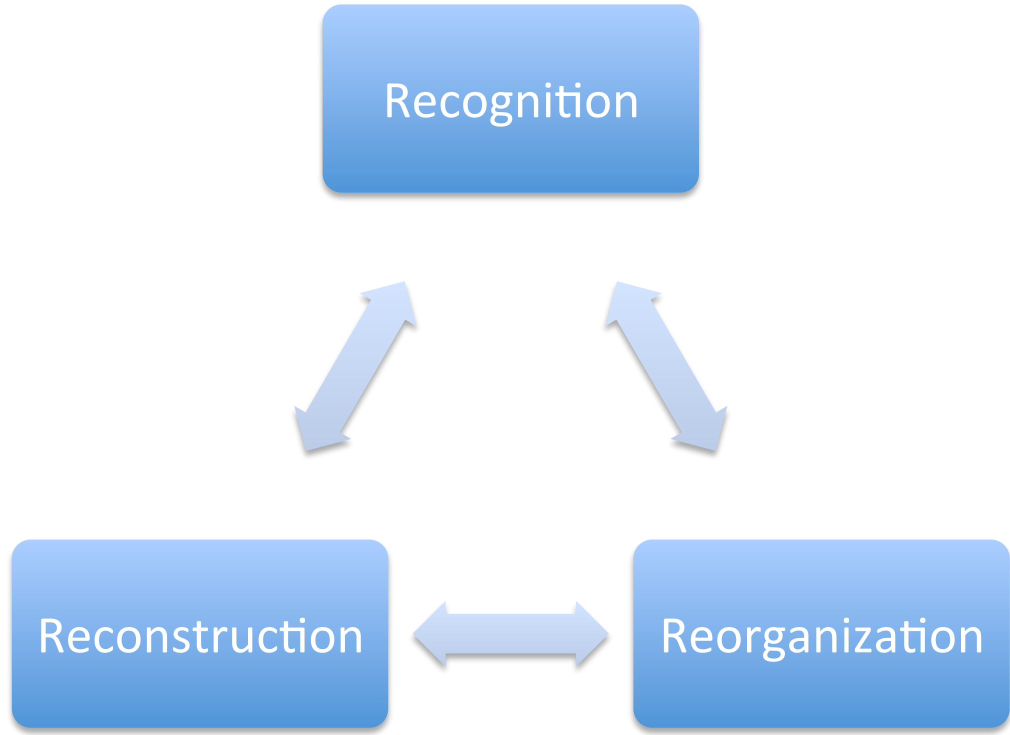
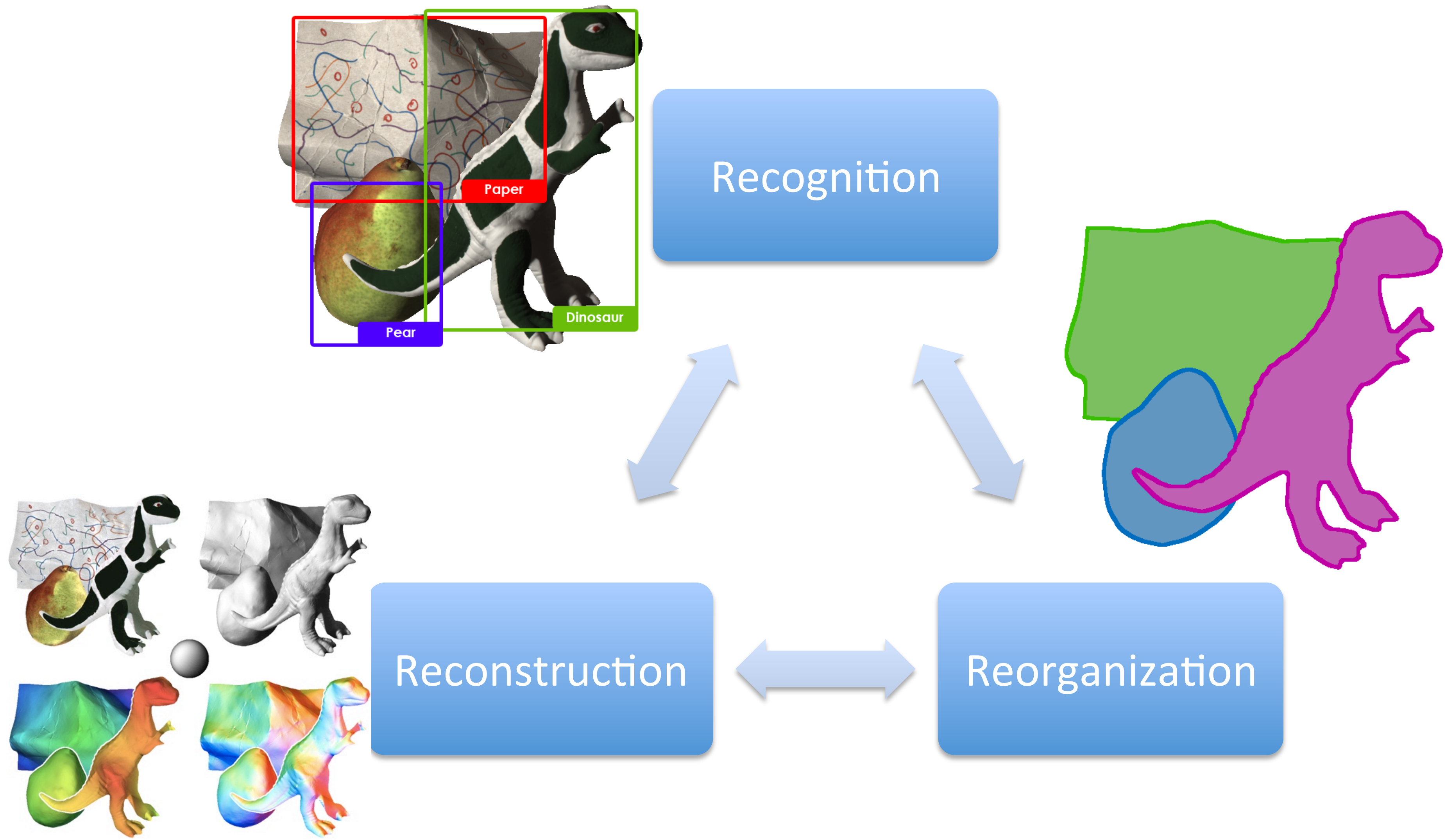
The Three R's of Vision



Jitendra Malik
UC Berkeley

Recognition, Reconstruction & Reorganization



Theories of Visual Perception in the 20th century

- Behaviorism emphasized stimulus generalization and association. Aligns well with machine learning approaches to recognition.
- Gestaltists emphasized perceptual organization- grouping and figure/ground phenomena. Natural home for those who regard reorganization of the stimulus – from pixels to entities- as primary.
- Gibson's ecological optics emphasized "information pickup" by a moving observer. Introduced optic flow and texture gradients as powerful 3d cues. Consistent with a view that there is enough information for 3d reconstruction of the world.

Marr's paradigm (1980)

Image

↓

Primal Sketch

↓

2.5D Sketch

↓

Part-based Models using Generalized Cylinders

Ignored texture, grouping factors

Shape-from-X didn't work for monocular cues

Generalized cylinders not rich enough as primitives

Overall approach violated the principle of least commitment, that Marr had himself advocated. Didn't use probabilistic inference or learning.

Computer vision since 1990…

- Significant progress **<u>without</u>** an overarching theory
- Has made considerable use of models drawn from
  - Geometry
  - Statistics/Machine learning
  - Optimization

Review

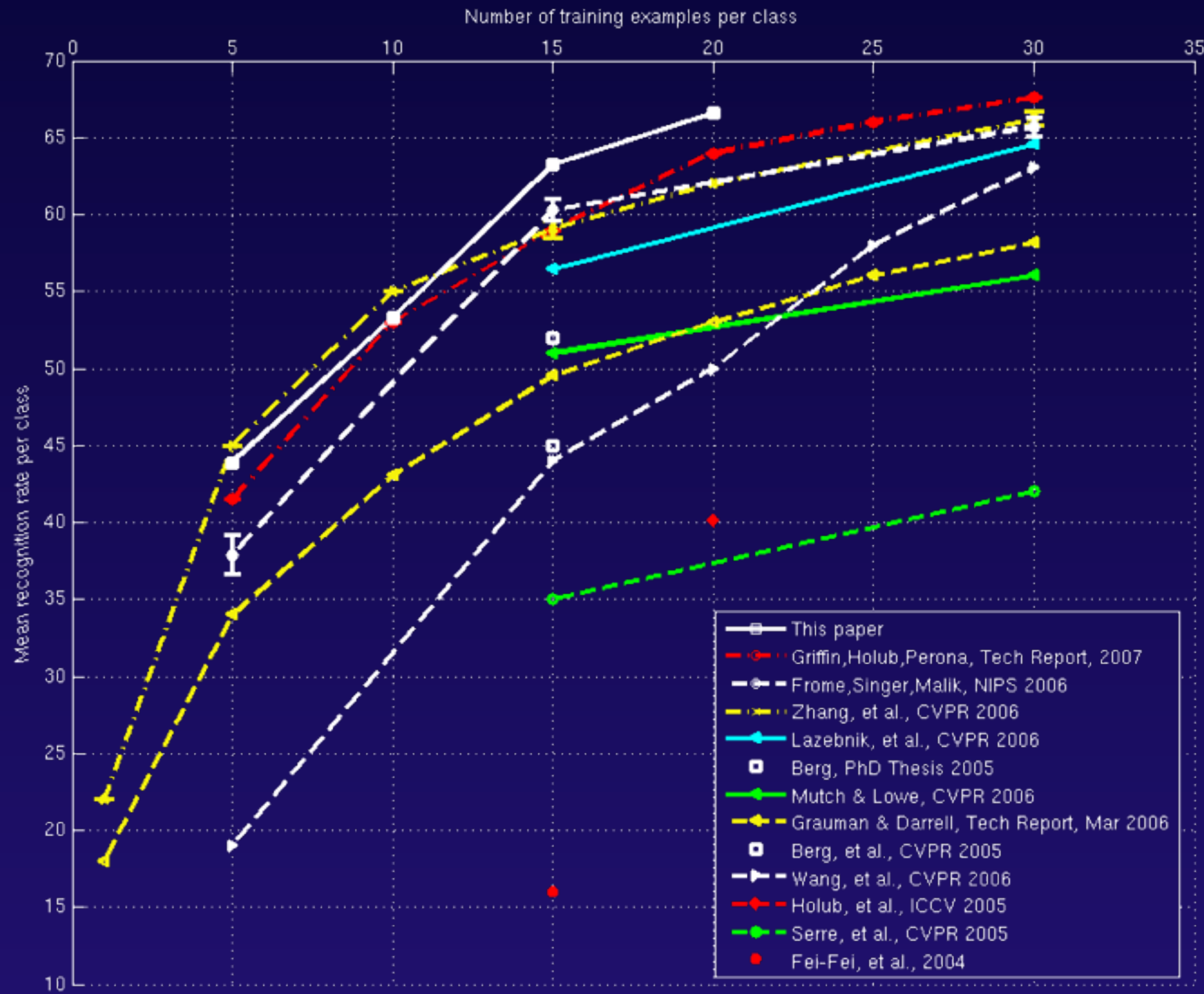
- Recognition
  - 2D problems such as handwriting recognition, face detection
  - Partial progress on 3d object category recognition
- Reconstruction
  - Feature matching + multiple view geometry has led to city scale point cloud reconstructions
- Reorganization
  - Graphcuts for interactive segmentation
  - Bottom up boundaries and regions/superpixels

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

- 102 classes, 31-300 images/class

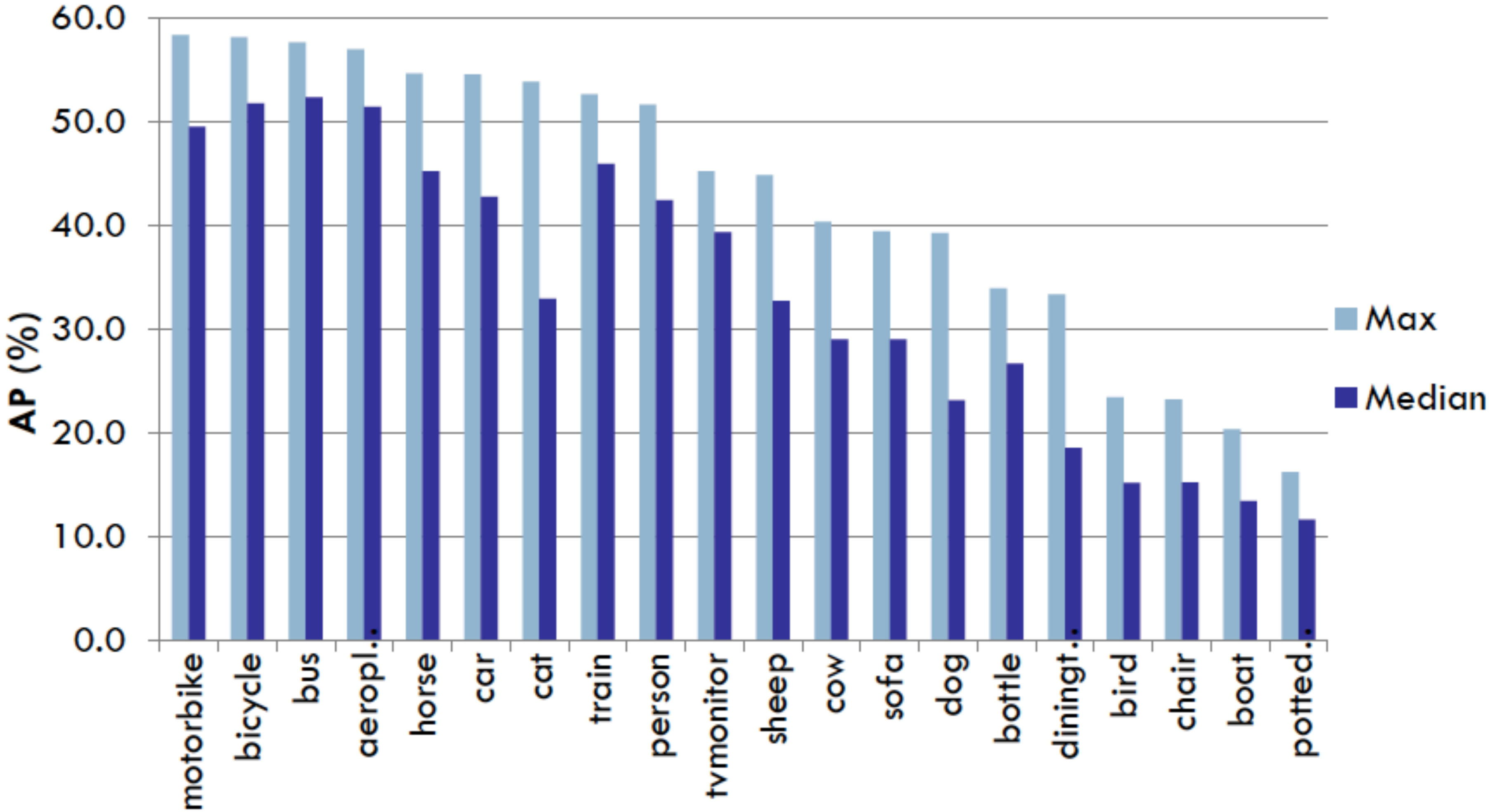
Caltech 101 classification results

(even better by combining cues..)



