A new framework for understanding vision from the perspective of the primary visual cortex

Li Zhaoping
University of Tübingen, Max Planck Institute for Biological Cybernetics, Tübingen, Germany

published in *Current Opinion in Neurobiology*, 2019, 58:1-10

Abstract:

Visual attention selects only a tiny fraction of visual input information for further processing. Selection starts in the primary visual cortex (V1), which creates a bottom-up saliency map to guide the fovea to selected visual locations via gaze shifts. This motivates a new framework that views vision as consisting of encoding, selection, and decoding stages, placing selection on center stage. It suggests a massive loss of non-selected information from V1 downstream along the visual pathway. Hence, feedback from downstream visual cortical areas to V1 for better decoding (recognition), through analysis-by-synthesis, should query for additional information and be mainly directed at the foveal region. Accordingly, non-foveal vision is not only poorer in spatial resolution, but also more susceptible to many illusions.

Highlights

- A new framework for vision: encoding, selection, and decoding
- A saliency map in V1 for primates guides attentional selection exogenously
- Massive loss of input information, i.e., the attentional bottleneck, starts at V1
- Peripheral/central vision mainly for looking (selection)/seeing (decoding)
- Feedback from higher visual areas to V1 is mainly directed at fovea for decoding

Introduction: considerations from the attentional bottleneck and the primary visual cortex

A common view frames vision as composed of low-level, mid-level, and high-level vision. This framework is imprecise, and is thus unfalsifiable and ineffective to guide research. David Marr[1] advocated a more falsifiable framework: vision is composed of a primal sketch, a 2.5-dimensional sketch, and a three-dimensional model of the visual world. This framework suffers from omitting the attentional bottleneck, which selects a tiny fraction of visual input information for further processing and makes us almost blind[2] to unattended inputs (Box 1). We are intuitively unaware of this dramatic blindness since we do not know what it is like without this blindness[3]. However, gating our perception, selection should take center stage in vision.

Box 1: Key observations and concepts

- Human retina receives about 100 megabytes (MB) (10^9 bits) per second (s) of visual input data[4], from about $\sim 10^7$ cones. 100 MB can store about 40,000 pages of text uncompressed.
- Human retina sends about one MB/s of visual data to the central brain[5••] from 10^6 retinal ganglion cells. This rate is lower than the retinal input data rate largely by efficient encoding or compression of retinal inputs.
- Visual attention selects a tiny fraction of input data for further processing, often by overtly directing gaze, head, limbs, whiskers, snout, and/or other body parts to the selected visual field or object location. Visual inputs at the selected location are said to be within the attentional spotlight, evoking faster and more accurate neural and behavioral responses[6, 7, 5••].
- Inattentional blindness: humans are often unaware of the presence of unattended objects in the visual scenes[2].
- Human attentional bottleneck is about 40 bits/second[8, 4], due to the inattentional blindness. 40 bits can store one to two short sentences of English text. This implies that more than 99% of the visual input information does not enter into our perceptual awareness, see Figure 1.
- Central versus peripheral vision: peripheral vision is not only worse in spatial acuity, but is also vulnerable to contextual crowding[9, 10], see Box 2.
- The primary visual cortex (V1) is the only visual cortical area with a substantial fraction of neurons that are monocular or eye-dominant so that their responses depend on the eye-of-origin of visual inputs in binocular visual fields[11, 12]. Most humans cannot discriminate the eye-of-origin of input. It has been argued that V1 activities are unrelated to visual awareness[13, 14].
- Attentional selection and V1: An eye-of-origin singleton, e.g., an item uniquely shown to the right eye in a background of items shown to the left eye, captures human gaze and attention even when the singleton appears non-distinctive[15, 16] (Figure 2b). Fluctuations of monkey V1's initial responses to salient singletons are correlated with the speeds of the subsequent gaze shifts to these singletons[17••] (Figure 2c).

A new framework[5••] views vision as having encoding, selection, and decoding stages (Figure 1). Encoding measures or samples visual inputs, and represents the inputs more efficiently (by data compression without substantial information loss). Selection chooses a tiny fraction of input information for further processing. Decoding infers from the selected information the properties of the three-dimensional visual world, such as objects and people.

In primates, selection is often by gaze shifts to place the selected visual field location at the fovea in the central visual field. It can be guided by top-down knowledge or goals, such that gaze is directed to an anticipated object location (e.g., when reading a book) or by task requirements (e.g., looking for red when searching for a red cup). It is also guided in a bottom-up manner, e.g., when gaze involuntarily shifts to a moving object outside the page of the book being read. Selecting the destination for gaze shifts can be achieved *before* the recognition of the object at this destination, particularly during the bottom-up selection (e.g., [18, 15]). The attentional bottleneck precludes recognizing all objects in the scene before selecting an object for scrutiny.

Selection implies that non-selected information is deleted. This starts at the primary visual cortex (V1), the first cortical area along the visual pathway[19•, 20••, 17••]. V1 creates a saliency

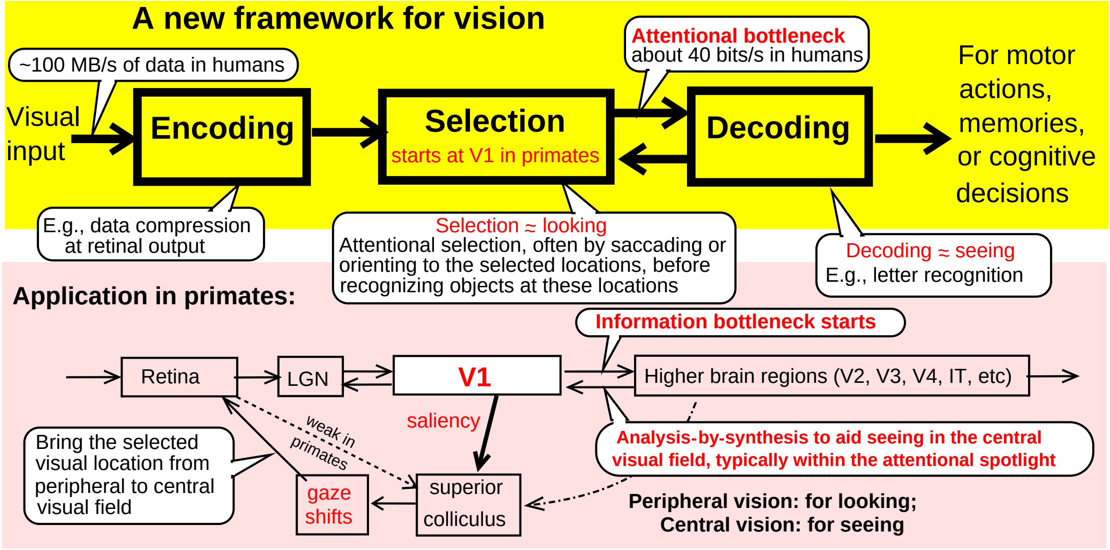

A new framework for vision
~100 MB/s of data in humans
Visual input
Encoding
Selection
starts at V1 in primates
Attentional bottleneck
about 40 bits/s in humans
Decoding
For motor actions, memories, or cognitive decisions
E.g., data compression at retinal output
Selection ≈ looking
Attentional selection, often by saccading or orienting to the selected locations, before recognizing objects at these locations
Decoding ≈ seeing
E.g., letter recognition
Application in primates:
Information bottleneck starts
Retina
LGN
V1
Higher brain regions (V2, V3, V4, IT, etc)
saliency
weak in primates
Bring the selected visual location from peripheral to central visual field
gaze shifts
superior colliculus
Analysis-by-synthesis to aid seeing in the central visual field, typically within the attentional spotlight
Peripheral vision: for looking;
Central vision: for seeing

