Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin

Extracting scene and object information from natural stimuli: the influence of scene structure and eye movements

Dissertation

zur Erlangung des akademischen Grades

Doktor(in) der Naturwissenschaften (Dr. rer. nat.)

vorgelegt von

M.Sc.
Häberle, Greta

Berlin, März 2023

First Reviewer: Prof. Dr. Radoslaw M. Cichy
Second Reviewer: Prof. Dr. Peter König

Date of Defense: August 22nd 2023

Acknowledgements

First and foremost would like to express my gratitude to my supervisor, Radoslaw Cichy, for his guidance and feedback throughout the course of my doctoral research. I appreciate the time and effort you put into providing me with constructive criticism and pushing me to improve my work.

I would like to thank Daniel Kaiser for making my first lab rotation more productive than I could have ever imagined. I thank Prof. Felix Blankenburg and Prof. Rasha Abdel Rahman for their valuable input and feedback during the Einstein Center Supervision Meeting. I thank Prof. Annette Kinder, Prof. Daniel Kaiser, Prof. Peter König, and Dr. Marleen Haupt for kindly agreeing to join my doctoral committee. I thank the Einstein Center for Neurosciences for their financial support.

I would like to thank the people from the Neurodynamics of Visual Cognition Lab for their open ears throughout these years. Thank you Daniela Satici-Thies, for your dedication to solving all administrative struggles. A special thanks goes to Marleen and Agnessa. Completing a Ph.D. during normal times is challenging, completing a Ph.D. during a pandemic would have been impossible without you. Thank you for your constant support, guidance, proofreading, and open ears.

Thank you Silja, Mariella, Geeske and Becca for constantly supporting me from afar and taking my mind off of work.

Felix - this thesis would not have happened without you. Thank you for your support, your empathy, your abundance of patience, and clear words when they were needed. Thank you, Thor and Loki, for making our lives brighter.

Zu guter letzt möchte ich mich noch bei meinen Eltern und meinem Bruder bedanken. Danke, dass ihr mich durch die vielen Jahre des Studiums und der Promotion begleitet habt und niemals den Glauben an mich aufgegeben habt.

Abstract

When we observe a scene in our daily lives, our brains seemingly effortlessly extract various aspects of that scene. This can be attributed to different aspects of the human visual system, including but not limited to (1) its tuning to natural regularities in scenes and (2) its ability to bring different parts of the visual environment into focus via eye movements. While eye movements are a ubiquitous and natural behavior, they are considered undesirable in many highly controlled visual experiments. Participants are often instructed to fixate but cannot always suppress involuntary eye movements, which can challenge the interpretation of neuroscientific data, in particular for magneto- and electroencephalography (M/EEG).

This dissertation addressed how scene structure and involuntary eye movements influence the extraction of scene and object information from natural stimuli. First, we investigated when and where real-world scene structure affects scene-selective cortical responses. Second, we investigated whether spatial structure facilitates the temporal analysis of a scene's categorical content. Third, we investigated whether the spatial content of a scene aids in extracting task-relevant object information. Fourth, we explored whether the choice of fixation cross influences eye movements and the classification of natural images from EEG and eye tracking. The first project showed that spatial scene structure impacts scene-selective neural responses in OPA and PPA, revealing genuine sensitivity to spatial scene structure starting from 255 ms, while scene-selective neural responses are less sensitive to categorical scene structure. The second project demonstrated that spatial scene structure facilitates the extraction of the scene's categorical content within 200 ms of vision. The third project showed that coherent scene structure facilitates the extraction of object information if the object is task-relevant, suggesting a task-based modulation. The fourth project showed that choosing a centrally presented bullseye instead of a standard fixation cross reduces eye movements on the single image level and subtly removes systematic eye movement related activity in M/EEG data. Taken together, the results advanced our understanding of (1) the impact of real-world structure on scene perception as well as the extraction of object information and (2) the influence of eye movements on advanced analysis methods.

Zusammenfassung

Wenn wir in unserem täglichen Leben eine Szene beobachten, extrahiert unser Gehirn scheinbar mühelos verschiedene Aspekte dieser Szene. Dies kann auf verschiedene Aspekte des menschlichen Sehsystems zurückgeführt werden, unter anderem auf (1) seine Ausrichtung auf natürliche Regelmäßigkeiten in Szenen und (2) seine Fähigkeit, verschiedene Teile der visuellen Umgebung durch Augenbewegungen in den Fokus zu bringen. Obwohl Augenbewegungen ein allgegenwärtiges und natürliches Verhalten sind, werden sie in vielen stark kontrollierten visuellen Experimenten als unerwünscht angesehen. Die Teilnehmer werden oft angewiesen, zu fixieren, können aber unwillkürliche Augenbewegungen nicht immer unterdrücken, was die Interpretation neurowissenschaftlicher Daten, insbesondere der Magneto- und Elektroenzephalographie (M/EEG), in Frage stellen kann.

In dieser Dissertation wurde untersucht, wie Szenenstruktur und unbewusste Augenbewegungen die Extraktion von Szenen- und Objektinformationen aus natürlichen Stimuli beeinflussen. Zunächst untersuchten wir, wann und wo die Struktur einer realen Szene die szenenselektiven kortikalen Reaktionen beeinflusst. Zweitens untersuchten wir, ob die räumliche Struktur die zeitliche Analyse des kategorialen Inhalts einer Szene erleichtert. Drittens untersuchten wir, ob der räumliche Inhalt einer Szene bei der Extraktion aufgabenrelevanter Objektinformationen hilft. Viertens untersuchten wir, ob die Wahl des Fixationskreuzes die Augenbewegungen und die Klassifizierung natürlicher Bilder aus EEG und Eye-Tracking beeinflusst. Das erste Projekt zeigte, dass sich die räumliche Szenenstruktur auf szenenselektive neuronale Reaktionen in OPA und PPA auswirkt, wobei eine echte Empfindlichkeit für räumliche Szenenstrukturen ab 255 ms festgestellt wurde, während szenenselektive neuronale Reaktionen weniger empfindlich auf kategoriale Szenenstrukturen reagieren. Das zweite Projekt zeigte, dass die räumliche Szenenstruktur die Extraktion des kategorialen Inhalts der Szene innerhalb von 200 ms nach dem Sehen erleichtert. Das dritte Projekt zeigte, dass eine kohärente Szenenstruktur die Extraktion von Objektinformationen erleichtert, wenn das Objekt aufgabenrelevant ist, was auf eine aufgabenbezogene Modulation hindeutet. Das vierte Projekt zeigte, dass die Wahl eines zentral präsentierten Bullauges anstelle eines Standard-Fixationskreuzes Augenbewegungen auf Einzelbildebene reduziert und systematische Augenbewegungsaktivität in M/EEG-Daten auf subtile Weise beseitigt. Zusammengenommen haben die Ergebnisse unser Verständnis (1) der Auswirkungen der Struktur der realen Welt auf die Wahrnehmung der Szene und die Extraktion von Objektinformationen und (2) des Einflusses von Augenbewegungen auf fortgeschrittene Analysemethoden verbessert.

Contents

Acknowledgements **3**

Abstract **4**

Zusammenfassung **5**

List of Abbreviations **8**

1 General introduction **9**

1.1 A general introduction to object recognition 9

1.1.1 Interim summary 10

1.2 A general introduction to scene processing 10

1.2.1 Interim summary 13

1.3 A short definition of eye movements 13

1.3.1 Interim summary 14

1.4 The effect of eye movements on the analysis of neuroimaging data 14

1.4.1 Interim summary 16

1.5 Aim of this thesis 17

2 Project I: Cortical sensitivity to natural scene structure **18**

3 Project II: Real-world structure facilitates the rapid emergence of scene category information in visual brain signals **29**

4 Project III: Coherent natural scene structure facilitates the extraction of task-relevant object information in visual cortex **37**

5 Project IV: The influence of the bullseye versus standard fixation cross on eye movements and classifying natural images from EEG **46**

6 General Discussion **68**

6.1 Summary 68

6.2 Main findings 68

6.2.1 Project I 68

6.2.2 Project II 69

6.2.3 Project III 70

6.2.4 Project IV 71

6.3 Key implications across projects 71

6.3.1 The temporal dynamics of scene processing 72
6.3.2 The spatial dynamics of scene processing 73
6.3.3 The interplay between the object and scene network 75
6.3.4 The effect of eye movements on object and scene processing . . . 76
6.4 Methodological considerations and future directions 78
6.4.1 Experimental design considerations 78
6.4.2 Methodological considerations 79
6.5 Conclusion 81

References **82**

Supplementary Material **92**

Appendix **115**
List of publications 115
Summary of main results 116
Zusammenfassung der Ergebnisse 119
Declaration of author contributions 123
Eidesstattliche Versicherung/Affidavit 126

List of Abbreviations

EBA	Extrastriate Body Area
EEG	Electroencephalography
ERP	Event Related Potentials
EVC	Early Visual Cortex
fMRI	Functional Magnetic Resonance Imaging
ICA	Independent Component Analysis
LOC	lateral occipital complex
MEG	Magnetoencephalography
MVPA	Multivariate Pattern Analysis
OPA	Occipital Place Area
PPA	Parahippocampal Place Area
RSC	Retrosplinal Complex

Chapter 1

General introduction

The overarching aim of this thesis is to understand how scene structure and involuntary eye movements influence the extraction of scene and object information from natural stimuli. The general introduction will introduce the concepts necessary to achieve this goal. Therefore, the first section gives a general overview of object recognition and the difficulties and pitfalls associated with it. The second section introduces scene perception and explains how object recognition and scene perception are inherently interconnected. The third section explains the need to understand eye movements as a window into our cognitive world and establishes problems and considerations that need to be taken into account when investigating visual experiments in a lab environment. The general introduction concludes with the aim of this thesis and the derived research questions for the four studies presented in the core part of this thesis.

1.1 A general introduction to object recognition

During our daily lives, we constantly solve tasks such as recognizing the coffee cup in front of us. These tasks are so effortlessly and automatically that in the mind of most people, they would not even qualify as tasks. However, different complex processing steps are required to solve this seemingly easy problems. To be able to identify an object reliably, a person must recognize said object under different lighting conditions, viewpoints, and in front of different backgrounds in a fraction of a second (Logothetis & Sheinberg, 1996). Adding to the complexity of the problem, objects from the same category present a multitude of different features, e.g., a mug can have numerous different shapes or colors. This problem is called the invariance problem (DiCarlo & Cox, 2007).

Decades of research have established that the invariance problem is solved primarily in the ventral visual stream through a cascade of largely feedforward computations concluding in object representations in the inferior temporal cortex (IT) (DiCarlo et al., 2012; Goodale & Milner, 1992; Mishkin et al., 1983). Numerous studies have identified several object-selective areas in the inferior temporal cortex that are preferentially activated by specific object categories. These include the ventral fusiform gyrus as well as the lateral superior and middle temporal gyri for animals and tools (Chao et al., 1999; Martin et al., 1996), the fusiform face area for faces (Kanwisher et al., 1997), and the extrastriate body area (EBA) in the lateral occipitotemporal cortex for body parts (Downing et al., 2001).

However, the categorical representations of these objects are not strictly limited to specific areas. Representations often overlap across regions in the ventral temporal areas. Patterns that discriminate between different object categories could even be found within cortical regions whose maximal response was to only one category (Haxby et al., 2001). This research suggests that the ventral visual pathway does not contain purely category-specific modules but instead forms a continuous representation of information about objects (Ishai et al., 1999).

However, to form a comprehensive understanding of object processing, both, spatial and temporal information needs to be investigated. Even though object recognition is such a complex process, magnetoencephalography (MEG) and electroencephalography (EEG) studies have shown that within the first 100 ms individual object exemplars can be classified. By 240 ms, a clear categorical distinction between animate and inanimate objects evolved (Carlson et al., 2013; Cichy et al., 2014; Contini et al., 2017). A seminal study combined spatial and temporal information and showed that early signals correlated more with early visual cortex V1 while later stages correlated more with IT, indicating a cascade of computations in the ventral visual stream (Cichy et al., 2014).

1.1.1 Interim summary

Decades of neuroscientific research have revealed that the complex process of object processing is achieved within a few hundred milliseconds along the ventral visual stream. Several areas in the inferior temporal cortex selectively respond to specific object categories. Overall, object recognition is one of the core abilities that human beings use to navigate and understand a complex world.

1.2 A general introduction to scene processing

During everyday life, we rarely encounter individual objects in isolation. Most of the time, these objects are embedded in larger contexts (Oliva & Torralba, 2007). Epstein (2005) contrasted object perception to so-called scene perception using the following definition: “[S]cene perception can be usefully contrasted to object perception: whereas objects are spatially compact entities that one acts upon, scenes are spatially distributed entities that one acts within“.

Even though scenes are complex compositions, consisting of several objects as well as fore- and backgrounds, humans are able to efficiently extract information about objects and the gist of a scene within a few hundred milliseconds, or - in other words - within a single glance (Potter, 1975; Thorpe et al., 1996). This gist of a scene includes the scene’s basic level category (e.g., natural versus artificial), as well as an estimate of the basic feature distribution (Oliva & Schyns, 2000; Rousselet et al., 2005), allowing observers to classify the content of a scene within a few hundred milliseconds. This classification allows observers to rapidly judge, e.g., whether a scene contains an animal or a tool (Li et al., 2002).

It was long assumed that visual exploration and guided search of individual objects within a scene is necessary to understand the content of a scene as our object recognition processes are limited to very few objects at a time and are guided by low-level stimulus properties (Wolfe, 2007). However, the classic guided search cannot account for the rapid conceptualization of scenes under real-world circumstances. One partial explanation for this rapid conceptualization can be found in the fact that all real-world scenes follow predictable statistical regularities. Spatial regularities are one example of these predictable statistical regularities. The spatial context of a scene aids human participants in correctly identifying objects within a scene. The importance of the spatial context for the recognition and identification of objects was demonstrated early on by a so-called jumbling paradigm. A typical scene was divided into six quadrants and then rearranged while the rotation of the individual quadrants was kept constant. Jumbling reduced the accuracy of identifying an object within a scene and the accuracy of scene identification. Therefore, the perception of a scene seems to be more than the sum of the individual parts of that scene (Biederman, 1972; Biederman et al., 1974).

Spatial regularities particularly aid scene recognition and the recognition of objects within a scene. They provide a framework of where specific objects are most commonly found in a scene, e.g., planes are usually located in the sky, whereas cars are usually on roads. This effect has not only been demonstrated for objects (Kaiser et al., 2018) but also for faces and bodies (Chan et al., 2010; de Haas et al., 2016) and translated into behavioral recognition advantages (Quek & Peelen, 2020). This indicates that the human experience with real-world regularities leads to enhanced perceptual processing of objects when they appear at their predicted absolute locations. Several neuroimaging studies have corroborated that the real-world structure of scenes impacts visual cortex responses to everyday objects (Kaiser et al., 2018; Kaiser & Peelen, 2018; Kim & Biederman, 2011) and facilitated cortical processing in the object-selective lateral occipital cortex (LOC) and early visual cortex (Kaiser & Cichy, 2018).

While the disruption of spatial regularities mainly disrupts the global position of objects within a scene, their positioning can also be described in relative terms (Hock et al., 1974; Oliva & Torralba, 2007). It is statistically more likely, e.g., for chairs to be grouped around a table and then for them to be grouped around a trash can. These statistical regularities, like the global positioning of objects, affect visual processing. Both behavioral and neuroimaging studies have supported this. Participants are faster in detecting objects positioned in typical relative positions versus atypical relative positions toward each other (Hock et al., 1974; Stein et al., 2015) and greater activity can be found LOC when objects interact with each other instead of simply being described side by side (Kim & Biederman, 2011). This transition from individual to integrative object processing emerges along the posterior-anterior axis of the visual cortex (Kaiser & Peelen, 2018).

Apart from the local and global positioning of objects within a scene, semantic information has been shown to guide an observer (e.g. to determine whether a kitchen or a garden is observed). These semantic concepts do not seem to arise from categorizing each individual object in a scene but rather from the above-mentioned gist of a scene. Contextual information has been shown to facilitate object processing (Bar, 2004) via an efficient allocation of attention (Torralba et al., 2006; Võ et al., 2019; Wolfe et al., 2011) but also disambiguates object information under uncertainty (Brandman & Peelen, 2017; Oliva & Torralba, 2007).

This raises the question, of which mechanisms facilitate this rapid extraction of contextual information. Wolfe et al., 2011 suggested that a two-pathway architecture could explain the rapid contextualization of scene content. Based on this architecture, a non-selective pathway enables the rapid extraction of statistical information from an image. These statistical regularities, in turn, enable a certain amount of semantic processing in the visual system, such as extracting and categorizing scenes and basic spatial structures, but do not allow the precise recognition of objects. To be able to understand the full content of a scene, including its individual objects, an interaction between the non-selective and selective pathways is necessary. This selective pathway allows binding features and recognizing objects but is, therefore limited in processing capabilities (Wolfe et al., 2011). This theory facilitates the proposition that the gist of a scene is computed from the global properties of the scene instead of the linear recognition of objects within the scene (Greene & Oliva, 2009).

In sum, the visual brain adapted to spatial and contextual regularities in visual scenes, allowing fast and efficient processing of the global and local properties. However, which areas in the human brain are responsible for recognizing such scenes?

Several areas in the human brain aid explicitly with scene perception. More specifically, the parahippocampal place area (PPA) (Epstein & Kanwisher, 1998), the retrosplenial cortex (RSC) (RSC/MPA) (O'Craven & Kanwisher, 2000), and the Occipital Place Area (OPA) (Dilks et al., 2013; Hasson et al., 2003) are causally involved in the perception of scenes and strongly connected (Epstein & Baker, 2019). All three areas respond strongly to local spatial layout, e.g., an empty room (Epstein & Kanwisher, 1998) or a scene conveyed from Lego blocks (Epstein et al., 1999), suggesting a sensitivity toward scenes versus object-like layouts. More recently it has been proposed that human visual scene processing seems to be composed of at least two distinct systems in which PPA preferentially responds to the category of a scene instead of the location of a scene, and RSC and OPA show the exact opposite results, preferentially responding to the location of a scene (Persichetti & Dilks, 2018, 2019).

1.2.1 Interim summary

In sum, decades of research have shown a strong impact of real-world structure on the cortical processing of everyday objects and human beings. Local and global properties of a scene have a beneficial impact on extracting a scene's category and recognizing objects within that scene. Several seminal papers have shown that spatial and contextual regularities interact, leading to a meaningful distribution of visual content in real-world scenes. Consequently, cortical responses differ if scene elements violate that typical real-world structure. Taken together, these results show that the rapid extraction of scene content aids with the recognition and classification of objects and therefore demonstrate that object and scene processing mechanisms interact to enable the efficient processing of object and scene information.

1.3 A short definition of eye movements

Eye movements are integral to how we perceive the world and are used to explore scenes and objects (Schütz et al., 2011). They can roughly be divided into voluntary and involuntary eye movements. Most important for conscious human perception are voluntary eye movements, saccades, and smooth pursuit (Gegenfurtner, 2016). Big, voluntary eye movements (saccades) are used, e.g., to explore a scene. These saccades are rapid movements that abruptly change the point of fixation and are used to bring parts of a scene or image from the periphery into the fovea. These rapid movements are crucial for the exploration of an environment as the visual acuity quickly decreases when moving from the fovea to the periphery (Campbell & Green, 1965). The fovea only covers around 2 degrees of visual angle and receives disproportionally more cortical processing resources (Tootell et al., 1982). Saccades are ballistic movements achieved by three pairs of extraocular muscles attached to the eyeballs and allow horizontal, vertical, and diagonal movements (Walls, 1962). They elicit a cascade of activity from the retina via the visual cortex, the frontal eye fields, and the cerebellum to the oculomotor plant (Lisberger, 2010).

The underlying mechanisms that guide where saccades land still need to be fully understood. Saccades are guided by several different factors including saliency (Itti & Koch, 2000; Koch & Ullman, 1985), task demands (Thielen et al., 2019), spatial biases (Tatler, 2007), objects (Einhäuser et al., 2008), and geometric properties of the saccades (for differing opinions see Brockmann and Geisel, 1999 and Millidge and Shillcock, 2018). However, one example of open questions is the ongoing debate on whether objects or saliency are a better predictor of eye movements (Borji et al., 2013; Einhäuser et al., 2008). After a saccade is executed, fixations are used to keep the image stable in the foveal center. Nevertheless, even during fixations, the eyes never stay entirely still. They drift slowly with intermixed small involuntary eye movements (microsaccades), which are used to, e.g. keep the retinal image from fading out of perception (Rolfs, 2009).

1.3.1 Interim summary

Eye movements are integral to the exploration of our surroundings, and decades of researchers have dedicated their lives to their understanding. While they are opening a window into our cognitive world, they might also cloud our understanding of underlying cortical processes. During a saccade, the part of the retina that is exposed to a specific stimulus changes, and eye movement induced electric potentials might either vary systematically with the stimulus of interest or create artifacts in M/EEG data. To be able to understand whether and to which extent eye movements influence the results of neuroimaging studies, their effects need to be studied further.

1.4 The effect of eye movements on the analysis of neuroimaging data

As mentioned above, humans are able to process the gist of a scene within a single glance (Potter, 1975; Thorpe et al., 1996). However, this does not negate the importance of eye movements as an integral part of scene and object exploration (Schütz et al., 2011). If not instructed otherwise, people perform several saccades per second to bring the parts of the scene or object of interest into their foveal vision as the acuity of peripheral vision is limited. Nevertheless, in many neuroimaging studies, participants are instructed to fixate on a fixation cross or dot in the middle of the screen while viewing the stimulus. What is the reasoning behind this?

Evidence suggests that - both - large and small eye movements can lead to systematic and unsystematic effects in neuroimaging measurements such as EEG, MEG, or functional magnetic resonance imaging (fMRI) (Dijkstra et al., 2018; Dimigen et al., 2009; Mostert et al., 2018; Plöchl et al., 2012; Thielen et al., 2019). The neural activity measured with

EEG or MEG yields activity with amplitudes in the range of a few microvolts. It is, therefore, prone to masking from artifacts emerging from the eye, muscles, or electrical devices. These artifacts can be several magnitudes larger than the initial signal evoked from brain sources and might bury the signal in noise (Plöchl et al., 2012). Microsaccades as small as 0.15 degrees generate a field potential of around 100-150 ms after movement onset over the occipital cortex and mid-central scalp (Dimigen et al., 2009). An EEG can then pick up these field potentials as a saccadic spike potential (Thickbroom & Mastaglia, 1986). Therefore, even though humans inherently use their eyes to scan their surroundings efficiently, many neuroimaging studies control eye movements in one way or another.

Some experimental setup requires stimuli to be presented at specific locations on the retina (e.g. to investigate hemispheric or eccentricity-based differences in visual processing) while others might manipulate covert visual attention and therefore require the reduction of overt eye movements (Guzman-Martinez et al., 2009). Another reason altogether is to avoid excessive eye movements and the corresponding signal. Systematic effects of eye movements on M/EEG data have been found in several studies employing different analysis techniques. These include event-related potential (ERP) research (Dimigen & Ehinger, 2019; Dimigen et al., 2009), gamma band activity (Yuval-Greenberg et al., 2008), and multivariate pattern analysis (Dijkstra et al., 2018; Mostert et al., 2018; Quax et al., 2019; Thielen et al., 2019).

Several different methods are employed to avoid these confounds. Participants can be instructed ad-hoc to fixate on a fixation cross and to blink only during so-called catch trials. However, novice participants cannot accurately control their eye movements (Guzman-Martinez et al., 2009), and fixation behavior varies wildly even between experienced participants (Bargary et al., 2017; Guzman-Martinez et al., 2009; Thielen et al., 2019).

Another solution is removing eye movements post-hoc via algorithms like independent component analysis (ICA) or linear regression. Unfortunately, it has been shown that these algorithms are not sufficient to remove all residual signals generated by eye movements. One possible explanation is that these techniques assume a linear relation between the EEG and data and eye movements, whereas eye movements might induce strong non-linear effects (Quax et al., 2019).

Consequently, even if precautions are taken, eye movements might still confound the recorded data. There are at least two ways in which these large and small eye movements can affect neuroimaging measurements. On the one hand, if the eye movements are unsystematic and hence decrease the signal-to-noise ratio, they enhance the chance of incorrectly accepting the H0 (False negative, Type II error). On the other hand, if eye movements are systematic and co-vary with the experimental conditions, they increase the possibility of incorrectly rejecting the H0 (False positive, Type I error).

Depending on the techniques used for analysis, the influence of eye movement-related artifacts differs. This is, in part, due to the different computational steps used for analyses. Univariate M/EEG analyses evaluate differences in activation, by quantifying relative differences in average activity between experimental conditions. In comparison, multivariate analysis methods (e.g., multivariate pattern analysis (MVPA)) have the potential to examine differences in information, e.g., by comparing differences in distributed patterns of brain activation between experimental conditions (de-Wit et al., 2016; Grootswagers et al., 2017). For univariate analysis methods, trials are averaged within conditions and over electrodes to enhance the signal-to-noise ratio, in part, averaging out noise-related artifacts and reducing their effect on the data (Plöchl et al., 2012). However, saccades also elicit strong ERPs, related to the saccade offsets, which often temporally overlap with the condition of interest. While these ERPs are interesting in themselves, they also add complexity to the analysis pipeline by being convoluted with the stimulus ERP (Dimigen & Ehinger, 2019).

In contrast, multivariate analysis techniques analyze patterns of activation associated with experimental conditions from multiple voxels/sensors simultaneously. Multivariate methods, therefore, have the potential to detect differences in activation which are lost when averaging data for univariate analyses, making them more sensitive to patterns in the data that are relevant to the experimental task or stimulus. Consequently, multivariate analysis techniques are more sensitive to within-subject-trial-by-trial variance than their univariate counterpart (Carlson et al., 2003; Cox & Savoy, 2003; Grootswagers et al., 2017; Haxby et al., 2001; Haynes & Rees, 2006), increasing their sensitivity to stimulus-related eye movement confounds. In sum, both ERP research and MVPA are affected by eye movements, but to varying degrees.

1.4.1 Interim summary

Eye movements are integral to the exploration of our surroundings, and decades of researchers have dedicated their lives to their understanding. While they are opening a window into our cognitive world, they might also cloud our understanding of underlying cortical processes. During a saccade, the part of the retina that is exposed to a specific stimulus changes, and eye movement induced electric potentials might either vary systematically with the stimulus of interest or create artifacts in M/EEG data. To be able to understand whether and to which extent eye movements influence the results of neuroimaging studies, their effects need to be studied further.

1.5 Aim of this thesis

The overarching aim of this thesis is to further understand how the inherent structure of scenes in our world and involuntary eye movements influence the extraction of scene and object information from natural stimuli. To this end, we conducted four different studies using a mixture of EEG, fMRI, and eye tracking. We focused on answering four main questions:

Project I: Does real-world structure have an impact on scene-selective neural responses?

Project II: Does the spatial structure of a scene help facilitate the cortical analysis of the scene's categorical content?

Project III: Does the spatial structure of a scene's context aid in the extraction of task-relevant object information from the scene?

Project IV: Does the choice of fixation cross influence eye movements and the classification of natural images from EEG and eye tracking?

Chapter 2

Project I: Cortical sensitivity to natural scene structure

The current chapter comprises the research article entitled "Cortical sensitivity to natural scene structure" which was published in *Human Brain Mapping* in 2020. This first research project demonstrated that spatial (but not categorical) scene structure impacts cortical processing in scene-selective occipital and parahippocampal cortices and after 255ms, by accurately differentiating between spatially intact and jumbled scenes.

Authors:

Daniel Kaiser, Greta Häberle, Radoslaw M. Cichy

Contributions:

D. K. and R. M. C. designed research, D. K. and G. H. acquired data, D. K. and G. H. analyzed data, D. K., G. H., and R. M. C. interpreted results, D. K. prepared figures, D. K. drafted manuscript, D. K., G. H., and R. M. C. edited and revised manuscript.

Contributions to open and reproducible science:

To contribute to open and reproducible science, the paper is published in an open-access journal. The original article can be found here: doi: 10.1002/hbm.24875. Data are publicly available on OSF: doi: 10.17605/OSF.IO/ W9874.

Copyright note:

Human Brain Mapping is an open-access journal. All articles are published under a Creative Commons Attribution 4.0 International License and are free to re-use.

Received: 31 July 2019 | Revised: 7 November 2019 | Accepted: 7 November 2019

DOI: 10.1002/hbm.24875

WILEY

RESEARCH ARTICLE

Cortical sensitivity to natural scene structure

Daniel Kaiser[1,2] | **Greta Häberle**[2,3,4] | **Radoslaw M. Cichy**[2,3,4,5]

[1]Department of Psychology, University of York, York, UK

[2]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany

[3]Einstein Center for Neurosciences Berlin, Humboldt-Universität Berlin, Berlin, Germany

[4]Berlin School of Mind and Brain, Humboldt-Universität Berlin, Berlin, Germany

[5]Bernstein Center for Computational Neuroscience Berlin, Humboldt-Universität Berlin, Berlin, Germany

Correspondence
Daniel Kaiser, Department of Psychology, University of York, Heslington, York, YO10 5DD, UK.
Email: danielkaiser.net@gmail.com

Funding information
Deutsche Forschungsgemeinschaft, Grant/Award Numbers: CI241/1-1, CI241/3-1, KA4683/2-1; H2020 European Research Council, Grant/Award Number: ERC-2018-StG 803370

Abstract

Natural scenes are inherently structured, with meaningful objects appearing in predictable locations. Human vision is tuned to this structure: When scene structure is purposefully jumbled, perception is strongly impaired. Here, we tested how such perceptual effects are reflected in neural sensitivity to scene structure. During separate fMRI and EEG experiments, participants passively viewed scenes whose spatial structure (i.e., the position of scene parts) and categorical structure (i.e., the content of scene parts) could be intact or jumbled. Using multivariate decoding, we show that spatial (but not categorical) scene structure profoundly impacts on cortical processing: Scene-selective responses in occipital and parahippocampal cortices (fMRI) and after 255 ms (EEG) accurately differentiated between spatially intact and jumbled scenes. Importantly, this differentiation was more pronounced for upright than for inverted scenes, indicating genuine sensitivity to spatial structure rather than sensitivity to low-level attributes. Our findings suggest that visual scene analysis is tightly linked to the spatial structure of our natural environments. This link between cortical processing and scene structure may be crucial for rapidly parsing naturalistic visual inputs.

KEYWORDS

EEG, fMRI, multivariate decoding, scene representation, spatial structure, visual perception

1 | INTRODUCTION

Humans can efficiently extract information from natural scenes even from just a single glance (Potter, 1975; Thorpe, Fize, & Marlot, 1996). A major reason for this perceptual efficiency lies in the structure of natural scenes: for instance, a scene's spatial structure tells us where specific objects can be found and its categorical structure tells us which objects are typically encountered within the scene (Kaiser, Quek, Cichy, & Peelen, 2019; Oliva & Torralba, 2007; Võ, Boettcher, & Draschkow, 2019; Wolfe, Võ, Evans, & Greene, 2011).

The beneficial impact of scene structure on perception becomes apparent in jumbling paradigms, where the scene's structure is purposefully disrupted by shuffling blocks of information across the scene. For instance, jumbling makes it harder to categorize scenes (Biederman, Rabinowitz, Glass, & Stacy, 1974), recognize objects within them (Biederman, 1972; Biederman, Glass, & Stacy, 1973) or to detect subtle visual changes (Varakin & Levin, 2008; Zimmermann, Schnier, & Lappe, 2010). These findings suggest that typical scene structure contributes to efficiently perceiving a scene and its contents.

Such perceptual effects prompt the hypothesis that scene structure also impacts perceptual stages of cortical scene processing. However, while there is evidence that real-world structure impacts visual cortex responses to everyday objects (Kaiser & Cichy, 2018; Kaiser & Peelen, 2018; Kim & Biederman, 2011; Roberts & Humphreys, 2010) and human beings (Bernstein, Oron, Sadeh, & Yovel, 2014; Brandman & Yovel, 2016; Chan, Kravitz, Truong, Arizpe, & Baker, 2010), it is unclear whether real-world structure has a similar impact on scene-selective neural responses.

To answer this question, we conducted multivariate pattern analysis (MVPA) and univariate analyses on fMRI and EEG responses to intact and jumbled scenes, which allowed us to spatially and temporally resolve whether cortical scene processing is indeed sensitive to scene structure. During the fMRI and EEG experiments, participants viewed scene images in which we manipulated two facets of natural scene structure: We orthogonally jumbled the scene's spatial structure (i.e., whether the scene's parts appear in their typical positions or not) or its categorical structure (i.e., whether the scene's parts belong to the same category or different categories).

Our results provide three key insights into how scene structure affects scene representations: (a) Cortical scene processing is primarily sensitive to the scene's spatial structure, more so than to the scene's categorical structure. (b) Spatial structure impacts the perceptual analysis of scenes, in occipital and parahippocampal cortices (Epstein, 2014) and shortly after 200 ms (Harel, Groen, Kravitz, Deouell, & Baker, 2016). (c) Spatial structure impacts cortical responses more strongly for upright than inverted scenes, indicating robust sensitivity to spatial scene structure that goes beyond sensitivity to low-level features.

2 | MATERIALS AND METHODS

2.1 | Participants

In the fMRI experiment, 20 healthy adults participated in session 1 (mean age 25.5, *SD* = 4.0; 13 female) and 20 in session 2 (mean age 25.4, *SD* = 4.0; 12 female). Seventeen participants completed both sessions, three participants only session 1 or session 2, respectively. In the EEG experiment, 20 healthy adults (mean age 26.6, *SD* = 5.8; 9 female) participated in a single session. Samples sizes were determined based on typical samples sizes in related research; a sample of *N* = 20 yields 80% power for detecting effects sizes greater than *d* = 0.66.[1] All participants had normal or corrected-to-normal vision. Participants provided informed consent and received monetary reimbursement or course credits. All procedures were approved by the ethical committee of Freie Universität Berlin and were in accordance with the Declaration of Helsinki.

2.2 | Stimuli and design

Stimuli were 24 scenes from four different categories (church, house, road, supermarket; Figure 1a), taken from an online resource (Konkle, Brady, Alvarez, & Oliva, 2010); the complete scene image set can be found in the Appendix S1. We split each image into quadrants and systematically recombined the resulting parts in a 2 × 2 design, where both the scenes' spatial structure and their categorical structure could be either intact or jumbled (Figure 1b,c). This yielded four conditions:

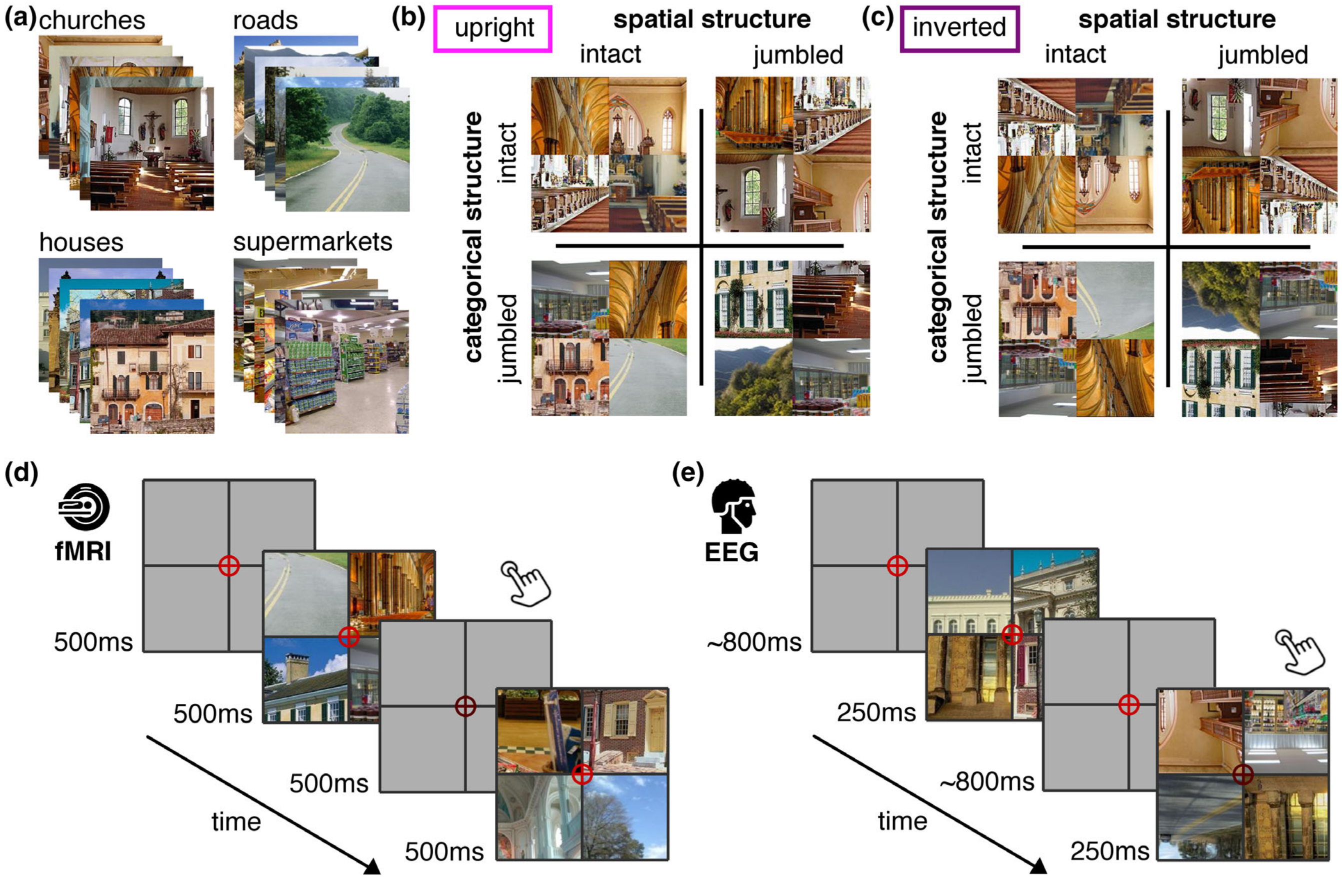

FIGURE 1 Stimuli and Paradigm. We combined parts from 24 scene images from four categories (a) to create a stimulus set where the scenes' structural (e.g., the spatial arrangements of the parts) and their categorical structure (e.g., the category of the parts) was orthogonally manipulated; all scenes were presented both upright and inverted (b, c). In the fMRI experiment, scenes were presented in a block design, where each block of 24 s exclusively contained scenes of a single condition (d). In the EEG experiment, all conditions were randomly intermixed (e). During both experiments, participants responded to color changes of the central crosshair [Color figure can be viewed at wileyonlinelibrary.com]

(a) In the "spatially intact & categorically intact" condition, parts from four scenes of the same category were combined in their correct locations. (b) In the "spatially intact & categorically jumbled" condition, parts from four scenes from different categories were combined in their correct locations. (c) In the "spatially jumbled & categorically intact" condition, parts from four scenes of the same category were combined, and their locations were exchanged in a crisscrossed way. (d) In the "spatially jumbled & categorically jumbled" condition, parts from four scenes from different categories were combined, and their locations were exchanged in a crisscrossed way. For each participant separately, 24 unique stimuli were generated for each condition by randomly drawing suitable fragments from different scenes.[2] During the experiment, all scenes were presented both upright and inverted.

2.3 | fMRI paradigm

The fMRI experiment (Figure 1d) comprised two sessions. In the first session, upright scenes were shown, in the second session inverted scenes were shown; the sessions were otherwise identical. Each session consisted of five runs of 10 min. Each run consisted of 25 blocks of 24 s. In 20 blocks, scene stimuli were shown with a frequency of 1 Hz (0.5 s stimulus, 0.5 s blank). Each block contained all 24 stimuli of a single condition. In five additional fixation-only blocks, no scenes were shown. Block order was randomized within every five consecutive blocks, which contained each condition (four scene conditions and fixation-only) exactly once.

Scene stimuli appeared in a black grid (4.5° visual angle), which served to mask visual discontinuities between quadrants. Participants were monitoring a central red crosshair, which twice per block (at random times) darkened for 50 ms; participants had to press a button when they detected a change. Participants on average detected 80.0% (*SE* = 2.5)[3] of the changes. Stimulus presentation was controlled using the Psychtoolbox (Brainard, 1997).

In addition to the experimental runs, each participant completed a functional localizer run of 13 min, during which they viewed images of scenes, objects, and scrambled scenes. The scenes were new exemplars of the four scene categories used in the experimental runs; objects were also selected from four categories (car, jacket, lamp, and sandwich). Participants completed 32 blocks (24 scene/object/scrambled blocks and 8 fixation-only blocks), with parameters identical to the experimental runs (24 s block duration, 1 Hz stimulation frequency, color change task).

2.4 | EEG paradigm

In the EEG experiment (Figure 1e), all conditions were randomly intermixed within a single session of 75 min (split into 16 runs). During each trial, a scene appeared for 250 ms, followed by an inter-trial interval randomly varying between 700 ms and 900 ms. In total, there were 3,072 trials (384 per condition), and an additional 1,152 target trials (see below).

As in the fMRI, stimuli appeared in a black grid (4.5° visual angle) with a central red crosshair. In target trials, the crosshair darkened during the scene presentation; participants had to press a button and blink when detecting this change. Participants on average detected 78.1% (*SE* = 3.6) of the changes. Target trials were not included in subsequent analyses.

2.5 | fMRI recording and preprocessing

MRI data was acquired using a 3 T Siemens Tim Trio Scanner equipped with a 12-channel head coil. T2*-weighted gradient-echo echo-planar images were collected as functional volumes (*TR* = 2 s, *TE* = 30 ms, 70° flip angle, 3mm^3 voxel size, 37 slices, 20% gap, 192 mm FOV, 64 × 64 matrix size, interleaved acquisition). Additionally, a T1-weighted anatomical image (MPRAGE; 1mm^3 voxel size) was obtained. Preprocessing was performed using SPM12 (www.fil.ion.ucl.ac.uk/spm/). Functional volumes were realigned, coregistered to the anatomical image, and normalized into MNI-305 space. Images from the localizer run were additionally smoothed using a 6 mm full-width-half-maximum Gaussian kernel.

2.6 | EEG recording and preprocessing

EEG signals were recorded using an EASYCAP 64-electrode[4] system and a Brainvision actiCHamp amplifier. Electrodes were arranged in accordance with the 10–10 system. EEG data was recorded at 1000 Hz sampling rate and filtered online between 0.03 Hz and 100 Hz. All electrodes were referenced online to the Fz electrode. Offline preprocessing was performed using FieldTrip (Oostenveld, Fries, Maris, & Schoffelen, 2011). EEG data were epoched from −200 ms to 800 ms relative to stimulus onset and baseline-corrected by subtracting the mean pre-stimulus signal. Channels and trials containing excessive noise were removed based on visual inspection. Blinks and eye movement artifacts were removed using independent component analysis and visual inspection of the resulting components. The epoched data were down-sampled to 200 Hz.

2.7 | fMRI region of interest definition

We restricted fMRI analyses to three regions of interest (ROIs): early visual cortex (V1), scene-selective occipital place area (OPA), and scene-selective parahippocampal place area (PPA) (Figure 2). We additionally localized scene-selective retrosplenial cortex (RSC), but did not observe reliable above-baseline activations to our scene stimuli in this region, all $t(19) < 0.14$, $p > .45$. The results for RSC can be found in the Appendix S1.

