Received: 8 December 2022 | Revised: 31 October 2023 | Accepted: 8 November 2023

DOI: 10.1111/mila.12497

ORIGINAL ARTICLE

WILEY

Pictorial syntax

Kevin J. Lande[1,2]

[1]Department of Philosophy, York University, Toronto, Ontario, Canada

[2]Centre for Vision Research, York University, Toronto, Ontario, Canada

Correspondence
Kevin J. Lande, Department of Philosophy, York University, 4700 Keele St, Toronto, ON M3J 1P3, Canada.
Email: lande@yorku.ca

Funding information
Canada First Research Excellence Fund (Vision: Science to Applications); Social Sciences and Humanities Research Council of Canada (Insight Development Grant: "Forms of Mind")

It is commonly assumed that images, whether in the world or in the head, do not have a privileged analysis into constituent parts. They are thought to lack the sort of syntactic structure necessary for representing complex contents and entering into sophisticated patterns of inference. I reject this assumption. "Image grammars" are models in computer vision that articulate systematic principles governing the form and content of images. These models are empirically credible and can be construed as literal grammars for images. Images can have rich syntactic structure, though of a markedly different form than sentences in language.

KEYWORDS
image grammar, images, perception, pictures, semantics, syntax

1 | INTRODUCTION

One can represent a situation by describing it with sentences in language ("Here is a black pyramid beside a grey cube …") or by depicting it with images (an image depicting a black pyramid beside a grey cube). Images encompass both public artifacts of communication—such as realist paintings, photographs, and line drawings—and arguably some psychological representations too, formed in the course of visual perception, memory, and imagining (Block, 1983; Burge, 2022; Kosslyn, 1980). Natural language sentences and logical formulae are paradigms of "discursive representations", which arguably also encompass psychological representations underlying propositional attitudes such as beliefs and desires (Fodor, 1975; Quilty-Dunn et al., 2022). Intuitively, images represent things in a radically different way than sentences.

Some have argued that this difference can help us to understand basic joints in the mind. Tyler Burge, for example, argues that all perception, but not all thought, has a "picture-like, imagistic format" (Burge, 2022, p. 715; see also Block, 2023; cf., Quilty-Dunn, 2020). But it is a substantial project to explain what makes images and sentences fundamentally different and what that difference can tell us about the nature of representation, depiction, and mind.

Historically, images have often been taken to lack the sort of structure necessary for expressing complex contents about the world and entering into sophisticated patterns of inference. In its modern form, this view holds that images, unlike sentences, lack anything like a substantive "grammar" or "syntax" (I use these terms interchangeably). For example, Flint Schier writes "pictures ... have no grammatical rules, natural or conventional" (Schier, 1986, p. 66); the psychologist Stephen Kosslyn notes that "Images do not seem to have a syntax" (Kosslyn, 1980, p. 32); and Jerry Fodor states, "it is having a canonical decomposition that distinguishes discursive representations from iconic ones ... a representation that has no canonical decomposition is an icon" (Fodor, 2007, p. 108). While Elisabeth Camp has argued that non-linguistic maps and diagrams can have a formal syntax and semantics, she joins the others in doubting that "a similarly formal semantics can be offered for pictures, let alone perception" (Camp, 2018, p. 42). Some take it to be constitutive of pictures that they lack syntactic structure; others merely doubt that images have syntax.[1]

That pictures lack syntax does not mean that they are simple. Unlike the words that make up sentences, pictures have multiple meaningful parts. But whereas complex phrases and sentences are composed in specific ways from a set of basic words, the common view is that there is no privileged, non-trivial way to analyze a picture into its parts. Images are complex and can be carved up in many ways according to one's purposes, but they are essentially unstructured. To be sure, popular works will sometimes mention the "grammar" or "syntax" of images, artworks, and visual design. But these terms are nearly always metaphorical; constantly embedded in scare-quotes, they refer more to heuristics for designing or describing images—a sort of style guide—than to fundamental, systematic principles governing the form and content of those images.

Against this received view, I will argue that images, whether in the world or in the mind, can have genuine syntactic structure. What would pictorial syntax look like? One strategy for answering this question would be to look at models of visual processing and argue both that the representations posited in these models are image-like and that they have syntactic structure (see, e.g., Burge, 2022; Clarke, 2022, 2023; Lande, 2023). I will pursue a complementary strategy in this paper. I will consider certain models in cognitive science and computer vision, called "image grammars" (see Zhu & Mumford, 2007), as a means to spelling out what it would be, in principle, for pictures to have syntax.

Image grammars make explicit use of grammatical formalisms in order to model image comprehension. In some cases, the models are intended to explain how observers parse and interpret artifactual pictures; in other cases, they are intended to explain how scenes are represented in human vision. I will argue that these models can be interpreted literally as offering grammars for images. They do not merely provide heuristics for using or describing images;

[1]There is little uniformity in terminology. Kulvicki (2014) uses the term "image" in a broader sense than I do, one which encompasses diagrams and maps, while reserving the term "picture" for a narrower class that includes photographs, realist paintings, and figurative line drawings. Fodor treats "icon" and "picture" as synonyms, while Camp views iconicity as a spectrum. I treat "image" and "picture" as synonyms and take these terms to include both public representations such as photographs, realist paintings, and figurative line drawings, as well as imagistic psychological representations. I reserve the term "icon" for a broader category that also includes diagrams and maps.

they have the potential to uncover core, non-trivial principles concerning the form and content of images, much as syntactic hypotheses in linguistics purport to do for language. Image grammars have the potential to explain the sorts of things that we expect grammars to explain, using the sorts of elements we expect grammars to contain.

My conclusion here is modest: Pictures can have syntactic structure. Image grammars provide instructive models of what that syntax could look like. It is therefore not constitutive of images that they lack syntax. Skepticism about pictorial syntax ought to be grounded in an empirical evaluation of structural theories of images, of which image grammars are an example. My argument will not rest on defending any particular image grammar, and I will neither argue that pictures necessarily have syntactic structure nor that every picture in fact has syntactic structure. Importantly, if images do have grammars of some sort, it does not follow that pictures are discursive. Pictures may be governed by substantially different kinds of syntactic rules and constraints than paradigm linguistic representations. An important upshot is that a representation's possessing a compositional syntax and semantics is consistent with its being distinctively imagistic in form. The fact that a perceptual representation, for example, is structured does not preclude it from being imagistic (cf., Fodor, 2007).

In Section 2, I survey the program of research into image grammars and I outline a toy image grammar as illustration. In Section 3, I discuss the core explanatory roles of syntactic theories and argue that image grammars can function to fulfill those roles in the domain of images. In Section 4, I respond to a potential objection that image grammars are not really grammars *of images*, but rather they are models for translating or describing images. In Section 5, I respond to a potential objection that image grammars are not really *grammars* since they do not (or cannot) fulfill the main roles of syntactic theory after all. Throughout, I intend the discussion to extend to both artifactual images as well as psychological ones.[2]

2 | IMAGE GRAMMARS

A theory does not have to be labeled explicitly as a "grammar" in order to offer an account of how representations are syntactically structured. Many models in vision science ascribe a compositional syntax and semantics to perceptual representations, often without explicitly invoking terms such as "grammar" and "syntax" (see Lande, 2021, 2023). Nevertheless, the overt use of

[2]There have been some previous arguments that images and the like can have something like a grammar. Some focus primarily on artifactual images. Ben Blumson (2014) suggests that if it is plausible that maps and chess diagrams have a compositional grammar, then it should be plausible that pictures do too. However, he does not directly tackle the question of what a pictorial grammar would look like. Moreover, the generalization from maps and diagrams to pictures is just what Camp expresses doubt about in the quote above. Recent work on the formal semantics of pictures tends to treat them as syntactically unstructured, though some, like Gabe Greenberg (2021), leave open the possibility of a more robust role for pictorial syntax. John Willats (1997; 2005) draws from early work on image grammars in order to characterize pictorial systems and styles, but he does not make a case that image grammars should be interpreted literally as providing syntactic hypotheses. Another set of authors have focused primarily on iconic psychological representations. Burge (2018, 2022) discusses artifactual images in the course of arguing that perceptual representations are imagistic and have semantic structure. In a similar vein, I have argued that perceptual representations have a compositional syntax and semantics and that they have hallmark features of iconicity (Lande, 2021, 2023), while Sam Clarke has argued that analog mental representations have privileged analyses into constituents (Clarke, 2022, 2023). The aim of the present discussion is to provide a philosophical justification for treating image grammar models as literal grammars of images and to explore thereby what it means, in the first place, to say that an imagistic representation is syntactically structured.

grammatical notions by image grammars makes them ideal devices for understanding what it means to say that images do or do not have syntax.