Figure 1: The new framework places selection at the center stage of vision, motivated by the following observations: (1) the attentional bottleneck selects only a tiny fraction of visual input for deeper processing, (2) selection is often through gaze shifts to place the selected visual locations in the central visual field, (3) selection starts at V1, which creates a bottom-up saliency map to guide saccadic gaze shifts. It suggests a dramatic loss of visual information starting at V1's outputs, qualitatively different functional roles for the central and peripheral visual fields, and that feedback from higher visual areas to V1 should mainly target the foveal representation to facilitate object recognition.

map to guide gaze to the most salient location, with saliency at a location defined as the strength of this location to attract bottom-up selection. V1's role in selection has not been realized until recently, this may be why research progress in extrastriate areas (which receive inputs from V1) has been much slower than that in V1, the lateral geniculate nucleus (LGN), and the retina over the last 50 years[21]. That is, if visual input information is not reproducibly transmitted beyond V1 due to the attentional bottleneck, measurements of neural receptive fields in extra-striate cortices, e.g., V2, V4, IT, etc, would be very difficult. For example, if a receptive field property at a visual location is measured by correlating visual inputs with neural responses over many trials, it would be difficult when this location is attentionally selected in some trials and not in other trials while visual inputs vary across trials. That bottom-up selection is performed by V1 further motivates placing selection on center stage for vision. Vision up to V1 can be largely understood as performing data compression (at the encoding stage) and data selection[22].

The new framework provides new perspectives and motivates new questions (Box 2). For example (Figure 1), it suggests that selecting and decoding could be understood as looking and seeing stages, respectively, so that looking selects the location for gaze direction, while seeing recognizes the object at the selected location. The selected location is typically in the peripheral visual field before being selected, suggesting that peripheral vision is mainly for selecting or looking, while central vision is mainly for seeing or recognizing. Indeed, this is demonstrated in a visual search task in an image designed to pit seeing (when the search target is recognized in the fovea) against looking (when the target attracts gaze from visual periphery before its shape is recognized)[18].

More generally, central and peripheral vision should differ qualitatively, rather than just quantitatively in (e.g.) visual spatial acuity. For instance, unlike foveal vision, peripheral vision is vulnerable to crowding, so that a peripheral object, e.g., a letter, is harder to recognize when it is in a cluttered background[9, 10] (Box 2). In the new framework, we postulate that central and peripheral vision differ, as manifested in crowding and other related phenomena, because of a feedback route from the decoding to the selection stage, as follows (Figure 1 and Box 2): Since selection starts at V1, feedback from higher visual areas to V1 should extract from V1 additional information (that was originally not sent through the feedforward route) to aid object recognition (using analysis-by-synthesis), especially in challenging recognition tasks. Since seeing is mainly for the central visual field, top-down feedback to V1 for object recognition (mainly in the ventral visual stream) should be weaker or absent in the peripheral visual field. Next, I review findings on V1's role in selection to motivate the new framework, and expand on the new perspectives and on the central-peripheral dichotomy as the first example of studies guided by this framework.

Box 2: New perspectives

- Dramatic visual input information loss starts at V1's outputs, it is desirable to quantify the amount of this loss at this first selection stage. The downstream areas along the visual pathway must be understood in light of the information selection at V1.

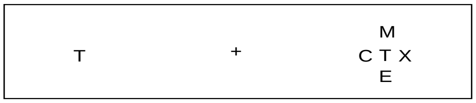
- Feedbacks to V1 from the downstream areas query for additional information for decoding using analysis-by-synthesis.

- The central-peripheral dichotomy:

Peripheral vision is mainly for looking, central vision is mainly for seeing.

Feedback to V1 for decoding or recognition may be weaker or absent in peripheral visual field, which is therefore more prone to crowding and visual illusions such as reversed depth in anticorrelated stereograms (Figure 3b) and feature misbinding.

Crowding in peripheral visual field:



fixating on the '+' above, one recognizes the 'T' to the left, but not the 'T' to the right

Visual illusions arising from feedforward V1 signals or from top-down feedback processes (for analysis-by-synthesis) should be more likely to occur in peripheral or central visual field, respectively.

The selection, loss, and processing of visual input information

A saliency map in V1 for primates to guide bottom-up selection

Noting that V1 neural activities can serve as a universal currency to bid for attentional selection regardless of the neural input selectivities, I proposed that the (bottom-up) saliency map is created in V1[19•, 20••] (Figure 2a). Accordingly, the saliency at any location in a scene is the highest V1 response to input at that location relative to the highest responses to the other locations. V1 responds more vigorously to salient feature singletons than to the non-salient background items because iso-feature suppression mediated by intracortical connections makes V1 neurons tuned to the same or similar features suppress each other[23, 19•]. For example, nearby V1 neurons tuned to the same orientation suppress each other. Thus, in an image (Figure 2a) containing many right-tilted (from vertical) bars in the background, the responses to the background bars are lower than the response

