularities in naturalistic scenes and search in a broader context. To this end, it may also be relevant to vary retinal size on a more fine-grained scale, as well as testing for the shape-specificity of these effects<sup>3</sup>. On the one hand, this would

---

<sup>3</sup> Within our experiment size-information per se was emphasized as the main feature of interest. Our decoding scheme focused on general size-

information, not tied to specific object shape and in fact shape in general could not be reliably decoded. As shape was blocked per run, different run-based

further reduce or eliminate any potential remaining influence of the visual overlap or simple conceptual associations between target objects and scene locations on size-decoding in LOC. Further, shape-specific effects (i.e. better decoding of the shape of the sought-for object when it also has the appropriate size) would indicate participants indeed formed templates reflecting the specific target objects in the search task. Ultimately however, independent of the specific processes involved, our findings likely reflect LOC's involvement in mechanisms helping to account for varying object appearance and accounting for scene context, underlining its importance in object recognition and search within naturalistic scenes.

In the current experiment there was however no evidence that scene-selective areas were involved in accounting for depth-related changes. Although above chance and consistent, decoding-accuracy for depth in both OPA and PPA was rather low compared to early visual cortex and did not correlate with size-decoding on a trial-by-trial basis. This may potentially be due to our specific depth-manipulation. To facilitate decoding of retinal size, depth in our experiment was dichotomous and did not require a very fine-grained estimate of distance. Further, searching near or far away did not change the overall depth-layout, navigational affordances or identity of the scenes, which are all known to strongly influence responses of scene-selective areas (Bonner & Epstein, 2017b; R. A. Epstein, 2005). Therefore, fixating near and far within the same scene may have elicited rather similar activation patterns compared to near fixations across scenes, leading to lower accuracy in

---

GLMs needed to be created for the different target objects. Thus, shape-information was likely captured by the general intercept of the GLMs rather than specific regressors, preventing shape-decoding. Further, if a target feature is kept constant across repeated searches, attentional capture and neuronal activation indicating active

depth across scenes. The fact that PPA generally showed the numerically lowest and slightly less consistent decoding may be in line with previous findings suggesting global and more identity-based scene-representation in PPA. In comparison, scene representations in OPA tend to be more sensitive to local aspects of a scene but generalizing better across different scene-identities (Bonner & Epstein, 2017b; Kamps et al., 2017).

Since the same scenes were used in the scene and search task, our cross-decoding captured scene-based distance information, but not necessarily an abstract representation of depth. Within our study, relatively low-level properties, as e.g. the spatial frequency content present in different parts of the scenes, could have reliably dissociated near from far fixations, explaining the high decoding accuracy in EVC. Such low-level features, processed by early visual regions may also have been sufficient to inform the representation of retinal size or distance in LOC, as suggested by their trialwise correlation (at least after accounting for potential influences of EVC and low-level scene-properties on size-decoding itself). Within even more complex and variable naturalistic environments, scene-selective areas may however still have a stronger contribution to visual search.

## General discussion

---

With two experiments, we aimed to investigate mechanisms underlying search in naturalistic scenes in which target size is modulated by distance. Within both tasks, behavioural evidence showed participants successfully took

maintenance in visual working memory decrease rapidly (Carlisle et al., 2011), suggesting transition to long term memory. Generally however, both size and shape information was required to identify targets and both therefore likely part of the template created.

retinal size changes into account during search. Using b-CFS, we were however not able to probe the template formed by participants, potentially due to a generally less visual nature of the template or insufficient sensitivity of our paradigm. Within our fMRI study, we found a consistent association between smaller objects and searching and fixating far away (and vice versa), specific to LOC. These effects were however not specific to our search instructions and can therefore not be interpreted as rescaled templates with certainty. They may however still reflect another mechanism through which the visual system could potentially account for depth-dependent size changes in template-based visual search.

