Perception: Psychophysics and Modeling

09 | Object recognition III

Felix Wichmann

Neural Information Processing Group
Eberhard Karls Universität Tübingen

Overview

The Problems of Perceiving and Recognising Objects **(VL07-Object Recognition 1)**

Mid-level vision **(VL07-Object Recognition 1)**

- What are ‘‘edges’’ and (illusionary) ‘‘contours’’?
- Gestalt psychology and ‘‘Gestalt laws’’ of perceptual organisation

More on *mid-level vision* **(VL08-Object Recognition 2)**

- Accidental viewpoint and non-accidental features
- Figure-ground, occlusion, wholes and parts
- Texture segmentation, grouping and camouflage

Neuroscience of object recognition **(VL08-Object Recognition 2)**

Object representation **(VL09-Object Recognition 3)**

- Structural description models
- View-based models

Object recognition by algorithms: DNNs **(VL10-Object Recognition 4)**

Overview

The Problems of Perceiving and Recognising Objects **(VL07-Object Recognition 1)**

Mid-level vision **(VL07-Object Recognition 1)**

- What are ‘‘edges’’ and (illusionary) ‘‘contours’’?
- Gestalt psychology and ‘‘Gestalt laws’’ of perceptual organisation

More on *mid-level vision* **(VL08-Object Recognition 2)**

- Accidental viewpoint and non-accidental features
- Figure-ground, occlusion, wholes and parts
- Texture segmentation, grouping and camouflage

Neuroscience of object recognition **(VL08-Object Recognition 2)**

Object representation **(VL09-Object Recognition 3)**

- Structural description models
- View-based models

Object recognition by algorithms: DNNs **(VL10-Object Recognition 4)**

Literature

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

DiCarlo, J. J., Zoccolan, D., and Rust, N. C. (2012). How does the brain solve visual object recognition? *Neuron*, 73(3): 415–434.

Supplementary Literature

Bülthoff, H.H. and Edelman, S. (1992). Psychophysical support for a two-dimensional view interpolation theory of object recognition. *Proceedings of the National Academy of Sciences*, 89: 60–64.

Gauthier, I., Curran, T., Curby, K. M., and Collins, D. (2003). Perceptual interference supports a non-modular account of face processing. *Nature Neuroscience*, 6(4): 428–432.

Kanwisher, N. G. and Yovel, G. (2006). The fusiform face area: a cortical region specialized for the perception of faces. *Philosophical Transactions of the Royal Society of London B*, 361: 2109–2128.

Rosch, E. H. (1973). Natural categories. *Cognitive Psychology*, 4(3):328–350.

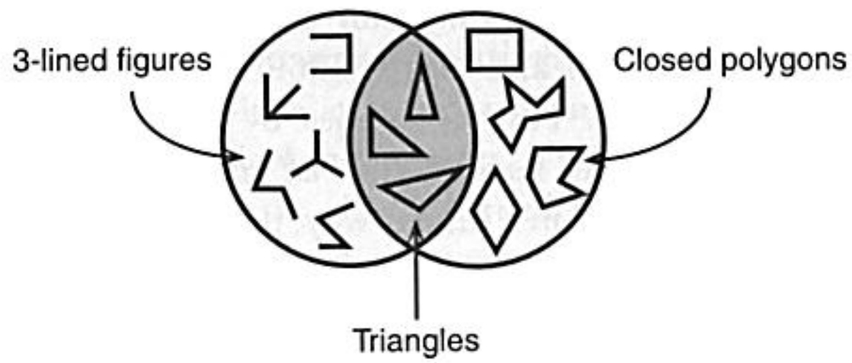
Tarr, M.J. and Pinker, S. (1989). Mental rotation and orientation-dependence in shape recognition. *Cognitive Psychology*, 21: 233–282.

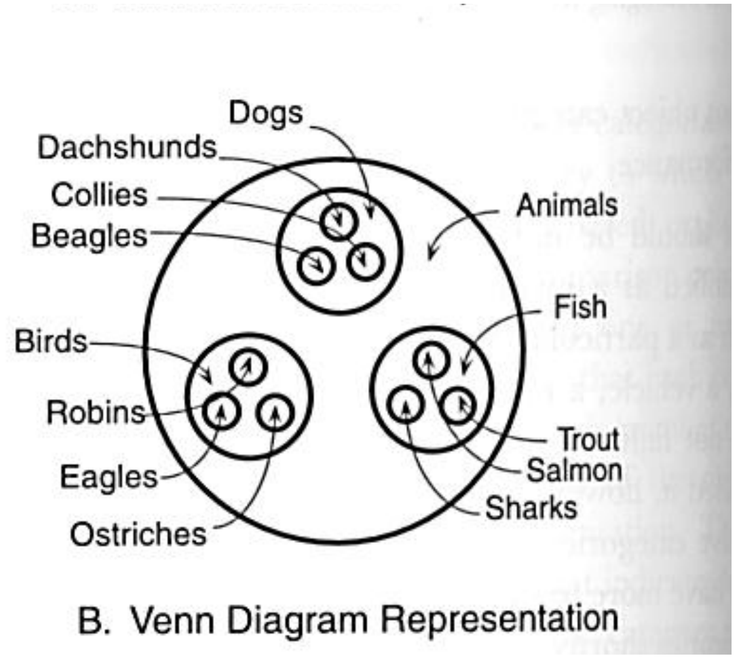
What do we mean by “object recognition”?

Object recognition is about categorising perceptual experiences. Because the same thing in the world can belong to multiple categories, this is by nature a hierarchical problem.

Aristotle: categories are defined by the necessary and sufficient conditions for membership

While logically true, is this a good theory of perceptual categories? Can dogs, fish and trees be characterised by necessary and sufficient conditions and binary category membership?





From Palmer (1999)

What do we mean by “object recognition”?

Wittgenstein (1953) argued persuasively that there is no such set of necessary and sufficient features that apply to natural categories.

He instead used the analogy of family resemblance: members of the same family look similar not because of particular features they share, but through global similarities that cannot be captured in simple logical rules.

What do we mean by "object recognition"?

Psychologist Eleanor Rosch conducted a series of studies (1973, 1975) that radically changed how cognitive scientists thought about categories, by defining **prototypes**: the "best example" of each category.

In most cases the prototype is the "average" member (the doggiest dog), but sometimes it is an "ideal" (the reddest red)—neither language nor most vision and cognitive scientists are often clear on the difference!

Prototype theories allow for graded category membership as distance to the prototype.

Category levels

Rosch had people rate category members as how ‘‘good’’ they were as examples for that category. For example, beagles are rated as ‘‘good’’ examples of dogs, but chihuahuas are not. Similarly, robins and sparrows are good examples of birds, but ostriches and penguins are not.

These ratings predict how quickly people respond to true / false judgments of verbal statements like ‘‘A robin is a bird’’ verses ‘‘A penguin is a bird’’. (similarly for picture classification).

Category levels

At what level of the hierarchy do people first / ‘‘naturally’’ classify objects?

Category levels

Rosch's experiments identified three levels of object categories,

Basic-level category: the intermediate level of hierarchy —e.g. *dog* | *table*. The highest level category for which members have similar shape, similar motor interaction and common attributes.

Subordinate-level category:
A more specific term for an object—e.g. *dalmatian* | *dining table*

Superordinate-level category:
A more general term for an object.—e.g. *animal* | *furniture*

Category levels

Entry-level category:
For an object, the label that comes to mind most quickly when we identify the object – *irrespective* of whether it's a "subordinate" or "basic" level category.

Basic level categories are defined for the entire category, but Jolicoeur, Gluck & Kosslyn (1984) showed that this depends on the category member.

What do we mean by "object recognition"?

What is generally called "object recognition" usually refers to the classification of objects into entry-level categories.

Some more precise definitions include

Object recognition: realising that you have seen a given object before regardless of whether you can name the object (i.e. object memory)

Object identification: recognising a particular known object ("my cat")

Object categorisation or classification: classifying objects into entry-level categories

... but these are not universally acknowledged

Recognition-by-components (RBC)

