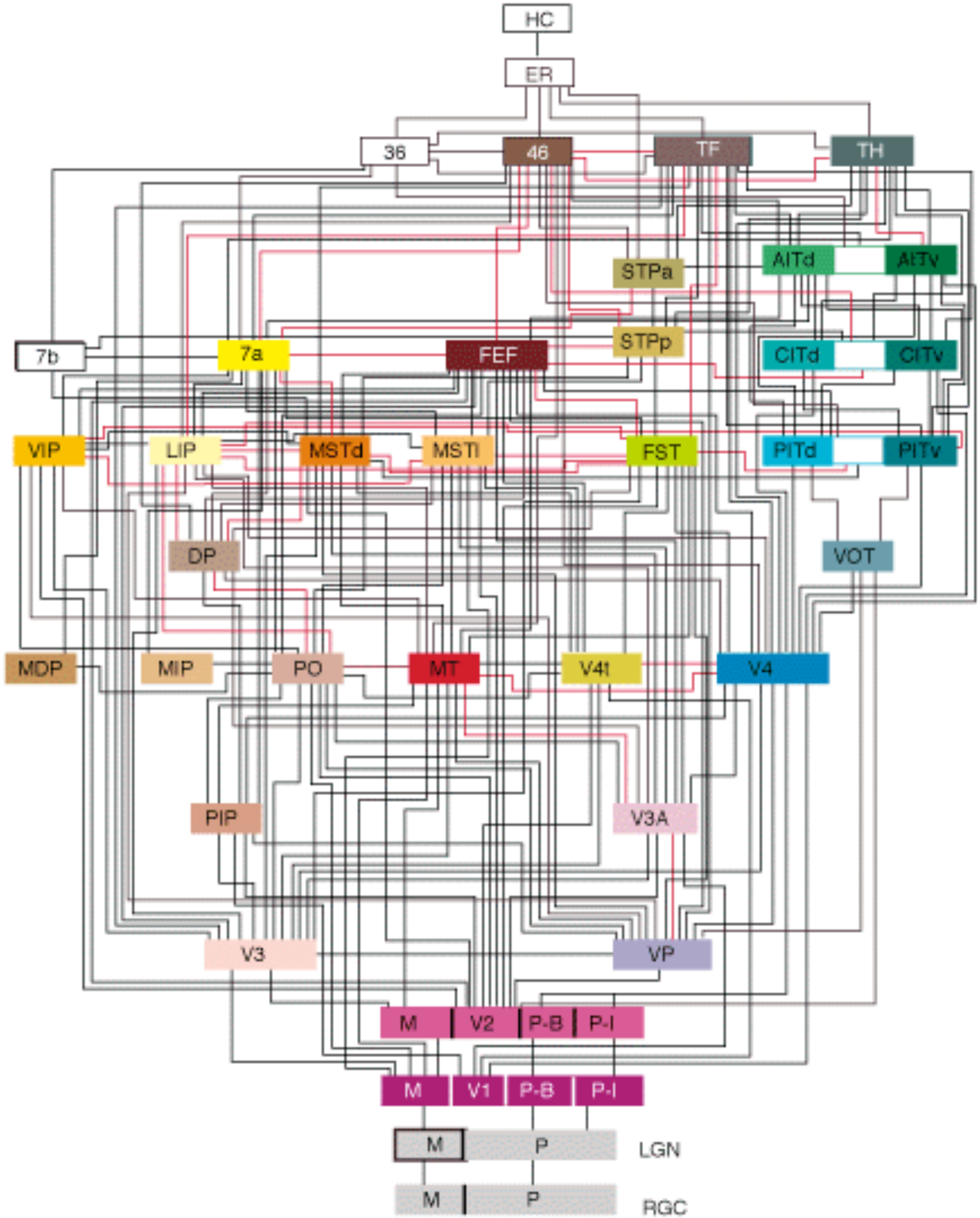
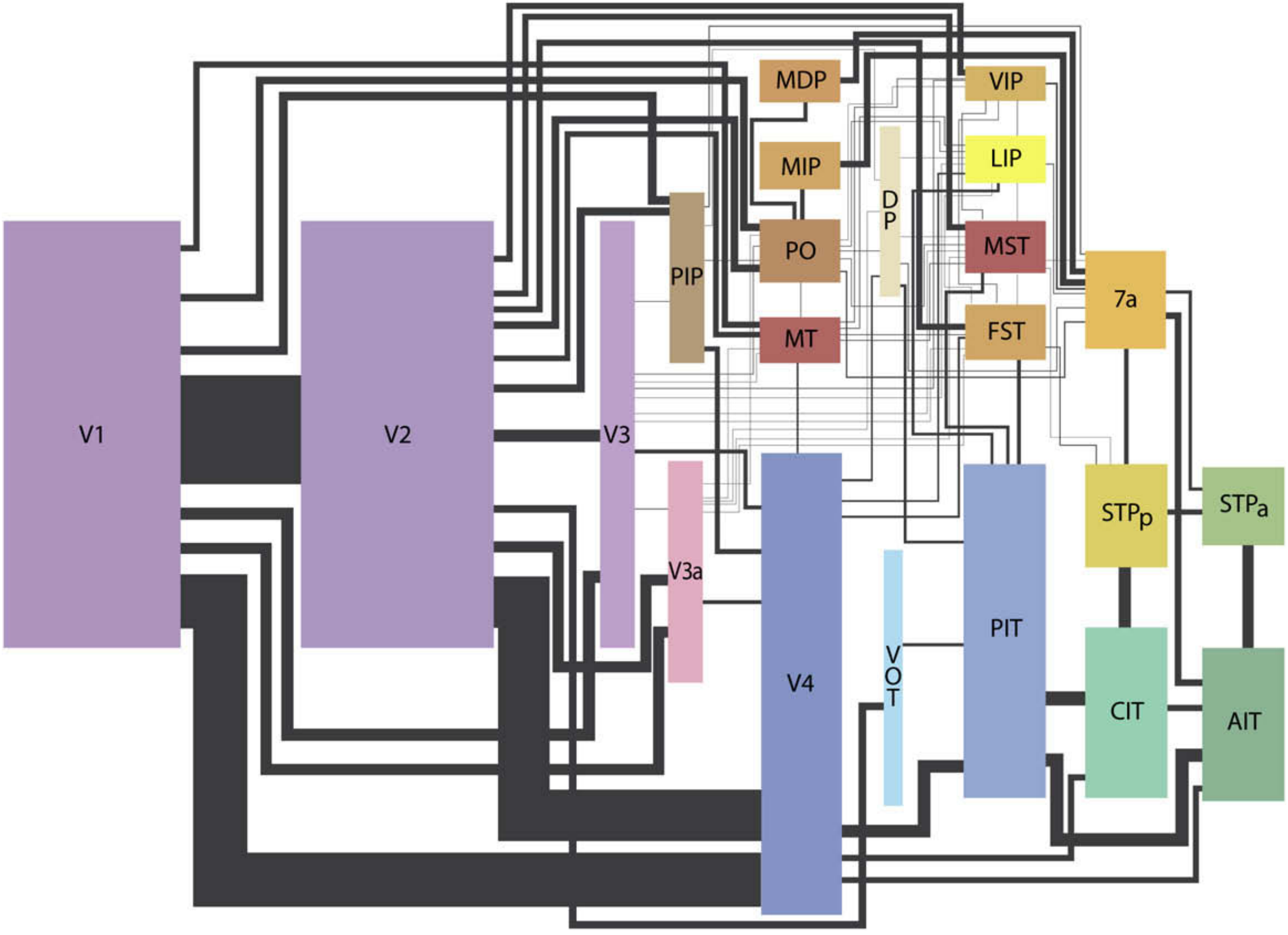
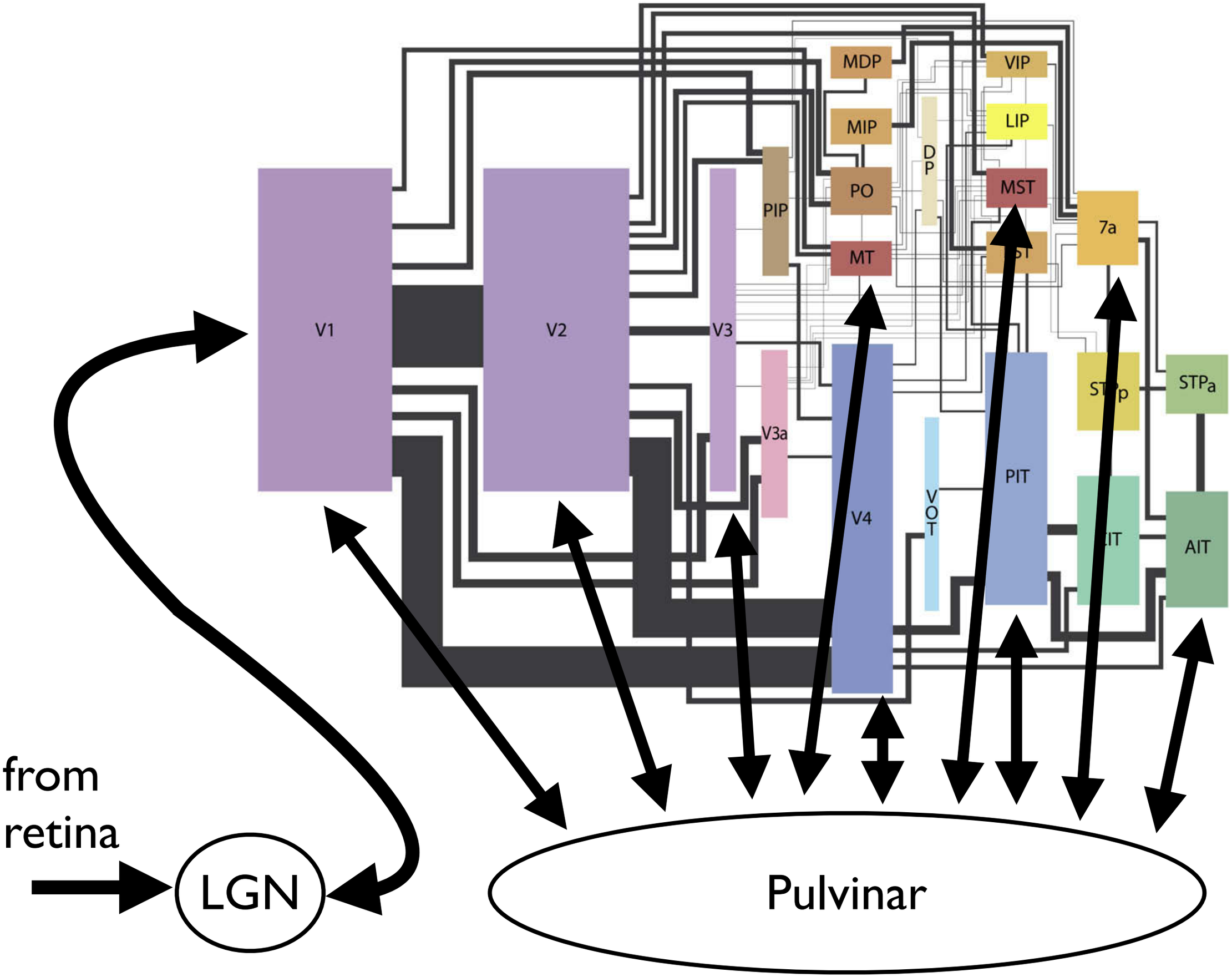
Hierarchical models

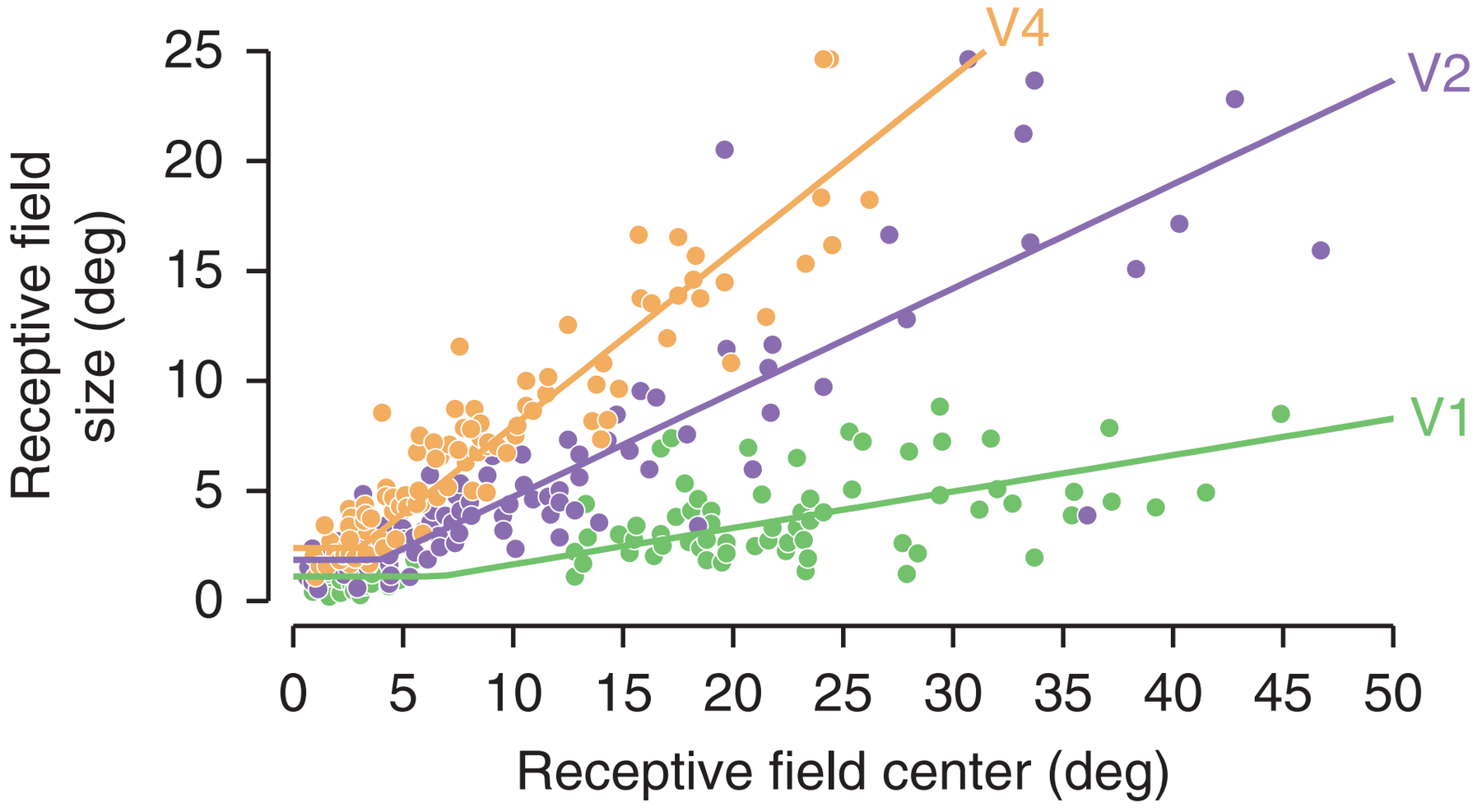


place cells
grid cells
.
.
face cells
.
invariant repr.
complex motion
.
.
?
.
.
.

