

Neural Images

Gabriel Greenberg
March 9, 2026

1 Introduction

In 1982, Tootell and his colleagues unveiled a new technique for displaying neural activity in the visual cortex of a macaque monkey (Tootell, Silverman, et al. 1982; Tootell, Switkes, et al. 1988). The experimenters arranged for the lab animals to stare at a stable visual display of a vivid circular grid. While looking at the grid, the monkeys were injected with a radioactive marker, and then euthanized. Their brains were quickly extracted, and their primary visual cortices flattened, then frozen. Once hardened, the cortex was cut into thin flat slices, and these slices were exposed to photosensitive paper. The paper reacted selectively to the radioactive markers, resulting in a directly registered image of brain activity at the time of the initial viewing. Remarkably, the original grid pattern to which the monkeys were exposed is visible to the naked eye in the developed image, implying a corresponding pattern of activity of the cortex itself. There were, it seemed, pictures hidden in the brain.

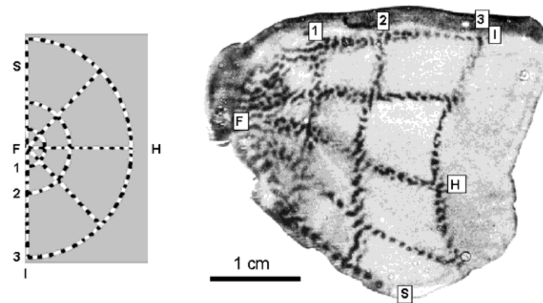


Figure 1: Visualizing retinotopy. Image source: Tootell, Switkes, et al. 1988.

The visually organized pattern of activity in the cortex, known as **retinotopy**, was first docu-

Acknowledgements: for discussions that have been essential to the development of this paper, I am grateful to Dave Barker-Plummer, Jacob Beck, Ned Block, Rosa Cao, Judy Fan, Kevin Lande, Jake Quilty-Dunn, Gabe Rabin, Dan Yamins, and audiences at NYU and Stanford. Special thanks to Elvie Lin for editorial and research assistance.

mented many decades before Tootell’s demonstration (Schwarzlose 2021, pp. 9–16). And evidence has since accumulated, revealing more than a dozen retinotopic areas in the visual processing stream (Wandell, Dumoulin, and Brewer 2007, p. 367; Brewer and Barton 2012, pp. 38–51). Can we infer from such findings that there are in fact pictures in the brain? Most philosophers and cognitive scientists have thought that we cannot. After all, there is no homoncular viewer inside the brain that might view and interpret such images. Instead, we would have to show that such brain states are actually *used* as images, but what would it be for a brain state to be used as an image anyway? These are challenging questions, but I believe they can be answered. In this paper, I will argue that the representations produced by retinotopic areas are in fact genuinely iconic, and that they function as picture-like representations of the visual world. Retinotopic states are used as picture-like representations, I will propose, by computational layers in the visual system that take the form of convolutional neural networks.¹

2 Retinotopy

Visual field maps, such as those found in V1, V2, or V3, are regions of the mammalian visual cortex whose component parts have receptive fields that (more or less) fully tile visual space. The receptive field of a neuron, or population of neurons, is understood as the region of visual space to which the neuron is responsive. Retinotopy, then, is a property of visual field maps in which the spatial organization of neural activity corresponds systematically with the spatial organization of the corresponding receptive fields. Neuroscientists typically describe this correspondence in terms of the preservation of adjacency relations between areas of the cortex and their receptive fields. Brewer and Barton’s (2012) definition of retinotopy is typical; for a visual area to be retinotopic:

Neurons whose visual receptive fields lie next to one another in visual space are located next to one another in cortex. (Brewer and Barton 2012, p. 29)

To understand what this means, and how it is possible, it is helpful to recall that the cortex is essentially a flat sheet of layered neurons, folded up within the skull. Physically speaking, then, the cortex itself is a two-dimensional surface. Distinct visual areas correspond to different regions of this sheet, wired together through gray matter projections that lie outside of the cortex. Retinotopy is a property of areas within the cortical sheet, in which their two-dimensional layout mirrors the layout of the visual field.

¹The discussion here assumes a broadly computationalist and externalist perspective on perception.

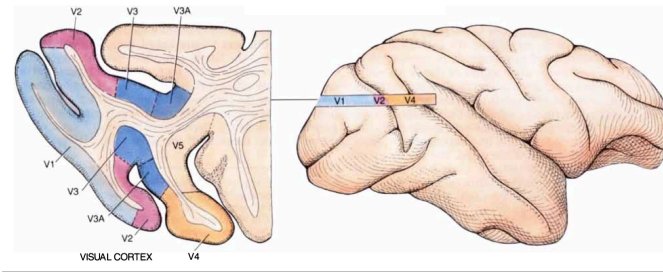


Figure 2: Cross section of visual areas. Early visual areas in brain of a macaque monkey.

It is worth emphasizing that retinotopy is, in part, a property of the *physical space* of the cortex. Standard definitions, like the one above, refer to the proximal *locations* of neurons within the cortical sheet. At the same time, retinotopy is not an intrinsic physical feature of cortical areas, but a relational one: it describes the correlation between spatial layout in the cortex and the layout of visual space. For a region to be retinotopic, it must normally bear a certain kind of relation to external stimuli.

To get a feel for the nature and extent of retinotopy, I highlight three key fact patterns. First, more than a dozen different retinotopic areas have been discovered in the visual systems of humans and macaque monkeys, and the number continues to grow (Wandell, Dumoulin, and Brewer 2007). (I will refer to the highly homologous visual systems of humans and macaques somewhat interchangeably.) The human visual system is arranged in a loose hierarchy, with visual information passing from the eyes to the LGN, and then, within the cortex, to V1, then V2, then branching out into ventral and dorsal streams (Cao and Yamins 2024). The retinotopic layouts of each area are fixed in meticulous spatial registration by tens of millions of precisely organized axonal projections from one area to the next (Zeki 1993).

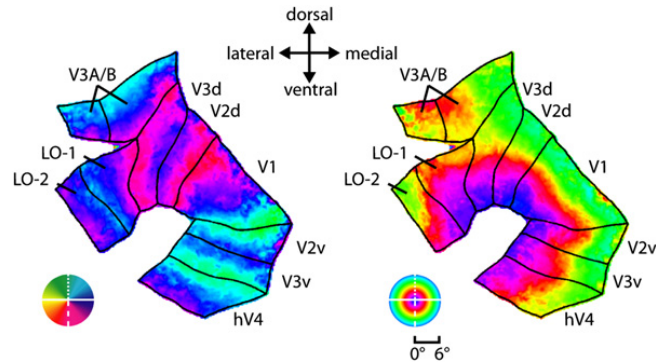


Figure 3: Multiple retinotopic areas. A flattened representation of some retinotopic areas in early- and mid-level vision, with angular position picture at left, and eccentricity pictured at right. Image source: Wandell, Dumoulin, and Brewer 2007.

Second, at each stage of the visual hierarchy, average receptive field size increases. Each stage appears to be composed of elements that are sensitive to larger, more abstract spatial patterns than those of its predecessor. This fact is reflected in the observation that retinotopy is most granular and robust in the early precortical and cortical areas and decreases in its specificity in later areas (Zeki 1993, Arcaro and Livingstone 2017).

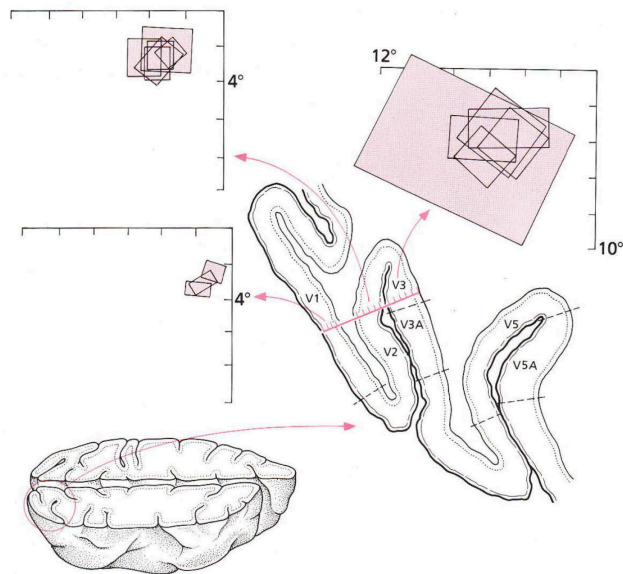


Figure 4: Increasing receptive field size along the ventral stream. Receptive field sizes as measured by a single probe passing through V1, V2, and V3 of a macaque monkey cortex. Image source: Zeki 1993.

A final observation is that the spatial layout of retinotopic maps in the cortex is not isometric, even under scaling, to the layout of visual space or the image on the surface of the retina. Instead, a phenomenon known as **cortical magnification** means that cortical areas whose receptive fields lie within the foveal region of vision are assigned much more space in the cortex, and those targeting the periphery, much less. This accounts for the characteristic stretching of visual space visible in the image from Tootell et al. (1988, pp. 1534–1536) with which we began.

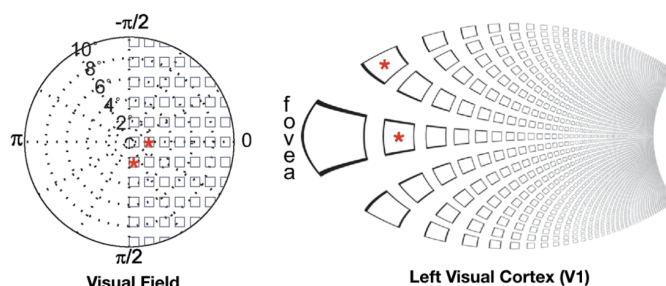


Figure 5: Cortical magnification. In retinotopic areas, central regions of the visual field are mapped by magnified regions of the cortex. The image above is based on a mathematical model. Image source: Schira et al. 2010.

