THE UNIVERSITY OF BRITISH COLUMBIA

Topics in AI (CPSC 532S):
Multimodal Learning with Vision, Language and Sound

Lecture 3: Introduction to Computer Vision

Computer vs. **human** vision

objects, scenes, people

Human Vision

*slide from V. Ordonex

Computer vs. **human** vision

objects, scenes, people

0	3	2	5	4	7	6	9	8
3	0	1	2	3	4	5	6	7
2	1	0	3	2	5	4	7	6
5	2	3	0	1	2	3	4	5
4	3	2	1	0	3	2	5	4
7	4	5	2	3	0	1	2	3
6	5	4	3	2	1	0	3	2
9	6	7	4	5	2	3	0	1
8	7	6	5	4	3	2	1	0

matrix of numbers

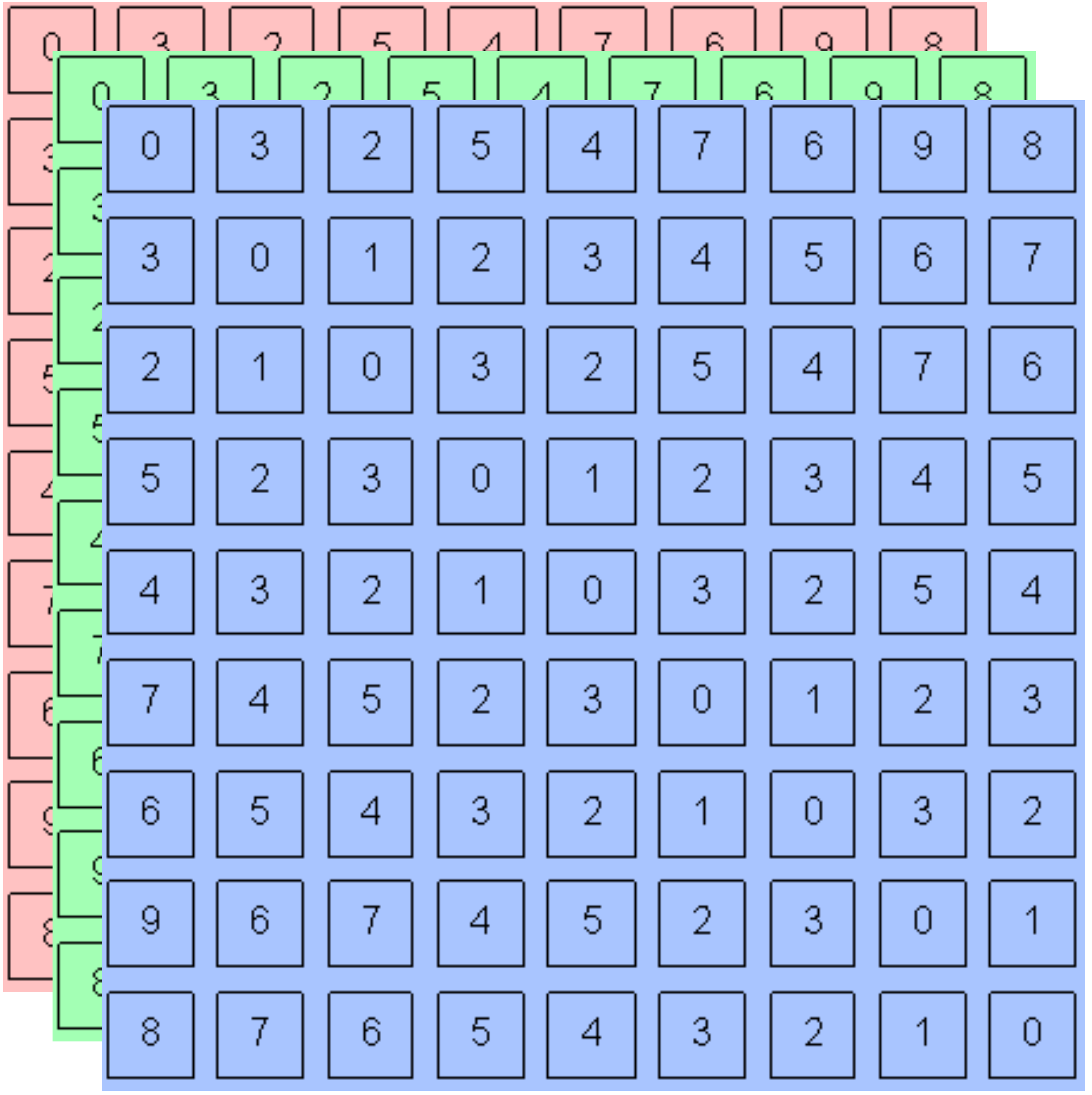
Human Vision

Computer Vision

*slide from V. Ordonex

Computer vs. **human** vision

objects, scenes, people



tensor of numbers

Human Vision

Computer Vision

*slide from V. Ordonex

Computer Vision

Computer vision studies the **tools and theories** that enable the design of machines that can **extract useful information from imagery data** (images and videos) toward the goal of **interpreting the world**

*curtesy of Peter Meer

Vision is Amazing Feat of **Natural Intelligence**

~ 55% of **cerebral cortex** in humans (13 billion neurons) are devoted to vision

more human brain devoted to vision than anything else

Challenges: Viewpoint invariance

Michelangelo 1475-1564

*slide credit Fei-Fei, Fergus & Torralba

Challenges: Lighting

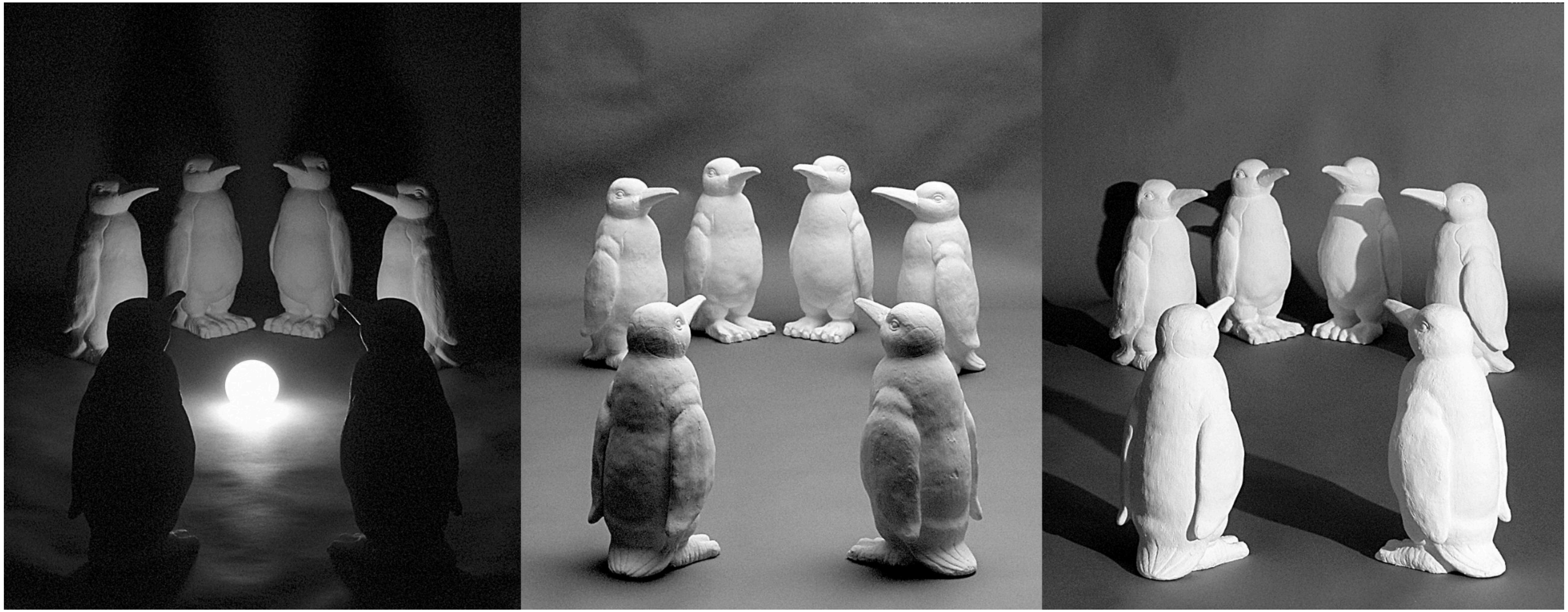


*image credit J. Koenderink

Challenges: Scale

*slide credit Fei-Fei, Fergus & Torralba

Challenges: Deformation

*image credit Peter Meer

Challenges: Occlusions

Rene Magritte 1965

Challenges: Background clutter

Kilmeny Niland 1995

Challenges: Local ambiguity and context

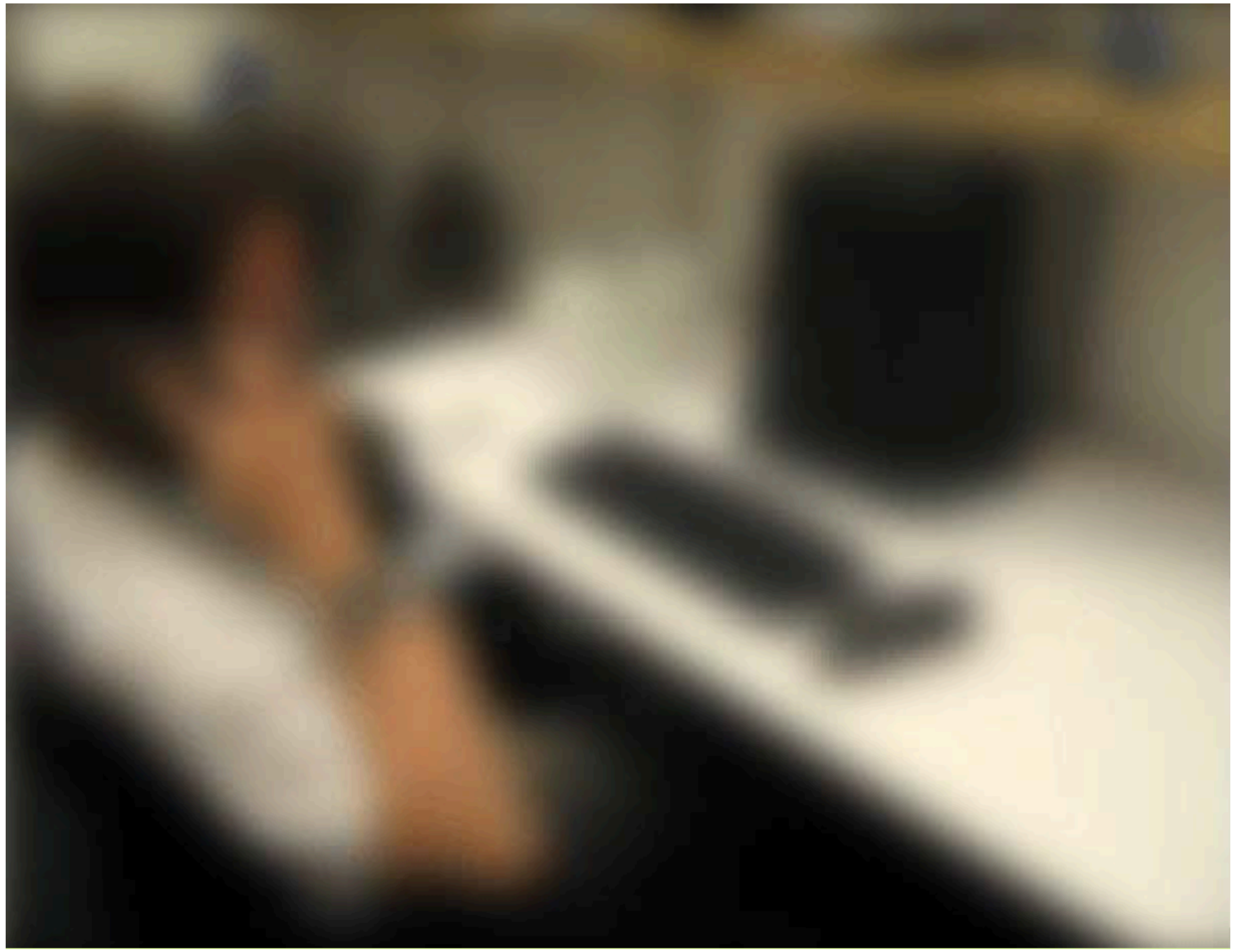


*image credit Fergus & Torralba

Challenges: Local ambiguity and context

*image credit Fergus & Torralba

Challenges: Motion

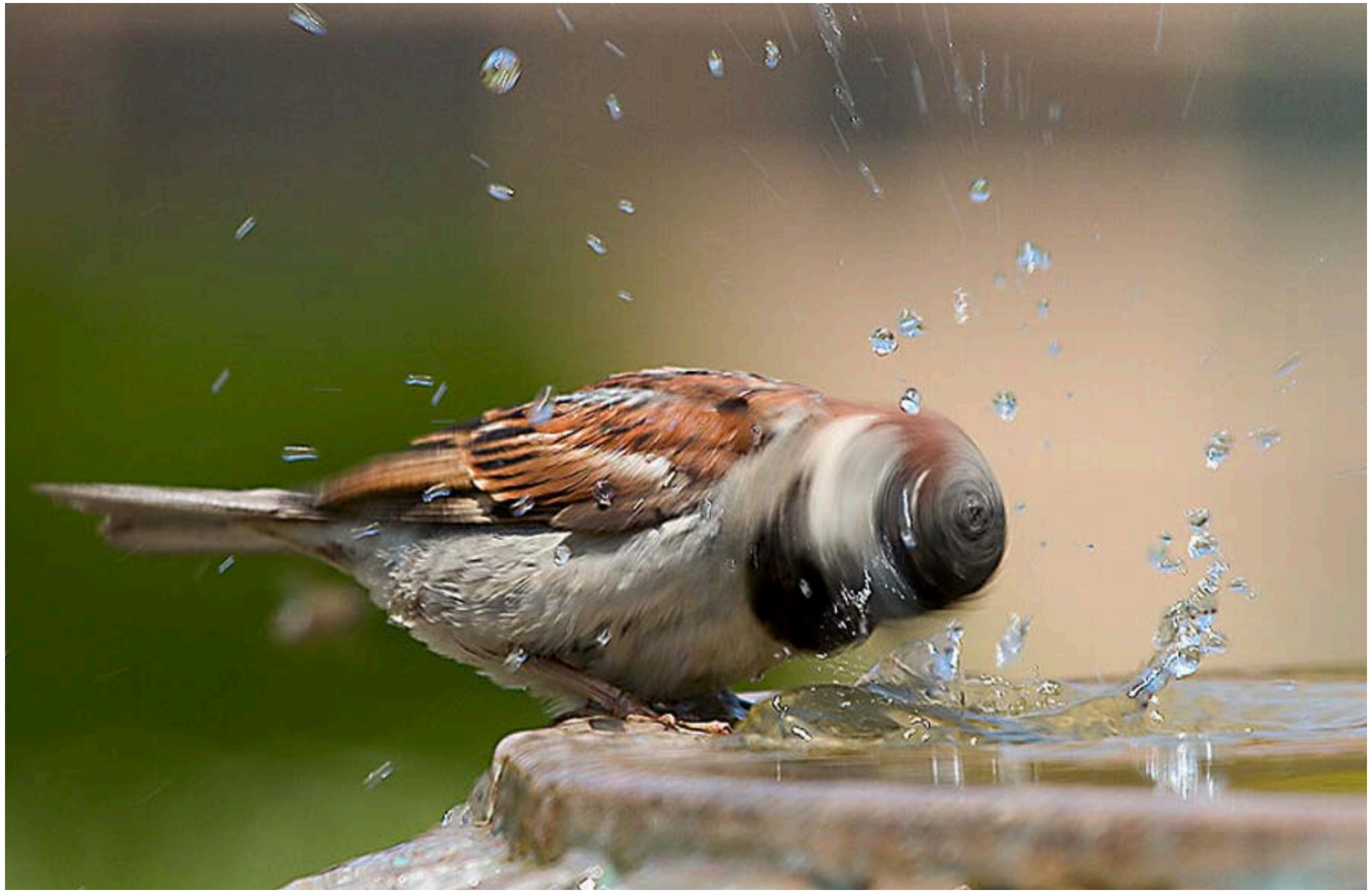


*image credit Peter Meer

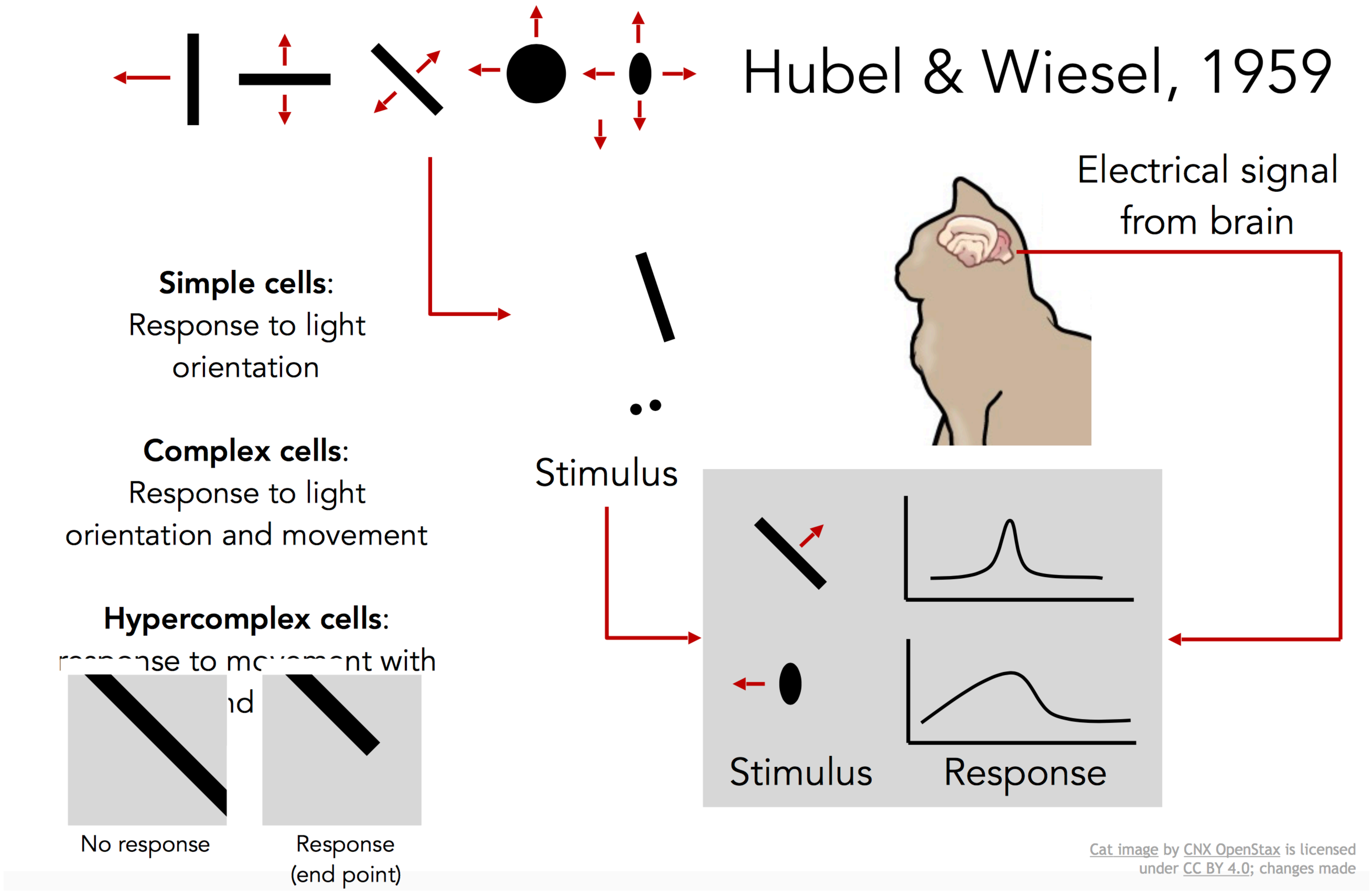
Challenges: Object inter-class variation






*slide credit Fei-Fei, Fergus & Torralba

Human vision …



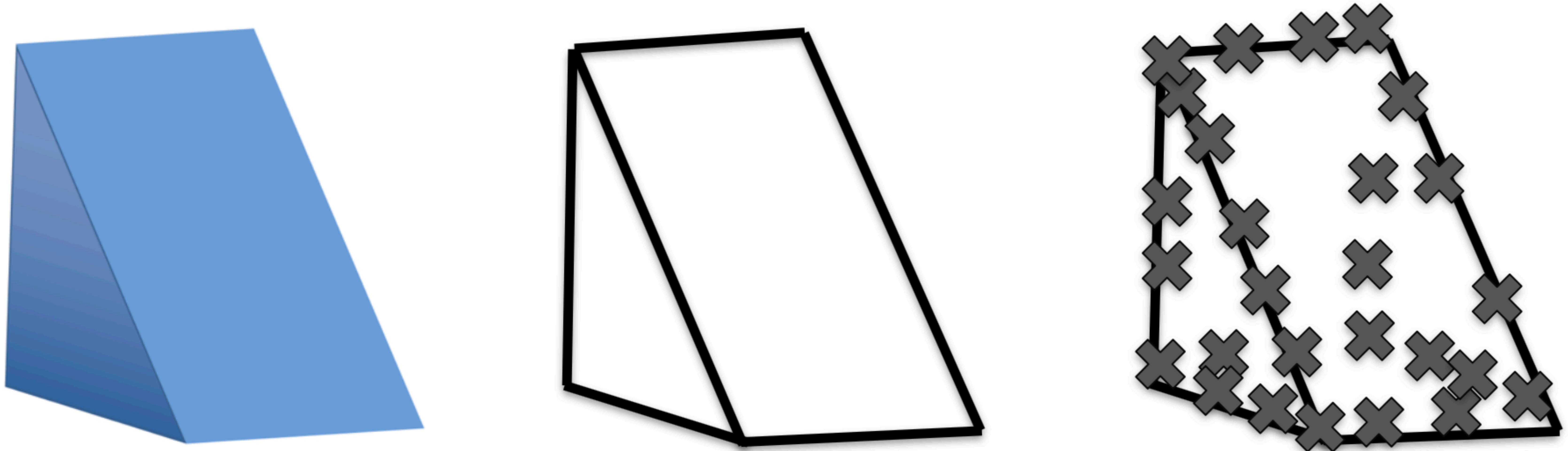
* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Computer vision … the beginning …

Larry Roberts

Blocks World. first thesis in computer vision, 1963

“the perception of **solid objects** is a process which can be based on the **properties of three-dimensional** transformations and **the laws of nature**"



(a) Original picture (b) Differentiated picture (c) Feature points selected

Computer vision … the beginning …

Larry Roberts

Blocks World. first thesis in computer vision, 1963

“the perception of **solid objects** is a process which can be based on the **properties of three-dimensional** transformations and **the laws of nature**"