PASCAL Visual Object Challenge (Everingham et al)

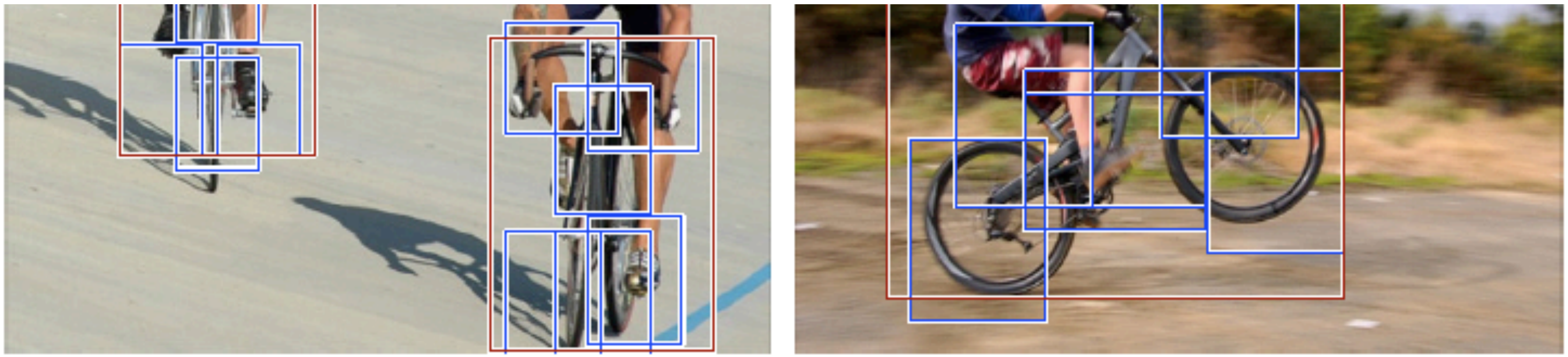
AP by Class

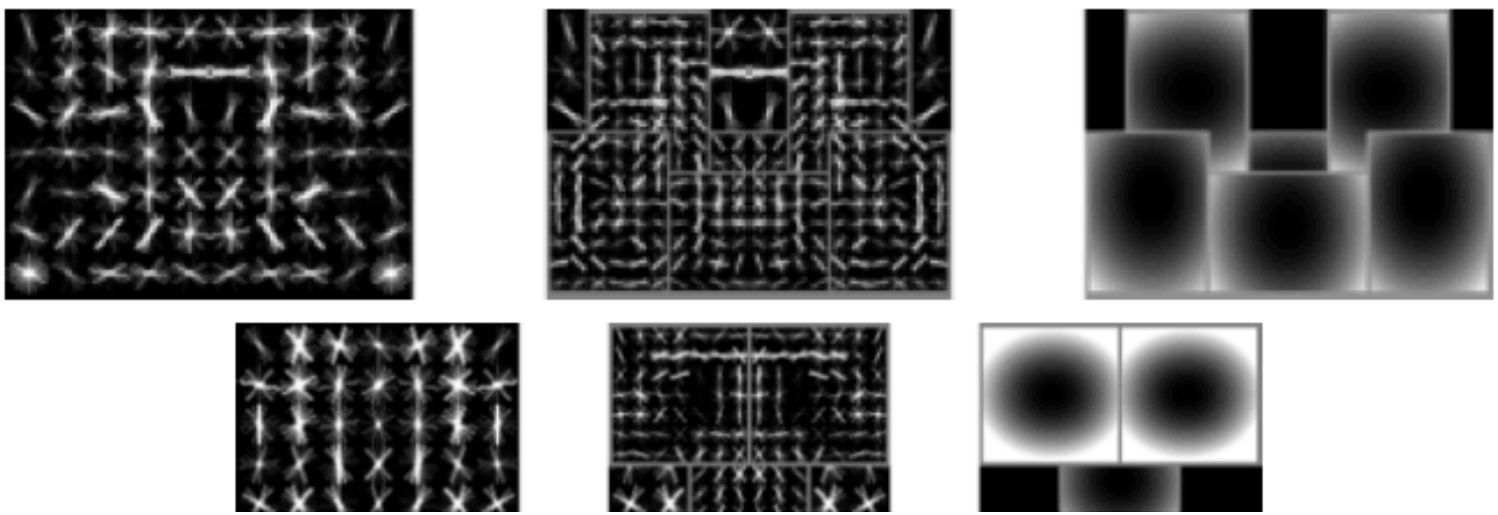


- Max AP: 58.3% (motorbike) ... 16.2% (potted plant)

Object Detection with Discriminatively Trained Part Based Models

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





Builds on Dalal & Triggs HOG detector (2005)

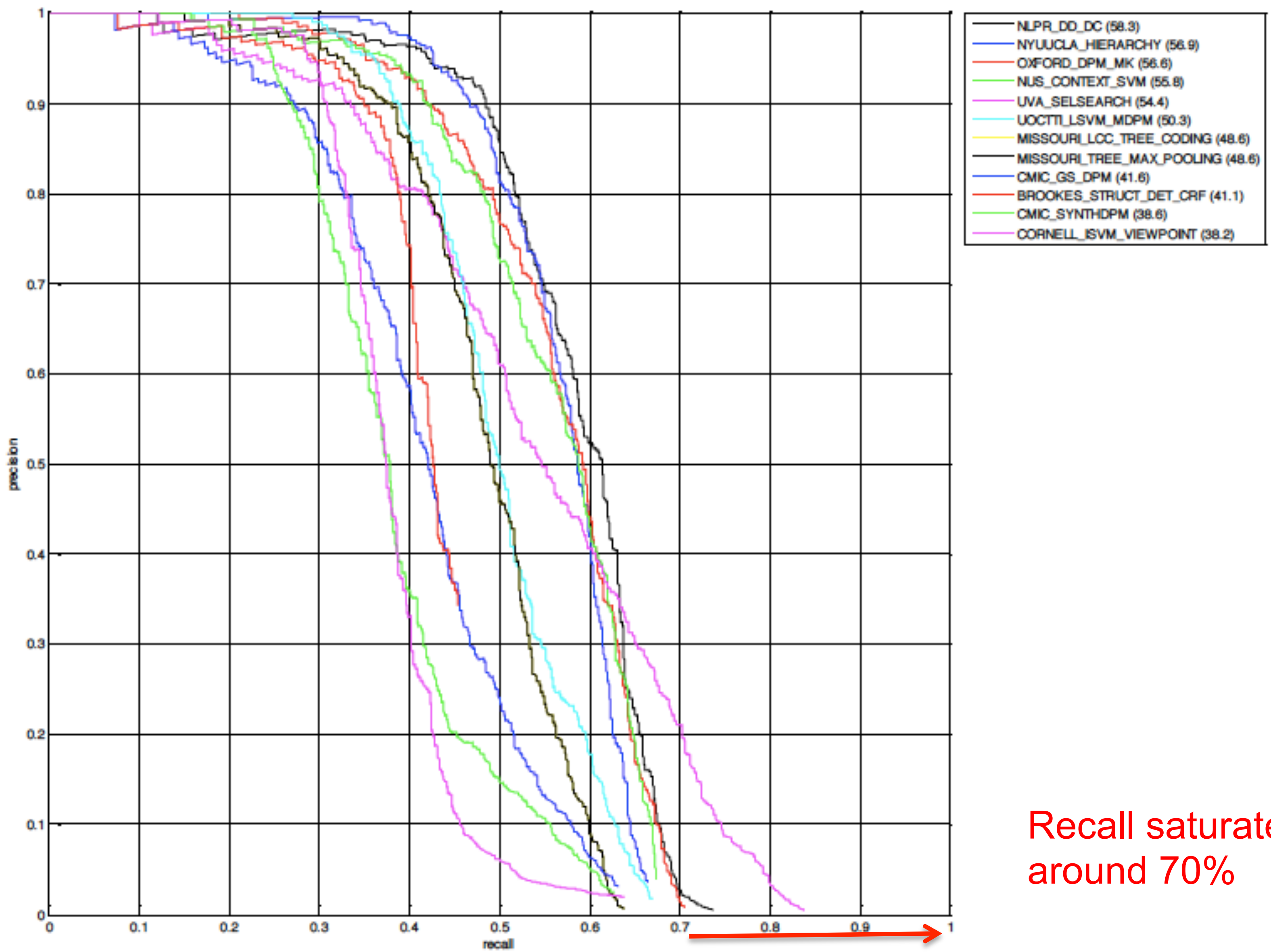
What did we learn from these datasets?

- Lazebnik, Schmid & Ponce's approach – Spatial Pyramid Matching - was validated by Caltech 101.
- Felzenszwalb et al's approach – Deformable Part Models - was validated by the PASCAL VOC challenge.
- There were other interesting and well-performing approaches that came up in these competitions. These two are noteworthy for their combination of (relative) simplicity combined with good performance.

Critique of the State of the Art

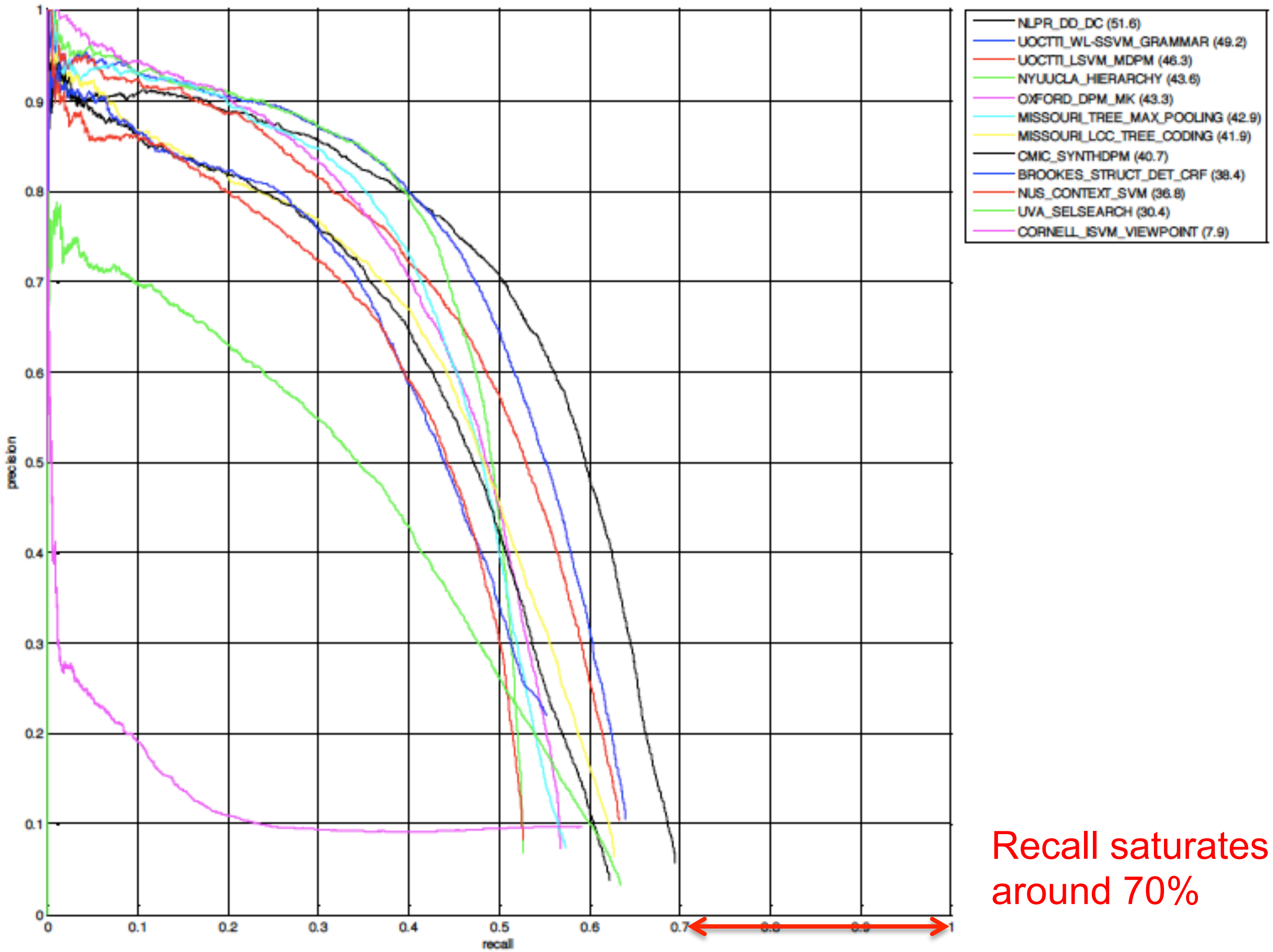
- Performance is quite poor compared to that at 2d recognition tasks and the needs of many applications.
- Pose Estimation / Localization of parts or keypoints is even worse. We can't isolate decent stick figures from radiance images, making use of depth data necessary.
- Progress has slowed down. Variations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