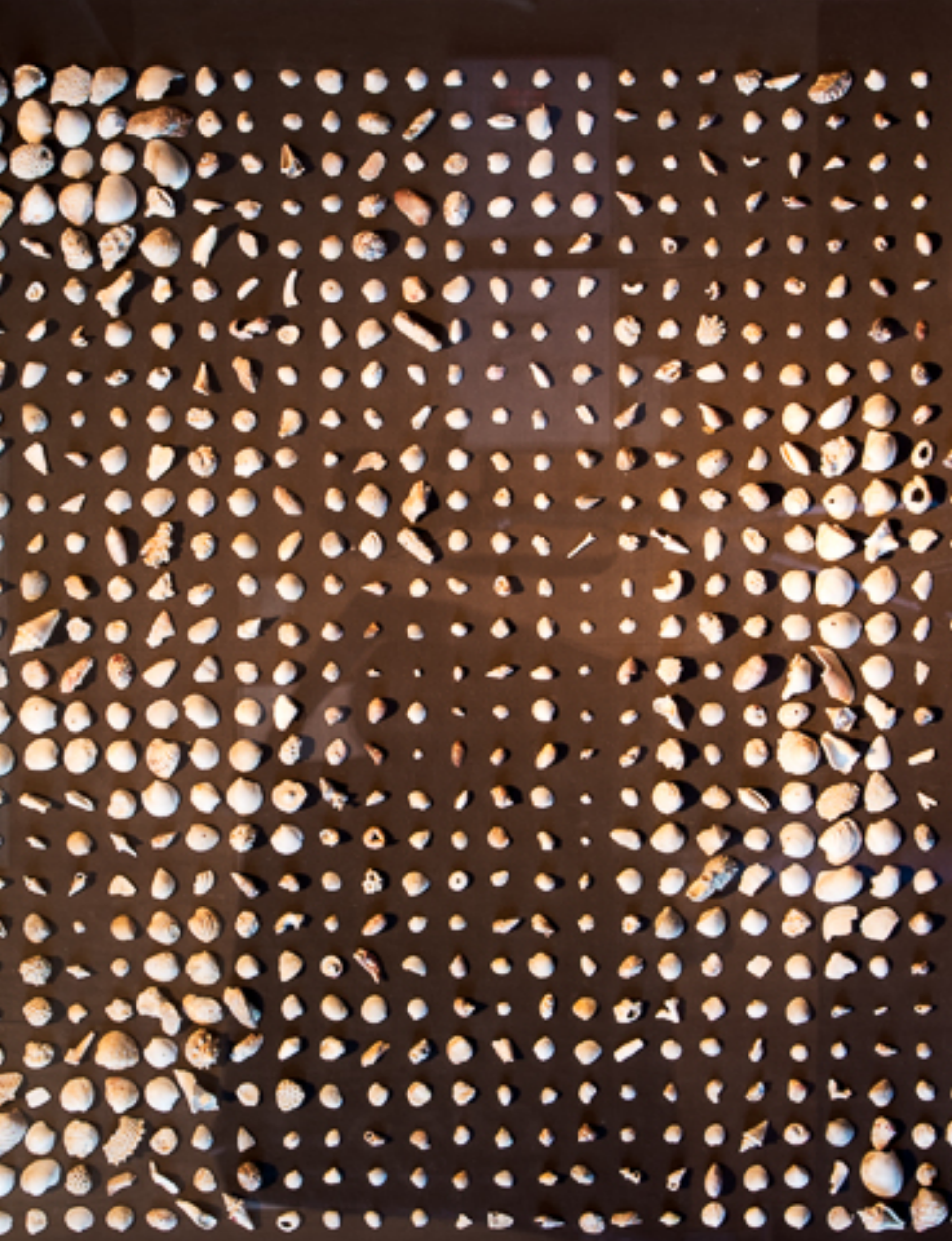
‘Gabor filters’


V1
V2
V3
V3a
PIP
MDP
MIP
PO
MT
DP
VIP
LIP
MST
FST
7a
V4
VOT
PIT
STPp
STPa
CIT
AIT


V1
V2
V3
V3a
PIP
MDP
MIP
PO
MT
DP
VIP
LIP
MST
7a
V4
VOT
PIT
AIT
STPa
from
retina
LGN
Pulvinar



From Freeman & Simoncelli (2011)
(data from Gattass et al. 1981; 1988)



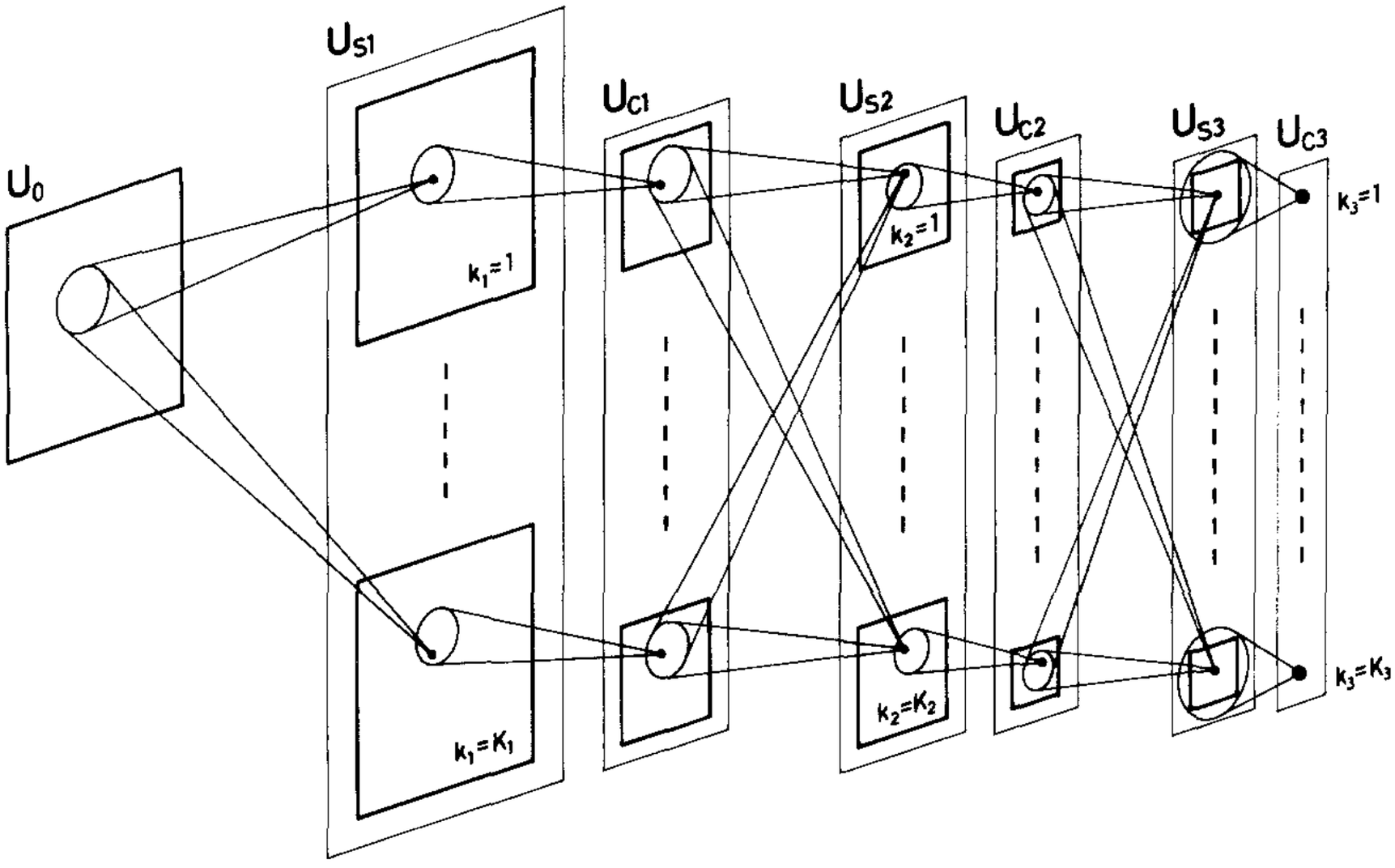
Pareidolia

Biol. Cybernetics 36, 193–202 (1980)

Neocognitron: A Self-organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position

Kunihiko Fukushima

NHK Broadcasting Science Research Laboratories, Kinuta, Setagaya, Tokyo, Japan



Neocognitron: rationale

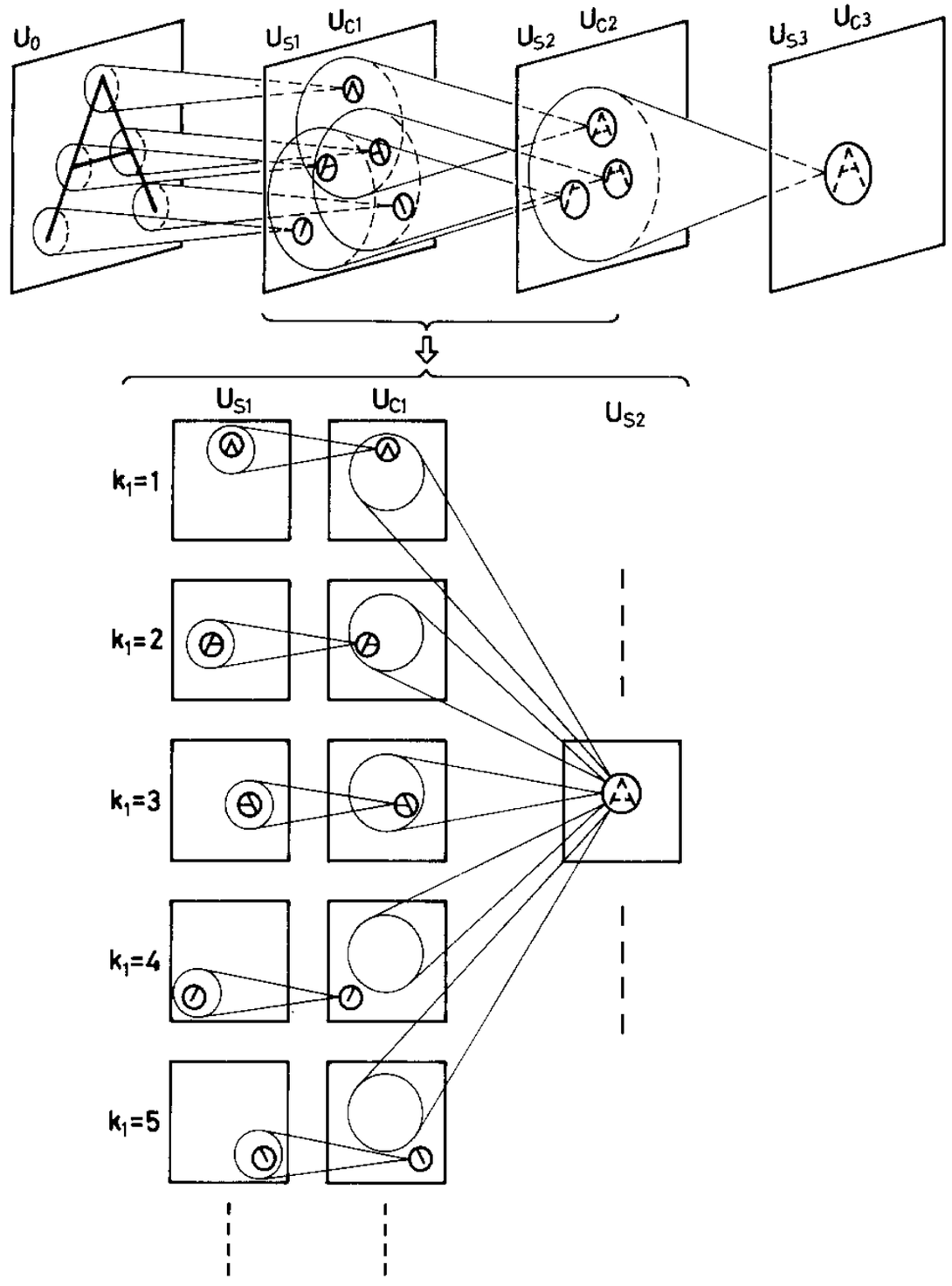


Fig. 5. An example of the interconnections between cells and the response of the cells after completion of self-organization

Neocognitron: activation rule

convolution

$$u_{Sl}(k_l, \mathbf{n}) = r_l \cdot \varphi\left[\frac{1 + \sum_{k_{l-1}=1}^{K_{l-1}} \sum_{\mathbf{v} \in S_l} a_l(k_{l-1}, \mathbf{v}, k_l) \cdot u_{Cl-1}(k_{l-1}, \mathbf{n}+\mathbf{v})}{1 + \frac{2r_l}{1+r_l} \cdot b_l(k_l) \cdot v_{Cl-1}(\mathbf{n})} - 1\right]$$

divisive normalization

where

$$\varphi[x] = \begin{cases} x & x \geqq 0 \\ 0 & x < 0. \end{cases}$$

Relu

$$v_{Cl-1}(\mathbf{n}) = \sqrt{\sum_{k_{l-1}=1}^{K_{l-1}} \sum_{\mathbf{v} \in S_l} c_{l-1}(\mathbf{v}) \cdot u_{Cl-1}^2(k_{l-1}, \mathbf{n}+\mathbf{v})},$$

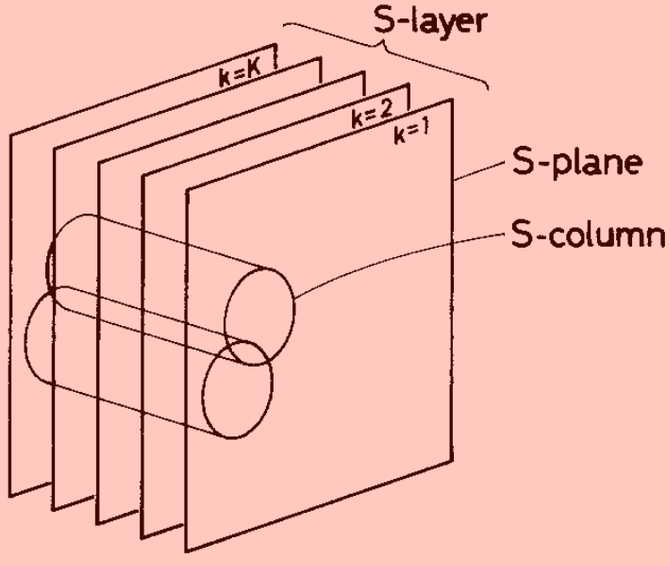
Neocognitron: learning rule

Let cell $u_{Sl}(\hat{k}_l, \hat{\mathbf{n}})$ be selected as a representative.

$$\Delta a_l(k_{l-1}, \mathbf{v}, \hat{k}_l) = q_l \cdot c_{l-1}(\mathbf{v}) \cdot u_{Cl-1}(k_{l-1}, \hat{\mathbf{n}} + \mathbf{v}),$$

← **Hebbian learning**

From each S-column, every time when a stimulus pattern is presented, the S-cell which is yielding the largest output is chosen as a candidate for the representatives. Hence, there is a possibility that a number of candidates appear in a single S-plane. If two or more candidates appear in a single S-plane, only the one which is yielding the largest output among them is selected as the representative from that S-plane. In