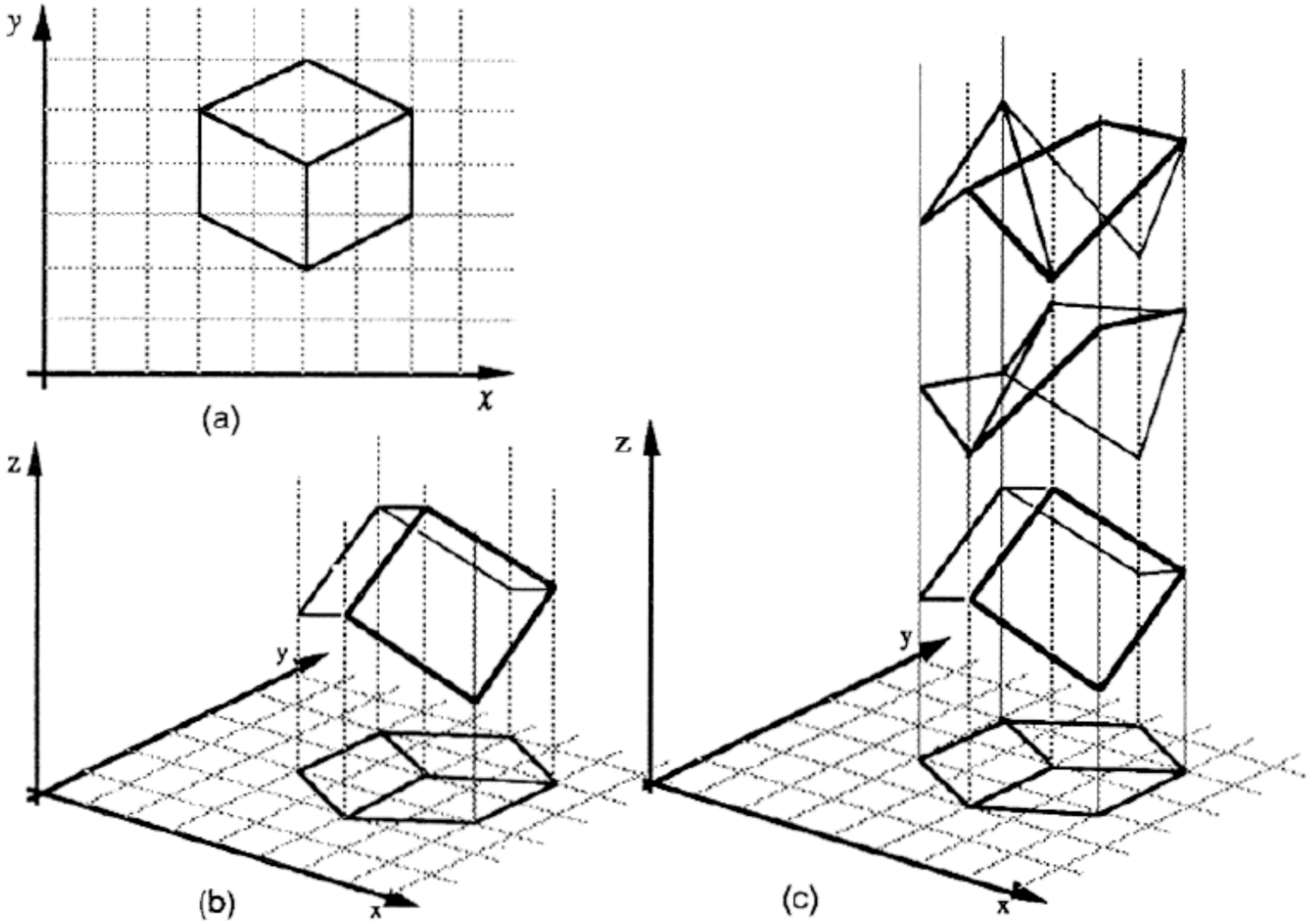
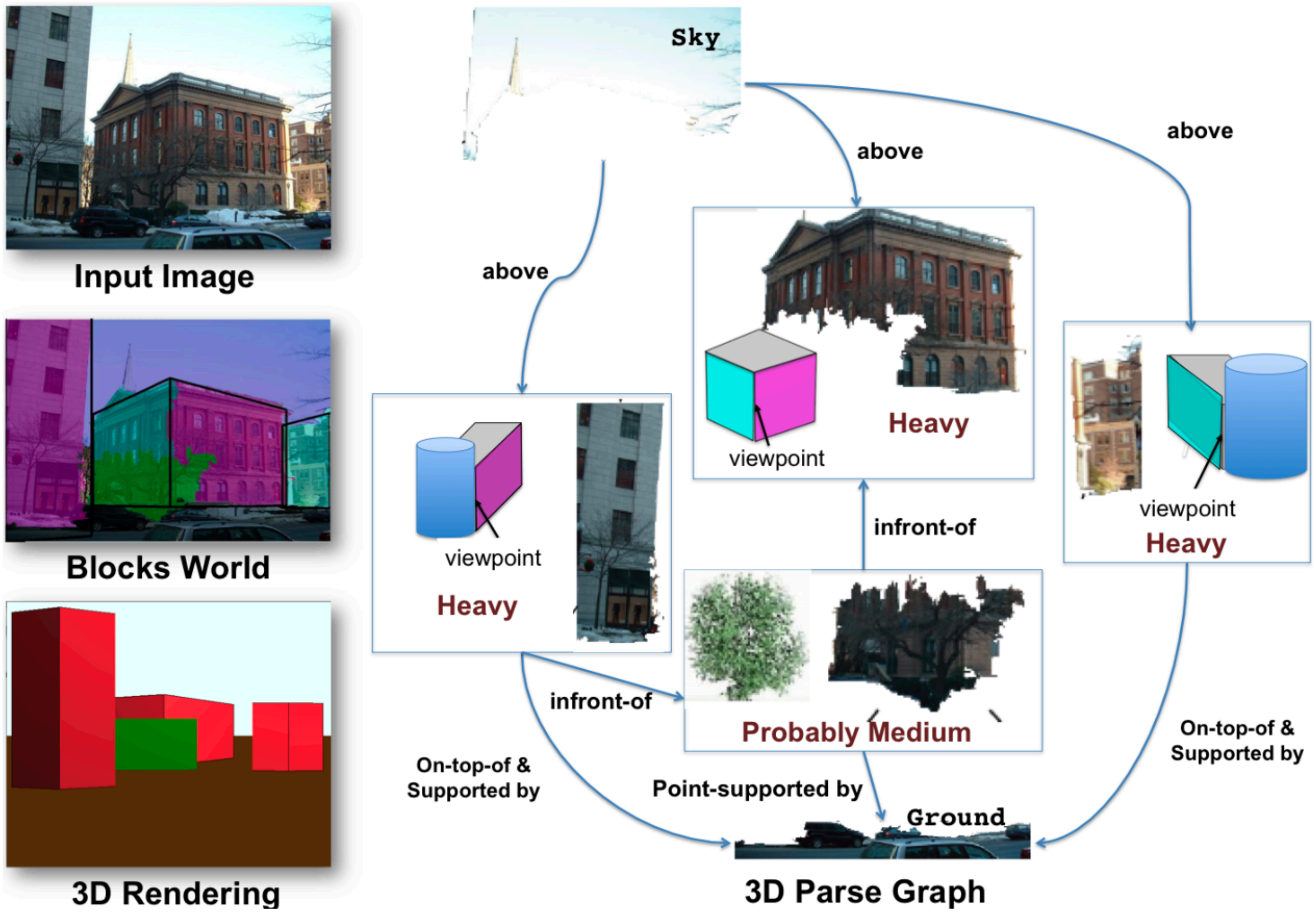


Figure 1. (a) A line drawing provides information only about the x, y coordinates of points lying along the object contours. (b) The human visual system is usually able to reconstruct an object in three dimensions given only a single 2D projection (c) Any planar line-drawing is geometrically consistent with infinitely many 3D structures.

[Since & Adelson, 1993]

Computer vision … the beginning …



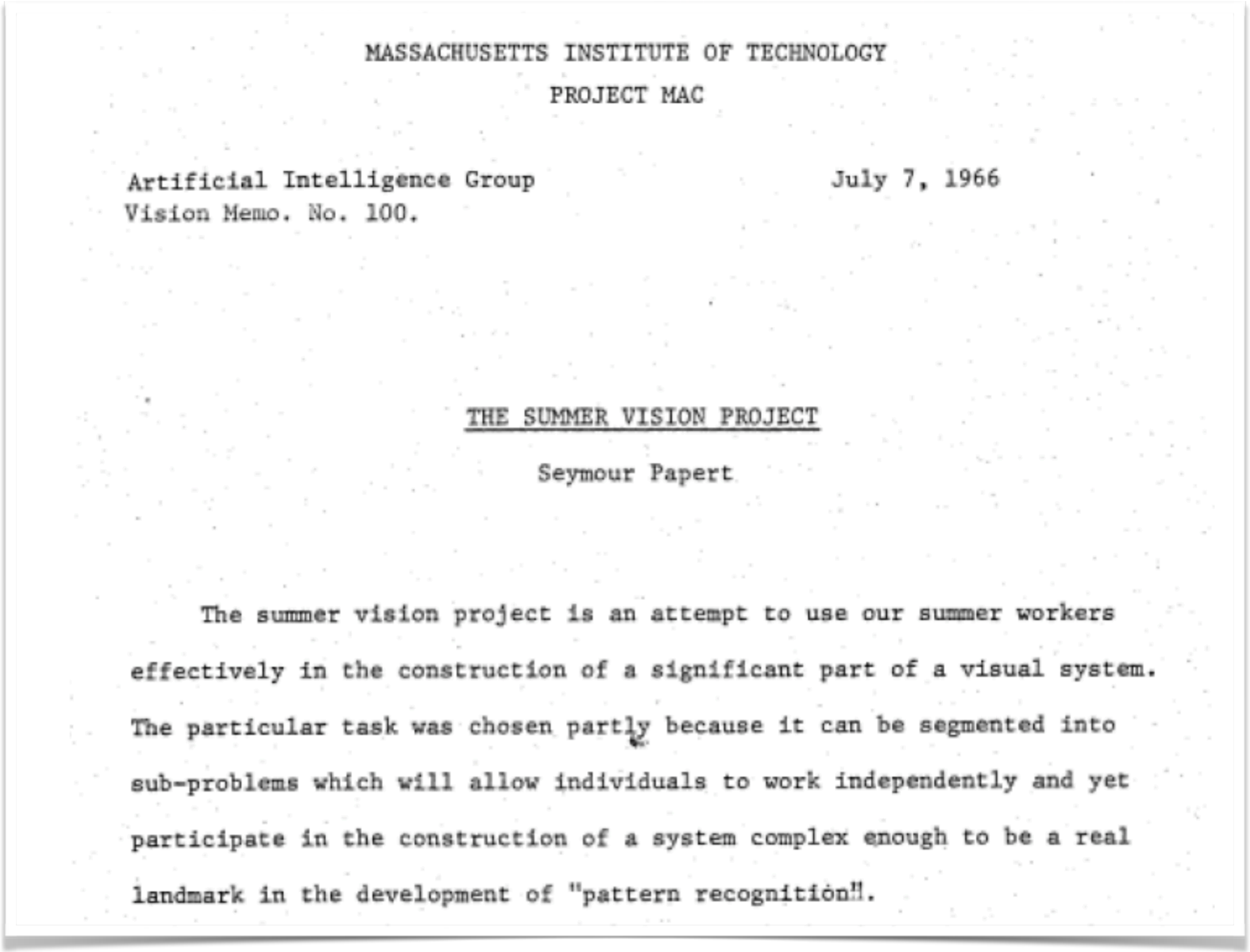
Static Equilibrium: Forces and torques acting on a block should cancel each other out.

Support Force Constraint: Supporting object should have enough strength to provide contact reactionary forces

Volumetric Constraints: All objects in the world must have finite volume & cannot penetrate each other

[Gupta, Efros & Hebert, 2010]

Computer vision … the beginning …



MASSACHUSETTS INSTITUTE OF TECHNOLOGY
PROJECT MAC

Artificial Intelligence Group July 7, 1966
Vision Memo. No. 100.

THE SUMMER VISION PROJECT

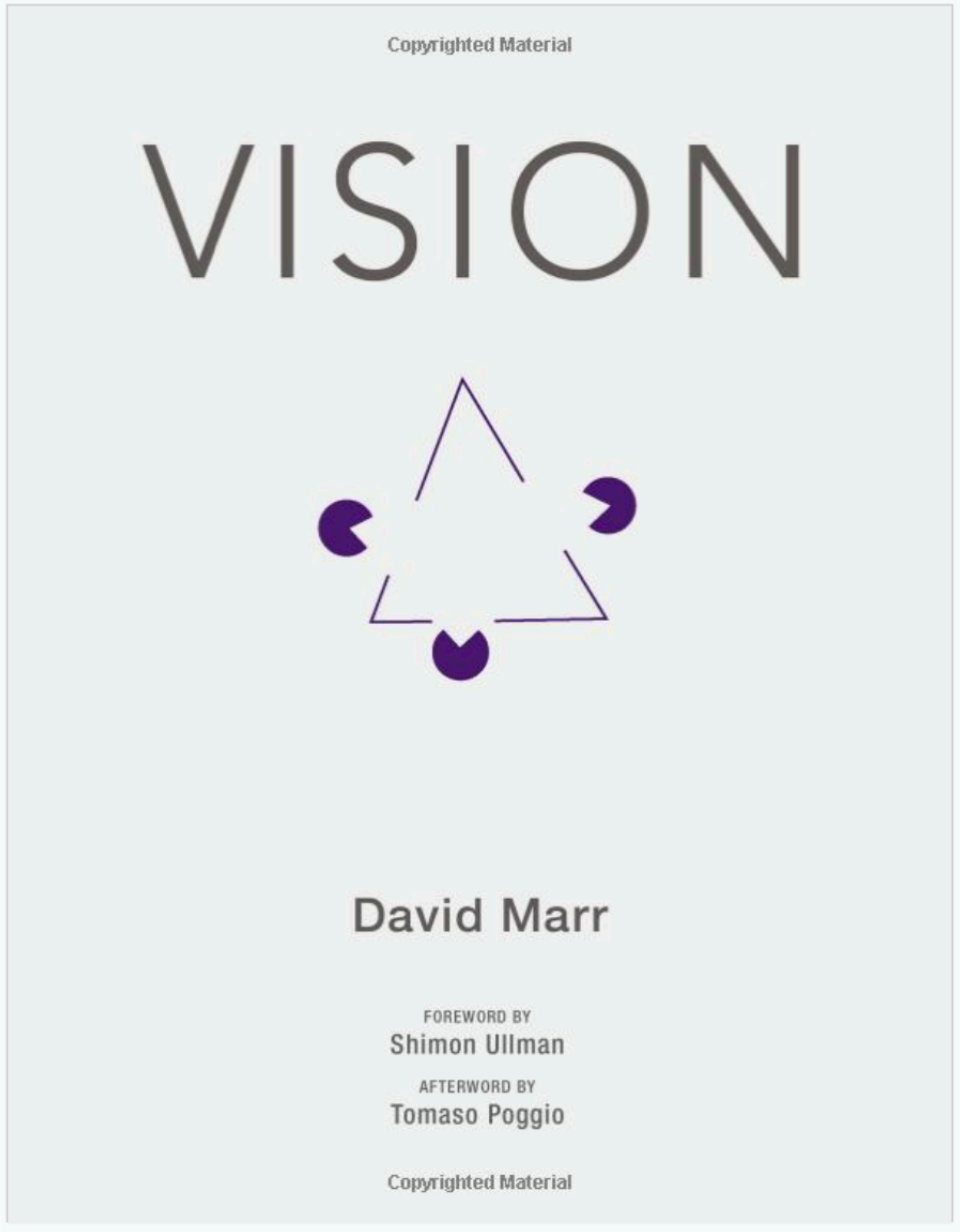
Seymour Papert

The summer vision project is an attempt to use our summer workers effectively in the construction of a significant part of a visual system. The particular task was chosen partly because it can be segmented into sub-problems which will allow individuals to work independently and yet participate in the construction of a system complex enough to be a real landmark in the development of "pattern recognition".

In 1966, Marvin Minsky at MIT asked his undergraduate student Gerald Jay Sussman to "spend the summer linking a camera to a computer and getting the computer to describe what it saw"

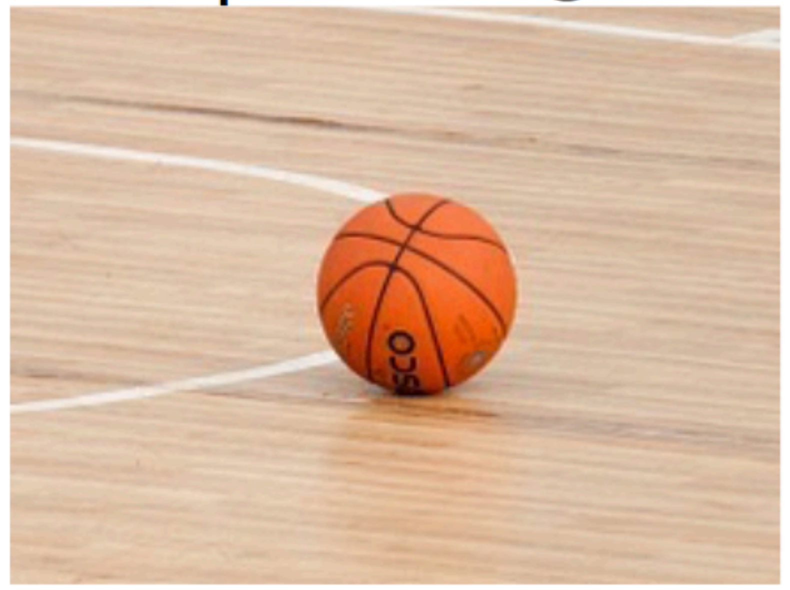
[Szeliski 2009, Computer Vision]

David **Marr**, 1970s

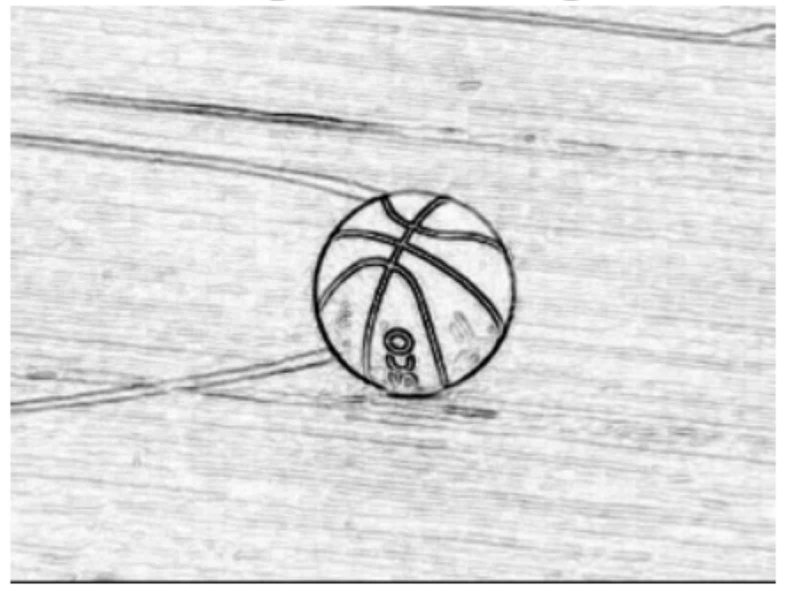


David **Marr**, 1970s

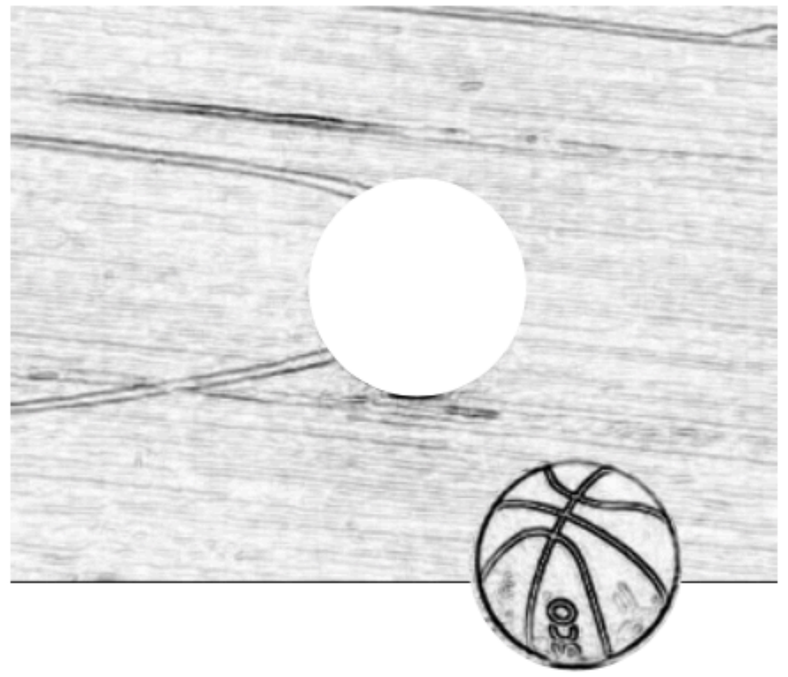
Input image



Edge image



2 ½-D sketch



3-D model

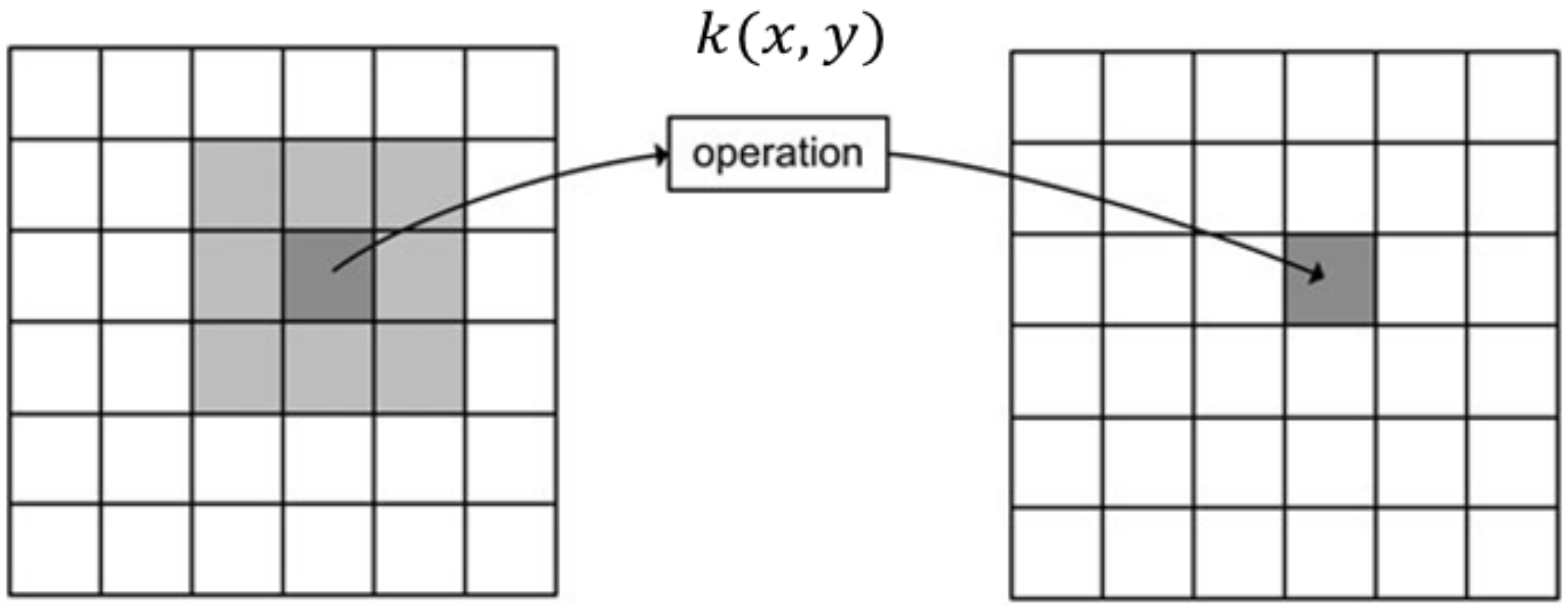
Input Image	→	Primal Sketch	→	2 ½-D Sketch	→	3-D Model Representation
Perceived intensities		Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries		Local surface orientation and discontinuities in depth and in surface orientation		3-D models hierarchically organized in terms of surface and volumetric primitives

[Stages of Visual Representation, **David Marr**]

* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

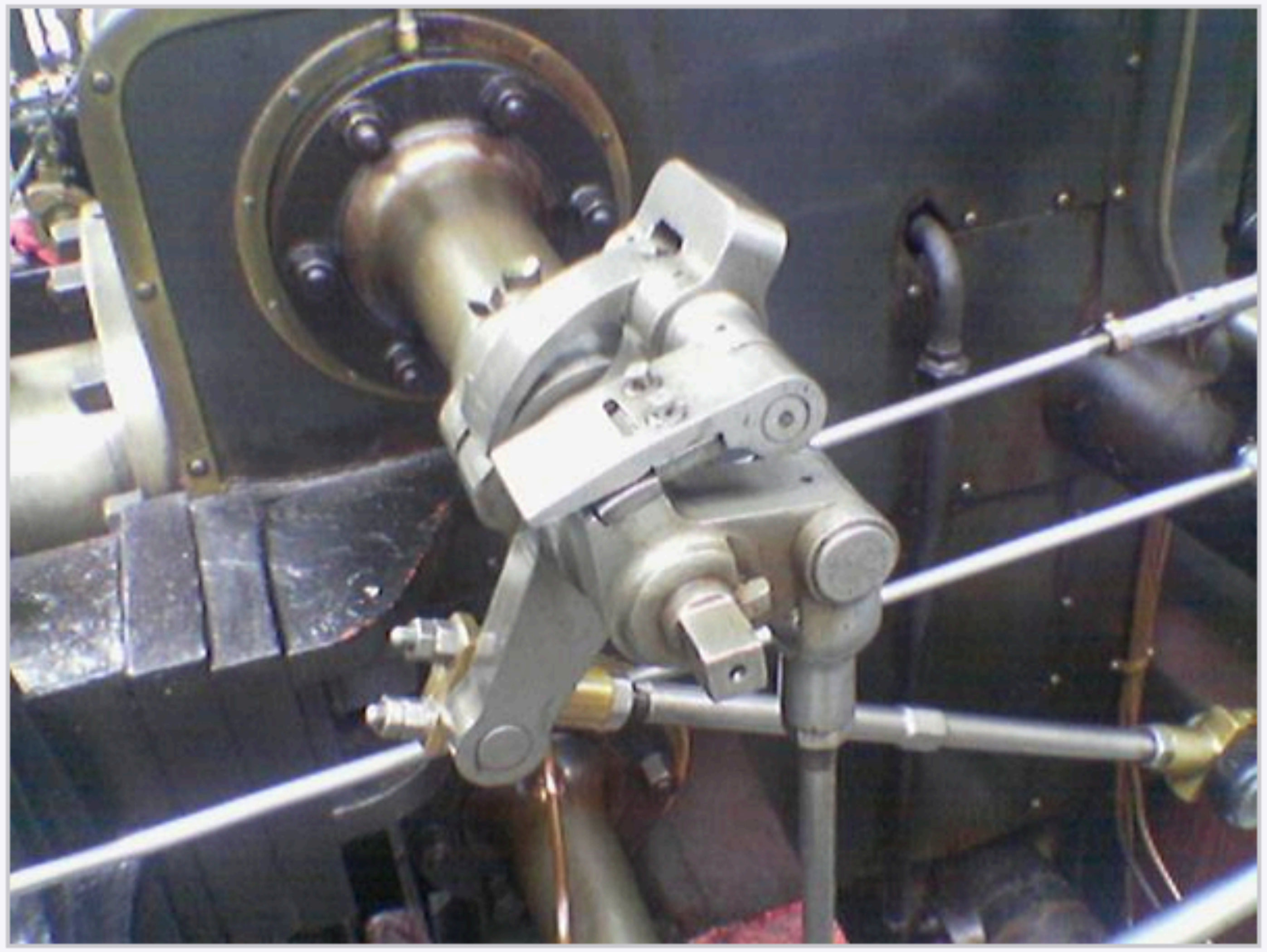
Edges

1	0	-1
2	0	-2
1	0	-1

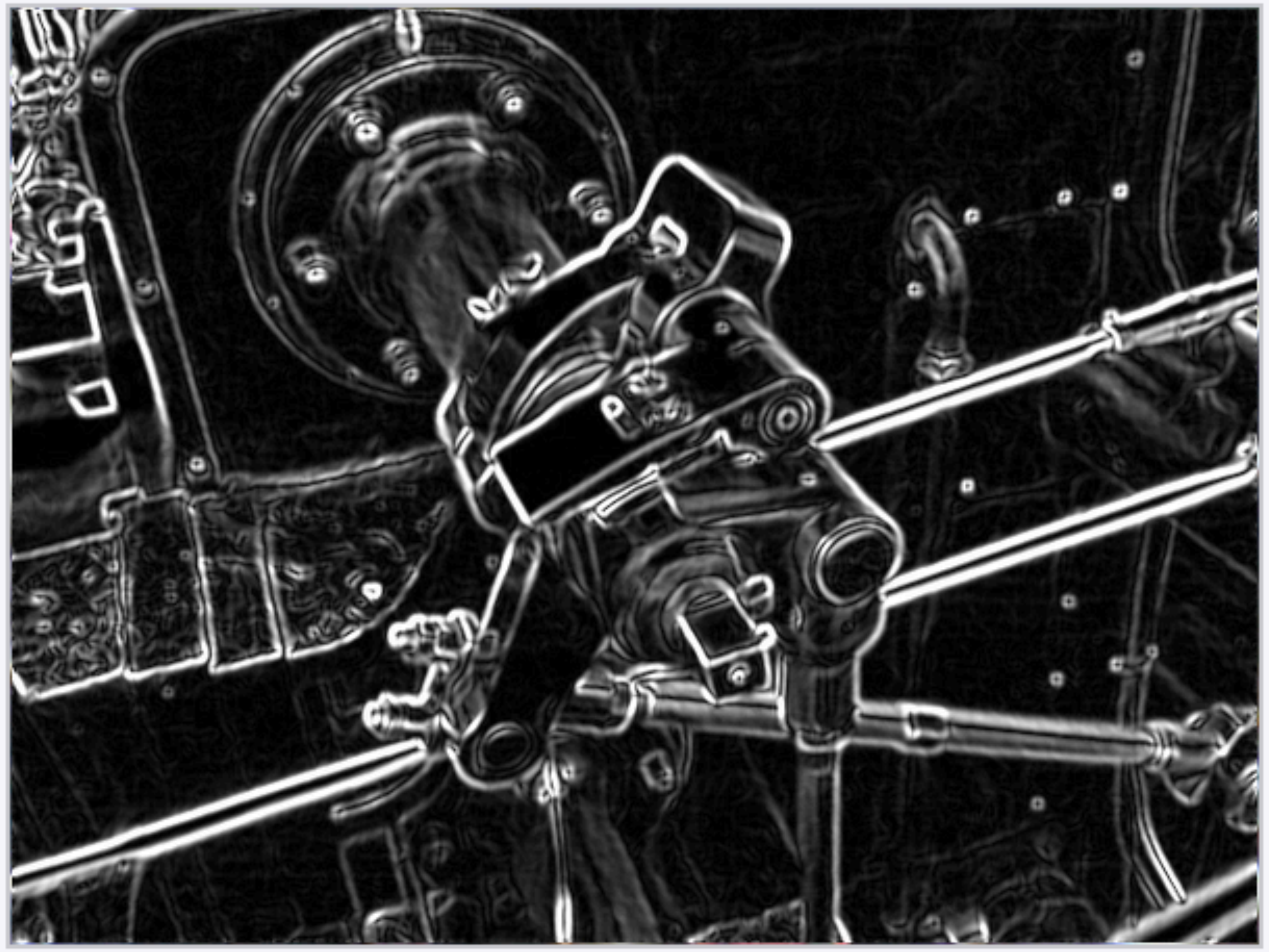


Input image
f(x,y)

Output image
g(x,y)



A color picture of a steam engine

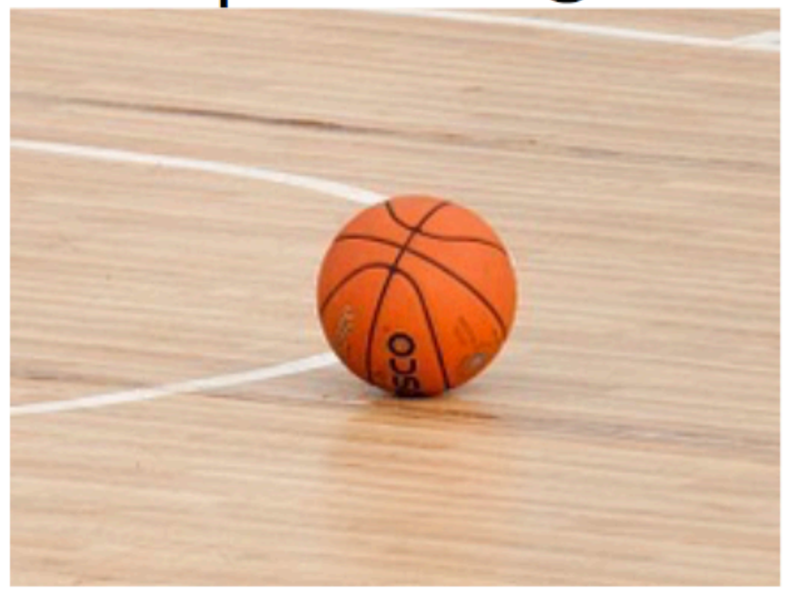


The Sobel operator applied to that image

*content from V. Ordonex

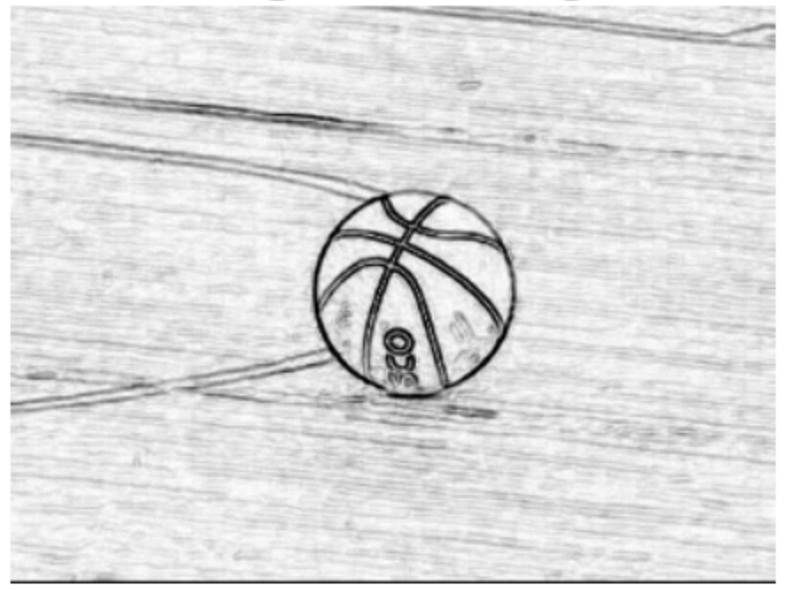
David **Marr**, 1970s

Input image

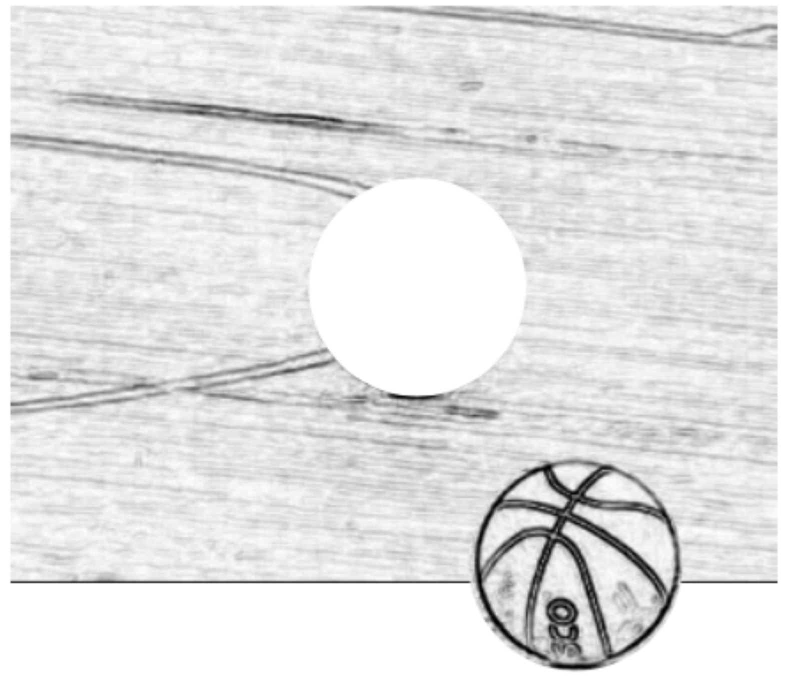


This image is CC0 1.0 public domain

Edge image



2 ½-D sketch



3-D model

This image is CC0 1.0 public domain

Input Image	Primal Sketch	2 ½-D Sketch	3-D Model Representation
Perceived intensities	Zero crossings, blobs, edges, bars, ends, virtual lines, groups, curves boundaries	Local surface orientation and discontinuities in depth and in surface orientation	3-D models hierarchically organized in terms of surface and volumetric primitives

[Stages of Visual Representation, **David Marr**]

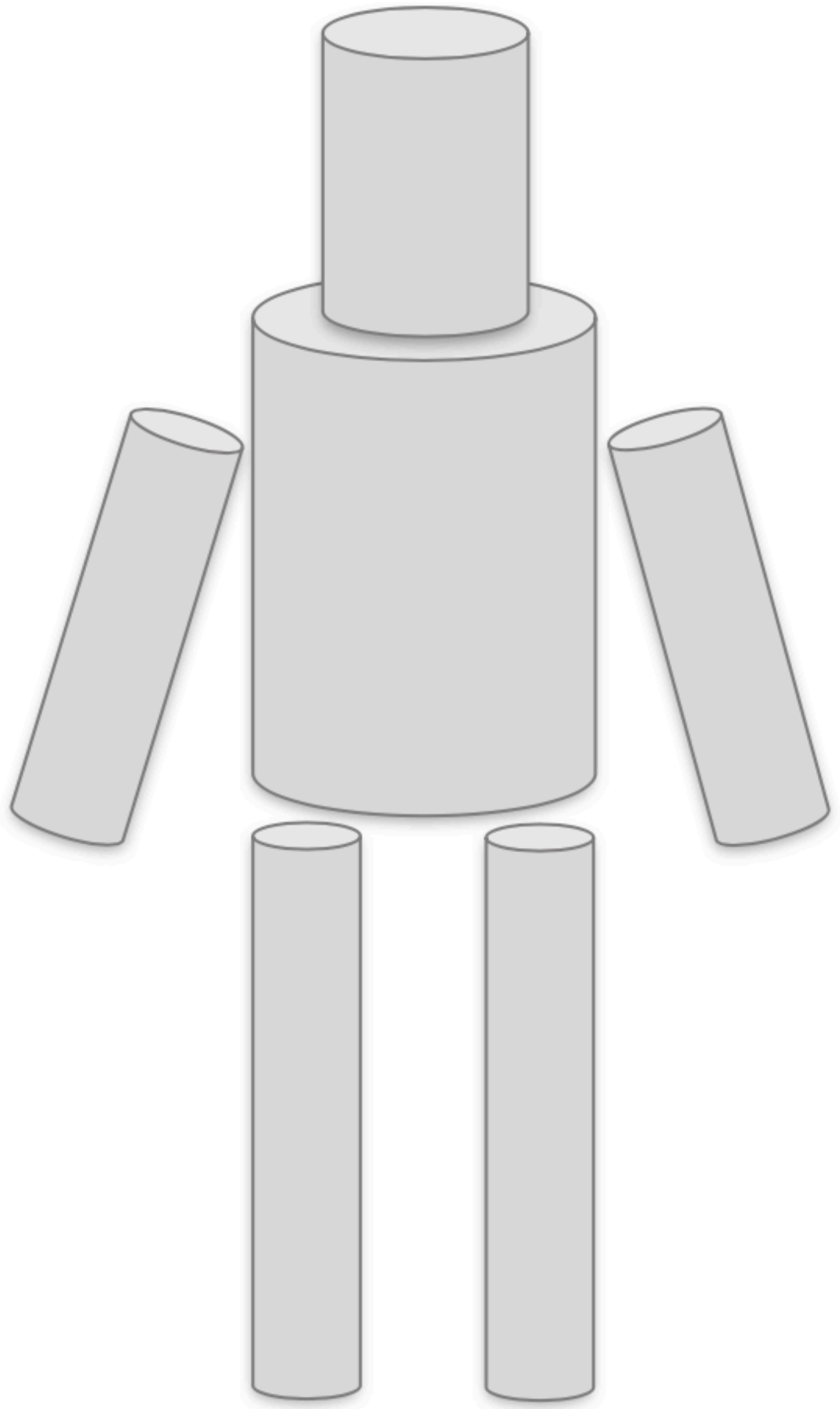
* slide from Fei-Dei Li, Justin Johnson, Serena Yeung, **cs231n Stanford**

Segmentation - GraphCuts

[Shi & Malik, 2000]

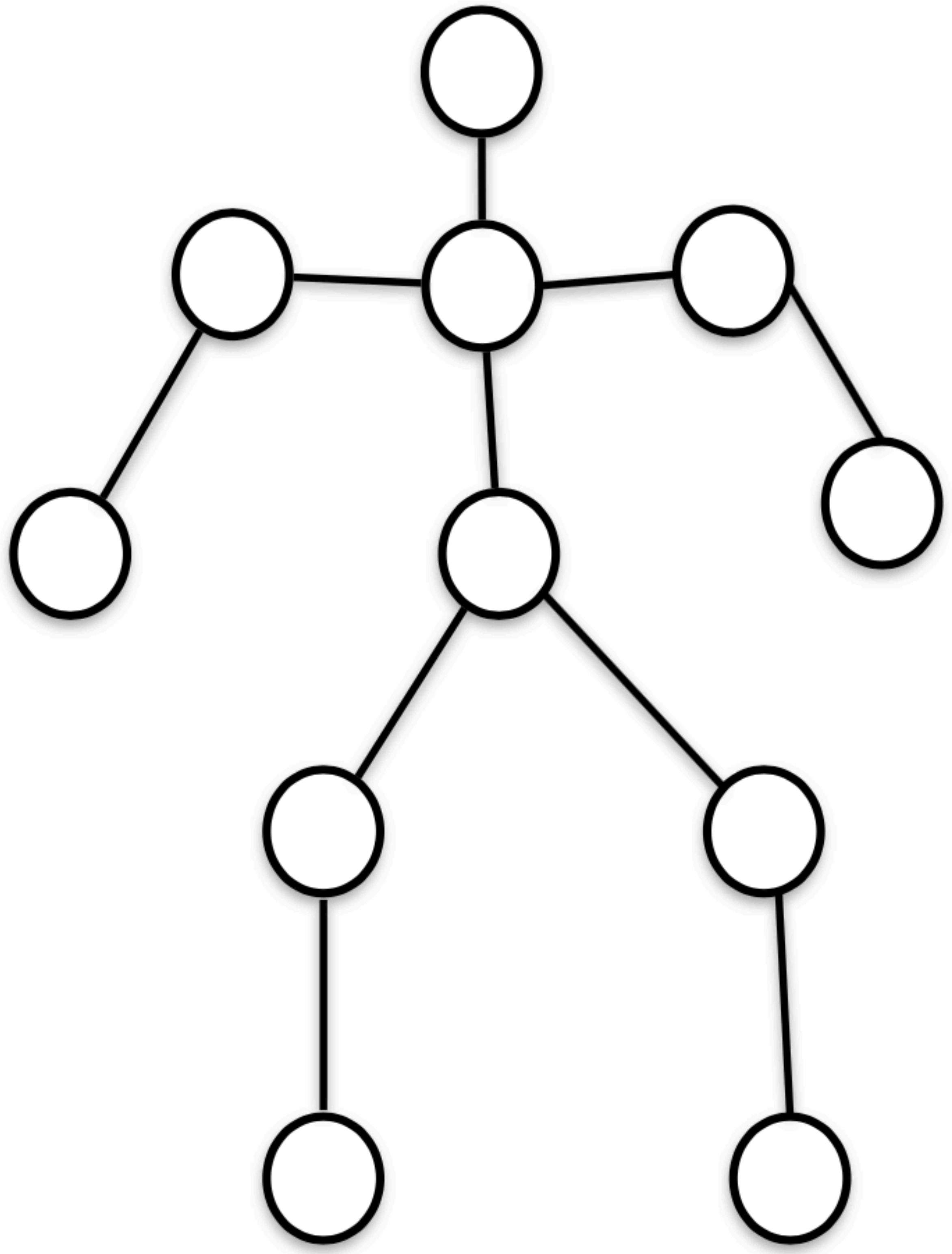
Part-based Models

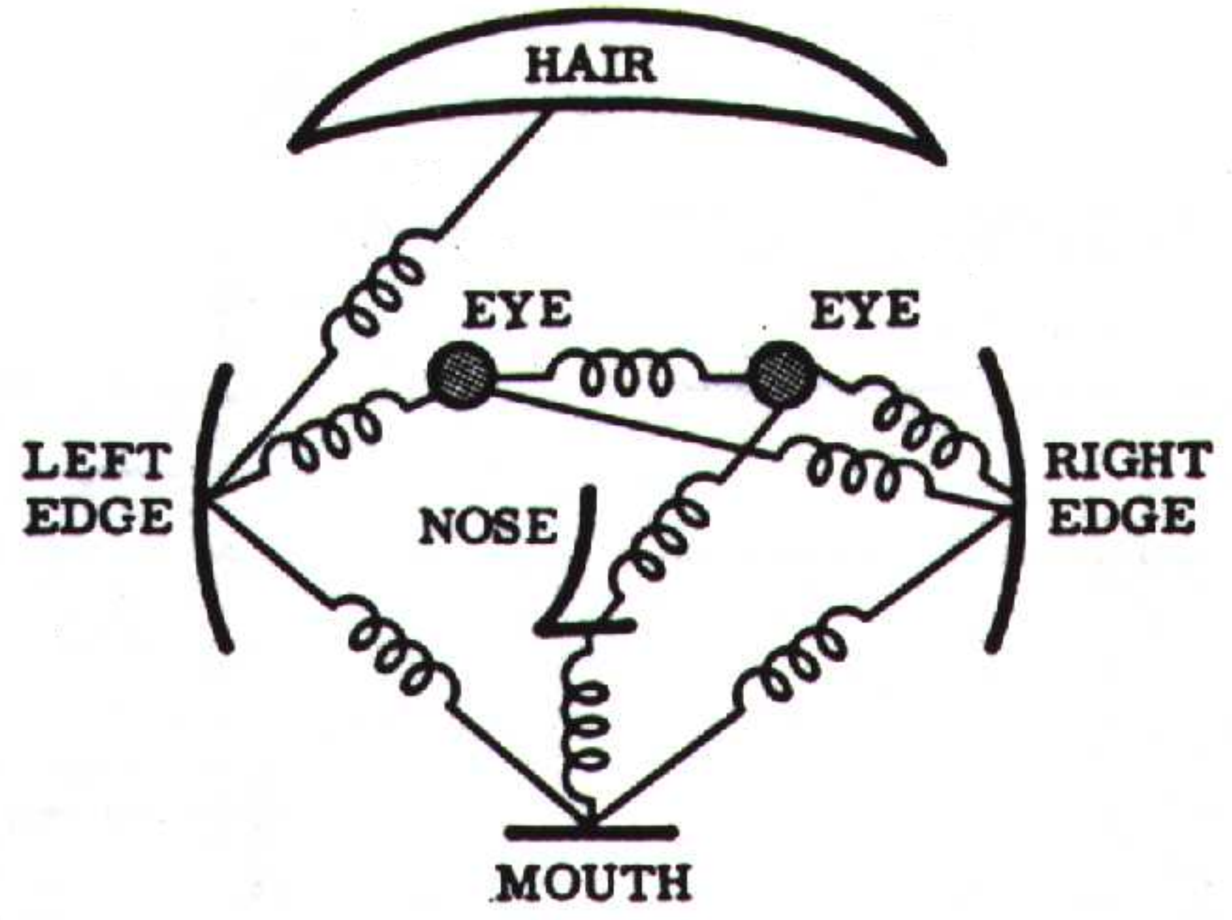
Generalized Cylinders



[Brooks & Binford, 1979]

