Bidirectional processing II:

feedforward & feedback networks for recognition

Focus on feedback computations

computational problems

Inferences about the image involve various inferences:

- types of features & attributes (shapes, material)
- recognition over levels of abstraction (parts, objects, actions, scenes)
  - spatial scales
  - relationships

Descriptions are inferences of object properties and relationships — i.e. causes of image intensities, not of image intensity patterns

A crucial assumption is that these inferences are based on deep, generative knowledge of how virtually any natural image could be produced

computational problems

Need to solve scalability

Solving toy (low-dimensional) problems rarely scales up to deal with the complexity of natural images.

In object recognition, humans have the capacity to *quickly* deal with an enormous space of possible objects (30 to 300K) as they appear in different contexts in natural images for different tasks.

computational problems

Need to model uncertainty

vision is concerned with causes of image intensity patterns, but the causes of behavioral relevance are encrypted and confounded

many hypotheses about cause can be consistent with the same local image evidence

local variations in image evidence can be consistent with the same cause

accurate perceptual decisions resolve these ambiguities by combining lots of image evidence with built-in knowledge

computational problems

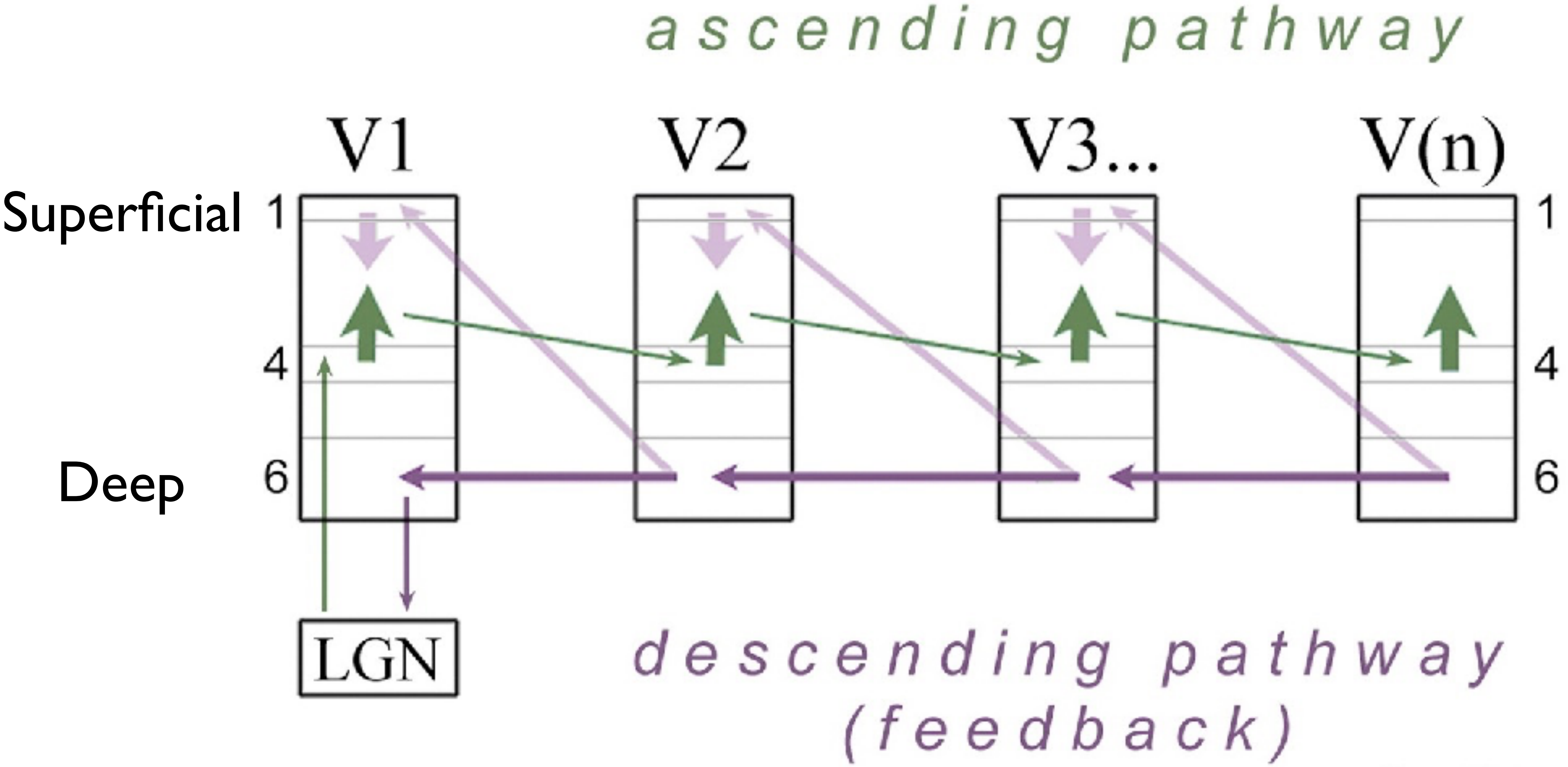
Need to solve task flexibility

Vision stimulates and support answers to a limitless range of questions. Human vision doesn’t just recognize, it interprets scenes.

e.g. description of the fox

“One can see that there is an animal, a fox–in fact a baby fox. It is emerging from behind the base of a tree not too far from the viewer, is heading right, high-stepping through short grass, and probably moving rather quickly. Its body fur is fluffy, reddish-brown, relatively light in color, but with some variation. It has darker colored front legs and a dark patch above the mouth. Most of the body hairs flow from front to back...and what a cute smile, like a dolphin.”

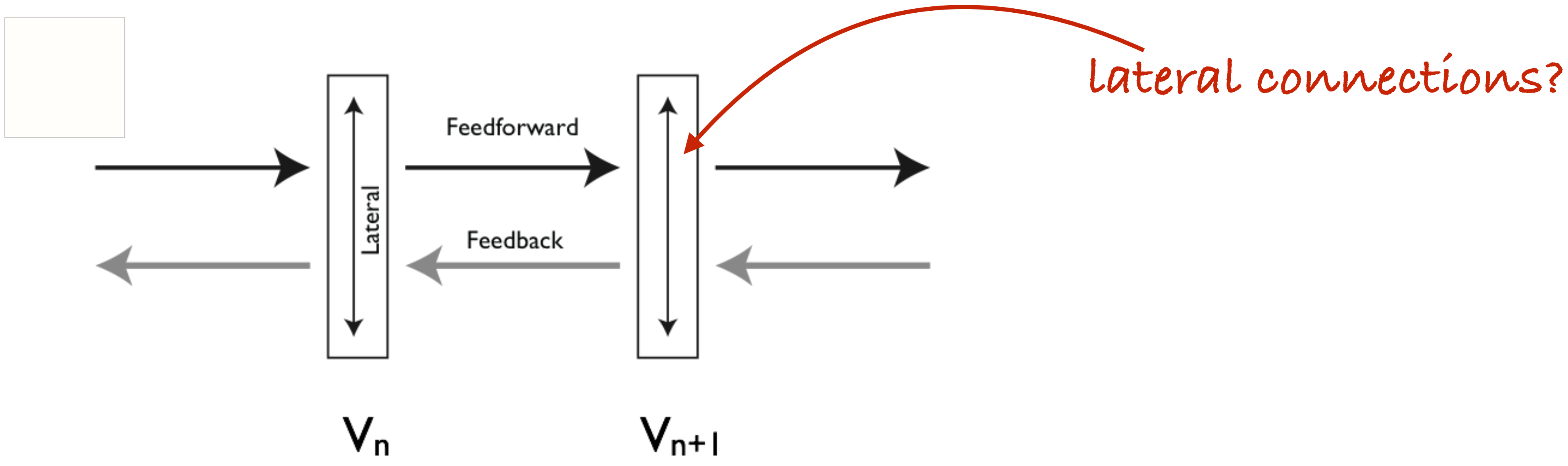
What is missing from feedforward models?



Current Biology

Shipp, S. (2007). Structure and function of the cerebral cortex. CURBIO, 17(12), R443–9. doi:10.1016/j.cub.2007.03.044

What is missing from feedforward models?

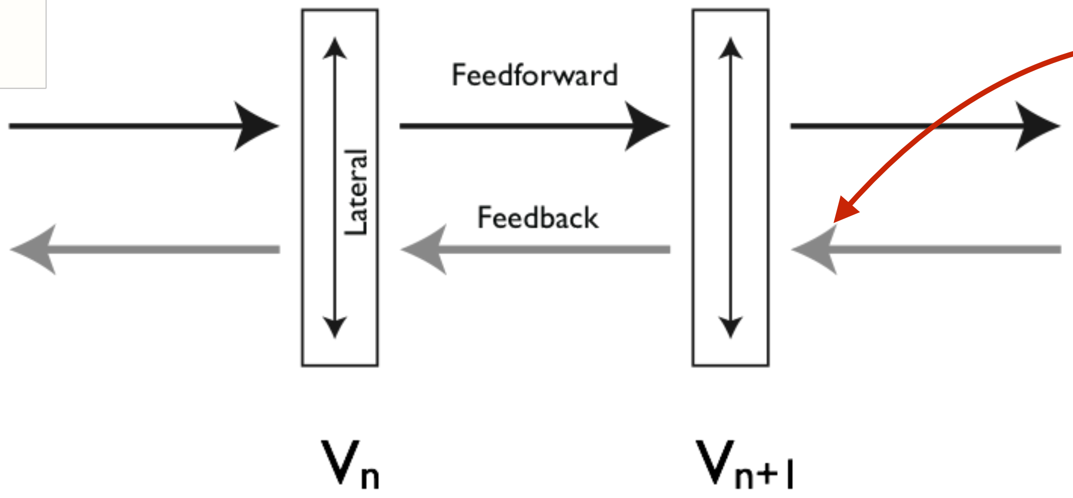


lateral connections?

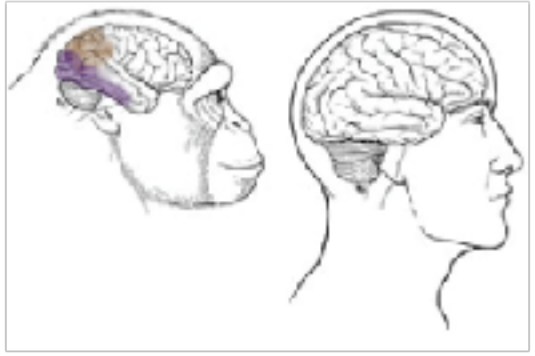
Lateral organization

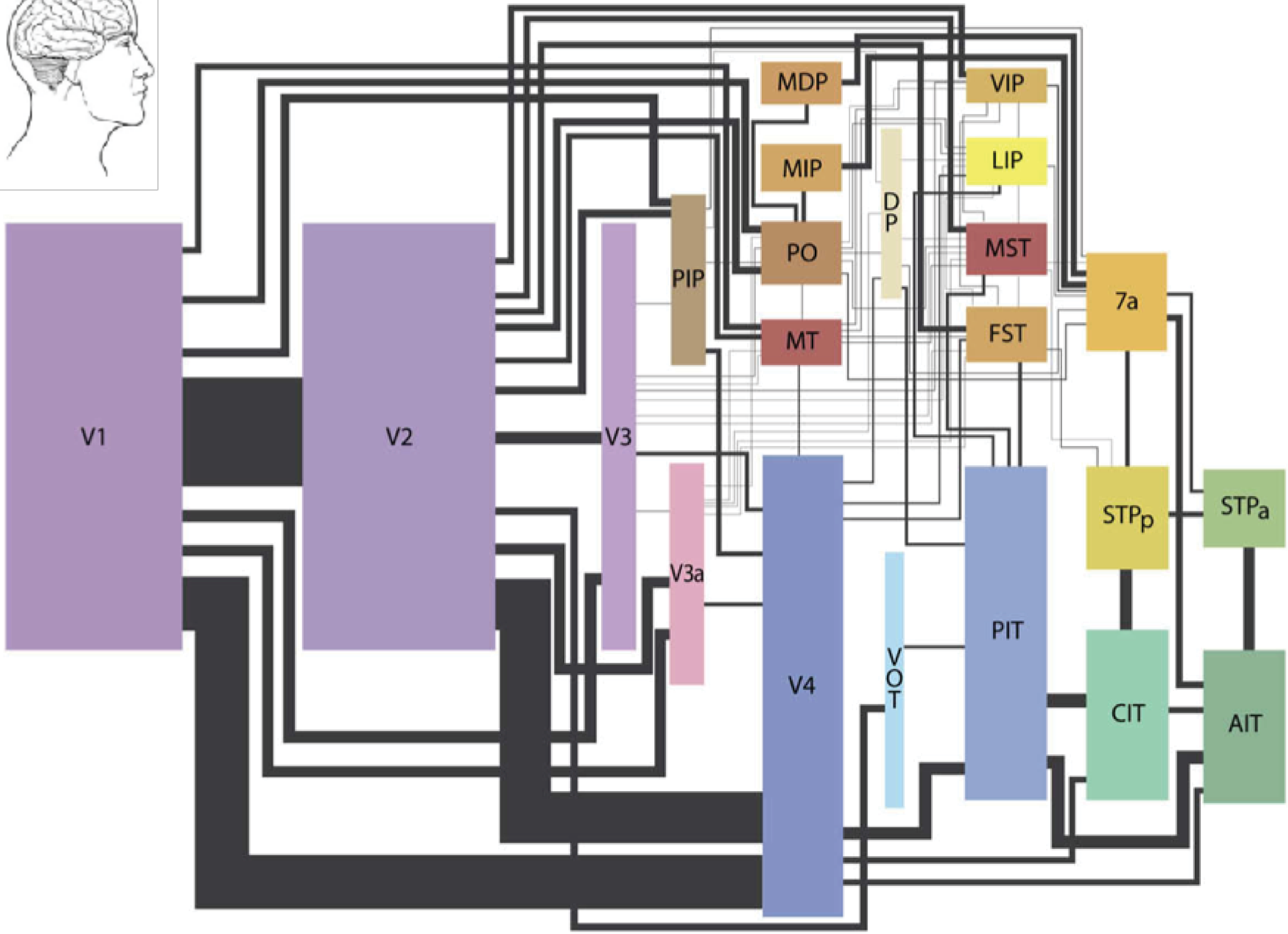
- representation and linking of features at a similar level of abstraction
- self-organizing topographical maps
- efficient image coding to explain receptive field properties
- machine learning methods for grouping

What is missing from feedforward models?



feedback





dorsal

ventral

feedforward

feedback

relation to Bayes?

$$p(S|I) = \frac{p(I|S)p(S)}{p(I)}$$

$$p(S|I) \propto p(I - f(S))p(S)$$

does the visual system use built-in knowledge of how images are naturally generated to predict the input I, based on candidate “explanations” f(S)?

If so, such a mechanism could be used to test and sort through competing explanations

Bayesian perspective: two computational strategies

Discriminative mechanisms

p(object | image)
feedforward

- Computational/behavioral speed and accuracy requires effective diagnostic features to deal with the enormous variation within a pattern/object category

VanRullen, R., & Thorpe, S. J. (2001). The time course of visual processing: from early perception to decision-making. *Journal of Cognitive Neuroscience*, *13*(4), 454–461.

Generative mechanisms

p(image | object) x p(object)*
feedback

- Provide flexibility, generalization beyond training

* recall bayes: $p(\text{object} \mid \text{image}) \propto p(\text{image} \mid \text{object}) \times p(\text{object})$

Can feedback help with the local uncertainty, scalability and flexibility problems

Fine-scale recognition and segmentation

Unfamiliar objects/appearances

Learning given only a few examples

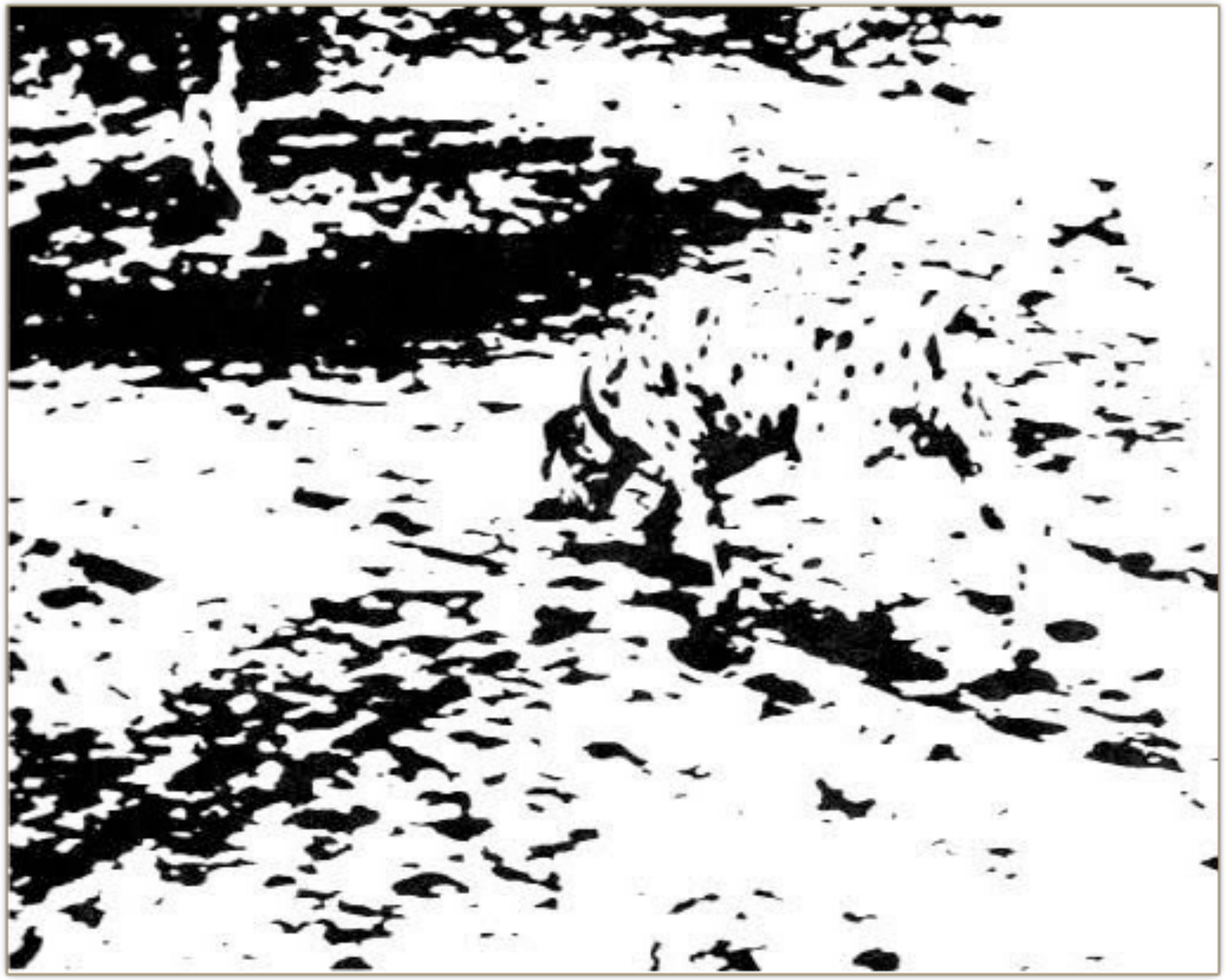
Bootstrap learning problem:
How to learn when objects aren’t experienced in isolation?