Local WTA

Neocognitron: performance

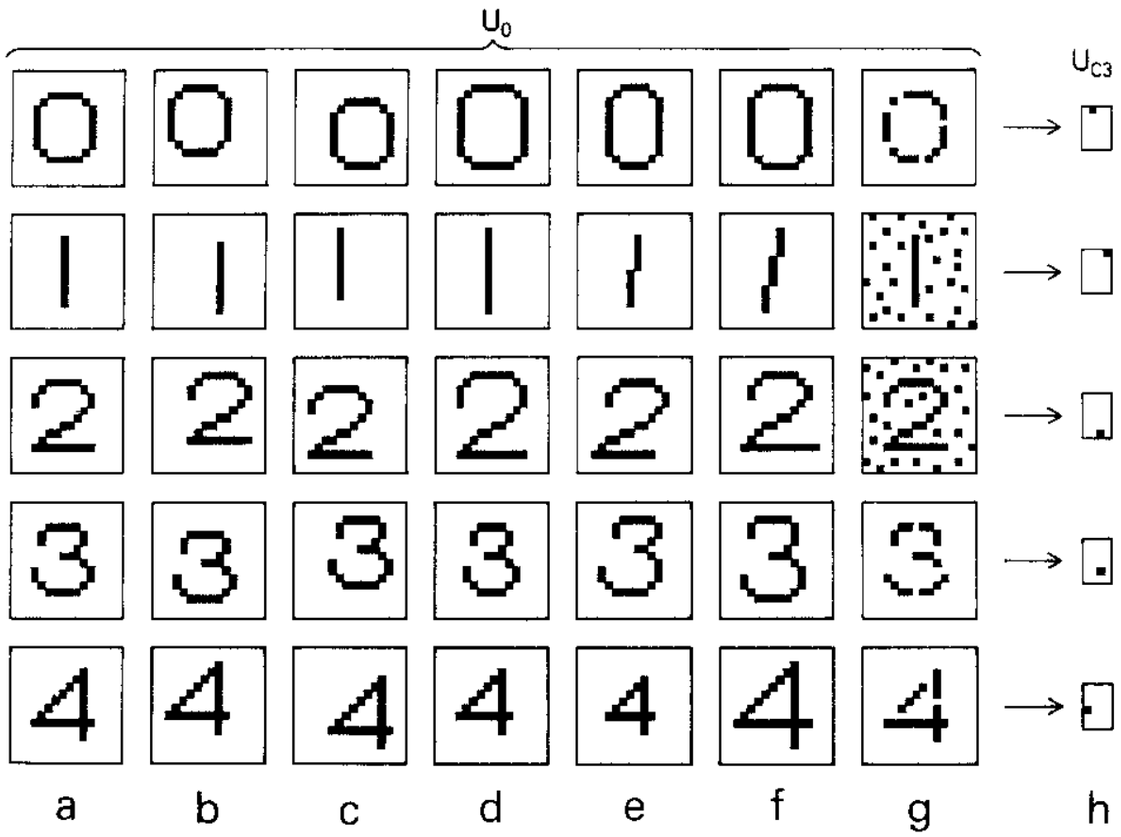


Fig. 6. Some examples of distorted stimulus patterns which the neocognitron has correctly recognized, and the response of the final layer of the network

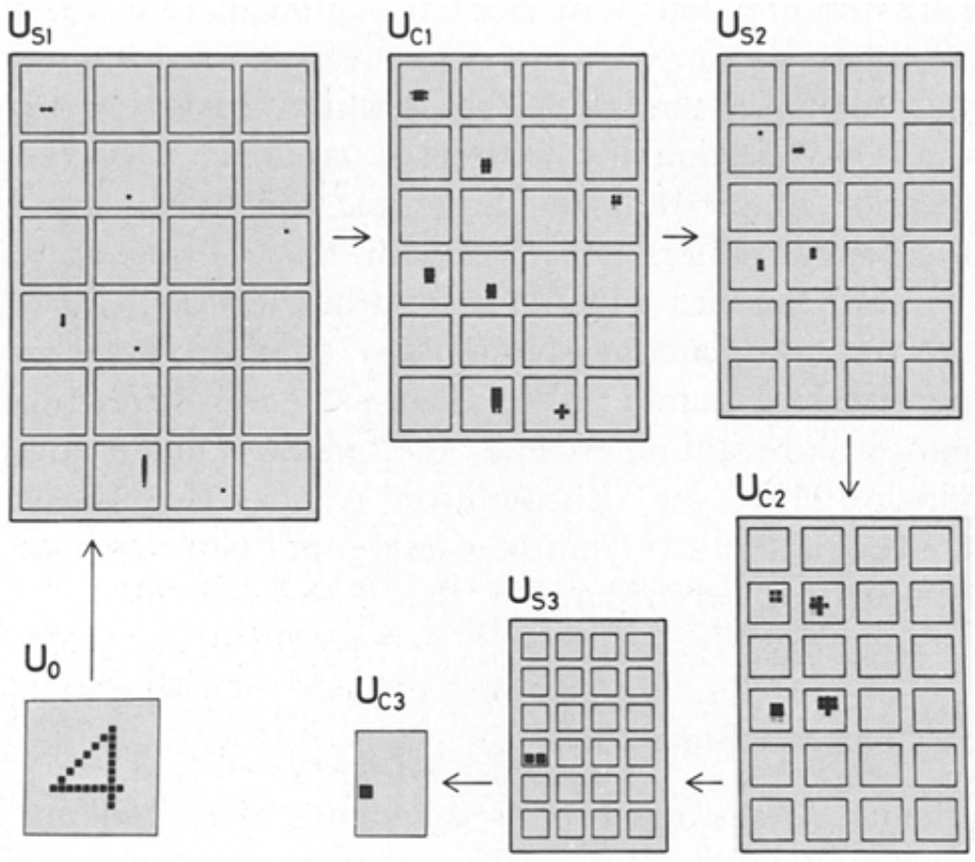


Fig. 7. A display of an example of the response of all the individual cells in the neocognitron

‘AlexNet’
(Krizhevsky, Sutskever & Hinton 2012)

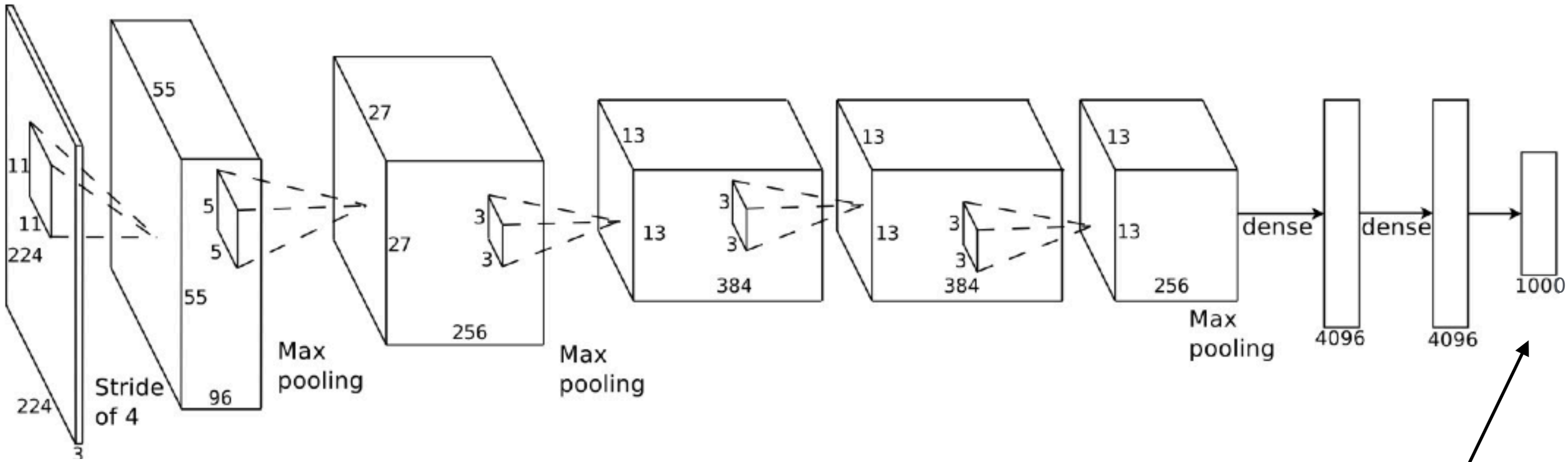
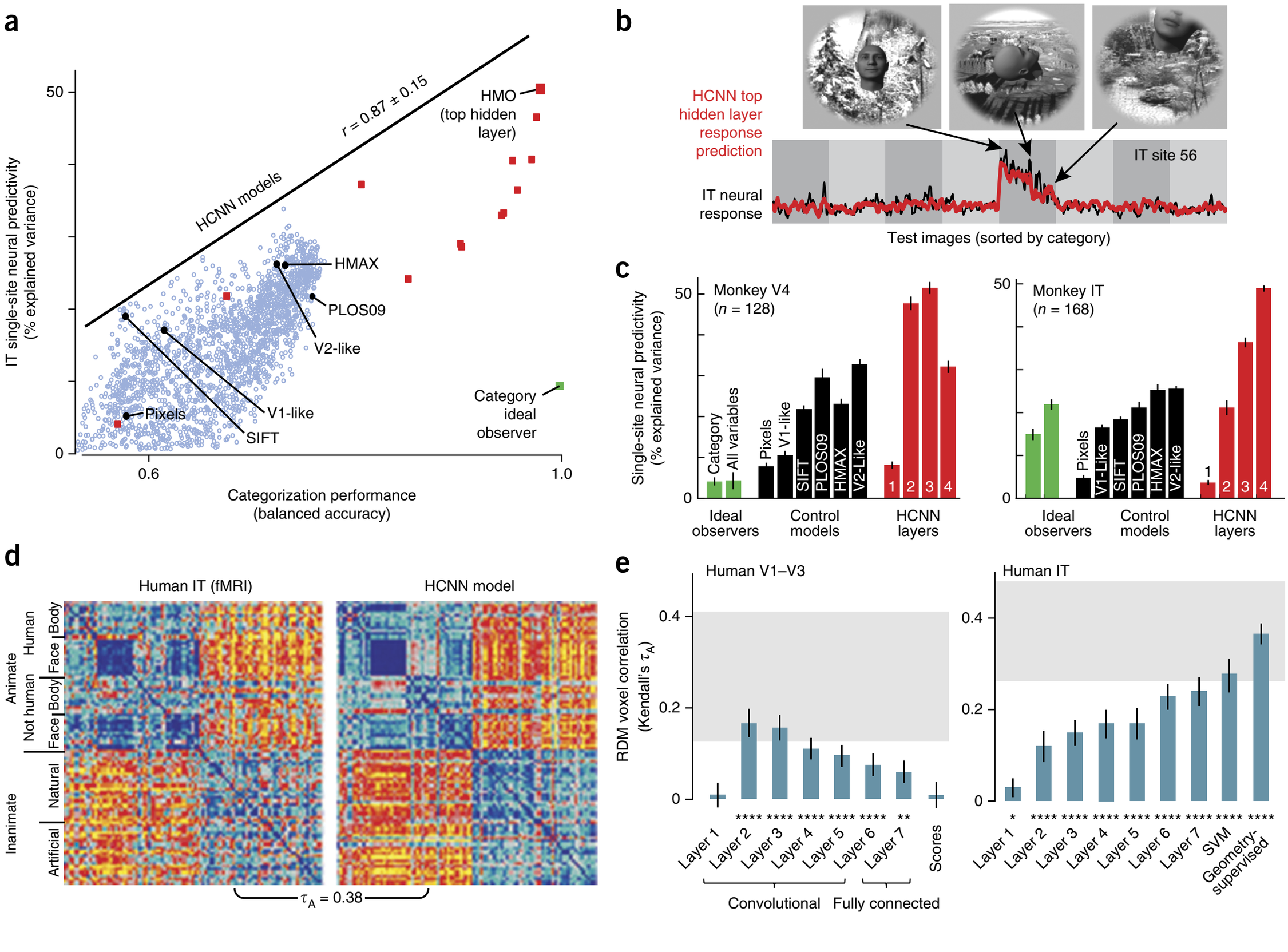
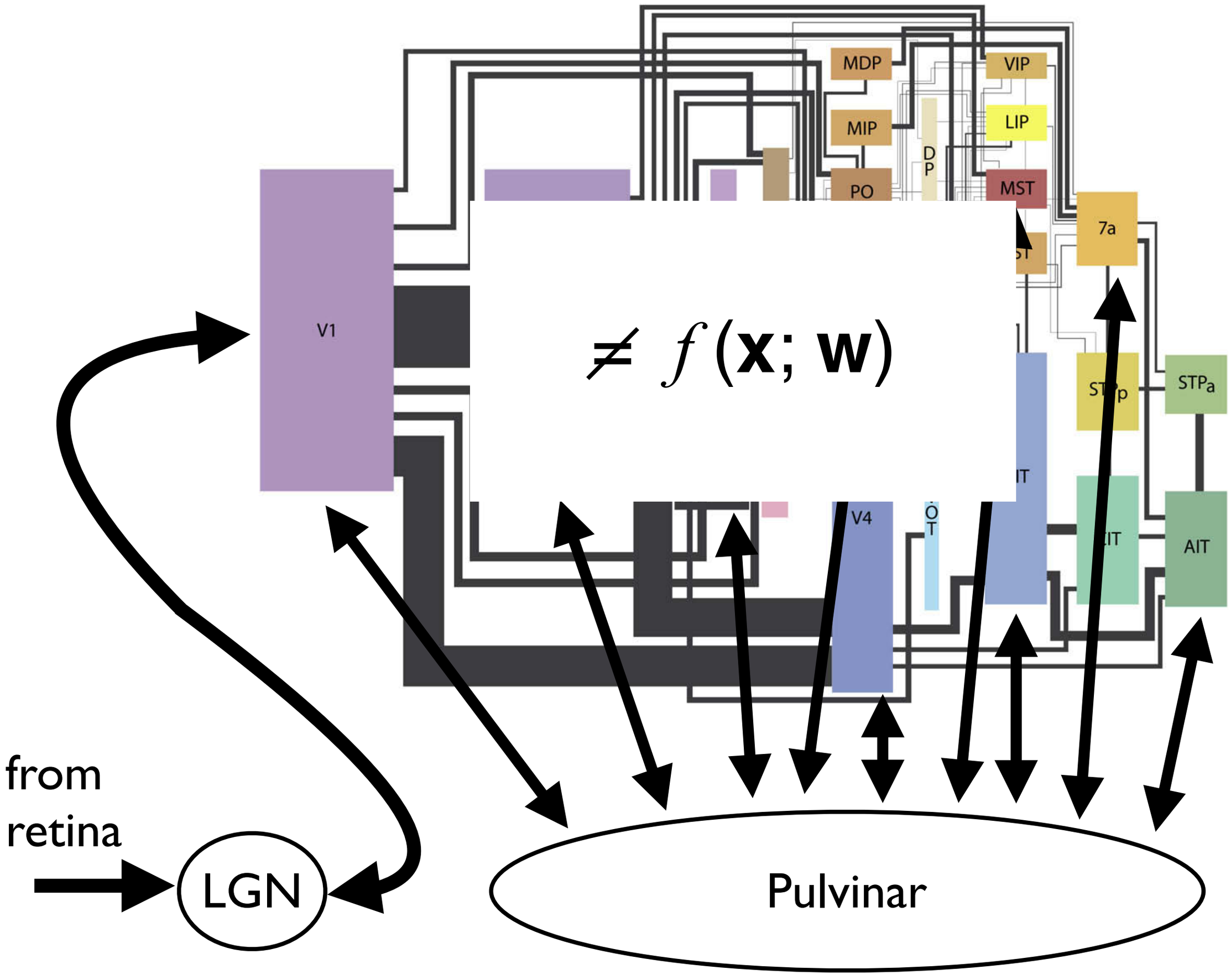


image feature extraction and pooling classification

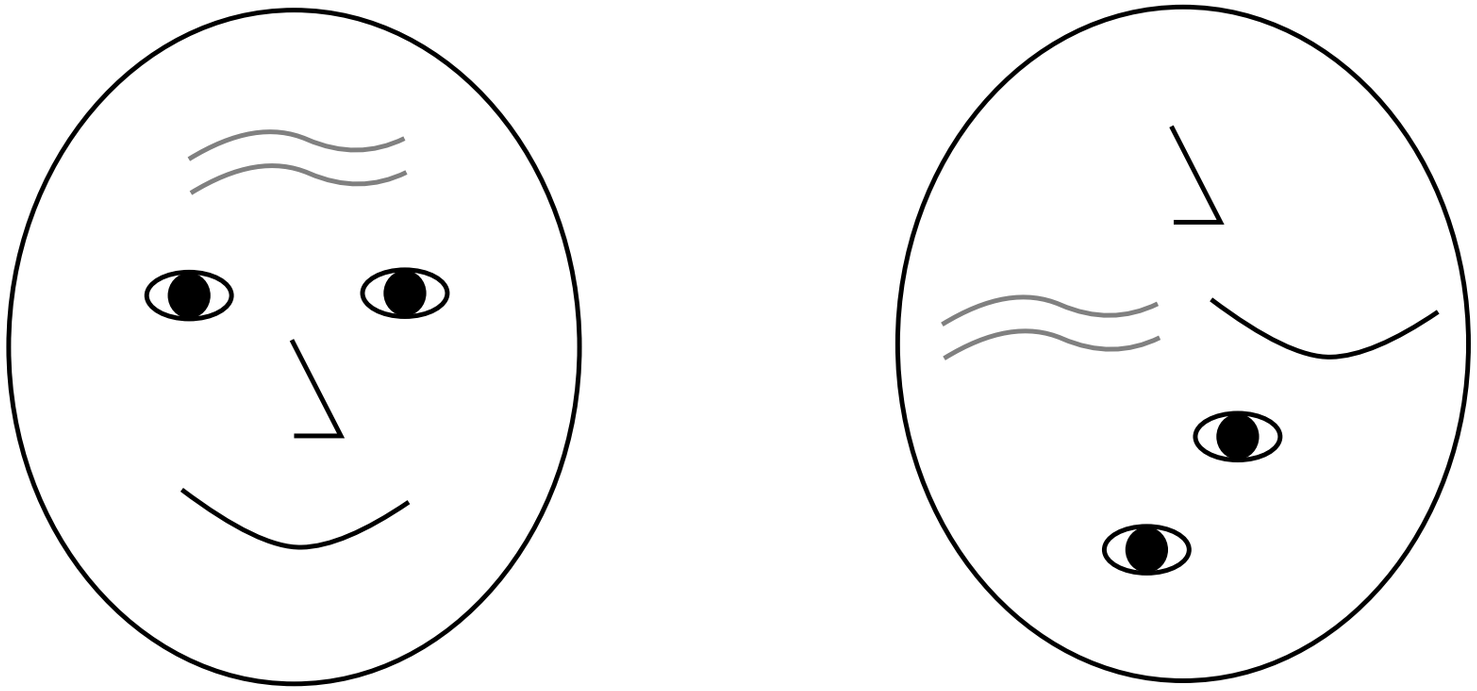
Deep networks appear to predict responses of V4 and IT neurons (Yamins & DiCarlo 2016)



