What and where: A Bayesian inference theory of attention

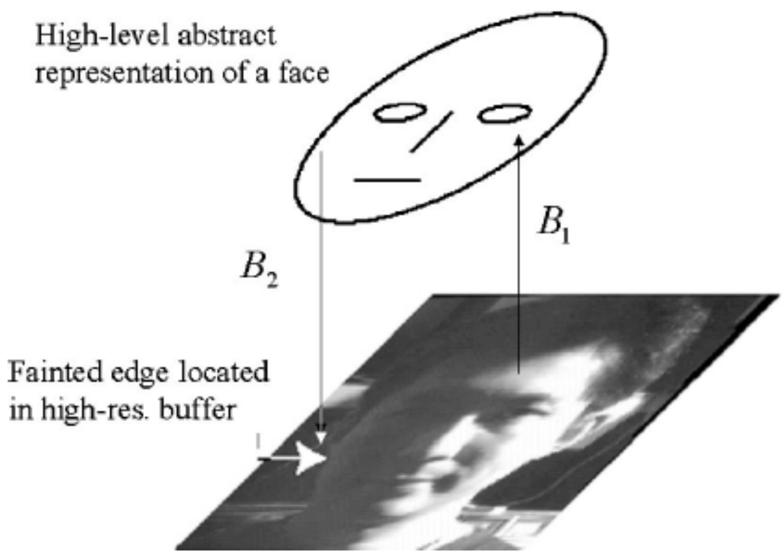
Sharat Chikkerur, Thomas Serre, Cheston Tan & Tomaso Poggio
CBCL, McGovern Institute for Brain Research, MIT

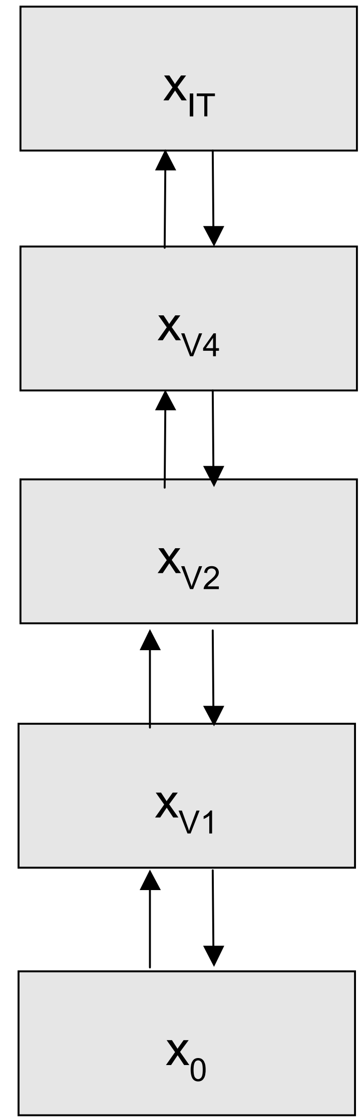
Outline

- Preliminaries
  - Perception & Bayesian inference
- Background & motivation
- Theory
  - Attention as inference
  - Bayesian model
- Computational model
  - Model properties
- Applications on real-world images
  - Predicting human eye movements
  - Improving object recognition

Perception as Bayesian inference

- Mumford and Lee, “Hierarchical Bayesian Inference in the Visual Cortex”, JOSA, 20(7), 2003
- Recurrent feed-forward/feedback loops integrate bottom up information with top down priors
- Bottom-up signals : Data dependent
- Top-down signals : Task dependent
- Top down signals provide context information and help to disambiguate bottom-up signals





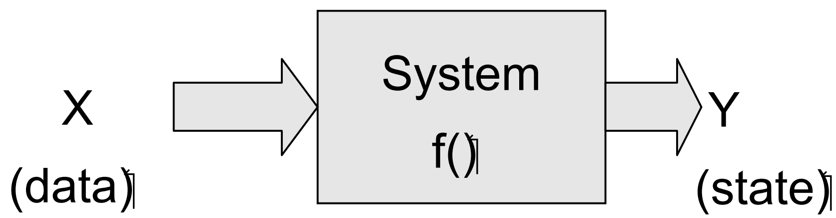
Bottom up vs. top-down

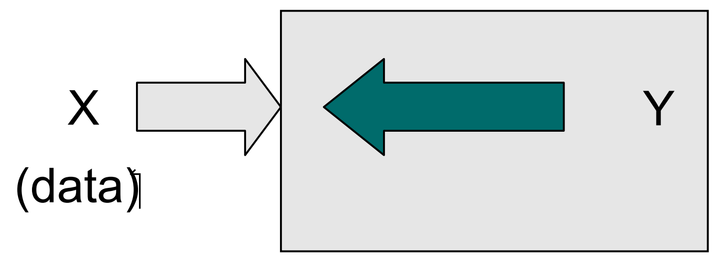
Hegde J. and Felleman J., Reappraising the functional implications of the primate visual anatomical hierarchy, Neuroscientist, 13(5), 2007

Bottom up vs. top-down

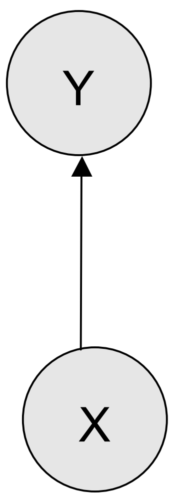
Mathematical framework

Bayesian generative models



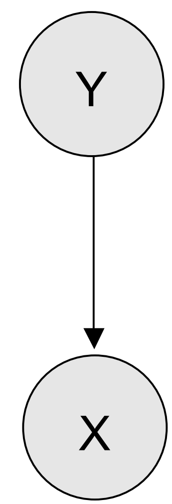


- Statistical learning view:
  - Y = f(X) , X-data, Y-class

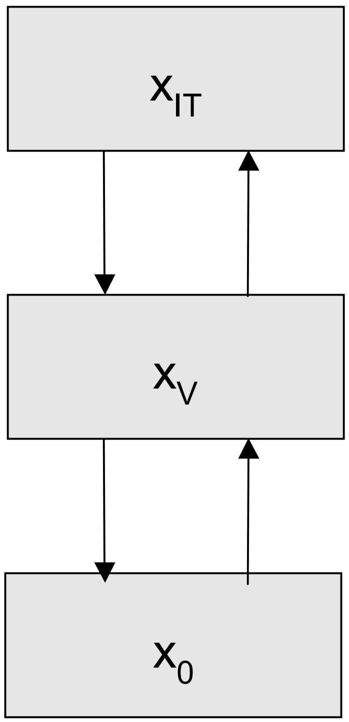


- Generative model view:
  - X-random, Y-random
  - X~P(X|Y) ,

$$P(Y \mid X) \propto P(X \mid Y) P(Y)$$



Perception: bottom-up & top-down



- Recall,

$$P(A,B)=P(A\mid B)P(B)=P(B\mid A)P(A)$$

$$P(B\mid A)=\frac{P(A\mid B)P(B)}{P(A)}$$

$$P(A)=\sum_B P(A,B)$$

- For the given network,

$$P(x_{IT},x_V,x_0)=P(x_0\mid x_V)P(x_V\mid x_{IT})P(x_{IT})$$

$$P(x_V,x_0\mid x_{IT})=P(x_0\mid x_V,x_{IT})P(x_V\mid x_{IT})$$
$$=P(x_0\mid x_V)P(x_V\mid x_{IT})$$

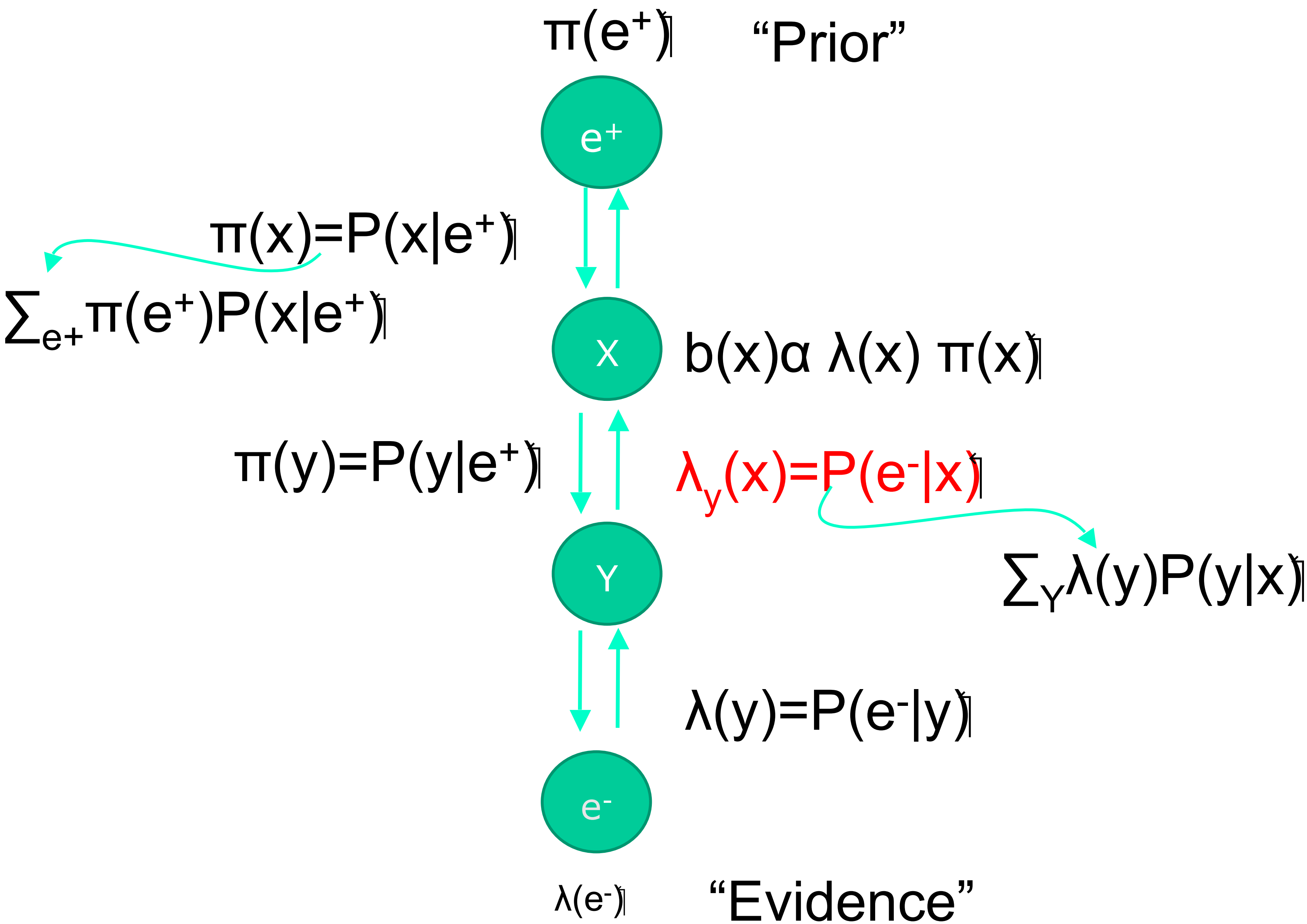
- Inference:

$$P(x_V\mid x_0,x_{IT})=\frac{P(x_0\mid x_V,x_{IT})P(x_V\mid x_{IT})}{P(x_0\mid x_{IT})}$$

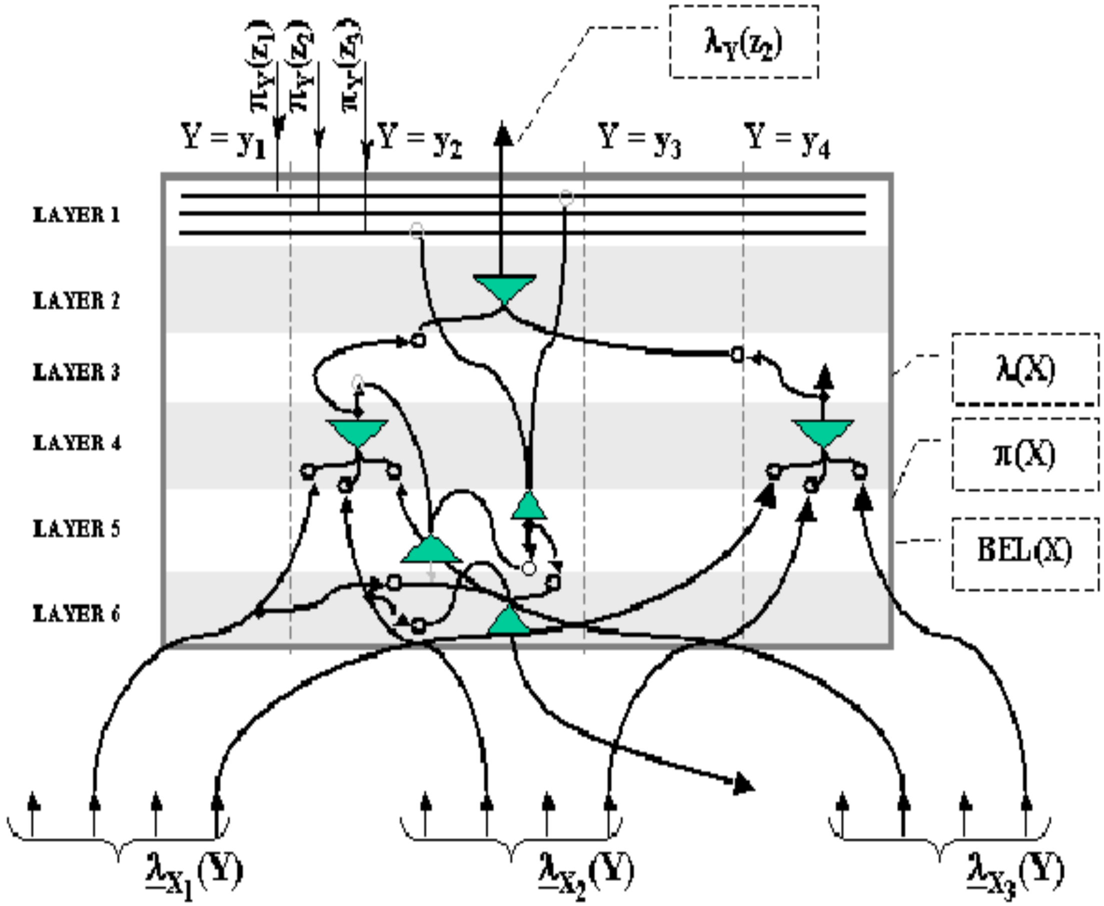
$$P(x_0\mid x_V)P(x_V\mid x_{IT})$$

Bottom-up Top-down

Belief propagation



Biological plausibility

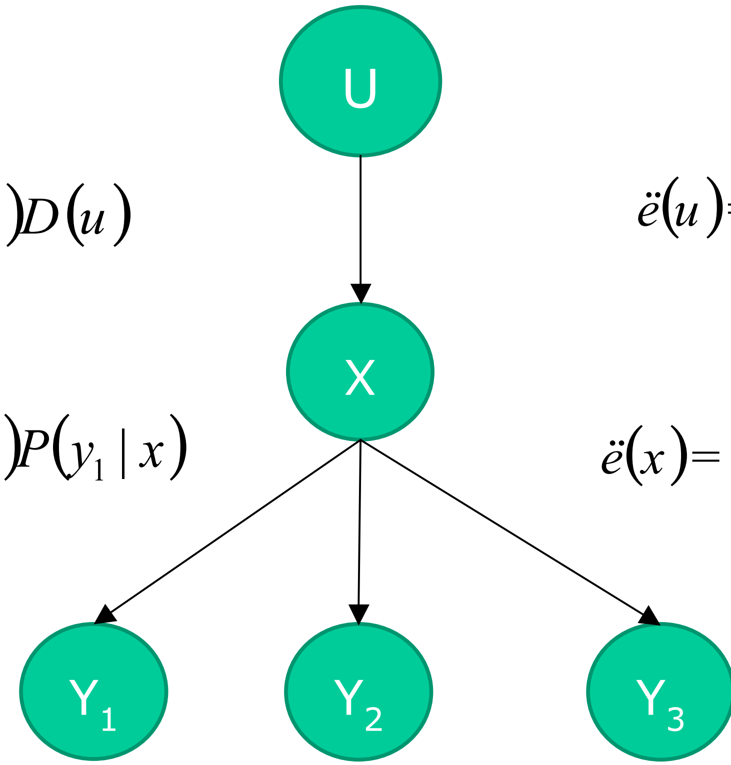


George D. and Hawkins J., A hierarchical Bayesian model of invariant pattern recognition in the visual cortex, IJCNN, 2005

