Visual Saliency

CPS592 – Visual Computing and Mixed Reality

Outline

- What is visual saliency?
- Saccade vs. fixation
- Visual Phenomena
- Applications

What is Visual Saliency?

- Something is said to be **salient** if it **stands out**
- **E.g. road signs should have high saliency**

Visual Saliency

- It is important as it drives a decision we make a couple hundred thousand times a day - where we decide to look.

- The role of Cognitive Science is to create a working model of visual saliency.

Finding “interesting” information

- In principle, very complex task:
  - Need to attend to all objects in scene?
  - Then recognize each attended object?
  - Finally evaluate set of recognized objects against behavioral goals?
- In practice, survival depends on ability to quickly locate and identify important information.
- Need to develop simple heuristics or approximations:
  - bottom-up guidance towards salient locations
  - top-down guidance towards task-relevant locations

Saliency in human eyes

Slow attention process – example:

Firs focus here:

And then notice the cat and Baby.

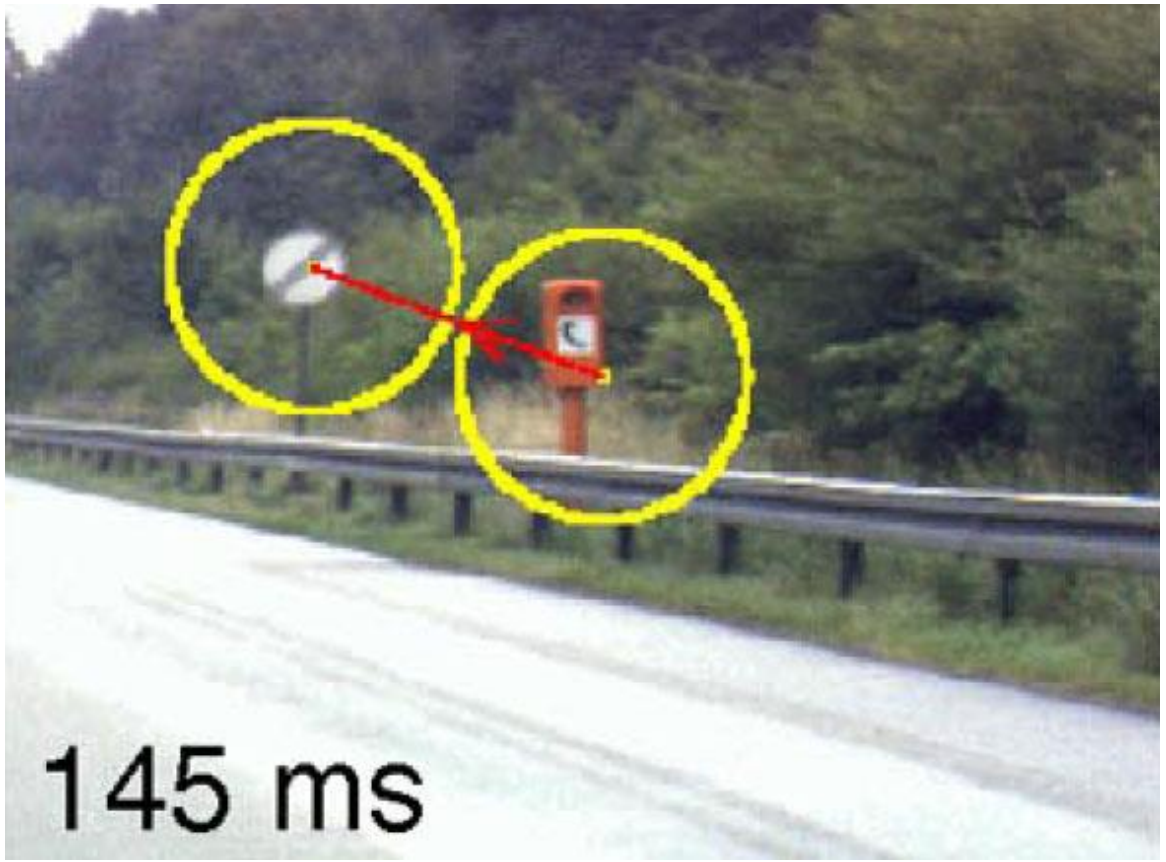
Introduction

- Trying to model visual attention
- Find locations of **Focus of Attention** in an image
- Use the idea of saliency as a basis for their model
- For primates focus of attention directed from:
  - **Bottom-up: rapid, saliency driven, task-independent**
  - Top-down: slower, task dependent

Results of the Model

- Only considering “Bottom-up”

  → task-independent



Fixations

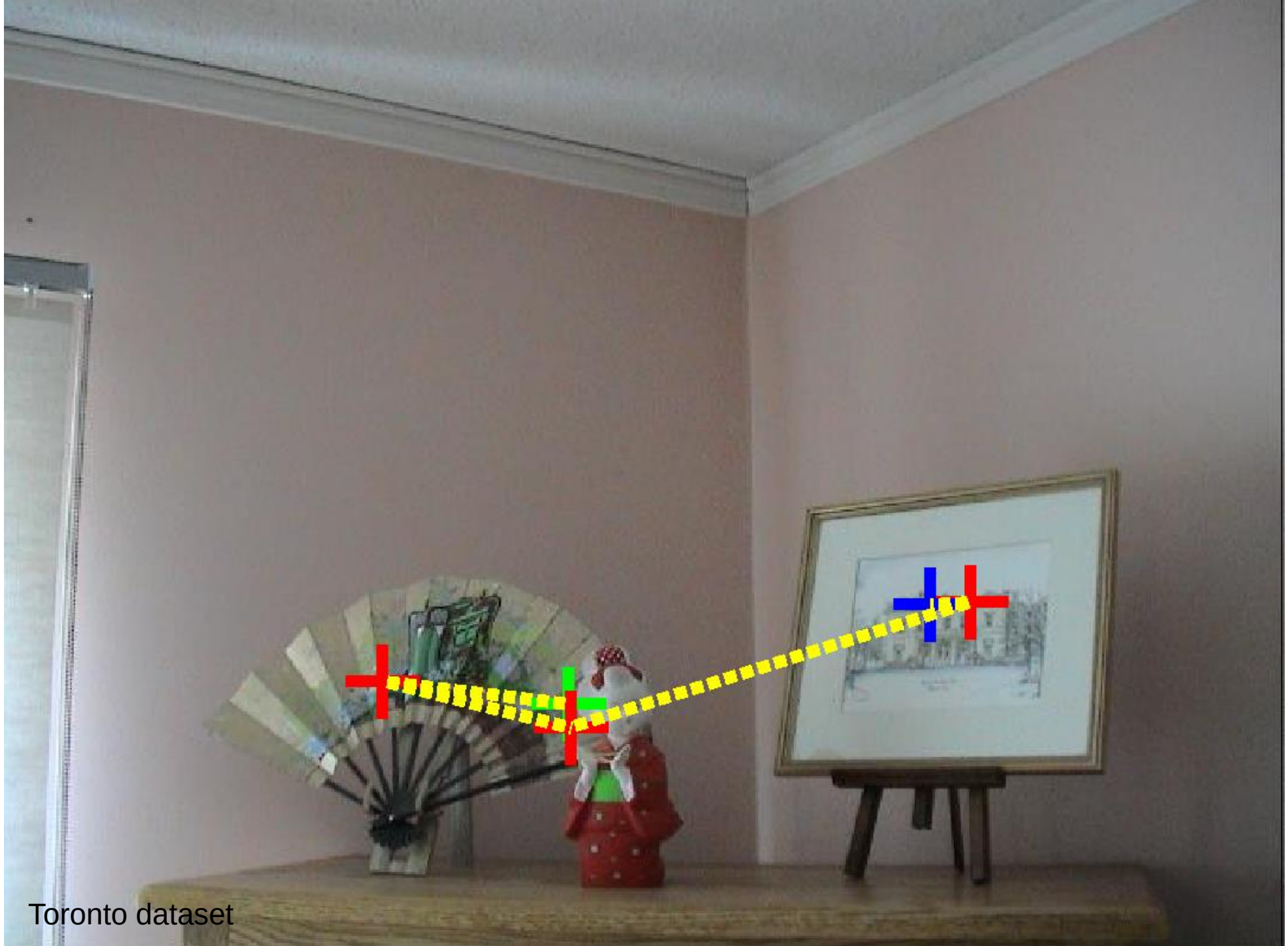
Fixation: period when eye is relatively stationary between saccades.

Saccades

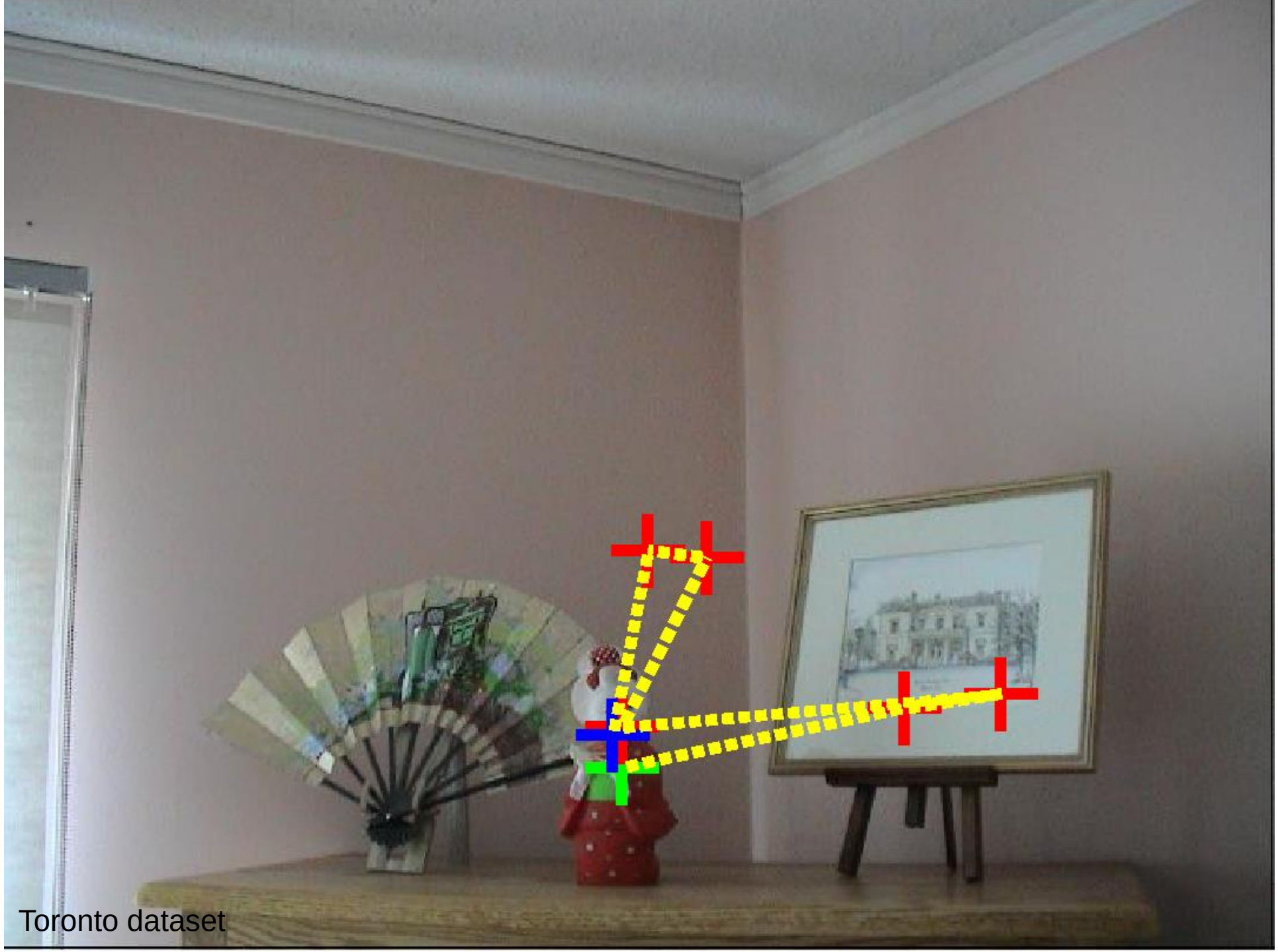
- Scope: 2 deg (poor spatial res beyond this)
- Duration: 50-500 ms (mean 250 ms)
- Length: 0.5 to 50 degrees (mean 4 to 12)
- Various types (e.g., regular, tracking, micro)

