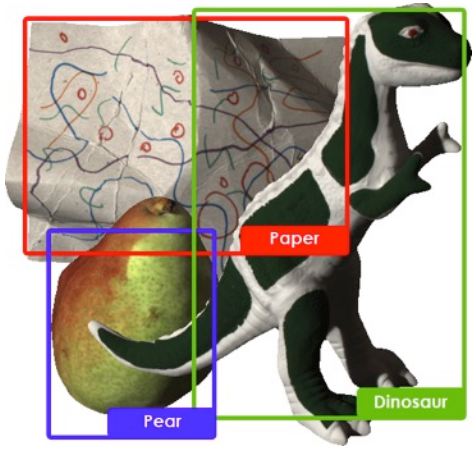
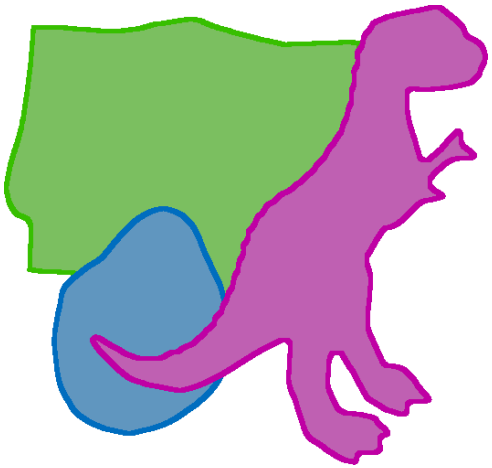
The Three R's of Computer Vision:
Recognition, Reconstruction & Reorganization

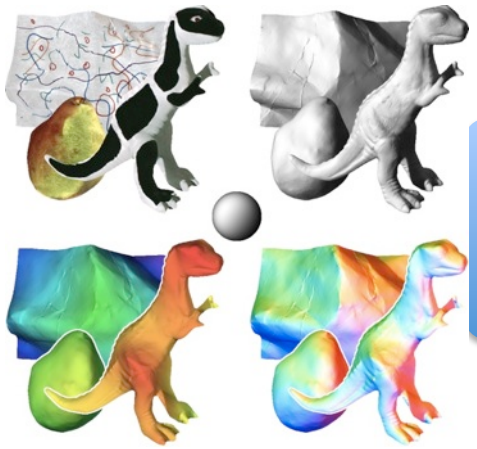
Jitendra Malik
UC Berkeley

Recognition, Reconstruction & Reorganization



Recognition

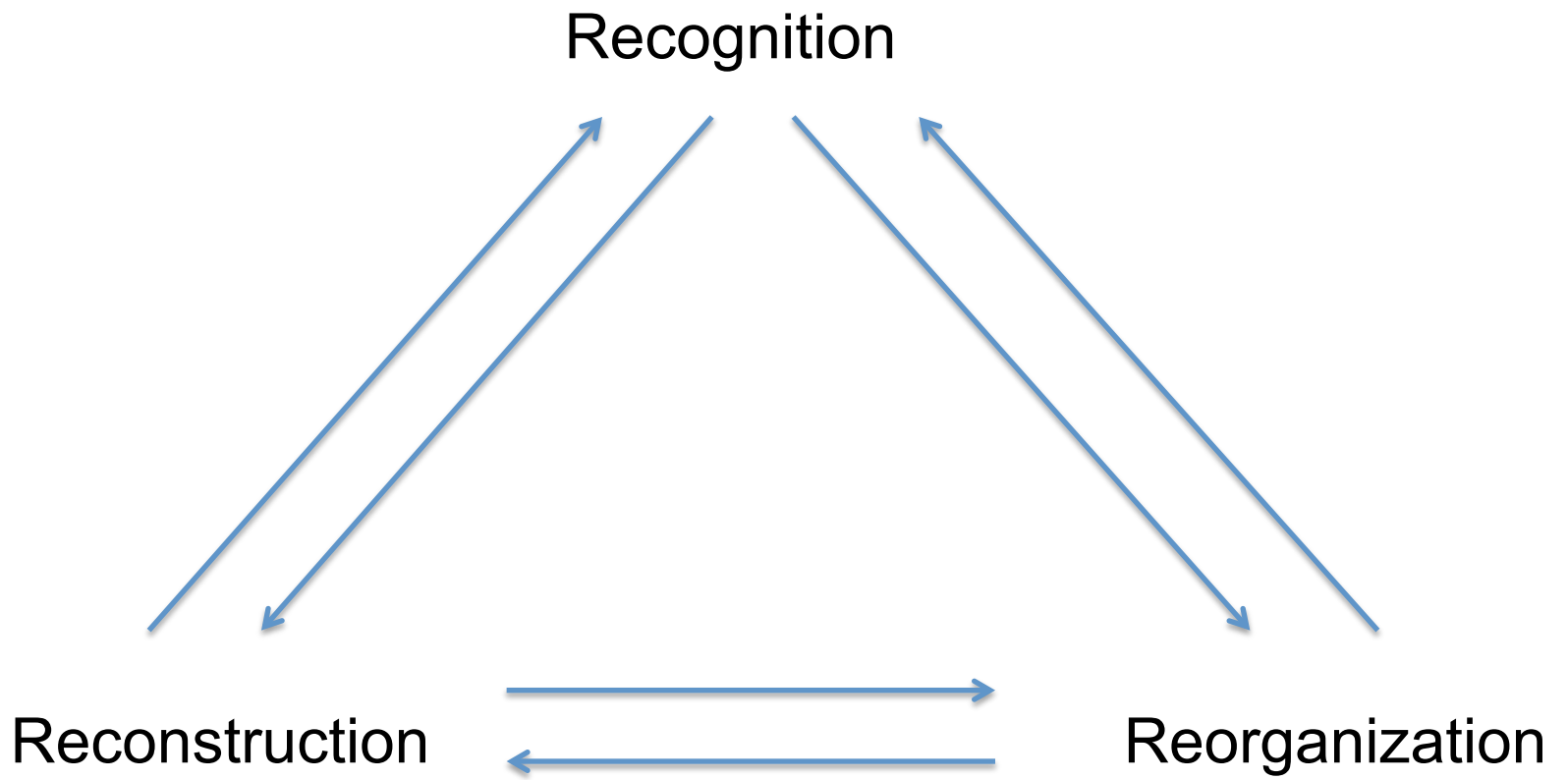




Reconstruction

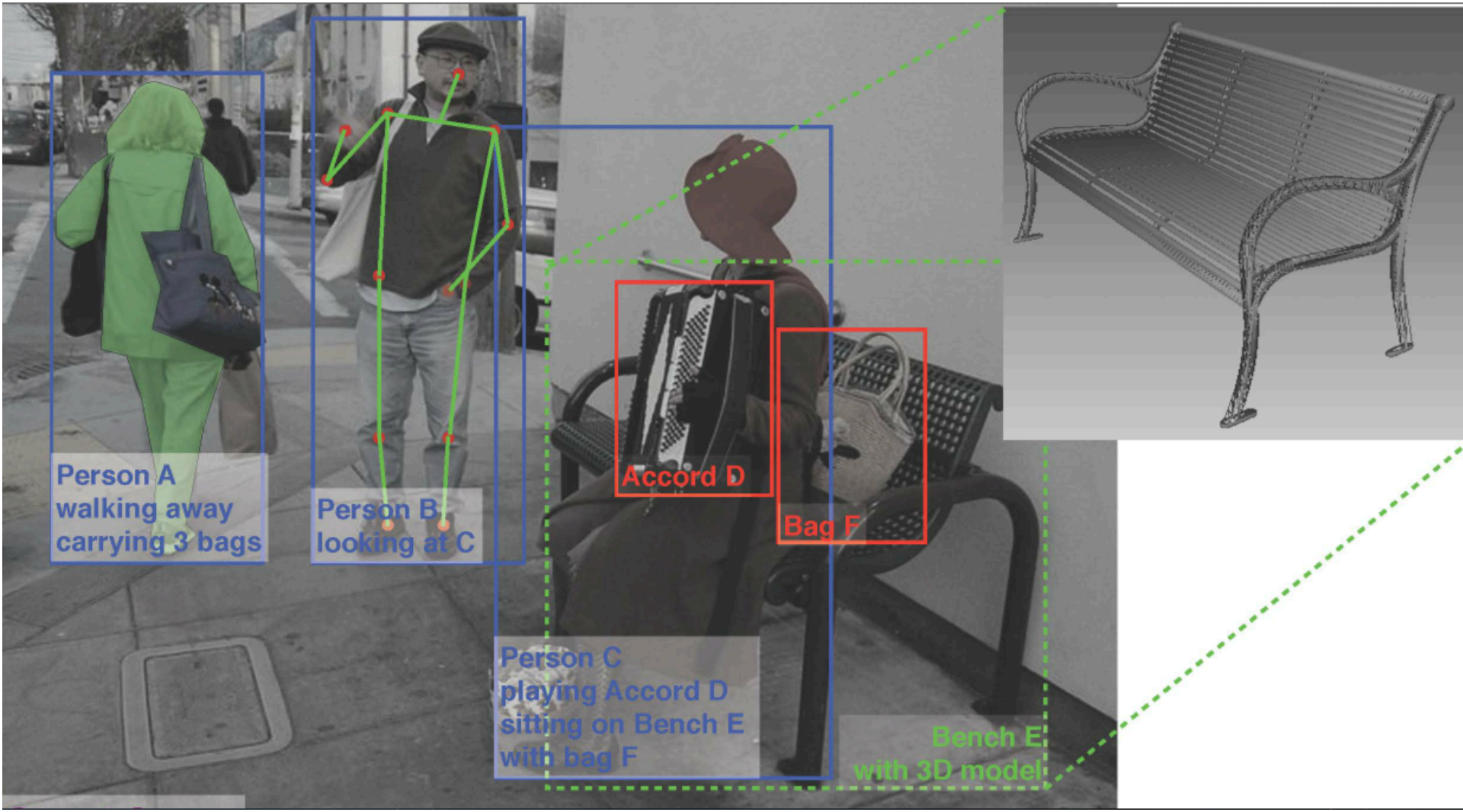
Reorganization

The Three R's of Vision



Each of the 6 directed arcs in this diagram is a useful direction of information flow

What we would like to infer…



Will person B put some money into Person C's tip bag?

Different aspects of vision

- Perception: study the “laws of seeing” -predict what a human would perceive in an image.
- Neuroscience: understand the mechanisms in the retina and the brain
- Function: how laws of optics, and the statistics of the world we live in, make certain interpretations of an image more likely to be valid

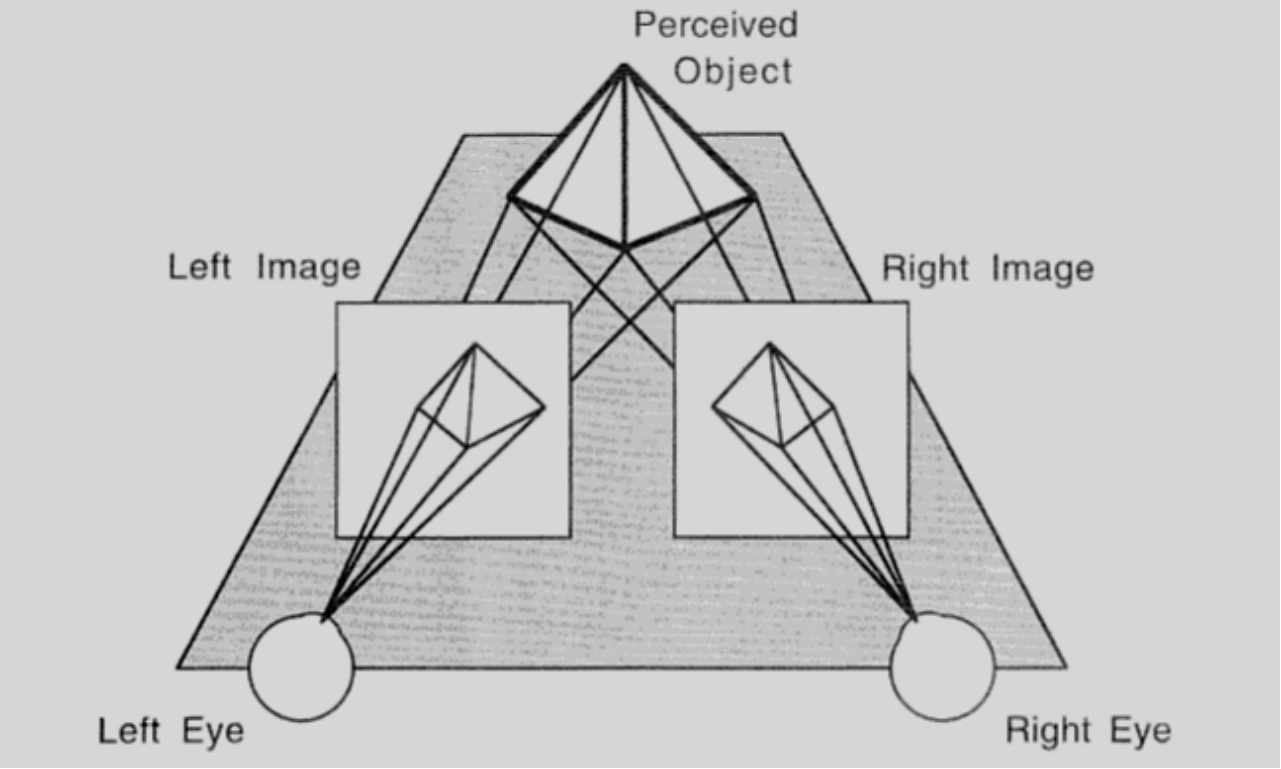
The match between human and computer vision is strongest at the level of function, but since typically the results of computer vision are meant to be conveyed to humans makes it useful to be consistent with human perception. Neuroscience is a source of ideas but being bio-mimetic is not a requirement.

Facts about the Visual World

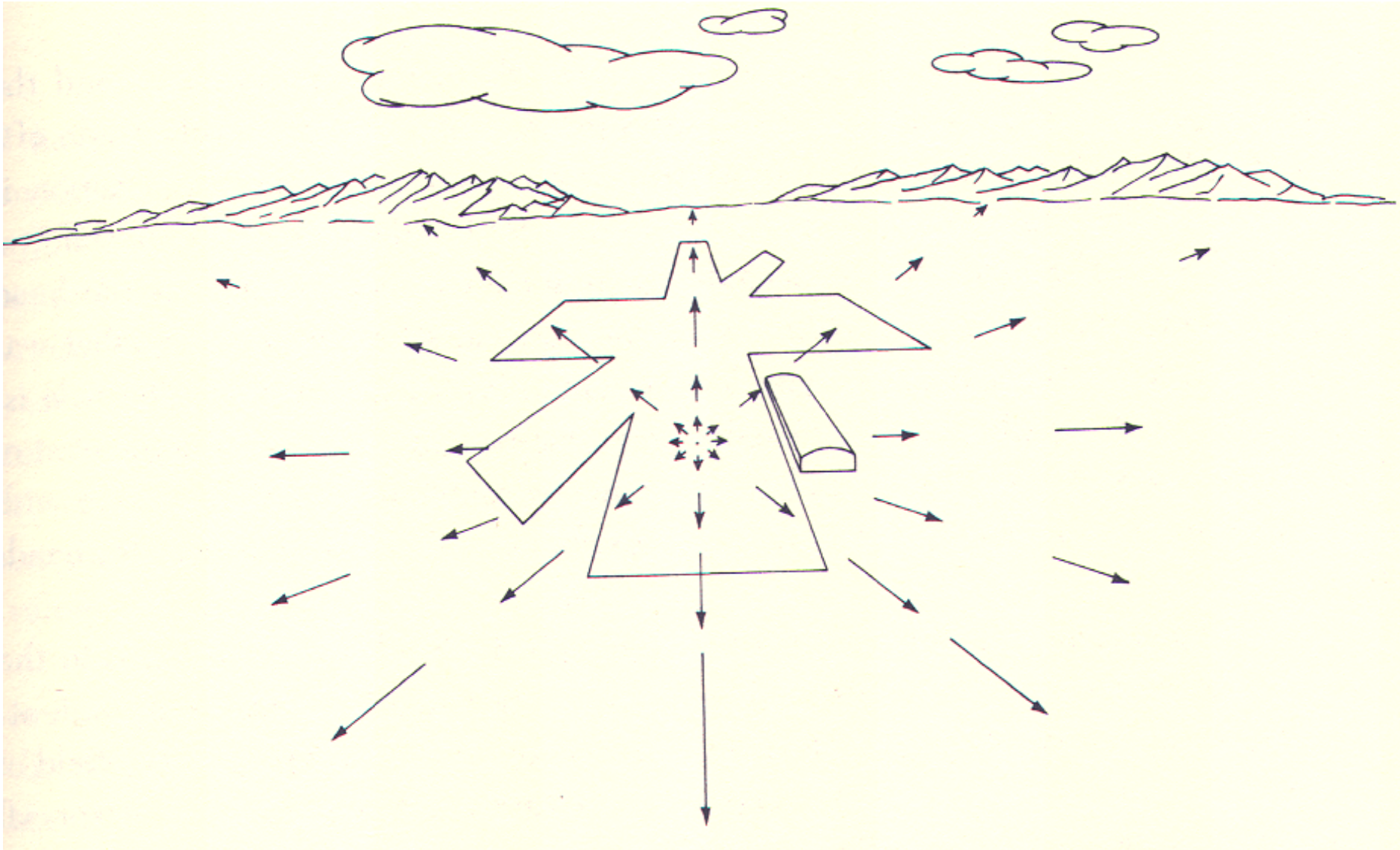
- The world consists of rigid, or piecewise rigid, objects
- Object surfaces have piecewise constant color and texture
- Objects in a category share parts
- Projection from the 3D world to 2D image is uniquely defined
- Objects are opaque & nearer objects occlude farther objects.
- Objects occur at varying distances and locations in the image
- Objects occur in context, stereotypical relations to each other
- Actions are performed by agents with goals and intentions
- ….