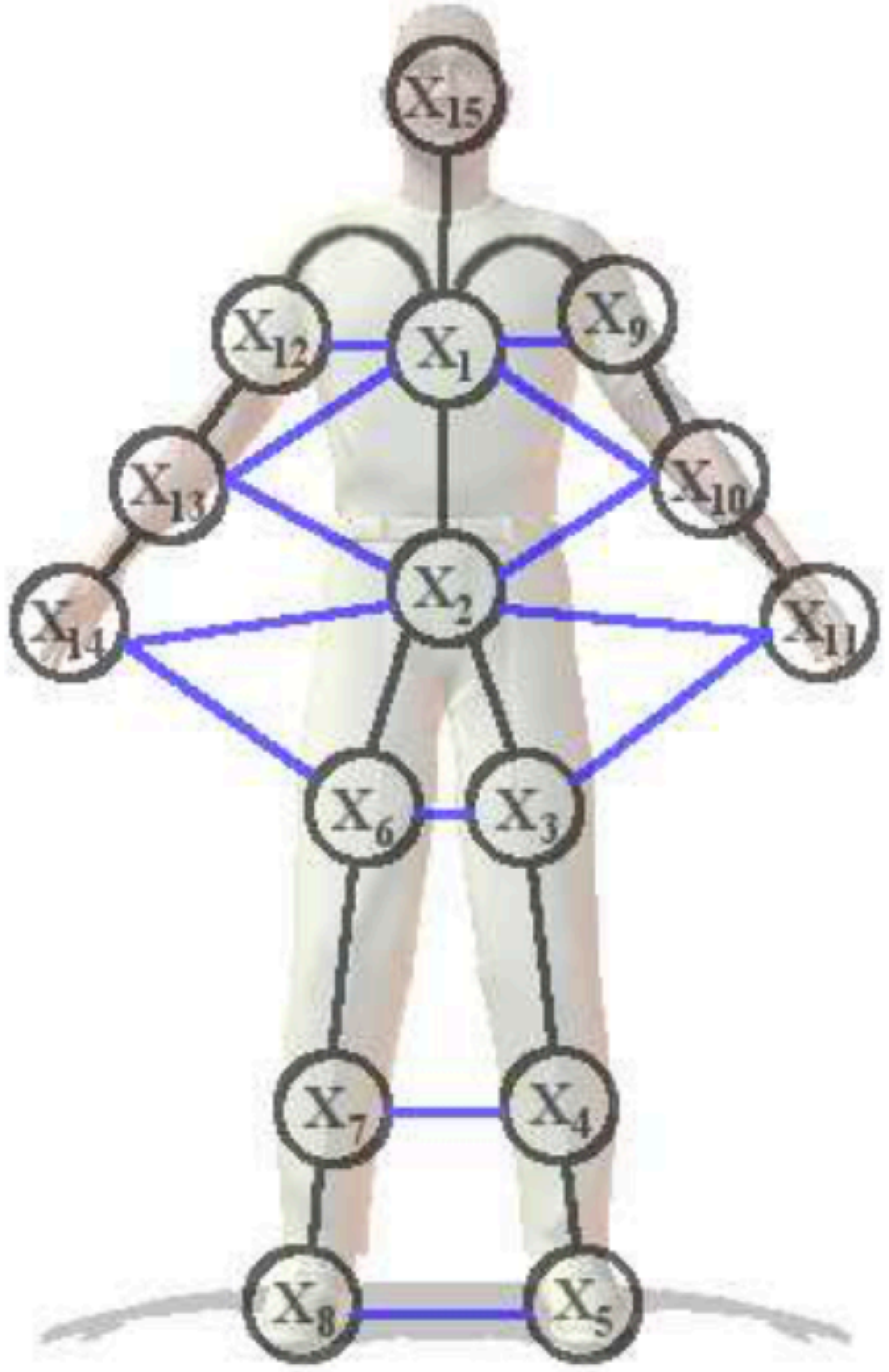
Pictorial Structures



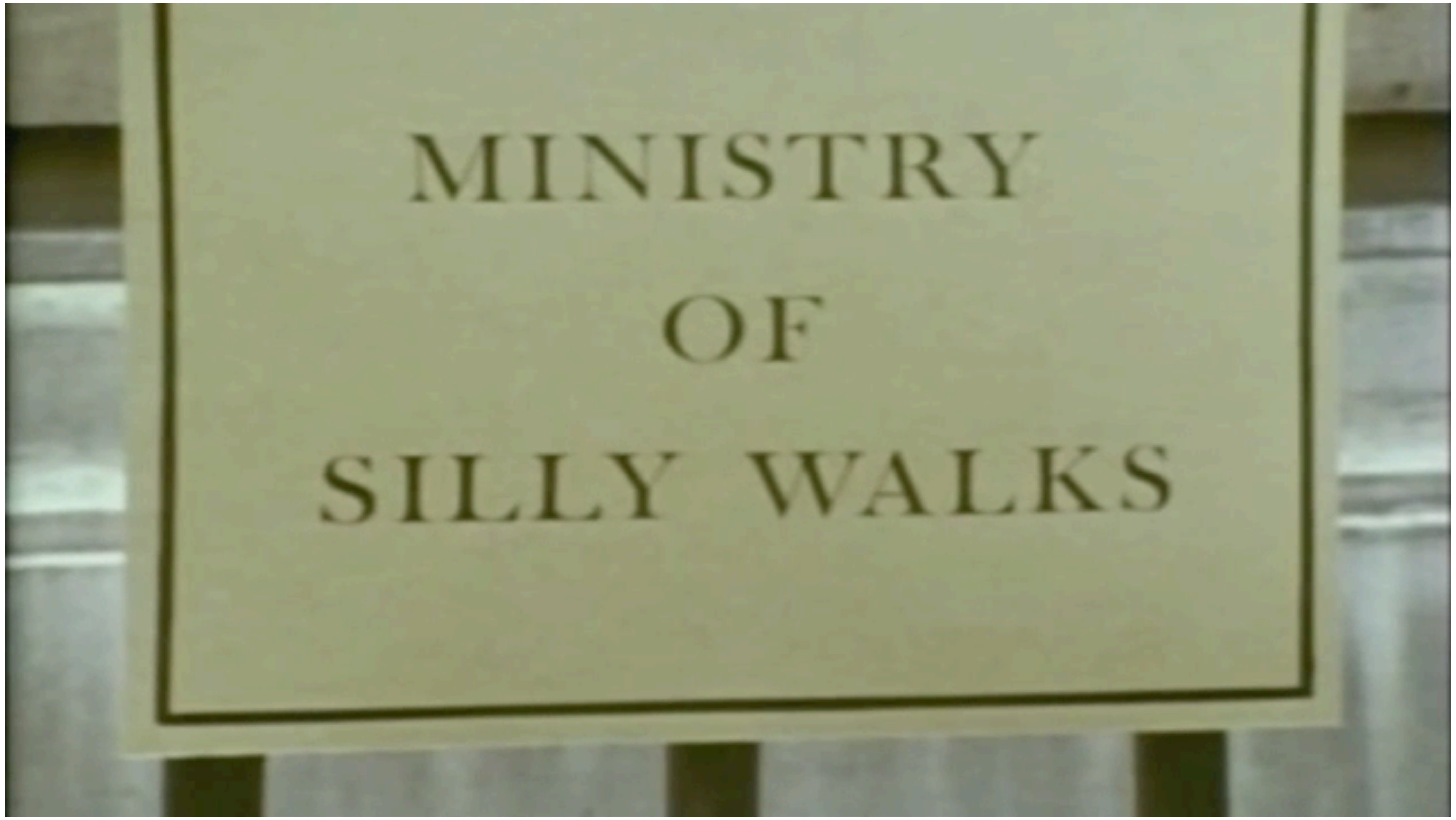


[Fischler & Elschlager, 1973]

Part-based Models

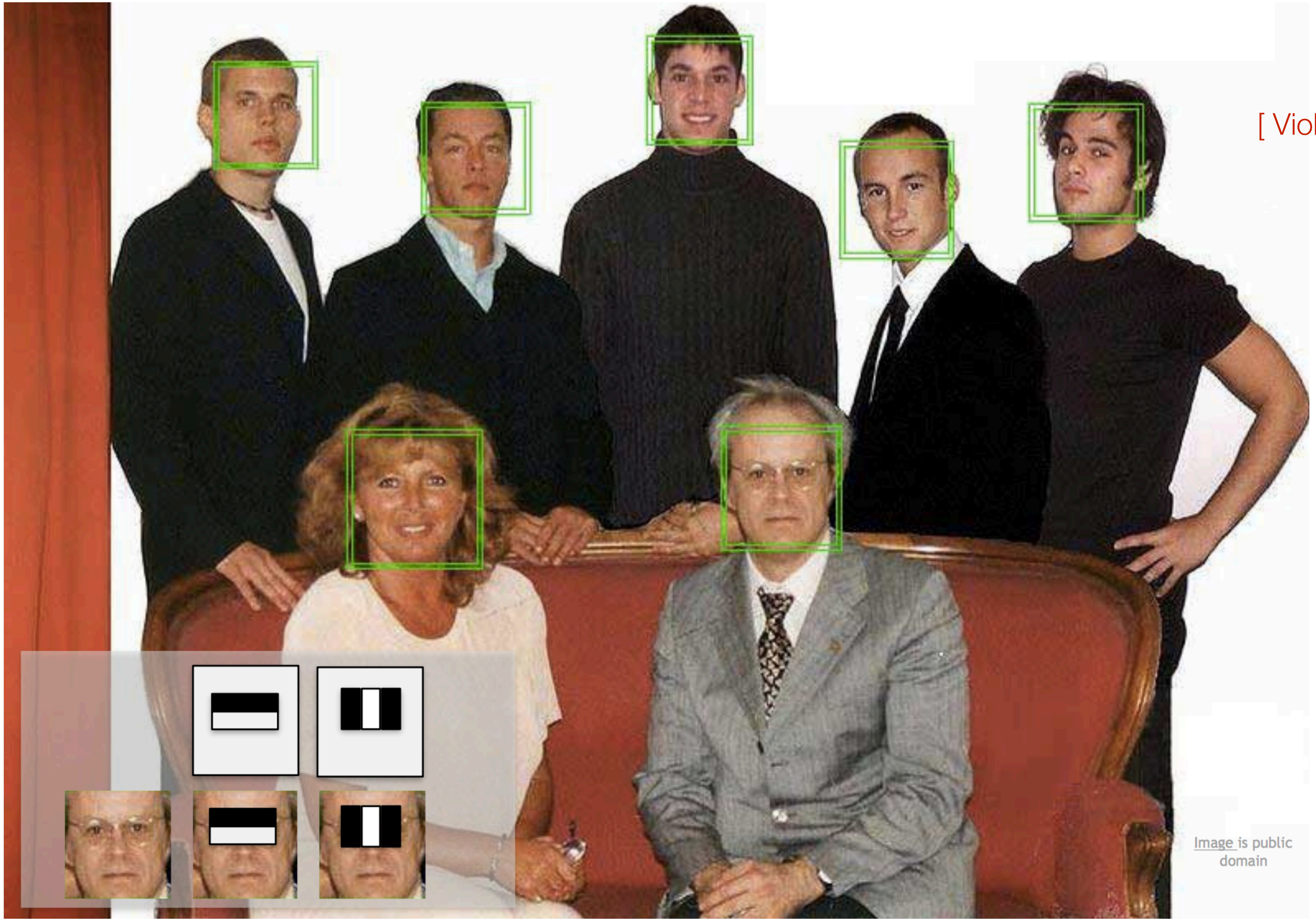


[Sigal et al. 2004]



Monty Python's **Ministry of Silly Walks**

Face Detection 1999-2000



[Viola & Jones, 2001]

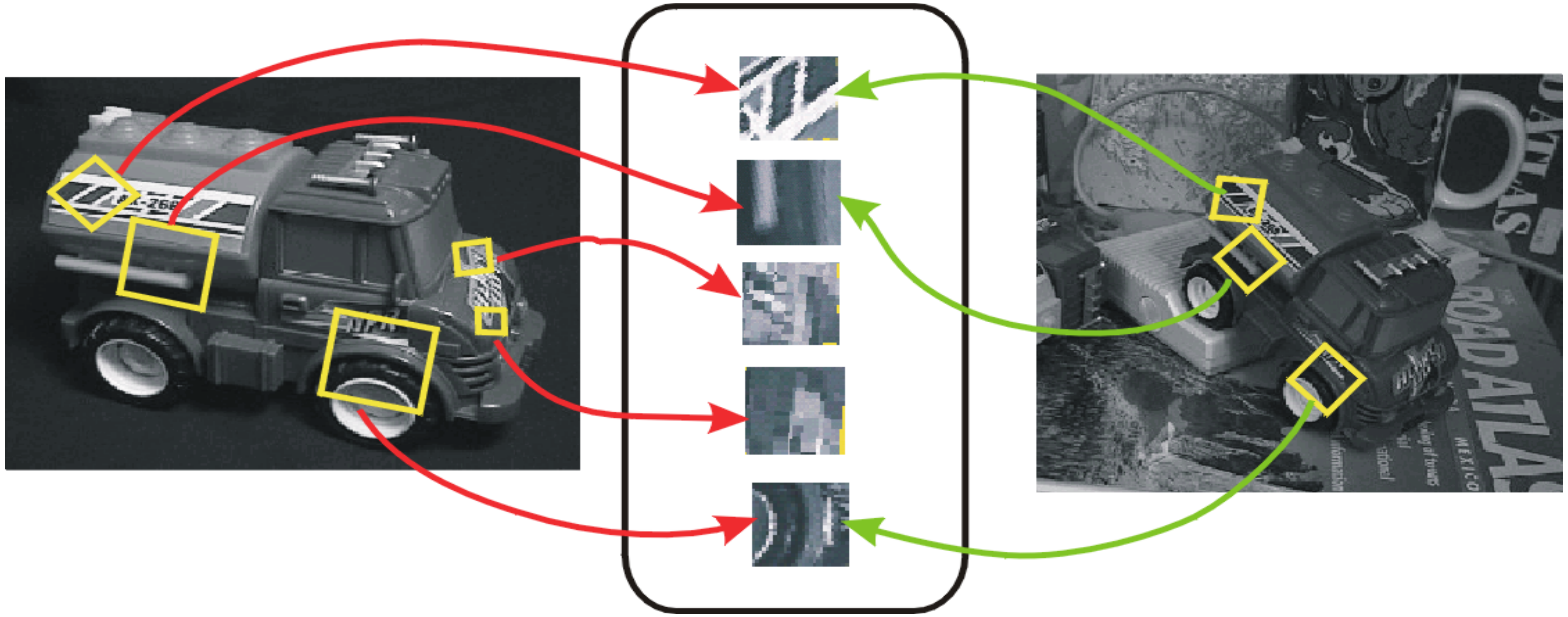
Image is public domain

Feature-based Vision

[**David Lowe**, 1999]

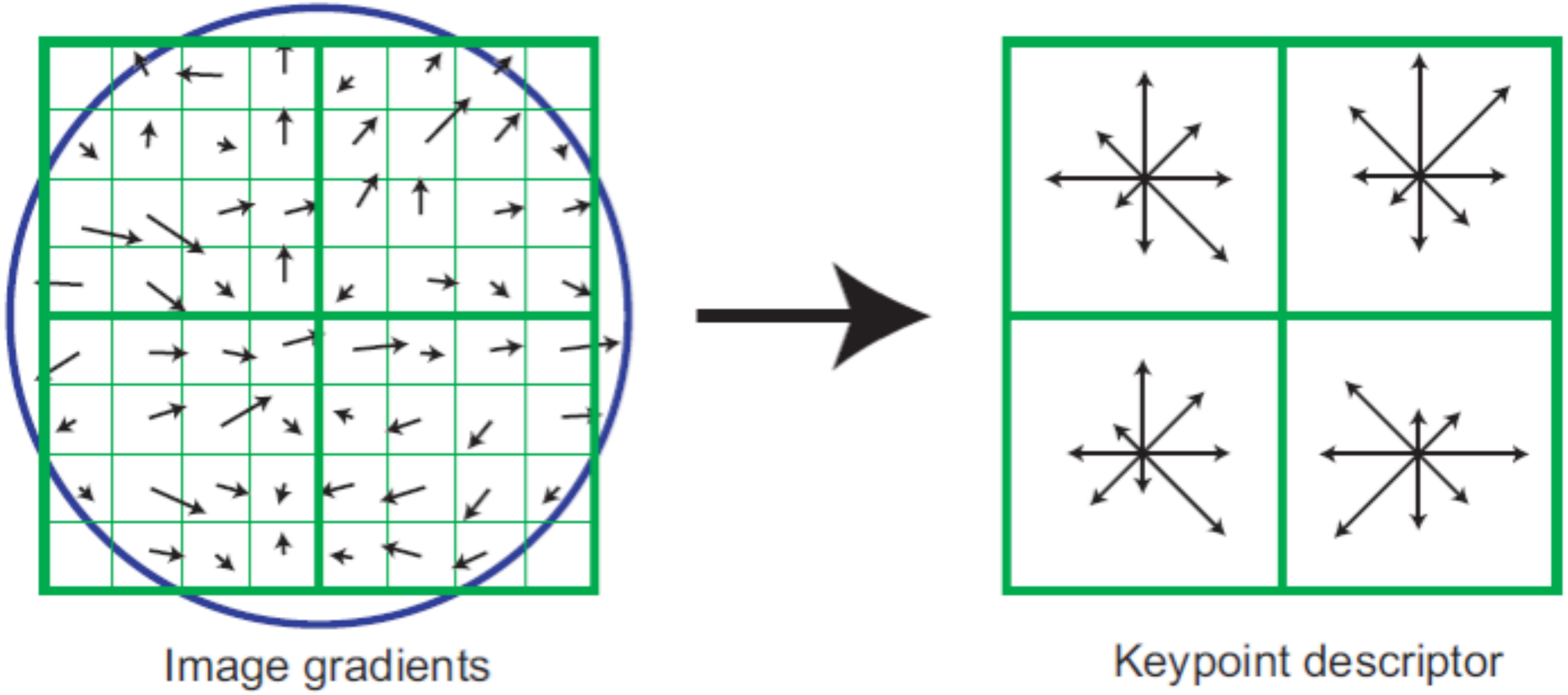
SIFT Idea

Image content is transformed into local feature coordinates that are **invariant** to translation, rotation, scale and imaging parameters



[**David Lowe**, 1999]

SIFT Discriptor



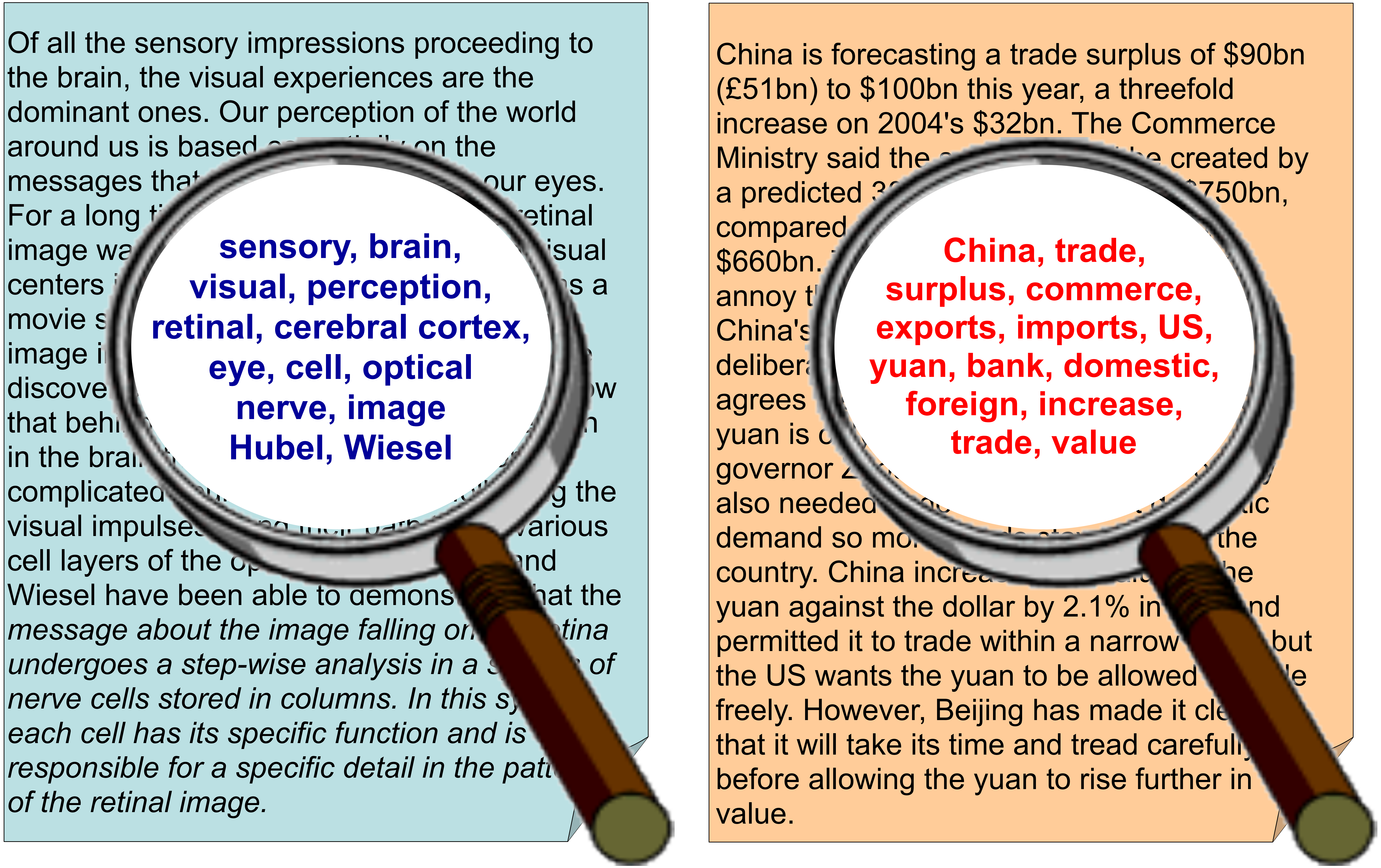
[**David Lowe**, 1999]

Massive 3D Reconstructions

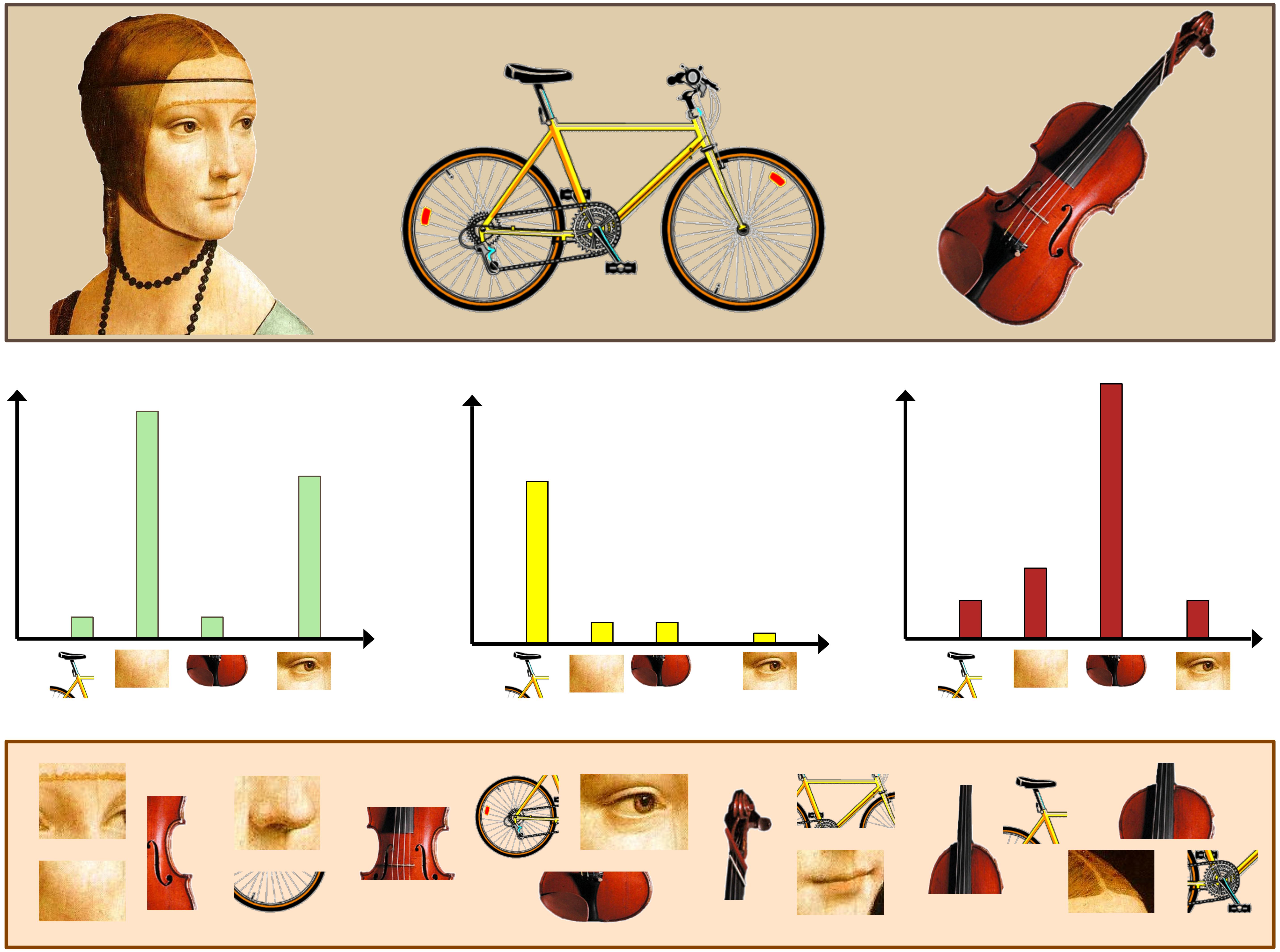
[Agarwal, Furukawa, Snavely, Curless, Seitz, Szeliski, 2010]