V1 was defined based on a functional group atlas (Wang et al., 2015), from which we selected all voxels that had a higher probability of belonging to V1 than belonging to another region in the atlas (905 voxels). Changing the number of voxels included did not qualitatively change the results in V1 (see Appendix S1).

Scene-selective ROIs were defined using the localizer data, which were modeled in a general linear model (GLM) with nine predictors (three regressors for the scene/object/scrambled blocks and six movement regressors). Scene-selective ROI definition was

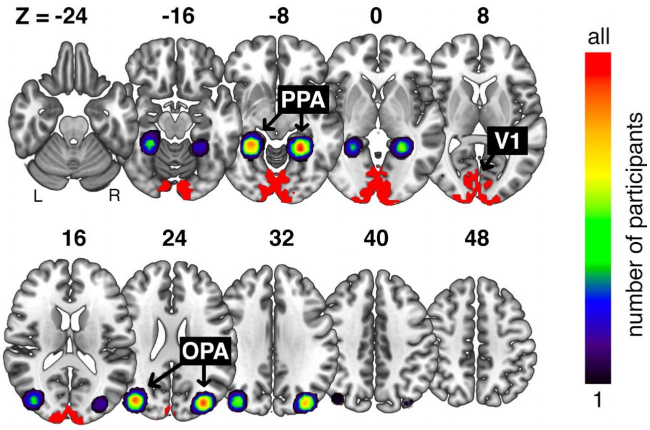

FIGURE 2 Location of the fMRI regions of interest (ROIs). fMRI data analysis was restricted to three ROIs: primary visual cortex (V1), the occipital place area (OPA) and the parahippocampal place area (PPA). The V1 ROI was based on a functional atlas (Wang, Mruczek, Arcaro, & Kastner, 2015), and identical for all participants. The scenes-selective regions were defined as spheres around each participant's peak activation in a separate scene-localizer run, constrained by functional group masks (Julian, Fedorenko, Webster, & Kanwisher, 2012). The colormap represents the consistency of ROI locations across participants (i.e., how many participants' ROIs covered the respective voxels) [Color figure can be viewed at wileyonlinelibrary.com]

constrained by group-level activation masks for OPA and PPA (Julian et al., 2012). Within these masks, we first identified the voxel exhibiting the greatest *t*-value in a scene>object contrast, separately for each hemisphere, and then defined the ROI as a 125-voxel sphere around this voxel (similar results were obtained for different ROI sizes, see Appendix S1). Left- and right-hemispheric ROIs were concatenated for further analysis.[5]

2.8 | fMRI decoding

fMRI response patterns for each ROI were extracted directly from the volumes recorded during each block. After shifting the activation time course by three TRs (i.e., 6 s) to account for the hemodynamic delay, we extracted voxel-wise activation values from the 12 TRs corresponding to each block of 24 s. Activation values for these 12 TRs were then averaged, yielding a single response pattern across voxels for each block. To account for activation differences between runs, the mean activation across all blocks was subtracted from each voxel's values, separately for each run. Decoding analyses were performed using CoSMoMVPA (Oosterhof, Connolly, & Haxby, 2016), and were carried out separately for each ROI and participant. We used data from four runs to train linear discriminant analysis (LDA) classifiers to discriminate multi-voxel response patterns (i.e., patterns of voxel activations across all voxels of an ROI) for two conditions (e.g., spatially intact versus spatially jumbled scenes). Classifiers were tested using response patterns for the same two conditions from the left out, fifth run. This classification routine was done repeatedly until every run was left out once and decoding accuracy was averaged across these repetitions.

2.9 | fMRI univariate analysis

To establish univariate activation differences, we modeled the fMRI data in a GLM analysis. For this analysis, all functional volumes were smoothed using a 6 mm full-width-half-maximum Gaussian kernel. For each run, we constructed a GLM with 10 predictors (four regressors reflecting the four scene conditions and six movement regressors). For each of the four scene conditions, this analysis yielded five beta maps (one for each run) for the upright scenes (from Session 1), and five beta maps (one for each run) for the inverted scenes (from Session 2). We first averaged beta weights for every condition across runs. These beta weights were then averaged across all voxels of each ROI, yielding one activation value for each condition, ROI, and participant. For each ROI (V1, OPA, PPA), and separately for the two stimulus orientations (upright, inverted), we computed three effects: (a) The main effect of spatial structure, reflecting the difference between the two spatially intact and the two spatially jumbled scenes, (b) the main effect of categorical structure, reflecting the difference between the two categorically intact and the two categorically jumbled scenes, and (c) the interaction effect of spatial and categorical structure. Subsequently, to uncover inversion effects, we compared these effects across the upright scenes and inverted scenes.

2.10 | EEG decoding

EEG decoding was performed separately for each time point (i.e., every 5 ms) from −200 ms to 800 ms relative to stimulus onset, using CoSMoMVPA (Oosterhof et al., 2016). We used data from all-but-one trials for two conditions to train LDA classifiers to discriminate topographical response patterns (i.e., patterns across all electrodes) for two conditions (e.g., spatially intact versus spatially jumbled scenes). Classifiers were tested using response patterns for the same two conditions from the left-out trials. This classification routine was done repeatedly until each trial was left out once and decoding accuracy was averaged across these repetitions. Classification time series for individual participants were smoothed using a running average of five time points (i.e., 25 ms).

2.11 | EEG univariate analysis

To establish univariate EEG response differences (i.e., ERP effects) between conditions, we averaged evoked responses for all trials of each condition. Based on a previous study on scene-selective ERPs (Harel et al., 2016), we then averaged these responses across six posterior-lateral EEG electrodes (P4, P8, O2, P7, P3, O1), yielding one ERP response for each condition and participant. For these ERPs, we computed the same effects as outlined above for the fMRI data: a main effect of spatial structure, a main effect of categorical structure, and interactions with scene inversion.[6]

2.12 | Statistical testing

For the fMRI data, we used *t*-tests to compare decoding against chance and between conditions. For the univariate data, we used ANOVAs to tests for differences in activations. To Bonferroni-correct for comparisons across ROIs, all *p*-values were multiplied by 3. For the EEG data, given the larger number of comparisons, we used a threshold-free cluster enhancement procedure (Smith & Nichols, 2009) and multiple-comparison correction based on a sign-permutation test (with null distributions created from 10,000 bootstrapping iterations), as implemented in CoSMoMVPA (Oosterhof et al., 2016). The resulting statistical maps were thresholded at $z > 1.96$ (i.e., $p_{corr} < .05$).

2.13 | Data availability

Data are publicly available on OSF (doi.org/10.17605/OSF.IO/W9874). Materials and code are available from the corresponding author upon request.

3 | RESULTS

For both the fMRI and EEG data, we performed two complimentary decoding analyses. In the first analysis, we tested sensitivity for spatial structure by decoding spatially intact from spatially jumbled scenes (Figure 3a). In the second analysis, we tested sensitivity for categorical structure by decoding categorically intact from categorically jumbled scenes (Figure 3d). To investigate whether successful decoding indeed reflected sensitivity to scene structure, we performed both analyses separately for the upright and inverted scenes. Critically, inversion effects (i.e., better decoding in the upright than in the inverted condition) indicate genuine sensitivity to natural scene structure that goes beyond purely visual differences.

3.1 | Sensitivity to spatial scene structure

First, to uncover where and when cortical processing is sensitive to spatial structure, we decoded between scenes whose spatial structure was intact or jumbled (Figure 3a).

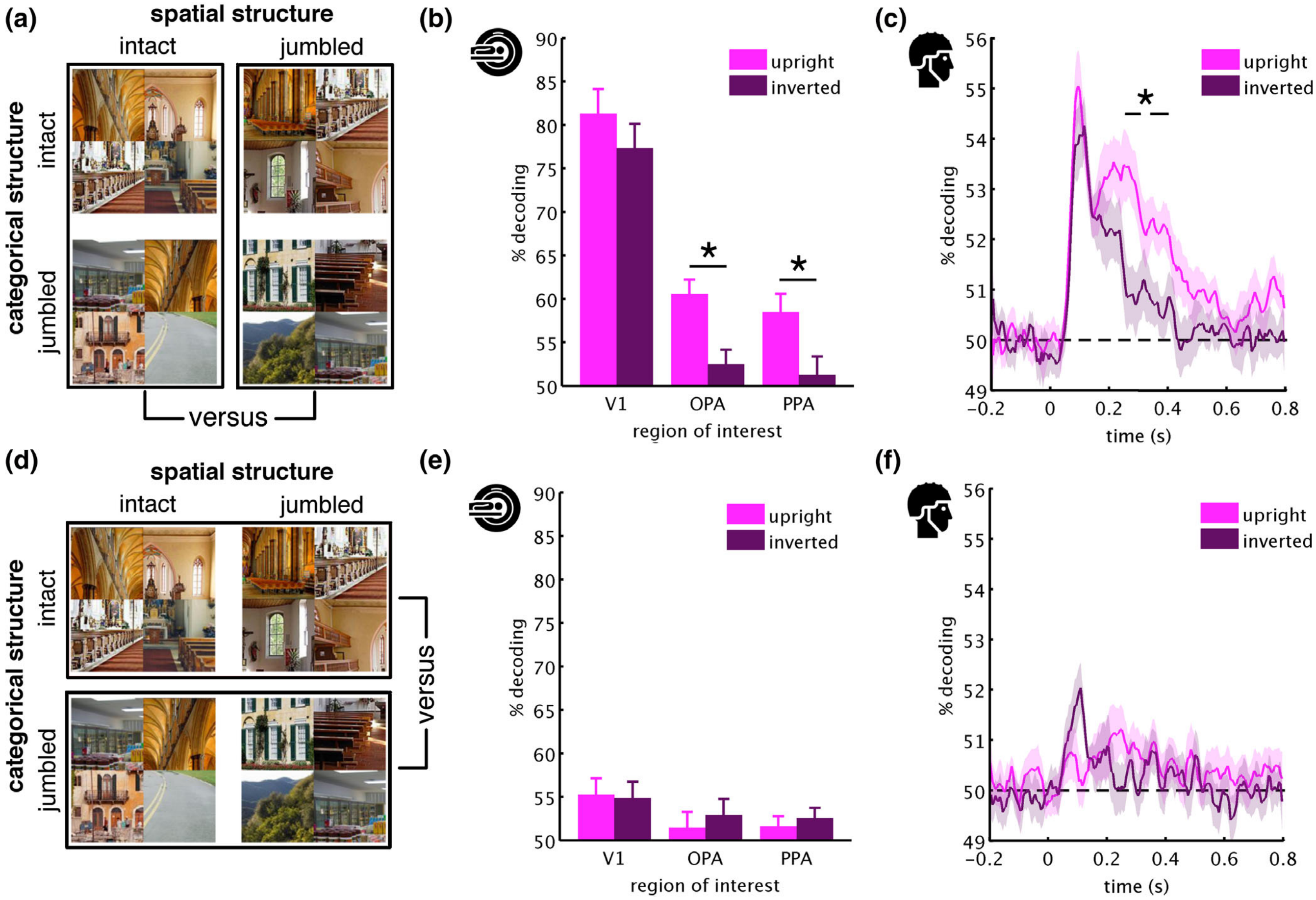

FIGURE 3 MVPA results. To reveal sensitivity to spatial scene structure, we decoded between scenes with spatially intact and spatially jumbled parts (a). Already during early processing (in V1 and before 200 ms) spatially intact and jumbled scenes could be discriminated well, both for the upright and inverted conditions. Critically, during later processing (in OPA/PPA and from 255 ms) inversion effects (i.e., better decoding for upright than inverted scenes) revealed genuine sensitivity to spatial scene structure (b, c). To reveal sensitivity to categorical scene structure, we decoded between scenes with categorically intact and categorically jumbled parts (d). In this analysis, no pronounced decoding and no inversion effects were found, neither across space (e) nor time (f). Error margins reflect standard errors of the difference. Significance markers denote inversion effects ($p_{corr} < .05$) [Color figure can be viewed at wileyonlinelibrary.com]

For the fMRI data (Figure 3b), we found highly significant decoding between spatially intact and spatially jumbled scenes. For upright scenes, significant decoding emerged in V1, $t(19) = 13.03$, $p_{corr} < .001$, OPA, $t(19) = 7.61$, $p_{corr} < .001$, and PPA, $t(19) = 5.92$, $p_{corr} = .002$, and for inverted scenes in V1, $t(19) = 9.92$, $p_{corr} < .001$, but not in OPA, $t(19) = 2.08$, $p_{corr} = .16$, and PPA, $t(19) = 0.85$, $p_{corr} > 1$. Critically, we observed inversion effects (i.e., better decoding for the upright scenes) in the OPA, $t(16) = 4.41$, $p_{corr} = .001$,[7] and PPA, $t(16) = 3.67$, $p_{corr} = .006$, but not in V1, $t(16) = 1.32$, $p_{corr} = .62$. Therefore, decoding in V1 solely reflects visual differences, whereas OPA and PPA exhibit genuine sensitivity to the spatial scene structure. This result was confirmed by further ROI analyses and a spatially unconstrained searchlight analysis (see Appendix S1).

For the EEG data (Figure 3c), we also found strong decoding between spatially intact and jumbled scenes. For upright scenes, this decoding emerged between 55 ms and 465 ms, between 505 ms and 565 ms, and between 740 ms and 785 ms, peak $z > 3.29$, $p_{corr} < .001$, and for inverted scenes between 65 ms and 245 ms, peak $z > 3.29$, $p_{corr} < .001$. As in scene-selective cortex, we observed inversion effects, indexing stronger sensitivity to spatial structure in upright scenes, between 255 ms and 300 ms and between 340 ms and 395 ms, peak $z = 2.78$, $p_{corr} = .005$.

Together, these results show that in scene-selective OPA and PPA, and after 255 ms, cortical activations are sensitive to the spatial structure of natural scenes. Critically, this sensitivity becomes apparent in inversion effects, and thus cannot be attributed to image-specific differences between intact and jumbled scenes, as these are identical for the upright and inverted scenes. Our findings rather indicate a genuine sensitivity to spatial structure consistent with real-world experience.

3.2 | Sensitivity to categorical scene structure

Second, to uncover where and when cortical processing is sensitive to categorical structure, we decoded between scenes whose categorical structure was intact or jumbled (Figure 3a).

For the fMRI (Figure 3e), the upright scenes' categorical structure could be decoded only from V1, $t(19) = 3.11$, $p_{corr} = .017$, but not the scene-selective ROIs, both $t(19) < 2.15$, $p_{corr} > .13$. Similarly, for the inverted scenes, significant decoding was only observed in V1, $t(19) = 4.58$, $p_{corr} < 0.001$, but not in the scene-selective ROIs, both $t(19) < 2.29$, $p_{corr} > .10$. No inversion effects were observed, all $t(16) < 0.60$, $p_{corr} > 1$.

For the EEG (Figure 3f), we found only weak decoding between the categorically intact and jumbled scenes. In the upright condition, decoding was significant between 165 ms and 175 ms and between 215 ms and 265 ms, peak $z = 2.32$, $p_{corr} = .02$, and in the inverted condition at 120 ms, peak $z = 1.97$, $p_{corr} = .049$. No significant inversion effects were observed, peak $z = 1.64$, $p_{corr} = .10$.[8]

Together, these results reveal no substantial sensitivity to the categorical structure of a scene, at least when none of the scenes are fully coherent and when they are not relevant for behavior. Please note that this absence of an effect does not in no way entail that there is no representation of category during scene analysis. In our analysis, we did not decode between different scene categories, but between scenes whose categories were intact or shuffled (collapsed across their categorical content); as a consequence, our analysis only reveals an absence of sensitivity for categorical structure, but not an absence of sensitivity for category per se.

This absence of sensitivity for categorical scene structure is in marked contrast with sensitivity for spatial scene structure, which is observed in the absence of behavioral relevance and is disrupted by stimulus inversion.

3.3 | Enhanced responses to spatially structured scenes

Our decoding analyses show that scene-selective cortex exhibits a profound sensitivity to spatial scene structure. To further understand this sensitivity, we conducted a univariate analysis in which we compared the magnitude of responses evoked by intact and jumbled scenes (Figure 4a,c). Critically, this analysis allowed us to disentangle two opposing interpretations: On one side, sensitivity to scene structure could indeed reflect a visual tuning to real-world properties—in this case, enhanced responses to intact scenes, compared to jumbled scenes, are expected. On the other side, sensitivity to scene structure could mainly reflect the coding of stimuli that are incoherent with real-world experience, reflecting a type of “surprise” response— in this case, enhanced responses to jumbled scenes, compared to intact scenes, are expected. Analyzing response magnitudes across space (fMRI) and time (EEG) allowed us to arbitrate these two interpretations.

In the fMRI, we found significant main effects of spatial structure in the upright condition in OPA, $F(1,19) = 21.00$, $p_{corr} < .001$, and PPA, $F(1,19) = 55.30$, $p_{corr} < .001$, but not in V1, $F(1,19) = 5.11$, $p_{corr} = .11$ (Figure 4b). No main effects of categorical structure, all $F(1,19) < 5.69$, $p_{corr} > .08$, and no interactions between spatial and categorical structure were found, all $F(1,19) < 1.18$, $p_{corr} > .88$. In the inverted condition, we observed no significant effects, all $F(1,19) < 1.12$, $p_{corr} > .92$ (Figure 4e). Critically, we inversion effects revealed greater effects of spatial structure in the upright than in the inverted condition in OPA, $F(1,16) = 17.04$, $p_{corr} = .002$, and PPA, $F(1,16) = 21.82$, $p_{corr} < .001$. In accordance with the MVPA results, this finding indicates genuine sensitivity to spatial scene structure in OPA and PPA. Additionally, the univariate results highlight that scene-selective cortex preferentially responds to the spatially intact scenes, rather than the spatially jumbled scenes.

In the EEG, we only found a significant main effect of spatial structure for the upright scenes (Figure 4c,f), which emerged between 225 ms and 425 ms, peak $z = 3.09$, $p_{corr} = .002$. None of the other main effects or interactions were significant. However, we observed trending inversion effects (at a more liberal threshold of $p_{corr} < .1$), which emerged between 260 ms and 270 ms, and at 305 ms, peak $z = 1.72$, $p_{corr} = .086$. Although not significant, these trending effects qualitatively resemble the findings obtained in the more sensitive

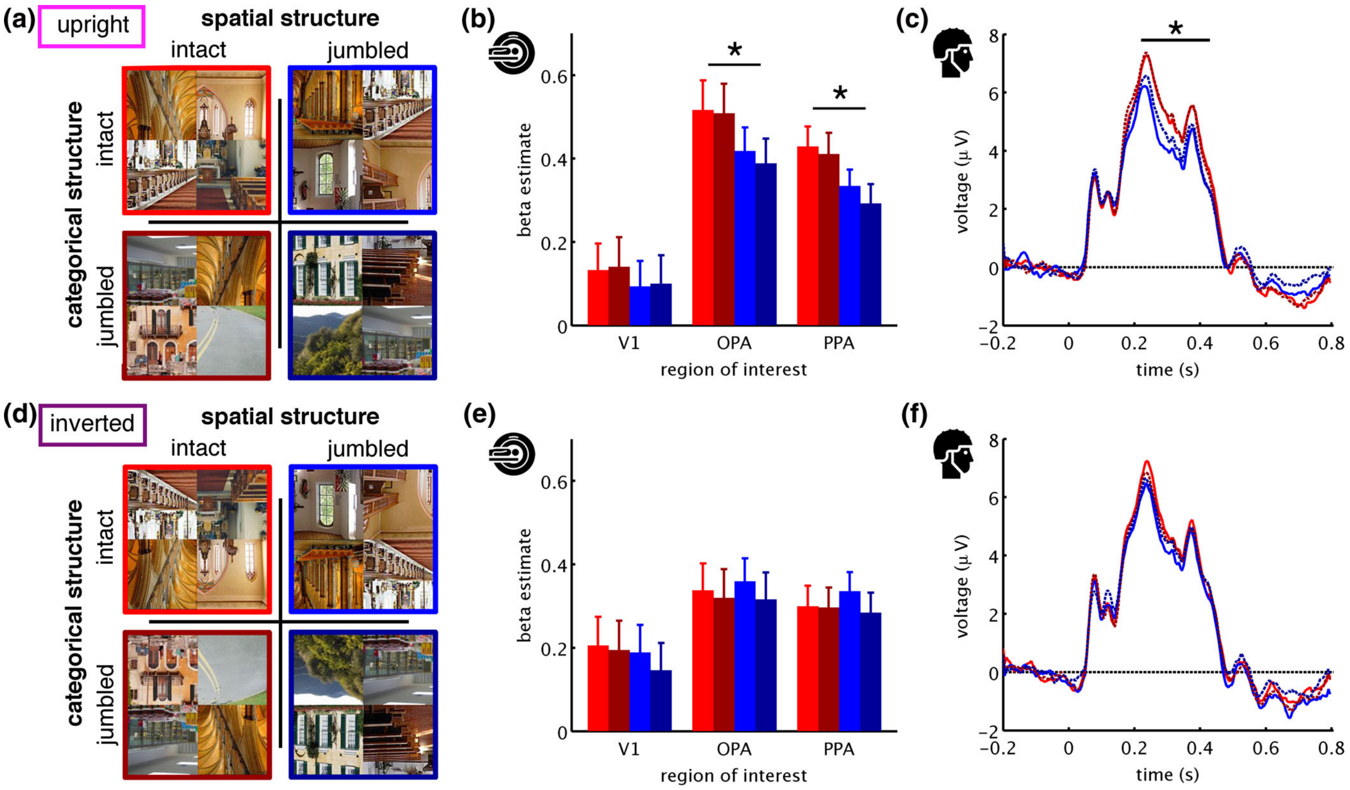

FIGURE 4 Univariate results. To reveal sensitivity to scene structure in univariate response magnitudes, we looked at average responses to each of the four conditions, separately for the upright scenes (a) and the inverted scenes (d). For the upright scenes, we found main effects of spatial structure in OPA and PPA (b) and between 225 ms and 425 ms (c), while no effects of spatial structure were found for the inverted scenes (e, f). Supporting our MVPA results, inversion effects (i.e., greater effects of spatial structure in the upright, compared to the inverted scenes) were found in OPA and PPA (at $p_{corr} < .05$) and from 260 ms (at a more liberal $p_{corr} < .1$), indicating increased responsiveness to spatially structured scenes. No main effects of categorical structure and no interaction effects were found. Error margins reflect standard errors of the mean. Significance markers denote main effects of spatial structure ($p_{corr} < .05$) [Color figure can be viewed at wileyonlinelibrary.com]

MVPA, which showed that from 255 ms responses become sensitive to spatial scene structure.

Together, the univariate results highlight that responses to natural scenes are stronger for scenes that are spatially structured. This suggests a preferential processing of scenes that are composed in accordance with real-world experience—rather than an enhanced response to scenes that do not adhere to this experience.

4 | DISCUSSION

Our findings provide the first spatiotemporal characterization of cortical sensitivity to natural scene structure. As the key result, we observed sensitivity to spatial (but not categorical) scene structure, which emerged in scene-selective cortex and from 255 ms of vision. By showing that this effect is stronger for upright than for inverted scenes, we provide strong evidence for genuine sensitivity to spatial structure, rather than low-level properties.

Sensitivity to spatial structure may index mechanisms enabling efficient scene understanding. Previous work on object processing shows that in order to efficiently parse the many objects contained in natural scenes, the visual system exploits regularities in the environment, such as regularities in individual objects' positions (Kaiser & Cichy, 2018; Kaiser, Moeskops, & Cichy, 2018), relationships between objects (Kaiser & Peelen, 2018; Kaiser, Stein, & Peelen, 2014; Kim & Biederman, 2011; Roberts & Humphreys, 2010), and relationships between objects and scenes (Brandman & Peelen, 2017; Faivre, Dubois, Schwartz, & Mudrik, 2019). Further, a recent fMRI study suggests that low-level representations of small and incomplete scene fragments partly depend on the fragment's typical position within the visual world (Mannion, 2015). Relatedly, we recently showed that in scene-selective occipital cortex and after 200 ms of vision, the representations of such scene fragments are sorted with respect to their typical location in the world (Kaiser, Turini, & Cichy, 2019). Focusing on the interplay of multiple scene elements, the current study shows that on higher levels of the scene processing hierarchy, the visual system uses spatial regularities to concurrently process the multiple elements of complex scenes in an efficient way. This result is in line with the emerging view that real-world structure facilitates processing in the visual system across diverse naturalistic contents (Kaiser, Quek, Cichy, & Peelen, 2019).

What mechanism underlies the preferential processing of spatially structured scenes? As one possibility, a scene's intact spatial structure may trigger integrative processing across the scene, akin to integrative processing of multiple objects that are positioned in accordance with spatial regularities (Baldassano, Beck, & Fei-Fei, 2017; Kaiser & Peelen, 2018). Alternatively, spatially structured scenes may contain typical global properties (Oliva & Torralba, 2006) that are absent in spatially jumbled scenes, and the sensitivity to spatial structure may partly reflect sensitivity to the formation of such global properties. At this point, more studies are needed to understand which types of features drive the sensitivity to spatial structure.

Our results also shine new light on the temporal processing cascade during scene perception. Sensitivity to spatial structure emerged after 255 ms of processing, which is only after scene-selective peaks in ERPs (Harel et al., 2016; Sato et al., 1999)[9] and after basic scene attributes are computed (Cichy, Khosla, Pantazis, & Oliva, 2017). Interestingly, after 250 ms brain responses not only become sensitive to scene structure, but also to object-scene consistencies (Draschkow et al., 2018; Ganis & Kutas, 2003; Mudrik et al., 2010; Võ & Wolfe, 2013). Together, these results suggest a dedicated processing stage for the structural analysis of objects, scenes, and their relationships, which is different from basic perceptual processing. However, whether these different findings indeed reflect a common underlying mechanism requires further investigation. For instance, future investigations need to clarify which of these findings reflect enhanced processing of consistent structure (as our finding does) and which primarily reflect responses to inconsistencies.

Further, our results suggest more pronounced sensitivity to spatial structure than to categorical structure. This is in line with studies showing that scene-selective responses are mainly driven by spatial layout, rather than scene content (Dillon, Persichetti, Spelke, & Dilks, 2018; Harel, Kravitz, & Baker, 2013; Henriksson, Mur, & Kriegeskorte, 2019; Kravitz, Peng, & Baker, 2011). However, our results need not to be taken as evidence that categorical structure is not represented at all during visual analysis.[10] It is conceivable that visual processing is less sensitive to categorical structure when, as in our study, all scenes are jumbled to some extent and not behaviorally relevant.

On the contrary, robust sensitivity to spatial scene structure emerged in the absence of behavioral relevance. This suggests that spatial structure is analyzed automatically during perceptual processing and is not strongly dependent on attentional engagement with the scene. As in real-world situations, we cannot explicitly engage with all aspects of a scene concurrently, this automatic analysis of spatial structure may be crucial for rapid scene understanding.

ACKNOWLEDGMENTS

We thank Sina Schwarze for help in EEG data collection and manuscript preparation. D.K. and R.M.C. are supported by Deutsche Forschungsgemeinschaft (DFG) grants (KA4683/2-1, CI241/1-1, CI241/3-1). R.M.C. is supported by an European Research Council Starting Grant (ERC-2018-StG 803370). G.H. was supported by a PhD fellowship of the Einstein Center for Neurosciences.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

D. K. and R. M. C. designed research, D. K. and G. H. acquired data, D. K. and G. H. analyzed data, D. K., G. H., and R. M. C. interpreted results, D. K. prepared figures, D. K. drafted manuscript, D. K., G. H., and R. M. C. edited and revised manuscript. All authors approved the final version of the manuscript.

DATA AVAILABILITY STATEMENT

Data are publicly available on OSF (doi.org/10.17605/OSF.IO/W9874). Materials and code are available from the corresponding author upon request.

ORCID

Daniel Kaiser https://orcid.org/0000-0002-9007-3160
Radoslaw M. Cichy https://orcid.org/0000-0003-4190-6071

ENDNOTES

[1] Related studies on object-object and object-scene consistencies typically yield large effect sizes which exceed this value, both for fMRI responses, d = 0.72 (Brandman & Peelen, 2017), d = 0.67 (Kaiser & Peelen, 2018), d = 2.14 (Kim & Biederman, 2011), d = 0.94 (Roberts & Humphreys, 2010), and EEG responses, d = 0.71 (Draschkow, Heikel, Võ, Fiebach, & Sassenhagen, 2018), d = 0.88 (Ganis & Kutas, 2003), d = 0.67 (Mudrik, Lamy, & Deouell, 2010), d = 0.69 (Võ & Wolfe, 2013).

[2] Note that all scenes were jumbled to some extent, as also in the categorically intact scenes four different exemplars were intermixed.

[3] For two participants, due to technical problems, no button presses were recorded.

[4] For two participants, due to technical problems, only data from 32 electrodes was recorded.

[5] Analyzing the data from the two hemispheres separately did not yield any significant differences between hemispheres (F < 2.04, p > .17, for all interactions with hemisphere).

[6] For using the same statistical tests as for the decoding results, interactions in the univariate EEG analyses were computed by testing the differences between conditions against each other (e.g., the difference between intact and jumbled scenes in the upright condition versus the difference between intact and jumbled scenes in the inverted conditions).

[7] Statistics for fMRI inversion effects are based on the 17 participants who completed both sessions.

[8] Note that the strongest tendency towards an inversion effect (at 115 ms) was against the predicted direction.

[9] In our study, ERP responses in posterior-lateral electrodes peaked at 235 ms.

[10] In the Appendix S1, we show that the four scene categories can be successfully decoded from the EEG signals.

REFERENCES

Baldassano, C., Beck, D. M., & Fei-Fei, L. (2017). Human-object interactions are more than the sum of their parts. *Cerebral Cortex*, *27*, 2276–2288.

Bernstein, M., Oron, J., Sadeh, B., & Yovel, G. (2014). An integrated face-body representation in the fusiform gyrus but not the lateral occipital cortex. *Journal of Cognitive Neuroscience*, *26*, 2469–2478.

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*, 77–80.

Biederman, I., Glass, A. L., & Stacy, E. W. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, *97*, 22–27.

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

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

Brandman, T., & Peelen, M. V. (2017). Interaction between scene and object processing revealed by human fMRI and MEG decoding. *Journal of Neuroscience*, *37*, 7700–7710.

Brandman, T., & Yovel, G. (2016). Bodies are represented as wholes rather than their sum of parts in the occipital-temporal cortex. *Cerebral Cortex*, *26*, 530–543.

Chan, A. W., Kravitz, D. J., Truong, S., Arizpe, J., & Baker, C. I. (2010). Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature Neuroscience*, *13*, 417–418.

Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, *153*, 346–358.

Dillon, M. R., Persichetti, A. S., Spelke, E. S., & Dilks, D. D. (2018). Places in the brain: Bridging layout and object geometry in scene-selective cortex. *Cerebral Cortex*, *28*, 2365–2374.

Draschkow, D., Heikel, E., Võ, M. L.-H., Fiebach, C. J., & Sassenhagen, J. (2018). No evidence for different processes underlying the N300 and N400 incongruity effects in object-scene processing. *Neuropsychologia*, *120*, 9–17.

Epstein, R. A. (2014). Neural systems for visual scene recognition. In M. Bar & K. Keverага (Eds.), *Scene Vision* (pp. 105–134). Cambridge: MIT Press.

Faivre, N., Dubois, J., Schwartz, N., & Mudrik, L. (2019). Imaging object-scene relations processing in visible and invisible natural scenes. *Scientific Reports*, *9*, 4567.

Ganis, G., & Kutas, M. (2003). An electrophysiological study of scene effects on object identification. *Cognitive Brain Research*, *16*, 123–144.

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *eNeuro*, *3*, ENEURO.0139-16.2016.

Harel, A., Kravitz, D. J., & Baker, C. I. (2013). Deconstructing visual scenes in cortex: Gradients of object and spatial layout information. *Cerebral Cortex*, *23*, 947–957.

Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *Neuron*, *103*, 161–171.e3. https://doi.org/, https://doi.org/10.1016/j.neuron.2019.04.014

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

Kaiser, D., & Cichy, R. M. (2018). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology*, *120*, 848–853.

Kaiser, D., Moeskops, M. M., & Cichy, R. M. (2018). Typical retinotopic locations impact the time course of object coding. *NeuroImage*, *176*, 372–379.

Kaiser, D., & Peelen, M. V. (2018). Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *NeuroImage*, *169*, 334–341.

Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences*, *23*, 672–685.

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

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

Kim, J. G., & Biederman, I. (2011). Where do objects become scenes? *Cerebral Cortex*, *21*, 1738–1746.

Konkle, T., Brady, T. F., Alvarez, G. A., & Oliva, A. (2010). Scene memory is more detailed than you think: The role of categories in visual long-term memory. *Psychological Science*, *21*, 1551–1556.

Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *Journal of Neuroscience*, *31*, 7322–7333.

Mannion, D. J. (2015). Sensitivity to the visual field origin of natural image patches in human low-level visual cortex. *PeerJ*, *3*, e1038.

Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object-scene processing. *Neuropsychologia*, *48*, 507–517.

Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in Brain Research*, *155*, 23–36.

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

Oostenveld, R., Fries, P., Maris, E., & Schoffelen, J. M. (2011). FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Computational Intelligence and Neuroscience*, *2011*, 156869.

Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU octave. *Frontiers in Neuroinformatics*, *10*, 20.

Potter, M. C. (1975). Meaning in visual search. *Science*, *187*, 965–966.

Roberts, K. L., & Humphreys, G. W. (2010). Action relationships concatenate representations of separate objects in the ventral visual cortex. *NeuroImage*, *52*, 1541–1548.

Sato, N., Nakamura, K., Nakamura, A., Sugiura, M., Iko, K., Fukuda, H., & Kawashima, R. (1999). Different time course between scene processing and face processing: A MEG study. *Neuroreport*, *10*, 3633–3637.

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

Thorpe, S., Fize, D., & Marlot, D. (1996). Speed of processing in the human visual system. *Nature*, *381*, 520–522.

Varakin, D. A., & Levin, D. T. (2008). Scene structure enhances change detection. *The Quarterly Journal of Experimental Psychology*, *61*, 543–551.

Võ, M. L.-H., Boettcher, S. E. P., & 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., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science*, *24*, 1816–1823.

Wang, L., Mruczek, R. E., Arcaro, M. J., & Kastner, S. (2015). Probabilistic maps of visual topography in human cortex. *Cerebral Cortex*, *25*, 3911–3931.

Wolfe, J. M., Võ, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*, 77–84.

Zimmermann, E., Schnier, F., & Lappe, M. (2010). The contribution of scene context on change detection performance. *Vision Research*, *50*, 2062–2068.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Kaiser D, Häberle G, Cichy RM. Cortical sensitivity to natural scene structure. *Hum Brain Mapp*. 2019;1–10. https://doi.org/10.1002/hbm.24875

Chapter 3

Project II: Real-world structure facilitates the rapid emergence of scene category information in visual brain signals

The current chapter comprises the research article entitled "Real-world structure facilitates the rapid emergence of scene category information in visual brain signals" that was published in the *Journal of Neurophysiology* in 2020. This second research project demonstrated that the presence of intact scene structure facilitates the analysis of the scenes' categorical content within 200ms of vision.

Authors:

Daniel Kaiser, Greta Häberle, Radoslaw M. Cichy

Contributions:

D.K. and R.M.C. conceived and designed research; G.H. performed experiments; D.K. and G.H. analyzed data; D.K., G.H., and R.M.C. interpreted results of experiments; D.K. prepared figures; D.K. drafted manuscript; D.K., G.H., and R.M.C. edited and revised manuscript.

Contributions to open and reproducible science:

To contribute to open and reproducible science, the paper is published in an open access journal. The original article can be found here: doi: 10.1152/jn.00164.2020. Data are publicly available on OSF: doi: 10.17605/OSF.IO/ W9874.

Copyright note:

Journal of Neurophysiology is an open-access journal. All articles are published under a Creative Commons Attribution 4.0 International License and are free to re-use.

J Neurophysiol 124: 145–151, 2020.
First published June 10, 2020; doi:10.1152/jn.00164.2020.

RAPID REPORT | *Sensory Processing*

Real-world structure facilitates the rapid emergence of scene category information in visual brain signals

Daniel Kaiser,[1] Greta Häberle,[2,3,4] and Radoslaw M. Cichy[2,3,4,5]

[1]*Department of Psychology, University of York, York, United Kingdom;* [2]*Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany;* [3]*Charité — Universitätsmedizin Berlin, Einstein Center for Neurosciences Berlin, Berlin, Germany;* [4]*Berlin School of Mind and Brain, Humboldt-Universität zu Berlin, Berlin, Germany; and* [5]*Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany*

Submitted 26 March 2020; accepted in final form 8 June 2020

Kaiser D, Häberle G, Cichy RM. Real-world structure facilitates the rapid emergence of scene category information in visual brain signals. *J Neurophysiol* 124: 145–151, 2020. First published June 10, 2020; doi:10.1152/jn.00164.2020.—In everyday life, our visual surroundings are not arranged randomly but structured in predictable ways. Although previous studies have shown that the visual system is sensitive to such structural regularities, it remains unclear whether the presence of an intact structure in a scene also facilitates the cortical analysis of the scene's categorical content. To address this question, we conducted an EEG experiment during which participants viewed natural scene images that were either "intact" (with their quadrants arranged in typical positions) or "jumbled" (with their quadrants arranged into atypical positions). We then used multivariate pattern analysis to decode the scenes' category from the EEG signals (e.g., whether the participant had seen a church or a supermarket). The category of intact scenes could be decoded rapidly within the first 100 ms of visual processing. Critically, within 200 ms of processing, category decoding was more pronounced for the intact scenes compared with the jumbled scenes, suggesting that the presence of real-world structure facilitates the extraction of scene category information. No such effect was found when the scenes were presented upside down, indicating that the facilitation of neural category information is indeed linked to a scene's adherence to typical real-world structure rather than to differences in visual features between intact and jumbled scenes. Our results demonstrate that early stages of categorical analysis in the visual system exhibit tuning to the structure of the world that may facilitate the rapid extraction of behaviorally relevant information from rich natural environments.

NEW & NOTEWORTHY Natural scenes are structured, with different types of information appearing in predictable locations. Here, we use EEG decoding to show that the visual brain uses this structure to efficiently analyze scene content. During early visual processing, the category of a scene (e.g., a church vs. a supermarket) could be more accurately decoded from EEG signals when the scene adhered to its typical spatial structure compared with when it did not.

EEG; multivariate pattern analysis; real-world structure; scene representation; visual processing

INTRODUCTION

IN EVERYDAY SITUATIONS, the input to our visual system is not random; rather, it rather arises from highly organized scenes, which follow a predictable structure. In practically every real-word scene, visual information (such as the scene's layout properties or the objects contained in a scene) is distributed in meaningful ways across space (Bar 2004; Kaiser et al. 2019a; Oliva and Torralba 2007; Võ et al. 2019; Wolfe et al. 2011). Neuroimaging studies have shown that the visual system is sensitive to this structure, with cortical responses differing when scene elements do or do not adhere to typical real-world structure (Abassi and Papeo 2020; Baldassano et al. 2017; Bilalić et al. 2019; Kaiser et al. 2014; Kaiser and Peelen 2018; Kim and Biederman 2011; Roberts and Humphreys 2010). Although such studies suggest that the presence of real-world structure aids efficient scene representation, it is unclear how real-world structure impacts the representation of scene content. Specifically, does the presence of real-world structure facilitate the extraction of categorical information from a scene?

Evidence for an increase of visual category information in the presence of real-world regularities has already been reported for individual object processing. Several studies showed that typical real-world positioning enhances the neural representation of object category (Chan et al. 2010; de Haas et al. 2016; Kaiser and Cichy 2018; Kaiser et al. 2018); for example, neural responses to an airplane are better discriminable from responses to other objects when the airplane is shown in the upper visual field, where it is typically encountered in the real world. Does the presence of real-world structure similarly facilitate the representation of categorical scene content in scenes?

To address this question, we used a jumbling paradigm (Biederman 1972; Biederman et al. 1974) that manipulates natural scenes' spatial structure. Individual parts of the scene could either appear in their typical, "intact" positions or in atypical, "jumbled" positions (Fig. 1). In a recent neuroimaging study (Kaiser et al. 2020a), we employed this paradigm to show that in scene-selective visual cortex (fMRI) and after 250 ms of vision (EEG), spatially intact scenes were represented differently from jumbled scenes. Here, we analyzed the EEG data from this jumbling paradigm to investigate whether the typical real-world structure, in contrast to an atypical structure, facilitates the visual representation of scene category.

Correspondence: D. Kaiser (danielkaiser.net@gmail.com).

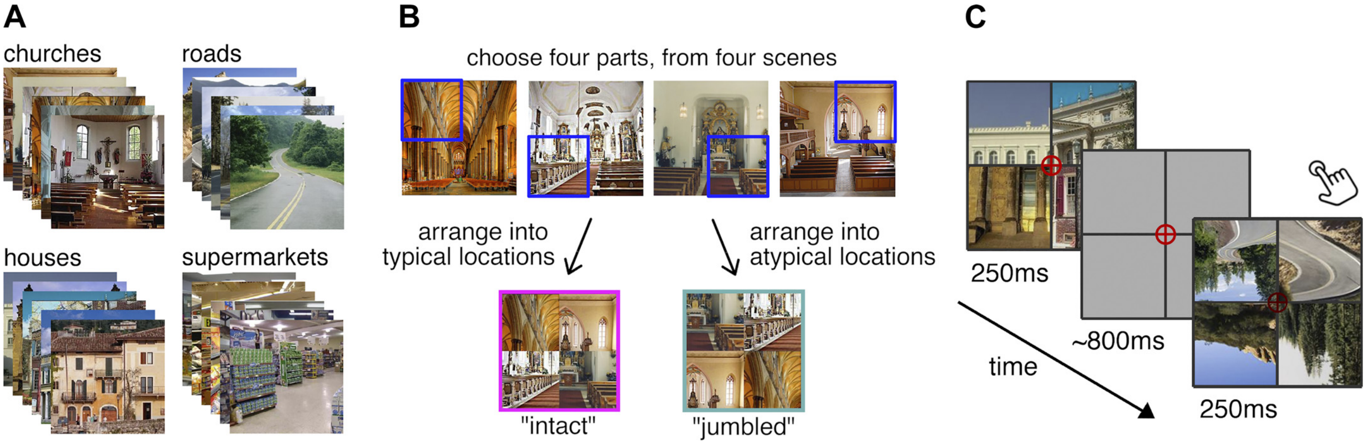
A
churches
roads
houses
supermarkets
B
choose four parts, from four scenes
arrange into typical locations
arrange into atypical locations
"intact"
"jumbled"
C
250ms
~800ms
time
250ms

Fig. 1. Experimental design. *A*: stimulus set was constructed from natural scene photographs of 4 categories. *B*: intact and jumbled scenes were created by combining parts of 4 different scenes of the same category in either typical locations or in atypical locations (with positions swapped in a crisscrossed way). *C*: during the EEG experiment, participants viewed the scenes in upright and inverted orientation for 250 ms each, in random order. Participants performed an orthogonal task, where they responded whenever the fixation cross darkened.

To extract differences in category information between intact and jumbled scenes with high sensitivity, we used a cumulative multivariate decoding approach (Ramkumar et al. 2013), which maximizes the amount of data available at every time point along the processing cascade. In line with previous reports (Dima et al. 2018; Kaiser et al. 2019b, 2020b; Lowe et al. 2018), this analysis showed that scene category information emerges rapidly (within the first 100 ms of vision). Critically, the early emergence of scene category information was facilitated for intact compared with jumbled scenes. This benefit was only present for upright but not inverted scenes, indicating that the early facilitation of scene analysis is related to the presence of real-world structure rather than differences in basic visual features.