This isn’t a good model of perception

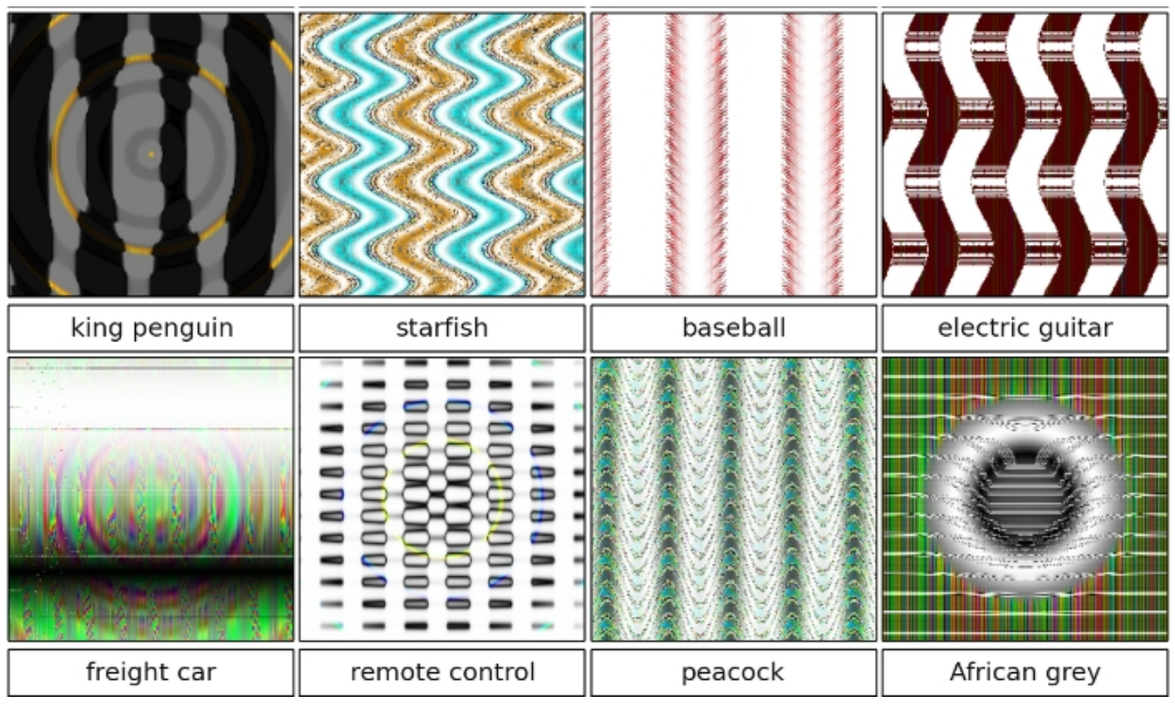

MDP
VIP
MIP
LIP
PO
MST
V1
7a
V4
AIT
from
retina
LGN
Pulvinar

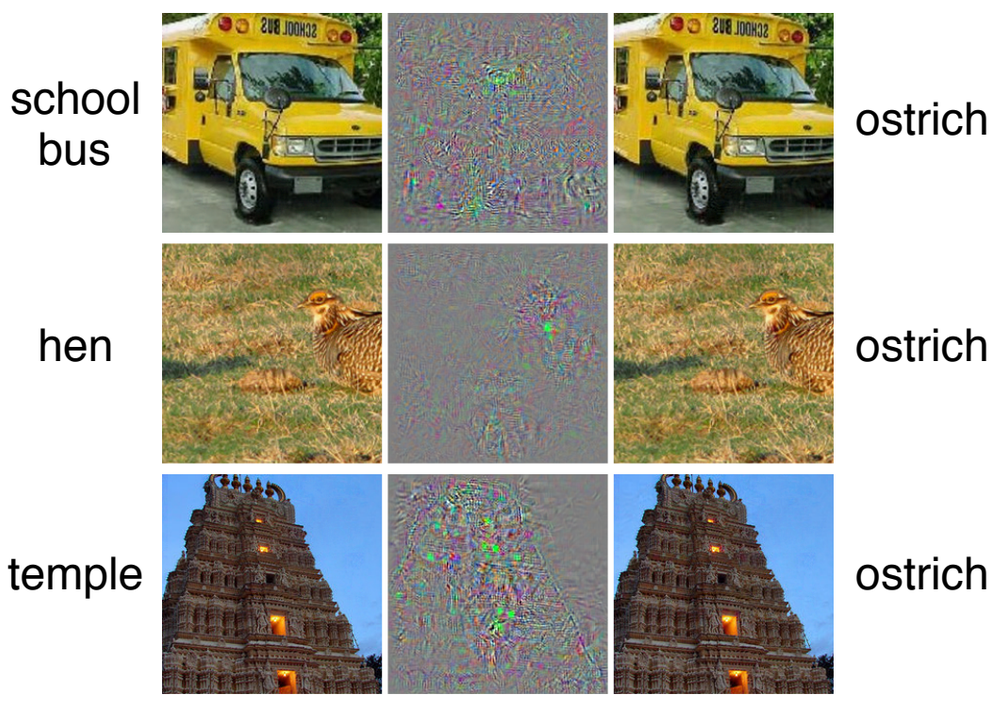
Relative spatial relationships are important



The invariant representations produced by deep convnets are…

easily fooled and brittle





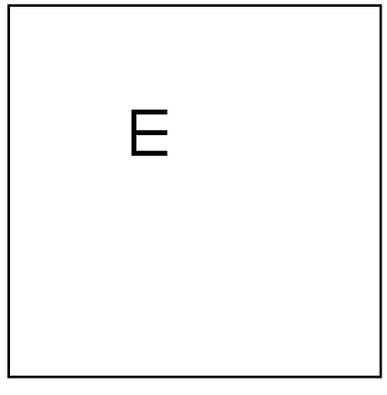
Images are not bags of features
(BagNet - Brendel & Bethge 2019)

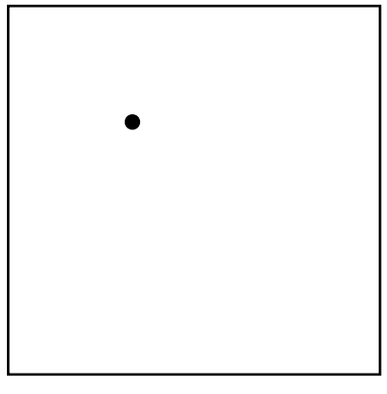
original

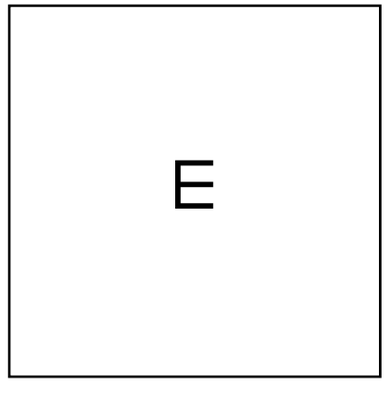
texturised images

Simple example: translation via Fourier phase-shifting

$I(\vec{x})$ $\quad = \quad \delta(\vec{x} - \Delta\vec{x}) \quad * \quad S(\vec{x})$

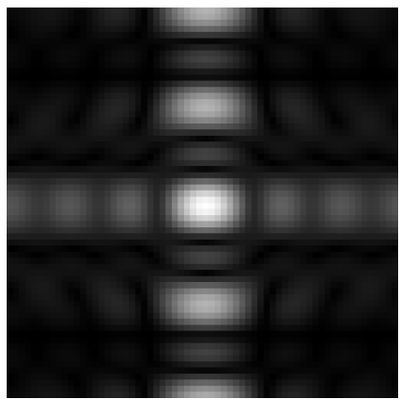


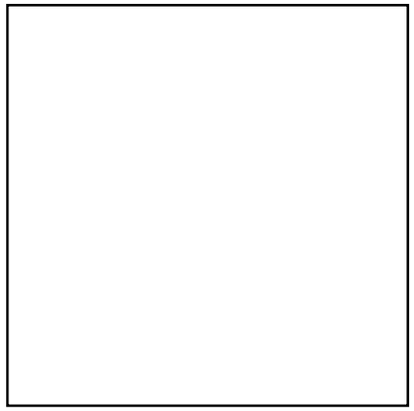


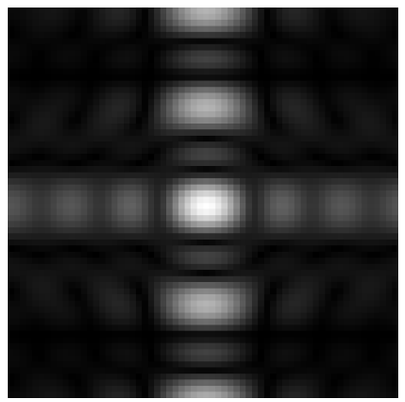


$I(\vec{\omega}) \quad = \quad e^{-j\vec{\omega}\cdot\Delta\vec{x}} \quad * \quad S(\vec{\omega})$

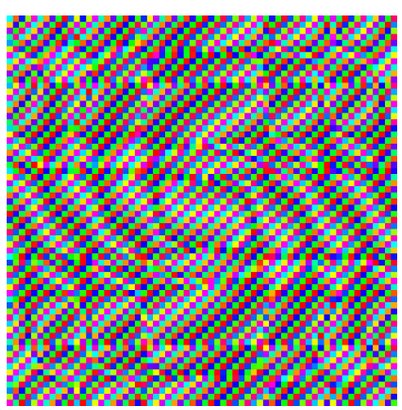
amplitude

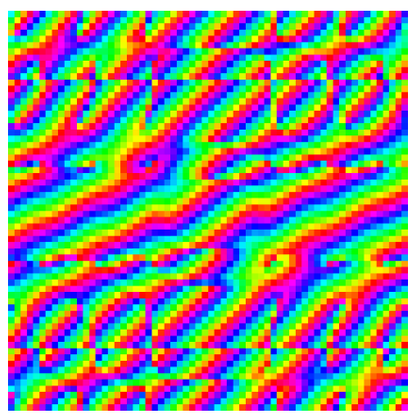






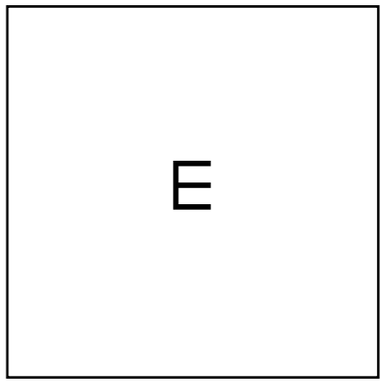
phase

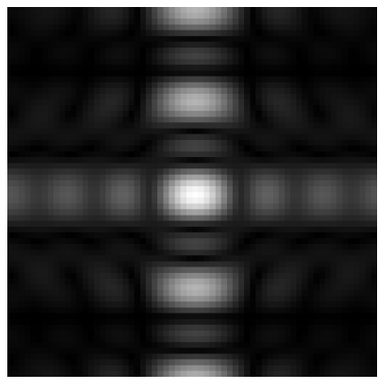


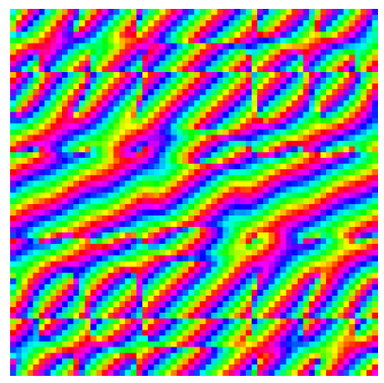


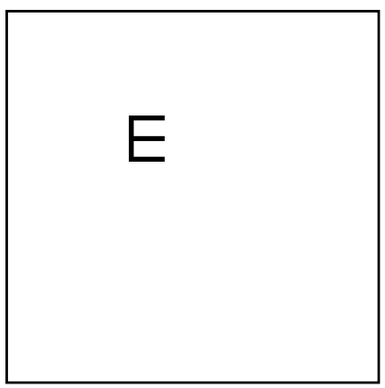
Amplitude spectrum is invariant to shift, but excessively so.