Source: CVPR Tutorial 2013

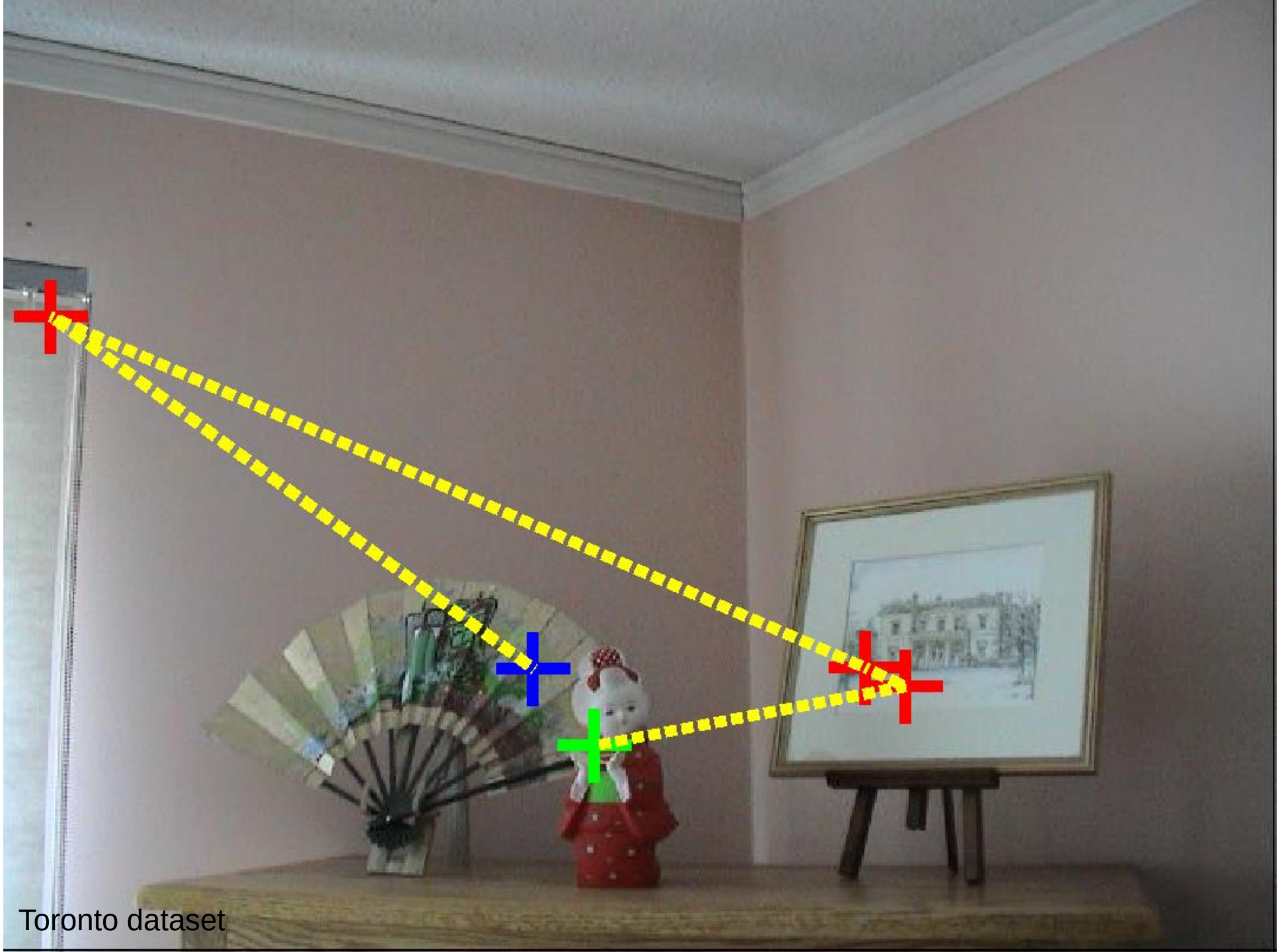
Toronto dataset



Toronto dataset



Toronto dataset



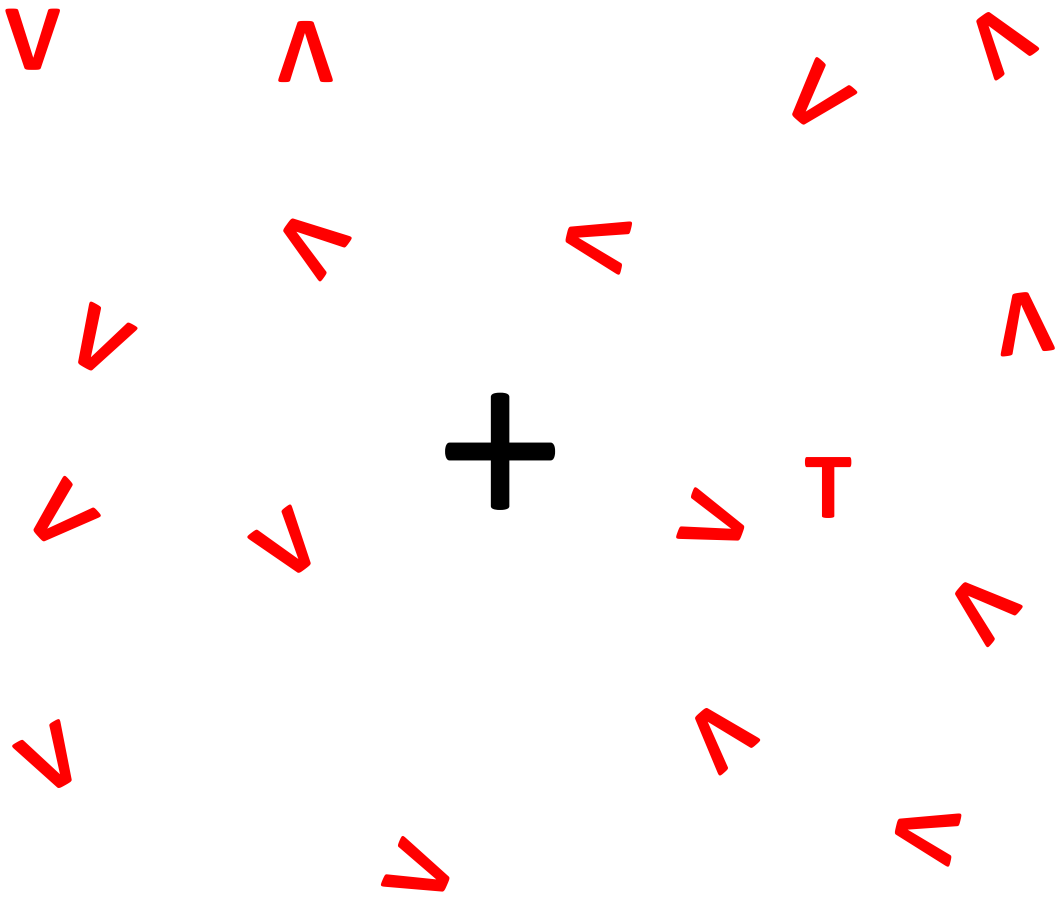
Toronto dataset

Toronto dataset

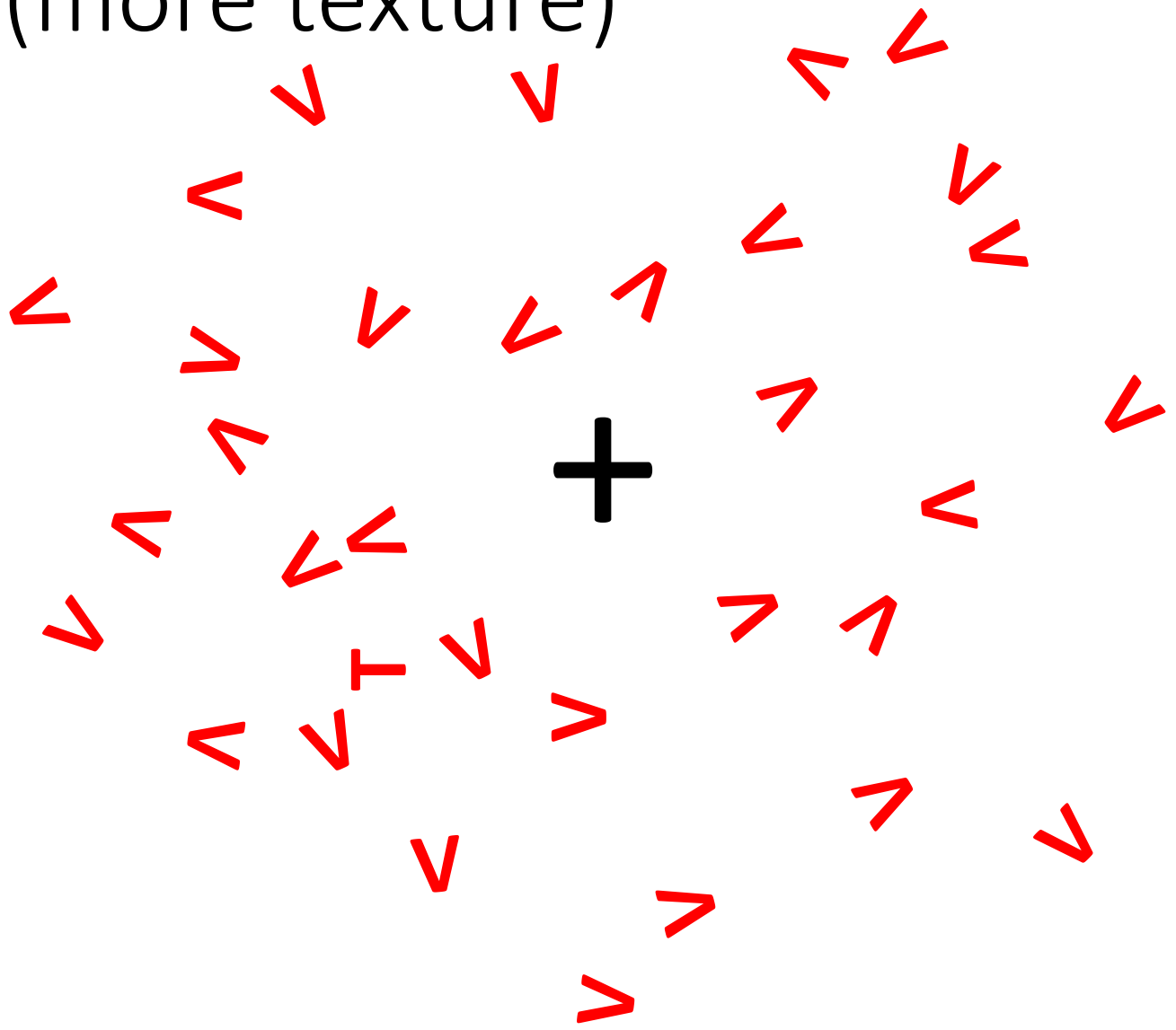
Visual Phenomena

- Pop-out
- Attentional blindness
- Change blindness

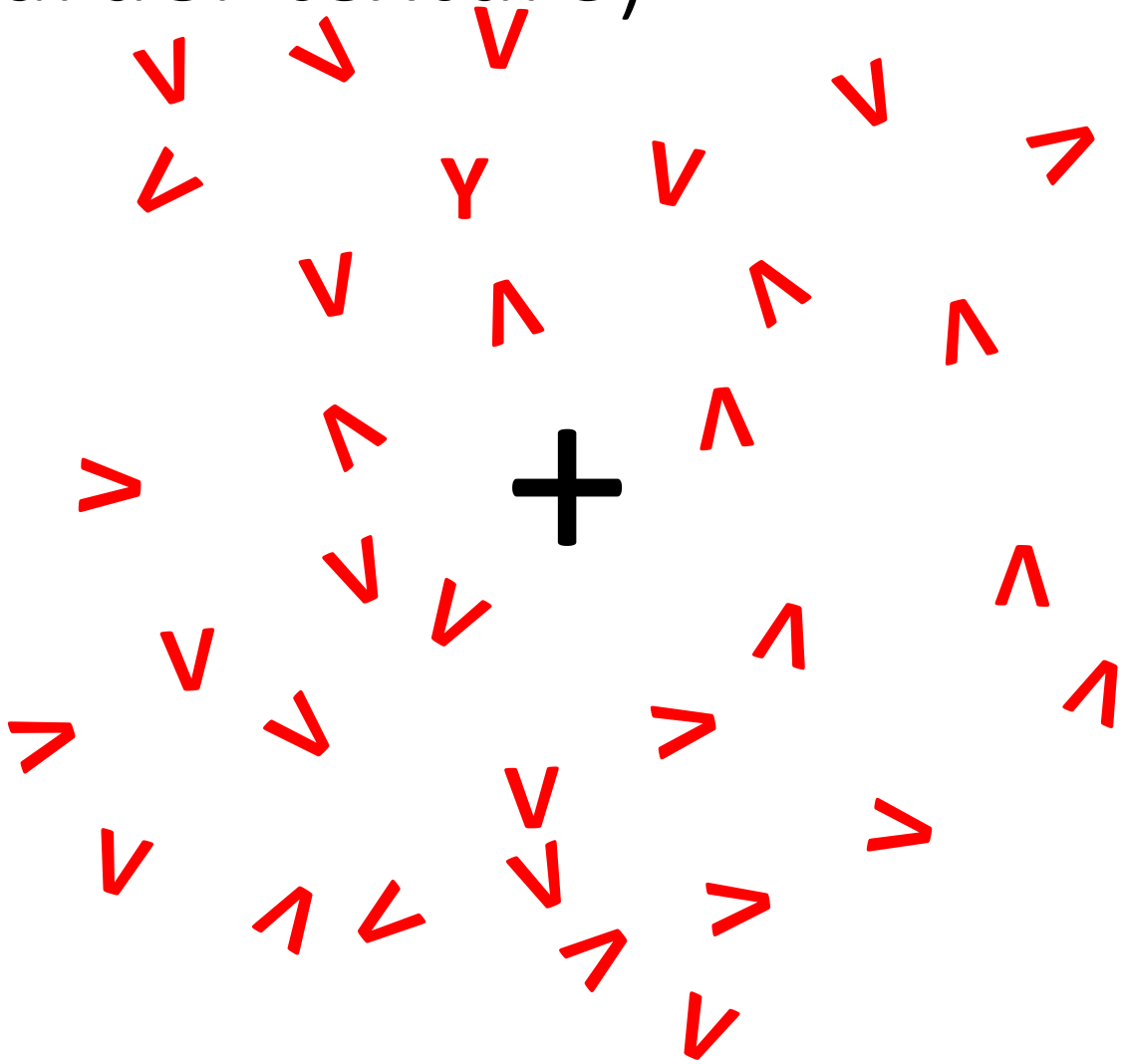
Pop-out (texture)



