



Natalia Rutkowska

**Scene–object interactions in naturalistic vision: insights from
behavioural, neurostimulation, and computational studies**

PhD thesis

Completed in the Laboratory of Brain Imaging
of the Nencki Institute of Experimental Biology
Polish Academy of Sciences

SUPERVISORS:

Dr. Michał Bola, PhD, Dsc

Prof. Marius Peelen, PhD

Warsaw, 2025

Acknowledgements

During my long and challenging PhD journey, I have been fortunate to meet many wonderful people. Although only some can be mentioned here, I am grateful to everyone whose support, advice, or companionship helped me through this time.

As a substantial part of the research presented in this thesis was carried out in the Netherlands, I would like to begin by expressing my deep gratitude to the people who made this work possible. I am especially thankful to Prof. Marius Peelen for welcoming me into his lab and for creating a truly stimulating and supportive scientific environment. I am grateful to Dr. Marco Gandolfo for sharing his expertise in TMS research and for guiding each stage of the experimental process. I would also like to thank Aaron Schnippe for his involvement and dedication to this project, and for his support during its critical moments. I am deeply grateful to the members of the Visual Cognitive Neuroscience Lab, who instantly made me feel at home and whose thoughtful discussions greatly enriched my understanding of visual cognition. Finally, I am very thankful to Sjors Reith for serving as the first pilot in the TMS project and for supporting my stay in the Netherlands throughout the work on this study.

My sincere thanks go to Dr. Maksymilian Bielecki for sharing his expertise in data analysis and for his unfailing openness and patience with all my questions.

I am also grateful to the Nencki Open Lab community for many important discussions and initiatives that enriched my understanding of neuroscience and encouraged me to consider it from multiple perspectives.

I am deeply thankful to my colleagues from the Laboratory of Brain Imaging. There were moments when I doubted that this project would ever be finished, and your support and confidence in me were deeply moving, helping me endure the most difficult periods and bring this work to completion.

Finally, I am grateful to my friends outside the lab for their support, encouragement and understanding throughout this journey. It is deeply reassuring to know that I can rely on such people, even in the most difficult and unpredictable moments.

This research was funded by the National Science Centre Poland (Grant No. 2018/29/B/HS6/02152 awarded to Dr. Michał Bola, Ph.D., D.Sc.) and by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie fellowship (grant agreement No. 101033489 awarded to Dr. Marco Gandolfo).



Funded by the Horizon 2020
Framework Programme of the
European Union

Table of contents

Abstract	6
Streszczenie	8
Abbreviations	10
1. Introduction	12
1.1. Naturalistic vision in scientific research	12
1.2. Real-world scenes as a subject of scientific investigation	14
1.3. Neural mechanisms of real-world scene perception	16
1.4. Research methods in real-world scene perception	20
1.5. Temporal dynamics of real-world scene perception	22
1.5.1. Analytic view	22
1.5.2. Holistic view	23
1.5.3. Predictive processing view	25
1.6. Aims of the thesis	28
2. Experimental studies: scene–object interactions in naturalistic vision	30
2.1. Study 1: testing the hierarchical models of scene–object interactions	30
2.1.1. Research question and hypotheses	30
2.1.2. Methods	32
2.1.2.1. Participants	32
2.1.2.2. Stimuli	33
2.1.2.3. Apparatus	34
2.1.2.4. Procedure	34
2.1.2.4.1. Experiment 1 – procedure	34
2.1.2.4.2. Experiment 2 – procedure	35
2.1.2.5. Data analysis	36
2.1.2.5.1. Analysis of behavioral data	36
2.1.2.5.1.1 Accuracy	37
2.1.2.5.1.2 Reaction times	37
2.1.2.5.1.3 Balanced Integration Score (BIS)	38
2.1.2.5.1.4. Statistical analysis	38
2.1.3. Behavioural results	39
2.1.3.1 Experiment 1: go/no-go	39
2.1.3.1.1. Accuracy	39
2.1.3.1.2. Reaction times	41
2.1.3.1.3. BIS	46
2.1.3.2 Experiment 2: 2AFC	48
2.1.3.2.1. Accuracy	48
2.1.3.2.2. Reaction times	49
2.1.3.2.3. BIS	52

2.1.4. Discussion	54
2.2. Study 2: testing the causal role of object representations in disambiguating scenes	57
2.2.1. Research questions and hypotheses	57
2.2.2. Methods	62
2.2.2.1. Participants	62
2.2.2.2. Stimuli	63
2.2.2.2.1. Creation of stimuli	63
2.2.2.2.1.1. Selection of images	64
2.2.2.2.1.2. Pilot study for the TMS experiment	68
2.2.2.3. Apparatus	69
2.2.2.4. Procedure	71
2.2.2.4.1. Four-pulse TMS study	71
2.2.2.4.1.1. Object recognition task	71
2.2.2.4.1.2. Scene classification task	74
2.2.2.4.1.2. Selection procedure	76
2.2.2.4.2. Chronometric TMS study	76
2.2.2.5. Data analysis	77
2.2.2.5.1. Four-pulse TMS study	77
2.2.2.5.2. Chronometric TMS study	77
2.2.3. Results	78
2.2.3.1. Within-subject analyses	78
2.2.3.1.1. OPA experiment	78
2.2.3.1.1.1. Accuracy	78
2.2.3.1.1.2. Reaction times	79
2.2.3.1.1.3. LISAS	80
2.2.3.1.2. LOC experiment	81
2.2.3.1.2.1. Accuracy	81
2.2.3.1.2.2. Reaction times	83
2.2.3.1.2.3. LISAS	83
2.2.3.3. Between-subject analysis	84
2.2.3.3.1. Accuracy	84
2.2.3.3.2. Reaction times	86
2.2.3.3.3. LISAS	87
2.2.4. Discussion	90
2.3. Study 3: object-based facilitation of scene recognition in humans and a computer model of human vision	94
2.3.1. Research questions and hypotheses	94
2.3.2. Methods	95
2.3.2.1. Participants	95

2.3.2.2. Stimuli	95
2.3.2.3. Procedure	96
2.3.2.4. Data analysis	97
2.3.3. Results	98
2.3.3.1. Human performance	98
2.3.3.2. Places365-GoogLeNet network performance	99
2.3.3.3. Comparison between human and Places365-GoogLeNet performance	100
2.3.4. Discussion	101
3. General discussion	103
3.1. Summary of the discussed studies and their outcomes	104
3.2. Implications of the present studies	105
3.3. Limitations and future directions	107
4. Summary and conclusions	109
5. References	111
6. Appendix	128
A. Pre-registered analyses and results of the four-pulse TMS study (#153165 AsPredicted)	128
Participants	128
Analysis	128
Results	129
Accuracy	129
Reaction Times	130
LISAS	131
B. Indoor/outdoor classification of Place365-GoogLeNet Network responses	134
7. Publications of the PhD candidate	141

Abstract

Perception of the visual world is strikingly fast and seemingly effortless. How such efficiency is neurally and computationally implemented remains a long-standing question. Neuroscientific research has identified partially segregated scene- and object-selective pathways in the human visual system, yet it is still unclear how and when these pathways interact. This thesis investigated their interaction in human vision, drawing on existing theoretical accounts. As a complementary benchmark, a computer model of human vision was evaluated to assess the extent to which the observed effects can be reproduced in a feedforward artificial architecture. In all studies, naturalistic photographs served as proxies for real-world input.

Study 1 tested whether one of the representations – scene context or objects – exhibits a temporal processing advantage and in what sequence these elements influence one another. Participants classified scenes and objects in two tasks: a go/no-go task and a two-alternative forced-choice (2AFC) task. Reaction times were analysed alongside the integrated speed–accuracy measure (Balanced Integration Score) to control for speed–accuracy trade-offs. After controlling for these trade-offs, no reliable temporal advantage was observed for either representation. The influences were mutual, with incongruent objects slowing scene context recognition and incongruent context slowing object recognition.

Study 2 examined whether object representations play a causal role in disambiguating scenes. Using chronometric TMS, the study assessed whether the object-selective lateral occipital complex (LOC) makes a causal, time-specific contribution to the categorization of ambiguous scenes, in parallel to the established role of the scene-selective occipital place area (OPA) in the classification of ambiguous objects. It was further assessed whether the effective temporal window of the LOC coincides with the one previously reported for the OPA. Accuracy, reaction times, and integrated speed–accuracy measure (LISAS) were compared across three stimulation windows. Results support a causal contribution of the LOC to the disambiguation of scenes, although its precise timing could not be conclusively established.

Study 3 investigated whether object-facilitated scene recognition requires a coherent scene layout or can still occur when only low-level scene statistics are preserved. It further examined whether this dependency is uniquely human or can also be instantiated within a feedforward artificial architecture. Human participants classified scenes with objects placed

on ambiguous scenes, phase-scrambled scenes, and neutral backgrounds. Performance showed that a coherent layout is critical for object-based facilitation. The same images were classified by Places365-GoogLeNet, and model performance was not significantly different from human behaviour under these manipulations.

Taken together, the results provide little support for strictly hierarchical (“object-first” or “scene-first”) accounts of real-world scene perception – there is no temporal processing advantage for either scene or object representation, and their influences are mutual. Critically, it is shown for the first time that object representations in the LOC make a causal contribution to the recognition of ambiguous scenes. Pinpointing the precise temporal window of this contribution will be essential for evaluating whether scene–object interactions are implemented via a shared bidirectional predictive processing mechanism. It is further demonstrated that object-based facilitation of scene recognition depends on a coherent scene layout, and that classification accuracy across conditions does not differ significantly between human observers and the artificial neural network. An important task for future work will therefore be to specify the conditions under which human and model processing diverge, and to identify the computational mechanisms that give rise to these differences.

Streszczenie

Percepcja otaczającego nas świata jest zdumiewająco szybka i bezwysiłkowa. Neuronalne i obliczeniowe podstawy tej niezwykłej sprawności wciąż pozostają jednak nieznane. Badania neurobiologiczne zidentyfikowały w ludzkim systemie wzrokowym częściowo odrębne ścieżki odpowiedzialne za rozpoznawanie scen i obiektów, ale nadal nie wiadomo, jak i na jakich etapach przetwarzania szlaki te współdziałają. Niniejsza praca doktorska bada interakcje między nimi, opierając się przewidywaniach ugruntowanych modeli teoretycznych oraz wykorzystując komputerowy model widzenia, by ocenić w jakim stopniu zaobserwowane efekty może odtworzyć prostsza architektura dół–góra. We wszystkich badaniach naturalistyczne fotografie posłużyły jako substytut rzeczywistego świata wizualnego.

Pierwsze badanie testowało który element, kontekst sceny (tło) czy obiekt, jest przetwarzany szybciej i w jakiej kolejności oddziałują one na siebie. Uczestnicy klasyfikowali sceny i obiekty w dwóch zadaniach: go/no-go oraz wymuszonego wyboru (two-alternative forced-choice, 2AFC). Czasy reakcji analizowano równolegle ze zintegrowaną miarą łączącą szybkość i dokładność (Balanced Integration Score), aby kontrolować kompromisy między tymi wymiarami. Wyniki nie wykazały istotnej przewagi czasowej dla żadnej z reprezentacji. Wpływy były wzajemne: niezgodne obiekty spowalniały rozpoznawanie sceny, a niezgodny kontekst spowalniał rozpoznawanie obiektów.

Drugie badanie sprawdzało, czy reprezentacje obiektów mogą przyczyniać się do rozpoznawania niejednoznacznych scen. W tym celu zastosowano chronometryczną przezczaszkową stymulację magnetyczną (transcranial magnetic stimulation, TMS), aby określić, czy obszar bocznej części płata potylicznego (lateral occipital complex, LOC), wyspecjalizowany w przetwarzaniu obiektów, wywiera istotny i czasowo specyficzny wpływ na kategoryzację takich scen – analogiczny do udokumentowanej roli obszaru OPA (occipital place area) w identyfikacji niejednoznacznych obiektów. Wyniki w postaci trafności, czasu reakcji oraz zintegrowanej miary łączącej szybkość i dokładność (LISAS) porównano dla trzech okien czasowych stymulacji. Uzyskane dane potwierdzają, że obszar LOC ma istotny, przyczynowy udział w kategoryzacji niejednoznacznych scen, choć nie udało się ustalić dokładnego momentu, w którym ten wpływ zachodzi.

Trzecie badanie miało na celu ustalenie, czy rozpoznanie sceny na podstawie obiektu zależy od jej spójnej struktury, czy może zachodzić nawet wtedy, gdy zachowane są jedynie

jej niskopoziomowe statystyki. Dodatkowo sprawdzono, czy zaobserwowany wynik jest specyficzny dla ludzi, czy też można go odtworzyć w sztucznej sieci neuronowej. Uczestnicy klasyfikowali obiekty umieszczone na trzech rodzajach tła: niejednoznacznym, zdegradowanym i neutralnym. Wyniki wykazały, że spójna struktura sceny jest kluczowa dla rozpoznania sceny opartego na obiekcie. Skuteczność klasyfikacji tych samych bodźców przez model Places365-GoogLeNet nie różniła się istotnie od zachowania ludzi.

Podsumowując, uzyskane wyniki nie potwierdzają założeń ścisłych modeli hierarchicznych w odniesieniu do percepcji scen naturalistycznych – żadna z reprezentacji nie ma czasowej przewagi, a ich wpływy są wzajemne. Badania wykazały również, że reprezentacje obiektów w obszarze LOC mają przyczynowy udział w rozpoznawaniu niejednoznacznym scen, choć nie udało się precyzyjnie określić okna czasowego tego efektu. Ustalenie przebiegu czasowego tego wpływu pozostaje kluczowe dla wyjaśnienia czy interakcje scena–obiekt są wspierane przez dwukierunkowy mechanizm predykcyjny. Rozpoznawanie sceny oparte na obiekcie okazało się zależeć od spójnej struktury sceny, a poprawność klasyfikacji w poszczególnych warunkach nie różniła się istotnie między ludźmi a sztuczną siecią neuronową. Kwestią otwartą pozostaje więc w jakich warunkach dochodzi do rozbieżności między przetwarzaniem ludzkim a modelowym i jakie dokładnie procesy obliczeniowe za te różnice odpowiadają.

Abbreviations

ANN - artificial neural network
ALU - arbitrary linear unit
BOLD - blood-oxygen-level dependent
BIS - Balanced Integration Score
CSF - cerebrospinal fluid
DNN - deep neural network
EBA - extrastriate body area
EEG - electroencephalography
ERP - event-related potential
FFA - fusiform face area
fMRI - functional magnetic resonance imaging
IO - *isolated object* condition
IS - *isolated scene* condition
LISAS - Linear Integrated Speed-Accuracy Score
LOC - lateral occipital complex
NO - *scene-with-no-object* condition
MEG - magnetoencephalography
MS - magnetic stimulation
MSO - maximum stimulator output
OO - *only-object* condition
OPA - occipital place area
OS - *object-with-scene* condition
PC - proportion of correct responses
pFs - posterior fusiform gyrus
PPA - parahippocampal place area
PT - phosphene threshold
RHT - Reverse Hierarchy Theory
RSC - retrosplenial cortex
RSVP - rapid serial visual presentation
RT - reaction time
RM-ANOVA - repeated-measures analysis of variance

SAT - speed–accuracy trade-off
SEM - standard error of the mean
SD - standard deviation
SDT - signal detection theory
TMS - transcranial magnetic stimulation
TOS - transverse occipital sulcus
TUS - transcranial ultrasound stimulation
2AFC - two-alternative forced-choice
WO - *scene-with-object* condition
WOs - scrambled *scene-with-object* condition
V1 - primary visual cortex

1. Introduction

1.1. Naturalistic vision in scientific research

Our visual environment is highly dynamic and cluttered with objects of diverse shapes, colors, and textures. Despite this complexity, perceiving the world feels both natural and instantaneous. For example, when we open the door to an unfamiliar room, we easily determine whether it is a bedroom or a kitchen and readily locate key objects such as a bed or a refrigerator. Another defining characteristic of real-world perception is its remarkable speed, as evidenced by our ability to rapidly switch TV channels or scroll through a news feed on a smartphone. Even when viewing a complex real-world image for only a fraction of a second – a single glance – we can categorize it and identify its main elements, including primary objects and individuals. Although these impressive perceptual abilities have clear adaptive value, facilitating effortless interactions with our environment, the cognitive, computational, and neural mechanisms underlying them remain subjects of interdisciplinary debate.

The question of how the visual system processes the complexity of the natural world has intrigued scientists since the earliest days of vision research. Initially, vision science was grounded in the assumption that real-world perception could be explained based on responses to elementary visual patterns. Consequently, most studies on visual perception employed simple, parametric stimulus sets (Felsen & Dan, 2005; Kayser et al., 2004). Similarly, visual recognition was typically studied using single, isolated objects surrounded by homogeneous space or an array of unrelated items (Biederman, 1972; Bar, 2014). This approach laid the groundwork for our current understanding of the visual system by revealing orientation-tuned receptive fields and retinotopic maps at the neuronal level (review: Hubel & Wiesel, 1979). At the systems level, it further established a hierarchically organized and functionally specialized architecture (review: Gross, 1994). Nevertheless, within this framework, visual processing was investigated using material that markedly differs from the rich, complex input we naturally encounter in our everyday environments.

How visual perception operates in response to more naturalistic, complex, and behaviourally relevant stimuli was not systematically investigated until the 1970s, when the first researchers began to move beyond the established paradigm of vision research. Pioneering this shift, Biederman (1972) demonstrated that the spatial context in which

an object appears significantly influences its perception. Even when images are presented briefly, people are sensitive to meaningful visual organization, recognizing objects more accurately when embedded in coherent, real-world settings than randomly rearranged ones. Further, in a classic rapid serial visual presentation (RSVP) study, Potter (1975) showed that participants could detect target real-world images with high accuracy, even when each was shown for just 1/8th of a second. Crucially, performance accuracy was equal regardless of whether the target was described verbally or shown beforehand, suggesting that brief visual exposure is sufficient to extract the meaning of a scene. Supporting this, research in the 1990s using electroencephalography (EEG) revealed that the brain distinguishes target real-world images (e.g., containing animals or vehicles) from non-targets as early as 150 ms after image onset (Thorpe et al., 1996; VanRullen & Thorpe, 2001). Finally, Li et al. (2002) found that participants could rapidly detect animals or vehicles in briefly presented naturalistic images, even while performing another attentionally demanding task. In contrast, they could not discriminate between simpler stimuli, such as large T's vs. L's or bisected disks vs. their mirror images, under the same conditions.

Collectively, these findings were considered to have “upset the visual applecart” (Braun, 2003), as they suggested that laboratory-based results derived from isolated features and objects might not necessarily generalize to real-world perception (Felsen & Dan, 2005; Kayser et al., 2004). This shift underscored the need to regard naturalistic perception as a separate and significant area of investigation within vision science. The resulting increase in methodological complexity sparked a vivid debate among vision scientists, centering on how to balance ecological validity with experimental control, how to interpret neural responses to richly structured stimuli, and how to relate rich natural inputs to mechanistic models capable of generating interpretable and testable predictions (see: Felsen & Dan, 2005; Rust & Movshon, 2005).

In the early 21st century, research on naturalistic vision has expanded considerably. Foundational studies by Biederman (1972) and Potter (1975), conducted before the introduction of computer-based methods, have since been replicated using modern techniques, consistently demonstrating the importance of context for object recognition (e.g., Kaiser et al., 2020) and the remarkable speed of a real-world perception (Potter, 2012). Further, in line with the findings of Li et al. (2002), the initial processing of real-world information has been shown to remain intact under reduced attention conditions (Groen et al., 2016). Yet, despite this progress, the neural mechanisms and computational processes

enabling rapid perception of richly structured, naturalistic input remain only partially understood.

In parallel to experimental work on human vision, computational efforts to model the visual system have advanced rapidly. The early inspiration for this work stemmed from Hubel and Wiesel (1959, 1962), whose discoveries of orientation-selective receptive fields, local sampling, and hierarchical organization in the cat visual cortex provided a biological blueprint for layered, feature-based architectures. Building on these insights, Rosenblatt's Perceptron (1958) introduced the first trainable artificial neuron. This concept later evolved into modern artificial neural networks (ANNs) through the development of multi-layer architectures and backpropagation (Rumelhart et al., 1986), as well as convolutional models that explicitly instantiate local receptive fields and hierarchical feature integration (Fukushima, 1980; LeCun et al., 1989). While such models, particularly convolutional deep neural networks (DNNs), have become preeminent in visual computational neuroscience (e.g., Yamins & DiCarlo, 2016; Doerig et al., 2023), they still fail to capture several core aspects of human cognition, such as robustness and data-efficient learning. This has prompted serious questions regarding the validity of DNNs as complete models of human visual processing (Bowers et al., 2022; Doerig et al., 2023). Consequently, the field is increasingly focused on pinpointing the specific architectural and computational elements that must be incorporated to bridge this gap between artificial and biological vision.

Below, I first briefly discuss how naturalistic perception has been studied and what constitutes a real-world scene in light of contemporary research (1.2). Secondly, I review state-of-the-art knowledge concerning the neural mechanisms underlying scene processing (1.3), and I introduce key methods that are used to study naturalistic scene perception (1.4). Thirdly, I provide a more detailed explanation of three main frameworks describing the temporal dynamics of scene processing (1.5). Finally, I introduce the aims of the present thesis (1.6).

1.2. Real-world scenes as a subject of scientific investigation

One of the most prevalent methods for investigating naturalistic visual perception, first employed in the seminal studies by Biederman (1972) and Potter (1975), involves examining how participants process photographs representative of real-world environments under different experimental conditions. These images, often termed “scenes”, serve as approximate

models of everyday visual experience. But what defines an image as a real-world scene within the context of visual perception research?

The definition of a real-world scene varies across the literature, ranging from broad to more specific conceptualizations. The most general definitions classify most of our visual experience as a scene, for example: “for most of the time that our eyes are open, the information occupying the visual field comprises a scene” (Foulsham, 2015), or “everywhere we look, a visual scene is in view” (Chun, 2003). More formal definitions emphasize the semantic and structural organization of scenes. Henderson and Hollingworth (1999) define a scene as “a semantically coherent (and often namable) view of a real-world environment comprising background elements and multiple discrete objects arranged in a spatially licensed manner.” Similarly, Foulsham (2015) describes the scene as “a pictorial view of the environment where someone might act,” noting that it consists of “multiple items arranged in a regular and meaningful manner.” Castelhana and Krzyś (2020) further emphasize the structural components of the scene, referring to it as “a view of the natural world, which is made up of space-defining surfaces and smaller objects.”

In line with these definitions, objects constitute a fundamental component of scenes. However, scenes and objects might also be conceptualized as distinct entities. Although the distinction may appear intuitive, precise definitions of what constitutes a scene or an object remain conceptually challenging (Bartnik & Groen, 2023; Peelen et al., 2024). One important difference lies in their visual properties. A defining feature of scenes is their global structure, such as the large-scale arrangement of surfaces, light, and immovable objects that define navigable spaces (Oliva & Torralba, 2001). According to the influential Spatial Envelope model, the dominant spatial structure of scenes can be captured through five perceptual dimensions: naturalness (natural vs. man-made), openness (open vs. enclosed), roughness (smooth vs. textured), expansion (perspective geometry), and ruggedness (depth complexity; Oliva & Torralba, 2001). Objects, in contrast, are defined by local features, such as shape contours, texture gradients, and part boundaries, that rely on high spatial frequencies for recognition (Biederman, 1987). Another key difference between those two is the way humans interact with them: they act on objects but act within scenes (Epstein, 2005). Finally, the distinction between a scene and an object is, in part, determined by spatial scale. Scenes are typically broad, egocentric views of human-scaled environments (Oliva & Torralba, 2001), whereas objects are discrete entities located within reaching distance (Josephs & Konkle, 2020). However, this can be fluid – an object at one scale (e.g., a desk) may become a scene

at another, depending on the observer's perspective and goal (Peelen et al., 2024). While the distinction between scenes and objects can sometimes be debated, it has nonetheless proven valuable for investigating real-world processing, including at the neural level (see Chapter 1.3).

In daily life, however, scenes and objects do not occur in isolation – they co-occur and interact, and these interactions are governed by rules that can be described as “scene grammar” (Vo & Wolfe, 2013; Vo et al., 2019). Just as linguistic grammar dictates word order and meaning, scene grammar defines both the semantic (the congruency between the object and the scene) and syntactic (positional regularities of the object within a scene) rules that structure our environments. Violations of these rules – whether semantic (e.g., a polar bear in a living room) or syntactic (e.g., a chair floating mid-air) – have been shown to slow down and impair scene processing (Vo et al., 2019). Further, the interactions between scene and object representations have also been shown to influence the perceptual level. For instance, objects are perceived with greater clarity when embedded in a congruent scene compared to an incongruent one (Rossel et al., 2022), and ambiguous objects are recognized more readily when presented within their corresponding scene rather than in isolation (Brandman & Peelen, 2017).

In the context of scene processing, it is essential to consider the gist concept. Gist is defined as the rapid extraction of a scene's overall meaning or essence, capturing its fundamental attributes without detailed analysis (Oliva, 2005). In this thesis, the term gist is used interchangeably with scene or scene context (excluding object information); however, it is worth noting that in certain cases, object information can also contribute to forming a scene's gist (Mack & Palmieri, 2010).

1.3. Neural mechanisms of real-world scene perception

How does the brain transform incoming light into a coherent representation of the external world? Visual information is conveyed when light, captured by the retina, is transmitted via the optic nerve to the occipital cortex, a principal region at the posterior part of the brain devoted to visual processing. Neurophysiological studies in animals, pioneered by Hubel and Wiesel (1979), alongside investigations of patients with focal brain lesions (e.g., Bodamer, 1947; Goodale et al., 1991), provided initial evidence for two fundamental principles of visual cortex organization: hierarchical processing and functional specialization (Peelen, 2024).

Hierarchical processing posits that visual perception is achieved through a gradual, stagewise integration of information: from detecting simple features such as edges and orientations to synthesizing complex representations like shapes and objects (Gross, 1994). Functional specialization, in turn, indicates that distinct neural pathways are dedicated to processing specific attributes of the visual scene, enabling the brain to analyze diverse aspects of visual input simultaneously (Grill-Spector & Malach, 2004). The advent of neuroimaging techniques, particularly functional magnetic resonance imaging (fMRI), has further refined our understanding of visual cortex organization by allowing the precise localization of brain regions involved in visual perception. Notably, neuroimaging research has revealed that scene contexts and objects are processed in distinct networks within the visual cortex (Epstein & Baker, 2019; Grill-Spector & Malach, 2004).

The scene-selective network is currently thought to comprise three main brain regions: the parahippocampal place area (PPA), the occipital place area (OPA), and the retrosplenial cortex (RSC). These regions have been delineated by contrasting fMRI activity evoked by scene images with responses elicited by other categories, such as faces, isolated objects, or body parts (Epstein & Baker, 2019).

The PPA, first identified by Epstein and Kanwisher (1998), is localized within the ventral visual stream, encompassing regions around the anterior lingual and collateral sulcus (Aguirre et al., 1998; Ishai et al., 1999; Weiner et al., 2018). This region is critically involved in encoding the spatial layout of scenes and representing navigationally relevant landmarks, such as walls, floors, and buildings (Aguirre & D'Esposito, 1999; Kravitz et al., 2011). The PPA also processes scene identity by analyzing low spatial frequency properties, such as openness and enclosure, to distinguish between different environments (e.g., kitchens versus forests; Walther et al., 2009).

The RSC, situated in the medial parietal cortex near the parieto-occipital sulcus (Epstein, 2008), is crucial in situating local scenes within broader spatial environments. It supports the formation of cognitive maps and the development of viewpoint-independent representations, phenomena that are fundamental for effective spatial navigation. The RSC is critical for spatial memory as it integrates egocentric (viewer-centered) and allocentric (world-centered) spatial frameworks, thereby facilitating the recall of routes and landmarks (Vann et al., 2009). Furthermore, the RSC interfaces with hippocampal networks to enhance memory consolidation and support the long-term retention of spatial information (Aggleton, 2014).

The OPA, originally designated as the transverse occipital sulcus (TOS), is localized on the lateral occipital surface near the parieto-occipital junction (Dilks et al., 2013). The OPA is hypothesized to play a critical role in processing the spatial structure of scenes (Dilks et al., 2011; Persichetti & Dilks, 2018) and delineating environmental boundaries (Julian et al., 2016). Transcranial magnetic stimulation (TMS) of the OPA disrupts multiple facets of scene processing, including scene categorization (Ganaden et al., 2013), spatial layout-based scene discrimination (Dilks et al., 2013), and scene-related expectations (Gandolfo & Downing, 2019).

The selective response of the PPA, RSC, and OPA to scene images, as opposed to other image categories, has been replicated across multiple studies, firmly establishing the existence of a scene-selective network in the human brain (Bartnik & Groen, 2023). Nonetheless, questions remain regarding the precise nature of the visual information represented in these regions. Some accounts propose that the posterior PPA together with the OPA support the rapid visual analysis of scene properties, whereas the anterior PPA and the RSC underpin scene memory and map-based navigation (Baldassano et al., 2016). In contrast, Dilks et al. (2022) argue for a different division of labour: the PPA is exclusively responsible for scene categorization, while the OPA and the RSC support visually guided and map-based navigation, respectively.

Further, the activity in the PPA, OPA, and RSC has been linked to low-level representations of scenes, such as high spatial frequencies, rectilinear features, cardinal orientations (Kauffmann et al., 2015; Nasr et al., 2014), contour junctions (Choo & Walther, 2016), and orthogonal edges (Bonner & Epstein, 2018). This raises questions regarding the extent to which these regions encode low- versus high-level visual properties, and given the inherent correlations between these features in natural scenes, disentangling their independent contributions remains a considerable challenge (Malcolm et al., 2016). While these observations caution against a simplistic view of scene-selective regions as exclusively high-level processors, neural sensitivity to low-level features may underpin the computations required to construct higher-level, abstract representations of scenes (Groen et al., 2017; Rajimehr et al., 2014).

Finally, while the PPA, OPA, and RSC encode spatial properties (e.g., boundary geometry; Kravitz et al., 2011) and scene categories (Walther et al., 2009), they also exhibit object-related coding. The PPA and RSC represent contextual object information like landmarks (Aminoff et al., 2007) and object co-occurrence statistics (Stansbury et al., 2013),

while the OPA responds selectively to scene-defining objects (MacEvoy & Epstein, 2007) or their spatial density (Julian et al., 2016). This object sensitivity persists alongside their responses to low-level features (spatial frequencies, rectilinear contours; Kauffmann et al., 2015), suggesting these regions integrate multiple visual cues to construct scene representations.

Object-selective cortical regions are principally localized within the lateral occipital complex (LOC) and the posterior fusiform gyrus (pFs). These regions demonstrate significantly greater activation in response to intact objects relative to their scrambled counterparts or unstructured textures, implicating their involvement in the encoding of object-specific attributes, including geometric shape and categorical identity (Grill-Spector, 2003). Notably, while certain subregions exhibit domain specificity, such as the fusiform face area (FFA) for facial processing (Kanwisher et al., 1997) and the extrastriate body area (EBA) for the perception of body parts (Downing et al., 2001), others, particularly the LOC, subserve a more generalized role in object recognition (Grill-Spector, 2003). The LOC is functionally robust, responding invariantly to objects defined by disparate low-level visual cues (e.g., luminance gradients, texture boundaries, or motion coherence; Grill-Spector et al., 1998) as well as higher-order representations, such as schematic line drawings (Ishai et al., 2000), illusory contour-defined shapes (e.g., Kanizsa figures; Mendola et al., 1999), and stimuli conveying animacy or intentionality (Gobbini et al., 2011; Wheatley et al., 2007).

Overall, a substantial body of fMRI research indicates that object-selective and scene-selective regions encode distinct aspects of visual scenes. Notably, this separation has been further supported by TMS studies. Firstly, Dilks et al. (2013) demonstrated a double dissociation in isolated recognition tasks by showing that TMS over the scene-selective OPA impaired scene recognition without affecting object recognition, whereas TMS over the object-selective LOC selectively disrupted object recognition. These findings were subsequently confirmed in a preregistered replication by Wischniewski and Peelen (2021a), providing compelling causal evidence that scene and object information are processed via separable neural pathways in the visual cortex.

Nevertheless, despite this apparent neuronal segregation, a wealth of behavioral studies indicates that scene context and object information interact during perception (Bartnik & Groen, 2023; Peelen et al., 2024). Neuroimaging research further shows that responses in the scene-selective cortex are modulated by the objects present within a scene, while activity in the object-selective cortex is influenced by the surrounding scene (Bartnik & Groen, 2023;

Peelen et al., 2024). Consequently, the question arises regarding how and when these interactions are implemented: a topic that continues to be the focus of ongoing investigation.

1.4. Research methods in real-world scene perception

The study of scene perception draws on a range of methodological approaches from experimental psychology, neuroscience, and computer science – each offering unique insights into how the brain processes complex visual environments (see Fig. 1).

Experimental psychology uses a family of paradigms, such as rapid serial visual presentation (RSVP) and speeded categorization, to quantify perceptual performance from accuracy, response times, and error profiles. These measures provide an indirect readout of the perceptual content available to observers and the efficiency with which it is extracted. In this approach, stimulus properties are manipulated under tightly controlled conditions while task demands are held constant, which enables causal inferences about which visual information supports performance. Concretely, one can orthogonally vary the presence of objects and scene context, their spatial arrangement/relative position, their semantic congruency/meaning, and the reliability of each source of evidence (e.g., by degradation or scrambling).

To characterise the temporal unfolding of perception in the brain, time-resolved techniques, including electroencephalography (EEG) and magnetoencephalography (MEG), are used. EEG, developed in the 1920s by Hans Berger, captures scalp-recorded electrical activity with millisecond precision. Event-related potential (ERP) analyses allow researchers to isolate distinct stages of perceptual and cognitive processing evoked by scene stimuli (Luck, 2014). MEG, which records the magnetic fields associated with neural currents, offers comparable temporal resolution with improved spatial localization of cortical sources (Hämäläinen et al., 1993). Using these methods, it is possible to trace the evolution of stimulus perception in the time domain.