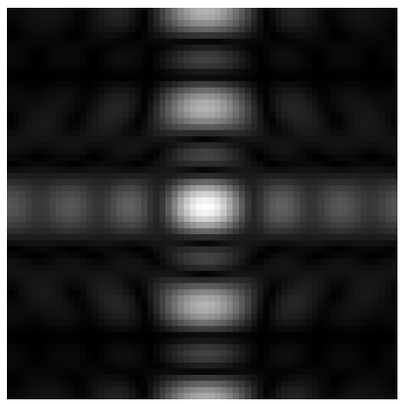
$I(\vec{x})$ | $|\tilde{I}(\vec{\omega})|$ | $\angle\tilde{I}(\vec{\omega})$

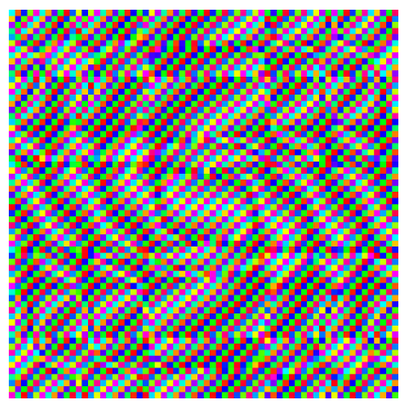


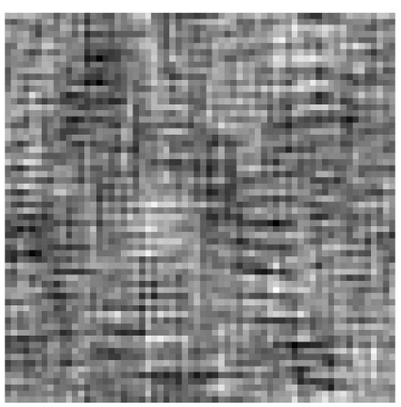


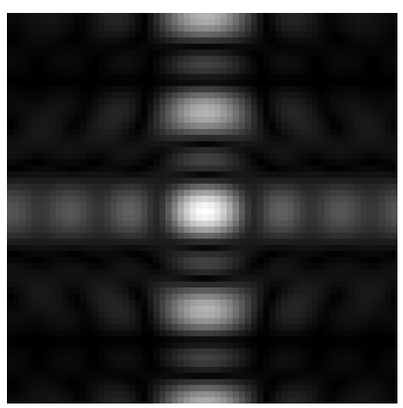


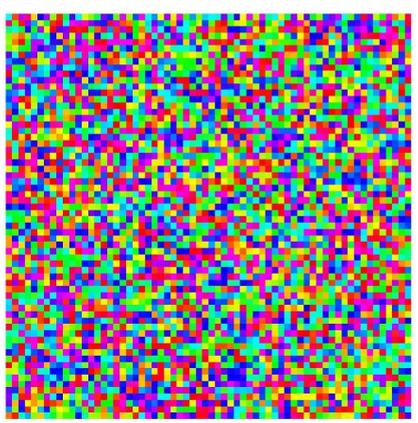








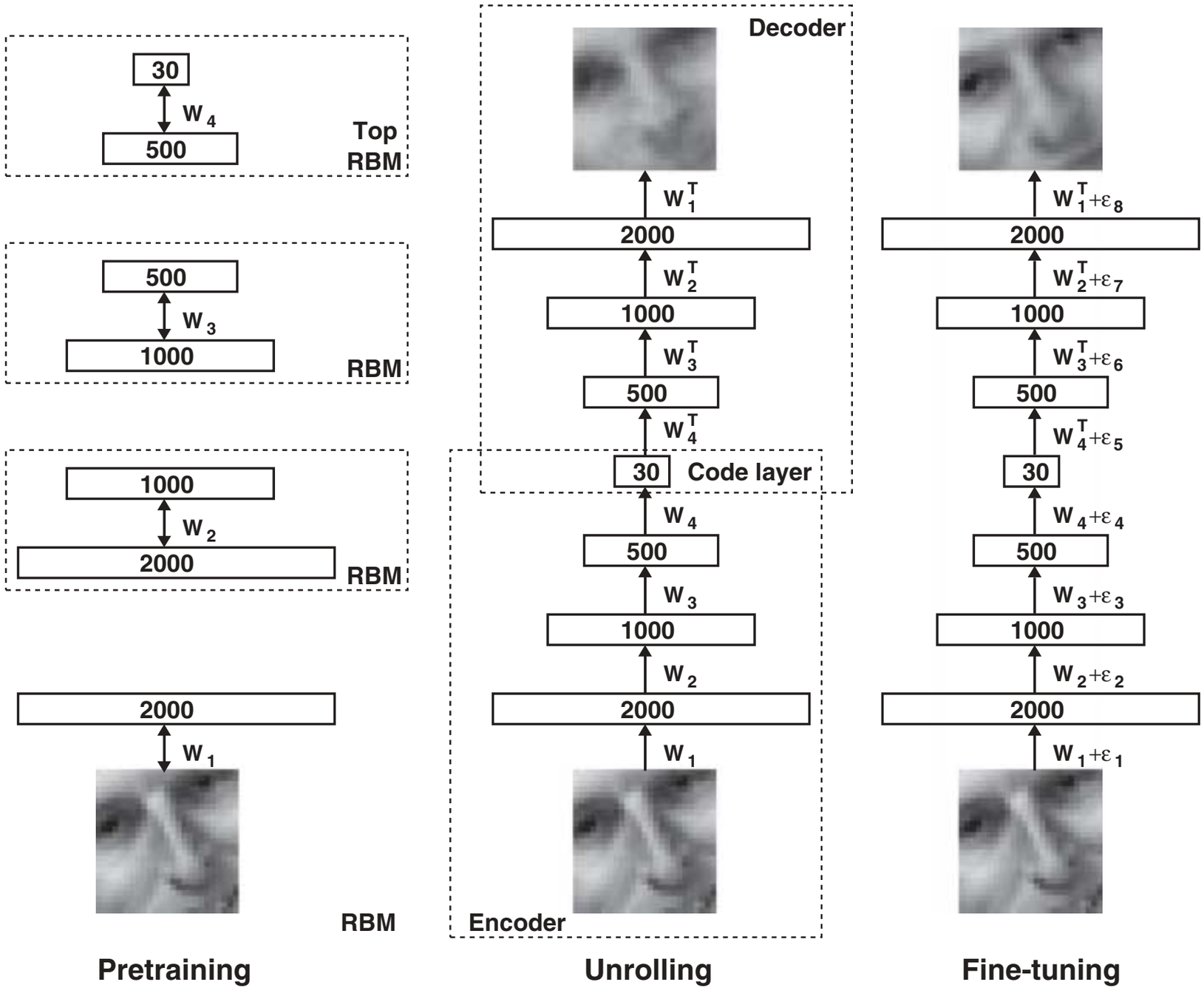




Structural information is contained in phase.
Factorization is required to extract it.

Reducing the Dimensionality of Data with Neural Networks

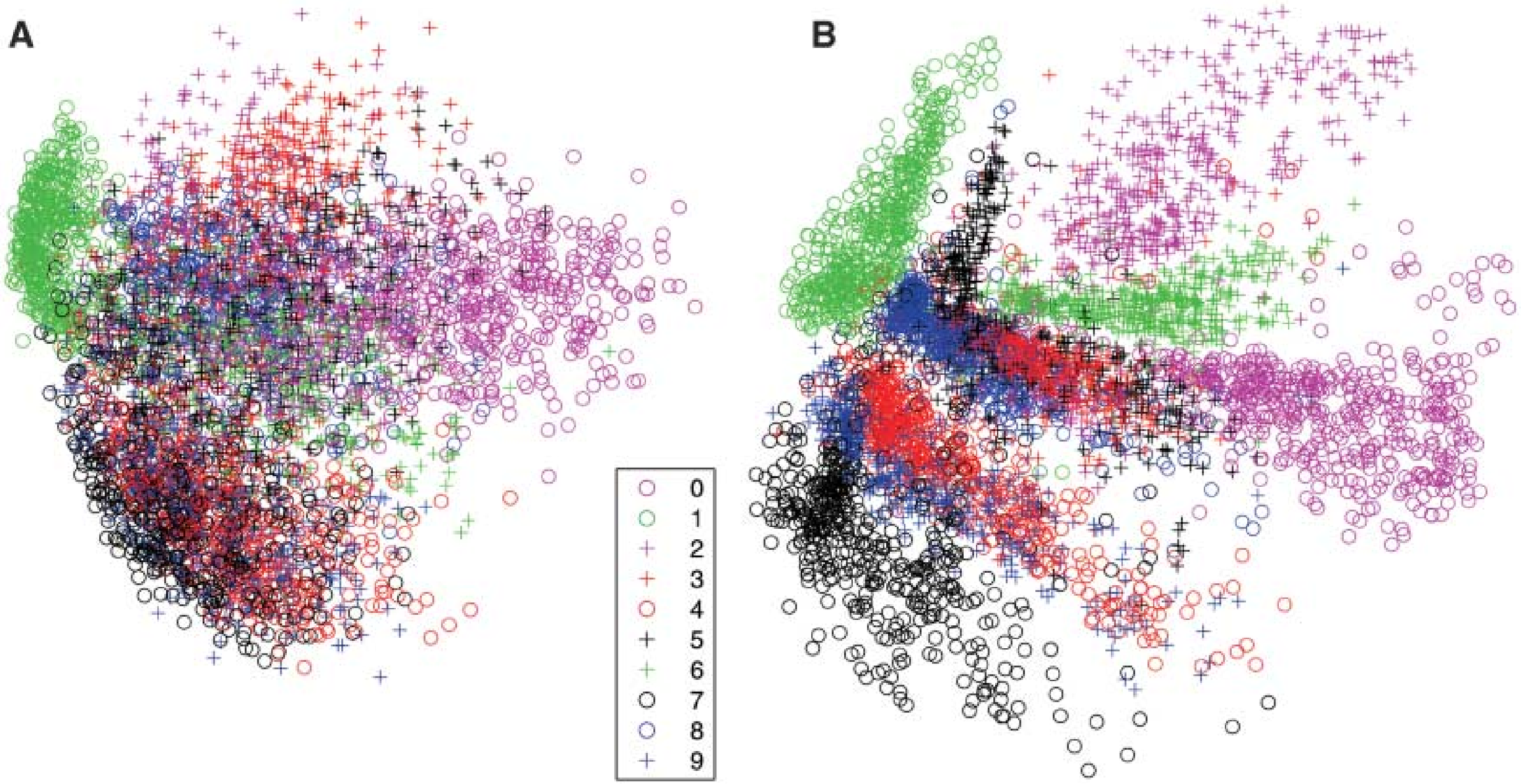
G. E. Hinton* and R. R. Salakhutdinov



$$E(\mathbf{v}, \mathbf{h}) = -\sum_{i \in \text{pixels}} b_i v_i - \sum_{j \in \text{features}} b_j h_j - \sum_{i,j} v_i h_j w_{ij}$$

$$\Delta w_{ij} = \varepsilon\left(\langle v_i h_j \rangle_{\text{data}} - \langle v_i h_j \rangle_{\text{recon}}\right)$$

Application to hand-written digits

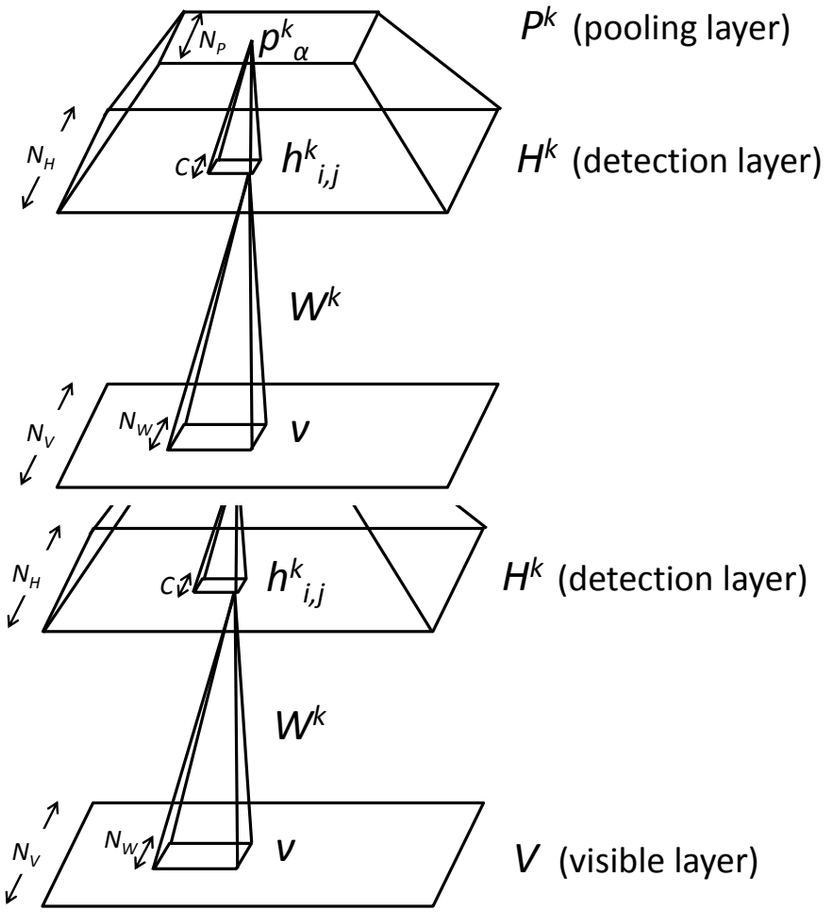


2D PCA

784-1000-500-250-2 autoencoder

Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations

Honglak Lee HLLEE@CS.STANFORD.EDU
Roger Grosse RGROSSE@CS.STANFORD.EDU
Rajesh Ranganath RAJESHR@CS.STANFORD.EDU
Andrew Y. Ng ANG@CS.STANFORD.EDU
Computer Science Department, Stanford University, Stanford, CA 94305, USA



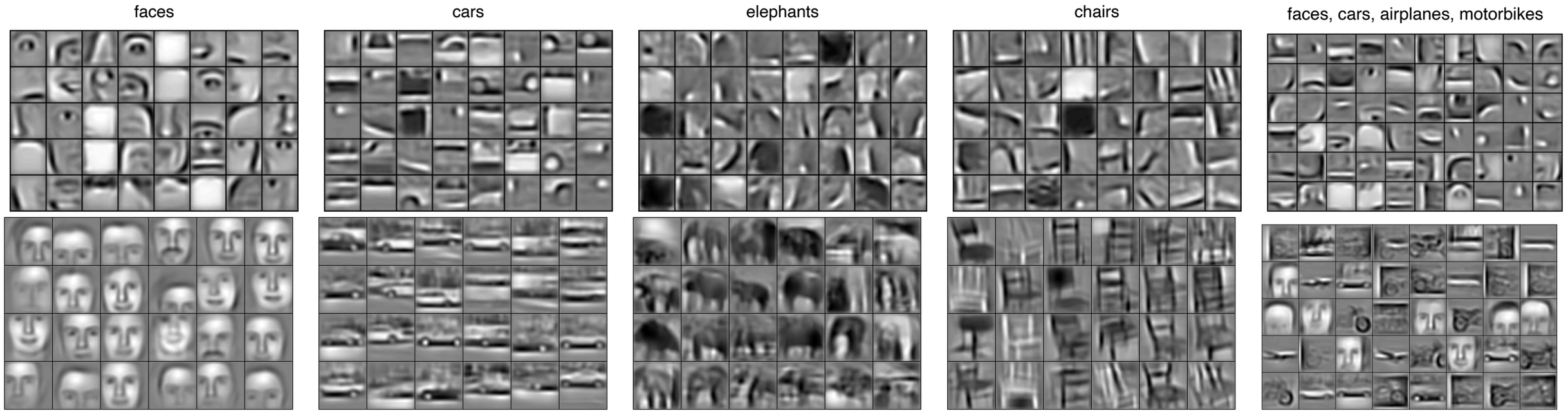
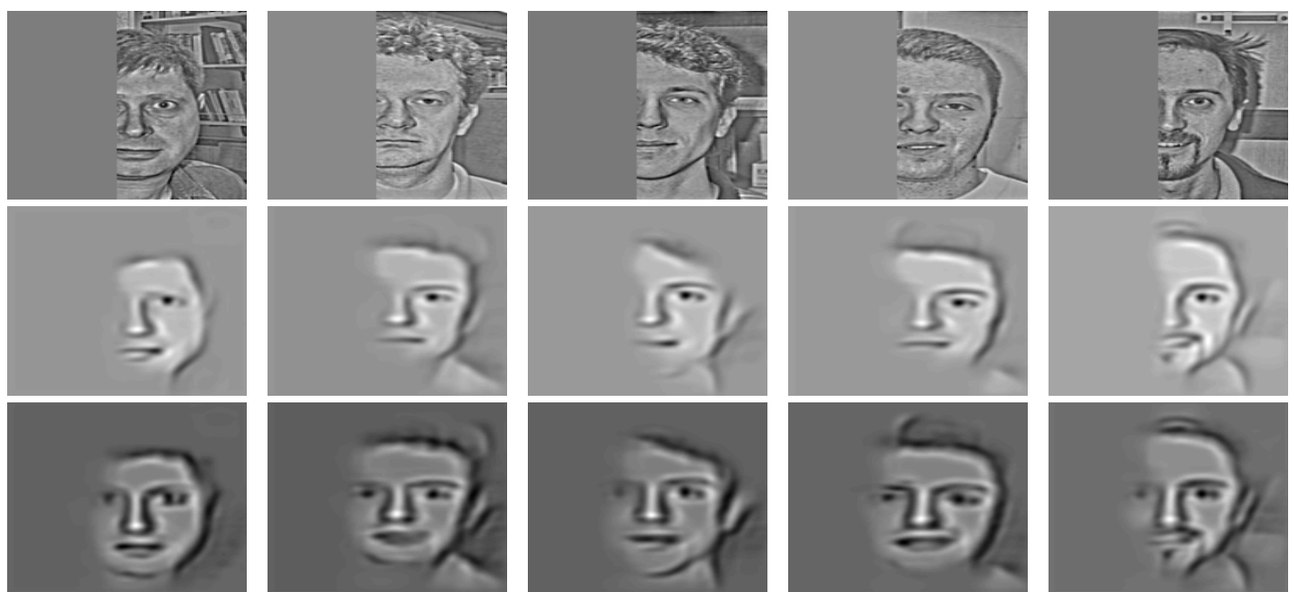
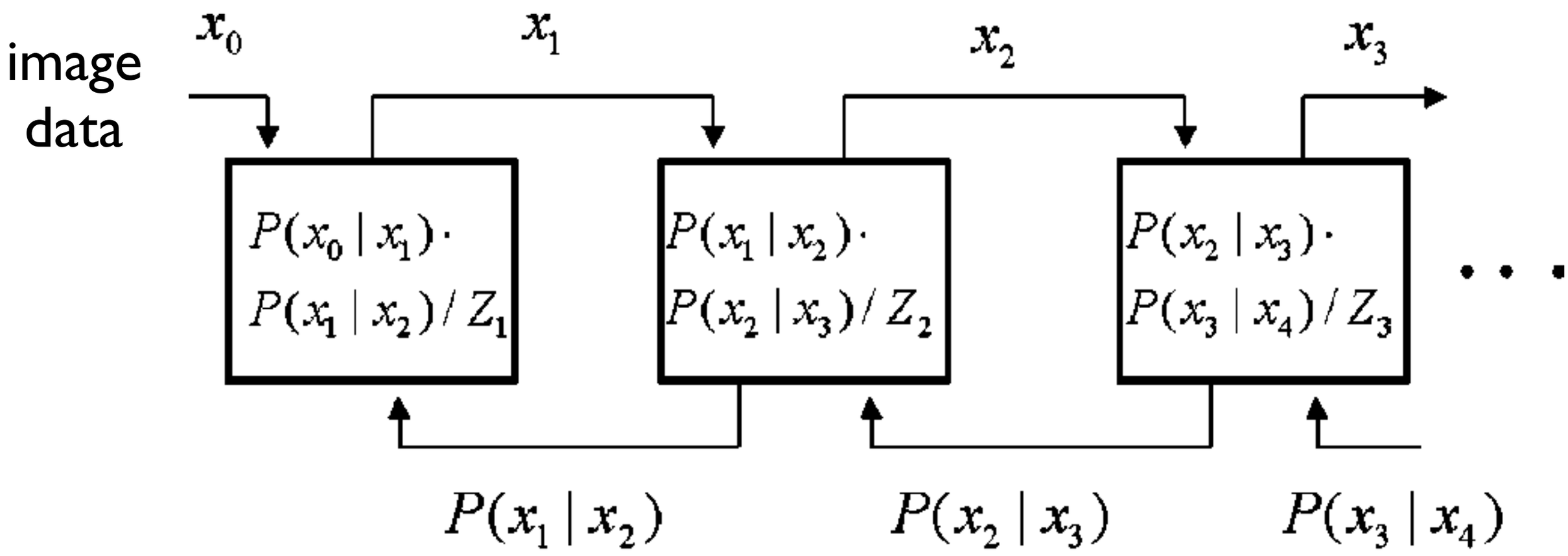


