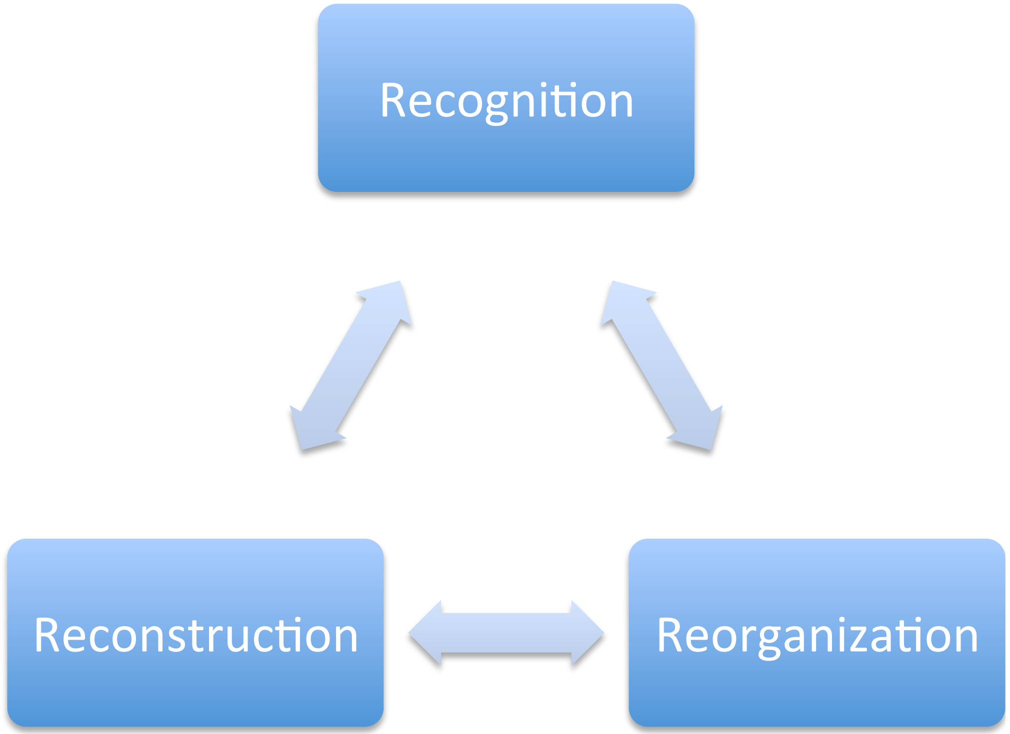
The Three R's of Vision

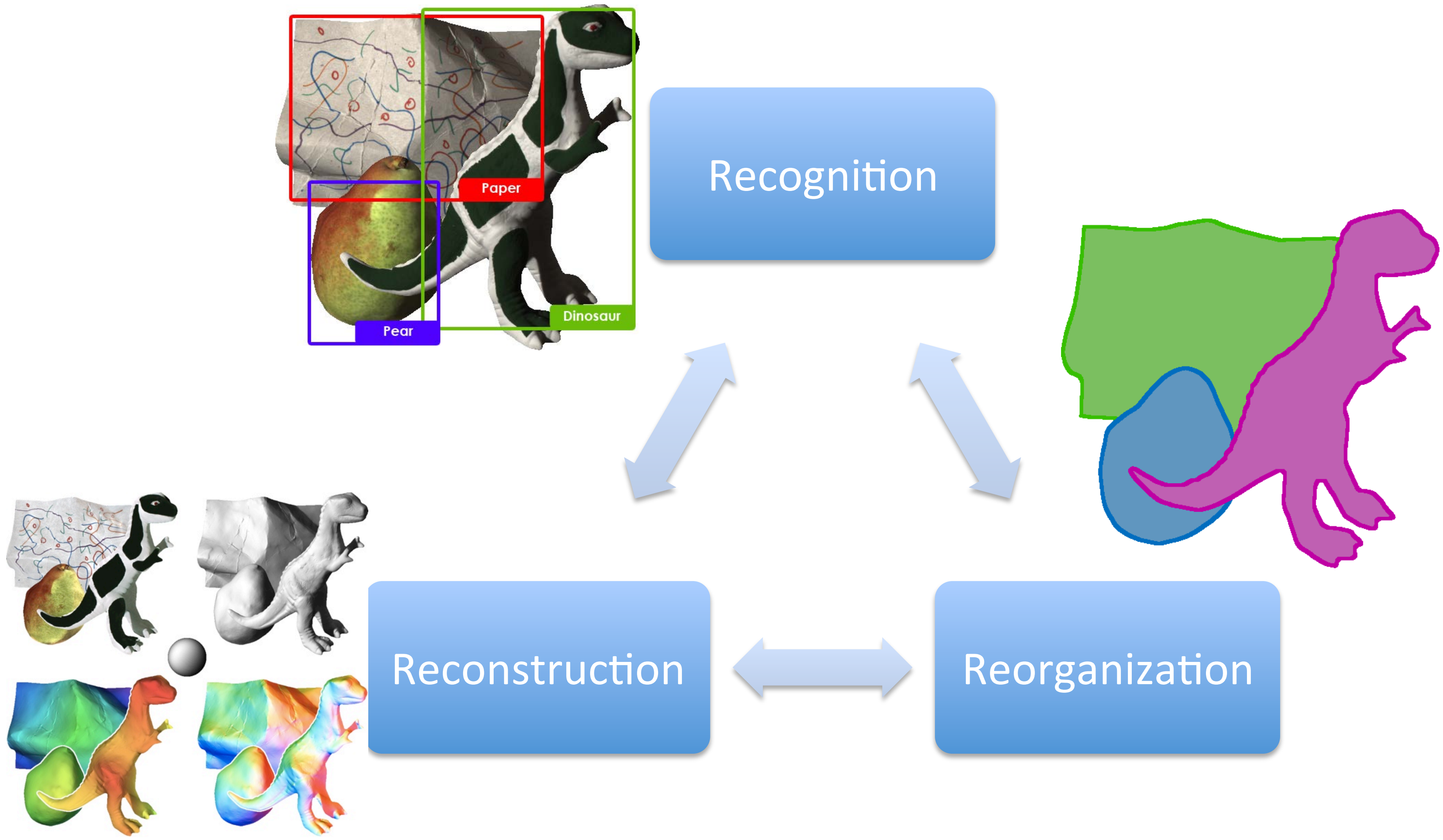


Jitendra Malik
UC Berkeley

Big Data is increasingly Visual Data

- Flickr has 5 billion images
- Facebook has 200 billion images
- YouTube has 72 hours of video uploaded every minute
- 30% of the internet traffic in the US is due to video
- Much of the explosion of data in science, engineering, and medicine is in the form of images.

Recognition, Reconstruction & Reorganization



Fifty years of computer vision 1963-2013

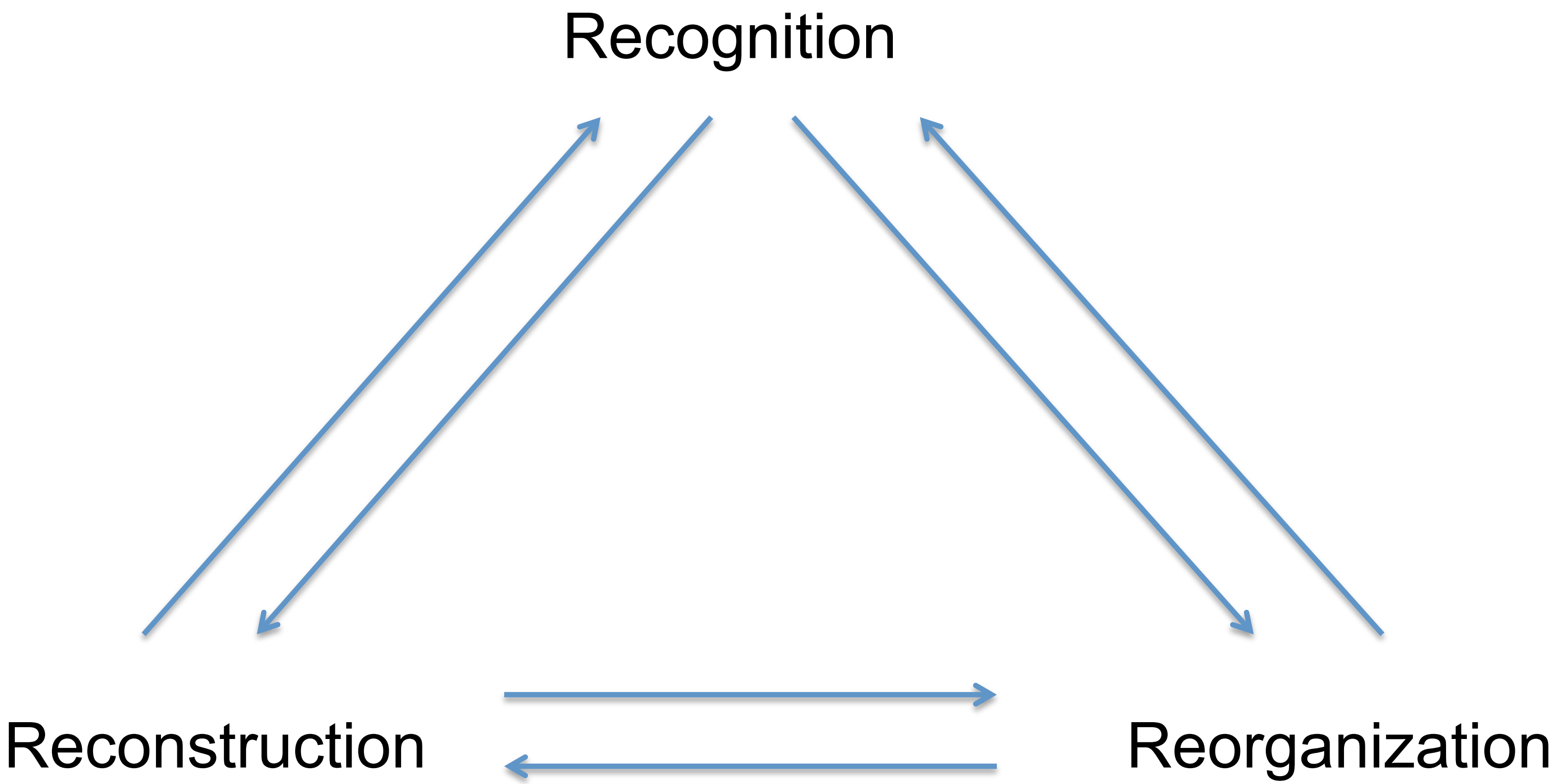
- 1960s: Beginnings in artificial intelligence, image processing and pattern recognition
- 1970s: Foundational work on image formation: Horn, Koenderink, Longuet-Higgins …
- 1980s: Vision as applied mathematics: geometry, multi-scale analysis, probabilistic modeling, control theory, optimization
- 1990s: Geometric analysis largely completed, vision meets graphics, statistical learning approaches resurface
- 2000s: Significant advances in visual recognition, range of practical applications

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.

The Three R's of Vision



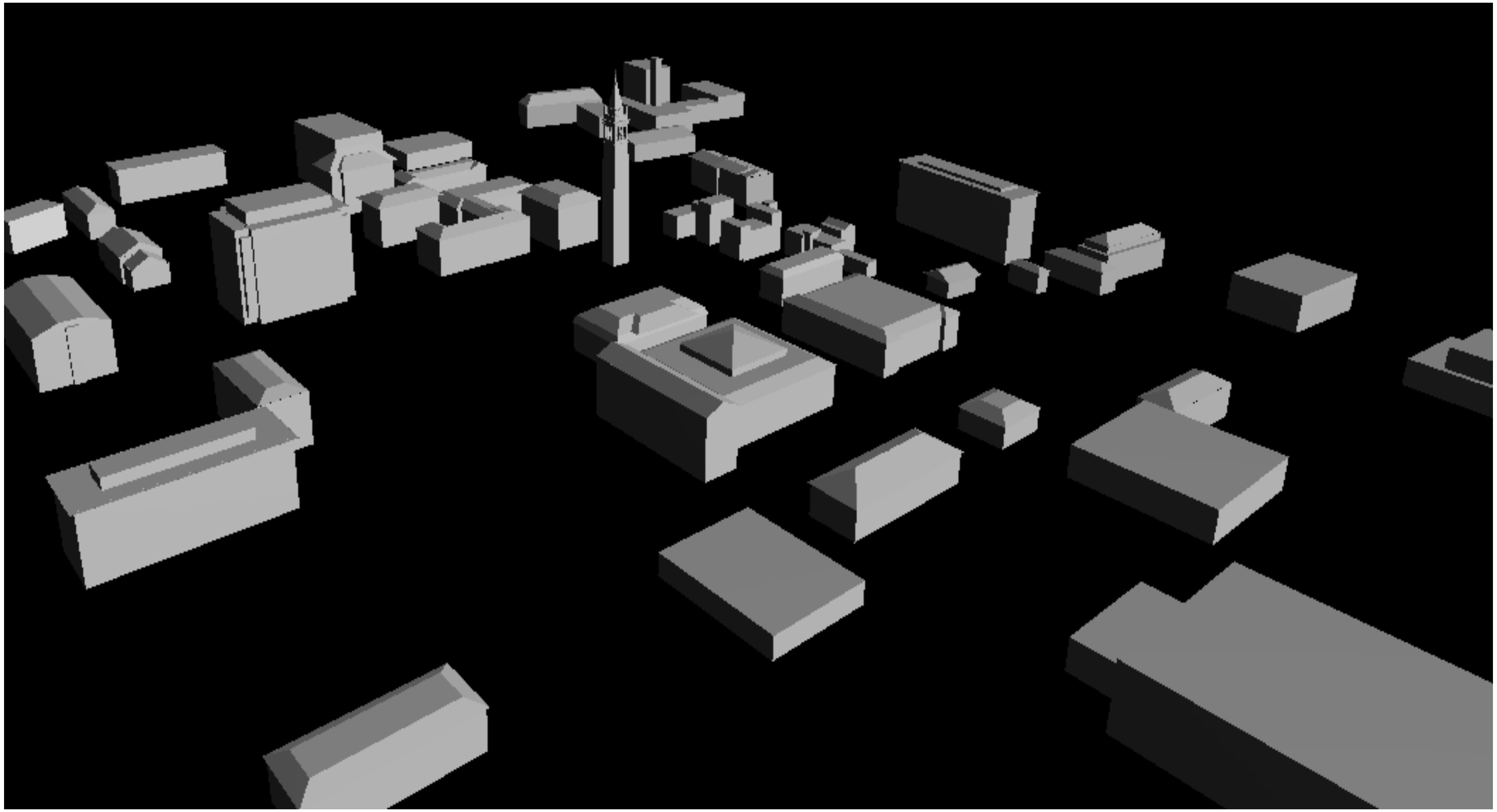
Review

- Reconstruction
  - Feature matching + multiple view geometry has led to city scale point cloud reconstructions
- Recognition
  - 2D problems such as handwriting recognition, face detection successfully fielded in applications.
  - Partial progress on 3d object category recognition
- Reorganization
  - Progress on bottom-up segmentation hitting diminishing returns
  - Semantic segmentation is the key problem now

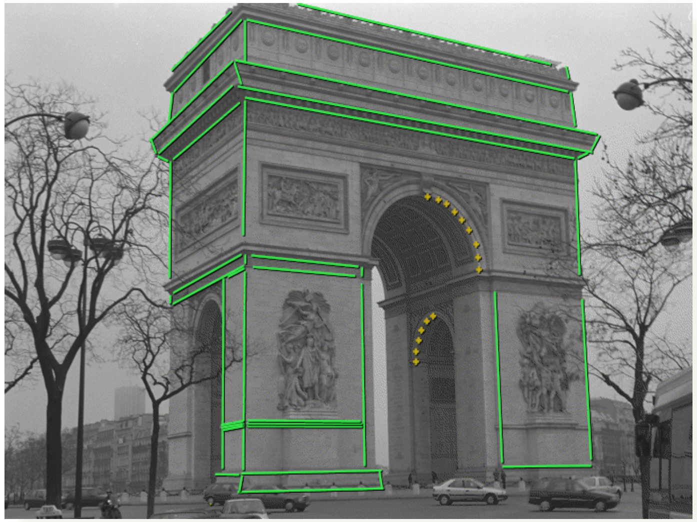
Image-based Modeling

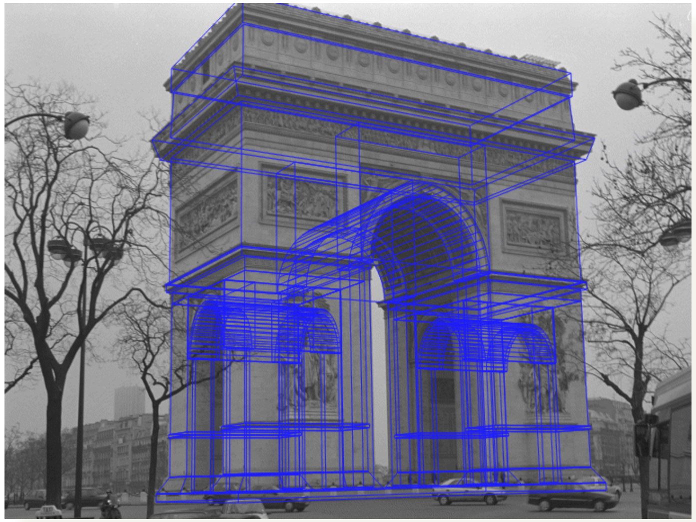
- Façade (1996) Debevec, Taylor & Malik
  - Acquire photographs
  - Recover geometry (explicit or implicit)
  - Texture map

Campus Model of UC Berkeley



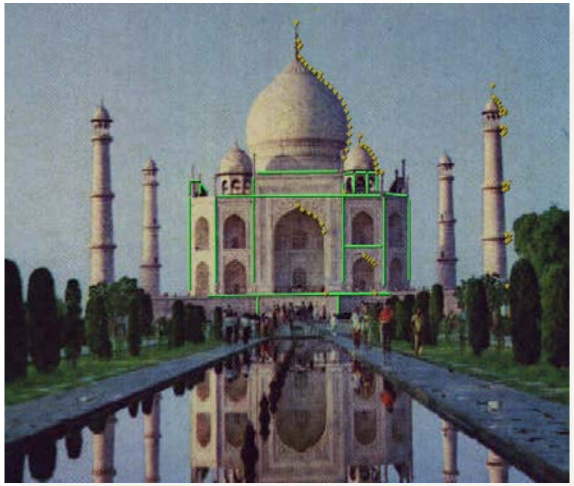
Campanile + 40 Buildings (Debevec et al, 1997)

