

Location representations of objects in cluttered scenes in the human brain

Dissertation

zur Erlangung des akademischen Grades
Doktorin der Naturwissenschaften (Dr. rer. nat.)

am Fachbereich Erziehungswissenschaft und Psychologie
der Freien Universität Berlin



vorgelegt von
Monika Graumann, M.Sc.

Berlin, 2022

Gutachter:

1. Prof. Dr. Radoslaw M. Cichy, Freie Universität Berlin
2. Prof. Dr. John-Dylan Haynes, Charité - Universitätsmedizin Berlin
3. Prof. Dr. Angelika Lingnau, Universität Regensburg

Datum der Disputation: 22. Februar 2023

Acknowledgements

First of all, I would like to thank my supervisor Prof. Radoslaw Cichy for his excellent guidance, availability and for providing me with plenty of opportunities to learn and grow as a scientist and to pursue my ideas. I thank Prof. John-Dylan Haynes for helpful input on my projects and advice on my work over the years. I also thank Prof. Annette Kinder, Prof. Liuba Papeo and Dr. Marleen Haupt for kindly agreeing to join my doctoral committee.

I am grateful for the friendly help from Christian Kainz regarding fMRI sessions and Daniela Satici-Thies with administrative questions.

A big thank you goes out to my colleagues Siying, Marleen and Kshitij for helping me with my research and for being great colleagues. Very special thanks go to Akritee, Lena, Gina, Nathan, Miriam, Sofie, Işil and Noor for their invaluable support.

May, 2022

Contents

Acknowledgements	iii
Abstract	vi
Zusammenfassung	vii
List of abbreviations	viii
List of original research articles	ix
1 Introduction	1
1.1 Divergent cortical loci of object location representations	2
1.2 The neural temporal dynamics of object recognition in realistic environments . .	6
1.3 Attentional modulation of object location representations	7
1.4 Research questions and hypotheses	9
1.5 Methods overview	10
2 Summary of Experiments	14
2.1 Study 1: The spatiotemporal neural dynamics of object location representations in the human brain	14
2.1.1 Experiment 1: fMRI and DNN experiment	14
2.1.2 Experiment 2: EEG experiment	15
2.1.3 Classification of object category in space and time	16
2.1.4 Contributions to open and reproducible science	16
2.2 Study 2: Independent spatiotemporal effects of spatial attention and background clutter on human object location representations	17
2.2.1 Experiment 1: EEG experiment	17
2.2.2 Experiment 2: fMRI experiment	18
2.2.3 Contributions to open and reproducible science	19
3 General Discussion	20
3.1 Summary	20
3.2 Location representations of objects with clutter emerge along the ventral visual stream	20
3.3 Location representations of objects with clutter emerge during the recurrent pro- cessing stage	23
3.4 Late attentional modulation of object location representations	25
3.5 Possible mechanisms of location encoding in the ventral stream	26
3.6 The role of location representations for object recognition	27
3.7 Limitations	29
3.8 Methodological strengths	31
3.9 Future directions	32
3.10 Conclusion	33
References	34
Appendix	47
Original publication of Study 1	48
Preprint of Study 2	69
Author contributions	107

Eidesstattliche Erklärung 111

Abstract

When we perceive a visual scene, we usually see an arrangement of multiple cluttered and partly overlapping objects, like a park with trees and people in it. Spatial attention helps us to prioritize relevant portions of such scenes to efficiently interact with our environments. In previous experiments on object recognition, objects were often presented in isolation, and these studies found that the location of objects is encoded early in time (before ~ 150 ms) and in early visual cortex or in the dorsal stream. However, in real life objects rarely appear in isolation but are instead embedded in cluttered scenes. Encoding the location of an object in clutter might require fundamentally different neural computations. Therefore this dissertation addressed the question of how location representations of objects on cluttered backgrounds are encoded in the human brain. To answer this question, we investigated *where* in cortical space and *when* in neural processing time location representations emerge when objects are presented on cluttered backgrounds and which role spatial attention plays for the encoding of object location. We addressed these questions in two studies, both including fMRI and EEG experiments. The results of the first study showed that location representations of objects on cluttered backgrounds emerge along the ventral visual stream, peaking in region LOC with a temporal delay that was linked to recurrent processing. The second study showed that spatial attention modulated those location representations in mid- and high-level regions along the ventral stream and late in time (after ~ 150 ms), independently of whether backgrounds were cluttered or not. These findings show that location representations emerge during late stages of processing both in cortical space and in neural processing time when objects are presented on cluttered backgrounds and that they are enhanced by spatial attention. Our results provide a new perspective on visual information processing in the ventral visual stream and on the temporal dynamics of location processing. Finally, we discuss how shared neural substrates of location and category representations in the brain might improve object recognition for real-world vision.

Zusammenfassung

Wenn wir uns umschauen, sehen wir Objekte, die von weiteren Objekten umgeben sind. Frühere Forschung zu Objekterkennung zeigte Objekte auf leeren Hintergründen und fand heraus, dass die Position von Objekten früh im Gehirn verarbeitet wird (vor ~ 150 ms) und, dass die Verarbeitung in frühen Arealen der visuellen Sehrinde oder dem dorsalen Pfad stattfand. Allerdings erscheinen Objekte in der realen Welt selten auf leeren Hintergründen und die Objektposition könnte im Gehirn fundamental anders verarbeitet werden, wenn Objekte von vielen anderen Objekten umgeben sind. Daraus leitet sich die Frage ab, wie die Position von Objekten im Gehirn repräsentiert wird, wenn Objekte von anderen Objekten umgeben sind. Um diese Frage zu beantworten, haben wir untersucht wo im Kortex und wann während der neuronalen Verarbeitungszeit Repräsentationen der Objektposition entstehen, wenn Objekte von anderen Objekten umgeben sind. Außerdem untersuchten wir, wie diese Repräsentationen von räumlicher Aufmerksamkeit beeinflusst werden. Diese Fragen beantworteten wir in zwei Studien mit je einem fMRT- und einem EEG-Experiment. Die Resultate der ersten Studie zeigten, dass Objektpositionsrepräsentationen im ventralen Pfad verarbeitet werden, wenn Objekte von anderen Objekten im Hintergrund umgeben sind und, dass diese Repräsentationen am höchsten im Areal LOC sind. Die Repräsentationen benötigten eine längere Verarbeitungszeit aufgrund von rekurrenten Verarbeitungsschritten. Die zweite Studie zeigte, dass diese Repräsentationen in mittleren und höheren Arealen des ventralen Pfads und während späterer Verarbeitungszeiten (nach ~ 150 ms) von räumlicher Aufmerksamkeit verstärkt werden. Dies war unabhängig davon, ob Objekte auf leeren oder mit Objekten gefüllten Hintergründen gezeigt wurden. Diese Resultate zeigen, dass Objektpositionsrepräsentationen in höheren Arealen im ventralen Pfad und spät in der neuronalen Verarbeitungszeit verarbeitet werden, wenn Objekte von mehr Objekten umgeben sind und, dass diese Repräsentationen von Aufmerksamkeit verstärkt werden. Unsere Resultate bieten eine neue Perspektive auf etablierte Theorien visueller Verarbeitung im ventralen Pfad und auf die zeitliche Dynamik von Positionsverarbeitung im Gehirn. Am Ende dieser Dissertation werden mögliche Vorteile der gemeinsamen Verarbeitung von Objektkategorie und -position im gleichen kortikalen Pfad für Objekterkennung besprochen.

List of abbreviations

DNN – deep neural network

EEG - electroencephalography

EVC – early visual cortex (V1, V2, V3)

fMRI – functional magnetic resonance imaging

IPS – intraparietal sulcus

IT – inferior temporal cortex

LOC – lateral occipital complex

pRF – population receptive field

RDM – representational dissimilarity matrix

RF – receptive field

RSA – representational similarity analysis

SPL – superior parietal lobule

SVM – support vector machine

TGA – time-generalization analysis

List of original research articles

Study 1

Graumann, M., Ciuffi, C., Dwivedi, K., Roig, G. & Cichy, R. M. (2022). The spatiotemporal neural dynamics of object location representations in the human brain. *Nature Human Behaviour*. doi: [10.1038/s41562-022-01302-0](https://doi.org/10.1038/s41562-022-01302-0).

Study 2

Graumann, M., Wallenwein, L. A., & Cichy, R. M. (submitted). Independent spatiotemporal effects of spatial attention and background clutter on human object location representations. *bioRxiv*. doi: [10.1101/2022.05.02.490141](https://doi.org/10.1101/2022.05.02.490141) .

1 Introduction

When we observe the visual world around us, we effortlessly and automatically perceive a large number of objects arranged into a visual scene: for example, during a walk in the park, we see trees, benches, flowers, people and dogs. To localize individual objects within a scene, our visual system needs to group single object parts into one object, assign overlapping objects to separate entities and dissect objects from the background. During these processes, spatial attention helps us to prioritize certain portions of the visual field while ignoring others to allocate resources to the relevant parts of the scene (Desimone & Duncan, 1995). This thesis addresses the question what the neural mechanisms are that allow us to localize an individual object within multiple background objects in the visual world.

The neural computations needed to perceive objects in cluttered scenes occur so quickly, that we usually do not even realize they are happening. The fragility of these complex brain mechanisms only becomes evident once they fail. After suffering a stroke in the occipital and ventral temporal lobes, a neuropsychological patient lost his ability to group object parts into individual objects and had in particular problems with identifying overlapping and cluttered objects (Humphreys & Riddoch, 1994; Riddoch & Humphreys, 1987; Humphreys & Riddoch, 1987). This neuropsychological disorder called integrative agnosia, has been observed in other patients with similar ventral stream lesions (Behrmann et al., 1994). The failure of the perception of cluttered objects demonstrates that there are distinct neural substrates that support this mechanism and that lesioning these substrates leads to the failure of the perception of objects in clutter.

Studying the neural mechanisms of location encoding in the human brain is important because, besides category, the location of an object is arguably one of the most fundamental object properties that we need to interact with objects in daily life (Groen et al., 2022; Malcolm et al., 2016). However, to date it is still unclear *where* and *when* the location of objects is encoded in the brain and which role spatial attention plays for location encoding. In this thesis, we took the background of objects into account to systematically uncover the mechanisms underlying object location encoding in the human brain.

Decades of neuroscientific research on visual object recognition have established a comparatively clear and comprehensive picture of the steps involved in how human brains form representations of isolated objects along the ventral visual stream (DiCarlo & Cox, 2007; DiCarlo

et al., 2012; Goodale & Milner, 1992; Mishkin et al., 1983). This research has established the basis for subsequent work on more complex representations of objects, e.g. of objects that are partially occluded (Kar et al., 2019; Rajaei et al., 2019; Spoerer et al., 2017; Tang et al., 2014) or objects in complex real-world scenes (Brandman & Peelen, 2017, 2019; Groen et al., 2018; Kaiser et al., 2016, 2019; Peelen & Kastner, 2014; Seijdel et al., 2020). However, this research has focused on the representation of object category (e.g. car, animal). Previous studies studying location representations typically presented objects in isolation (Carlson, Hogendoorn, Fonteijn, & Verstraten, 2011; Carlson, Hogendoorn, Kanai, et al., 2011; Cichy et al., 2011, 2013; Kay et al., 2015). Therefore it is still an open question how the location of an object is represented in the brain under more realistic circumstances, such as when the object’s visual surroundings are cluttered.

Addressing this open question is crucial for understanding object location encoding in the human brain because in the real world, objects rarely appear in isolation. Perceiving an object in isolation does not require the grouping and segmentation operations which are needed to segregate an object from clutter (Poort et al., 2016; Scholte et al., 2008; Seijdel et al., 2021), suggesting that the neural mechanisms supporting these two cases are fundamentally different (Groen et al., 2018; Hong et al., 2016; Li et al., 2009; Reddy & Kanwisher, 2007; Seijdel et al., 2021). In isolation, objects automatically pop-out, which means that the object’s location should be processed in a bottom-up manner (Itti & Koch, 2001). In contrast, with clutter, top-down spatial attention might be beneficial to perceive the object. Therefore, the present dissertation investigated the neural mechanisms of *where* and *when* the location of objects in clutter is encoded in the human brain and which role spatial attentional modulation plays for location encoding.

1.1 Divergent cortical loci of object location representations

Unlike for location, there is a clear consensus on where in the brain object category representations emerge. Longstanding research has established that object category representations emerge in a succession of hierarchical transformations that occur in sequential stages along the ventral visual stream (Cichy, Khosla, et al., 2016; DiCarlo & Cox, 2007; DiCarlo et al., 2012; Goodale & Milner, 1992; Mishkin et al., 1983). In a first step, early visual regions V1, V2 and V3 process low-level features like line orientations and retinotopic location (Hubel & Wiesel, 1959, 1977; Wandell et al., 2007; Wandell & Winawer, 2015). Subsequently, higher-order shape descriptors and colour constancy are encoded in V4 (Pasupathy & Connor, 2002; Zeki & Marini, 1998). Fi-

nally, object representations emerge in the lateral occipital complex (LOC) (Grill-Spector et al., 2001; Malach et al., 1995) or in its primate object-selective equivalent inferior temporal cortex (IT) (Cichy et al., 2014; Hung, 2005; Kiani et al., 2007; Kriegeskorte, Mur, Ruff, et al., 2008). Category representations in these regions have reached a level of abstraction in which they are invariant to transformations like the angle from which an object is perceived, its size on the retina, clutter in the background of the object and its location in the visual field (Carlson, Hogenboom, Fonteijn, & Verstraten, 2011; Cichy et al., 2011; DiCarlo & Cox, 2007; Li et al., 2009; Schwarzlose et al., 2008).

When looking at how the location of an object is represented in the brain, the picture that emerges from previous literature is more heterogeneous. Therefore, the first research question of this dissertation was *where* in cortical space the location of objects is represented in the brain when objects are presented under more realistic circumstances. Overall, previous studies linked object location representations to three main brain areas: studies finding location representations in early visual cortex (EVC) (Cichy et al., 2013; Golomb & Kanwisher, 2012; Wandell & Winawer, 2015), in the dorsal stream (Kravitz et al., 2011; Ungerleider & Haxby, 1994; Zachariou et al., 2015) or in the ventral stream (Cichy et al., 2011, 2013; Golomb & Kanwisher, 2012; Hong et al., 2016; Schwarzlose et al., 2008; Xu & Vaziri-Pashkam, 2021). Here, I will present three testable hypotheses derived from these studies. Then I will explain the relevance of comparing representations of objects on cluttered and isolated backgrounds to distinguish between these hypotheses.

The first hypothesis posits that location representations are encoded in EVC (Fig. 1a, c, H1). This view is derived from seminal studies, showing that visual space can be mapped out retinotopically in V1 (Engel et al., 1994; Holmes, 1918; Tootell et al., 1988). For example, an early lesion study showed that the visual field is flipped horizontally and vertically in V1, thus representing visual images reversed and inverted on its cortical surface (Holmes, 1918). Furthermore, V1 contains cells that selectively fire when corresponding regions of the visual field are stimulated (Hubel & Wiesel, 1977; Wandell et al., 2007), demonstrating its retinotopic organization. Since receptive fields (RF) in V1 are small compared to RF in higher-level brain regions, V1 maps out visual space with high spatial resolution (Wandell & Winawer, 2015). More recently, fMRI studies showed that location information of isolated objects was highest in EVC (Cichy et al., 2013; Golomb & Kanwisher, 2012), supporting the view that EVC is the main locus of location representations in the brain.

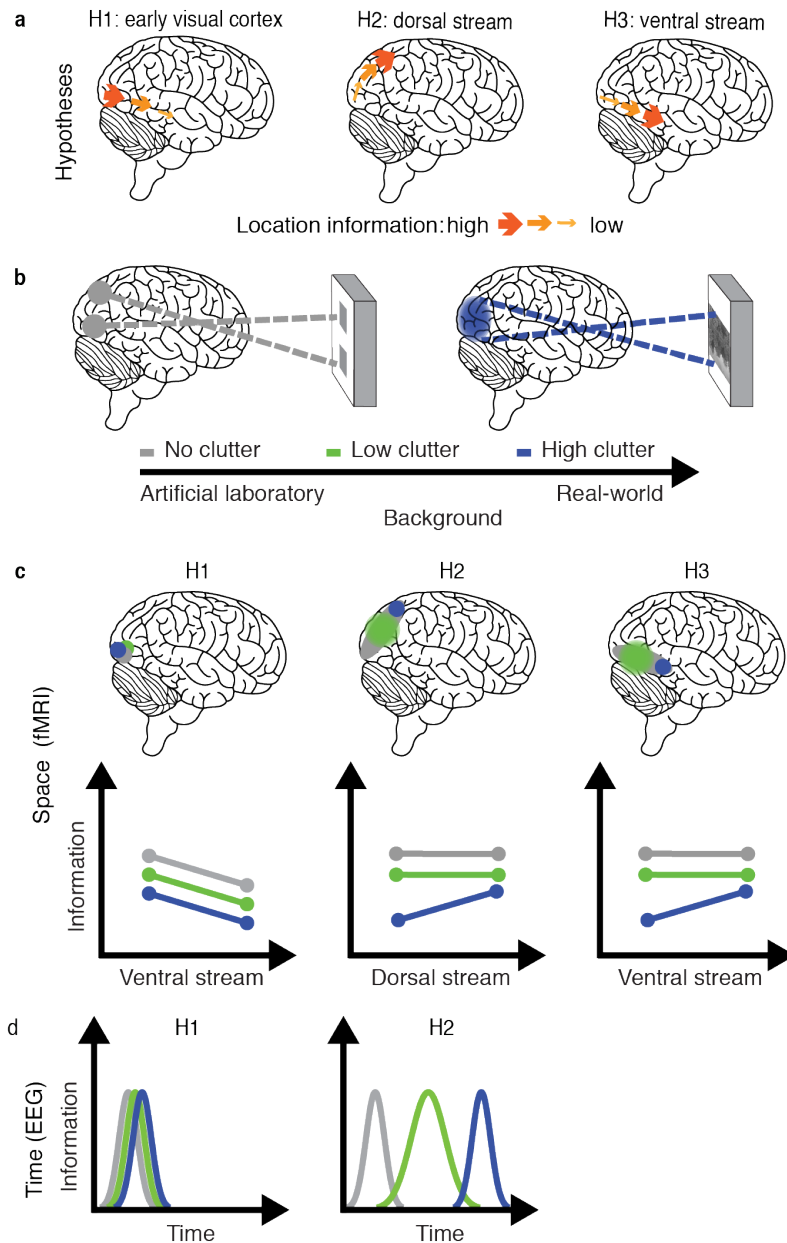


Figure 1: Hypotheses and predictions in study 1. **a**, The three hypotheses predict that location representations emerge in EVC (H1), along the dorsal stream (H2) or along the ventral stream (H3). **b**, When objects appear on blank backgrounds, the stimulation in the visual field directly maps onto stimulated portions of EVC, allowing for direct read-out of the location. With clutter, this is not possible because large parts of EVC are activated. **c**, fMRI predictions. H1 predicts that location information peaks in EVC. With high clutter, H2 and H3 predict that location information emerges along the dorsal and the ventral stream, respectively. **d**, EEG predictions. H1 predicts no delay for the emergence of location representations between background conditions. H2 predicts that location representations emerge later when clutter is present in the background than when it is not.

The second hypothesis posits that location representations are encoded in the dorsal stream (Fig. 1a, c, H2). This view is based on seminal neuropsychological case studies of patients with lesions in the dorsal stream, whose object localization behaviour was impaired, although object

recognition performance was intact (Goodale & Milner, 1992; James et al., 2003; Milner et al., 1991; Ungerleider & Haxby, 1994). These studies led to the dual-stream theory which proposed that object category and location are processed along two separate pathways in the brain (Ungerleider & Haxby, 1994). In this framework, location representations emerge along the dorsal stream, which is called the “where” pathway. Object category and identity emerge along the ventral stream which called the “what” pathway in this framework.

Finally, we propose a novel, third hypothesis in which location representations emerge along the ventral stream (Fig. 1a, c, H3). This view is based on a primate study in which location representations were higher in primate object-selective area IT than in V4 (Hong et al., 2016), suggesting an increase of location information from lower to higher-level ventral stream regions. In humans, location representations were found in high-level ventral visual region LOC (Baeck et al., 2013; Cichy et al., 2011, 2013; Golomb & Kanwisher, 2012; Schwarzlose et al., 2008). Traditionally, both IT and LOC are known as object-selective regions which encode category representations that are invariant to view-dependent properties like an object’s location (DiCarlo & Cox, 2007; Grill-Spector et al., 2001). This suggests that location information is gradually removed along the ventral stream and makes these regions rather counterintuitive candidates for the encoding of location representations. However, LOC’s two subregions LO1 and LO2 are selective for the processing of object shape and orientations (Silson et al., 2013), suggesting that LOC could be sensitive to other category-independent properties like location, too. Furthermore, LOC has large RF (Wandell & Winawer, 2015) with a peripheral bias (Sayres & Grill-Spector, 2008) which are beneficial properties for encoding the location of objects in the periphery via distributed patterns (Eurich & Schwegler, 1997; Snippe & Koenderink, 1992).

There are two reasons why previous studies could not dissociate between these three hypotheses and answer the question where location representations emerge. First, most studies described above presented objects on blank backgrounds (Baeck et al., 2013; Cichy et al., 2011, 2013; Golomb & Kanwisher, 2012; Schwarzlose et al., 2008). While this enhanced experimental control and established first insights about the encoding of object category and location, it limits the generalizability to real-world vision. The clutter of real-world environments turns object categorization and localization into a much more challenging problem for the visual system than in the lab. When objects are presented on blank backgrounds in the lab, the retinotopic portion of EVC is stimulated that corresponds to the position of the stimulus in the visual field, allowing for direct location read-out from EVC (Fig. 1b, left). When an object is presented on a cluttered

background however, this one-to-one mapping does not exist anymore because the entire EVC is visually stimulated (Fig. 1b, right) and direct read-out is not possible anymore. Consequently, presenting objects on cluttered backgrounds is especially important to distinguish the EVC hypothesis (H1) from the other two hypotheses (H2, H3). Furthermore, when objects are embedded in a cluttered scene, the visual system needs to perform additional grouping and segmentation operations to disentangle objects from the background and from each other (Poort et al., 2016; Scholte et al., 2008; Seijdel et al., 2021). It is likely that these operations require more in-depth processing that can only be accomplished along the ventral or dorsal hierarchy instead of EVC.

The second reason why previous research could not dissociate between the three hypotheses was that the only study that presented objects on natural scenes backgrounds was performed in primates and had limited coverage of the brain (Hong et al., 2016). In this study electrodes were placed in areas V4 and IT. However testing our three hypotheses additionally requires coverage of EVC and of the dorsal stream. This can easily be accomplished in an fMRI experiment with whole brain coverage. To address our first question *where* in cortical space the location of objects is represented, we tested our three hypotheses in study 1.

1.2 The neural temporal dynamics of object recognition in realistic environments

Visual representations in the human brain can not only be mapped out across the cortex but can also be measured emerging over time. Studying brain representations over time can yield important insights about the temporal sequence in which object representations develop. Some visual processes require more processing time compared to others and this allows us to infer which representations require additional processing steps (Carlson et al., 2013; Cichy et al., 2014; Isik et al., 2014; King & Dehaene, 2014). Object location has traditionally been regarded as a low-level feature in visual neuroscience (Rice et al., 2014) and low-level features are thought to be processed during the fast feedforward sweep of early visual processing stages (Contini et al., 2017). For example, on blank backgrounds, location representations of objects can be read out as early as 60 ms from the MEG signal (Carlson, Hogendoorn, Kanai, et al., 2011). This early emergence might be related to the high salience of the object on a blank background which allows for the encoding of the object’s location in a bottom-up manner (Itti & Koch, 2001). However, no prior studies investigated when location representations emerge when objects are presented on cluttered backgrounds. This is a crucial question, because disentangling an object

from a cluttered background requires operations that group visual information into separate objects and segments objects from the background and from other overlapping objects (Poort et al., 2016; Scholte et al., 2008; Seijdel et al., 2021). These grouping and segmentation operations have been related to recurrent processing (Lamme & Roelfsema, 2000; Seijdel et al., 2021, 2020) which is known to require additional processing time (Camprodon et al., 2010; Rajaei et al., 2019; Tang et al., 2014, 2018).

Studies on the temporal dynamics of category encoding demonstrate that presenting objects under more realistic and complex circumstances can dramatically alter their temporal dynamics (Kar et al., 2019), suggesting that this might also be the case for object location representations. For example, category representations of objects on cluttered backgrounds emerge later than on simple or on blank backgrounds (Groen et al., 2018; Seijdel et al., 2021). Similarly, presenting objects that are partially occluded can delay the emergence of their category in neural signals by up to 220 ms (Rajaei et al., 2019; Tang et al., 2014). These results highlight the importance of studying object representations with more ecological validity and complexity. Simultaneously, they raise the question how the temporal dynamics of location representations would be affected when objects would be presented on cluttered backgrounds. Based on the research above on category representations, our hypothesis was that the encoding of location representations of objects in clutter would require additional processing steps resulting in an increased processing time (Fig. 1d, H2). We tested this hypothesis in study 1.

1.3 Attentional modulation of object location representations

The previous two sections introduced the questions *where* and *when* the brain encodes object location representations. However, vision is not merely the processing of incoming sensory information in the brain. Visual representations in the brain can also be modulated by internal, cognitive processes (Kastner & Ungerleider, 2000). A cognitive process of paramount importance to visual perception that has been extensively studied is visual attention (Mangun, 1995; Maunsell, 2015; Squire et al., 2013). The role of spatial attention for location encoding of objects on clutter is particularly important because spatial attention focuses neural resources on important parts of the visual field and helps us ignore irrelevant parts (Desimone & Duncan, 1995), thereby alleviating the computational costs that clutter creates for the visual system (Reddy & Kanwisher, 2007; Wolfe, 1994; Wolfe et al., 2011). Behaviourally, this has been demonstrated in conjunction search (Treisman & Gelade, 1980) where a target letter is embedded among other

letters which differ from the target by two or more features (e.g. colour, shape). The complexity of the search display triggers top-down attentional resources and finding the target takes some time (Treisman & Gelade, 1980). In contrast, finding the target in a pop-up search display where the target differs by just one feature from distractors, is considerably faster because the target’s salience captures attention in a bottom-up manner (Braun, 1994; Itti & Koch, 2001; Treisman & Gelade, 1980; Wolfe et al., 2003). Similar neural processes might be at play when we perceive natural objects on blank vs. on cluttered backgrounds: on blank backgrounds, the object pops out and, therefore, its location should also be processed in a bottom-up manner during the feed-forward sweep of visual processing, requiring minimal attentional and computational resources (Itti & Koch, 2001; Treisman & Gelade, 1980; Wolfe, 1994; Wolfe et al., 2003). In contrast with clutter, the cognitive and neural processes of perceiving the object might be more similar to a conjunction search with covert spatial attention. This notion finds support in studies showing stronger attentional modulation of high-level visual areas when objects are embedded in clutter compared to when they are not (Lee & Maunsell, 2010; Reddy & Kanwisher, 2007).

Spatial coding in visual areas is essentially characterized by RF size and eccentricity bias of neurons (Groen et al., 2022). Attention is known to modulate neural responses by increasing the neural firing rates (Briggs et al., 2013; Desimone & Duncan, 1995; Reynolds & Chelazzi, 2004) and by increasing the RF size and eccentricity bias of neurons that have a receptive field in the attended location (Kay et al., 2015). Therefore, spatial attention might be tightly linked to representing object location.

Previous research showed mixed results concerning the processing stage when attentional modulation can be observed in brain measurements. Attentional modulation of visual representations has been observed both in low-level visual areas (Briggs et al., 2013; Herrero et al., 2013; Khayat et al., 2006; Lakatos et al., 2008; Martínez et al., 2001; Noesselt et al., 2002; Roelfsema et al., 1998) and high-level visual areas (Buffalo et al., 2010; Kay et al., 2015; Peelen & Kastner, 2011). In neural processing time, attentional modulation has been found both in early time windows (Hillyard, Teder-Sälejärvi, & Münte, 1998; Hillyard, Vogel, & Luck, 1998; Luck et al., 2000; Mangun, 1995) before the end of the feedforward sweep at ~ 150 ms (Camprodon et al., 2010; Fahrenfort et al., 2007; Koivisto et al., 2011; Lamme & Roelfsema, 2000; VanRullen & Thorpe, 2001) and in a late time window after the feedforward sweep (Battistoni et al., 2020; Groen et al., 2016; Kaiser et al., 2016; Wyatte et al., 2014).

The heterogeneity of these studies raises the question which role attentional modulation plays for the encoding of object location. Specifically, we asked during which processing stage in neural time and space covert spatial attention modulates location representations and whether attentional modulation depends on the clutter level of the background. Based on studies using naturalistic stimuli, we hypothesized that we would find attentional modulation of location representations during late processing stages in neural processing time and in cortical space (Battistoni et al., 2020; Kaiser et al., 2016; Kay et al., 2015; Peelen & Kastner, 2011). Based on the ubiquity of attentional modulation across paradigms and stimuli, we hypothesized that we would find attentional modulation independent of the background. We investigated these hypotheses in study 2.

1.4 Research questions and hypotheses

In sum, the previous sections brought forward three research questions that were addressed in this dissertation. The first question of this dissertation was *where* in cortical space the location of objects is represented in the human brain when objects are presented in cluttered environments. We formulated three hypotheses about *where* in cortical space the location of objects is encoded in the human brain (Cichy et al., 2013; Golomb & Kanwisher, 2012; Hong et al., 2016; Kravitz et al., 2011; Ungerleider & Haxby, 1994). According to hypothesis 1 this was in EVC, according to hypothesis 2 it was in the dorsal stream and according to hypothesis 3 in the ventral stream (Fig. 1a, c). We tested these hypotheses in an fMRI experiment in study 1 and further explored the role of spatial attention for these representations in study 2.

The second question of this dissertation was *when* location representations emerge in the human brain when objects are presented in cluttered environments. We hypothesized that disentangling an object from the background requires additional processing steps which delay the emergence of object location representations in time (Fig. 1d, H2; Groen et al., 2018; Seijdel et al., 2021; Thorat et al., 2021). We tested this hypothesis in an EEG experiment in study 1 and subsequently investigated the role of spatial attention for the temporal dynamics of location representations in study 2.

The third question of this dissertation was which role spatial attentional modulation plays for the encoding of location representations in the human brain. We specifically investigated during which processing stage attention modulates location representations and whether attentional modulation depends on the clutter level of the background or not. Since earlier studies found

attentional modulation of category representations during late processing stages in cortical space and neural time (Battistoni et al., 2020; Kaiser et al., 2016; Kay et al., 2015; Peelen & Kastner, 2011), we hypothesized that also location representations would be modulated by attention during late processing stages, independent of the background. We investigated this hypothesis in study 2 in an EEG and an fMRI experiment to characterize the processing stage of attentional modulation in neural time and cortical space.

1.5 Methods overview

In sum, this dissertation addressed the three experimental questions 1) *where* across the cortex location representations emerge when objects are presented on cluttered backgrounds 2) *when* those representations emerge and 3) which role spatial attentional modulation plays during location encoding. To answer these three questions, we combined different neuroscientific methods and analysis techniques to comprehensively investigate our experimental questions in cortical space and in neural time.

In both studies we recorded an fMRI experiment to map out the cortical distribution of location representations across the human brain and an EEG experiment to characterize their temporal neural dynamics. We combined these methods to exploit their spatial and temporal resolution, respectively. Although EEG can provide some coarse spatial information, fMRI has much higher spatial resolution (Dale & Halgren, 2001; Huettel et al., 2009; Luck, 2014). Conversely, in typical fMRI settings the sluggishness of the hemodynamic response precludes measuring temporal dynamics at the millisecond scale which was necessary here to distinguish between feedforward and recurrent processing stages. In short, fMRI has a good spatial and poor temporal resolution and the reverse is true for EEG (Dale & Halgren, 2001). Therefore, we combined these two methods to characterize the spatiotemporal processing of object location representations with high spatial and high temporal resolution.

In addition, in study 1 we modelled location representations in deep neural network models (DNN). These models currently represent the best performing computational tools to predict visual object representations in the ventral visual stream (Kubilius et al., 2019; Schrimpf et al., 2020; Yamins & DiCarlo, 2016). They show, for example, a hierarchical correspondence with neural processing in time and cortical space (Cichy, Khosla, et al., 2016; Güçlü & van Gerven, 2015; Yamins et al., 2014). Moreover, comparing different DNN architectures to brain represent-

ations can provide information about the computations that underlie those brain representations (Cichy & Kaiser, 2019; Kietzmann et al., 2019; Kriegeskorte & Douglas, 2018).

To analyse EEG and fMRI data in both studies, we applied a common analysis framework using multivariate pattern classification (Carlson, Hogendoorn, Kanai, et al., 2011; Haynes, 2015; Haynes & Rees, 2006; Isik et al., 2014). The advantage of this method compared to univariate treatments of data lies in its increased sensitivity by reading out information from neural data that lie in the joint pattern of combined fMRI voxels or EEG channels, rather than evaluating the outcome at each voxel or channel individually (Friston et al., 1995). In this dissertation, this analysis approach was additionally crucial to read out location information of objects on cluttered backgrounds. With a univariate approach, the stimulation of broad parts of the visual field would create unspecific activations in large portions of visual cortex that would not provide sufficient information about the object’s location on a cluttered background (Fig. 1b, right).

Hypothesis 3, regarding the first question *where* location representations emerge, predicts a shared neural substrate for object category and location representations. Therefore, it was pivotal to measure location information that was not confounded by category information. To accomplish this, we used a combination of a fully-crossed stimulus design with a cross-classification approach. Specifically, in all experiments, we presented each object exemplar once in each location and with each background (Fig. 2a). This stimulus design allowed us classify location information from brain measurements while training and testing on different object categories (Fig. 2b). The result of this cross-classification approach indicates that location information derived from training in one object category generalizes to testing in another object category and is, therefore, independent of category information. This common classification framework of pairwise classification of object location across categories was applied on data from the 4 experiments in this dissertation and on the activations of the DNNs. For fMRI data, it resulted in a spatial profile of location representations in ROIs and in a searchlight across the whole brain (Haynes & Rees, 2006, 2005). For EEG data, this resulted in time courses of location representations (Carlson, Hogendoorn, Kanai, et al., 2011; Isik et al., 2014). In the DNN activations, it resulted in a mechanistic profile of location representations (Kriegeskorte & Douglas, 2018).

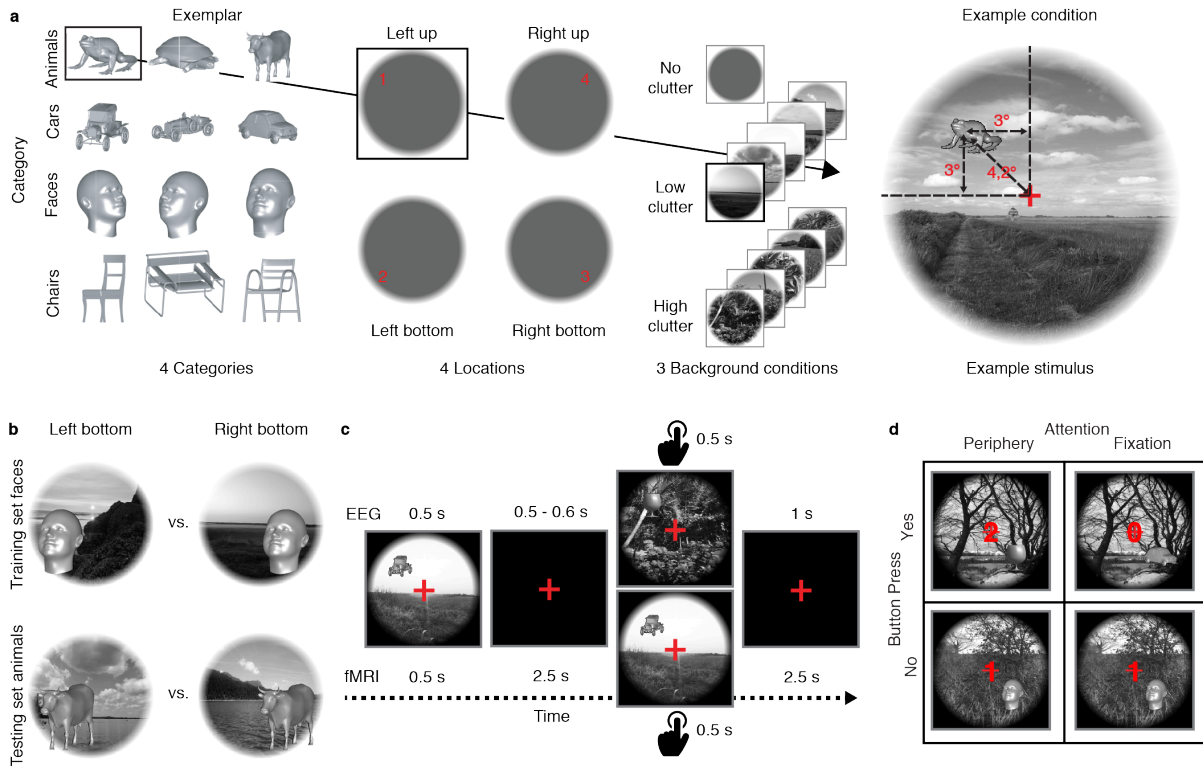


Figure 2: Experimental design, analysis scheme and tasks. **a**, Experimental design in study 1. Three object exemplars from four categories were presented in four locations with no, low and highly cluttered backgrounds. This amounted to 144 individual stimuli that were presented during the experiment. The experimental design and stimuli in study 2 were comparable to study 1. **b**, Cross-classification scheme. In all experiments, location information was classified from neural data by training a classifier on data associated with objects presented in two locations in the same category (here faces). This classifier was then tested on data associated with the same locations in a different object category (here animals). Objects are enlarged for visibility and did not extend into another quadrant in the original stimuli. **c**, Experimental tasks in study 1. In the EEG experiment, participants pressed a button when a catch object (glass) appeared. In the fMRI experiment, participants pressed a button when a stimulus was repeated (1-back task). Catch trials were removed from the analysis. **d**, Experimental tasks in study 2 in the EEG and fMRI experiments. In the condition with covert attention on the periphery, participants pressed a button when a catch object (glass) appeared. Digits on fixation were task-irrelevant in this condition. In the condition with attention on fixation, participants pressed a button when the digit 0 appeared on fixation. Objects were task-irrelevant in this condition. Catch trials were removed from the analysis.

Beyond the individual strengths of fMRI, EEG and DNNs, these methods can also be combined to bring forward new insights that cannot be provided by each method alone. To compare representations across methods, we used representational similarity analysis (RSA) (Cichy & Oliva, 2020; Cichy et al., 2014; Kriegeskorte, Mur, & Bandettini, 2008). RSA-based EEG-fMRI fusion allowed us to make inferences about lateral recurrent processes within LOC (Cichy & Oliva, 2020; Cichy et al., 2014). Comparing representations from fMRI and DNNs using RSA further

helped us fortify these conclusions to derive an algorithmic view of the underlying mechanism of the empirical effects.

Finally, to characterize the neural dynamics of location representations across time, we applied temporal-generalization analysis (TGA) (King & Dehaene, 2014). It has been suggested that this method can reveal distinct underlying neural processes that cannot be distinguished in the time courses (King & Dehaene, 2014). However, thus far, no methods existed to quantify and distinguish the different patterns that can emerge within a time-generalization matrix. Here, we developed a new analysis tool to assess whether two processes share neural information over a temporal delay (King and Dehaene, 2014). Our tool quantifies asymmetric off-diagonal patterns in the results of TGA when training and testing across two conditions with different temporal profiles.

2 Summary of Experiments

This section summarizes the two studies that this dissertation is based on. First I will introduce the methods that were applied to all four experiments in the two studies, with slight variations depending on imaging modality or paradigm. Each study consists of one EEG and one fMRI experiment. In our experimental designs, we fully crossed the factors object category, object location and background condition (Fig. 2a). Participants had to respond with a button press on trials where a catch object appeared (Fig. 2c, d). Catch trials were not included in the analysis and served to engage participants' attention during the experiment. For data analysis, we cross-classified location across categories as described above (Fig. 2b). The output of these analyses was category-independent location information, measured in percent classification accuracy, within background condition (study 1) or within background and attention condition (study 2). The first study is published, and the second study is submitted to a peer-reviewed journal. For further details on the methods, please refer to the original papers which are attached to this thesis.