Bag-of-Words

*slide credit Li Fei-Fei

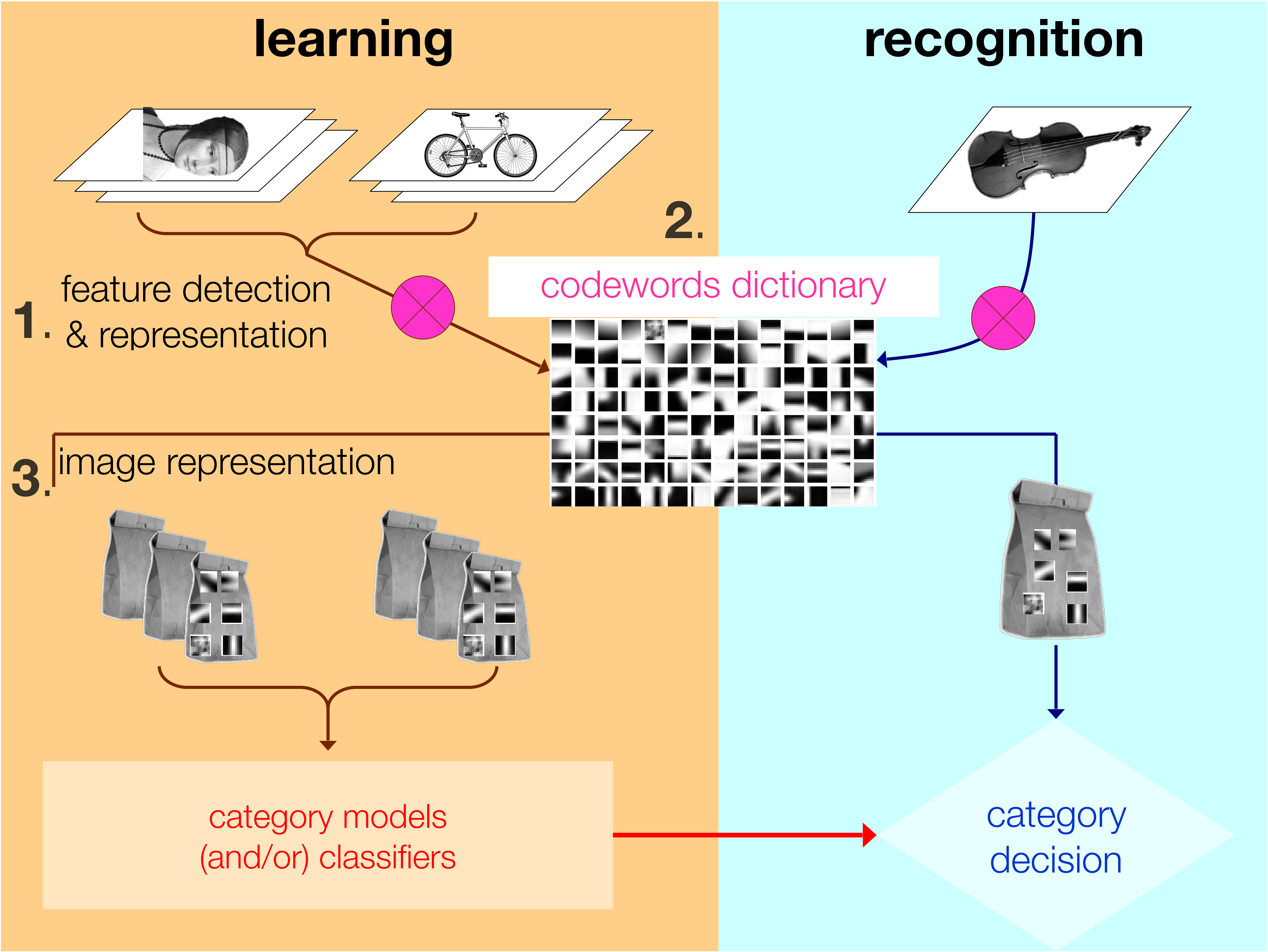


Bag-of-**Visual**-Words



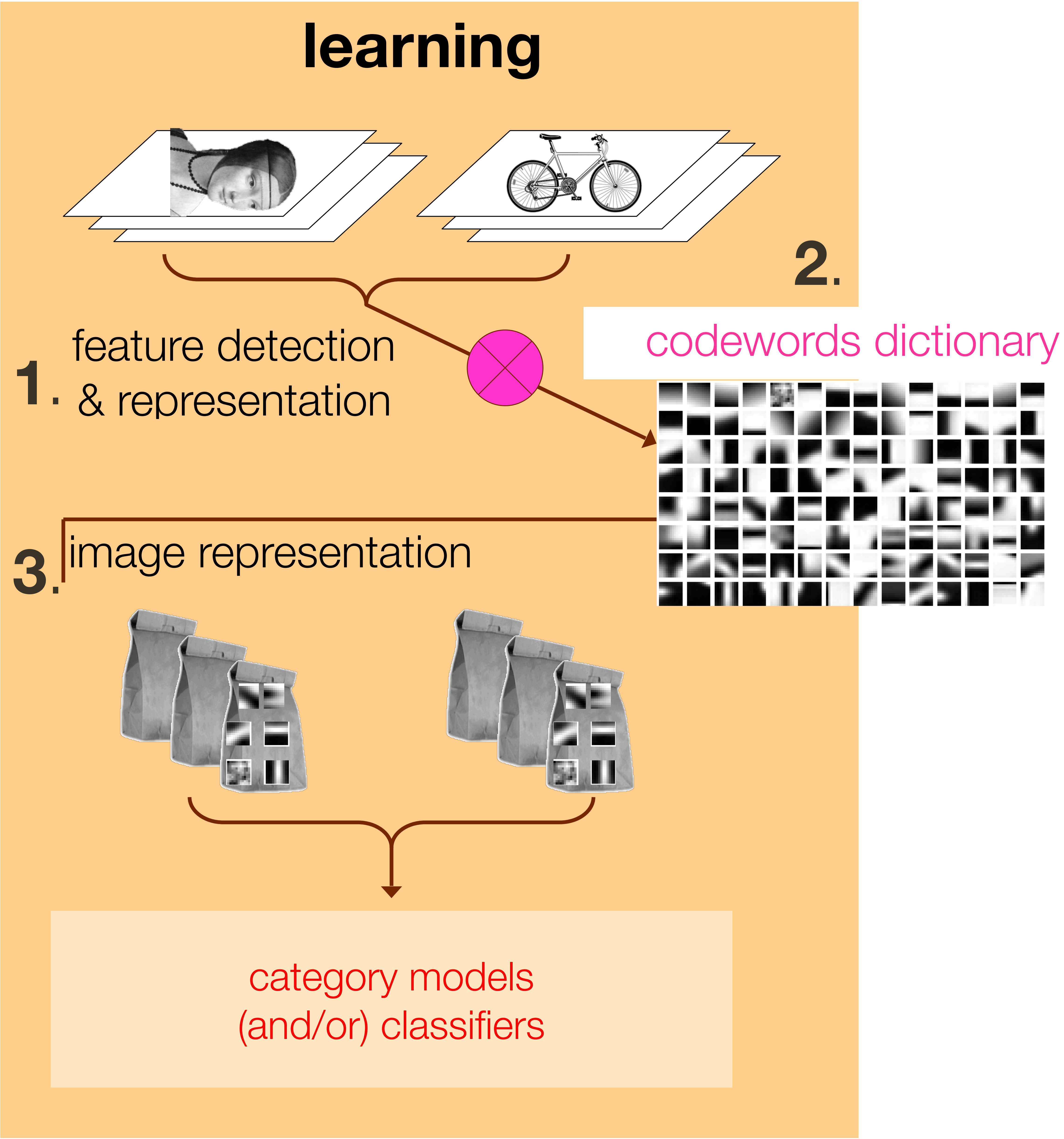
*slide credit Li Fei-Fei

Bag-of-**Visual**-Words



*slide credit Li Fei-Fei

Bag-of-**Visual**-Words: Learning



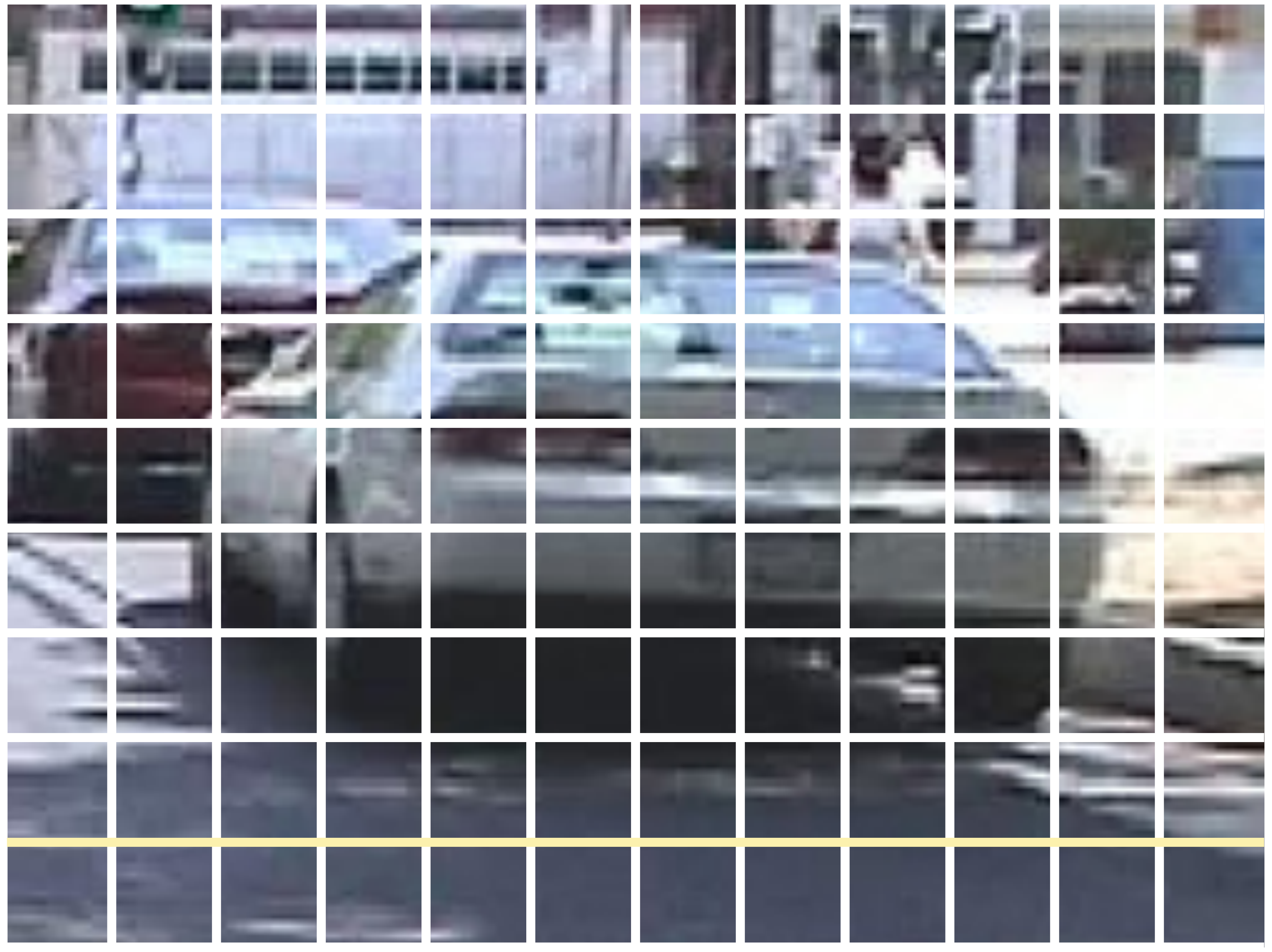
*slide credit Li Fei-Fei

Feature Detection & Representation

Regular Grid

— Vogel et al. 2003

— Fei-Fei et al. 2005



*slide credit Li Fei-Fei

Feature Detection & Representation

Regular Grid

— Vogel et al. 2003

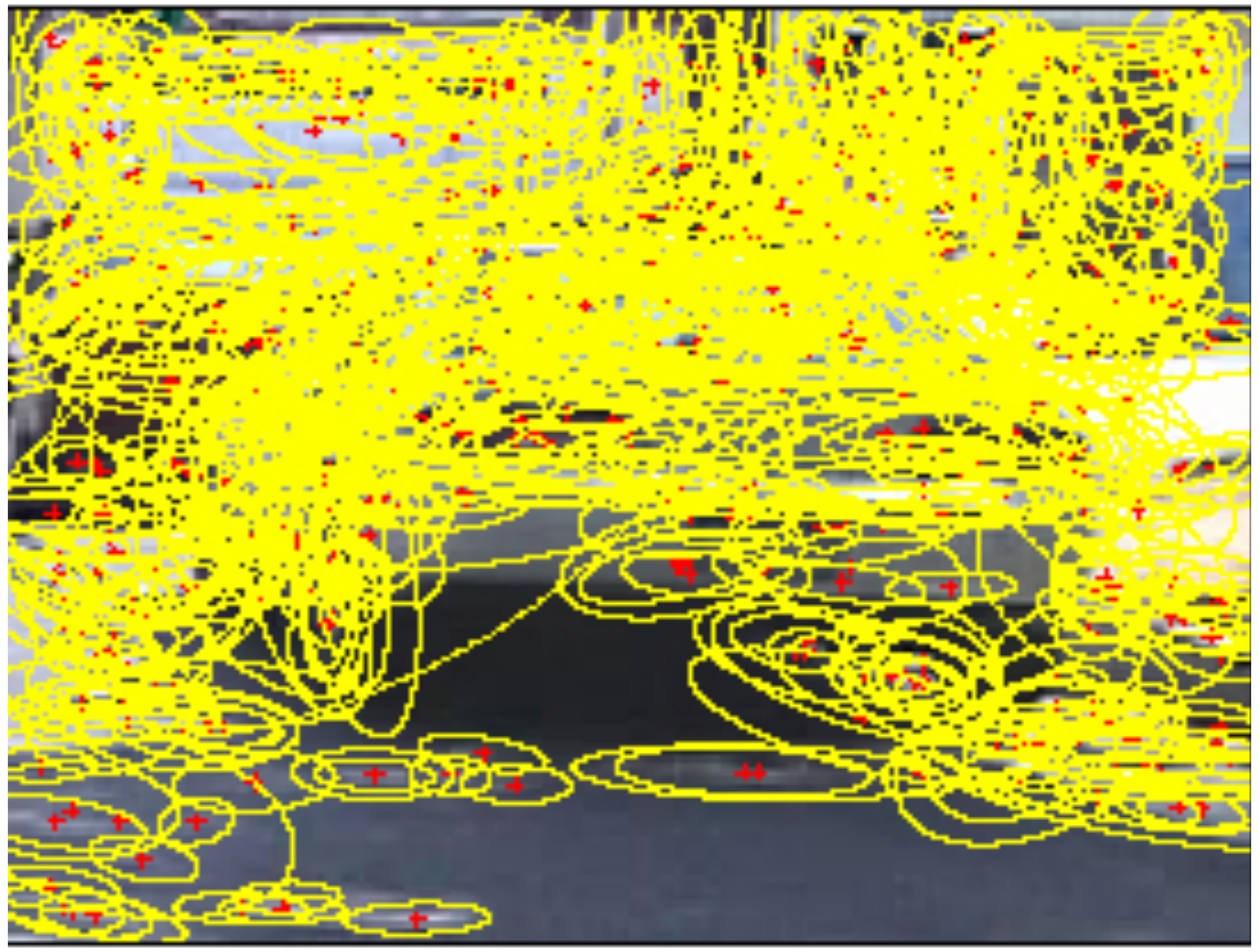
— Fei-Fei et al. 2005

Interest Point Detector

— Csurka et al. 2004

— Fei-Fei et al. 2005

— Sivic et al. 2005



*slide credit Li Fei-Fei

Feature Detection & Representation

Regular Grid

— Vogel et al. 2003

— Fei-Fei et al. 2005

Interest Point Detector

— Csurka et al. 2004

— Fei-Fei et al. 2005

— Sivic et al. 2005

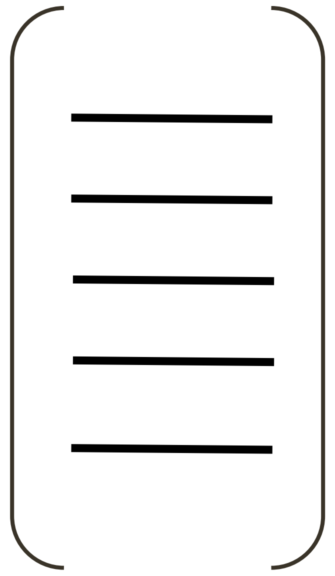
Other Methods

— Randome sampling (Ullman et al. 2002)

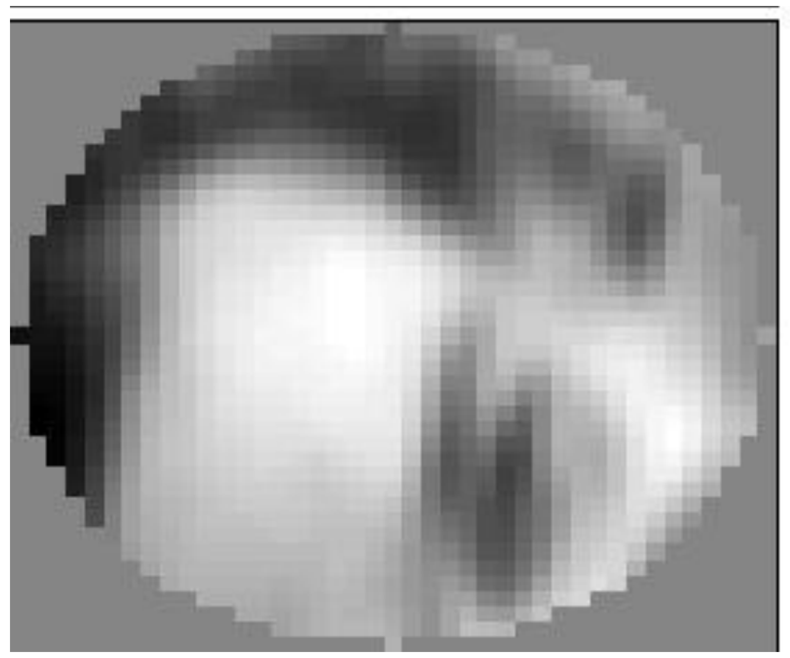
— Segmentation based patches (Barnard et al. 2003)

