Parts and Wholes in Scene Processing

Daniel Kaiser[1,2,3] and Radoslaw M. Cichy[4,5,6]

Abstract

■ During natural vision, our brains are constantly exposed to complex, but regularly structured, environments. Real-world scenes are defined by typical part–whole relationships, where the meaning of the whole scene emerges from configurations of localized information present in individual parts of the scene. Such typical part–whole relationships suggest that information from individual scene parts is not processed independently, but that there are mutual influences between the parts and the whole during scene analysis. Here, we review recent research that used a straightforward, but effective approach to study such mutual influences: By dissecting scenes into multiple arbitrary pieces, these studies provide new insights into how the processing of whole scenes is shaped by their constituent parts and, conversely, how the processing of individual parts is determined by their role within the whole scene. We highlight three facets of this research: First, we discuss studies demonstrating that the spatial configuration of multiple scene parts has a profound impact on the neural processing of the whole scene. Second, we review work showing that cortical responses to individual scene parts are shaped by the context in which these parts typically appear within the environment. Third, we discuss studies demonstrating that missing scene parts are interpolated from the surrounding scene context. Bridging these findings, we argue that efficient scene processing relies on an active use of the scene's part–whole structure, where the visual brain matches scene inputs with internal models of what the world should look like. ■

INTRODUCTION

The ability to efficiently parse visual environments is critical for successful human behavior. Efficient scene analysis is supported by a specialized brain network spanning the occipital and temporal cortices (Epstein & Baker, 2019; Baldassano, Esteva, Fei-Fei, & Beck, 2016; Epstein, 2014). Over the last decade, functional neuroimaging has revealed that this network represents multiple key properties of visual scenes, including basic-level scene category (e.g., a beach vs. a mountain; Walther, Chai, Caddigan, Beck, & Fei-Fei, 2011; Walther, Caddigan, Fei-Fei, & Beck, 2009), high-level visual characteristics of the scene (e.g., how open or cluttered a scene is; Henriksson, Mur, & Kriegeskorte, 2019; Park, Konkle, & Oliva, 2015), and the type of actions that can be performed within a specific environment (e.g., in which directions people can navigate within the scene; Park & Park, 2020; Bonner & Epstein, 2017). Complementary magnetoencephalography (MEG) and EEG studies have shown that many of these properties are computed within only a few hundred milliseconds (Henriksson et al., 2019; Groen et al., 2018; Lowe, Rajsic, Ferber, & Walther, 2018; Cichy, Khosla, Pantazis, & Oliva, 2017), demonstrating that critical scene information is extracted already early during visual analysis in the brain.

Scene analysis inherently relies on the typical part–whole structure of the scene: Many key properties of scenes cannot be determined from localized scene parts alone—they rather become apparent through the analysis of meaningful configurations of features across different parts of the whole scene.[1] Such configurations arise from the typical spatial distribution of low-level visual attributes (Purves, Wojtach, & Lotto, 2011; Geisler, 2008; Torralba & Oliva, 2003), environmental surfaces (Henriksson et al., 2019; Lescroart & Gallant, 2019; Spelke & Lee, 2012), and objects (Castelhano & Krzyś, 2020; Kaiser, Quek, Cichy, & Peelen, 2019; Võ, Boettcher, & Draschkow, 2019). For instance, the navigability of a scene can only be determined by integrating a set of complimentary features that appear in different characteristic parts of the scene (Bonner & Epstein, 2018): The lower parts of the scene convey information about horizontal surfaces near us, which determine our immediate options for navigational movement. Conversely, the upper parts of the scene contain information about more distant obstacles and passageways that are often constrained by vertical boundaries, which determine our subsequent options for navigating the scene. Thus, to successfully analyze the possibilities for navigating the environment, the visual system needs to analyze and integrate different pieces of information across the different scene parts. The need for analyzing such configurations of information across scene parts prompts the hypothesis that scenes and their individual constituent parts are not processed independently. Instead, they mutually influence each other: The representation of a scene should be

[1]Justus-Liebig-Universität Gießen, Germany, [2]Philipps-Universität Marburg, Germany, [3]University of York, United Kingdom, [4]Freie Universität Berlin, Germany, [5]Humboldt-Universität zu Berlin, Germany, [6]Bernstein Centre for Computational Neuroscience Berlin, Germany

Journal of Cognitive Neuroscience 34:1, pp. 4–15
https://doi.org/10.1162/jocn_a_01788

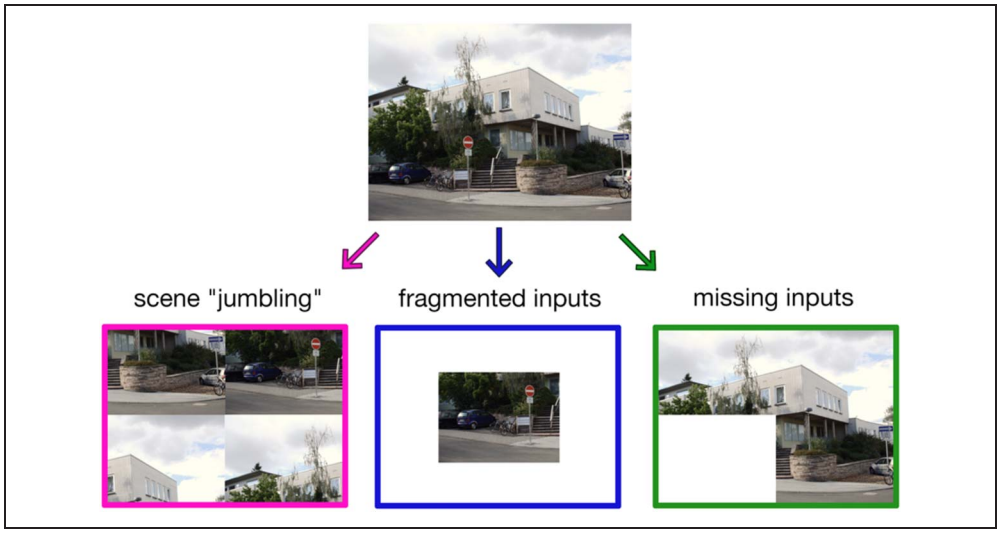

Figure 1. Approaches to studying part–whole relationships in natural scenes. The characteristic distribution of visual information across scenes prompts the hypothesis that the representation of whole scenes and their individual parts hinges on typically experienced part–whole relationships. Here, we review three complimentary approaches to test this hypothesis: (1) studies that used "jumbling" paradigms to investigate the role of coherent multipart structure on scene-selective cortical responses, (2) studies that presented fragmented inputs to infer how the representation of individual scene parts is determined by their role in the full scene, and (3) studies that obscured scene parts to investigate how missing inputs are interpolated by the visual system. Example image obtained from https://de.m.wikipedia.org/wiki/Datei:Giessen_Arndtstrasse_2.png under a Creative Commons License.

determined not only by an independent analysis of its localized parts but also by the way in which these parts are configured across visual space. In turn, the representation of a scene part should not be determined by its visual contents alone, but also by where the part typically appears in the context of the whole scene.

In this review, we will highlight recent research that utilized a simple, yet effective approach to investigate such mutual influences between the whole scene and its constituent parts. In this approach, scene images are dissected into multiple, arbitrary image parts, which can then be recombined into new scenes or presented on their own. Through variations of this straightforward manipulation, researchers have now gained novel insights into how part–whole relationships in natural scenes affect scene analysis in the brain. We will review three facets of this research (Figure 1): First, we will discuss how recent studies that have used "jumbling" paradigms, in which scene parts are systematically shuffled, have revealed the critical role of multipart structure for cortical scene processing. Second, we will review work demonstrating that typical part–whole structure aids the contextualization of individual, fragmented scene parts. Third, we will discuss studies showing that when parts of a scene are missing, the visual brain uses typical part–whole structure to "fill in" information that is currently absent. Synthesizing these findings, we argue that the mutual influences between the whole scene and its constituent parts are well captured by a framework of scene processing in which the visual system actively matches visual inputs with internal models of the world.

MULTIPART STRUCTURE IN SCENE PROCESSING

To reveal how the spatial configuration of scene parts shapes the representation of the whole scene, researchers have used "jumbling" paradigms (Biederman, 1972), in which scenes are dissected into multiple parts that are then either re-assembled into their typical configurations or shuffled to appear in atypical configurations. If the part–whole structure of a scene indeed plays a critical role for its cortical representation, then we should expect that such manipulations profoundly impair scene processing. Classical studies have shown that scene jumbling reduces behavioral performance in scene and object categorization (Biederman, Rabinowitz, Glass, & Stacy, 1974; Biederman, 1972), as well as object recognition within a scene (Biederman, Glass, & Stacy, 1973). More recently, jumbling paradigms have been used to demonstrate that change detection performance benefits from coherent scene structure (Zimmermann, Schnier, & Lappe, 2010; Varakin & Levin, 2008; Yokosawa & Mitsumatsu, 2003). Together, these studies show that scene perception benefits from typical part–whole relationships across the scene.

