PHILOSOPHICAL TRANSACTIONS B

royalsocietypublishing.org/journal/rstb

Review

Cite this article: Domini F. 2022 The case against probabilistic inference: a new deterministic theory of 3D visual processing. *Phil. Trans. R. Soc. B* **378**: 20210458. https://doi.org/10.1098/rstb.2021.0458

Received: 14 April 2022
Accepted: 3 October 2022

One contribution of 18 to a discussion meeting issue ‘New approaches to 3D vision’.

Subject Areas:
behaviour, cognition

Keywords:
depth perception, cue integration, 3D vision, computational models

Author for correspondence:
Fulvio Domini
e-mail: fulvio_domini@brown.edu

The case against probabilistic inference: a new deterministic theory of 3D visual processing

Fulvio Domini

CLPS, Brown University, 190 Thayer Street Providence, Rhode Island 02912-9067, USA

FD, 0000-0002-5510-0397

How the brain derives 3D information from inherently ambiguous visual input remains the fundamental question of human vision. The past two decades of research have addressed this question as a problem of probabilistic inference, the dominant model being maximum-likelihood estimation (MLE). This model assumes that independent depth-cue modules derive noisy but statistically accurate estimates of 3D scene parameters that are combined through a weighted average. Cue weights are adjusted based on the system representation of each module’s output variability. Here I demonstrate that the MLE model fails to account for important psychophysical findings and, importantly, misinterprets the just noticeable difference, a hallmark measure of stimulus discriminability, to be an estimate of perceptual uncertainty. I propose a new theory, termed Intrinsic Constraint, which postulates that the visual system does not derive the most probable interpretation of the visual input, but rather, the most stable interpretation amid variations in viewing conditions. This goal is achieved with the Vector Sum model, which represents individual cue estimates as components of a multi-dimensional vector whose norm determines the combined output. This model accounts for the psychophysical findings cited in support of MLE, while predicting existing and new findings that contradict the MLE model.

This article is part of a discussion meeting issue ‘New approaches to 3D vision’.

1. Introduction

The spatial organization of the visual world in three dimensions is a fundamental aspect of our conscious perception. The objects we interact with appear to have a definite and unambiguous 3D structure. This appearance, however, hides the inherent ambiguity of the retinal stimulation, since there are in principle infinite 3D configurations that may have produced the same retinal image and infinite images that may arise from the same 3D configuration. Take for instance the picture on figure 1*a*: it is immediately perceived as a smooth bump ‘bulging out’ of the image plane. What we see, however, is only one of infinite structures that may have produced that projection. Selecting only one of these possible interpretations is therefore a central concern of any theory of 3D vision.

A prominent theory that has dominated the field of vision science for over two decades addresses this issue as a problem of probabilistic inference. This theory postulates that among all possible interpretations of an image the visual system chooses that most probable [1–37]. This theory further assumes that the visual system is endowed with specialized and independent visual modules that choose the most probable depth interpretation from individual *depth cues*, for example, the texture and shading patterns on figure 1*a*. Since these interpretations are only ‘best guesses’ of what is ‘out there’, the outputs

THE ROYAL SOCIETY
PUBLISHING

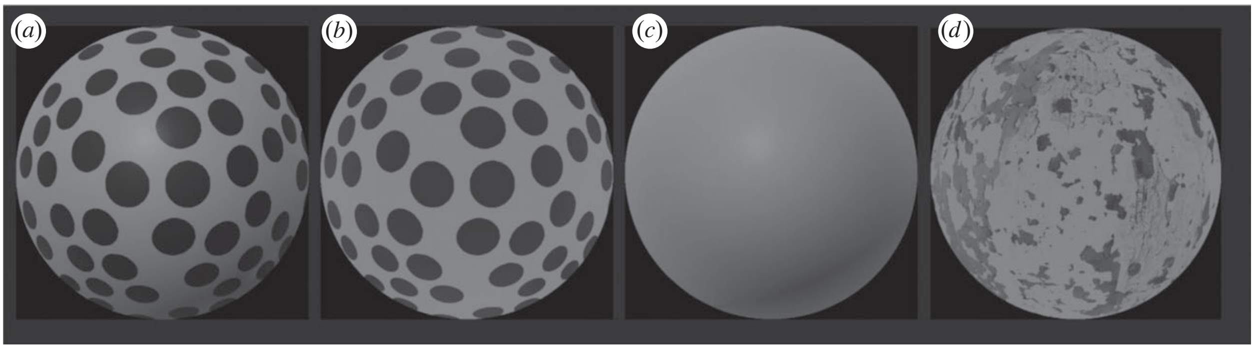

Figure 1. Renderings of the same 3D surface with different 3D information: texture and shading (*a*), only texture (*b*), only shading (*c*) and ineffective texture (*d*).

of the depth modules randomly and independently fluctuate across views of the same distal structure. These fluctuations are often referred to as estimation noise. Nevertheless, an important assumption of the theory is that, despite estimation noise, the outputs from these modules (i.e. their 3D estimates) are *unbiased*. This means that the *average* output of the modules (or, more generally, the expected value) coincides with the ground truth. Probabilistic theories (e.g. Bayesian theories) of *cue integration* specify how unbiased outputs should be combined in a statistically optimal manner, which results in the reduction of the estimation noise affecting each individual output. Specifically, the maximum-likelihood estimation (MLE) model combines the unbiased outputs of individual modules through a *weighted* average, where larger weights are assigned to less noisy outputs. In addition to the assumption of unbiased outputs, this model therefore requires that depth modules provide explicit information about their output variability [15] (e.g. by encoding the noise distribution with neural populations).

These fundamental assumptions, however, are problematic. First, the assumption that outputs are unbiased has been falsified in numerous studies [38–66]. An informal test can be done by directly observing figure 1. Here the four pictures are renderings of the same underlying 3D structure and yet they are all perceived to have a different 3D structure. For example, the images containing only shading or texture information look flatter than the image containing both shading and texture information. Moreover, the surface with polka-dot texture (figure 1*b*) looks deeper than the one with only shading information (figure 1*c*) and also deeper than the surface covered with a less effective texture pattern (figure 1*d*). Second, the assumption that module outputs are noisy also clashes with simple phenomenological observations. Repeated viewing of any of the three surfaces in figure 1 does not seem to produce any detectable changes in their perceived 3D structure. This is also echoed in the fact that the perceived shape of objects in the real world appears stable and unambiguous and does not appear to change over successive viewings. In this paper, I will describe a new theory of cue integration that does not require any of the assumptions of the probabilistic inference model of depth cues and depth-cue integration. First, the new theory assumes that the output of individual depth modules is *linearly* related to distal 3D properties, but that this linear relationship is not necessarily accurate. Second, it assumes that this linear mapping between distal 3D properties and module outputs is *deterministic*, since the 'noise' intrinsic to this mapping is negligible and independent of the particular visual stimulus. The two theories also disagree fundamentally in postulating the source of module output variability. Probabilistic theories assume that visual modules reverse the process of image formation by finding all the 3D structures in the world that may have produced the retinal input and picking among those the most probable. If a cue is unreliable then even small variations of the visual input results in large variations of the output. Specifically, an 'unreliable' cue is one which provides only a weak signal regarding the 3D property, therefore eliciting a 'noisy' or more variable response. The model proposed here, in contrast, assumes linear input–output mappings that are stable, but often inaccurate, where in most cases individual modules result in an underestimate of distal properties. According to the new theory 'noise' or variability is not related to the space of possible 3D interpretations, where 'weak' cues accommodate a larger range of possible 3D interpretations, and are therefore deemed 'noisy'. Instead, a weaker cue does not generate a noisier output, but generates an output with a shallower linear relationship between distal 3D properties and the module output. This is because the image measurement that carries 3D information from that cue delivers a small readout. For instance, 3D information from texture is given by the *texture gradient*, which quantifies the rate of change of the two-dimensional shape of texture elements on the image. In figure 1*b*, the texture gradient is larger than the texture gradient in figure 1*d*, hence the texture cue in figure 1*b* is *stronger* than the texture cue in figure 1*d*. Note, however, that figure 1*d* does not look 'noisier' than figure 1*b*, but it definitely looks shallower. For the new theory the source of a module output variability is the change of 'nuisance' parameters such as the object material properties, the illumination conditions, the object distance from the observer, etc., which can in principle cause sizable changes in the cue strength. In the examples of figure 1, the nuisance variable affecting the texture cue is the material composition of the objects and the nuisance variable affecting the shading cue is the illumination condition. The new theory, name Intrinsic Constraint (IC) theory, postulates that processing of independent depth modules and their combination thereof, is done in a way to minimize the sensitivity to nuisance parameters while maximizing the response to changes of distal 3D properties. The combination of module outputs is achieved by the *Vector Sum* model, which combines the output of individual modules through a vector sum. In contrast with the MLE model, this vector sum does not weigh the individual module outputs and, therefore, does not require any estimate of variability or strength of individual cues. The Vector Sum model predicts systematic biases arising from cue combination since combining cues results in a stronger combined signal, and, therefore a larger magnitude in the estimate. As a consequence, combined-cue stimuli always appear deeper than single-cue stimuli, a phenomenon that can be readily observed on figure 1 by comparing the texture-shading stimulus (figure 1*a*) with the texture-only (figure 1*b*) and shading-only (figure 1*c*) stimuli.

In addition to not requiring an internal estimate of cue variability, the IC theory and the Vector Sum model have also a theoretical advantage over probabilistic theories and the MLE model. Probabilistic theories need to clarify how the visual system has learnt either through evolution or during development to associate the visual input to the accurate geometrical description of the viewed objects. In fact, the brain has no way to access the geometrical structure of the outside world since all information comes through our inherently ambiguous proximal sense data. The IC theory, instead, is based on the conjecture that the visual system 'discovers' (ontogenetically of phylogenetically) distal properties of objects from the *intrinsic* relationships existing among independent image cues. In principle, the correlation among these cues enables the bootstrapping of individual depth modules so as to produce outputs that are *linearly* related to distal properties. This learning process is therefore unsupervised, since it does not require explicit access to the ground truth. Although the existence of this unsupervised learning process is a conjecture that needs to be verified through machine learning algorithms, there is already evidence that unsupervised learning can be very effective in spontaneously clustering images according to distal properties such as reflectance and illumination [67–69]. The Vector Sum model is also more parsimonious than the MLE model since (i) it assumes a linear input–output mapping instead of an accurate or veridical mapping, which can account for the biases observed in depth estimation, and (ii) it does not require an estimate of the output variability of individual modules.