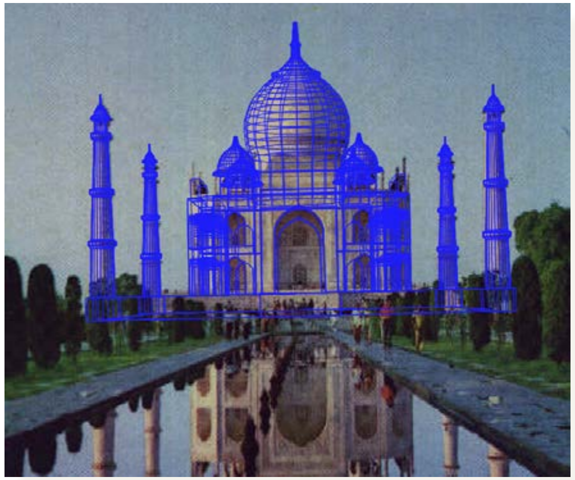




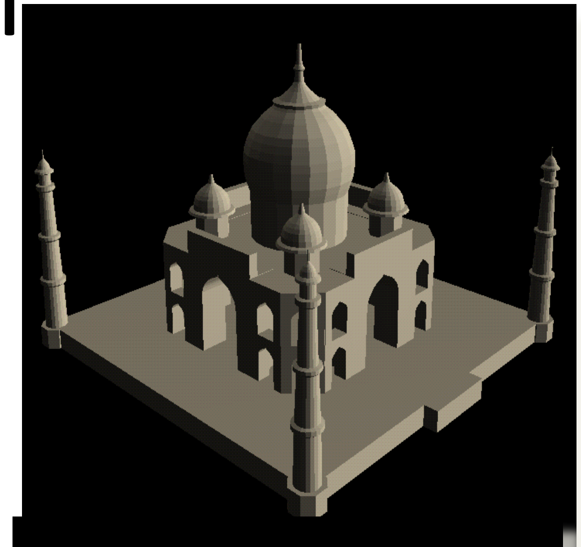
Arc de Triomphe

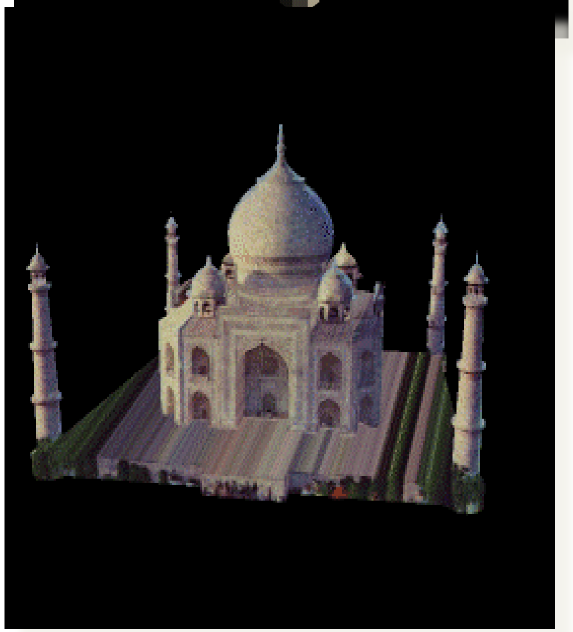
The Taj Mahal





Taj Mahal modeled from one photograph by G. Borshukov





State of the Art in Reconstruction

- Multiple photographs

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

Agarwal et al (2010)

Frahm et al, (2010)

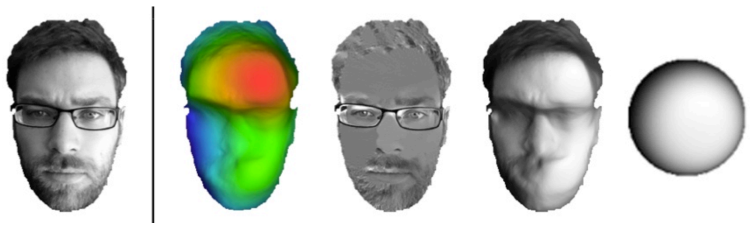
- Range Sensors

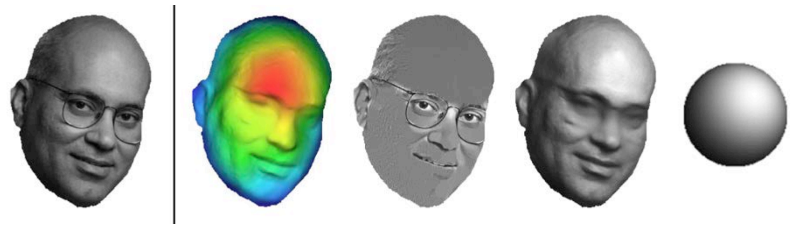
Kinect (PrimeSense)

Velodyne Lidar

Semantic Segmentation is needed to make this more useful…

Shape, Albedo, and Illumination from Shading



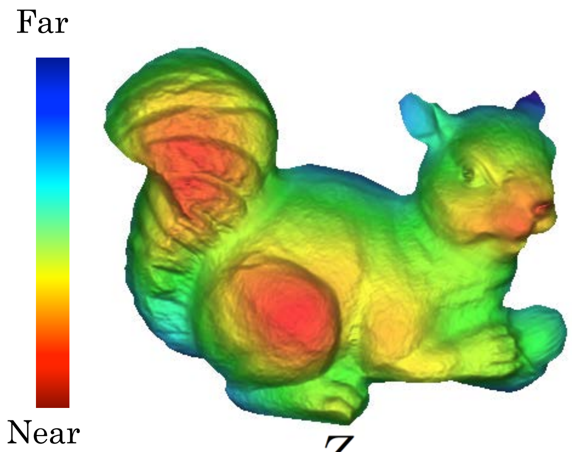


Jonathan Barron

Jitendra Malik

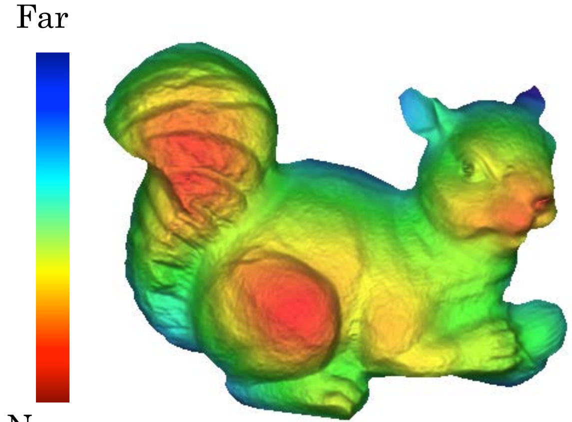
UC Berkeley

Forward Optics



Z
shape / depth

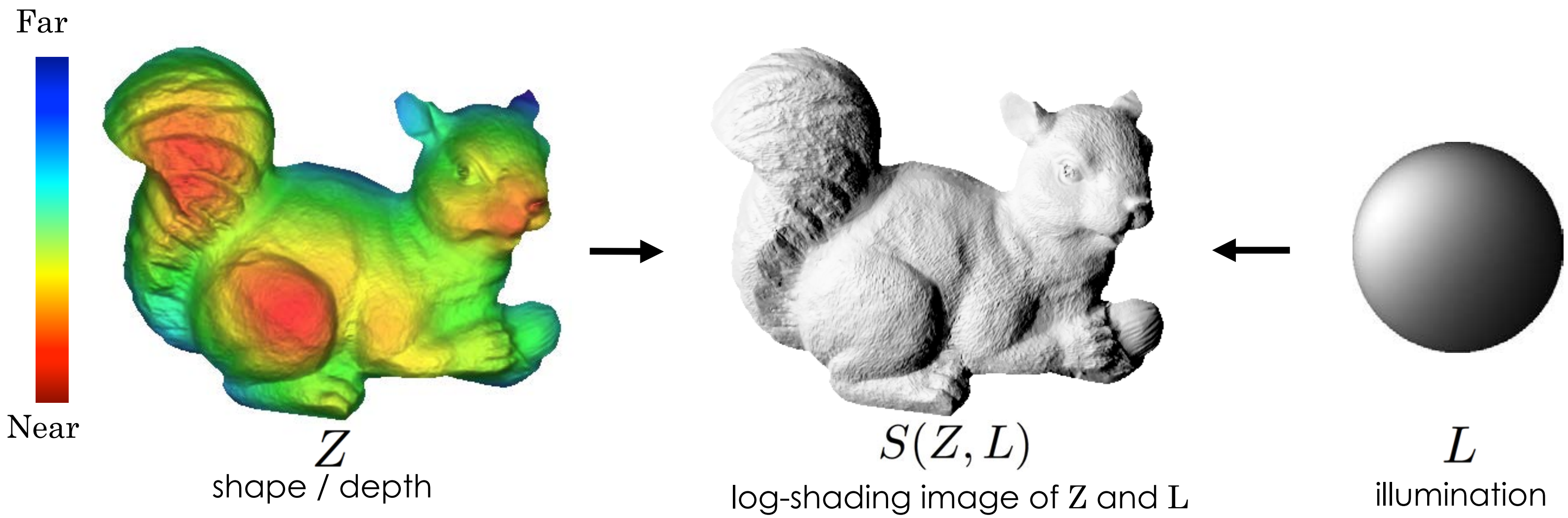
Forward Optics



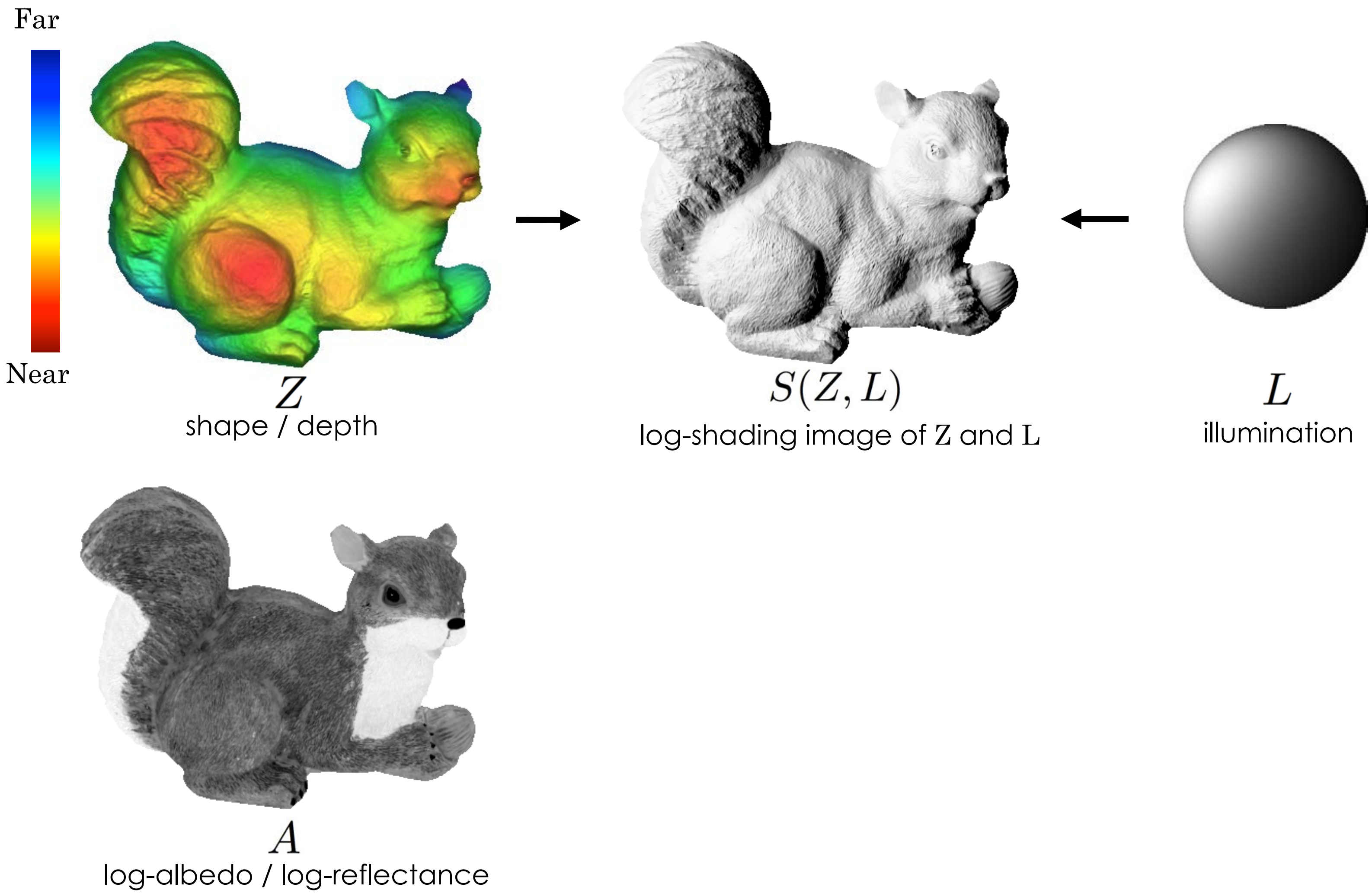
Z
shape / depth

L
illumination

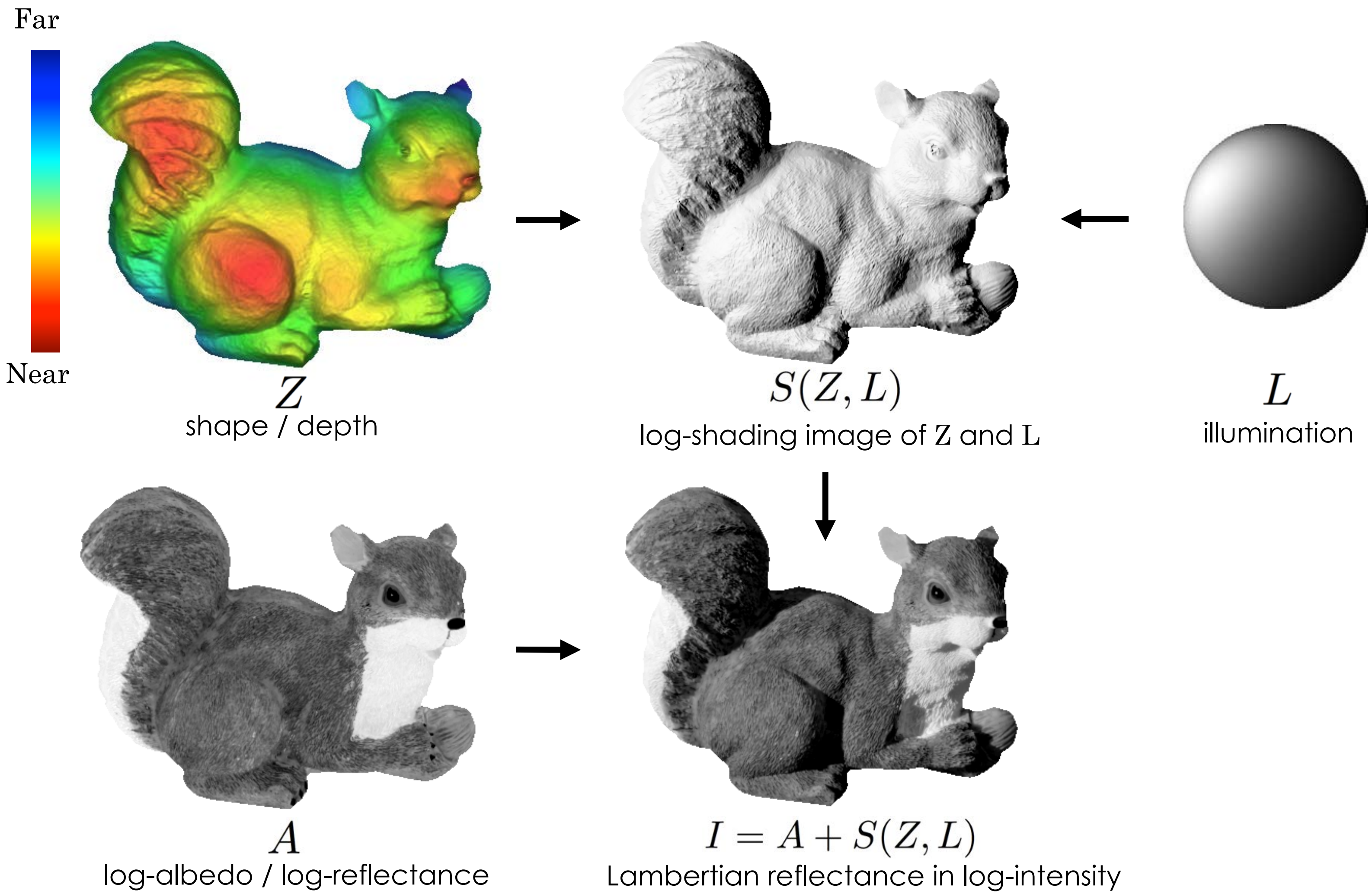
Forward Optics



Forward Optics

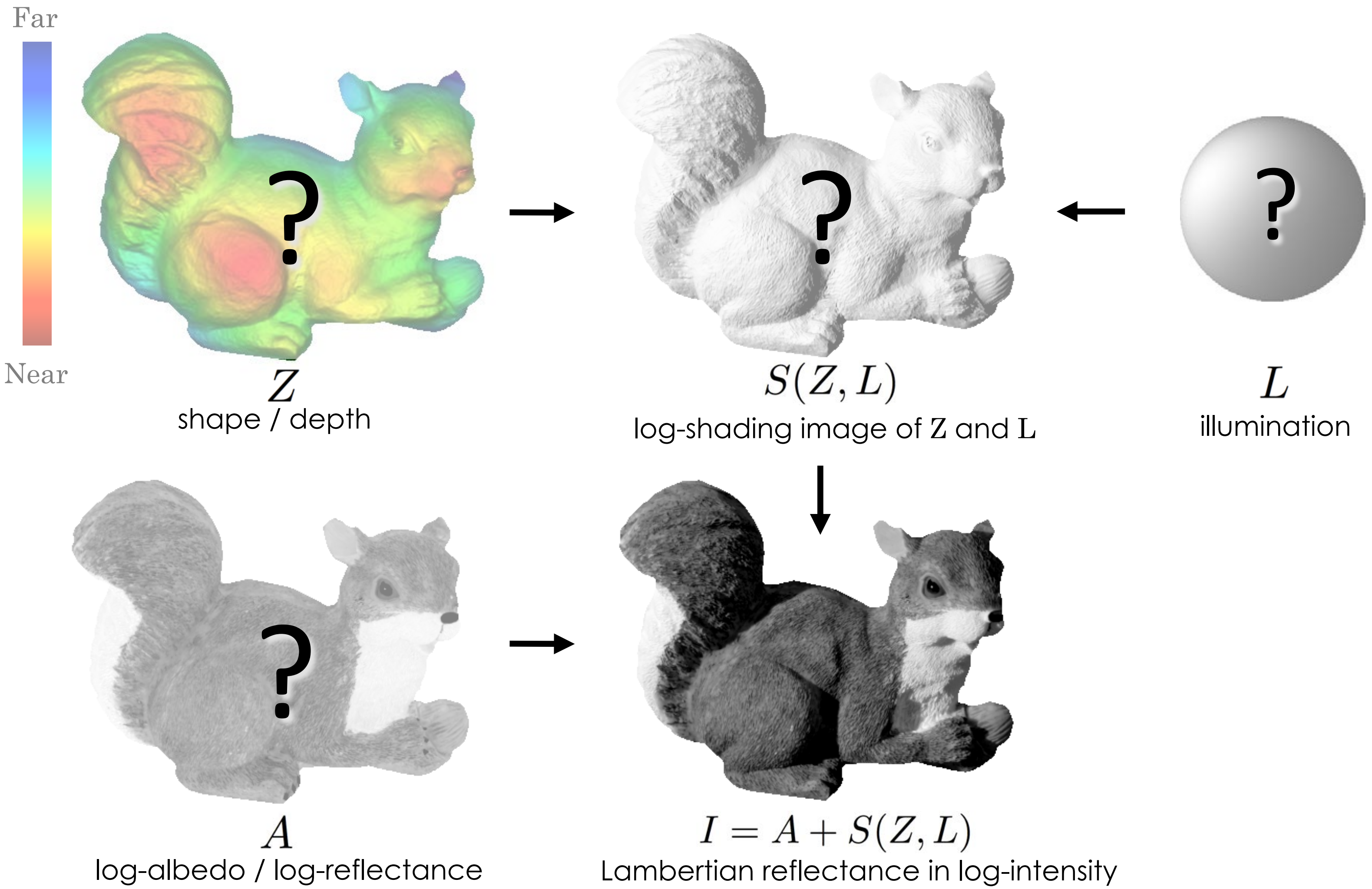


Forward Optics

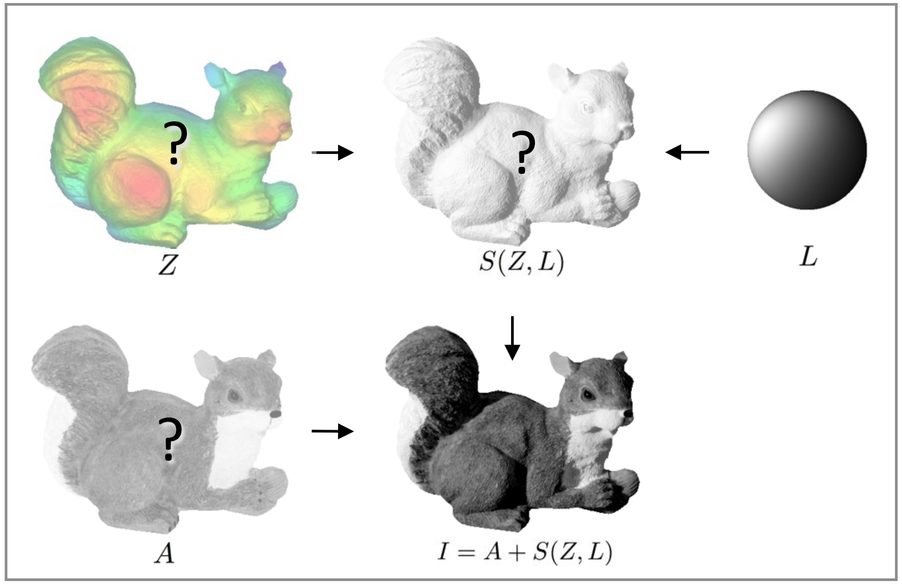


Shape, Albedo, and Illumination from Shading

SAIFS (“safes”)



Problem Formulation: Known Lighting



$$\begin{array}{ll} \underset{Z,A}{\text{maximize}} & P(A|Z, L)P(Z) \\ \text{subject to} & I = A + S(Z, L) \end{array}$$

"Find the most likely explanation (shape Z and log-albedo A) that together exactly reconstructs log-image I, given rendering engine $S()$ and known illumination L."