From such behavioral results, one predicts that responses in scene-selective visual cortex should also be sensitive to part–whole structure. A recent neuroimaging study (Kaiser, Häberle, & Cichy, 2020a) put this prediction to the test. In this study, participants viewed intact and jumbled scenes (Figure 2A) while their brain activity was recorded with fMRI and EEG. Using multivariate classification analysis, the intact and jumbled scenes were discriminable across early and scene-selective cortex (fMRI) and across processing time (EEG), revealing that the visual system is broadly sensitive to the scenes' part–whole structure (Figure 2B). Interestingly, a much greater difference between intact and jumbled scenes was found when the scenes were presented in their upright orientation, compared to when they were presented upside–down. In the fMRI, such an inversion effect was specifically found in scene-selective occipital place area

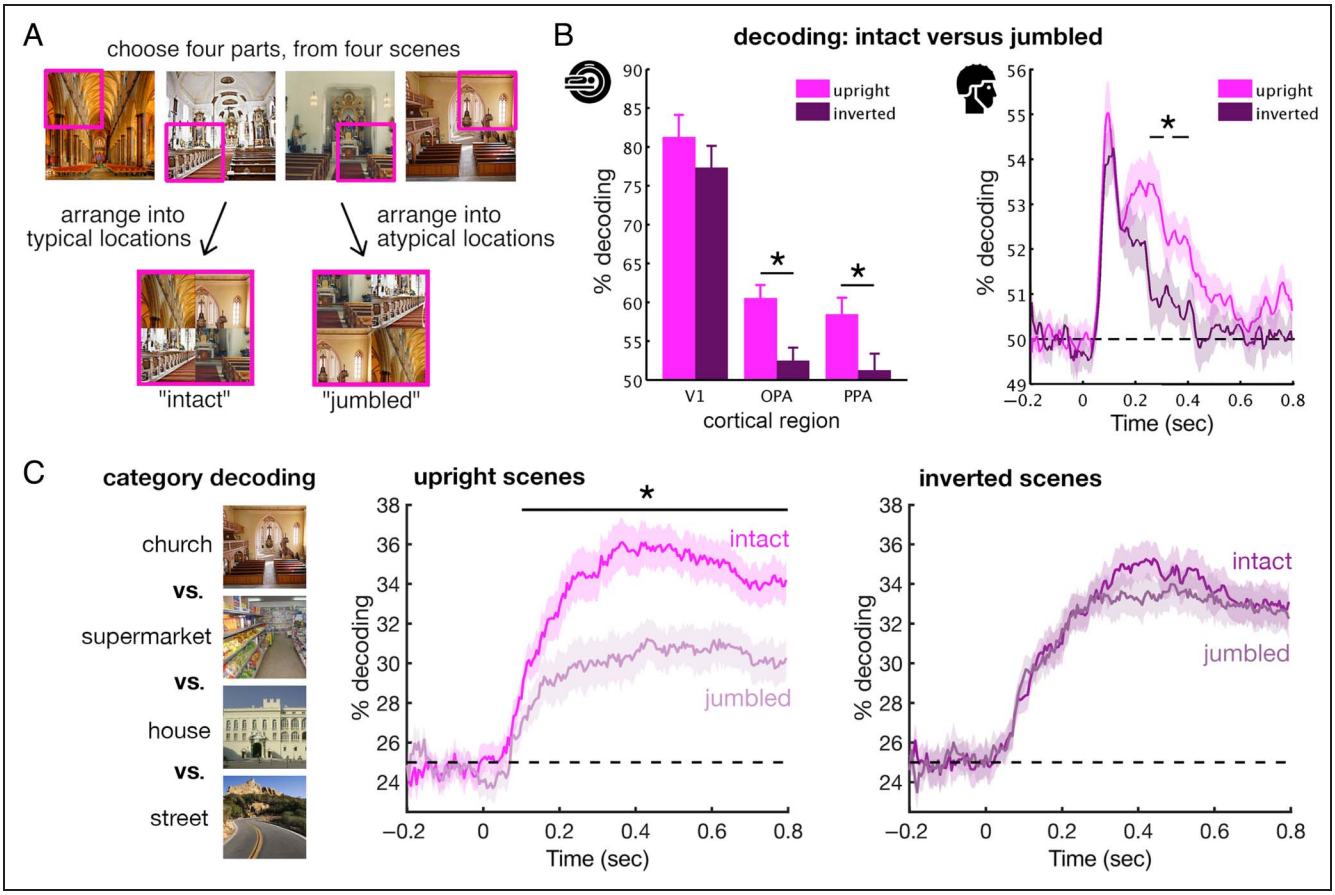

Figure 2. Sensitivity to part–whole structure revealed by scene jumbling. (A) In this study, intact and jumbled versions of natural scenes were created by placing the four quadrants of scenes in their typical or atypical positions. Quadrants were always taken from four different scene exemplars to equate the presence of visual discontinuities. All scenes were presented upright and upside–down. (B) To reveal sensitivity to typical part–whole structure, linear classifiers were used to decode between intact and jumbled scenes, either based on multivoxel fMRI response patterns (left) or on multi-electrode EEG response patterns (right). Intact and jumbled scenes were discriminable across visual regions of interest (fMRI) and in a temporally sustained way (EEG). Critically, this sensitivity to scene structure was more pronounced for upright than inverted scenes in scene-selective regions OPA and parahippocampal place area (PPA), and around 250 msec of processing. (C) In a follow-up EEG study, multivariate decoding was used to track the accumulation of category information over time. Category information was accumulating faster for intact than for jumbled scenes, starting within the first 200 msec of processing. However, when the scenes were inverted, no such effect of scene structure was found. All significance markers donate differences between upright and inverted. Panels were reproduced from Kaiser et al. (2020a, 2020b).

(OPA) and parahippocampal place area, whereas, in the EEG, this difference emerged at around 250 msec, shortly after the time during which scene-selective waveform components are first observed (Harel, Groen, Kravitz, Deouell, & Baker, 2016). The timing and localization of these effects suggests that they occur during the early stages of scene representation in specialized regions of the visual cortex, showing that, already, the initial perceptual coding of a scene—rather than only postperceptual processes such as attentional engagement—is altered depending on the availability of scene structure. The inversion effects therefore indicate that scene-selective responses are fundamentally sensitive to the part–whole structure of scenes that we frequently experience in the world, rather than only to visual differences between intact and jumbled scenes.

Although these results reveal a strong sensitivity to scene structure for typically oriented scenes, it is unclear whether they also index a richer representation of upright and structured scenes. Specifically, does the typical structure of a scene facilitate the analysis of its contents? To resolve this question, a follow-up EEG study (Kaiser, Häberle, & Cichy, 2020b) tested whether coherent scene structure facilitates the emergence of scene category information. In this study, scene category (e.g., whether the participant had seen an image of a church or a supermarket) could indeed be decoded more accurately from EEG response patterns within the first 200 msec of processing when the image was intact than when it was jumbled (Figure 2C). Critically, this benefit was restricted to upright scenes: When the scenes were inverted, category decoding was highly similar for intact and jumbled scenes. This suggests that the scene structure specifically available in intact and upright scenes facilitates the rapid readout of meaningful category information from the scene.

