# Using the Forest to see the Trees: A computational model relating features, objects and scenes

Antonio Torralba

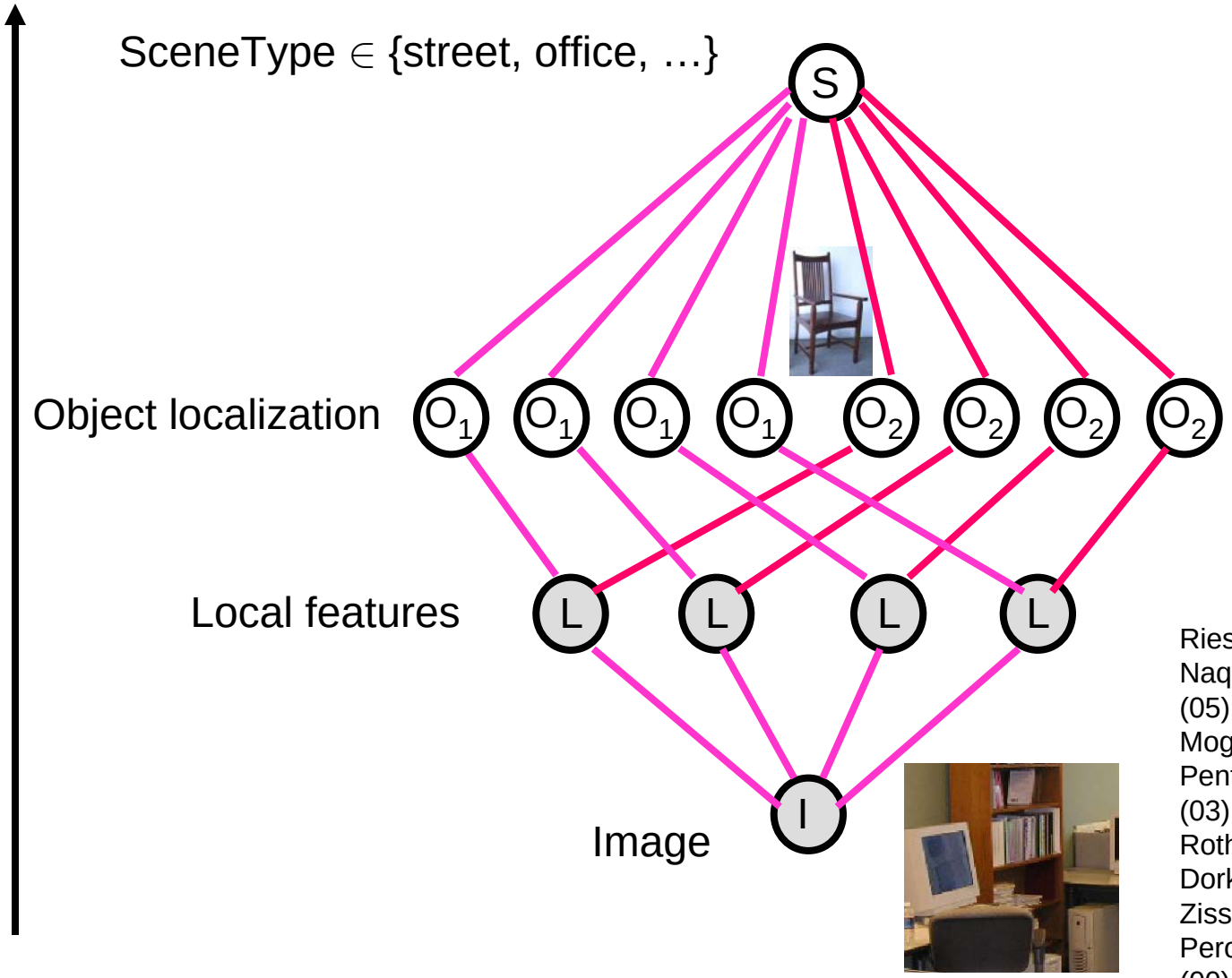
CSAIL-MIT

Joint work with

Aude Oliva, Kevin Murphy, William Freeman

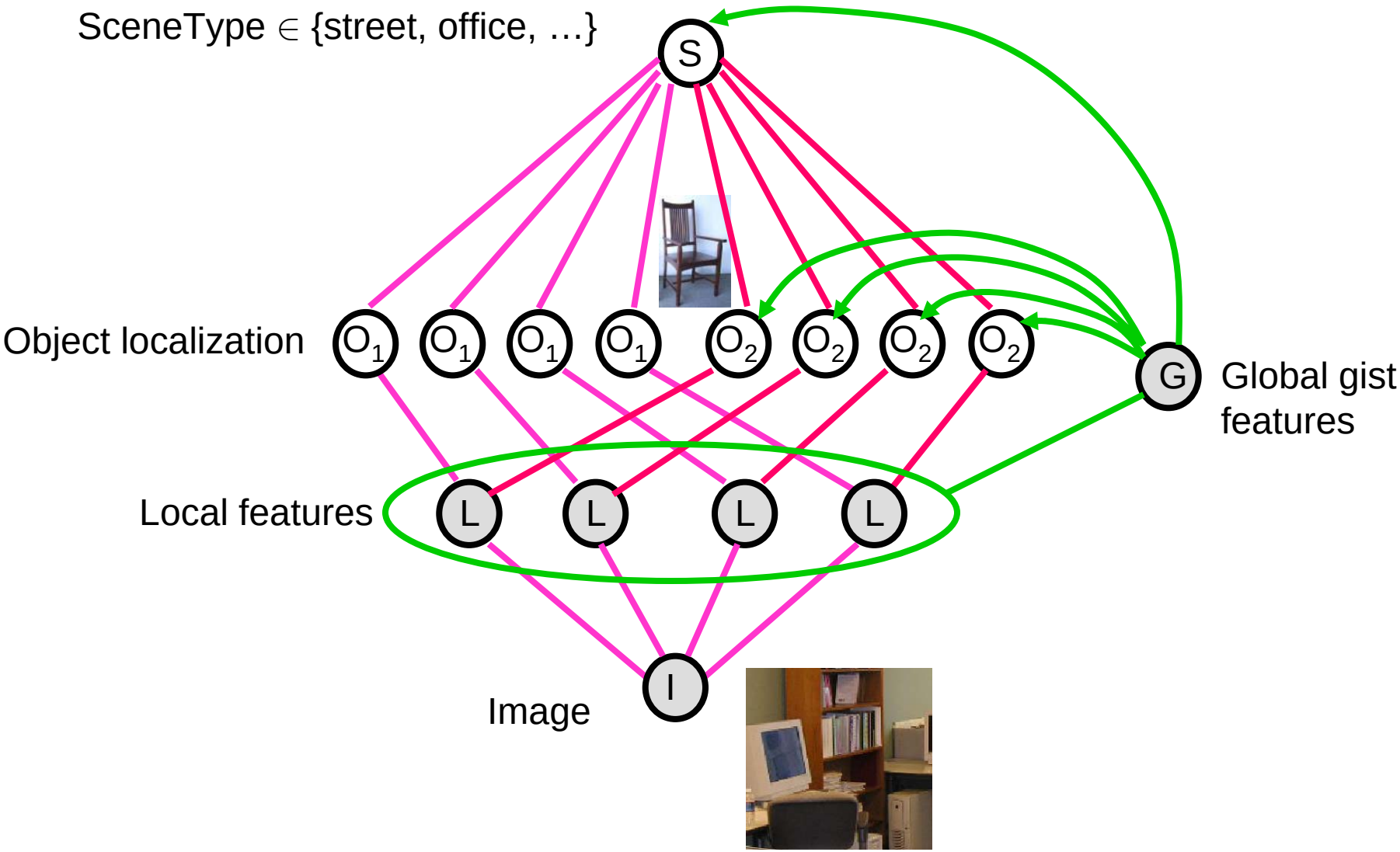
Monica Castelhano, John Henderson

# From objects to scenes

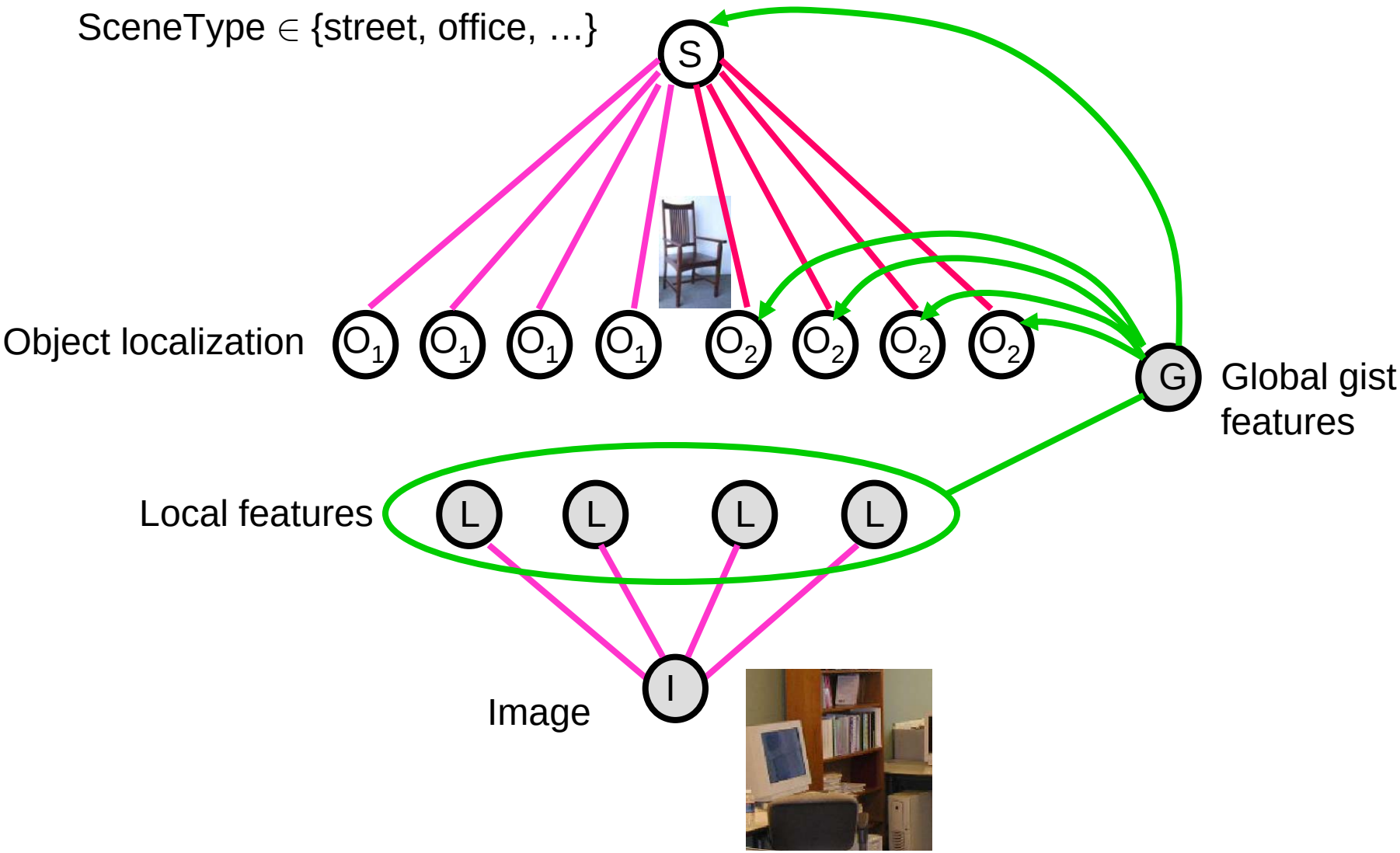



Riesenhuber & Poggio (99); Vidal-Naquet & Ullman (03); Serre & Poggio, (05); Agarwal & Roth, (02), Moghaddam, Pentland (97), Turk, Pentland (91),Vidal-Naquet, Ullman, (03) Heisele, et al, (01), Agarwal & Roth, (02), Kremp, Geman, Amit (02), Dorko, Schmid, (03) Fergus, Perona, Zisserman (03), Fei Fei, Fergus, Perona, (03), Schneiderman, Kanade (00), Lowe (99)