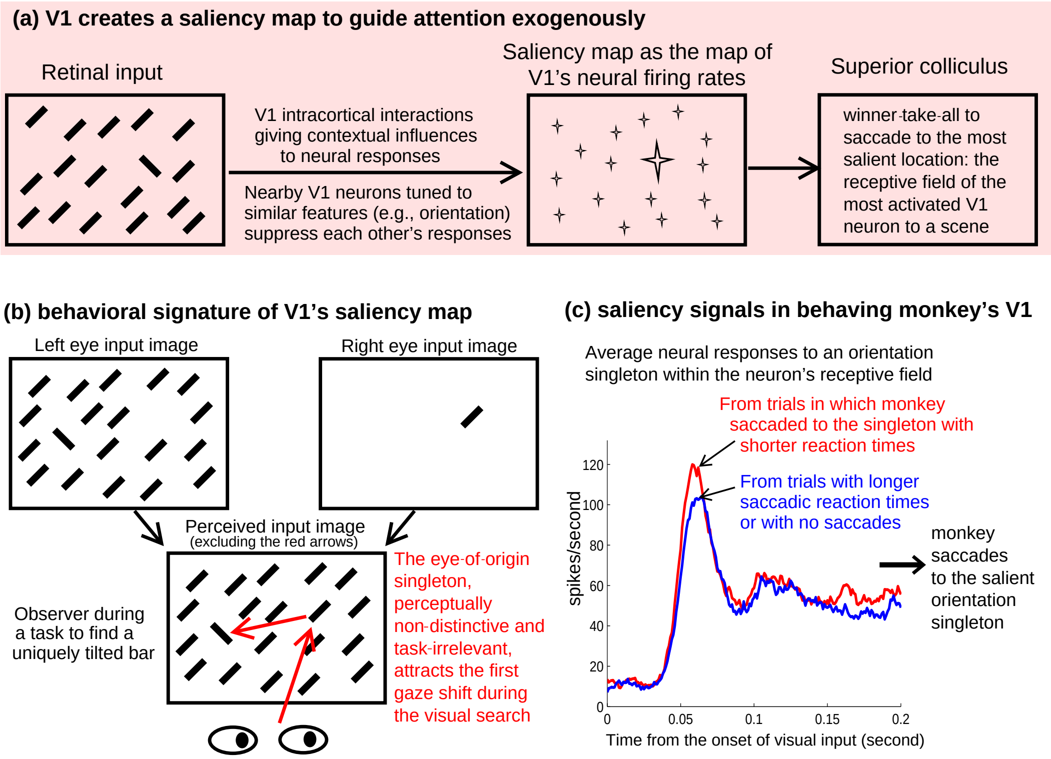

(a) V1 creates a saliency map to guide attention exogenously
Retinal input
V1 intracortical interactions giving contextual influences to neural responses
Nearby V1 neurons tuned to similar features (e.g., orientation) suppress each other's responses
Saliency map as the map of V1's neural firing rates
Superior colliculus
winner-take-all to saccade to the most salient location: the receptive field of the most activated V1 neuron to a scene
(b) behavioral signature of V1's saliency map
Left eye input image
Right eye input image
Perceived input image
(excluding the red arrows)
Observer during a task to find a uniquely tilted bar
The eye-of-origin singleton, perceptually non-distinctive and task-irrelevant, attracts the first gaze shift during the visual search
(c) saliency signals in behaving monkey's V1
Average neural responses to an orientation singleton within the neuron's receptive field
From trials in which monkey saccaded to the singleton with shorter reaction times
From trials with longer saccadic reaction times or with no saccades
monkey saccades to the salient orientation singleton
spikes/second
0
20
40
60
80
100
120
0
0.05
0.1
0.15
0.2
Time from the onset of visual input (second)

Figure 2: V1 saliency hypothesis and its supporting evidence. (a) V1 is proposed to transform visual input to saliency signals via contextual influences: a neuron's response to inputs within its classical receptive field (RF) is suppressed when similar inputs (e.g., inputs with similar orientations) are outside its RF. Hence, feature singletons, e.g., orientation singletons, tend to evoke the highest response to a scene. V1's firing rate is the saliency signal read by superior colliculus to guide saccade. (b) an ocular singleton, unique in its eye-of-origin, is predicted to be salient and to attract gaze (red arrows indicate gaze shifts), as confirmed subsequently[15, 16]. (c) monkey V1 neural responses to an orientation singleton (within the RF) tilted 90^o away from uniformly oriented background bars (as in (a)), while the monkey tries to find and saccade to an orientation singleton as quickly as possible (usually ≥ 200 ms after stimulus onset)[17••]. Given the singleton, faster saccades towards the singleton are typically preceded by higher initial peak responses (around 40-50 ms after stimulus onset) of the V1 neurons (tuned to orientations close to that of the singleton) to the singleton within their RFs.

to the salient orientation singleton, the uniquely left-tilted bar, because the V1 neuron responding to this singleton escapes the iso-orientation suppression. Analogously, iso-color suppression and iso-motion-direction suppression make salient a color and motion-direction singleton, respectively.

Data supporting this hypothesis have since emerged. In particular, they confirmed its surprising prediction that an eye-of-origin singleton is salient since it escapes the iso-eye-of-origin suppression. This singleton is for example an item uniquely shown to the right eye among many items shown to the left eye (Figure 2b). If it is unique only in terms of its eye-of-origin, it can scarcely be distinguished by its appearance. Nevertheless, it is so salient as to attract gaze more strongly than a salient and highly distinctive orientation singleton favored by the top-down task goal, as schematized in Figure 2b[15, 16]. This, an example of looking without seeing, provides a hallmark that the saliency map is in V1 rather than extrastriate cortical areas, since V1 is the only cortical area tuned to eye-of-origin[11, 12] to make this feature impact saliency.

More direct support for this V1 theory comes from monkey V1 responses to an orientation singleton while the animal searches for the singleton to saccade to it[17••]. Given the same visual input, neural and behavioral responses fluctuate between trials stochastically. When the monkey saccades to the target by a shorter saccadic reaction time (RT) in a trial, the initial neural response to this target before the saccade is more likely higher (Figure 2c). These neural responses, at a short latency of 40-50 ms after the visual input appears, are unlikely due to feedback from downstream regions, such as the superior colliculus[24•] or the frontal eye field; they are more likely direct causes of the saccade via monosynaptic projections to the superior colliculus (Figure 1).

Lesioning V1 in monkeys eliminates visually guided saccades, but not memory guided saccades (via the frontal eye field[25]), for at least several weeks[25, 26]. Hence, retinal projections to the superior colliculus are normally insufficient for selection in primates. LGN, not projecting to the superior colliculus (Figure 1), is unlikely to impact gaze shifts directly, although massive top-down inputs from the cortex to LGN suggest that brain states such as arousal could influence visual information transmission via LGN.

A more stringent test of the V1 saliency hypothesis comes via a parameter-free quantitative prediction[27], briefly derived as follows: First, from the hypothesis, the saliency of a visual location is determined by the highest V1 neural response to the input at that location relative to the highest responses to the other locations. Second, consider three images of bars of the same size, they have the same green-horizontal background bars and differ only in the singleton bar, which is red-vertical (unique in both color and orientation) in image 1, red-horizontal (unique in color) in image 2, and green-vertical (unique in orientation) in image 3. Imagine that V1 had no neurons tuned to both color and orientation, by the hypothesis, the saliency at the double-feature singleton (in image 1) equals the maximum of the saliency values at the two single-feature singletons in the other two images. Since a higher saliency means a shorter RT for an attentional shift, one predicts, without parameters, the RT to the double-feature singleton in image 1 as the shorter of the two RTs to the respective single-feature singletons in the two other images. Although V1 does have neurons tuned simultaneously to color and orientation, it lacks neurons tuned simultaneously to color, orientation, and motion direction. Extending the derivations above, one predicts, parameter-free, the probability distribution of the RTs to the triple-feature (color, orientation, and motion direction) singleton from the experimentally measured RTs to the corresponding single-feature and double-feature singletons. This prediction is invalid for extrastriate cortices since they do have neurons tuned simultaneously to the three feature dimensions. The confirmation[27, 28] of this prediction suggests that the downstream areas do not contribute to computing saliency.