It's widely assumed that the retinotopic layout of visual areas can be explained as the combined product of their computational function and a general principle of efficient wiring in the brain. Explanations of this form have been used to predict other organizational motifs in the visual cortex, and they have been conjectured also to explain retinotopy. (REF) Of course, by this logic, *all* parts of the brain are subject to the same combination of functional and wiring constraints, so the question in this case becomes: what are the particular functional constraints that give rise to retinotopic layout?

Here, I think we could imagine two broad styles of explanation, noting that they are not mutually exclusive. Retina-oriented accounts hold that the retinotopic organization of visual areas can be explained primarily in terms of spatial organization of input from the retina. Computation-oriented explanations hold that the retinotopic organization is maintained primarily because of the computational function it serves in subsequent processing. The account given here explores the second kind of hypothesis.

3 What is the meaning of retinotopy?

Here is the plan for this section:

- **Opening question:** What is the significance of retinotopy to information processing?
- **Wiring optimization + function:** Common wisdom in neuroscience: retinotopy is the result of

wiring optimization and function.

- On its own, this claim is uninformative, since presumably the same is true for all brain structures.
- Helps draw a distinction between kinds of explanatory factors: function-based, and implementation-based (wiring optimization).
- What is the dominant explanation of retinotopy?
- **Hypothesis 0:** The spatial information carried by retinotopic locations has no functional significance.
 - **Developmental explanation:** Retinotopy is driven by development and evolution from the organization of the retina itself plus wiring optimization, but not at all by computational function.
 - **The Tree Ring Model:** Retinotopic layout is like tree rings: physical encoding of information that is vivid to the observer, but functionally inert for the organism.
 - **Cortex–Perception correlations as counter-example:** This model is contradicted by a variety of cortex–perception correlations:
 - * Lesions in cortex = scotomas in the visual field.
 - * Activations of cortex = phosphenes in the visual field.
 - * Relative size in cortex = judgements of size in perception.
 - **Spatial information is preserved.** These examples show that retinotopic information is not thrown away, but used by later cognition.
 - What does this show about representation?
- **Hypothesis 1:** Retinotopic locations in cortex function as iconic representations of locations in visual space.
 - **Kosslyn–Burge argument:** correlations → registration → representation → iconicity
 - **Why this argument is inconclusive:** jump from information registration to representation.
- **Hypothesis 2:** Retinotopic locations in cortex (merely) register information about locations in visual space.
 - Retinotopic locations as memory addresses; information via architecture.
 - Quotes from Pylyshyn.
- **What’s at issue (roughly):** Do the spatial properties of the retinotopic maps play a computational role for their consumer computation?
 - **Mental calipers:** The idea of mental calipers, the fact that there is none.
 - **Direct computation requirement:** For a representational system to be iconic, (1) the consumer computation must be sensitive to the structural features of the representation; (2) the

consumer computation must exploit an iconic semantics defined over these structural features.

- **What's called for:** To settle the debate about the iconicity of retinotopic representations, we need a substantive account of the computations they enter into.

4 The pictorial hypothesis

Here's the plan for this section:

- **Key claims:**
 - **The iconic hypothesis:** The representations tokened by retinotopic areas are iconic.
 - **The pictorial hypothesis:** The representations tokened by retinotopic areas are 2D forms that represent 3D visual space via picture-like projection semantics.
- **Theory of signs:**
 - **Outline theory of iconicity:** Uniform, sign-dependent semantic rules. (Highlight the role of semantics)
 - **Outline theory of pictorial semantics:** Viewpoint-based projection semantics.
- **Mental iconicity:**
 - **From signs to mental representations:** Semantic rules cannot be grounded in mental representations themselves, but must reflect the relationship of the mind to the environment.
 - **A functional account of iconicity:**
 - **A functional account of picture-like representation:**
- **The burdens of proof**
 - **Retinotopic representations are 2D structures.**
 - **Consumer computation is sensitive to this 2D structure.**
 - **Consumer computations exploit a projection-semantics defined over these 2D structures.**
- **The form of the argument**
 - **Format depends on use by consumer computations.** So to settle the question of iconicity for retinotopic representations, we need a model of the consumer computation.
 - **Identify a plausible model:** Convolutional neural networks (CNNs).
 - **Show how the CNN model exploits projection-semantics.**
- **Contrast with other kinds of arguments for iconicity**

- **Contrast with analogue arguments** (Weber’s law, etc.)
- **Contrast with Beuhler’s argument**
- **Contrast with Clarke and Block’s arguments**
- **Complementary, not contradictory, methods:** Converging lines of evidence about iconicity in early vision.

Here’s the un-revised text from my last draft:

The third hypothesis is that retinotopy reflects the use of functional space within the cortex to iconically represent the organization of visual space. In particular, it reflects a picture-like use of functional space within the cortex to represent visual space. This is the hypothesis defended in this paper, broadly in line with the conclusions of Kosslyn (2006, pp. 100–112), Burge (2022, pp. 305–311), and Beuhler (2024).

To articulate and support this hypothesis we need at least a working account of iconicity in hand. In previous work, I’ve argued for a view of iconicity in public signs, in which iconicity is distinguished by the kind of semantics used to map signs to contents (Greenberg 2023). According to this view, roughly, in iconic representation, the content of a sign bears a natural dependence to the form of that sign. In the specific cases of pictorial representation, the content of a picture bears a relation of perspective projection to the geometrical form of the picture itself (Greenberg 2013, 253–82, 2021, secs. 1-3).

I believe the basic outline of this view carries over to the mental domain, except that, instead of thinking of representational systems as conventions among communicators, we consider them as components of cognitive processes which function to token representations. Such representational systems can be distinguished, with some idealization, from the computational systems which take representations as inputs, and produce them as outputs. With this notion of representational system in mind, I propose to adopt the following provisional definition of iconic mental representation.

- (1) **Mental iconicity:** A representational system S is iconic iff for any representation ϕ that S has the capacity to token, the content of ϕ is determined by a natural dependency on the form of $R(\phi, E)$.

For now, I suppress the issue of representational order. In general, representations may be understood to be composed at different structural orders, where first-order representations are bearers of content that have no meaningful parts, and second-order representations put first-order elements into meaningful structural arrangements. Because different semantic rules apply at each order, whether or not they are iconic may vary from order to order. In the present context, I mean to argue that the representations produced by retinotopic areas are iconic at the level of their 2nd-order structure (roughly, their geometrical structure). This claim is independent from the issue of

whether lower-order constituent representations, such as representations of individual features or objects, are themselves iconic (cf. Block 2023).

The conception of mental iconicity adopted here diverges in various ways from the definitions which have been deployed in recent discussions of perceptual representation (Beck 2018; Burge 2018; Quilty-Dunn 2020; Green and Quilty-Dunn 2021; Clarke 2022; Block 2023). Still, for present purposes, I think the spirit is much the same, since all invoke some kind of correspondence between representation and content; and the different accounts largely agree on prototypical cases. I’ve argued elsewhere (Greenberg 2023) that the semantic approach to iconicity can explain the key signatures of iconicity noted by other authors.

To make sense of the definition of iconicity proposed above, we must reconceive the theoretical role of semantic rules. In the case of public representation, these are rules of *interpretation*—rules which an interpretive agent explicitly or implicitly encodes and follows in the process of interpreting a signal. But for mental representations, of course, there is no privileged interpreter or homunculus to follow a rule of interpretation. Rather, I will assume, the rules which map representations to their contents arise from a representational system’s history of interactions with the environment, and with its own computational context.

I propose that, at least in the case of the perceptual system, we may usefully understand semantic rules as deriving from underlying informational functions. These are functions of a system to establish correlations between internal variables and external variables (Neander 2017, Shea 2018). **Iconic functions**, in particular, are informational functions to establish internal-external correlations *by* producing internal states whose form depends on the properties of an external state.² For example, analogue magnitude systems would have iconic functions because they function to token neural states (analogue magnitude representations) with activation levels that stand in logarithmic relations to environmental magnitudes (Carey 2009; Beck 2015).

The specific conjecture of this essay is that retinotopic visual areas function to token representations whose 2D metric properties stand in relations of perspective projection to visual space. Thinking of iconic semantics in terms of iconic informational functions helps ground ascriptions of content in predictions about empirically observable covariations. To say that a state’s content depends on its form is not just a philosophical abstraction, but implies, under functionally optimal conditions, concrete spatial correlations between optical space and cortical space.

If we assume the foregoing approach to mental iconicity, and its basis in iconic functions, then any argument that retinotopic areas have picture-like iconic semantics incurs three key burdens of proof. (i) First, we must provide evidence for picture-like representational structures in the visual cortex, in the form of computationally functional spaces governed by a 2D metric. (ii) Second, we must show that subsequent visual computations make use of this functional 2D metric. (And, as discussed, such computations must be the immediate consumers of the representations

²Compare to Shea 2018 on functional structural correspondence.

in question.) (iii) Third, we must argue that subsequent visual computations exploit a relation of perspective projection between visual space and the functional 2D metric. It is not enough to show that there are functional 2D metrics in the cortex, or that visual computations make use of these metrics. According to the conception of iconicity assumed here, we must also argue that these functional metrics stand in the right kind of dependency relation to their visual content.

The remainder of the paper takes on each of these burdens of proof in turn. Of course, my argument is not intended as a proof. Instead, I hope at least to establish the plausibility of the iconic hypothesis, and to marshal a range of evidence for it. I begin with the notion of functional space, then turn to the visual computations which might use this space, and finally to the informational relations that connect functional space to the external world.

5 Physical space and functional space

5.1 From physical space to functional space

Retinotopy, as it figures in neuroscientific literature, is a property of the physical space of the cortex. Recall Brewer and Barton's (2012, p. 29) standard characterization: "neurons whose visual receptive fields lie next to one another in visual space are located next to one another in cortex."