# From scenes to objects



# From scenes to objects



# The context challenge

What do you think are the hidden objects?


Biederman et al 82; Bar & Ullman 93; Palmer, 75;

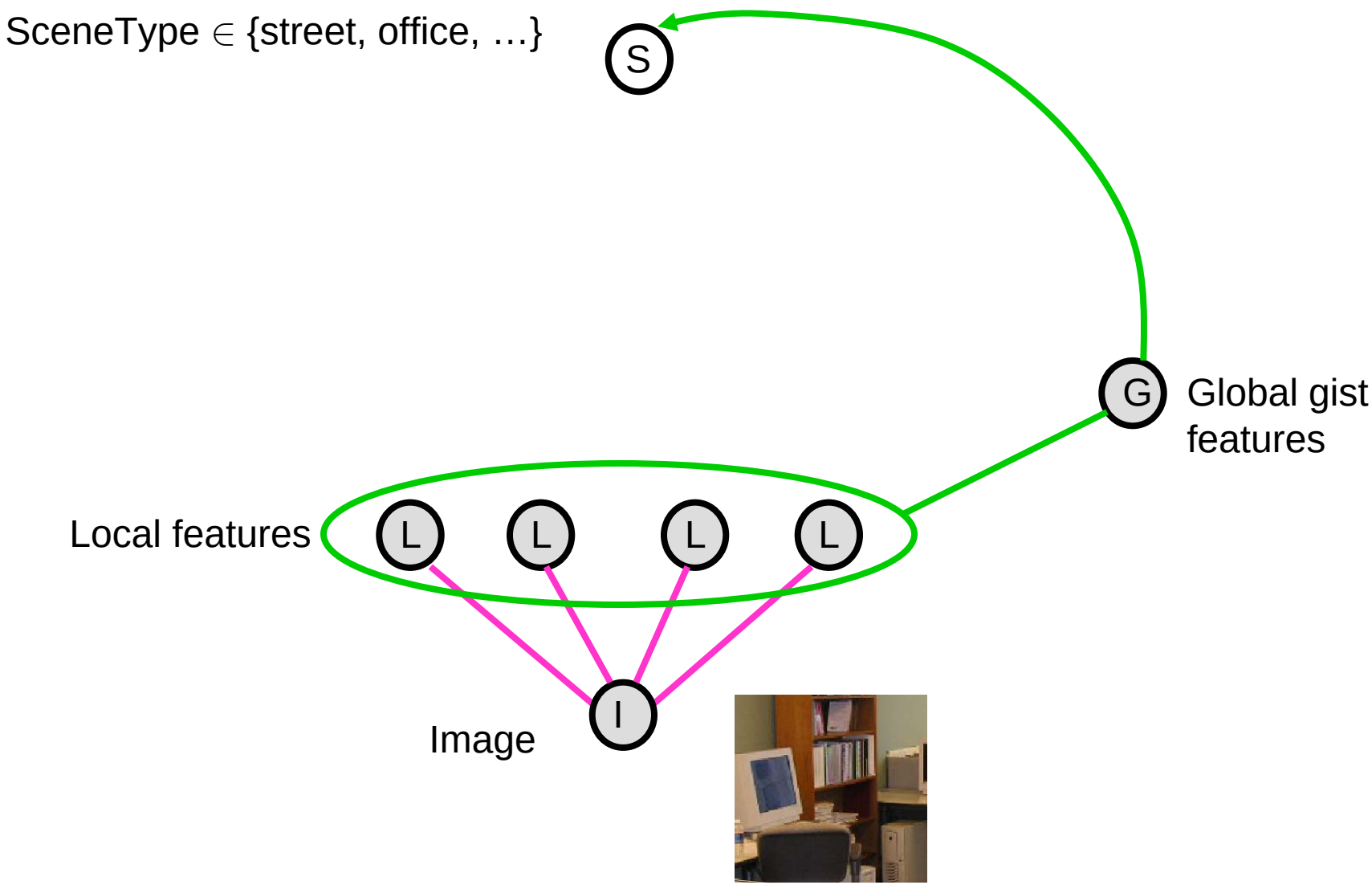
# The context challenge

What do you think are the hidden objects?

Chance ~ 1/30000

Answering this question does not require knowing how the objects look like. It is all about context.

# From scenes to objects



# Scene categorization

Office

Corridor

Street

Oliva & Torralba, IJCV'01; Torralba, Murphy, Freeman, Mark, CVPR 03.