An alternative to rescaling the template and matching it to varying visual input is to rescale an object representation based on its distance before comparing it to a template of fixed depth. A recent study indeed found evidence that scene context modulated the representation of object size before this object representation interacted with a memory template (Gayet & Peelen, 2019). Our current findings may generally be in line with either mechanism, reflecting differently-sized search templates formed also in the absence of search instructions or computing the distance of isolated objects and their respective scene-locations to rescale the perceptual input during object recognition. More research will be needed to fully understand which processes our findings in LOC reflect, and both may be flexibly used during real-world search, but they may generally highlight this area's relevance in recognizing and searching for objects despite their varying appearance.

## References

- Amit, E., Mehoudar, E., Trope, Y., & Yovel, G. (2012). Do object-category selective regions in the ventral visual stream represent perceived distance information? *Brain and Cognition*, *80*(2), 201–213. <https://doi.org/10.1016/j.bandc.2012.06.006>
- Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. <https://doi.org/10.1038/nrn1476>
- Battistoni, E., Stein, T., & Peelen, M. V. (2017). Preparatory attention in visual cortex. *Annals of the New York Academy of Sciences*, 1–16. <https://doi.org/10.1111/nyas.13320>
- Bonner, M. F., & Epstein, R. (2017a). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLOS Computational Biology* (Vol. 14). <https://doi.org/10.1101/177329>
- Bonner, M. F., & Epstein, R. A. (2017b). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(18), 4793–4798. <https://doi.org/10.1073/pnas.1618228114>
- Brainard, D. H. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4), 433–436. <https://doi.org/10.1163/156856897X00357>
- Carlisle, N. B., Arita, J. T., Pardo, D., & Woodman, G. F. (2011). Attentional templates in visual working memory. *Journal of Neuroscience*, *31*(25), 9315–9322. <https://doi.org/10.1523/JNEUROSCI.1097-11.2011>
- Castelhano, M. S., & Henderson, J. M. (2007). Initial Scene Representations Facilitate Eye Movement Guidance in Visual Search. *Journal of Experimental Psychology: Human Perception and Performance*, *33*(4), 753–763. <https://doi.org/10.1037/0096-1523.33.4.753>
- Cate, A. D., Goodale, M. A., & Köhler, S. (2011). The role of apparent size in building- and object-specific regions of ventral visual cortex. *Brain Research*, *1388*, 109–122. <https://doi.org/10.1016/j.brainres.2011.02.022>
- Chelazzi, L., Duncan, J., Miller, E. K., & Desimone, R. (1998). Responses of neurons in inferior temporal cortex during memory-guided visual search. *Journal of Neurophysiology*, *80*(6), 2918–2940.

