Doctoral thesis

Investigating the role of image meaning and prior knowledge in human eye-movements control

Marek Antoni Pędziwiatr

ORCID iD: 0000-0002-3959-8666

Supervisor: Dr Christoph Teufel

A thesis submitted in fulfillment of the requirements for the degree of

Doctor of Philosophy (Psychology)

March 2021

Table of Contents

Summary

Humans sample visual information by making eye movements towards different parts of their surroundings. Understanding what guides this sampling process is an important goal of vision science, and the present thesis is a contribution to this endeavour. Chapter One provides an overview of factors influencing human eye movements, which are typically divided into bottom-up (stimulus-dependent) and top-down (observer-dependent) processes. One of the challenges in studying these factors stem from the fact that they are often difficult to operationalize in a precise, unambiguous way. This is particularly problematic for semantic information contained in visual scenes ('image meaning'), a top-down factor which is the backbone of the recently proposed framework for understanding human eye movements: the meaning maps approach. Chapter Two evaluates this approach and demonstrates that meaning maps – a crowd-sourced method designed to quantify the distribution of meaning in natural scenes – might be sensitive to complex visual features, rather than meaning. Chapter Three builds on that finding and shows that contextualized meaning maps, the most recent variant of the original meaning maps, share the limitations of their predecessors. Chapter Four adopts a novel perspective on eye-movement control and focuses on the interactions between image features (a bottom-up factor) and prior object-knowledge possessed by an observer (a top-down factor). Specifically, it shows that the same stimuli – black and white, Mooney-style two-tone images – are looked at differently depending on whether the observer possesses object-knowledge that enables them to bind images into coherent percepts of objects. The final chapter summarizes the thesis and maps the future directions for studies on eye movements. Taken together, findings reported here indicate that while top-down factors such as prior object-knowledge play a crucial role in guiding human gaze, the tools to study them offered by the meaning maps approach still need to be improved.

Acknowledgments

Kraków, 8.01.2021
setRng(2706)

Writing this thesis was possible because of the tremendous help and support from many people, to whom I am greatly indebted. Acknowledging their role in this endeavour is a bare minimum I can do to express my deep gratitude to them.

First, I would like to thank my advisor, Dr. Christoph Teufel. Four years ago, Christoph believed in my determination, despite my limited knowledge about human vision, and has supported me grow and develop into the researcher I am today, who now knows a thing or two about (vision) science. Being mentored by Christoph has undoubtedly been the most valuable aspect of my PhD journey. Now, as this journey is coming to an end, and I am moving on from Christoph's lab, I am left with a feeling there is still so much I can learn from him. I am sure I will continue to seek invaluable advice and expertise from Christoph as I progress through my career as a scientist.

Next, I would like to thank my second thesis advisor, Dr. Elisabeth von dem Hagen and my collaborators, with whom I worked on various projects related to my PhD (listed in alphabetical order): Prof. Matthias Bethge, Adelina-Mihaela Halchin, Dr. Matthias Kümmerer, and Prof. Thomas S.A. Wallis. I owe special thanks to Prof. Michał Wierzchoń for giving me the opportunity to be a satellite member of his research group C-Lab, and for his help and support on various occasions.

As a PhD student at Cardiff University, I have had the privilege of belonging to several inspiring communities: Christoph's lab group, Cardiff University Brain Research Imaging Centre (CUBRIC) community, and the School of Psychology community. I owe a lot to each of them, as each enriched me in a unique way.

During the time of my PhD, personal relationships were equally important as the professional ones. First, I owe immense gratitude to my parents, who have provided me with unconditional support throughout this process. They have taught me resilience, inspired me, and equipped me with the resources to overcome obstacles I encountered along this PhD journey. Second, I would like to thank all my friends and other people whose presence in my life during the last four years remained a source of joy, support, and the opportunities to grow (listed in alphabetical order): Adela I., Agata B., Agnieszka S., Aleksandra S., Bob. D., David M., David McG., Dominik K., Isobel W., Jerzy Antoni N., Jeżowiec A., Kora J., Mariola U., Mateusz Ś., Monika D., Nadziejka F., Natalia M., Paulina K., Rychu P., Stefan N., Tomasz K., and Wojciech Z. I would particularly like to single out Monika D. and Stefan N, invaluable friends who – also being doctoral students – accompanied me at all stages of my studies.

Finally, I would like to thank the community of academics on Twitter, where I have felt connected to, and inspired by, the global academic world.

Chapter One – general introduction

Eye movements and vision

Eye movements are integral to our ability to see. Our visual system is structured in such a way that we can see fine detail only in a small patch in the centre of our visual field called the fovea. The resolution drops off rapidly as a stimulus moves away from central vision (Anstis, 1974; Rosenholtz, 2016; Sloan, 1961; Stewart, Valsecchi, & Schütz, 2020). This restriction of our visual system is thought to be due to the metabolic costs of high-resolution perception, and the spatial constraints of the human skull (Akbas & Eckstein, 2017; Schwartz, 1994). A brain that would allow us to enjoy the same amount of detail found in the fovea in all locations of the visual field, would have to be much larger than it is now. Humans and many other visually-oriented animals have found an elegant and efficient way that allows combining a large visual field with high-acuity vision: we use saccades – rapid, stereotypical eye movements – to orient the high-resolution part of our visual system successively to different parts of a visual scene. Information about these small scene parts is extracted during fixations – short periods in which the eyes are relatively stable – and this local information is used to gradually build up the scene representation that we consciously experience (Rolfs, 2015; Wurtz, Joiner, & Berman, 2011). Thus, due to the structure of our visual system, human vision depends on eye movements. How the brain decides where to look is therefore an important question that has attracted considerable attention from a wide range of different fields, ranging from cognitive psychology and neuroscience to computer science and machine vision. This thesis is a part of this interdisciplinary endeavour. It makes a contribution towards understanding human eye-movements control in the context of natural-scene viewing – the situation when individuals look at static depictions of real-world environments or objects (e.g., a photograph).

Studying eye movements: in the lab and in the wild

Studying oculomotor behaviour involves recording the eye movements of individuals. The current eye-tracking technology allows these recordings in two different settings. In the first setting, participants wear eye-tracking glasses and move around in the environment in an unconstrained fashion (e. g., walk down the street; Foulsham et al., 2011) or perform real-world

tasks requiring complex motor activity, e. g. make a sandwich (Hayhoe et al., 2003) or play cricket (Land & McLeod, 2000). In the second setting, eye movements are recorded in the laboratory, where the participants – frequently called 'observers', in line with the convention used in vision science – look at stimuli displayed on a computer screen. The key difference between both settings pertains to the extent to which the participants can control the visual input in ways other than by their eye movements. In the first setting, participants control the possible inputs that are then explored by eye movements. In the second setting, this is controlled by the experimenter. While the first, 'real world', setting is ideal to investigate the overall sampling strategy of the organism, it provides challenges when a researcher wants to isolate the distinct contribution of the eye movements to that strategy.

This thesis describes experiments conducted in a laboratory. Such setting has a number of advantages compared to the first approach. The laboratory-based studies enable exposing observers to a large range of stimuli, including artificial stimuli, which are unlikely to be encountered in the real world. Moreover, a laboratory-based approach provides far more control over stimulus properties and enables studying eye-movement behaviours which are uncommon in everyday life, such as making saccades in the direction opposite to the direction indicated by a visual cue (antisaccade task; Hallett, 1978).

Despite these advantages, lab-based studies have been criticized because of the lack of ecological validity (Foulsham & Kingstone, 2017; Tatler, Hayhoe, Land, & Ballard, 2011b). This critique was targeted predominantly at studies that used a similar approach to the experiments described in Chapters One and Two of this thesis, in which observers viewed photographs of natural scenes on a computer screen. Such situation creates the conditions for visual exploration which differ a lot from a major part of everyday human experience. Specifically, the laboratory-based studies impose strong constraints on what the observer is allowed to do – for example, they usually cannot move their head. Next, usually, the onset of images presented in experiments is sudden, which, again, is uncommon 'in the wild' and is known to affect visual processing (Dorr, Martinetz, Gegenfurtner, & Barth, 2010; see also Wu et al., 2013). Further aspects of viewing images in a laboratory which makes this situation different from exploring visual environment naturally are the fact that images are framed by the screen (which

introduce viewing biases; Bindemann, 2010) and the lack of binocular depth cues resulting from the two-dimensionality of images.

Indeed, there are cases when individuals exhibit different behaviours in the lab and 'in the wild'. A striking example here is the finding that while people have a strong tendency to direct their eyes at faces on images (Cerf, Paxon Frady, & Koch, 2009; Flechsenhar & Gamer, 2017), in the real world they rather avoid looking at strangers directly and resort to monitoring other people's behaviour by means of covert attention (Dosso, Huynh, & Kingstone, 2020). This finding clearly warrants caution when generalising from lab-based effects to real-world behaviour. Note, however, that this example pertains only to a specific kind of stimulus and a specific effect. It is not clear whether similar caveats exist for more general effects.

The concerns regarding lab-based eye-movement studies are particularly pressing if it is assumed that investigating real-world situations, which involve acting in the environment and interacting with it by means of a whole body, is the main goal of research on eye movements in natural-scene viewing. However, this definition of what constitutes a 'real-world situations' does not take into account the fact that looking at a screen with rapidly changing images is a common human activity, at least in industrialised societies, where the access to computers, television sets and smartphones is ubiquitous. Despite the availability of technology that allows studying eye movements in real-world settings, the ubiquity of this image-viewing situation keeps fuelling the interest in both basic and applied research on this process. Furthermore, photographs in laboratory-based eye-tracking experiments can be viewed as stimuli in and of themselves, and not just the imperfect proxies of the real-world environments. For example, when treated as such, they have been proven to be useful tools for revealing the individual differences in gaze behaviour related to cultural origin (Chua, Boland, & Nisbett, 2005; Goh, Tan, & Park, 2009) or personality traits (De Haas, Iakovidis, Schwarzkopf, & Gegenfurtner, 2019). For instance, a recent study by de Haas and colleagues (2019) demonstrated that individuals exhibit large and robust differences in the amount of time they spend fixating elements of scenes possessing certain semantic attributes, like 'being text', 'being related to motion' or 'being a face'. Moreover, the proportion of first fixations after image onset made by individuals towards faces was correlated with their face recognition skills.

Eye movements and attention

Before proceeding to outlining the possible ways of addressing the question of what guides human gaze in natural-scene viewing, it is worthwhile asking a meta-question: what does one really investigate, when investigating human gaze? One typical answer would be that eye movements provide a means to index 'spatial attention'. The tight coupling of gaze location and the locus of attention is both an intuition common among laymen (Guterstam, Kean, Webb, Kean, & Graziano, 2019) and a well-supported empirical finding (Deubel & Schneider, 1996). However, there are two reasons for which I refrain from using the term 'attention' when referring to gaze position in this thesis. First, attention can operate at the locations, which are not fixated: the distinction between overt (i.e., operating at the fixated location) and covert (operating at the location which is not fixated) attention is one of the cornerstones of cognitive psychology (Posner, 1980). Therefore, even if eye movements and attention are often linked, equating changes of gaze position with changes of the locus of attention is not always justified. Second, even in cases when the locus of spatial attention and a fixation-location are aligned, there are still many possible types of attention to be considered, for example object-based attention (Duncan, 1984; Egly, Driver, & Rafal, 1994) or features-based attention (Maunsell & Treue, 2006; Rossi & Paradiso, 1995). More generally speaking, there is an abundance of theories of attention (Logan, 2004), so this term can be interpreted in different ways, depending on the theory of attention, to which a reader subscribes. Therefore, considering the allocation of fixations as a sampling process, carries fewer theoretical assumptions and is better grounded in the data than using the term 'attention'. This terminological choice is shared by the majority of authors, who I cite in this thesis (albeit not by all of them – for example, see Bylinskii et al., 2015 or Henderson, 2020).

Bottom-up vs. top-down dichotomy: an overview

Focusing on the role the eye movements play in vision – propelled and propagated by an approach to studying vision called active vision (Berman & Colby, 2009; Gilchrist & Findlay, 2001) – led to recognising that understanding this sampling process is essential for understanding the primate visual system. Therefore, answering the questions about what guided eye movements became an object of intensive investigation. In almost all of the works,

which I cite in this thesis, the factors that are thought to guide eye movements are divided into two broad categories. The first category encompasses image-computable visual features, which are processed in a bottom-up manner. The second category relates to internal states of the observer, influencing stimulus processing in a top-down way. This 'bottom-up vs. top-down' dichotomy is not specific to this particular research field; it rather reflects the historic development of how researchers think about perception in domains such as psychology, vision science, and philosophy (Cavanagh, 2011; Firestone & Scholl, 2015; Pylyshyn, 1980; Teufel & Nanay, 2016) but see Awh, Belopolsky, & Theeuwes, 2012).

However, before proceeding to that review, it must be noted that the dichotomy between bottom-up and top-down processes does not capture the nature of all factors influencing oculomotor behaviour. In many studies, the additional factors, considered separately, are spatial biases (for example, see Kollmorgen et al., 2010 or Benjamin W. Tatler et al., 2005). Spatial biases are tendencies to fixate certain scene regions more than others. The most well-known bias is the centre bias: observers have the tendency to fixate central image region more often than the regions closer to image edges (Tatler & Vincent, 2008). This bias is especially prominent at the beginning of image viewing which might indicate that it is a part of an involuntary response elicited by sudden image onset (Bindemann, 2010; Rothkegel, Trukenbrod, Schütt, Wichmann, & Engbert, 2017). Its persistence throughout image viewing might, in turn, reflect the fact that interesting objects on photographs tend to be located in their central region (Tseng et al., 2009, but see Benjamin W. Tatler, 2007). A second, less prominent, bias is a leftwards bias: the tendency to direct more fixations towards the left-hand side of an image (Nuthmann & Matthias, 2014). It has been hypothesized that this bias – at least in samples from Western cultures – is related to the fact that observers read texts starting from the left-hand side (Foulsham, Gray, Nasiopoulos, & Kingstone, 2013). An alternative, less culture-specific, explanation of this bias is that it originates from the hemispheric asymmetries in the brain's attentional system (Ossandón, Onat, & König, 2014).

Bottom-up processing, visual features, and saliency models

The idea that image features can guide eye movements is rooted in the Feature Integration Theory, an influential theory of attention proposed by Treisman and Gelade (1980). In brief, according to this theory, different low-level features of the input – such as colour or orientation

– are analysed in parallel and combined into a 'master map' (A. Treisman & Gormican, 1988). This map codes for the 'interestingness' or 'saliency' of each location in the visual field based on the low-level features in that area. Feature Integration Theory posits that attention – like a spotlight – serially visits the most salient locations on the master map and integrates the features encountered there into objects.

The concept of a saliency map easily lends itself to computational modelling. This fact, together with a wider access to modern computers, sparked an outpouring of attempts to build computational models of attentional selection (Itti & Koch, 2000, 2001; Itti, Koch, & Niebur, 1998; Koch & Ullman, 1987). The core assumption of these models was that attention operates like an outlier detector: it is attracted to locations which are salient because they are distinct from the surroundings in terms of their visual properties. This assumption, as well as many other aspects of these early models, were derived from experimental results obtained using electrophysiological and behavioural methods (Itti & Koch, 2000; Koch & Ullman, 1985). Therefore, the early saliency models were aimed at modelling one specific, highly abstract process: pre-attentive selection of locations within the visual field for attentional inspection.

In the early days of saliency modelling, researchers attempting to compare the predictions of saliency models to human behaviour struggled due to the challenging task of operationalizing pre-attentive visual selection (Wloka, Kotseruba, & Tsotsos, 2017). These difficulties led to a gradual shift towards modelling the spatial distributions of fixations on visual stimuli such as images (for example, see Itti et al., 1998). This problem turned out to be more tractable, and the models built to predict the location of fixations could be more readily evaluated and compared with each other (Borji & Itti, 2013; Kümmerer et al., 2020) than their more ambitious predecessors. A side-effect of this shift towards modelling eye movements rather than pre-attentive selection was a gradual change of meaning of the word saliency – it drifted towards indicating any image property (indexed by a model) that is predictive of human fixations (Schütt, Rothkegel, Trukenbrod, Engbert, & Wichmann, 2019).

Initial studies that tested predictions about the relationship between fixation locations and image features, which were derived from the derived from the saliency modelling framework, yielded promising results (Einhäuser & König, 2003; Krasovskaya & MacInnes, 2019; Krieger,

Rentschler, Hauske, Schill, & Zetzsche, 2000; Kümmerer et al., 2020; Reinagel & Zador, 1999). For instance, the models correctly predicted differences in visual properties of regions that were fixated compared to those that were not fixated (Reinagel & Zador, 1999), and their predictions regarding the locations of fixation were consistently better than chance (Kümmerer et al., 2020). Moreover, the models carried the potential to be developed further and, in consequence, be able to account for even a larger share of human behaviour (Itti & Koch, 2001; Krasovskaya & MacInnes, 2019). This potential was mainly due to the map-like format of the predictions they generate – it made it easy to extend the models by implementing mechanisms operating 'on top' of the maps generated by the models, in order to account for additional processes guiding oculomotor behaviour without sacrificing the main advantage of the map-like format, that is, the ease with which model predictions can be compared against human data (for example, see Adeli et al., 2017; Torralba et al., 2006).

This initial enthusiasm was curbed by a string of studies highlighting both the limitations of saliency models and problems with the theoretical framework within which they were conceptualised (Tatler, Hayhoe, Land, & Ballard, 2011a). Here, I list four such issues, some of which are further explored in the next sections of this thesis. First, observers performing a visual tasks, when it is necessary, are able to 'decouple' their gaze from the visual features highlighted as salient by the models. In other words, they are able to ignore saliency, if this if necessary to perform a given task (Einhäuser, Rutishauser, & Koch, 2008; Foulsham & Underwood, 2007). Second, specific classes of stimuli – for example, social signals such as faces – tend to strongly attract fixations, an effect not modelled by early saliency models (Cerf et al., 2009; Flechsenhar & Gamer, 2017). Third, it has been demonstrated that the predictive power of saliency models is moderate at best, especially when they are test on a wide range of natural scene stimuli (Kümmerer et al., 2020). Fourth, the idea that fixations should land primarily on visually salient image-locations results in predictions, which are counter-intuitive and turned out to be inaccurate. For example, individuals viewing scenes should direct their gaze at the sources of light because they are much brighter than their surroundings and thereby highly salient. Another example is the model prediction that fixations should frequently land on object edges, because they are indicated by highly salient sudden changes in image-feature values (colour, for example). However, it is now known that human observer do not exhibit either of these two predicted behaviours (Nuthmann & Einhäuser, 2015; Nuthmann &

Henderson, 2010; Nuthmann, Schütz, & Einhäuser, 2020; Stoll, Thrun, Nuthmann, & Einhäuser, 2015; Vincent, Baddeley, Correani, Troscianko, & Leonards, 2009).

In order to address these shortcomings, more recently developed models took three different approaches. First, some researchers tried to improve saliency models by incorporating characteristics of the human visual system, which earlier attempts had ignored (Adeli et al., 2017; Bruce, Bruce, Tsotsos, & Tsotsos, 2009; Zhang et al., 2008). A good example of this approach is the Adaptive-Whitening Saliency (AWS) model, which I use in Chapter Two of this thesis (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012; Garcia-Diaz, Leboran, et al., 2012). It implements a biologically-inspired mechanism, which removes correlations between maps indexing the distributions of different visual features (whitens the image representation used by the model). Implementing this mechanism results in the model being able to predict fixations better than the previous models. This example is part of a line of research that is actively pursued by a number of research groups, with novel models that implement ever more sophisticated biologically-inspired mechanisms being constantly proposed (for example, see Berga & Otazu, 2018 or Uejima et al., 2020).

A second approach in the development of saliency models involved moving away from biologically-inspired principles, with the sole aim of maximizing the model's predictive power (Wloka et al., 2018). The Graph-Based Visual Saliency (GBVS) model I use in Chapter Two (Harel, Koch, & Perona, 2007) is an example of this approach. This model relies on detecting 'outliers' on maps that index different image features similar to the earlier saliency models. However, the way in which these maps are combined is derived from graph theory – a branch of mathematics – rather than biology or psychology.

In the most recent developments of saliency modelling, the models move even further away from the initial notion of saliency. These models are purely data-driven. The first model developed in this line of research was the one by Kienzle, Wichmann, Schölkopf, and Franz (2007). These authors collected eye-tracking data for a set of natural scenes and extracted image-patches around the locations fixated by human observers. They used these patches to train a machine learning algorithm which served as a basis for predicting fixation locations in novel images. This initial attempt inspired further development of data-driven, machine-

learning based saliency models. The latest developments within this line of research are models based on training of deep neural networks (Kümmerer, Wallis, Gatys, & Bethge, 2017; Thomas, 2016). These models outperform all their predecessors by a large margins. I use two such models in Chapter Two, and they are described in greater detail there. In general, models belonging to that class capitalize on two properties of deep neural networks: their ability to reliably extract vast numbers of visual features from images and the relative ease with which these networks can be fine-tuned to specific applications (Krizhevsky, Sutskever, & Hinton, 2012; Storrs & Kriegeskorte, 2019). The first of these properties is acquired by a network during training, that is, gradual adjusting of the network's parameters which is guided by an optimization process aiming at maximizing performance in a given task. The second property is related to the fact that parts of such a trained network can serve as a mechanism for extracting visual features which are then applied in some other task. For example, for an object recognition task, training would require providing a network with a large set of labelled images, and a goal of the optimization process would be to minimize the number of images from this set that are labelled by the network incorrectly (Simonyan & Zisserman, 2015). Once properly trained, the network acquires sensitivity to different visual features which enable correct labelling of novel images. Next, the part responsible for extracting these features can be 'transplanted' to another network which, after training on appropriate data, can predict where people look at images (see Chapter Two). Importantly, the first, 'original' training equips the network with a sensitivity to a vast number of features, and the second one requires much less data.

Saliency models – evaluation metrics

A side-product of the intense research into saliency models was the development of methods for evaluating and comparing their performance. The problem of model evaluation comes down to the question of how well a smooth distribution over an image – a saliency map produced by a model – predicts the distribution of discrete fixation-points on that image. There are many model-performance metrics, which address this issue and their properties are thoroughly characterized (Bylinskii, Judd, Oliva, Torralba, & Durand, 2016; Kümmerer, Wallis, & Bethge, 2015; Wilming, Betz, Kietzmann, & König, 2011). In this thesis, I relied on two of them: sAUC (in Chapter Two) and correlation (in Chapter Two, Three, and Four).

The abbreviation sAUC stands for 'shuffled area under the curve', where the 'curve' refers to a receiver operating characteristic (ROC) curve. This measure is derived from signal detection theory. To calculate sAUC (Tatler et al., 2005; Zhang et al., 2008), a saliency map is thresholded at different values. For each threshold, the true positive rate is calculated as the proportion of fixation points which fall within map regions having values higher than the threshold. The false positive rate is calculated in an equivalent way. Note that for predictions of a hypothetical saliency model that would be devoid of any predictive power (that is, a model completely unable to discriminate between fixated and not fixated image locations) the sAUC value would amount to 0.5. For a perfect saliency model, in turn, it would amount to 1. Importantly false positives in sAUC are not based on fixations from the image, for which model predictions are generated. Rather, fixations from other images, usually presented as part of the same experiment, are considered. The reason for this procedure relates to the centre bias, the tendency to allocate more fixations to the central regions of an image (Tatler, 2007). Models, which treat central image regions as more salient irrespective of their content, can therefore increase their predictive power compared to models, which do not up-weight central regions, without increasing their sensitivity to image-properties relevant for eye movements. Because of the centre bias, fixations of images other than the one, for which a prediction is generated, should clustered around the image centre without being related to image properties. Therefore, using such fixations to calculate false positive rates, as done in sAUC, penalises models that rely too much on biases when generating predictions.

The second measure I use in all empirical Chapters of this thesis is Pearson's linear correlation coefficient, dubbed correlation (Bylinskii et al., 2016). This metric is calculated by generating a 2D Pearson's linear correlation for two smooth distributions over an image (typically, both derived from fixations or one derived from fixations and the other being model predictions). It measures the degree of linear dependence between them. Conceptually, each distribution is treated as a variable with values arranged according to the same pixel grid, which allows to establish a one-to-one mapping between values of both variables which makes calculating correlation easy (Wilming et al., 2011). In this thesis, I rely on Matlab implementation of 2D correlation provided in *corr2* function. Correlation can be used either to assess the predictive power of a saliency model (when calculated for smoothed human fixations and saliency map, as in Chapters One and Two) or to quantify the similarity between two sets of smoothed

human fixations (as in Chapter Three). Correlation has the advantage of being easy to compute and intuitive. Additionally, it is symmetric (commutative): a correlation between distributions A and B is no different from the correlation between B and A. This property makes correlation particularly suitable for comparing fixation maps, where the measure of their similarity is needed, not a 'unidirectional' information about how well one predicts the other.

To summarize, bottom-up approaches to oculomotor control inspired a large body of experimental work and a rich toolbox of computational models, some of which predict human gaze exceptionally well. Furthermore, the maturation of the field of saliency modelling resulted in the establishment of standard metrics for model comparison. The idea that the brain uses a map-like representation of the visual field to code the 'interestingness' of different locations has not been disproved until today. Currently, however, it is believed that such representation take into account not only the feature-based saliency but also other factors determining the importance of certain locations, such as the influence of a task (Bisley & Mirpour, 2019; Zelinsky & Bisley, 2015).

Top-down processing, task, and free viewing

Arguably, the most widely studied top-down factor, which can affect human eye movements is the necessity to perform a task, i.e., to sample visual information from a scene in order to accomplish a certain goal (Castelhano, Mack, & Henderson, 2009; Henderson, Shinkareva, Wang, Luke, & Olejarczyk, 2013; Yarbus, 1967). When the task is well-specified – for example, when it involves searching for a specific object in a scene – observers tend to restrict their fixation locations to scene regions which have a high probability of containing the target object (Pereira & Castelhano, 2019; Torralba et al., 2006). Importantly, the characteristics of oculomotor behaviour change as the function of task specification even when the viewed stimulus remains the same (Mills, Hollingworth, Van der Stigchel, Hoffman, & Dodd, 2011). This observation illustrates that humans can flexibly deploy their knowledge about the task and the environment – the factor which is intrinsic to them, and not to the visual input – to guide their gaze.

Apart from search tasks, many scenes-viewing studies involve less-constrained tasks, such as memorization (Foulsham & Underwood, 2008; Schütt et al., 2019; Tatler et al., 2005) or

aesthetic judgements (Nuthmann & Henderson, 2010). An extreme example of a loosely defined task is free viewing, in which observers are told to look at the image without any further instruction. Free viewing has been criticised for leaving too much freedom to observers, which might result in different observers adopting different viewing strategies which is effectively equivalent to performing different tasks (Tatler et al., 2005). Despite these concerns, free viewing has been routinely used in eye movements studies (Koehler, Guo, Zhang, & Eckstein, 2014), including the studies reported in this thesis. Moreover, the unconstrained nature of this approach is often also seen as an advantage: the basic notion is that, without instructions to perform a task, observers move their eyes because they are motivated by the intrinsic drive to seek information about in the environment (Baranes et al., 2015; J. Gottlieb, 2012; but see Benjamin W. Tatler et al., 2011). This idea therefore suggests that free viewing elicits spontaneous, naturalistic behaviour, comparable to the one exhibited in real-world situation in which someone looks at the world without any specific purpose other than to gain information.

While a task can significantly affect gaze control, it is not the only top-down factor capable of doing that. Numerous eye tracking studies demonstrated that humans exhibit many strong tendencies to look at certain kinds of image content, such as texts (Cerf et al., 2009; Wang & Pomplun, 2012), content related to social interactions, such as faces (Flechsenhar & Gamer, 2017; Rösler, End, & Gamer, 2017), the presence of animals (Drewes, Trommershäuser, & Gegenfurtner, 2011), or content eliciting strong emotions (Pilarczyk & Kuniecki, 2014). As alluded to in the previous sections, the fact that observers looked at these image regions and that this behaviour was not predicted by saliency models has been often treated as evidence against the bottom-up, saliency-based theories of oculomotor control, and in favour of the existence of top-down influences on gaze behaviour (for the criticism of this approach, explaining why strong conclusions might be premature here, see the 'Beyond a dichotomy' section of this Chapter).

In addition to the suggestion that specific stimulus classes have effects on eye movements, a number of studies suggested a more generalised effect of top-down processing, namely that eye movements might be guided by the representations of objects or by proto-objects (feature clusters indexed by peripheral vision as having a high chance for being an object; Rensink,

2000). Two kinds of experimental findings led to this suggestion. Firstly, maps indexing object locations were demonstrated to predict fixations better than sophisticated biologically-inspired saliency models (Stoll et al., 2015). Secondly, fixations usually land on the central regions of objects (Anderson & Donk, 2017; Foulsham & Kingstone, 2013; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013). This so-called preferred viewing-position effect contrasts with predictions of the traditional saliency-based approach: that saliency should be highest at the edges of objects, where the sharp changes in values of visual features occur. In Chapter Four, I discuss the relationship between object representations and visual features in a greater detail.

Scene Meaning

The picture, which emerges from these studies, is that human eye movements are predominantly control by the top-down factors. This conception is expressed in the cognitive relevance theory (Henderson, Malcolm, & Schandl, 2009), which posits that image features are largely irrelevant for oculomotor control, because it either operates at the level of semantic interpretations of the visual input or strongly relies on knowledge-based predictions about the possible semantic content of the scene (Henderson, 2017). This idea gave rise to meaning maps (Henderson & Hayes, 2017), a method of quantifying the distribution of semantic content in visual scenes. The proponents of this method claim that meaning – as measured by meaning maps – guides human gaze. In the Chapters One and Two, I report experimental findings, which serve as the basis for a critique of this claim. Here, I provide a more general overview of this topic and highlight several themes related to it, which I will revisit at various points throughout this thesis.

'Meaning' or 'semantics' are of interest to many disciplines, ranging from philosophy to cognitive linguistics. This ubiquity makes the attempt to provide a general and precise definition of meaning difficult. However, within cognitive psychology and neuroscience, there is an emerging consensus regarding how to understand meaning (Constantinescu, OReilly, & Behrens, 2016; Huth, De Heer, Griffiths, Theunissen, & Gallant, 2016; Kumaran, Summerfield, Hassabis, & Maguire, 2009; Mirman, Landrigan, & Britt, 2017; Sadeghi, McClelland, & Hoffman, 2015). According to this consensus, meaning of a certain concept (where the concept is understood as a noun or a verb) is its location in an abstract, multidimensional space occupied by other concepts, a so-called conceptual space. The fact that concepts share the same space

implies that is possible to measure the degree of relatedness or the distance between different concepts. Intuitively, in this space concepts such as 'cooking' and 'kitchen' are close to each other, but far from 'car' and 'driving', which, in turn are close together.

Only a subspace of conceptual space understood in that way is directly relevant to visual scenes. This subspace is determined by two inherent characteristics of scenes. Firstly, scenes can contain only those entities, which can be depicted visually. For example, concepts like 'failure' or 'destiny' do not meet this criterion but 'cat' does (Crutch & Warrington, 2005; Mkrtychian et al., 2019). Secondly, a single scene is not able to capture a sequence of steps involved in activities like, for example, 'cooking'. Of course, a scene still can depict a person who cooks but this is different from depicting 'cooking' itself. These two properties of scenes narrow down the space of concepts to the subspace of concrete nouns, which directly refer to objects. Therefore, in natural scenes, meaning is necessarily object-based, although it must be noted that i) objects in this context are understood very broadly, and scene elements typically belonging to the background – such as sky – are treated as objects too (Sadeghi et al., 2015) and ii) objects in the scene still remain in the relationships to the remaining parts of the space; they are not isolated in any sense.

The way in which the human mind organizes knowledge about objects, and the relationships between mental representations of different objects are the topics of active investigation (Cichy, Pantazis, & Oliva, 2014; Clarke & Tyler, 2014; Kriegeskorte, Mur, & Bandettini, 2008). The dominant approach in this research area involves testing how well different formal models of knowledge structure can account for experimental results obtained via behavioural or neuroimaging methods. These models are very diverse, and range from simple, hierarchical structures of categories (in which, for example, birds and mammals are sub-categories of animals, see a landmark study by Cichy et al., 2014) to complex multidimensional spaces based on data provided by individuals performing different behavioural tasks (Devereux, Tyler, Geertzen, & Randall, 2014; Hebart, Zheng, Pereira, & Baker, 2020; McRae, Cree, Seidenberg, & Mcnorgan, 2005). This general conceptualization of 'meaning' provides a good backdrop against which different use-cases (or ways of understanding) of that term, coming from the literature on natural-scene viewing, can be presented. In Chapter Two, I enumerate specific

examples of such use-cases derived from different studies and in the following sections focus on three more general ones.

Meaning related to object-context consistency

Objects can be semantically more or less consistent with the scene in which they are presented (Biederman, Mezzanotte, & Rabinowitz, 1982). A good example of a semantically inconsistent object is a shoe on a bathroom sink. Such objects are known to strongly attract human fixations (Mackworth & Morandi, 1967) and this effect is routinely cited as an example of semantic influence on eye-movement (Coco et al., 20202019; Võ et al., 2019; Williams & Castelhano, 2019). Object-scene inconsistency can be described in terms of the conceptual space under the assumption that objects, which are typical for a given context (for example, the context of a bathroom) are clustered together (Rose & Bex, 2020). Any object, which is located far from the cluster that defines the scene, is semantically inconsistent with the scene (although note that the relationship between typicality of an object for a given context and the location of this objects in the conceptual space is not necessarily a straightforward one). This understanding of meaning is further discussed in Chapters One and Two.

Meaning related to object individuation and recognition

The implicit assumption behind the notion of meaning described above is that objects in a scene can be individuated. In the study described in Chapter Four, I used two-tones images: stimuli which are meaningless when viewed for a first time but become meaningful after the observers acquire relevant prior object-knowledge. The initial meaninglessness of these images to observers can be understood as the inability to segment them into objects, which are linked to concepts in semantic space. Acquiring relevant object-knowledge, in turn, enables these processes. This dependence of scene meaningfulness on the knowledge possessed by observers highlights the fact that object individuation and recognition in this case cannot be achieved solely by processing image-features. Instead, these processes require the interaction between image features and object- knowledge. I elaborate on this point in Chapter Four.

Meaning as measured by meaning maps

A third way of understanding scene meaning present in this thesis is the one derived from the cognitive relevance theory (Henderson et al., 2009) and embodied in the meaning maps approach (Henderson, Hayes, Peacock, & Rehrig, 2019). It assumes that 'scene meaning' is a property, which is smoothly distributed over a scene and can be inferred simply by asking humans to rate the meaningfulness of image regions (for more details of this rating procedure see below). This perspective therefore takes a 'whatever works' approach, and makes no a prior assumptions about how meaning is constituted other than that human observers are able to rate the amount of meaning present in image parts. What comes at the expense of theoretical precision is the spatial nature of image meaning: meaning as measured by meaning maps is expressed in the format of a map, which is convenient from a practical perspective because it allows comparing meaning maps to saliency maps generated by saliency models (Henderson & Hayes, 2017, 2018). This conceptualisation of scene meaning advocated by the meaning maps approach is, however, hard to reconcile with the idea of conceptual space: concepts within this space cannot be ranked as more or less meaningful in and of themselves. Yet, the assumption of 'ranked' meaningfulness lies at the core of the meaning maps approach. This point is revisited in the final Discussion section.

Meaning maps are the core tool for measuring the kind of meaning described above. They are constructed from ratings of the meaningfulness of local image patches provided by many individuals. These ratings are combined into a smooth distribution over an image. Meaning maps come in two different versions, both of which have been evaluated for this thesis (see Chapters Two and Three for more details); the points made here are equally relevant for both of them. As alluded to above, meaning maps rest on the assumption that ratings provided by humans index cognitive processes related to the semantic analysis of the scene. While meaning maps have been used in numerous studies (reviewed in Henderson, 2020, and in Henderson et al., 2019), this central assumption (and the more general claim that meaning maps measure 'image meaning') has never been critically evaluated. Chapters Two and Three of the current thesis fill this gap in the literature by scrutinising these assumptions, and demonstrating the limitations of this method.

In sum, the top-down factors that influence human gaze control can be divided into two broad categories: one category of effects, which are task-related, and a second category of factors that exert their influence independently of any task (social signals, texts, object etc.). The top-down effects belonging to the second category have been demonstrated in studies aiming at demonstrating the limitations of saliency models (I discuss this issue in the next section). In other words, these studies were not guided by a general theory of top-down influences on oculomotor control and, to date, a theory that binds different experimental results together and is able to generate novel hypotheses is largely missing. The meaning maps approach had the ambition for being such theory. However, the experimental results, which I describe in Chapters Two and Three, seriously challenge its potential.

Beyond a dichotomy

The bottom-up vs. top-down dichotomy dominates the literature on natural-scene viewing and eye movements (Berga & Otazu, 2020; Henderson et al., 2019). It has been used as a conceptual framework for investigating various topics within that field, for example eye movements of patients with different impairments of visual processing (Charles Leek, Patterson, Paul, Rafal, & Cristino, 2012; Fellrath & Ptak, 2015; Ossandón et al., 2012), similarities of eye-movements between different species (Wilming et al., 2017), and developmental trajectories of changes in oculomotor behaviour (Franchak, Heeger, Hasson, & Adolph, 2016; van Renswoude, Raijmakers, & Visser, 2020). Despite its popularity, however, this dichotomy has a number of serious conceptual limitations and comes with methodological difficulties. Many studies that were inspired by this dichotomy attempted to quantify the unique influences of factors belonging to either of the two components by comparing the predictive power of different operationalisations of 'bottom-up' and 'top-down' factors in different conditions. The conclusions of these studies hold only to the extent to which the specific operationalisation on which they rest are a good proxy for the underlying construct (see Chapter Four for the elaboration of this point).

A more fundamental problem that I will revisit at various points throughout this thesis is the fact that the 'high-level' content of a scene usually supervenes on certain, very specific visual features. This issue complicates meaningful operationalisations that tease apart bottom-up and

top-down components. For example, locations of objects which people find interesting (and, thereby, 'worth' fixating) can be predicted with a reasonable accuracy with saliency models, which exclusively index local visual features and are therefore blind to information about objecthood (Elazary & Itti, 2008; Masciocchi, Mihalas, Parkhurst, & Niebur, 2009). In Chapters Two and Three I demonstrate that this limitation haunts even the meaning maps approach, which can be considered one of the more sophisticated operationalisations of semantic scene content. Specifically, instead of measuring image meaning, these maps might measure complex visual features and owe their ability to predict fixation locations to the fact that these features typically are the carriers meaning. In this sense, they are very similar to saliency models that are based on deep neural networks but lack their ease of use and automation.