Demo!

What do we know about **reflectance**?

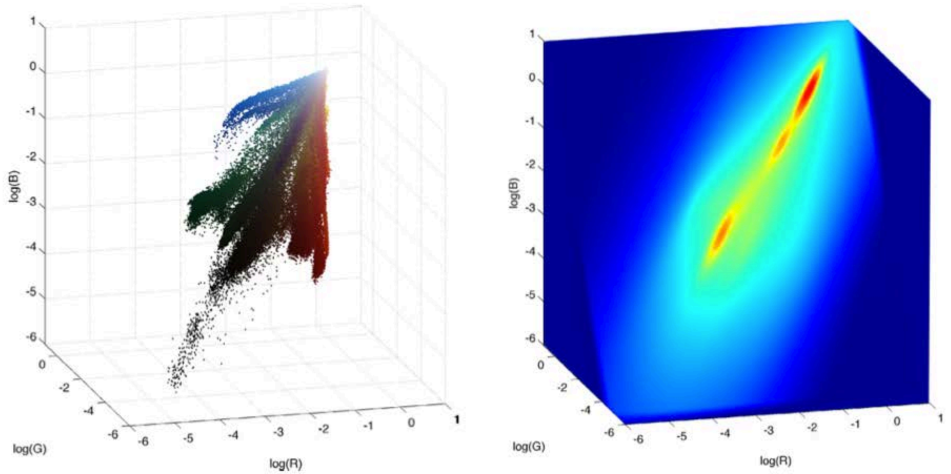
1) Piecewise smooth
(variation is small and sparse)

2) Palette is small
(distribution is low-entropy)

3) Some colors are common
(maximize likelihood under density model)

$$g(R) = \lambda_s \sum_i \sum_{j \in N(i)} \log \left(\sum_{k=1}^{K} \boldsymbol{\alpha}_k \mathcal{N}\left(R_i - R_j \,; \mathbf{0}, \boldsymbol{\sigma}_k\right) \right) - \lambda_e \log \left(\sum_i \sum_j \exp\left(-\frac{\left(R_i - R_j\right)^2}{4\sigma_e^2} \right) \right) + \lambda_a \sum_i \mathrm{F}(R_i)$$

Reflectance: Absolute Color



(a) Training reflectances (b) Our PDF of reflectance

What do we know about **shapes**?

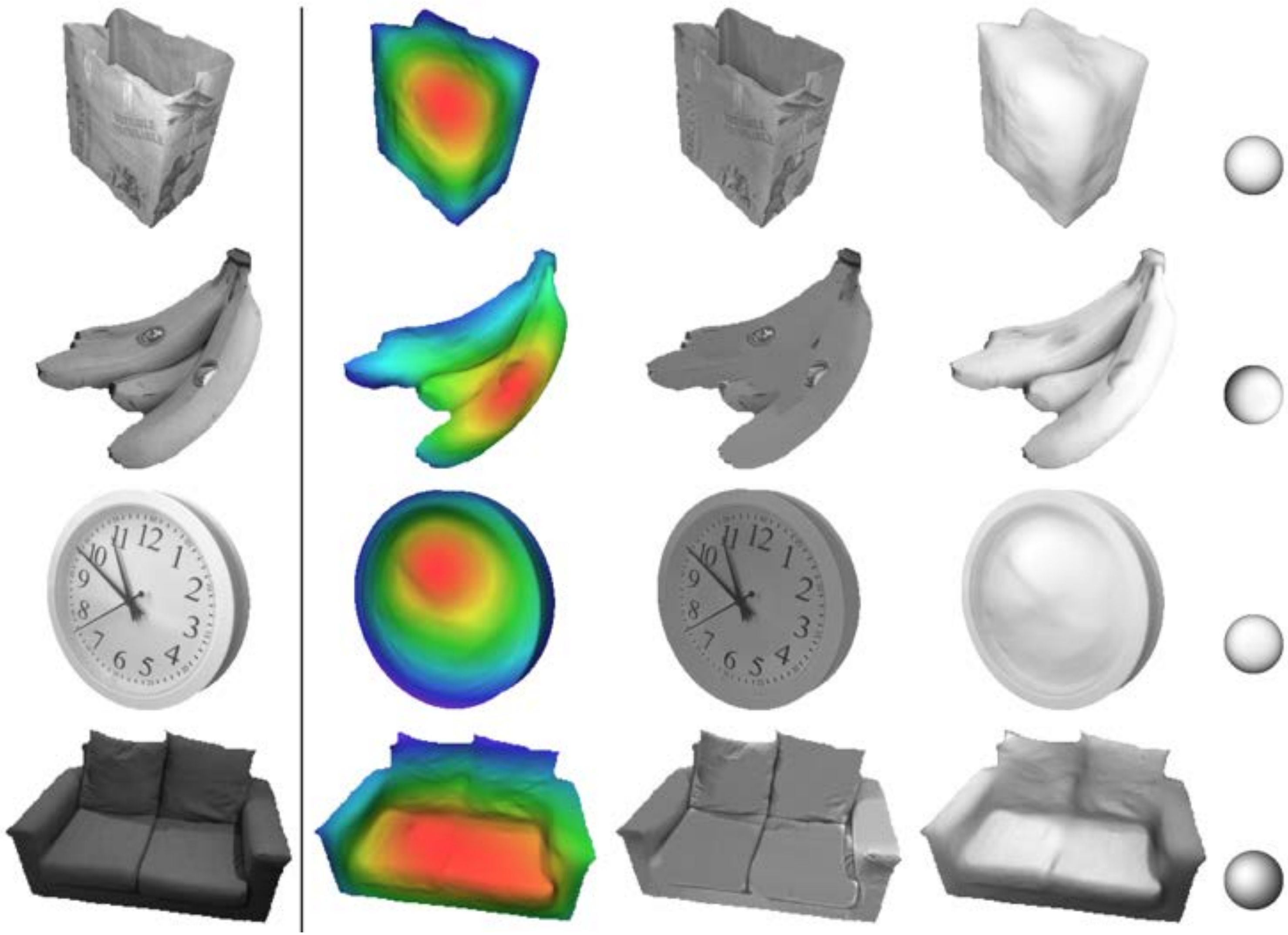
1) Piecewise smooth
(variation in mean curvature is small and sparse)

2) Face outward at the occluding contour

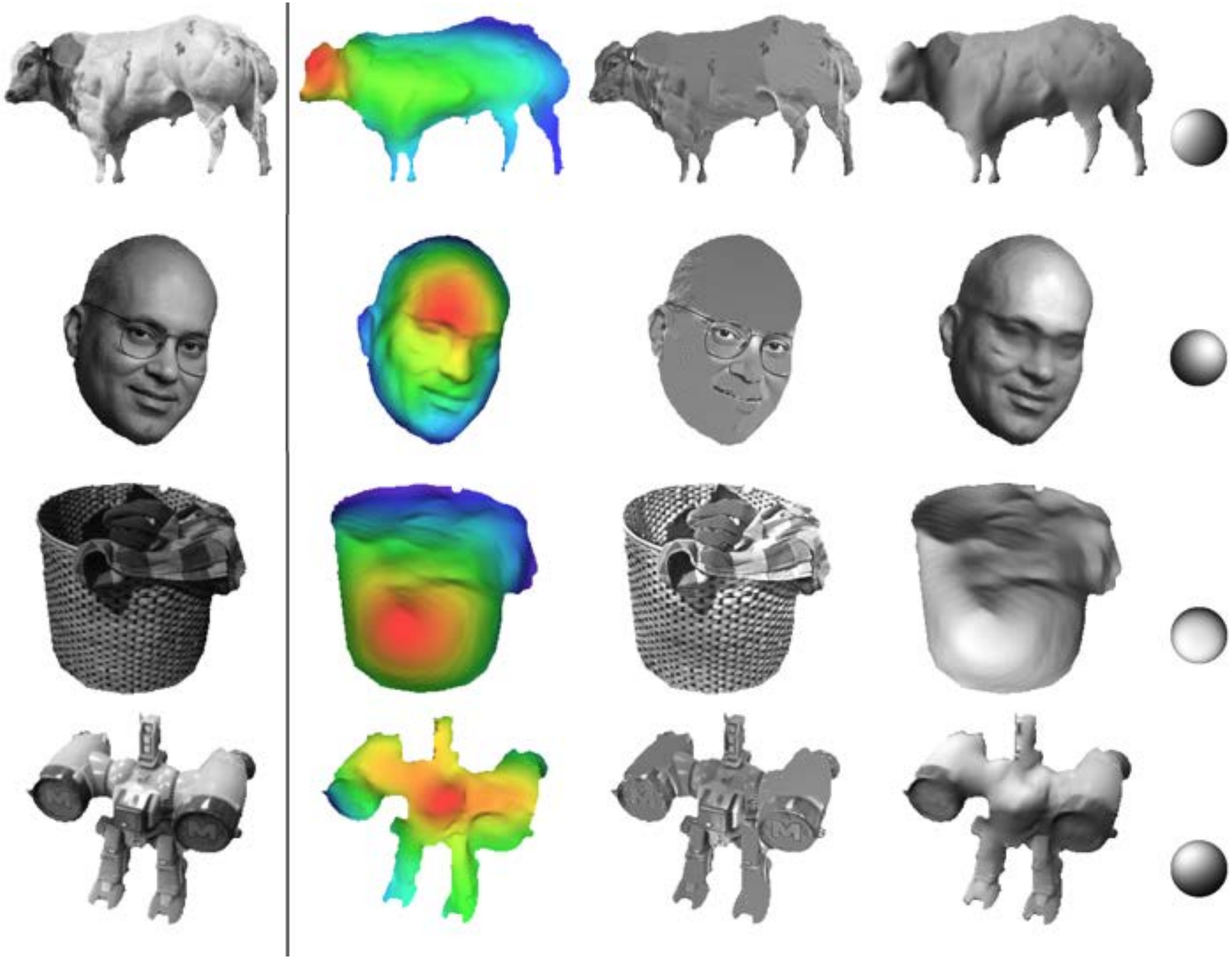
3) Tend to be fronto-parallel
(slant tends to be small)

$$f(Z) = \lambda_k \sum_i \sum_{j \in N(i)} \log\left(\sum_{k=1}^{K} \boldsymbol{\alpha}_k \mathcal{N}\left(H(Z)_i - H(Z)_j \,; 0, \boldsymbol{\sigma}_k\right)\right) + \lambda_c \sum_{i \in C} \sqrt{\left(N_i^x(Z) - n_i^x\right)^2 + \left(N_i^y(Z) - n_i^y\right)^2} - \lambda_f \sum_{x,y} \log\left(2N_{x,y}^z(Z)\right)$$

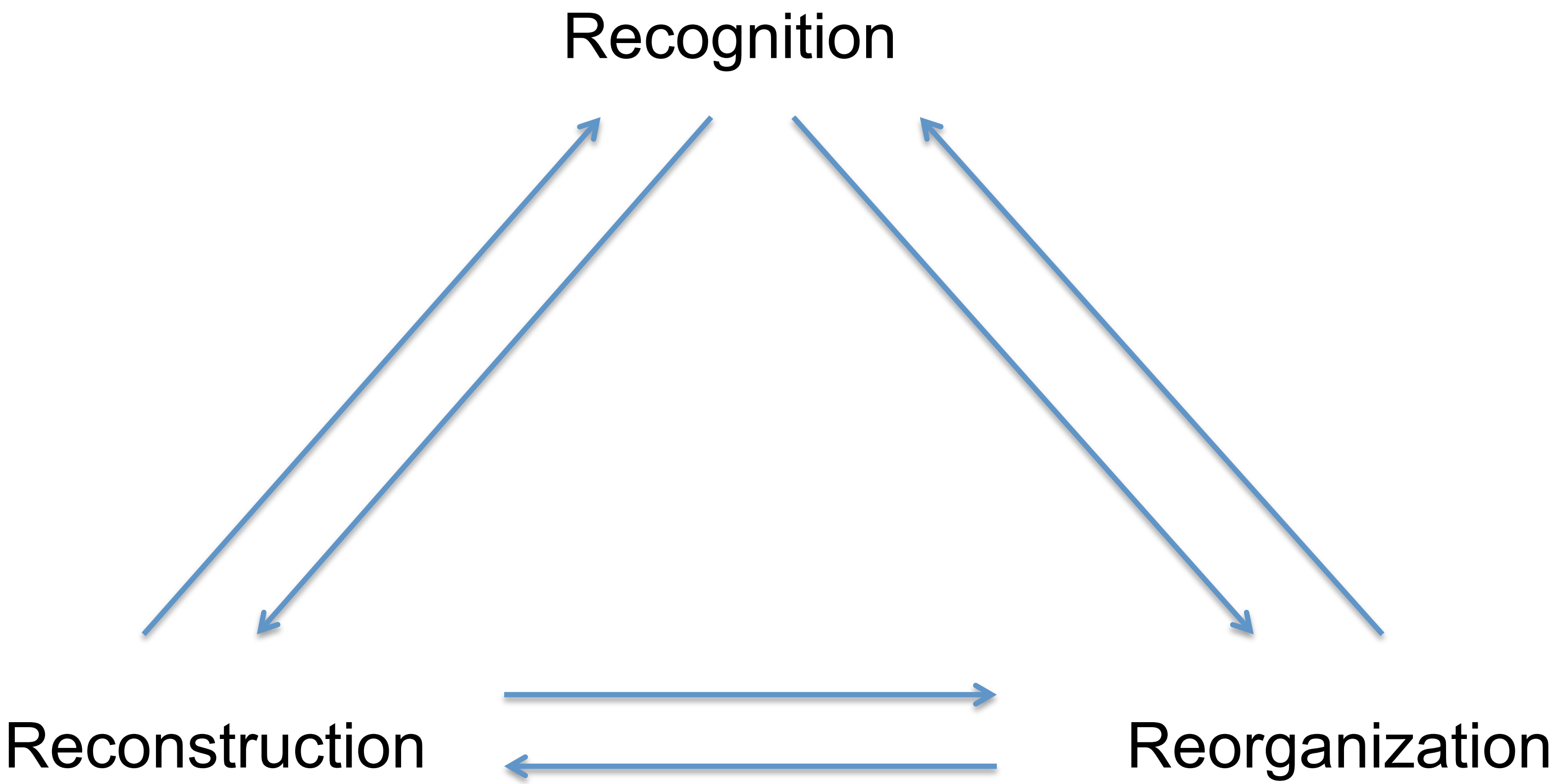
Evaluation: Real World Images



Evaluation: Real World Images



The Three R’s of Vision

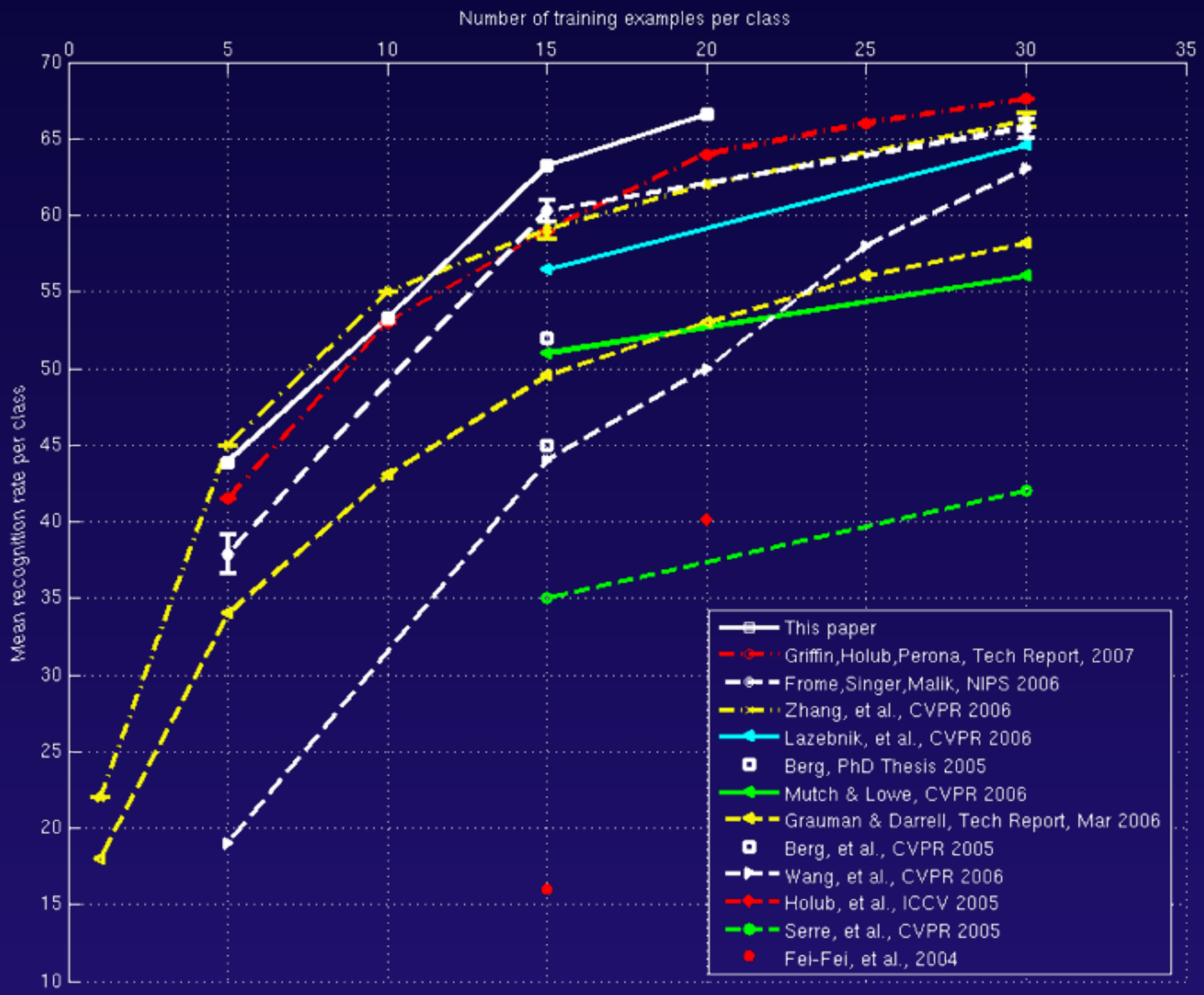


Caltech-101 [Fei-Fei et al. 04]

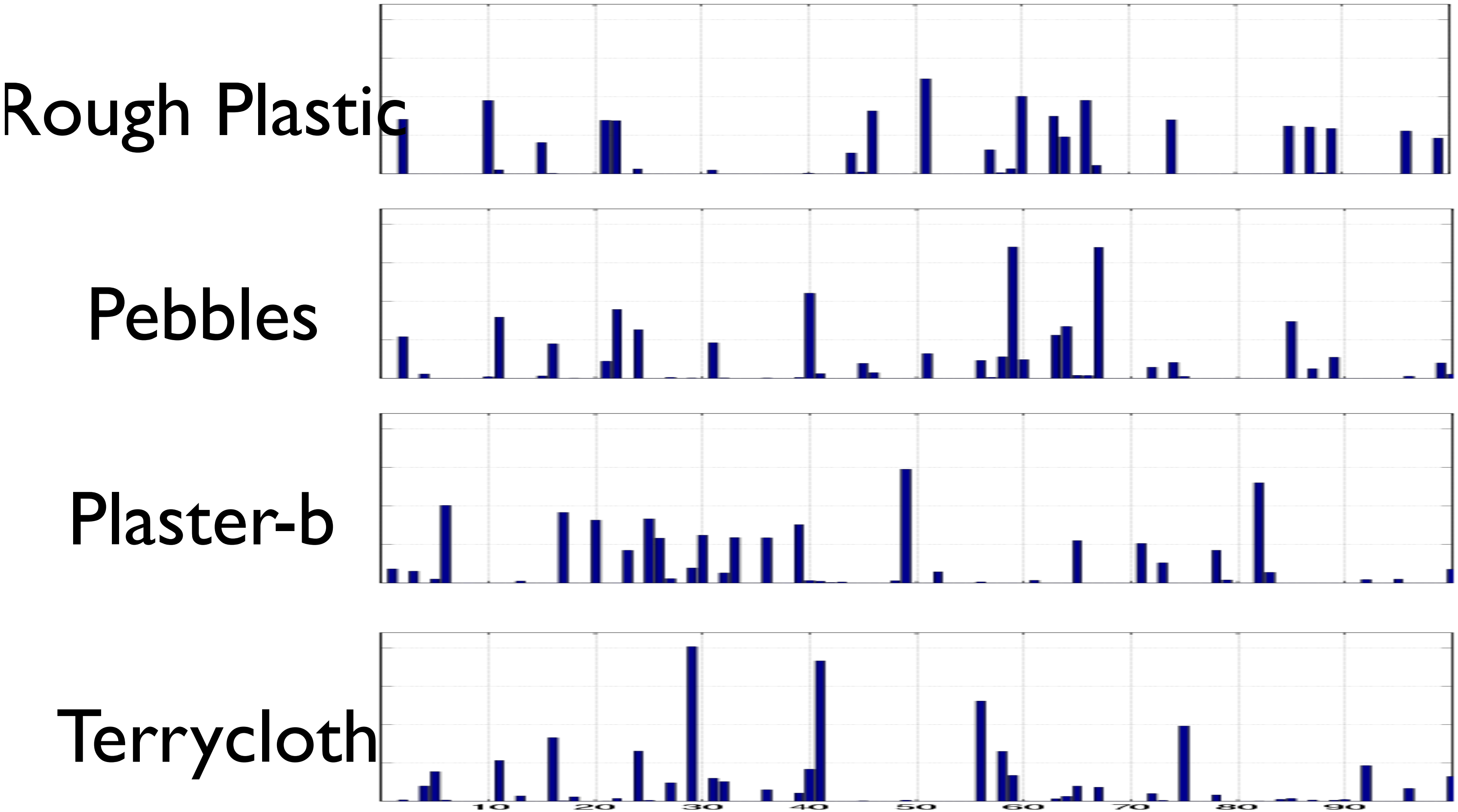
- 102 classes, 31-300 images/class

Caltech 101 classification results

(even better by combining cues..)