Finally, there is a major difference in methodological assumptions at the heart of the two theories. According to the probabilistic inference theory and MLE model, variability in perceptual estimates in psychophysical tasks as measured by the just noticeable difference (JND) directly reflects the internal 'noise' or reliability of module outputs or cues. Instead, according to the IC theory and Vector Summation model (which assume that outputs of depth modules are not stochastic but deterministic), variability in perceptual estimates is simply a reflection of task-specific demands, such as memory limitations intrinsic to psychophysical tasks. Instead, differences in measured JNDs depend simply on the differences in strength of the linear mapping between distal property and of single or combined module outputs. Shallower mappings result in higher JNDs while steeper linear mappings result in lower JNDs. Thus, JNDs do not appear in the main equation of the Vector Summation model.

In the following sections, I will detail the differences between the IC theory and Probabilistic Inference theory. In §2, I first describe the differences between Vector Sum and the MLE models. In §3, I will describe an entirely new interpretation of JND compatible with the Vector Sum assumption that the output of depth modules is not based on a stochastic process that depends on the stimulus reliability, as assumed by the MLE. I will then describe a novel methodology to test the validity of this new interpretation of JND and empirical evidence that supports it. In §4, I will show new results indicating that the output of single-cue modules is systematically biased and provide further empirical evidence that corroborates the predictions of the Vector Sum model. Finally, in §5, I will describe results of a classic cue integration experiment that quantitatively agree with the predictions of the Vector Sum model and, at the same time, falsify the predictions of the MLE model.

2. Maximum-likelihood estimation versus Vector Sum model of depth-cue integration

Probabilistic models provide mathematical methods for assigning a probability value to each 3D interpretation of an image based on knowledge about the image formation process and prior statistical knowledge of the material composition of objects in the world, their structure and the viewing parameters [15]. As depicted in figure 2*a* (top left), this probabilistic mapping between the distal depth values (horizontal axis) and possible perceptual interpretations (vertical axis) is represented as a probability distribution (solid black). If we disregard the influence of prior knowledge, this probability distribution coincides with the *likelihood* function. MLE models assume that the *most likely* 3D structure determines our perception [15]. In figure 2*a* (top), this perceptual solution is the peak of the probability distribution. Note that this choice is only a 'best guess', since it often does not correspond to the ground truth (Z_A, indicated in green). However, if the same depth-from-texture analysis is repeated over a large number of images of the same 3D surface then the average estimate will correspond to the ground truth. This is a fundamental assumption of MLE models: *3D estimates are unbiased* [15]. The extent to which these solutions fluctuate around the ground truth depends on the *reliability* of the texture pattern, determined by how sharply peaked the probability distribution is. Another image pattern carrying 3D information that can be seen on the rendering of figure 2*a* is the *shading gradient*, determined by the smooth luminance change across the image (figure 2*a*, bottom left). In the example, the shading gradient is characterized by a wider probability distribution. In summary, the module outputs are modelled as unbiased estimates with additive noise reflecting their *reliability* (figure 2*a*, equations (A1)). The MLE model combines the individual outputs in a statistical optimal fashion in order to minimize the combined output noise [15]. In the example of figure 2*a*, combining the texture and shading cues produces a sharper probability distribution whose peak is on average closer to the ground truth. Since the two cues are independent, the combined probability distribution is simply the product of the two individual probability distributions (figure 2*a*, right). This is achieved through a weighted average where the weights are inversely proportional to the variance of the noise of the single-cue estimates (figure 2*b*, equations (1) and (2)). The variance of the combined estimate is smaller and can be predicted from the variance of the single-cue estimates [3] (figure 2*b*, equation (3)).

The Vector Sum model, in contrast, assumes a *linear* mapping between distal 3D properties and the output of individual modules (figure 2*d*, left). This assumption is based on the conjecture that the developing visual system may learn statistical co-occurrences among independent

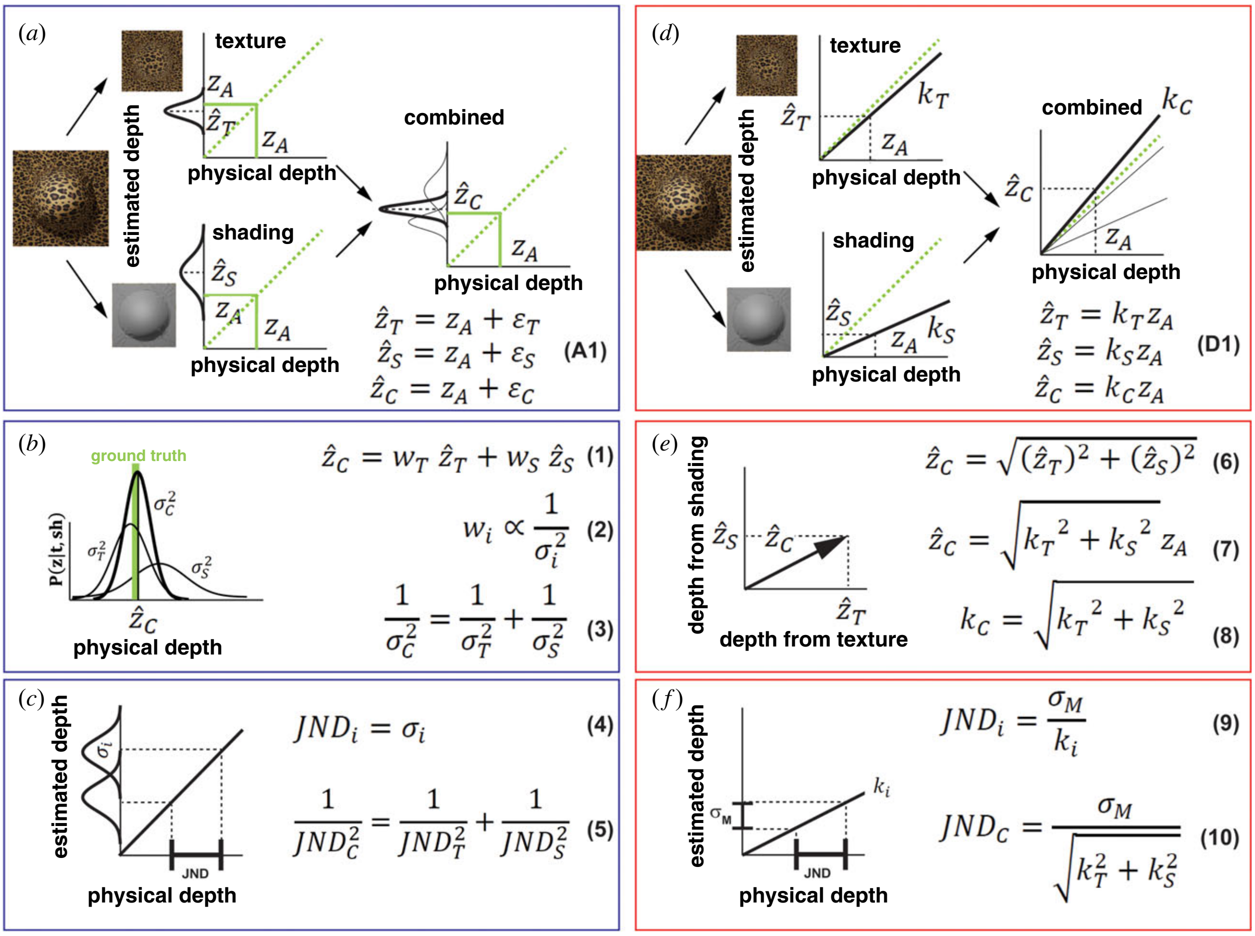

Figure 2. (*a*) The MLE model maps 3D properties (horizontal axes) to estimated properties (vertical axes) with probability distributions (solid black). Individual distributions are associated with independent cues in an image (left), which in this example are texture (top) and shading (bottom). The probability distribution of the combined estimate is sharper than those of individual estimates (right). Although the distributions associated with a particular visual input are peaked at values different from the ground truth z_A, they are on average accurate. (*b*) For linear MLE models, the combined likelihood function is the product of the individual likelihood functions (left) and can be obtained through a weighted average of single-cue estimates, where the weights are inversely proportional to the standard deviation of the noise affecting each cue (equations (1) and (2)). The result is a less noisy combined estimate (equation (3)). (*c*) For MLE models, the JND obtained in discrimination tasks is a proxy measure of the standard deviation of the likelihood function (left, equation (4)). JNDs can therefore be used to predict the s.d. of the combined estimate (equation (5)). (*d*) The Vector Sum model maps distal 3D properties to 3D estimates with a deterministic mapping. Assuming a linear function, the slope of this function is not unitary, as for MLE models, but depends on the strength of individual cues. In this example, texture (top left) is stronger than shading (bottom left). The strength of the combined-cue (right) is always larger than that of the single-cues. (*e*) For the Vector Sum model, individual cue estimates are the components of a multi-dimensional vector (left). The combined estimate is therefore the vector sum of the individual estimates (equations (6) and (7)) and its strength (slope) is also the vector sum of the individual strengths (equation (8)). (*f*) For the Vector Sum model, the noise in a discrimination task is related to memory encoding and retrieval and it is independent of the particular stimulus. The JND measures the change in the distal depth producing a change in perceived depth sufficient to overcome this noise (left). The JND is therefore inversely related to the strength of the stimulus (equation (9)). This explains why the JND for combined stimuli is smaller than the JND of single-cue stimuli, since the strength of combined stimuli is always larger (equation (10)). This prediction is formally identical to that of the linear MLE models (equation (5)).

image patterns and attribute *these co-occurrences* to external causes encoded as latent variables (e.g. a depth map). Since this learning process does not correlate image cues with the ground truth, the linear mapping is in general not accurate (i.e. the slope is not 1) and, instead, depends on the 'quality' of the visual input. Consider for instance in figure 3, the two images of the same 3D bump arising from objects of different material compositions. Whereas the image on the left is immediately perceived as a protruding surface, the image on the right appears almost flat because it displays a much weaker texture gradient. This degraded input will yield a shallow linear function. I term the slope of this function *cue strength*. In the example of figure 3, the *strength* of the texture cue is therefore the slope k_T of the observed mapping between the ground truth z and the texture output $\hat{z}_T$. Therefore, k_T is much smaller for the image on the right of figure 3 than for the image on the left. What is evident from this example is that the magnitude of the output of single-cue modules depends both on distal depth and the cue strength. In turn, cue strengths depend on nuisance scene parameters such as the material composition of objects, ambient illumination, observer's motion, etc. A desirable cue combination rule is therefore one that maximizes the sensitivity to distal 3D properties while at the same time minimizing the influence of nuisance scene parameters. This can be achieved by considering the individual depth estimates as scalar components of a multi-dimensional vector, where the length of the vector corresponds to the combined-cue depth (and the orientation is ignored). Consider figure 4 for the particular case of texture and shading information. For illustration purposes suppose

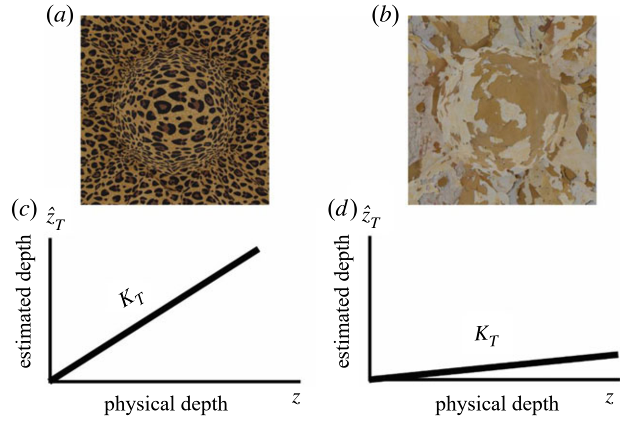

Figure 3. (*a*,*b*) 3D information from individual cues varies across images. (*c*,*d*) Assuming a linear relationship between a 3D estimate ($\hat{z}_T$) and the encoded distal property (z), a degraded input results in a smaller slope (k_T) of this relationship. Termed strength, this slope is larger for the strong texture pattern on the left than for the weak texture pattern on the right.

that in the stimulus on the left, texture and shading have the same strength. This means that the output of the texture module ($\hat{z}_T$) and the output of the shading module ($\hat{z}_S$) are the same (figure 4*a*). In the stimulus in the centre, the strength of texture is much larger than the strength of shading. In the stimulus on the right, the strength of shading is much larger than the strength of texture. It can be shown mathematically that if the axes of this multi-dimensional space of cue estimates are scaled to take into account the natural variability of the scene parameters associated with each individual cue, then the length of the multi-dimensional vector will be maximally sensitive to changes in distal depth values and minimally sensitive to variations of the strength of image cues (see Appendix A for the mathematical proof).