The enhanced representation of typically structured scenes may indicate that the brain integrates information from different parts of the scene, but only when these parts are positioned correctly. On a mechanistic level, this integration of information across the scene may be achieved by neural assemblies that have a shared tuning for both the content of the individual parts and their relative positioning across the scene. Such shared tuning could be prevalent in scene-selective regions of the visual cortex, where neurons' large receptive field coverage (Silson, Chan, Reynolds, Kravitz, & Baker, 2015) enables them to simultaneously receive and integrate information across different parts of the scene. If the neurons are sensitive to the typical multipart structure of the scene, they would specifically integrate information across the scene when the scene is arranged in a typical way. Because of the additional involvement of such neurons in the analysis of typically configured scenes, the resulting scene representation will be qualitatively different from the representations of the individual parts. In fMRI response patterns, such information integration can become apparent in nonlinearities in the way that responses to multiple parts approximate the response to the whole (Kaiser, Quek, et al., 2019; Kubilius, Baeck, Wagemans, & Op de Beeck, 2015): Whenever multiple, unrelated stimuli are presented, the response patterns to the whole display can be predicted by a linear combination of the response patterns to the constituent stimuli (Kliger & Yovel, 2020; MacEvoy & Epstein, 2009). By contrast, when the stimuli form meaningful configurations, response patterns to the whole display become different from the linear combination of individual response patterns. When the meaningful whole is presented, additional tuning to the stimulus configuration cannot be predicted by a linear combination of the individual patterns—although the response patterns to the pairs are themselves reliable across participants. Such integrative effects have been shown in object-selective cortex, for multi-object displays that convey meaningful real-world relationships (Kaiser & Peelen, 2018; Baldassano, Beck, & Fei-Fei, 2017; Baeck, Wagemans, & Op de Beeck, 2013), suggesting that meaningful object groups are indeed represented as a whole rather than independently. Similar conclusions have been reached using fMRI adaptation techniques (Hayworth, Lescroart, & Biederman, 2011). One study has so far looked into multi-object processing within complex scenes (MacEvoy & Epstein, 2011). This study revealed that in object-selective cortex, responses to scenes that contain multiple objects can be approximated by a linear combination of the responses to the individual objects in isolation. By contrast, in scene-selective cortex, the scene response was not well approximated by the same linear combination. This result suggests that object responses are not linearly combined in scene-selective cortex when the objects are part of a complex natural environment. Whether, and to which extent, this absence of an effect can be attributed to integration processes that are enabled by typical multi-object relationships within the scene needs to be investigated in future studies.

More generally, the tuning to typical part–whole structure reinforces the view that the visual system is fundamentally shaped by visual input statistics (Purves et al., 2011). Adaptations to typically structured inputs can be observed across the visual hierarchy, from simple features (Geisler, 2008) to objects (Kaiser, Quek, et al., 2019) and people (Papeo, 2020). The findings reviewed here show that such experience-based adaptations extend to natural scenes. On what timescale these adaptations emerge during development and how flexibly they can be altered during adulthood needs to be addressed in future studies.

In summary, the reviewed findings show that typical part–whole structure plays a critical role in scene representation. They establish that multiple scene parts are represented as a meaningful configuration, rather than as independently coded pieces of information. Next, we turn to studies that probed the representation of individual scene parts and discuss how typical part–whole structure aids the visual system in coping with situations in which only fragments of a scene are available for momentary analysis.

DEALING WITH FRAGMENTED INPUTS

During natural vision, we do not have simultaneous access to all the visual information in our surroundings. Important pieces of information become visible or invisible as we navigate the environment and as we attend to spatially confined pieces of information. At each moment, we therefore only have access to an incomplete snapshot of the world. How are these snapshots put into the context of the current environment? To experimentally mimic this situation, researchers presented individual scene parts of natural scenes in isolation and subsequently looked at how cortical responses to these isolated parts are shaped by the role the parts play in the context of the whole scene.

When individual scene parts are presented on their own, they are not only defined by their content, but they also carry implicit information about where that content typically appears within the environment. As a consequence of the typical part–whole structure of natural scenes, specific scene parts reliably appear in specific parts of the visual field: Skies are more often encountered in the upper regions of the visual field, whereas grass appears in the lower regions of the visual field. To study whether such statistical associations between scene parts and visual-field locations influence processing, researchers probed cortical responses to individual scene parts across the visual field. If the visual system is tuned to the typical positioning of individual scene parts, responses in visual cortex should be stronger when the parts are shown in visual-field positions that

correspond to the positions in which we encounter them during natural vision. In a recent fMRI study (Mannion, 2015), multiple small fragments of a scene were presented in their typical locations in the visual field (e.g., a piece of sky in the upper visual field, in which it is typically encountered when viewed under real-world conditions), or in atypical locations (e.g., a piece of sky in the lower visual field). The positioning of these scene fragments determined activations in retinotopically organized early visual cortex, with stronger overall responses to typically positioned fragments than to atypically positioned ones. Complementary evidence comes from studies that probed the processing of basic visual features that are typically found in specific parts of a scene and thus most often fall into specific parts of the visual field. These studies found that distributions of basic visual features across natural environments are associated with processing asymmetries in visual cortex. For example, low spatial frequencies are more commonly found in the lower visual field, whereas high spatial frequencies are more common in the upper visual field. Following this natural distribution, discrimination of low spatial frequencies is better in the lower visual field, and discrimination of high spatial frequencies is better in the upper visual field; this pattern was associated with response asymmetries in near- and far-preferring columns of visual area V3 (Nasr & Tootell, 2020). Other tentative associations between natural feature distributions and cortical response asymmetries have been reported for cortical visual responses to stimulus orientation (Mannion, McDonald, & Clifford, 2010), texture density (Herde, Uhl, & Rauss, 2020), and surface geometry (Vaziri & Connor, 2016). Together, such findings suggest that areas in retinotopic early visual cortex exhibit a tuning to the typical visual-field location in which parts—and their associated features—appear within the whole scene. To date, such tuning has not been shown for scene-selective regions in high-level visual cortex. However, stronger activations to typically positioned stimuli have been shown in other category-selective regions, such as object-selective lateral occipital cortex (Kaiser & Cichy, 2018) and in face- and body-selective regions of the occipitotemporal cortex (de Haas et al., 2016; Chan, Kravitz, Truong, Arizpe, & Baker, 2010), suggesting that similar tuning properties could also be present in scene-selective areas.

A complementary way to study how the multipart structure of scenes determines the representation of their individual parts is to test how responses to scene parts vary solely as a function of where they *should* appear in the world. Here, instead of experimentally varying the location of scene parts across the visual field, all parts are presented in the same location. The key prediction is that parts stemming from similar real-world locations are coded similarly in the visual system, because they share a link to a common real-world position (Figure 3A). Two recent studies support this prediction (Kaiser, Inciuraite, & Cichy, 2020; Kaiser, Turini, & Cichy, 2019): In these studies, cortical representations are more similar among parts that stem from the same locations along the vertical scene axis, for instance, an image displaying the sky is coded more similarly to an image displaying a ceiling than to an image displaying a floor. Such effects were apparent in fMRI response patterns in scene-selective OPA (Kaiser, Turini, et al., 2019), as well as in EEG response patterns after 200 msec of processing (Kaiser, Inciuraite, et al., 2020; Kaiser, Turini, et al., 2019; Figure 3B). Critically, these effects were not accounted for by low-level visual feature differences between these fragments, such as possible color or orientation differences between scene parts appearing in different locations within the scene context. It rather seems like the brain uses an intrinsic mapping between the varied content of scene parts and their typical real-world locations to sort inputs according to their position within the whole scene. Interestingly, such a sorting is strongly found along the vertical dimension, where information in the world and the behaviors this information affords diverge as a function of distance in visual space (Yang & Purves, 2003; Previc, 1990). Alternatively, statistical regularities may be less prevalent or more subtle along the horizontal dimension: For instance, gravity organizes contents very strongly along the vertical axis, whereas the positioning of objects along the horizontal axis is often arbitrary. To arbitrate between these different accounts, future studies could study situations where the organization along the horizontal axis is indeed meaningful (e.g., in road traffic).

These results point toward an active use of scenes' part–whole structure, whereby the visual system contextualizes inputs with respect to the typical composition of the environment. This contextualization does not just constitute a representational organization by categorical content—it rather constitutes an organization that is based on our typical visual impression of the world. This is consistent with ideas from Bayesian theories of vision, where inputs are interpreted with respect to experience-based priors about the structure of the world (Yuille & Kersten, 2006; Kayser, Körding, & König, 2004). In this case, the observer has a prior of where the current fragment of visual information should appear in the context of the environment, and the representation of the fragment is then determined by this prior: Fragments that yield similar priors for their typical location are consequently coded in a similar way. This representational organization also yields predictions for behavior, where scene fragments stemming from different real-world locations should be better discriminable than those stemming from similar locations.