We can incorporate these into the design of visual processing architectures. Parameters should be learnt.

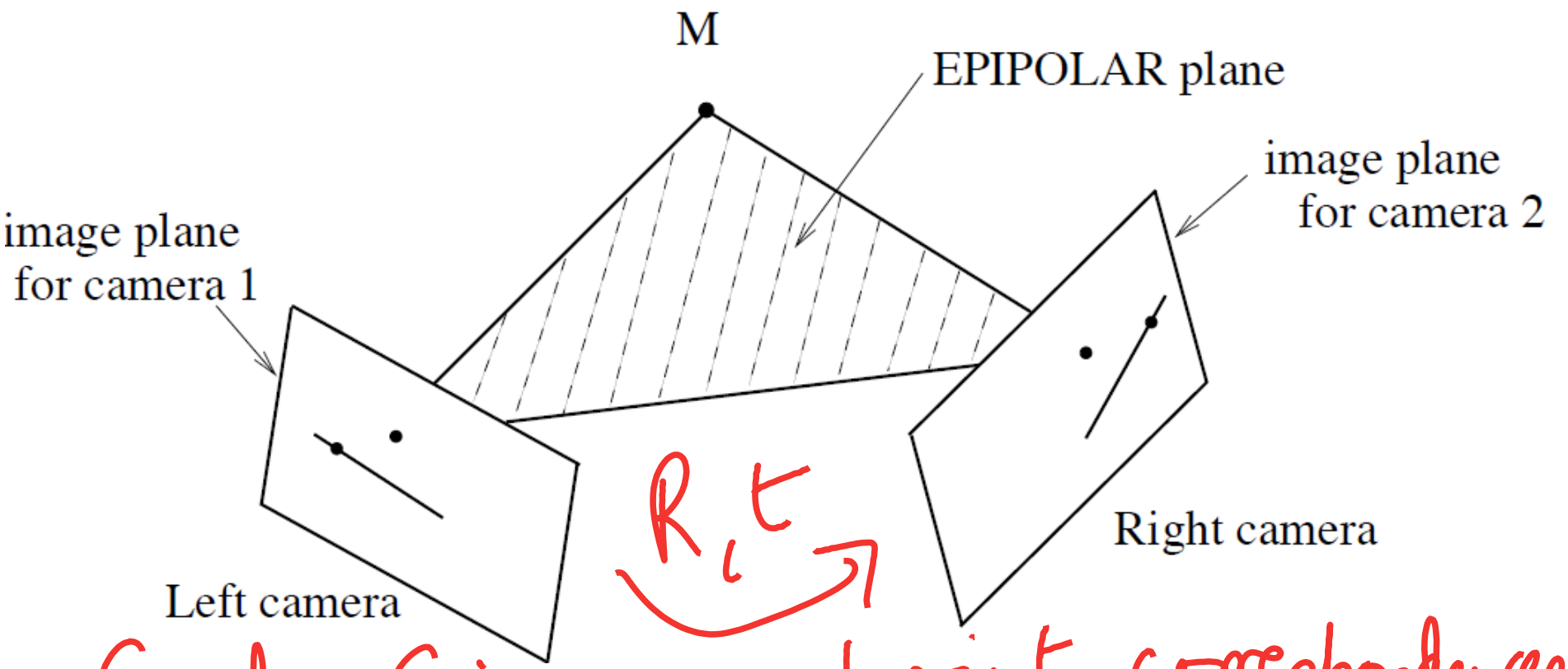
Binocular Stereopsis



Optical flow is a basic cue for all animals



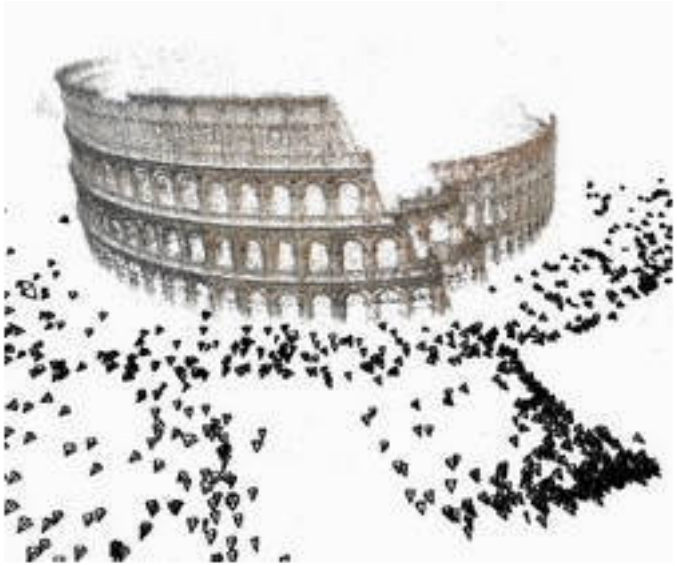
Epipolar geometry for cameras in general position



Goal: Given n point correspondences, estimate R, t and depths at the n points

State of the Art in Reconstruction

- Multiple photographs



Credit: http://grail.cs.washington.edu/rome/

Agarwal et al (2010)

Frahm et al, (2010)

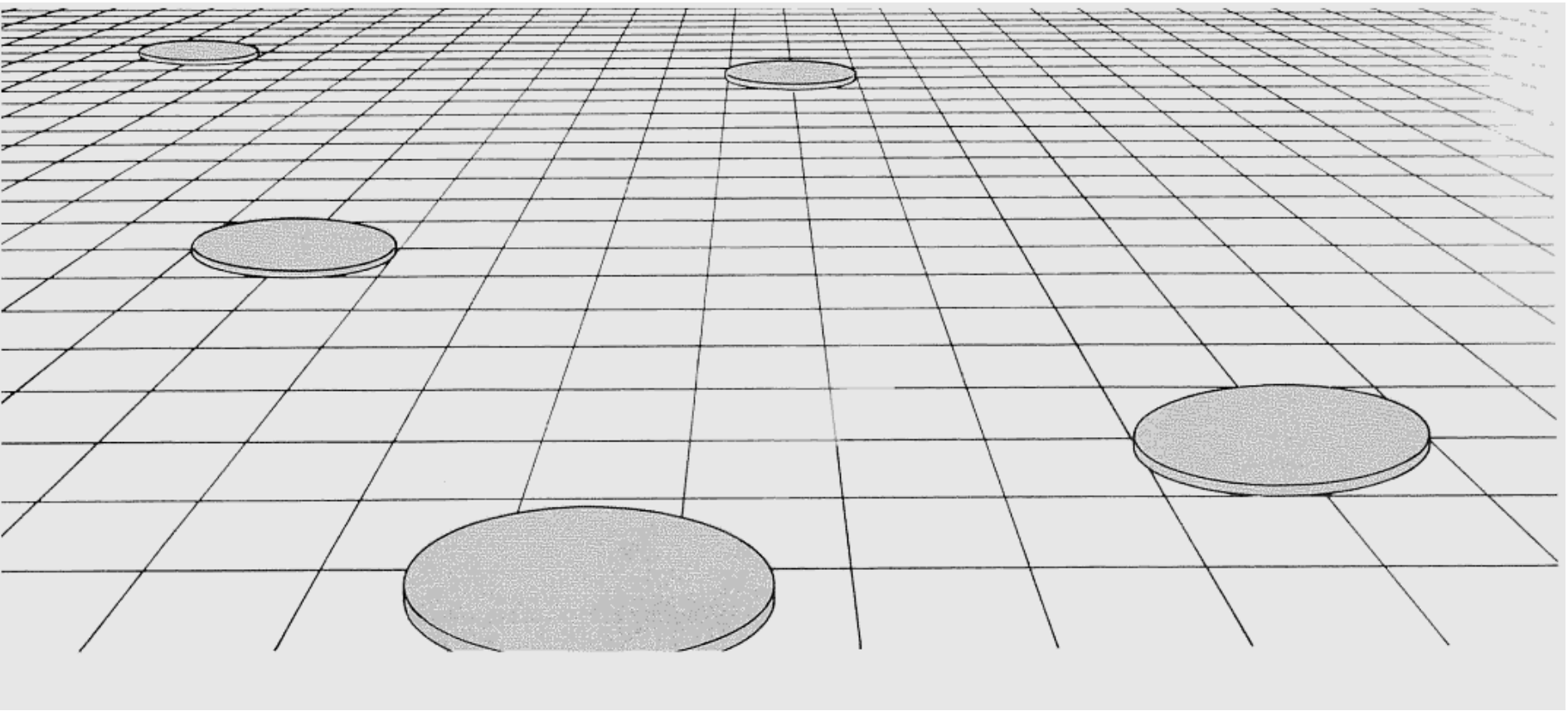
- Range Sensors

Kinect (PrimeSense)