To determine the localised representation of perceptual content in the brain, researchers typically rely on functional magnetic resonance imaging (fMRI). Introduced in the early 1990s (Ogawa et al., 1990), fMRI measures blood-oxygen-level-dependent (BOLD) signals to indirectly infer changes in regional neural activity. While temporally coarse, fMRI offers high spatial resolution, enabling more precise, compared to EEG and MEG, localization of functionally specialized regions involved in scene analysis.

To infer causal relationships, researchers employ transcranial magnetic stimulation (TMS). This technique delivers brief magnetic pulses that can transiently disrupt neural processing in targeted cortical regions (Barker et al., 1985). Therefore, TMS allows for assessing whether a given region is necessary for a specific perceptual function (Walsh & Cowey, 2000). Moreover, when applied at varying delays relative to stimulus onset, chronometric TMS can also reveal *when* a region is functionally involved in a task, bridging the gap between spatial and temporal dynamics. However, TMS is limited to superficial cortical areas accessible from the scalp.

While neuroscientific methods help to establish when and where various scene properties are processed in the brain, they do not explain *how* this information is computed. To address this question, researchers increasingly rely on computational modelling within the field of computer vision. Among the most powerful tools to emerge in this field are convolutional deep neural networks (DNNs): computational architectures inspired by the hierarchical organization of the visual cortex (Kriegeskorte, 2015). These networks process visual input by applying a series of convolutional filters that detect local patterns such as edges, orientations, and textures. Through successive layers, these features are combined into increasingly complex and abstract representations, enabling the network to recognize complete objects and scenes (Kriegeskorte, 2015; LeCun et al., 2015). Trained on large-scale labelled datasets, these models have achieved – and in some aspects even surpassed – human-level accuracy in visual recognition (He et al., 2015, Krizhevsky et al., 2012, Russakovsky et al., 2014). Furthermore, subsequent studies have demonstrated a strong correspondence between the representational structures of these models and those found in the human visual cortex (e.g., Cichy et al., 2016; Güçlü & van Gerven, 2015; Khaligh-Razavi & Kriegeskorte, 2014).

Despite major progress, current models still differ from biological vision in both architecture and performance (review: Bowers et al., 2022; Doerig et al., 2023). Given these differences, DNNs cannot yet be treated as direct models of how the brain processes visual information. Instead, they can be viewed as a useful tool for formulating and testing falsifiable theories about brain computation, as proposed within the emerging neuroconnectionism programme (Doerig et al., 2023). This framework enables the systematic testing of hypotheses about how the visual system transforms sensory input into meaningful percepts, through comparisons between human and model performance, alignment of neural and network representations, and targeted computational simulations.

For real-world scene perception, the key open question is which aspects of scene–object interactions depend on recurrent, two-stream processing and which are achievable within a one-stream, feedforward model.

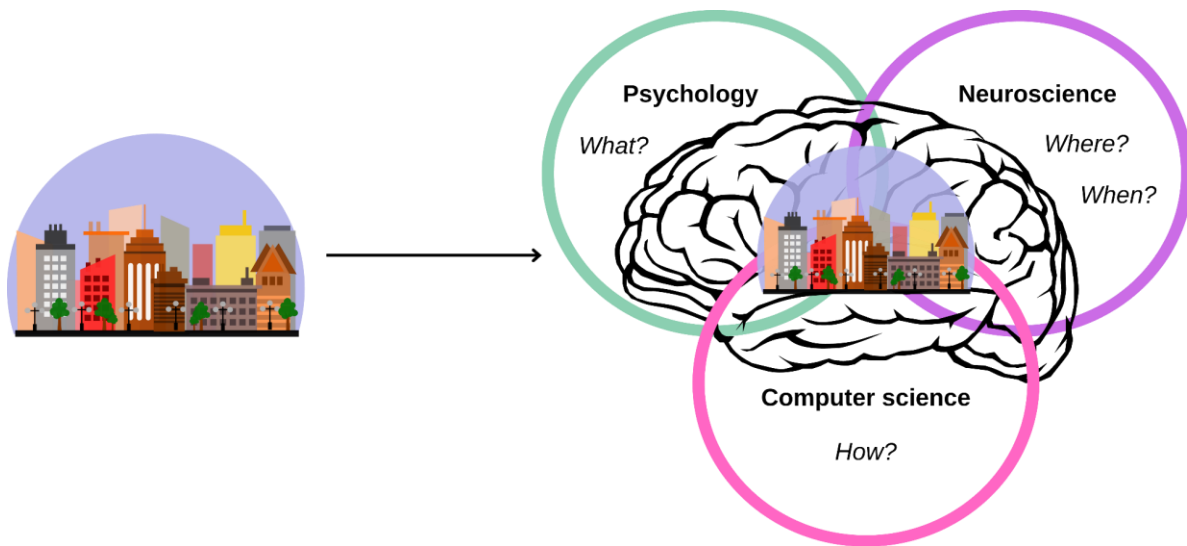


Figure 1. Research on real-world scene perception utilizes methodologies from experimental psychology, neuroscience, and computer science, each offering unique but complementary insights into the mechanisms underlying naturalistic perception.

1.5. Temporal dynamics of real-world scene perception

Real-world scene perception, particularly its temporal unfolding, has been theorized within three principal frameworks (analytic, holistic, and predictive processing), which differ in their assumptions regarding the sequence of processing stages and the interplay between scene and object representations. This chapter will shortly discuss the three frameworks, presenting their concepts and predictions arising from them. These predictions will later form the basis of research questions and hypotheses tested in the scope of this thesis.

1.5.1. Analytic view

The analytic view of visual perception is rooted in the tradition of associationism (Wundt, 1874, Titchener, 1902), and posits that perception results from the progressive integration of elementary features into increasingly complex structures until a complete

recognition (Gibson, 1966, Riesenhuber & Poggio, 1999; DiCarlo et al., 2012; see Fig. 2). This atomistic conceptualization found strong neurobiological support in the early neurophysiological investigations into the organization of visual receptive fields (Hubel & Wiesel, 1962).

Within the domain of real-world scene perception, the analytic account assumes that objects are building blocks of scenes. In fact, within this framework, scene contexts are thought to be formed by the sets of objects that co-occur together (Biederman et al., 1982; Bar & Ullman, 1996; Liu et al., 2009; MacEvoy & Epstein, 2011; review: Bartnik & Groen, 2023). Consequently, the analytic account predicts that object recognition should precede the interpretation of broader scene context. Initial behavioral and electrophysiological evidence supports this view. The EEG studies have shown that object-related neural responses can be detected remarkably early, with differential activity between target and non-target objects emerging as soon as 150 ms post-stimulus (Thorpe et al., 1996; VanRullen & Thorpe, 2001). Complementary behavioral findings indicated that objects can be accurately categorized within 250–400 ms following brief visual presentations (Fabre-Thorpe et al., 2001; Rousset et al., 2002, 2003; Joubert et al., 2008). Importantly, when scenes contained semantically incongruent objects, recognition of the scene context was impaired – participants exhibited slower and less accurate categorization of the background scene when it included an atypical object (Joubert et al., 2007; Mack & Palmeri, 2010). These findings might suggest that objects are detected early enough to shape the ongoing construction of scene meaning, consistent with the predictions of the analytic account.

1.5.2. Holistic view

The holistic view of visual perception, rooted in Gestalt psychology (Koffka, 1922; Wagemans et al., 2012), emphasizes the primacy of global structure in shaping perceptual experience. This perspective posits that the visual system tends to prioritize the organization of stimuli into coherent wholes before processing their constituent parts. A seminal extension of this idea was offered by Navon (1977), who demonstrated through compound letter stimuli – large letters composed of smaller ones – that observers typically process global features (large letter) before local ones (small letters), a phenomenon now known as the global precedence effect.

This global-before-local tendency was further formalized in the Reverse Hierarchy Theory (RHT), which distinguished between “vision at a glance” and “vision with scrutiny” (Hochstein & Ahissar, 2002). According to the RHT, initial perceptual impressions originate in the high-level visual areas that encode coarse, global representations, while subsequent feedback enables access to the lower-level areas for fine-grained analysis (see Fig.2). Importantly, neurophysiological research has shown that extensive feedback connectivity provides the neural infrastructure required for global-to-local processing (Lamme & Roelfsema, 2000; Bullier, 2001).

Empirical support for the temporal primacy of global information came first from studies using artificial stimuli. While early findings (e.g., Navon, 1977; review: Kimchi, 1992) primarily demonstrated perceptual dominance of global features, more recent work has additionally provided evidence for their temporal precedence. In particular, Campana et al. (2016) showed that global-level representations can emerge earlier in the visual processing stream than local ones, even in controlled laboratory conditions with abstract stimuli. These findings suggest that global precedence is not merely a result of decisional or attentional biases but may reflect intrinsic temporal dynamics of the visual system.

In the context of real-world scene perception, the holistic framework predicts that scene context is processed rapidly and exerts top-down influence on object recognition (Bar, 2004; Bar et al., 2006; Campana and Tallon-Baudry, 2013; Oliva & Torralba, 2006). Scenes typically span the entire visual field and convey low spatial frequency information, which is efficiently transmitted via the magnocellular pathway. This pathway allows for the rapid extraction of scene gist, a coarse, abstract representation of the overall layout and meaning of a scene, which can subsequently bias and facilitate object recognition in the slower, higher-resolution parvocellular pathway along the ventral stream.

Numerous behavioral and electrophysiological studies support this framework in the context of scene processing. First of all, humans can categorize the gist of a scene extremely quickly, with a differential ERP wave emerging within 150–200 ms of stimulus onset (Goffaux et al., 2005) and mean reaction times to target scenes around 400 ms (Joubert et al., 2007). Further, numerous studies are showing that when objects are embedded within semantically incongruent scenes, object recognition is impaired (e.g., Boyce et al., 1989; Davenport & Potter, 2004; Davenport, 2007; Furtak et al., 2022; Leroy et al., 2020) and delayed (Fize et al., 2011; Joubert et al., 2008; Remy et al., 2013, 2014, 2020) relative to objects placed in congruent scenes. These findings might suggest that the initial, global

interpretation of the scene biases and constrains the subsequent perception of its local elements.

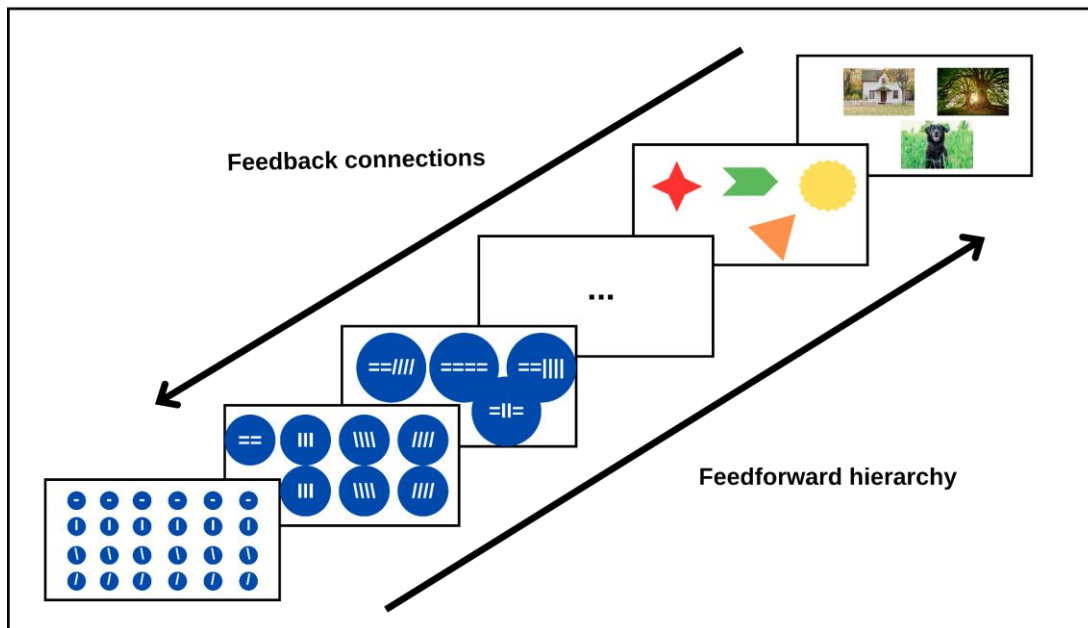


Figure 2. Schematic depiction of analytic and holistic accounts of visual perception. Both frameworks conceptualize perception as hierarchically organized, yet they diverge in the processing sequence: the analytic account posits that local elements are processed first, whereas the holistic perspective suggests that global structure is accessed early and constrains the interpretation of local detail. Adapted from Hochstein & Ahissar (2002).

1.5.3. Predictive processing view

Predictive coding is a neurocomputational framework that conceptualizes perception as an active inferential process, in which the brain continuously generates top-down predictions about sensory input based on prior knowledge and internal models (Clark, 2013; Peelen et al., 2024). These predictions are continuously evaluated against incoming sensory input. When a mismatch occurs between what is expected and what is sensed, the resulting prediction error is used to adjust the brain's internal model, refining perception to better match the external world. At the neurobiological level, such an architecture is thought to be implemented through feedback connections that convey predictive signals from higher-

to lower-order areas and feedforward pathways that transmit residual errors (i.e. the remaining, unexplained parts of the sensory input) allowing perceptual hypotheses to be dynamically updated based on the reliability of sensory information (Friston, 2005, Keller & Mrcic-Flogel, 2018; Rao & Ballard, 1999, review: Peelen et al., 2024).

The predictive processing account of the real-world scene perception proposes that scene and object information are initially processed through anatomically distinct pathways that operate in parallel. In both systems, processing is hierarchical, progressing from low-level sensory to high-level semantic analysis. Such a hierarchical structure is considered essential for implementing predictive coding principles, with reciprocal connections enabling continuous bottom-up and top-down modulation (Friston, 2005; Rao & Ballard, 1999). This parallel processing perspective aligns with experimental findings showing a similar timeline – around 200 ms after stimulus onset – for the emergence of scene and object representations (Cichy et al., 2014, 2017; Henriksson et al., 2019; Isik et al., 2014; Kaiser et al., 2016).

Despite being initially processed along distinct pathways, object and scene representations can still engage in mutual, facilitatory interactions. According to the predictive processing view, these interactions can be conceptualized as a form of non-hierarchical Bayesian inference (Doya et al., 2006; Ma et al., 2012), and their direction and magnitude depend on the observer's goal and the reliability of visual information (Peelen et al., 2024). Thus, in contrast to classical hierarchical accounts that assume a fixed processing order ("object-first" or "scene-first"), the predictive processing approach emphasizes an adaptive architecture in which scene and object representations dynamically influence each other depending on the current sensory context. Importantly, in line with the predictive processing approach, the interactions between scene and objects' representations should be observed mainly when either scene or object information is ambiguous. When both sources of information are highly reliable, the scene-to-object and object-to-scene influences should be reduced, and if such effects are still observed (e.g., slower or less accurate reactions to incongruent objects), they should be interpreted as a result of post-perceptual processes (Peelen et al., 2024).

Numerous studies provide evidence in favour of this account. Firstly, behavioral studies demonstrate that coherent scene contexts facilitate the recognition of degraded or ambiguous objects (Brandman & Peelen, 2017), and conversely, that reliable object cues improve the interpretation of visually degraded or semantically ambiguous scenes (Brandman & Peelen, 2019). Second, neuroimaging studies reveal that such mutual facilitation reflects

sharpened neural representations: ambiguous objects elicit stronger and more discriminable responses in object-selective cortex when embedded in coherent scenes (Brandman & Peelen, 2017), while diagnostic objects enhance scene-selective activity when the scene itself is degraded (Brandman & Peelen, 2019). Furthermore, MEG decoding studies show that both scene-to-object and object-to-scene disambiguation unfold with similar temporal dynamics, peaking around 300 ms post-stimulus onset (Brandman & Peelen, 2017, 2023). Finally, causal evidence from TMS indicates that the scene-selective occipital place area (OPA) is necessary for object disambiguation between 160–200 ms after stimulus onset (Wischnewski & Peelen, 2021b).

1.6. Aims of the thesis

The overarching aim of this thesis was to clarify how the representations of scene contexts and objects interact in human vision, drawing on existing theoretical accounts. As a targeted complement, the third study benchmarked these interactions against a feedforward computer model of human vision.

Study 1. Testing the hierarchical models of scene–object interactions

The main research questions were:

- 1) Does one type of representation – scene context or object – exhibit a temporal processing advantage?
- 2) What is the sequence in which scene context and object information influence each other?

Study 2. Testing the causal role of object representations in disambiguating scenes

The main research questions were:

- 1) Does the object information represented in the object-selective LOC play a causal role in disambiguating scenes?
- 2) Is the involvement of the object-selective LOC in scene disambiguation time-specific, and does its effective temporal window align with the previously established timing of scene-selective OPA involvement in object disambiguation (Brandman & Peelen, 2017)?

These questions were addressed using chronometric TMS, which allows for causal and time-resolved interference with specific cortical regions.

Study 3. Object-based facilitation of scene recognition in humans and a computer model of human vision

The main research questions were:

- 1) Does object-based facilitation of scene recognition depend on a coherent scene layout, or can it still occur when only low-level scene statistics are preserved?
- 2) Does a deep neural network (Places365-GoogLeNet) exhibit sensitivity to contextual structure comparable to that of human observers?

2. Experimental studies: scene–object interactions in naturalistic vision

2.1. Study 1: testing the hierarchical models of scene–object interactions

2.1.1. Research question and hypotheses

Scene–object interactions in real-world vision have predominantly been explained using two divergent theoretical models: analytic and holistic. Within both these frameworks, scenes are viewed as hierarchical structures, with the scene *gist* (the general meaning of a scene) representing the global level and single objects representing the local level. In line with the analytic approach, local objects and details are perceived first, and the global meaning of a scene is created based on their subsequent integration (Hubel & Wiesel, 1962; Gibson, 1966; Riesenhuber & Poggio, 1999; Liu et al., 2009; MacEvoy & Epstein, 2011; Di Carlo et al., 2012). In contrast, the holistic approach assumes the temporal precedence of the global scene level, which is recognized early and automatically, and influences the subsequent recognition of local parts (Navon, 1977; Kimchi, 1992; Hochstein & Ahissar, 2002; Bar et al., 2006; Oliva and Torralba, 2006; Campana & Tallon-Baudry, 2013). These accounts are therefore often described as “object-first” and “scene-first,” respectively.

Interestingly, there is evidence in support of both approaches. Firstly, parallel lines of research indicated that the categorization speed of both scenes (Goffaux et al., 2005; Joubert et al., 2007; Rousselet et al., 2005) and objects (Fabre-Thorpe et al., 2001; Rousselet et al., 2002; Rousselet et al., 2003; Thorpe et al., 1996; VanRullen & Thorpe, 2001) is extremely fast, even when images were displayed for no longer than 30 ms. Specifically, the differential neural activity between target and distractor images was observed as early as around 150 ms post-stimulus, with minimal and mean reaction times to targets reported at around 250 ms and 400 ms post-stimulus onset, respectively. Further, numerous studies indicated that objects semantically incongruent with the scene gist, namely those with a very low probability of occurring in a given context (Biederman et al., 1982), were recognized more slowly than congruent ones (Fize et al., 2011; Joubert et al., 2008; Remy et al., 2013, 2014, 2020). However, although most studies have focused on the effects of context on object

recognition, there is also some evidence indicating that scenes containing semantically incongruent objects are recognized more slowly than those with congruent ones (Joubert et al., 2007; Mack & Palmieri, 2010).

Thus, despite fundamental differences in analytic and holistic predictions concerning the temporal unfolding of visual perception, it remains unclear which, if any, of these frameworks offers a good model for describing the scene recognition process. It is worth noting, however, that previous research has one important limitation: it only investigated one level, either global (scene) or local (objects), without controlling how the other contributed to the categorization speed. As a result, direct comparisons between the speed of scene and object processing, crucial to uncovering their potential hierarchical relationship, could not be conducted.

To enable such direct comparisons, the study should meet several conditions. First, the recognition of both levels should be assessed using the same task within the same participant group. Further, the set of stimuli should manipulate global and local levels in a fully factorial design, ensuring that the two dimensions are orthogonal and not predictive of each other. Finally, recognition should be evaluated at only one of the two levels at a time, so that the other remains task-irrelevant. This approach ensures that performance limits for each level are accurately assessed and minimizes the risk of participants adopting biases or strategies to attend primarily to one of the levels (at the cost of the other).

To my knowledge, only a single previous study using hierarchical artificial stimuli met the above criteria (Campana et al., 2016). Its results supported the holistic view by demonstrating that global properties of stimuli were systematically reported faster than the local ones and that global information biased the reports of local properties despite being task-irrelevant (Campana et al., 2016). It is yet to be established whether a similar global-to-local pattern can be observed in the context of real-world scene perception.

The present study was designed to address this question. It consisted of two experiments: a go/no-go task and a two-forced-choice alternative (2AFC) task. Participants were presented with briefly displayed images of real-world scenes in each task. A set of stimuli in which all scenes depicted either a natural or man-made context and a single natural (animal) or man-made (furniture) foreground object, combined in congruent or incongruent ways (thus not predictive of each other), was used. Images were displayed for 67 ms, as previous studies showed that performance in both tasks plateaus at this display time, reaching a d' value > 3 (Bacon-Mace et al., 2005; Furtak et al., 2022). In the go/no-go task, participants were asked

to react by pressing a button as fast as possible to either scene contexts (natural or man-made) or objects (natural or man-made) in separate blocks. In the 2AFC task, they performed a speeded classification of either scene contexts or objects (in separate blocks) as natural or man-made.

Based on these experiments, two main predictions of hierarchical views on visual perception were tested. First, it was examined whether one level – either global gist (scene context) or local object – is recognized faster. Second, it was investigated whether a mismatch between scene context and object categories (i.e., semantic incongruency) slows target classification and, if so, whether this influence is unidirectional, either local-to-global or global-to-local. Additionally, the time course of interference from semantic incongruency was assessed, focusing on whether this effect is already present in the fastest responses or develops over time. Finally, given evidence for the special status of animals in visual perception (Crouzet et al., 2012), the target category (natural vs. man-made) was included as a factor to assess its role in classification speed.

Mean reaction times (RTs) were used as the primary measure of performance, reflecting processing speed. However, because the tasks involved potential speed–accuracy trade-offs, Balanced Integration Scores (BIS) were additionally computed and served as the main index for interpreting the behavioral effects.

2.1.2. Methods

2.1.2.1. Participants

Thirty-six healthy participants aged 18 to 40 were invited to participate in the study. All experimental procedures were approved by the Research Ethics Committee at Nicolaus Copernicus University (24/2022). To participate, individuals were required to have normal or corrected-to-normal vision, no history or current diagnosis of mental or neurological disorders, and no current use of psychoactive substances, including medications. All participants provided written consent and received monetary compensation for their time (150 PLN = c.a. 32 EUR).

For Experiment 1, data from 29 participants (16 women, mean age = 23.8 years; SD = 4.0; range: 19–37 years, 5 left-handed) were included in the behavioral analysis. Data of 7 additional participants were collected but excluded due to a lack of compliance with

the task instructions (2) or their performance in one or more stimulus conditions below 3 SDs from the overall mean in the task (5).

For Experiment 2, data from 31 participants (15 women, mean age = 23.9 years; SD = 3.99; range: 19–37 years, 6 left-handed) were included in the behavioral analysis. Data of 5 additional participants were collected but excluded due to the lack of compliance with the task instructions (2) or their performance in one or more stimulus conditions below 3 SDs from the overall mean in the task (3).

As in the previous study (Furtak et al., 2022), the aim was to collect at least 30 valid datasets for both experiments. A sensitivity analysis (Campbell & Thompson, 2012) indicated that the smallest possible effect size (η_p^2) that can be detected with these sample sizes (33 and 31), assuming alpha level .05 and power .80, is equal to .207 and 0.218, respectively.

2.1.2.2. Stimuli

A stimulus set developed by Remy et al. (2013, 2020) was used. The original set included 400 real-life color images, each depicting a single foreground object (an animal or a piece of furniture) placed in a context (outdoor natural, or indoor man-made). The 400 stimuli were organized into 100 sets, each containing 4 stimuli. These sets were created by alternately placing 2 foreground objects into 2 different contexts, which resulted in 2 congruent object-context pairings (an animal in a natural setting and a piece of furniture in a man-made setting) and 2 incongruent object-context pairings (an animal in a man-made setting and a piece of furniture in a natural setting). Within each set, the stimuli were matched for contrast and luminance, with the objects having the same pixel size and positioned identically across both contexts. Across all 400 stimuli, the average object size was relatively large, covering $12.7 \pm 4.7\%$ of the image.

Given that the present study, as opposed to Remy's experiments (2013, 2020), included context categorization tasks, twenty images in which the context included both natural and man-made elements (e.g., a balcony with a view of the sky; a room with a window depicting trees) were excluded, in all of their versions. Further, the initial analysis of physical features was conducted, which indicated differences in the luminance ($F(3,316) = 2.839$, $p = .038$, $\eta^2 = .026$) and the red-green spectrum ($F(3,316) = 13.196$, $p < .001$, $\eta^2 = .111$) between the four categories of stimuli (natural congruent/incongruent and man-made congruent/incongruent). The differences in the former were addressed by excluding five man-

made background images with the lowest luminance and five natural background images with the highest luminance, in all their versions (no significant differences in luminance were observed after exclusion: $F(3,276) = 0.238$, $p = .870$, $\eta^2 = .003$). To address the latter, the images were additionally converted to grayscale. The luminance and contrast of the black and white stimuli were equated using the `histMatch` function from the SHINE toolbox (Willenbockel et al., 2010) in Matlab (R2022b). After this operation, all images had a luminance of 128.2 and a contrast of 55.4 ALU (Arbitrary Linear Units) as estimated by the SHINE toolbox. The final stimulus set consisted of 280 fully counterbalanced grayscale images.

An example of a set of 4 stimuli used in the study is shown in Fig. 3A.

2.1.2.3. Apparatus

The experimental procedure was written in the Presentation software (Neurobehavioral Systems, Albany, CA, USA) and presented on an LCD monitor MSI Optix MAG251RX (25") with 1920*1080 resolution and 240 Hz refresh rate. Participants were seated comfortably in a dimly lit room with a viewing distance of 60 cm, maintained by an adjustable chinrest.

2.1.2.4. Procedure

The trial sequence for both experiments is shown in Fig. 3B.

2.1.2.4.1. Experiment 1 – procedure

Experiment 1 employed a classic go/no-go paradigm. Participants performed context and object categorization tasks in separate blocks. During each task, they were asked to press a response pad key as fast as possible when they saw a natural or a man-made context (scene context categorization task) or object (object categorization task). Thus, overall, the experiment consisted of 4 blocks (2 tasks x 2 target categories). The order of blocks was randomized. Each block contained half of the target trials (go trials) and half of the distractor trials (no-go trials). Additionally, in each trial subset (go and no-go), half of the stimuli were congruent (i.e., context and object of the same category) and the other half were incongruent (i.e., context and object of mismatching categories).

The procedure consisted of 1120 trials in total (560 per task). In each natural and man-made block, all 280 stimuli were presented, resulting in the full stimulus set being shown twice per task. The procedure took approximately 40 minutes to complete. The order of trials was random. Each trial began with a white fixation cross (subtending $0.9^\circ \times 0.9^\circ$ of visual angle) displayed at the center of a black screen for a randomly chosen duration between 600 ms and 900 ms. Then, the scene image was presented centrally for 67 ms. The image presentation was followed by a blank screen (1000 ms). Scene images subtended $20.8^\circ \times 13.8^\circ$ of visual angle. Participants were asked to respond as fast and accurately as possible and were given maximally 1000 ms for a response. The next trial sequence always began when the maximal time elapsed, regardless of whether a response was provided. Participants were given 4 self-paced breaks during each block.

2.1.2.4.2. Experiment 2 – procedure

Experiment 2 employed a two-alternative forced-choice (2AFC) categorization task with 2 blocks. Participants were asked to classify scene contexts or objects (depending on the block) as natural or man-made. Responses were made by pressing one of two buttons using the index finger of their left or right hand. Button assignment was fixed within the experimental session but counterbalanced across participants. Each block contained half of the natural targets and half of the man-made ones. Additionally, half of the stimuli in both natural and man-made subsets were congruent (i.e., context and object of the same category), and the other half were incongruent (i.e., context and object of mismatching categories). The order of blocks (context/object classification) was counterbalanced across participants.

The procedure consisted of 560 trials (280 per task, with the full stimulus set presented twice) and lasted approximately 25 minutes. The order of trials was random. Each trial began with a white fixation cross (subtending $0.9^\circ \times 0.9^\circ$ of visual angle) displayed at the center of a black screen for a randomly chosen duration between 1100 ms and 1500 ms. The change in pre-stimulus time between experiments was introduced due to additional subsequent analysis planned for Experiment 2 (not described in the present thesis). As in Experiment 1, scene images subtended $20.8^\circ \times 13.8^\circ$ of visual angle and were presented centrally for 67 ms. The stimulus presentation was followed by a blank screen (1000 ms). Participants were asked to respond as fast and accurately as possible and were given maximally 1000 ms for a response. The next trial sequence always began when the maximal time elapsed, regardless

of whether a response was provided. Participants were given 4 self-paced breaks during each block.

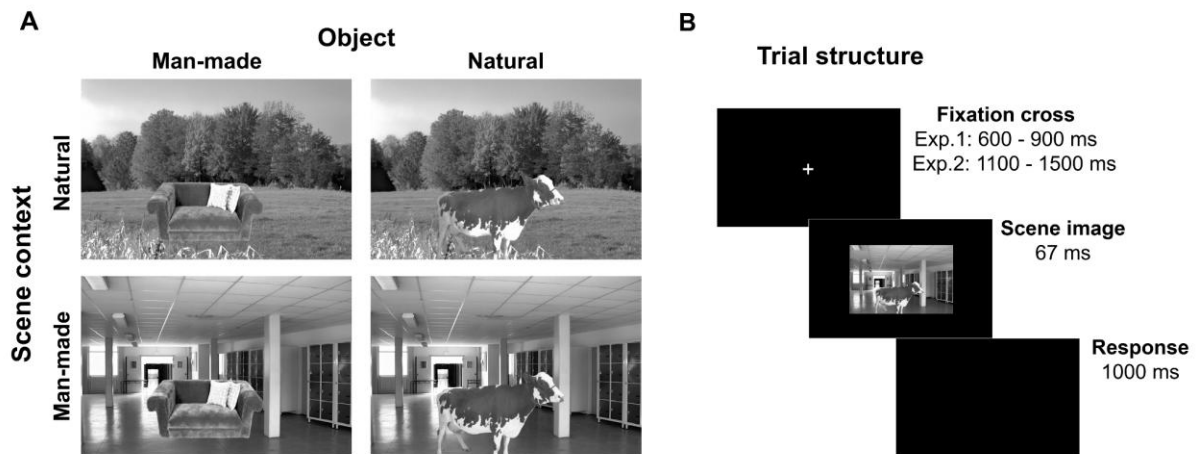


Figure 3. A) Examples of scene images used in the study. Each scene depicted a natural (outdoors) or man-made (indoors) context, and a natural (animal) or man-made (furniture) foreground object. The set was created in a full-factorial way: each context and object appeared in a congruent and incongruent version. B) A schematic depiction of a trial sequence. Experiment 1 consisted of 4 blocks of a go-nogo task, with scene contexts (natural or man-made) or objects (natural or man-made) as targets. Experiment 2 consisted of 2 blocks of the 2AFC task: participants categorized scene contexts and objects (in separate blocks) as natural or man-made.

2.1.2.5. Data analysis

2.1.2.5.1. Analysis of behavioral data

All analyses of behavioral data were conducted using custom-made R and Python scripts. The following measures of behavioral performance were analysed: accuracy (the d' index and the proportion of correct responses), reaction times, and Balanced Integration Score (BIS).

2.1.2.5.1.1 Accuracy

In Experiment 1, the d' index was calculated as a measure of perceptual sensitivity. The d' index, based on signal detection theory (SDT), quantifies the observer's ability

to discriminate signal from noise. It is computed as the difference between the z-scores of hit rates (proportion of correct "go" responses) and false alarm rates (proportion of incorrect "go" responses).

Hit and false alarm rates were determined for each participant and each stimulus condition, as defined by task (context/object), congruency (congruent/incongruent), and target category (natural/man-made). "Go" responses occurring within 200 ms post-stimulus onset were considered anticipatory and incorrect. All hit and false alarm rates were corrected using the log-linear rule, considered the least biased method to address extreme values (i.e., hit and false alarm rates of 0 or 1; Hautus, 1995; Stanislaw & Todorov, 1999). Higher d' indicates better performance.

In Experiment 2, the proportion of correct responses (PC) was calculated for each participant and each stimulus condition. As in Experiment 1, responses occurring within 200 ms post-stimulus onset were classified as incorrect, and the stimulus condition was defined by task, congruency, and target category.

2.1.2.5.1.2 Reaction times

For both experiments, only the correct responses that occurred more than 200 ms post-stimulus onset were included in the analyses. Only "go" responses were analyzed for Experiment 1. The mean reaction times (RTs) were calculated for each participant and each stimulus condition, as defined by task (scene context/object), congruency (congruent/incongruent), and target category (natural/man-made). All RTs were calculated in ms.

Further, to examine how quickly interference from semantic incongruency emerges during the categorization process, the RT distributions for each participant and stimulus condition were first divided into three quantiles (bins), following the Vincentization procedure introduced by Ratcliff (1979). Second, mean RTs were computed for each participant, stimulus condition, and bin, and then averaged across participants. Finally, an interference index was calculated by subtracting the mean RTs of congruent trials from those of incongruent trials for each condition, as defined by bin, task, and target category.

