Published in final edited form as:
Nat Rev Psychol. 2024 January ; 3: 13–26. doi:10.1038/s44159-023-00254-0.

Predictive processing of scenes and objects

Marius V. Peelen[1], **Eva Berlot**[1], **Floris P. de Lange**[1]
[1]Donders Institute for Brain, Cognition and Behaviour, https://ror.org/016xsfp80Radboud University, Nijmegen, The Netherlands.

Abstract

Real-world visual input consists of rich scenes that are meaningfully composed of multiple objects which interact in complex, but predictable, ways. Despite this complexity, we recognize scenes, and objects within these scenes, from a brief glance at an image. In this review, we synthesize recent behavioral and neural findings that elucidate the mechanisms underlying this impressive ability. First, we review evidence that visual object and scene processing is partly implemented in parallel, allowing for a rapid initial gist of both objects and scenes concurrently. Next, we discuss recent evidence for bidirectional interactions between object and scene processing, with scene information modulating the visual processing of objects, and object information modulating the visual processing of scenes. Finally, we review evidence that objects also combine with each other to form object constellations, modulating the processing of individual objects within the object pathway. Altogether, these findings can be understood by conceptualizing object and scene perception as the outcome of a joint probabilistic inference, in which "best guesses" about objects act as priors for scene perception and vice versa, in order to concurrently optimize visual inference of objects and scenes.

The apparent ease with which we recognize visual scenes and the objects within them is one of the most remarkable feats of human cognition. This ability is supported by a wealth of low- and high-level regularities embedded in natural scenes[1–6]. Examples of high-level regularities include the natural dependencies of how objects tend to co-occur with other objects (e.g., cars with traffic signs) and scenes (e.g., cars on roads). Furthermore, objects usually appear in specific positions relative to other objects and relative to the scene (e.g., a car on the ground rather than in the sky). Finally, the retinal image of an object changes in predictable ways as a function of the object's position in a scene. Here, we review evidence showing that such object-object and scene-object regularities impact the visual processing of scenes and objects. We propose that these influences can be understood within the common framework of predictive processing.

In this review, we will adopt the frequently made distinction between objects and scenes. Although this distinction is intuitive, defining what counts as a scene or an object is not trivial. One possible definition distinguishes them based on their visual properties: scenes typically consist of larger-scale and global surfaces or environments (e.g., office, forest),

Correspondence to: Marius V. Peelen.
Correspondence: marius.peelen@donders.ru.nl (M.V. Peelen).

Europe PMC Funders Author Manuscripts

while objects are smaller-scale, local entities that are arranged in a lawful manner within scenes (e.g., computer on a desk, bird on a tree branch)[6,7]. A second distinction stems from how we interact with objects vs. scenes – we act *upon* objects, and interact *within* scenes[8]. While we will focus on work examining how scenes and objects are processed when presented as two-dimensional images, this marks an important distinction between objects and scenes in daily life. Additionally, what is considered a scene or an object partly depends on the spatial scale; for instance, an office can be considered a scene with a desk as an object within the 'office scene'. However, when zooming in, the desk can become a scene itself, with a computer as one of the objects within that 'desk scene'. Even further, the computer could be the scene within which the text on the monitor is the object of interest for a reader. While the definition of scenes and objects is thus dependent on the perspective and goal of the observer, distinguishing between the two has nevertheless proven useful for understanding visual perception.

For example, human cognitive neuroscience has revealed that objects and scenes are processed in distinct regions of the visual cortex. This distinction between object and scene processing has in fact been described as one of the main organizing principles of the human high-level visual cortex[8,9]. Within the ventral visual cortex, scenes are processed in a medial pathway, while objects are processed in a lateral pathway (Fig 1b). This follows a center-periphery organization, with the medial (scene) pathway most responsive to input from the peripheral visual field, and the lateral (object) pathway most responsive to input from the central visual field[9,10]. This organization reflects the relevance of peripheral (large-scale and coarse) versus foveated (small-scale and detailed) visual information for scene and object recognition, respectively[11,12]. Furthermore, neuroimaging studies have discovered several focal regions within the visual cortex that respond selectively to either scenes or objects (Fig 1a)[13–15]. Scene-selective regions include the parahippocampal place area (PPA[14]), the medial place area (MPA; also labeled the retrosplenial complex[16]), and the occipital place area (OPA; near the transverse occipital sulcus[17]). These regions respond more strongly to pictures of scenes than to a wide range of control pictures, including objects, faces, and scrambled scenes[14,15]. Scene-selective regions also respond strongly to "empty" scenes, such as an empty room or an open field, showing that object content is not required to activate these regions. Object-selective regions are found in the lateral occipital (LO) cortex and the posterior fusiform gyrus (pFs)[13]. These regions respond more strongly to intact objects than scrambled objects or textures, and are thought to encode object-specific properties, such as shape and category[13]. Functional magnetic resonance imaging (fMRI) studies using multivariate pattern analysis have shown that object-and scene-selective regions represent distinct aspects of visual scenes, with object-selective cortex representing the identity or category of objects in a scene (e.g., whether the objects are natural or man-made), and scene-selective cortex representing scene layout (e.g., open or closed scene boundaries)[18–20] and scene category[21]. Finally, the distinction between object- and scene-selective regions has been confirmed with transcranial magnetic stimulation (TMS): TMS over scene-selective OPA impairs the recognition of scenes (e.g., recognizing whether a scene is a beach or a forest), but not the recognition of objects (e.g., recognizing that an object is a shoe or a car)[22–24]. Conversely, TMS over object-selective LO selectively impairs the recognition of objects but not of scenes [22,23,25]. Taken together, there is convincing

evidence that objects and scenes are processed separately in the visual cortex. However, as we will discuss in this review, there are strong interactions between object and scene processing, such that responses in scene-selective cortex are modulated by the objects present in a scene[26,27] and responses in object-selective cortex are modulated by the scene surrounding an object[28].

Object and scene processing pathways are hierarchically organized, combining low-level visual features into mid-level and high-level representations. At a global level, the large-scale organization of the visual cortex can thus be described by two main orthogonal axes: a posterior-to-anterior hierarchy and a center-periphery organization[13], even though the fine-grained connectivity of the visual cortex is more complex[31]. Object- and scene-selective responses emerge at higher stages of the center- and periphery-biased pathways, respectively (Fig 1b). Note, however, that while representations in object- and scene-selective regions are tuned to relatively high-level categorical information, responses in these regions still also reflect low- or mid-level visual features[32–34].

Both object and scene processing pathways also contain extensive feedback connections, allowing higher-level representations to modulate lower-level representations. Importantly, however, while scenes and objects are processed in parallel in the visual cortex, behavioral studies have shown that scenes can influence the recognition of objects, and objects can influence the recognition of scenes. The mechanisms of such object-scene interactions have long been debated[35]. This debate is centered around the question of whether scene context influences the visual perception of objects[36,37] or whether it only influences post-perceptual processes at the level of decision making and responding[38].

In this Review, we will revisit object-scene interactions through the lens of predictive processing, arguing that perceptual modulations follow naturally from this framework. Predictive processing is a neurocomputational framework of information processing, which proposes that the brain contains internal, generative models, which send “downward” signals embodying predictions about sensory input. These signals are compared with incoming signals. The mismatch between the two, i.e. the prediction errors, are communicated “upward” to update predictions. This leads to an efficient encoding and transmission strategy, and allows the organism to infer the distal causes behind the proximal sensory input in a probabilistically optimal manner. Predictive processing thus casts perception as a process of probabilistic, knowledge-driven inference[39]. It is a process theory of information processing that can be realized at the algorithmic level in several different ways[40] and for which specific neural implementations have been proposed (Box 1). Predictions can be formed based on regularities in the input that exist both in time and space. While it may be more intuitive to think of predictions in time (e.g., predicting future locations of a moving object), predictions in space are particularly relevant for object and scene perception. Indeed, object-scene and object-object interactions can all be thought of as special cases of the general principle that computations at all levels of the processing hierarchy may generate predictions that inform and constrain computations at other levels of the hierarchy.

We first review evidence for the influence that scene processing exerts on object perception (Fig 2a), emphasizing the perceptual consequences of such influences. Thereafter, we

examine the reverse influence, that is, the influence that objects exert on the perception of scenes (Fig 2b). Comparing these two influences points to several commonalities and shared principles. Next, we review research on co-occurrences between objects in scenes, and how orderly groups of objects can be integrated to form scene-like object constellations. We discuss these findings in the context of the predictive processing framework, arguing that perception of both scenes and objects can mutually influence each other, depending on the observer's goals and the reliability of the visual signals. Last, we outline challenges and opportunities for future research, including how research on human scene perception can steer the growing line of computer vision research (Box 2) to develop computational models that could robustly recognize scenes and objects in the way humans do.

Scenes influence object perception

In one of the first studies to demonstrate the influence of scene context on object recognition, participants were asked to identify objects in briefly presented photographs, where the background was either intact or jumbled[41]. Observers were strongly impaired in selecting the shown object from a set of alternatives when the background was jumbled. These results were interpreted as evidence that the presence of scene background influenced object recognition. Since this pioneering work, many studies have replicated and extended these findings, conclusively showing that meaningful scene context facilitates the recognition of objects that are frequently found within those scenes[2,3]. These findings raise the question of *how* scene context interacts with object processing. Specifically, does scene context influence the perceptual processing of objects[3,35]?

Recent studies have provided evidence that scene context can influence the *perceptual experience* of within-scene objects (Fig 2a). One study examined the influence of scene context on the perceived sharpness of objects in a perceptual matching task[42] (Fig 3ab). Participants were presented with two blurred images of an object, side-by-side, and had to adjust one until it perceptually matched the blurriness of the other image. The authors manipulated contextual expectations afforded by the background; one of the images contained predictable information (upright intact scenes) whereas the other did not (phase-scrambled or inverted scenes). Interestingly, participants added more blur to objects in predictable scenes to match them to objects in unpredictable backgrounds, indicating that they perceived those objects as subjectively sharper. This serves as an elegant demonstration of how expectations derived from scenes influence not only semantic judgements, but also how sharply we perceive objects.

Further evidence for perceptual effects of scene context comes from neuroimaging studies showing that scene context modulates the representation of objects in the visual cortex. In an fMRI study[28], degraded objects were either presented alone or within congruent scenes (e.g., a helicopter presented in the sky). Activity patterns in the object-selective visual cortex (LO and pFs; Fig 1a) in response to degraded objects became more similar to activity patterns evoked by intact (i.e., non-degraded) objects when objects were presented within scenes (Fig 4a). These results are in line with the perceptual sharpening observed in the behavioral study reviewed above[42]. Furthermore, this effect was correlated with the amount of activity concurrently observed in scene-selective areas, suggesting that these areas could be the

origin of prediction signals facilitating object processing. This was further corroborated by a TMS study that disrupted processing in object- and scene-selective brain regions at specific time points[43]. The recognition of degraded objects in scenes was impaired when TMS was applied over scene-selective OPA 160-200 ms after scene onset. Object recognition was also impaired when applying TMS over object-selective LO at 260-300 ms after scene onset, suggesting that feedback to LO causally supports context-based object recognition. Together, these results show that scene-based expectations modulate visual object processing.