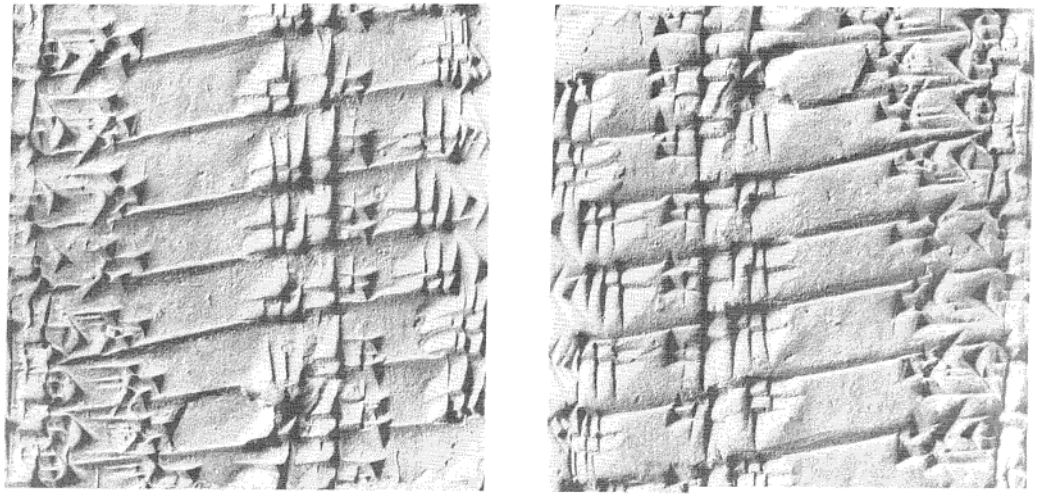
Velodyne Lidar

Semantic Segmentation is needed to make this more useful…

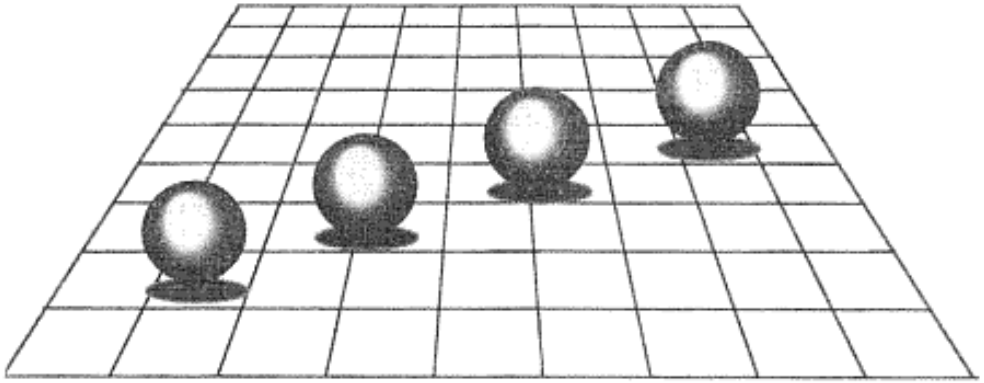
Some Pictorial Cues



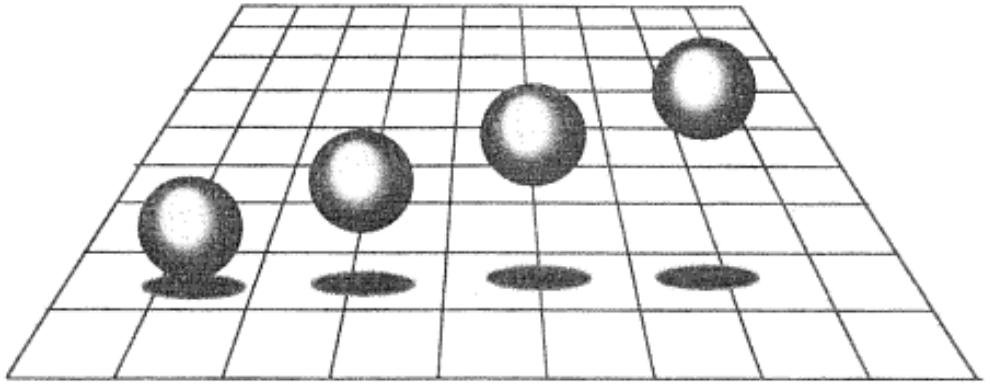
Shading



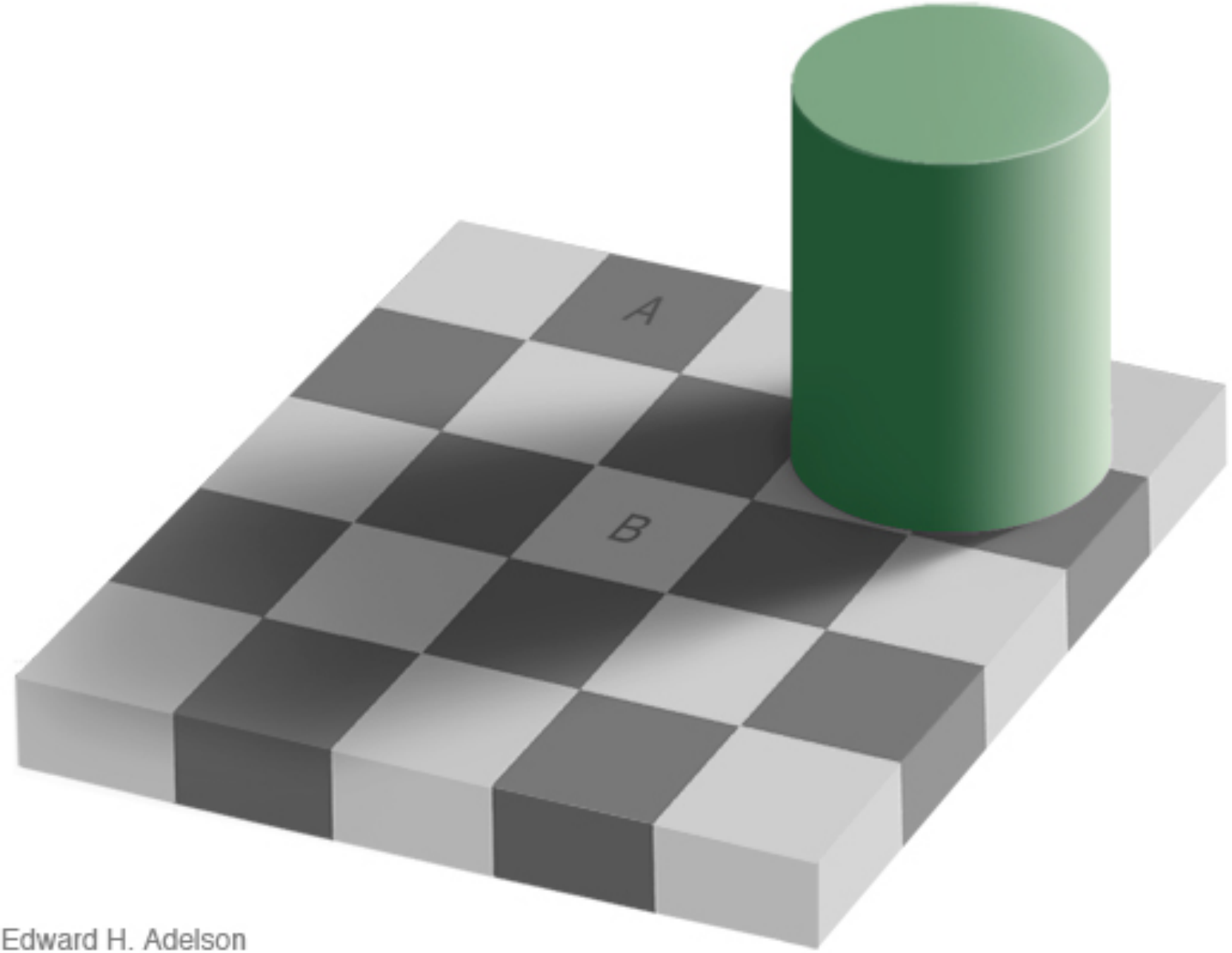
Cast Shadows



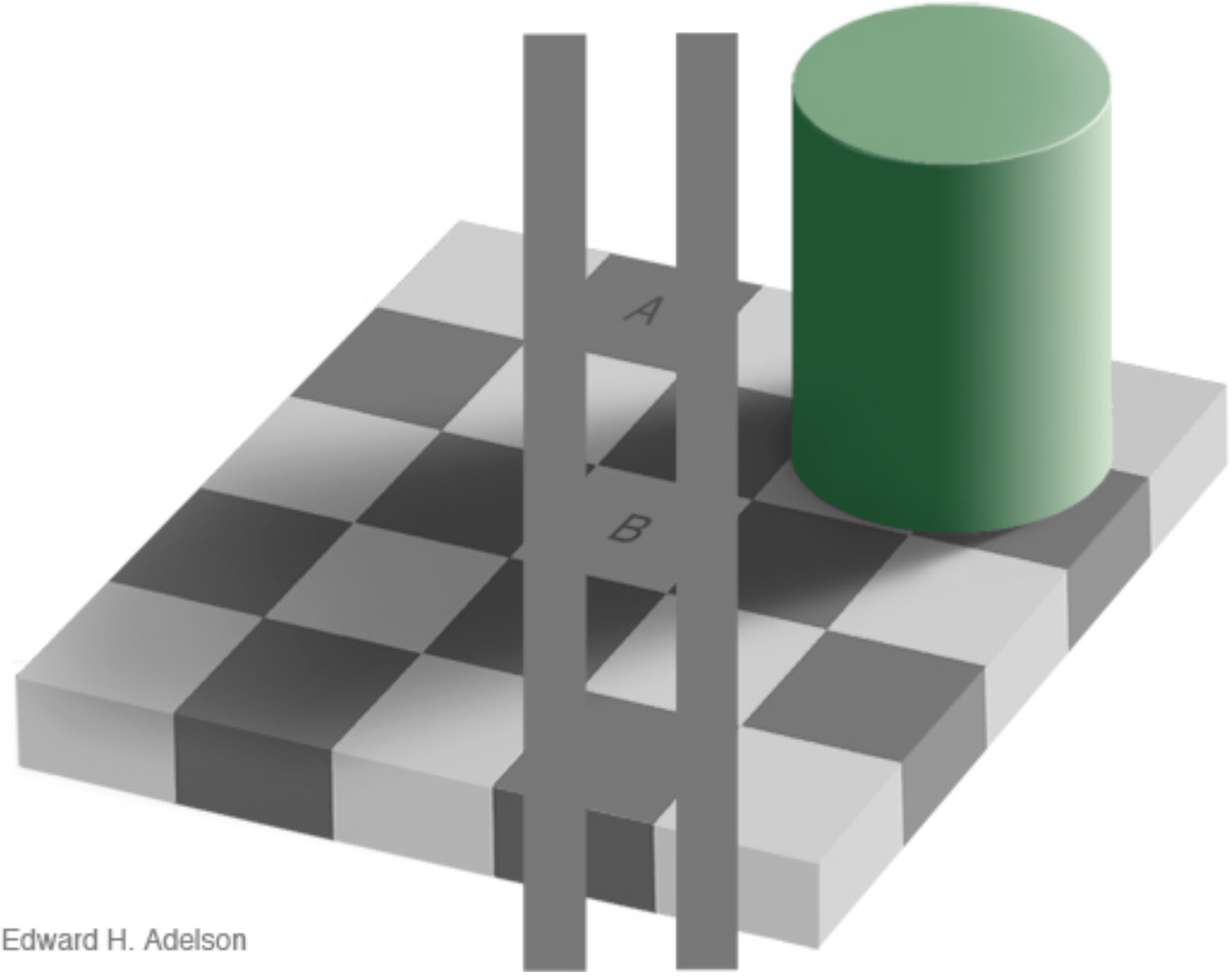
A



B

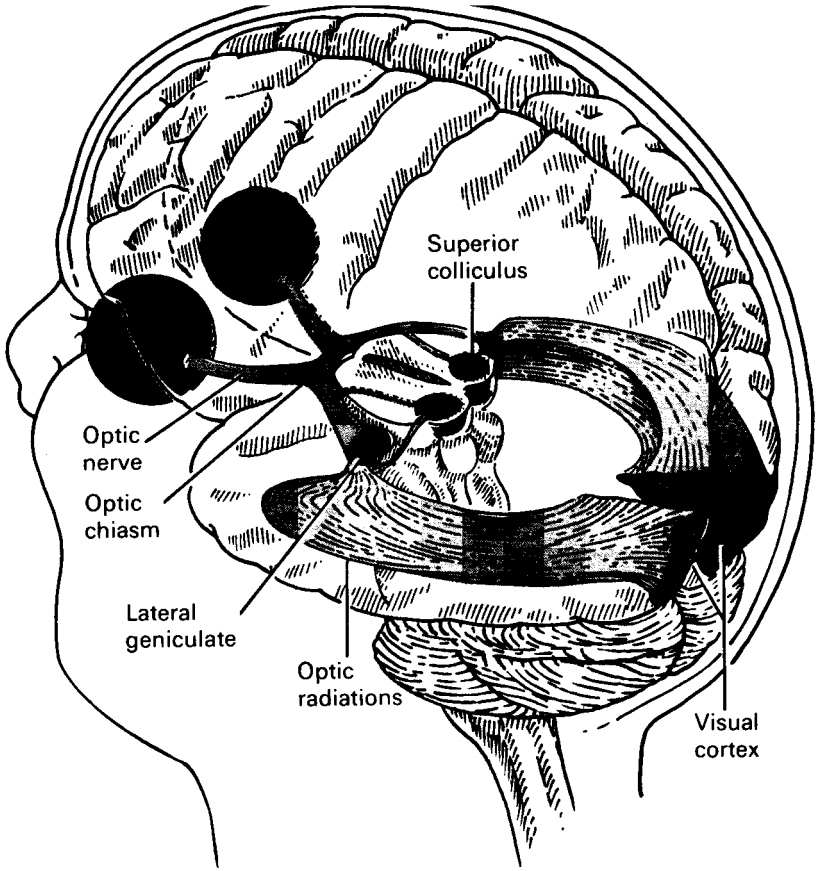


Edward H. Adelson

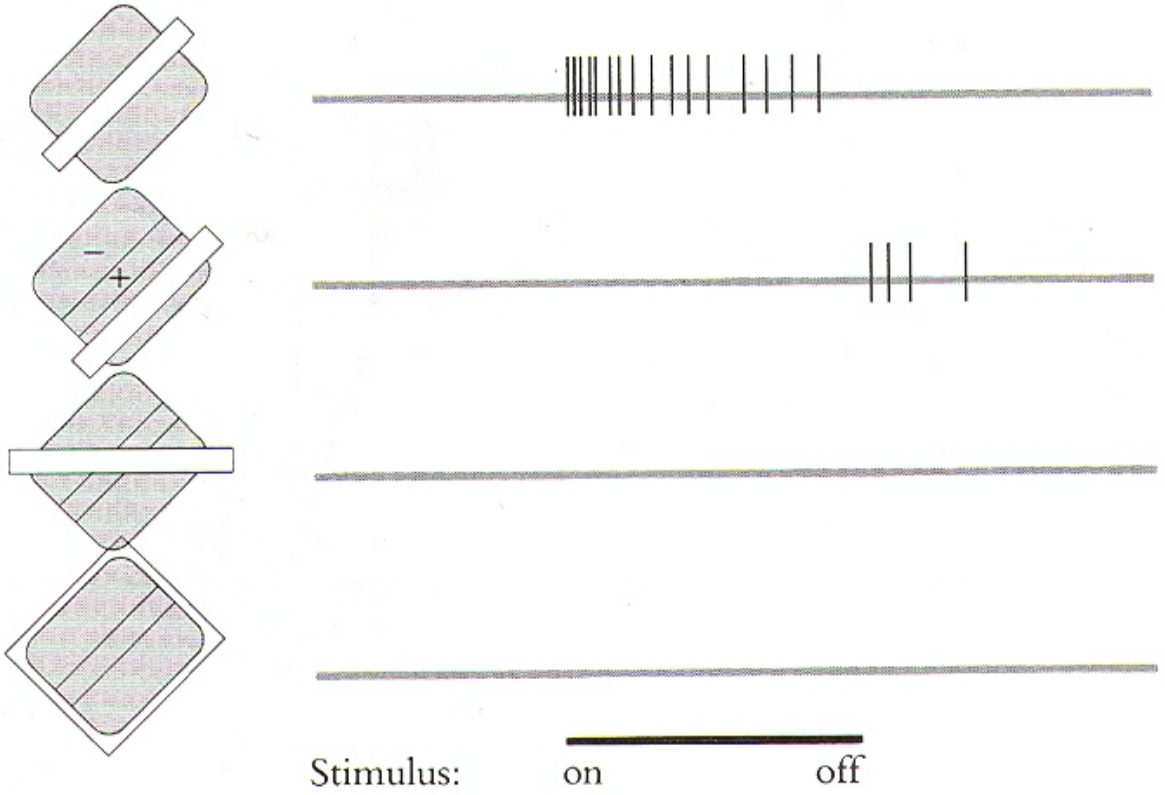


Edward H. Adelson

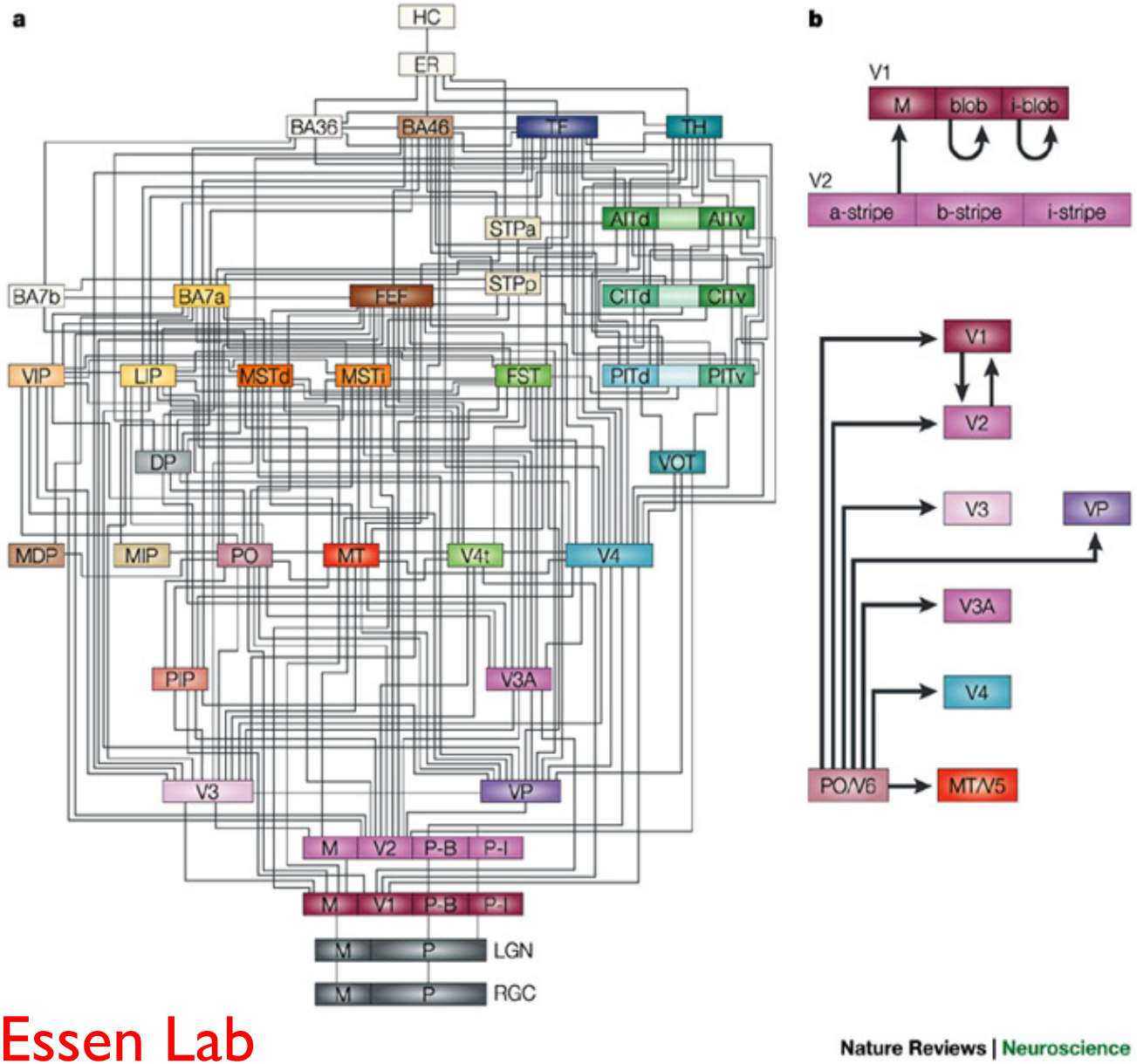
The Visual Pathway



Hubel and Wiesel (1962) discovered orientation sensitive neurons in V1

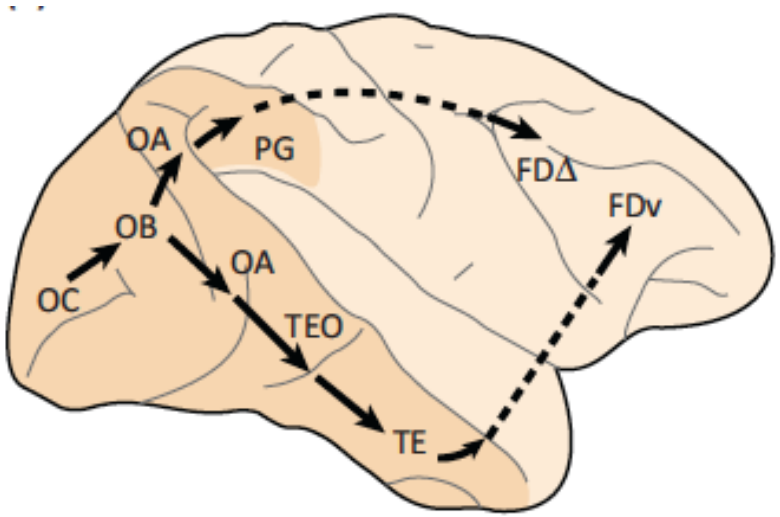


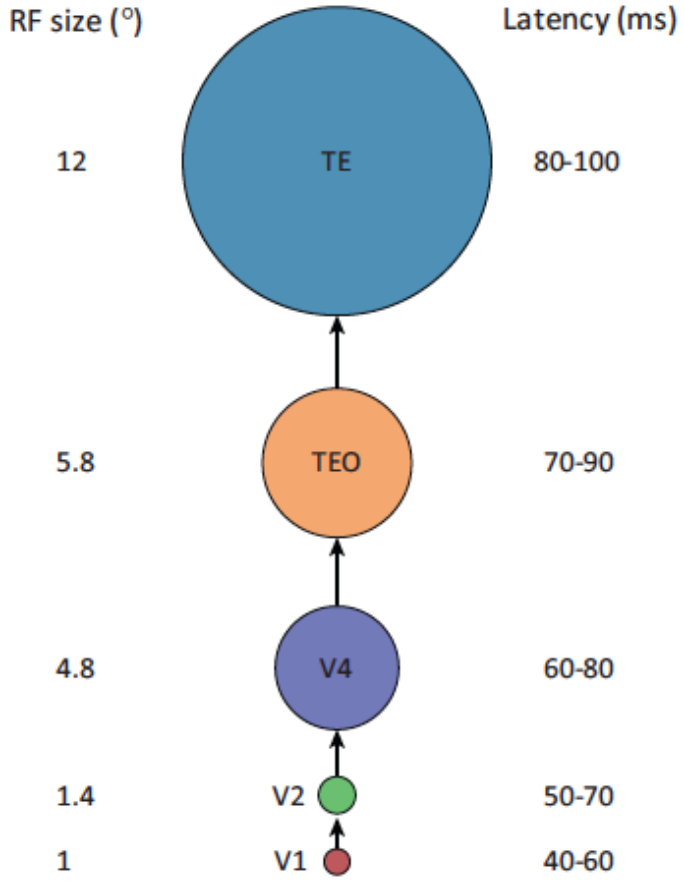
Block Diagram of the Primate Visual System