Trees

$$D(x) = \sum_{u} P(x \mid u) D(u)$$

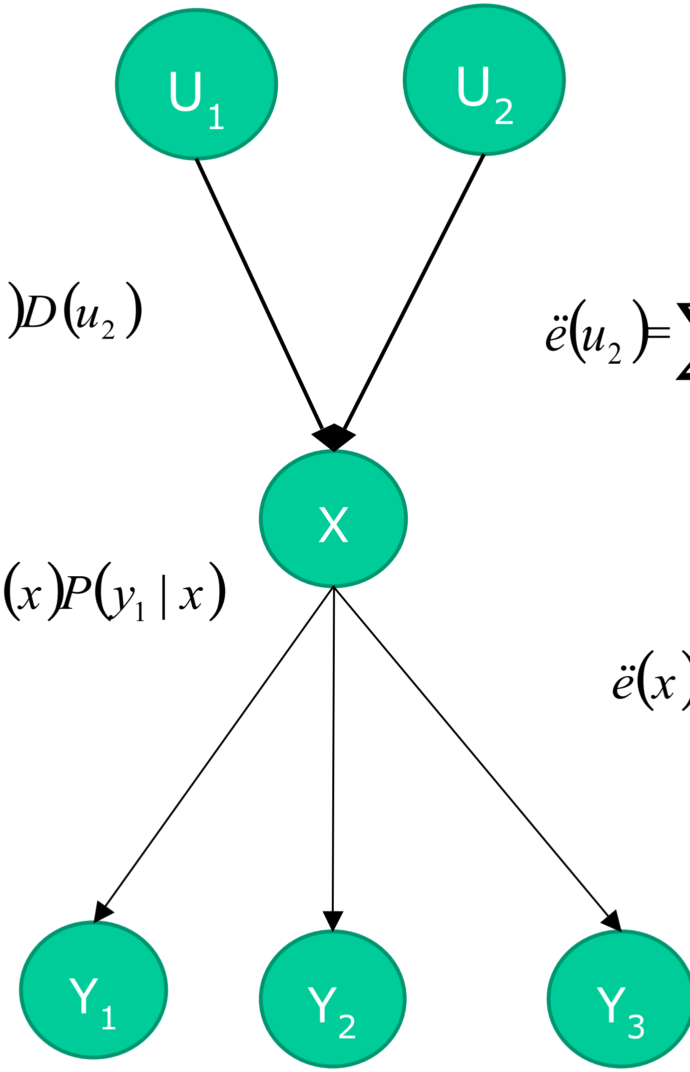
$$\ddot{e}(u) = \sum_{x} \ddot{e}(x) P(x \mid u)$$



$$D(y_1) = c \sum_{x} D(x) \ddot{e}_{y_1}(x) \ddot{e}_{y_2}(x) P(y_1 \mid x)$$

$$\ddot{e}(x) = \ddot{e}_{y_1}(x) \ddot{e}_{y_2}(x) \ddot{e}_{y_3}(x)$$

Polytrees



$$D(x) = \sum_{u_1 u_2} P(x \mid u_1 u_2) D(u_1) D(u_2)$$

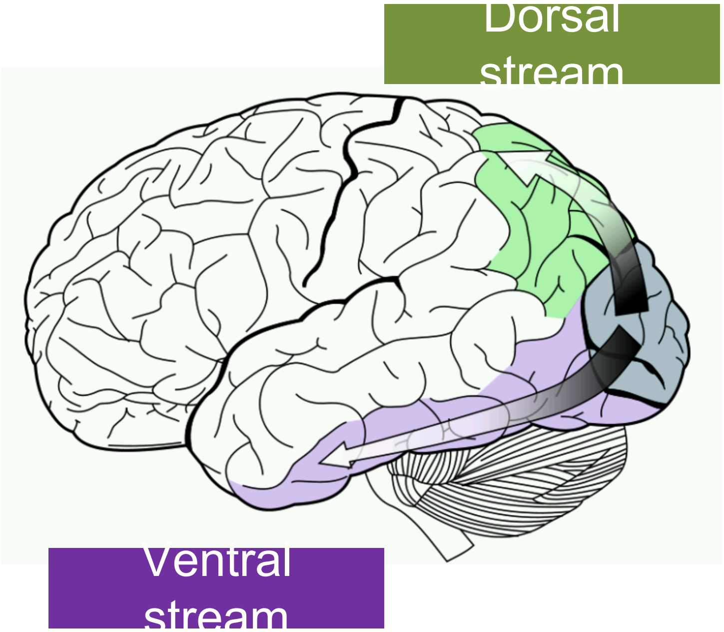
$$\ddot{e}(u_2) = \sum_{x} \ddot{e}(x) \sum_{u_1} P(x \mid u_1, u_2) D(u_1)$$

$$D(y_1) = c \sum_{x} D(x) \ddot{e}_{y_1}(x) \ddot{e}_{y_2}(x) P(y_1 \mid x)$$

$$\ddot{e}(x) = \prod \ddot{e}_{y_1}(x) \ddot{e}_{y_2}(x) \ddot{e}_{y_3}(x)$$

Attention
Background & motivation

Visual processing: what and where

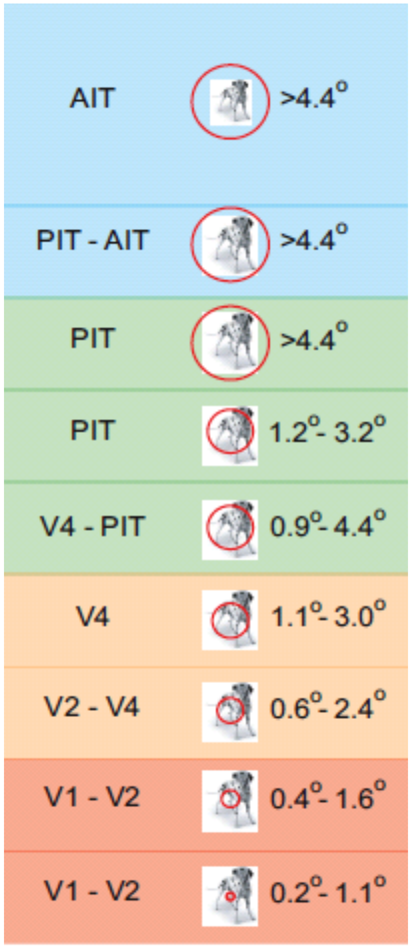


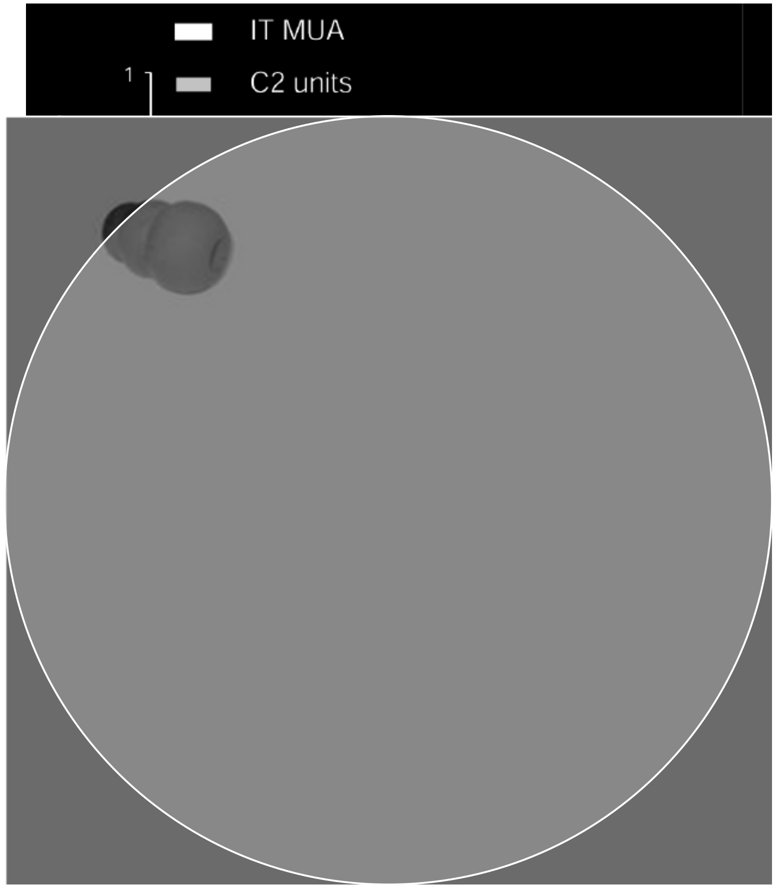
- Ventral ('what') stream:
  - Processes shape information
  - Responsible for object recognition
  - Progressive loss of location information
- Dorsal ('where') stream:
  - Processes location and motion information
  - Progressive loss of form information

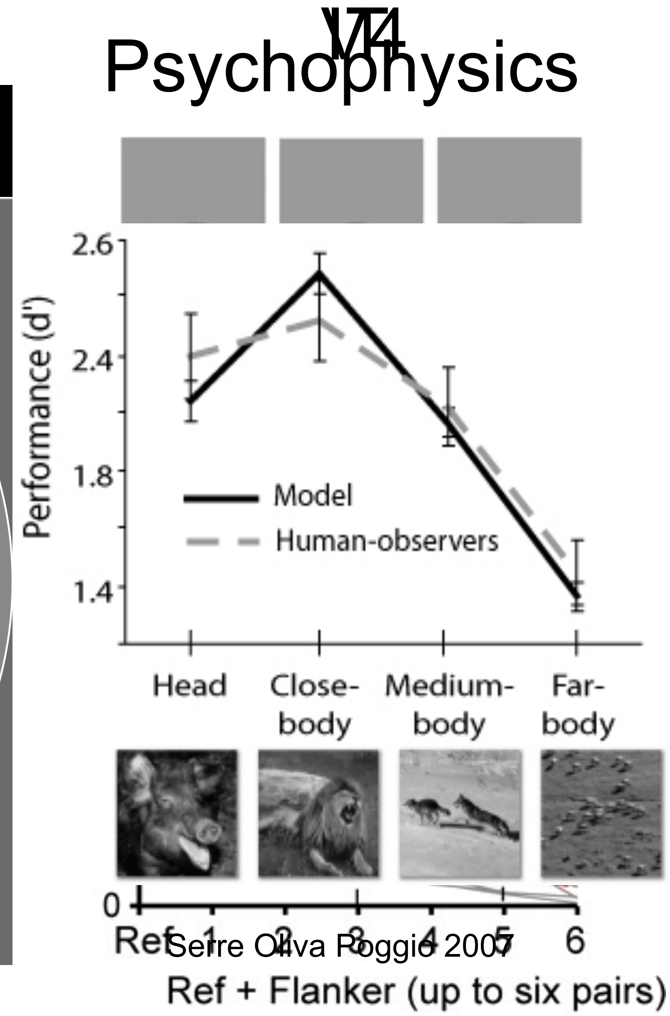
•Form and location is processed concurrently and (almost) independently of each other

•How does the brain combine form and location information?

Ventral stream: invariant recognition





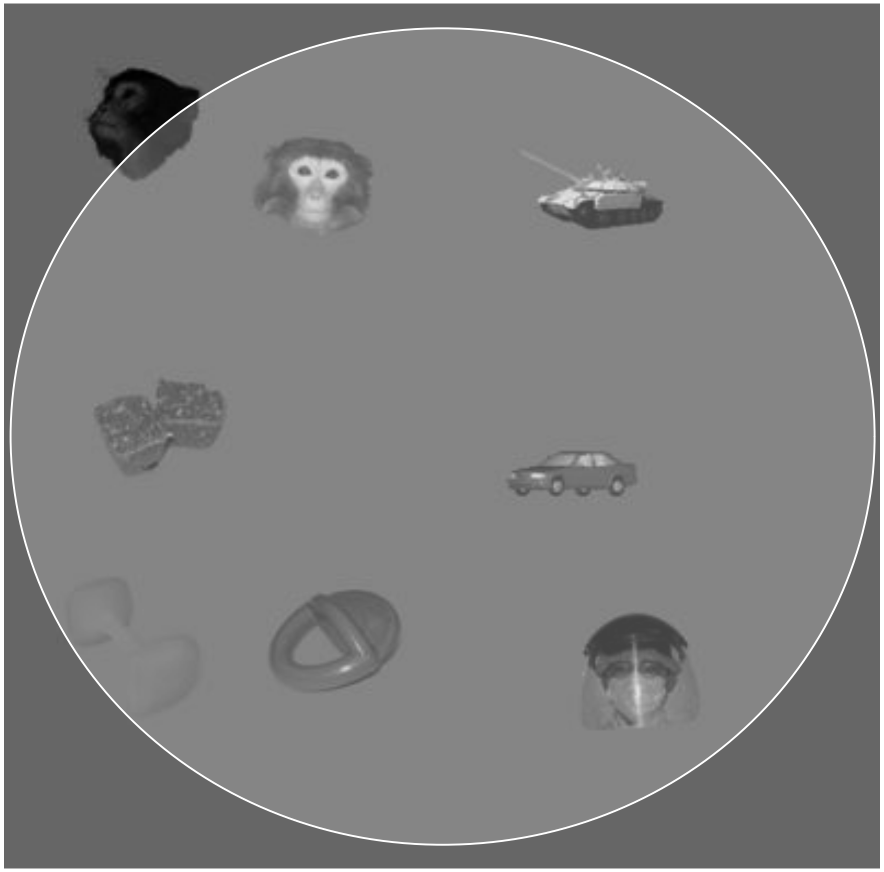


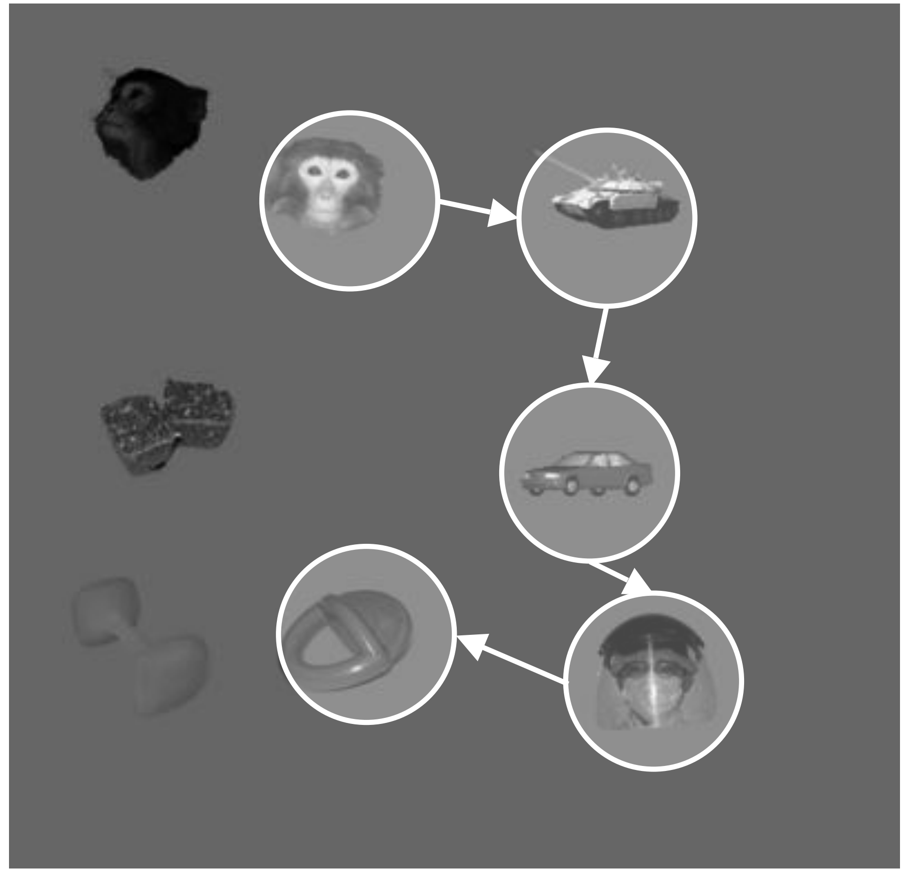
Zoccolan Kouh Poggio DiCarlo 2007

•How does the brain recognize objects under clutter?

Figures from Serre et al, Hung et al.