To recap, the Vector Sum model represents single-cue estimates $\hat{z}_i = k_i z$ as the components of a multi-dimensional vector (figure 2*e*). The magnitude of this vector $\hat{z}_C$ is the combined estimate described by the *Vector Sum equation*:

$$\hat{z}_C = \sqrt{(\hat{z}_1)^2 + (\hat{z}_2)^2 + \ldots (\hat{z}_n)^2} = \sqrt{(k_1 z)^2 + (k_2 z)^2 + \ldots (k_n z)^2}.$$

In summary, the MLE model and Vector Sum models describe different mechanisms of 3D processing based on fundamentally different assumptions about how the visual system deals with the ambiguity of the visual input.

The MLE model assumes that the goal of the visual system is to provide a 3D interpretation of the visual input that is as close as possible to the ground truth. Independent modules output the most likely interpretation from single cues. Repeated viewing of a series of equally reliable stimuli will cause the output to 'jump around' the ground truth. Since on average the output coincides with the ground truth (i.e. it is unbiased), the MLE model combines cues in order to minimize the variability of the combined output.

The Vector Sum model strives for the most stable interpretation of the visual input amid variations of viewing conditions. Single cues deliver a 3D output that is linearly related to the distal property. This linear function depends on the *cue strength*. For instance, a cue that is referred to as 'unreliable' or 'noisy' in the MLE model is instead, according to the Vector Summation Model, one that yields a shallow slope of the linear function. In contrast with the MLE model, the input–output mapping is *deterministic*, only affected by negligible noise that is stimulus-independent Therefore, repeated viewing of stimuli with the same strength will yield a nearly constant output (negligible noise). The output only changes when the stimulus itself changes, for example for a given distal 3D structure in response to changes in viewing-dependent nuisance parameters (viewpoint, lighting, material pattern, etc.). The goal of the Vector Sum rule of cue combination is to minimize the influence of the nuisance parameters on the 3D derivation process.

In the following sections, I will describe in detail how these different assumptions about the mechanisms of 3D processing can be tested and provide converging empirical evidence confirming the predictions of the Vector Sum model.

3. What is the source of perceptual variability according to maximum-likelihood estimation and Vector Sum models? Interpretation of just noticeable differences

(a) Maximum-likelihood estimation model and Vector Sum model interpretation of just noticeable difference

The most important empirical tests of the MLE model have used discrimination thresholds to measure the variability of depth judgements. These thresholds, termed JNDs, are usually measured with a 2 Alternative Forced Choice (2AFC) task where a standard stimulus, which is kept fixed throughout the experiment, is being compared to a comparison stimulus that the experimenter varies either through a staircase procedure or through a range of constant stimuli. The discrimination threshold is the minimum difference in depth magnitude between the comparison and standard yielding a reliable discrimination (e.g. 84% correct discriminations). After I summarize the interpretation of JND according to the MLE model, I will describe an alternative interpretation compatible with the assumptions of the Vector Sum model.

The MLE model derives from the input the most likely 3D interpretation that varies across repeated views of similar inputs to an extent that depends on the stimuli reliability. An important prediction of this model is, therefore, that the variability of the MLE interpretation causes the variability of depth judgements. To illustrate, consider the plaster texture in figure 5*a*. The likelihood function associated with this pattern is expected to be wide since there are many possible 3D surfaces that are likely to have produced that pattern [13,23]. Due to this ambiguity of the plaster texture, slightly different texture patterns will yield likelihood functions that have different peaks (figure 5*b*). Therefore, the MLE output of modules will vary considerably across trials and, consequently, also the predicted perceptual judgements (figure 5*c*). If the likelihood distribution is Gaussian, then the distribution of depth judgements will also be Gaussian with the same variance and centred at the true depth value (figure 5*d*). If the same

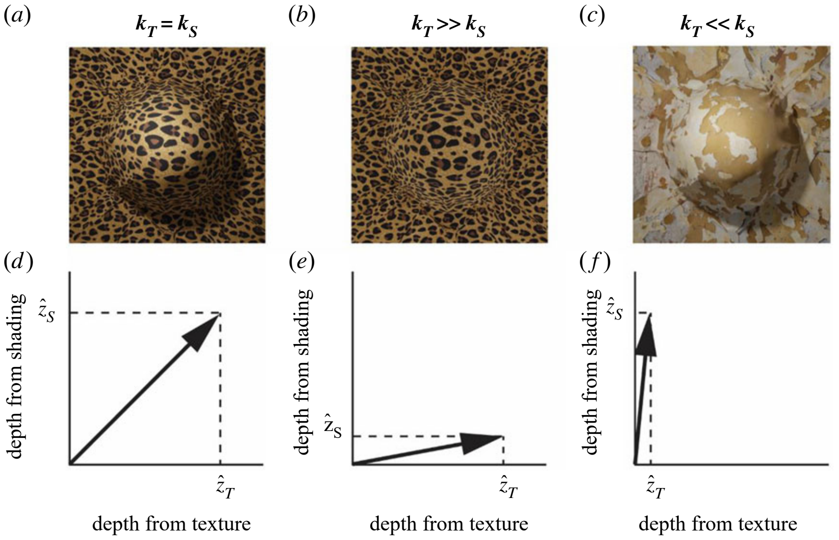

Figure 4. (*a–c*) The same 3D structure as in figure 3 rendered with different combinations of texture and shading information. (*d–f*) The vector sum of individual cue estimates changes among the three viewing conditions, but this change is smaller than the relative change of each single-cue estimate.

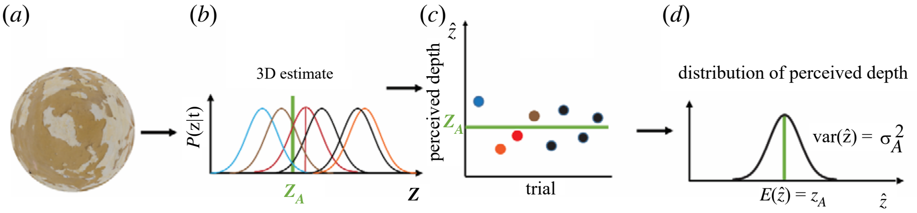

Figure 5. When the same unreliable texture stimulus (*a*) is viewed on different trials, it will determine likelihood functions peaked at different depth values. In this example, these functions, depicted in different colours, are assumed to be Gaussian (*b*). If perceived depth is predicted to be a MLE then it will vary from trial to trial (*c*). The distribution of depth estimates is in this case also Gaussian, centred at the veridical depth value z_A and with variance σ_A^2equal to the variance of the likelihood functions (*d*).

experiment is repeated with the cheetah-texture pattern (figure 3*a*), then the MLE model predicts a much smaller variation in depth judgements, since the cheetah texture is more reliable and therefore characterized by a much sharper likelihood. Since the distribution of depth judgements is assumed to be unbiased, in a discrimination task, the JND will coincide with the discriminable separation between the means of the two distributions being compared (figure 2*c* left, equation (4)).

By contrast, for the Vector Sum model, the input–output mapping of visual modules is *deterministic*. Consequently, an observer is expected to perceive the same depth magnitude from a sequence of equivalent stimuli (e.g. repeated viewing of the plaster stimulus). Most importantly, whatever negligible amount of neural noise may affect the output, it is independent of the 3D information contained in the input. In other words, a sequence of cheetah-texture stimuli will produce the same output variability as a sequence of plaster-texture stimuli. This basic assumption of the Vector Sum model requires a radical re-interpretation of JNDs from that of the MLE model. The Vector Sum model assumes that the source of noise that determines the JND is entirely task-related. For instance, in a 2AFC task, the observer must store in memory the perceived depth magnitude of the stimulus in the first interval and then retrieve this information to compare it with the perceived depth magnitude of the second stimulus. We can assume that this process of memory storage and retrieval introduces noise to the *perceived* depth estimate of the first stimulus, a sort of memory ‘smearing’ with s.d. σ_M. Therefore, the difference of perceived depth between the two stimuli $(\Delta\hat{z})$ must be large enough to overcome this task-related memory noise $(\Delta\hat{z} = \sigma_M)$. As just mentioned above, what is fundamental here is that according to the Vector Sum model this difference is *independent* of the type of stimuli being discriminated. For instance, a cheetah texture and a plaster texture must bring about the same perceived depth difference $\Delta\hat{z}$ for a reliable discrimination. However, since the JND is defined in terms of *distal* depth values, the change in simulated depth of the cheetah-texture stimulus ($\mathrm{JND}_{\mathrm{cheetah}}$) producing a perceived depth difference $\Delta\hat{z} = \sigma_M$ depends on its strength (k_{cheetah}): $\Delta\hat{z} = k_{\mathrm{cheetah}}\mathrm{JND}_{\mathrm{cheetah}}$. In the same way, $\Delta\hat{z} = k_{\mathrm{plaster}}\mathrm{JND}_{\mathrm{plaster}}$. Therefore, the two stimuli yield the same perceived depth difference $\Delta\hat{z}$ if $\mathrm{JND}_{\mathrm{cheetah}} < \mathrm{JND}_{\mathrm{plaster}}$ since $k_{\mathrm{cheetah}} > k_{\mathrm{plaster}}$. An important consequence of this interpretation of JND is that it is inversely proportional to the strength of a stimulus (figure 2*f*, equation (9)).