It is worth noting that many of the part–whole regularities in scenes, such as mutual relationships among objects, are conveyed in a world-centered frame of reference (i.e., they are largely preserved when viewpoints are changed). By contrast, the normalization process discussed here is contingent on a viewer-centered reference frame:

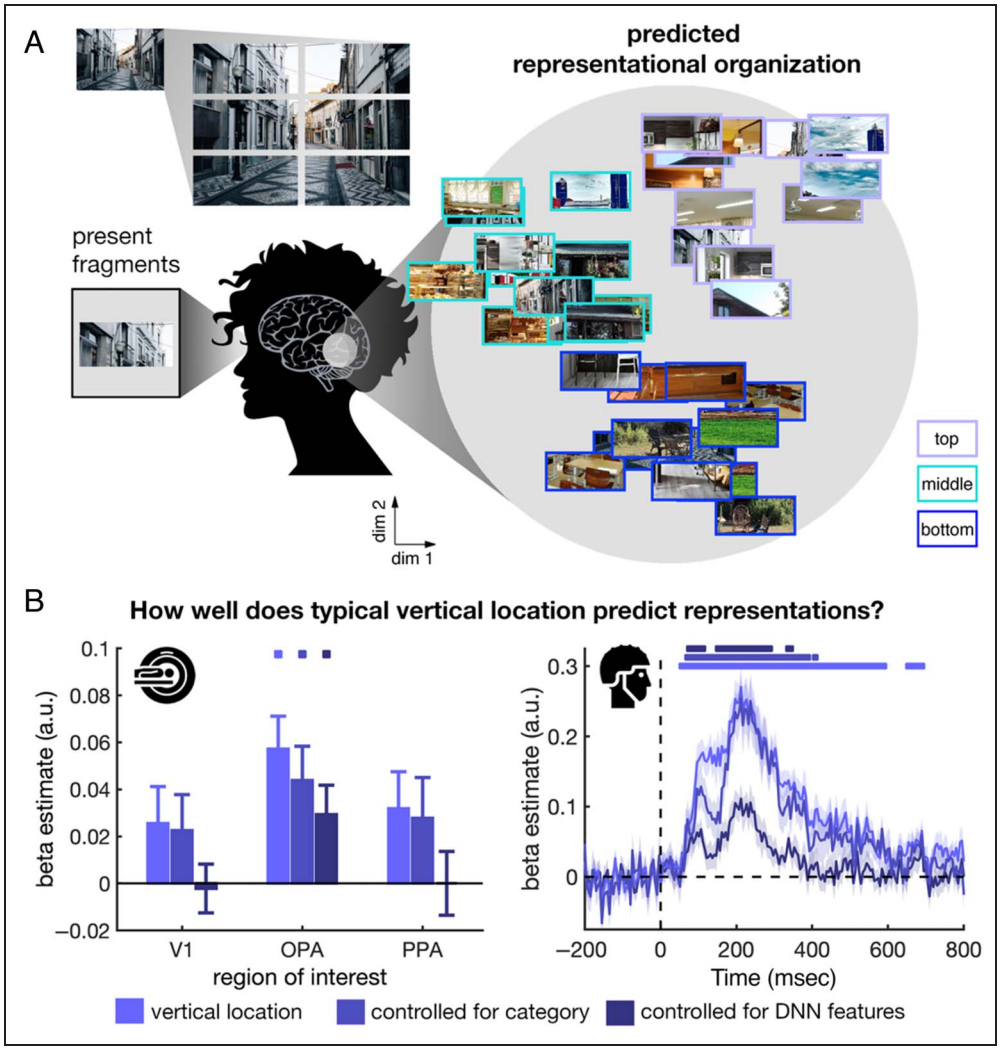

Figure 3. Representations of fragmented inputs are sorted with respect to the scene's part–whole structure. (A) In this study, the incomplete inputs received during natural vision were mimicked by centrally presenting isolated parts from different locations in a scene. The key prediction was that cortical representations of the scene parts would reflect their typical, implicit location within the world, based on how scenes are normally structured. This should be visible in a greater similarity among representations of parts that appear in similar vertical locations within the scenes. (B) To test this prediction, similarities among response patterns in the fMRI (left) and EEG (right) were modelled as a function of the parts' similarity in vertical location. Vertical location predicted responses in the OPA (fMRI) and soon after onset, peaking at 200 msec poststimulus (EEG). As shown, this organization persisted when controlling for category similarity (indoor vs. outdoor scenes), and when controlling for visual features in an artificial deep neural network (DNN) model of visual categorization. These analyses suggest that differences in simple visual features across the scene parts cannot explain the differences in cortical representations. There rather seems to be an intrinsic mapping between visual content and typical real-world locations that persists across a range of visually diverse scenes. Significance markers denote significant prediction of response patterns by vertical location. Panels were reproduced from Kaiser, Turini, et al. (2019).

Fragmented inputs are organized in the visual system in the same way as they are spatially organized in a typical viewer-centered perspective, most likely one that we have experienced pertinently in the past. This normalization process allows us to assemble the whole environment from the different visual snapshots we accumulate over time: By organizing the individual snapshots by their spatial position in the world, it becomes easier to piece them together in a coherent representation of the world around us. Additionally, representing scene information in a typical viewer-centered perspective also allows us to readily make inferences about current behavioral possibilities: For instance, the typical location of an object in the world—rather than its current location in the visual field—offers additional information on which actions can be performed on it. Although normalizing scene inputs to concur with typical real-life views may be a beneficial processing strategy in many everyday situations, it also alters representations so that they become less veridical.

Such alterations of representations become apparent in another pervasive phenomenon in scene perception: In boundary extension, the visual system extrapolates information outside the currently available view of the scene, leading participants to report additional content around the scene when subsequently remembering it (Park, Intraub, Yi, Widders, & Chun, 2007; Intraub, Bender, & Mangels, 1992; Intraub & Richardson, 1989). Interestingly, recent findings show that the degree of boundary extension is stimulus-dependent (Park, Josephs, & Konkle, 2021; Bainbridge & Baker, 2020): For some scene images, their original boundaries are indeed extended during scene recall, whereas, for others, boundaries are contracted. This pattern of results may arise as a consequence of adjusting scene inputs to their typically experienced structure, relative to a typical viewpoint: When the scene view is narrower than typically experienced, boundaries are extended, and when it is wider than typically experienced, boundaries are compressed. This result fits well with an active use of scene structure

in organizing cortical representations, where scene inputs are "normalized" to a typical real-world view. How much this normalization changes as a function of internal states and current task demands needs to be explored in more detail.

Together, these results show that the representation of individual scene parts is actively influenced by their role within the typical part–whole structure of the full scene. If the part–whole structure of scenes indeed influences the representation of their parts, the effect of the whole on the representation of local information should also be apparent when a part of the scene is missing. We turn to research addressing this issue in the next section.

DEALING WITH MISSING INPUTS

The notion that part–whole structure is actively used by the visual system is most explicitly tested in studies that probe visual processing under conditions where inputs from parts of the scene are absent. In such cases, can the visual system use typical scene structure to interpolate the missing content?

We know from neurophysiological studies that neurons in early visual cortex actively exploit context to interpolate the nature of missing inputs (Albright & Stoner, 2002). This is strikingly illustrated by studies of visual "filling-in" (Komatsu, 2006; de Weerd, Gattass, Desimone, & Ungerleider, 1995): For instance, even V1 neurons whose receptive fields are unstimulated display orientation-specific responses, driven by neurons that respond to orientation information in the surrounding spatial context. Such cortical filling-in of information for unstimulated regions of the retina is well established for low-level attributes such as orientation. Can similar cortical filling-in effects from contextual information be observed for high-level contents? If the visual system actively uses information about the part–whole structure of scenes, then we should be able to find neural correlates of a contextual filling-in process, in which the missing input is compensated by a cortical representation of what *should* be there.

A series of recent fMRI studies has probed such contextual effects in scene vision (Morgan, Petro, & Muckli, 2019; Muckli et al., 2015; Smith & Muckli, 2010). In these studies, participants viewed scenes in which a quarter of the image was occluded. Using retinotopic mapping techniques, the authors then measured multivariate response patterns across voxels in early visual cortex that were specifically responsive to the occluded quadrant but not surrounding areas of visual space (Figure 4A). What they found is that these voxels still allowed linear classifiers to discriminate between the different scenes, suggesting that information from the stimulated quadrants leads the visual system to fill in scene-specific information for the unstimulated quadrant (Figure 4B). In another study (Morgan et al., 2019), the authors could show that the information represented in the obscured quadrant concurs with participants' expectations of what should be appearing in this part of the scene: Participants' drawings of the expected content of the occluded quadrant predicted cortical activations in the retinotopically corresponding region of V1.

From where does the interpolated information found in early visual cortex originate? One possibility is that downstream regions in visual cortex provide content-specific feedback to early visual cortex. Using cortical layer-specific analysis of 7 T fMRI recordings, Muckli et al. (2015) provided evidence that such filling-in processes are mediated by top–down connections. By performing decoding analyses across cortical depth, they found that multivoxel response patterns in the superficial layer allowed for discriminating the scenes, even when again looking at only the unstimulated portion of V1 (Figure 4C). In the superficial layer, top–down connections to V1 terminate, suggesting that cortical responses for the missing input are interpolated by means of feedback information from higher cortical areas. This result thus suggests that the typically experienced multipart structure of a scene allows the visual system to actively feed back information that is missing in the input. Although these effects are observed in early visual areas, they are mediated by top–down connections that carry information about which information *should* be there.