Information loss starts at V1

A saliency map in V1 for selection suggests that non-selected information is lost starting from the output of V1. For example, the information about the eye-of-origin of inputs is lost by V2, which has mainly or only binocular neurons[11, 12]. Consequently, typical humans cannot discriminate eye of origin, even after long term training[29]. Information about the exact spatial positions or contrast polarites of visual input is also partly lost when complex (rather than simple) cells convey outputs from V1, since these cells' responses are insensitive or invariant to such information. Future work should examine further the nature of the information loss[30•], and quantify how much of the total information loss by selection has already occurred by the output of V1.

This loss continues downstream, for example, in fine spatial information. There is almost a 3-fold increase in the average receptive field size in monkey V2 as compared to V1[31], and in V2, but not in V1, simple cells prefer smaller spatial frequencies than the complex cells[31]. V2 responses are more invariant than V1 responses to details in visual textures[32]. Meanwhile, visual cortical areas become smaller downstream along the visual pathway, each of them may receive and process less information, although they could process information parallelly. Examining the amount and the nature of progressive information selection along the visual pathway by the attentional bottleneck should help to understand the underlying computation.

The central-peripheral dichotomy

The progressive information loss by selection, starting at V1, and the selection by gaze shifts motivate the following two proposals[33••]. First, central and peripheral vision are mainly for seeing (decoding or recognizing) and looking (selecting), respectively. Second, feedback from higher to lower visual areas such as V1 to facilitate seeing is mainly directed to the central visual field. Note that saliency to shift gaze from the current gaze position is irrelevant for fovea, which is the current gaze position. Hence, the saliency computation in V1 is only significant for non-fovea locations (see discussions in [5••]), leaving V1's foveal representation better focused on decoding and on top-down selection. The top-down feedback should be part of the task-dependent feedback route from decoding to selection (Figure 1), allowing extraction of more information about the currently attended object (e.g., Figure 3a) or segmentation of the attended object from the sensory background. Such top-down feedback can find analogies in other senses such as olfaction[34].

Hypothesis: Top-down feedback for object recognition is weaker in the peripheral visual field

The top-down feedback can be probed by biased perceptions of the ambiguous dichoptic inputs like that in Figure 3a[33••]. These inputs are such that the summation and difference of visual inputs to the two eyes are respectively gratings tilted in opposite directions from horizontal. The ocular summation and difference signals are V1's representation of left and right eye inputs[36]. In any case, such dichoptic inputs evoke an ambiguous percept of the direction of the grating's tilt. The visual system presumably deals with the ambiguity by analysis-by-synthesis as follows[33••]. First, V1 feeds forward the initial hypotheses about the visual input properties; e.g., in Figure 3a, the directions of tilt in ocular summation and difference channels are fed forward as two conflicting initial hypotheses — clock- and anti-clockwise — about the tilt direction. Second, using the internal knowledge of the visual world, the higher visual areas synthesize visual inputs consistent with each hypothesis. Third, the synthesized inputs are fed back to V1 to verify the match between the synthesized and the actual visual inputs. Fourth, each hypothesis is reinforced or weakened if the respective match is good or poor, and the alternative hypotheses compete with each other for the perceptual outcome. This process is termed feedforward-feedback-verify-(re)weight, or FFVW[33••].

Before applying the FFVW process to Figure 3a, we note that V1 can send information to downstream areas about the input tilt or other non-ocular visual features from both the ocular summation and ocular difference channels[37, 38], without indicating which channel was responsible. This is because V1 does not send eye-of-origin information downstream. The two tilt signals (from the summation and difference channels) are likely treated by vision as two views of the same scene property, since in natural scenes they are typically consistent with each other in tilt direction although the one from the ocular difference channel is typically much weaker in signal-to-noise[33••]. The stimulus in Figure 3a is designed to violate the typical consistency between the two tilt signals to allow us to probe the top-down feedback. For such inputs, the higher brain areas generate for each of V1's feedforward hypotheses, clock- and anti-clockwise tilts, a synthe-