Image grammars first appeared in the 1960s and 1970s as part of an approach to image recognition in machines. Inspired by work on generative grammars for natural languages, King-Sun Fu introduced his monograph, *Syntactic methods in pattern recognition* (Fu, 1974), by describing the importance for pattern recognition of having "a capability for describing a large set of complex patterns by using small sets of simple pattern primitives and of grammatical rules" (p. 3). For example, Fu proposed that the recognition of handwritten Chinese characters could be accomplished by specifying a set of primitives (stroke segments) and rules for assembling sub-patterns from those primitives, along with algorithms for matching subsets of pixels to those primitives and for checking the application of those rules. While many researchers had noted the importance of encoding relations among pattern elements in order to perform recognition tasks, the syntactic approach explicitly drew from the concepts and tools of formal language theory to encode those relationships via combinatorial operations over primitive elements. I will provide a brief introduction to the theoretical aims of image grammars before offering a toy example of one.

2.1 | Pictorial competences and pictorial systems

Much of the contemporary research on image grammars pursues engineering solutions to practical problems in computer vision, such as image recognition. Nevertheless, a longstanding theoretical goal of this research program has been to model and explain how human observers, and their visual systems, group and interpret images. Max Clowes, a formative figure in the program, aspired to provide an account of the "structural description of pictorial objects" that would reflect our intuitions about the organization of pictures and our ability to extract core content about scenes from certain classes of pictures, such as line drawings (Clowes, 1967). For Clowes, the goal was to model and explain our "pictorial competences" (Schier, 1986; Sober, 1976; Willats, 1997).

To have a pictorial competence is, in the first place, to have productive and systematic capacities to treat images and their parts as clustering into image classes. One can intuitively classify whether an image is a natural image, block drawing, stick figure drawing, architectural sketch, and so on (Figure 1). The space of possible images is not uniformly distributed. The kinds of elements and configurations that line drawings tend to contain are different from the kinds that natural photographs tend to contain, even when they depict the same type of scene. Lines are more likely to terminate in junctions in a block drawing system than in a stick figure system.[3]

An ability to group images into an image class is "productive" insofar as one can normally, other things equal, determine whether arbitrary images can belong to that class (whether or not one has the skill to produce the image oneself). To say that one's ability to group images is "systematic" is to say that if one were able to classify an image as belonging to a class then normally one would also be able to classify other images as belonging to that class (Schier, 1986). A core aim of image grammars is to provide generative models that "identify, map out, and catalog high-density clusters" in image space in a way that reflects our productive and systematic abilities with those clusters (Zhu & Wu, 2023, p. 9).

[3]I take it to be a live empirical question what the different image systems will turn out to be. I leave it open whether distinctions among pictorial "styles" of artworks (e.g., cubism vs. pointillism vs. realism) are a matter of differences in image systems and their grammars (Willats & Durand, 2005).

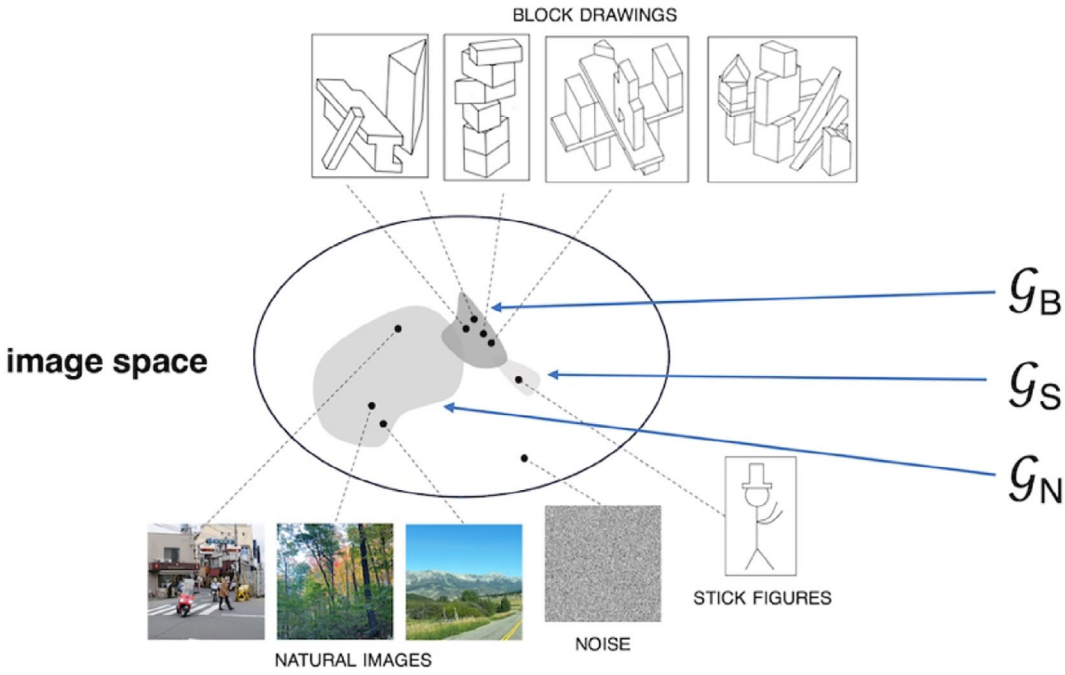

FIGURE 1 The space of possible images is not uniformly distributed. Some images cluster into classes (as rough examples: natural images, stick figures, and block drawings [adapted from Guzmán, 1968, with kind permission from the Association for Computing Machinery]) that observers can productively and systematically identify and comprehend. Image grammars ($\mathcal{G}_B$, $\mathcal{G}_S$, $\mathcal{G}_N$) provide generative models of these image classes, which potentially reflect one's competences with those classes. (Based on Zhu & Wu, 2023, pp. 10, 21).

Pictorial competences also consist in abilities to comprehend any given image within a system as depicting a certain type of scene. It is helpful to distinguish, at least in the case of artifactual pictures, between the core semantic content of an image and more pragmatic aspects of its interpretation. An image's semantic content depends on systematic principles and biases that govern the relevant image class. Pragmatic interpretation is relatively unsystematic and depends on adventitious features of background context, communicative intentions, task demands, or examination of extrinsic representational features such as captions (Abusch, 2020; Fan et al., 2019). For example, observers can normally comprehend a realist image or photograph as depicting a more or less determinate layout of surfaces of different lightnesses, colors, shapes, and orientations in depth (Sober, 1976). Determining that the image is of an elementary school student, and that the student is Julian rather than his identical twin Lucas, would require additional contextual information. Clowes (1971) argued that understanding pictorial competence is integral to understanding our broader performance with pictures: the set of processes involved in producing, parsing, and using images in the full context of communication.

There is no commitment here to a universal set of principles governing all images or to the existence of a modular "picture faculty". Competences for public images might well consist in combinations of visual capacities for performing "inverse optics" (Greenberg, 2021) and for perceptual organization together with learned conventions and biases specific to a given pictorial system. Image grammars might be innate, learned, universal, heterogeneous, or some combination of these.

I have so far been discussing the pictorial competences of individuals; but a closely related matter is understanding the pictorial capacities of an organism's visual system. The history of image grammar research is closely married to research on structured representations in biological vision. The main input to our visual system is an array of light (the "retinal image") that stimulates photoreceptors laid out across the curved surface of the retina. This array of light is the product of a variety of distal factors including the orientation, shape, size, distance, color, and illumination of the surfaces in the scene. By many accounts, the visual system proceeds to form a variety of separate (though interacting) "intrinsic images" dedicated to representing the color, illumination, texture, outline shapes, and so on of distal items (e.g., Anderson & Winawer, 2008; Barrow & Tenenbaum, 1978; Zucker, 2014). Information within and across these images is organized to compose representations of scenes, objects, and surfaces with multiple features. One can think of the visual system as embodying competences for different kinds of intrinsic images.

A core theoretical objective of image grammars, then, has been to specify the primitives and combinatorial rules that govern a given type of psychological image in biological vision as well as the rules for combining information within and across such images (Yuille & Kersten, 2006; Zhu et al., 2011). Fittingly, the suggestion that vision has a grammar has been echoed repeatedly in vision science (Biederman, 1987; Cavanagh, 2021; Gregory, 1970; Võ, 2021), though not always accompanied by the explicit use of formal language theory. Of course, there are important differences between artifactual images and psychological images. Artifactual images are endowed with content by the people who make and use them; psychological images have their contents intrinsically. But in either case, we can ask about the form of the representation and how that form constrains the content of the representation.