Parallel vs. serial processing





Attention is needed to recognize objects under clutter

•Filter theory (Broadbent)
•Biased competition (Desimone)
•Feature integration theory (Treisman)
•Guided search (Wolfe)
•Scanpath theory (Noton)

•Bayesian surprise (Itti)
•Bottleneck (Tsotsos)

Computational Role

Attention

Biology

Effects

Everybody knows what attention is…
-William James, 1907

• V1
• V4
• MT
• LIP
• FEF

•Contrast gain
•Response gain
•Modulation under spatial attention
•Modulation under feature attention

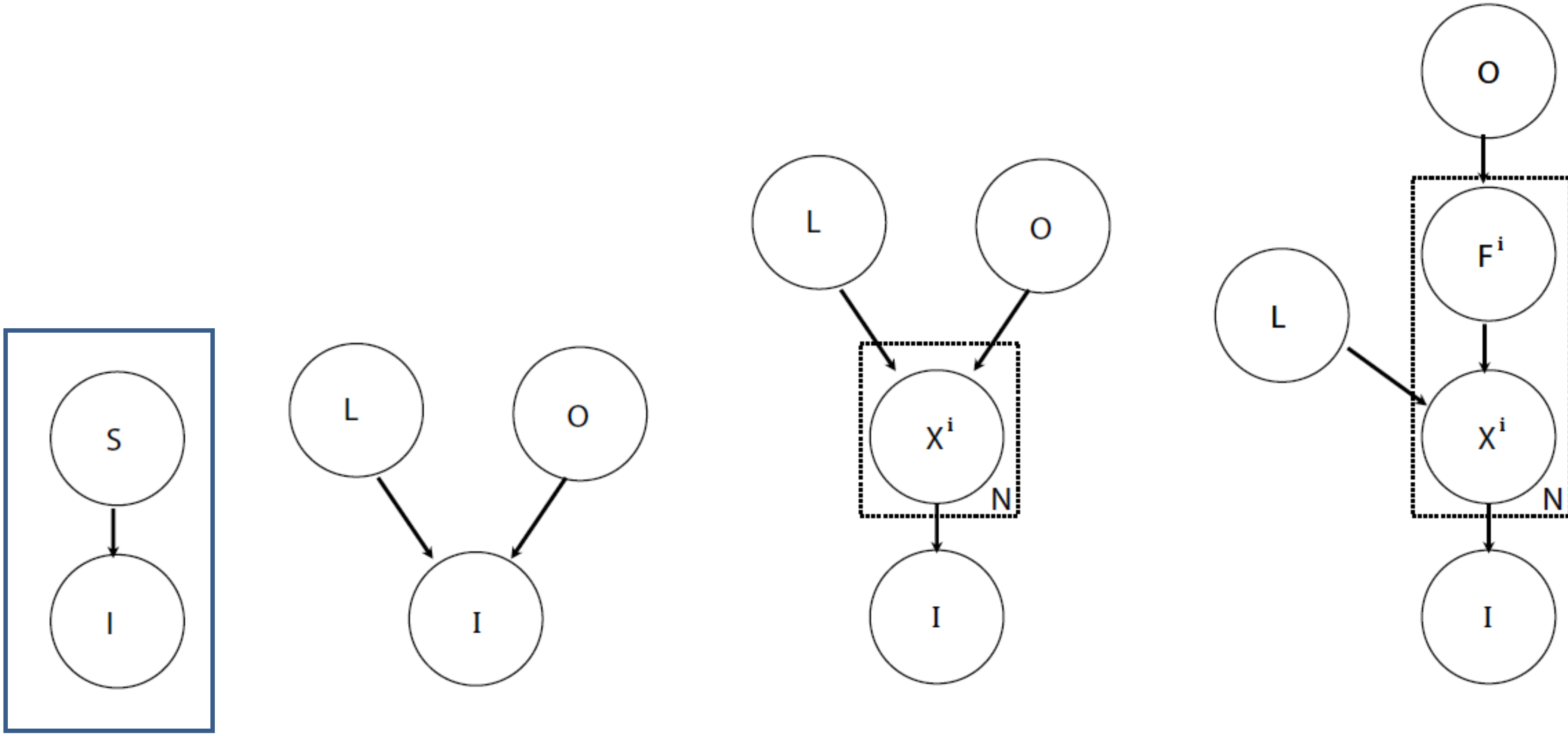
•Pop-out
•Serial vs. Parallel
•Bottom-up vs. Top-down

Bridging the gap

- Conceptual models (theories)
  - Provide justifications not implementations
- Computational models
  - Model behavior (eye-movements)
  - Cannot model physiological effects
- Phenomenological models
  - Model specific physiological effects
  - Cannot provide theory
- Bridging the gap
  - Phenomenological, predicts behavior, theory

A theoretical framework

Bayesian model



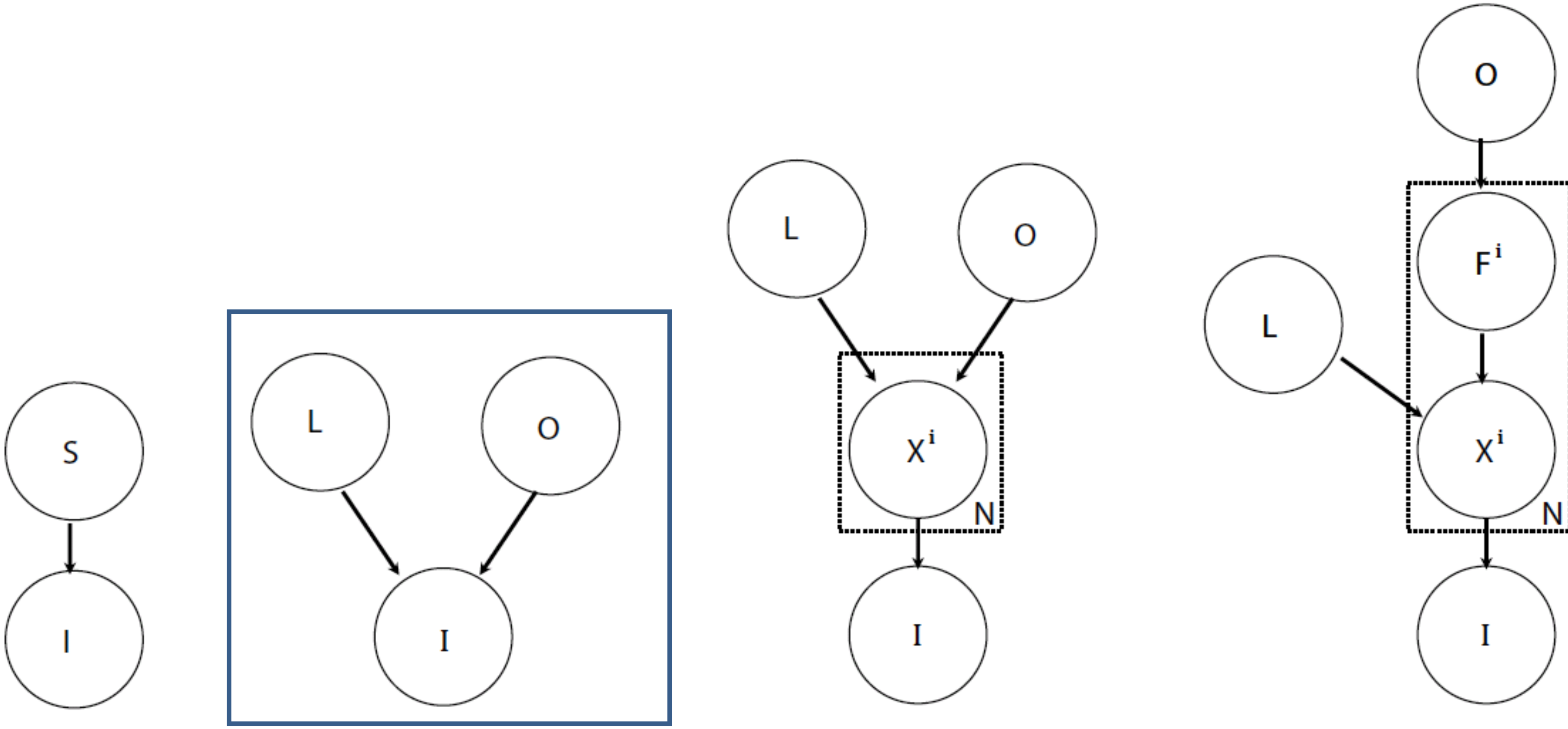
$$P(S, I) = P(I|S)P(S)$$ Kersten & Yuille '04

$$S = \{O_1, O_2, \cdots, O_n, L_1, L_2, \cdots, L_n\}$$

$$P(S, I) = P(O_1, L_1, O_2, L_2, \cdots, O_n, L_n, I)$$

$$P(O_1, L_1, I) = \sum_{O_2 \cdots O_n, L_2 \cdots L_n} P(O_1, L_1, O_2 \cdots, O_n, L_2, \cdots, L_n, I)$$ ects, one at

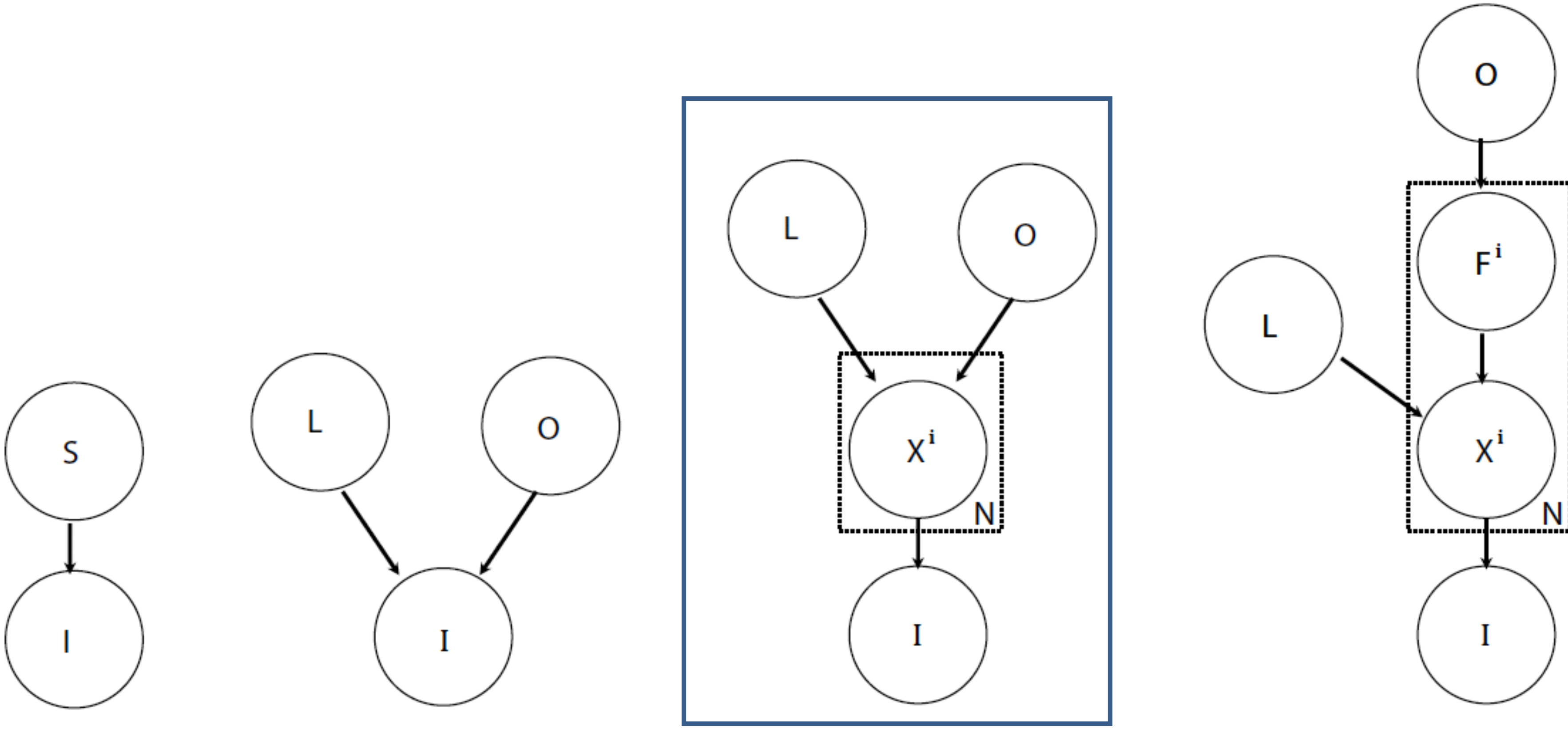
Bayesian model



Assumption: object location and identity are marginally independent of each other

$$P(O, L, I) = P(O)P(L)P(I|L, O)$$

Bayesian model

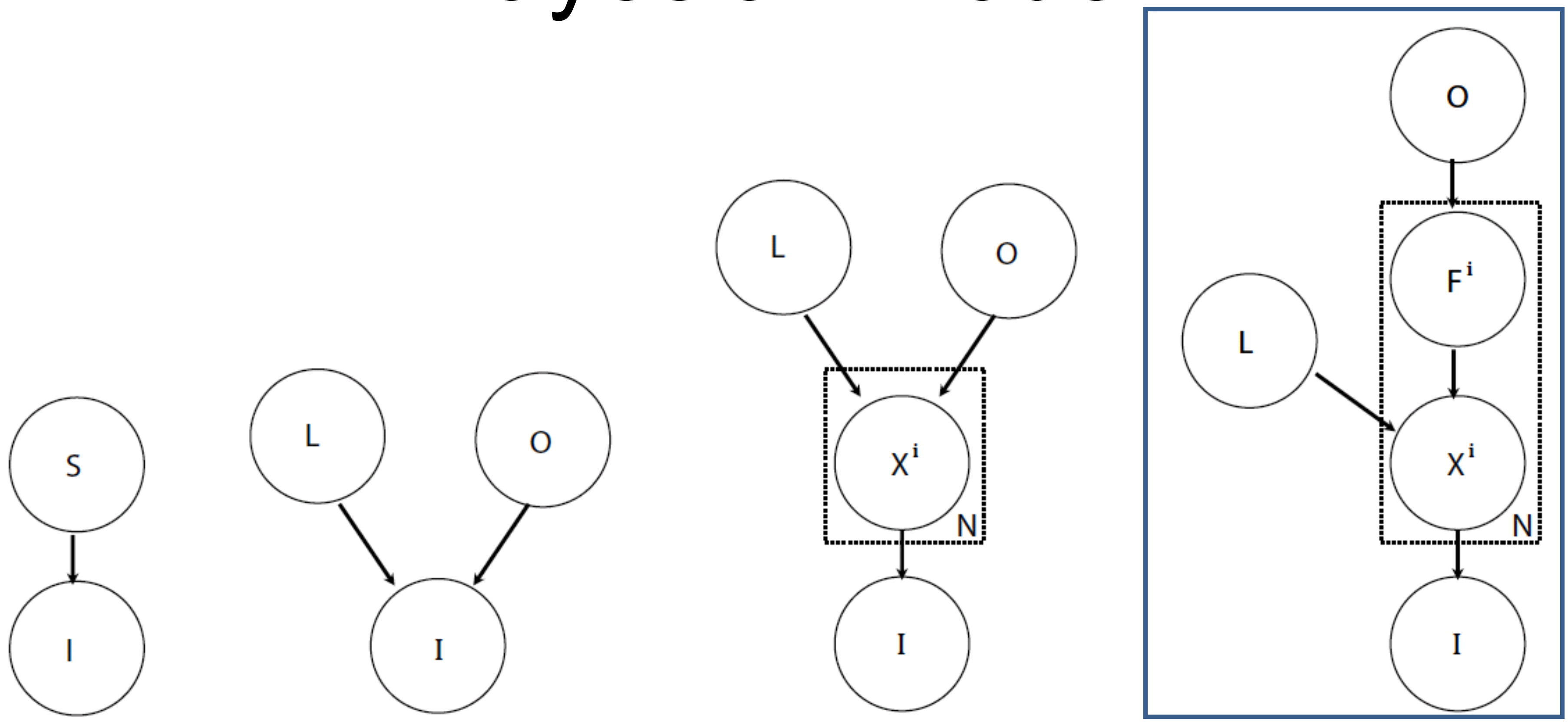


Assumption: Every object is generated using a set of N complex features each of which may be present or absent

$$P(O, L, X^1, \cdots, X^N, I)$$

$$= P(O)P(L)\left\{\prod_{i=1}^{i=N}\{P(X^i|L,O)\}\right\}P(I|X^1,\cdots,X^N)$$

Bayesian model

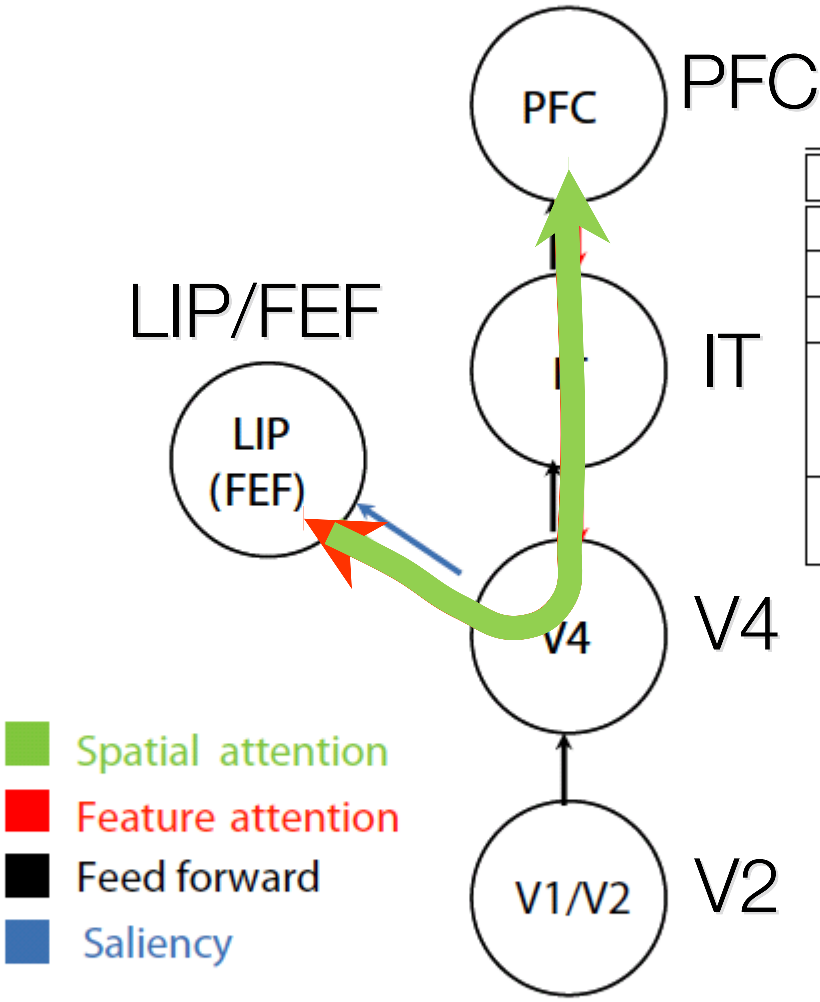


F^i: Location/scale invariant features

$$P(O, L, X^1, \cdots, X^N, F^1, \cdots, F^N, I)$$
$$= P(O)P(L)\left\{\prod_{i=1}^{i=N}\{P(X^i|L, F^i)P(F^i|O)\}\right\}P(I|X^1, \cdots, X^N)$$