Most studies that are inspired by the bottom-up vs. top-down dichotomy implicitly assume that it is possible to independently examine the respective influence of these components on eye movements. They therefore treat the correlation between 'high-level' factors and local visual features as a nuisance impeding their experimental separation. A broader perspective emerges once we consider a non-dichotomous approach (for example, see Borji & Tanner, 2016 or Nuthmann et al., 2020). The limitations of the dichotomy, as well as its over-simplistic nature, has been already pointed out by other authors. For example, Awh, Belopolsky, and Theeuwes (2012; see also Wolfe & Horowitz, 2017) describe variants of a visual search-task, in which this dichotomy is not able to account for all aspects of human performance. Further arguments for the need of a more nuance approach to discussions about top-down and bottom-up aspects of perception come from the literature on object perception (Driver, Davis, Russell, Turatto, & Freeman, 2001; Teufel & Fletcher, 2020). One of the themes, which emerge from this literature (reviewed in Chapter Four), is that image-computable features and mental object-representations formed by observers remain in a complex relationship (Nuthmann et al., 2020). Specifically, for the representation of objects to arise, an interaction between prior object-knowledge and early visual mechanisms must take place: a process I have labelled knowledge-driven perceptual organization. This object-as-interaction way of thinking has been successfully used in behavioural task (also with combination with neuroimaging) and allowed for gaining new insights into how the brain processes visual information (Chang, Baria, Flounders, & He, 2016; Flounders, González-García, Hardstone, & He, 2019; González-García, Flounders, Chang, Baria, & He, 2018; Gorlin et al., 2012; Teufel, Dakin, & Fletcher, 2018; Teufel et al., 2015). For

example, Teufel and colleagues (Teufel et al., 2018) demonstrated that forming an object representation facilitates perceptual processing of local image-features, on which this representation supervenes.

Given the important role of objects in eye-movements guidance (Nuthmann & Henderson, 2010; Nuthmann et al., 2020; Pajak & Nuthmann, 2013), it is therefore likely that knowledge-driven perceptual organization affects eye movements as well. In Chapter Four, I report experiments exploring this possibility and demonstrating that acquiring object-knowledge necessary for binding image features into objects indeed changes many aspects of human oculomotor behaviour.

Chapter Two – meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations

Introduction

As highlighted in the previous Chapter, a long-standing hypothesis suggests that semantic content of image regions is important in guiding eye movements. Recent work presented meaning maps (MMs) as a tool to test this hypothesis (Henderson & Hayes, 2017, 2018). This technique aims to index the spatial distribution of meaning across an image, which has potential applications far beyond eye-movement research. In this Chapter, I assess and challenge central assumptions of this novel tool.

A classic finding in eye-movement research shows that the specific task of an observer has an influence on where they direct their eyes (Yarbus, 1967; Hayhoe & Ballard, 2005). But in everyday life, we frequently move our eyes without any goal other than to explore the environment. In the lab, this behaviour is examined in free-viewing paradigms, during which eye movements are recorded while images are viewed without an explicit task (Koehler, Guo, Zhang, & Eckstein, 2014, but see Tatler, Hayhoe, Land, & Ballard, 2011). To explain what guides eye movements during free viewing, two opposing accounts have been put forward. Both are described in greater detail in the previous Chapter; here, they are presented only briefly.

According to the first account, eye movements are guided primarily by image characteristics (Borji, Sihite, & Itti, 2013; Itti & Koch, 2001; Parkhurst, Law, & Niebur, 2002). Potential support for this view comes from saliency models: algorithms, which exclusively use visual features of an image to predict human fixations. Although early models, which used only simple features such as local intensity or colours (Itti & Koch, 2000), are now deemed only moderately

successful (Bylinskii et al., 2014), more recent saliency models achieve a remarkably high performance (Kümmerer, Wallis, Gatys, & Bethge, 2017). These models harness deep convolutional neural networks – biologically inspired machine learning algorithms, that somewhat resemble the human visual system (Kietzmann et al., 2019). However, even such models rely solely on visual features, albeit high-level ones.

In contrast to the idea underlying saliency models, several authors have argued that during free viewing, eye movements are mainly guided by the semantic content of the visual scene (Henderson, Malcolm, & Schandl, 2009; Nyström & Holmqvist, 2008; Onat, Açik, Schumann, & König, 2014; Rider, Coutrot, Pellicano, Dakin, & Mareschal, 2018; Stoll, Thrun, Nuthmann, & Einhäuser, 2015). This perspective differs fundamentally from the saliency-based approach. Attributing meaning to certain parts of the scene is impossible without prior knowledge of the world, i.e., a factor that is independent of the visual input (Hegde & Kersten, 2010; Teufel, Dakin, & Fletcher, 2018). Consequently, the notion that semantic content guides eye-movements is inconsistent with the idea that the allocation of fixations is dependent solely on the distribution of image features. Given that meaning is not image-computable, the notion that semantic content guides eye-movements is inconsistent with the idea that the eye-movements are dependent solely on the distribution of image features.

A string of recent studies has focused on providing support for the role of meaning in driving eye movements (Hayes & Henderson, 2019; Henderson & Hayes, 2017, 2018; Henderson, Hayes, Rehrig, & Ferreira, 2018; Peacock, Hayes, & Henderson, 2018). These studies (reviewed in Henderson, Hayes, Peacock, & Rehrig, 2019) are based on a novel technique called meaning maps (MMs). A MM for a given image is created by breaking it down into small isolated patches, which are rated for their meaningfulness independently from the rest of the visual scene. These ratings are pooled together into a smooth map, which is supposed to capture the distribution of meaning across the image. Compared to outputs from a simple saliency model (GBVS, Harel et al., 2006), MMs were more predictive of human fixations. On that basis it has been claimed that meaning guides human fixations in natural scene viewing (Henderson & Hayes, 2017, 2018). In the present Chapter, I examined central predictions of this claim.

First, if MMs measure meaning and if meaning guides human eye-movements, MMs should be better in predicting locations of fixations than saliency models because these models rely solely on image features. Therefore, I compared MMs to a range of classic and state-of-the-art models. I replicate the finding that MMs perform better than some of the most basic saliency models. Contrary to the prediction, however, DeepGaze II (DGII; Kümmerer, Wallis, & Bethge, 2016; Kümmerer et al., 2017), a model based on a deep convolutional neural network, outperforms MMs.

A second prediction is that if MMs are sensitive to meaning and if meaning guides human gaze, differences in eye movements that result from changes in meaning should be reflected in equivalent differences in MMs. I probed this prediction experimentally using a well-established effect: the same object, when presented in an atypical context (e.g., a shoe on a bathroom sink) attracts more fixations than when presented in a typical context because of the change in the semantic object-context relationship (Henderson, Weeks, & Hollingworth, 1999; Öhlschläger & Võ, 2017). Replicating previous studies, image regions attracted more fixations when they contained context-inconsistent compared to context-consistent objects. Crucially, however, MMs of the modified scenes did not attribute more 'meaning' to these regions. DGII also failed to adjust its predictions accordingly.

Together, these findings suggest that semantic information contained in visual scenes is critical for the control of eye movements. However, this information is captured neither by MMs nor DGII. I suggest that similar to saliency models, MMs index the distribution of visual features rather than meaning.

Method

I conducted a single experiment in which human observers free-viewed natural scenes while their eye-movements were being recorded. The obtained data was analysed in two complimentary ways. First, I compared how well MMs and different saliency models predict locations of human fixations in natural scenes. Subsequently, I assessed the sensitivity of MMs and the best-performing saliency model to manipulations of scene meaning. The data, the code

to create MMs, and all openly available resources used in the study described in the present Chapter can be accessed via the links provided in the Supplement (see page 37).

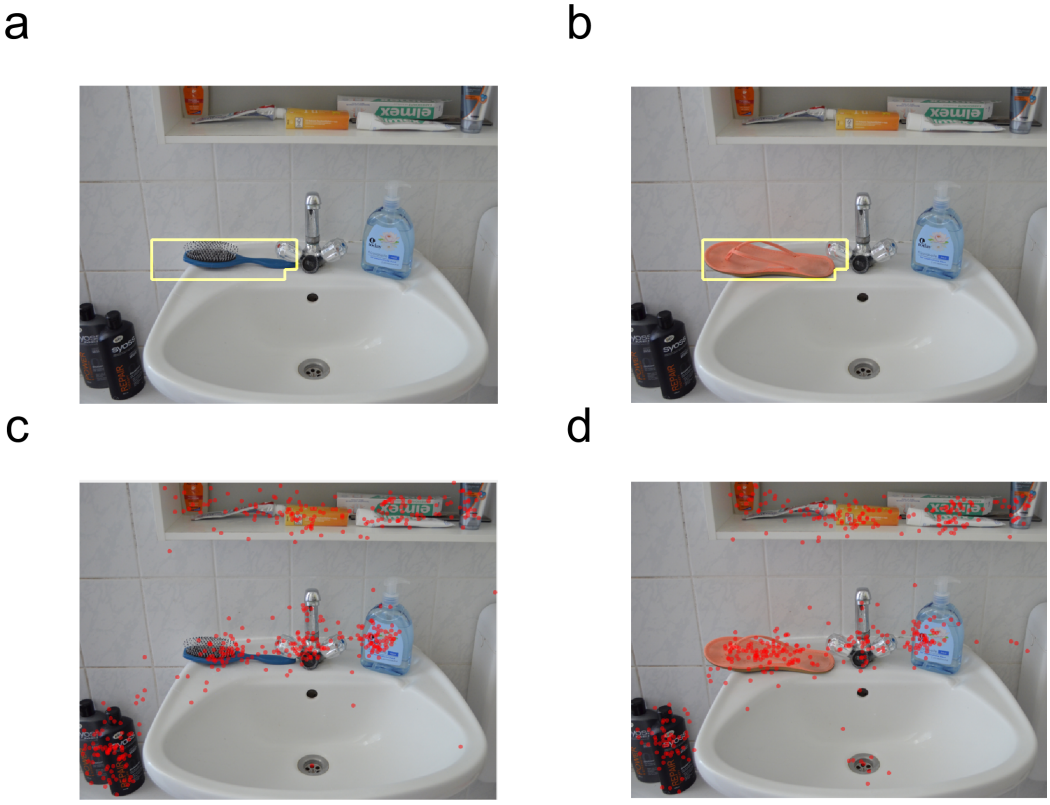

Fig. 1. Illustration of sample stimuli in (a) the Consistent and (b) the Inconsistent condition with the Critical Region outlined in yellow and (c, d) human fixations recorded in both conditions. In this example, a hair brush on a bathroom sink (a) – an object consistent with the scene context – has been exchanged for a shoe (b) to introduce semantic inconsistency.

Stimuli

I used images from two conditions of the SCEGRAM database (Öhlschläger & Võ, 2017): the Consistent and the Semantically Inconsistent conditions (called 'Inconsistent' here). In the Consistent condition (used in both analyses), scenes contain only objects that are typical for a given context. In the Inconsistent condition (used only in the second analysis), one of the objects is contextually inconsistent. For example, a hairbrush in the context of a bathroom sink from the Consistent condition is replaced with a flip-flop in the Inconsistent condition (see Figs. 1a and 1b). Such changes in object-context relationship alter the meaning attached to the manipulated object. For every scene, I indexed the location of the consistent and inconsistent objects with the superimposed bounding boxes for both objects (see Figs. 1a and 1b). I refer to this location as the Critical Region, because it is the only part of the image that changes

between Consistent and Inconsistent conditions. I used 36 selected scenes in both conditions (72 photographs in total, listed in the Supplement to the present Chapter together with the selection criteria). I also replicated the main finding of the first analysis in an additional set of 30, very different, images (reported in the Supplement).

Procedure

The procedure consisted of 3 blocks, interleaved with breaks. Each participant viewed all images from both conditions (Consistent and Inconsistent – no counterbalancing was applied) and was instructed to 'look carefully' at each of them. Experimental blocks began with an eye tracker calibration/validation. Within each block, observers free-viewed a series of 24 photographs from both SCEGRAM conditions, each for 7 seconds. After image offset, observers were required to press a button to view the next image. Then, a fixation point appeared centrally on a screen and once observers fixate on it (as determined online by their eye-trace), the actual image was displayed. Before starting the experiment, observers viewed a sample image in an identical regime to familiarize themselves with the procedure. Each stimulus was shown once and the order of presentation was fully randomized. The stimuli were presented against a uniform grey background and had a width of 688 pixels and a height of 524 pixels, which subtended approximately 19.7 and 15 degrees of visual angle, respectively. My choice of task (free viewing) and stimulus parameters for size (measured in degrees of visual angle – note that to to achieve this, the absolute size of the stimuli had to be modified which resulted in a slight change in their aspect ratio: from 1.33 to 1.31) and presentation time were adopted from the original study developing the SCEGRAM stimuli (Öhlschläger & Võ, 2017). These design characteristics fall within the typical range used in this literature (e.g. Wilming et al., 2017).

Observers

20 volunteers (3 male; mean age 19.4) recruited from the Cardiff University undergraduate population took part in the study. All reported normal or corrected-to-normal vision, provided written consent, and received course credits in return for participation. The study was approved by the Cardiff University School of Psychology Research Ethics Committee. The primary units of interest in my analyses were the distributions of fixations over images. The

number of observers I recruited guarantees that including more observers would not change these distributions significantly (demonstrated in the Supplement to the present Chapter).

Apparatus

The study was conducted in a dimly lit room. SCEGRAM images from both conditions were presented on an LCD monitor (Iiyama ProLite B2280HS, resolution 1920 by 1080 pixels, 21 inches diagonal). Chin and forehead rests were used to ensure that observers maintained the constant distance of 49 cm from the screen. Their eye movements were recorded with the frequency of 500 Hz using an EyeLink 1000+ eye tracker placed on a tower mount. The experiment was controlled by custom-written Matlab (R2017a version) scripts using Psychophysics Toolbox Version 3 (Kleiner, Brainard, & Pelli, 2007).

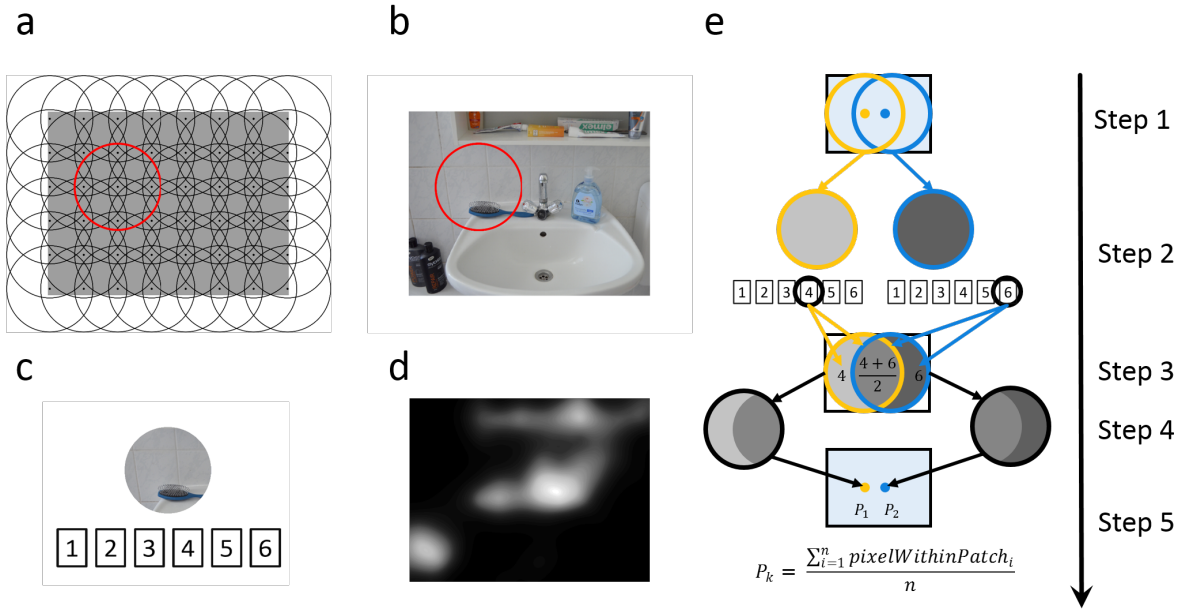

*Fig. 2. Illustration of the stimuli and procedure used for creating meaning maps. (**a**) Grids of equally spaced circles were used to cut images into fine and coarse patches (only the latter are illustrated here). The red circle indicates a sample patch in the grid. (**b**) Here, the sample patch is highlighted in one of the scenes from the Consistent condition. (**c**) Patches were presented in isolation and rated for their meaningfulness by three independent observers on a scale from 1 to 6. The panel has illustrative purpose only – the scale presented to observers included additional labels (ranging from 'Very Low' to 'Very High'). (**d**) Illustration of a meaning map with greyscale values indicating 'meaningfulness'. (**e**) Simplifying illustration of how meaning maps are*

generated from ratings. For simplicity sake, only two patches are shown (step 1). Each patch is rated in isolation (step 2; here only one rating per patch is shown). All pixels within an image area are then assigned average rating values, taking into account all ratings for patches that overlap with this area (step 3). For the area of the original patch (step 4), all pixels are then averaged and the resulting value is assigned to the centre of the patch (step 5). Finally, the patch centres were used as interpolation nodes for thin-plate spline interpolation producing a smooth distribution of values over the image (not illustrated). This procedure was conducted separately for the fine and coarse grid, and the meaning map for a given image was created by averaging the two outcomes and normalizing the result to a range between 0 and 1.

Creating MMs

To create MMs for my stimuli, I followed the procedure described by Henderson & Hayes (2017, 2018; for details see Fig. 2). Each image was segmented into partially overlapping patches of two sizes: fine patches had a diameter of 107 pixels (3 degrees of the visual angle, or 16 % of the image width), coarse patches of 247 pixels (7 degrees or 36% of the image width) (Fig. 2a and b). Their centres were 58 pixels (fine) and 97 pixels (coarse) apart from each other.

Next, I collected meaningfulness ratings from human subjects for all patches. Each patch was presented in isolation and rated for its meaningfulness on a 6 point Likert scale (Fig. 2). As in Henderson and Hayes (2017), I used a Qualtrics survey completed by naive observers recruited via the crowdsourcing platform Amazon Mechanical Turk (see Supplement for eligibility criteria). Each participant provided ratings for 305 or 303 patches of both sizes (selected randomly from all images), on average spent approximately 14 min on the task, and received 2.18 USD as remuneration. In total, 69 individuals were used as raters, with three individuals rating each individual patch. The collected ratings were then used to create MMs (see Fig. 2).

When creating MMs for images from both conditions, I exploited the fact that photographs from the Consistent and Inconsistent conditions differ only in the Critical Region (the part of the image containing the manipulated object) while the remaining parts overlap. I collected meaningfulness ratings for the patches belonging to overlapping areas only once, and the separate sets of ratings for Consistent and Inconsistent condition were collected only for those

patches that contained at least one pixel belonging to the Critical Region. In total, the number of patches rated in the study amounted to 7013: 4840 fine patches (of which 520 belonged to the images from the Inconsistent condition) and 2173 coarse patches (445 Inconsistent).

Saliency models

In the first analysis, I compared predictive performance of MMs to four saliency models of different complexity. The first two models – GBVS (Harel et al., 2006) and AWS (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012) – rely on simple visual features, such as local colors and edge orientations, and share the assumption that fixations land on image regions distinct from their surroundings in terms of values of these features. By contrast to GBVS, AWS includes a statistical whitening procedure to improve performance. Both these models were previously used to estimate the influence of image features relative to cognitive factors on the deployment of fixations: GBVS in the previous studies with MMs, AWS elsewhere (Stoll et al., 2015).

Two other models that I compared to MMs – Intensity Contrast Features (ICF) and DeepGaze II (DGII) – were designed in a data-driven manner (Kümmerer et al., 2017). Both have the same architecture, consisting of a fixed network that extracts sets of features from images and a readout network that is trained on human fixations to combine the features in a way to maximize the models' predictive power. While the fixed network of ICF extracts only simple visual features (local intensity and contrast), DGII is tuned to features extracted by a deep convolutional neural network pre-trained for object recognition (VGG; Simonyan & Zisserman, 2015). The key characteristic of these models that distinguishes them from models such as GBVS and AWS is that they have been trained on human fixations. Specifically, during the training phase, the read-out network receives its respective features as an input, generates a prediction about where human observers will look in the image, and gradually adjusts its parameters based on feedback comparing its prediction to human fixation data to maximise the predictive power of each model.

The predictions of all four models were obtained by running Matlab (for GBVS and AWS) or python (for DGII and ICF) scripts provided by models' authors. These scripts did not require

providing any parameter values and their predictions – as all having a form of smooth maps that predict the probability of image regions to be fixated – could be used in my analyses directly.

Human observers have the tendency to look at the centre of images (Tatler, 2007), and therefore this probability is usually higher in the central region of the image. This 'centre bias' has important consequences for the evaluation of saliency models. Their performance differs depending on whether they are evaluated using a metric expecting some form of this bias or not (Kümmerer, Wallis, & Bethge, 2018). Here, for the sake of simplicity, I do not incorporate centre bias in the models or in the MMs (unlike the original authors) and use an appropriate metric for this situation (see Performance metrics section).

Data pre-processing

Fixation locations from the eye tracker recordings were extracted using the algorithm provided by the device manufacturer operating with the default parameter values. This algorithm analyses the stream of incoming data about eye position and segments it into events (saccades, fixations, and blinks) in real time. Its core is the mechanism for detecting saccade onsets and offsets which relies on thresholds for velocity, acceleration, and motion (eye displacement) applied to subsequent data samples. These thresholds have the values of 30 degrees of visual angle per second (deg/s), 8000 deg/sec^2, and 0.15 deg, respectively. Using this algorithm, I obtained a discrete distribution of fixations on each image (see Fig. 1c and 1d). Then, in line with the previous MMs studies, I smoothed these discrete distributions with a Gaussian filter with a cut-off frequency of -6 dB, using the function provided by Bylinskii and colleagues (2014).

Next, smooth distributions from fixations, models, and MMs were separately normalized to a range from 0 to 1 for each image. Finally, for each scene, histograms of all distributions from both conditions were matched to histograms of smoothed fixations from Consistent condition using the Matlab *imhistmatch* function, as in the original MMs studies. Histogram matching makes distributions directly comparable as it ensures that they differ only with respect to their shape, and not their total mass.

Performance metrics

To compare the ability of MMs and models to predict locations of human fixations in Experiment 1, I use two well-established metrics (Bylinskii, Judd, Oliva, Torralba, & Durand, 2016): Correlation and Shuffled Area Under ROC curve (sAUC; Zhang et al., 2008) with the implementations provided by Bylinskii and colleagues (2014).

Correlation, used in the previous studies on MMs, is calculated as Pearson's linear correlation coefficient between a smoothed distribution of observers' fixations over the image and predictions of a saliency model or MMs. I additionally used sAUC (Zhang et al., 2008), which, unlike Correlation, guarantees that the measured differences in performance between models are driven by their sensitivity to factors guiding fixations, and not by the degree to which they include human centre bias in their predictions, even implicitly (Kümmerer, Wallis, & Bethge, 2015; Kümmerer et al., 2018).

Comparing meaning maps and saliency models – results

In the first analysis, I compared performance of four saliency models to MMs in predicting human fixations in the Consistent condition, i.e., when viewing typical scenes with no obvious object-context inconsistencies (Tab. 1, Fig. 3). If human gaze is guided by meaning, and if MMs provide an index for the distribution of meaning, I would expect MMs to outperform all saliency models because these models are based solely on image features. Please note that for the sake of this comparison, I aggregated fixations from all observers for each image and analyzed the data on a per-image basis.

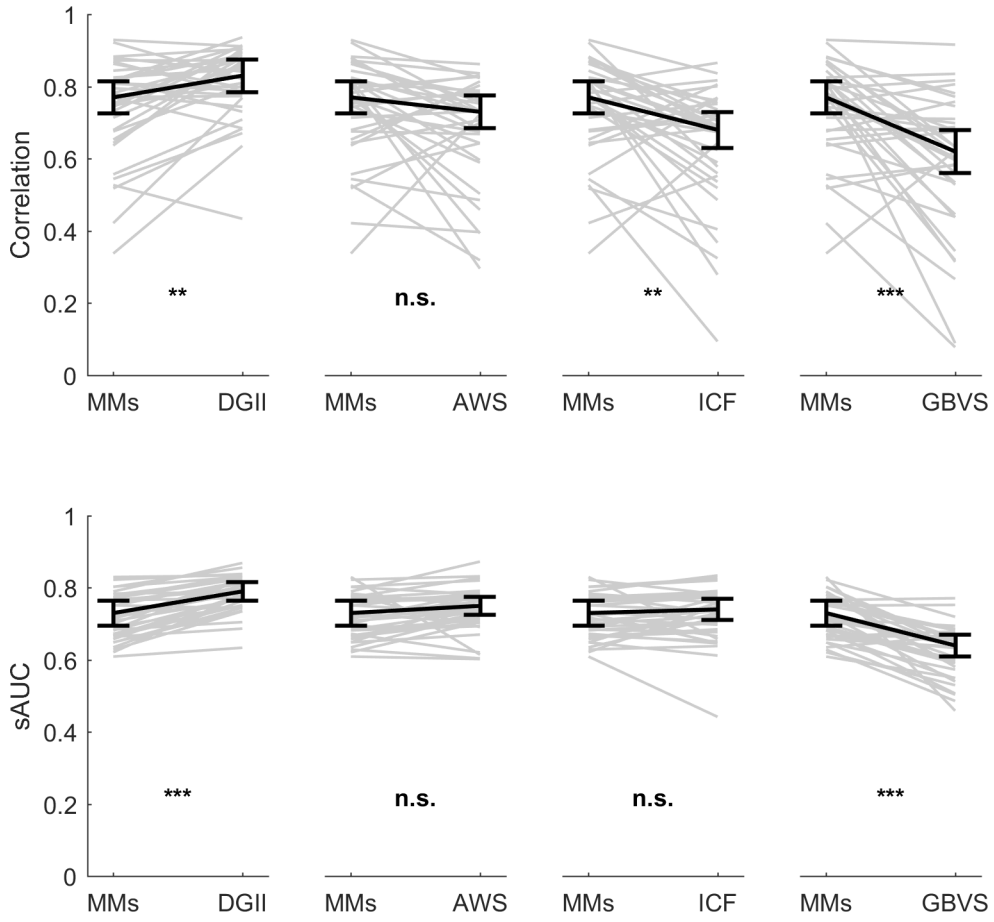

*Fig. 3. Performance of MMs and saliency models in predicting human fixations according to (a) Correlation and (b) sAUC metrics. Note that according to both metrics DGII predicted human fixations better than MMs. Asterisks indicate p-values from statistical tests comparing MMs to different models (reported in Table 1.): * indicates p ≤ .05, ** p ≤ .01, *** ≤ .001 and 'n.s.' indicates the lack of statistical significance. Grey lines connect values obtained for individual images. Black vertical bars indicate 95% confidence intervals for the medians.*

Predictive power

Correlation and sAUC values obtained for MMs and for each of the models were compared using Bonferroni-corrected paired Wilcoxon tests (Fig. 3; Tab. 1). I used non-parametric tests because for some of the distributions a visual inspection of Q-Q plots indicated that the assumptions of normality might be not met. For the same reason I chose a median as a measure of centrality (I calculate confidence intervals for median using a bootstrapping method – see details in the Supplement). Additionally, I calculated Jeffreys–Zellner–Siow (JZS) Bayes Factor (Rouder et al., 2009) to quantify the evidence for (or against) the differences

between models and MMs (Tab. 1). While deviations from normality can be problematic for Bayes factor analyses, they are most likely not an issue in the current situation: the Bayes factors for the key finding are large and the deviations from normality are small. When interpreting BF values, I adopted a convention that only BFs greater than 3 (and, consequently, smaller than 1/3) are interpreted as informative. This convention, although subjective (which is inevitable when using BFs), is frequently adopted in the literature (Jarosz & Wiley, 2014).

As shown in Tab. 1 and on Fig. 3, according to both measures, MMs outperformed GBVS in predicting human fixations, thereby replicating the results of Henderson and Hayes (2017, 2018) using new images and new participants. Contrary to expectations, however, both metrics indicated that DGII predicted fixations better than MMs. Furthermore, performance of AWS and MMs did not differ significantly irrespective of the metrics. Finally, MMs outperformed ICF according to Correlation, but not sAUC. In fact, for the latter metric, JZS-Bayes Factor indicated support for the null hypothesis. When interpreting sAUC scores, be mindful that values of 0.5 indicate chance performance (see Chapter One for details).

Table 1. Comparison of Predictive Power of Saliency Models and MMs Using Correlation and sAUC.

Model	Median of prediction values with 95% confidence intervals	Median of differences from MMs with 95% confidence intervals	W statistic	p-value (Bonferroni-corrected)	JZS Bayes Factor
Correlation					
DGII	0.83 [0.78, 0.87]	0.07 [0.03, 0.11]	526	0.00738	32.26
MMs	0.77 [0.72, 0.81]	–	–	–	–
AWS	0.73 [0.67, 0.76]	-0.06 [-0.12, -0.01]	192	0.10412	1.48
ICF	0.68 [0.61,	-0.12 [-0.18, -0.06]	144	0.00936	16.90

	0.71]					
GBVS	0.62	[0.56,	-0.11 [-0.26, -0.05]	94	< .001	396.96
	0.68]					
sAUC						
DGII	0.79	[0.77,	0.06 [0.05, 0.08]	662	< .001	> 1000
	0.82]					
MMs	0.73	[0.69,	–	–	–	–
	0.76]					
AWS	0.75	[0.72,	0.02 [0.01, 0.04]	490	0.0507	0.60
	0.77]					
ICF	0.74	[0.70,	0.01 [-0.01, 0.02]	383	1.00	0.19
	0.76]					
GBVS	0.64	[0.60,	-0.10 [-0.12, -0.08]	13	< .001	> 1000
	0.66]					

Semi-partial correlations

Because predictions of models and MMs overlap, I quantified their distinct predictive power using semi-partial correlations. I conducted these analyses for GBVS (used in the original MMs studies) and DGII (the only model which markedly outperformed MMs). For each scene from the Consistent condition, I calculated two semi-partial correlations with the distribution from smoothed fixations: one for MMs while controlling for GBVS, and one for GBVS while controlling for MMs (see Fig. 4). Consistent with findings by Henderson and Hayes (2018), MMs explain more unique variance than GBVS (Fig. 6a), as indicated by the significantly higher coefficients in the former than the latter case (mean difference 0.28, 95% confidence interval (CI) [0.17, 0.39]; paired t-test, $t(35) = 5.22$, $p < 0.001$). Interestingly, the identical analysis with DGII revealed that DGII explained significantly more unique variance than MMs (mean difference 0.15, 95% CI [0.07, 0.24]; $t(35) = 3.60$, $p < 0.001$, see also Fig. 4b).

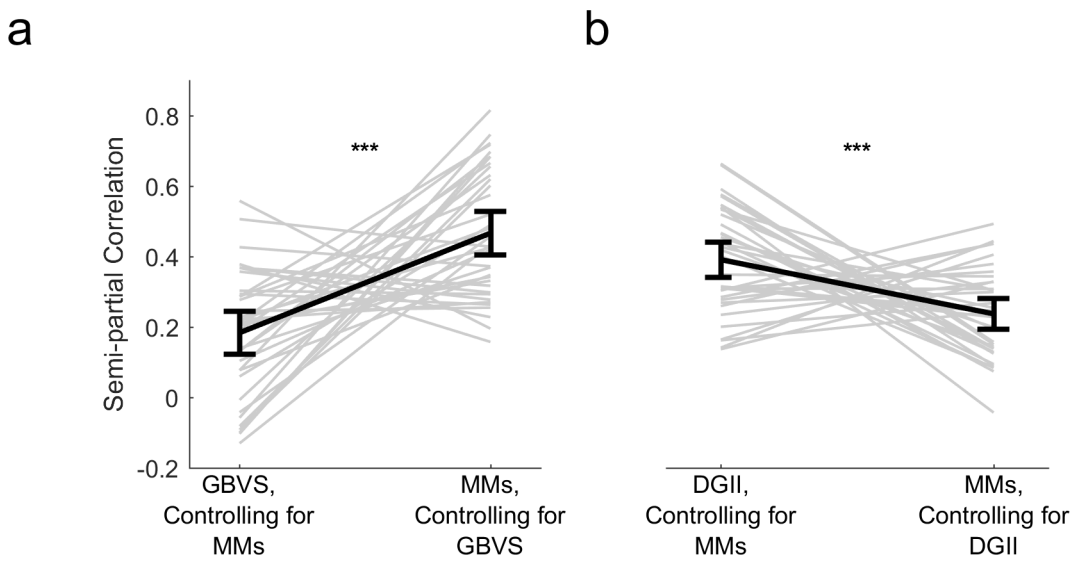

Fig. 4. Comparison of semi-partial correlations with smoothed human fixations for (a) MMs and GBVS and for (b) MMs and DGII. The obtained coefficients were significantly higher when assessing MMs while controlling for GBVS compared to when assessing GBVS when controlling for MMs. The opposite was true for the analyses with DGII. All figure characteristics are as in Fig. 3. except that means instead of medians are presented.

Internal replication

To demonstrate the generalizability of my conclusions beyond SCEGRAM images, I replicated the main results with a different stimulus set (see the Supplement).

Comparing meaning maps and saliency models – discussion

If human gaze is guided by meaning, and if MMs index the distribution of meaning across an image, MMs should outperform saliency models that are exclusively based on image features. My first analysis showed that this prediction does not hold. In fact, DGII generated better predictions and explained more unique variance than MMs. Therefore, at least one of the two premises of my prediction does not hold: either human eye-movements are not sensitive to meaning or MM do not index meaning. The second analysis allowed us to distinguish between these alternatives.

Analysing the effects of semantic inconsistencies within scenes – method

In the second analysis, I assessed how human observers, DGII, and MMs respond to experimental changes in meaning induced by altered object-context relationships. I used eye-movement data from both the Consistent and the Inconsistent condition. These conditions differed solely in the Critical Region, an area that either contained an object that was either consistent with the scene context or induce semantic conflict. For each scene, I calculated the mass of the distributions of human gaze, DGII, and MMs falling into the Critical Region, respectively, and divided it by the Region's area for normalization. My primary interest was the comparison between conditions: to the extent to which humans, DGII, and MMs are sensitive to meaning, they should fixate more (humans) or predict more fixations (DGII and MMs) on the Critical Region in the Inconsistent than the Consistent condition.

Analysing the effects of semantic inconsistencies within scenes – results

My comparison indicated that, as predicted, observers fixated more on inconsistent than consistent objects (Fig. 5a). By contrast, behavior of both MMs and DGII did not change across conditions (Fig. 5b and c). These impressions were confirmed by a 2x3 ANOVA, with condition (Consistent vs. Inconsistent) as a within-subjects factor and the distribution source (human fixations vs. MMs vs. DGII) as a between-subjects factor. I found a statistically significant main effect of distribution source, $F(2, 105) = 13.09$, $p < 0.001$, $\omega^2 = 0.16$ and condition, $F(1, 105) = 7.41$ $p = 0.0076$, $\omega^2 = 0.005$. These main effects were qualified by a significant interaction, $F(2, 105) = 16.90$, $p < 0.001$, $\omega^2 = 0.026$. Tukey post-hoc tests showed that human observers looked more at the Critical Regions in the Inconsistent, than the Consistent condition, $t(105) = -6.22$, $p < 0.001$. In contrast, no significant differences between conditions were found for DGII, $t(105) = -0.09$ $p = 1.0$, and MMs, $t(105) = 1.60$ $p = 0.6028$. Comparisons within conditions indicated that human fixations differed from MMs in the Inconsistent condition, $t(129.91) = 5.78$ $p < 0.001$, but not the Consistent condition, $t(129.91) = 2.16$ $p = 0.2662$. A significant difference between DGII and

human fixations was detected in both Consistent, t(129.91) = -2.96 p = 0.0420, and Inconsistent conditions, t(129.91) = -5.79 p < 0.001.

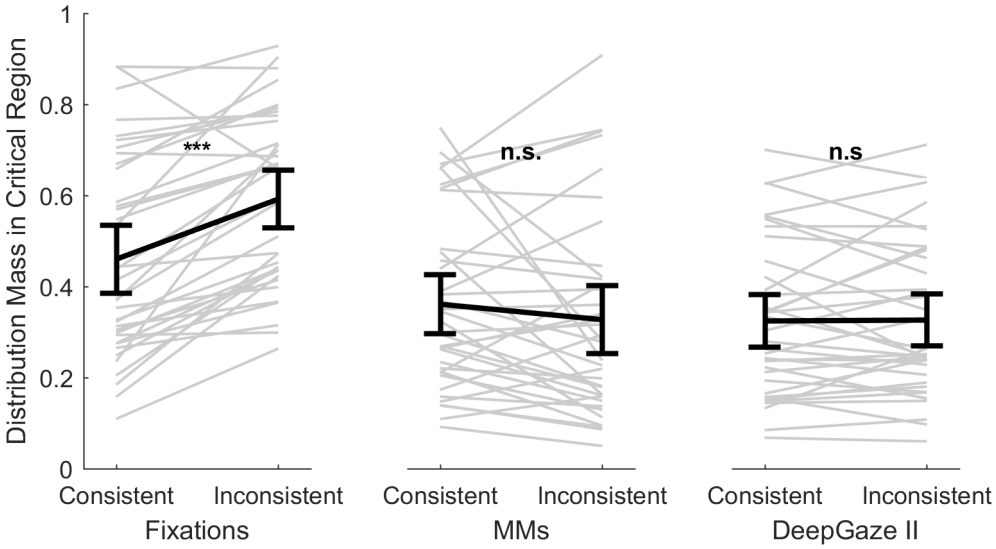

Fig. 5. Normalized distribution mass falling within Critical Regions in both conditions for (a) smoothed human fixations, (b) MMs, and (c) DGII. All figure characteristics are as in Fig. 3.