Yet physical space, by and large, is not functional for brains.³ The principal computational elements of the central nervous system are neurons and neural connections. There are no standard mechanisms by which one neuron may be sensitive to the physical distance between upstream neurons. But if this is so, then retinotopy as stated *can't* be functional for brains—pushing us back toward the skeptical Hypothesis 1. Instead, if a space is to be functional within the cortex, that space must be defined by neural connections, rather than actual physical distance (or even, physical distance measured along the surface of the cortex). This was Block's objection, as reported by Kosslyn (2006, p. 131):⁴

One might argue that the crucial variable is not the actual physical distance on the cortex, but rather a kind of "functional distance" defined by connectivity among neurons in a brain area. Start by taking the extreme case, where physically distant neurons are connected directly to each other and thus function as if they were adjacent, whereas physically nearby neurons are connected indirectly to each other via intermediate neurons and thus function as if they were very far apart. If the neurons in an area were so arranged, would the area in fact depict? Now take a less extreme case, where physically distant neurons tend to have more intermediate neurons between them—and

³There are various well-known exceptions, see, for example Catherine Carr and Masakazu Konishi (1988 in Science, 1990 in J. Neuroscience).

⁴The objection was first put to me by Rosa Cao (p.c.).

thus, the crucial variable is not actual distance on the cortex but rather the number of connections that intervene among neurons. Would such an area still depict?

Kosslyn himself adopted a notion of a “functional space” which he has defined in various ways in different publications (Kosslyn 1981, 1994; Kosslyn, Thompson, and Ganis 2006; Tye 1991). But it is fair to say that these definitions were vague on key details and did little to satisfy philosophically-oriented critics (Beck 2018; Pylyshyn 2003, 2007). In some cases, Kosslyn’s definitions of functional space seemed to lean heavily on the geometrical nature of the subject matter— in which case, a functional space is just a representation whose content is spatial; then the question of iconicity does not arise. In others, the definitions replace physical distance with other structural relations that have equally little purchase in neural computation, such as the numbers of elements between nodes. Some of these vices are legible in a representative passage from Kosslyn 1981 (and highlighted in Tye 1991).

The primary characteristic of representations in this format is that every portion of the representation must correspond to a portion of the object such that the relative inter-portion distances on the object are preserved by the distances among the corresponding portions of the representation. (50)

Importantly, distance in the medium can be defined without reference to actual physical distance but merely in terms of the number of locations intervening between any two locations. (50)

Suffice for now to say that I share Kosslyn’s conviction that there is a viable notion of functional space by which we can make sense of computationally effective spatial properties of retinotopic areas. But the notion of functional space must be defined by abstraction from the layer of physical implementation rather than as a constraint on content, in order to circumvent the pit that critics have already highlighted. This is the strategy that I’ll pursue in all follows.

5.2 Wiring optimization

The idea that retinotopy is likely to reflect key functional organization at the neural level is suggested by the widely discussed principle of **wiring optimization** from theoretical neuroscience (Durbin and Mitchison 1990; Chklovskii, Schikorski, and Stevens 2002; Margalit et al. 2024). The principle holds that neural circuits will, all things equal, employ the shortest neural paths that are required to fulfill their computational functions. This idea is motivated by general considerations having to do with materials and energy needed to construct axonal connections, and by the increased time required for a signal to propagate through a longer axon.

Wiring optimization implies that neurons will be wired in physical proximity proportional to the regularity with which their outputs must be compared. As it were, neurons that fire together,

wire close together. Assuming this principle, we can defeasibly infer from spatial proximity in cortex to functional convergence of outputs. And thus, from retinotopy to a topology of functioning adjacency relations in the neural network. From this perspective, wiring optimization provides a principled bridge between retinotopic physical space and functional organization (Chklovskii and Koulakov 2004).

5.3 Functional space

Our goal now is to identify a notion of functional space appropriate to early visual field maps that avoids the fallacy of importing features of represented space to the spatial features of the representation (Pylyshyn 2003).

I now introduce the concept of an **accessibility metric**, understood as a measure of how informationally accessible one node is from another, relative to a given network of nodes. ("Nodes" here may be understood to refer to neurons, or populations of neurons.) In a feedforward neural network, we can think of the accessibility distance between two input nodes as proportional to the number of output nodes they both feed into, or the relative weight given to their inputs when inputs are compared. Intuitively, accessibility distance reflects the degree to which the information from distinct nodes may be combined by the network in the process of computation.

For simple illustration, suppose we have a 1-dimensional feedforward neural network, where every pair of nodes feeds into a single output node. Then we can define the accessibility metric by assuming the following equation. Extensions of the metric to greater distances are achieved by applying standard metric axioms.

- (2) $d(x, y) = 1$ iff x and y feed into the same output node

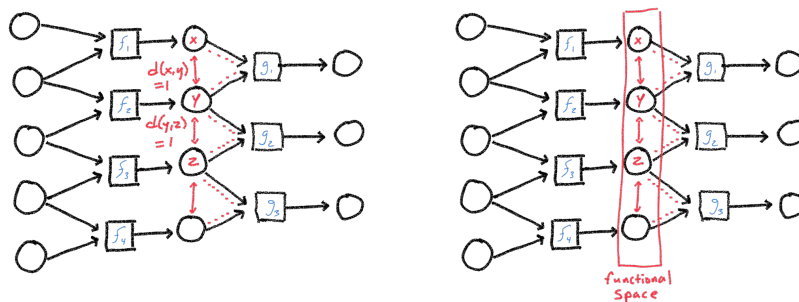


Figure 6: Accessibility metrics and functional space. The accessibility metric for a 1D layer of a feedforward network and the corresponding functional space.

Now consider a slightly more realistic case; we have a 2-dimensional feedforward network, in which every 3x3 array of input nodes feeds into a single output node. This will result in a 2D

city-block metric, as visualized below.

- (3) a. $d(x, y) = 0$ iff x and y feed into the same *nine* output nodes
b. $d(x, y) = 1$ iff x and y feed into the same *six* output nodes
c. $d(x, y) = 2$ iff x and y feed into the same *four* output nodes
d. $d(x, y) = 3$ iff x and y feed into the same *two* output nodes
e. $d(x, y) = 4$ iff x and y feed into the same *one* output nodes
f. $d(x, y) > 4$ iff x and y feed into *none* of the same output nodes

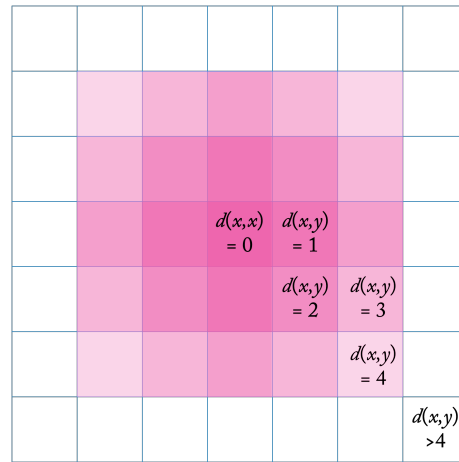


Figure 7: Accessibility metric for a 2D network layer. A 3x3 kernel produces six degrees of accessibility differentiation, which imply a functional space with a 2D city-block metric.

For biologically realistic cases, more complex accessibility metrics might take into account a variety of structural features reflected in neural computation. For example, in a model of V1, we should expect layer 4 cells from a single hyper-column to be effectively distance 0 from one another (even if they belong to distinct orientation columns), while the distance between hyper-columns would be determined by the reach of their complex cells. My presumption here is not that there is one analytical definition of accessibility metric than can be mechanically applied to all cases. But I do expect that, in any given computational environment, there will be a sensible way of developing an accessibility metric. This is especially plausible in the neatly layered, feed-forward framework of early visual processing.

In general, the structure of a given space is determined by its metric. We can now define the **functional space** of a visual layer by its accessibility metric. This gives us a notion of space for a neural population that is functional, since it reflects informational flow rather than physical

distance. But it is still a feature of the representation itself, and not derived from the features of a notional represented space. I have not yet argued that the functional space of visual areas does in fact work iconically; what I maintain, for now, is that if there is to be a kind of space in visual areas that works iconically, then it will be this kind of functional space.

Several consequences of this approach are worth highlighting. First, an accessibility metric applies to a collection of neurons only relative to another collection of neurons that they feed into. Functional space isn't an intrinsic feature of a neural population, but, appropriately enough, reflects its informational functions with respect to downstream computation.

Second, any given collection of neurons can form an accessibility metric. But we must still ask in each case, (a) what kind of metric do they form? and (b) is this set of connections actually used *as* a metric? The accessibility metric of an arbitrary collection of neurons will not necessarily be coherent or even functional for the system. Picking out the right population to be viewed in this way is a matter of identifying perspicuous computational explanations.

In the context of early vision— at least our simplified model of it— we have large, flat, feed-forward neural networks with sparse local connections. Such configurations give rise to roughly Euclidean, two-dimensional accessibility metrics, not unlike a picture plane. It remains to show how these metrics are exploited.

A final complication to note has to do with the presence of massive feedback connections in early vision (Kandel et al. 2000). My definition of informational accessibility privileges feed-forward connections, which give rise to the characteristic hierarchical view of visual processing. So the mere existence of feedback connections doesn't upset the basic picture of accessibility. Still, if we focus, for example, on V1 layer 5, which sends strong retinotopically organized feedback connections to LGN, we might wonder if we can still read off a sensible accessibility metric in the same way. This is, of course, a speculative matter. But the fact that backward projections strictly preserve retinotopy, leading from V1 back to LGN in precise spatial registration, suggests that even here, we may employ the same basic conception of an accessibility metric to define a 2D space.

To this end, let us adopt a simplifying model of early visual areas, and their representations, as layers in a sparsely connected feedforward neural network (Cao and Yamins 2024, Kriegeskorte 2015). While there are many shortcomings of such a model, it is at least an abstraction drawn from life. And it bears a passing resemblance to the computational processes within V1, in particular the passage from simple cells in layer 4 to complex cells in layers 2/3, as well as, more globally, the transition from processing in V1 to processing in V2, or from V2 to higher areas, such as V3v or V5 (Kandel et al. 2000).