Computational model

Relation to biology



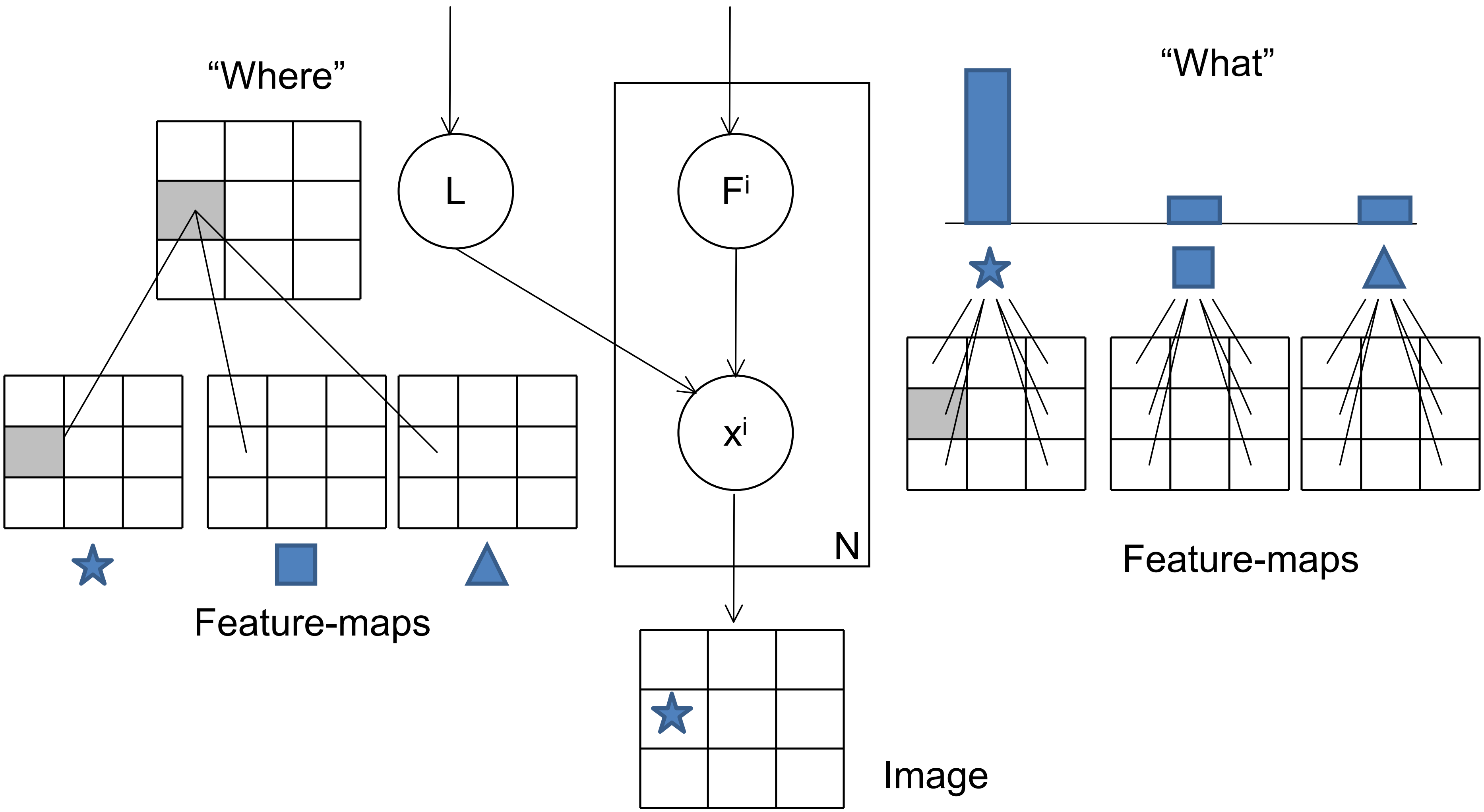
Model unit	Brain area	Representation
L	LIP/FEF	Discrete: $L \in \{1, 2 \cdots \lvert L \rvert\}$
O	PFC	Discrete: $O \in \{1, 2 \cdots \lvert O \rvert\}$
F^i	IT	Binary : $F^i \in \{0, 1\}$
X^i	V4	Discrete: $X^i \in \{0, 1, 2 \cdots \lvert L \rvert\}$ Value $\{1, 2, \cdots, \lvert L \rvert\}$. F^i is present Value $X^i = 0$. F^i is absent
I	V1/V2	Feedforward evidence obtained from lower areas

$$
\begin{aligned}
m_{O \to F^i} &= P(O) \\
m_{F^i \to X^i} &= \sum_{O} P(F^i|O)P(O) \\
m_{L \to X^i} &= P(L) \\
m_{I \to X^i} &= P(I|X^i) \\
m_{X^i \to F^i} &= \sum_{L} \sum_{X^i} P(X^i|F^i, L)(m_{L \to X^i})(m_{I \to X^i}) \\
m_{X^i \to L} &= \sum_{F^i} \sum_{X^i} P(X^i|F^i, L)(m_{F^i \to X^i})(m_{I \to X^i})
\end{aligned}
$$

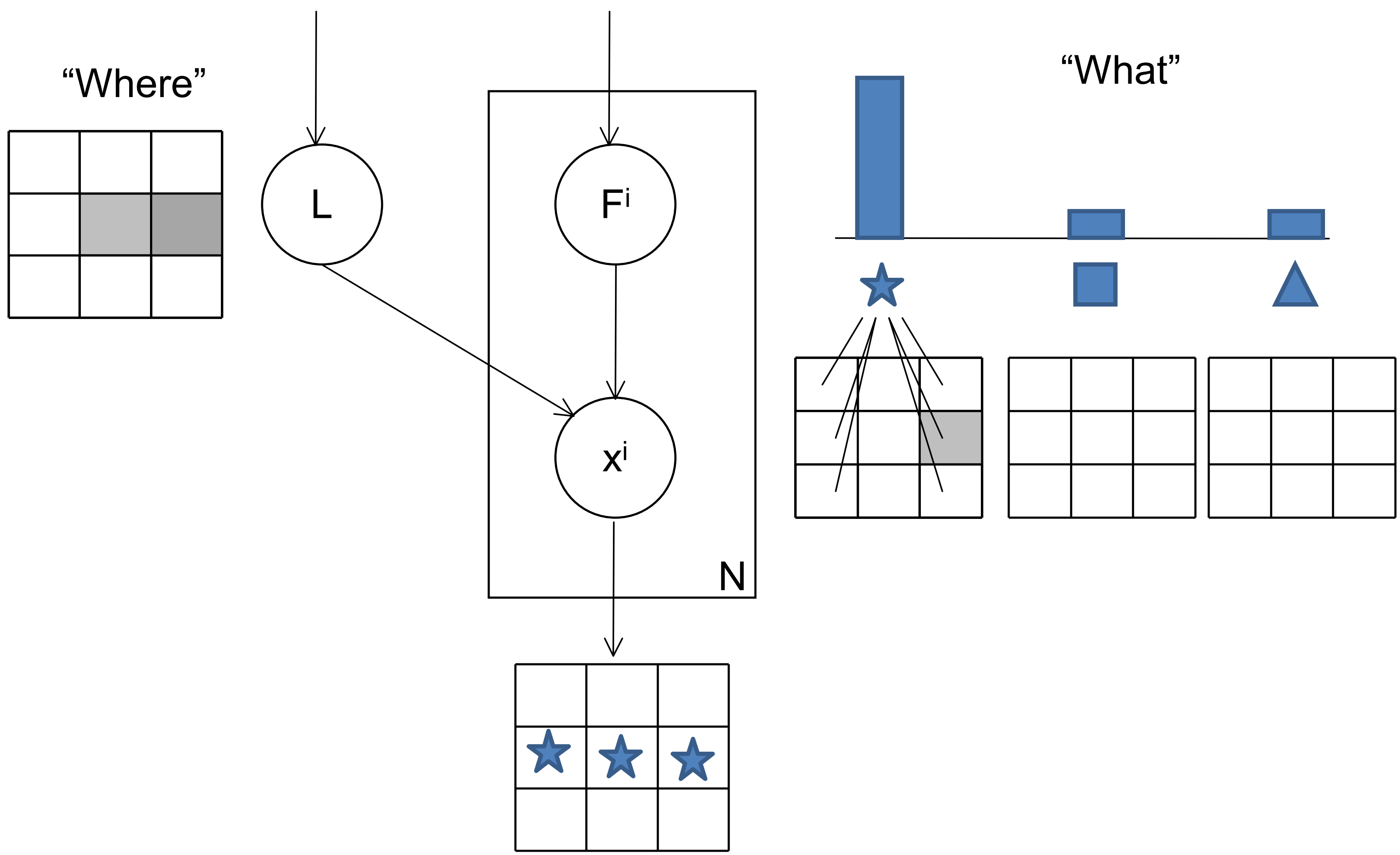
Spatial attention: What is at location L?

Feature-based attention: Where is object O?

