Compositionality in Perception: A Framework

Kevin J. Lande
Department of Philosophy & Centre for Vision Research
York University

Abstract: Perception involves the processing of content or information about the world. In what form is this content represented? I argue that perception is widely compositional. The perceptual system represents many stimulus features (including shape, orientation, and motion) in terms of arrangements of other features (shape parts, slant and tilt, grouped and residual motion vectors). But compositionality can take a variety of forms. The ways in which perceptual representations compose are markedly different from the ways in which sentences or thoughts are thought to be composed. Throughout, I suggest that the thesis that perception is compositional is not itself a concrete hypothesis with specific predictions, but rather it affords a framework for developing and evaluating empirical hypotheses about the nature of perceptual representations. The question is not just *whether* perception is compositional, but *how*. Answering this latter question can provide fundamental insights into the place of perception within the mind.

1. INTRODUCTION

One of the great challenges of modern science is to understand the immense creative powers of thought and language—our abilities to produce, understand, and communicate a seemingly infinite variety of spontaneous thoughts and plans, which can be arbitrarily unmoored from present concerns and circumstances. But impressive too is the immense receptivity of the mind to the immediate conditions in which it finds itself. One is capable of perceptually representing patterns that one has never encountered before—endlessly variable arrangements of figures, of different shapes, sizes, and textures, changing and interacting over time. One can at once perceptually distinguish these patterns while also being sensitive to their similarities. How does the mind administer such a finely differentiated yet well-catalogued repertoire of representations?

The creativity of thought and language is thought to be powered by the capacity for *composition*—roughly, the ability to represent something (that Ida loves Una) by combining more elementary representations (of Ida, Una, and loving). Is the receptivity of perception also rooted in the power of compositionality? Many philosophers have thought that the compositional powers of language or thought are primary and make possible our abilities to distinguish and represent objective patterns in the world. One possibility is that perception itself wholly lacks compositional structure. Perception without thought is blind; a blooming, buzzing confusion of unstructured information (Dretske, 1981). Another possibility is that insofar as there is compositionality in perception, it will be similar in form to the type of compositionality that resides in language and thought (McDowell, 1996). This might be the case if the role of compositionality in perception is to encode perceptual content in a format that is readable by cognition (Cavanagh, 2021; Quilty-Dunn et al., 2022). These possibilities have corollaries in contemporary deep neural network models of vision, where it sometimes happens that either the representations of a network are high-dimensional vectors with uninterpretable bases or that what structure they have is impressed on them by the task of assigning conceptual labels or descriptions.

I will argue that perception is compositional in its own right. The compositionality of perception is implicit in a wide variety of theories in perception science. David Marr once characterized perceptual representations as constituting a “formal scheme,” comprising a set of elementary representations and “rules for putting them together” (Marr, 1982, p. 21), with Stephen Palmer writing that “perceptual representations are selectively organized data structures” (Palmer, 1977, p. 442). But compositionality can come in a variety of forms. I argue that compositionality in perception is quite unlike the compositionality that is thought to characterize language and thought. An underlying moral is that the thesis that perception is compositional is not itself a concrete hypothesis that makes specific predictions. Instead, compositionality, like Bayesianism (Knill et al., 1996; Griffiths et al., 2012) or neuroconnectionism (Doerig et al., 2023), should be construed as a theoretical framework for posing and answering empirical questions about the nature of perceptual representations. The question is not simply *whether* perception is compositional, but *how*. Answering this latter question provides special insights into the place of perception within the mind.

2. What is compositionality?

When one asks about the “composition” of a photograph, a novel, or a mathematical function, one is asking how one thing depends for certain of its properties on the way it is combined from its parts and certain of their properties. The relevant notions of “thing,” “part,” and “combination” can vary depending on the subject matter (a rock song and the temporal arrangement of its musical phrases vs. a sample of rock and its physical aggregation of constituent minerals), as can the relevant properties (one can ask about a sculpture’s aesthetic composition or its chemical composition). *Semantic compositionality* is specific to representations—roughly, things (sentences, maps, pictures) or psychological states (perceptions, memories, thoughts) that matter in large part because of what they are about. A newspaper headline is a representation: it may be written in a visually striking font, but critically it also has content about some world event. Your visual state is also a representation: it might correspond to relative influx of oxygen to certain parts of your brain, but critically it also carries content about the shape, orientation, and motion of the object that is spinning past you.

Some representations, such as the newspaper headline, have other representations (words) as parts. Visual representations, I will argue, also have parts. Semantic compositionality concerns the relationship between what a whole representation is about and what its parts are about. The concept of a *semantically compositional* *representation* has its origins in philosophy of language, logic, and linguistics, where it is typically traced back to the work of Gottlob Frege. It is standardly defined as follows (Partee, 2004; Janssen, 2011):

> **Compositional Representation:** A representation is *compositional* if and only if its content depends wholly on the contents of its constituent parts and the way they are structurally combined.

For example, the phrase "red balloon" is a compositional expression. We can characterize the meaning of the phrase by noting that it is true of anything that is both red and a balloon. It applies to just these things because (i) "red balloon" is a combination of an adjective "red" and a noun "balloon," (ii) "red" is true of anything that is red and "balloon" refers to the set of things that are balloons, and (iii) in the standard case, a combination of an adjective, A, and a noun, N, is true of anything that is a member of the set designated by N and of which A is true. More generally, if a representation is compositional, then one can exhaustively characterize its content (what it is about) by answering what I will call the "Three Cs":

The Three Cs:

(1) What are the elementary, or primitive, *constituents* of the representation?

(2) In what way are the constituents structurally *combined*?

(3) How does the *content* of a representation depend on the contents of its constituents, given the manner in which they are combined?

We can also pose the Three Cs for a system as a whole, which is tantamount to asking about the system's "lexicon," its "syntax," and its "compositional semantics," respectively. A set of representations is a *compositional system* just in case each structured representation in the system is compositional and, moreover, there are common principles—answers to the Three Cs—that determine the contents of multiple representations in the system. For example, "white rabbit," "square hat," and "litigious pony" are all governed by the A-N principle.

"Content," "constituent part," and "structural combination" all refer to theoretical concepts. While theorists differ in how they understand and even label these concepts, most agree on some core points. First, an essential feature of a representation is that it has the function of being about something. The "content" of a representation––as with the content of a newspaper headline—determines what it is about. The content of the sentence "Ida loves Una" is about Ida's loving Una; the content of the monkey's current visual state is about (among other things) the estimated orientation of the target Gabor patch.