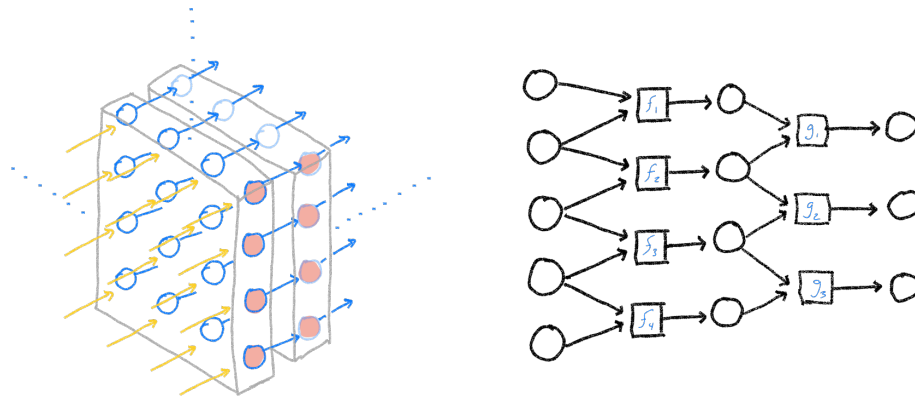


Figure 8: Visual computation as a feedforward neural network. Visual computation on 2D layers (left) and 1D layers (right).

5.4 Visuotopy

The physical signature of retinotopy is what first made it amenable to scientific study, and provocative to philosophical reflection. But we have seen that if some kind of space is to play a role in neural computation, it must be functional space, not physical space. It is now time to kick away the ladder. Let us introduce the functional concept of a **visuotopic map** to supplant our talk of retinotopic maps.⁵ It is visuotopy, not retinotopy, which I will ultimately argue is the proper bearer of picture-like iconicity in the brain. We may roughly define visuotopy as follows:

- (4) A visual area V is **visuotopic** iff the locations of nodes in the 2D functional space of V are approximately smooth and invertible transformations of the locations of their receptive fields in visual space.

Defined this way, visuotopy departs from retinotopy in three ways.

(i) First, visuotopy describes a property of a cortical area's functional space, rather than its physical space. By the principle of wiring optimization, we may take retinotopy as good evidence for visuotopy. But visuotopy could also arise, perhaps at later layers, with no straightforward spatial signature.

(ii) Second, whereas the standard definitions of retinotopy are cast in terms of adjacency preservation, visuotopy is defined in terms of smooth, invertible transformation. As a matter of descriptive fact, the physical space of each retinotopic area stands in a relationship of approximately smooth, invertible transformation to the retinal space. And it is reasonable to assume that the same would hold for the relationship between the functional space of these areas and regions of the visual field. This still carries the implication that functional proximity is preserved, but also

⁵The term "visuotopic" is used in the empirical literature, more or less interchangeably with "retinotopic." I co-opt the term here to serve a specific theoretical role.

clarifies that the transformation in question preserves a rich array of other geometric information beyond local proximity.

The adjacency-based formulation may have been adopted in neuroscience partly in light of the facts of cortical magnification. Cortical magnification implies that visual field space is not isometric under scaling with retinotopic space—there are, as it were, “distortions” in the latter. So metric relations in general are not preserved, but adjacency is.

This might be taken to imply that local adjacency relations are the *only* spatial information that is available in the retinotopic map, which in turn might engender skepticism about the functional significance of retinotopy.⁶ This is because preservation of adjacency alone is compatible with any amount of smooth stretching; but mere adjacency is an insufficient base for the extraction of geometrical information that could ground metrically precise actions and phenomenology. In other words, the mere preservation of adjacency relations isn’t the kind of information which must be functional, if the iconicity hypothesis is correct.

If we instead note that the transformation from visual space to functional cortical space is smooth and invertible, then it is clear that no 2D spatial information is lost. In principle, one can use locations in the functional space to work backwards to regions of visual space; and this is exactly what I hypothesize that vision does. Cortical magnification, and the loss of isometry, is no threat to this project, so long as the metric transformation remains approximately invertible.

(iii) Third, the definition of visuotopy drops reference to retinal locations in favor of locations in visual space. Here, it seems, retinotopy was always mislabeled. Even though scientists do sometimes describe retinotopy as a relationship between space on the cortex and space on the retina (Chklovskii and Koulakov 2004), their experimental paradigms equally often invoke binocular information—meaning there are really two retinas involved, not one (Wandell, Dumoulin, and Brewer 2007). What’s more, the topographic layout characteristic of retinotopy in fact applies at once to the input from both eyes (Kandel et al. 2000). That is, adjacent points on the cortex correspond to adjacent points in *both* retinas, where they overlap. Thus, the more encompassing statement of “retinotopy” should really advert to locations in the **visual field**, rather than locations on the retina. And indeed, this is the notion that scientists often invoke, even under the title of “retinotopy.” This shift to visual field position is made explicit in the definition of visuotopy above.

I should also clarify the notion of a receptive field, which I defined earlier as a region of visual space. My baseline assumption is that visual space can be thought of as a solid angle extending away from the perceiver, and anchored at a cyclopean viewpoint (Peacocke 1992, Greenberg MS). The viewpoint defines an angular space in which “locations” are solid angles. So a neuron’s receptive field is just such a region of angular space; it isn’t necessary, at this point, to reify a 2D visual field in which receptive fields are sub-regions.

⁶Thanks to Kevin Lande for pressing this point.

With these points in place, I propose to shift focus from retinotopy and visuotopy. Just as retinotopy may be considered an established neuroscientific fact, independent of the debate about its information processing function, we may consider visuotopy to be a purported fact established more or less by the same body of evidence. But just as before, the mere existence of visuotopy does not itself establish its relevance to information processing, which must be motivated by further considerations.

We can also now see that cortical retinotopy is an indicator of visuotopy, due to the principle of wiring optimization. Cortical proximity indicates proximity in functional space. And stable proximity mappings across visual episodes indicate the presence of the general metric transformations that sustain visuotopy. Thus, while I'll argue that retinotopic *areas* produce iconic representations, it is their visuotopy, not their retinotopy, which ultimately functions iconically.

6 The convolutional model of visual processing

6.1 From representation to computation

The proposal that visuotopic maps are iconic representations presupposes and understanding of these maps as complex spatial representations with simpler constituent parts. This assumption contrasts with the attitude that the only perceptual representations encoded in early vision are the constituent perceptual elements whose physical vehicles are arranged retinotopically. And much philosophical literature similarly focuses on unitary representations—be they percepts, feature bundles, or object files.

The focus on featural representation is complemented by a view of visual computation as essentially local: transformations that can be characterized in terms of the connections between individual nodes at a single location in cortical space. The paradigmatic example is the transition from simple to complex cells described by Hubel and Wiesel (1962), where a complex cell's response is explained by its connectivity to a handful of simple cells with overlapping receptive fields. Such local computations can be fully characterized without reference to the two-dimensional layout of the visuotopic map in which they are embedded. The 2D spatial organization of the map plays no essential role in explaining what any individual neuron computes; it is, from the perspective of local computation, merely an implementation detail.

In order to motivate the turn to global visuotopic representations, we must look to a correspondingly global computation: computation that takes whole visuotopic maps as input and produces whole visuotopic maps as output. It is at this level of description that the two-dimensional structure of the map becomes computationally relevant—not just as a container for individual feature detectors, but as a structured representational format that shapes how information is transformed from one processing stage to the next. In this section I'll argue that convolutional neu-

ral networks, an architecture for visual processing derived from contemporary machine learning, provides an account of visual computation which is at once plausible and suitably global. By definition, it takes whole feature maps as inputs, and produces feature maps as outputs. Likewise, I will argue that these global convolutional algorithms will substantiate the claim that global visuotopic maps are being *used* or *function* iconically. Thus, in what follows, we will shift our focus from individual visuotopic maps to the computations that they serve as input to.

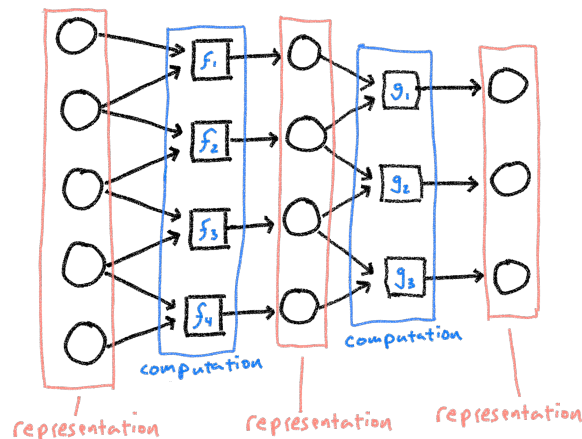


Figure 9: Visuotopic maps as representations. The computations of a feedforward neural network may be divided into representational and computational layers. Each node in a representational layer may store a first-order representation, but the set of nodes and a functional metric compose a second-order representation, which is the input to a computational layer.

6.2 Convolutional neural networks

Hierarchical Convolutional Neural Networks (CNNs) are one of the most successful computer vision architectures currently in use.⁷ CNNs were developed incrementally over many decades by computer scientists who were directly inspired by the hierarchical organization of the visual system, as reported by Hubel and Wiesel (1962). The original neurobiological insights were formalized in an abstract, even philosophical fashion by Fukushima (1980) and LeCun et al. (1998), before bursting into mainstream computer engineering with 2012’s AlexNet (Krizhevsky, Sutskever, and Hinton 2012). As we’ll see, modern CNNs exhibit a number of key parallels with the human visual system.

⁷The main competition comes from vision transformer models, which eschew much of the explicit visual structure built into CNNs. However, transformers make much of their relevant structure implicit, so that it is difficult to say, at present, just how similar CNNs and vision transformers are in their underlying mechanism.

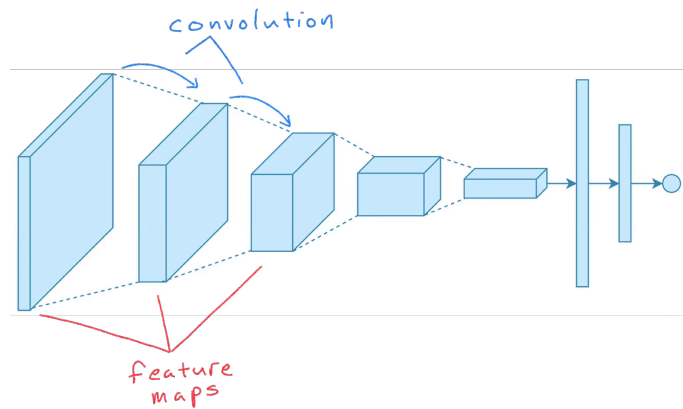


Figure 10: Convolutional neural network. A hierarchical convolutional neural network with four convolutional layers.