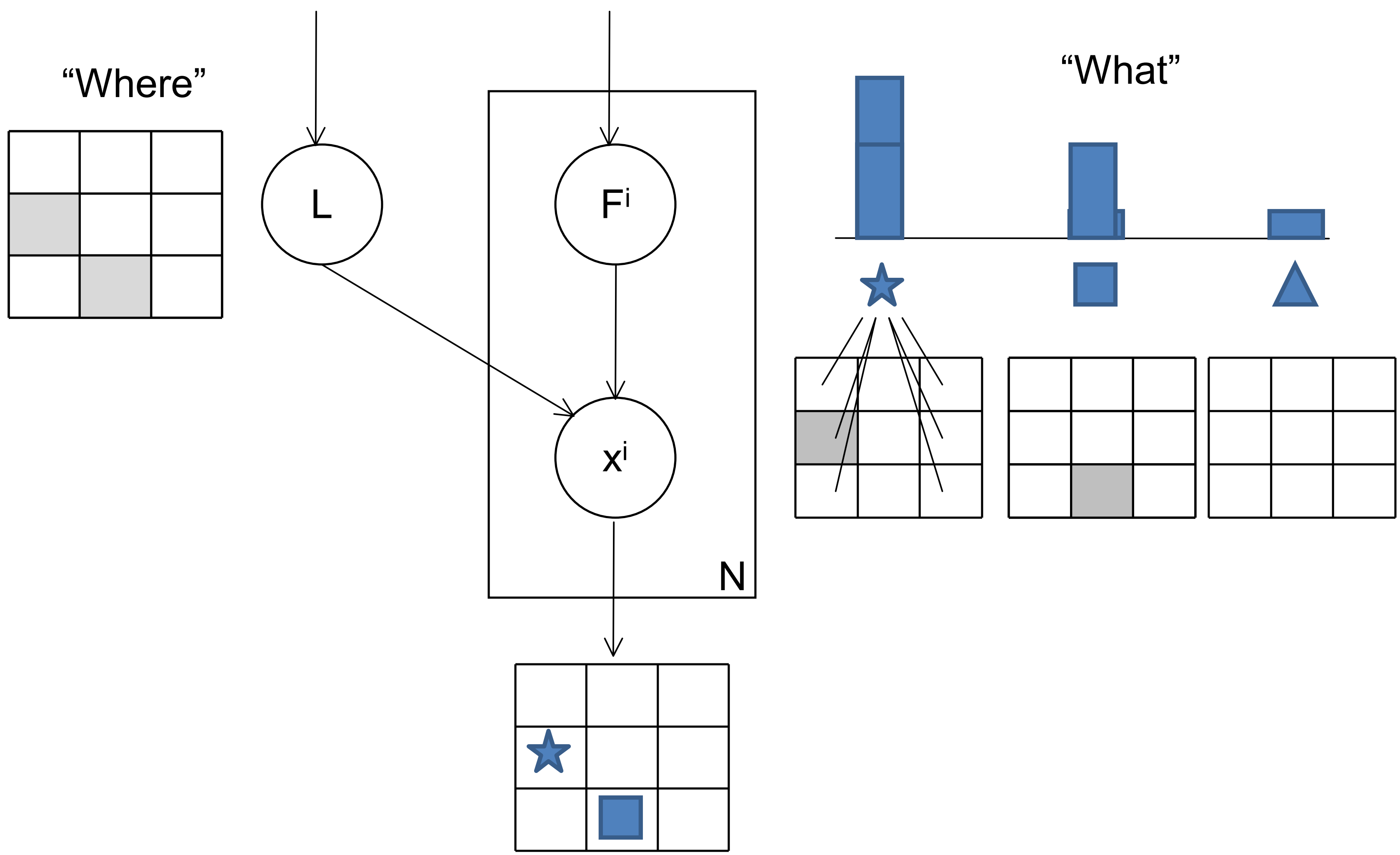
Model description



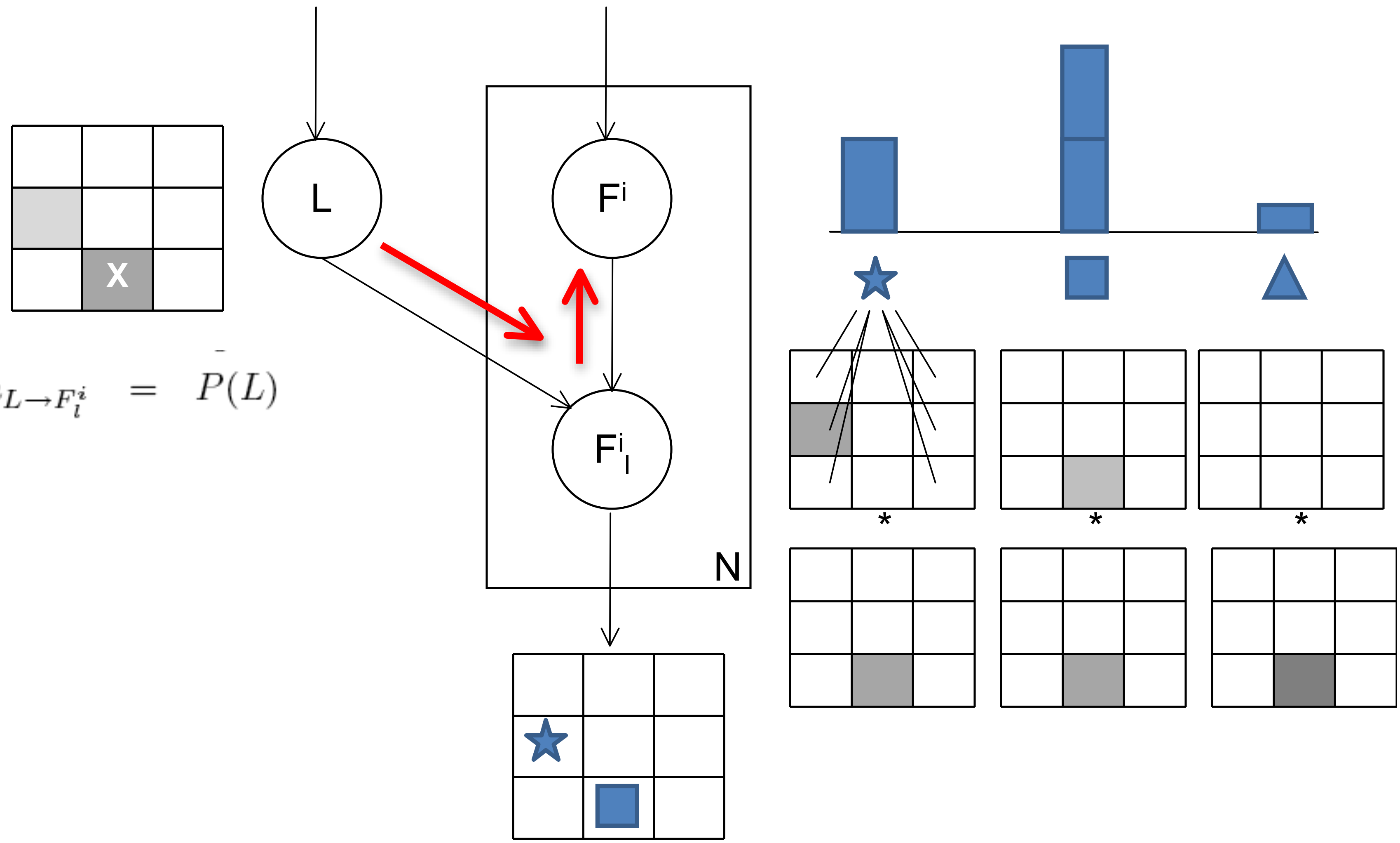
Model properties: invariance



Model properties: crowding



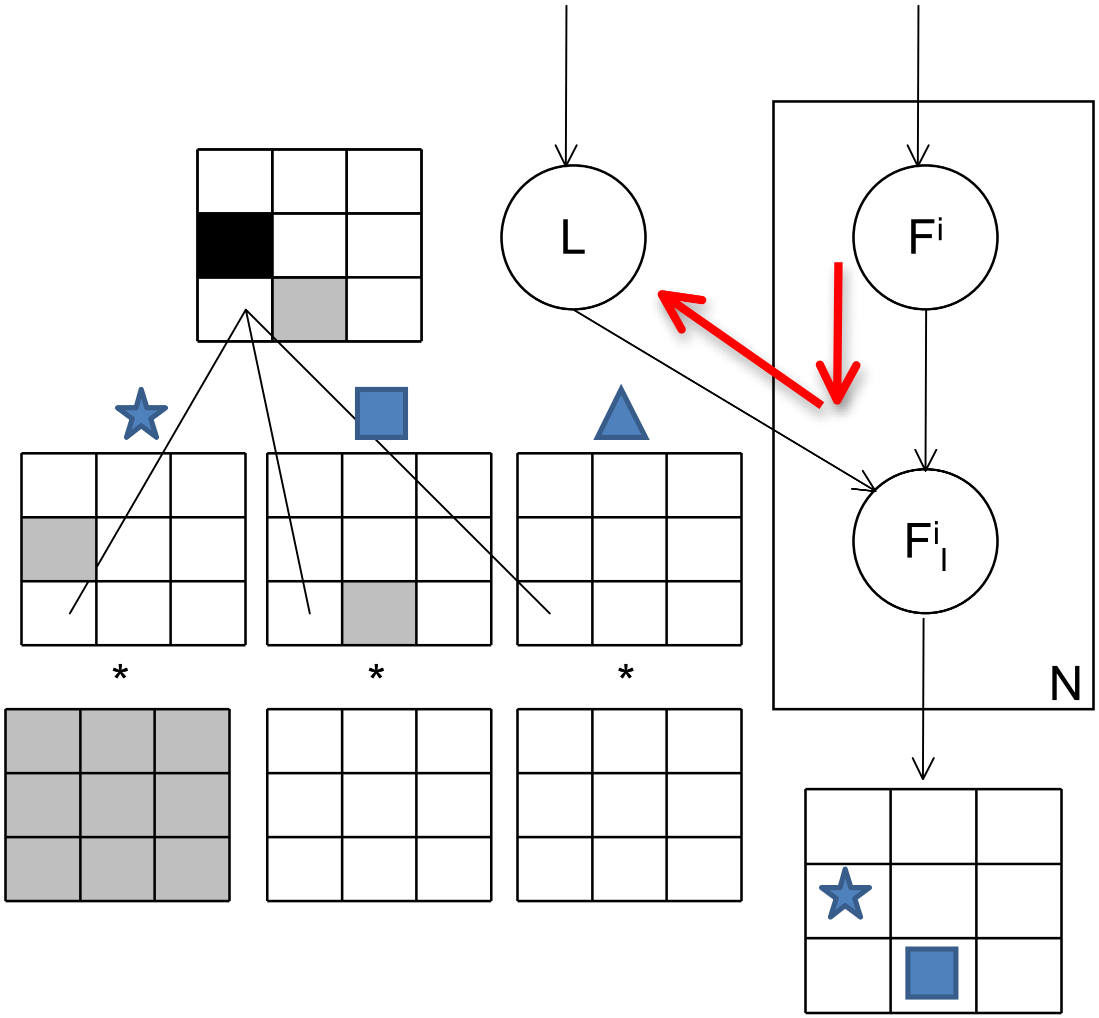
Model: spatial attention

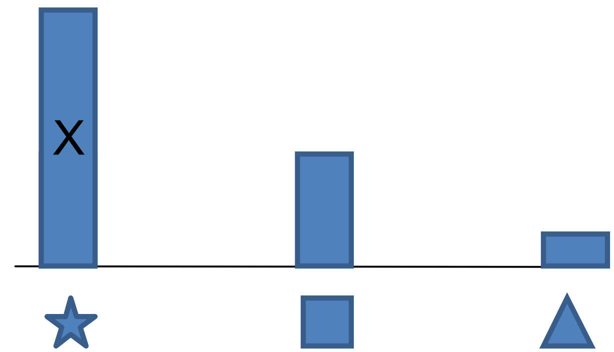


$$m_{L \to F_l^i} = \bar{P}(L)$$

- What is at location X?

Model: feature-based attention





$$m_{O \to F^i} = P(O)$$

$$m_{F^i \to F^i_l} = \sum_O P(F^i|O)P(O)$$

- Where is object X?