Second, just as chemical laws determine what molecules are structurally possible, psychological constraints determine what representations are structurally possible. "Ida loves Una" is a possible sentence in my idiolect because it has the right kind of constituents, combined in the right kind of way. “Ida Una balloon luft” doesn’t, and so isn’t a possible sentence in my idiolect (though it is physically possible for me to pronounce and perceive it). A theory of representational structure (a “grammar”) aims to explain what representations are possible in a system—and therefore available for computing or inferring with—in terms of how representations can, cannot, or must combine to be constituents of a common, composite representation (a data structure, if you like). A theory of representational content (“semantics”) aims to explain what those representations are about (Lande, 2023; Larson & Segal, 1995). In a compositional system, the same principles that delimit which representations are possible also delimit what those representations are about.

That a system is compositional is not a concrete hypothesis that makes specific empirical predictions. It is a theoretical framework for formulating such hypotheses. To hypothesize that a system is compositional is simply to say that the space of its representations can be spanned by filling in the Three Cs. Predictions are made when one hypothesizes specific values for these parameters. Still, it is an empirical question whether the framework is feasible. Filling in the Three Cs is no trivial task and it is not self-evident that it will always prove to be a productive endeavour. Prior to the research programs championed by Noam Chomsky and Richard Montague, many doubted that meaning in natural language could be explained in terms of the Three Cs. Many still doubt it. The claim that thought is compositional has proven to be controversial (Fodor, 1975; Quilty-Dunn et al., 2022). Many have especially doubted the compositionality of psychological capacities outside of language and thought.

3. Compositionality in vision

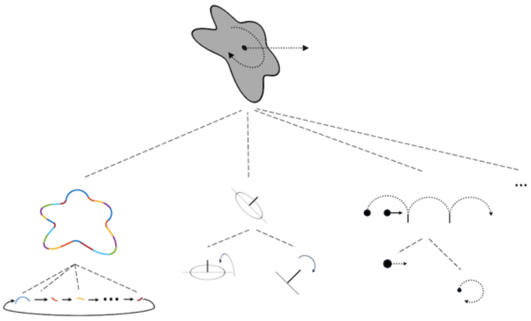

Figure 1: Vision is compositional. A spinning object moving through the scene is represented in terms of (among other things) its shape, 3D orientation, and motion, which themselves are represented, respectively, in terms of configurations of shape parts, slant and tilt, and the boundary’s rotational motion around the object’s centre with the linear motion of the object’s centre.

Vision is compositional. Vision science contains a host of hypotheses about how one type of stimulus feature is represented “in terms of” other features. These constitute hypotheses about how one representation is composed of others (Figure 1). Here are some examples:

- A representation of a surface's orientation in depth is composed from representations of its *slant* and *tilt* (Stevens, 1983; Nguyenkim & DeAngelis, 2003).
- A representation of an element's motion (e.g. the trajectory of part of a wheel rolling across the ground) is composed from a representation of the common motion that it shares with other elements (the linear motion of the wheel's centre) and a representation of the element's residual motion (its rotation about the centre; Johansson, 1973; Gershman et al., 2016).
- A representation of visual texture is composed from representations of summary statistics of luminance, orientation, and so on (Portilla & Simoncelli, 2000; Balas et al., 2009).
- A representation of an object's orientation relative to an extrinsic object (e.g. the orientation of my pen on the notepad) is composed from representations of "(a) a correspondence between object axes and external axes, (b) the tilt of the object axes relative to the external axes, and (c) the relationship between the polarity of the object axes and the polarity of the external axes" (McCloskey et al., 2006, p. 685).
- A representation of an object is composed from representations of features—including its orientation, motion, texture, etc.—perhaps with a separate "index" for the object that is invariant across feature changes (A. Treisman, 1986; Pylyshyn, 1989; Kahneman et al., 1992; Green & Quilty-Dunn, 2017).
- A representation of a scene is composed from representations of objects and their relations (Biederman et al., 1982; Zhu & Mumford, 2006; Võ, 2021; Hafri & Firestone, 2021); or from representations of summary statistics of features, e.g. the mean depths of surfaces within the scene (Greene & Oliva, 2009).

Even if none of these hypotheses is exactly right, that they all presuppose compositionality suggests that this is a productive framework for zeroing in on the truth.

3.1 Compositionality in Shape Perception

Compositionality has played an especially central role in approaches to one of the fundamental problems in vision science: how humans and primates represent shape. Since the representation of three-dimensional shape has proven to be elusive to model, researchers have often focused on the more constrained problem of how we represent the two-dimensional outlines of things (Elder, 2018; Todd & Petrov, 2022). Many approaches to this problem hinge on how they answer the Three Cs (Figure 2; Table 1).