2.1 Study 1: The spatiotemporal neural dynamics of object location representations in the human brain

In this study, we addressed the first two questions of this dissertation: *where* and *when* in the human brain object location representations emerge, when objects are presented on cluttered backgrounds compared to blank backgrounds.

2.1.1 Experiment 1: fMRI and DNN experiment

To answer the first question, *where* in cortical space location representations emerge, we tested three hypotheses in an fMRI experiment. The first hypothesis predicted location representations to be highest in EVC, the second hypothesis in the dorsal, and the third hypothesis in the ventral stream (Fig. 1a, c). To investigate the algorithmic plausibility of our results, we additionally modelled the results in a DNN. To test our three hypotheses, we compared the amount of location information across background conditions in EVC, in regions along the ventral and in regions along the dorsal stream.

During the experiment, participants passively viewed the stimuli and catch trials were excluded from the analysis (1-back task, Fig. 2c, bottom). To test if location information was

highest in EVC (hypothesis 1), increased along the dorsal stream (hypothesis 2) or increased along the ventral stream (hypothesis 3; Fig. 1a, c) when objects are presented with cluttered backgrounds, we performed 2 repeated-measures ANOVAs along the ventral and dorsal streams, comparing ROIs and background conditions.

Additionally, we determined the peak in location information across the whole cortex using a searchlight analysis. Our results of the ROI analysis showed that location representations in the high clutter condition increased along the ventral stream towards LOC. A spatially unbiased searchlight analysis confirmed this peak in LOC in the high clutter condition. A follow-up analysis showed that this effect was stronger for the classification of location across- than within-hemifields. These results were mirrored in a DNN, where location representations increased towards higher layers in the high clutter condition.

Together, these results demonstrate that when objects are presented on cluttered backgrounds, location representations emerge not in EVC or the dorsal, but in the ventral stream, peaking in high-level region LOC. This provides evidence for hypothesis 3 (Fig. 1a, c; ventral stream) and rules out hypothesis 1 (EVC) and 2 (dorsal stream).

2.1.2 Experiment 2: EEG experiment

The second question of this dissertation was *when* location representations emerge in time. Our hypothesis was that location representations of objects on cluttered backgrounds emerge later than with blank backgrounds in an EEG experiment (Fig. 1d). We tested this hypothesis by determining the classification peaks in the EEG time courses of each background condition and then testing the peak-to-peak latency differences between them.

During the experiment, participants passively viewed the stimuli, and all catch trials were excluded from the analysis (glass task, Fig. 2c, top). To quantify location representations over time, we classified object location across category as described above, and subsequently determined the classification peaks in the time course of each background condition and tested the delays between them for significance. The results showed that in the high clutter condition, location information peaked significantly later than in the no and low clutter conditions. In a subsequent time-generalization analysis we cross-classified location information across the no and high clutter conditions to determine if and when location information was shared between background conditions. This analysis showed that location information in the no and high clutter condi-

tions emerged during the same processing stage but with a delay in the high clutter condition. EEG-fMRI fusion localized this shared processing stage to LOC.

Thus, the results from the EEG experiment provide evidence for the hypothesis that location representations of objects in cluttered backgrounds emerge later than on blank backgrounds (Fig. 1d). Since delays within the same processing stage cannot be explained by a feedforward account, together the results from the time-resolved analysis, time-generalization analysis and from the EEG-fMRI fusion analysis provide evidence that location representations emerge during the recurrent processing stage in LOC.

2.1.3 Classification of object category in space and time

To investigate whether location and category representations share similar spatial and temporal neural substrates, we performed the analyses described above in parallel also for the classification of category, independent of location information. These analyses represented a replication of earlier findings because the emergence of object category along the ventral stream (Cichy, Pantazis, & Oliva, 2016; DiCarlo & Cox, 2007; DiCarlo et al., 2012; Goodale & Milner, 1992; Mishkin et al., 1983) with a temporal delay when viewing conditions are challenging has already been demonstrated in numerous studies (Groen et al., 2018; Kar et al., 2019; Rajaei et al., 2019; Seijdel et al., 2021; Tang et al., 2014).

We performed the classification analysis described above, reversing the factors category and location. Thus, we trained an SVM to classify between data associated with two object categories, training on data from objects in one location and testing on data from objects in another location, thereby quantifying the amount of category information that was independent of location information. This was done in ROIs and within and across time points. The fMRI results of this analysis showed that category information likewise increased along the ventral stream. In time, category information peaked later in the high clutter condition and emerged within the same processing stage as with no clutter, suggesting the involvement of recurrent processing.

Overall, these results demonstrate that location and category representations share similar neural substrates in space and time.

2.1.4 Contributions to open and reproducible science

To contribute to open science and increase the reproducibility of these results, we published the preprocessed fMRI, DNN and EEG data and the experimental stimuli on [OSF](#). The corres-

ponding analysis code is publicly available via github.com/graumannm/ObjectLocationRepresentations.

2.2 Study 2: Independent spatiotemporal effects of spatial attention and background clutter on human object location representations

Study 1 showed that location representations of objects on cluttered backgrounds emerged in LOC involving recurrent processing. Behavioral research showed a benefit of spatial attention for the perception of objects in clutter (Treisman & Gelade, 1980; Wolfe, 1994), suggesting that attention modulates objects on cluttered backgrounds during late stages of processing where they emerge. However, neuroscientific evidence is mixed concerning whether attention modulates neural responses during early (Briggs et al., 2013; Herrero et al., 2013; Hillyard, Vogel, & Luck, 1998; Hillyard, Teder-Sälejärvi, & Münte, 1998; Khayat et al., 2006; Lakatos et al., 2008; Luck et al., 2000; Mangun, 1995; Martínez et al., 2001; Noesselt et al., 2002; Roelfsema et al., 1998) or during late processing stages in space and time (Battistoni et al., 2020; Buffalo et al., 2010; Groen et al., 2016; Kaiser et al., 2016; Kay et al., 2015; Peelen & Kastner, 2011; Wyatte et al., 2014), suggesting that both early representations with blank backgrounds and late representations with cluttered backgrounds could be modulated by attention. Therefore, in study 2, we addressed the question which role attentional modulation plays during location encoding and whether the processing stage of attentional modulation depends on the clutter level of the background or not. We hypothesized that we would find evidence for attentional modulation of location representations at late stages of processing, independent of which background an object was presented on. We investigated this hypothesis in an EEG and an fMRI experiment to characterize the stage of attentional modulation in neural processing time and in cortical space.

2.2.1 Experiment 1: EEG experiment

We began our investigation by addressing the third question of this dissertation about attentional modulation in time, in an EEG experiment. We hypothesized that location representations would be modulated by attention during late processing stages, independent of the background. We tested this hypothesis by quantifying the difference between attention conditions over time, in background conditions with no and with high clutter.

During the experiment, participants performed tasks that either directed their covert spatial attention towards the objects in the periphery or that withdrew their spatial attention from the

objects towards fixation (Fig. 2d). Both tasks involved the presentation of a catch object, either on fixation or in the periphery. All stimuli were presented in both attention conditions, thus visual stimulation was equal across attention conditions. To determine at which processing stage in time attention modulates location representations, we classified location information across categories as described above, within background and within attention condition. We found that attention modulated both background conditions during late stages of processing after the feedforward loop. Furthermore, this study replicated EEG results from study 1, showing that location information with high clutter emerged with a temporal delay compared to no clutter. Additionally, this study showed that this delay occurred in both attention condition and was thus independent of attention.

In sum, the results of the EEG experiment confirmed the hypothesis that attention modulates location representations after the feedforward loop, independent of background. Additionally, they replicate the result from study 1 that location representations emerge with a delay when objects are presented with cluttered backgrounds and extended the finding by showing that this delay was independent of spatial attention.

2.2.2 Experiment 2: fMRI experiment

We then addressed the same, third question of this dissertation on the role of covert, spatial attention for location encoding, in space in an fMRI experiment. Our hypothesis was that location representations would be modulated by attention during late processing stages along the ventral stream, independent of the background. We tested this hypothesis in ANOVAs with factors attention and background in ROIs along the ventral stream.

The experimental design and tasks were the same as in the EEG experiment, with small adjustments to accommodate the longer trial duration required for fMRI while keeping the session duration within reasonable limits. We performed the same classification analysis as in the EEG experiment across ventral stream ROIs and tested for main and interaction effects of the factors background and attention with repeated-measures ANOVAs within each ROI. The results showed that location representations in mid- and high-level areas V3 and LOC were modulated by attention, independent of background. V4 was modulated by both attention and background. In line with the results of study 1, location representations in EVC regions V1 and V2 were not robust to background clutter and thus showed a main effect of background. The main effect of background was independent of attention since we found no evidence for attentional modulation

in V1 and V2. Also in line with study 1, we did not find evidence for widespread location representations in the dorsal stream.

Overall, these results show that attention modulates location representations in mid- and high-level ventral visual regions, independent of background. This provides evidence for our hypothesis that attention modulates location representations during late processing stages in space, independent of the background on which an object is presented.

2.2.3 Contributions to open and reproducible science

To contribute to open science and increase the reproducibility of these results, we publish the preprocessed EEG and fMRI data on [OSF](#). The corresponding analysis code will be publicly available on github via github.com/graumannm/AttentionLocation.

3 General Discussion

3.1 Summary

This dissertation investigated the spatiotemporal neural dynamics of location representations when objects are presented on blank vs. cluttered backgrounds and the role of spatial attention therein. This was examined by breaking down the problem into three questions, each comparing location representations on blank vs. on cluttered backgrounds. First, we asked *where* the location of objects is represented in cortical space. Second, we asked *when* the location of objects is represented in time. Third, we asked which role spatial attentional modulation plays for the encoding of object location. We addressed the first and second question in study 1 and the third question in study 2.

Overall, study 1 demonstrated that location representations of objects on cluttered backgrounds emerge along the ventral stream, peaking in LOC with a temporal delay that was linked to recurrent processing. Study 2 showed that these representations late in time and in mid- and high-level ventral stream regions benefit from attentional modulation, independent of background. These results provide evidence for the hypotheses that location representations emerge towards late stages of the ventral stream, emerge late in time and are enhanced by attentional modulation. This resolves long standing debates as to *where* and *when* location representations are processed in the human brain and which role attentional modulation plays during location encoding.

Our findings show that location information is not treated like a low-level property by the brain when an object's environment is cluttered. As opposed to what was previously suggested, object location and category representations both emerge along the ventral stream hierarchy. These findings demonstrate how approximating the complexity of real-world environments in an experimentally controlled manner can reveal new insights about object representations in the brain.

3.2 Location representations of objects with clutter emerge along the ventral visual stream

The first question of this dissertation was *where* in cortical space the location of objects is represented when objects are presented on blank vs. on cluttered backgrounds. We addressed this question in an fMRI experiment in study 1 and found that location representations of objects on

cluttered backgrounds emerged along the ventral visual stream, peaking in LOC. Study 2 showed that these representations were enhanced by spatial attention. These results clearly support the hypothesis that location representations emerge along the ventral visual stream and allowed us to reject the hypotheses that location representations are encoded in EVC or the dorsal stream.

Our findings are consistent with previous studies finding location representations of objects on blank backgrounds along the ventral visual stream, including LOC (Baeck et al., 2013; Cichy et al., 2011, 2013; Golomb & Kanwisher, 2012; Schwarzlose et al., 2008; Xu & Vaziri-Pashkam, 2021). They also align with a primate study showing that location representations of objects on natural scene backgrounds are higher in IT than in V4 (Hong et al., 2016). However, our results go beyond previous findings in three important ways. First, comparing location representations of objects on blank and cluttered backgrounds allowed us to rule out EVC as a candidate region. Second, as opposed to the primate study which recorded activity only from V4 and IT (Hong et al., 2016), we were able to compare results across three candidate regions by recording whole brain fMRI data and to reliably pinpoint the locus of location representations in the human brain to the ventral, rather than the dorsal stream or EVC. Third, we show that spatial attention helps to encode the location of objects by focusing neural resources on the relevant parts of the visual field and enhancing neural responses to the object's location.

Our results have important implications for a seminal theory of visual perception, the dual-stream theory (Mishkin et al., 1983; Ungerleider & Mishkin, 1982; Ungerleider & Haxby, 1994). This theory posits that object category and location representations emerge along two separate processing streams in the brain: according to this theory, category representations emerge along the ventral stream and location representations emerge along the dorsal stream (Mishkin et al., 1983; Ungerleider & Mishkin, 1982; Ungerleider & Haxby, 1994). However, recent studies showed, that the division of location and category information into these two streams might not be as strictly separated as initially proposed (Konen & Kastner, 2008). Those studies found that category information is also present in the dorsal and location information in the ventral stream (Carlson, Hogendoorn, Fonteijn, & Verstraten, 2011; Cichy et al., 2011, 2013; Konen & Kastner, 2008; Kourtzi et al., 2002). Our results are consistent with these studies and go beyond them by showing that the ventral stream is the primary locus of object location representations.

How can the discrepancies between our finding and the dual-stream theory be explained? One way to resolve these discrepancies is by looking at neuropsychological studies since the dual stream theory was, among other evidence, based on neuropsychological findings. I will first

discuss how case studies of patients with ventral and dorsal lesions are not contradicting our results upon closer look. Then I will discuss what type of information the dorsal stream might encode instead of object location representations.

The theory that the ventral stream is the "what" pathway and encodes object category representations is partly based on neuropsychological case studies showing impaired object recognition in patients with ventral stream lesions (James et al., 2003; Ungerleider & Haxby, 1994). However, impairments of object recognition due to lesions in the ventral stream (agnosia) come in many different forms (de Haan, 2019). For example, integrative agnosia patients are impaired at tasks requiring figure-ground segmentation and grouping and have problems recognizing objects in clutter (Behrmann et al., 1994; De Renzi & Lucchelli, 1993; Humphreys & Riddoch, 1987; Riddoch & Humphreys, 1987). Thus, consistent with our findings, ventral stream lesions can lead to impairments of the operations necessary to recognize cluttered objects. Grouping and figure-ground segmentation are crucial for segregating objects from cluttered backgrounds and, therefore, our results predict impaired object location perception in these patients, too. Future studies could explicitly assess not only object category, but also location perception in patients with ventral lesions to test this prediction.

Together, our results and the neuropsychological findings above suggest that the ventral, not the dorsal stream encodes the location of objects in cluttered scenes. This however raises the question what type of information is encoded by the dorsal stream. Early neuropsychological findings that gave rise to the dual-stream theory found that patients with damage to dorsal areas performed well in object categorization tasks but were impaired during object localization (Goodale, 2011; Goodale & Milner, 1992; Ungerleider & Haxby, 1994). How can our findings be consolidated with this? When taking a closer look at these studies, it becomes clear that localization behavior rather than location perception was impaired in these patients, which does not contradict our results (Goodale & Milner, 1992; Milner et al., 1991). The distinction between behavior and perception is important because phenomena like blind sight have shown that it is possible to use visual information, which is not consciously processed, for behavior (Weiskrantz, 1986). In the case of blind sight, information from spared EVC can be used to guide behavior (Fendrich et al., 1992). Since our results showed that location representations of objects with blank backgrounds are robustly encoded in EVC, it is possible that patients with ventral stream lesions were able to use location information from spared EVC for behavior. In line with this notion, later theories updated the name for the dorsal stream from "where" to "how" pathway,

proposing that it supports vision for action rather than for spatial perception (Goodale & Milner, 1992; Milner & Goodale, 2006). Consistent with this notion, regions in parietal cortex play an important role for coordinating vision and movement during reaching behaviors (Filimon et al., 2009; Rossit et al., 2013). In our study, the intraparietal sulcus (IPS) hosted only very limited amounts of location information. IPS is known to be activated during eye-movements (Pierrot-Deseilligny et al., 2004) and selectively responds to tools (Chao & Martin, 2000). Tools have affordances and consequently, their perception is directly linked to a possible action (Osiurak et al., 2010). Our studies excluded both eye movements and the object category of tools, which might be a reason why we did not find more location information in IPS. More recent work suggests that the dorsal stream and the ventral stream are highly interconnected (Cloutman, 2013; Milner, 2017) and that the dorsal stream uses visual information from the ventral stream for subsequent actions (Milner, 2017; van Polanen & Davare, 2015).

In sum, our results demonstrate that the ventral visual stream, rather than the dorsal stream or EVC hosts location representations when objects are perceived on cluttered backgrounds and are not linked to an action. Our results are consistent with neuropsychological studies finding impaired grouping and figure-ground segmentation in patients with ventral stream lesions (De Renzi & Lucchelli, 1993; Humphreys & Riddoch, 1987; Riddoch & Humphreys, 1987). More recent research suggests that the dorsal stream is relevant for visually guided action rather than location perception (Goodale & Milner, 1992; Milner & Goodale, 2006).

3.3 Location representations of objects with clutter emerge during the recurrent processing stage

The second question of this dissertation was *when* location representations emerge when objects are presented on cluttered backgrounds. We addressed this question in an EEG experiment in study 1 and found that location representations of objects on cluttered backgrounds emerged later in time than when objects were presented on blank backgrounds. This delay was linked to recurrent processing in LOC. Study 2 additionally showed that the delays occurred independently of spatial attention. Our findings are notable for two reasons. First, for a long time core object recognition was thought to rely on feedforward processing, without strong dependence on recurrence (DiCarlo et al., 2012; Riesenhuber & Poggio, 1999). Second, although more recent studies have demonstrated the involvement of recurrent processing in object recognition (Kar et

al., 2019; Kietzmann et al., 2019; Rajaei et al., 2019; Tang et al., 2018), these studies focused only on category, not location representations.

According to the feedforward account, core object recognition was the result of a series of successive signal transformations along the ventral visual hierarchy, where information becomes increasingly complex towards higher-level stages (DiCarlo et al., 2012). Information processing was one-directional and recurrent processing was not necessary to solve object recognition in the feedforward view (Serre et al., 2007). This notion was supported by studies showing that both location and category representations of objects emerged within the first ~ 150 ms of processing (Carlson, Hogendoorn, Kanai, et al., 2011; Cichy et al., 2014; Isik et al., 2014; Thorpe et al., 1996; VanRullen & Thorpe, 2001) which is a time window that can be linked to the initial fast feedforward sweep of visual processing (Camprodon et al., 2010; Fahrenfort et al., 2007; Koivisto et al., 2011; Lamme & Roelfsema, 2000; VanRullen & Thorpe, 2001) from the retina, via the lateral geniculate nucleus (LGN) and EVC up to category selective ventral regions like IT in primates or LOC in humans (Cichy et al., 2014; Isik et al., 2014; Lamme & Roelfsema, 2000). However, the visual system contains a large number of lateral and top-down connections (Felleman & Van Essen, 1991) and more recent studies demonstrated that such recurrent connections play an integral part during object recognition (Kar & DiCarlo, 2020; Kar et al., 2019; Kietzmann et al., 2019). Here, we extend these findings by showing that not only category representations, but also location representations are computed involving recurrent processing. This suggests that recurrence might be a ubiquitous mechanism involved in the processing of visual objects, including their category-independent features (Thorat et al., 2021).

Consistent with our findings, previous work showed that recurrent processing is particularly important when objects are presented in clutter and when objects are partially occluded (Groen et al., 2018; Rajaei et al., 2019; Seijdel et al., 2020, 2021; Spoerer et al., 2017; Tang et al., 2014, 2018). For example, category representations of partially occluded objects emerge between 60 ms (Rajaei et al., 2019) and 220 ms (Tang et al., 2014) later than non-occluded objects. These latencies are comparable to the delays for location representations observed in our EEG experiments (study 1: 177 ms; study 2: 116 ms and 79 ms). In some of the previous studies, a mask that was presented after the stimulus, interrupted recurrent information flow, resulting in impaired object recognition performance, demonstrating the contribution of recurrent processing (Fahrenfort et al., 2007; Rajaei et al., 2019; Seijdel et al., 2021). These findings are further supported by computational modelling work, where recurrent DNNs perform better than feedforward DNNs on object

categorization tasks with partially occluded objects or objects on complex backgrounds (Seijdel et al., 2020, 2021; Spoerer et al., 2020, 2017). This lends further algorithmic plausibility to the necessity of a recurrent neural architecture to solve object recognition under challenging viewing conditions. As for category, in our study recurrent DNNs outperformed feedforward DNNs when comparing location information magnitude and prediction strength of representations in LOC. Modelling work in recurrent DNNs trained on object categorization showed that the location of objects in clutter becomes increasingly explicit with each recurrent iteration (Thorat et al., 2021). This indicates that recurrent loops directly maintain location information.

Thus, previous research extensively demonstrated the involvement of recurrent processing for the representation of object category when viewing conditions are challenging. Our study 1 shows that this is also the case for location representations. This raises the question which visual features of clutter trigger recurrent processing and which operations are performed during the recurrent loop to process those features. A possible answer to this question is that cluttered images are high in spatial coherence and contrast energy (Groen et al., 2018; Scholte et al., 2009). These visual features drive recurrent responses in cluttered images (Groen et al., 2018; Seijdel et al., 2021) and might therefore have triggered the recurrent computations observed in our study. Furthermore, cluttered backgrounds require segmentation as well as grouping operations and these operations are performed during the recurrent processing stage (Lamme & Roelfsema, 2000; Seijdel et al., 2020, 2021). Together, these empirical and modelling studies suggest that clutter creates the need for recurrent computations which in turn increase both location and category information over time.

3.4 Late attentional modulation of object location representations

Following up on the results of study 1, the third question of this dissertation was which role spatial attentional modulation plays for the representation of object location. We addressed this question in study 2 and found attentional modulation in a late time window after the feedforward sweep and in mid- and high-level ventral areas, independent of background. Additionally, the EEG experiment showed that the temporal delays between clutter conditions that were observed in study 1, also occurred independent of attentional modulation in study 2.

The EEG experiment in study 2 showed that attention enhanced location representations after the feedforward sweep ending at ~ 150 ms (Camprodon et al., 2010; Fahrenfort et al., 2007; Koivisto et al., 2011; Lamme & Roelfsema, 2000; VanRullen & Thorpe, 2001) but attention

had no significant influence on the timing of the responses. The late emergence of location representations was related to recurrent processing in study 1. Together, both studies suggest that recurrence and attention operate independently of each other, but within the same time window. Background clutter triggered recurrent processing which we measured in delayed responses. In contrast, attention enhanced responses which we measured in higher location information. These two processes likely have different underlying mechanisms: while delays due to recurrence are related to iterative processing (Spoerer et al., 2020), attentional response enhancement is driven by increased firing rates (Briggs et al., 2013; Maunsell, 2015).

Attention and recurrence have been proposed to operate on separate time scales, with recurrence starting before attention (Wyatte et al., 2014). Our results suggest instead that attention and recurrence can operate independently of each other but within the same time window. An interesting question for future research is why both processes coincide in time. One possibility is that attention operates in a late time window because long-range feedback from PFC and parietal cortices needs to travel to visual areas to trigger attentional modulation (Corbetta & Shulman, 2002; Squire et al., 2013) and that attention is fully independent of recurrent processing. Another possibility is that attentional response enhancement depends on computations performed during the recurrent loop.

In line with the EEG results showing attentional modulation at late processing stages in time, the fMRI experiment in study 2 showed attentional modulation at late processing stages in cortical space, in mid- and high-level visual areas. The next section will discuss possible reasons why both location representations and attentional modulation increase along the ventral visual stream and how this could explain improved location encoding in those areas.

3.5 Possible mechanisms of location encoding in the ventral stream

What are the mechanisms behind the encoding of location representations with attention in high-level ventral visual cortex? The experiments in this thesis are not designed to answer this question, but combined with previous research, our results can provide a basis for a theoretical, mechanistic account.

The smallest building blocks of visual spatial processing in the brain are RF size and eccentricity of neurons and neuronal populations (Groen et al., 2022). High-level ventral visual areas might robustly encode location representations of objects in clutter because these regions might have RF properties that are advantageous. In particular, high-level ventral visual regions

have RFs that are larger than in EVC (Wandell & Winawer, 2015). Large and overlapping RFs could provide better spatial resolution for objects on clutter by capturing the object’s location via distributed patterns (Kay et al., 2015). Such a mechanism has been described within the coarse coding framework (Eurich & Schwegler, 1997; Snippe & Koenderink, 1992). In contrast, the small RFs found in EVC (Wandell & Winawer, 2015) can provide high-spatial resolution to encode objects on blank backgrounds via direct retinotopic activation of the corresponding location in the visual field (Fig. 1b, left). With clutter, however, no such mapping is possible because large parts of the visual field and hence of EVC are stimulated (Fig. 1b, right). Thus, within this framework, large RF represent an advantage for spatial coding. Consistent with this view, RFs are known to increase in size and overlap progressively along the ventral stream (Groen et al., 2022; Wandell & Winawer, 2015), which could thus account for the increase of location information along the ventral stream via the proposed coarse coding mechanism.

How could attention be beneficial within this framework? Attention increases RF size and eccentricity (Kay et al., 2015) and this effect is stronger for larger RFs (Klein et al., 2014). Thus, attention increases large RFs more than small RFs. Since RF size increases along the ventral stream (Wandell & Winawer, 2015), this consequently suggests that the RFs of high-level regions increase more with attentional modulation than small RFs in EVC (Klein et al., 2014). Hence, a possible reason why both location representations and attentional modulation increase along the ventral stream might be that attention aids location encoding in high-level ventral visual areas by increasing their RF’s size (Kay et al., 2015; Klein et al., 2014).

In sum, coarse coding and attentional modulation of RF properties together provide a parsimonious mechanistic framework for location encoding along the ventral stream and provide a possible explanation why both location representations and attentional modulation increase along the ventral stream (Eurich & Schwegler, 1997; Snippe & Koenderink, 1992). This framework could be tested with a combination of population receptive field (pRF) mapping and modelling work.

3.6 The role of location representations for object recognition

In study 1 we not only found location representations to increase along the ventral stream, but we also replicated that location-invariant category representations are encoded in high-level ventral visual cortex (Baeck et al., 2013; Cichy et al., 2011, 2013; Rust & DiCarlo, 2010; Schwarzlose et al., 2008). How and why does the ventral stream encode location-invariant category representations and category-invariant location representations simultaneously (Cichy et al., 2011,

2013; Schwarzlose et al., 2008)? The coarse coding mechanism described in the previous section provides robustness to location translation needed for location-invariant category encoding (Spirkowska & Reid, 1993), therefore accounting for the simultaneous, yet invariant encoding of location and category. Thus, coarse coding provides a framework for achieving higher spatial resolution despite clutter on one hand and location invariance on the other hand. Here I will discuss three different lines of research which describe how location information could benefit object recognition.

The first line of research shows how spatial coding and category encoding are connected in the topography of visual areas. For scientific research it is important to establish category-invariant location representations that are free of confounding category information. However, in the real world we cannot separate a perceived object from its location and vice versa. Therefore, the coexistence of location and category representations in the same brain regions is not surprising. In fact, recent studies suggest that the coding of visual space and category might be tightly linked. For example, many category-selective regions in high-level ventral visual cortex exhibit eccentricity biases that are optimized for the categories that they encode (Groen et al., 2022). For example the FFA, which is selectively activated for faces (Kanwisher et al., 1997) exhibits a foveal bias and has small RFs (Finzi et al., 2021) which is beneficiary for recognizing faces which are usually foveated. LOC and the PPA, which is selectively activated for scenes (Epstein & Kanwisher, 1998), both exhibit a peripheral bias and have large RFs (Levy et al., 2001; Sayres & Grill-Spector, 2008; Silson et al., 2015, 2016; Wandell & Winawer, 2015), which is advantageous for perceiving objects and scenes because they are often perceived in the periphery (Groen et al., 2017). Recently, it has been shown that these RF biases likely emerge along with the category-selectivity of these brain regions and constrain the functional topography of high-level category selective areas (Gomez et al., 2019).

The second line of research shows how recurrent processing of location representations might improve object categorization. Specifically, a recent DNN study suggests that location information is used during recurrent processing to focus neural resources on the part of a cluttered scene containing the object and thereby improving categorization performance (Thorat et al., 2021). Such a process might explain why both category and location processing depend on recurrence (Graumann et al., 2022; Kietzmann et al., 2019; Seijdel et al., 2021). Thus, it could be an advantage to explicitly and independently represent location information in the brain to focus subsequent neural processing steps on the relevant part of the visual field to create more fine-

grained representations of object category that then become independent of location (DiCarlo & Cox, 2007; Thorat et al., 2021).

Finally, the third line of research shows that location and category information might interact in the brain to exploit the spatial structure of our environments to facilitate object recognition (Kaiser et al., 2019; Vö et al., 2019). For example, object category representations are enhanced when objects are presented at their typical locations, such as planes in the upper visual field and carpets in the lower visual field (Kaiser & Cichy, 2018; Kaiser et al., 2018, 2014). Furthermore, object recognition is faster and more accurate when objects are presented on congruent scenes compared to incongruent scenes (Bar, 2004; Biederman, 1972; Oliva & Torralba, 2007; Seijdel et al., 2020) and when they are presented in expected locations (Kaiser et al., 2019). Since the studies in this thesis did not congruently and predictably combine objects and scenes, the above-mentioned research poses limitations on the conclusions that we can draw from our results. These limitations will be discussed in the next section.

3.7 Limitations

We highlight three relevant limitations of the studies presented here. The first limitation of our studies is that it is unclear in how far our results would generalize to images where objects congruently fit into the scene, e.g. a car on a street. Object-scene congruency could potentially influence results because it has been shown that congruency between objects, scenes and locations enhances both behavioral recognition performance (Kaiser et al., 2014; Vö et al., 2019) and neural processing (Kaiser & Cichy, 2018; Kaiser et al., 2018), as described in the previous section. In detail, it has been shown that presenting objects in congruent scenes enhances detection speed, accuracy of object recognition and neural processing (Brandman & Peelen, 2017, 2019; Kaiser et al., 2019; Vö et al., 2019). Scene context is used as information for recognizing objects (Wischnewski & Peelen, 2021) and can aid object recognition when objects are degraded (Brandman & Peelen, 2017). These interactions are causally related to activity in scene selective cortex (Wischnewski & Peelen, 2021).

Despite this potential limitation, we chose to randomly pair objects, locations and background scenes to avoid these systematic congruency effects. This was necessary to manipulate only clutter while keeping all other influences constant and attribute our results to the contrast between objects on blank vs. on cluttered backgrounds. Another reason to randomly pair objects and scenes was that it was pivotal for our design to precisely control the positions of objects in the

visual field to avoid noise or confounds from varying positions that objects have in naturally occurring locations.

Future experiments could go beyond our design by systematically comparing location representations in the brain when objects are presented on congruent vs. incongruent cluttered backgrounds. However, based on a study on object categorization which found feedback signals during the processing of objects in complex, congruent scenes, we predict that our results would replicate with congruent scenes (Groen et al., 2018).

A second limitation is that we inferred the involvement of recurrent processing based on converging evidence from multiple analyses, but we did not directly manipulate recurrence in study 1. One possible direct manipulation to assess whether a visual process requires recurrent processing is to include a condition during which the presented image is followed by a visual mask (e.g. a scrambled image) (Fahrenfort et al., 2007; Seijdel et al., 2021). This mask interrupts ongoing recurrent processes of a first stimulus by presenting a new stimulus (Fahrenfort et al., 2007). Behavioral or neural measurements can then be compared between conditions with and without masking. The rationale is that if a process depends on recurrent processing, behavioral performance will be impaired and neural responses will be reduced in the masking condition, compared to a condition with no masking (Fahrenfort et al., 2007; Rajaei et al., 2019; Seijdel et al., 2021). Other possibilities to directly manipulate recurrent information flow include TMS (Wischniewski & Peelen, 2021) and pharmacological interventions in monkeys (Kar & DiCarlo, 2020). Therefore, given our design we have indirect, but no direct evidence that the measured signals late in time reflect recurrent processing.

Although we did not directly manipulate recurrence, we did find indirect, converging evidence for the involvement of recurrence during the processing of location representations by combining EEG, fMRI and DNN data in a number of analyses. First, the time-generalization analysis in study 1 (Fig. 5e) shows that location representations of objects with cluttered backgrounds are delayed but emerge during the same processing stage as with blank backgrounds. Second, the subsequent EEG-fMRI fusion analysis (Fig. 5f) showed that this processing stage could be localized to LOC. Since temporal delays within the same processing stage cannot be explained by a purely feedforward account, these results strongly suggest the involvement of recurrence via lateral connections within LOC (Graumann et al., 2022). Third, location representations between objects with blank and cluttered backgrounds were shared in LOC, but not in other regions (Supplementary Fig. 7), confirming the shared, delayed processing stages found in time

to also exist in cortical space. Fourth, for objects with cluttered backgrounds, DNNs with a recurrent architecture showed an advantage compared to shallow, feedforward neural networks for encoding location representations of objects with cluttered backgrounds (Supplementary Fig. 3c). Fifth, recurrent DNNs additionally performed better than shallow feedforward DNNs in predicting brain representations in LOC (Supplementary Fig. 3d). Together, this indicates that recurrent architectures provide a better model for brain responses of objects in clutter than feedforward architectures and thus add plausibility for the proposed mechanism. Based on this converging evidence, we predict that a similar experiment including a masking condition would result in reduced neural responses with masking, indicating recurrent processing.

Finally, the third limitation of the studies in this thesis is, that we cannot draw conclusions as to whether our findings are behaviorally relevant, since we did not include a task on object location that was part of the analysis. Decodable information in the brain and across time does not automatically imply that this information is also used by the brain and is relevant for behavior (De-Wit et al., 2016; Grootswagers et al., 2018; Williams et al., 2007). A possibility to assess the behavioral relevance of these results would be to conduct an experiment that includes a behavioral task that could subsequently be integrated with EEG and fMRI results using RSA. In the behavioral task, objects could be presented in a large number of different locations on blank or cluttered backgrounds and participants would have to perform a speeded object detection task. Location representations could subsequently be classified from reaction time data, EEG data and fMRI data separately, building RDMs for each of these modalities. The reaction time RDM could be compared to the neural RDMs using RSA. This would yield an estimate of the behavioral relevance of the neural representations across different brain regions and over time. An alternative to this analysis would be the approach described in Grootswagers et al. (2018) during which the reaction times are correlated to the hyperplane distances generated during classification of EEG or fMRI data. Based on a previous study finding category representations in LOC to be behaviorally relevant (Williams et al., 2007), we predict that also location representations in LOC will be behaviorally relevant.

3.8 Methodological strengths

Our results demonstrate how combining methods can lead to a more comprehensive picture of the underlying mechanisms. In particular, in study 1, combining fMRI and EEG results in a common analysis framework (EEG-fMRI fusion; Cichy et al., 2014; Cichy & Teng, 2017; Cichy &

Oliva, 2020) allowed us to draw conclusions about the cortical sources of the time course results. This in turn allowed us to conclude that an early and a late process were related to the same brain region (LOC) and infer the involvement of recurrence, which would not have been possible with either of our methods (EEG or fMRI) alone. In general, converging evidence from different methods can provide stronger evidence and clearer insights into neural mechanisms (Dale & Halgren, 2001; Jorge et al., 2014). Single methods could be biased or yield a limited view on neural mechanisms. One very obvious advantage is that EEG provides high temporal and low spatial resolution while the opposite is true for fMRI (Dale & Halgren, 2001). However, the underlying challenges of combining methods are the different temporal scales at which responses are measured and units in which they are measured. EEG-fMRI fusion overcomes these challenges by bringing EEG and fMRI results into a common representational space, that can be combined independent of univariate response unit and magnitude. Therefore, it provides a suited method to overcome methodological differences and combine results into a common framework. A limitation of this method is that its signal-to-noise-ratio depends on the richness of the condition set and that it can only be applied to parameters to which both modalities are sensitive (Cichy & Oliva, 2020). For example it has not been established yet for parametric designs, where different intensities, frequencies or speeds of the same variable are being compared. This could be achieved, e.g. by creating RDMs using a support vector regression for continuous as opposed to categorical variables.

3.9 Future directions

This thesis lays the groundwork for future research directions. Here, I propose two directions that future studies could take to gain new insights about the encoding of location representations in the brain.

The first direction is to investigate a causal relationship between recurrent processing in LOC and location perception. This could be done for example in a TMS experiment. The experiment could present participants with objects in various locations on blank or cluttered backgrounds and subsequently present a grid and ask participants in which of the locations the object was presented. TMS could be applied 1) at different time points to probe recurrent processing and 2) over different areas such as EVC, dorsal regions and LOC, to probe the causal relationship between activations in a region and location perception. Based on this thesis, the prediction for

such an experiment is that TMS over LOC after ~ 200 ms should impair location perception of objects on clutter.

Another way of investigating a causal relationship between location perception and activity in LOC would be by conducting case studies with neuropsychological patients with lesions in this region. Previous studies in agnosia patients already showed impaired object categorization when recognition required grouping and segmentation because of clutter or overlapping objects (De Renzi & Lucchelli, 1993; Riddoch & Humphreys, 1987). New studies could explicitly test patients' location perception for objects on cluttered backgrounds. Based on this thesis I predict that patients with ventral damage will be impaired in both object recognition and the perception of their location when objects are presented in cluttered environments.

The second direction that future research could take is to investigate which neural properties allow LOC to encode location representations better than other regions. This could be done using a pRF mapping paradigm or modelling experiment. Both methods could test the coarse-coding theory described above which states that regions with larger, overlapping RFs can encode location with higher resolution and are better suited to segment an object from its background. A pRF mapping experiment could for example correlate pRF size and eccentricity to location information in voxel groups across the ventral stream with and without spatial attention. A modelling experiment could model RF size and overlap as kernel size and stride in a shallow DNN, and compare how kernel size and stride affect location information across DNN layers. The proposed mechanism above predicts that larger and overlapping RFs should be related to increased location information for objects with cluttered backgrounds.

3.10 Conclusion

When we perceive the world around us, we rarely see an object in isolation, but we usually see objects surrounded by a large number of other objects, resulting in cluttered environments that we effortlessly navigate through in everyday life. This dissertation investigated the spatiotemporal neural dynamics of object location representations when the object's backgrounds are cluttered. Our results show that location representations emerge along the ventral stream, peaking in LOC involving recurrent processing when objects appear on cluttered backgrounds. Spatial attention modulates location representations at late stages of processing in neural time and along the ventral visual hierarchy, independent of the object's background.