D. Van Essen Lab

Feed-forward model of the ventral stream





Kravitz et al, Trends in Cognitive Science 2013

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

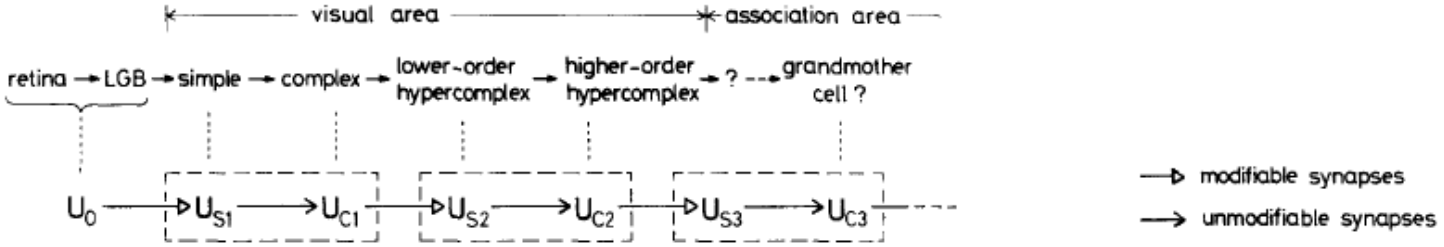


Fig. 1. Correspondence between the hierarchy model by Hubel and Wiesel, and the neural network of the neocognitron

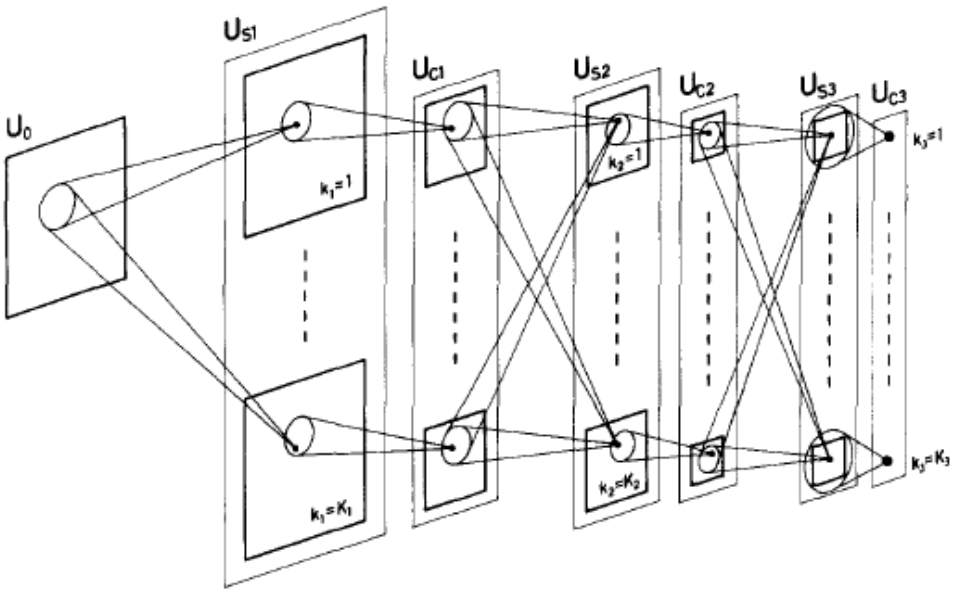


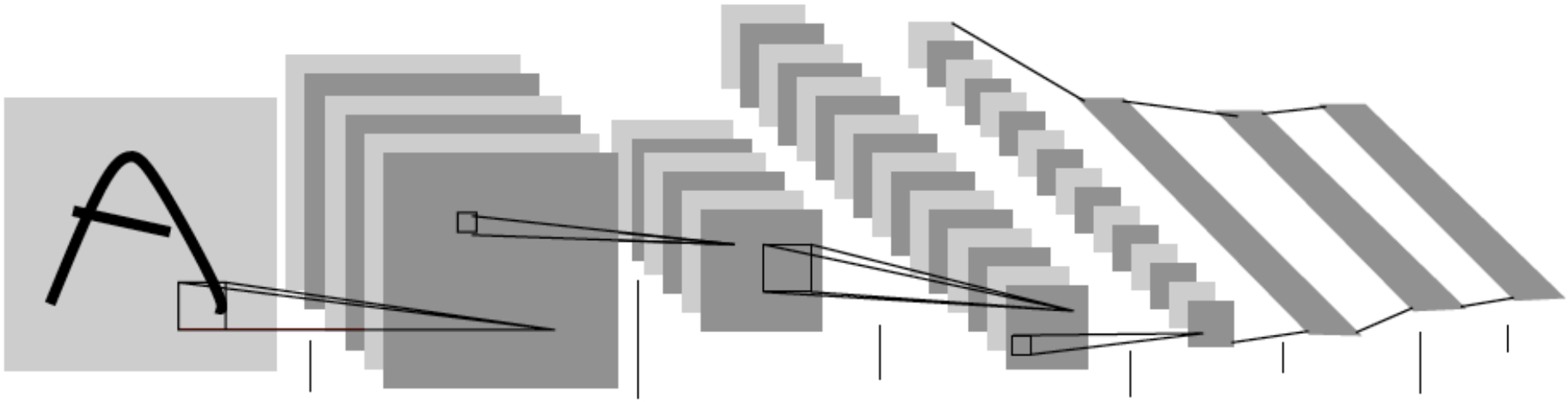
Fig. 2. Schematic diagram illustrating the interconnections between layers in the neocognitron

Biol. Cybernetics 36, 193–202 (1980)

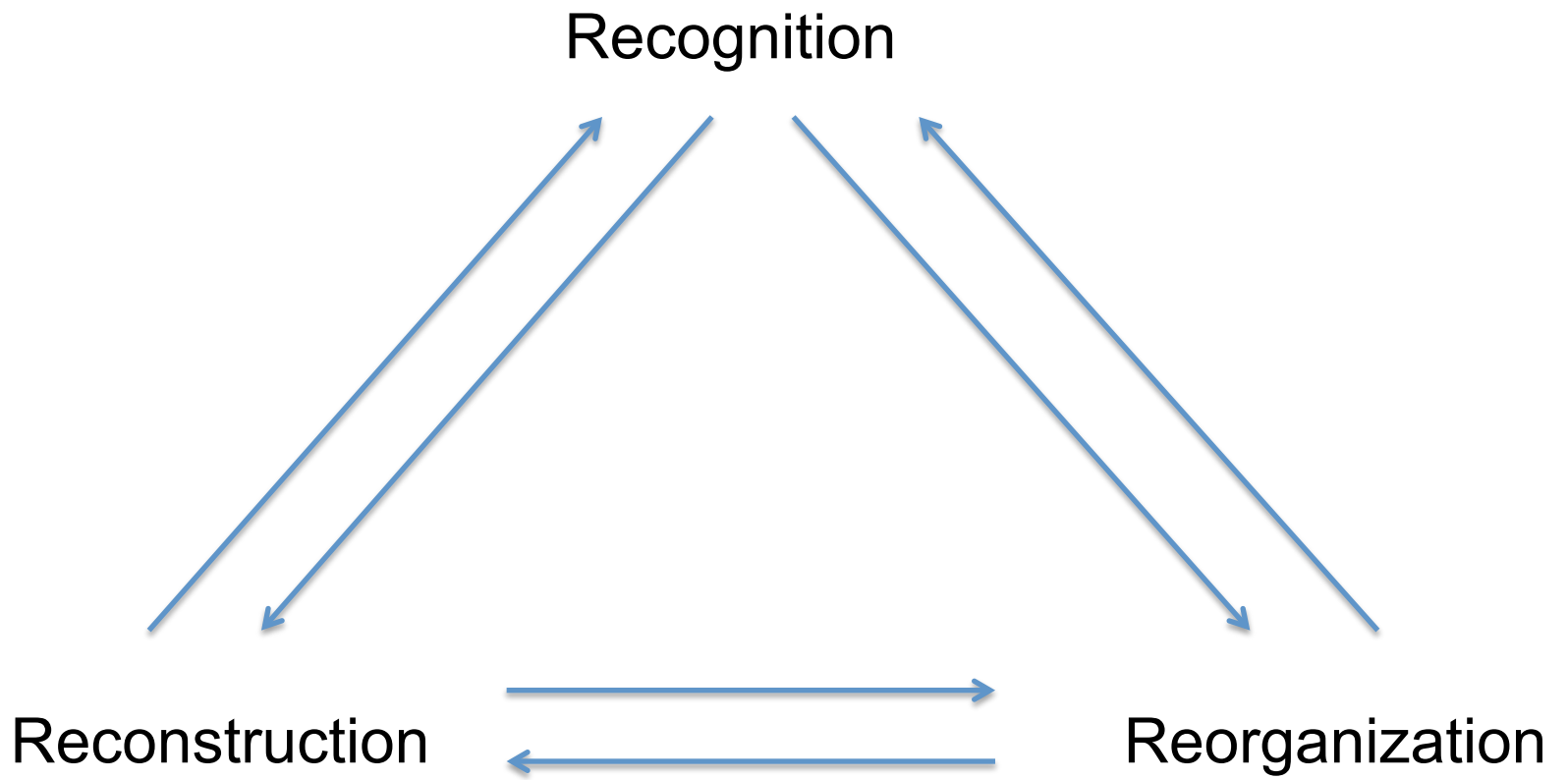
Convolutional Neural Networks (LeCun et al)

Used backpropagation to train the weights in this architecture

- First demonstrated by LeCun et al for handwritten digit recognition(1989)
- Krizhevsky, Sutskever & Hinton showed effectiveness for full image classification on ImageNet Challenge (2012)
- Girshick, Donahue, Darrell & Malik (arxiv, 2013)(CVPR 2014) showed that these features were also effective for object detection
- And many others…

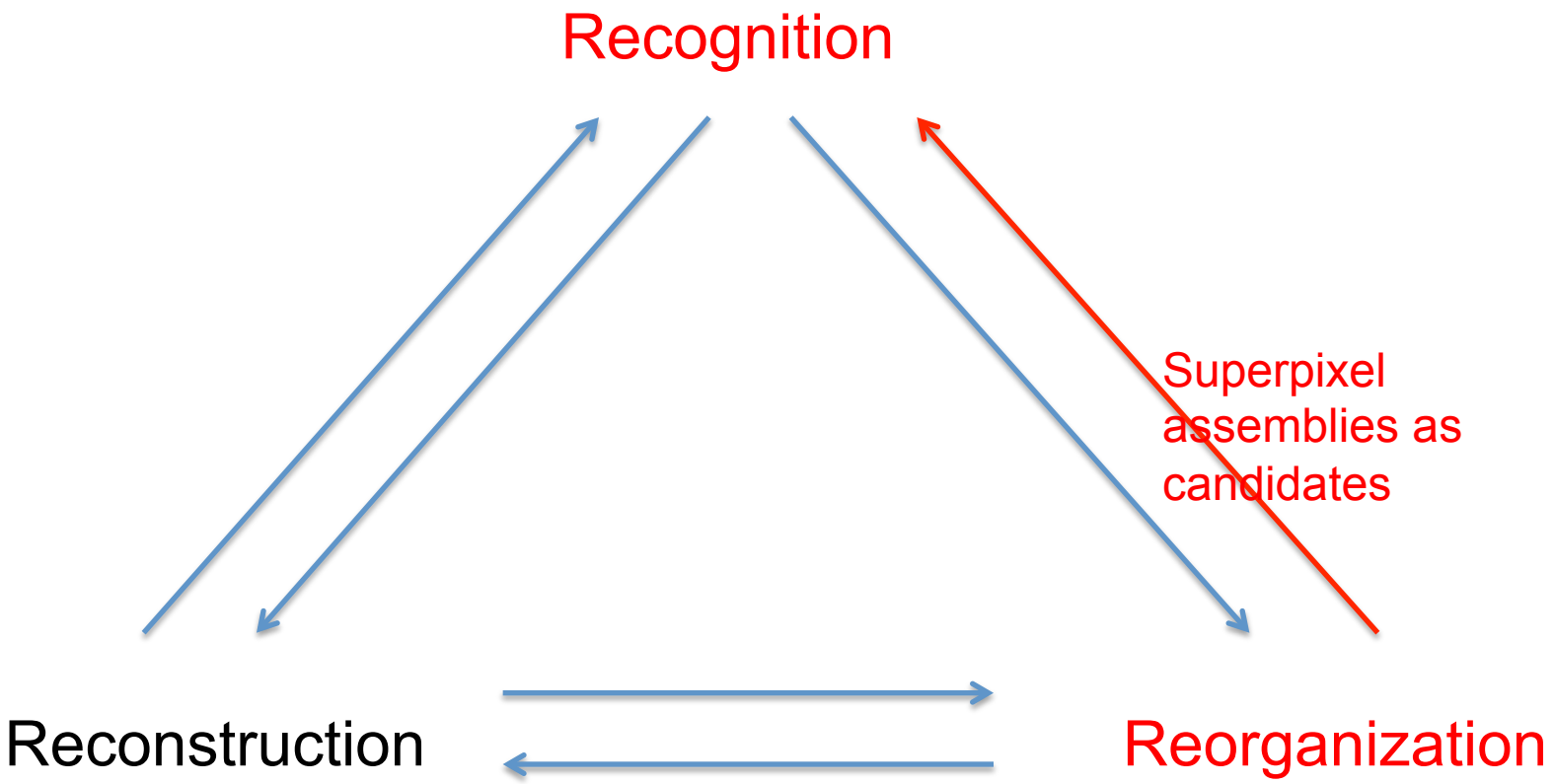


The Three R's of Vision

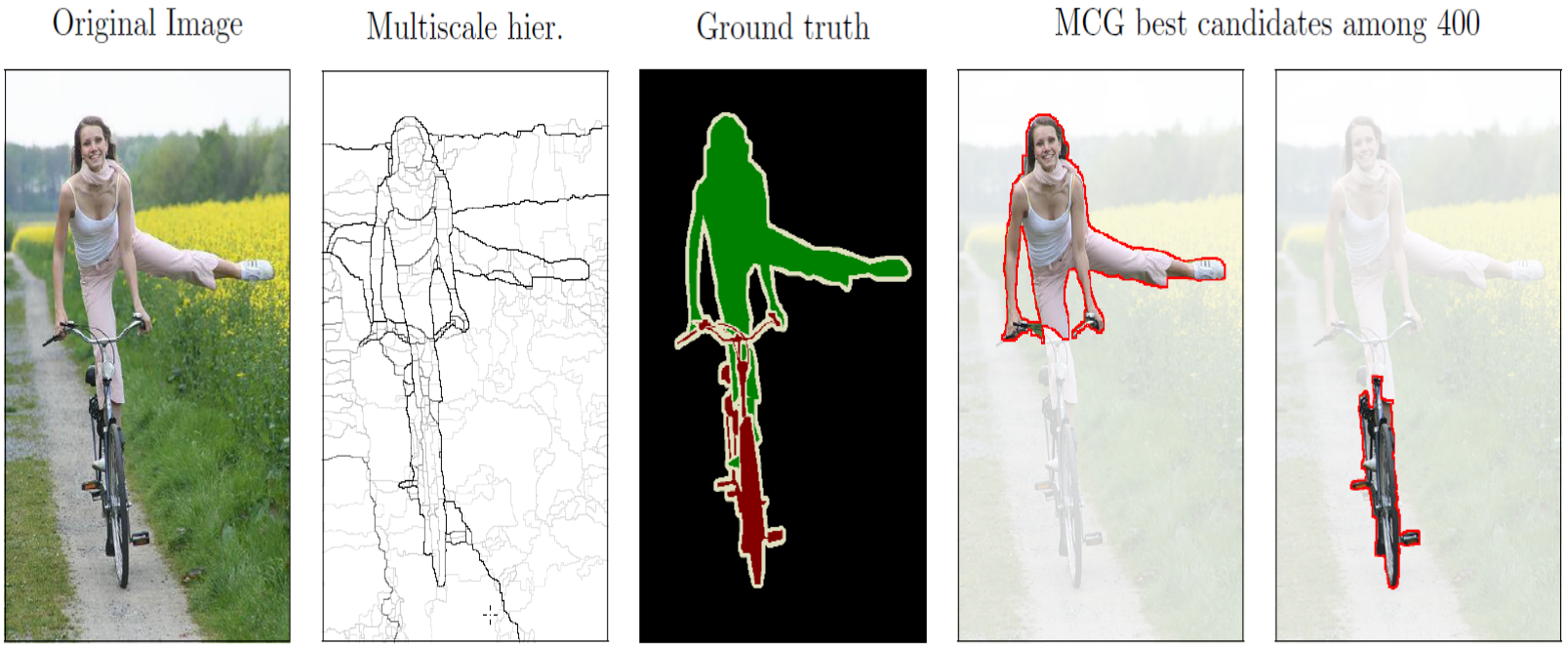


Each of the 6 directed arcs in this diagram is a useful direction of information flow