Another influence that scenes exert on object perception is by guiding our attention. This has facilitatory consequences for visual search - observers find specific objects more quickly if they are embedded in scenes that are either semantically or spatially congruent[2,44–47]. As an example, a toaster is found more easily if placed in a kitchen than in a bathroom (semantic congruence), and if it is positioned on a kitchen counter rather than the floor (spatial congruence). In the absence of a specific search task, however, attention and eye movements have been shown to be directed earlier to (and dwell longer on) semantically incongruent objects in a scene[48–54]. One reason for this increase in attention may be the increased difficulty of recognizing incongruent objects[51]. Together, these findings lead to the somewhat counterintuitive prediction that objects congruent with the scene may be processed *less well* than objects incongruent with the scene. Indeed, it has been shown that participants detect visual changes slower in scene-congruent than in scene-incongruent objects[52,53,55], unless the scene-congruent objects are central to understanding the scene[56].

Building on this notion, a recent study tested whether scene context can influence even more basic aspects of object perception – discriminating between visually similar exemplars[57] (Fig 3cd). Participants were shown photographs containing objects that were either congruent or incongruent with the scene context (e.g., a cup/toilet roll in a dishwasher/toilet-paper holder[58]). Afterwards, they were presented with two images of an exemplar from the same category (e.g., two images of a toilet roll) and asked to determine which of the two items had appeared in the scene. Participants' judgements were *less* accurate for objects congruent with the scene (e.g., toilet roll in a toilet-roll holder) than for incongruent objects (e.g., toilet roll in a dishwasher), indicating that scene-derived expectations impaired the report of expected items.

Altogether, this body of work demonstrates that scene context not only primes semantic representations of objects, but also influences how objects are perceptually experienced and represented in the visual cortex.

Objects influence scene perception

While many studies have focused on the effects of scene context on object recognition, there is also evidence for the reverse influence, with objects affecting scene recognition[59]. In one study[60], participants viewed scenes with objects that were semantically congruent or incongruent. Objects were named more accurately when they were semantically congruent with the background scene, providing another demonstration of scene context influencing object recognition. Interestingly, an equally strong congruency effect was found when participants had to name the background scene, such that scenes (e.g., a church) were recognized better when shown together with a semantically congruent object (priest) than

a semantically incongruent object (baseball player). Subsequent studies replicated[61] and extended these findings, showing that object-to-scene congruency effects persist when the object and the scene are presented simultaneously but in different images[62]. Furthermore, the semantic congruency of objects also influenced rapid categorization judgements of very briefly presented and backward masked images of scenes[63,64], suggesting interactions at relatively early stages of processing. These results indicate bidirectional interactions between object and scene processing, with semantic congruency facilitating the recognition of both objects and scenes.

In addition to influencing scene recognition, objects have also been shown to affect the representation of *visuo-spatial* scene properties. In one study[65], participants judged the spaciousness of indoor scenes after adapting participants to other, either more or less spacious, scenes from the same category (e.g., bathrooms). Results revealed a negative aftereffect, with test scenes judged as more spacious after viewing of less spacious scenes. Importantly, this effect was modulated by the visibility of scene-informative objects in the adapting scenes: when informative objects (e.g., a bathtub in a bathroom) were visible in the adapting scenes, the aftereffect was reduced. This suggests that the objects in the scene biased the perceived spaciousness of the scene towards the average spaciousness of the object-associated scene category. This interpretation was further supported by an fMRI study investigating the representation of scenes in the scene-selective parahippocampal place area (PPA). Activity patterns in the PPA differentiated between low- and high-spaciousness scenes, replicating previous findings showing PPA's sensitivity to spatial layout[18–20]. In line with the behavioral findings, the response patterns to these low- and high-spaciousness scenes became more similar to the category average when scene-associated objects were visible. Together, these results show that scene-informative objects trigger expectations about the scene that affect not only the encoding of scene identity, but also of spatial scene properties.

Similar to the influence of scenes on object recognition, these expectation effects are particularly powerful when information is ambiguous. Indeed, object information allows to disambiguate poorly visible scenes (Fig 2b), just like scene information allows to disambiguate poorly visible objects (Fig 2a). A recent fMRI study[66] (Fig 4b) showed that this object-to-scene disambiguation affects the visual representation of scenes in the PPA. When ambiguous scenes (e.g., a foggy road) contained a scene-congruent object (a car), activity patterns in the left PPA more closely resembled activity patterns evoked by clearly visible scenes of the same category. These results suggest that objects perceptually disambiguate scenes, leading to a sharper representation of these scenes in the scene-selective visual cortex. Interestingly, the object presented alone also evoked a weak scene representation (Fig 4b), suggesting that even a grey background can be interpreted as a scene[66]. Such a top-down explanation may also account for other findings of isolated object responses in scene-selective PPA. For example, PPA has been shown to respond to objects that prime a sense of space[67] and to objects that are strongly associated with a specific context[16]. Moreover, it was recently shown that voxelwise PPA responses reflect the statistical associations between objects and their visual contexts[68]. Altogether, these findings provide further evidence for interactions between object and scene processing, showing that expectations derived from objects can modulate activity in the scene-selective cortex.

Europe PMC Funders Author Manuscripts

Object-scene interactions

The previous sections indicate that scenes can influence object perception and objects can influence scene perception. This bidirectional exchange of information prompts the question of how exactly these two types of information processing interact and whether there is an asymmetrical relationship between these levels of processing.

A prominent hypothesis put forward in the past two decades has been the “scene first” hypothesis. Under this view, perception operates in a global-to-local, or coarse-to-fine[69–71], manner. Scenes typically encompass the whole visual field, and contain information at low spatial frequencies[72]. This low spatial frequency information may be rapidly processed by a magnocellular pathway, extracting the “scene gist” information (the general meaning of a scene), which can subsequently bias object computations in the slower parvocellular ventral visual pathway[73,74]. This view contrasts with an “object first” view, which conceptualizes scene comprehension as a rapid serial cascade from low-level sensory information combining into object features, and further into objects[75–77], which can then be combined to inform scene understanding[78].

On balance, neither of these views may be correct. For instance, when observers were asked to describe what they saw in briefly presented and masked pictures of scenes, they were equally likely to list semantic descriptions of objects and scenes at a given exposure time[79,80]. Rather than a temporal advantage for scene-level over object-level recognition, reports of sensory- and low-level information of scenes and objects (such as shading and shape) consistently preceded reports of high-level semantic information of either category. This finding aligns well with neuroscientific evidence that both object and scene processing comprise multiple stages, ranging from the analysis of low-level sensory to high-level semantic attributes[13]. Therefore, processing of low-level aspects of both objects and scenes precede the higher-level semantic categorization of these categories. Indeed, electrophysiological work showed that high-level representations of both objects and scenes emerge approximately 200 ms after stimulus onset[81–83]. Interestingly, scene-based sharpening of object representations in visual cortex has been observed significantly later, from around 300 ms after stimulus onset[28], reflecting feedback signals after the initial perceptual analysis of scenes and objects. Importantly, the reverse influence - with objects sharpening scene representations - was found at the same latency[84], in line with a common predictive processing mechanism for bidirectional object-scene interactions.

The evidence reviewed in the previous paragraphs argues for parallel processing pathways subserving object and scene perception. Each of these pathways is hierarchically organized, with specific combinations of simple features like edges and colors forming increasingly complex structures like objects and scenes. This hierarchical organization is an important element of predictive processing theories (Box 1), in which hierarchically ‘higher’ neural regions generate hypotheses about the expected input at hierarchically ‘lower’ regions. Mismatch between those expectations and the input is then sent forward from lower to higher regions to update expectations[85,86]. These models can provide a natural explanation for the interactions between processing stages – between stages dedicated to low-level features, like edges and contours in early visual cortex, and later processing stages dedicated to visual and semantic representations of objects and scenes[9]. High-level representations can

help to disambiguate processing of low-level elements. For example, when simple Gabor patches are aligned such that they form a global shape, this not only leads to increased activity in higher-order shape-related area LO[87], but also to activity modulations in the early visual cortical areas dedicated to the processing of the local elements. This modulation of activity can be understood as resulting from feedback from LO, providing a disambiguation that helps perceptual inference at earlier levels[88]. Similar to the hierarchical structure of object processing, scene processing is also hierarchical in nature, and also here processing at hierarchically higher levels can influence processing at hierarchically lower levels[89,90].

If objects and scenes are processed in parallel in the visual cortex, how, then, do these processing streams interact to affect visual perception? We suggest that the bidirectional interaction between object and scene processing can be understood as a form of non-hierarchical Bayesian inference[91,92]. An intuitive example of such an inference process is provided by multisensory integration, where perceptual inference is achieved by combining cues from different modalities. Human observes can integrate signals from multiple modalities in a near-optimal reliability-weighted fashion, adhering to the normative principles of Bayesian inference[93]. Likewise, auditory information that carries information about visual content (e.g., barking) can facilitate the neural and perceptual representation of ambiguous visual information (e.g., a dog)[94,95]. Analogous to the example of multisensory facilitation, object and scene perception, while processed in parallel, can still engage in mutual and facilitatory interactions. These interactions may then shape the feedback signals propagated within each pathway, modulating activity in hierarchically lower levels of the visual system, thereby resulting in overall reduced uncertainty and improved visual perception.

The reliability-weighted nature of the integration of signals also provides an intuitive explanation for the direction and magnitude of object-scene interactions. As illustrated in Figure 2, scenes influence object perception particularly when scene information is clear while object information is uncertain. This is because the relatively low uncertainty information provided by the scene is able to “move” the relatively broad probability distribution elicited by the object perception. Thereby, the scene-first vs object-first conflict disappears when viewed through the lens of Bayesian inference: the influence of each source of information is weighted as a function of its reliability. Usually, scene information is more stable over time and therefore has a larger influence on object perception than vice versa. But under situations where object information is reliable and scene information is unreliable (Fig 2b, top row), this predominant scene-to-object influence can reverse. When there is no ambiguity (both object and scene cues are highly reliable), the influence of predictive processes on shaping perceptual object and scene representations will be reduced. Nevertheless, impaired or slower recognition (e.g., naming) of contextually inconsistent objects may still be observed due to post-perceptual influences[35].

Bayesian models can also account for other apparent discrepancies in the literature, with respect to how scene context influences object perception. For example, Rossel et al.[42] (Fig 3ab) observed that contextual scene-related information disambiguated object information and rendered it perceptually “sharper”. In seeming contradiction, Spaak et al.[57] (Fig 3cd) found a perceptual disadvantage for objects that were congruent with scene context.

Importantly, the stimuli in the study by Rossel et al. were marked by large uncertainty (the objects were strongly blurred), allowing for contextual information to influence the perceptual interpretation of the input. The stimuli used by Spaak et al., on the contrary, were unambiguous, and the perceptual interpretation of the individual objects may not have been altered by the scene context. Rather, objects that were unexpected in scenes may have elicited surprise, leading to prolonged attentional processing of these objects and subsequent superior perceptual performance. A similar effect was recently reported using the perceptual matching paradigm (Fig 3a), with a perceptual benefit for expected objects when they were ambiguous but a benefit for unexpected objects when they were unambiguous[96]. These findings are in line with a two-process model of how contextual expectations can both bias towards expected, but also enhance unexpected information[97], leading us to broadly perceive what we expect, unless unexpected signals are both very reliable and surprising. In this case, perceptual inference will be dominated by the evidence (in view of its high precision), rather than the prior, and indeed may call for an update of one's prior beliefs.

Object-object interactions

The evidence discussed so far points to the interactive processing of objects and scenes, with expectations derived from scenes modulating object processing and vice versa. On the definition of objects and scenes we have adopted so far, representations of objects and scenes are not hierarchically related: object and scene processing rely on different visual cues (e.g., foveal vs peripheral input) and are processed in distinct (parallel) pathways.