2.2 | Aspects of image grammars: An example

I now offer a toy illustration of an image grammar, focusing on simplicity and vividness over formal details. The goal is to fix our focus and show in rough outline what an image grammar might look like, not to articulate a fully working model.[4]

Consider a system, B, of polygonal line drawings that depict scenes containing block-objects with flat surfaces (Figure 1). There are indefinitely many images that intuitively fall under this class. Observers can productively make out the shapes of the objects depicted in images of this sort as well as how most of the surfaces and objects are ordered in depth. An image grammar, $\mathcal{G}_B$, for such a class of images will specify a set of basic image elements together with rules or constraints on how image elements can combine to form more complex configurations.[5]

[4]I draw here on work from Guzmán (1968), Huffman (1971), Waltz (1975), Mackworth (1976), and especially Clowes (1971). Zhu and Mumford (2007) provide a valuable overview of the history and contemporary directions of image grammar research.

[5]While different formal frameworks are used, an image grammar is standardly modeled as a 4-tuple $\mathcal{G} = (V_N, V_T, R, S)$. V_N is a finite set of "non-terminal nodes" (corresponding to types of complex image patterns). V_T is a set of terminal nodes (corresponding to basic image elements). $S \in V_N$ is the root node (corresponding to a whole image). R is a set of production rules ($R = \{\gamma : \alpha \longrightarrow \beta\}$), which specify how non-terminal nodes α decompose into child nodes, β. In a stochastic or probabilistic grammar, these production rules are weighted or associated with probabilities—that is, there is a probability that a node decomposes in one way rather than another. Nodes may have "attributes" (corresponding, say, to spatial coordinates or to surface properties such as hue-saturation-brightness) and production rules may place constraints on the attributes of the child nodes (e.g., that their spatial coordinates be related in a certain way).

A basic condition on the adequacy of such an image grammar is that it span approximately all and only the images that one would count as members of the target image system.

The syntax, $\mathcal{G}_B$, is dedicated to generating the set of images within the target class B, while a semantics, $\mathcal{I}_B$, determines (or constrains) the type of scene that an image in that class can accurately depict. An image system is *semantically compositional* just in case the core scene type that an image depicts, in that system, is solely a function of what the basic image elements represent and how they are syntactically combined to form the image (see Lande, 2023). In a compositional system, syntax plays an integral role in determining the contents of representations. All image grammars of which I am aware are meant to be semantically compositional, though they do not always come with an explicit semantic theory.

Figure 2 illustrates a syntactic analysis that $\mathcal{G}_B$ might assign to an image, together with one possible assignment of image elements to types of scene elements, given $\mathcal{I}_B$. The basic image elements of $\mathcal{G}_B$ are straight line segments, which are characterized by values along dimensions such as size, orientation, and position in the picture (bottom row of Figure 2). In the semantic domain, the basic types of scene elements are straight edges at which flat surfaces intersect. Edges are organized into three types: edges formed at convex intersections between two surfaces (abbreviated "+"); edges formed at concave intersections between surfaces ("−"); and convex edges formed when one surface occludes another (">," where the occluded surface is to the left of the visible surface in the direction of the arrow). The semantics therefore distinguishes between types of scenes that differ in the topology and depth relations of edges and surfaces.

Line elements in the image can combine into at least four types of junction—fork-junctions, arrow-junctions, L-junctions, and T-junctions—as a function of the number, locations, and relative orientations of those elements (Figure 3). Just as not all combinations of words can yield a well-formed phrase, not all intersections of lines yield admissible junctions in the system. The system does not, for example, recognize "¥-junctions". However, it may be that additional junction-types, such as "X-junctions", do prove to be necessary for an empirically adequate grammar (Guzmán, 1968). Junctions and other compound elements have both syntactic and semantic attributes such as location, orientation, and size in the image or in the depicted scene, which are inherited or synthesized, according to certain rules, from the attributes of their constituents (see, e.g., Han & Zhu, 2009; Liu et al., 2018).

As Figure 3 illustrates, fork, arrow, and L-junctions can each accurately represent several corresponding types of corners. These happen to be the corners that are geometrically possible for three-dimensional block-objects. T-junctions represent cases in which one surface or body occludes another surface or body behind it.

Ascending the hierarchy, an image "region" is formed by a sequence of junctions just in case some of the line segments that compose those junctions form a closed cycle. Image regions function to depict the visible surface of a body in a scene. A "region set" is formed from a set of adjacent regions that share junctions, and functions to depict sets of adjacent surfaces that belong to the same material body ("body sets"). A "frame" is a set of region sets, and functions to depict a scene type, or set of body sets arranged in space.

Several features of this system are worth noting. First, as Clowes emphasized, the syntax and semantics of the system have distinct explanatory roles. The syntax purports to explain which images and image-parts are possible in the image system. The semantics purports to explain what those images and their parts can accurately depict. Syntactic well-formedness can come apart from semantic interpretability, as can be seen by considering some additional rules of the system. As a syntactic rule, a frame can only contain line segments that belong to exactly two junctions, one at each endpoint. Call this the "line rule". The configuration of lines in

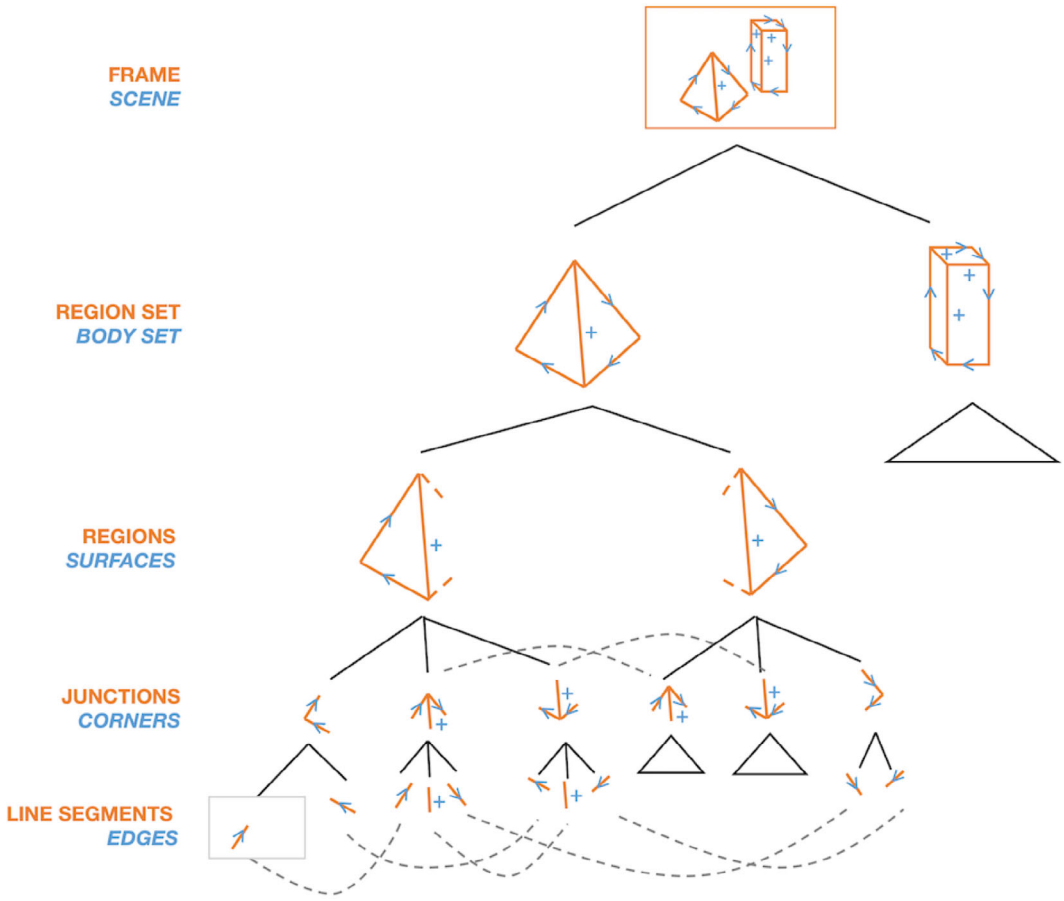

FIGURE 2 An abridged decomposition of an image, according to $\mathcal{G}_{\mathrm{B}}$, and a sample assignment of syntactic elements (orange lines) to types of scene elements (blue symbols) given a semantics, $\mathcal{I}_{\mathrm{B}}$. "+" indicates a convex edge type, "−" a concave edge type, and ">" a type of edge that occludes what is on the left in the direction of the arrow. Each node has syntactic and semantic attributes such as position and orientation (illustrated by the left-most leaf; omitted elsewhere). Dashed arcs indicate the occurrence of the same token element within different constituents. Triangles are used on branches with abbreviated analyses.