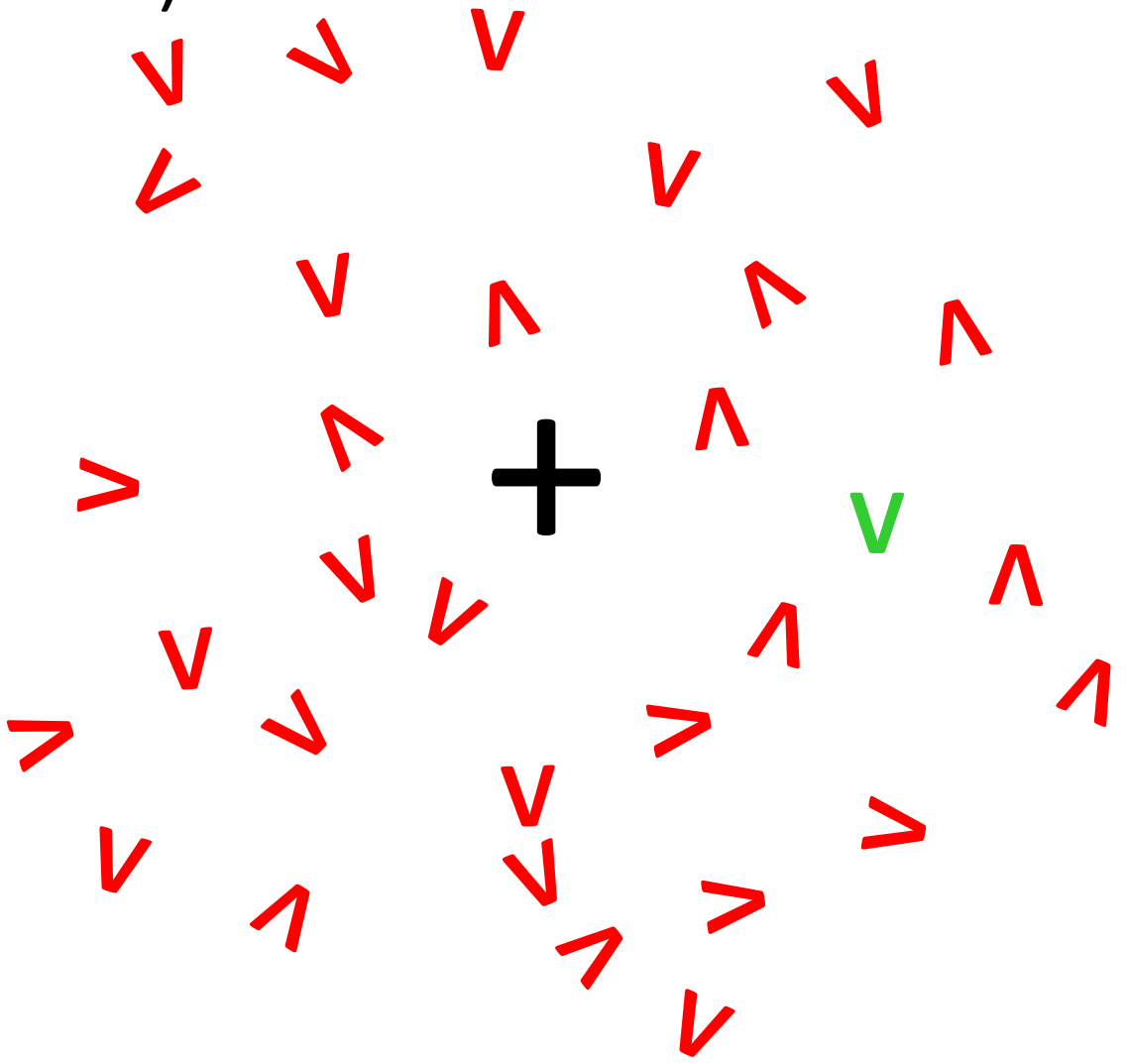
Pop-out (more texture)



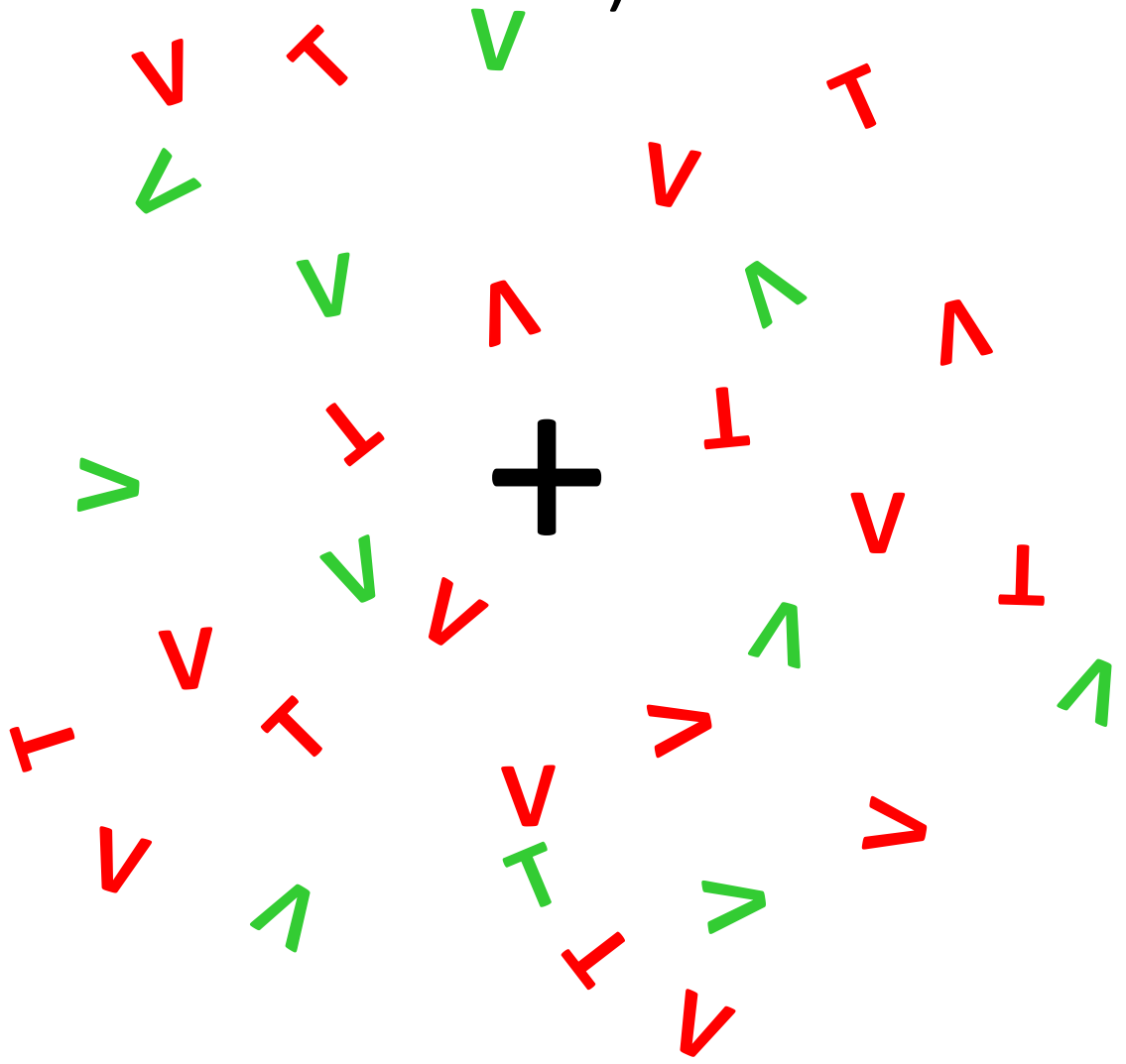
Pop-out (harder texture)



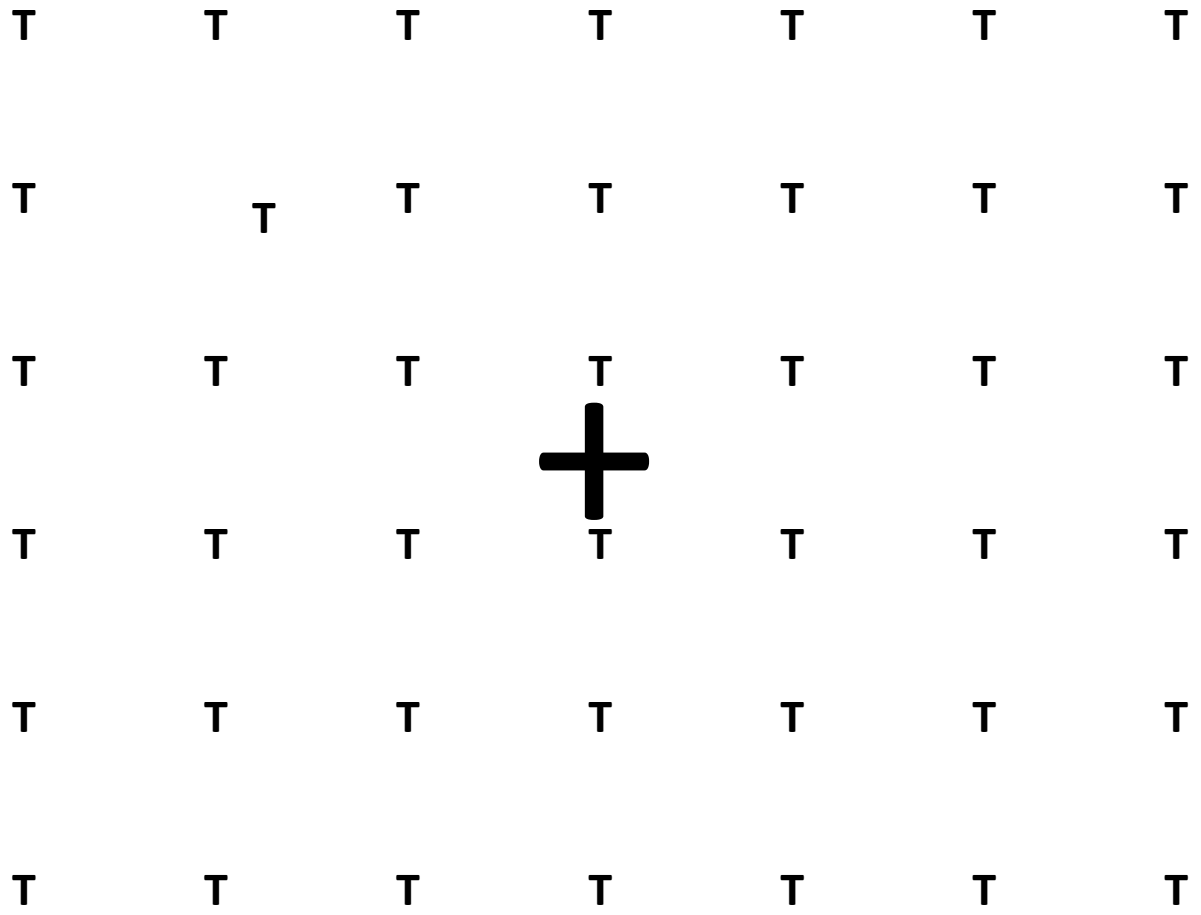
Pop-out (color)



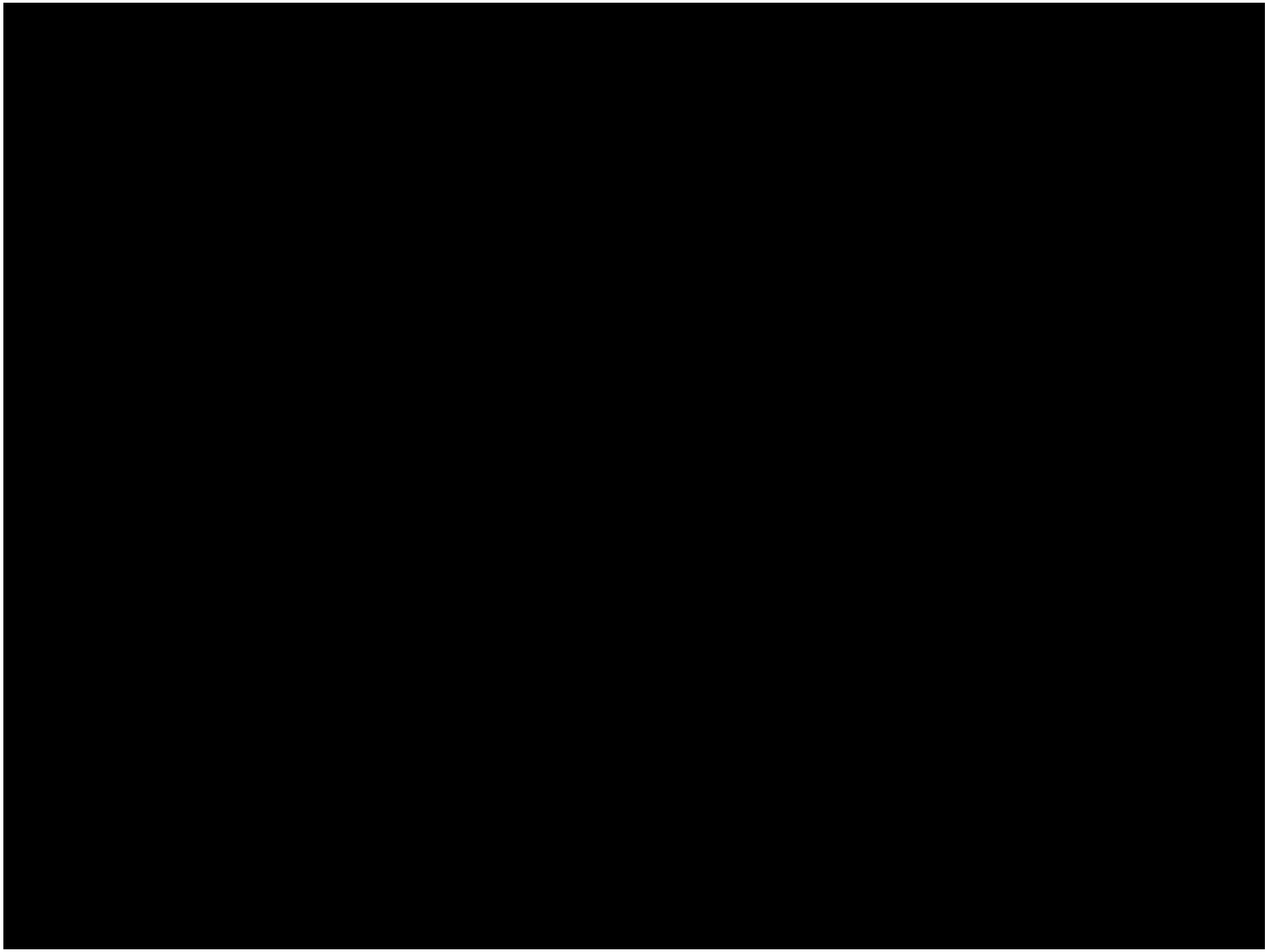
Pop-out (color + texture)



Pop-out (layout)



Attentional blindness



Visual Attention (Change Blindness)

Visual Attention (Change Blindness)

Visual Attention (Change Blindness)

Visual Attention (Change Blindness)

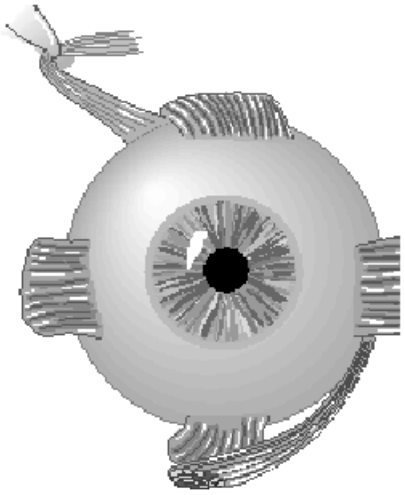
Visual Attention (Change Blindness)

Why?