However, objects do not only systematically occur relative to the global scene, but also relative to other objects. As a result, objects in scenes are not processed independently, but mutually inform and constrain each other. Unlike the parallel scene-object processing reviewed above, objects can be combined in a hierarchical manner, where multiple objects build an "object constellation" (Fig 5a). Recent studies have shown that the brain capitalizes on regularities between objects, both in terms of objects' co-occurrences as well as their relative spatial positions.

First, objects predict the presence of frequently co-occurring objects: an eraser is likely found in the vicinity of a pencil, and a marked crosswalk is likely found together with a traffic light. These co-occurrence probabilities can be used to guide visual search[98], which is particularly useful when large "anchor" objects (e.g., a sink) predict the position of smaller objects (e.g., a toothbrush)[99]. Object co-occurrences also facilitate recognition: seeing a pencil helps to disambiguate the blocky object next to it and recognize it as an eraser. Indeed, behavioral experiments have shown that objects that are typically found in the same context (e.g., sofa and lamp) were named more accurately than objects from different contexts (e.g., sofa and tractor)[61]. Similarly, objects were identified more easily when the previously fixated object was contextually related[100,101], and an object surrounded by contextually related objects was identified more accurately than the same object surrounded by unrelated objects in a forced-choice recognition task[102]. Finally, there is also neural evidence that contextual associations between objects can be learned, with objects priming representations of associated objects in the visual cortex[103,104].

Second, many co-occurring objects in everyday scenes appear in regular configurations: lamps appear above tables and mirrors above bathroom sinks (Fig 5a). Such spatial regularities between objects have been shown to facilitate perception[6]. For example, recognition is faster and more accurate for two objects shown in a regular (e.g., chair facing table) than an irregular configuration[105,106]. Furthermore, studies in patients with parietal lesions have demonstrated that extinction of an object (e.g., a bottle) presented on the contralesional side is reduced when shown together with a correctly positioned partner object (e.g., a bottle opener) on the ipsilesional side[107].

Other studies have provided evidence that regularly positioned objects are perceptually grouped, potentially reflecting an integrated representation. For example, objects in highly familiar configurations break into awareness more quickly than the same objects in unfamiliar configurations[108] (Fig 5b). similar to effects of Gestalt grouping[109]. However, unlike Gestalt grouping, these effects did not reflect low-or mid-level visual cues, as they were specific to upright (vs inverted) displays. Regularly positioned objects are also easier to reject as distractors in visual search tasks[110,111], suggesting a grouping-based reduction in distractor numerosity. Finally, neuroimaging studies have provided evidence for conjoint representations of identity-based and positional associations between objects[112–117]. For example, activity patterns in the high-level visual cortex elicited by a visual display containing multiple *unrelated* objects can be accurately modeled as a linear combination of the individual object response patterns[118], but this approximation is much less accurate when the objects are positioned to form meaningful constellations[115,116]. These results suggest that the whole is different from the sum of its parts: object constellations may activate neural populations that are sensitive to the object group, shifting the response pattern away from the combination of individual object response patterns.

These results raise the intriguing possibility that object constellations – objects that are frequently seen together in specific spatial configurations – could be a relevant representational stage in visual perception[115,116]. This representational stage would differ from the representation of whole scenes, in that it would constitute a combination of individual objects in a specific spatial arrangement (e.g., monitor and keyboard) rather than a more global “ensemble” representation of a whole scene. As such, it may be similar to how object parts are combined to form whole-object representations (cf. a laptop, for which monitor and keyboard are parts). Similar to the effects of scene-object associations, object constellations allow for perceptual facilitation: recognizing a constellation (e.g., monitor + keyboard) facilitates the processing of objects within the constellation (e.g., keyboard). However, unlike scene-object associations, object constellations are hierarchically related to individual objects, with constellations (but not scenes) built from objects. The facilitatory effects of object constellations can thus be understood within a hierarchical predictive processing framework, in which hierarchically higher levels provide a “prior” that aids the processing of the evidence at hierarchically lower levels. In general, Bayesian inference models formalize how an agent can make optimal perceptual inferences by combining different sources of information, weighted by their uncertainty. Given that low-level perception is marked by more ambiguity than high-level perception, expectations derived from hierarchically higher levels (e.g., “this input is consistent with a living room set”) can

act as a prior for disambiguating and facilitating the perceptual analysis at lower levels (e.g., "the object in the periphery may correspond to a chair").

Interestingly, similar grouping mechanisms have been described for written language, where letter stimuli are combined to form *n*-grams and words. Similar to objects, letter stimuli are perceived more accurately when they are embedded in words than when viewed in isolation[119], and word context can lead to a perceptually sharper experience of individual letters within words[120]. Concomitant with this behavioral improvement, a recent fMRI study observed a sharpened neural representation of letters in early visual areas when embedded within real words, compared to non-words[121]. We propose that such hierarchical interactions similarly take place between objects and object constellations. This provides a natural explanation for why the neural response pattern in high-level visual cortex evoked by meaningful constellations of objects are no longer accurately approximated as a linear combination of the individual object response patterns[115,116].

Summary and future directions

The ability to rapidly recognize scenes and objects is essential for navigation and decision-making. When we synthesize the findings reviewed in the previous sections, a scheme with both integrated parallel and hierarchical relations emerges (Fig 6). Initially, there is rapid parallel processing of visual input in the object processing (operating on foveal input) and scene processing (operating on peripheral input) pathways. A key advantage of these parallel and modular pathways is that the visual recognition process is compartmentalized into modules, allowing for quick and efficient inference[122]. Within each processing pathway, there is a hierarchy of processing stages, from low-level simple sensory features to high-level complex semantic features. This initial rapid feedforward process sets up a hypothesis landscape of objects, constellations of objects, and scenes. Following this first pass of processing, interactions can emerge between different levels of processing. First, interactions arise within a pathway, either within a specific level of processing (e.g., one object may activate other congruent objects), or between levels of processing (e.g., the high-level representation of a chair may sharpen the perceived sensory attributes of the chair). These interactions can be implemented by horizontal connections between neurons within one level of processing and feedback modulations between hierarchically related processing regions within one pathway. Second, direct interactions may emerge between pathways (e.g., a kitchen scene may enhance the representations of a dinner table and chairs). Finally, both object and scene processing can interact by activating a conceptual schema (e.g., a kitchen scene may activate the "kitchen schema", which in turn enhances activation of schema-congruent objects). In terms of neural implementation, several studies have pointed at hippocampus, the medial prefrontal cortex, and the orbitofrontal cortex as having an important role in integrating new information within a pre-existing schema[3,74,123]. Predictions derived from this activated schema could then be sent back to the separate pathways[74], to facilitate processing in both pathways, in line with predictive processing.

Our review raises several questions that can be addressed in future work. First, it is unknown which scene or object cues are most important for the perceptual effects reviewed here. For example, when considering the disambiguation example of Fig 2a, the object may be

perceived as a car because of the scene's overall meaning (a road scene), because of the position of the object within the scene (on the road), because of the size implied by the distance within the scene, or because of all these factors combined. Future studies could systematically manipulate these factors to test which are needed for the perceptual effects described here. Another question concerns the automaticity of these effects. For example, studies could test whether attention[124] or conscious recognition[125,126] is required for the perceptual disambiguation of objects and scenes (Fig 2a). Relatedly, by comparing responses across tasks, future studies could test whether feedback signaling within each pathway is automatic or whether it depends on task demands. Finally, studies could train participants on new scene-object associations and test how much exposure is required for scene and object priors to become effective.

At the neural level, it is still largely unknown how object and scene pathways are connected. For example, while both PPA and pFs/LO are anatomically connected to earlier retinotopic areas, there is currently no evidence for direct connections between PPA and pFs/LO[20,31,127]. One possibility is that the interactive effects we have reviewed here are the result of interactions between pathways within the visual cortex at multiple levels. Alternatively, the main connection could travel through non-visual areas such as the hippocampus, orbitofrontal or medial prefrontal cortex[3,123]. Much is also still unknown about the representation of object constellations. For example, are there holistic representations of familiar constellations, or does grouping merely link individual object representations (similar questions have been addressed for the representation of letters and words[128])? And what is the role of recently described regions that are tuned to reachable-scale environments[129], and that are sensitive to the presence of multiple objects, in the inter-object effects described here? Another open question is whether the role of different layers in conveying prediction and error signals (Box 1) can be extended to explain the interactions between object- and scene-selective cortical regions. Finally, an important avenue for future research is to improve modeling of contextual effects in neural networks, which will help to increase our understanding of these effects at the computational level, improve performance of these networks, and make them behave more similar to humans (Box 2).

Moving forward, we will need to consider how we process scenes and objects in daily life, where the environment is typically stable and where we actively explore scenes in the context of goal-directed behavior (e.g., visual search). Recent advances in virtual and augmented reality, in combination with mobile imaging[130], may serve to approach more naturalistic conditions in experimental research[131]. In daily life, the spatial and temporal context within which we perceive objects is much richer than when we are viewing briefly flashed images[132–134]. As a consequence, the contribution of predictive processes in perception is likely even stronger than revealed in the studies reviewed here.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreements No. 725970 to M.V.P. and 101000942 to F.P.d.L.). E.B. is supported by The EMBO Postdoctoral Fellowship (award no. ALTF 579-2021).

References

1. Barlow HB. The knowledge used in vision and where it comes from. Phil Trans R Soc Lond B. 1997; 352: 1141–1147. [PubMed: 9304681]
2. Oliva A, Torralba A. The role of context in object recognition. Trends in Cognitive Sciences. 2007; 11: 520–527. [PubMed: 18024143]
3. Bar M. Visual objects in context. Nat Rev Neurosci. 2004; 5: 617–629. [PubMed: 15263892]
4. Purves D, Wojtach WT, Lotto RB. Understanding vision in wholly empirical terms. Proc Natl Acad Sci USA. 2011; 108: 15588–15595. [PubMed: 21383192]
5. Simoncelli EP, Olshausen BA. Natural Image Statistics and Neural Representation. Annu Rev Neurosci. 2001; 24: 1193–1216. [PubMed: 11520932]
6. Kaiser D, Quek GL, Cichy RM, Peelen MV. Object Vision in a Structured World. Trends in Cognitive Sciences. 2019; 23: 672–685. [PubMed: 31147151]
7. Malcolm GL, Groen IIA, Baker CI. Making Sense of Real-World Scenes. Trends in Cognitive Sciences. 2016; 20: 843–856. [PubMed: 27769727]
8. Epstein, RA. Neural Systems for Visual Scene Recognition. In: Kveraga, K, Bar, M, editors. Scene Vision. The MIT Press; 2014. 105–134.
9. Grill-Spector K, Weiner KS. The functional architecture of the ventral temporal cortex and its role in categorization. Nat Rev Neurosci. 2014; 15: 536–548. [PubMed: 24962370]
10. Levy I, Hasson U, Avidan G, Hendler T, Malach R. Center–periphery organization of human object areas. Nat Neurosci. 2001; 4: 533–539. [PubMed: 11319563]
11. Larson AM, Loschky LC. The contributions of central versus peripheral vision to scene gist recognition. Journal of Vision. 2009; 9: 6. [PubMed: 19761321]
12. Trouilloud A, et al. Rapid scene categorization: From coarse peripheral vision to fine central vision. Vision Research. 2020; 170: 60–72. [PubMed: 32259648]
13. Grill-Spector K, Malach R. The human visual cortex. Annu Rev Neurosci. 2004; 27: 649–677. [PubMed: 15217346]
14. Epstein R, Kanwisher N. A cortical representation of the local visual environment. Nature. 1998; 392: 598–601. [PubMed: 9560155]
15. Epstein RA, Baker CI. Scene Perception in the Human Brain. Annu Rev Vis Sci. 2019; 5: 373–397. [PubMed: 31226012]
16. Bar M, Aminoff E. Cortical Analysis of Visual Context. Neuron. 2003; 38: 347–358. [PubMed: 12718867]
17. Hasson U, Levy I, Behrmann M, Hendler T, Malach R. Eccentricity Bias as an Organizing Principle for Human High-Order Object Areas. Neuron. 2002; 34: 479–490. [PubMed: 11988177]
18. Kravitz DJ, Peng CS, Baker CI. Real-World Scene Representations in High-Level Visual Cortex: It's the Spaces More Than the Places. Journal of Neuroscience. 2011; 31: 7322–7333. [PubMed: 21593316]
19. Park S, Brady TF, Greene MR, Oliva A. 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. 2011; 31: 1333–1340. [PubMed: 21273418]
20. Harel A, Kravitz DJ, Baker CI. Deconstructing Visual Scenes in Cortex: Gradients of Object and Spatial Layout Information. Cerebral Cortex. 2013; 23: 947–957. [PubMed: 22473894]
21. Walther DB, Caddigan E, Fei-Fei L, Beck DM. Natural Scene Categories Revealed in Distributed Patterns of Activity in the Human Brain. J Neurosci. 2009; 29: 10573–10581. [PubMed: 19710310]
22. Dilks DD, Julian JB, Paunov AM, Kanwisher N. The Occipital Place Area Is Causally and Selectively Involved in Scene Perception. Journal of Neuroscience. 2013; 33: 1331–1336. [PubMed: 23345209]
23. Wischnewski M, Peelen MV. Causal Evidence for a Double Dissociation between Object- and Scene-Selective Regions of Visual Cortex: A Preregistered TMS Replication Study. J Neurosci. 2021; 41: 751–756. [PubMed: 33262244]