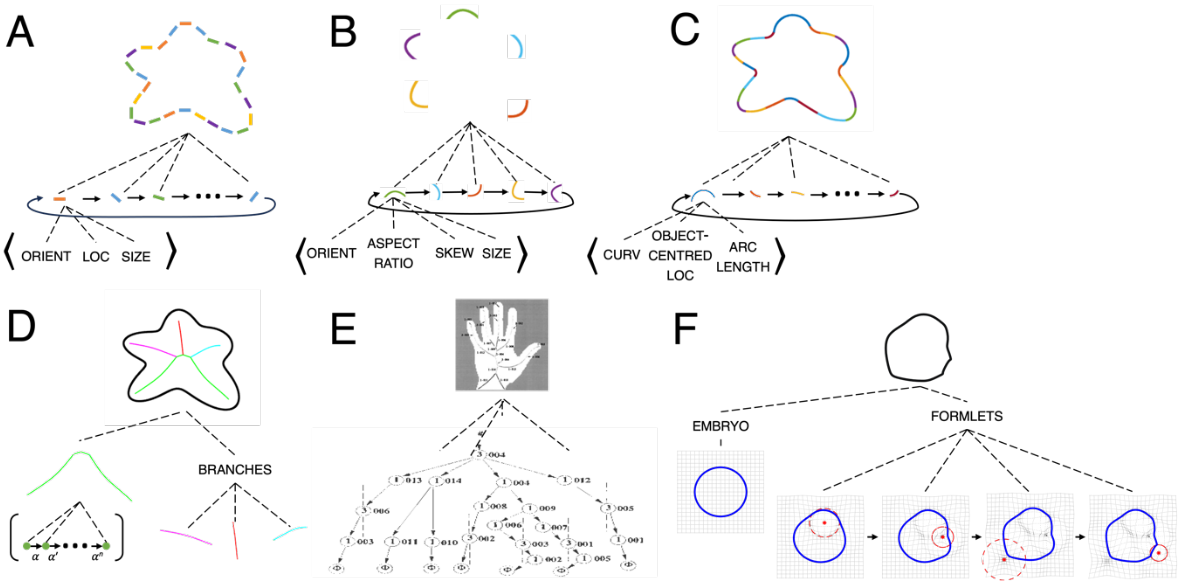

Figure 2: Different models of compositional representations of two-dimensional shape. Dashed lines indicate constituency. Solid lines indicate structural relationships between constituents, i.e. the "manner in which they are combined" (e.g., that one representation precedes another within an ordered sequence). Some constituents are taken to be feature-vectors that decompose into component coordinates, representing orientation, location, size, etc. See Table 1 for corresponding combinatorial constraints and their semantic import. A-D are part-based schemes. A: An edge code, in which shape representations are composed from an ordered sequence of representations of local oriented line segments (Elder & Goldberg, 2002; Geisler & Super, 2000). B: A convexity code, in which shapes are represented in terms of convex segments of their contour (Hoffman & Richards, 1984; Richards & Hoffman, 1985; Schmidtmann et al., 2015). C: A constant curvature code, in which shapes are represented in terms of segments of constant curvature, or "arclets" (N. Baker et al., 2020; N. Baker & Kellman, 2021; Kellman et al., 2013). D: A medial axis code, in which a shape is represented in terms of local axes of symmetry, which are related as principal and branching axes (Blum & Nagel, 1978; Feldman & Singh, 2006; Feldman et al., 2013; Green, 2023). E-F are transformational schemes. E: Shock grammars represent shapes in terms of a hierarchical tree of expansions, contractions, collisions, and mergings of symmetric shapes (Siddiqi & Kimia, 1996; Kimia, 2003) [Adapted from (Siddiqi et al., 1999, p. 26) with kind permission from Springer]. F: A formlet scheme takes shapes to be coded in terms of sequences of warping transformations applied to an embryonic template shape (Elder et al., 2013) [Adapted from (Elder et al., 2013, p. 5) with kind permission from Elsevier]. [Thanks to Nick Baker for assistance with the figure.]

System	**1. Constituents**	**2. Combination**	**3. Content**
A. Edge code	Representations of oriented edges, composed from representations of – Orientation – Location – Size	– *Form*: Constituents are arranged in an ordered sequence (or cycle). – *Constraint*: Adjacent constituents must satisfy good continuation and proximity.	A shape that has each represented edge as a part, such that adjacent constituents represent neighbouring, approximately collinear edges.
B. Convexity code	Representations of contour segments, composed from representations of – Orientation – Aspect ratio – Skew – Size	– *Form:* Ordered sequence. – *Constraint:* Good continuation and proximity.	A shape that has each represented segment as a part along with smooth connections between those segments, such that adjacent constituents represent neighbouring convex parts.
C. Constant curvature code	Arc-units, composed from representations of – Curvature – Object-centred location – Arc-length	– *Form:* Ordered sequence. – *Constraint:* Good continuation and proximity.	A shape that has each represented segment as a part, such that adjacent constituents represent neighbouring, approximately collinear parts.
D. Medial axis code	Local axes of symmetry, composed of a series of points related by turning angles.	– *Form:* Hierarchical. – *Smoothness Constraint:* The shallower the turning angle between represented points, the more the representations of those points can be combined into an axis representation. – *Simplicity Constraint:* There is a fixed cost to adding constituents. Structures with fewer constituents are better than structures with more constituents.	The main axis represents the global shape and branching axes represent component regions of the shape, such for each axis there is a region in the shape that is approximately symmetric about that axis.

E. Shock grammar	Representatiosn of shock types, corresponding to different ways symmetric shapes might expand, collide, merge, or collapse.	– *Form*: Hierarchical. – *Constraint:* A set of production rules constraining the possible combinations and orderings of shocks.	A shape that results from the hierarchical ordering of represented shocks.
F. Formlet code	– *Embryos*, i.e. representations of basic template shapes. – *Formlets*, i.e. representations of localized warping transformations.	– *Form:* Ordered sequence. – *Constraint:* Initial element is an embryo. Subsequent elements are formlets (in any arbitrary sequence).	A shape that results from the sequential application of warping transformations to the represented embryonic shape.

Table 1: Compositional schemes for representing shape differ in (1) what they take to be the constituents of shape representations, (2) how those constituents can be combined (including their form, or the type of structural relations they stand in when combined, and the constraints they must satisfy to stand in those relations), and (3) how the content of the whole shape representation is derived from the contents of its constituents and their manner of combination. See Figure 2 for illustrations of these different schemes.

3.1.1 Contour Parts

"Part-based" accounts hold that the representation of a shape is composed from constituent representations of parts of that shape and their configuration. According to one family of theories (Figure 2A-C; Table 1A-C), the relevant shape-parts are segments along the object's contour. Some models take the constituents to be representations of local oriented edges (Geisler et al., 2001; Elder & Goldberg, 2002). Most take the constituents to be representations of curved segments, whether just the convex segments or bumps (Hoffman & Richards, 1984; Richards & Hoffman, 1985; De Winter & Wagemans, 2006; Schmidtmann et al., 2015; Figure 1B); or the concavities (dents) and sides as well (Bell et al., 2010); or segments of constant curvature (Kellman et al., 2013; N. Baker et al., 2020; N. Baker & Kellman, 2021; Figure 1C). These part-representations themselves decompose further into coordinates on dimensions of curvature, orientation, size, and so on.

Many contour-based schemes assume that shape representations have the form of ordered sequences (or cycles, in the case of closed figures) in which one part-representation is structurally adjacent to another (Elder & Goldberg, 2002; Figure 2A-C; Table 1A-C). Standardly, whether two representations can combine as adjacent constituents is a function of whether they represent segments as having positions and orientations that are relatively collinear and nearby (Field et al., 1993; Boucart et al., 1994; Elder, 2018; Kellman & Fuchser, 2023). That is, the principle of combination reflects the Gestalt rule of "good continuation" (Lande, 2021).