2.1.2.5.1.3 Balanced Integration Score (BIS)

For both experiments, a Balanced Integration Score (BIS; Liesefeld & Janczyk, 2019, 2023) was calculated for each participant and each condition. The BIS combines speed and accuracy data in a way that attenuates the influence of speed–accuracy trade-off (SAT) while maintaining true effects. It is calculated according to the following formula:

$$(1) \quad BIS_{i,j} = z_{i,j}^{PC} - z_{i,j}^{RT} = \frac{PC_{i,j} - \underline{PC}}{S^{PC}} - \frac{RT_{i,j} - \underline{RT}}{S^{RT}},$$

where $z_{i,j}^x$ represents the z-standardized performance (either mean RT or PC) for a participant i in condition j , S^{RT} denotes the standard deviation (SD) of the mean RTs used in calculating BIS, the grand mean RT (\underline{RT}) refers to the overall average of mean RTs across all conditions and participants, and \underline{PC} indicates the overall average proportion of correct responses across all conditions and participants. Higher BIS indicates better performance.

In the go/nogo task, only RTs for “go” trials could be used. The accuracy was calculated as the total number of correct responses (hits on "go" trials and correct inhibitions on "no-go" trials) divided by the total number of trials.

2.1.2.5.1.4. Statistical analysis

Repeated-measure analysis of variance (rm-ANOVA) was used to analyze mean accuracy (d' index in Experiment 1, proportions of correct responses in Experiment 2), mean RTs, and mean BIS. The following variables were included in the analysis as factors: task (scene context/object categorization), congruency (congruent/incongruent), and target category (natural/man-made). Rm-ANOVA with the task, target category, and bin (first/second/third) as factors was used to analyze the timing of the interference effect. All variables were introduced in the ‘within-subject’ design. All statistical analyses were conducted using JASP 0.18.3.0 software (JASP Team, 2023). Values are reported as Mean \pm SEM. For all statistical tests, probability values were reported (p), and the standard 0.05 alpha level was used as a threshold for rejecting the null hypothesis. The Greenhouse-Geisser (GG) correction was applied when necessary to account for violations of sphericity.

2.1.3. Behavioural results

2.1.3.1 Experiment 1: go/no-go

2.1.3.1.1. Accuracy

The ANOVA revealed a main effect of *task* ($F(1,28) = 9.349$, $p = .005$, $\eta_p^2 = .250$), indicating that scene contexts were discriminated better than objects (context: 3.81 ± 0.12 ; object: 3.56 ± 0.10). Additionally, a main effect of *congruency* ($F(1,28) = 49.984$, $p < .001$, $\eta_p^2 = .641$) was observed, suggesting that discrimination of both scene contexts and objects was better when the scene was congruent than incongruent (congruent: 3.88 ± 0.11 ; incongruent: 3.48 ± 0.11). A significant interaction between *congruency* and *target category* ($F(1,28) = 22.465$, $p = .001$, $\eta_p^2 = .445$) was also found.

To explore the interaction, the analysis of simple main effects was conducted. It revealed that discrimination of both natural (congruent: 3.75 ± 0.10 ; incongruent: 3.58 ± 0.11 ; $F(1,28) = 5.178$, $p = .031$, $\eta_p^2 = .156$) and man-made (congruent: 4.02 ± 0.11 ; incongruent: 3.39 ± 0.11 ; $F(1,28) = 76.263$, $p < .001$, $\eta_p^2 = .731$) targets was significantly better when trials were congruent than incongruent. Further, while natural targets were discriminated better than man-made targets when incongruent trials were compared (natural: 3.58 ± 0.11 ; man-made: 3.39 ± 0.11 ; $F(1,28) = 14.117$, $p < .001$, $\eta_p^2 = .335$), man-made targets were discriminated significantly better than natural ones in congruent trials (natural: 3.75 ± 0.10 ; man-made: 4.02 ± 0.11 ; $F(1,28) = 6.583$, $p = .016$, $\eta_p^2 = .190$).

Figure 4 depicts the interaction between congruency and target category. Table 1 summarizes the ANOVA results.

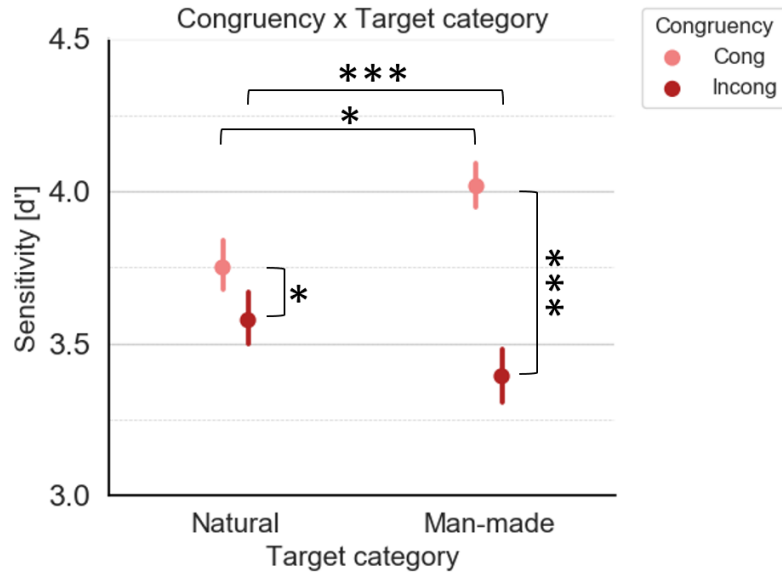


Figure 4. The interaction between congruency and target category. Congruent trials were discriminated better than incongruent ones for natural and man-made targets. Natural targets were discriminated better than man-made targets in incongruent trials, and man-made targets were discriminated better than natural ones in congruent trials. Dots represent means, whiskers designate standard error of the mean (SEM), and asterisks (*, **, ***) indicate $p < .05$, $p < .01$, and $p < .001$, respectively.

Table 1. A summary of a 2x2x2 rm-ANOVA, with the task (scene context, object), congruency (congruent, incongruent), and target category (natural, man-made) as within-subject factors. D' index is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	28	9.349	.005	.250
congruency	1	28	49.984	<.001	.641
task * congruency	1	28	1.615	.214	.055
target category	1	28	0.395	.535	.014
task * target category	1	28	0.056	.814	.002
congruency * target category	1	28	22.465	<.001	.445
task*congruency*target category	1	28	3.081	.090	.099

2.1.3.1.2. Reaction times

The analysis of mean RTs revealed a significant main effect of *task* ($F(1,28) = 17.051$, $p < .001$, $\eta_p^2 = .378$), indicating that responses were faster to objects compared to scene contexts (context: 379 ± 12 ms; object: 363 ± 10 ms). Further, there was a significant main effect of *congruency* ($F(1,28) = 18.564$, $p < .001$, $\eta_p^2 = .399$), with faster responses to congruent as compared to incongruent trials (congruent: 368 ± 11 ms; incongruent: 375 ± 11 ms) and a significant main effect of *target category* ($F(1,28) = 18.564$, $p < .001$, $\eta_p^2 = .239$), indicating faster responses to natural compared to man-made targets (natural: 366 ± 10 ms; man-made: 376 ± 11 ms). Additionally, significant interactions were observed between *task* and *congruency* ($F(1,28) = 12.049$, $p = .002$, $\eta_p^2 = .301$), and *congruency* and *target category* ($F(1,28) = 58.958$, $p < .001$, $\eta_p^2 = .678$).

The simple main effects analysis of the interaction between *task* and *congruency* revealed that objects were classified faster than scene contexts in both congruent (context: 373 ± 11 ms; object: 363 ± 10 ms; $F(1,28) = 6.660$, $p = .015$, $\eta_p^2 = .192$) and incongruent trials (context: 386 ± 12 ms; object: 364 ± 10 ms; $F(1,28) = 24.518$, $p < .001$, $\eta_p^2 = .467$). Further, scene contexts were classified faster in congruent than incongruent trials (congruent: 373 ± 11 ms; incongruent: 386 ± 12 ms; $F(1,28) = 32.726$, $p < .001$, $\eta_p^2 = .539$), but no such effect was observed in the object task (congruent: 363 ± 10 ms; incongruent: 364 ± 10 ms; $F(1,28) = 0.439$, $p = .513$, $\eta_p^2 = .015$).

The simple main effects analysis of the interaction between *congruency* and *target category* revealed that congruent trials were classified significantly faster than incongruent ones for man-made targets (congruent: 367 ± 11 ms; incongruent: 386 ± 11 ms; $F(1,28) = 54.466$, $p < .001$, $\eta_p^2 = .660$), while for natural ones the reverse pattern was observed, with incongruent trials being classified faster than congruent ones (congruent: 368 ± 11 ms; incongruent: 364 ± 10 ms; $F(1,28) = 4.396$, $p = .045$, $\eta_p^2 = .136$). Further, natural targets were classified significantly faster than man-made ones, but this effect was found only when incongruent trials were compared (natural: 364 ± 10 ms; man-made: 386 ± 11 ms; $F(1,28) = 33.162$, $p < .001$, $\eta_p^2 = .542$). The comparison of congruent trials yielded no significant differences between target categories (natural: 368 ± 11 ms; man-made: 367 ± 11 ms; $F(1,28) = 0.160$, $p = .692$, $\eta_p^2 = .006$).

Figure 5 depicts the interactions between task and congruency (A) and congruency and target category (B). Table 2 summarizes the ANOVA results.

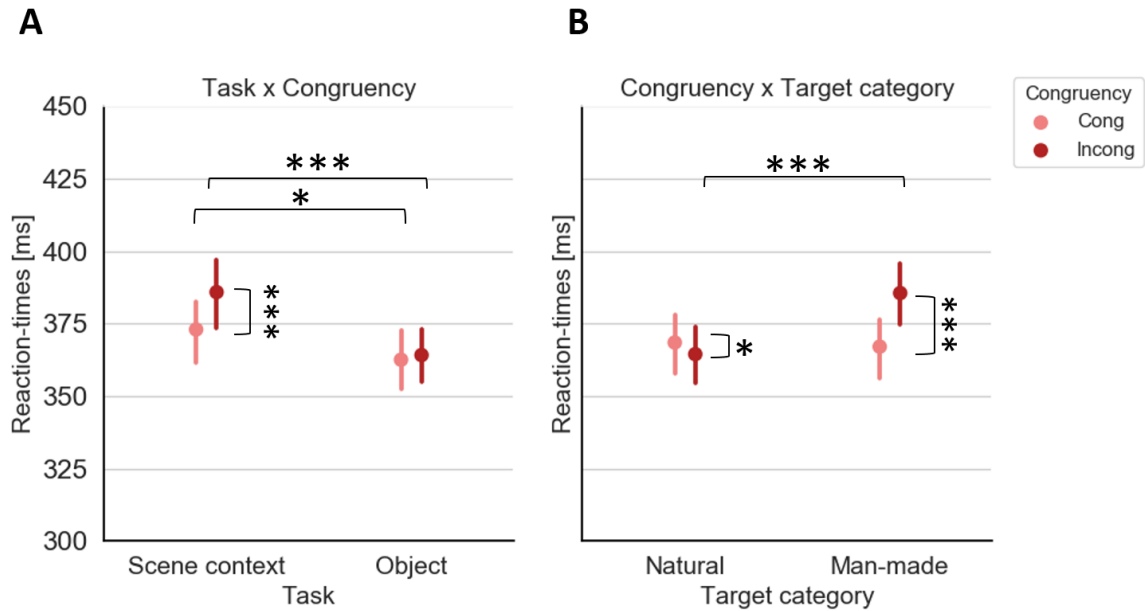


Figure 5. A) The interaction between task and congruency. Objects were classified faster than scene contexts in both congruent and incongruent trials. Congruent trials were classified faster than incongruent ones only in the scene context task. B) The interaction between target category and congruency. Congruent trials were classified faster than incongruent ones for man-made targets. Incongruent trials were classified faster than congruent ones for natural targets. Natural targets were classified faster than man-made ones, but only in incongruent trials. Dots represent means, whiskers designate standard error of the mean (SEM), and asterisks (*, **, ***) indicate $p < .05$, $p < .01$, and $p < .001$, respectively.

Table 2. A summary of a 2x2x2 rm-ANOVA, with the task (scene context, object), congruency (congruent, incongruent), and target category (natural, man-made) as within-subject factors. Mean RT is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	28	17.051	<.001	.378
congruency	1	28	18.564	<.001	.399
task * congruency	1	28	12.049	.002	.301
target category	1	28	8.816	.006	.239
task * target category	1	28	0.522	.476	.018
congruency * target category	1	28	58.958	<.001	.678
task*congruency*target category	1	28	1.091	.305	.038

The analysis of interference indices revealed significant main effects of *task* ($F(1,28) = 11.167$, $p = .002$, $\eta_p^2 = .285$), *target category* ($F(1,28) = 57.658$, $p < .001$, $\eta_p^2 = .673$), and *bin* ($F(1.194,33.422) = 6.539$, $p = .012$, $\eta_p^2 = .189$). Additionally, significant interactions were observed between *task* and *bin* ($F(1.327,37.146) = 7.789$, $p = .005$, $\eta_p^2 = .218$), and *target category* and *bin* ($F(1.260,35.293) = 10.578$, $p = .001$, $\eta_p^2 = .274$).

Post-hoc t-tests with Bonferroni corrections indicated that the interference effect was significantly larger in the scene context compared to the object task (context: 13 ± 5 ms; object: 2 ± 5 ms; $t(28) = 3.342$, $p_{adj} = .002$, $d = 0.488$), for man-made targets compared to natural targets (natural: -4 ± 4 ms; man-made: 18 ± 5 ms; $t(28) = 7.593$, $p_{adj} < .001$, $d = 0.995$), and for the third compared to the first bin (first: 3 ± 3 ms; third: 12 ± 7 ms; $t(28) = 3.601$, $p_{adj} = .002$, $d = 0.389$). No significant differences were found between the second and third bin (second: 7 ± 3 ms; third: 12 ± 7 ms; $t(28) = -2.085$, $p_{adj} = .125$, $d = 0.225$) and between the first and the second bin ($t(28) = -1.517$, $p_{adj} = .405$, $d = 0.164$).

The simple main effects analysis of the interaction between *task* and *bin* revealed that the effect of bin was significant in the scene context task (first: 5 ± 3 ms; second: 10 ± 3 ms; third: 23 ± 7 ms; $F(2,56) = 21.579$, $p < .001$, $\eta_p^2 = .435$), but not in the object task (first: 1 ± 2 ms; second: 3 ± 3 ms; third: 1 ± 7 ms; $F(2,56) = 0.170$, $p = .844$, $\eta_p^2 = .006$). Further, the difference between tasks in the interference effect was significant between the second bins (context: 10 ± 3 ms; object: 3 ± 3 ms; $F(1,28) = 7.221$, $p = .012$, $\eta_p^2 = .205$) and the third

bins (context: 23 ± 7 ms; object: 1 ± 7 ms; $F(1,28) = 11.245$, $p = .002$, $\eta_p^2 = .287$), but not the first ones (context: 5 ± 3 ms; object: 1 ± 2 ms; $F(1,28) = 2.511$, $p = .124$, $\eta_p^2 = .082$).

The simple main effects analysis of the interaction between *target category* and *bin* revealed that the effect of bin was significant for man-made (first: 9 ± 3 ms; second: 17 ± 3 ms; third: 29 ± 7 ms; $F(2,56) = 13.255$, $p < .001$, $\eta_p^2 = .321$), but not natural targets (first: -3 ± 2 ms; second: -3 ± 3 ms; third: -5 ± 7 ms; $F(2,56) = 0.313$, $p = .732$, $\eta_p^2 = .011$). Further, the difference in the interference effect between target categories was significant between all the bins: the first (natural: -3 ± 2 ms; man-made: 9 ± 3 ms; $F(1,28) = 27.081$, $p < .001$, $\eta_p^2 = .492$), the second (natural: -3 ± 3 ms; man-made: 17 ± 3 ms; $F(1,28) = 51.418$, $p < .001$, $\eta_p^2 = .647$), and the third (natural: -5 ± 7 ms; man-made: 29 ± 7 ms; $F(1,28) = 33.354$, $p < .001$, $\eta_p^2 = .544$).

Figure 6 depicts the interactions between task and bin (A), and target category and bin (B). Table 3 summarizes the ANOVA results.

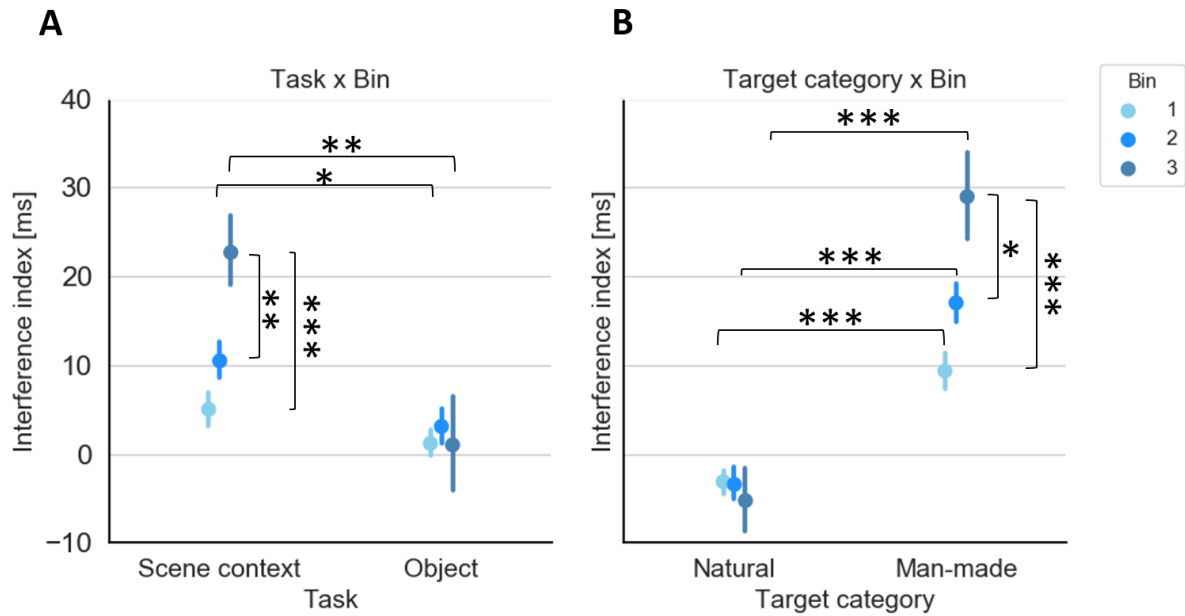


Figure 6. A) The interaction between bin and task. The interference effect was larger in the third than in the second and first bins in the scene context task. No significant differences between bins were observed in the object task. The interference was larger in the scene context compared to the object task in the second and third bins. B) The interaction between bin and target category. The interference effect was larger in the third bin than in the second and first bins for man-made targets. No significant differences between bins were observed for natural targets. The interference effect was larger for man-made than natural targets in all bins. Dots represent means, whiskers designate standard error of the mean (SEM), and asterisks (*, **, ***) indicate $p < .05$, $p < .01$, and $p < .001$, respectively.

Table 3. A summary of a 2x2x3 rm-ANOVA, with the task (scene context, object), congruency (congruent, incongruent), and bin (first, second, third) as within-subject factors. The difference in mean RT between incongruent and congruent trials is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	28	11.167	.002	.285
target category	1	28	57.658	<.001	.673
task *target category	1	28	1.054	.313	.036
bin	1.194	33.422	6.539	.012	.189
task * bin	1.327	37.146	7.789	.005	.218
target category*bin	1.260	35.293	10.578	.001	.274
task*target category*bin	1.075	30.091	2.200	.147	.073

2.1.3.1.3. BIS

The analysis revealed a main effect of *congruency* ($F(1,28) = 46.284, p < .001, \eta_p^2 = .623$), indicating that performance was better for congruent compared to incongruent trials (congruent: 0.41 ± 0.21 ; incongruent: -0.41 ± 0.22). Additionally, a significant interaction of *congruency* and *target category* ($F(1,28) = 67.529, p < .001, \eta_p^2 = .707$) was observed.

The simple main effects analysis revealed that performance was better in congruent than incongruent trials, but only for man-made (congruent: 0.62 ± 0.22 ; incongruent: -0.80 ± 0.22 ; ($F(1,28) = 85.158, p < .001, \eta_p^2 = .753$) and not natural targets (congruent: 0.20 ± 0.19 ; incongruent: -0.02 ± 0.20 ; ($F(1,28) = 3.235, p = .083, \eta_p^2 = .104$). Further, a better performance was observed for man-made compared to natural targets in congruent trials (natural: 0.20 ± 0.19 ; man-made: 0.62 ± 0.22 ; ($F(1,28) = 6.413, p = .017, \eta_p^2 = .186$), while in incongruent ones the reverse pattern was observed, with better performance for natural than man-made targets (natural: -0.02 ± 0.20 ; man-made: -0.80 ± 0.22 ; ($F(1,28) = 44.119, p < .001, \eta_p^2 = .612$).

Figure 7 depicts the interaction between congruency and target category. Table 4 summarizes the ANOVA results.

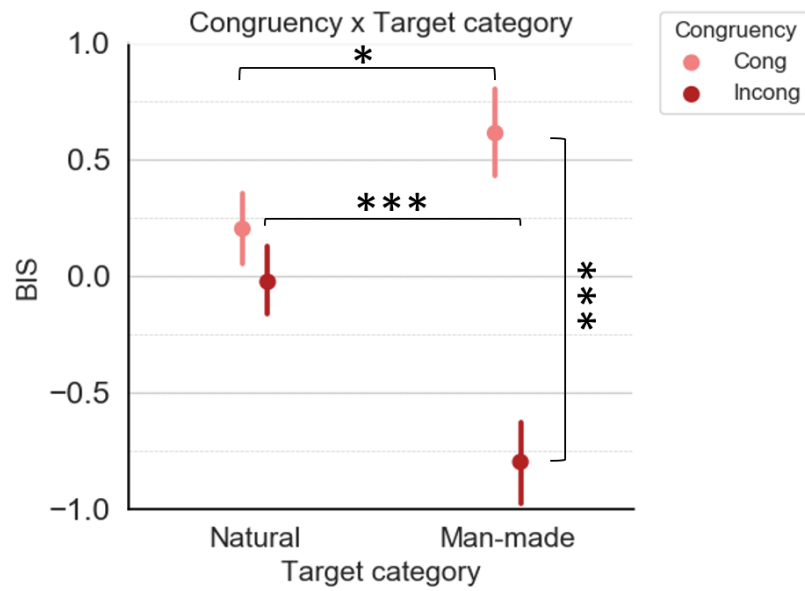


Figure 7. Interaction between target category and congruency. Congruent trials were classified better than incongruent ones for man-made targets, but not for natural ones. Natural targets were detected better than man-made ones in incongruent trials, and man-made targets were detected better than natural ones in congruent trials. Dots represent means, whiskers designate standard error of the mean (SEM), and asterisks (*, **, ***) indicate $p < .05$, $p < .01$, and $p < .001$, respectively.

Table 4. A summary of a 2x2x2 rm-ANOVA, with the task (scene context, object), congruency (congruent, incongruent), and target category (natural, man-made) as within-subject factors. The BIS index is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	28	0.053	.819	.002
congruency	1	28	46.284	<.001	.623
task * congruency	1	28	0.091	.766	.003
target category	1	28	2.224	.147	.074
task * target category	1	28	1.253	.272	.043
congruency * target category	1	28	67.529	<.001	.707
task*congruency*target category	1	28	2.945	.097	.095

2.1.3.2 Experiment 2: 2AFC

2.1.3.2.1. Accuracy

The analysis revealed a main effect of *task* ($F(1,30) = 8.717$, $p = .006$, $\eta_p^2 = .225$), indicating that the categorization of scene contexts was more accurate than the categorization of objects (context: 0.921 ± 0.009 ; object: 0.904 ± 0.010). Additionally, a main effect of *congruency* ($F(1,30) = 26.078$, $p < .001$, $\eta_p^2 = .465$) was observed, indicating better performance for congruent compared to incongruent trials (congruent: 0.926 ± 0.009 ; incongruent: 0.900 ± 0.010). Finally, a main effect of *target category* ($F(1,30) = 5.060$, $p = .032$, $\eta_p^2 = .144$) was found, with responses being more accurate for man-made targets compared to natural ones (natural: 0.905 ± 0.010 ; man-made: 0.920 ± 0.010). None of the interactions between factors were significant.

Table 5 summarizes the ANOVA results.

Table 5. A summary of a 2x2x2 rm-ANOVA, with the task (scene context, object), congruency (congruent, incongruent), and target category (natural, man-made) as within-subject factors. The mean proportion of correct responses is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	30	8.717	.006	.225
congruency	1	30	26.078	<.001	.465
task * congruency	1	30	1.881	.180	.059
target category	1	30	5.060	.032	.144
task * target category	1	30	0.002	.962	<.001
congruency * target category	1	30	3.516	.071	.105
task*congruency*target category	1	30	0.492	.488	.016

2.1.3.2.2. Reaction times

The analysis of mean RTs revealed a main effect of *task* ($F(1,30) = 7.926, p = .009, \eta_p^2 = .209$), indicating that responses were faster to objects compared to scene contexts (context: 473 ± 13 ms; object: 455 ± 11 ms). Additionally, a main effect of *congruency* ($F(1,30) = 25.041, p < .001, \eta_p^2 = .455$) was observed, with faster responses to congruent compared to incongruent trials (congruent: 459 ± 12 ms; incongruent: 469 ± 12 ms). A significant interaction between *congruency* and *target category* ($F(1,30) = 12.256, p = .001, \eta_p^2 = .290$) was also found.

The simple main effects analysis revealed that congruent trials were classified significantly faster than incongruent ones for man-made targets (congruent: 454 ± 12 ms; incongruent: 472 ± 12 ms; ($F(1,30) = 28.825, p < .001, \eta_p^2 = .490$), but not for natural ones (congruent: 464 ± 13 ms; incongruent: 466 ± 12 ms; ($F(1,30) = 0.292, p = .593, \eta_p^2 = .010$). Further, significantly faster responses to man-made targets compared to natural ones were observed in congruent trials (natural: 464 ± 13 ms; man-made: 454 ± 12 ms; ($F(1,30) = 5.036, p = .032, \eta_p^2 = .144$), but not in incongruent ones (natural: 466 ± 12 ms; man-made: 472 ± 12 ms; ($F(1,30) = 1.718, p = .200, \eta_p^2 = .054$).

Figure 8 depicts the interaction between congruency and target category. Table 6 summarizes the ANOVA results.

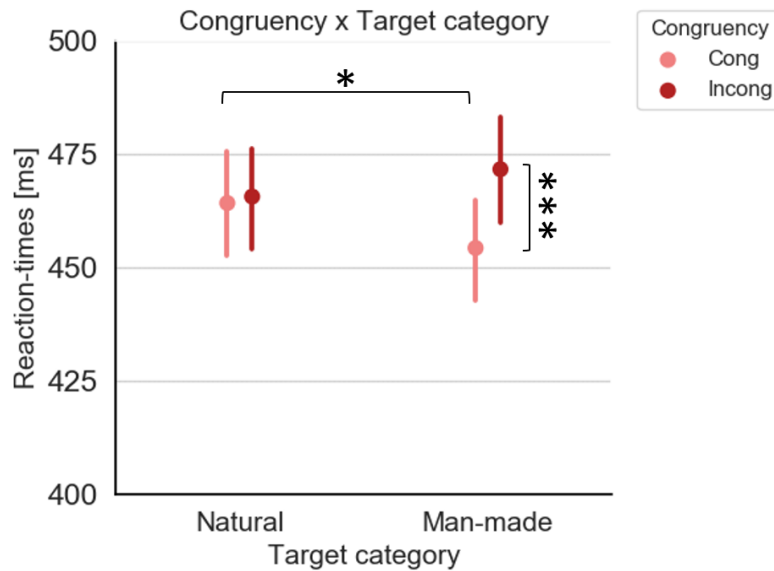


Figure 8. The interaction between congruency and target category. Congruent trials were classified significantly faster than incongruent ones for man-made targets, but not for natural ones. Responses were faster to man-made compared to natural targets, but only in congruent trials. Dots represent means, whiskers designate standard error of the mean (SEM), and asterisks (*, **, ***) indicate $p < .05$, $p < .01$, and $p < .001$, respectively.

Table 6. A summary of a 2x2x2 rm-ANOVA, with the task (scene context, object), congruency (congruent, incongruent), and target category (natural, man-made) as within-subject factors. Mean RT is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	30	7.926	.009	.209
congruency	1	30	25.041	<.001	.455
task * congruency	1	30	0.002	.962	<.001
target category	1	30	0.249	.622	.008
task * target category	1	30	0.762	.390	.025
congruency * target category	1	30	12.256	.001	.290
task*congruency*target category	1	30	3.910	.057	.115

The analysis of interference indices revealed a main effect of *target category* ($F(1,30) = 12.659, p = .001, \eta_p^2 = .297$) and a main effect of *bin* ($F(1.161, 34.836) = 15.097, p < .001, \eta_p^2 = .335$). Post-hoc t-tests with Bonferroni correction confirmed that the effect of interference was significantly higher for man-made compared to natural targets (natural: 1 ± 5 ms; man-made: 17 ± 6 ms; $t(30) = 3.558, p_{adj} = .001, d = 0.544$). Further, the interference effect was significantly higher in the third bin compared to the first bin (first: 2 ± 3 ms; third: 18 ± 8 ms; $t(30) = 5.428, p_{adj} < .001, d = 0.556$) and the second bin (second: 8 ± 4 ms; third: 18 ± 8 ms; $t(30) = 3.455, p_{adj} = .003, d = 0.354$). No significant difference was found between the first and the second bin ($t(30) = -1.973, p_{adj} < .159, d = -0.202$).

Figure 9 depicts the main effect of bin. Table 7 summarizes the ANOVA results.

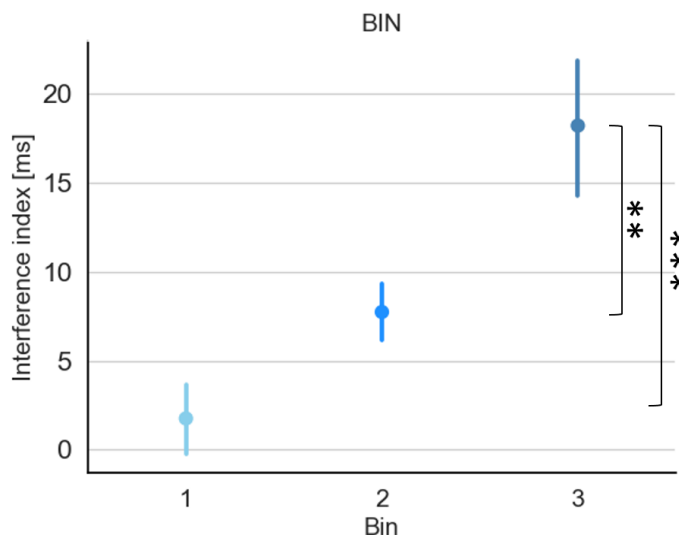


Figure 9. The main effect of bin. The interference effect was significantly higher in the third bin compared to the second and first bins. Dots represent means, whiskers designate standard error of the mean (SEM), and asterisks (**, ***) indicate $p < .01$ and $p < .001$, respectively.

Table 7. A summary of a 2x2x3 rm-ANOVA, with the task (scene context, object), congruency (congruent, incongruent), and bin (first, second, third) as within-subject factors. The difference in mean RT between incongruent and congruent trials is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	30	0.001	.970	<.001
target category	1	30	12.659	.001	.297
task *target category	1	30	4.087	.052	.120
bin	1.161	34.836	15.097	<.001	.335
task * bin	1.251	37.515	3.043	.081	.092
target category*bin	1.190	35.689	2.729	.102	.083
task*target category*bin	1.233	36.988	0.784	.407	.025

2.1.3.2.3. BIS

The analysis revealed a main effect of *congruency* ($F(1,30) = 35.880$, $p < .001$, $\eta_p^2 = .545$), indicating that performance was better for congruent compared to incongruent trials (congruent: 0.31 ± 0.26 ; incongruent: -0.31 ± 0.25). Additionally, a significant interaction of *congruency* and *target category* ($F(1,30) = 8.677$, $p = .006$, $\eta_p^2 = .224$) was observed.

The simple main effects analysis revealed that performance was better in congruent than incongruent trials for both natural (congruent: 0.03 ± 0.29 ; incongruent: -0.32 ± 0.25 ; ($F(1,30) = 5.873$, $p = .022$, $\eta_p^2 = .164$) and man-made targets (congruent: 0.60 ± 0.23 ; incongruent: -0.30 ± 0.25 ; ($F(1,30) = 45.308$, $p < .001$, $\eta_p^2 = .602$). Further, a better performance was observed for man-made compared to natural targets in congruent trials (natural: 0.03 ± 0.29 ; man-made: 0.60 ± 0.23 ; ($F(1,30) = 9.490$, $p = .004$, $\eta_p^2 = .240$), but not in incongruent ones (natural: -0.32 ± 0.25 ; man-made: -0.30 ± 0.25 ; ($F(1,30) = 0.028$, $p = .868$, $\eta_p^2 = .001$).

Figure 10 depicts the interaction between congruency and target category. Table 8 summarizes the ANOVA results.