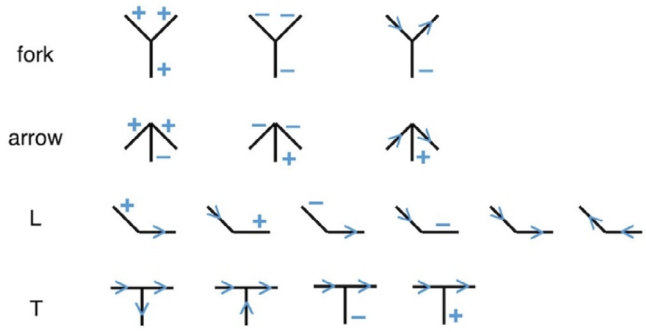

FIGURE 3 For each type of junction admitted by the grammar $\mathcal{G}_{\mathrm{B}}$, the semantics $\mathcal{I}_{\mathrm{B}}$ specifies the types of corners or occlusions that a junction of that type can accurately depict. (Based on Huffman, 1971).

Figure 4a is ill-formed with respect to $\mathcal{G}_B$. It neither satisfies the line rule nor corresponds to a recognized junction-type. This conforms to the intuition that even if Figure 4a belongs to some cohesive image class, it is not in the target class of images, B.[6]

Now consider what Clowes called the "co-occurrence rule". This is a semantic rule dictating that any given line can depict at most one type of edge (concave, convex, or occluding). By this rule, patterns such as the "blivet" in Figure 4b do not depict any scene type within the semantic domain, because in order to compose the relevant junctions the very same line segment would have to depict one and the same edge as both occluding and convex. Figure 4b is syntactically well-formed according to $\mathcal{G}_B$, but it does not have a semantic value according to $\mathcal{I}_B$.

Second, while the rules of $\mathcal{G}_B$ and $\mathcal{I}_B$ provide a backbone for parsing and interpreting images in B, they do not completely determine how a given image in B will be parsed and interpreted. Like utterances, images may be considered members of multiple systems. As with language, even within one system an image may admit of several syntactic analyses. And while a given syntactic analysis will significantly constrain the types of scenes that an image could accurately represent, there will also be some degree of nonspecificity about which scene type is represented. Additional background and context can play a role in determining which grammar should be operative in interpreting an image, which admissible structure to assign to the image given a particular grammar, and what specific messages one might take away from the image.

Third, the system assigns hierarchical structures to images. Images are composed from intermediate constituents (such as regions and junctions), which are eventually decomposed into basic elements (such as lines). Some models treat images as having "flat" structures (Zhu et al., 2011, pp. 124–126; see also Camp, 2018), so that each image is decomposed immediately into an array of primitive line segments, pixels, or what have you, without intermediate structures such as junctions and regions. Flat structure should not be confused for a lack of structure. A flat grammar will provide a privileged analysis of an image into its constituent primitives while excluding analyses that posit intermediate constituents. If images truly lack structure, then nothing would privilege an analysis that precludes intermediate constituents over one that includes them.

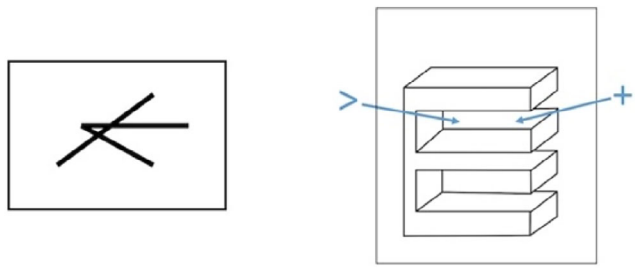

FIGURE 4 (a) A picture that violates the line rule of $\mathcal{G}_B$. (b) A drawing of an "impossible figure".

[6]Alternatively, in order to consider Figure 4a as a member of system B, one would have to treat it as containing unmarked lines connecting the open endpoints while treating the cross-cutting line as being an "incidental" non-syntactic mark (see Kulvicki, 2014, p. 93).

Finally, the properties of the image grammar are quite different from the properties typically found in grammars for natural languages. While linguistic structures have been argued to be binary branching, images are not assumed to be binary branching. While distinct linguistic phrases cannot share the very same constituent, image elements might be shared by multiple constituents at a time (Zhu & Mumford, 2007, pp. 286–287; Geman et al., 2002, p. 716). In language, the way a word combines with others depends on its categorical syntactic features such as person, number, gender, case, and tense, which govern agreement with other items in the structure. Image elements have values (attributes) on various dimensions (such as position, orientation, and color), which partly determine how those elements can combine into types of compound configurations such as junctions. Image elements are analog in the sense that their attributes lie on dimensions that "mirror" the dimensions of the depicted scene elements (much like the elements of the "analog maps" discussed by Clarke, 2022).[7]

2.3 | Advances in image grammars

This toy model has many limitations, as do the more sophisticated models on which it is based. But the limitations of these models are not endemic to image grammars in general.

Image grammars are not inherently limited to abstract, stylized, or conventionalized image systems. While $\mathcal{G}_B$ is limited to a certain class of highly simplified line drawings, most contemporary image grammars deal with complex natural images, as in everyday photographs. Some of these models analyze the outline shapes present in photographs while abstracting away from "surface properties" such as color and texture, others model the surface properties themselves, while yet others attempt to model both the geometry and surface properties of the image and depicted scene (Bear et al., 2020; Liu et al., 2018; Zhu & Mumford, 2007).

The primitives and rules of $\mathcal{G}_B$ are specified "by hand" and the enumeration of permissible primitives, junction types, and so on, is based on intuition. Contemporary image grammar research can harness efficient coding principles to specify the primitive elements of an image system, which may include small line segments, textured image regions, or other configurations of pixels.

In general, the rules for how images in the target system can and cannot be put together are not always obvious and their discovery can require theoretical ingenuity and empirical confirmation. More recent work builds on research on perceptual organization (Yuille & Kersten, 2006) and employs statistical "structure learning" techniques to discover the combinatorial principles of a system (Tenenbaum et al., 2011; Zhu et al., 2011). These models have even been extended in attempts to handle sequences of images, such as video clips (Astolfi et al., 2021; Zhu & Huang, 2021). While early image grammars had more or less deterministic syntactic rules, almost all contemporary image grammars work with probability distributions over possible parses of an image.

Lastly, while our toy grammar delivers content only about the topology and depth-ordering of surfaces, more sophisticated accounts aim to deliver more specific metric content about the shape, position, and surface properties of items in a scene (e.g., Liu et al., 2018).

[7]The "dictionary" of basic image elements might be either discrete or "dense". If dense, then for any two elements with different values on a given syntactic dimension, there is a third type of element with an intermediate value (Goodman, 1968). Geman et al. (2002) provide a general framework for describing compositional structures over a dense dictionary of elements.

3 | SYNTACTIC EXPLANATION

Why insist on calling these models *grammars*? On one reading, "grammar" is a natural kind term that refers to the type of syntax that characterizes natural languages. As we have seen, properties that commonly characterize natural language grammars may well not characterize image grammars. Still, there is a more basic notion of "grammar" or "syntax", of which natural language grammar is just one species, which applies also to invented formal languages of logic as well as diagrams and maps (Camp, 2007; Shin, 1994). I take it that those who are skeptical that images can have a grammar are not merely skeptical that images can have a specific type of grammar characteristic of natural language; they are skeptical that images can have any kind of grammar whatsoever. Instead of attempting to define "grammar" or "syntax", in this broader sense, I will suggest that image grammars can explain the sorts of things we expect a grammar to explain using the sorts of elements we expect grammars to contain.

What does syntactic structure explain? Distinguish, first, between "syntactic features (or parts)" and "syntactic structure". Syntactic features (or parts) are features (or parts) of the representation that make a difference for the semantic content of the representation (Goodman, 1968; Kulvicki, 2014)—the lightness of a region in a black and white photo or the presence of a closed circular contour, for example, but not the weight and gloss of the paper on which the photo is printed. The syntactic structure of a representation is a higher-order syntactic feature, which concerns how semantically significant features and parts of the representation are related to each other and to those of other representations in a system. A representation's syntactic structure, if it has one, indicates how that representation is composed from basic features and parts according to determinate principles of combination. The basic and compound parts and features that enter into the composition of the image are "constituents" of that image.