# Place identification

Office 610

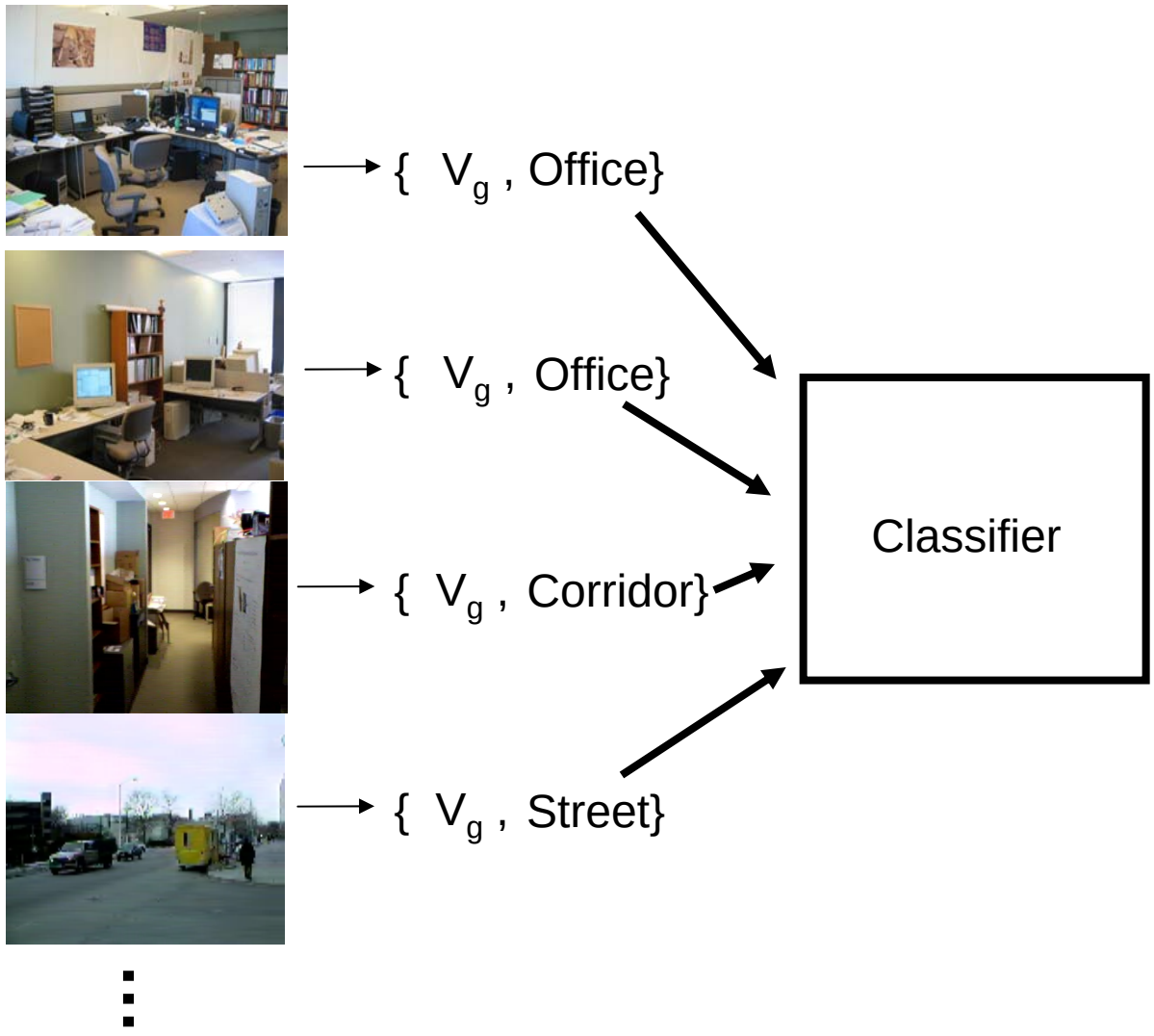
Office 615

Draper street

59 other places…

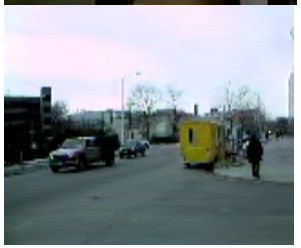
Scenes are categories, places are instances

# Supervised learning



# Supervised learning



$\rightarrow$ { $V_g$ , Office}

$\rightarrow$ { $V_g$ , Office}

$\rightarrow$ { $V_g$ , Corridor}

$\rightarrow$ { $V_g$ , Street}

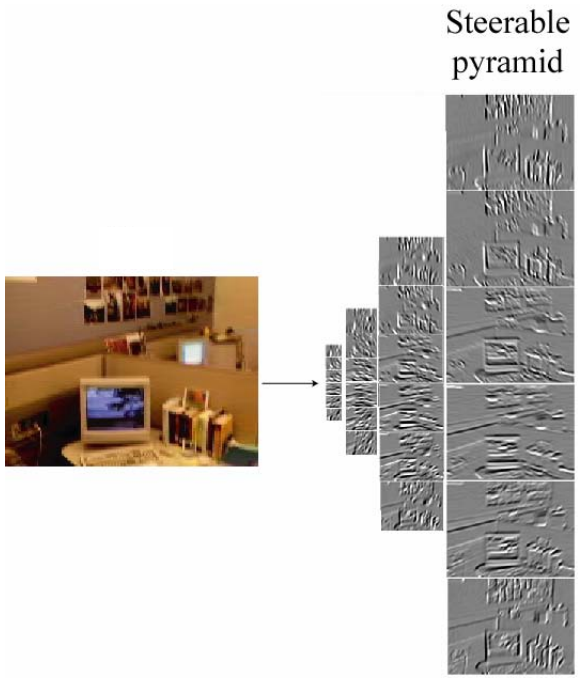
Classifier

Which feature vector for a whole image?

⋮

# Global features (gist)

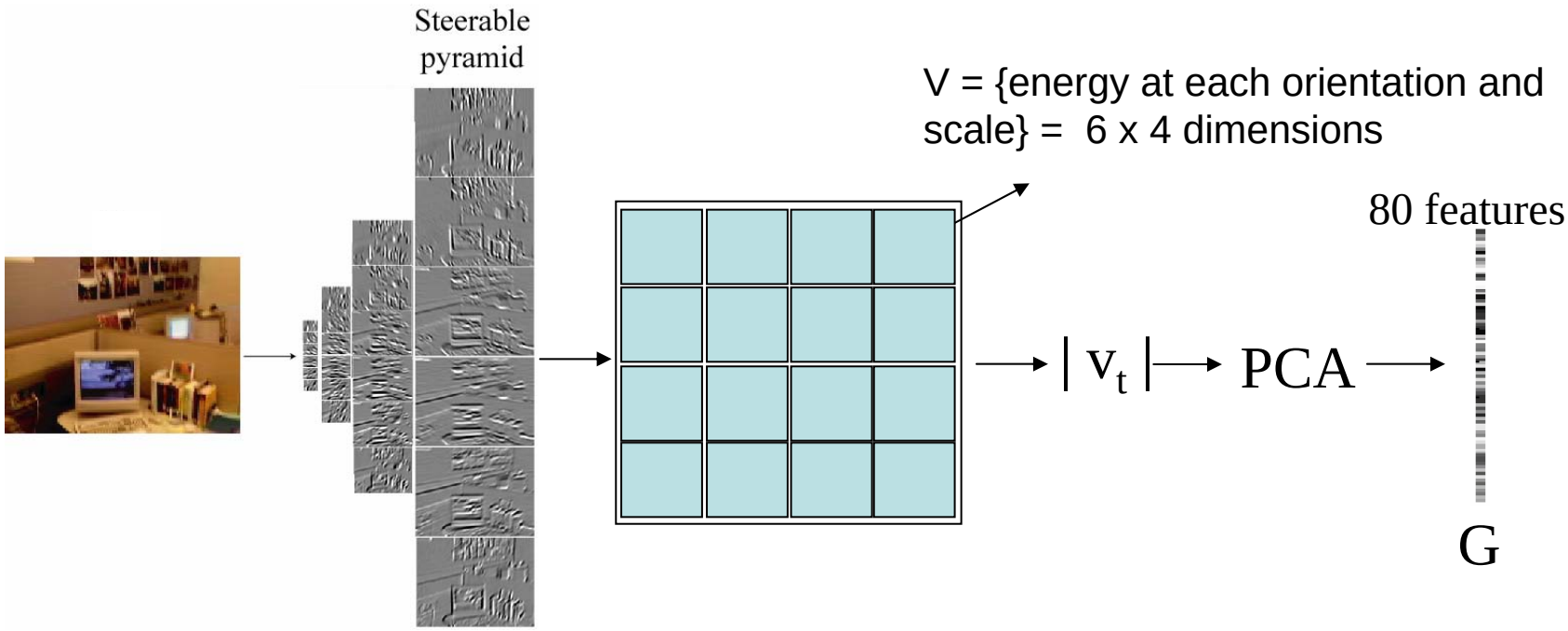
First, we propose a set of features that do not encode specific object information




Oliva & Torralba, IJCV'01; Torralba, Murphy, Freeman, Mark, CVPR 03.

# Global features (gist)

First, we propose a set of features that do not encode specific object information




Oliva & Torralba, IJCV'01; Torralba, Murphy, Freeman, Mark, CVPR 03.

# Example visual gists

I

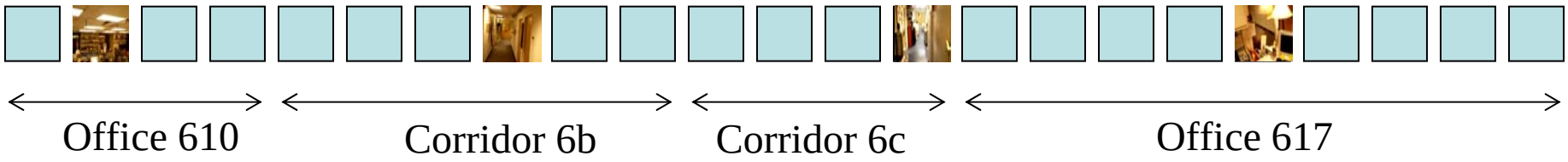
I’

Global features (I) ~ global features (I’)

Cf. “Pyramid Based Texture Analysis/Synthesis”, Heeger and Bergen, *Siggraph*, 1995

# Learning to recognize places

We use annotated sequences for training




- Hidden states = location (63 values)
- Observations = $v^G_t$ (80 dimensions)
- Transition matrix encodes topology of environment
- Observation model is a mixture of Gaussians centered on prototypes (100 views per place)