CNNs work by applying hierarchical **convolution** operations to a sequence of feature maps, where each convolution can be conceived as a transitional layer, connecting input and output layers. Thus, the input to a given convolutional layer is a **feature map**, or matrix, that covers the whole visual field with values. For example, in a simple digital implementation, a feature map might have the dimensions 255×255 pixels. Then a **kernel** (also known as a **filter**) of limited 2D size (e.g. 3×3) passes over each position on the feature map.

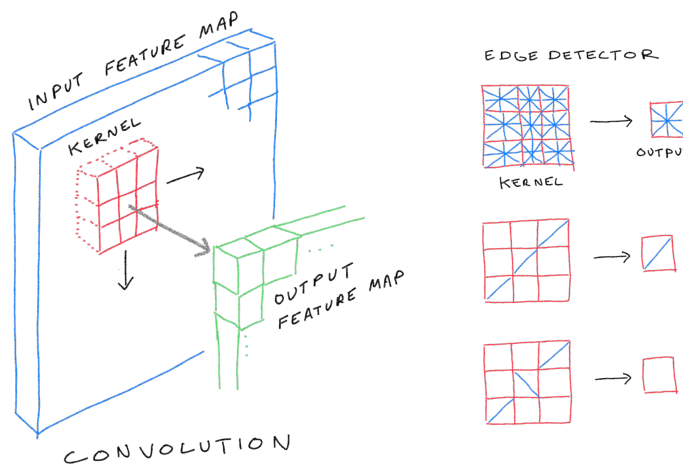


Figure 11: Convolution. Convolution with a 3×3 kernel.

For each position on the feature map, the kernel takes the values of the input cells and maps them onto an output value at a corresponding location in a new feature map. Typically, several kernels are applied to a region simultaneously, producing a corresponding stack (vector) of output

values (Prince 2023). The output of convolution, then, is a new feature map which typically has a smaller height and width than the input map but increases in depth.

On digital computers, convolution is performed serially, with the same set of instructions, corresponding to the kernel, applied to successive samples of the feature map matrix. Intuitively, this operation can be thought of as “sliding” the kernel across the input feature map, while iteratively generating the output map, piece by piece. Yet the same algorithm can be implemented in parallel as a feedforward circuit. Then the kernel corresponds to a computational function which is reduplicated in space, rather than time, across the feature map surface.

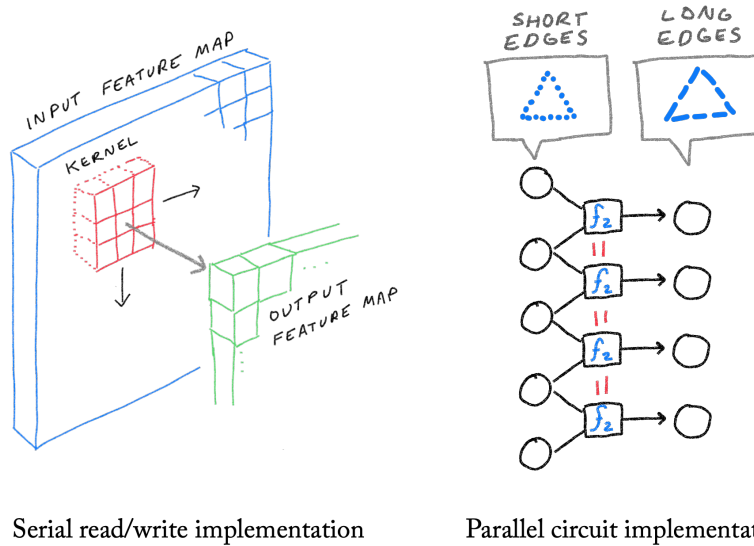


Figure 12: Serial and parallel implementations of a convolutional algorithm.

Keeping such parallel implementations in mind, we can see that a familiar simplified model of visual processing is already convolutional. On this model, a bank of transducers turns light stimuli into neural (or digital) signals; these signals are then transformed, in sequence, first into registration of luminance values, then simple features of the environment (such as short edges), then more complex features (such as long edges). Provided that each transformation is achieved by applying the same operations repeatedly across the visual field at each stage, then this kind of simple visual circuit is also a form of convolution.

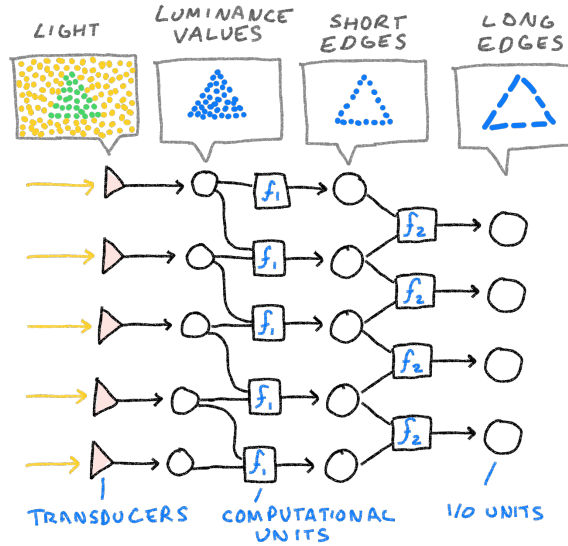


Figure 13: Stylized visual circuitry as convolutional.

6.3 Visual computation as convolution

We may now consider the conjecture that inter-map computations in early visual areas take the form of convolutions. This is a familiar but debated idea among vision scientists (Kriegeskorte 2015, Yamins and DiCarlo 2016, Bowers et al. 2023). I propose to endorse the hypothesis, which is supported by the remarkable success of convolutional models across a range of tasks, and by extensive parallels between CNN models and ventral stream processing at both global and local scales.

To begin, let us highlight five points of global analogy. (i) CNNs have achieved human-level categorization accuracy on large-scale image recognition tasks (Krizhevsky, Sutskever, and Hinton 2012). (ii) The layout of feature maps in CNNs is straightforwardly visuotopic, mirroring the apparent visuotopy of visual areas, and even predicting the spatial organization of contour maps in early visual cortex (**margalit2023unifying**). (iii) Receptive field size in CNNs increases along the hierarchy, matching a key observation about the ventral stream, with larger patterns and objects detected later in processing. (iv) Feature detection also increases in abstraction along the hierarchy, so that in both CNNs and the ventral stream, relatively low-level spatial properties are decodable from early layers, while high-level categorical properties are decodable from later layers (Zeiler and Fergus 2014). Finally, (v) an ongoing research program in computational neuroscience has uncovered robust correlations between activation patterns in early, middle, and late stages of CNNs and the corresponding stages of the ventral stream (Kriegeskorte 2015, Yamins and DiCarlo 2016, Zhuang et al. 2021).

At the local level, we find strong parallels between the internal workings of isolated convolutional layers and those of specific visual areas, as exemplified by the primary visual cortex V1. Consider three points of analogy. (i) In both cases, the same operation is applied uniformly over the visual field. In V1, for example, hypercolumns more or less homogeneously tile the visuotopic map; orientation and edge detection appear to work more or less the same across the visual field.⁸ (ii) Computational operations only take local inputs in each case, spanning at most a few cells (in the CNN) or a few hypercolumns (in V1) in each direction.⁹ (iii) Processing is hierarchical in both cases, where nodes that detect simple elements are mapped into nodes that detect complex elements.

I propose to take the model literally: layer-to-layer computations in human visual processing are convolutional. The layers in question may mark intra-area transitions, such as the transition from cortical layer 4 to cortical layer 2/3, within V1; but they may also belong to inter-area transitions, such as the transition from V1 to V2. Taking the model literally also requires transposing the serial implementation of convolution to a parallel feedforward circuit, of the kind characteristic of ventral stream processing. Then the kernel corresponds to a computational function which is reduplicated in space, rather than time, across the feature map surface. The kernel of the convolution in V1 can be identified with the hypercolumn, and idealization of V1's key repeating computational motif; more carefully, we may equate the regular distribution of the kernel with the repeating pinwheels of orientation columns (and color blobs) that tile cortical surface.

To be sure, valid concerns have been raised about the conjectured parallel between visual processing and hierarchical convolution (Bowers et al. 2023). For instance, CNNs have been “distracted” by irrelevant patterns in visual statistics that seem to have no correlate in vision, especially in the domain of feature-binding and object representation (REF). These are legitimate concerns, especially for later layers of visual processing, though there is countervailing evidence that even late layers of CNNs *can* manage key components of object representation (Hong et al. 2016). And there is always the concern that the kind of neural net under consideration is simply too crude or poorly trained for the task at hand. In any case, while acknowledging these worries, I think that they are less applicable to early stages of visual processing (V1, V2, V3v), so do not impact the present argument unduly.

A related set of concerns has to do with the biological plausibility of the conjecture. There are a number of notable architectural differences between ventral stream processing and any standard CNN. One has to do with back-projections; in the visual system, feed-back wiring outnumbers feed-forward wiring by manyfold, though the function of this feedback proliferation is still de-

⁸This generalization must be qualified for difference between foveal and peripheral vision, though it is still true, to my knowledge, of each separately (REF). There is also evidence for some differentiation of computation between upper and lower regions of the visual field, but not to an extent that it disturbs our core observation (REF).

⁹I'll use the term **hypercolumn** somewhat loosely, to refer both to single, monocular pinwheels of orientation columns, and to the binocular pairs of such pinwheels which share the same receptive field (Kandel et al. 2000).

bated (Kandel et al. 2000). In CNNs, on the other hand, there are no feedback connections.¹⁰ Another issue has to do with the apparent absence of parallel modularity in CNNs, compared with the extensive parallel modularity in the visual system. CNNs are modular in the sense of having distinct convolutional layers, but there is no evidence, at this point, for the kind of functional segregation we see in vision, such as distinct areas for motion and color perception.