Precision/Recall - Motorbike



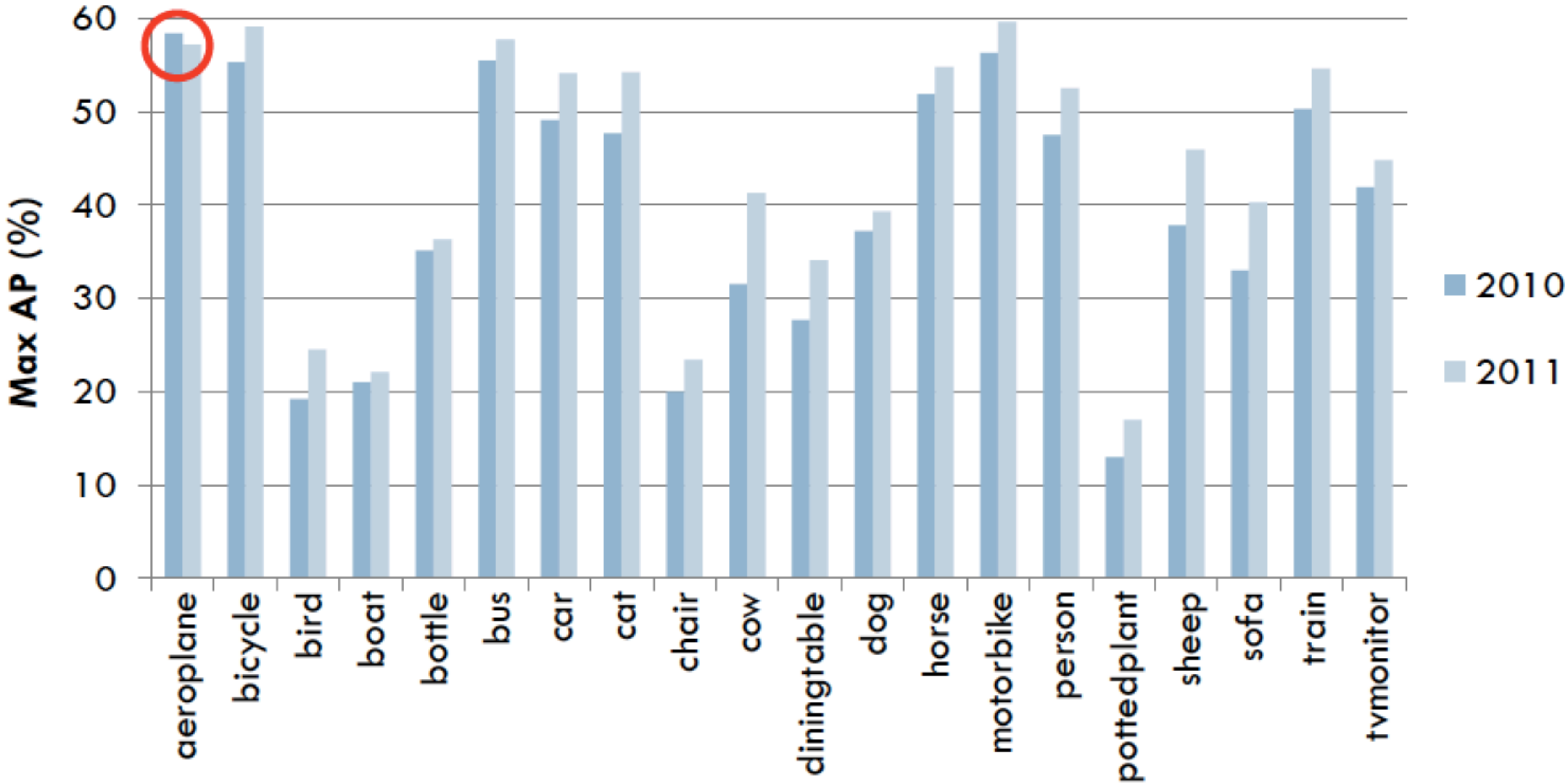
Recall saturates around 70%

Precision/Recall - Person



Recall saturates around 70%

Progress 2010-2011



- Results on 2010 data improve for best 2011 methods for all but one category (aeroplane)
  - Caveats: More training data + re-use of test data

Some categories are visually incoherent

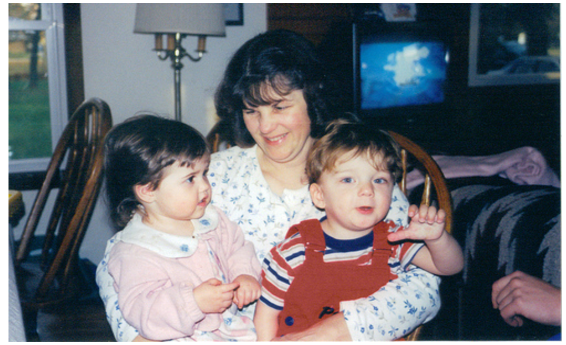
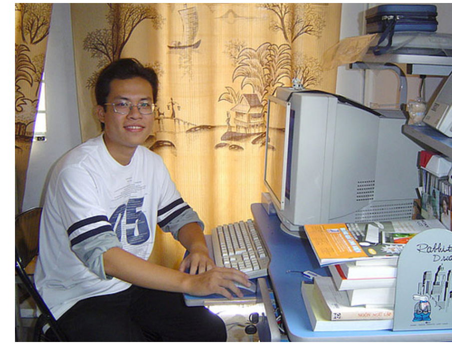







AP=0.23

We are not going to find chairs with HOG templates!




State of the Art in Reconstruction

- Multiple photographs

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

Agarwal et al (2010)

- Range Sensors

Kinect (PrimeSense)

Velodyne Lidar

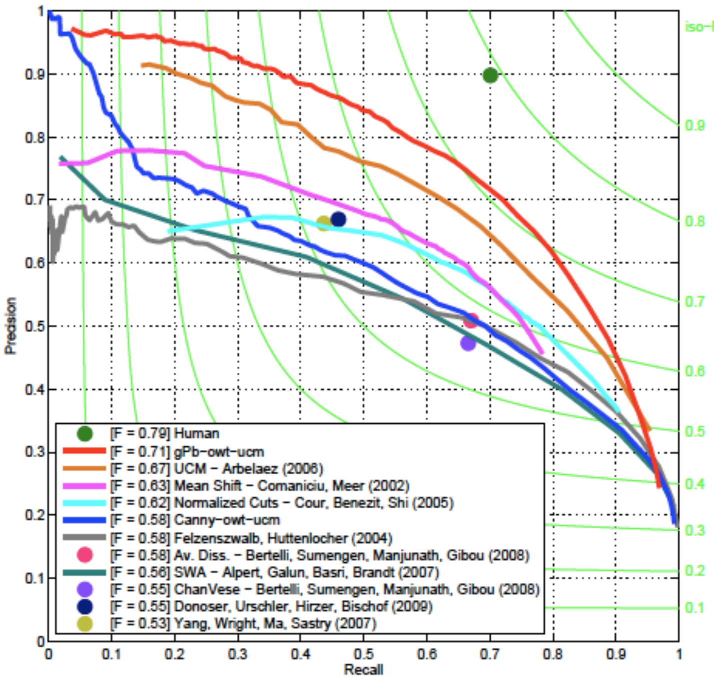
Critique: Semantic Segmentation is needed to make this more useful…

State of the Art in Reorganization

- Interactive segmentation using graph cuts

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

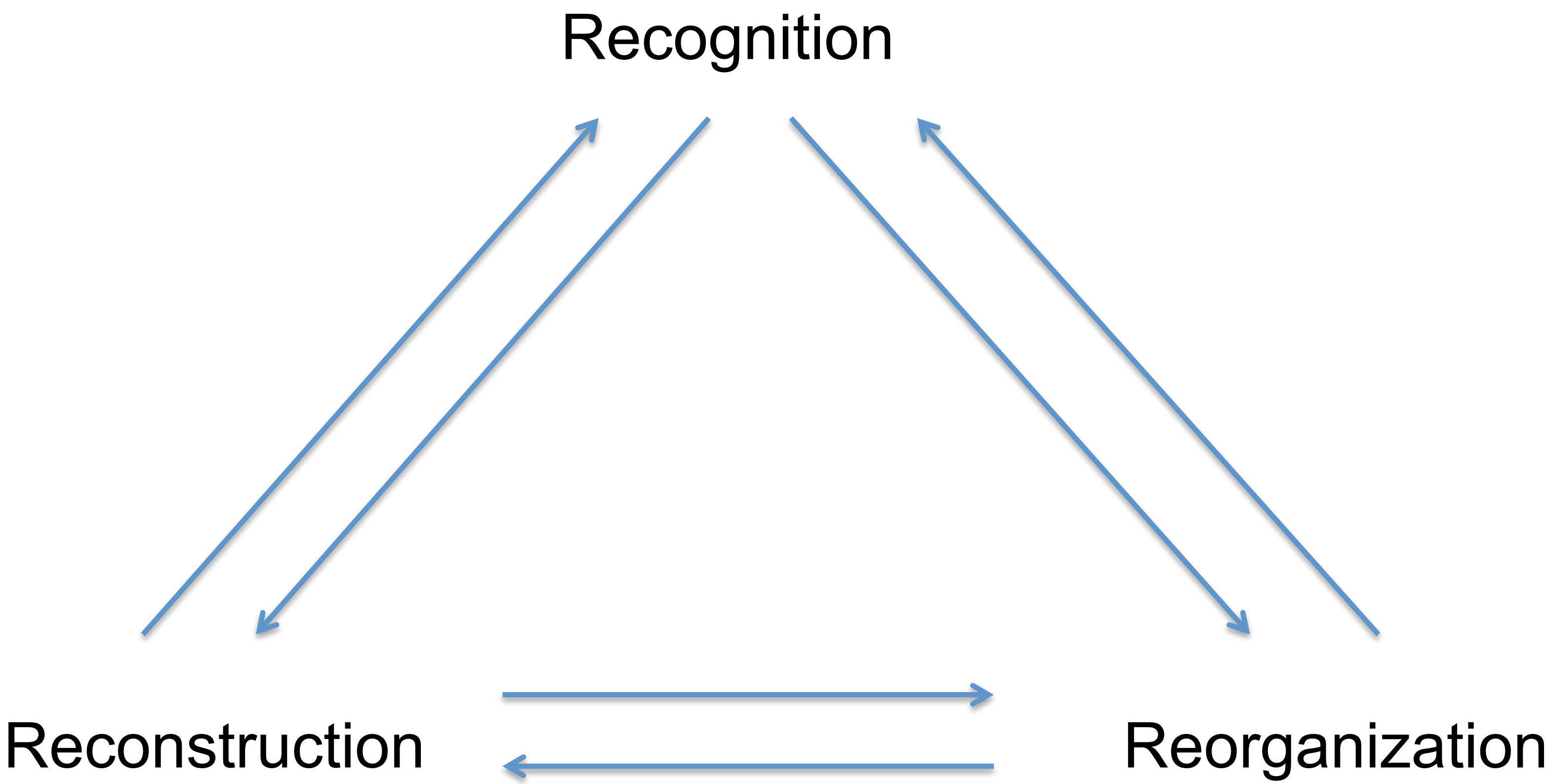
- Berkeley gPb edges & regions



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

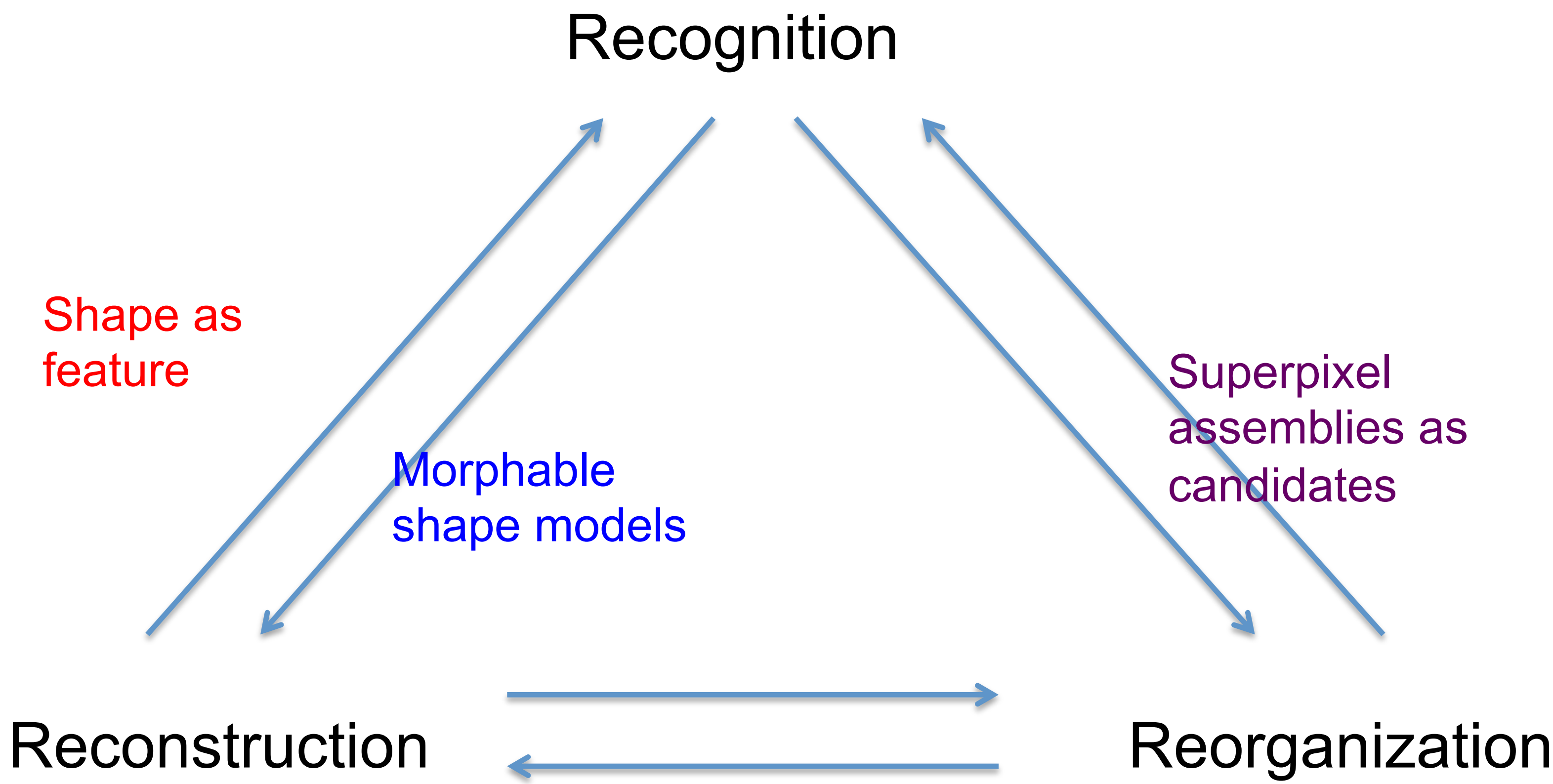


Each of the 6 directed arcs in this diagram is a useful direction of information flow

Theory vs. Models

- Evolution is a theory; structure of DNA is a model. Models have limited scope & are easily testable. Theory is less precise but broader in scope.
- The value of this "theory" is
  - Conceptual framework that points to most fruitful research directions in vision
  - Pedagogic value for students
  - Someday, there may be a grand reunification, such as what Maxwell brought to electromagnetism (we may dream, can't we?)

The Three R's of Vision



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.