MATERIALS AND METHODS

Participants. Twenty healthy adults (mean age 26.6 yr, SD = 5.8; 9 female) participated. All participants had normal or corrected-to-normal vision. Participants provided written informed consent and received either monetary reimbursement or course credits. All procedures were approved by the ethical committee of the Department of Psychology at Freie Universität Berlin and were in accordance with the Declaration of Helsinki.

Stimuli. Stimuli were scenes from four different categories: churches, houses, roads, and supermarkets (Fig. 1*A*). The stimuli were taken from an online resource (Konkle et al. 2010). For each category, six different exemplars were used. To manipulate scenes' adherence to real-world structure, we first split each original image into quadrants. We then systematically recombined parts (quadrants) from different scenes such that the scenes' spatial structure was either intact or jumbled (Fig. 1*B*). For the intact scenes, four parts from four different scenes of the same scene category were combined in their correct spatial locations. For the jumbled scenes, four parts from four different scenes of the same scene category were combined, but their spatial locations were arranged in a crisscrossed way. This jumbling manipulation simultaneously disrupted multiple structural regularities in the scene, such as visual feature distributions, scene geometry, absolute and relative object positions, and cues to three-dimensional structure. Additionally, the stimulus set entailed scenes that were jumbled in their categorical content (with the individual scene parts stemming from different categories); these scenes were created to answer a different research question (see Kaiser et al. 2020a) and not used in the analyses reported in this paper. In both conditions relevant for this paper, we used parts from four different scenes to equate the presence of visual discontinuities between fragments. Separately for each participant, 24 unique intact and 24 unique jumbled stimuli were generated by randomly drawing suitable fragments from different scenes. Each scene was presented upright and upside down. Although the key manipulation was the positioning of the individual scene parts relative to each other, it is worth noting that stimuli from the four resulting conditions adhered to, or violated, real-world structure on different levels: *1*) upright intact scenes featured typical orientation of the individual parts, typical absolute locations of the parts, and typical relative positions of the parts; *2*) upright jumbled scenes featured typical orientation of the individual parts, atypical absolute locations of the parts, and atypical relative positions of the parts; *3*) inverted intact scenes featured atypical orientation of the individual parts, atypical absolute locations of their individual parts, and typical relative positions of the parts; and *4*) inverted jumbled scenes featured atypical orientation of the individual parts, typical absolute locations of the parts, and atypical relative positions of the parts.

Paradigm. During the EEG experiment, the different stimuli were randomly intermixed within a single session. Within each trial, a scene appeared for 250 ms. Stimuli appeared in a black grid (4.5° visual angle), which served to mask visual discontinuities between quadrants (Fig. 1*C*). Each trial was followed by an intertrial interval that varied randomly between 700 ms and 900 ms. For this paper, only parts of the collected data (spatially intact and spatially jumbled scenes in upright and upside-down orientation) were analyzed. Each of these four conditions covered 384 trials (96 trials per scene category). Additionally, 1,152 target trials were measured. During the target trials, the crosshair changed into a slightly darker red at the same time the scene was presented. When detecting a target, participants had to press a button; additionally, they were asked to blink during the target trials, making it easier for them to refrain from blinking during nontarget trials. Target detection was purposefully made challenging to ensure sufficient attentional engagement (mean accuracy 78.1%, SE = 3.6%). Target trials were not included in subsequent analyses. Furthermore, 1,536 trials where the scenes' categorical structure was altered were measured. This data has been analyzed elsewhere (see Kaiser et al. 2020a). Furthermore, participants were instructed to maintain central fixation throughout the experiment. Stimulus presentation was controlled using the Psychtoolbox (Brainard 1997).

EEG recording and preprocessing. The EEG data were the same as in Kaiser et al. (2020a). EEG signals were recorded using an EASYCAP 64-electrode system and a Brainvision actiCHamp amplifier. For two participants, only data from 32 electrodes were recorded because of technical problems. Electrodes were arranged in accordance with the 10–10 system. EEG data was recorded at 1,000 Hz sampling rate and filtered online between 0.03 Hz and 100 Hz. All electrodes were referenced online to the Fz electrode. Offline preprocessing was performed using FieldTrip (Oostenveld et al. 2011). EEG data were epoched from −200 ms to 800 ms relative to stimulus onset and were baseline corrected by subtracting the mean prestimulus signal. Channels and trials containing excessive noise were removed based on visual inspection. Blink and eye movement artifacts were removed using independent components analysis and visual inspection of the resulting components (Jung et al. 2000). The epoched data were downsampled to 200 Hz.

EEG decoding. Decoding analyses were performed using CoSMoMVPA (Oosterhof et al. 2016). To track cortical representations across time, we used a cumulative classification approach that takes into account all time points before the current time point for each time point across the epoch (Ramkumar et al. 2013). This classification technique uses larger amounts of data at each subsequent time point while maintaining temporal precision in the forward direction (i.e., it only collapses across information backward in time but not forward). Cumulative decoding may thus provide increased sensitivity for detecting decoding onsets compared with standard timeseries decoding (Grootswagers et al. 2017).

We used such cumulative classifiers to discriminate between the four scene categories. This analysis was done separately for the intact and jumbled scenes. Classification analyses were performed repeatedly, with the amount of information available to the classifier accumulating across time (Fig. 2); that is, for the first time point in the epoch, the classifier was trained and tested on response patterns across the electrodes at this time point. At the second time point in the epoch, the classifier was trained and tested on response patterns across the electrodes at the first and second time point in this epoch. Finally, at the last time point in the epoch, the classifier was trained on response patterns across all electrodes and at all time points in this epoch.

The richer information contained in these cumulative response patterns comes at the expense of a higher dimensionality of the data, which potentially harms classification. To reduce the dimensionality of the data at each time point, we performed principal component analyses (PCAs). These PCAs were always done on the classifier training set, and the PCA solution was projected onto the testing set (Grootswagers et al. 2017). For each PCA, we retained as many components as needed to explain 99% of the variance in the training set data (average number of components retained at example time points; at 0 ms: 225, SE = 11; at 200 ms: 250, SE = 10; at 800 ms: 269, SE = 10).

For classification, we used linear discriminant analysis classifiers. For each classifier, the covariance matrix was regularized by adding the identity matrix scaled by 1% of the mean of the diagonal elements (as implemented in the *cosmo_classify_lda* function in CoSMoMVPA; Oosterhof et al. 2016). Classification was performed in a cross-validation scheme with 12 distinct folds. Classifiers were trained on data from 11 of these folds and tested on data from the left-out fold. The amount of data in the training set was always balanced across the four categories. Classification was done repeatedly until every fold was left out once. Classification accuracies were averaged across these repetitions. These analyses resulted in separate decoding timeseries for intact and jumbled scenes, which reflect the temporal accrual of category information (i.e., how well the four categories are discriminable from the neural data).

Statistical testing. To compare decoding timeseries against chance level and the different conditions' decoding timeseries against each other, we used a threshold-free cluster enhancement (TFCE) proce-

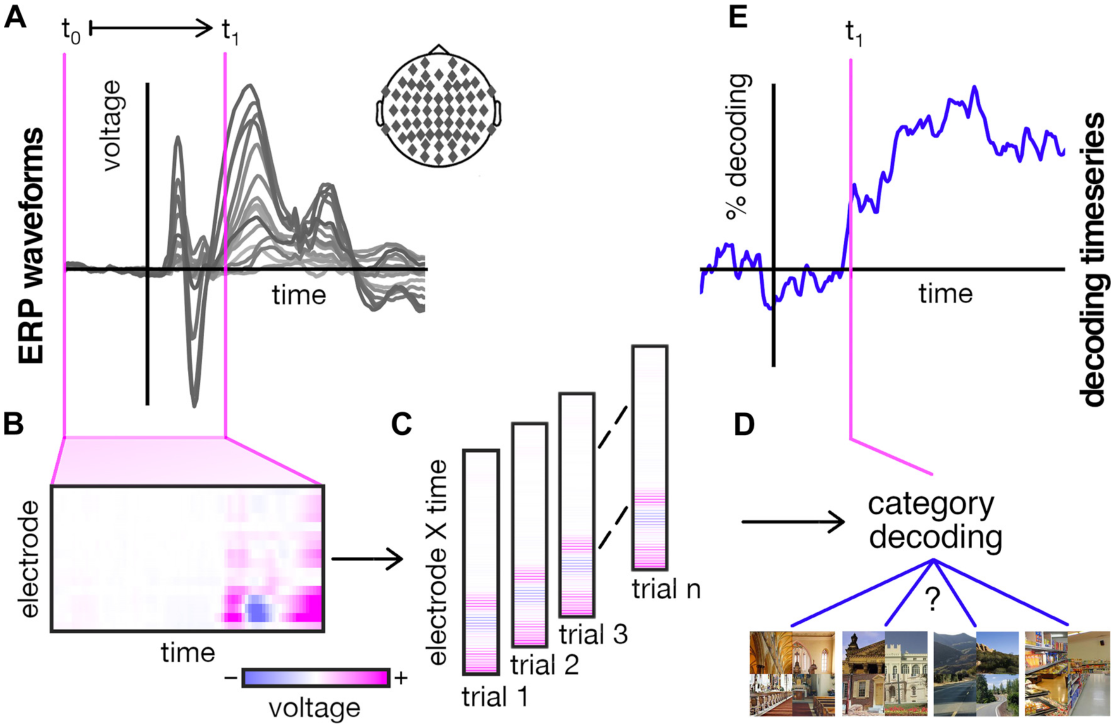

Fig. 2. Schematic depiction of the cumulative decoding approach. *A*: for each time point t_1 across the epoch, a separate decoding analysis was performed. *B*: for each of these analyses, we aggregated event-related potential waveforms across all EEG electrodes and all time points between t_1 and the beginning of the epoch (t_0). *C*: for each trial, we then unfolded these two-dimensional response patterns across electrodes and time into a one-dimensional response pattern. *D*: these one-dimensional response patterns were first subjected to principal component analysis to reduce dimensionality (see MATERIALS AND METHODS) and then fed to linear discriminant analysis classifiers, which were trained to discriminate the 4 scene categories. Decoding accuracy was computed by repeatedly assessing classifier performance on single trials left out during classifier training. *E*: repeating this analysis across time yielded a decoding timeseries with 200 Hz resolution. Importantly, the cumulative nature of this analysis allowed us to increase power by increasing the amount of data available to the classifier without losing temporal precision regarding the onset of category information.

dure (Smith and Nichols 2009). Multiple-comparison correction was based on a sign-permutation test (with null distributions created from 10,000 bootstrapping iterations) as implemented in CoSMoMVPA (Oosterhof et al. 2016). The resulting statistical maps were thresholded at $z > 1.96$ (i.e., $P_{corr} < 0.05$). However, the onset of statistical significance for TFCE methods may be biased by the presence of strong clusters following the onset (as expected from the cumulative decoding performed here) and can therefore not be directly interpreted (Sassenhagen and Draschkow 2019). We thus additionally provide statistics for conventional one-sample t tests, which we corrected for multiple comparisons using false discovery rate (FDR) corrections. For all tests, only clusters of at least 4 consecutive significant time points (i.e., more than 20 ms) were considered.

Data availability. Data are publicly available on OSF (https://doi.org/10.17605/OSF.IO/ECMA4).

RESULTS

We first analyzed data from the upright scenes, where we expected a facilitation of category information for spatially intact, compared with jumbled, scenes. We found that EEG signals conveyed robust scene category information. Categories were discriminable for both intact scenes (significant decoding obtained from TFCE statistics: between 75 ms and 800 ms; significant decoding obtained from FDR-corrected statistics: between 75 ms and 800 ms) and jumbled scenes (TFCE: between 120 ms and 800 ms; FDR: between 135 ms and 800 ms) (Fig. 3*A*). Crucially, we found significantly enhanced decoding for the spatially intact scenes compared with the jumbled scenes (TFCE: between 105 ms and 800 ms; FDR: between 105 ms and 800 ms) (Fig. 3*C*).

The inclusion of inverted scenes allowed us to investigate whether the effects of scene structure were genuinely related to the scenes adhering to real-world structure rather than differences in their low-level visual attributes. If the enhanced category information for spatially intact scenes is indeed related to their adherence with real-world structure, then no effects should be seen when the same scenes are viewed upside

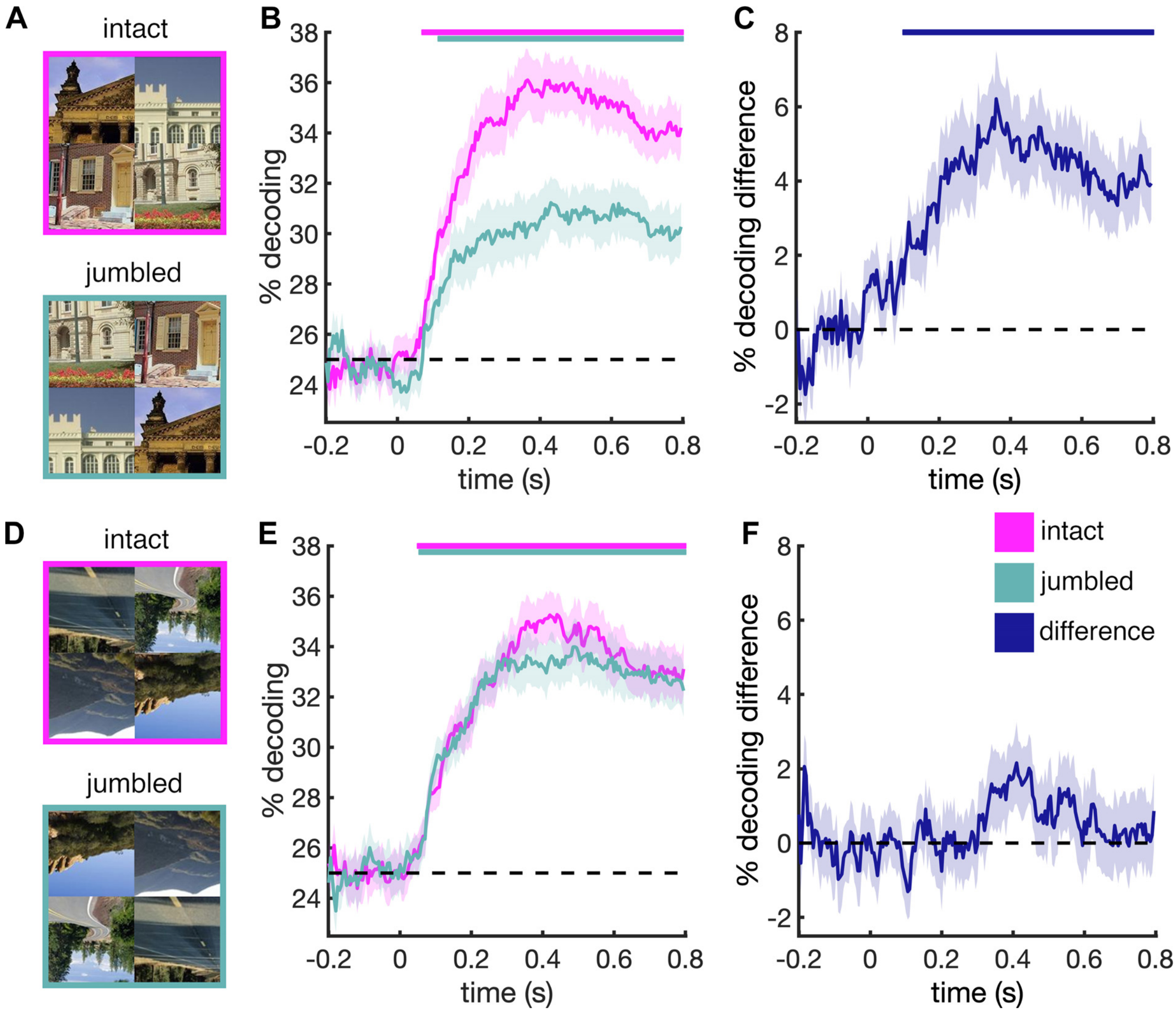

Fig. 3. Decoding of scene category for intact and jumbled scenes. *A*: first, we decoded the category of intact and jumbled scenes when they were presented upright. *B*: this analysis revealed widespread clusters of category decoding for both intact and jumbled scenes. *C*: critically, we found more accurate decoding of scene category when the scene was intact, suggesting that adherence to real-world structure boosts early visual category information. *D*: second, we decoded the category of upside-down scenes. *E*: for upside-down scenes, category could be similarly decoded from the EEG signals. *F*: however, there was no benefit of intact scene structure when the scenes were inverted, suggesting that adherence to real-world structure, rather than low-level differences, explains the enhanced category decoding for structured scenes when they are upright. Error margins indicate standard errors of the difference. Significance markers (colored horizontal lines) indicate $P < 0.05$, corrected for multiple comparisons using threshold-free cluster enhancement.

down, as both types of inverted scenes do not adhere to real-world structure in the same way as upright scenes: (*1*) although their individual parts appear in typical relative positions, the inverted intact scenes have parts that are themselves inverted and each appear in atypical absolute locations, and *2*) although their individual parts appear in typical absolute positions, the inverted jumbled scenes have parts that are themselves inverted and each appear in atypical relative positions.

Performing the category decoding analysis on the inverted scenes (Fig. 3*D*) revealed a qualitative difference to the upright scenes. The effect of scene structure was significantly stronger for the upright scenes (TFCE: between 170 ms and 800 ms; FDR: between 95 ms and 115 ms and between 185 ms and 800 ms). Indeed, no significant differences between intact and jumbled scenes were observed for the inverted scenes, although the category of both intact scenes (TFCE: between 55 ms and 800 ms; FDR: between 60 ms and 800 ms) and jumbled scenes (TFCE: between 60 ms and 800 ms; FDR: between 75 ms and 800 ms) could be decoded from the EEG signals (Fig. 3, *E* and *F*). This indicates that the early facilitation of scene category information for spatially structured scenes can be attributed to the scenes adhering to typical real-world structure, rather than to low-level features differing between the intact and jumbled scenes.

Our results establish that for processing of upright scenes, scene structure matters more than for processing inverted scenes. Additionally, one can also ask how robustly category information emerges as a function of whether the scene is presented upright or upside down. To answer this question, we directly compared category information for the intact upright scenes, the jumbled upright scenes, and the inverted scenes (Fig. 4*A*). For the inverted scenes we averaged across the intact and jumbled conditions, because there were no statistical differences between them. We found that category decoding accuracy for the inverted scenes was numerically in between the intact and jumbled upright scenes (Fig. 4*B*). When directly comparing the decoding time courses (Fig. 4*C*), we found that category decoding was not significantly stronger in the intact upright scenes compared with the inverted scenes. By contrast, category decoding for the upright jumbled scenes was significantly weaker than for the inverted scenes (TFCE: between 170 ms and 800 ms; FDR: between 200 ms and 800 ms). This result suggests that for the inverted scenes, category can be decoded similarly as for the intact upright scenes. However, once the structure of an upright scene is destroyed, only weaker categorical representations emerge in the visual system.

DISCUSSION

Our results provide evidence that real-world regularities facilitate the extraction of scene category information during visual analysis. We show that this facilitation of category information emerges within the first 200 ms of vision. Our findings highlight the pervasive role of real-world structure in perceptual processing, suggesting that already at relatively early processing stages cortical scene representations are tightly linked to the typical composition of our daily surroundings.

Here, we used a cumulative decoding technique to establish differences in the initial emergence of information in EEG signals. This technique uses all the available historical data (i.e., data before the current time point) for classification. Together with using PCA for dimensionality reduction, the availability of this larger amount of data promises high detection sensitivity. The availability of historical data at later time points may also hold true for the brain, where downstream regions have access to information coded earlier in upstream regions. However, as a note of caution, classifiers may also use temporally distinct information that is not necessarily available in the same way in the brain, particularly when looking at late processing stages. Cumulative decoding nonetheless provides a useful approach to reveal early differences in cortical information processing.

The early facilitation of category information is consistent with results from single-object processing, where representa-

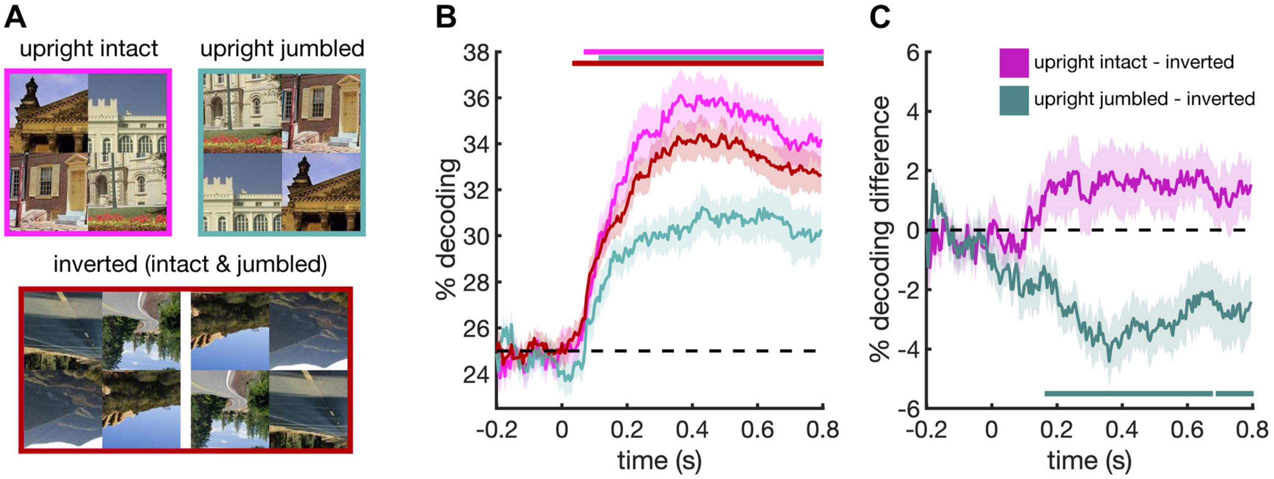

Fig. 4. Comparing category decoding between upright and inverted scenes. *A*: we compared the emergence of category information for the intact upright scenes, the jumbled upright scenes, and the inverted scenes; for the inverted scenes, we averaged across the intact and jumbled scenes, as no significant differences between the two were found. *B*: numerically, category decoding accuracy for the inverted scenes was in between the accuracies observed for the intact and jumbled upright scenes. *C*: when subtracting decoding in the inverted condition from decoding in the upright conditions, we found that statistically, category information was comparable for intact upright scenes and inverted scenes. By contrast, weaker category information was found for the jumbled upright scenes, compared with the inverted scenes, suggesting that jumbling specifically harms the emergence of category information in upright scenes. Error margins indicate standard errors of the difference. Significance markers (colored horizontal lines) indicate $P < 0.05$, corrected for multiple comparisons using threshold-free cluster enhancement.

tions of individual objects are rapidly enhanced (within the first 150 ms of vision) when the objects appear in their typical real-world locations, such as an eye in the upper visual field (Issa and DiCarlo 2012) or a shoe in the lower visual field (Kaiser et al. 2018). Together, these findings therefore support the idea that real-world structure can boost basic visual analysis across diverse stimuli and processing levels (Kaiser et al. 2019a).

When directly comparing neural category information in upright and inverted scenes, we found that it was equally pronounced when the scenes were intact and upright and when the scenes were inverted, regardless of their structural arrangement—only when the upright scenes were jumbled did we find significantly reduced category information. One interpretation of this result is that jumbling causes a specific disruption for upright scenes because for these scenes, the jumbling manipulation may be perceptually more salient. Alternatively, the pattern of results may be explained by an interaction of two different effects. The inverted intact scenes still retain the intact relative positioning of their parts, which may explain why they are better decodable than the upright jumbled scenes. The inverted jumbled scenes do not have this intact relative positioning, but by means of inversion they gain an intact absolute positioning of their parts (e.g., a piece of sky would be in the upper part of an inverted jumbled scene, which is where it belongs); this may explain why these scenes yield better category decoding than upright jumbled scenes. At this point, further studies are needed to fully understand this pattern of results. Challenges with interpreting inversion effects in the current paradigm may necessitate the inclusion of other low-level stimulus controls in these future studies.

Although our effects demonstrate an enhanced early representation of scenes that adhere to real-world structure compared with scenes that do not, studies on object-scene consistency suggest that EEG waveforms only become affected by typical object positioning after around 250 ms of vision (Coco et al. 2020; Draschkow et al. 2018; Ganis and Kutas 2003; Mudrik et al. 2010, 2014; Võ and Wolfe 2013). How do these early and late effects of scene structure relate to each other?

As one possibility, later effects may partly reflect increased responses to inconsistencies rather than an enhanced processing of consistent scene-object combinations (Faivre et al. 2019). Together with our results, these findings may suggest that early responses are biased toward scenes that predictably follow real-world structure, whereas later responses may be more biased toward violations of this structure. This idea is consistent with a recent proposal in predictive processing, which suggests a temporal succession of more general processing biases, first toward the expected and then toward the surprising (Press et al. 2020).

Alternatively, the beneficial effects of real-world regularities may not immediately result in consistency signals. Whether visual inputs generally are consistent with our real-world experience may only be analyzed following more basic visual analysis. Supporting this idea, generic consistency signals in our data only emerge later than the enhanced category processing. As previously reported, intact and jumbled scenes (independent of their category) evoked reliably different responses only after 255 ms of processing (Kaiser et al. 2020a).

More broadly, the findings can add to our understanding of efficient everyday vision. Even under challenging real-world conditions, human vision is remarkably efficient; in fact, it is much more efficient than findings from simplified laboratory experiments would predict (Wolfe et al. 2011; Peelen and Kastner 2014). Behavioral studies using jumbling paradigms have suggested that typical scene structure contributes to this efficiency. When scenes are structurally intact, observers can better categorize them (Biederman et al. 1974), recognize objects within them (Biederman 1972), or detect visual changes in the scene (Varakin and Levin 2008). These perceptual benefits may be linked to the rapid facilitation of neural category information for typical scenes observed in the current study. However, our participants performed an orthogonal fixation task, which precludes directly linking brain and behavior here. Future studies combining neural recordings with naturalistic behavioral tasks may reveal that the early cortical tuning to real-world structure may be a crucial asset for solving complex real-world tasks.

ACKNOWLEDGMENTS

We thank Sina Schwarze for help in EEG data collection.

GRANTS

D. Kaiser and R. M. Cichy are supported by Deutsche Forschungsgemeinschaft (DFG) grants (KA4683/2-1, CI241/1-1, CI241/3-1). R. M. Cichy is supported by a European Research Council Starting Grant (ERC-2018-StG 803370). G. Häberle is supported by a PhD fellowship of the Einstein Center for Neurosciences Berlin.

DISCLOSURES

No conflicts of interest, financial or otherwise, are declared by the authors.

AUTHOR CONTRIBUTIONS

D.K. and R.M.C. conceived and designed research; G.H. performed experiments; D.K. and G.H. analyzed data; D.K., G.H., and R.M.C. interpreted results of experiments; D.K. prepared figures; D.K. drafted manuscript; D.K., G.H., and R.M.C. edited and revised manuscript; D.K., G.H., and R.M.C. approved final version of manuscript.

REFERENCES

Abassi E, Papeo L. The representation of two-body shapes in the human visual cortex. *J Neurosci* 40: 852–863, 2020. doi:10.1523/JNEUROSCI.1378-19.2019.

Baldassano C, Beck DM, Fei-Fei L. Human-object interactions are more than the sum of their parts. *Cereb Cortex* 27: 2276–2288, 2017. doi:10.1093/cercor/bhw077.

Bar M. Visual objects in context. *Nat Rev Neurosci* 5: 617–629, 2004. doi:10.1038/nrn1476.

Biederman I. Perceiving real-world scenes. *Science* 177: 77–80, 1972. doi:10.1126/science.177.4043.77.

Biederman I, Rabinowitz JC, Glass AL, Stacy EW Jr. On the information extracted from a glance at a scene. *J Exp Psychol* 103: 597–600, 1974. doi:10.1037/h0037158.

Bilalić M, Lindig T, Turella L. Parsing rooms: the role of the PPA and RSC in perceiving object relations and spatial layout. *Brain Struct Funct* 224: 2505–2524, 2019. doi:10.1007/s00429-019-01901-0.

Brainard DH. The psychophysics toolbox. *Spat Vis* 10: 433–436, 1997. doi:10.1163/156856897X00357.

Chan AW, Kravitz DJ, Truong S, Arizpe J, Baker CI. Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nat Neurosci* 13: 417–418, 2010. doi:10.1038/nn.2502.

Coco MI, Nuthmann A, Dimigen O. Fixation-related brain potentials during semantic integration of object-scene information. *J Cogn Neurosci* 32: 571–589, 2020. doi:10.1162/jocn_a_01504.

de Haas B, Schwarzkopf DS, Alvarez I, Lawson RP, Henriksson L, Kriegeskorte N, Rees G. Perception and processing of faces in the human brain is tuned to typical feature locations. *J Neurosci* 36: 9289–9302, 2016. doi:10.1523/JNEUROSCI.4131-14.2016.

Dima DC, Perry G, Singh KD. Spatial frequency supports the emergence of categorical representations in visual cortex during natural scene perception. *Neuroimage* 179: 102–116, 2018. doi:10.1016/j.neuroimage.2018.06.033.

Draschkow D, Heikel E, Võ ML, Fiebach CJ, Sassenhagen J. No evidence from MVPA for different processes underlying the N300 and N400 incongruity effects in object-scene processing. *Neuropsychologia* 120: 9–17, 2018. doi:10.1016/j.neuropsychologia.2018.09.016.

Faivre N, Dubois J, Schwartz N, Mudrik L. Imaging object-scene relations processing in visible and invisible natural scenes. *Sci Rep* 9: 4567, 2019. doi:10.1038/s41598-019-38654-z.

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

Grootswagers T, Wardle SG, Carlson TA. Decoding dynamic brain patterns from evoked responses: a tutorial on multivariate pattern analysis applied to time series neuroimaging data. *J Cogn Neurosci* 29: 677–697, 2017. doi:10.1162/jocn_a_01068.

Issa EB, DiCarlo JJ. Precedence of the eye region in neural processing of faces. *J Neurosci* 32: 16666–16682, 2012. doi:10.1523/JNEUROSCI.2391-12.2012.

Jung TP, Makeig S, Humphries C, Lee TW, McKeown MJ, Iragui V, Sejnowski TJ. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology* 37: 163–178, 2000. doi:10.1111/1469-8986.3720163.

Kaiser D, Cichy RM. Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *J Neurophysiol* 120: 848–853, 2018. doi:10.1152/jn.00229.2018.

Kaiser D, Häberle G, Cichy RM. Cortical sensitivity to natural scene structure. *Hum Brain Mapp* 41: 1286–1295, 2020a. doi:10.1002/hbm.24875.

Kaiser D, Inciuraite G, Cichy RM. Rapid contextualization of fragmented scene information in the human visual system. *Neuroimage* 117045, 2020. doi:10.1016/j.neuroimage.2020.117045.

Kaiser D, Moeskops MM, Cichy RM. Typical retinotopic locations impact the time course of object coding. *Neuroimage* 176: 372–379, 2018. doi:10.1016/j.neuroimage.2018.05.006.

Kaiser D, Peelen MV. Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *Neuroimage* 169: 334–341, 2018. doi:10.1016/j.neuroimage.2017.12.065.

Kaiser D, Quek GL, Cichy RM, Peelen MV. Object vision in a structured world. *Trends Cogn Sci* 23: 672–685, 2019a. doi:10.1016/j.tics.2019.04.013.

Kaiser D, Stein T, Peelen MV. Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proc Natl Acad Sci USA* 111: 11217–11222, 2014. doi:10.1073/pnas.1400559111.

Kaiser D, Turini J, Cichy RM. A neural mechanism for contextualizing fragmented inputs during naturalistic vision. *eLife* 8: e48182, 2019b. doi:10.7554/eLife.48182.

Kim JG, Biederman I. Where do objects become scenes? *Cereb Cortex* 21: 1738–1746, 2011. doi:10.1093/cercor/bhq240.

Konkle T, Brady TF, Alvarez GA, Oliva A. Scene memory is more detailed than you think: the role of categories in visual long-term memory. *Psychol Sci* 21: 1551–1556, 2010. doi:10.1177/0956797610385359.

Lowe MX, Rajsic J, Ferber S, Walther DB. Discriminating scene categories from brain activity within 100 milliseconds. *Cortex* 106: 275–287, 2018. doi:10.1016/j.cortex.2018.06.006.

Mudrik L, Lamy D, Deouell LY. ERP evidence for context congruity effects during simultaneous object-scene processing. *Neuropsychologia* 48: 507–517, 2010. doi:10.1016/j.neuropsychologia.2009.10.011.

Mudrik L, Shalgi S, Lamy D, Deouell LY. Synchronous contextual irregularities affect early scene processing: replication and extension. *Neuropsychologia* 56: 447–458, 2014. doi:10.1016/j.neuropsychologia.2014.02.020.

Oliva A, Torralba A. The role of context in object recognition. *Trends Cogn Sci* 11: 520–527, 2007. doi:10.1016/j.tics.2007.09.009.

Oostenveld R, Fries P, Maris E, Schoffelen JM. FieldTrip: Open source software for advanced analysis of MEG, EEG, and invasive electrophysiological data. *Comput Intell Neurosci* 2011: 156869, 2011. doi:10.1155/2011/156869.

Oosterhof NN, Connolly AC, Haxby JV. CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Front Neuroinform* 10: 27, 2016. doi:10.3389/fninf.2016.00027.

Peelen MV, Kastner S. Attention in the real world: toward understanding its neural basis. *Trends Cogn Sci* 18: 242–250, 2014. doi:10.1016/j.tics.2014.02.004.

Press C, Kok P, Yon D. The perceptual prediction paradox. *Trends Cogn Sci* 24: 13–24, 2020. doi:10.1016/j.tics.2019.11.003.

Ramkumar P, Jas M, Pannasch S, Hari R, Parkkonen L. Feature-specific information processing precedes concerted activation in human visual cortex. *J Neurosci* 33: 7691–7699, 2013. doi:10.1523/JNEUROSCI.3905-12.2013.

Roberts KL, Humphreys GW. Action relationships concatenate representations of separate objects in the ventral visual system. *Neuroimage* 52: 1541–1548, 2010. doi:10.1016/j.neuroimage.2010.05.044.

Sassenhagen J, Draschkow D. Cluster-based permutation tests of MEG/EEG data do not establish significance of effect latency or location. *Psychophysiology* 56: e13335, 2019. doi:10.1111/psyp.13335.

Smith SM, Nichols TE. Threshold-free cluster enhancement: addressing problems of smoothing, threshold dependence and localisation in cluster inference. *Neuroimage* 44: 83–98, 2009. doi:10.1016/j.neuroimage.2008.03.061.

Varakin DA, Levin DT. Scene structure enhances change detection. *Q J Exp Psychol (Hove)* 61: 543–551, 2008. doi:10.1080/17470210701774176.

Võ ML, Boettcher SE, Draschkow D. Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Curr Opin Psychol* 29: 205–210, 2019. doi:10.1016/j.copsyc.2019.03.009.

Võ ML, Wolfe JM. Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychol Sci* 24: 1816–1823, 2013. doi:10.1177/0956797613476955.

Wolfe JM, Võ ML, Evans KK, Greene MR. Visual search in scenes involves selective and nonselective pathways. *Trends Cogn Sci* 15: 77–84, 2011. doi:10.1016/j.tics.2010.12.001.

Chapter 4

Project III: Coherent natural scene structure facilitates the extraction of task-relevant object information in visual cortex

The current chapter comprises the research article entitled "Coherent natural scene structure facilitates the extraction of task-relevant object information in visual cortex" which was published in *Neuroimage* in 2021. This third research project demonstrated that the presence of intact scene structure facilitates the extraction of object information from natural scenes in a task-dependent way.

Authors:

Daniel Kaiser, Greta Häberle, Radoslaw M. Cichy

Contributions:

Daniel Kaiser: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing –review and editing, Visualization, Supervision, Project administration, Funding acquisition. Greta Häberle: Investigation, Writing –review and editing. Radoslaw M. Cichy: Resources, Writing –review and editing, Supervision, Project administration, Funding acquisition.

Contributions to open and reproducible science:

To contribute to open and reproducible science, the paper is published in an open-access journal. The original article can be found here: doi: 10.1016/j.neuroimage.2021.118365. Data are publicly available on OSF: doi: 10.17605/osf.io/gs2t5.

Copyright note:

According to Elsevier, the author has the right to include this article in a thesis or dissertation provided that it is not separately downloadable and the thesis is not published commercially. The article is published under a CC BY-NC-ND license.

Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Coherent natural scene structure facilitates the extraction of task-relevant object information in visual cortex

Daniel Kaiser[a,*], Greta Häberle[b,c,d], Radoslaw M. Cichy[b,c,d,e]

[a] Department of Psychology, University of York, York, UK
[b] Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany
[c] Charité – Universitätsmedizin Berlin, Einstein Center for Neurosciences Berlin, Berlin, Germany
[d] Humboldt-Universität zu Berlin, Faculty of Philosophy, Berlin School of Mind and Brain, Berlin, Germany
[e] Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany

ARTICLE INFO

Keywords:
Object perception
Natural scene categorization
Real-world structure
Functional magnetic resonance imaging
Multivariate pattern analysis

ABSTRACT

Looking for objects within complex natural environments is a task everybody performs multiple times each day. In this study, we explore how the brain uses the typical composition of real-world environments to efficiently solve this task. We recorded fMRI activity while participants performed two different categorization tasks on natural scenes. In the object task, they indicated whether the scene contained a person or a car, while in the scene task, they indicated whether the scene depicted an urban or a rural environment. Critically, each scene was presented in an “intact” way, preserving its coherent structure, or in a “jumbled” way, with information swapped across quadrants. In both tasks, participants’ categorization was more accurate and faster for intact scenes. These behavioral benefits were accompanied by stronger responses to intact than to jumbled scenes across high-level visual cortex. To track the amount of object information in visual cortex, we correlated multi-voxel response patterns during the two categorization tasks with response patterns evoked by people and cars in isolation. We found that object information in object- and body-selective cortex was enhanced when the object was embedded in an intact, rather than a jumbled scene. However, this enhancement was only found in the object task: When participants instead categorized the scenes, object information did not differ between intact and jumbled scenes. Together, these results indicate that coherent scene structure facilitates the extraction of object information in a task-dependent way, suggesting that interactions between the object and scene processing pathways adaptively support behavioral goals.

1. Introduction

Despite the complexity of our everyday environments, perceiving objects embedded in natural scenes is remarkably efficient. This efficiency is illustrated by studies that require participants to categorize objects under conditions of limited visual exposure: For instance, participants can tell whether a scene contains an animal or not from just a single glance (Thorpe et al., 1996; Potter, 1975, 2012), and even when only limited attentional resources are available (Li et al., 2002).

The ability to effortlessly make such categorization responses is underpinned by the efficient extraction of object information in visual cortex. Neuroimaging research has shown that the category of task-relevant objects can be reliably decoded from fMRI activity patterns in visual cortex, even when the objects are embedded in complex natural scenes (Peelen et al., 2009; Peelen and Kastner, 2011; Seidl et al., 2012) or movies (Cukur et al., 2013; Nastase et al., 2017; Shahdloo et al., 2020). M/EEG studies demonstrate that object category is represented well within the first 200ms of vision, even when the object is shown under such naturalistic conditions (Cauchoix et al., 2014; Kaiser et al., 2016; VanRullen and Thorpe, 2001; Thorpe et al., 1996). Together, these results highlight that the cortical processing of objects appearing within rich real-world environments is surprisingly efficient.

This processing efficiency becomes less surprising if scene context is not just considered as a nuisance that puts additional strain on our visual resources. Indeed, contextual information can facilitate object processing (Bar, 2004): For instance, scene context allows for efficient allocation of attention (Torralba et al., 2006; Wolfe et al., 2011; Võ et al., 2019), or for disambiguating object information under uncertainty (Brandmann and Peelen, 2017; Oliva and Torralba, 2007). Such findings demonstrate that object and scene processing mechanisms interact with each other to enable the efficient processing of object information.

[*] Corresponding author at: Department of Psychology, University of York, Heslington, York, YO10 5DD, UK.
E-mail address: danielkaiser.net@gmail.com (D. Kaiser).

https://doi.org/10.1016/j.neuroimage.2021.118365.
Received 2 December 2020; Received in revised form 22 April 2021; Accepted 3 July 2021
Available online 4 July 2021.

Here, we investigated how the coherent spatial structure of the scene context aids the extraction of object information from the scene. To this end, we used a jumbling paradigm, in which we disrupted the scenes' coherent structure by dividing them into multiple rectangular pieces and shuffling those pieces. Classical studies suggest that jumbling drastically impairs participants' ability to categorize both the scene itself (Biederman et al., 1974), and the object embedded within the scene (Biederman et al., 1972, 1973). Such impairments can be linked to changes in cortical scene processing: We have recently shown that scene-selective brain responses are less pronounced and contain less scene category information when the scene is jumbled (Kaiser et al., 2020a, 2020b). However, it is unclear how these changes in scene-selective activations modulate the representation of objects within the scene.

In the current study, we thus set out to characterize how the presence of an intact – versus a jumbled – scene context modulates object representations in visual cortex. First, we asked whether cortical object processing is indeed facilitated by the presence of a coherent scene context. Second, we asked whether such facilitation effects depend on the objects being relevant or irrelevant for current behavioral goals.

To answer these questions, we recorded fMRI activity while participants categorized objects contained in intact or jumbled scenes. We found that fMRI responses across high-level visual cortex were generally higher for intact scenes than for jumbled scenes, revealing widespread sensitivity to scene structure. When analyzing object category information in multi-voxel response patterns, we found that coherent scene structure enhanced object information in object-selective visual cortex. However, this enhancement was task-specific: When participants categorized the scenes instead of the objects, we found no such enhancement of object information. These results suggest that the visual brain uses coherent real-world structure to more efficiently extract task-relevant object information from complex scenes.

2. Materials and methods

2.1. Participants

Twenty-five healthy adults (mean age 26.4 years, SD=5.3; 15 female, 10 male) participated. All participants had normal or corrected-to-normal vision. They all provided informed written consent and received either monetary reimbursement or course credits. Procedures were approved by the ethical committee of the Department of Psychology at Freie Universität Berlin and were in accordance with the Declaration of Helsinki.

2.2. Stimuli

The stimulus set consisted of colored natural scene photographs (640 × 480 pixels resolution). Scenes were selected to cover three independent manipulations. First, each scene contained one of two object categories: half of the scenes contained a person (or multiple people), whereas the other half contained a car (or multiple cars). Second, the person or car appeared equally often in each of the quadrants of the scene. Third, each scene belonged to one of two scene categories: half of the scenes depicted urban environments, the other half depicted rural environments. For each possible combination of these factors (e.g., a person appearing in the bottom left quadrant of a rural scene), 10 unique scene exemplars were available, yielding 160 scenes in total (2 object categories × 4 object locations × 2 scene categories × 10 exemplars). During the experiment, the scenes could be presented in their original orientation or mirrored along their vertical axis (as in Kaiser et al., 2016), yielding a total of 320 different scene stimuli. Example scenes are shown in Fig. 1a.