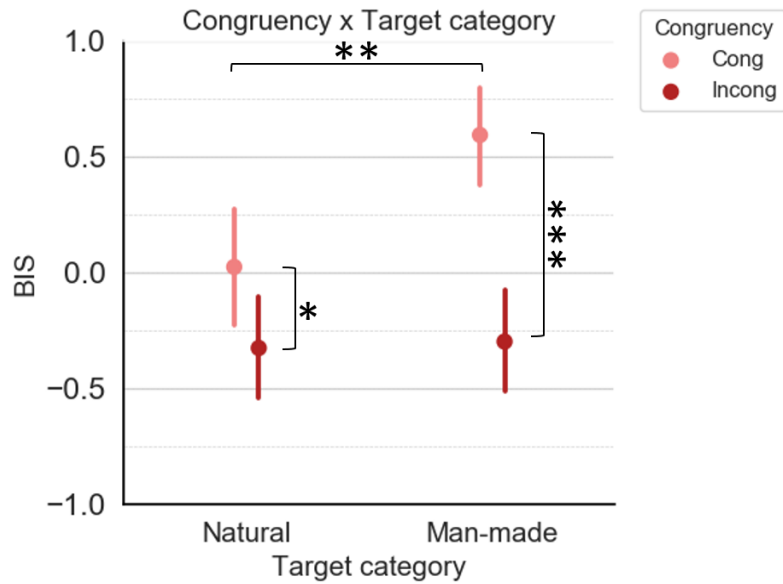


Figure 10. Interaction between congruency and target category. Congruent targets were detected better than incongruent ones for both natural and man-made targets. Man-made targets were detected better than natural ones in congruent trials. Dots represent means, whiskers designate standard error of the mean (SEM), and asterisks (*, **, ***) indicate $p < .05$, $p < .01$, and $p < .001$, respectively.

Table 8. A summary of a 2x2x2 rm-ANOVA, with the task (scene context, object), congruency (congruent, incongruent), and target category (natural, man-made) as within-subject factors. The BIS index is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	30	0.098	.757	.003
congruency	1	30	35.880	<.001	.545
task * congruency	1	30	1.667	.206	.053
target category	1	30	3.986	.055	.117
task * target category	1	30	0.143	.708	.005
congruency * target category	1	30	8.677	.006	.224
task*congruency*target category	1	30	1.354	.254	.043

2.1.4. Discussion

The present study examined predictions derived from hierarchical accounts of real-world scene perception, contrasting analytical (local-first) and holistic (global-first) perspectives on scene–object interactions. Analytic accounts predict faster object recognition and greater costs to scene classification under object–scene incongruency (local-to-global interference), whereas holistic accounts predict faster scene processing and greater costs to object classification under scene–object incongruency (global-to-local interference). Target category (natural vs. man-made) was included as an exploratory factor. To my knowledge, this is the first study to directly compare the speed of scene context and object processing within the same participants across tasks, using the same stimulus set and an orthogonal manipulation of scene–object congruency.

The results revealed a clear speed–accuracy trade-off (SAT) in both paradigms (go/no-go, 2AFC): participants were more accurate when classifying scene context but faster when responding to objects. To accommodate this trade-off, the Balanced Integration Score (BIS; Liesefeld & Janczyk, 2019, 2023), which combines speed and accuracy into one performance metric, was computed. After controlling for the SAT with BIS, performance no longer differed between scene and object tasks in either paradigm.

The absence of primacy for scenes or objects challenges hierarchical accounts positing a fixed processing advantage and stands in contrast to reports of global dominance for artificial stimuli (Campana et al., 2016). Prior work conducted in separate literatures suggests that the time course of scene context categorization can match that of object categorization when the scene target is defined at the superordinate level (e.g., natural vs. man-made; Joubert et al., 2007) but is slower at the basic level (e.g., beach, mountain, city; Goffaux et al., 2005; Rousselet et al., 2005). Using the same stimulus set and the same level of target definition for scenes and objects, the study provides direct evidence that their time courses converge once the SAT is controlled. These parallel scene- and object-processing dynamics for unambiguous images are consistent with predictive processing accounts of real-world scene perception.

Further, across both paradigms (go/no-go; 2AFC), congruent images were classified more accurately and faster than incongruent ones. In the go/no-go experiment, a congruency effect on RTs appeared only for scene context judgments. However, after controlling the SAT with BIS, the effect no longer depended on the task, emerging similarly for scene and object

classification. This pattern argues against hierarchical accounts positing an unidirectional influence and instead points to a bidirectional interplay between scene context and object information. Such a result stays in line with previous findings, showing the reciprocal character of scene–object influences (e.g., Furtak et al., 2022; Leroy et al., 2020).

In terms of timing, vincentized RT distributions showed minimal incongruency costs in the fastest third of responses and substantially higher costs in the slowest third. This pattern indicates that interference builds over time. The temporal stage of scene–object interplay is debated. Perceptual accounts ascribe the cost to early disruptions of natural scene statistics in incongruent images, which impair initial encoding and slow down evidence accumulation (Mack & Palmeri, 2010, Oliva and Torralba, 2001). Other views place the effect at later semantic (Leroy et al., 2020) or even decisional stages (Hollingworth & Henderson, 1998). In the presented data, the near absence of costs among the fastest responses favors the latter. Interestingly, these results contrast with reports of context incongruency affecting even the earliest behavioral responses to target objects (e.g., Joubert et al., 2008). The discrepancies with the previous studies, however, might also stem from the differences in the stimulus characteristics. The magnitude of incongruency effects has been shown to depend on stimulus properties: the effect decreases as objects become larger (Fize et al., 2011). In the stimulus set used in the present study, both scene and object cues were reliable, and objects were relatively large (Remy et al., 2013). Under these conditions, viewed through predictive processing lenses, scene–object interactions at an early level may not have been necessary to complete the tasks.

Finally, the impact target category on scene perception was examined. Based on prior findings (Kirchner & Thorpe, 2006; Crouzet et al., 2012; Rémy et al., 2013, 2014; Rousselet et al., 2005; but see: Joubert et al., 2007; Rémy et al., 2020; VanRullen & Thorpe, 2001), it was hypothesized that natural targets – particularly animals (Crouzet et al., 2012) – might be processed preferentially. Using BIS to control the SAT, it was found that across both paradigms (go/no-go, 2AFC), congruent man-made targets were classified better than congruent natural targets. On incongruent trials, performance in the go/no-go experiment was better for natural than for man-made targets, whereas in the 2AFC the two target types did not differ. Overall, target-type effects were congruency-dependent, not task-dependent, and there was no general natural target advantage.

Taken together, these findings do not support a hierarchical (global-first vs. local-first) organization of perception for naturalistic stimuli. After controlling for the SAT, scene

and object classification proceeded at comparable rates. Congruency effects were symmetric across tasks and increased with response time: minimal in the fastest trials and largest in the slowest. This temporal profile points to a bidirectional interplay that arises at late, decisional stages rather than from an early perceptual interaction. The pattern stays in line with the predictive processing account: when neither scene nor object is ambiguous, parallel processing suffices, and early interactions are unnecessary. Finally, no general advantage for natural (animal) targets was found. Instead, a more nuanced, congruency-dependent pattern emerged across conditions.

2.2. Study 2: testing the causal role of object representations in disambiguating scenes

2.2.1. Research questions and hypotheses

In recent years, an increasing number of behavioural studies have provided evidence for bidirectional interactions between scene context and objects, showing that not only scenes facilitate object recognition but objects can also influence scene interpretation (Furtak et al., 2022; Joubert et al., 2007; Leroy et al., 2020; Mack & Palmieri, 2010). Study 1 extends this behavioural evidence by directly comparing the time courses of scene and object processing and demonstrating their mutual influences.

Consistent with this behavioural work, the reciprocal nature of these interactions has also been investigated at the neural level: fMRI and MEG decoding studies using ambiguous images show that scene context sharpens object representations in the visual cortex (Brandman & Peelen, 2017), whereas objects sharpen scene representations (Brandman & Peelen, 2019). Moreover, the temporal profiles of scene-to-object and object-to-scene influences are broadly comparable, with disambiguation of either representation emerging around 300 ms after stimulus onset (Brandman & Peelen, 2017, 2023). Crucially, TMS work has demonstrated that the scene-selective cortex (OPA) is causally and selectively involved in context-based object disambiguation between 160 and 200 ms after image onset (Wischnewski & Peelen, 2021b). It remains unknown, however, whether object representations in the object-selective cortex (LOC) also causally disambiguate scenes and, if so, how the LOC and the OPA interact over time during object-based scene recognition. Study 2 was designed to address these questions.

The study consisted of two pre-registered TMS experiments conducted on the same participants: a four-pulse TMS study ([#153165 | AsPredicted](#)) and a chronometric TMS study ([#153967 | AsPredicted](#)). The latter included two experiments: the OPA and the LOC. A schematic overview of the two-session TMS study is presented in Figure 11.

During the first experimental session, participants performed an object recognition task (Wischnewski and Peelen, 2021b) and a scene classification task while receiving four TMS pulses, starting from image onset, on the left LOC, the left OPA, and the vertex. The choice of the left OPA and the left LOC was based on previous fMRI and MEG decoding studies showing that the effect of object-based facilitation of scene representation was restricted to

the left hemisphere (Brandman & Peelen, 2019) and strongest over left posterior sensors (Brandman & Peelen, 2023). The images were presented in two conditions in each task. In the object recognition task, the picture depicted either an intact object on a gray background (*isolated object* condition, IO) or a degraded object presented within a congruent scene context (*object-with-scene* condition, OS). In the scene classification task, either a slightly degraded scene without an object (*isolated scene* condition, IS) or a degraded scene presented with one single intact object (*scene-with-object* condition, WO) was presented (see Section 2.2.2.4. and Figures 14A and 15A).

The individual data from the first session were used to assign participants to stimulation sites in the chronometric TMS study (see Section 2.2.2.4.1.2.). Such a TMS-based assignment procedure increases the chances of finding time-sensitive TMS effects within a given region by reducing the impact of interindividual variation due to factors, such as, for example, skull-thickness, localization of the area based on a template or subject-specific gyral folding pattern (Wischnewski and Peelen, 2021b). For the present thesis, the four-pulse TMS study was described solely as a selection procedure for the subsequent chronometric TMS experiment. The pre-registered analyses and results for the first study are described in Appendix A, but not discussed as part of this work.

In the chronometric TMS study, participants performed the same scene classification task as in the first session, while receiving two TMS pulses at one of the three time points (early: 60–100ms; middle: 160–200ms; or late: 260–300ms) relative to stimulus onset. Pulses were delivered over either the left LOC or the left OPA.

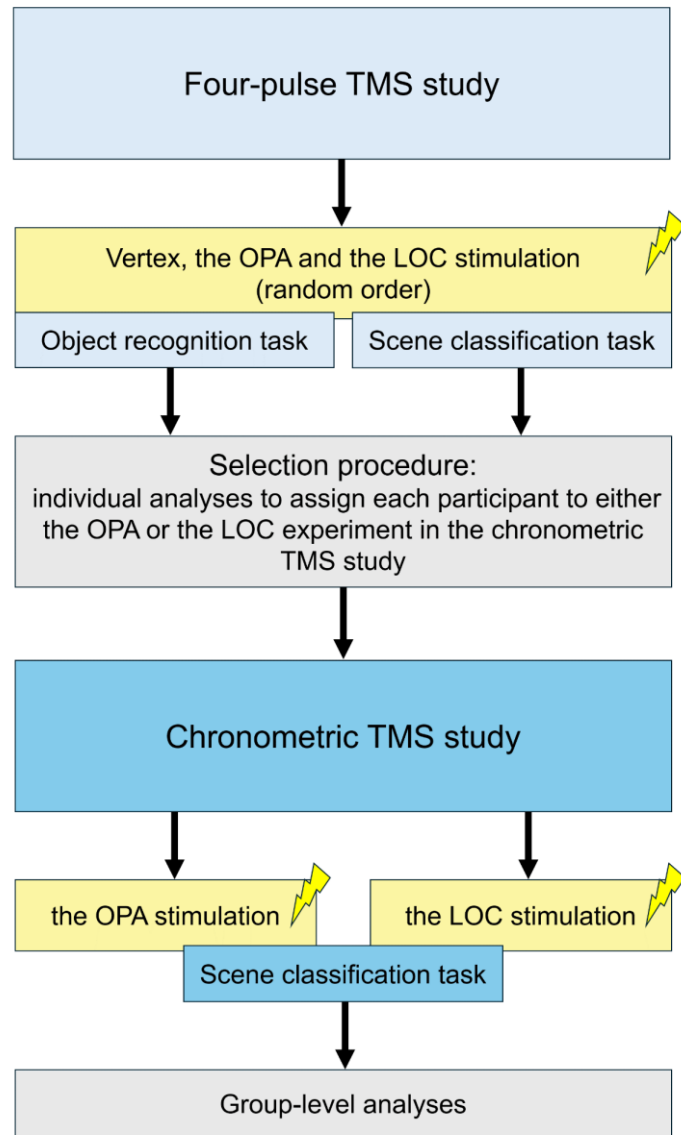


Figure 11. Schematic overview of the two-session TMS study. The four-pulse TMS study consisted of an object recognition task and a scene classification task (the order of tasks was counterbalanced across participants). The first TMS study was used here as a selection procedure: the data were analyzed individually to assign each participant to a stimulation site condition (either the OPA or the LOC) in the chronometric TMS study. During the chronometric TMS experiment, participants performed the same scene classification task as in the selection procedure. The data from the second session were analyzed on a group level.

This design allowed to investigate whether distinct neural mechanisms underlie scene recognition when it is based on its global features (*isolated scene* condition) and when it is based on a contained object (*scene-with-object* condition). It was hypothesized that while the recognition of intact scenes would rely on the OPA, object-based scene recognition would also be causally supported by the LOC. Thus, the LOC stimulation was expected to impair scene recognition in the scene-with-object, but not in the isolated scene condition.

Further, by manipulating the time point at which TMS was delivered, it could be examined at which stage – and, for object-based scenes, in which relative order – the OPA and the LOC causally contribute to scene recognition. Previous research indicated that high-level representations of both scenes (Cichy et al., 2017; Harel et al., 2016) and objects (Cichy et al., 2014; Isik et al., 2014; Kaiser et al., 2016) emerge around 200 ms after stimulus onset. Thus, for the isolated scene condition, it was expected that performance would be significantly worse in the middle relative to the early and late time windows during the OPA stimulation. In the scene-with-object condition, where the object mediates scene recognition, it was hypothesized that the LOC stimulation in the middle relative to the early and late time windows would lead to significantly worse performance. Additionally, it was predicted that performance in the scene-with-object condition during the OPA stimulation would also be worse in the middle compared to the early time window, reflecting the processing of ambiguous scene cues. Finally, according to the existing evidence, the integration of the scene and object information, mediated by the feedback connections between the OPA and the LOC, occurs at around 300 ms (Brandman & Peelen, 2023). Thus, the performance in the scene-with-object condition was expected to be even worse in the late as relative to the middle time window during the OPA stimulation. The summary of predictions for the chronometric TMS study is shown in Figure 12.

Such a pattern of results would be consistent with the previous study showing the causal involvement of the OPA in context-based object recognition (Wischnewski & Peelen, 2021b). It would also provide further support for the existence of a common predictive processing mechanism for bidirectional scene–object interactions (Peelen et al., 2024).

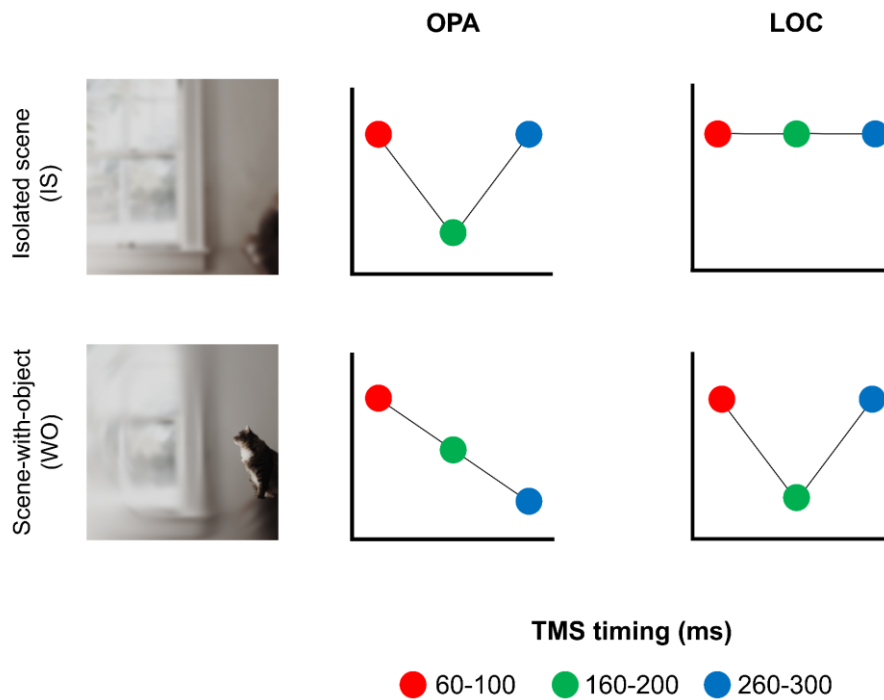


Figure 12. Time-specific predictions for the chronometric TMS experiment. It was hypothesized that the recognition of scenes in the isolated condition (top row) would be causally supported by the scene-selective OPA between 160–200 ms after stimulus onset (middle time point), with no involvement from the object-selective LOC in this process. For scenes in the with-object condition (bottom row), it was predicted that the OPA would again play a causal role at 160–200 ms, and crucially, the LOC would also contribute to object-based scene recognition at this time. Lastly, object-based scene recognition was hypothesized to occur in the OPA between 260–300 ms after stimulus onset (late time point).

2.2.2. Methods

2.2.2.1. Participants

The aim was to collect the pre-registered sample size of 48 participants (24 for the OPA and 24 for the LOC experiments). Sample sizes were chosen to match those of a previous similar TMS study (Wischnewski & Peelen, 2021b).

Healthy, right-handed volunteers aged 18 to 35 were invited to participate in the experiment. Before each experimental session, participants were informed about the procedures, completed a screening form, and provided written consent. Individuals were excluded from the study if they reported any of the following: pregnancy, previous neurosurgical treatments, a pacemaker or intracardiac lines, an implanted neuro-stimulator or medication infusion device, cochlear implants or other metal implants in the head or neck area, epilepsy or a history of cerebral seizures (including family history), tinnitus, use of CNS-acting medication, or sleep deprivation. Additionally, the use of any psychoactive substances or drugs within 48 hours preceding each session, and excessive consumption of alcohol or caffeine within 24 or 12 hours, respectively, were considered exclusion criteria.

To account for potential withdrawals or exclusions, 56 volunteers were initially recruited, exceeding the planned sample size of 48. Out of these 56 participants, 52 completed both experimental sessions. Four participants were identified as outliers and excluded from the analysis: based on either mean accuracy (2 participants) or mean reaction times (2 participants) falling below 2.5 standard deviations from the overall mean in the second session. After exclusion, 25 valid datasets remained for the OPA experiment, and 23 for the LOC experiment. To achieve the planned minimum sample size for the LOC experiment, an additional volunteer was tested, who completed only the second session and was pre-assigned to the LOC condition. Therefore, the data from a total of 49 participants (32 women; mean age \pm SD = 23.06 \pm 3.50), with 25 in the OPA condition (20 women; mean age \pm SD = 22.28 \pm 2.97) and 24 in the LOC condition (12 women; mean age \pm SD = 23.88 \pm 3.88), were included in the analysis.

All participants were reimbursed (15 euros per hour) or received course credits for their participation. The study procedures were approved by the Radboud Ethics Committee (NL72752.091.20).

2.2.2.2. Stimuli

The stimulus set created by Wischnewski & Peelen (2021b) was used for the object recognition task. The set consisted of 128 scene photographs, each depicting a single object belonging to one of the eight categories: airplane, bird, car, fish, human, mammal, ship, and train. The photographs were presented in two versions. In the first version, the intact object was cropped out of the scene and placed at its original location on a gray background (*isolated object* condition, IO). In the second version, the scene background was preserved, but the object was pixelated to remove its local features (*object-with-scene* condition, OS). For a detailed description of the stimuli, see Wischnewski & Peelen (2021b).

A newly created stimulus set, previously validated in two pilot behavioral studies (see Section 2.2.2.2.1. for details), was used for the scene classification task. This set consisted of 64 photographs of scenes (32 indoors and 32 outdoors) also presented in two versions. The scene was largely intact in the first version, but its main object was cropped out (*isolated scene* condition, IS). In the second version, the main object was preserved in its original form, but the scene context was heavily degraded (*scene-with-object* condition, WO).

Examples of the stimuli used in both procedures can be found in Figures 14A and 15A.

2.2.2.2.1. Creation of stimuli

The chronometric TMS experiment required a set of images depicting the same scenes in two versions: one with global features preserved and no object present (*isolated scene condition*, IS), and another with the object intact but global features degraded (*scene-with-object condition*, WO).

In the initial stage of stimulus preparation, scene images were selected and modified to ensure that they were easy to classify as indoor or outdoor when the object was presented within the scene, but difficult to classify when the scene and object were shown separately. Importantly, the image set was experimentally validated to confirm that the objects indeed disambiguated the scenes (see Section 2.2.2.2.1.1).

Before starting the TMS experiment, it was also crucial to ensure that baseline accuracy was not significantly different between isolated scene and scene-with-object conditions. Any main effect of stimulus condition could complicate evaluating the impact of stimulation on task outcomes. Therefore, a behavioral study was designed to compare

the performance across those two conditions and adjust the difficulty of the isolated scene condition, as needed, to match the scene-with-object one (see Section 2.2.2.2.1.2.).

The validation of the stimuli involved two pilot behavioral experiments. All experiments were approved by the Radboud University Ethics Committee (ECSW – 2022-079). Participants gave informed consent before the study and received financial compensation for their participation (4£).

2.2.2.2.1.1. Selection of images

The first step in the stimulus preparation involved selecting real-world photographs featuring a single main foreground object embedded in either an indoor or an outdoor scene context. The selection focused on objects that could plausibly appear in both indoor and outdoor settings, thereby ensuring that images could not be classified as indoor or outdoor based solely on the object. Additionally, care was taken to include an equal number of categories and exemplars representing both animate and inanimate objects. The scene images were sourced from Unsplash (unsplash.com) and categorized into 6 groups: 3 animate (cats, dogs, and people) and 3 inanimate (chairs, lamps, and plants). Further, all photographs were modified using Adobe Photoshop (21.1.3 Release), and each image was created in three versions.

In the first version, the object remained intact but was presented within a degraded context (*scene-with-object* condition; WO). In the second version, the object was removed from the scene, and the scene was heavily degraded (*scene-with-no-object* condition; NO). In the third version, the intact object was cropped out of the scene and placed at its original location on a gray background (*only-object* condition; OO). To degrade scene context, saturation was reduced, and Gaussian and radial blurs were applied. Objects were not blurred, but their saturation was adjusted to match the context, preventing a pop-out effect. The size of all the images in all their versions was standardized to 500 x 500 pixels using a custom-made R script. Example stimuli for each version are shown in Fig. 13A.

An online behavioral study was designed to determine whether the image manipulations were effective. Specifically, it was assessed whether participants would perform significantly better in the indoor–outdoor classification in the WO condition than in the NO and OO conditions. The performance in the NO and OO conditions was expected to be similar and close to the chance level in accuracy.

The initial set of stimuli included 96 photographs (16 per category: 8 indoor and 8 outdoor) in three versions, resulting in 288 images in total. To prevent participants from recognizing the degraded scenes alone based on having seen their version with an object, the stimulus set was divided into two halves: for each participant, half of the stimuli were used in the WO condition, and the other half was used in both the NO and the OO conditions. Thus, two sets of 144 stimuli (77 indoor, 77 outdoor) were created (A and B), and the assignment to a set was counterbalanced across participants.

The experimental procedure was programmed using PsychoPy (version 2023.1.3; Peirce et al., 2019) and administered online via Pavlovia (<https://pavlovia.org>).

The procedure comprised 288 trials (144 stimuli set was presented twice) and took around 20 minutes to complete. Each trial began with a display of a white fixation cross in the middle of the gray screen (1000 ms), followed by a briefly presented (50 ms) image and a blank screen (500 ms). The images were fit to a size of 10x10 cm, irrespective of the monitor's resolution. Participants were asked to decide whether it depicted an indoor or outdoor scene by pressing "f" or "j", respectively. The response window lasted maximally 2000 ms. The next trial sequence began immediately after the response was given or when the maximal time elapsed. Before starting the experiment, participants received detailed instructions and performed 12 practice trials, each followed by feedback about their performance. During the experiment, feedback was not given after each trial, but participants were informed about their accuracy and mean reaction time during breaks, which occurred after every 24 trials. There were 12 short breaks during the experiment.

The trial sequence is shown in Fig. 13B.

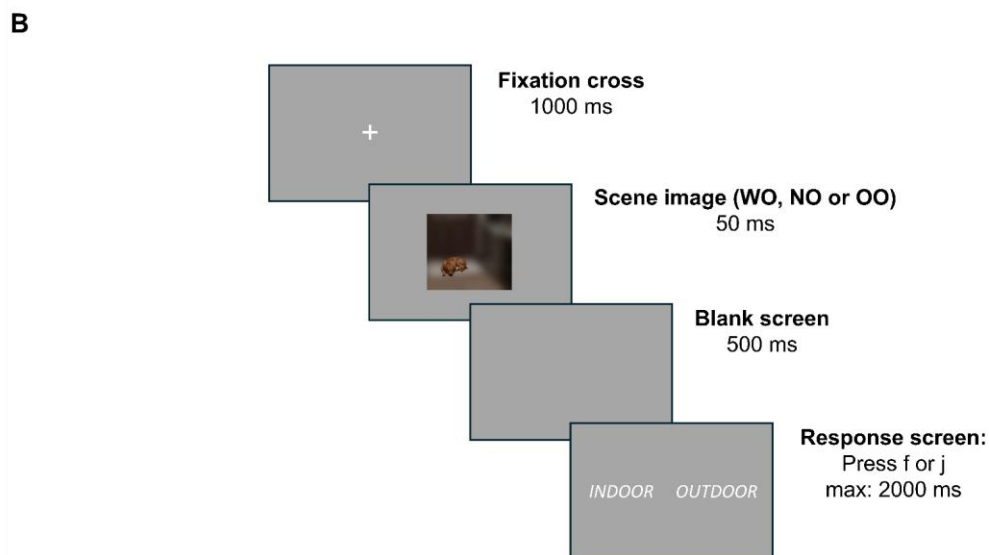
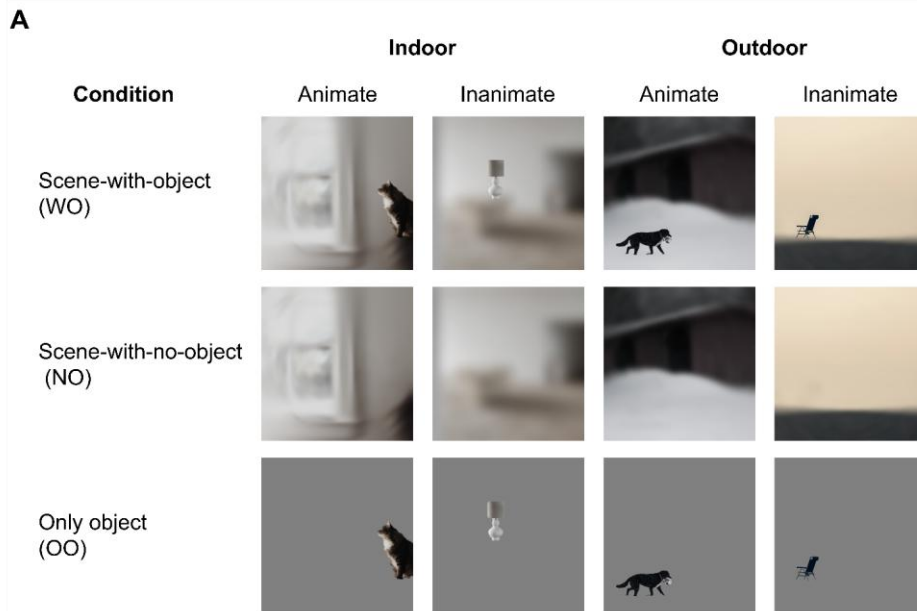


Figure 13. A) Examples of scene images used in the study in three versions: scene with-object (WO), scene-with-no-object (NO), and only object (OO). The objects belonged to either the animate (cats, dogs, people) or inanimate (chairs, lamps, plants) categories, and were balanced across indoor and outdoor scenes. B) Pilot behavioral study: a schematic depiction of a trial sequence. After a fixation cross (1000 ms), the stimulus was presented for 50 ms. Then a blank screen appeared (500 ms), followed by a 2000 ms response window (the indoor/outdoor labels are presented for illustrative purposes and were not displayed in the experiment). The next trial began after a response was given or the maximum time elapsed.

For the first pilot study, 32 participants were recruited via Prolific. Seven participants were excluded from the sample based on binomial tests, as their performance was indistinguishable from chance level. Thus, the final sample consisted of 25 participants (20 women, age range: 19-50, mean age \pm SD = 30.52 \pm 9.51; 12 for set A and 13 for set B). The repeated-measures ANOVA revealed a significant main effect of image condition on mean accuracy ($F(2,48) = 21.297$, $p < .001$, $\eta_p^2 = .470$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that mean accuracy was significantly higher for the WO condition (mean accuracy \pm SD = 0.669 \pm 0.075) than for the NO condition (mean accuracy \pm SD = 0.594 \pm 0.075; $t(24) = 5.047$, $p_{adj} < .001$, $d = 1.10$) and the OO condition (mean accuracy \pm SD = 0.579 \pm 0.056; $t(24) = 6.107$, $p_{adj} < .001$, $d = 1.31$). The NO and the OO conditions did not differ significantly in the mean accuracy ($t(24) = 1.060$, $p_{adj} = .883$, $d = 0.23$).

While the improvement in the scene classification performance due to the presence of objects was observed at the mean accuracy level, the analysis of accuracy at the individual image level revealed an imbalance between indoor (24) and outdoor (11) stimuli that had the expected pattern of results: WO accuracy higher than NO and OO. Therefore, a new set of stimuli was created, in which all 48 outdoor and 27 indoor images were replaced, and the pilot study was rerun using the same experimental procedure.

For the second experiment, 42 participants were recruited from Prolific. Eight participants were excluded from the sample based on binomial tests, as their performance was indistinguishable from chance level. Thus, the final sample consisted of 34 participants (19 women, age range: 18–62; mean age \pm SD = 35.27 \pm 13.75; 17 for set A and 17 for set B). The rm-ANOVA revealed a main effect of image condition on mean accuracy ($F(2,66) = 33.573$, $p < .001$, $\eta_p^2 = .504$), with performance in the WO condition (mean accuracy \pm SD = 0.725 \pm 0.088) being significantly better than in the NO (mean accuracy \pm SD = 0.650 \pm 0.090; $t(33) = 5.182$, $p_{adj} < .001$, $d = 0.95$) and in the OO (mean accuracy \pm SD = 0.607 \pm 0.057; $t(33) = 8.088$, $p_{adj} < .001$, $d = 1.48$) conditions. The analysis also revealed a significant difference between the NO and OO conditions, with significantly higher mean accuracy for the former ($t(24) = 2.907$, $p_{adj} = .015$, $d = 0.53$).

From all the stimuli tested in the pilot behavioral studies, those that either met or, in the case of a few outdoor stimuli, were close to meeting the criteria mentioned above (i.e., WO accuracy higher than NO and OO), were selected. The final set included 64 stimuli: 32 indoor (15 from the animate category: 6 cats, 6 dogs, and 3 people; and 17 from the inanimate

category: 6 chairs, 7 lamps, and 4 plants) and 32 outdoor (18 from the animate category; 5 cats, 7 dogs, and 6 people; and 14 from inanimate category: 5 chairs, 7 lamps, and 2 plants). Additionally, all stimuli were saved in three versions: WO, NO, and OO. Thus, the final set consisted of 192 images.

2.2.2.2.1.2. Pilot study for the TMS experiment

In the second stage of the stimuli preparation process, the isolated scene (IS) versions for all 64 images in the final set were created. This involved removing the objects and filling in the remaining space using a content-aware fill tool. The revised stimulus set, consisting of 128 images (64 isolated scenes (IS) and 64 scenes with objects (WO)), was then used for a second pilot behavioral study. Specifically, the goal was to achieve similar baseline accuracy levels for both image versions, preferably between 70% and 80%, as reported in previous TMS studies (Dilks et al., 2013; Wischniewski & Peelen, 2021b).

Similar to the previous behavioral pilot study, the experimental procedure was programmed using PsychoPy (version 2023.1.3; Peirce et al., 2019) and then administered online via Pavlovia (<https://pavlovia.org>). To prevent participants from recognizing scenes with objects based on having seen their intact versions, the stimulus set was divided into two halves: for each participant, half of the stimuli were used in their WO version, and the other half were used in their IS version. Thus, two sets (A and B), each consisting of 64 stimuli (32 WO and 32 IS; 16 indoor and 16 outdoor images), were created, with the assignment to a set counterbalanced across participants.

The procedure comprised 192 trials (one set was presented three times) and took around 13 minutes to complete. The trial sequence and timing mirrored those of the previous behavioral study (Figure 1B): each trial began with a 1000 ms fixation cross, followed by a 50 ms image presentation and 500 ms blank screen. Participants had up to 2000 ms to classify the scene as indoor or outdoor by pressing “f” or “j”. The next trial started immediately after a response or when the time elapsed. Instructions and 12 practice trials, each with feedback, were provided before the task. During the task, accuracy and reaction-time feedback were given every 24 trials. There were 8 short breaks during the experiment.

For the first pilot study, 30 participants were recruited via Prolific. One participant was excluded from the sample based on binomial tests due to performance indistinguishable from chance level. Thus, the final sample consisted of 29 participants (19 women, age range:

18– 47; mean age \pm SD = 33.32 ± 7.90 ; 14 for set A and 15 for set B). In line with the expectations, the paired-sample t-test comparison indicated that the mean accuracy for the IS condition (mean accuracy \pm SD = 0.796 ± 0.090) and the WO condition (mean accuracy \pm SD = 0.773 ± 0.093) was not significantly different ($t(28) = 1.654$, $p = .109$, $d = 0.307$). However, the analysis at the single-image level revealed substantial differences between versions for some stimuli. Therefore, for those images where the difference was apparent, the level of blur applied to the IS condition was adjusted, depending on the direction of the difference. Then the study was rerun with a new sample of participants.