Texton Histogram Model for Recognition (Leung & Malik, 1999) cf. Bag of Words



Lazebnik, Schmid & Ponce (2006)

Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories

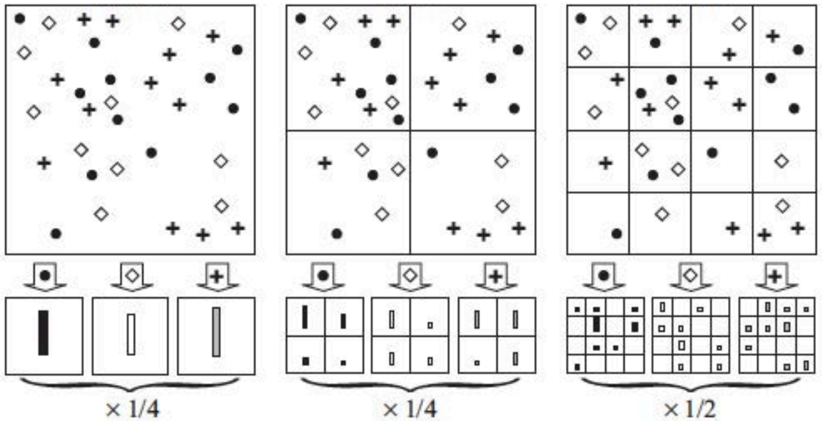
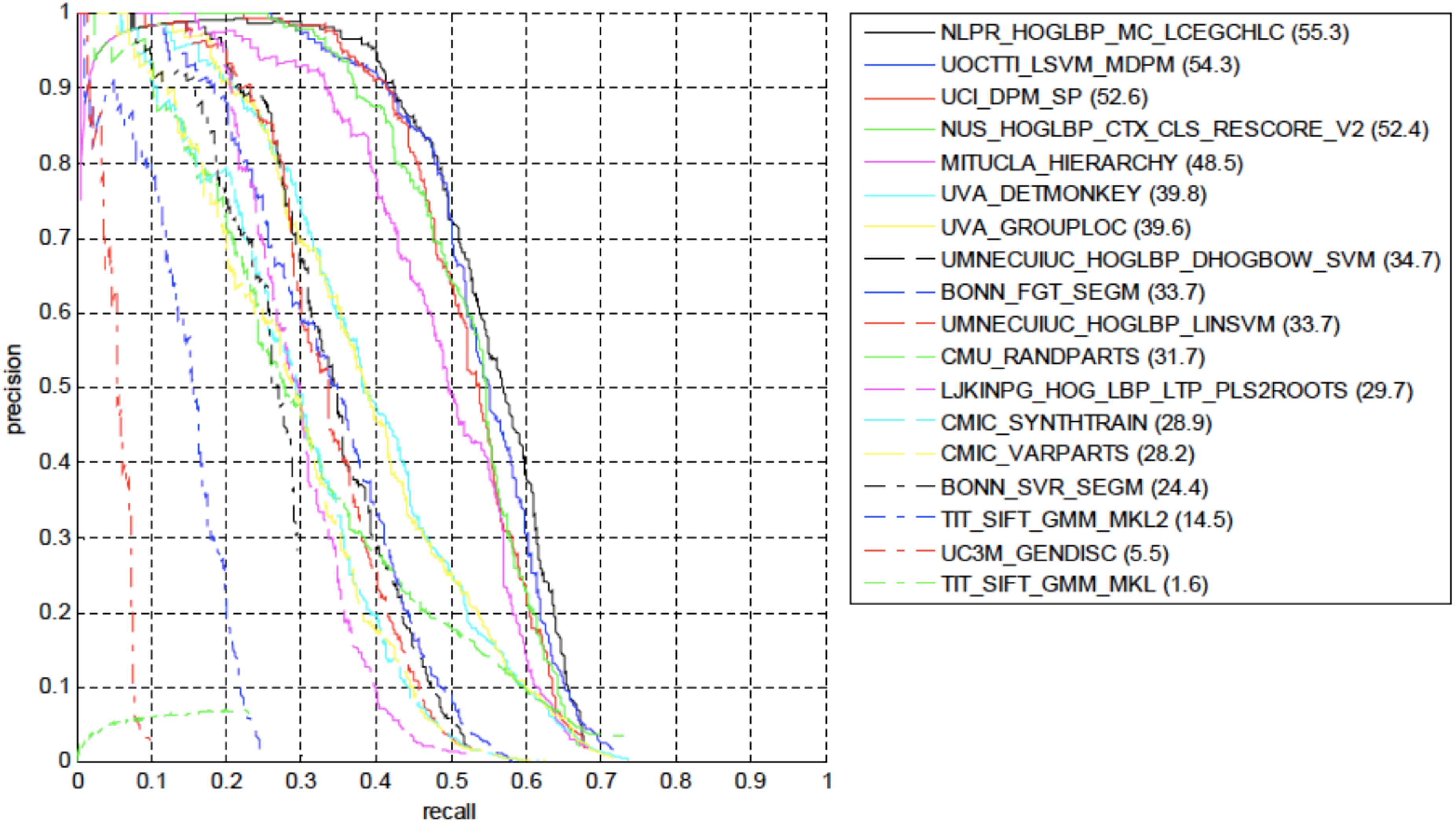


Figure 1. Toy example of constructing a three-level pyramid. The image has three feature types, indicated by circles, diamonds, and crosses. At the top, we subdivide the image at three different levels of resolution. Next, for each level of resolution and each channel, we count the features that fall in each spatial bin. Finally, we weight each spatial histogram according to eq. (3).

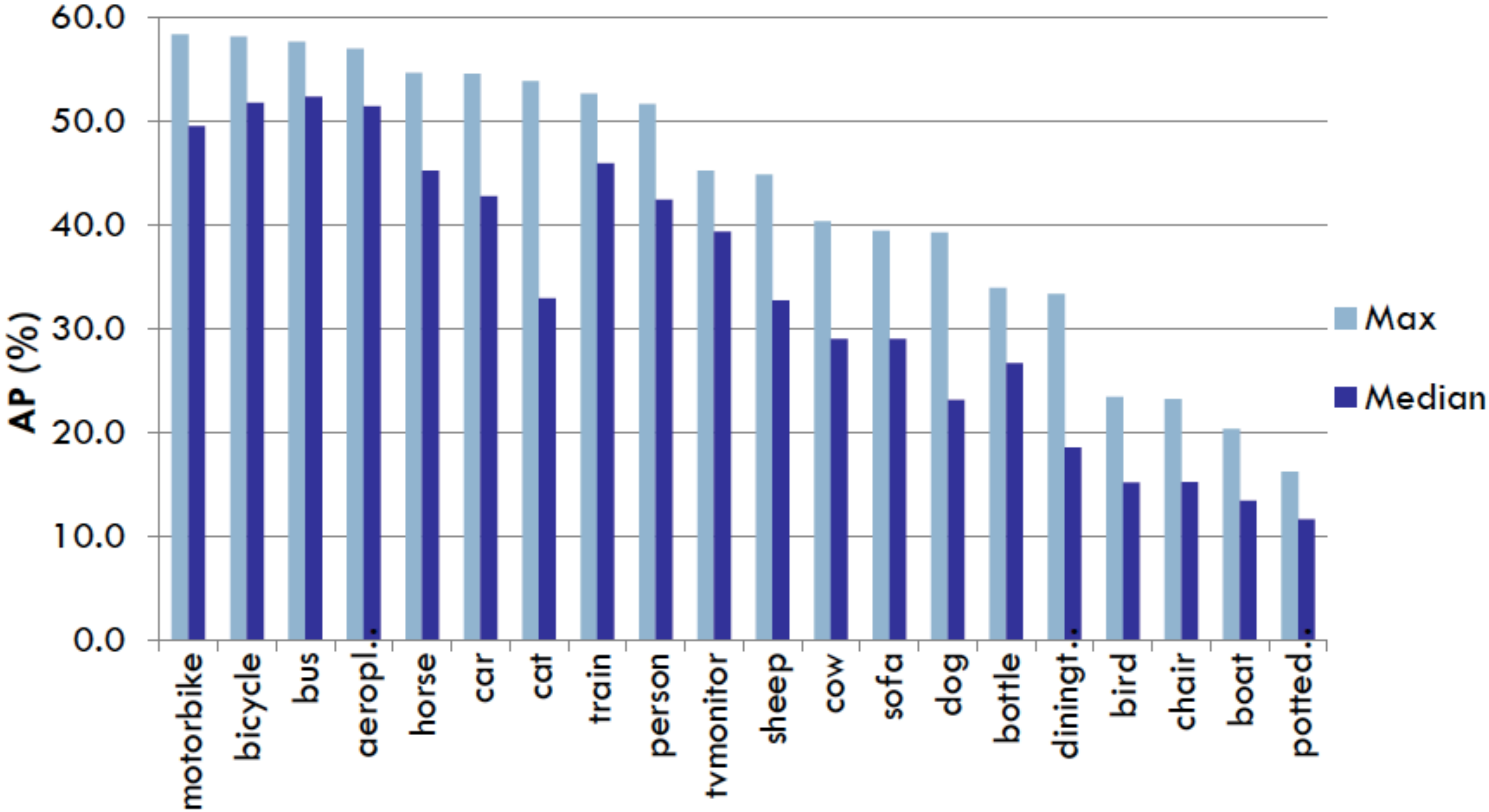
They proposed using vector-quantized SIFT descriptors as "words"

PASCAL Visual Object Challenge (Everingham et al)

Precision/Recall - Bicycle

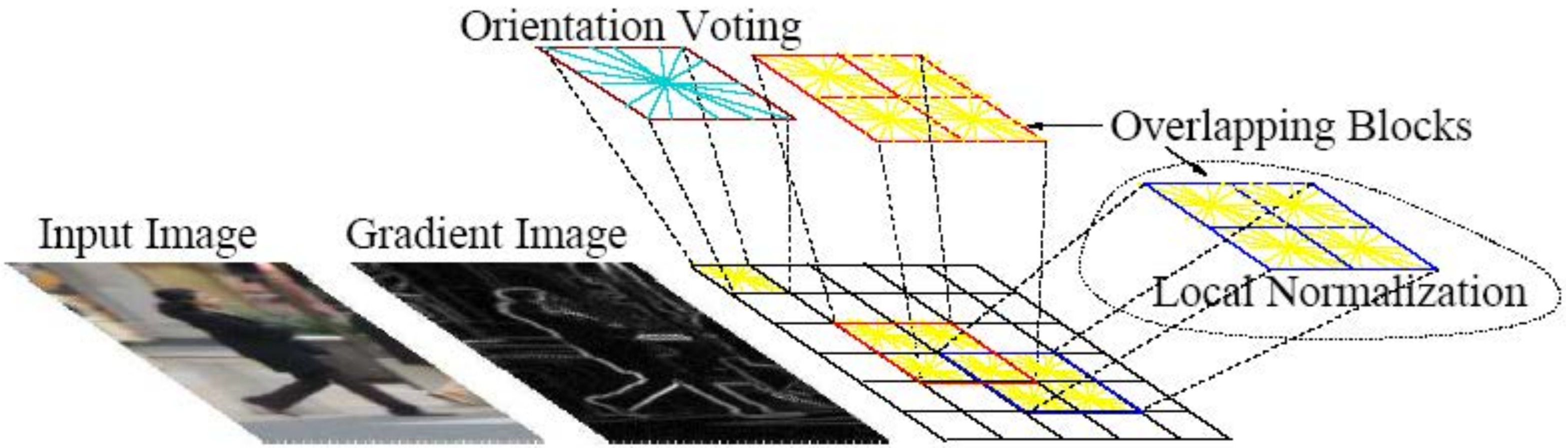


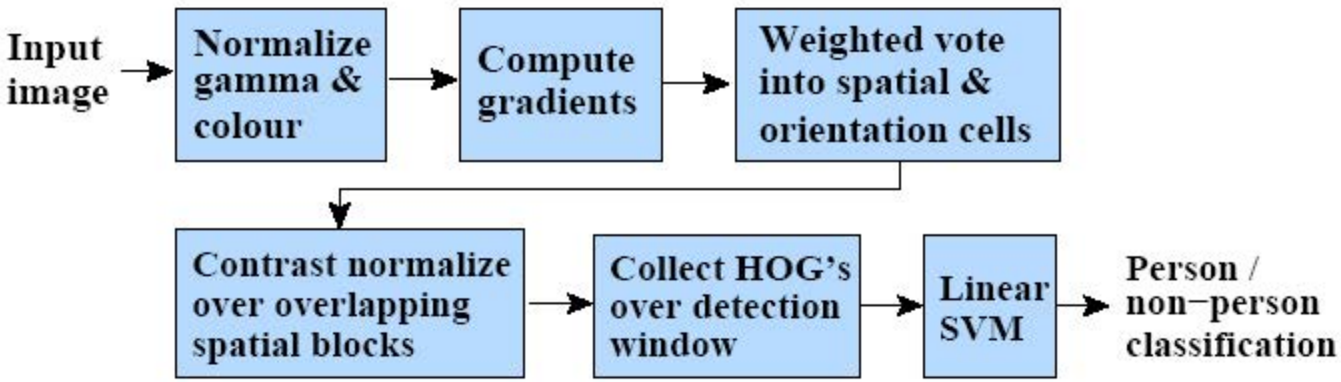
AP by Class



- Max AP: 58.3% (motorbike) ... 16.2% (potted plant)

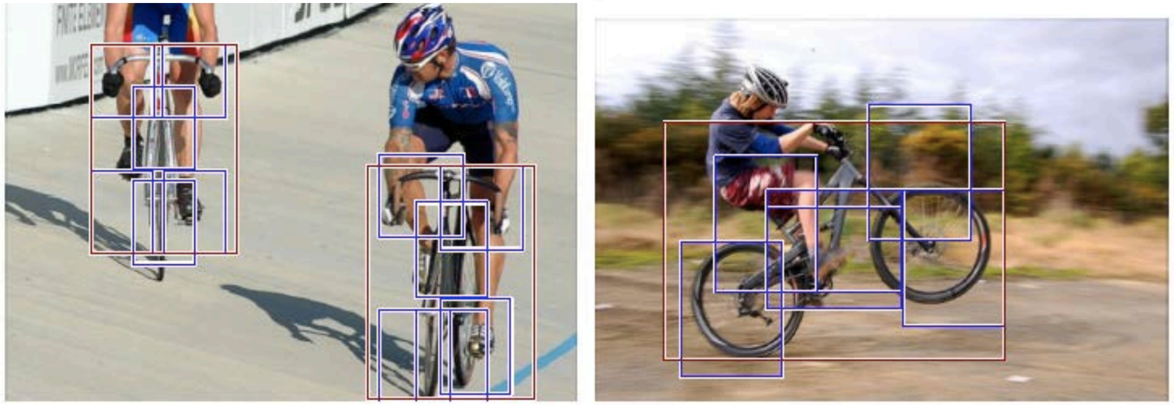
A good building block is a linear SVM trained on HOG features (Dalal & Triggs)

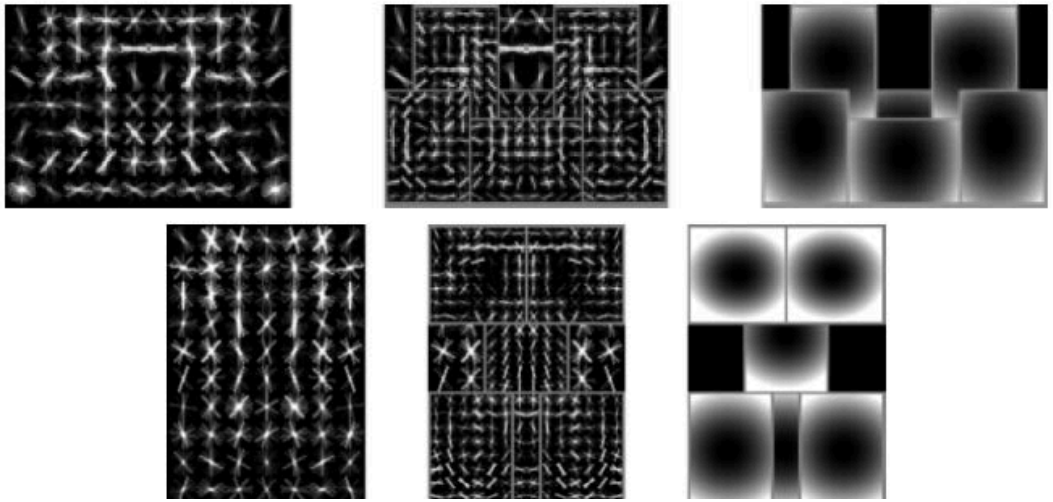




Object Detection with Discriminatively Trained Part Based Models

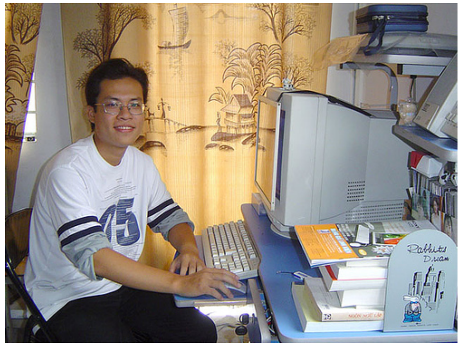
Pedro F. Felzenszwalb, Ross B. Girshick, David McAllester and Deva Ramanan





AP=0.23








Problems with current recognition approaches

- Performance is quite poor compared to that at 2d recognition tasks and the needs of many applications.
- Pose Estimation / Localization of parts or keypoints is even worse. We can't isolate decent stick figures from radiance images, making use of depth data necessary.
- Progress has slowed down. Variations of HOG/Deformable part models dominate.