To manipulate scene structure, we either presented the scenes in a coherent, "intact" condition or in an incoherent, "jumbled" condition. Jumbled scenes were generated by shuffling the four quadrants of the image in a crisscrossed way (i.e., top-left was swapped with bottom-right, and top-right was swapped with bottom-left; Fig. 1b). This manipulation solely affected the scene's structure, but not the people or cars contained in the scene: First, as the objects never straddled the boundary between quadrants, the objects themselves always remained unaltered. Second, as the objects appeared equally often in each quadrant before jumbling the scenes, they also appeared equally often in each quadrant after jumbling them.

In total, 640 scene images were used, which covered 320 intact scenes and 320 jumbled scenes. Additionally, 200 colored texture masks (Kaiser et al., 2016) were used to visually mask the scenes during the experiment (see below).

2.3. Experimental paradigm

Each participant completed four experimental runs of 17 minutes each. Each run contained 320 experimental trials, corresponding to 320 unique scene stimuli. Both intact and jumbled scenes were included in each run. For half of the participants, the even runs only contained the original scenes, while the odd runs only contained the horizontally mirrored scenes; for the other half of the participants, the odd runs only contained the original scenes, while the even runs only contained the horizontally mirrored scenes. Each of the scenes was presented once during the run. Trial order was fully randomized for each participant and run.

On each trial, the scene was presented for 83ms, immediately followed by a visual mask (chosen randomly from the 200 available masks) for 800ms. Masks were shown to establish a sensitive performance range for reasonably long presentation times, as they disrupt ongoing visual processing after the offset of the stimulus. All images were shown within a black rectangle (10deg X 7.5deg visual angle). After an inter-trial interval of 1,617ms, during which a pink fixation dot was shown, the next trial started. An example trial is illustrated in Fig. 1c. In addition to the experimental trials, each run contained 80 fixation-only trials, during which only the fixation dot was displayed. Runs started and ended with a brief fixation period.

In two of the four runs, participants were asked to categorize the object contained in each scene as either a person or a car ("object task"). In the other two runs, participants were asked to categorize the scene as either a rural or an urban environment ("scene task"). Participants were instructed to respond as accurately and quickly as possible, with an emphasis on accuracy. Button-press responses were recorded during the whole inter-trial interval (i.e., until 2,500s after stimulus onset). The four runs were alternating between the object and scene tasks. The task in the first run was counter-balanced between participants. Notably, physical stimulation was completely identical across the object and scene tasks.

All stimuli were back-projected onto a translucent screen mounted to the head end of the scanner bore. Participants viewed the stimulation through a mirror attached to the head coil. Stimulus presentation was controlled using the Psychtoolbox (Brainard, 1997).

2.4. Benchmark localizer paradigm

In addition to the experimental runs, each participant completed a benchmark localizer run, which was designed to obtain "benchmark" patterns in response to people and cars in isolation (Peelen et al., 2009; Peelen and Kastner, 2011). During this run, participants viewed images of bodies, cars, and scrambled images of bodies and cars. For each of the three categories, 40 images were used. All images were different than the ones used in the main experiment. These images were presented in a block design. Each block lasted 20 seconds and contained 20 images of one of the three categories, or only a fixation cross. Images were presented for 500ms (5deg × 5deg visual angle), separated by a 500ms inter-stimulus interval. The benchmark localizer run consisted

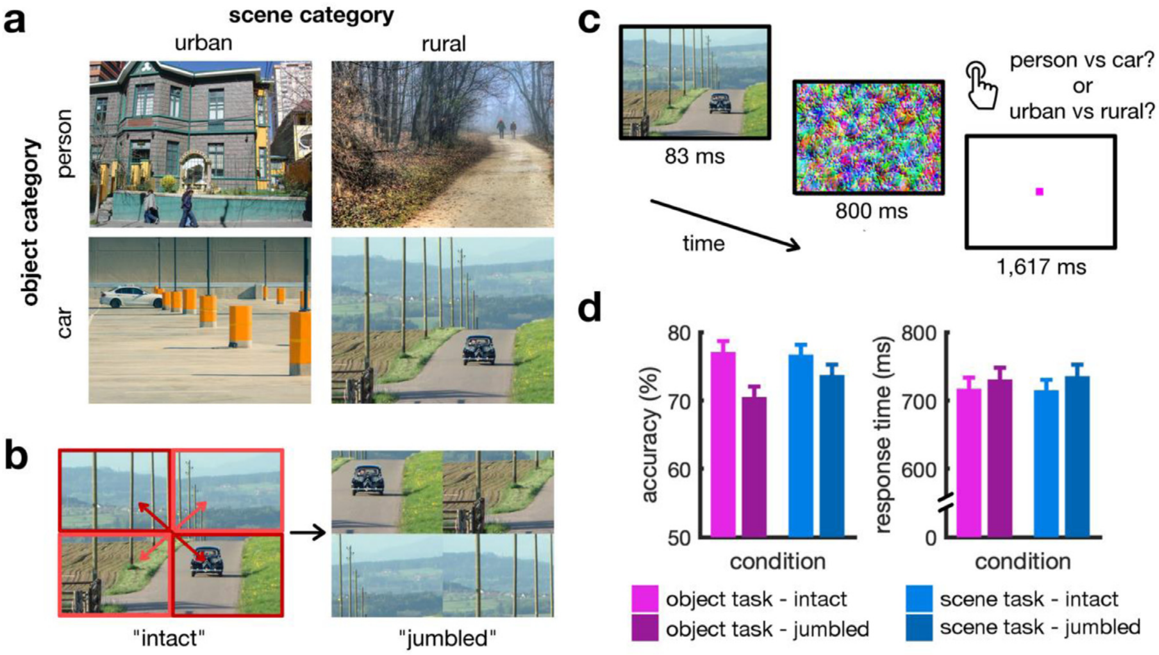

Fig. 1. Stimuli, paradigm, and behavioral results. **a)** Stimuli consisted of natural scene images from two categories: urban or rural environments. Each of the scenes contained one of two object categories: people or cars. **b)** During the experiment, these scenes were shown in an unaltered way ("intact" condition) or with their quadrants intermixed ("jumbled" condition). The jumbled scenes were created by shuffling the quadrants in a crisscrossed way, as illustrated. **c)** Participants viewed each scene briefly, followed by a visual mask. In separate runs, they performed two different tasks: They were either asked to indicate whether the scene contained a person or a car ("object task") or whether the scene depicted an urban or a rural environment ("scene task"). **d)** In both tasks, scene structure impacted behavioral performance: Participants were significantly more accurate and faster for the intact scenes than for the jumbled scenes. Error bars represent standard errors of the mean.

of a total of 24 blocks (6 blocks for each of the three stimulus categories, and 6 fixation-only blocks). Four consecutive blocks always contained the four different conditions in random order. Participants were instructed to respond to one-back image repetitions (i.e., two identical images back-to-back), which happened once during each non-fixation block. The benchmark localizer run lasted 8:30 minutes and was completed halfway through the experiment, after two of the four experimental runs.

2.5. fMRI recording and preprocessing

MRI data was acquired using a 3T Siemens Magnetom Tim Trio Scanner equipped with a 12-channel phased-array head coil. T2*-weighted gradient-echo echo-planar images were collected as functional volumes, with the following parameters: TR=2s, TE=30ms, 70° flip angle, 3mm3 voxel size, 37 slices, 20% slice gap, 192mm FOV, 64 × 64 matrix size, interleaved acquisition, A/P phase encoding, acquisition time 17min (main experiment) / 8:20min (benchmark localizer), whole-brain coverage, ACPC orientation. Additionally, a T1-weighted 3D MPRAGE image was obtained as an anatomical reference, with the following parameters: TR=1.9s, TE=2.52ms, 9° flip angle, 1mm3 voxel size, 176 slices, 50% slice gap, 256mm FOV, ascending acquisition, A/P phase encoding, acquisition time 4:26min, whole-brain coverage. All acquisitions contained four initial dummy volumes that were discarded later.

Preprocessing and hemodynamic response modelling was performed using SPM12 (www.fil.ion.ucl.ac.uk/spm/). Functional volumes were realigned and coregistered to the anatomical image. Further, transformation parameters to MNI-305 standard space were obtained using the "segmentation" routine in SPM12.

Functional data from each experimental run were modelled in a general linear model (GLM) with 16 experimental predictors (2 object categories × 4 object locations × 2 scene categories). Additionally, we included the six movement regressors obtained during realignment. Data from the benchmark localizer run were modelled in a GLM with three experimental predictors (person, car, scrambled) and six movement regressors.

2.6. Region of interest definition

We restricted fMRI analyses to five regions of interest (ROIs): early visual cortex (EVC), object-selective lateral occipital cortex (LO), body-selective extrastriate body area (EBA), scene-selective occipital place area (OPA), and scene-selective parahippocampal place area (PPA). ROIs masks were defined using group-level activation masks from functional brain atlases: For EVC, we selected all voxels that were most probably assigned to primary visual cortex (V1v, V1d) in the Wang et al. (2015) atlas, and for LO, EBA, OPA, and PPA we selected region masks from the Julian et al. (2012) atlas. ROIs were defined separately for each hemisphere. All ROI masks were inverse-normalized into individual-participant space using the parameters obtained during T1 segmentation. Average voxel counts in individual-participant space amounted to 248/271 (EVC; SD=42/41, left/right), 929/947 (LO; SD=103/102), 402/443 (EBA; SD=45/52), 26/47 (OPA; SD=5/8), and 140/105 (PPA; SD=14/10). Notably, the LO and EBA ROIs overlapped to some extent (300/406 voxels overlap, left/right); the inclusion of the EBA allowed us to see whether the results hold in a smaller cortical region with a narrower category preference for bodies. As we did not have any hypothesis related to hemispheric differences, all results for the left- and right-hemispheric ROIs were averaged before statistical analysis. Separate results for the right- and left-hemispheric ROIs are reported in the Supplementary Information.

2.7. Univariate analysis

Response magnitudes during the experimental runs were analyzed separately for each ROI. We first averaged beta values across the two object-task and scene-task runs, respectively. We then averaged beta values across object categories, object locations, and scene categories.

This way, we obtained response magnitudes for four conditions: (1) responses to intact scenes in the object task, (2) responses to jumbled scenes in the object task, (3) responses to intact scenes in the scene task, and (4) responses to jumbled scenes in the scene task. These four conditions allowed us to separately estimate the effects of task (object task versus scene task) and scene structure (intact versus jumbled) on neural responses across the five ROIs. For a univariate analysis of category-specific responses across the two tasks, see the Supplementary Information.

2.8. Multivariate pattern analysis

Multivariate pattern analysis (MVPA) was carried out in CoSMoMVPA (Oosterhof et al., 2016). Our MVPA approach closely followed similar fMRI studies that investigated the representation of objects in natural scenes (Peelen et al., 2009; Peelen and Kastner, 2011). We first computed a one-sample t-contrasts for every condition against baseline (i.e., against the fixation trials). In the benchmark localizer run, there were 2 such t-contrasts (one for people versus baseline, and one for cars versus baseline). In the object task and scene task runs, there were 16 t-contrasts each (one contrast for each experimental condition against baseline, reflecting 2 object categories × 4 object locations × 2 scene categories). For each of the three tasks (benchmark localizer, object task, and scene task), the resulting t-values were normalized for each voxel by subtracting the average t-value across conditions. For each ROI, multi-voxel response patterns were constructed by concatenating the t-values across all voxels belonging to the ROI.

To obtain an index of object discriminability (i.e., how discriminable people and cars in scenes are based on multi-voxel response patterns), we performed a correlation-based MVPA. The goal of this analysis was to quantify how “person-like” or “car-like” the cortical representation of each of the scenes was, thereby isolating the amount of object category information in visual cortex (note that each of the scenes either contained a person or a car). To this end, we correlated multi-voxel response patterns evoked by people and cars in isolation (from the benchmark localizer) with response patterns evoked by people and cars contained in a scene (from one of the experimental tasks). These correlations were Fisher-transformed. To quantify object discriminability, we then subtracted the correlations between different categories (e.g., person in isolation and car within a scene) from correlations between the same categories (e.g., person in isolation and person within a scene). This yielded an index of category-discriminability, with values greater than zero indicating that the two categories are represented differently (Haxby et al., 2001). Results for different analysis routines (using Spearman correlations and no mean-removal across conditions) can be found in the Supplementary Information.

Before performing this analysis, response patterns in the main experiment were averaged across object locations and scene categories. This way, we obtained an index of object category-discriminability for four separate conditions: (1) category-discriminability for intact scenes in the object task, (2) category-discriminability for jumbled scenes in the object task, (3) category-discriminability for intact scenes in the scene task, and (4) category-discriminability for jumbled scenes in the scene task. The resulting four conditions allowed us to estimate the effects of scene structure on the quality of object representations in visual cortex, both when the objects were task-relevant and task-irrelevant.

2.9. Statistical testing

To compare behavioral performance, univariate responses, and multi-voxel pattern information across conditions, we used repeated-measures ANOVAs and paired-sample t-tests. We report partial eta-squared (η_p^2, for F-tests) and Cohen’s d (for t-tests) as measures of effect size. Descriptive statistics (means and standard errors) are reported in the Supplementary Information.

2.10. Data availability

Data are publicly available on OSF (doi.org/10.17605/osf.io/gs2t5). Other materials are available from the corresponding author upon request.

3. Results

3.1. Coherent scene structure facilitates the perception of objects within scenes

We first analyzed participants’ behavioral performance in the object and scene tasks, separately for the intact and jumbled scenes (Fig. 1d). In the object task, participants’ categorization (person versus car) of objects within the intact scenes was more accurate, t(24)=8.28, p<.001, d=1.61, and faster, t(24)=3.26, p=.0033, d=0.65, compared to the jumbled scenes. In the scene task, participants’ categorization (rural versus urban) of the intact scenes was more accurate, t(24)=4.77, p<.001, d=0.95, and faster, t(24)=3.26, p=.0033, d=0.65, compared to the jumbled scenes. These results are in line with classical findings on object and scene categorization in jumbling paradigms (Biederman, 1972; Biederman et al., 1973, 1974), showcasing that scene jumbling has a profound impact on perception.

Further, when directly comparing the two tasks, we did not find differences in accuracy, F(1,24)=3.13, p=.090, or response times, F(1,24)=0.04, p=.84. Any differences in neural responses are therefore unlikely to reflect differences in task difficulty, and therefore attentional engagement, between the two tasks.

Together, these results demonstrate that jumbling similarly impairs the perception of the scene and the objects contained in it, demonstrating a cross-facilitation between scene and object vision that can be observed on the behavioral level.

3.2. Scene structure impacts univariate responses across object- and scene-selective cortex

To quantify the effects of scene jumbling on the neural level, we first ran univariate analyses. In these analyses, we compared fMRI response magnitudes across the intact and jumbled scenes and across the two tasks (Fig. 2). To do so, we performed a 2 × 2 repeated measures ANOVA with the factors scene structure (intact versus jumbled) and task (object task versus scene task). The analysis was performed separately and in turn for each of the five ROIs: EVC, LO, EBA, OPA, and PPA. Detailed results for these analyses can be found in Table 1.

In EVC, responses were comparable across all conditions, all F<1.25, p>.27, η_p^2<0.06, suggesting that EVC is not sensitive to typical scene composition.

In all extrastriate ROIs, we found a main effect of scene structure, which indicated stronger responses to intact than to jumbled scenes, all F(1,24)>7.95, p<.010, η_p^2>0.24. Comparing this effect across regions, we found that it was more pronounced in the scene-selective regions, OFA versus LO/EBA, both F(1,24)>31.17, p<.001, η_p^2>0.56, and PPA versus LO/EBA, both F(1,24)>35.54, p<.001, η_p^2>0.59. This finding confirms our previous fMRI results, which revealed particularly strong effects of scene jumbling in scene-selective areas of visual cortex (Kaiser et al., 2020a).

In all ROIs, scene structure affected univariate responses similarly across the two tasks, as indexed by no significant interaction effects, all F<2.46, p>.12, η_p^2<0.10. This pattern of results mirrors the pattern observed in behavior, where scene jumbling produced comparable effects in the object and scene tasks.

PPA was the only region that additionally showed an effect of task, F(1,24)=6.51, p=.017, η_p^2=0.21, with stronger responses in the scene task compared to the object task. This suggests an increased importance of computations in higher-level scene-selective cortex when scene attributes were behaviorally relevant.

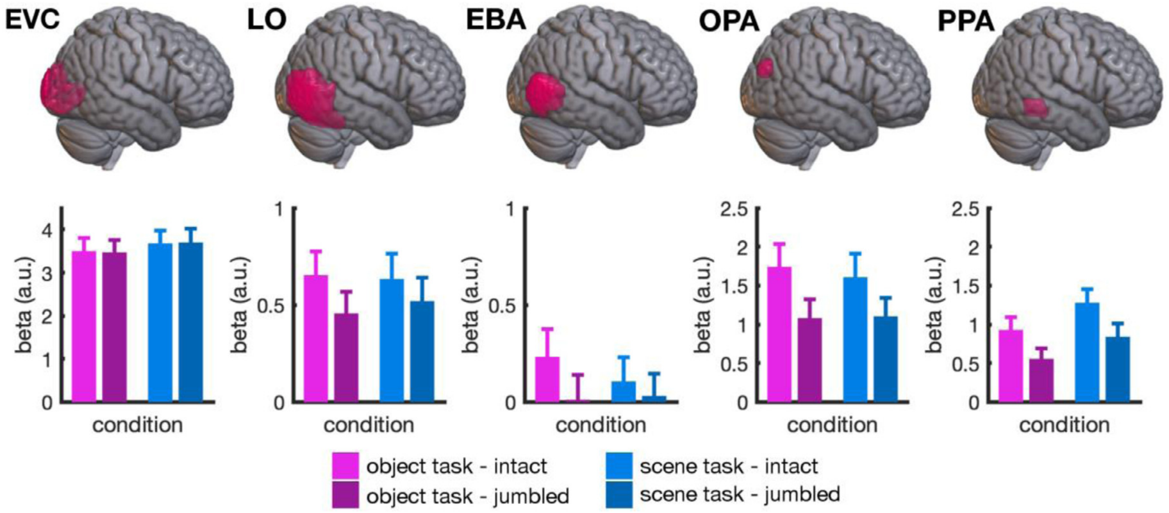

Fig. 2. Univariate results. In all extrastriate regions, but not in EVC, we found a significant main effect of scene structure: Intact scenes led to significantly stronger responses than jumbled scenes. This effect was comparable across the two tasks and most pronounced in scene-selective ROIs. PPA was the only region that additionally showed a modulation by task, with significantly stronger responses when participants were categorizing the scenes, compared to when they were categorizing the objects within them. For illustration purposes, ROI masks are shown on the right hemisphere of a standard-space template using MRIcroGL (Li et al., 2016); the displayed results are averaged across ROIs in both hemispheres. Error bars represent standard errors of the mean.

Table 1
Univariate responses, analyzed in a 2 × 2 repeated measures ANOVA with the factors scene structure (intact versus jumbled) and task (object task versus scene task). Significant effects are highlighted in bold.

ROI	*Main effect scene structure*	*Main effect task*	*Interaction* effect structure × *Task*
EVC	F(1,24)<0.01, p=.98, $\eta_p{}^2$<0.01	F(1,24)=1.25, p=.28, $\eta_p{}^2$=0.05	F(1,24)=0.09, p=.76, $\eta_p{}^2$<0.01
LO	**F(1,24)=9.74, p=.005, $\eta_p{}^2$=0.29**	F(1,24)=0.04, p=.85, $\eta_p{}^2$<0.01	F(1,24)=0.97, p=.33, $\eta_p{}^2$=0.04
EBA	**F(1,24)=7.95, p=.009, $\eta_p{}^2$=0.25**	F(1,24)=0.21, p=.65, $\eta_p{}^2$<0.01	F(1,24)=2.46, p=.13, $\eta_p{}^2$=0.09
OPA	**F(1,24)=27.18, p<.001, $\eta_p{}^2$=0.53**	F(1,24)=0.09, p=.77, $\eta_p{}^2$<0.01	F(1,24)=0.97, p=.34, $\eta_p{}^2$=0.04
PPA	**F(1,24)=48.02, p<.001, $\eta_p{}^2$=0.67**	**F(1,24)=6.51, p=.017, $\eta_p{}^2$=0.21**	F(1,24)=0.51, p=.48, $\eta_p{}^2$=0.02

Having established that scene structure enhanced cortical responses across object- and scene-selective cortex, and similarly for both tasks, we next asked how scene structure contributes to the extraction of object information – both when the objects are behaviorally relevant and when they are not.

3.2. *Coherent scene structure enhances task-relevant object information in multi-voxel response patterns*

To understand how the coherent spatial structure of the scene impacts cortical object processing, we performed a correlation-based multivariate pattern analysis (MVPA). In this analysis, we correlated the multi-voxel response patterns evoked by objects embedded in scenes (from the object and scene tasks) with the patterns evoked by the objects in isolation (from the benchmark localizer) (Fig. 3a). This approach allowed us to quantify how "person-like" or "car-like" the cortical representation of each of the scene conditions was, thereby isolating the amount of object information present in visual cortex (note that each of the scenes either contained a person or a car). When object information is operationalized in this way, it can be separated from differences in the scene context (as in the benchmark localizer no scene context is presented) and task-related differences (as in the benchmark localizer participants perform a different task).

To quantify object information, we computed a correlation measure by subtracting correlations between different categories (e.g., person in isolation and car within a scene) from correlations between the same categories (e.g., person in isolation and person within a scene) (Fig. 3a). This measure was computed separately for each of the object and scene tasks, the intact and jumbled scenes, and all ROIs.

To test whether multi-voxel response patterns contained any information at all about the object contained in the scenes, we first averaged the correlation measure across all conditions. We then tested whether the average category information was significantly different from zero, separately for each ROI. As expected, people and cars could be reliably discriminated from response patterns in the object-selective LO, t(24)=7.56, p<.001, d=1.51, and body-selective EBA, t(24)=8.00, p<.001, d=1.60, but not from response patterns in EVC, t(24)=0.80, p=.43, d=0.18, or scene-selective OPA, t(24)=0.49, p=.63, d=0.10, and PPA, t(24)=0.70, p=.49, d=0.14.

Given that we only found robust object information in LO and EBA, we only performed further analyses for these two regions (Fig. 3b). Data were again analyzed in a 2 × 2 ANOVA with factors scene structure (intact vs jumbled) and task (object task vs scene task), separately for LO and EBA.

When analyzing the amount of object information contained in LO response patterns, we found a significant interaction between task and scene structure, F(1,24)=5.63, p=.026, $\eta_p{}^2$=0.19: When participants performed the object task, object information in LO was more pronounced for objects embedded in intact compared to jumbled scenes, t(24)=2.65, p=.014, d=0.53. This effect was absent when participants performed the scene task, t(24)=1.22, p=.24, d=0.24. A similar interaction effect was found in the EBA, F(1,24)=5.19, p=.032, $\eta_p{}^2$=0.18: Object information was again enhanced for intact scenes during the object task, t(24)=2.30, p=.030, d=0.46, but not during the scene task, t(24)=0.92, p=.37, d=0.18. These results demonstrate that coherent scene structure indeed enhances object representations in visual cortex. However, this enhancement depends on the behavioral relevance of the object: When scene category, rather than object category, was task-relevant, no such enhancement was observed.

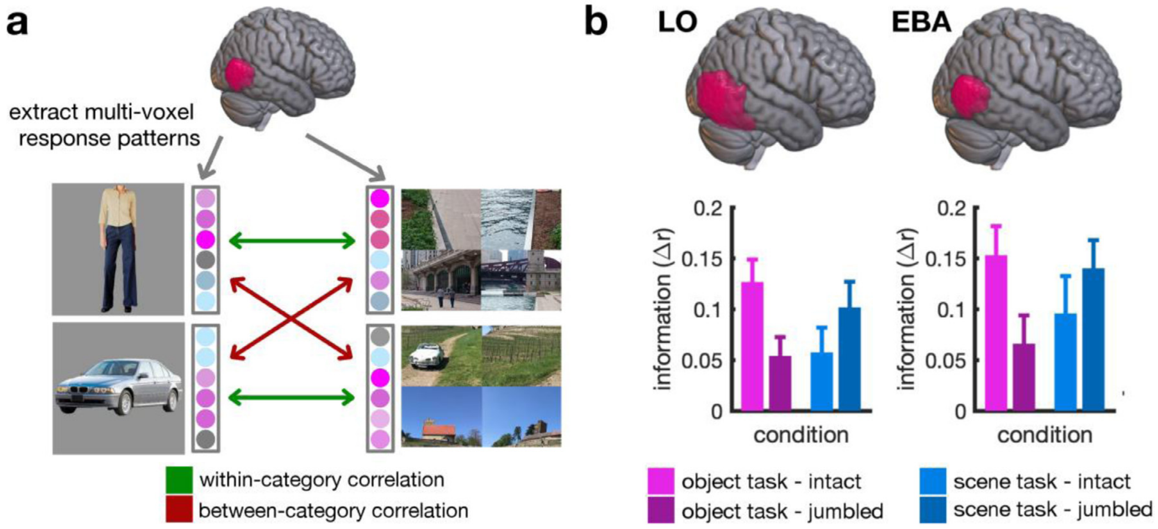

Fig. 3. Correlation MVPA logic and results. **a)** To measure object discriminability, we extracted multi-voxel response patterns for each ROI, separately for objects in isolation (from the benchmark localizer) and objects appearing within the scenes (from the main experiment). We then computed within- and between-category correlations. By subtracting the between-category from the within-category correlations, we obtained an index of category information (Δr). **b)** In both LO and EBA, category information was significantly higher for objects that were embedded in intact scenes than for objects embedded in jumbled scenes. However, this was only true when participants performed the object task; when they performed the scene task, no significant difference in object category information was observed when comparing intact and jumbled scenes. For illustration purposes, ROI masks are shown on the right hemisphere of a standard-space template using MRIcroGL (Li et al., 2016); the displayed results are averaged across ROIs in both hemispheres. Error bars represent standard errors of the mean.

4. Discussion

4.1. *Coherent scene structure facilitates task-relevant object processing*

In this study, we shed light on neural object processing in situations where the object is either embedded within a coherent, intact scene or an incoherent, jumbled scene. Consistent with classical studies (Biederman, 1972; Biederman et al., 1973, 1974), our participants were more accurate and faster in perceiving intact, compared to jumbled scenes, both when performing an object categorization task and a scene categorization task. Our univariate findings are consistent with previous fMRI work (Kaiser et al., 2020a): We replicate the finding that intact scenes yield stronger neural responses than jumbled scenes, across high-level visual cortex and prominently in scene-selective regions. This suggests a widespread sensitivity to typical scene structure in the visual system. Importantly, our current results show that scene structure also matters when it comes to the neural representation of objects within the scene: When analyzing the amount of object information contained in multi-voxel response patterns in object and body-selective visual cortex, we found an enhancement of object information when the objects were embedded within intact scenes, compared to jumbled scenes. Critically, this enhancement only emerged in the object categorization task, suggesting that coherent scene structure facilitates the extraction of object information only when the objects are relevant for current behavioral goals.

4.2. *Interactions between object and scene processing are mediated by scene structure*

Our findings support the view that the scene and object processing pathways are not functionally separate, but that scene information can aid the extraction of object information (Brandmann and Peelen, 2017). Theories of contextual facilitation propose that scene structure is analyzed rapidly, potentially based on coarse low-spatial frequency information (Bar, 2004; Bar et al., 2006). This idea is consistent with the observation that an initial representation of scene meaning – the scene's "gist" – can be extracted from just a single glance (Greene and Oliva, 2009; Oliva and Torralba, 2006, 2007). Contextual facilitation theories argue that detailed object analysis is facilitated by this more readily available information about scene gist (Bar, 2004; Hochstein and Ahissar, 2002). Informing object analysis through the analysis of coarse scene properties may be particularly useful when perception is challenged by the presence of many distracter items and limited visual exposure. Probing perception with such a challenging task, our study shows that the cross-facilitation between object and scene processing is mediated by the scene's structural coherence: When the analysis of scene gist is disrupted by jumbling the scene, contextual information cannot amplify object processing in the same way as it can for intact scenes.

The enhanced extraction of object information from the intact scenes suggests that useful information about scene gist is extracted less efficiently from the jumbled scenes. Indeed, the rapid analysis of scene gist depends on our priors about typical scene composition (Csathó et al., 2015; Greene et al., 2015). Neuroimaging studies suggest that the cortical scene processing network is tuned to these priors (Kaiser et al., 2020a; Torralbo et al., 2013), and that the early extraction of properties like the scene's basic-level category depends on the structural coherence of the scene (Kaiser et al., 2020b). Jumbling is a strong manipulation in the sense that is disrupts multiple aspects of the scene's spatial coherence at the same time: it disrupts the spatial positioning of individual pieces of information in visual space (Kaiser and Cichy, 2018; Mannion, 2015), the positioning of objects relative to each other (Kaiser et al., 2019; Kaiser and Peelen, 2018), as well as the typical geometry of the scene (Dillon et al., 2018; Spelke and Lee, 2012). Future research is needed to disentangle these different factors, and how much they each contribute to the facilitation of object representation.

Alternatively, one could argue that the jumbling manipulation generates a more general "artificiality" in the stimuli (through the salient borders between quadrants of the jumbled images) that puts additional strain on the visual system. Based on this assertion, one would predict lower responses for jumbled scenes. In previous studies (Kaiser et al., 2020a, 2020b), we have shown that strong effects of scene jumbling are also obtained when introducing similar artificial discontinuities to the typical scenes, suggesting that the degree of image artificiality introduced by the jumbling manipulation alone cannot explain the results.

However, although jumbling is a strong manipulation that conflates multiple factors of scene structure, it preserves critical characteristics of the objects: First, the objects remain completely unaltered across the intact and jumbled scenes. Second, the objects' absolute positions in visual space were matched across the intact and jumbled scenes. Finally, each object's local visual context remains constant across the intact and jumbled scenes. These properties allow us to attribute differences in object representations to facilitates effects from cortical scene analysis: If the visual brain would not take global scene context into account and would only analyze the objects in their local visual surroundings, our paradigm should yield comparable results for structurally coherent, intact scenes and incoherent, jumbled scenes.

4.2. Attention mediates contextual facilitation effects

Unlike task-relevant objects, task-irrelevant objects were not processed differently as a function of scene coherence. This finding shows that contextual facilitation of object processing is not an automatic process. On the contrary, interactions between the object and scene processing systems seem to be mediated by attention. This observation fits well with previous results from studies on object detection in natural scenes. Compared to task-relevant objects, multi-voxel response patterns in visual cortex contain far less information about unattended objects (Peelen et al., 2009; Peelen and Kastner, 2011). Further, MEG decoding results suggest strong differences in the representation of attended and unattended object categories (Kaiser et al., 2016): Particularly at early stages of processing, within the first 200ms after stimulus onset, the category of unattended objects is represented less accurately. Beyond the visual brain, differences in task demands also affect more widespread activations across the cortex (Cukur et al., 2013; Harel et al., 2014; Hebart et al., 2018; Nastase et al., 2017), potentially causing substantial task-related changes in processing dynamics. One such change may be an alteration of the crosstalk between representations in different visual domains. Our data indeed suggests that the exchange of information between the object and scene processing pathways is not mandatory, but rather constitutes an adaptive mechanism for improving task performance. Under this view, interactions between the scene and object processing pathways may be specifically "switched on" when objects are part of current attentional templates (Battistoni et al., 2017; Peelen and Kastner, 2011). The specific mechanism underlying this adaptive control of the crosstalk between scene and object processing needs further investigation.

How does the apparent importance of attention tie in with previous studies that reported a cross-facilitation between the object and scene-processing systems (Brandmann and Peelen, 2017, 2019)? While these studies did not use object categorization tasks, they still explicitly asked participants to attend to the objects appearing within the scene (either by asking them to memorize them or through one-back tasks). In our scene categorization task, the situation was entirely different, as the objects were completely irrelevant for solving the task. In fact, this orthogonality of object and scene category in our design may have introduced an active suppression of object information when participants performed the scene categorization task. Previous studies suggest that task-irrelevant distracter objects can be suppressed effectively and quickly (Seidl et al., 2012; Hickey et al., 2019). During the scene task, we indeed found numerically better object representations for jumbled scenes. This tentative reversal of the facilitation effect could potentially hint at a more efficient suppression of object information when the object is embedded in a structurally coherent scene. However, as this reversal is not statistically significant in our data and is somewhat susceptible to changes in analysis choices (see Supplementary Information), this assertion is largely speculative at this point. As another interesting observation, object information for the jumbled scenes was comparable between the object and scene tasks, suggesting that attention cannot as efficiently amplify object information when the scene is jumbled. However, some caution needs to be applied when directly comparing representations across the two tasks (rather than comparing differences between conditions), because different task-specific demand characteristics and attentional requirements complicate the interpretation of such comparisons.

4.3. Conclusion

In conclusion, our results show that the object and scene processing pathways can interact to facilitate the processing of task-relevant object information embedded in coherent scenes. However, such interactions are not mandatory. They rather seem to be guided by current behavioral goals. Our findings therefore suggest that the visual brain adaptively exploits coherent scene context to resolve object perception in challenging real-world situations.

Data availability: Data are publicly available on OSF (doi.org/10.17605/osf.io/gs2t5). Other materials are available from the corresponding author upon request.

Credit authorship contribution statement

Daniel Kaiser: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing – review & editing, Visualization, Supervision, Project administration, Funding acquisition. **Greta Häberle:** Investigation, Writing – review & editing. **Radoslaw M. Cichy:** Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Acknowledgements

We thank Sina Schwarze for helping with the stimulus collection. D.K. and R.M.C. are supported by Deutsche Forschungsgemeinschaft (DFG) grants (KA4683/2-1, CI241/1-1, CI241/3-1). G.H. is supported by a PhD fellowship of the Einstein Center for Neurosciences Berlin. R.M.C. is supported by a European Research Council Starting Grant (ERC-2018-StG 803370).

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.neuroimage.2021.118365.

References

Bar, M., 2004. Visual objects in context. Nat. Neurosci. 5, 617–629.

Bar, M, Kassam, KS, Ghuman, AS, Boshyan, J, Schmid, AM, Dale, AM, Hämäläinen, MS, Marinkovic, K, Schacter, DL, Rosen, BR, Halgren, E., 2006. Top-down facilitation of visual recognition. Proc. Natl. Acad. Sci. USA, 103, 449–454.

Battistoni, E, Stein, T, Peelen, MV., 2017. Preparatory attention in visual cortex. Ann. N. Y. Acad. Sci. 1396, 92–107.

Biederman, I., 1972. Perceiving real-world scenes. Science 177, 77–80.

Biederman, I, Glass, AL, Stacy, EW., 1973. Searching for objects in real-world scenes. J. Exp. Psychol. 97, 22–27.

Biederman, I, Rabinowitz, JC, Glass, AL, Stacy, EW., 1974. On the information extracted from a glance at a scene. J. Exp. Psychol. 103, 597–600.

Brandman, T, Peelen, MV., 2017. Interaction between scene and object processing revealed by human fMRI and MEG decoding. J. Neurosci. 37, 7700–7710.

Brandman, T, Peelen, MV., 2019. Signposts in the fog: objects facilitate scene representations in left scene-selective cortex. J. Cogn. Neurosci. 31, 390–400.

Brainard, DH., 1997. The psychophysics toolbox. Spat. Vis. 10, 433–436.

Cauchoix, M, Barragan-Jason, G, Serre, T, Barbeau, EJ., 2014. The neural dynamics of face detection in the wild revealed by MVPA. J. Neurosci. 34, 846–854.

Csathó, Á, van der Linden, D, Gács, B., 2015. Natural scene recognition with increasing time-on-task: the role of typicality and global image properties. Q. J. Exp. Psychol. 68, 814–828.

Cukur, T, Nishimoto, S, Huth, AG, Gallant, JL., 2013. Attention during natural vision warps semantic representation across the human brain. Nat. Neurosci. 16, 763–770.

Dillon, MR, Persichetti, AS, Spelke, ES, Dilks, DD., 2018. Places in the brain: bridging layout and object geometry in scene-selective cortex. Cereb. Cortex 28, 2365–2374.

Greene, MR, Botros, AP, Beck, DM, Fei-Fei, L., 2015. What you see is what you expect: rapid scene understanding benefits from prior experience. Atten. Percept. Psychophys. 77, 1239–1251.

Greene, MR, Oliva, A., 2009. Recognition of natural scenes from global properties: seeing the forest without representing the trees. Cogn. Psychol. 58, 137–176.

Harel, A, Kravitz, DJ, Baker, CI., 2014. Task context impacts visual object processing differentially across the cortex. Proc. Natl. Acad. Sci. USA, 111, 962–971.

Haxby, JV, Gobbini, IM, Furey, ML, Ishai, A, Schouten, JL, Pietrini, P., 2001. Distributed and overlapping representations of faces and objects in ventral temporal cortex. Science 293, 2425–2430.

Hebart, MN, Bankson, BB, Harel, A, Baker, CI, Cichy, RM., 2018. eLife 7, e32816.

Hickey, C, Pollicino, D, Bertazzoli, G, Barbaro, L., 2019. Ultrafast object detection in naturalistic vision relies on ultrafast distractor suppression. J. Cogn. Neurosci. 31, 1563–1572.

Hochstein, S, Ahissar, M., 2002. View from the top: hierarchies and reverse hierarchies in the visual system. Neuron 36, 791–804.

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

Kaiser, D, Cichy, RM., 2018. Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. J. Neurophysiol. 120, 848–853.

Kaiser, D, Häberle, G, Cichy, RM., 2020a. Cortical sensitivity to natural scene structure. Hum. Brain Mapp. 41, 1286–1295.

Kaiser, D, Häberle, G, Cichy, RM., 2020b. Real-world structure facilitates the rapid emergence of scene category information in visual brain signals. J. Neurophysiol. 124, 145–151.

Kaiser, D, Oosterhof, NN, Peelen, MV., 2016. The neural dynamics of attentional selection in natural scenes. J. Neurosci. 36, 10522–10528.

Kaiser, D, Quek, GL, Cichy, RM, Peelen, MV., 2019. Object vision in a structured world. Trends Cogn. Sci. 23, 672–685.

Kaiser, D, Peelen, MV., 2018. Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. Neuroimage 169, 334–341.

Li, FF, VanRullen, R, Koch, C, Perona, P., 2002. Rapid natural scene categorization in the near absence of attention. Proc. Natl. Acad. Sci. USA 99, 9596–9601.

Li, X, Morgan, PS, Ashburner, J, Smith, J, Rorden, C., 2016. The first step for neuroimaging data analysis: DICOM to NIfTI conversion. J. Neurosci. Methods 264, 47–56.

Mannion, DJ., 2015. Sensitivity to the visual field origin of natural image patches in human low-level visual cortex. PeerJ 3, e1038.

Nastase, SA, Connolly, AC, Oosterhof, NN, Halchenko, YO, Guntupalli, JS, Visconti di Oleggio Castello, M, Gobbini, I, Haxby, JV, 2017. Attention selectively reshapes the geometry of distributed semantic representation. Cereb. Cortex 27, 4277–4291.

Oliva, A, Torralba, A., 2006. Building the gist of a scene: the role of global image features in recognition. Prog. Brain Res. 155, 23–36.

Oliva, A, Torralba, A., 2007. The role of context in object recognition. Trends Cogn. Sci. 11, 520–527.

Oosterhof, NN, Connolly, AC, Haxby, JV., 2016. CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. Front. Neuroinform. 10, 20.

Peelen, MV, Fei-Fei, L, Kastner, S., 2009. Neural mechanisms of rapid natural scene categorization in human visual cortex. Nature 460, 94–97.

Peelen, MV, Kastner, S., 2011. A neural basis for real-world visual search in human occipitotemporal cortex. Proc. Natl. Acad. Sci. USA 108, 12125–12130.

Potter, MC., 1975. Meaning in visual search. Science 187, 965–966.

Potter, MC., 2012. Recognition and memory for briefly presented scenes. Front. Psychol. 3, 32.

Seidl, KN, Peelen, MV, Kastner, S., 2012. Neural evidence for distracter suppression during visual search in real-world scenes. J. Neurosci. 32, 11812–11819.

Shahdloo, M, Celik, E, Cukur, T., 2020. Biased competition in semantic representation during natural visual search. Neuroimage 216, 116383.

Spelke, ES, Lee, SA., 2012. Core systems of geometry in animal minds. Philos. Trans. R. Soc. Lond. B. 367, 2784–2793.

Thorpe, S, Fize, D, Marlot, C., 1996. Speed of processing in the human visual system. Nature 381, 520–522.

Torralba, A, Oliva, A, Castelhano, MS, Henderson, JM., 2006. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychol. Rev. 113, 766–786.

Torralbo, A, Walther, DB, Chai, B, Caddigan, E, Fei-Fei, L, Beck, DM., 2013. Good exemplars of natural scene categories elicit clearer patterns than bad exemplars but not greater BOLD activity. PLoS One 8, e58594.

VanRullen, R, Thorpe, SJ., 2001. The time course of visual processing: from early perception to decision-making. J. Cogn. Neurosci. 13, 454–461.

Võ, MLH, Boettcher, SEP, Draschkow, D., 2019. Reading scenes: How scene grammar guides attention and aids perception in real-world environments. Curr. Opin Psychol. 29, 205–210.

Wang, L, Mruczek, RE, Arcaro, MJ, Kastner, S., 2015. Probabilistic maps of visual topography in human cortex. Cereb. Cortex 25, 3911–3931.

Wolfe, JM, Võ, ML-H, Evans, KK, Greene, MR., 2011. Visual search in scenes involves selective and nonselective pathways. Trends Cogn. Sci. 15, 77–84.

Chapter 5

Project IV: The influence of the bullseye versus standard fixation cross on eye movements and classifying natural images from EEG

The current chapter comprises the manuscript entitled entitled "The influence of the bullseye versus standard fixation cross on eye movements and classifying natural images from EEG" which has been submitted to *Scientific Reports* and is available on the BioRxiv preprint server. In this manuscript, we compare the effect of two different fixation crosses on eye movements and the classification of natural images from EEG. We showed that the classification of neuroscientific data is influenced to a small degree by systematic eye movements at the level of single images for the standard but not for the bullseye fixation cross.

Authors:

Greta Häberle, Aynur Pelin Çelikkol, Radoslaw M. Cichy

Contributions:

G. H. and R. M. C. designed research; G. H. and A. P. C. acquired data; G. H. and A. P. C. analyzed data; G. H., and R. M. C. interpreted results; G. H. prepared figures; G. H. drafted the manuscript; G. H., A. P. C., and R. M. C. edited and revised the manuscript.