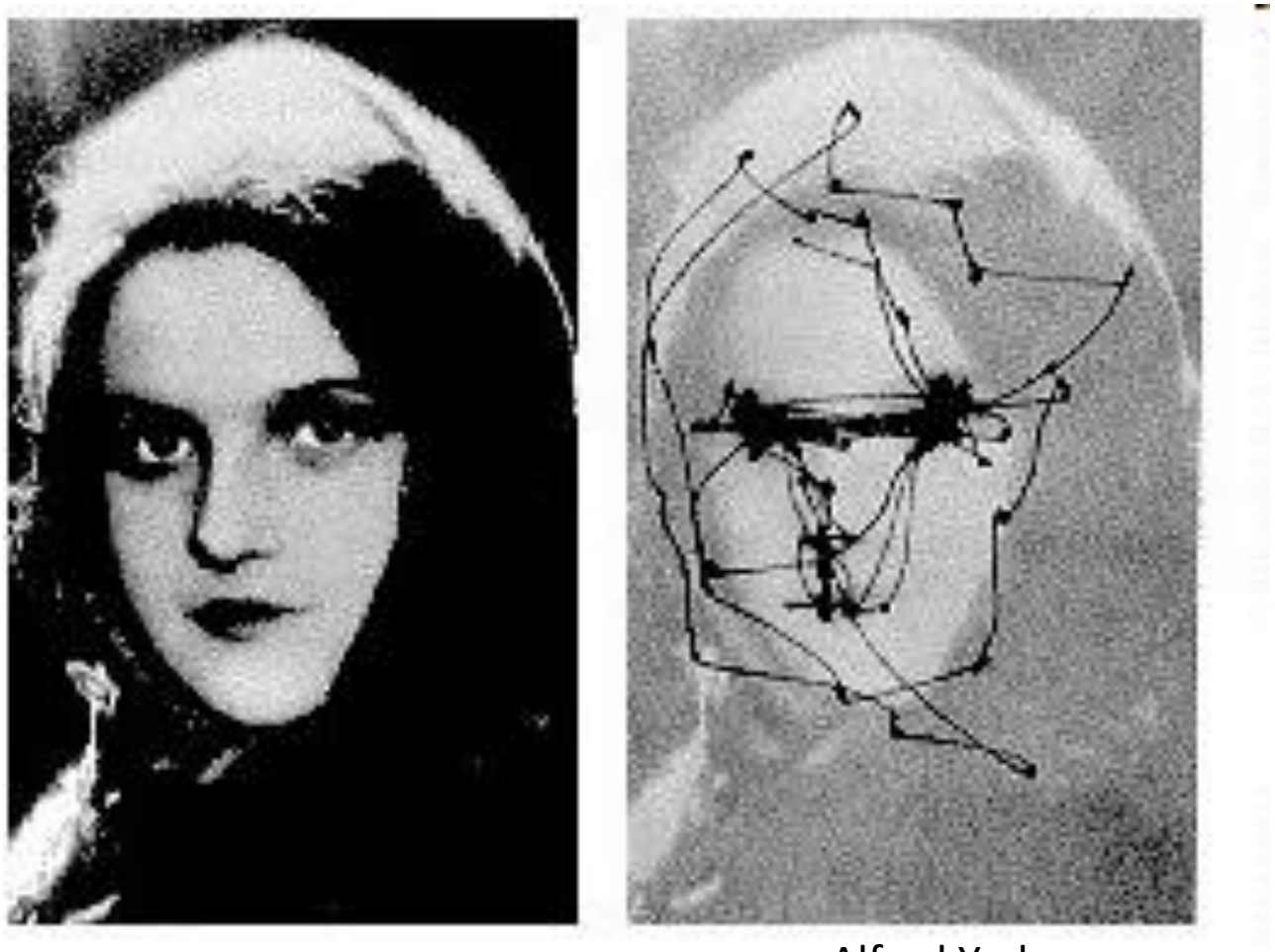
A: Limitations of the eye – only fovea is high-res enough for many tasks

The Eye

- 120 million rods (intensity)
- 7 million cones (color)
- Fovea: 2 degrees of cones



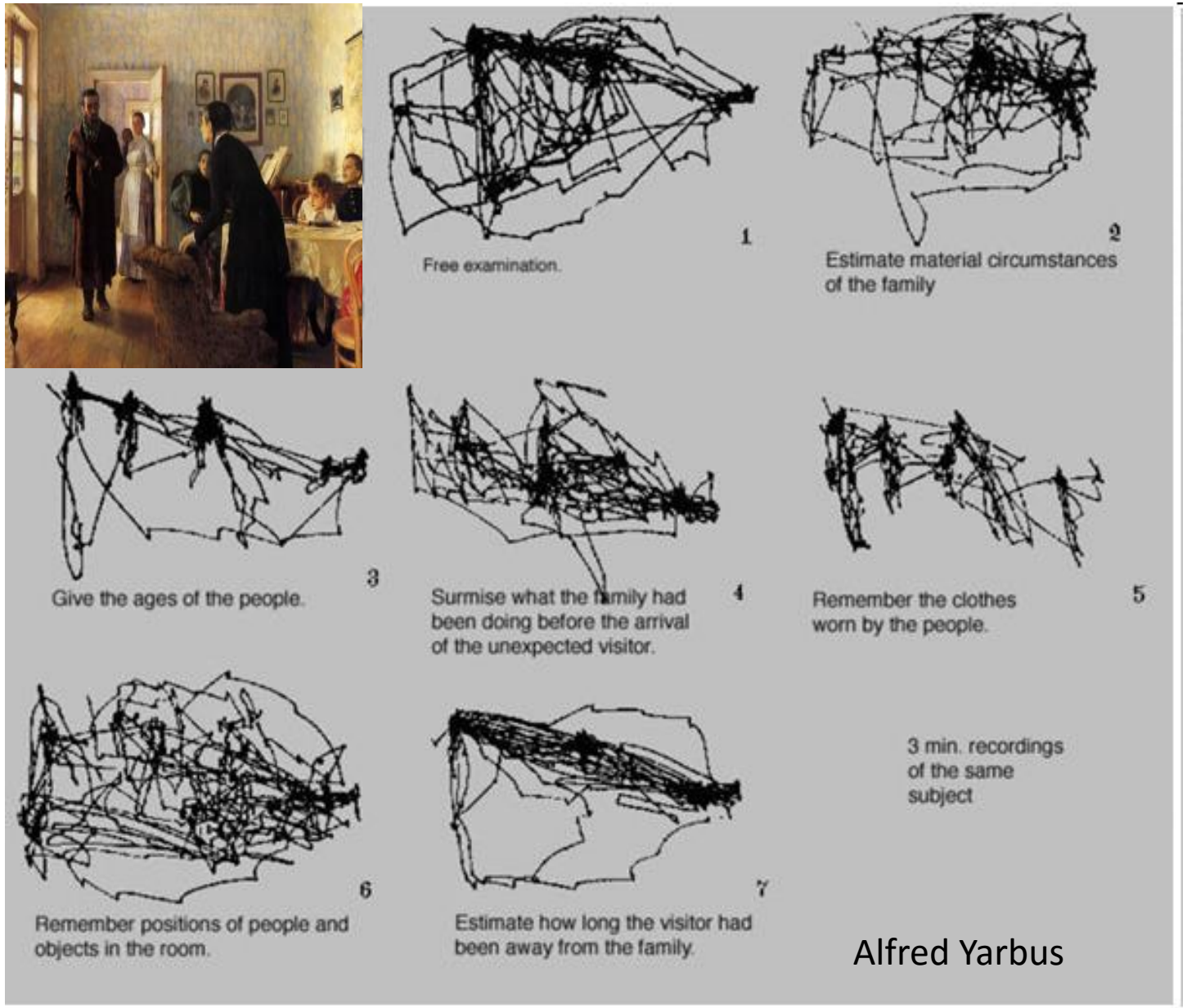
Eye-Tracking



Alfred Yarbus

Experiment

goal-attenuated



Alfred Yarbus

Purpose of visual saliency models

- Warning (animals, flashes, sudden motion)
- Exploration (find objects, verification)
- Inspection

Motivation - application

Image mosaicking: the salient details are preserved, with the use of smaller building blocks.

Input

Mosaic

Motivation - application

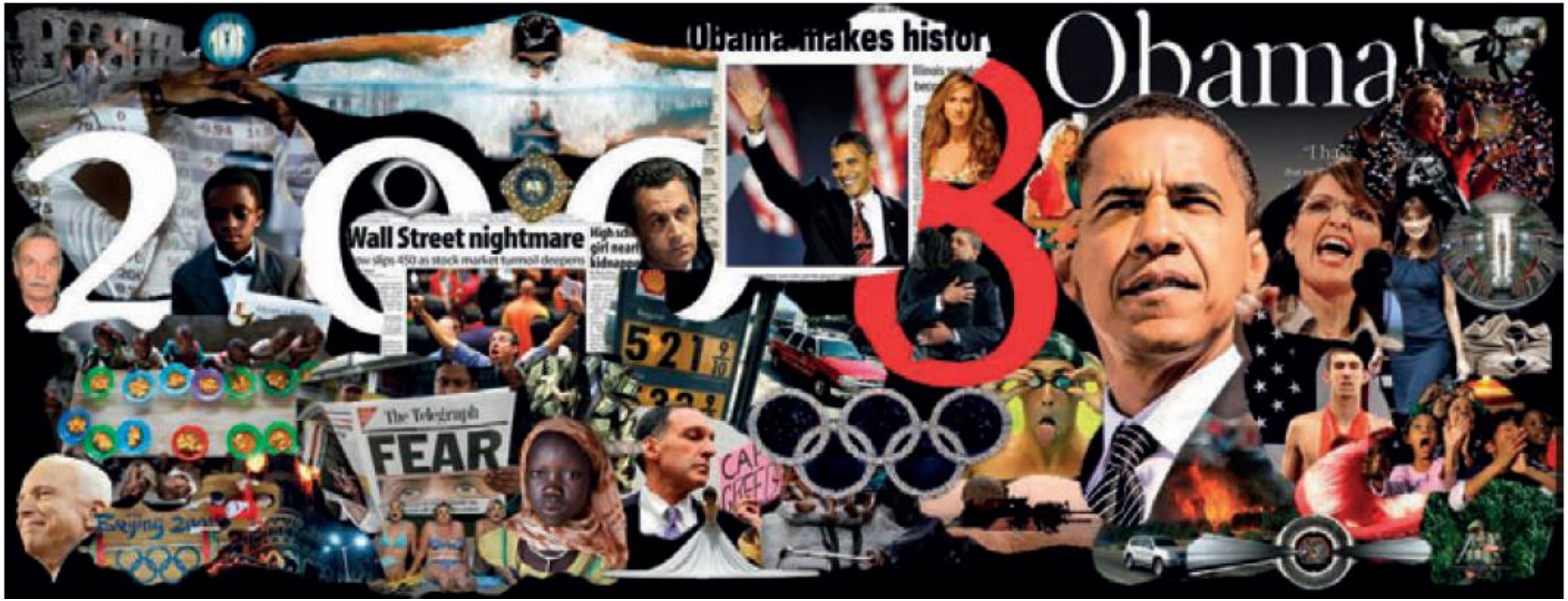
Painterly rendering – the fine details of the dominant objects are maintained, abstracting the background

Input

Painterly rendering

Motivation - application

Puzzle-like collage:

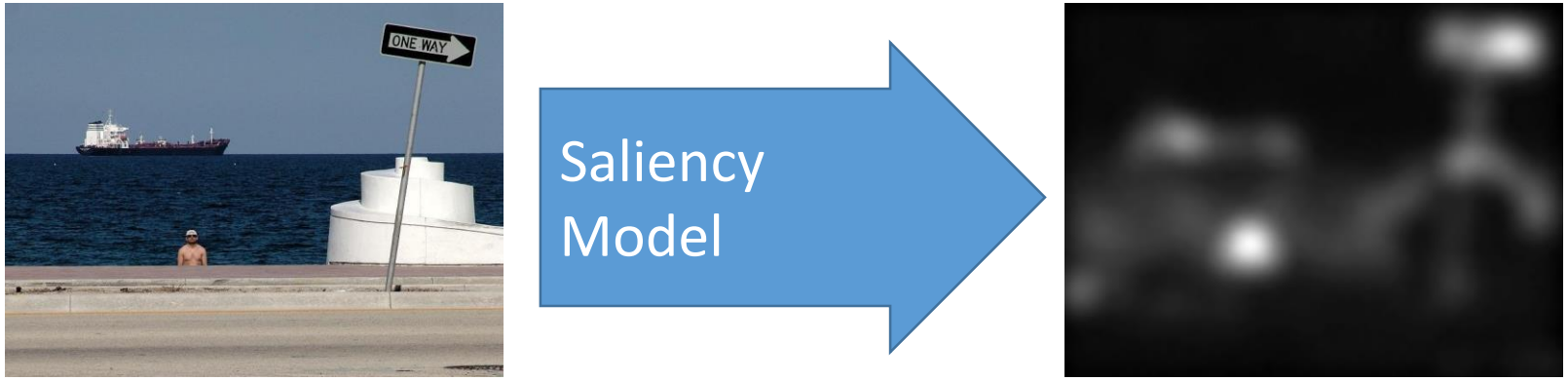


Outline

- Computational models
- Itti’s model
- Frequency Tuned model
- Context aware model

Computational saliency models

Researchers create computational models to predict where people look.