Next steps in recognition

- Incorporate the “shape bias” known from child development literature

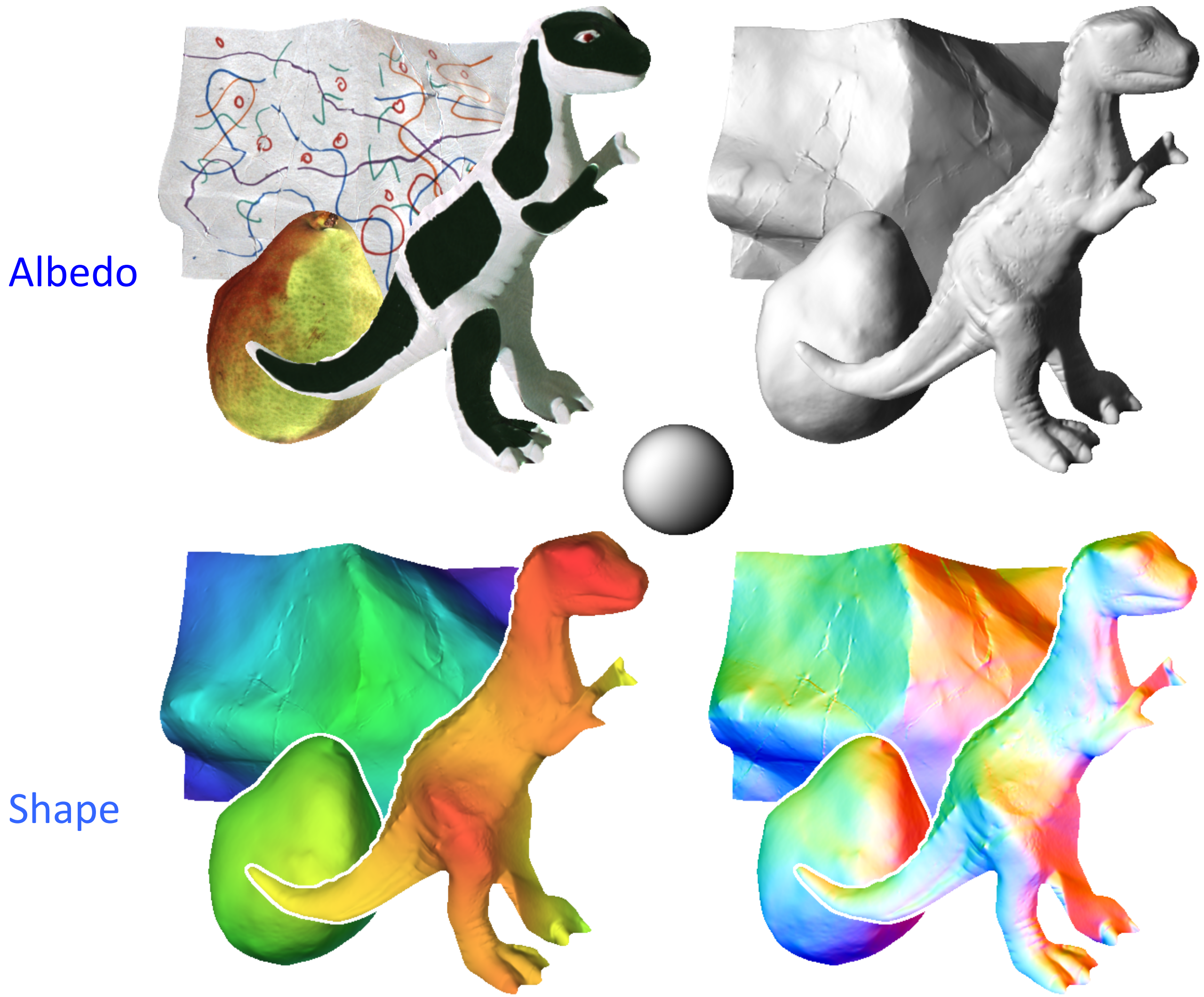
  Barron & Malik, CVPR 2012

  - 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 Poselets: Bourdev & Malik, 2009 & later
- Recognition and figure-ground inference need to co-evolve. Occlusion is signal, not noise.