Additionally, conditions differed regarding the number of fixations per image, t(35) = 5.67 p < 0.001. On average, there were 6% fewer fixations in the Inconsistent condition. This excludes the possibility that higher number of fixations in this condition might drive the observed increase in the distribution mass falling within the Critical Regions.

Any systematic differences in object size between Consistent and Inconsistent conditions also could affect my results because larger objects may attract more fixations solely because they occupy a larger image area. However, this factor was minimized by showing each object in a consistent and an inconsistent context. Yet, the same object might be shown in a slightly different position in the two conditions and might therefore occupy slightly different amounts of the image. This was, however, not the case: the JZS Bayes Factor of 4.26 indicated that the two conditions did not differ in the size of the bounding boxes of each manipulated object (objects in the Inconsistent condition were on average 1562.28 pixels larger; 95% confidence interval: [-2582.74, 5707.29]).

Next, please note that I employed a within-subject design, which might have led to carry-over effects: observer viewing a given scene in the Inconsistent condition first could be biased to look at the Critical Region in the Consistent condition when they viewed the same scene for a second time. Note that even if this unwanted phenomenon occurred despite a randomised order of stimuli presentation, it could only decrease the magnitude of the effects of interest.

Finally, it is possible that my observers implicitly engaged in a task. Specifically, once the observers realized that the stimuli contain object-context inconsistencies, they might have started actively searching for them. Engaging in this semantic oddball-search task would result in very different spatial distributions of fixations compared to the ones, which would be obtained during free-viewing. This prediction was not supported by my findings: I replicated my main experiment in a different set of observers with images that did not contain semantic inconsistencies, and found that DGII still predicted fixation locations better than MMs. This separate data set, therefore, suggests that observers did not engage in an oddball search task and that the superiority of DGII is not specific to SCEGRAM images only (details to be found in the Supplement).

To summarize, semantic changes induced by altering object-context relationships elicited changes in distributions of human fixations, but neither MMs nor DGII could predict them. These results suggest that both models might be sensitive to image features, which are frequently correlated with image meaning, rather than to meaning itself.

Discussion

A long-standing debate in visual perception concerns the extent to which visual features vs. semantic content guide human eye-movements in free viewing of natural scenes. To distinguish these hypotheses, indexing the distributions both of features and meaning across an image is critical. While image-based saliency models have been used to index features for two decades, measuring semantic importance has been difficult until meaning maps (MMs) have recently been proposed. Here, I assessed the extent to which MMs indeed capture the distribution of meaning across an image. First, I demonstrate that despite the claims about the importance of

meaning as measured by MMs for gaze control, MMs are not better predictors of locations of human fixations than at least some saliency models, which are based solely on image features. In fact, DeepGaze II (DGII), a model using deep neural network features, outperformed MMs. Second, I assessed the sensitivity of human eye-movements, MMs, and DGII to changes in image meaning induced by violations of typical object-context relationships. Observers fixated more often on regions containing objects inconsistent with scene context (thus replicating previous findings) but these regions were not indexed as more meaningful by MMs, or as more salient by DGII. Together, these findings challenge central assumptions of MMs, suggesting that they are insensitive to the semantic information contained in the stimulus.

The good performance of DGII in predicting human gaze might be attributable to the high-level features it extracts from images. Three other models, which use low-level features, failed to decisively outperform MMs. However, unlike two of them (GBVS and AWS), DGII is trained with data on human fixations to optimize performance (Kümmerer et al., 2016, 2017). Yet, training alone cannot explain the difference in performance. The third low-level feature model (ICF) is trained in the same way (Kümmerer et al., 2017) but still achieves a lower performance than DGII. These findings suggest that feature type is indeed critical for a model's performance. Importantly, however, while DGII uses high-level features transferred from a deep neural network trained on object recognition (Simonyan & Zisserman, 2014), this is not equivalent to indexing meaning. Rather, the good performance of DGII is likely due to meaning supervening on, or correlating with, some of the features indexed by this model.

Correlation between visual features and meaning as the source of good performance in saliency models has already been considered by the authors of MMs (Henderson & Hayes, 2017). My findings suggest that MMs might share this characteristic with saliency models. Specifically, the ratings used to construct MMs might be based on visual properties in such a way that highly structured patches that contain high-level features receive high ratings. These features often correlate with meaning, but in and of themselves do not amount to meaning. According to this interpretation, both DGII and MMs index high-level features. Their success in predicting human behaviour derives from the typically strong correlation between high-level features and meaning, with a higher correlation for the features extracted by DGII than MMs.

An alternative interpretation of the finding that DGII outperforms MMs is that image features rather than meaning guide human fixations. However, this interpretation is inconsistent with my second analysis. Here, observers clearly exhibited sensitivity to meaning, as indicated by changes in gaze-patterns elicited by introducing semantic inconsistencies into the images. This experimental manipulation targets a type of meaning that is based on how objects relate to the broader context in which they occur. While specific, it is precisely this kind of meaning that is of high theoretical importance in eye-movement research (Henderson, 2017; Henderson et al., 2009). Natural scenes are by definition composed of multiple objects, and the physical and semantic relationships between these objects as well as their relationship to the scene gist, determine the meaning of a scene (Kaiser et al., 2019; Malcolm et al., 2016; Võ, 2021). Thus, the fact that MMs are not sensitive to the meaning derived from object-context relationships seriously limits their usefulness.

It is, however, possible that – as has been already suggested (Henderson et al., 2018) – MMs capture some form of 'local' meaning that is important for oculomotor control. Evaluating my results in this respect is complicated by the correlation between features and meaning (Elazary & Itti, 2008), which I already alluded to above. Yet, at the very least, the fact that MMs do not consistently outperform even simple saliency models such as AWS that by design rely on low-level image features warrants caution. This finding indicates that either the purported kind of meaning indexed by MMs is not of primary importance for guidance of eye-movements, or that it is almost perfectly correlated with the features indexed by the models.

A similar issue relates to DGII: while my study shows that this model does not index meaning derived from object-context relationships, one might argue that it acquires sensitivity to some (local) form of meaning by virtue of being trained on human data. Specifically, if eye-movements are guided by the semantic content of images, then training on eye-movement data might lead to developing 'meaning-sensitivity' in the model. While this scenario cannot be ruled out for the same reasons as in the case of MMs, recall that the ICF model – which uses simpler features than DGII – is also trained on human data but fails to reach the high performance of DGII. Therefore, if the high performance of DGII is based on some form of 'local' meaning, then it is not training per se that leads to the development of this meaning but an interaction of training and specific features.

These considerations indicate the urgent need for developing a more nuanced conceptual approach and terminology to capture the intricacies of different types of 'meaning', and a more appropriate language to talk about the relationship between 'features' and 'meaning'. Without a clearer theoretical framework, it will be difficult to experimentally settle debates regarding the role of 'meaning' in natural-scenes perception.

In any case, the insensitivity to semantic inconsistencies reveals inherent limitations of both MMs and DGII. The way in which MMs are constructed implicitly assumes that meaning is a local image-property, which is not true for object-context (in)consistency. This limitation may potentially be alleviated by 'contextualized MMs' (Peacock, Hayes, & Henderson, 2019), a recently suggested modification of the 'standard' MMs. These novel maps are created from meaningfulness ratings by observers who see the whole scenes from which the to-be-rated patches were derived (the next Chapter of this thesis is dedicated to assessing them). DGII, in turn, does not explicitly encode semantic information, and was not trained on the relationship between eye movements and semantic (in)consistency. But its failure highlights an opportunity to improve saliency models by incorporating semantic relationships (Bayat et al., 2018).

Taken together, my results suggest that, contrary to their core promise as a methodology, meaning maps (MMs) do not offer a way to measure the spatial distribution of meaning across an image. Instead of meaning per-se, they seem to index high-level features that have the potential to carry meaning in typical natural scenes. They share this characteristic with state-of-the-art saliency models, which are easier to use, do not require human annotation, and yet predict locations of human fixations better than MMs.

Supplement to Chapter Two

Internal replication

I replicated the finding that DGII outperforms MMs in predicting human fixations in a separate experiment with a different set of stimuli and new observers. This second experiment was

identical to the first one except for the stimuli, the number of observers (21 instead of 20), and the number and the duration of images presented.

Methods

I used 30 photographs from Corel Photo Library depicting mainly single animals. The images were converted to grayscale and resized to 788 by 526 pixels (22.5 by 15 degrees of visual angle). I presented them for 3 seconds each, in blocks of 10. 21 new observers took part in the study (1 male, mean age 19.05). To create MMs, the images were fragmented into 4200 fine and 1620 coarse patches in total. Each rater recruited via Amazon Mechanical Turk rated 291 patches. Because DGII requires RGB images as input, to generate model's predictions I converted the stimuli to RGB by copying greyscale pixel values to the three color channels.

Results

DGII again outperformed MMs in predicting human fixations (see Table S1).

Table S1. Comparison of Predictive Power of Saliency Models and MMs Using Correlation and sAUC – Internal Replication

Model	Median of prediction values with 95% CIs	Median of differences from MMs with 95% CIs	W statistic	p-value (Bonferroni -corrected)	JZS Bayes Factor
Correlation					
DGII	0.86 [0.71, 0.89]	0.11 [0.04, 0.17]	386	0.0042	6.34
MMs	0.74 [0.62, 0.80]	–	–	–	–
sAUC					
DGII	0.75 [0.73, 0.78]	0.04 [0.02, 0.06]	409	< .001	63.52
MMs	0.70 [0.68, 0.74]	–	–	–	–

Number of observers

I tested if fixations from 20 observers who viewed images in my experiment are sufficient to closely approximate the theoretical ground truth distributions of fixations which would have been obtained from an infinite number of observers. Visual inspection of the data revealed that including fixations from about 10 observers results in distributions which remain virtually unchanged when fixations from more observers are added. This observation was confirmed by a more formal analysis. For each observer, I randomly selected a subset of 12 other observers 10 times and calculated how well smoothed fixations from these subsets predict fixations of the observer for each image using correlation. Averaging all the obtained values over observers and subsets resulted in an estimate of how well 12 observers predict fixations of a single observer viewing a given image. The value obtained for all SCEGRAM images amounted to 0.835 (SD = 0.094). Next, again for each observer viewing each image, I calculated how well their fixations can be predicted using fixations of the remaining 19 observers. The average of the obtained values equaled to 0.840 (SD = 0.097). This number is close to the value obtained for 12 observers, thus indicating that the underlying distributions of fixations are similar, and that increasing the number of observers would not affect them substantially.

SCEGRAM scenes

I used SCEGRAM photographs from two conditions: Consistent and (Semantically) Inconsistent. These photographs are divided into quadruples: object A in scene A (scene and object are consistent), object A in scene B (scene and objects are inconsistent), object B in scene B, and object B in scene A. In order to avoid potential distortions of the results, in my experiment I included only these quadruples in which both manipulated objects do not contain any digits or text observers might be trying to read. Applying this selection criterion resulted in retaining the following SCEGRAM scenes: 1-4, 7, 8, 11, 12, 17, 18, 21, 22, 27, 28, 33-44, 47-50, 55, 56, 59-62.

Eligibility criteria for Amazon Mechanical Turk raters

Raters who rated the meaningfulness of image patches had to meet the following requirements: they had to have an approval rating grater than 96%, be located in the United

States, have more than 500 tasks ('HITs') approved, and not have completed the task for images from a given set of images (SCEGRAM or from the replication) before.

Statistical software

All statistical analyses were conducted in R (R Core Team, 2016) with the help of functions from the following packages: BayesFactor (Morey & Rouder, 2015), jmv (The jamovi project, 2019), ppcor (Kim, 2015), and boot (Canty & Ripley, 2019; Davison & Hinkley, 1997).

Confidence intervals for medians

All confidence intervals for medians were calculated using adjusted bootstrap percentile (BCa) method with 10000 bootstrap replicates, implemented in R package boot (Canty & Ripley, 2019; Davison & Hinkley, 1997).

Openly available materials

Deep Gaze II and ICF: https://deepgaze.bethgelab.org/

AWS: http://persoal.citius.usc.es/xose.vidal/research/aws/AWSmodel.html

GBVS: http://www.vision.caltech.edu/~harel/share/gbvs.php

SCEGRAM database: https://www.scenegrammarlab.com/research/scegram-database/

Code for creating MMs used in this paper: DOI: 10.5281/zenodo.3490592

Data from this study: DOI: 10.5281/zenodo.3490434

Chapter Three – human eye-movements are not guided by meaning as measured by contextualised meaning maps

Introduction

In the previous Chapter, I demonstrated that even when observers view images without a task, the spatial allocation of fixations can be by guided by factors which are not captured by saliency models, namely, semantic content of the visual scene (see Henderson et al., 2019; Wu et al., 2014 for reviews). It seems fair to say that, to date, there is no conceptual framework to capture precisely what authors refer to when they talk about the semantic content, or meaning in images. With respect to oculomotor control, these terms are used to label factors such as identifiability of the depicted objects (Luke & Henderson, 2016; Pilarczyk & Kuniecki, 2014; see also Williams & Castelhano, 2019), specific properties of objects (Xu, Jiang, Wang, Kankanhalli, & Zhao, 2014), image parts best conveying the information of the whole image (Nyström & Holmqvist, 2008), or social signals such as the presence of human faces (Rider, Coutrot, Pellicano, Dakin, & Mareschal, 2018).

There is, however, one well-studied effect in eye movement research, for which the notion of semantic content, or meaning is more clearly defined. Specifically, objects can be more or less semantically consistent with the scene, within which they appear, and, as shown in the previous Chapter, the extent of object-scene consistency has an effect on eye movement behaviour. For instance, in one of the seminal studies (Loftus & Mackworth, 1978), one example stimulus showed a farmyard scene either with a (semantically consistent) tractor, or a (semantically inconsistent) octopus. Inconsistent objects such as the octopus were looked at earlier, attracted more fixations, and were inspected for longer in comparison to consistent objects. While mixed results have since been found with respect to the timing of eye movements (or, more precisely, with respect to the existence of semantic processing outside foveal vision that might be indicated by the earlier fixations to inconsistent objects; Wu, Wick,

et al., 2014), there is robust evidence demonstrating that object-scene inconsistencies lead to more and longer fixations (Coco, Nuthmann, & Dimigen, 2020; Friedman, 1979; Henderson, Weeks, Phillip A., & Hollingworth, 1999).

Two types of mechanisms are thought to underpin these effects: first, scene context influences object processing (Bar, 2004; Kaiser, Quek, Cichy, & Peelen, 2019) and objects that are viewed in inconsistent contexts are processed less effectively (Biederman, Mezzanotte, & Rabinowitz, 1982; Munneke, Brentari, & Peelen, 2013). Consequently, more fixations towards, and longer inspection times of inconsistent objects are thought to reflect the increased processing resources needed to process these stimuli (Bonitz & Gordon, 2008; De Graef, Christiaens, & D'Ydewalle, 1990).

A second, and possibly complementary, explanation for the effects of object-scene inconsistencies on eye movements is based on the notion that inconsistent objects are 'more informative' (Loftus & Mackworth, 1978), 'semantically informative' (Henderson, 2011; Henderson et al., 1999), or 'contain greater meaning' (Peacock, Hayes, & Henderson, 2019, page 6). According to this idea, the oculomotor system drives fixations towards inconsistent objects in an effort to maximise extraction of 'meaning' from a scene.

This second interpretation has recently gained increased attention, in particular with the development of meaning maps, a method to quantify the spatial distribution of 'meaning' across an image described in the previous Chapter (Henderson & Hayes, 2017, 2018). To reiterate, meaning maps are created by first partitioning an image into many circular, partially-overlapping patches. These patches are then presented to individuals (called raters) who view them without knowing the scene from which they were extracted, and are asked to rate their meaningfulness on a Likert scale. Finally, these ratings are combined into a smooth distribution over the image to create a map. 'Meaning' indexed by this method has been demonstrated to be a better predictor of fixations than a simple saliency model, a finding that has been interpreted as evidence to suggest that semantic information rather than image-computable features control eye movements (Henderson & Hayes, 2017, 2018). The meaning maps approach is rapidly gaining in popularity and these maps have been used to study eye movements in contexts such as virtual reality (Haskins, Mentch, Botch, & Robertson, 2020) or

mind-wandering (Krasich, Huffman, Faber, & Brockmole, 2020; Zhang, Anderson, & Miller, 2020).

The previous Chapter tested key assumptions underpinning MMs, and compared them to a wider range of saliency models. This comparison has highlighted a number of limitations (see also Pedziwiatr et al., 2021). One of these issues relates to semantic object-scene inconsistencies: contrary to the idea that inconsistent objects attract fixations because they are richer in meaning, meaning maps in their originally proposed form do not ascribe more meaning to scene regions occupied by objects that are inconsistent with the global scene context compared to consistent objects presented in the same region and matched in terms of low-level features. Moreover, meaning maps were outperformed in the prediction of fixation locations by Deep Gaze II, a saliency model based on a deep neural network (Kümmerer, Wallis, & Bethge, 2016; Kümmerer, Wallis, Gatys, & Bethge, 2017). Together, the results of this study led to the conclusion that MMs do not index semantic information *per se*, but high-level visual features that are highly correlated with semantics. In this respect, the original form of MMs are similar to modern saliency models.

It seems plausible to assume that many of the limitations of the original approach stem from the fact that MMs ignore the global context of the scene – recall that they are created from ratings of isolated, 'context-free' image patches. To overcome these issues, contextualised meaning maps have recently been proposed (Peacock et al., 2019). They differ from their predecessors in one important detail: during rating, each patch is presented alongside the full scene from which it originated, so raters have access to global scene-context when assessing the meaningfulness of the patch. Given the critical importance of context in scene semantics (Biederman et al., 1982; Võ, Boettcher, & Draschkow, 2019), the contextualised meaning maps that are created from such context-sensitive ratings might be better suited to quantify semantic information within the visual scene.

In the present Chapter, I applied the same straightforward but critical test to contextualised meaning maps, to which the original meaning maps were subjected: I assessed the extent to which they are sensitive to semantic object-scene inconsistencies, assigning higher 'meaning' to inconsistent objects, and – consequently – predicting increased fixations on such objects. To

that end, I created contextualised meaning maps for two types of indoor scenes: each scene either contained a semantically consistent object such as a hair brush on a bathroom sink, or this object was replaced with an inconsistent object such as a shoe on the sink. Analysing these maps and comparing them to fixation-patterns of human observers viewing the corresponding scenes revealed that the maps are not able to predict the gaze changes elicited by the manipulation of semantic object-context consistency. Moreover, this experiment provided initial evidence that contextualised meaning maps might attribute *less* meaning to scene regions that contain inconsistent compared to consistent objects. Given this surprising result, in a second experiment, I asked a large number of raters to provide meaningfulness ratings for a carefully controlled set of image patches. The results of this second experiment suggest that objects that are semantically inconsistent with the global scene are judged as less meaningful than consistent objects. This finding is diametrically opposed not only to the predictions of the meaning maps approach but also – when taken at face value – to the more general assumption that semantically inconsistent objects are 'more meaningful'. In addition to these main findings, my sample of 122 individuals provided the means to reveal a substantial between-individuals variability in meaningfulness ratings. Overall, my results presented in the Chapter challenge the meaning maps approach but point towards new directions for research on individual differences in scene perception.

Experiment 1 – Methods

The main goal of Experiment 1 was to compare contextualised meaning maps and human eye-movements. In particular, I was interested in the extent to which contextualised meaning maps and fixations respond to local changes in semantic information within a scene, resulting from the presence of objects that are semantically consistent vs. inconsistent with the overall scene-context.

Stimuli

I used the same stimuli as in the previous Chapter: photographs of 36 indoor scenes, taken from the openly available dataset SCEGRAM (Öhlschläger & Võ, 2017). To remind, each scene

was photographed in two conditions: Consistent and (semantically) Inconsistent which resulted in two images per scene (72 images in total). Images from the Consistent condition contained only objects typically found in certain contexts. In the Inconsistent condition, one of these objects was replaced with an object unusual in the context provided by the whole scene, thus introducing a semantic inconsistency. For example, in one of the scenes, a hair brush on a bathroom sink (Consistent condition) was replaced with a flip-flop (Inconsistent condition) – see Fig. 1b. The SCEGRAM dataset is construct in such a way that, across scenes, consistent and inconsistent objects are matched for low-level properties (Öhlschläger & Võ, 2017). In each scene, consistent and inconsistent objects occupy the same image locations, and the superposition of the bounding boxes of both conditions constituted what I here call a Critical Region. These Critical Regions are important for the data analyses I report further below because they contain the only image regions that differ between conditions. Please refer to the previous Chapter for more details.

Observers

I used the same eye-tracking data as in Chapter Two – no new observers were recruited for the present Experiment.

Eye-movement data

In this Chapter I used the same eye-movement data as in the previous one. Therefore, here, I report only the key characteristics of its collection and pre-processing, asthey are described in detail in the previous Chapter. For all 72 images, I collected eye-tracking data from 20 observers. Each observer free-viewed the full set of images displayed in a random order while their eyes were tracked with an EyeLink 1000+ eye-tracker. The images had a width of 688 pixels and a height of 524, corresponding to, respectively, 19.7 and 15 degrees of a visual angle. Each image was presented for 7 seconds, which is similar to the presentation duration of 8 s used in the original contextualised meaning maps study (Peacock et al., 2019). To analyse the eye-movements data, fixation locations were extracted from raw eye-tracker recordings using a standard EyeLink algorithm. The discrete fixations on each image were transformed into continuous distributions by means of Gaussian smoothing (filter cut-off frequency: -6 dB;

implemented in Matlab – see Kümmerer et al., 2020) followed by a normalization to the [0-1] range.

Creating contextualised meaning maps – overview

The procedure of creating contextualised meaning maps is almost exactly the same as that used to generate the original meaning maps. I closely followed the procedure described in detail in previous publications (Henderson & Hayes, 2017, 2018; Pedziwiatr et al., 2021). In summary, a pre-defined grid is used to segment the image into circular, partially overlapping patches (Fig. 1 A). Next, in a crowdsourced online experiment, each patch is presented next to the image from which it was derived, and human raters are asked to rate the meaningfulness of the patch. Presenting the full image next to the patch ensures that the rater knows the context when providing their responses (Fig. 1B; see this figure for details of the rating procedure itself). The presence of context is the only element differentiating contextualised meaning maps from their predecessors, meaning maps. Each individual patch was rated by three individuals. They were instructed in the same way as in the original contextualised meaning maps study: to rate how "meaningful" they think scene patch is. In a third step, the ratings from individual patches are combined into a smooth distribution over the image by means of averaging and interpolation (Fig. 1C). For each image, these three steps are conducted twice: once for bigger 'coarse' patches and once for smaller 'fine' patches. Then, the maps resulting from coarse and fine patches are averaged. Finally, the regions of the average map, which are close to the edges are down-weighted (Fig. 1D; see *Creating contextualised meaning maps – modelling centre-bias* section for details). This manipulation accounts for the centre-bias of human eye-movements, i.e., the tendency to look more at the central region of an image (Tatler, 2007). Note that in the study described in the previous Chapter, the centre bias model was not included in the meaning maps. I did include it here because I was primarily interested in replicating the steps from the original contextualised meaning maps article (where the bias was included), rather than in comparing the maps to saliency models regarding their predictive power, as I did in the previous Chapter (to remind, this comparison might be affected by the presence of the centre-bias model in the maps).

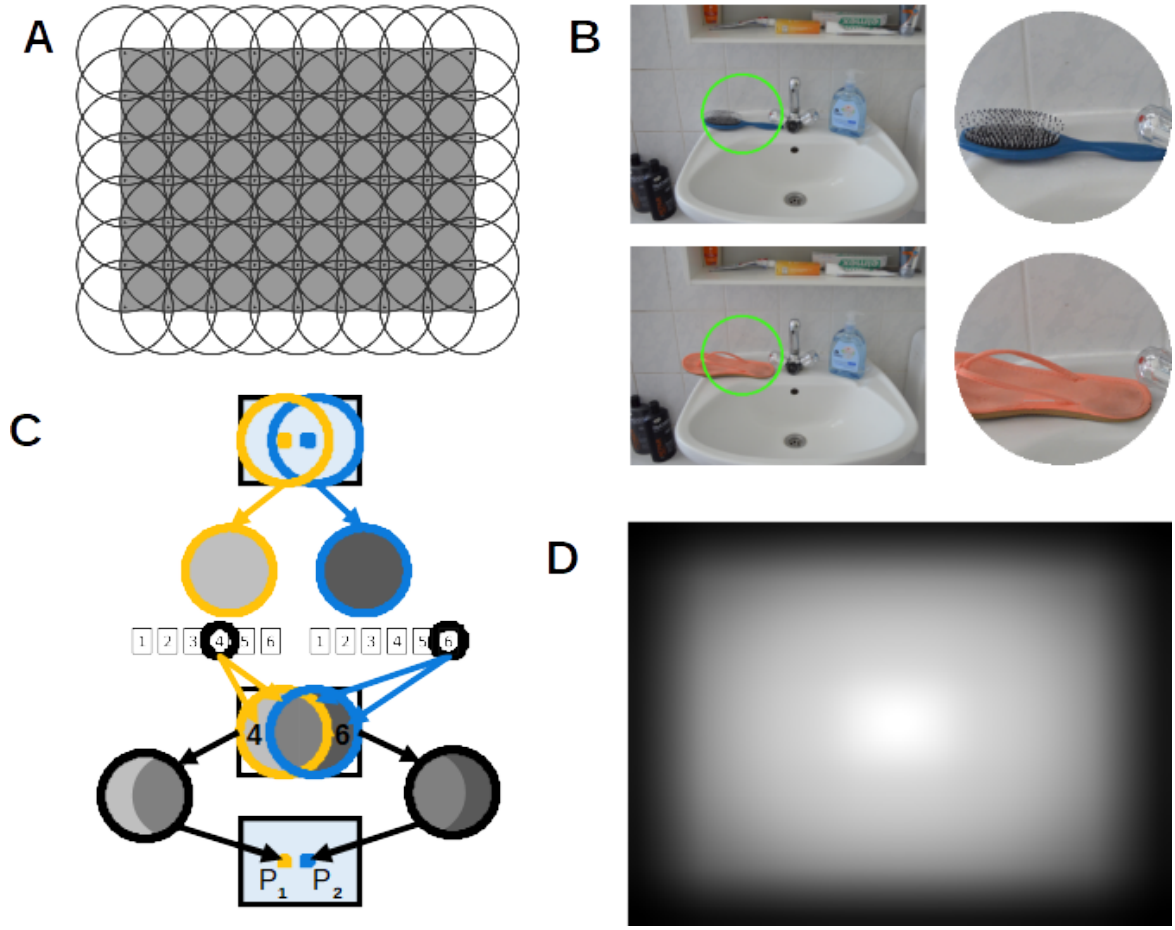

Fig. 1. Generating contextualised meaning maps.

A) Grid used to segment images into coarse patches. Grey rectangle represents image area. B) *Sample stimuli from the patch-rating task used for creating contextualised meaning maps. The patch, which raters were asked to rate for its meaningfulness, was always presented next to the image from which it originated to provide the relevant context. A green circle on the context image indicated the location of the patch. Both panels show the same scene in the Consistent (upper part of the panel) and the Inconsistent (lower part) condition. The images on both panels differ only with respect to the object shown in the patch. The hair brush on the upper part is a semantically consistent object for a bathroom scene, the shoe on the lower parts is semantically inconsistent. In the task, raters were asked to assess the meaningfulness of the patches by means of selecting a value on a six-point rating scale. C) Simplified schematic illustration of combining patch ratings into contextualised meaning maps (to be read from the top to the bottom). For the sake of simplicity, the illustration includes only two patches of a single size, and each patch received only one rating. Rating values assigned to different image regions in the subsequent steps of the procedure are reflected by the intensity of grey colour. Points* P_1 *and* P_2 *serve as nodes*

for thin plate spline interpolation (not shown), which is the final stage of ratings processing. For a more detailed description of the procedure, please refer to the previous Chapter. D) Centre bias model used in contextualised meaning maps. To account for the human inclination to allocate fixation predominantly to central image-regions (a so-called centre bias), creating in contextualised meaning maps includes assigning different weights to different pixels of the maps depending on their location. This re-weighting is done by convolving the maps with a model of centre bias shown on this panel, in which brighter pixels indicate higher pixel-weights.

Creating contextualised meaning maps – parameter value selection

When creating contextualised meaning maps for my stimuli, the aim was to match as closely as possible the procedure used in the original study by Peacock and colleagues (2019). My images, however, differed in size from the stimuli used in that study and were viewed from a different distance during the eye-movements data collection. In order to account for these differences, I matched the two studies with respect to the size of coarse and fine patches in degrees of visual angle (deg), and with respect to patch density of coarse and fine patches expressed in the number of patches per square degree of visual angle (p/deg^2). Under the constraint that the centres of each two adjacent patches have to be equidistant horizontally and vertically, these four values fully specify the grids necessary for creating contextualised meaning maps. In terms of absolute values, matching the two studies with respect to these parameters was perfect for patch diameter and resulted in 5.26 deg (coarse patches) and 2.26 deg (fine patches), which corresponded to 187 pixels and 79 pixels, respectively. The patch densities closest to the original I could possibly achieve were 0.56 p/deg^2 and 0.21 p/deg^2 (compared to 0.57 p/deg^2 and 0.2 p/deg^2 in the original study). Given the size of my stimuli, these values correspond to 63 coarse and 165 fine patches per image. The resulting grid for creating coarse patches is shown on Fig. 1a.

Creating contextualised meaning maps – data collection

The procedure described in the previous sections resulted in a total of 16 416 patches (4 536 coarse and 11 880 fine patches). As described in detail in the caption for Fig. 1, each patch was rated for its meaningfulness by three human raters on a 6 point Likert-scale. Patches were divided into 54 blocks of 304 patches each, and each block was assigned to three different raters (see details below).

Recall that each scene exists in a Consistent and an Inconsistent version, differing only with respect to the identity of a single object. If the raters were to view the same scene in both conditions, there would be a high chance that they guess the main focus of the study and, in turn, adjust their rating strategy (by, for example, conditioning the rating values assigned to patches on the presence – or absence – of the semantic inconsistency in the context image). Such situation would invalidate my results because the raters would be performing a task different than intended. To prevent that from happening, I assigned patches to blocks in such a way that each rater never saw a scene in both the Consistent and Inconsistent conditions. Specifically, I created two sets of blocks. First one contained half of the patches from the Consistent condition and half from the Inconsistent, with the patches in both these halves derived from different scenes. The other set of blocks contained the remaining patches. Because of this division, raters were never exposed to the same scene in both conditions. Within each set of blocks, patches were allocated to blocks randomly.

Each block was rated by three unique raters, and 162 raters were recruited in total. The order of patch-presentation was randomised for each rater separately. Data collection was conducted online. The raters were recruited using the crowdsourcing platform Prolific (www.prolific.co) and the patch-rating task was implemented as a Qualtrics survey (Qualtrics, Provo, UT). All my raters had to meet the following eligibility criteria: they had to be of U.S. nationality (as in the original contextualised meaning maps study), they had to have submitted at least 100 tasks to Prolific before, had to have an approval rate of 95% or more, and had to use a laptop or a personal computer to complete the task. They were financially reimbursed for their time and were allowed to participate in my study only once.

Creating contextualised meaning maps – modelling centre-bias

Recall that the final step of creating contextualised meaning maps involves reweighting the map with a model of centre bias. Such models have the form of smooths distributions over the image, with higher values clustering closer to image centre (Clarke & Tatler, 2014). When creating contextualised meaning maps I followed the original authors and relied on a model provided with the saliency model GBVS (Harel, Koch, & Perona, 2007) (to be precise, on the inverse of centre-bias model included in *invCenterBias.mat* file, which I inverted-back by

subtracting it from one). This model is shown on Fig. 1d. The effects of applying it are illustrated on Fig. 2d and e.

Creating contextualised meaning maps – histogram matching

For each image, I matched the histogram of its contextualised meaning map to the histogram of the distribution obtained by smoothing human-fixations registered on this image. This was done using *imhistmatch* Matlab function. Histogram matching – also used in the original meaning maps studies – ensures that values from both distributions are directly comparable because they have been aligned to the same scale (see Fig. 2b, c, d).

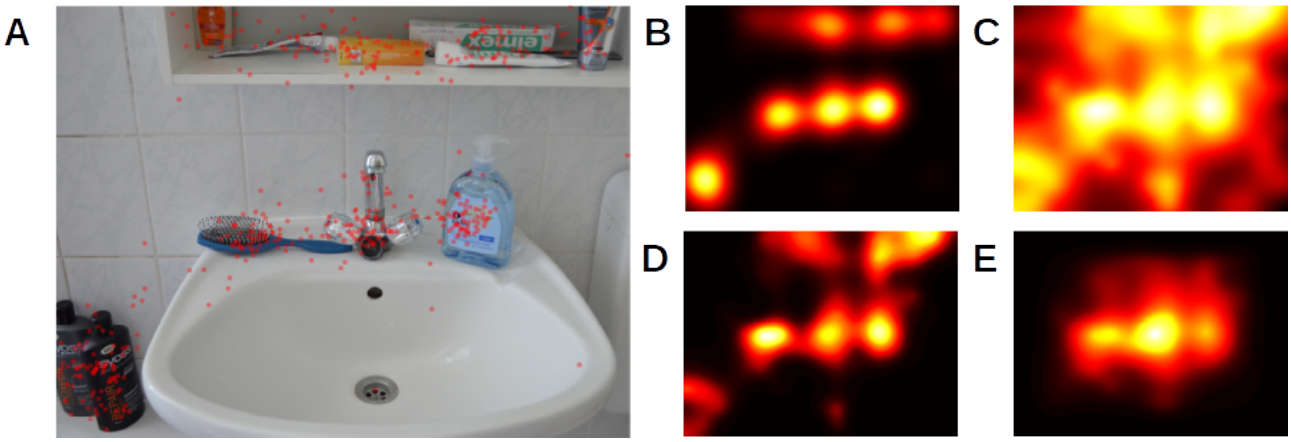

Fig. 2 Contextualised meaning maps – illustration.

A) Singles scene from the Consistent condition of my study, with fixations registered on it marked with red dots. B) Smoothed fixations from panel A). The histogram of this distribution served as a reference to which the histogram of the contextualised meaning map was matched (see next panel). This procedure ensures the comparability of values from both distributions by aligning these values to the same scale. C) 'Raw' map for the scene from panel A). Since this map has not been subjects to histogram matching, colour-values on it are not comparable to values on the remaining panels. D) The map from panel C), after histogram-matching but before including centre bias. Contextualised meaning maps were better predictors of fixations before including centre bias in them than afterwards (see Soundness check: general predictive power of contextualised meaning maps *section). E) The map from panel D), with centre bias model included. Such maps are used in all my analyses, unless otherwise stated.*

Data analysis software

Data from this study was handled using Matlab R2020a (Mathworks Inc., Natick, MA) and R (R Core Team, 2020). In particular, I relied on the R packages belonging to the tidyverse collection (Wickham et al., 2019); for pre-processing and plots generation, as well on packages jmv (The jamovi project, 2020; for running ANOVAS). Other R packages I used are cited in the relevant places in the text.

Data and code availability

The eye movement data used in this study are openly accessible via the following link: https://zenodo.org/record/3490434). SCEGRAM stimuli are available under the following link: https://www.scenegrammarlab.com/research/scegram-database/.

Experiment 1 – Results

Soundness check: general predictive power of contextualised meaning maps

I tested how well the patterns of human fixations on images could be predicted by the contextualised meaning maps I created. To quantify their predictive power, I applied a standard technique (Bylinskii, Judd, Oliva, Torralba, & Durand, 2016), used also by Peacock and colleagues (2019): for each image, I calculated the correlation between its contextualised meaning map and smoothed fixations registered on this image. For images from the Consistent condition, the average correlation amounted to 0.60 (SD = 0.17). The average percent of the explained variance in the eye-movement data amounted to 39%. In the Inconsistent condition, contextualized meaning maps performed slightly worse (M = 0.57, SD = 0.20, 37% of the variance explained). Additionally, I investigated the effects of removing centre bias from contextualised meaning maps and, interestingly, found that contextualised meaning maps performed better without it (Consistent: M = 0.71, SD = 0.13, 52% of the variance explained; Inconsistent: M = 0.66, SD = 0.17, 47% of the variance explained). This might be related to the fact that in SCEGRAM scenes, the content fixated by observers was distributed more uniformly than the specific model of a centre bias included in contextualised meaning maps assumes (see example in Fig. 2).

Either way, all these results are similar to what is reported in the original study (where the maps explained 40% of the variance in human data) and thus provide an important soundness check for my study. A lower quality of predictions in my study than in the original contextualised meaning maps study (Peacock et al., 2019) could have indicated that either the procedure of creating contextualised meaning maps is sensitive to aspects of the design which were different between my study and the original study (such as absolute image size), or that there were some technical problems with my implementation of it.

Sensitivity of contextualised meaning maps and eye movements to semantic manipulations

In the first key part of my analysis, I compared contextualised meaning maps and smoothed human-fixations with respect to their sensitivity to semantic manipulations. For this analysis, I focused on Critical Regions – image regions which, depending on the condition, contained a semantically consistent or inconsistent objects (see *Stimuli* section for details). For each scene, I first performed histogram matching (see previous section) and then calculated the mass of each distribution (contextualised meaning maps and smoothed fixations) falling within the Critical Region and divided that value by the Region's area for normalisation (see Fig. 3). These values were then analysed using a mixed 2×2 ANOVA with the condition (Consistent vs. Inconsistent) as a within-subjects factor and the distribution source (contextualised meaning maps vs. smoothed fixations) as a between-subjects factor. Please note that here a 'subject' indicates a single scene. Such an approach is typical for studies comparing fixation-prediction methods and is grounded in the observation that different observers agree to a large extent in their selection of fixation targets in images (De Haas, Iakovidis, Schwarzkopf, & Gegenfurtner, 2019). This analysis revealed that both the distribution sources and conditions, differed from each other statistically (distribution source: $F(1, 70) = 23.05$, $p < 0.001$, $\omega^2 = 0.22$; condition: $F(1, 70) = 5.34$, $p = 0.0238$, $\omega^2 \approx 0$). Importantly, however, these main effects were qualified by an interaction ($F(1, 70) = 23.83$, $p = p < 0.001$, $\omega^2 = 0.02$). Tukey post-hoc test showed that human eye-movements were sensitive to the change in semantic relationship between object and scene, as indicated by the fact that more mass of the smoothed-fixations distribution fell within the Critical regions in the Inconsistent condition compared to the Consistent condition

(Consistent – Inconsistent: M = -0.09, SE = 0.02, p < 0.001). The same comparison, however, did not yield statistically significant differences for the contextualised meaning maps (M = 0.03, SE = 0.02, p = 0.2737), suggesting that the amount of 'meaning' they assigned to the Critical Region did not differ between conditions.