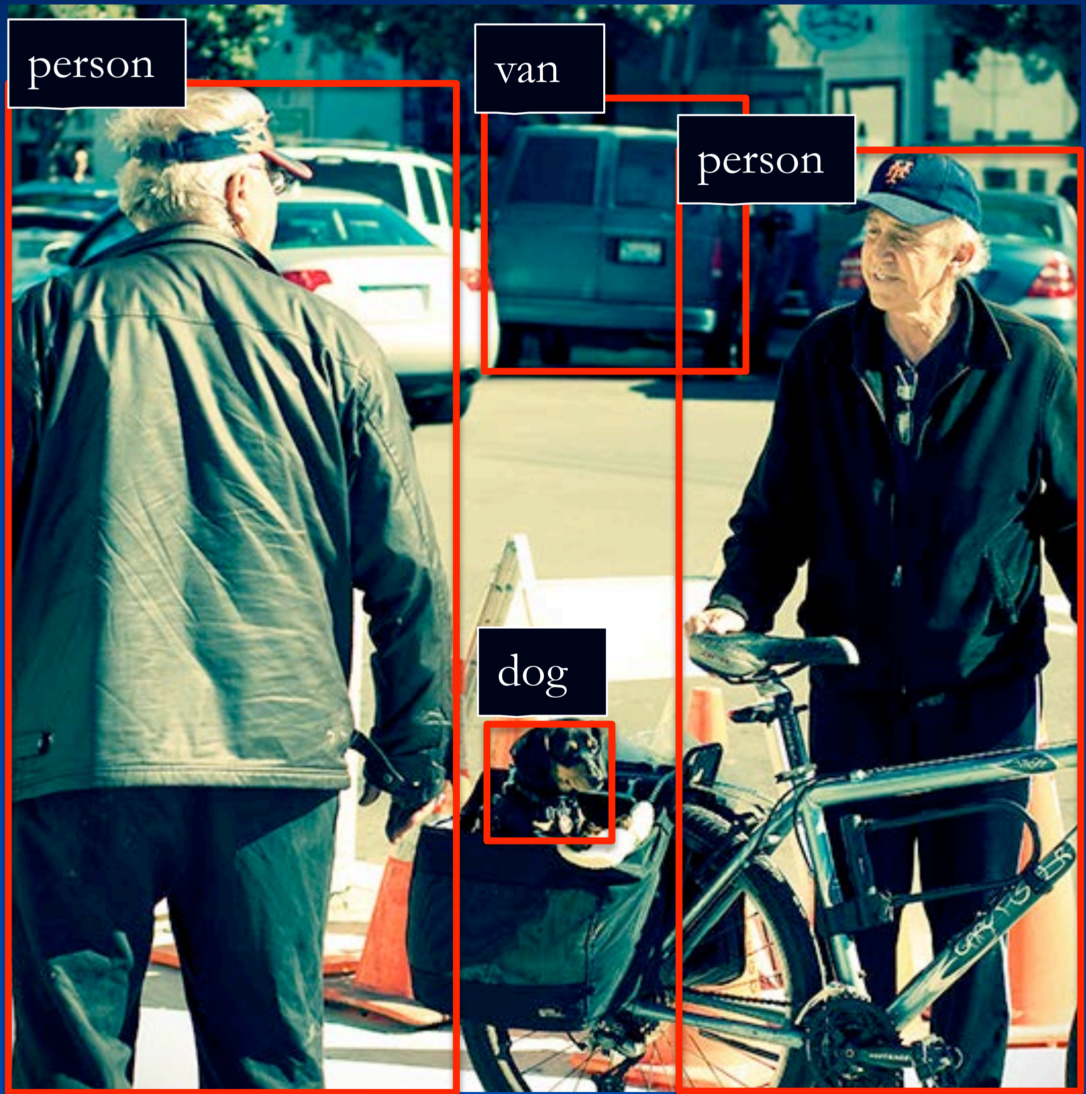
Next steps in recognition

- Incorporate the “shape bias” known from child development literature to improve generalization
  - This requires monocular computation of shape, as once posited in the 2.5D sketch, and distinguishing albedo and illumination changes from geometric contours
- Top down templates should predict keypoint locations and image support, not just information about category
- Recognition and figure-ground inference need to co-evolve. Occlusion is signal, not noise.


DESIRE
RACK

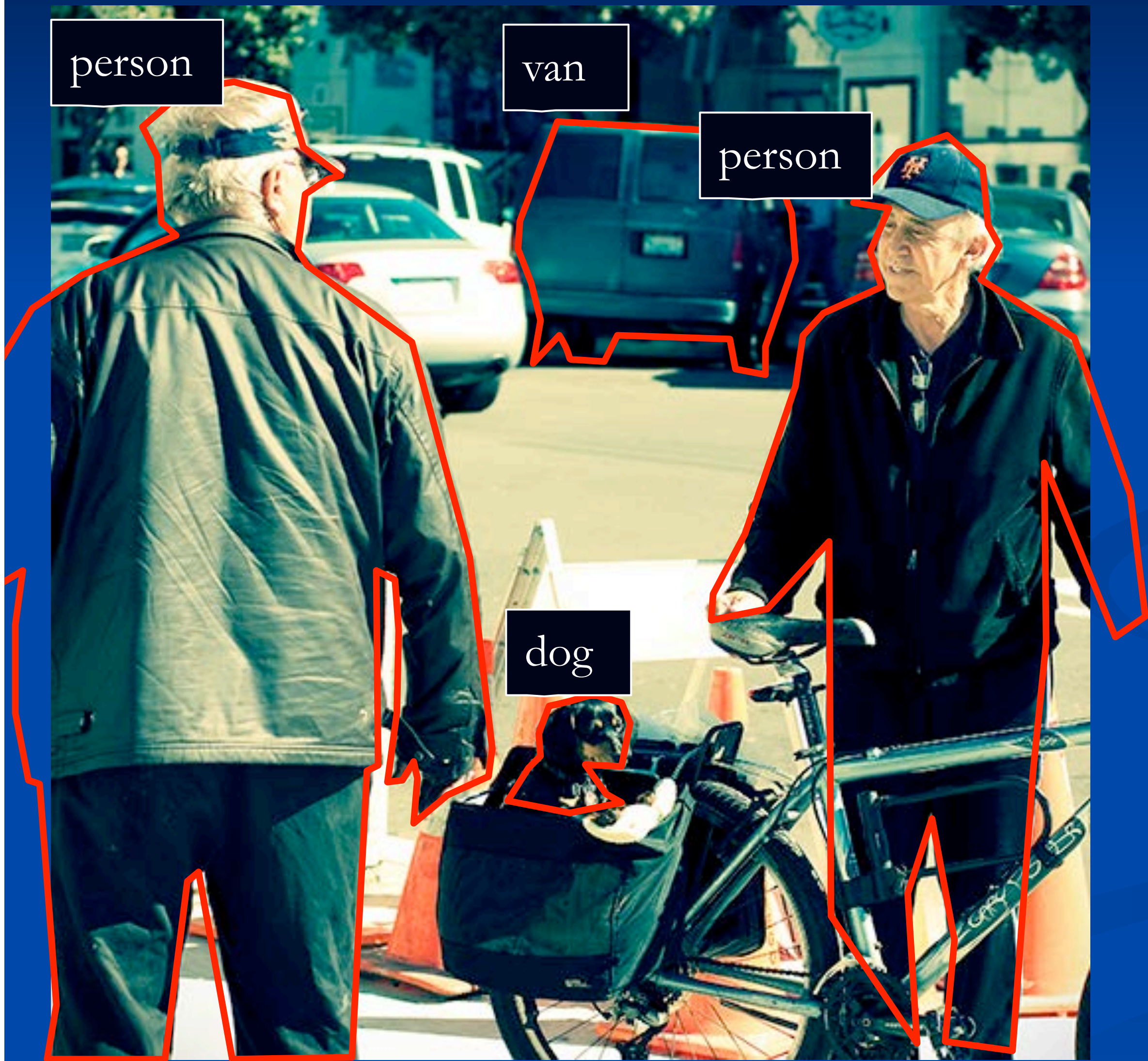
High-Level Computer Vision

High-Level Computer Vision



Object Recognition

High-Level Computer Vision



Object Recognition

Semantic Segmentation

High-Level Computer Vision

Object Recognition
Semantic Segmentation
Pose Estimation

High-Level Computer Vision

Object Recognition
Semantic Segmentation
Pose Estimation
Action Recognition

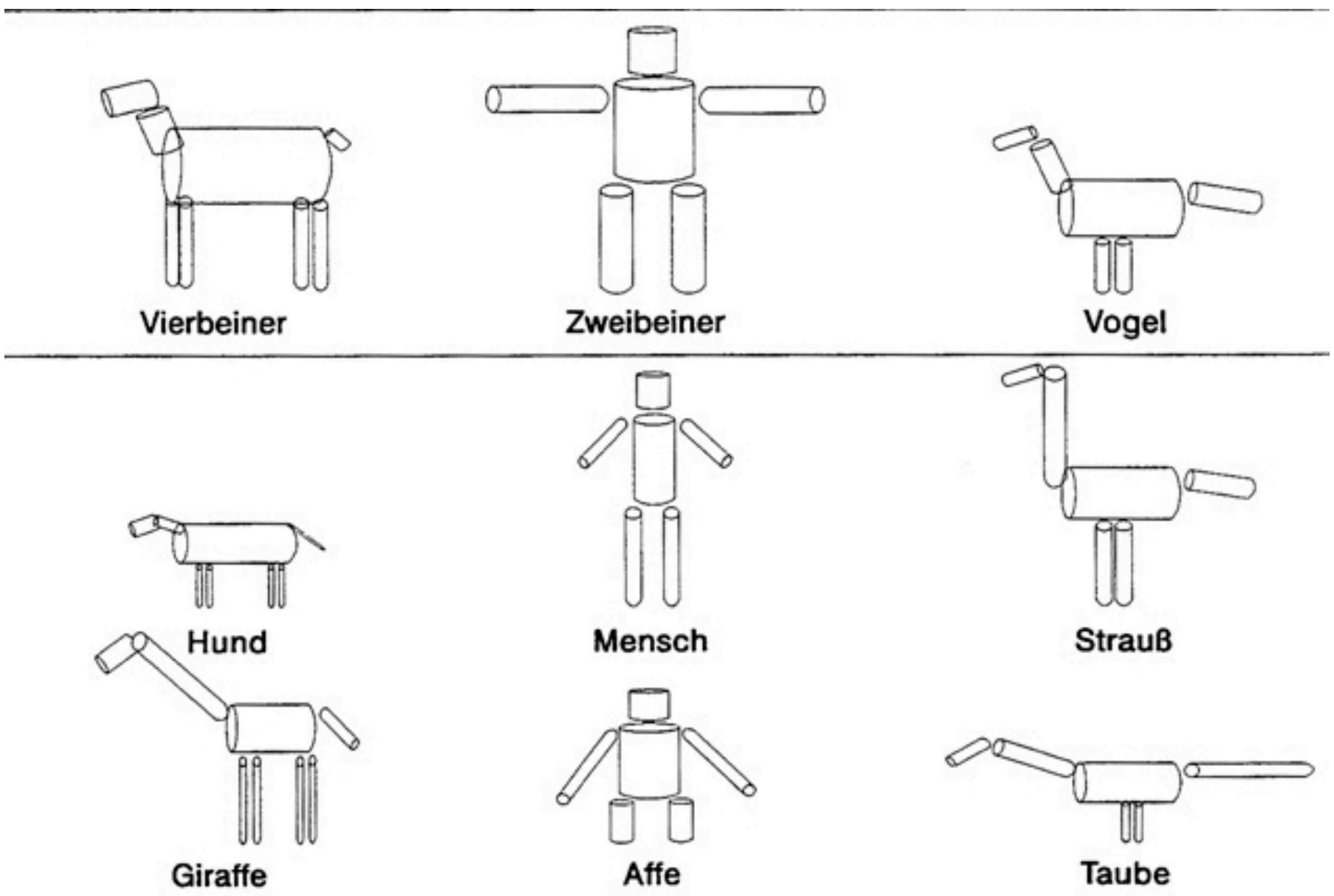
High-Level Computer Vision

Object Recognition
Semantic Segmentation
Pose Estimation
Action Recognition
Attribute Classification

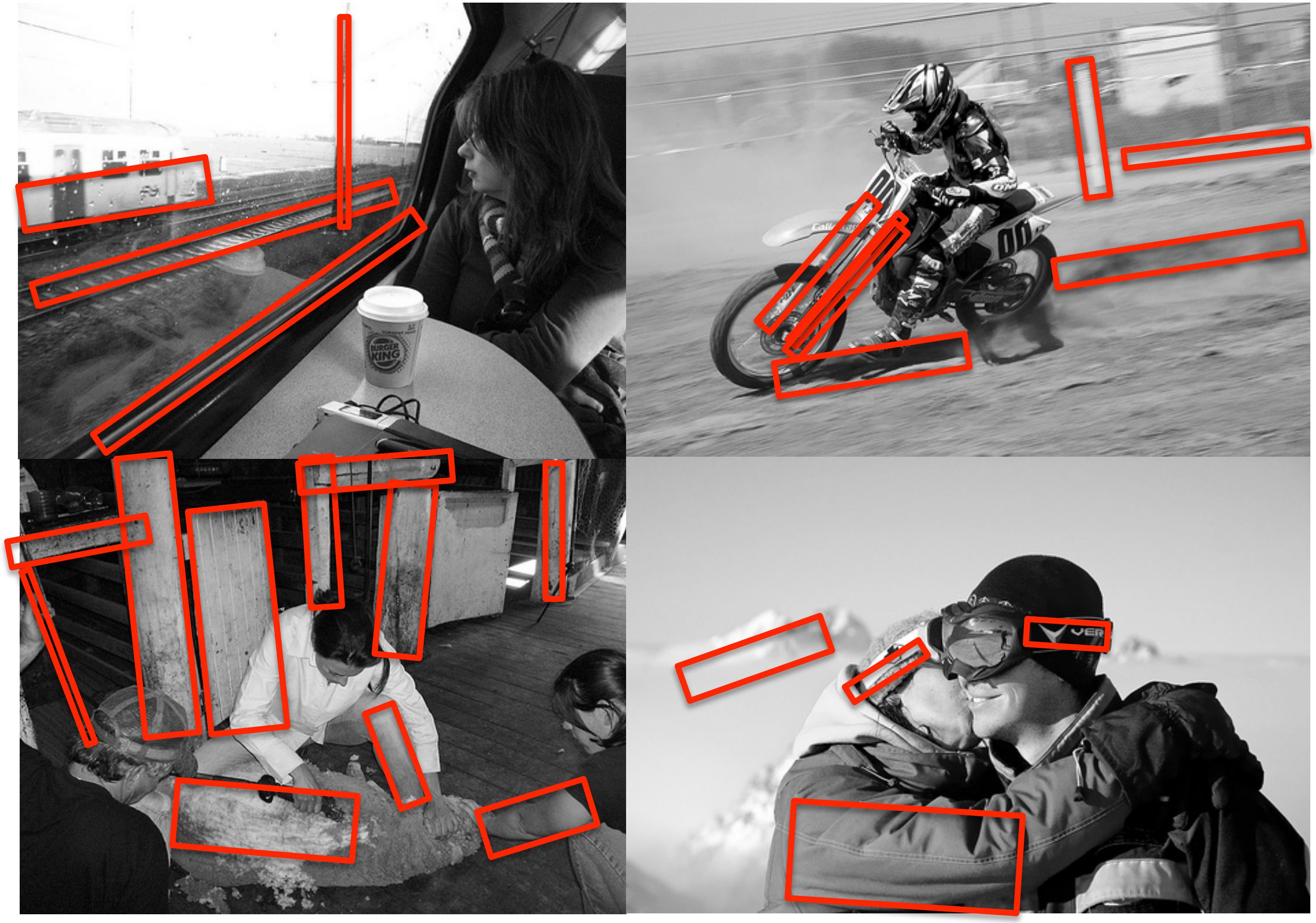
High-Level Computer Vision

Object Recognition
Semantic Segmentation
Pose Estimation
Action Recognition
Attribute Classification

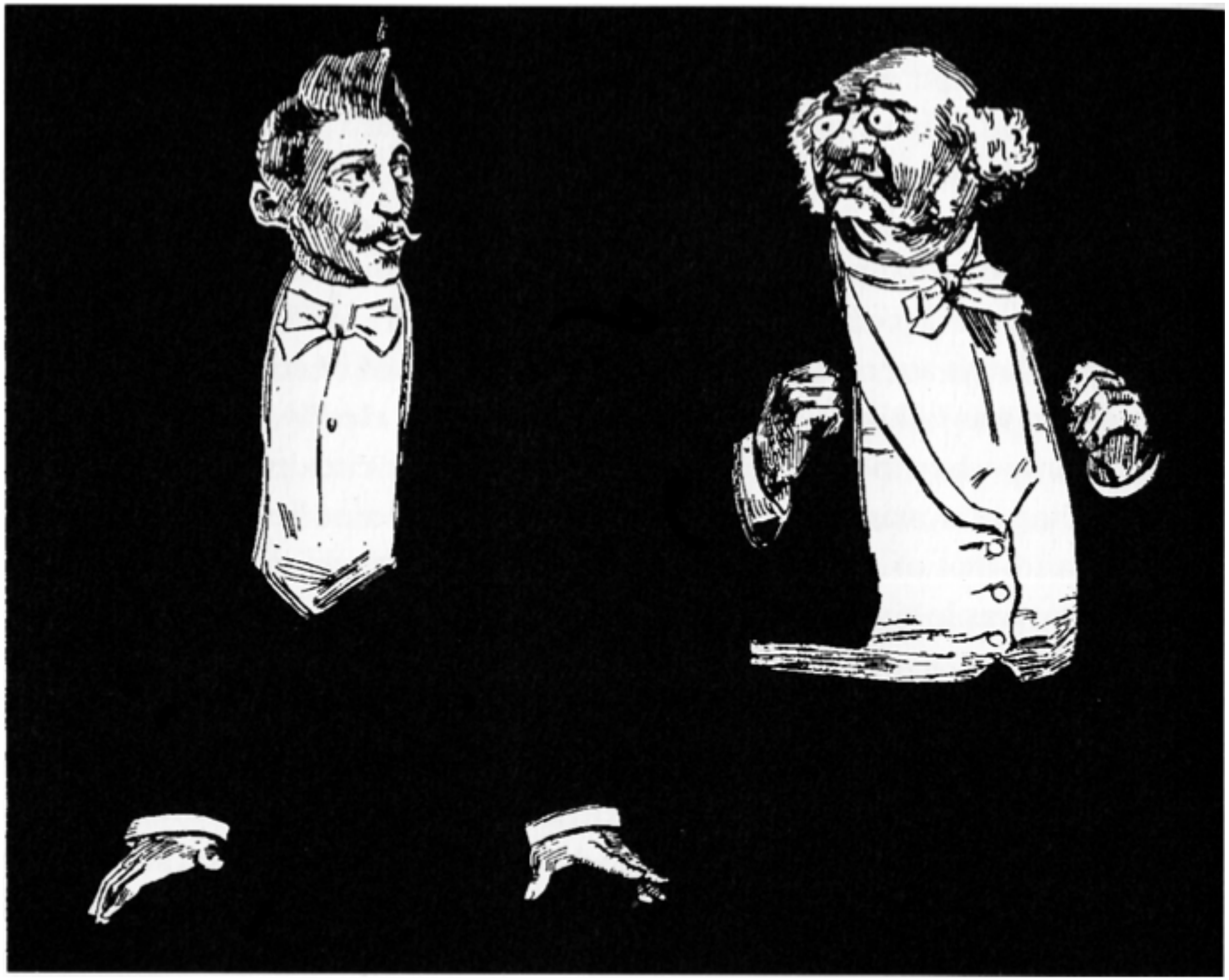
Trying to extract stick figures is hard (and unnecessary!)