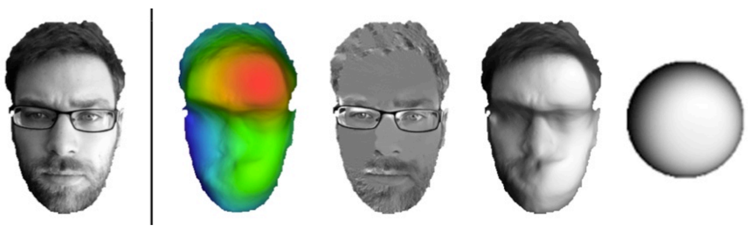
Arbelaez et al, CVPR 2012

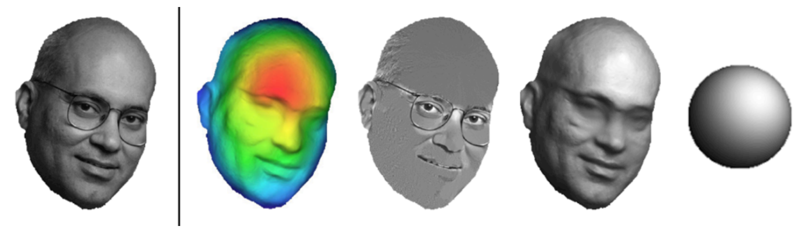
Reconstruction

Shape, Albedo & Illumination



Shape, Albedo, and Illumination from Shading



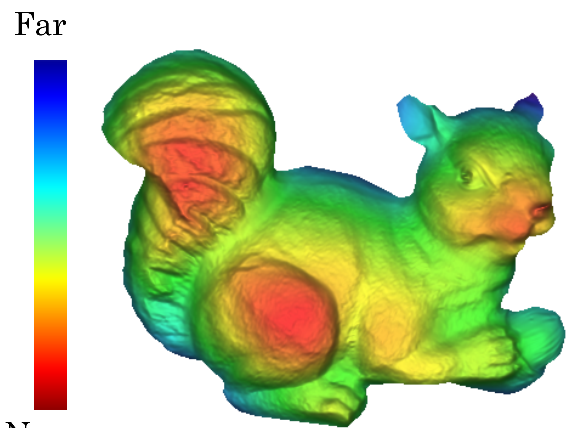


Jonathan Barron

Jitendra Malik

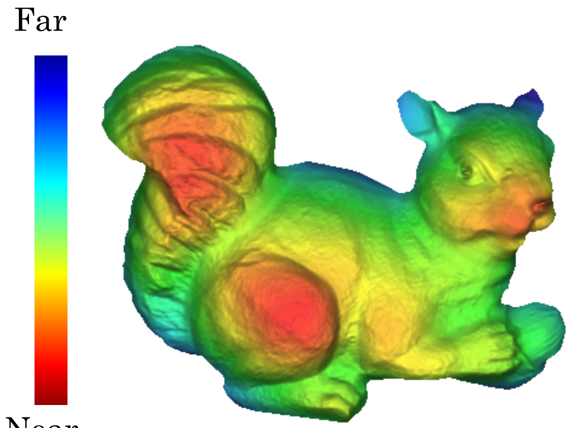
UC Berkeley

Forward Optics

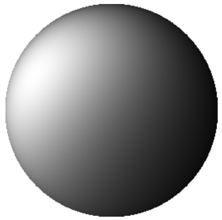


Z
shape / depth

Forward Optics

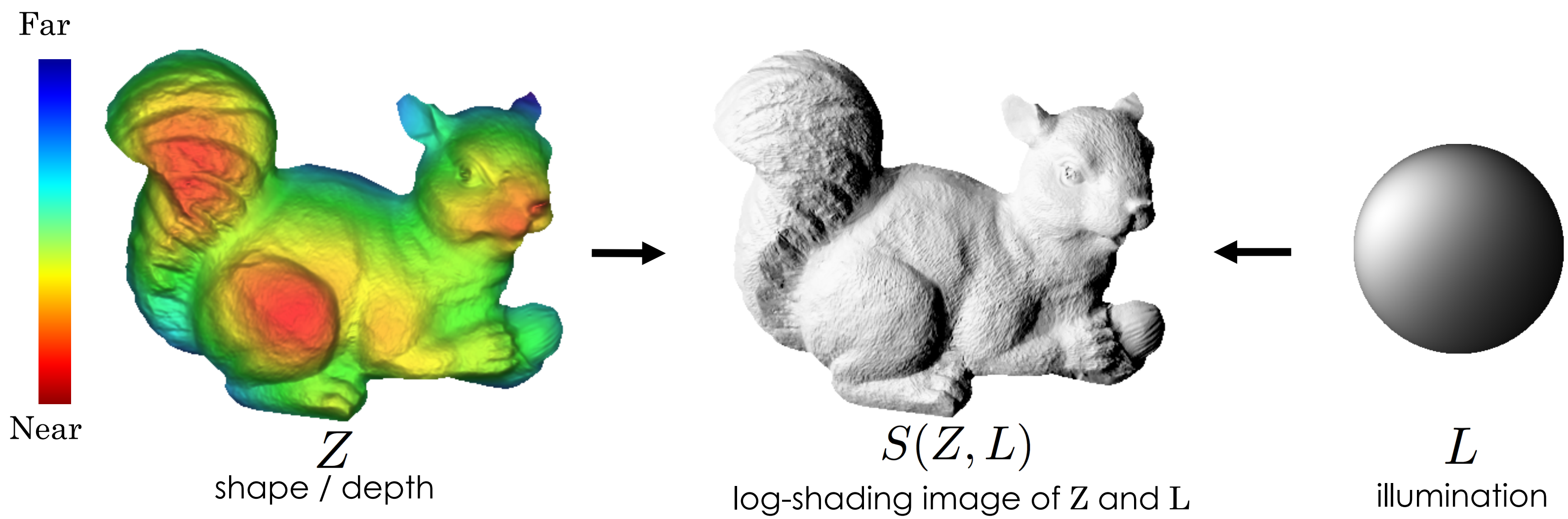


Z
shape / depth

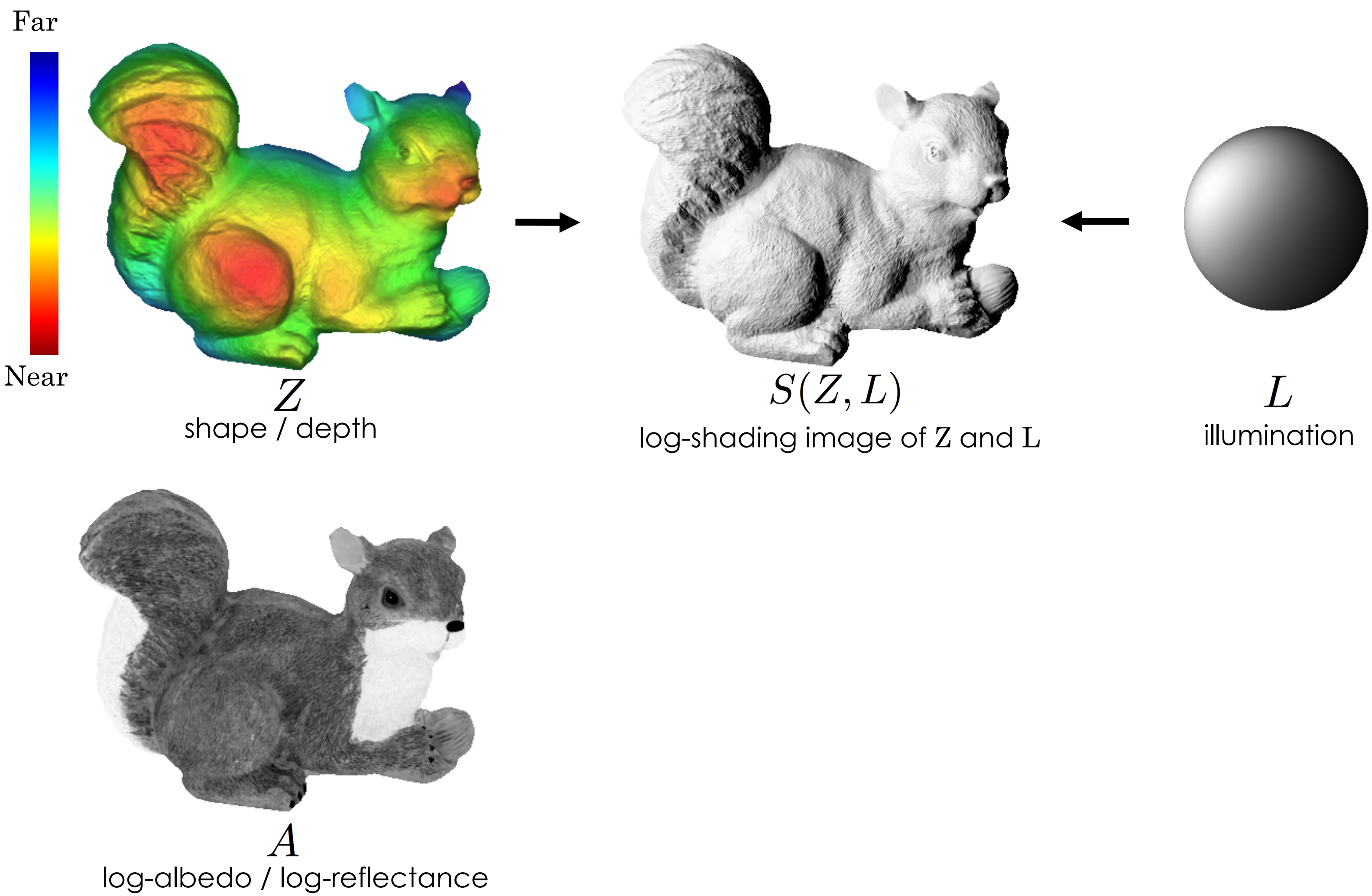


L
illumination

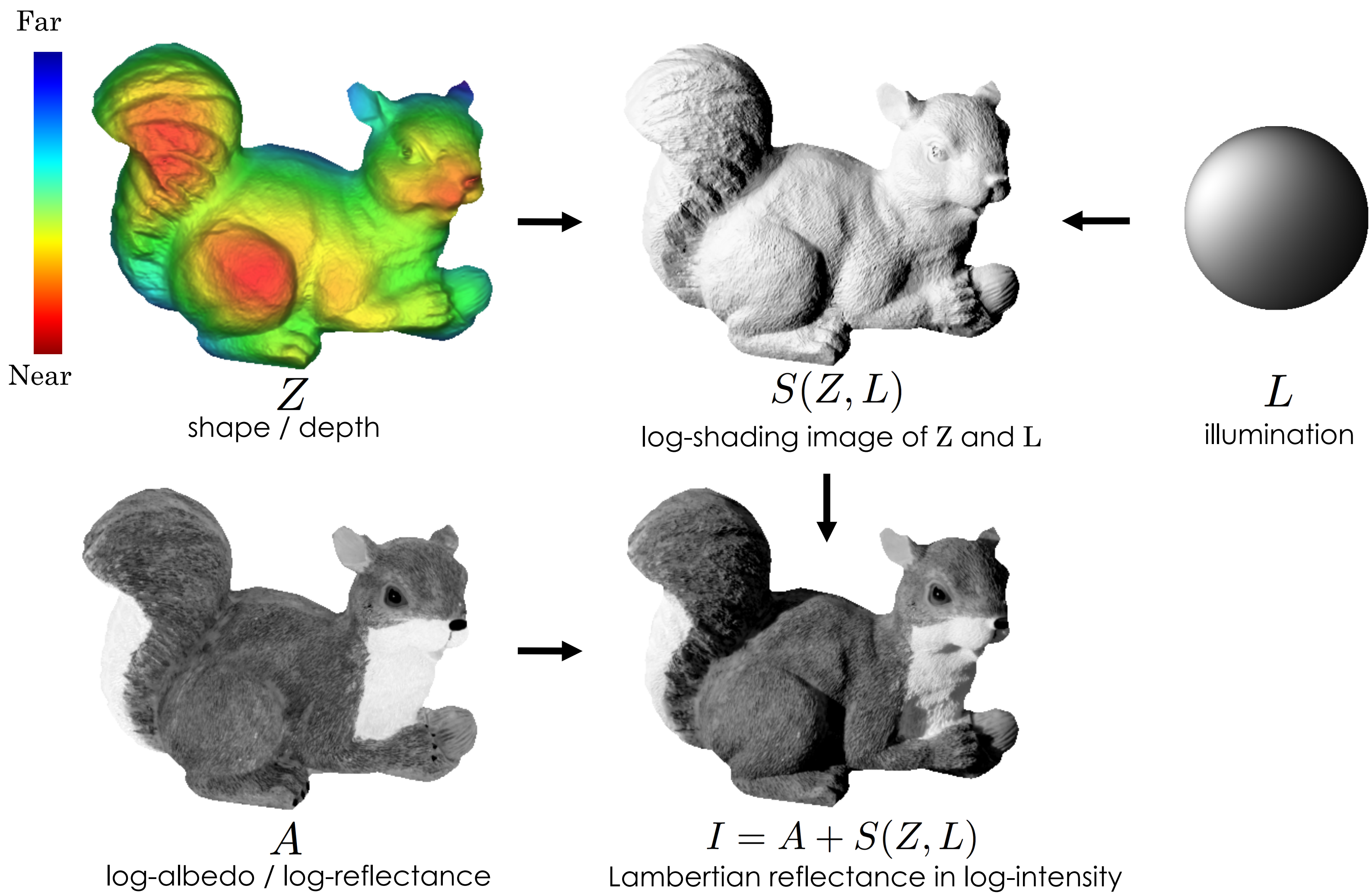
Forward Optics



Forward Optics

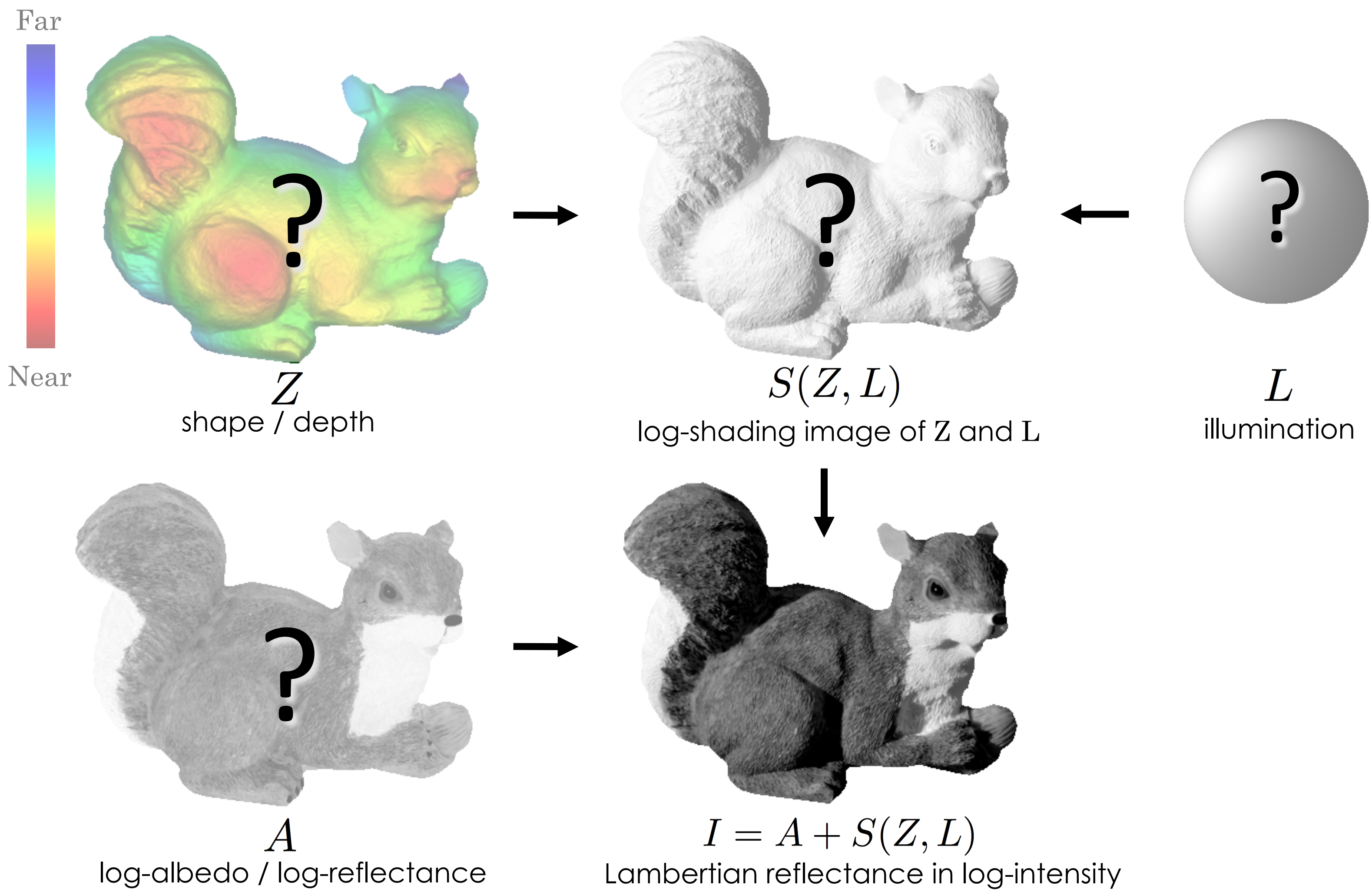


Forward Optics



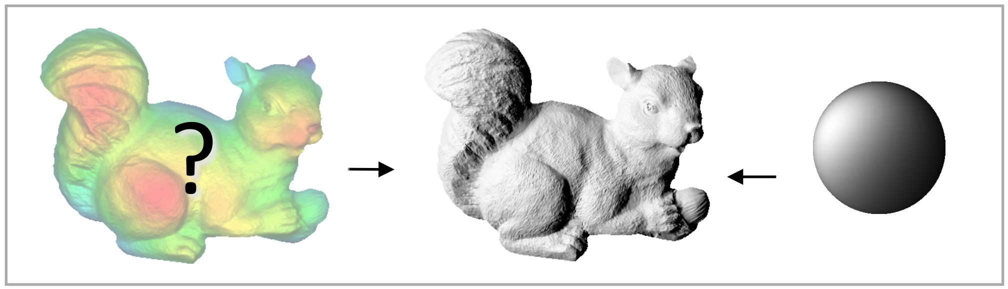
Shape, Albedo, and Illumination from Shading

SAIFS ("safes")



Past Work

Shape from Shading



Assume illumination and albedo are known, and solve for the shape

Intrinsic Images

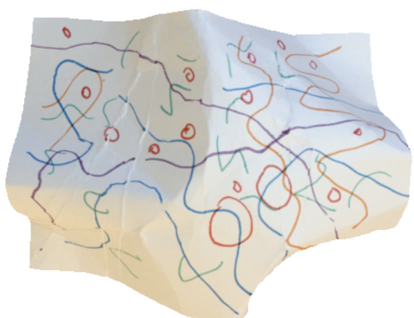
Ignore shape and illumination, and classify edges as either shading or albedo

Problem Formulation

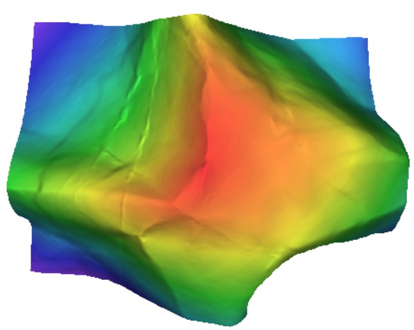
$$\begin{aligned} \underset{Z,A}{\text{maximize}} \quad & P(A|Z,L)P(Z) \\ \text{subject to} \quad & I = A + S(Z,L) \end{aligned}$$

Given a single image, search for the most likely *explanation*
(shape, albedo, and illumination) that together exactly reproduces that image.

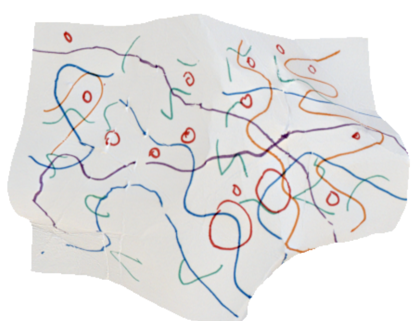
Input:



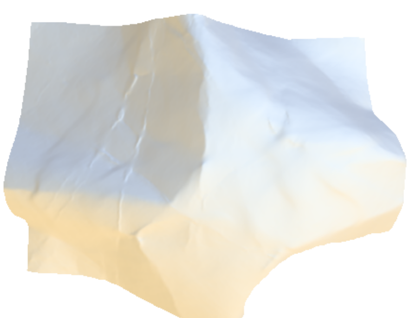
Output:



Shape



Albedo

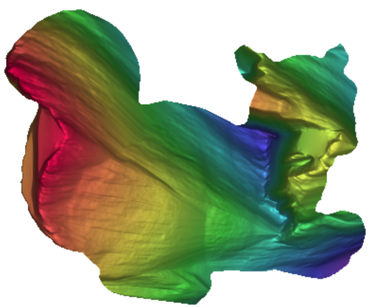


Shading

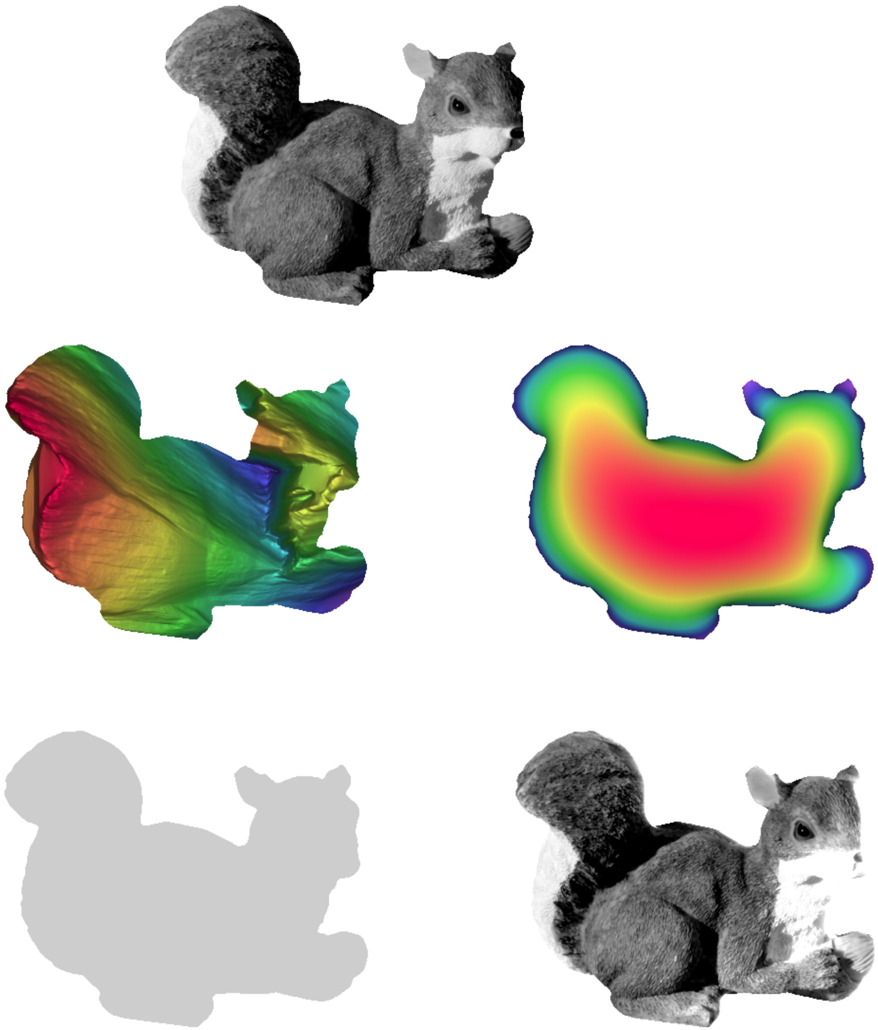
Illumination

Some Explanations

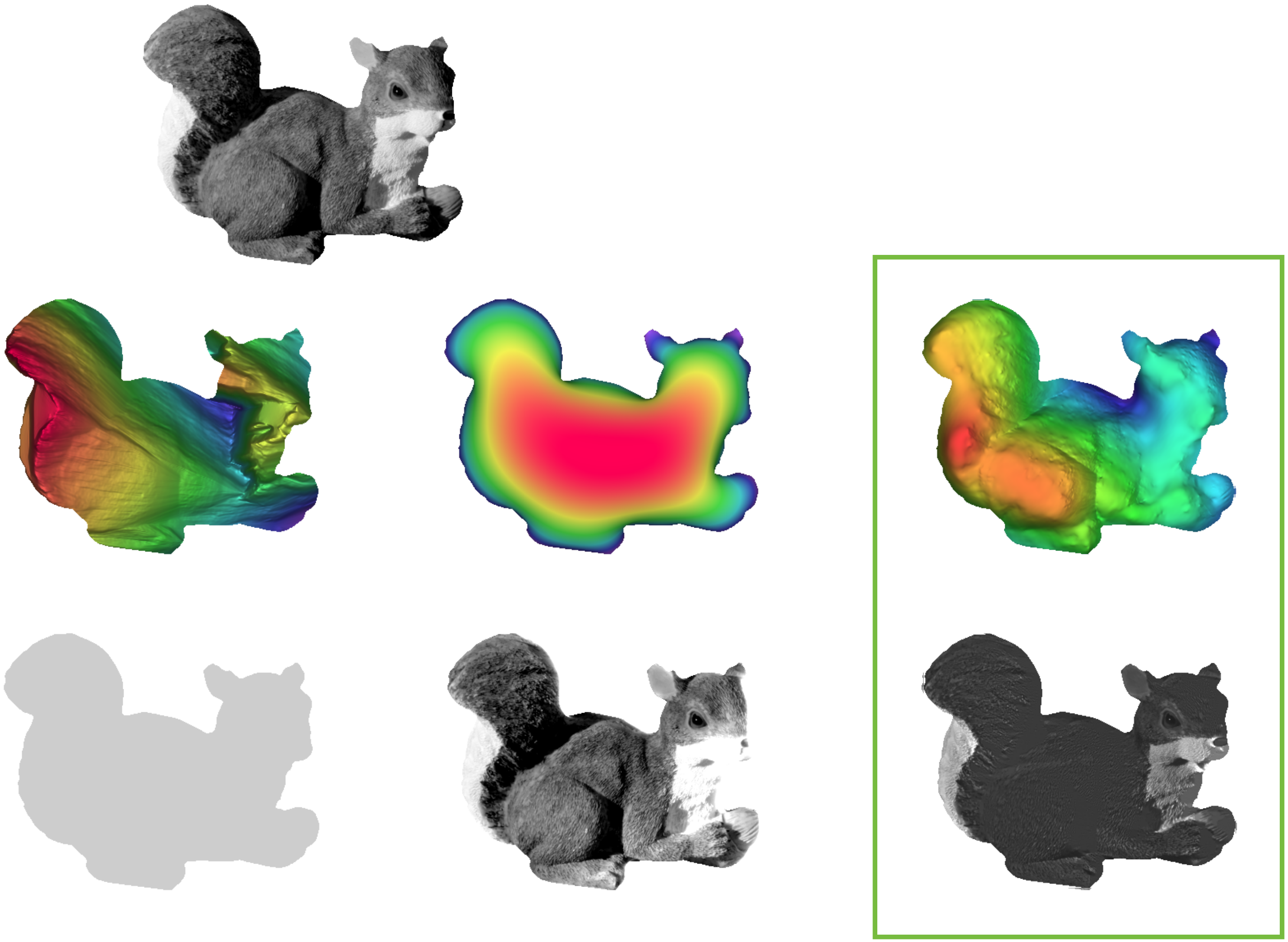
Some Explanations



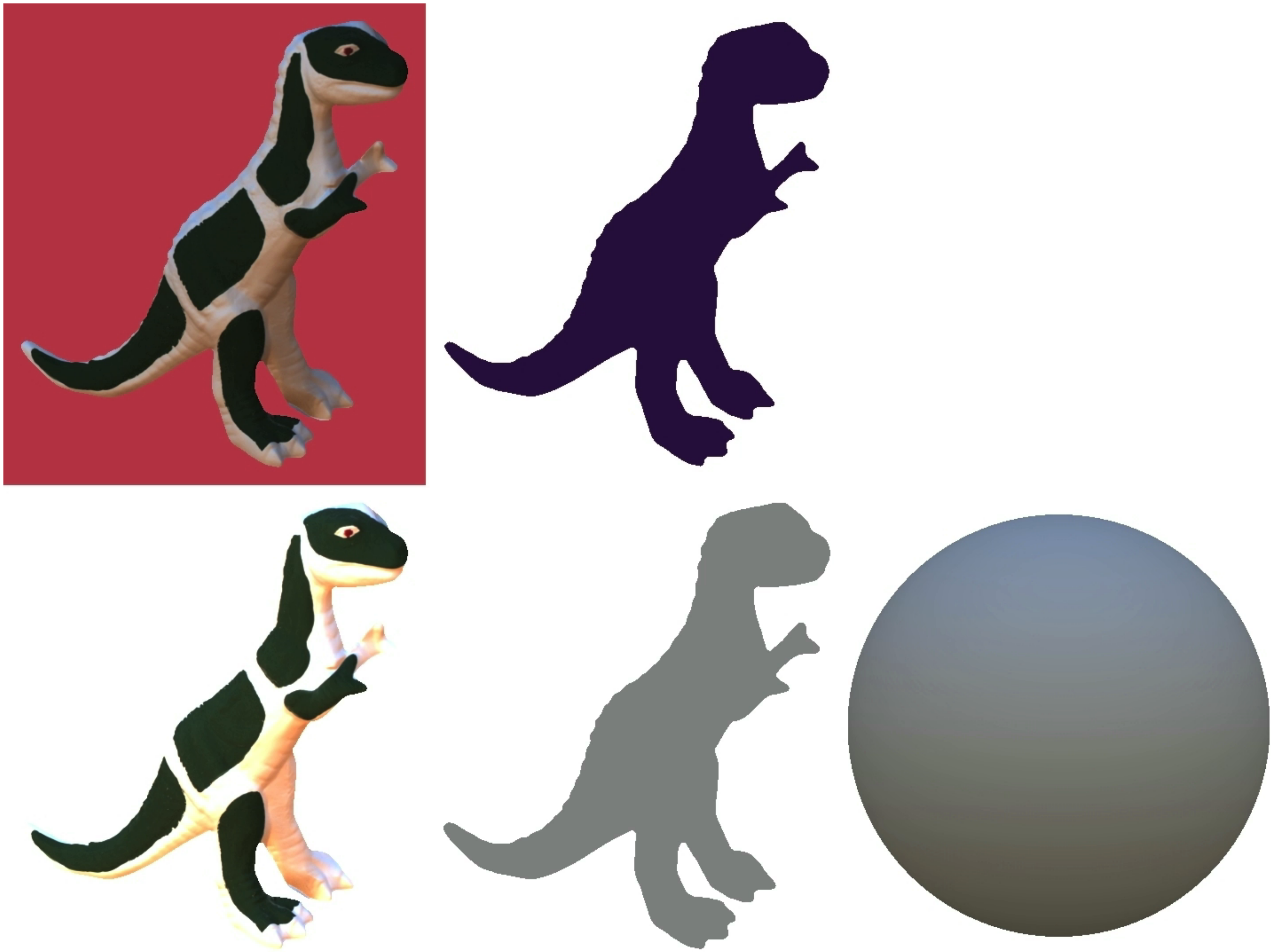
Some Explanations



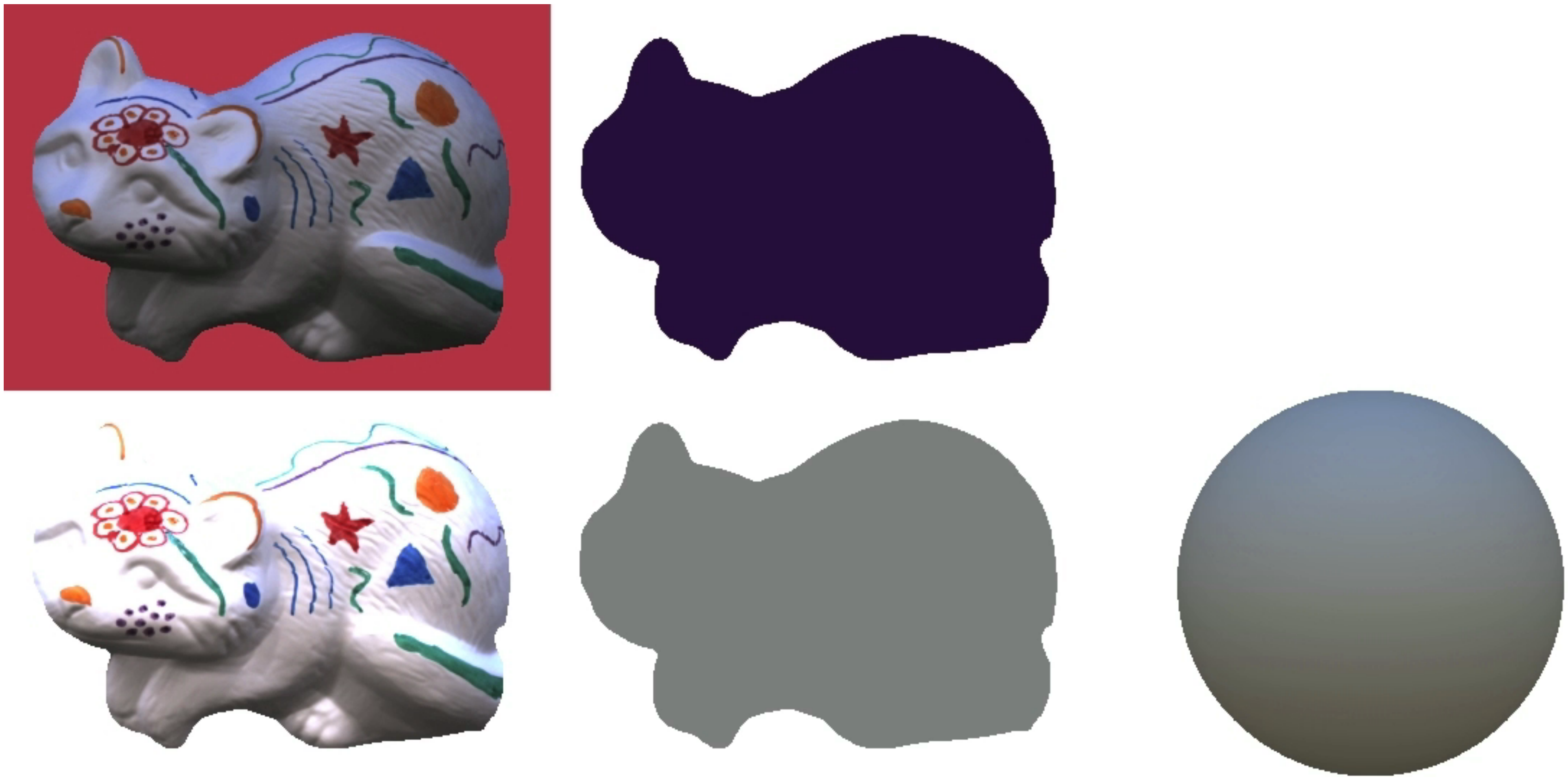
Some Explanations



Demo!



Demo!



What do we know about **albedo**?

1) Piecewise smooth at all scales and orientations
(variation is small and sparse)

2) Takes on a few discrete values everywhere in an image
(distribution is low-entropy)

$$g(A) = \sum_{k=1}^{K} 4^{k-1} \sum_{x,y} c\left(\|\nabla \mathcal{G}(A,k)\|_{x,y} \,; \boldsymbol{\alpha}_A^k, \boldsymbol{\sigma}_A^k \right) \;\; - \;\; \lambda_e \sum_{k=1}^{K} \log \left(\sum_{i=1}^{N} \sum_{j=1}^{N} \exp\left(- \frac{(\mathcal{G}(A,k)_i - \mathcal{G}(A,k)_j)^2}{4\sigma_A^2} \right) \right)$$

What do we know about **shapes**?

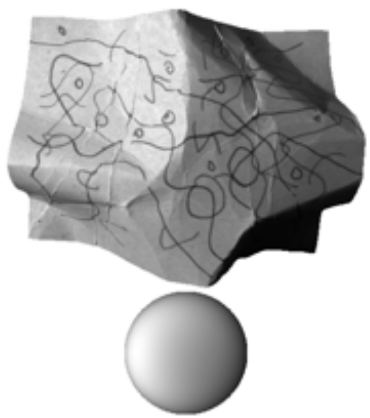
1) Piecewise smooth at all scales and orientations
   (variation of mean curvature is small and sparse)

2) Face outwards at the occluding contour

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

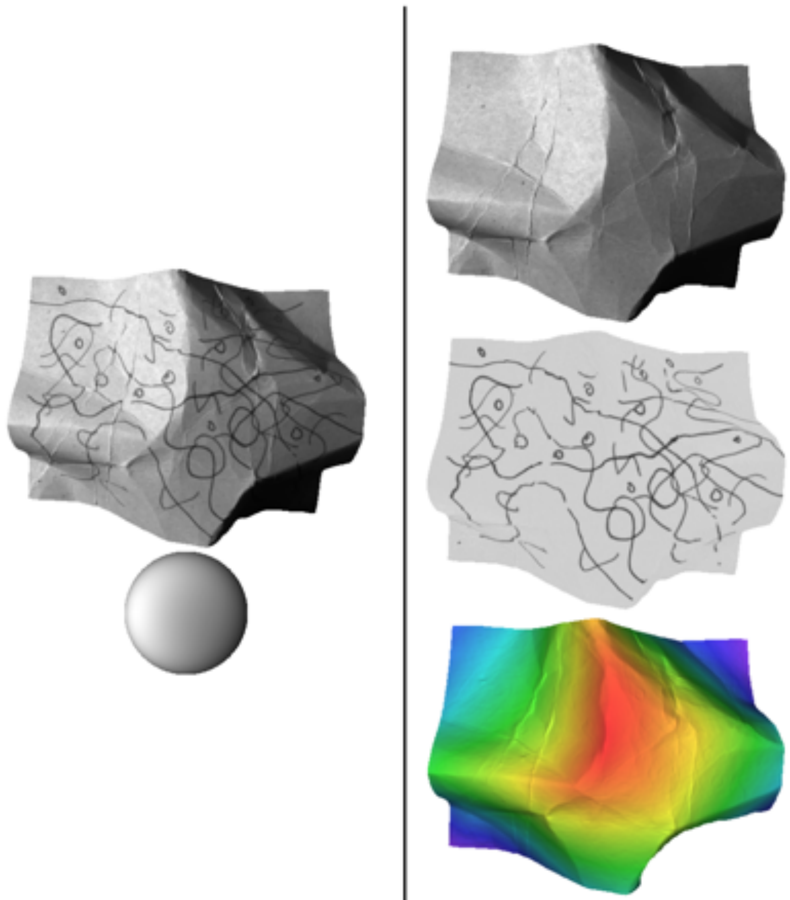
$$f(Z) = \lambda_s \sum_{k=1}^{K} 4^{k-1} \sum_{x,y} c\left(\left\| \nabla H\left(\frac{\mathcal{G}(Z,k)}{2^{k-1}} \right) \right\|_{x,y} ; \boldsymbol{\alpha}_Z^k, \boldsymbol{\sigma}_Z^k \right) \; + \; \lambda_c \sum_{i \in C} \sqrt{\left(N_i^x(Z) - n_i^x\right)^2 + \left(N_i^y(Z) - n_i^y\right)^2} \; - \; \lambda_f \sum_{x,y} \log\left(2 N_{x,y}^z(Z)\right)$$