In summary, the interpretations of JND according to the MLE and Vector Sum models are profoundly different since they assume different sources of noise affecting the JND. Nevertheless, in specific experimental conditions they make the same predictions. For instance, both models predict that discriminating cheetah-texture stimuli yields a smaller JND than discriminating plaster-texture stimuli. This is because, on the one hand, the MLE model assumes

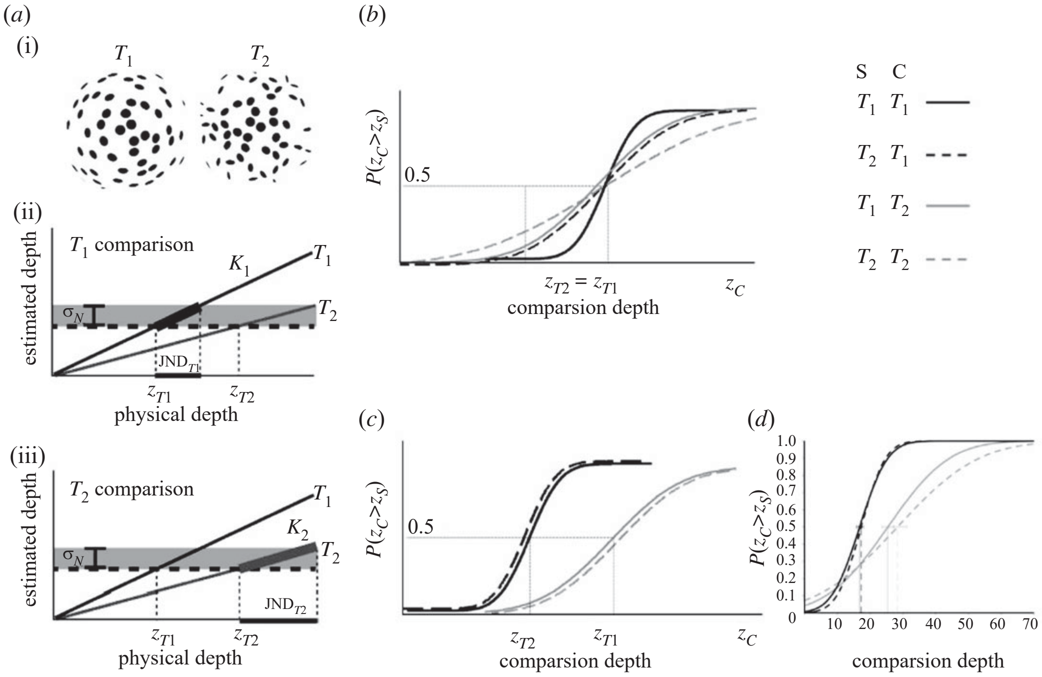

Figure 6. (*a*) Two texture stimuli of different strengths (T_1 and T_2, (i)). For the Vector Sum model, the JND measured in a discrimination task only depends on the strength of the comparison stimulus. If the comparison is the strong stimulus (T_1, (ii)) then the JND, equal to the distal depth difference necessary to overcome the task-related noise ($\boldsymbol{\sigma_N}$), will be smaller than the JND obtained when the comparison stimulus is the weak stimulus (T_2, (iii)). (*b*) Psychometric functions predicted by the Bayesian model (the curves are for illustrative purposes only and not based on actual data). In the legend S indicates the standard stimulus and C the comparison stimulus. The slope of the psychometric functions depends on both the comparison and standard stimuli. (*c*) Psychometric functions predicted by the Vector Sum model. The slope of the psychometric functions only depends on the comparison stimulus. (*d*) Empirical data [70].

that cheetah-texture stimuli produce a less variable output than plaster-texture stimuli whereas, on the other hand, the Vector Sum model assumes that the strength of cheetah-texture stimuli is larger than the strength of plaster-texture stimuli. In the next sections, I will describe a novel methodology that allows us to discriminate between the two models.

(b) Just noticeable difference according to the maximum-likelihood estimation and Vector Sum models: empirical evidence with single cues

Fundamental to the following discussion is how the two models interpret the role played by the standard stimulus, which is kept fixed during a discrimination task, and the comparison stimulus, which is varied through a staircase procedure or the constant stimuli method. I will now show that the two models predict the same JNDs only in experimental conditions where the standard and comparison stimuli contain the same cue or set of cues (e.g. discriminating two cheetah-texture stimuli). When 3D information composing the standard and comparison differs (e.g. discriminating a cheetah-texture stimulus from a plaster-texture stimulus) the predictions of the two models diverge. Figure 6*a*(i) shows a stimulus T_1 that is more informative than a stimulus T_2. In what follows I will describe the predictions of the two models for a discrimination task in all possible conditions where each of the stimuli T_1 and T_2 are either assigned to be the fixed standard or the variable comparison. In particular, I will formulate the specific predictions for the point of subjective equality (PSE) and *slope* of the *psychometric* function, which relates the depth of the comparison stimulus to the proportion of responses where the comparison is perceived to be deeper than the standard.

Figure 6*b* shows the predictions of the MLE model for all possible combinations of standard and comparison stimuli. The first prediction is that all the psychometric curves have identical PSEs since the MLE model predicts that the perceived depth of the two texture stimuli is the same when the distal depth is the same (figure 6*b*, PSEs: $z_{T1} = z_{T2}$). Second, the slope of the psychometric function should depend on the reliability of *both* the fixed standard and the varying comparison. This is because the depth estimates of both the standard and comparison stimuli are subject to estimation noise. As discussed in §3(a), these are described by probability distributions centred at the simulated depth values with s.d. reflecting the stimuli reliabilities. In the example of figure 6, the texture stimulus T_1 is more reliable (i.e. less noisy) than the texture stimulus T_2. Judging above chance which stimulus is deeper requires a depth difference between the simulated depth of the two stimuli that separates the means of these probability distributions by an amount overcoming the pooled variance (i.e. the sum of squares of the two variances). For instance, if a reliable comparison is being discriminated from a reliable standard then this difference will be smaller than when it is compared to an unreliable standard. Therefore, in the first case, the slope of the resulting psychometric function will be larger than in the second case. In summary, according to the MLE model, we should expect (i) the steepest slope when standard and comparison have both high reliability (figure 6*d*, solid black); (ii) intermediate slopes when standard and comparison have different reliabilities (figure 6*d*, dashed black and solid grey); and (iii) the shallowest slope when both standard and comparison are unreliable (figure 6*d*, dashed grey).

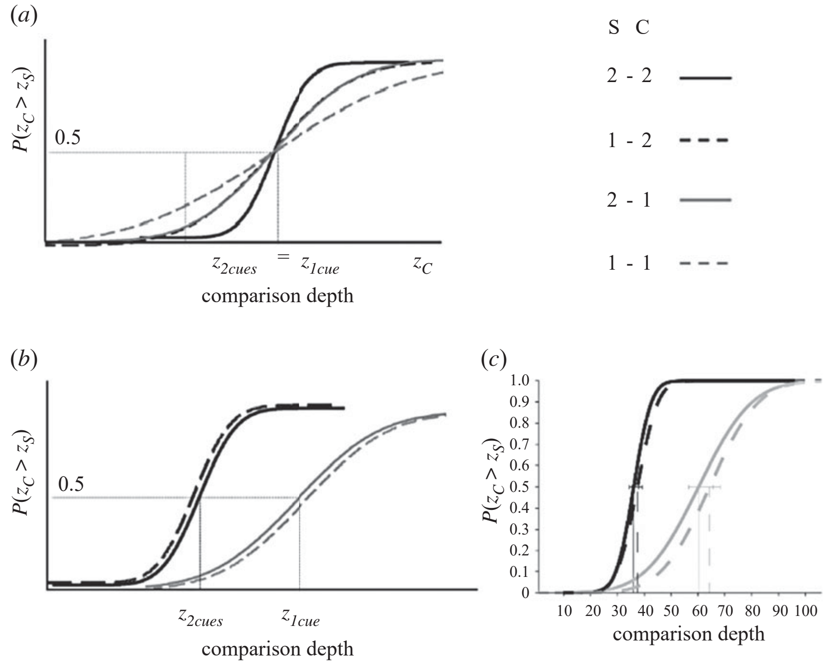

Figure 7. (*a*) MLE model predictions of the psychometric functions for a depth-discrimination task. In the legend, S indicates the standard stimulus and C the comparison stimulus. '2' refers to the combined-cue stimuli and '1' to the single-cue stimuli. The slope of the psychometric functions depends on both the standard and comparison stimuli. (*b*) Vector Sum model predictions. The slope of the psychometric functions depends only on the comparison stimuli. (*c*) Empirical data [70].

For the Vector Sum model, the strength of the texture stimulus T_1 is larger than the strength of the texture stimulus T_2. According to the Vector Sum model (i) 3D estimates from both standard and comparison stimuli are deterministic, that is, they do not vary across trials, and (ii) depth discrimination is achieved when the perceived depth of the comparison stimulus differs from the perceived depth of the standard stimulus by an amount that overcomes the task-related noise σ_M. The Vector Sum model therefore predicts that the comparison stimulus is what determines the JND. Consider figure 6*a*(ii) where T_1 is the comparison stimulus and T_2 is the fixed standard. First, note that a perceptual match (horizontal dashed line) is achieved when the simulated depth values of the two stimuli are different (z_{T1} and z_{T2} in the figure), because they have different cue strengths. Second, for a measurable discrimination, the simulated depth of T_1 (comparison) must increase from the pedestal value z_{T1} by an amount yielding a perceived depth difference equal to the standard deviation of the task-related noise σ_M. This amount is the JND (JND_{T_1}, figure 6*a*(ii), black horizontal bar on the *x*-axis), which we saw is inversely proportional to the cue strength (figure 1*f*, equation (9)). If the comparison is the weaker stimulus (i.e. texture T_2) then the *JND* will be larger (JND_{T_2} on figure 6*a*(iii)). In figure 4*c*, I show the predicted psychometric functions based on the IC model hypothesis that the JND only depends on the cue strength k of the comparison stimulus. When the comparison is the strong-cue stimulus, we expect *steep* and identical psychometric functions when both (a) the standard is a strong-cue stimulus (solid black) and (b) the standard is a weak-cue stimulus (dashed black). Similarly, when the comparison is the weak-cue stimulus, we expect shallow and identical psychometric functions for both a strong (solid grey) and a weak (dashed grey) standard. The results of a recent study shown in figure 6*d* confirm the predictions of the Vector Sum model and clearly disagree with those of the MLE model [70].

(c) Just noticeable difference according to the maximum-likelihood estimation and Vector Sum models: empirical evidence with multiple cues

The MLE model and Vector Sum model predict different outcomes of cue integration with regard to the variance of the combined output. The MLE model predicts that combining cues yields a reduction of the variance of the combined output (figure 2*b*, equation (3)) [3]. This prediction has been tested extensively in depth-discrimination experiments where standards and comparisons are the same stimulus type: either single-cue stimuli or combined-cue stimuli.