Domain-specific compositional models

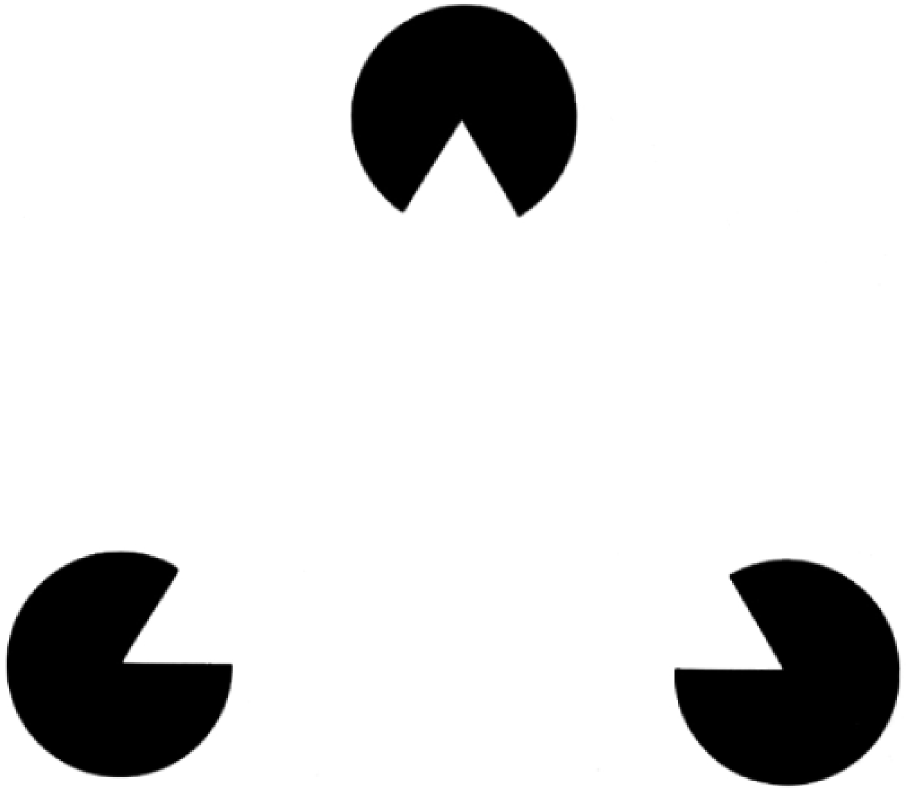
Automatic or consciously driven?

The executive metaphor
expertise at various levels of abstraction

local uncertainty, missing data

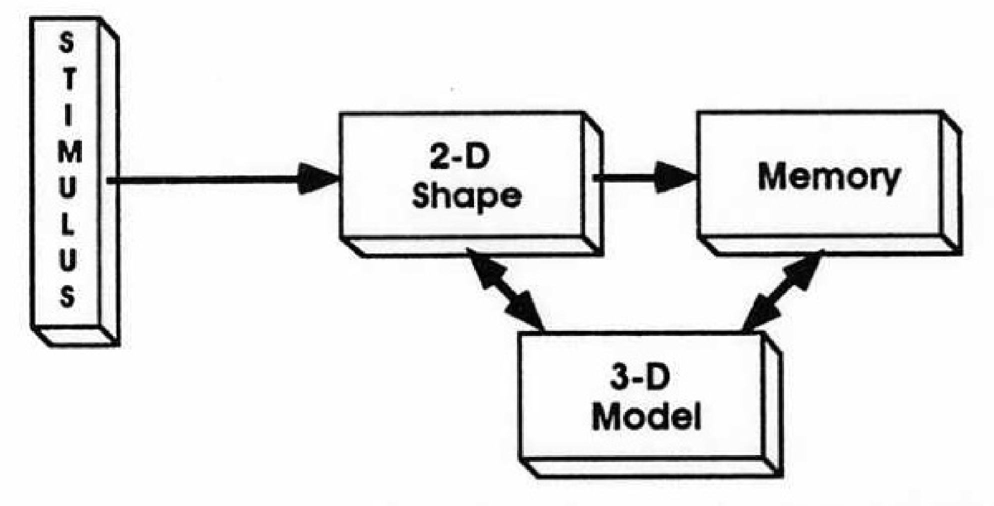


Top-down, generative models?



"explaining away"

Extraneous data: recognition despite cast shadows

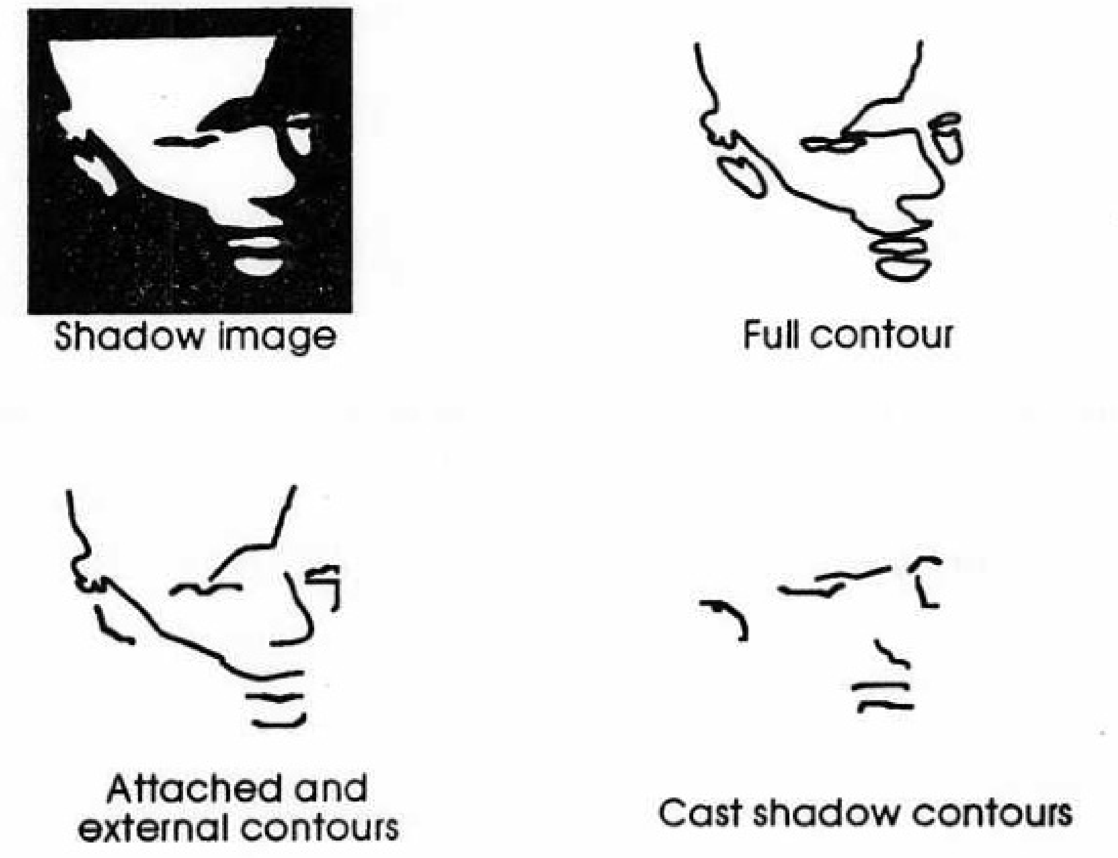


STIMULUS

2-D Shape

Memory

3-D Model



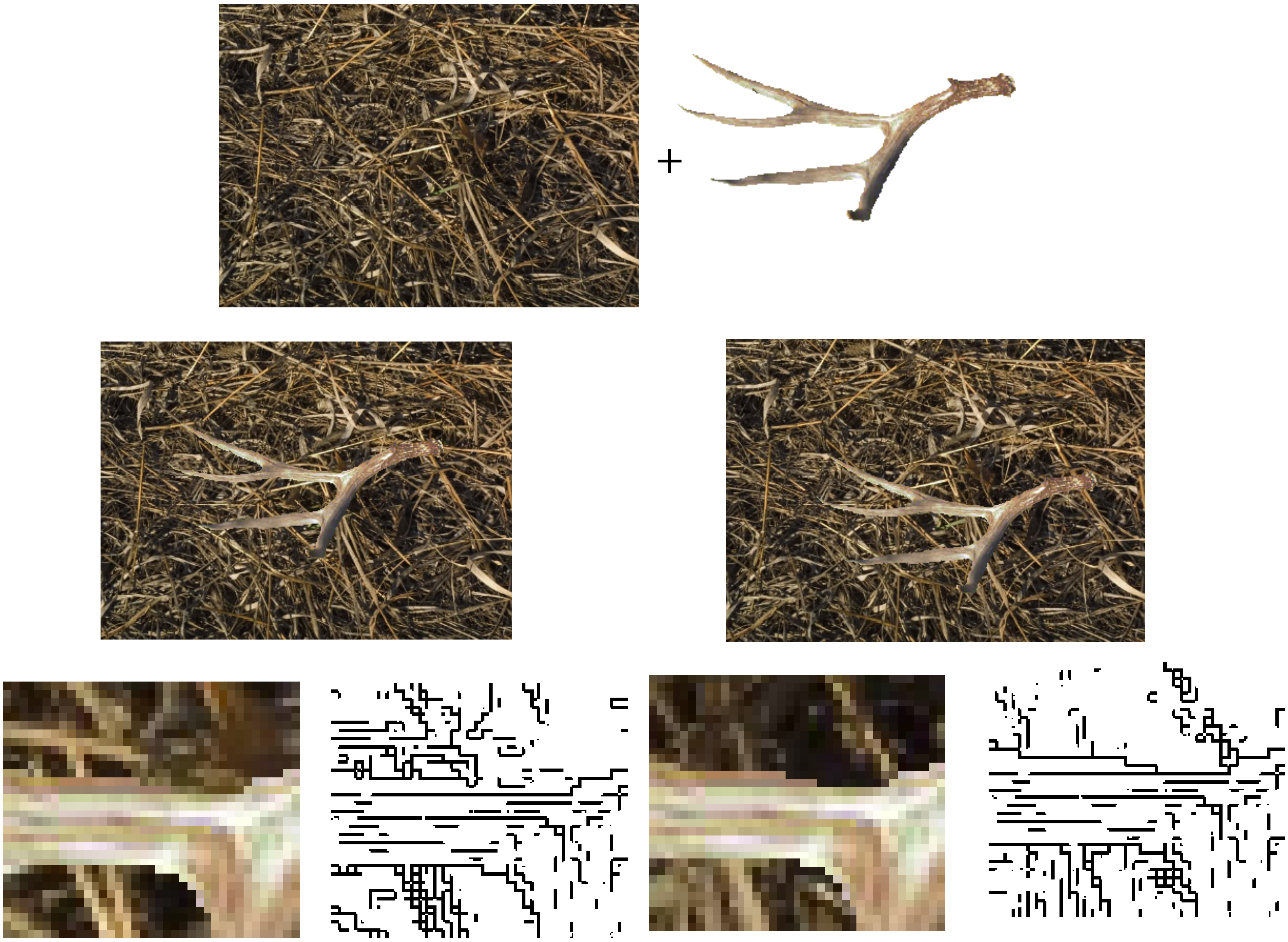
Shadow image

Full contour

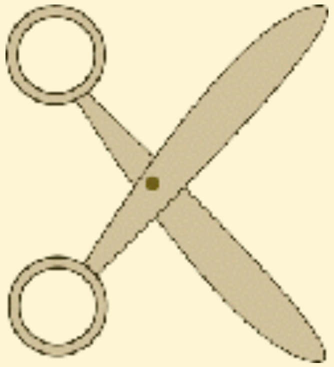
Attached and external contours

Cast shadow contours

Cavanagh P (1991) What's up in top-down processing? In: Representations of Vision: Trends and tacit assumptions in vision research (Gorea A, ed), pp 295-304. Cambridge, UK: Cambridge University Press.

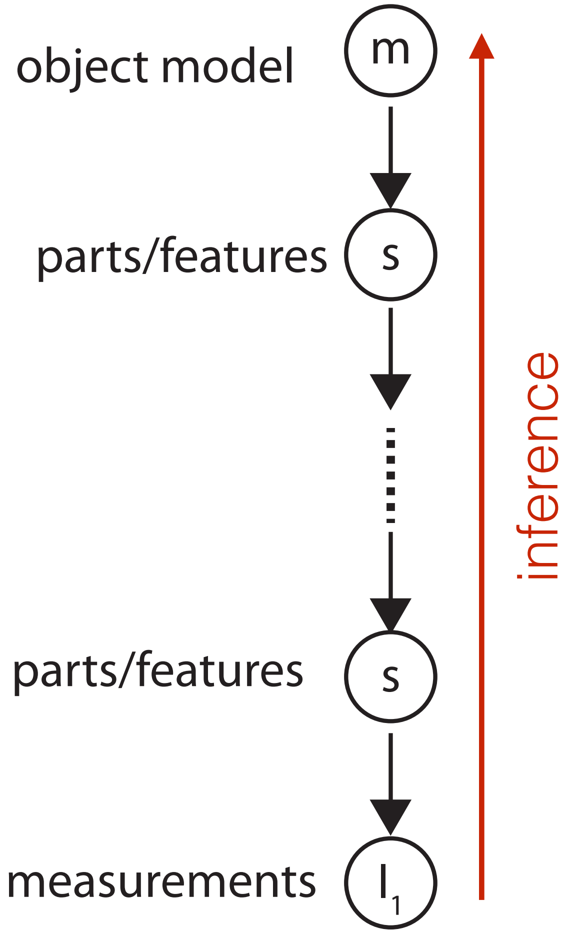


Object variations that haven't been seen before

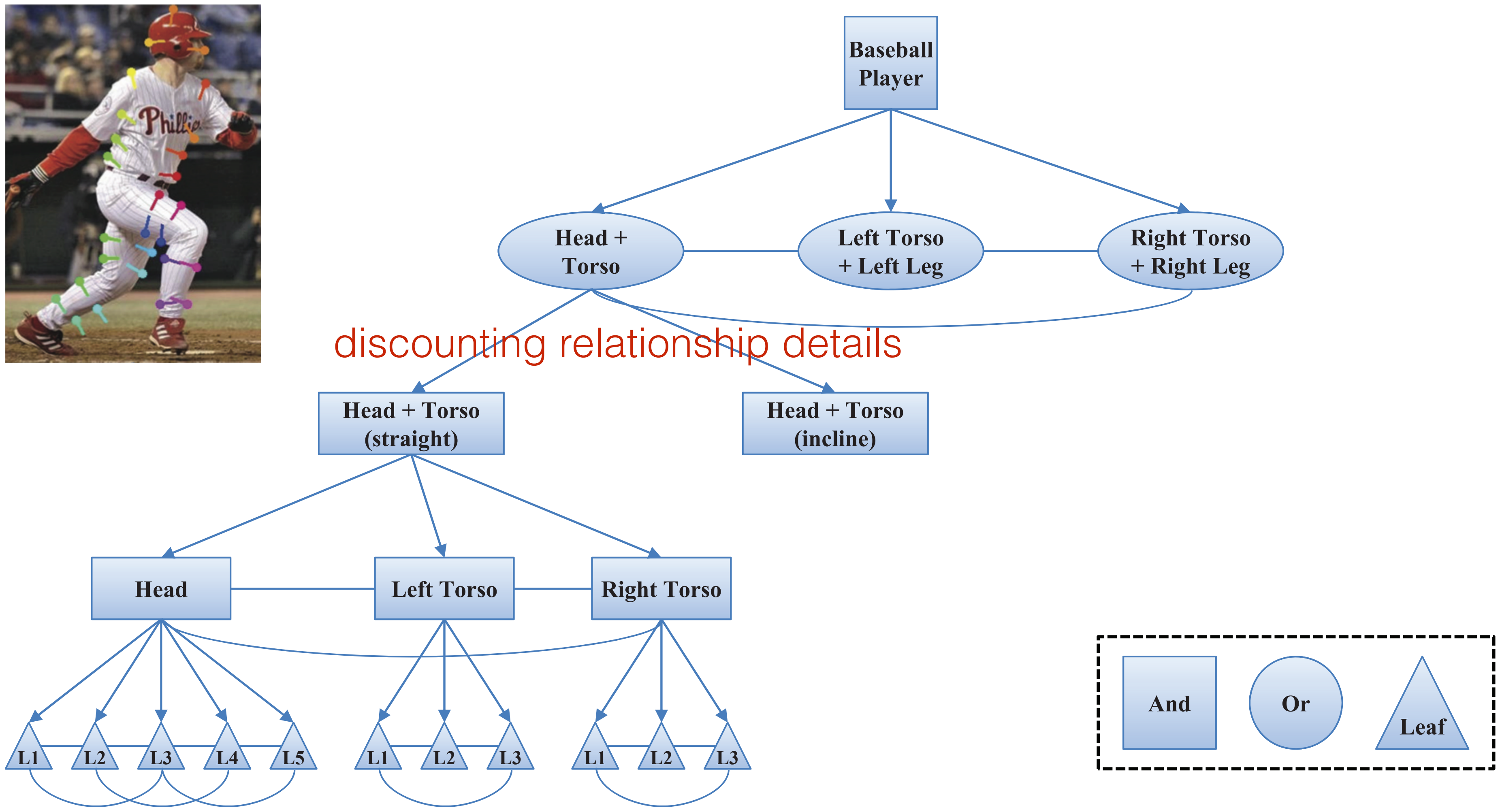


can recognize as scissors AND estimate an articulation

hard to allow for articulation without an object model



Object variations that haven't been seen before: Compositional architectures for representation



discounting relationship details

parts, patches, fragments

basic logical operation: detect "disjunctions of conjunctions"