Thirty-one participants were recruited via Prolific. One participant was excluded from the sample based on binomial tests, due to the performance being indistinguishable from the chance level. Thus, the final sample consisted of 30 participants (17 women, age range: 22– 69; mean age \pm SD = 36.5 ± 13.9 ; 15 for set A and 15 for set B). As in the previous pilot study, the paired sample t-test comparison revealed no significant difference ($t(29) = -1.243$, $p = .224$, $d = -0.227$) between the mean accuracy in the IS condition (mean accuracy \pm SD = 0.780 ± 0.094) and the mean accuracy in the WO condition (mean accuracy \pm SD = 0.791 ± 0.085).

To select the IS images for the final set, the mean accuracy of their two versions was compared with the mean accuracy of the WO condition averaged across two experiments. Then those IS images that were the closest in accuracy estimates to their WO versions were chosen. After the final selection was made, four stimuli (1 lamp outdoors, 1 chair outdoors, and 2 people indoors) were removed from the set in both their versions, as it was noticed that the mean accuracy estimates for their WO condition indicated consistent misclassification or chance-level performance across behavioral pilot studies. These stimuli were replaced with four newly created images of the same categories. Thus, the final set included 128 images (64 WO and 64 IS).

2.2.2.3. Apparatus

The experimental procedures were written in Matlab (R2020b) using the Psychophysics Toolbox Version 3 extensions (Brainard, 1997) and displayed on a BenQ Mobiuz (27") computer screen with 1920*1080 resolution and 120 Hz refresh rate.

Magnetic stimulation (MS) was applied via a C-B60 butterfly-shaped figure-of-8 coil with an outer diameter of 75 mm, which received input from a Magpro-X-100 magnetic

stimulator (MagVenture, Farum, Denmark). For 2 participants, an MC-B70 butterfly-shaped figure-of-8 coil with an outer diameter of 96 mm was used instead. The TMS coil was placed with the help of an infrared-based neuronavigation system (Localite, Bonn, Germany, and, for the first four participants, Visor 2, ANT Neuro, Hengelo, the Netherlands) using an individually adapted standard brain model. The size and shape of a participant's head were modeled by marking the nasal bridge,inion, left and right preauricular points, left and right canthi of the eyes, and a few hundred additional points across the scalp (just the first four points were used in the Visor2 system). Based on these points, a standard MNI brain was fit.

The stimulation location was identified through Talairach coordinates set in the neuronavigation system for the left OPA and the left LOC. The homologs of the right OPA (Julian et al., 2016) and the right LOC (Pitcher et al., 2009) coordinates were used for this study. Thus, the coordinates were - 34, -77, 21, and - 45, -74, 0 for the left OPA and the left LOC, respectively. Vertex was chosen as a control site and localized individually as the midpoint between the inion and nasion.

During the first experimental session, participants received a train of four pulses at 10 Hz over the left LOC, left OPA, and vertex. MS was delivered at the onset of each object or scene stimulus at 60% of the maximum stimulator output (MSO).

During the second experimental session, the intensity of stimulation was adjusted to 85% of the individual phosphene threshold (PT). PT was established by increasing stimulator output targeting the early visual cortex until 50% of the pulses resulted in the perception of a phosphene while participants fixated on a gray screen in a dimly lit room. On average, the PT was $49.20\% \pm 9.38\%$ of MSO, ranging between 35% and 89% of MSO. The TMS design mirrored the one used by Wischnewki & Peelen (2021b), with two MS pulses (biphasic, wavelength: 280 μ s) at 25 Hz, separated by 40 ms, applied, depending on the condition, at 3 different time points after stimulus onset. The pulses were applied over the left OPA or the left LOC.

2.2.2.4. Procedure

2.2.2.4.1. Four-pulse TMS study

The first TMS session consisted of an object-recognition task, based on Wischniewski and Peelen (2021b), and a two-alternative forced-choice scene classification task. The order of tasks was counterbalanced across participants. During both tasks, a sequence of four TMS pulses at 10Hz (0-100-200-300 ms starting from image onset) was delivered at 60% of maximum stimulator output over the left LOC, left OPA, and vertex in separate blocks. The entire session lasted approximately 120 minutes.

Before the experiment, the participants were introduced to the TMS technique. They were given a few single pulses at the back of the head to become familiar with the sensation induced by the stimulation. Further, they received detailed instructions and completed the 32 practice trials for each task. The stimuli used during the practice trials were not included in the experimental procedures but were identical to the experimental conditions.

2.2.2.4.1.1. Object recognition task

In the object recognition task, each trial sequence began with a fixation cross at the center of the screen (500 ms), followed by a brief presentation of an image (50 ms) and a blank screen (500 ms). The images were 400 x 300 pixels in size. Next, an untimed response window was shown, and once a response was given, a variable inter-trial interval ranging from 2 to 5 seconds followed. This extended interval was selected to prevent the coil from overheating during the trials and to minimize cumulative TMS effects across trials. Such long intervals were also used in previous online TMS experiments (e.g., Ganaden et al., 2013, Gandolfo and Downing, 2019, Gandolfo et al., 2024).

The goal of the participants in the object recognition task was to indicate whether an object depicted in the briefly displayed image belonged to one of eight categories (airplane, bird, car, fish, human, mammal, ship, or train) by pressing one of eight possible keys. A key appeared on the screen throughout the response window, indicating which number corresponded to each object category. Participants performed this task for clearly visible objects (*isolated object* condition) and degraded objects presented within a congruent scene context (*object-with-scene* condition). Four images per category were chosen randomly for each participant, resulting in 64 unique stimuli (32 per condition). To prevent familiarity

and carry-over effects, each image was shown to participants in only one condition (either the *isolated object* or the *object-with-scene*).

The chosen stimulus subset was repeated three times (once for each stimulation site), resulting in 192 trials. The task was divided into 6 blocks of 32 trials, each lasting approximately 3 min. There were breaks between blocks, lasting between 2 and 5 min, to prevent participants' fatigue and the coil from overheating. The order of stimuli within each block was randomized. The two blocks of each stimulation site followed one another, but the order of stimulation sites was randomly determined for each participant.

The trial sequence for the object recognition task is shown in Fig. 14B.

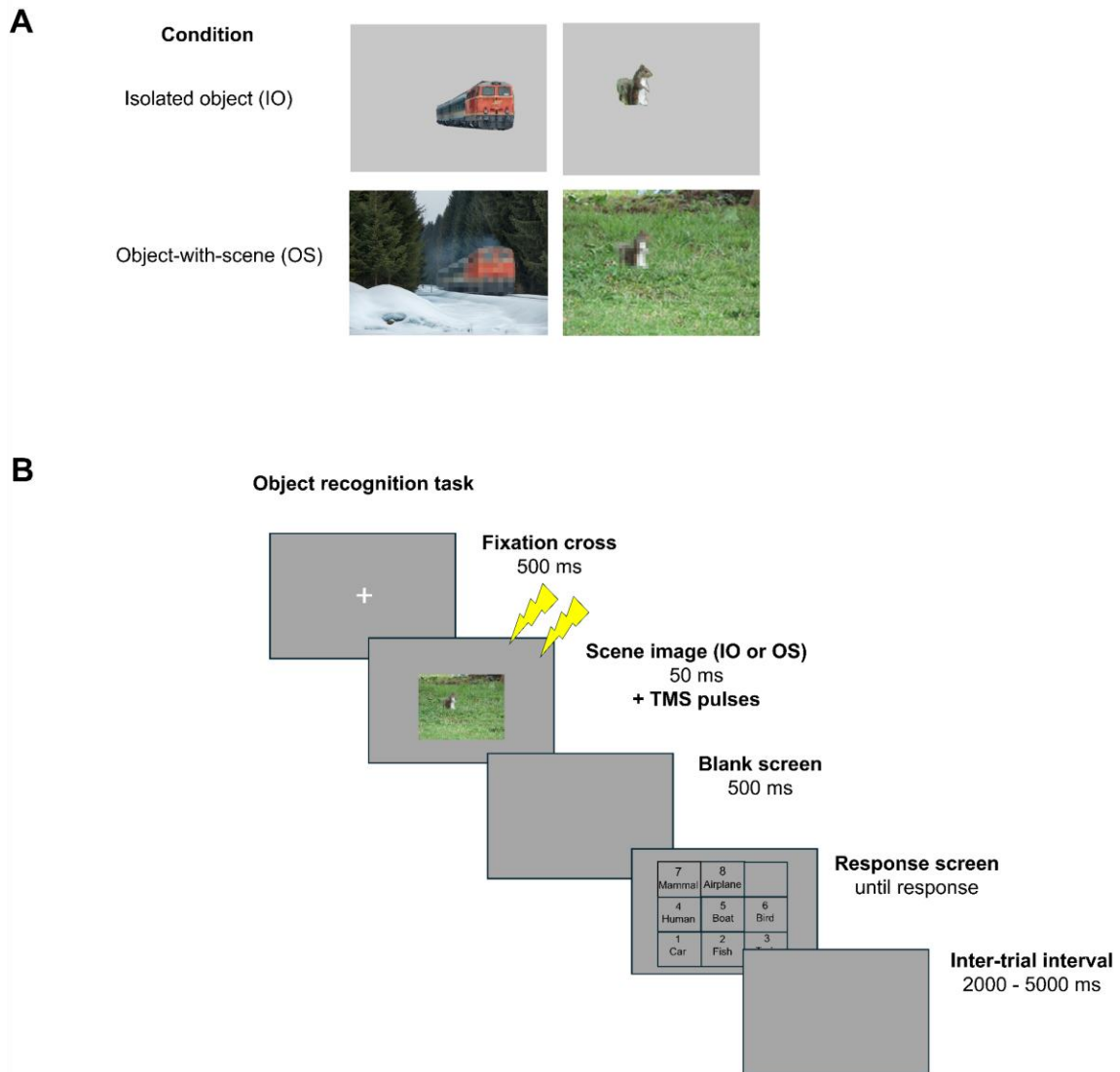


Figure 14. Object recognition task: A) Examples of stimuli used in the procedure. Images could be presented in two conditions: isolated object and object-with-scene B) Schematic trial of the object recognition task used in the selection procedure. During the task, four TMS pulses at 10Hz, starting at the stimulus onset, were delivered over the left LOC, left OPA, and vertex in separate blocks.

2.2.2.4.1.2. Scene classification task

In the scene classification task, each trial sequence began with a fixation cross (500 ms), followed by a brief presentation of an image (50 ms) and a blank screen (500 ms). The scene images were 500 x 500 pixels in size. Then, participants had a response period of two seconds. When the response was recorded or a maximum time elapsed, a variable inter-trial interval of 2-5 seconds started. As in the object recognition task, the inter-trial interval was relatively long to prevent the coil from overheating during the trials and to minimize cumulative TMS effects across trials.

The goal of the participants in the scene classification task was to decide whether the briefly displayed image depicted an indoor or outdoor scene by pressing “f” or “j”, respectively. Participants performed this task for clearly visible scenes (*isolated scene* condition) and degraded scenes presented with a congruent object (*scene-with-object* condition). For each participant, a list of 64 unique stimuli was generated (32 per condition). To prevent familiarity and carry-over effects, each image was shown to participants in only one condition (either the *isolated scene* or the *scene-with-object*). The chosen stimulus subset was repeated three times (once for each stimulation site), resulting in a total of 192 trials. The task was divided into 6 blocks of 32 trials, each lasting approximately 3 min. There were breaks between blocks, lasting between 2 and 5 min, to prevent participants’ fatigue and the overheating of the coil. The order of stimuli within each block was randomized. The two blocks of each stimulation site followed one another, but the order of stimulation sites was randomly determined for each participant.

The trial sequence for the scene classification task is shown in Fig. 15B.

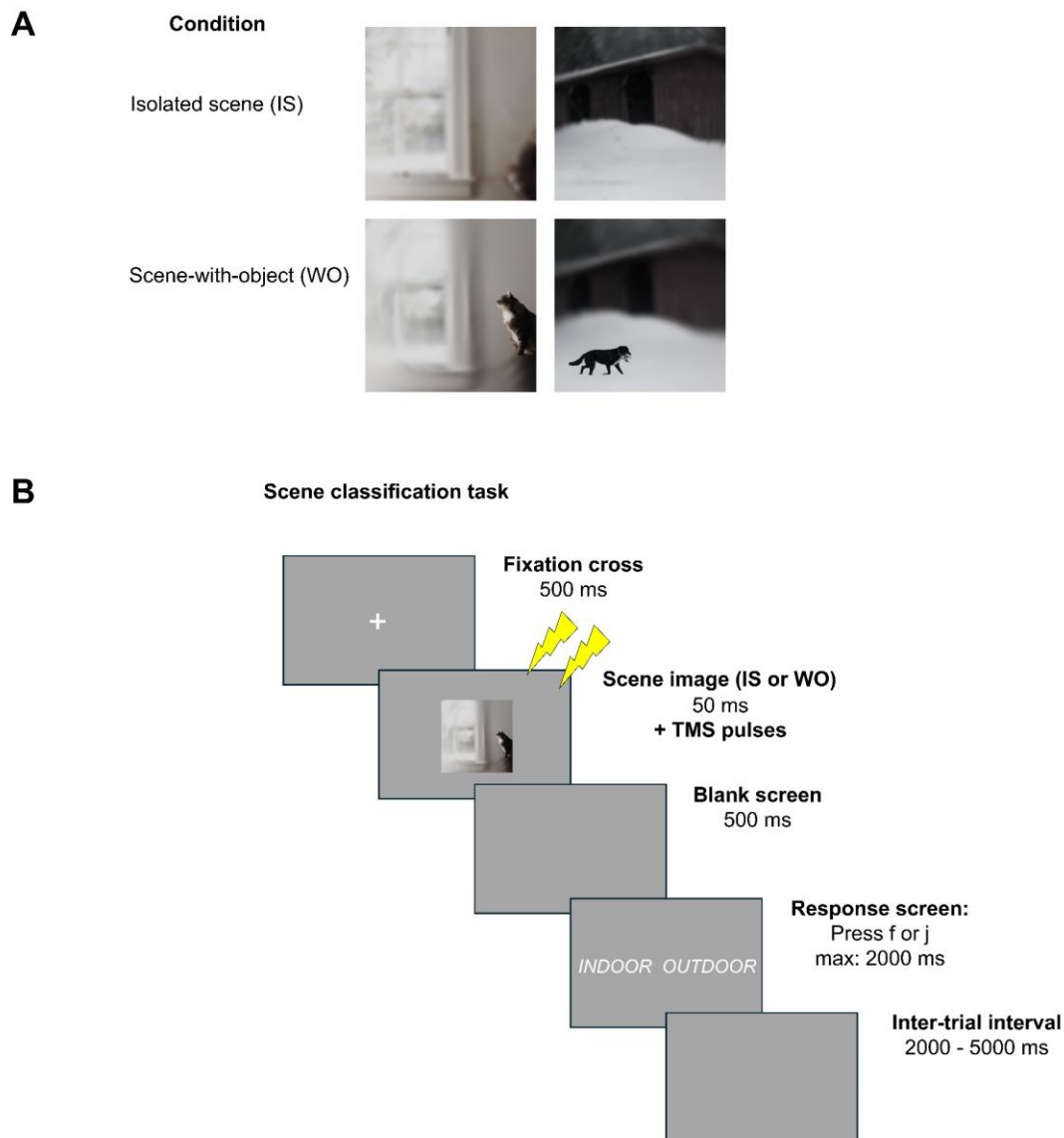


Figure 15. Scene classification task: A) Examples of stimuli used in the procedure. Images could be presented in two conditions: isolated scene and scene-with-object. B) Schematic trial of the scene classification task used in the selection procedure and chronometric TMS study. During the selection procedure, participants received four TMS pulses at 10Hz, starting at the stimulus onset, over the left LOC, the left OPA, and the vertex in separate blocks. In the chronometric TMS study, two TMS pulses (40ms apart) were administered in each trial at one of three time points (60–100ms, 160–200ms, 260–300ms) relative to stimulus onset over the left LOC or the left OPA. The indoor/outdoor labels in the response window are presented for illustrative purposes and were not displayed in the experiment.

2.2.2.4.1.2. Selection procedure

Group assignment (OPA vs. LOC) was based on comparisons of performance during the LOC versus vertex and the LOC versus the OPA stimulation in the object-recognition task, and during the OPA versus vertex and the OPA versus the LOC stimulation in the scene-classification task. Participants who demonstrated lower accuracy in the isolated object condition during the LOC relative to the vertex and the OPA stimulation were selected to receive the LOC stimulation in the chronometric TMS study. Participants who exhibited lower accuracy in the isolated scene condition during the OPA relative to the vertex and the LOC stimulation were assigned to receive the OPA stimulation in the chronometric TMS study. If the pattern of results based on accuracy was ambiguous, mean reaction times and LISAS (a measure combining speed and accuracy; see Section 2.2.5.2.) were also considered to support the decision process. Each participant was assigned to one group or the other: if a participant showed effects at both sites, the site with the larger and more consistent effects across all indices was chosen; if neither site showed a clear effect, the site with results trending in the expected direction was selected.

2.2.2.4.2. Chronometric TMS study

Participants were invited to the second TMS session after a minimum of 3 days following the completion of the first session (range: 3–11 days; mean interval \pm SD = 6.1 ± 1.5 days). The second session consisted solely of a scene classification task. The trial sequence was identical to that of the first session (see Figure 15B), and the same unique list of 64 scene images (half in the isolated scene condition and half in the with-object condition) was used for each participant.

In this session, however, each participant was stimulated at only one site throughout the experiment: either the left OPA or the left LOC, based on their results from the first session. Further, instead of the sequence of four pulses (0-100-200-300 ms starting from image onset), participants in the second session received double TMS pulses at one of three different time points: early (60 ms and 100 ms after stimulus onset), middle (160 ms and 200 ms after stimulus onset), or late (260 ms and 300 ms after stimulus onset). TMS pulses were delivered at the desired onset times (60 ms, 160 ms, and 260 ms) with a variability of 3-4 ms, and the stimulation strength was set to 85% of the individually determined phosphene threshold.

Every image was repeated twice for each stimulation onset, resulting in 384 trials. The trials were presented in random order. The task was divided into 12 blocks of 32 trials, each lasting approximately 3 minutes. After each block, participants were given a short break. The total duration of the experiment, including PT determination, preparation, and practice trials, was approximately 90 minutes.

2.2.2.5. Data analysis

All analyses were conducted using custom-made R scripts and JASP 0.18.3.0 software (JASP Team, 2023). The mean accuracy, reaction times, and LISAS for each experimental condition can be found in Table 12.

2.2.2.5.1. Four-pulse TMS study

The pre-registered analyses of the four-pulse TMS study are attached in Appendix A and will not be discussed in the present thesis.

2.2.2.5.2. Chronometric TMS study

For both OPA and LOC experiments, the pre-registered 3x2 repeated-measures ANOVAs were conducted with stimulation onset (early, middle, late) and stimulus condition (isolated scene, scene-with-object) as within-subject factors. Accuracy was considered the primary dependent variable. Reaction times of correct trials and linear integrated speed-accuracy scores (LISAS; Vandierendonck, 2017, 2018) were treated as secondary dependent variables and analyzed in separate ANOVAs.

The LISAS is a method to combine speed and accuracy data, which increases RTs as a function of error rates. It is defined as:

$$(2) \quad LISAS = RT_{ij} + PE_{ij} \times \frac{SRT_i}{SPE_i}$$

with RT_{ij} being the average reaction time in correct trials for participant i in condition j , PE_{ij} participant i 's error rate in condition j , SRT_i the standard deviation of response time (for participant i), and SPE_i the standard deviation of the proportion of errors of this participant. Lower LISAS indicate better performance.

Further, in line with the hypotheses, the pre-registered pairwise t-tests were computed for each experiment. For the OPA experiment, the performance in the isolated scene condition was compared between early TMS onset and middle TMS onset, and between middle TMS onset and late TMS onset. In the scene-with-object condition, the performance was compared between early TMS onset and late TMS onset. For the LOC experiment, the performance in the isolated scene condition was compared between all stimulation onsets. In the scene-with-object condition, the performance was compared between early TMS onset and middle TMS onset, and between middle TMS onset and late TMS onset.

Finally, to examine whether there was a differential effect of stimulation site depending on the stimulus condition, the mixed-design 2x2x3 ANOVAs on accuracy, reaction times, and LISAS were also conducted. The between-subjects factor was stimulation site (OPA, LOC), while the within-subjects factors were stimulus condition (isolated scene, scene-with-object) and stimulation onset (early, middle, late).

Values are reported as Mean \pm SD. Probability values were reported (p) for all statistical tests, and the standard .05 alpha level was used as a threshold for rejecting the null hypothesis. The Greenhouse-Geisser (GG) correction was applied when necessary to account for violations of sphericity.

2.2.3. Results

2.2.3.1. Within-subject analyses

2.2.3.1.1. OPA experiment

2.2.3.1.1.1. Accuracy

The analysis revealed no significant interaction between stimulus condition and stimulation onset ($F(1.989, 47.724) = 0.211, p = .809, \eta_p^2 = .009$). Further, there were no significant main effects of stimulus condition ($F(1,24) = 0.707, p = .409, \eta_p^2 = .029$) or stimulation onset ($F(1.606, 38.545) = 0.079, p = .886, \eta_p^2 = .003$).

The pre-registered pairwise t-tests revealed no significant differences between early and middle TMS onsets ($t(24) = 0, p = 1, d = 0$) or middle and late TMS onsets ($t(24) = 0.473, p = .641, d = 0.095$) in the isolated scene condition. Similarly, no significant difference was

observed between early and late TMS onsets in the scene-with-object condition ($t(24) = -0.730, p = .472, d = -0.146$).

2.2.3.1.1.2. Reaction times

The analysis revealed no significant interaction between stimulus condition and stimulation onset ($F(2,48) = 0.195, p = .823, \eta_p^2 = .008$) and no main effect of stimulus condition ($F(1,24) = 1.358, p = .255, \eta_p^2 = .054$). However, the main effect of stimulation onset was observed ($F(2,48) = 4.921, p = .011, \eta_p^2 = .170$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that reaction times were significantly faster when TMS pulses were applied in the early as relative to the middle time window (early: 0.639 ± 0.088 s, middle: 0.653 ± 0.099 s; $t(24) = -2.918, p_{adj} = .016, d = -0.144$). The difference between early and late time windows was on the level of statistical trend (early: 0.639 ± 0.088 s, late: 0.651 ± 0.097 s; $t(24) = -2.456, p_{adj} = .053, d = -0.121$) and there was no significant difference between middle and late time windows (middle: 0.653 ± 0.099 , late: 0.651 ± 0.097 ; $t(24) = 0.462, p_{adj} = 1, d = 0.023$).

The pre-registered pairwise t-tests revealed a significant difference between early and middle TMS onsets ($t(24) = -3.553, p = .002, d = -0.707$), and an insignificant difference between middle and late TMS onsets ($t(24) = 0.172, p = .865, d = 0.034$) in the isolated scene condition. No significant difference was observed between early and late TMS onsets in the scene-with-object condition ($t(24) = -1.294, p = .208, d = -0.259$).

Figure 16 depicts the main effect of stimulation onset.

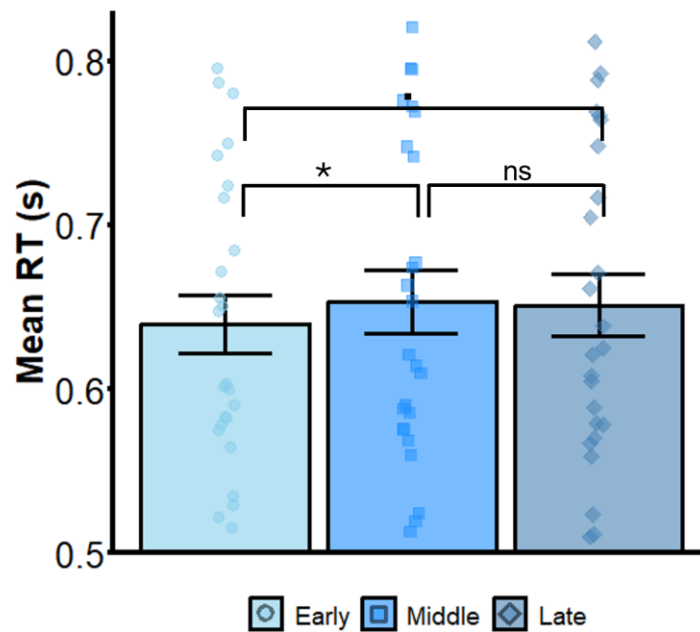


Figure 16. The main effect of stimulation onset on the mean reaction times in the OPA experiment. The reaction times were shortest when stimulation was applied in the early time window. The bars denote mean reaction times, the error bars indicate the standard error of the mean (SEM), the points represent individual participants, and the dot and the asterisk (·,*) indicate statistical trend ($.05 < p < .1$) and statistical significance ($p < .05$), respectively.

2.2.3.1.1.3. LISAS

The analysis revealed no significant interaction between stimulus condition and stimulation onset ($F(2,48) = 1.193, p = .312, \eta_p^2 = .047$). Further, as in accuracy analysis, there were no main effects of stimulus condition ($F(1,24) = 0.010, p = .921, \eta_p^2 = 0$) or stimulation onset ($F(2,48) = 1.740, p = .186, \eta_p^2 = .068$).

The pre-registered pairwise t-tests revealed a significant difference between early and middle TMS onsets ($t(24) = -2.633, p = .015, d = -0.527$), and an insignificant difference between middle and late TMS onsets ($t(24) = 0.524, p = .605, d = 0.105$) in the isolated scene condition. No significant difference was observed between early and late TMS onsets in the scene-with-object condition ($t(24) = -0.436, p = .666, d = -0.087$).

2.2.3.1.2. LOC experiment

2.2.3.1.2.1. Accuracy

The analysis revealed no significant interaction between stimulus condition and stimulation onset ($F(2,46) = 0.794$, $p = .458$, $\eta_p^2 = .033$). However, the main effects of stimulus condition ($F(1,23) = 7.895$, $p = .010$, $\eta_p^2 = .256$) and stimulation onset were observed ($F(2,46) = 4.517$, $p = .016$, $\eta_p^2 = .164$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that scenes with objects were classified significantly worse than isolated scenes (WO: 0.883 ± 0.068 , IS: 0.909 ± 0.044 ; $t(23) = -2.810$, $p_{adj} = .010$, $d = -0.449$). Further, accuracy in the late time window was significantly lower as relative to the accuracy in the middle time window (late: 0.886 ± 0.050 , middle: 0.902 ± 0.052 ; $t(23) = -2.782$, $p_{adj} = .023$, $d = 0.272$) The difference between late and early time windows was on the level of statistical trend (late: 0.886 ± 0.050 , early: 0.899 ± 0.047 ; $t(23) = -2.376$, $p_{adj} = .065$, $d = -0.232$) and the difference between early and middle time windows was not significant (early: 0.899 ± 0.047 , middle: 0.902 ± 0.052 ; $t(23) = -0.406$, $p_{adj} = 1$, $d = -0.040$).

The pre-registered pairwise t-tests revealed no significant differences between early and middle TMS onsets ($t(23) = -0.232$, $p = .819$, $d = -0.047$), middle and late TMS onsets ($t(23) = 1.152$, $p = .261$, $d = 0.235$), or early and late TMS onsets ($t(23) = 0.805$, $p = .429$, $d = 0.164$) in the isolated scene condition. In the scene-with-object condition, there was an insignificant difference between early and middle TMS onsets ($t(23) = -0.266$, $p = .793$, $d = -0.054$) and a significant difference between middle and late TMS onsets ($t(23) = 2.473$, $p = .021$, $d = 0.505$).

Figure 17 depicts the main effect of stimulus condition (A) and stimulation onset (B).

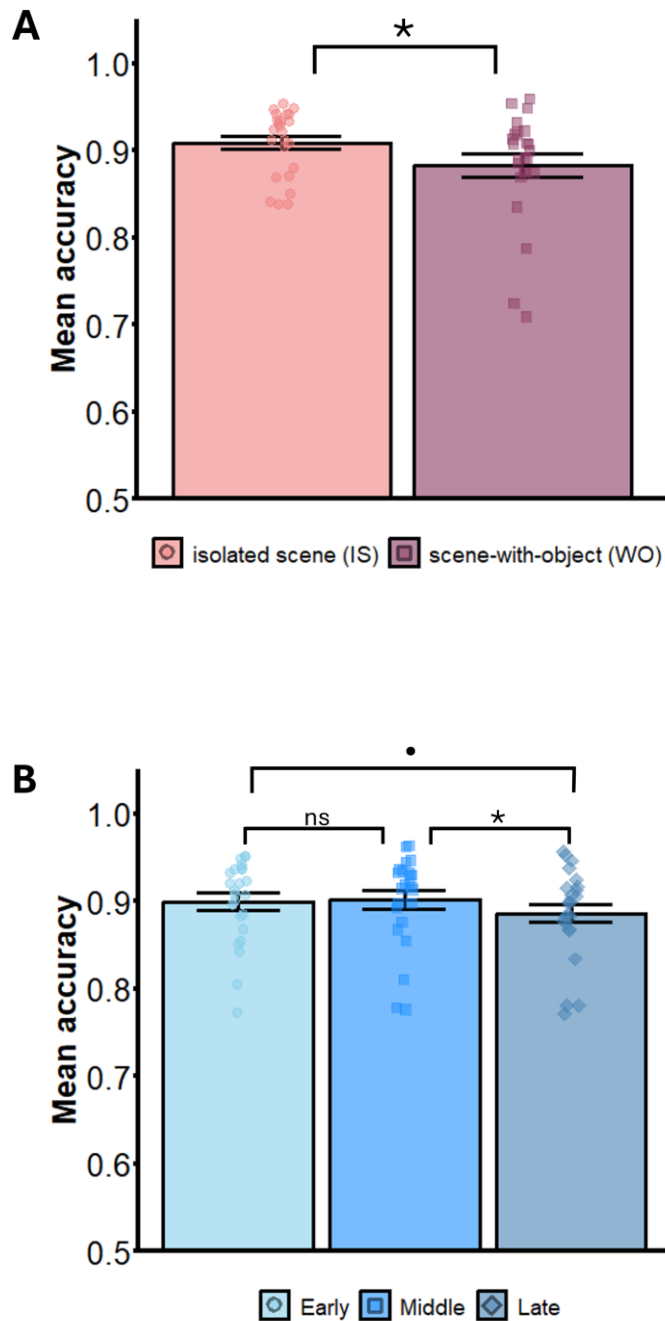


Figure 17. A) The main effect of stimulus condition: scenes in the with-object (WO) condition were classified significantly worse than scenes in the isolated (IS) condition B) The main effect of stimulation onset: accuracy was lowest in the late time window. The bars denote mean accuracy, the error bars indicate the standard error of the mean (SEM), the points represent individual participants, and the dot and the asterisk (·,*) represent statistical trend ($.05 < p < .1$) and statistical significance ($p < .05$), respectively.

2.2.3.1.2.2. Reaction times

The analysis revealed no significant interaction between stimulus condition and stimulation onset ($F(2,46) = 0.368, p = .694, \eta_p^2 = .016$), and no main effect of stimulation onset ($F(2,46) = 0.267, p = .767, \eta_p^2 = .011$). There was, however, a main effect of the stimulus condition ($F(1,23) = 10.553, p = .004, \eta_p^2 = .315$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that scenes with objects were classified significantly slower than isolated scenes (WO: 0.702 ± 0.110 s, IS: 0.681 ± 0.109 s; $t(23) = 3.249, p_{adj} = .004, d = 0.188$).

The pre-registered pairwise t-tests revealed no significant differences between early and middle TMS onsets ($t(23) = 0.703, p = .489, d = 0.144$), middle and late TMS onsets ($t(23) = -0.889, p = .383, d = -0.182$), or early and late TMS onsets ($t(23) = -0.077, p = .939, d = -0.016$) in the isolated scene condition. Similarly, no significant differences were observed between early and middle ($t(23) = -0.207, p = .838, d = -0.042$) or middle and late ($t(23) = -0.432, p = .670, d = -0.088$) TMS onsets in the scene-with-object condition.

2.2.3.1.2.3. LISAS

The analysis revealed no significant interaction between stimulus condition and stimulation onset ($F(2,46) = 0.521, p = .598, \eta_p^2 = .022$), and no main effect of stimulation onset ($F(2,46) = 2.512, p = .092, \eta_p^2 = .098$). There was, however, a main effect of the stimulus condition ($F(1,23) = 14.520, p < .001, \eta_p^2 = 0.387$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that scenes with objects were classified significantly worse than isolated scenes (WO: 0.796 ± 0.130 , IS: 0.765 ± 0.128 ; $t(23) = 3.811, p_{adj} < .001, d = 0.240$).

The pre-registered pairwise t-tests revealed no significant differences between early and middle TMS onsets ($t(23) = 0.494, p = .626, d = 0.101$), middle and late TMS onsets ($t(23) = -1.160, p = .258, d = -0.237$), or early and late TMS onsets ($t(23) = -0.631, p = .534, d = -0.129$) in the isolated scene condition. Similarly, no significant differences were observed between early and middle ($t(23) = -0.487, p = .631, d = -0.099$) or middle and late ($t(23) = -1.658, p = .111, d = -.338$) TMS onsets in the scene-with-object condition.

2.2.3.3. Between-subject analysis

2.2.3.3.1. Accuracy

The analysis revealed no significant effect of stimulus condition, which indicated that scenes were well equated in terms of difficulty (IS: 0.905 ± 0.044 ; WO: 0.897 ± 0.067 ; $F(1,47) = 1.427$, $p = .238$, $\eta_p^2 = .029$). The interaction between stimulation site, stimulus condition, and stimulation onset was insignificant ($F(2,94) = 0.866$, $p = .424$, $\eta_p^2 = .018$). A significant interaction between stimulation site and stimulus condition was found ($F(1,47) = 6.071$, $p = .017$, $\eta_p^2 = .114$). The analysis of simple main effects indicated that scenes in the with-object condition were classified significantly worse than scenes in the isolated condition during the LOC stimulation (WO: 0.883 ± 0.068 , IS: 0.909 ± 0.044 ; $F(1,23) = 7.895$, $p = .010$, $\eta_p^2 = .256$). During the OPA stimulation, there was no significant difference between stimulus conditions (WO: 0.910 ± 0.063 , IS: 0.901 ± 0.044 ; $F(1,24) = 0.707$, $p = .409$, $\eta_p^2 = .029$).