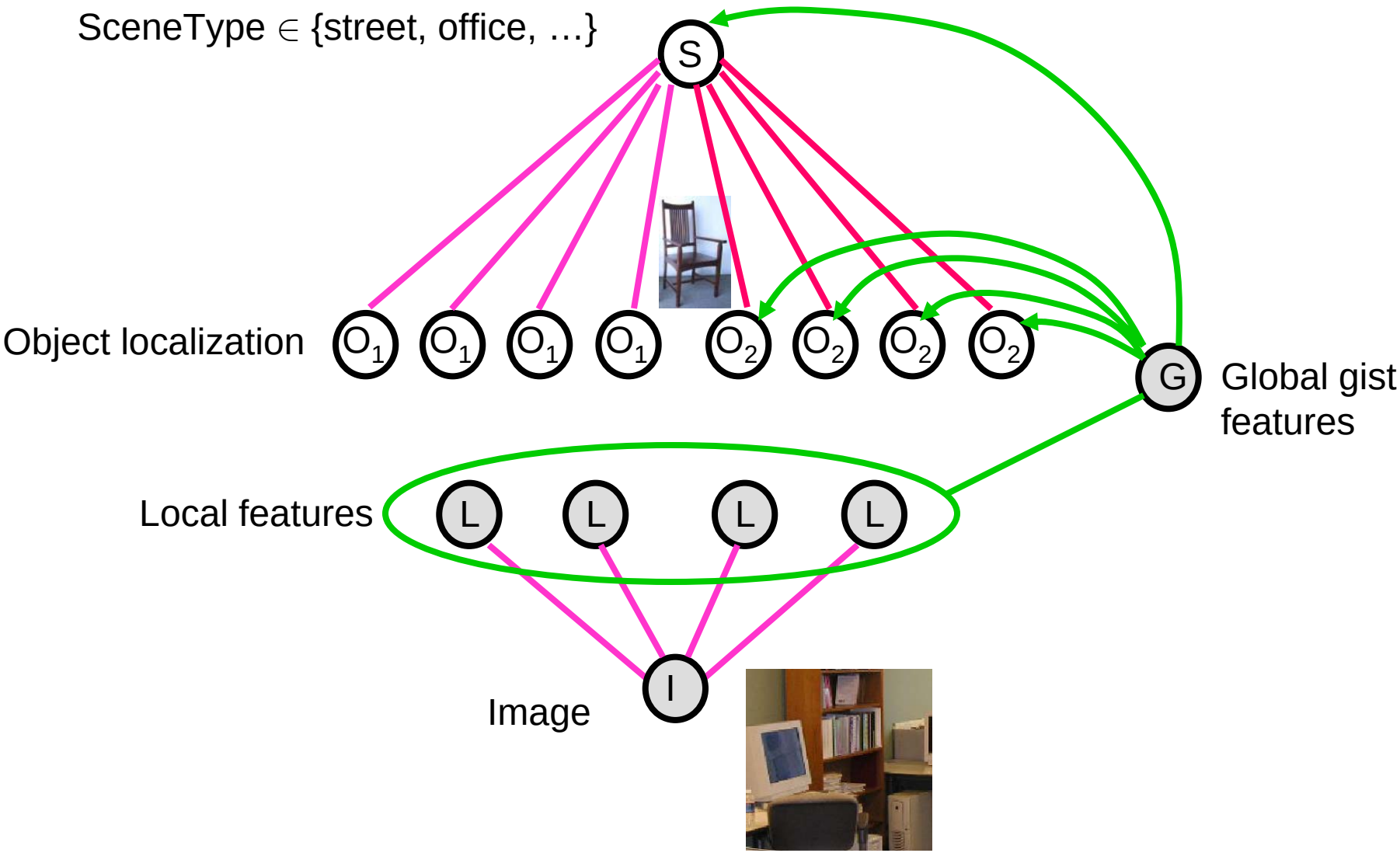
# Wearable test-bed v1

# Wearable test-bed v2

# Place/scene recognition demo

t=930, truth = 400-fl6-visionArea1

# From scenes to objects



# Global scene features predicts object location

New image

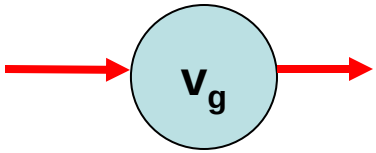



Image regions likely to contain the target

# Global scene features predicts object location

**Training set (cars)**

$\rightarrow \{v_g^1, x^1\}$

$\rightarrow \{v_g^2, x^2\}$

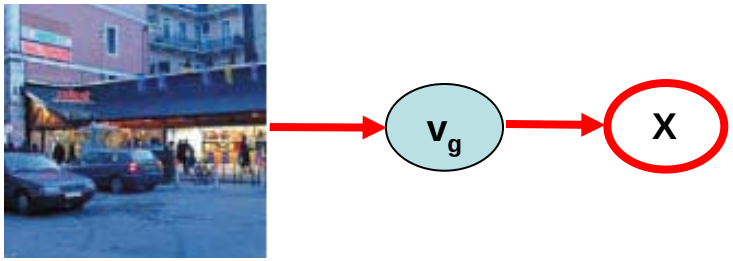
$\rightarrow \{v_g^3, x^3\}$

$\rightarrow \{v_g^4, x^4\}$

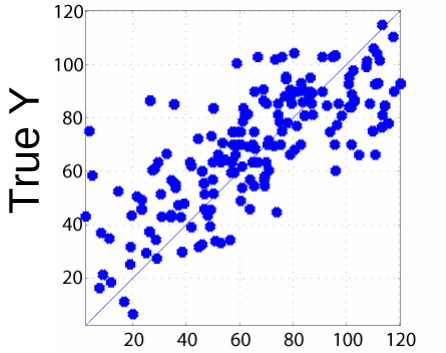
⋮

The goal of the training is to learn the association between the location of the target and the global scene features

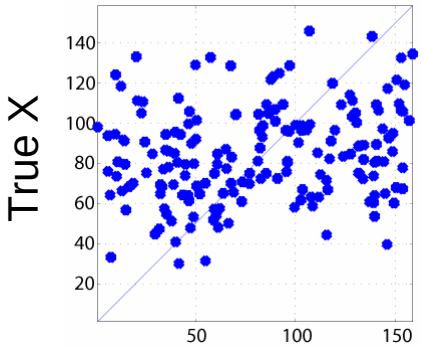
# Global scene features predicts object location



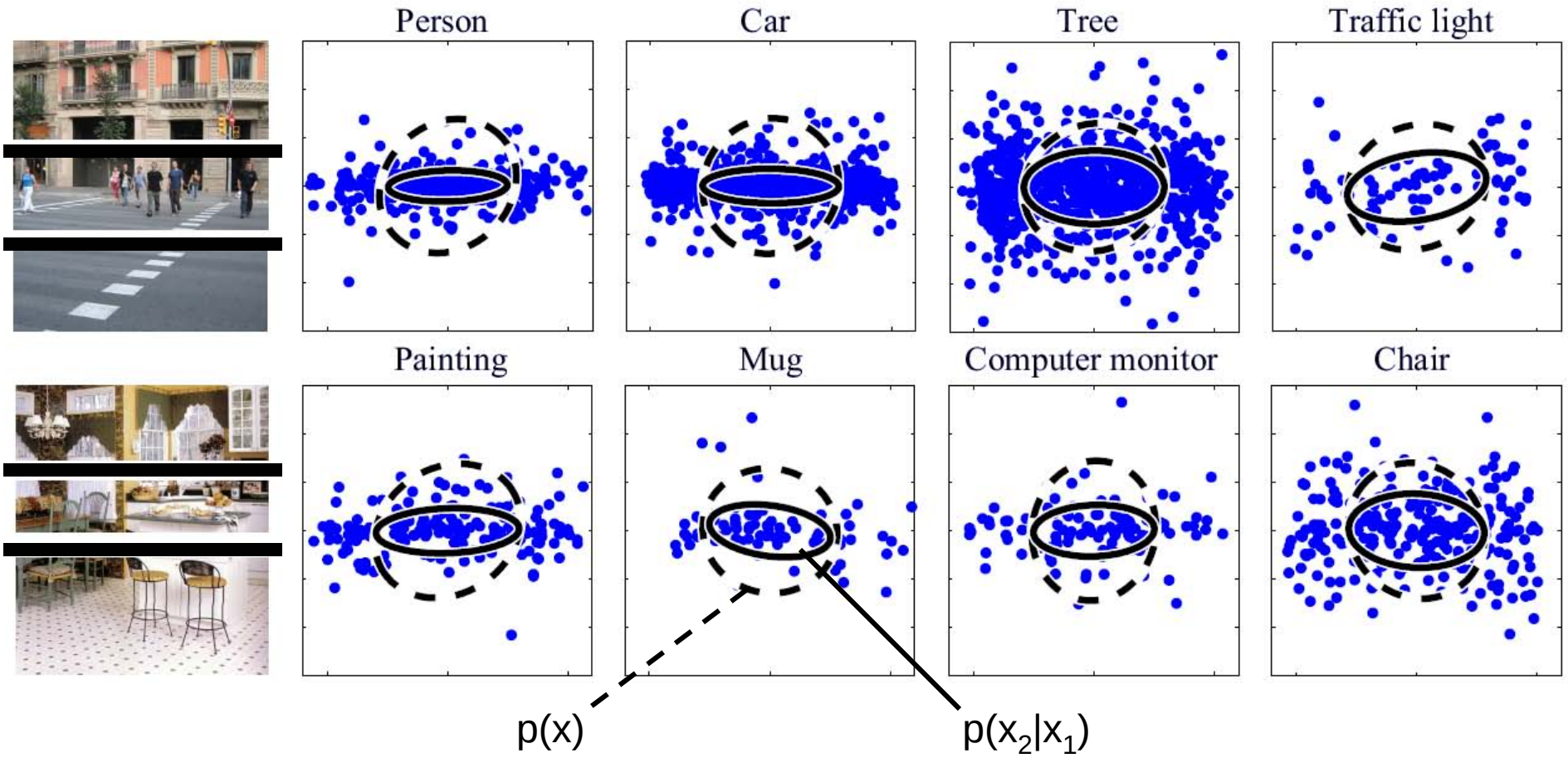

Results for predicting the vertical location of people




Results for predicting the horizontal location of people

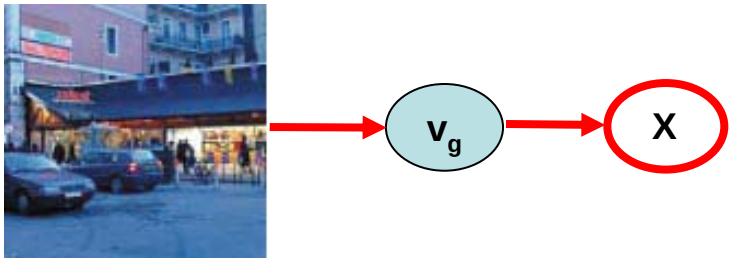


# The layered structure of scenes




In a display with multiple targets present, the location of one target constraints the ‘y’ coordinate of the remaining targets, but not the ‘x’ coordinate.