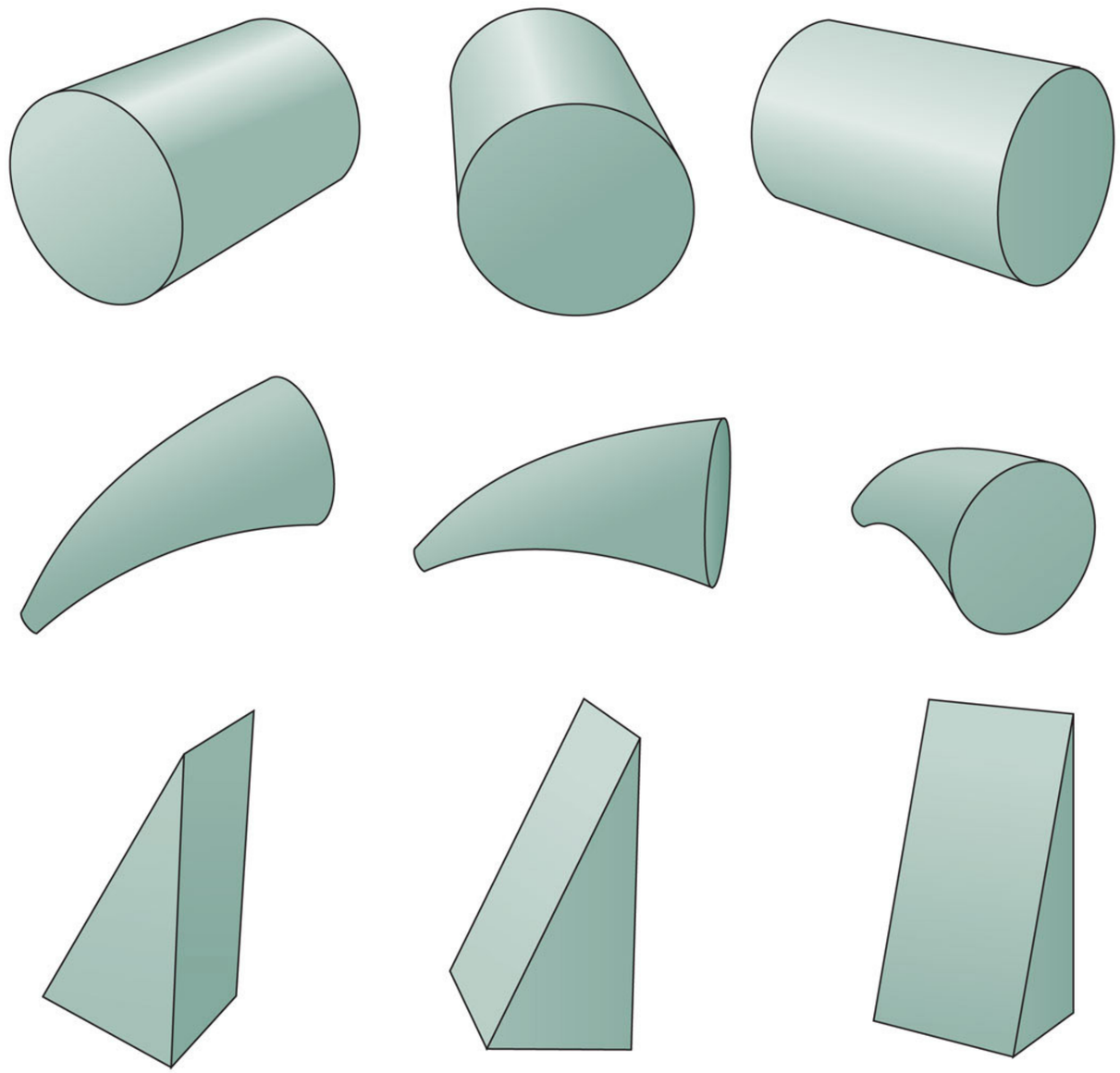
Biederman's model of object recognition: Holds that objects are recognised by the identities and relationships of their component parts.

Biederman's ideas are in a tradition popularized by Marr (1982), initially put forward in Marr & Nishihara (1978).

A ''four-stage'' theory of vision: image-based, surface-based, object-based and category-based processing.

Recognition-by-components (RBC)

Geons: The ‘‘geometric ions’’ out of which objects are built.



***SENSATION & PERCEPTION 4e*, Figure 4.42**

Psychological Review
1987, Vol. 94, No. 2, 115–147

0033-295X/87/$00.75

Recognition-by-Components: A Theory of Human Image Understanding

Irving Biederman
State University of New York at Buffalo

The perceptual recognition of objects is conceptualized to be a process in which the image of the input is segmented at regions of deep concavity into an arrangement of simple geometric components, such as blocks, cylinders, wedges, and cones. The fundamental assumption of the proposed theory, recognition-by-components (RBC), is that a modest set of generalized-cone components, called geons ($N \leq 36$), can be derived from contrasts of five readily detectable properties of edges in a two-dimensional image: curvature, collinearity, symmetry, parallelism, and cotermination. The detection of these properties is generally invariant over viewing position and image quality and consequently allows robust object perception when the image is projected from a novel viewpoint or is degraded. RBC thus provides a principled account of the heretofore undecided relation between the classic principles of perceptual organization and pattern recognition: The constraints toward regularization (Pragnanz) characterize not the complete object but the object's components. Representational power derives from an allowance of free combinations of the geons. A Principle of Componential Recovery can account for the major phenomena of object recognition: If an arrangement of two or three geons can be recovered from the input, objects can be quickly recognized even when they are occluded, novel, rotated in depth, or extensively degraded. The results from experiments on the perception of briefly presented pictures by human observers provide empirical support for the theory.

Biederman's analogy between speech and object recognition

In speech perception, phonemes form basic primitives that can be combined to generate a vast number of words.

For example, only about 55 phonemes are needed to represent all words in all spoken languages in the world.

The hypothesis explored here is that a roughly analogous system may account for our capacities for object recognition. In the visual domain, however, the primitive elements would not be phonemes but a modest number of simple geometric components—generally convex and volumetric—such as cylinders, blocks, wedges, and cones. Objects are segmented, typically at regions of sharp concavity, and the resultant parts matched against the best fitting primitive. The set of primitives derives from combinations of contrasting characteristics of the edges in a two-dimensional image (e.g., straight vs. curved, symmetrical vs. asymmetrical) that define differences among a set of simple volumes (viz., those that tend to be symmetrical and lack sharp

not taxed. The particular properties of edges that are postulated to be relevant to the generation of the volumetric primitives have the desirable properties that they are invariant over changes in orientation and can be determined from just a few points on each edge. Consequently, they allow a primitive to be extracted with great tolerance for variations of viewpoint, occlusion, and noise.

Biederman, 1987, p. 116

Some Nonaccidental Differences Between a Brick and a Cylinder

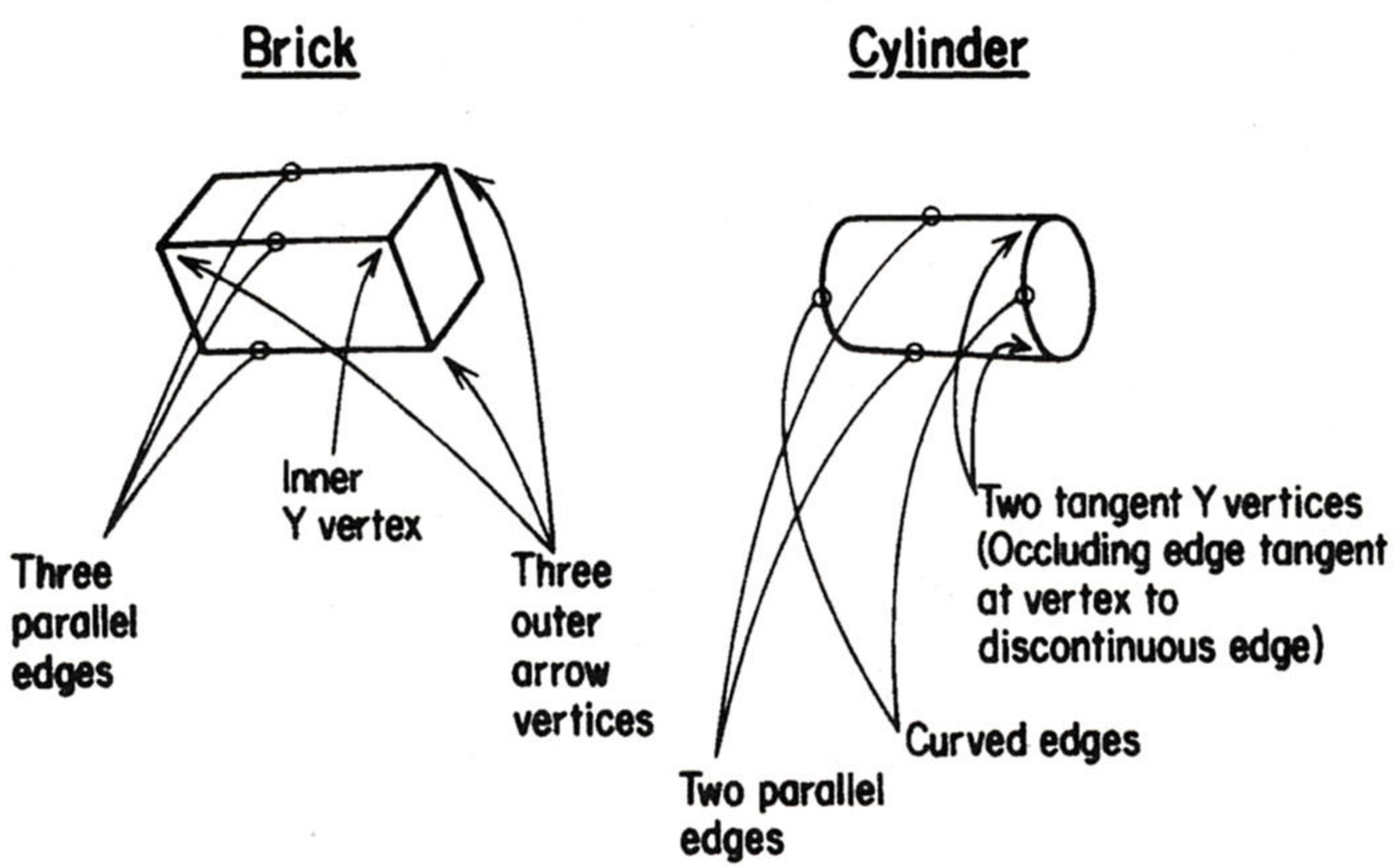


Figure 5. Some differences in nonaccidental properties between a cylinder and a brick.