24. Ganaden RE, Mullin CR, Steeves JKE. Transcranial Magnetic Stimulation to the Transverse Occipital Sulcus Affects Scene but Not Object Processing. Journal of Cognitive Neuroscience. 2013; 25: 961–968. [PubMed: 23410031]
25. Mullin CR, Steeves JKE. TMS to the Lateral Occipital Cortex Disrupts Object Processing but Facilitates Scene Processing. Journal of Cognitive Neuroscience. 2011; 23: 4174–4184. [PubMed: 21812554]
26. Troiani V, Stigliani A, Smith ME, Epstein RA. Multiple Object Properties Drive Scene-Selective Regions. Cerebral Cortex. 2014; 24: 883–897. [PubMed: 23211209]
27. Aminoff EM, Durham T. Scene-selective brain regions respond to embedded objects of a scene. Cerebral Cortex. 2022; bhac399 doi: 10.1093/cercor/bhac399
28. Brandman T, Peelen MV. Interaction between Scene and Object Processing Revealed by Human fMRI and MEG Decoding. J Neurosci. 2017; 37: 7700–7710. [PubMed: 28687603]
29. Bainbridge WA, Hall EH, Baker CI. Distinct Representational Structure and Localization for Visual Encoding and Recall during Visual Imagery. Cerebral Cortex. 2021; 31: 1898–1913. [PubMed: 33285563]
30. Malach R, Levy I, Hasson U. The topography of high-order human object areas. Trends in Cognitive Sciences. 2002; 6: 176–184. [PubMed: 11912041]
31. Kravitz DJ, Saleem KS, Baker CI, Ungerleider LG, Mishkin M. The ventral visual pathway: an expanded neural framework for the processing of object quality. Trends in Cognitive Sciences. 2013; 17: 26–49. [PubMed: 23265839]
32. Andrews TJ, Clarke A, Pell P, Hartley T. Selectivity for low-level features of objects in the human ventral stream. NeuroImage. 2010; 49: 703–711. [PubMed: 19716424]
33. Groen IIA, Silson EH, Baker CI. Contributions of low- and high-level properties to neural processing of visual scenes in the human brain. Phil Trans R Soc B. 2017; 372 20160102 [PubMed: 28044013]
34. Op de Beeck HP, Haushofer J, Kanwisher NG. Interpreting fMRI data: maps, modules and dimensions. Nat Rev Neurosci. 2008; 9: 123–135. [PubMed: 18200027]
35. Henderson JM, Hollingworth A. High-level scene perception. Annu Rev Psychol. 1999; 50: 243–271. [PubMed: 10074679]
36. Biederman I, Mezzanotte RJ, Rabinowitz JC. Scene perception: Detecting and judging objects undergoing relational violations. Cognitive Psychology. 1982; 14: 143–177. [PubMed: 7083801]
37. Palmer, SE. Explorations in cognition. Freeman; 1975. 279–307.
38. Hollingworth A, Henderson JM. Does Consistent Scene Context Facilitate Object Perception? Journal of Experimental Psychology: General. 1998; 127: 398–415. [PubMed: 9857494]
39. Clark A. Whatever next? Predictive brains, situated agents, and the future of cognitive science. Behav Brain Sci. 2013; 36: 181–204. [PubMed: 23663408]
40. Spratling MW. A review of predictive coding algorithms. Brain and Cognition. 2017; 112: 92–97. [PubMed: 26809759]
41. Biederman I. Perceiving real-world scenes. Science. 1972; 177: 77–80. [PubMed: 5041781]
42. Rossel P, Peyrin C, Roux-Sibilon A, Kauffmann L. It makes sense, so I see it better! Contextual information about the visual environment increases its perceived sharpness. Journal of Experimental Psychology: Human Perception and Performance. 2022; 48: 331–350. [PubMed: 35130017]
43. Wischnewski M, Peelen MV. Causal neural mechanisms of context-based object recognition. eLife. 2021; 10 e69736 [PubMed: 34374647]
44. Wolfe JM, Võ ML-H, Evans KK, Greene MR. Visual search in scenes involves selective and nonselective pathways. Trends in Cognitive Sciences. 2011; 15: 77–84. [PubMed: 21227734]
45. Peelen MV, Kastner S. Attention in the real world: toward understanding its neural basis. Trends in Cognitive Sciences. 2014; 18: 242–250. [PubMed: 24630872]
46. Castelhano MS, Krzyś K. Rethinking Space: A Review of Perception, Attention, and Memory in Scene Processing. Annu Rev Vis Sci. 2020; 6: 563–586. [PubMed: 32491961]

47. Võ ML-H, Boettcher SE, Draschkow D. Reading scenes: how scene grammar guides attention and aids perception in real-world environments. Current Opinion in Psychology. 2019; 29: 205–210. [PubMed: 31051430]
48. Friedman A. Framing pictures: The role of knowledge in automatized encoding and memory for gist. Journal of Experimental Psychology: General. 1979; 108: 316–355. [PubMed: 528908]
49. Henderson JM, Weeks PA Jr, Hollingworth A. The Effects of Semantic Consistency on Eye Movements During Complex Scene Viewing. Journal of Experimental Psychology: Human Perception and Performance. 1999; 25: 210–228.
50. Cornelissen THW, Võ ML-H. Stuck on semantics: Processing of irrelevant object-scene inconsistencies modulates ongoing gaze behavior. Atten Percept Psychophys. 2017; 79: 154–168. [PubMed: 27645215]
51. Underwood G, Templeman E, Lamming L, Foulsham T. Is attention necessary for object identification? Evidence from eye movements during the inspection of real-world scenes. Consciousness and Cognition. 2008; 17: 159–170. [PubMed: 17222564]
52. LaPointe MRP, Lupianez J, Milliken B. Context congruency effects in change detection: Opposing effects on detection and identification. Visual Cognition. 2013; 21: 99–122.
53. Ortiz-Tudela J, Jiménez L, Lupiáñez J. Scene-object semantic incongruity across stages of processing: From detection to identification and episodic encoding. Front Cognit. 2023; 2 1125145
54. Loftus GR, Mackworth NH. Cognitive determinants of fixation location during picture viewing. Journal of Experimental Psychology: Human Perception and Performance. 1978; 4: 565–572. [PubMed: 722248]
55. Hollingworth A, Henderson JM. Semantic Informativeness Mediates the Detection of Changes in Natural Scenes. Visual Cognition. 2000; 7: 213–235.
56. Rensink RA, O'Regan JK, Clark JJ. To See or not to See: The Need for Attention to Perceive Changes in Scenes. Psychol Sci. 1997; 8: 368–373.
57. Spaak E, Peelen MV, de Lange FP. Scene Context Impairs Perception of Semantically Congruent Objects. Psychological Science. 2022; 33: 299–313. [PubMed: 35020519]
58. Öhlschläger S, Võ ML-H. SCEGRAM: An image database for semantic and syntactic inconsistencies in scenes. Behav Res. 2017; 49: 1780–1791.
59. Wiesmann SL, Võ ML-H. What makes a scene? Fast scene categorization as a function of global scene information at different resolutions. Journal of Experimental Psychology: Human Perception and Performance. 2022; 48: 871–888. [PubMed: 35708933]
60. Davenport JL, Potter MC. Scene Consistency in Object and Background Perception. Psychol Sci. 2004; 15: 559–564. [PubMed: 15271002]
61. Davenport JL. Consistency effects between objects in scenes. Memory & Cognition. 2007; 35: 393–401. [PubMed: 17691140]
62. Leroy A, Faure S, Spotorno S. Reciprocal semantic predictions drive categorization of scene contexts and objects even when they are separate. Scientific Reports. 2020; 10 8447 [PubMed: 32439874]
63. Joubert OR, Rousselet GA, Fize D, Fabre-Thorpe M. Processing scene context: Fast categorization and object interference. Vision Research. 2007; 47: 3286–3297. [PubMed: 17967472]
64. Furtak M, Mudrik L, Bola M. The forest, the trees, or both? Hierarchy and interactions between gist and object processing during perception of real-world scenes. Cognition. 2022; 221 104983 [PubMed: 34968994]
65. Linsley D, MacEvoy SP. Encoding-Stage Crosstalk Between Object- and Spatial Property-Based Scene Processing Pathways. Cereb Cortex. 2015; 25: 2267–2281. [PubMed: 24610116]
66. Brandman T, Peelen MV. Signposts in the Fog: Objects Facilitate Scene Representations in Left Scene-selective Cortex. Journal of Cognitive Neuroscience. 2019; 31: 390–400. [PubMed: 29561241]
67. Mullally SL, Maguire EA. A New Role for the Parahippocampal Cortex in Representing Space. J Neurosci. 2011; 31: 7441–7449. [PubMed: 21593327]
68. Bonner MF, Epstein RA. Object representations in the human brain reflect the co-occurrence statistics of vision and language. Nat Commun. 2021; 12 4081 [PubMed: 34215754]