Figure 18 depicts the interaction between stimulation site and stimulus condition. Table 9 summarizes the ANOVA results.

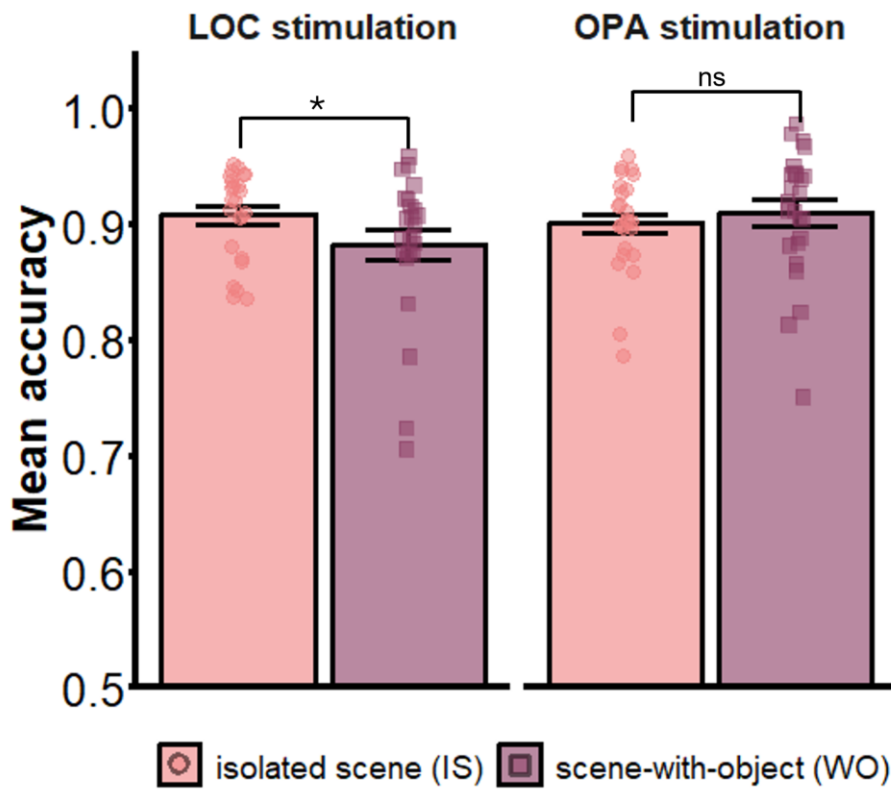


Figure 18. Accuracy in the scene-with-object condition was significantly worse than in the isolated scene condition during the LOC stimulation, but not during the OPA stimulation. The bars denote mean accuracy, the error bars indicate the standard error of the mean (SEM), the points represent individual participants, and the asterisk (*) indicates statistical significance ($p < .05$).

Table 9. A summary of a mixed-design 2x2x3 ANOVA, with stimulation site (OPA, LOC) as between-subject factor, and stimulus condition (IS: isolated scene, WO: scene-with-object) and stimulation onset (early, middle, late) as within-subject factors. Accuracy is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
site	1	47	0.621	.435	.013
condition	1	47	1.427	.238	.029
site * condition	1	47	6.071	.017	.114
onset	2	94	1.871	.160	.038
site * onset	2	94	2.463	.091	.050
condition * onset	2	94	0.155	.857	.003
site * condition * onset	2	94	0.866	.424	.018

2.2.3.3.2. Reaction times

The analysis revealed no significant interaction between stimulus condition, stimulation site, and stimulation onset ($F(2,94) = 0.406, p = .667, \eta_p^2 = .009$), but a significant main effect of stimulus condition was found ($F(1,47) = 9.213, p = .004, \eta_p^2 = .164$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that scenes with objects were classified significantly slower than isolated scenes (WO: 0.676 ± 0.107 s; IS 0.662 ± 0.101 s; $t(47) = 3.035, p_{adj} = .004, d = 0.141$).

A summary of the ANOVA results can be found in Table 10.

Table 10. A summary of a mixed-design 2x2x3 ANOVA, with stimulation site (OPA, LOC) as a between-subject factor, and stimulus condition (isolated scene, scene-with-object) and stimulation onset (early, middle, late) as within-subject factors. Reaction times are a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
site	1	47	2.268	.139	.046
condition	1	47	9.213	.004	.164
site * condition	1	47	1.716	.197	.035
onset	2	94	2.155	.122	.044
site * onset	2	94	2.274	.109	.046
condition * onset	2	94	0.161	.852	.003
site * condition * onset	2	94	0.406	.667	.009

2.2.3.3.3. LISAS

The analysis revealed no significant interaction between stimulus condition, stimulation site, and stimulation onset ($F(2,94) = 1.400, p = .252, \eta_p^2 = .029$) and a significant interaction between stimulation site and stimulus condition ($F(1,47) = 4.395, p = .041, \eta_p^2 = .086$). The analysis of simple main effects indicated that while performance for isolated scenes did not differ significantly between the LOC and the OPA stimulation (LOC: 0.765 ± 0.128 , OPA: 0.712 ± 0.125 ; $F(1,47) = 2.285, p = .137, \eta_p^2 = .046$), scenes with objects were classified significantly worse during the LOC as relative to the OPA stimulation (LOC: 0.796 ± 0.130 , OPA: 0.712 ± 0.141 ; $F(1,47) = 4.828, p = .033, \eta_p^2 = .093$). Further, performance for scenes in the with-object condition was significantly worse as compared to scenes in the isolated condition during the LOC stimulation (WO: 0.796 ± 0.130 , IS: 0.765 ± 0.128 ; $F(1,23) = 14.520, p < .001, \eta_p^2 = .387$). During the OPA stimulation, there was no significant difference in performance between stimulus conditions (WO: 0.712 ± 0.14 , IS: 0.712 ± 0.125 ; $F(1,24) = 0.010, p = .921, \eta_p^2 = 0$).

Table 11 summarizes the ANOVA results.

Table 11. A summary of a mixed-design 2x2x3 ANOVA, with stimulation site (OPA, LOC) as a between-subject factor, and stimulus condition (isolated scene, scene-with-object) and stimulation onset (early, middle, late) as within-subject factors. LISAS is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
site	1	47	3.615	.063	.071
condition	1	47	5.100	.029	.098
site * condition	1	47	4.395	.041	.086
onset	2	94	2.903	.060	.058
site * onset	2	94	1.317	.273	.027
condition * onset	2	94	0.079	.924	.002
site * condition * onset	2	94	1.400	.252	.029

Table 12. Mean accuracy, mean reaction times, and mean LISAS for each condition in the OPA and the LOC experiment.

Site	Condition	Onset	Accuracy (M \pm SD)	RT (s) (M \pm SD)	LISAS (M \pm SD)	N
LOC	IS	early	0.910 \pm 0.045	0.682 \pm 0.103	0.765 \pm 0.121	24
LOC	IS	middle	0.912 \pm 0.043	0.677 \pm 0.112	0.760 \pm 0.132	24
LOC	IS	late	0.904 \pm 0.044	0.683 \pm 0.115	0.771 \pm 0.135	24
LOC	WO	early	0.889 \pm 0.068	0.700 \pm 0.107	0.787 \pm 0.123	24
LOC	WO	middle	0.891 \pm 0.064	0.701 \pm 0.111	0.793 \pm 0.137	24
LOC	WO	late	0.868 \pm 0.073	0.704 \pm 0.116	0.809 \pm 0.136	24
OPA	IS	early	0.902 \pm 0.045	0.634 \pm 0.085	0.699 \pm 0.112	25
OPA	IS	middle	0.902 \pm 0.044	0.649 \pm 0.095	0.719 \pm 0.136	25
OPA	IS	late	0.900 \pm 0.046	0.648 \pm 0.096	0.715 \pm 0.129	25
OPA	WO	early	0.907 \pm 0.060	0.645 \pm 0.092	0.709 \pm 0.131	25
OPA	WO	middle	0.911 \pm 0.065	0.658 \pm 0.105	0.714 \pm 0.152	25
OPA	WO	late	0.913 \pm 0.067	0.654 \pm 0.101	0.713 \pm 0.146	25

2.2.4. Discussion

The present study examined whether object representations in the object-selective cortex (LOC) play a causal role in disambiguating scenes, and how the temporal contribution of the LOC relates to that of the scene-selective OPA. The results showed that recognition accuracy for scenes disambiguated by objects was significantly reduced during the LOC stimulation, but not during the OPA stimulation. This pattern was further supported by the LISAS analysis, which likewise revealed poorer performance for object-based scenes during the LOC relative to the OPA stimulation. The convergence of accuracy and LISAS results rules out a speed–accuracy trade-off as an explanation for the observed effect.

Thus, the present experiment provides the first TMS evidence that the LOC is selectively and causally involved in scene recognition when it is facilitated by object information. This result is consistent with prior work demonstrating the causal involvement of the OPA in context-based object recognition (Wischnewski & Peelen, 2021b). Critically, because the object category in the present study was not diagnostic of the scene category, the observed effect is likely to reflect sensitivity to spatial properties of objects – such as their position, size, and viewpoint – that can be processed within the LOC and subsequently inform the scene-selective cortex, aiding in the disambiguation of scene layout. This interpretation aligns with previous literature showing that the LOC is sensitive to such object attributes (Eger et al., 2008; Sayres & Grill-Spector, 2008; Andresen et al., 2009).

Nonetheless, based on the present experiment, no firm conclusions can be drawn regarding the involvement of the OPA or the precise timing of the LOC effect. In particular, it remains unclear whether the temporal dynamics of object-to-scene influences mirror those of scene-to-object influences (Wischnewski & Peelen, 2021b), leaving the question of a common predictive processing mechanism for bidirectional scene–object interactions unresolved (Peelen et al., 2024).

The lack of a significant effect of OPA stimulation on scene recognition is inconsistent with previous TMS findings (Dilks et al., 2013; Wischnewski & Peelen, 2021a). Several factors may account for this discrepancy. First of all, recent evidence indicates that the effect of object-based scene recognition is observed mostly in the left hemisphere (Brandman & Peelen, 2019, 2023). Consequently, in this study, in contrast to previous experiments (Dilks et al., 2013; Wischnewski & Peelen, 2021a; Gandolfo & Downing, 2019), the left, rather than the right, OPA was selected for stimulation. Further, to localize the left OPA and the left LOC the

homologues of their coordinates defined in the right hemisphere were used. Thus, it is possible that stimulation sites were inaccurate due to interhemispheric differences in the position of target areas. To ensure that the localization of the left OPA and the left LOC is precise, it might be beneficial to first identify fMRI coordinates of these areas for each participant and then use them for neuronavigation.

Notably, there is evidence that inhibiting the LOC in one hemisphere impairs object processing in the contralateral LOC, whereas such interhemispheric dependence does not appear to hold for the OPA (Rafique et al., 2015). Considering these findings, it is possible that stimulation of the left LOC in the present study resulted in a more complete suppression of object processing, while the right OPA may have compensated for the left OPA during stimulation. Such differences in the interhemispheric connectivity could therefore contribute to the absence of the expected OPA effects. However, it is worth mentioning that the effect of the left OPA stimulation on scene categorization was observed in the study by Ganaden et al. (2013), which localized the target area in 8 participants using either individual fMRI coordinates or anatomical data.

Finally, the scene tasks used in the previous TMS studies involved the assessment of scene orientation (Gandolfo & Downing, 2019), the spatial memory for navigationally relevant objects (Julian et al., 2016), the classification of scenes into four predefined categories (Dilks et al., 2013, Wischnewski & Peelen, 2021a) and the categorization of scenes as natural or non-natural (Ganaden et al., 2013). Thus, they all differ from the indoor-outdoor judgment required in this experiment. Existing research suggests that both the OPA and the PPA are involved in representing the spatial layout of the scene (Epstein & Baker, 2019). However, while the OPA is hypothesized to represent scenes in terms of their navigational affordances (Bonner & Epstein 2017, Bonner & Epstein, 2018, Persichetti & Dilks, 2018), the PPA is thought to represent scene in terms of categorical distinctions (Dillon et al., 2018; Park et al., 2011; Persichetti & Dilks, 2018) and was also shown to integrate object the most in its representation (Aminoff & Durham, 2023). Therefore, it is not ruled out that solving the indoor-outdoor task in this study relied more on the PPA than OPA, and, for object-based scenes, more on feedback connections between the LOC and the PPA than between the LOC and the OPA, as initially hypothesized. In such a scenario, the input from the LOC concerning the object would be used by the PPA to disambiguate scene spatial layout. This hypothesis might be worth testing in the follow-up studies. However, as the TMS stimulation of the PPA

is not currently feasible due to its anatomical location, putting this hypothesis to the test would require using more experimental methods, such as transcranial ultrasound stimulation (TUS).

When comparing the present study to previous experiments, it is also important to note that although the same TMS stimulator (MagPro X100) was used, the coil differed: most participants in the current study were stimulated with the C-B60 coil, whereas Wischnewski and Peelen (2021a, 2021b) used the Cool-B65 coil in their investigations of the double dissociation between the OPA and LOC and of context-based object recognition. Several studies have shown that both the TMS device and coil can influence stimulation outcomes (Corp et al., 2021; Deng et al., 2013; Wang et al., 2024). In particular, Wang et al. (2024) reported significant differences in resting motor thresholds when measured with the C-B60 versus the Cool-B65 coil. However, because the same coil (C-B60 or MC-B70) was used both to estimate each participant's PT and to deliver stimulation, it is unlikely that coil differences relative to previous studies account for the absence of the OPA or timing effects in the current experiment. Nonetheless, future studies may wish to adopt the Cool-B65 coil, as in Wischnewski and Peelen (2021a, 2021b), given its active cooling system, which reduces overheating.

Another factor that is widely discussed in the TMS literature is the impact of inter-individual variability on stimulation effects. A range of subject-specific characteristics related to brain anatomy, such as gyral folding and conductivity anisotropy, as well as non-brain anatomy, such as brain–scalp distance and cerebrospinal fluid (CSF) thickness, have been shown to affect the intensity and spread of stimulation (Lee et al., 2018; Opitz et al., 2013). Recent evidence further suggests that age can significantly modulate stimulation effects: corticospinal excitability appears to be lower in individuals under 20 years of age (an initial phase of hypoexcitability) and to increase until approximately 25–35 years (Corp et al., 2021). There are also indications that gender may influence TMS outcomes (Hanlon & McCalley, 2022; Shibuya et al., 2016). To minimise the impact of such inter-individual factors on the present results, participants were assigned to experimental conditions using a TMS-based selection procedure that has previously been shown to be effective for this purpose (Wischnewski & Peelen, 2021b). In this procedure, participants received stimulation at a fixed intensity of 60% of the maximum stimulator output. However, this fixed intensity may have been suboptimal for some individuals, particularly given the age range in the sample (18–33 years). Thus, it cannot be ruled out that inter-individual variability reduced the sensitivity of the selection procedure. In the subsequent chronometric TMS experiment, however, stimulation

intensity was individually adjusted based on each participant's phosphene threshold. Therefore, it seems unlikely that individual differences in susceptibility to stimulation alone can account for the absence of the expected effects in the present experiment.

It is also worth noting that the mean accuracy in the scene classification task during the TMS procedure was higher than in two preceding pilot behavioral studies (0.901 vs 785 / 0.786) and higher than in the object recognition task used to investigate context-based object recognition (Wischnewski & Peelen, 2021b). Importantly, however, as in the pilot experiments, no significant differences between stimulus conditions were observed in the main study. Given that the effect of the left OPA stimulation was reported in the previous study with a comparable level of task difficulty (Ganaden et al., 2013), and that the stimulation effect for the LOC was detectable in the present study, it is unlikely that high overall accuracy alone accounts for the absence of the OPA effect. Nonetheless, very high performance may, in general, make stimulation effects more difficult to detect. Considering the differences between online and laboratory testing environments, future studies would likely benefit from piloting the procedure in controlled laboratory settings.

To conclude, the present study provides the first causal evidence that the LOC contributes to scene classification, demonstrating that activity in the object-selective cortex can support scene recognition when it is facilitated by object information. Follow-up TMS studies are needed to clarify the role of the OPA in this process and to determine the temporal window in which the LOC exerts its influence. Using individual fMRI coordinates to localise the target regions is recommended to increase the likelihood of detecting these effects.

2.3. Study 3: object-based facilitation of scene recognition in humans and a computer model of human vision

2.3.1. Research questions and hypotheses

While most studies have focused on scene-to-object influences, a growing body of evidence shows that these interactions are reciprocal, with objects also shaping scene recognition (e.g., Furtak et al., 2022; Joubert et al., 2007; Leroy et al., 2020). Study 1 provides converging evidence for this bidirectionality. Further, recent research showed that not only can scene context facilitate ambiguous object recognition (context-based object recognition; Brandman & Peelen, 2017), but also objects can facilitate ambiguous scene recognition – a phenomenon termed object-based scene recognition (Brandman & Peelen, 2019, 2023). In Study 2, this effect was replicated with a new stimulus set. Further, the causal contribution of the object-selective LOC to the scene disambiguation process was demonstrated.

Building on these findings, the present study examined whether the facilitatory effect of objects on scene recognition depends on the presence of a coherent scene layout. Two hypotheses were considered: (i) that object-based facilitation requires intact global structure, such that disrupting the spatial layout of the scene should reduce or eliminate the effect; or (ii) that facilitation may persist even when the scene is degraded, provided that low-level visual statistics are preserved. To evaluate these accounts, participants were asked to perform the scene-categorization task with images presented in three conditions: an object on a neutral background (the only-object condition), an object placed in a scrambled scene (the scrambled-scene-with-object condition), and an object placed in an ambiguous scene (the scene-with-object condition). Scrambled backgrounds were created by applying phase scrambling (Thomson, 1999) to scene images from the stimulus set developed and validated in Study 2 (Section 2.2.2.2). Phase scrambling renders the spatial layout unrecognizable while preserving low-level properties, such as colour, luminance, and the overall distribution of spatial frequencies (Oliva & Schyns, 1997, 2000).

In addition, from the neuroconnectionist perspective, the study examined whether phase scrambling differentially affects scene recognition in humans and in a scene-trained deep neural network. To this end, human performance was compared with that of GoogLeNet, a 22-layer Inception-based convolutional model trained on the Places365 dataset to recognise 365 scene categories (Zhou et al., 2017; see Appendix, Table B1). The network analyzes

images at multiple spatial scales and then aggregates this information to produce a scene label (Szegedy et al., 2015). In contrast to the human visual system, which is characterised by functionally specialised cortical streams and extensive recurrent (feedback) connectivity, GoogLeNet implements a single feedforward pathway with no separate “object” and “scene” routes. Given these differences, the study tested whether the network would process ambiguous and scrambled scenes with objects differently, and whether this pattern would resemble human performance.

2.3.2. Methods

2.3.2.1. Participants

Fifty-eight participants were recruited via Prolific. Seven participants were excluded from the sample based on binomial tests, as their performance was indistinguishable from chance level. Thus, the final sample consisted of 51 participants (29 women, age range: 19– 70, mean age \pm SD = 37.9 \pm 13.6). A sensitivity analysis (Campbell & Thompson, 2012) indicated that the smallest possible effect size (η_p^2) that can be detected with this sample size, assuming alpha level .05 and power .80, is equal to 0.094.

All participants gave informed consent before the study and received financial compensation for their participation (4£). This study was approved by the Radboud University Ethics Committee (ECSW – 2022-079).

2.3.2.2. Stimuli

A subset of 66 stimuli (33 indoor and 33 outdoor) was selected from the newly created stimulus set that had originally been developed for the TMS study (see Section 2.2.2.2. for details). From this pool, two pre-existing versions of each image were used: the only-object condition (OO), in which the intact object had been cropped from the scene and placed at its original location on a uniform gray background, and the scene-with-object condition (WO), in which the object was preserved but the background was degraded. In addition, a scrambled version of the scene-with-object condition was created by applying Fourier phase scrambling, yielding three conditions in total: OO, WO, and WOs (scrambled-scene-with-object), with the intact object presented on a scrambled background. The final stimulus pool, therefore,

comprised 198 images (see Fig. 19), equally balanced between indoor and outdoor scenes. To minimize familiarity and carry-over effects, only one condition of a given image was presented to each participant. Accordingly, three stimulus sets of 66 images each were created, and their assignment was counterbalanced across participants.

2.3.2.3. Procedure

The experimental procedure was programmed using PsychoPy (version 2023.1.3; Peirce et al., 2019) and administered online via Pavlovia (<https://pavlovia.org>). It comprised 264 trials (the set of 66 stimuli was presented four times) and took around 20 minutes to complete. The trial sequence and timing followed the procedure described previously (see Section 2.2.2.2.1.1. for details) and illustrated in Figure 13B. Participants first completed 12 practice trials with feedback. During the main task, feedback was provided only during breaks, which occurred every 24 trials, resulting in a total of 11 short breaks.

Concurrently, the Deep Learning Toolbox™ model for Places365-GoogLeNet was used to classify the same set of images. For each image (in each condition), the top five predictions were extracted. These predictions were then divided into two categories: indoor and outdoor (see Appendix B). For each image, the number of correct predictions was summed, resulting in a score ranging from 0 (all predictions incorrect) to 5 (all predictions correct).

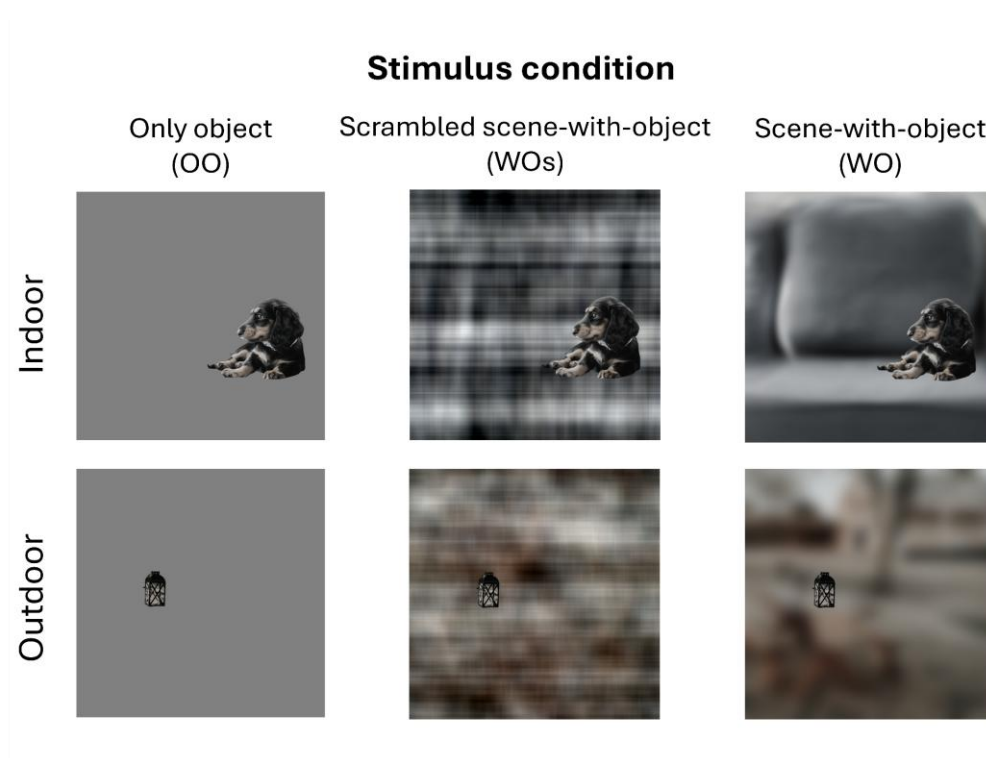


Figure 19. Examples of scene images used in the study in three versions: only-object (OO), scrambled-scene-with-object (WOs), and scene-with-object (WO). The objects belonged to either the animate (cats, dogs, people) or inanimate (chairs, lamps, plants) categories, and were balanced across indoor and outdoor scenes.

2.3.2.4. Data analysis

Analyses of variance (ANOVAs) were performed on recognition accuracy.

First, a one-way repeated-measures ANOVA was conducted to examine whether human performance differed across stimulus conditions (only-object [OO], scene-with-object [WO], and scrambled-scene-with-object [WOs]). A corresponding one-way repeated-measures ANOVA was performed to assess differences across the same conditions for the Places365-GoogLeNet network. Finally, a mixed-design 2×3 ANOVA was conducted to directly compare human and model performance, with classifier type (human participants vs. DNN) as a between-subjects factor and stimulus condition (OO, WO, WOs) as a within-subjects factor.

All analyses used mean recognition accuracy as the dependent variable, and post-hoc comparisons were corrected using the Bonferroni method. The Greenhouse-Geisser (GG)

correction was applied when necessary to account for violations of sphericity. Values are reported as Mean \pm SEM.

2.3.3. Results

2.3.3.1. Human performance

The analysis revealed a significant effect of stimulus condition on mean accuracy ($F(1.700,110.531) = 30.939$, $p < .001$, $\eta_p^2 = .322$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that the images in the scene-with-object (WO) condition (0.752 ± 0.020) were classified significantly better than the images in the scrambled-scene-with-object (WOs; 0.643 ± 0.029 ; $t(65) = 6.070$, $p_{adj} < .001$, $d = 0.499$) and the only-object (OO; 0.619 ± 0.031 ; $t(65) = 7.368$, $p_{adj} < .001$, $d = 0.605$) conditions. The difference between the WOs and OO conditions was not significant ($t(65) = 1.298$, $p_{adj} = .590$, $d = 0.107$)

The results are shown in Fig. 20.

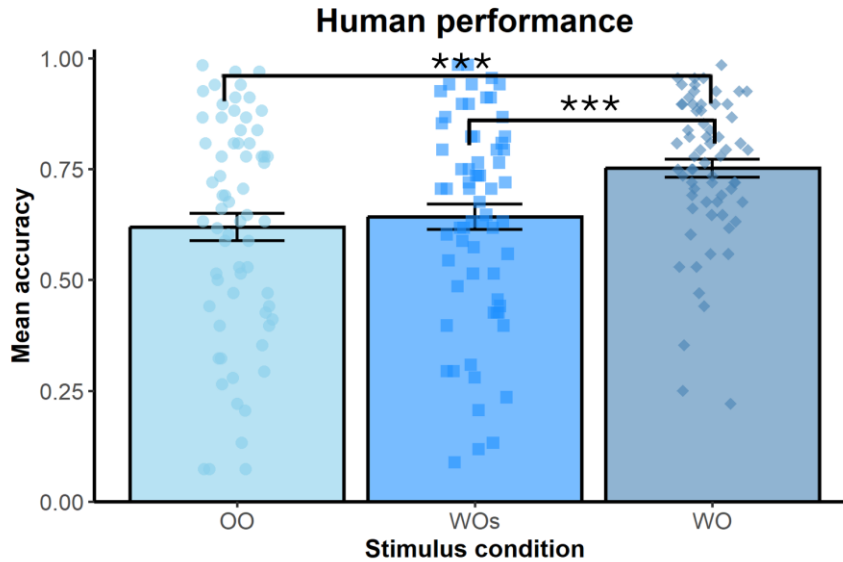


Figure 20. Participants classified images in the scene-with-object (WO) condition significantly more accurately than in the scrambled-scene-with-object (WOs) and only-object (OO) conditions. No significant difference was observed between the WOs and OO conditions. Bars represent mean accuracy; error bars denote the standard error of the mean (SEM); individual points indicate accuracy for each image within a given condition. Asterisks indicate (***) statistical significance ($p < .001$).

2.3.3.2. Places365-GoogLeNet network performance

The analysis revealed a significant effect of stimulus condition on mean accuracy ($F(2,130) = 4.562$, $p = .012$, $\eta_p^2 = .066$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that the images in the scene-with-object condition (0.724 ± 0.040) were classified significantly better than the images in the only-object condition (0.573 ± 0.042 ; $t(65) = 2.970$, $p_{adj} = .011$, $d = 0.449$). Accuracy in the scrambled-scene-with-object condition (0.673 ± 0.043) did not differ significantly from either the scene-with-object condition ($t(65) = 1.010$, $p_{adj} = .943$, $d = 0.153$) or the only-object condition ($t(65) = 1.960$, $p_{adj} = .156$, $d = 0.296$).

The results are shown in Fig. 21.

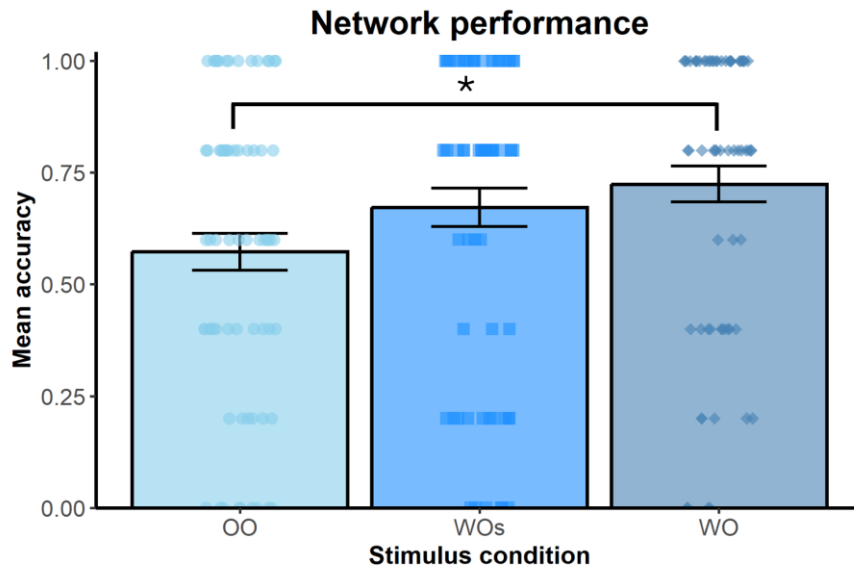


Figure 21. The Places365-GoogLeNet network classified images in the scene-with-object (WO) condition significantly more accurately than in the only-object (OO) condition. Accuracy in the scrambled-scene-with-object (WOs) condition did not differ significantly from either the WO or OO conditions. Bars represent mean accuracy; error bars denote the standard error of the mean (SEM); individual points indicate accuracy for each image within a given condition. Asterisk (*) indicates statistical significance ($p < .05$).

2.3.3.3. Comparison between human and Places365-GoogLeNet performance

The analysis revealed no significant difference in classification accuracy between human participants (0.671 ± 0.028) and the Places365-GoogLeNet network (0.657 ± 0.042 ; $F(1,130) = 0.151$, $p = .698$, $\eta_p^2 = .001$). A significant main effect of stimulus was observed ($F(2,260) = 13.889$, $p < .001$, $\eta_p^2 = .097$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that the images in the scene-with-object (WO) condition (0.738 ± 0.022) were classified significantly more accurately than those in the scrambled-scene-with-object (WOs; 0.658 ± 0.026 ; $t(131) = 2.973$, $p_{adj} = .010$, $d = 0.283$) and the only-object (OO; 0.596 ± 0.026 ; $t(131) = 5.253$, $p_{adj} < .001$, $d = 0.501$) conditions. The difference between the WOs and the OO conditions was not significant ($t(131) = 2.280$, $p_{adj} = .070$, $d = 0.217$). There was also no significant interaction between classifier type and stimulus condition ($F(2,260) = 1.090$, $p = .338$, $\eta_p^2 = .008$).

2.3.4. Discussion

The present study had two main aims. First, we tested whether object-based scene recognition depends on a coherent scene layout or can be accounted for by low-level scene statistics alone. Second, we assessed whether a scene-trained deep network (Places365–GoogLeNet) exhibits human-like sensitivity to contextual structure.

In humans, object-based facilitation was absent for phase-scrambled scenes: performance with an object on a scrambled background was significantly worse than with an object on an ambiguous background and indistinguishable from an object on a neutral background. This pattern is consistent with the first, layout-mediated account: object-based facilitation appears to depend on intact global scene structure, as disrupting the spatial layout abolished the benefit despite preserved low-level scene statistics. Under this interpretation, disambiguation likely arises at later stages, after object processing, with object properties (e.g., position) biasing the parsing of scene structure. This temporal profile accords with evidence that scene-based predictions bias object processing at later stages (Wischnewski & Peelen, 2021b). However, since based on Study 2 (Chapter 2.2), the precise timing of the LOC contribution to the disambiguation process could not be established, further research is needed to test and confirm this interpretation.

For Places365-GoogLeNet, classification was significantly better for the scene-with-object displays than for only-object displays, whereas performance with an object on a scrambled background was intermediate and did not differ reliably from either condition. A direct comparison between humans and the network showed no significant group differences across conditions. Taken together, this pattern suggests that, despite architectural differences, both systems can exploit object-based cues to support scene classification and are at least partly sensitive to spatial layout information.

The absence of clear differences between human and network performance is noteworthy, given mixed prior results. Humans often outperform DNNs on object recognition under a range of image degradations, and their error patterns diverge progressively as the signal strength drops (Geirhos et al., 2018). Further, deep networks can be made to misclassify otherwise correctly recognised images by adding near-imperceptible perturbations that maximise the model prediction error (Szegedy et al., 2013). By contrast, networks can also assign high-confidence labels to inputs that are unrecognisable to humans (so-called *fooling images*; Nguyen et al., 2015). To probe whether humans and the network relied on similar cues in scrambled scenes, an informative next step is trial-by-trial error-

consistency analysis (Geirhos et al., 2020), which tests whether the two systems tend to fail on the same images. High consistency would suggest shared strategies despite architectural differences, whereas low consistency would indicate different information use despite comparable mean accuracy.