Evaluation: Known Lighting



(a) Input Image &
Illumination

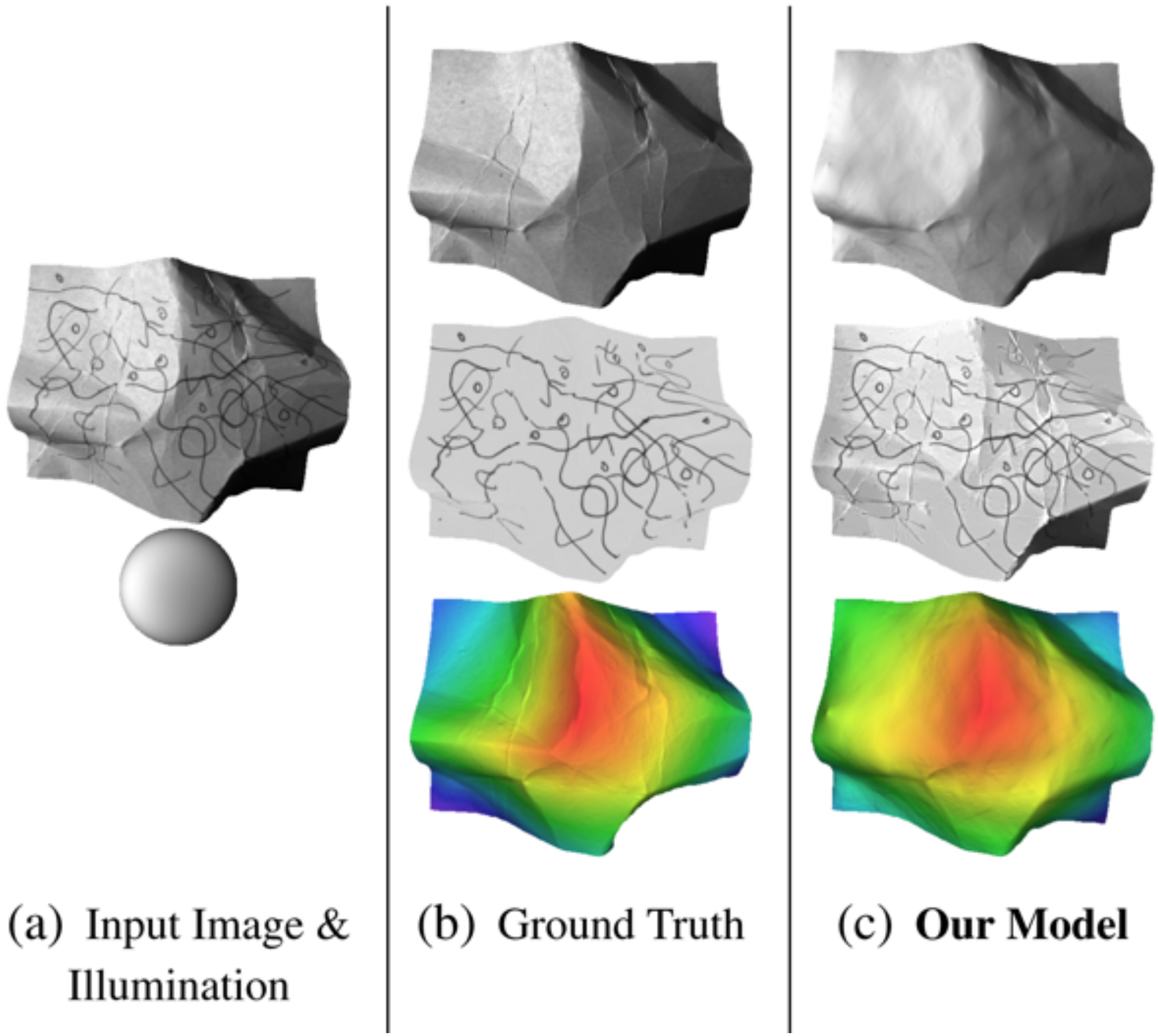
Evaluation: Known Lighting



(a) Input Image & Illumination

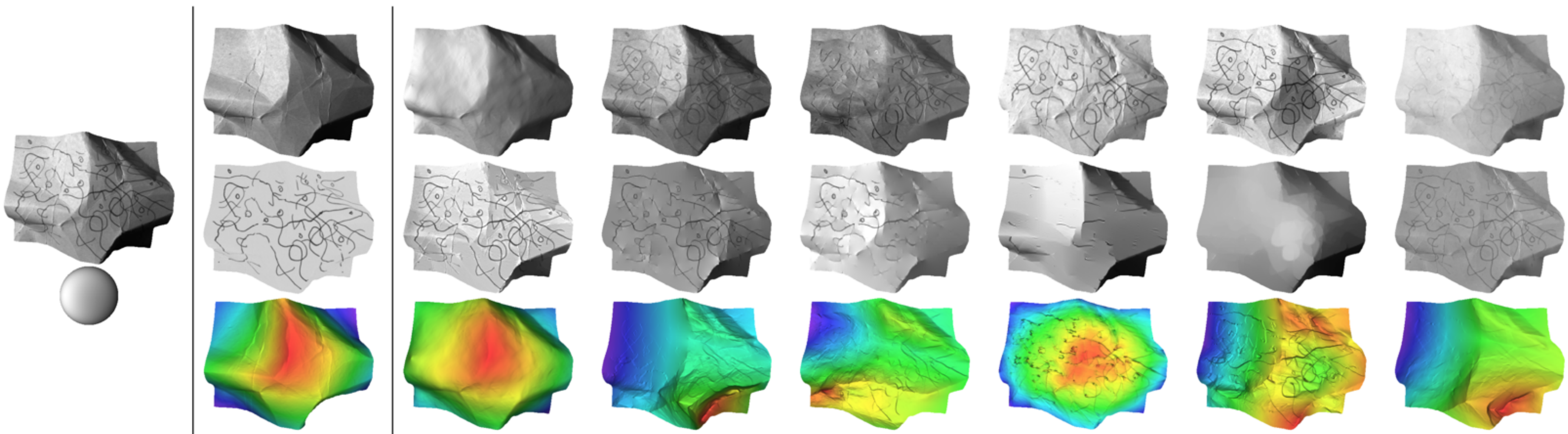
(b) Ground Truth

Evaluation: Known Lighting



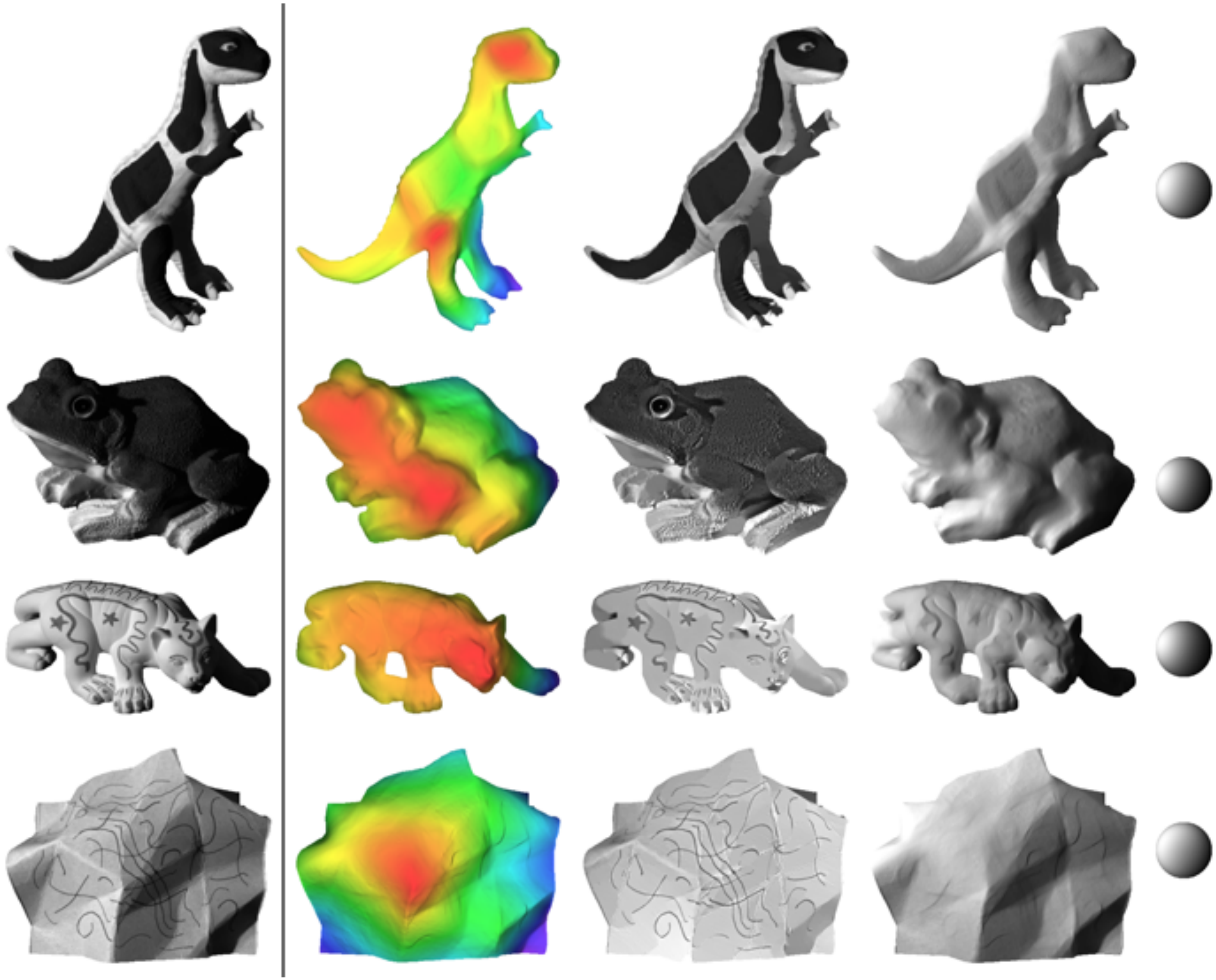
(a) Input Image & Illumination | (b) Ground Truth | (c) **Our Model**

Evaluation: Known Lighting

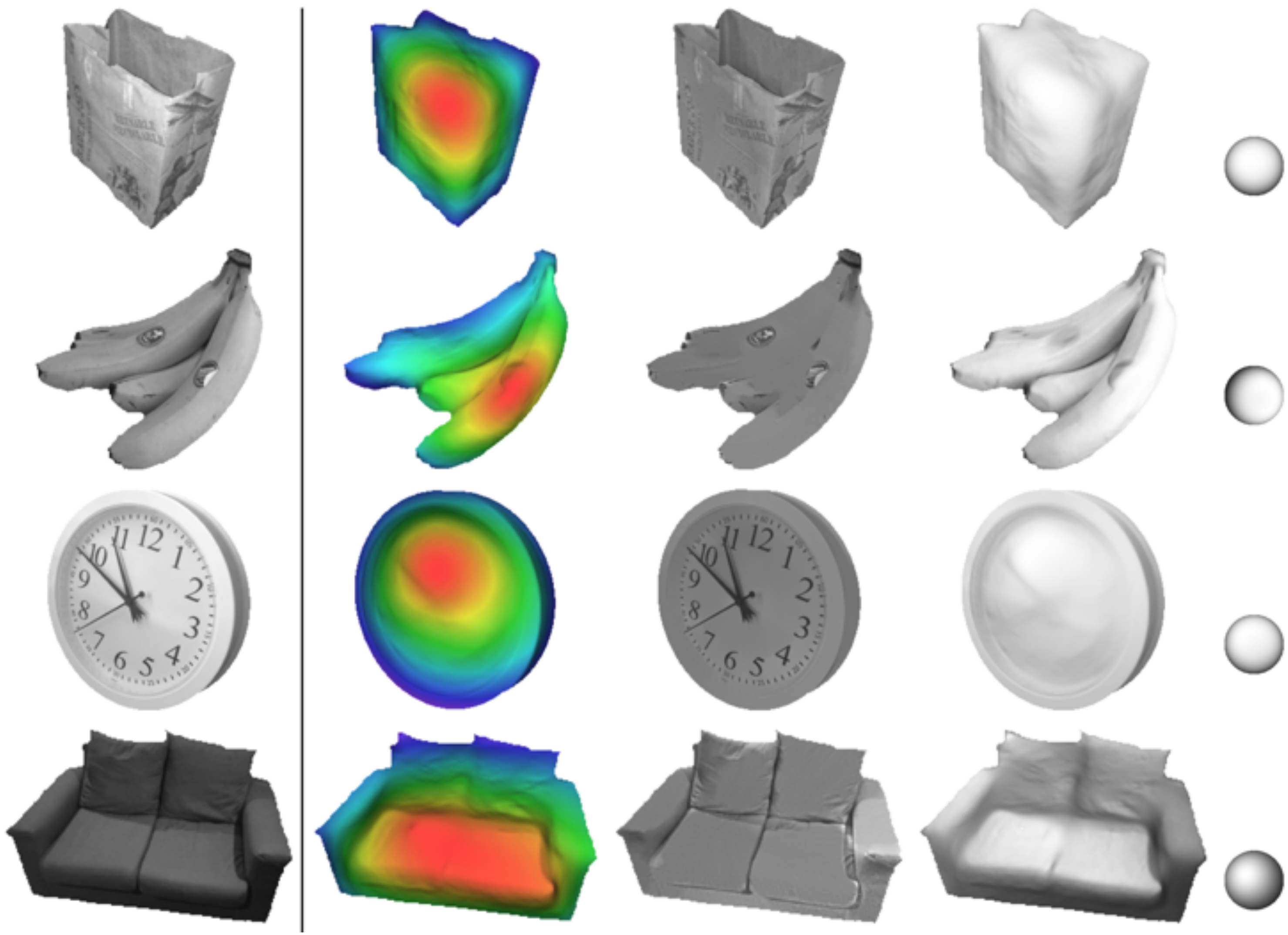


(a) Input Image & Illumination | (b) Ground Truth | (c) **Our Model** | (d) Retinex[5, 4] + SFS | (e) Tappen *et al.* 2005[10] + SFS | (f) Barron & Malik 2010[1] | (g) J. Shen *et al.* 2010[8] + SFS | (h) L. Shen & Yeo 2010[9] + SFS