Using a similar experimental paradigm to the one illustrated above, I will extend the predictions of the MLE model to a discrimination task where single-cue or combined-cue stimuli are either assigned to be the fixed standard or the variable comparison. Figure 7*a* shows the psychometric curves predicted by the MLE model. First, since the MLE model assumes accurate outputs in all conditions, the PSE of the psychometric curves is predicted to be the same for all combinations of standard and comparison stimuli (figure 7*a*, PSEs: $z_{2\text{cues}} = z_{1\text{cue}}$). Second, as for the case of single-cue stimuli described in §3(b), the slope of the psychometric function depends on both the reliability of standard and comparison. We therefore expect: (i) the steepest slope when standard and comparison are both (the more reliable) double-cue stimuli (figure 7*a*, solid black); (ii) intermediate slopes when standard and comparison are different (figure 7*a*, dashed black and solid grey); and (iii) the shallowest slope when both standard and comparison are both (the less reliable) single-cue stimuli (figure 7*a*, dashed grey).

For the Vector Sum model, combining cues increases the depth-cue strength (figure 2*e*, equation (8)). Since the JND only depends on the cue strength, we expect steeper psychometric curves for combined-cue comparisons (figure 7*b*, black curves) and shallower psychometric curves for single-cue comparisons (figure 7*b*, grey curves). In both cases, the slope does not depend on the standard stimulus. We also expect that when the cue strengths of comparison and standard stimuli are different, the PSEs are also different, since depth estimates of combined-cue stimuli are predicted to be *larger* than those of single-cue stimuli because the strength of combined cues is larger (figure 2*e*, equations (6) and (7)). Therefore, the simulated depth $z_{2\text{cues}}$ of combined stimuli must be smaller than the simulated depth $z_{1\text{cue}}$ of single-cue stimuli to obtain a perceptual match (figure 7*b*). In a recent experiment, I tested these predictions with disparity and texture cues. The single-cue stimuli were random-dot stereograms simulating 3D bumps similar to those shown on figure 1. The combined-cue stimuli provided both texture and disparity information. In figure 7*c*, I show the results of a representative observer [70]. The results accurately match the predictions of the Vector Sum model and contradict the predictions of the MLE model.

These findings impose a radical re-interpretation of previous results on depth-discrimination experiments used as the strongest evidence in support of the MLE model. As mentioned before, previous results seem to indicate that the variance of combined-cue estimates is smaller than the variance of single-cue estimates as predicted by MLE (figure 1*c*, equations (4) and (5)) [15,30]. However, in these experiments, depth discrimination was studied separately for single-cue and combined-cue stimuli. Note, however, that in these experimental conditions the Vector Sum model makes identical predictions for the slope of the psychometric function (figure 7*a*,*b*) where the smaller JND observed for combined-cue discriminations is due to the larger combined-cue strength, which amplifies the response to distal depth changes (figure 2*e*, equations (7) and (8)). Most importantly, when strengths are expressed in terms of JNDs (figure 2*f*, equations (9) and (10)) the relationship between the predicted JND of the combined-cue stimulus and the JNDs of the single-cue stimuli is identical to the prediction of the MLE model (figure 2*c*, equation (5))[1].

In summary, the Vector Sum model can predict previous results from depth-discrimination experiments used to confirm the predictions of the MLE model. However, the Vector Sum model can also predict the outcomes of untested experimental conditions where the cues displayed in the standard and comparison stimuli are different. In these newly tested conditions the predictions of the MLE model are falsified. From these results we can conclude that variability of depth judgements measured in depth-discrimination tasks does not reflect the output variability of 3D processing modules, as predicted by the MLE model. Instead, as predicted by the Vector Sum model, the JND is influenced by task-related noise and the cue strength.

4. Are 3D estimates from single-cues accurate? Evidence from a depth-matching experiment

The MLE model is based on the fundamental assumption that the output of depth modules is on average accurate (i.e. it is *unbiased*). The Vector Sum model, instead, assumes that individual outputs are only linearly related to distal 3D properties, but, in general, do not correspond the ground truth (i.e. they are *biased*). It is only the 'lucky' circumstance of encountering the ideal stimulus that yields a cue strength $k_i = 1$ that will *coincidentally* produce a veridical output. Typically, cue strengths elicit under-estimation of 3D properties, but can in some specific viewing conditions also produce over-estimations (figure 2*d*, left panels, equations (D1)).

If the assumption of the MLE model is true then the perceived depth from one cue should match the perceived depth from a different cue as long as both cues specify the same 3D structure. Specifically, consider a depth-matching task where a fixed single-cue (e.g. motion) standard stimulus specifying a depth magnitude z_1 is compared to a comparison stimulus specified by a different depth cue (e.g. disparity). The simulated depth of the comparison stimulus (z_2) is varied until a depth match is obtained. If the single-cue outputs are unbiased, as assumed by the MLE model, then a perceptual match is obtained when $z_1 = z_2$. By contrast, the Vector Sum model predicts that the distal depth values resulting in a perceptual match are not in general the same, but depend on the cue strength of each cue. Since perceived depth $\hat{z}_1 = k_1 z_1$ and $\hat{z}_2 = k_2 z_2$ (figure 2*d*, equation D1) a perceptual match is obtained when $k_1 z_1 = k_2 z_2$, where k_1 and k_2 are the cue strengths. Therefore, since $z_2 = (k_1/k_2) z_1$, the depth match can be predicted from the cue strength ratio (k_1/k_2). Unfortunately, this ratio cannot be determined since the values of the cue strengths are unknown. Remember, however, that according to the Vector Sum model JNDs are inversely proportional to the cue strengths (figure 2f, equation (9)). Therefore, the ratio k_1/k_2 can be determined from the two discrimination thresholds JND_1 and JND_2 obtained in independent discrimination tasks performed on each cue separately. As a consequence, the ratio between the cue strengths is inversely proportional to the ratio between the JNDs: $k_1\ /k_2 = \text{JND}_2/\text{JND}_1$. Given this additional quantity, the prediction of the Vector Sum model is $z_2 = (\text{JND}_2/\text{JND}_1) z_1$. I tested this prediction for motion and disparity cues in two separate studies [71,72] and indeed found that the perceptual match was not veridical as predicted by the MLE model. As can be seen in figure 8 (top), simulated depth from disparity (z_2) producing a perceptual depth match is much smaller than the simulated depth from motion (z_1). However, as can be seen in figure 8 (bottom), the simulated depth from motion scaled by the JND ratio (such that $z_2 = (\text{JND}_2/\text{JND}_1) z_1$) accurately predicts the empirical data and therefore confirms the predictions of the Vector Sum model.

On a final note, these results not only reject the assumption of the MLE model that 3D perception is accurate, but also provide further converging evidence that JNDs do not measure the s.d. of single-cue outputs. In fact, according to the MLE theory, the PSEs estimated in a matching task and the JNDs resulting from discrimination tasks should be completely *unrelated parameters*. This is because the PSEs are estimates of the means of the likelihood functions associated with each cue (which should correspond to the ground truth) and the JNDs are estimates of the s.d. of these likelihood functions. By contrast, for the Vector Sum model, both PSEs and JNDs are univocally determined by the cue strength. The results shown in figure 8 clearly confirm this second interpretation of JNDs.

5. Predicting cue combination

The MLE and Vector Sum models make different predictions about the outcome of *cue integration* experiments. In reference

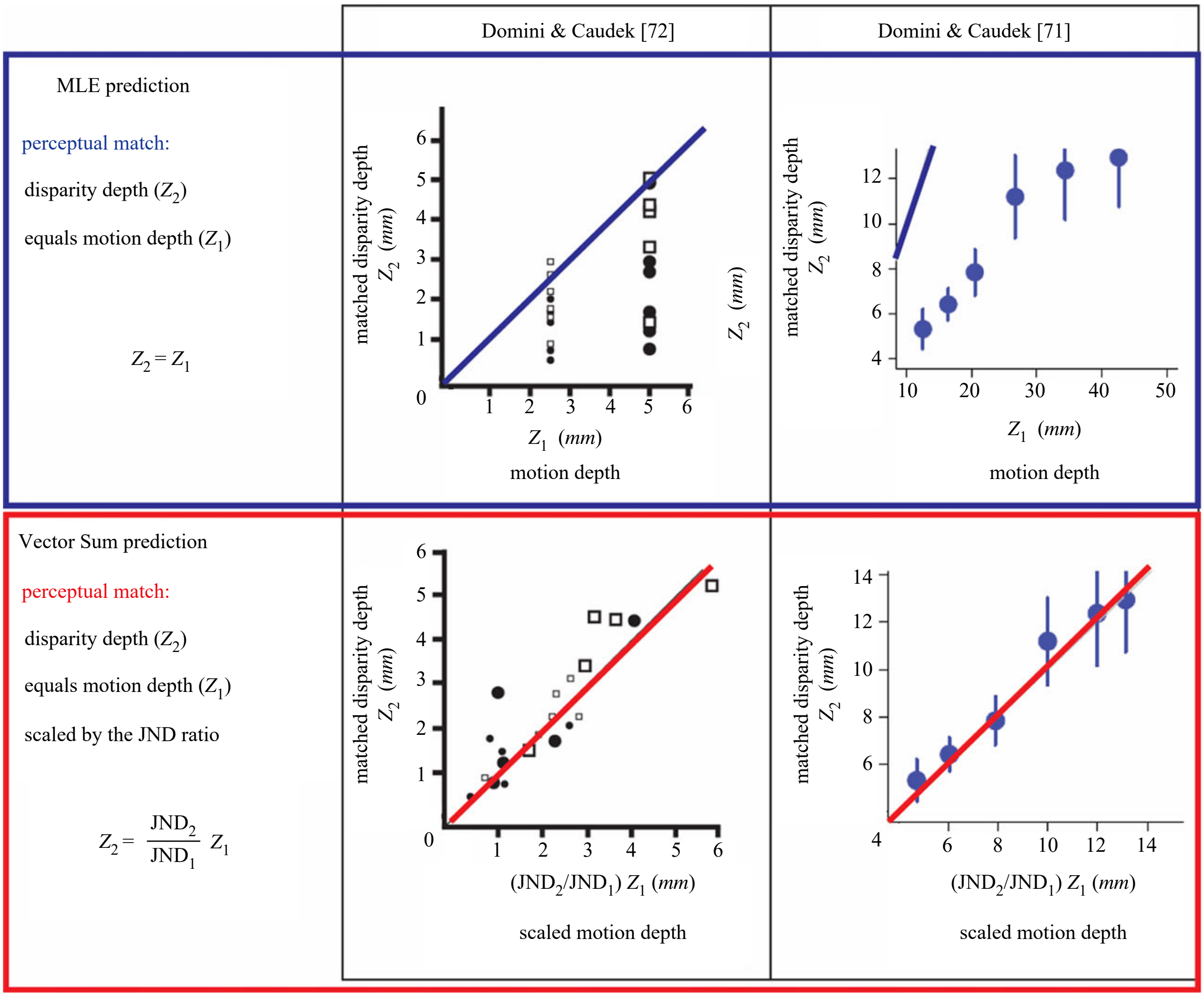

Figure 8. Results of two studies by Domini & Caudek ([71], right column, and [72], left column) where subjects matched a motion stimulus of depth z_1 with a disparity stimulus. z_2 is the disparity depth that yields a perceptual match. (Top) The MLE model predicts accuracy: a perceptual match is obtained when the disparity depth (z_2) equals the motion depth (z_1). This prediction, represented by a blue line, is shown on the two graphs where matched disparity depth is plotted as function of motion depth. The symbols on the left plot indicate individual observers. The symbols on the right plot represent data averaged across observers. The data show that responses are highly inaccurate and, therefore, in clear disagreement with the MLE predictions. (Bottom) The Vector Sum model predicts a perceptual match when the disparity depth (z_2) is equal to the motion depth (z_1) scaled by the JND ratio ($\text{JND}_2/\text{JND}_1$). This prediction, represented by the red lines, is shown on the two graphs where the same results shown on the top graphs have been replotted so that the matched disparity depth is shown as function of the scaled motion depth ($\text{JND}_2/\text{JND}_1)z_1$. The data clearly agree with the predictions of the Vector Sum model.