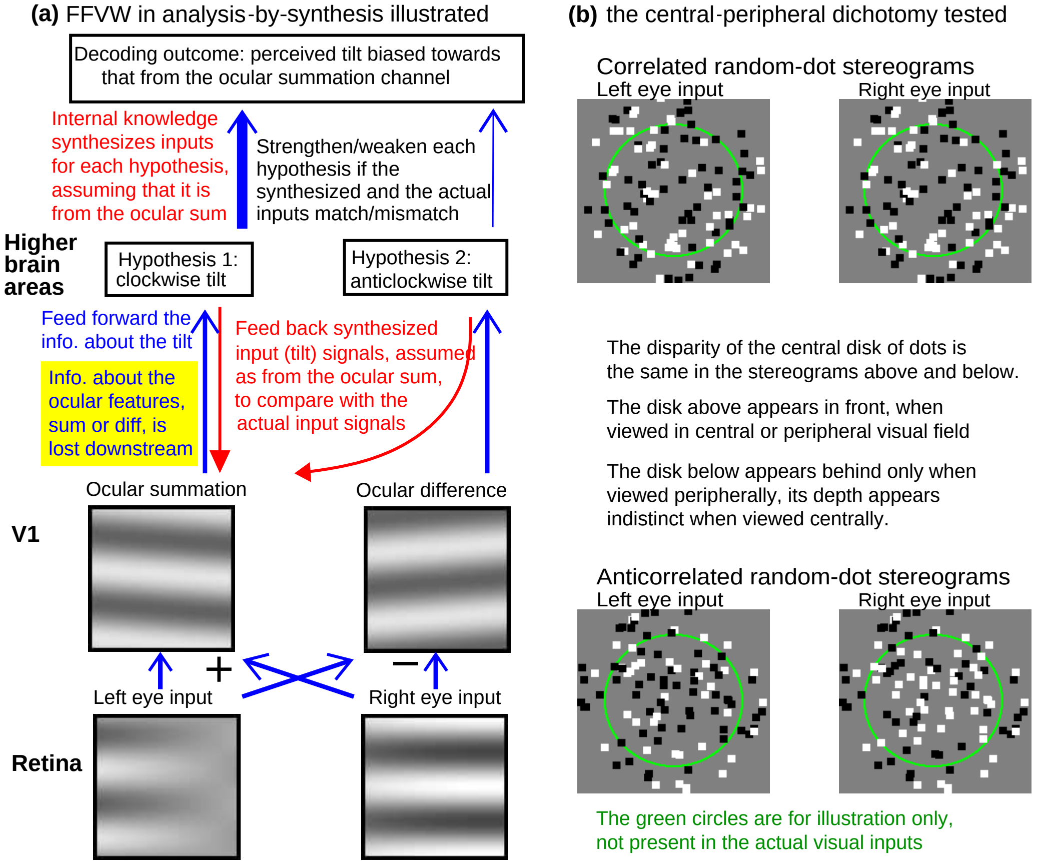

(a) FFVW in analysis-by-synthesis illustrated
Decoding outcome: perceived tilt biased towards that from the ocular summation channel
Internal knowledge synthesizes inputs for each hypothesis, assuming that it is from the ocular sum
Strengthen/weaken each hypothesis if the synthesized and the actual inputs match/mismatch
Higher brain areas
Hypothesis 1: clockwise tilt
Hypothesis 2: anticlockwise tilt
Feed forward the info. about the tilt
Info. about the ocular features, sum or diff, is lost downstream
Feed back synthesized input (tilt) signals, assumed as from the ocular sum, to compare with the actual input signals
Ocular summation
Ocular difference
V1
+
−
Left eye input
Right eye input
Retina
(b) the central-peripheral dichotomy tested
Correlated random-dot stereograms
Left eye input
Right eye input
The disparity of the central disk of dots is the same in the stereograms above and below.
The disk above appears in front, when viewed in central or peripheral visual field
The disk below appears behind only when viewed peripherally, its depth appears indistinct when viewed centrally.
Anticorrelated random-dot stereograms
Left eye input
Right eye input
The green circles are for illustration only, not present in the actual visual inputs

Figure 3: Top-down feedback for analysis-by-synthesis probed by perceptual biases in ambiguous perception (a) and tested for the central-peripheral dichotomy (b). (a) the feedforward-feedback-verify-(re)weight (FFVW) process for analysis-by-synthesis in the example of ambiguous perception of the tilt direction in a dichoptic stimulus. Two conflicting hypotheses (clockwise and anti-clockwise) about the direction of the grating's tilt from horizontal are fed forward from V1. Higher visual areas verify each hypothesis by synthesizing the would-be visual inputs for the hypothesis. The internal model of the visual world presumes that the would-be input is in the ocular summation channel and dictates the synthesis accordingly. The would-be input is compared with the actual input in the ocular summation channel in V1 to verify whether they match in tilt direction. Then the weight of the original hypothesis in the ultimate perceptual outcome is enhanced or suppressed when the match is good or bad, respectively. This FFVW process biases the percept towards visual inputs (e.g., from the ocular summation channel) that are consistent with the internal model. (b) both correlated and anticorrelated random-dot stereograms (RDSs) excite V1 neurons, which feed forward to report the disparity (depth) information in the RDSs. However, anticorrelated RDSs disagree with the internal model of the visual world, making the reported reversed depth an illusion vetoed by the top-down feedback in the FFVW process. Hence, this illusion is more visible in peripheral vision[35••] where the feedback is weaker or absent.

sized input of this tilt signal. Furthermore, the synthesis process lacks the eye-of-origin information from V1 and just assumes that the tilt signal arises from (or should be consistent with) the ocular summation channel, regardless of whether this is actually the case. This assumption is based on the brain's internal knowledge that inputs to the two eyes tend to be identical or similar to each other[36]. Hence, the synthesized tilt signal is compared with the actual tilt signal in the ocular summation channel to verify whether their visual feature values, e.g., tilt directions, match. In Figure 3a, this gives a good or poor match, respectively, for the hypothesis arising from the actual summation or difference channel. Consequently, the respective hypothesis is reinforced or weakened, biasing the percept to the tilt in the summation channel. Hence, this bias can be used as a measure of the top-down feedback in the FFVW process. Psychophysically, this bias is observed when the inputs were viewed in the central visual field, but is weaker or absent in the peripheral visual field[33••]. This central-peripheral difference in the bias is also found in motion and color perception using analogous stimuli, prompting the hypothesis that this top-down feedback is generally weaker or absent in peripheral visual field[33••].

The central-peripheral dichotomy in top-down feedback may underlie the central-peripheral difference in response latencies in inferotemporal cortex (IT)[39•], the central-peripheral difference in V1's connections to the rest of the brain[40], the focus on central vision by IT[41], and the fact that cortical mappings of a retinal location across V1, V2, V3, V4 are physically closer or further from each other when this retinal location is foveal or peripheral, respectively, for more economical connections intercortically. The dichotomy can be tested if the top-down influences in some recent studies (e.g., [42••, 43•]) can be assessed across visual eccentricities.

Feedforward sensory input, the top-down feedback, and visual illusions

According to FFVW, feedback should weaken or veto a perceptual hypothesis that arises from V1 signals that do not largely agree with the internal model of the visual world. Such V1 signals arise from retinal images that are almost always unrealizable from real visual objects, and if the corresponding hypothesis is not vetoed then the resulting percept is visible and is an illusion. Our proposal, that top-down feedback is weaker peripherally, predicts that such illusions are more visible in peripheral visual field. Illusions that are stronger in the peripheral field include the rotating snake illusion[44], the Hermann grid illusion[45], the furrow illusion[46], the curved ball illusion[47], and the reversed Phi motion illusion[48]. Predicting a previously unknown or unseen illusion that can be stronger in peripheral field can provide a falsifiable test to our central-peripheral dichotomy.

This prediction comes from an analogue of the reversed Phi motion illusion — reversed depth, using random-dot stereograms (RDSs). Past observations have shown that humans cannot see the reversed depth; in retrospect, they have involved only the central visual field and are comprehensible since the reversed depth is an illusion from V1 signals that violate our brain's internal models. In the correlated RDS of Figure 3b, the two monocular images are identical to each other except for a disparity between the corresponding retinal locations of the central disk of random dots (enclosed by an illustrative green circle not shown to observers) in the two eyes, making the disk of dots appear in front of the surrounding dots. The anticorrelated RDS in Figure 3b is made in the same way using the same disparity, except that, in the central disk, a white dot in the left eye corresponds to a black dot in the right eye and vice versa, violating the internal model. However, V1 neurons do respond to the anticorrelated RDS, but they report a reversed depth that the central disk is behind the surrounding dots. This is because the preferred disparity of a V1 neuron in anti-correlated RDSs is the negative of its preferred disparity in correlated RDSs[49]. As predicted[35••], peripheral vision sees this illusion of reversed depth more readily, since weaker or absent feedback makes it difficult to verify and veto the illusion.