Model properties

Spatial Invariance

$F^1, \cdots, F^4 \sim$Orientations

$X^i \in \{0, 1, \cdots, 25\}$

$L \in \{1, 2, \cdots, 25\}$

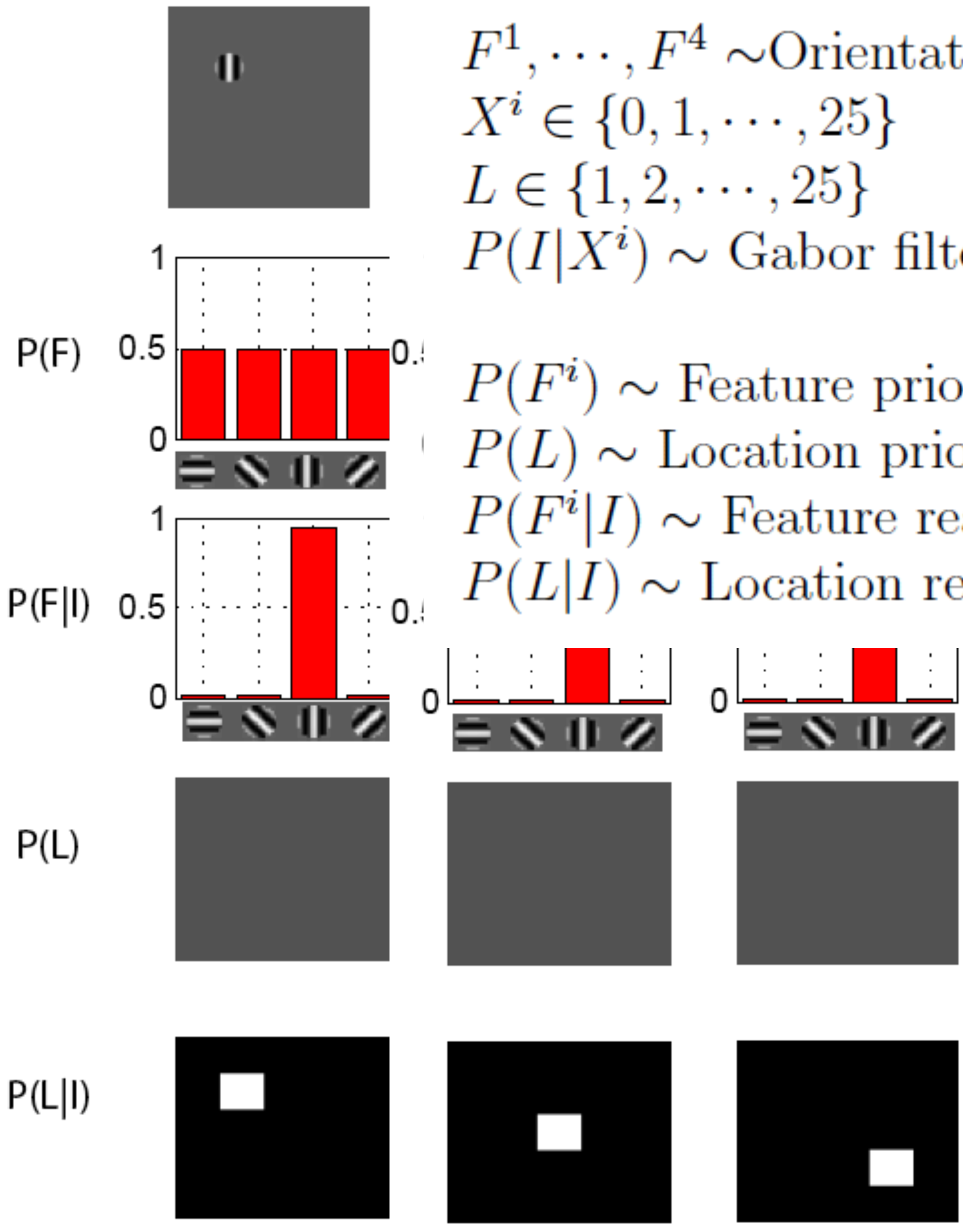
$P(I|X^i) \sim$ Gabor filter outputs

$P(F^i) \sim$ Feature prior

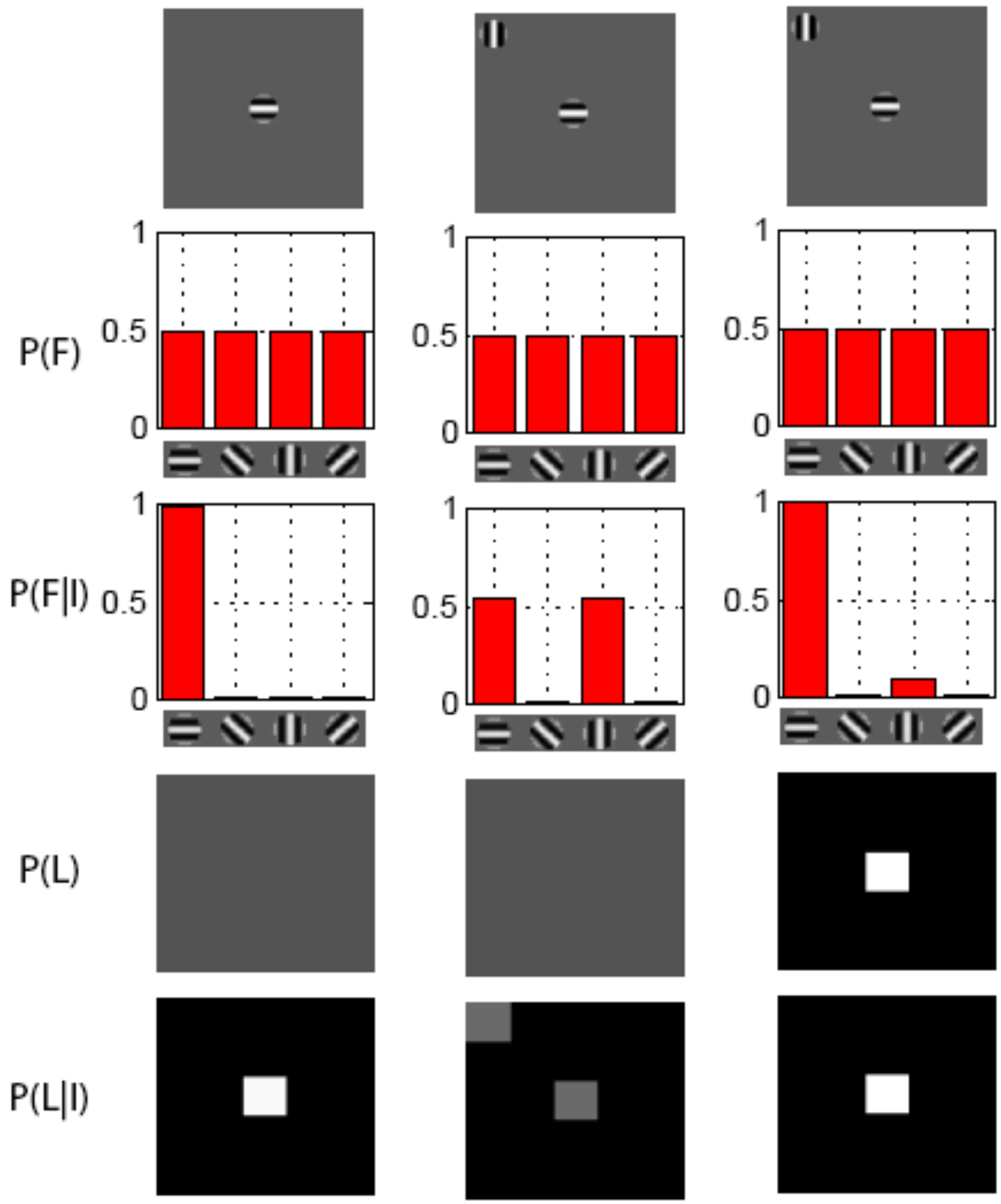
$P(L) \sim$ Location prior

$P(F^i|I) \sim$ Feature readout

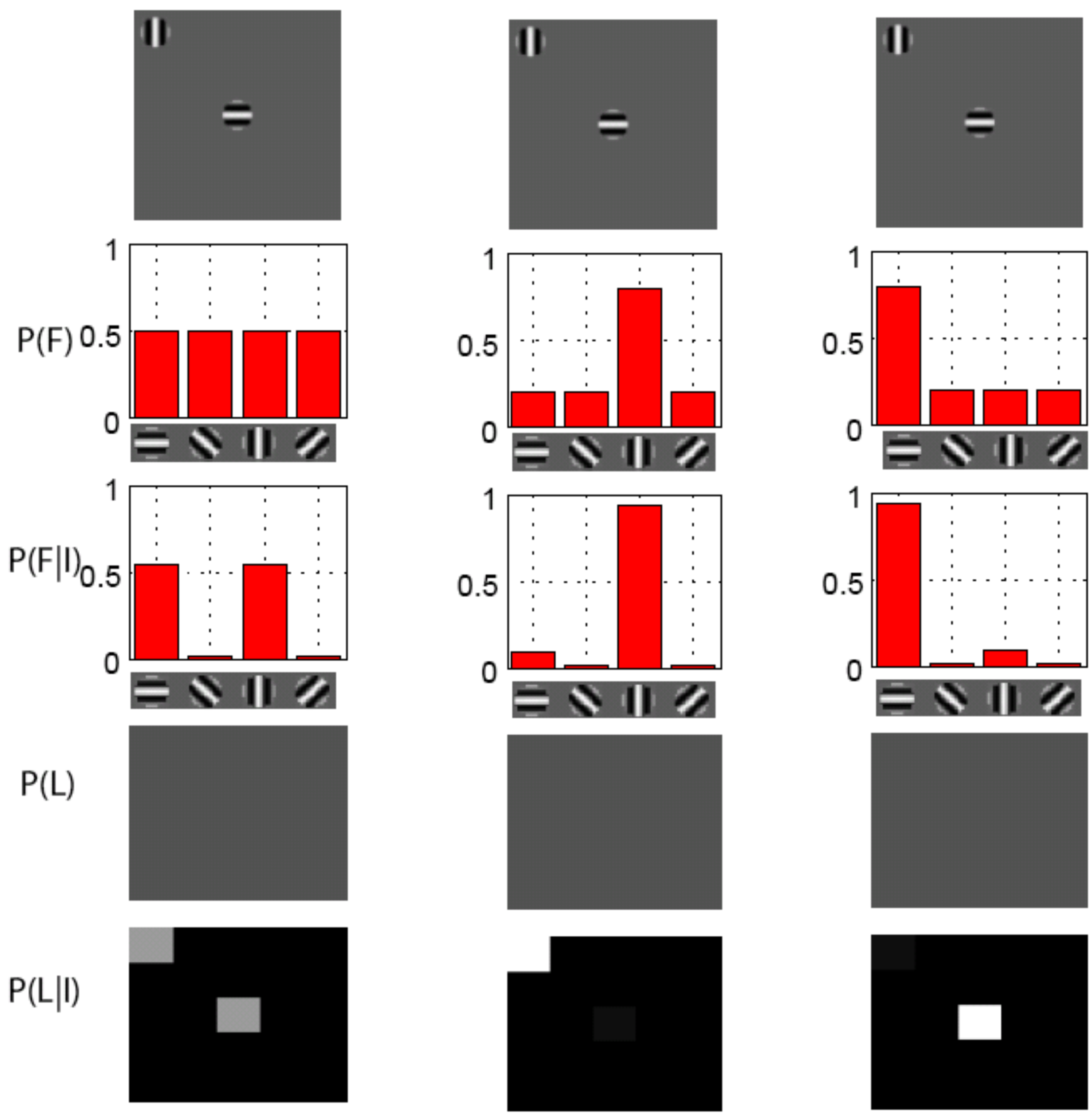
$P(L|I) \sim$ Location readout



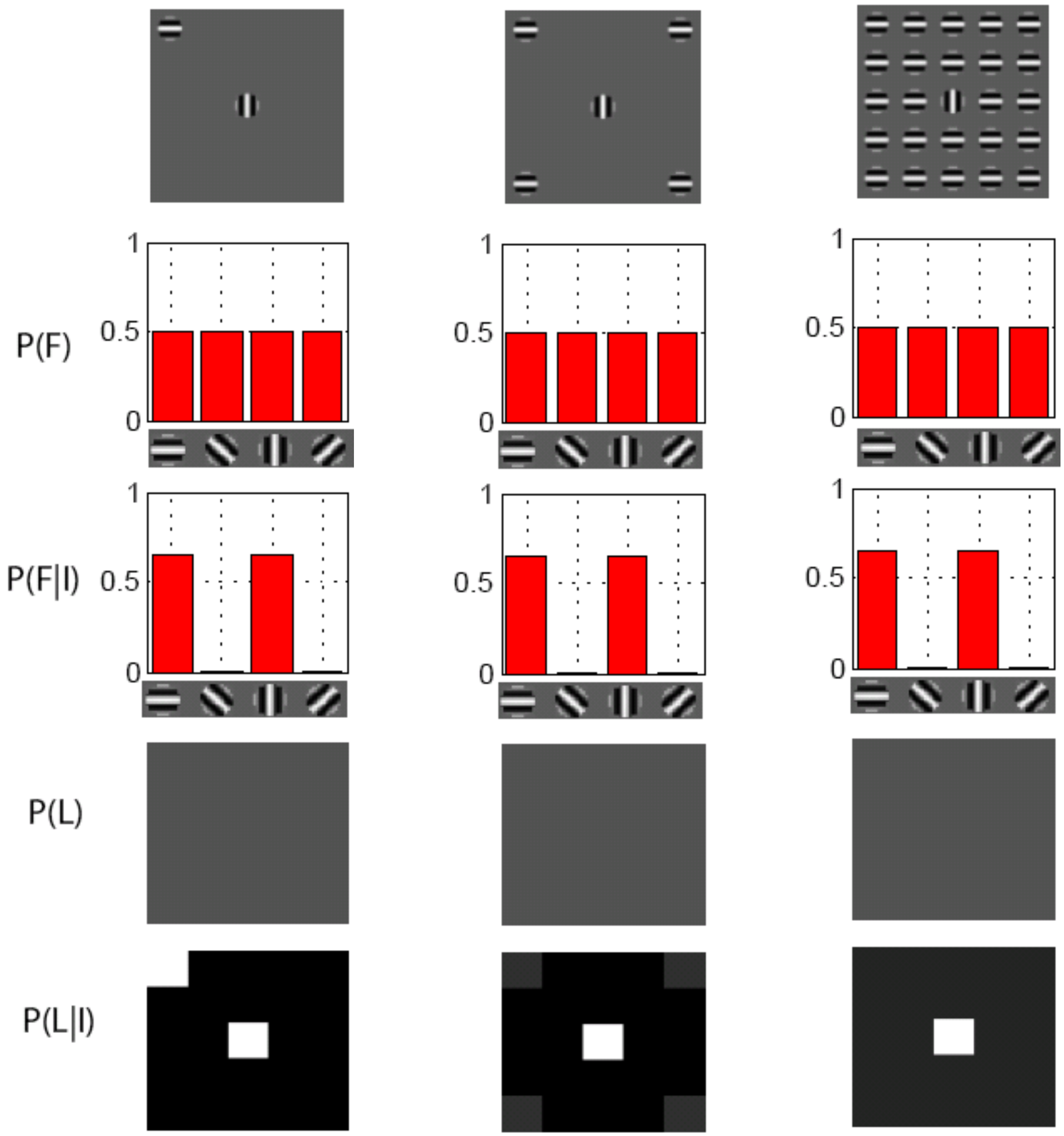
Spatial Attention



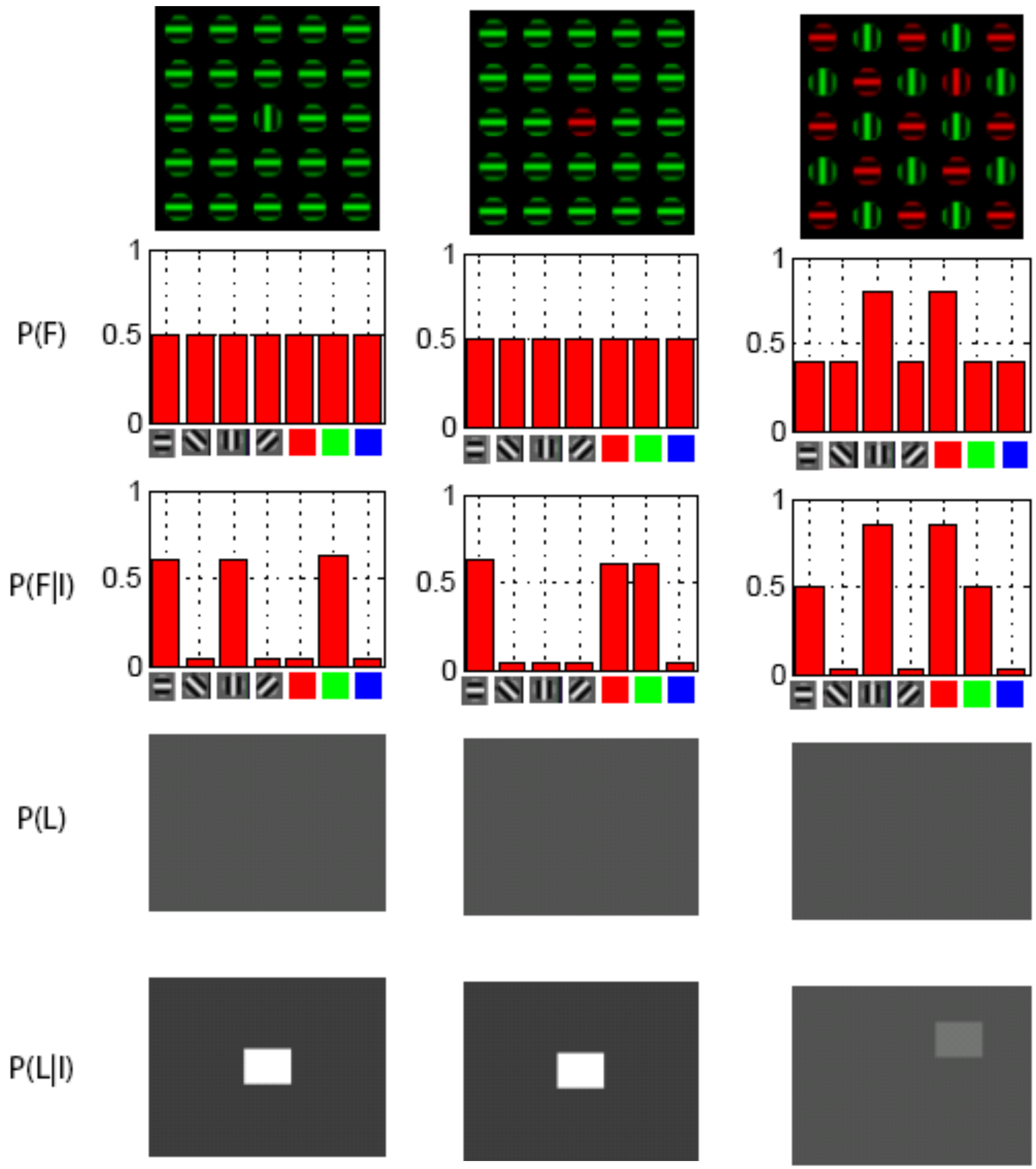
Feature Attention



Feature Popout



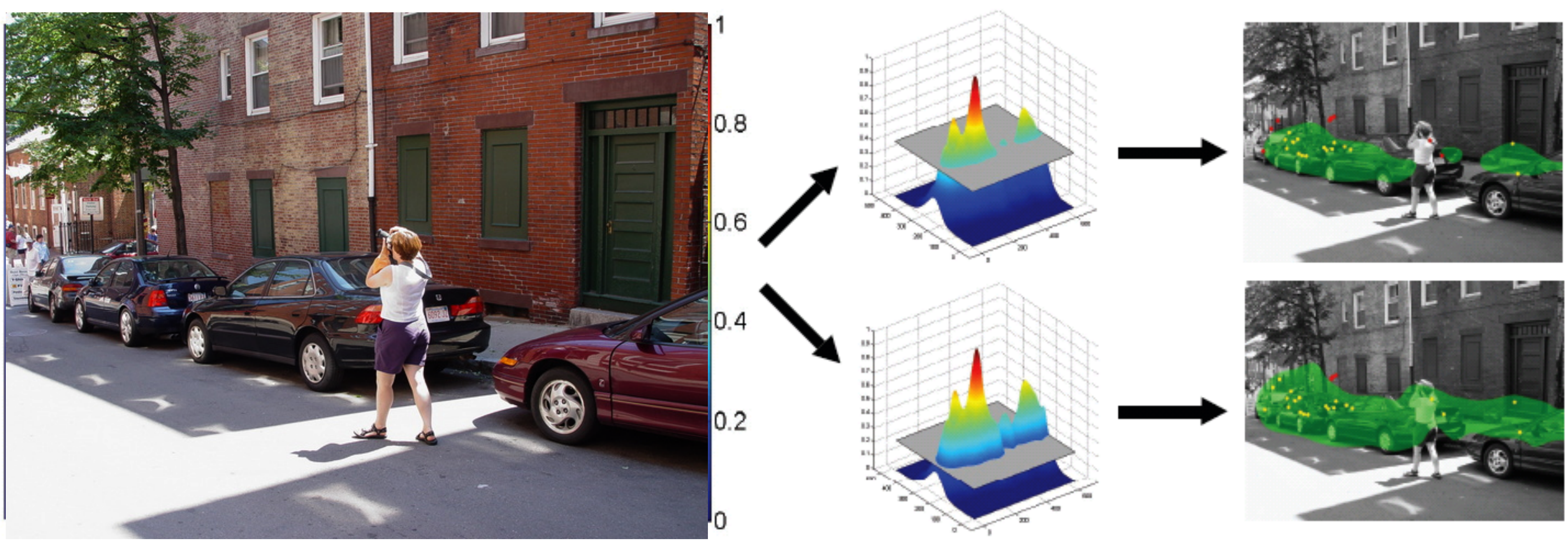
Parallel vs. Serial Search



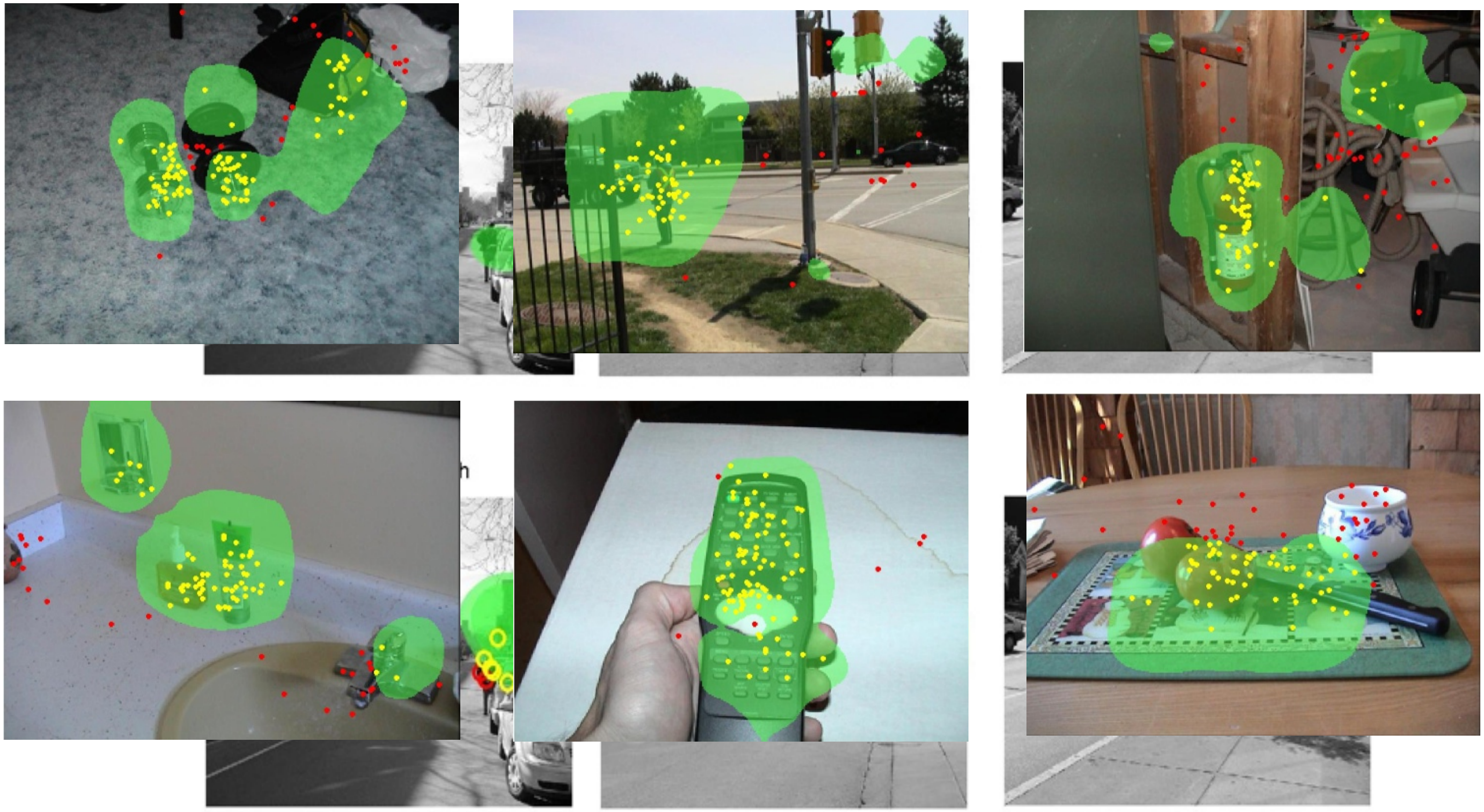
Application I: Predicting eye-movements

Predicting eye movements

- Eye movements can be considered as a proxy for attention
- Cues influencing eye-movements
  - Bottom-up image saliency
  - Top-down feature biases
  - Top-down spatial bias



Model can predict human eye-movements

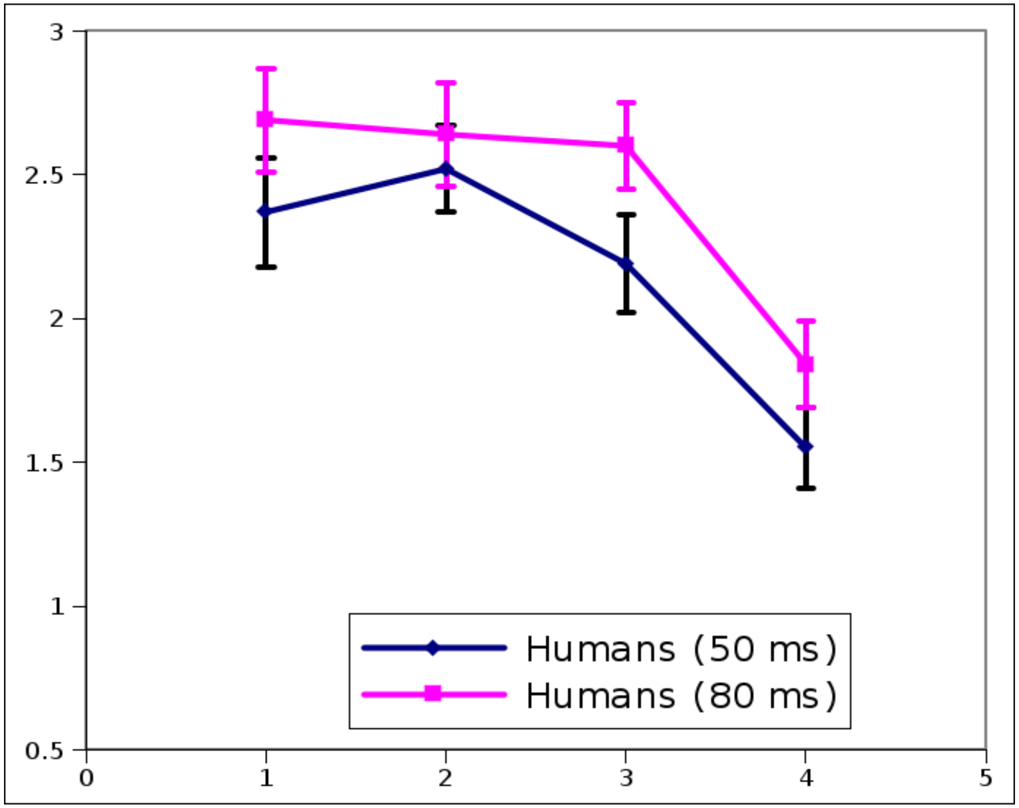


Top-down feature attention
Bottom-up spatial attention

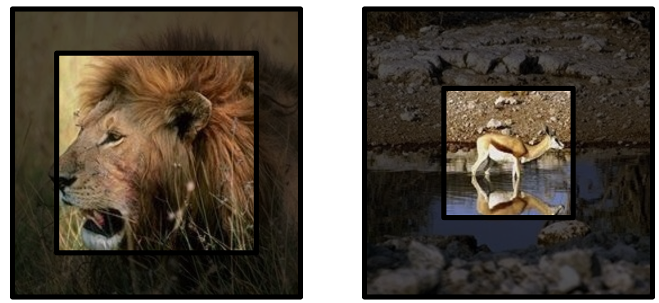
Method	ROC area (Cars)	ROC area (Pedestrian)
Itti et al. '01	42.3%	42.3%
Torralba et al.	78.9%	77.1%
Proposed	80.4%	80.1%
Humans	**87.8%**	**87.4%**