In sum, while contemporary CNNs still only provide a simplified model of visual processing, a range of features do suggest that the basic architecture of computation in the visual system is indeed convolutional. And the evidence for this claim is perhaps strongest for early stages of visual cognition.

7 Convolution exploits projection semantics

The goal of this section is to argue that the content of visuotopic maps is given by a projection semantics, the same family of semantics that govern pictures, and a core exemplar of iconicity. For such a semantics to be used by a system, it must be exploited by the immediate downstream computation. So in this section, I'll aim to establish that the visual function of convolution requires a projection semantics. In informational terms, my claim is that projective information about the spatial environment is not merely available in visuotopic maps, but exploited by their immediate consumer computations.

Here is the plan for this section:

- **Why does convolution work?**
 - **“To work”**: to fulfill the function that promotes the well-being of the system.
 - **For visual computations:**
 - * **Spatial regularity**: Compute X from Y (e.g. shape from shading)
 - * **Inferential baseline**: Preserve truth
 - * **Inferential ideal**: Maximize completeness
 - **Spatial regularity:**
 - * **Definition**: A regularity in which X 's at location $loc(X)$ normally implies the existence of Y 's at a location related to $loc(X)$.
 - * **Example**: shape from shading, contour integration, face recognition.
- **Convolution as a spatial algorithm**
 - **Spatial aspects of the kernel:**
 - * Dimension

¹⁰Check the literature for RNNs used for vision tasks.

- * Internal locations
- **Spatial aspects of the convolution:**
 - * Spatial uniformity
 - * Location mapping (translational equivariance)
- **Spatial algorithm, meet spatial regularity:** The role of semantics: connecting spatial representations with spatial content.
- **Convolution exploits projection semantics**
 - **Projection as spatial alignment:** Projection supplies the spatial mapping from feature maps to visual space that would make convolution reliable.
 - **Projection as exploitable relation:** (1) Cortical space stands in a projective relation to visual space; (2) the spatial locations in cortical space inform consumer computation.
 - **Conditions on exploiting projection:** (1) The projection relation is exploitable; (2) convolution is a reliable algorithm for tracking spatial regularities if and only if the projection relation holds.
- **Contour integration as a case study**
 - Spatial regularity
 - Computation
 - Function
 - Semantics

Following are the subsections from an earlier draft of this section. The material gets increasingly obscure as you go along!

7.1 Why does convolution work?

If a visual field map is visuotopic, then, by definition, its 2D functional space is an approximately smooth and invertible transformation of angular space centered on a privileged viewpoint. For an embodied perceiver, visuotopy also implies a certain relationship to objects in the environment. We can think of this relationship as having two parts. First, object locations are projected to regions of the visual field; then region locations are transformed into locations in the visuotopic map. Then, as a descriptive generalization, each map stands in a particular relationship of transformed projection to objects in the environment. The spatial organization of visuotopic maps thereby carries correlational information about the spatial organization of objects in visual space.

But is this information actually functional for the consumer computational system? If it is, then the spatial content of a visuotopic map likewise depends on a projective transformation of its form. The abductive argument of this section is that the projective information supported by visuotopic maps is functional because the best explanation of convolutional computation depends

upon it. If functions are selected effects, then the relevant effect of tokening a visuotopic map is that it makes available to the consumer projective information about the environment.

7.2 A spatial algorithm

Let us suppose that key stages of early visual computation take the approximate form of convolutions. The next step in my overall argument is to show that visual computations of this kind make use of the functional 2D metric properties of its input maps. To make this argument concrete and precise, it will be necessary to lean somewhat heavily on the CNN model of early vision. I'll imagine that we are dealing with an organism whose visual system actually has the structure of CNN, and make the argument for 2D maps in this context. My aspiration is that this kind of idealized and simplified model is still sufficiently life-like to allow us to reasonably draw the same conclusions about biological retinotopy.

To begin, convolutional algorithms, by definition, take (layers of) two-dimensional matrixes as inputs and produce (layers of) two-dimensional matrixes as outputs. To see why these matrixes should be thought of as genuine 2D metric spaces, we must look beyond the input-output mappings of convolutional layers. I'll argue in this section that the convolutional algorithm itself uses the 2D metric properties of the input to compute the 2D metric properties of its output. The core operations of a convolutional algorithm are the iterated application of the kernel, and the movement of the kernel between operations. Key aspects of both operations are defined spatially.

Consider first a single application of an $n \times n$ kernel. In terms of computational capacity, a kernel is just a single layer neural network with m input nodes (where $m = n \times n$) and one output node. A generic neural network with this profile differentiates between each input, assigning each a unique weight; but they are only differentiated, as it were, by identity. Asked to apply the "same" network to a different set of m input values, there is no fixed way of knowing which input values go with which input node. Kernels are not generic in this sense, since they specify their inputs spatially. Each input is given a unique location in the $n \times n$ space. This kind of **position sensitivity** is a critical feature for the kernel's role in the overall convolutional computation. It means that a kernel which detects, say, edges at 45° angles, will perform the same spatialized function when applied to any other $n \times n$ array of cells from the feature map. This in turn means that similarly uniform operations can be applied to the output layer. Thus the kernel's function is sensitive to the relative 2D spatial position of its inputs, as embedded in the broader 2D space of the input feature map.

Kernels are not only position sensitive; they are also **local**, meaning that they only look at input nodes that are near one another in the feature map. Such locality is a defining feature of CNNs (and distinguishes them, for example, from transformers). Each node of the output map is determined by a small set of mutually local nodes in the input.

Next consider the pattern of kernel applications across the feature map. Once again, it is

helpful to consider the operation abstractly: we have a large collection of input values, which we wish to iteratively feed into a single layer network of limited input size. In what order should we present the inputs? Convolution, of course, follows a strict spatial order. The kernel begins in one corner of an $N \times N$ feature map, and then “slides” across the remainder of the map, where “sliding” is spatial translation by a given increment, leftward and downward, without rotation. Each of these constraints can only be imposed against a background 2D metric, far richer than the unordered collection of input values imagined above.

In a parallel implementation, as in the cortex, there is no mechanism that literally slides across the surface of the feature map. But we can observe functionally identical repetitions of the same mechanism across space. The same kind of kernel is applied uniformly over the 2D space: that is, the same function is applied to local regions that uniformly tile the input map. **Spatial uniformity**, then, is a third signature of convolution, definable only by assuming a metric on the input map.

Finally, each application of the kernel to the input map produces a value at the “same location” on the output map, a property I’ll call **location mapping**. Location mapping is the essential operation which allows convolutional computations not only to build an output map from an input map, but to build an output map which stands in perfect spatial registration with the input map. Location mapping, on its face, seems crucially to involve reference to spatial location, but this contribution is obscured, to a degree, in the standard “layer cake” visualization of convolutional layers, as in Figure X. This kind of diagram makes the alignment of input and output layers appear inevitable, exploiting the 2D metric of the page and thereby obscuring the algorithm’s reliance on functional metric properties.

Instead, I find it helpful to visualize the mapping problem more or less as it is faced in the cortex itself, with the two layers represented side by side. Then we can see that there are any number of ways of mapping inputs from A to outputs in B . In formal implementations of CNNs, straightforward location mapping follows from an implicit coordinate system for input and output layers. Of course, in the cortex, there are no intrinsic coordinates, so, strictly speaking, the mapping algorithm doesn’t compare locations. Instead, we can say that (in CNNs and cortex) the computational mapping from A to B is approximately smooth and invertible, so that locations (and distances) in A can be mapped to locations (and distances) in B and back again. Implementing a particular such transformation is part of the function of the convolutional computation.

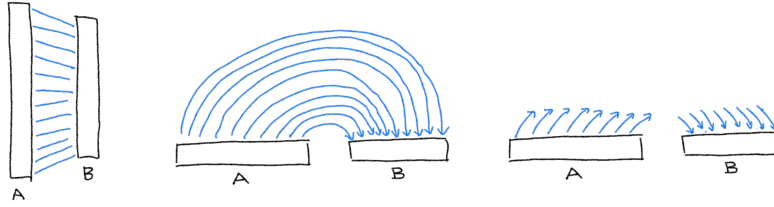


Figure 14: Visualizations of location mapping.

In sum, through position sensitivity, locality, spatial uniformity, and location mapping we can see that the convolutional algorithm is essentially defined over an input endowed with 2D metric structure. This, I suggest, shows that, in a visual system based on convolution, functional 2D metrics are actually put to use by the immediate downstream computation. By parallel reasoning, we should expect the same of the functional 2D metric realized by retinotopic areas.

7.3 Projection semantics

A semantics based on projection can be motivated by the skeletal architecture of convolution, regardless of how the kernel is implemented. In the next section, I'll show how the same lessons arise for the internal design of the kernel.

According to the classic “tower-bridge” model of computation, computation is essentially conceived as operating at two levels (REF: Cummins, Turing). At a lower, syntactic level, we have a physically instantiated computational system, which executes an algorithm. It takes structured representations as inputs and produces structured representations as outputs. At a higher, semantic level, we have a mapping between input and output contents, and it is this mapping that a computer functions to compute *by* executing an algorithm.¹¹ Semantics, in this setting, are the rules that determine contents from representations, for both input and output. A machine successfully computes a mapping, given a semantics, if its inputs and outputs, under the semantics, match those of the target mapping.

Applied to the present case, we have, at the lower algorithmic level, a convolutional algorithm that takes functional 2D maps as both inputs and outputs. At the higher, semantic level, we have the contentful mapping that it computes. These kinds of computations are described informally as “computing shape from shading,” or “computing depth from relative size.” As a coarse generalization, these mappings tend to relate arrangements of F s in visual space to arrangements of G s in visual space, for various F s and G s. To explain how the lower level convolution realizes the higher level computational function, semantic rules map the 2D space of the visuotopic maps to attributions of F s and G s. We may then ask what kinds of semantics this sort of computational

¹¹I'm going to use the term “mapping” to designate what logicians and mathematicians usually call “functions,” since the latter is overburdened in the present context.

function requires.