Syntactic structure is never assigned to a representation in isolation; a grammar assigns structures to a whole system of representations. A fundamental role of a grammar is to explain the "distribution" of representations in the relevant system: which representations are possible in the system and how they can, cannot, or must co-vary (Lande, 2021). Where semantics concerns the meanings or contents of representations and processing theories concern the transitions from one representation to another, the syntactician's prime question is: Which representations are possible within the system in the first place? The syntactic analysis of a representation is a proof of its belonging to the system, explaining how that representation is a possible member of that system. A representation belongs in the system only if its syntactic features are organized in the right way. In accordance with this explanatory role of grammars, image grammars purport to explain which images do and do not belong to the target system of images and how the features of images in that system can and cannot co-vary by specifying the primitives and rules of combination that characterize the image system. For example, Figure 4a does not belong to system B because it fails to satisfy the line rule.

On top of explaining the distribution of images in a class, we expect a grammar to contribute to explaining the semantics of the system. Disambiguating the structure of a representation should contribute to disambiguating its meaning. In the previous section, we saw that the type of corner that $\mathcal{I}_B$ assigns to a junction depends on the type of junction that those lines form. So, if an image is taken to be a construction of certain kinds of junction types rather than others, this will have ramifications for what kinds of scenes the image can be understood as accurately depicting.

As Clowes (1971) emphasized, a grammar is not by itself a model of the exact order of operations by which an image is produced, parsed, and interpreted. However, we do expect a

grammar to help explain how the representations in the systems are formed and used. Image grammars normally are developed as components within computational models for classifying, transforming, and generating images.

In summary, grammars standardly function to fill three explanatory roles (see also Larson & Segal, 1995, pp. 67–76): (1) to explain the distribution of representations in a system, (2) to help explain the semantics of those representations, and (3) to support models of the production and consumption of those representations. Image grammars have the potential to jointly fulfill these explanatory roles by appeal to the sorts of elements that normally make up a grammar (e.g., primitives and combinatorial constraints). Image grammars could therefore offer literal grammars of images. A successful image grammar would articulate a picture's syntactic structure.

4 | ARE IMAGE GRAMMARS *GRAMMARS OF IMAGES*?

There are various ways in which one might deny that image grammars bear on the question of whether images have grammar. One natural thought is that image grammars do not characterize the syntax of images themselves. Rather they are grammars for translations, descriptions, or ways of labeling images.

To be sure, some image grammars have merely descriptive or practical purposes. However, in principle they have the potential to explain systemic features of how images are placed (distributionally, semantically, and inferentially) within broader systems of images. Conceived as explanatory hypotheses, image grammars function not just to describe an image or set of images (considered "in extension"), but also to explain generative abilities to group and grasp images in a set (considered "in intension"). If an image grammar succeeds in explaining these systemic features of our competence with images, its analysis of an image into different types of parts under different types of structural relations is not mere labeling, but rather a limning of the image's underlying structure as a member of a representational system.

Still, one might doubt that the structure assigned by an image grammar is a feature of the image itself. One way to press this doubt is to note that the paradigmatic content-carrying features of images (line segments, colored regions, and so on) are visible, rendered by intrinsic marks in an image. (It is not clear how to extend this point to psychological images, but we can set this worry aside for the moment.) By contrast, while line segments or junctions may be visible, without explicit labels there are no marks specifically dedicated to indicating that an element is a primitive, or a compound, or that it satisfies the line rule. Adding such marks to an image would merely be to annotate the image with symbols.

The syntax of sentences is not, strictly speaking, audible; so, why must the syntactic structure of images be visible? Perhaps one motivation is the intuition that images are distinctively visual. One condition on the "visual" character of artifactual images might be that every syntactic feature must correspond to some visible feature. This condition can easily be met by an image grammar for which every syntactic structure terminates in visible elements (as in Figure 2). Image grammars are intimately tied with the visual character of images. The parsing of an image purports to determine which of the image's visual features are semantically significant (rather than merely incidental) and how the image groups with other images on the basis of those features. But the core intuition does not force the implausible view that each and every node and relation in an image's syntactic structure must receive its own distinct visual mark.

A representation's structure—whether an image or a sentence—is a matter of how it functions within a wider system (Block, 1983, p. 513; Burge, 2018). Syntactic structure functions to explain how a representation is distributionally, semantically, and computationally related to other representations in the system. An image grammar functions to characterize the syntax of an image—and not merely to describe or annotate the image—insofar as it purports to explain these functional and systemic features of the image.

5 | ARE IMAGE GRAMMARS *GRAMMARS*?

I have suggested that image grammars are capable of providing substantive explanations of the distribution, semantics, and use of images. I will now consider some potential worries about an image grammar's ability to perform each one of these functions. More globally, the worry is that image grammars are either redundant or else that they do not do the work associated with literal grammars.

5.1 | Distribution: Well-formedness and systematicity

A grammar explains which representations (and combinations thereof) are well-formed—that is, which are possible elements of the representational system. But many have doubted that there is a substantive distinction to be drawn between "ill-formed" and "well-formed" pictures. Hence, John Haugeland says of photographs that "every ... shape is allowed—nothing is ill-formed" (Haugeland, 1981, p. 220). There may be distinctions in the neighborhood, but they do not require syntactic explanation.

For example, in response to the question of what would count as an "ill-formed picture", Schier writes, "[t]he only thing I can think of is that a symbol is ill-formed, from the point of view of iconicity [pictoriality], if it isn't really an icon [picture] at all" (Schier, 1986, p. 66). To be an ill-formed picture, on this account, is to not be a picture. For Schier, the core difference between a picture and a non-picture is that the former, but not the latter, trigger appropriate perceptual recognitional capacities in observers under certain conditions. If such perceptual-recognitional dispositions are the hallmark of pictoriality, he writes, then "there is no place for a grammar or syntax of pictures" (Schier, 1986, p. 65). There is little need to appeal to rules of pictorial syntax to explain why a picture cannot be red and blue all over (that is just a law of nature or metaphysics), or why objects such as chairs, universities, and short stories do not trigger the relevant recognitional capacities in the relevant ways. By contrast, even if "impossible pictures" (e.g., Figure 4b) are befuddling, they are pictures nonetheless—well-formed, though perhaps semantically deviant.

But it is a mistake to equate pictorial ill-formedness with non-pictoriality. Constraints on well-formedness are aspects of one's pictorial competence and, as Schier himself pointed out, "pictorial competence is, in effect, system-relative" (Schier, 1986, p. 47). It is a plausible empirical conjecture that some images systematically cluster with other images. Within a given system, elements co-occur in systematic ways (e.g., in system B, lines must terminate at junctions). Image grammars aim to identify and explain these patterns by providing constraints on being a well-formed image of a given system. If an image is ill-formed with respect to an image system, that simply means that it does not belong to that particular system. A stick figure drawing will

be ill-formed with respect to the block-drawing system B; but it may be a well-formed member of a different system.

Perhaps there is a non-syntactic explanation for the uneven distribution of images in image space. For example, perhaps images of similar scenes tend to cluster together. On the other hand, the same scene could be represented by images belonging to intuitively different systems. A more nuanced approach would be to distinguish pictorial systems by the types of features that images in those systems tend to possess, the kinds of elements those features tend to designate, and perhaps the geometry that relates the arrangement of those features in the image to an arrangement of elements in the scene (Greenberg, 2021; Willats, 1997; Willats & Durand, 2005). For example, one can explain the difference between sepia-toned pictures and color pictures simply by noting that color pictures, unlike sepia ones, employ color to depict colors in the scene.

However, this sort of approach fails to explain the following observation. Arbitrary combinations of marks from a given system are not guaranteed to yield an image belonging to that same system: Image systems are not closed under arbitrary recombination. For example, the line segments that are basic marks in the system B can be arranged to form Figure 4a, which does not intuitively belong to B. Adjoining two natural images often does not yield something that we would register as a single natural image, as opposed to a collage. Image grammars explain these transgressions in terms of combinatorial constraints on image structures.

One might resist the observation above, arguing that in fact image systems are closed under arbitrary recombination. Adjoining two or more natural images will normally not yield something that we would take to be a natural image—colors change suddenly at the border, lines and regions are cut off, odd junctions are produced—though we may take it to be a valid collage. But with enough imagination, one can envision a sort of scene for which the juxtaposition would be an accurate image. (Searching the Internet for the phrase "this is a single picture" reveals many examples of what appear to be collages of separate images, but which in fact are photographs of cleverly composed scenes.)

Consider the most extreme case: an image that is just random white noise (Figure 1). We may have pragmatic reasons not to treat the highly atypical noise image as a natural image. Yet one could tell a story about how to interpret the image as a photograph of a blizzard, say. If the pattern of noise is in principle interpretable as depicting a scene type, it would be overly restrictive to count that pattern as ill-formed. And if even a random distribution of pixels is a well-formed natural image, then it is unlikely that there are any substantive constraints on which combinations of pixels are admissible natural images. A similar argument could be given for just about any other system.