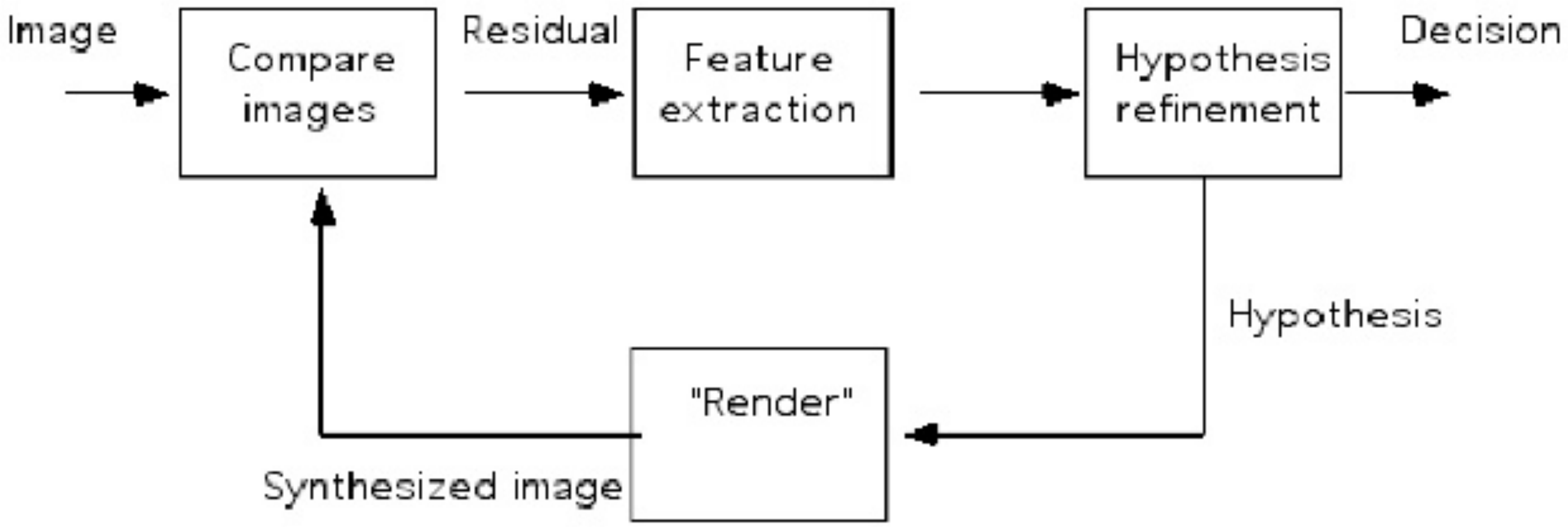
explicit recognition at multiple levels

Doesn’t mean that feedback is necessary for recognition (Thorpe et al.)

But top-down feedback may be important for

- achieving high-performance given uncertainty, noise, clutter
- task flexibility
- learning new object models

Contrast predictive coding with strictly feedforward



Image

Compare images

Residual

Feature extraction

Hypothesis refinement

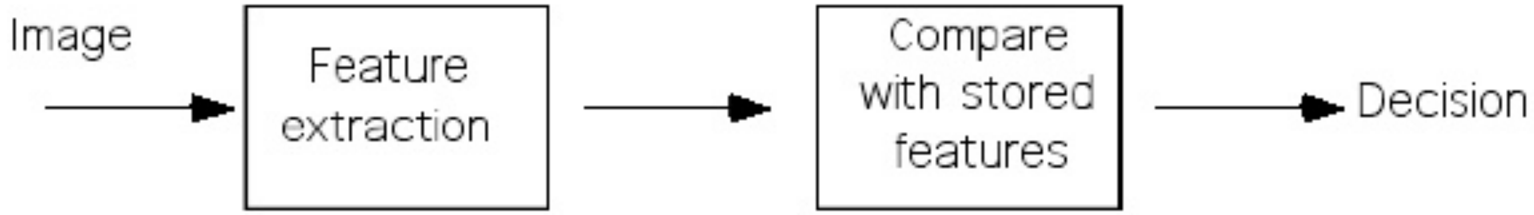
Decision

Hypothesis

"Render"

Synthesized image

Bottom-up / Top-down



Image

Feature extraction

Compare with stored features

Decision

Bottom-up

Disambiguation?

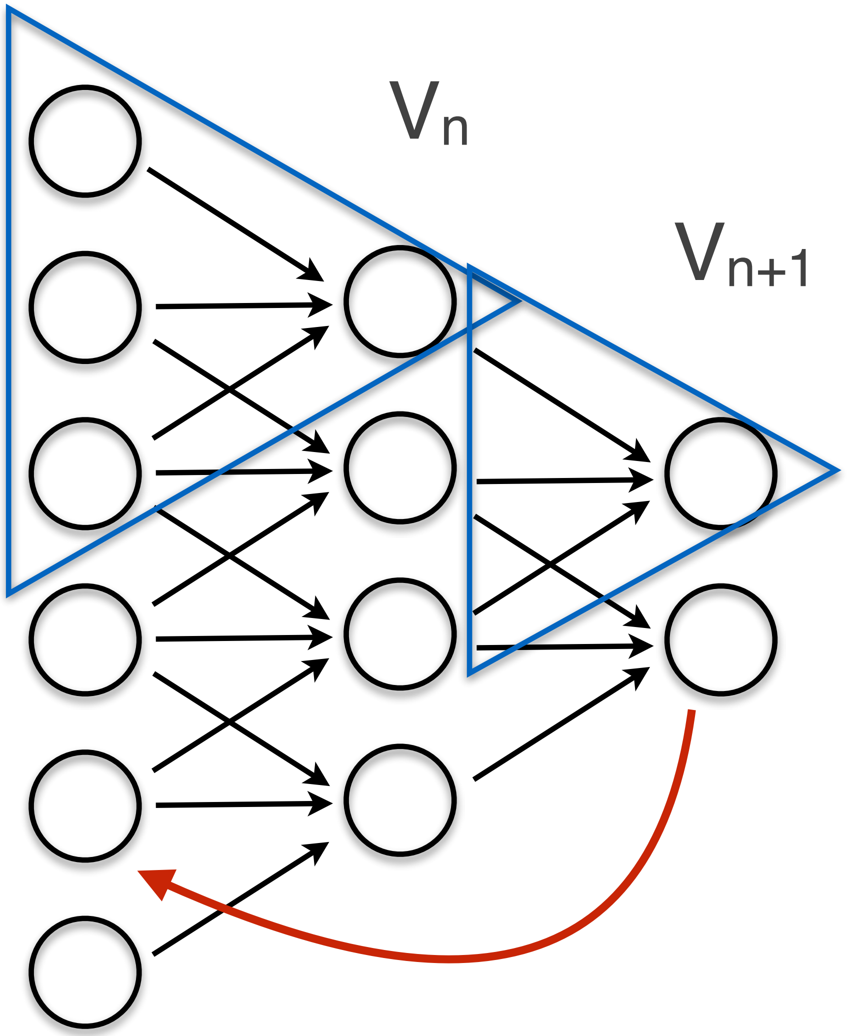
Predictive coding: suppress lower-level features that are consistent with a confident high-level interpretation. Reduce metabolic costs, signal new unexplained incoming information.

Analysis-by-synthesis. Bind lower-level information that might be required for executive tasks, e.g. fine-grain. : enhance lower-level consistent features and/or suppress inconsistent ones. Useful for representation and interpretation of novel patterns? Dealing with clutter?

How can one study feedback in humans? Psychophysics? Large-scale imaging?

take advantage of the hierarchical structure of visual cortical areas

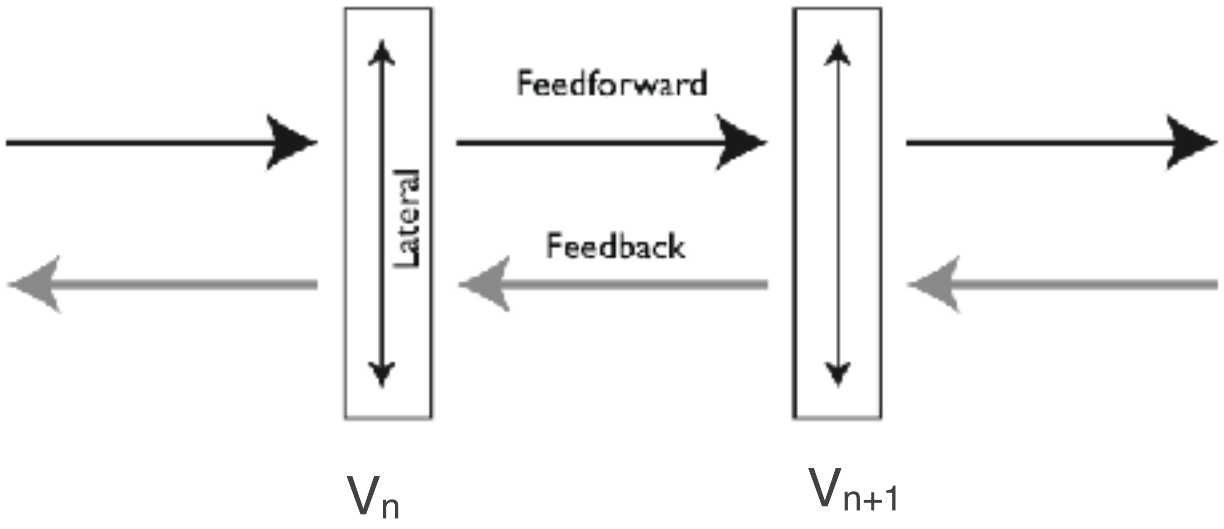
look for effects of spatial context on early, local processing



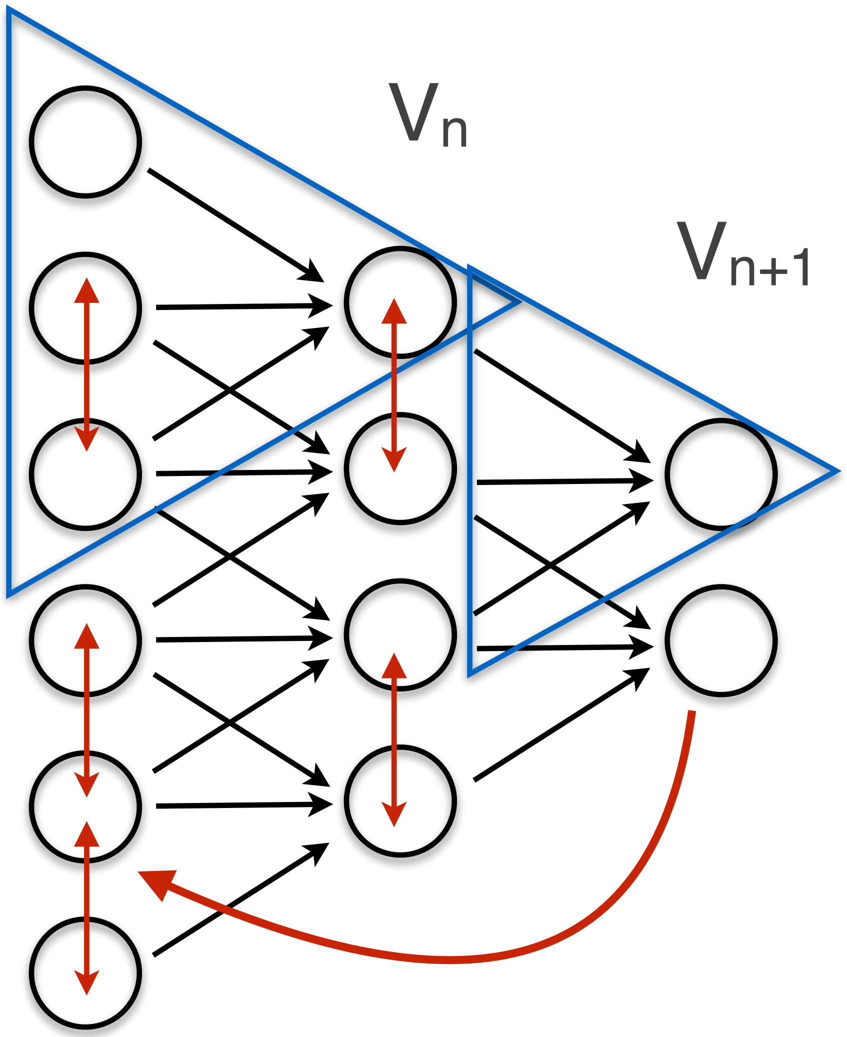
small receptive fields,
local features

larger receptive fields,
integration of features into
global forms

…some caveats

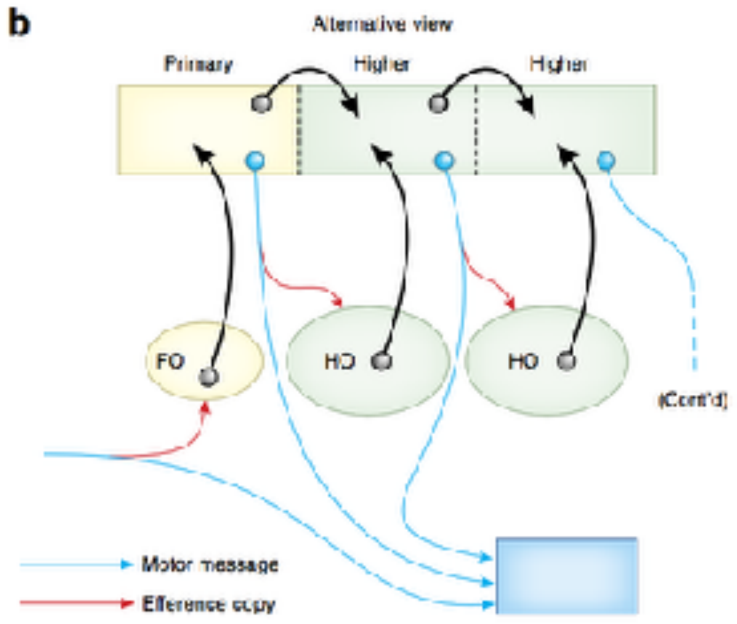


contextual information can be integrated feedforward, laterally within an area, and through feedback