to figure 1, consider an experiment exploring the combination of texture and shading information. In the single-cue condition, only one cue carries information about the 3D structure of distal objects. The bump in figure 1*b* viewed in a uniformly illuminated environment contains only a texture gradient. Conversely, if the material composition of the bump produces a negligible texture gradient but it is illuminated by a distant light source it carries only shading information (figure 1*c*). In the combined-cue conditions, both texture and shading are present (figure 1*a*).

The MLE model combines in a statistical optimal fashion the single-cue outputs through a weighted average, where the weights are inversely proportional to variance of the output noise (figure 2*b*, equations (1) and (2)). The perceived depth magnitude of a cue-combined stimulus is therefore predicted to be 'in between' the perceived depth magnitudes of the single-cue stimuli and closer to that of the more reliable cue. In the example of figure 1, the perceived depth of the texture-shading stimulus should be intermediate to the perceived depth of the shading-only and texture-only stimuli.

The Vector Sum model treats the single-cue outputs as orthogonal components of a multi-dimensional vector. The length of this vector is the combined-cue estimate (figure 2*e*, equations (6)–(8)). Therefore, the perceived depth magnitude of the combined-cue stimulus is always predicted to be *larger* than the perceived depth of single-cue stimuli. This phenomenon can be directly observed in figure 1 where the texture-shading stimulus appears deeper than the texture-only and shading-only stimuli.

In a recent experiment where we studied the combination of texture and binocular-disparity information, we confirmed the predictions of the Vector Sum model [73]. Observers judged the depth of a sinusoidal corrugation by adjusting the amplitude of a two-dimensional sinusoidal probe in three conditions (figure 9*a*). In the *Texture* condition, only the texture cue was present. In the *Disparity* condition, we showed subjects a random-dot stereogram, so that only binocular disparities were present. In the *Combined* condition, both cues were present.

Figure 9 shows the results of the experiment repeated at two viewing distances (40 cm and 80 cm) [73]. Note first

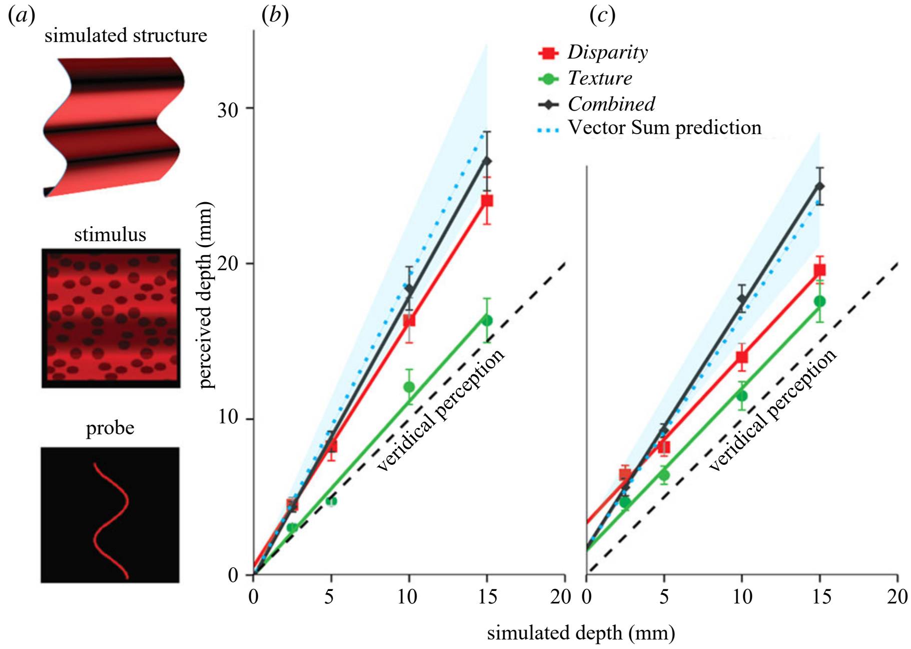

Figure 9. (*a*) The stimuli used in a study of cue integration simulated the projection of a sinusoidal corrugation in depth providing texture and disparity information [73]. The observers adjusted a two-dimensional sinusoidal probe until its amplitude matched the perceived amplitude of the sinusoidal depth corrugation. (*b*,*c*) Perceived depth as a function of simulated depth in the three experimental conditions: disparity (red), texture (green) and combined (black). The stimuli were seen at two distances: 40 cm (*b*) and 80 cm (*c*). The light-blue areas represent the 95% confidence intervals of the Vector Sum model predictions. The MLE model predicts that the combined-cue judgements should fall in between the single-cue judgements [73].

that at 40 cm the single-cue estimates differ from each other. Perceived depth from disparity (red) is larger than perceived depth from texture (green) and is significantly overestimated. However, depth in the combined condition (black) is even more overestimated. At 80 cm, perceived depth from disparity appears to be close to veridicality and close to perceived depth from texture, as seen by their matching slopes. But when the cues are combined, perceived depth is again larger and overestimated. These results are predicted by the Vector Sum model: the shaded light-blue areas on figure 9 indicate the confidence intervals of the predictions for the combined-cue results obtained directly from the single-cue data. The results clearly match this prediction and deviate systematically from a weighted average prediction of the MLE model (which should fall in-between the single-cue estimates) [73].

6. Concluding remarks

In this paper, I proposed a new computational theory of cue integration, termed IC, as an alternative to the current theories based on probabilistic inference. The IC theory is implemented with the Vector Sum model, a *deterministic* model that represents the outputs of independent processing modules as the components of a multi-dimensional vector. The norm of this vector is the estimate of a 3D property (e.g. depth, slant or curvature). This model maximizes the sensitivity to changes in distal properties while minimizing the undesired influence of nuisance scene parameters, such as the motion of the observer, illumination conditions, material composition of objects, etc. There are two main reasons why I argue that the Vector Sum model of depth-cue integration is a more viable alternative to the MLE models. First, the Vector Sum model is ***more parsimonious***, since it does not require two strong foundational assumptions of the MLE approach. The first assumption is that independent modules provide *accurate* estimates of 3D properties of distal objects. Instead, the Vector Sum model only postulates a *linear* mapping between distal properties and each module output. The second assumption of the MLE model is that the outputs of the single-cue modules are combined through a weighted average that favours the most reliable cues. This requires that each module also provides information about cue reliability. In order to hypothesize possible neural mechanisms that implement this combination process, it has been suggested that cortical neurons directly represent probability distributions and then combine those distributions in a statistical optimal fashion [19]. By contrast, the rule of combination of the Vector Sum model does not necessitate any estimate of the 'quality' of the output of single-cue modules.

Second, the IC model has ***more explanatory power***, since it predicts previous results in support of the MLE model and, most importantly, accounts for novel results that cannot be predicted by these models without the introduction of ad-hoc assumptions and priors [71–84]. From an empirical standpoint, the MLE model makes accurate predictions only in restricted experimental conditions, but fails to account for systematic biases in 3D judgements and does not explain simple phenomenological observations. The experimental tests that do agree with the MLE model hinge on the measurement and interpretation of discrimination thresholds, commonly termed JND. The methodological assumption of the MLE model is that JNDs are measures of the s.d. of the noise of 3D estimates. We showed that this methodological assumption, however, is incorrect, and provided new empirical evidence consistent with a radically different interpretation of JND. Through this new interpretation, the Vector Sum model can both predict previous data in support of the MLE model and new results incompatible with the MLE predictions.

It is important to highlight that both the IC theory and the probabilistic inference theories only specify how 3D estimates from separate modules are combined, but do not specify *how* the depth modules are implemented. This is certainly work for future research, but the way this problem should be approached must be guided by the assumptions the two theories make about the nature of the computation of depth modules and the output information. Take, for instance, the problem of deriving 3D shape from texture information. Probabilistic theories, and in particular Bayesian models, postulate that the visual system reverses the process of retinal projection to find among infinite structures in the world the most probable. This requires that the texture module embeds information about the statistical distribution of textures in the environment and also *a priori* probabilities of 3D structures in the world. If this information is correct then the module can compute the truthful probability distribution of 3D shapes that have generated a given visual input. In reference again to figure 1, this means that this output would generate a sharp probability distribution for the texture in figure 1*b* and a wide distribution for the texture in figure 1*d*. Moreover, the peaks of these distributions are expected to vary across views, with small variations for the reliable texture (figure 1*b*) and large variations for the unreliable texture (figure 1*d*). However, both outputs are expected to be on average veridical. Although attempts have been made in the past to develop probabilistic models of shape from texture, these have been confined to the very special case of planar surfaces [4,23]. In general, there are no existing working models of shape from texture that predict human behaviour. In contrast with probabilistic theories, the IC theory postulates the existence of depth modules that output a 3D estimate that is only linearly related to the ground truth, but it is not in general accurate. Unlike the probabilistic models, this estimate is stable across multiple views of the same type of stimulus. What it is expected, instead, is that the magnitude of this response depends on the strength of the cue. In the example of figure 1, the texture module will output a large response for the texture in figure 1*b* and a weak response for the texture in figure 1*d*. In contrast with the probabilistic models, the output estimate of depth modules does not carry any information about the 'quality' of the visual input, such that the response to a weak cue (the texture cue in figure 1*d*) and the response to a strong cue projected from a shallow surface are indistinguishable. Therefore, according to the IC theory, single-cue modules should be attuned to image properties that are *invariant* with the nuisance scene parameters (e.g. the material property that determines the surface texture). A consequence of this assumption is that, in stark contrast with the MLE model, the visual system does not 'weigh' the outputs of individual modules based on some measure of 'reliability' or strength of cues, since such information is absent from the module output. In other words, the visual system is *blind* to the distal causes of the visual input and simply combines all the module outputs through a vector sum. It is the logic of the vector sum regime that guarantees that the combined output will always optimally track the underlying distal depth variable, regardless of the strength of the individual cues.