But just as the uninterpretability of impossible pictures does not entail their ill-formedness, so too the interpretability of some pictures does not entail their well-formedness. Even ill-formed representations can be pragmatically interpreted in context, given additional reflection, imagination, and background information. Is the noise image well-formed though unexpected, or ill-formed though pragmatically interpretable? This is a live empirical question. Our intuitions about what is well-formed or not are evidential, but they are not demonstrative. Syntactic well-formedness is not a theory-neutral designation. One must develop tests to distinguish these possibilities and one must consider the overall explanatory power of a theory that designates the image as well-formed or ill-formed. In short, one cannot dismiss the explanatory value of pictorial syntax a priori.

5.2 | Semantics: Disambiguating pictures

Even if there were no such thing as an ill-formed picture, pictorial syntax might nevertheless play an important role in explaining the contents of pictures. Many are skeptical, however. "Carve the picture up however you like", or "take any jigsaw puzzle you could create from the picture", the thinking goes, it makes no difference to the picture's content. The claim, in short, is that disambiguating a picture's structure does not disambiguate its meaning. I will consider three related arguments that pictorial syntax is semantically idle.

5.2.1 | The picture principle

The claim that pictures lack syntactic structure is often taken to be an implication of the "picture principle", which characterizes how the contents of images are semantically related to the contents of their parts. Many take the picture principle to imply that all ways of "carving up an image"—all possible syntactic analyses of an image—are alike in identifying parts of the image from which the whole image's content can be composed. Images do not have semantic joints. Here is one formulation of this principle:

> **Picture principle:** If P is a picture of X, then all the parts of P are pictures of parts of X (Fodor, 2007, p. 108).[8]

For example, in a line drawing (Figure 2) the line segments are parts of junctions. Those lines depict edges which are parts of the corners that the junctions depict. By contrast, "The 16th President of the United States" refers to Abraham Lincoln, but "the United States" does not refer to something that is part of Abraham Lincoln (Block, 1983, p. 513).

The picture principle does not immediately entail that pictures lack syntactic structure. On the face of it, the toy grammar for the line drawing system B conforms to the picture principle—lines depict parts of what junctions depict. But that grammar permits some structural analyses of drawings and disallows others, and the semantics of that system runs on the analyses that are permitted by the grammar. Yet Fodor seems to infer from the picture principle that there is no "canonical" analysis, or "decomposition", of an image into a set of constituents under specific structural relationships. How does he get there? Fodor writes, "a representation has a *canonical* decomposition iff (at the appropriate level of grain) its parts have content under some *but not all* of the ways of carving it up" (Fodor, 2003, p. 37). From this, it becomes apparent that Fodor's inference rests on two further assumptions that are not contained in the picture principle itself.

The first assumption is that any arbitrary composite of contentful picture-parts itself has content—namely, it depicts the composite of what the picture-parts depict. Only some parts of the phrase, "The 16th President of the United States", have meaning: "the United States" has a meaning but not "of the" or "16th ... States". By contrast, for any set of lines in a line drawing you can specify the set of edges that are depicted. The second assumption is that being a contentful part of a representation is sufficient for being a constituent of that representation. Pictures "have interpretable parts, but they don't have constituents. Or, if you prefer, all the parts of a picture are *ipso facto* among its constituents" (Fodor, 2007, p. 108). The idea is that a

[8]For variations of this principle, see Sober (1976), Block (1983), and Kulvicki (2015).

central function of a grammar is to identify the meaningful parts from which the meaning of a whole representation is composed. Identifying those parts (and their relationships) is tantamount to giving a "canonical decomposition" of the representation. But if all the parts of an image are contentful parts from which the content of the whole could be composed, then a grammar for that image would be vacuous: Any decomposition will do, so no decomposition is privileged. I will target the second assumption here.[9] Even if all parts of an image have content, they do not all play an equal role in explaining the content of the whole.

Suppose, for the sake of argument, that arbitrary parts of an image have depictive content. It is nevertheless possible that some contentful parts contribute to explaining the content of the whole, while others have their contents derivatively and are epiphenomenal or redundant in the composition of the image's content. Perhaps one can point to an arbitrary set of lines in a drawing and say what kinds of edges they depict: "This set of lines depicts a convex edge here, a concave edge there ...." But it is entirely possible that what that set of lines depicts depends on how the image has been parsed into specific regions made up of specific junctions made up of specific lines. Consider a fork-junction in a line drawing. A competent viewer will only find a few admissible ways to interpret that junction (Figure 3). By contrast, if the three line segments that make up the junction are taken independently, not as members of a fork junction, then that set of line segments should be interpretable as depicting any of $3^3 = 27$ possible configurations of edge types. An analysis that elides junctions is not in the same position to explain why an image can depict some of these configurations but not others. The junction structure explains the content of the whole in a way that the arbitrary sets of lines do not. Once that structure is given, the specific contents of arbitrary parts may be abstracted.

Moreover, different analyses of an image might all yield parts that have content but that compositionally determine different contents for the whole. Zhu and Mumford (2007) suggest that the root pattern in Figure 5 is structurally ambiguous. On one decomposition (Figure 5a), the pattern "represents an occlusion configuration with two layers" (Zhu & Mumford, 2007, p. 283)—imagine a wire passing behind a telephone pole. On another decomposition (Figure 5b), the pattern "represents a butting/alignment configuration at one layer"—imagine two corkscrews facing opposite each other on a counter. Each of these parses is consistent with the picture principle and each yields contentful parts. Nevertheless, different parses yield different contents for the whole. Disambiguating the picture's structure contributes to disambiguating its content. A decomposition may have a privileged role in explaining one interpretation but not another.

The picture principle is typically associated with the claim that parts of pictures are semantically uniform or homogeneous, as Fodor says. Sentences can be made up of syntactically and semantically distinct types of phrases. In the case of pictures, it is pictures all the way down: pictures are made up of pictures, and all the parts of a picture are said to make the same type of contribution to the content of the whole. John Kulvicki writes that the parts of a picture "will not be grammatically marked as playing different roles in constituting what the picture as a whole means" (Kulvicki, 2020, p. 39). But while the picture principle implies that every part of a picture is itself a picture of something, these picture-parts may nevertheless have different combinatorial and semantic powers. Different types of image-parts may play different roles in

[9]Burge (2018) rejects the first assumption, which follows from the picture principle itself only on a liberal interpretation of what count as "parts" of an image and "parts" of a scene. For the assumption to follow from the picture principle, "all the parts of P" must range over not just, for example, the constituents recognized by an image grammar (individual lines, the junctions they compose, and so on), but also arbitrary composites of these.

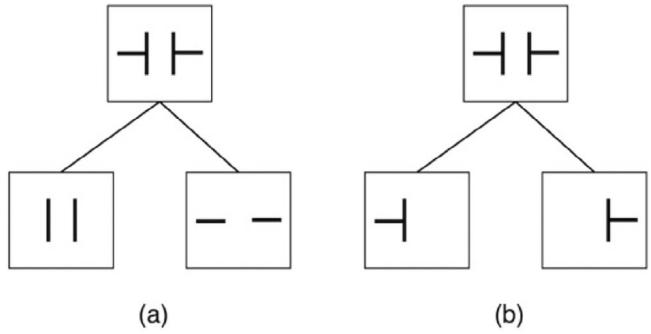

FIGURE 5 Two structural analyses of the same pattern. (Based on Zhu & Mumford, 2007, p. 282).

explaining the content of the whole. Line segments depict edges, not corners; junctions depict corners, not surfaces; fork junctions depict different kinds of corners than L-junctions. Kulvicki writes that images "have no grammar for identifying one part as a subject, another an object, or predicate" (Kulvicki, 2020, p. 49; cf., Burge, 2018). But these are not the only semantic roles worth distinguishing.

These examples highlight the source of the gap in the inference from the picture principle to the claim that images lack syntactic structure. Being a contentful part is not sufficient for being an explanatory part. Not all parts of an image are explanatorily equal and different decompositions into constituents can correspond to the depiction of spatially different types of scenes. Indeed, the picture principle is just silent about how images come to depict fine-grained properties and relations in the scene, such as topology and depth. On the face of it, the picture principle offers a necessary but not sufficient condition on the contents of pictures. The syntax and semantics for B and the parses in Figure 5 all satisfy the picture principle; but they also do more, accounting for the depiction of topological properties and depth relations in the scene. The argument from the picture principle to the absence of pictorial syntax only becomes promising if we deny images these sorts of contents.