References

- Baeck, A., Wagemans, J., & Op de Beeck, H. P. (2013). The distributed representation of random and meaningful object pairs in human occipitotemporal cortex: The weighted average as a general rule. *NeuroImage*, *70*, 37–47. doi: 10.1016/j.neuroimage.2012.12.023
- Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, *5*(8), 617–629. doi: 10.1038/nrn1476
- Battistoni, E., Kaiser, D., Hickey, C., & Peelen, M. V. (2020). The time course of spatial attention during naturalistic visual search. *Cortex*, *122*, 225–234. doi: 10.1016/j.cortex.2018.11.018
- Behrmann, M., Moscovitch, M., & Winocur, G. (1994). Intact Visual Imagery and Impaired Visual Perception in a Patient With Visual Agnosia. *Journal of Experimental Psychology: Human Perception and Performance*, *20*(5), 1068–1087. doi: 10.1037/0096-1523.20.5.1068
- Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*(4043), 77–80. doi: 10.1126/science.177.4043.77
- Brandman, T., & Peelen, M. V. (2017). Interaction between scene and object processing revealed by human fMRI and MEG decoding. *The Journal of Neuroscience*, *37*(32), 7700–7710. doi: 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. doi: 10.1162/jocn_a_01258
- Braun, J. (1994). Visual search among items of different salience: Removal of visual attention mimics a lesion in extrastriate area V4. *Journal of Neuroscience*, *14*(2), 554–567. doi: 10.1523/jneurosci.14-02-00554.1994
- Briggs, F., Mangun, G. R., & Usrey, W. M. (2013). Attention enhances synaptic efficacy and the signal-to-noise ratio in neural circuits. *Nature*, *499*(7459), 476–480. doi: 10.1038/nature12276
- Buffalo, E. A., Fries, P., Landman, R., Liang, H., & Desimone, R. (2010). A backward progression of attentional effects in the ventral stream. *Proceedings of the National Academy of Sciences*, *107*(1), 361–365. doi: 10.1073/pnas.0907658106
- Camprodon, J. A., Zohary, E., Brodbeck, V., & Pascual-Leone, A. (2010). Two phases of V1 activity for visual recognition of natural images. *Journal of Cognitive Neuroscience*, *22*(6), 1262–1269. doi: 10.1162/jocn.2009.21253
- Carlson, T. A., Hogendoorn, H., Fonteijn, H., & Verstraten, F. A. (2011). Spatial coding and invariance in object-selective cortex. *Cortex*, *47*(1), 14–22. doi: 10.1016/j.cortex.2009.08.015

- Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J., & Turret, J. (2011). High temporal resolution decoding of object position and category. *Journal of Vision*, *11*(10), 1–17. doi: 10.1167/11.10.1
- 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), 1–19. doi: 10.1167/13.10.1
- Chao, L. L., & Martin, A. (2000). Representation of manipulable man-made objects in the dorsal stream. *NeuroImage*, *12*(4), 478–484. doi: 10.1006/nimg.2000.0635
- Cichy, R. M., Chen, Y., & Haynes, J. D. (2011). Encoding the identity and location of objects in human LOC. *NeuroImage*, *54*(3), 2297–2307. doi: 10.1016/j.neuroimage.2010.09.044
- Cichy, R. M., & Kaiser, D. (2019). Deep neural networks as scientific models. *Trends in Cognitive Sciences*, *23*(4), 305–317. doi: 10.1016/j.tics.2019.01.009
- Cichy, R. M., Khosla, A., Pantazis, D., Torralba, A., & Oliva, A. (2016). Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. *Scientific Reports*, *6*, 1–13. doi: 10.1038/srep27755
- Cichy, R. M., & Oliva, A. (2020). A M/EEG-fMRI fusion primer: Resolving human brain responses in space and time. *Neuron*, *107*(5), 772–781. doi: 10.1016/j.neuron.2020.07.001
- Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature Neuroscience*, *17*(3), 455–462. doi: 10.1038/nn.3635
- Cichy, R. M., Pantazis, D., & Oliva, A. (2016). Similarity-based fusion of MEG and fMRI reveals spatio-temporal dynamics in human cortex during visual object recognition. *Cerebral Cortex*, *26*(8), 1–17. doi: 10.1093/cercor/bhw135
- Cichy, R. M., Sterzer, P., Heinzle, J., Elliott, L. T., Ramirez, F., & Haynes, J.-D. (2013). Probing principles of large-scale object representation: Category preference and location encoding. *Human Brain Mapping*, *34*(7), 1636–1651. doi: 10.1002/hbm.22020
- 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). doi: 10.1098/rstb.2016.0108
- Cloutman, L. L. (2013). Interaction between dorsal and ventral processing streams: Where, when and how? *Brain and Language*, *127*(2), 251–263. doi: 10.1016/j.bandl.2012.08.003
- 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*(10), 165–176. doi: 10.1016/j.neuropsychologia.2017.02.013
- Corbetta, M., & Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, *3*(3), 201–215. doi: 10.1038/nnr755

- Dale, A. M., & Halgren, E. (2001). Spatiotemporal mapping of brain activity by integration of multiple imaging modalities. *Current Opinion in Neurobiology*, *11*(2), 202–208. doi: 10.1016/S0959-4388(00)00197-5
- De Renzi, E., & Lucchelli, F. (1993). The Fuzzy Boundaries of Apperceptive Agnosia. *Cortex*, *29*(2), 187–215. doi: 10.1016/S0010-9452(13)80176-1
- de Haan, E. (2019). *Impaired vision: how the visual world may change after brain damage*. NJ: Wiley-Blackwell. doi: 10.2469/cfm.v26.n1.5
- Desimone, R., & Duncan, J. (1995). Selective visual attention. *Annual Review of Neuroscience*, *18*, 193–222. doi: 10.4135/9781483328768.n2
- De-Wit, L., Alexander, D., Ekroll, V., & Wagemans, J. (2016). Is neuroimaging measuring information in the brain? *Psychonomic Bulletin and Review*, *23*(5), 1415–1428. doi: 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. doi: 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. doi: 10.1016/j.neuron.2012.01.010
- Engel, S. A., Rumelhart, D. E., Wandell, B. A., Lee, A. T., Glover, G. H., Eduardo-Jose, C., ... Wandell (1994). fMRI of human visual cortex. *Nature*, *369*(6481), 525. doi: 10.1038/369525a0
- Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, *392*, 598–601. doi: 10.1016/s1053-8119(18)31174-1
- Eurich, C. W., & Schwegler, H. (1997). Coarse coding: Calculation of the resolution achieved by a population of large receptive field neurons. *Biological Cybernetics*, *76*(5), 357–363. doi: 10.1007/s004220050349
- Fahrenfort, J. J., Scholte, H. S., & Lamme, V. A. (2007). Masking disrupts reentrant processing in human visual cortex. *Journal of Cognitive Neuroscience*, *19*(9), 1488–1497. doi: 10.1162/jocn.2007.19.9.1488
- Felleman, D., & Van Essen, D. C. (1991). Distributed hierarchical processing in the primate cerebral cortex. *Cerebral Cortex*, *1*(1), 1–47. doi: 10.1093/cercor/1.1.1-a
- Fendrich, R., Wessinger, C. M., & Gazzaniga, M. S. (1992). Residual vision in a scotoma: implications for blindsight. *Science*, *258*(5087), 1489–1491. doi: 10.1126/science.1439839
- Filimon, F., Nelson, J. D., Huang, R. S., & Sereno, M. I. (2009). Multiple parietal reach regions in humans: Cortical representations for visual and proprioceptive feedback during on-line reaching. *Journal of Neuroscience*, *29*(9), 2961–2971. doi: 10.1523/JNEUROSCI.3211-08.2009

- Finzi, D., Gomez, J., Nordt, M., Rezai, A. A., Poltoratski, S., & Grill-Spector, K. (2021). Differential spatial computations in ventral and lateral face-selective regions are scaffolded by structural connections. *Nature Communications*, *12*(1), 1–14. doi: 10.1038/s41467-021-22524-2
- Friston, K. J., Holmes, A. P., Worsley, K. J., Poline, J. Frith, C. D., & Frackowiak, R. S. (1995). Statistical parametric maps in functional imaging: A general linear approach. *Human Brain Mapping*, *2*(4), 189–210. doi: 10.1002/hbm.460020402
- Golomb, J. D., & Kanwisher, N. (2012). Higher level visual cortex represents retinotopic, not spatiotopic, object location. *Cerebral Cortex*, *22*(12), 2794–2810. doi: 10.1093/cercor/bhr35
- Gomez, J., Barnett, M., & Grill-Spector, K. (2019). Extensive childhood experience with Pokémon suggests eccentricity drives organization of visual cortex. *Nature Human Behaviour*, *3*(6), 611–624. doi: 10.1038/s41562-019-0592-8
- Goodale, M. A. (2011). Transforming vision into action. *Vision Research*, *51*(13), 1567–1587. doi: 10.1016/j.visres.2010.07.027
- Goodale, M. A., & Milner, A. D. (1992). Separate visual pathways for perception and action. *Essential Sources in the Scientific Study of Consciousness*, *15*(1), 20–25. doi: 10.1016/0166-2236(92)90344-8
- Graumann, M., Ciuffi, C., Dwivedi, K., Roig, G., & Cichy, R. M. (2022). The spatiotemporal neural dynamics of object location representations in the human brain. *Nature Human Behaviour*, 1–38. doi: 10.1038/s41562-022-01302-0
- Grill-Spector, K., Kourtzi, Z., & Kanwisher, N. (2001). The lateral occipital complex and its role in object recognition. *Vision Research*, *41*(10-11), 1409–1422. doi: 10.1016/S0042-6989(01)00073-6
- Groen, I. I. A., Dekker, T. M., Knapen, T., & Silson, E. H. (2022). Visuospatial coding as ubiquitous scaffolding for human cognition. *Trends in Cognitive Sciences*, *26*(1), 81–96. doi: 10.1016/j.tics.2021.10.011
- Groen, I. I. A., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2016). The time course of natural scene perception with reduced attention. *Journal of Neurophysiology*. doi: 10.1152/jn.00896.2015
- Groen, I. I. A., Jahfari, S., Seijdel, N., Ghebreab, S., Lamme, V. A. F., & Scholte, H. S. (2018). Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLoS Computational Biology*, *14*(12), e1006690. doi: 10.1371/journal.pcbi.1006690
- Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1714), 20160102. doi: 10.1098/rstb.2016.0102

- Groetswagers, T., Cichy, R. M., & Carlson, T. A. (2018). Finding decodable information that can be read out in behaviour. *NeuroImage*, *179*, 252–262. doi: 10.1016/j.neuroimage.2018.06.022
- Güçlü, U., & van Gerven, M. A. J. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *The Journal of Neuroscience*, *35*(27), 10005–10014. doi: 10.1523/JNEUROSCI.5023-14.2015
- Haynes, J.-D. (2015). A Primer on Pattern-Based Approaches to fMRI: Principles, Pitfalls, and Perspectives. *Neuron*, *87*(2), 257–270. doi: 10.1016/j.neuron.2015.05.025
- Haynes, J.-D., & Rees, G. (2005). Predicting the stream of consciousness from activity in human visual cortex. *Current Biology*, *115*(14), 1301–1307. doi: 10.1016/j.cub.2005.06.026
- Haynes, J.-D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*, 523–534.
- Herrero, J. L., Gieselmann, M. A., Sanayei, M., & Thiele, A. (2013). Attention-induced variance and noise correlation reduction in macaque V1 is mediated by NMDA receptors. *Neuron*, *78*(4), 729–739. doi: 10.1016/j.neuron.2013.03.029
- Hillyard, S. A., Teder-Sälejärvi, W. A., & Münte, T. F. (1998). Temporal dynamics of early perceptual processing. *Current Opinion in Neurobiology*, *8*(2), 202–210. doi: 10.1016/S0959-4388(98)80141-4
- Hillyard, S. A., Vogel, E. K., & Luck, S. J. (1998). Sensory gain control (amplification) as a mechanism of selective attention: Electrophysiological and neuroimaging evidence. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *353*(1373), 1257–1270. doi: 10.1098/rstb.1998.0281
- Holmes, G. (1918). Disturbances of vision by cerebral lesions. *The British Journal of Ophthalmology*, *2*(7), 353–384. doi: 10.1136/bjo.2.7.353
- Hong, H., Yamins, D. L. K., Majaj, N. J., & DiCarlo, J. J. (2016). Explicit information for category-orthogonal object properties increases along the ventral stream. *Nature Neuroscience*, *19*, 613–622. doi: 10.1038/nn.4247
- Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *J. Physiol.*, *148*, 574–591. doi: 10.1109/SOCC.2011.6085109
- Hubel, D. H., & Wiesel, T. N. (1977). Functional architecture of macaque monkey visual cortex. *Proc R Soc Lond B*, *198*(1130), 1–59.
- Huettel, S. A., Song, A. W., & McCarthy, G. (2009). *Functional Magnetic Resonance Imaging* (Second Edi ed.). Sunderland, MA: Sinauer Associates, Inc.
- Humphreys, G. W., & Riddoch, M. J. (1987). The fractionation of visual agnosia. In G. W. Humphreys & M. J. Riddoch (Eds.), *Visual object processing: A cognitive neuropsychological approach*. Hove, England: Erlbaum.

- Humphreys, G. W., & Riddoch, M. J. (1994). Intermediate visual processing and visual agnosia. In M. J. Farah & R. Ratcliff (Eds.), *The neuropsychology of high-level vision: Collected tutorial essays*. Hillsdale, NJ: NJ.
- Hung, C. P. (2005). Fast readout of object identity from macaque inferior temporal cortex. *Science*, *310*(5749), 863–866. doi: 10.1126/science.1117593
- 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. doi: 10.1152/jn.00394.2013
- Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, *2*(3), 194–203. doi: 10.1038/35058500
- James, T. W., Culham, J., Humphrey, G. K., Milner, A. D., & Goodale, M. A. (2003). Ventral occipital lesions impair object recognition but not object-directed grasping: An fMRI study. *Brain*, *126*(11), 2463–2475. doi: 10.1093/brain/awg248
- Jorge, J., Van der Zwaag, W., & Figueiredo, P. (2014). EEG-fMRI integration for the study of human brain function. *NeuroImage*, *102*, 24–34. doi: 10.1016/j.neuroimage.2013.05.114
- 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. doi: 10.1152/jn.00229.2018
- Kaiser, D., Moeskops, M. M., & Cichy, R. M. (2018). Typical retinotopic locations impact the time course of object coding. *NeuroImage*, *176*, 372–379. doi: 10.1016/j.neuroimage.2018.05.006
- Kaiser, D., Oosterhof, N. N., & Peelen, M. V. (2016). The neural dynamics of attentional selection in natural scenes. *Journal of Neuroscience*, *36*(41), 10522–10528. doi: 10.1523/JNEUROSCI.1385-16.2016
- Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences*, *23*(8), 672–685. doi: 10.1016/j.tics.2019.04.013
- Kaiser, D., Stein, T., & Peelen, M. V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences*, *111*(30), 11217–11222. doi: 10.1073/pnas.1400559111
- Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *The Journal of neuroscience*, *17*(11), 4302–11. doi: 10.1523/JNEUROSCI.17-11-04302.1997
- Kar, K., & DiCarlo, J. J. (2020). Fast recurrent processing via ventrolateral prefrontal cortex is needed by the primate ventral stream for robust core visual object recognition. *Neuron*, *109*(1), 164–176. doi: 10.1016/j.neuron.2020.09.035

- Kar, K., Kubilius, J., Schmidt, K., Issa, E. B., & DiCarlo, J. J. (2019). Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nature Neuroscience*, *22*(6), 974–983. doi: 10.1038/s41593-019-0392-5
- Kastner, S., & Ungerleider, L. G. (2000). Mechanisms of visual attention in the human cortex. *Annual review of neuroscience*, *23*, 315–341. doi: 10.1146/annurev.neuro.23.1.315
- Kay, K. N., Weiner, K. S., & Grill-Spector, K. (2015). Attention reduces spatial uncertainty in human ventral temporal cortex. *Current Biology*, *25*(5), 595–600. doi: 10.1016/j.cub.2014.12.050
- Khayat, P. S., Spekreijse, H., & Roelfsema, P. R. (2006). Attention lights up new object representations before the old ones fade away. *Journal of Neuroscience*, *26*(1), 138–142. doi: 10.1523/JNEUROSCI.2784-05.2006
- Kiani, R., Esteky, H., Mirpour, K., & Tanaka, K. (2007). Object category structure in response patterns of neuronal population in monkey inferior temporal cortex. *Journal of Neurophysiology*, *97*(6), 4296–4309. doi: 10.1152/jn.00024.2007
- Kietzmann, T. C., Spoerer, C. J., Sörensen, L. K., Cichy, R. M., Hauk, O., & Kriegeskorte, N. (2019). Recurrence is required to capture the representational dynamics of the human visual system. *Proceedings of the National Academy of Sciences of the United States of America*, *116*(43), 21854–21863. doi: 10.1073/pnas.1905544116
- King, J. R., & Dehaene, S. (2014). Characterizing the dynamics of mental representations: The temporal generalization method. *Trends in Cognitive Sciences*, *18*(4), 203–210. doi: 10.1016/j.tics.2014.01.002
- Klein, B. P., Harvey, B. M., & Dumoulin, S. O. (2014). Attraction of position preference by spatial attention throughout human visual cortex. *Neuron*, *84*(1), 227–237. doi: 10.1016/j.neuron.2014.08.047
- Koivisto, M., Railo, H., Revonsuo, A., Vanni, S., & Salminen-Vaparanta, N. (2011). Recurrent processing in V1/V2 contributes to categorization of natural scenes. *Journal of Neuroscience*, *31*(7), 2488–2492. doi: 10.1523/JNEUROSCI.3074-10.2011
- Konen, C. S., & Kastner, S. (2008). Two hierarchically organized neural systems for object information in human visual cortex. *Nature Neuroscience*, *11*(2), 224–231. doi: 10.1038/nn2036
- Kourtzi, Z., Bühlhoff, H. H., Erb, M., & Grodd, W. (2002). Object-selective responses in the human motion area MT/MST. *Nature Neuroscience*, *5*(1), 17–18. doi: 10.1038/nn780
- Kravitz, D. J., Saleem, K. S., Baker, C. I., & Mishkin, M. (2011). A new neural framework for visuospatial processing. *Nature Reviews Neuroscience*, *12*(4), 217–30. doi: 10.1038/nrn3008
- Kriegeskorte, N., & Douglas, P. K. (2018). Cognitive computational neuroscience. *Nature Neuroscience*, *21*(9), 1148–1160. doi: 10.1038/s41593-018-0210-5

- Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. doi: 10.3389/neuro.06.004.2008
- Kriegeskorte, N., Mur, M., Ruff, D. A., Kiani, R., Bodurka, J., Esteky, H., . . . Bandettini, P. A. (2008). Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, *60*(6), 1126–1141. doi: 10.1016/j.neuron.2008.10.043
- Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., . . . DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs. In H. Wallach, H. Larochelle, A. Beygelzimer, F. Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32, pp. 12805–12816). Curran Associates, Inc.
- Lakatos, P., Karmos, G., Mehta, A. D., Ulbert, I., & Schroeder, C. E. (2008). Entrainment of neuronal oscillations as a mechanism of attentional selection. *Science*, *320*(5872), 110–113. doi: 10.1126/science.1154735
- Lamme, V. A., & Roelfsema, P. R. (2000). The distinct modes of vision offered by feedforward and recurrent processing. *Trends in neurosciences*, *23*(11), 571–579. doi: 10.1016/s0166-2236(00)01657-x
- Lee, J., & Maunsell, J. H. R. (2010). Attentional modulation of MT neurons with single or multiple stimuli in their receptive fields. *Journal of Neuroscience*, *30*(8), 3058–3066. doi: 10.1523/JNEUROSCI.3766-09.2010
- Levy, I., Hasson, U., Avidan, G., Hendler, T., & Malach, R. (2001). Center-periphery organization of human object areas. *Nature Neuroscience*, *4*(5), 533–539. doi: 10.1038/8749
- Li, N., Cox, D. D., Zoccolan, D., & DiCarlo, J. J. (2009). What Response Properties Do Individual Neurons Need to Underlie Position and Clutter "Invariant" Object Recognition? *Journal of Neurophysiology*, *102*(1), 360–376. doi: 10.1152/jn.90745.2008.
- Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd editio ed.). Cambridge: MIT Press.
- Luck, S. J., Woodman, G. F., & Vogel, E. K. (2000). Event-related potential studies of attention. *Trends in Cognitive Sciences*, *4*(11), 432–440. doi: 10.1016/S1364-6613(00)01545-X
- Malach, R., Reppas, J. B., Benson, R. R., Kwong, K. K., Jlang, H., Kennedy, W. A., . . . Tootell, R. B. H. (1995). Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proceedings of the National Academy of Sciences*, *92*(18), 8135–8139. doi: 10.1073/pnas.92.18.8135
- Malcolm, G. L., Groen, I. I. A., & Baker, C. I. (2016). Making Sense of Real-World Scenes. *Trends in Cognitive Sciences*, *20*(11), 843–856. doi: 10.1016/j.tics.2016.09.003

- Mangun, G. R. (1995). Neural mechanisms of visual selective attention. *Psychophysiology*, *32*(1), 4–18. doi: 10.1111/j.1469-8986.1995.tb03400.x
- Martínez, A., DiRusso, F., Anllo-Vento, L., Sereno, M. I., Buxton, R. B., & Hillyard, S. A. (2001). Putting spatial attention on the map: Timing and localization of stimulus selection processes in striate and extrastriate visual areas. *Vision Research*, *41*(10-11), 1437–1457. doi: 10.1016/S0042-6989(00)00267-4
- Maunsell, J. H. (2015). Neuronal Mechanisms of Visual Attention. *Annual Review of Vision Science*, *1*(1), 373–391. doi: 10.1146/annurev-vision-082114-035431
- Milner, A. D. (2017). How do the two visual streams interact with each other? *Experimental Brain Research*, *235*(5), 1297–1308. doi: 10.1007/s00221-017-4917-4
- Milner, A. D., & Goodale, M. A. (2006). *The visual brain in action*. Oxford: Oxford University Press.
- Milner, A. D., Perrett, D. I., Johnston, R. S., Benson, P. J., Jordan, T. R., Heeley, D. W., ... Davidson, D. L. (1991). Perception and action in 'visual form agnosia'. *Brain*, *114*(1), 405–428. doi: 10.1093/brain/114.1.405
- Mishkin, M., Ungerleider, L. G., & Macko, K. A. (1983). Object vision and spatial vision: two cortical pathways. *Trends in Neurosciences*, *6*, 414–417. doi: 10.1016/0166-2236(83)90190-X
- Noesselt, T., Hillyard, S. A., Woldorff, M. G., Schoenfeld, A., Hagner, T., Jäncke, L., ... Heinze, H. J. (2002). Delayed striate cortical activation during spatial attention. *Neuron*, *35*(3), 575–587. doi: 10.1016/S0896-6273(02)00781-X
- Oliva, A., & Torralba, A. (2007). The role of context in object recognition. *Trends in Cognitive Sciences*, *11*(12), 520–527. doi: 10.1016/j.tics.2007.09.009
- Osiurak, F., Jarry, C., & Le Gall, D. (2010). Grasping the Affordances, Understanding the Reasoning: Toward a Dialectical Theory of Human Tool Use. *Psychological Review*, *117*(2), 517–540. doi: 10.1037/a0019004
- Pasupathy, A., & Connor, C. E. (2002). Population coding of shape in area V4. *Nature Neuroscience*, *5*(12), 1332–1338. doi: 10.1038/nm972
- Peelen, M. V., & Kastner, S. (2011). A neural basis for real-world visual search in human occipitotemporal cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *108*(29), 12125–12130. doi: 10.1073/pnas.1101042108
- Peelen, M. V., & Kastner, S. (2014). Attention in the real world: Toward understanding its neural basis. *Trends in Cognitive Sciences*, *18*(5), 242–250. doi: 10.1016/j.tics.2014.02.004
- Pierrot-Deseilligny, C., Milea, D., & Müri, R. M. (2004). Eye movement control by the cerebral cortex. *Current Opinion in Neurology*, *17*(1), 17–25. doi: 10.1097/00019052-200402000-0000

- Poort, J., Self, M. W., Van Vugt, B., Malkki, H., & Roelfsema, P. R. (2016). Texture Segregation Causes Early Figure Enhancement and Later Ground Suppression in Areas V1 and V4 of Visual Cortex. *Cerebral Cortex*, *26*(10), 3964–3976. doi: 10.1093/cercor/bhw235
- Rajaei, K., Mohsenzadeh, Y., Ebrahimpour, R., & Khaligh-Razavi, S.-M. (2019). Beyond core object recognition: Recurrent processes account for object recognition under occlusion. *PLOS Computational Biology*, *15*(5), e1007001. doi: 10.1371/journal.pcbi.1007001
- Reddy, L., & Kanwisher, N. (2007). Category Selectivity in the Ventral Visual Pathway Confers Robustness to Clutter and Diverted Attention. *Current Biology*, *17*(23), 2067–2072. doi: 10.1016/j.cub.2007.10.043
- Reynolds, J. H., & Chelazzi, L. (2004). Attentional modulation of visual processing. *Annual Review of Neuroscience*, *27*, 611–647. doi: 10.1146/annurev.neuro.26.041002.131039
- Rice, G. E., Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Low-level image properties of visual objects predict patterns of neural response across category-selective regions of the ventral visual pathway. *Journal of Neuroscience*, *34*(26), 8837–8844. doi: 10.1523/JNEUROSCI.5265-13.2014
- Riddoch, M. J., & Humphreys, G. W. (1987). A case of integrative visual agnosia. *Brain*, *110*(6), 1431–1462. doi: 10.1093/brain/110.6.1431.
- Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature*, *2*(11), 1019–1025. doi: 10.1038/14819
- Roelfsema, P. R., Lamme, V. A. F., & Spekreijse, H. (1998). Object-based attention in the primary visual cortex of the macaque monkey. *Nature*, *395*, 376–381. doi: 10.1038/26475
- Rossit, S., McAdam, T., Mclean, D. A., Goodale, M. A., & Culham, J. C. (2013). fMRI reveals a lower visual field preference for hand actions in human superior parieto-occipital cortex (SPOC) and precuneus. *Cortex*, *49*(9), 2525–2541. doi: 10.1016/j.cortex.2012.12.014
- Rust, N. C., & DiCarlo, J. J. (2010). Selectivity and tolerance ("invariance") both increase as visual information propagates from cortical area V4 to IT. *Journal of Neuroscience*, *30*(39), 12978–12995. doi: 10.1523/JNEUROSCI.0179-10.2010
- Sayres, R., & Grill-Spector, K. (2008). Relating retinotopic and object-selective responses in human lateral occipital cortex. *Journal of Neurophysiology*, *100*(1), 249–267. doi: 10.1152/jn.01383.2007
- Scholte, H. S., Ghebreab, S., Waldorp, L., Smeulders, A. W., & Lamme, V. A. (2009). Brain responses strongly correlate with Weibull image statistics when processing natural images. *Journal of Vision*, *9*(4), 1–15. doi: 10.1167/9.4.29
- Scholte, H. S., Jolij, J., Fahrenfort, J. J., & Lamme, V. A. (2008). Feedforward and recurrent processing in scene segmentation: Electroencephalography and functional magnetic resonance imaging. *Journal of Cognitive Neuroscience*, *20*(11), 2097–2109. doi: 10.1162/jocn.2008.20142

- Schrimpf, M., Kubilius, J., Lee, M. J., Murty, N. A. R., Ajemian, R., & DiCarlo, J. J. (2020). Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron*, *108*(3), 413–423. doi: 10.1016/j.neuron.2020.07.040
- Schwarzlose, R. F., Swisher, J. D., Dang, S., & Kanwisher, N. (2008). The distribution of category and location information across object-selective regions in human visual cortex. *Proceedings of the National Academy of Sciences*, *105*(11), 4447–4452. doi: 10.1073/pnas.0800431105
- Sejdel, N., Loke, J., van de Klundert, R., van der Meer, M., Quispel, E., van Gaal, S., ... Scholte, H. S. (2021). On the necessity of recurrent processing during object recognition: It depends on the need for scene segmentation. *Journal of Neuroscience*, *41*(29), 6281–6289. doi: 10.1523/JNEUROSCI.2851-20.2021
- Sejdel, N., Tsakmakidis, N., De Haan, E. H., Bohte, S. M., & Scholte, H. S. (2020). Depth in convolutional neural networks solves scene segmentation. *PLoS Computational Biology*, *16*(7), e1008022. doi: 10.1371/journal.pcbi.1008022
- Serre, T., Lior, W., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on pattern analysis and machine intelligence*, *29*(3), 411–426. doi: 10.12816/0047609
- Silson, E. H., Chan, A. W. Y., Reynolds, R. C., Kravitz, D. J., & Baker, C. I. (2015). A retinotopic basis for the division of high-level scene processing between lateral and ventral human occipitotemporal cortex. *Journal of Neuroscience*, *35*(34). doi: 10.1523/JNEUROSCI.0137-15.2015
- Silson, E. H., Groen, I. I. A., Kravitz, D. J., & Baker, C. I. (2016). Evaluating the correspondence between face-, scene-, and object-selectivity and retinotopic organization within lateral occipitotemporal cortex. *Journal of Vision*, *16*(6), 1–21. doi: 10.1167/16.6.14
- Silson, E. H., McKeefry, D. J., Rodgers, J., Gouws, A. D., Hymers, M., & Morland, A. B. (2013). Specialized and independent processing of orientation and shape in visual field maps LO1 and LO2. *Nature Neuroscience*, *16*(3), 267–269. doi: 10.1038/nn.3327
- Snippe, H. P., & Koenderink, J. J. (1992). Discrimination thresholds for channel-coded systems. *Biological Cybernetics*, *66*(6), 543–551. doi: 10.1007/BF00204120
- Spirkovska, L., & Reid, M. B. (1993). Coarse-coded higher-order neural networks for PSRI object recognition. *IEEE Transactions on Neural Networks*, *4*(2), 276–283.
- Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I., & Kriegeskorte, N. (2020). Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Computational Biology*, *16*(10), e1008215. doi: 10.1371/journal.pcbi.1008215
- Spoerer, C. J., McClure, P., & Kriegeskorte, N. (2017). Recurrent convolutional neural networks: A better model of biological object recognition. *Frontiers in Psychology*, *8*(SEP), 1551. doi: 10.3389/fpsyg.2017.01551

- Squire, R. F., Noudoost, B., Schafer, R. J., & Moore, T. (2013). Prefrontal contributions to visual selective attention. *Annual Review of Neuroscience*, *36*, 451–466. doi: 10.1146/annurev-neuro-062111-150439
- Tang, H., Buia, C., Madhavan, R., Crone, N. E., Madsen, J. R., Anderson, W. S., & Kreiman, G. (2014). Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron*, *83*(3), 736–748. doi: 10.1016/j.neuron.2014.06.017
- Tang, H., Schrimpf, M., Lotter, W., Moerman, C., Paredes, A., Caro, J. O., . . . Kreiman, G. (2018). Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences of the United States of America*, *115*(35), 8835–8840. doi: 10.1073/pnas.1719397115
- Thorat, S., Aldegheri, G., & Kietzmann, T. C. (2021). Category-orthogonal object features guide information processing in recurrent neural networks trained for object categorization. *arXiv preprint arXiv:2111.07898*.
- Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, *381*(6582), 520–522. doi: 10.1038/381520a0
- Tootell, R. B., Switkes, E., Silverman, M. S., & Hamilton, S. L. (1988). Functional anatomy of macaque striate cortex. II. Retinotopic organization. *Journal of Neuroscience*, *8*(5), 1569–1593. doi: 10.1523/jneurosci.08-05-01569.1988
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive psychology*, *12*(1), 97–136. doi: 10.1016/0010-0285(80)90005-5
- Ungerleider, L. G., & Haxby, J. V. (1994). 'What' and 'where' in the human brain. *Current Opinion in Neurobiology*, *4*(2), 157–165. doi: 10.1016/0959-4388(94)90066-3
- Ungerleider, L. G., & Mishkin, M. (1982). Two cortical systems. In *Analysis of visual behavior*. Cambridge: MIT Press.
- van Polanen, V., & Davare, M. (2015). Interactions between dorsal and ventral streams for controlling skilled grasp. *Neuropsychologia*, *79*, 186–191. doi: 10.1016/j.neuropsychologia.2015.07.010
- VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: From early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461. doi: 10.1162/08989290152001880
- Vö, 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. doi: 10.1016/j.copsyc.2019.03.009
- Wandell, B. A., Dumoulin, S. O., & Brewer, A. A. (2007). Visual field maps in human cortex. *Neuron*, *56*(2), 366–383. doi: 10.1016/j.neuron.2007.10.012

- Wandell, B. A., & Winawer, J. (2015). Computational neuroimaging and population receptive fields. *Trends in Cognitive Sciences*, *19*(6), 349–357. doi: 10.1016/j.tics.2015.03.009
- Weiskrantz, L. (1986). *Blindsight: a case study and implications*. Oxford, England: Oxford University Press.
- Williams, M. A., Dang, S., & Kanwisher, N. G. (2007). Only some spatial patterns of fMRI response are read out in task performance. *Nature Neuroscience*, *10*(6), 685–686. doi: 10.1038/nn1900
- Wischnewski, M., & Peelen, M. V. (2021). Causal neural mechanisms of context-based object recognition. *eLife*, *10*, 1–13. doi: 10.7554/ELIFE.69736
- Wolfe, J. M. (1994). Visual search in continuous, naturalistic stimuli. *Vision Research*, *34*(9), 1187–1195. doi: 10.1016/0042-6989(94)90300-x
- Wolfe, J. M., Butcher, S. J., Lee, C., & Hyle, M. (2003). Changing Your Mind: On the Contributions of Top-Down and Bottom-Up Guidance in Visual Search for Feature Singletons. *Journal of Experimental Psychology: Human Perception and Performance*, *29*(2), 483–502. doi: 10.1037/0096-1523.29.2.483
- Wolfe, J. M., Võ, M. L., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, *15*(2), 77–84. doi: 10.1016/j.tics.2010.12.001
- Wyatte, D., Jilk, D. J., & O'Reilly, R. C. (2014). Early recurrent feedback facilitates visual object recognition under challenging conditions. *Frontiers in Psychology*, *5*(JUL), 1–10. doi: 10.3389/fpsyg.2014.00674
- Xu, Y., & Vaziri-Pashkam, M. (2021). Examining the coding strength of object identity and nonidentity features in human occipito-temporal cortex and convolutional neural networks. *Journal of Neuroscience*, *41*(19), 4234–4252. doi: 10.1523/JNEUROSCI.1993-20.2021
- Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. doi: 10.1038/nn.4244
- Yamins, D. L. K., Hong, H., Cadieu, C. F., Solomon, E. A., Seibert, D., & DiCarlo, J. J. (2014). Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(23), 8619–8624. doi: 10.1073/pnas.1403112111
- Zachariou, V., Nikas, C. V., Safiullah, Z. N., Behrmann, M., Klatzky, R. L., & Ungerleider, L. G. (2015). Common dorsal stream substrates for the mapping of surface texture to object parts and visual spatial processing. *Journal of Cognitive Neuroscience*, *27*(12), 2442–2461. doi: 10.1162/jocn_a_00871
- Zeki, S., & Marini, L. (1998). Three cortical stages of colour processing in the human brain. *Brain*, *121*(9), 1669–1685. doi: 10.1093/brain/121.9.1669

Appendix

Original publication of Study 1







Graumann, M., Ciuffi, C., Dwivedi, K., Roig, G. & Cichy, R. M. (2022). The spatiotemporal neural dynamics of object location representations in the human brain. *Nature Human Behaviour*. doi: [10.1038/s41562-022-01302-0](https://doi.org/10.1038/s41562-022-01302-0).

This article is licensed under a Creative Commons Attribution 4.0 International License creativecommons.org/licenses/by/4.0/, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original authors and the source, provide a link to the Creative Commons license, and indicate if changes were made.



OPEN

The spatiotemporal neural dynamics of object location representations in the human brain

Monika Graumann^{1,2}  , Caterina Ciuffi¹, Kshitij Dwivedi^{1,3} , Gemma Roig³  and Radoslaw M. Cichy^{1,2,4}  

To interact with objects in complex environments, we must know what they are and where they are in spite of challenging viewing conditions. Here, we investigated where, how and when representations of object location and category emerge in the human brain when objects appear on cluttered natural scene images using a combination of functional magnetic resonance imaging, electroencephalography and computational models. We found location representations to emerge along the ventral visual stream towards lateral occipital complex, mirrored by gradual emergence in deep neural networks. Time-resolved analysis suggested that computing object location representations involves recurrent processing in high-level visual cortex. Object category representations also emerged gradually along the ventral visual stream, with evidence for recurrent computations. These results resolve the spatiotemporal dynamics of the ventral visual stream that give rise to representations of where and what objects are present in a scene under challenging viewing conditions.

To interact with objects in our environments, the two arguably most basic questions that our brains must answer are what objects are present and where they are. To address the first question and identify an object, we must recognize objects independently of the viewing conditions of a given scene, such as where the object is located. A large body of research has shown that the ventral visual stream^{1–4}, a hierarchically interconnected set of regions, achieves this by transforming retinal input in successive stages marked by increasing tolerance and complexity. At its high stages in high-level ventral visual cortex, object representations are tolerant to changes in retinotopic location^{5–7}.

In contrast, we know considerably less about how the brain determines where an object is located. Current empirical data imply three different theoretical accounts.

One hypothesis (H1) is that object location representations are already present at the early stages of visual processing (H1, Fig. 1a) and thus no further computation is required. Given the idea that ventral stream representations become successively more tolerant to changes in viewing conditions such as location¹, it seems plausible that object location representations are to be found at the early stages of the processing hierarchy. Consistent with this view, human studies using multivariate analysis have shown that object location is often strongest in early visual cortex^{8,9}, likely related to its small receptive field size which allows for spatial coding with high resolution¹⁰.

An alternative account (H2) is that location representations emerge in the dorsal visual stream (H2, Fig. 1a)¹¹. This view is supported by findings from neuropsychology^{2,4,11} and by studies finding object location information along the dorsal pathway^{2,12}.

A third possibility is that location representations emerge through extensive processing but in the ventral visual stream (H3, Fig. 1a). This view receives support from the observation that object location information was found across the entire ventral visual stream including high-level ventral visual cortex in human^{5,8,9,13} and non-human primates¹⁴. In line with these observations, high-level

ventral visual cortex is known to be retinotopically organized^{15–17} and exhibits an eccentricity bias^{18–20}.

How can we adjudicate between these hypotheses given the mixed empirical support? We propose that it is key to acknowledge the importance of assessing object location representations under conditions that increase the complexity of the visual scene to increase ecological validity. Previous research typically investigated object location representations by presenting cut-out objects on blank backgrounds. This creates a direct mapping between the location of visual stimulation and the active portions of retinotopically organized cortex (Fig. 1b, left). In contrast, in daily life, objects appear on backgrounds cluttered by other elements^{21,22}. This activates a large swath of cortex, independently of where the object is (Fig. 1b, right). Whereas in the former case location information can be directly accessible through retinotopic activation in early visual areas (supporting H1), in the latter case additional processing might be required to distil out location information (supporting H2 or H3).

Taking the importance of background into consideration, we used a combination of methods to distinguish between the proposed theoretical hypotheses. We used functional MRI (fMRI), deep neural networks (DNNs) and electroencephalography (EEG) to assess where, how and when location representations emerge in the human brain. We quantified the presence of location representations by the performance of a multivariate pattern classifier to predict object location from brain measurements.

Assessed in this way, the predictions for the hypotheses are as follows: If H1 is correct, independent of the nature of the object's background, object location information peaks in early visual cortex (Fig. 1c, left), early in the DNN processing hierarchy (Fig. 1d, left) and early during visual processing (Fig. 1e, left). For H2 and H3, the prediction of peak location information depends on the background. For cut-out isolated objects, location information is high across the entire dorsal and ventral pathways, and the processing hierarchy of the DNN (Fig. 1c,d, middle and right, grey).

¹Department of Education and Psychology, Freie Universität Berlin, Berlin, Germany. ²Berlin School of Mind and Brain, Faculty of Philosophy, Humboldt-Universität zu Berlin, Berlin, Germany. ³Department of Computer Science, Goethe Universität, Frankfurt am Main, Germany. ⁴Bernstein Center for Computational Neuroscience Berlin, Berlin, Germany. [✉]e-mail: monika.graumann@fu-berlin.de; rmcichy@zedat.fu-berlin.de

In contrast, for objects appearing on cluttered backgrounds, object location information emerges late in the DNN hierarchy (Fig. 1d, right, blue) and late in time (Fig. 1e, middle and right, blue). H2 and H3 differ in predicting location information to peak in dorsal (Fig. 1c, middle, blue) or ventral visual cortex (Fig. 1c, right, blue), respectively.

To anticipate, our results strongly support H3. When objects appear on cluttered backgrounds, object location representations emerge late in the hierarchy of the ventral visual stream and of the DNN, as well as late in time, indicating recurrent processing. A corresponding analysis of object category representations revealed an equivalent pattern of results with emergence along the ventral visual stream and temporal dynamics suggesting recurrence. Taken together, our results resolve where, when and how object representations emerge in the human brain when objects are viewed under more challenging viewing conditions.

Results

To investigate where, how and when representations of object location emerge in the brain, we created a visual stimulus set (Fig. 2a) with the three orthogonal factors objects (three exemplars each in four object categories), locations (four quadrants) and backgrounds (three kinds: uniform grey, low- and high-cluttered natural scenes, referred to as 'no', 'low' and 'high' clutter). Collapsing across exemplars, we used a fully crossed design with four categories \times four locations \times three background conditions, resulting in 48 stimulus conditions. This design allowed us to also investigate representations of object category as a secondary question of the study.

To resolve human brain responses with high spatial and temporal resolution, participants viewed images from the stimulus set while we recorded fMRI ($N=25$) and EEG ($N=27$) data in separate sessions. Experimental parameters were optimized for each

imaging modality (Fig. 2b). On each trial, participants viewed individual stimuli while fixating on a central fixation cross and performing a one-back (fMRI) or a detection task (EEG) to direct participants' attention towards the images (Fig. 2b). Response trials were excluded from analysis.