Contributions to open and reproducible science:

To contribute to open and reproducible science, upon acceptance the paper will be published in an open-access journal. Data are publicly available on OSF: https://osf.io/4ekct/. All code can be found here: https://github.com/Neural-Dynamics-of-Visual-Cognition-FUB/FixEyeEEG. The paper is made available as a preprint to the scientific community. doi: 10.1101/2023.03.21.532944

Copyright note:

The influence of the bullseye versus standard fixation cross on eye movements and classifying natural images from EEG

Greta Häberle[1,2,3*], Aynur Pelin Çelikkol[4], Radoslaw M. Cichy[1,2,3,5]

[1]Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany
[2]Charité – Universitätsmedizin Berlin, Einstein Center for Neurosciences Berlin, Berlin, Germany
[3]Humboldt-Universität zu Berlin, Faculty of Philosophy, Berlin School of Mind and Brain, Berlin, Germany
[4]Germany Cognitive Sciences, University of Potsdam, Germany
[5]Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany

*Corresponding author: greta.haeberle@gmail.com

1 Abstract

Eye movements are a ubiquitous and natural behavior, but in many tightly controlled experimental visual paradigms, eye movements are undesirable. Their occurrence can pose challenges to the interpretation of behavioral and neuroscientific data, in particular for magneto- and electroencephalography (M/EEG), which is sensitive to signals created by eye muscle movement. Here we compared the effect of two different fixation symbols – the standard fixation cross and the bullseye fixation cross – in the context of a visual paradigm with centrally presented naturalistic object images. We investigated eye movements and EEG data recorded simultaneously using behavioral and multivariate analysis techniques. Our findings comparing the bullseye to the standard fixation cross are threefold. First, the bullseye fixation cross reduces the number of saccades and amplitude size of microsaccades. Second, the bullseye fixation cross subtly reduces classification accuracy in both eye tracking and EEG data for the classification of single object images, but not for the superlevel category animacy. Third, using representational similarity analysis, we found a systematic relationship between eye tracking and EEG data at the level of single images for the standard, but not for the bullseye fixation cross. In conclusion, we recommend the bullseye fixation cross in experimental paradigms with fixation when particularly tight control of fixation is beneficial.

2 Introduction

Eye movements are a diverse, ubiquitous, and integral part of visual behavior[1]. For example, we use saccades, i.e., large voluntary eye movements, to explore a scene, and microsaccades, i.e., small involuntary eye movements, to keep the retinal image from fading[2].

However, for human cognitive neuroscience experiments that aim to establish statistical dependencies between tightly controlled visual input and brain activity eye movements pose experimental challenges. Eye movements change the visual input to the brain and also influence the recordings of brain activity by magneto- and electroencephalography (M/EEG)[3,4] through the currents created by eye muscle movements. Eye movements can thus introduce noise, add confounds, or both simultaneously into the experimental settings as shown for analyses of event-related potentials[5,6], frequency-resolved responses[7], and multivariate activation patterns[8–10].

A straight-forward way to reduce the effect of eye movements is to avoid them in the first place and ask participants to fixate. However, participants do not follow such instructions perfectly - novice participants, in particular, do not control their eye movements accurately[11], and fixation behavior varies widely among participants[11–13].

The amount of residual eye movements depends on the type of visual symbol used as a fixation target. An influential study[14] systematically evaluated the effect of different fixation symbols on eye movements when presented on a uniform gray background. The combination of a bullseye and cross hair fixation cross was associated with the smallest number of eye movements[14]. Throughout this paper, we will refer to this combination as the bullseye fixation cross. If this observation generalized to situations in which the fixation symbol appeared on top of visual stimuli for which brain responses were recorded, it should reduce the effect of eye movements on the neural analysis.

To investigate we collected eye tracking and EEG data simultaneously while participants viewed naturalistic still images of everyday objects overlaid with one of two different fixation symbols: a classical fixation cross and the bullseye fixation cross. We used time-resolved multivariate analysis methods to establish statistical dependencies between the experimental stimuli and the EEG data.

As anticipated, we replicate the behavioral advantage of the bullseye fixation cross over a standard fixation cross for experimental setups involving naturalistic still images[14]. We further show that the bullseye fixation cross reduced the influence of eye movements on the analysis of EEG brain data by reducing stimulus-specific eye movements. We thus recommend using the bullseye fixation cross to avoid unwanted eye movements and their effects on brain measurements in studies involving static, naturalistic still images.

3 Materials and methods

3.1 Participants

30 healthy adults (20 female, 9 male, 1 diverse, mean age = 24.77, SD = 4.08) were recruited at Freie Universität Berlin. Participants provided informed consent and received either course credit or monetary reimbursement for their participation. All participants had normal vision and no history of neurological disorders. The experiment was approved by the Ethics committee of Freie Universität Berlin and was conducted in accordance with the Declaration of Helsinki.

3.2 Stimulus set and stimulus presentation

The main stimulus set consisted of 40 images of everyday objects on natural backgrounds (Fig. 1a). Twenty images depicted animate and twenty inanimate objects. The set is a subset of the stimulus set used in a previous study[15].

Stimuli subtending 5-degree visual angle were presented centrally on a gray background on a Samsung Screen (SyncMaster 2233) (Fig. 1b) and either overlayed with a standard or bullseye fixation cross subtending 0.6-degrees visual angle (Fig. 1c).

Stimulus presentation was controlled by custom-made scripts in MATLAB 2021a[16] and the Psychtoolbox extension[17–19].

3.3 Experimental design

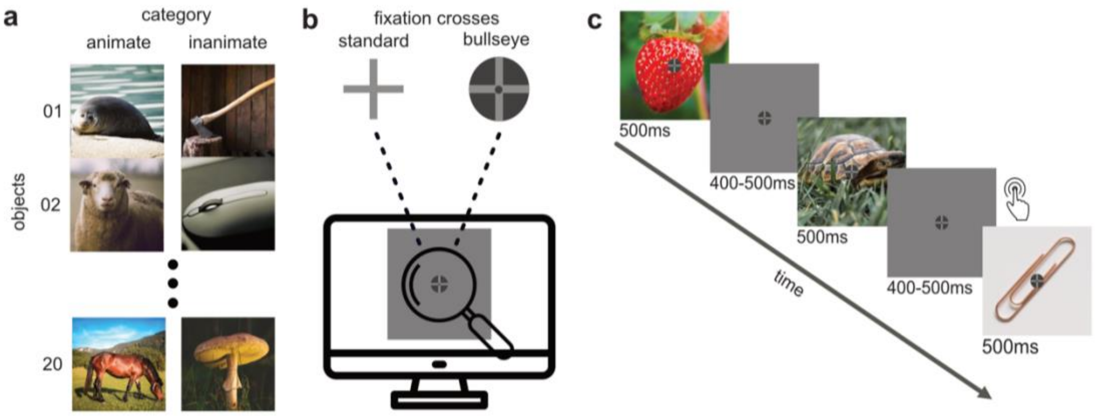

Fig.1: Stimulus set and paradigm. a) The stimulus set consisted of 40 images with natural objects, split equally into the animate and inanimate subcategories. **b)** We used two different fixation crosses overlaid onto the stimuli: the standard fixation cross and the bullseye fixation cross. **c)** Trials were blocked by fixation symbol. Participants were instructed to fixate the fixation symbol while objects appeared for 500 ms in random order with an inter-stimulus-interval of 400-500 ms and not to blink. Every 4-6 trials the image of a paperclip was shown for which participants were asked to press a button and to blink their eyes. Due to copyright, the stimuli in this figure are not the exact stimuli used in the experiment but resemble the original.

The combined EEG and eye tracking study consisted of one session, partitioned into 14 blocks. In each block only one fixation symbol was used, resulting in 7 blocks for the fixation cross and seven blocks for the bullseye fixation cross. The order of blocks was randomized.

During each block, stimuli were presented with object images in random order. On each trial, an image overlaid with a fixation symbol was presented for 500 ms, followed by an inter-trial interval randomly varying between 400 and 500 ms.

Participants were instructed to fixate on the fixation symbol in the middle of the screen and not to blink their eyes. Every 4th-6th trial an image of a paper clip was shown that was not part of the main stimulus set. Participants were instructed to press a button, blink, and refrain from blinking otherwise. Responses were collected with a standard keyboard attached to the presentation computer. Participants could take self-paced breaks after each block. On average participants detected 81,9% of catch trials.

In each block each image of the main stimulus set was repeated six times, resulting in 240 trials plus 60 additional paperclip trials. Over the course of the experimental session, this resulted in 3380 trials for the main stimulus set and 840 paperclip target trials. Paperclip trials were excluded from further analysis

3.4 Eye tracking recording and preprocessing

3.4.1 Recording

We monocularly recorded the right eye of each participant using the Eyelink 1000 Tower Mount (SR Research Ltd., Osgoode, Ontario, Canada) and the Eyelink Toolbox extensions[20] with a sampling rate of 1000 Hz. Participants were seated comfortably in front of the screen and instructed to rest their chins on the chinrest, 60cm away from the monitor. Before each block, we calibrated the eye tracker using a nine-point calibration. The data was recorded in gaze position coordinates.

3.4.2 General preprocessing

Before applying the saccade detection algorithm, we cleaned the eye tracking data from artifacts by excluding (i) all trials containing blinks[21], (ii) all data samples outside the screen range, (iii) all negative data samples, and (iv) additionally all data samples in a 100 ms period around excluded samples from steps one to three[22].

After artifact exclusion we converted the eye tracking data samples from screen coordinates in pixels to spherical angles in degree, using the following formula:

$$\beta_x = 2 * atan2(p_x * m, d)$$

where β_x is the azimuth angle in visual degree from the monitor center, p_x represents the horizontal position relative to the center of the monitor and is measured in pixels, m is the conversion factor to convert pixels to millimeters and d represents the distance from the observer's eyes to the monitor, and is measured in millimeters[21]. An equivalent procedure was applied to the y-coordinate. This yielded a data frame with x and y coordinates for the position of the eye in degree visual angle (dva) for each time point for all included trials.

3.4.3 Saccade and microsaccade detection

We used a velocity-based algorithm[23] to detect saccades and microsaccades. Numerous studies have shown that the magnitude of microsaccades mostly falls below one degree[23–26]. Therefore, we defined all saccades with amplitudes smaller than one degree as microsaccades and all saccades with amplitudes larger or equal to one degree as saccades.

To detect saccades and microsaccades, the position vector (x and y positions of the eye in dva) was transformed into a two-dimensional velocity space. The velocity of the eye was required to exceed eight standard deviations of the eye’s velocity during the trial for at least 8 ms to be detected as an event. These values are higher than the typically proposed values[23] to minimize noise emerging from our monocular recording setup (R. Engbert, personal communication).

To further boost the signal-to-noise ratio, we implemented a ratio criterion where data points were only considered if the ratio of path length to amplitude exceeded 0.5[14]. Together the procedures yielded a data frame with all detected saccades and microsaccades and their amplitudes for all trials.

3.4.4 Preprocessing for multivariate pattern analysis

We took specific steps to preprocess the eye tracking data for multivariate pattern analysis (MVPA). We epoched the data from -200 to +1000 ms around stimulus onset, downsampled the data to 200 Hz, and baseline corrected each epoch by subtracting the mean of the 200 ms prestimulus interval from the entire epoch. Each epoch thus contained a time course of x and y positions of the eye.

We only kept trials for MVPA that were neither excluded during the eye tracking data preprocessing, nor the EEG data preprocessing (see below). This amounted to, on average, 40 trials per object image and 802 trials per category (animate/inanimate) per participant.

3.5 EEG recording and preprocessing

3.5.1 Recording

We recorded EEG data using the ActiCap64 electrodes system and Brainvision actiChamp amplifier. 64 Electrodes were placed according to the 10-10 system[27] with an additional ground and reference electrode placed on the scalp. We recorded the data with Brainvision recorder software, using a 1000 Hz sampling rate and online filtering between 0.03 Hz and 100 Hz[28]. We kept all impedances below 10kΩ.

3.5.2 Preprocessing

We preprocessed the EEG data offline using Fieldtrip[29] in MATLAB 2021a[16]. We epoched the data between -200 and +1000 ms relative to stimulus onset, notch filtered it at 50 Hz, downsampled it to 200 Hz, and performed baseline correction by subtracting the mean of the 200 ms prestimulus interval from the entire epoch. Each epoch thus contained a 64-dimensional EEG time course. Subsequently, we excluded all trials which were excluded during the eye tracking preprocessing (see above for details) from the EEG data and additionally manually removed all channels and trials containing excessive noise. We then interpolated missing channels by using the average of all surrounding channel neighbors. This procedure resulted in the equivalent number of matched EEG and eye tracking trials, i.e., on average, 40 trials per object image and 802 trials per category (animate/inanimate) per participant.

3.6 Generalized linear mixed model

We analyzed saccade and microsaccade numbers and amplitudes using Generalized Linear Mixed Models (GLMMs) with the lmer4 package[30] in R version 1.3.1093[31]. Saccades and microsaccades were analyzed separately. A general description of the models is given by the following formula:

$$g(model_{baseline}) = \beta_0 + u_{0,j} + e_0$$
$$g(model_{fixation}) = \beta_0 + \beta_1 * cross + u_{0,j} + u_{1,j} * cross + e_0$$

with g() defining the link function, β_0 the intercepts, $\beta_1 * cross$ the fixed effect fixation cross, $u_{0,j}$ the subjects' random intercept, $u_{1,j} * cross$ the random slopes for the factor fixation cross and e_0 the error term. We included random intercepts to capture variances in the individual subject means and random slopes to allow for participant-specific effect magnitudes[32].

We modeled the effect of the factor fixation cross on amplitude size and the number of saccades and microsaccades separately. In each case, we fitted two models (baseline and full model), with a linear link function for the amplitude and a Poisson link function for the number of saccades and microsaccades. The full model included the factor fixation cross, whereas the baseline model did not. We used sum coding (-0.5 and 0.5) for all contrasts to be able to interpret differences in condition means.

3.7 Multivariate pattern analysis

To characterize the time course with which object and category representations emerge, we conducted MVPA using linear support vector machine (SVM)[33] as implemented in libsvm[34].

We conducted 8 separate analyses: 2 (modality: EEG, eye tracking) * 2 (classification type: image identity, object animacy) * 2 (fixation cross: standards, bullseye), and each participant was analyzed separately.

Each analysis had three main steps, with each step performed independently for each time point in the epoch. First, we averaged over individual trials to create pseudo-trials, thereby increasing the signal-to-noise ratio[35]. Second, we trained an SVM on all but one pseudo-trial to differentiate either image identity or image category in a pairwise fashion. Third, we tested the prediction accuracy of the SVM on left-out data.

We conducted each analysis in two ways: time-resolved and time-generalized MVPA. For time-resolved analysis[36,37], the SVM was trained and tested on data from the same time points only, yielding a single time course with which object information emerges as a result. For time-generalized analysis[38], the SVM was trained and tested for all possible combinations of time points, yielding a two-dimensional result array, indicating how stable EEG activation patterns are. We describe the details of each analysis type below.

3.7.1 Time-resolved MVPA

We used time-resolved MVPA to determine the time course with which object identity and animacy representations emerged. We conducted equivalent analyses based on EEG and eye tracking data. For each time point, we extracted trial-specific EEG channel activations (64 channels for EEG classification) or trial-specific eye positions (x and y coordinates in dva for eye tracking classification).

We averaged trials aggregated for each object (for object classification) or by animacy (for classification animate vs. inanimate) into six pseudo trials[35]. We used multivariate noise normalization[39] to whiten the data and further improve the signal-to-noise ratio[40]. For this, we multiplied the data by the inverse of the square root of the covariance matrix of electrode activations from the entire epoch. We trained the SVM classifier in a pairwise fashion on data of all but one pseudo trial and tested the SVM classifier on the left-out trial. We repeated this procedure 100 times, each time with a different random assignment of trials to pseudo-trials. To ensure that the SVM classifier was not biased by an excess of data for one or the other condition, we included the same number of trials in the creation of pseudo-trials for each condition. We averaged the resulting decoding accuracies over the 100 repetitions.

Across our analysis space, this resulted in one decoding accuracy number per participant, modality, classification type, and fixation cross and time point.

3.7.2 Time generalization MVPA

We used time generalization MVPA[38] to determine the temporal stability of object and animacy representations. The procedure was equivalent to the time-resolved MVPA described above, except that we tested the SVM classifier trained on any one specific time point iteratively on all time points. Further, for object classification, the EEG and eye tracking data were downsampled to 50 Hz for time and memory efficiency.

Across our analysis space, this resulted in a 2D matrix of classification accuracies, indexed in rows and columns by the time points of the epoch, per participant, modality, classification type, and fixation cross.

3.8 Representational similarity analysis

We performed representational similarity analysis (RSA)[41] to compare data quantitatively across modalities (EEG, eye tracking). This was a two-step process. First, for each participant, we constructed representational dissimilarity matrices (RDMs) for each modality and fixation cross type separately. Second, we related the RDMs directly by calculating their similarity. We describe each of the two steps in detail below.

3.8.1 Construction of RDMs

We created RDMs using the preprocessed EEG and eye tracking data for each fixation cross type separately. For each time point, we calculated dissimilarity for all pairs of conditions (i.e., the 40 different object images) by using 1-Pearson correlations[39] (Fig. 5a). This resulted for each time point in a 40×40 RDM, indexed in rows and columns by the conditions compared. We equalized the number of trials across conditions by randomly subsampling trials with the minimum number of trials across conditions. We repeated the analysis 100 times, calculating an RDM each time, and averaged the RDMs across the iterations. We used the vectorized upper triangular matrix of the symmetric RDM (without the diagonal) for further analysis.

3.8.2 RDM comparison

We determined how visual representations measured with EEG and eye movements measured with eye tracking relate. For this, we correlated the EEG RDMs and eye tracking RDMs using Spearman's r for each participant, time point, and fixation cross separately. Averaging over participants, we obtained a time course of similarity between the EEG and eye tracking data for each fixation cross type.

3.8.3 Noise ceiling

The RSA results are bound by measurement noise, as the calculated correlation values are affected by both the variance in the data as well as by noise in the data. To determine how much variance could, in principle, be explained by the model if we knew the true data-generating-model[42,43] we determined the noise ceilings for the EEG and eye tracking data. We estimated an upper and lower bound for the noise ceiling[43] in a time-resolved fashion. For the lower bound, we singled out a participant RDM and compared it with the average of the RDMs for all participants except the one singled out, iterated this for each participant, and averaged. For the upper bound, we singled out a participant's RDM and compared it with the average of the RDMs for all participants, including the one singled out, iterated for each participant, and averaged.

3.9 Statistical testing

MVPA-related statistical analyses were performed using MATLAB. We evaluated the statistic of interest (classification accuracy, correlation coefficient) using nonparametric tests. The null hypothesis stated that the statistic of interest was equal to chance level (i.e., 50% classification accuracy or a Spearman's r of 0). To estimate a null distribution we used a sign permutation test, multiplying participant-specific data randomly with +1 or -1 and recomputing the statistic of interest 10,000 times. To obtain p-values, we calculated the rank of the test statistic with respect to the null distribution.

We controlled the family-wise error rate across time points using cluster size inference[44] with a $p < 0.05$ cluster-definition threshold and $p < 0.05$ cluster threshold. For all tests, both thresholds were right-sided, with the exception of EEG classification difference curves where the thresholds were two-sided.

We used bootstrapping to compute 95% confidence intervals for peak latencies. For this, we sampled the participant pool 10,000 times with replacement and calculated peak latencies for each sample. This created an empirical distribution of peaks on which we determined 95% confidence intervals.

For the GLMMs, we performed statistics using R version 1.3.1093[31]. The derivation of p-values for the different predictors in GLMMs is debated with two prominent candidate approaches- the Wald test and the likelihood ratio test (LRT). For Wald tests, the z distribution is used to obtain p-values from Wald t-values. This is generally appealing as the t distribution approximates the z distribution for increasing numbers of degrees of freedom (identical for infinite degrees of freedom). LRTs are classically used to test whether a certain predictor should be part of a model or whether the predictor can be excluded. LRTs determine which and whether one of two models fits the data better. We report the outcome of both tests that here, in all cases, concur.

4 Results

We investigated the effect of fixation cross type on naturalistic still images in a three-step procedure. First, we assessed eye movement behavior alone, analyzing the number and amplitudes of saccades and microsaccades using GLMMs. Second, we, in parallel, assessed EEG and eye tracking data using MVPA, delineating how fixation cross type influences classification of visual information presented alongside the fixation cross. Third, to directly and quantitatively evaluate systematic relationships between eye tracking data to EEG data, we used RSA.

4.1 The bullseye fixation cross reduces eye movements

We analyzed the influence of fixation cross type on eye movements. We used a velocity-based algorithm[23] to detect saccades and microsaccades, with microsaccades being smaller than one degree of visual angle.

Visual inspection of the descriptive statistics for the numbers and amplitudes of saccades and microsaccades (Fig 2a, b) revealed similar distributions, indicating that any potential effects of fixation crosses are subtle. To statistically evaluate the effect of fixation cross type on saccade and microsaccade amplitudes and numbers, we fitted pairs of GLMMs, once with the predictor fixation cross type included and once excluded. We evaluated statistical significance with both the likelihood ratio test (LRT) and the Wald test, yielding equivalent results (see Table 1 for amplitudes, Table 2 for numbers).

Using the LRT, we found that the standard fixation cross was associated with a 3.4% increase in saccades ($\chi^2(1) = 10.35,\ p = 0.001297$) and a 0.023 dva increase in amplitudes ($\chi^2(1) = 7.96,\ p = 0.004785$). There was no evidence that fixation cross type affected saccade amplitudes and microsaccade numbers.

Together this shows that the bullseye fixation cross reduces eye movements compared to the standard fixation cross when participants are asked to fixate on naturalistic still images. Our results contextualize the effect size as subtle with respect to the overall observed distribution of eye movements and generalize previous findings from fixation on uniform backgrounds[14].

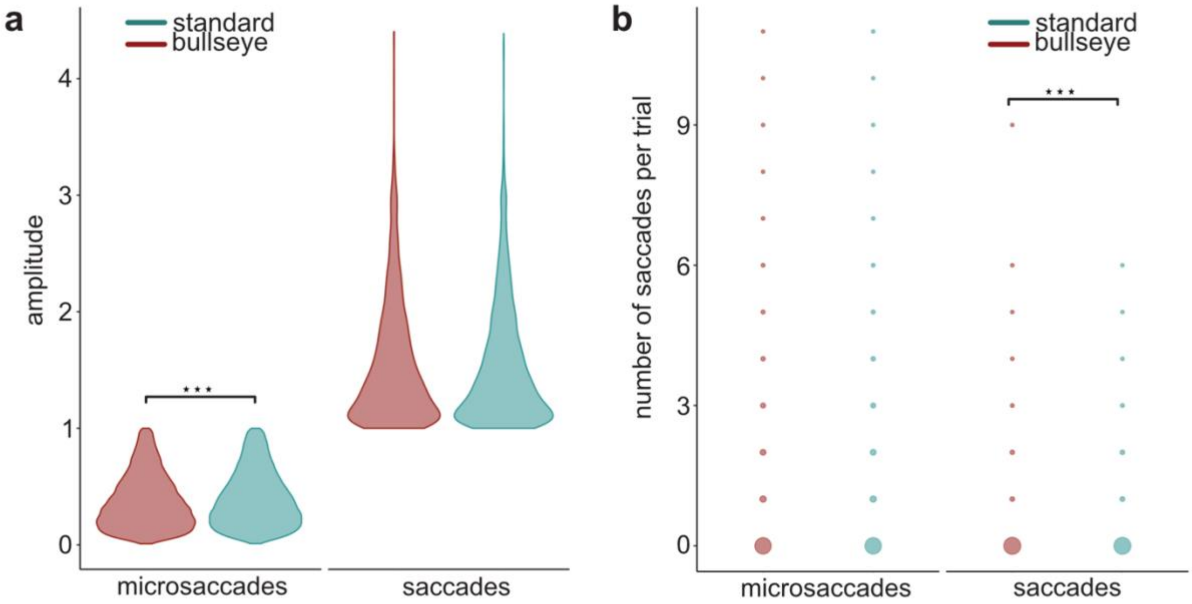

Fig. 2: Descriptive statistics for the amplitude and number of microsaccades and saccades per trial. **a)** Violin plots of saccade and microsaccade amplitudes. **b)** Distribution of number of saccades and microsaccades per trial. Stars indicate significant fixed effects with p < 0.01.

	saccades		microsaccades	
	(1)	(2)	(3)	(4)
intercept	1.553***	1.551***	0.369***	0.378***
	(1.485, 1.620)	(1.483, 1.619)	(0.342, 0.395)	(0.351, 0.405)
fixation cross		-0.006		0.023***
		(-0.040, 0.027)		(0.008, 0.037)
log likelihood	-8,805.197	-8,805.133	7,123.959	7,127.938
Akaide inf. crit.	17,620.400	17,622.270	-14,237.920	-14,243.880
Bayesian inf. crit.	17,657.220	17,666.460	-14,191.040	-14,191.040

Note: *p<0.1; **p<0.05; ***p<0.01

Table 1: Summary of model fits, intercepts, and fixed effect estimates for saccade and microsaccade amplitude models. Estimates are in real values and represent the difference between conditions. Stars represent significance, calculated with Wald's t-as-z approach. Confidence intervals are stated in parentheses below.

	saccades		microsaccades	
	(1)	(2)	(3)	(4)
intercept	0.011***	0.012***	0.158***	0.149***
	(0.008, 0.015)	(0.008, 0.017)	(0.139, 0.181)	(0.130, 0.170)
fixation cross		1.256***		1.034
		(1.109, 1.421)		(0.988, 1.082)
log likelihood	-51,131.250	-51,126.070	-266,203.600	-266,202.800
Akaide inf. crit.	102,270.500	102,262.100	532,415.200	532,415.600
Bayesian inf. crit.	102,315.000	102,317.800	532,459.700	532,471.200

Note: *p<0.1; **p<0.05; ***p<0.01

Table 2: Summary of model fits, intercepts, and fixed effect estimates for saccade and microsaccade number models. Estimates are in rate ratios and represent the difference between conditions. Stars represent significance, calculated with Wald's t-as-z approach. Confidence intervals are stated in parentheses below.

4.2 The impact of fixation cross type on MVPA of visual information from eye tracking and EEG data

We investigated how fixation cross type influences the classification of visual information available in the naturalistic object images on which the fixation cross symbols were overlaid. For this, we analyzed eye tracking and EEG data in parallel to identify common patterns that might suggest an influence of eye movements on EEG data. We assessed visual information at two levels commonly queried in current cognitive neuroscience experiments: at the level of single object images and at the higher categorization level by classifying object animacy (i.e., animate vs. inanimate). We present the classification results for the eye tracking data first.

4.2.1 Classifying visual information from eye tracking data

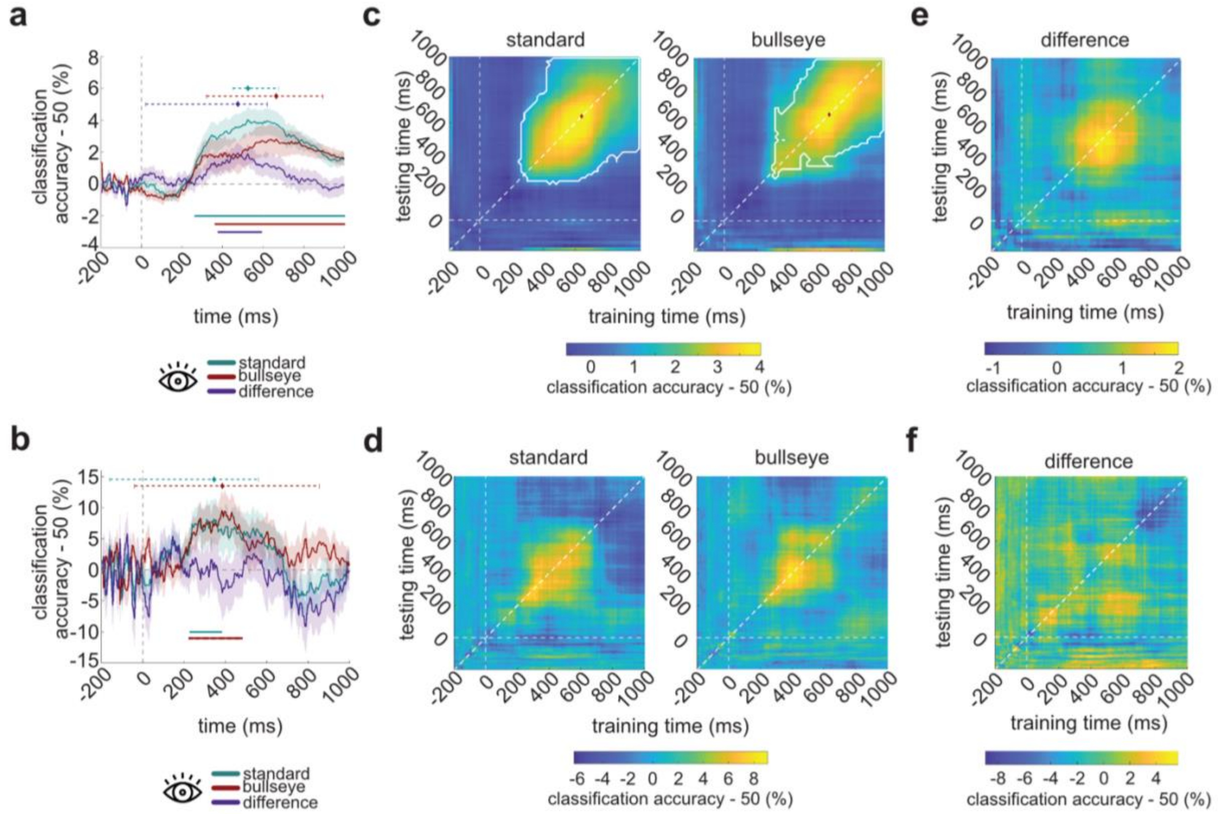

Fig. 3: Eye tracking MVPA results. a) Time-resolved object and **b)** animacy classification. The vertical dotted line shows stimulus onset, error bars indicate SEMs across participants. Peak latencies and 95% confidence intervals are indicated above the curves as dotted lines color coded as the result curves. The lines below the curves denote significant time points. **c)** Object and **d)** animacy time-generalization analyses, and **e, f**) differences between fixation crosses for each. Statistically significant time points are outlined in white. The diamond shapes indicate peak latencies. The dashed lines indicate stimulus onset and the diagonal. Results are corrected for multiple comparisons by cluster correction (cluster definition threshold $p < 0.05$, cluster threshold $p < 0.05$). Detailed information about cluster extents can be found in Table 3.

We determined whether there is a systematic relationship between eye movements and the visual material presented. For this, we applied time-resolved multivariate pattern analysis to the eye tracking data classifying both object/image identity (Fig. 3a) and object animacy (Fig. 3b). We conducted the classification analyses separately for the two fixation cross types and compared the outcome. We assessed statistical significance with cluster permutation tests (cluster-definition threshold $p < 0.05$; cluster threshold $p < 0.05$) and reported peak latencies with 95% confidence intervals in square brackets.

For object identity classification, we found significant information for both fixation crosses (Fig. 3a, turquoise and red curves). The result curves increased gradually from 200 ms after stimulus onset, followed by a prolonged plateau and slow decay over the duration of the trial with peaks at 525 ms [450 675] and 665 ms [320 895]. The difference curve between those results (Fig. 3a, purple curve) had a similar shape, showing higher accuracy for the standard than the bullseye fixation cross with a peak at 475 ms [20 620].

For animacy classification, we also found significant information for both fixation crosses (Fig. 3b, turquoise and red curve) with peaks at 345 ms [-160 560] and 385 ms [-40 855], respectively. However, the difference curve fluctuated around chance level and was not significantly different from it.

Together, these results indicate a systematic relationship between eye movements and the visual material presented at the level of single images and image identity, but not at the level of more abstract, categorical object divisions.

4.2.2. Time-generalized visual information classification on eye tracking data

The finding of prolonged object information in the time-resolved eye tracking classification analysis poses the question about the temporal stability of the data patterns underlying this result: is the prolonged effect due to a data pattern stable over time, or due to a rapidly evolving data pattern? To assess temporal stability, we conducted time-generalized MVPA[38], classifying object information across time points in the epoch.

Visual inspection of the results suggested a stable data pattern as classification generalized across data points for both object identity (Fig. 3c) and object animacy (Fig. 3d) and for both fixation crosses. However, statistically, results were significant only for object identity.

As expected from the time-resolved analysis, visually we observed higher classification accuracy for the standard than the bullseye fixation cross (Fig. 3e), with strong off-diagonal classification results again indicating temporal stability of the underlying data patterns. However, neither the difference curve for object identity nor for object animacy (Fig. 3e, f) was statistically significant, precluding interpretation.

Together this shows that object identity classification from eye tracking data depends on a temporally stable rather than strongly dynamic data patterns.

4.2.3 Classifying visual information from EEG data

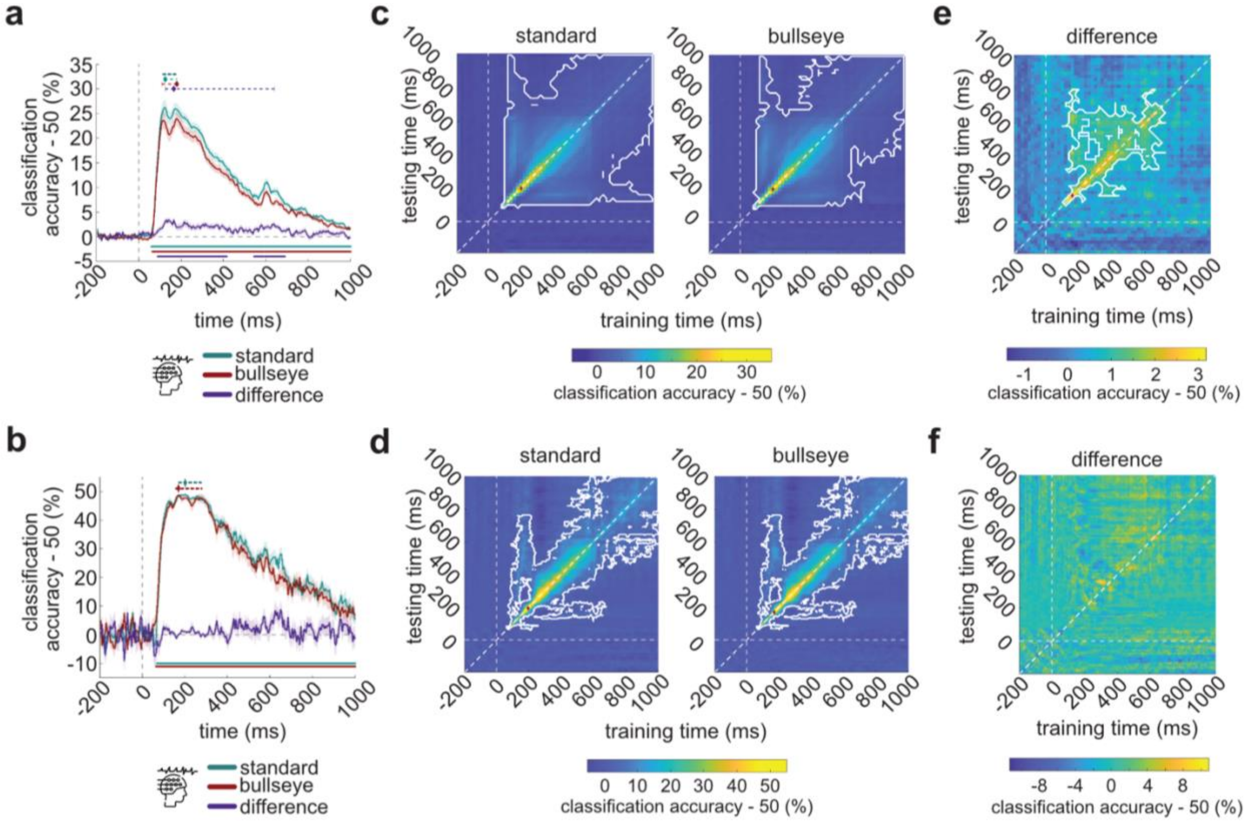

Fig. 4: EEG MVPA results. a) Time-resolved object and **b)** animacy classification. The vertical dotted line shows stimulus onset, error bars indicate SEMs across participants. Peak latencies and 95% confidence intervals are indicated above the curves as dotted lines color coded as the result curves. The lines below the curves denote significant time points. **c)** Object and **d)** animacy time-generalization analyses, and **e, f**) differences between fixation crosses for each. Statistically significant time points are outlined in white. The diamond shapes indicate peak latencies. The dashed lines indicate stimulus onset and the diagonal. Results are corrected for multiple comparisons by cluster correction (cluster definition threshold $p < 0.05$, cluster threshold $p < 0.05$). Detailed information about cluster extents can be found in Table 3.

Using an equivalent analysis strategy as for the eye tracking data, we determined whether there is a systematic relationship between EEG data and the visual material presented and to which degree this is influenced by fixation cross type

For object identity classification, we found significant information for both fixation crosses (Fig. 4a, turquoise and red curves). The result curves increased rapidly from 60 ms after stimulus onset, followed by peaks at 125 ms [120 180] (standard fixation cross), at 180 ms [110 185] (bullseye fixation cross), and a gradual decline. This result mirrors the commonly observed MVPA results pattern for classifying visual information at the image level from M/EEG[45–49]. The difference curve between results of classification for the different fixation crosses (Fig. 4a, purple curve) had a similar, though highly down-scaled shape with a peak at 165 ms [125 640] with higher accuracy for the standard fixation cross.

For animacy classification, we also found significant information for both fixation crosses (Fig. 4b, turquoise and red curve) with peaks at 200 ms [170 280] (standard fixation cross) and 170 ms [160 280] (bullseye fixation cross), consistent with previous results[45–47,50]. However, the difference curve (Fig. 4b, purple curve) fluctuated around chance level and was not significantly different from it.

In sum, these results indicate an influence of fixation symbols on classification of visual information from EEG at the level of single images, but not more abstract, categorical object divisions. This result pattern parallels that from eye tracking classification, suggesting that observations of stronger eye movements for the standard fixation cross compared to the bullseye fixation cross are related to the increased classification accuracy in the EEG data analysis.

		fixation cross type		
method	**classification**	standard	bullseye	difference
eye tracking	object identity	265-1000 ms	365-1000 ms	380-590 ms
	animacy	230-1000 ms	225-1000 ms	no significant clusters
EEG	object identity	60-1000 ms	65-1000 ms	90-415 ms, 545-715 ms
	animacy	70-1000 ms	65-1000 ms	no significant clusters

Table 3: Cluster extent of time-resolved eye tracking and EEG MVPA. Overview of earliest and latest time points of significant clusters for the different fixation cross types and classification schemes (cluster definition threshold p < 0.05, cluster threshold p < 0.05).

4.2.4 Time-generalized visual information classification from EEG data

We assessed temporal stability of EEG activation patterns underlying time-resolved EEG classification with the same methodology as for eye tracking data patterns using time generalization analysis.

For both object identity and object animacy, we observed evidence for both rapidly changing as well as more stable activation pattern dynamics for both fixation cross types (Fig. 4c, d), consistent with previous studies investigating perception with time-generalized analysis[8,45,46,51]. The rapidly changing dynamics were indicated by the relatively high classification accuracy along the diagonals and the stable aspects by significant effects far beyond the diagonal, in particular, a broadening of effects after 200 ms.

The difference curve for object identity classification (Fig. 4e) comparing results based on fixation cross symbols revealed both rapidly changing and stable activation pattern dynamics underlying the higher classification accuracy for the standard compared to the bullseye fixation cross. The difference curve for animacy classification (Fig. 4f), as predicted from the time-resolved analysis (Fig. 4b), was not significant.

Together these results show that fixation cross type influences EEG classification in both rapidly changing as well as stable neural dynamics, with higher classification for the standard fixation cross.

4.3 The bullseye fixation cross reduces systematic eye movement-related confounds in EEG classification

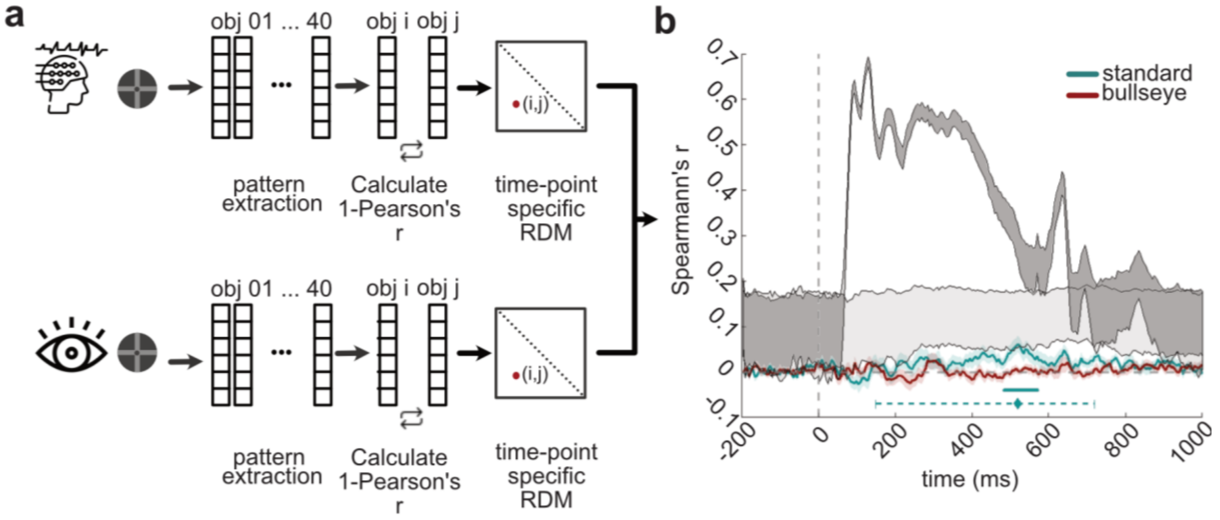

Fig. 5: Relating EEG and eye tracking data using RSA. a) Analysis pipeline for time-resolved RSA. In the first step, we built RDMs separately for EEG and eye tracking data and for both fixation crosses. For this, we calculated the pairwise dissimilarity (1 – Pearson's r) between all 40 individual object conditions, resulting in 40*40 RDMs for each time point. In the second step, we compared RDMs using Spearman's r. **b)** RSA results. The vertical dotted line shows stimulus onset, error bars indicate SEMs across participants. The lines below the curves denote significant time points. Peak latencies and the corresponding CIs are indicated by the dotted line below the curve. The dark and light grey shaded areas delineate the upper and lower bounds on the noise ceilings for the EEG and the eye tracking data, respectively.

Our analyses revealed - both for eye movement and for EEG data - a decrease in classification accuracy for object identity classification for the bullseye fixation cross compared to the standard fixation cross. This suggests the hypothesis that the decrease in EEG classification accuracy is systematically related to a decrease in eye movement-related effects, such as confounding eye muscle activity or differences in neural processing as a consequence of eye movements.

A prediction of this hypothesis would be that images eliciting similar activation patterns in the EEG data should also elicit similar activation patterns in the eye tracking data. To test this prediction directly and quantitatively, we used representation similarity analysis[41,52,53]. In a time-resolved fashion, and for each fixation symbol separately, we aggregated all pairwise dissimilarities between object identities in representational dissimilarity matrices that abstract away from the disparate measurement spaces of eye tracking and EEG into a common similarity space (Fig 5a). We then related the eye tracking and EEG data by calculating the similarity between their RDMs.

This analysis revealed a systematic relationship between EEG and eye tracking data for the bullseye fixation cross, but not for the standard fixation cross (Fig. 5b). This shows that for the standard fixation cross condition eye movements confound EEG data to a small but systematic degree. In contrast, the difference curve between the RSA results for the two fixation symbols was not significant.

5 Discussion

5.1 Summary

We compared the effect of two fixation cross types - the standard and the bullseye fixation cross - on eye movements and EEG data in the context of a paradigm presenting naturalistic object images that were either animate or inanimate. We made three key observations. First, we showed that the bullseye fixation cross reduced eye movements compared to the standard fixation cross. Second, we showed that the bullseye fixation cross reduced classification of object identity, but not animacy from both eye tracking and EEG data. Third, we established a systematic relationship between classification results from eye tracking and from EEG data for the standard fixation cross, but not the bullseye fixation cross.