5.2.2 | Bare bones content and alternative interpretations

Kulvicki (2020) argues that an observer's pictorial competence allows them to systematically recover only what he calls the "bare bones" content of an image, which on first pass corresponds to the set of scenes of which that image is a projection. Bare bones content does not specify "obvious features of pictorial content like depth and shape" (Greenberg, 2021, p. 859). Bare bones content conforms to the picture principle, but it is not "syntax-driven" (Abusch, 2020, p. 27). No reference to the syntax of the image is required to specify the set of scenes that could have projected it.

One reason for ascribing such content to images is that images permit a range of different interpretations. While one might naturally interpret a still-life image as depicting a flat, round plate viewed at an angle, one could also take it to be an accurate depiction of an ellipse seen face-on, or of an Ames room (a distorted scene that appears normal from a specific viewpoint), or even just of another image (Kulvicki, 2020, pp. 22–25; see also Fodor, 1975, p. 189). Bare bones content is supposed to encode what is common to all the possible alternative interpretations of an image.

However, it is not clear that bare bones content must be part of the core semantics of images. By hypothesis, the "intrinsic images" posited in vision science are committed to

representing rather specific layouts of shapes, colors, depths, and so on. I have argued elsewhere that bare bones, or "projective", content does not figure into explanatory accounts of perceptual representations, including intrinsic images (Lande, 2018). Insofar as intrinsic images are images, not all images are confined to bare bones content. This opens space for some images to have more substantial, syntax-driven content.

What about artifactual images that do admit of a range of alternative interpretations? One possibility is that some of these alternative interpretations are in fact syntax-driven, as with the alternative interpretations of the image pattern in Figure 5. Additionally, the interpretation of an image might be modulated away from its core syntax-driven content (Recanati, 2004). Rules of projective geometry may systematically constrain the permissible ways one broadens, narrows, or replaces the types of scenes that one interprets an image as depicting. The suggestion that bare bones content plays only a modulatory role in picture interpretation could explain why nonstandard interpretations of realist images are normally optional and effortful, whereas certain interpretations in terms of specific layouts of shapes in depth tend to be automatic and difficult to ignore, even when it is prudent to do so (Perdreau & Cavanagh, 2011).

Even if pictures do have bare bones content, this does not preclude them from having richer, syntax-driven content as well. Kulvicki suggests that one is only "keyed in" to richer scene contents on the basis of adventitious features of background context, cognitive expectations, and communicative intentions. But no argument is given for why images cannot systematically depict features such as shape and depth. Vision science and computer vision contain countless proofs, sometimes in the form of image grammars, that contents about the shapes and depths of elements in a scene can be systematically derived from images, without appeal to features of background context or communicative intention.

Indeed, I think it plausible in many cases that one's pictorial competence involves a systematic ability to grasp certain images as depicting determinate layouts of shapes in depth. Consider the realist still-life of a tilted plate on a table. An observer who is unable to grasp the image as depicting a tilted circular object would have a different competence than one who did grasp this content. Moreover, it is plausibly part of an observer's competence that they treat this as the standard and default content of the image, even if an alternative interpretation is more appropriate, all things considered. The point is not just that a normal observer would expect the image to depict a certain arrangement of shapes in depth. Instead, a systematic ability to grasp the image as depicting a tilted circular object, and to treat this as the standard interpretation, is a constitutive feature of many individuals' competences with realist images. Image grammars provide credible models of such competences.[10]

5.2.3 | Interpretation first?

The idea that a picture's syntax helps to drive its semantics might seem to reverse the order of explanation. Maybe we do not interpret junctions as depicting certain kinds of corners because we take them to have a certain syntax; rather we classify junctions as having a certain syntax

[10]Some take a purely pragmatist view of pictorial interpretation, arguing that not even bare bones interpretations are systematic. Rather, the interpretation of pictures is always a matter of pragmatic interpretation based on background expectations, the interpretation of communicative intentions, and adventitious features of the context (e.g., Abell, 2009; Hopkins, 2023). My argument that some images (and certainly psychological images) can systematically have richer-than-bare-bones semantic content also stands against the purely pragmatist view of pictorial interpretation.

based on what types of corners we take them to depict. Camp, for example, writes that "insofar as we can discern syntactic 'parts' to a picture at all, these are either just points in a two-dimensional array, or else regions whose boundaries are given by salient boundaries in the scene being represented" (Camp, 2007, p. 156; see also Hopkins, 2023, pp. 11–12). On the "interpretation-first" view, an image's syntax (such as it is) is driven by its interpretation; the syntax is redundant to, or parasitic on, the interpretation.

It is important to distinguish the interpretation-first view from the more anodyne observation that there is a close fit between the syntax and the semantics of a picture. If the syntax of a system is to serve as an efficient vehicle for representing contents, then we would expect that the syntactic categories and their principles of combination would correspond to "salient" types of features and relations in the represented domain. Junctions and the line rule are well-suited for representing block objects. It does not follow that the syntax is redundant or has no role in interpretation.

The interpretation-first view requires that a competent observer can recover the relevant interpretation of an image independently of any syntactic analysis and that any ascription of structure to an image itself depends on prior interpretation of what the image depicts. How do we come to the prior interpretation? One possibility is that we interpret pictures by, very roughly, determining what objects or scenes the pictorial patterns "resemble" (Abell, 2009). Another possibility is that we interpret a picture as depicting a certain kind of object because the picture triggers in us the same recognition responses as would the depicted object (Schier, 1986). It is question-begging to deny, without argument, that pictorial syntax plays a role in such interpretive processes.

The syntactic structure of an image may well play a role in pictorial interpretation on either theory of depiction. Take the case of resemblance theories. If pictures have syntactic structure, then this may serve as a core source of resemblance: What the picture depicts is constrained by whether the depicted content has a structure that resembles the syntactic structure of the picture. Likewise for recognitional accounts of pictorial interpretation. Image grammars purport to help explain the very recognitional capacities to which Schier, for example, appeals. The interpretation-first view owes an independent argument that pictorial syntax can play no role in explaining how competent observers interpret images.

5.3 | Computation: Using pictures

Some have doubted that pictorial syntax could play a role in supporting computations, inferences, or uses of pictures. For example, Mariela Aguilera writes in passing that "it is rather unlikely that the inferential power of pictures depends on their internal structure" (Aguilera, 2016). However, contemporary work on image grammars explicitly details how image structures can be used to compute decisions about an object or scene's category, for example, not to mention making predictions about a scene and generating plausible images within a given system. In the case of the visual system, theories detail how intrinsic images are formed, transformed, and operated on in virtue of how they are structured (e.g., Yuille & Kersten, 2006; Zucker, 2014).

It may be that some computational processes are insensitive to an image's syntactic structure. For example, feature-matching processes might compare images by measuring the similarity between their syntactic features, irrespective of the structural relations between those features. However, the possibility of structure-insensitive processes does not imply the absence of processes that are sensitive to the core structure of the image.

6 | CONCLUSION

Images have traditionally been thought to lack the sort of structure necessary for expressing complex contents and entering into sophisticated patterns of inference. I have tried to show, by contrast, how images, whether in the world or in the mind, can be structured vehicles of representation. It is not constitutive of images that they lack syntactic structure. Image grammars offer empirically credible models of what the syntax of an image might be like and they illuminate what it would mean, in the first place, for images to have syntax. If images are structured, it does not follow that they are just like sentences in language. There may be marked differences between the syntax of pictures and that of sentences. The lesson is that representations, images among them, can be structured in radically different ways.

It could be argued that a distinctive feature of images is just how far one can go in characterizing an image's content "pre-syntactically," without appeal to anything like a grammar. Perhaps some level of pictorial content abstracts from, or does not depend on, a picture's syntax. Whereas truth conditions cannot be recovered from the words in a sentence independent of their syntactic combination, perhaps one can at least specify bare bones constraints on the accuracy of an image without recourse to its structure. Even if this is so, an image's syntactic structure can play a substantial role in determining how the image is placed within a wider system, how it systematically represents the specific layout of a scene, and how it is manipulated in certain inferences and computations.

ACKNOWLEDGMENTS

For comments and helpful discussion, I am grateful to Louise Antony, Brandon Ashby, Lance Balthazar, Jacob Beck, Denis Buehler, Sam Clarke, Gabe Dupre, Judy Fan, Chris Gauker, Seth Goldwasser, Gabe Greenberg, Bill Kowalsky, Bence Nanay, Joan Ongchoco, Jake Quilty-Dunn, an anonymous reviewer, and audiences at the 3rd Joint Meeting of the Society for Philosophy and Psychology & European Society for Philosophy and Psychology in Milan (July 2022), the Imagistic Cognition Workshop at the University of Salzburg (May 2023), and the annual meeting of the Canadian Philosophical Association (June 2023).