We used multivariate pattern classification to track the emergence of object location representations. We consider the peaks in information, that is, in classification, as indicators of where (fMRI) and when (EEG) location representations become most untangled and are thus explicitly represented¹. In each case, we trained a support vector machine (SVM) to pairwise classify between activation patterns belonging to one object category shown at two different locations (Fig. 3a, faces at bottom left and right). We then tested the SVM on activation patterns of the same locations with a new object category (Fig. 3a, animals at bottom left and right). Repeated for all combinations of locations and categories, the averaged classification

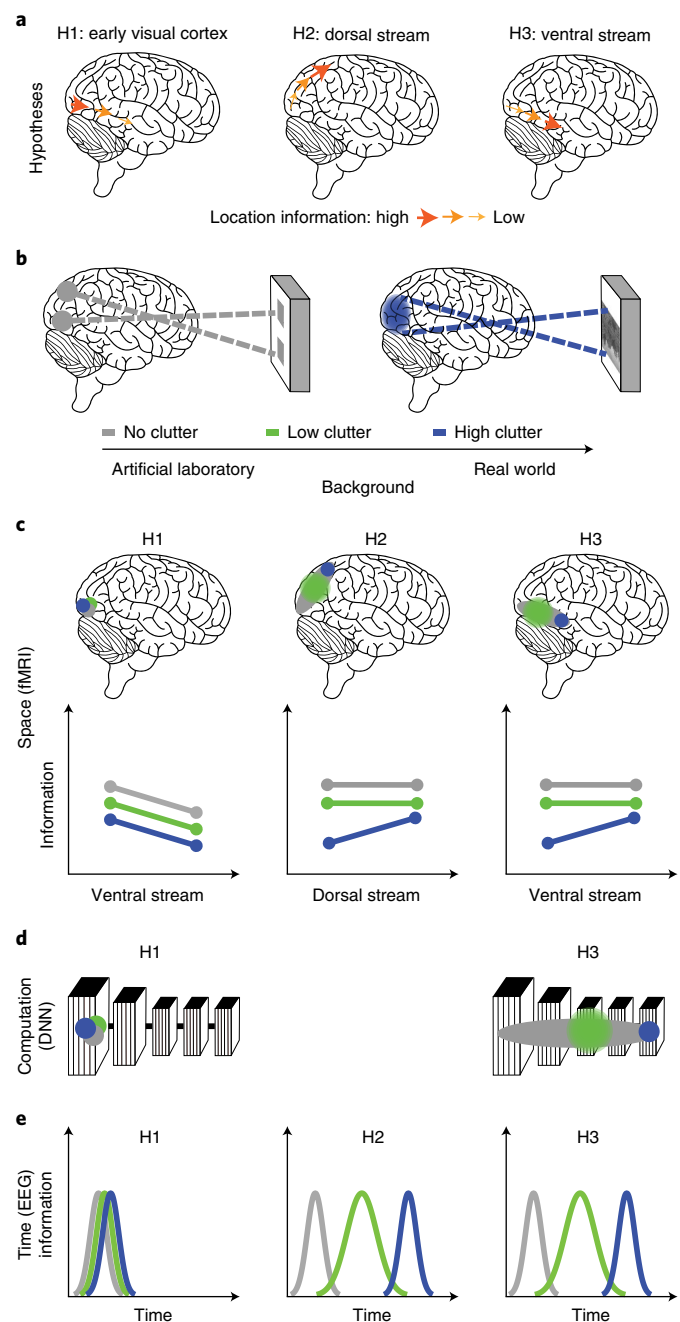


Fig. 1 | Hypotheses and predictions about the pathway of object location representations in the human brain. **a**, H1: representations of object location emerge in early visual cortex and degrade along further processing stages. H2 and H3: object location representations emerge gradually along the dorsal (H2) or ventral (H3) visual stream. **b**, Left: when objects are presented on a blank background, object location in the visual field maps retinotopically onto early visual cortex, allowing for direct location read-out (grey). Right: when objects appear in a cluttered scene, large parts of early visual cortex are activated, hindering a direct read-out (blue). Representations are quantified as linearly classifiable object location information from brain or model activity patterns¹. **c**, Predictions in space, colour-coded by background condition: no (grey), low (green) and high (blue) clutter. H1 predicts that independent of the object's background, location information for the object is highest in early processing stages in space. H2 and H3 predict similar levels of location information with no clutter across the entire processing pathway in all assessments.

For highly cluttered backgrounds, H2 and H3 predict the emergence of location representations in late processing stages of the dorsal (H2, **c**) and ventral (H3, **c**) stream. Location information in the low-clutter condition is expected to be in between the no- and the high-clutter condition. **d**, Computational model of the ventral visual stream. H1 (left) predicts highest location information in early layers of the model in all conditions. H3 (right) predicts high location information across all layers with no clutter and highest location information in late layers with high clutter. Location information in the low-clutter condition is expected to be in between the other two conditions. Since this is a model of the ventral stream, it does not make predictions about the dorsal stream (H2). **e**, Location information in time. H1 predicts that location information peaks early in time in all conditions. Both H2 and H3 predict an early peak with no and a late peak with high clutter. The peak for low clutter is expected to be in between no and high clutter.

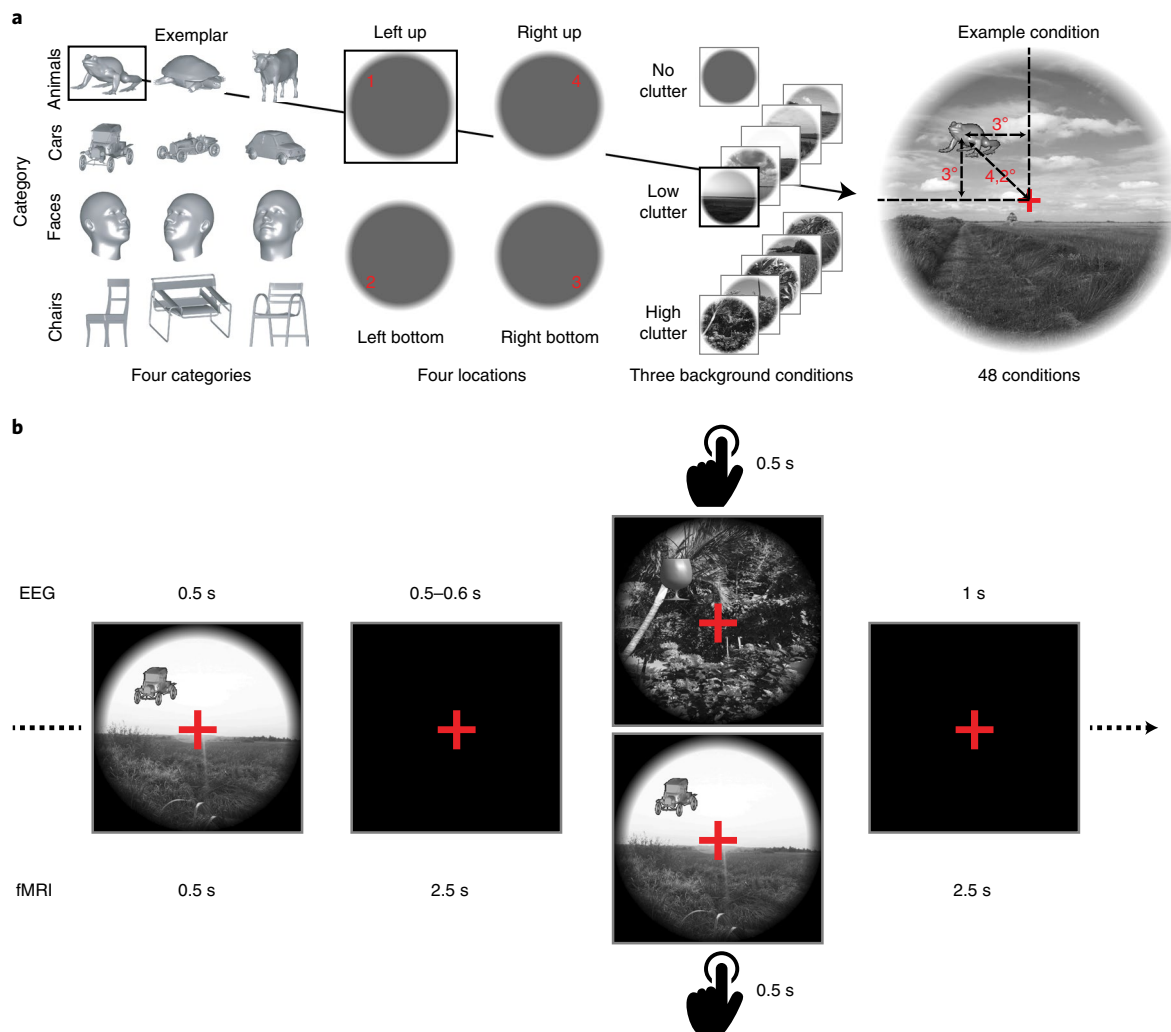


Fig. 2 | Experimental design and tasks. a, Experimental design. We used a fully crossed design with factors of object category, location and background. Note that, for copyright reasons, all example backgrounds shown are for illustrative purposes and were not used in the experiment. **b**, Tasks. The experimental design was adapted to the specifics of each modality by adjusting the interstimulus interval. On each trial, participants viewed images for 500 ms followed by a blank interval (0.5–0.6 s in EEG, 2.5 s in fMRI). The task was to respond with button press to catch trials that were presented on every fourth trial on average. Catch trials were marked by the presence of a probe (glass) in the EEG experiment and by an image repetition (one-back) in the fMRI experiment. Image presentation was followed by blank screen (1 s in EEG, 2.5 s in fMRI).

accuracy quantifies object location information independent of object category. This procedure was performed in a space-resolved fashion for fMRI and in a time-resolved fashion for EEG (see Supplementary Fig. 1a for details).

The locus of object location representations. To determine the locus of object location representations, we used a regions of interest (ROI) fMRI analysis, including early visual regions (V1, V2 and V3) shared to the hierarchy of the ventral (V4 and LOC²³) and the dorsal visual stream (intraparietal sulcus: IPS0, IPS1, IPS2 and superior parietal lobule (SPL)).

As expected, we found that most regions contained above-chance level location information in all background clutter conditions (Fig. 3b; $N=25$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected; see Supplementary Table 1 for P values). However, the amount of location information depended critically on the brain region and background condition.

Focusing on the ventral visual stream first, we observed similar amounts of location information across regions when objects were presented without clutter (Fig. 3b, grey bars). In contrast, when

objects were presented on cluttered backgrounds, location information emerged along the ventral visual processing hierarchy with less information in early visual areas than in LOC (Fig. 3b, green and blue bars; $N=25$, 5×3 repeated-measures ANOVA, post hoc t tests Tukey corrected; see Supplementary Table 2 for P values). These results are at odds with H1, which predicts that location information decreases along the ventral stream independent of background condition. Instead, the observed increase of location information along the ventral visual stream with cluttered backgrounds is consistent with H3.

We ascertained these observations statistically with a 5×3 repeated-measures ANOVA with factors ROI (V1, V2, V3, V4 and LOC) and background (no, low and high clutter). Besides both main effects (ROI: $F_{(4,96)}=18.30$, $P<0.001$, partial $\eta^2=0.43$; background: $F_{(1,44,34,48)}=64.11$, $P<0.001$, partial $\eta^2=0.73$), we crucially found the interaction to be significant ($F_{(8,192)}=5.40$, $P<0.001$, partial $\eta^2=0.18$). As the interaction makes the main effects difficult to interpret, we conducted post hoc paired t tests (all reported in Supplementary Table 2, Tukey corrected). The statistical analysis confirmed all the qualitative observations: There were no significant

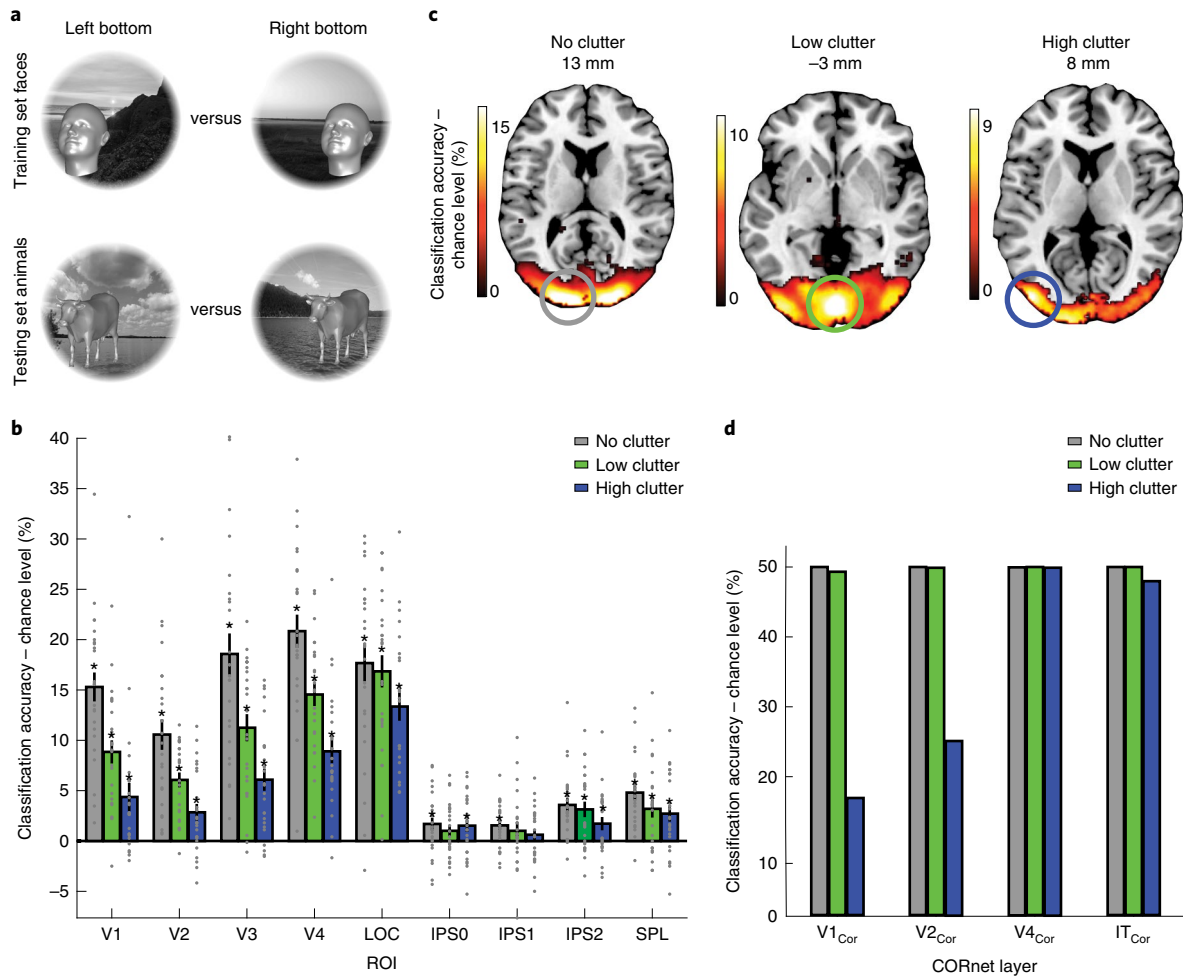


Fig. 3 | fMRI results of location classification. **a**, Classification scheme for object location across category. We trained an SVM to distinguish between brain activation patterns evoked by objects of a particular category presented at two locations (here: faces bottom left and right) and tested the SVM on activation patterns evoked by objects of another category (here: animals) presented at the same locations. Objects are enlarged for visibility and did not extend into another quadrant in the original stimuli. **b**, Location classification in early visual cortex, ventral and dorsal visual ROIs ($N=25$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected). With no clutter, location information was high across early visual cortex and ventral ROIs. In the low- and high-clutter conditions, location representations emerged gradually along the ventral stream. In dorsal ROIs, location information was low, independent of background condition. Stars above bars indicate significance above chance (see Supplementary Tables 1, 2 and 3 for P values). Error bars represent s.e.m. Dots represent single subject data. **c**, fMRI searchlight result for classification of object location ($N=25$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected). Peak classification accuracy is indicated by colour-coded circles (no clutter: left V3 (grey, XYZ coordinates -19 mm, -97 mm, 13 mm); low clutter: left V1 (green, -5 mm, -86 mm, -3 mm); high clutter: left LOC (blue, -44 mm, -83 mm, 8 mm)). Millimetres (mm) indicate axial slice position along z axis in Montreal Neurological Institute space. **d**, Location classification in a DNN. In the high-clutter condition, location information emerged along the processing hierarchy, analogous to the ventral visual stream.

differences between ROIs in the no-clutter condition, except between V2 and V3 ($P=0.009$) and between V2 and V4 ($P=0.001$). There was more location information in LOC than in V1, V2 and V3 when background clutter (both low and high) was present than when it was not (Fig. 3b; all $P < 0.03$, see Supplementary Table 2 for P values, Tukey corrected). This effect was robust for the comparison of locations across, but not within, visual hemifields (Fig. 4a,b): post hoc tests comparing early visual areas versus LOC in the high-clutter condition were significant for the cross-hemifield classification (Fig. 4a; V1: $P=0.003$; V2: $P < 0.001$; V3: $P=0.004$, Tukey corrected), but not for the within-hemifield classification (Fig. 4b; V1: $P=0.697$; V2: $P=0.281$; V3: $P=1.00$, Tukey corrected).

Focusing next on the dorsal visual stream, we observed low object location information independent of background condition (Fig. 3b; $N=25$, 7×3 repeated-measures ANOVA). In the no- and low-clutter conditions, location information was higher in early

visual cortex than in dorsal regions ($N=25$, post hoc t tests, Tukey corrected; see Supplementary Table 3 for P values). This is inconsistent with H2, which predicts an increase of object location information along the dorsal stream.

Consistent with these qualitative observations, statistical testing by 7×3 repeated-measures ANOVA with factors ROI (V1, V2, V3, IPS0, IPS1, IPS2 and SPL) and background (no, low and high clutter) did not provide statistical evidence for H2. We found significant main (ROI: $F_{(3,16,75,93)} = 36.2$, $P < 0.001$, partial $\eta^2 = 0.60$; background: $F_{(2,48)} = 35.8$, $P < 0.001$, partial $\eta^2 = 0.60$) and interaction effects ($F_{(6,25,149,89)} = 14.5$, $P < 0.001$, partial $\eta^2 = 0.38$). The post hoc tests showed that location information was higher in V1, V2 and V3 compared with dorsal regions in the no- and low-clutter conditions (Fig. 3b, grey and green, except V1 and V2 versus IPS2 and SPL with low clutter, which were n.s.; see Supplementary Table 3 for P values). With high clutter, there was more location information in V3

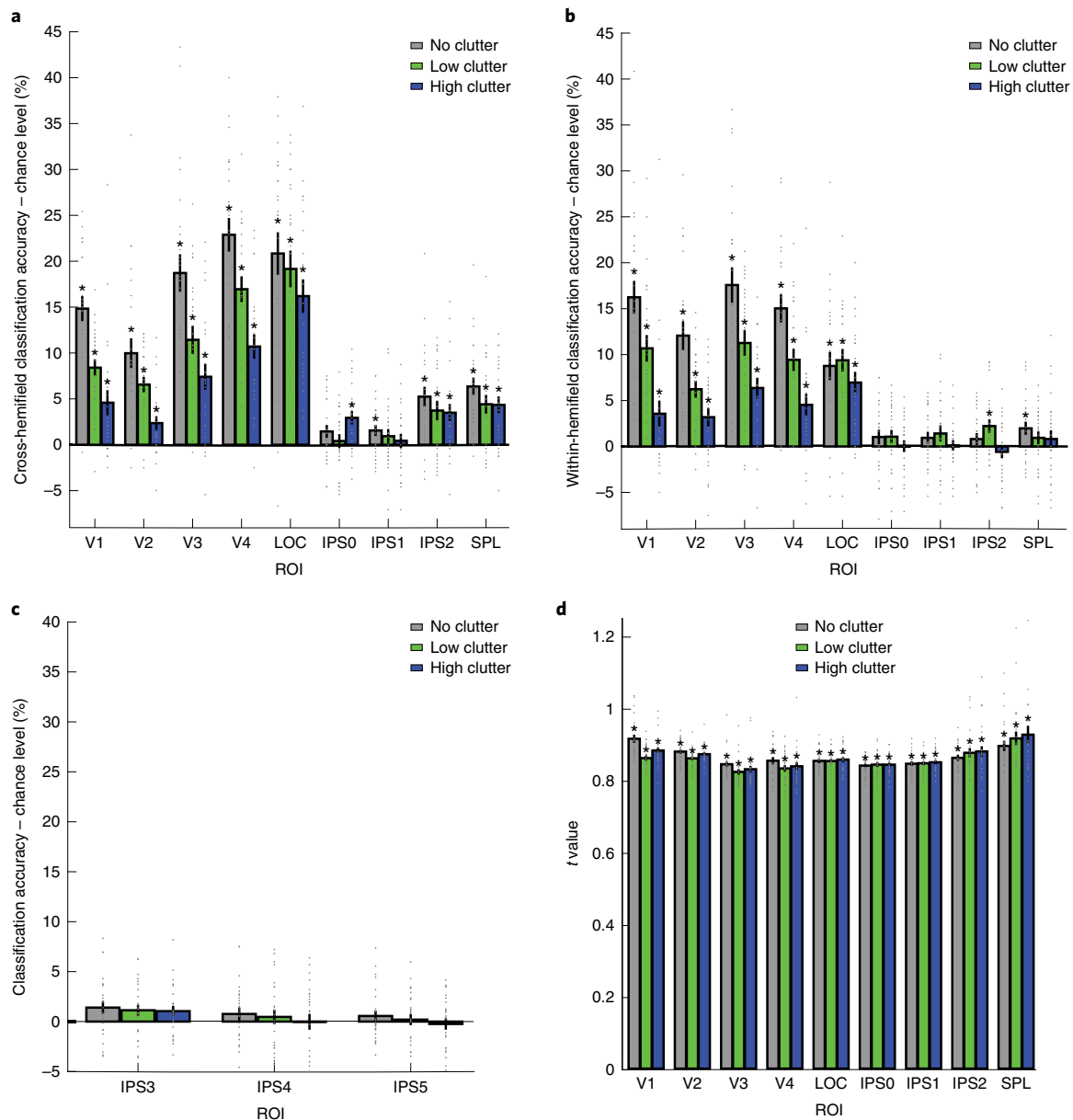


Fig. 4 | Location classification within and across hemifields, in IPS3-5 and univariate ROI results. **a**, Results of location classification across categories between visual hemifields (left up versus right up, left bottom versus right bottom). Similar to the classification across four locations, the repeated-measures ANOVA along the ventral stream (five ROIs \times three clutter levels) yielded significant main (ROI: $F_{(4,96)} = 24.62, P < 0.001$, partial $\eta^2 = 0.51$; background: $F_{(1,49,35,85)} = 45.34, P < 0.001$, partial $\eta^2 = 0.65$) and interaction effects ($F_{(8,192)} = 2.95, P = 0.004$, partial $\eta^2 = 0.11$). Post hoc tests yielded results comparable to the main results (V1, V2 and V3 $<$ LOC with high clutter). Stars above bars indicate significance above chance ($N = 25$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected). **b**, Location classification across categories within visual hemifields (left up versus left bottom, right up versus right bottom). As for the main analysis, the ANOVA yielded significant main (ROI: $F_{(4,96)} = 4.16, P = 0.004$, partial $\eta^2 = 0.15$; background: $F_{(1,60,38,43)} = 57.90, P < 0.001$, partial $\eta^2 = 0.71$) and interaction effects ($F_{(8,192)} = 5.84, P < 0.001$, partial $\eta^2 = 0.20$). The post hoc tests revealed a significant difference between V3 and LOC in the noclutter condition ($P = 0.030$). Stars above bars indicate significance above chance ($N = 25$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected). **c**, Classification accuracies in IPS3, IPS4 and IPS5 were not significantly higher than chance level in all background conditions ($N = 25$, two-sided Wilcoxon signed-rank test, $P > 0.05$, FDR corrected). Error bars represent s.e.m. Dots represent single/subject data. **d**, Absolute t values in each background condition and ROI, averaged across locations and categories. A 9×3 repeated-measures ANOVA with factors ROI and clutter revealed a significant main effect of ROI ($F_{(2,60,62,43)} = 9.19, P < 0.001$, partial $\eta^2 = 0.18$) and a significant interaction effect ($F_{(3,40,81,64)} = 9.89, P < 0.001$, partial $\eta^2 = 0.03$). Significant post hoc tests are listed in Supplementary Table 4. Overall, post hoc tests showed no clear pattern of results between early, ventral and dorsal areas, except for higher activation in V1 than in dorsal areas and LOC with no clutter. Stars above bars indicate significance above chance ($N = 25$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected).

than in IPS0, IPS1 and IPS2. Location classification in IPS3, IPS4 and IPS5 did not reveal significant information above chance level (Fig. 4c; $N = 25$, two-tailed Wilcoxon signed-rank test, all $P > 0.05$ FDR corrected, see Supplementary Table 1 for P values). Univariate

responses were comparable across regions overall (Fig. 4d). Post hoc tests to a 9×3 repeated-measures ANOVA with factors ROI (V1, V2, V3, V4, LOC, IPS0, IPS1, IPS2 and SPL) and background (no, low and high clutter) revealed that responses were significantly

higher in V1 compared with the other ROIs in the no-clutter condition (all $P < 0.03$; all P values listed in Supplementary Table 4, Tukey corrected), but there was no significant difference in activation between LOC and dorsal areas (Fig. 4d, all P values in Supplementary Table 4, Tukey corrected).

To explore whether any other brain regions beyond the investigated ROIs contain location information, we used a spatially unbiased fMRI searchlight analysis²⁴. We did not find statistical evidence for location information beyond the ventral and dorsal stream, and the pattern of results was consistent with the outcome of the ROI analysis (Supplementary Fig. 2). There was widespread location information ($N=25$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected) from the occipital cortex up into the dorsal (precuneus, superior parietal lobule) and ventral (fusiform gyrus) visual stream. Depending on background condition, location information peaked in different visual areas. In the no-clutter condition, the peak was in left V3, in the low-clutter condition in left V1 and in the high-clutter condition in left LOC (Fig. 3c, see caption for coordinates). Distances between peaks were significantly larger than chance ($N=25$, bootstrapping of condition labels, 10,000 bootstraps, $P < 0.05$ one-tailed bootstrap test against chance level, Bonferroni corrected) between the no- and the high-clutter condition (Euclidean distance 15.9, CI 1.0–3.6, $P < 0.001$) and between the low- and the high-clutter condition (Euclidean distance 22.0, CI 2.0–16.3, $P = 0.002$), but not for the no- and low-clutter condition (Euclidean distance 13.6, CI 1.4–14.7, $P = 0.275$).

Together, these results provide consistent evidence for the hypothesis that representations of object location across visual hemifields emerge in the ventral visual stream (H3) when objects appear in cluttered scenes.

Computational modelling. DNNs trained on object categorization are currently the best predicting models of ventral visual stream representations^{25–27} and show a spatiotemporal correspondence in their processing hierarchy to the visual brain^{25,28–30}. Therefore, they constitute feasible biologically inspired models for computing complex visual representations^{28,31}. If such DNNs are useful models of visual processing in human visual cortex, they should show a similar pattern of results as the ventral visual stream in the representation of object location, too.

To evaluate this prediction, we chose the recurrent CORnet-S model because it is among the best-performing models on a benchmark for predicting neural responses in monkey inferior temporal cortex (IT)^{26,27} and approximates explicitly the hierarchy of the ventral visual system. Each region of the ventral stream is modelled as one processing block with a corresponding name (V1_{Cor}, V2_{Cor}, etc.). Analogous to the fMRI analysis, we extracted the unit activation patterns to our stimulus set at the last layer of each block and classified object location across category to identify the processing stage of the DNN at which object location representations emerge (Fig. 3d).

We found that in the no- and low-clutter conditions, location information was at or close to ceiling in all layers. In the high-clutter condition however, location information was low in V1_{Cor} and emerged along the processing hierarchy. Qualitatively equivalent results were obtained in three other DNNs (Alexnet, ResNet-50 and CORnet-Z; Supplementary Fig. 3a–c), demonstrating the generalizability of the results pattern. This result was still robust in all four DNNs when limiting the classification to either horizontal or vertical location comparisons (Supplementary Fig. 3e,f).

In sum, we found that DNNs trained on object categorization show a similar pattern of location representations along their processing hierarchy as the human brain. This demonstrates how object location representations might be computed in biological systems. This result lends independent evidence against H1 and yields plausibility to H3 since CORnet-S was built to model the ventral stream. However, this result cannot disambiguate between H2 and H3, as

models of this kind have been found to predict human brain activity in both the ventral and dorsal stream^{32,33}.

Temporal dynamics of object location representations. We conducted time-resolved multivariate EEG analysis to determine the time course with which object location representations emerge. The general analysis scheme was the same as for the fMRI analysis presented above (Fig. 3a) but applied to time-specific EEG channel activation patterns rather than fMRI activation patterns.

The analysis revealed location information for all background clutter levels (Fig. 5a, $N=27$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected), but with different dynamics (Fig. 5b, see Supplementary Table 5 for classification onsets and peak values). We report peak latencies with 95% confidence intervals ($N=27$, 10,000 bootstraps). Whereas the peak latency was similar for the no- (140 ms (133–147 ms)) and the low-clutter (133 ms (121–233 ms)) condition, it was delayed in the high-clutter condition (317 ms (250–336 ms)). Statistical analysis ($N=27$, bootstrap test, 10,000 bootstraps, $P < 0.05$, one-tailed bootstrap test against zero, FDR corrected) ascertained that the peak latency difference was significant between the high-clutter and the no-clutter conditions ($N=27$, 177 ms (94–190 ms), $P < 0.001$) and between the high- and the low-clutter conditions (184 ms (16–196 ms), $P = 0.023$), but not between the no- and the low-clutter conditions (7 ms (–11–156 ms), $P = 0.620$). These delays were also robust when classifying locations across or within visual hemifields (Supplementary Fig. 4). A searchlight in EEG sensor space showed that location information at the peaks of the three background conditions was highest at occipital, occipito-parietal and occipito-temporal electrodes (Fig. 5c; $N=27$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected across electrodes and time points; see Supplementary Fig. 5a–c for time courses), suggesting sources in those areas, which is in line with the fMRI searchlight results (Supplementary Fig. 2) and with univariate EEG topographies (Supplementary Fig. 5d–f).

In sum, this result shows that object location representations emerge later when objects appear on cluttered backgrounds than when they appear on blank backgrounds. This provides further concurrent evidence against H1 and is consistent with H2 and H3, that is, that object location representations emerge at late stages of visual processing when objects are viewed under complex visual conditions.

How is the delay in the peak latencies of the no- and the high-clutter conditions to be interpreted? Assuming that in object processing the brain runs through a series of distinct stages, we see two possible explanations.

One explanation is that the peak latency delay indicates a change in the processing stage at which object location representations emerge. This would mean that in the no-clutter condition, location representations emerge in an early stage whereas with high clutter they emerge during a different, later processing stage (the ‘change’ hypothesis). An alternative explanation is that the processing stage at which object location representations emerge remains the same, but its emergence is delayed in time in the high-clutter condition (the ‘delay’ hypothesis).

To distinguish between these explanations, we used temporal generalization analysis³⁴, comparing the representational dynamics with which object location representations emerge in the no- and the high-clutter conditions across time (Fig. 5d). Used in this way, the time generalization analysis yields a two-dimensional matrix indexed in time, indicating at which time points location representations in the no- and the high-clutter conditions are similar. We implemented time generalization by classifying object location across category and background condition for all time point combinations (Fig. 5d and Supplementary Fig. 1b). Overall, we observed a large significant cluster of above-chance classification accuracies across the time generalization matrix ($N=27$, two-tailed Wilcoxon

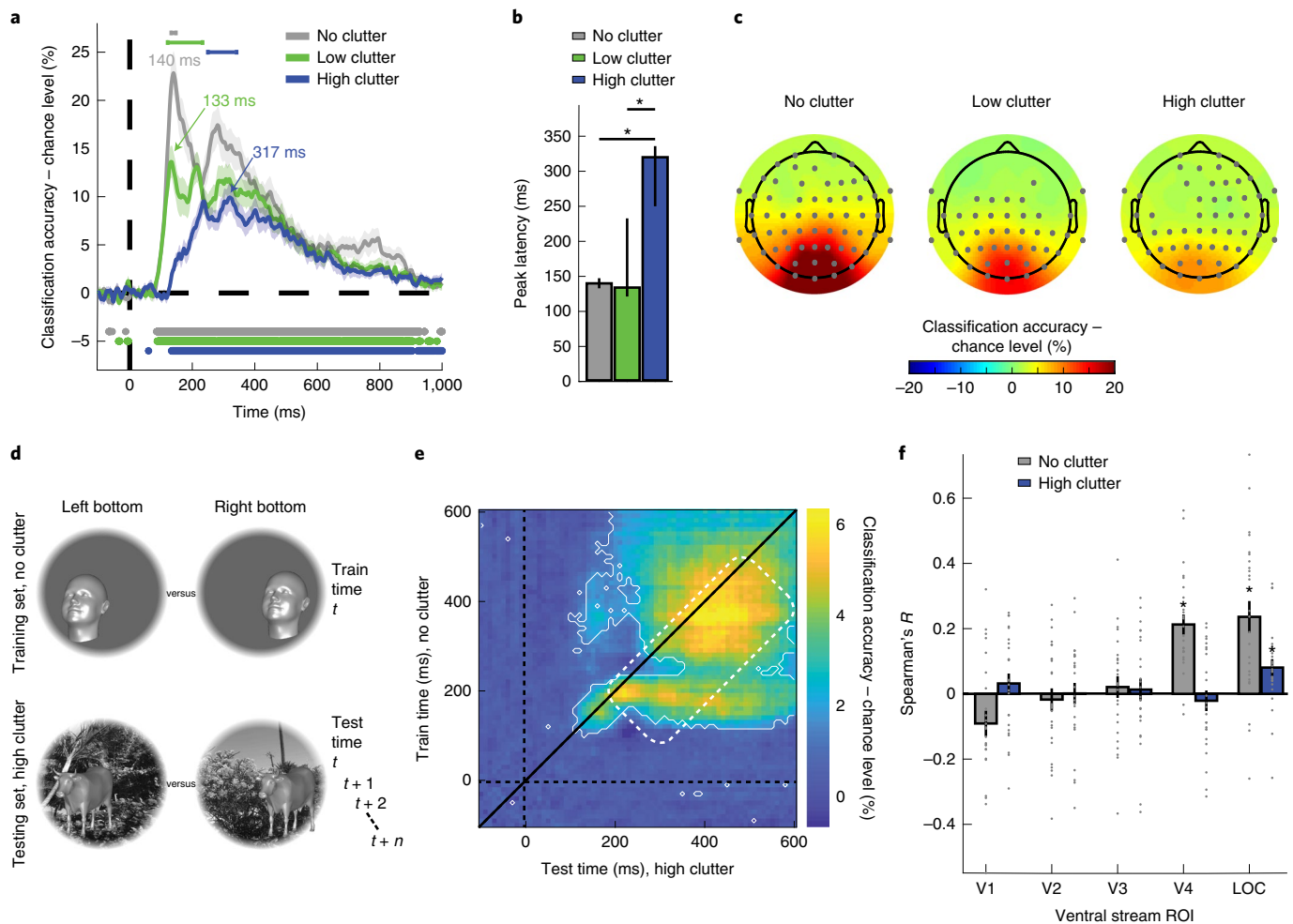


Fig. 5 | Temporal dynamics of object location representations. **a**, Results of time-resolved location classification across category from EEG data. Results are colour coded by background condition, with significant time points indicated by lines below curves ($N=27$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected), 95% CI of peak latencies indicated by lines above curves. Shaded areas around curves indicate *s.e.m.* Inset text at arrows indicates peak latency (140 ms, 133 ms and 317 ms in the no-, low- and high-clutter condition, respectively). **b**, Comparison of peak latencies of curves in **a**. Error bars represent 95% CI. Stars indicate significant peak latency differences ($P<0.05$; $N=27$, bootstrap test with 10,000 bootstraps). **c**, Results of location across category classification searchlight in EEG channel space at peak latencies in no-, low- and high-clutter condition, down-sampled to 10 ms steps. Significant electrodes are marked in grey ($N=27$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected across electrodes and time points). **d**, Time generalization analysis scheme for classifying object location across category and background condition. The classification scheme was the same as in **a** with the differences that (i) the training set conditions always came from the no-clutter while the testing set conditions came from the high-clutter condition and (ii) training and testing was repeated across all combinations of time points for a peri-stimulus time window between -100 and 600 ms (see Supplementary Fig. 1b for details). Objects are enlarged for visibility and did not extend into another quadrant in the original stimuli. **e**, Results of the time generalization analysis. Dashed black lines indicate stimulus onset; oblique black line highlights the diagonal. Solid white outlines indicate significant time points ($N=27$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected). Dashed white outline highlights delayed clusters. **f**, EEG-fMRI fusion. Results represent the correlations between single-subject fMRI RDVs of classification accuracies and group-averaged RDVs of the EEG peaks in **a**. Stars above bars indicate significance above chance ($N=27$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected). Error bars represent the *s.e.m.* Dots represent single-subject data points.

signed-rank test, $P<0.05$, FDR corrected). While the ‘change’ hypothesis predicts highest classification accuracies on the diagonal, the ‘delay’ hypothesis predicts highest classification accuracies below the diagonal. The results are reported in Fig. 5e. We found that peak latencies in location information as tested across subjects were significantly shifted below the diagonal (mean Euclidean distance 56.31 ms; $N=27$, two-tailed Wilcoxon signed-rank test, $P<0.001$, $r=0.65$, *s.e.m.* 1.55; see Supplementary Fig. 6a for single-subject peaks), indicating that location representations in the no-clutter condition generalized to the high-clutter condition at later time points (Fig. 5e, white dashed outline). This result was confirmed

in a supplementary analysis on the group-averaged peak in Fig. 5e (Euclidean distance 49.50 ms; 10,000 bootstraps; one-tailed bootstrap test against zero, $P=0.010$; 95% CI 14.14–77.78). Classification accuracies were significantly higher below than above the diagonal between ~ 120 and 240 ms in the no-clutter condition and from ~ 200 ms in the high-clutter condition ($N=27$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected; Supplementary Fig. 6b).

Together, these results provide evidence for the ‘delay’ hypothesis and demonstrate that object location representations in the no- and the high-clutter condition emerge at the same processing stage with a temporal delay.

Spatiotemporal similarity of location representations. Temporal delays for the same processing stage cannot be explained by a purely feedforward process, suggesting instead the involvement of recurrent processing. Recurrent processes could account for the observed delay with lateral connections within the same area^{35,36}. The shared processing stage underlying early and late location representations in the no- and the high-clutter conditions should have a common origin in space, too. Based on the fMRI results, we hypothesized that this origin would be in LOC. To test this hypothesis directly, we used EEG–fMRI fusion based on representational similarity of object location representations^{37–39}.

The processing stage at which location representations emerge corresponds to the peak latency of location classification in the EEG for the no- and the high-clutter condition. We thus determined whether location representations identified with EEG at these time points are representationally similar to those identified with fMRI in ventral stream regions for the no- and the high-clutter condition separately. Specifically, we averaged the representational dissimilarity vectors (RDVs) of the time-resolved EEG classification accuracies in Fig. 5a across subjects and time points within the 95% confidence intervals over the peaks. This yielded one RDV per background condition that was then correlated with the single-subject RDV of an fMRI ROI in the same background condition. Results within background and ROI were averaged across fMRI participants.

We found a spatiotemporal correspondence with EEG peak latency for the no-clutter condition in V4 and LOC but for the high-clutter condition in LOC only (Fig. 5f; $N=25$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected). This establishes LOC as the cortical locus at which object location representations emerge independent of background condition, but involving additional recurrent processing when the background is cluttered. Post hoc tests to a 5×2 repeated-measures ANOVA with factors ROI (V1, V2, V3, V4 and LOC) and clutter (no and high) additionally showed that correlations were higher in V4 and LOC than in V1, V2 and V3 with no clutter (see Supplementary Table 6 for P values; main effect of ROI: $F_{(4,96)}=14.30$, $P<0.001$, partial $\eta^2=0.37$; n.s. main effect of background: $F_{(1,24)}=3.62$, $P=0.069$; interaction: $F_{(4,96)}=8.17$, $P<0.001$, partial $\eta^2=0.25$). The notion that location representations emerge in LOC with recurrence when background is cluttered finds further support from a supplementary analysis showing that location representations with no and high clutter were significantly similar in LOC, but not in other regions (Supplementary Fig. 7; $N=25$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected). Furthermore, recurrent DNNs showed an advantage compared with shallow feedforward DNNs for the classification of location with high clutter and for the prediction of location representations in LOC (Supplementary Fig. 3c,d; $N=25$, 4×2 repeated-measures ANOVA). Together, these results suggest that location information of objects on highly cluttered scenes emerges in LOC with local recurrent processes.