The Three R's of Vision



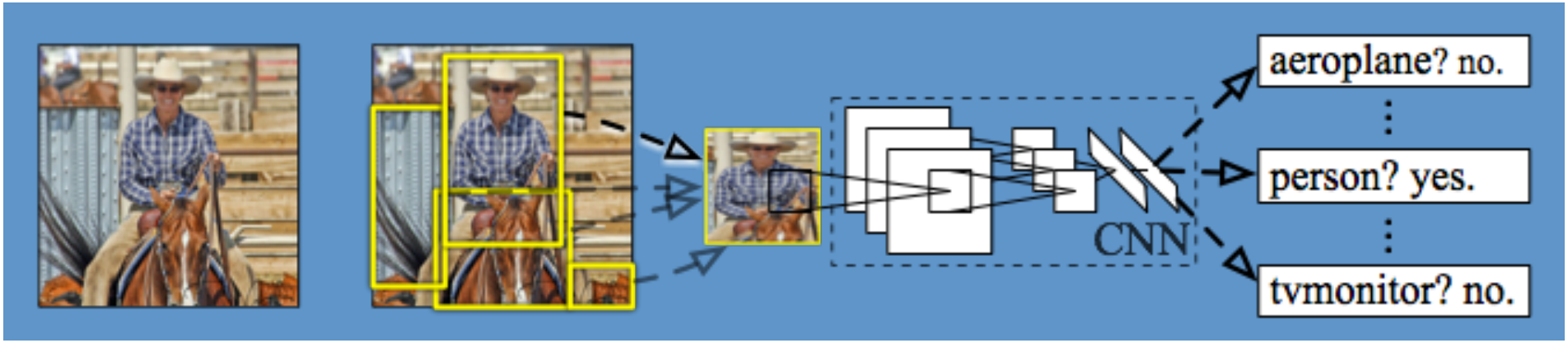
Bottom-up grouping as input to recognition



We produce superpixels of coherent color and texture first, then combine neighboring ones to generate object candidates

R-CNN: Regions with CNN features

Girshick, Donahue, Darrell & Malik (CVPR 2014)



Input image

Extract region proposals (~2k / image)

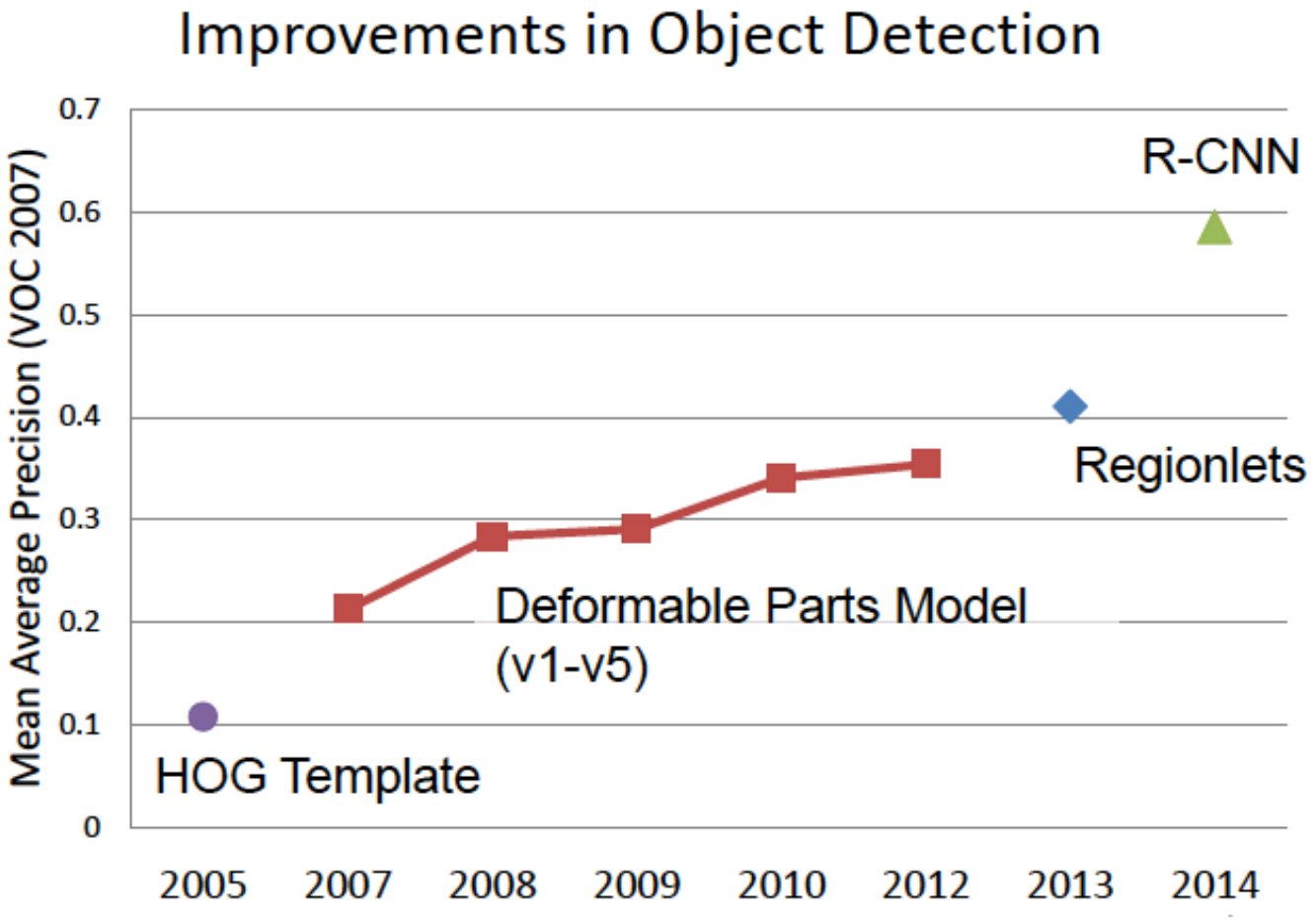
Compute CNN features

Classify regions (linear SVM)

CNN features are inspired by the architecture of the visual system

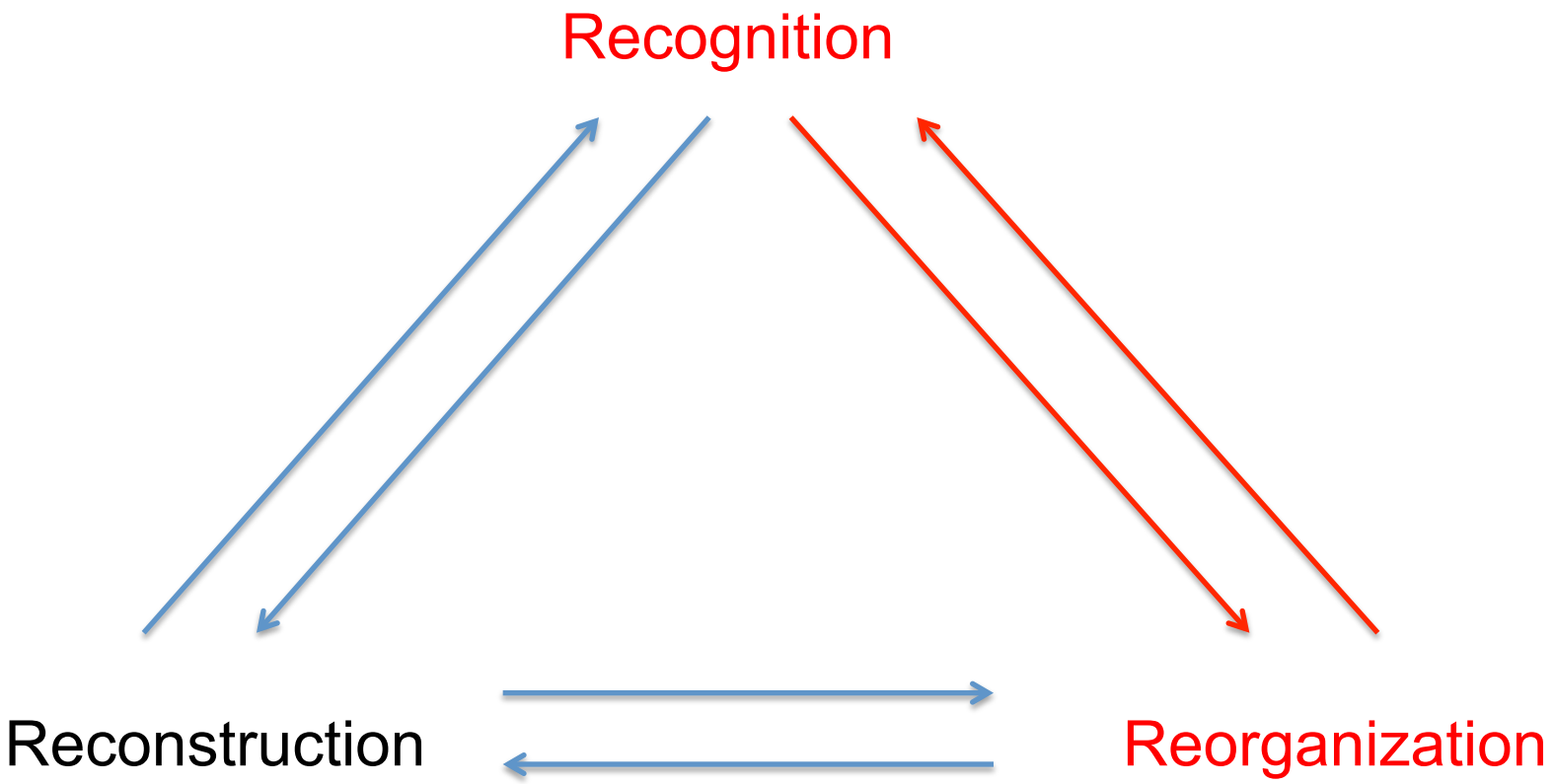
PASCAL Visual Object Challenge (Everingham et al)

State of the Art in Recognition



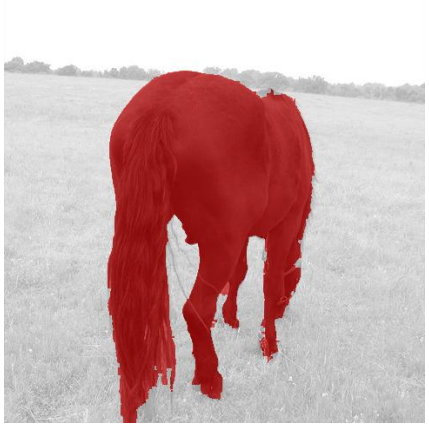
(Slide from D. Hoiem)

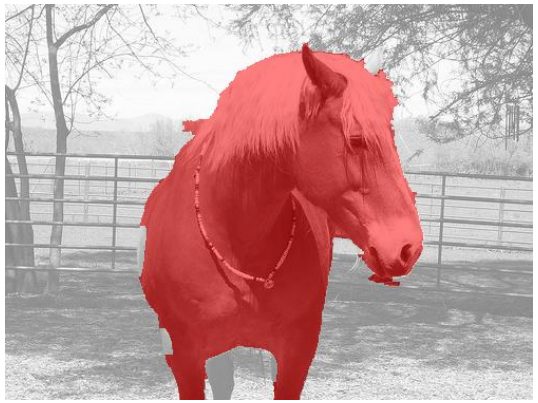
How about the other direction…



Recognition Helps Reorganization

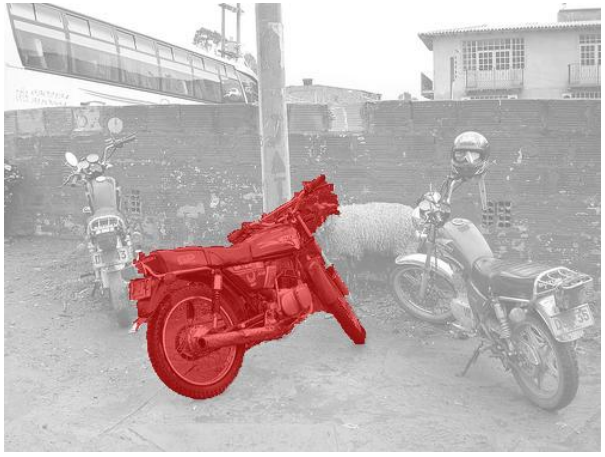
Results of Simultaneous Detection and Segmentation
Hariharan, Arbelaez. Girshick & Malik (2014)



We mark the pixels corresponding to an object instance, not just its bounding box.