Recall that one step in creating contextualised meaning maps involved averaging the two maps derived from the coarse patches and the fine patches. I repeated my mixed ANOVA analysis separately for the coarse and fine maps. The pattern of results of the ANOVA for both fine and coarse patches was similar to that reported in the previous section (fine patches: distribution source: $F(1, 70) = 32.64$, $p < 0.001$, $\omega^2 = 0.26$, condition: $F(1, 70) = 0.08$, $p = 0.7769$, interaction: $F(1, 70) = 31.56$, $p < 0.001$, $\omega^2 = 0.04$; coarse patches: distribution source: $F(1, 70) = 41.85$, $p < 0.001$, $\omega^2 = 0.3$; condition: $F(1, 70) = 3.71$, $p = 0.0581$; interaction: $F(1, 70) = 5.87$, $p = 0.018$, $\omega^2 = 0.01$). Importantly, however, I obtained an unexpected outcome in the post-hoc tests for the fine patches maps: this analysis revealed that fine maps attributed *less* meaning to Critical Regions in the Inconsistent condition than the Consistent condition (M = 0.08, SE = 0.02, p = 0.0019). Coarse maps did not exhibit this puzzling tendency; the pattern of results for them was the same as in the first analysis (post-hoc for contextualised meaning maps: M = 0.01, SE = 0.03, p = 0.9851).

Additionally, as a side note to my main considerations, I examined the temporal evolution of the influences of semantic inconsistencies on eye-movements. Other studies, also comparing fixations on consistent and inconsistent objects which occupied the same image location, yielded conflicting findings regarding whether the inconsistent objects are fixated earlier or not (see Wu, Wang, et al., 2014 for review). To help to clarify this issue, I conducted an additional analysis of my eye-movements data, in which I compared the ordinal numbers of first fixations landing within the Critical Regions between out experimental condition. This comparison revealed that it took observers 5.03 fixations on average to look at the inconsistent objects for a first time, and 5.97 for consistent (data pooled over scenes and over observers). The finding that the inconsistent objects are not fixated immediately after image onset but still earlier than consistent replicates the results of a recent study by Coco, Nuthmann and Dimigen (2019). These authors supplemented gaze recordings with electroencephalography (EEG) and concluded that object semantics can be at least partially accessed via peripheral vision.

To summarize, human eye-movements changed in response to local changes in semantic information: inconsistent objects attracted more fixations than the consistent ones, and were fixated earlier. The analogous effect was not detected for contextualised meaning maps and for their coarse component: I did not find the differences between conditions for neither of the two. For fine component of contextualised meaning maps, introducing semantic inconsistency to a scene region elicited a change in the amount of 'meaning' ascribed to this region. Intriguingly, however, it had the opposite direction. That is, contrary to predictions of the meaning maps approach, the fine maps ascribed less meaning to scene regions when they contained inconsistent objects.

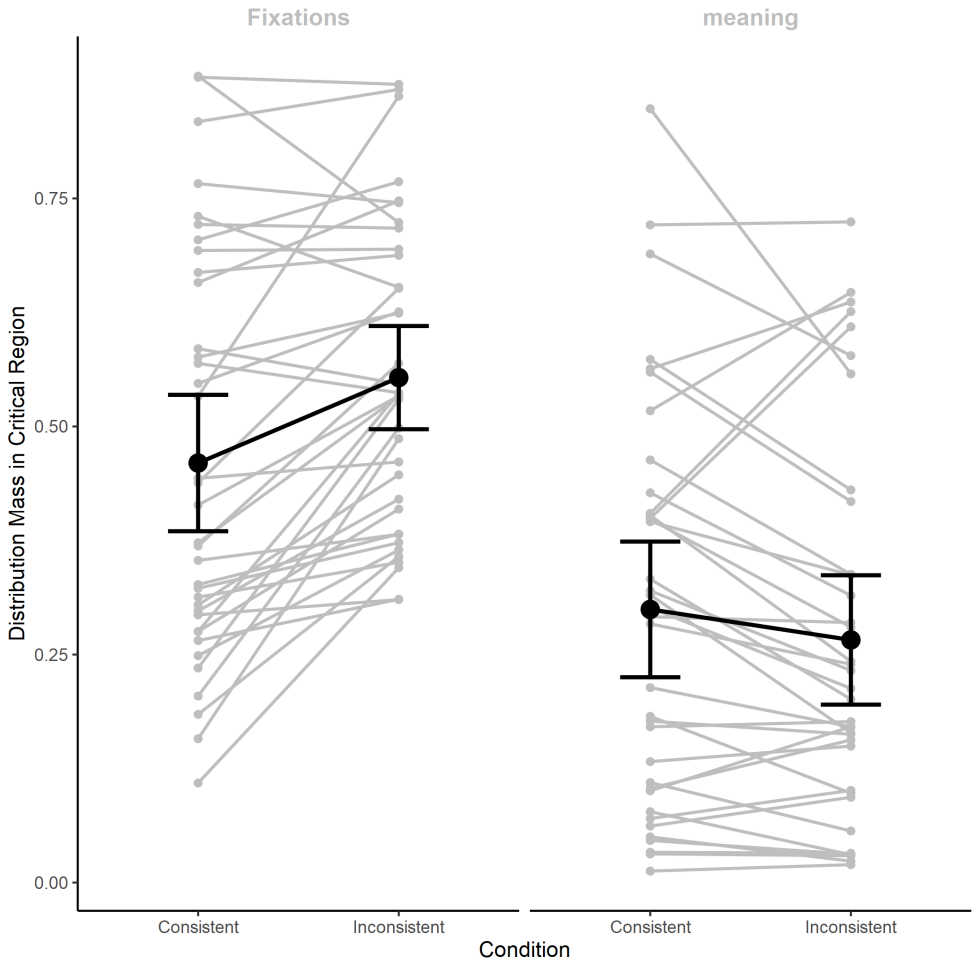

Fig. 3 Comparison of eye movements data and contextualised meaning maps

For each scene in each condition, for smoothed fixations and for the contextualised meaning maps, I calculated the amount of distribution-mass falling within the Critical Region (the region, in which the manipulated objects were located) divided by the Region's area. Comparing these values between conditions revealed that observers tend to fixate the Critical Regions more when they contained semantic inconsistencies (Inconsistent condition), as compared to the situation when they did not (Consistent condition; left plot). This effect was not predicted by contextualised meaning maps (right plot), as they did not attribute more 'meaning' to semantic inconsistencies. Each grey line indicates a single scene, black oblique lines connect the means, black vertical lines indicate standard errors. Grey lines indicate single images; black vertical bars indicates means with 95% confidence intervals.

Sensitivity of patch ratings to semantic manipulations

It should be noted that transforming patch ratings into contextualised meaning maps involves a number of steps, some of which include non-linear transformations. In order to exclude the possibility that these steps mask real between-conditions differences in the full maps and in their coarse components, or unintentionally introduce incidental between-conditions differences in the fine components, I conducted two analyses on the raw rating data. In the first analysis, I selected all patches, which had an overlap of at least one pixel with the Critical Regions, and discarded the remaining patches. The ratings for patches from each condition were averaged for each scene, separately for coarse and fine patches. Averaging allowed us to account for between-scene differences in the number of patches overlapping with Critical Regions and guaranteed that the data from each scene had an equal contribution to the subsequent analyses. A comparison of these average ratings between conditions did not provide any evidence to suggest that between-condition differences were present in the raw data but were masked in the processes of assembling contextualised meaning maps (see Table 1 rows 1 and 4 and Fig. 4).

Note, however, that in this analysis, I included all patches with at least one pixel overlap with the Critical Regions. These regions were derived from the bounding boxes of the objects (see *Stimuli and eye-movements data* section for details). Consequently, some patches showed only small parts of the manipulated objects, or none at all. Averaging ratings for such patches with

those that clearly depict the manipulated objects might cover subtle effects. I therefore repeated my analysis of ratings with more stringent criteria for patch inclusion. In order for a given patch to be included in this second analysis, the percentage of its area overlapping with a Critical Region (dubbed Overlap Percentage henceforth) had to be above a certain threshold (see Table 1 and Fig. 4). For patches of each size, I tested two threshold values. These values were selected as 34th and 67th percentiles of all above-zero Overlap Percentage values. The motivation for using percentiles to determine the thresholds was to make sure that the consecutive analyses differ from each other by approximately the same percentage of retained patches: while in the first analysis I included 100% of patches which had above-zero Overlap Percentage, the thresholds resulted in including 66% (for 34th percentile) and 33% (for 67th percentile) of them. For each threshold and each scene, I averaged ratings of the retained patches, separately for each combination of experimental condition and patch size. Next, I compared these per-scene values between conditions (see Table 1 for full results). Only one of the resulting tests reached statistical significance: for the most conservative threshold, fine patches from the Inconsistent condition were rated as *less* meaningful than their equivalents from the Consistent one. The magnitude of this difference was small: it amounted to 0.28 points, which corresponded to 5.6% of the full range of the rating scale (spanning from 1 to 6).

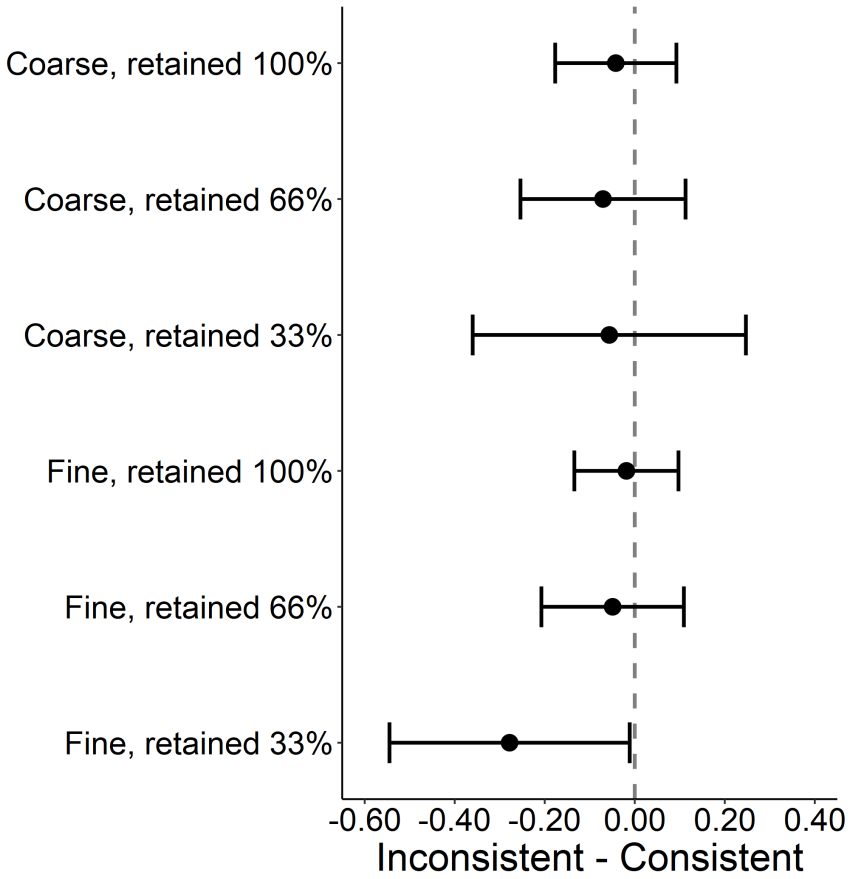

Fig. 4 Comparison of patch ratings between conditions – visualization.

For each scene in each condition, I averaged ratings from patches, which covered some parts of the Critical Regions, separately for coarse and fine patches. I then subtracted the values for Inconsistent from Consistent. Averages of these per-scene differences are presented on this figure, together with their 95% confidence intervals. For each patch-size, I conducted this analysis three times, including all patches overlapping with the Critical Regions, or only the top 66% and top 33% of patches ordered by the amount of their overlap with Critical Regions. As a result, the subsequent analyses were restricted to patches presenting larger parts of the manipulated objects. In all analyses, patches derived from the Inconsistent condition had the tendency to be rated as less meaningful than patches form the Consistent condition but this unexpected effect was statistically significant only for those fine patches, which shared the largest overlap with Critical Regions (see Table 1).

Table 1: Comparison of patch ratings between conditions – statistical results

Patch size	Percent of patches having above-zero Overlap Percentage included	Patch inclusion threshold: minimal Overlap Percentage	Number of included scenes[1]	Mean difference in ratings (Inconsistent – Consistent) with 95% confidence intervals	Paired t-test results[2]
Coarse	100%	>0	36	-0.04 [-0.18, 0.09	t(35) = -0.63, p = 0.530
	66%	0.07	35	-0.07 [-0.25, 0.11]	t(34) = -0.78, p = 0.440
	33%	0.21	27	-0.06 [-0.36, 0.25]	t(26) = -0.38, p = 0.705
Fine	100%	>0	36	-0.02 [-0.13, 0.1]	t(35) = -0.33, p = 0.747
	66%	0.18	36	-0.05 [-0.21, 0.11]	t(35) = -0.63, p = 0.533
	33%	0.56	30	-0.28 [-0.54, -0.01]	t(29) = -2.13, p = 0.042

[1] Because some scenes had small Critical Regions, for more conservative thresholds none of the patches derived from them had an Overlap Percentage high enough to be included in the analysis.

[2] I did not apply any correction for multiple comparisons here.

In my first experiment, I used a dedicated image data set to evaluate the sensitivity of contextualised meaning maps and human eye movements to manipulations of the semantic relationship between objects and scenes. As expected, human observers looked more at objects that are semantically inconsistent with the scene context compared to consistent objects. Contrary to predictions of the meaning maps approach, however, my results indicate that contextualised meaning maps assign similar 'meaning' to consistent and inconsistent objects. This insensitivity to manipulations of semantic object-scene relationships was present already at the level of the raw rating data. It therefore seems unlikely that it is merely due to the way in which ratings are combined into contextualised meaning maps. When I split the analyses of contextualised meaning maps or the raw data by the size of the patches (fine vs. coarse), there was, however, a slight indication that the effects of semantic inconsistencies on meaningfulness ratings and on eye-movements might go in opposite directions. Specifically,

when I only considered the contextualised meaning maps resulting from ratings of fine patches, the maps assigned *less* 'meaning' to the Critical Region, which was defined by the bounding boxes of the consistent/inconsistent objects. A similar effect was observable in the raw data of small patches (here called fine) that contained large parts of, or whole objects: those patches depicting objects that are inconsistent with the scene context were rated as *less* meaningful than the patches depicting consistent objects. If confirmed, this finding would not only challenge the meaning maps approach but would contrast starkly with typical explanation of the semantic inconsistency effect in eye-movement research which assumes that human observers look more at semantically inconsistent objects because the object-scene inconsistency results in these objects being semantically more informative or conveying larger amounts of 'meaning' (Henderson, 2011; Henderson et al., 1999; Loftus & Mackworth, 1978; Peacock et al., 2019). Given that the evidence of my first experiment was patchy and, at best, preliminary, I decided to conduct a second, more targeted experiment. I considered two hypotheses for why I found the effect only for a subset of fine patches. Firstly, it could simply be a false positive. Secondly, there might be a general but subtle tendency to rate semantic inconsistencies as less meaningful, but the subtlety of this effect might have meant that it could not be detected in ratings of coarse patches because of their low number. To adjudicate between these two hypotheses, I conducted Experiment 2. In this experiment, I created a single, well-controlled set of coarse patches derived from scenes with consistent and inconsistent objects, and collected ratings for them from multiple raters.

Experiment 2 – Methods

Stimuli and design

In this experiment, I used the same 72 photographs (of 36 scenes) as in Experiment 1. For each scene, I manually selected two coarse patches that fully contained the consistent and inconsistent objects (see Fig. 5). The locations of these patches were the same in both conditions but their content changed. These patches were dubbed Con and Incon. Con-patches were derived from scenes in the Consistent condition, Incon in the Inconsistent condition. In this experiment, I were primarily interested in the ratings associated with these two types of patches. Con- and Incon-patches were presented interleaved with other patches to mimic the

circumstances of a rating task used for creating contextualised meaning maps. The rationale for this interleaved presentation is explained in more detail at the end of the next paragraph.

When creating the set of additional patches, I relied on the ratings from Experiment 1 and selected six patches from each scene (see Fig. 5): two patches, which received the lowest meaningfulness ratings (dubbed L), one, which received the highest (dubbed H), and three patches, for which the ratings were midway between these extremes (dubbed M). This selection was carried out as follows. For each scene, I took all the coarse patches that did not overlap with the Critical Region. For each location occupied by these patches, I averaged ratings across the Consistent and Inconsistent conditions. I sorted the patches according to these average ratings in an increasing order and selected two from the bottom (L), one from the top (H), and the three closest to the median (M). Therefore, I selected eight patches for each scene in total: six patches, which were identical between conditions with respect to content (L, M, and H), and two patches, which differed (Con and Incon). Recall that my main goal was to compare ratings for the last two patches for each scene. The purpose for including the remaining patches in the patch-rating task was to ensure that raters have the opportunity to use all values from the meaningfulness-rating scale. Additionally, I wanted all the values to be used approximately equally frequently. Since I expected Con- and Incon-patches to be rated rather high, I included only one H-patch but two L-patches in order to maintain this intended balance.

For stimulus presentation, each L-, M-, and H-patch was paired with the full images from both conditions. In contrast, Con- and Incon-patches were paired only with either the consistent or the inconsistent scenes, respectively. This resulted in a set of 504 patch-contexts pairs (36 scenes x 2 conditions x 6 L/M/H-patches + 36 Con-patches + 36 Incon-patches). I split this set into two equally large subsets, each containing half of the patch-context pairs from one condition and half from the other in order to avoid the situation that raters would be exposed to the same scene in both conditions.

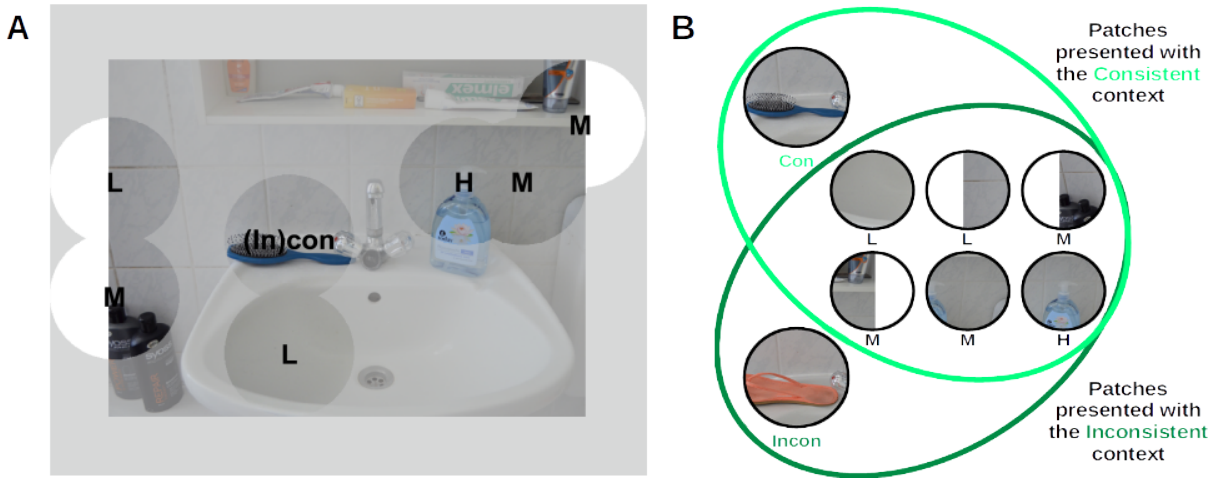

Fig. 5 Stimuli used in Experiment 2.

A, B) In the second experiment, I tested the hypothesis that in the patch-rating task used for creating contextualised meaning maps, patches depicting semantically inconsistent objects tend to be rated as less meaningful than their counterparts which depict consistent objects. To this end, for each scene, I selected two patches, which contained the consistent (Con) or the inconsistent (Incon) object, with the intention to compare the ratings they would receive. Testing my hypothesis, however, required mimicking the overall context in which the ratings were being provided when collecting data for contextualised meaning maps. Therefore, I additionally included six patches, which did not differ depending on whether the scene contained inconsistent object or not. These patches – according to the ratings they received when creating contextualised meaning maps for this scene – were either low in meaning (labelled L on the figure, two patches), high (H, one patch) or midway between these extremes (M, three patches). Some of the patches that were selected were close to image edges and were therefore clipped. [add or not: in the experiment,] A pool of raters viewed these patches paired with either the Consistent and Inconsistent context and provided meaningfulness ratings.

Sample-size justification

I planned to compare ratings for Con- and Incon-patches after averaging them per-rater and, therefore, treating data provided by each rater as two repeated measurements that I would compare using a paired t-test. Initially, I planned to resort to the analysis of statistical power to determine the number of raters to be recruited for Experiment 2. Following the logic that contextualised meaning maps were conceived as a method of predicting fixation-distributions,

I tied my power analysis to the magnitude of the effect of object-context inconsistencies observed in the eye-movements data. This effect was quite big and, in consequence, the calculated required sample size was small. Given how surprising the preliminary findings of Experiment 1 were, however, I decided that smaller effect sizes, which would require larger sample sizes, would be also interesting. Ultimately, I therefore decided not to determine my sample-size by means of power-analysis but instead by considering feasibility constraints, namely, the amount of resources I deemed reasonable for running Experiment 2. These considerations resulted in the recruitment of 140 raters, out of which 18 were excluded (see the *Rater inclusion criteria and inter-rater agreement* section). The final sample-size of 122 raters allowed detecting effects having the magnitude of Cohen's $D_z = 0.32$ with 95% power, when using paired, two-tailed t-test and when adopting a significance level of 0.05 (as indicated by the G-Power software; Faul et al., 2007).

Collecting meaningfulness ratings

Data collection was conducted identically to Experiment 1. I used the same patch-rating task (with the order of stimuli presentation randomized individually for each rater) and the same method of recruiting raters (Prolific platform).

Rater inclusion criteria and inter-rater agreement

I assumed that raters, who followed the task instructions, would agree in their ratings to a large degree. For example, I assumed that they would consistently rate M-patches higher than L-patches. Following that logic, I excluded raters, whose ratings vastly disagreed with the ratings provided by the majority of participants. I operationalized this idea by first measuring the agreement of ratings within each possible pair of raters who had viewed the same subset of patches using Krippendorff's α (Hayes & Krippendorff, 2007). Values of α span from negative values to 1, where 1 indicates perfect agreement, 0 indicates the degree of agreement achievable by chance, and the negative values indicate systematic disagreement. I calculated pairwise α for my raters using the function kripp.alpha from the R package irr (Gamer, Lemon, Fellows, & Singh, 2019), with the option *scaleType* set to 'interval' to indicate that my raters were using an interval scale. Next, for each rater, I averaged the α values from all pairs to which this rater belonged. These per-rater average α values (dubbed R_α from now on) indicated the

degree to which a given rater agreed with other raters who rated the same subset of patches. I visually inspected the histogram of R_α values calculated for all raters and decided that in my final sample, I would include only raters having R_α larger than 0.4. This resulted in excluding 18 raters and retaining 122. The average R_α for the retained raters was 0.70 (SD = 0.06). Additionally, I calculated R_α values for the excluded raters, using only data provided by them. These values indicated the agreement being close to the chance level (mean = -0.06, SD = 0.20) which means that these raters were most likely responding at random, rather than using a common rating-strategy, consistently differentiating them from the majority of my sample.

Experiment 2 – Results

Patches that were identical between condition (L, M, and H)

As a soundness check, I first tested whether L, M, and H-patches were rated as low, medium and high in meaning, respectively. I used Page's test, a non-parametric, rank-based statistical test assessing the ordering of values obtained in repeated measurements (Page, 1963), and compared the null hypothesis that there were no differences between ratings for all three types of patches against the alternative stating that L-patches (mean rating M = 1.5, SD = 0.58) were rated lower than M-patches (M = 2.5, SD = 0.64) which, in turn, were rated lower than H-patches (M = 4.6, SD = 0.72). I conducted it separately for patches from the Consistent and the Inconsistent conditions. In both cases the results were identical ($L = 1708$, $p < 0.001$) and indicated that the pattern of obtained results matched my expectations.

To evaluate whether the presence of consistent or inconsistent objects in a scene affects the ratings for all patches in that scene, I analysed whether ratings for L-, M-, and H-patches differed between consistent and inconsistent conditions. For each rater, I averaged ratings provided for each patch type per condition (see Fig. 6), and analysed the averages with a 2×3 repeated-measures ANOVA (with a Greenhouse-Geisser correction) with the two within-subjects factors Condition (Consistent and Inconsistent) and Patch-Type (L-, M-, and H-patches). As expected based on the preceding analysis, this analysis also showed that ratings differed according to patch type, as indicated by a main effect for this factor ($F(1.33, 160.45) =$

1376.33, $p < 0.001$). The other main effect and the interaction showed no significant differences (Condition: $F(1, 121) = 0.56$, $p = 0.457$; interaction: $F(1.33, 160.67)$, $p = 0.5$).

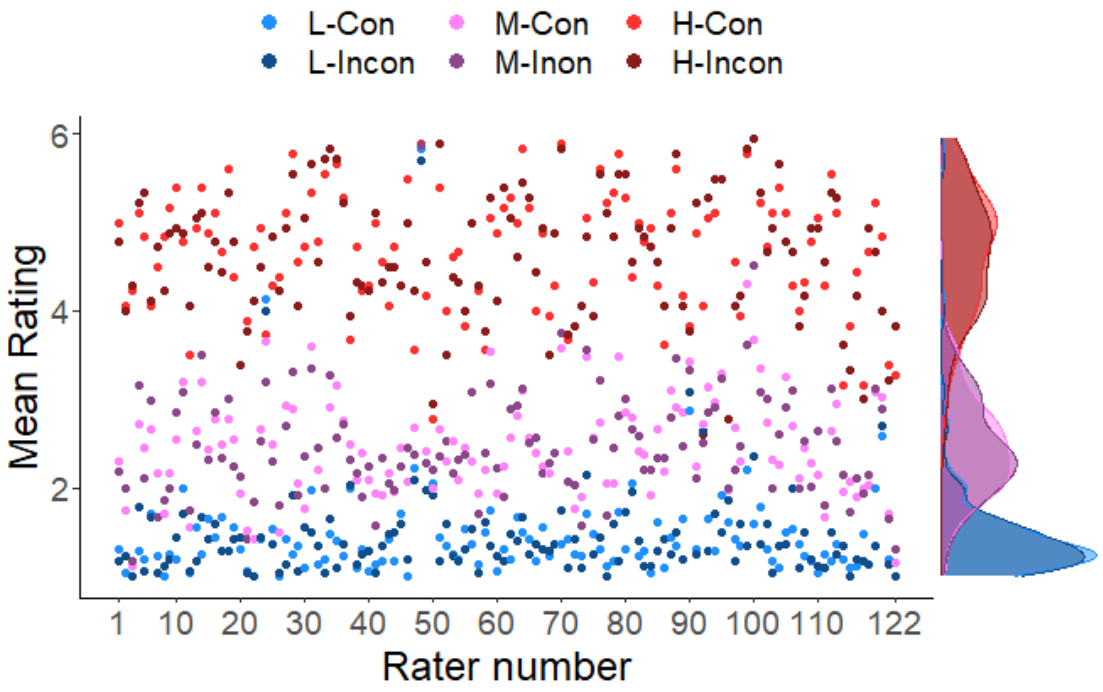

Fig. 6. Meaningfulness ratings obtained for L-, M-, and H-patches, averaged per rater over scenes and segregated by condition. Brighter colours indicate mean ratings from the Consistent condition, darker from the Inconsistent. On the right-hand side, density plots are shown.

In the final analysis of the L-, M-, and H-patches, I focussed on potential differences between individual scenes. The previous analyses reported in this section averaged patch ratings per rater over scenes. In my final analysis, I took a different approach and compared ratings provided for individual L-, M-, and H-patches across conditions. Individual patches were rated by a separate set of raters in the Consistent and Inconsistent conditions (see section *Stimuli and design*). I therefore used a between-subjects Welsh test to compare the ratings for each patch individually across conditions and found statistically significant differences only for 16 out of 216 patches. These patches constituted 7.4% of all L-, M-, and H-patches and were derived from different scenes. Therefore, in the vast majority of cases, the condition from which the context image was derived did not influence the ratings for individual patches. In fact, the number of detected differences is close to the level expected by chance, and none would

survive correction for multiple testing. Moreover, the few differences that arose did not come consistently from the same scenes and did not show consistent directionality (in 6 out of 16 cases where the differences were statistically significant, patches associated with the Inconsistent context were rated higher).

Overall, these analyses have two implications: first, they indicate that the task was successful in generating meaningful ratings, as suggested by the expected ordering of values for L-, M-, and H-patches. Second, exchanging a single object that is semantically consistent with the scene for an inconsistent object does not have general effects on the rating of patches that do not contain the manipulated object, neither on average nor on a scene-by-scene level.

Patches that were manipulated between conditions (Con and Incon)

The main focus of Experiment 2 was to assess whether objects that are semantically inconsistent with the scene context are rated differently with respect to the amount of meaning they convey compared to consistent object. To address this question, I used a similar approach to that employed for the analysis of the ratings for L-, M-, and H-patches. I averaged ratings provided by each rater, separately for all Con- and all Incon-patches, and compared them with a paired-samples t-test. In line with the preliminary findings of Experiment 1, and in direct opposition to previous assumptions (Henderson, 2011; Henderson et al., 1999; Loftus & Mackworth, 1978; Peacock et al., 2019), the results demonstrate that semantically inconsistent objects are rated as less meaningful compared to consistent objects ($t(121) = 5.87$, $p < 0.001$, mean of the differences: $M = -0.21$, 95% confidence interval [-0.14, -0.28], effects size = 0.53 (D_z).

To assess the contribution of the consistent vs. the inconsistent condition to this effect in a subject-by-subject approach, I ordered the raters by the difference between their average rating for Con- and Incon-patches. As shown in Fig. 7, this difference seems to be largely due to changes in ratings of inconsistent patches: while there was no clear subject-by-subject difference in the ratings for Con-patches, raters who contributed to the group-level effect showed decreased ratings for D-Incon patches.

This impression was corroborated by a statistical analyses that showed a significant correlation between Con/Incon differences and the Incon ratings ($r(111) = .49$ [.33; .62], $p < 0.001$), but no

such relationship for Con ratings (r(110) = 0.01 [-.17; .20], p = 0.885). Note that - for each analysis separately – I excluded points which had a Cook's distance higher than 3 times the mean Cook distance for all points. For Con ratings, this exclusion threshold amounted to 0.02 (0.03 for Incon) and resulted in 10 exclusions (9 for Incon). Data without the excluded points is shown on Fig 7. I applied these exclusion criteria because the initial inspection of the data suggested that, in each case, the effects might be driven by a small number of points, which would have a disproportionately large influence on regression. However, repeating the analyses with all the data included resulted in the same pattern of outcomes (Con: r(120) = -.09, p = 0.342; Incon: r(111) = 0.49 [.33; .62], p < 0.001).

These findings suggest that there is high consistency across rater regarding their evaluation of the meaningfulness of objects that are semantically consistent with their scene context. Ratings for inconsistent objects, in contrast, revealed substantial variability in raters behaviour. Different individuals tended to rate these objects as either lower, similar, or higher in meaning than the consistent objects. Ultimately, this difference not only offers interesting insights into individual differences but also suggests that the group-level effect is mainly driven by changes in the ratings of inconsistent objects.

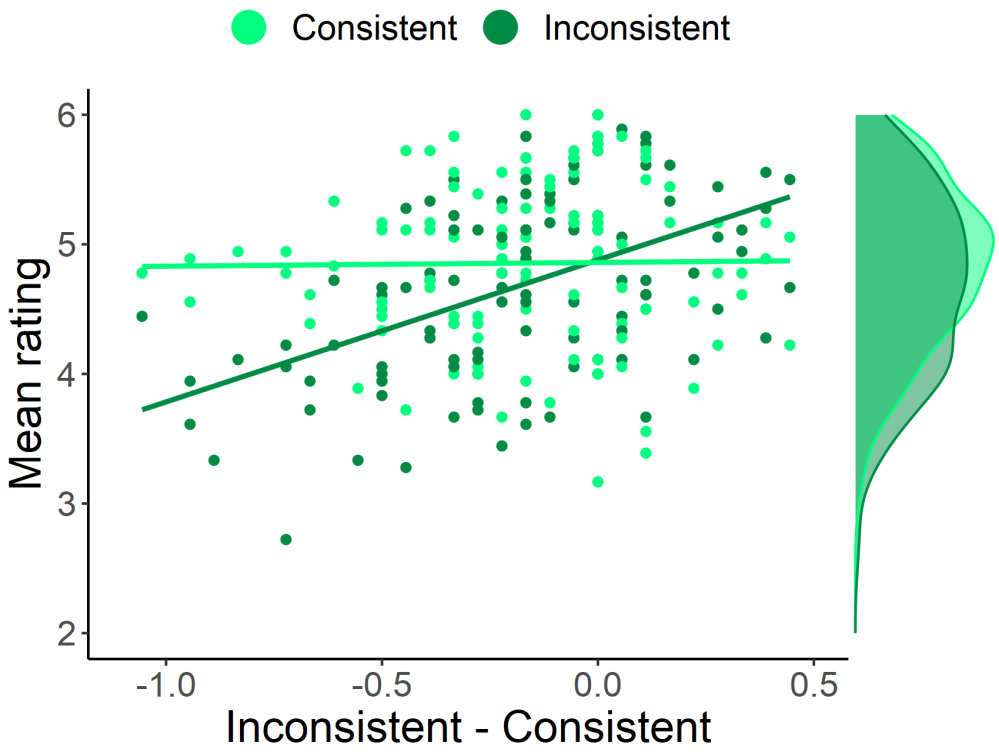

Fig. 7 *Meaningfulness ratings obtained for Con- and Incon-patches.*
For each rater, I averaged ratings provided for Con-patches (light-green points) and for Incon-patches (dark-green points). Next, I subtracted the average ratings for Incon-patches from Con-patches and ordered the raters according to these difference scores. The ratings for Incon-patches, but not for Con-patches, increase along this axis. Regression analyses conducted for both types patches separately confirmed this impression: the correlation between Con/Incon differences and ratings was significant for the Incon-patches, but not for Con- (see regressions lines on the plot). Therefore, the tendency to rate Incon-patches as less meaningful than Con-patches varied substantially across raters. Please note that this figure was generated using data not containing points identified as outliers based on their Cook's distance (see the main text).

My final analysis focused on the individual scenes, rather than individual raters, comparing ratings for Con- and Incon-patches derived from the same scenes. To that end, for each scene, I conducted a separate between-subjects Welsh test comparing ratings received by Con- and Incon- patches, similar to the analysis conducted for L/M/H-patches. Without the correction for multiple comparisons, 15 out of 36 of these tests yielded statistically significant results (this number was reduced to 3 after applying the correction). Out of these 15 cases, in 14 (39% of all scenes) the Incon-patch was rated as less meaningful than the Con-patch. Therefore, the tendency of Incon-patches to be rated as less meaningful than Con-patches was observable at the level of scenes too, which corroborates the finding from the rater-level analysis.

Discussion

In this Chapter, I tested the hypothesis that objects, which are semantically inconsistent with the scene context, strongly attract human fixations because they are more informative (carry more 'meaning'). In different forms, this hypothesis has been proposed by a number of authors for at least four decades (Henderson, 2011; Henderson et al., 1999; Loftus & Mackworth, 1978; Peacock et al., 2019) but has recently gained increasing attention with the development of meaning maps, a novel tool to index the distribution of 'meaning' across an image (Henderson & Hayes, 2017, 2018). Of particular interest are contextualised meaning maps, a recently proposed variant of the original technique. The distribution of 'meaning' indexed by these maps is based on aggregating crowd-sourced judgements about the meaningfulness of

multiple image-patches cut from an image. Importantly, when providing meaningfulness ratings for the patches, the individuals view the full scenes from which the patches have been derived. By contrast to the original meaning maps, the raters can therefore take the scene context into account when evaluating the meaningfulness of image parts. In a first experiment, I created contextualised meaning maps for images of scenes with objects, which are either semantically consistent or inconsistent, and compared them to eye-movement data for the same stimuli. While observers tended to look more at inconsistent compared to consistent objects, contextualised meaning maps did not attribute higher amounts of 'meaning' to the former than the latter. Even more surprising, my first experiment provided preliminary evidence to suggest that the same scene location might even be indexed as less rich in 'meaning' when it contains semantic inconsistencies. In a second experiment, I asked 122 raters to provide meaningfulness ratings for a carefully controlled set of image patches, including patches that showed semantically consistent or inconsistent objects. The results of this second experiment indicate that humans have a tendency to judge objects that are semantically inconsistent with the scene as less meaningful than their consistent counterparts.

The tendency of human observers to look more at semantically inconsistent objects is considered to be a prototypical example of semantic influences on eye movements. Previous explanations of this effect implicitly or explicitly assumed that semantic inconsistency increases the amount of semantic information, or 'meaning' that is conveyed (Henderson, 2011; Henderson et al., 1999; Loftus & Mackworth, 1978; Peacock et al., 2019). This interpretation has been strongly expressed within the recently developed meaning maps approach (Henderson & Hayes, 2017; Henderson et al., 2019). In contrast to this notion, my direct evaluation of contextualised meaning maps suggests that, while showing an overall good ability to predict human gaze patterns, they are unable to predict influences of semantic inconsistencies, showing no difference between consistent and inconsistent conditions. At the most basic level, my findings therefore show that contextualised meaning maps fail to capture at least one critical factor that guides eye-movement control. It might be that contextualised meaning maps, similarly to the original meaning maps, index complex local features that often act as carriers of semantic information in visual scenes, but fail to directly measure 'meaning' *per se* – see previous Chapter. Note that the raters were instructed to base their meaningfulness judgements on 'how informative or recognizable' they think an image patch is and this explicit

reference to recognizability could encourage the raters to implicitly considering the presence (or absence) of complex, object-related features in the patch when providing the judgement.