All the wrong limbs…

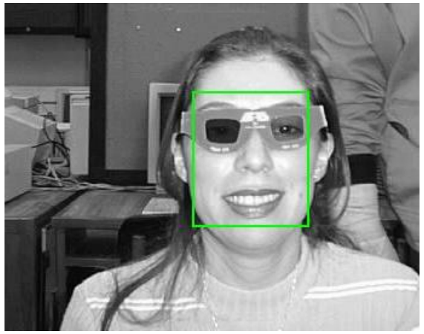


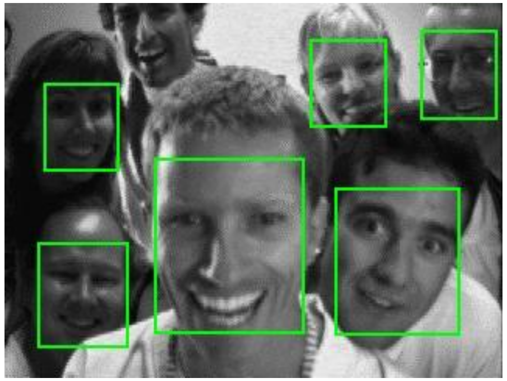
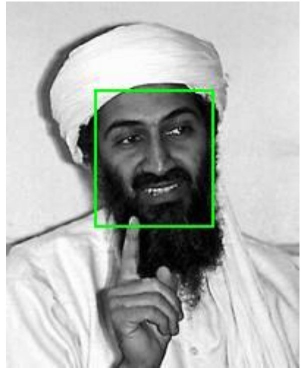
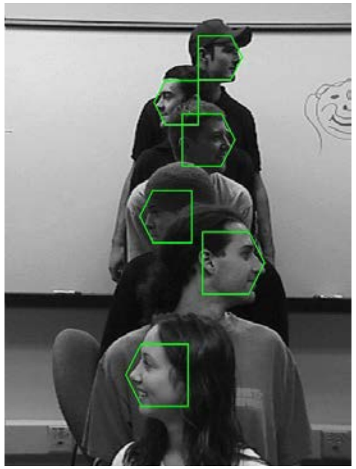
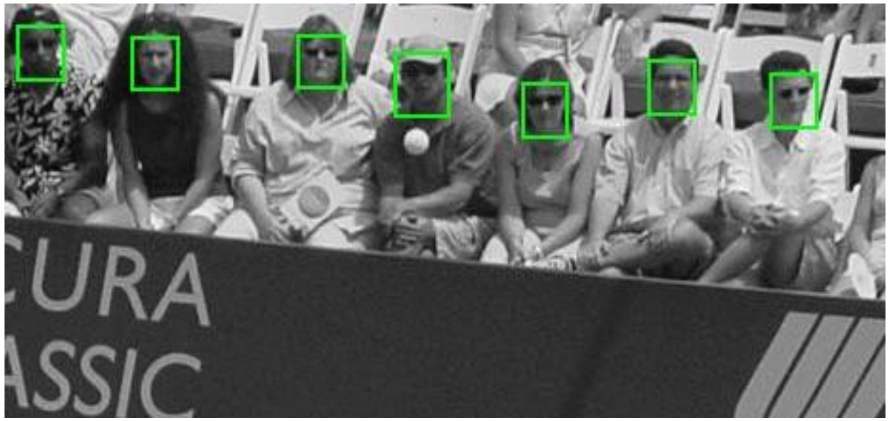
Motivation

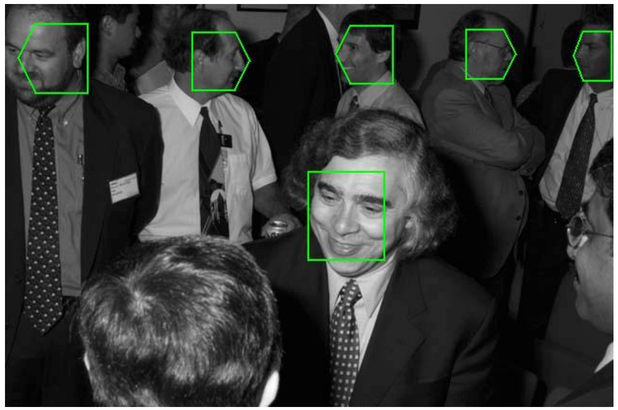


Face Detection

Carnegie Mellon University

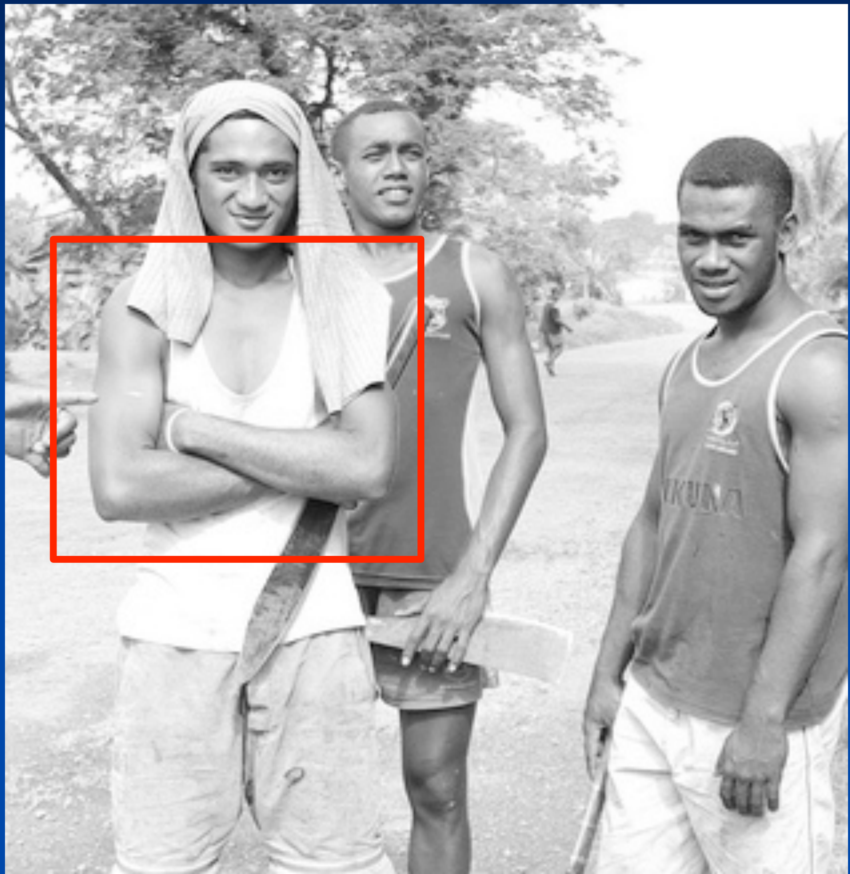


Examples of poselets (Bourdev & Malik , 2009)

Patches are often far **visually**, but they are close **semantically**

How do we train a poselet for a given pose configuration?

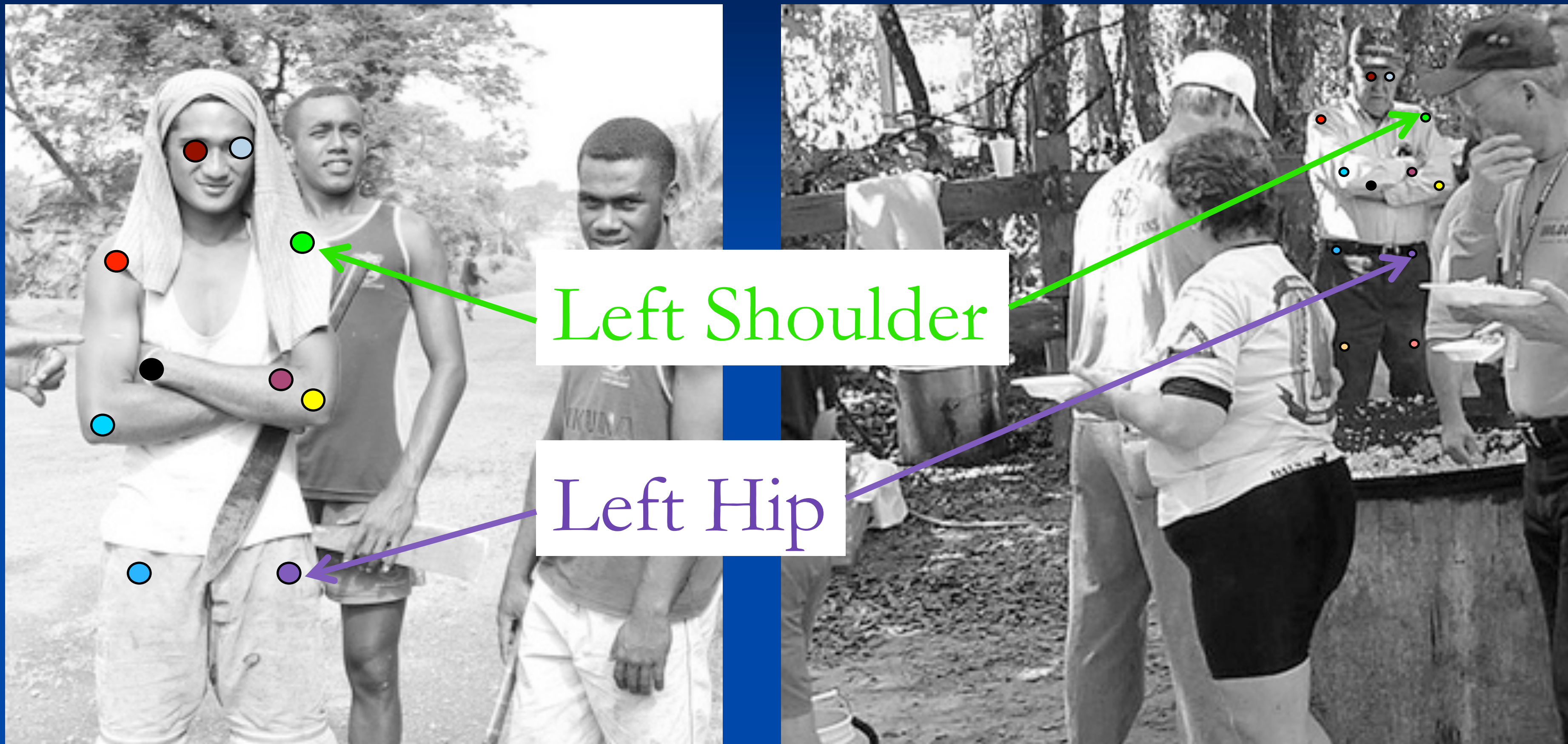
Finding Correspondences



Given part of a human pose

How do we find a similar pose configuration in the training set?

Finding Correspondences



We use keypoints to annotate the joints, eyes, nose, etc. of people

Finding Correspondences

Residual Error

Training poselet classifiers

Residual Error: 0.15 0.20 0.10 0.85 0.15 0.35

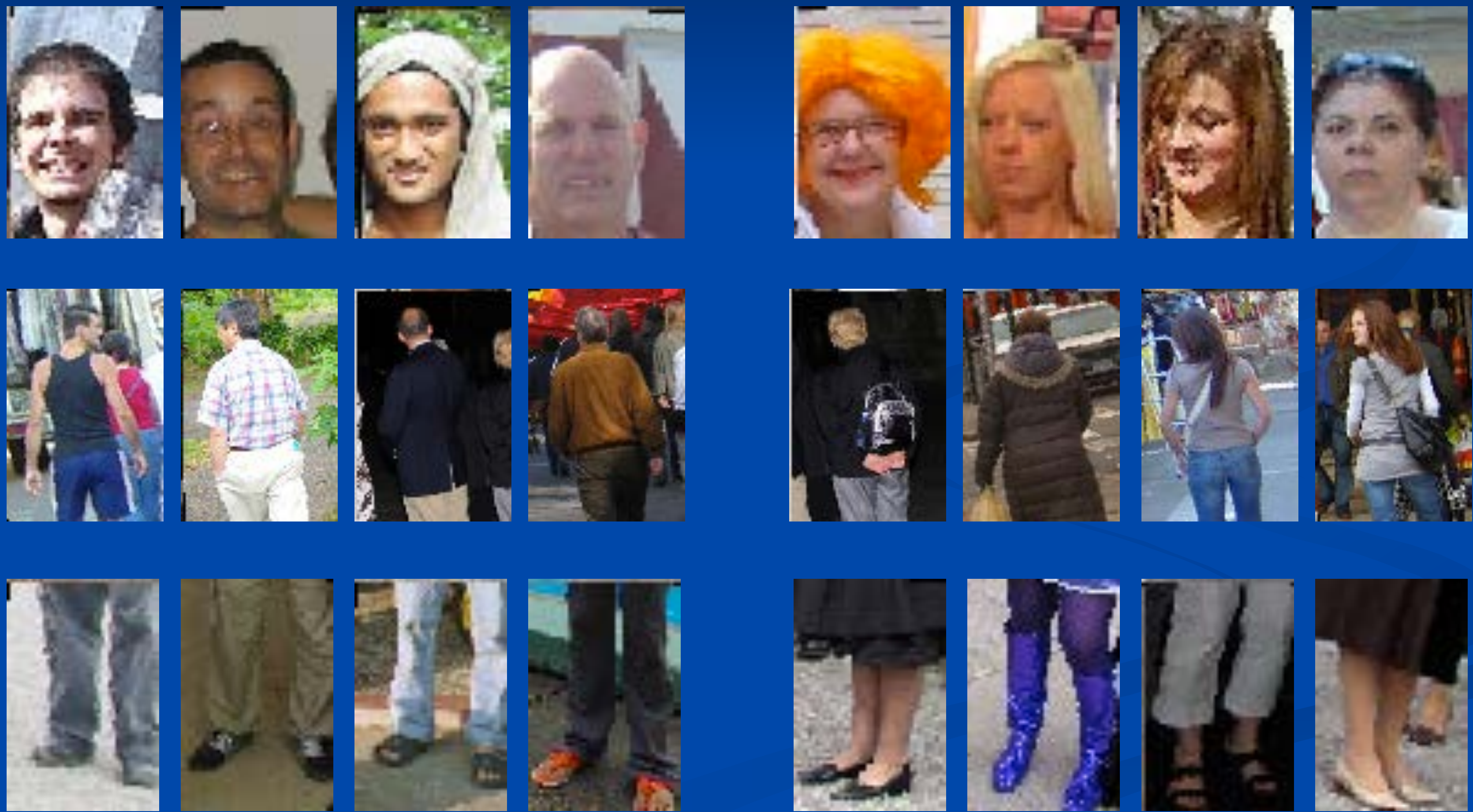
1. Given a seed patch
2. Find the closest patch for every other person
3. Sort them by residual error
4. Threshold them

Male or female?

How do we train attribute classifiers "in the wild"?

- Effective prediction requires inferring the pose and camera view
- Pose reconstruction is itself a hard problem, but we don't need perfect solution.
- We train attribute classifiers for each poselet
- Poselets implicitly decompose the pose

Gender classifier per poselet is much easier to train



Is male

Has long hair

Wears a hat

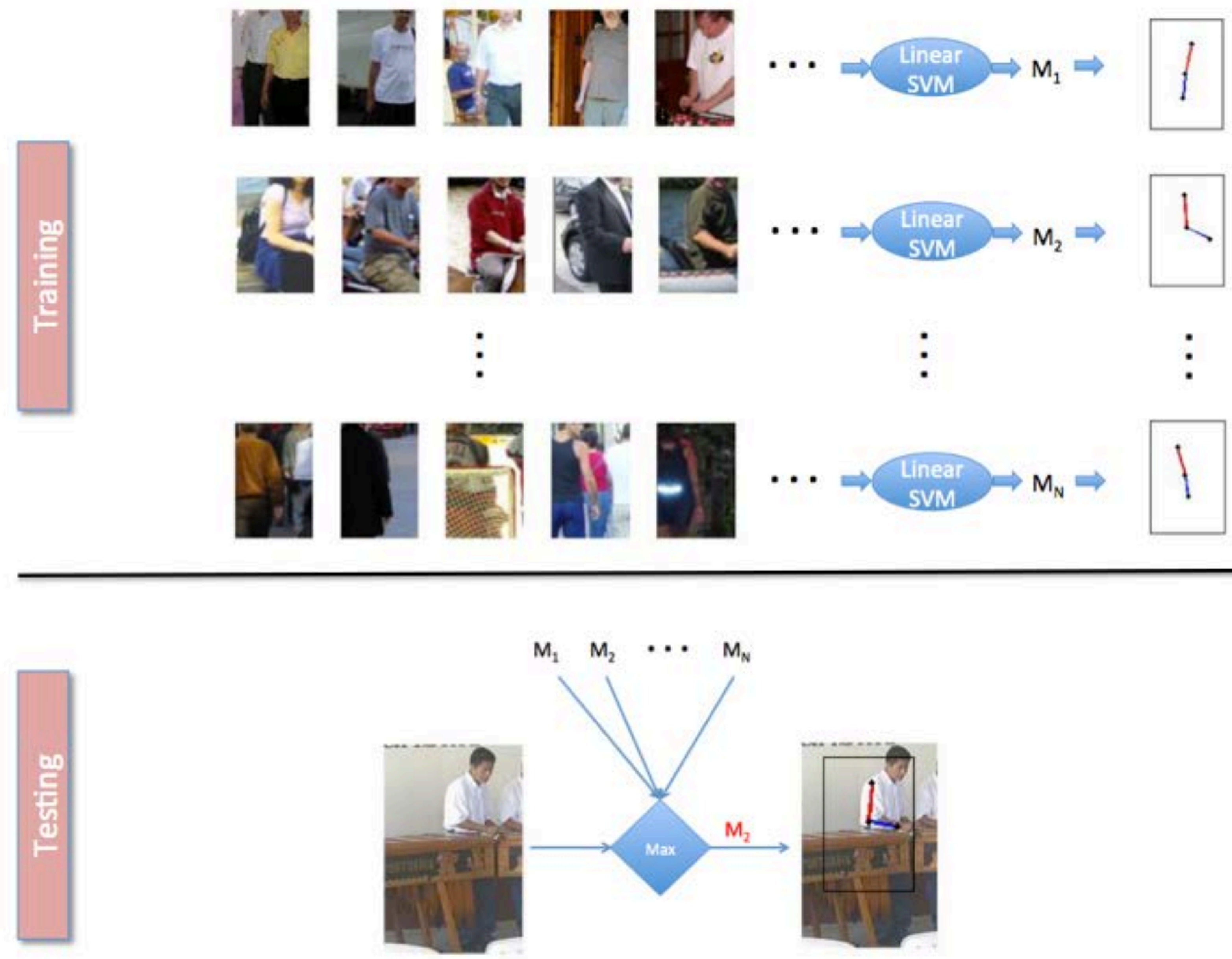
Wears glasses

Wears long pants

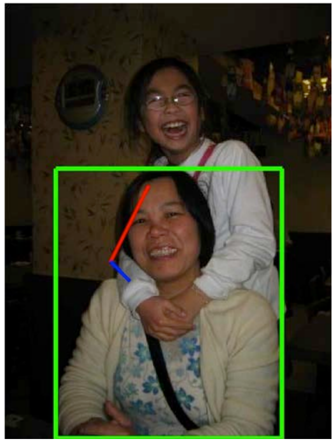
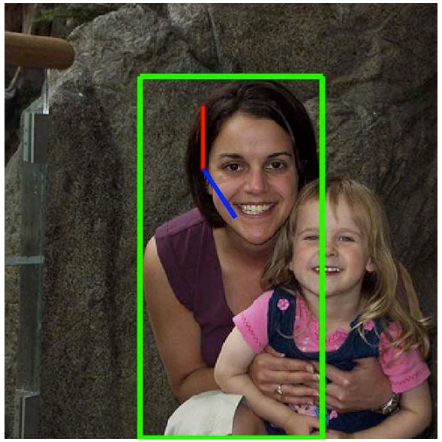
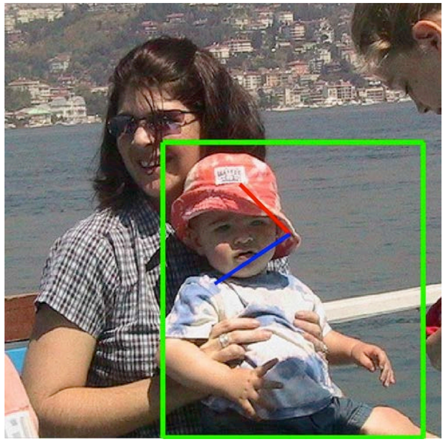
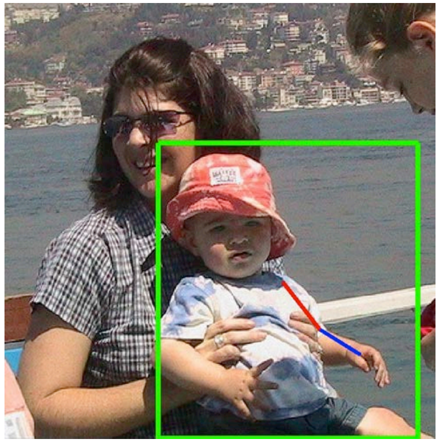
Wears long sleeves