In our framework, illusions are stronger in peripheral vision when they are caused by V1's feedforward responses to violating inputs. On the other hand, if an illusion or phenomenon is caused by top-down feedback for analysis-by-synthesis, then it should be stronger foveally. For example, visual backward masking, in which perception of a briefly presented target is disrupted by a mask presented around 40-100 ms after, but not before, the target, is believed to arise from a mismatch between the top-down synthesized inputs about the target and the feedforward input arising from the trailing mask[50, 51]. We thus predict that such a backward masking should be weaker in peripheral vision. It will be instructive to categorize systematically visual illusions and phenomena according to whether they are stronger in central or peripheral field, and relate this to whether they arise from the presence or absence of the top-down feedback.

Relation with other works and concluding remarks

Treisman's feature integration by top-down analysis-by-synthesis

Treisman's feature integration theory posits that different object features are integrated or bound to the object when the object is within, but not when the object is outside, the attentional spotlight[52••]. For example, in an image of many red-vertical bars and green-horizontal bars, illusory conjunctions, red-horizontal and green-vertical bars, can be perceived outside the spotlight. In our framework, V1 feeds forward the feature values (red, green, horizontal, and vertical) with impoverished details as to which features are spatially together, giving the following initial hypotheses of color-orientation conjunctions: red-vertical, red-horizontal, green-vertical, and green-horizontal. The top-down synthesis to veto the illusory ones is weaker or absent peripherally; hence, if the fovea has insufficient time to scan the scene, illusory conjunctions are likely. Hence, the analysis-by-synthesis process in FFVW is the mechanism behind Treisman's feature integration, consistent with recent data[53].

Top-down processing in FFVW has the general purpose to distinguish between alternative hypotheses about visual objects, with feature binding being a special example. We distinguish between the central and peripheral visual fields for this processing, while Treisman distinguished between inside and outside the attentional spotlight for feature integration. Central or peripheral field is usually, but not always, inside or outside the attentional spotlight, respectively. In our data, the top-down processing is weak or absent in peripheral vision even when attention is directed to it[33••, 35••]. Whether the top-down processing in the central visual field becomes weaker when attention is directed elsewhere should be examined in future studies.

Analysis-by-synthesis

FFVW is based on the long-standing idea (e.g., [54, 55, 56]) that the brain uses both bottom-up and top-down processes, and uses the internal knowledge for sensory inference. Meanwhile, our framework explicitly motivates the feedback process by the information loss in the feedforward direction starting at V1's output: analysis-by-synthesis in our FFVW is a form of top-down query for additional information from (e.g.,) V1. By confirming (enhancing) or vetoing (weakening) a perceptual hypothesis via feedback verification, neural activities in V1 provide additional information to the perceptual decision stage in the higher brain areas. Furthermore, this top-down query is absent or weaker in the peripheral visual field.

Crowding, summary statistics, and metamers in peripheral visual field

Our information loss starting at V1, particularly in the peripheral visual field which is prone to certain illusions, can be related to observations that pre-attention vision (typically outside the fovea) provides merely shapeless bundles or summary statistics of basic features[57, 58], creating metamers sharing the same summary statistics[59]. In monkeys, relative to V1's receptive fields, V2's receptive fields are larger peripherally[60]. This suggests that the information loss in the periphery becomes progressively greater along the visual pathway, while the summary statistics can also be about higher level visual features[61•]. This information loss in the periphery and the inability to resolve it by top-down feedback should cause visual crowding[10, 9].

Feedforward artificial neural networks imitate only the peripheral vision

Artificial neural networks (ANNs) with feedforward architectures imitate biological visual cortex without (essentially) imitating the recurrent and feedback processes. They can achieve human-like capabilities in visual categorizations that can be rapidly done by humans, suggesting that such categorizations involve only feedforward processing[62]. However, they are much worse than primates in recognizing more challenging visual inputs, e.g., of partially occluded objects. In primate visual areas like IT, discriminative neural responses emerge later by dozens of milliseconds to challenging than non-challenging inputs[63•], suggesting an involvement of recurrent (bottom-up and top-down) processes. Hence, these ANNs imitate only the peripheral-field human vision, lacking the top-down analysis-by-synthesis to fill-in the missing sensory information.

Visual understanding via top-down analysis-by-synthesis

Feedforward ANNs easily misclassify visual inputs when subtle inputs, almost undetectable to humans, are added to original images[64]. The top-down, analysis-by-synthesis-based, prediction of what inputs should be for a perceptual hypothesis is a hallmark of visual understanding. Accordingly, human central vision has this understanding, while feedforward ANNs do not. Peripheral retinotopic regions in higher visual areas seem to send feedback to central retinotopic locations in lower visual areas during recognition of peripheral objects[65, 66]. In the natural behavior of saccading to an attention-grabbing object in the periphery, such feedbacks can serve a generalized FFVW process by directing to the expected foveal location of the same object after the saccade, and thus could endow peripheral vision with visual understanding.

Concluding remark and extensions to other animal species and senses

Motivated by the early visual selection by the attentional bottleneck, vision should be seen as composed of encoding, selection, and decoding stages. The selection and decoding stages, respectively, can be viewed as looking and seeing, and therefore seeing should be understood in light of the selection. In primates with a fovea, selecting is typically via shifting gaze to the selected location, making central vision special for seeing. We hypothesize that vision in the peripheral visual field has a weaker or absent top-down feedback and is thus less able to verify visual input properties (e.g., to carry out feature binding), and more prone to certain visual illusions.

Lower animals are likely to have a more restrictive attentional bottleneck, viewing their vision as made of encoding, selecting, and decoding stages is still helpful, even when they, e.g., rodent or fish, do not have a pronounced fovea or even a visual cortex. Their selection is via visual orienting by the head, limbs, whiskers, tentacles, snout, and/or body in addition to, or instead of, eye movements, and should have a stronger impact on overt behavior (e.g., navigation and predation). The brain regions involved can be evolutionary adaptive (e.g., the bottom-up selection likely involves optic tectum rather than V1[67•, 68•]); and seeing may involve a lesser degree of top-down analysis-by-synthesis. Our framework should also be extended to multiple senses. In rodents, for example, the whole of vision may function only like peripheral vision in primates, while the touch sense by the whiskers may function like the central vision in primates, while olfaction could serve both "peripheral" and "central" functions, particularly by the main olfactory and vomeronasal systems, respectively.

Acknowledgement: I am grateful to Peter Dayan and Máté Lengyel for very helpful comments on the manuscript. This work is partly supported by funding from the Max Planck Society and the University of Tübingen

References

[1] Marr D: **Vision, a computational investigation into the human representation and processing of visual information**. *MIT Press* 2010, .

[2] Simons D, Chabris C: **Gorillas in our midst: sustained inattentional blindness for dynamic events**. *Perception* 1999, **28**:1059–1074.