Europe PMC Funders Author Manuscripts

69. Schyns PG, Oliva A. From Blobs to Boundary Edges: Evidence for Time- and Spatial-Scale-Dependent Scene Recognition. Psychol Sci. 1994; 5: 195–200.
70. Oliva A, Torralba A. Building the Gist of a Scene: The Role of Global Image Features in Recognition. Progress in Brain Research. 2006; 155: 23–36. [PubMed: 17027377]
71. Hochstein S, Ahissar M. View from the Top: Review Hierarchies and Reverse Hierarchies in the Visual System. Neuron. 2002; 36: 791–804. [PubMed: 12467584]
72. Kauffmann L, RamanoÃ«l S, Peyrin C. The neural bases of spatial frequency processing during scene perception. Front Integr Neurosci. 2014; 8: 1–14. [PubMed: 24474908]
73. Bullier J. Integrated model of visual processing. Brain Research Reviews. 2001; 36: 96–107. [PubMed: 11690606]
74. Bar M, et al. Top-down facilitation of visual recognition. Proc Natl Acad Sci USA. 2006; 103: 449–454. [PubMed: 16407167]
75. Liu H, Agam Y, Madsen JR, Kreiman G. Timing, Timing, Timing: Fast Decoding of Object Information from Intracranial Field Potentials in Human Visual Cortex. Neuron. 2009; 62: 281–290. [PubMed: 19409272]
76. Thorpe S, Fize D, Marlot C. Speed of processing in the human visual system. Nature. 1996; 381: 520–522. [PubMed: 8632824]
77. Crouzet SM, Joubert OR, Thorpe SJ, Fabre-Thorpe M. Animal Detection Precedes Access to Scene Category. PLoS ONE. 2012; 7 e51471 [PubMed: 23251545]
78. MacEvoy SP, Epstein RA. Constructing scenes from objects in human occipitotemporal cortex. Nat Neurosci. 2011; 14: 1323–1329. [PubMed: 21892156]
79. Fei-Fei L, Iyer A, Koch C, Perona P. What do we perceive in a glance of a real-world scene? Journal of Vision. 2007; 7: 10.
80. Chuyin Z, Koh ZH, Gallagher R, Nishimoto S, Tsuchiya N. What can we experience and report on a rapidly presented image? Intersubjective measures of specificity of freely reported contents of consciousness. F1000Res. 2022; 11: 69. [PubMed: 36176545]
81. Carlson T, Tovar DA, Alink A, Kriegeskorte N. Representational dynamics of object vision: The first 1000 ms. Journal of Vision. 2013; 13: 1.
82. Cichy RM, Pantazis D, Oliva A. Resolving human object recognition in space and time. Nat Neurosci. 2014; 17: 455–462. [PubMed: 24464044]
83. Kaiser D, Azzalini DC, Peelen MV. Shape-independent object category responses revealed by MEG and fMRI decoding. Journal of Neurophysiology. 2016; 115: 2246–2250. [PubMed: 26740535]
84. Brandman T, Peelen MV. Objects sharpen visual scene representations: evidence from MEG decoding. 2023; doi: 10.1101/2023.04.06.535903 http://biorxiv.org/lookup/doi/10.1101/2023.04.06.535903
85. Rao RPN, Ballard DH. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. Nat Neurosci. 1999; 2: 79–87. [PubMed: 10195184]
86. Friston K. A theory of cortical responses. Phil Trans R Soc B. 2005; 360: 815–836. [PubMed: 15937014]
87. Altmann CF, Bülthoff HH, Kourtzi Z. Perceptual Organization of Local Elements into Global Shapes in the Human Visual Cortex. Current Biology. 2003; 13: 342–349. [PubMed: 12593802]
88. Teufel C, Dakin SC, Fletcher PC. Prior object-knowledge sharpens properties of early visual feature-detectors. Sci Rep. 2018; 8 10853 [PubMed: 30022033]
89. Neri P. Global Properties of Natural Scenes Shape Local Properties of Human Edge Detectors. Front Psychology. 2011; 2: 1–20.
90. Smith FW, Muckli L. Nonstimulated early visual areas carry information about surrounding context. Proceedings of the National Academy of Sciences. 2010; 107: 20099–20103.
91. Doya, K, Ishii, S, Pouget, A, Rao, RPN. Bayesian Brain: Probabilistic Approaches to Neural Coding. MIT Press; 2006.
92. Ma WJ. Organizing probabilistic models of perception. Trends in Cognitive Sciences. 2012; 16: 511–518. [PubMed: 22981359]

Europe PMC Funders Author Manuscripts

93. Ernst MO, Banks MS. Humans integrate visual and haptic information in a statistically optimal fashion. Nature. 2002; 415: 429–433. [PubMed: 11807554]
94. Werner S, Noppeney U. Distinct Functional Contributions of Primary Sensory and Association Areas to Audiovisual Integration in Object Categorization. Journal of Neuroscience. 2010; 30: 2662–2675. [PubMed: 20164350]
95. Brandman T, Avancini C, Leticevscaia O, Peelen MV. Auditory and Semantic Cues Facilitate Decoding of Visual Object Category in MEG. Cerebral Cortex. 2020; 30: 597–606. [PubMed: 31216008]
96. Rossel P, Peyrin C, Kauffmann L. Subjective perception of objects depends on the interaction between the validity of context-based expectations and signal reliability. Vision Research. 2023; 206 108191 [PubMed: 36773476]
97. Press C, Kok P, Yon D. The Perceptual Prediction Paradox. Trends in Cognitive Sciences. 2020; 24: 13–24. [PubMed: 31787500]
98. Brockmole JR, Henderson JM. Short Article: Recognition and Attention Guidance during Contextual Cueing in Real-World Scenes: Evidence from Eye Movements. Quarterly Journal of Experimental Psychology. 2006; 59: 1177–1187.
99. Boettcher SEP, Draschkow D, Dienhart E, Võ ML-H. Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. Journal of Vision. 2018; 18: 11.
100. de Graef P, de Troy A, d'Ydewalle G. Local and global contextual constraints on the identification of objects in scenes. Canadian Journal of Psychology / Revue canadienne de psychologie. 1992; 46: 489–508.
101. Henderson JM, Pollatsek A, Rayner K. Effects of Foveal Priming and Extrafoveal Preview on Object Identification. Journal of Experimental Psychology: Human Perception and Performance. 1987; 13: 449–463. [PubMed: 2958593]
102. Auckland ME, Cave KR, Donnelly N. Nontarget objects can influence perceptual processes during object recognition. Psychonomic Bulletin & Review. 2007; 14: 332–337. [PubMed: 17694922]
103. Meyer T, Olson CR. Statistical learning of visual transitions in monkey inferotemporal cortex. Proc Natl Acad Sci USA. 2011; 108: 19401–19406. [PubMed: 22084090]
104. He T, Richter D, Wang Z, de Lange FP. Spatial and Temporal Context Jointly Modulate the Sensory Response within the Ventral Visual Stream. Journal of Cognitive Neuroscience. 2022; 34: 332–347. [PubMed: 34964889]
105. Bar M, Ullman S. Spatial Context in Recognition. Perception. 1996; 25: 343–352. [PubMed: 8804097]
106. Green C, Hummel JE. Familiar interacting object pairs are perceptually grouped. Journal of Experimental Psychology: Human Perception and Performance. 2006; 32: 1107–1119. [PubMed: 17002525]
107. Riddoch MJ, Humphreys GW, Edwards S, Baker T, Willson K. Seeing the action: neuropsychological evidence for action-based effects on object selection. Nat Neurosci. 2003; 6: 82–89. [PubMed: 12469129]
108. Stein T, Kaiser D, Peelen MV. Interobject grouping facilitates visual awareness. Journal of Vision. 2015; 15: 10.
109. Wang L, Weng X, He S. Perceptual Grouping without Awareness: Superiority of Kanizsa Triangle in Breaking Interocular Suppression. PLoS ONE. 2012; 7 e40106 [PubMed: 22768232]
110. Kaiser D, Stein T, Peelen MV. Object grouping based on real-world regularities facilitates perception by reducing competitive interactions in visual cortex. Proc Natl Acad Sci USA. 2014; 111: 11217–11222. [PubMed: 25024190]
111. Thorat S, Quek GL, Peelen MV. Statistical learning of distractor co-occurrences facilitates visual search. Journal of Vision. 2022; 22: 2.
112. Gronau N, Neta M, Bar M. Integrated Contextual Representation for Objects' Identities and Their Locations. Journal of Cognitive Neuroscience. 2008; 20: 371–388. [PubMed: 18004950]
113. Kim JG, Biederman I. Where Do Objects Become Scenes? Cerebral Cortex. 2011; 21: 1738–1746. [PubMed: 21148087]

114. Roberts KL, Humphreys GW. Action relationships concatenate representations of separate objects in the ventral visual system. NeuroImage. 2010; 52: 1541–1548. [PubMed: 20580845]
115. Baldassano C, Beck DM, Fei-Fei L. Human–Object Interactions Are More than the Sum of Their Parts. Cereb Cortex. 2016; bhw077 doi: 10.1093/cercor/bhw077
116. Kaiser D, Peelen MV. Transformation from independent to integrative coding of multi-object arrangements in human visual cortex. NeuroImage. 2018; 169: 334–341. [PubMed: 29277645]
117. Quek GL, Peelen MV. Contextual and Spatial Associations Between Objects Interactively Modulate Visual Processing. Cerebral Cortex. 2020; 30: 6391–6404. [PubMed: 32754744]
118. MacEvoy SP, Epstein RA. Decoding the Representation of Multiple Simultaneous Objects in Human Occipitotemporal Cortex. Current Biology. 2009; 19: 943–947. [PubMed: 19446454]
119. Reicher GM. Perceptual recognition as a function of meaningfulness of stimulus material. Journal of Experimental Psychology. 1969; 81: 275–280. [PubMed: 5811803]
120. Lupyan G. Objective Effects of Knowledge on Visual Perception. Journal of Experimental Psychology: Human Perception and Performance. 2017; 43: 794–806. [PubMed: 28345946]
121. Heilbron M, Richter D, Ekman M, Hagoort P, de Lange FP. Word contexts enhance the neural representation of individual letters in early visual cortex. Nat Commun. 2020; 11: 321. [PubMed: 31949153]
122. Lee TS. The Visual System's Internal Model of the World. Proc IEEE. 2015; 103: 1359–1378.
123. van Kesteren MTR, Ruiter DJ, Fernández G, Henson RN. How schema and novelty augment memory formation. Trends in Neurosciences. 2012; 35: 211–219. [PubMed: 22398180]
124. Munneke J, Brentari V, Peelen MV. The influence of scene context on object recognition is independent of attentional focus. Front Psychol. 2013; 4
125. Mudrik L, Breska A, Lamy D, Deouell LY. Integration Without Awareness: Expanding the Limits of Unconscious Processing. Psychol Sci. 2011; 22: 764–770. [PubMed: 21555524]
126. Faivre N, Dubois J, Schwartz N, Mudrik L. Imaging object-scene relations processing in visible and invisible natural scenes. Sci Rep. 2019; 9 4567 [PubMed: 30872607]
127. Kim M, et al. Anatomical correlates of the functional organization in the human occipitotemporal cortex. Magnetic Resonance Imaging. 2006; 24: 583–590. [PubMed: 16735179]
128. Dehaene S, Cohen L, Sigman M, Vinckier F. The neural code for written words: a proposal. Trends in Cognitive Sciences. 2005; 9: 335–341. [PubMed: 15951224]
129. Josephs EL, Konkle T. Large-scale dissociations between views of objects, scenes, and reachable-scale environments in visual cortex. Proc Natl Acad Sci USA. 2020; 117: 29354–29362. [PubMed: 33229533]
130. De Vos M, Debener S. Mobile EEG: Towards brain activity monitoring during natural action and cognition. International Journal of Psychophysiology. 2014; 91: 1–2. [PubMed: 24144634]
131. Helbing J, Draschkow D, L-H, Võ M. Auxiliary Scene-Context Information Provided by Anchor Objects Guides Attention and Locomotion in Natural Search Behavior. Psychol Sci. 2022; 33: 1463–1476. [PubMed: 35942922]
132. Shamay-Tsoory SG, Mendelsohn A. Real-Life Neuroscience: An Ecological Approach to Brain and Behavior Research. Perspect Psychol Sci. 2019; 14: 841–859. [PubMed: 31408614]
133. Matusz PJ, Dikker S, Huth AG, Perrodin C. Are We Ready for Real-world Neuroscience? Journal of Cognitive Neuroscience. 2019; 31: 327–338. [PubMed: 29916793]
134. Willems RM, Peelen MV. How context changes the neural basis of perception and language. iScience. 2021; 24 102392 [PubMed: 33997677]
135. Keller GB, Mrsic-Flogel TD. Predictive Processing: A Canonical Cortical Computation. Neuron. 2018; 100: 424–435. [PubMed: 30359606]
136. Lawrence SJD, Formisano E, Muckli L, de Lange FP. Laminar fMRI: Applications for cognitive neuroscience. NeuroImage. 2019; 197: 785–791. [PubMed: 28687519]
137. Lawrence SJD, et al. Laminar Organization of Working Memory Signals in Human Visual Cortex. Current Biology. 2018; 28: 3435–3440. e4 [PubMed: 30344121]
138. Muckli L, et al. Contextual Feedback to Superficial Layers of V1. Current Biology. 2015; 25: 2690–2695. [PubMed: 26441356]