3. General discussion

How do we so swiftly make sense of the visual world? This long-standing question has puzzled researchers in cognitive psychology, neuroscience, and, more recently, artificial intelligence. Early research focused on the neural processing of simple visual patterns, producing key insights into cortical organization. However, this reductionist strategy raised concerns about ecological validity and generalization to everyday vision. In contrast to this approach, a parallel trend emerged in the 1970s: researchers began using naturalistic photographs as experimental stimuli, moving from elementary patterns to images that approximate real-world input. Although this shift complicated experimental control, introducing greater variability and potential confounds, natural scenes quickly became a widely adopted proxy for everyday vision, and in the ensuing decades, real-world scene processing emerged as a distinct domain of inquiry (see Section 1.1). One of the most striking findings within the scene vision field was that the intuitive divide between objects and scene context is reflected at the neural level (see Section 1.3). Yet, despite substantial progress, the timing, mechanisms, and neural implementation of scene-object interactions remain a matter of active debate.

One of the reasons for this is that real-world scene perception has long been pursued along two largely separate lines: one centred on scenes, the other on objects. These traditions have often been guided by implicit assumptions – most notably that scenes, as the predominant source of information, influence object perception, while the reverse, object-to-scene influence, has rarely been treated as equally important. As a result, the very interactions that are central to real-world vision have rarely been examined within a single, coherent framework. This split, together with heterogeneous stimulus sets and single-task paradigms, makes it difficult to integrate findings and to draw principled conclusions about scene–object interactions.

The present thesis tackles this problem by focusing directly on these interactions across a series of experiments that draw on established theoretical frameworks of scene understanding. In line with contemporary work on scene perception, naturalistic photographs were used as proxies for real-world visual input, and carefully controlled manipulations were applied to probe how scene and object representations influence each other. Behaviour was quantified using multiple indices – accuracy, RT, and speed–accuracy–integrated measures (BIS/LISAS) – to limit speed–accuracy trade-off confounds and to enable robust comparison

of scene-to-object and object-to-scene influences. In addition, neurostimulation was employed to assess the causal contribution of scene- and object-selective cortex to the observed effects, and human performance was compared with that of a scene-trained deep neural network (Places365–GoogLeNet). This comparison provided a feedforward benchmark for the behavioural patterns reported in this thesis.

3.1. Summary of the discussed studies and their outcomes

Study 1 sought to determine whether one type of representation – scenes or objects – exhibits temporal precedence in naturalistic vision, and whether semantic congruency effects reflect primarily unidirectional (either scene-to-object or object-to-scene) or bidirectional influences. These questions were situated within hierarchical accounts of scene perception by contrasting analytic (“object-first”) and holistic (“scene-first”) perspectives, which make opposing claims about the temporal ordering of scene and object processing. A key strength of the study lies in its within-subjects design: the same group of participants completed scene- and object-categorization tasks using a shared stimulus set in which scene–object congruency was orthogonally manipulated. This approach enabled direct comparisons between scene and object processing within the same individuals. The use of two task types also enabled an assessment of whether the findings generalise across different task contexts.

After accounting for speed–accuracy trade-offs, the results showed no temporal advantage for either scenes or objects. Congruency effects were symmetric across tasks, indicating bidirectional rather than strictly unidirectional influences between scene context and object processing. Taken together, these findings do not support the notion of a fixed hierarchy in naturalistic scene–object processing.

Study 2 investigated whether object representations can causally contribute to the disambiguation of scene representations. Previous chronometric TMS work demonstrated that disrupting the scene-selective occipital place area (OPA) impairs context-based disambiguation of objects, indicating that scene information in OPA plays a causal role in object recognition (Wischnewski & Peelen, 2021b). What remained unresolved, however, was whether an analogous mechanism operates in the reverse direction, whereby objects contribute to the disambiguation of scene representations.

To address this question, Study 2 focused on object-based scene recognition, in which an otherwise ambiguous scene becomes categorizable in the presence of an object. The key test was whether the object-selective lateral occipital complex (LOC) contributes causally to this effect and whether the temporal profile of this influence parallels that previously reported for the OPA in the context-based object disambiguation. Chronometric TMS was applied over the LOC or the OPA while participants categorised scenes without objects and ambiguous scenes with objects. Disrupting the LOC selectively impaired disambiguation when an object was present, thereby providing the first direct evidence that object-selective cortex plays a causal role in object-based scene recognition.

The precise time window for LOC involvement could not be isolated, and it remains unresolved whether – and when – the scene-selective OPA contributes to this effect. Consequently, the present pattern offers only partial support for a shared bidirectional predictive mechanism.

Study 3 aimed to deepen understanding of the object-based scene recognition effect by examining whether scene disambiguation depends on preserved scene layout or can be driven by low-level scene statistics alone. To this end, phase-scrambled scenes were employed, in which global layout was disrupted while low-level image statistics were preserved. Under these conditions, no object-based facilitation was observed: performance on scrambled scenes closely approximated the object-only baseline. The performance of the scene-trained DNN (Places365–GoogLeNet) did not differ significantly from that of human observers in any condition. Together, these results indicate that, in both humans and a feedforward model, object-based gains emerge only when a scene’s global structure is preserved, suggesting that scene–object interactions rely on layout-anchored relations rather than on the mere aggregation of low-level features.

3.2. Implications of the present studies

The results obtained in the described series of experiments can be viewed as a foundation for a more fine-grained characterisation of the mechanisms underlying scene–object interactions. A key next step is to determine the boundary conditions of the observed effects. Future studies should systematically manipulate the ambiguity of both scene and object cues to quantify their respective contributions to congruency costs – a natural prediction is that greater ambiguity will increase reliance on contextual information.

In parallel, it will be important to test whether a syntactic relation between the scene and the object is indeed crucial for object-based scene recognition. One approach would be to present an object that originally belongs to a given scene at a different location within that scene, thereby disrupting the syntactic, while preserving the semantic, scene–object relationship. If the facilitation effect depends on syntactic structure, such a position change should eliminate or at least attenuate object-based scene facilitation. By contrast, if object-based scene recognition is driven primarily by semantic association between scene and object, manipulating object position should not affect either the presence or the magnitude of the facilitation effect. Notably, unlike the human visual system (Gayet & Peelen, 2022), current deep networks have been reported to show relative insensitivity to contextual priors such as expected real-world size, performing similarly for canonical and anomalous sizes (Eckstein et al., 2017). This makes spatial position – another form of contextual prior – a particularly suitable candidate for probing human–network divergences: a spatial position manipulation may reveal a dissociation, with human performance depending strongly on an object’s location within the scene structure and a scene-trained network remaining comparatively unaffected by it.

Further, it is worth underscoring that the present findings were obtained with brief stimulus exposures under conditions that promote attentive, conscious viewing. Inferences should therefore be restricted to such conditions. Future work should systematically manipulate attentional load and awareness (e.g., via dual-task paradigms or masking/continuous flash suppression combined with objective awareness checks) to evaluate whether the observed effects – such as object-based scene disambiguation – require attention or consciousness to occur, or whether they can arise under reduced awareness or increased cognitive load.

Finally, it would be informative to examine individual differences (e.g., global- vs. local-processing preference; Chamberlain et al., 2015) and relate them to behavioural and neurophysiological indices during real-world scene perception.

3.3. Limitations and future directions

From a broader perspective, although the shift from simple patterns to naturalistic photographs was transformative, it remains a limited approach to perception. Viewing static images on a screen is far from everyday vision: real-world experience is multisensory, temporally continuous, and shaped by our capacity to act within the environment. These features of perception are central to an affordance-centered approach to real-world visual processing (review: Bartnik & Groen, 2023). The core idea is that environments invite specific behaviors and that the visual system is ecologically tuned to detect invariant properties “to perceive what they afford” (Gibson, 1977, p. 127). Within this framework, objects and structural properties of scenes matter for perception insofar as they specify opportunities for action, which vary across organisms and across individuals as goals change. A natural next step in naturalistic perception research is therefore to determine whether – and how – scene representations measured in laboratory, picture-based paradigms are actually engaged when people (inter)act in real-world environments. Advancing this question, however, requires moving scene perception research into the wild, where maintaining experimental control over the visual environment and acquiring brain measurements is substantially harder. Thus, further progress in this area is likely to rely not only on new theoretical frameworks, but also on the development and application of novel methodologies (Bickle, 2016). Ecologically valid research on real-world vision is increasingly feasible due to technological developments, including mobile EEG (Djebbara et al., 2019), active tasks performed in real settings (Draschkow & Vö, 2016), virtual reality (David et al., 2021), and vision–action paradigms in which participants carry out naturalistic tasks (such as simulated driving during fMRI; Zhang & Gallant, 2020).

At the neural level, one particularly promising technique is layer-resolved fMRI (Lawrence et al., 2019), which may enable the dissociation of feedforward and feedback processes in the brain, thereby offering more fine-grained insights into the communication between object- and scene-selective pathways. Another emerging method is transcranial ultrasound stimulation, which offers superior spatial resolution and greater depth of penetration compared to traditional non-invasive brain stimulation techniques (Darmani et al., 2022). This approach holds promise for elucidating the causal roles of deeper cortical and subcortical structures, such as the PPA and the RSC processing.

The insights gained from these advanced methodologies could be further integrated and tested within the neuroconnectionist framework (Doerig et al., 2023), offering a computational platform for simulating neural interactions underlying scene and object processing. A deeper understanding of the principles underlying scene–object interactions and their neural implementation might also inform the development of more human-like deep neural networks (DNNs), which, in turn, could serve as testbeds for hypotheses that are challenging to examine empirically in human participants.

4. Summary and conclusions

To sum up, the results reported in this thesis indicate that:

- No temporal precedence. After controlling for speed–accuracy trade-offs, neither scenes nor objects representations showed a stable behavioural timing advantage across tasks.
- Bidirectional influence. After controlling for speed–accuracy trade-offs, incongruency effects were symmetric: incongruent scene context delayed object recognition, and incongruent objects delayed scene recognition.
- Causal role of object representations in disambiguating scenes. Chronometric TMS over the lateral occipital complex (LOC) selectively impaired object-based scene recognition for ambiguous images, confirming a causal contribution of the LOC-mediated object representations to scene disambiguation.
- Scene layout is necessary. Object-based scene recognition occurred only when a coherent global layout was preserved; integrating low-level scene statistics with object cues was insufficient.
- OPA involvement and temporal profile remain unresolved. Chronometric TMS revealed no reliable effects of stimulation over the occipital place area (OPA), and the precise post-stimulus window during which the LOC contributes to the object-based scene recognition could not be isolated.
- Human–model parity under scrambling. Under phase scrambling, no significant differences were observed between human observers and a feedforward, scene-trained DNN; this parity indicates that low-level scene statistics alone are insufficient – in both architectures – to support object-based scene recognition.

In conclusion, the present thesis, which aimed to characterise the nature of scene–object interactions in naturalistic vision, yields a convergent pattern of findings that is difficult to reconcile with strictly hierarchical accounts of scene perception: neither scene nor object representations exhibit fixed temporal precedence, and their influences are mutually constraining. Moreover, in line with this bidirectional view, the TMS results show that object representations in the LOC can causally disambiguate scene representations, paralleling prior evidence that scene representations can disambiguate object perception. At the same time, further experiments are needed to determine whether a common predictive mechanism underlies these interactions, including confirming the causal contribution of the OPA to object-based scene disambiguation and specifying the precise temporal windows of both the OPA and the LOC involvement. The finding that low-level scene statistics are insufficient to produce object-facilitated scene recognition motivates future work aimed at dissecting the sources of the observed effect. It also underscores the need to determine which aspects of object-based scene recognition are specific to the human visual system and which can be reproduced in feedforward architectures.

5. References

1. Aggleton, J. P. (2014). Looking beyond the hippocampus: old and new neurological targets for understanding memory disorders. *Proceedings of the Royal Society B: Biological Sciences*, 281(1786), 20140565. <https://doi.org/10.1098/rspb.2014.0565>
2. Aguirre, G. K., & D'Esposito, M. (1999). Topographical disorientation: a synthesis and taxonomy. *Brain*, 122(9), 1613-1628. <https://doi.org/10.1093/brain/122.9.1613>
3. Aguirre, G. K., Zarahn, E., & D'Esposito, M. (1998). An area within human ventral cortex sensitive to "building" stimuli: evidence and implications. *Neuron*, 21(2), 373-383.
4. Aminoff, E. M., & Durham, T. (2023). Scene-selective brain regions respond to embedded objects of a scene. *Cerebral Cortex*, 33(9), 5066-5074. <https://doi.org/10.1093/cercor/bhac399>
5. Aminoff, E., Gronau, N., & Bar, M. (2007). The parahippocampal cortex mediates spatial and nonspatial associations. *Cerebral cortex*, 17(7), 1493-1503. <https://doi.org/10.1093/cercor/bhl078>
6. Andresen, D. R., Vinberg, J., & Grill-Spector, K. (2009). The representation of object viewpoint in human visual cortex. *Neuroimage*, 45(2), 522-536. <https://doi.org/10.1016/j.neuroimage.2008.11.009>
7. Bacon-Macé, N., Macé, M. J. M., Fabre-Thorpe, M., & Thorpe, S. J. (2005). The time course of visual processing: Backward masking and natural scene categorisation. *Vision research*, 45(11), 1459-1469. <https://doi.org/10.1016/j.visres.2005.01.004>
8. Baldassano, C., Esteva, A., Fei-Fei, L., & Beck, D. M. (2016). Two distinct scene-processing networks connecting vision and memory. *Eneuro*, 3(5). <https://doi.org/10.1523/ENEURO.0178-16.2016>
9. Bar, M. (2004). Visual objects in context. *Nature Reviews Neuroscience*, 5(8), 617-629. <https://doi.org/10.1038/nrn1476>
10. Bar, M., Kassam, K. S., Ghuman, A. S., Boshyan, J., Schmid, A. M., Dale, A. M., Hämäläinen, M.S., Marinkovic, K., Schacter, D.L., Rosen, B.R., & Halgren, E. (2006). Top-down facilitation of visual recognition. *Proceedings of the national academy of sciences*, 103(2), 449-454. <https://doi.org/10.1073/pnas.0507062103>
11. Bar, M. (2014) .The current Scene. In K. Kveraga & M. Bar (Eds.), *Scene vision: Making sense of what we see* (pp. 1-3). MIT Press.
12. Bar, M., & Ullman, S. (1996). Spatial context in recognition. *Perception*, 25(3), 343-352. <https://doi.org/10.1068/p250343>

13. Barker, A. T., Jalinous, R., & Freeston, I. L. (1985). Non-invasive magnetic stimulation of human motor cortex. *The Lancet*, 325(8437), 1106-1107.
14. Bartnik, C. G., & Groen, I. I. (2023). Visual perception in the human brain: How the brain perceives and understands real-world scenes. In *Oxford research encyclopedia of neuroscience*. <https://doi.org/10.1093/acrefore/9780190264086.013.437>
15. Bickle, J. (2016). Revolutions in neuroscience: Tool development. *Frontiers in systems neuroscience*, 10, 24. <https://doi.org/10.3389/fnsys.2016.00024>
16. Biederman, I. (1972). Perceiving real-world scenes. *Science*, 177(4043), 77-80. DOI: [10.1126/science.177.4043.77](https://doi.org/10.1126/science.177.4043.77)
17. Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). *Scene perception: Detecting and judging objects undergoing relational violations*. *Cognitive psychology*, 14(2), 143-177. [https://doi.org/10.1016/0010-0285\(82\)90007-X](https://doi.org/10.1016/0010-0285(82)90007-X)
18. Biederman, I. (1987). Recognition-by-components: a theory of human image understanding. *Psychological review*, 94(2), 115.
19. Bodamer, J. (1947). Die prosop-agnosie. *Archiv für Psychiatrie und Nervenkrankheiten*, 179, 6–53. <https://doi.org/10.1007/BF00352849>
20. Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS computational biology*, 14(4), e1006111. <https://doi.org/10.1371/journal.pcbi.1006111>
21. Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J.E., Heaton, R. F., Evans, B.D., Mitchell, J. & Blything, R. (2023). Clarifying status of DNNs as models of human vision. *Behavioral and Brain Sciences*, 46. <https://doi.org/10.1017/S0140525X23002777>
22. Boyce, S. J., Pollatsek, A., & Rayner, K. (1989). Effect of background information on object identification. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3), 556.
23. Brainard, D. H., & Vision, S. (1997). The psychophysics toolbox. *Spatial vision*, 10(4), 433-436.
24. Brandman, T., & Peelen, M. V. (2017). Interaction between scene and object processing revealed by human fMRI and MEG decoding. *Journal of Neuroscience*, 37(32), 7700-7710. <https://doi.org/10.1523/JNEUROSCI.0582-17.2017>
25. Brandman, T., & Peelen, M. V. (2019). Signposts in the fog: objects facilitate scene representations in left scene-selective cortex. *Journal of cognitive neuroscience*, 31(3), 390-400. https://doi.org/10.1162/jocn_a_01258

26. Brandman, T., & Peelen, M. V. (2023). Objects sharpen visual scene representations: evidence from MEG decoding. *Cerebral Cortex*, 33(16), 9524-9531. <https://doi.org/10.1093/cercor/bhad222>
27. Braun, J. (2003). Natural scenes upset the visual appercept. *Trends in cognitive sciences*, 7(1), 7-9.
28. Bullier, J. (2001). Integrated model of visual processing. *Brain research reviews*, 36(2-3), 96-107. [https://doi.org/10.1016/S0165-0173\(01\)00085-6](https://doi.org/10.1016/S0165-0173(01)00085-6)
29. Campana, F., Rebollo, I., Urai, A., Wyart, V., & Tallon-Baudry, C. (2016). Conscious vision proceeds from global to local content in goal-directed tasks and spontaneous vision. *Journal of Neuroscience*, 36(19), 5200-5213. <https://doi.org/10.1523/JNEUROSCI.3619-15.2016>
30. Campana, F., & Tallon-Baudry, C. (2013). Anchoring visual subjective experience in a neural model: the coarse vividness hypothesis. *Neuropsychologia*, 51(6), 1050-1060. <https://doi.org/10.1016/j.neuropsychologia.2013.02.021>
31. Campbell, J. I., & Thompson, V. A. (2012). MorePower 6.0 for ANOVA with relational confidence intervals and Bayesian analysis. *Behavior research methods*, 44(4), 1255-1265. <https://doi.org/10.3758/s13428-012-0186-0>
32. Castelhana, M. S., & Krzyś, K. (2020). Rethinking space: A review of perception, attention, and memory in scene processing. *Annual Review of Vision Science*, 6(1), 563-586. <https://doi.org/10.1146/annurev-vision-121219-081745>
33. Chamberlain, R., Van der Hallen, R., Huygelier, H., Van de Cruys, S., & Wagemans, J. (2017). Local-global processing bias is not a unitary individual difference in visual processing. *Vision research*, 141, 247-257. <https://doi.org/10.1016/j.visres.2017.01.008> [Get rights and content](#)
34. Choo, H., & Walther, D. B. (2016). Contour junctions underlie neural representations of scene categories in high-level human visual cortex. *Neuroimage*, 135, 32-44. <https://doi.org/10.1016/j.neuroimage.2016.04.021>
35. Chun, M. M. (2003). Scene perception and memory. In *Psychology of Learning and Motivation* (Vol. 42, pp. 79-108). Academic Press. [https://doi.org/10.1016/S0079-7421\(03\)01003-X](https://doi.org/10.1016/S0079-7421(03)01003-X)
36. Cichy, R. M., Khosla, A., Pantazis, D., & Oliva, A. (2017). Dynamics of scene representations in the human brain revealed by magnetoencephalography and deep neural networks. *NeuroImage*, 153, 346-358. <https://doi.org/10.1016/j.neuroimage.2016.03.063>
37. Cichy, R. M., Pantazis, D., & Oliva, A. (2014). Resolving human object recognition in space and time. *Nature neuroscience*, 17(3), 455-462. <https://doi.org/10.1038/nn.3635>

38. 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), 27755. <https://doi.org/10.1038/srep27755>
39. Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, 36(3), 181-204. 10.1017/S0140525X12000477
40. Corp, D. T., Bereznicki, H. G., Clark, G. M., Youssef, G. J., Fried, P. J., Jannati, A., Davies, A. B., Gomes-Osman, J., Kirkovski, M., Albein-Urios, N., Fitzgerald, P. B., Koch, G., Di Lazzaro, V., Pascual-Leone, A. & Enticott, P. G. (2021). Large-scale analysis of interindividual variability in single and paired-pulse TMS data. *Clinical Neurophysiology*, 132(10), 2639-2653. <https://doi.org/10.1016/j.clinph.2021.06.014>
41. Crouzet, S. M., Joubert, O. R., Thorpe, S. J., & Fabre-Thorpe, M. (2012). Animal detection precedes access to scene category. *PLoS One*, 7(12), e51471. <https://doi.org/10.1371/journal.pone.0051471>
42. Darmani, G., Bergmann, T. O., Pauly, K. B., Caskey, C. F., De Lecea, L., Fomenko, A., Fouragnan, E., Legon, W., Murphy, K. R., Nandi, T., Phipps, M.A., Pinton, G., Ramezanpour, H., Sallet, J., Yaakub, S.N., Yoo, S.S & Chen, R. (2022). Non-invasive transcranial ultrasound stimulation for neuromodulation. *Clinical Neurophysiology*, 135, 51-73. <https://doi.org/10.1016/j.clinph.2021.12.010>
43. Davenport, J. L. (2007). Consistency effects between objects in scenes. *Memory & Cognition*, 35(3), 393-401. <https://doi.org/10.3758/BF03193280>
44. Davenport, J. L., & Potter, M. C. (2004). Scene consistency in object and background perception. *Psychological science*, 15(8), 559-564. <https://doi.org/10.1111/j.0956-7976.2004.00719.x>
45. David, E. J., Beitner, J., & Vő, M. L. H. (2021). The importance of peripheral vision when searching 3D real-world scenes: A gaze-contingent study in virtual reality. *Journal of Vision*, 21(7), 3-3. <https://doi.org/10.1167/jov.21.7.3>
46. Deng, Z. D., Lisanby, S. H., & Peterchev, A. V. (2013). Electric field depth–focality tradeoff in transcranial magnetic stimulation: simulation comparison of 50 coil designs. *Brain stimulation*, 6(1), 1-13. <https://doi.org/10.1016/j.brs.2012.02.005>
47. DiCarlo, J. J., Zoccolan, D., & Rust, N. C. (2012). How does the brain solve visual object recognition?. *Neuron*, 73(3), 415-434.
48. Dilks, D. D., Julian, J. B., Kubilius, J., Spelke, E. S., & Kanwisher, N. (2011). Mirror-image sensitivity and invariance in object and scene processing pathways. *Journal of Neuroscience*, 31(31), 11305-11312. <https://doi.org/10.1523/JNEUROSCI.1935-11.2011>

49. Dilks, D. D., Julian, J. B., Paunov, A. M., & Kanwisher, N. (2013). The occipital place area is causally and selectively involved in scene perception. *Journal of Neuroscience*, 33(4), 1331-1336. <https://doi.org/10.1523/JNEUROSCI.4081-12.2013>
50. Dilks, D. D., Kamps, F. S., & Persichetti, A. S. (2022). Three cortical scene systems and their development. *Trends in cognitive sciences*, 26(2), 117-127.
51. Dillon, M. R., Persichetti, A. S., Spelke, E. S., & Dilks, D. D. (2018). Places in the brain: bridging layout and object geometry in scene-selective cortex. *Cerebral Cortex*, 28(7), 2365-2374. <https://doi.org/10.1093/cercor/bhx139>
52. Djebbara, Z., Fich, L. B., Petrini, L., & Gramann, K. (2019). Sensorimotor brain dynamics reflect architectural affordances. *Proceedings of the National Academy of Sciences*, 116(29), 14769-14778. <https://doi.org/10.1073/pnas.1900648116>
53. Doerig, A., Sommers, R.P., Seeliger, K. *et al.* (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 24(7), 431-450. <https://doi.org/10.1038/s41583-023-00705-w>
54. Downing, P. E., Jiang, Y., Shuman, M., & Kanwisher, N. (2001). A cortical area selective for visual processing of the human body. *Science*, 293(5539), 2470-2473. [DOI: 10.1126/science.1063414](https://doi.org/10.1126/science.1063414)
55. Doya, K. (Ed.). (2007). *Bayesian brain: Probabilistic approaches to neural coding*. MIT press.
56. Draschkow, D., & Vö, M. L. H. (2016). Of “what” and “where” in a natural search task: Active object handling supports object location memory beyond the object’s identity. *Attention, Perception, & Psychophysics*, 78(6), 1574-1584. <https://doi.org/10.3758/s13414-016-1111-x>
57. Eckstein, M. P., Koehler, K., Welbourne, L. E., & Akbas, E. (2017). Humans, but not deep neural networks, often miss giant targets in scenes. *Current Biology*, 27(18), 2827-2832.
58. Eger, E., Kell, C. A., & Kleinschmidt, A. (2008). Graded size sensitivity of object-exemplar-evoked activity patterns within human LOC subregions. *Journal of Neurophysiology*, 100(4), 2038-2047. <https://doi.org/10.1152/jn.90305.2008>
59. Epstein, R. (2005). The cortical basis of visual scene processing. *Visual Cognition*, 12(6), 954-978. <https://doi.org/10.1080/13506280444000607>
60. Epstein, R. A. (2008). Parahippocampal and retrosplenial contributions to human spatial navigation. *Trends in cognitive sciences*, 12(10), 388-396.
61. Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual review of vision science*, 5(1), 373-397. doi.org/10.1146/annurev-vision-091718-014809

62. Epstein, R., & Kanwisher, N. (1998). A cortical representation of the local visual environment. *Nature*, 392(6676), 598-601. <https://doi.org/10.1038/33402>
63. Fabre-Thorpe, M. (2011). The characteristics and limits of rapid visual categorization. *Frontiers in psychology*, 2, 243. <https://doi.org/10.3389/fpsyg.2011.00243>
64. Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. (2001). A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of cognitive neuroscience*, 13(2), 171-180. <https://doi.org/10.1162/089892901564234>
65. Felsen, G., & Dan, Y. (2005). A natural approach to studying vision. *Nature neuroscience*, 8(12), 1643-1646. <https://doi.org/10.1038/nn1608>
66. Fize, D., Cauchoix, M., & Fabre-Thorpe, M. (2011). Humans and monkeys share visual representations. *Proceedings of the National Academy of Sciences*, 108(18), 7635-7640. <https://doi.org/10.1073/pnas.1016213108>
67. Foulsham, T. (2015). Scene perception. *The handbook of attention*, 257-280.
68. Friston, K. (2005). A theory of cortical responses. *Philosophical transactions of the Royal Society B: Biological sciences*, 360(1456), 815-836. <https://doi.org/10.1098/rstb.2005.1622>
69. Fukushima, K. (1980). Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological cybernetics*, 36(4), 193-202. <https://doi.org/10.1007/BF00344251>
70. Furtak, M., Mudrik, L., & Bola, M. (2022). The forest, the trees, or both? Hierarchy and interactions between gist and object processing during perception of real-world scenes. *Cognition*, 221, 104983. <https://doi.org/10.1016/j.cognition.2021.104983>
71. Ganaden, R. E., Mullin, C. R., & Steeves, J. K. (2013). Transcranial magnetic stimulation to the transverse occipital sulcus affects scene but not object processing. *Journal of cognitive neuroscience*, 25(6), 961-968. https://doi.org/10.1162/jocn_a_00372
72. Gandolfo, M., & Downing, P. E. (2019). Causal evidence for expression of perceptual expectations in category-selective extrastriate regions. *Current Biology*, 29(15), 2496-2500.
73. Gandolfo, M., Abassi, E., Balgova, E., Downing, P. E., Papeo, L., & Koldewyn, K. (2024). Converging evidence that left extrastriate body area supports visual sensitivity to social interactions. *Current Biology*, 34(2), 343-351.
74. Gayet, S., & Peelen, M. V. (2022). Preparatory attention incorporates contextual expectations. *Current Biology*, 32(3), 687-692.

75. Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of CNNs and humans by measuring error consistency. *Advances in neural information processing systems*, 33, 13890-13902.
76. Geirhos, R., Temme, C. R., Rauber, J., Schütt, H. H., Bethge, M., & Wichmann, F. A. (2018). Generalisation in humans and deep neural networks. *Advances in neural information processing systems*, 31.
77. Gibson, J. J. (1966). *The senses considered as perceptual systems*. Boston: Houghton Mifflin.
78. Gibson, J. J. (1977). The theory of affordances. In R. Shaw & J. Bransford (Eds.), *Perceiving, acting, and knowing: Toward an ecological psychology* (pp. 67–82). Lawrence Erlbaum.
79. Gobbini, M. I., Gentili, C., Ricciardi, E., Bellucci, C., Salvini, P., Laschi, C., Guazzelli, M. & Pietrini, P. (2011). Distinct neural systems involved in agency and animacy detection. *Journal of cognitive neuroscience*, 23(8), 1911-1920. <https://doi.org/10.1162/jocn.2010.21574>
80. Goffaux, V., Jacques, C., Mouraux, A., Oliva, A., Schyns, P., & Rossion, B. (2005). Diagnostic colours contribute to the early stages of scene categorization: Behavioural and neurophysiological evidence. *Visual cognition*, 12(6), 878-892. <https://doi.org/10.1080/13506280444000562>
81. Goodale, M. A., Milner, A. D., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, 349(6305), 154-156. <https://doi.org/10.1038/349154a0>
82. Grill-Spector, K. (2003). The neural basis of object perception. *Current opinion in neurobiology*, 13(2), 159-166. [https://doi.org/10.1016/S0959-4388\(03\)00040-0](https://doi.org/10.1016/S0959-4388(03)00040-0)
83. Grill-Spector, K., Kushnir, T., Edelman, S., Itzchak, Y., & Malach, R. (1998). Cue-invariant activation in object-related areas of the human occipital lobe. *Neuron*, 21(1), 191-202.
84. Grill-Spector, K., & Malach, R. (2004). The human visual cortex . *Annual Review of Neuroscience*, 27(1), 649–677. <https://doi.org/10.1146/annurev.neuro.27.070203.144220>
85. Groen, I. I., Ghebreab, S., Lamme, V. A., & Scholte, H. S. (2016). The time course of natural scene perception with reduced attention. *Journal of neurophysiology*, 115(2), 931-946. <https://doi.org/10.1152/jn.00896.2015>
86. Groen, I. I., 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. <https://doi.org/10.1098/rstb.2016.0102>

87. Gross, C. G. (1994). How inferior temporal cortex became a visual area. *Cerebral cortex*, 4(5), 455-469. <https://doi.org/10.1093/cercor/4.5.455>
88. Güçlü, U., & Van Gerven, M. A. (2015). Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27), 10005-10014. <https://doi.org/10.1523/JNEUROSCI.5023-14.2015>
89. Hämäläinen, M., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography—theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of modern Physics*, 65(2), 413. DOI: <https://doi.org/10.1103/RevModPhys.65.413>
90. Hanlon, C. A., & McCalley, D. M. (2022). Sex/gender as a factor that influences transcranial magnetic stimulation treatment outcome: three potential biological explanations. *Frontiers in psychiatry*, 13, 869070. <https://doi.org/10.3389/fpsy.2022.869070>
91. Harel, A., Groen, I. I., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *Eneuro*, 3(5). <https://doi.org/10.1523/ENEURO.0139-16.2016>
92. Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of d' . *Behavior research methods, instruments, & computers*, 27(1), 46-51. <https://doi.org/10.3758/BF03203619>
93. He, K., Zhang, X., Ren, S., & Sun, J. (2015). Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In *Proceedings of the IEEE international conference on computer vision* (pp. 1026-1034).
94. Henderson, J. M., & Hollingworth, A. (1999). High-level scene perception. *Annual review of psychology*, 50(1), 243-271. <https://doi.org/10.1146/annurev.psych.50.1.243>
95. Hollingworth, A. & Henderson, J. M. Does Consistent Scene Context Facilitate Object Perception? *Journal of Experimental Psychology: General* 127, 398–415 (1998)
96. Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *Neuron*, 103(1), 161-171.
97. Hochstein, S., & Ahissar, M. (2002). View from the top: Hierarchies and reverse hierarchies in the visual system. *Neuron*, 36(5), 791-804. [https://doi.org/10.1016/S0896-6273\(02\)01091-7](https://doi.org/10.1016/S0896-6273(02)01091-7)
98. Hubel, D. H., & Wiesel, T. N. (1959). Receptive fields of single neurones in the cat's striate cortex. *The Journal of physiology*, 148(3), 574.