5.2 The bullseye fixation cross reduces eye movements

Previous research established that the bullseye fixation cross reduces eye movements compared to the standard fixation cross (and other fixation crosses) when presented in isolation on a uniform gray background[14]. Here we extend this research by reproducing the advantage of the bullseye fixation cross when superimposed on naturalistic object images. This result generalizes the previous findings[14] to a commonly used basic visual paradigm in cognitive neuroscience. This is further not a trivial result, as the naturalistic images were neither controlled for color, luminance, nor salience, while saccades are known to be modulated by perceptual attention[54], background[55], and salience[56].

While significant, the reduction in eye movements by fixation cross choice was subtle in effect size, and overall the distribution of saccades and microsaccades was similar irrespective of the fixation cross chosen. This highlights the need for additional measures to reduce the occurrence of eye movements in future studies.

5.3 The impact of fixation cross type on MVPA of visual information from eye tracking and EEG data

Our results are broadly consistent with previous studies that reported successful classification of diverse conditions of interest from eye movement alone, such as object category[8] and grating orientation[13] during perception, as well as stimulus information during the delay period of a working memory match to sample task[9] or during the perception and attention phase of a working memory match to sample task[10]. We go beyond those studies by assessing the differential effect of fixation cross type on multivariate analysis. We observed that in multivariate pattern analysis on both eye tracking and EEG data, accuracy was reduced for the bullseye fixation cross compared to the standard fixation cross. However, this effect was limited to object identity, and did not extend to object animacy. One reason for this might be that in our data set single object images were associated systematically with different eye movements that supported classification, whereas object image sets at the supra-category level of animacy were not and eye movements averaged out into similar distributions for animate vs. inanimate objects. However, this is likely a function of the stimulus set used here. In an experimental setup with two categories that have low intra-group image variability at the pixel level (e.g., highly controlled images of front-view faces and houses), we would expect systematic eye movement effects at the level of category.

The parallel reduction of classification accuracy for the bullseye fixation cross in the classification analyses based on eye tracking and on EEG data suggests a relationship between those observations. We substantiated this hypothesis by establishing that for the standard fixation cross, but not for the bullseye fixation cross, EEG and eye tracking were systematically related at the object image level. This is consistent with a previous study[8] using a bullseye fixation cross that also observed both classification of visual stimulus information from eye tracking and MEG data, but did not find a systematic relationship between them. We thus recommend the use of the bullseye fixation cross as a measure to reduce the confounding effect of eye movements in M/EEG, in particular but not limited to multivariate classification studies.

To put the effect into context, the identified confounding effect of eye movements on classification analysis was significant, but the effect size comparing the two fixation crosses (3.41%) was small compared to the overall classification results (26.22% and 23.98%, Fig. 4a). Together with the observation that the systematic relationship between EEG and eye tracking data was small or absent depending on the fixation cross, (Fig. 5b), this speaks for a limited confounding impact of eye movements systematically related to experimental conditions on EEG classification analyses under well-controlled laboratory conditions involving fixation control.

One open question is how exactly eye movements systematically related to object images influenced the EEG data and thus the classification results. One possibility is that the EEG acquisition picked up electrical activity created by eye movements. Another possibility is that the eye movements led to changes in neural processing picked up by the EEG, e.g., elicited by changes in the visual input. Future studies relating eye movements to predicted changes in visual input combined with electromyography to isolate the effect of eye muscle activity are needed.

5.4 Limitations

Our results and conclusions are subject to several limitations. First, we measured data only from one eye, limiting the ability to distinguish measurement noise from microsaccades[14,57]. Future studies recording both eyes are needed for further scrutiny. Second, it is unclear how well our results generalize from the particular experimental setup of centrally presented naturalistic images. Moving images, images presented in the periphery, or images presented alongside with other sensory cues[58] might affect other eye movement patterns that might be differently affected by the choice of a fixation symbol. Further studies assessing the effects of eye movements and their interaction with fixation crosses on M/EEG data are needed. Our data guide this research with the a priori hypothesis that the bullseye fixation cross will reduce eye movements compared to the standard fixation cross.

5.5 Conclusion

We conclude that the bullseye fixation cross reduced eye movements and their effects on M/EEG data for a visual setup with centrally presented naturalistic images. While the effect of eye movements on M/EEG data observed here was limited, and the reduction of eye movements and their associated effects on M/EEG is subtle, we recommend its use when tight control of eye movements is key for the experimental design and tested hypothesis.

6 Data availability

The code used for this project can be found at: https://github.com/Neural-Dynamics-of-Visual-Cognition-FUB/FixEyeEEG.
The data set can be found at https://osf.io/4ekct/.

7 References

1. Schütz, A. C., Braun, D. I. & Gegenfurtner, K. R. Eye movements and perception: A selective review. *J. Vis.* **11**, 9 (2011).
2. Rolfs, M. Microsaccades: Small steps on a long way. *Vision Res.* **49**, 2415–2441 (2009).
3. Plöchl, M., Ossandón, J. P. & König, P. Combining EEG and eye tracking: identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Front. Hum. Neurosci.* **6**, 278 (2012).
4. Thickbroom, G. W. & Mastaglia, F. L. Presaccadic spike potential. Relation to eye movement direction. *Electroencephalogr. Clin. Neurophysiol.* **64**, 211–214 (1986).
5. Dimigen, O., Valsecchi, M., Sommer, W. & Kliegl, R. Human Microsaccade-Related Visual Brain Responses. *J. Neurosci.* **29**, 12321–12331 (2009).
6. Dimigen, O. & Ehinger, B. V. Regression-based analysis of combined EEG and eye-tracking data: Theory and applications. *J. Vis.* **21**, 3 (2021).
7. Yuval-Greenberg, S., Tomer, O., Keren, A. S., Nelken, I. & Deouell, L. Y. Transient Induced Gamma-Band Response in EEG as a Manifestation of Miniature Saccades. *Neuron* **58**, 429–441 (2008).
8. Dijkstra, N., Mostert, P., de Lange, F. P., Bosch, S. & van Gerven, M. A. Differential temporal dynamics during visual imagery and perception. *eLife* **7**, e33904 (2018).
9. Mostert, P. *et al.* Eye Movement-Related Confounds in Neural Decoding of Visual Working Memory Representations. *eNeuro* **5**, (2018).
10. Quax, S. C., Dijkstra, N., van Staveren, M. J., Bosch, S. E. & van Gerven, M. A. J. Eye movements explain decodability during perception and cued attention in MEG. *NeuroImage* **195**, 444–453 (2019).
11. Guzman-Martinez, E., Leung, P., Franconeri, S., Grabowecky, M. & Suzuki, S. Rapid eye-fixation training without eye tracking. *Psychon. Bull. Rev.* **16**, 491–496 (2009).
12. Bargary, G. *et al.* Individual differences in human eye movements: An oculomotor signature? *Vision Res.* **141**, 157–169 (2017).
13. Thielen, J., Bosch, S. E., van Leeuwen, T. M., van Gerven, M. A. J. & van Lier, R. Evidence for confounding eye movements under attempted fixation and active viewing in cognitive neuroscience. *Sci. Rep.* **9**, 17456 (2019).
14. Thaler, L., Schütz, A. C., Goodale, M. A. & Gegenfurtner, K. R. What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Res.* **76**, 31–42 (2013).
15. Cichy, R. M., Pantazis, D. & Oliva, A. Similarity-Based Fusion of MEG and fMRI Reveals Spatio-Temporal Dynamics in Human Cortex During Visual Object Recognition. *Cereb. Cortex* **26**, 3563–3579 (2016).
16. MATLAB *version 9.10.0.1602886 (R2021a)*. (The Mathworks, Inc., 2021).
17. Brainard, D. H. The Psychophysics Toolbox. *Spat. Vis.* **10**, 433–436 (1997).
18. Kleiner, M., Brainard, D. & Pelli, G. “What’s new in Psychtoolbox-3?” *Percept. 36 ECVP Abstr. Suppl.* (2007).
19. Pelli, D. G. The VideoToolbox software for visual psychophysics: transforming numbers into movies. *Spat. Vis.* **10**, 437–442 (1997).
20. Cornelissen, F. W., Peters, E. M. & Palmer, J. The Eyelink Toolbox: Eye tracking with MATLAB and psychophysics toolbox. *Behav. Res. Methods Instrum. Comput.* **34**, 613–617 (2002).
21. Ehinger, B. V., Groß, K., Ibs, I. & König, P. A new comprehensive eye-tracking test battery concurrently evaluating the Pupil Labs glasses and the EyeLink 1000. *PeerJ* **7**, e7086 (2019).
22. Dimigen, O., Sommer, W., Hohlfeld, A., Jacobs, A. M. & Kliegl, R. Coregistration of eye movements and EEG in natural reading: analyses and review. *J. Exp. Psychol. Gen.* **140**, 552–572 (2011).
23. Engbert, R. & Kliegl, R. Microsaccades uncover the orientation of covert attention. *Vision Res.* **43**, 1035–1045 (2003).

24. Engbert, R. & Mergenthaler, K. Microsaccades are triggered by low retinal image slip. *Proc. Natl. Acad. Sci.* **103**, 7192–7197 (2006).
25. Otero-Millan, J., Troncoso, X. G., Macknik, S. L., Serrano-Pedraza, I. & Martinez-Conde, S. Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator. *J. Vis.* **8**, 21 (2008).
26. Rolfs, M., Engbert, R. & Kliegl, R. Crossmodal coupling of oculomotor control and spatial attention in vision and audition. *Exp. Brain Res.* **166**, 427–439 (2005).
27. Nuwer, M. R. *et al.* IFCN standards for digital recording of clinical EEG. *Electroencephalogr. Clin. Neurophysiol.* **106**, 259–261 (1998).
28. van Driel, J., Olivers, C. N. L. & Fahrenfort, J. J. High-pass filtering artifacts in multivariate classification of neural time series data. *J. Neurosci. Methods* **352**, 109080 (2021).
29. Oostenveld, R., Fries, P., Maris, E. & Schoffelen, J.-M. FieldTrip: Open Source Software for Advanced Analysis of MEG, EEG, and Invasive Electrophysiological Data. *Comput. Intell. Neurosci.* **2011**, e156869 (2010).
30. Bates, D., Mächler, M., Bolker, B. & Walker, S. Fitting Linear Mixed-Effects Models Using lme4. *J. Stat. Softw.* **67**, 1–48 (2015).
31. RStudio Team. RStudio: Integrated Development Environment for R}. (2020).
32. Winter, B. Linear models and linear mixed effects models in R with linguistic applications. Preprint at http://arxiv.org/abs/1308.5499 (2013).
33. Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K. & Schölkopf, B. An Introduction to Kernel-Based Learning Algorithms. *IEEE Trans. NEURAL Netw.* **12**, 21 (2001).
34. Lin, C.-J. & Chang, C.-C. LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol. TIST 23 1–27* (2011).
35. Stehr, D. A., Garcia, J. O., Pyles, J. A. & Grossman, E. D. Optimizing multivariate pattern classification in rapid event-related designs. *J. Neurosci. Methods* 109808 (2023).
36. Carlson, T. A., Grootswagers, T. & Robinson, A. K. An introduction to time-resolved decoding analysis for M/EEG. *ArXiv190504820 Q-Bio* (2019).
37. Grootswagers, T., Wardle, S. G. & Carlson, T. A. Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *J. Cogn. Neurosci.* **29**, 677–697 (2017).
38. King, J.-R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210 (2014).
39. Guggenmos, M., Sterzer, P. & Cichy, R. M. Multivariate pattern analysis for MEG: A comparison of dissimilarity measures. *NeuroImage* **173**, 434–447 (2018).
40. Kriegeskorte, N. & Diedrichsen, J. Peeling the Onion of Brain Representations. *Annu. Rev. Neurosci.* **42**, 407–432 (2019).
41. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational Similarity Analysis – Connecting the Branches of Systems Neuroscience. *Front. Syst. Neurosci.* **2**, 1–28 (2008).
42. Lage-Castellanos, A., Valente, G., Formisano, E. & Martino, F. D. Methods for computing the maximum performance of computational models of fMRI responses. 25 (2019).
43. Nili, H. *et al.* A Toolbox for Representational Similarity Analysis. *PLoS Comput. Biol.* **10**, e1003553 (2014).
44. Maris, E. & Oostenveld, R. Nonparametric statistical testing of EEG- and MEG-data. *J. Neurosci. Methods* **164**, 177–190 (2007).
45. Carlson, T. A., Tovar, D. A., Alink, A. & Kriegeskorte, N. Representational dynamics of object vision: The first 1000 ms. *J. Vis.* **13**, 1 (2013).
46. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat. Neurosci.* **17**, 455–462 (2014).
47. Contini, E. W., Wardle, S. G. & Carlson, T. A. Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia* **105**, 165–176 (2017).

48. Kaneshiro, B., Perreau Guimaraes, M., Kim, H.-S., Norcia, A. M. & Suppes, P. A Representational Similarity Analysis of the Dynamics of Object Processing Using Single-Trial EEG Classification. *PLOS ONE* **10**, e0135697 (2015).
49. Karapetian, A. *et al.* Empirically identifying and computationally modelling the brain-behaviour relationship for human scene categorization. Preprint at https://doi.org/10.1101/2023.01.22.525084 (2023).
50. Jozwik, K. M. *et al.* Disentangling five dimensions of animacy in human brain and behaviour. *Commun. Biol.* **5**, 1–15 (2022).
51. Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R. & Khaligh-Razavi, S.-M. Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLOS Comput. Biol.* **15**, e1007001 (2019).
52. Diedrichsen, J. & Kriegeskorte, N. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Comput. Biol.* **13**, e1005508 (2017).
53. Kriegeskorte, N. Relating population-code representations between man, monkey, and computational models. *Front. Neurosci.* **3**, (2009).
54. Kowler, E., Anderson, E., Dosher, B. & Blaser, E. The role of attention in the programming of saccades. *Vision Res.* **35**, 1897–1916 (1995).
55. Hicheur, H., Zozor, S., Campagne, A. & Chauvin, A. Microsaccades are modulated by both attentional demands of a visual discrimination task and background noise. *J. Vis.* **13**, 18–18 (2013).
56. Itti, L. & Koch, C. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Res.* **40**, 1489–1506 (2000).
57. Fang, Y., Gill, C., Poletti, M. & Rucci, M. Monocular microsaccades: Do they really occur? *J. Vis.* **18**, 18 (2018).
58. Rolfs, M., Kliegl, R. & Engbert, R. Toward a model of microsaccade generation: The case of microsaccadic inhibition. *J. Vis.* **8**, 5–5 (2008).
59. Bennett, L., Melchers, B. & Proppe, B. Curta: A General-purpose High-Performance Computer at ZEDAT, Freie Universität Berlin. (2020).

8 Author contributions

G. H. and R. M. C. designed research; G. H. and A. P. C. acquired data; G. H. and A. P. C. analyzed data; G. H., and R. M. C. interpreted results; G. H. prepared figures; G. H. drafted manuscript; G. H., A. P. C., and R. M. C. edited and revised the manuscript.

9 Competing Interests

The authors declare no competing interests.

10 Acknowledgements

G.H. is supported by a Ph.D. fellowship provided by the European Research Council (ERC) starting grant (ERC-StG-2018-803370). A.P.C. is supported by the Research Focus Cognitive Science (RFCS). R.M.C. is supported by German Research Council (DFG) grants (CI 241/1-1, CI 241/3-1, CI 241/1-7) and the European Research Council (ERC) starting grant (ERC-StG-2018-803370). The authors would like to thank the HPC Service of ZEDAT, Freie Universität Berlin, for computing time[59] and Agnessa Karapetian and Marleen Haupt for their valuable input. We thank Antoniya Boyanova for her help with data collection. Icons are made by Freepik from www.flaticon.com.

Chapter 6

General Discussion

6.1 Summary

This dissertation investigated how scene structure and involuntary eye movements influence the extraction of scene and object information from natural stimuli in the visual system. Projects I-III used a combination of EEG and fMRI to investigate the effect of natural scene structure on scene perception and the extraction of object information from natural scenes. Project IV used a combination of EEG and eye tracking to quantify if and to which extent eye movements influence the extraction of object and category information from natural stimuli. The first section of the general discussion offers an overview of the main findings and conclusions of the four studies conducted during this dissertation. The second section will discuss the overarching implications of all projects. The third section evaluates methodological considerations, resulting limitations and derives future directions.

6.2 Main findings

6.2.1 Project I

In project I, we investigated the impact of spatial and categorical regularities on scene representations in healthy human adults (Kaiser et al., 2020a). Humans efficiently extract information from natural scenes (Potter, 1975; Thorpe et al., 1996). Several studies have shown that one reason for this efficiency can be found in the inherent structure of natural scenes. When this structure is interrupted, perception and categorization of these scenes are strongly impaired (Biederman, 1972; Biederman et al., 1974). However, the impact of spatial and categorical regularities on scene-selective neural responses have not been investigated in the past. Participants participated in two fMRI sessions (upright scenes n=20, inverted scenes n=20) and one EEG session (n=20). Seventeen participants participated in both fMRI sessions, and three participants only participated in sessions one or two. Participants passively viewed natural scene stimuli during all three experimental sessions. At the same time, they completed a demanding orthogonal task by responding to a subtle color change of the fixation cross. Each stimulus was partitioned into four quadrants. Each of the four quadrants was drawn from 24 scenes belonging to four categories. All scene parts were either drawn from the same category (categorically intact) or from different categories (categorically jumbled). Those four quadrants were then either kept in their correct spatial location (spatially intact) or systematically recombined (spatially jumbled). This manipulation generated a 2x2 design, answering how categorical and spatial coherences are reflected in the neural sensitivity to scene structure. Stimuli were presented both upright and inverted.

Using EEG and fMRI multivariate and univariate analyses, we tested for the sensitivity to spatial and categorical structure using two complementary analyses. To test for spatial sensitivity, we decoded spatially intact from spatially jumbled scenes (irrespective of category). To test for categorical sensitivity, we decoded categorically intact from categorically jumbled scenes (irrespective of spatial structure). We showed that sensitivity to spatial (but not categorical) scene structure emerged in OPA and PPA and after 255 ms. This effect was stronger for upright than inverted scenes facilitating the interpretation that this effect shows genuine sensitivity to spatial scene structure instead of just reflecting differences in low-level properties of the scenes.

6.2.2 Project II

Building upon these findings, project II aimed to investigate whether the presence of an intact scene structure facilitates the cortical analysis of the categorical content of that scene (Kaiser et al., 2020b). Previous studies have shown that the visual system is sensitive to the inherent structure of our natural world (Abassi & Papeo, 2020; Baldassano et al., 2017; Kaiser et al., 2014; Kim & Biederman, 2011; Roberts & Humphreys, 2010). However, it was still unclear how and whether this structure aids in the extraction of a scene's categorical content.

Project I did not reveal cortical sensitivity to categorical scene structure. Importantly, even though no effect of categorical intact versus categorical jumbled scenes on scene-selective responses could be detected, the scenes category could still be decoded from the EEG data between 45 ms and 660 ms (Kaiser et al., 2020a). In project II, we reused the EEG data collected for project I (n=20) to test whether real-world structure facilitates the emergence of scene categories. Stimuli were drawn from four categories: churches, houses, roads, and supermarkets. Four parts from different scenes drawn from the same scene category were combined in their correct spatial locations for the categorical intact scenes. Four parts from four different scenes of the same category were combined for the jumbled scenes, with the spatial location jumbled in a crisscrossed way. All stimuli were included upright and inverted. To track cortical representations across time, we used a cumulative decoding approach. This approach uses a larger amount of data for decoding than standard decoding techniques by considering all time points prior to the currently decoded time point. Therefore, more data were available at each subsequent step while maintaining temporal precision in the forward direction. Consequently, this provided increased sensitivity for detecting decoding onsets compared to standard time series decoding (Ramkumar et al., 2013). The analysis was conducted separately for the intact and jumbled scenes. For upright scenes, we found that the EEG signal conveyed robust category information for both, the spatially intact and spatially jumbled scenes. Between 105 and 800 ms, significantly enhanced decoding for the spatially intact versus jumbled scenes emerged. No significant difference between the intact and jumbled scenes could be found for the inverted scenes, even though the category could be decoded from both. Interestingly, category information was statistically comparable for the intact upright and inverted scenes, suggesting that the jumbling manipulation specifically harms category information in the upright scenes. Our results, therefore, show that scene structure matters more for the processing of upright scenes than for the processing of inverted scenes. Overall, our results provide evidence that the facilitation of category information by real-world structure emerges within 200 ms of vision. In line with project I, we were able to show that this facilitation can be attributed to the adherence to the real-world structure instead of differences in low-level properties.

6.2.3 Project III

Projects I and II showed that cortical scene representations are tightly linked to real-world structure (Kaiser et al., 2020a, 2020b). Participants in both studies were instructed to perform an orthogonal fixation task which did not allow us to directly assess the behavioral relevance of the spatial regularities we observed in the brain data. Project III sought to investigate the behavioral relevance of the previously described neural findings by combining neural recordings with a more naturalistic task. In detail, we investigated whether typical real-world environments help participants to efficiently solve an object (person versus car) and a scene (rural versus urban) categorization task while recording fMRI (n=25). The stimuli set consisted of colored natural scene photographs. In each photograph, a person or a car was depicted in a rural or urban environment in either of the four stimuli quadrants. Spatial regularities of the scenes were interrupted by jumbling the four quadrants in a crisscrossed way. Using a combination of univariate and correlation-based multivariate analysis techniques, we were able to show that participants were faster and more accurate in performing the object and scene categorization task when perceiving intact versus jumbled scenes. Object information was enhanced for intact versus jumbled scenes only when the objects were relevant to the current behavioral goals. These findings revealed that early real-world structure is a crucial asset for solving complex real-world tasks (Kaiser et al., 2021).

6.2.4 Project IV

During the data collection for projects I, II, and III, participants were instructed to fixate on a centrally presented fixation cross. Project IV sought to investigate the influence of two different fixation crosses (a bullseye versus a standard fixation cross) on eye movements and the classification of natural images from EEG. While eye movements are a ubiquitous and natural behavior, they are undesirable in many highly controlled experimental visual paradigms. Previous studies revealed that eye movements affect various analysis techniques, including MVPA (Mostert et al., 2018; Quax et al., 2019) and univariate analysis techniques (Dimigen & Ehinger, 2021; Dimigen et al., 2009).

In project IV, we used a combination of EEG and eye tracking to compare the effect of two different fixation symbols – the standard fixation cross and the bullseye fixation cross – in the context of a visual paradigm with centrally presented naturalistic object images, using behavioral and multivariate analysis techniques. Participants (n=30) viewed natural object stimuli while performing an orthogonal task to keep them engaged. Our findings were threefold. First, the bullseye fixation cross reduced the number of saccades and amplitude size of microsaccades. Second, the bullseye fixation cross subtly reduced classification accuracy in both eye tracking and EEG data for the classification of single object images, but not for the super-level category animacy. Third, using representational similarity analysis, we found a systematic relationship between eye tracking and EEG data at the level of single images for the standard, but not for the bullseye fixation cross. These findings suggest that systematic eye movements indeed influence the results of MVPA, albeit to a small degree. Therefore, we recommend the bullseye fixation cross in experimental paradigms with fixation, particularly when control of fixation is beneficial.

6.3 Key implications across projects

In summary, projects I-III aimed at answering three interconnected questions to further our understanding of scene processing. While project I showed that intact spatial structure impacts scene-selective cortical responses in space and time, project II provided evidence that spatial structure facilitates the extraction of scene categories. Project III added a brain-behavior link by investigating whether and how spatial regularities aid object extraction from a scene, while manipulating attention through an object and a scene classification task. The project results show that intact spatial structure enhances the representation of objects in a scene only if the objects are behaviorally relevant. Data from all three studies were analyzed using univariate and multivariate analysis techniques, while participants were instructed to fixate on a centrally presented fixation cross. Not only has it been shown that the choice of fixation cross affects participants' eye movement patterns (Thaler et al., 2013), there has been a growing discussion in recent years about how univariate and multivariate analysis techniques are influenced by participants' voluntary and involuntary eye movements (Mostert et al., 2018; Quax et al., 2019). Even though participants are often instructed to fixate during stimulus presentations, participants still exhibit involuntary eye movements (Quax et al., 2019; Thaler et al., 2013). Project IV investigated whether and to which extent eye movements influence EEG decoding during a simple perceptual fixation task. We showed that the classification of neuroscientific data is influenced to a small degree by systematic eye movements at the level of single images for the standard but not for the bullseye fixation cross. This indicates that lower-but not higher-level order stimulus properties might be influenced to a small degree by systematic eye movements.

6.3.1 The temporal dynamics of scene processing

Projects I and II expanded our understanding of the temporal and spatial dynamics of scene processing, tying into an already extensive knowledge base. In the following paragraph, I will shortly outline the timeline of scene processing and highlight where our results provide new insight.

Scene perception is aided by several different neural mechanisms. Human observers are as fast in global context categorization as in object categorization (Fabre-Thorpe et al., 2001; Joubert et al., 2007), even when the scene is only presented briefly. Such a performance cannot be explained by the individual processing of every single object in a scene sequentially because this would require significantly longer processing durations. This suggests that scene perception is not a sequential process and seems to be more than the combination of the scenes' individual parts. Single images of natural scenes are discriminated early, starting from 50 ms with a peak at 97 ms by visual representations similar to single images with other visual content (Carlson et al., 2013; Cichy et al., 2014; Isik et al., 2014). We could show that within 200 ms of vision, the extraction of a scene's categorical content is facilitated by spatial regularities (Kaiser et al., 2020b). These results go hand in hand with results from single object processing, which showed that object representations are enhanced after 140 ms if they appear in their typical real-world location (Issa & DiCarlo, 2012; Kaiser et al., 2018). This indicates that the adherence to typical real-world location of scene parts facilitates the extraction of scene-relevant content.

Apart from facilitating the extraction of the scenes' categories content, the adherence to typical-real-world structure also modifies scene-selective neural responses. Using univariate analysis, we found a significant main effect of spatial structure for the upright scenes, emerging between 225 and 425 ms, with a peak at 235 ms (Kaiser et al., 2020a). These results align with earlier studies, showing that approximately 220 ms after stimulus onset, ERP components elicited a significantly stronger response to scenes than to other categories, corresponding to the P2 component. This component is sensitive to scenes at both the categorical level (open versus closed natural scenes) and the single image level, where it reflected scene statistics and behavioral ratings of naturalness and spatial expanse (Harel et al., 2016). We replicated these findings in our study, with the difference that this marker appeared slightly later at 235 ms, indicating an effect of spatial structure. These findings align with earlier experiments showing that MEG signals are responsive to scenes between 200 and 300 ms (Sato et al., 1999), strengthening our understanding of scene-selective neural responses.

Several previous studies have shown that the jumbling of scene structure strongly impairs perception (Biederman, 1972; Biederman et al., 1974). We showed that this impairment in perception is also reflected in scene-selective neural responses. Sensitivity to spatial structure emerged after 255 ms of processing, after scene-selective peaks in ERPs (Harel et al., 2016; Sato et al., 1999) and shortly after scene layout properties like scene size at around 250 ms (Cichy et al., 2016). Together, these results enhance our understanding of the impact of real-world scene structure on the analysis of the content of a scene. More specifically, we showed that higher-level scene properties are analyzed during a dedicated processing stage. We expanded the understanding of the temporal processing cascade underlying scene perception by identifying at which time points the spatial structure of a scene facilitates the analysis of both its structural features and categorical content.

6.3.2 The spatial dynamics of scene processing

We did not only assess the temporal but also the spatial dynamics underlying scene processing. Several previous studies have identified scene-selective areas that respond more when viewing scenes compared with objects or faces and may be specialized for representing specific aspects of the environment (Dilks et al., 2013; Epstein et al., 1999; Persichetti & Dilks, 2018) including PPA (Epstein & Kanwisher, 1998), OPA (Dilks et al., 2013; Hasson et al., 2003), and the RSC (O'Craven & Kanwisher, 2000). The experiments conducted in projects I and III extended our understanding of the contribution of these areas to scene perception.

Our fMRI results showed that the structural analysis of scene content was reflected through activity in PPA and OPA, showing a stronger response to spatially intact compared to spatially jumbled scenes (Kaiser et al., 2020a). This aligns with findings showing that PPA is responsive to viewpoint specificity and discriminates between different views (Park & Chun, 2009). Findings from lesion studies further complement these findings. If

PPA is damaged, e.g., through a stroke, patients report losing their sense of a scene as a coherent whole and problems identifying places and landmarks (Aguirre & D'Esposito, 1999).

In addition to the structural content of a scene, PPA might also contain information about the scene category being viewed. Previous studies revealed a network of regions, including PPA, RSC, and LOC, contributing to the human ability to categorize natural scenes (Walther et al., 2009). Irrespective of PPAs involvement in natural scene categorization, project I did not reveal sensitivity to categorical scene structure, indicating that spatial structure impacts cortical responses more strongly than categorical scene structure (Kaiser et al., 2020a). Importantly, these results do not contradict the findings of the involvement of PPA in natural scene categorization. Unfortunately, due to the nature of the design, we were not able to classify the scene's categorical content from the fMRI data directly, as scene categories were intermixed in the fMRI blocks. However, in a supplementary analysis, we showed that the scene category could be decoded from EEG data, even though sensitivity to categorical scene structure could not be found. This suggests that scene category can still be classified from the used stimulus set, implying that the analysis of categorical scene content is independent of the effects of categorical scene structure on the cortical response in PPA and OPA.

Project III replicated the sensitivity of both PPA and OPA to the structural content of a scene with a different set of stimuli. PPA was the only area showing an additional modulation by task demand. PPA showed significantly stronger responses when participants were asked to categorize scenes compared to being asked to categorize objects (Kaiser et al., 2021). This expands our understanding of PPA's involvement in scene processing by suggesting an increasing importance of computation in higher-level scene-selective cortex when the attributes of the scenes are relevant for behavior.

We were additionally able to show a main effect of spatial structure in LO and EBA. These results are in line with earlier studies showing that positional regularities within the real-world influence object perception, with more accurate decoding within LOC for objects appearing in their typical real-world positions (Kaiser & Cichy, 2018). Importantly, we showed that spatial scene structure not only matters for scene-related neural responses but also enhances object information in LO and EBA when objects are embedded in intact versus jumbled scenes. Critically, this enhancement only occurred when the participants were asked to identify the object instead of the scene, indicating that scene structure only facilitated the extraction of objects from a scene background if the object is task-relevant (Kaiser et al., 2021).

6.3.3 The interplay between the object and scene network

The results from project III show that the interplay between attention and scene structure aids in extracting task-relevant object information. The behavioral results of this study revealed that incoherent scene structure impact task performance for both - the object and scene categorization tasks. Additionally, object information was significantly enhanced in LO and EBA, if the scene's structure was coherent and the objects task-relevant. When participants were instructed to perform the scene task, this facilitation could not be observed. If the scene and object network would be functionally separate, we would expect to see the the task-based enhancement of object information in both, spatially coherent and jumbled scenes. Therefore, the findings from project III further support available evidence that the object and scene networks are not functionally separate (Brandman & Peelen, 2017, 2019; Wischnewski & Peelen, 2021). Recent studies revealed that that scene context strongly enhances the category representation of degraded objects, which are hard to recognize in isolation. Object-selective areas LO and pFs additionally showed strong scene-based facilitation, demonstrating that degraded objects can be fully recognized when aided by scene-based context. The facilitation of object information was correlated with activity in scene-related areas RSC and PPA, showing that object and scene processing mechanisms effectively interact with each other to help with the efficient processing of object information (Brandman & Peelen, 2017). This context-based object recognition is causally supported by OPA and LOC (Wischnewski & Peelen, 2021). We expanded these results by showing that the interaction between the scene and object networks undergoes a task-based modulation and is mediated by attention (Kaiser et al., 2021).

At a first glance, these results could be interpreted as contradictory to the study's results by Brandman and Peelen, 2017, who showed a direct interaction between the object and scene networks. These differing results could e.g., be explained by the different task demands. Participants performed an orthogonal oddball task where they had to respond every time a number was presented instead of a scene while being instructed to attend to the object (Brandman & Peelen, 2017). In contrast, participants in project III were instructed to perform an object or a scene categorization task, efficiently manipulating attention toward the scene or the object content of the image. The objects themselves were completely irrelevant if participants were asked to categorize the scenes effectively manipulating attention away from the object. We found a strong enhancement of object information in LO and EBA when the task was to attend to the object, while this enhancement was not present when participants were asked to attend to the scenes. Our findings align with two studies that have already shown that task-relevant objects are represented more accurately than task-irrelevant objects (Peelen et al., 2009; Peelen & Kastner, 2011).

However, this raises the question of why an intact scene structure enhances task-relevant object information. According to the two-pathway hypothesis (Wolfe et al., 2011), the combination of a selective and a nonselective pathway allows for efficient processing of scenes. First, the observer rapidly extracts statistical information from the entire image through global non-selective image processing. These statistical regularities include the mean and distribution of several visual feature dimensions (e.g., size (Chong & Treisman, 2003), orientation (Parkes et al., 2001), some contrast texture descriptors (Chubb et al., 2007), and the presence of classes of objects (Vanrullen, 2009)). Second, the image details are analyzed in a selective step.

Due to the jumbling manipulation in project III, some statistical regularities are interrupted and might hinder the rapid extraction of such regularities. Indeed, we showed that people are more accurate and faster in identifying objects and scenes in an intact versus jumbled image. We also find that in both LO and EBA, object category information was significantly higher for objects that were embedded in intact scenes than for objects embedded in jumbled scenes. These findings underline the interpretation that adherence to statistical regularities facilitates the extraction of other aspects of the scene, possibly through global non-selective image processing (Greene & Oliva, 2009; Oliva & Torralba, n.d., 2007).

6.3.4 The effect of eye movements on object and scene processing

The previous paragraph discussed the impact of real-world regularities on scene-selective responses and the extraction of object and category information from natural stimuli. While testing participants in experimental settings, we often simplify the stimuli they are confronted with and make them look at these pictures in a highly standardized setting. In most cases, these experimental settings involve the participants sitting in front of a monitor with their heads stabilized on a chin rest and their eyes fixated on a fixation cross in the center of the screen. Importantly, our eyes constantly move not only during real-world interactions but also during fixation tasks. Therefore, project IV set out to investigate how participants‘ eye movements during fixation influence multivariate pattern analysis of time-resolved EEG data. We showed that for a highly controlled experiment where participants were instructed to fixate, classification at the level of single images is influenced by systematic stimulus-specific eye movements. In contrast, the classification of category, and hence higher-level conceptual aspects of the stimulus, was not influenced by stimulus-specific eye movements. Single object images were associated systematically with eye movements that supported classification. In contrast, object image sets at the supra-category level of animacy were not and eye movements averaged out into similar distributions for animate vs. inanimate objects. This might be a function of the stimulus set used. In an experimental setup with two categories with low intra-group image variability at the pixel level (e.g., highly controlled images of front-view faces and houses), we would expect systematic eye movement effects also at the category level. Alternatively, the causal relationship may also operate in reverse, with brain activity being the instigator of eye movements. In this scenario, neural activity could arise from low-level stimulus characteristics, which trigger neural activity that can subsequently be decoded. As a result, the stimulus-dependent eye movements may be caused by the neural activity, rather than being its consequence (Thielen et al., 2019). However, the reduction of classification accuracies when a fixation cross is associated with fewer eye movements speaks in favor of the first interpretation, even though an interaction of the two effects cannot be discarded.

Naturally, the question arises whether the results from project IV provide us with insight into the interpretation of the results of projects I and II. Due to the narrowly tested stimulus set in project IV, which only included 40 exemplars of natural object stimuli on natural backgrounds, we cannot automatically assume that these findings generalize to the analysis of scene stimuli. In line with scene stimuli from project I and II, the stimuli used in project IV were not controlled for luminance, color, or background. This was deliberately chosen as we aimed to replicate a setup with naturalistic stimuli closely.

Nevertheless, while this particular stimulus set up did not allow for a systematic variation of stimulus features, let us - for the sake of this discussion - assume that the findings of project IV translate to other setups and that lower-level properties primarily drive stimulus-relevant eye movements. What would this imply for interpreting the temporal results from project I and II?

Scenes contain a lot of higher-level representations like scene categories (Walther et al., 2009; Walther et al., 2011) and familiar places (Marchette et al., 2015; Park & Chun, 2009). However, several studies have shown that the recognition and classification of scenes do not primarily rely on higher, but also on lower-level properties (Cant & Xu, 2012; Kauffmann et al., 2015; Nasr et al., 2014; Rajimehr et al., 2011). Other studies have explicitly emphasized the role of midlevel features in scene recognition (Choo & Walther, 2016). Recently, it has been suggested that low-, mid-, and high-level features of scenes can largely explain the same variance in brain responses in scene-selective regions (Lescroart et al., 2015).

Taken together, scene analysis does not only rely on high-level but also on low-level properties. In turn, it is cogitable that lower-level properties of the scene stimuli could drive stimulus-relevant eye movements. Based on the findings from project IV, it may be possible that participants participating in projects I and II displayed more eye movements with the standard fixation cross than they would have with implementing a bullseye fixation cross. That being said, we are optimistic that, even if present, the effects of stimulus-relevant eye movements on our results would be marginal. There are several reasons for that. First, we classified global scene properties like the spatial arrangement and category of a scene which are higher-level properties. Project IV revealed an effect of systematic eye movements for the classification of single images only, not for the classification of the category. Second, in line with the stimuli used in project IV variability between the different scene stimuli was high (images were not highly controlled at the pixel level). Third, we found that the decoding of spatial coherence and category in projects I and II are affected by the jumbling manipulation only for upright but not inverted stimuli, suggesting that the effect is driven by high and not low-level stimulus properties. Fourth, the effect of eye movements on the classification accuracy of single images in project IV was much smaller than the overall effect sizes (3.41% reduction compared to the overall classification results of 26.22% and 23.98% for the standard and bullseye fixation cross).

The precise effect should be investigated by adopting the paradigm of project IV to incorporate natural scene stimuli instead of natural object stimuli.

6.4 Methodological considerations and future directions

After already touching upon open questions based on our findings in the previous chapter, I will now discuss further avenues for future studies taking into account experimental design and methodological considerations.

6.4.1 Experimental design considerations

The jumbling paradigm

We manipulated spatial and categorical scene structure in projects I-III with a jumbling paradigm. This paradigm introduces a strong manipulation that conflates several aspects of a scene's inherent structure (Biederman, 1972; Biederman et al., 1974). Jumbling the different scene quadrants disrupts the typical positioning of the individual pieces of a scene, the positioning of objects relative to each other, and the typical geometry of the scene itself. Altogether, this results in non-naturalistic scene stimuli as artificial discontinuities are introduced. We controlled for these artificial discontinuities in projects I and II by introducing comparable discontinuities in the stimuli showing intact scenes. Using this method, we could show that the effects of scene jumbling persists with artificial discontinuities present in both intact and jumbled scenes (Kaiser et al., 2020a, 2020b).

Even though these artificial discontinuities were not controlled for in project III, these findings strengthen the interpretation that the effect of natural scene structure and attention on the enhancement of object information extraction in project III is due to a genuine difference between intact and jumbled scenes and cannot solely be explained by discontinuities in the stimulus presentation when the different quadrants of the image were jumbled. Nevertheless, future studies are needed to disentangle the different features that drive the sensitivity to spatial scene structure.

Real-world testing conditions of scenes and tightly controlled laboratory settings

One common disadvantage to all four projects is the generalizability of the findings to real-world scenarios. While we used the term real-world scenes in the context of this thesis, this term is related to predictable distributions of information across natural 2D scenes. However, a scene in the real world entails much more than a 2D visual representation. In addition to containing cues from odor and sound (for an opinion piece on how to integrate different modalities in scene perception, see Cichy and Teng, 2017), scenes are 3D representations and can be acted within. Additionally, observers are usually unrestricted in whether and how they move their eyes to understand the visual world around them. Translating effects from the lab to "the wild" has, for example, shown that the well-established N170 effect in ERP research translates to a more ecological setting (Gert et al., 2021). A logical next step would be to test whether the effect of spatial scene structure persists when tested in a more ecologically valid framework. To do so, one could make use of the recently advanced VR and mobile EEG possibilities.

6.4.2 Methodological considerations

Eye tracking data recording

Eye movement data collected for project IV was recorded using an Eyelink 1000 Tower Mount, which only allowed for monocular recordings of one eye. Recently, there has been a growing discussion about monocular and binocular microsaccades (Fang et al., 2018). A monocular setup limits the ability to discriminate microsaccades from noise as the binocular criterion cannot be taken advantage of (Ciuffreda & Tannen, 1995; Fang et al., 2018; Krauskopf et al., 1960; Schulz, 1984; Thaler et al., 2013). This could be circumvented by utilizing a setup that allows for binocular recordings, which was not accessible for this thesis.

The interaction between EEG preprocessing pipelines and eye movements

In addition to the consideration that stimuli from projects I and II might not directly translate to the experimental setup of project IV, different preprocessing pipelines were utilized. This might limit the extraction and transfer of information from Project IV to the other projects. EEG preprocessing in project IV was limited to baseline correction, filtering, and the removal of excessively noisy channels and trials. The preprocessing pipeline of projects I and II additionally included the removal of eye movements from the collected EEG data using ICA (Jung et al., 2000) (Project I) or the combination of ICA (Jung et al., 2000) and principal component analysis (Jolliffe, 2011; Jolliffe & Cadima, 2016) (Project II). This might further mitigate eye movements' effects on decoding accuracies.

Several arguments can be made to justify the exclusion of further preprocessing steps in the pipeline for project IV. First, ICA can detect and separate several sources within EEG data, including eye movements, heart and muscle artifacts, and brain data. However, using ICA without an automated detection algorithm for eye movement reduction hinders reproduction, and common transformation techniques like ICA are sensitive to early preprocessing pipeline choices and data preparation. Amplitude features often vary greatly from headset to headset and session to session and, therefore, generally do not improve the signal-to-noise ratio (Bigdely-Shamlo et al., 2015). Second, ICA and regression alone are insufficient to remove systematic eye movements in M/EG data (Mostert et al., 2018; Quax et al., 2019). Third, it has been a long-standing challenge in neuroscience to be able to forgo as many manual steps in the preprocessing pipelines as possible. Fourth, a recent paper showed that automated rejection of ICA eye and muscle artifacts did not increase performance reliability for ERPs (Delorme, 2023). As these findings cannot directly be transferred to MVPA, future studies should systematically test the interactions between eye movements, removal techniques, and fixation cross usage.

Interindividual differences in eye movements

Another limitation on the interpretation of eye movement effects on EEG MVPA is interindividual differences between observers. Several studies showed varying fixation behavior between individual participants (Bargary et al., 2017; Thielen et al., 2019). Thielen et al., 2019 demonstrated that the amplitude of observers' eye movements significantly correlated with classification accuracy from the eye movement data, indicating that larger eye movements correlate with higher classification accuracy. A subset of participants showed eye movements that covaried with the stimulus in question. One possible explanation could be the (in)ability of participants to focus appropriately. In this particular study, mainly saccades with large amplitudes drove classification accuracies (Thielen et al., 2019). Interestingly, the reduction in decoding accuracies in project IV was mainly driven by fewer saccades in general and smaller microsaccade amplitudes. However, we cannot ascertain whether individual performances drive this effect as we did not control for interindividual differences.

This needs to be addressed in future studies to substantiate the size of the effect and to show whether the fixation cross has a differential impact on differently skilled observers.