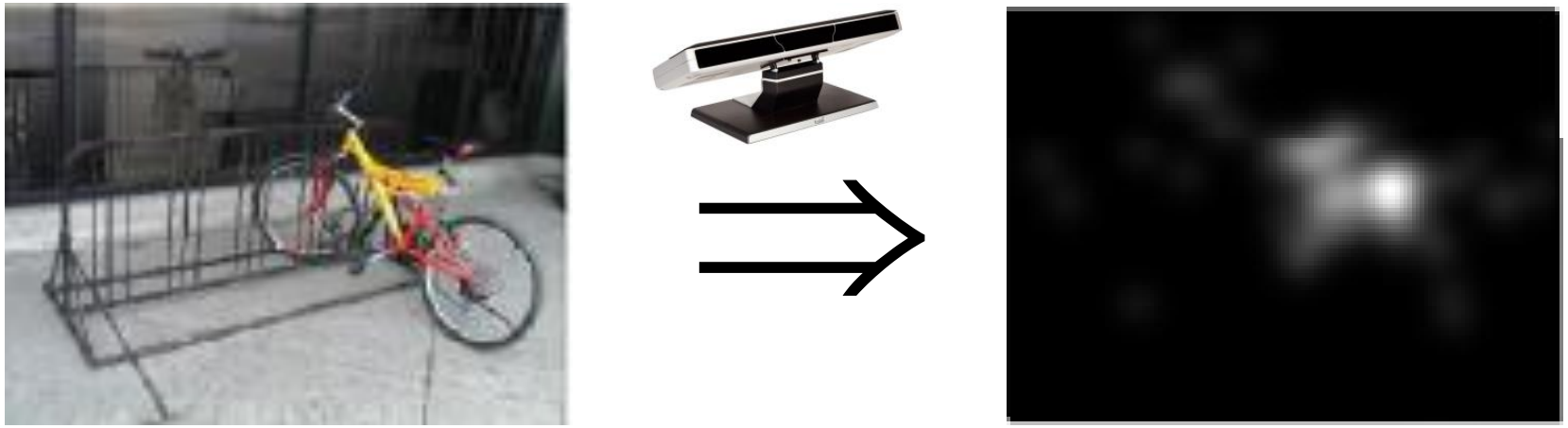


Saliency Maps

- Itti et al proposed that bottom-up attention can be predicted from low-level visual features.
- Eye-tracking can be used to validate the predictions

Fixation maps

Example for saliency map generated by eye tracker:



A Model of Saliency-Based Visual Attention for Rapid Scene Analysis

Laurent Itti, Christof Koch, and Ernst Niebur

Itti's model

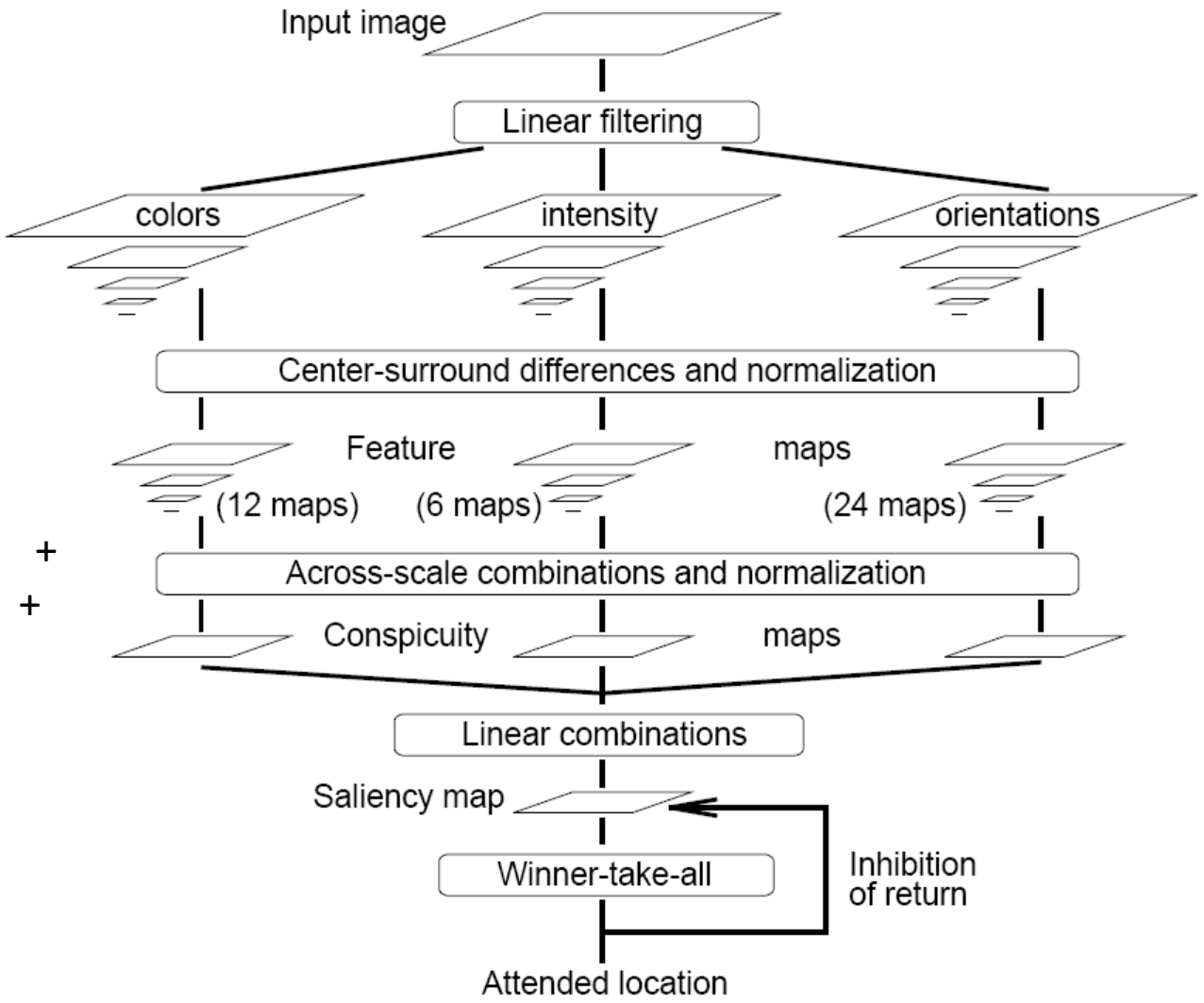
Gabor Pyramid +
Orientation Filters

Subtract low-res (3-4 octaves) from higher res

Normalize (0..1) +
map * (1 – max_{ave})^2 +
add maps

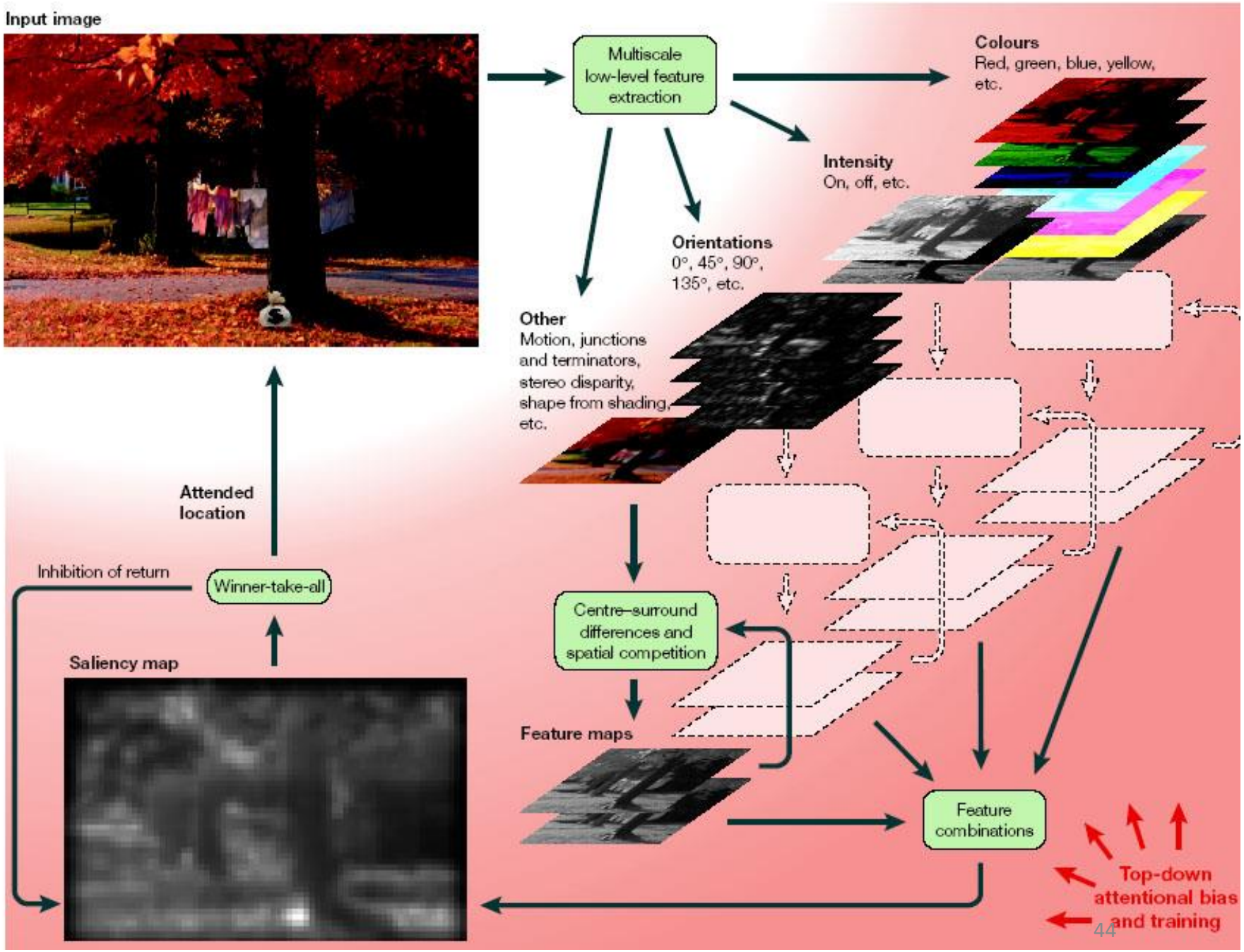
Average Maps

Inhibition + Excitation



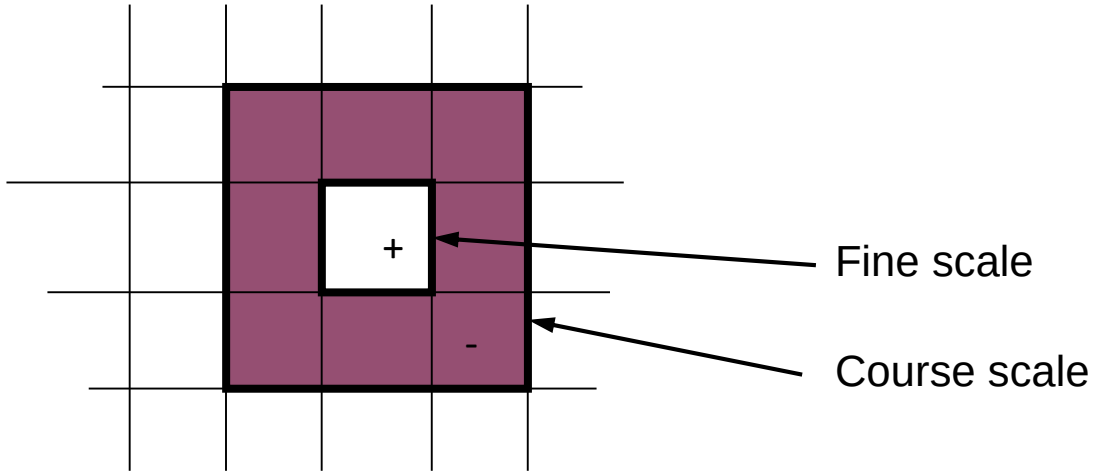
[Itti Koch Niebur 1998]

How it works


Input image
Multiscale low-level feature extraction
Colours
Red, green, blue, yellow, etc.
Intensity
On, off, etc.
Orientations
0°, 45°, 90°, 135°, etc.
Other
Motion, junctions and terminators, stereo disparity, shape from shading, etc.
Attended location
Inhibition of return
Winner-take-all
Saliency map
Centre–surround differences and spatial competition
Feature maps
Feature combinations
Top-down attentional bias and training