139. Aitken F, et al. Prior expectations evoke stimulus-specific activity in the deep layers of the primary visual cortex. PLoS Biol. 2020; 18 e3001023 [PubMed: 33284791]
140. Kok P, Bains LJ, van Mourik T, Norris DG, de Lange FP. Selective Activation of the Deep Layers of the Human Primary Visual Cortex by Top-Down Feedback. Current Biology. 2016; 26: 371–376. [PubMed: 26832438]
141. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. Commun ACM. 2012; 60: 84–90.
142. VanRullen R. Perception Science in the Age of Deep Neural Networks. Front Psychol. 2017; 8: 1–6. [PubMed: 28197108]
143. Khaligh-Razavi S-M, Kriegeskorte N. Deep Supervised, but Not Unsupervised, Models May Explain IT Cortical Representation. PLoS Comput Biol. 2014; 10 e1003915 [PubMed: 25375136]
144. Cichy RM, Khosla A, Pantazis D, Torralba A, Oliva A. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Sci Rep. 2016; 6 27755 [PubMed: 27282108]
145. Guclu U, van Gerven MAJ. Deep Neural Networks Reveal a Gradient in the Complexity of Neural Representations across the Ventral Stream. Journal of Neuroscience. 2015; 35: 10005–10014. [PubMed: 26157000]
146. Eckstein MP, Koehler K, Welbourne LE, Akbas E. Humans, but Not Deep Neural Networks, Often Miss Giant Targets in Scenes. Current Biology. 2017; 27: 2827–2832. e3 [PubMed: 28889976]
147. Gayet S, Peelen MV. Preparatory attention incorporates contextual expectations. Current Biology. 2022; 32: 687–692. e6 [PubMed: 34919809]
148. Katti H, Peelen MV, Arun SP. Machine vision benefits from human contextual expectations. Sci Rep. 2019; 9 2112 [PubMed: 30765753]
149. Zhu Z, Xie L, Yuille A. Object Recognition with and without Objects. IJCAI. 2017; 7
150. Daucé E, Albiges P, Perrinet LU. A dual foveal-peripheral visual processing model implements efficient saccade selection. Journal of Vision. 2020; 20: 22.
151. Akbas E, Eckstein MP. Object detection through search with a foveated visual system. PLoS Comput Biol. 2017; 13 e1005743 [PubMed: 28991906]
152. Pramod RT, Katti H, Arun SP. Human peripheral blur is optimal for object recognition. Vision Research. 2022; 200 108083 [PubMed: 35830763]
153. Chen G, et al. A Survey of the Four Pillars for Small Object Detection: Multiscale Representation, Contextual Information, Super-Resolution, and Region Proposal. IEEE Trans Syst Man Cybern, Syst. 2022; 52: 936–953.
154. Xiang, W; Zhang, D-Q; Yu, H; Athitsos, V. Context-Aware Single-Shot Detector; 2018 IEEE Winter Conference on Applications of Computer Vision (WACV); 2018. 1784–1793.
155. Wang AY, Kay K, Naselaris T, Tarr MJ, Wehbe L. Incorporating natural language into vision models improves prediction and understanding of higher visual cortex. 2022; doi: 10.1101/2022.09.27.508760 http://biorxiv.org/lookup/doi/10.1101/2022.09.27.508760

Europe PMC Funders Author Manuscripts

Box 1

Neurobiology of predictive processing

The neurobiological implementation of predictive processing is still an active topic of investigation[135], and it is plausible that a multitude of implementation schemes exist, depending on the exact properties of the inference process. Nevertheless, a general neural motif has emerged, suggesting that feedback prediction signals are encapsulated from incoming and error signals by virtue of residing in distinct layers of each cortical module. Namely, bottom-up connections originate from superficial layers of a "lower" area and terminate in the middle layer of a "higher" area, while top-down connections from a higher to lower area originate from deep layers and avoid the middle layer. This neuro-anatomical organization, together with recent advances in neuroimaging technology, now allows researchers to isolate bottom-up and top-down processing[136]. Within the context of predictive processing, top-down processing embodies the generative model, which can internally generate patterns of activity that specific external stimuli would elicit "from the bottom up". Indeed, it has been found that internally generated stimulus representations (i.e., visual imagery) leads to stimulus-specific activity in the deep and superficial, but not middle, layers of the primary visual cortex[137,138]. Interestingly, the mere anticipation of the occurrence of a specific oriented grating stimulus leads to stimulus-specific activity selectively in the deep layers of the primary visual cortex[139]. Similar stimulus-specific activity in the deep layers of the visual system is observed when an expectation of stimulus input is generated by a visual illusion[140]. It is an open question whether object expectations generated by scene information, or vice versa, results in similar layer-specific activity modulations. Therefore, layer-resolved fMRI may hold great promise for elucidating the interactions between object- and scene-selective processing.

Box 2

Incorporating context into neural network models

Deep learning has dramatically advanced the field of computer vision in the last decade[141]. This success is largely due to the new computational architectures of visual processing, specifically convolutional deep neural networks (DNNs), which are trained to perform object categorization using large-scale datasets with labeled examples. The achieved object classification of these trained networks is impressive, achieving human-level performance[142]. Interestingly, the representations that emerge in deeper network layers exhibit similarities to how objects are represented in the human ventral visual cortex[143–145]. Nonetheless, there are many outstanding differences between DNNs and human visual processing, especially in the processing of scenes.

While humans have separate pathways for processing objects and scenes, this distinction is usually not present in DNNs. As a result, the networks do not systematically exploit scene information for object processing the way humans do. For instance, humans use scene information to form expectations about the size of objects[146,147], facilitating the detection of congruently sized objects. In contrast, DNNs are typically insensitive to such information: they are as likely to identify objects correctly when they are of an expected, i.e. real-world, size, as when they are sized abnormally[146].

Recent work has explored various approaches to incorporate context into DNNs and has shown that this can benefit the performance of DNNs. One approach is to augment DNNs with human-derived expectations about the position, scale and likelihood of specific objects in scenes. Such expectations are image-computable and can facilitate object classification[148]. It has also been shown that training DNNs separately on the foreground vs. background of image contents (i.e., scenes with and without objects in place) improves object recognition[149]. Yet another approach is by adopting the local vs. global processing of object- vs. scene-processing: we typically fixate our gaze on objects, resulting in high-frequency information, while the periphery (scene background) is more blurred. Studies have used this information to increase the efficiency of visual search models[150,151]. Furthermore, training DNNs on such “foveated” images, mimicking human vision, has been shown to improve their object classification compared to when they are trained on full resolution images[152].

While all of the aforementioned examples rely on either altering the input or the training procedure of DNNs, some recent developments have focused on changing the architecture of DNNs so as to incorporate context-processing modules, analogous to how humans have separate pathways for object and scene processing. These context modules rely on receptive fields of varying spatial scales which are later fused to create global context. These approaches appear promising, leading to superior object detection accuracy, with improvement most pronounced for small objects[153,154]. Yet another route to model representations of scenes may be through the use of language; multimodal models such as CLIP (Contrastive Language-Image Pre-training) appear promising at capturing higher-level visual representations[155]. Thus, language may be a pathway to access rich semantic, or schema-like, representations of scenes (Fig 6).

Altogether these examples serve to illustrate how taking insights from human cognitive neuroscience into account may lead to both improved and more human-like object identification in DNNs. Building-in context sensitivity would also make machine vision more human-like. This is relevant for the usefulness of these networks as scientific models of human visual processing, but potentially also for applications in which their output is used by human operators.

Europe PMC Funders Author Manuscripts

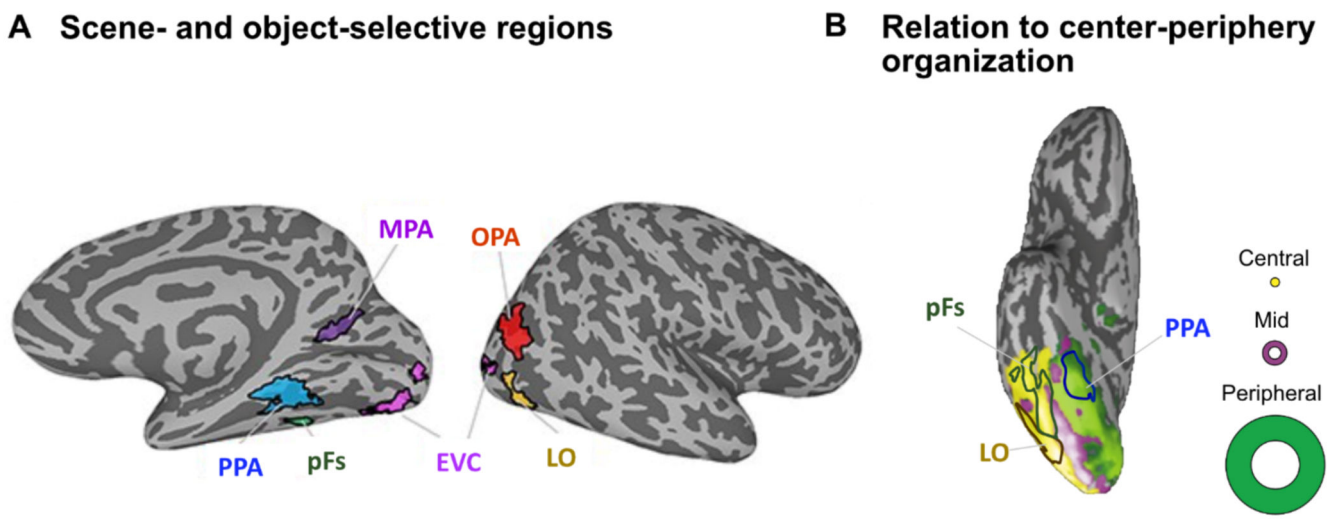

Fig. 1. Scene- and object-selective regions in human visual cortex and their relation to the center-periphery organization.

a) Medial (left) and lateral (right) views of scene- and object-selective regions in the human visual cortex. Scene-selective regions include MPA (medial place area), OPA (occipital place area), and PPA (parahippocampal place area). Object-selective regions include pFs (posterior fusiform gyrus) and LO (lateral occipital cortex). EVC: early visual cortex (adapted from[29]). **b**) Ventral view of scene- and object-selective regions and their relation to the center-periphery organization. The scene-selective PPA is biased towards peripheral visual input, while the object-selective pFs and LO are biased towards central visual input (adapted from[10,30]).