Some discriminative poselets

Armlets (Gkioxari et al, CVPR 2013)



Multiple Instances

	Right Arm		Left Arm	
Yang & Ramanan				
Our method				

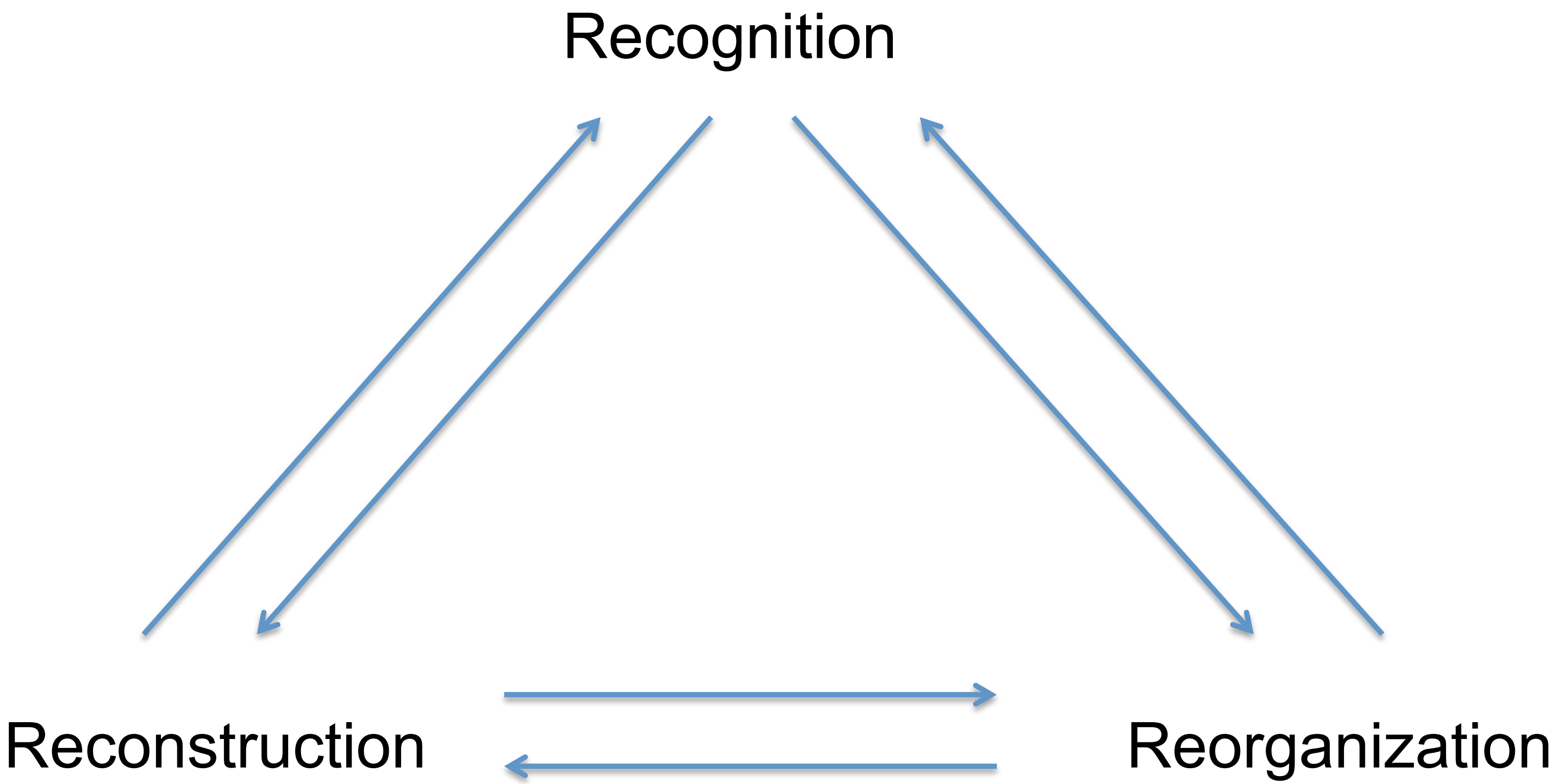
Results

- Results of Augmented Armlets and Comparison with baseline[1]

PCP	Yang & Ramanan [1]	Our model
R_UpperArm	38.9	50.2
R_Lower Arm	21.0	25.0
L_Upper Arm	36.9	49.2
L_Lower Arm	19.1	25.4
Average	29.0	37.5

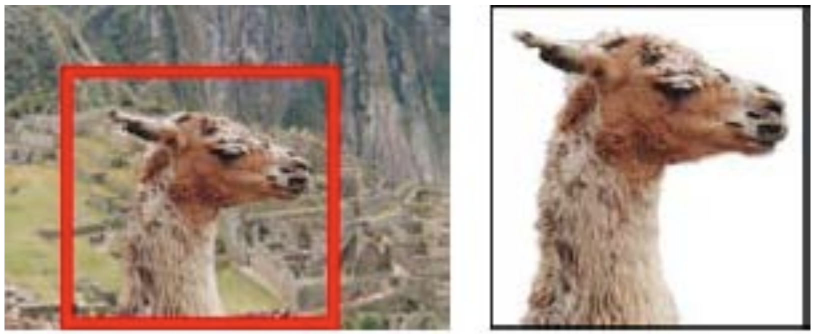
[1] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. CVPR, 2011

The Three R’s of Vision



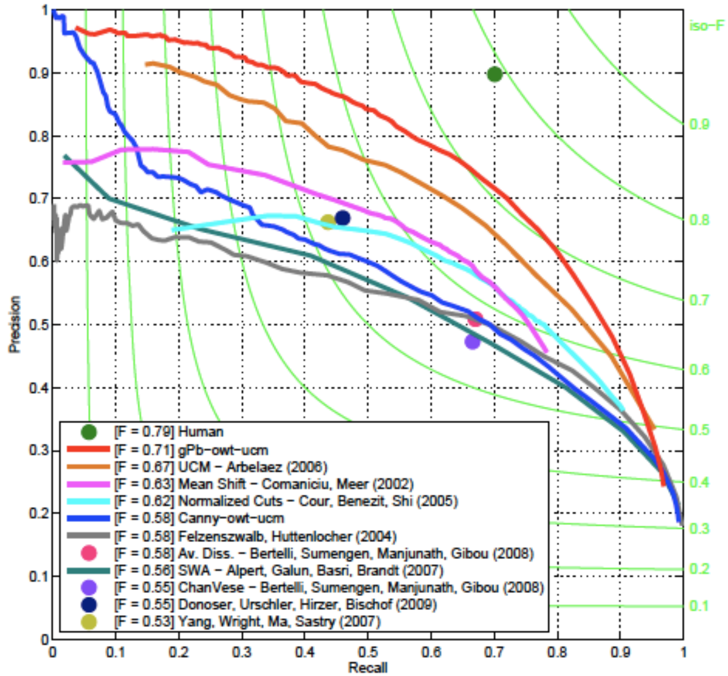
State of the Art in Reorganization

- Interactive segmentation using graph cuts



Rother, Kolmogorov & Blake (2004), Boykov & Jolly (2001), Boykov, Veksler & Zabih(2001)

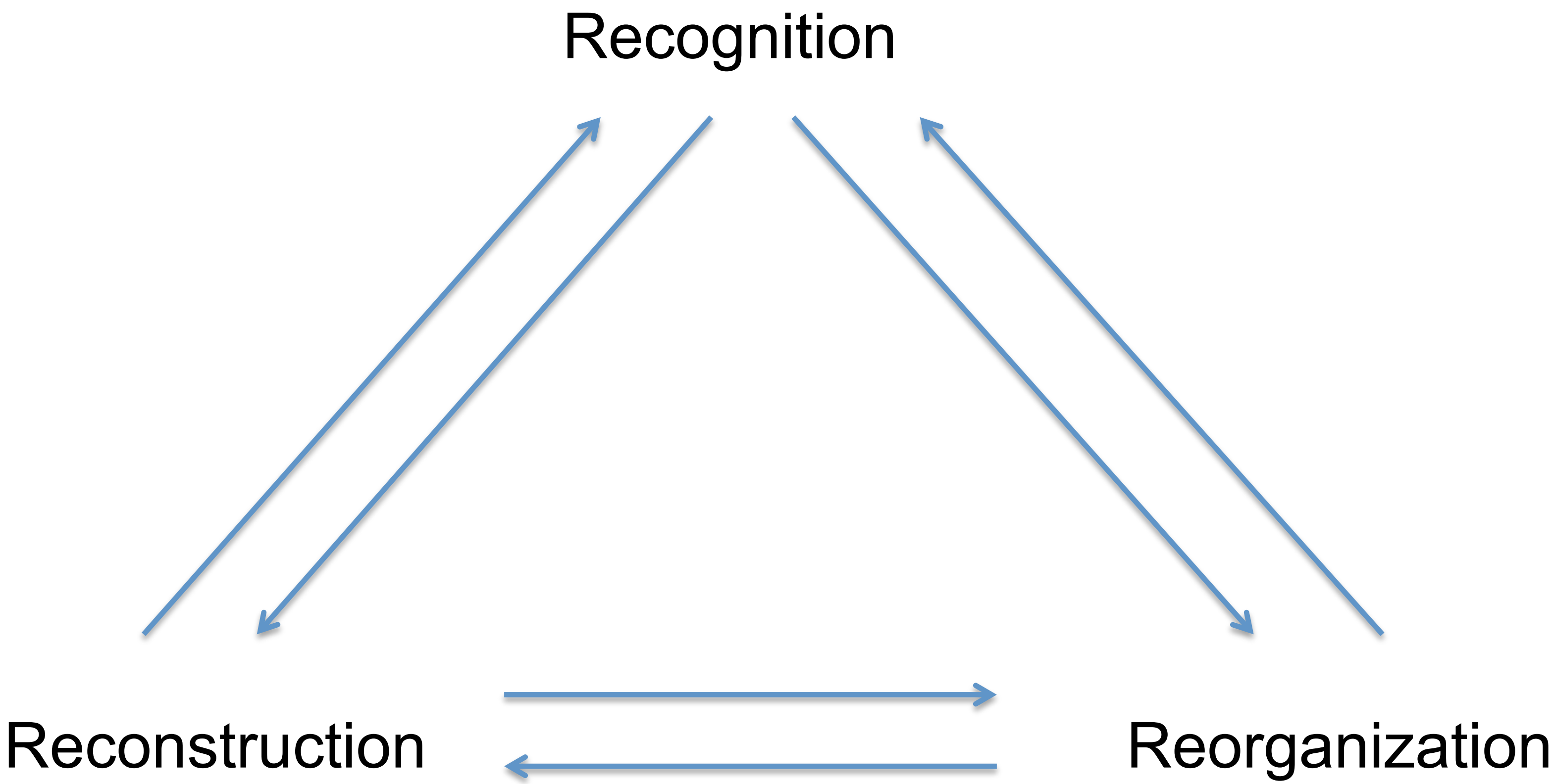
- Berkeley gPb edges & regions



Arbelaez et al (2009), Martin, Fowlkes, Malik (2004), Shi & Malik (2000)

Critique: What is needed is fully automatic semantic segmentation

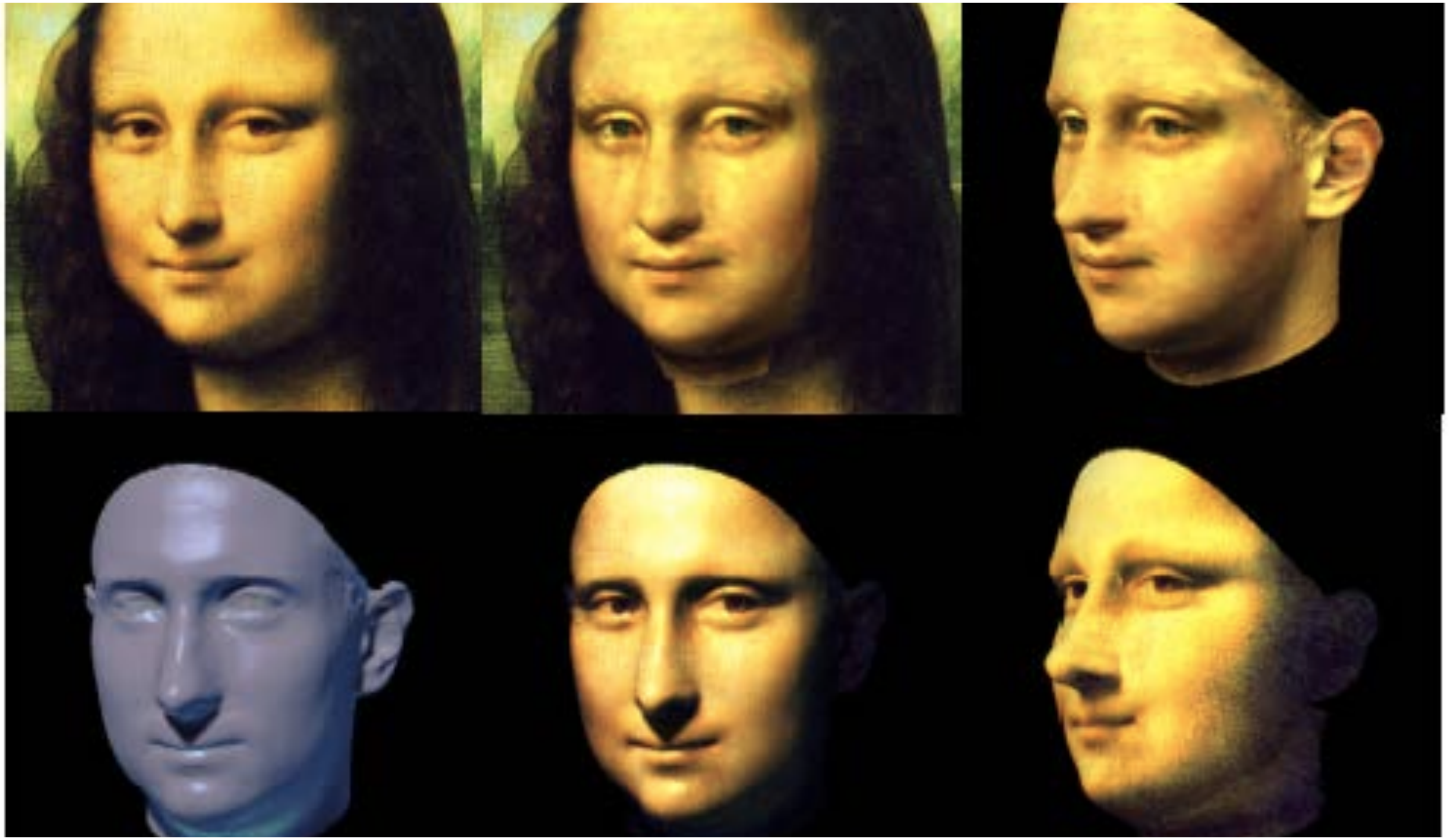
The Three R's of Vision



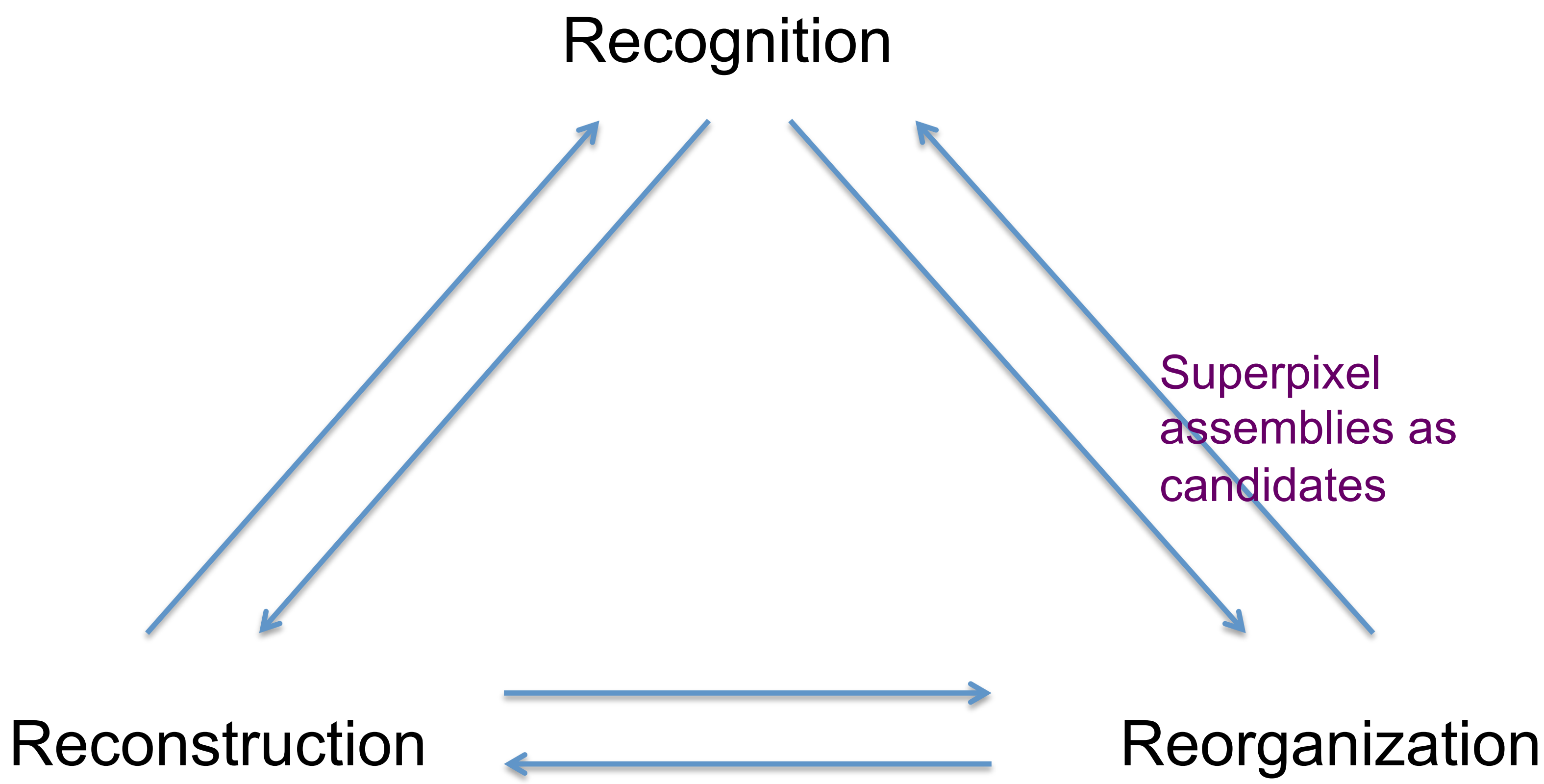
Each of the 6 directed arcs in this diagram is a useful direction of information flow

Recognition Helps Reorganization

Recognition helps reconstruction
Blanz & Vetter (1999)