Biederman, 1987, p. 120

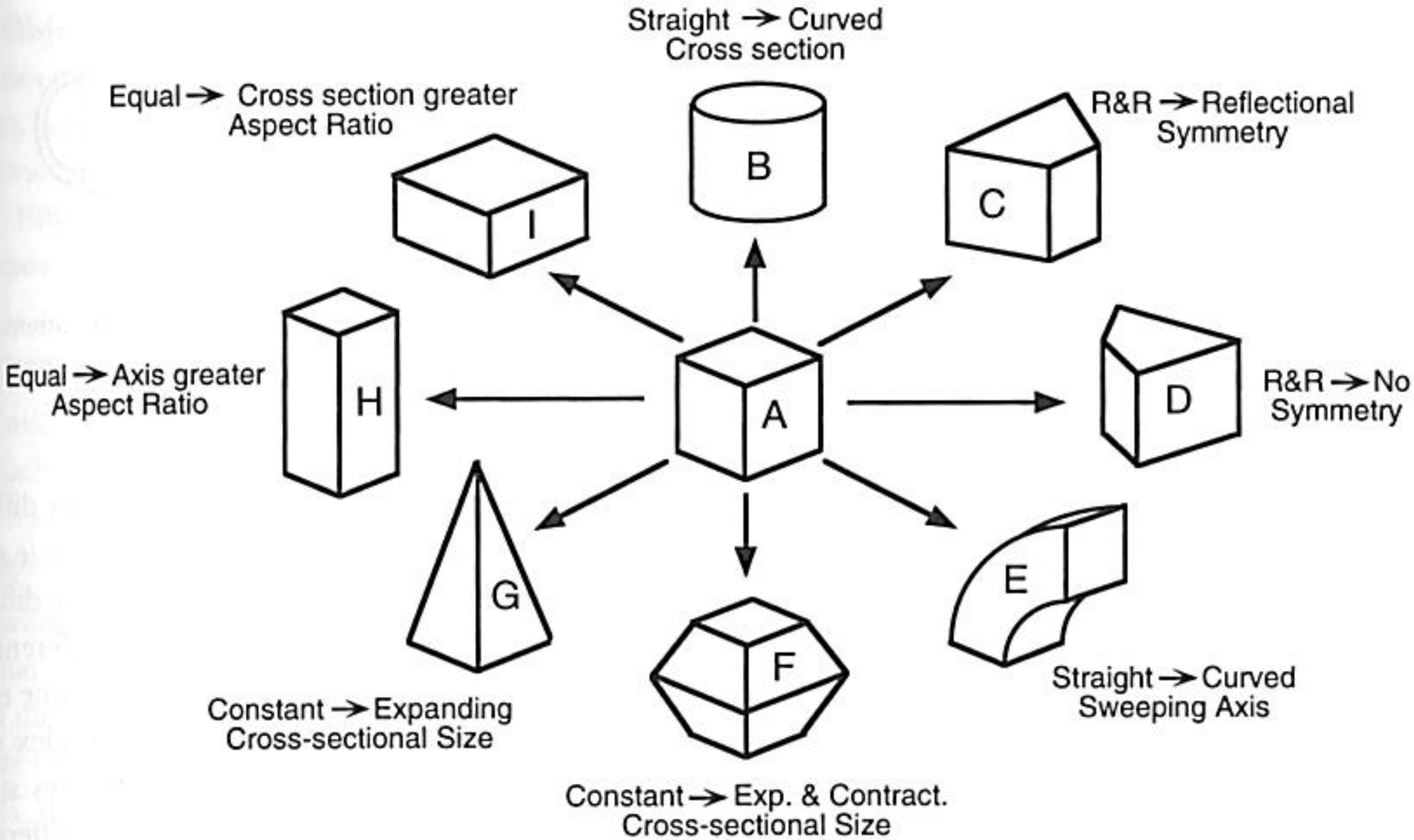


Figure 9.3.2 Illustrations of five variables in constructing generalized cylinders. The central cube (A) can be modified to construct the eight other geons shown by changing just one of five parameters: curvature of cross-sectional edges (B), cross-sectional symmetry (C and D), curvature of sweeping axis (E), diameter of sweeping rule (or cross-sectional size) (F and G), and aspect ratio (H and I). R&R = rotational and reflectional symmetry.

From Palmer (1999)

Partial Tentative Geon Set Based on Nonaccidentalness Relations

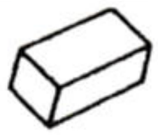
Geon	Edge Straight S Curved C	Symmetry Rot & Ref ++ Ref + Asymm –	Size Constant ++ Expanded – Exp & Cont – –	Axis Straight + Curved –	CROSS SECTION
	S	+ +	+ +	+	
	C	+ +	+ +	+	
	S	+	–	+	
	S	+ +	+	–	
	C	+ +	–	+	
	S	+	+	+	

Figure 7. Proposed partial set of volumetric primitives (geons) derived from differences in nonaccidental properties.

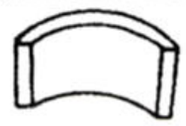
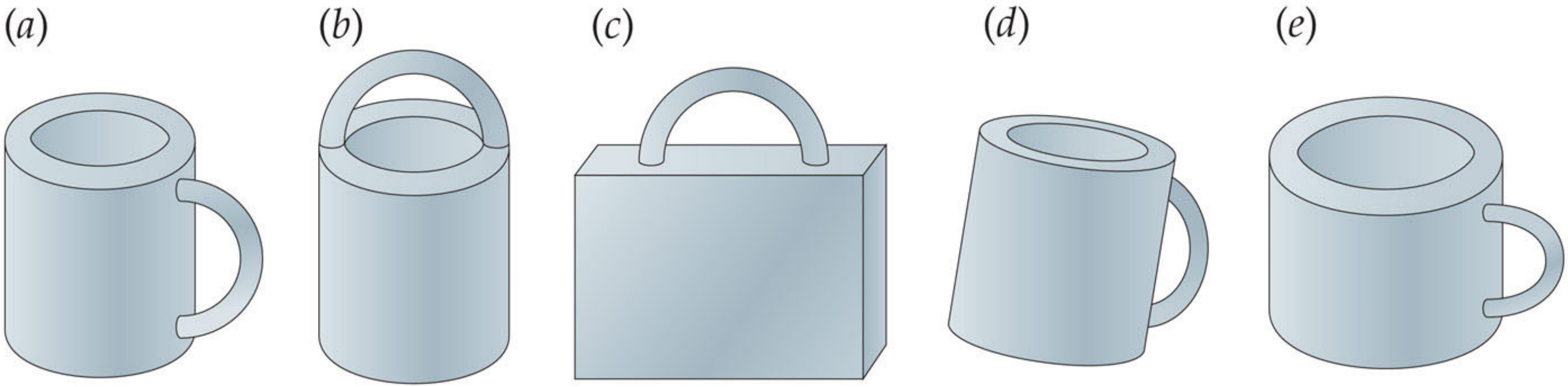
Geon	Edge Straight S Curved C	Symmetry Rot & Ref ++ Ref + Asymm –	Size Constant ++ Expanded – Exp & Cont – –	Axis Straight + Curved –	CROSS SECTION
	S	+	+ +	–	
	C	+	+ +	–	
	S	+ +	–	–	
	C	+ +	–	–	
	S	+	–	–	
	C	+	–	–	

Figure 9. Geons with curved axis and straight or curved cross sections. (Determining the shape of the cross section, particularly if straight, might require attention.)

Altogether the theory states that there are 36 geons—plenty of elements together with the visual "grammar" of how to combine them to generate thousands and thousands of objects!

Biederman, 1987, p. 122f.

Combining geons can create a wide variety of object representations



***SENSATION & PERCEPTION 4e*, Figure 4.43**

The power of a *generative* system

Table 1
Generative Power of 36 Geons

Value		Component
36		First component (G_1)
	×	
36		Second component (G_2)
	×	
3		Size ($G_1 \gg G_2$, $G_1 \ll G_2$, $G_1 = G_2$)
	×	
2.4		G_1 top or bottom or side (represented for 80% of the objects)
	×	
2		Nature of join (end-to-end [off center] or end-to-side [centered])
	×	
2		Join at long or short surface of G_1
	×	
2		Join at long or short surface of G_2
		Total: 74,649 possible two-geon objects

Note. With three geons, 74,649 × 36 × 57.6 = 154 million possible objects. Equivalent to learning 23,439 new objects every day (approximately 1465/waking hr or 24/min) for 18 years.