The conclusion that meaning maps and contextualised meaning maps do not measure meaning is rooted in a specific notion of meaning, focusing on the degree to which certain objects are probable for certain contexts (see Loftus & Mackworth, 1978). This definition captures at least one way, in which semantic processing is relevant to oculomotor control, as demonstrated in my study and many other cited here. While this definition pin-points the nature of object-context inconsistencies and was useful for revealing the limitations of the meaning maps approach, it has its own limitations, too. Most importantly, it does not highlight (but does also not preclude) the fact that different objects can be semantically related to each other to different degrees. For example, consider a corkscrew, a fork, and a knife. Intuitively, while all three are highly consistent with the contexts of a kitchen, the fork and the knife are more closely related to each other than any of them is to the corkscrew. Such a relational understanding of 'meaning' has already provided interesting insights into human oculomotor control (reviewed in Wu, Wick, et al., 2014; see also Boettcher et al., 2018). For example, it has been demonstrated that, during scene viewing, consecutive fixations tend to land on objects which are semantically related to each other (Hwang, Wang, & Pomplun, 2011; Wu, Wang, et al., 2014). Note that this relational conceptualization of meaning is distinct from the conceptualization proposed by the meaning maps approach: the former abandons the idea which is the core of the latter, namely, that 'meaning' is gradual (that is, high for some objects and low for others) and, consequently, directly comparable to saliency (which also might be high or low for different image regions). Looking from a broader perspective, it is becoming increasingly clear that there are a number of areas in vision research, in which the notion of 'meaning' is in dire need for conceptual clarification. The particularly pressing problem in the current context is the idea that 'meaning' is a unitary concept. The distinction between 'prior-probabilities meaning' and 'relational meaning' suggests that, in fact, it might be not. The possibility that there might be several subtypes of meaning that are important for eye movements has recently already been suggested by other authors (Henderson, Hayes, Rehrig, & Ferreira, 2018; Williams & Castelhano, 2019). In line with this idea, while contextualised meaning maps and patch ratings might measure one type of meaning, they might ignore other types. The critical question then is what type of meaning, or what information, the patch-rating

task provides access to. Answering this question is impeded by the fact that it is far from clear what raters are doing when asked to provide meaningfulness judgments for image patches. Currently, participants are a black box with respect to their meaningfulness judgments, and it is unclear what information is taken into account or what processes lead to different ratings. A related difficulty concerns the possible dependence of ratings on the specifics of the instructions given to raters. While my study did not address this question, it seem plausible that changing the instructions might affect meaningfulness rating. In both experiments, I modelled my instructions on those used in the original contextualised meaning maps study by Peacock et. al (2019). These instructions do not provide strong guidance as to the raters' task, because the intention was to compare the contextualised meaning maps created from these ratings to eye-movements measured during free-viewing, a similar, weakly-constrained context. However, it is possible that for a more precisely defined task, raters' behaviour would be very different. For instance, imagine observers would have been told that the images in the study show crime scenes. It seems possible if not likely that raters would pick out the semantically inconsistent objects as being particularly meaningful in this context. These considerations illustrate not only the potential sensitivity of the patch-rating task to the changes in instruction but also how drastically the amount of 'meaning' carried by different image parts can change as a function of a context.

Given the limitations of human rating data, current developments in computational approaches might provide a more fruitful or, at least, complementary alternative that could contribute to a better understanding of the role of high-level factors in eye-movement control, including semantic information and 'meaning'. A number of authors have attempted to develop indices of these high-level aspects of visual input by applying techniques to images that have originally been developed in natural-language processing (Hwang et al., 2011; Lüddecke, Agostini, Fauth, Tamosiunaite, & Wörgötter, 2019; Rose & Bex, 2020; Treder, Mayor-Torres, & Teufel, 2020), in particular in the field of distributional semantics (Harris, 1954). While, of course, these computational methods come with their own limitations, they have a number of key advantages over human rating data. To begin with, they are comparably inexpensive, fast, and easy to use, and can comfortably be applied to large image data sets due to their automation. More importantly, if used wisely, computational tools have the potential to be less opaque compared to human rating data, and might be more amenable to detailed analyses of which

specific statistical aspect of high-level scene contents contributes to eye-movement control. For instance, the finding that humans look more and longer at 'semantically' inconsistent objects might be based purely on a statistical analysis of object co-occurrences in visual scenes (Wang, Hwang, & Pomplun, 2010). Not surprisingly, recent analyses of image datasets with more than 20000 images indicate that different scene categories indeed show a highly consistent clustering of object types (Treder et al., 2020), and the oculomotor system might exploit these regularities for outlier detection. This interpretation of the influence of object-scene inconsistencies on eye movements is similar in spirit to earlier notions of saliency (Bruce & Tsotsos, 2009), but transfers this idea from a low-level (feature-based) to a high-level (object- and scene-based) analysis of the visual input. While – most likely – being an important contributor, co-occurrence *per se* does not necessarily amount to a semantic relationship between objects, or 'meaning'. And some computational approaches, such as the one developed by Treder and colleagues (2020), might have the potential to determine whether oculomotor control relies purely on basic co-occurrence or transforms these raw data further into a type of information that is closer to what I might label 'meaning'.

The caveat to bear in mind when interpreting all my results is that eye-movements data and meaningfulness ratings were collected from different people. However, at least at the level of between-group comparisons, I observed that introducing semantic inconsistency to a certain scene region elicits the increase in the number of fixations registered on this region and, at the same time, the decrease in the meaningfulness ratings provided for this region. Interestingly, this decrease was underpinned by a considerable variability in ratings revealed in Experiment 2: while the majority of raters, on average, rated patches from the Inconsistent condition as less meaningful than the patches from the Consistent condition, a small number of raters rated both kinds of patches as almost equally meaningful, or even had a reversed tendency. On the one hand, this high variability severely limits the usefulness of the patch-rating task for its original purpose. Recall that creating contextualised meaning maps involves averaging and pooling the ratings provided by different raters. These procedures implicitly assume that the raters are interchangeable with each other and my results clearly show that they are not. On the other hand, the high variability in responses makes this task a potentially interesting tool for indexing individual differences (Hedge, Powell, & Sumner, 2018). While currently the clarity regarding the processes underpinning the selection of rating values is lacking, further research,

combining the patch-rating task with other measures, might shed more light on this issue, and – thereby – on how different individuals process the content of natural scenes. This topic is still understudied in the context of eye movements, despite the evidence showing that such individual differences exist (De Haas et al., 2019; see also Kröger et al., 2020). For example, humans exhibit idiosyncratic biases regarding the kinds of semantic information contained in scene-regions fixated first and these biases are linked to other characteristics of individuals (De Haas et al., 2019). First fixations might reveal differences in more involuntary processes related to scene processing, so the patch-rating task can supplement them by offering insights into more deliberative ones. Importantly, the patch-ratings and eye movements, as shown in my study, are not always in concordance with each other.

To summarize, in my first experiment, introducing a semantic inconsistency to a scene region by replacing a semantically consistent object with one that is semantically inconsistent did not change the amount of meaning attributed to this region by contextualised meaning maps, despite increasing the number of human-fixations landing on this region. Therefore, even though the maps predicted human fixations well for scenes containing only typical objects, the 'meaning' they measure was not able to account for semantic influences on human gaze allocation linked to object-context inconsistencies. In fact, data from this experiment provided preliminary evidence suggesting that people might have the tendency to treat semantically inconsistent objects as less meaningful than their consistent counterparts. The second experiment corroborated this conclusion: individuals performing a patch-rating task – the backbone of contextualised meaning maps – indeed had the inclination towards rating image-patches depicting inconsistent objects as less meaningful. The strength of this inclination, however, varied substantially across individuals. Results of both my experiments may serve as a springboard for the much-needed in-depth discussions about the meaning maps approach which inspired my study and, more generally, the role of semantic information in human oculomotor control and individual differences in processing it.

Chapter Four – knowledge-driven perceptual organisation reshapes information sampling via eye-movements

Introduction

Both previous Chapters highlighted the intricate and complex relationship between image-computable features and semantic information, which is carried primarily by objects (see Chapter One). While features are necessary for visual object representations to arise, they are often not sufficient. Indeed, a growing number of studies suggests that in order for object representations to emerge, prior object-knowledge has to flexibly interact with early visual mechanisms (Christensen, Bex, & Fiser, 2015; Flounders, González-García, Hardstone, & He, 2019; Hsieh, Vul, & Kanwisher, 2010; Lengyel et al., 2019; Neri, 2017; Ongchoco & Scholl, 2019; Teufel, Dakin, & Fletcher, 2018). In effect, prior object-knowledge reorganizes sensory processing of low-level, image-computable features in order to carve up the inputs into meaningful units (Teufel & Fletcher, 2020). While these object-oriented effects of information sampling are well-established, the current literature provides little consensus as to what specific aspect of objects influence programming of eye movements (Borji & Tanner, 2016; Federico & Brandimonte, 2019; Henderson, Malcolm, & Schandl, 2009; Nuthmann, Schütz, & Einhäuser, 2020; Van der Linden, Mathôt, & Vitu, 2015). The present Chapter assesses implications resulting from this novel conceptualisation of objecthood for the understanding of information sampling via eye movements in human observers.

Conventional accounts of object perception in both biological and machine vision are feature-based (Kourtzi & Connor, 2011; Kriegeskorte, 2015; Lee, 2015; Marr & Nishihara, 1978): extraction of visual boundaries or edges formed by feature dissimilarities in images is thought to be among the core processes involved in segmenting a figure from its background. The extracted features provide the input to a hierarchical system of down-stream visual mechanisms, which combine them into object representations. Recent developments in object

perception, however, demonstrate that the relationship between image-computable features and object representations is substantially more complex than advocated by this conventional account. This complexity is often concealed by the fact that object locations are strongly correlated with clusters of specific features, such as edges (Elazary & Itti, 2008; Masciocchi, Mihalas, Parkhurst, & Niebur, 2009). Yet, this correlation should not be mistaken for causation. Disregarding cases of hallucinations (Horga & Abi-Dargham, 2019; Teufel et al., 2015), image-based features are clearly necessary for visual object representations to arise. That they are not sufficient, however, is demonstrated by work indicating that identical images can lead to categorically different object representations depending on the observer's prior knowledge. In these cases, prior object-knowledge effectively generates objecthood by flexibly guiding extraction, processing, and organisation of lower-level features (Christensen et al., 2015; Christensen, Bex, & Fiser, 2019; Flounders et al., 2019; Hsieh et al., 2010; Lengyel et al., 2019; Neri, 2017; Ongchoco & Scholl, 2019; Teufel et al., 2018).

The most influential early models of information sampling via eye-movements (saliency models) have largely disregarded objects, arguing that programming of eye-movements is controlled by an analysis of low-level features such as luminance, colour, and orientation (see Chapter One). According to these early accounts, the visual system computes a number of maps on the basis of featural analyses, which highlight areas in the image that attract fixations (Zelinsky & Bisley, 2015). Over the past decade, however, a number of studies have emphasised the importance of "objects" or "semantic information" in guiding information sampling (Einhauser, 2013; Henderson & Hayes, 2017; Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013; Rider, Coutrot, Pellicano, Dakin, & Mareschal, 2018; Stoll, Thrun, Nuthmann, & Einhäuser, 2015). In one of the early studies, Einhäuser and colleagues (2008) found that maps of object locations outperform maps derived from low-level feature models in predicting human fixations. Moreover, human observers show a tendency to look at the centre of objects rather than their edges, contrasting with predictions from (some) low-level feature models (Nuthmann & Henderson, 2010; Pajak & Nuthmann, 2013; Stoll et al., 2015). These effects have been interpreted as demonstrating the importance of objects in oculomotor control. An even more ambitious approach is based on a novel technique called meaning maps (Henderson & Hayes, 2017). Such maps are created by segmenting a visual scene into small, isolated patches, which are rated for their meaningfulness independently from the rest of the image. These

ratings are pooled together into a smooth map, which is supposed to capture the distribution of meaning across the image. Meaning maps are better at predicting human fixations in comparison to a low-level feature model (GBVS; Harel, Koch, & Perona, 2007), a finding that has been used as evidence to suggest that eye-movements are controlled by the semantic properties of images (Henderson & Hayes, 2017).

The notion that eye-movements are controlled by object locations, or by the meaning of image parts has not remained unchallenged. For instance, a careful psychophysical study has recently suggested that the tendency of human observers to focus on the centre of objects might be controlled by a relatively simple process that programs eye-movements towards homogeneous luminance surfaces on the basis of luminance-defined edges (Kilpelaïnen & Georgeson, 2018; see also Van der Linden et al., 2015). More generally, a potential limitation of almost all previous studies that aim to show the contribution of objects, or semantic meaning to oculomotor control is their reliance on a comparison to models that calculate low-level feature maps as their null hypothesis. The specific choice of model has been shown to be critical, with changes in the model sometimes demanding dramatic reversals in interpretation (Borji, Sihite, & Itti, 2013; Pedziwiatr, Kümmerer, Wallis, Bethge, & Teufel, 2021).

Independently of the favoured interpretation of these findings, there is a more fundamental aspect that is easily overlooked. The emphasis on contrasting outputs of low-level feature models with "objects" or "semantics", and the tendency to conceptualise these as categorically different interpretations, has concealed a fundamental similarity between low-level models and those studies that have aimed at showing the importance of objects or semantic meaning in oculomotor control. Specifically, comparable to how low-level models deal with simple features, these studies implicitly treat "objects" or "semantic information" as image-computable properties. This notion is also the basis for state-of-the-art computer vision models that aim to predict human fixations: these models use deep convolutional neural networks trained on object recognition in order to extract high-level features that are directly computed from the image (Kümmerer, Wallis, Gatys, & Bethge, 2017b; Thomas, 2016). In other words, rather than providing diametrically opposed interpretations, the different approaches in the current eye-movement literature can be understood as lying on a continuum, with their position being defined by the features they emphasise. This notion is made explicit in a recent

study by Schütt and colleagues (Schütt, Rothkegel, Trukenbrod, Engbert, & Wichmann, 2019). The authors explicitly conceptualised objects as high-level features that are computed in a bottom-up fashion and contrasted their contribution to the guidance of eye-movements with the contribution of low-level features.

While the theoretical precision of the study by Schütt and colleagues is exceedingly helpful in clarifying the different positions, conceptualising objects as high-level features conflicts with current developments in object perception. Two aspects of the complex relationship between features and objects are particularly relevant: first, a number studies demonstrate that features are not necessarily sufficient for object representations to arise. Rather, objecthood emerges as a consequence of the interaction between current visual input and prior object-knowledge. Second, once object representations have been generated, top-down influences re-shape the way in which even some of the earliest cortical mechanisms process low-level visual features (Christensen et al., 2015, 2019; Flounders et al., 2019; Hsieh et al., 2010; Lengyel et al., 2019; Neri, 2017; Ongchoco & Scholl, 2019; Teufel et al., 2018). For instance, psychophysical studies show that early feature-detector units are sharpened for currently relevant input based on top-down influences from object representations that emerge as an interaction between input and prior object-knowledge (Teufel et al., 2018). The re-shaping of information processing is detectable in early retinotopic cortices (Flounders et al., 2019; Hsieh et al., 2010). Overall, these findings thus cast serious doubt on the notion that the human visual system computes image features independently of the inferred object structure of the environment (Neri, 2017), regardless of whether they are low- or high-level.

This novel perspective of object perception has fundamental implications for the understanding of information sampling via eye movement. First, if objecthood emerges from the interaction of prior knowledge and image-computable features, then the question of whether objects guide eye movements cannot be answered by an approach that exclusively focuses on how the oculomotor system carves up image-computable feature space, regardless of whether the considered features are low- or high-level. Second, the novel perspective of object perception means that a full understanding of the role of objects in eye-movement control has to move away from regarding feature space as static, taking into account the plasticity of low-level sensory processing introduced by dynamic interactions with object

representations. Here I address both of these limitations. I analysed gaze data from human observers viewing two-tone images. On initial viewing, two-tone images are experienced as a collection of meaningless black and white patches. After gaining relevant object knowledge, however, the observers' visual system organises the sensory input into meaningful object representations. I demonstrate that this knowledge-driven perceptual organization of inputs substantially re-shapes eye-movement patterns, with the selection of fixation locations being driven by a combination of image-computable features and the knowledge-dependent object representations. In summary, I show that a fundamental human visual behaviour – information sampling via eye movements – is guided by object representations that emerge when prior object-knowledge restructures sensory input, rather than being based solely on image-computable features, regardless of whether they are low- or high-level.

Two-tone images

The history of using stimuli similar to two-tones images in psychology dates back to 1957 (Mooney, 1957), when Mooney used black and white images of human faces in a test measuring 'closure': the ability to spontaneously perceptually bind image-features into an object. In this test, as well as in its revised version proposed recently (Verhallen & Mollon, 2016), observers have to look at two-tone face until they decide if they can recognize it as a face or not (therefore, this test, although designed to measure 'closure', is also sensitive to individual differences in face processing). Because of this study by Mooney, it became customary among researchers to call black-and white, degraded images 'Mooney images'. This name, however, is misleading. It suggests that these images – similarly to Mooney faces – can be recognised spontaneously, while the purpose of creating them is often opposite: researchers do not want them to be recognisable for observers who have not been exposed to their templates before. Therefore, to avoid this confusion, in this Chapter I refrain from using the term 'Mooney images' and use 'two-tone' images instead.

Thus far, two-tones images have been used mainly to investigate object perception (Flounders et al., 2019; González-García, Flounders, Chang, Baria, & He, 2018; Moore & Cavanagh, 1998), individual differences (Teufel et al., 2015; Tulver, Aru, Rutiku, & Bachmann, 2019), and memory and learning (a so-called one-shot learning; Ishikawa & Mogi, 2011). Importantly, they were also

used in eye-tracking studies (M. E. M. Król & Król, 2018; M. Król & Król, 2019; Loth, Gómez, & Happé, 2010). I refer to these studies later in this Chapter.

Experiment 1 – Methods

Overview

In Experiment 1, observers viewed two-tone images while their eye movements were recorded. Two-tones are derived from photographs of natural scenes ('templates') by smoothing the image, and binarising pixel values around a threshold. Each two-tone appears as meaningless patches on initial viewing. Once an observer has acquired relevant prior object-knowledge by viewing the corresponding template, however, processes of perceptual organization in the visual system bind the patches of the two-tone image into a coherent percept of an object (see caption of Fig. 1 for instructions of how to experience the effect).

Two-tone images provide a tool to manipulate object perception without changing the visual features of the stimulus. They are therefore ideally suited to test the hypothesis that human oculomotor control is determined by object representations that are not constituted by image-computable features but emerge via an interaction between features and prior object-knowledge. According to this idea, eye movements in response to two-tone images should be determined by the observer's subjective object percept rather than the objectively measurable features. Specifically, when an observer binds a given two-tone into a meaningful object percept, patterns of fixations should be more similar to those measured in response to the corresponding template, compared to when the two-tone image was perceived as meaningless.

To test these predictions, I recorded eye-movements of 36 human observers who viewed two-tone images before (Before condition) and after (After condition) being exposed to the relevant templates (Template condition) – see Fig. 1. In the Before condition, observers perceive two-tone images as meaningless black and white patches. In the After condition, they have received the relevant object-knowledge to bind patches into meaningful object percepts. It is critical to note that any potential differences in eye movements between the Before and the After conditions cannot be explained by image-computable features because these are

identical across conditions. The only aspect that has changed is the prior object-knowledge that observers have access to.

Observers

In this study, the primary unit of analysis was not a single observer, but the distribution of fixations on a single image (pooled across observers). Therefore, I selected the number of observers based on the estimated approximation of my empirical fixation distributions to the theoretical distributions, obtained from the population of infinitely many observers. Previous work has shown that fixations from 18 observers provide a sufficiently good approximation, and that further increasing the number of observers results only in marginal improvements (Judd, Durand, & Torralba, 2012). I therefore set my minimal number of observers to 18. However, one of my analyses required splitting my sample into two groups and I therefore recruited 36 observers in total, ensuring sufficient amounts of data in each groups after the split. All participants were Cardiff University students, had normal or corrected-to-normal vision, participated in the study voluntarily, and received either money or study-credits as a reimbursement. This experiment (as well as the other two reported in thisChapter) was approved by the Cardiff University School of Psychology Research Ethics Committee.

Stimuli

I used 30 pairs of images, where each pair consisted of a two-tone and its template. These stimuli were a subset of stimuli used in previous studies (Teufel et al., 2015) and the details of selecting templates and deriving the two-tones from them can be found in the respective article. In brief, template images – predominantly photographs of animals in their natural environments – were taken from the Corel Photo library. In order to derive the two-tones, templates were smoothed and binarized. A good two-tone image should be perceived as a collection of meaningless patches prior to seeing its template but observers should be able to easily bind the stimulus into a coherent percept of an object after they see the template. Extensive tests on naïve observers were conducted to select both the template images, and the parameters of smoothing and binarization that guarantee that the created two-tones have these desired properties.

Experimental setup

The experiment was conducted in a dark testing room. Participants sat 56 cm from the monitor, with their head supported by a chin and forehead. Their eye-movements were recorded using an EyeLink 1000+ eye-tracker placed on a tower mount and working with the sampling frequency of 500 Hz. The procedure was programmed in Matlab R2016b (Mathworks, Natick, MA) with the Psychophysics Toolbox Version 3 (Brainard, 1997; Kleiner, Brainard, & Pelli, 2007). Images were presented centrally on the screen, against a mid-grey background. They measured 21.9 degrees of the visual angle (788 pixels) horizontally and 14.6 (526 pixels) vertically. Templates were presented in greyscale.

Procedure

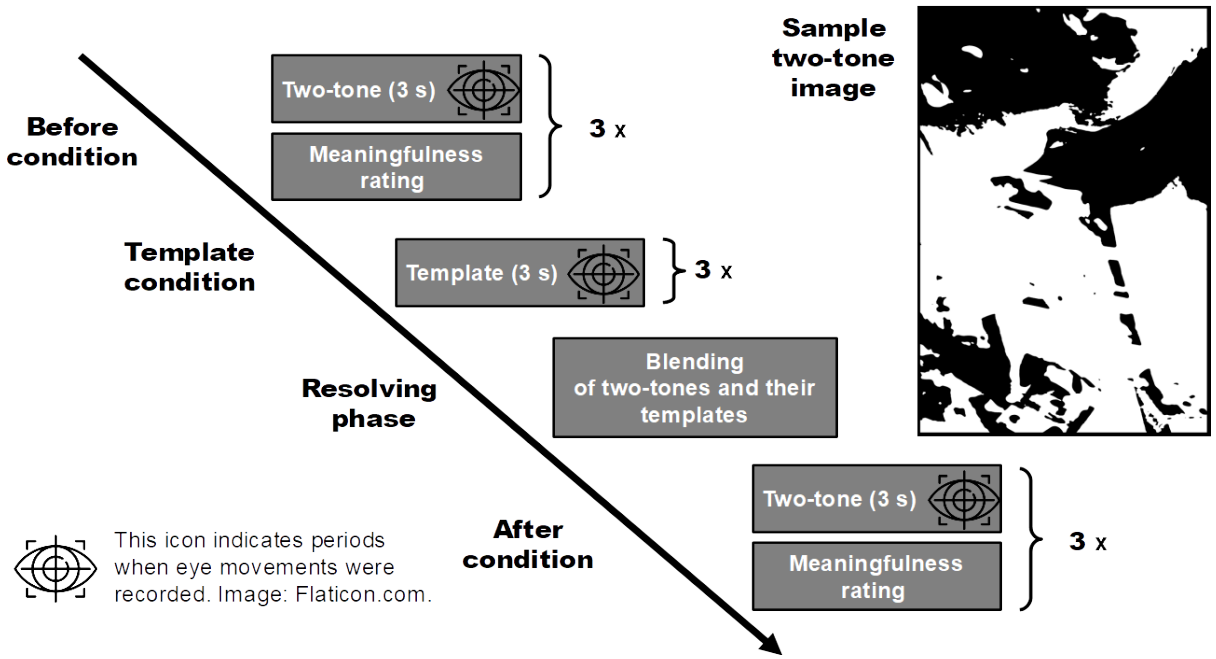

Fig. 1. Experiment 1 – Outline of a single experimental block and sample two-tone image.

In each block, observers free-viewed three two-one images while their eye movements were recorded (Before condition). After presentation of each image, they were asked to provide a rating of its perceived meaningfulness by adjusting a visual analog scale. This task was included as a manipulation check. Next, the grayscale templates of these three two-tones were presented while recording eye movements (Template condition). In order to ensure that observers acquired the relevant object-knowledge essential to bind the two-tone image into a meaningful percept, in the next part of the block, observers viewed the two-tones gradually blended with their

templates multiple times (Resolving phase). The After condition formed the final part of each block. It was identical to the Before condition in all aspects except for the order of presentation of the two-tone images, which was randomized for each condition. The whole experiment consisted of 10 blocks (30 two-tones in total) and throughout each block, the eye movements of observers were being recorded. In the upper-right corner of the figure, a sample two-tone image is presented (copyrights owner: author C. T.). For a naive viewer, this image appears as meaningless black and white patches. To be able to perceptually organize it into a meaningful percept, the reader is advised to first carefully look at the template image from which this two-tone was derived, presented on Fig. 2. This two-tone image was not used in the study.

The experiment consisted of ten blocks – a single block is schematically illustrated in Fig. 1. Before the start of the procedure, a 13-point eye-tracker calibration and validation was conducted. Each block started with the Before condition, in which three two-tones were presented in sequence, each for 3 seconds. Observers were instructed to carefully look at these images. Each of the two-tones was preceded with a centrally located fixation-dot displayed for 1 second and followed by a screen with a visual analog scale ('slider') used for collecting meaningfulness ratings. Observers adjusted the scale by pressing two buttons on a computer keyboard ('z' and 'm') and used the space bar to confirm their response. Then, a blank screen was displayed for one second. After providing a rating for the third two-tone, the Before condition was finished. It was followed by the Template condition, in which template images were displayed – again, each for 3 seconds, preceded by a fixation dot. In order to proceed to the next part of the experiment, the Resolving phase, observers had to press the space bar. It consisted of six cycles of dynamic blending between two-tones and their templates. Each cycle began with the presentation of a template image for two seconds. Then, it was linearly blended into the corresponding two-tone. During the blending, the value of each pixel was the weighted average of values from two-tone and template, and the weights – always summing to one – were changing dynamically. The full transition from template to two-tone was accomplished after 4 seconds. The two-tone remained on the screen for 2 seconds and then was blended-back into the template, remaining on the screen for another 2 seconds. Each of the three image-pairs used in a block was presented in two such blending cycles, but never twice in a row. The subsequent cycles of blending were separated with a 500 ms period, in which a blank screen was presented. Each block ended with the After condition, which was

identical to the Before condition except that images were presented in a newly randomized order. Observers had a break after each two blocks, and the eye-tracker was re-calibrated after each break. For each observer, images were assigned to blocks randomly and were presented in a pseudo-random order within each block. The pseudo-randomization ensured that the image shown last in the resolving phase was never presented at the beginning of After condition. The total time of the experiment was about 50 minutes.

The experiment began with instructions, which were delivered both verbally by the experimenter and written on the screen. The instructions were accompanied by a visual illustration of the key elements of the procedure: observers viewed a single two-tone image (not used in the actual experiment) and were asked to rate its meaningfulness on the visual analog scale. Second, they viewed its blending with the corresponding template. Finally, they viewed the same two-tone and were again asked to provide the meaningfulness rating.

Data pre-processing and analysis methods

The default EyeLink algorithm was used to extract fixation-locations from the eye-trace recordings. Further data pre-processing was done in Matlab. For each image, I discarded the first fixation, because first fixations were directed at the fixation-dot presented before image onset. I also discarded fixations not landing within the image-boundaries. Further details regarding data exclusions can be found in the *Data exclusions* section of the Appendix. For each image in each condition, I generated heatmaps (see examples on Fig. 3E) by first smoothing the discrete distribution of fixation with a Gaussian filter with a cutoff frequency of –6dB and then normalizing the smoothed distribution to the zero-one range.

The majority of my analyses focused on the similarity between two heatmaps. To quantify this similarity, I used Pearson's linear correlation coefficient calculated using the Matlab function *corr2*. This measure has previously been used used in the literature (Bylinskii, Judd, Oliva, Torralba, & Durand, 2016; Wilming, Betz, Kietzmann, & König, 2011), and its values have a straightforward interpretation. Specifically,, values ranged between zero and one, with one indicating that two heatmaps are identical and zero indicating a maximal dissimilarity. For statistical comparisons, I primarily relied on standard null hypothesis significance testing

techniques implemented in R (R Core Team, 2020) and Matlab. Because the majority of statistical comparisons I report here were conducted within-subjects (with single images serving as the 'subjects'), the t-tests reported throughout the text are the paired ones, unless otherwise stated. In order to assess the amount of evidence for a lack of a difference between groups of measurements, in some instances these analyses were supplemented with Bayes factors calculated using bayesFactor R package (Morey & Rouder, 2018).

Experiment 1 – Results

In the following part of this Chapter, I report several analyses of data from Experiment 1. First, the comparison of meaningfulness ratings provided by observers in the Before and After conditions. Second, two different analyses of spatial distributions of fixations: one focused on overall similarity and one relying on regions of interests. Third, I repeat these analyses but for fixations after image onset only. Fourth, I report an attempt to disentangle the contributions of image features and object representations to gaze guidance in the After condition. Fifth, I demonstrate that the key effects found in the gaze-pattern similarity analysis are observable also between subjects. Finally, I analyse how different aspects of oculomotor behaviour change between experimental conditions.

Manipulation check: prior object-knowledge changes perceived meaningfulness of two-tone images

In the Before and After conditions, observers rated the perceived meaningfulness of two-tone images. These ratings suggested that the perceptual experience of observers differed between these two conditions (see Fig. 2B and C). Averaging the ratings per image and comparing the obtained values between conditions revealed that the same images presented in After condition were perceived as more meaningful than when presented in the Before one ($t(29) = 23.84$, $p < 0.001$; mean difference $M_{diff} = 0.36$, 95% confidence interval CI = [0.4, 0.33]). The same pattern of results held when the ratings were averaged per observer: again, they were higher in the After condition compared to the Before condition ($t(35) = 14.42$, $p < 0.001$; $M_{diff} = 0.37$, 95% CI = [0.42, 0.31]). These results suggest that observers are able to bind two-tone images into meaningful object representations after – but not before – acquiring relevant prior-knowledge.

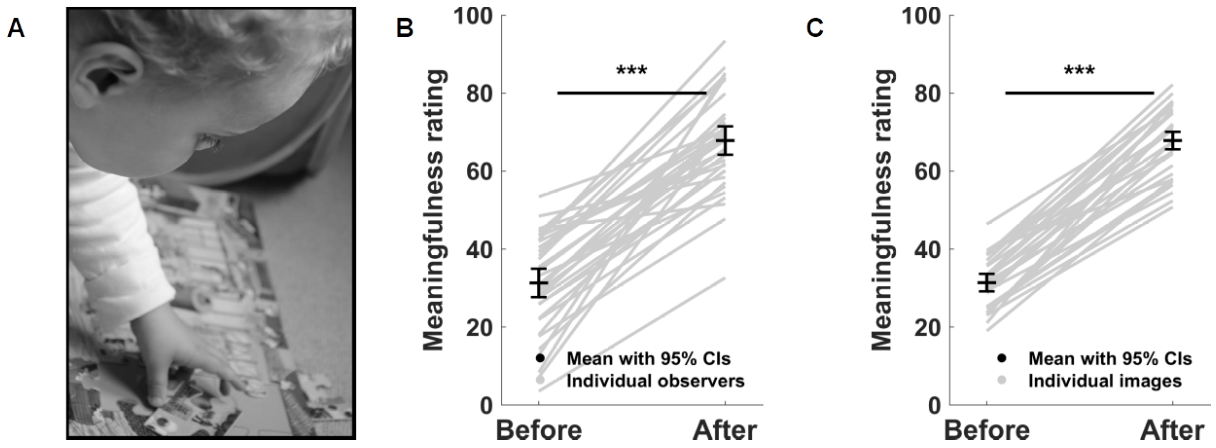

Fig. 2. Sample template image and image meaningfulness ratings.

*A) Template of the two-tone image from Fig. 1 (image copyrights owner: Christoph Teufel). B), C) Meaningfulness ratings for two-tone images provided a manipulation check. As expected, two-tone images were rated as more meaningful in the After than the Before condition, both when the ratings were averaged per observer (B) and per image (C). This finding demonstrates that acquiring relevant object-knowledge changes perception. Asterisks on plots indicate p-values: *** indicates p ≤ 0.001, ** indicates p ≤ 0.01, * indicates p ≤ 0.05, and 'n.s.' indicates the lack of statistical significance. Confidence intervals were Black horizontal bars indicate means. They are surrounded with 95% confidence intervals for within-subjects designs, calculated using Cousineau-Morey method (Cousineau, 2011; Morey, 2008). These conventions are used in all the remaining figures.*

Knowledge-dependent object representations control the spatial distributions of fixations

Observers perceived meaningful objects in the After but not the Before condition despite the stimuli being identical in both conditions. In the After condition, the experienced object representations were similar to those of the template images. The spatial distribution of fixations in the After condition should therefore resemble that from the Template condition, if knowledge-dependent object representations drive eye movements. I therefore compared the similarities of heatmaps between two pairs of conditions: Template-Before (mean correlation M = 0.72, SD = 0.13) and Template-After (M = 0.9, SD = 0.07) – see Fig. 3A. As predicted, I found a higher similarity between Template-After than Template-Before ($t(29) = 8.39$, $p < 0.001$; $M_{diff} = 0.18$, 95% CI = [0.14, 0.22]). This result suggests that gaze patterns in response to two-tone images resemble eye movements from the templates to a larger degree when the two-tones

were perceived as containing meaningful objects, as compared to when they were perceived as meaningless patches.

The distribution of fixations on images is not only determined by the characteristics of the visual input but also by general factors that are independent of image-specific content. One important general factor that is known to influence oculomotor control is the centre bias, a tendency of humans to visually inspect the centre of an image rather than regions closer to the edges (Tatler, 2007). A meaningful evaluation of the difference in similarities between Template-Before and Template-Before therefore requires a baseline that takes this centre bias into account. I modelled a centre bias for my data by creating a single heatmap from all fixations registered throughout the experiment. This heat map (labelled 'Centre') was then correlated with each individual heatmap from the Template condition. I found a statistically robust difference between the Template-Centre and Template-Before similarity scores (Template-Centre: M = 0.64, SD = 0.16 ; Template-Before: M = 0.72, SD = 0.13; t(29) = 2.40, p = 0.023 ; M_{diff} = 0.08, 95% CI = [0.14, 0.01]). Importantly, however, this difference was small (0.08 on average) suggesting that a general centre bias explained most (but not all) of the similarity between Template and Before.

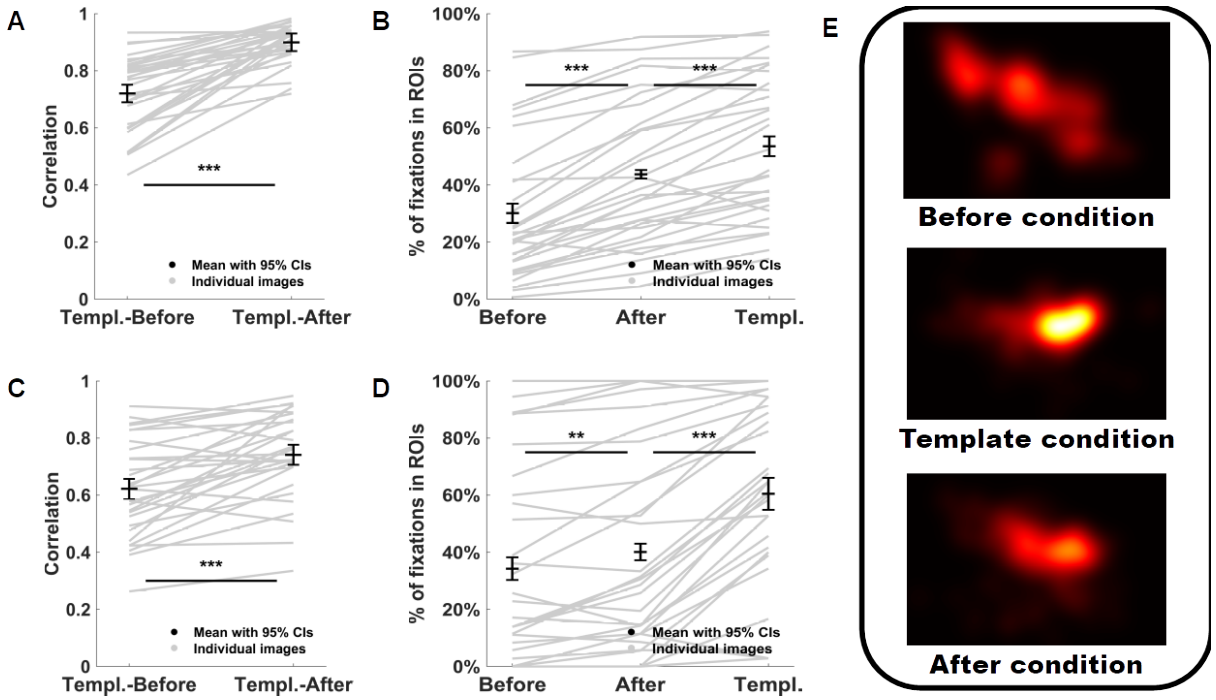

Figure 3. Results of Experiment 1.

A) The similarities of heatmaps from template images to heatmaps from two-tones, where the two-tones were viewed either in Before or in After condition. The patterns of fixations on the same two-tone images differed between the experimental conditions with respect to their similarity to patterns from templates. Specifically, after acquiring prior knowledge necessary for triggering knowledge-driven perceptual organization, when the observers were able to perceptually organize the two-tones into coherent percepts of objects, their gaze patterns on two-tones became more similar to the gaze patterns registered on the corresponding template images. This effect is illustrated for a single two-tone and its template on panel E. B) Percentage of fixations landing within the regions of interests (ROIs) in each condition. ROIs were defined for template images and covered their most informative parts, like the heads of depicted animals. It can be seen that the parts of two-tones encompassed by ROIs attracted more fixations in After condition than in the Before, which was consistent with the hypothesis that in After condition the eye-movements were guided by the representations of objects. C, D) The same analyses as on panels A and B but conducted including only first fixations from After condition. In both analyses, the effects of knowledge-drive perceptual organization – albeit weaker – were still evident, which indicated that they occurred fast enough to influence already the first eye-movement of observers. E) Sample heatmaps illustrating the distributions of fixations in all three conditions of Experiment 1. Top: heatmp from a two-tone image viewed in Before condition. Middle: heatmap from the template of this two-tone. Bottom: heatmap from the same two-tone as on top but from After condition. These maps were created from all fixations registered on the images. The values of all three maps were jointly normalised to zero-one range, so colour values are comparable between panels.