One possible solution would be to divide participants into groups of better and worse observers. This would answer whether fixation improvement for different fixation crosses has a differential impact on participants who perform better in fixations from the start and whether participants who are better at suppressing saccades exhibit worse classification accuracy for EEG data. Even though it might not be feasible to only test participants with a better oculomotor signature, these steps could ensure to consider inter-individual differences when interpreting decoding accuracies.

6.5 Conclusion

Humans can efficiently extract information from scenes in the environment. One reason for this efficient behavior is the inherent natural structure of our surroundings. In projects I-III, we investigated the effect of spatial and categorical regularities on scene and object processing using a mixture of EEG and fMRI. Another factor that allows humans to extract information from natural stimuli efficiently is eye movements. In project IV, we investigated if and to which extent eye movements influence MVPA of EEG data. In detail, we aimed to answer four main questions: (1) Does real-world structure impact scene-selective neural responses? (2) Does the spatial structure of a scene facilitate the cortical analysis of the scene's categorical content? (3) Does the spatial structure of a scene's context aid in the extraction of task-relevant object information from the scene? (4) Does the choice of different fixation crosses influence eye movements and the classification of natural images from EEG and eye tracking?

We showed that: (1) Spatial scene structure impacts scene-selective neural responses in OPA and PPA and reveals genuine sensitivity to spatial scene structure from 255 ms on, while scene-selective neural responses are less sensitive to categorical scene structure. (2) Spatial scene structure facilitates the extraction of the scene's categorical content within 200 ms of vision. (3) Coherent scene structure facilitates the extraction of object information if the object is task-relevant, suggesting a task-based modulation. (4) The bullseye fixation cross reduces eye movements on the single image level and subtly removes systematic eye movement related activity in M/EEG data.

Overall, this thesis advanced our understanding of the impact of real-world structure and eye movements on the extraction of scene and object information from natural stimuli.

References

Abassi, E., & Papeo, L. (2020). The Representation of Two-Body Shapes in the Human Visual Cortex. *Journal of Neuroscience*, *40*(4), 852–863. https://doi.org/10.1523/JNEUROSCI.1378-19.2019

Aguirre, G. K., & D'Esposito, M. (1999). Topographical disorientation: A synthesis and taxonomy. *Brain*, *122*(9), 1613–1628. https://doi.org/10.1093/brain/122.9.1613

Baldassano, C., Beck, D. M., & Fei-Fei, L. (2017). Human–Object Interactions Are More than the Sum of Their Parts. *Cerebral Cortex*, *27*(3), 2276–2288. https://doi.org/10.1093/cercor/bhw077

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

Bargary, G., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Hogg, R. E., & Mollon, J. D. (2017). Individual differences in human eye movements: An oculomotor signature? *Vision Research*, *141*, 157–169. https://doi.org/10.1016/j.visres.2017.03.001

Biederman, I. (1972). Perceiving Real-World Scenes. *Science*, *177*(4043), 77–80. https://doi.org/10.1126/science.177.4043.77

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

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K.-M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*. https://doi.org/10.3389/fninf.2015.00016

Borji, A., Sihite, D. N., & Itti, L. (2013). Objects do not predict fixations better than early saliency: A re-analysis of Einhäuser et al.'s data. *Journal of Vision*, *13*(10), 18. https://doi.org/10.1167/13.10.18

Brandman, T., & Peelen, M. V. (2017). Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding. *Journal of Neuroscience*, *37*(32), 7700–7710. https://doi.org/10.1523/JNEUROSCI.0582-17.2017

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

Brockmann, D., & Geisel, T. (1999). Are human scanpaths Levy flights?, 263–268. https://doi.org/10.1049/cp:19991119

Campbell, F. W., & Green, D. G. (1965). Optical and retinal factors affecting visual resolution. *The Journal of Physiology*, *181*(3), 576–593. https://doi.org/10.1113/jphysiol.1965.sp007784

Cant, J. S., & Xu, Y. (2012). Object Ensemble Processing in Human Anterior-Medial Ventral Visual Cortex. *Journal of Neuroscience*, *32*(22), 7685–7700. https://doi.org/10.1523/JNEUROSCI.3325-11.2012

Carlson, T. A., Schrater, P., & He, S. (2003). Patterns of activity in the categorical representations of objects. *Journal of Cognitive Neuroscience*, *15*(5), 704–717. https://doi.org/10.1162/089892903322307429

Carlson, T. A., Tovar, D. A., Alink, A., & Kriegeskorte, N. (2013). Representational dynamics of object vision: The first 1000 ms. *Journal of Vision*, *13*(10), 1. https://doi.org/10.1167/13.10.1

Chan, A. W.-Y., Kravitz, D. J., Truong, S., Arizpe, J., & Baker, C. I. (2010). Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature Neuroscience*, *13*(4), 417–418. https://doi.org/10.1038/nn.2502

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortex for perceiving and knowing about objects. *Nature Neuroscience*, *2*(10), 913–919. https://doi.org/10.1038/13217

Chong, S. C., & Treisman, A. (2003). Representation of statistical properties. *Vision Research*, *43*(4), 393–404. https://doi.org/10.1016/S0042-6989(02)00596-5

Choo, H., & Walther, D. B. (2016). Contour junctions underlie neural representations of scene categories in high-level human visual cortex. *NeuroImage*, *135*, 32–44. https://doi.org/10.1016/j.neuroimage.2016.04.021

Chubb, C., Nam, J.-H., Bindman, D. R., & Sperling, G. (2007). The three dimensions of human visual sensitivity to first-order contrast statistics. *Vision Research*, *47*(17), 2237–2248. https://doi.org/10.1016/j.visres.2007.03.025

Cichy, R., Khosla, A., Pantazis, D., & Oliva, A. (2016). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, *53*. https://doi.org/10.1016/j.neuroimage.2016.03.063

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

Cichy, R. M., & Teng, S. (2017). Resolving the neural dynamics of visual and auditory scene processing in the human brain: A methodological approach. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160108. https://doi.org/10.1098/rstb.2016.0108

Ciuffreda, K. J., & Tannen, B. (1995). *Eye movement basics for the clinician*. Mosby OCLC: 31295173.

Contini, E. W., Wardle, S. G., & Carlson, T. A. (2017). Decoding the time-course of object recognition in the human brain: From visual features to categorical decisions. *Neuropsychologia*, *105*, 165–176. https://doi.org/10.1016/j.neuropsychologia.2017.02.013

Cox, D. D., & Savoy, R. L. (2003). Functional magnetic resonance imaging (fMRI) "brain reading": Detecting and classifying distributed patterns of fMRI activity in human visual cortex. *NeuroImage*, *19*(2 Pt 1), 261–270. https://doi.org/10.1016/s1053-8119(03)00049-1

de Haas, B., Schwarzkopf, D. S., Alvarez, I., Lawson, R. P., Henriksson, L., Kriegeskorte, N., & Rees, G. (2016). Perception and Processing of Faces in the Human Brain Is Tuned to Typical Feature Locations. *Journal of Neuroscience*, *36*(36), 9289–9302. https://doi.org/10.1523/JNEUROSCI.4131-14.2016

Delorme, A. (2023). EEG is better left alone. *Scientific Reports*, *13*(1), 2372. https://doi.org/10.1038/s41598-023-27528-0

de-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin & Review*, *23*, 1415–1428. https://doi.org/10.3758/s13423-016-1002-0

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

DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, *73*(3), 415–434. https://doi.org/10.1016/j.neuron.2012.01.010

Dijkstra, N., Mostert, P., de Lange, F. P., Bosch, S., & van Gerven, M. A. (2018). Differential temporal dynamics during visual imagery and perception (S. Kastner, Ed.). *eLife*, *7*, e33904. https://doi.org/10.7554/eLife.33904

Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The Occipital Place Area Is Causally and Selectively Involved in Scene Perception. *Journal of Neuroscience*, *33*(4), 1331–1336. https://doi.org/10.1523/JNEUROSCI.4081-12.2013

Dimigen, O., & Ehinger, B. V. (2019). Analyzing combined eye-tracking/EEG experiments with (non)linear deconvolution models. *bioRxiv*, 735530. https://doi.org/10.1101/735530

Dimigen, O., & Ehinger, B. V. (2021). Regression-based analysis of combined EEG and eye-tracking data: Theory and applications. *Journal of Vision*, *21*(1), 3. https://doi.org/10.1167/jov.21.1.3

Dimigen, O., Valsecchi, M., Sommer, W., & Kliegl, R. (2009). Human Microsaccade-Related Visual Brain Responses. *Journal of Neuroscience*, *29*(39), 12321–12331. https://doi.org/10.1523/JNEUROSCI.0911-09.2009

Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A Cortical Area Selective for Visual Processing of the Human Body. *Science*, *293*(5539), 2470–2473. https://doi.org/10.1126/science.1063414

Einhäuser, W., Spain, M., & Perona, P. (2008). Objects predict fixations better than early saliency. *Journal of Vision*, *8*(14), 18. https://doi.org/10.1167/8.14.18

Epstein, R., Harris, A., Stanley, D., & Kanwisher, N. (1999). The parahippocampal place area: Recognition, navigation, or encoding? *Neuron*, *23*(1), 115–125. https://doi.org/10.1016/s0896-6273(00)80758-8

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

Epstein, R. A., & Baker, C. I. (2019). Scene Perception in the Human Brain. *Annual Review of Vision Science*, *5*(1), 373–397. https://doi.org/10.1146/annurev-vision-091718-014809

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

Fang, Y., Gill, C., Poletti, M., & Rucci, M. (2018). Monocular microsaccades: Do they really occur? *Journal of Vision*, *18*(3), 18. https://doi.org/10.1167/18.3.18

Gegenfurtner, K. R. (2016). The Interaction Between Vision and Eye Movements. *Perception*, *45*(12), 1333–1357. https://doi.org/10.1177/0301006616657097

Gert, A. L., Ehinger, B. V., Timm, S., Kietzmann, T. C., & König, P. (2021). Wild lab: A naturalistic free viewing experiment reveals previously unknown EEG signatures of face processing. https://doi.org/10.1101/2021.07.02.450779

Goodale, M. A., & Milner, A. D. (1992). Seperate visual pathways for perception and action. *Trends in Cognitive Sciences*, *15*(1), 20–25. https://doi.org/10.1016/0166-2236(92)90344-8

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

Grootswagers, T., Wardle, S. G., & Carlson, T. A. (2017). Decoding Dynamic Brain Patterns from Evoked Responses: A Tutorial on Multivariate Pattern Analysis Applied to Time Series Neuroimaging Data. *Journal of Cognitive Neuroscience*, *29*(4), 677–697. https://doi.org/10.1162/jocn_a_01068

Guzman-Martinez, E., Leung, P., Franconeri, S., Grabowecky, M., & Suzuki, S. (2009). Rapid eye-fixation training without eye tracking. *Psychonomic bulletin & review*, *16*(3), 491–496. https://doi.org/10.3758/PBR.16.3.491

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The Temporal Dynamics of Scene Processing: A Multifaceted EEG Investigation. *eneuro*, *3*(5), ENEURO.0139–16.2016. https://doi.org/10.1523/ENEURO.0139-16.2016

Hasson, U., Harel, M., Levy, I., & Malach, R. (2003). Large-Scale Mirror-Symmetry Organization of Human Occipito-Temporal Object Areas. *Neuron*, *37*(6), 1027–1041. https://doi.org/10.1016/S0896-6273(03)00144-2

Haxby, J. V., Gobbini, M. I., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*(5539), 2425–2430. https://doi.org/10.1126/science.1063736

Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534. https://doi.org/10.1038/nrn1931

Hock, H. S., Gordon, G. P., & Whitehurst, R. (1974). Contextual relations: The influence of familiarity, physical plausibility, and belongingness. *Perception & Psychophysics*, *16*(1), 4–8. https://doi.org/10.3758/BF03203242

Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences of the United States of America*, *96*(16), 9379–9384. https://doi.org/10.1073/pnas.96.16.9379

Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of Neurophysiology*, *111*(1), 91–102. https://doi.org/10.1152/jn.00394.2013

Issa, E. B., & DiCarlo, J. J. (2012). Precedence of the Eye Region in Neural Processing of Faces. *Journal of Neuroscience*, *32*(47), 16666–16682. https://doi.org/10.1523/JNEUROSCI.2391-12.2012

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

Jolliffe, I. (2011). Principal Component Analysis. In M. Lovric (Ed.), *International Encyclopedia of Statistical Science* (pp. 1094–1096). Springer. https://doi.org/10.1007/978-3-642-04898-2_455

Jolliffe, I. T., & Cadima, J. (2016). Principal component analysis: A review and recent developments. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, *374*(2065). https://doi.org/10.1098/rsta.2015.0202

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

Jung, T.-P., Makeig, S., Humphries, C., Lee, T.-W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, *37*(2), 163–178. https://doi.org/10.1111/1469-8986.3720163

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

Kaiser, D., & Cichy, R. M. (2018). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology*, *120*(2), 848–853. https://doi.org/10.1152/jn.00229.2018

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

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

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

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

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

Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The Fusiform Face Area: A Module in Human Extrastriate Cortex Specialized for Face Perception. *Journal of Neuroscience*, *17*(11), 4302–4311. https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997

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

Kim, J. G., & Biederman, I. (2011). Where do objects become scenes? *Cerebral Cortex (New York, N.Y.: 1991)*, *21*(8), 1738–1746. https://doi.org/10.1093/cercor/bhq240

Koch, C., & Ullman, S. (1985). Shifts in selective visual attention: Towards the underlying neural circuitry. *Human Neurobiology*, *4*(4), 219–227.

Krauskopf, J., Cornsweet, T. N., & Riggs, L. A. (1960). Analysis of Eye Movements during Monocular and Binocular Fixation*. *Journal of the Optical Society of America*, *50*(6), 572. https://doi.org/10.1364/JOSA.50.000572

Lescroart, M. D., Stansbury, D. E., & Gallant, J. L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. *Frontiers in Computational Neuroscience*, *9*. https://doi.org/10.3389/fncom.2015.00135

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

Lisberger, S. G. (2010). Visual guidance of smooth pursuit eye movements: Sensation, action, and what happens in between. *Neuron*, *66*(4), 477–491. https://doi.org/10.1016/j.neuron.2010.03.027

Logothetis, N. K., & Sheinberg, D. L. (1996). Visual Object Recognition. *Annual Review of Neuroscience*, *19*(1), 577–621. https://doi.org/10.1146/annurev.ne.19.030196.003045

Marchette, S. A., Vass, L. K., Ryan, J., & Epstein, R. A. (2015). Outside Looking In: Landmark Generalization in the Human Navigational System. *Journal of Neuroscience*, *35*(44), 14896–14908. https://doi.org/10.1523/JNEUROSCI.2270-15.2015

Martin, A., Wiggs, C. L., Ungerleider, L. G., & Haxby, J. V. (1996). Neural correlates of category-specific knowledge. *Nature*, *379*(6566), 649–652. https://doi.org/10.1038/379649a0

Millidge, B., & Shillcock, R. (2018). Human scanpaths are not Levy Flights, 8.

Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: Two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. https://doi.org/10.1016/0166-2236(83)90190-X

Mostert, P., Albers, A. M., Brinkman, L., Todorova, L., Kok, P., & de Lange, F. P. (2018). Eye Movement-Related Confounds in Neural Decoding of Visual Working Memory Representations. *eNeuro*, *5*(4). https://doi.org/10.1523/ENEURO.0401-17.2018

Nasr, S., Echavarria, C. E., & Tootell, R. B. H. (2014). Thinking Outside the Box: Rectilinear Shapes Selectively Activate Scene-Selective Cortex. *Journal of Neuroscience*, *34*(20), 6721–6735. https://doi.org/10.1523/JNEUROSCI.4802-13.2014

O'Craven, K. M., & Kanwisher, N. (2000). Mental Imagery of Faces and Places Activates Corresponding Stimulus-Specific Brain Regions. *Journal of Cognitive Neuroscience*, *12*(6), 1013–1023. https://doi.org/10.1162/08989290051137549

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

Oliva, A., & Torralba, A. (n.d.). Building the Gist of a Scene: The Role of Global Image Features in Recognition, 19.

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

Park, S., & Chun, M. M. (2009). Different roles of the parahippocampal place area (PPA) and retrosplenial cortex (RSC) in panoramic scene perception. *NeuroImage*, *47*(4), 1747–1756. https://doi.org/10.1016/j.neuroimage.2009.04.058

Parkes, L., Lund, J., Angelucci, A., Solomon, J. A., & Morgan, M. (2001). Compulsory averaging of crowded orientation signals in human vision. *Nature Neuroscience*, *4*(7), 739–744. https://doi.org/10.1038/89532

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

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

Persichetti, A. S., & Dilks, D. D. (2018). Dissociable Neural Systems for Recognizing Places and Navigating through Them. *Journal of Neuroscience*, *38*(48), 10295–10304. https://doi.org/10.1523/JNEUROSCI.1200-18.2018

Persichetti, A. S., & Dilks, D. D. (2019). Distinct representations of spatial and categorical relationships across human scene-selective cortex. *Proceedings of the National Academy of Sciences*, *116*(42), 21312–21317. https://doi.org/10.1073/pnas.1903057116

Plöchl, M., Ossandón, J. P., & König, P. (2012). Combining EEG and eye tracking: Identification, characterization, and correction of eye movement artifacts in electroencephalographic data. *Frontiers in Human Neuroscience*, *6*, 278. https://doi.org/10.3389/fnhum.2012.00278

Potter, M. C. (1975). Meaning in Visual Search. *Science*, *187*(4180), 965–966. https://doi.org/10.1126/science.1145183

Quax, S. C., Dijkstra, N., van Staveren, M. J., Bosch, S. E., & van Gerven, M. A. (2019). Eye movements explain decodability during perception and cued attention in MEG. *NeuroImage*, *195*, 444–453. https://doi.org/10.1016/j.neuroimage.2019.03.069

Quek, G. L., & Peelen, M. V. (2020). Contextual and Spatial Associations Between Objects Interactively Modulate Visual Processing. *Cerebral Cortex*, *30*(12), 6391–6404. https://doi.org/10.1093/cercor/bhaa197

Rajimehr, R., Devaney, K. J., Bilenko, N. Y., Young, J. C., & Tootell, R. B. H. (2011). The "Parahippocampal Place Area" Responds Preferentially to High Spatial Frequencies in Humans and Monkeys. *PLOS Biology*, *9*(4), e1000608. https://doi.org/10.1371/journal.pbio.1000608

Ramkumar, P., Jas, M., Pannasch, S., Hari, R., & Parkkonen, L. (2013). Feature-Specific Information Processing Precedes Concerted Activation in Human Visual Cortex. *Journal of Neuroscience*, *33*(18), 7691–7699. https://doi.org/10.1523/JNEUROSCI.3905-12.2013

Roberts, K. L., & Humphreys, G. W. (2010). Action relationships concatenate representations of separate objects in the ventral visual system. *NeuroImage*, *52*(4), 1541–1548. https://doi.org/10.1016/j.neuroimage.2010.05.044

Rolfs, M. (2009). Microsaccades: Small steps on a long way. *Vision Research*, *49*(20), 2415–2441. https://doi.org/10.1016/j.visres.2009.08.010

Rousselet, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the "gist" of real-world natural scenes? *Visual Cognition*, *12*(6), 852–877. https://doi.org/10.1080/13506280444000553

Sato, N., Nakamura, K., Nakamura, A., Sugiura, M., Ito, K., Fukuda, H., & Kawashima, R. (1999). Different time course between scene processing and face processing: A MEG study. *Neuroreport*, *10*(17), 3633–3637. https://doi.org/10.1097/00001756-199911260-00031

Schulz, E. (1984). Binocular micromovements in normal persons. *Graefe's Archive for Clinical and Experimental Ophthalmology*, *222*(2), 95–100. https://doi.org/10.1007/BF02150640

Schütz, A. C., Braun, D. I., & Gegenfurtner, K. R. (2011). Eye movements and perception: A selective review. *Journal of Vision*, *11*(5), 9. https://doi.org/10.1167/11.5.9

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

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), 4. https://doi.org/10.1167/7.14.4

Thaler, L., Schütz, A. C., Goodale, M. A., & Gegenfurtner, K. R. (2013). What is the best fixation target? The effect of target shape on stability of fixational eye movements. *Vision Research*, *76*, 31–42. https://doi.org/10.1016/j.visres.2012.10.012

Thickbroom, G. W., & Mastaglia, F. L. (1986). Presaccadic spike potential. Relation to eye movement direction. *Electroencephalography and Clinical Neurophysiology*, *64*(3), 211–214. https://doi.org/10.1016/0013-4694(86)90167-7

Thielen, J., Bosch, S. E., van Leeuwen, T. M., van Gerven, M. A. J., & van Lier, R. (2019). Evidence for confounding eye movements under attempted fixation and active viewing in cognitive neuroscience. *Scientific Reports*, *9*(1), 17456. https://doi.org/10.1038/s41598-019-54018-z

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

Tootell, R. B., Silverman, M. S., Switkes, E., & De Valois, R. L. (1982). Deoxyglucose analysis of retinotopic organization in primate striate cortex. *Science (New York, N.Y.)*, *218*(4575), 902–904. https://doi.org/10.1126/science.7134981

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

Vanrullen, R. (2009). Binding hardwired versus on-demand feature conjunctions. *Visual Cognition*, *17*(1-2), 103–119. https://doi.org/10.1080/13506280802196451

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

Walls, G. (1962). The evolutionary history of eye movements. *Vision Research*, *2*(1-4), 69–80. https://doi.org/10.1016/0042-6989(62)90064-0

Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. *Journal of Neuroscience*, *29*(34), 10573–10581. https://doi.org/10.1523/JNEUROSCI.0559-09.2009

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences*, *108*(23), 9661–9666. https://doi.org/10.1073/pnas.1015666108

Wischnewski, M., & Peelen, M. V. (2021). Causal neural mechanisms of context-based object recognition (R. G. O'Connell, J. I. Gold, R. G. O'Connell, & P. Kok, Eds.). *eLife*, *10*, e69736. https://doi.org/10.7554/eLife.69736

Wolfe, J. M. (2007). Guided Search 4.0: Current progress with a model of visual search. In *Integrated models of cognitive systems* (pp. 99–119). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195189193.003.0008

Wolfe, J. M., Vo, M. L.-H., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and non-selective pathways. *Trends in cognitive sciences*, *15*(2), 77–84. https://doi.org/10.1016/j.tics.2010.12.001

Yuval-Greenberg, S., Tomer, O., Keren, A. S., Nelken, I., & Deouell, L. Y. (2008). Transient Induced Gamma-Band Response in EEG as a Manifestation of Miniature Saccades. *Neuron*, *58*(3), 429–441. https://doi.org/10.1016/j.neuron.2008.03.027

Supplementary Material Project I

Supplementary material for Project I "Cortical sensitivity to natural scene structure".

Authors:

Daniel Kaiser, Greta Häberle, Radoslaw M. Cichy

Contributions:

D. K. and R. M. C. designed research, D. K. and G. H. acquired data, D. K. and G. H. analyzed data, D. K., G. H., and R. M. C. interpreted results, D. K. prepared figures, D. K. drafted manuscript, D. K., G. H., and R. M. C. edited and revised manuscript.

Contributions to open and reproducible science:

To contribute to open and reproducible science, the paper is published in an open-access journal. The original article can be found here: doi: 10.1002/hbm.24875. Data are publicly available on OSF: doi: 10.17605/OSF.IO/ W9874.

Copyright note:

Supplementary Information

Cortical Sensitivity to Natural Scene Structure

Daniel Kaiser, Greta Häberle, Radoslaw M. Cichy

Contents:

- Complete scene image set
- fMRI decoding – searchlight analysis
- fMRI decoding – additional ROIs
- fMRI decoding – varying voxel counts in V1
- fMRI decoding – varying voxel counts in OPA / PPA
- EEG decoding – classifying scene category

Complete scene image set

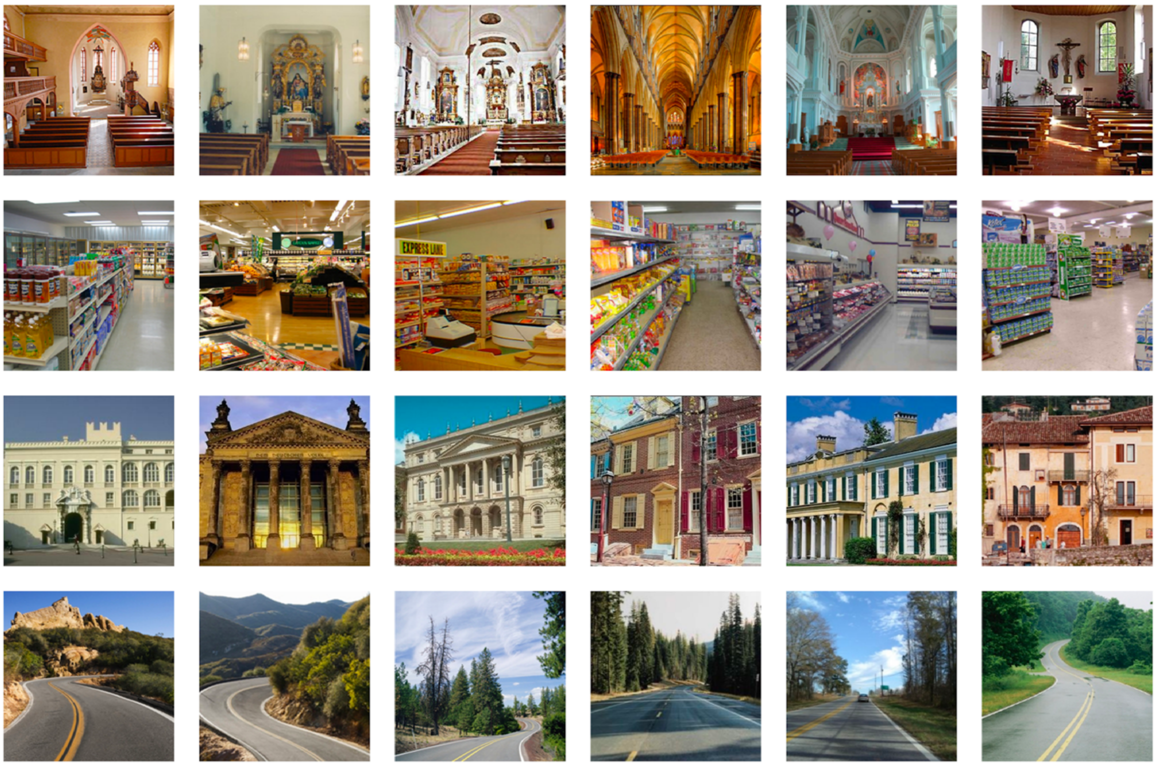

Figure S1. Scene images used in the study. The intact and jumbled scene stimuli were generated from 24 scene images from four categories: churches, supermarkets, houses, and streets. All images were chosen to depict easily recognizable scenes, photographed from a typical real-life viewpoint.

fMRI decoding – searchlight analysis

To substantiate our ROI analyses, we additionally ran a searchlight MVPA, where we probed sensitivity to spatial and categorical scene structure across the whole occipitotemporal visual cortex.

For this searchlight MVPA, we repeatedly performed the two decoding analyses (i.e., decoding spatial or categorical scene structure; see Method) for a moving sphere of 250 voxels, which was centered on every voxel within an anatomical mask of the occipital and temporal cortices (taken from WFU PickAtlas for SPM12). This procedure allowed us to map sensitivity to spatial and categorical scene structure across the whole visual cortex, separately for the upright and inverted scenes, and separately for each participant. By testing decoding against chance across participants, we computed six effects: (1) sensitivity to spatial structure for the upright scenes, (2) sensitivity to spatial structure for the inverted scenes, (3) sensitivity to categorical structure for the upright scenes, (4) sensitivity to categorical structure for the inverted scenes, (5) an inversion effect for spatial structure (i.e., the difference between (1) and (2)), and (6) an inversion effect for categorical structure (i.e., the difference between (3) and (4)). Significance was established using a threshold-free cluster enhancement procedure (as used for the EEG data). The resulting statistical maps were thresholded at z>1.96 (i.e., p_{corr}<.05).

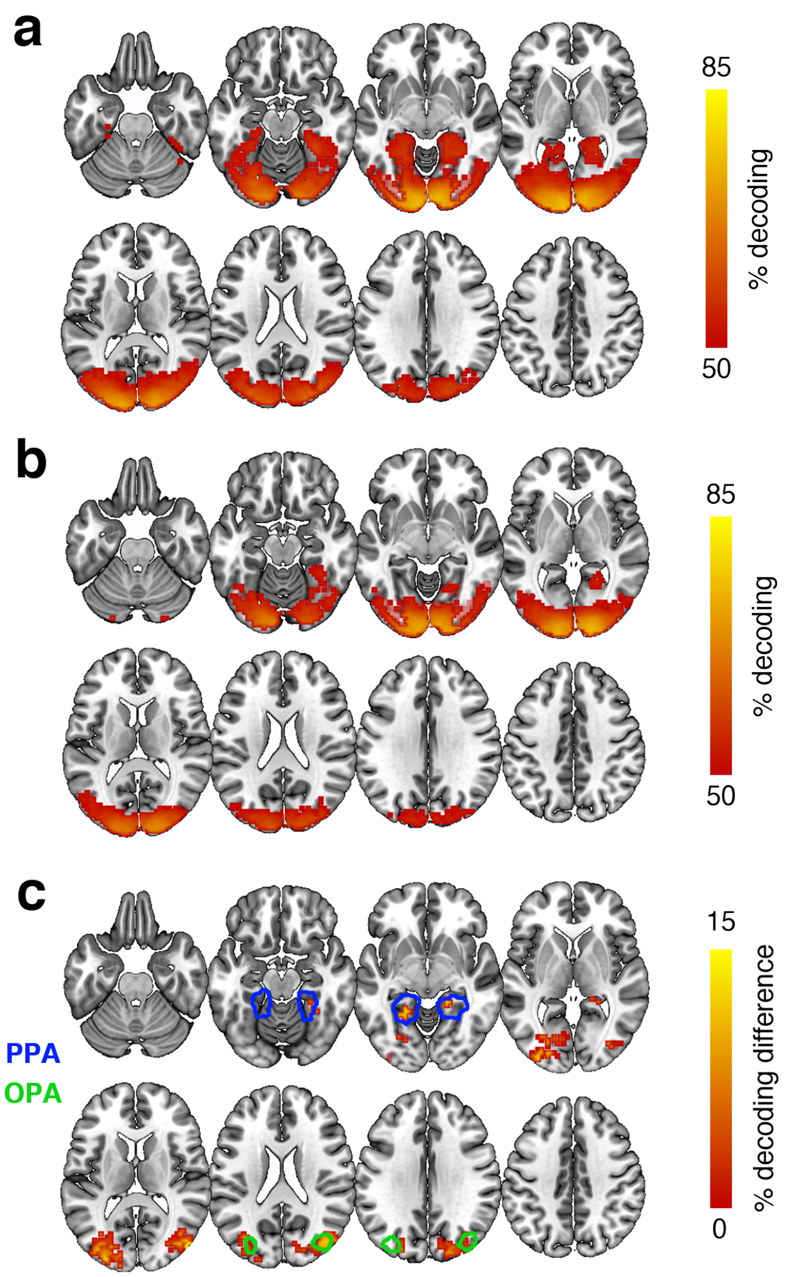

Figure S2. MVPA searchlight results in occipitotemporal cortex. Spatially intact and spatially jumbled scenes were discriminable in widespread regions of the visual cortex, both when presented upright (a) and inverted (b). Critically, when subtracting decoding in the upright and inverted conditions, we found inversion effects in regions overlapping with the typical locations of OPA and PPA (masks taken from Julian et al., 2012) (c), which indicated genuine sensitivity to spatial scene structure. Only voxels exhibiting significant effects (p_{corr}<.05) are shown.

The searchlight analysis yielded widespread significant decoding between spatially intact and jumbled upright scenes, covering early and high-level visual cortex (Figure S2a). For inverted scenes, this decoding was less pronounced (Figure S2b). Critically, we found inversion effects (i.e., better decoding for the upright, compared to the inverted scenes) in areas around the transverse occipital sulcus, corresponding to the location of OPA and areas around the parahippocampal cortex, corresponding to the location of PPA (Figure S2c). No significant inversion effects were found for categorical scene structure.

These analyses strongly support the results of our ROI-based analysis (Figure 3b/e), which revealed genuine sensitivity to spatial scene structure in the OPA and PPA.

fMRI decoding – additional ROIs

In addition to the scene-selective OPA and PPA, we also performed MVPA on responses in scene-selective retrosplenial cortex (RSC) and object-selective lateral occipital cortex (LO). These ROIs were defined similarly to OPA and PPA: For both regions, we used a functional template mask (Julian et al., 2012), and within this mask defined the voxel exhibiting the greatest *t*-value in a scene>object (RSC) or an object>scrambled (LO) contrast. Then, the ROIs were constructed as 125-voxel spheres around this peak voxel, and concatenated for the left and right hemispheres. After extracting responses from these ROIs, we performed the same decoding analyses (Figure S3a/c) as for the other ROIs.

For RSC, we did not find significant decoding between the spatially intact and spatially jumbled scenes (Figure S3b), neither in the upright, $t(19)=2.49$, $p_{corr}=.066$, nor the inverted condition, $t(19)=0.13$, $p_{corr}>1$. No significant inversion effect was observed, $t(16)=1.82$, $p_{corr}=.26$. Similarly, no significant effects were found when decoding between categorically intact and categorically jumbled scenes (Figure S3d), all $t<1.64$, $p_{corr}>.35$. These results suggest that scene structure is not represented in RSC.

For LO, we found significant decoding between spatially intact and spatially jumbled scenes (Figure S3b), both in the upright, $t(19)=4.19$, $p_{corr}=.001$, and in the inverted condition, $t(19)=3.48$, $p_{corr}=0.008$. However, no inversion effect was found, $t(16)=0.47$, $p_{corr}>1$. No significant effects were found when decoding between categorically intact and categorically jumbled scenes (Figure S3d), all $t<2.19$, $p_{corr}>.12$. These results suggest that only scene-selective regions, but not object-selective regions of the occipital cortex are genuinely sensitive to spatial scene structure.

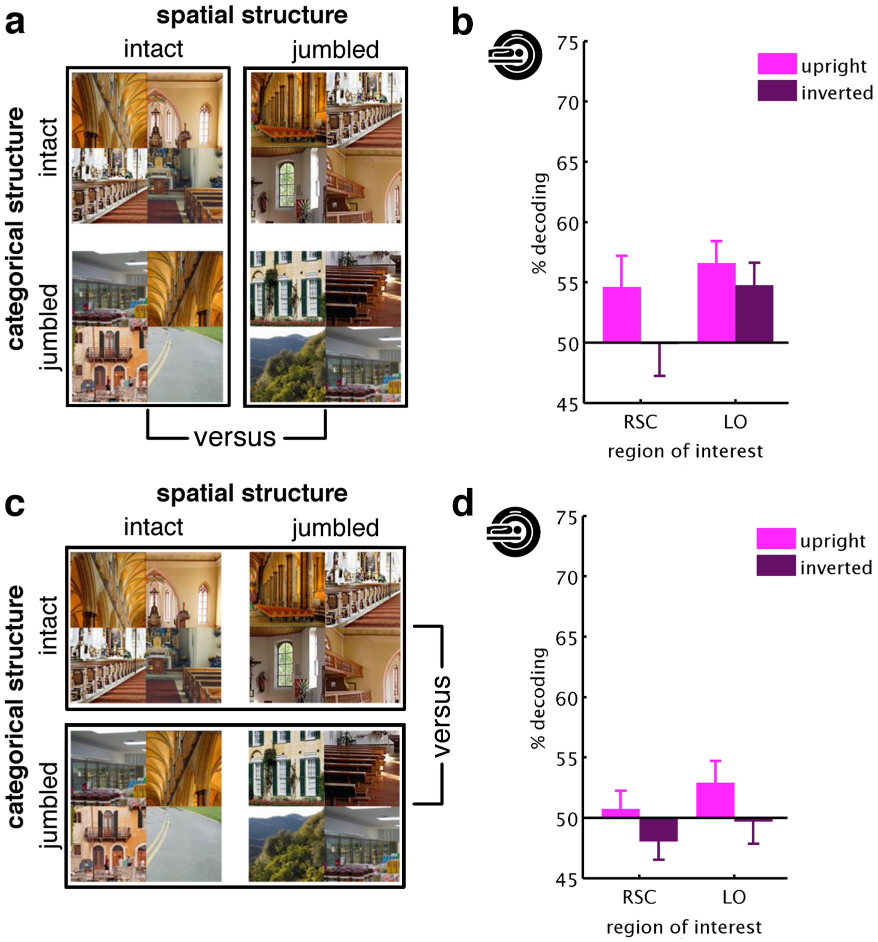

Figure S3. MVPA results in RSC and LO. To reveal sensitivity to spatial scene structure, we decoded between scenes with spatially intact and spatially jumbled parts (a). Scene-selective RSC did not show any significant decoding of spatial scene structure and no inversion effects. Object-selective LO showed significant decoding between spatially intact and spatially jumbled scenes, but no significant inversion effect (b). To reveal sensitivity to categorical scene structure, we decoded between scenes with categorically intact and categorically jumbled parts (c). In this analysis, no significant decoding and no inversion effects were found for both regions (d).

fMRI decoding – varying voxel counts in V1

To explore whether the results in V1 changed as a function of ROI size, we selected different numbers of voxels, depending on their probability to belong to V1, taken from the Wang et al. (2015) atlas. The resulting V1 sizes varied between 1032 (10% probability cutoff) and 87 voxels (60% probability cutoff).

For each of these voxel counts, we re-performed the main decoding analysis, where we decoded between (1) spatially intact and spatially jumbled scenes and (2) categorically intact and categorically scrambled scenes (Figure S4a/c).

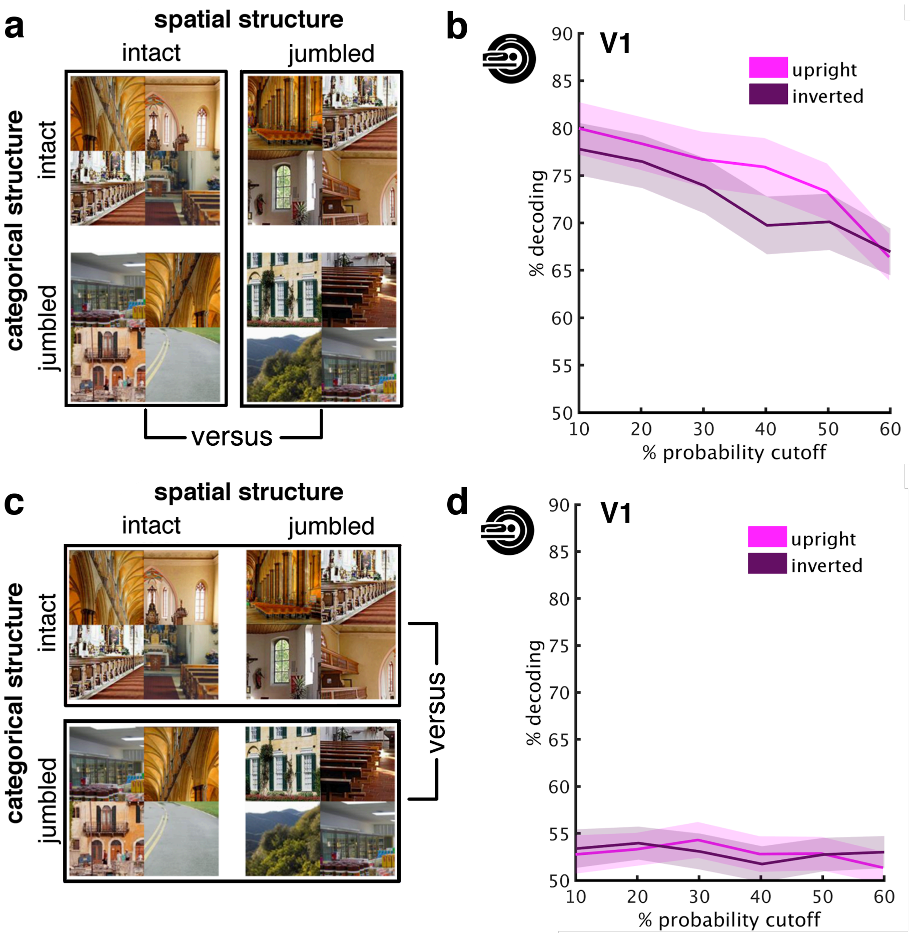

Figure S4. Results for different voxel counts in V1. Results were highly similar across different V1 sizes, ranging from voxels belonging to V1 with probabilities ≥10% (1032 voxels) to voxels belonging to V1 with probabilities ≥60% (87 voxels). This shows that

independently of the region's size, there is no reliable sensitivity to scene structure (i.e., no robust inversion effects) in early visual cortex. Error margins reflect standard errors of the difference.

Independently of the voxel counts, we found similar results as in the main analysis (Figure 3b/e). Spatial structure could be decoded reliably from V1 activations, both in the upright and inverted conditions, all $t(19)>7.32$, $p_{corr}<.001$ (Figure S4b). We observed an inversion effect only for the 40% probability cutoff, $t(16)=3.01$, $p_{corr}=.025$, but not all other cutoffs, all $t(16)<1.62$, $p_{corr}>.37$. Across the different voxel counts, we did not find a difference between the upright and inverted conditions, $F(1,16)=2.19$, $p_{corr}=.47$, suggesting no genuine inversion effects in V1. Similarly, we did not observe any significant inversion effects when looking at categorical scene structure, all $t(16)<1.07$, $p_{corr}>.90$ (Figure S4d). These results corroborate our finding that V1 does not exhibit robust sensitivity to scene structure.

fMRI decoding – varying voxel counts in OPA / PPA

To explore whether the results in OPA and PPA changed as a function of ROI size, we selected different numbers of voxels by varying the number of voxels selected around the localizer peak activation of each hemisphere (see Materials and Methods). Each ROI's size was varied between 25 voxels and 225 voxels for each hemisphere (i.e., 50 to 450 voxels for the collapsed ROI).

For each of these voxel counts, we re-performed the main decoding analysis, where we decoded between (1) spatially intact and spatially jumbled scenes and (2) categorically intact and categorically scrambled scenes (Figure S5a/c).