# Global scene features predicts object location

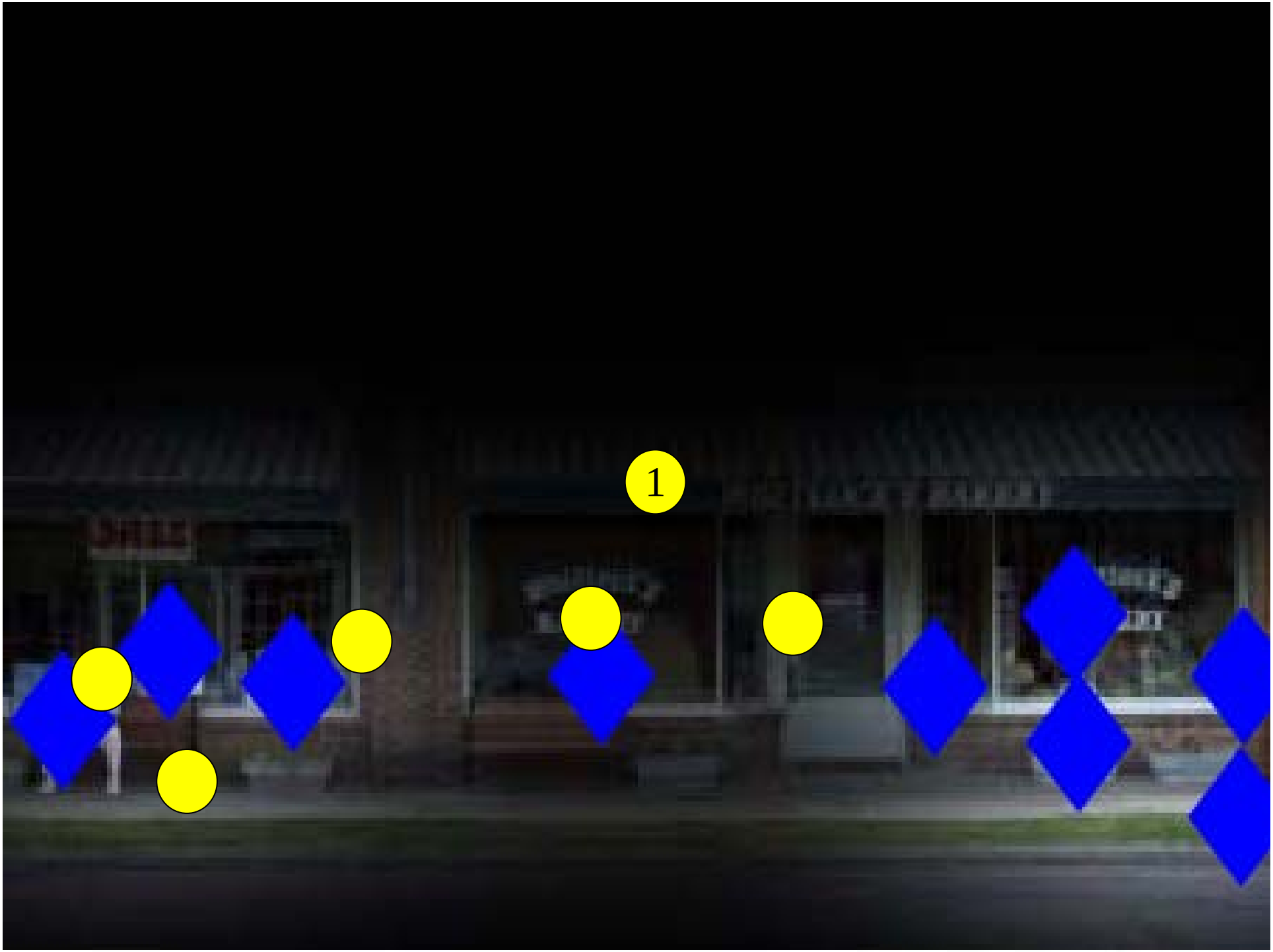



Stronger contextual constraints can be obtained using other objects.