- <https://doi.org/10.1152/jn.1998.80.6.2918>
- Chelazzi, L., Miller, E. K., Duncan, J., & Desimone, R. (1993). A neural basis for visual search in inferior temporal cortex. *Nature*, *363*(6427), 345–347. <https://doi.org/10.1038/363345a0>
- Chiou, R., & Ralph, M. A. L. (2016). Task-related dynamic division of labor between anterior temporal and lateral occipital cortices in representing object size. *Journal of Neuroscience*, *36*(17), 4662–4668. <https://doi.org/10.1523/JNEUROSCI.2829-15.2016>
- Cohen, L., Gray, F., Meyrignac, C., Dehaene, S., & Degos, J.-D. (1994). Selective deficit of visual size perception: two cases of hemimicropsia. *Journal of Neurology, Neurosurgery, and Psychiatry*, *57*, 73–78.
- Desimone, R. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*(1), 193–222. <https://doi.org/10.1146/annurev.neuro.18.1.193>
- Desimone, R. (1998). Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353*(1373), 1245–1255. <https://doi.org/10.1098/rstb.1998.0280>
- Desimone, R., & Duncan, J. (1995). Neural Mechanisms of Selective Visual Attention. *Annual Review of Neuroscience*, *18*, 193–222. <https://doi.org/10.1146/annurev-psych-122414-033400>
- Duncan, J., & Humphreys, G. W. (1989). Visual Search and Stimulus Similarity. *Perception & Psychophysics*, *54*(6), 716–732. <https://doi.org/10.3758/BF03211797>
- Eckstein, M. P., Drescher, B. A., & Shimozaki, S. S. (2006). Attentional cues in real scenes, saccadic targeting, and Bayesian priors. *Psychological Science*, *17*(11), 973–980. <https://doi.org/10.1111/j.1467-9280.2006.01815.x>
- Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes. *Current Biology*, *27*(18), 2827–2832.e3. <https://doi.org/10.1016/J.CUB.2017.07.068>
- Ehinger, K. A., Hidalgo-Sotelo, B., Torralba, A., & Oliva, A. (2009). Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual Cognition*, *17*(6–7), 945–978. <https://doi.org/10.1080/13506280902834720>
- Eimer, M. (2014). The neural basis of attentional control in visual search. *Trends in Cognitive Sciences*, *18*(10), 526–535.
- Epstein, R. A. (2005). The cortical basis of visual scene processing. *Visual Cognition*, *12*(6), 954–978. <https://doi.org/10.1080/13506280444000607>
- Epstein, R., & Kanwisher, N. (1998). The parahippocampal place area: A cortical representation of the local visual environment. *NeuroImage*, *7*(4 PART II), 6–9. [https://doi.org/10.1016/s1053-8119\(18\)31174-1](https://doi.org/10.1016/s1053-8119(18)31174-1)
- Ferrara, K., & Park, S. (2016). Neural representation of scene boundaries. *Neuropsychologia*, *89*, 180–190. <https://doi.org/10.1016/j.neuropsychologia.2016.05.012>
- Gayet, S., Douw, I., van der Burg, V., Van der Stigchel, S., & Paffen, C. L. E. (2018). Hide and seek: Directing top-down attention is not sufficient for accelerating conscious access. *Cortex*. <https://doi.org/10.1016/J.CORTEX.2018.08.027>
- Gayet, S., Paffen, C. L. E., & Van der Stigchel, S. (2013). Information Matching the Content of Visual Working Memory Is Prioritized for Conscious Access. *Psychological Science*, *24*(12), 2472–2480. <https://doi.org/10.1177/0956797613495882>
- Gayet, S., & Peelen, M. V. (2019). Scenes Modulate Object Processing Before Interacting With Memory Templates. *Psychological Science*, *095679761986990*. <https://doi.org/10.1177/0956797619869905>
- Gayet, S., & Stein, T. (2017). Between-subject variability in the breaking continuous flash suppression paradigm: Potential causes, consequences, and solutions. *Frontiers in Psychology*, *8*(MAR), 1–11. <https://doi.org/10.3389/fpsyg.2017.00437>
- Gayet, S., Van Der Stigchel, S., & Paffen, C. L. E. (2014). Breaking continuous flash suppression: Competing for consciousness on the pre-semantic battlefield. *Frontiers in Psychology*, *5*(MAY), 1–10. <https://doi.org/10.3389/fpsyg.2014.00460>
- Gayet, S., van Maanen, L., Heilbron, M., Paffen, C.