If two part-representations are structurally adjacent, this signifies that the represented parts are topologically adjacent and more or less well-aligned. So, the constituents contribute content about the parts of the shape (their curvature, orientation, and so on), and the structural relationship between the constituents (the manner in which they are combined) contributes content about the configuration of those parts.

3.1.2 Medial Axes

Hafri et al. (2023) have recently proffered another part-based scheme, medial axis representations, as an example *par excellence* of compositional perceptual representation. Your palm and your fingers are each locally symmetric about their own axes. A medial axis representation of the shape of your hand is composed from representations of these different axes (Blum, 1973; Feldman & Singh, 2006; Feldman et al., 2013; Green, 2023; Figure 2D; Table 1D). The axis representations are related hierarchically, with a main axis that captures the main body of the shape (your palm) and branch axes for the parts (fingers). An axis might itself be coded as a sequence of points (Feldman et al., 2013, p. 57). A smoothness constraint entails that an axis representation is more likely to decompose into a collinear set of points (related by shallower turning angles) than a ragged or twisty set of points. A simplicity constraint entails that there is a fixed cost to adding axes to a shape representation. The constituents of a medial axis representation carry content about the local axes around which parts of the shape are symmetrical; the way those constituents are combined contributes content about the topological relationships among those parts.

3.1.3 Transformational Models

Not all theories take shape representations to be part-based. Two examples are "shock grammars" (Siddiqi & Kimia, 1996; Figure 2E; Table 1E), which are closely related to medial axis representations, and "formlets" (Elder et al., 2013; Figure 2F; Table 1F). In these schemes, constituents represent distinct stages in some abstract evolution of the shape. In a shock grammar, a shape is represented in terms of a hierarchy of operations of growth, contraction, collision, and mergings of shapes. The representations of these operations are combined according to grammar-like production rules. In a formlet system, a shape representation has the form of an ordered sequence, where an initial constituent represents an embryonic shape such as an ellipse and other constituents represent spatial warping functions. In either case, the representation is a composite data structure that specifies each of the relevant shape transformations, and which represents the shape that would be generated from the actual realization of those transformations in the relevant order.

3.2 Diagnosing Structure

There is broad, tacit consensus about how to empirically evaluate hypotheses about each of the Three Cs. I focus here on evaluating claims about constituency and combination. The experimental logic for tests of constituency and combinability rest on a few defeasible principles (Lande, 2021; Schwartz & Sanchez Giraldo, 2017):

(1) A representation's constituents are necessary for forming that representation.

Without the parts, you can't have the whole. So, as Schwartz and Sanchez Giraldo write, "Coarsely, if some aspect of a task (e.g., adaptation or forgetting) affects the activity of a representational unit, then it will affect decisions, storage, and recall of all aspects of the input that share this representational unit either directly, or through the hierarchical construction of a representation" (Schwartz & Sanchez Giraldo, 2017, p. 6). Hence, a count against an "oriented edge" code (Figure 2A) is that local edge information can be substantially varied without variation in the overall shape representation (Boucart et al., 1994; N. Baker & Kellman, 2018). By contrast, when curvature information is masked, this substantially impacts shape representation (Biederman & Cooper, 1991; De Winter & Wagemans, 2006; Habak et al., 2004; Bell et al., 2010; Schmidtmann et al., 2015). This suggests that even if edges are *cues* to shape, shape is *coded in terms of* curvature.

(2) Distinct constituents are approximately conditionally independent of each other, given that they belong to the same representation.

If different parts receive different representations, then one should be able to introduce internal or external noise to the representation of one part without thereby inducing variability in the representation of the others (Garner, 1974; Bell et al., 2009; Denisova et al., 2016). Conversely, the same constituent can be preserved ("reused") despite variation in the other constituents with which it is combined (Palmer, 1977; Ankrum & Palmer, 1991). To be clear, one shouldn't expect perfect or absolute independence. One of the strengths of the compositional framework is that one can express *interdependencies*, or *co-determination*, among constituents (Zhu & Mumford, 2006; Schwartz & Sanchez Giraldo, 2017). If good continuation is a constraint on the combinability of two contour representations, then the visual system might be biased toward representing one contour's orientation as more collinear to the others so that their representations can be combined (Keemink & van Rossum, 2016; Schwartz et al., 2009). But the represented orientations are approximately independent if we conditionalize on their being encoded by parts of an integrated representation.

While much of the focus in vision science is on identifying the distinct constituents of a representation (the first C), it is just as important to spell out the constraints on how constituents can combine (the second C).

(3) Combinatorial constraints determine which representations are possible in the system as a function of how those constituents are related.

For example, if good continuation is a constraint on combining contours, then without any good continuation at any spatial scale, you won't be able to represent an integrated contour (Field et al., 1993; Lande, 2021). Conversely, one's theory should not "over-generate" representations: if all the combinatorial constraints in your model allow a given representation, then it should be possible in principle to induce that representation in subjects (notwithstanding limits on memory, attention, and so on).

Compositionality provides an effective and tractable framework for formulating and testing hypotheses about the way content is encoded in vision. This is good reason to think that vision is compositional. Still, the framework and its experimental logic are often only implicit in vision research. Making these explicit might encourage more systematic and consistent development, testing, and comparison of models.

4. DOES VISION HAVE A LANGUAGE?

Some hold that insofar as perceptual representations are compositional, they have a similar form as representations in cognition and language. Indeed, it is sometimes suggested that perceptual representations have a "grammar" (Gregory, 1970; Võ, 2021) or "logic" (Rock, 1985), that they are like "descriptions" (Hummel, 2013) with "logical predicates" (Feldman, 1997), or that there is a "language of vision" (Cavanagh, 2021). On the weak interpretation, the claim is simply that perceptual representations are compositional. But as Elisabeth Camp (2007) points out, this does not imply a stronger interpretation that the syntax and semantics of perception are substantially like that of language or thought. In fact, there seem to be marked differences.