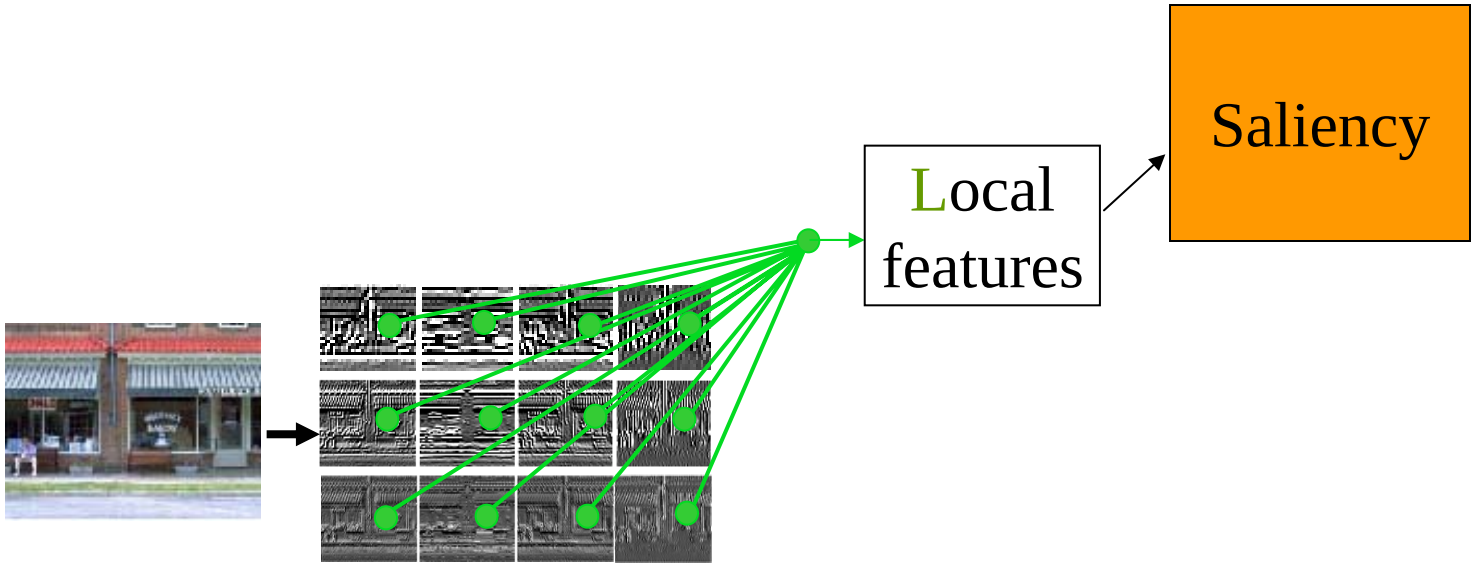

SALE
$15.00
WHITLOCK'S BAKERY
Whitlock's
BAKERY
Whitlock's
BAKERY


SALE
$15.00
WHITLOCK'S BAKERY
Whitlock's
BAKERY
Whitlock's
BAKERY


1

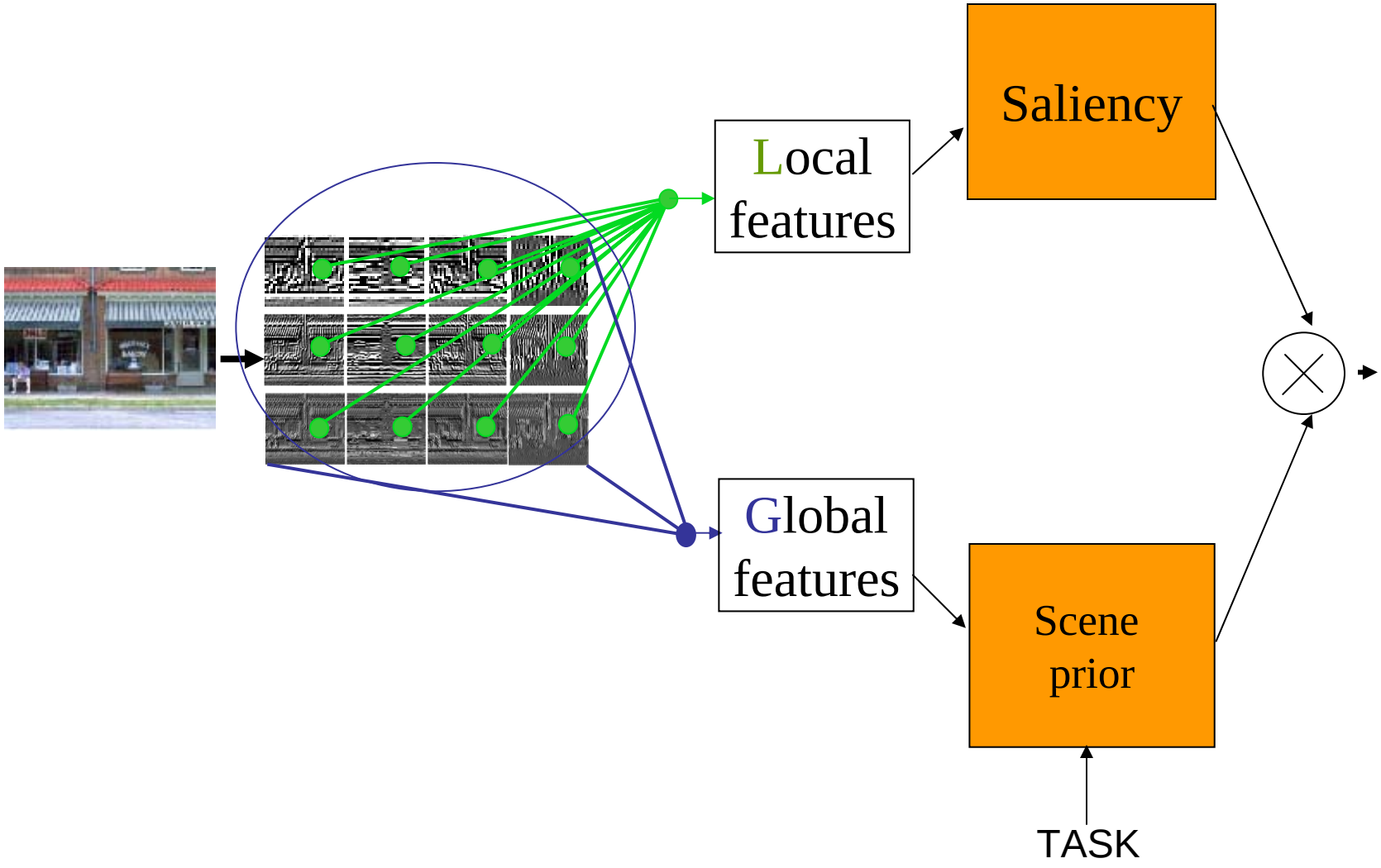

1

# Attentional guidance



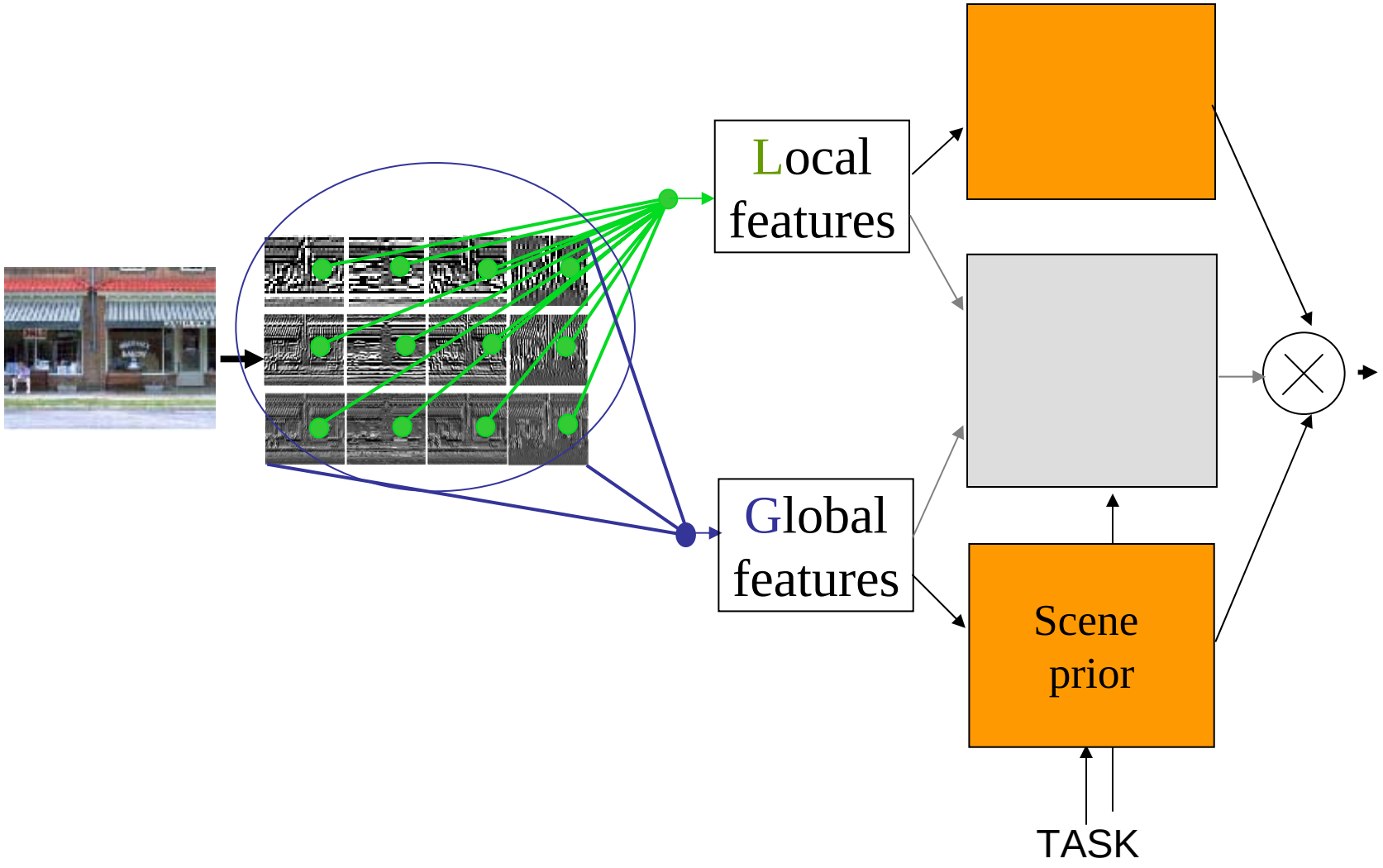

Saliency models: Koch & Ullman, 85; Wolfe 94; Itti, Koch, Niebur, 98; Rosenholtz, 99