Figure 3. Columns 1-4: the second layer bases (top) and the third layer bases (bottom) learned from specific object categories. Column 5: the second layer bases (top) and the third layer bases (bottom) learned from a mixture of four object categories (faces, cars, airplanes, motorbikes).



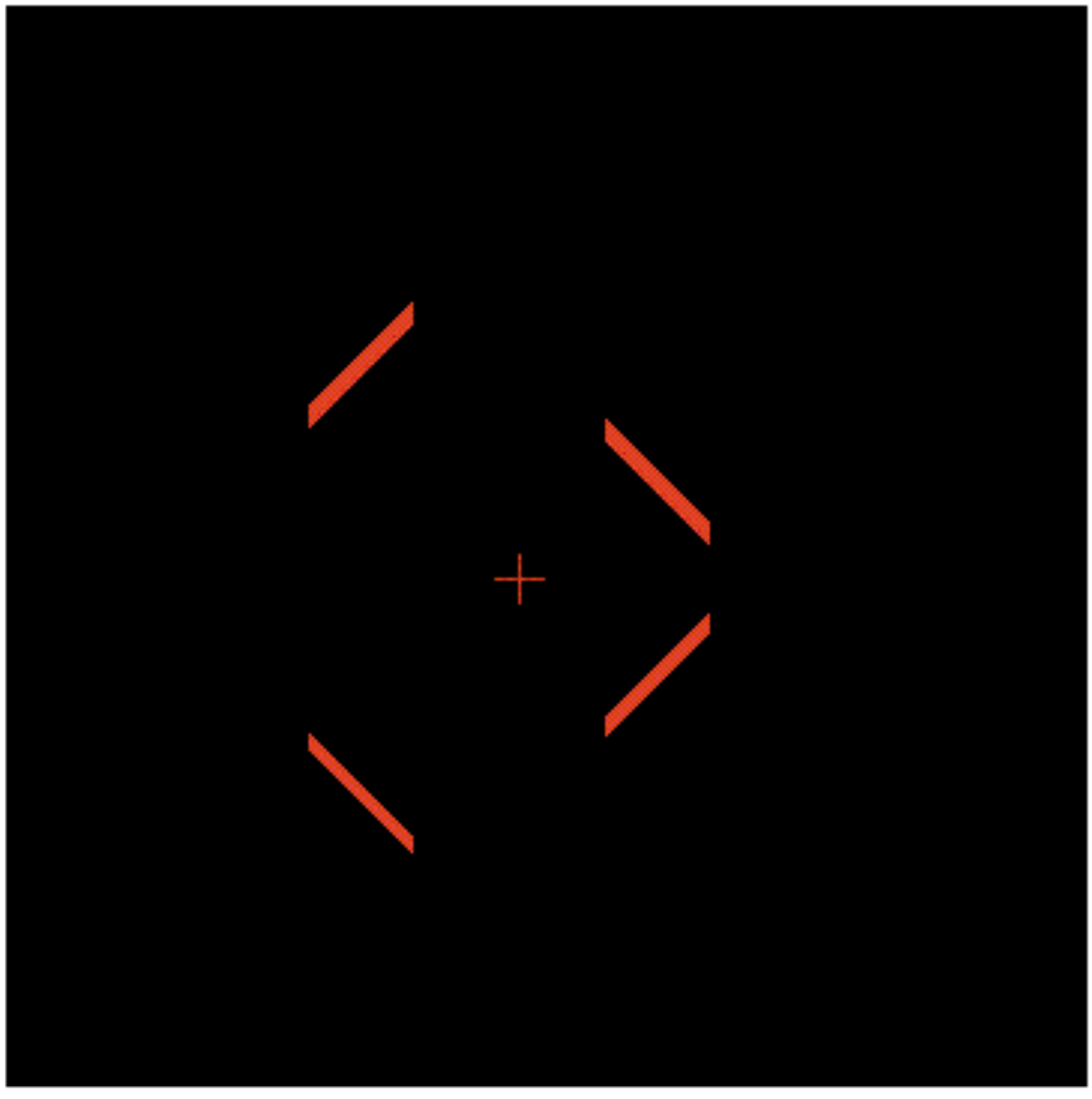
Hierarchical Bayesian inference in visual cortex

(Lee & Mumford, 2003)



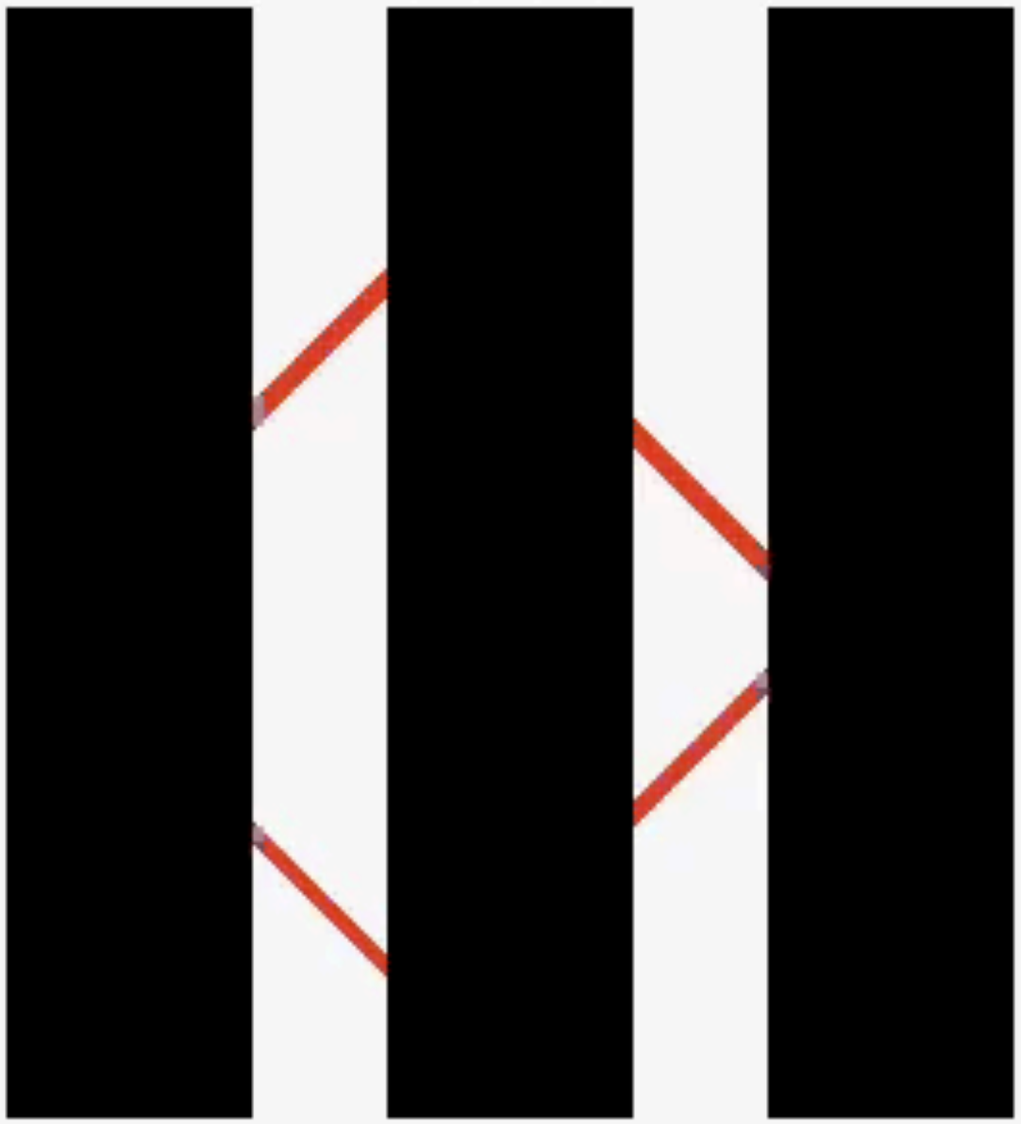
What do you see?

How do neurons in V1 encode this?

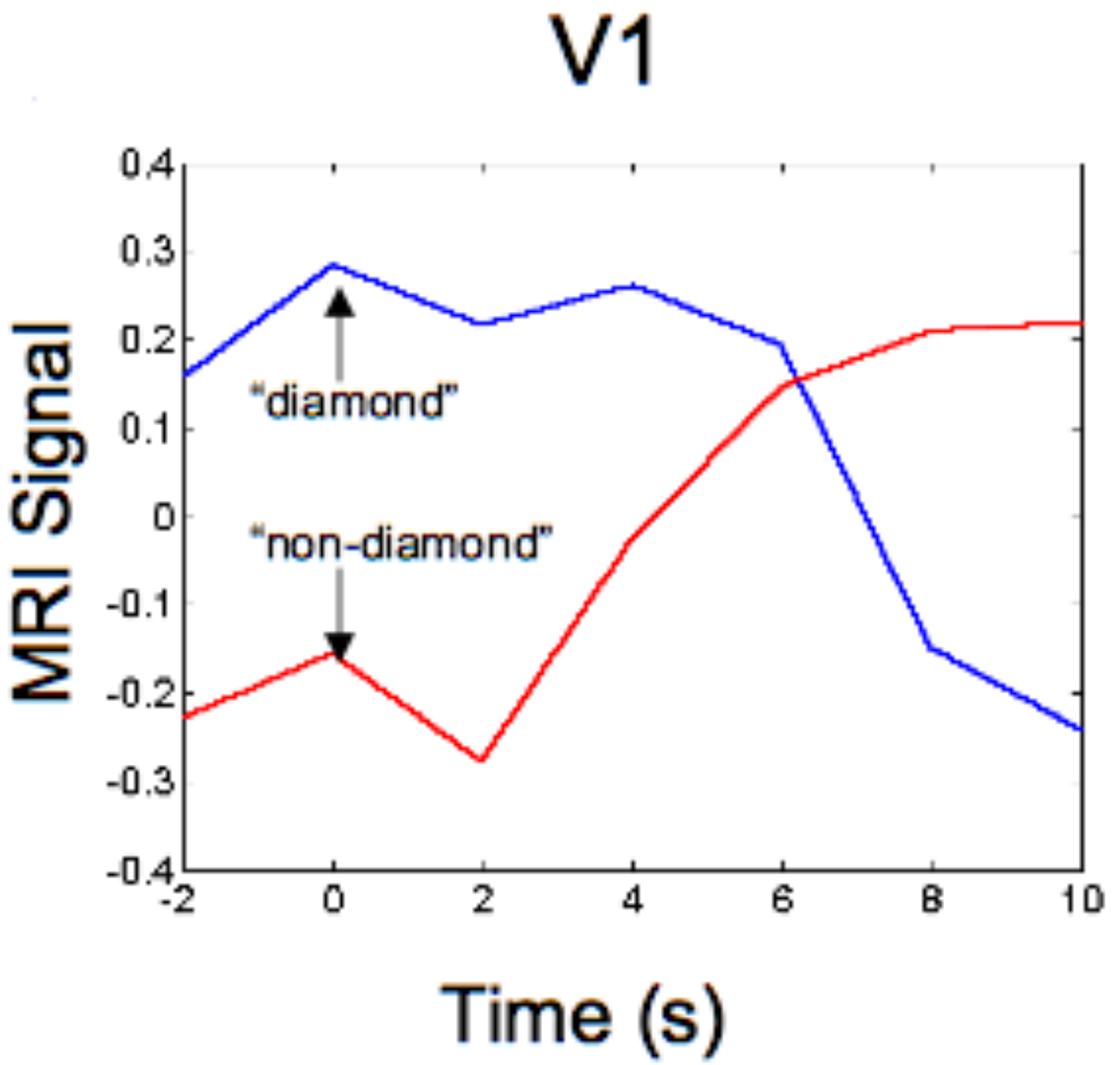


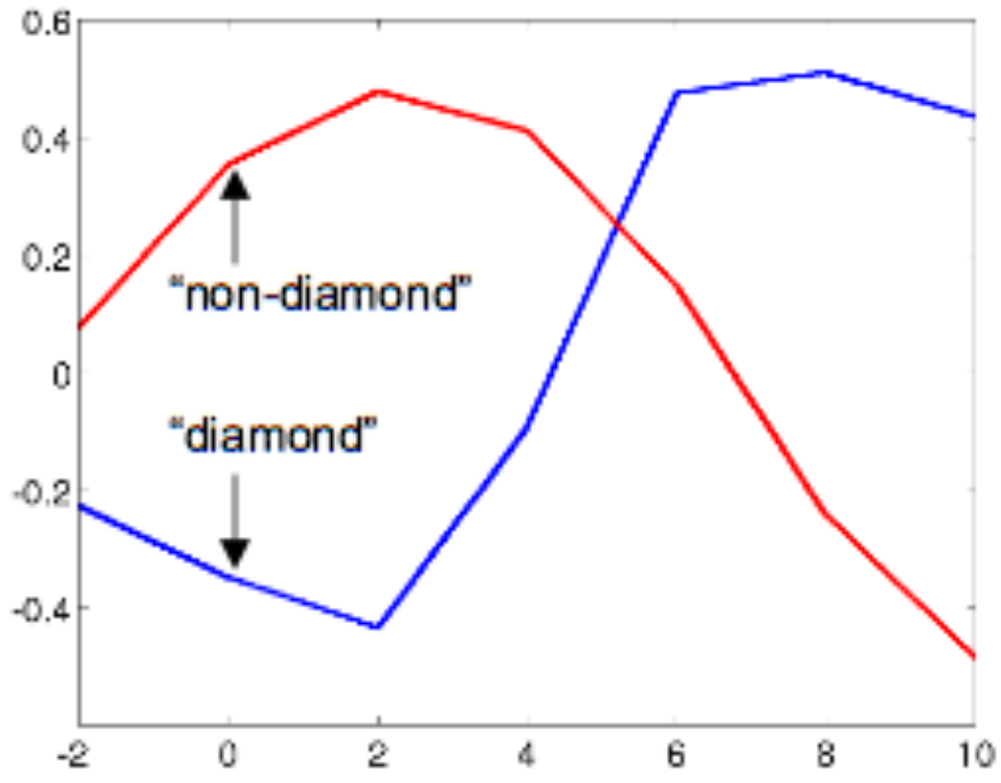
Murray, Kersten, Schrater, Olshausen, Woods, *PNAS* 2002.