*slide credit Li Fei-Fei

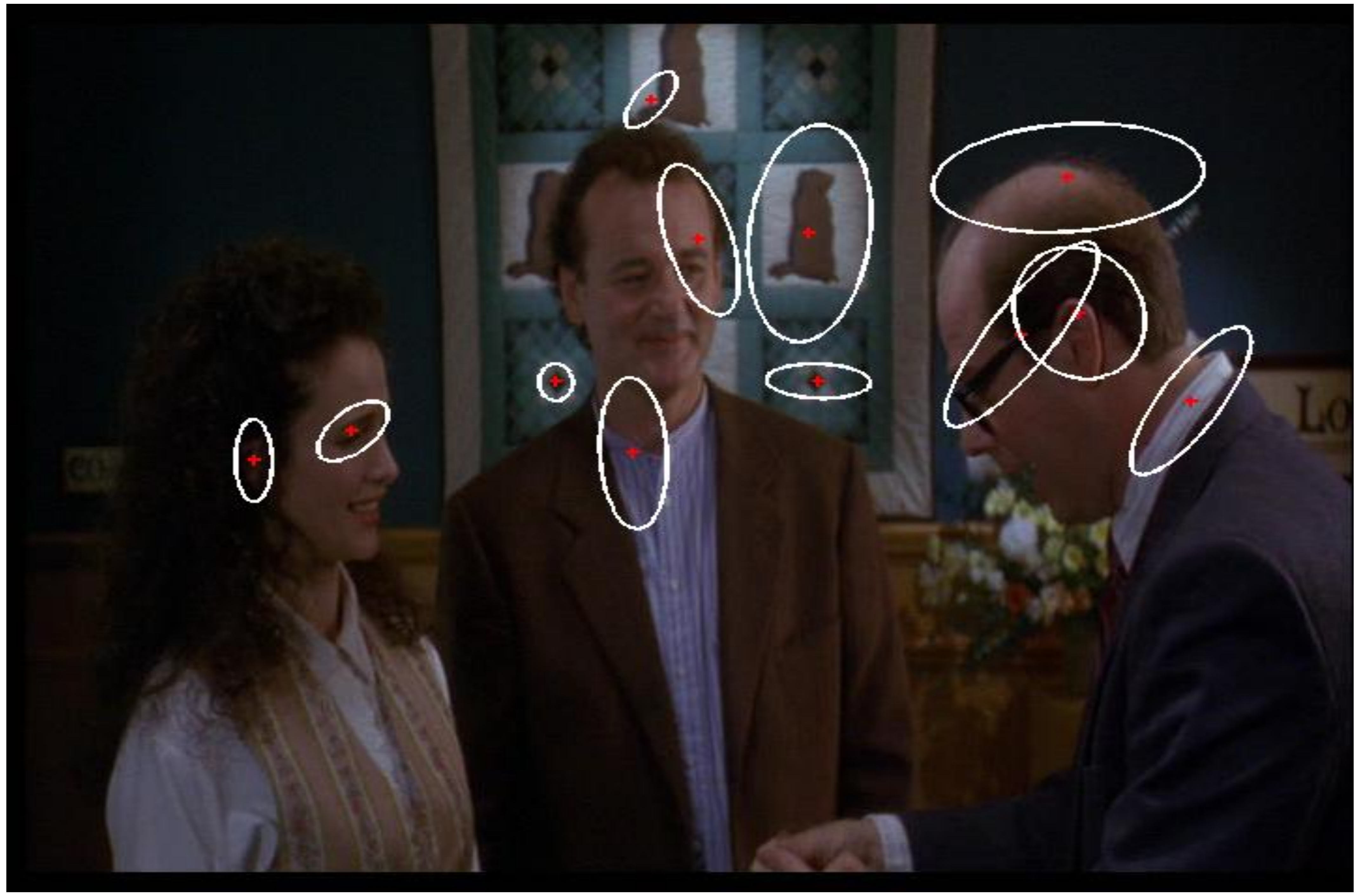
Feature Detection & **Representation**



Compute **SIFT** descriptor



Normalize Patch

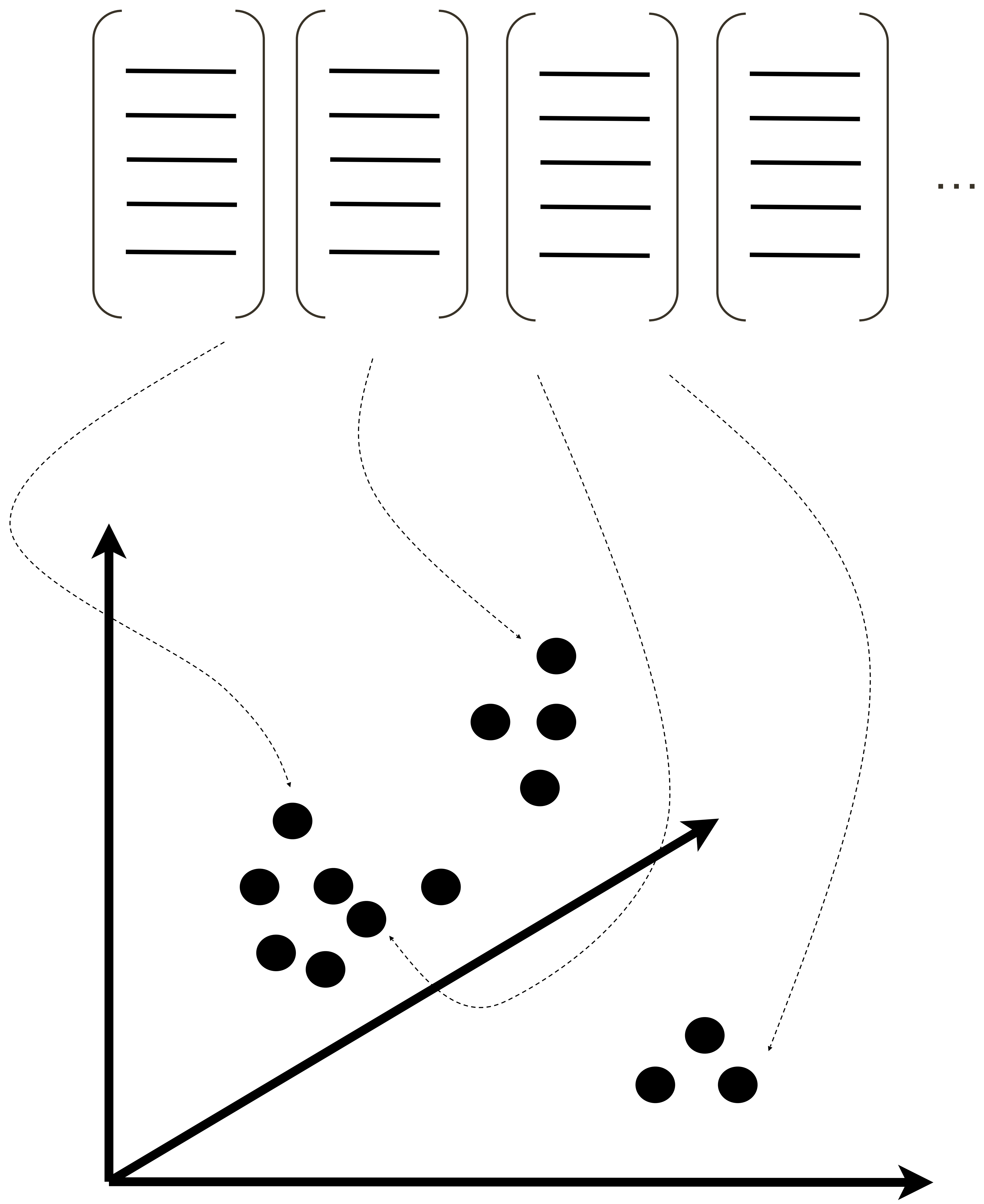


*slide credit Josef Sivic

Feature Detection & **Representation**

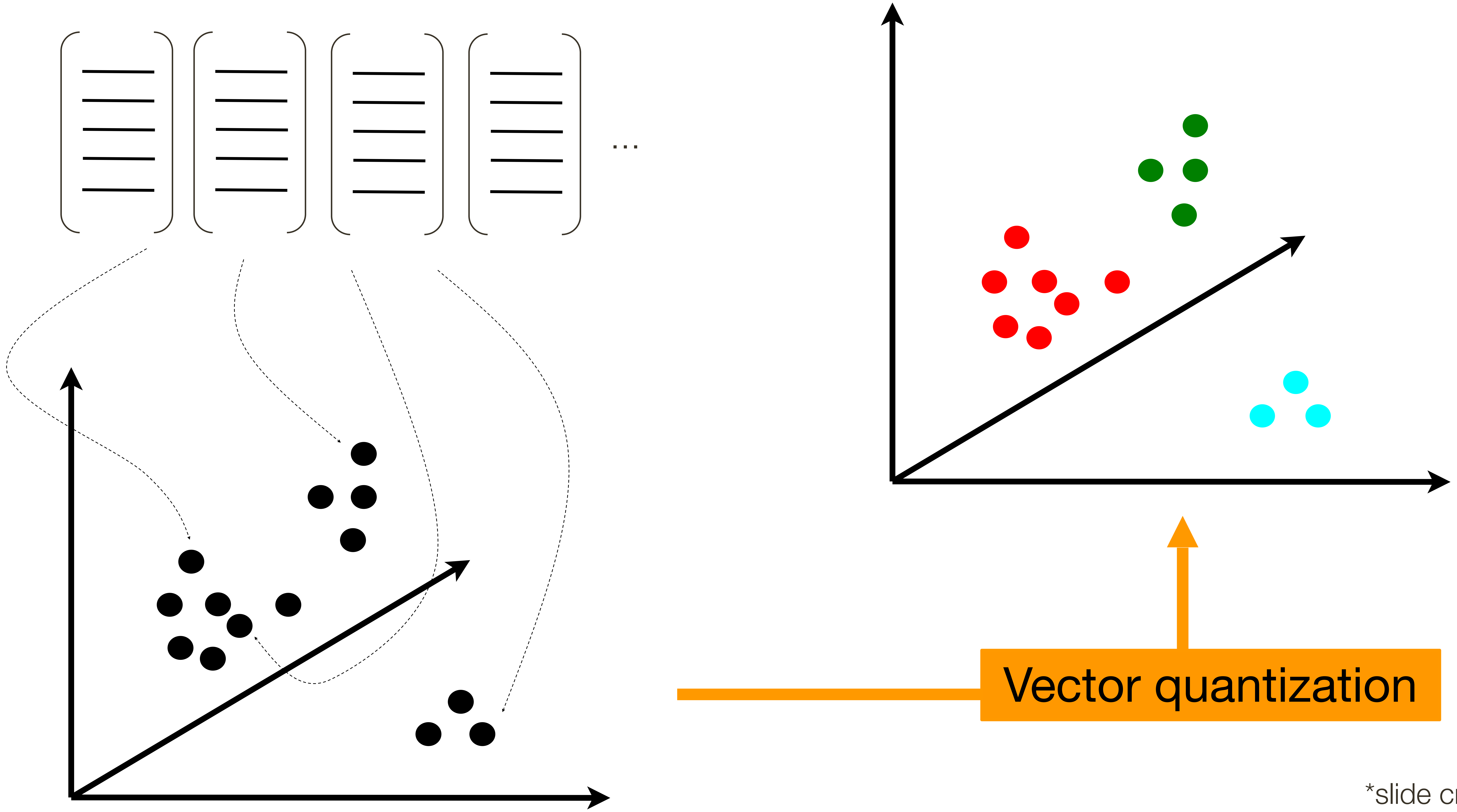
*slide credit Josef Sivic

Feature Detection & **Representation**



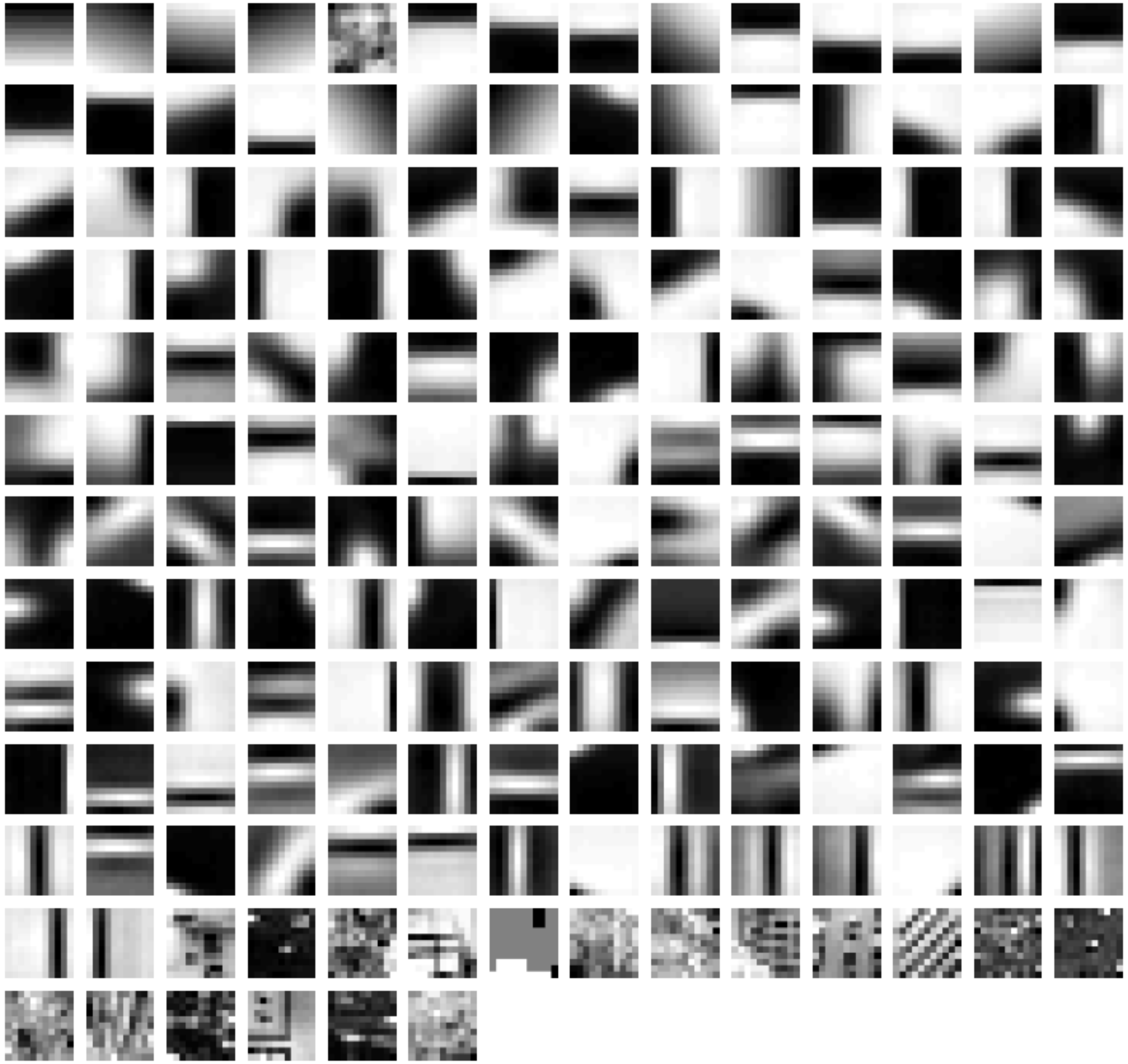
*slide credit Josef Sivic

Feature Detection & **Representation**



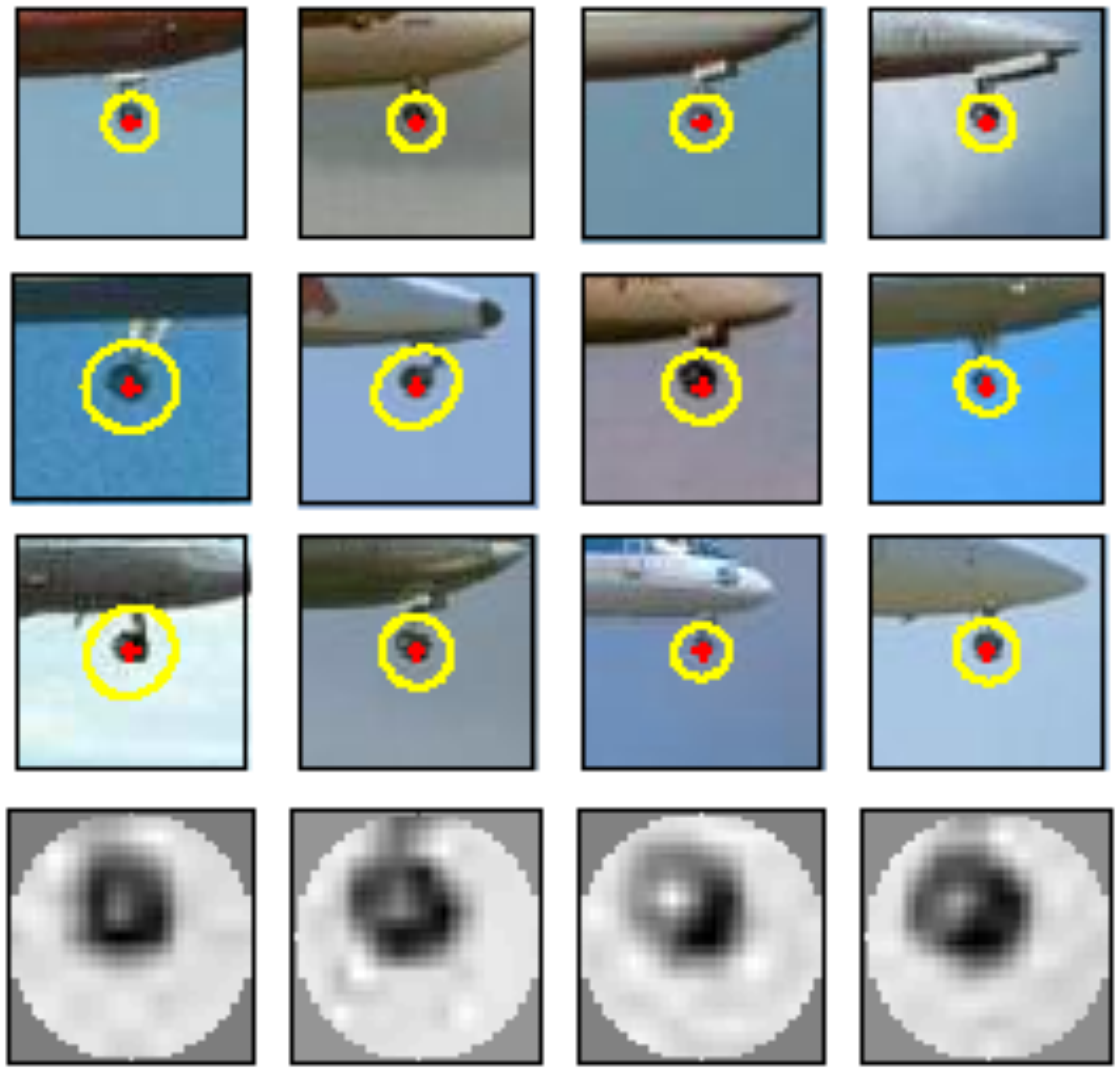
*slide credit Josef Sivic

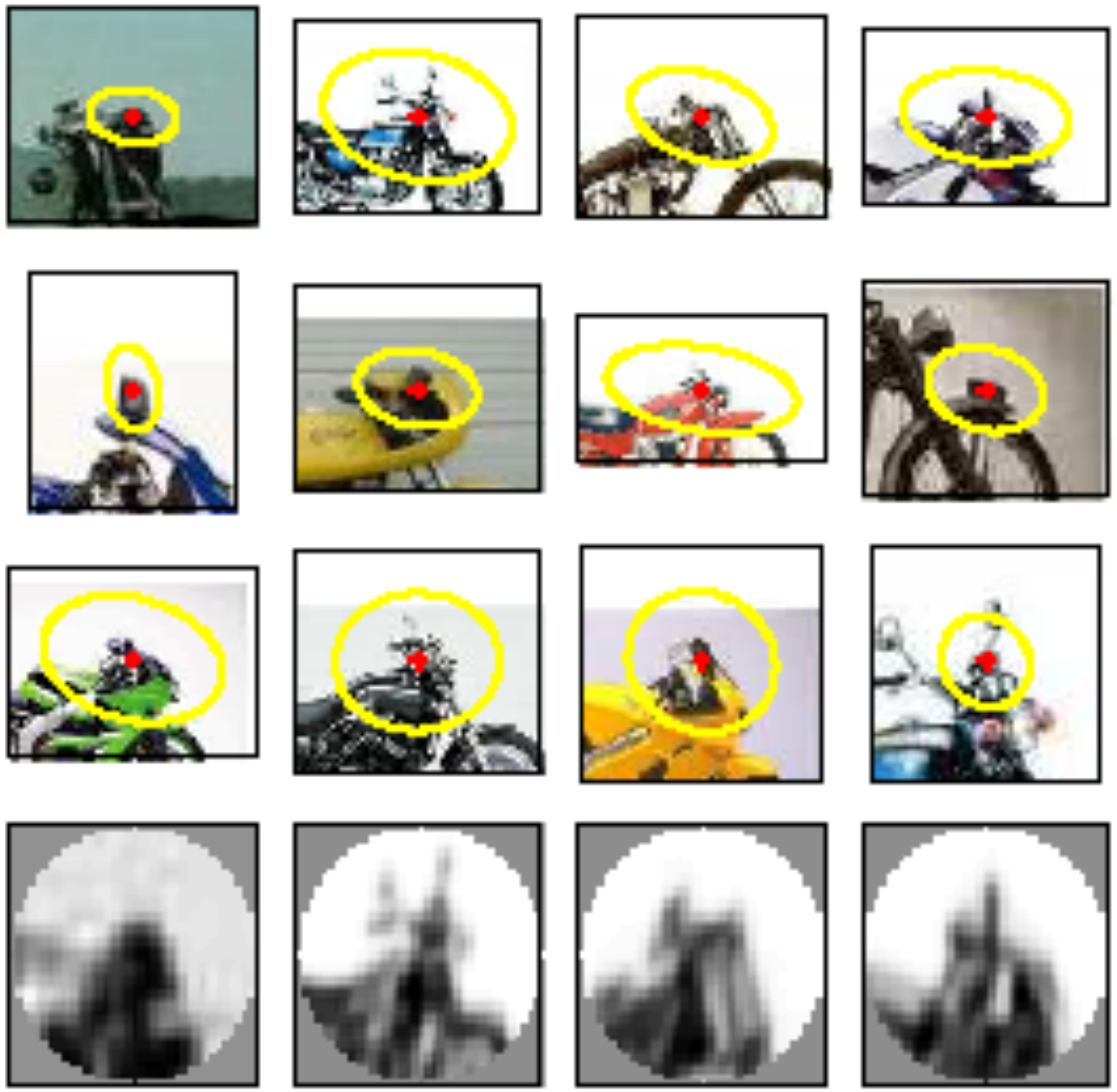
Codeword **Dictionary**



[Fei-Fei et al., 2005]

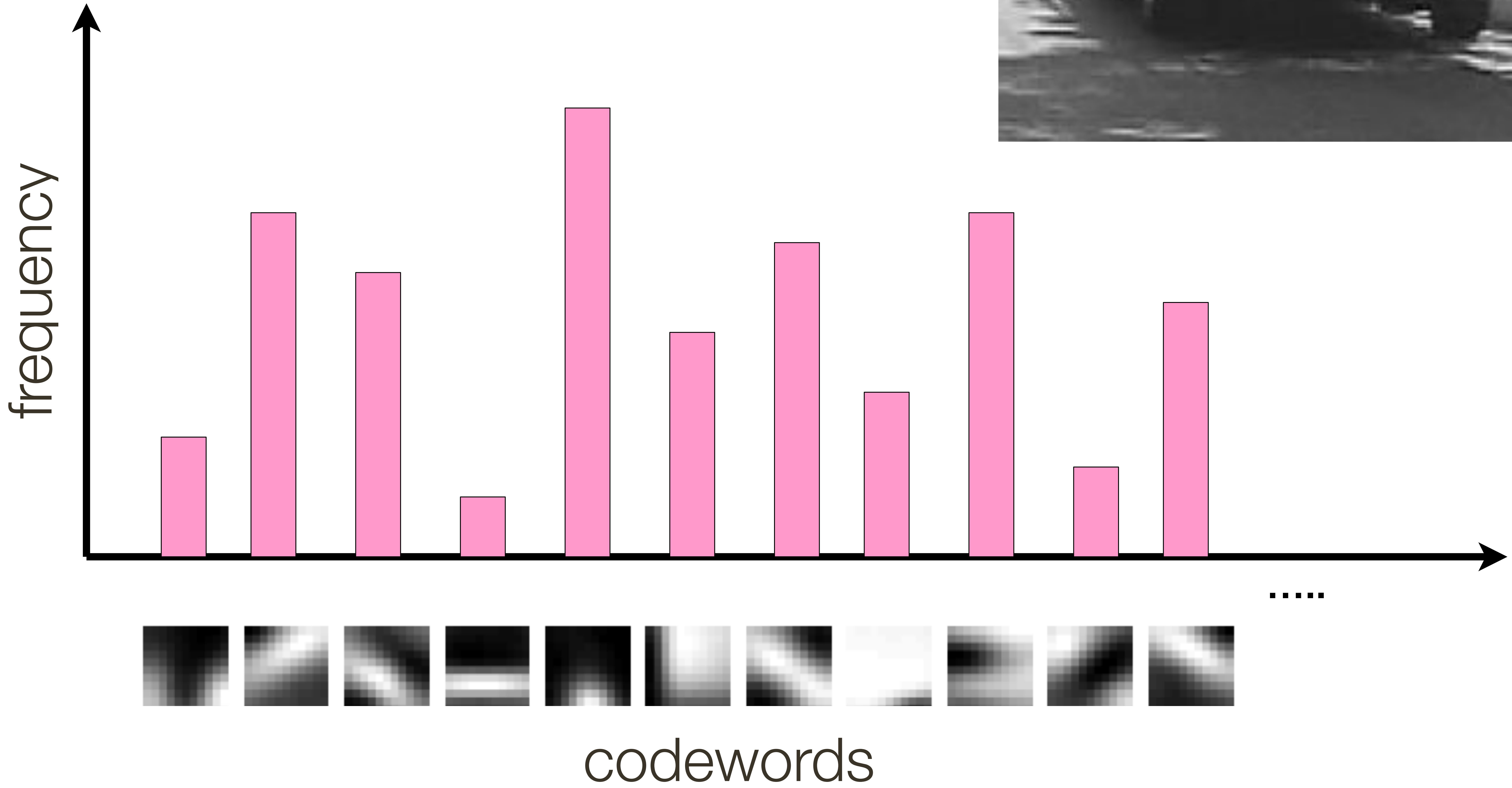
Image Patch Examples of Code Words





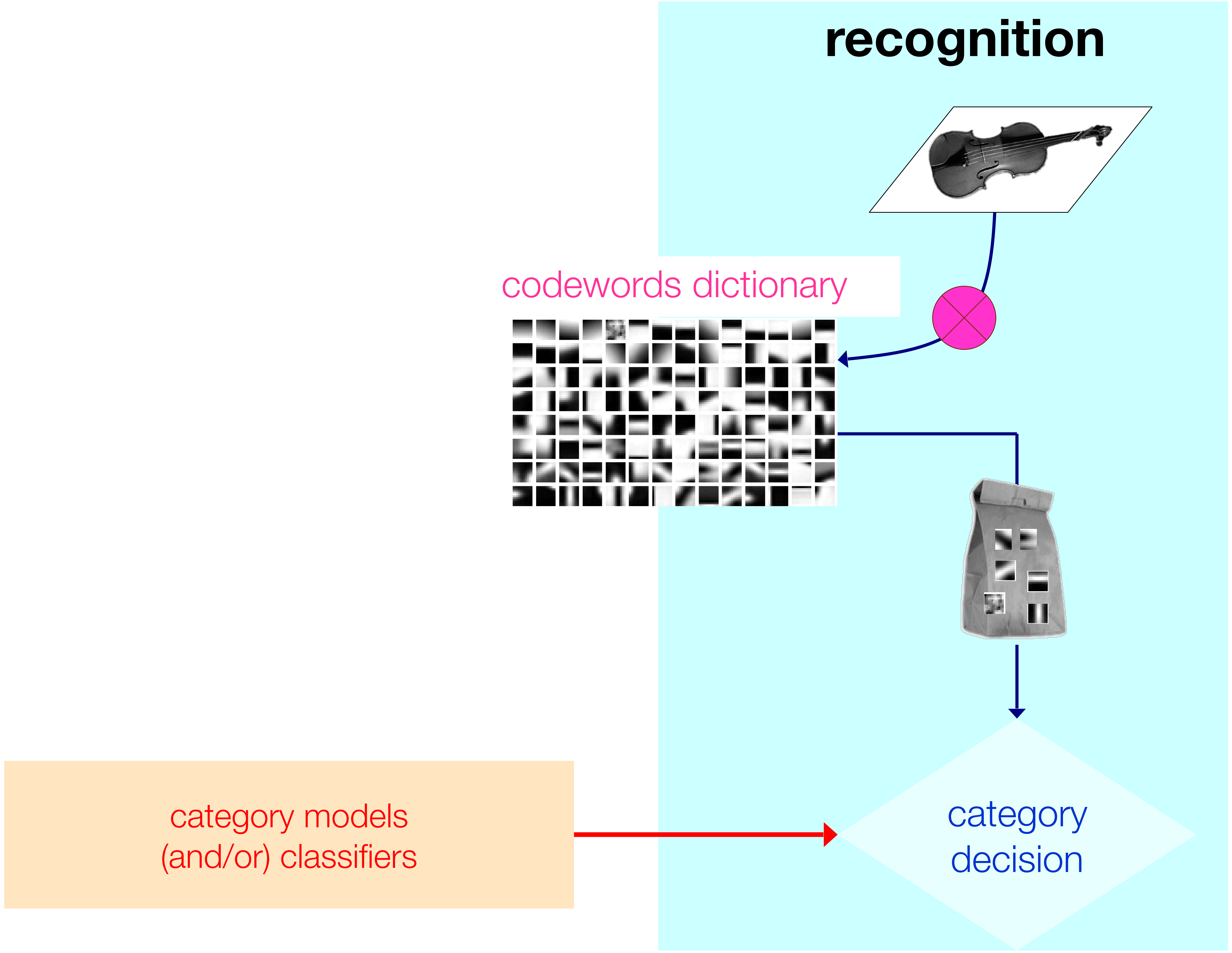
[Sivic et al., 2005]