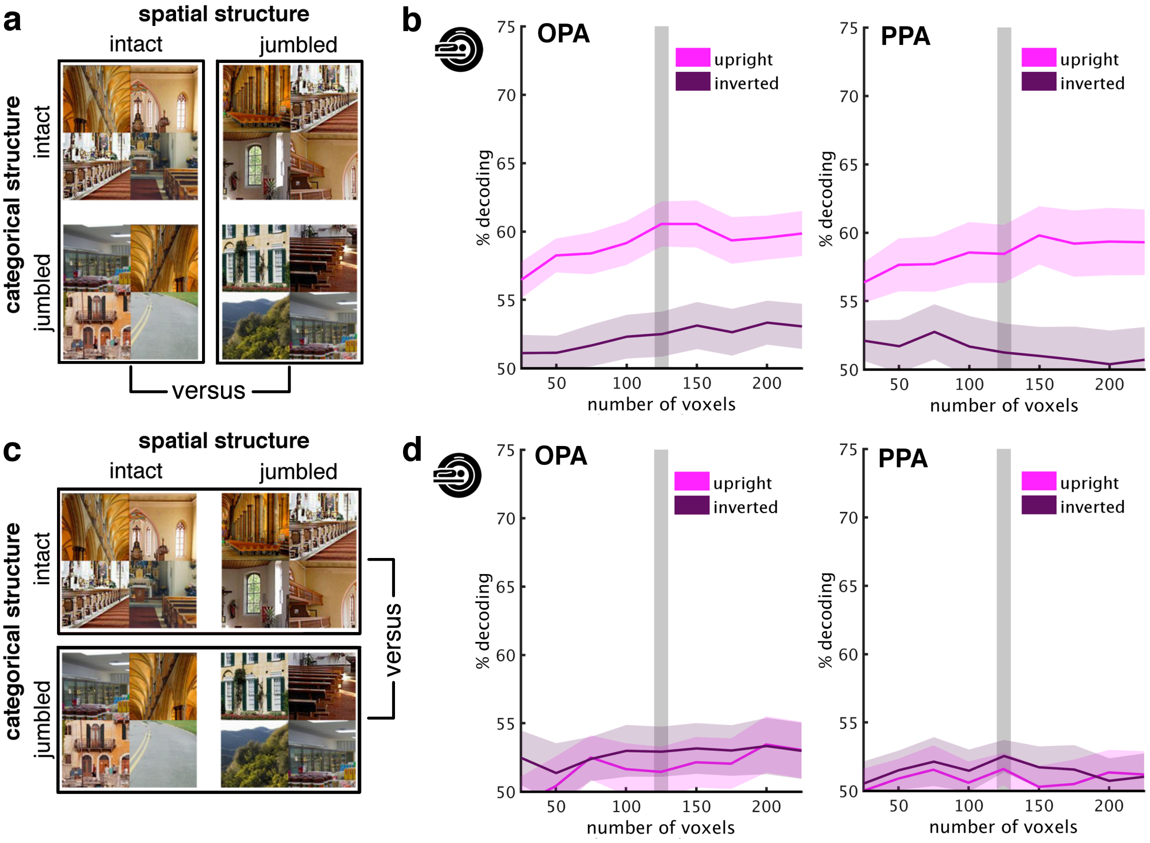

Figure S5. To explore whether the results in scene-selective cortex change as a function of ROI size, we selected the nearest 25 to 225 voxels around the individual participants' localizer peaks. Critically, the results were highly similar across the different ROI sizes, showing that the observed effect of sensitivity to spatial structure

in scene-selective cortex is not strongly dependent on the number of voxels considered part of each ROI. Error margins reflect standard errors of the difference. Shaded gray bars mark the 125-voxel spheres used in the main analysis.

We found that results were largely independent of region size. Spatial structure could be reliably decoded for upright scenes in OPA, all $t(19)>3.70$, $p_{corr}<.005$, and PPA, all $t(19)>5.19$, $p_{corr}<.003$ (Figure S5b). This decoding was significantly weaker for in the inverted condition, both in OPA, all $t(16)>3.37$, $p_{corr}<.012$, and PPA, all $t(16)>2.50$, $p_{corr}<.071$, indicating inversion effects across all voxels counts. By contrast, we did not observe any significant inversion effects when looking at categorical scene structure, neither in OPA, all $t(16)<0.61$, $p_{corr}>1$, nor in PPA, all $t(16)<0.51$, $p_{corr}>1$ (Figure S5d). These results show that our finding of robust sensitivity to spatial scene structure in scene-selective cortex cannot be attributed to the ROI definition criteria applied.

EEG decoding – classifying scene category

Our data suggest that categorically intact and categorically shuffled scenes were not represented differently during the experiments. Could this result be explained by an absence of category information from neural signals in the first place?

To investigate how well cortical representations tracked the scenes' categorical content, we performed a decoding analysis on the EEG data in which we classified scenes into the four categories used in the experiment (church, house, supermarket, street). Note that this analysis could not be performed for the fMRI experiment, where scenes of all categories were intermixed within each block of the block design.

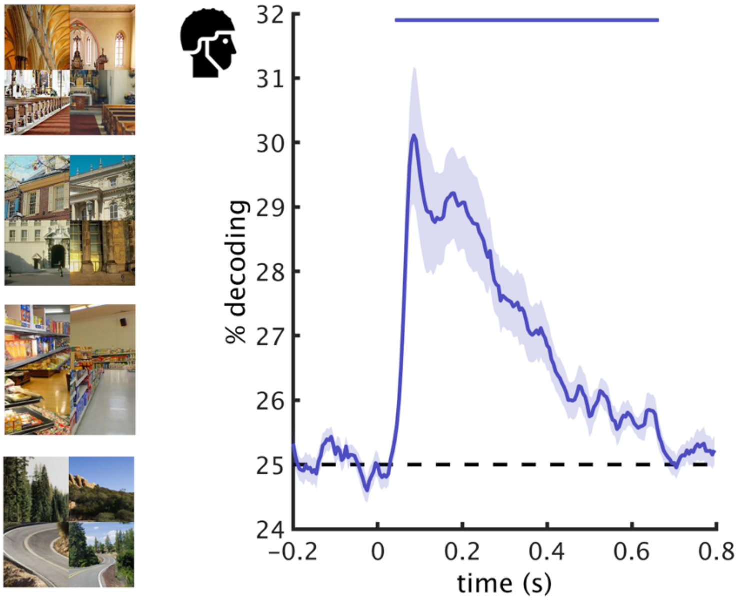

Figure S6. Decoding of scene categories from EEG signals. The four scene categories (example stimuli on the left) could be reliably decoded from the EEG data between 45ms and 660ms, showing that the neural data contained robust information about scene category. Error margins reflect standard errors of the mean. Significance markers denote above-chance decoding (p_{corr}<.05).

For decoding scene category from the EEG signals, we only used the conditions where scene category remained intact across the four scene parts. For each of these four conditions separately, we then performed a four-way decoding analysis in a leave-one-trial-out fashion (see Materials and Methods for details on the decoding procedure), and subsequently averaged across these analyses. Note that the purpose of this analysis was to show that the EEG signals contained reliable category information. Further analyses on the nature of this category information are beyond the scope of the current paper.

Across the conditions analyzed, we found that scene category information was robustly decodable between 45ms and 660ms after scene onset, peak z>3.71, p_{corr}<.001 (Figure S6). This shows that there was robust category information in the EEG signals, although across scenes there was no sensitivity to the scenes' categorical structure (i.e., whether categorical content matched within a scene or not).

Supplementary Material Project III

Supplementary material to Project III "Coherent natural scene structure facilitates the extraction of task-relevant object information in visual cortex."

Authors:

Daniel Kaiser, Greta Häberle, Radoslaw M. Cichy

Contributions:

Daniel Kaiser: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Resources, Data curation, Writing – original draft, Writing –review and editing, Visualization, Supervision, Project administration, Funding acquisition. Greta Häberle: Investigation, Writing –review and editing. Radoslaw M. Cichy: Resources, Writing –review and editing, Supervision, Project administration, Funding acquisition.

Contributions to open and reproducible science:

To contribute to open and reproducible science, the paper is published in an open-access journal. The original article can be found here: doi: 10.1016/j.neuroimage.2021.118365. Data are publicly available on OSF: doi: 10.17605/osf.io/gs2t5.

Copyright note:

According to Elsevier, the author has the right to include this article in a thesis or dissertation provided that it is not separately downloadable and the thesis is not published commercially. The article is published under a CC BY-NC-ND license.

Supplementary Information

Coherent natural scene structure facilitates the extraction of task-relevant object information in visual cortex

Daniel Kaiser, Greta Häberle, Radoslaw M. Cichy

Supplementary Contents Page

Figure S1. Univariate results – category-selective responses 2

Figure S2. Univariate results separately for both hemispheres 3

Figure S3. MVPA results separately for both hemispheres 4

Figure S4. MVPA results with alternative analysis routines 5

Table S1. Descriptive statistics – behavior 6

Table S2. Descriptive statistics – univariate analysis 7

Table S3. Descriptive statistics – MVPA 8

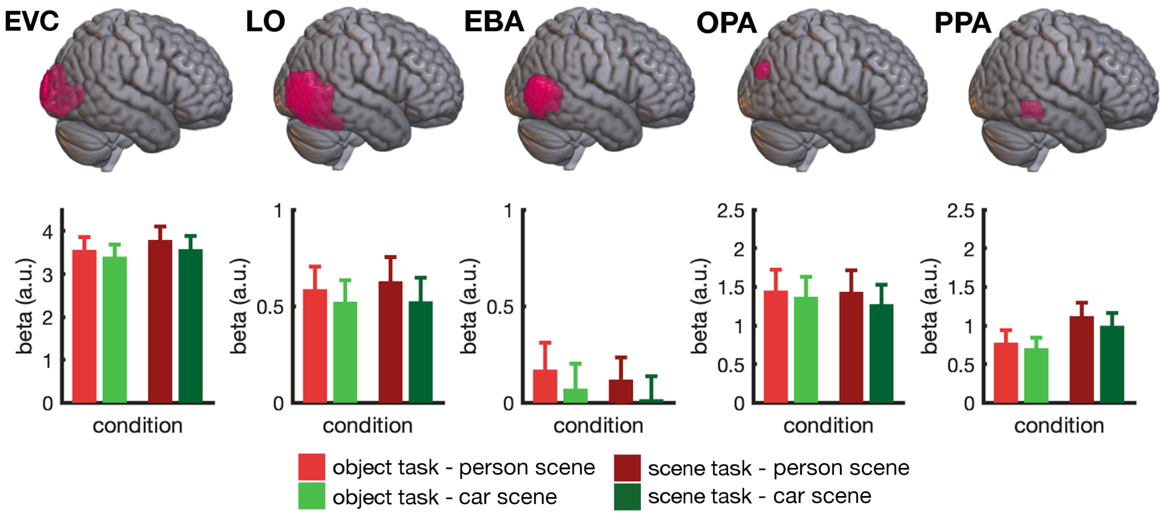

Figure S1. Univariate results – category-selective responses. *Across all regions, we found stronger responses to scenes that contained a person than to scenes that contained a car (collapsed across intact and jumbled scenes), main effect of category across ROIs, F(1,24)=8.50, p=0.008,* η_p^2*=0.62. This general bias towards person-scenes was not modulated by participants' current task, category × task interaction, F(1,24)=0.30, p=0.59,* η_p^2*=0.01. For illustration purposes, ROI masks are shown on the right hemisphere of a standard-space template using MRIcroGL (Li et al., 2016); the displayed results are averaged across ROIs in both hemispheres. Error bars represent standard errors of the mean.*

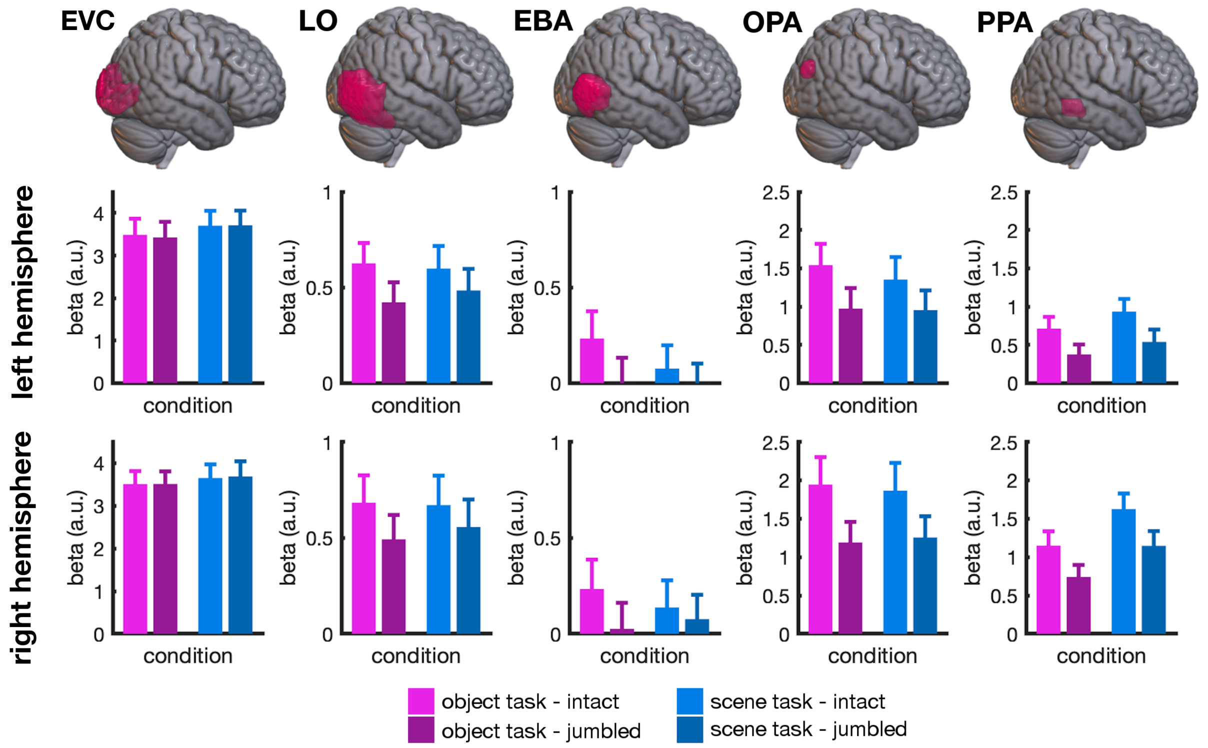

Figure S2. Univariate results separate for both hemispheres. *Results for the left and right hemispheres were highly similar and closely resembled the results across both hemispheres (Figure 2). No qualitative difference between hemispheres was found, as indicated by non-significant main effects and interactions in all ROIs. The only exception was a hemisphere × scene structure interaction in OPA, F(1,24)=6.31, p=0.019,* η_p^{2}*=0.21, with a stronger effect of scene structure in the right hemisphere. For illustration purposes, ROI masks are shown on the right hemisphere of a standard-space template using MRIcroGL (Li et al., 2016). Error bars represent standard errors of the mean.*

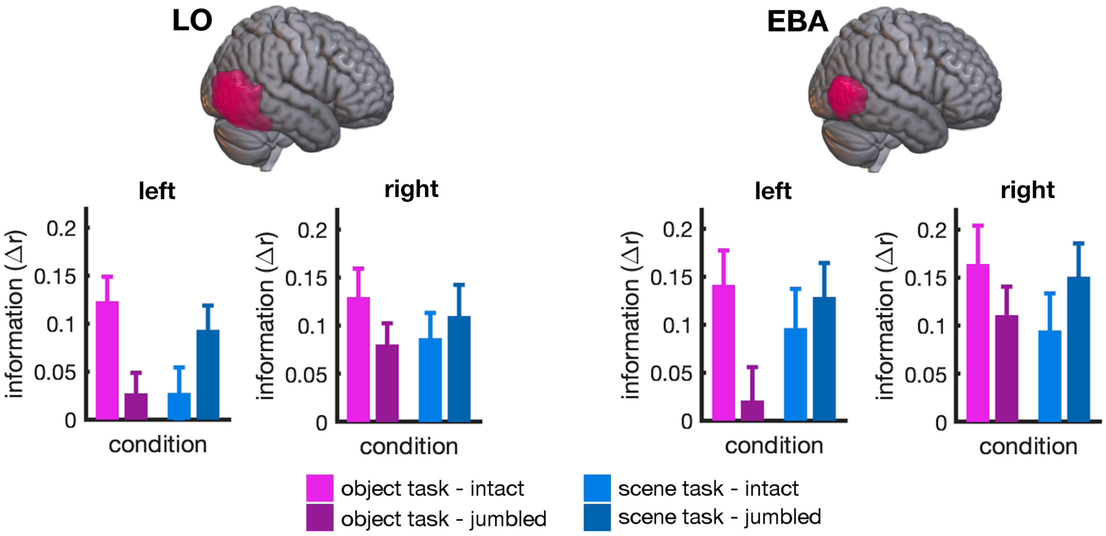

Figure S3. MVPA results separately for both hemispheres. *Results for the left and right hemispheres closely resembled the results across both hemispheres (Figure 3). No qualitative difference between hemispheres was found, all interactions with hemisphere, LO: F(1,24)<2.03, p>0.16, η_p^2<0.08, EBA: F(1,24)<2.49, p>0.12, η_p^2<0.10. In LO, category information was generally stronger in the right-hemispheric than in the left-hemispheric ROI, main effect of hemisphere, F(1,24)=5.34, p=0.030, η_p^2=0.18. For illustration purposes, ROI masks are shown on the right hemisphere of a standard-space template using MRIcroGL (Li et al., 2016). Error bars represent standard errors of the mean.*

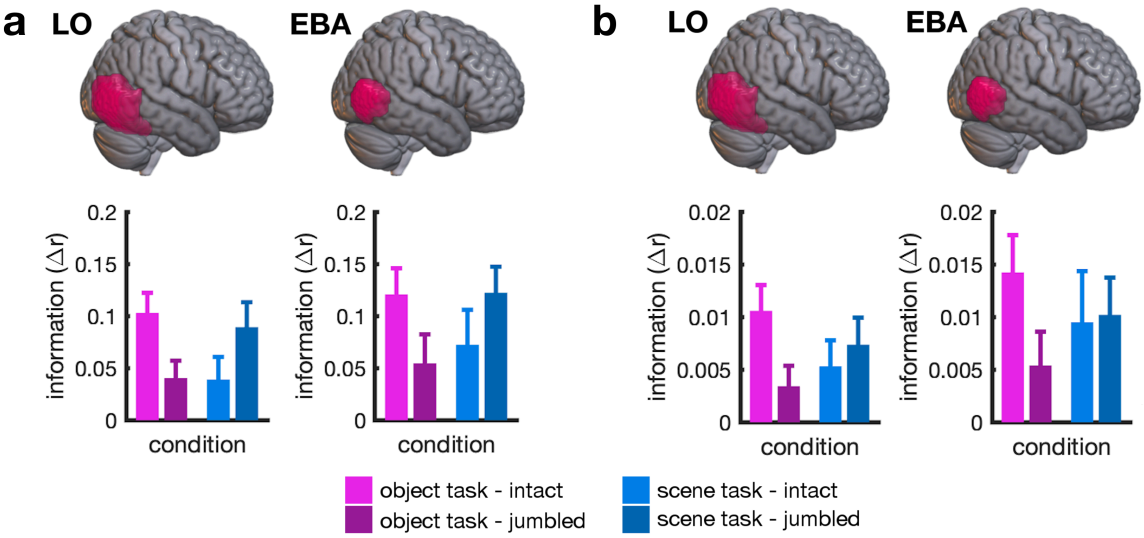

Figure S4. MVPA results with alternative analysis routines. a) *Category information in LO and EBA for intact and jumbled scenes across the two tasks, computed using Spearman correlations instead of Pearson correlations. As in the main analysis (Figure 3), an interaction between task and scene structure was observed for both regions, LO: F(1,24)=6.29, p=0.019,* η_p^2*=0.21, EBA: F(1,24)=5.10, p=0.033,* η_p^2*=0.18.* ***b)*** *Category information, as in (a), but computed without removing the voxel-wise mean activation across conditions. Again, an interaction effect was observed for LO: F(1,24)=4.48, p=0.045,* η_p^2*=0.16, but did not reach significance in EBA: F(1,24)=2.33, p=0.14,* η_p^2*=0.09. For illustration purposes, ROI masks are shown on the right hemisphere of a standard-space template using MRIcroGL (Li et al., 2016); the displayed results are averaged across ROIs in both hemispheres. Error bars represent standard errors of the mean.*

Table S1. Descriptive statistics – behavior. *Means (M) and standard errors (SE) for the accuracies (in % correct) and response times (in ms), as shown in Figure 1.*

	Object Task Intact	***Object Task Scrambled***	***Scene Task Intact***	***Scene Task Scrambled***

Accuracy	M=77.1
	SE=1.6
		M=70.5
		SE=1.5
			M=76.7
			SE=1.4
				M=73.7
				SE=1.5

Response Time	M=717
	SE=16
		M=731
		SE=17
			M=715
			SE=15
				M=736
				SE=17

Table S2. Descriptive statistics – univariate analysis. *Means (M) and standard errors (SE) for univariate activations in each ROI, as shown in Figure 2.*

ROI	***Object Task Intact***	***Object Task Scrambled***	***Scene Task Intact***	***Scene Task Scrambled***

EVC	M=3.50
	SE=0.30
		M=3.47
		SE=0.29
			M=3.68
			SE=0.30
				M=3.70
				SE=0.32

LO	M=0.65
	SE=0.12
		M=0.46
		SE=0.11
			M=0.63
			SE=0.13
				M=0.52
				SE=0.12

EBA	M=0.23
	SE=0.14
		M=0.01
		SE=0.13
			M=0.11
			SE=0.12
				M=0.03
				SE=0.11

OPA	M=1.74
	SE=0.29
		M=1.08
		SE=0.24
			M=1.61
			SE=0.30
				M=1.10
				SE=0.24

PPA	M=0.93
	SE=0.16
		M=0.56
		SE=0.13
			M=1.28
			SE=0.17
				M=0.84
				SE=0.17

Table S3. Descriptive statistics – MVPA. *Means (M) and standard errors (SE) for object discriminability (difference of within- and between-category correlations) in each ROI, as shown in Figure 3.*

ROI	***Object Task Intact***	***Object Task Scrambled***	***Scene Task Intact***	***Scene Task Scrambled***

LO	M=0.13
	SE=0.02
		M=0.05
		SE=0.02
			M=0.06
			SE=0.02
				M=0.10
				SE=0.03

EBA	M=0.15
	SE=0.03
		M=0.07
		SE=0.03
			M=0.10
			SE=0.04
				M=0.14
				SE=0.03

Appendix

List of publications

Kaiser D., **Häberle, G.**, Cichy RM. (2020a) Cortical sensitivity to natural scene structure. *Human Brain Mapping*. doi:10.1002/hbm.24875

Kaiser, D., **Häberle, G.**, Cichy, RM. (2020b). Real-world structure facilitates the rapid emergence of scene category information in visual brain signals. *Journal of Neurophysiology*. doi: 10.1152/jn.00164.2020

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

Submitted manuscripts

Häberle, G., Çelikkol, AP., Cichy RM., (2023) The influence of the bullseye versus standard fixation cross on eye movements and classifying natural images from EEG. *BioRxiv*. doi: https://doi.org/10.1101/2023.03.21.532944

Govaart, G. H., Schettino, A., Helbling, S., Mehler, D. M., Ngiam, W. X. Q., Moreau, D., Chiossi, F., Zanesco, A. P., Yang, Y.-F., Gau, R., Bartlett, J. E., Alanis, J. C. G., Gutsell, J., **Häberle, G.**, Pavlov, Y. G., Šoškić, A., Ehinger, B. V., Mouseli, P., Algermissen, J., … Paul, M. (2022). EEG ERP Preregistration Template. *MetaArXiv*. doi: 10.31222/osf.io/4nvpt

Summary of main results

When we observe a scene in our daily lives, our brains seemingly effortlessly extract various aspects of that scene. This can be attributed to different aspects of the human visual system, including but not limited to (1) its tuning to natural regularities in scenes and (2) its ability to bring different parts of the visual environment into focus via eye movements. While eye movements are a ubiquitous and natural behavior, they are considered undesirable in many highly controlled visual experiments. Participants are often instructed to fixate but cannot always suppress involuntary eye movements, which can challenge the interpretation of neuroscientific data, in particular for magneto- and electroencephalography (M/EEG).

This dissertation investigated the effect of scene structure and eye movements on the extraction of scene and object information from natural stimuli in the visual system. Projects I-III used a combination of EEG and fMRI to investigate the effect of natural scene structure on scene perception and extracting object information from natural scenes. Project IV used a combination of EEG and eye tracking to quantify how and to which extent eye movements influence the extraction of object and category information from natural stimuli. In detail, we aimed to answer four main questions: (1) Does the real-world structure impact scene-selective neural responses? (2) Does the spatial structure of a scene help facilitate the cortical analysis of the scene's categorical content? (3) Does the spatial structure of a scene's context aid in extracting task-relevant object information from the scene? (4) Does the choice of different fixation crosses influence eye movements the classification of natural images from EEG and eye tracking? In project I, we investigated the impact of spatial and categorical regularities on scene representations in healthy human adults (Kaiser et al., 2020a). Humans efficiently extract information from natural scenes (Potter, 1975; Thorpe et al., 1996). Several studies have shown that one reason for this efficiency can be found in the inherent structure of natural scenes. When this structure is interrupted, perception and categorization of these scenes are strongly impaired (Biederman, 1972; Biederman et al., 1974). However, the impact of spatial and categorical regularities on scene-selective neural responses has yet to be investigated.

Using EEG and fMRI multi- as well as univariate analyses, we tested for the cortical sensitivity to spatial and categorical structure by using two complementary analyses. To test for spatial sensitivity, we classified spatially intact from spatially jumbled scenes (irrespective of category). To test for categorical sensitivity we classified categorically intact from categorically jumbled scenes (irrespective of spatial structure). We showed that sensitivity to spatial (but not categorical) scene structure emerged after 255ms and in OPA and PPA. This effect was more substantial for upright than inverted scenes facilitating the

interpretation that this effect shows genuine sensitivity to spatial scene structure instead of just reflecting differences in low-level properties of the scenes.

To answer the second question, project II investigated whether the presence of an intact scene structure facilitates the cortical analysis of the categorical content of that scene (Kaiser et al., 2020). Previous studies have shown that the visual system is sensitive to the inherent structure of our natural world (Abassi & Papeo, 2020; Baldassano et al., 2017; Kaiser et al., 2014; Kim & Biederman, 2011; Roberts & Humphreys, 2010). However, it is still unclear how and whether this structure aids in the extraction and representation of a scene's categorical content.

To track cortical representations across time, we used a cumulative decoding approach. This approach uses a larger amount of data for decoding than standard decoding techniques by considering all time points prior to the currently decoded time point. Therefore, more data were available at each subsequent step while maintaining temporal precision in the forward direction. Consequently, this provided increased sensitivity for detecting decoding onsets compared to standard time series decoding (Ramkumar et al., 2013) Our results provide evidence that the facilitation of category information by real-world structure emerges within 200 ms of vision. While the category of the intact scenes could reliably be decoded within the first 100 ms, within 200 ms category decoding was more pronounced for the spatially intact versus the spatially jumbled scenes. In line with project I, we were able to show that this facilitation can be attributed to the adherence to the real-world structure instead of differences in low-level properties. Critically, we showed that for upright scenes the jumbling manipulation had a greater effect than for inverted scenes.

To answer the third question, project III investigated the behavioral relevance of the previously described neural findings by combining neural recordings with a more naturalistic task. In detail, we investigated whether typical real-world environments help participants to efficiently solve an object (person versus car) and a scene (rural versus urban) categorization task while recording fMRI. Using a combination of univariate and correlation-based multivariate analysis techniques, we were able to show that participants were faster and more accurate in performing the object and scene categorization task when perceiving intact versus jumbled scenes. Object information was enhanced for intact versus jumbled scenes only when the objects were relevant to the current behavioral goals. These findings revealed that early cortical tuning to the real-world structure is a crucial asset for solving complex real-world tasks (Kaiser et al., 2021).

During the recordings of projects I, II, and III, participants were instructed to fixate on a centrally presented fixation cross. To answer the fourth question, project IV sought to investigate the influence of two different fixation crosses (a bullseye versus a standard fixation cross) on eye movements and the classification of natural images from EEG. While eye movements are a ubiquitous and natural behavior, they are undesirable in many highly controlled experimental visual paradigms. Previous studies revealed that eye movements affect various analysis techniques, including MVPA (Mostert et al., 2018; Quax et al., 2019). In the combined EEG and eye tracking study, we compared the effect of two different fixation symbols – the standard fixation cross and the bullseye fixation cross – in the context of a visual paradigm with centrally presented naturalistic object images, using behavioral and multivariate analysis techniques. Our findings were threefold. First, the bullseye fixation cross reduced the number of saccades and amplitude size of microsaccades. Second, the bullseye subtly reduced classification accuracy in eye tracking and EEG data for the classification of single object images, but not for the super-level category animacy. Third, using representational similarity analysis, we found a systematic relationship between eye tracking and EEG data at the level of single images for the standard, but not for the bullseye fixation cross. These findings suggest that systematic eye movements indeed influence the results of MVPA, albeit to a small degree. Therefore, we recommend the bullseye fixation cross in experimental paradigms with fixation, particularly when control of fixation is beneficial.

In summary, projects I, II, and III aimed at answering three interconnected questions to further our understanding of scene processing. While project I showed that showed that scene-selective neural responses are sensitive to spatial scene structure, project II provided evidence that spatial structure facilitates the extraction of scene categories. Project III added a brain-behavior link by investigating whether and how spatial regularities aid object extraction from a scene while manipulating attention through an object and a scene classification task. The project results show that intact spatial structure enhances the representation of objects in a scene only if the objects are behaviorally relevant. Project IV suggest that systematic eye movements indeed influence classification results for single object images, albeit to a small degree.

Zusammenfassung der Ergebnisse

Wenn wir in unserem täglichen Leben eine Szene beobachten, extrahiert unser Gehirn scheinbar mühelos verschiedene Aspekte dieser Szene. Dies kann auf verschiedene Aspekte des menschlichen Sehsystems zurückgeführt werden, unter anderem auf (1) seine Abstimmung auf natürliche Regelmäßigkeiten in Szenen und (2) seine Fähigkeit, verschiedene Teile der visuellen Umgebung durch Augenbewegungen in den Fokus zu bringen. Obwohl Augenbewegungen ein allgegenwärtiges und natürliches Verhalten sind, werden sie in vielen stark kontrollierten visuellen Experimenten als unerwünscht angesehen. Die Teilnehmer werden oft angewiesen, zu fixieren, können aber unwillkürliche Augenbewegungen nicht immer unterdrücken, was die Interpretation neurowissenschaftlicher Daten, insbesondere der Magneto- und Elektroenzephalographie (M/EEG), in Frage stellen kann.

In dieser Dissertation wurde der Einfluss von Szenenstruktur und Augenbewegungen auf die Extraktion von Szenen- und Objektinformationen aus natürlichen Reizen im visuellen System untersucht. In den Projekten I-III wurde eine Kombination aus EEG und fMRI verwendet, um die Auswirkungen der natürlichen Szenenstruktur auf die Szenenwahrnehmung und die Extraktion von Objektinformationen aus natürlichen Szenen zu untersuchen. In Projekt IV wurde eine Kombination aus EEG und Eye Tracking eingesetzt, um zu quantifizieren, wie und in welchem Ausmaß Augenbewegungen die Extraktion von Objekt- und Kategorieinformationen aus natürlichen Reizen beeinflussen. Im Einzelnen wollten wir vier Hauptfragen beantworten: (1) Wirkt sich die Struktur der realen Welt auf szenenselektive neuronale Reaktionen aus? (2) Hilft die räumliche Struktur einer Szene dabei, die kortikale Analyse des kategorialen Inhalts der Szene zu erleichtern? (3) Hilft die räumliche Struktur des Kontextes einer Szene bei der Extraktion aufgabenrelevanter Objektinformationen aus der Szene? (4) Beeinflusst die Wahl unterschiedlicher Fixationskreuze Augenbewegungen und die Klassifikation natürlicher Bilder aus EEG und Eye-Tracking? In Projekt I untersuchten wir den Einfluss räumlicher und kategorialer Regelmäßigkeiten auf die Szenenrepräsentationen gesunder erwachsener Menschen (Kaiser et al., 2020a). Der Mensch extrahiert effizient Informationen aus natürlichen Szenen (Potter, 1975; Thorpe et al., 1996). Mehrere Studien haben gezeigt, dass ein Grund für diese Effizienz in der inhärenten Struktur natürlicher Szenen zu finden ist. Wenn diese Struktur unterbrochen wird, sind Wahrnehmung und Kategorisierung dieser Szenen stark beeinträchtigt (Biederman, 1972; Biederman et al., 1974). Der Einfluss räumlicher und kategorialer Regelmäßigkeiten auf szenenselektive neuronale Reaktionen muss jedoch noch untersucht werden.

Mithilfe von EEG- und fMRI multi- sowie univariaten Analysen testeten wir die kortikale Sensitivität für räumliche und kategoriale Strukturen mit Hilfe von zwei komplementären Analysen. Um die räumliche Sensitivität zu testen, klassifizierten wir räumlich intakte von räumlich durcheinander geworfenen Szenen (unabhängig von der Kategorie). Um die kategoriale Empfindlichkeit zu testen, haben wir kategorial intakte von kategorial durcheinander geworfenen Szenen (unabhängig von der räumlichen Struktur) unterschieden. Wir konnten zeigen, dass die Sensitivität für räumliche (aber nicht kategoriale) Szenenstrukturen in OPA und PPA und nach 255 ms auftrat. Dieser Effekt war bei aufrechten Szenen ausgeprägter als bei invertierten Szenen, was die Interpretation erleichtert, dass dieser Effekt eine echte Sensitivität für die räumliche Szenenstruktur zeigt und nicht nur Unterschiede in den niedrigen Eigenschaften der Szenen widerspiegelt.

Zur Beantwortung der zweiten Frage wurde in Projekt II untersucht, ob das Vorhandensein einer intakten Szenenstruktur die kortikale Analyse des kategorialen Inhalts dieser Szene erleichtert (Kaiser et al., 2020). Frühere Studien haben gezeigt, dass das visuelle System für die inhärente Struktur unserer natürlichen Welt empfindlich ist (Abassi & Papeo, 2020; Baldassano et al., 2017; Kaiser et al., 2014; Kim & Biederman, 2011; Roberts & Humphreys, 2010). Es ist jedoch noch unklar, wie und ob diese Struktur bei der Extraktion und Darstellung des kategorialen Inhalts einer Szene hilft.

Um die kortikalen Repräsentationen über die Zeit zu verfolgen, haben wir einen kumulativen Dekodierungsansatz verwendet. Bei diesem Ansatz wird eine größere Menge an Daten für die Dekodierung verwendet als bei Standarddekodierungstechniken, da alle Zeitpunkte vor dem aktuell dekodierten Zeitpunkt berücksichtigt werden. Daher standen bei jedem nachfolgenden Schritt mehr Daten zur Verfügung, während die zeitliche Präzision in Vorwärtsrichtung beibehalten wurde. Dies führte zu einer höheren Sensitivität bei der Erkennung von Dekodierungsanfängen im Vergleich zur standardmäßigen Zeitreihendekodierung (Ramkumar et al., 2013) Unsere Ergebnisse belegen, dass die Erleichterung der Kategorieninformation durch die Struktur der realen Welt innerhalb von 200 ms nach dem Sehen einsetzt. Während die Kategorie der intakten Szenen innerhalb der ersten 100 ms zuverlässig dekodiert werden konnte, war die Kategoriedekodierung innerhalb von 200 ms für die räumlich intakten Szenen ausgeprägter als für die räumlich durcheinandergeworfenen Szenen. In Übereinstimmung mit Projekt I konnten wir zeigen, dass diese Erleichterung auf das Festhalten an der Struktur der realen Welt zurückzuführen ist und nicht auf Unterschiede in den Eigenschaften auf niedriger Ebene. Entscheidend war, dass wir zeigen konnten, dass die Manipulation des Durcheinanders bei aufrechten Szenen einen größeren Effekt hatte als bei invertiert Szenen.

Zur Beantwortung der dritten Frage untersuchten wir in Projekt III die Verhaltensrelevanz der zuvor beschriebenen neuronalen Befunde, indem wir neuronale Ableitungen mit einer eher naturalistischen Aufgabe kombinierten. Im Einzelnen untersuchten wir, ob typische reale Umgebungen den Teilnehmern helfen, eine Objekt- (Person versus Auto) und eine Szenen-Kategorisierungsaufgabe (ländlich versus städtisch) effizient zu lösen, während wir fMRI-Aufnahmen machten. Mithilfe einer Kombination aus univariaten und korrelationsbasierten multivariaten Analysetechniken konnten wir zeigen, dass die Teilnehmer die Objekt- und Szenenkategorisierungsaufgabe schneller und genauer lösten, wenn sie intakte Szenen im Vergleich zu ungeordneten wahrnahmen. Die Objektinformation war bei intakten Szenen nur dann besser als bei durcheinandergeworfenen Szenen, wenn die Objekte für die aktuellen Verhaltensziele relevant waren. Diese Ergebnisse zeigten, dass eine frühe kortikale Abstimmung auf die Struktur der realen Welt ein entscheidender Vorteil für das Lösen komplexer Aufgaben in der realen Welt ist (Kaiser et al., 2021).

Während der Aufnahmen der Projekte I, II und III wurden die Teilnehmer angewiesen, auf ein zentral präsentiertes Fixationskreuz zu fixieren. Um die vierte Frage zu beantworten, wurde in Projekt IV der Einfluss von zwei verschiedenen Fixationskreuzen (ein Bullauge und ein Standard-Fixationskreuz) auf die Augenbewegungen und die Klassifizierung natürlicher Bilder aus dem EEG untersucht. Obwohl Augenbewegungen ein allgegenwärtiges und natürliches Verhalten sind, sind sie in vielen stark kontrollierten experimentellen visuellen Paradigmen unerwünscht. Frühere Studien haben gezeigt, dass Augenbewegungen verschiedene Analysetechniken beeinträchtigen, darunter MVPA (Mostert et al., 2018; Quax et al., 2019). In der kombinierten EEG- und Eye-Tracking-Studie verglichen wir die Wirkung von zwei verschiedenen Fixationssymbolen - dem Standard-Fixationskreuz und dem Bullseye-Fixationskreuz - im Rahmen eines visuellen Paradigmas mit zentral präsentierten naturalistischen Objektbildern, wobei wir verhaltensbasierte und multivariate Analysetechniken verwendeten. Unsere Ergebnisse waren dreigeteilt. Erstens reduzierte das Bullseye-Fixationskreuz die Anzahl der Sakkaden und die Amplitudengröße der Mikrosakkaden. Zweitens verringerte das Bullauge die Klassifizierungsgenauigkeit in den Eye-Tracking- und EEG-Daten für die Klassifizierung von Einzelobjekten, nicht aber für die übergeordnete Kategorie der Lebendigkeit. Drittens fanden wir mit Hilfe einer repräsentativen Ähnlichkeitsanalyse eine systematische Beziehung zwischen Eye-Tracking- und EEG-Daten auf der Ebene der Einzelbilder für das Standard-, nicht aber für das Bullseye-Fixationskreuz. Diese Ergebnisse deuten darauf hin, dass systematische Augenbewegungen tatsächlich die Ergebnisse der MVPA beeinflussen, wenn auch nur in geringem Maße. Daher empfehlen wir das Bullseye-Fixationskreuz in experimentellen Paradigmen mit Fixation, insbesondere wenn die Kontrolle der Fixation von Vorteil ist.

Zusammenfassend lässt sich sagen, dass die Projekte I, II und III darauf abzielten, drei miteinander verknüpfte Fragen zu beantworten, um unser Verständnis der Szenenverarbeitung zu erweitern. Während Projekt I zeigte, dass szenenselektive neuronale Reaktionen empfindlich auf die räumliche Szenenstruktur reagieren, lieferte Projekt II Beweise dafür, dass die räumliche Struktur die Extraktion von Szenekategorien erleichtert. Projekt III stellte eine Verbindung zwischen Gehirn und Verhalten her, indem es untersuchte, ob und wie räumliche Regelmäßigkeiten die Objektextraktion aus einer Szene unterstützen, während die Aufmerksamkeit durch eine Objekt- und eine Szenenklassifikationsaufgabe manipuliert wurde. Die Projektergebnisse zeigen, dass eine intakte räumliche Struktur die Darstellung von Objekten in einer Szene nur dann verbessert, wenn die Objekte verhaltensrelevant sind. Projekt IV deutet darauf hin, dass systematische Augenbewegungen tatsächlich die Klassifikationsergebnisse für einzelne Objektbilder beeinflussen, wenn auch nur in geringem Maße.

Author contributions

Declaration pursuant to Sec. 7 (3), fourth sentence, of the Doctoral Study Regulations regarding my own share of the submitted scientific or scholarly work that has been published or is intended for publication within the scope of my publication-based work

I. Last name, first name: Häberle, Greta
Institute: Department of Education and Psychology, Freie Universität Berlin
Doctoral study subject: Psychology
Title: Extracting scene and object information from natural stimuli: the influence of scene structure and eye movements

II. Numbered listing of works submitted (title, authors, where and when published and/or submitted):

1. Kaiser, D., Häberle, G., Cichy, R.M., (2020a). Cortical sensitivity to natural scene structure. Published in Human Brain Mapping

2. Kaiser, D., Häberle, G., Cichy, R.M., (2020b). Real-world structure facilitates the rapid emergence of scene category information in visual brain signals. Published in Journal of Neurophysiology

3. Kaiser, D., Häberle, G., & Cichy, R.M. (2021). Coherent natural scene structure facilitates the extraction of task-relevant object information in visual cortex. Published in NeuroImage

4. Häberle, G., Çelikkol, A.P., Cichy, R.M., (2023). The influence of the bullseye versus standard fixation cross on eye movements and classifying natural images from EEG. Submitted to Scientific Reports March 2023, uploaded on Bioarxiv.

III. Explanation of own share of these works:

Regarding II. 1.: programming of paradigm (part), data collection (part), data analysis (part), interpretation of results (part), revising and editing of manuscript (part)

Regarding II. 2.: programming of paradigm (all), data collection (vast majority), data analysis (part), interpretation of results (part), revising and editing of manuscript (part)

Regarding II. 3.: data collection (most), revising and editing of manuscript (part)

Regarding II. 4.: Study conceptualisation (vast majority), design (all), programming of paradigm (vast majority), data collection (vast majority), data analysis (all), discussion of results (vast majority), writing of original draft (all), revising and editing the manuscript (most)

IV. Names, addresses, and e-mail addresses or fax numbers for the relevant co-authors:

Regarding II. 1.: Kaiser, D., (1) ,

Häberle, G., (2),

Cichy, R. M., (2),

Regarding II. 2.: Kaiser, D., (1) ,

Häberle, G., (2),

Cichy, R. M., (2),

Regarding II. 3.: Kaiser, D., (1) ,

Häberle, G., (2),

Cichy, R. M., (2),

Regarding II. 4.: Häberle, G., (1),

Çelikkol, A.P., (3),

Cichy, R. M., (1),

(1) Neural Dynamics of Visual Cognition Lab
Department of Education and Psychology Freie Universität Berlin
Habelschwerdter Allee 45
14195 Berlin

(2) Mathematical Institute
Justus-Liebig-University Gießen
Arndtstraße 2
35392 Gießen

(3) Research Focus Cognitive Sciences
University of Potsdam
Campus Golm
Karl-Liebknecht-Str. 24-25
14469 Potsdam

Date, doctoral candidate signature .... ……………

The information in III. must be confirmed in writing by the co-authors. I confirm the declaration made by Greta Häberle under III.:

Name: Radoslaw Martin Cichy Signature: ....

Name: Daniel Kaiser Signature: ....

Name: Aynur Pelin Ćelikkol Signature: ....

Selbstständigkeitserklärung/Affidavit

Hiermit versichere ich,

- dass ich die vorliegende Arbeit eigenständig und ohne unerlaubte Hilfe verfasst habe,
- dass Ideen und Gedanken aus Arbeiten anderer entsprechend gekennzeichnet wurden,
- dass ich mich nicht bereits anderwärtig um einen Doktorgrad beworben habe und keinen Doktorgrad in dem Promotionsfach Psychologie besitze, sowie
- dass ich die zugrundeliegende Promotionsordnung vom 08.08.2016 anerkenne.

Berlin, 24. März 2023

Berlin, 24th March 2023 Greta Häberle