As Hafri et al. (2023) and Lande (2023) have noted, perception lacks the expressive power of cognition. Consider the "lexicon" of constituents available in perception. There is no evidence for explicit representations of logical relations (*not*, *or, if…then…*) in perception (Block, 2023, ch. 3). Moreover, many categories of words and concepts are open class—you can endlessly expand your stock of nouns and verbs, thing-concepts and property-concepts. Correspondingly, there are few constraints on what a noun (or thing-concept) can represent: a person, a transfinite cardinal, or a weather event. By contrast, the dimensions along which we can perceptually represent things (shape, texture, and so on) are arguably quite fixed (Green, 2020).

These differences in the constituents available to perception and language correspond to differences in the ways those constituents can combine. Nouns with quite different meanings can be interchanged within a sentence without threatening its grammaticality ("Ida loves [Una / aleph-null / the weather]"). This pattern of intersubstitutability defines the syntactic role of nouns. There is much less freedom for substitution between perceptual representations. For example, Green and Quilty-Dunn (2017) argue that object files have "feature-specific slots": they can be composed from a limited number of shape representations, a limited number of orientation representations, and so on. Shapes can be swapped for shapes and orientations for orientations; but shapes cannot enter into the orientation slot or *vice versa* (see also Ashby, 2020). The compositional roles of perceptual representations are much finer-grained than those of nouns and adjectives, say.

The grammar of language permits us to express unexpected and even contradictory claims. This incurs the risk of saying and thinking highly inaccurate things, but it also enables us to state and argue for deep and surprising insights—that the apple and the moon are governed by the same laws of motion, or that there are infinitely many prime numbers. Perception is notably more conservative. It is unlikely that the motion of the moon across the sky and the motion of the apple falling to the ground would ever be grouped together visually. Likewise, the sorts of contours that can be represented given the combinatorial constraint of good continuation are just the sorts of contours that tend to arise in natural scenes (Geisler et al., 2001; Elder & Goldberg, 2002). In language, well-formedness guarantees neither plausibility nor even meaningfulness. In perception, the combinability of representations tends to track the likelihood that the compound representation would be accurate.

One motivation for thinking that compositionality in perception would be like compositionality in cognition is the idea that compositional representations are a way of packaging perceptual content into a format that is readable by language and thought (Cavanagh, 2021; Quilty-Dunn et al., 2022). But we have seen that compositionality is posited throughout vision, not just at its interface with cognition. The constituents of orientation, texture, motion, and shape representations may be far removed from the kinds of features we tend to talk about and think about when describing our visual world. Moreover, the compositional principles of perception seem to subserve a central function of perception itself: to accurately represent the distal causes of sensory stimulations (Helmholtz, 1925; Knill et al., 1996; Graham, 2014). The perceptual code ensures that the hypothesis space of perception is well-fitted to its ecological niche, containing hypotheses only insofar as they are plausible given the laws and regularities by which distal objects cause proximal sensations. No such constraints bind the hypothesis space of thought. The compositional principles governing perception may in some cases serve to facilitate the interface with language and thought; but in many cases they may be better understood as subserving endogenous functions of perception itself.

5. WHAT COMPOSITIONALITY IS NOT

It remains to dispel some potential confusion about what compositionality in perception entails. In perceptual psychology and computer vision, the term "compositional representation" is often taken as synonymous with "part-based representation," whereby something is represented in terms of its parts and their relationships. But semantically compositional representations need not strictly be part-based. The phrase, "The 16th President of the United States," is a compositional representation of Abraham Lincoln; but the constituent, "the United States," does not represent a literal part of Abraham Lincoln (Block, 1983). Visual representations of texture are semantically compositional, but their constituents represent the statistics of a texture rather than its individual parts (Balas et al., 2009). Transformational models of shape perception represent shapes in terms of transformations that would produce the shape, rather than parts of the shape itself. If some perceptual representations are in fact part-based, then this is a substantive fact about the way they are compositional.

Talk of "combination" and "composition" might encourage a picture of processing, according to which the constituents are formed first and then put together (A. M. Treisman & Gelade, 1980). But compositionality does not require an assembly line. Many compositional theories of vision posit an "analysis-by-synthesis" architecture in which representations are refined through the convergence of bottom-up cues and prior expectations (Yuille & Kersten, 2006). By analogy, that the human body is composed of limbs and other body parts does not mean that those parts were pre-fabricated and then assembled like building blocks. Moreover, while we can expect attention and memory to be able to operate on the constituents of a compositional representation, there is no requirement that they must always do so. Recognitional processes, for example, need not match each constituent of a visual representation to a representation in memory (Barenholtz & Tarr, 2006). Semantic compositionality is a feature of representations; it is consistent with different ways in which those representations might be formed and used (Lande, 2021).

Finally, it is helpful to distinguish *psychological* (or *cognitive*) representation from *neural* representation. Psychological representations are typically posited as elements of models that explain psychophysical patterns in how individuals experience and behaviorally respond to stimuli (Palmer, 1978; Burge, 2014). Neural representations are patterns of neural activity that are investigated through various neuroscientific probes, such as neuroimaging and extracellular recordings (Cao, 2022). A central task of cognitive neuroscience is to validate "linking hypotheses" that relate the sorts of representations posited in psychological models to the sorts of representations probed in neuroscience (Teller, 1984; Shea, 2018; B. Baker et al., 2022). In the case of compositionality, these linking hypotheses might be straightforward—for example, associating distinct psychological constituents with distinct cells or neural ensembles (Vaziri et al., 2009). But the links might be more obscure. Likewise, while some deep neural network models of vision seem to eschew compositionality, structural diagnostics, of the sort canvassed above, can be applied to neural networks, possibly providing evidence of representations with compositional structure like that diagnosed in primate vision or in human language (Zerroug et al., 2022; Hupkes et al., 2020). It is a matter of active investigation what properties of neural networks might give rise to such structure. In any case, uncertainty about either biological or computational linking hypotheses does not invalidate the existing psychophysical methods for evaluating hypotheses about the structures of psychological representations. Indeed, such methods are critical for validating candidate linking hypotheses.

Conclusion

Perception involves encoding content or information about the world. In what form is this content represented? I have argued that perception is compositional: in many cases, the content of a perceptual representation depends wholly on the contents of its constituents and the manner in which they are combined. Vision science contains manifold models of how shape, orientation, motion, and other aspects of a scene are "coded in terms of" various features. These models explain the space of possible representations in a system by answering the Three Cs: what are the constituents, how can they combine, and what is the semantic import of combining them that way? Compositionality provides a productive framework for understanding perceptual representation. While I think compositionality is pervasive in perception, that does not mean that it is universal. It is a live question *how many* perceptual capacities are compositional and, for any given capacity, *how* it is compositional.