Image Representation



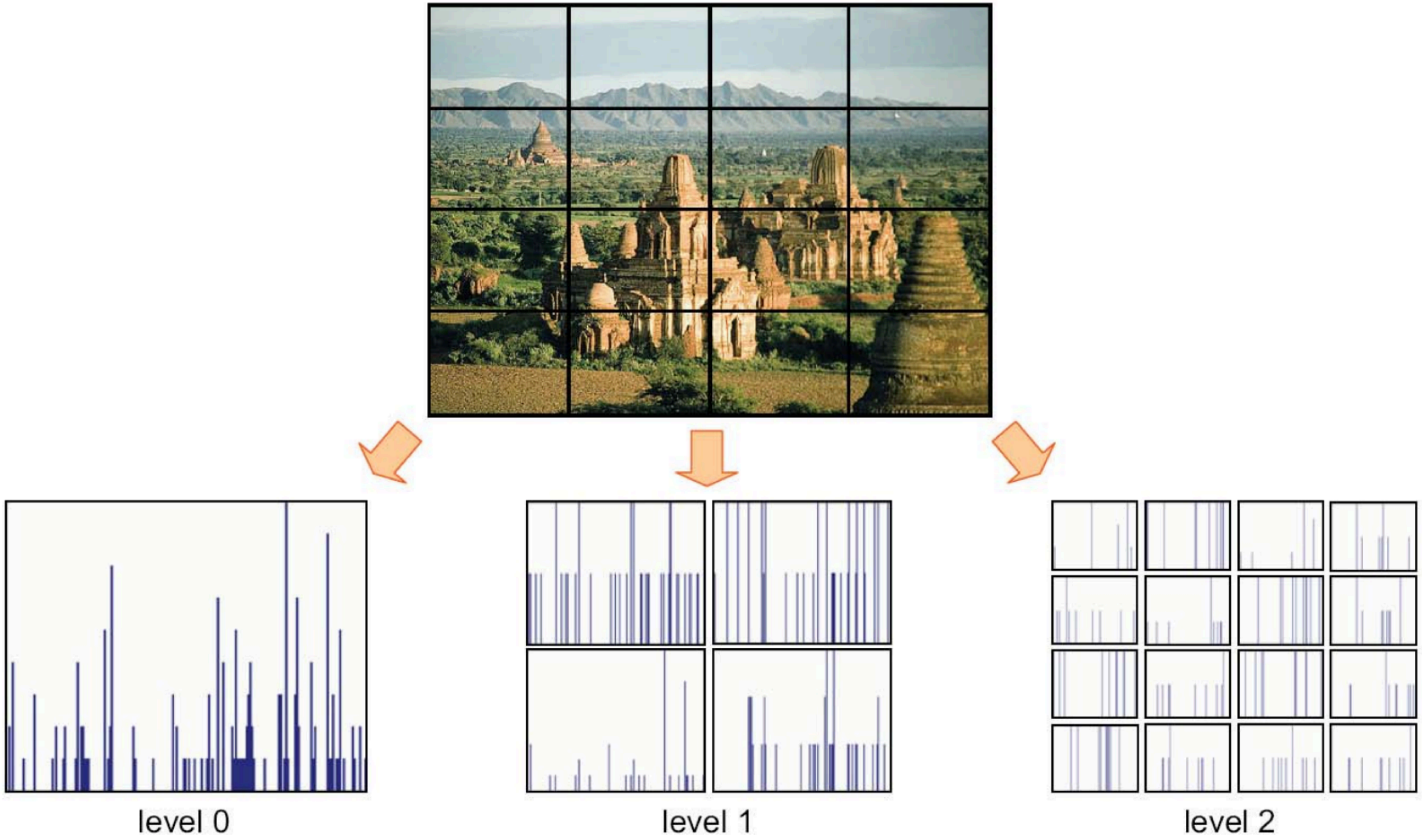
*slide credit Li Fei-Fei

Bag-of-**Visual**-Words



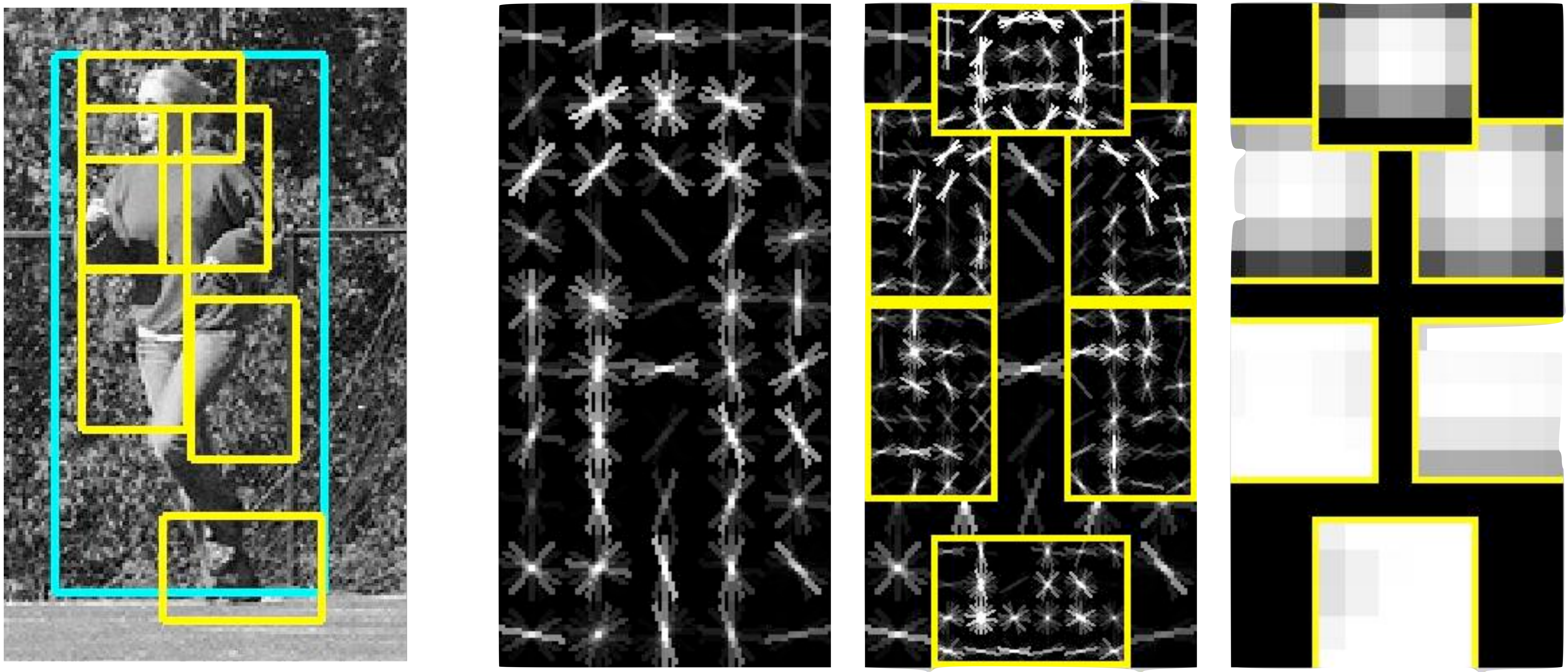
*slide credit Li Fei-Fei

Beyond Bag of Features



[Lazebnik, Schmid, Ponce, 2006]

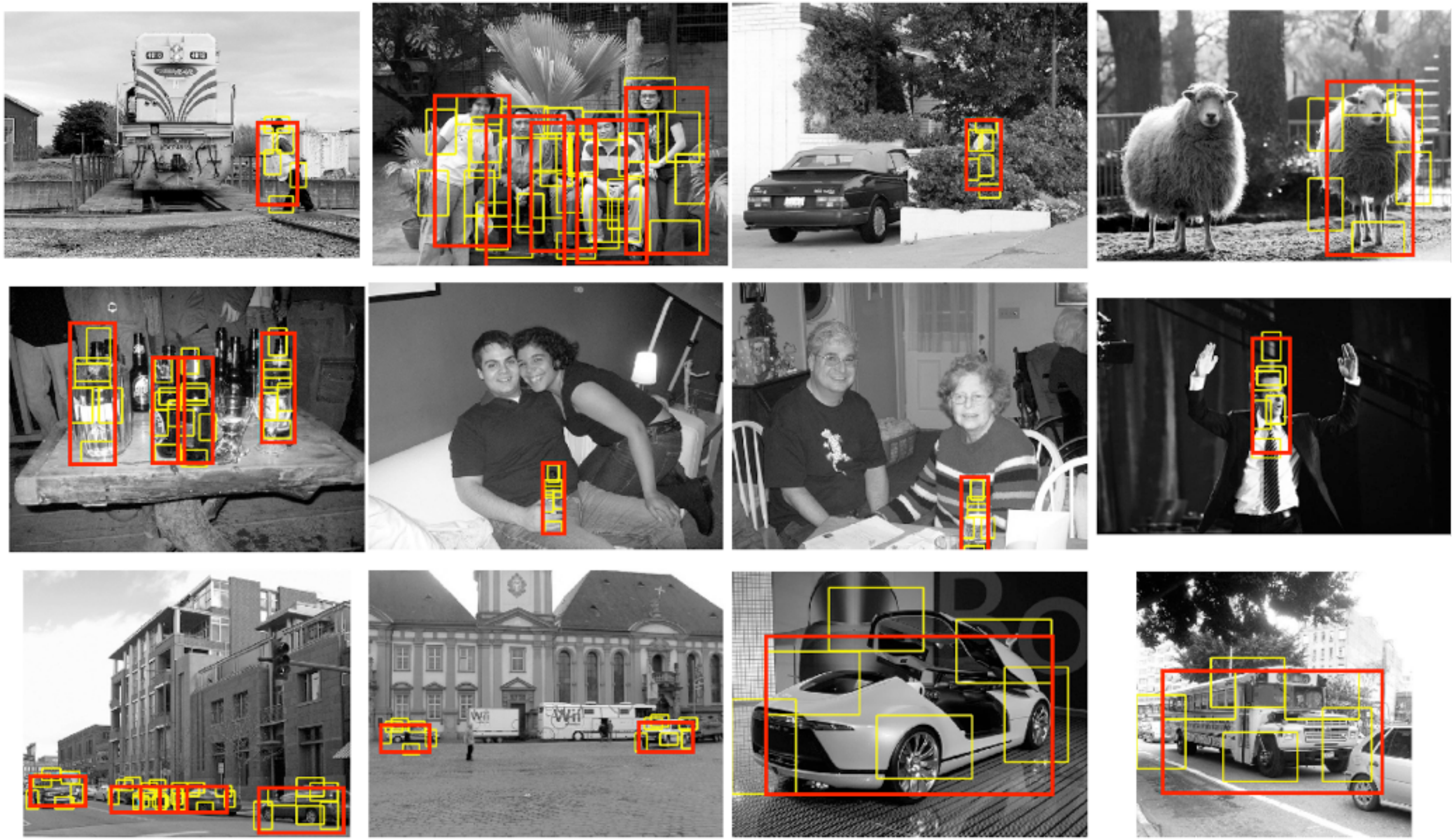
Deformable Part Models



Detection Root Filter Part Filters Deformations

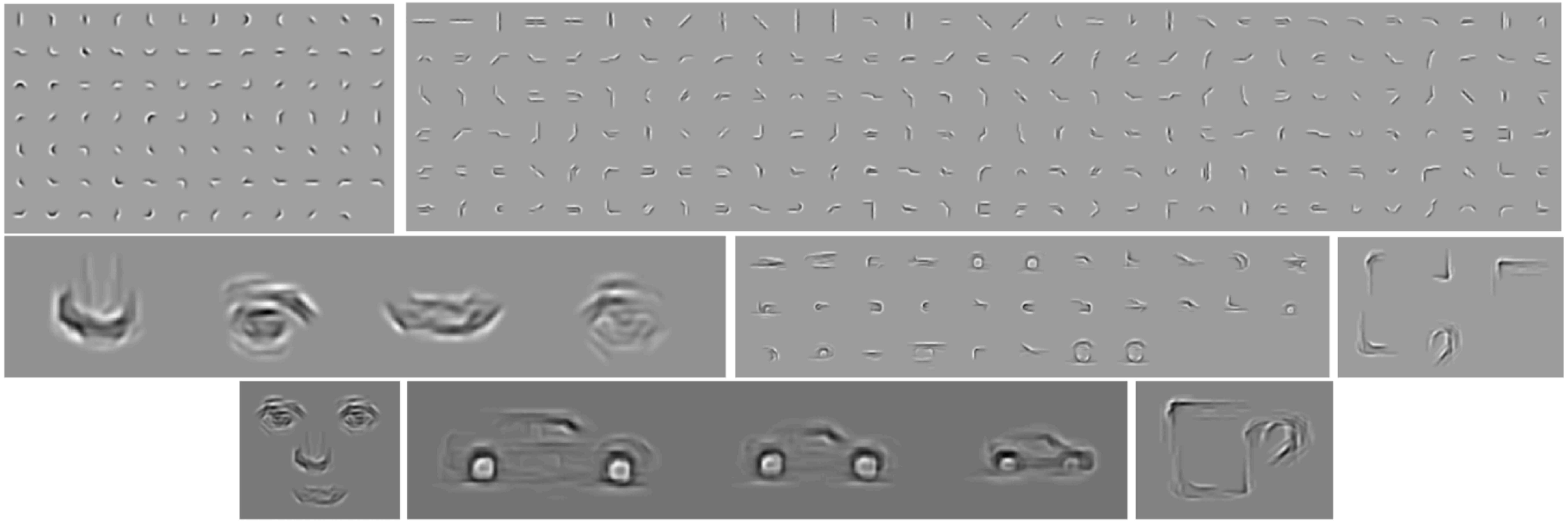
[Felzenswalb, McAllester, Ramanan, 2009]

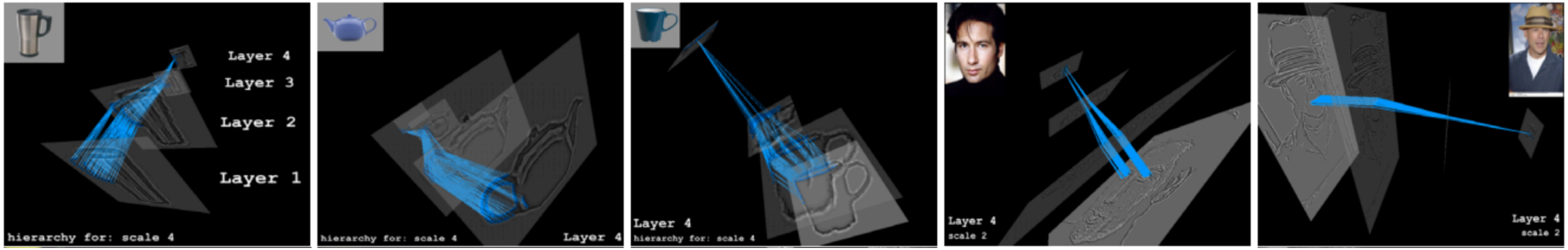
Deformable Part Models



[Felzenswalb, McAllester, Ramanan, 2009]

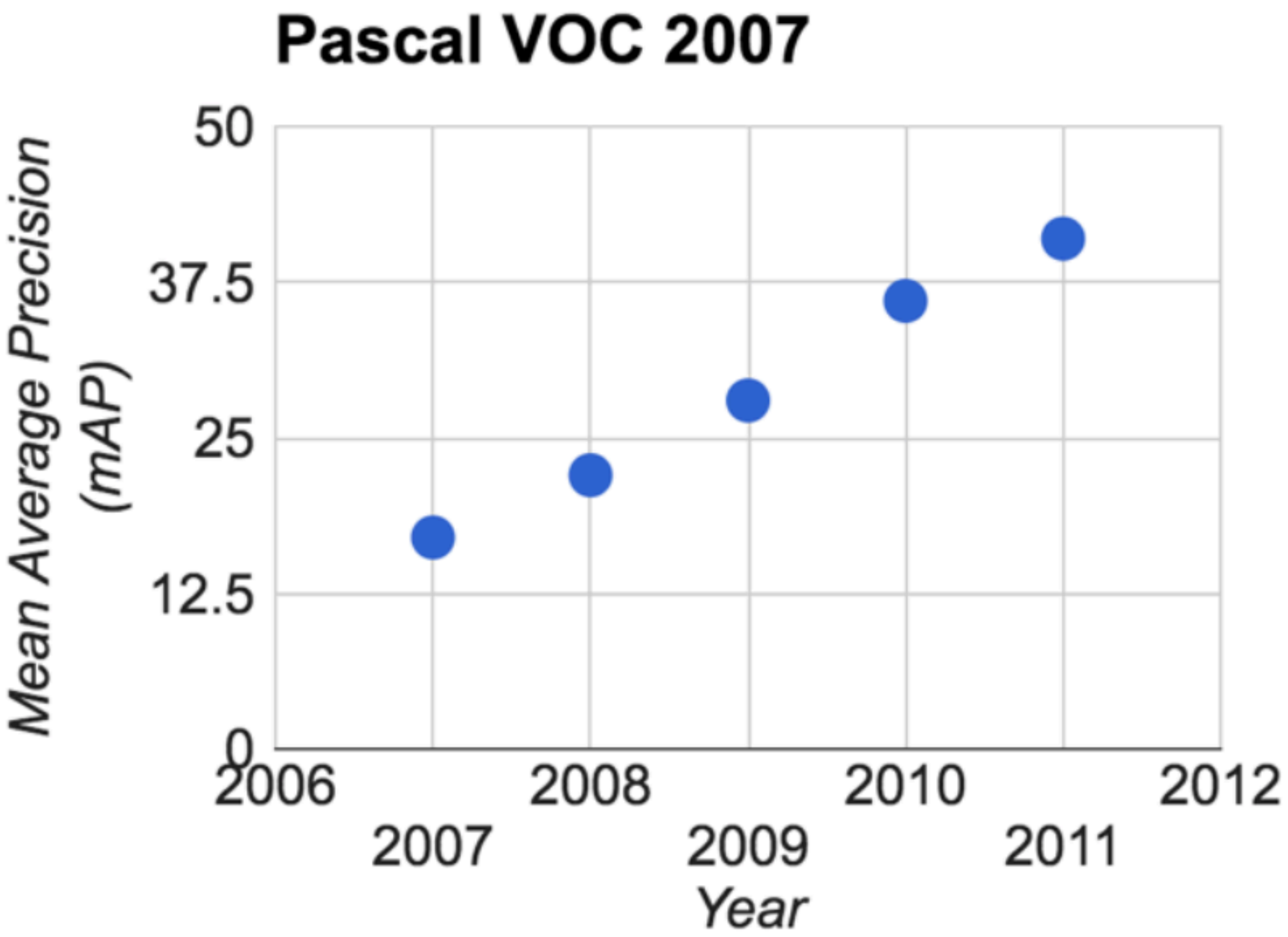
Hierarchical Models





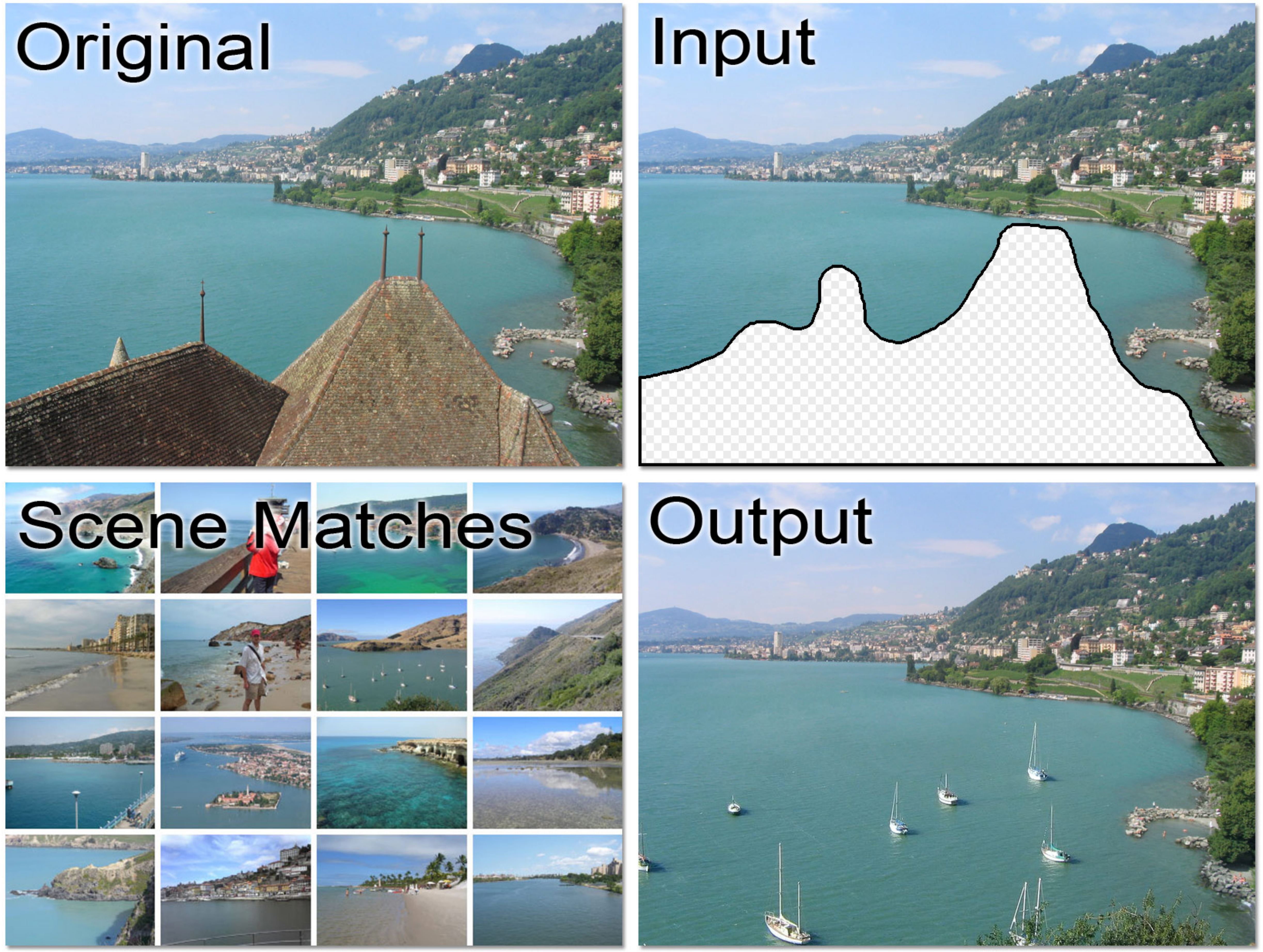
[Fidler, Leonardis, CVPR 2007]

PASCAL Visual Object Challenge (VOC)

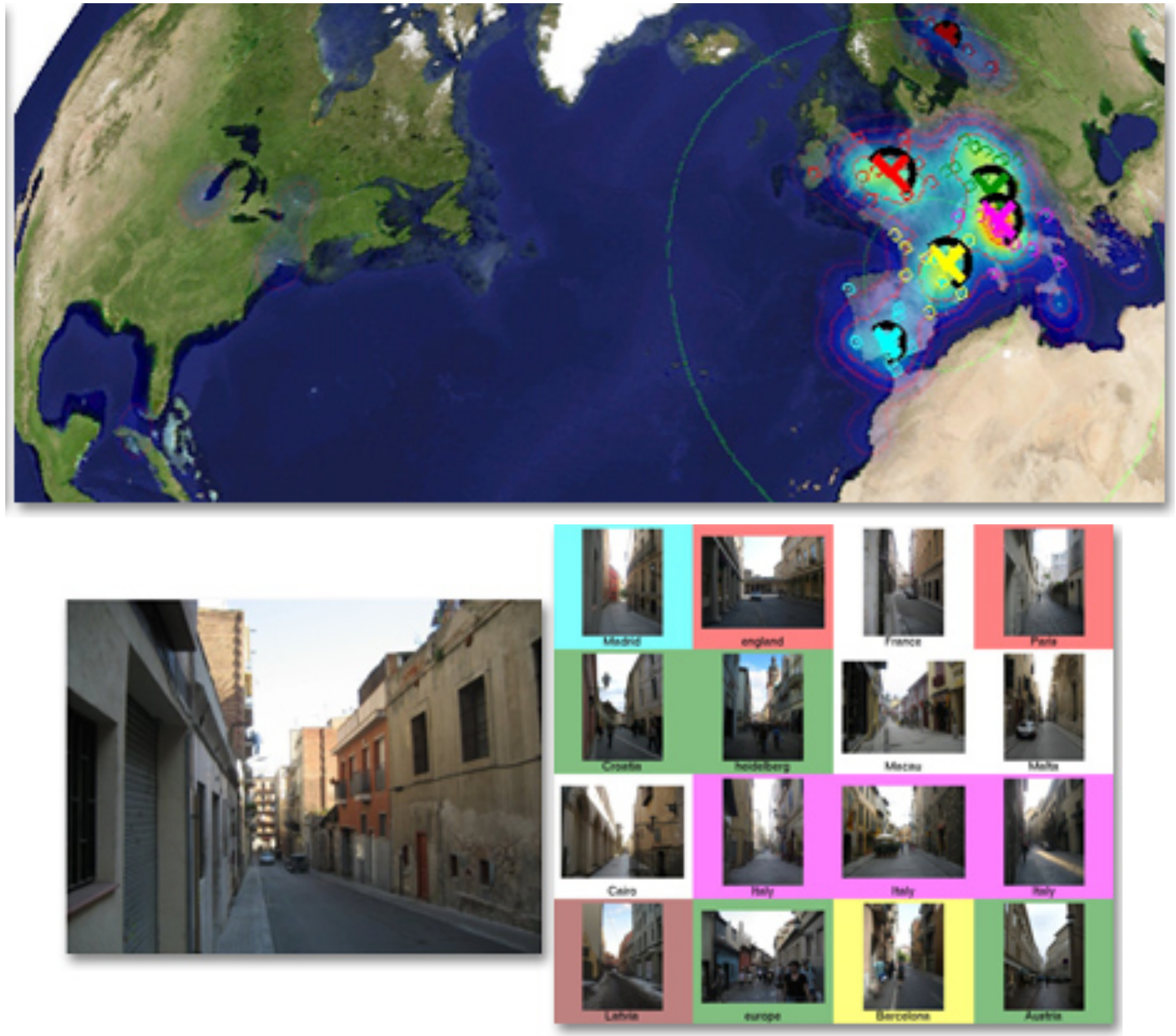


[Everingham et al. 2006-2012]