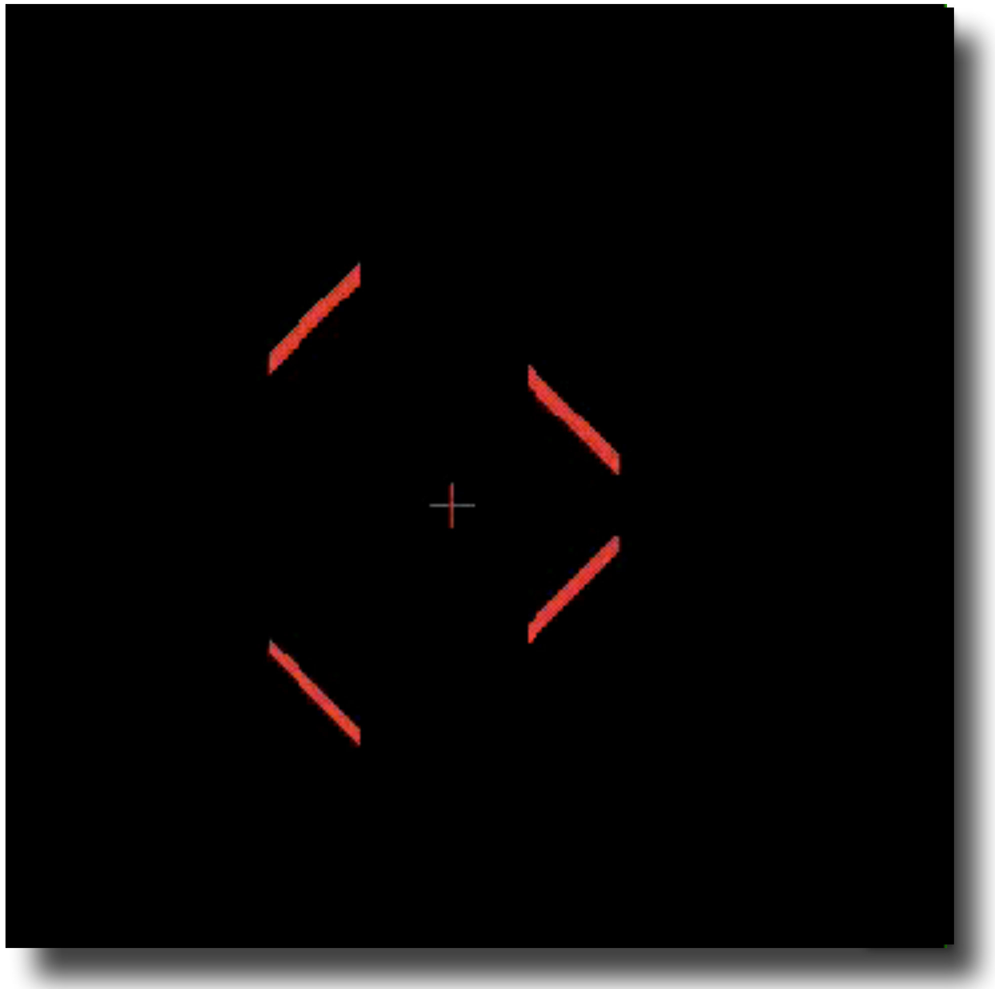


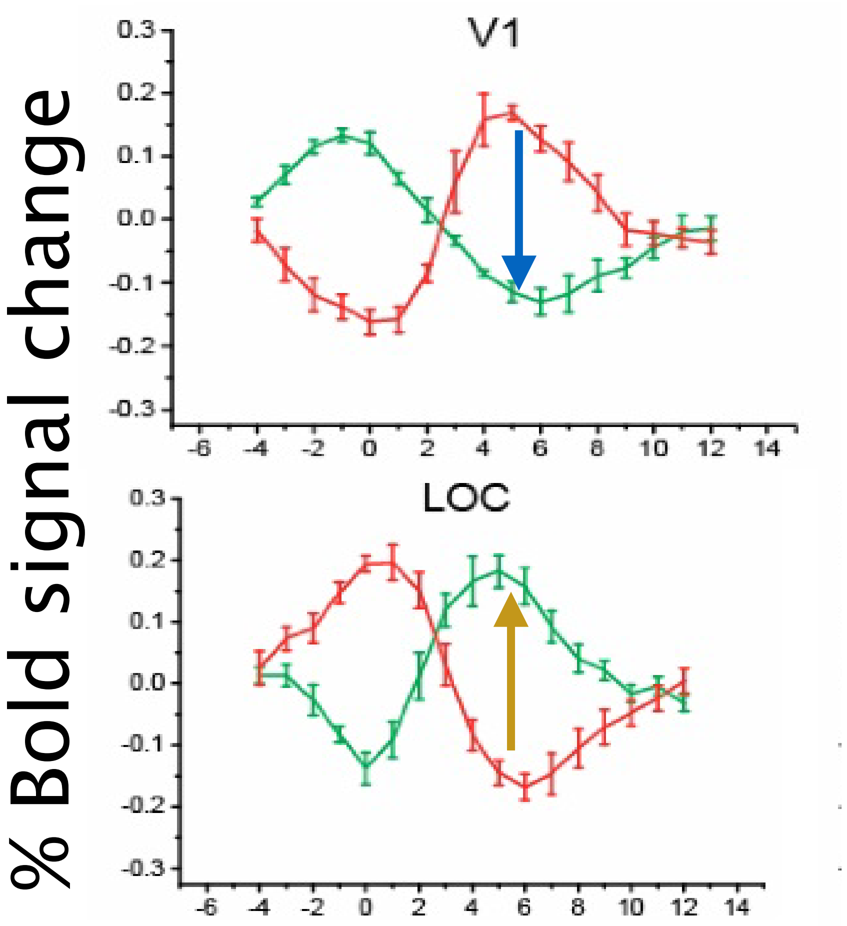
..and the elephant in the room

Sherman and Guillery



fMRI activity in V1



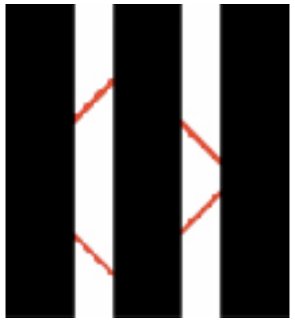


Diamond shape perceived

Line fragments perceived

V1 activity decreases when the diamond shape is perceived

one of the perceptual states - a “diamond” shape

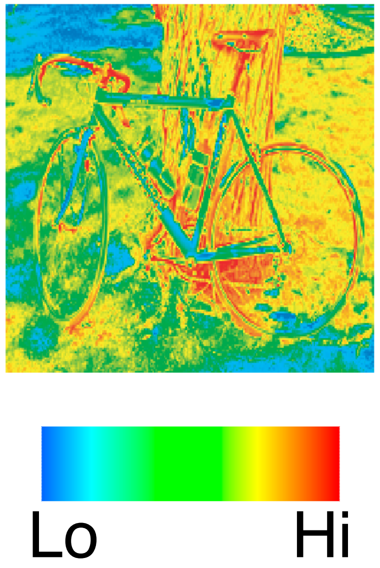


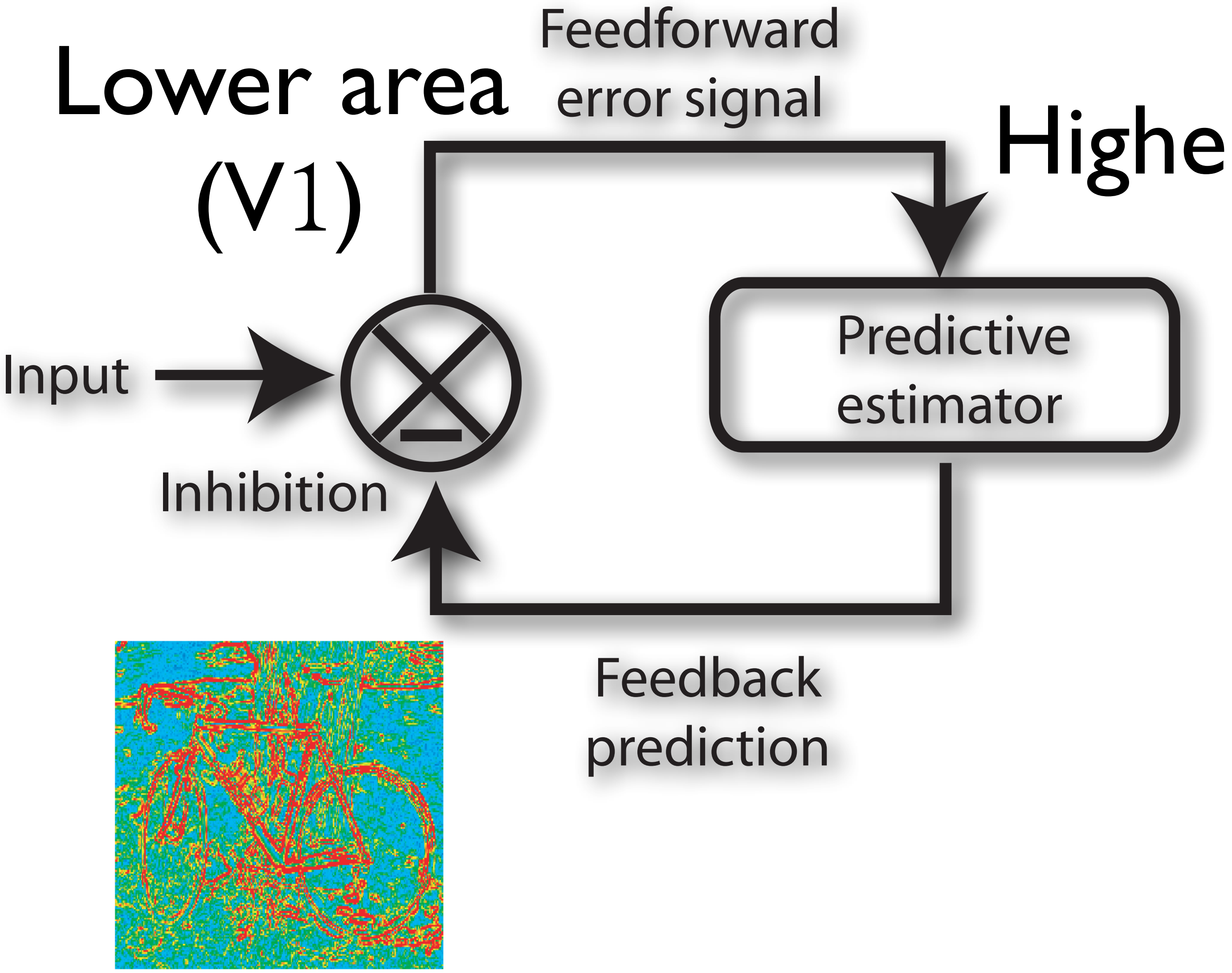
LOC—a high-level object area— activity is increases when the diamond shape is perceived

Murray, Kersten, Olshausen, Schrater, & Woods (2002)

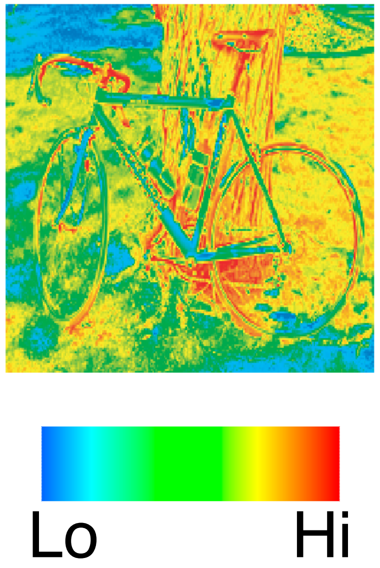
Fang, Boyaci, Kersten, Murray (2008)

“predictive coding” through suppression of consistent features at lower levels





e.g. Rao, R. P., & Ballard, D. H. (1997). Dynamic model of visual recognition predicts neural response properties in the visual cortex. Neural Comput, 9(4), 721-763.

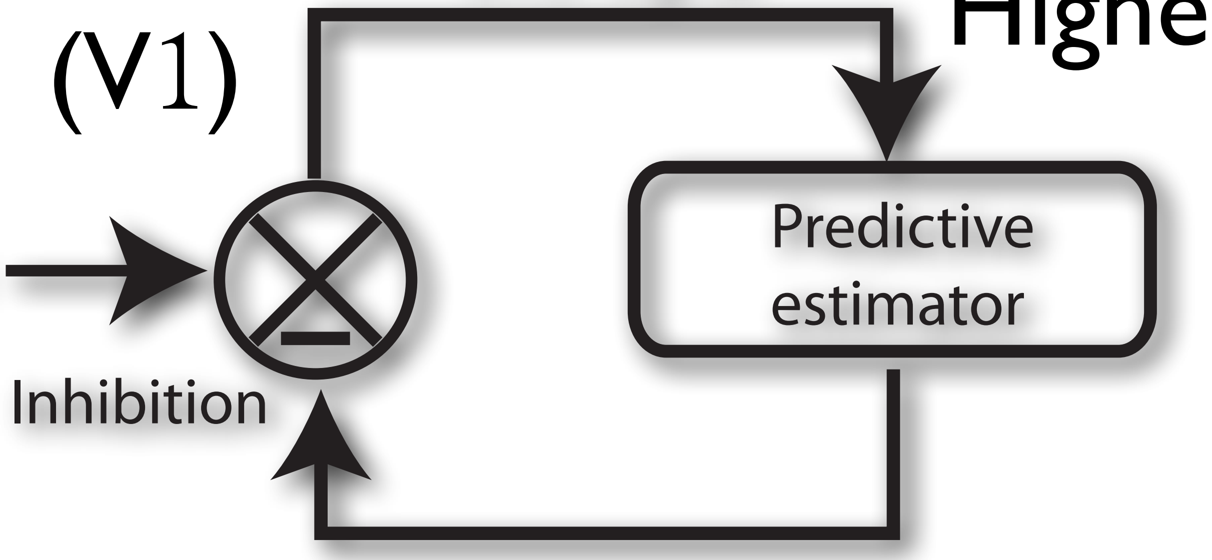


Lower area (V1)

Feedforward error signal

Higher area

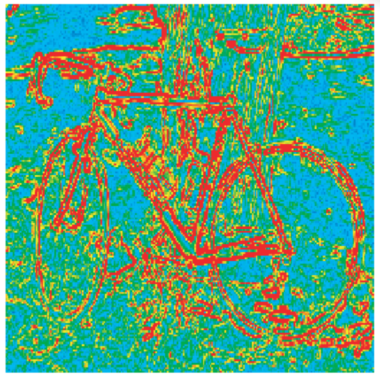
Input

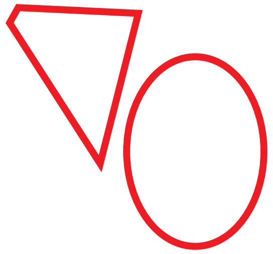


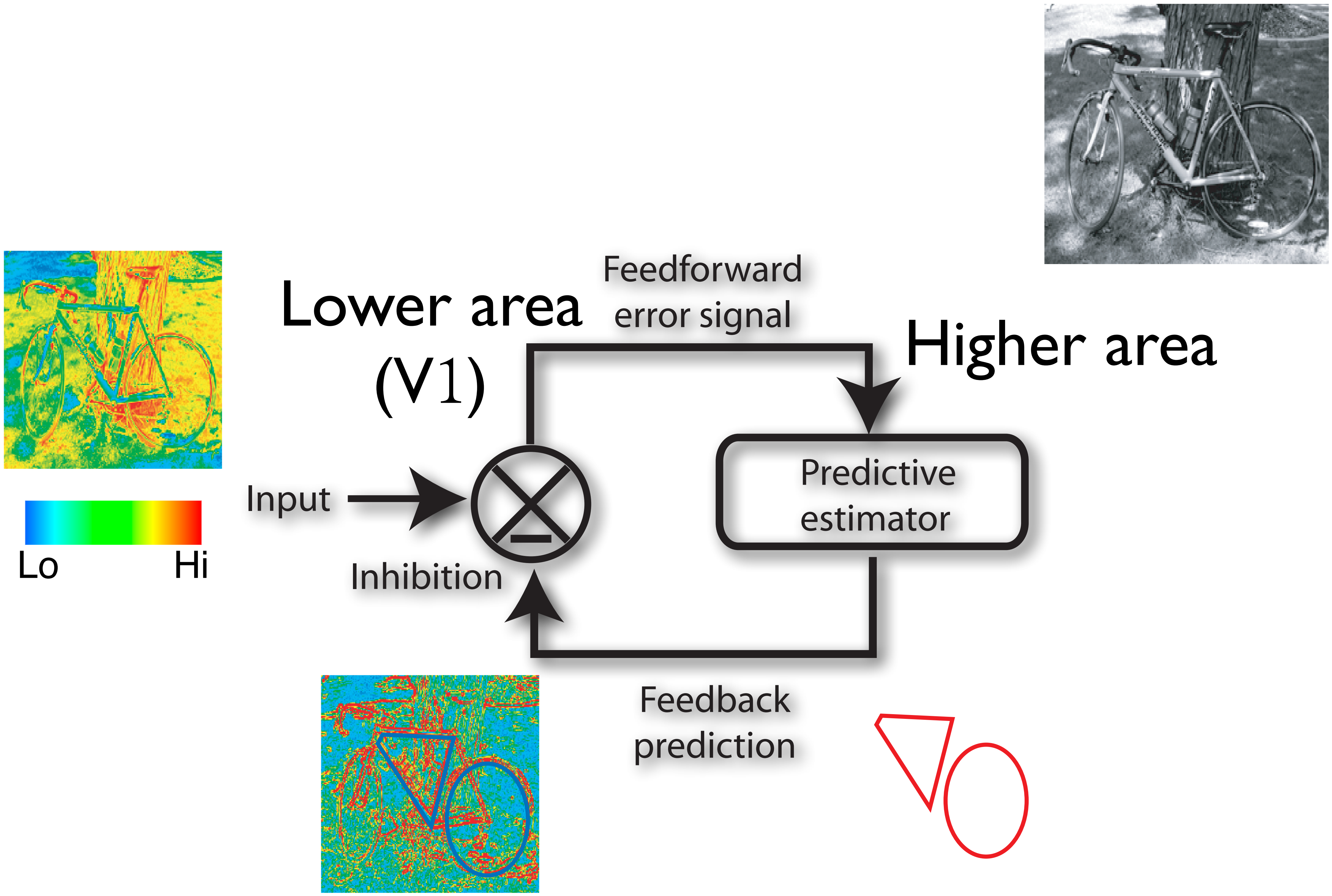
Predictive estimator

Inhibition

Feedback prediction







Lower area (V1)

Feedforward error signal

Higher area

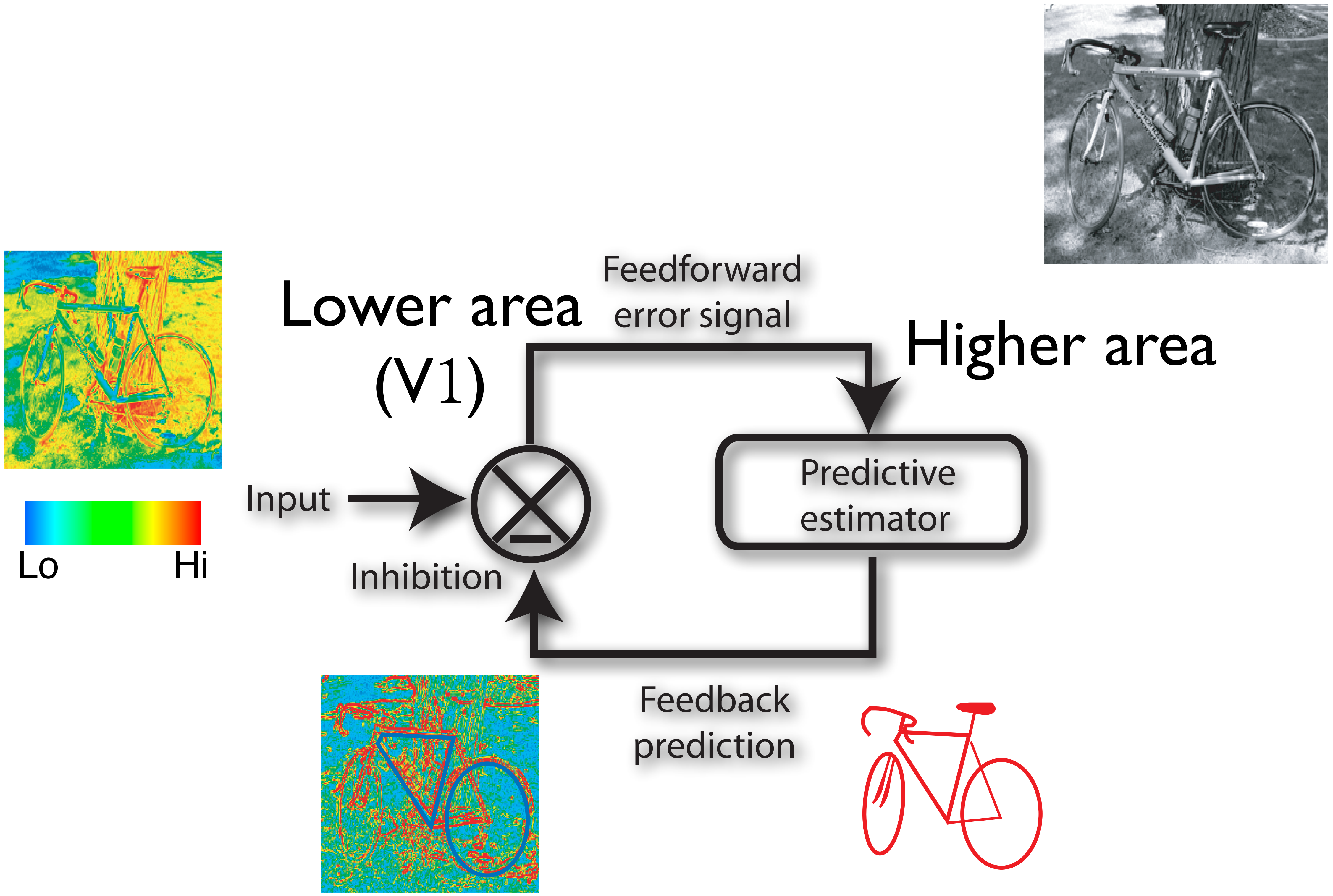
Input

Predictive estimator

Inhibition

Feedback prediction

Lo Hi



Lower area (V1)