I close by highlighting three principal conclusions. First, the view that there is no compositionality in perception does not sit well with much of contemporary perception science. Second, compositional representations in perception are markedly different than those in cognition. Structured representations in perception do not simply sit at the interface with cognition, serving to relay messages in a format readable to it. The form in which perceptual content is represented supports perception's own proprietary ends. Finally, I have stressed that the thesis that perception is compositional does not constitute a concrete hypothesis with specific empirical predictions. It instead offers a theoretical framework for posing and answering empirical questions about the nature of representations in a system. While that framework could in principle turn out to be a dead end, it currently supports live and productive programs of research. Fundamental insights arise within that framework, when one asks not just *whether* perception is compositional, but *how*.

Funding Information

This work was supported by funding from an Insight Development Grant from the Social Sciences and Humanities Research Council of Canada and from the Canada First Research Excellence Fund through York University's 'Vision: Science to Applications' program.

Acknowledgments

Thanks to Brandon Ashby, Lance Balthazar, Jake Beck, Gabe Dupre, and Joan Ongchoco for helpful feedback.

References

Ankrum, C., & Palmer, J. (1991). Memory for objects and parts. *Perception & Psychophysics*, *50*(2), 141–156.

Ashby, B. (2020). Rainbow's end: The structure, character, and content of conscious experience. *Mind & Language*. https://doi.org/10.1111/mila.12316

Baker, B., Lansdell, B., & Kording, K. P. (2022). Three aspects of representation in neuroscience. *Trends in Cognitive Sciences*, *26*(11), 942–958. https://doi.org/10.1016/j.tics.2022.08.014

Baker, N., Garrigan, P., & Kellman, P. J. (2020). Constant curvature segments as building blocks of 2D shape representation. *Journal of Experimental Psychology: General*. https://doi.org/10.1037/xge0001007

Baker, N., & Kellman, P. J. (2018). Abstract Shape Representation in Human Visual Perception. *Journal of Experimental Psychology: General*, *147*(9), 1295–1308. https://doi.org/10.1037/xge0000409

Baker, N., & Kellman, P. J. (2021). Constant curvature modeling of abstract shape representation. *PLOS ONE*, *16*(8), e0254719. https://doi.org/10.1371/journal.pone.0254719

Balas, B., Nakano, L., & Rosenholtz, R. (2009). A summary-statistic representation in peripheral vision explains visual crowding. *Journal of Vision*, *9*(12), 13–13. https://doi.org/10.1167/9.12.13

Barenholtz, E., & Tarr, M. J. (2006). Reconsidering the Role of Structure in Vision. *Psychology of Learning and Motivation*, *47*, 157–180. https://doi.org/10.1016/s0079-7421(06)47005-5

Bell, J., Hancock, S., Kingdom, F. A. A., & Peirce, J. W. (2010). Global Shape Processing: Which Parts Form the Whole? *Journal of Vision*, *10*(6), 1–13. https://doi.org/10.1167/10.6.16

Bell, J., Wilkinson, F., Wilson, H. R., Loffler, G., & Badcock, D. R. (2009). Radial Frequency Adaptation Reveals Interacting Contour Shape Channels. *Vision Research*, *49*(18), 2306–2317. https://doi.org/10.1016/j.visres.2009.06.022

Biederman, I., & Cooper, E. E. (1991). Priming contour-deleted images: Evidence for intermediate representations in visual object recognition. *Cognitive Psychology*, *23*(3), 393–419. https://doi.org/10.1016/0010-0285(91)90014-f

Biederman, I., Mezzanotte, R. J., & Rabinowitz, J. C. (1982). Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, *14*(2), 143–177. https://doi.org/10.1016/0010-0285(82)90007-x

Block, N. (1983). Mental Pictures and Cognitive Science. *The Philosophical Review*, *92*(4), 499–541.

Block, N. (2023). *The Border between Seeing and Thinking*. Oxford University Press.

Blum, H. (1973). Biological shape and visual science (part I). *Journal of Theoretical Biology*, *38*(2), 205–287.

Blum, H., & Nagel, R. N. (1978). Shape descriptions using weighted symmetric axis features. *Pattern Recognition*, *10*(3), 167–180.

Boucart, M., Delord, S., & Giersch, A. (1994). The Computation of Contour Information in Complex Objects. *Perception*, *23*(4), 399–409. https://doi.org/10.1068/p230399

Burge, T. (2014). Perception: Where Mind Begins. *Philosophy*, *89*(3), 385–403.

Camp, E. (2007). Thinking with Maps. *Philosophical Perspectives*, *21*(1), 145–182. https://doi.org/10.1111/j.1520-8583.2007.00124.x

Cao, R. (2022). Putting representations to use. *Synthese*, *200*(2). https://doi.org/10.1007/s11229-022-03522-3

Cavanagh, P. (2021). The Language of Vision. *Perception*, *50*(3), 195–215. https://doi.org/10.1177/0301006621991491

De Winter, J., & Wagemans, J. (2006). Segmentation of object outlines into parts: A large-scale integrative study. *Cognition*, *99*(3), 275–325.

Denisova, K., Feldman, J., Su, X., & Singh, M. (2016). Investigating Shape Representation Using Sensitivity to Part- and Axis-Based Transformations. *Vision Research*, *126*, 347–361. https://doi.org/10.1016/j.visres.2015.07.004

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., Gerven, M. A. J. van, Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, *24*(7), 431–450. https://doi.org/10.1038/s41583-023-00705-w

Dretske, F. I. (1981). *Knowledge & the Flow of Information*. The MIT Press. https://doi.org/10.1086/289062

Elder, J. H. (2018). Shape from Contour: Computation and Representation. *Annual Review of Vision Science*, *4*, 423–450.