Effectiveness of **Data**

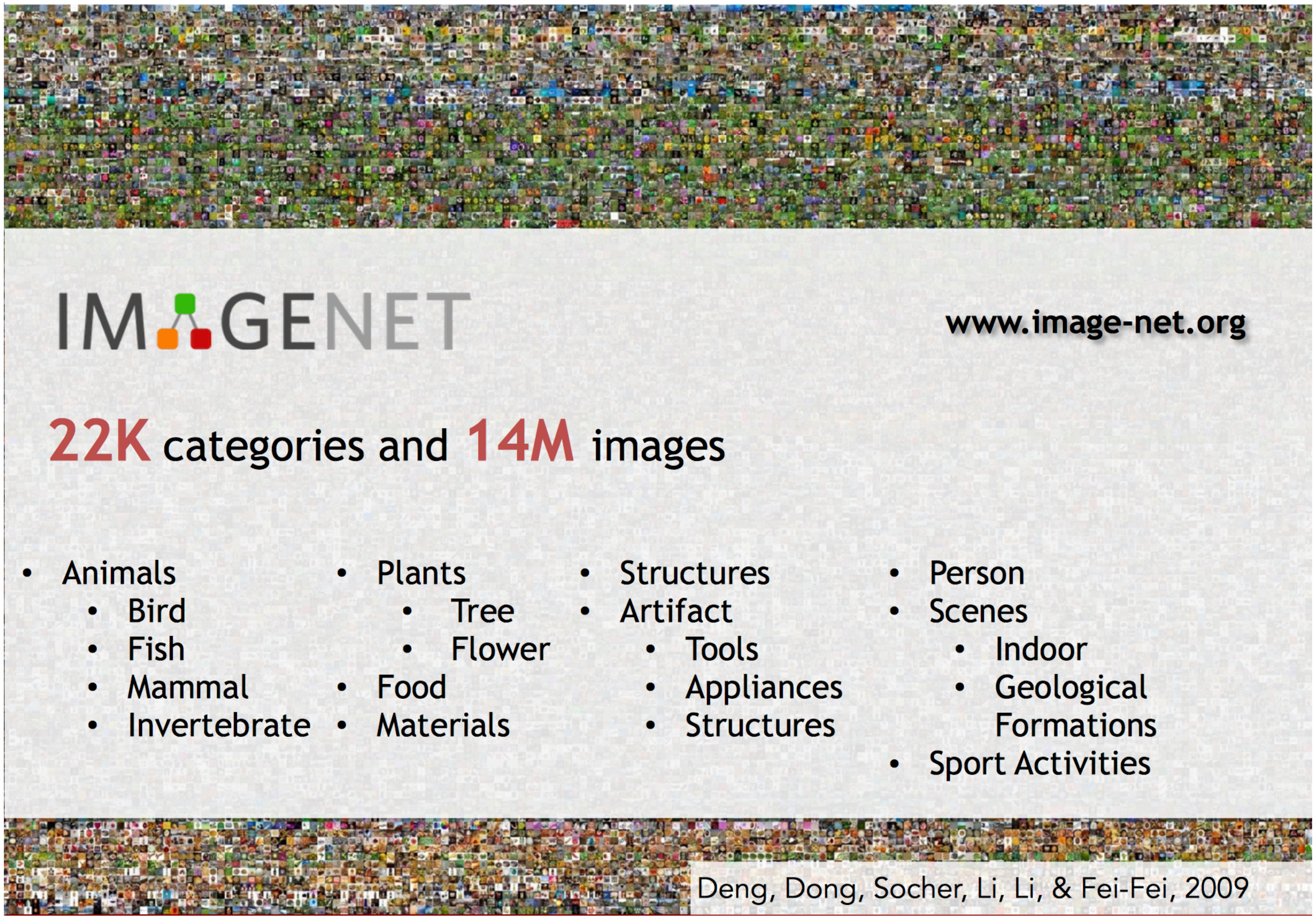


[Hays, Efros, ACM Siggraph 2007]

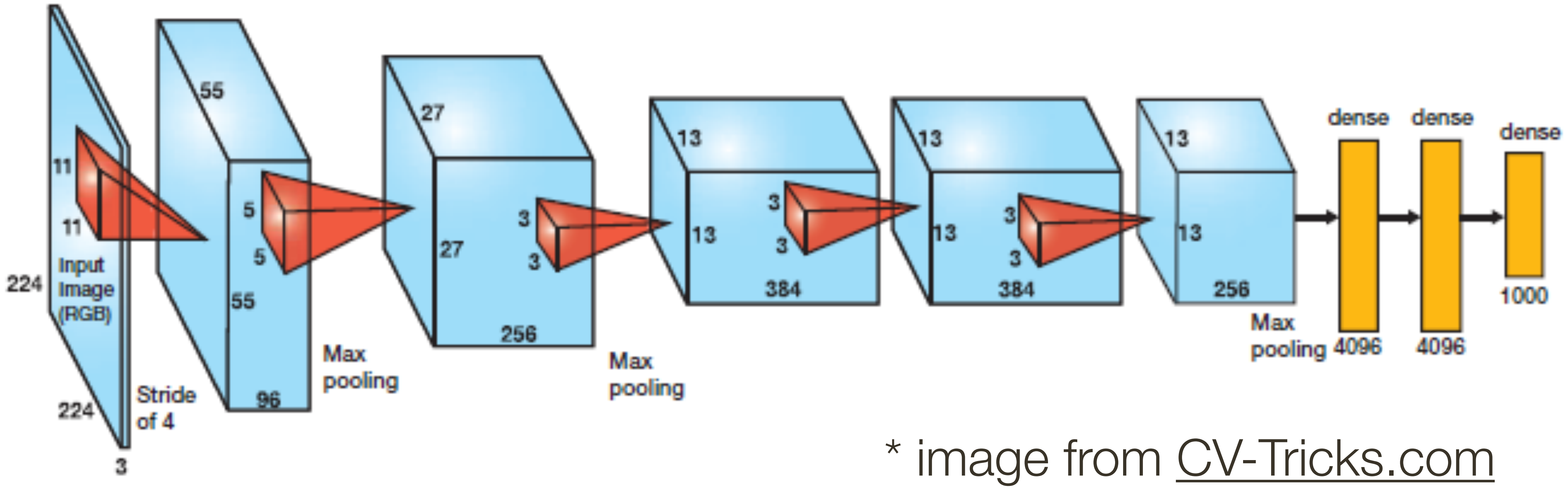


[Hays, Efros, CVPR 2008]

ImageNet Bechmark

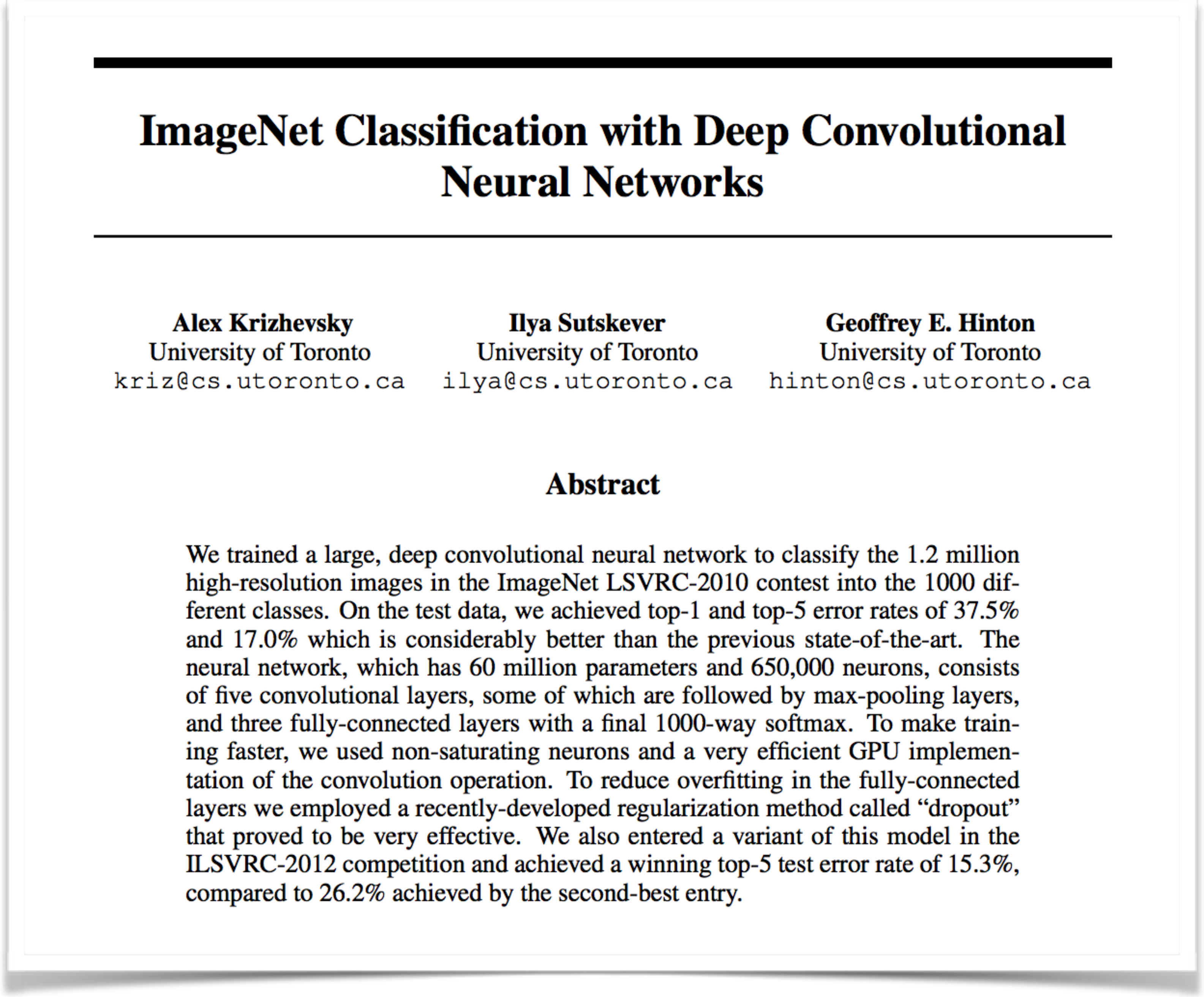


AlexNet on ImageNet



* image from CV-Tricks.com

Model	Top-1	Top-5
Sparse coding [2]	*47.1%*	*28.2%*
SIFT + FVs [24]	*45.7%*	*25.7%*
CNN	**37.5%**	**17.0%**



ImageNet Classification with Deep Convolutional Neural Networks

Alex Krizhevsky
University of Toronto
kriz@cs.utoronto.ca

Ilya Sutskever
University of Toronto
ilya@cs.utoronto.ca

Geoffrey E. Hinton
University of Toronto
hinton@cs.utoronto.ca

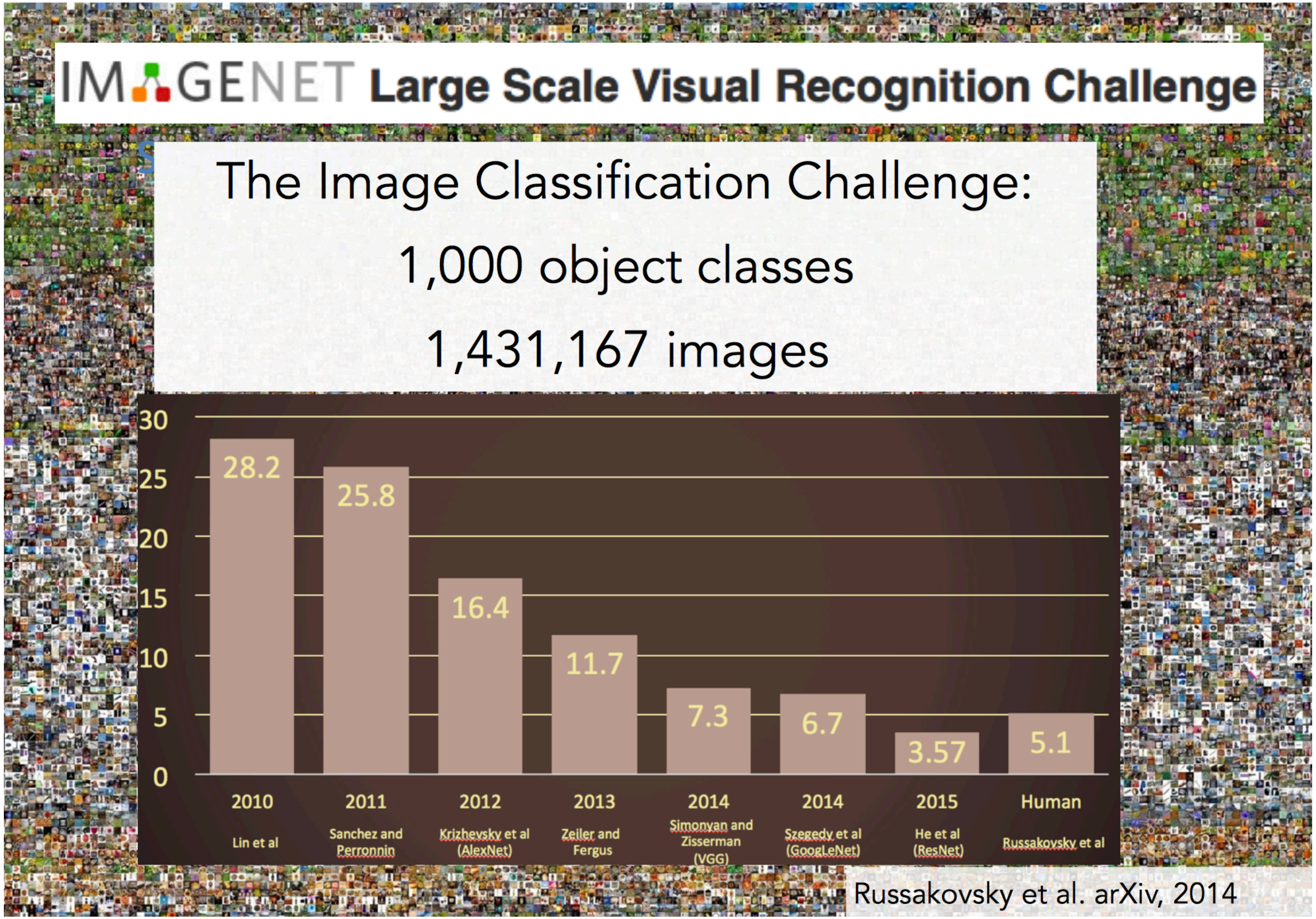
Abstract

We trained a large, deep convolutional neural network to classify the 1.2 million high-resolution images in the ImageNet LSVRC-2010 contest into the 1000 different classes. On the test data, we achieved top-1 and top-5 error rates of 37.5% and 17.0% which is considerably better than the previous state-of-the-art. The neural network, which has 60 million parameters and 650,000 neurons, consists of five convolutional layers, some of which are followed by max-pooling layers, and three fully-connected layers with a final 1000-way softmax. To make training faster, we used non-saturating neurons and a very efficient GPU implementation of the convolution operation. To reduce overfitting in the fully-connected layers we employed a recently-developed regularization method called "dropout" that proved to be very effective. We also entered a variant of this model in the ILSVRC-2012 competition and achieved a winning top-5 test error rate of 15.3%, compared to 26.2% achieved by the second-best entry.

Model	Top-1 (val)	Top-5 (val)	Top-5 (test)
SIFT + FVs [7]	—	—	*26.2%*
1 CNN	40.7%	18.2%	—
5 CNNs	38.1%	16.4%	**16.4%**
1 CNN*	39.0%	16.6%	—
7 CNNs*	36.7%	15.4%	**15.3%**

[Krizhevsky, Sutskever, Hinton, NIPS 2012]

Success of **Deep Learning**



Final thought …

— Model based, compositional, primitives, inverse graphics

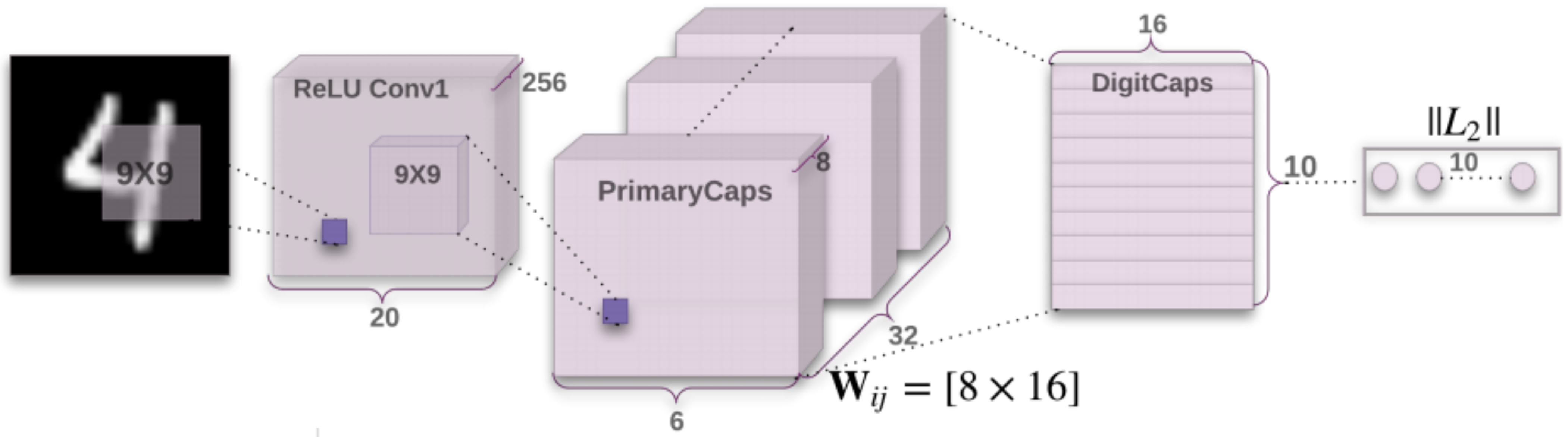
— Hand-crafted features for given invariances & matching

— Hand-crafted features with learned statistical models on top

— Joint learning of features and statistical models for recognition

Capsule**NET**

Going **back to inverse graphics**



[Sabour, Frosst, Hinton, NIPS 2017]

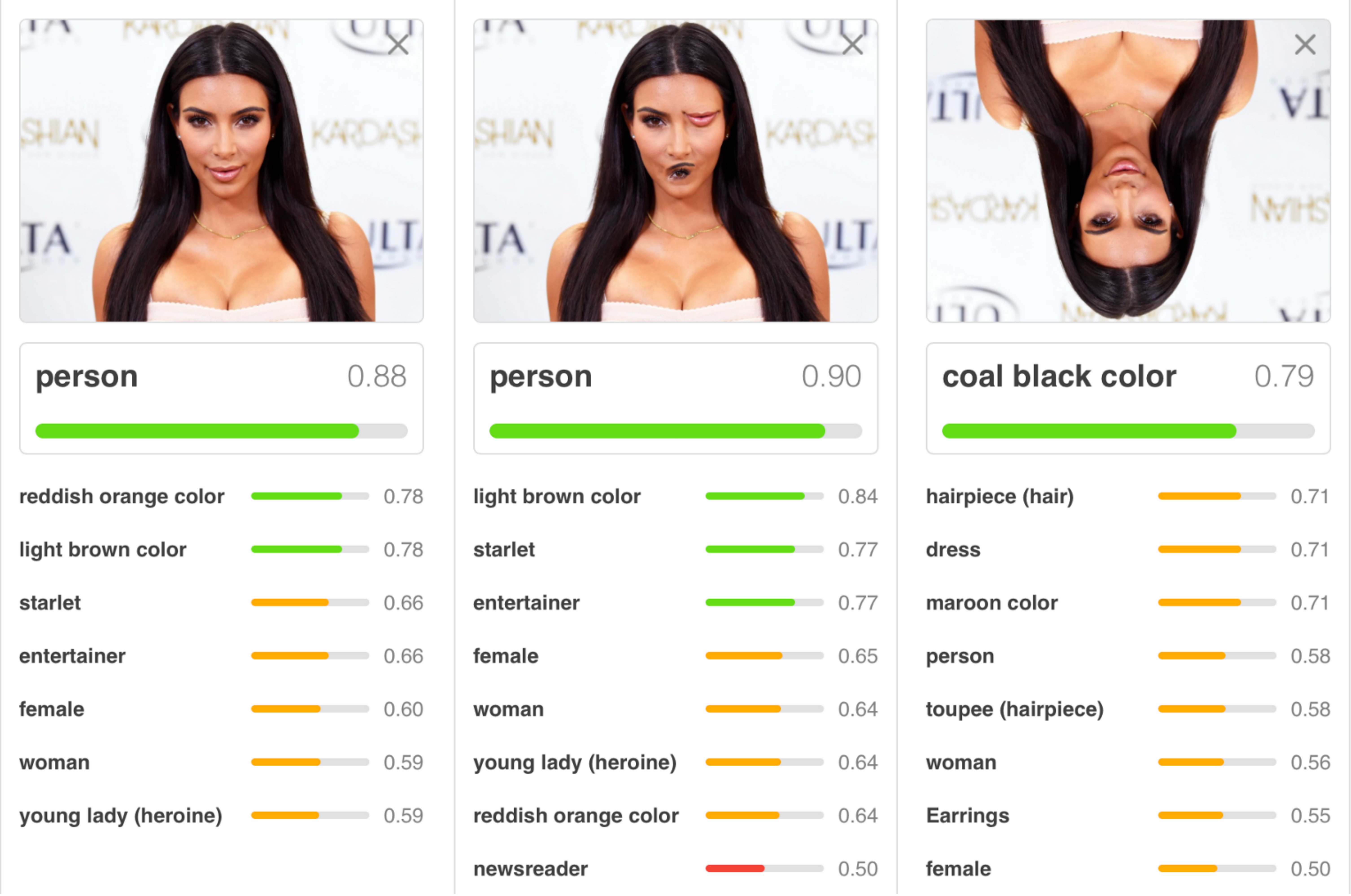
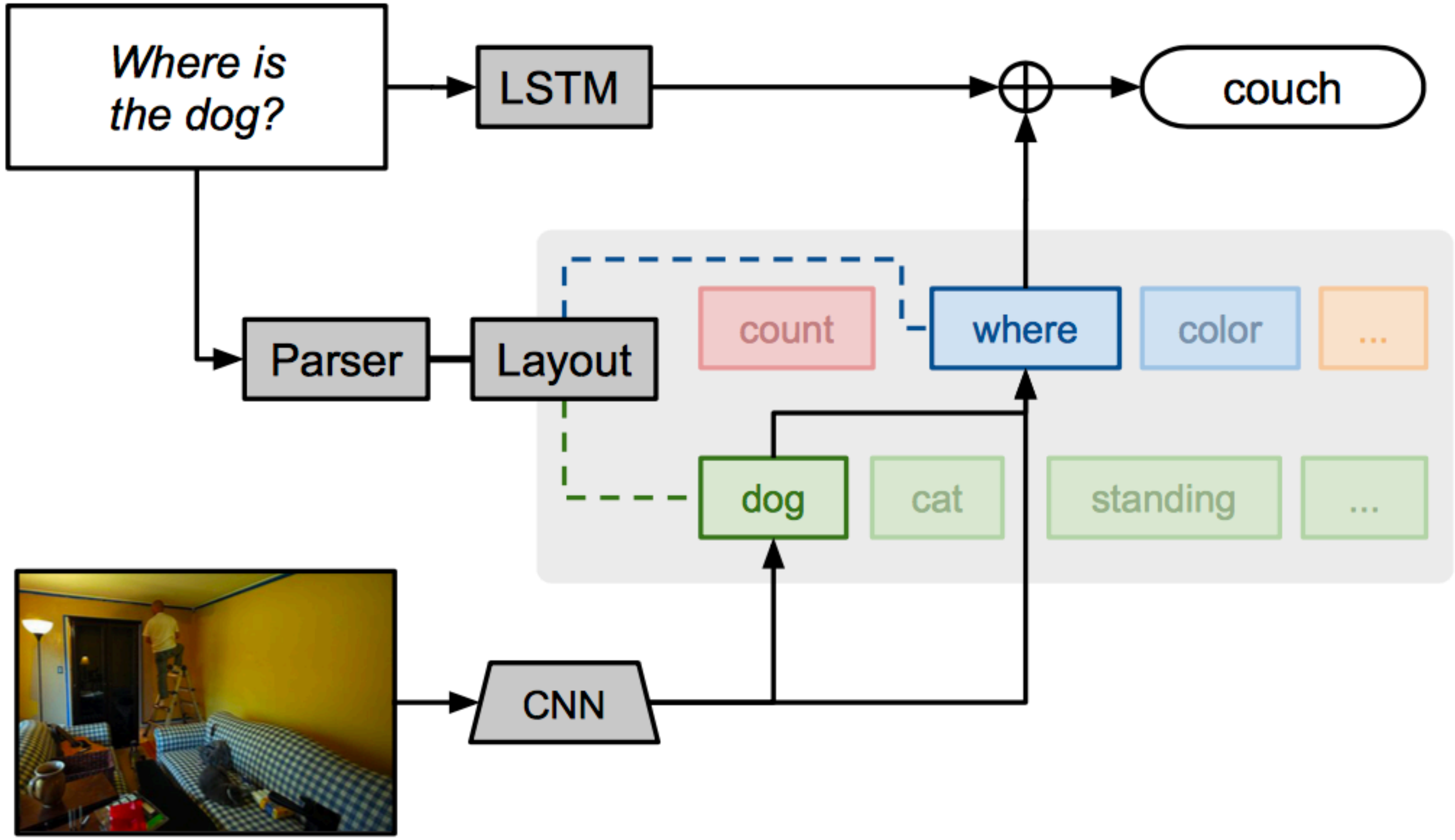


*image credit medium.com

Neural **Modular Networks**



[Andreas, Rohrbach, Darrell, Klein, CVPR 2016]