Object category representations. The observation that representations of object location depend on the background on which the object appears immediately raises the question of whether representations of object category are affected by background, too. Previous research suggests opposite answers to this question. One line of research demonstrated that object representations in the ventral stream are modulated by the presence of other objects and the background on which they are viewed^{40–43}. Another line of research has provided strong evidence that the ventral stream constructs object representations that are increasingly tolerant to changes in viewing conditions^{1,5,8}, suggesting that object category representations should be unaffected by the background of the objects. Here we bring these two lines of research together by explicitly investigating how background impacts object category representations that are tolerant to location. To do this, we analysed EEG and fMRI data as

described in previous sections but exchanging the role of experimental factors location and category. In essence, we performed cross-classification analyses of category across location (Fig. 6a) to determine where and when location-tolerant object category representations emerge in the human brain.

The locus of object category representations. We investigated object category representations tolerant to changes in location using an ROI-based fMRI analysis. We observed that location-tolerant object category could be classified in the ventral stream in V4 and LOC (Fig. 6b; $N=25$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected, all P values in Supplementary Table 1), but not at earlier stages and not in dorsal ROIs except IPS0 with high clutter ($P=0.005$). This pattern was not influenced by the level of clutter, suggesting that object category representations that are tolerant to location variations are unaffected by the clutter level of the background on which the object appears.

These observations were statistically ascertained in a 5×3 ANOVA along the ventral stream with factors ventral ROIs (V1, V2, V3, V4 and LOC) and background (no, low and high clutter), revealing a significant main effect of ROI ($F_{(2,42,58,03)}=21.97$, $P<0.001$, partial $\eta^2=0.48$), but not of background ($F_{(2,48)}=0.68$, $P=0.510$) and no interaction ($F_{(8,192)}=1.85$, $P=0.070$, see Supplementary Fig. 8 for searchlight results and Supplementary Table 7 for post hoc tests, Tukey corrected). In the 7×3 repeated-measures ANOVA along the dorsal stream with factors ROI (V1, V2, V3, IPS0, IPS1, IPS2 and SPL) and background (no, low and high clutter) we found no significant main effect (ROI: $F_{(6,144)}=1.38$, $P=0.227$; background: $F_{(2,48)}=0.94$, $P=0.396$) or interaction effect ($F_{(12,288)}=0.96$, $P=0.463$).

In sum, our results confirm that the ventral stream constructs object representations that are robust to changes in viewing conditions and show in particular that location-tolerant category representations emerge in the ventral stream unaffected by the clutter level in the object's background.

Object category representations in time. Emergence of object category representations can be delayed, for example when objects are occluded or are hard to categorize^{44–46}. This suggests that object category representations might emerge with a delay also when objects appear on cluttered backgrounds, for example because additional grouping and segmentation operations are necessary that depend on recurrence and hence require additional time^{47–49}.

We therefore investigated whether background clutter influences the timing with which location-tolerant category representations emerge using time-resolved multivariate EEG analysis (Fig. 6c). We found that object category could be reliably classified for all background conditions from the EEG data (Fig. 6c, $N=27$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected), but with distinct temporal dynamics (see Supplementary Table 5 for classification onsets and peak values). Classification peaks were 18 ms later in the high-clutter than in the no- and the low-clutter conditions (no clutter: 215 ms (213–219 ms); low clutter: 215 ms (203–236 ms); high clutter: 233 ms (214–303 ms)). The delay (95% difference CI no clutter: 16–173 ms; $P<0.001$; low clutter: 13–171 ms; $P=0.029$) was significant ($N=27$, bootstrap test, 10,000 bootstraps, $P<0.05$, one-tailed bootstrap test against zero, FDR corrected; Fig. 6d). Location-independent category information at the peaks of the three background conditions was most pronounced at occipital and temporal electrodes as revealed in the EEG searchlight in sensor space (Fig. 6e and Supplementary Fig. 5g–i; $N=27$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR corrected across electrodes and time points). This is in line with the results from the fMRI searchlight analysis (Supplementary Fig. 8), together suggesting neural sources of the peaks in Fig. 6c in occipital and temporal regions. Univariate EEG activity was strongest in occipital rather

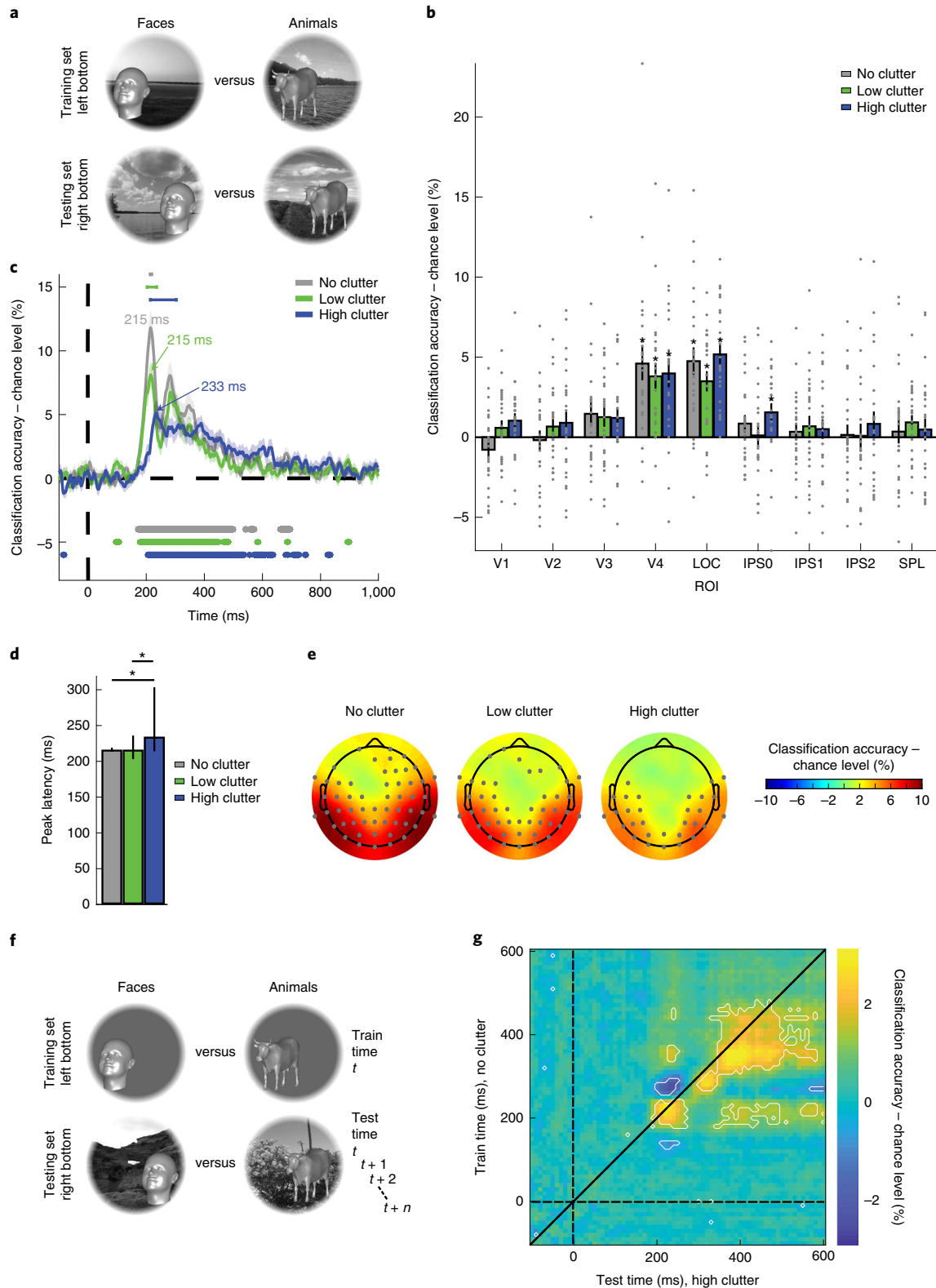


Fig. 6 | Spatial and temporal dynamics of object category representations. **a**, Classification scheme of category across location. **b**, Location-tolerant category representations in the ventral and dorsal streams. Stars indicate classification above chance level (two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected). Conventions as in Fig. 3b. **c**, Results of the time-resolved category classification across locations from EEG activation patterns. Conventions and statistics as in Fig. 5a. **d**, Peak latencies of curves in **c**. Statistics and conventions as in Fig. 5b. **e**, Results of searchlight in EEG channel space at peak latencies in no-, low- and high-clutter condition, down-sampled to 10 ms steps. Significant electrodes are marked in grey ($N = 27$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected across electrodes and time points). **f**, Time generalization analysis scheme for classifying object category across location and background condition. **g**, Results of the time generalization analysis ($N = 27$, two-tailed Wilcoxon signed-rank test, $P < 0.05$, FDR corrected). Conventions as in Fig. 5e.

than in temporal electrodes (Supplementary Fig. 5,j,k,l). Together, this shows that cortical processing of object category requires more time when objects appear in cluttered scenes compared with artificial blank backgrounds.

Analogous to the delay in location processing (Fig. 5a,b,e), we asked whether this delay indicates a temporal shift in the processing cascade or reflects a change to a later processing stage. To disambiguate, we classified object category across locations in a time generalization analysis across the no- and the high-clutter conditions (Fig. 6f).

We identified three main clusters of high classification accuracy with timing corresponding roughly to the timing of the three peaks observed in the time courses of the no- and the high-clutter conditions (Fig. 6g; see Supplementary Table 8 for timing details). To test whether category information in the no-clutter condition generalized to later time points in the high-clutter condition and thus was shifted below the diagonal, we computed the single-subject distances from the peak in the time generalization matrix to the diagonal. Category information peaks were significantly shifted below the diagonal as tested across subjects (mean Euclidean distance 27.24 ms; $N=27$, two-sided Wilcoxon signed-rank test, $P=0.025$, $r=0.43$, *s.e.m.* 2.50; single-subject peaks shown in Supplementary Fig. 6c), but not as tested for the group-averaged peak (Euclidean distance 28.28 ms; 10,000 bootstraps; one-tailed bootstrap test against zero, $P=0.230$; 95% CI -7.07 to 35.35). Classification accuracies were significantly higher below than above the diagonal from ~190 ms (no clutter) and ~240 ms (high clutter) until ~360 ms (no clutter) and ~400 ms (high clutter) (Supplementary Fig. 6d). This pattern of results suggests that object category representations of objects on blank and cluttered backgrounds emerge at a similar processing stage. This stage emerges with a delay when objects are presented on cluttered backgrounds, indicating recurrent processing.

Discussion

Using multivariate analysis of fMRI and EEG data and computational model comparison, we resolved where, how and when object location representations emerge in the human brain. Our results are three fold and depend crucially on whether objects appeared on cluttered backgrounds or on blank backgrounds. First, location representations emerged along the ventral visual pathway and peaked in region LOC when viewed on cluttered backgrounds. Second, this pattern of results was mirrored in DNNs trained on object categorization. Third, location representations emerged later in time when objects were viewed on cluttered backgrounds than when viewed on blank backgrounds. In-depth analysis suggested that this delay indexed recurrent processing in LOC. Together, these results provide converging evidence against the hypothesis that object location is processed in early visual cortex (H1), and in addition the results in space provide evidence for the hypothesis that object location emerges along the ventral stream (H3, Fig. 1a). A corresponding analysis of object category representations revealed equivalently an emergence in the ventral visual stream, and a delay when objects appear on cluttered backgrounds due to a temporal shift in the processing cascade, related to recurrent processing. Thus, the two arguably most fundamental properties of objects, that is, what the object is and where it is, emerge in the ventral visual stream with a similar spatiotemporal processing pattern.

Our fMRI results single out the ventral stream with a peak in LOC (H3), rather than early visual areas (H1) or the dorsal stream (H2), as the processing hierarchy responsible for computing object location in the human brain when objects appear on cluttered backgrounds. This concurs with a primate study¹⁴ that found category-orthogonal object representations to emerge in IT (the putative homologue of human LOC⁵⁰) rather than V4. Together, these results indicate that object location representations emerge along the ventral stream towards LOC when viewing conditions are realistic and challenging.

We observed that location representations with high clutter increased along the ventral stream for the classification of cross-but not within-hemifield locations. This pattern of results might be due to several factors. For one, statistical power is reduced when assessing results of cross- and within-hemifield location classification separately rather than combined, the test for which our study was originally planned. Second, cross-hemifield location representations might be more distinguishable as there is less integration of location information across than within hemispheres: cross-hemifield integration requires trans-callosal connections, whereas within-hemifield integration does not. Third, factors unrelated to location representations that however affect hemispheres differently, such as possible vascular changes, can contribute to the effect. Importantly, we do not see a difference between within- versus across-hemifield classification in the high-clutter condition in the EEG and DNN results, supporting our main conclusions and suggesting that the discrepancy in the fMRI results might be related to a decreased signal-to-noise ratio.

When objects are viewed on blank backgrounds rather than on cluttered backgrounds, location information can be read out from V1 because there is a direct mapping from stimulus location to the retinotopic location in V1 that is activated. With clutter, there is no such mapping (Fig. 1b) and therefore visual input is processed through the ventral visual stream cascade where LOC but not V1 reliably indicates object location representations. Under this assumption, location information in V1 might be an epiphenomenon caused by artificial stimulation conditions, revealing information that can be measured by the experimenter but is not necessarily used by the brain^{51–53} and relevant for behaviour at this stage of processing. Our results thus further emphasize the importance of increasing image complexity to increase the ecological validity of experimental stimuli²¹. While our study was designed to establish the presence and nature of object location representations in the brain, it cannot establish the behavioural relevance of those representations. Future studies could investigate this, for example, by using speeded detection tasks for objects presented in different locations and relating detection speed and performance to location representations across the brain.

Our results are seemingly at odds with neuropsychological findings showing that patients with ventral lesions performed well on localization tasks². However, later studies showed that in fact just localization behaviour was intact in those patients^{54–56}, but not location perception. It is conceivable that these patients recruited sparse location information from spared early visual areas to accomplish the localization tasks (similar to blindsight) and that tasks involving more cluttered displays would have been more challenging for these patients. In line with this, other patients with occipito-temporal lesions had problems with tasks requiring figure-ground segmentation⁵⁷ or perceptual grouping⁵⁸, both of which are essential to dissect an object from its background in a cluttered scene. Thus, neuropsychological studies taking background clutter into account are necessary to resolve this issue.

While we do observe location information in dorsal and ventral regions anterior and medial from LOC, the fMRI searchlight analysis (Supplementary Fig. 2) shows the peak in LOC. Why did location information not peak in other high-level ventral or dorsal areas? It is possible that IPS would represent object location more prominently if we optimized our stimulus selection for it by including tools⁵¹. However, the univariate response profile of the dorsal and ventral ROIs in our study tentatively suggests comparable activations across ROIs (Fig. 4d and Supplementary Table 4), indicating that univariate activation was not the source of lower information in IPS. Likewise, it is possible that different stimuli (for example, faces) would have yielded stronger effects in other high-level, category-selective ventral regions (for example, fusiform face area or occipital face area). Another possibility is that LOC has optimal

receptive field properties for the eccentricities used in this study^{59,60}, which allows it to encode object location on clutter better than other high-level ventral ROIs. These questions need more investigation in future research.

Our empirical findings were reinforced by the observation that representations of object location emerge in DNNs in a similar way as they emerge in the human brain. Importantly, the DNNs used here were trained on object categorization and not localization. Our results thus show that representations of object properties for which the network is not optimized can emerge in such networks¹⁴. One limitation of our approach is that the models used here were specifically designed to model the ventral visual stream^{25–30}, even though they have been shown to predict brain responses in the dorsal stream, too^{32,33}. Therefore, the presented modelling results cannot distinguish between H2 and H3. Future studies could compare location representations in DNNs that model the dorsal versus the ventral stream and investigate how the model's representations relate to brain representations in the two streams.

The time-resolved EEG analyses and the EEG–fMRI fusion analysis³⁸ revealed together that location representations of objects with high clutter were delayed due to a temporal shift within the same processing stage in LOC. Since temporal delays at the same processing stage cannot be explained purely by a feedforward neural architecture, this indicates the involvement of recurrence. Physiologically, this might be implemented via lateral connections within LOC, resulting in slower information accumulation^{61,62}. Furthermore, we found not only location but also object category representations to be delayed when objects were superimposed on natural scenes. Together with previous reports that object category processing can be delayed when objects are degraded, occluded or are hard to categorize^{44–46,48}, our results add to the emergent view that recurrent computations are critically involved in the processing of fundamental object properties such as what objects are⁶² and where they are in real world vision. Future studies could provide more direct evidence for recurrence by manipulating it experimentally, for example, by adding a masking condition to the study design used here.

We find that both object category and object location representations emerged gradually along the ventral visual stream. This might seem counter-intuitive, given that transformations that lead to the emergence of category representations in LOC have been linked to building increasing tolerance to viewing conditions, in particular to changes in object location^{5–7}. However, this apparent contradiction is qualified by the observation that the observed tolerance to changes in viewing conditions is graded rather than absolute⁶³, mirrored by the presence of cells in high-level ventral visual cortex with large overlapping receptive fields^{10,17}. Such tuning properties provide the spatial resolution needed for localization⁶⁴, while also providing robustness to location translation⁶⁵, needed for object categorization.

In this study, we deliberately avoided congruence between objects and backgrounds, which is known to lead to interaction effects with category processing⁴⁰. However, this deviation from normality in our stimulus set might have triggered mismatch responses that lead to additional recurrent processing for disambiguation or attentional responses triggered by atypical object appearance (for example, size and texture). Further, because objects and backgrounds did not form a coherent scene, objects and backgrounds might have been represented more independently. Another design limitation is that we constrain the number of locations to four to fully cross all stimulus conditions while maintaining a feasible session duration. Future research will have to establish whether congruent versus incongruent scene–object pairings yield different location representations on cluttered backgrounds and whether our results generalize to more locations.

What an object is and where an object is are arguably the two most fundamental properties that we need to know to be able to

interact with objects in our environment. Our results reveal the basis of this knowledge by revealing representations of location and category in the human brain when viewing conditions are challenging, as encountered outside of the laboratory. Both object location and category representations emerge along the ventral visual stream towards LOC and depend on recurrent processing. Together, our results provide a spatiotemporally resolved account of object vision in the human brain when viewing conditions are cluttered.

Methods

Participants in EEG and fMRI experiments. The experiment was approved by the ethics committee of the Department of Education and Psychology of the Freie Universität Berlin (ethics reference number 104/2015) and was conducted in accordance with the Declaration of Helsinki. Twenty-nine participants participated in the EEG experiment, of whom two were excluded because of equipment failure ($N=27$, mean age 26.8 years, *s.d.* 4.3 years, 22 female). Twenty-five participants (mean age 28.8 years, *s.d.* 4.0 years, 17 female) completed the fMRI experiment. The participant pools of the experiments did not overlap except for two participants. Sample size was chosen to exceed comparable magnetoencephalography, EEG and fMRI classification studies to enhance power^{8,9,43,66–68}. All participants had normal or corrected-to-normal vision and no history of neurological disorders. All participants provided informed consent prior to the studies and received a monetary reward or course credit for their participation.

Experimental design. To enable us to investigate the representation of object location, category and background independently, we used a fully crossed design with factors of category (four values: animals, cars, faces and chairs; Fig. 2a, left, with three exemplars per category), location (four values: left up, left bottom, right up and right bottom; Fig. 2a left centre) and background clutter (three values: no, low and high clutter; Fig. 2a, right centre). This amounted to 144 individual condition combinations (12 object exemplars \times 4 locations \times 3 background clutter levels). We analysed the data at the level of category, effectively resulting in 48 experimental conditions (4 categories \times 4 locations \times 3 background clutter levels).

Stimulus set generation. The stimulus material was created by superimposing three-dimensional (3D) rendered objects (Fig. 2a, left) with Gouraud shading in one of four image locations (Fig. 2a, left centre) onto images of real-world backgrounds (Fig. 2a, right centre).

In detail, in each category, one of the objects was rotated by 45°, one by 22.5° and the third by –45° with respect to the frontal view to introduce equal variance in the viewing angle for each category. Locations were in the four quadrants of the screen (Fig. 2a, left centre). Expressing locations in degrees of visual angle, the object's centre was 3° visual angle away from the vertical and horizontal central midlines (that is, 4.2° from image centre; Fig. 2a, right). The size of the objects was adjusted so that all of them fitted into one quadrant of the aperture, while maintaining a similar size (mean (*s.d.*) size: vertical, 2.4° (0.4°); horizontal, 2.2° (0.6°)).

We used backgrounds with three different clutter levels: no, low and high (Fig. 2a, right centre; note that example backgrounds shown here are for illustrative purposes and were not used in the experiment. The original stimulus material is available for download together with the data). We defined clutter as the organization and quantity of objects that fill up a visual scene⁶⁹. In the no-clutter condition, the background was uniform grey. In the low- and the high-clutter condition, we selected a set of 60 natural scene images each from the Places365 database (<http://places2.csail.mit.edu/download.html>) that had low or high clutter, respectively, and did not contain objects of the categories defined in our experimental design (that is, no animals, cars, faces or chairs). We converted the images to greyscale and superimposed a circular aperture of 15° visual angle. The visual angle was the same in the EEG and fMRI experiments.

We confirmed that our selection of low- and high-clutter images was appropriate by an independent behavioural rating experiment ($N=10$) in which participants rated clutter level on a scale from 1 to 6 (mean (*s.d.*) clutter image rating: low clutter, 2.52 (0.85); high clutter, 5.04 (0.87); the difference was significant: $N=10$, paired-sample *t* test, $P<0.0001$, $t=14.96$).

From the set of 60 low- and high-clutter images, we selected 48, one for each experimental condition of our experimental design. We then randomly paired objects to background images to avoid systematic congruencies between backgrounds and objects. This was done for each of the 20 runs of the EEG experiment and for the 10 runs of the fMRI experiment. This resulted in 144 individual images per run, one for each condition (that is, 12 object exemplars \times 4 locations \times 3 background clutter levels). The remaining set of 12 low- and high-clutter images was used separately to create catch trials in the EEG experiment (see details below).

Experimental procedures. fMRI main experiment. Each participant completed one fMRI recording session consisting of ten runs (run duration 552 s), resulting

in 92 min of fMRI recording of the main experiment. During each run, each of the 144 images of the stimulus set was shown once (denoted here as 'regular' trials) in random order. Image duration was 0.5 s, with a 2.5 s inter-stimulus interval (ISI). Images were presented at the centre of a black screen, overlaid with a red fixation cross in the centre. Participants were asked to fixate their eyes on the central cross at all times. Regular trials were interspersed every third to fifth trial (equally probable, in total 36 per run) with catch trials. Catch trials repeated the image shown on the previous trial (Fig. 2b, bottom). Participants were instructed to respond with a button press to these repetitions (that is, a one-back task). Catch trials were excluded from further analysis. Since this was a repeated-measures design, data collection and analysis were not performed blind to the conditions of the experiment.

fMRI localizer experiment. To define ROIs in early visual, dorsal and ventral visual stream areas, we performed a separate localizer experiment prior to the main fMRI experiment with images in three experimental conditions: faces, objects and scrambled objects. Each image shown in the localizer experiment consisted of four identical versions of an object presented at the four locations as defined in the main experiment (for example, one particular face shown in all four quadrants) to approximate the stimulation conditions of the main experiment.

The localizer experiment consisted of a single run lasting 384 s, comprising six blocks of presentation of faces, objects, scrambled objects and a blank background as baseline. Each stimulation block was 16 s long with presentations of 20 different objects (500 ms on, 300 ms off), including two one-back repetitions that participants were instructed to respond to with a button press. Stimulation block order was first order counterbalanced, with triplets of stimulation blocks being presented in random order and being interspersed regularly with blank background blocks.

EEG main experiment. The EEG experiment was a modified version of the fMRI main experiment with adjusted timing parameters and a different task (Fig. 2b, top). The EEG recording session consisted of 20 runs of 205 s each (that is, in total 68 min). Twenty-three participants completed all 20 runs, while four participants completed fewer runs due to technical problems (12 runs, 17 runs and 2×13 runs). Image duration was 0.5 s, with a 0.5 or 0.6 s ISI (equally probable) on regular trials. Participants were asked to fixate their eyes on the central cross at all times. Catch trials consisted of the presentation of the target object (a glass) at any of the four locations and on any type of background. Participants were instructed to respond with a button press to the glass (that is, a detection task), and to blink their eyes to minimize eye blink contamination on regular trials. To avoid contamination of movement and eye blink artefacts on subsequent trials, the ISI was 1 s on catch trials. Catch trials were excluded from further analysis. Since this was a repeated-measures design, data collection and analysis were not performed blind to the conditions of the experiment.

Pre-processing and univariate fMRI analysis. *fMRI acquisition and pre-processing.* We acquired MRI data on a 3-T Siemens Tim Trio scanner with a 12-channel head coil. We obtained a structural image using a T1-weighted sequence (magnetization-prepared rapid gradient-echo, 1 mm³ voxel size). For the main experiment and the localizer run, we obtained functional images covering the entire brain using a T2*-weighted gradient-echo planar sequence (repetition time 2 ms, echo time 30 ms, 70° flip angle, 3 mm³ voxel size, 37 slices, 20% gap, 192 mm field of view, 64 × 64 matrix size, interleaved acquisition).

We pre-processed fMRI data using SPM8 (<https://www.lion.ucl.ac.uk/spm/>). This involved realignment, coregistration and normalization to the structural Montreal Neurological Institute template brain. fMRI data from the localizer was smoothed with an 8 mm full-width at half-maximum Gaussian kernel, but the main experiment data was left unsmoothed.

Univariate fMRI analysis. For the main experiment, we modelled the fMRI responses to the 48 experimental conditions for each run using a general linear model (GLM). The onsets and durations of each image presentation entered the GLM as regressors and were convolved with a haemodynamic response function. Movement parameters entered the GLM as nuisance regressors. For each of the 48 conditions, we converted GLM parameter estimates into *t* values by contrasting each parameter estimate against the implicit baseline. This resulted in 48 condition-specific *t* value maps per run and participant.

For the localizer experiment, we modelled the fMRI response to the three experimental conditions, entering block onsets and durations as regressors of interest and movement parameters as nuisance regressors before convolving with the haemodynamic response function. From the resulting three parameter estimates, we generated two contrasts. The first contrast served to localize activations in early, mid-level ventral and dorsal visual regions (V1, V2, V3, V4, IPS0, IPS1, IPS2 and SPL) and was defined as objects + scrambled objects > baseline. The second contrast served to localize activations in object-selective area LOC and was defined as objects > scrambled objects. In sum, this resulted in two *t* value maps for the localizer run per participant.

Definition of ROIs. To identify regions along the ventral and dorsal visual streams, we defined ROIs in a two-step procedure. We first defined ROIs using anatomical

masks from a probabilistic atlas⁷⁰ for both hemispheres combined (three early visual ROIs for regions shared between the ventral and dorsal stream (V1, V2 and V3), two ROIs in mid- and high-level ventral visual cortex (V4 and LOC) and four ROIs in dorsal visual cortex (IPS0, IPS1, IPS2 and SPL)). To avoid overlap between the ROI masks we removed all overlapping voxels. In a second step we selected the 325 most activated voxels of the participant-specific localizer results within the masks, using the objects > scrambled contrast for LOC and the objects & scrambled objects > baseline contrast for the remaining ROIs. This yielded participant-specific ROI definitions.

EEG acquisition and pre-processing. We recorded EEG data using an EASYCAP 64-channel system and a Brainvision actiCHamp amplifier at a sampling rate of 1,000 Hz. The electrodes were placed according to the standard 10–10 system. The data were filtered online between 0.03 and 100 Hz and re-referenced online to FCz.

Offline pre-processing was conducted using the EEGLAB toolbox (version 14)⁷¹ and incorporated a low-pass filter with a cut-off at 50 Hz and epoching trials between –100 ms and 999 ms with respect to stimulus onset. Epochs were baseline corrected by subtracting the mean of the 100 ms prestimulus time window from the entire epoch. To clean the data from artefacts such as eye blinks, eye movements and muscular contractions, we used independent component analysis as implemented in the EEGLAB toolbox. SASICA⁷² was used to guide the visual inspection of components for removal. Components related to horizontal eye movements were identified using two lateral frontal electrodes (F7 and F8). In the last six participants, additional external electrodes were available that allowed for the direct recording of the horizontal electro-oculogram to identify and remove components related to horizontal eye movements. For blink artefact detection based on the vertical electro-oculogram, we used two frontal electrodes (Fp1 and Fp2). On average, 11 (*s.d.* 4) components were removed per participant. As a final step, we applied multivariate noise normalization to improve the signal-to-noise ratio and reliability of the data (following the recommendation of Guggenmos et al.⁷³).

Object location classification from brain measurements. To determine the amount of location information independent of category present in multivariate brain measurements, we applied a common multivariate cross-classification scheme^{68,66–69}. In essence, separately for each background condition, we classified location while assigning data from different object categories to the training and testing sets (Supplementary Fig. 1a). All classification analyses relied on binary c-support vector classification with a linear kernel as implemented in the libsvm toolbox⁷⁴ (<https://www.csie.ntu.edu.tw/~cjlin/libsvm>). Furthermore, all analyses were conducted in a participant-specific manner.

Spatially resolved multivariate fMRI analysis. We conducted an ROI-based and a spatially unbiased volumetric searchlight procedure^{24,75}. For the ROI-based analysis, for each ROI separately, we extracted and arranged *t* values into pattern vectors for each of the 48 conditions and 10 runs. To increase the signal-to-noise ratio, we randomly binned run-wise pattern vectors into five bins of two runs, which were averaged, resulting in five pseudo-run pattern vectors. We then performed five-fold leave-one-pseudo-run-out cross-validation, training on four and testing on one pseudo-trial per classification iteration. In detail, we assigned four pseudo-trials per location condition of the same category to the training set (Supplementary Fig. 1a). We then tested the SVM on one pseudo-trial for each of the same two location conditions, but now from a different category, yielding per cent classification accuracy (50% chance level) as output. Equivalent SVM training and testing was repeated for all combinations of location and category pairs. With four locations that were all classified pairwise once, this resulted in six pairwise location classifications. In addition, each pairwise location classification was iterated across all possible training and testing combinations of the four categories. This yielded an additional 12 iterations per location classification across training and testing pairs of categories. Therefore, in total 72 (6 × 12) classification accuracies were averaged during each of the five-fold cross-validation iterations, resulting in 360 averaged accuracies in total. The result reflects how much category-tolerant location information was present for each ROI, participant and background condition separately.

The searchlight procedure was conceptually equivalent to the ROI-based analysis with the difference of the selection of voxel patterns entering the analysis. For each voxel v_i in the 3D *t* value maps, we defined a sphere with a radius of four voxels centred around voxel v_i . For each condition and run, we extracted and arranged the *t* values for each voxel of the sphere into pattern vectors. Classification of location across category proceeded as described above. This resulted in one average classification accuracy for voxel v_i . Iterated across all voxels, this yielded a 3D volume of classification accuracies across the brain for each participant and background condition separately.

Time-resolved classification of location from EEG data. To determine the timing with which category-independent location information emerges in the brain, we conducted time-resolved EEG classification^{68,76}. This procedure was conceptually equivalent to the fMRI location classification in that it classified location while assigning data from different categories to the training and testing sets

and was conducted separately for each background condition and participant (Supplementary Fig. 1a).

For each time point of the epoched EEG data, we extracted 63 EEG channel activations and arranged them into pattern vectors for each of the 48 conditions and 60 raw trials. To increase the signal-to-noise ratio, we randomly assigned raw trials into four bins of 15 trials each and averaged them into four pseudo-trials. The classification was conducted on those four pseudo-trials. We trained the SVM on three pseudo-trials and tested it on the remaining pseudo-trial, yielding per cent classification accuracy (50% chance level, binary classification) as output. This procedure was repeated 100 times with random assignment of trials to pseudo-trials, and across all combinations of location and all category pairs. As for the fMRI classification, in total 72 (6 location pairs \times 12 category train–test pairs) classification accuracies were averaged. With 100 iterations to randomly assign trials to training and testing bins, this yielded a total of 7,200 classification accuracies, which were averaged per background condition and participant. The result reflects how much category-tolerant location information was present at each time point, participant and background condition separately.

Time-resolved EEG searchlight in sensor space. We conducted an EEG searchlight analysis resolved in time and sensor space (that is, across EEG channels) to gain insights into which EEG channels contained the highest amount of location information and therefore contributed most to the results of the time-resolved analysis described above. For the EEG searchlight, we conducted the time-resolved EEG classification as described above with the following difference: For each EEG channel c , we conducted the classification procedure on the five closest channels surrounding c . The classification accuracy was stored at the position of c . After iterating across all channels and down-sampling the time points to a 10 ms resolution, this yielded a classification accuracy map across all channels and down-sampled time points, for each participant and background condition separately.

Time generalization analysis of location from EEG data. To determine when object location representations are similar across background conditions and time, we used temporal generalization analysis^{34,38,68,76}.

The procedure was equivalent to the multivariate time-resolved EEG location classification analysis but with two crucial differences. First, data from the no-clutter condition were assigned to the training set while data from the high-clutter condition were assigned to the testing set (Supplementary Fig. 1b). The second difference was that the SVM was not only tested on data from the same time point as that from which the testing data were derived, but additionally on data from each time point from the -100 to 600 ms peri-stimulus time window (in 10 ms steps). Like previously, training was conducted on three and testing on one pseudo-trial, resulting in 7,200 classification accuracies (6 location pairs \times 12 category train–test pairs \times 100 randomization iterations), which were averaged per time point and participant. This resulted in a two-dimensional matrix of classification accuracies indicating the combination of time points in the no- and high-clutter conditions at which object location representations were similar in the no- and the high-clutter conditions.

Off-diagonal peak shift in time generalization matrix. To quantify whether classification accuracies were significantly higher below than on or above the diagonal, we computed the distance from the post-stimulus classification peak to the diagonal for single subjects. For this, we first determined the peak coordinates (p_x, p_y) along the x and y axes. We then computed the coordinates of the point on the diagonal that was closest to the peak using

$$b_x = \frac{(p_x + p_y)}{2}$$

since on the diagonal, $b_x = b_y$. This allowed us to compute the shortest perpendicular Euclidean distance between the peak and the diagonal as

$$d_{\text{Euclidean}} = \sqrt{(p_x - b_x)^2 + (p_y - b_x)^2}.$$

To be able to later test group distances against zero, we set

$$d_{\text{Euclidean}} = d_{\text{Euclidean}} \times -1$$

for all cases where $p_x < p_y$, which is the case for all peaks above the diagonal.

Diagonal difference in temporal generalization matrix. To obtain a temporally resolved estimate of the time points at which the classification accuracy was higher below than above the diagonal, we subtracted the classification accuracies above the diagonal from the accuracies below the diagonal. Specifically, we subtracted each time point from the time point with the equivalent coordinates mirrored along the diagonal. For example, the time point with coordinates 300 ms in the no-clutter (y axis) and 100 ms in the high-clutter (x axis) condition (above diagonal) was subtracted from the time point with coordinates 100 ms in the no-clutter (y axis) and 300 ms in the high-clutter (x axis) condition (below diagonal).

EEG–fMRI fusion. To determine the spatiotemporal correspondence between object location representations revealed at particular time points in the EEG signals and localized in particular cortical regions using fMRI, we used representational similarity analysis-based EEG–fMRI fusion^{77–79}. We focused the analysis on representations emerging at peak latencies in the EEG and on ventral stream ROIs. The rationale for this approach is that time points and ROIs are linked if they represent object locations similarly, that is, if their representational geometries (dissimilarity relations between representations) are comparable.

As a measure of (dis-)similarity relations between location representations, we used the classification results from the multivariate analyses conducted. This choice assumes that representations for two locations will be classified more easily if they are more dissimilar. In detail, we considered the pairwise classification accuracies between all pairs of locations (six) and all training and testing pairs across categories (six) in both training and testing directions (two), resulting in a 72×1 RDV. For EEG, we extracted the RDVs for the time points within the confidence intervals around the EEG peak latency, averaged them across time points and, following the method employed previously^{32,77,78}, averaged them across participants, resulting in one EEG RDV per background condition. For fMRI ROIs, we extracted the RDVs for each participant and background condition separately.

We compared fMRI and EEG RDVs for representational similarity by correlating (using Spearman's R) the averaged EEG RDV with the subject-specific fMRI ROI RDVs, resulting in one correlation per subject, background condition and ROI.

Multivariate classification of category. We conducted a set of spatially resolved (fMRI: ROI and searchlight), time-resolved and temporally generalized analyses (EEG) of object category. The analyses were equivalent to the procedures described above with the crucial difference that the role of the experimental factors location and category was reversed (Fig. 6a,f).

Object location classification in DNNs. We investigated whether DNNs trained on object categorization display a similar pattern of gradually emerging location representations along their processing hierarchy as we observed in the human brain.

We selected the DNN CORnet-S for investigation, on the basis of its top performance in predictivity of neural responses in the ventral stream as quantified on the Brain-Score platform²⁷. CORnet-S is a shallow recurrent DNN consisting of four computational blocks referred to as areas, analogous to ventral visual areas V1, V2, V4 and IT. Each block consists of four convolutional layers with self-recurrence and a skip connection followed by group normalization and a rectified linear unit. The response of the final IT block is averaged over the entire receptive field and mapped to categories using a fully connected linear decoder.

To investigate the representation of object location in CORnet-S, we performed multivariate pattern analysis analogous to the analysis performed on brain data, classifying object location across category separately for each background condition. For this, we extracted unit activations of the last layer in each block of the DNN after running a forward pass of the stimulus material from the 20 runs of the EEG experiment.

For the top layer of each block, we arranged the unit activations into pattern vectors for each of the 48 conditions and 60 trials. We then proceeded with the analysis as done with the EEG data (Supplementary Fig. 1a). We randomly assigned raw trials into four bins of 15 trials each and averaged them into four pseudo-trials. We trained the SVM on three pseudo-trials and tested it on the remaining pseudo-trial. This procedure was repeated 100 times with random assignment of trials to pseudo-trials, and across all combinations of location and all category pairs before results were averaged. This resulted in one averaged classification accuracy value per top layer of each CORnet-S block and per background condition. The result reflects how much category-tolerant location information was present in CORnet-S.

Statistical testing. Wilcoxon signed-rank test. We performed non-parametric two-tailed Wilcoxon signed-rank tests to test for above-chance classification accuracy at time points in the EEG time courses, in the EEG time generalization matrix, for Euclidean distances from peak to diagonal in the time generalization matrices, for above-chance classification in the ROI and fusion results and for significant voxels in the fMRI searchlight results. In each case, the null hypothesis was that the observed parameter (classification accuracy, correlation or Euclidean distance) came from a distribution with a median of chance-level performance (that is, 50% for pairwise classification and zero correlation or Euclidean distance). The resulting P values were corrected for multiple comparisons using false discovery rate (FDR) at 5% level if more than one test was conducted.

Bootstrap tests. We used bootstrapping to compute confidence intervals and to determine the significance of peak-to-peak differences in EEG latencies, peak-to-peak distances of fMRI searchlight classification peaks and for the distance from the group-averaged classification peak in the temporal generalization matrix to the diagonal in Figs. 5e and 6g. In each case, we sampled the participant pool 10,000 times with replacement and for each sample calculated the statistic of interest.

For the fMRI searchlight peak distances, we first shuffled condition labels of two background conditions to then generate a distribution of peak distances under the null hypothesis.

To determine whether peak-to-peak Euclidean distances in searchlight classification maps were significantly longer than expected independent of background, we set $P < 0.05$. If the computed P value was smaller than this threshold with Bonferroni correction, we rejected the null hypothesis of no peak-to-peak distance.

For the EEG peak-to-peak latency differences, we bootstrapped the latency difference between two background conditions, yielding an empirical distribution that could be compared with zero.