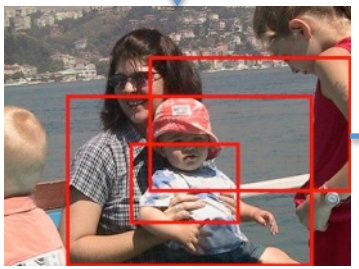
More results



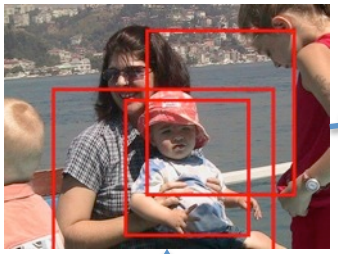
We train classifiers to predict top-down the pixels belonging to the object

Original detection

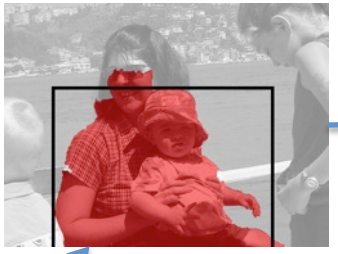
Search nearby



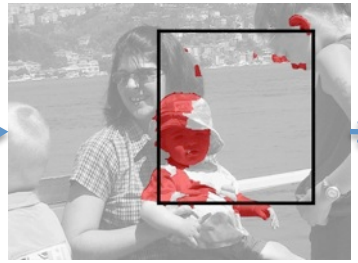
Regress boxes



Segment



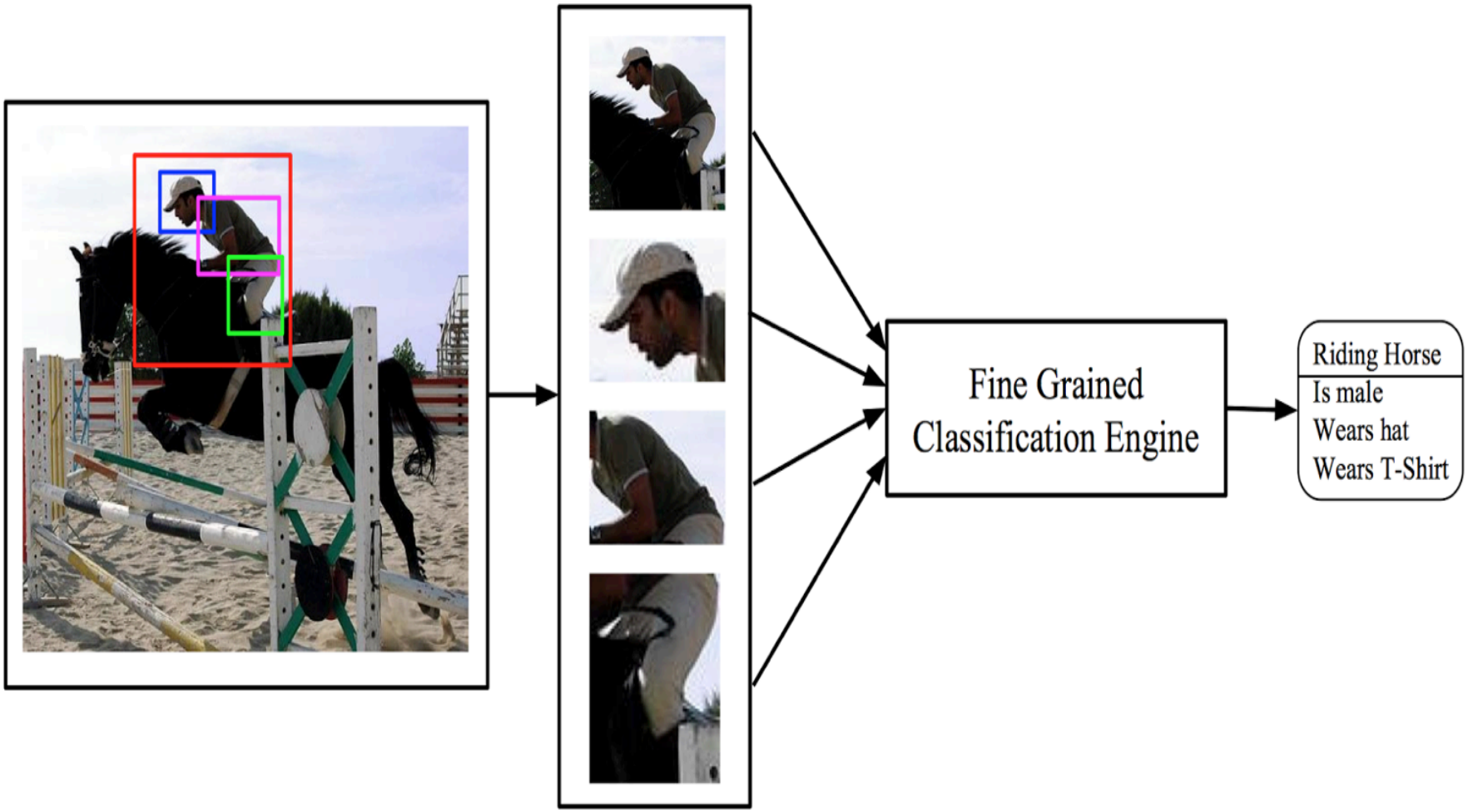
Score



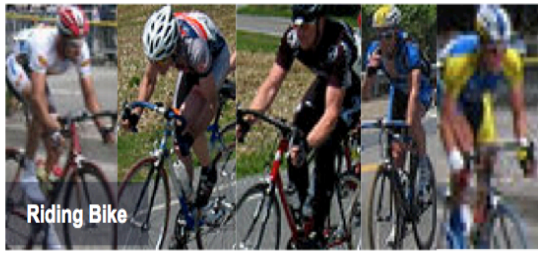
Score

Score

Actions and Attributes from Wholes and Parts
G. Gkioxari, R. Girshick & J. Malik




Jumping

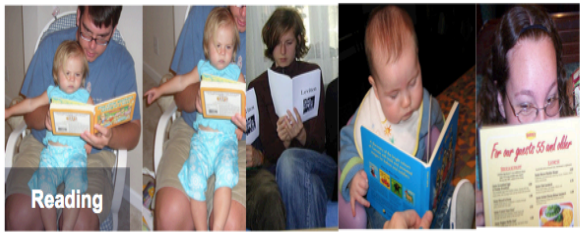

Riding Bike


Phoning


Riding Horse


Playing Instrument


Running


Reading


Using Computer


Taking Photo


Walking


Is Male


Has Long Hair


Has Glasses


Has Hat


Has T-Shirt


Has Long Sleeves


Has Shorts


Has Jeans


Has Long Pants

Finding Human Body Joints

Jonathan Tompson, Ross Goroshin, Arjun Jain, Yann LeCun, Christopher Bregler
New York University
tompson/goroshin/ajain/lecun/bregler@cims.nyu.edu

Viewpoint Prediction for Objects

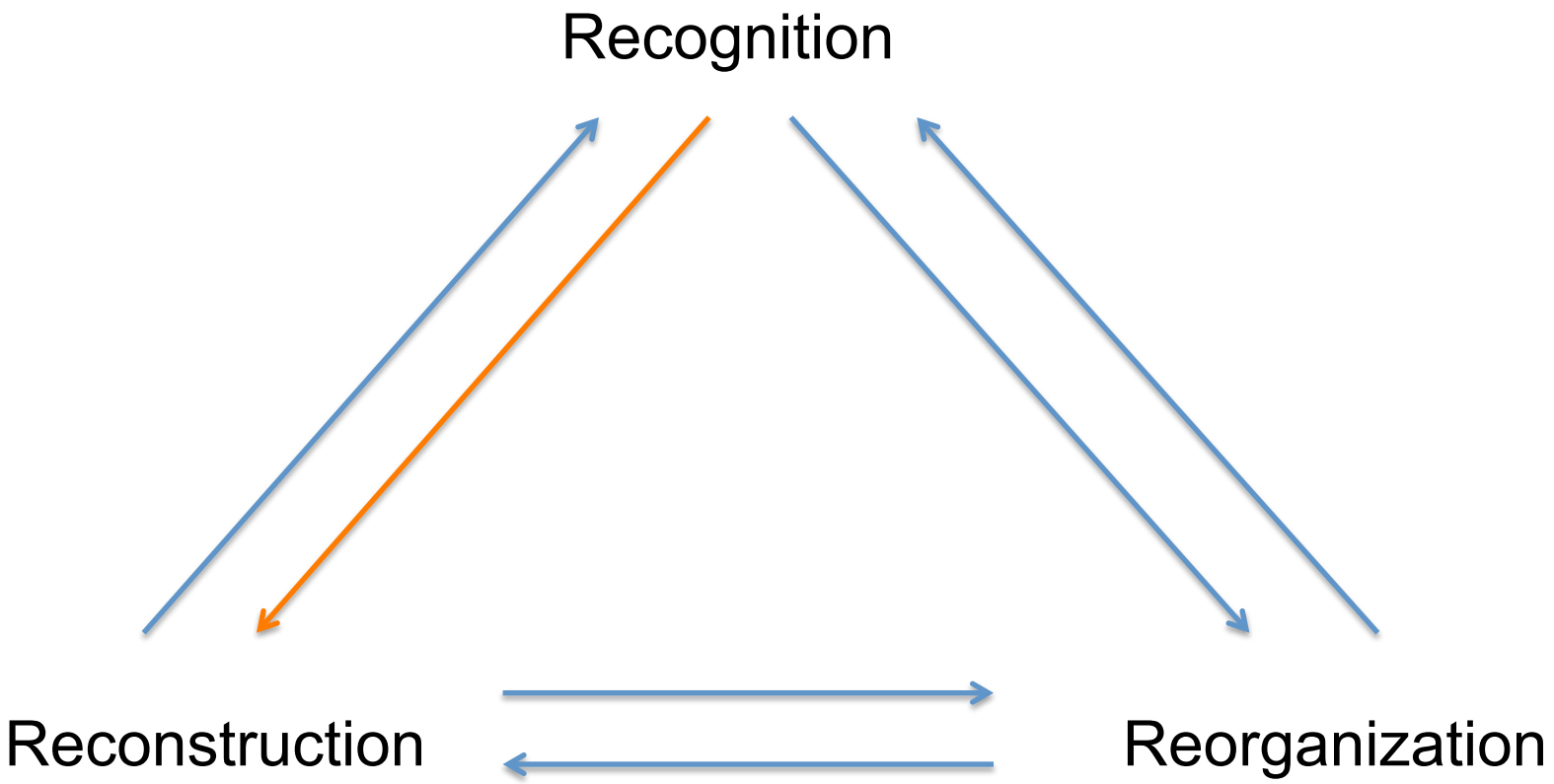
Tulsiani & Malik (2014)

The columns show 15th, 30th, 45th, 60th, 75th and 90th percentile instances respectively in terms of the error.

Keypoint Prediction for Objects

Visualization of keypoints predicted in the detection setting. We sort the keypoints detections by their prediction score and visualize every 15th detection for 'Nosetip' of aeroplanes, 'Left Headlight' of cars and 'Crankcentre' of bicycles.

The Three R's of Vision



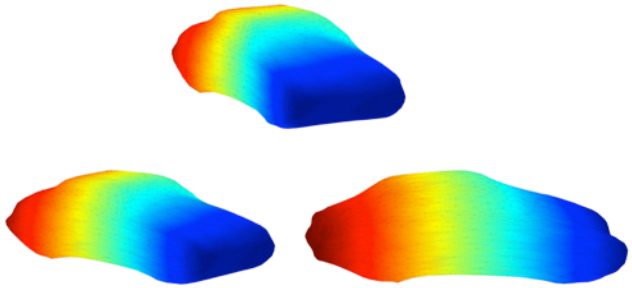
We have explored category-specific 3D reconstruction.

Category Specific Object Reconstruction
Kar, Tulisiani, Carreira & Malik

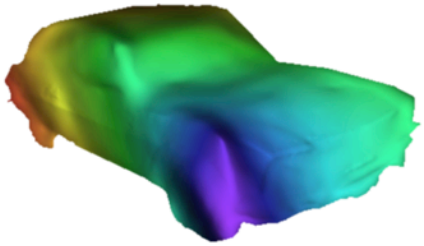
Input Image

Instance Segmentation

Viewpoint estimation

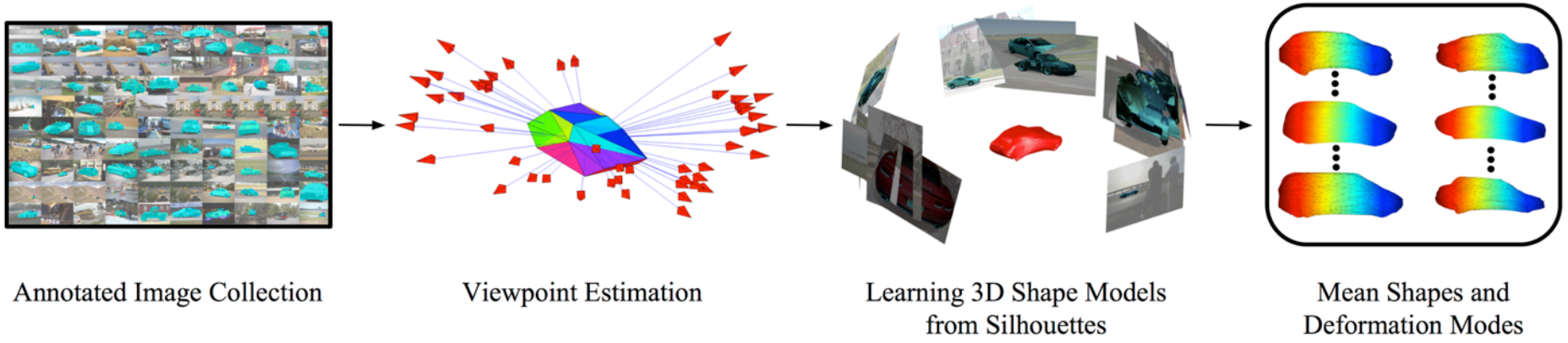


Full 3D Reconstruction



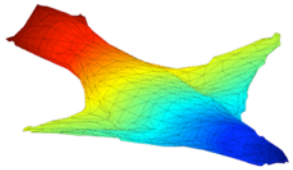
High Frequency 2.5D Reconstruction

Deformable 3D Model Learning

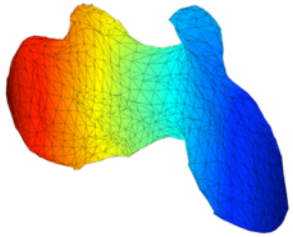


- Viewpoint estimation – NRSfM on keypoint correspondences
- Idea - *Deform* a mesh to satisfy *silhouettes* from different viewpoints
- Energies - Consistency, Coverage, **S**moothness, **K**eypoint
- Intra-class Variation - *Linear* deformation modes