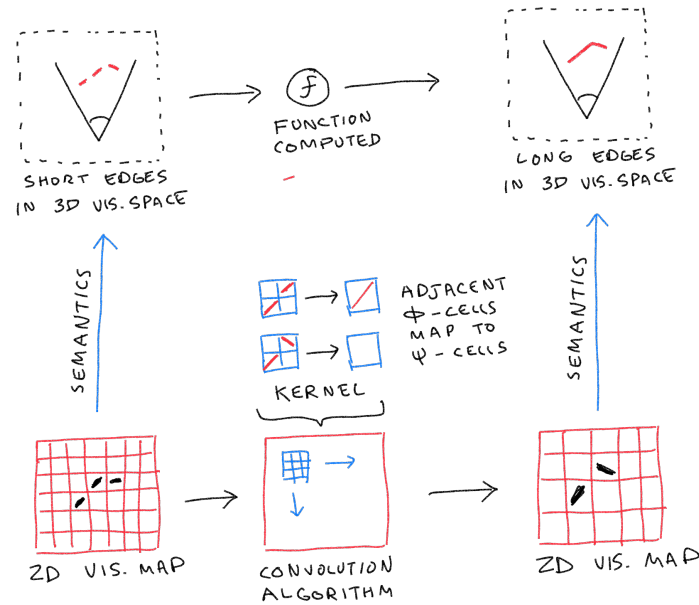


Figure 15: The role of semantics in visual computation.

In a standard perceptual case, let us assume, computational layers have an ür -function of producing accurate and complete representations of a given set of features. And it achieves this, in part, by functioning to compute an accuracy-preserving mapping from visual contents to visual contents. Given the inverse nature of vision, the system must leverage environmental regularities in order to make reliable inferences possible. [Note that truth-preservation is an idealized case, and probabilification will usually be sufficient.]

We have to explain why the convolutional algorithm is truth-preserving. To sharpen the problem, it helps to focus on specific aspects of the algorithm that make use of the functional space of the input. Recall the property of locality, associated with convolutional networks: each output node is determined by a set of input nodes that are near one another, in the functional space of the input. But why is a local algorithm suited to be truth-preserving? For some kinds of representations and semantics, it is clearly not. Suppose we attempted to apply a local convolution to a series of logical formulae (arranged in a 2D matrix), with the aim of making truth-preserving inferences. Our machine would be a failure because spatial locality is not, in general, relevant to the semantics or logic of symbolic representations.

For locality to be truth-preserving, it has to be the case that nearby nodes carry information about F s that is evidentially relevant to the presence of G s. Since we are dealing with visual space, and the identity of visual events generally depends on their local parts, we know that the kind

of F-information that is relevant to G's is spatial adjacency. So making sense of locality requires a semantics that smoothly transforms adjacency in the VM's to adjacency in the represented visual space. Of course, this is just what a semantics based on projective transformation yields.

[Similar lines of thought will be sketched for other spatial properties of convolution: uniformity and location mapping.]

Shea (2018) gives us the useful notion of an **exploitable relation**. For a system to produce a representation with a given content, it must exploit a relationship between representations and environmental variables in order to fulfill its function. In the case of iconic functions, it is relations of natural dependency that are exploited. And in the specific case of visuotopic maps, it is a relation of projective transformation to the environment. I made the case earlier that this information is available to the convolutional computer; we now see one way in which it is used.

7.4 Edge detection: a case study

The motivations for projection semantics outlined above are recapitulated at the level of the kernel. The argument can be made more concrete by focusing on a case study. Here I will continue down the path of idealization by picking out one, very simple function: that of computing long oriented edges from short oriented edges.¹² Something like this is at least part of the function of the transition from simple to complex cells in V1. If this basic visual computation is sufficient to establish the iconicity of visuotopy, then, I hope, the same conclusions may be extended to more naturalistic and complex visual computations.

So we want to say that a computational layer C computes long edges from short edges. To make this statement more explicit, we can say that C functions to compute f , where f is a mapping from visual contents to visual contents. In particular, f is a mapping from a visual content V that locates short edges in the environment to a visual content U that locates long edges in the environment.

- (5) $f(U) = V$ iff V is just like U except that, for every pair of short edges that are angularly adjacent and collinear in the angular space of U , there is a long edge that occupies the same relative location in the angular space of V .

Normally, we expect that C will only function to compute f if doing so yields a positive outcome for the organism. And an important positive outcome for many kinds of computation is preserving accuracy; preserving accuracy and completeness may be even more valuable. (And of course, it will often suffice for the input to merely probabilify the output.) In our hypothetical environment, we might discover that f is often truth-preserving because the following two environmental generalizations typically hold.

- (6) **Environmental regularity 1 (E1):**

¹²Cf. Lande 2023.

Given a viewpoint v in physical space, if short edges are angularly collinear and adjacent in the angular space defined by v , then they are collinear and adjacent in physical space.

(7) **Environmental regularity 1 (E2):**

If short edges are collinear and adjacent in physical space, then they are parts of a single long edge in physical space at the same location.

As the reader may verify, if E1 and E2 both hold, and if U is accurate, then $f(U)$ is also accurate. In this sense, C leverages or implicitly encodes the environmental regularities E1 and E2. They play an essential part of the explanation for why C computes f , by fulfilling the necessary conditions for C to produce accurate representations of long edges (cf. Johnson 2020).

C computes long edges from short edges. And, by hypothesis, it achieves this by convolution. We may further assume that the kernel k has the stereotyped behavior you might expect for a task like this. Roughly: if adjacent cells in the input matrix contain the appropriate orientation features, then the output value contains the same orientation feature at the corresponding position of the output matrix.

Imagining ourselves as scientists discovering the mechanisms of C , we may ask: why is convolution with k a reliable way to compute f ? And by extension: why is convolution with k a way of implementing an accuracy-preserving inference? To answer these questions, we must specify the semantics, or informational function, that associates the inputs and outputs to C with environmental conditions.

By now it is clear that *if* we assume that the inputs and outputs to C are related by projective transformation, then it follows that applying k is truth preserving. The following reasoning, somewhat idealized from the case, illustrates the point. I use $[x]$ to refer to feature representations contained in node x .

- (8) kernel k : if $[x]=1$ and $[y]=1$ and x and y are adjacent in M , then output $[o]=1$.
- (9) Let x and y be nodes in input map M :
 - a. Suppose $[x]=1$ and $[y]=1$
 - b. Suppose x and y are adjacent in M .
 - c. Assume projection semantics, $[[x]]$ and $[[y]]$ are angularly adjacent in visual space.
 - d. By E1 and E2, there is an edge in visual space.
 - e. By covariational semantics for o , there is an edge in visual space.
 - f. So k is truth-preserving.

Applying k in a single instance instructs us to search for appropriate adjacent nodes in the input. By projective transformation, activity in these nodes covaries with the adjacent angular positions of short edges, so it carries information about them. The kernel then produces an activation

of the output node that functions to covary with the presence of an appropriate long edge. The kernel’s algorithmic mapping can only harness the force of the environmental regularities E1 and E2, to compute function f , if the semantics for the feature map relate locations in VM to angular locations in visual space, just as projection semantics does.

8 Pictorial semantics for visuotopic maps

I now wish to outline a semantics for visuotopic maps that is descriptively plausible within the confines of our idealized case, and answers to the theoretical role outlined above. It is this kind of semantics that is required in order for basic visual computations to perform their functions.

A semantics is a mapping from representations to contents. In the present case, the representations are visuotopic maps, consisting of a matrix of nodes arranged in a 2D functional metric, with particular activation values (or features) stored at each node. And the contents are visual spaces, abstract structures which locate objects and features in directional space, centered on a viewpoint. I’ll provisionally model such visual contents as sets of world-viewpoint pairs, w, v .

The analytic strategy pursued here is inspired by a projection-based approach to the semantics for pictures that I’ve defended elsewhere (Greenberg 2021). The key idea there is that, for a picture to be accurate at a world and viewpoint, it must be a projection of that world from that viewpoint; to interpret a picture is to work backwards to the set of conditions that must have projected it. The view of visuotopic maps taken here doesn’t depend on the validity of this way of thinking about picture semantics, but the convergence is suggestive. In the present case, the rough idea is that, for a visuotopic map to be accurate at a world and a viewpoint, its functional 2D shape must stand in a relation of transformed projection to the world from that viewpoint.

I’ll begin with an outline of the semantics, and then explain its various components. For any visuotopic map M :

- (10) $\llbracket M \rrbracket$ = the set of $\langle w, v \rangle$ such that:
for every node n in V , there is an object o in w such that:
- a. $proj(o, v) = r$ and $transform(r) = n$;
 - b. o has the property $\llbracket n \rrbracket$ at w, v }.

I focus first on clause (a), which specifies the location of the quantified object o . Let us assume that, when all aspects of the perceptual apparatus are functioning properly, there is an idealized relation of geometrical projection which holds between worlds and viewpoints, on one hand, and regions of the visual field, on the other. We can say that the projection of an object, at a world, yields a subregion of the visual field, formalized as $proj(o, w, v) = r$.

In addition, let us assume (per the definition of visuotopy) that there is a smooth transformation from the visual field to the 2D surface of the visuotopic map. This transformation like-

wise takes regions of the visual field to regions of the visuotopic map; in the extreme, small regions of the visual field are transformed to individual nodes of the map. In this case, I'll write $transform(r) = n$, where n is (a region containing) an individual node. Note that these relations aren't causal; rather they are the formal relationships that the causal processes of optical projection and neural transformation are supposed to realize. Putting these elements together, (a) specifies the location of an object o as that angular region in w, v that projects to a region that transforms to the location of n .

Clause (b) specifies the features attributed to the quantified object o . Here I will just stipulate a simple semantic mapping, $\llbracket \cdot \rrbracket$, which associates internal feature representations with external features. (In a formal treatment, we may model external features and relations as intensions.) Thus, in a basic scheme, we might associate activation values 0, 1 with the properties *edge*, *non-edge*; in a richer scheme, we might associate real values between 0 and 1 with edges at a range of orientations; and so on. These feature representations are, in effect, a kind of lower-order representation which are the constituents of the higher-order structural representation that is the visuotopic map itself. For a given map M and a given node n that is part of M , we can say that n **contains** a set of features. In clause (b) I use the expression $\llbracket n \rrbracket$ to refer to the (intersected) contents of the features contained in n .