To determine whether peak-to-peak latencies in the EEG time courses were significantly different from zero, we computed the proportion of values that were equal to or smaller than zero and corrected them for multiple comparisons using FDR at $P = 0.05$. To compute 95% confidence intervals for single peak latencies in the EEG time courses, we bootstrapped the peaks for each background condition and determined the 95% percentiles of this distribution.

ANOVAs. We ran sets of ANOVAs to test for main effects and the interaction between ROIs along the ventral and dorsal stream and background condition, which we detail below. For all reported ANOVAs, we tested whether the assumption of sphericity had been met using Mauchly's test. Below, we report the effects for which the assumption of sphericity had been violated and for which the Greenhouse–Geisser estimates of sphericity were used to correct the degrees of freedom. For all remaining effects, the assumption of sphericity had been met.

To test for main effects and the interaction between ROIs along the ventral stream and background condition, we ran two 5×3 repeated-measures ANOVAs with within-subject factors of ROI (V1, V2, V3, V4 and LOC) and background (no, low and high clutter). The first ANOVA tested the results of location classification across categories. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of background ($P = 0.003$). Therefore, the degrees of freedom were corrected using the Greenhouse–Geisser estimates of sphericity ($\epsilon = 0.72$). The second ANOVA tested the results of category classification across locations. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of ROI ($P < 0.001$). The degrees of freedom were corrected using the Greenhouse–Geisser estimates of sphericity ($\epsilon = 0.61$).

To test for main effects and the interaction between ROIs along the dorsal stream and background condition, we ran two 7×3 repeated-measures ANOVAs with within-subject factors of ROI (V1, V2, V3, IPS0, IPS1, IPS2 and SPL) and background (no, low and high clutter). The first ANOVA tested the results of location classification across categories. Mauchly's test indicated that the assumption of sphericity had been violated for the main effect of ROI ($P < 0.001$) and for the interaction ($P = 0.028$). Therefore, the degrees of freedom were corrected using the Greenhouse–Geisser estimates of sphericity ($\epsilon = 0.53$ for the main effect of ROI, $\epsilon = 0.52$ for the interaction). The second ANOVA tested the results of category classification across locations. Mauchly's test indicated that the assumption of sphericity had been violated for the interaction ($P < 0.001$). The degrees of freedom were corrected using the Greenhouse–Geisser estimates of sphericity ($\epsilon = 0.59$).

To test for main effects and the interaction in the results of the EEG–fMRI fusion, we ran a 5×2 repeated-measures ANOVA with factors of ROI (V1, V2, V3, V4 and LOC) and clutter (no, high). The assumption of sphericity had been met for all main and interaction effects.

All post hoc tests were conducted using pairwise t tests, and P values were corrected for multiple comparisons using Tukey correction.

Effect sizes. For the main and interaction effects of the ANOVAs, we computed the partial η^2 using

$$\text{Partial } \eta^2 = \frac{\text{Sum of squares (SS)}_{\text{Effect}}}{\text{SS}_{\text{Effect}} + \text{SS}_{\text{Residual}}}$$

and the effect size estimate r (ref.⁷⁹) for the off-diagonal peak shifts across subjects, as tested with the Wilcoxon signed-rank test, using

$$r = \frac{Z}{\sqrt{N}}$$

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

The experimental stimuli, fMRI data, EEG data and the neural network activations are publicly available via https://osf.io/7zswm/?view_only=21a714db58584ffeb2837fc0548bf659.

Code availability

Analysis code is publicly available via <https://github.com/graumannm/ObjectLocationRepresentations>.

Received: 26 March 2021; Accepted: 14 January 2022;
Published online: 24 February 2022

References

- DiCarlo, J. J. & Cox, D. D. Untangling invariant object recognition. *Trends Cogn. Sci.* **11**, 333–341 (2007).
- Ungerleider, L. & Haxby, J. V. 'What' and 'where' in the human brain. *Curr. Opin. Neurobiol.* **4**, 157–165 (1994).
- DiCarlo, J. J., Zoccolan, D. & Rust, N. C. How does the brain solve visual object recognition? *Neuron* **73**, 415–434 (2012).
- Milner, A. D. & Goodale, M. A. *The Visual Brain in Action* (Oxford Univ. Press, 2006).
- Schwarzlose, R. F., Swisher, J. D., Dang, S. & Kanwisher, N. The distribution of category and location information across object-selective regions in human visual cortex. *Proc. Natl Acad. Sci. USA* **105**, 4447–4452 (2008).
- Rust, N. C. & DiCarlo, J. J. Selectivity and tolerance ('invariance') both increase as visual information propagates from cortical area V4 to IT. *J. Neurosci.* **30**, 12978–12995 (2010).
- Baek, A., Wagemans, J. & Op de Beeck, H. P. The distributed representation of random and meaningful object pairs in human occipitotemporal cortex: the weighted average as a general rule. *Neuroimage* **70**, 37–47 (2013).
- Cichy, R. M. et al. Probing principles of large-scale object representation: category preference and location encoding. *Hum. Brain Mapp.* **34**, 1636–1651 (2013).
- Golomb, J. D. & Kanwisher, N. Higher level visual cortex represents retinotopic, not spatiotopic, object location. *Cereb. Cortex* **22**, 2794–2810 (2012).
- Wandell, B. A. & Winawer, J. Computational neuroimaging and population receptive fields. *Trends Cogn. Sci.* **19**, 349–357 (2015).
- Kravitz, D. J., Saleem, K. S., Baker, C. I. & Mishkin, M. A new neural framework for visuospatial processing. *Nat. Rev. Neurosci.* **12**, 217–30 (2011).
- Zachariou, V. et al. Common dorsal stream substrates for the mapping of surface texture to object parts and visual spatial processing. *J. Cogn. Neurosci.* **27**, 2442–2461 (2015).
- Xu, Y. & Vaziri-Pashkam, M. Examining the coding strength of object identity and nonidentity features in human occipito-temporal cortex and convolutional neural networks. *J. Neurosci.* **41**, 4234–4252 (2021).
- Hong, H., Yamins, D. L. K., Majaj, N. J. & DiCarlo, J. J. Explicit information for category-orthogonal object properties increases along the ventral stream. *Nat. Neurosci.* **19**, 613–622 (2016).
- Brewer, A. A., Liu, J., Wade, A. R. & Wandell, B. A. Visual field maps and stimulus selectivity in human ventral occipital cortex. *Nat. Neurosci.* **8**, 1102–1109 (2005).
- Larsson, J. & Heeger, D. J. Two retinotopic visual areas in human lateral occipital cortex. *J. Neurosci.* **26**, 13128–13142 (2006).
- Groen, I. I. A., Silson, E. H. & Baker, C. I. Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philos. Trans. R. Soc. B* **372**, 20160102 (2017).
- Grill-Spector, K., Kourtzi, Z. & Kanwisher, N. The lateral occipital complex and its role in object recognition. *Vis. Res.* **41**, 1409–1422 (2001).
- Malach, R., Levy, I. & Hasson, U. The topography of high-order human object areas. *Trends Cogn. Sci.* **6**, 176–184 (2002).
- Levy, I., Hasson, U., Avidan, G., Hendler, T. & Malach, R. Center-periphery organization of human object areas. *Nat. Neurosci.* **4**, 533–539 (2001).
- Sonkusare, S., Breakspear, M. & Guo, C. Naturalistic stimuli in neuroscience: critically acclaimed. *Trends Cogn. Sci.* **23**, 699–714 (2019).
- Henderson, J. M. & Hollingworth, A. High-level scene perception. *Annu. Rev. Psychol.* **50**, 243–271 (1999).
- Malach, R. et al. Object-related activity revealed by functional magnetic resonance imaging in human occipital cortex. *Proc. Natl Acad. Sci. USA* **92**, 8135–8139 (1995).
- Kriegeskorte, N., Goebel, R. & Bandettini, P. Information-based functional brain mapping. *Proc. Natl Acad. Sci. USA* **103**, 3863–3868 (2006).
- Yamins, D. L. K. & DiCarlo, J. J. Using goal-driven deep learning models to understand sensory cortex. *Nat. Neurosci.* **19**, 356–365 (2016).
- Kubilius, J. et al. in *Advances in Neural Information Processing Systems* (eds. Wallach, H. et al.) **32**, 12805–12816 (Curran Associates, 2019).
- Schrimpf, M. et al. Integrative benchmarking to advance neurally mechanistic models of human intelligence. *Neuron* **108**, 413–423 (2020).
- Kriegeskorte, N. & Douglas, P. K. Cognitive computational neuroscience. *Nat. Neurosci.* **21**, 1148–1160 (2018).
- Yamins, D. L. K. et al. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proc. Natl Acad. Sci. USA* **111**, 8619–8624 (2014).
- Güçlü, U. & van Gerven, M. A. J. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *J. Neurosci.* **35**, 10005–10014 (2015).

31. Cichy, R. M. & Kaiser, D. Deep neural networks as scientific models. *Trends Cogn. Sci.* **23**, 305–317 (2019).
32. 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**, 1–17 (2016).
33. Güçlü, U. & van Gerven, M. A. J. Increasingly complex representations of natural movies across the dorsal stream are shared between subjects. *Neuroimage* **145**, 329–336 (2017).
34. King, J. R. & Dehaene, S. Characterizing the dynamics of mental representations: the temporal generalization method. *Trends Cogn. Sci.* **18**, 203–210 (2014).
35. Spoerer, C. J., McClure, P. & Kriegeskorte, N. Recurrent convolutional neural networks: a better model of biological object recognition. *Front. Psychol.* **8**, 1551 (2017).
36. Spoerer, C. J., Kietzmann, T. C., Mehrer, J., Charest, I. & Kriegeskorte, N. Recurrent neural networks can explain flexible trading of speed and accuracy in biological vision. *PLoS Comput. Biol.* **16**, e1008215 (2020).
37. Cichy, R. M. & Oliva, A. A M/EEG-fMRI fusion primer: resolving human brain responses in space and time. *Neuron* **107**, 772–781 (2020).
38. Cichy, R. M., Pantazis, D. & Oliva, A. Resolving human object recognition in space and time. *Nat. Neurosci.* **17**, 455–462 (2014).
39. Kriegeskorte, N., Mur, M. & Bandettini, P. Representational similarity analysis – connecting the branches of systems neuroscience. *Front. Syst. Neurosci.* **2**, 4 (2008).
40. Kaiser, D., Quek, G. L., Cichy, R. M. & Peelen, M. V. Object vision in a structured world. *Trends Cogn. Sci.* **23**, 672–685 (2019).
41. Vö, M. L. H., Boettcher, S. E. & Draschkow, D. Reading scenes: how scene grammar guides attention and aids perception in real-world environments. *Curr. Opin. Psychol.* **29**, 205–210 (2019).
42. Biederman, I., Mezzanotte, R. J. & Rabinowitz, J. C. Scene perception: detecting and judging objects undergoing relational violations. *Cogn. Psychol.* **14**, 143–177 (1982).
43. Brandman, T. & Peelen, M. V. Interaction between scene and object processing revealed by human fMRI and MEG decoding. *J. Neurosci.* **37**, 7700–7710 (2017).
44. Tang, H. et al. Spatiotemporal dynamics underlying object completion in human ventral visual cortex. *Neuron* **83**, 736–748 (2014).
45. Kar, K., Kubilius, J., Schmidt, K., Issa, E. B. & DiCarlo, J. J. Evidence that recurrent circuits are critical to the ventral stream's execution of core object recognition behavior. *Nat. Neurosci.* **22**, 974–983 (2019).
46. 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).
47. Lamme, V. A. F. & Roelfsema, P. R. The distinct modes of vision offered by feedforward and recurrent processing. *Trends Neurosci.* **23**, 571–579 (2000).
48. Groen, I. I. A. et al. Scene complexity modulates degree of feedback activity during object detection in natural scenes. *PLoS Comput. Biol.* **14**, e1006690 (2018).
49. Seijdel, N., Tsakmakidis, N., De Haan, E. H. F., Bohte, S. M. & Scholte, H. S. Depth in convolutional neural networks solves scene segmentation. *PLoS Comput. Biol.* **16**, e1008022 (2020).
50. Kriegeskorte, N. et al. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron* **60**, 1126–1141 (2008).
51. Williams, M. A., Dang, S. & Kanwisher, N. G. Only some spatial patterns of fMRI response are read out in task performance. *Nat. Neurosci.* **10**, 685–686 (2007).
52. Grootswagers, T., Cichy, R. M. & Carlson, T. A. Finding decodable information that can be read out in behaviour. *Neuroimage* **179**, 252–262 (2018).
53. de-Wit, L., Alexander, D., Ekroll, V. & Wagemans, J. Is neuroimaging measuring information in the brain? *Psychon. Bull. Rev.* **23**, 1415–1428 (2016).
54. Milner, A. D. et al. Perception and action in 'visual form agnosia'. *Brain* **114**, 405–428 (1991).
55. James, T. W., Culham, J., Humphrey, G. K., Milner, A. D. & Goodale, M. A. Ventral occipital lesions impair object recognition but not object-directed grasping: an fMRI study. *Brain* **126**, 2463–2475 (2003).
56. Goodale, M. A. & Milner, A. D. Separate visual pathways for perception and action. *Essent. Sources Sci. Stud. Consciousness* **15**, 20–25 (1992).
57. De Renzi, E. & Lucchelli, F. The fuzzy boundaries of apperceptive agnosia. *Cortex* **29**, 187–215 (1993).
58. Riddoch, M. J. & Humphreys, G. W. A case of integrative visual agnosia. *Brain* **110**, 1431–1462 (1987).
59. Sayres, R. & Grill-Spector, K. Relating retinotopic and object-selective responses in human lateral occipital cortex. *J. Neurophysiol.* **100**, 249–267 (2008).
60. Alvarez, I., de Haas, B., Clark, C. A., Rees, G. & Samuel Schwarzkopf, D. Comparing different stimulus configurations for population receptive field mapping in human fMRI. *Front. Hum. Neurosci.* **9**, 1–16 (2015).
61. Felleman, D. & Van Essen, D. C. Distributed hierarchical processing in the primate cerebral cortex. *Cereb. Cortex* **1**, 1–47 (1991).
62. Kietzmann, T. C. et al. Recurrence is required to capture the representational dynamics of the human visual system. *Proc. Natl Acad. Sci. USA* **116**, 21854–21863 (2019).
63. Eger, E., Kell, C. A. & Kleinschmidt, A. Graded size sensitivity of object-exemplar-evoked activity patterns within human LOC subregions. *J. Neurophysiol.* **100**, 2038–2047 (2008).
64. Eurich, C. W. & Schwegler, H. Coarse coding: calculation of the resolution achieved by a population of large receptive field neurons. *Biol. Cybern.* **76**, 357–363 (1997).
65. Spirkovska, L. & Reid, M. B. Coarse-coded higher-order neural networks for PSRI object recognition. *IEE Trans. Neural Netw.* **4**, 276–283 (1993).
66. Cichy, R. M., Chen, Y. & Haynes, J. D. Encoding the identity and location of objects in human LOC. *Neuroimage* **54**, 2297–2307 (2011).
67. Carlson, T., Hogendoorn, H., Fonteijn, H. & Verstraten, F. A. J. Spatial coding and invariance in object-selective cortex. *Cortex* **47**, 14–22 (2011).
68. Isik, L., Meyers, E. M., Leibo, J. Z. & Poggio, T. The dynamics of invariant object recognition in the human visual system. *J. Neurophysiol.* **111**, 91–102 (2014).
69. Park, S., Konkle, T. & Oliva, A. Parametric coding of the size and clutter of natural scenes in the human brain. *Cereb. Cortex* **25**, 1792–1805 (2015).
70. Wang, L., Mruczek, R. E. B., Arcaro, M. J. & Kastner, S. Probabilistic maps of visual topography in human cortex. *Cereb. Cortex* **25**, 3911–3931 (2015).
71. Delorme, A. & Makeig, S. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *J. Neurosci. Methods* **134**, 9–21 (2004).
72. Chaumon, M., Bishop, D. V. M. & Busch, N. A. A practical guide to the selection of independent components of the electroencephalogram for artifact correction. *J. Neurosci. Methods* **250**, 47–63 (2015).
73. Guggenmos, M., Sterzer, P. & Cichy, R. M. Multivariate pattern analysis for MEG: a comparison of dissimilarity measures. *Neuroimage* **173**, 434–447 (2018).
74. Chang, C.-C. & Lin, C.-J. Libsvm: a library for support vector machines. *ACM Trans. Intell. Syst. Technol.* **2**, 1–27 (2011).
75. Haynes, J. D. et al. Reading hidden intentions in the human brain. *Curr. Biol.* **17**, 323–328 (2007).
76. Carlson, T. A., Hogendoorn, H., Kanai, R., Mesik, J. & Turret, J. High temporal resolution decoding of object position and category. *J. Vis.* **11**, 1–17 (2011).
77. Mohsenzadeh, Y., Qin, S., Cichy, R. M. & Pantazis, D. Ultra-rapid serial visual presentation reveals dynamics of feedforward and feedback processes in the ventral visual pathway. *eLife* **7**, 1–23 (2018).
78. Cichy, R. M. & Teng, S. Resolving the neural dynamics of visual and auditory scene processing in the human brain: a methodological approach. *Philos. Trans. R. Soc. B* **372**, 1714 (2017).
79. Rosenthal, R. *Meta-analytic Procedures for Social Research* (Sage, 1991).

Acknowledgements

We thank D. Kaiser for comments and support. We thank S. Shrestha for helpful conversations on the math. Computing resources were provided by the high-performance computing facilities at ZEDAT, Freie Universität Berlin. EEG and fMRI data were acquired at the Center for Cognitive Neuroscience (CCNB), Freie Universität Berlin, Berlin. The study was supported by the German Research Council (DFG) (CI241/1-1, CI241/3-1, R.M.C. and M.G.), by the European Research Council (ERC-StG-2018-803370, R.M.C.) and by the Alfons and Gertrud Kassel Foundation (G.R. and K.D.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

M.G., C.C. and R.M.C. designed research. M.G. and C.C. performed experiments. M.G. performed data analyses. K.D. performed computational modelling. M.G., K.D. and R.M.C. wrote the manuscript. G.R. acquired funding.

Funding

Open access funding provided by Freie Universität Berlin.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41562-022-01302-0>.

Correspondence and requests for materials should be addressed to Monika Graumann or Radoslaw M. Cichy.

Peer review information *Nature Human Behaviour* thanks Talia Brandman, Christian Olivers and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to

the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2022

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our [Editorial Policies](#) and the [Editorial Policy Checklist](#).

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

- | n/a | Confirmed |
|-------------------------------------|--|
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> The statistical test(s) used AND whether they are one- or two-sided
<i>Only common tests should be described solely by name; describe more complex techniques in the Methods section.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> A description of all covariates tested |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals) |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> For null hypothesis testing, the test statistic (e.g. F , t , r) with confidence intervals, effect sizes, degrees of freedom and P value noted
<i>Give P values as exact values whenever suitable.</i> |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings |
| <input checked="" type="checkbox"/> | <input type="checkbox"/> For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes |
| <input type="checkbox"/> | <input checked="" type="checkbox"/> Estimates of effect sizes (e.g. Cohen's d , Pearson's r), indicating how they were calculated |

Our web collection on [statistics for biologists](#) contains articles on many of the points above.

Software and code

Policy information about [availability of computer code](#)

Data collection The data was collected using Matlab and the experimental paradigms were presented using the Psychophysics Toolbox Version 3.0.12 (PTB-3).

Data analysis For the data preprocessing and analysis we used the following software: MATLAB R2018b, EEGLAB toolbox (version 14), SASICA plugin for EEGLAB, LIBSVM-3.11, SPM8 toolbox, CoSMoMVPA toolbox.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research [guidelines for submitting code & software](#) for further information.

Data

Policy information about [availability of data](#)

All manuscripts must include a [data availability statement](#). This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The experimental stimuli used in this study, the fMRI and EEG data as well as neural network activations are publicly available via [https://osf.io/7zswn/?view_only=21a714db58584ffeb2837fc0548bf659](https://osf.io/7zswm/?view_only=21a714db58584ffeb2837fc0548bf659).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

Life sciences Behavioural & social sciences Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see [nature.com/documents/nr-reporting-summary-flat.pdf](https://www.nature.com/documents/nr-reporting-summary-flat.pdf)

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	In this study we recorded quantitative data separately from two experiments. 1) 3 Tesla functional magnetic resonance imaging (fMRI) data to acquire human brain activity data with high spatial resolution. 2) Electroencephalography (EEG) data to acquire human brain activity data with high temporal resolution. In both experiments, participants were performing a visual task while we recorded data.
Research sample	29 participants participated in the EEG experiment of which two were excluded because of equipment failure (N=27, mean age 26.8 years, SD=4.3, 22 female). 25 participants (mean age 28.8, SD=4.0, 17 female) completed the fMRI experiment. The participant pools of the experiments did not overlap except for two participants. All participants provided informed consent prior to the studies and received a monetary reward or course credit for their participation.
Sampling strategy	Participants were selected according to the following requirements: 18-40 years old, with normal or corrected-to-normal vision, fulfillment of the MR security criteria (no implants or metal parts, tattoos, non-removable piercing, claustrophobia, pregnancy, neurological disorders, etc.). Sample size was chosen to exceed comparable M/EEG and fMRI classification studies to enhance power.
Data collection	During both experiments, participants' responses were recorded with a computer, while the ongoing brain activity during the task was recorded using the 3T fMRI scanner (experiment 1) and the EEG (experiment 2). No one was present in the room together with the participants during the experiments. Blinding to the experimental conditions or the study hypothesis was not possible, but data was analyzed using a single pipeline for all subjects.
Timing	1) fMRI experiment: the data collection started February 2019 and ended in March 2019. 2) EEG experiment: the data collection started in May 2017 and ended in November 2017, with a short gap from July to September 2017 for data analysis.
Data exclusions	1) No participants were excluded in the fMRI experiment. 2) Two participants were excluded in the EEG experiment because of equipment failure.
Non-participation	No participants declined participation or dropped out.
Randomization	Participants were not allocated into experimental groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> Antibodies
<input checked="" type="checkbox"/>	<input type="checkbox"/> Eukaryotic cell lines
<input checked="" type="checkbox"/>	<input type="checkbox"/> Palaeontology and archaeology
<input checked="" type="checkbox"/>	<input type="checkbox"/> Animals and other organisms
<input type="checkbox"/>	<input checked="" type="checkbox"/> Human research participants
<input checked="" type="checkbox"/>	<input type="checkbox"/> Clinical data
<input checked="" type="checkbox"/>	<input type="checkbox"/> Dual use research of concern

Methods

n/a	Included in the study
<input checked="" type="checkbox"/>	<input type="checkbox"/> ChIP-seq
<input checked="" type="checkbox"/>	<input type="checkbox"/> Flow cytometry
<input type="checkbox"/>	<input checked="" type="checkbox"/> MRI-based neuroimaging

Human research participants

Policy information about [studies involving human research participants](#)

Population characteristics	See above.
Recruitment	Participants were recruited using the mailing lists for study participation of the psychology program, of the cognitive

Recruitment neuroscience program and of the medical studies program from the following Berlin universities: Freie Universität Berlin, Humboldt Universität zu Berlin, Charité.

Ethics oversight The study was approved by the ethics committee of the Department of Education and Psychology of the Freie Universität Berlin, Germany.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Magnetic resonance imaging

Experimental design

Design type Event-related fMRI design.

Design specifications Each participant completed one fMRI recording session consisting of 10 runs (run duration: 552 s), resulting in 92 minutes of fMRI recording of the main experiment. During each run, each of the 144 images of the stimulus set was shown once (regular trials). Image duration was 0.5 s, with a 2.5 s inter-stimulus-interval (ISI). Regular trials were interspersed every 3rd to 5th trial (equally probable, in total 36 per run) with catch trials. Catch trials repeated the image shown on the previous trial. Participants were instructed to respond with a button press to these repetitions (i.e. a one-back task).

Behavioral performance measures Button presses and response times were recorded for each subject during the experiment. Responses were recorded to ensure that participants were directing their attention towards the stimuli. Response trials were excluded from analysis.

Acquisition

Imaging type(s) functional and structural MRI

Field strength 3 Tesla

Sequence & imaging parameters We acquired functional images covering the entire brain using a T2*-weighted gradient-echo planar sequence (TR=2, TE=30 ms, 70° flip angle, 3-mm³ voxel size, 37 slices, 20% gap, 192-mm field of view, 64 × 64 matrix size, interleaved acquisition).

Area of acquisition Whole brain.

Diffusion MRI Used Not used

Preprocessing

Preprocessing software We preprocessed fMRI data using SPM8. This involved realignment, coregistration and normalization to the structural MNI template brain. fMRI data from the localizer was smoothed with an 8 mm FWHM Gaussian kernel, but the main experiment data was left unsmoothed.

Normalization The normalization method applied on all functional brain data was non-linear. We entered the subject specific T1 structural image as source image and the MNI standard T1 provided in the SPM8 toolbox as template image.

Normalization template We used the T1 template in MNI space provided in the SPM8 toolbox.

Noise and artifact removal To remove movement artifacts from the fMRI time-series, we realigned the functional brain images in SPM8 using default parameters. In the GLM, movement parameters were entered as nuisance regressors. We applied no artifact removal for heart rate and respiration.

Volume censoring Was not applied.

Statistical modeling & inference

Model type and settings We performed multivariate pattern analysis on the brain activity data. Specifically, we trained and tested support-vector machines on the individual participants' data and performed a statistical analysis on classification results.

Effect(s) tested Whole-brain: for all voxels, we tested whether classification accuracies significantly exceeded chance level. This was done separately for three background conditions (no, low and high background clutter). ROI: using a repeated-measures ANOVA with a 5×3 design, we tested for the interaction between 5 regions-of-interest in the ventral stream (V1, V2, V3, V4, LOC) and 3 background conditions (no, low and high cluttered backgrounds). Another repeated measures ANOVA with 7 ×3 design tested the interaction between 7 regions-of-interest in the dorsal stream (V1,V2,V3,IPS0,IPS1,IPS2,SPL) and 3 background conditions (no, low and high cluttered backgrounds). When the assumption of sphericity was violated, the degrees of freedom were corrected using the Greenhouse-Geisser estimates of sphericity.

Specify type of analysis: Whole brain ROI-based Both

Anatomical location(s)

We first defined ROIs in early visual cortex (V1, V2, V3), in the ventral stream (V4, LOC) and in the dorsal stream (IPSO, IPS1, IPS2, SPL) using anatomical masks from a probabilistic atlas (Wang et al., 2015) for both hemispheres combined. To avoid overlap between the ROI masks we removed all overlapping voxels. In a second step we selected the 325 most activated voxels in the participant-specific localizer results, using the objects > scrambled contrast for LOC and the objects & scrambled objects > baseline contrast for the remaining ROIs. This yielded participant-specific ROI definitions.

Statistic type for inference
(See [Eklund et al. 2016](#))

We tested whether classification accuracies significantly exceeded chance-level. This was done per ROI and in the whole-brain searchlight it was done voxel-wise. In both cases we tested this with non-parametric, two-tailed Wilcoxon signed rank tests. In each case the null hypothesis was that the observed classification accuracies came from a distribution with a median of chance level performance (i.e., 50% for pairwise classification).

Correction

The P-values resulting from the Wilcoxon signed rank tests were corrected for multiple comparisons using false discovery rate at 5% level under the assumption of independent or positively correlated tests.

Models & analysis

n/a | Involved in the study

- Functional and/or effective connectivity
 Graph analysis
 Multivariate modeling or predictive analysis

Multivariate modeling and predictive analysis

For the ROI-based analysis, for each ROI separately we extracted and arranged t-values into pattern vectors for each of the 48 conditions and 10 runs. To increase the SNR, we randomly binned run-wise pattern vectors into five bins of two runs which were averaged, resulting in five pseudo-run pattern vectors. We then performed 5-fold leave-one-pseudo-run-out-cross validation. In detail, we assigned four pseudo-trials per location condition of the same category to the training set. We then tested the SVM on one pseudo-trial for each of the same two location conditions, but now from a different category yielding percent classification accuracy (50% chance level) as output. Equivalent SVM training and testing was repeated for all combinations of location and category pairs before results were averaged. The result reflects how much category-tolerant location information was present for each ROI, participant and background condition separately.

The searchlight procedure was conceptually equivalent to the ROI-based analysis with the difference of the selection of voxel patterns entering the analysis.

Preprint of Study 2

Graumann, M., Wallenwein, L. A., & Cichy, R. M. (submitted). Independent spatiotemporal effects of spatial attention and background clutter on human object location representations. *bioRxiv*. doi: [10.1101/2022.05.02.490141](https://doi.org/10.1101/2022.05.02.490141) .

This article is licensed under a CC-BY-NC-ND 4.0 International License creativecommons.org/licenses/by-nc-nd/4.0/

1 **Independent spatiotemporal effects of spatial attention and background clutter on**
2 **human object location representations**

3

4 Monika Graumann^{1,2,*}, Lara A. Wallenwein³, Radoslaw M. Cichy^{1,2,4,*}

5

6

7 1 Department of Education and Psychology, Freie Universität Berlin, 14195 Berlin, Germany

8 2 Berlin School of Mind and Brain, Faculty of Philosophy, Humboldt-Universität zu Berlin,

9 10117 Berlin, Germany

10 3 Department of Psychology, Universität Konstanz, 78457 Konstanz, Germany

11 4 Bernstein Center for Computational Neuroscience Berlin, 10115 Berlin, Germany

12

13 *Correspondence to:

14 monika.graumann@fu-berlin.de

15 **1 Abstract**

16 Spatial attention helps us to efficiently localize objects in cluttered environments. However,
17 the processing stage at which spatial attention modulates object location representations
18 remains unclear. Here we investigated this question identifying processing stages in time and
19 space in an EEG and fMRI experiment respectively. As both object location representations
20 and attentional effects have been shown to depend on the background on which objects appear,
21 we included object background as an experimental factor. During the experiments, human
22 participants viewed images of objects appearing in different locations on blank or cluttered
23 backgrounds while either performing a task on fixation or on the periphery to direct their covert
24 spatial attention away or towards the objects. We used multivariate classification to assess
25 object location information. Consistent across the EEG and fMRI experiment, we show that
26 spatial attention modulated location representations during late processing stages (>150ms, in
27 middle and high ventral visual stream areas) independent of background condition. Our results
28 clarify the processing stage at which attention modulates object location representations in the
29 ventral visual stream and show that attentional modulation is a cognitive process separate from
30 recurrent processes related to the processing of objects on cluttered backgrounds.

31

32 **2 Introduction**

33 Spatial attention helps us to focus visual processing on the relevant portions of the visual field
34 while ignoring its irrelevant portions (Desimone and Duncan, 1995). For example, spatial
35 attention helps during navigation to determine where in visual space objects are located,
36 allowing us to avoid obstacles and to reach desired targets better.

37

38 While the importance of spatial attention is widely acknowledged, its neural basis remains
39 incompletely understood. An important open question is, at which stage of the visual
40 processing hierarchy attention modulates object location representations. Previous research
41 yielded contradictory results. Considering the temporal emergence of attentional effects, some
42 studies found attentional modulation early (Mangun, 1995; Hillyard et al., 1998a, 1998b; Luck
43 et al., 2000) within a time window that corresponds to the initial bottom-up response within
44 the first 150 ms (Lamme and Roelfsema, 2000; VanRullen and Thorpe, 2001; Fahrenfort et al.,
45 2007; Camprodon et al., 2010; Koivisto et al., 2011) while others found such effects only later
46 (Wyatte et al., 2014; Groen et al., 2016; Kaiser et al., 2016; Battistoni et al., 2020). Similarly,
47 considering the locus in the visual processing hierarchy some studies found attentional
48 modulation already in V1 (Roelfsema et al., 1998; Martínez et al., 2001; Noesselt et al., 2002;
49 Khayat et al., 2006; Lakatos et al., 2008; Briggs et al., 2013; Herrero et al., 2013) while others
50 found such effects only or predominantly in higher-level brain regions (Buffalo et al., 2010;
51 Peelen and Kastner, 2011; Kay et al., 2015).

52

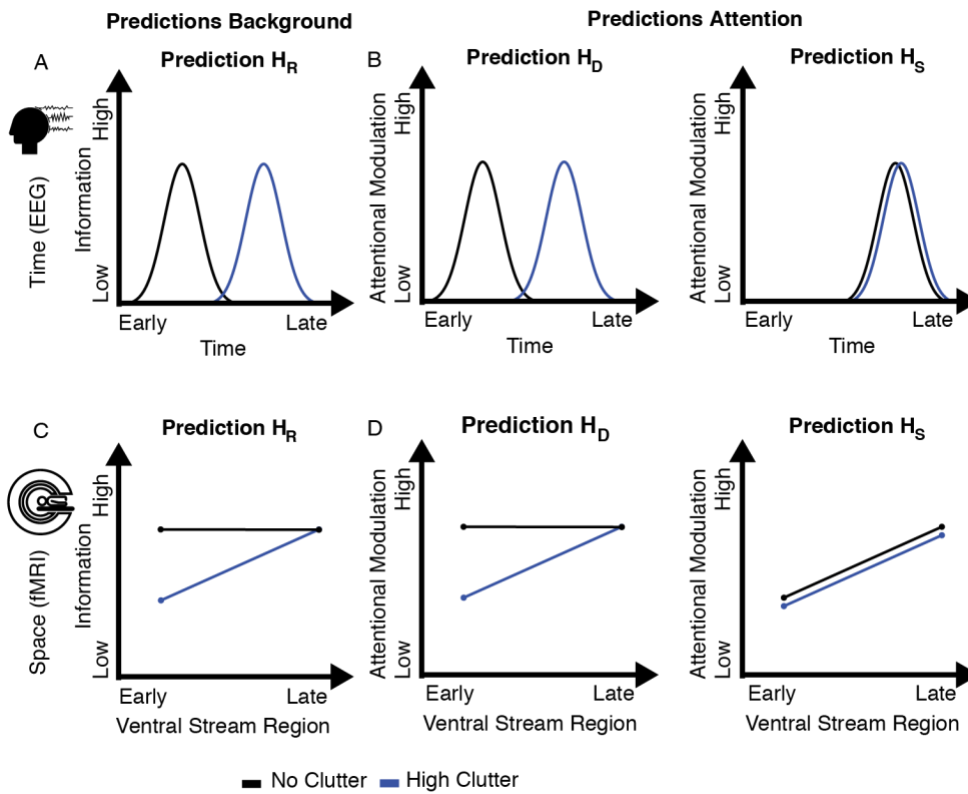
53 The contradiction might be resolved when considering together the processing stage at which
54 object location representations emerge, the object's viewing conditions and attentional
55 modulation. Recent research has shown that viewing conditions influence the processing stage
56 at which object location representations emerge. For example, object location representations
57 emerge early for objects on blank and late on cluttered backgrounds (Hong et al., 2016;
58 Graumann et al., 2022). Further, the surroundings of an object modulates the employment of
59 spatial attention: spatial attention is more relevant for the localization of objects in clutter than
60 in isolation (Treisman and Gelade, 1980; Wolfe, 1994).

61

62 Here we set out to untangle the complex link between the processing stage at which object
63 location representations emerge, its viewing conditions, and the effect of attentional
64 modulation.

65
66
67
68
69
70

Our hypotheses are as follows. We set the stage by hypothesizing based on recent findings that the processing stage at which object location representations emerge depends on the object's viewing conditions in particular its background (Graumann et al., 2022) independent of spatial attention. This replication hypothesis was termed $H_{\text{Replication}}$ (abbreviated H_R ; Fig. 1A,C).



71
72
73
74
75
76
77
78
79
80
81
82
83
84
85
86
87
88
89
90

Figure 1. Experimental predictions based on hypotheses. **A**, Predictions for the effect of background on location information in the EEG experiment. H_R predicts a delay in time for location information with high clutter compared to no clutter. **B**, Predictions for the effect of attention on location information in the EEG experiment. Predictions are based on H_R in A. H_D predicts that the time point when attentional modulation is highest depends on background: attentional modulation is highest at time points when location information is highest, depending on background condition. H_S predicts that attentional modulation is always highest during late processing stages, independent of background condition. **C**, Predictions for the effect of background on location information in the fMRI experiment. H_R predicts an increase along the ventral stream for location information with high clutter compared to no clutter. **D**, Predictions for the effect of attention on location information in the fMRI experiment. Predictions are based on H_R in C. H_D predicts that the region where attentional modulation is highest depends on background: attentional modulation is highest in regions where location information is highest, depending on background condition. H_S predicts that attentional modulation always increases along the ventral stream, independent of background condition.

On this basis we then theorize how an objects background impacts when (in time with respect to stimulus onset) and where (in the cortical processing hierarchy) attention modulates location representations. We propose two alternative hypotheses.

91 The first hypothesis is that attention and background interact: attention dynamically modulates
92 location representations at the processing stage at which they first emerge, resulting in an
93 interaction between background and attention (H_{Dynamic} , abbreviated H_D ; Fig. 1B,D). The
94 alternative hypothesis is that attention modulates location representations statically and always
95 during a late processing stage (Wyatte et al., 2014; Kay et al., 2015; Groen et al., 2016; Kaiser
96 et al., 2016; Battistoni et al., 2020), independent of the background (H_{Static} , abbreviated H_S ;
97 Fig. 1B,D).

98

99 We investigated these hypotheses in an integrated research project consisting of an EEG and
100 fMRI experiment in combination with multivariate pattern analysis methods. We manipulated
101 background by presenting objects on backgrounds of different clutter levels, and attention by
102 task instruction that attracted or diverted spatial attention from an object's location.

103

104 To anticipate, we first confirmed H_R , i.e., object location representations emerged later in time
105 and space when the object appeared on a cluttered background than on a blank background,
106 independent of attention. We then found strong empirical support for H_S . That is, attention
107 modulates object location representations late in both time and space, independent of
108 background.

109

110 3 Results

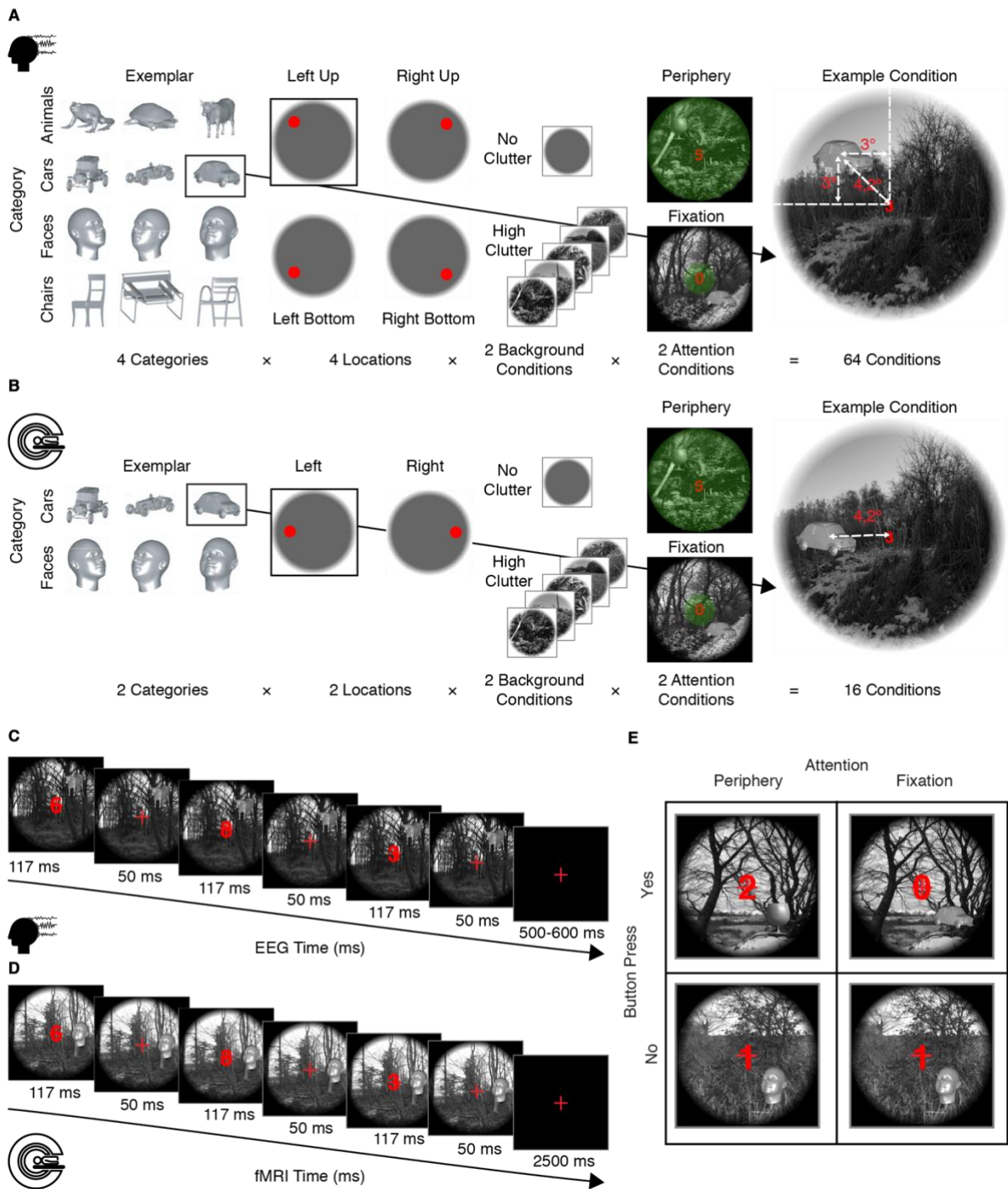
111 For both the EEG and fMRI experiments the strategy to determine when and where attention
112 modulates location representations was equivalent: first we sought to establish H_R , i.e., that
113 location representations of objects emerge at a later processing stage when objects are
114 presented on cluttered backgrounds compared to blank backgrounds, independent of attention.
115 On this basis we then arbitrated between H_D and H_S , i.e., whether attention dynamically
116 modulates object location representations at different processing stages depending on
117 background, or whether it statically modulated object location representations always at a late
118 processing stage.