Changes to the spatial distribution of fixations are specific to image content

The analyses of heatmap similarities shows that the distributions of fixations on two-tone images becomes more similar to the distributions from their templates after observers have had access to prior object-knowledge. This result suggests that object representations that emerge when the features of two-tones interact with prior object-knowledge contribute to eye-movement control. To provide further support for this interpretation, I assessed in a more fine-grained manner the extent to which changes in fixation patterns related directly to object representations. For this analysis, I exploited two facts. First, template images depicted either

animals only (25 images), animals and humans (2 images) or humans only (3 images). Second, animal and human heads are known to attract fixations in natural scenes (Cerf, Paxon Frady, & Koch, 2009; Drewes, Trommershäuser, & Gegenfurtner, 2011). If the specific characteristics of knowledge-dependent object representations are important in the changes I observe between Before and After conditions, image regions that contain heads should attract more fixations in the After (and Template) conditions than in Before.

I tested this hypothesis by means of a regions-of-interest (ROIs) analysis. First, on each template, we manually labelled each pixel associated with a head of an animal or a human. For both the template and its associated two-tone image, the resulting mask served as the ROI (no 'buffer' was added around the masks). The masks covered 9% of image area on average (SD = 12%, median = 3%). For each image and condition, I calculated the percentage of fixations landing within the ROI by calculating what fraction of all fixations from a given image landed within the ROI (see Fig. 3B). The results showed an increase in the percentage of fixations landing within ROIs in the Before compared to the After condition, indicating that changes in fixations were object-specific (Before: M = 30%, SD = 24; After: M = 44%, SD = 25; $t(29) = 8.64$, $p < 0.001$; $M_{diff} = 0.14$, 95% CI = [0.1, 0.17]). In the Template condition, even more fixations landed within the ROI compared to After (Template: M = 54%, SD = 25; $t(29) = 6.02$, $p < 0.001$; $M_{diff} = 0.1$, 95% CI = [0.06, 0.13]). Overall, the ROI analysis provide clear evidence to suggest that the influence of knowledge-dependent object representations on fixation patterns is object specific.

Changes to the spatial distribution of fixations occur shortly after image onset

In order to assess whether the influence of knowledge-dependent object representations on oculo-motor control requires time to emerge or is present early on, I repeated my previous analyses but, instead of including all fixations for each image and condition, I focused exclusively on first fixations. Interestingly, this restriction did not change the overall pattern of results (see Fig. 3 C and D), suggesting that even first fixations were influenced by object representations that emerged as a consequence of the observer's prior knowledge. Specifically, the statistical analysis showed that for first fixations, the similarity between Template and After was higher than for Template and Before (Template-After: M = 0.74, SD = 0.15; Template-Before: M = 0.62, SD = 0.17; $t(29) = 4.91$, $p < 0.001$; $M_{diff} = 0.12$, 95% CI = [0.07,

0.17]). This conclusion was corroborated by an ROI analysis of first fixations: similar to the results obtained when all fixations analysed, the percentage of first fixations landing on ROIs was higher in After than in Before, and also higher in Template than in After (Before: M = 34%, SD = 34; After: M = 40%, SD = 35; Template: M = 60%, SD = 32; Before-After: $t(29) = 3.61$, $p = 0.001$; $M_{diff} = 0.06$, 95% CI = [0.09, 0.03]; Template-After: $t(29) = 6.41$, $p < 0.001$; $M_{diff} = 0.2$, 95% CI = [0.27, 0.14]). Taken together, these results provide evidence to suggest that knowledge-dependent object representations emerge fast enough to influence even the first eye-movements after stimulus onset.

Knowledge-dependent object representations and image features act in synergy

I demonstrated that object-representations formed at the basis of prior knowledge were affecting gaze control. This demonstration, however, did not reveal much about the role image features may play in this process. To shed some light on this, I compared the distinct contributions of features and objects to eye-movements guidance. I relied on the fact that the features of the two-tone images remained unchanged while I manipulated the presence of representation of objects. Additionally, I adopted three simplifying assumptions. Firstly, I assumed that eye-movements in the Before condition, when the representations of objects were not yet formed, were mainly driven by the visual features of the stimuli. Secondly, I assumed that in the Template condition, object representations played a dominant role in determining fixation locations. Thirdly, I assumed that in the After condition, both factors contributed to the oculomotor control: stimuli features were the same as in the Before condition but objects representations were identical to the ones from Template condition.

Resting on these three assumptions, for each image I generated a series of new heatmaps. Each new heatmap was a linear combination of the heatmaps from the Before condition and the Template condition, using the formula: $w_{Template}$*$heatmap_{Template}$ + w_{Before}*$heatmap_{Before}$, where w is a weight for the heatmap indicated by the subscript. Incorporating the normalization assumption ($w_{Template} + w_{Before} = 1$), I created a continuum of heatmaps spanning between the two extremes of being fully determined by the Template heatmap to being fully determined by the Before heatmap. This continuum was uniformly sampled with a step-size of

0.05, with each sample being a distinct heatmap. I assessed the similarity of each of these new heatmaps to the heatmap from the After condition. Note that the heatmaps resulting from linear combinations always incorporated all fixations. To capture potential temporal changes in the balance of 'objects' and 'features', I compared these combined heatmaps to the heatmaps of only the first fixations, or of all fixations from the After condition.

The pattern of results suggests that fixations are guided synergistically by two factors, namely, the image-computable features of the two-tone image and the object representation resulting from the interaction between features and prior knowledge (see Fig. 4). This observation was corroborated by a statistical analysis. The optimal linear combinations for first fixations ($w_{Template}$ = 0.4; w_{Before} = 0.6) were significantly more similar to heatmaps from the After conditions than either the Template or the Before conditions (Optimal-After vs. Before-After: $t(29) = -2.67$, $p = 0.012$; $M_{diff} = 0.03$, 95% CI = [0.04, 0.01]; Optimal-After vs. Template-After: $t(29) = 5.70$, $p < 0.001$; $M_{diff} = 0.11$, 95% CI = [0.15, 0.07]). For the linear combination for the remaining fixations ($w_{Template}$ = 0.65; w_{Before} = 35), the same pattern of results was observed (Optimal-After vs. Before-After: $t(29) = 6.49$, $p < 0.001$; $M_{diff} = 0.09$, 95% CI = [0.12, 0.06]; Optimal-After vs. Template-After: $t(29) = 5.48$, $p < 0.001$; $M_{diff} = 0.05$, 95% CI = [0.06, 0.03]).

This finding suggests that the two-tone features and object representation worked in a synergistic manner when guiding eye movements. The contribution of these two factors varied over viewing-time with features playing a larger role in first fixations than for later fixations, for which object representations were dominant.

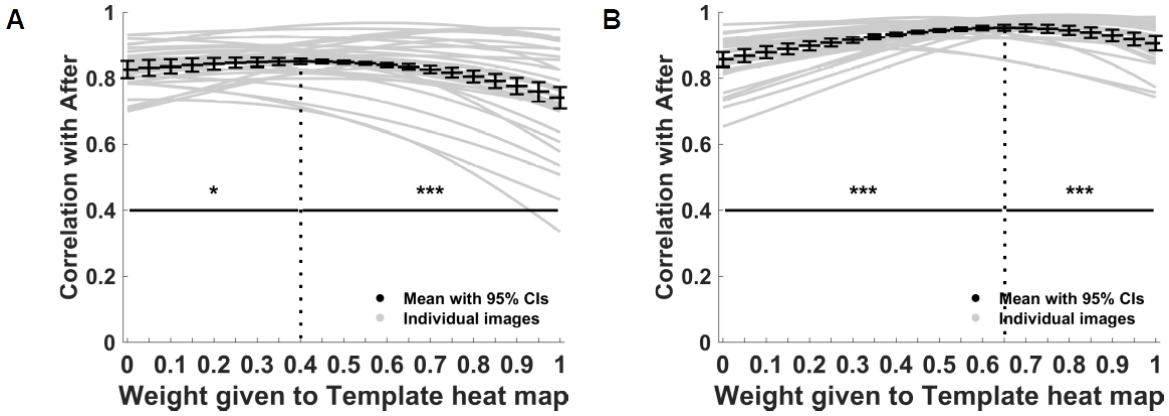

Figure 4. *Similarities of heatmaps from the After condition to different linear-combinations of heatmaps from the Template and Before conditions. A) Similarities obtained when only first fixations from the Before condition are considered. B) The same analysis but for all the remaining fixations (i.e., without first) from After condition. The weights of the linear combinations for which the similarity is maximal (indicated by the dotted vertical lines) are shifted more strongly to the left for first fixations than for the remaining fixations, demonstrating that the influence of the features of the two-tone image is stronger at the beginning of viewing compared to later on.*

Effects of knowledge-dependent object representations are observable despite the high similarities of gaze-patterns between After and Before conditions

The analyses reported in the preceding section revealed that the average similarity between heatmaps from Before and After conditions amounted to 0.83 (SD = 0.06) for first fixations and to 0.86 (SD = 0.08) for the remaining ones. These values can be interpreted as indicating a rather high similarity between these conditions which, at first glance, might seem difficult to reconcile with the idea that knowledge-dependent object representations are an important factor in determining the distribution of fixations in the After but not in the Before condition. However, interpreting the aboslute correlation values obtained when measuring similarity of fixations patterns in these two conditions requires providing a meaningful context. For this purpose, I randomly split my sample of 36 observers into two equally large groups, and compared the similarity of fixation patterns in the Before-After pair to the Before-Before and After-After pairs. If object representations affect oculomotor control, then the similarity between fixation patterns for the Before-After pairs should be lower than the similarity for the Before-Before and the After-After pairs. The results of my analysis confirmed these expectations. Specifically, the similarity between heatmaps from the Before-After pairs was lower than the similarities from the Before-Before and the After-After pairs (Before-Before: M = 0.94, SD = 0.02; Before-After: M = 0.84, SD = 0.07; After-After: M = 0.95, SD = 0.02; Before-Before vs. Before-After: $t(29) = 7.76$, $p < 0.001$; $M_{diff} = 0.1$, 95% CI = [0.12, 0.07]; After-After vs. Before-After: $t(29) = 8.43$, $p < 0.001$; $M_{diff} = 0.11$, 95% CI = [0.13, 0.08]). This finding provides further evidence for the influence of knowledge-dependent object representations in oculomotor control. Note that while I refer to this analysis as a between-groups analysis, I relied on paired t-tests. The reason for this choice of test is the fact that the unit of analysis still

is a single image. In order to warrant that the outcome of this analysis did not depend on a specific composition of the two groups, I repeated the split 20 times, each time assigning observers to the groups randomly. For each spilt, I obtained the same patterns of results which indicates the robustness of the effects in question.

Knowledge-dependent object representations affect multiple characteristics of oculomotor behaviour

A previous study investigating eye movements of observers viewing unresolved and resolved two-tones found that observers made fewer fixations when viewing resolved images, while these fixations were longer and landed closer to each other (M. Król & Król, 2019; see also Loth et al., 2010). To check if the same pattern is present in my data, I compared my experimental conditions with respect to the three characteristics of oculomotor behaviour analysed in the aforementioned study: number of fixations, average fixation duration (in seconds), and average euclidean distance between consecutive fixations (interfixation distance, in degrees of visual angle). All these characteristics were calculated per-image and their average values were compared between conditions (see Fig. 5). In the Before-After comparison, for the After conditions I found the decrease in the number of fixations (Before: M = 281.37, SD = 13.22; After: M = 240.1, SD = 19.32; $t(29) = 12.76$, $p < 0.001$; $M_{diff} = 41.27$, 95% CI = [34.65, 47.88]), increase in the fixation duration Before: M = 0.28, SD = 0.01; After: M = 0.3, SD = 0.02; $t(29) = -8.22$, $p < 0.001$; $M_{diff} = -0.02$, 95% CI = [-0.03, -0.02], and decrease in interfixation distance (Before: M = 4.09, SD = 0.45; After: M = 3.34, SD = 0.55; $t(29) = 11.24$, $p < 0.001$; $M_{diff} = 0.75$, 95% CI = [0.61, 0.89]). In the After-Template comparison, I did not find statistically significant effects for any of the three characteristics I analysed (number of fixations: $t(29) = -0.50$, $p = 0.621$; $M_{diff} = -2.67$, 95% CI = [-13.58, 8.25]; fixation duration: $t(29) = -0.24$, $p = 0.816$; $M_{diff} = 0$, 95% CI = [-0.01, 0.01]; interfixation distance: $t(29) = 0.32$, $p = 0.755$; $M_{diff} = 0.04$, 95% CI = [-0.19, 0.27]; descriptive statistics for the these three respective characteristics for Template condition: M = 242.77, SD = 31.76; M = 0.3, SD = 0.03; M = 3.3, SD = 0.96).

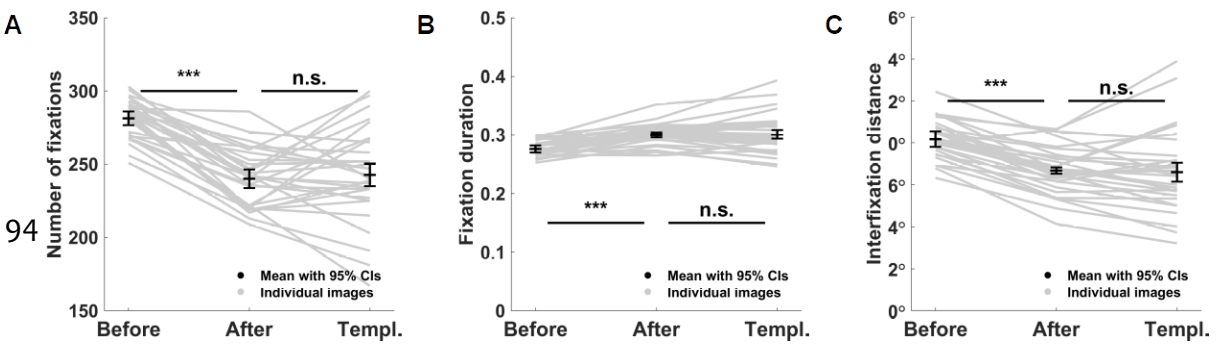

Fig. 5 Number of fixations (A), fixation duration (B), and interfixation distance measured in degrees of a visual angle (C). All three were calculated per image and compared between conditions.

In summary, my results replicate findings reported by Król and Król (2019). These authors interpreted such pattern of results to be in line with their "economies of experience" approach (M. E. M. Król & Król, 2018). According to this approach, the deployment of prior knowledge results in the optimization of different aspects of stimulus processing, including the optimization of the eye-movements which results in sampling information primarily from image regions containing objects. This interpretation is consistent with the idea that while observers in the Before condition explore the whole stimulus, in the After condition, they rather extract information ('exploit') only from selected parts of it. To acquire further evidence for this claim, I resorted to entropy – an abstract construct derived from physics and associated with the amount of 'disorder' in a given system. Entropy calculated for a heatmap indexes the extent to which observer's behaviour can be described as exploratory (Gameiro, Kaspar, König, Nordholt, & König, 2017; Kaspar et al., 2013), with higher values indicating more exploratory behaviour. Specifically, I estimated the normalized entropy of each heatmap (see Fig. 6A). Normalized entropy quantifies the spread of a given heatmap in a way which is indifferent to the heatmap's specific shape. This measure has values ranging from zero to one, with higher values indicating a larger spread (please refer to Appendix for details). As expected, entropy was the lowest in Template condition, increased in After condition, and was the highest in Unresolved (Before: $M = 0.56$, $SD = 0.05$; After: $M = 0.48$, $SD = 0.06$; Template: $M = 0.42$, $SD = 0.07$; Before-After: $t(29) = 10.70$, $p < 0.001$; $M_{diff} = 0.09$, 95% CI = [0.07, 0.1]; After-Template: $t(29) = 6.59$, $p < 0.001$; $M_{diff} = 0.06$, 95% CI = [0.04, 0.07]).

Taken together, the three gaze characteristics analysed thus far indicate that in After condition (that is, after acquiring prior object-knowledge), observers viewing two-tones fixated only selected image regions, instead of exploring the whole stimulus like in the Before condition. In the Template condition, this pattern was even more evident. Following up on that finding, I hypothesised that in the Before condition each observer exhibited more idiosyncratic behaviour than in the After condition because the only factor guiding gaze common for all observers were the image features; in the After and Template condition, an additional such

factor was present, namely, object knowledge. Therefore, I expected that acquiring object-knowledge resulted in observers exhibiting a more homogenous gaze behaviour. To test that, I quantified between-observers consistency, that is, the extent to which different observers tend to fixate the same image-locations (see Fig. 6B). I measured it using a method used previously in the literature (Lyu et al., 2020): by averaging the values obtained after calculating for each observer how similar the heatmap created from their fixations was to the heatmap created from the fixations of all the remaining observers. Comparing consistency revealed that it increased both between Before and After conditions and between After and Template (Before: M = 0.66, SD = 0.05; After: M = 0.7, SD = 0.05; Template: M = 0.76, SD = 0.05; Before-After: $t(29) = 3.96$, $p < 0.001$; $M_{diff} = 0.04$, 95% CI = [0.06, 0.02]; After-Template $t(29) = 6.96$, $p < 0.001$; $M_{diff} = 0.06$, 95% CI = [0.07, 0.04]), thus confirming my predictions. Summarising, the analysis of the different gaze characteristics, in line with the ROI-based analyses, indicate that when object representations can be formed (that is, in Template and After conditions), observers primarily attend to image locations containing objects and the gaze patterns of different observers become more similar to each other.

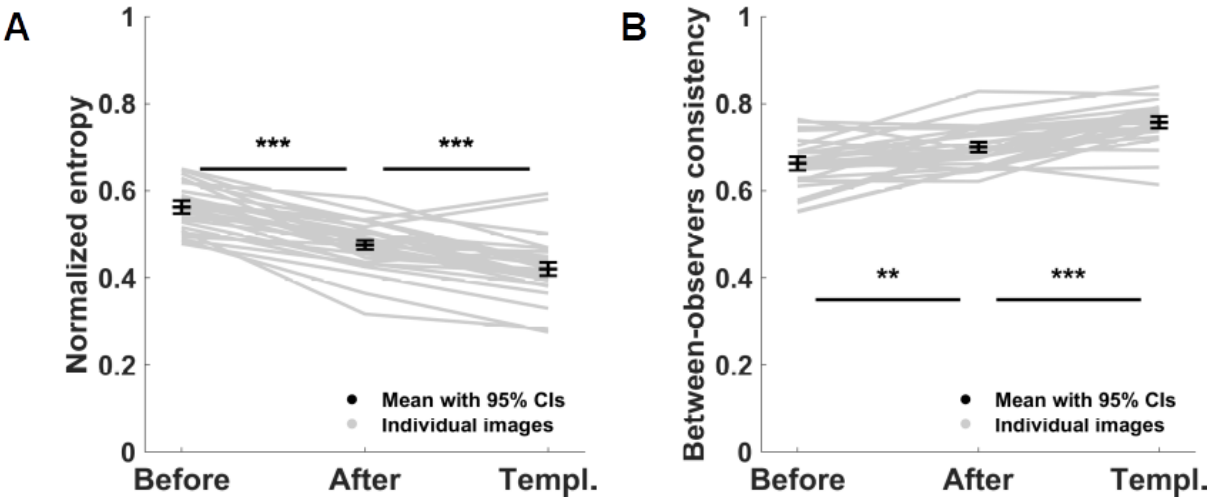

Fig. 6 Normalized entropy and between-observers consistency.

A) Normalized entropy of fixation distributions (in arbitrary units), measuring their spread. Higher values indicate more exploratory behaviour of observers. B) Between-observers consistency in selecting fixation targets measured by how similar (on average) fixations of a single observer were to fixations of all the remaining observers pooled together.

Experiment 1 – Discussion

In Experiment 1, I measured eye-movements in response to two-tone images. These stimuli are derived by manipulating so-called template images – grayscale photographs of objects – in such a way that, on initial viewing, two-tones are experienced as (relatively) meaningless black-and-white patches. Once the observer has acquired relevant prior object-knowledge, however, perceptual organization processes in the visual system bind the patches into a coherent percept of an object. I demonstrate that, when a two-tone image is perceived as showing a coherent object rather than meaningless patches, fixation patterns are more similar to those measured in response to the original template. Moreover, between groups of observers, the same two-tone images lead to a lower similarity in fixation patterns if one group perceives the stimuli as meaningless patches and the other as coherent objects, compared to when the presence of object representations is consistent across groups. Overall, these results suggest that object representations that are not fully determined by the image-computable aspects of the stimulus but depend on the observer's prior object-knowledge have an important influence on eye movements.

As a potential alternative explanation for the results of Experiment 1, it could be argued that the observed change in fixation patterns was caused by a memory process unrelated to perceptual organization. Specifically, it has been suggested that eye movements preformed during memory retrieval of an image resemble the eye movements performed when seeing this stimulus for the first time (Noton & Stark, 1971; see Wynn et al., 2019 for a recent review). Therefore, it is possible that viewing of two-tone images in the After condition acted as a cue that triggered the retrieval of the corresponding template, and that this retrieval was accompanied by the re-enactment of gaze behaviour from the Template condition. A closely related alternative explanation of my results from Experiment 1 is that memory-retrieval of template images resulted in the observers voluntarily directing their gaze towards display locations they remembered from the Template condition to be occupied by the objects. According to both of these explanations, the factor driving the changes in eye movements in the After condition was the objects-to-locations mapping remembered by the observers from the Template condition, rather than the perceptual organization induced by prior object-knowledge. Specifically, the observers might fixate display-locations overlapping with objects in the templates because of processes related to retrieving templates from memory, and not

because of their visual system organised the two-tone images into meaningful scenes based on prior knowledge. To exclude these alternative interpretations, I conducted Experiment 2.

Experiment 2

Experiment 2 was identical to Experiment 1 in all aspect except that the template images were mirror-flipped along the vertical axis (in the resolving phase, when the blending of two-tones and templates was presented, both images were mirror-flipped). Consequently, the screen locations occupied by objects differed between the Template condition and the remaining conditions. If, when viewing two-tones in the After condition, observers merely revisited the parts of the display, which contained objects during the presentation of template images, I would expect a high similarity between heatmaps from the After and Template conditions, despite the lack of overlap in spatial location of objects in these two conditions. If, however, the effects observed in Experiment 1 were attributable to knowledge-dependent object representations, I would expect the similarity between After and Template to be low. Moreover, the similarity should increase when the heatmaps from the mirror-flipped templates would be flipped back to align the object locations between template and two-tone images.

Experiment 2 – Method

The design of Experiment 2 was identical to that of Experiment 1 except that the template images were presented mirror-flipped from left to right during all parts of the experiment (i.e., during instructions, the blending phase, and the Template condition). This condition is labelled Flipped Template. A separate set of 18 Cardiff University students (mean age 19.5 years, 5 males), who had not participate in Experiment 1, served as observers.

Experiment 2 – Results and Discussion

Memory-retrieval of objects-to-locations mapping does not explain changes in eye movements

Similar to Experiment 1, the meaningfulness ratings provided by the observers after viewing each two-tone were higher in the After condition than the Before condition both when I

averaged them per observer (t(17) = 6.62, p < 0.001; M_{diff} = 0.24, 95% CI = [0.31, 0.16]) and per image (t(29) = -16.74, p < 0.001; M_{diff} = -0.24, 95% CI = [-0.27, -0.21]). This result indicates that observers were able to bind the two-tone images into meaningful percepts despite viewing mirror-flipped templates. The analysis of the eye-movements data was inconsistent with the objects-to-locations hypothesis but provided support for the idea that knowledge-dependent object representations influence eye movements (see Fig. **7**). In particular, by contrast to the analogous analysis in Experiment 1, heatmap similarities did not differ when comparing the Flipped Template-Before pair vs. the Flipped Template-After pair (Flipped Template-Before: M = 0.46, SD = 0.22; Flipped Template-After: M = 0.48, SD = 0.22; t(29) = 1.45, p = 0.158; M_{diff} = 0.03, 95% CI = [-0.01, 0.06]). A Bayes factor (BF) of 0.5 suggested that this difference is rather unlikely to exist. Moreover, the similarity in heatmaps between Flipped Template and After increased after flipping-back the heatmaps of the Flipped Template condition, that is, spatially aligning the two-tone and templates (Template-Before: M = 0.68, SD = 0.15; Template-After M = 0.8, SD = 0.11; t(29) = 7.77, p < 0.001; M_{diff} = 0.13, 95% CI = [0.09, 0.16]).

Summarising, the results of Experiment 2 suggest that when observers view two-tones After that they experience as containing meaningful objects (After condition), their eye movements are guided by knowledge-dependent object representations. This finding thus excludes an interpretation of my results in terms of an objects-to-locations mapping, where retrieval of information regarding the locations occupied by objects in the Template condition determines fixations.

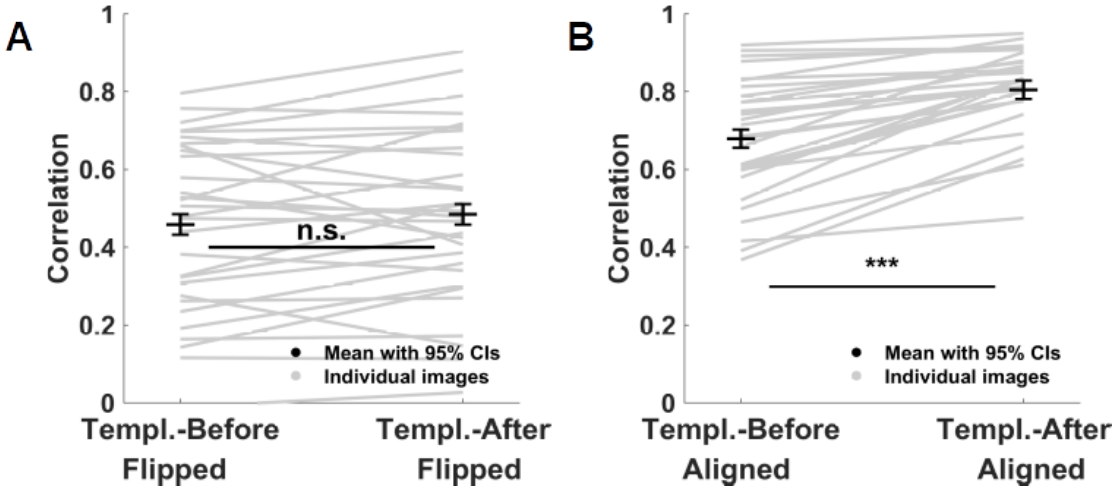

Fig. 7. Results of Experiment 2.

The plot shows the similarities between heatmaps from the resolved two-tones to heatmaps from the mirror-flipped templates before (panel A) and after (panel B) 'flipping them back' to realign the locations of objects. The increase in similarities observed on panel B (but not present on A) indicates that in the After condition, observer were not merely revisiting display-parts they remembered as containing objects – in such case, the flipping-back would result in the decrease in the similarities. Instead, the fixations of observers were directed at objects on the two-tones.

In a final experiment, I addressed two further alternative explanations. First, it is possible that during the blending phase, observers learn to associate specific image-features in the two-tone images with the objects from the templates. When viewing two-tone images in the After condition, these feature-object associations might guide fixations towards these specific visual patterns, irrespective of transformations such as those introduced by flipping (for instance, changes in the relative location on screen, or mirroring). While this possibility might seem not seem intuitively plausible, there is evidence suggesting that associative learning processes are an important factor in oculomotor control (Alfandari, Belopolsky, & Olivers, 2019). A final alternative explanation of my results from both Experiment 1 and 2 relates to potential order effects. It is possible that the changes in fixation patterns between Before and After conditions resulted from viewing two-tones for a second time, rather than from knowledge-based perceptual organization. In other words, observers might sample information from different image regions on second compared to first viewing, irrespective of the kind of information they acquire in the meantime. Therefore, I conducted a final experiment to exclude the possibility that (i) feature-objects associations or, (ii) any order effects could explain the effects of Experiments 1 and 2.

Experiment 3

Experiment 3 adopted the same procedure as the previous experiments except that the templates from Experiment 1 ('original templates') were replaced with different images that were unrelated to the two-tones ('dummy templates'). This experimental design allowed us to test whether feature-object associations provide a plausible explanation for the findings of Experiment 1 and 2. Specifically, observers might associate certain features in the two-tone

images with objects in the templates during the resolving phase. When viewing two-tone images in the Resolving condition, these feature-object associations drive fixations towards image locations in the two-tones that overlap with objects in the respective templates. Importantly, these effects should be observable despite the fact that observers did not acquire the prior-knowledge required to bind the patches of the two-tones into a coherent object percept. The design also allowed us to address the issue of order effects, because, despite the modification in the design, each two-tone was still viewed by observers twice.

Experiment 3 – Method

Experiment 3 was completed by 20 observers (mean age 19.55, 5 males) who did not participate in the previous two experiments. Again, they were Cardiff University students.

The procedure closely resembled the one from both previous experiments. The only difference was that in each block, in the Template condition and in the resolving phase, instead of the templates from which the two-tones presented in this block were derived ('original templates', like in Experiments 1 and 2), different images – 'dummy templates' – were presented. Each two-tone had a unique dummy template paired with it and this pairing was fixed throughout the experiment (that is, for all observers). Crucially, for each two-tone, a dummy template paired with it was an original template of some other two-tone, also presented in the experiment. The assignment of stimuli to experimental blocks was pseudo-randomized for each observer individually. Crucially, the pseudo randomization was always done in a way that guaranteed that dummy templates presented in any given block were the original templates of two-tones presented in the preceding block. This was to ensure that the images belonging to any given triplet of a dummy template, a two-tone paired with this dummy template, and the original template of this two-tone were always viewed by the same observers. For the same reason, I undertook two further steps. First, I always discarded fixations registered on the two-tones presented in the last experimental block, because the original templates of these two-tones were never presented to a given observer. Second, in the first block, which, obviously, was not preceded by any other block, the dummy templates were always the same (they were not related to any of the two-tones and fixations on them were discarded); only the two-tones were different in each run of the procedure. With 20 observers in total, after taking into the

account discarding fixations in the first and last block, for each two-tone (viewed in Before and After condition) and its dummy and original templates, I retained fixations from 18 observers.

Experiment 3 – Results

Lack of relevant object-knowledge prevents the emergence of knowledge-dependent object representations

The analysis of meaningfulness ratings demonstrated that, as expected, observers were not able to bind the two-tone images into coherent object percepts even in the After condition (see Fig. 8A and B). In particular, the differences in ratings between Before and After conditions were not statistically significant, both when the data were averaged per observer ($t(19) = 1.49$, $p = 0.152$; $M_{diff} = 0.02$, 95% CI = [-0.01, 0.06]) or per image ($t(29) = 1.97$, $p = 0.058$; $M_{diff} = 0.02$, 95% CI = [0, 0.05]). In the former case, Bayes factor analysis suggested weak evidence for the lack of differences (BF = 0.6), while in the latter no clear conclusions could be drawn (BF = 1.07). Taken together, these results suggest that when the two-tones were viewed for a second time, they were as meaningless as when they were seen for the first time.

Memory-retrieval of feature-object associations might lead to small changes in eye movements but cannot explain key findings of Experiments 1 and 2

Experiment 3 was designed to test the idea that the effects observed in the two previous experiments might be explainable by an association between features in two-tones and object locations on templates. The analysis of eye-movement data revealed that there was evidence to suggest that, indeed, such an association might take place and might guide oculomotor control to a limited extent. Specifically, the similarity in heatmaps in the Dummy Template-After pair was higher compared to the Dummy Template-Before pair Before (see Fig. 8C). This increase in similarity, although significant in a statistical sense, was small (Template-Before: $M = 0.46$, $SD = 0.21$; Template-After: $M = 0.52$, $SD = 0.22$; $t(29) = 4.70$, $p < 0.001$; $M_{diff} = 0.06$, 95% CI = [0.03, 0.08]). I resorted to equivalence tests (Lakens, Scheel, & Isager, 2018) to assess whether the effect was of a comparable size as the analogous increase in similarities observed in Experiment 1 (between the Template-Before and Template-After pairs). In essence, these tests evaluate whether the size of an observed effect falls in a specified range of effect sizes, which

are judged by the researcher to be too small as to be of interest (and are 'practically' equivalent to zero, hence the name). Here, I relied on the "two one-sided tests" (TOST) procedure (Shuirmann, 1987) implemented in R package TOSTER (Lakens, 2017; Lakens et al., 2018). This procedure requires specification of an upper and a lower bound of the range of effect sizes, and tests two null hypotheses: that the observed effect is smaller than the lower bound and greater than the upper bound. When both are statistically rejected, it can be concluded that the observed effect falls within equivalence-region defined by the bounds and, therefore, is equivalent to zero. I used TOST to determine whether the increase in similarities observed in Experiment 3, when compared with the analogous increase from Experiment 1, was statistically equivalent to the lack of an effect. Specifically, I tested if the magnitude of the increase in similarities from Experiment 3 fell within the equivalence region with an upper bound determined by the lower bound of the 95% confidence interval of the mean increase in similarities in Experiment 1 (equal to 0.14). The lower bound – irrelevant here – was set to – 0.14 to make the equivalence region symmetrical around zero. Note that here, I were not using any standardised effect sizes, but operated on raw values, derived directly from the data. With these upper and lower bounds, the effect from Experiment 3 was statistically equivalent to zero according to the equivalence tests ($t(29) = 6.685$, $p < 0.001$). This result indicates that the processes responsible for changing gaze-patterns between the Before and After conditions in Experiment 3 most likely could not be responsible for the analogous changes in Experiment 1.

The key findings of Experiment 1 and 2 cannot be attributed to order effects

The second reason for conducting Experiment 3 was to exclude the possibility that order effects could explain the key findings of Experiment 1 and 2. In particular, I wanted to address the concern that viewing the same two-tones for a second time changed fixation patterns in such a way so that they started to resemble the patterns from the (original) templates. Recall that in each block of Experiment 3, the images used as dummy templates were the original templates of the two-tones presented in a previous block. This design allowed us to record, in the same observers, fixation patterns for the original templates as well as for two-tone images viewed twice without prior object-knowledge (Before and After conditions, respectively). If the findings in the previous experiments resulted from the fact that two-tones were viewed for the

second time, I would expect that the fixation patterns in the After condition were more similar to the patterns recorded in response to the original template compared to the similarity between Before and the original template. Importantly, these changes in fixation patterns would occur despite the fact that observers did not acquire any relevant object knowledge between the two two-tone image conditions.

The results are inconsistent with this 'second-viewing' hypothesis (see Fig. 8D). I calculated heatmaps similarities between the original templates and the corresponding two-tones viewed in Before and After conditions and found that these similarities did not differ significantly in a statistical sense (Template-Before M = 0.64, SD = 0.15; Template-After M = 0.64, SD = 0.14; $t(29) = 0.22$, $p = 0.83$; $M_{diff} = 0$, 95% CI = [-0.03, 0.03]). Moreover, a Bayes factor analysis provided direct evidence to support a lack of a difference (BF = 0.2).

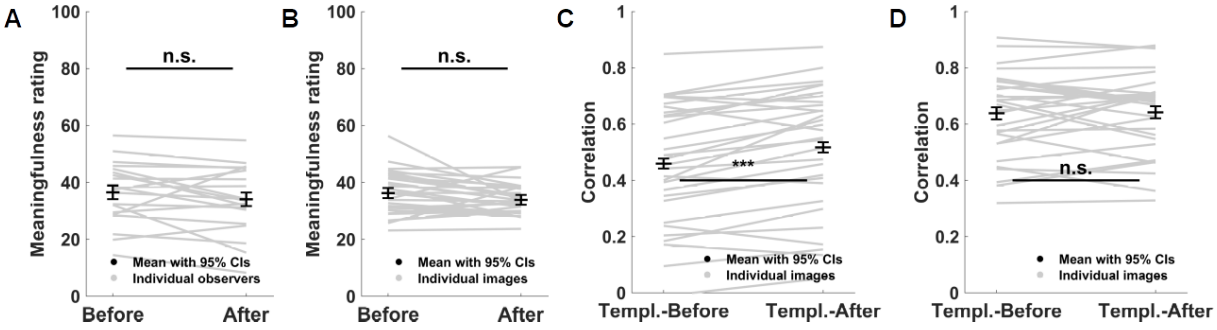

Figure 8. Results of Experiment 3.

A, B) Meaningfulness ratings averaged per observer (A) and per image (B). C) Comparison of heatmap similarities between two-tones (viewed in Before and After conditions) and their dummy templates, that is, unrelated images. D) Comparison of heatmap-similarities between two-tones (viewed in Before and After conditions) and their original templates.

Discussion

Control of eye movements is typically considered within a dichotomy between bottom-up processing of low-level features and top-down control via factors such as an observer's high-level object representations (see Chapter One). In the current Chapter, I abandoned this

simplifying framework in light of emerging evidence highlighting the complex and intricate relationship between features and high-level object representations. I recorded eye movements in response to so-called two-tone images – clusters of black and white patches derived from images of natural scenes called templates. When viewed by naive observers, two-tone images appear as meaningless patches. After the observer is exposed to the template and thereby acquires relevant object-knowledge, these patches are bound into coherent and meaningful percepts of objects. As a result, two-tone images provide a means to study the intricate interplay between prior object-knowledge and image-computable features in bringing about object representations. In this study, observers viewed two-tones images twice: before (Before condition) and after (After condition) viewing the templates (Template condition). In all three conditions, their eye movements were recorded. Across three experiments and on a number of different metrics, fixation patterns on the two-tone images differed substantially depending on whether observers were able to bind images into meaningful percepts of objects. In particular, fixations patterns were more similar to the patterns on templates, more focussed on pre-specified regions of interest, less dispersed, and more consistent across observers when the same two-tone images were organised into object percepts compared to when they were not. Importantly, these effects were evident from the first moments of image viewing. My results contribute to the mounting body of evidence that knowledge-driven perceptual organisation of visual features into object representations fundamentally alters processing of these feature (González-García et al., 2018; Ongchoco & Scholl, 2019; Teufel et al., 2018). Specifically, I demonstrate that the emergence of objecthood determines which features are selected for further inspection by means of fixations.