Higher area

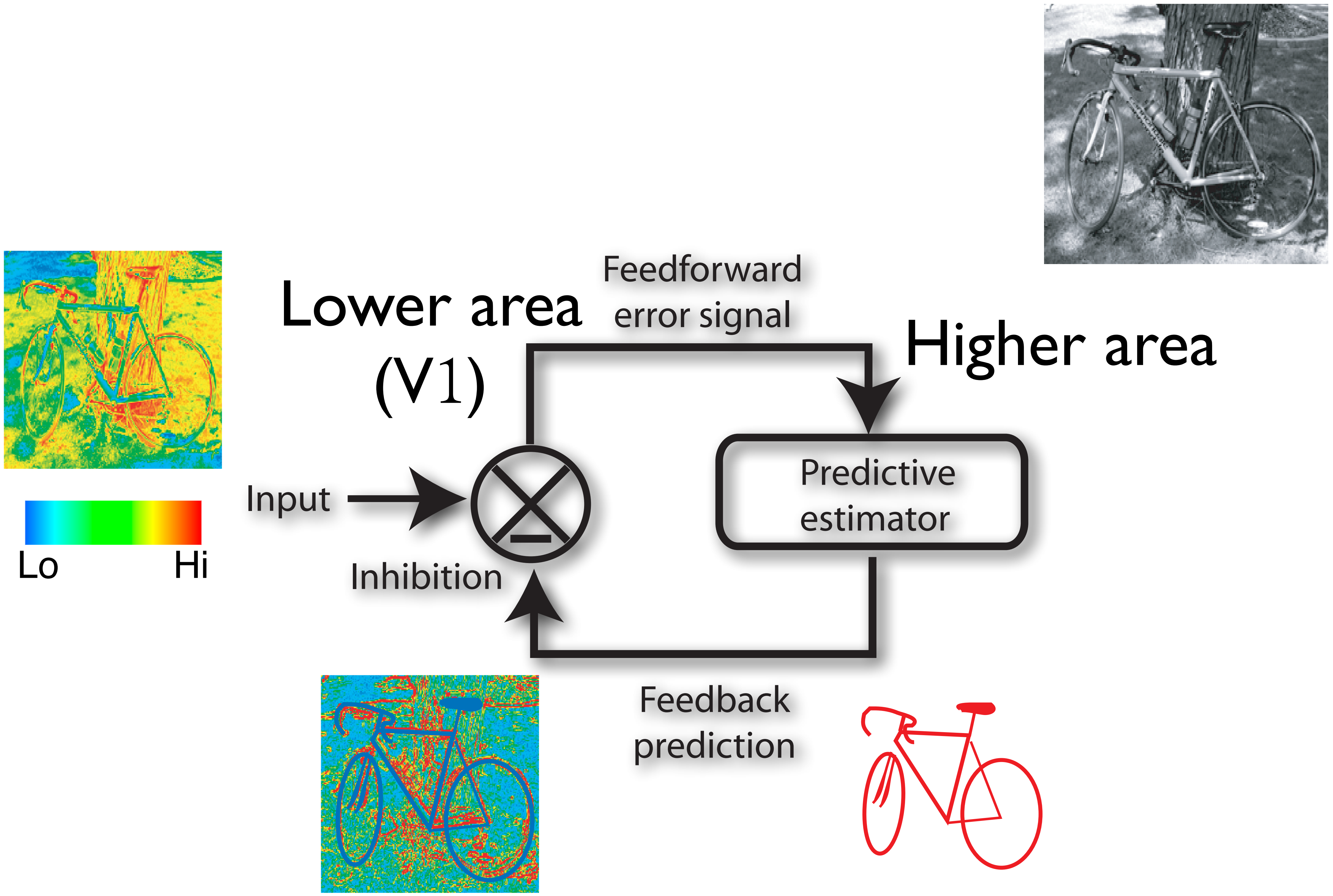
Feedforward error signal

Predictive estimator

Feedback prediction

Input

Inhibition



Lower area (V1)

Feedforward error signal

Higher area

Input

Predictive estimator

Inhibition

Feedback prediction

Lo

Hi

summary: resolve ambiguity using high-level knowledge

Exploit the hierarchical organization of object knowledge, and use feedback to solve ambiguity through "explaining away"

"predictive coding" as top-down error detection

- suppress lower-level responses to features "explained" by a higher-level interpretation

  and/or amplify those responses ("residuals") that are not explained

cf. Mumford, 1992; Rao & Ballard, 1999

Bastos, A. M., Usrey, W. M., Adams, R. A., Mangun, G. R., Fries, P., & Friston, K. J. (2012). Canonical Microcircuits for Predictive Coding. *Neuron*, *76*(4), 695–711.

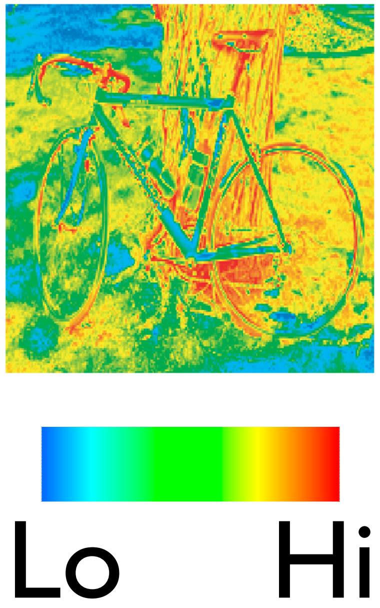
…summary so far

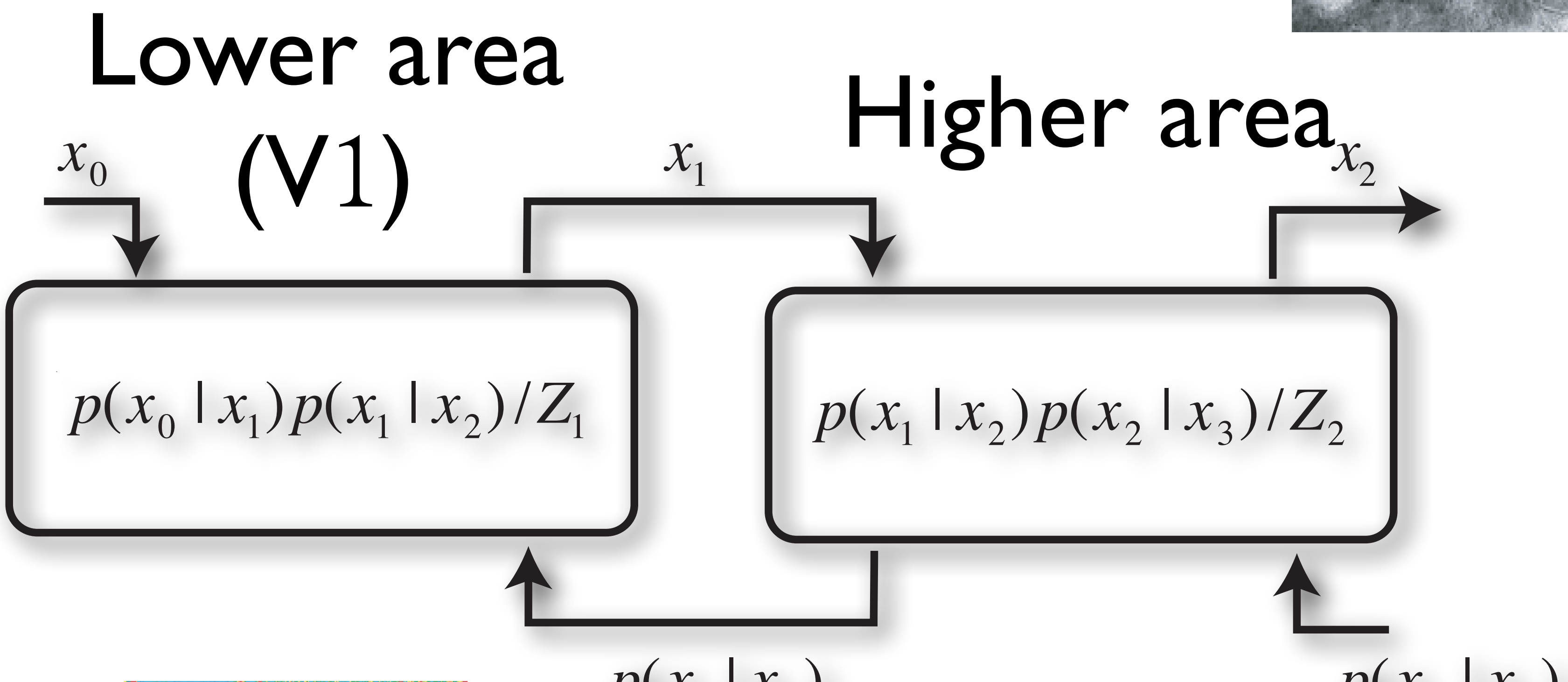
Evidence for suppression of local activity in V1 as a consequence of higher-level, global perceptual organization—i.e. suppression when all the local features have been "explained".

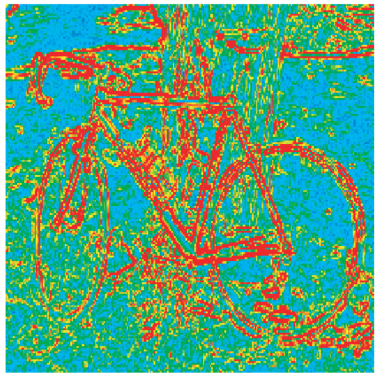
$$p(S|I) \propto p(I - f(S))p(S)$$

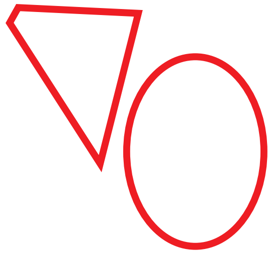
But is there another explanation for the fMRI results?

binding through enhancement of consistent features at lower levels

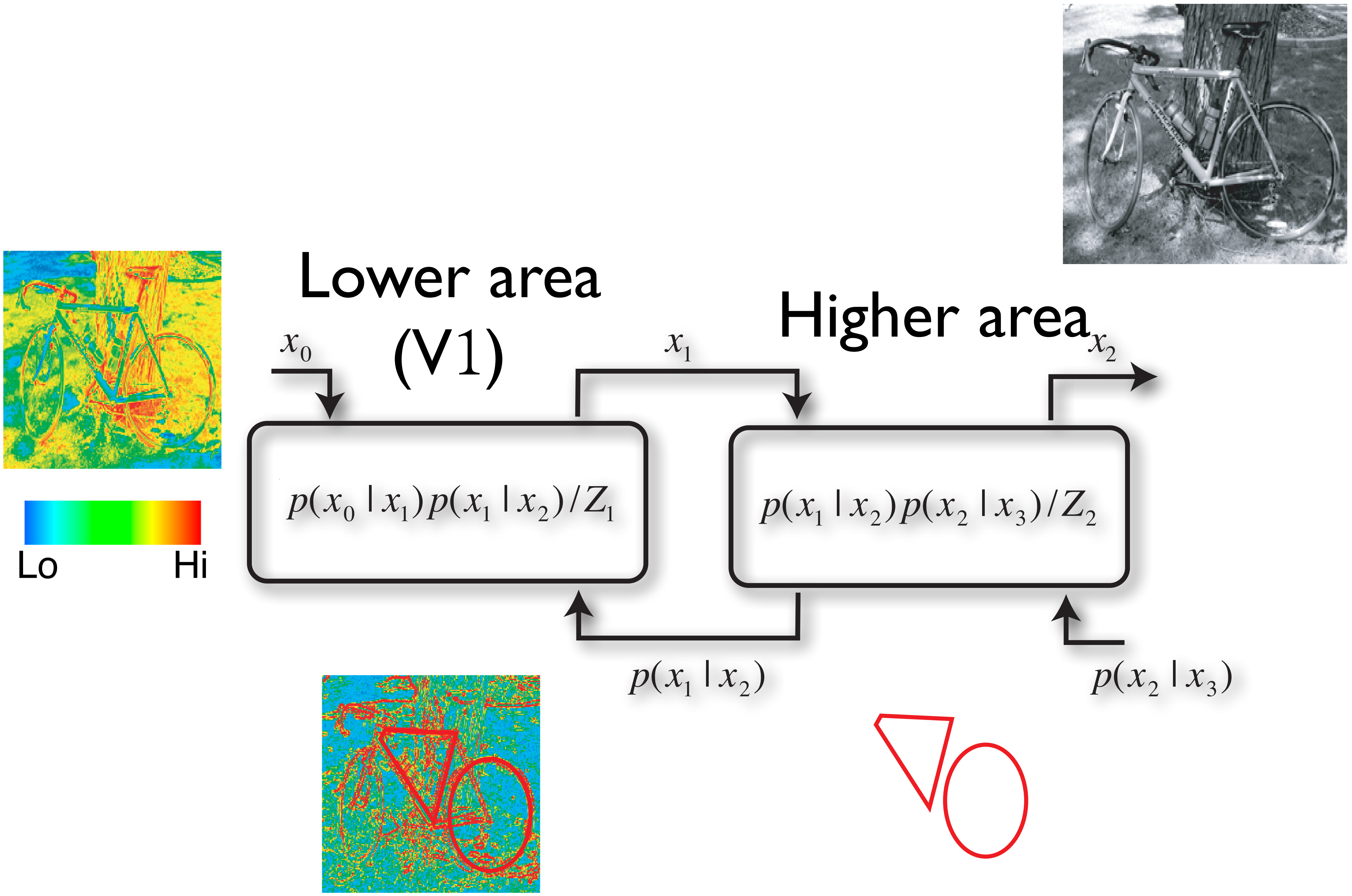








Lee & Mumford, 2003, JOSA



Lower area (V1)