What enables the visual brain to feed back the missing information accurately? In low-level feature filling-in, missing information is typically interpolated by means of the surrounding information—the same feature present in the stimulated regions of the visual field is filled into neighboring unstimulated regions (Komatsu, 2006). This mechanism is not sufficient for interpolating missing information in natural scenes, which not only are defined by complex features, but for which these features also vary drastically across different parts of the scene. Missing information thus needs to be interpolated from more downstream regions, presumably from memory and knowledge systems where detailed scene schemata are stored. Candidate regions for schema storage are memory regions of the medial temporal lobe, as well as a recently discovered memory-related system in anterior scene-selective cortex (Steel, Billings, Silson, & Robertson, 2021). Whether these regions indeed feed back missing scene information to early visual cortex needs to be tested in future studies.

CONCLUSION AND OUTLOOK

Together, the recent findings establish that parts and wholes substantially influence each other during scene processing, which suggests that the efficiency of real-world scene vision lends itself to the exploitation of typical distributions of information across the environment. From the reviewed work, we distill out two key conclusions.

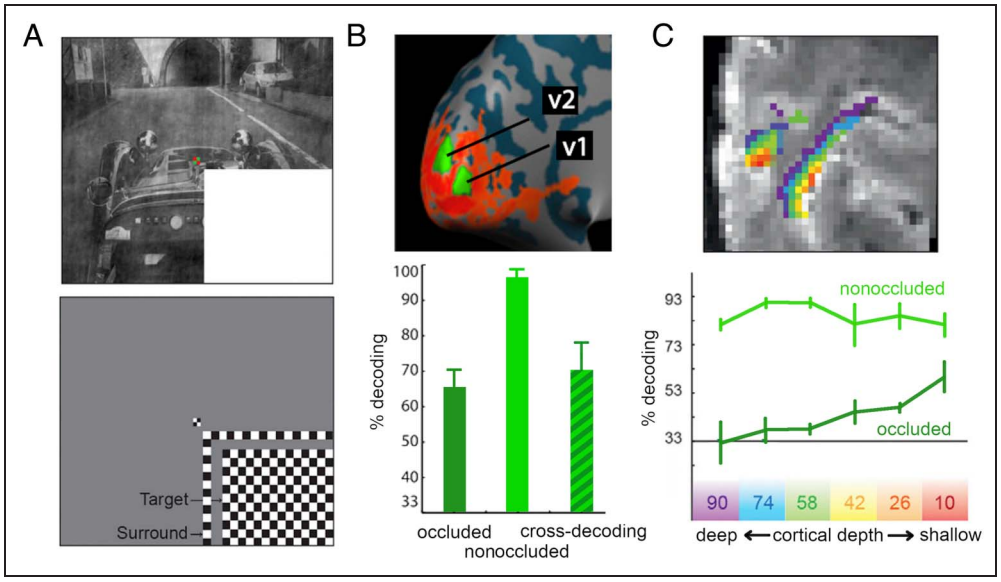

Figure 4. Missing scene inputs are filled in via cortical feedback connections. (A) In a series of fMRI studies, participants viewed scenes, in which one quadrant was occluded. Using retinotopic mapping techniques, the authors could isolate regions of early visual cortex that precisely responded to the occluded area, but not surrounding areas of visual space. (B) Multivoxel patterns across V1 and V2, corresponding to the occluded areas, allowed for accurately decoding between different scenes, both when the area was occluded and when it was fully visible. Cross-decoding analysis revealed that cortical representations in both cases were similar: Classifiers trained on the nonoccluded condition could accurately discriminate the scenes in which the quadrant was missing, suggesting that the filled-in information accurately approximates the information that is actually present. (C) In a high-field fMRI study, these effects were dissected across cortical depth (color-coded). Whereas nonoccluded scenes could be discriminated from V1 response patterns across cortical depth, the occluded scenes were only discriminable from patterns in the superficial layers of V1. As cortical top–down connections terminate in superficial layers of the cortex, this result suggests that missing scene parts are filled in via cortical feedback connections. Panels were reproduced from Smith and Muckli (2010) and Muckli et al. (2015).

First, these studies highlight the importance of typical part–whole structure for cortical processing of natural scenes. When their part–whole structure is broken, scenes are represented less efficiently; when scene parts are presented in isolation, part–whole structure is used to actively contextualize them; and when information from scene parts is missing, typical part–whole structure is used to infer the missing content. These findings are reminiscent of similar findings in the brain's face and body processing systems, in which neurons are tuned to typical part–whole configurations (Brandman & Yovel, 2016; Liu, Harris, & Kanwisher, 2010), and where representations of individual face and body parts are determined by their role in the full face or body, respectively (de Haas, Sereno, & Schwarzkopf, 2021; de Haas et al., 2016; Henriksson, Mur, & Kriegeskorte, 2015; Chan et al., 2010). The current work therefore suggests a similarity between the analysis of the "parts and wholes" in face recognition (Tanaka & Simonyi, 2016) and scene processing, and hints toward a configural mode of processing in the scene network that needs to be explored further. Contrary to faces and bodies, however, the individual parts of a scene are not so straightforward to define, and the reviewed work has used an arguably quite coarse approach to define arbitrary parts of a scene. In reality, scenes vary in more intricate ways and across a multitude of dimensions, including typical distributions of low- and mid-level scene properties (Groen, Silson, & Baker, 2017; Nasr, Echavarria, & Tootell, 2014; Watson, Hartley, & Andrews, 2014), the category and locations of objects contained in the scene (Bilalić, Lindig, & Turella, 2019; Kaiser, Stein, & Peelen, 2014; Kim & Biederman, 2011), relationships between objects and the scene context (Faivre, Dubois, Schwartz, & Mudrik, 2019; Preston, Guo, Das, Giesbrecht, & Eckstein, 2013; Võ & Wolfe, 2013; Mudrik, Lamy, & Deouell, 2010), and scene geometry (Henriksson et al., 2019; Lescroart & Gallant, 2019; Harel, Kravitz, & Baker, 2013; Kravitz, Peng, & Baker, 2011). At this point, a systematic investigation of how regularities across these dimensions contribute to efficient information analysis across natural scenes is still lacking. Another defining aspect of face perception is that it is sensitive not only to the relative positioning of different face features but also to their precise distances (Maurer, Le Grand, & Mondloch, 2002). This distance-based feature organization is also apparent in responses in the face processing network (Henriksson et al., 2015; Loffler, Yourganov, Wilkinson, & Wilson, 2005). In our recent study (Kaiser, Turini, et al., 2019), we have shown that also the typical Euclidean distance between coarse scene parts can explain the representational organization of the individual scene parts presented in isolation. Whether more fine-grained typical distances between different scene elements (e.g., distances between individual objects) similarly shape representations in scene-selective visual cortex needs further investigation.

Second, the reviewed findings support a view on which scene vision is accomplished by matching sensory inputs with internal models of the world, derived from our experience with natural scene structure. This idea has first been highlighted by schema theories (Mandler, 1984; Biederman, Mezzanotte, & Rabinowitz, 1982), which

assume that the brain maintains internal representations that carry knowledge of the typical composition of real-world environments. More recently, theories of Bayesian inference reinforced this view, suggesting that priors about the statistical composition of the world determine the representation of visual inputs (Yuille & Kersten, 2006; Kayser et al., 2004). The reviewed studies indeed suggest that the coding of fragmented and incomplete inputs is constrained by the typical part–whole structure of scenes. On a mechanistic level, this process may be implemented through active mechanisms of neural prediction: Efficient coding of scenes may be achieved by a convergence between the bottom–up input and top–down predictions about the structure of this input (Keller & Mrsic-Flogel, 2018; Clark, 2013; Huang & Rao, 2011). Establishing the precise mechanisms that govern this convergence is a key challenge for future research. Empirical results with simple visual stimuli suggest that expected stimuli can be processed efficiently because top–down predictions suppress sensory signals that are inconsistent with current expectations, leading to a sharpening of neural responses (de Lange, Heilbron, & Kok, 2018; Kok, Jehee, & de Lange, 2012). However, what needs further exploration is how the brain balances between the need for efficiently processing expected inputs and the complimentary need for detecting novel and unexpected stimuli that violate our expectations—after all, reacting fast and accurately to the unexpected is critical in many real-life situations (e.g., while driving). To find the right balance between favoring the expected and the novel, the brain may dynamically adjust the relative weights assigned to visual inputs and to top–down predictions, for example, based on current internal mental states (Herz, Baror, & Bar, 2020) and the precision of both the visual input and our predictions in a given situation (Yon & Frith, 2021). A recent complimentary account suggests that during the perceptual processing cascade, processing is, in turn, biased toward the expected and then the surprising (Press, Kok, & Yon, 2020). When and how natural vision is biased toward the expected structure of the world and toward novel, unexpected information and how this balance is controlled on a neural level are exciting questions for future investigation.