Biederman, 1987, p. 128

Relations of RBC to Principles of Perceptual Organization

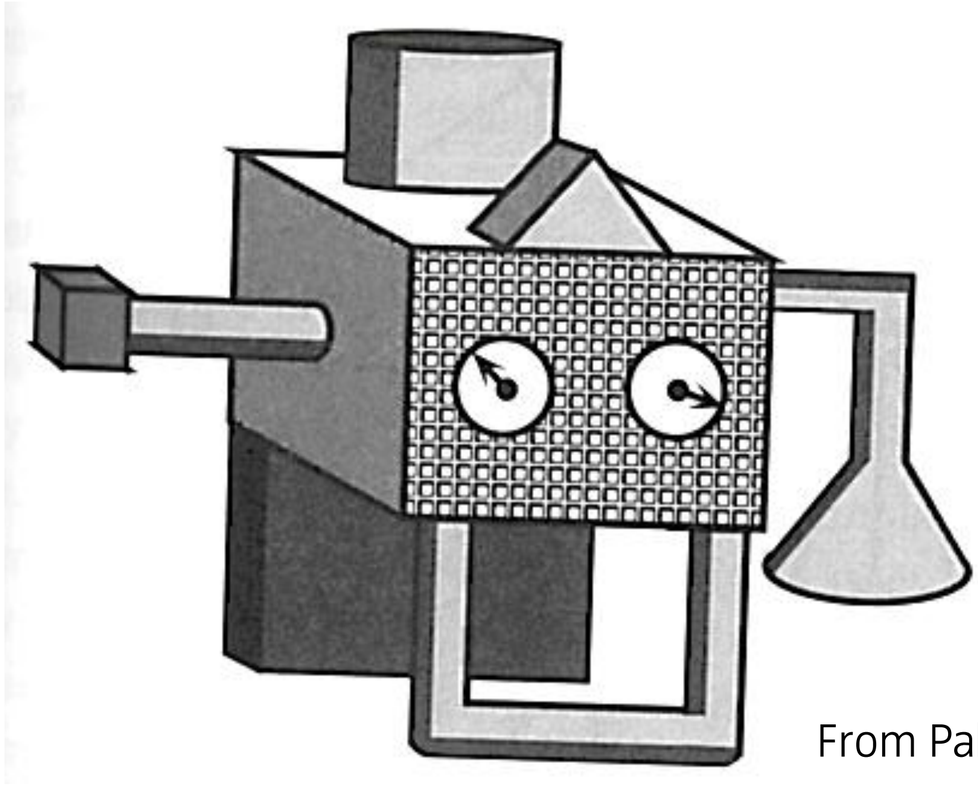
Textbook presentations of perception typically include a section of Gestalt organizational principles. This section is almost never linked to any other function of perception. RBC posits a specific role for these organizational phenomena in pattern recognition. As suggested by the section on generating geons through nonaccidental properties, the Gestalt principles, particularly those promoting Pragnanz (Good Figure), serve to determine the individual geons, rather than the complete object. A complete object, such as a chair, can be highly complex and asymmetrical, but the components will be simple volumes. A consequence of this interpretation is that it is the components that will be stable under noise or perturbation. If the components can be recovered and object perception is based on the components, then the object will be recognizable.

Biederman, 1987, p. 126

Object classification within RBC

Once the shape of an object has been represented (via geons and their spatial relationships), object classification reduces to matching the structural description of the input with stored structural descriptions of known entry-level categories.

Biederman (1987) estimates that most people know about 30,000 different object categories. This implies there are arrangements of geons that do not correspond to known categories (allowing novel object recognition).



From Palmer (1999)

Object classification within RBC

There are also combinations of features that do not constitute legal geons. The brain may use this constraint to improve segmentation / edge extraction via feedback — what edge information would we expect to be present given a geon.

Similarly: category-level expectations could affect what other geons are expected to be present.

Hummel & Biederman (1992) proposed a neural network model of RBC—but skipped one of the hardest steps (by taking edge sketches, not images, as input)

Accounting for empirical phenomena

Typicality effects (Rosch's evidence for prototypes): prototypical instances (robin) activate categorical representation (bird) more strongly than atypical instances (ostrich)

Entry-level categories: atypical examples (ostrich) activate their subordinate category more than their basic-level category

Limitations of RBC

Count Versus Mass Noun Entities: The Role of Surface Characteristics

There is a restriction on the scope of this approach of volumetric modeling that should be noted. The modeling has been limited to concrete entities with specified boundaries. In English, such objects are typically designated by count nouns. These are concrete objects that have specified boundaries and to which we can apply the indefinite article and number. For example, for a count noun such as "chair" we can say "a chair" or "three chairs." By contrast, mass nouns are concrete entities to which the indefinite article or number cannot be applied, such as water, sand, or snow. So we cannot say "a water" or "three sands," unless we refer to a count noun shape, as in "a drop of water," "a bucket of water," "a grain of sand," or "a snowball," each of which does have a simple volumetric description. We conjecture that mass nouns are identified primarily through surface characteristics such as texture and color, rather than through volumetric primitives.

Biederman, 1987, p. 116

... more limitations of RBC

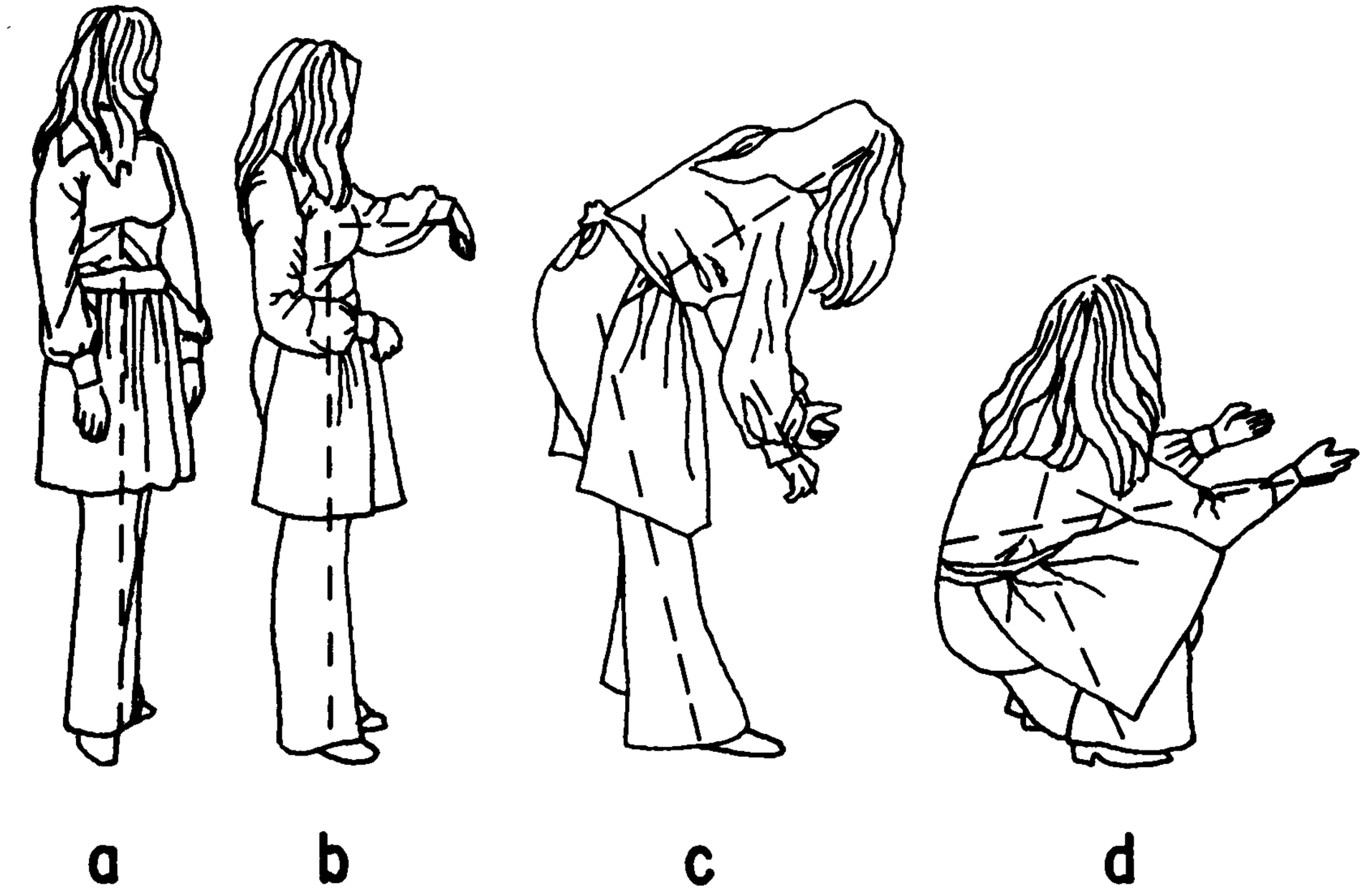


Figure 29. Four configurations of a nonrigid object.