# Attentional guidance




Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003

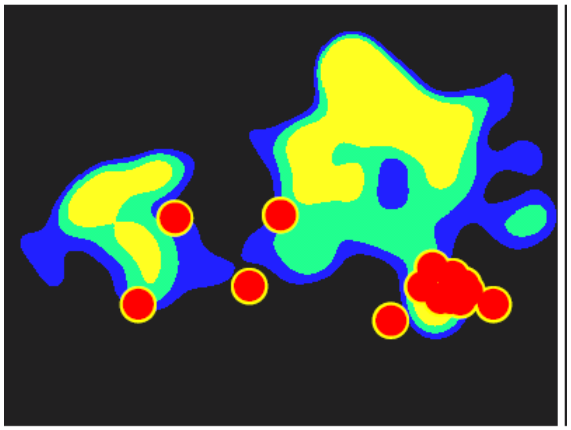
# Attentional guidance



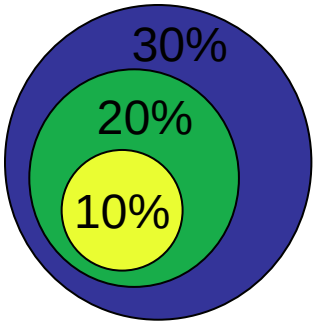
# Comparison regions of interest

Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003

# Comparison regions of interest

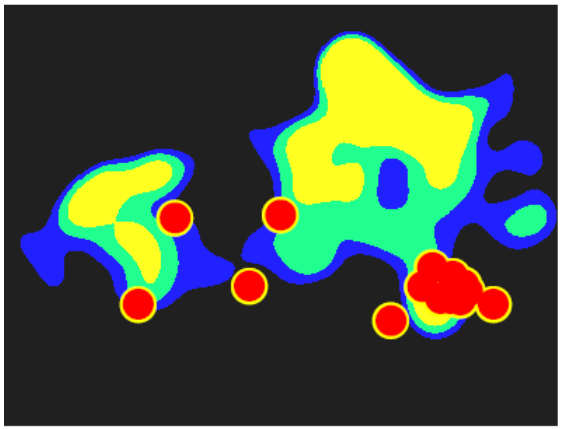


Saliency predictions

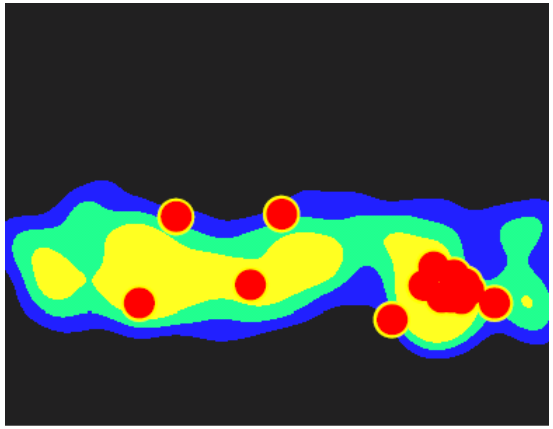



Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003

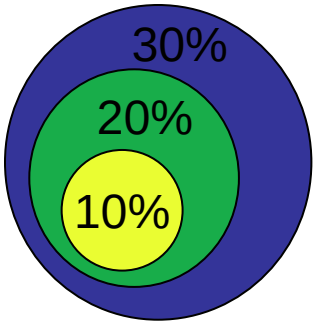
# Comparison regions of interest



Saliency predictions

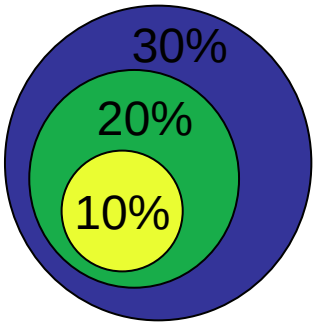


Saliency and Global scene priors

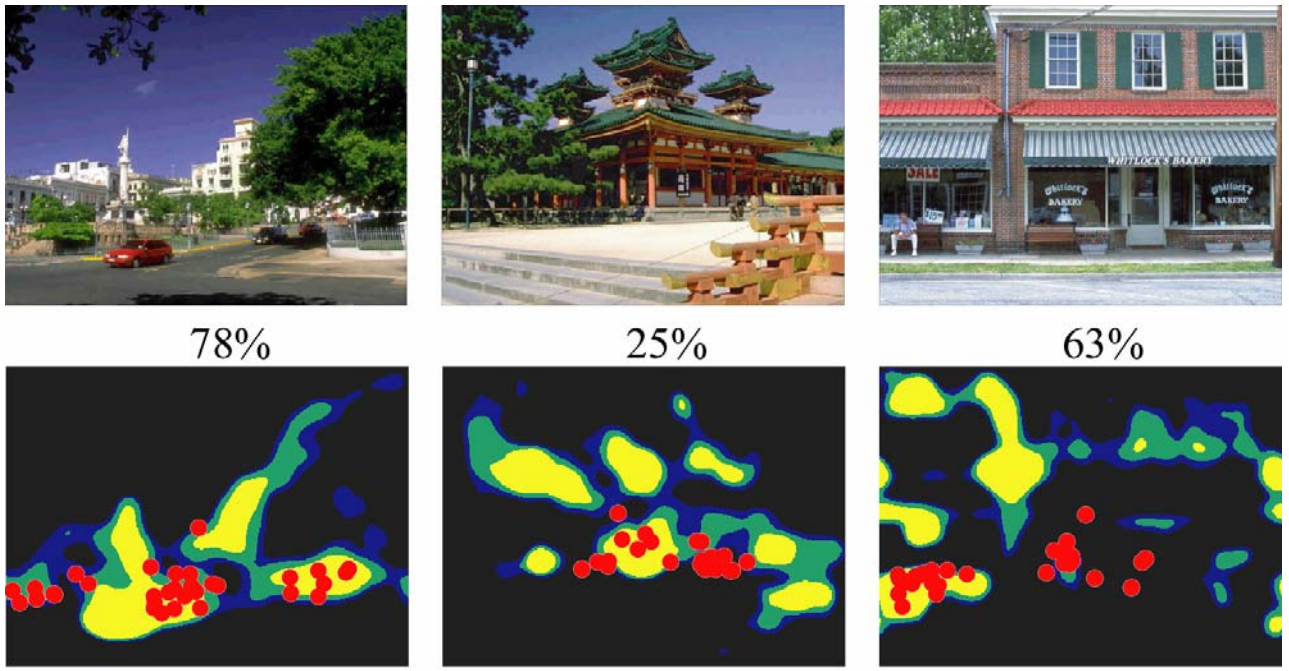



Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003

# Comparison regions of interest



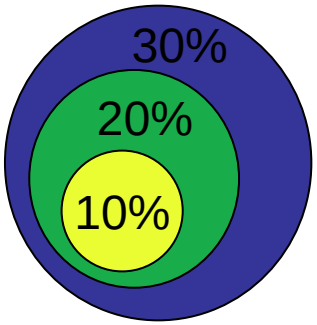

Saliency predictions




Dots correspond to fixations 1-4

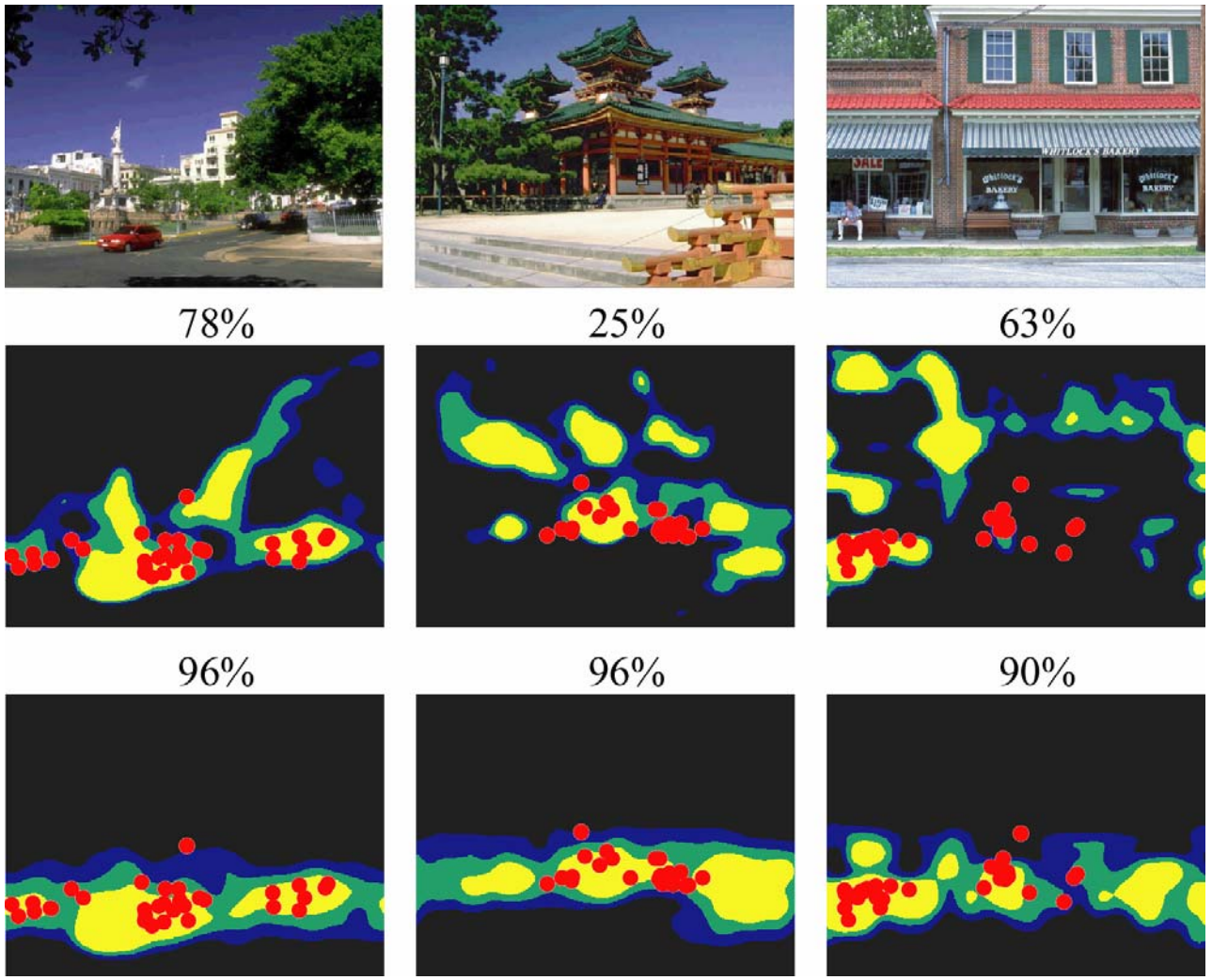
Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003