Higher area

x_0

x_1

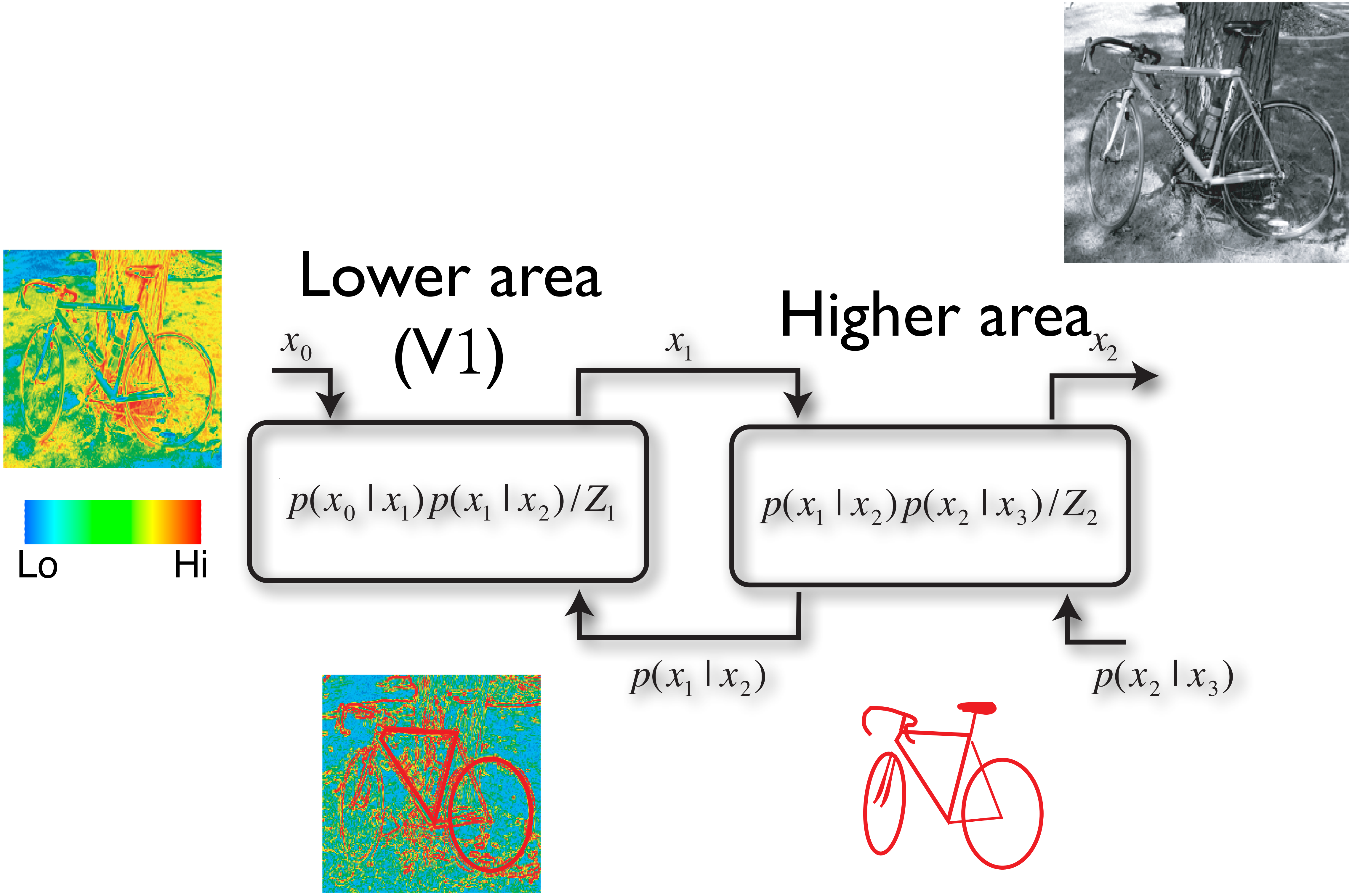
x_2

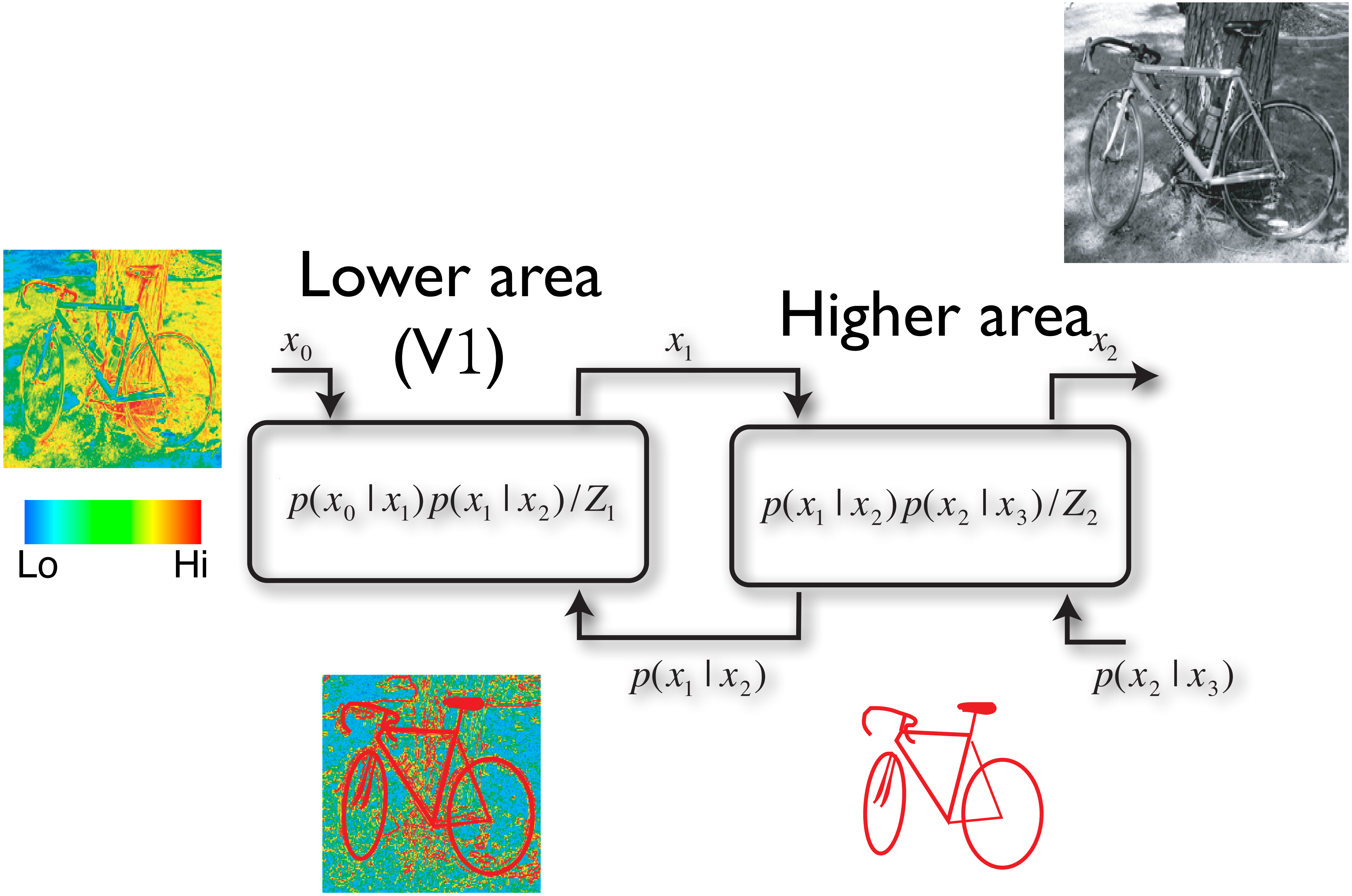
$p(x_0 \mid x_1)p(x_1 \mid x_2)/Z_1$

$p(x_1 \mid x_2)p(x_2 \mid x_3)/Z_2$

$p(x_1 \mid x_2)$

$p(x_2 \mid x_3)$





Lower area (V1)

Higher area

x_0

x_1

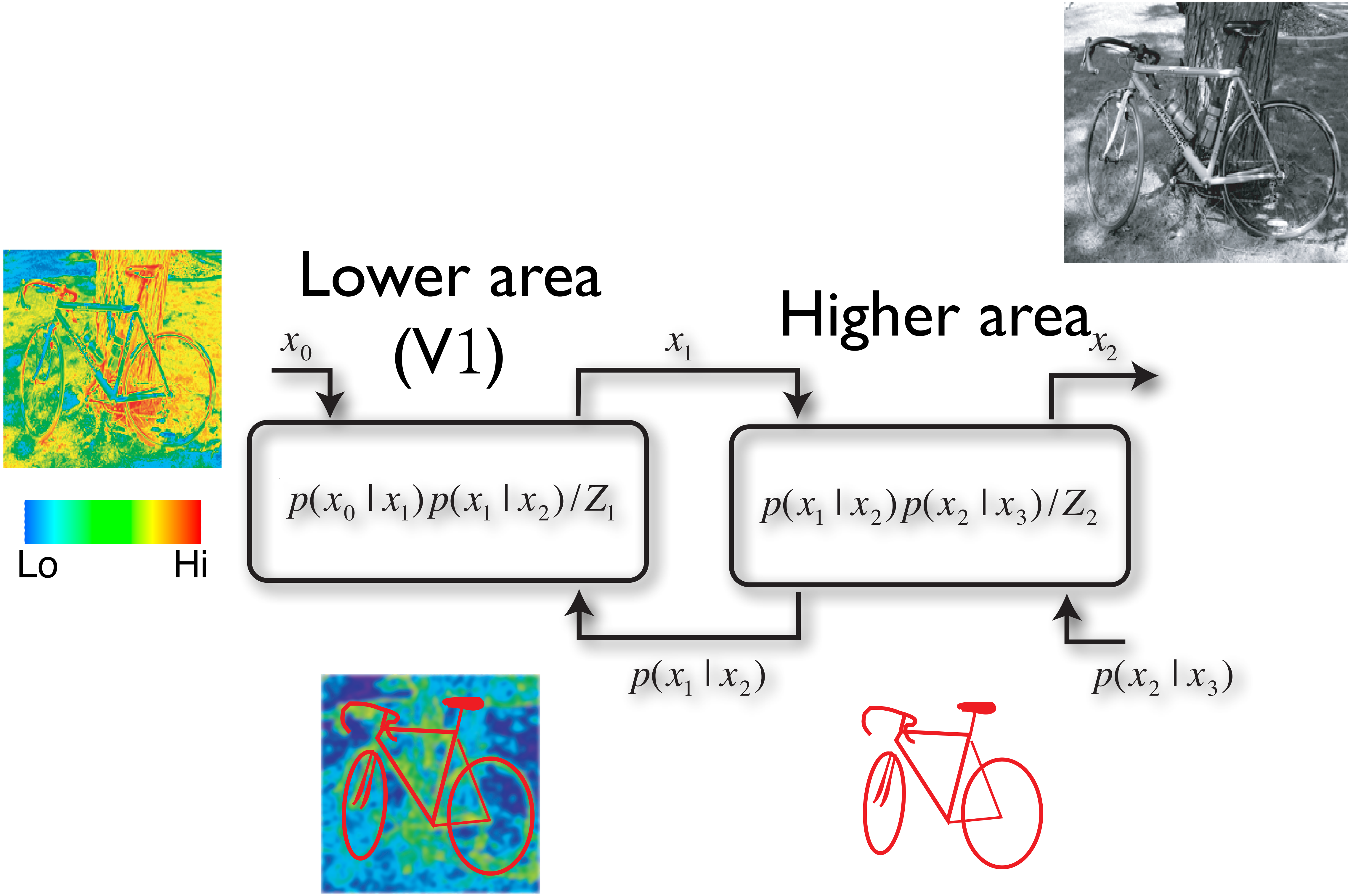
x_2

$p(x_0 \mid x_1) p(x_1 \mid x_2) / Z_1$

$p(x_1 \mid x_2) p(x_2 \mid x_3) / Z_2$

$p(x_1 \mid x_2)$

$p(x_2 \mid x_3)$



Lower area (V1)

Higher area

x_0

$p(x_0 | x_1) p(x_1 | x_2) / Z_1$

x_1

$p(x_1 | x_2) p(x_2 | x_3) / Z_2$

x_2

$p(x_1 | x_2)$

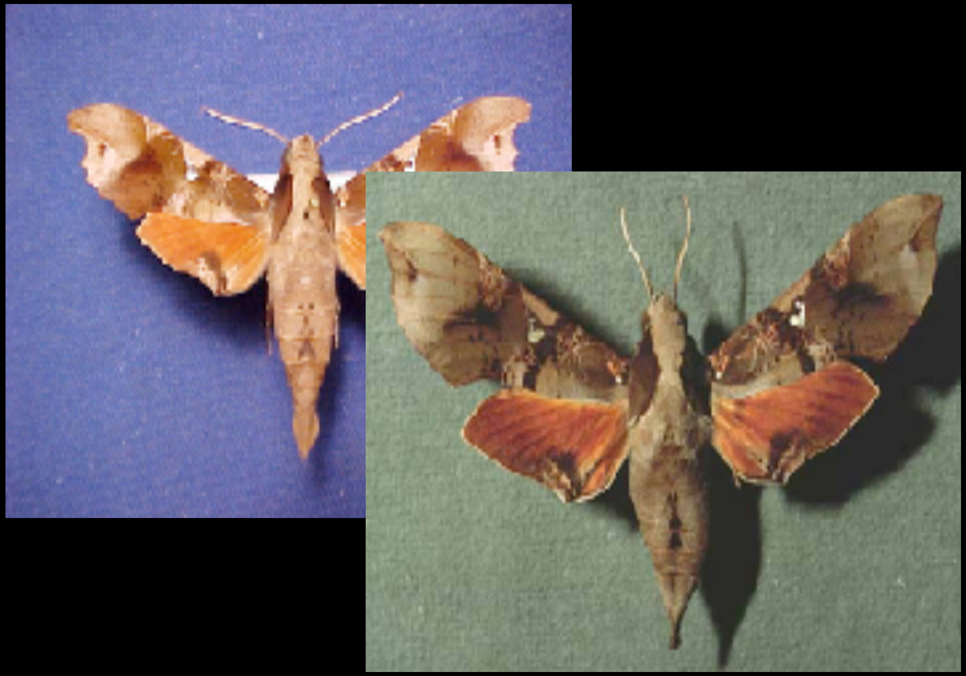
$p(x_2 | x_3)$

Return to the challenge of task flexibility

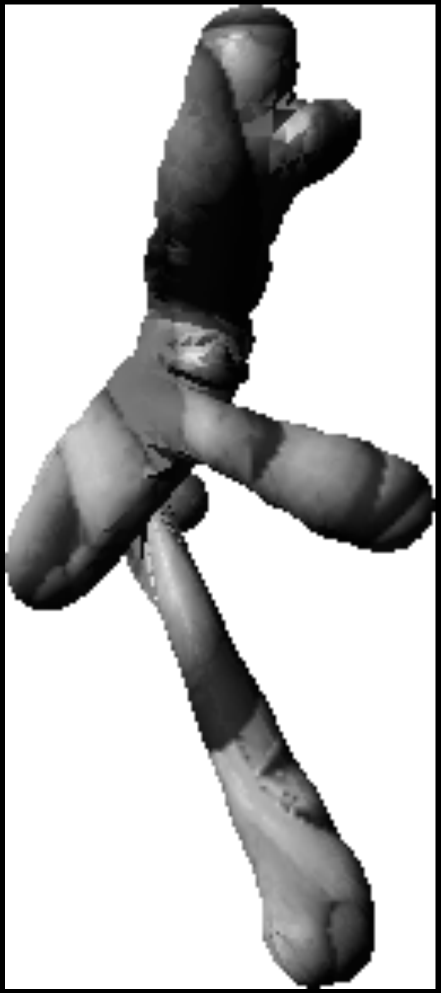
Humans can not only localize and recognize object categories, they can

- parse, describe and precisely segment an image, and lots more, such as measure attributes and relations, infer intent, …
- rapidly learn new object models under difficulty segmentation conditions

Learning to recognize and segment camouflaged novel objects can be done quickly



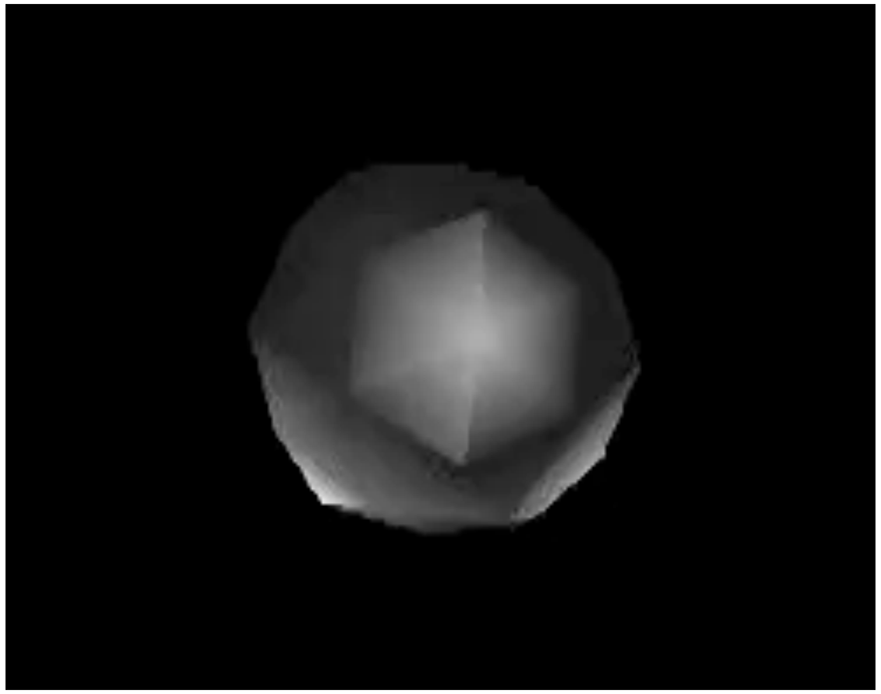
200-Gram Thinsulate Insulation
NEW COLOR
Available in Tall, Regulars & Stouts
Sizes to 15



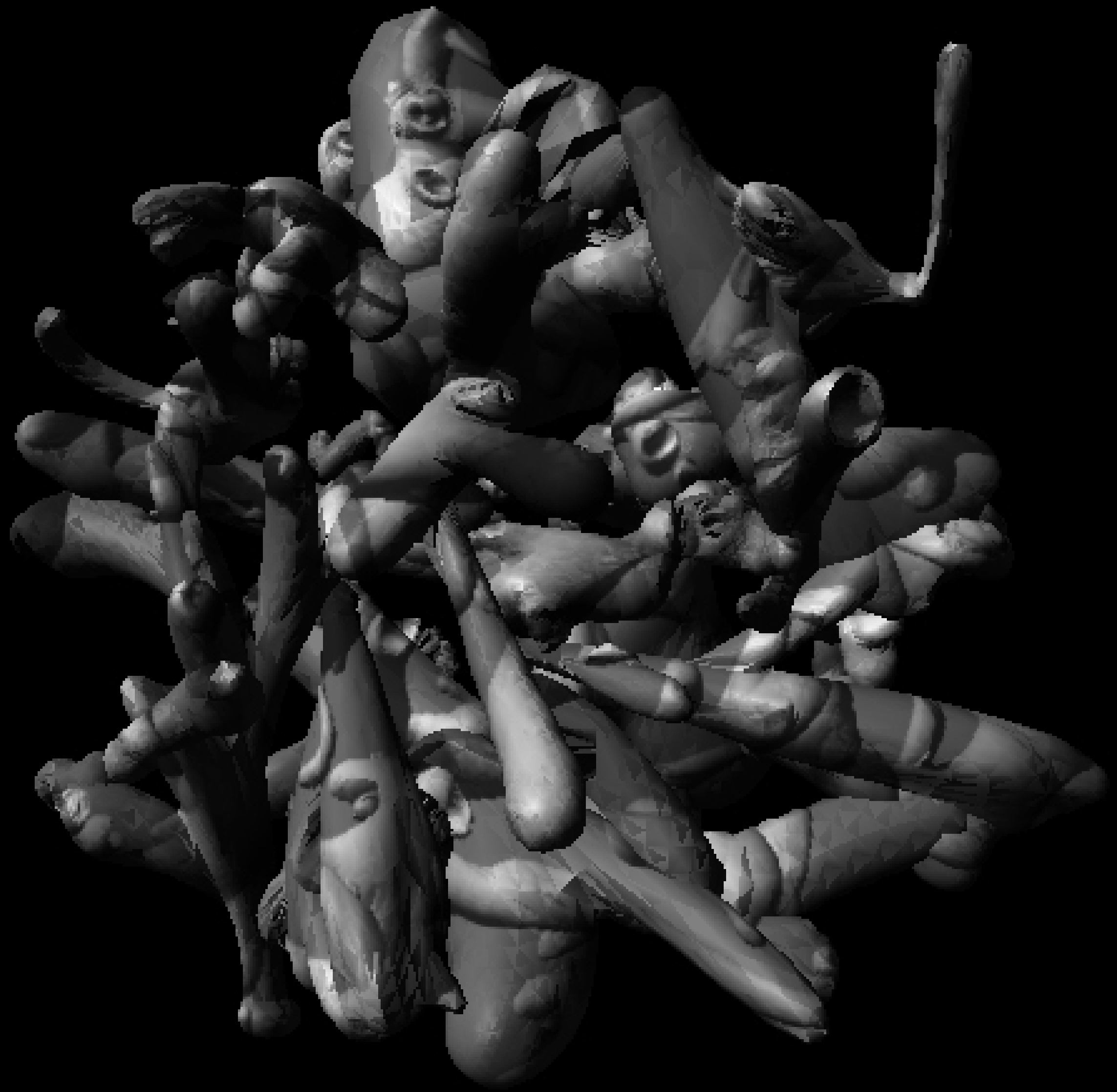
Bringing nature to the lab. Callionima moths on the left show disruptive camouflage patterns. The garment in the middle image replicate images of the woodland background. Texture mapping was used by Brady and Kersten (2003) to mimic this for the digital embryo on the right. This introduces false positive object boundaries, and apparent shape from shading cues.