Biederman, 1987, p. 145

Viewpoint invariance

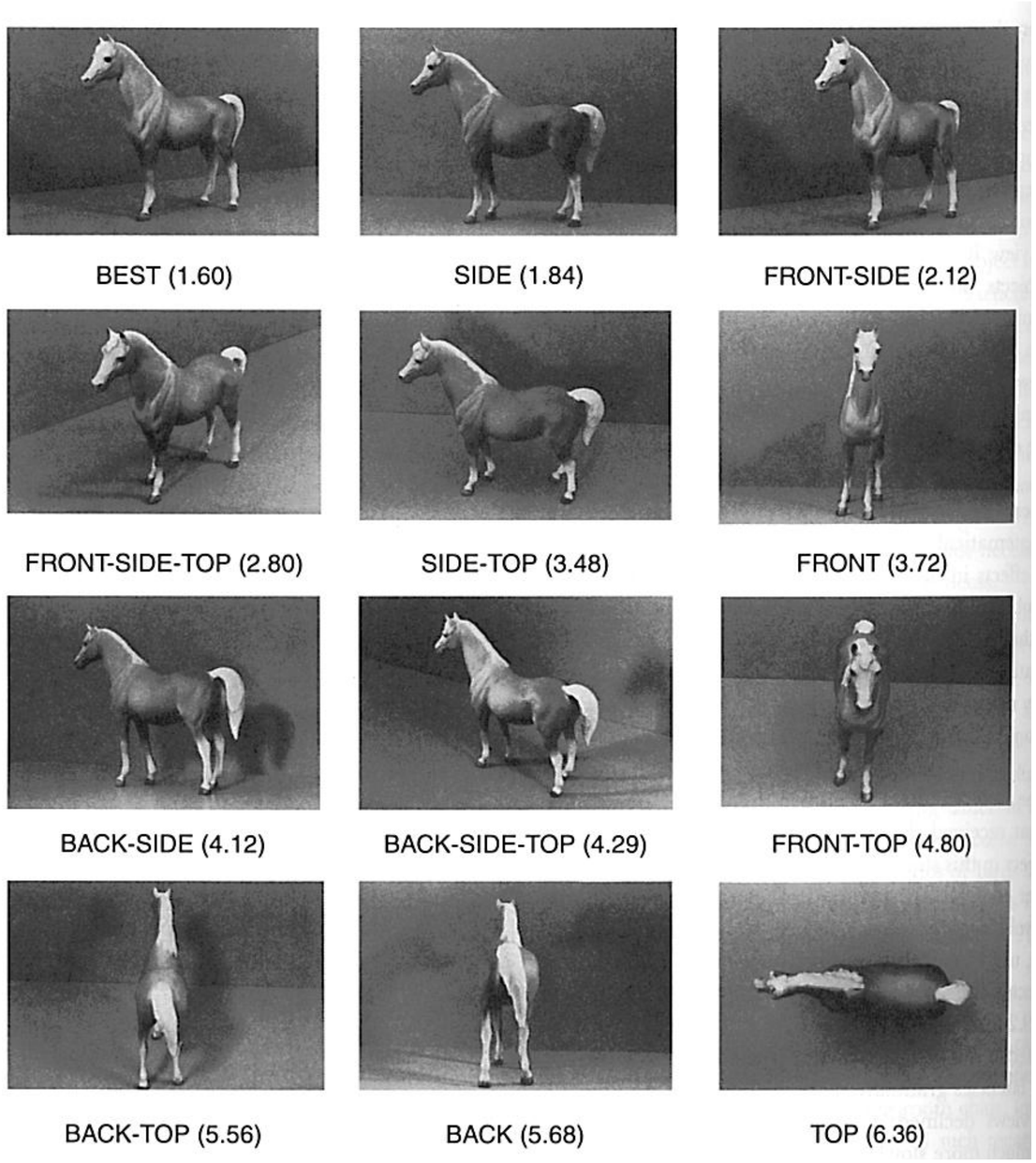
1. A property of an object that does not change when an observer changes viewpoint

2. A class of theories of object recognition that proposes representations of objects that do not change when viewpoint changes (e.g. RBC)

3. The observation that humans are good at recognising 3D objects despite variations in perspective

Viewpoint invariance and canonical views

...but. Object recognition is not completely viewpoint invariant. Objects shown in ‘‘canonical views’’ are classified and recognised faster and more accurately.



From Palmer (1999)

Problems with structural-description theories

All of the above (canonical view effects, nonrigid objects) is often cited as a problem for a structural-description theories

but, to be honest, I disagree: RBC theory, for example, does not argue every possible geon arrangement is equally easy to parse from the input and the calculated edges. Perhaps it is easier to see the geons themselves if they are presented in non-ambiguous ways.

e.g. viewing a brick geon directly from the side means you are missing the crucial Y vertex. Can be confused with other geons!

This is one potential strength of the theory, because humans also show canonical view effects. To my knowledge no systematic comparison of human and RBC errors has been performed.

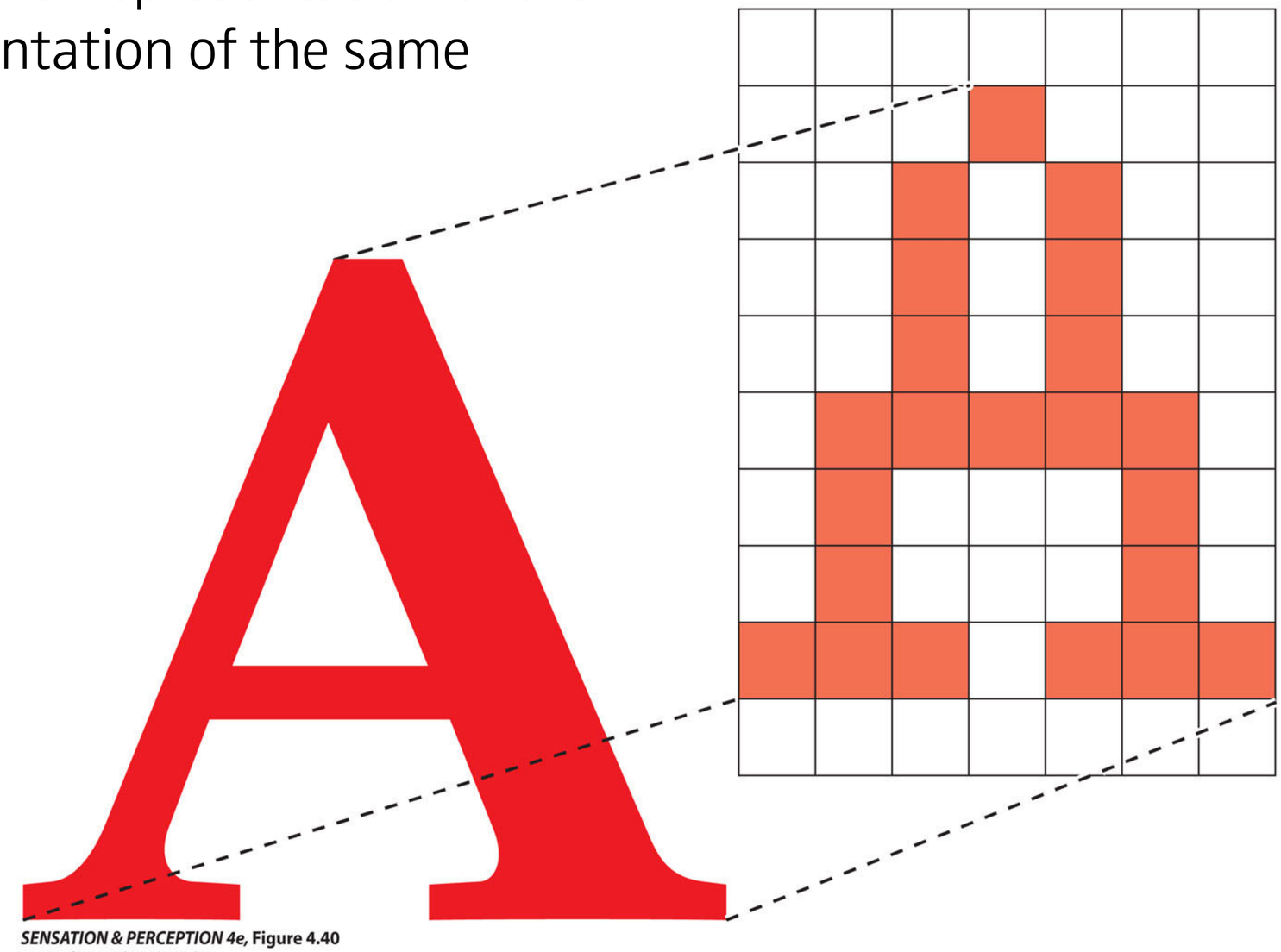
Nevertheless, geons aren't always the best—or even a good—descriptions of objects, particularly for non-rigidly deforming, non-geometric, objects and natural shapes—thus for almost anything not man-made!