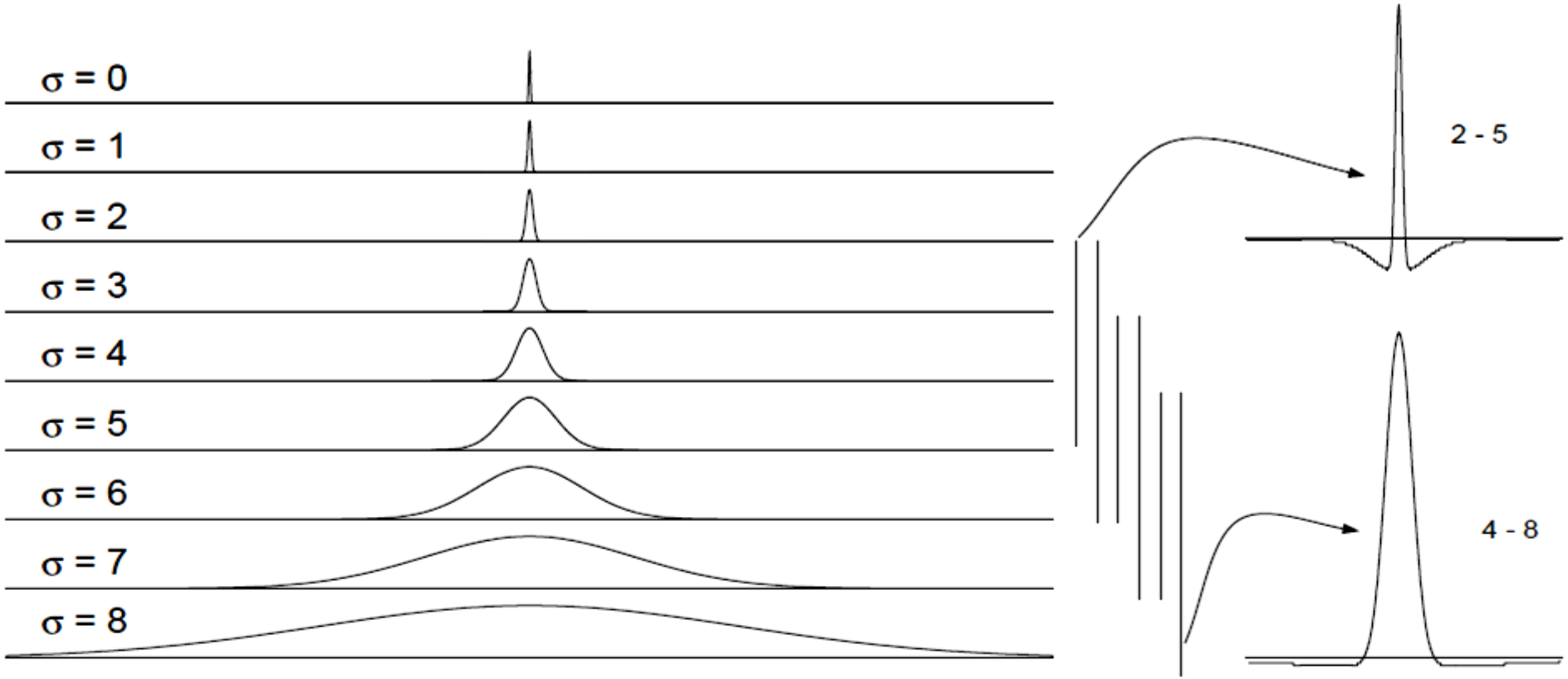
Model

- Input: static images (640x480)
- Each image at 8 different scales (640x480, 320x240, 160x120, …)
- Use different scales for computing “centre-surround” differences (similar to assignment)



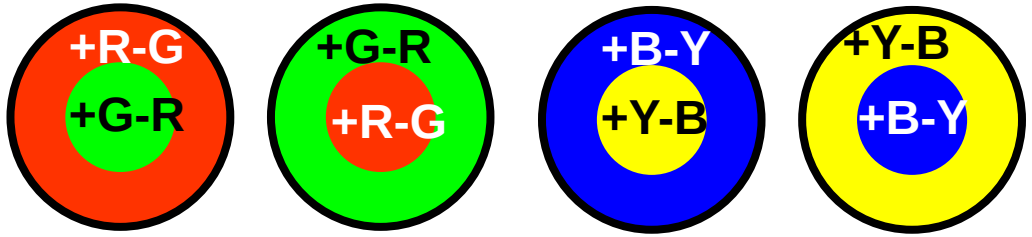
Center-surround Difference

- Achieve center-surround difference through across-scale difference



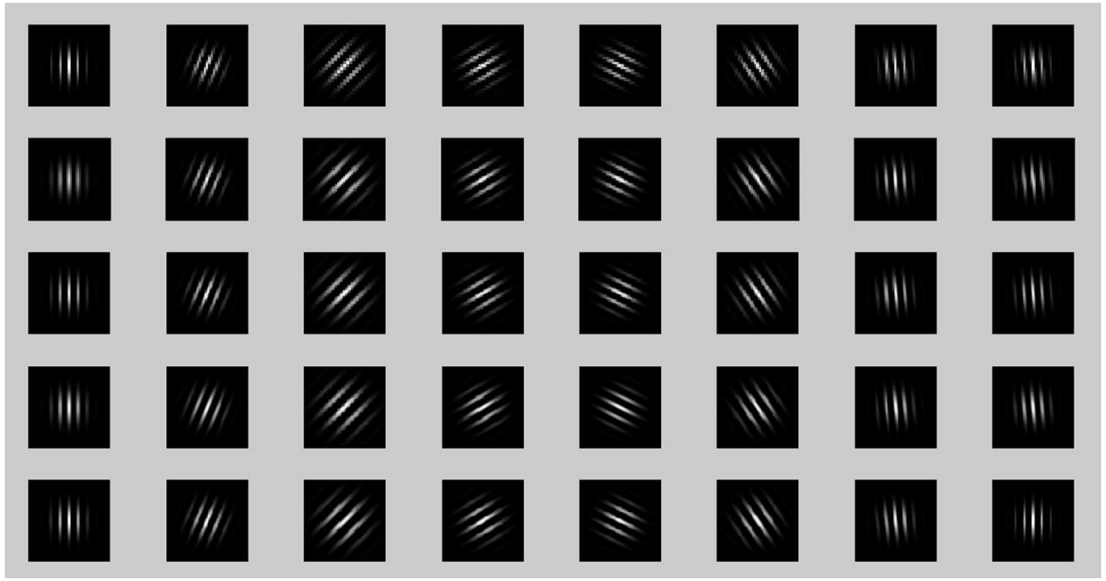
Feature Maps

1. Intensity contrast (6 maps)
   - Using “centre-surround”
   - Similar to neurons sensitive to **dark centre, bright surround**, and opposite
2. Color (12 maps)
   - Similar to intensity map, but using different color channels
   - E.g. high response to **centre red, surround green**



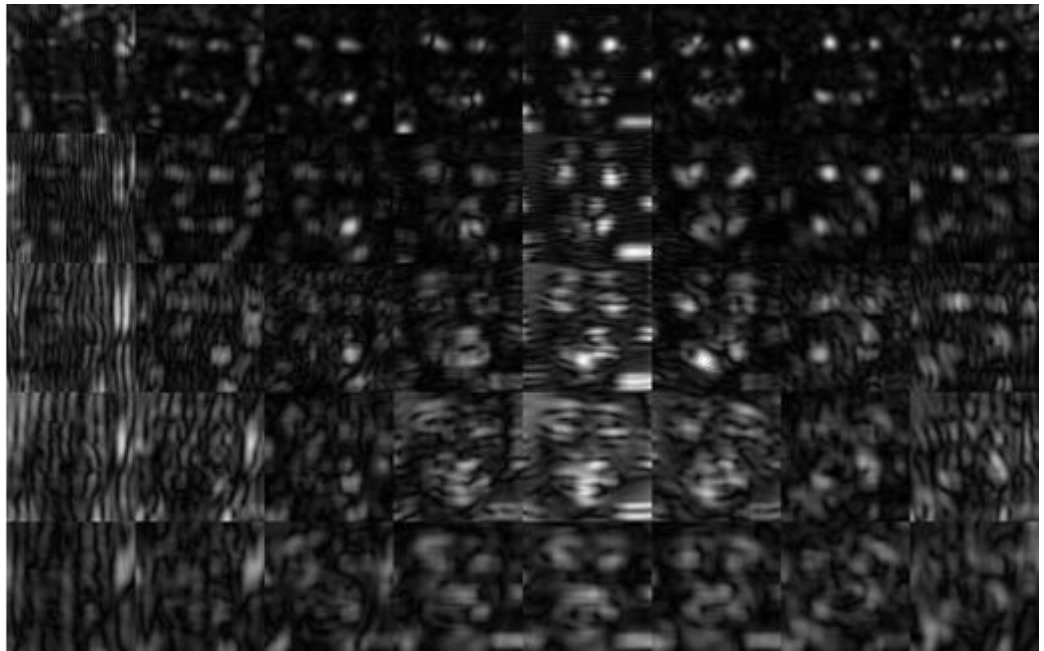
$$RG(c, s) = |\,(\mathrm{R}(c) - \mathrm{G}(c))\ \Theta\ (\mathrm{G}(s) - \mathrm{R}(s))\,|$$
$$BY(c, s) = |\,(\mathrm{B}(c) - \mathrm{Y}(c))\ \Theta\ (\mathrm{Y}(s) - \mathrm{B}(s))\,|$$

Feature Maps



3. Orientation maps (24 maps)
   - Gabor filters at 0º, 45º, 90º, and 135º
   - Also at different scales

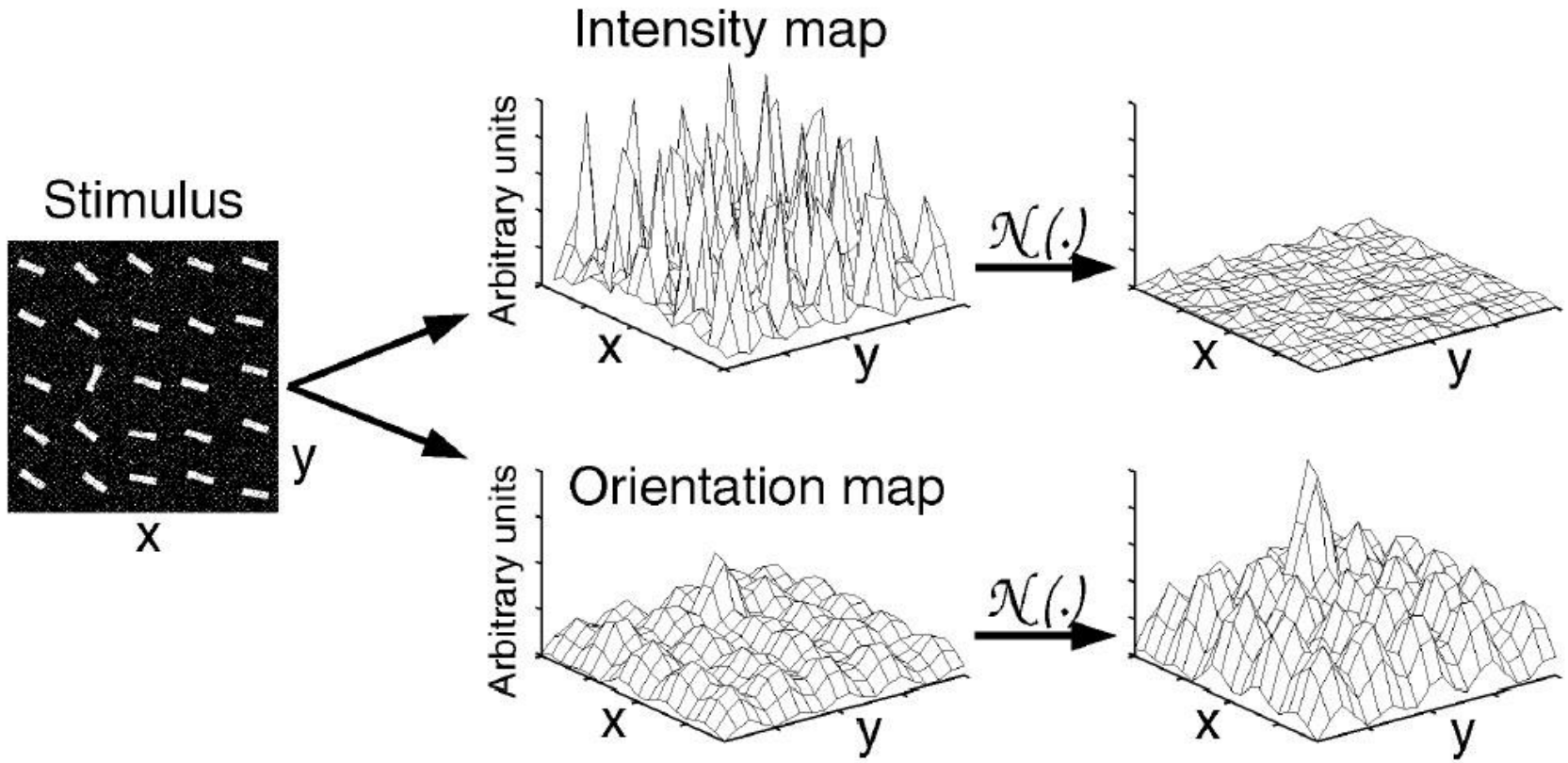
→ Total of 42 feature maps are combined into the saliency map