Application II: Improving recognition

Effect of clutter on detection

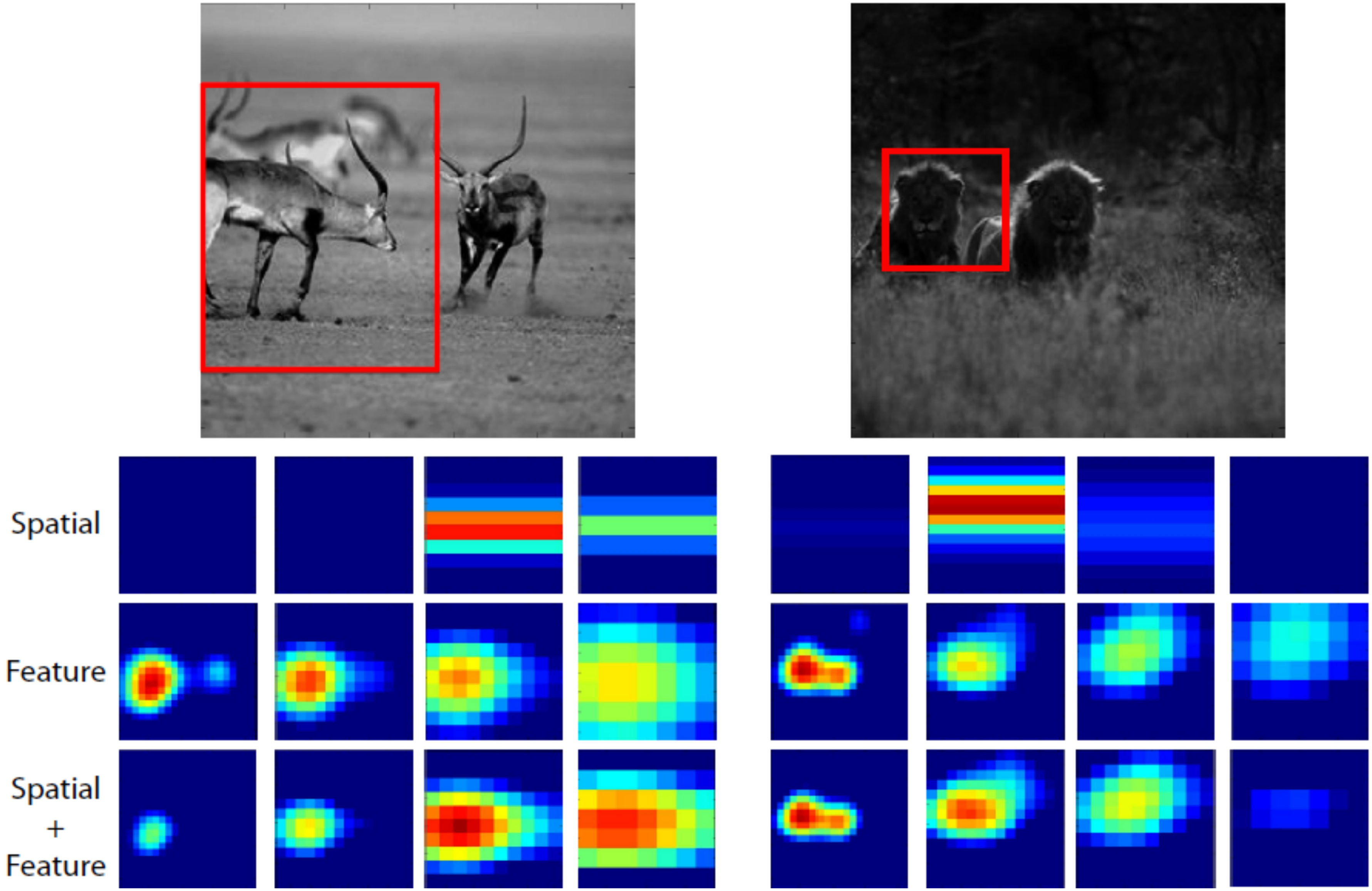


recognition without attention



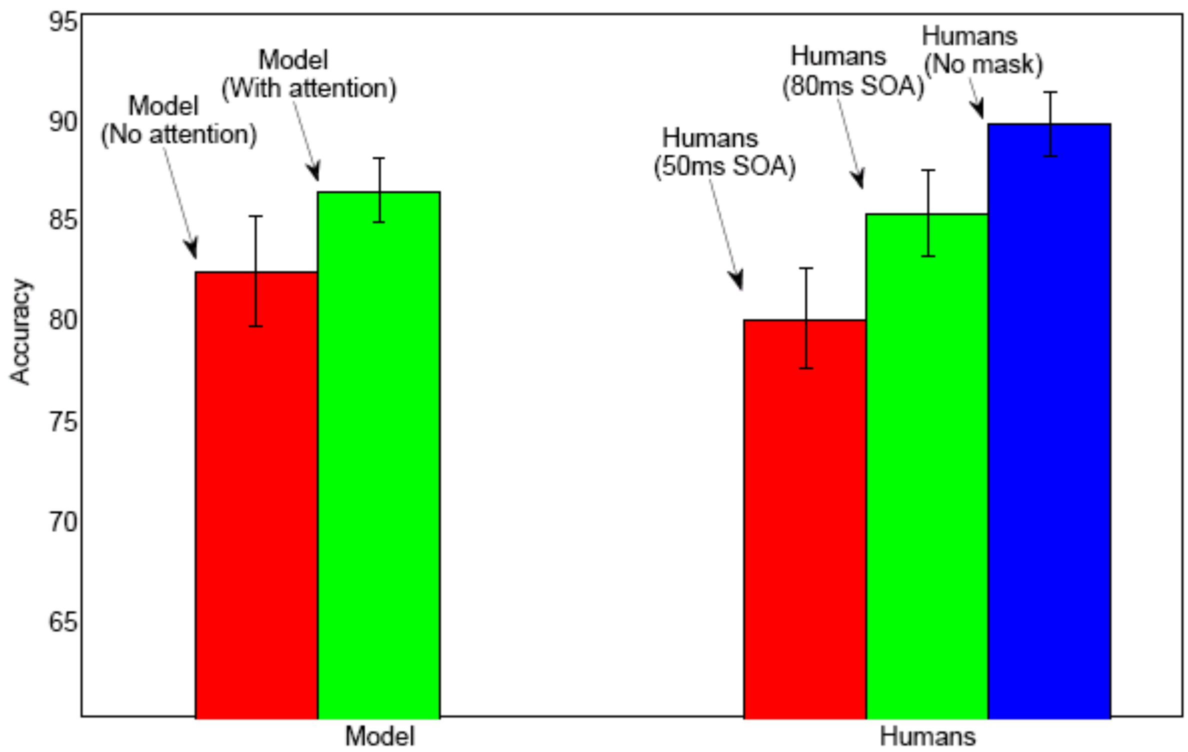
recognition under attention

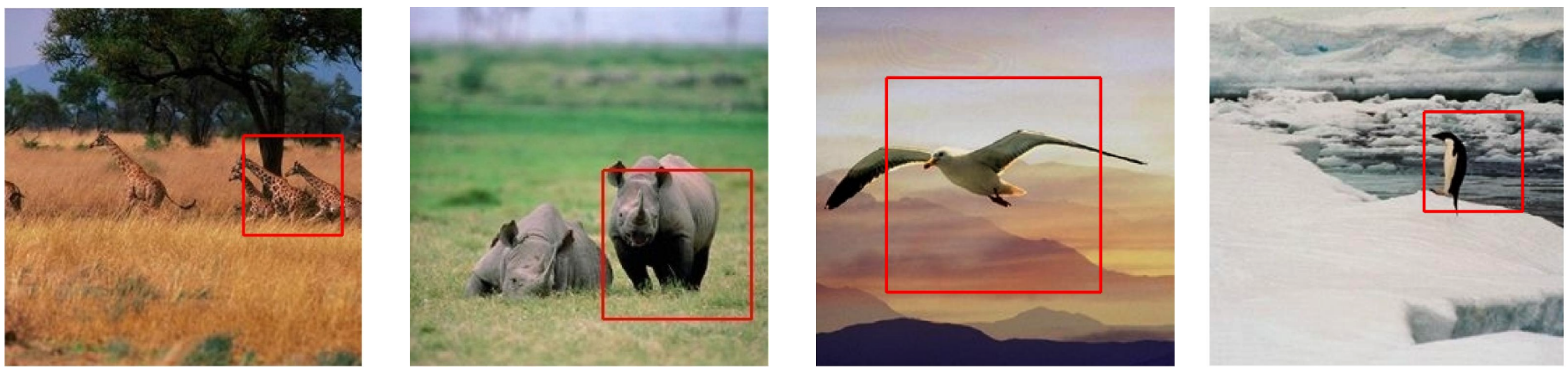
Recognition performance improves with attention



Chikkerur, Serre, Tan & Poggio (in submission, Vision Research)

Recognition performance improves with attention





Chikkerur, Serre, Tan & Poggio (in prep)

Summary

- Theory
  - Attention is part of the inference process that solves the problem of what is where.
- Computational model
  - We describe a computational model and relate it to functional anatomy of attention.
  - Attentional phenomena (pop-out, multiplicative modulation, contrast response) are ‘predicted’ by the model.

Applications

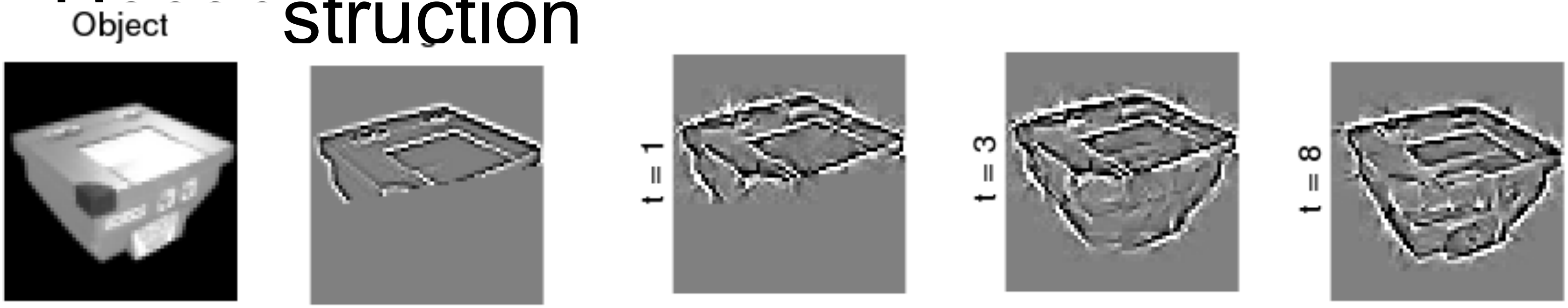
  - Predicting human eye movements.
  - Improving object recognition

Thank you!

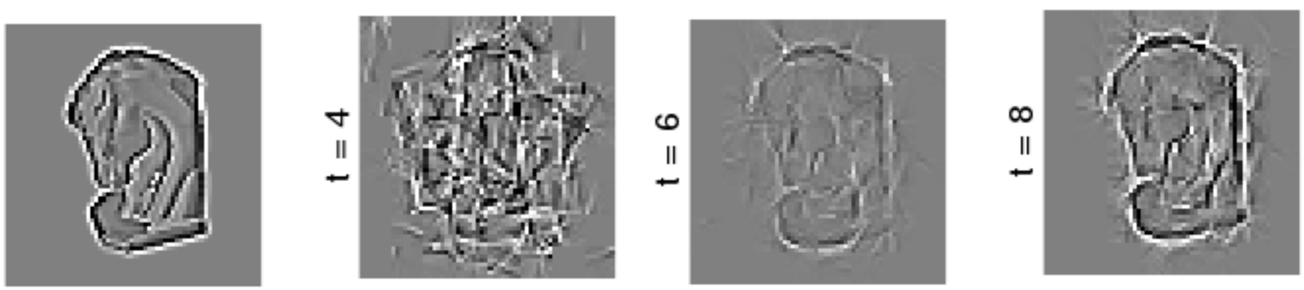
Relation to prior work

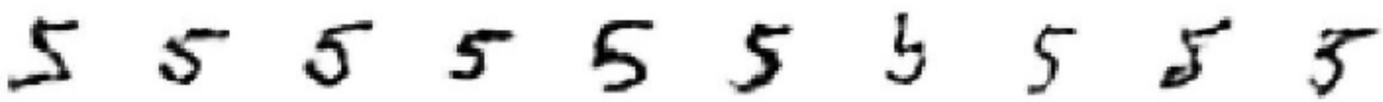
	Proposed	[Bruce and Tsotsos, 2006]	[Zhang et al., 2008]	[Deco and Rolls, 2004]	[K. et al., 2009]	[Fukushima, 1986]	[Hou and Zhang, 2007]	[Harel et al., 2007]	[Itti and Koch, 2001]	[Rao, 2005]	[Torralba, 2003]	[Walther and Koch, 2007]	[Wolfe, 2007]	[Yu and Dayan, 2005]
Biologically plausible	✓	✓	✓	✓	×	✓	×	×	✓	✓	✓	✓	✓	✓
Real world stimuli	✓	✓	✓	×	✓	×	✓	✓	✓	×	✓	✓	×	×
Pop-out	✓	✓	✓	×	✓	×	✓	✓	✓	×	✓	✓	×	×
Feature-based attention	✓	×	×	✓	✓	✓	×	×	×	×	✓	×	✓	✓
Spatial attention	✓	×	×	×	✓	×	×	×	×	✓	✓	✓	×	✓
Parallel vs. serial search	✓	×	×	×	×	×	×	×	×	×	×	×	✓	×
Explicitly models ventral/parietal	✓	×	×	✓	×	×	×	×	×	✓	×	×	×	×

Reconstruction



- **Mental Imagery**

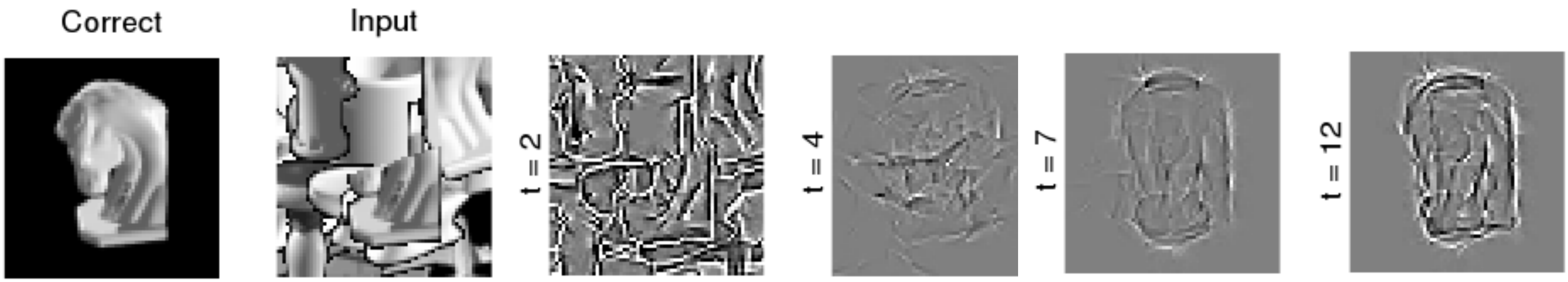


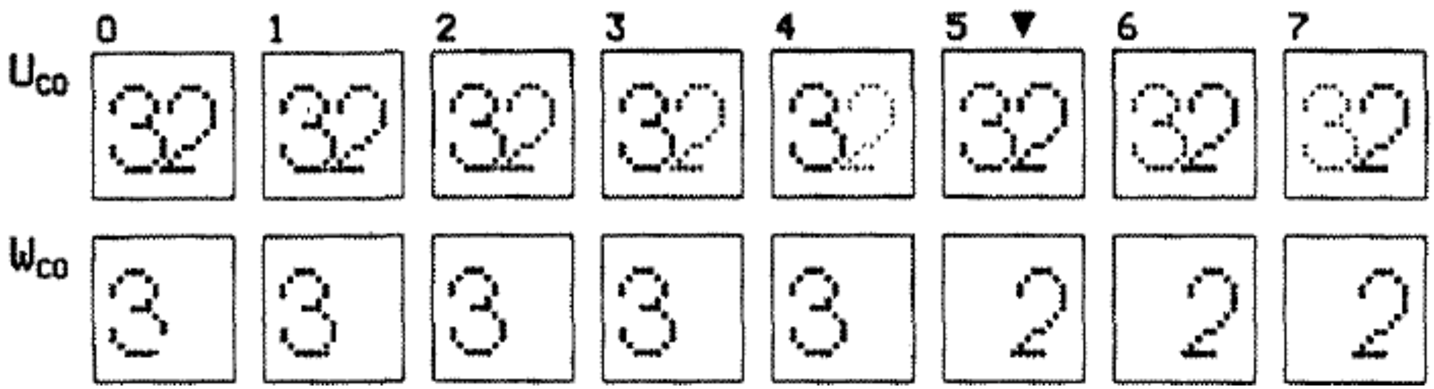


Murray J. F., Visual recognition, inference and coding using learned sparse overcomplete representations, PhD thesis, UCSD, 2005
Hinton G. E, Osindero S. and Teh. Y, A fast learning algorithm for deep belief nets, Neural computation, vol. 18, 2006

[redacted]

- **Visual attention/Segmentation**





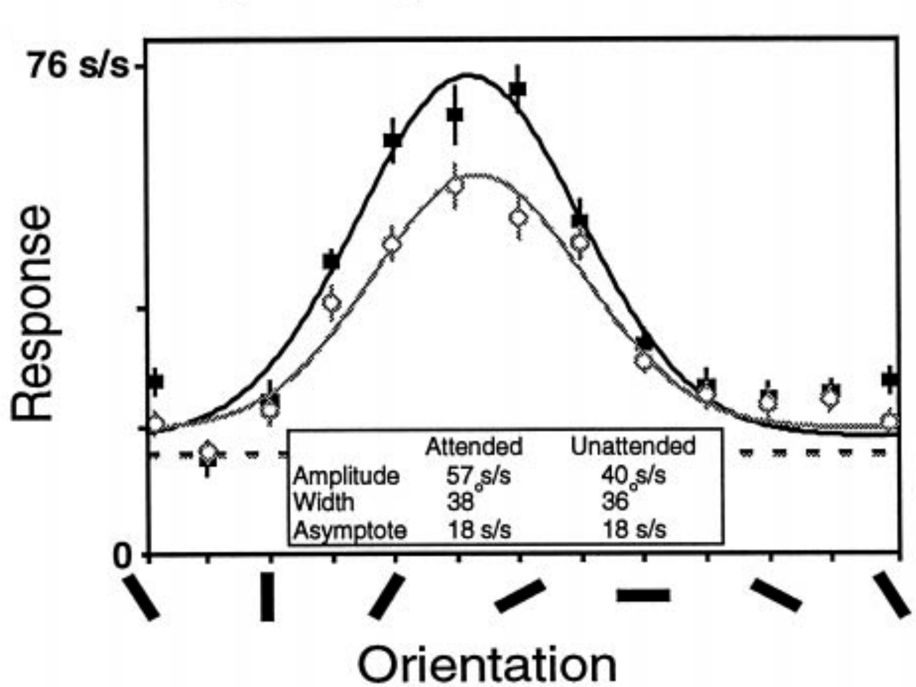
Murray J. F., Visual recognition, inference and coding using learned sparse overcomplete representations, PhD thesis, UCSD, 2005