(easy version)



BOLD signal in V1 and LOC

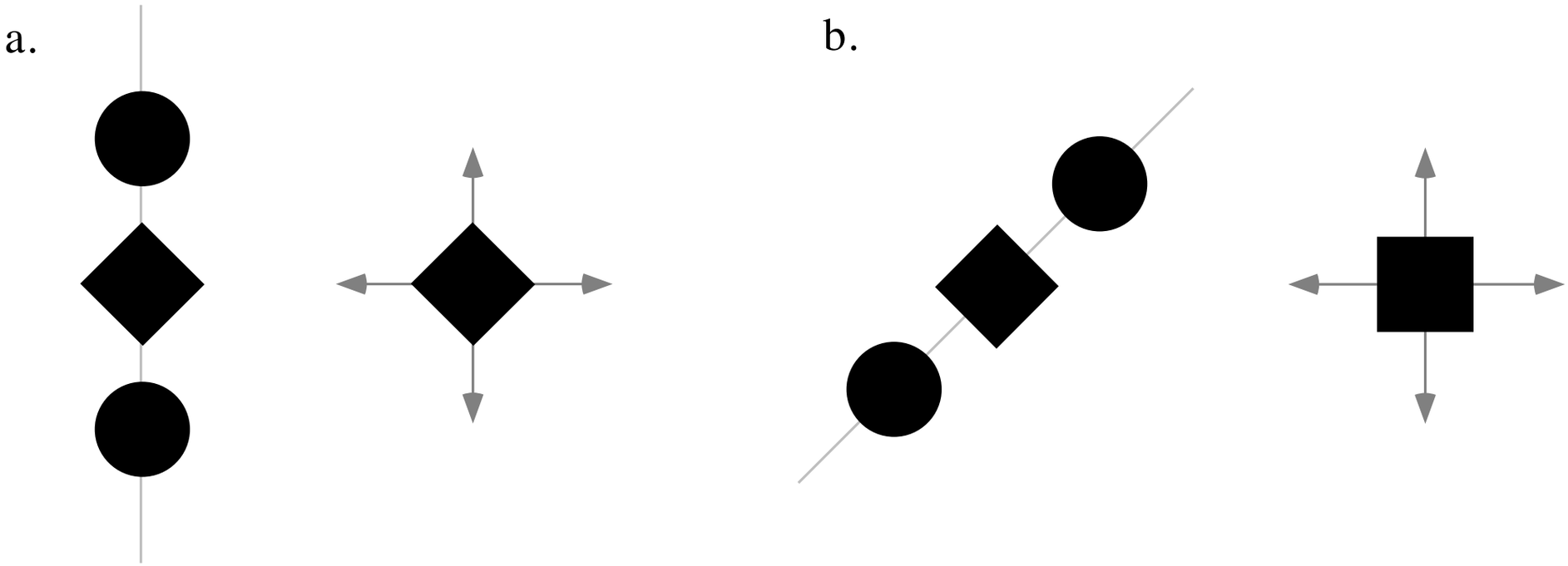




How to form invariant object representations?

Reference frame effects in perception

Diamond or square?

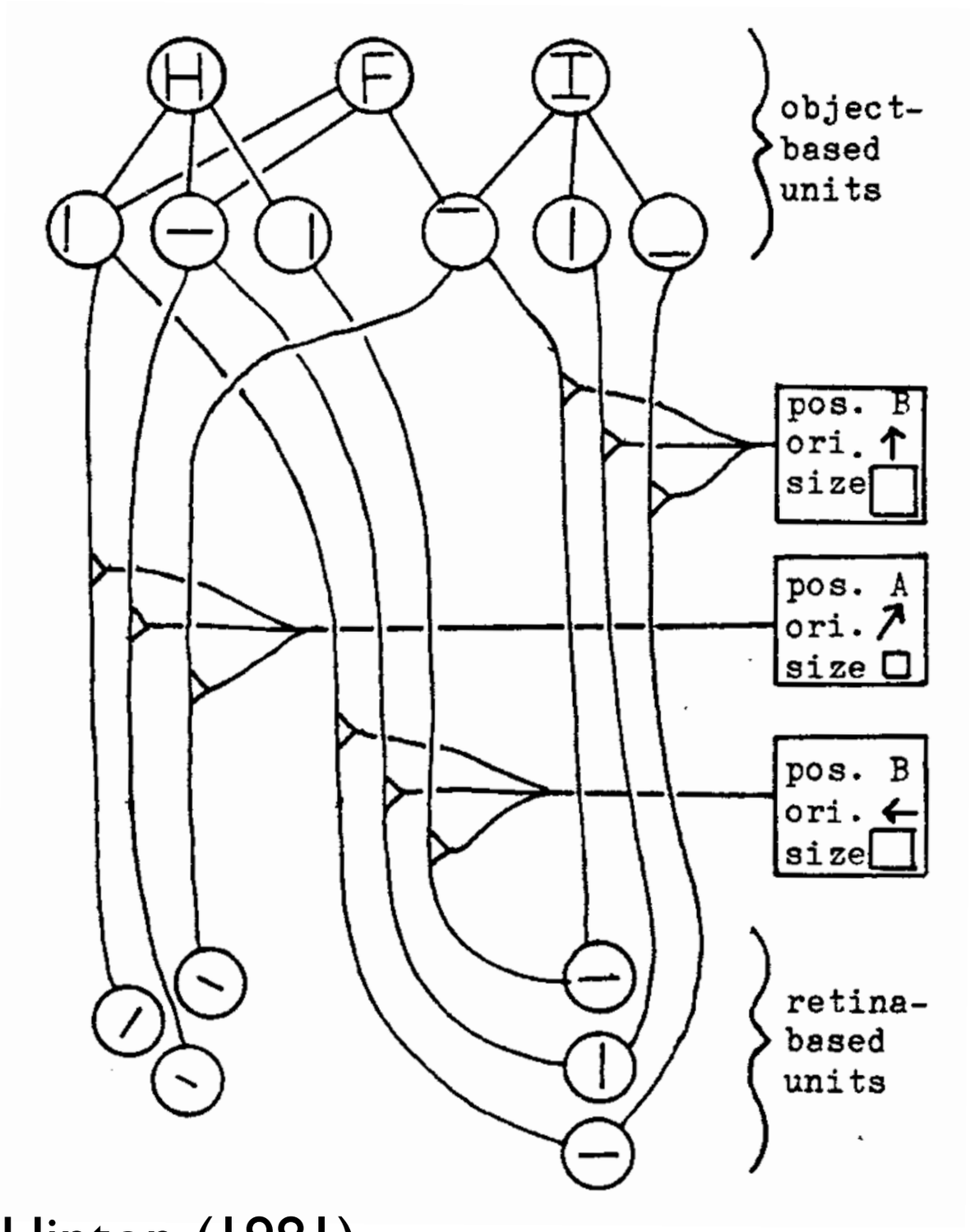


Reference frames require *structured representations*

The meaning of the triangular symbol in fig. 1 is quite complex. It stands for two rules:

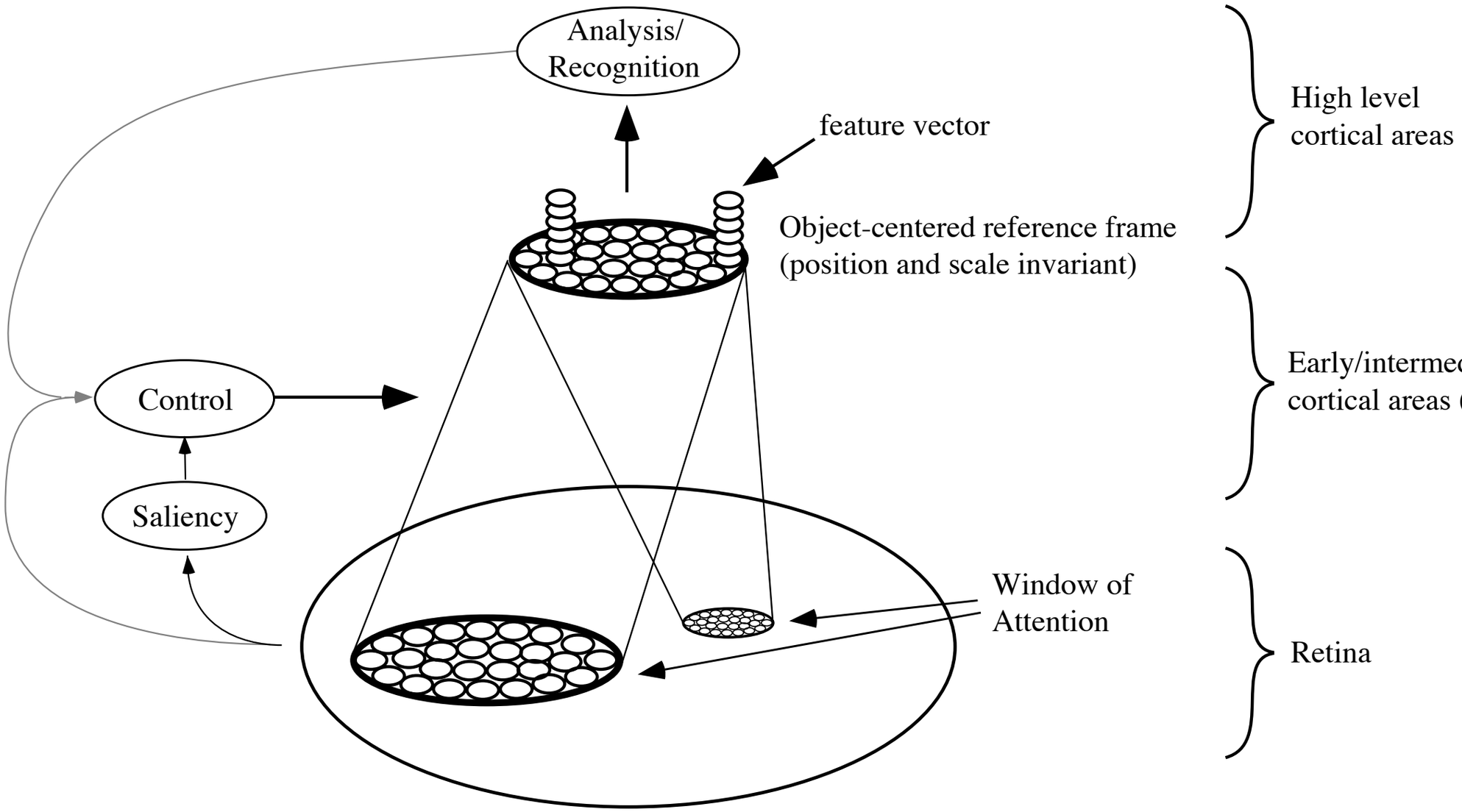
1. Multiply the activity level in the retina-based unit by the activity level in the mapping unit and send the product to the object-based unit.

2. Multiply the activity level in the retina-based unit by the activity level in the object-based unit and send the product to the mapping unit.

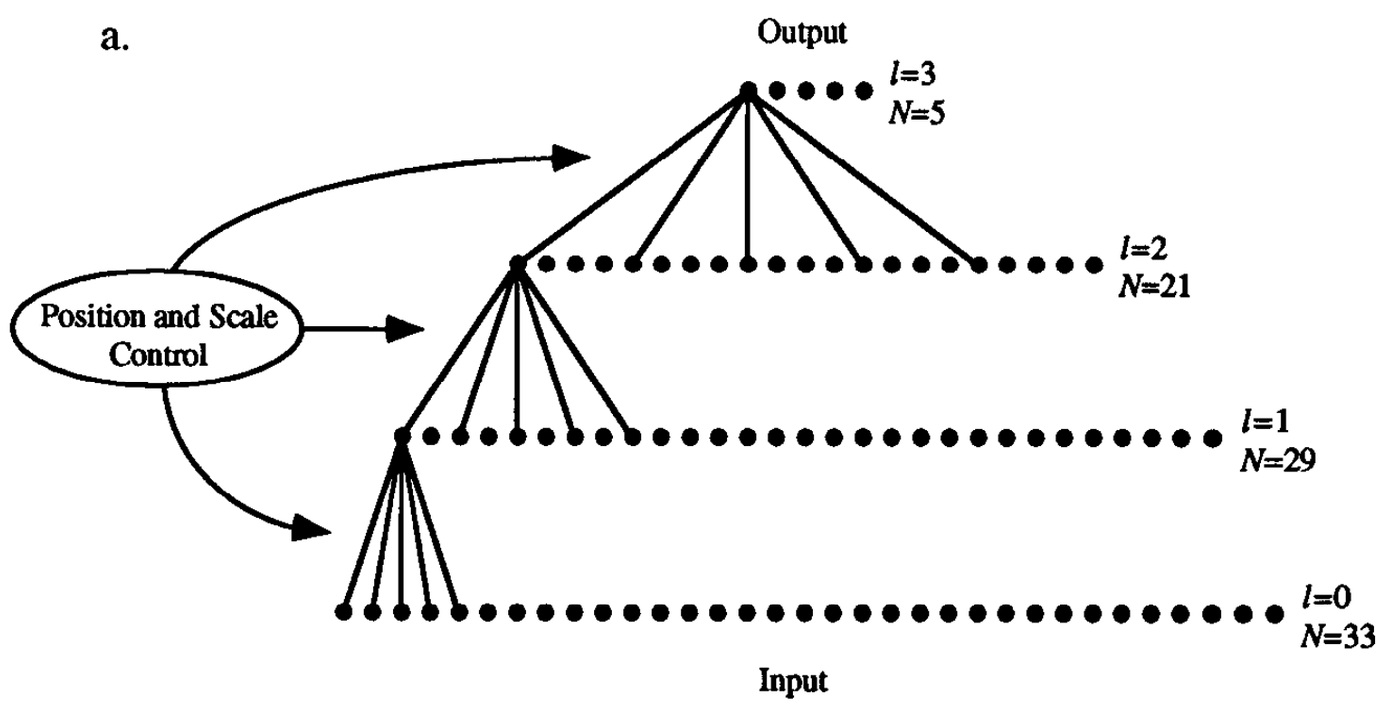


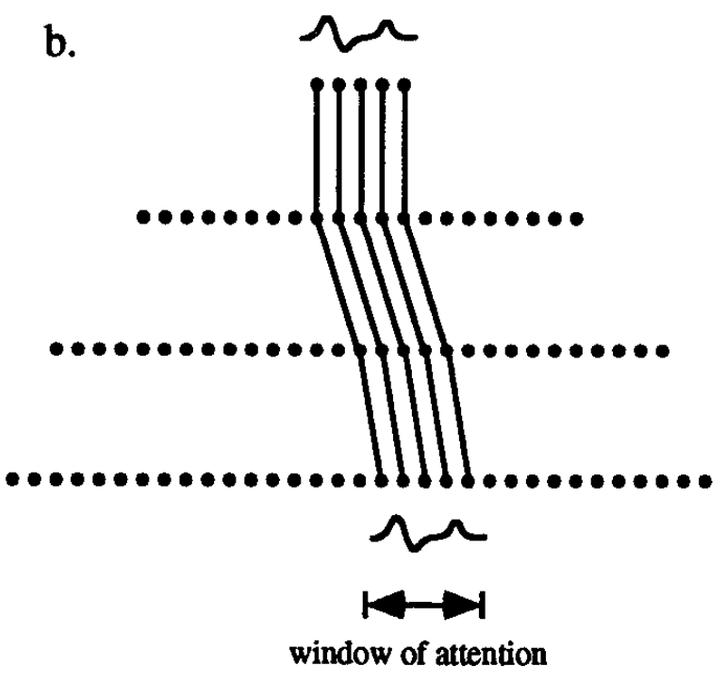
Hinton (1981)