This semantics, while admittedly schematic, gives rough shape to the present proposal. It is at least explicit enough to support the following observations.

First, this semantics is clearly iconic in my preferred sense. This is because the content of the visuotopic map, defined above as a set of accuracy conditions, is determined by a natural relation of projection plus transformation to the functional 2D form of the map itself. In particular, it is the spatial content of the map— where it locates represented objects and properties— that depends on the spatial form of the map itself.

Second, the projection semantics for visuotopic maps is a second-order rule: it specifies the content of representation types, which are made up of spatial arrangements of constituent representations. This can be seen from the fact that the content of the map, $\llbracket M \rrbracket$ depends in part on the featural content of each of its nodes, $\llbracket n \rrbracket$. The claim of iconicity should be correspondingly restricted. As far as the argument of this paper is concerned, the semantics for visuotopic maps are second-order iconic; the iconic status of the first-order rule is not at issue here, but the question is closely related to cases of perceptual iconicity that have been discussed in the literature (Green and Quilty-Dunn 2021, Block 2023, cf. Clarke 2022).

8.1 Alternative and deviant semantics

- (Each configuration of nodes is symbolically associated with a full set of visually directions.)
- Each node is symbolically associated with a specific visual direction - Implies each node has its own function, not derived from the function of the whole.
- Different sizes of cortex, different

9 The first pictures

The preceding discussion suggests that retinotopic visual areas have the function of tokening representational states whose functional 2D metrics are smoothly transformed projections of 3D regions in visual space. As a consequence, the visual content of a visuotopic map depends, via a relation of transformed projection, on the map's functional form. By present lights, then, visuotopic representational systems clearly exemplify iconicity.

From these points I also conclude that visuotopic maps are genuinely picture-like representations. Of course, visuotopic maps and pictures differ at a basic level, since one derives its content by being viewed, while the other does not. Nevertheless, it is part of the essence of pictorial representation that it is based on a semantics that encodes 3D visual content in 2D signs via perspective projection from an implicit viewpoint (Greenberg 2021). This expressive strategy, exploited by photographs, drawings, and maps alike, marks an extremely efficient way of storing 3D information, and 2D surfaces are relatively easy to come by, and easy to inscribe. Retinotopic areas seem to store the same kind of 3D information in 2D surfaces, via the same kinds of sign-to-world relations. The cortex provides a ready-to-hand 2D surface, and projection offers a precise and efficient method for storing 3D information in two dimensions.

At a high level, we can see artists and nature as trying to solve the same problem: how to encode rich spatial information about a 3D world on a 2D surface? And both arrive at roughly the same solution: via projection of the 3D world onto a 2D surface, relative to a viewpoint. We may be tempted to think that visuotopic maps are not just iconic and picture-like, they are the first pictures.

References

- Arcaro, M. J. and M. S. Livingstone (2017). "Retinotopic organization of scene areas in macaque inferior temporal cortex". In: *Journal of Neuroscience* 37.31, pp. 7373–7389.
- Beck, J. (2015). "Analogue magnitude representations: A philosophical introduction". In: *The British Journal for the Philosophy of Science*.
- (2018). "Analog mental representation". In: *Wiley Interdisciplinary Reviews: Cognitive Science* 9.6, e1479.
- Block, N. (2023). *The Border Between Seeing and Thinking*. New York, US: Oxford University Press.
- Bowers, J. S. et al. (2023). "Deep problems with neural network models of human vision". In: *Behavioral and Brain Sciences* 46, e385.
- Brewer, A. A. and B. Barton (2012). "Visual field map organization in human visual cortex". In: *Visual Cortex: Current Status and Perspectives*. Ed. by S. Molotchnikoff and J. Rouat. Rijeka, Croatia: InTech, pp. 29–60.

- Buehler, D. (2024). "The Priority Map". In: *Australasian Journal of Philosophy*, pp. 1–26. DOI: 10.1080/00048402.2024.2412243. URL: <https://doi.org/10.1080/00048402.2024.2412243>.
- Burge, T. (2018). "Iconic Representation: Maps, Pictures, and Perception". In: *The Map and the Territory: Exploring the Foundations of Science, Thought and Reality*. Ed. by W. Shyam and F. A. Dorio. Springer.
- (2022). *Perception: First Form of Mind*. New York, NY: Oxford University Press.
- Cao, R. and D. L. Yamins (2024). "Explanatory models in neuroscience, Part 1: Taking mechanistic abstraction seriously". In: *Cognitive Systems Research* 87, p. 101244.
- Carey, S. (2009). *The Origin of Concepts*. New York: Oxford Academic.
- Chklovskii, D. B. and A. A. Koulakov (2004). "Maps in the brain: what can we learn from them?" In: *Annu. Rev. Neurosci.* 27.1, pp. 369–392.
- Chklovskii, D. B., T. Schikorski, and C. F. Stevens (2002). "Wiring optimization in cortical circuits". In: *Neuron* 34.3, pp. 341–347.
- Clarke, S. (2022). "Mapping the visual icon". In: *The Philosophical Quarterly* 72.3, pp. 552–577.
- Durbin, R. and G. Mitchison (1990). "A dimension reduction framework for understanding cortical maps". In: *Nature* 343.6259, pp. 644–647.
- Fukushima, K. (1980). "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position". In: *Biological cybernetics* 36.4, pp. 193–202.
- Green, E. J. and J. Quilty-Dunn (2021). "What is an Object File?" In: *British Journal for the Philosophy of Science* 72.3, pp. 665–699. DOI: 10.1093/bjps/axx055.
- Greenberg, G. (2013). "Beyond Resemblance". In: *Philosophical Review* 122.2, pp. 215–287. DOI: 10.1215/00318108-1963716.
- (2021). "Semantics of Pictorial Space". In: *Review of Philosophy and Psychology* 1.4, pp. 847–887. DOI: 10.1007/s13164-020-00513-6.
- (2023). "The iconic-symbolic spectrum". In: *Philosophical Review* 132.4, pp. 579–627.
- (MS). "The Struture of Visual Content". In.
- Hong, H. et al. (2016). "Explicit information for category-orthogonal object properties increases along the ventral stream". In: *Nature neuroscience* 19.4, pp. 613–622.
- Hubel, D. H. and T. N. Wiesel (1962). "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex". In: *The Journal of physiology* 160.1, p. 106.
- Johnson, G. M. (2020). "The structure of bias". In: *Mind* 129.516, pp. 1193–1236.
- Kandel, E. R. et al. (2000). *Principles of neural science*. Vol. 4. McGraw-hill New York.
- Kosslyn, S. M. (1981). "The medium and the message in mental imagery: A theory." In: *Psychological review* 88.1, p. 46.
- (1994). *Image and brain: The resolution of the imagery debate*. MIT press.
- Kosslyn, S. M., W. L. Thompson, and G. Ganis (2006). *The case for mental imagery*. Oxford University Press.
- Kriegeskorte, N. (2015). "Deep neural networks: a new framework for modeling biological vision and brain information processing". In: *Annual review of vision science* 1.1, pp. 417–446.
- Krizhevsky, A., I. Sutskever, and G. E. Hinton (2012). "Imagenet classification with deep convolutional neural networks". In: *Advances in neural information processing systems* 25.
- Lande, K. J. (2023). "Contours of Vision: Towards a Compositional Semantics of Perception". In: *The British Journal for the Philosophy of Science*. Advanced online publication. DOI: 10.1086/725094. URL: <https://doi.org/10.1086/725094>.
- LeCun, Y. et al. (1998). "Gradient-based learning applied to document recognition". In: *Proceedings of the IEEE* 86.11, pp. 2278–2324.

- Margalit, E. et al. (2024). “A unifying framework for functional organization in early and higher ventral visual cortex”. In: *Neuron* 112.14, pp. 2435–2451.
- Neander, K. (2017). *A mark of the mental: In defense of informational teleosemantics*. MIT press.
- Peacocke, C. (1992). *A study of concepts*. The MIT Press.
- Prince, S. J. (2023). *Understanding deep learning*. MIT press.
- Pylyshyn, Z. W. (2003). *Seeing and Visualizing: It’s Not What You Think*. Bradford.
- (2007). *Things and Places: How the Mind Connects with the World*. MIT Press.
- Quilty-Dunn, J. (2020). “Perceptual pluralism”. In: *Noûs* 54.4, pp. 807–838.
- Schira, M. M. et al. (2010). “Modeling magnification and anisotropy in the primate foveal confluence”. In: *PLoS computational biology* 6.1, e1000651.
- Schwarzlose, R. (2021). *Brainscapes: The Warped, Wondrous Maps Written in Your Brain-and how They Guide You*. Mariner Books.
- Shea, N. (2018). *Representation in cognitive science*. Oxford University Press.
- Tootell, R. B., M. S. Silverman, et al. (1982). “Deoxyglucose analysis of retinotopic organization in primate striate cortex”. In: *Science* 218.4575, pp. 902–904.
- Tootell, R. B., E. Switkes, et al. (1988). “Functional anatomy of macaque striate cortex. II. Retinotopic organization”. In: *The Journal of Neuroscience* 8.5, pp. 1531–1568.
- Tye, M. (1991). *The Imagery Debate*. Cambridge: MIT Press.
- Wandell, B. A., S. O. Dumoulin, and A. A. Brewer (2007). “Visual field maps in human cortex”. In: *Neuron* 56.2, pp. 366–383.
- Yamins, D. L. and J. J. DiCarlo (2016). “Using Goal-Driven Deep Learning Models to Understand Sensory Cortex”. In: *Nature Neuroscience* 19.3, pp. 356–365. DOI: 10.1038/nn.4244. URL: <https://www.nature.com/articles/nn.4244>.
- Zeiler, M. D. and R. Fergus (2014). “Visualizing and understanding convolutional networks”. In: *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part I* 13. Springer, pp. 818–833.
- Zeki, S. (1993). *A vision of the brain*. Blackwell scientific publications.
- Zhuang, C. et al. (2021). “Unsupervised neural network models of the ventral visual stream”. In: *Proceedings of the National Academy of Sciences* 118.3, e2014196118.