- L. E., & Van der Stigchel, S. (2016). Visual input that matches the content of visual working memory requires less (not faster) evidence sampling to reach conscious access. *Journal of Vision, 16*(11), 26. <https://doi.org/10.1167/16.11.26>
- Greene, Michelle, M., & Oliva, Aude, M. (2009). The Briefest of Glances: The Time Course of Natural Scene Understanding. *Psychological Science, 20*(4), 464–472.
- Grill-Spector, K. (2003). The neural basis of object perception. *Current Opinion in Neurobiology, 13*(2), 159–166. [https://doi.org/10.1016/S0959-4388\(03\)00040-0](https://doi.org/10.1016/S0959-4388(03)00040-0)
- Grill-Spector, K., & Malach, R. (2001). fMR-adaptation: A tool for studying the functional properties of human cortical neurons. *Acta Psychologica, 107*(1–3), 293–321. [https://doi.org/10.1016/S0001-6918\(01\)00019-1](https://doi.org/10.1016/S0001-6918(01)00019-1)
- Gunseli, E., Meeter, M., & Olivers, C. N. L. (2014). Is a search template an ordinary working memory? Comparing electrophysiological markers of working memory maintenance for visual search and recognition. *Neuropsychologia, 60*(1), 29–38. <https://doi.org/10.1016/j.neuropsychologia.2014.05.012>
- Hebart, M. N., Görden, K., & Haynes, J. D. (2015). The decoding toolbox (TDT): A versatile software package for multivariate analyses of functional imaging data. *Frontiers in Neuroinformatics, 8*(JAN), 1–18. <https://doi.org/10.3389/fninf.2014.00088>
- Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid Invariant Encoding of Scene Layout in Human OPA. *Neuron, 103*(1), 161–171.e3. <https://doi.org/10.1016/j.neuron.2019.04.014>
- Humphrey, N. K., & Weiskrantz, L. (1969). Size constancy in monkeys with inferotemporal lesions. *The Quarterly Journal of Experimental Psychology, 21*(3), 225–238. <https://doi.org/10.1080/14640746908400217>
- Jiang, Y., Costello, P., & He, S. (2007). Processing of Invisible Stimuli: Advantage of Upright Faces and Recognizable Words in Overcoming Interocular Suppression. *Psychological Science, 18*(4), 349–355. <https://doi.org/10.1111/j.1467-9280.2007.01902.x>
- Kamps, F. S., Julian, J. B., Kubilius, J., Kanwisher, N., & Dilks, D. D. (2017). The occipital place area represents the local elements of scenes. *Physiology & Behavior, 176*(1), 139–148. <https://doi.org/10.1016/j.physbeh.2017.03.040>
- Kastner, S., & Ungerleider, L. G. (2001). The neural basis of biased competition in human visual cortex. *Neuropsychologia, 39*(12), 1263–1276. [https://doi.org/10.1016/S0028-3932\(01\)00116-6](https://doi.org/10.1016/S0028-3932(01)00116-6)
- Konkle, T., & Oliva, A. (2012). A Real-World Size Organization of Object Responses in Occipitotemporal Cortex. *Neuron, 74*(6), 1114–1124. <https://doi.org/10.1016/J.NEURON.2012.04.036>
- Kornblith, S., Cheng, X., Ohayon, S., & Tsao, D. Y. (2013). A network for scene processing in the macaque temporal lobe. *Neuron, 79*(4), 766–781. <https://doi.org/10.1016/j.neuron.2013.06.015>
- Liu, D., Wang, L., Wang, Y., & Jiang, Y. (2016). Conscious Access to Suppressed Threatening Information Is Modulated by Working Memory. *Psychological Science, 27*(11), 1419–1427. <https://doi.org/10.1177/0956797616660680>
- Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage, 59*(3), 2636–2643. <https://doi.org/10.1016/j.neuroimage.2011.08.076>
- Neider, M. B., & Zelinsky, G. J. (2006). Scene context guides eye movements during visual search. *Vision Research, 46*(5), 614–621. <https://doi.org/10.1016/j.visres.2005.08.025>
- Pan, Y., Lin, B., Zhao, Y., & Soto, D. (2014). Working memory biasing of visual perception without awareness. *Attention, Perception, and Psychophysics, 76*(7), 2051–2062. <https://doi.org/10.3758/s13414-013-0566-2>
- Park, S., Konkle, T., & Oliva, A. (2015). Parametric Coding of the Size and Clutter of Natural Scenes in the Human Brain. *Cerebral Cortex, 25*(7), 1792–1805. <https://doi.org/10.1093/cercor/bht418>
- Peelen, M. V., & Kastner, S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex. *Proceedings of the*