99. Hubel, D. H., & Wiesel, T. N. (1962). Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*, 160(1), 106-154. doi: [10.1113/jphysiol.1959.sp006308](https://doi.org/10.1113/jphysiol.1959.sp006308)
100. Hubel, D. H., & Wiesel, T. N. (1979). Brain mechanisms of vision. *Scientific American*, 241(3), 150-163.
101. Ishai, A., Ungerleider, L. G., Martin, A., & Haxby, J. V. (2000). The representation of objects in the human occipital and temporal cortex. *Journal of cognitive neuroscience*, 12(Supplement 2), 35-51. <https://doi.org/10.1162/089892900564055>
102. Ishai, A., Ungerleider, L. G., Martin, A., Schouten, J. L., & Haxby, J. V. (1999). Distributed representation of objects in the human ventral visual pathway. *Proceedings of the National Academy of Sciences*, 96(16), 9379-9384. <https://doi.org/10.1073/pnas.96.16.9379>
103. Isik, L., Meyers, E. M., Leibo, J. Z., & Poggio, T. (2014). The dynamics of invariant object recognition in the human visual system. *Journal of neurophysiology*, 111(1), 91-102. <https://doi.org/10.1152/jn.00394.2013>
104. Joubert, O. R., Fize, D., Rousselet, G. A., & Fabre-Thorpe, M. (2008). Early interference of context congruence on object processing in rapid visual categorization of natural scenes. *Journal of Vision*, 8(13), 11-11. doi:<https://doi.org/10.1167/8.13.11>
105. Joubert, O. R., Rousselet, G. A., Fize, D., & Fabre-Thorpe, M. (2007). Processing scene context: Fast categorization and object interference. *Vision research*, 47(26), 3286-3297. <https://doi.org/10.1016/j.visres.2007.09.013>
106. Josephs, E. L., & Konkle, T. (2020). Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. *Proceedings of the National Academy of Sciences*, 117(47), 29354-29362. <https://doi.org/10.1073/pnas.1912333117>
107. Julian, J. B., Ryan, J., Hamilton, R. H., & Epstein, R. A. (2016). The occipital place area is causally involved in representing environmental boundaries during navigation. *Current Biology*, 26(8), 1104-1109.
108. Kaiser, D., Azzalini, D. C., & Peelen, M. V. (2016). Shape-independent object category responses revealed by MEG and fMRI decoding. *Journal of neurophysiology*, 115(4), 2246-2250. <https://doi.org/10.1152/jn.01074.2015>
109. Kaiser, D., Häberle, G., & Cichy, R. M. (2020). Real-world structure facilitates the rapid emergence of scene category information in visual brain signals. *Journal of Neurophysiology*, 124(1), 145-151. <https://doi.org/10.1152/jn.00164.2020>
110. Kanwisher, N., McDermott, J., & Chun, M. M. (1997). The fusiform face area: a module in human extrastriate cortex specialized for face perception. *Journal of*

- neuroscience*, 17(11), 4302-4311. <https://doi.org/10.1523/JNEUROSCI.17-11-04302.1997>
111. Kauffmann, L., Ramanoël, S., Guyader, N., Chauvin, A., & Peyrin, C. (2015). Spatial frequency processing in scene-selective cortical regions. *NeuroImage*, 112, 86-95. <https://doi.org/10.1016/j.neuroimage.2015.02.058>
 112. Kayser, C., Körding, K. P., & König, P. (2004). Processing of complex stimuli and natural scenes in the visual cortex. *Current opinion in neurobiology*, 14(4), 468-473. <https://doi.org/10.1016/j.conb.2004.06.002>
 113. Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: a canonical cortical computation. *Neuron*, 100(2), 424-435.
 114. Khaligh-Razavi, S. M., & Kriegeskorte, N. (2014). Deep supervised, but not unsupervised, models may explain IT cortical representation. *PLoS computational biology*, 10(11), e1003915. <https://doi.org/10.1371/journal.pcbi.1003915>
 115. Kimchi, R. (1992). Primacy of wholistic processing and global/local paradigm: a critical review. *Psychological bulletin*, 112(1), 24.
 116. Koffka, K. (1922). Perception: an introduction to the Gestalt-Theorie. *Psychological bulletin*, 19(10), 531.
 117. Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: it's the spaces more than the places. *Journal of Neuroscience*, 31(20), 7322-7333. <https://doi.org/10.1523/JNEUROSCI.4588-10.2011>
 118. Kriegeskorte, N. (2015). Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1(1), 417-446. <https://doi.org/10.1146/annurev-vision-082114-035447>
 119. Kirchner, H., & Thorpe, S. J. (2006). Ultra-rapid object detection with saccadic eye movements: Visual processing speed revisited. *Vision research*, 46(11), 1762-1776. <https://doi.org/10.1016/j.visres.2005.10.002>
 120. Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25.
 121. 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.
 122. Lawrence, S. J., Formisano, E., Muckli, L., & de Lange, F. P. (2019). Laminar fMRI: Applications for cognitive neuroscience. *Neuroimage*, 197, 785-791. <https://doi.org/10.1016/j.neuroimage.2017.07.004>

123. LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4), 541-551. doi: 10.1162/neco.1989.1.4.541.
124. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, 521(7553), 436-444. <https://doi.org/10.1038/nature14539>
125. Lee, E. G., Rastogi, P., Hadimani, R. L., Jiles, D. C., & Camprodon, J. A. (2018). Impact of non-brain anatomy and coil orientation on inter-and intra-subject variability in TMS at midline. *Clinical Neurophysiology*, 129(9), 1873-1883. <https://doi.org/10.1016/j.clinph.2018.04.749>
126. Leroy, A., Faure, S., & Spotorno, S. (2020). Reciprocal semantic predictions drive categorization of scene contexts and objects even when they are separate. *Scientific reports*, 10(1), 8447. <https://doi.org/10.1038/s41598-020-65158-y>
127. Li, F. F., VanRullen, R., Koch, C., & Perona, P. (2002). Rapid natural scene categorization in the near absence of attention. *Proceedings of the National Academy of Sciences*, 99(14), 9596-9601. <https://doi.org/10.1073/pnas.092277599>
128. Liesefeld, H. R., & Janczyk, M. (2019). Combining speed and accuracy to control for speed-accuracy trade-offs (?). *Behavior Research Methods*, 51, 40-60. <https://doi.org/10.3758/s13428-018-1076-x>
129. Liesefeld, H. R., & Janczyk, M. (2023). Same same but different: Subtle but consequential differences between two measures to linearly integrate speed and accuracy (LISAS vs. BIS). *Behavior Research Methods*, 55(3), 1175-1192. <https://doi.org/10.3758/s13428-022-01843-2>
130. Liu, H., Agam, Y., Madsen, J. R., & Kreiman, G. (2009). Timing, timing, timing: fast decoding of object information from intracranial field potentials in human visual cortex. *Neuron*, 62(2), 281-290. <https://doi.org/10.1016/j.neuron.2009.02.025>
131. Luck, S. J. (2014). *An introduction to the event-related potential technique*. MIT press.
132. Ma, W. J. (2012). Organizing probabilistic models of perception. *Trends in cognitive sciences*, 16(10), 511-518.
133. MacEvoy, S. P., & Epstein, R. A. (2007). Position selectivity in scene- and object-responsive occipitotemporal regions. *Journal of Neurophysiology*, 98(4), 2089-2098. <https://doi.org/10.1152/jn.00438.2007>
134. MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature neuroscience*, 14(10), 1323-1329. <https://doi.org/10.1038/nn.2903>
135. Malcolm, G. L., Groen, I. I., & Baker, C. I. (2016). Making sense of real-world scenes. *Trends in cognitive sciences*, 20(11), 843-856.

136. Mack, M. L., & Palmeri, T. J. (2010). Modeling categorization of scenes containing consistent versus inconsistent objects. *Journal of Vision*, 10(3), 11-11. <https://doi.org/10.1167/10.3.11>
137. Mendola, J. D., Dale, A. M., Fischl, B., Liu, A. K., & Tootell, R. B. (1999). The representation of illusory and real contours in human cortical visual areas revealed by functional magnetic resonance imaging. *Journal of Neuroscience*, 19(19), 8560-8572. <https://doi.org/10.1523/JNEUROSCI.19-19-08560.1999>
138. Nasr, S., Echavarria, C. E., & Tootell, R. B. (2014). Thinking outside the box: rectilinear shapes selectively activate scene-selective cortex. *Journal of Neuroscience*, 34(20), 6721-6735. <https://doi.org/10.1523/JNEUROSCI.4802-13.2014>
139. Navon, D. (1977). Forest before trees: The precedence of global features in visual perception. *Cognitive psychology*, 9(3), 353-383. [https://doi.org/10.1016/0010-0285\(77\)90012-3](https://doi.org/10.1016/0010-0285(77)90012-3)
140. Nguyen, A., Yosinski, J., & Clune, J. (2015). Deep neural networks are easily fooled: High confidence predictions for unrecognizable images. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 427-436).
141. Ogawa, S., Lee, T. M., Kay, A. R., & Tank, D. W. (1990). Brain magnetic resonance imaging with contrast dependent on blood oxygenation. *proceedings of the National Academy of Sciences*, 87(24), 9868-9872. <https://doi.org/10.1073/pnas.87.24.9868>
142. Oliva, A. (2005). Gist of the scene. In *Neurobiology of attention* (pp. 251-256). Academic press. <https://doi.org/10.1016/B978-012375731-9/50045-8>
143. Oliva, A., & Torralba, A. (2001). Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42, 145-175. <https://doi.org/10.1023/A:1011139631724>
144. Oliva, A., & Torralba, A. (2006). Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, 155, 23-36. [https://doi.org/10.1016/S0079-6123\(06\)55002-2](https://doi.org/10.1016/S0079-6123(06)55002-2)
145. Opitz, A., Legon, W., Rowlands, A., Bickel, W. K., Paulus, W., & Tyler, W. J. (2013). Physiological observations validate finite element models for estimating subject-specific electric field distributions induced by transcranial magnetic stimulation of the human motor cortex. *Neuroimage*, 81, 253-264. <https://doi.org/10.1016/j.neuroimage.2013.04.067>
146. Park, S., Brady, T. F., Greene, M. R., & Oliva, A. (2011). Disentangling scene content from spatial boundary: complementary roles for the parahippocampal place area and lateral occipital complex in representing real-world scenes. *Journal of Neuroscience*, 31(4), 1333-1340. <https://doi.org/10.1523/JNEUROSCI.3885-10.2011>

147. Peelen, M. V. (2024). Visual Cognitive Neuroscience. In M. C. Frank & A. Majid (Eds.), *Open Encyclopedia of Cognitive Science*. MIT Press. <https://doi.org/10.21428/e2759450.d3438ecd>
148. Peelen, M. V., Berlot, E., & de Lange, F. P. (2024). Predictive processing of scenes and objects. *Nature Reviews Psychology*, 3(1), 13-26. <https://doi.org/10.1038/s44159-023-00254-0>
149. Persichetti, A. S., & Dilks, D. D. (2018). Dissociable neural systems for recognizing places and navigating through them. *Journal of Neuroscience*, 38(48), 10295-10304. <https://doi.org/10.1523/JNEUROSCI.1200-18.2018>
150. Peirce, J., Gray, J.R., Simpson, S. *et al.* PsychoPy2: Experiments in behavior made easy. *Behav Res* 51, 195–203 (2019). <https://doi.org/10.3758/s13428-018-01193-y>
151. Potter MC (1975) Meaning in visual search. *Science* 187:965–966. DOI: [10.1126/science.1145183](https://doi.org/10.1126/science.1145183)
152. Potter, M. C. (2012). Recognition and memory for briefly presented scenes. *Frontiers in psychology*, 3, 32. <https://doi.org/10.3389/fpsyg.2012.00032>
153. Rafique, S. A., Solomon-Harris, L. M., & Steeves, J. K. (2015). TMS to object cortex affects both object and scene remote networks while TMS to scene cortex only affects scene networks. *Neuropsychologia*, 79, 86-96. <https://doi.org/10.1016/j.neuropsychologia.2015.10.027>
154. Rajimehr, R., Nasr, S., & Tootell, R. (2014). Deconstructing scene selectivity in visual cortex. *Scene vision: Making sense of what we see*, 73-84.
155. Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, 2(1), 79-87. <https://doi.org/10.1038/4580>
156. Ratcliff, R. (1979). Group reaction time distributions and an analysis of distribution statistics. *Psychological bulletin*, 86(3), 446.
157. Rémy, F., Saint-Aubert, L., Bacon-Macé, N., Vayssière, N., Barbeau, E., & Fabre-Thorpe, M. (2013). Object recognition in congruent and incongruent natural scenes: A life-span study. *Vision research*, 91, 36-44. <https://doi.org/10.1016/j.visres.2013.07.006>
158. Rémy, F., Vayssière, N., Pins, D., Boucart, M., & Fabre-Thorpe, M. (2014). Incongruent object/context relationships in visual scenes: Where are they processed in the brain?. *Brain and cognition*, 84(1), 34-43. <https://doi.org/10.1016/j.bandc.2013.10.008>
159. Rémy, F., Vayssière, N., Saint-Aubert, L., Bacon-Macé, N., Pariente, J., Barbeau, E., & Fabre-Thorpe, M. (2020). Age effects on the neural processing of object-context associations in briefly flashed natural scenes. *Neuropsychologia*, 136, 107264. <https://doi.org/10.1016/j.neuropsychologia.2019.107264>

160. Riesenhuber, M., & Poggio, T. (1999). Hierarchical models of object recognition in cortex. *Nature neuroscience*, 2(11), 1019-1025. <https://doi.org/10.1038/14819>
161. Rosenblatt, F. (1958). The perceptron: A probabilistic model for information storage and organization in the brain. *Psychological Review*, 65(6), 386–408. <https://doi.org/10.1037/h0042519>
162. Rossel, P., Peyrin, C., & Kauffmann, L. (2023). Subjective perception of objects depends on the interaction between the validity of context-based expectations and signal reliability. *Vision Research*, 206, 108191. <https://doi.org/10.1016/j.visres.2023.108191>
163. Rousset, G. A., Fabre-Thorpe, M., & Thorpe, S. J. (2002). Parallel processing in high-level categorization of natural images. *Nature neuroscience*, 5(7), 629-630. <https://doi.org/10.1038/nn866>
164. Rousset, G. A., Macé, M. J. M., & Fabre-Thorpe, M. (2003). Is it an animal? Is it a human face? Fast processing in upright and inverted natural scenes. *Journal of vision*, 3(6), 5-5. <https://doi.org/10.1167/3.6.5>
165. Rousset, G., Joubert, O., & Fabre-Thorpe, M. (2005). How long to get to the “gist” of real-world natural scenes?. *Visual cognition*, 12(6), 852-877. <https://doi.org/10.1080/13506280444000553>
166. Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *nature*, 323(6088), 533-536. <https://doi.org/10.1038/323533a0>
167. Russakovsky, O., Deng, J., Su, H. *et al.* ImageNet Large Scale Visual Recognition Challenge. *Int J Comput Vis* **115**, 211–252 (2015). <https://doi.org/10.1007/s11263-015-0816-y>
168. Rust, N. C., & Movshon, J. A. (2005). In praise of artifice. *Nature neuroscience*, 8(12), 1647-1650. <https://doi.org/10.1038/nn1606>
169. Sayres, R., & Grill-Spector, K. (2008). Relating retinotopic and object-selective responses in human lateral occipital cortex. *Journal of neurophysiology*, 100(1), 249-267. <https://doi.org/10.1152/jn.01383.2007>
170. Shibuya, K., Park, S. B., Geevasinga, N., Huynh, W., Simon, N. G., Menon, P., Howells, J., Vucic, S. & Kiernan, M. C. (2016). Threshold tracking transcranial magnetic stimulation: effects of age and gender on motor cortical function. *Clinical Neurophysiology*, 127(6), 2355-2361. <https://doi.org/10.1016/j.clinph.2016.03.009>
171. Stanislaw, H., & Todorov, N. (1999). Calculation of signal detection theory measures. *Behavior research methods, instruments, & computers*, 31(1), 137-149. <https://doi.org/10.3758/BF03207704>

172. Stansbury, D. E., Naselaris, T., & Gallant, J. L. (2013). Natural scene statistics account for the representation of scene categories in human visual cortex. *Neuron*, 79(5), 1025-1034.
173. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1-9).
174. Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., & Fergus, R. (2013). Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*. <https://doi.org/10.48550/arXiv.1312.6199>
175. Thomson, M. G. (1999). Visual coding and the phase structure of natural scenes. *Network: Computation in Neural Systems*, 10(2), 123. DOI 10.1088/0954-898X/10/2/302
176. Thorpe, S., Fize, D., & Marlot, C. (1996). Speed of processing in the human visual system. *Nature*, 381(6582), 520-522. <https://doi.org/10.1038/381520a0>
177. Titchener, E. B. (1902). *An outline of psychology*. Macmillan.
178. Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior research methods*, 49(2), 653-673. <https://doi.org/10.3758/s13428-016-0721-5>
179. Vandierendonck, A. (2018). Further tests of the utility of integrated speed-accuracy measures in task switching. *Journal of cognition*, 1(1). doi: 10.5334/joc.6
180. Vann, S. D., Aggleton, J. P., & Maguire, E. A. (2009). What does the retrosplenial cortex do?. *Nature reviews neuroscience*, 10(11), 792-802. <https://doi.org/10.1038/nrn2733>
181. 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. <https://doi.org/10.1162/08989290152001880>
182. Vő, M. L.-H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing . *Psychological Science*, 24(9), 1816–1823. <https://doi.org/10.1177/0956797613476955>
183. Vő, M. L. H., Boettcher, S. E., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current opinion in psychology*, 29, 205-210. <https://doi.org/10.1016/j.copsyc.2019.03.009>
184. Wagemans, J., Feldman, J., Gepshtein, S., Kimchi, R., Pomerantz, J. R., Van der Helm, P. A., & Van Leeuwen, C. (2012). A century of Gestalt psychology in visual perception: II. Conceptual and theoretical foundations. *Psychological bulletin*, 138(6), 1218.

185. Wang, Y., Vora, I., Huynh, B. P., Picard-Fraser, M., Daneshzand, M., Nummenmaa, A., & Kimberley, T. J. (2024). Coils are not created equal: Effects on TMS thresholding. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, 17(1), 1-3. doi: [10.1016/j.brs.2023.11.017](https://doi.org/10.1016/j.brs.2023.11.017)
186. Walsh, V., & Cowey, A. (2000). Transcranial magnetic stimulation and cognitive neuroscience. *Nature Reviews Neuroscience*, 1(1), 73-80. <https://doi.org/10.1038/35036239>
187. Walther, D. B., Caddigan, E., Fei-Fei, L., & Beck, D. M. (2009). Natural scene categories revealed in distributed patterns of activity in the human brain. *Journal of neuroscience*, 29(34), 10573-10581. <https://doi.org/10.1523/JNEUROSCI.0559-09.2009>
188. Weiner, K. S., Barnett, M. A., Witthoft, N., Golarai, G., Stigliani, A., Kay, K. N., Gomez, J., Natu, V.S., Amunts, K., Zilles, K. & Grill-Spector, K. (2018). Defining the most probable location of the parahippocampal place area using cortex-based alignment and cross-validation. *Neuroimage*, 170, 373-384. <https://doi.org/10.1016/j.neuroimage.2017.04.040>
189. Willenbockel, V., Sadr, J., Fiset, D., Horne, G. O., Gosselin, F., & Tanaka, J. W. (2010). Controlling low-level image properties: the SHINE toolbox. *Behavior research methods*, 42(3), 671-684. <https://doi.org/10.3758/BRM.42.3.671>
190. Wischniewski, M., & Peelen, M. V. (2021a). Causal evidence for a double dissociation between object-and scene-selective regions of visual cortex: a preregistered TMS replication study. *Journal of Neuroscience*, 41(4), 751-756. <https://doi.org/10.1523/JNEUROSCI.2162-20.2020>
191. Wischniewski, M., & Peelen, M. V. (2021b). Causal neural mechanisms of context-based object recognition. *Elife*, 10, e69736. <https://doi.org/10.7554/eLife.69736>
192. Wheatley, T., Milleville, S. C., & Martin, A. (2007). Understanding animate agents: distinct roles for the social network and mirror system. *Psychological science*, 18(6), 469-474. <https://doi.org/10.1111/j.1467-9280.2007.01923.x>
193. Wundt, W. M., & Wundt, W. (1874). *Grundzüge der physiologischen Psychologie* (Vol. 1). Engelmann.
194. Yamins, D. L., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3), 356-365. . <https://doi.org/10.1038/nn.4244>
195. Zhang, T., & Gallant, J. L. (2020). A naturalistic navigation task reveals rich distributed representations of information across the human cerebral cortex. *Journal of Vision*, 20(11), 462-462. <https://doi.org/10.1167/jov.20.11.462>

196. Zhou, B., Lapedriza, A., Khosla, A., Oliva, A., & Torralba, A. (2017). Places: A 10 million image database for scene recognition. *IEEE transactions on pattern analysis and machine intelligence*, 40(6), 1452-1464. doi: 10.1109/TPAMI.2017.2723009

6. Appendix

A. Pre-registered analyses and results of the four-pulse TMS study ([#153165 | AsPredicted](#))

Participants

Out of 56 participants, who initially qualified for the study, 53 completed the first experimental procedure. Three participants were identified as outliers and subsequently excluded from the analysis. Outliers were defined based on mean accuracy (1 participant) or mean reaction times (2 participants) falling below 2.5 standard deviations from the overall mean across conditions in the first session. The final sample included 50 participants (32 women, 18 men; mean age \pm SD = 23.04 \pm 3.53).

Analysis

The experiment employed a pre-registered 2x2x3 repeated-measures ANOVA design. Task (object, scene), stimulus condition (isolated, degraded), and stimulation site (LOC, OPA, and vertex) were within-subject factors. Accuracy was considered the primary dependent variable. Reaction times of correct trials and linear integrated speed-accuracy scores (LISAS; Vandierendonck, 2017) were treated as secondary dependent variables and analyzed in separate ANOVAs.

In line with the hypotheses, six pre-registered planned pairwise t-tests were also conducted comparing: 1) vertex vs LOC for isolated object recognition, 2) vertex vs LOC for degraded object recognition, 3) vertex vs LOC for degraded scenes recognition, and 4) vertex vs OPA for isolated scene recognition, 5) vertex vs OPA for degraded objects recognition, 6) vertex vs OPA for degraded scenes recognition. Further, the performance between LOC and OPA in the isolated object condition (7) and OPA and LOC in the isolated scene condition (8) were compared.

Values are reported as Mean \pm SD. Probability values were reported (p) for all statistical tests, and the standard .05 alpha level was used as a threshold for rejecting the null hypothesis.

Results

The mean accuracy, reaction times, and LISAS for each experimental condition can be found in Table A4.

Accuracy

The interaction between task, stimulus condition, and stimulation site was insignificant ($F(2,98) = 2.120$, $p = .125$, $\eta_p^2 = .041$). A significant main effect of site ($F(2,98) = 3.965$, $p = .022$, $\eta_p^2 = .075$) and a significant interaction between task and stimulus condition were found ($F(1,49) = 50.382$, $p < .001$, $\eta_p^2 = .507$). Post-hoc pairwise comparisons with a Bonferroni correction indicated that accuracy during the LOC stimulation was significantly worse than during the OPA stimulation (LOC: 0.849 ± 0.088 , OPA: 0.863 ± 0.083 ; $t(49) = -2.797$, $p_{adj} = .019$, $d = -0.194$). There were no significant differences between accuracy during the OPA and vertex (OPA: 0.863 ± 0.083 , vertex: 0.857 ± 0.085 ; $t(49) = 1.113$, $p_{adj} = .806$, $d = 0.077$) and the LOC and vertex stimulation (LOC: 0.849 ± 0.088 , vertex: 0.857 ± 0.085 ; $t(49) = -1.684$, $p_{adj} = .286$, $d = -0.117$). The analysis of simple main effects indicated that accuracy was significantly worse for objects relative to scenes, but only in the isolated stimulus condition (object: 0.785 ± 0.083 , scene: 0.892 ± 0.063 ; $F(1,49) = 110.559$, $p < .001$, $\eta_p^2 = .659$). For the degraded stimulus condition, accuracy did not differ significantly between tasks (object: 0.871 ± 0.069 , scene: 0.878 ± 0.082 ; $F(1,49) = 0.391$, $p = .535$, $\eta_p^2 = .009$). Further, while in the object task degraded stimuli were recognized better than isolated ones (degraded: 0.871 ± 0.069 , isolated: 0.785 ± 0.083 ; $F(1,49) = 75.778$, $p < .001$, $\eta_p^2 = .555$), in the scene task there were no significant differences between stimulus conditions (degraded: 0.878 ± 0.082 , isolated: 0.892 ± 0.063 ; $F(1,49) = 2.751$, $p = .104$, $\eta_p^2 = .033$).

In line with our hypotheses, eight pre-registered pairwise t-tests were also conducted: 1) vertex vs LOC for isolated object recognition ($t(49) = 0.660$, $p = .513$, $d = 0.093$) 2) vertex vs LOC for degraded object recognition ($t(49) = 1.319$, $p = .193$, $d = 0.187$) 3) vertex vs LOC for degraded scene recognition ($t(49) = -0.454$, $p = .652$, $d = -0.064$) 4) vertex vs OPA for isolated scene recognition ($t(49) = -1.066$, $p = .292$, $d = -0.151$) 5) vertex vs OPA for degraded object recognition ($t(49) = 1.476$, $p = .146$, $d = 0.209$) 6) vertex vs OPA for degraded scene recognition ($t(49) = -2.520$, $p = .015$, $d = -0.356$) 7) LOC vs OPA for isolated object recognition ($t(49) = -0.717$, $p = .477$, $d = -0.101$) 8) OPA vs LOC for isolated scene recognition ($t(49) = 3.100$, $p = .003$, $d = 0.438$).

Table A1 summarizes the ANOVA results.

Table A1. A summary of a 2x2x3 rm-ANOVA, with task (object, scene), stimulus condition (isolated, degraded), and stimulation site (LOC, OPA, vertex) as within-subject factors. Accuracy is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	49	51.392	<.001	.512
condition	1	49	37.131	<.001	.431
site	2	98	3.965	.022	.075
task * condition	1	49	50.382	<.001	.507
task*site	2	98	2.286	.107	.045
condition * site	2	98	0.336	.716	.007
task * condition * site	2	98	2.120	.125	.041

Reaction Times

In line with the expectations, a significant interaction between task, stimulus condition, and stimulation site was found ($F(2,98) = 3.851, p = .025, \eta_p^2 = .073$). However, contrary to the hypotheses, the simple main effects analysis indicated that participants were slower to respond to degraded as relative to isolated stimuli only in the scene task during vertex stimulation ($F(1,49) = 14.258, p < .001, \eta_p^2 = .225$).

As a follow-up, eight pre-registered pairwise t-tests were also conducted: 1) vertex vs LOC for isolated object recognition ($t(49) = 0.731, p = .469, d = 0.103$) 2) vertex vs LOC for degraded object recognition ($t(49) = -0.143, p = .887, d = -0.020$) 3) vertex vs LOC for degraded scene recognition ($t(49) = 0.221, p = .826, d = 0.031$) 4) vertex vs OPA for isolated scene recognition ($t(49) = -2.087, p = .042, d = -0.295$) 5) vertex vs OPA for degraded object recognition ($t(49) = -0.428, p = .671, d = -0.060$) 6) vertex vs OPA for degraded scene recognition ($t(49) = 0.529, p = .599, d = 0.075$) 7) LOC vs OPA for isolated object recognition ($t(49) = -0.129, p = .898, d = -0.018$) 8) OPA vs LOC for isolated scene recognition ($t(49) = 0.629, p = .532, d = 0.089$). The only significant difference was observed

for isolated scene recognition which was slower during OPA than during vertex stimulation (vertex: 0.868 ± 0.083 s; OPA: 0.893 ± 0.079 s).

Table A2 summarizes the ANOVA results.

Table A2. A summary of a 2x2x3 rm-ANOVA, with task (object, scene), stimulus condition (isolated, degraded), and stimulation site (LOC, OPA, vertex) as within-subject factors. Reaction times are a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	49	333.620	<.001	.872
condition	1	49	7.354	.009	.130
site	2	98	0.040	.961	.001
task * condition	1	49	0.101	0.751	.002
task*site	2	98	0.133	0.875	.003
condition * site	2	98	0.043	0.958	.001
task * condition * site	2	98	3.851	0.025	.073

LISAS

The interaction between task, stimulus condition, and stimulation site was insignificant ($F(2,98) = 2.510$, $p = .086$, $\eta_p^2 = .049$). A significant interaction between task and stimulus condition was found ($F(1,49) = 35.514$, $p < .001$, $\eta_p^2 = .420$). The analysis of simple main effects indicated that performance was significantly worse for objects relative to scenes in both the isolated (object: 1.800 ± 0.458 , scene: 1.105 ± 0.268 ; $F(1,49) = 223.870$, $p < .001$, $\eta_p^2 = .820$) and the degraded stimulus condition (object: 1.560 ± 0.399 , scene: 1.152 ± 0.322 ; $F(1,49) = 83.537$, $p < .001$, $\eta_p^2 = .636$). However, while in the object task, degraded stimuli were recognized better than isolated ones (degraded: 1.560 ± 0.399 , isolated: 1.800 ± 0.458 ; $F(1,49) = 56.050$, $p < .001$, $\eta_p^2 = .534$), in the scene task, there were no significant differences between stimulus conditions (degraded: 1.152 ± 0.322 , isolated: 1.105 ± 0.268 ; $F(1,49) = 2.422$, $p = .126$, $\eta_p^2 = .047$)

In line with the hypotheses, eight pre-registered pairwise t-tests were also conducted: 1) vertex vs LOC for isolated object recognition ($t(49) = -0.451, p = .654, d = -0.064$) 2) vertex vs LOC for degraded object recognition ($t(49) = -0.643, p = .523, d = -0.091$) 3) vertex vs LOC for degraded scene recognition ($t(49) = 1.237, p = .222, d = 0.175$) 4) vertex vs OPA for isolated scene recognition ($t(49) = 0.925, p = .359, d = 0.131$) 5) vertex vs OPA for degraded object recognition ($t(49) = -1.033, p = .307, d = -0.146$) 6) vertex vs OPA for degraded scene recognition ($t(49) = 2.885, p = .006, d = 0.408$) 7) LOC vs OPA for isolated object recognition ($t(49) = 0.413, p = .681, d = 0.058$) 8) OPA vs LOC for isolated scene recognition ($t(49) = -2.188, p = .033, d = -0.309$).

Table A3 summarizes the ANOVA results.

Table A3. A summary of a 2x2x3 rm-ANOVA, with task (object, scene), stimulus condition (isolated, degraded), and stimulation site (LOC, OPA, vertex) as within-subject factors. LISAS is a dependent variable.

Factor	df1	df2	<i>F</i>	<i>p</i>	η_p^2
task	1	49	203.087	<.001	.806
condition	1	49	22.990	<.001	.319
site	2	98	0.870	.422	.017
task * condition	1	49	35.514	<.001	.420
task*site	2	98	2.046	.135	.040
condition * site	2	98	0.801	.452	.016
task * condition * site	2	98	2.510	.086	.049

Table A4. Mean accuracy, mean reaction times, and mean LISAS for each experimental condition.

Task	Condition	Site	Accuracy (M±SD)	RT (s) (M± SD)	LISAS (M ± SD)	N
object	degraded	LOC	0.781 ±0.082	1.131 ±0.190	1.810 ±0.417	50
object	isolated	LOC	0.865 ± 0.075	1.149 ± 0.186	1.564 ± 0.360	50
object	degraded	OPA	0.787 ± 0.079	1.134 ± 0.217	1.795 ± 0.470	50
object	isolated	OPA	0.868 ± 0.069	1.158 ± 0.225	1.581 ± 0.426	50
object	degraded	vertex	0.787 ± 0.089	1.150 ± 0.238	1.791 ± 0.492	50
object	isolated	vertex	0.880 ± 0.064	1.145 ± 0.245	1.535 ± 0.415	50
scene	degraded	LOC	0.876 ± 0.074	0.745 ± 0.146	1.138 ± 0.250	50
scene	isolated	LOC	0.873 ± 0.084	0.759 ± 0.142	1.152 ± 0.277	50
scene	degraded	OPA	0.904 ± 0.053	0.752 ± 0.145	1.072 ± 0.252	50
scene	isolated	OPA	0.893 ± 0.079	0.756 ± 0.140	1.097 ± 0.299	50
scene	degraded	vertex	0.894 ± 0.058	0.731 ± 0.145	1.103 ± 0.300	50
scene	isolated	vertex	0.868 ± 0.083	0.762 ± 0.148	1.207 ± 0.379	50

B. Indoor/outdoor classification of Place365-GoogLeNet Network responses

Table B1. Binary (indoor/outdoor) classification of Places365-GoogLeNet Network outputs.

INDOOR	OUTDOOR
Alcove	Airfield
Aquarium	Alley
Arch	Amusement park
Archive	Army base
Arena (performance)	Bamboo forest
Art gallery	Barn
Attic	Barndoor
Ballroom	Baseball field
Basement	Beach
Bathroom	Botanical garden

Beauty salon

Castle

Bedroom

Cemetery

Bow window (indoor)

Coast

Catacomb

Construction site

Classroom

Cottage

Closet

Crevasse

Coffee shop

Crosswalk

Computer room

Dam

Conference centre

Desert (sand)

Conference room

Field (cultivated)

Corridor

Field (road)

Dining room

Field (wild)

Dressing room

Forest (broadleaf)

Elevator (door)

Forest path

Elevator shaft	Forest road
Florist shop (indoor)	Fountain
Gift shop	Gas station
Greenhouse (indoor)	Glacier
Home office	Golf course
Home theater	Harbor
Hospital room	Highway
Hotel room	Hot spring
Ice skating rink (indoor)	Ice floe
Jail cell	Ice shelf
Jewelry shop	Ice skating rink (outdoor)
Lecture room	Iceberg
Martial arts gym	Igloo
Movie theater (indoor)	Islet

Museum (indoor)

Kennel (outdoor)

Music studio

Lake (natural)

Natural history museum

Landing deck

Nursery

Lighthouse

Office

Mansion

Parking garage (indoor)

Marsh

Patio

Ocean

Pharmacy

Oil rig

Playroom

Orchard

Reception

Parking lot

Recreation room

Pasture

Science museum

Picnic area

Server room

Pier

Shower

Playground

Stage (indoor)

Pond

Staircase

Raceway

Subway station (platform)

Railroad track

Veterinarian's office

Rainforest

Waiting room

River

Wet bar

Ruin

Youth hostel

Runway

Schoolhouse

Shed

Ski slope

Sky

Skyscraper

Slum

Snowfield

Snowy mountain

Street

Tower

Train station (platform)

Tree house

Tundra

Underwater (ocean deep)

Valley

Vegetable garden

Volcano

Water tower

Waterfall

Watering hole

Wheat field

Wind farm

Windmill

Zen garden

7. Publications of the PhD candidate

Jakubowska, N., Dobrowolski, P., **Rutkowska, N.**, Skorko, M., Myśliwiec, M., Michalak, J., & Brzezicka, A. (2021). The role of individual differences in attentional blink phenomenon and real-time-strategy game proficiency. *Heliyon*, 7(4).

Rutkowska, N., Doradzińska, Ł., & Bola, M. (2022). Attentional Prioritization of Complex, Naturalistic Stimuli Maintained in Working-Memory—A Dot-Probe Event-Related Potentials Study. *Frontiers in Human Neuroscience*, 16, 838338.

Okruszek, Ł., **Rutkowska, N.**, Jakubowska, N., & Mąka, S. (2023). Communicative intentions automatically hold attention—evidence from event-related potentials. *Social Neuroscience*, 18(3), 123-131.