Europe PMC Funders Author Manuscripts

Europe PMC Funders Author Manuscripts

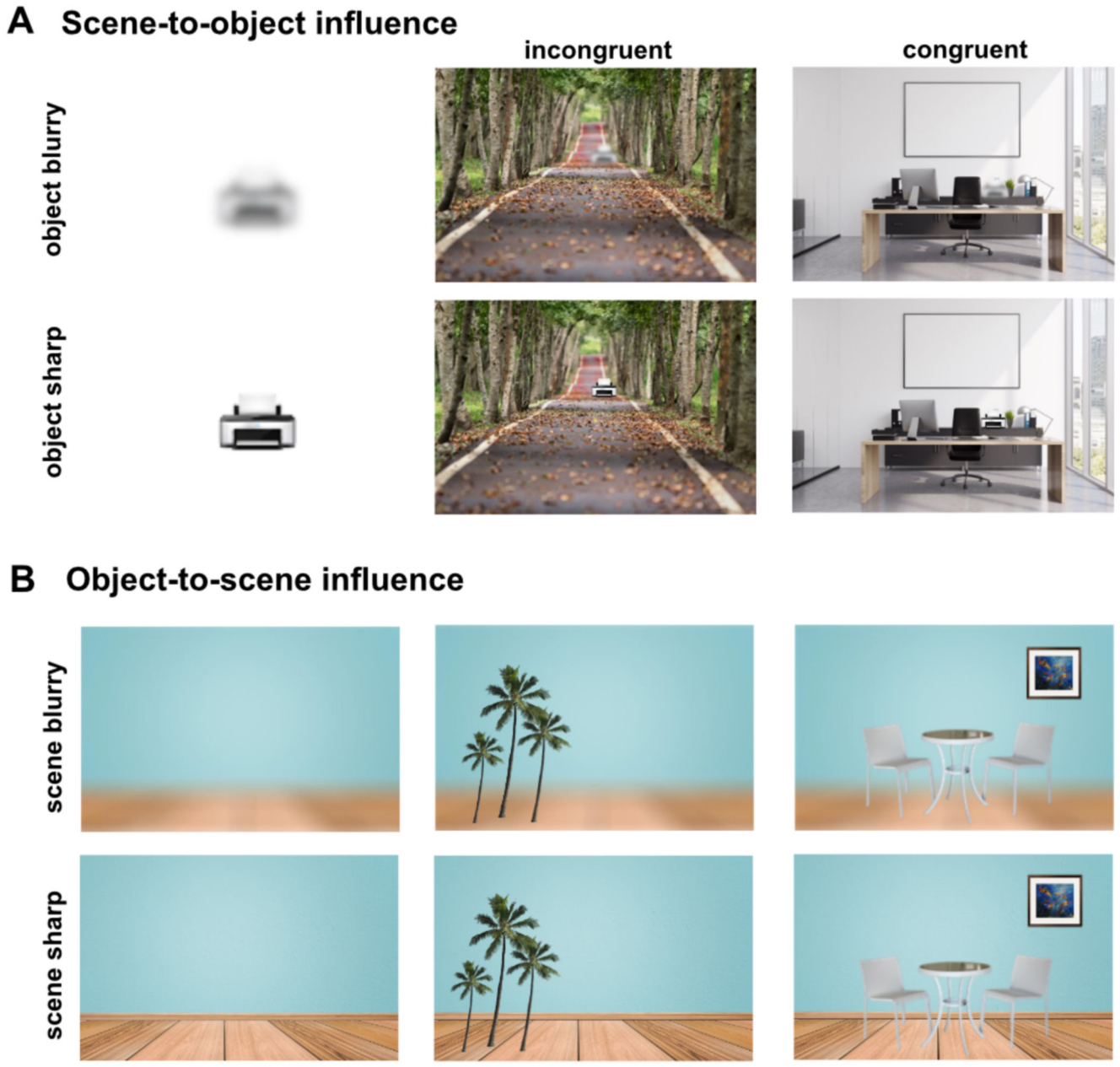

Fig. 2. Bidirectional interactions between objects and scenes.

a) Scene context can shape object perception, particularly when an object is ambiguous ("object blurry", top row). In this example, the ambiguous object is perceived either as a car or as a printer depending on the scene context. However, when the object is sharp (bottom row), it can be recognized based on local features alone, reducing the influence of context. In that scenario, an incongruent object (a printer on a road) is surprising and receives more attention. **b**) Objects can also shape scene perception, following similar principles. Here, the ambiguous scene (top row) is perceived as an outdoor (open) or indoor (closed) space depending on the objects. When the scene is sharp, the object influence is reduced (bottom row).

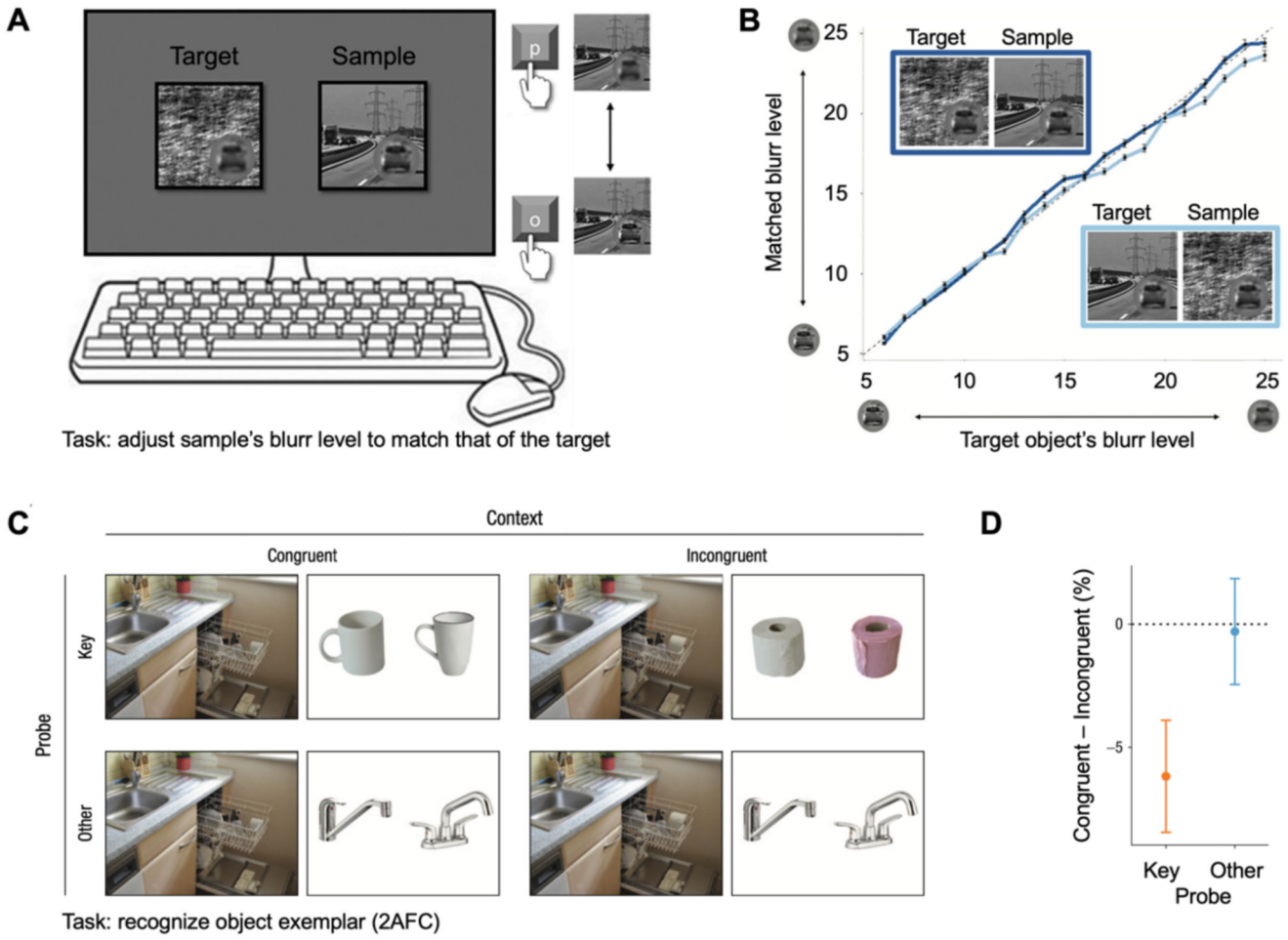

Fig. 3. Scene context can both facilitate and impair object perception.

a) Paradigm to investigate the perceived sharpness of objects in scenes. Participants adjusted the blur level of a sample object (the car) to match that of a target object[42]. **b**) More blur was added to objects when they were viewed within a coherent scene context, indicating that those objects were initially perceived as sharper. Congruent scene context thus facilitates object perception. **c**) Paradigm used to investigate the perception of unambiguous objects as a function of semantic congruency[57]. After viewing a scene for 2.5 s, participants had to indicate which of the two exemplars had been presented in the scene. *Key* objects could be congruent or incongruent with the scene (top row). To control for general effects of congruence, control objects (*Other*) were also tested (bottom row). *Other* objects were always congruent with the scene but were presented in scenes that also contained congruent or incongruent *Key* object. **d**) Results showed a congruency cost, such that participants were more accurate at recognizing objects that were incongruent with the scene. No effect of congruency was found for the control objects. In this case, congruent scene context impaired object perception. Error bars show 95% confidence intervals.