View-based theories

Templates versus structural descriptions

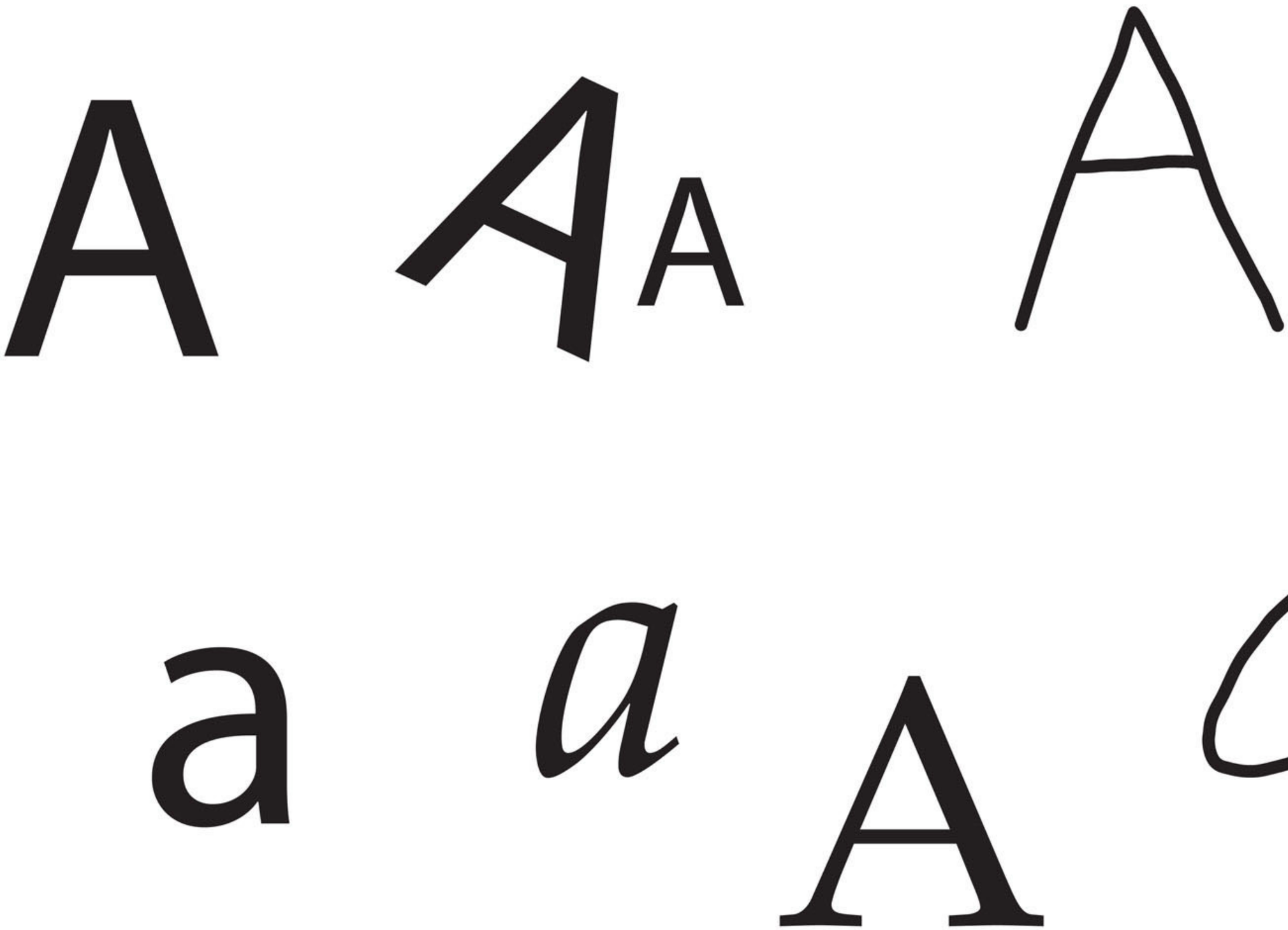
Structural description: A description of an object in terms of its parts and the relationships between those parts.

Naïve template theory: The proposal that the visual system recognises objects by matching the neural representation of the image with a stored representation of the same "shape" in the brain.



***SENSATION & PERCEPTION 4e,* Figure 4.40**

The problem with templates is that we need a lot of them



View-Based Models

Image-driven models; visual system "stores" snap-shots of 2D "images" of objects—often referred to as templates—and inter- or extrapolates from them to novel views.

Thus not quite as naïve in that they do not argue you need a template for exactly every object and every novel view, because of "inter- and extrapolation."

But ... the exact processes are often left unspecified.

How can you recognise a novel 3D object, if all you store are snapshots of 2D views of previously-seen objects?

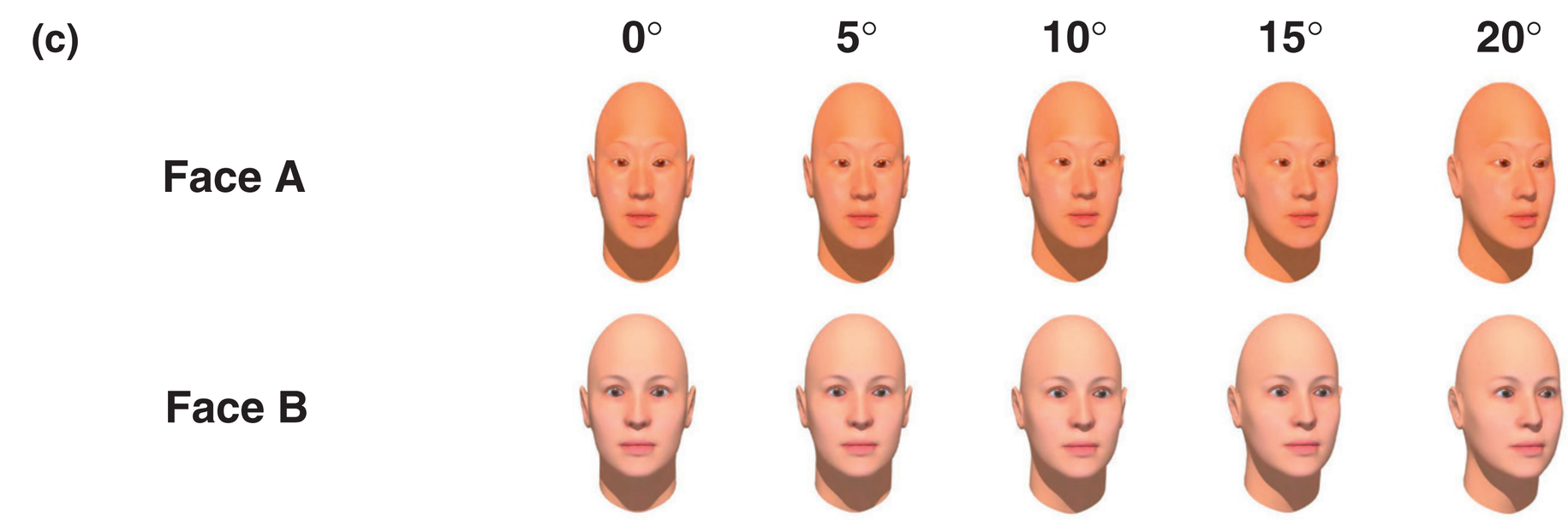
View-based models (as with RBC) have trouble with nonrigid objects for the same reason: store lots of 3D models?

Furthermore, it is extremely unlikely that any template could be simply in the input space (= image space or pixel space)—this is explained in the following slides.

Cox (2014)

(a)

(c)



Cox (2014)

Neuron
Perspective

How Does the Brain Solve Visual Object Recognition?

James J. DiCarlo,[1,*] Davide Zoccolan,[2] and Nicole C. Rust[3]
[1]Department of Brain and Cognitive Sciences and McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Cognitive Neuroscience and Neurobiology Sectors, International School for Advanced Studies (SISSA), Trieste, 34136, Italy
[3]Department of Psychology, University of Pennsylvania, Philadelphia, PA 19104, USA
*Correspondence: dicarlo@mit.edu
DOI 10.1016/j.neuron.2012.01.010

Mounting evidence suggests that 'core object recognition,' the ability to rapidly recognize objects despite substantial appearance variation, is solved in the brain via a cascade of reflexive, largely feedforward computations that culminate in a powerful neuronal representation in the inferior temporal cortex. However, the algorithm that produces this solution remains poorly understood. Here we review evidence ranging from individual neurons and neuronal populations to behavior and computational models. We propose that understanding this algorithm will require using neuronal and psychophysical data to sift through many computational models, each based on building blocks of small, canonical subnetworks with a common functional goal.