Elder, J. H., & Goldberg, R. M. (2002). Ecological statistics of Gestalt laws for the perceptual organization of contours. *Journal of Vision*, *2*(4), 324–353. https://doi.org/10:1167/2.4.5

Elder, J. H., Oleskiw, T. D., Yakubovich, A., & Peyré, G. (2013). On Growth and Formlets: Sparse Multi-Scale Coding of Planar Shape. *Image and Vision Computing*, *31*(1), 1–13. https://doi.org/10.1016/j.imavis.2012.11.002

Feldman, J. (1997). Regularity-Based Perceptual Grouping. *Computational Intelligence*, *13*(4), 582–623.

Feldman, J., & Singh, M. (2006). Bayesian Estimation of the Shape Skeleton. *Proceedings of the National Academy of Sciences*, *103*(47), 18014–18019. https://doi.org/10.1073/pnas.0608811103

Feldman, J., Singh, M., Briscoe, E., Froyen, V., Kim, S., & Wilder, J. (2013). An integrated Bayesian approach to shape representation and perceptual organization. In *Shape Perception in Human and Computer Vision* (pp. 55–70). Springer. https://doi.org/10.1007/978-1-4471-5195-1_4

Field, D. J., Hayes, A., & Hess, R. F. (1993). Contour Integration by the Human Visual System: Evidence for a Local "Association Field." *Vision Research*, *33*(2), 173–193. https://doi.org/10.1016/0042-6989(93)90156-q

Fodor, J. (1975). *The Language of Thought*. Harvard University Press.

Garner, W. R. (1974). *The Processing of Information and Structure*. Lawrence Elbaum Associates. https://doi.org/10.4324/9781315802862

Geisler, W. S., Perry, J. S., Super, B. J., & Gallogly, D. P. (2001). Edge co-occurrence in natural images predicts contour grouping performance. *Vision Research*, *41*(6), 711–724. https://doi.org/10.1016/s0042-6989(00)00277-7

Geisler, W. S., & Super, B. J. (2000). Perceptual organization of two-dimensional patterns. *Psychological Review*, *107*(4), 677–708. https://doi.org/10.1037//0033-295x.107.4.677

Gershman, S. J., Tenenbaum, J. B., & J/"akel, F. (2016). Discovering hierarchical motion structure. *Vision Research*, *126*, 232–241.

Graham, P. J. (2014). The Function of Perception. In *Virtue Epistemology Naturalized* (pp. 13–31). Springer International Publishing. https://doi.org/10.1007/978-3-319-04672-3_2

Green, E. J. (2020). The Perception-Cognition Border: A Case for Architectural Division. *The Philosophical Review*, *129*(3), 323–393.

Green, E. J. (2023). A Pluralist Perspective on Shape Constancy. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/727427

Green, E. J., & Quilty-Dunn, J. (2017). What is an Object File? *British Journal for the Philosophy of Science*, *72*(3), 665–699. https://doi.org/10.1093/bjps/axx055

Greene, M. R., & Oliva, A. (2009). Recognition of Natural Scenes From Global Properties: Seeing the Forest Without Representing the Trees. *Cognitive Psychology*, *58*(2), 137–176. https://doi.org/10.1016/j.cogpsych.2008.06.001

Gregory, R. (1970). The Grammar of Vision. *The Listener*, *83*(2134), 242–244.

Griffiths, T. L., Tenenbaum, J. B., & Kemp, C. (2012). Bayesian Inference. In K. J. Holyoak & R. G. Morrison (Eds.), *The Oxford Handbook of Thinking and Reasoning* (pp. 22–35). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199734689.013.0003

Habak, C., Wilkinson, F., Zakher, B., & Wilson, H. R. (2004). Curvature Population Coding for Complex Shapes in Human Vision. *Vision Research*, *44*(24), 2815–2823. https://doi.org/10.1016/j.visres.2004.06.019

Hafri, A., & Firestone, C. (2021). The Perception of Relations. *Trends in Cognitive Sciences*, *25*(6), 475–492. https://doi.org/10.1016/j.tics.2021.01.006

Hafri, A., Green, E. J., & Firestone, C. (2023). *Compositionality in visual perception*. https://doi.org/10.31234/osf.io/trg7q

Helmholtz, H. von. (1925). *Treatise on Physiological Optics* (J. P. C. Southall, Ed.; Vol. 3). The Optical Society of America.

Hoffman, D. D., & Richards, W. A. (1984). Parts of Recognition. *Cognition*, *18*(1), 65–96. https://doi.org/10.1016/0010-0277(84)90022-2

Hummel, J. E. (2013). Object Recognition. In D. Reisberg (Ed.), *The Oxford Handbook of Cognitive Psychology* (pp. 32–45). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195376746.013.0003

Hupkes, D., Dankers, V., Mul, M., & Bruni, E. (2020). Compositionality Decomposed: How do Neural Networks Generalise? *Journal of Artificial Intelligence Research*, *67*, 757–795. https://doi.org/10.1613/jair.1.11674

Janssen, T. M. V. (2011). Compositionality. In J. van Benthem & A. ter Meulen (Eds.), *Handbook of Logic and Language* (2nd ed., pp. 495–553). Elsevier.

Johansson, G. (1973). Visual perception of biological motion and a model for its analysis. *Perception & Psychophysics*, *14*(2), 201–211.

Kahneman, D., Treisman, A., & Gibbs, B. J. (1992). The Reviewing of Object Files: Object-Specific Integration of Information. *Cognitive Psychology*, *24*(2), 175–219. https://doi.org/10.1016/0010-0285(92)90007-o

Keemink, S. W., & van Rossum, M. C. W. (2016). A unified account of tilt illusions, association fields, and contour detection based on elastica. *Vision Research*, *126*, 164–173. https://doi.org/10.1016/j.visres.2015.05.021

Kellman, P. J., & Fuchser, V. (2023). Visual Completion and Intermediate Representations in Object Formation. In A. Mroczko-Wąsowicz & R. Grush (Eds.), *Sensory Individuals: Unimodal and Multimodal Perspectives* (pp. 55–76). Oxford University Press. https://doi.org/10.1093/oso/9780198866305.003.0004

Kellman, P. J., Garrigan, P., & Erlikhman, G. (2013). Challenges in Understanding Visual Shape Perception and Representation: Bridging Subsymbolic and Symbolic Coding. In S. J. Dickinson & Z. Pizlo (Eds.), *Shape Perception in Human and Computer Vision: An Interdisciplinary Perspective* (pp. 249–274). Springer-Verlag. https://doi.org/10.1007/978-1-4471-5195-1_18

Kimia, B. B. (2003). On the role of medial geometry in human vision. *Journal of Physiology-Paris*, *97*(2–3), 155–190.