In summary, our review highlights that the cortical scene processing system analyzes the meaning of natural scenes by strongly considering their typical part–whole structure. The reviewed research also highlights that natural vision is an active process that strongly draws from prior knowledge about the world. By further scrutinizing this process, future research can bring us closer to successfully modelling and predicting perceptual efficiency in real-life situations.

Acknowledgments

D. K. and R. M. C. are supported by Deutsche Forschungsgemeinschaft grants (KA4683/2-1, CI241/1-1, CI241/3-1, CI241/7-1). R. M. C. is supported by a European Research Council Starting Grant (ERC-2018-StG 803370). The authors declare no competing interests exist.

Reprint requests should be sent to Daniel Kaiser, Mathematical Institute, Justus-Liebig-University Gießen, Arndtstraße 2, 35392 Gießen, Germany, or via e-mail: danielkaiser.net@gmail.com.

Author Contributions

Daniel Kaiser: Conceptualization; Funding acquisition; Project administration; Visualization; Writing—Original draft; Writing—Review & editing. Radoslaw M. Cichy: Conceptualization; Funding acquisition; Writing—Review & editing.

Funding Information

Daniel Kaiser and Radoslaw M. Cichy, Deutsche Forschungsgemeinschaft (https://dx.doi.org/10.13039/501100001659), grant numbers: KA4683/2-1, CI241/1-1, CI241/3-1, CI241/7-1. Radoslaw M. Cichy, H2020 European Research Council (https://dx.doi.org/10.13039/100010663), grant number: ERC-2018-StG 803370.

Diversity in Citation Practices

A retrospective analysis of the citations in every article published in this journal from 2010 to 2020 has revealed a persistent pattern of gender imbalance: Although the proportions of authorship teams (categorized by estimated gender identification of first author/last author) publishing in the *Journal of Cognitive Neuroscience* (*JoCN*) during this period were M(an)/M = .408, W(oman)/M = .335, M/W = .108, and W/W = .149, the comparable proportions for the articles that these authorship teams cited were M/M = .579, W/M = .243, M/W = .102, and W/W = .076 (Fulvio et al., *JoCN*, 33:1, pp. 3–7). Consequently, *JoCN* encourages all authors to consider gender balance explicitly when selecting which articles to cite and gives them the opportunity to report their article's gender citation balance.

Note

1. In everyday situations, humans do not have visual access to the whole environment. In the following, when we talk about "wholes" in scene perception, we refer to a typical full-field scene input that we experience during our everyday lives.

REFERENCES

Albright, T. D., & Stoner, G. R. (2002). Contextual influences on visual processing. *Annual Review of Neuroscience*, *25*, 339–379. https://doi.org/10.1146/annurev.neuro.25.112701.142900, PubMed: 12052913

Baeck, A., Wagemans, J., & Op de Beeck, H. P. (2013). The distributed representation of random and meaningful object pairs in human occipitotemporal cortex: The weighted average as a general rule. *Neuroimage*, *70*, 37–47. https://doi.org/10.1016/j.neuroimage.2012.12.023, PubMed: 23266747

Bainbridge, W. A., & Baker, C. I. (2020). Boundaries extend and contract in scene memory depending on image properties. *Current Biology*, *30*, 537–543. https://doi.org/10.1016/j.cub.2019.12.004, PubMed: 31983637

Baldassano, C., Beck, D. M., & Fei-Fei, L. (2017). Human–object interactions are more than the sum of their parts. *Cerebral Cortex*, *27*, 2276–2288. https://doi.org/10.1093/cercor/bhw077, PubMed: 27073216

Baldassano, C., Esteva, A., Fei-Fei, L., & Beck, D. M. (2016). Two distinct scene-processing networks connecting vision and memory. *eNeuro*, *3*, ENEURO.0178-16.2016. https://doi.org/10.1523/ENEURO.0178-16.2016, PubMed: 27822493

Biederman, I. (1972). Perceiving real-world scenes. *Science*, *177*, 77–80. https://doi.org/10.1126/science.177.4043.77, PubMed: 5041781

Biederman, I., Glass, A. L., & Stacy, E. W., Jr. (1973). Searching for objects in real-world scenes. *Journal of Experimental Psychology*, *97*, 22–27. https://doi.org/10.1037/h0033776, PubMed: 4704195

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*, 143–177. https://doi.org/10.1016/0010-0285(82)90007-X, PubMed: 7083801

Biederman, I., Rabinowitz, J. C., Glass, A. L., & Stacy, E. W., Jr. (1974). On the information extracted from a glance at a scene. *Journal of Experimental Psychology*, *103*, 597–600. https://doi.org/10.1037/h0037158, PubMed: 4448962

Bilalić, M., Lindig, T., & Turella, L. (2019). Parsing rooms: The role of the PPA and RSC in perceiving object relations and spatial layout. *Brain Structure and Function*, *224*, 2505–2524. https://doi.org/10.1007/s00429-019-01901-0, PubMed: 31317256

Bonner, M. F., & Epstein, R. A. (2017). Coding of navigational affordances in the human visual system. *Proceedings of the National Academy of Sciences, U.S.A.*, *114*, 4793–4798. https://doi.org/10.1073/pnas.1618228114, PubMed: 28416669

Bonner, M. F., & Epstein, R. A. (2018). Computational mechanisms underlying cortical responses to the affordance properties of visual scenes. *PLoS Computational Biology*, *14*, e1006111. https://doi.org/10.1371/journal.pcbi.1006111, PubMed: 29684011

Brandman, T., & Yovel, G. (2016). Bodies are represented as wholes rather than their sum of parts in the occipito-temporal cortex. *Cerebral Cortex*, *26*, 530–543. https://doi.org/10.1093/cercor/bhu205, PubMed: 25217470

Castelhano, M. S., & Krzyś, K. (2020). Rethinking space: A review of perception, attention, and memory in scene processing. *Annual Review of Vision Science*, *6*, 563–586. https://doi.org/10.1146/annurev-vision-121219-081745, PubMed: 32491961

Chan, A. W.-Y., Kravitz, D. J., Truong, S., Arizpe, J., & Baker, C. I. (2010). Cortical representations of bodies and faces are strongest in commonly experienced configurations. *Nature Neuroscience*, *13*, 417–418. https://doi.org/10.1038/nn.2502, PubMed: 20208528

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, PubMed: 27039703

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences*, *36*, 181–204. https://doi.org/10.1017/S0140525X12000477, PubMed: 23663408

de Haas, B., Schwarzkopf, D. S., Alvarez, I., Lawson, R. P., Henriksson, L., Kriegeskorte, N., et al. (2016). Perception and processing of faces in the human brain is tuned to typical feature locations. *Journal of Neuroscience*, *36*, 9289–9302. https://doi.org/10.1523/JNEUROSCI.4131-14.2016, PubMed: 27605606

de Haas, B., Sereno, M. I., & Schwarzkopf, D. S. (2021). Inferior occipital gyrus is organized along common gradients of spatial and face-part selectivity. *Journal of Neuroscience*, *41*, 5511–5521. https://doi.org/10.1523/JNEUROSCI.2415-20.2021, PubMed: 34016715

de Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in Cognitive Sciences*, *22*, 764–779. https://doi.org/10.1016/j.tics.2018.06.002, PubMed: 30122170

de Weerd, P., Gattass, R., Desimone, R., & Ungerleider, L. G. (1995). Responses of cells in monkey visual cortex during perceptual filling-in of an artificial scotoma. *Nature*, *377*, 731–734. https://doi.org/10.1038/377731a0, PubMed: 7477262

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

Epstein, R. A., & Baker, C. I. (2019). Scene perception in the human brain. *Annual Review of Vision Science*, *5*, 373–397. https://doi.org/10.1146/annurev-vision-091718-014809, PubMed: 31226012

Faivre, N., Dubois, J., Schwartz, N., & Mudrik, L. (2019). Imaging object–scene relations processing in visible and invisible natural scenes. *Scientific Reports*, *9*, 4567. https://doi.org/10.1038/s41598-019-38654-z, PubMed: 30872607

Geisler, W. S. (2008). Visual perception and the statistical properties of natural scenes. *Annual Review of Psychology*, *59*, 167–192. https://doi.org/10.1146/annurev.psych.58.110405.085632, PubMed: 17705683

Groen, I. I. A., Greene, M. R., Baldassano, C., Fei-Fei, L., Beck, D. M., & Baker, C. I. (2018). Distinct contributions of functional and deep neural network features to representational similarity of scenes in human brain and behavior. *eLife*, *7*, e32962. https://doi.org/10.7554/eLife.32962, PubMed: 29513219