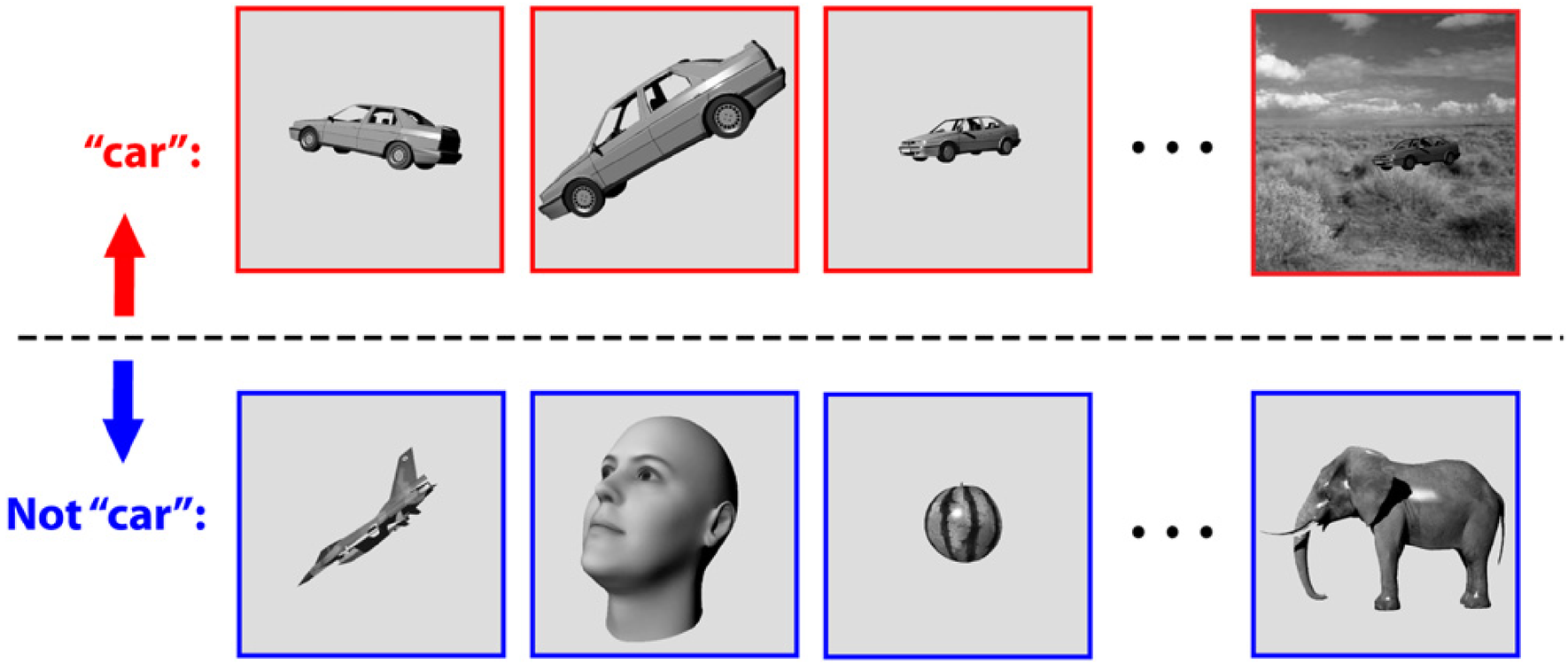
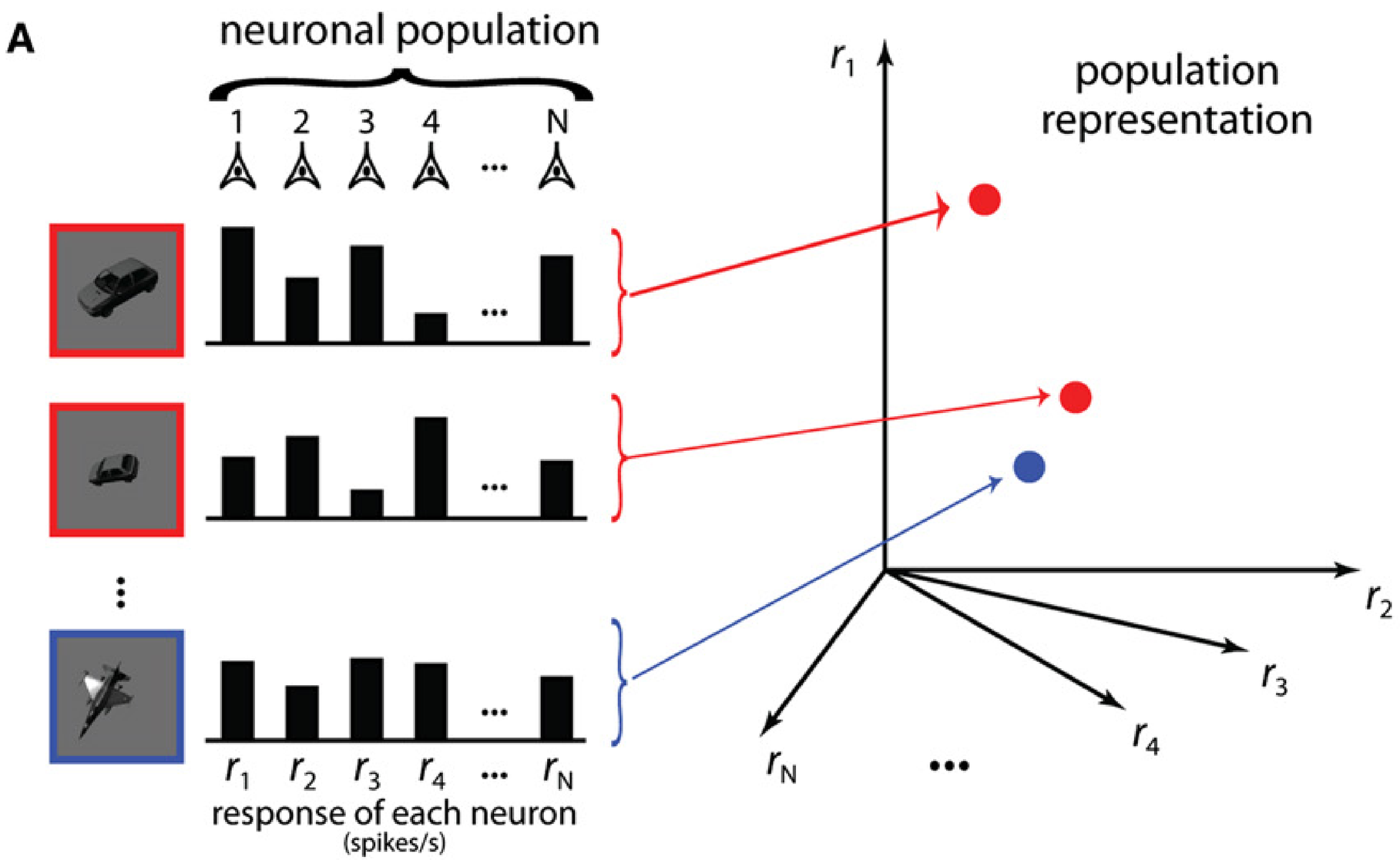


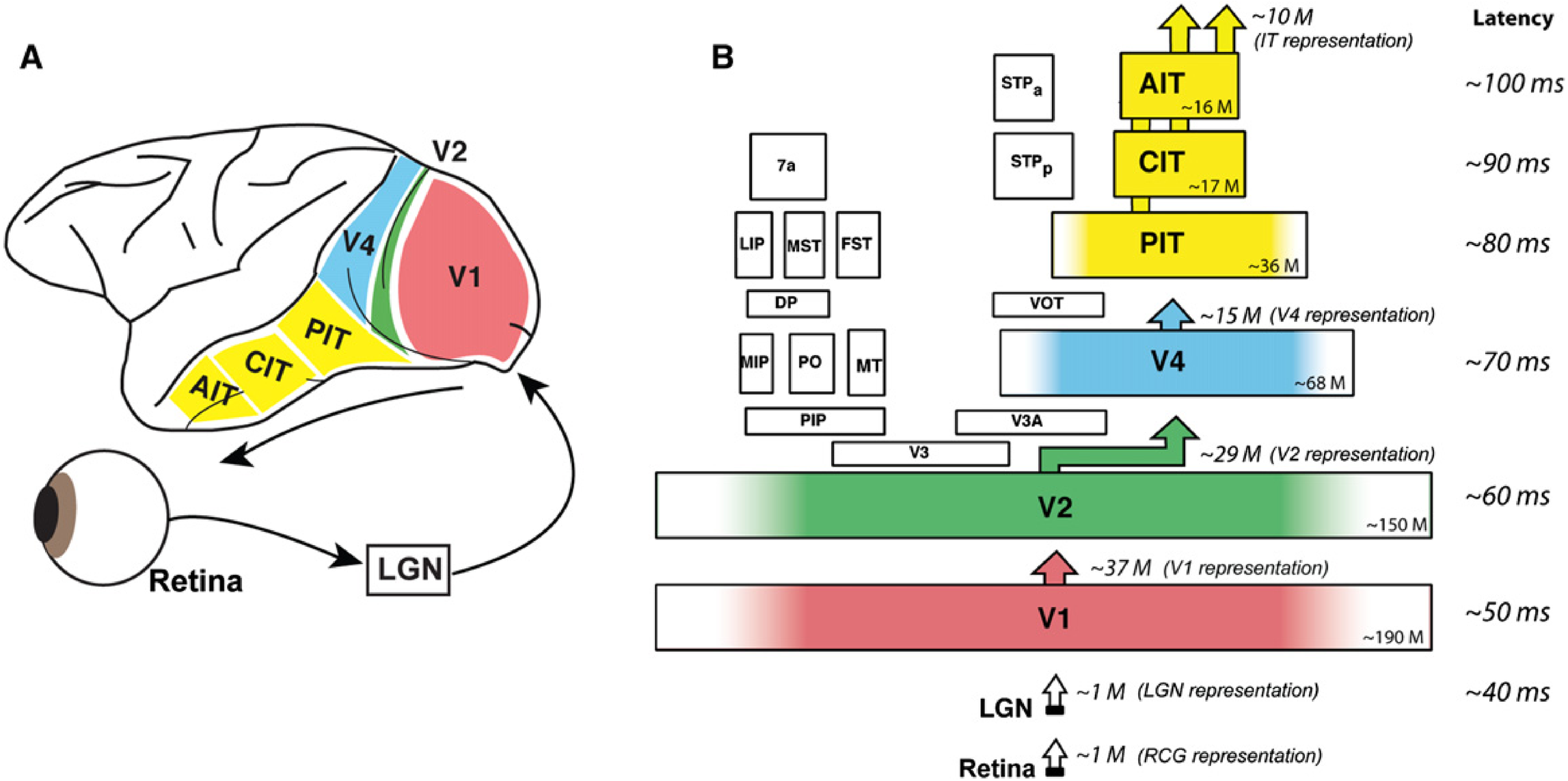
Figure 1. Core Object Recognition

Core object recognition is the ability to rapidly (<200 ms viewing duration) discriminate a given visual object (e.g., a car, top row) from all other possible visual objects (e.g., bottom row) without any object-specific or location-specific pre-cuing (e.g., DiCarlo and Cox, 2007). Primates perform this task remarkably well, even in the face of identity-preserving transformations (e.g., changes in object position, size, viewpoint, and visual context).