Knill, D. C., Kersten, D., & Yuille, A. (1996). Introduction: A Bayesian Formulation of Visual Perception. In D. C. Knill & W. Richards (Eds.), *Perception as Bayesian Inference* (pp. 1–21). Cambridge University Press.

Lande, K. J. (2021). Mental structures. *Noûs*, *55*(3), 649–677. https://doi.org/10.1111/nous.12324

Lande, K. J. (2023). Contours of Vision: Towards a Compositional Semantics of Perception. *The British Journal for the Philosophy of Science*. https://doi.org/10.1086/725094

Larson, R., & Segal, G. (1995). *Knowledge of Meaning: An Introduction to Semantic Theory*. MIT Press.

Marr, D. (1982). *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. The MIT Press. https://doi.org/10.7551/mitpress/9780262514620.001.0001

McCloskey, M., Valtonen, J., & Cohen Sherman, J. (2006). Representing Orientation: A Coordinate-System Hypothesis and Evidence from Developmental Deficits. *Cognitive Neuropsychology*, *23*(5), 680–713. https://doi.org/10.1080/02643290500538356

McDowell, J. (1996). *Mind and World*. Harvard University Press.

Nguyenkim, J. D., & DeAngelis, G. C. (2003). Disparity-Based Coding of Three-Dimensional Surface Orientation by Macaque Middle Temporal Neurons. *The Journal of Neuroscience*, *23*(18), 7117–7128. https://doi.org/10.1523/jneurosci.23-18-07117.2003

Palmer, S. E. (1977). Hierarchical structure in perceptual representation. *Cognitive Psychology*, *9*(4), 441–474. https://doi.org/10.1016/0010-0285(77)90016-0

Palmer, S. E. (1978). Fundamental Aspects of Cognitive Representation. In E. Roach & B. B. Lloyd (Eds.), *Cognition and Categorization* (pp. 259–303). Lawrence Elbaum Associates.

Partee, B. (2004). Compositionality. In *Compositionality in Formal Semantics* (pp. 153–181). Blackwell.

Portilla, J., & Simoncelli, E. P. (2000). A Parametric Texture Model Based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision*, *40*(1), 49–70. https://doi.org/10.1023/a:1026553619983

Pylyshyn, Z. (1989). The role of location indexes in spatial perception: A sketch of the FINST spatial-index model. *Cognition*, *32*(1), 65–97. https://doi.org/10.1016/0010-0277(89)90014-0

Quilty-Dunn, J., Porot, N., & Mandelbaum, E. (2022). The Best Game in Town: The Re-Emergence of the Language of Thought Hypothesis Across the Cognitive Sciences. *Behavioral and Brain Sciences*, 1–55. https://doi.org/10.1017/S0140525X22002849

Richards, W., & Hoffman, D. D. (1985). Codon Constraints on Closed 2D Shapes. *Computer Vision, Graphics, and Image Processing*, *31*(3), 265–281. https://doi.org/10.1016/0734-189X(85)90031-3

Rock, I. (1985). *The Logic of Perception*. The MIT Press.

Schmidtmann, G., Jennings, B. J., & Kingdom, F. A. A. (2015). Shape Recognition: Convexities, Concavities and Things In Between. *Scientific Reports*, *5*, 17142. https://doi.org/10.1038/srep17142

Schwartz, O., & Sanchez Giraldo, L. G. (2017). Behavioral and Neural Constraints on Hierarchical Representations. *Journal of Vision*, *17*(3). https://doi.org/10.1167/17.3.13

Schwartz, O., Sejnowski, T. J., & Dayan, P. (2009). Perceptual Organization in the Tilt Illusion. *Journal of Vision*, *9*(4). https://doi.org/10.1167/9.4.19

Shea, N. (2018). *Representation in Cognitive Science*. Oxford University Press. https://doi.org/10.1093/oso/9780198812883.001.0001

Siddiqi, K., & Kimia, B. B. (1996). A shock grammar for recognition. *Proceedings CVPR IEEE Computer Society Conference on Computer Vision and Pattern Recognition*. https://doi.org/10.1109/cvpr.1996.517119

Siddiqi, K., Shokoufandeh, A., Dickinson, S. J., & Zucker, S. W. (1999). Shock Graphs and Shape Matching. *International Journal of Computer Vision*, *35*(1), 13–32. https://doi.org/10.1023/a:1008102926703

Stevens, K. A. (1983). Slant-tilt: The visual encoding of surface orientation. *Biological Cybernetics*, *46*(3), 183–195. https://doi.org/10.1007/BF00336800

Teller, D. Y. (1984). Linking Propositions. *Vision Research*, *24*(10), 1233–1246.

Todd, J. T., & Petrov, A. A. (2022). The many facets of shape. *Journal of Vision*, *22*(1), 1. https://doi.org/10.1167/jov.22.1.1

Treisman, A. (1986). Features and objects in visual processing. *Scientific American*, *255*(5), 106–115.

Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, *12*(1), 97–136. https://doi.org/10.1016/0010-0285(80)90005-5

Vaziri, S., Pasupathy, A., Brincat, S. L., & Connor, C. E. (2009). Structural Representation of Object Shape in the Brain. In S. J. Dickinson, A. Leonardis, B. Schiele, & Mi. J. Tarr (Eds.), *Object Categorization: Computer and Human Vision Perspectives* (pp. 182–195). Cambridge University Press. https://doi.org/10.1017/cbo9780511635465.011

Võ, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, *181*, 10–20. https://doi.org/10.1016/j.visres.2020.11.003

Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.

Zerroug, A., Vaishnav, M., Colin, J., Musslick, S., & Serre, T. (2022). A Benchmark for Compositional Visual Reasoning. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, & A. Oh (Eds.), *Advances in Neural Information Processing Systems* (Vol. 35, pp. 29776–29788). https://proceedings.neurips.cc/paper_files/paper/2022/file/c08ee8fe3d19521f3bfa4102898329fd-Paper-Datasets_and_Benchmarks.pdf

Zhu, S. C., & Mumford, D. (2006). A Stochastic Grammar of Images. *Foundations and Trends in Computer Graphics and Vision*, *2*(4), 259–362. https://doi.org/10.1561/0600000018