The typical approach to distinguish between the two factors of the dichotomous framework outlined above is (i) to compute a saliency maps based on certain features of images used in the experiment, (ii) to generate a map of semantically important regions or object locations in these images, and (iii) to assess which of the two maps better predicts human fixations (for example, see Henderson, 2017; see also Pilarczyk & Kuniecki, 2014 and; Rider et al., 2018). To the extent to which one of the two different types of maps better explains human fixations, the respective factor is considered to be critical to guide eye movements. This approach led to important insights regarding oculomotor control. However, it is hampered by its dependence on specific operationalisations of visual features and high-level factors, up to the point when

different operationalisation have led to qualitatively different conclusions. For example, Einhäuser and colleagues (Einhäuser, Spain, & Perona, 2008) compared the extent to which two maps were able to predict human fixations: a saliency map created using the model developed by Itti and Koch (Itti & Koch, 2000), and a map based on manually labelled object regions. Given the superiority of the object map, the authors' initial conclusion was that object recognition, rather than low-level features, drive eye movements. This conclusion was challenged by a re-analysis of the data with other saliency models, including the AWS model (Garcia-Diaz, Fdez-Vidal, Pardo, & Dosil, 2012), which outperformed the high-level object model in predicting human performance (Borji et al., 2013; see also Einhauser, 2013). However, it turned out that a more sophisticated representation of objects can outperform even AWS (Stoll et al., 2015;), it turned out that a more sophisticated representation of objects can outperform even AWS see also Nuthmann et al., 2020).

Since publication of these studies, models that outperform AWS by a large margin have been developed (Kümmerer, Wallis, Gatys, & Bethge, 2017a; Thomas, 2016). They no longer rely on local image conspicuity, but harness high-level visual features extracted from deep neural networks trained for object recognition. As described in detail in Chapters Two and Three, these models played a critical role in the latest instalment of the debate about low-level vs. high-level factors in oculomotor control. This dispute had a very similar structure to previous discussions. It was initially ignited by the first meaning maps study claiming that eye-movements are driven by meaning rather than image features because meaning maps were better at predicting fixations than one specific saliency model – the GBVS model (Harel et al., 2007). However, when my subsequent work compared meaning maps to more advanced saliency models like DeepGaze II (Kümmerer et al., 2017b), they perform less well. Therefore the conclusions that can be drawn from this approach depend on the choice of model. Consequently, virtually every result that has been obtained with this approach can be called into question: it is always logically possible that the conclusions would have been different, had different operationalisations been used.

A further limitation of this approach stems from the fact that high-level factors like object representations supervene on visual features: in other words, the two are (often but not always) functionally and spatially correlated (Elazary & Itti, 2008; Masciocchi et al., 2009). To

illustrate, consider the following two hypotheses: 'certain clusters of visual features attract fixations (in and of themselves)' and 'certain clusters of visual features attract fixations only if they are bound into the representation of object'. For typical natural scenes, both hypotheses generate identical predictions regarding which regions would be fixated. This abstract illustration is reflected in empirical findings: objects which are interesting for human observers and the clusters of visual features indexed as salient by saliency models tend to occupy the same image locations (Elazary & Itti, 2008; Masciocchi et al., 2009).

The approach used in my study circumvents the difficulties of this conventional approach. Specifically, using two-tone images does not involve disentangling and quantifying the contributions of bottom-up and top-down factors post-hoc, after recording the eye-movements. Rather, I manipulated the presence of the representations of objects directly, and without changing the visual features of images.

The results of Experiment 1 and 2 revealed that the similarity between fixation patterns registered on tow-tones and templates were higher when the two-tones were viewed in the After condition, as compared to the Before condition. I found this effect both in the general assessment of similarity between fixations distributions and in a more specific analysis focusing on fixations within the regions-of-interest. The fact that the manipulation altered the distributions of fixations provides strong support for the hypothesis that mental representations of objects contribute to the process of selecting fixation targets in images. Crucially, formation of these representations is dependent on possessing prior object-knowledge, image-computable visual features are insufficient for this formation to occur. Therefore, my results not only rule out the possibility that human oculomotor control relies solely on visual features but also highlight the fact that the emergence of objecthood requires an interaction between prior-knowledge and the visual input.

Fixation locations, however, are not the only aspect of oculomotor control influenced by object representations. I found that binding features into objects changes also how the visual system approaches the exploration-exploitation dilemma (Ehinger, Kaufhold, & König, 2018; Gameiro et al., 2017; Hills et al., 2015). During every fixation, the visual system must constantly decide whether to keep the eyes still – and be able to further inspect the currently fixated scene

region (exploit it) – or to perform a saccade, and thereby begin inspecting another region (explore the scene). In the context of this dilemma, the fact that in the After condition (as compared to Before) I observed a decrease in interfixation distance and the increase in fixation duration marks the shift from exploration to exploitation (Gameiro et al., 2017). This shift, together with the increase in the amount of fixations landing on objects (also occurring in the After condition, as revealed by the analysis of spatial distributions of fixations), suggests that after resolving a two-tone, the visual system prioritizes objects: it 'exploits them', while giving up exploring the remaining parts of the image. Therefore, clusters of features that provide support for object representations become interesting for the system (for a similar finding, see Król & Król, 2019).

The speed with which object representations acted to influence eye-movement is a very striking finding: already the first fixations after image onset were affected by these representations. This finding suggests that knowledge-dependent object representations, or at least a coarse form of them, emerge very quickly based solely on the information available at the beginning of image viewing, when the eyes are stationary. At this moment, only a limited part of the image can be inspected with the high-resolution foveal vision, and access to the remaining part is possible only via peripheral vision. Information from peripherial vision therefore has to contribute to the emergence of the initial representation of object. Peripheral vision mainly provides access to low spatial frequencies. There is a body of evidence suggesting that LSF play a facilitating role in segmenting the incoming visual input into objects (Bar et al., 2006; Bar, 2004; Bullier, 2001). Specifically, rapidly extracted LSF representations may serve as a basis for narrowing down the search space of possible hypotheses about object identities in the input, and thereby scaffolding more precise object identification. In light of these studies, it is tempting to speculate that in my experiment first fixations were guided by coarse LSF representations while later fixations might be guided by fuller object representations.

This idea rests on the assumption that two-tone images provide enough information to peripheral vision to form LSF representations. There are several reasons to believe that this assumption holds. First, recall that the process of creating two-tones involves blurring and binarisation of templates. These operations drastically change the high spatial frequency content of the images but have less impact on low spatial frequencies. Second, visual

inspection of fixation distributions from my study indicates that in the Before condition, image regions containing torsos (but not heads) of animals were fixated more frequently than regions belonging to the backgrounds. This observation suggests that converting templates into two-tones, despite being effective in concealing the identities of depicted objects (as indicated by the meaningfulness ratings), could not completely conceal their locations. Crucially, the information about the animal torsos is (most likely) carried mainly by the LSF, which suggests that they are to large extend preserved in the two-tones.

Another intriguing pattern of results was evident in almost all between-condition comparisons: compared to the Before condition, gaze patterns registered in the After condition were more similar to those in the Template condition but there was still a substantial difference in eye-movements between After and Template conditions. One possible explanation for this phenomenon might be incomplete perceptual organization in some trials. According to this idea, as long as visual features give rise to the same object representation, these representations guide eye movements towards the same locations. Therefore, different eye-movement patterns in the After and Template conditions are due to differences in the object representations. This view on oculomotor control has been expressed in the cognitive relevance theory (Henderson et al., 2009), which proposes that visual features do not contribute to oculomotor control directly but only create a 'flat landscape' on top of which top down-factors operate and determine fixations locations.

Another possibility is that the differences between After and Template conditions are not merely a 'bug' resulting from the variability in strength of perceptual organization experienced by the observer, but rather a manifestation of an inherent characteristic of the oculomotor system. Specifically, the visual features, instead of being merely potential carriers of object representations might exert influence on eye movements even after being bound into the representations of objects. According to this hypothesis, I would expect eye movements to differ between After and Templates conditions because in each condition the visual features with which the same object representation interact are different.

I did not plan to directly test these alternative explanations of the differences between gaze patterns in the After and Template conditions. Yet, the results of my linear weighting analysis

are incompatible with the idea that these two conditions differ due to incompleteness of perceptual organization and provide support for a persisting influence of visual features even for fully organised percepts. Recall that the crux of this analysis was to compare the similarity of different linear combinations of heatmaps from the Before and Template conditions to heatmaps from the After condition. These linear combinations included varying proportions of both components and ranged from 100% Before + 0% Template to 0% Before + 100% Template. If object representations drive human eye-movements and features are only potential carriers for such representations along the lines of the 'flat landscape' idea proposed by the cognitive/behavioural relevance theory, then features of two-tones and templates are 'interchangeable': as long as they support the same object representation, eye-movements on two-tones and templates should be identical. In this case, I would expect a monotonic increase of similarity in the linear combination analysis. That is, with the decrease in the weight given to the Before component, the similarity of the linear combination to After heatmaps should increase. This prediction is, however, not supported by my data. Instead, the linear combinations peaked at a point, at which they included both components, thus demonstrating that both the heatmaps from the Before and the Template condition are critical in explaining the fixation patterns in the After condition. This finding suggests that even if the same or very similar object representations are experienced when viewing two-tones and templates, the fact that these representations are supported by different features matters for oculomotor control. Interestingly, the linear combination analysis indicated that influence of features is stronger at the beginning of image viewing. This effect might reflect the fact that perceptual organisation takes time to fully unveil. Note that the linear combination analysis was conducted on a per-image basis. The finding that features and object representations contribute to eye-movement control can therefore not be explained by averaging across different images, with some leading to purely feature-driven and other to purely representations-driven eye-movement control.

Another interesting aspect of my study is the relationship between object-knowledge driven perceptual organization, eye-movements, and memory. It is obvious that prior-knowledge acquired when viewing templates must first be stored in memory and then retrieved from it in After condition. Numerous studies demonstrated a tight link between the eye-movements and memory retrieval: gaze shifts made during recall of a specific stimulus resemble those made

during encoding of this stimulus (Noton & Stark, 1971; Wynn, Ryan, & Buchsbaum, 2020; Wynn et al., 2019). It is, therefore, conceivable that in Experiment 1, the two-tone images in the After condition served as a cue eliciting the retrieval of the corresponding template from memory, and that it is this retrieval – not the emergence of objecthood – which changed the eye-movements of observers. Even if such phenomenon indeed occurs, my Experiments 2 and 3 demonstrated that it is not able to account for all the effects I attribute to object representations. Specifically, Experiment 2 showed that the changes in fixation location observed in After condition in Experiment 1 cannot be explained by revisiting screen locations remembered as containing meaningful parts of the templates, which I labelled the objects-to-locations-mapping hypothesis in the main text). Experiment 3, in turn, demonstrated that when the perceptual organization is not taking place – due to the lack of object-knowledge – memory of where object presented on templates were and what visual features they corresponded to (objects-to-features mapping) were not sufficient to change the eye movements to the same degree as in Experiment 1, when the perceptual organization took place.

Summarising, in this Chapter I demonstrated that prior object-knowledge and the process of perceptual organisation driven by that knowledge play a crucial role in oculomotor control. First, object-knowledge largely determines which image locations are being sampled. Second, it influences the whole sampling strategy of the visual system and leads to the prioritization of extracting information from only a subset of image locations over exploring the entire image. Taken together, these findings provide evidence that eye movements control is based on the interaction between knowledge already stored in the visual system and the visual features of the input.

Appendix to Chapter Four

Data exclusions

Recall that in each experiment, each participant viewed each image in each condition for 3 seconds. Some of such viewing session were discarded from my analyses because of the low amount of data recorded throughout them – see Table S1. First, some viewing sessions were discarded because no fixations were registered throughout their duration (because, for example, observers did not move their gaze from the fixation point). Next, for each of the

remaining viewing sessions, I calculated the percentage of the eye-tracker data-samples in which the eye-position was not recorded (for example, due to blinking of the observers). After visually inspecting histograms of the obtained values, I decided to exclude from further analyses all viewing sessions for which more than 30% of the position data was missing. The steps described in this section were identical for all my experiments and their results are summarized in Table S1. To reiterate, each observer provided meaningfulness rating for each two-tone image twice: once in the After, and once in the Before condition. When analysing these ratings, whenever I encountered a rating provided after viewing session marked as excluded in one condition, I excluded both this rating, and the rating provided for the same image by the same observer in the other condition.

Table S1. Excluded viewing-session per experiment

Experiment Number	Number of viewing sessions with no fixations	Number of viewing sessions excluded because of the missing data	Total number of viewing sessions	Total percent of excluded viewing sessions
1	12	46	3240	1.79%
2	7	60	1620	4.14%
3	6	13	1800	1.33%

Normalized entropy calculation

Entropy calculated for a heatmap provides the measure of its spread. This measure, importantly, is not dependent on the map's specific shape. However, its values are dependent on the number of fixations used to create the heatmap (Gameiro et al., 2017; Wilming et al., 2011). Given that the heatmaps from my experiments differed with respect to total the number of fixations underlying them, I estimated entropy values by means of a bootstrapping procedure which accounts for these differences (Gameiro et al., 2017). Specifically, for a given image, I first randomly selected 50 fixations from the pool of all fixations registered on it, converted them into a heatmap, and calculated its entropy using a standard Matlab function

(entropy). This procedure was then repeated 50 times and the entropy values obtained in all the iterations were averaged.

The absolute value of heatmap's entropy depends also on the specific binning of a heatmap, i.e. the range of possible pixel values. Because I was interested only in the changes of entropy between conditions, rather than in the absolute values, I normalized – hence the term normalized entropy I use here – the values from the bootstrapping procedure so that they belonged to a range from zero to one. The normalization was performed by dividing them by the maximal entropy-value possible to obtain for a heatmap, given the size of my images and the binning. This theoretical maximal value was calculated as the entropy of a heatmap being a uniform random distribution.

Chapter Five – general discussion

Summary

In the present thesis, I investigated factors influencing human oculomotor behaviour during natural scene-viewing. The starting point for all of the questions addressed in my research was the bottom-up vs. top-down dichotomy used to characterise factors influencing human gaze and described in Chapter One. Chapters from Two to Four report experiments I conducted using eye-tracking, computational modelling, and crowd-sourced data collection methods. The first two of these Chapters are devoted to assessing the meaning maps approach: a theoretical and methodological stance according to which human eye movements are controlled primarily by one of the top-down factors – image meaning. In Chapter Four, I make an attempt to reshape the dominant way of thinking about oculomotor control and, instead of focusing on disentangling the contributions of both components of the aforementioned dichotomy, I focus on their interactions. The main findings reported in this thesis are summarised below (see also a bullet-point summary at the end of this Chapter).

Chapter Two

In this chapter, I evaluated the fundamental assumptions underpinning meaning maps (Henderson & Hayes, 2017, 2018), a tool designed to measure the distribution of semantic information in images. I demonstrated that these maps might be sensitive to complex image features, rather than semantic information. To create meaning maps, images are segmented into partially overlapping patches, which are rated for their meaningfulness by multiple individuals (raters). These ratings are combined into a smooth distribution over the image. Recently, meaning maps have been used to provide support for the claim that meaning – rather than image-computable features – guides human eye-movements (Henderson et al., 2019). If meaning maps capture the distribution of meaning, and if the deployment of eye-movements in humans is guided by meaning, two predictions arise: first, meaning maps should be better predictors of gaze position than saliency models, which use image features rather than meaning to predict fixations; second, differences in eye movements that result from changes in meaning should be reflected in equivalent differences in meaning maps. This

Chapter describes experiments testing these predictions. Their results showed that meaning maps performed better in predicting fixation locations than the simplest saliency model (GBVS; (Harel et al., 2006), were similar to a more advanced model (AWS; Garcia-Diaz et al., 2012) and were outperformed by DeepGaze II – a model using a deep neural network trained on object recognition to carve up feature space (Kümmerer et al., 2017). These data suggest that, similar to saliency models, meaning maps might not measure meaning but index the distribution of complex features. I tested this notion directly by comparing scenes containing consistent object-context relationships with identical images, in which one object was semantically inconsistent, thus changing its meaning (e.g., a kitchen with a mug swapped for a toilet roll). Replicating previous studies, regions containing inconsistencies attracted more fixations from observers than the same regions in consistent scenes. Crucially, however, meaning maps of the modified scenes did not attribute more 'meaning' to these regions. DeepGaze II exhibited the same insensitivity to meaning. I conclude that both methods are thus unable to capture changes in the deployment of eye-movements induced by changes of an image's meaning that are based on object-context relationships.

Chapter Three

This Chapter further evaluated the meaning maps approach and demonstrated that contextualized meaning maps (Peacock et al., 2019), the modification of the original meaning maps, are also not able to account for the effects of semantic inconsistencies on human eye movements. The basic rationale of the chapter is that the limitations of meaning maps, which I highlight in Chapter Two, may result from the fact that meaning maps do not provide the raters with the opportunity to consider the influence of context on the meaningfulness of a patch – recall that the patches are being shown to them in isolation, without the context scene from which they were derived. This limitation of meaning maps had been anticipated by the authors of this method (Henderson et al., 2018) and, after publishing its initial version, they proposed a modification of the original method: contextualised meaning maps (Peacock et al., 2019). These maps carry the potential to overcome the limitation of their predecessors because they are constructed from ratings provided by raters who know the context scene of each patch and, therefore, can take context information into account in their meaningfulness judgements. In this Chapter, in Experiment 1, I put the contextualised meaning maps to the same test as the original meaning maps. Specifically, I assessed whether they are able to account for the human

tendency to fixate semantically inconsistent objects more than consistent objects. The experiment revealed that this is not the case. Moreover, the experiment provided an indication that introducing semantic inconsistency to an image region results in lower meaningfulness ratings (for this specific region). This observation was confirmed in Experiment 2, in which a carefully selected set of image patches was rated by 140 raters. Interestingly, the average decrease in ratings for patches containing semantic inconsistencies was underpinned by a considerable between-rater variability. Together, these results demonstrate that while contextualised meaning maps share the limitations of the original meaning maps, they (or the patch-rating task used to create them) might be useful as a tool for investigating individual differences in scene processing.

Chapter Four

In the two previous Chapters, I demonstrated that quantifying semantic information in images – one of the top-down factors thought to guide human gaze – remains a challenging task. The idea of this quantification is rooted in the conviction that top-down and bottom-up factors influencing gaze can be disentangled and measured independently. In this Chapter, I adopt a different perspective and focus on their interaction. Specifically, I consider the process of knowledge-driven perceptual organization: the situation when prior object-knowledge possessed by an individual is rapidly deployed to determine the way in which visual input is segmented into objects. As a tool to study to this process, I used two-tone images: black and white, Mooney-style versions of photographs of natural scenes ('templates'). The two-tones are perceived as meaningless by observers who have not seen the templates. Only observers who know the templates are able to perceptually organize the two-tone images into coherent scenes. In three experiments, I compared the eye movements of observers viewing two-tones before and after viewing their templates. The key finding is that gaze patterns on the two-tones resemble those from the templates to a larger extent when the two-tones are bound into object percepts, as compared to when they are not. This result suggests that eye-movements are determined, to a large extent, by the object representations that result from the interaction between image-computable features and prior object-knowledge. These effects were observable already in the first eye-movements made by observers after stimulus onset. Furthermore, knowledge-driven perceptual organization changed various characteristics of gaze behaviour, such as the number of fixations and fixation duration. In summary, I

demonstrated that the interaction between image-computable features and prior object-knowledge possessed by the observers affects human gaze behaviour in a multifaceted way.

Future directions

In the process of writing this thesis, I identified a number of theoretical, methodological and practical issues related to human oculomotor control during natural-scene viewing which remain understudied or unnoticed. Some of them are specific to the content of this thesis, while others are more general. Below, I outline several directions for further research, which – in my view – carry the potential to address these issues.

Reconciling the relational nature of meaning with the spatial nature of images and eye movements

Chapters Two and Three, focusing on the assessment of the meaning maps approach, revealed that to investigate the role of semantic information in visual scenes, stronger theoretical foundations and more precise definitions of terms such as 'meaning' are necessary. One of the key issues highlighted by these two Chapters is that the conceptualisation of semantics commonly used in psychology is not easily adapted to the way in which gaze behaviour is typically modelled. Specifically, there is a tension between the assumptions of this conceptualisation and the way in which the distributions of fixations in natural scenes are typically modelled. The dominant approach to modelling fixations – inspired by saliency modelling – boils down to providing the distribution of some property (saliency, for example) over an image which determines how likely certain pixels are to be fixated. The models of semantics (see Chapter One), on the other hand, assume that (i) meaning is carried by objects, (ii) that the meaning of a certain object is constituted by its place in an abstract, multidimensional conceptual space, occupied by other objects, and (iii) that the distances between objects in that space reflect the degree of their semantic relatedness (Rose & Bex, 2020; Sadeghi et al., 2015). Incorporating semantics into the process of generating distributions predicting fixations (akin to the outputs generated by saliency models) would – for example – require proposing a way in which, for a given object, its location in the conceptual space affects its chance for being fixated when it is shown in the image. One promising avenue for building

'saliency models with semantics' might be to incorporate one additional factor in them, apart from (image-based) 'saliency' and 'semantics': an observer's internal states and priorities. For example, it is known that observers performing visual search tasks rely on semantic relationships between the target (an item to be found) and objects in the display when selecting subsequent fixations targets (Hwang et al., 2011; although see Wu et al., 2014). It is conceivable that this behaviour could be modelled using a combination of image-based saliency model, a model of conceptual space, and a model of internal state of the observer, which would store the identity of the target. In fact, recent studies demonstrate the feasibility of such approach (Rose & Bex, 2020; Treder et al., 2020).

Taking differences between images into account

The rarely spelled out but ubiquitous assumption of saliency models is that they are a general-purpose tool and should be able to predict fixations for almost any given image. This assumed broad scope of applicability is reflected in the way, in which models are routinely assessed (Kümmerer et al., 2020; see also Chapter One): the values of some metric of quality of predictions are calculated for multiple images and averaged. The resulting value serves an indicator of a model's predictive power. The limitation of this approach is that it neglects the differences both between individual images and image categories, while it is clear that that such differences exist (Torralba & Oliva, 2001) and that they are relevant for oculomotor control. For example, Onat and colleagues (Onat et al., 2014) showed that the extent to which simple visual features predict human fixations on images varies as a function of a general category (e. g., urban or natural scene) to which the images belong. Therefore, an interesting avenue for future research would be to investigate in more detail – both experimentally and by means of developing dedicated saliency models – how scene category or some characteristics of individual images contribute to eye-movement guidance. Given that a lot of information can be extracted from images rapidly (Thorpe et al., 1996, see also Chapter Four), and that the oculomotor system is highly flexible in its behaviour (Rothkegel et al., 2019), it can be hypothesised that the information extracted from a scene initially might determine the mode of operation of the oculomotor system adopted when inspecting this scene which, in turn, affects subsequent characteristics of gaze behaviour.

Clarifying the role of computational models

In Chapter Two, I used a variety of saliency models. The intense reliance on these models as tools to study oculomotor control, on the one hand, resulted in many important developments (reviewed in Chapter One). On the other hand, their popularity and diversity resulted in these models being used – and even built – for purposes which are often unclear. This problem is related to the gradual evolution of saliency models, from being the expressions of constructs postulated by specific theories (for example, Feature Integration Theory; A. Treisman, 1985; A. M. Treisman & Gelade, 1980), to being algorithms designed for accurately predicting where people would look in images. Below, I use three examples to outline this evolution and highlight how this evolution affected the usability of saliency models for theory development. A more in-depth analysis of this issue is – in my opinion – much needed to advance our understanding of oculomotor control.

Early saliency models were derived from specific theories about the relationships between image features and fixations allocation and, as such, were convenient tools for generating testable predictions of these theories. Testing these predictions, in turn, lead to the modifications of the theories. For example, consider a study by Einhäuser and Konig (2003; see also Parkhurst & Niebur, 2004). These authors tested the hypothesis that image locations with high luminance contrast attract fixations. In order to do this, they recorded eye movements of observers viewing scenes, in which contrast was manipulated. Then, they assessed the influence of these manipulations and fixations allocation. One of the main findings of this study was that image locations for which local contrast is strongly reduced attract fixations. This finding resulted in the refinenement of theory: the idea that only high-contrast areas attract gaze turned out to be too simplistic. In this study, the distribution of luminance contrast played a role of a 'saliency model': it was used to generate predictions based on a theory positing that high values of this specific feature attract fixations.

Next, consider the line of research focusing on object locations (Borji & Tanner, 2016; Nuthmann et al., 2020; Nuthmann & Henderson, 2010). The maps indexing the locations of objects within an image can be treated as a saliency models too (assuming that saliency is understood in a broad sense). This line of research provided evidence that object locations predict fixations well, and that this effect does not supervene on saliency indexed both by the

simplest and more advanced models. Yet, the theoretical consequences of these studies are less clear than in the previous example. The question about which aspects of objects are important for eye-movements (their specific, complex visual features? the interaction of features with the prior knowledge about the world?) remains open. The result that object locations predict fixations well, rather than leading to a straightforward theoretical refinement, may be treated rather as a heuristic and inspiration for further experiments.

Finally, consider the successes of models based on deep neural networks. The quality of predictions they generate is superb (Kümmerer et al., 2020), but the reasons for this good performance are yet to be fully elucidated. Deep neural networks are characterised by an enormous number of parameters and the way in which they process the input is still only partially understood. Therefore, using saliency models based on these networks is an extreme case of the situation illustrated in the two previous examples: the increase in the predictive power of a saliency model happens at the expense of understanding why the model performs so well, thus reducing its heuristic usefulness.

The picture which emerges from these three examples is that, currently, models can be very successful at making valid predictions, but nevertheless lack explanatory power. They may, however, provide good heuristics about future avenues for research. This problem of tension between ‘explaining’ and ‘predicting’ present in the eye-movement literature is only briefly outlined here. A more detailed elaboration of it would be undoubtedly beneficial both for the researchers who build models and for those, who rely on them in their studies.

Bibliography

Adeli, H., Vitu, F., & Zelinsky, G. J. (2017). A Model of the Superior Colliculus Predicts Fixation Locations during Scene Viewing and Visual Search. *The Journal of Neuroscience*, *37*(6), 1453–1467. https://doi.org/10.1523/JNEUROSCI.0825-16.2016

Akbas, E., & Eckstein, M. P. (2017). Object detection through search with a foveated visual system. *PLoS Computational Biology*, *13*(10), e1005743. https://doi.org/10.1371/journal.pcbi.1005743

Alfandari, D., Belopolsky, A. V., & Olivers, C. N. L. (2019). Eye movements reveal learning and information-seeking in attentional template acquisition. *Visual Cognition*, *27*(5–8), 467–486. https://doi.org/10.1080/13506285.2019.1636918

Anderson, N. C., & Donk, M. (2017). Salient object changes influence overt attentional prioritization and object-based targeting in natural scenes. *Plos One*, *12*(2), e0172132. https://doi.org/10.1371/journal.pone.0172132

Anstis, S. M. (1974). A chart demonstrating variations in acuity with retinal position. *Vision Research*, *14*(7), 589–592. https://doi.org/10.1016/0042-6989(74)90049-2

Awh, E., Belopolsky, A. V., & Theeuwes, J. (2012). Top-down versus bottom-up attentional control: A failed theoretical dichotomy. *Trends in Cognitive Sciences*, *16*(8), 437–443. https://doi.org/10.1016/j.tics.2012.06.010

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

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

Baranes, A., Oudeyer, P. Y., & Gottlieb, J. (2015). Eye movements reveal epistemic curiosity in human observers. *Vision Research*, *117*, 81–90. https://doi.org/10.1016/j.visres.2015.10.009

Bayat, A., Nand, A. K., Koh, D. H., Pereira, M., & Pomplun, M. (2018). Scene grammar in human and machine recognition of objects and scenes. *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, *2018-June*(June), 2073–2080. https://doi.org/10.1109/CVPRW.2018.00268

Berga, D., & Otazu, X. (2018). A Neurodynamic model of Saliency prediction in V1. *ArXiv*.

Berga, D., & Otazu, X. (2020). Modeling bottom-up and top-down attention with a neurodynamic model of V1. *Neurocomputing*, *417*, 270–289. https://doi.org/10.1016/j.neucom.2020.07.047

Berman, R., & Colby, C. (2009). Attention and active vision. *Vision Research*, *49*(10), 1233–1248. https://doi.org/10.1016/j.visres.2008.06.017

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

Bindemann, M. (2010). Scene and screen center bias early eye movements in scene viewing. *Vision Research*, *50*(23), 2577–2587. https://doi.org/10.1016/j.visres.2010.08.016

Bisley, J. W., & Mirpour, K. (2019). The neural instantiation of a priority map. *Current Opinion in Psychology*, *29*, 108–112. https://doi.org/10.1016/j.copsyc.2019.01.002

Boettcher, S. E. P., Draschkow, D., Dienhart, E., & Võ, M. L. H. (2018). Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of Vision*, *18*(13), 1–13. https://doi.org/10.1167/18.13.11

Bonitz, V. S., & Gordon, R. D. (2008). Attention to smoking-related and incongruous objects during scene viewing. *Acta Psychologica*, *129*(2), 255–263. https://doi.org/10.1016/j.actpsy.2008.08.006

Borji, A., & Itti, L. (2013). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(1), 185–207. https://doi.org/10.1109/TPAMI.2012.89

Borji, A., Sihite, D. N., & Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of einẗauser et al.'s data. *Journal of Vision*, *13*(10), 1–4. https://doi.org/10.1167/13.10.18

Borji, A., & Tanner, J. (2016). Reconciling Saliency and Object Center-Bias Hypotheses in Explaining Free-Viewing Fixations. *IEEE Transactions on Neural Networks and Learning Systems*, *27*(6), 1214–1226. https://doi.org/10.1109/TNNLS.2015.2480683

Brainard, D. (1997). The Psychophysics Toolbox. *Spatial Vision*, *10*(4).

Bruce, N. D. B., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, *9*(3). https://doi.org/10.1167/9.3.5

Bullier, J. (2001). Integrated model of visual processing. *Brain Research Reviews*, *36*(2–3), 96–107. https://doi.org/10.1016/S0165-0173(01)00085-6

Bylinskii, Z., DeGennaro, E. M. M., Rajalingham, R., Ruda, H., Zhang, J., & Tsotsos, J. K. K. (2015). Towards the quantitative evaluation of visual attention models. *Vision Research*, *116*, 258–268. https://doi.org/10.1016/j.visres.2015.04.007

Bylinskii, Z., Judd, T., Oliva, A., Torralba, A., & Durand, F. (2016). What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *41*(3), 740–757. https://doi.org/10.1109/TPAMI.2018.2815601

Castelhano, M. S., Mack, M. L., & Henderson, J. M. (2009). Viewing task influences eye movement control during active scene perception. *Journal of Vision*, *9*(3), 1–15. https://doi.org/10.1167/9.3.6

Cavanagh, P. (2011). Visual cognition. *Vision Research*, *51*(13), 1538–1551. https://doi.org/10.1016/j.visres.2011.01.015

Cerf, M., Paxon Frady, E., & Koch, C. (2009). Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of Vision*, *9*(12), 1–15. https://doi.org/10.1167/9.12.1

Chang, R., Baria, A. T., Flounders, M. W., & He, B. J. (2016). Unconsciously elicited perceptual prior. *Neuroscience of Consciousness*, *2016*(1), niw008. https://doi.org/10.1093/nc/niw008

Christensen, J. H., Bex, P. J., & Fiser, J. (2015). Prior implicit knowledge shapes human threshold for orientation noise. *Journal of Vision*, *15*(9), 1–15. https://doi.org/10.1167/15.9.24

Christensen, J. H., Bex, P. J., & Fiser, J. (2019). Coding of low-level position and orientation information in human naturalistic vision. *PLoS ONE*, *14*(2), 1–23. https://doi.org/10.1371/journal.pone.0212141

Chua, H. F., Boland, J. E., & Nisbett, R. E. (2005). Cultural variation in eye movements during scene perception. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(35), 12629–12633. https://doi.org/10.1073/pnas.0506162102

Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. https://doi.org/10.1038/nn.3635

Clarke, A. D. F., & Tatler, B. W. (2014). Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, *102*, 41–51. https://doi.org/10.1016/j.visres.2014.06.016

Clarke, A., & Tyler, L. K. (2014). Object-Specific Semantic Coding in Human Perirhinal Cortex. *Journal of Neuroscience*, *34*(14), 4766–4775. https://doi.org/10.1523/JNEUROSCI.2828-13.2014

Coco, M. I., Nuthmann, A., & Dimigen, O. (2020). Fixation-related Brain Potentials during Semantic Integration of Object–Scene Information. *Journal of Cognitive Neuroscience*, *32*(4), 571–589. https://doi.org/10.1162/jocn_a_01504

Constantinescu, A. O., OReilly, J. X., & Behrens, T. E. J. (2016). Organizing conceptual knowledge in humans with a gridlike code. *Science*, *352*(6292), 1464–1468. https://doi.org/10.1126/science.aaf0941

Cousineau, D. (2011). Randomization test of mean is compuationally inaccessible when the number of groups exceeds two. *Tutorials in Quantitative Methods for Psychology*, *7*(1), 15–18.