In a final note, it should be emphasized that although the two theories do not specify how depth modules may be learnt, the depth modules postulated by the IC theory do not necessarily require information about the ground truth for their mechanisms to be learnt. Instead, *invariant* properties of independent image signals that covary with distal properties must also necessarily covary among themselves and, likewise, covary with signals arising from other modalities such as touch or proprioception. Discovering these covariations may be a plausible learning process for a biological system that never has direct access to the distal geometric structure of the environment [85,86].

Data accessibility. This article has no additional data.

Authors' contributions. F.D.: writing—original draft and writing—review and editing.

Conflict of interest declaration. I declare I have no competing interests.

Funding. This research was funded by NSF grant no. BCS-2120610.

Endnote

[1]This can be verified by combining equations (9) and (10) of figure 2*f*.

Appendix A

Appendix 1: *The Vector Sum model maximizes the Signal-to-Noise-Ratio*. For simplicity consider only two signals $s_1 = \lambda_1 z$ and $s_2 = \lambda_2 z$, where λ_i are unknown multipliers depending on confounding variables and z is the magnitude of the 3D property. These signals are the visual system encoding of 3D information from single cues. We seek an estimate $\hat{z}_C = f(s_1,s_2)$ (1) proportional to z and (2) most sensitive to 3D information and least sensitive to random fluctuations ε_i of λ_i. If λ_{i0} is the unperturbed value of λ_i then $\lambda_i = \lambda_{i0} + \ \varepsilon_i$ and $s_{i0} = \lambda_{i0} z$. We assume small random perturbations due to changes in viewing conditions such that ε_i are Gaussian distributions with zero mean and standard deviations σ_i. Taking the derivative of

$$\frac{\mathrm{d}f(s_1,s_2)}{\mathrm{d}z} = \frac{\mathrm{d}f}{\mathrm{d}s_1}(\lambda_{10} + \ \varepsilon_1) + \frac{\mathrm{d}f}{\mathrm{d}s_2}(\lambda_{20} + \ \varepsilon_2),$$

where $\mathrm{d}f/\mathrm{d}s_i$ are calculated at s_{i0}, we observe a signal term $S = f_1\lambda_{10} + f_2\lambda_{20}$ (where $f_i = \mathrm{d}f\ /\mathrm{d}s_i$) and a noise term $\mathrm{E} = f_1\varepsilon_1 + f_2\varepsilon_2$ having s.d. $\sigma_\mathrm{E} = \sqrt{f_1^2\sigma_1^2 + f_2^2\sigma_2^2}$. If we minimize the Noise to Signal Ratio $NSR = \frac{\sigma_\mathrm{E}}{S}$ with respect to f_i (by solving for f_i the equation dSNR $/\mathrm{d}f_i = 0$), we find that the first derivatives of the function are $\mathrm{d}f\ /\mathrm{d}s_i \propto \lambda_{i0}\ /\sigma_i^2$. It can be shown that the derivatives $\mathrm{d}\hat{z}_C\ /\mathrm{d}s_i$ of the equation $\hat{z}_C = \beta\sqrt{\left(\frac{s_1}{\sigma_1}\right)^2 + \left(\frac{s_2}{\sigma_2}\right)^2}$ (calculated at s_{i0}) meet this requirement. By substituting $k_i = \beta(\lambda_i/\sigma_i)$, we obtain the Vector Sum equation $\hat{z}_C = \sqrt{(k_1 z)^2 + (k_2 z)^2}$ (easily generalizable to n signals).

References

1. Buelthoff HH, Yuille AL. 1991 Shape-from-X: psychophysics and computation. In *Proc. SPIE 1383, Sensor Fusion III: 3D Perception and Recognition (1 April 1991)*, pp. 235–246. Bellingham, WA: International Society for Optics and Photonics.

2. Freeman WT. 1994 The generic viewpoint assumption in a framework for visual perception. *Nature* **368**, 542–545. (doi:10.1038/368542a0)
3. Landy MS, Maloney LT, Johnston EB, Young, M. 1995 Measurement and modeling of depth cue combination: in defense of weak fusion. *Vision Res.* **35**, 389–412. (doi:10.1016/0042-6989(94)00176-M)
4. Knill DC, Richards W (eds). 1996 *Perception as Bayesian inference*. Cambridge, UK: Cambridge University Press.
5. Pizlo, Z. 2001 Perception viewed as an inverse problem. *Vision Res.* **41**, 3145–3161. (doi:10.1016/S0042-6989(01)00173-0)
6. Mamassian P, Landy M, Maloney LT. 2002 Bayesian modelling of visual perception. In *Probabilistic models of the brain: perception and neural function* (eds BA Olshausen, MS Lewicki), pp. 13–36. Cambridge, MA: The MIT Press.
7. Knill DC. 2003 Mixture models and the probabilistic structure of depth cues. *Vision Res.* **43**, 831–854. (doi:10.1016/S0042-6989(03)00003-8)
8. Geisler WS. 2003 Ideal observer analysis. In *The visual neurosciences,* vol. 1 (eds LM Chalupa, JS Werner). (doi:10.7551/mitpress/7131.003.0061)
9. Kersten D, Yuille A. 2003 Bayesian models of object perception. *Curr. Opin Neurobiol.* **13**, 150–158. (doi:10.1016/S0959-4388(03)00042-4)
10. Kersten D, Mamassian P, Yuille, A. 2004 Object perception as Bayesian inference. *Annu. Rev. Psychol.* **55**, 271–304. (doi:10.1146/annurev.psych.55.090902.142005)
11. Ernst MO, Banks MS. 2002 Humans integrate visual and haptic information in a statistically optimal fashion. *Nature* **415**, 429–433. (doi:10.1038/415429a)
12. Yuille A, Kersten D. 2006 Vision as Bayesian inference: analysis by synthesis? *Trends Cogn. Sci.* **10**, 301–308. (doi:10.1016/j.tics.2006.05.002)
13. Knill DC. 2007 Bayesian models of sensory cue integration. In *Bayesian brain: Probabilistic approaches to neural coding* (eds K Doya, S Ishii, A Pouget, RPN Rao), pp. 189–206. Cambridge, MA: MIT.
14. Maloney LT, Zhang H. 2010 Decision-theoretic models of visual perception and action. *Vision Res.* **50**, 2362–2374. (doi:10.1016/j.visres.2010.09.031)
15. Landy MS, Banks MS, Knill DC. 2011 Ideal-observer models of cue integration. In *Sensory cue integration* (eds J Trommershäuser, K Kording, MS Landy), pp. 5–29. (doi:10.1093/acprof:oso/9780195387247.003.0001)
16. Geisler WS. 2011 Contributions of ideal observer theory to vision research. *Vision Res.* **51**, 771–781. (doi:10.1016/j.visres.2010.09.027)
17. Vilares I, Kording K. 2011 Bayesian models: the structure of the world, uncertainty, behavior, and the brain. *Ann. N Y Acad. Sci.* **1224**, 22. (doi:10.1111/j.1749-6632.2011.05965.x)
18. Li Y, Sawada T, Shi Y, Kwon T, Pizlo, Z. 2011 A Bayesian model of binocular perception of 3D mirror symmetrical polyhedra. *J. Vis.* **11**, 11. (doi:10.1167/11.4.11)
19. Ma WJ. 2012 Organizing probabilistic models of perception. *Trends Cogn. Sci.* **16**, 511–518. (doi:10.1016/j.tics.2012.08.010)
20. Clark JJ, Yuille AL. 2013 *Data fusion for sensory information processing systems*. Vol. **105**. Berlin, Germany: Springer Science & Business Media.
21. Erdogan G, Jacobs RA. 2017 Visual shape perception as Bayesian inference of 3D object-centered shape representations. *Psychol. Rev.* **124**, 740. (doi:10.1037/rev0000086)
22. Mamassian P, Landy MS. 1998 Observer biases in the 3D interpretation of line drawings. *Vision Res.* **38**, 2817–2832. (doi:10.1016/S0042-6989(97)00438-0)
23. Knill DC. 1998 Discrimination of planar surface slant from texture: human and ideal observers compared. *Vision Res.* **38**, 1683–1711. (doi:10.1016/S0042-6989(97)00325-8)
24. Jacobs RA. 1999 Optimal integration of texture and motion cues to depth. *Vision Res.* **39**, 3621–3629. (doi:10.1016/S0042-6989(99)00088-7)
25. Schrater PR, Kersten D. 2000 How optimal depth cue integration depends on the task. *Int. J. Comput. Vision* **40**, 71–89. (doi:10.1023/A:1026557704054)
26. Saunders JA, Knill DC. 2001 Perception of 3D surface orientation from skew symmetry. *Vision Res.* **41**, 3163–3183. (doi:10.1016/S0042-6989(01)00187-0)
27. Jacobs RA. 2002 What determines visual cue reliability? *Trends Cogn. Sci.* **6**, 345–350. (doi:10.1016/S1364-6613(02)01948-4)
28. Hillis JM, Ernst MO, Banks MS, Landy MS. 2002 Combining sensory information: mandatory fusion within, but not between, senses. *Science* **298**, 1627–1630. (doi:10.1126/science.1075396)
29. Geisler WS, Kersten D. 2002 Illusions, perception and Bayes. *Nat. Neurosci.* **5**, 508. (doi:10.1038/nn0602-508)
30. Knill DC, Saunders JA. 2003 Do humans optimally integrate stereo and texture information for judgments of surface slant? *Vision Res.* **43**, 2539–2558. (doi:10.1016/S0042-6989(03)00458-9)
31. Hillis JM, Watt SJ, Landy MS, Banks MS. 2004 Slant from texture and disparity cues: optimal cue combination. *J. Vis.* **4**, 1. (doi:10.1167/4.12.1)
32. Adams WJ, Mamassian P. 2004 Bayesian combination of ambiguous shape cues. *J. Vis.* **4**, 7. (doi:10.1167/4.10.7)
33. Adams WJ, Graf EW, Ernst MO. 2004 Experience can change the 'light-from-above' prior. *Nat. Neurosci.* **7**, 1057–1058. (doi:10.1038/nn1312)
34. Knill DC. 2007 Learning Bayesian priors for depth perception. *J. Vis.* **7**, 13. (doi:10.1167/7.8.13)
35. Welchman AE, Lam JM, Bülthoff HH. 2008 Bayesian motion estimation accounts for a surprising bias in 3D vision. *Proc. Natl Acad. Sci. USA* **105**, 12 087–12 092. (doi:10.1073/pnas.0804378105)
36. Lovell PG, Bloj M, Harris JM. 2012 Optimal integration of shading and binocular disparity for depth perception. *J. Vis.* **12**, 1. (doi:10.1167/12.1.1)
37. Chen Z, Saunders JA. 2019 Perception of 3D slant from textures with and without aligned spectral components. *J. Vis.* **19**, 7. (doi:10.1167/19.4.7)
38. Todd JT, Bressan P. 1990 The perception of 3-dimensional affine structure from minimal apparent motion sequences. *Percept. Psychophys.* **48**, 419–430. (doi:10.3758/BF03211585)
39. Tittle JS, Todd JT, Perotti VJ, Norman JF. 1995 Systematic distortion of perceived three-dimensional structure from motion and binocular stereopsis. *J. Exp. Psychol. Percept. Perf.* **21**, 663. (doi:10.1037/0096-1523.21.3.663)
40. Norman JF, Todd JT, Phillips, F. 1995 The perception of surface orientation from multiple sources of optical information. *Percept. Psychophys.* **57**, 629–636. (doi:10.3758/BF03213268)
41. Todd JT, Tittle JS, Norman JF. 1995 Distortions of three-dimensional space in the perceptual analysis of motion and stereo. *Perception* **24**, 75–86. (doi:10.1068/p240075)
42. Norman JF, Todd JT, Perotti VJ, Tittle JS. 1996 The visual perception of three-dimensional length. *J. Exp. Psychol.* **22**, 173. (doi:10.1037/0096-1523.22.1.173)
43. Phillips F, Todd JT. 1996 Perception of local three-dimensional shape. *J. Exp. Psychol.* **22**, 930. (doi:10.1037/0096-1523.22.4.930)
44. Perotti VJ, Todd JT, Lappin JS, Phillips, F. 1998 The perception of surface curvature from optical motion. *Percept. Psychophys.* **60**, 377–388. (doi:10.3758/BF03206861)
45. Todd JT, Chen L, Norman JF. 1998 On the relative salience of Euclidean, affine, and topological structure for 3-D form discrimination. *Perception* **27**, 273–282. (doi:10.1068/p270273)
46. Domini F, Braunstein ML. 1998 Recovery of 3-D structure from motion is neither euclidean nor affine. *J. Exp. Psychol.* **24**, 1273. (doi:10.1037/0096-1523.24.4.1273)
47. Domini F, Caudek C, Richman, S. 1998 Distortions of depth-order relations and parallelism in structure from motion. *Percept. Psychophys.* **60**, 1164–1174. (doi:10.3758/BF03206166)
48. Caudek C, Domini F. 1998 Perceived orientation of axis rotation in structure-from-motion. *J. Exp. Psychol.* **24**, 609. (doi:10.1037/0096-1523.24.2.609)
49. Domini F, Caudek C. 1999 Perceiving surface slant from deformation of optic flow. *J. Exp. Psychol.* **25**, 426. (doi:10.1037/0096-1523.25.2.426)
50. Todd JT, Norman JF. 2003 The visual perception of 3-D shape from multiple cues: are observers capable of perceiving metric structure? *Percept. Psychophys.* **65**, 31–47. (doi:10.3758/BF03194781)
51. Domini F, Caudek C. 2003 3-D structure perceived from dynamic information: a new theory. *Trends Cogn. Sci.* **7**, 444–449. (doi:10.1016/j.tics.2003.08.007)
52. Liu B, Todd JT. 2004 Perceptual biases in the interpretation of 3D shape from shading. *Vision Res.* **44**, 2135–2145. (doi:10.1016/j.visres.2004.03.024)
53. Todd JT. 2004 The visual perception of 3D shape. *Trends Cogn. Sci.* **8**, 115–121. (doi:10.1016/j.tics.2004.01.006)
54. Norman JF, Todd JT, Orban GA. 2004 Perception of three-dimensional shape from specular highlights, deformations of shading, and other types of visual