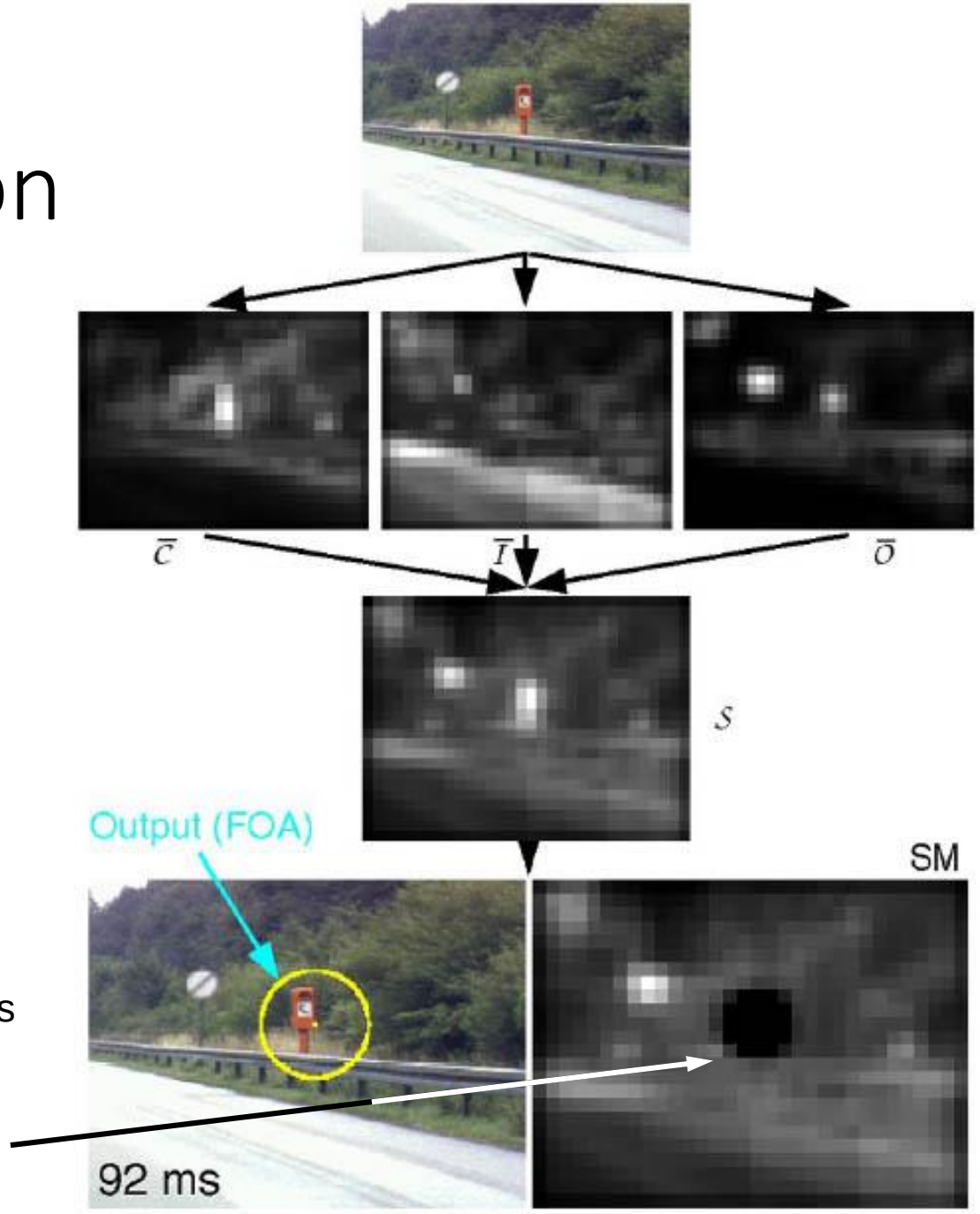
Computing Saliency Map

- Feature maps combined into three “conspicuity maps”
  - Intensity (I)
  - Color (C)
  - Orientation (O)
- Before they are combined they need to be normalized

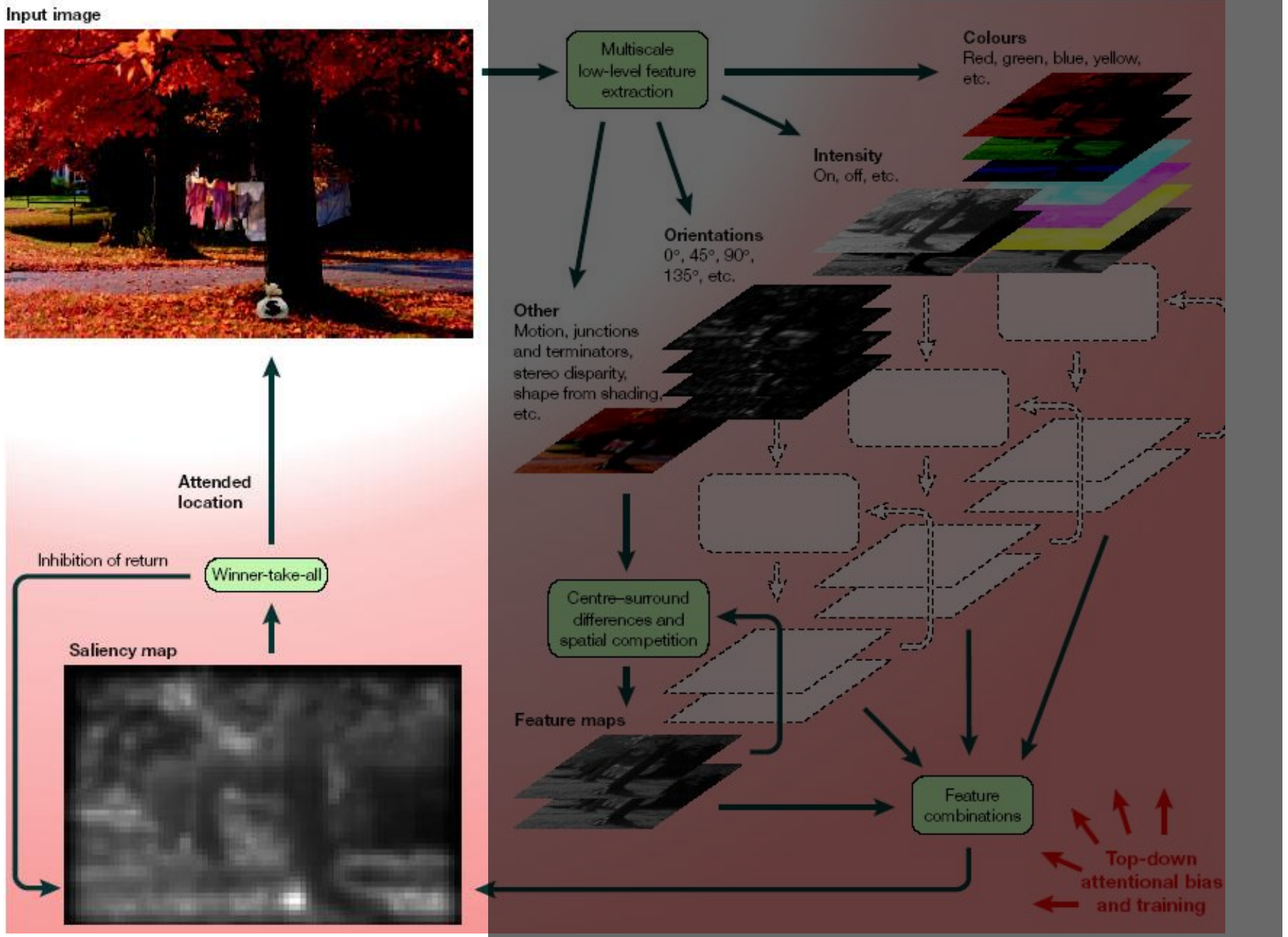
Normalization Operator



Example of operation

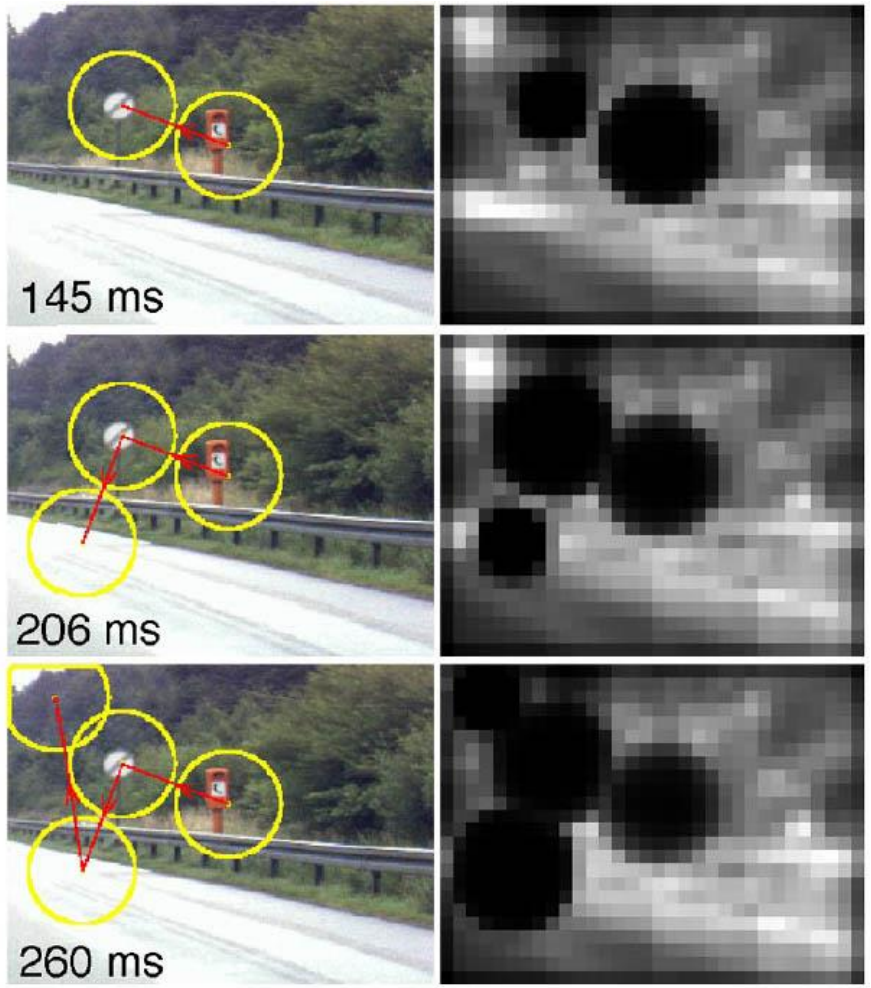


Leaky *integrate-and-fire* neurons
“Inhibition of return”


Input image
Multiscale low-level feature extraction
Colours
Red, green, blue, yellow, etc.
Intensity
On, off, etc.
Orientations
0°, 45°, 90°, 135°, etc.
Other
Motion, junctions and terminators, stereo disparity, shape from shading, etc.
Attended location
Inhibition of return
Winner-take-all
Saliency map
Centre–surround differences and spatial competition
Feature maps
Feature combinations
Top-down attentional bias and training

Example of operation

- Using 2D “winner-take-all” neural network at scale 4
- FOA shifts every 30-70 ms
- Inhibition lasts 500-900 ms

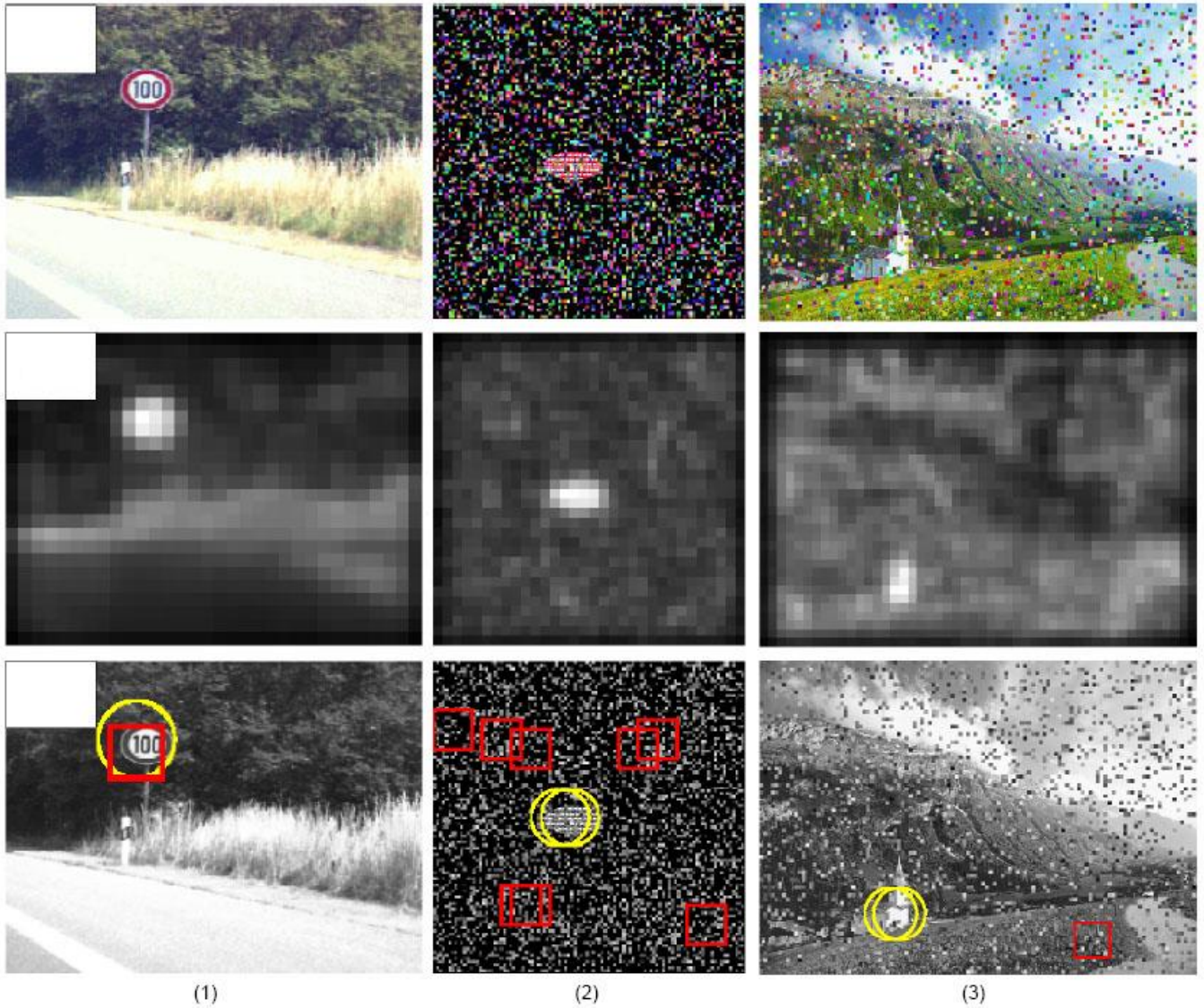


Results

Image

Saliency Map

High saliency Locations (yellow circles)