Crutch, S. J., & Warrington, E. K. (2005). Abstract and concrete concepts have structurally different representational frameworks. *Brain*, *128*(3), 615–627. https://doi.org/10.1093/brain/awh349

De Graef, P., Christiaens, D., & D'Ydewalle, G. (1990). Perceptual effects of scene context on object identification. *Psychological Research*, *52*(4), 317–329. https://doi.org/10.1007/BF00868064

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

Deubel, H., & Schneider, W. X. (1996). Saccade target selection and object recognition: Evidence for a common attentional mechanism. *Vision Research*, *36*(12), 1827–1837. https://doi.org/10.1016/0042-6989(95)00294-4

Devereux, B. J., Tyler, L. K., Geertzen, J., & Randall, B. (2014). The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, *46*(4), 1119–1127. https://doi.org/10.3758/s13428-013-0420-4

Dorr, M., Martinetz, T., Gegenfurtner, K. R., & Barth, E. (2010). Variability of eye movements when viewing dynamic natural scenes. *Journal of Vision*, *10*(10), 28–28. https://doi.org/10.1167/10.10.28

Dosso, J. A., Huynh, M., & Kingstone, A. (2020). I spy without my eye: Covert attention in human social interactions. *Cognition*, *202*(June), 104388. https://doi.org/10.1016/j.cognition.2020.104388

Drewes, J., Trommershäuser, J., & Gegenfurtner, K. R. (2011). Parallel visual search and rapid animal detection in natural scenes. *Journal of Vision*, *1120*(2), 1–21. https://doi.org/10.1167/11.2.20

Driver, J., Davis, G., Russell, C., Turatto, M., & Freeman, E. (2001). Segmentation, attention and phenomenal visual objects. *Cognition*, *80*(1–2), 61–95. https://doi.org/10.1016/S0010-0277(00)00151-7

Duncan, J. (1984). Selective attention and the organization of visual information. *Journal of Experimental Psychology: General*, *113*(4), 501–517. https://doi.org/10.1037/0096-3445.113.4.501

Egly, R., Driver, J., & Rafal, R. D. (1994). Shifting Visual Attention Between Objects and Locations: Evidence From Normal and Parietal Lesion Subjects. *Journal of Experimental Psychology: General*, *123*(2), 161–177. https://doi.org/10.1037/0096-3445.123.2.161

Ehinger, B. V., Kaufhold, L., & König, P. (2018). Probing the temporal dynamics of the exploration- exploitation dilemma of eye movements. *Journal of Vision*, *18*(3), 1–24. https://doi.org/10.1167/18.3.6

Einhäuser, W. (2013). Objects and saliency: Reply to Borji et al. *Journal of Vision*, *13*(10), 20–20. https://doi.org/10.1167/13.10.20

Einhäuser, W., & König, P. (2003). Does luminance-contrast contribute to a saliency map for overt visual attention? *European Journal of Neuroscience*, *17*(5), 1089–1097. https://doi.org/10.1046/j.1460-9568.2003.02508.x

Einhäuser, W., Rutishauser, U., & Koch, C. (2008). Task-demands can immediately reverse the effects of sensory-driven saliency in complex visual stimuli. *Journal of Vision*, *8*(2), 1–19. https://doi.org/10.1167/8.2.2

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 18.1-26. https://doi.org/10.1167/8.14.18

Elazary, L., & Itti, L. (2008). Interesting objects are visually salient. *Journal of Vision*, *8*(3), 1–15. https://doi.org/10.1167/8.3.3

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Federico, G., & Brandimonte, M. A. (2019). Tool and object affordances: An ecological eye-tracking study. *Brain and Cognition*, *135*(May), 103582. https://doi.org/10.1016/j.bandc.2019.103582

Fellrath, J., & Ptak, R. (2015). The role of visual saliency for the allocation of attention: Evidence from spatial neglect and hemianopia. *Neuropsychologia*, *73*, 70–81. https://doi.org/10.1016/j.neuropsychologia.2015.05.003

Firestone, C., & Scholl, B. J. (2015). Cognition does not affect perception: Evaluating the evidence for top-down effects. *Behavioral and Brain Sciences*, (2016). https://doi.org/10.1017/S0140525X15000965

Flechsenhar, A. F., & Gamer, M. (2017). Top-down influence on gaze patterns in the presence of social features. *PLoS ONE*, *12*(8), 1–20. https://doi.org/10.1371/journal.pone.0183799

Flounders, M. W., González-García, C., Hardstone, R., & He, B. J. (2019). Neural dynamics of visual ambiguity resolution by perceptual prior. *ELife*, *8*, 1–25. https://doi.org/10.7554/eLife.41861

Foulsham, T., Gray, A., Nasiopoulos, E., & Kingstone, A. (2013). Leftward biases in picture scanning and line bisection: A gaze-contingent window study. *Vision Research*, *78*, 14–25. https://doi.org/10.1016/j.visres.2012.12.001

Foulsham, T., & Kingstone, A. (2013). Optimal and preferred eye landing positions in objects and scenes. *Quarterly Journal of Experimental Psychology*, *66*(9), 1707–1728. https://doi.org/10.1080/17470218.2012.762798

Foulsham, T., & Kingstone, A. (2017). Are fixations in static natural scenes a useful predictor of attention in the real world? *Canadian Journal of Experimental Psychology*, *71*(2), 172–181. https://doi.org/10.1037/cep0000125

Foulsham, T., & Underwood, G. (2007). How does the purpose of inspection influence the potency of visual salience in scene perception? *Perception*, *36*(8), 1123–1138. https://doi.org/10.1068/p5659

Foulsham, T., & Underwood, G. (2008). What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition. *Journal of Vision*, *8*(2), 1–17. https://doi.org/10.1167/8.2.6

Foulsham, T., Walker, E., & Kingstone, A. (2011). The where, what and when of gaze allocation in the lab and the natural environment. *Vision Research*, *51*(17), 1920–1931. https://doi.org/10.1016/j.visres.2011.07.002

Franchak, J. M., Heeger, D. J., Hasson, U., & Adolph, K. E. (2016). Free Viewing Gaze Behavior in Infants and Adults. *Infancy*, *21*(3), 262–287. https://doi.org/10.1111/infa.12119

Friedman, A. (1979). Framing Pictures: The Role of Knowledge in Automatized Encoding and Memory for Gist. *Journal of Experimental Psychology: General*, *108*(3), 316–355. https://doi.org/10.1037/0096-3445.108.3.316

Gameiro, R. R., Kaspar, K., König, S. U., Nordholt, S., & König, P. (2017). Exploration and Exploitation in Natural Viewing Behavior. *Scientific Reports*, *7*(1), 1–23. https://doi.org/10.1038/s41598-017-02526-1

Gamer, M., Lemon, J. and, Fellows, I., & Singh, P. (2019). *irr: Various Coefficients of Interrater Reliability and Agreement*. Retrieved from https://cran.r-project.org/package=irr

Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, *30*(1), 51–64. https://doi.org/10.1016/j.imavis.2011.11.007

Garcia-Diaz, A., Leboran, V., Fdez-Vidal, X. R., & Pardo, X. M. (2012). On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision*, *12*(6). https://doi.org/10.1167/12.6.17

Gilchrist, I. D., & Findlay, J. M. (2001). Visual Attention: The Active Vision Perspective. In M. Jenkin & L. Harris (Eds.), *Vision and Attention*. New York, NY: Springer New York. https://doi.org/10.1007/978-0-387-21591-4

Goh, J. O., Tan, J. C., & Park, D. C. (2009). Culture Modulates Eye-Movements to Visual Novelty. *PLoS ONE*, *4*(12), e8238. https://doi.org/10.1371/journal.pone.0008238

González-García, C., Flounders, M. W., Chang, R., Baria, A. T., & He, B. J. (2018). Content-specific activity in frontoparietal and default-mode networks during prior-guided visual perception. *ELife*, *7*, 1–25. https://doi.org/10.7554/eLife.36068

Gorlin, S., Meng, M., Sharma, J., Sugihara, H., Sur, M., & Sinha, P. (2012). Imaging prior information in the brain. *Proceedings of the National Academy of Sciences*, *109*(20), 7935–7940. https://doi.org/10.1073/pnas.1111224109

Gottlieb, J. (2012). Attention, Learning, and the Value of Information. *Neuron*, *76*(2), 281–295. https://doi.org/10.1016/j.neuron.2012.09.034

Guterstam, A., Kean, H. H., Webb, T. W., Kean, F. S., & Graziano, M. S. A. (2019). Implicit model of other people's visual attention as an invisible, force-carrying beam projecting from the eyes. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(1), 328–333. https://doi.org/10.1073/pnas.1816581115

Hallett, P. E. (1978). Primary and secondary saccades to goals defined by instructions. *Vision Research*, *18*(10), 1279–1296. https://doi.org/10.1016/0042-6989(78)90218-3

Harel, J., Koch, C., & Perona, P. (2007). Graph-Based Visual Saliency. In *Advances in Neural Information Processing Systems 19* (Vol. 19, pp. 545–552). The MIT Press. https://doi.org/10.7551/mitpress/7503.003.0073

Harris, Z. S. (1954). Distributional Structure. *WORD*, *10*(2–3), 146–162. https://doi.org/10.1080/00437956.1954.11659520

Haskins, A. J., Mentch, J., Botch, T. L., & Robertson, C. E. (2020). Active vision in immersive, 360° real-world environments. *Scientific Reports*, *10*(1), 1–11. https://doi.org/10.1038/s41598-020-71125-4

Hayes, A. F., & Krippendorff, K. (2007). Answering the Call for a Standard Reliability Measure for Coding Data. *Communication Methods and Measures*, *1*(1), 77–89. https://doi.org/10.1080/19312450709336664

Hayhoe, M. M., Shrivastava, A., Mruczek, R., & Pelz, J. B. (2003). Visual memory and motor planning in a natural task. *Journal of Vision*, *3*(1), 49–63. https://doi.org/10.1167/3.1.6

Hebart, M. N., Zheng, C. Y., Pereira, F., & Baker, C. I. (2020). Revealing the multidimensional mental representations of natural objects underlying human similarity judgements. *Nature Human Behaviour*, *4*(November). https://doi.org/10.1038/s41562-020-00951-3

Hedge, C., Powell, G., & Sumner, P. (2018). The reliability paradox: Why robust cognitive tasks do not produce reliable individual differences. *Behavior Research Methods*, *50*(3), 1166–1186. https://doi.org/10.3758/s13428-017-0935-1

Henderson, J. M. (2011). Eye movements and scene perception. In S. P. Liversedge, I. D. Gilchrist, & S. Everling (Eds.), *The Oxford Handbook of Eye Movements*. Oxford: Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199539789.013.0033

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

Henderson, J. M. (2020). Meaning and attention in scenes. In *Psychology of Learning and Motivation - Advances in Research and Theory* (1st ed., Vol. 73). Elsevier Inc. https://doi.org/10.1016/bs.plm.2020.08.002

Henderson, J. M., & Hayes, T. R. (2017). Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nature Human Behaviour*, *1*(October). https://doi.org/10.1038/s41562-017-0208-0

Henderson, J. M., & Hayes, T. R. (2018). Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *Journal of Vision*, *18*(6), 10. https://doi.org/10.1167/18.6.10

Henderson, J. M., Hayes, T. R., Peacock, C. E., & Rehrig, G. (2019). Meaning and Attentional Guidance in Scenes: A Review of the Meaning Map Approach. *Vision*, *3*(2), 19. https://doi.org/10.3390/vision3020019

Henderson, J. M., Hayes, T. R., Rehrig, G., & Ferreira, F. (2018). Meaning Guides Attention during Real-World Scene Description. *Scientific Reports*, *8*(1), 13504. https://doi.org/10.1038/s41598-018-31894-5

Henderson, J. M., Malcolm, G. L., & Schandl, C. (2009). Searching in the dark: Cognitive relevance drives attention in real-world scenes. *Psychonomic Bulletin & Review*, *16*(5), 850–856. https://doi.org/10.3758/PBR.16.5.850

Henderson, J. M., Shinkareva, S. V, Wang, J., Luke, S. G., & Olejarczyk, J. (2013). Predicting Cognitive State from Eye Movements. *PLoS ONE*, *8*(5). https://doi.org/10.1371/journal.pone.0064937

Henderson, J. M., Weeks, Phillip A., J., & Hollingworth, A. (1999). The Effects of Semantic Consistency on Eye Movements During Complex Scene Viewing. *Journal of Experimental*

Psychology: Human Perception and Performance, *25*(1), 210–228. https://doi.org/10.1037/0096-1523.25.1.210

Hills, T. T., Todd, P. M., Lazer, D., Redish, A. D., Couzin, I. D., Bateson, M., … Wolfe, J. W. (2015). Exploration versus exploitation in space, mind, and society. *Trends in Cognitive Sciences*, *19*(1), 46–54. https://doi.org/10.1016/j.tics.2014.10.004

Horga, G., & Abi-Dargham, A. (2019). An integrative framework for perceptual disturbances in psychosis. *Nature Reviews Neuroscience*, *20*(12), 763–778. https://doi.org/10.1038/s41583-019-0234-1

Hsieh, P. J., Vul, E., & Kanwisher, N. (2010). Recognition alters the spatial pattern of fMRI activation in early retinotopic cortex. *Journal of Neurophysiology*, *103*(3), 1501–1507. https://doi.org/10.1152/jn.00812.2009

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

Hwang, A. D., Wang, H.-C., & Pomplun, M. (2011). Semantic guidance of eye movements in real-world scenes. *Vision Research*, *51*(10), 1192–1205. https://doi.org/10.1016/j.visres.2011.03.010

Ishikawa, T., & Mogi, K. (2011). Visual one-shot learning as an "anti-camouflage device": A novel morphing paradigm. *Cognitive Neurodynamics*, *5*(3), 231–239. https://doi.org/10.1007/s11571-011-9171-z

Itti, L., & Koch, C. (2000). A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, *40*(10–12), 1489–1506. https://doi.org/10.1016/S0042-6989(99)00163-7

Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. https://doi.org/10.1038/35058500

Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *20*(11), 1254–1259. https://doi.org/10.1109/34.730558

Jarosz, A. F., & Wiley, J. (2014). What Are the Odds? A Practical Guide to Computing and Reporting Bayes Factors. *The Journal of Problem Solving*, *7*(1), 2–9. https://doi.org/10.7771/1932-6246.1167

Judd, T., Durand, F., & Torralba, A. (2012). A Benchmark of Computational Models of Saliency to Predict Human Fixations. In *MIT Technical Report*. SCITEPRESS - Science and Technology Publications. https://doi.org/10.5220/0005678701340142

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

Kaspar, K., Hloucal, T. M., Kriz, J., Canzler, S., Gameiro, R. R., Krapp, V., & König, P. (2013). Emotions' Impact on Viewing Behavior under Natural Conditions. *PLoS ONE*, *8*(1). https://doi.org/10.1371/journal.pone.0052737

Kienzle, W., Wichmann, F. A., Schölkopf, B., & Franz, M. O. (2007). A nonparametric approach to bottom-up visual saliency. *Advances in Neural Information Processing Systems*, (May 2014), 689–696.

Kietzmann, T. C., McClure, P., Kriegeskorte, N., Kietzmann, T. C., McClure, P., & Kriegeskorte, N. (2019). Deep Neural Networks in Computational Neuroscience. *Oxford Research Encyclopedia of Neuroscience*. https://doi.org/10.1093/acrefore/9780190264086.013.46

Kilpelaïnen, M., & Georgeson, M. A. (2018). Luminance gradient at object borders communicates object location to the human oculomotor system. *Scientific Reports*, *8*(1), 1–11. https://doi.org/10.1038/s41598-018-19464-1

Kleiner, M., Brainard, D., & Pelli, D. G. (2007). What's new in Psychtoolbox-3 ? *Perception*, *36*(1), 1. https://doi.org/10.1016/S0140-6736(13)62162-5

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: towards the Neural Circuitry. In *Human neurobiology* (Vol. 4, pp. 219–227).

Koch, C., & Ullman, S. (1987). Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. *Matters of Intelligence*, pp. 115–141. https://doi.org/10.1007/978-94-009-3833-5_5

Koehler, K., Guo, F., Zhang, S., & Eckstein, M. P. (2014). What do saliency models predict? *Journal of Vision*, *14*(3). https://doi.org/10.1167/14.3.14

Kollmorgen, S., Nortmann, N., Schröder, S., & König, P. (2010). Influence of low-level stimulus features, task dependent factors, and spatial biases on overt visual attention. *PLoS Computational Biology*, *6*(5). https://doi.org/10.1371/journal.pcbi.1000791

Kourtzi, Z., & Connor, C. E. (2011). Neural representations for object perception: Structure, category, and adaptive coding. *Annual Review of Neuroscience*, *34*, 45–67. https://doi.org/10.1146/annurev-neuro-060909-153218

Krasich, K., Huffman, G., Faber, M., & Brockmole, J. R. (2020). Where the eyes wander: The relationship between mind wandering and fixation allocation to visually salient and semantically informative static scene content. *Journal of Vision*, *20*(9), 1–30. https://doi.org/10.1167/JOV.20.9.10

Krasovskaya, S., & MacInnes, W. J. (2019). Salience Models: A Computational Cognitive Neuroscience Review. *Vision*, *3*(4), 56. https://doi.org/10.3390/vision3040056

Krieger, G., Rentschler, I., Hauske, G., Schill, K., & Zetzsche, C. (2000). Object and scene analysis by saccadic eye-movements: An investigation with higher-order statistics. *Spatial Vision*, *13*(2–3), 201–214. https://doi.org/10.1163/156856800741216

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

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*(NOV), 1–28. https://doi.org/10.3389/neuro.06.004.2008

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). ImageNet Classification with Deep Convolutional Neural Networks. *Advances In Neural Information Processing Systems*, *60*(6), 1–9. https://doi.org/10.1145/3065386

Kröger, J. L., Lutz, O. H.-M., & Müller, F. (2020). What Does Your Gaze Reveal About You? On the Privacy Implications of Eye Tracking. In *IFIP Advances in Information and Communication Technology*: *Vol. 576 LNCS* (pp. 226–241). Springer International Publishing. https://doi.org/10.1007/978-3-030-42504-3_15

Król, M. E. M., & Król, M. E. M. (2018). "Economies of Experience"-Disambiguation of Degraded Stimuli Leads to a Decreased Dispersion of Eye-Movement Patterns. *Cognitive Science*, *42*, 728–756. https://doi.org/10.1111/cogs.12566

Król, M., & Król, M. (2019). The world as we know it and the world as it is: Eye-movement patterns reveal decreased use of prior knowledge in individuals with autism. *Autism Research*, *12*(9), 1386–1398. https://doi.org/10.1002/aur.2133

Kumaran, D., Summerfield, J. J., Hassabis, D., & Maguire, E. A. (2009). Tracking the Emergence of Conceptual Knowledge during Human Decision Making. *Neuron*, *63*(6), 889–901. https://doi.org/10.1016/j.neuron.2009.07.030

Kümmerer, M., Bylinskii, Z., Judd, T., Borji, A., Itti, L., Durand, F., … Torrabla, A. (2020). MIT/Tübingen Saliency Benchmark. Retrieved from https://saliency.tuebingen.ai/

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2015). Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, *112*(52), 16054–16059. https://doi.org/10.1073/pnas.1510393112

Kümmerer, M., Wallis, T. S. A., & Bethge, M. (2016). *DeepGaze II: Reading fixations from deep features trained on object recognition*. Retrieved from http://arxiv.org/abs/1610.01563

Kümmerer, M., Wallis, T. S. A., Gatys, L. A., & Bethge, M. (2017a). Understanding Low- and High-Level Contributions to Fixation Prediction. *Proceedings of the IEEE International Conference on Computer Vision*, *2017-Octob*, 4799–4808. https://doi.org/10.1109/ICCV.2017.513

Lakens, D. (2017). Equivalence Tests: A Practical Primer for t Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Lakens, D., Scheel, A. M., & Isager, P. M. (2018). Equivalence Testing for Psychological Research: A Tutorial. *Advances in Methods and Practices in Psychological Science*, *1*(2), 259–269. https://doi.org/10.1177/2515245918770963

Land, M. F., & McLeod, P. (2000). From eye movement to actions: how batsman hit the ball. *Nature Neuroscience*, *3*(12), 1340–1345.

Lee, T. S. (2015). The Visual System's Internal Model of the World. *Proceedings of the IEEE*, *103*(8), 1359–1378. https://doi.org/10.1109/JPROC.2015.2434601

Leek, Ch., E., Patterson, C., Paul, M. A., Rafal, R., & Cristino, F. (2012). Eye movements during object recognition in visual agnosia. *Neuropsychologia*, *50*(9), 2142–2153. https://doi.org/10.1016/j.neuropsychologia.2012.05.005

Lengyel, G., Žalalytė, G., Pantelides, A., Ingram, J. N., Fiser, J., Lengyel, M., & Wolpert, D. M. (2019). Unimodal statistical learning produces multimodal object-like representations. *ELife*, *8*, 1–21. https://doi.org/10.7554/eLife.43942

Loftus, G. R., & Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human Perception and Performance*, *4*(4), 565–572. https://doi.org/10.1037/0096-1523.4.4.565

Logan, G. D. (2004). Cumulative progress in formal theories of attention. *Annual Review of Psychology*, *55*, 207–234. https://doi.org/10.1146/annurev.psych.55.090902.141415

Loth, E., Gómez, J. C., & Happé, F. (2010). When seeing depends on knowing: Adults with Autism Spectrum Conditions show diminished top-down processes in the visual perception of degraded faces but not degraded objects. *Neuropsychologia*, *48*(5), 1227–1236. https://doi.org/10.1016/j.neuropsychologia.2009.12.023

Lüddecke, T., Agostini, A., Fauth, M., Tamosiunaite, M., & Wörgötter, F. (2019). Distributional semantics of objects in visual scenes in comparison to text. *Artificial Intelligence*, *274*, 44–65. https://doi.org/10.1016/j.artint.2018.12.009

Luke, S. G., & Henderson, J. M. (2016). The influence of content meaningfulness on eye movements across tasks: Evidence from scene viewing and reading. *Frontiers in Psychology*, *7*(MAR), 1–10. https://doi.org/10.3389/fpsyg.2016.00257

Lyu, M., Choe, K. W., Kardan, O., Kotabe, H. P., Henderson, J. M., & Berman, M. G. (2020). Overt attentional correlates of memorability of scene images and their relationships to scene semantics. *Journal of Vision*, *20*(9), 2. https://doi.org/10.1167/jov.20.9.2

Mackworth, N. H., & Morandi, A. J. (1967). The gaze selects information details within picutres. *Perception & Psychophysics*, *2*(11), 547–552.

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

Marr, D., & Nishihara, H. K. (1978). Representation and recognition of the spatial organization of three-dimensional shapes. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character. Royal Society (Great Britain)*, *200*(1140), 269–294. https://doi.org/10.1098/rspb.1978.0020

Masciocchi, C. M., Mihalas, S., Parkhurst, D., & Niebur, E. (2009). Everyone knows what is interesting: Salient locations which should be fixated. *Journal of Vision*, *9*(11), 1–22. https://doi.org/10.1167/9.11.1

Maunsell, J. H. R., & Treue, S. (2006). Feature-based attention in visual cortex. *Trends in Neurosciences*, *29*(6), 317–322. https://doi.org/10.1016/j.tins.2006.04.001

McRae, K., Cree, G. S., Seidenberg, M. S., & Mcnorgan, C. (2005). Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, *37*(4), 547–559. https://doi.org/10.3758/BF03192726

Mills, M., Hollingworth, A., Van der Stigchel, S., Hoffman, L., & Dodd, M. D. (2011). Examining the influence of task set on eye movements and fixations. *Journal of Vision*, *11*(8), 17. https://doi.org/10.1167/11.8.17

Mirman, D., Landrigan, J. F., & Britt, A. E. (2017). Taxonomic and thematic semantic systems. *Psychological Bulletin*, *143*(5), 499–520. https://doi.org/10.1037/bul0000092

Mkrtychian, N., Blagovechtchenski, E., Kurmakaeva, D., Gnedykh, D., Kostromina, S., & Shtyrov, Y. (2019). Concrete vs. Abstract Semantics: From Mental Representations to Functional Brain Mapping. *Frontiers in Human Neuroscience*, *13*(August), 1–6. https://doi.org/10.3389/fnhum.2019.00267

Mooney, C. M. (1957). Age in the development of closure ability in children. *Canadian Journal of Psychology*, *11*(4), 219–226. https://doi.org/10.1037/h0083717

Moore, C., & Cavanagh, P. (1998). Recovery of 3D volume from 2-tone images of novel objects. *Cognition*, *67*, 45–71. https://doi.org/S0010-0277(98)00014-6 [pii]

Morey, R. D. (2008). Confidence Intervals from Normalized Data: A correction to Cousineau (2005). *Tutorials in Quantitative Methods for Psychology*, *4*(2), 61–64. https://doi.org/10.20982/tqmp.04.2.p061

Morey, R. D., & Rouder, J. N. (2018). *BayesFactor: Computation of Bayes Factors for Common Designs*. Retrieved from https://cran.r-project.org/package=BayesFactor

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

Neri, P. (2017). Object segmentation controls image reconstruction from natural scenes. In *PLoS Biology* (Vol. 15). https://doi.org/10.1371/journal.pbio.1002611

Noton, D., & Stark, L. (1971). Scanpaths in Eye Movements during Pattern Perception. *Science*, *171*(3968), 308–311. https://doi.org/10.1126/science.171.3968.308

Nuthmann, A., & Einhäuser, W. (2015). A new approach to modeling the influence of image features on fixation selection in scenes. *Annals of the New York Academy of Sciences*, *1339*(1), 82–96. https://doi.org/10.1111/nyas.12705

Nuthmann, A., & Henderson, J. M. (2010). Object-based attentional selection in scene viewing. *Journal of Vision*, *10*(8), 20. https://doi.org/10.1167/10.8.20

Nuthmann, A., & Matthias, E. (2014). Time course of pseudoneglect in scene viewing. *Cortex*, *52*(1), 113–119. https://doi.org/10.1016/j.cortex.2013.11.007

Nuthmann, A., Schütz, I., & Einhäuser, W. (2020). Salience-based object prioritization during active viewing of naturalistic scenes in young and older adults. *Scientific Reports*, *10*(1), 22057. https://doi.org/10.1038/s41598-020-78203-7

Nyström, M., & Holmqvist, K. (2008). Semantic override of low-level features in image viewing–both initially and overall. *Journal of Eye Movement Research*, *2*(2), 1–11. https://doi.org/10.16910/jemr.2.2.2

Öhlschläger, S., & Võ, M. L.-H. (2017). SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes. *Behavior Research Methods*, *49*(5), 1780–1791. https://doi.org/10.3758/s13428-016-0820-3

Onat, S., Açik, A., Schumann, F., & König, P. (2014). The contributions of image content and behavioral relevancy to overt attention. *PLoS ONE*, *9*(4). https://doi.org/10.1371/journal.pone.0093254

Ongchoco, J. D. K., & Scholl, B. J. (2019). How to Create Objects With Your Mind: From Object-Based Attention to Attention-Based Objects. *Psychological Science*, *30*(11), 1648–1655. https://doi.org/10.1177/0956797619863072

Ossandón, J. P., Onat, S., Cazzoli, D., Nyffeler, T., Müri, R., & König, P. (2012). Unmasking the contribution of low-level features to the guidance of attention. *Neuropsychologia*, *50*(14), 3478–3487. https://doi.org/10.1016/j.neuropsychologia.2012.09.043

Ossandón, J. P., Onat, S., & König, P. (2014). Spatial biases in viewing behavior. *Journal of Vision*, *14*(2), 1–26. https://doi.org/10.1167/14.2.20

Page, E. B. (1963). Ordered Hypotheses for Multiple Treatments: A Significance Test for Linear Ranks. *Journal of the American Statistical Association*, *58*(301), 216–230. https://doi.org/10.1080/01621459.1963.10500843

Pajak, M., & Nuthmann, a. (2013). Object-based saccadic selection during scene perception: Evidence from viewing position effects. *Journal of Vision*, *13*(2013), 1–21. https://doi.org/10.1167/13.5.2.doi

Parkhurst, D. J., & Niebur, E. (2004). Texture contrast attracts overt visual attention in natural scenes. *European Journal of Neuroscience*, *19*(3), 783–789. https://doi.org/10.1111/j.0953-816X.2003.03183.x

Peacock, C. E., Hayes, T. R., & Henderson, J. M. (2019). The role of meaning in attentional guidance during free viewing of real-world scenes. *Acta Psychologica*, *198*(December 2018), 102889. https://doi.org/10.1016/j.actpsy.2019.102889

Pedziwiatr, M. A., Kümmerer, M., Wallis, T. S. A., Bethge, M., & Teufel, C. (2021). Meaning maps and saliency models based on deep convolutional neural networks are insensitive to image meaning when predicting human fixations. *Cognition*, *206*(10), 104465. https://doi.org/10.1016/j.cognition.2020.104465

Pereira, E. J., & Castelhano, M. S. (2019). Attentional capture is contingent on scene region: Using surface guidance framework to explore attentional mechanisms during search. *Psychonomic Bulletin and Review*, *26*(4), 1273–1281. https://doi.org/10.3758/s13423-019-01610-z

Pilarczyk, J., & Kuniecki, M. J. (2014). Emotional content of an image attracts attention more than visually salient features in various signal-to-noise ratio conditions. *Journal of Vision*, *14*(12), 4–4. https://doi.org/10.1167/14.12.4

Posner, M. I. (1980). Orienting of attention. *The Quarterly Journal of Experimental Psychology*, *32*(1), 3–25. https://doi.org/10.1080/00335558008248231

Pylyshyn, Z. W. (1980). Computation and cognition: Issues in the foundations of cognitive science. *Behavioral and Brain Sciences*, *3*(1), 111–132. https://doi.org/10.1017/S0140525X00002053

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna: R Foundation for Statistical Computing. Retrieved from https://www.r-project.org/

Reinagel, P., & Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, *10*, 341–350. https://doi.org/10.1088/0954-898X

Rensink, R. A. (2000). Seeing, sensing, and scrutinizing. *Vision Research*, *40*(10–12), 1469–1487. https://doi.org/10.1016/S0042-6989(00)00003-1

Rider, A. T., Coutrot, A., Pellicano, E., Dakin, S. C., & Mareschal, I. (2018). Semantic content outweighs low-level saliency in determining children's and adults' fixation of movies. *Journal of Experimental Child Psychology*, *166*(February), 293–309. https://doi.org/10.1016/j.jecp.2017.09.002

Rolfs, M. (2015). Attention in Active Vision: A Perspective on Perceptual Continuity Across Saccades. *Perception*, *44*(8–9), 900–919. https://doi.org/10.1177/0301006615594965

Rose, D., & Bex, P. (2020). The Linguistic Analysis of Scene Semantics: LASS. *Behavior Research Methods*, *52*(6), 2349–2371. https://doi.org/10.3758/s13428-020-01390-8

Rosenholtz, R. (2016). Capabilities and Limitations of Peripheral Vision. *Annual Review of Vision Science*, *2*, 437–457. https://doi.org/10.1146/annurev-vision-082114-035733

Rösler, L., End, A., & Gamer, M. (2017). Orienting towards social features in naturalistic scenes is reflexive. *PLoS ONE*, *12*(7), 1–14. https://doi.org/10.1371/journal.pone.0182037

Rossi, A. F., & Paradiso, M. A. (1995). Feature-specific effects of selective visual attention. *Vision Research*, *35*(5), 621–634. https://doi.org/10.1016/0042-6989(94)00156-G

Rothkegel, L. O. M., Schütt, H. H., Trukenbrod, H. A., Wichmann, F. A., & Engbert, R. (2019). Searchers adjust their eye-movement dynamics to target characteristics in natural scenes. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-018-37548-w

Rothkegel, L. O. M., Trukenbrod, H. A., Schütt, H. H., Wichmann, F. A., & Engbert, R. (2017). Temporal evolution of the central fixation bias in scene viewing. *Journal of Vision*, *17*(13), 1–18. https://doi.org/10.1167/17.13.3

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian t tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin and Review*, *16*(2), 225–237. https://doi.org/10.3758/PBR.16.2.225

Sadeghi, Z., McClelland, J. L., & Hoffman, P. (2015). You shall know an object by the company it keeps: An investigation of semantic representations derived from object co-occurrence in visual scenes. *Neuropsychologia*, *76*, 52–61. https://doi.org/10.1016/j.neuropsychologia.2014.08.031

Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R., & Wichmann, F. A. (2019). Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *Journal of Vision*, *19*(3), 1–25. https://doi.org/10.1167/19.3.1

Schwartz, E. L. (1994). Computational Studies of the Spatial Architecture of Primate Visual Cortex. In *Primary Visual Cortex in Primates*. Springer Science.

Shuirmann, D. J. (1987). A comparison of the two one-sided tests procedure and the power. *Journal of Pharmacokinetics and Biopharmaceutics*, *15*(6), 657–680.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings*, 1–14.

Sloan, L. L. (1961). Area and luminance of test object as variables in examination of the visual field by projection perimetry. *Vision Research*, *1*(1–2). https://doi.org/10.1016/0042-6989(61)90024-4

Stewart, E. E. M., Valsecchi, M., & Schütz, A. C. (2020). A review of interactions between peripheral and foveal vision. *Journal of Vision*, *20*(12), 2. https://doi.org/10.1167/jov.20.12.2

Stoll, J., Thrun, M., Nuthmann, A., & Einhäuser, W. (2015). Overt attention in natural scenes: Objects dominate features. *Vision Research*, *107*, 36–48. https://doi.org/10.1016/j.visres.2014.11.006

Storrs, K. R., & Kriegeskorte, N. (2019). *Deep Learning for Cognitive Neuroscience*. Retrieved from http://arxiv.org/abs/1903.01458

Tatler, B. W. (2007). The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, *7*(14). https://doi.org/10.1167/7.14.4

Tatler, B. W., Baddeley, R. J., & Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision Research*, *45*(5), 643–659. https://doi.org/10.1016/j.visres.2004.09.017

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011a). Eye guidance in natural vision: reinterpreting salience. *Journal of Vision*, *11*(5), 5. https://doi.org/10.1167/11.5.5

Tatler, B. W., & Vincent, B. T. (2008). Systematic tendencies in scene viewing. *Journal of Eye Movement Research*, *2*(2), 1–18. https://doi.org/10.16910/jemr.2.2.5

Teufel, C., Dakin, S. C., & Fletcher, P. C. (2018a). Prior object-knowledge sharpens properties of early visual feature-detectors. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-28845-5

Teufel, C., & Fletcher, P. C. (2020). Forms of prediction in the nervous system. *Nature Reviews Neuroscience*, *21*(4), 231–242. https://doi.org/10.1038/s41583-020-0275-5

Teufel, C., & Nanay, B. (2016). How to (and how not to) think about top-down influences on visual perception. *Consciousness and Cognition*, *47*(February), 1–4. https://doi.org/10.1016/j.concog.2016.05.008

Teufel, C., Subramaniam, N., Dobler, V., Perez, J., Finnemann, J., Mehta, P. R., … Fletcher, P. C. (2015). Shift toward prior knowledge confers a perceptual advantage in early psychosis and psychosis-prone healthy individuals. *Proceedings of the National Academy of Sciences*, *112*(43), 13401–13406. https://doi.org/10.1073/pnas.1503916112

The jamovi project. (2020). *jamovi*. Retrieved from https://www.jamovi.org

Thomas, C. (2016). *OpenSalicon: An Open Source Implementation of the Salicon Saliency Model*. Retrieved from http://arxiv.org/abs/1606.00110

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

Torralba, A., & Oliva, A. (2001). Modeling the shape of the scene: a holistic representation of the spatial envelope. *International Journal of Computer Vision*, *42*(3), 145–175.

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

Treder, M. S., Mayor-Torres, J., & Teufel, C. (2020). *Deriving Visual Semantics from Spatial Context: An Adaptation of LSA and Word2Vec to generate Object and Scene Embeddings from Images*. Retrieved from http://arxiv.org/abs/2009.09384

Treisman, A. (1985). Preattentive processing in vision. *Computer Vision, Graphics and Image Processing*, *31*(2), 156–177. https://doi.org/10.1016/S0734-189X(85)80004-9

Treisman, A., & Gormican, S. (1988). Feature Analysis in Early Vision: Evidence From Search Asymmetries. *Psychological Review*, *95*(1), 15–48. https://doi.org/10.1037/0033-295X.95.1.15

Treisman, A. M., & Gelade, G. (1980). A Feature-Integration Theory of Attention. *Cognitive Psychology*, *12*, 97–136.

Tseng, P. H., Carmi, R., Cameron, I. G. M., Munoz, D. P., & Itti, L. (2009). Quantifying center bias of observers in free viewing of dynamic natural scenes. *Journal of Vision*, *9*(7), 1–16. https://doi.org/10.1167/9.7.4

Tulver, K., Aru, J., Rutiku, R., & Bachmann, T. (2019). Individual differences in the effects of priors on perception: A multi-paradigm approach. *Cognition*, *187*(September 2018), 167–177. https://doi.org/10.1016/j.cognition.2019.03.008

Uejima, T., Niebur, E., & Etienne-Cummings, R. (2020). Proto-Object Based Saliency Model With Texture Detection Channel. *Frontiers in Computational Neuroscience*, *14*(September). https://doi.org/10.3389/fncom.2020.541581

Van der Linden, L., Mathôt, S., & Vitu, F. (2015). The role of object affordances and center of gravity in eye movements toward isolated daily-life objects. *Journal of Vision*, *15*(5), 1–18. https://doi.org/10.1167/15.5.8

van Renswoude, D. R., Raijmakers, M. E. J., & Visser, I. (2020). Looking (for) patterns: Similarities and differences between infant and adult free scene-viewing patterns. *Journal of Eye Movement Research*, *13*(1), 1–20. https://doi.org/10.16910/jemr.13.1.2

Verhallen, R. J., & Mollon, J. D. (2016). A new Mooney test. *Behavior Research Methods*, *48*(4), 1546–1559. https://doi.org/10.3758/s13428-015-0666-0

Vincent, B. T., Baddeley, R., Correani, A., Troscianko, T., & Leonards, U. (2009). Do we look at lights? Using mixture modelling to distinguish between low- and high-level factors in natural image viewing. *Visual Cognition*, *17*(6–7), 856–879. https://doi.org/10.1080/13506280902916691

Võ, M. L.-H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210. https://doi.org/10.1016/j.copsyc.2019.03.009

Võ, M. L. H. (2021). The meaning and structure of scenes. *Vision Research*, *181*(August 2019), 10–20. https://doi.org/10.1016/j.visres.2020.11.003

Wang, H.-C., Hwang, A. D., & Pomplun, M. (2010). Object Frequency and Predictability Effects on Eye Fixation Durations in Real-World Scene Viewing. *Journal of Eye Movement Research*, *3*(3), 1–10. https://doi.org/10.16910/jemr.3.3.3

Wang, H. C., & Pomplun, M. (2012). The attraction of visual attention to texts in real-world scenes. *Journal of Vision*, *12*(6), 1. https://doi.org/10.1167/12.6.1

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., … Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, *4*(43), 1686. https://doi.org/10.21105/joss.01686

Williams, C. C., & Castelhano, M. S. (2019). The changing landscape: High-level influences on eye movement guidance in scenes. *Vision (Switzerland)*, *3*(3), 1–20. https://doi.org/10.3390/vision3030033

Wilming, N., Betz, T., Kietzmann, T. C., & König, P. (2011). Measures and Limits of Models of Fixation Selection. *PLoS ONE*, *6*(9), e24038. https://doi.org/10.1371/journal.pone.0024038

Wilming, N., Kietzmann, T. C., Jutras, M., Xue, C., Treue, S., Buffalo, E. A., & König, P. (2017). Differential contribution of low- And high-level image content to eye movements in monkeys and humans. *Cerebral Cortex*, *27*(1), 279–293. https://doi.org/10.1093/cercor/bhw399

Wilming, N., Onat, S., Ossandón, J. P., Açik, A., Kietzmann, T. C., Kaspar, K., … König, P. (2017). An extensive dataset of eye movements during viewing of complex images. *Scientific Data*, *4*, 1–11. https://doi.org/10.1038/sdata.2016.126

Wloka, C., Kotseruba, I., & Tsotsos, J. K. (2018). Active fixation control to predict saccade sequences. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3184-3193. https://doi.org/10.1109/CVPR.2018.00336

Wloka, C., Kunić, T., Kotseruba, I., Fahimi, R., Frosst, N., Bruce, N. D. B., & Tsotsos, J. K. (2018). SMILER: Saliency Model Implementation Library for Experimental Research. *ArXiv*.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Publishing Group*, *1*(March), 1–8. https://doi.org/10.1038/s41562-017-0058

Wu, C.-C., Wang, H. C., & Pomplun, M. (2014). The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes. *Vision Research*, *105*, 10–20. https://doi.org/10.1016/j.visres.2014.08.019

Wu, C. C., Wang, H. C., & Pomplun, M. (2014). The roles of scene gist and spatial dependency among objects in the semantic guidance of attention in real-world scenes. *Vision Research*, *105*, 10–20. https://doi.org/10.1016/j.visres.2014.08.019

Wu, C. C., Wick, F. A., & Pomplun, M. (2014). Guidance of visual attention by semantic information in real-world scenes. *Frontiers in Psychology*, *5*(FEB). https://doi.org/10.3389/fpsyg.2014.00054

Wu, E. X. W. W., Gilani, S. O., van Boxtel, J. J. A. J., Amihai, I., Chua, F. K., & Yen, S.-C. C. (2013). Parallel programming of saccades during natural scene viewing: evidence from eye movement positions. *Journal of Vision*, *13*(12), 17. https://doi.org/10.1167/13.12.17

Wurtz, R. H., Joiner, W. M., & Berman, R. A. (2011). Neuronal mechanisms for visual stability: Progress and problems. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *366*(1564), 492–503. https://doi.org/10.1098/rstb.2010.0186

Wynn, J. S., Ryan, J. D., & Buchsbaum, B. R. (2020). Eye movements support behavioral pattern completion. *Proceedings of the National Academy of Sciences of the United States of America*, *117*(11), 6246–6254. https://doi.org/10.1073/pnas.1917586117

Wynn, J. S., Shen, K., & Ryan, J. D. (2019). Eye movements actively reinstate spatiotemporal mnemonic content. *Vision (Switzerland)*, *3*(2), 1–19. https://doi.org/10.3390/vision3020021

Xu, J., Jiang, M., Wang, S., Kankanhalli, M. S., & Zhao, Q. (2014). Predicting human gaze beyond pixels. *Journal of Vision*, *14*(1), 1–20. https://doi.org/10.1167/14.1.28

Yarbus, A. L. (1967). *Eye Movements and Vision*. New York: Plenum Press.

Zelinsky, G. J., & Bisley, J. W. (2015). The what, where, and why of priority maps and their interactions with visual working memory. *Annals of the New York Academy of Sciences*, *1339*(1), 154–164. https://doi.org/10.1111/nyas.12606

Zhang, H., Anderson, N. C., & Miller, K. F. (2020). *Mind-wandering during Scene Perception: On the Role of Meaning and Salience*. https://doi.org/https://doi.org/10.31234/osf.io/9fc2u

Zhang, L., Tong, M. H., Marks, T. K., Shan, H., Cottrell, G. W., Tong, M. H., … Cottrell, G. W. (2008). SUN: A Bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7), 1–20. https://doi.org/10.1167/8.7.32