Groen, I. I. A., Silson, E. H., & Baker, C. I. (2017). Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *372*, 20160102. https://doi.org/10.1098/rstb.2016.0102, PubMed: 28044013

Harel, A., Groen, I. I. A., Kravitz, D. J., Deouell, L. Y., & Baker, C. I. (2016). The temporal dynamics of scene processing: A multifaceted EEG investigation. *eNeuro*, *3*, ENEURO.0139-16.2016. https://doi.org/10.1523/ENEURO.0139-16.2016, PubMed: 27699208

Harel, A., Kravitz, D. J., & Baker, C. I. (2013). Deconstructing visual scenes in cortex: Gradients of object and spatial layout information. *Cerebral Cortex*, *23*, 947–957. https://doi.org/10.1093/cercor/bhs091, PubMed: 22473894

Hayworth, K. J., Lescroart, M. D., & Biederman, I. (2011). Neural encoding of relative position. *Journal of Experimental Psychology: Human Perception and Performance*, *37*, 1032–1050. https://doi.org/10.1037/a0022338, PubMed: 21517211

Henriksson, L., Mur, M., & Kriegeskorte, N. (2015). Faciotopy—A face-feature map with face-like topology in the human occipital face area. *Cortex*, *72*, 156–167. https://doi.org/10.1016/j.cortex.2015.06.030, PubMed: 26235800

Henriksson, L., Mur, M., & Kriegeskorte, N. (2019). Rapid invariant encoding of scene layout in human OPA. *Neuron*, *103*, 161–171. https://doi.org/10.1016/j.neuron.2019.04.014, PubMed: 31097360

Herde, L., Uhl, J., & Rauss, K. (2020). Anatomic and functional asymmetries interactively shape human early visual cortex responses. *Journal of Vision*, *20*, 3. https://doi.org/10.1167/jov.20.6.3, PubMed: 32503040

Herz, N., Baror, S., & Bar, M. (2020). Overarching states of mind. *Trends in Cognitive Sciences*, *24*, 184–199. https://doi.org/10.1016/j.tics.2019.12.015, PubMed: 32059121
Huang, Y., & Rao, R. P. N. (2011). Predictive coding. *Wiley Interdisciplinary Reviews: Cognitive Science*, *2*, 580–593. https://doi.org/10.1002/wcs.142, PubMed: 26302308
Intraub, H., Bender, R. S., & Mangels, J. A. (1992). Looking at pictures but remembering scenes. *Journal of Exprimental Psychology: Learning, Memory, and Cognition*, *18*, 180–191. https://doi.org/10.1037/0278-7393.18.1.180, PubMed: 1532019
Intraub, H., & Richardson, M. (1989). Wide-angle memories of close-up scenes. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 179–187. https://doi.org/10.1037/0278-7393.15.2.179, PubMed: 2522508
Kaiser, D., & Cichy, R. M. (2018). Typical visual-field locations enhance processing in object-selective channels of human occipital cortex. *Journal of Neurophysiology*, *120*, 848–853. https://doi.org/10.1152/jn.00229.2018, PubMed: 29766762
Kaiser, D., Häberle, G., & Cichy, R. M. (2020a). Cortical sensitivity to natural scene structure. *Human Brain Mapping*, *41*, 1286–1295. https://doi.org/10.1002/hbm.24875, PubMed: 31758632
Kaiser, D., Häberle, G., & Cichy, R. M. (2020b). Real-world structure facilitates the rapid emergence of scene category information in visual brain signals. *Journal of Neurophysiology*, *124*, 145–151. https://doi.org/10.1152/jn.00164.2020, PubMed: 32519577
Kaiser, D., Inciuraite, G., & Cichy, R. M. (2020). Rapid contextualization of fragmented scene information in the human visual system. *Neuroimage*, *219*, 117045. https://doi.org/10.1016/j.neuroimage.2020.117045, PubMed: 32540354
Kaiser, D., Quek, G. L., Cichy, R. M., & Peelen, M. V. (2019). Object vision in a structured world. *Trends in Cognitive Sciences*, *23*, 672–685. https://doi.org/10.1016/j.tics.2019.04.013, PubMed: 31147151
Kaiser, D., & Peelen, M. V. (2018). Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. *Neuroimage*, *169*, 334–341. https://doi.org/10.1016/j.neuroimage.2017.12.065, PubMed: 29277645
Kaiser, D., Stein, T., & Peelen, M. V. (2014). Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. *Proceedings of the National Academy of Sciences, U.S.A.*, *111*, 11217–11222. https://doi.org/10.1073/pnas.1400559111, PubMed: 25024190
Kaiser, D., Turini, J., & Cichy, R. M. (2019). A neural mechanism for contextualizing fragmented inputs during naturalistic vision. *eLife*, *8*, e48182. https://doi.org/10.7554/eLife.48182, PubMed: 31596234
Kayser, C., Körding, K. P., & König, P. (2004). Processing of complex stimuli and natural scenes in the visual cortex. *Current Opinion in Neurobiology*, *14*, 468–473. https://doi.org/10.1016/j.conb.2004.06.002, PubMed: 15302353
Keller, G. B., & Mrsic-Flogel, T. D. (2018). Predictive processing: A canonical cortical computation. *Neuron*, *100*, 424–435. https://doi.org/10.1016/j.neuron.2018.10.003, PubMed: 30359606
Kim, J. G., & Biederman, I. (2011). Where do objects become scenes? *Cerebral Cortex*, *21*, 1738–1746. https://doi.org/10.1093/cercor/bhq240, PubMed: 21148087
Kliger, L., & Yovel, G. (2020). The functional organization of high-level visual cortex determines the representation of complex visual stimuli. *Journal of Neuroscience*, *40*, 7545–7558. https://doi.org/10.1523/JNEUROSCI.0446-20.2020, PubMed: 32859715
Kok, P., Jehee, J. F. M., & de Lange, F. P. (2012). Less is more: Expectation sharpens representations in the primary visual cortex. *Neuron*, *75*, 265–270. https://doi.org/10.1016/j.neuron.2012.04.034, PubMed: 22841311
Komatsu, H. (2006). The neural mechanisms of perceptual filling-in. *Nature Reviews Neuroscience*, *7*, 220–231. https://doi.org/10.1038/nrn1869, PubMed: 16495943
Kravitz, D. J., Peng, C. S., & Baker, C. I. (2011). Real-world scene representations in high-level visual cortex: It's the spaces more than the places. *Journal of Neuroscience*, *31*, 7322–7333. https://doi.org/10.1523/JNEUROSCI.4588-10.2011, PubMed: 21593316
Kubilius, J., Baeck, A., Wagemans, J., & Op de Beeck, H. P. (2015). Brain-decoding fMRI reveals how wholes relate to the sum of parts. *Cortex*, *72*, 5–14. https://doi.org/10.1016/j.cortex.2015.01.020, PubMed: 25771992
Lescroart, M. D., & Gallant, J. L. (2019). Human scene-selective areas represent 3D configurations of surfaces. *Neuron*, *101*, 178–192. https://doi.org/10.1016/j.neuron.2018.11.004, PubMed: 30497771
Liu, J., Harris, A., & Kanwisher, N. (2010). Perception of face parts and face configurations: An fMRI study. *Journal of Cognitive Neuroscience*, *22*, 203–211. https://doi.org/10.1162/jocn.2009.21203, PubMed: 19302006
Loffler, G., Yourganov, G., Wilkinson, F., & Wilson, H. R. (2005). fMRI evidence for the neural representation of faces. *Nature Neuroscience*, *8*, 1386–1391. https://doi.org/10.1038/nn1538, PubMed: 16136037
Lowe, M. X., Rajsic, J., Ferber, S., & Walther, D. B. (2018). Discriminating scene categories from brain activity within 100 milliseconds. *Cortex*, *106*, 275–287. https://doi.org/10.1016/j.cortex.2018.06.006, PubMed: 30037637
MacEvoy, S. P., & Epstein, R. A. (2009). Decoding the representation of multiple simultaneous objects in human occipitotemporal cortex. *Current Biology*, *19*, 943–947. https://doi.org/10.1016/j.cub.2009.04.020, PubMed: 19446454
MacEvoy, S. P., & Epstein, R. A. (2011). Constructing scenes from objects in human occipitotemporal cortex. *Nature Neuroscience*, *14*, 1323–1329. https://doi.org/10.1038/nn.2903, PubMed: 21892156
Mandler, J. M. (1984). *Stories, scripts and scenes: Aspects of schema theory*. Hillsdale, NJ: Lawrence Erlbaum Associates Inc.
Mannion, D. J. (2015). Sensitivity to the visual field origin of natural image patches in human low-level visual cortex. *PeerJ*, *3*, e1038. https://doi.org/10.7717/peerj.1038, PubMed: 26131378
Mannion, D. J., McDonald, J. S., & Clifford, C. W. G. (2010). Orientation anisotropies in human visual cortex. *Journal of Neurophysiology*, *103*, 3465–3471. https://doi.org/10.1152/jn.00190.2010, PubMed: 20410358
Maurer, D., Le Grand, R., & Mondloch, C. J. (2002). The many faces of configural processing. *Trends in Cognitive Sciences*, *6*, 255–260. https://doi.org/10.1016/S1364-6613(02)01903-4, PubMed: 12039607
Morgan, A. T., Petro, L. S., & Muckli, L. (2019). Scene representations conveyed by cortical feedback to early visual cortex can be described by line drawings. *Journal of Neuroscience*, *39*, 9410–9423. https://doi.org/10.1523/JNEUROSCI.0852-19.2019, PubMed: 31611306
Muckli, L., De Martino, F., Vizioli, L., Petro, L. S., Smith, F. W., Ugurbil, K., et al. (2015). Contextual feedback to superficial layers of V1. *Current Biology*, *25*, 2690–2695. https://doi.org/10.1016/j.cub.2015.08.057, PubMed: 26441356
Mudrik, L., Lamy, D., & Deouell, L. Y. (2010). ERP evidence for context congruity effects during simultaneous object–scene processing. *Neuropsychologia*, *48*, 507–517. https://doi.org/10.1016/j.neuropsychologia.2009.10.011, PubMed: 19837103