DATA AVAILABILITY STATEMENT

There are no data available.

ORCID

Kevin J. Lande https://orcid.org/0000-0002-9543-7787

REFERENCES

Abell, C. (2009). Canny resemblance. *Philosophical Review*, *118*(2), 183–223.

Abusch, D. (2020). Possible-worlds semantics for pictures. In D. Gutzmann, L. Matthewson, C. Meier, H. Rullmann, & T. Zimmermann (Eds.), *The Wiley Blackwell companion to semantics* (pp. 1–31). Wiley-Blackwell.

Aguilera, M. (2016). Cartographic systems and non-linguistic inference. *Philosophical Psychology*, *29*(3), 349–364.

Anderson, B. L., & Winawer, J. (2008). Layered image representations and the computation of surface lightness. *Journal of Vision*, *8*(7), 1–22.

Astolfi, G., Rezende, F. P. C., Porto, J., V. D. A., Matsubara, E. T., & Pistori, H. (2021). Syntactic pattern recognition in computer vision: A systematic review. *ACM Computing Surveys*, *54*(3), 1–35.

Barrow, H., & Tenenbaum, J. (1978). Recovering intrinsic scene characteristics from images. In A. Hanson & E. Riseman (Eds.), *Computer vision systems* (pp. 3–26). Academic Press.

Bear, D., Fan, C., Mrowca, D., Li, Y., Alter, S., Nayebi, A., Schwartz, J., Fei-Fei, L. F., Wu, J., Tenenbaum, J., & Yamins, D. L. (2020). Learning physical graph representations from visual scenes. In *Proceedings of the 34th international conference on neural information processing systems* (pp. 6027–6039). Curran Associates, Inc.

Biederman, I. (1987). Recognition-by-components: A theory of human image understanding. *Psychological Review*, *94*(2), 115–147.

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.

Blumson, B. (2014). *Resemblance and representation: An essay in the philosophy of pictures*. Open Book Publishers.

Burge, T. (2018). Iconic representation: Maps, pictures, and perception. In S. Wuppuluri & F. Doria (Eds.), *The map and the territory* (pp. 79–100). Springer International Publishing.

Burge, T. (2022). *Perception: First form of mind*. Oxford University Press.

Camp, E. (2007). Thinking with maps. *Philosophical Perspectives*, *21*(1), 145–182.

Camp, E. (2018). Why maps are not propositional. In A. Grzankowski & M. Montague (Eds.), *Non-propositional intentionality* (pp. 19–45). Oxford University Press.

Cavanagh, P. (2021). The language of vision. *Perception*, *50*(3), 195–215.

Clarke, S. (2022). Mapping the visual icon. *The Philosophical Quarterly*, *72*(3), 552–577.

Clarke, S. (2023). Compositionality and constituent structure in the analogue mind. *Philosophical Perspectives*. Advanced online publication. https://onlinelibrary.wiley.com/doi/full/10.1111/phpe.12182

Clowes, M. (1967). Perception, picture processing and computers. In N. Collins & D. Michie (Eds.), *Machine intelligence* (Vol. 1, pp. 181–197). Oliver & Boyd.

Clowes, M. (1971). On seeing things. *Artificial Intelligence*, *2*(1), 79–116.

Fan, J. E., Hawkins, R. D., Wu, M., & Goodman, N. D. (2019). Pragmatic inference and visual abstraction enable contextual flexibility during visual communication. *Computational Brain & Behavior*, *3*(1), 86–101.

Fodor, J. (1975). *The language of thought*. Harvard University Press.

Fodor, J. (2007). The revenge of the given. In B. P. McLaughlin & J. D. Cohen (Eds.), *Contemporary debates in philosophy of mind* (pp. 105–116). Blackwell.

Fodor, J. A. (2003). *Hume variations*. Oxford University Press.

Fu, K. S. (1974). *Syntactic methods in pattern recognition*. Academic Press.

Geman, S., Potter, D. F., & Chi, Z. (2002). Composition systems. *Quarterly of Applied Mathematics*, *60*(4), 707–736.

Goodman, N. (1968). *Languages of art: An approach to a theory of symbols*. The Bobbs-Merril Company, Inc.

Greenberg, G. (2021). Semantics of pictorial space. *Review of Philosophy and Psychology*, *12*(4), 847–887.

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

Guzmán, A. (1968). Decomposition of a visual scene into three-dimensional bodies. In *Proceedings of the December 9–11, 1968, fall joint computer conference, part 1* (pp. 291–304). Association for Computing Machinery.

Han, F., & Zhu, S.-C. (2009). Bottom-up/top-down image parsing with attribute grammar. *IEEE Transactions in Pattern Analysis and Machine Intelligence*, *31*(1), 59–74.

Haugeland, J. (1981). Analog and analog. *Philosophical Topics*, *12*(1), 213–225.

Hopkins, R. (2023). Design and syntax in pictures. *Mind & Language*. Advance online publication. https://doi.org/10.1111/mila.12485

Huffman, D. A. (1971). Impossible objects as nonsense sentences. In B. Meltzer & D. Michie (Eds.), *Machine intelligence* (Vol. 6, pp. 295–323). Halsted.

Kosslyn, S. M. (1980). *Image and mind*. Harvard University Press.

Kulvicki, J. (2020). *Modeling the meanings of pictures*. Oxford University Press.

Kulvicki, J. V. (2014). *Images*. Routledge.

Kulvicki, J. V. (2015). Analog representation and the parts principle. *Review of Philosophy and Psychology*, *6*(1), 165–180.

Lande, K. J. (2018). The perspectival character of perception. *The Journal of Philosophy*, *115*(4), 187–214.

Lande, K. J. (2021). Mental structures. *Noûs*, *55*(3), 649–677.

Lande, K. J. (2023). Contours of vision: Towards a compositional semantics of perception. *The British Journal for the Philosophy of Science*. Advance online publication. https://doi.org/10.1086/725094

Larson, R., & Segal, G. (1995). *Knowledge of meaning: An introduction to semantic theory*. MIT Press.
Liu, X., Zhao, Y., & Zhu, S.-C. (2018). Single-view 3D scene reconstruction and parsing by attribute grammar. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *40*(3), 710–725.
Mackworth, A. K. (1976). Model-driven interpretation in intelligent vision systems. *Perception*, *5*, 349–370.
Perdreau, F., & Cavanagh, P. (2011). Do artists see their retinas? *Frontiers in Human Neuroscience*, *5*(171), 1–10.
Quilty-Dunn, J. (2020). Perceptual pluralism. *Noûs*, *54*(4), 807–838.
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*, *46*, E261.
Recanati, F. (2004). *Literal meaning*. Cambridge University Press.
Schier, F. (1986). *Deeper into pictures: An essay on pictorial representation*. Cambridge University Press.
Shin, S.-J. (1994). *The logical status of diagrams*. Cambridge University Press.
Sober, E. (1976). Mental representations. *Synthese*, *33*(2–4), 101–148.
Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: Statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.
Võ, M. L.-H. (2021). The meaning and structure of scenes. *Vision Research*, *181*, 10–20.
Waltz, D. (1975). Understanding line drawings of scenes with shadows. In P. Winston (Ed.), *The psychology of computer vision* (pp. 19–91). McGraw-Hill.
Willats, J. (1997). *Art and representation: New principles in the analysis of pictures*. Princeton University Press.
Willats, J., & Durand, F. (2005). Defining pictorial style: Lessons from linguistics and computer graphics. *Axiomathes*, *15*(3), 319–351.
Yuille, A., & Kersten, D. (2006). Vision as Bayesian inference: Analysis by synthesis? *Trends in Cognitive Sciences*, *10*(7), 301–308.
Zhu, L. L., Chen, Y., & Yuille, A. (2011). Recursive compositional models for vision: Description and review of recent work. *Journal of Mathematical Imaging and Vision*, *41*(1–2), 122–146.
Zhu, S.-C., & Huang, S. (2021). *Computer vision: Stochastic grammars for parsing objects, scenes, and events*. Springer.
Zhu, S. C., & Mumford, D. (2007). A stochastic grammar of images. *Foundations and Trends in Computer Graphics and Vision*, *2*(4), 259–362.
Zhu, S.-C., & Wu, Y. N. (2023). *Computer vision: Statistical models for Marr's paradigm*. Springer.
Zucker, S. W. (2014). Stereo, shading, and surfaces: Curvature constraints couple neural computations. *Proceedings of the IEEE*, *102*(5), 812–829.

How to cite this article: Lande, K. J. (2023). Pictorial syntax. *Mind & Language*, 1–22. https://doi.org/10.1111/mila.12497