[3] Zhaoping L: **Brains studying brains: look before you think in vision**. *Physical Biology* 2016, **13**:035002.

[4] Kelly DH: **Information capacity of a single retinal channel**. *IRE Transactions on Information Theory* 1962, **8**:221–226.

•• [5] Zhaoping L: **Understanding vision: theory, models, and data**. *Oxford University Press* 2014, .

The original introduction of vision as composed of encoding, selection, and decoding stages

[6] Desimone R, Duncan J: **Neural mechanisms of selective visual attention**. *Annual Review of Neuroscience* 1995, **18**:193–122.

[7] Carrasco M: **Visual attention: The past 25 years**. *Vision research* 2011, **51**:1484–1525.

[8] Sziklai G: **Some studies in the speed of visual perception**. *IRE Transactions on Information Theory* 1956, **2**:125–8.

[9] Strasburger H, Rentschler I, Jüttner M: **Peripheral vision and pattern recognition: A review**. *Journal of vision* 2011, **11**:Article 13.

[10] Whitney D, Levi DM: **Visual crowding: A fundamental limit on conscious perception and object recognition**. *Trends in cognitive sciences* 2011, **15**:160–168.

[11] Hubel DH, Wiesel TN: **Ferrier lecture: Functional architecture of macaque monkey visual cortex**. *Proceedings of the Royal Society of London Series B, Biological Sciences* 1977, **198**:1–59.

[12] Burkhalter A, Van Essen DC: **Processing of color, form and disparity information in visual areas vp and V2 of ventral extrastriate cortex in the macaque monkey**. *The Journal of Neuroscience* 1986, **6**:2327–2351.

[13] Crick F, Koch C: **Are we aware of neural activity in primary visual cortex?** *Nature* 1995, **375**:121–3.

[14] Watanabe M, Cheng K, Murayama Y, Ueno K, Asamizuya T, Tanaka K, Logothetis N: **Attention but not awareness modulates the bold signal in the human V1 during binocular suppression**. *Science* 2011, **334**:829–831.

[15] Zhaoping L: **Attention capture by eye of origin singletons even without awareness—a hallmark of a bottom-up saliency map in the primary visual cortex**. *Journal of Vision* 2008, **8**:article 1.

[16] Zhaoping L: **Gaze capture by eye-of-origin singletons: Interdependence with awareness**. *Journal of Vision* 2012, **12**:article 17.

•• [17] Yan Y, Zhaoping L, Li W: **Bottom-up saliency and top-down learning in the primary visual cortex of monkeys**. *Proceedings of the National Academy of Sciences* 2018, **115**:10499-10504.

Higher V1 responses tend to precede faster saccades in monkeys

[18] Zhaoping L, Guyader N: **Interference with bottom-up feature detection by higher-level object recognition**. *Current Biology* 2007, **17**:26–31.

• [19] Li Z: **Contextual influences in V1 as a basis for pop out and asymmetry in visual search**. *Proceedings of the National Academy of Sciences of the USA* 1999, **96**:10530–10535.

The first journal article to propose a saliency map in V1 for exogenous attentional guidance

•• [20] Li Z: **A saliency map in primary visual cortex**. *Trends in Cognitive Sciences* 2002, **6**:9–16.

An early paper to formulate and propose the V1 saliency hypothesis using a V1 circuit model, before the subsequent experimental tests of the falsifiable predictions

[21] Hubel D, Wiesel T: **Q & A, David Hubel and Torsten Wiesel**. *Neuron* 2012, **75**:182–184.

[22] Zhaoping L: **Theoretical understanding of the early visual processes by data compression and data selection**. *Network: Computation in Neural Systems* 2006, **17**:301–334.

[23] Allman J, Miezin F, McGuinness E: **Stimulus specific responses from beyond the classical receptive field: neurophysiological mechanisms for local-global comparisons in visual neurons**. *Annual Review of Neuroscience* 1985, **8**:407–30.

• [24] White BJ, Kan JY, Levy R, Itti L, Munoz DP: **Superior colliculus encodes visual saliency before the primary visual cortex**. *Proceedings of the National Academy of Sciences* 2017, **114**:9451–9456.

An alternative view on the brain region for the saliency map

[25] Schiller P: **The neural control of visually guided eye movements**. In *Cognitive Neuroscience of Attention, a Developmental Perspective*, Edited by Richards JE, Lawrence Erlbaum Associates, Inc., Mahwah, New Jersey, USA; 1998, 3–50. 1998.

[26] Isa T, Yoshida M: **Saccade control after V1 lesion revisited**. *Current Opinion in Neurobiology* 2009, **19**:608–614.

[27] Zhaoping L, Zhe L: **Primary visual cortex as a saliency map: A parameter-free prediction and its test by behavioral data**. *PLoS Comput Biol* 2015, **11**:e1004375.

[28] Koene A, Zhaoping L: **Feature-specific interactions in salience from combined feature contrasts: Evidence for a bottom-up saliency map in V1**. *Journal of Vision* 2007, **7**:article 6.

[29] Zhaoping L, Xiao Z: **Without informative cues, little can be learned to discriminate eye of origin of visual inputs after multiple weeks of training**. *Journal of Vision* 2016, **16**:440.

• [30] Semedo JD, Zandvakili A, Machens CK, Byron MY, Kohn A: **Cortical areas interact through a communication subspace**. *Neuron* 2019, **102**:249–259.

Suggesting from electrophysiological data that V1 to V2 communications may have a lower dimension than that in V1 activities

[31] Foster K, Gaska JP, Nagler M, Pollen D: **Spatial and temporal frequency selectivity of neurones in visual cortical areas V1 and V2 of the macaque monkey.** *The Journal of physiology* 1985, **365**:331–363.

[32] Ziemba CM, Freeman J, Movshon JA, Simoncelli EP: **Selectivity and tolerance for visual texture in macaque V2**. *Proceedings of the National Academy of Sciences* 2016, **113**:E3140–E3149.

•• [33] Zhaoping L: **Feedback from higher to lower visual areas for visual recognition may be weaker in the periphery: Glimpses from the perception of brief dichoptic stimuli**. *Vision Research* 2017, **136**:32–49.

The central-peripheral dichotomy is proposed from computational considerations, past experimental data, and psychophysical studies specifically designed to probe the dichotomy in the top-down feedback

[34] Zhaoping L: **Olfactory object recognition, segmentation, adaptation, target seeking, and discrimination by the network of the olfactory bulb and cortex: computational model and experimental data**. *Current Opinion in Behavioral Sciences* 2016, **11**:30–39.

•• [35] Zhaoping L, Ackermann J: **Reversed depth in anticorrelated random-dot stereograms and the central-peripheral difference in visual inference**. *Perception* 2018, **47**:531–539.