Results

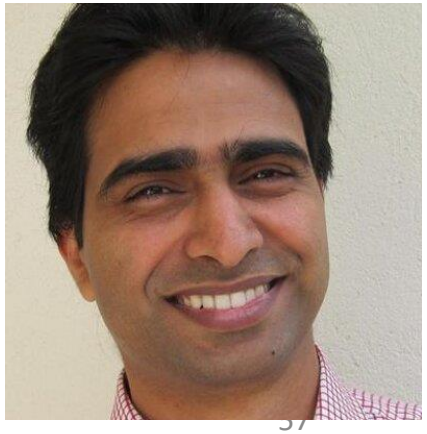
- Tested on both synthetic and natural images
- Typically finds objects of interest, e.g. traffic signs, faces, flags, buildings…
- Generally robust to noise (less to multicoloured noise)

Summary

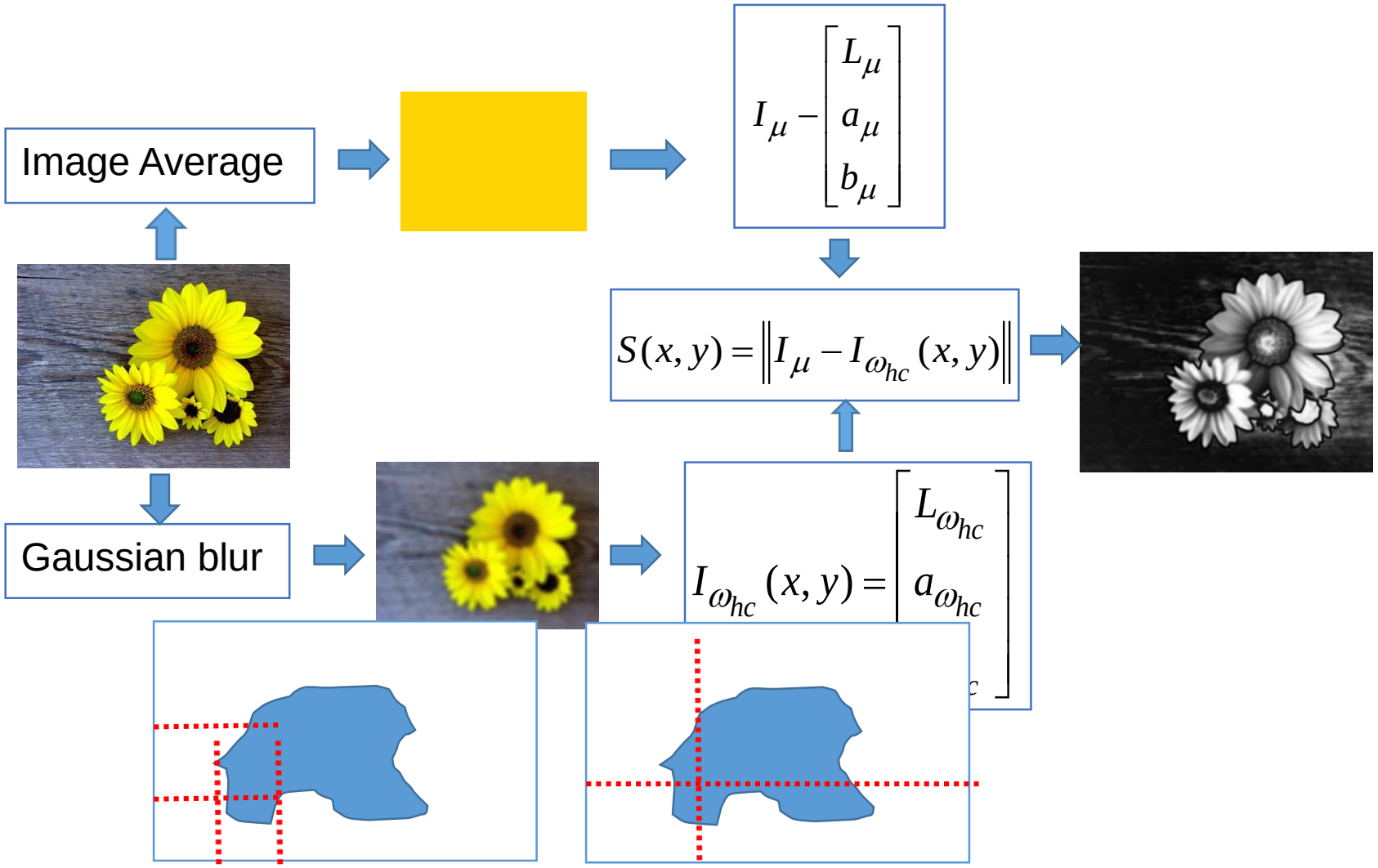
- Basic idea:
  - Find multiple saliency measures in parallel
  - Normalize
  - Combine them to a single map
  - Use 2D integrate-and-fire layer of neurons to determine position of FOA
- Model appears to work accurately and robustly (but difficult to evaluate)
- Can be extended with other feature maps

Frequency-tuned Salient Region Detection

Radhakrishna Achanta[†], Sheila Hemami[‡], Francisco Estrada[†], and Sabine Süsstrunk[†]
[†]School of Computer and Communication Sciences (IC)
Ecole Polytechnique Fédérale de Lausanne (EPFL), CH-1015, Switzerland.
[radhakrishna.achanta,francisco.estrada,sabine.susstrunk]@epfl.ch
[‡]School of Electrical and Computer Engineering
Cornell University, Ithaca, NY 14853, U.S.A.
hemami@ece.cornell.edu



Frequency-tuned



Context-Aware Saliency Detection

Stas Goferman
Technion
stasix@gmail.com

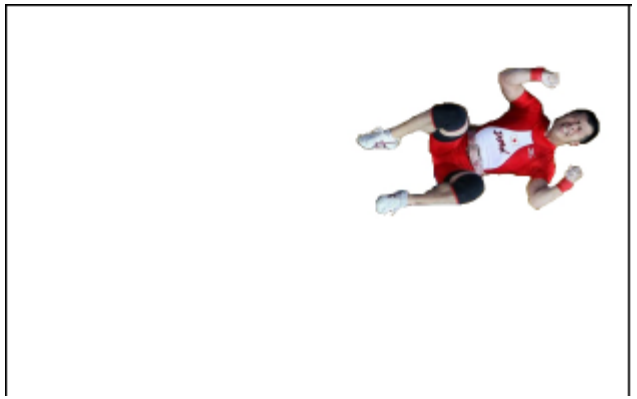
Lihi Zelnik-Manor
Technion
lihi@ee.technion.ac.il

Ayellet Tal
Technion
ayellet@ee.technion.ac.il

Context aware saliency

One algorithm that do so, uses a new kind of definition for saliency, were the salient part in the picture is not only a single object but it's surroundings too.
This definition is named Context aware saliency

What do you see?



And now?

Context aware saliency algorithm

(1) Local low-level considerations,
including factors such as contrast and color

(2) Global considerations, which suppress frequently Occurring features

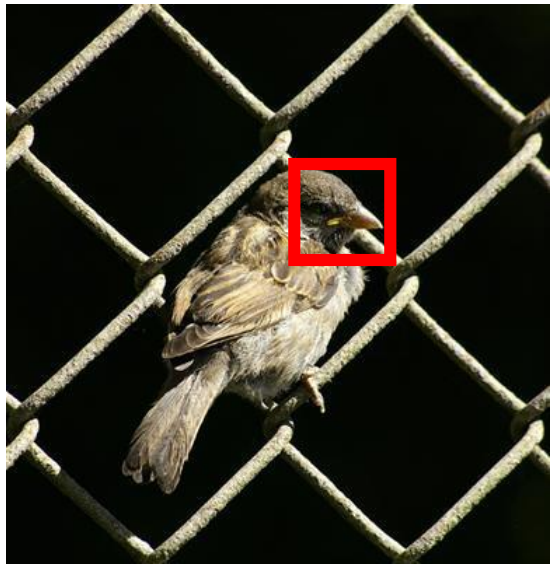
(3) Visual organization rules, which state that visual Forms may possess one or several centers of attention.

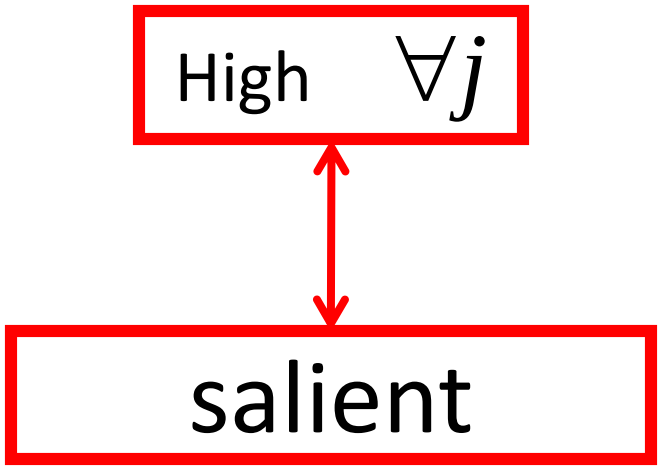
(4) High- level factors, such as priors on the salient Object location.

Context-Aware

- Distance between a pair of patches:

$$d(p_i, p_j) = \frac{d_{color}(p_i, p_j)}{1 + c \cdot d_{position}(p_i, p_j)}$$



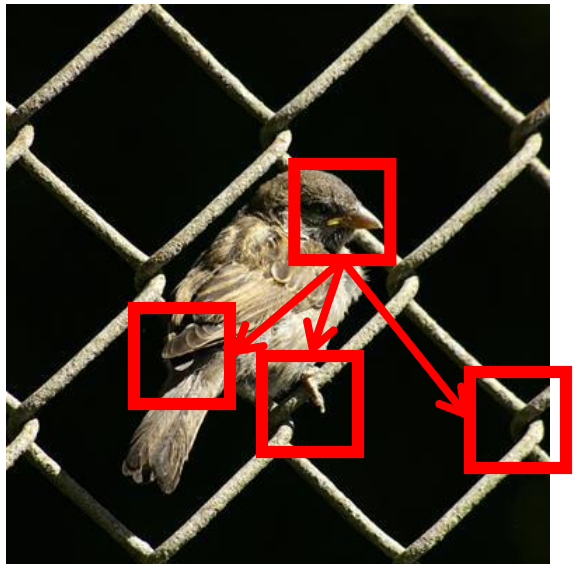


Context-Aware

- Distance between a pair of patches:

$$S_i^r = 1 - \exp\left[-\frac{1}{K}\sum_{k=1}^{K} d(p_i^r, q_j^r)\right]$$

$q_k^r =$ K most similar patches at scale r



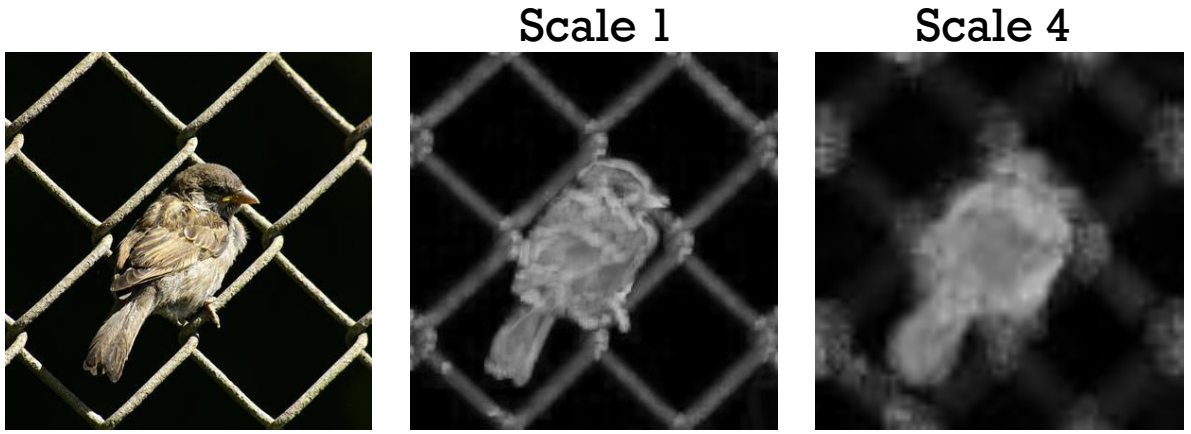
High for K most similar

Saliency

Context-Aware

- Salient at:
  - Multiple scales ➔ foreground
  - Few scales ➔ background

$$\bar{S}_i = \frac{1}{M} \sum_{r=r_1}^{r_M} S_i^r$$



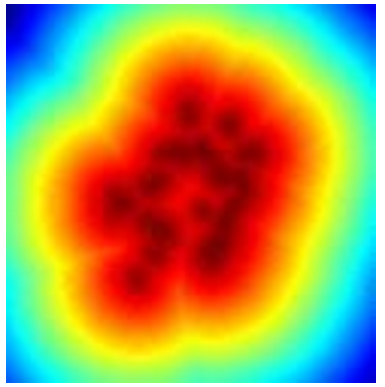
Context-Aware

- Foci = $\bar{S}_i > 0.8$

- Include distance map

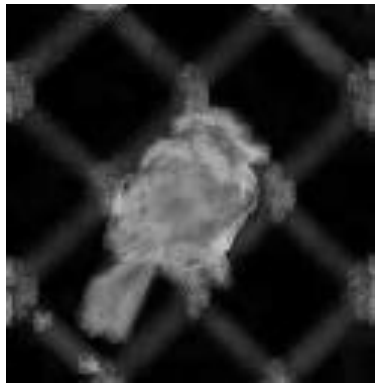
$$\hat{S}_i = \bar{S}_i \left(1 - d_{foci}(i)\right)$$

$1 - d_{foci}(i)$



x

$\bar{S}_i$



Q&A