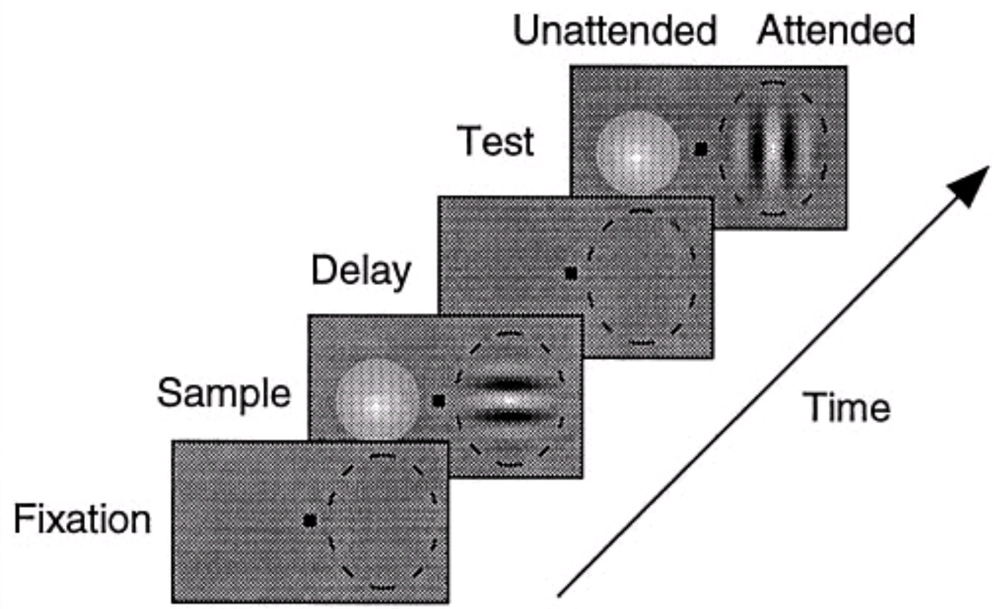
Fukushima K., A neural network model for selective attention in visual pattern recognition, Biological Cybernetics, vol. 55, 1986

‘Predicting’ physiological effects

Spatial attention

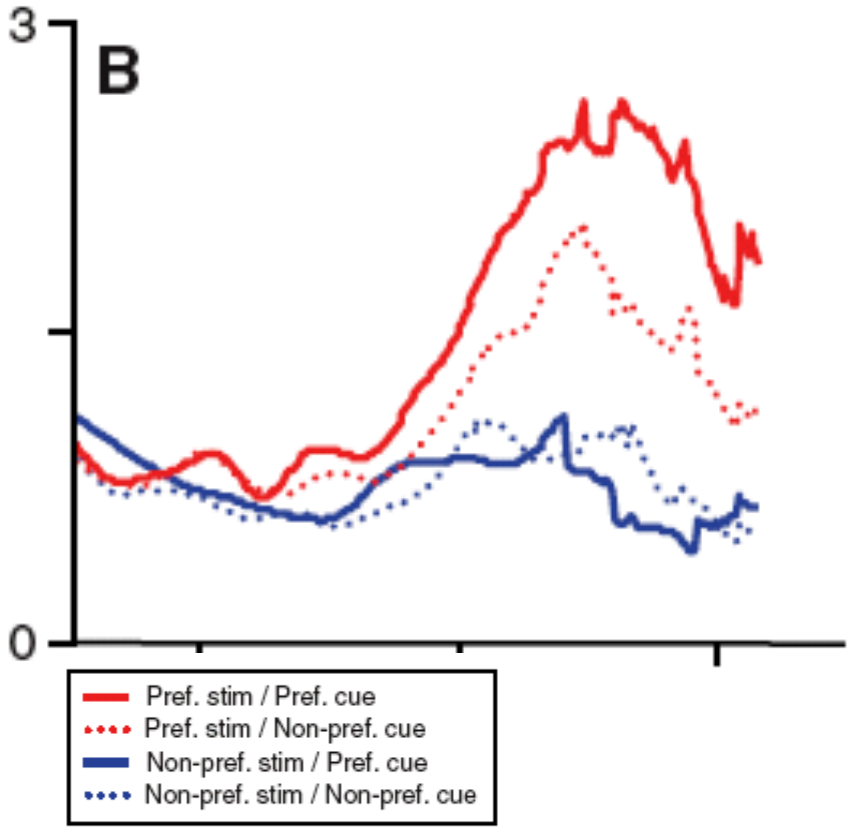
McAdams and Maunsell '99

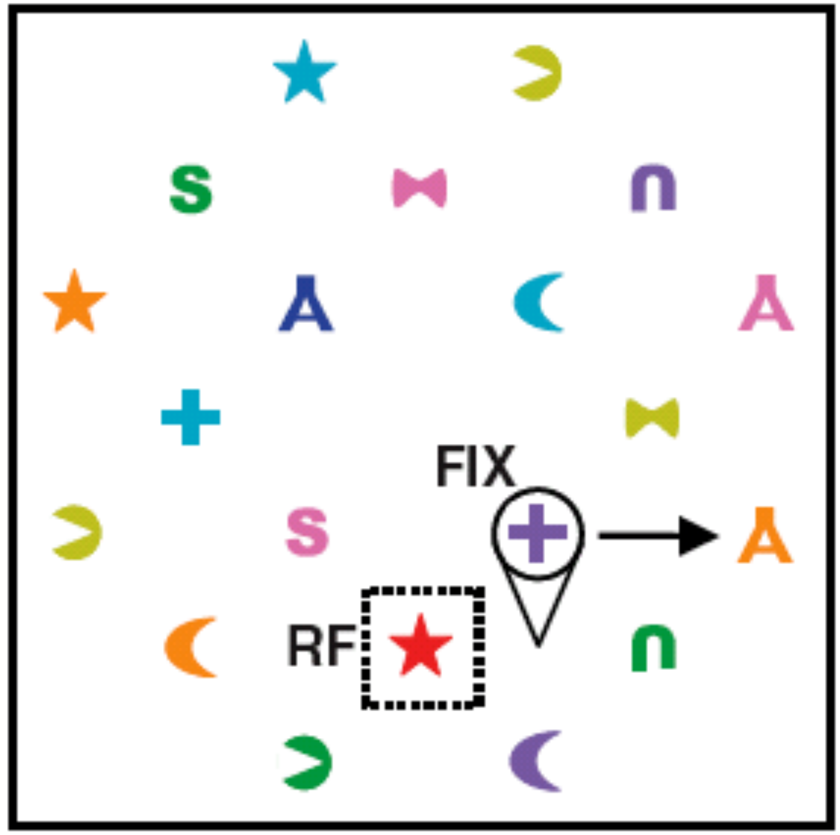




Feature-based attention

Bichot and Desimone '05

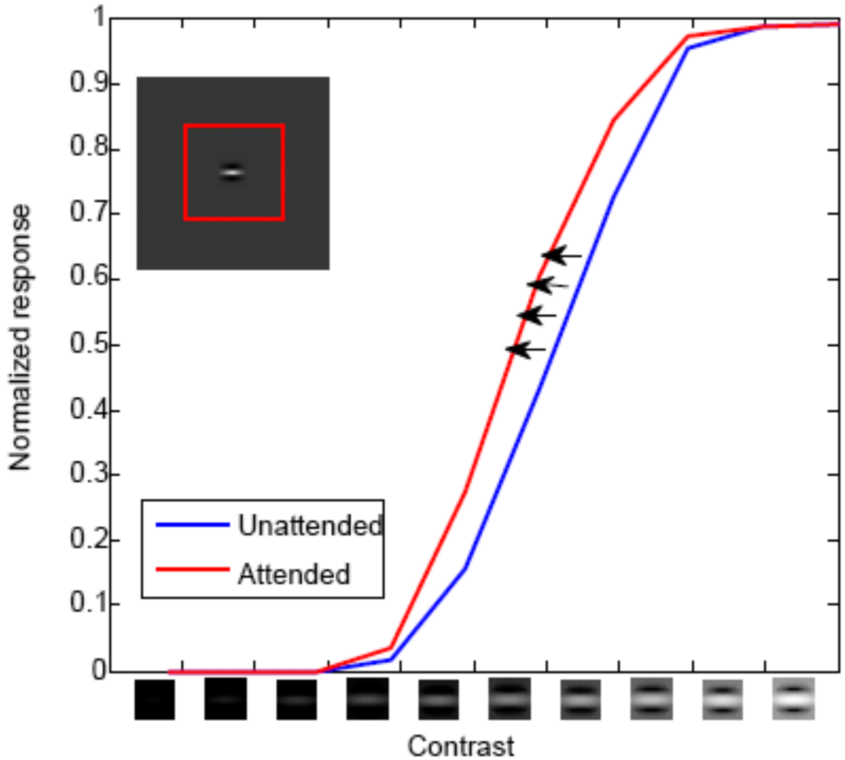




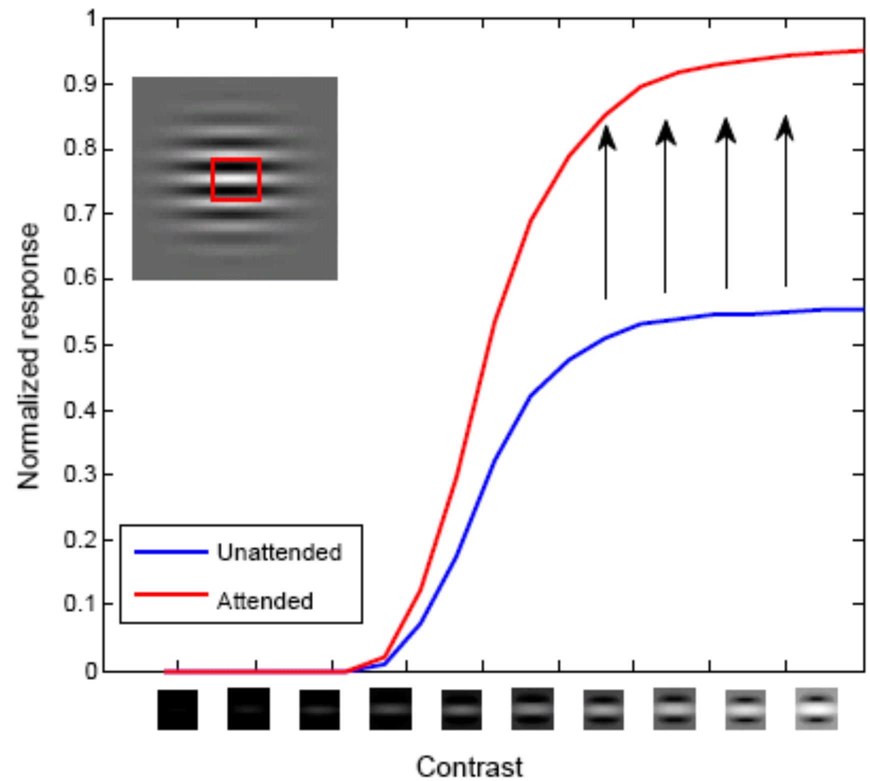
i

Contrast gain vs. Response gain

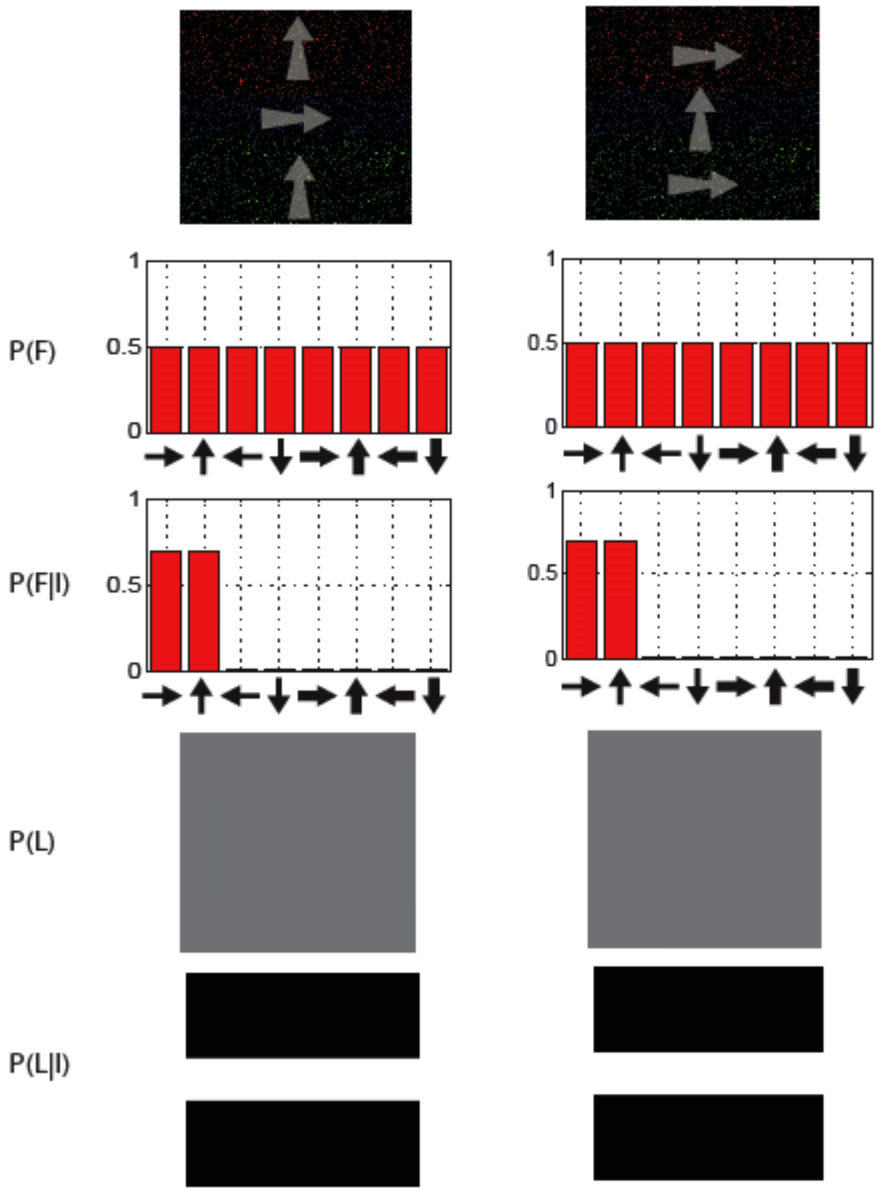
Trujillo and Treue '02



Mc Adams and Maunsell'99



Attentional effects in MT: popout



ections,2 speeds

}

odel outputs (Simoncelli

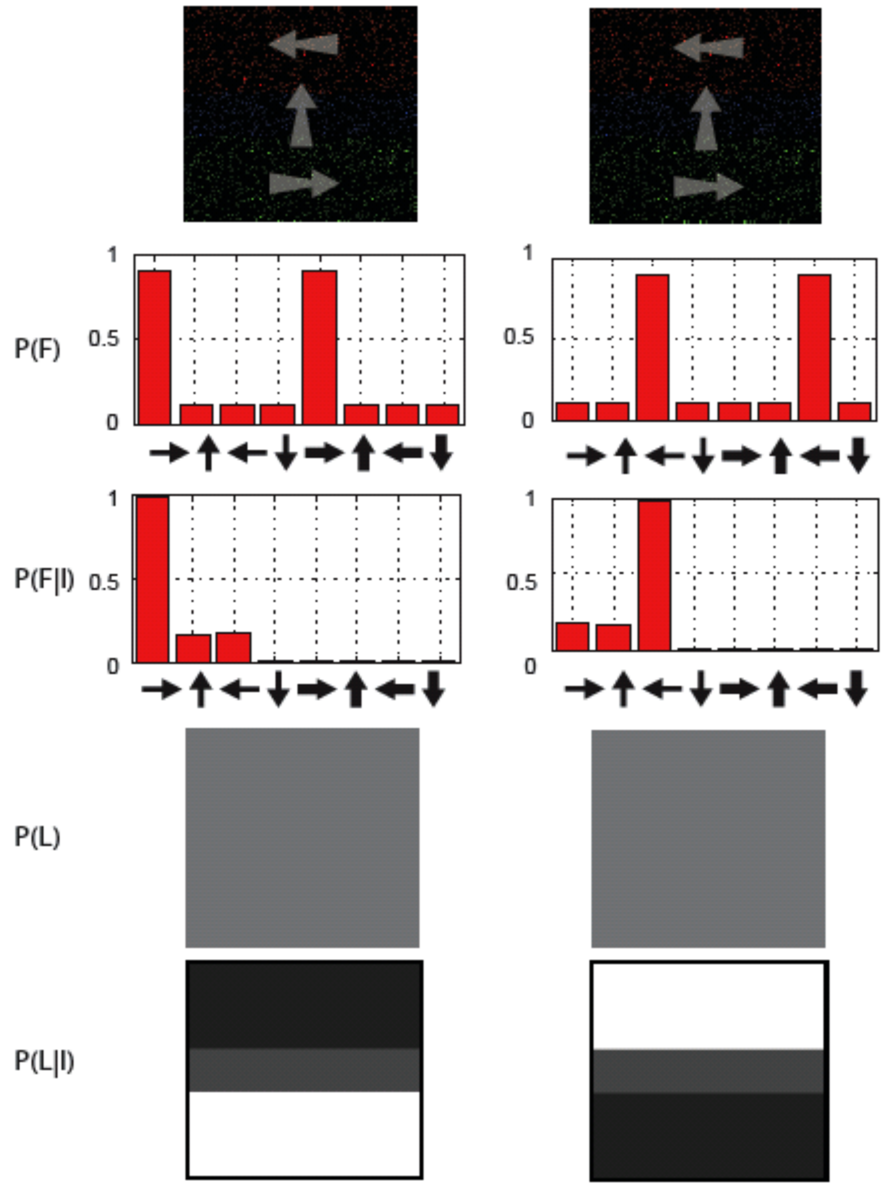
prior

prior

e readout

n readout

MT: Feature based attention



MT: Multi-modal interaction