Nasr, S., Echavarria, C. E., & Tootell, R. B. H. (2014). Thinking outside the box: Rectilinear shapes selectively activate scene-selective cortex. *Journal of Neuroscience*, *34*, 6721–6735. https://doi.org/10.1523/JNEUROSCI.4802-13.2014, PubMed: 24828628

Nasr, S., & Tootell, R. B. H. (2020). Asymmetries in global perception are represented in near- versus far-preferring clusters in human visual cortex. *Journal of Neuroscience*, *40*, 355–368. https://doi.org/10.1523/JNEUROSCI.2124-19.2019, PubMed: 31744860

Papeo, L. (2020). Twos in human visual perception. *Cortex*, *132*, 473–478. https://doi.org/10.1016/j.cortex.2020.06.005, PubMed: 32698947

Park, J., Josephs, E., & Konkle, T. (2021). Systematic transition from boundary extension to contraction along an object–scene continuum. *Journal of Vision*, *21*, 2124. https://doi.org/10.1167/jov.21.9.2124

Park, J., & Park, S. (2020). Coding of navigational distance and functional constraint of boundaries in the human scene-selective cortex. *Journal of Neuroscience*, *40*, 3621–3630. https://doi.org/10.1523/JNEUROSCI.1991-19.2020, PubMed: 32209608

Park, S., Intraub, H., Yi, D.-J., Widders, D., & Chun, M. M. (2007). Beyond the edges of a view: Boundary extension in human scene-selective visual cortex. *Neuron*, *54*, 335–342. https://doi.org/10.1016/j.neuron.2007.04.006, PubMed: 17442252

Park, S., Konkle, T., & Oliva, A. (2015). Parametric coding of the size and clutter of natural scenes in the human brain. *Cerebral Cortex*, *25*, 1792–1805. https://doi.org/10.1093/cercor/bht418, PubMed: 24436318

Press, C., Kok, P., & Yon, D. (2020). The perceptual prediction paradox. *Trends in Cognitive Sciences*, *24*, 13–24. https://doi.org/10.1016/j.tics.2019.11.003, PubMed: 31787500

Preston, T. J., Guo, F., Das, K., Giesbrecht, B., & Eckstein, M. P. (2013). Neural representations of contextual guidance in visual search of real-world scenes. *Journal of Neuroscience*, *33*, 7846–7855. https://doi.org/10.1523/JNEUROSCI.5840-12.2013, PubMed: 23637176

Previc, F. H. (1990). Functional specialization in the lower and upper visual fields in humans: Its ecological origins and neurophysiological implications. *Behavioral and Brain Sciences*, *13*, 519–542. https://doi.org/10.1017/S0140525X00080018

Purves, D., Wojtach, W. T., & Lotto, R. B. (2011). Understanding vision in wholly empirical terms. *Proceedings of the National Academy of Sciences, U.S.A.*, *108(Suppl. 3)*, 15588–15595. https://doi.org/10.1073/pnas.1012178108, PubMed: 21383192

Silson, E. H., Chan, A. W.-Y., Reynolds, R. C., Kravitz, D. J., & Baker, C. I. (2015). A retinotopic basis for the division of high-level scene processing between lateral and ventral human occipitotemporal cortex. *Journal of Neuroscience*, *35*, 11921–11935. https://doi.org/10.1523/JNEUROSCI.0137-15.2015, PubMed: 26311774

Smith, F. W., & Muckli, L. (2010). Nonstimulated early visual areas carry information about surrounding context. *Proceedings of the National Academy of Sciences, U.S.A.*, *107*, 20099–20103. https://doi.org/10.1073/pnas.1000233107, PubMed: 21041652

Spelke, E. S., & Lee, S. A. (2012). Core systems of geometry in animal minds. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, *367*, 2784–2793. https://doi.org/10.1098/rstb.2012.0210, PubMed: 22927577

Steel, A., Billings, M. M., Silson, E. H., & Robertson, C. E. (2021). A network linking scene perception and spatial memory systems in posterior cerebral cortex. *Nature Communications*, *12*, 2632. https://doi.org/10.1038/s41467-021-22848-z, PubMed: 33976141

Tanaka, J. W., & Simonyi, D. (2016). The "parts and wholes" of face recognition: A review of the literature. *Quarterly Journal of Experimental Psychology*, *69*, 1876–1889. https://doi.org/10.1080/17470218.2016.1146780, PubMed: 26886495

Torralba, A., & Oliva, A. (2003). Statistics of natural image categories. *Network: Computation in Neural Systems*, *14*, 391–412. https://doi.org/10.1088/0954-898X_14_3_302, PubMed: 12938764

Varakin, D. A., & Levin, D. T. (2008). Scene structure enhances change detection. *Quarterly Journal of Experimental Psychology*, *61*, 543–551. https://doi.org/10.1080/17470210701774176, PubMed: 18300186

Vaziri, S., & Connor, C. E. (2016). Representation of gravity-aligned scene structure in ventral pathway visual cortex. *Current Biology*, *26*, 766–774. https://doi.org/10.1016/j.cub.2016.01.022, PubMed: 26923785

Võ, M. L.-H., Boettcher, S. E. P., & Draschkow, D. (2019). Reading scenes: How scene grammar guides attention and aids perception in real-world environments. *Current Opinion in Psychology*, *29*, 205–210. https://doi.org/10.1016/j.copsyc.2019.03.009, PubMed: 31051430

Võ, M. L.-H., & Wolfe, J. M. (2013). Differential electrophysiological signatures of semantic and syntactic scene processing. *Psychological Science*, *24*, 1816–1823. https://doi.org/10.1177/0956797613476955, PubMed: 23842954

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*, 10573–10581. https://doi.org/10.1523/JNEUROSCI.0559-09.2009, PubMed: 19710310

Walther, D. B., Chai, B., Caddigan, E., Beck, D. M., & Fei-Fei, L. (2011). Simple line drawings suffice for functional MRI decoding of natural scene categories. *Proceedings of the National Academy of Sciences, U.S.A.*, *108*, 9661–9666. https://doi.org/10.1073/pnas.1015666108, PubMed: 21593417

Watson, D. M., Hartley, T., & Andrews, T. J. (2014). Patterns of response to visual scenes are linked to the low-level properties of the image. *Neuroimage*, *99*, 402–410. https://doi.org/10.1016/j.neuroimage.2014.05.045, PubMed: 24862072

Yang, Z., & Purves, D. (2003). A statistical explanation of visual space. *Nature Neuroscience*, *6*, 632–640. https://doi.org/10.1038/nn1059, PubMed: 12754512

Yokosawa, K., & Mitsumatsu, H. (2003). Does disruption of a scene impair change detection? *Journal of Vision*, *3*, 5. https://doi.org/10.1167/3.1.5, PubMed: 12678624

Yon, D., & Frith, C. D. (2021). Precision and the Bayesian brain. *Current Biology*, *31*, R1026–R1032. https://doi.org/10.1016/j.cub.2021.07.044, PubMed: 34520708

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*, 301–308. https://doi.org/10.1016/j.tics.2006.05.002, PubMed: 16784882

Zimmermann, E., Schnier, F., & Lappe, M. (2010). The contribution of scene context on change detection performance. *Vision Research*, *50*, 2062–2068. https://doi.org/10.1016/j.visres.2010.07.019, PubMed: 20682328