Europe PMC Funders Author Manuscripts

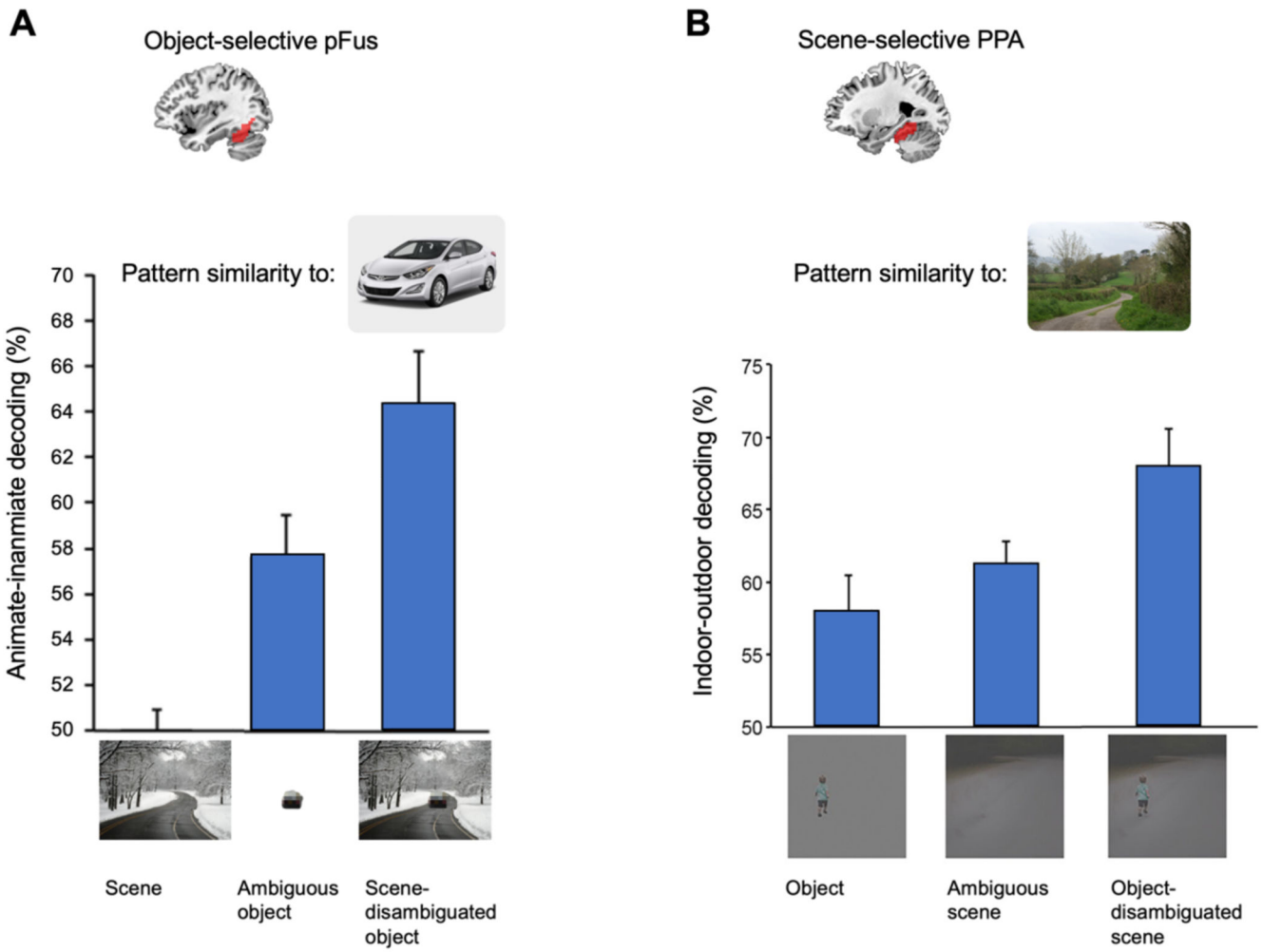

Fig. 4. Neural evidence for bidirectional interactions between object and scene processing in the visual cortex.

a) To test whether scene context modulates representations in the visual cortex, participants viewed ambiguous objects with ("Scene-disambiguated object") and without ("Ambiguous object") scene context while brain activity was measured using fMRI. Multivariate activity patterns in the object-selective visual cortex in response to the ambiguous objects were classified as animate vs inanimate categories based on activity patterns evoked by clearly visible objects (illustrated by the picture in inset), presented in a separate experimental run. Results showed that the presence of scene context increased decoding accuracy (i.e., third bar higher than second bar)[28]. These results may reflect a neural correlate of the perceptual sharpening illustrated in Fig. 3ab. **b**) Similar effects were observed for the reverse influence, with objects modulating scene representations in the scene-selective cortex. In this case, an object disambiguated the scene, such that response patterns evoked by ambiguous scenes became more similar to clearly visible scenes presented in a separate experimental run[66]. Error bars indicate standard error of the mean.

Europe PMC Funders Author Manuscripts

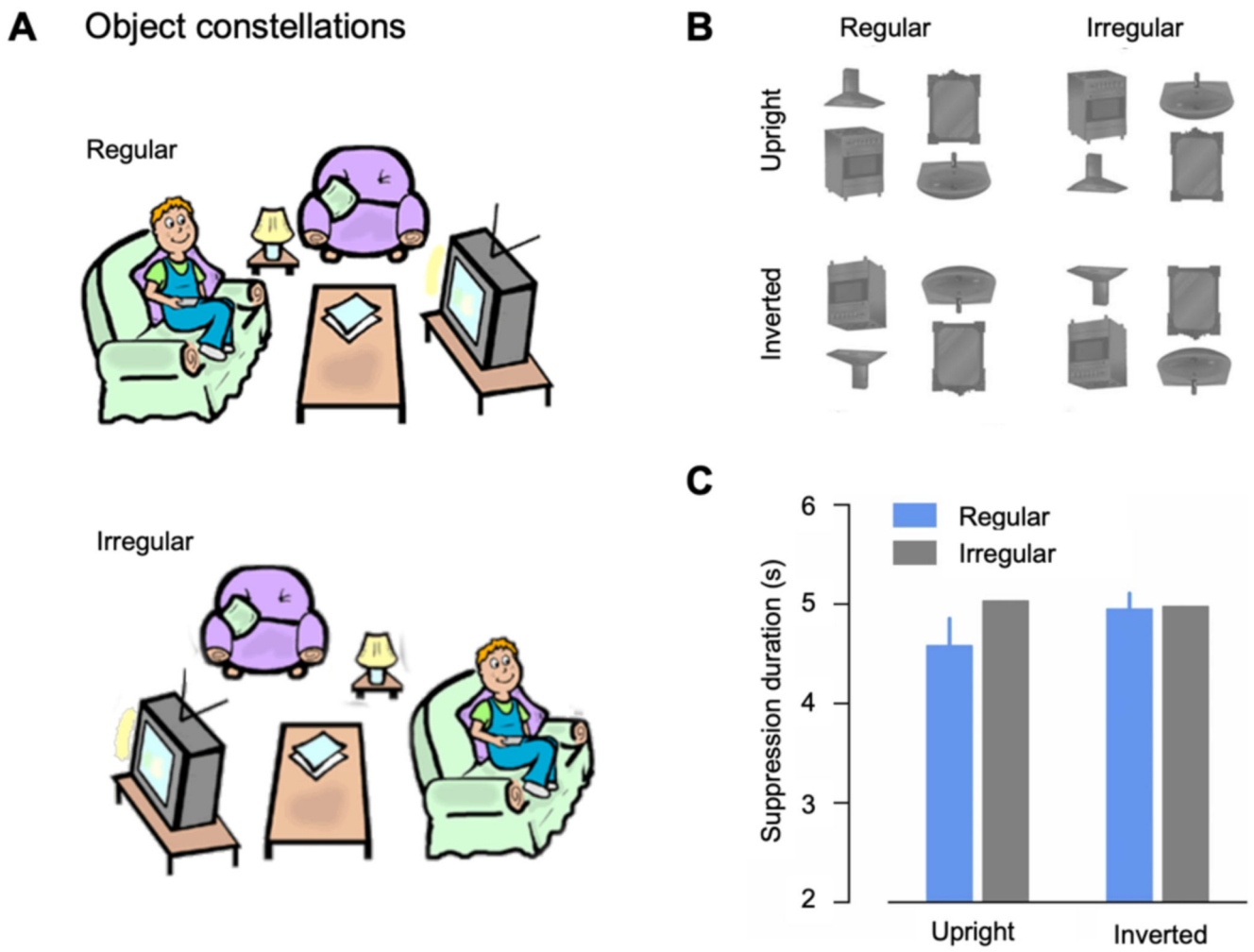

Fig. 5. Object constellations.

a) Objects are often seen together with other objects in familiar spatial arrangements, as illustrated here for a living room set. Neuroimaging studies have provided evidence for integrative representations of regular object arrangements in the ventral visual cortex[115,116]. **b**) Example stimuli used to test whether regular object arrangements are detected more quickly than irregular object arrangements in a breaking continuous flash suppression experiment[108]. Inverting the objects serves as a control for possible low-level stimulus differences. **c**) Results showing that regular object displays broke suppression (i.e., were detected more quickly) than irregular displays. No such effect was found for inverted controls. Error bars show 95% confidence intervals of the mean difference between regular and irregular conditions.

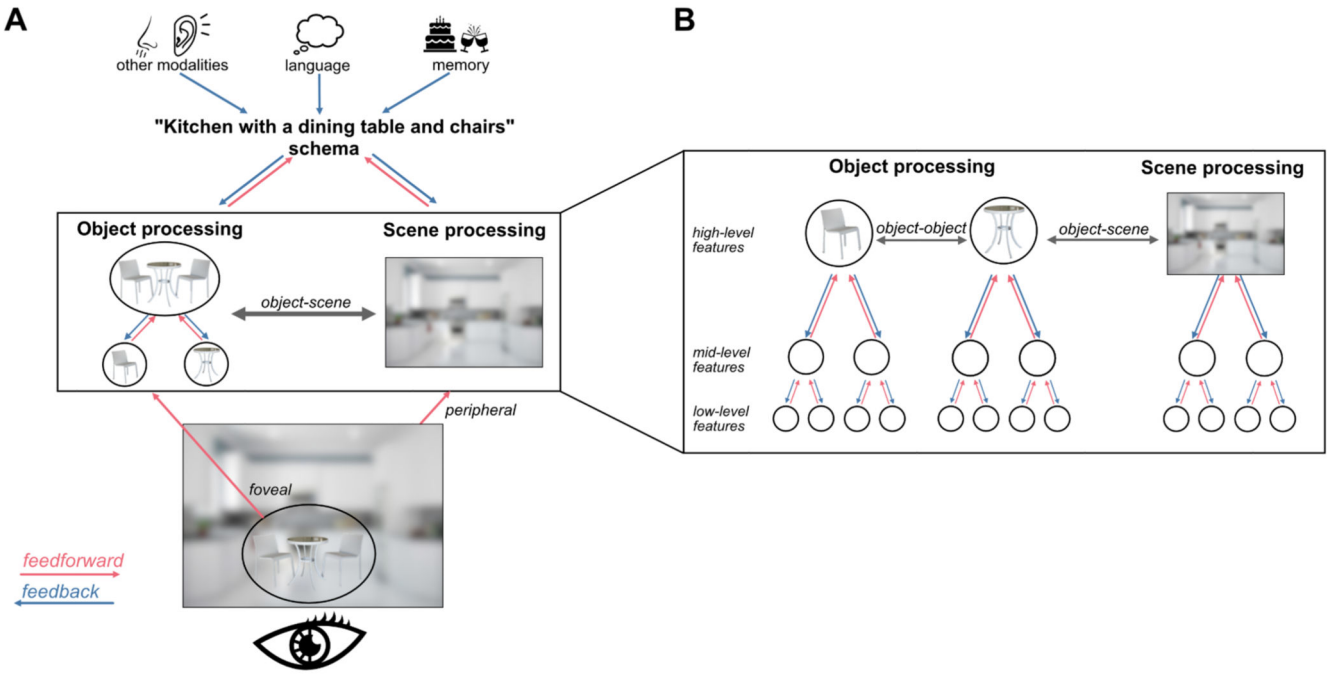

Fig. 6. An integrated model of object-scene and object-object interactions.

a) General overview of visual processing, from initial perceptual processing (bottom) to semantic-level representations of scene schemas (top). Object processing is primarily informed by foveal (local) input, while scene processing is primarily informed by peripheral (global) input. Each processing pathway is hierarchically organized, with feedforward and feedback connections, indicated by red and blue arrows, respectively. Both pathways project to, and receive information from, higher-order regions containing semantic information of object-scene schemas (mental models in long-term memory). **b**) Schematic illustration of proposed interactions within and between object and scene processing pathways. Hierarchical organization of each pathway is indicated with circles, which contain low-, mid-, and high-level features of object and scene processing. Cross-pathway interactions may be most effective at higher levels of the hierarchy, but may also exist at lower levels (not illustrated). Subsequent feedback within each pathway can result in modulations at lower levels of the processing hierarchy and result in perceptual sharpening.

Europe PMC Funders Author Manuscripts