# Comparison regions of interest




Saliency predictions

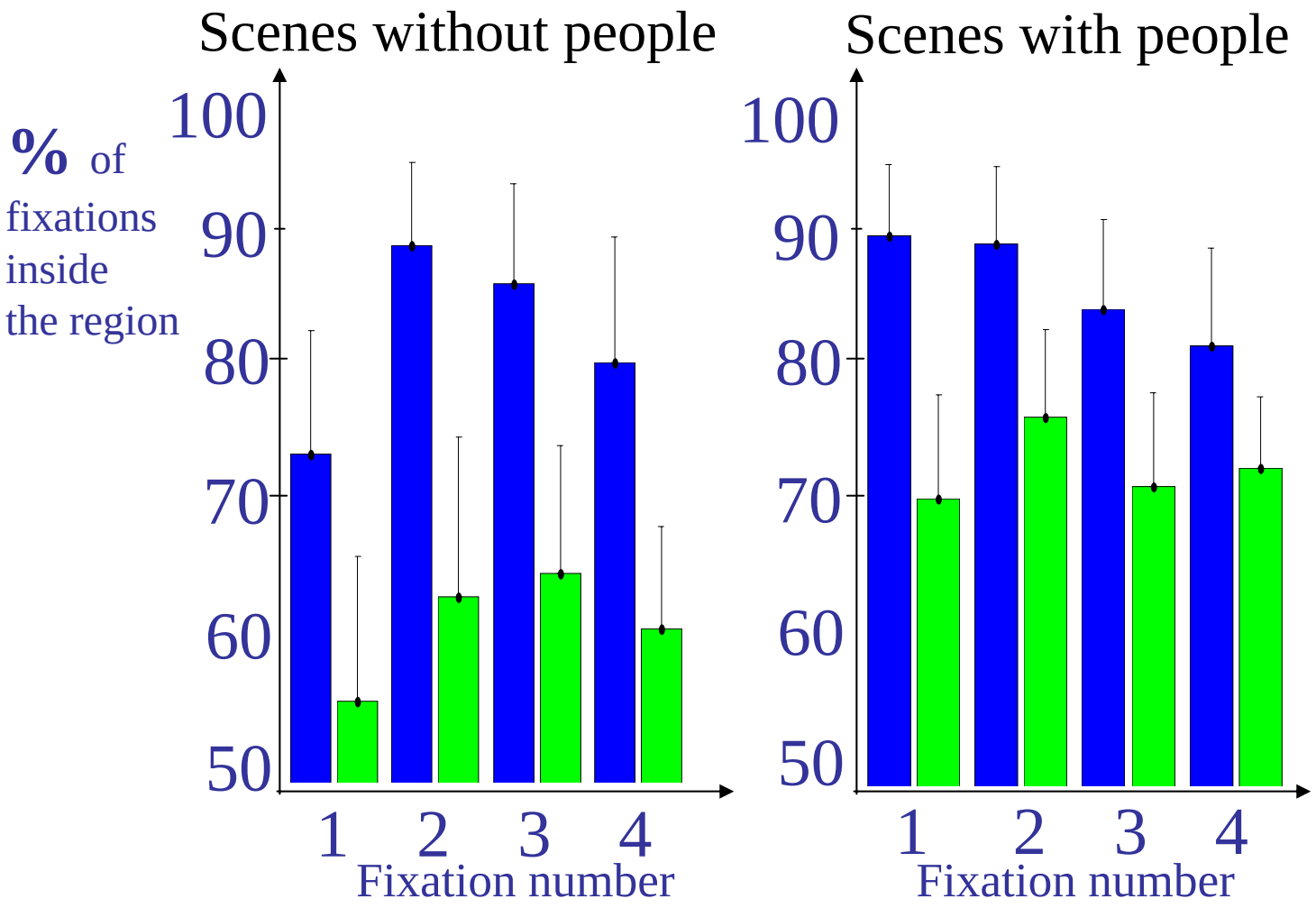
Saliency and Global scene priors



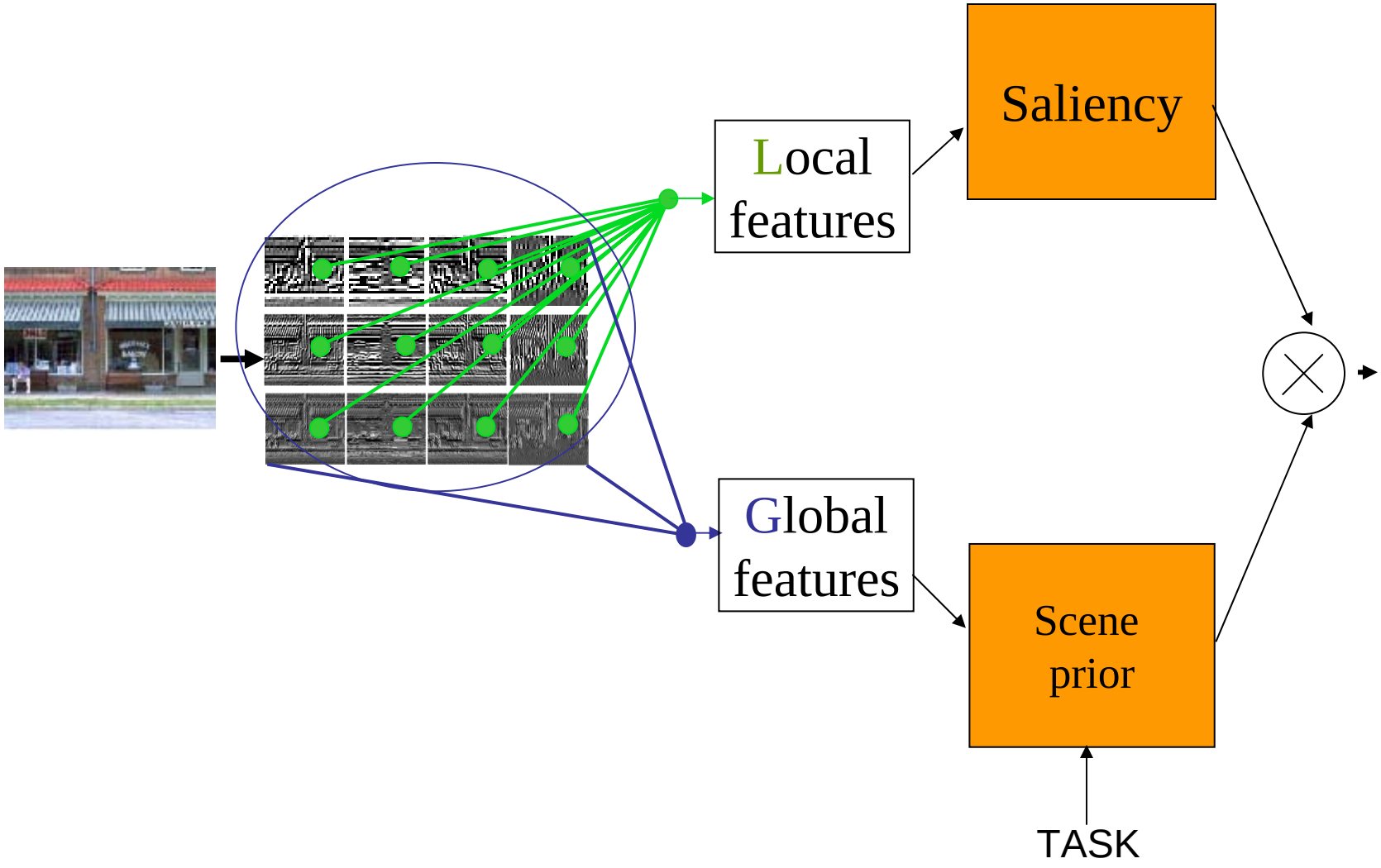

Dots correspond to fixations 1-4

Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003

# Results

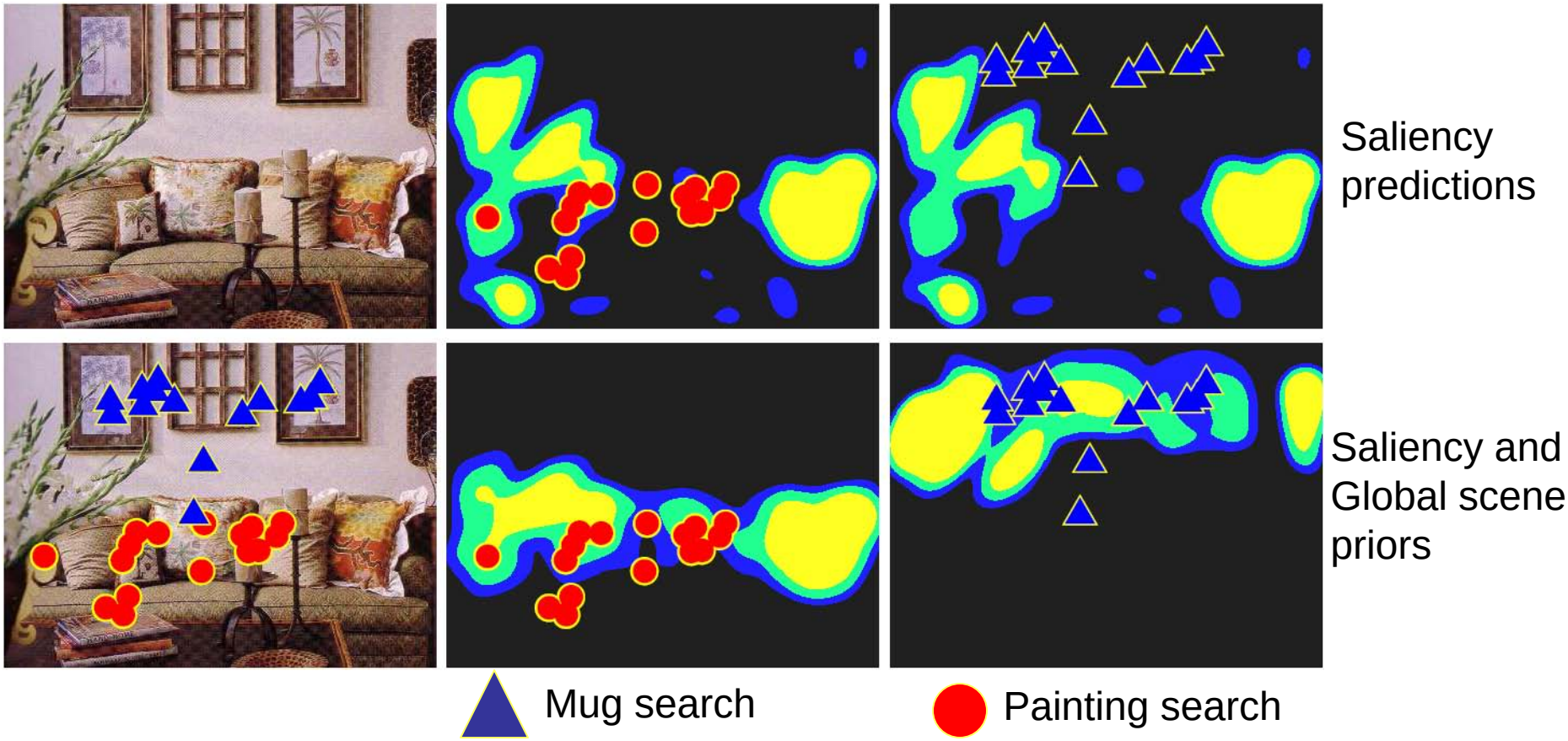

Scenes without people
Scenes with people
% of fixations inside the region
100
90
80
70
60
50
1
2
3
4
Fixation number
Chance level: 33 %
Saliency Region
Contextual Region

# Task modulation




Torralba, 2003; Oliva, Torralba, Castelhano, Henderson. ICIP 2003

# Task modulation



# Discussion

- From the computational perspective, scene context can be derived from global image properties and predict where objects are most likely to be.

- Scene context considerably improves predictions of fixation locations. A complete model of attention guidance in natural scenes requires both saliency and contextual pathways