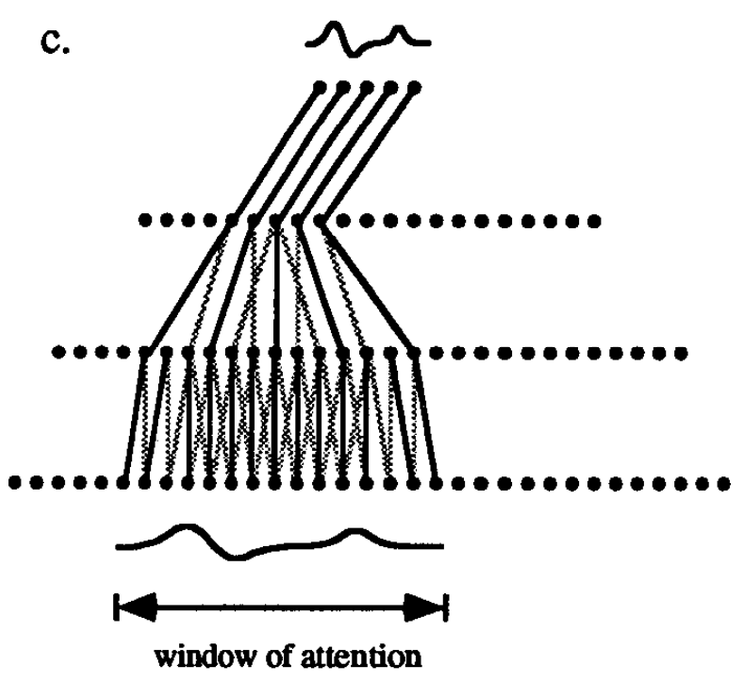
Dynamic routing
(Olshausen, Anderson, Van Essen 1993)



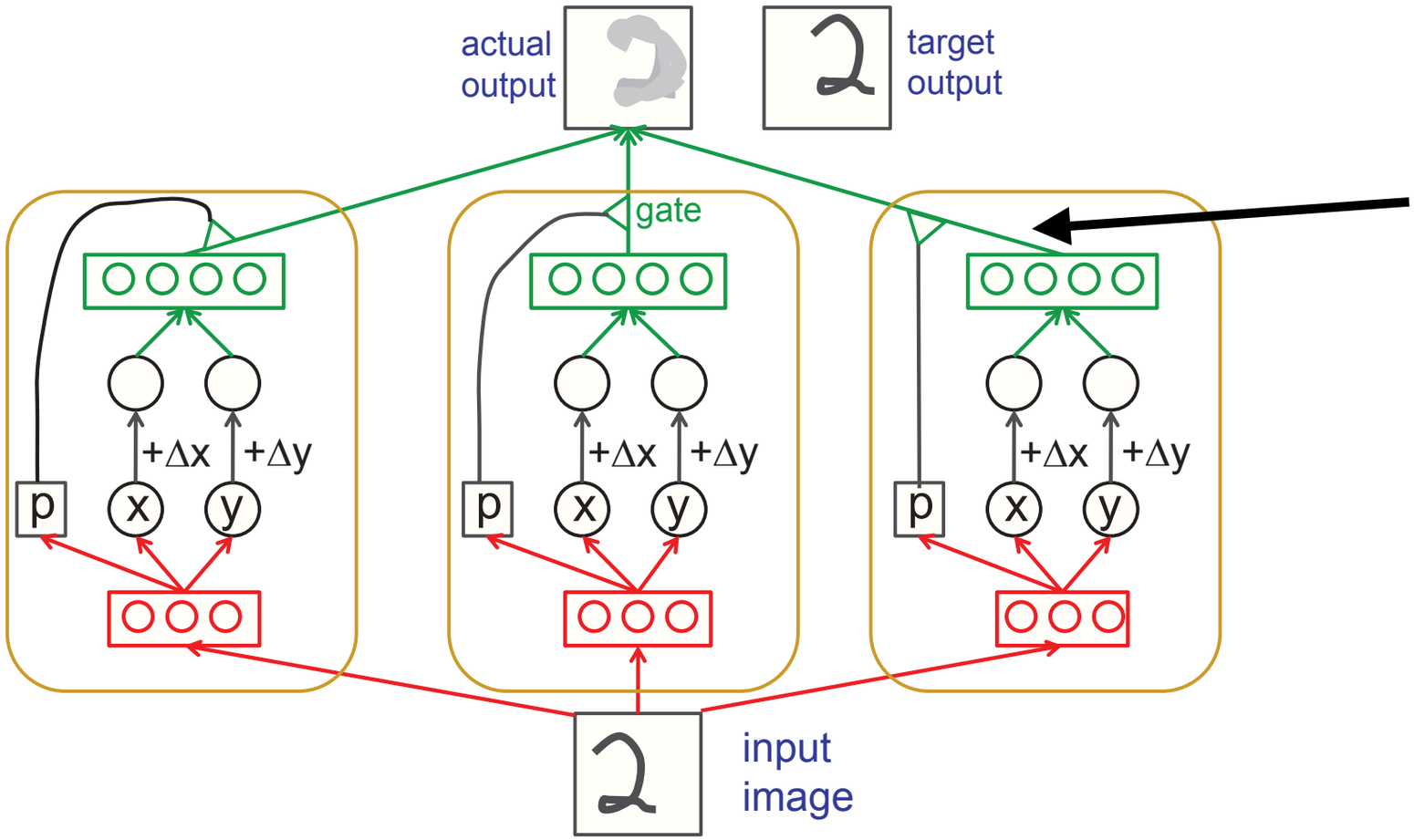
Dynamic routing circuit

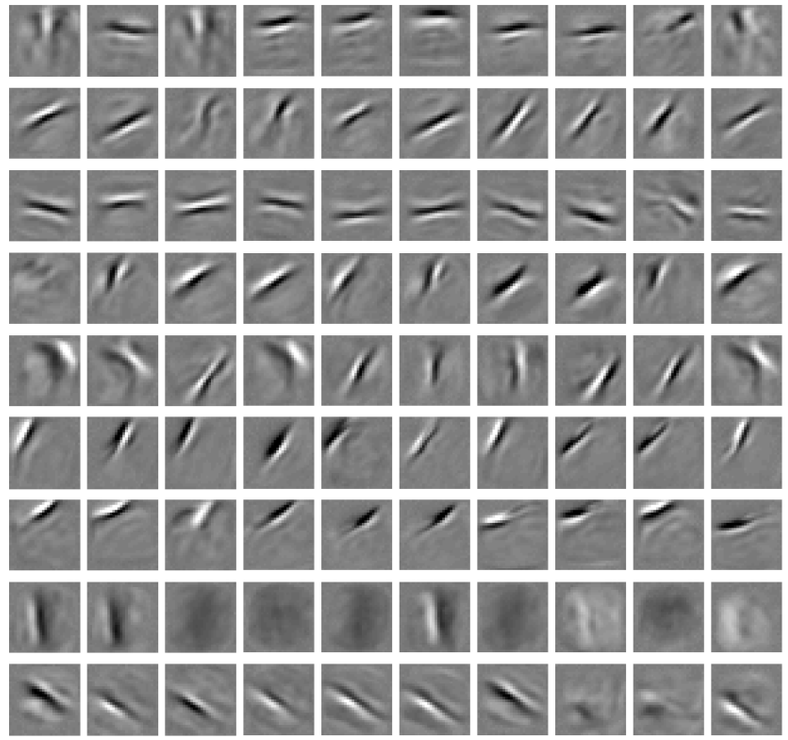




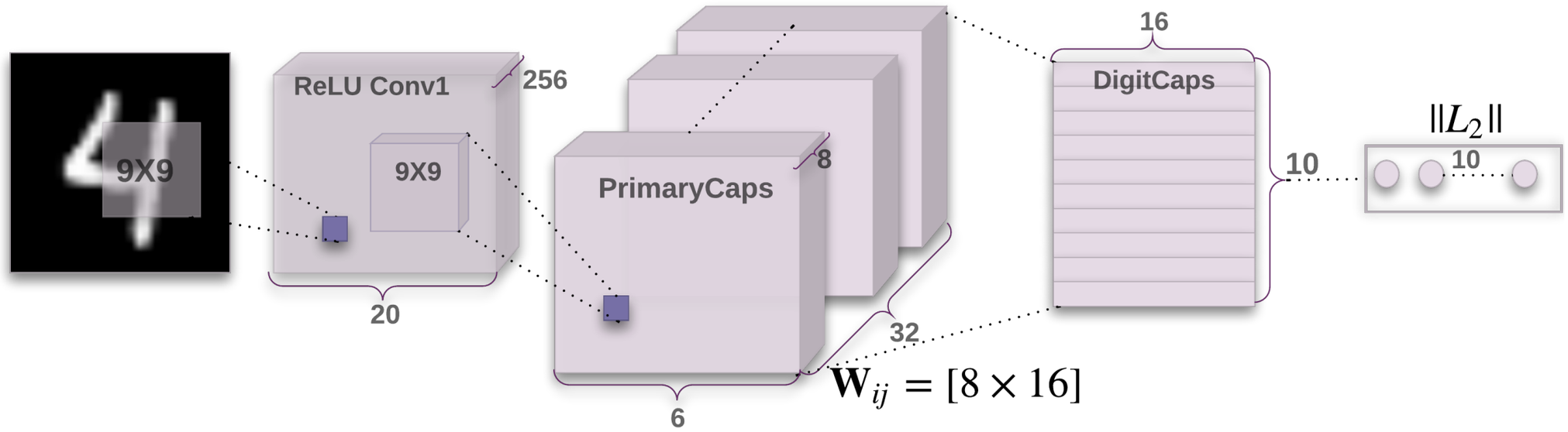


Transforming Auto-encoders
(Hinton, Krizhevsky & Wang 2011)





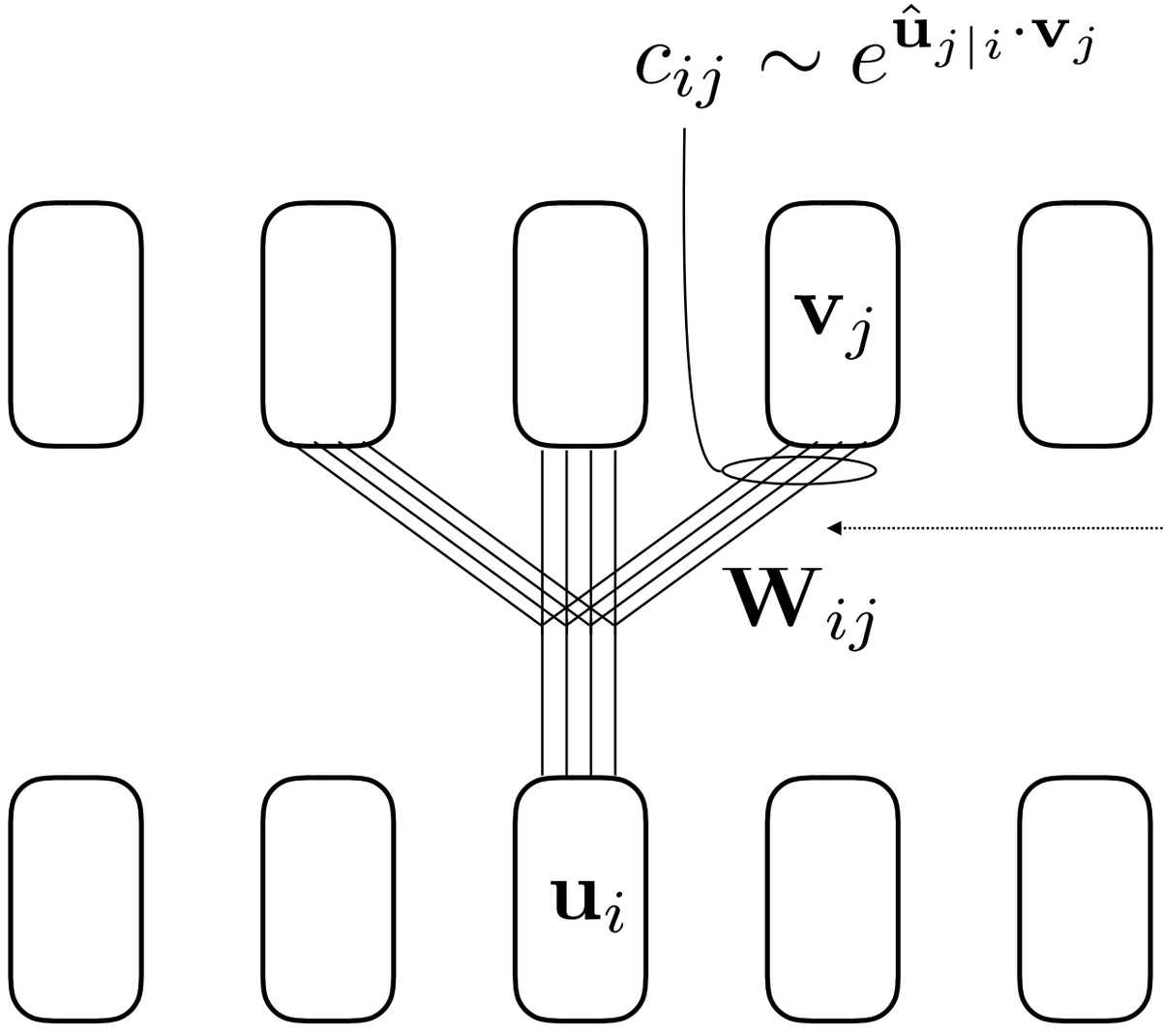
Dynamic routing between capsules
(Sabour, Frosst & Hinton 2017)



This type of "routing-by-agreement" should be far more effective than the very primitive form of routing implemented by max-pooling, which allows neurons in one layer to ignore all but the most active feature detector in a local pool in the layer below. We demonstrate that our dynamic routing mechanism is an effective way to implement the "explaining away" that is needed for segmenting highly overlapping objects.

For low level capsules, location information is "place-coded" by which capsule is active. As we ascend the hierarchy, more and more of the positional information is "rate-coded" in the real-valued components of the output vector of a capsule.

Dynamic routing between capsules
(Sabour, Frosst & Hinton 2017)



$$\mathbf{v}_j = \frac{||\mathbf{s}_j||^2}{1 + ||\mathbf{s}_j||^2} \frac{\mathbf{s}_j}{||\mathbf{s}_j||}$$

$$\mathbf{s}_j = \sum_i c_{ij} \hat{\mathbf{u}}_{j|i}$$

$$\hat{\mathbf{u}}_{j|i} = \mathbf{W}_{ij} \mathbf{u}_i$$