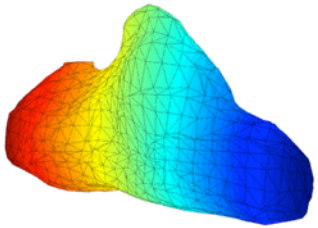
Basis Shape Models



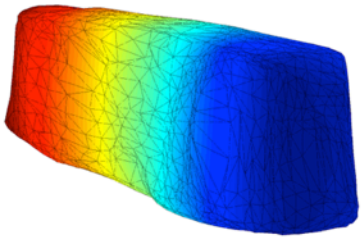
aeroplane



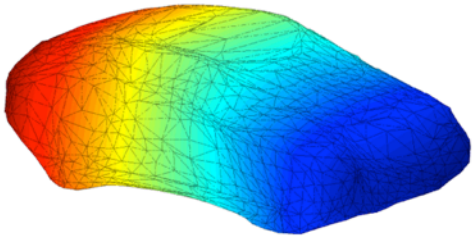
bicycle



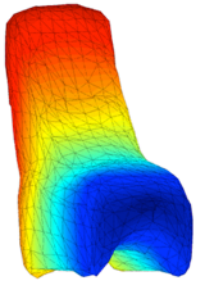
boat



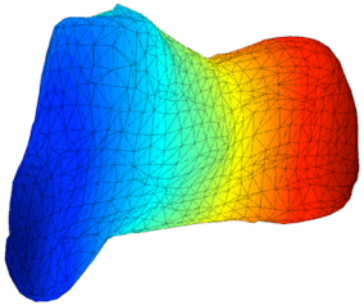
bus



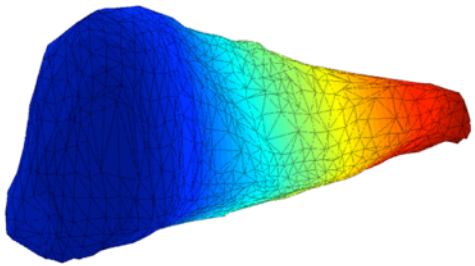
car



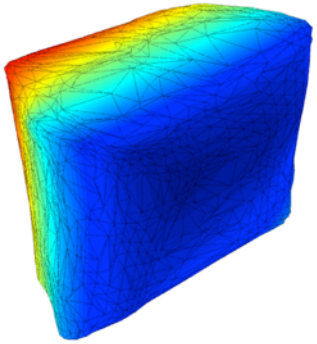
chair



motorbike

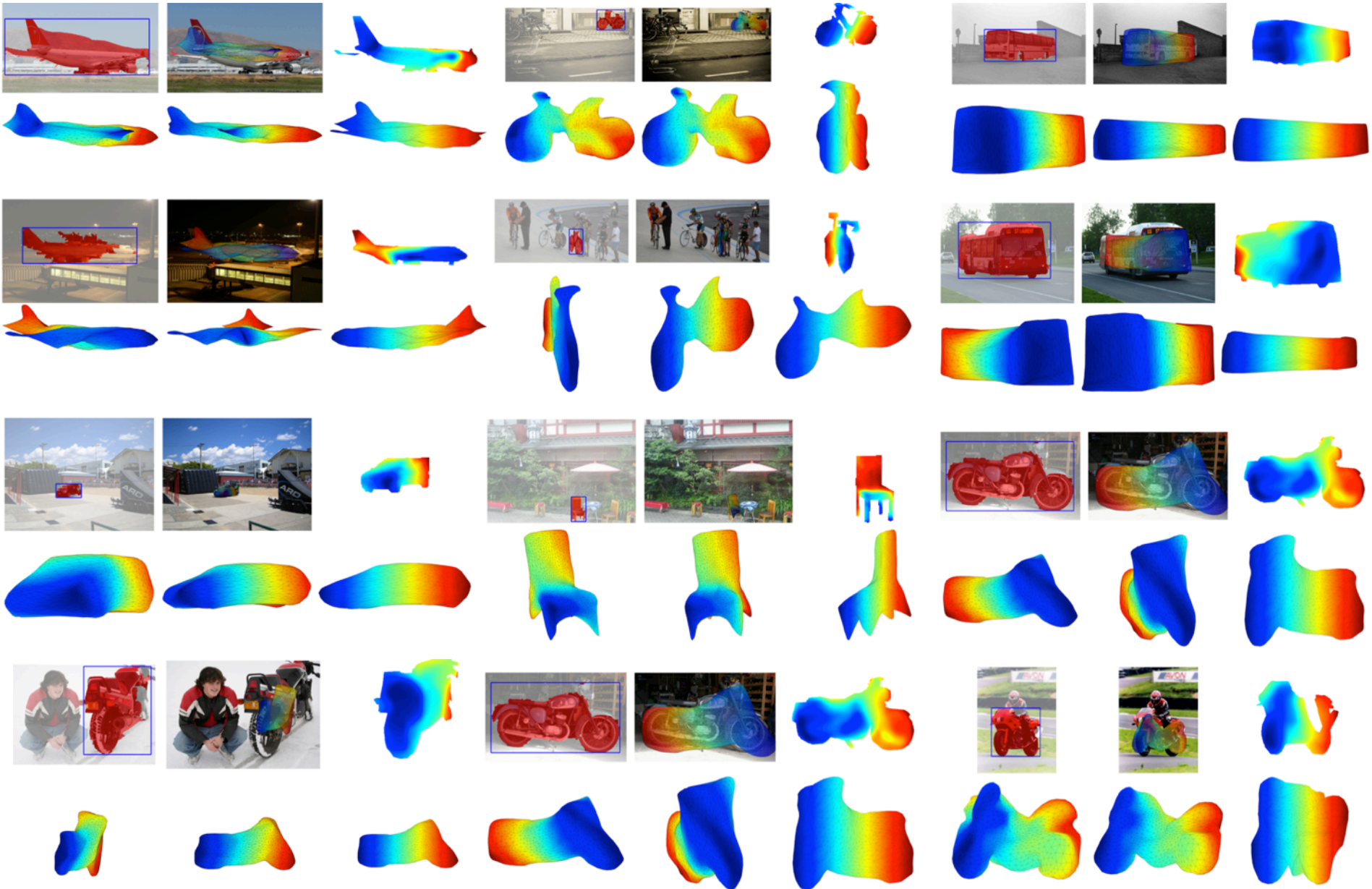


train

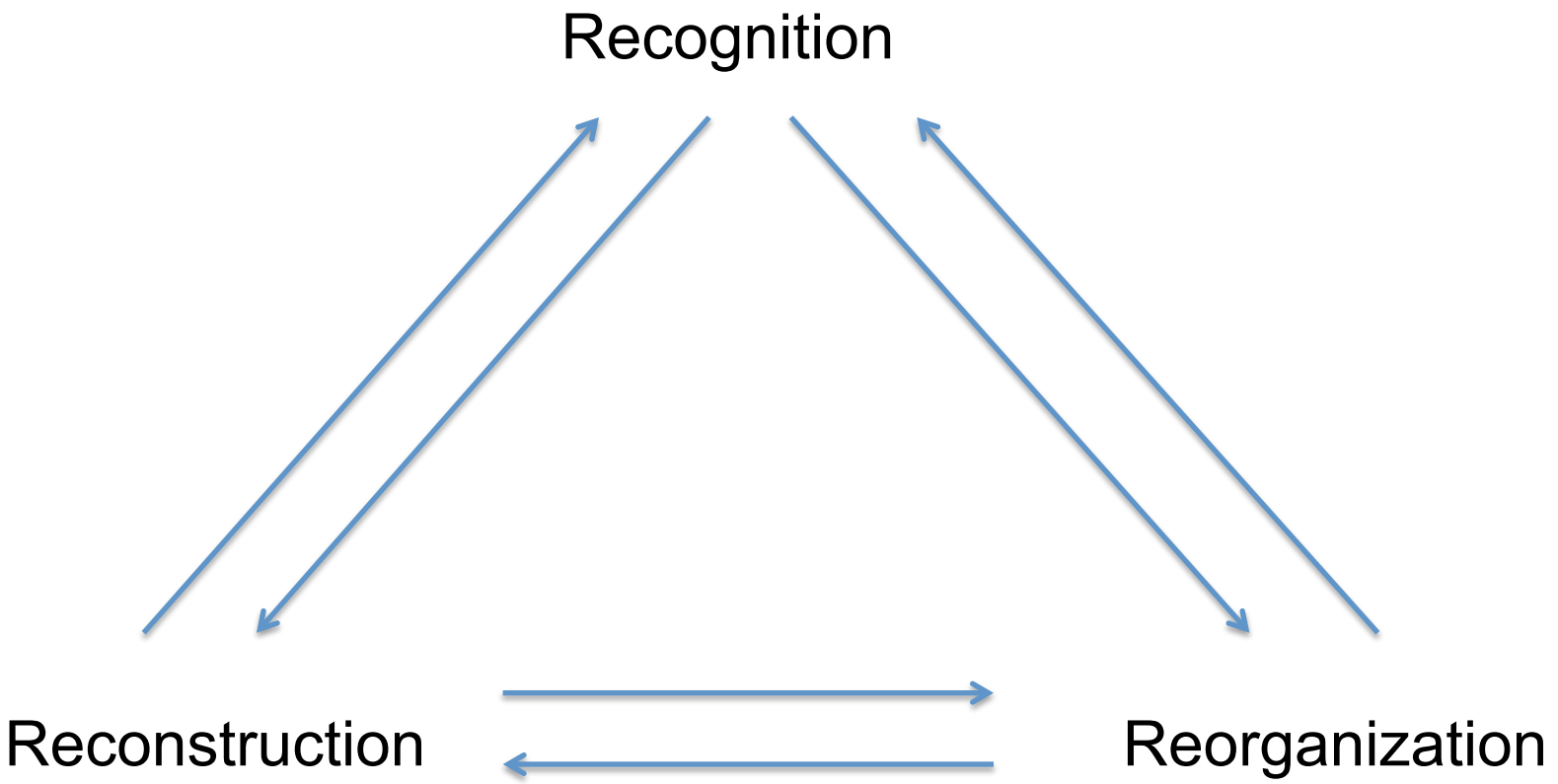


tvmonitor

Results



The Three R's of Vision



These ideas apply equally well in a video setting

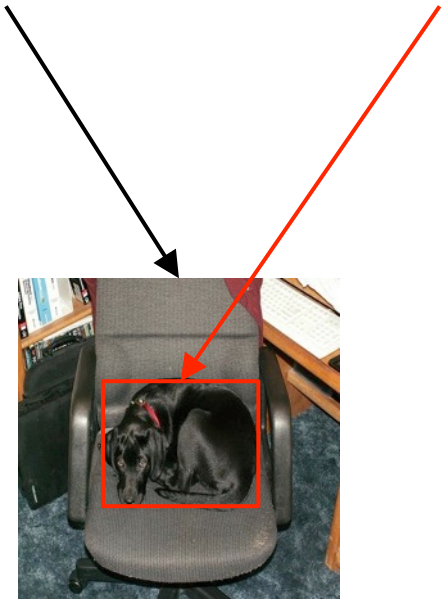
Images

Image classification

“Is there a dog in the image?”

Object detection

“Is there a dog and where is it in the image?”



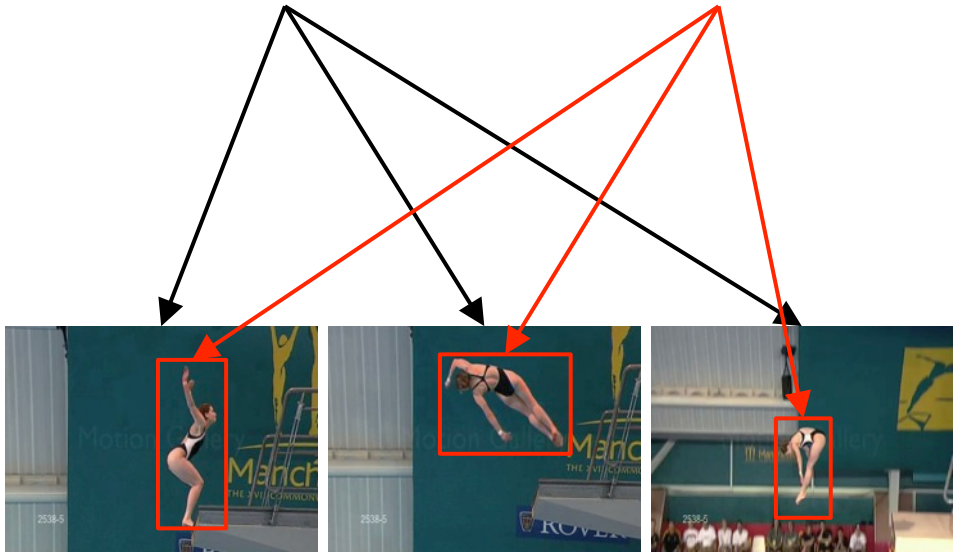
Video

Action classification

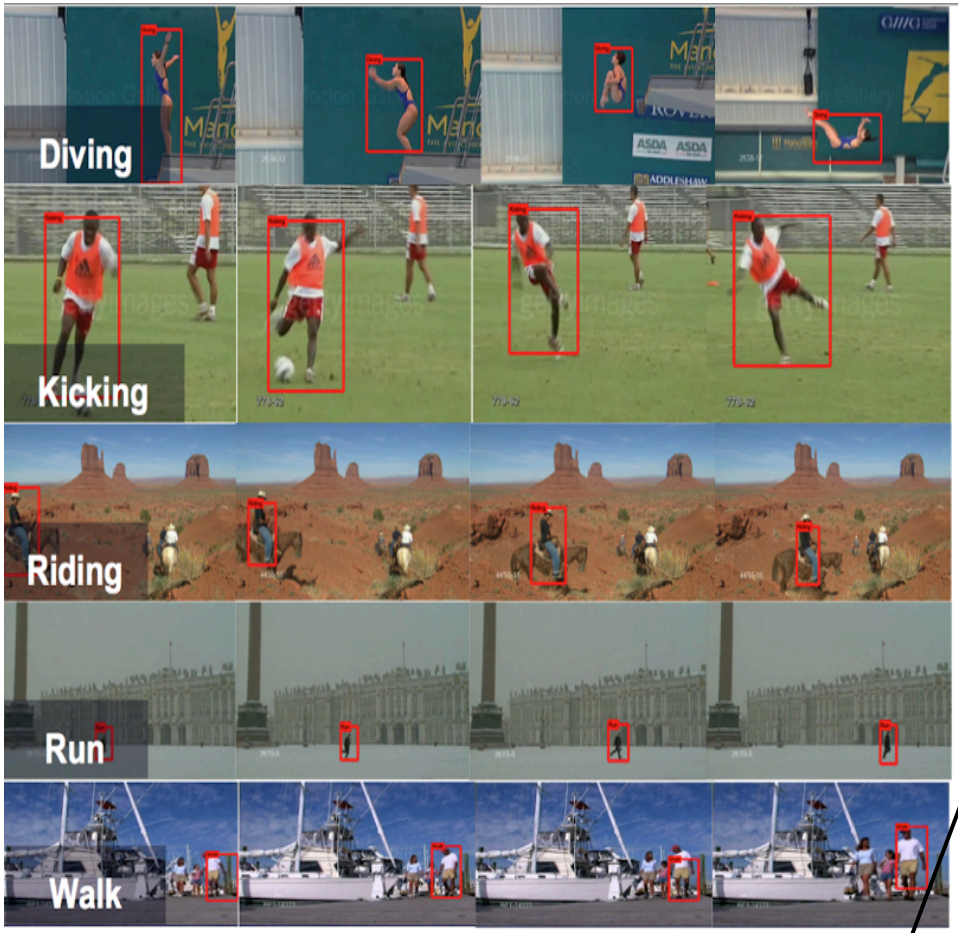
“Is there a person diving in the video?”

Action detection

“Is there a person diving and where is it in the video?”



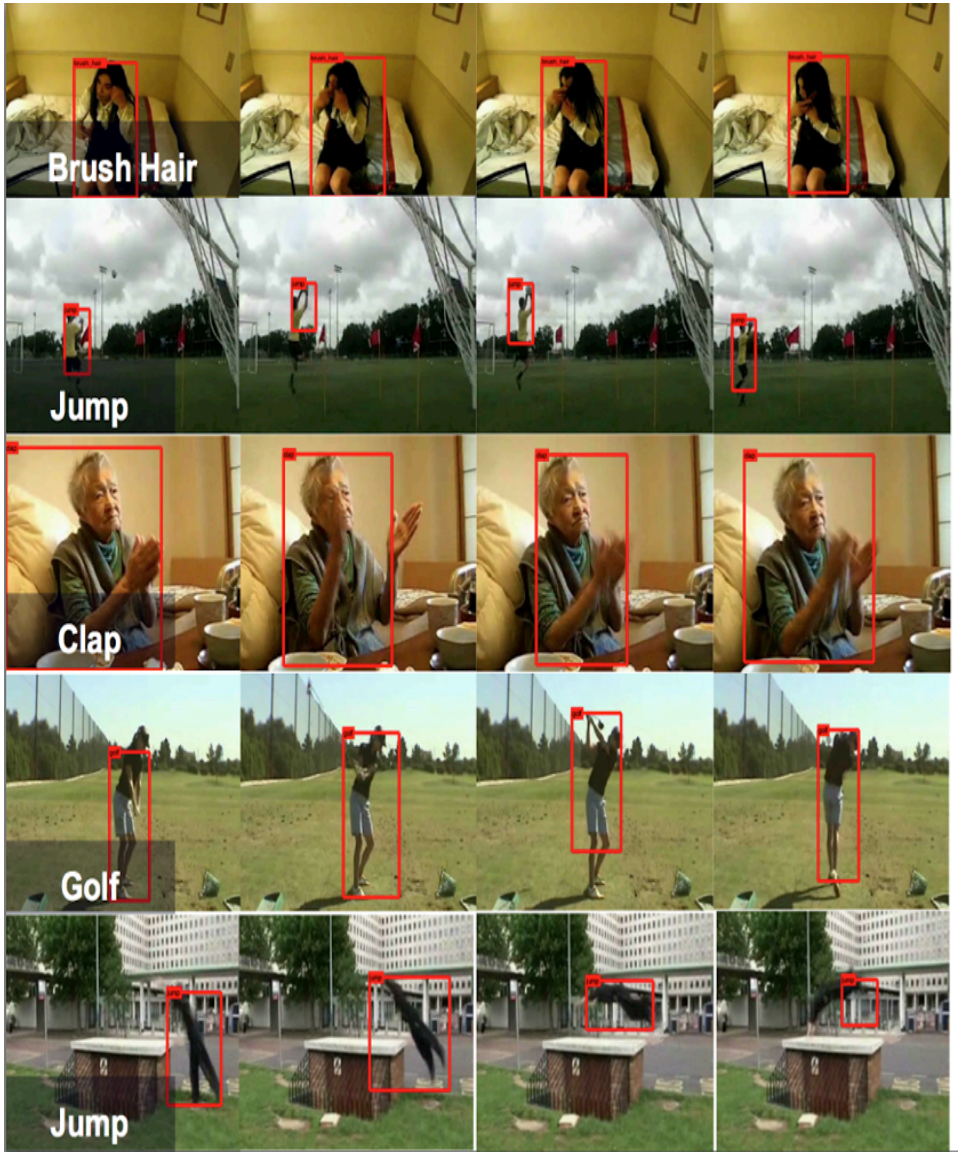
Results on UCF Sports (Gkioxari & Malik, 2014)