information. *Psychol. Sci.* **15**, 565–570. (doi:10.1111/j.0956-7976.2004.00720.x)

55. Todd JT, Thaler L, Dijkstra TM. 2005 The effects of field of view on the perception of 3D slant from texture. *Vision Res.* **45**, 1501–1517. (doi:10.1016/j.visres.2005.01.003)
56. Norman JF, Todd JT, Norman HF, Clayton AM, McBride TR. 2006 Visual discrimination of local surface structure: slant, tilt, and curvedness. *Vision Res.* **46**, 1057–1069. (doi:10.1016/j.visres.2005.09.034)
57. Todd JT, Thaler L, Dijkstra TM, Koenderink JJ, Kappers AM. 2007 The effects of viewing angle, camera angle, and sign of surface curvature on the perception of three-dimensional shape from texture. *J. Vis.* **7**, 9. (doi:10.1167/7.12.9)
58. Todd JT, Thaler L. 2010 The perception of 3D shape from texture based on directional width gradients. *J. Vis.* **10**, 17. (doi:10.1167/10.5.17)
59. Fantoni C, Caudek C, Domini, F. 2010 Systematic distortions of perceived planar surface motion in active vision. *J. Vis.* **10**, 12. (doi:10.1167/10.5.12)
60. Volcic R, Fantoni C, Caudek C, Assad JA, Domini F. 2013 Visuomotor adaptation changes stereoscopic depth perception and tactile discrimination. *J. Neurosci.* **33**, 17 081–17 088. (doi:10.1523/JNEUROSCI.2936-13.2013)
61. Todd JT, Egan EJ, Phillips F. 2014 Is the perception of 3D shape from shading based on assumed reflectance and illumination? *i-Perception* **5**, 497–514. (doi:10.1068/i0645)
62. Egan EJ, Todd JT. 2015 The effects of smooth occlusions and directions of illumination on the visual perception of 3-D shape from shading. *J. Vis.* **15**, 24. (doi:10.1167/15.2.24)
63. Bozzacchi C, Domini F. 2015 Lack of depth constancy for grasping movements in both virtual and real environments. *J. Neurophysiol.* **114**, 2242–2248. (doi:10.1152/jn.00350.2015)
64. Bozzacchi C, Volcic R, Domini, F. 2016 Grasping in absence of feedback: systematic biases endure extensive training. *Exp. Brain Res.* **234**, 255–265. (doi:10.1007/s00221-015-4456-9)
65. Campagnoli C, Croom S, Domini, F. 2017 Stereovision for action reflects our perceptual experience of distance and depth. *J. Vis.* **17**, 21. (doi:10.1167/17.9.21)
66. Kopiske KK, Bozzacchi C, Volcic R, Domini, F. 2019 Multiple distance cues do not prevent systematic biases in reach to grasp movements. *Psychol. Res.* **83**, 147–158. (doi:10.1007/s00426-018-1101-9)
67. Storrs KR, Anderson BL, Fleming RW. 2021 Unsupervised learning predicts human perception and misperception of gloss. *Nat. Hum. Behav.* **5**, 1402–1417. (doi:10.1038/s41562-021-01097-6)
68. Storrs KR, Fleming RW. 2021 Learning about the world by learning about images. *Curr. Dir. Psychol. Sci.* **30**, 120–128. (doi:10.1177/0963721421990334)
69. Fleming RW, Storrs KR. 2019 Learning to see stuff. *Curr. Opin. Behav. Sci.* **30**, 100–108. (doi:10.1016/j.cobeha.2019.07.004)
70. Kemp J, Domini F. 2021 Just noticeable differences in 3D shape perception: a measure of estimation noise or cue gain? *J. Vis.* **21**, 2889. (doi:10.1167/jov.21.9.2889)
71. Domini F, Caudek C. 2009 The intrinsic constraint model and Fechnerian sensory scaling. *J. Vis.* **9**, 25. (doi:10.1167/9.2.25)
72. Domini F, Caudek C. 2010 Matching perceived depth from disparity and from velocity: modeling and psychophysics. *Acta Psychol.* **133**, 81–89. (doi:10.1016/j.actpsy.2009.10.003)
73. Kemp J, Cesanek E, Domini F. 2022 Perceiving depth from texture and disparity cues: evidence for a non-probabilistic account of cue integration. *bioRxiv*, 2022.10.20.513044. (doi:10.1101/2022.10.20.513044)
74. Domini F, Caudek C, Tassinari H. 2006 Stereo and motion information are not independently processed by the visual system. *Vision Res.* **46**, 1707–1723. (doi:10.1016/j.visres.2005.11.018)
75. Tassinari H, Domini F, Caudek C. 2008 The intrinsic constraint model for stereo-motion integration. *Perception* **37**, 79–95. (doi:10.1068/p5501)
76. Domini F, Shah R, Caudek C. 2011 Do we perceive a flattened world on the monitor screen?. *Acta Psychol.* **138**, 359–366. (doi:10.1016/j.actpsy.2011.07.007)
77. Foster R, Fantoni C, Caudek C, Domini, F. 2011 Integration of disparity and velocity information for haptic and perceptual judgments of object depth. *Acta Psychol.* **136**, 300–310. (doi:10.1016/j.actpsy.2010.12.003)
78. Vishwanath D, Domini F. 2013 Pictorial depth is not statistically optimal. *J. Vis.* **13**, 613. (doi:10.1167/13.9.613)
79. Domini F, Caudek C, Trommershäuser J, Körding KP, Landy MS. 2011 Combining image signals before three-dimensional reconstruction: the intrinsic constraint model of cue integration. In *Sensory cue integration* (eds J Trommershäuser *et al.*), pp. 120–143. (doi:10.1093/acprof:oso/9780195387247.003.0007)
80. Domini F, Caudek C. 2013 Perception and action without veridical metric reconstruction: an affine approach. In *Shape perception in human and computer vision*, pp. 285–298. London, UK: Springer.
81. Anderson BL. 2015 Can computational goals inform theories of vision? *Topics Cogn. Sci.* **7**, 274–286. (doi:10.1111/tops.12136)
82. Marlow PJ, Anderson BL. 2021 The cospecification of the shape and material properties of light permeable materials. *Proc. Natl Acad. Sci. USA* **118**, e2024798118. (doi:10.1073/pnas.2024798118)
83. Di Luca M, Domini F, Caudek C. 2010 Inconsistency of perceived 3D shape. *Vision Res.* **50**, 1519–1531. (doi:10.1016/j.visres.2010.05.006)
84. Domini F, Vishwanath D. 2020 Computational models of 3D-cue integration. In *Encyclopedia of Computational Neuroscience* (eds D Jaeger, R Jung). New York, NY: Springer.
85. Campagnoli C, Hung B, Domini, F. 2022 Explicit and implicit depth-cue integration: evidence of systematic biases with real objects. *Vision Res.* **190**, 107961. (doi:10.1016/j.visres.2021.107961)
86. Vishwanath D, Hibbard PB. 2013 Seeing in 3-D with just one eye: stereopsis without binocular vision. *Psychol. Sci.* **24**, 1673–1685. (doi:10.1177/0956797613477867)