The Three R’s of Vision

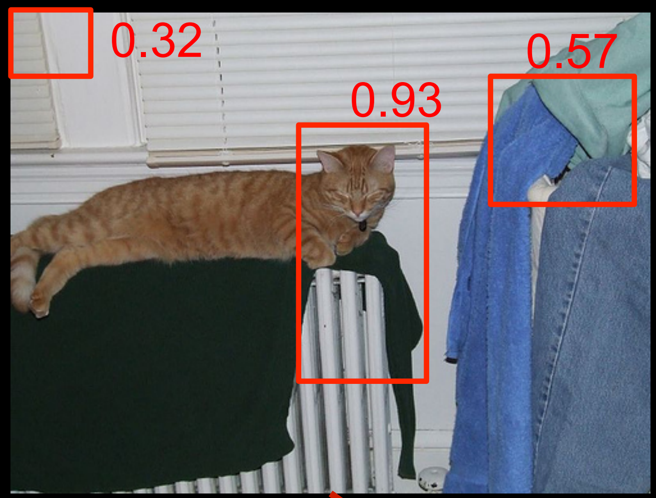


Semantic Segmentation using Regions and Parts

P. Arbeláez, B. Hariharan, S. Gupta,
C. Gu, L. Bourdev and J. Malik

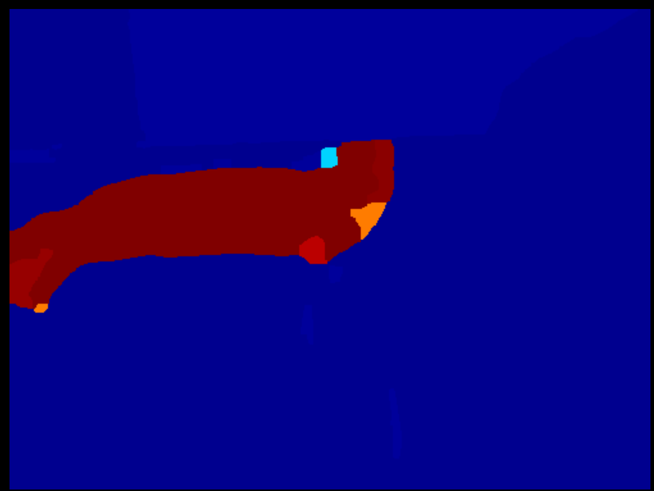
This Work

Top-down Part/Object Detectors



Bottom-up Region Segmentation

Cat Segmenter



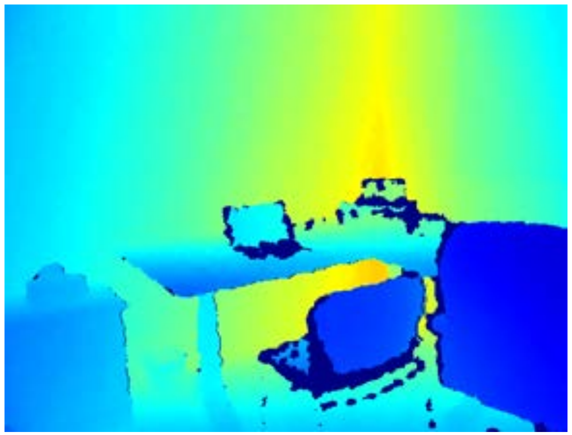
Results on PASCAL VOC

VOC(%)	[18]	[10]	[21]	[5]	SRL	UC3M	TTI	[23]	[9]	FULL	FULL +[14]
plane	51.6	**59.0**	31.0	52.6	38.8	45.9	36.7	49.4	43.8	50.2	48.1
bicycle	25.1	**28.0**	18.8	26.8	21.5	12.3	23.9	23.1	23.7	21.2	20.1
bird	**52.4**	44.0	19.5	37.7	13.6	14.5	20.9	19.2	30.4	38.8	42.2
boat	**35.6**	35.5	23.9	35.4	9.2	22.3	18.8	24.8	22.2	31.4	32.7
bottle	49.6	**50.9**	31.3	34.4	31.1	9.3	41.0	26.1	45.7	39.6	41.9
bus	66.7	**68.0**	53.5	63.3	51.8	46.8	62.7	52.4	56.0	58.9	58.0
car	55.6	53.5	45.3	**61.0**	44.4	38.3	49.0	44.9	51.9	52.1	52.5
cat	44.6	45.6	24.4	32.1	25.7	41.7	21.5	32.9	30.4	**48.1**	45.2
chair	10.6	**15.3**	8.2	11.9	6.7	0.0	8.3	6.5	9.2	7.7	9.2
cow	41.2	40.0	31.0	36.6	26.0	35.9	21.1	35.8	27.7	37.9	**42.2**
table	29.9	28.9	16.4	23.9	12.5	20.7	7.0	22.3	6.9	**30.9**	**37.8**
dog	25.5	33.5	15.8	33.7	12.8	34.1	16.4	25.5	29.6	**36.4**	**36.6**
horse	49.8	**53.1**	27.3	36.8	31.0	34.8	28.2	21.9	42.8	46.9	50.4
mbike	47.9	53.2	48.1	**61.6**	41.9	33.5	42.5	58.1	37.0	52.0	52.6
person	37.2	37.6	31.1	45.0	44.4	24.6	40.5	34.6	47.1	**47.3**	**47.6**
plant	19.3	**35.8**	31.0	26.6	5.7	4.7	19.6	26.8	15.1	24.9	28.7
sheep	45.0	48.5	27.5	40.5	37.5	25.6	33.6	39.9	35.1	**51.9**	**49.0**
sofa	24.4	23.6	19.8	20.4	10.0	13.0	13.3	17.5	23.0	**26.1**	**25.2**
train	37.2	39.3	34.8	**43.8**	33.2	26.8	34.1	38.0	37.7	36.4	41.5
tv	43.3	42.1	26.4	36.4	32.3	26.1	**48.5**	25.3	36.5	40.1	43.8
bgd	83.4	**84.6**	70.1	82.2	80.0	73.4	80.0	77.9	82.2	83.6	84.0
articulat	42.2	43.2	25.2	37.5	27.3	30.2	26.0	30.0	34.7	**43.9**	**44.8**
transp	45.7	48.1	36.5	**49.2**	34.4	32.3	38.2	41.5	38.9	43.2	43.7
indoors	29.5	**32.8**	22.2	25.6	16.4	12.3	23.0	20.8	22.7	28.2	31.1
mean	41.7	**43.8**	30.2	40.1	29.1	27.8	31.8	33.5	35.0	41.1	42.4

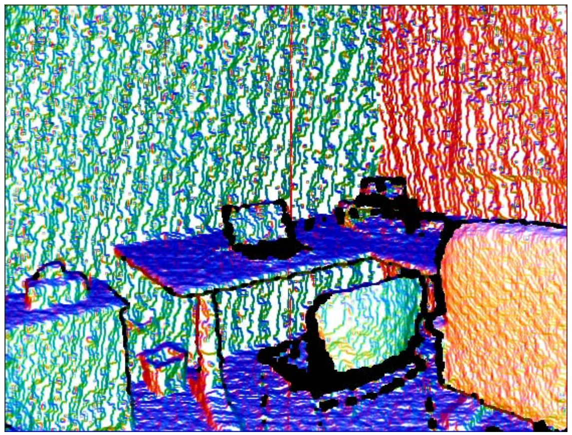
Reconstruction

Kinect - Active depth sensor based on triangulation

Color Image



Depth Image



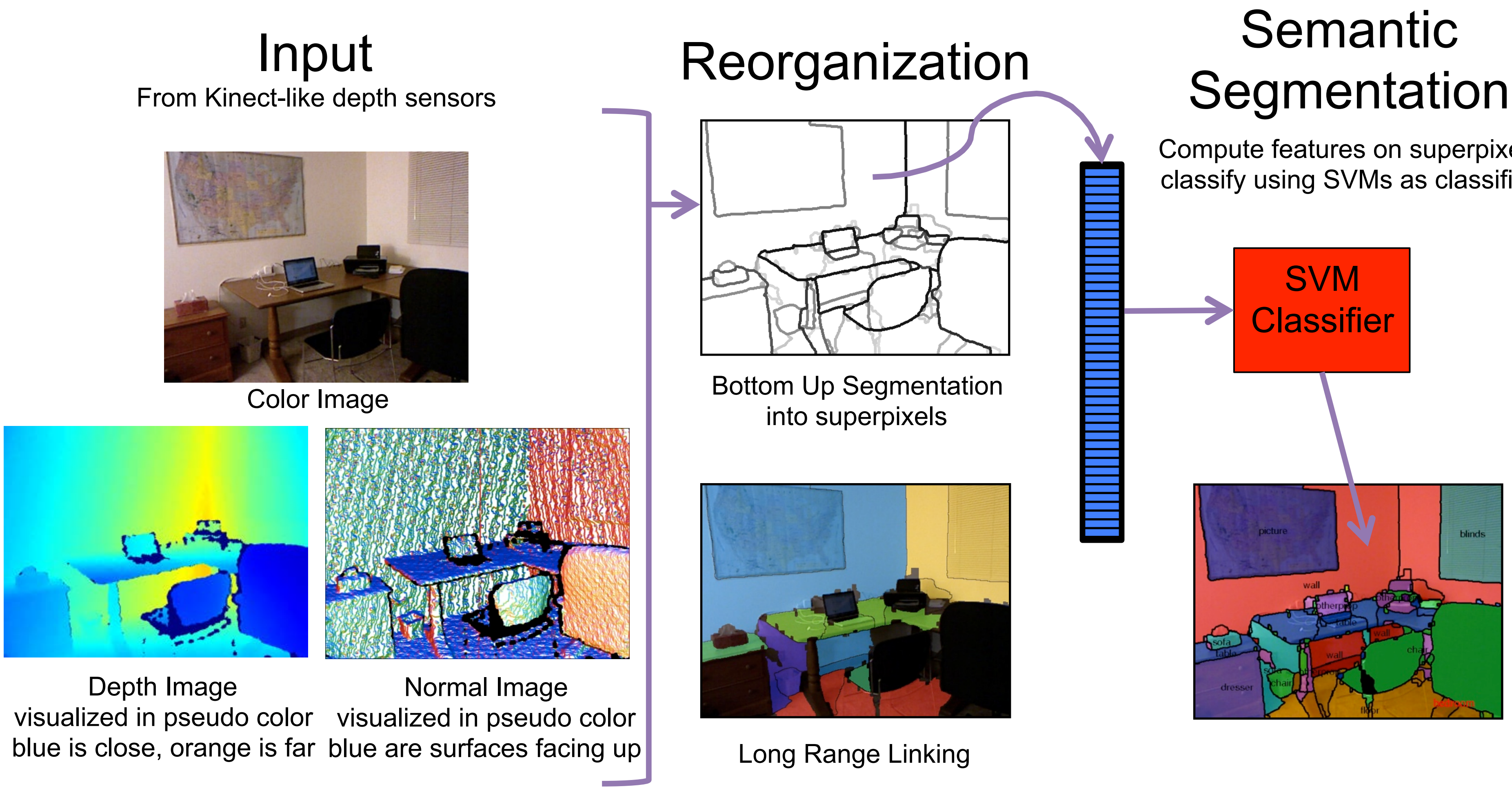
Normal Image

Perceptual Robotics

Using RGBD images to semantically parse scenes

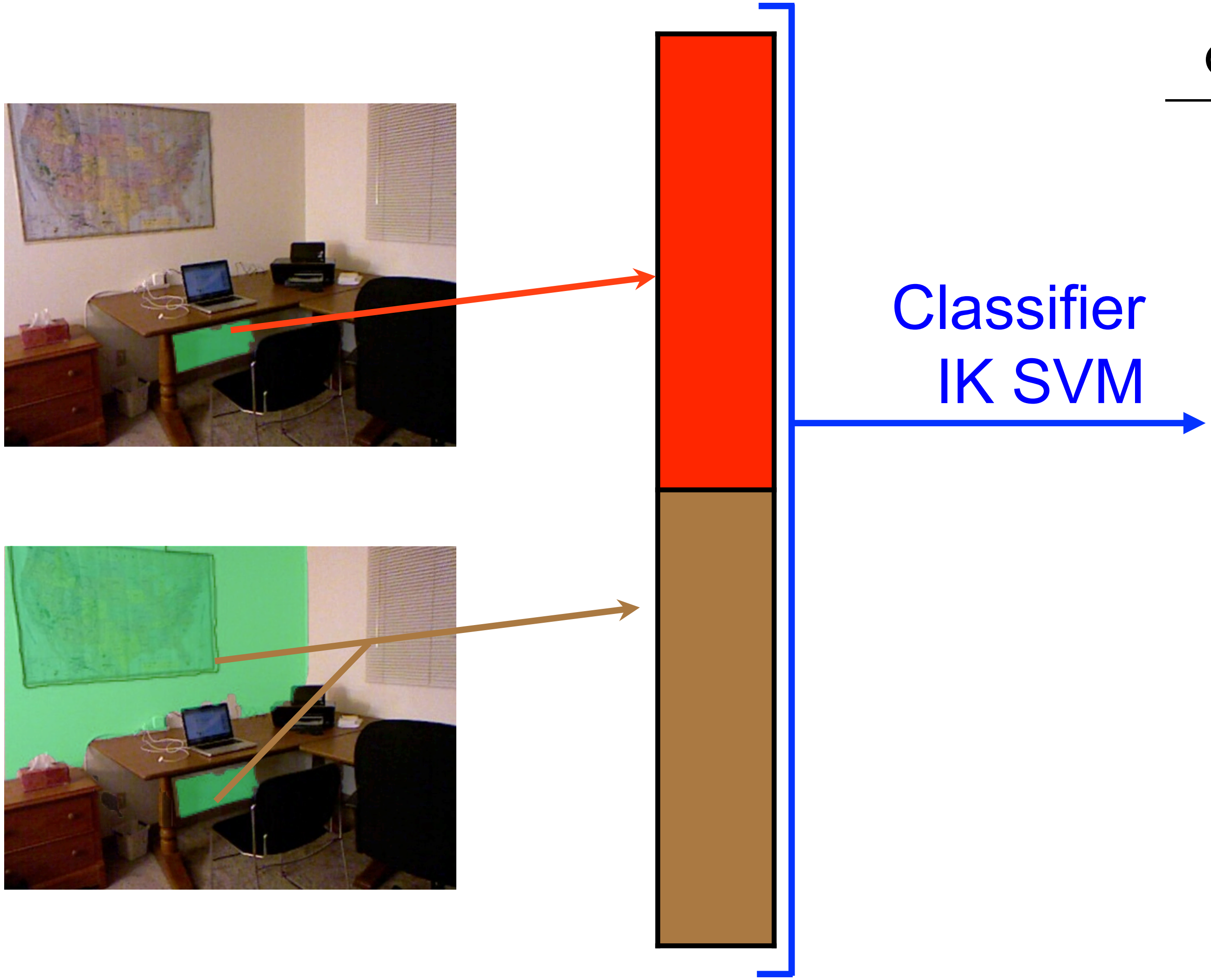
- S. Gupta, P. Arbeláez & J. Malik (CVPR 2013)

Using RGBD Images to Semantically Parse Scenes



Semantic Segmentation

Super Pixel Classification



Category	Pr
wall	0.90
cabinet	0.05
window	0.05
chair	0.0
table	0.0
-	
-	

Semantic Segmentation

Affordance Based Features

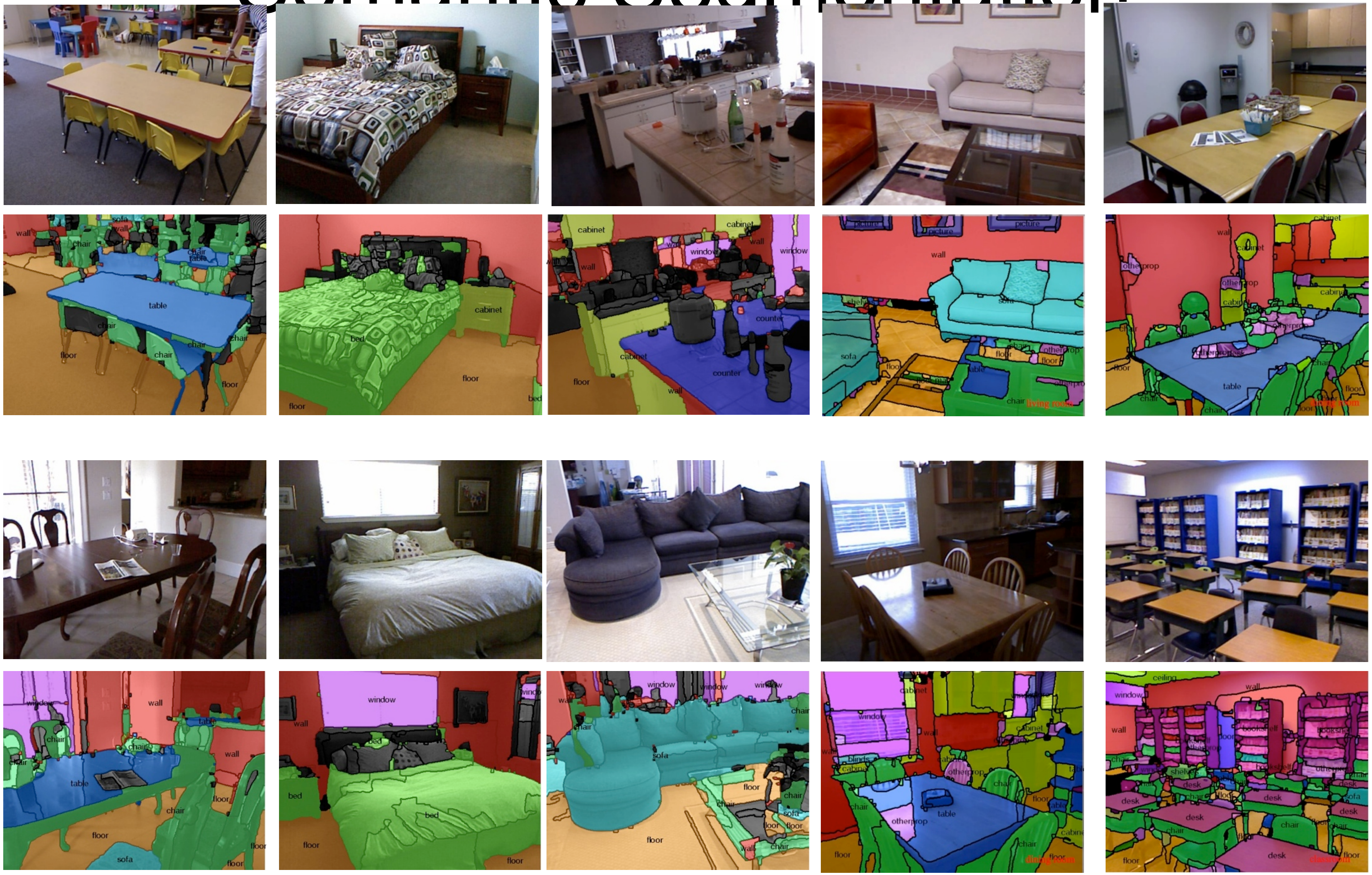
- Geocentric Pose
  - Orientation Features
  - Height above ground
- Size Features
  - Spatial extent
  - Surface Area
  - Is clipped/occluded
- Shape Features
  - Planarity
  - Strength of local geometric gradients

Use orientation with respect to gravity,
heights above ground,
actual sizes

Category Specific Features

- Scores of one-versus-rest SVMs using histogram of
  - Vector Quantized SIFT
  - Geocentric Textons

Semantic Segmentation



Semantic Segmentation

Aggregate Performance

[NYU]	Our
35.26	42.04

Category wise performance

	[NYU]	Our
wall	55.25	62.2
floor	73.08	75.9
cabinet	31.4	44.5
bed	38.87	49.4
chair	28.94	37.9
sofa	24.52	39.3
table	20.13	31.2
door	5.59	10.4
window	26.35	32.4
bookshelf	20.6	19

	[NYU]	Our
picture	34.31	39.5
counter	32.03	47.4
blinds	39.01	42.1
desk	4.52	9.4
shelves	3.07	3.3
curtain	26.43	32
dresser	13.08	19.9
pillow	18.34	27.1
mirror	4.08	18.9
floor mat	7.11	20.8

NYU [Silberman et al ECCV12] Indoor segmentation and support inference from RGBD images.

Semantic Segmentation

Performance – some more categories

	[NYU]	Our
clothes	6.27	8.5
ceiling	62.99	58.3
books	5.34	3.4
refrigerator	1.28	17.3
television	5.66	19.1
paper	12.6	12.5
towel	0.11	8
shower curtain	3.55	15
box	0.12	3.3
whiteboard	0	31.2

	[NYU]	Our
person	6.35	16.7
night stand	5.95	29
toilet	26.49	39.4
sink	24.66	25.2
lamp	14.99	23.5
bathtub	0	20.5
bag	0	0.1
otherstructure	5.75	2.6
otherfurniture	3.66	19.8
otherprop	20.29	25.5

[NYU] Silberman et al, ECCV12, Indoor segmentation and support inference from RGBD images.

Thank You

The Three R's of Vision