Confirmation of the prediction that the reversed depth in anticorrelated stereograms can be perceived in peripheral but not central visual field

[36] Li Z, Atick JJ: **Efficient stereo coding in the multiscale representation**. *Network: Computation in Neural Systems* 1994, **5**:157–174.

[37] May KA, Zhaoping L: **Efficient coding theory predicts a tilt aftereffect from viewing untilted patterns**. *Current Biology* 2016, **26**:1571–1576.

[38] May K, Zhaoping L, Hibbard P: **Perceived direction of motion determined by adaptation to static binocular images**. *Current Biology* 2012, **22**:28–32.

• [39] Obara K, OHashi K, Tanifuji M: **Mechanisms for shaping receptive field in monkey area te**. *Journal of Neurophysiology* 2017, **118**:2448–2457.

Showing that IT neurons have shorter response latencies to visual inputs to central than peripheral visual field during visual attention tasks

[40] Griffis JC, Elkhetali AS, Burge WK, Chen RH, Bowman AD, Szaflarski JP, Visscher KM: **Retinotopic patterns of functional connectivity between V1 and large-scale brain networks during resting fixation**. *NeuroImage* 2016, **146**:1071-1083. .

[41] Baizer JS, Ungerleider LG, Desimone R: **Organization of visual inputs to the inferior temporal and posterior parietal cortex in macaques**. *Journal of Neuroscience* 1991, **11**:168–190.

•• [42] Chen M, Yan Y, Gong X, Gilbert CD, Liang H, Li W: **Incremental integration of global contours through interplay between visual cortical areas**. *Neuron* 2014, **82**:682–694.

Showing bottom-up and top-down interactions between V4 to V1 in monkeys to detect contours in cluttered background

• [43] Chen R, Wang F, Liang H, Li W: **Synergistic processing of visual contours across cortical layers in V1 and V2**. *Neuron* 2017, **96**:1388–1402.

Bottom-up and top-down interactions between V2 to V1 in monkeys to detect contours in cluttered background

[44] Hisakata R, Murakami I: **The effects of eccentricity and retinal illuminance on the illusory motion seen in a stationary luminance gradient**. *Vision Research* 2008, **48**:1940–1948.

[45] Schiller PH, Carvey CE: **The Hermann grid illusion revisited**. *Perception* 2005, **34**:1375–1397.

[46] Anstis S: **The furrow illusion: Peripheral motion becomes aligned with stationary contours**. *Journal of vision* 2012, **12**:article 12.

[47] Shapiro A, Lu ZL, Huang CB, Knight E, Ennis R: **Transitions between central and peripheral vision create spatial/temporal distortions: A hypothesis concerning the perceived break of the curveball**. *PLoS One* 2010, **5**:e13296.

[48] Anstis S: **Phi movement as a subtraction process**. *Vision research* 1970, **10**:1411–1430.

[49] Cumming BG, Parker AJ: **Responses of primary visual cortical neurons to binocular disparity without depth perception**. *Nature* 1997, **389**:280–283.

[50] Enns JT: **Object substitution and its relation to other forms of visual masking**. *Vision research* 2004, **44**:1321–1331.

[51] Fahrenfort JJ, Scholte HS, Lamme VA: **Masking disrupts reentrant processing in human visual cortex**. *Journal of cognitive neuroscience* 2007, **19**:1488–1497.

•• [52] Treisman AM, Gelade G: **A feature-integration theory of attention**. *Cognitive Psychology* 1980, **12**:97–136.

Argues that illusory conjunctions of visual features occur outside the spotlight of attention

[53] Bouvier S, Treisman A: **Visual feature binding requires reentry**. *Psychological science* 2010, **21**:200–204.

[54] MacKay D: **Towards an information flow model of human behavior**. *British Journal of Psychology* 1956, **47**:30–43.

[55] Kawato M, Hayakawa H, Inui T: **A forward-inverse optics model of reciprocal connections between visual cortical areas**. *Network: Computation in Neural Systems* 1993, **4**:415–422.

[56] Yuille A, Kersten D: **Vision as Bayesian inference: analysis by synthesis?** *Trends in Cognitive Sciences* 2006, **10**:301–308.

[57] Wolfe JM, Bennett SC: **Preattentive object files: Shapeless bundles of basic features**. *Vision research* 1997, **37**:25–43.

[58] Balas B, Nakano L, Rosenholtz R: **A summary-statistic representation in peripheral vision explains visual crowding**. *Journal of vision* 2009, **9**:Article 13.

[59] Freeman J, Simoncelli EP: **Metamers of the ventral stream**. *Nature neuroscience* 2011, **14**:1195–1201.

[60] Gattass R, Gross C, Sandell J: **Visual topography of V2 in the macaque**. *Journal of Comparative Neurology* 1981, **201**:519–539.

• [61] Manassi M, Whitney D: **Multi-level crowding and the paradox of object recognition in clutter**. *Current Biology* 2018, **28**:R127–R133.

Showing that crowding can occur for high level object features such as faces

[62] Serre T, Oliva A, Poggio T: **A feedforward architecture accounts for rapid categorization**. *Proceedings of the national academy of sciences* 2007, **104**:6424–6429.

• [63] Tang H, Schrimpf M, Lotter W, Moerman C, Paredes A, Caro JO, Hardesty W, Cox D, Kreiman G: **Recurrent computations for visual pattern completion**. *Proceedings of the National Academy of Sciences* 2018, **115**:8835–8840.

Invasive brain recordings from human patients showing delayed ventral cortical responses to partially occluded visual objects

[64] Szegedy C, Zaremba W, Sutskever I, Bruna J, Erhan D, Goodfellow I, Fergus R: **Intriguing properties of neural networks**. *arXiv preprint arXiv:1312.6199*, 2013.

[65] Williams MA, Baker CI, De Beeck HPO, Shim WM, Dang S, Triantafyllou C, Kanwisher N: **Feedback of visual object information to foveal retinotopic cortex**. *Nature neuroscience* 2008, **11**:1439–1445.

[66] Fan X, Wang L, Shao H, Kersten D, He S: **Temporally flexible feedback signal to foveal cortex for peripheral object recognition**. *Proceedings of the National Academy of Sciences* 2016, **113**:11627–11632.

• [67] Ben-Tov M, Donchin O, Ben-Shahar O, Segev R: **Pop-out in visual search of moving targets in the archer fish**. *Nature Communications* 2015, **6**:article number 6476.

Visual search behavior and properties of the optic tectum neurons in archer fish

• [68] Zhaoping L: **From the optic tectum to the primary visual cortex: migration through evolution of the saliency map for exogenous attentional guidance**. *Current opinion in neurobiology* 2016, **40**:94–102.

Identifying visual attention in primates with visual orienting in lower vertebrates, the brain regions and neural processes for the bottom-up saliency map is analyzed comparatively