- National Academy of Sciences*, 108(29), 12125–12130.  
<https://doi.org/10.1073/pnas.1101042108>
- Reeder, R. R., & Peelen, M. V. (2013). The contents of the search template for category-level search in natural scenes. *Journal of Vision*, 13(3), 13–13.  
<https://doi.org/10.1167/13.3.13>
- Sawamura, H., Georgieva, S., Vogels, R., Vanduffel, W., & Orban, G. A. (2005). Using functional magnetic resonance imaging to assess adaptation and size invariance of shape processing by humans and monkeys. *Journal of Neuroscience*, 25(17), 4294–4306.  
<https://doi.org/10.1523/JNEUROSCI.0377-05.2005>
- Soon, C. S., Namburi, P., & Chee, M. W. L. (2013). Preparatory patterns of neural activity predict visual category search speed. *NeuroImage*, 66, 215–222.  
<https://doi.org/10.1016/J.NEUROIMAGE.2012.10.036>
- Sperandio, I., & Chouinard, P. A. (2015). The mechanisms of size constancy. *Multisensory Research*, 28(3–4), 253–283.  
<https://doi.org/10.1163/22134808-00002483>
- Stein, T. (2019). The Breaking continuous flash suppression paradigm. In *Transitions Between Consciousness and Unconsciousness* (pp. 1–39).
- Ungerleider, L. G., Ganz, L., & Pribram, K. H. (1977). Size constancy in rhesus monkeys: Effects of pulvinar, prestriate, and inferotemporal lesions. *Experimental Brain Research*, 27(3–4), 251–269.  
<https://doi.org/10.1007/BF00235502>
- Wolfe, J. M. (1994). Guided Search 2.0 A revised model of visual search, 1(2), 202–238.
- Wolfe, J. M. (2017). Dispatches Visual Attention : Size Matters. *Current Biology*, 27(18), R1002–R1003.  
<https://doi.org/10.1016/j.cub.2017.07.057>
- Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I., & Sherman, A. M. (2011). Visual search for arbitrary objects in real scenes. *Attention, Perception, & Psychophysics*, 73(6), 1650–1671.  
<https://doi.org/10.3758/s13414-011-0153-3>
- Wolfe, J. M., & Horowitz, T. S. (2004). What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6), 495–501.  
<https://doi.org/10.1038/nrn1411>
- Wolfe, J. M., Võ, M. L. H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, 15(2), 77–84.  
<https://doi.org/10.1016/j.tics.2010.12.001>
- Zeng, H., Fink, G. R., & Weidner, R. (2020). Visual size processing in early visual cortex follows lateral occipital cortex involvement. *The Journal of Neuroscience*, 40(22), JN-RM-2437-19.  
<https://doi.org/10.1523/jneurosci.2437-19.2020>

## Acknowledgement

---

There are many people who I would like to thank for their contribution. First, I was lucky to have two great advisors who have encouraged and challenged me, helped me to develop as a scientist and who always found time for me. I would like to thank Marius Peelen for giving me the opportunity to join his group, his continuous guidance and valuable advice related to the project and my academic future as well as his trust in me. Surya Gayet for his great day-to-day supervision, humour, patient explanations, and answering endless emails and questions. Writing this thesis made me realize again how much I was able to learn during this time and I am grateful for this experience.

Further, I would like to thank the entire Peelen lab for warmly welcoming me in their midst and making me feel part of the group. I greatly enjoyed stimulating discussions, group lunch, lab activities, drinks and gala dances. In particular my intern-room buddy Alexandra for shared moments of success and despair.

I would also like to thank my friends and classmates, especially Kirsten, Lisa, Vera and Yingjie who have supported me in the past two years and filled this time with countless memories and laughter.

Lastly, my family who have been there for me throughout my life, helped me make it through challenging times and the current pandemic, kept me sane (most of the time, sometimes the opposite) and supported all my endeavours. My mom for believing in me and always encouraging my curiosity. My brother for putting my feet back on the ground and the occasional glass of whiskey when it was needed. And Ulli. For everything you taught me in life, most importantly to trust myself, and your help making it to this point. Wish you were here.