119

120 The difference between the EEG and the fMRI analyses lies in the way that the processing
121 stages are determined: EEG determines the temporal delay with respect to image onset (Fig.
122 1A,B) and fMRI determines the region in the ventral visual stream (Fig. 1C,D) in which
123 experimental effects emerge.

124

125 In the following we give the specifics of the EEG and fMRI experiments, the precise
126 predictions, and the results. We begin with the EEG experiment determining the timing of
127 attentional modulation, followed by the fMRI experiment determining where in the visual
128 processing hierarchy the attentional modulation occurs.



129
130
131
132
133
134
135
136
137
138
139

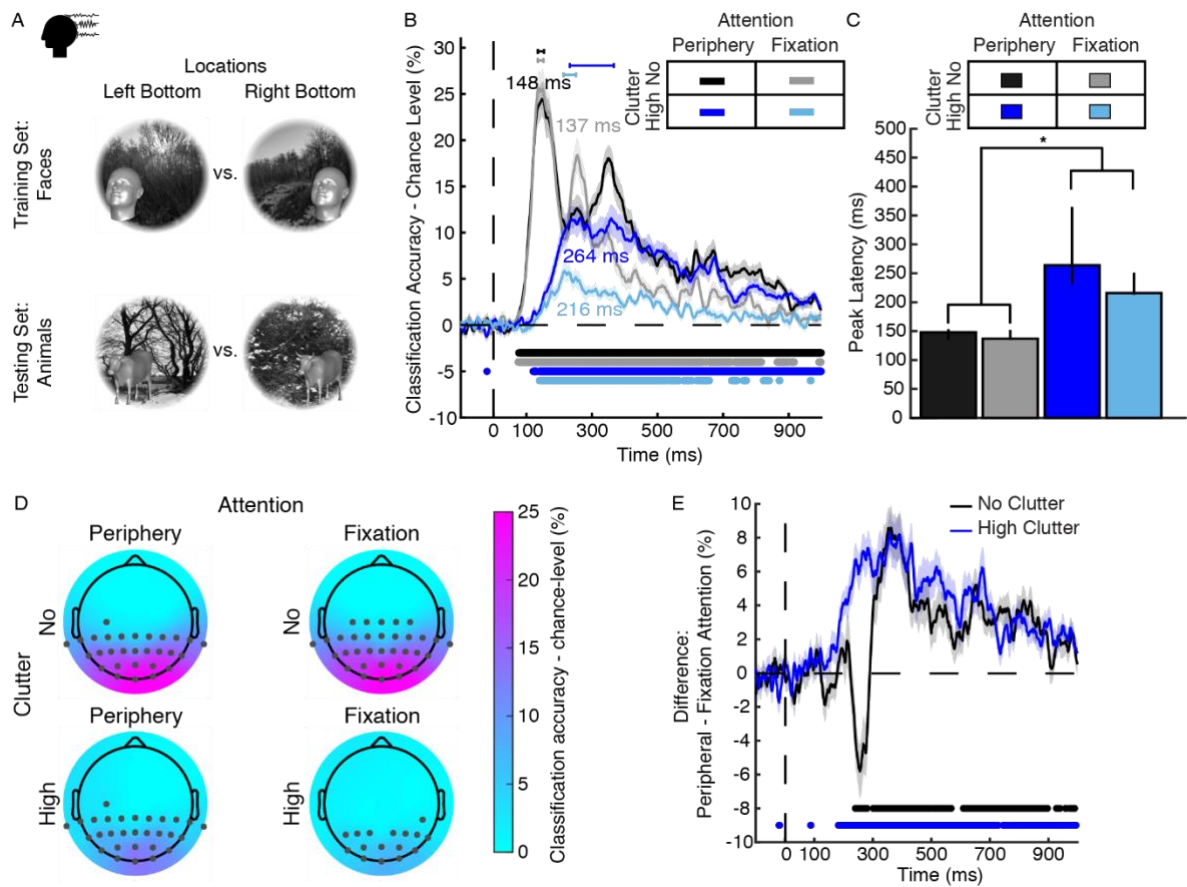
Figure 2. Experimental design and tasks. **A**, Experimental design in the EEG experiment. We used a fully crossed design with factors: object category, location, background and attention. Green translucent circles represent attentional width. **B**, Experimental design in fMRI experiment. The design was equivalent to the EEG experiment, except that the factors category and location had two levels. **C**, Trial timing and example condition in EEG experiment. **D**, Trial timing and example condition in fMRI experiment. **E**, Tasks. In the peripheral attention condition (left) participants responded with button press when a glass appeared in the periphery, while fixating their gaze on the central cross. Digits presented on fixation were task-irrelevant. In the fixation attention condition (right) participants responded with button press when the digit 0 appeared on fixation, while fixating on the central cross. Objects in the periphery were irrelevant in this task. Visual stimulation was the same in both tasks on regular trials (see bottom row ‘Button press: no’).

140 **3.1 Attentional modulation of object location representations in time**

141 For the EEG experiment we used an experimental design with fully crossed experimental
 142 factors background and attention with two levels per factor (2 background condition levels \times 2
 143 attention condition levels, Fig. 2A).

144
 145 In detail, participants saw objects from four different categories, each presented in four
 146 locations (Fig. 2A). The two background conditions were no and high clutter. Each object in
 147 each location was presented in both background conditions. Each combination of object
 148 category, location and background was then also crossed with the two levels of attention
 149 conditions: the peripheral and fixation attention conditions (Fig. 2A). Attention conditions
 150 were solely defined by the task that participants performed, while visual stimulation was
 151 identical (Fig. 2E). In the peripheral attention condition, participants directed their covert
 152 spatial attention to the periphery and responded to a catch object (glass) with button press (Fig.
 153 2A,E). In the fixation attention condition, participants performed a demanding task on fixation
 154 to remove their spatial attention from the objects in the periphery (Fig. 2A,E).

155



156

157 **Figure 3. Classification schemes and results of EEG location classification.** **A**, Scheme for the classification
158 of object location across categories within any background and attention condition. We trained a support vector
159 machine (SVM) to distinguish brain activation patterns evoked by objects of a particular category presented at
160 two locations (here: faces bottom left and right) and tested the SVM on activation patterns evoked by objects of
161 another category (here: animals) presented at the same locations. Objects are enlarged here for display purposes.
162 In the experiment objects did not extend across quadrants. **B**, Results of time-resolved location across category
163 classification from EEG data. Results are color-coded by background and attention condition, with significant
164 time points indicated by lines below curves ($N=26$, $P<0.05$, FDR-corrected), 95% confidence intervals of peak
165 latencies are indicated by lines above curves. Shaded areas around curves indicate SEM. **C**, Comparison of peak
166 latencies of curves in B. Error bars represent 95% CIs. Stars indicate significant peak latency differences ($N=26$,
167 bootstrap test with 10,000 bootstraps). **D**, Results of the location across category classification searchlight in EEG
168 channel space at peak latencies (as shown in B) in each condition. Significant electrodes are indicated by grey
169 dots ($N=26$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR-corrected across electrodes and time points). **E**,
170 Difference curves resulting from subtracting the time courses of the foveal from the peripheral attention condition
171 in each background condition. Conventions as in B.
172

173 In total, this 2×2 experimental design resulted in 4 factor combinations. We performed a time-
174 resolved and pair-wise classification analysis of location across category within each of these
175 four factor combinations separately (Fig. 3A,B). This meant training a classifier to distinguish
176 between millisecond-specific EEG pattern vectors associated with two locations and testing on
177 a held-out testing data set associated with the same two locations. We performed the
178 classification across object category, that is training on data associated with locations from one
179 object category and testing on data from another category (Fig. 3A). This ensured that location
180 classification results were not confounded with category information and allowed us to draw
181 conclusion about location representations independent of object category representations.

182 3.1.1 *The temporal dynamics of object location representations with blank and cluttered* 183 *backgrounds*

184 To lay the basis for later analyses on attentional modulation, we first tested H_R , i.e., that
185 location representations of objects with clutter emerge later than on blank backgrounds,
186 independent of attention. For this we determined and compared the latencies of the
187 classification peaks in the EEG time courses of both background conditions, assuming that the
188 peaks represent the time points at which representations become most differentiable (DiCarlo
189 and Cox, 2007). Our prediction was that location information would peak later in the high than
190 in the no clutter condition, because dissecting objects from the background requires additional
191 grouping and segmentation operations implemented in recurrent processing and thus increasing
192 processing time (Groen et al., 2018; Seijdel et al., 2020, 2021; Graumann et al., 2022).

193

194 The results of the time-resolved location classification are shown in Fig. 3B. We read out
195 location information from the EEG signal in all background and attention conditions above
196 chance level ($N=26$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR-corrected).

197

198 Focusing on peak latencies (95% confidence intervals reported in brackets, $N=26$, 10,000
199 bootstrap samples), we observed that time courses in the no and high clutter conditions peaked
200 at different times. In the no clutter condition, location information peaked early, regardless of
201 attention condition (Fig. 3B; peak latency peripheral condition: 148 ms (135–153.5 ms); peak
202 latency fixation condition: 137 ms (135–152 ms)). With high clutter, location information
203 peaked later in both attention conditions (Fig. 3B; peripheral condition: 264 ms (232–365 ms);
204 fixation condition: 216 ms (213–251 ms)). To test whether the peak latencies across
205 background conditions were significantly different, we bootstrapped the peak-to-peak latency
206 differences between pairs of no and high clutter condition peaks (Fig. 3C, 95% confidence
207 intervals in brackets, $N=26$, bootstrap test, 10,000 bootstraps, FDR-corrected). This was done
208 both within and across attention conditions. Overall, the results clearly and consistently support
209 Hr. Location information peaked significantly earlier in the no compared to the high clutter
210 conditions independent of attention condition: Within attention condition, the peak-to-peak
211 latency difference between background conditions was 116 ms (83–223 ms; $P<0.001$) in the
212 peripheral attention condition and 79 ms in the fixation attention condition (63–114 ms;
213 $P<0.001$). Across attention conditions, the delays between background condition peaks were
214 also significant (peripheral attention and no clutter condition vs. fixation attention and high
215 clutter condition: 68 ms delay, 63–105 m; $P<0.001$; fixation attention and no clutter condition
216 vs. peripheral attention and high clutter condition: 127 ms delay, 83–224 ms, $P<0.001$).

217

218 Additional analyses of the observed effects reproduced previously observed characteristics of
219 object location representations (Graumann et al., 2022) and thus further supported Hr. A
220 searchlight analysis in EEG sensor space (Fig. 3D) localized the sources of the peaks to
221 occipito-temporal electrodes (Fig. 3D), suggesting the locus of object location representations
222 to be in occipital and temporal cortices. A supplementary time-generalization analysis (King
223 and Dehaene, 2014) showed that location representations for objects on blank and cluttered
224 background emerged within the same processing stage, but with a delay with cluttered
225 backgrounds (Supplementary Fig. 1, Supplementary Methods 1).

226

227 Together, these results provide empirical evidence for Hr.

228 *3.1.2 Late attentional modulation of location representations independent of background*

229 Affirming H_R formed the basis for arbitrating between our main hypotheses H_D and H_S . H_D
230 predicts that attentional modulation is highest when location information is highest: with no
231 clutter, it predicts an early modulation in time of location representations and with high clutter
232 it predicts a late modulation in time (Fig. 1B). H_S states that spatial attention modulates location
233 representations always late, after the end of the bottom-up response at ~100-150 ms (Lamme
234 and Roelfsema, 2000; VanRullen and Thorpe, 2001; Fahrenfort et al., 2007; Camprodon et al.,
235 2010; Koivisto et al., 2011). Thus, H_D predicts an interaction between attention and background
236 and H_S predicts that they are independent.

237

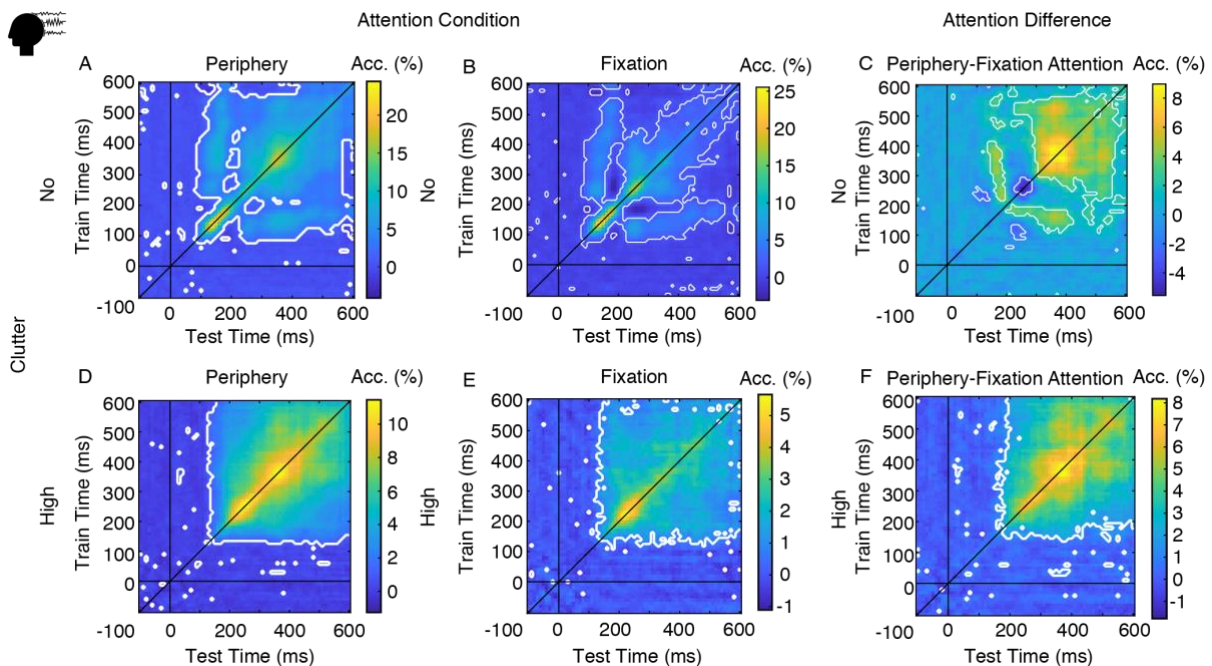
238 To assess H_S and H_D we determined the time course of attentional modulation in both
239 background conditions. Attentional modulation was defined as an enhancement of
240 representations (Desimone and Duncan, 1995; Reynolds and Chelazzi, 2004; Briggs et al.,
241 2013). To quantify attentional modulation, we subtracted classification accuracies in the
242 fixation attention condition from the peripheral attention condition, within each background
243 condition. Since visual stimulation was identical across attention conditions, we attributed
244 differences between them to attentional modulation.

245

246 Fig. 3E shows the result of this analysis. We found attentional modulation of location
247 representations in both background conditions in a late time window, providing clear evidence
248 for H_S . In detail, we observed a significant positive difference in the no clutter condition
249 starting from ~300 ms, reflecting attentional modulation (Fig. 3E; $N=26$, two-tailed Wilcoxon
250 signed-rank test, $P<0.05$, FDR-corrected). In the high clutter condition, we found evidence for
251 attentional modulation starting from 182 ms (Fig. 3E; $N=26$, two-tailed Wilcoxon signed-rank
252 test, $P<0.05$, FDR-corrected), as reflected in a significant positive difference that lasted until
253 the end of the time window.

254

255 Together, these results show that attention modulates object location representations in a late
256 time window after the bottom-up response, independent of background. This constitutes strong
257 evidence for H_S .



259
 260 **Figure 4. EEG results of time-generalization analyses within each background and attention condition.**
 261 **Rows represent background and columns represent attention conditions. A,** Location classification across
 262 categories and time points in the no clutter & peripheral attention condition. Horizontal and vertical black lines
 263 indicate stimulus onset, oblique black line highlights the diagonal. White outlines indicate significant time points
 264 ($N=26$, two-tailed Wilcoxon signed-rank test, $P<0.05$, FDR-corrected). **B,** Location classification across
 265 categories and time points in the no clutter & fixation attention condition. **C,** Difference matrix resulting from
 266 subtracting the matrices representing fixation (B) from peripheral attention (A) in the no clutter condition. Plot
 267 conventions as in A. **D,** Location classification across categories and time points in the high clutter & peripheral
 268 attention condition. **E,** Location classification across categories and time points in the high clutter & fixation
 269 attention condition. **F,** Difference matrix resulting from subtracting the matrices representing fixation (E) from
 270 peripheral attention (D) in the high clutter condition.
 271

272 While clearly supporting H_s , the results hitherto do not yet characterize the temporal dynamics
 273 underlying attentional modulation of location representations. Typically during visual
 274 perception, time-resolved multivariate results reflect a conglomerate of both rapidly changing
 275 transient information flow as well as persistent activity which maintains certain types of
 276 information over long stretches of time (Cichy et al., 2014; King and Dehaene, 2014).

277
 278 Thus, here we investigated whether attention and background modulate persistent, transient or
 279 both aspects of location representations. For this we conducted temporal generalization
 280 analysis (King and Dehaene, 2014). This resulted in two-dimensional time generalization
 281 matrices, indexed in both dimensions in time indicating similarities of object location
 282 representation across time. While transient representations are reflected as high information on
 283 the diagonal of such matrices, persistent representations are found off-diagonal.

284

285 As previously, we classified location representation within background and attention condition,
286 resulting in 4 time-generalization matrices (Fig. 4A,B,D,E), corresponding to the 4
287 classification time courses above (Fig. 3B). We first present the single results ordered by
288 background condition, before quantifying attentional modulation.

289

290 In the no clutter condition, we found similar results in both attention conditions (Fig. 4A,B):
291 location information peaked early at ~100 ms on the diagonal, representing transient
292 information flow ($N=27$, $P<0.05$, two-tailed Wilcoxon signed-rank test, FDR-corrected).
293 Starting from ~250 ms, location information generalized more broadly across time points,
294 indicating persistent information. In the high clutter condition (Fig. 4D,E) information
295 generalized broadly across time points starting from ~140 ms in both attention conditions,
296 ($N=27$, $P<0.05$, two-tailed Wilcoxon signed-rank test, FDR-corrected), indicating persistent
297 information. Transient information peaked on the diagonal starting from ~240 ms.

298

299 We quantified attentional modulation as above (Fig. 3E) by comparing the classification results
300 for the two attention conditions, subtracting the results of the fixation attention condition from
301 the peripheral attention condition.

302

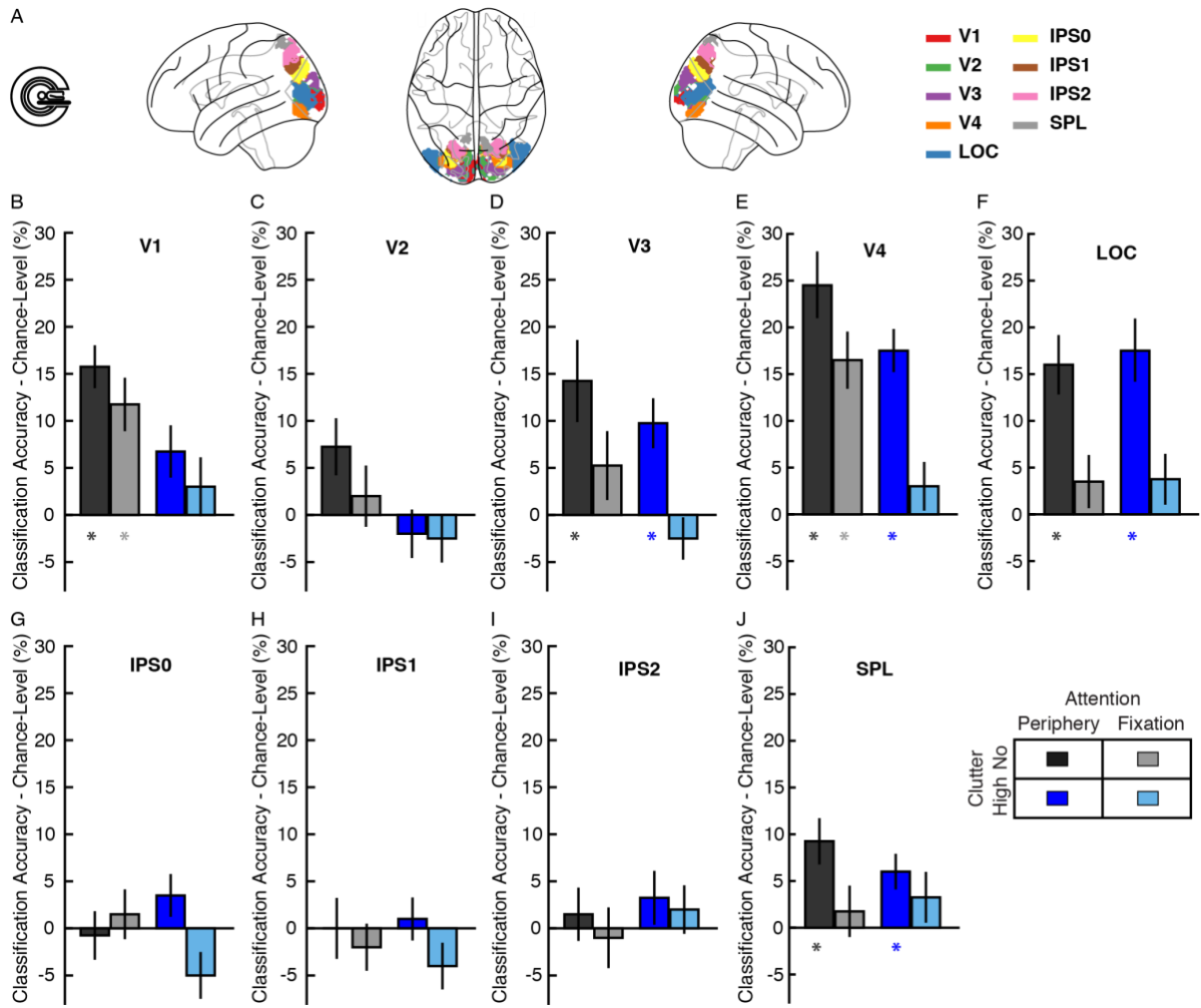
303 We found that spatial attention modulated both transient and persistent representations in late
304 time windows, independent of background. In the no clutter condition, attention modulated the
305 persistent clusters from ~230 ms and both transient and persistent information from ~300 ms
306 (Fig. 4C; $N=27$, $P<0.05$, two-tailed Wilcoxon signed-rank test, FDR-corrected). In the high
307 clutter condition, spatial attention modulated location representations across the entire time
308 window starting from ~180 ms (Fig. 4F; $N=27$, $P<0.05$, two-tailed Wilcoxon signed-rank test,
309 FDR-corrected).

310

311 In sum, we found that attention modulates both transient and persistent representations of
312 object location representations in late time beyond 150 ms.

313

314 **3.2 Clutter and attention independently affect location representations along the**
 315 **ventral visual stream**



316
 317 **Figure 5.** Location across category classification results in the four conditions in early (V1, V2, V3), ventral (V4,
 318 LOC) and dorsal (IPS0-2, SPL) visual ROIs. Stars below bars indicate significant above-chance classification
 319 ($N=20$, two-tailed Wilcoxon signed-rank test, $P<0.05$ false discovery rate (FDR) corrected). Error bars represent
 320 standard error of the mean (SEM). **A**, ROIs on cortical surface. **B**, V1. **C**, V2. **D**, V3. **E**, V4. **F**, LOC. **G**, IPS0. **H**,
 321 IPS1. **I**, IPS2. **J**, SPL.
 322

ROI	Main effect background	Main effect attention	Interaction effect
V1	$F_{(1,19)}=9.88, P=0.005^*$, partial $\eta^2=0.34$	$F_{(1,19)}=2.30, P=0.15$, partial $\eta^2=0.11$	$F_{(1,19)}=0.00, P=0.97$, partial $\eta^2=0.00$
V2	$F_{(1,19)}=11.60, P=0.003^*$, partial $\eta^2=0.38$	$F_{(1,19)}=0.94, P=0.35$, partial $\eta^2=0.38$	$F_{(1,19)}=0.58, P=0.46$, partial $\eta^2=0.03$
V3	$F_{(1,19)}=3.48, P=0.08$, partial $\eta^2=0.16$	$F_{(1,19)}=13.36, P=0.002^*$, partial $\eta^2=0.41$	$F_{(1,19)}=0.67, P=0.42$, partial $\eta^2=0.03$
V4	$F_{(1,19)}=16.64, P<0.001^*$, partial $\eta^2=0.47$	$F_{(1,19)}=45, P<0.001^*$, partial $\eta^2=0.70$	$F_{(1,19)}=2.21, P=0.15$, partial $\eta^2=0.10$

LOC	$F_{(1,19)}=0.08, P=0.78, \text{partial } \eta^2=0.00$	$F_{(1,19)}=24.04, P<0.001^*, \text{partial } \eta^2=0.56$	$F_{(1,19)}=0.06, P=0.81, \text{partial } \eta^2=0.00$
IPS0	$F_{(1,19)}=0.29, P=0.60, \text{partial } \eta^2=0.02$	$F_{(1,19)}=1.40, P=0.25, \text{partial } \eta^2=0.07$	$F_{(1,19)}=5.45, P=0.03, \text{partial } \eta^2=0.22$
IPS1	$F_{(1,19)}=0.03, P=0.87, \text{partial } \eta^2=0.00$	$F_{(1,19)}=2.41, P=0.14, \text{partial } \eta^2=0.11$	$F_{(1,19)}=0.53, P=0.47, \text{partial } \eta^2=0.03$
IPS2	$F_{(1,19)}=0.97, P=0.34, \text{partial } \eta^2=0.05$	$F_{(1,19)}=0.58, P=0.46, \text{partial } \eta^2=0.03$	$F_{(1,19)}=0.03, P=0.87, \text{partial } \eta^2=0.00$
SPL	$F_{(1,19)}=0.12, P=0.74, \text{partial } \eta^2=0.01$	$F_{(1,19)}=4.38, P=0.05, \text{partial } \eta^2=0.19$	$F_{(1,19)}=0.72, P=0.41, \text{partial } \eta^2=0.04$

Table 1. Results of the 2×2 repeated-measures ANOVA ($N=20$) with factors background (no clutter, high clutter) and attention (peripheral, fixation), analyzing location classification accuracies in 9 ROIs that were included in the analyses. Asterisks behind P -values indicate significance with FDR correction across number of comparisons (for 9 ROIs).

We proceed to investigate which visual processing stages are modulated by background and attention in an fMRI experiment, determining processing stages by localizing and assessing cortical regions of the ventral visual stream (Fig. 1C,D). In this context H_R predicts that location representations of objects emerge in higher regions along the ventral stream when objects are presented on cluttered backgrounds compared to blank backgrounds (Graumann et al., 2022; Fig. 1C). H_D predicts that attentional modulation is high where location information is high (Fig. 1D): with no clutter, attention modulates location representations throughout the ventral stream and with high clutter attention modulates location representations in mid- or high-level visual areas. H_S instead predicts that attentional modulation is high in mid- and high-level visual areas only, independent of background.

The design of the fMRI experiment was equivalent to the design in the EEG experiment with a reduced number of levels for the factors category and location. This adaptation was made to accommodate the longer trial duration required for our fMRI event-related design (Fig. 2D) while maintaining a feasible session duration. We presented objects from two categories (faces, cars) in two locations (left and right horizontally from fixation; Fig. 2B) instead of four categories and locations. To characterize the effects of background and spatial attention on location representations in visual cortex, we defined regions-of-interest (ROIs) along the ventral stream, since H_R predicted effects to emerge there based on previous studies (Hong et al., 2016; Graumann et al., 2022). We additionally included ROIs in the dorsal stream, since it has been implicated in visuospatial (Ungerleider and Haxby, 1994; Milner and Goodale, 2006;

349 Kravitz et al., 2011; Groen et al., 2022) and attentional processing (Silver et al., 2005;
350 Szczepanski et al., 2010; Sprague and Serences, 2013).

351

352 We classified location across category using an analogous classification scheme as in the EEG
353 experiment. We trained a classifier on fMRI patterns associated with two locations of one
354 object category and subsequently cross-validated the classifier on new testing data associated
355 with the same locations of a new object category. This classification was performed in each
356 ROI separately, for each level of the background condition (no clutter, high clutter) and for
357 each level of attention conditions (periphery, fixation) separately, resulting in four
358 classification accuracies per ROI and subject. We included three ROIs in early visual cortex
359 (V1, V2, V3), two ROIs in the ventral stream (V4, LOC) and four ROIs in the dorsal stream
360 (IPS0, IPS1, IPS2, SPL).

361

362 We tested H_R , H_S and H_D in 2×2 repeated-measures ANOVAs ($N=20$, FDR-correction for
363 multiple comparisons) with factors background (no clutter, high clutter) and attention
364 (peripheral, fixation) in all ROIs of the ventral and dorsal visual streams, focusing on the
365 ventral visual stream first.

366

367 Hypothesis H_R predicted a main effect of background in early visual areas, but not in high-
368 level visual areas of the ventral visual stream. Consistent with the predictions of H_R , we found
369 significant main effects of background in early visual areas V1 and V2 (Fig. 5A,B; V1:
370 $F_{(1,19)}=9.88$, $P=0.005$, partial $\eta^2=0.34$; V2: $F_{(1,19)}=11.56$, $P=0.003$, partial $\eta^2=0.3$), but not in
371 mid- and high-level visual areas V3 and LOC (Table 1), except for V4, which also showed a
372 main effect of background ($F_{(1,19)}=16.64$, $P<0.001$, partial $\eta^2=0.47$).

373

374 On this basis arbitrating between H_S and H_D we found clear evidence for H_S . Location
375 information in mid- and high-level ventral visual areas V3, V4 and LOC all showed a
376 significant main effect of attention (Fig. 5C,D,E; V3: $F_{(1,19)}=13.36$, $P=0.002$, partial $\eta^2=0.41$;
377 V4: $F_{(1,19)}=45$, $P<0.001$, partial $\eta^2=0.70$; LOC: $F_{(1,19)}=24.04$, $P<0.001$, partial $\eta^2=0.56$), but no
378 significant interaction between background and attention as would have been predicted by H_D .

379

380 Equivalent testing in the dorsal visual stream revealed no significant main or interaction effect
381 in any regions along the dorsal stream (Fig. 5F,G,H,I; Table 1), consistent with the observation

382 that object location representations emerge rather along the ventral than the dorsal stream
383 (Hong et al., 2016; Graumann et al., 2022).

384

385 In sum, the results of the fMRI experiment concur with the results of the EEG experiment in
386 providing evidence for H_R and H_s. Object location representations emerge gradually along the
387 processing hierarchy of the ventral visual stream, and attention modulates object location
388 representations in mid- and high-level ventral areas independent of the object's background.

389

390 **4 Discussion**

391 Using EEG and fMRI we investigated at which stage of the visual processing hierarchy
392 attention modulates object location representations. Our results converge across the two
393 experiments and imaging modalities into a common view. We reproduced the recent
394 observation that object location representations emerge at later processing stages when
395 presented on cluttered than on blank backgrounds (H_R) and showed that this holds independent
396 of attention. On this basis we examined the effect of attention on object location
397 representations, finding that attention modulated location representations statically during late
398 stages of visual processing in cortical time and space, independent of the object's background
399 (H_S).

400 **4.1 Disentangling the influences of background and attention on the temporal dynamics** 401 **of location representations**

402 Recent research has revealed that object location representations emerge later in the ventral
403 visual processing hierarchy when objects appear on cluttered rather than on blank backgrounds
404 (Hong et al., 2016; Graumann et al., 2022). However, it remained unclear to which degree this
405 effect relied on or was influenced by attention or not. Previous research has highlighted
406 attention as important for object perception under cluttered conditions (Treisman and Gelade,
407 1980; Wolfe, 1994; Reddy and Kanwisher, 2007; Lee and Maunsell, 2010). Further, both
408 temporal delays observed for object perception and attention have been related to recurrent
409 processes (Tang et al., 2014; Kar et al., 2019; Rajaei et al., 2019; van Bergen and Kriegeskorte,
410 2020), suggesting shared neural mechanisms.

411
412 Here, we clarify the relationship and find a dissociation: while cluttered viewing conditions
413 delay processing (see also time-generalization analysis in Supplementary Fig. 1,
414 Supplementary Methods 1), spatial attention in contrast increases information without
415 changing its timing. These results suggest that background clutter and attention have
416 differential effects on object location processing. Background clutter, like other factors that
417 increase image complexity, triggers local recurrent processes that can be measured in delayed
418 responses (Tang et al., 2014, 2018; Groen et al., 2018; Kar et al., 2019; Rajaei et al., 2019;
419 Seijdel et al., 2021; Graumann et al., 2022). In contrast, spatial attention triggers modulation
420 of neural responses that can be measured as enhancement of response magnitude (Desimone
421 and Duncan, 1995; Reynolds and Chelazzi, 2004; Briggs et al., 2013).

422 **4.2 Attention modulates location representations later than the initial bottom-up** 423 **response**

424 The EEG results revealed attentional modulation of location representations in a late time
425 window beyond the first 150 ms of the bottom-up response independent of the object's
426 background. In-depth investigation further revealed that both transient and persistent neural
427 components were modulated. This directly supports the hypothesis H_S that the processing stage
428 of attentional modulation is static and refutes the hypothesis H_D that the processing stage at
429 which attention modulates location representations changes dynamically.

430
431 Our results are seemingly at odds with earlier studies finding attentional modulation before 150
432 ms in the P1 (Hillyard et al., 1998b; Luck et al., 2000; Itthipuripat et al., 2019) and the N1
433 (Mangun, 1995; Hillyard et al., 1998a; Itthipuripat et al., 2019) component. How is this
434 discrepancy to be explained? We believe that the viewing conditions and the choice of the
435 stimulus are relevant. Above mentioned ERP studies employed simple artificial stimulation
436 conditions which might elicit attentional modulation already early. However, later studies
437 using naturalistic stimuli, comparable to the ones used here, did not find early attentional
438 modulation (VanRullen and Thorpe, 2001; Groen et al., 2016; Kaiser et al., 2016; Battistoni et
439 al., 2020). Together this questions the degree to which previously observed effects of early
440 attentional modulation generalize to more complex stimuli and naturalistic viewing conditions
441 encountered in the real world.

442
443 Another contributing factor to the discrepancy could be that attentional enhancement of early
444 neural responses is stronger when the visual task is more difficult or when visual processing is
445 overloaded (Spitzer et al., 1988; Lavie, 1995; Luck et al., 2000; Boudreau et al., 2006; Chen et
446 al., 2008) which might have been the case in earlier ERP studies e.g. by presenting stimuli in
447 faster sequences (Hillyard et al., 1998b). In contrast in our experiment, stimuli in the no clutter
448 condition were highly salient and presented long enough to be clearly visible. Future research
449 comparing attentional modulation of artificial vs. real-world stimuli with different levels of
450 task difficulty are needed to resolve this issue.

451 **4.3 Attentional modulation in mid- and high-level ventral visual areas**

452 Consistent with the EEG results indicating attentional modulation of later visual object
453 processing stages in time, the fMRI experiment localized those modulations to mid- and high-
454 level ventral visual areas. While our results do not exclude the existence of attentional

455 modulation also in early visual cortex as observed previously (Roelfsema et al., 1998; Gandhi
456 et al., 1999; Martínez et al., 2001; Noesselt et al., 2002; Khayat et al., 2006; Lakatos et al.,
457 2008; Briggs et al., 2013; Herrero et al., 2013; Itthipuripat et al., 2019), they suggest that the
458 modulation might be strongest and thus most likely to be detected at later stages of ventral
459 visual cortex (Murray and Wojciulik, 2004; Buffalo et al., 2010; Peelen and Kastner, 2014;
460 Kay et al., 2015). Our results add further evidence towards the view that attentional modulation
461 begins in higher processing stages and is then relayed back to lower stages (Buffalo et al.,
462 2010), which could be reflected as increasing attentional modulation along the ventral stream
463 (Kay et al., 2015).

464 **4.4 Location representations of objects on cluttered backgrounds in the ventral stream**

465 The fMRI results reveal a double dissociation between the effects of clutter and attention on
466 early and late ventral visual areas: early visual areas show an effect of background but not of
467 attention, while the reverse is true for mid- and high-level visual areas. Put differently, we find
468 that both robustness to clutter (Hong et al., 2016; Graumann et al., 2022) and attentional
469 modulation increase along the ventral visual stream (Buffalo et al., 2010; Kay et al., 2015). We
470 speculate that these phenomena depend on the common mechanistic and computational basis
471 of receptive field size increases along the ventral visual stream. Attention increases population
472 receptive field (pRF) size in higher-level ventral areas, thereby enhancing location sensitivity
473 (Kay et al., 2015). Such pRF size increase might also simultaneously benefit object
474 segmentation from cluttered backgrounds by encoding object location in global voxel patterns
475 (Eurich and Schwegler, 1997; Kay et al., 2015). This benefit for object segmentation might in
476 contrast not be present in early visual cortex where pRF size is small (Wandell and Winawer,
477 2015) and cells respond in a location unspecific way across all stimulated portions of the visual
478 field to both objects and background clutter.

479 **4.5 Limitations**

480 We highlight two limitations of our experimental designs that are important for the correct
481 interpretation of the results.

482

483 The first limitation is that in our experiment object locations and the content of the background
484 are randomly paired and thus incongruent. In contrast, in the real world objects typically appear
485 in locations congruent with the background scene. Attentional selection can exploit such
486 relations between objects and backgrounds (Wolfe et al., 2011; Kaiser et al., 2019; Vö et al.,

487 2019; Battistoni et al., 2020) on the basis of scene gist information (Oliva, 2005; Greene and
488 Oliva, 2009). In our experiment this type of information cannot be exploited. Thus, when object
489 locations and scene background are congruent, attentional modulation might be faster than
490 revealed here. The flipside of the limitation is that our experimental design isolates the effect
491 of clutter on visual processing and attentional modulation independent of congruency effects.
492 To determine the effect of congruency of object location and background on visual processing,
493 studies are needed that additionally investigate congruency as an experimental factor.

494

495 Another limitation is that we did not directly assess the behavioral effects of attentional
496 modulation on localization performance. Spatial attention benefits object localization in
497 cluttered displays (Treisman and Gelade, 1980; Wolfe, 1994; Wolfe et al., 2011) by increasing
498 processing speed. Future studies may combine assessment with brain imaging to link the effect
499 of attention for objects on cluttered backgrounds in brain and behavior.

500 **4.6 Conclusion**

501 In daily life, we use our spatial attention to help us focus on relevant portions of the visual field
502 in cluttered environments (Wolfe et al., 2011). Our results clarify that attention modulates
503 object location representations at late processing stages, using both spatial and temporal
504 markers. Furthermore, they establish that attentional modulation is a cognitive process which
505 is separate from recurrent processes which are engaged when objects appear in cluttered
506 environments.