Virtual morphogenesis

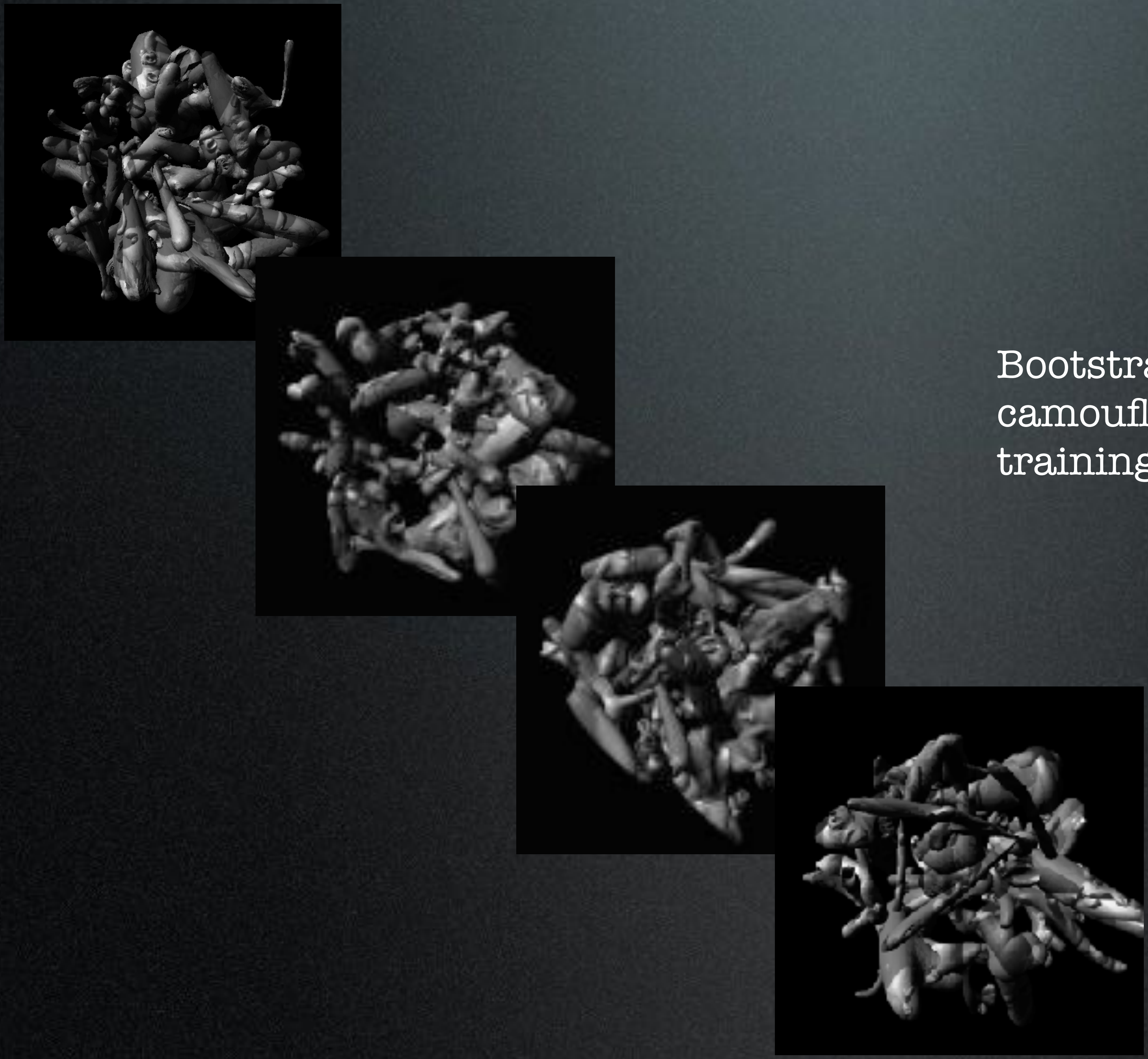
Brady, M. J., & Kersten, D. (2003).
Bootstrapped learning of novel objects.
Journal of Vision, 3(6), 413–422.



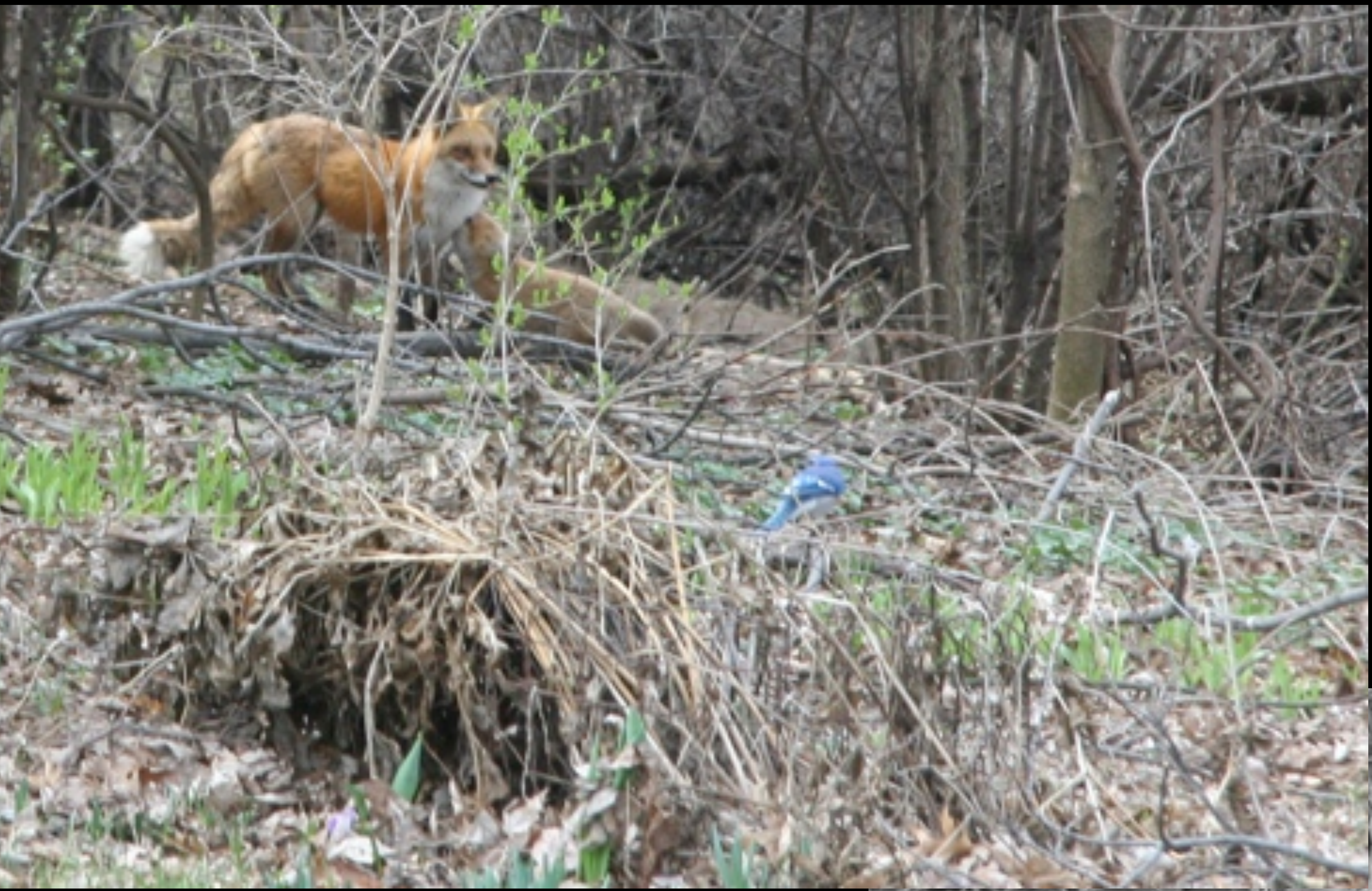
Brady, M. J., & Kersten, D. (2003).
Bootstrapped learning of novel objects.
Journal of Vision, 3(6), 413–422.

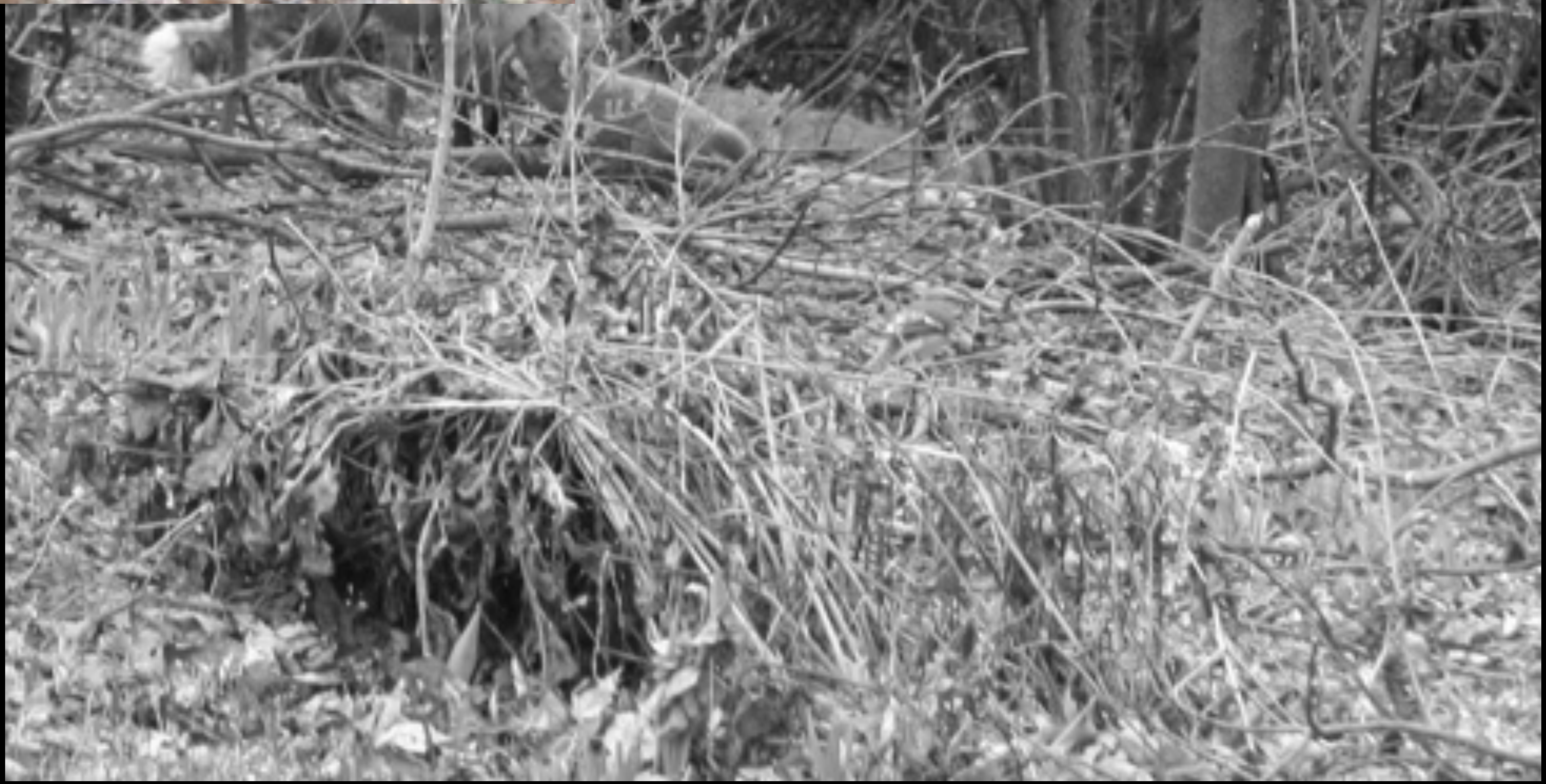


How do humans acquire prior knowledge of object classes? There is a target object in “plain view” in this figure. Without training, it is impossible to detect or draw a line around its boundary.

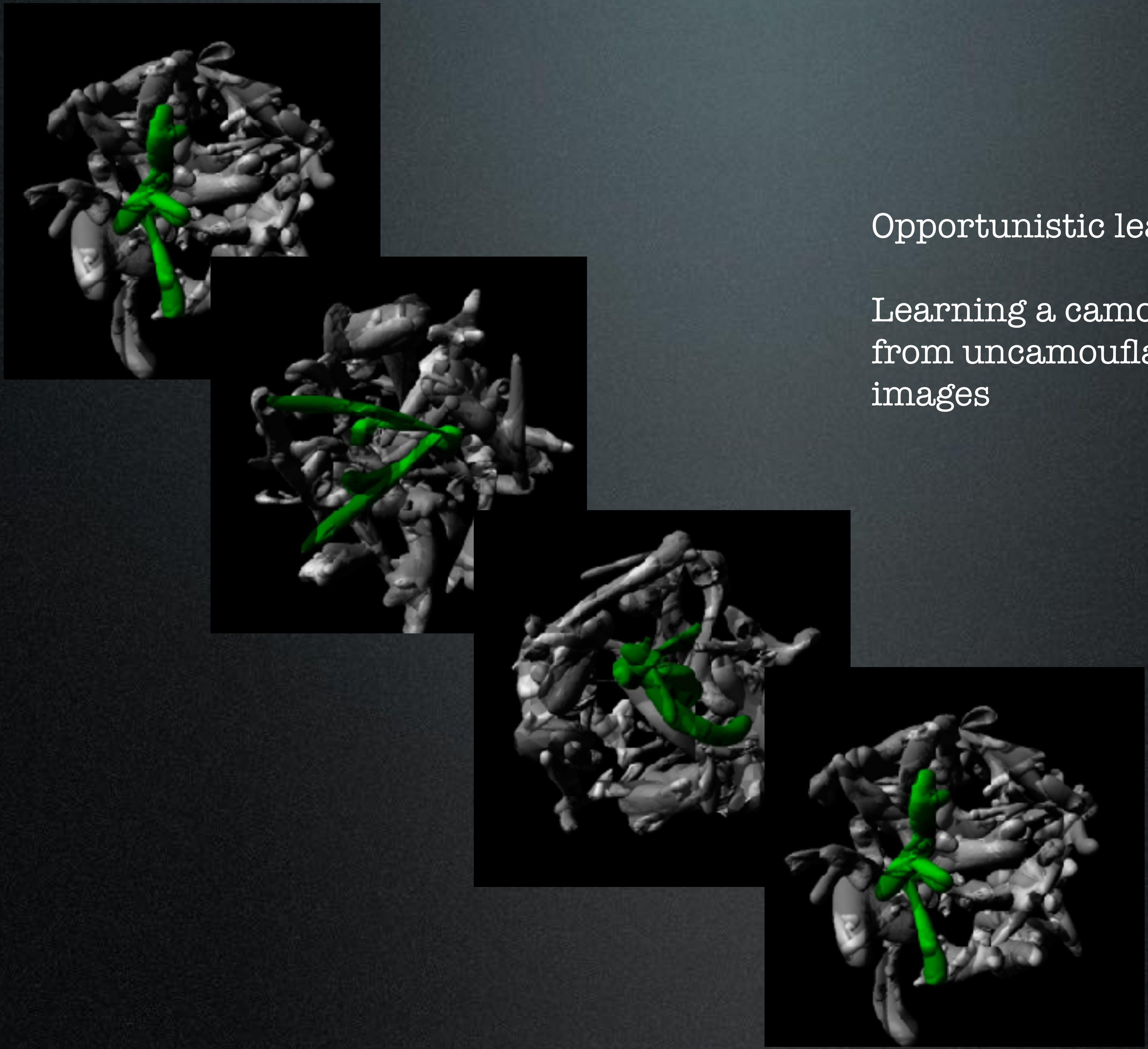


Bootstrapped learning: Learning a camouflaged object from camouflaged training images





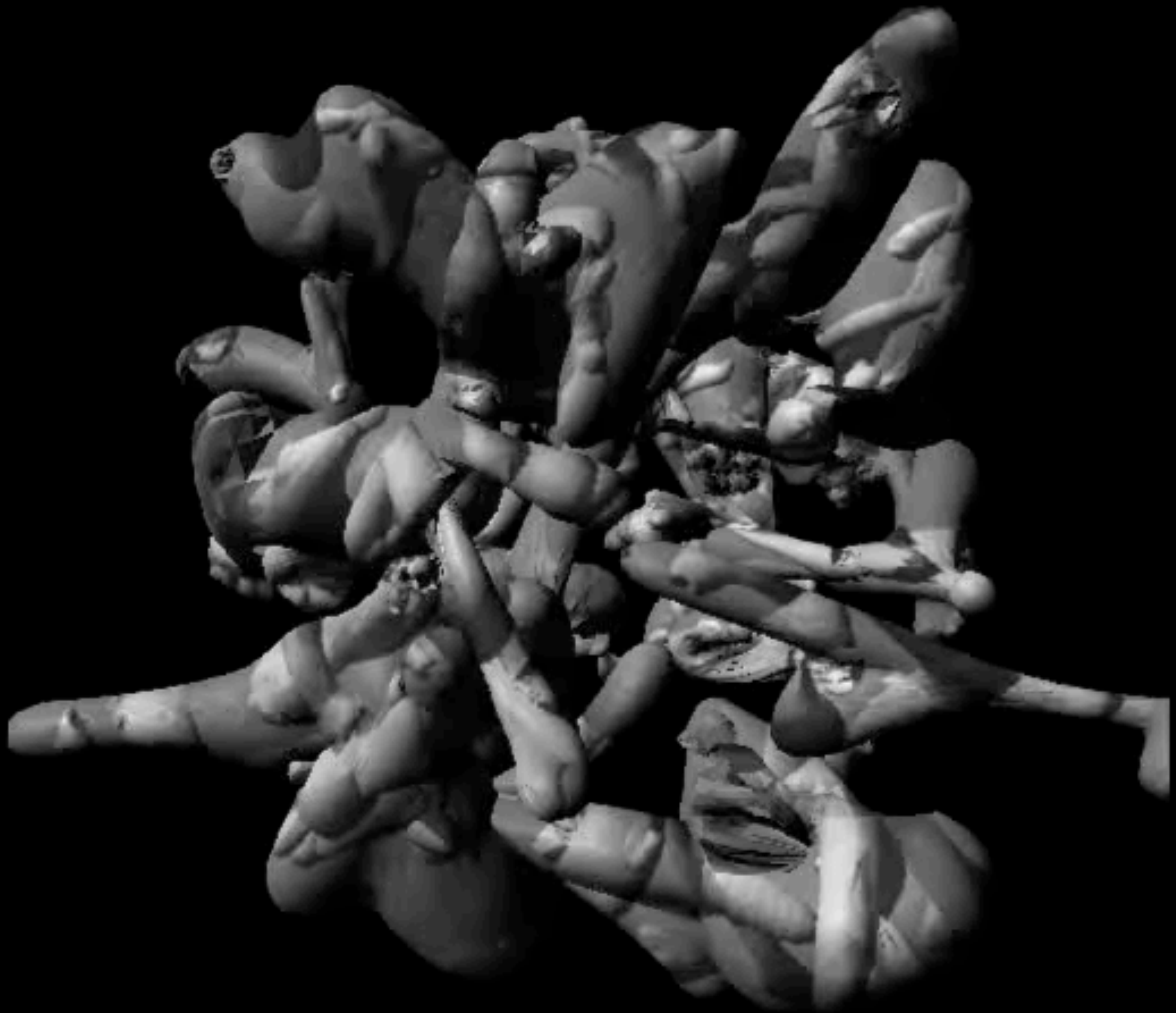
If an observer has to opportunities to see colored birds, this could help the observer to learn about the forms that birds can take. Then at some future time, it could use this knowledge to see birds whose color does not distinguish it from the background, e.g. a different kind of bird, or under more difficult viewing conditions, such as during the night or fog. This is “opportunistic learning”