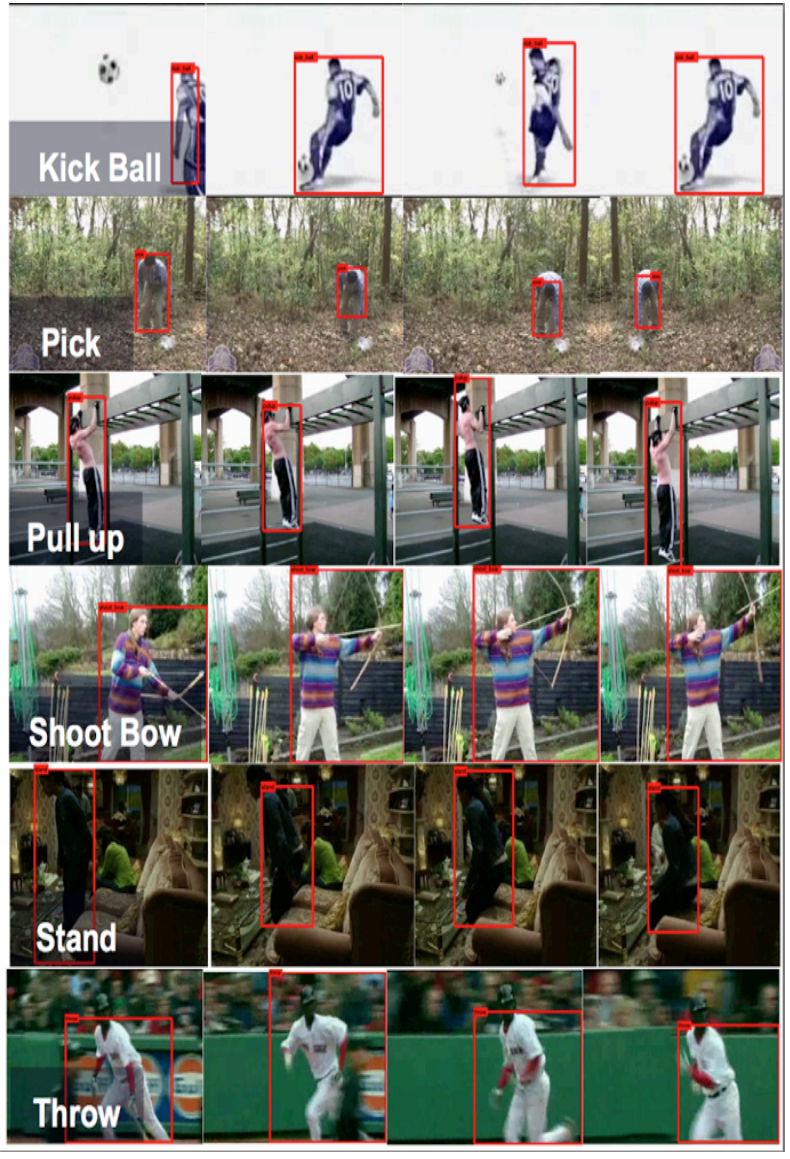


Tracking error

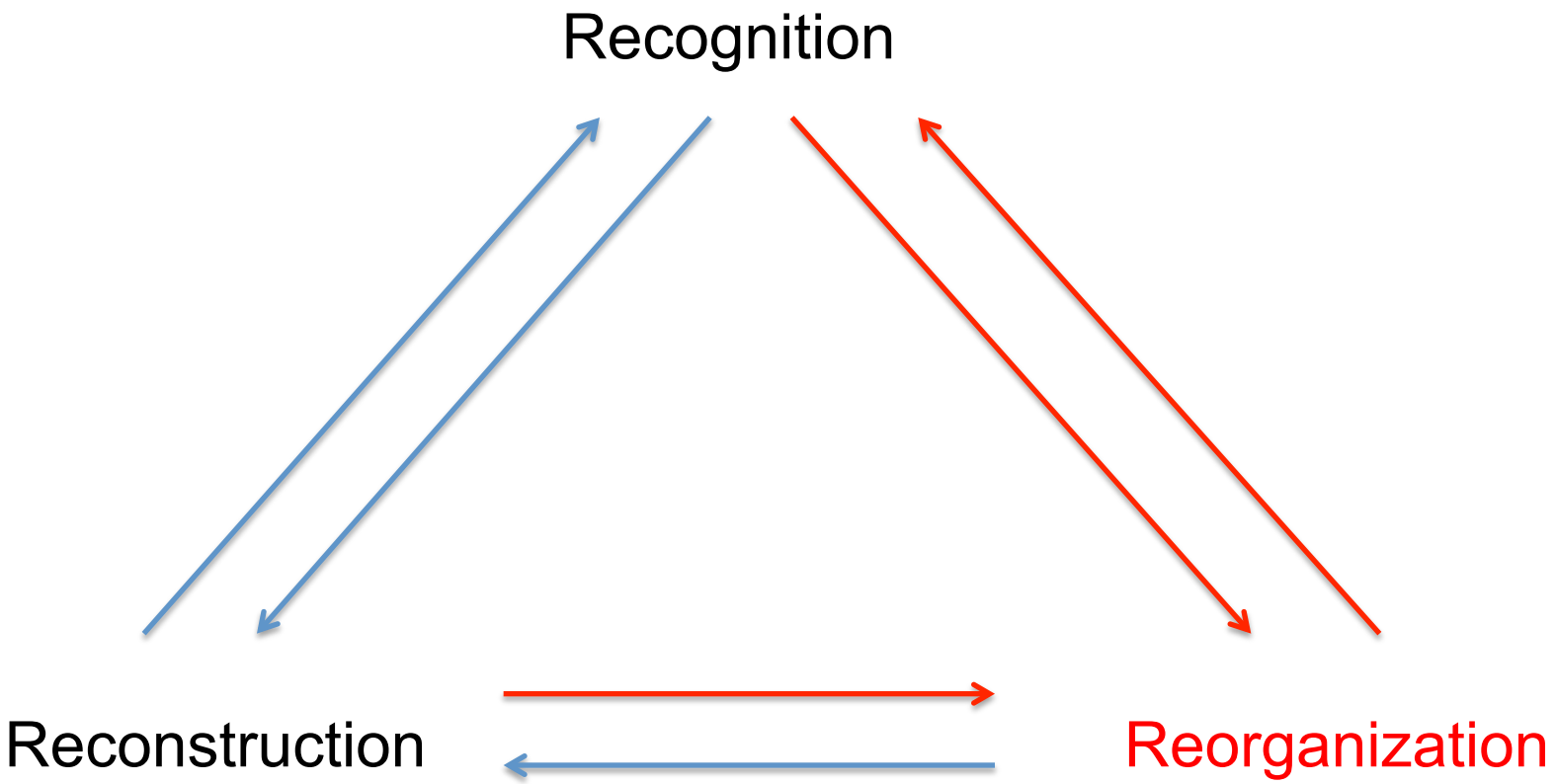
Action prediction error

Results on J-HMDB



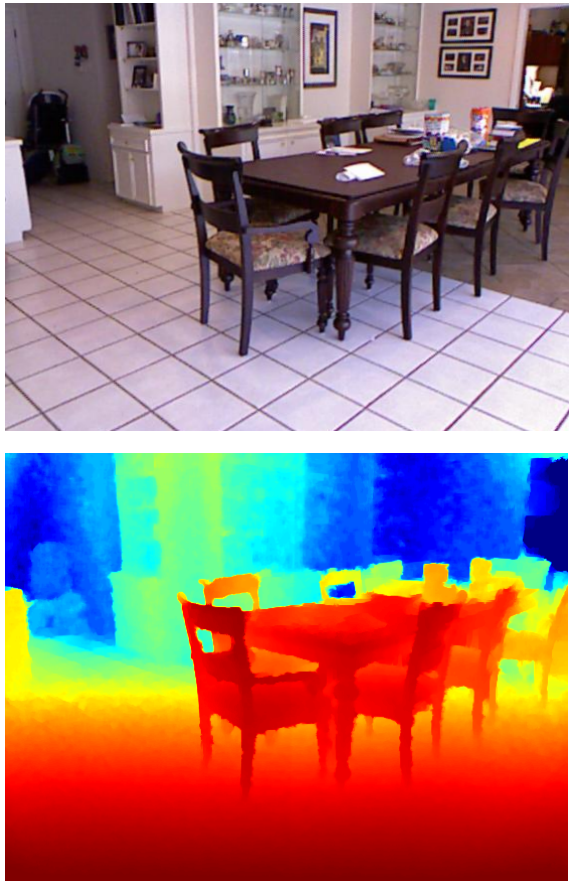


The Three R's of Vision

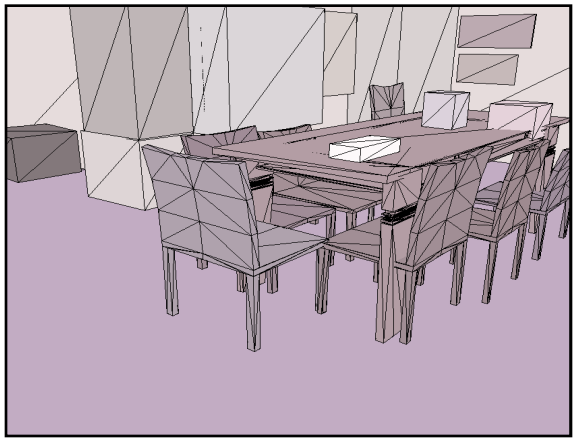


Scene Understanding using RGB-D data

Gupta, Girshick, Arbelaez, Malik (ECCV 2014)



Color and Depth Image Pair



Complete 3D Understanding

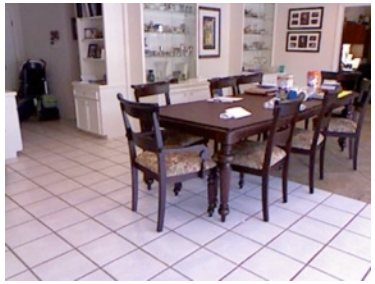
Overview

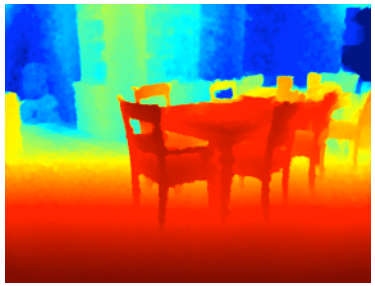
Input

Re-organization

Recognition

Detailed 3D Understanding

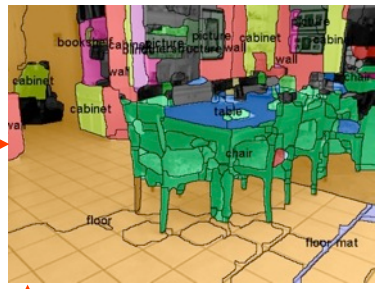




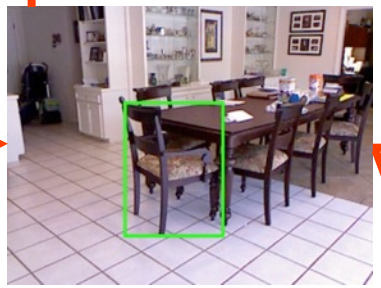
Color and Depth Image Pair

Contour Detection

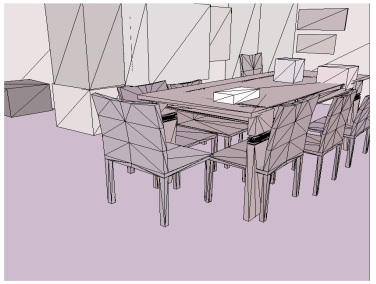
Region Proposal Generation

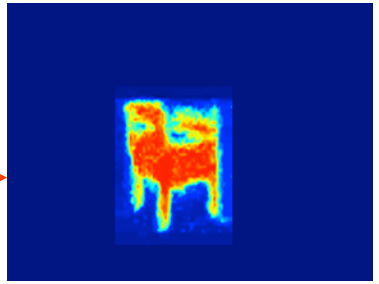


Semantic Segm.

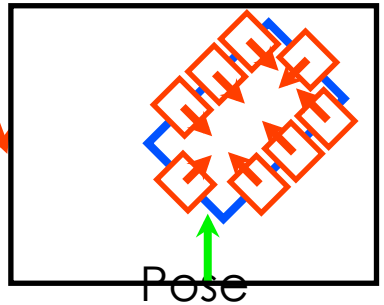


Object Detection



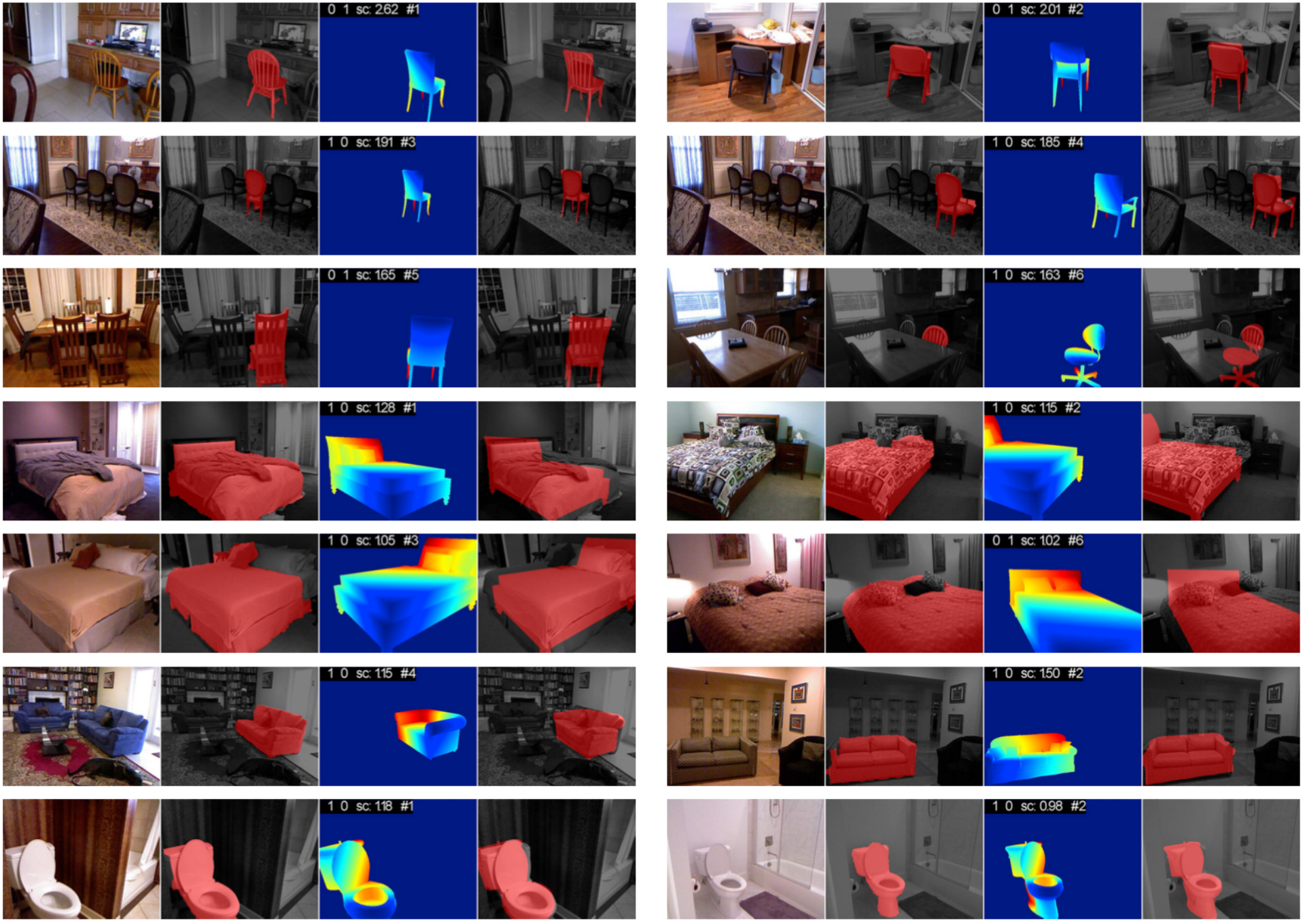


Instance Segm.



Pose Estimation

Instance Segmentation


0 1 sc: 2.62 #1
0 1 sc: 2.01 #2
1 0 sc: 1.91 #3
1 0 sc: 1.85 #4
0 1 sc: 1.65 #5
1 0 sc: 1.63 #6
1 0 sc: 1.28 #1
1 0 sc: 1.15 #2
1 0 sc: 1.05 #3
0 1 sc: 1.02 #6
1 0 sc: 1.15 #4
1 0 sc: 1.50 #2
1 0 sc: 1.18 #1
1 0 sc: 0.98 #2

I hope you enjoy the course!