507

508 5 Materials and Methods

509 5.1 Participants in EEG and fMRI experiment

510 27 participants completed the EEG experiment. One participant was excluded because of
511 technical problems, resulting in 26 participants (mean age 26.42 years, $SD=4.12$, 19 female)
512 included in the final EEG study. 23 participants completed the fMRI experiment, out of which
513 one also participated in the EEG experiment. Three participants were excluded because they
514 did not complete the whole experiment, resulting in 20 participants (mean age 26.71 years,
515 $SD=4.48$, 13 female) included in the final fMRI study.

516 All participants had no history of neurological disorders and normal or corrected-to-normal
517 vision. Participants provided informed consent prior to the studies and participation was
518 compensated with payment or course credit. The study was conducted in accordance with the
519 Declaration of Helsinki and the ethics committee of the Department of Education and
520 Psychology of the Freie Universität Berlin approved the study in advance.

521 5.2 Experimental design

522 5.2.1 EEG experimental design

523 The experimental design in the EEG study comprised the four factors object category (animals,
524 cars, faces, chairs, Fig. 2A left, with 3 exemplars per category), location (left up, left bottom,
525 right bottom, right up, Fig. 2A left center), background (no and high clutter, Fig. 2A center)
526 and attention (on periphery or on fixation, Fig. 2A right center). These four factors were fully
527 crossed, to investigate them independently of each other. In total, this created 192 individual
528 conditions (12 object exemplars \times 4 locations \times 2 background conditions \times 2 attention
529 conditions). For further analysis, data was collapsed across exemplars, so that data was
530 analyzed at the level of category. Thus, the number of conditions for further analysis was 64
531 (4 categories \times 4 locations \times 2 background conditions \times 2 attention conditions, Fig. 2A right).

532

533 5.2.2 fMRI experimental design

534 The experimental factors in the fMRI experiment were the same as in the EEG experiment, but
535 there were two instead of four levels for the factors category (cars, faces) and location (left,
536 right; Fig. 2B). This resulted in 48 individual conditions (6 object exemplars \times 2 locations \times 2
537 background conditions \times 2 attention conditions). For further analysis, data was likewise
538 collapsed across exemplars, so that data was analyzed at the level of category. Thus, the number

539 of conditions for further analysis was 16 (2 categories \times 2 locations \times 2 background conditions
540 \times 2 attention conditions).

541 **5.3 Stimulus set generation**

542 *5.3.1 Stimulus set generation: EEG experiment*

543 The experimental design in the EEG study comprised 96 individual stimulus conditions shown
544 in each attention condition, as detailed in the previous section. To create these stimuli, each
545 exemplar was superimposed onto backgrounds with or without scene images in four locations.
546 First, to position object exemplars onto the four image locations, we projected the 3D rendered
547 objects onto to the four quadrants of the screen (Fig. 2A, left center). Rendered objects did not
548 extend beyond a quadrant. Each object's center was positioned 3 degrees from the vertical and
549 3 degrees from the horizontal central midline (i.e., 4.2 degrees diagonally from image center
550 to fixation, Fig. 2A right), subtending 2.4 degrees ($SD=0.4$) in vertical and 2.2 degrees
551 ($SD=0.6$) in horizontal extent.

552

553 Second, each exemplar in each location was superimposed onto a background with no and with
554 high clutter (Fig. 2A, center; the backgrounds shown here are comparable to the original
555 backgrounds used in the experiments). We chose the background conditions no and high clutter
556 to compare visual stimuli with low and high image complexity, respectively (Groen et al.,
557 2018). The no clutter condition was a uniform gray background. In the high clutter condition,
558 we selected 60 natural scene images from the Places365 database
559 (<http://places2.csail.mit.edu/download.html>) that did not contain objects of the categories
560 included in our experimental design (i.e., no animals, cars, faces, chairs) and were highly
561 cluttered (as defined by 10 independent subject ratings; for methods and results see Graumann
562 et al., 2022). We converted the images to grayscale and superimposed a circular aperture of 15
563 degrees. Original backgrounds are not shown because of copyright reasons but are available
564 here: https://osf.io/85sak/?view_only=db183dde8f4b406aaba5dfc0dd0ae67d.

565

566 From the set of 60 scene images, we selected 48 scene images to go with the 48 stimulus
567 conditions within the high clutter condition (12 exemplars \times 4 locations). To avoid systematic
568 congruencies between objects and background images within the high clutter condition,
569 stimulus conditions and backgrounds were randomly paired for each of the 20 runs into which
570 the EEG experiment was divided (see below). Together with the 48 stimulus conditions in the

571 no clutter condition, this resulted in 96 individual images per run. The 12 remaining scene
572 images from the set of 60 were used to create catch trials.

573 5.3.2 Stimulus set generation: fMRI experiment

574 Stimulus set generation for the fMRI experiment was equivalent to the EEG experiment, with
575 the difference that objects were positioned on two instead of four image locations (Fig. 2B)
576 4.2° to the left or right of the image's center. In the fMRI experiment, each background
577 condition had 12 individual stimulus conditions (6 exemplars \times 2 locations). In combination
578 with the 12 stimulus conditions in the no clutter condition, this resulted in 24 individual images
579 per run. The 12 remaining scene images from the set of 60 were used to create 24 catch trials
580 (1 catch object \times 12 scene images \times 2 locations), which were randomly presented during the
581 fMRI experiment.

582 5.4 Experimental procedures

583 5.4.1 EEG main experiment

584 Each of the 26 participants completed one EEG recording session with 20 runs (run duration:
585 277 s). Overall, the EEG session lasted for 92 minutes. Participants performed attention tasks
586 on separate runs. The EEG recording session consisted of 10 periphery attention runs and 10
587 fixation attention runs in randomized order. Within each attention condition, there were 96
588 individual stimulus conditions (12 exemplars \times 4 locations \times background conditions). Runs
589 consisted of the presentation of regular trials and catch trials. In each run, there were 192
590 regular trials, representing the 96 stimulus condition images were presented twice. These trials
591 formed the basis for further analysis. On regular trials, digits between 1 and 9 were overlaid
592 for 117 ms each, followed by a 50 ms presentation of the image and fixation cross after each
593 digit (Fig. 2C). In total, stimuli were presented for 0.5 s followed by 0.5 or 0.6 s of ISI (equally
594 probable; Fig. 2C). Participants were asked to fixate their eyes on the central cross at all times.

595

596 On catch trials, a target was presented to which participants were asked to respond with button
597 press (Fig. 2E). These trials were excluded from the analyses. Catch trials were presented on
598 every 3rd to 5th trial (equally probable, in total 48 per run). Participants were instructed to
599 respond with button press to catch trials and to blink their eyes to minimize eye blink
600 contamination on subsequent trials. The ISI was 1s on catch trials to avoid contamination of
601 movement and eye blink artefacts on subsequent trials.

602

603 In the periphery and the fixation attention condition different trials were task-relevant catch
604 trials. In the periphery attention condition, catch trials were trials during which a target object
605 (a glass) was presented (Fig. 2E). The target could be presented at any of the four locations and
606 on any type of background. Digits on fixation were task-irrelevant in this attention condition.
607 In the fixation attention condition, catch trials were trials during which the digit 0 appeared
608 among any of the 3 digits that were presented on fixation during a single trial (Fig. 2E). The
609 presented object in the periphery was task-irrelevant in this attention condition (Fig. 2E). The
610 digit 0 never appeared on periphery attention runs and the glass never appeared on fixation
611 attention runs.

612 5.4.2 *fMRI main experiment*

613 Each of the 20 participants completed one fMRI recording session with 20 runs (run duration:
614 288 s). Overall, an fMRI recording in the main experiment lasted for 96 minutes. Each of the
615 24 images of the stimulus set was shown 3 times in random order without back-to-back
616 repetitions in each run. On each trial, the image was presented for 0.5 s at the center of a black
617 screen. The inter-stimulus-interval (ISI) was 2.5 s (Fig. 2D). Images were overlaid with a red
618 central cross for fixation. Participants were instructed to fixate their eyes on this cross
619 throughout the experiment. Every 3rd to 5th trial (equally probable, in total 18 per run) a catch
620 trial was presented. The tasks in the attention conditions and the catch objects were identical
621 to the EEG experiment (Fig. 2E). Catch trials were excluded from further analysis.

622

623 *fMRI localizer experiment.* Prior to the main fMRI experiment, participants completed a
624 separate localizer run to define ROIs in early visual, dorsal and ventral visual stream. We
625 presented images from three categories: faces, objects, and scrambled objects. Each image
626 showed identical versions of the same object located left and right of fixation to stimulate the
627 same retinotopic regions of visual cortex as the objects in the main experiment.

628

629 The localizer run lasted for 384 s, during which we presented 6 stimulation blocks. Each block
630 was 16 s long with presentations of 20 different objects from one of the three categories (500
631 ms on, 300 ms off) block-wise. Each block included two one-back image repetitions to which
632 participants had to respond to with a button press. The order of these blocks was first order
633 counterbalanced: triplets of stimulation blocks were presented in random order and
634 interspersed with blank background blocks.

635 **5.5 EEG acquisition and preprocessing**

636 To record EEG data, we used the EASYCAP 64-channel system with a Brainvision actiCHamp
637 amplifier at a sampling rate of 1,000 Hz and with an online filter between 0.03 and 100 Hz.
638 The signal was online re-referenced to FCz. Electrode placement followed the standard 10-10
639 system. Data was preprocessed offline with the EEGLAB toolbox version 14 (Delorme and
640 Makeig, 2004). This comprised a low-pass filter with a 50 Hz cut-off, trial epoching in a peri-
641 stimulus time window between -100 ms and 999 ms, and baseline-correction by subtracting
642 the mean of the 100 ms prestimulus time window from the entire epoch. We used independent
643 component analysis (ICA) to clean the data from ocular and muscular artefacts. To guide the
644 visual inspection of components for removal we used SASICA (Chaumon et al., 2015). To
645 identify horizontal eye movement components, we used external electrodes from the horizontal
646 electrooculogram (HEOG). We detected blink artefact and vertical eye movements using the
647 two frontal electrodes Fp1 and Fp2. On average, we removed 18 ($SD=5$) components per
648 participant. We finally applied multivariate noise normalization on the preprocessed data to
649 improve the signal-to-noise ratio and reliability of the data (Guggenmos et al., 2018).

650 **5.6 Preprocessing and univariate fMRI analysis**

651 *fMRI acquisition and preprocessing.* MRI data was recorded using a 12-channel head coil on
652 a 3T Siemens Tim Trio Scanner (Siemens, Erlangen, Germany). The structural image was
653 acquired with a T1-weighted sequence (MPRAGE; 1-mm³ voxel size). To acquire functional
654 data for the main experiment and the localizer run, we ran a T2*-weighted gradient-echo planar
655 sequence (TR=2, TE=30 ms, 70° flip angle, 3-mm³ voxel size, 37 slices, 20% gap, 192-mm
656 field of view, 64 × 64 matrix size, interleaved acquisition) on the entire brain.

657

658 fMRI data was preprocessed using SPM8 (<https://www.1ion.ucl.ac.uk/spm/>), involving
659 realignment, coregistration and normalization to the structural MNI template brain. We
660 smoothed functional data from the localizer run with an 8 mm FWHM Gaussian kernel, but
661 the data from the main experiment were not smoothed.

662

663 *Univariate fMRI analysis.* We modelled the fMRI responses of the experimental conditions at
664 the level of category. This was done for each run in the main experiment separately using a
665 general linear model (GLM). We entered onsets and durations of stimulus presentations per
666 category, pooling exemplars and repetitions. Thus, each GLM was estimated based on 9 trials
667 (3 exemplars × 3 condition repetitions per run) and was convolved with the hemodynamic

668 response function (hrf). We further entered movement parameters into the GLM as nuisance
669 regressors. This resulted in 8 beta maps per attention condition run (2 categories \times 2 locations
670 \times 2 backgrounds). For each run, we converted GLM parameter estimates into t -values by
671 contrasting each parameter estimate against the implicit baseline for each condition. This
672 resulted for each participant and attention condition run separately in 8 (2 categories \times 2
673 locations \times background conditions) t -value maps per condition. In sum, this resulted in 8 t -
674 value maps per 10 runs, per 2 attention conditions and per participant, which were later used
675 in the classification analysis.

676

677 For the fMRI responses to the localizer experiment, we modelled the responses to objects, faces
678 and scrambled objects by entering block onsets and durations as regressors of interest and
679 movement parameters as nuisance regressors into the GLM and convolved them with the hrf.
680 This resulted in three parameter estimates which we used to generate two contrasts that formed
681 part of ROI definitions. The first contrast was defined as objects and scrambled objects $>$
682 baseline and was used to localize activations in early, mid-level ventral and dorsal visual
683 regions (V1, V2, V3, V4, IPS0, IPS1, IPS2, SPL). The second contrast was defined as objects
684 and faces $>$ scrambled objects and was used to localize activations in object-selective area
685 LOC. Overall, this yielded two t -value maps for the localizer run for each participant.

686

687 *Definition of regions of interest.* To define ROIs, we first applied anatomical masks and then
688 selected voxels using appropriate contrasts from the functional localizer run. In detail, we first
689 defined ROIs using anatomical masks from a probabilistic atlas (Wang et al., 2015) and
690 combined these for both hemispheres. We included three masks in early visual cortex V1, V2
691 and V3. V4 and LOC served as ROIs in mid- and high-level ventral visual cortex. We also
692 included four ROIs from dorsal visual cortex: IPS0, IPS1, IPS2 and SPL. We removed all
693 overlapping voxels from these masks to avoid overlap between ROIs. The second step entailed
694 selecting the most activated voxels of the participant-specific t -value maps of the localizer run
695 within the previously defined anatomical masks. To keep the number of voxels constant
696 between ROIs and participants, we determined the smallest ROI in any participant when
697 overlaying the localizer t -value maps and the anatomical masks. This resulted in a minimum
698 ROI size of 288 voxels. This was then the number of most activated voxels to select of the
699 participant-specific localizer t -value maps within all anatomical masks and participants. To
700 select voxels in LOC we used the objects & faces $>$ scrambled contrast and to select voxels in
701 the remaining ROIs we used the objects & scrambled objects $>$ baseline contrast. This resulted

702 in ROI definitions that were specific to each participant with an equal number of voxels across
703 ROIs and participants.

704 **5.7 Object location classification from brain measurements**

705 To measure location information in time using EEG and in space using fMRI, we applied
706 multivariate classification (Carlson et al., 2011a; Cichy et al., 2011, 2013; Isik et al., 2014) of
707 object location. Since object location and object category have partly overlapping neural
708 fingerprints in time and space (Cichy et al., 2011; Graumann et al., 2022), we applied a cross-
709 classification scheme that avoided location information results to be confounded with category
710 information (Carlson et al., 2011b; Isik et al., 2014). For this, we cross-classified locations
711 across categories, meaning that during each classification of a given location pair, we trained
712 and tested on different object categories. For all classification analyses described, we employed
713 a binary c-support vector classification (C-SVC) with a linear kernel from the libsvm toolbox
714 (Chang and Lin, 2011) (<https://www.csie.ntu.edu.tw/~cjlin/libsvm>). This cross-classification
715 scheme was applied separately within each background condition, within each attention
716 condition and within each individual participant. The classification scheme was adapted to the
717 specifics of the methods used here: it was applied per time point on the EEG data and per ROI
718 in the fMRI data.

719

720 *Time-resolved classification of location from EEG data.* The time-resolved EEG classification
721 analysis (Carlson et al., 2011b; Isik et al., 2014) served to determine the temporal dynamics
722 with which category-independent location information emerged in the brain.

723

724 For each time point of the epoched EEG data, we extracted activations from 33 EEG channels.
725 We chose the 33 central and posterior channels starting from the central midline, because we
726 were interested in visual responses and previous studies had shown that location information
727 was most pronounced in those areas (Graumann et al., 2022). We arranged activations from
728 these channels into pattern vectors of 64 conditions and 60 raw trials. Raw trials were randomly
729 arranged into four bins of 15 trials each and averaged by bin into four pseudo-trials to increase
730 SNR. The classification procedure was repeated 100 times, each time assigning random trials
731 into the bins before averaging into pseudo-trials. For classification, three of the pseudo-trials
732 that came from two location conditions of the same category went into the training set. The
733 model resulting from SVM classifier training was then tested on other pseudo-trials coming
734 from the same two location conditions, but from a different category. The accuracy of the

735 classification procedure was measured in percent classification accuracy (50% chance level).
736 This amounted to 6 pairwise location classifications since we had 4 locations that were all
737 classified pairwise once. During each iteration of pairwise location classification, the SVM was
738 trained and tested across all combinations of the four categories in the training and testing set.
739 For example, for a given location classification, the SVM was trained on faces and tested on
740 animals (Fig. 3A). Then the same procedure was applied combining the remaining categories.
741 With four categories in total, this resulted in 6 classification iterations to combine all categories
742 into training and testing pairs. The direction of all training and testing pairs was reversed once
743 (e.g., training on animals and testing on faces and vice versa), yielding a total of 12
744 classification iterations per pairwise location classification. We averaged 72 (6 location pairs
745 \times 12 category train/test pairs) classification accuracies in total per iteration. With 100 iterations
746 with random trial assignment into pseudo-trials, this resulted in 7,200 classification accuracies
747 that were averaged per background condition, attention condition and participant. The result
748 reflects the amount of location information that is independent of category at each time point,
749 and within a background condition, attention condition and participant.

750

751 *Time-resolved EEG searchlight in sensor space.* To gain insights into which EEG channels
752 contained the highest amount of location information we conducted a time-resolved EEG
753 searchlight analysis in EEG channel space. This analysis followed the same scheme as the time-
754 resolved EEG classification described above but extended it by one step: For each EEG channel
755 c , the classification procedure was conducted not on all 33, but on the five closest channels
756 surrounding c . The resulting classification accuracy was stored at the position of c . Iterating
757 across all EEG channels with a temporal resolution downsampled to 10 ms steps, this yielded
758 a map of classification accuracy across all channels and downsampled time points, for each
759 participant, background condition and attention condition.

760

761 *Time generalization analysis of location from EEG data.* To characterize the neural dynamics
762 of object location representations across time, we used temporal generalization analysis
763 (Carlson et al., 2011b; Cichy et al., 2014; Isik et al., 2014; King and Dehaene, 2014).

764

765 In this analysis, the classification scheme was the same as in the time-resolved EEG
766 classification but with the following extension: besides training and testing the SVM on data
767 from the same time point, we additionally tested the SVM on data from all other time points
768 within a -100 to 600 ms peristimulus time window, downsampled to a 10 ms temporal

769 resolution. This resulted in a two-dimensional matrix of classification accuracies, indexed in
770 rows and columns by the time points of data used for training and testing the SVM. This matrix
771 indicates how much location information was shared at a given combination of time points.
772 This analysis was conducted within time point combination, background condition, attention
773 condition and participant.

774

775 *Multivariate fMRI ROI analysis.* The fMRI ROI classification analysis served to determine
776 where category-independent location information emerged in the brain. For each ROI of the
777 fMRI data, we extracted and arranged t -values into pattern vectors, one for each of the 16
778 conditions and 10 runs of the main experiment. Raw trials were randomly arranged into five
779 bins with two runs each and averaged by bin into five pseudo-runs to increase SNR. We then
780 proceeded with a 5-fold leave-one-pseudo-run-out-cross validation procedure. During each
781 classification iteration, we trained an SVM on 4 and tested it on one pseudo-trial. The
782 classification scheme was conceptually equivalent to the EEG classification. Training and
783 testing was conducted across the two different categories, with each being in the training set
784 once. We averaged across the two different training and testing directions of the two categories.
785 The result reflects how much category-tolerant location information was present for each ROI,
786 participant, background and attention condition separately.

787 **5.8 Statistical testing**

788 *Wilcoxon signed-rank tests.* To test for above-chance classification accuracy at time points in
789 the EEG time courses, in the EEG time-generalization matrix and for above-chance
790 classification in the fMRI ROI results, we performed non-parametric two-tailed Wilcoxon
791 signed-rank tests. The null hypothesis was always that the parameter being tested (i.e.,
792 classification accuracy) came from a distribution with a median of chance level (i.e., 50%
793 classification accuracy for pairwise classification). We corrected the resulting P -values for
794 multiple comparisons using false discovery rate at 5% level in every case where more than
795 one test was conducted.

796

797 *Bootstrap tests.* To estimate confidence intervals and to compute the significance of peak-to-
798 peak latency differences in the EEG time courses we used bootstrapping. We bootstrapped
799 the participant pool 10,000 times with replacement and calculated the statistic of interest for
800 each of the bootstrap samples.

801

802 For the peak-to-peak latency differences in the EEG time courses, we bootstrapped the latency
803 difference between the peaks of the two time courses being compared. This resulted in a
804 bootstrapped distribution that could be compared to zero. To determine the significance of
805 peak-to-peak latencies in the EEG time courses, we computed the proportion of values that
806 were equal to or smaller than zero and corrected them for multiple comparisons using FDR at
807 $P=0.05$. For computing the 95% confidence intervals of peak latencies of each time course, we
808 bootstrapped the peak and computed the 95% percentiles of this distribution.

809

810 *ANOVAs*. We used repeated-measures ANOVAs to test for main effects and the interaction
811 between the factors background and attention within ROIs. Since both factors had two levels,
812 the assumption of sphericity was always met.

813

814 All post-hoc tests were conducted using pairwise *t*-tests and *P*-values were corrected for
815 multiple comparisons using Tukey correction.

816

817 **Data availability**

818 The fMRI and EEG data will be publicly available at the time of publication via
819 <https://osf.io/hf6zp/>.

820

821 **Code availability**

822 Analysis code will be publicly available at the time of publication via
823 <https://github.com/graumannm/AttentionLocation>.

824

825 **Acknowledgements**

826 We thank Benjamin Lahner for the glass brain plots. Computing resources were provided by
827 the high-performance computing facilities at ZEDAT, Freie Universität Berlin. EEG and fMRI
828 data were acquired at the Center for Cognitive Neuroscience, Freie Universität Berlin, Berlin.
829 M.G. and R.M.C. are supported by German Research Council (DFG) (CI241/1-1, CI241/3-1,
830 CI241/7-1). R.M.C. is supported by the European Research Council (ERC-StG-2018-803370).
831 L.A.W. is supported by the University of Konstanz. The funders had no role in study design,
832 data collection and analysis, decision to publish or preparation of the manuscript.

833
834 **Author contributions**

835 M.G. and R.M.C. designed research. M.G. and L.A.W. performed experiments. M.G. and
836 L.A.W. performed EEG preprocessing. M.G. performed data analyses. M.G. and R.M.C. wrote
837 the manuscript.

838
839 **Competing interests**

840 The authors declare no competing interests.

841

842 **6 References**

- 843 Battistoni E, Kaiser D, Hickey C, Peelen M V (2020) The time course of spatial attention
844 during naturalistic visual search. *Cortex* 122:225–234.
- 845 Boudreau CE, Williford TH, Maunsell JHR (2006) Effects of task difficulty and target
846 likelihood in area V4 of macaque monkeys. *J Neurophysiol* 96:2377–2387.
- 847 Briggs F, Mangun GR, Usrey WM (2013) Attention enhances synaptic efficacy and the
848 signal-to-noise ratio in neural circuits. *Nature* 499:476–480.
- 849 Buffalo EA, Fries P, Landman R, Liang H, Desimone R (2010) A backward progression of
850 attentional effects in the ventral stream. *Proc Natl Acad Sci* 107:361–365.
- 851 Camprodon JA, Zohary E, Brodbeck V, Pascual-Leone A (2010) Two phases of V1 activity
852 for visual recognition of natural images. *J Cogn Neurosci* 22:1262–1269.
- 853 Carlson TA, Hogendoorn H, Fonteijn H, Verstraten FA (2011a) Spatial coding and
854 invariance in object-selective cortex. *Cortex* 47:14–22.
- 855 Carlson TA, Hogendoorn H, Kanai R, Mesik J, Turret J (2011b) High temporal resolution
856 decoding of object position and category. *J Vis* 11:1–17.
- 857 Chang C-C, Lin C-J (2011) Libsvm: A library for support vector machines. *ACM Trans Intell*
858 *Syst Technol* 2:1–27.
- 859 Chaumon M, Bishop DVM, Busch NA (2015) A practical guide to the selection of
860 independent components of the electroencephalogram for artifact correction. *J Neurosci*
861 *Methods* 250:47–63.
- 862 Chen Y, Martinez-Conde S, Macknik SL, Bereshpolova Y, Swadlow HA, Alonso JM (2008)
863 Task difficulty modulates the activity of specific neuronal populations in primary visual
864 cortex. *Nat Neurosci* 11:974–982.
- 865 Cichy RM, Chen Y, Haynes JD (2011) Encoding the identity and location of objects in
866 human LOC. *Neuroimage* 54:2297–2307.
- 867 Cichy RM, Pantazis D, Oliva A (2014) Resolving human object recognition in space and
868 time. *Nat Neurosci* 17:455–462.
- 869 Cichy RM, Sterzer P, Heinzle J, Elliott LT, Ramirez F, Haynes J-D (2013) Probing principles
870 of large-scale object representation: Category preference and location encoding. *Hum*
871 *Brain Mapp* 34:1636–1651.
- 872 Delorme A, Makeig S (2004) EEGLAB: An open source toolbox for analysis of single-trial
873 EEG dynamics including independent component analysis. *J Neurosci Methods* 134:9–
874 21.
- 875 Desimone R, Duncan J (1995) Selective visual attention. *Annu Rev Neurosci* 18:193–222.
- 876 DiCarlo JJ, Cox DD (2007) Untangling invariant object recognition. *Trends Cogn Sci*
877 11:333–341.
- 878 Eurich CW, Schwegler H (1997) Coarse coding: Calculation of the resolution achieved by a
879 population of large receptive field neurons. *Biol Cybern* 76:357–363.
- 880 Fahrenfort JJ, Scholte HS, Lamme VAF (2007) Masking disrupts reentrant processing in
881 human visual cortex. *J Cogn Neurosci* 19:1488–1497.
- 882 Gandhi SP, Heeger DJ, Boynton GM (1999) Spatial attention affects brain activity in human
883 primary visual cortex. *Proc Natl Acad Sci U S A* 96:3314–3319.
- 884 Graumann M, Ciuffi C, Dwivedi K, Roig G, Martin R (2022) The spatiotemporal neural
885 dynamics of object location representations in the human brain. *Nat Hum Behav*:1–38.
- 886 Greene M, Oliva A (2009) The Briefest of Glances: The Time Course of Natural Scene
887 Understanding. *Psychol Sci* 20:464–472.
- 888 Groen IIA, Dekker TM, Knapen T, Silson EH (2022) Visuospatial coding as ubiquitous
889 scaffolding for human cognition. *Trends Cogn Sci* 26:81–96.
- 890 Groen IIA, Ghebreab S, Lamme VAF, Scholte HS (2016) The time course of natural scene

891 perception with reduced attention. *J Neurophysiol*.

892 Groen IIA, Jahfari S, Seijdel N, Ghebreab S, Lamme VAF, Scholte HS (2018) Scene
893 complexity modulates degree of feedback activity during object detection in natural
894 scenes. *PLoS Comput Biol* 14:e1006690.

895 Guggenmos M, Sterzer P, Cichy RM (2018) Multivariate pattern analysis for MEG: A
896 comparison of dissimilarity measures. *Neuroimage* 173:434–447.

897 Herrero JL, Gieselmann MA, Sanayei M, Thiele A (2013) Attention-induced variance and
898 noise correlation reduction in macaque V1 is mediated by NMDA receptors. *Neuron*
899 78:729–739.

900 Hillyard SA, Teder-Sälejärvi WA, Münte TF (1998a) Temporal dynamics of early perceptual
901 processing. *Curr Opin Neurobiol* 8:202–210.

902 Hillyard SA, Vogel EK, Luck SJ (1998b) Sensory gain control (amplification) as a
903 mechanism of selective attention: Electrophysiological and neuroimaging evidence.
904 *Philos Trans R Soc B Biol Sci* 353:1257–1270.

905 Hong H, Yamins DLK, Majaj NJ, DiCarlo JJ (2016) Explicit information for category-
906 orthogonal object properties increases along the ventral stream. *Nat Neurosci* 19:613–
907 622.

908 Isik L, Meyers EM, Leibo JZ, Poggio T (2014) The dynamics of invariant object recognition
909 in the human visual system. *J Neurophysiol* 111:91–102.

910 Itthipuripat S, Sprague TC, Serences JT (2019) Functional MRI and EEG index
911 complementary attentional modulations. *J Neurosci* 39:6162–6179.

912 Kaiser D, Oosterhof NN, Peelen M V. (2016) The neural dynamics of attentional selection in
913 natural scenes. *J Neurosci* 36:10522–10528.

914 Kaiser D, Quek GL, Cichy RM, Peelen M V. (2019) Object vision in a structured world.
915 *Trends Cogn Sci* 23:672–685

916 Kar K, Kubilius J, Schmidt K, Issa EB, DiCarlo JJ (2019) Evidence that recurrent circuits are
917 critical to the ventral stream’s execution of core object recognition behavior. *Nat*
918 *Neurosci* 22:974–983.

919 Kay KN, Weiner KS, Grill-Spector K (2015) Attention reduces spatial uncertainty in human
920 ventral temporal cortex. *Curr Biol* 25:595–600.

921 Khayat PS, Spekreijse H, Roelfsema PR (2006) Attention lights up new object
922 representations before the old ones fade away. *J Neurosci* 26:138–142.

923 King JR, Dehaene S (2014) Characterizing the dynamics of mental representations: The
924 temporal generalization method. *Trends Cogn Sci* 18:203–210.

925 Koivisto M, Railo H, Revonsuo A, Vanni S, Salminen-Vaparanta N (2011) Recurrent
926 processing in V1/V2 contributes to categorization of natural scenes. *J Neurosci*
927 31:2488–2492.

928 Kravitz DJ, Saleem KS, Baker CI, Mishkin M (2011) A new neural framework for
929 visuospatial processing. *Nat Rev Neurosci* 12:217–230.

930 Lakatos P, Karmos G, Mehta AD, Ulbert I, Schroeder CE (2008) Entrainment of neuronal
931 oscillations as a mechanism of attentional selection. *Science* 320:110–113.

932 Lamme VAF, Roelfsema PR (2000) The distinct modes of vision offered by feedforward and
933 recurrent processing. *Trends Neurosci* 23:571–579.

934 Lavie N (1995) Perceptual Load as a Necessary Condition for Selective Attention. *J Exp*
935 *Psychol Hum Percept Perform* 21:451–468.

936 Lee J, Maunsell JHR (2010) Attentional modulation of MT neurons with single or multiple
937 stimuli in their receptive fields. *J Neurosci* 30:3058–3066.

938 Luck SJ, Woodman GF, Vogel EK (2000) Event-related potential studies of attention. *Trends*
939 *Cogn Sci* 4:432–440.

940 Mangun GR (1995) Neural mechanisms of visual selective attention. *Psychophysiology*

941 32:4–18.

942 Martínez A, DiRusso F, Anllo-Vento L, Sereno MI, Buxton RB, Hillyard SA (2001) Putting
943 spatial attention on the map: Timing and localization of stimulus selection processes in
944 striate and extrastriate visual areas. *Vision Res* 41:1437–1457.

945 Milner AD, Goodale MA (2006) *The visual brain in action*. Oxford: Oxford University Press.

946 Murray SO, Wojciulik E (2004) Attention increases neural selectivity in the human lateral
947 occipital complex. *Nat Neurosci* 7:70–74.

948 Noesselt T, Hillyard SA, Woldorff MG, Schoenfeld A, Hagner T, Jäncke L, Tempelmann C,
949 Hinrichs H, Heinze HJ (2002) Delayed striate cortical activation during spatial attention.
950 *Neuron* 35:575–587.

951 Oliva A (2005) *Gist of the scene*. Elsevier Inc.

952 Peelen M V., Kastner S (2011) A neural basis for real-world visual search in human
953 occipitotemporal cortex. *Proc Natl Acad Sci U S A* 108:12125–12130.

954 Peelen M V., Kastner S (2014) Attention in the real world: Toward understanding its neural
955 basis. *Trends Cogn Sci* 18:242–250

956 Rajaei K, Mohsenzadeh Y, Ebrahimpour R, Khaligh-Razavi S-M (2019) Beyond core object
957 recognition: Recurrent processes account for object recognition under occlusion. *PLOS*
958 *Comput Biol* 15:e1007001.

959 Reddy L, Kanwisher N (2007) Category Selectivity in the Ventral Visual Pathway Confers
960 Robustness to Clutter and Diverted Attention. *Curr Biol* 17:2067–2072.

961 Reynolds JH, Chelazzi L (2004) Attentional modulation of visual processing. *Annu Rev*
962 *Neurosci* 27:611–647.

963 Roelfsema PR, Lamme VAF, Spekreijse H (1998) Object-based attention in the primary
964 visual cortex of the macaque monkey. *Nature* 395:376–381.

965 Seijdel N, Loke J, van de Klundert R, van der Meer M, Quispel E, van Gaal S, de Haan EHF,
966 Scholte HS (2021) On the necessity of recurrent processing during object recognition: It
967 depends on the need for scene segmentation. *J Neurosci* 41:6281–6289.

968 Seijdel N, Tsakmakidis N, De Haan EHF, Bohte SM, Scholte HS (2020) Depth in
969 convolutional neural networks solves scene segmentation. *PLoS Comput Biol*
970 16:e1008022

971 Silver MA, Ress D, Heeger DJ (2005) Topographic maps of visual spatial attention in human
972 parietal cortex. *J Neurophysiol* 94:1358–1371.

973 Spitzer H, Desimone R, Moran J (1988) Increased attention enhances both behavioral and
974 neuronal performance. *Science* 240:338–340.

975 Spoerer CJ, Kietzmann TC, Mehrer J, Charest I, Kriegeskorte N (2020) Recurrent neural
976 networks can explain flexible trading of speed and accuracy in biological vision. *PLoS*
977 *Comput Biol* 16:e1008215.

978 Sprague TC, Serences JT (2013) Attention modulates spatial priority maps in the human
979 occipital, parietal and frontal cortices. *Nat Neurosci* 16:1879–1887.

980 Szczepanski SM, Konen CS, Kastner S (2010) Mechanisms of spatial attention control in
981 frontal and parietal cortex. *J Neurosci* 30:148–160.

982 Tang H, Buia C, Madhavan R, Crone NE, Madsen JR, Anderson WS, Kreiman G (2014)
983 Spatiotemporal dynamics underlying object completion in human ventral visual cortex.
984 *Neuron* 83:736–748.

985 Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, Caro JO, Hardesty W, Cox D,
986 Kreiman G (2018) Recurrent computations for visual pattern completion. *Proc Natl*
987 *Acad Sci U S A* 115:8835–8840.

988 Treisman AM, Gelade G (1980) A feature-integration theory of attention. *Cogn Psychol*
989 12:97–136.

990 Ungerleider L, Haxby J V. (1994) “What” and “where” in the human brain. *Curr Opin*

991 Neurobiol 4:157–165.
992 van Bergen RS, Kriegeskorte N (2020) Going in circles is the way forward: the role of
993 recurrence in visual inference. *Curr Opin Neurobiol* 65:176–193.
994 VanRullen R, Thorpe SJ (2001) The time course of visual processing: From early perception
995 to decision-making. *J Cogn Neurosci* 13:454–461.
996 Vö MLH, Boettcher SE, Draschkow D (2019) Reading scenes: How scene grammar guides
997 attention and aids perception in real-world environments. *Curr Opin Psychol* 29:205–
998 210.
999 Wandell BA, Winawer J (2015) Computational neuroimaging and population receptive
1000 fields. *Trends Cogn Sci* 19:349–357.
1001 Wang L, Mruczek REB, Arcaro MJ, Kastner S (2015) Probabilistic maps of visual
1002 topography in human cortex. *Cereb Cortex* 25:3911–3931.
1003 Wolfe JM (1994) Visual search in continuous, naturalistic stimuli. *Vision Res* 34:1187–1195.
1004 Wolfe JM, Vö MLH, Evans KK, Greene MR (2011) Visual search in scenes involves
1005 selective and nonselective pathways. *Trends Cogn Sci* 15:77–84.
1006 Wyatte D, Jilk DJ, O’Reilly RC (2014) Early recurrent feedback facilitates visual object
1007 recognition under challenging conditions. *Front Psychol* 5:1–10.
1008

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: Graumann, Monika

Institute: Department of Education and Psychology, Freie Universität Berlin

Doctoral study subject: Psychology

Title: Location representations of objects in cluttered scenes in the human brain

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

1. Graumann, M., Ciuffi, C., Dwivedi, D., Roig, G., & Cichy, R. M. (2022). The spatiotemporal neural dynamics of object location representations in the human brain. Published in Nature Human Behavior.

Published in Nature Human Behavior.

2. Graumann, M., Wallenwein, L. A., & Cichy R. M. (submitted). Independent spatiotemporal effects of spatial attention and background clutter on human object location representations. Submitted at eLife May 2022, uploaded on Biorxiv.

III. Explanation of own share of these works:

Regarding II. 1.: Study conceptualisation and design (vast majority), programming of paradigm (vast majority), data collection (vast majority), data analysis (all), discussion of results (vast majority), writing/revising the manuscript (vast majority)

Regarding II. 2.: Study conceptualisation and design (vast majority), programming of paradigm (all), data collection (vast majority), data analysis (all), discussion of results (vast majority), writing/revising the manuscript (vast majority)

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

Regarding II. 1.: Graumann, M., (1), monikag@zedat.fu-berlin.de
Ciuffi, C., (1),
Dwivedi, D., (2),
Roig, G., (2),
Cichy, R. M. (1), rmcichy@zedat.fu-berlin.de

Regarding II. 2.: Graumann, M.: see above
Lara A. Wallenwein,
AG Mier
Department of Psychology
Universität Konstanz
Universitätsstrasse 10
78457 Konstanz
Cichy, R. M.: see above

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

14195 Berlin
(2) Computational Vision & Artificial Intelligence
Department of Computer Science
Goethe Universität
Robert-Mayer-Str. 11-15
60325 Frankfurt am Main

Date, doctoral candidate signature 28.04.2022.....

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

Name: Radoslaw Martin Cichy Signature:

Name: Caterina Ciuffi Signature:

Name: Kshitij Dwivedi Signature:

Name: Gemma Roig Signature:

Name: Lara Alicia Wallenwein Signature:

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: Graumann, Monika

Institute: Department of Education and Psychology, Freie Universität Berlin

Doctoral study subject: Psychology

Title: Location representations of objects in cluttered scenes in the human brain

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

1. Graumann, M., Ciuffi, C., Dwivedi, D., Roig, G., & Cichy, R. M. (2022). The spatiotemporal neural dynamics of object location representations in the human brain. Published in Nature Human Behavior.

Published in Nature Human Behavior.

2. Graumann, M., Wallenwein, L. A., & Cichy R. M. (submitted). Independent spatiotemporal effects of spatial attention and background clutter on human object location representations. Submitted at eLife May 2022, uploaded on Biorxiv.

III. Explanation of own share of these works:

Regarding II. 1.: Study conceptualisation and design (vast majority), programming of paradigm (vast majority), data collection (vast majority), data analysis (all), discussion of results (vast majority), writing/revising the manuscript (vast majority)

Regarding II. 2.: Study conceptualisation and design (vast majority), programming of paradigm (all), data collection (vast majority), data analysis (all), discussion of results (vast majority), writing/revising the manuscript (vast majority)

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

Regarding II. 1.: Graumann, M., (1), monikag@zedat.fu-berlin.de
Ciuffi, C., (1),
Dwivedi, D., (2),
Roig, G., (2),
Cichy, R. M. (1), rmcichy@zedat.fu-berlin.de

Regarding II. 2.: Graumann, M.: see above
Lara A. Wallenwein,
AG Mier
Department of Psychology
Universität Konstanz
Universitätsstrasse 10
78457 Konstanz
Cichy, R. M.: see above

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

14195 Berlin
(2) Computational Vision & Artificial Intelligence
Department of Computer Science
Goethe Universität
Robert-Mayer-Str. 11-15
60325 Frankfurt am Main

Date, doctoral candidate signature 28.04.2022.....

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

Name: Radoslaw Martin Cichy Signature:

Name: Caterina Ciuffi Signature:

Name: Kshitij Dwivedi Signature:

Name: Gemma Roig Signature:

Name: Lara Alicia Wallenwein Signature:

Eidesstattliche Erklärung

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, 04. Mai 2022

Monika Graumann