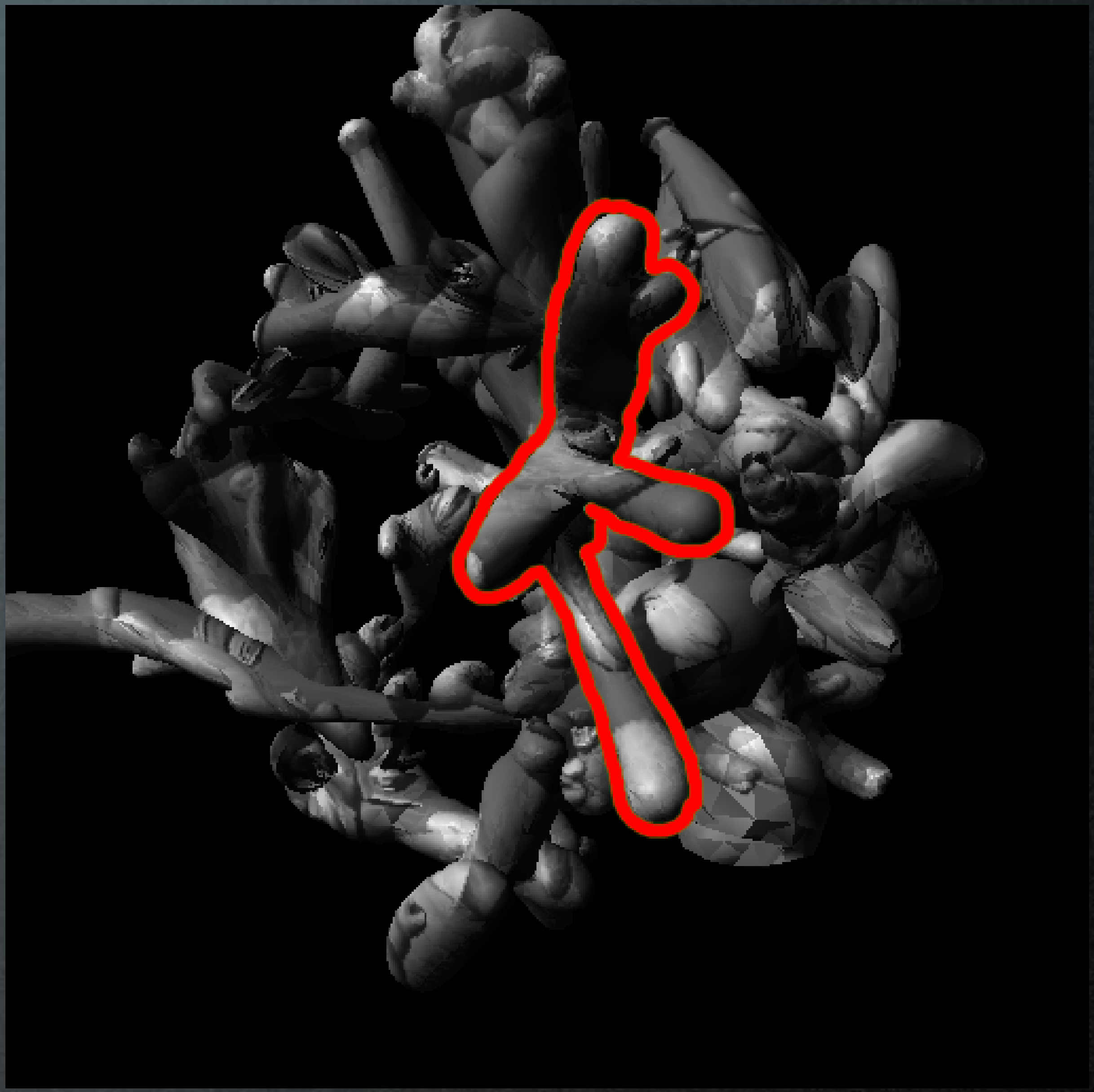


Opportunistic learning:

Learning a camouflaged object from uncamouflaged training images

First 4 scenes (out of 15) of a motion training movie.





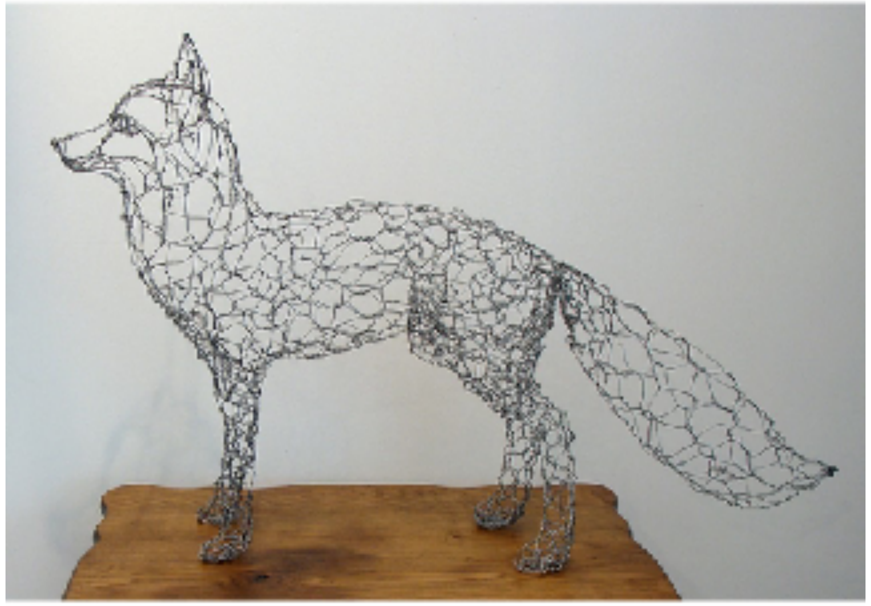
After training, observers were tested on test images in which the objects were given new camouflage, and presented against new backgrounds.

All observers were able to learn opportunistically, and some were also able to learn from the camouflaged training images. This figure shows a perfect segmentation by an observer after training.

Flexibility

Limitations to current recognition algorithms as models of biological/human vision?

Humans generalize far beyond training data to novel images/ forms

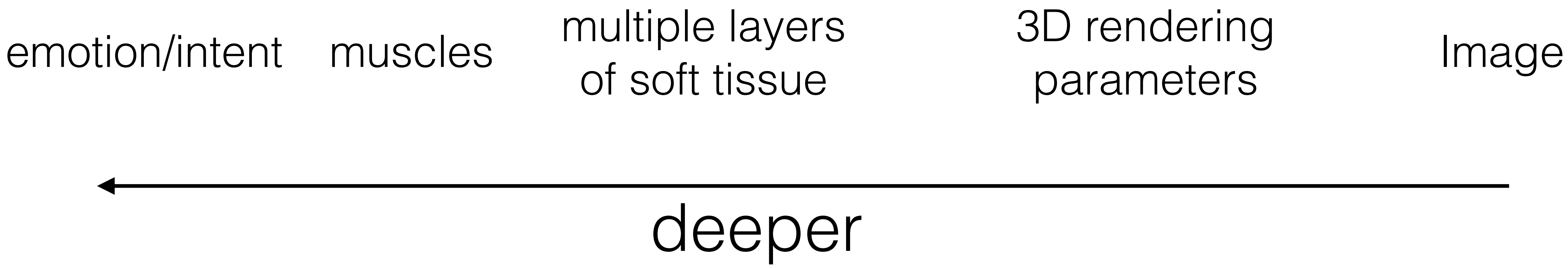


To what extent does human visual flexibility, ability to generalize rely on deep generative knowledge?

How deep?

depth in terms of causes,
not network depth

http://www.pauldebevec.com



emotion/intent

muscles

multiple layers
of soft tissue

3D rendering
parameters

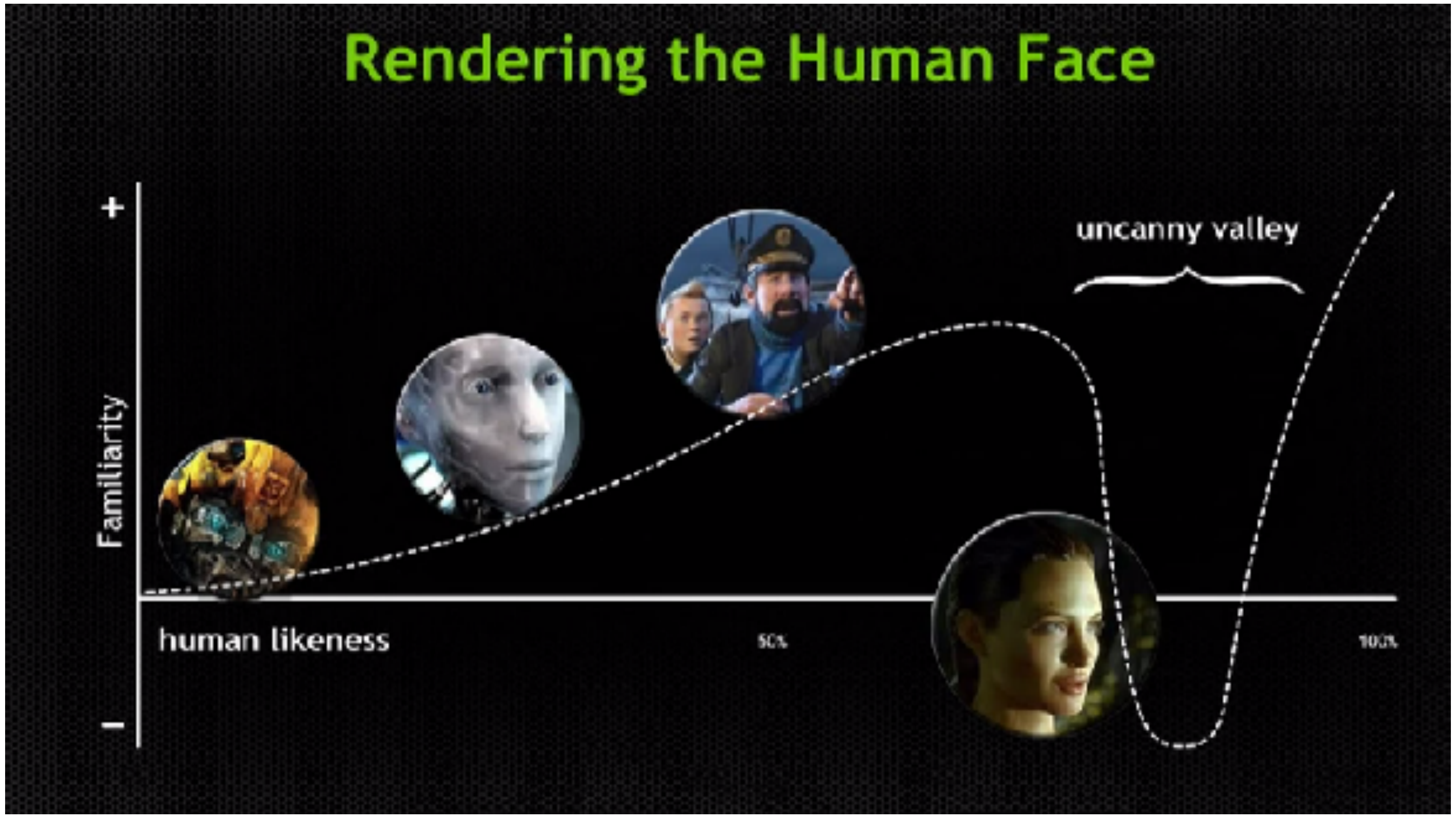
Image

deeper

Insights from computer graphics…

Take a look at faces, materials such as *hair and fluids*, and *body pose*

Message from computer graphics is as deep as you can given processing limitations

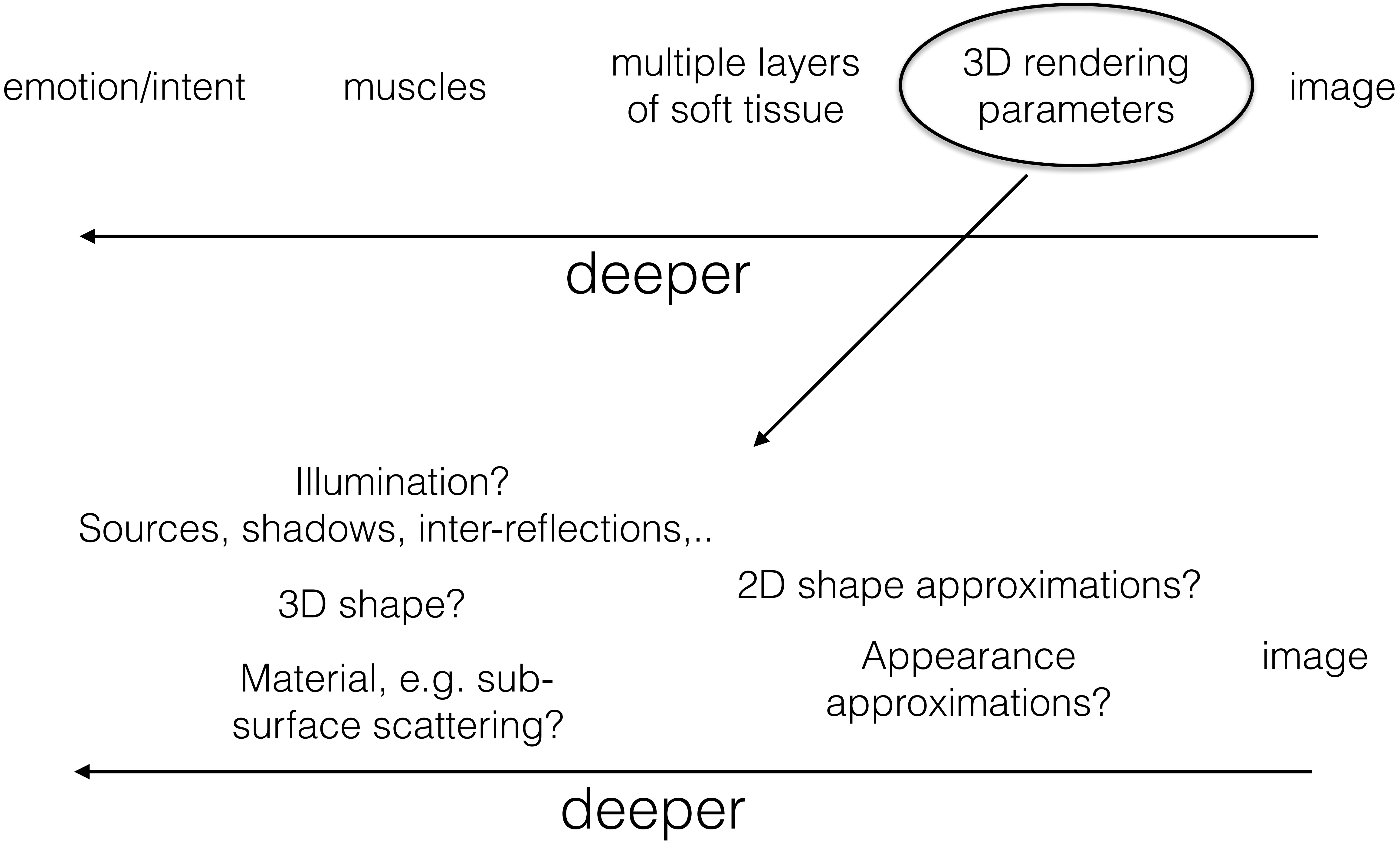


https://developer.nvidia.com/faceworks

General message for human visual neuroscience is “deep, but not too deep”.

“How to cheat and get away with it?”

How deep?



emotion/intent

muscles

multiple layers of soft tissue

3D rendering parameters

image

deeper

Illumination?
Sources, shadows, inter-reflections,..

3D shape?

Material, e.g. sub-surface scattering?

2D shape approximations?

Appearance approximations?

image

deeper

Ira — NVIDIA Face Works

https://youtu.be/7fqEAzMZhJ

Hair

hair care products have the highest sale volume of all non-food items in the US

What does it take to generate realistic hair?

'Pan'

G McKenna.

Sasquatch software trom www.worley.com

Hair can be…

- wavy, curly, straight, spiky, stiff, buzzed, shaved, parted, neatly-combed, tamed, long, short, cropped
- thick, full, lustrous, bushy, coarse, wiry
- thin, scraggly, fine, baby-fine, wispy, limp, flat, balding, receding
- black, brunette, brown, chestnut-brown, honey-blond, blond, golden-blond, ash-blond, auburn, red, strawberry-blond, gray, silver, white, salt-and-pepper
- permed, dyed, bleached, highlighted, weaved
- braids, ponytail, pigtails, bun, twist, bob, ringlets, flip, bangs, buzz
- layered, feathered, chopped, gelled, spiked, slicked down
- terminal and vellus

Viscous fluids

Discrete Viscous Threads

Miklós Bergou	Columbia University
Basile Audoly	UPMC Univ Paris 06 & CNRS
Etienne Vouga	Columbia University
Max Wardetzky	Universität Göttingen
Eitan Grinspun	Columbia University

Body pose, actions

Toshev, A., & Szegedy, C. (2013). Deeppose: Human pose estimation via deep neural networks. arXiv Preprint arXiv:1312.4659.

X. Chen and A.L. Yuille. Articulated Pose Estimation with Image-Dependent Preference on Pairwise Relations. NIPS 2014 global

http://astro.temple.edu/~tshipley/ptlt_movies/mlwalk2.mov

http://www.biomotionlab.ca/Demos/BMLwalker.html

local

how to get from local to global?

Current inferential models of human visual recognition are not very "deep" in the sense of relying on inductive biases, generative models that could allow rapid learning from few samples, the ability to deal with almost any image (familiar or not).

Need to understand the critical dimensions that avoid the uncanny valley without computations and representations unlikely to exist in the brain. I.e. the "right" kind of generative model.

Need to understand how to model statistical regularities in classes of natural images. Linear methods are inadequate.

Need for compositional models, grammars, in the spirit of "recognition-by-components"