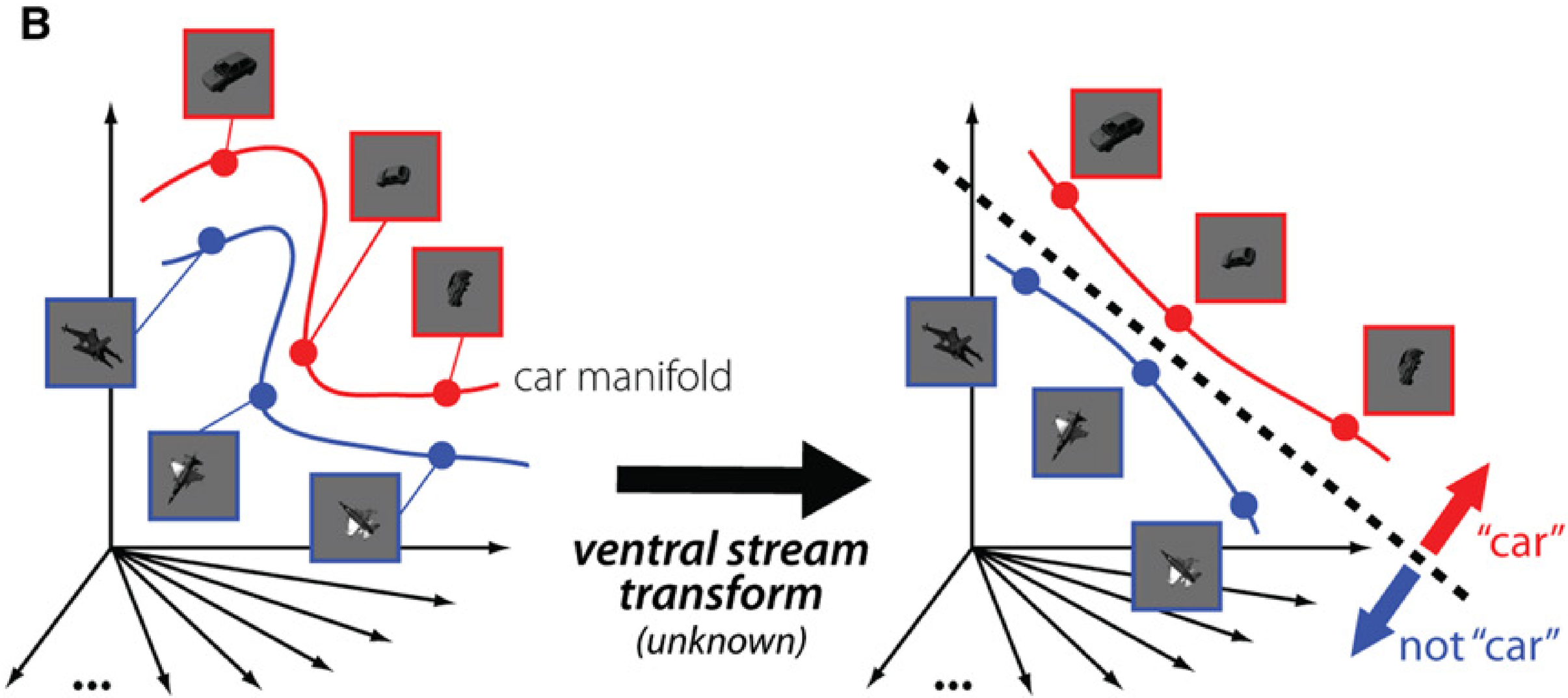
DiCarlo et al. (2012)



DiCarlo et al. (2012)

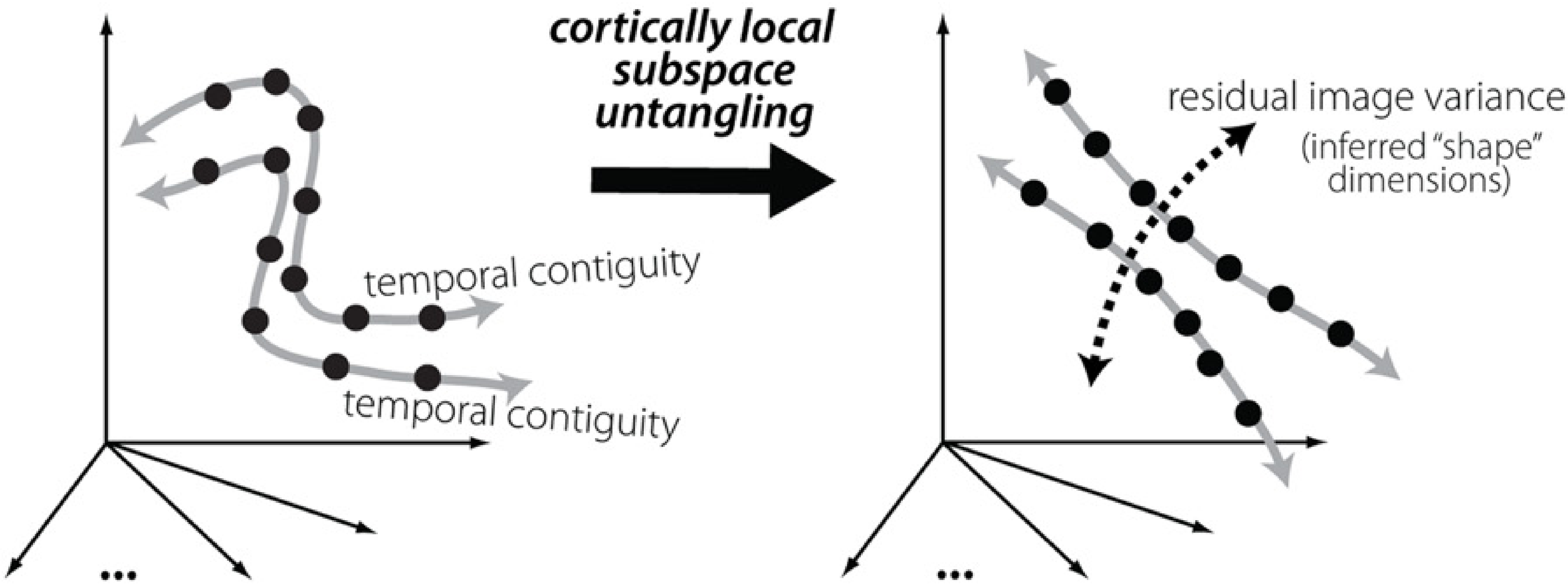


from: DiCarlo et al. (2012)

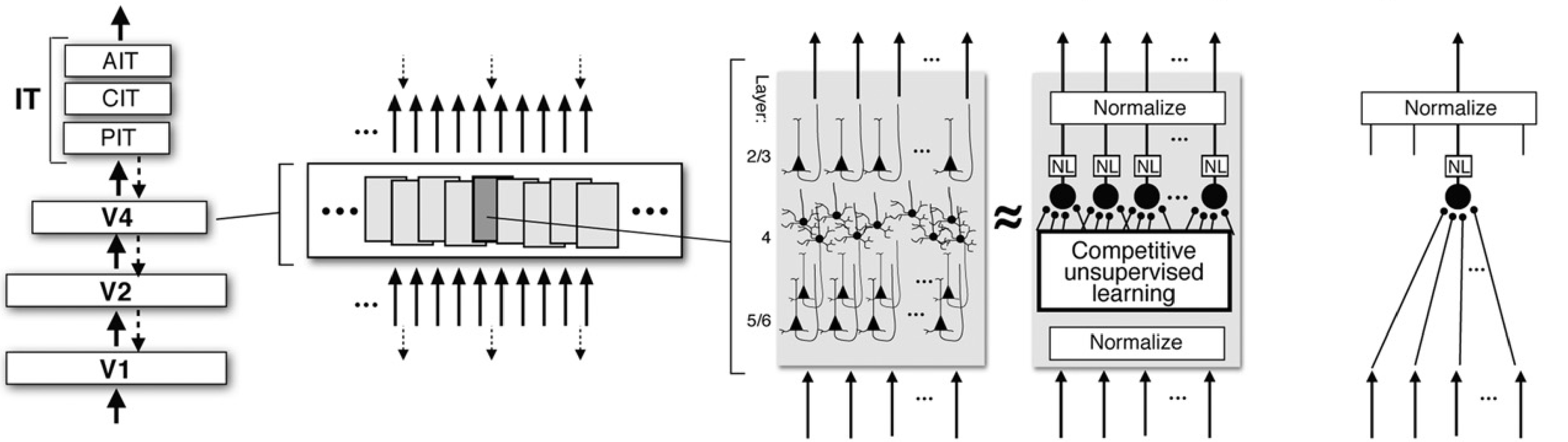


DiCarlo et al. (2012)

c



DiCarlo et al. (2012)



DiCarlo et al. (2012)

Multiple object recognition systems?

Perhaps there are several object recognition processes, depending on the type of object: Man-made versus others? Faces? Depending on the category level? (entry-level, subordinate, superordinate?

Perhaps view-based representations exist, and when the current input matches a stored view, recognition is fast and accurate. But to recognise objects under more challenging conditions, or novel objects, categorisation relies on a slower process of matching structural descriptions.

Faces: An illustrative special case

Could (e.g.) RBC theory account for face recognition? Can faces even be reduced to geon descriptions? Even if, they would have virtually identical geon descriptions.

Biederman: RBC is meant as a theory of object categorisation, not face identification.

Face recognition seems to be special and different from object recognition (Gauthier et al., 2003; Kanwisher & Yovel, 2006).

For example, for faces we have comparatively little up-down invariance, presumably because we see so few up-down inverted faces.

Which of these two photos has been altered?

***SENSATION & PERCEPTION 4e*, Figure 4.46**

Which of these two photos has been altered?

Object recognition processing assorted “common wisdom”

Perception is a two-way street: Feedback and reentrant processing

Initial object recognition can occur very quickly (150 ms), but that’s not the end of the story.

The brain continues to process information, sending signals up and down the ‘‘what-pathway’’.

Object recognition should be seen as a conversation among many parts of the brain rather than as a one-way progression.

The End

Felix Wichmann

Neural Information Processing Group
Eberhard Karls Universität Tübingen