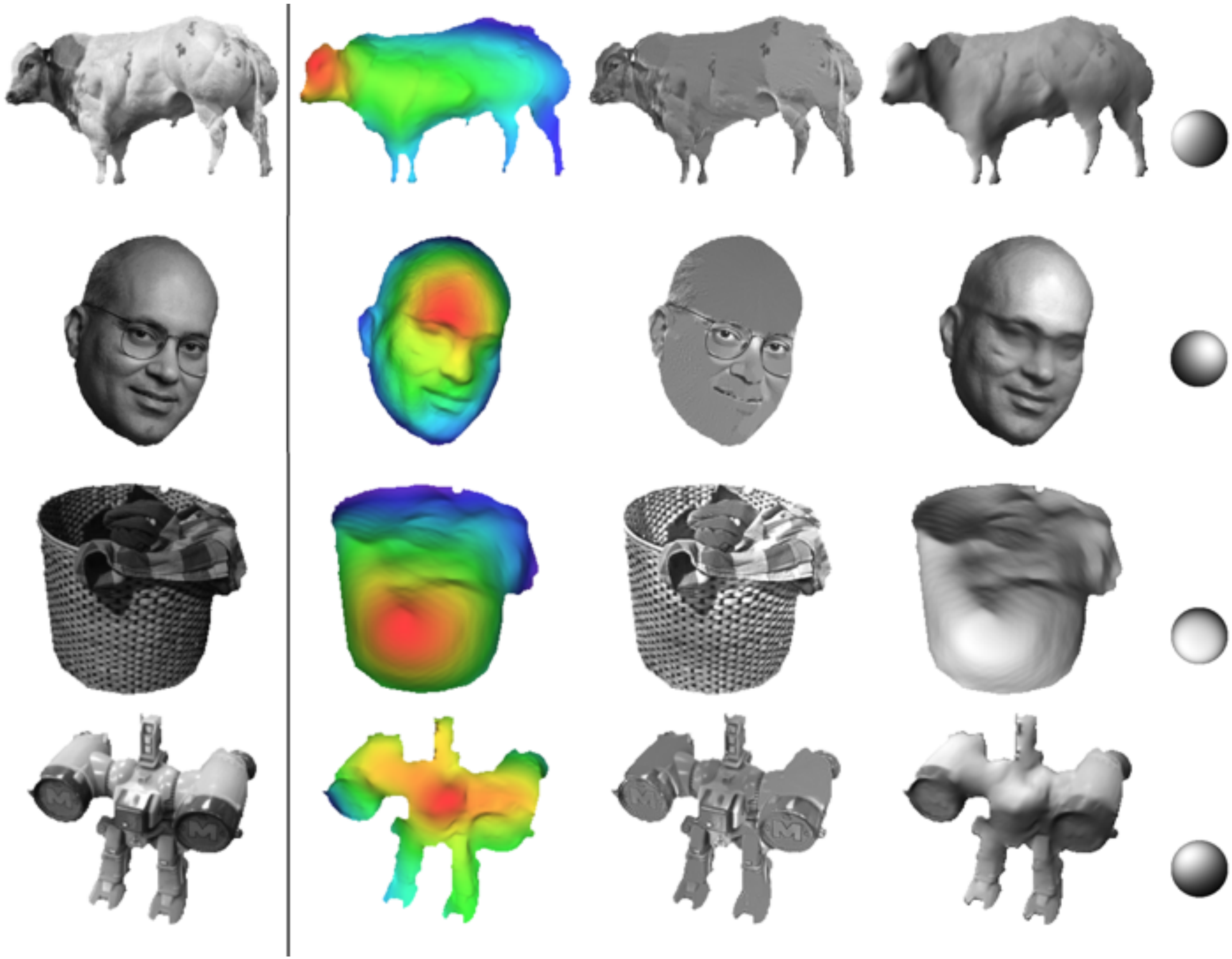
Evaluation: Unknown Lighting



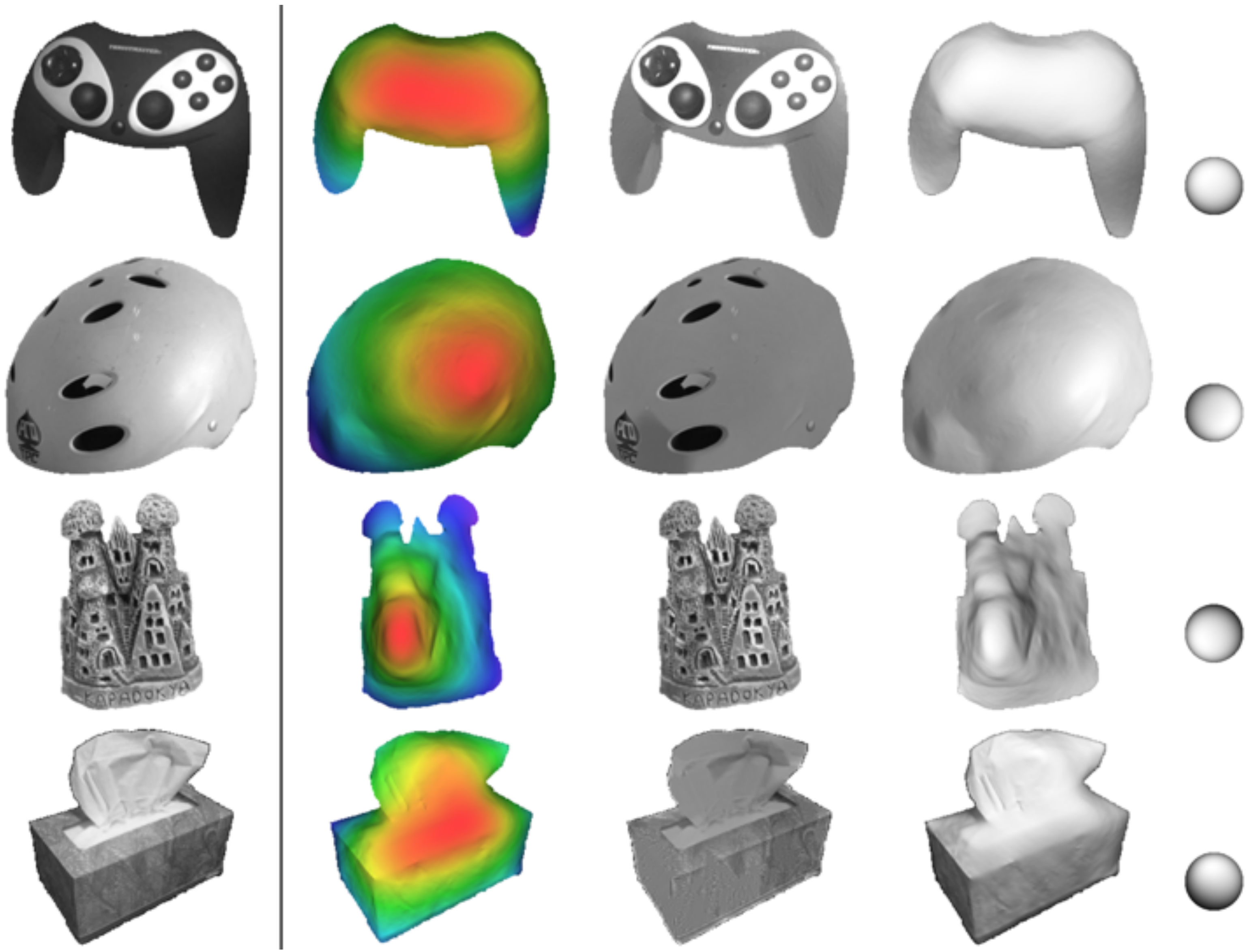
Evaluation: Real World Images



Evaluation: Real World Images



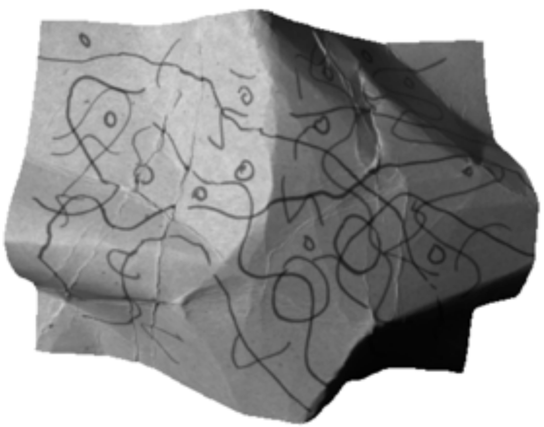
Evaluation: Real World Images



Evaluation: The Numbers

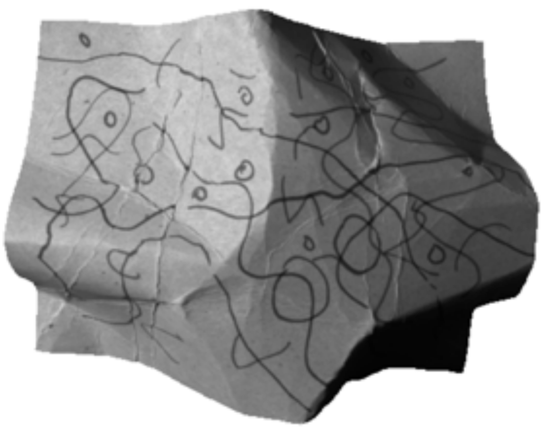
Algorithm	Avg.
Flat Baseline	0.2004
Retinex + SFS	0.2009
Tappen *et al*. 2005 + SFS	0.1761
Barron & Malik 2011	0.1682
J. Shen *et al*. 2011 + SFS	0.2376
Our Model (All Priors)	0.0856

Evaluation: Graphics!

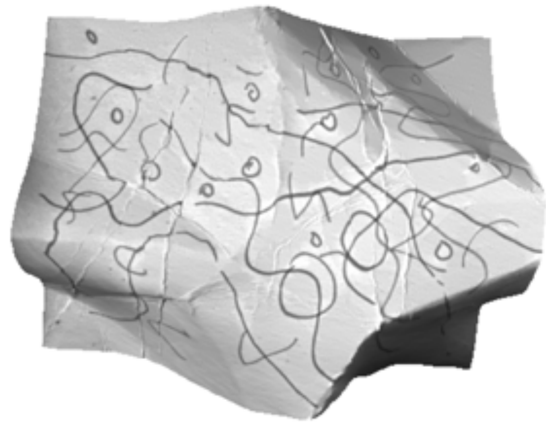


Input Image

Evaluation: Graphics!

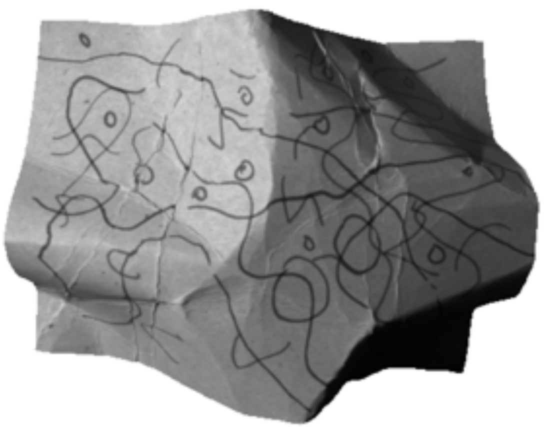


Input Image

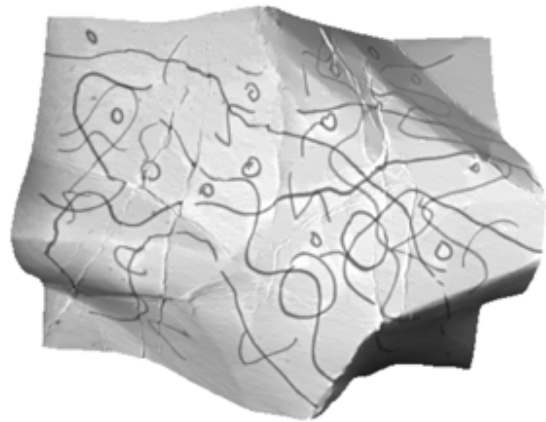


Modified illumination

Evaluation: Graphics!



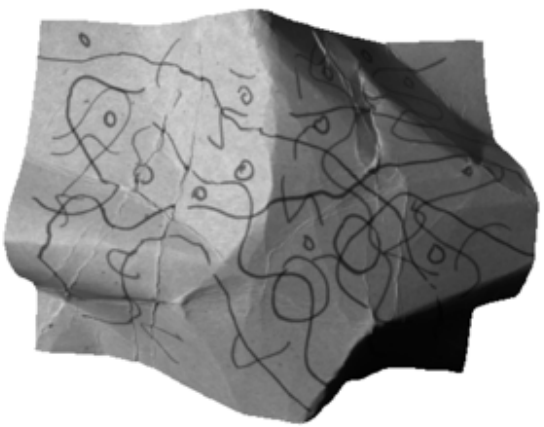
Input Image



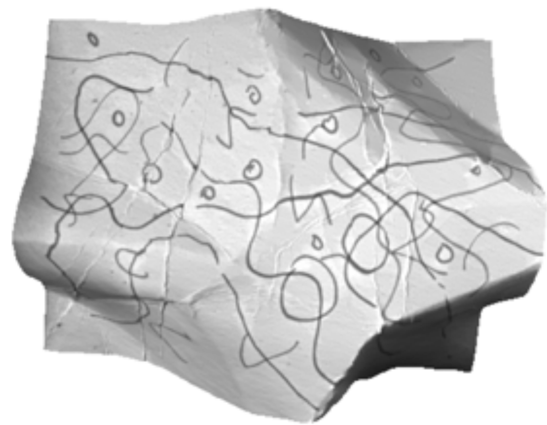
Modified illumination

Modified shape

Evaluation: Graphics!



Input Image

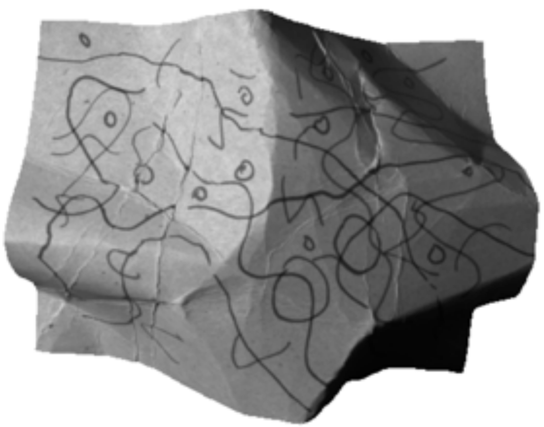


Modified illumination

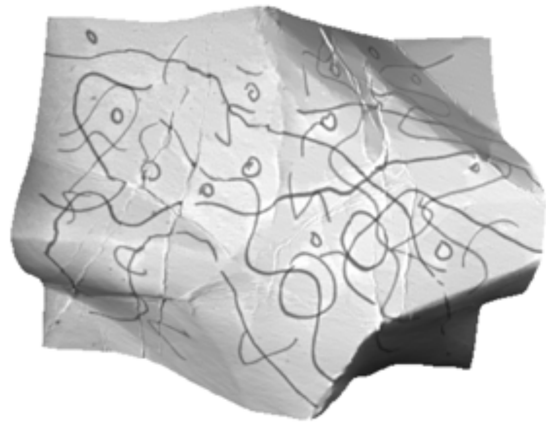
Modified shape

Modified albedo

Evaluation: Graphics!



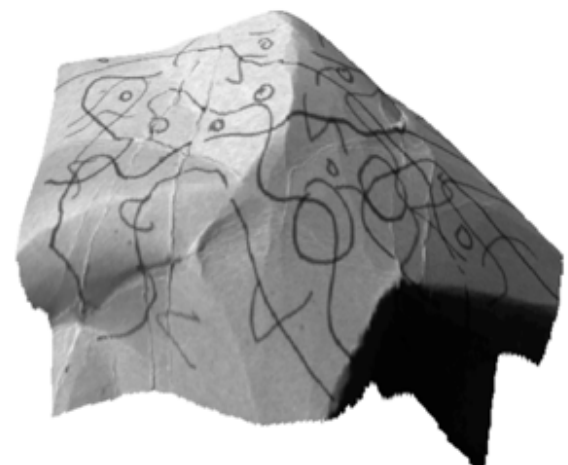
Input Image



Modified illumination

Modified shape

Modified albedo

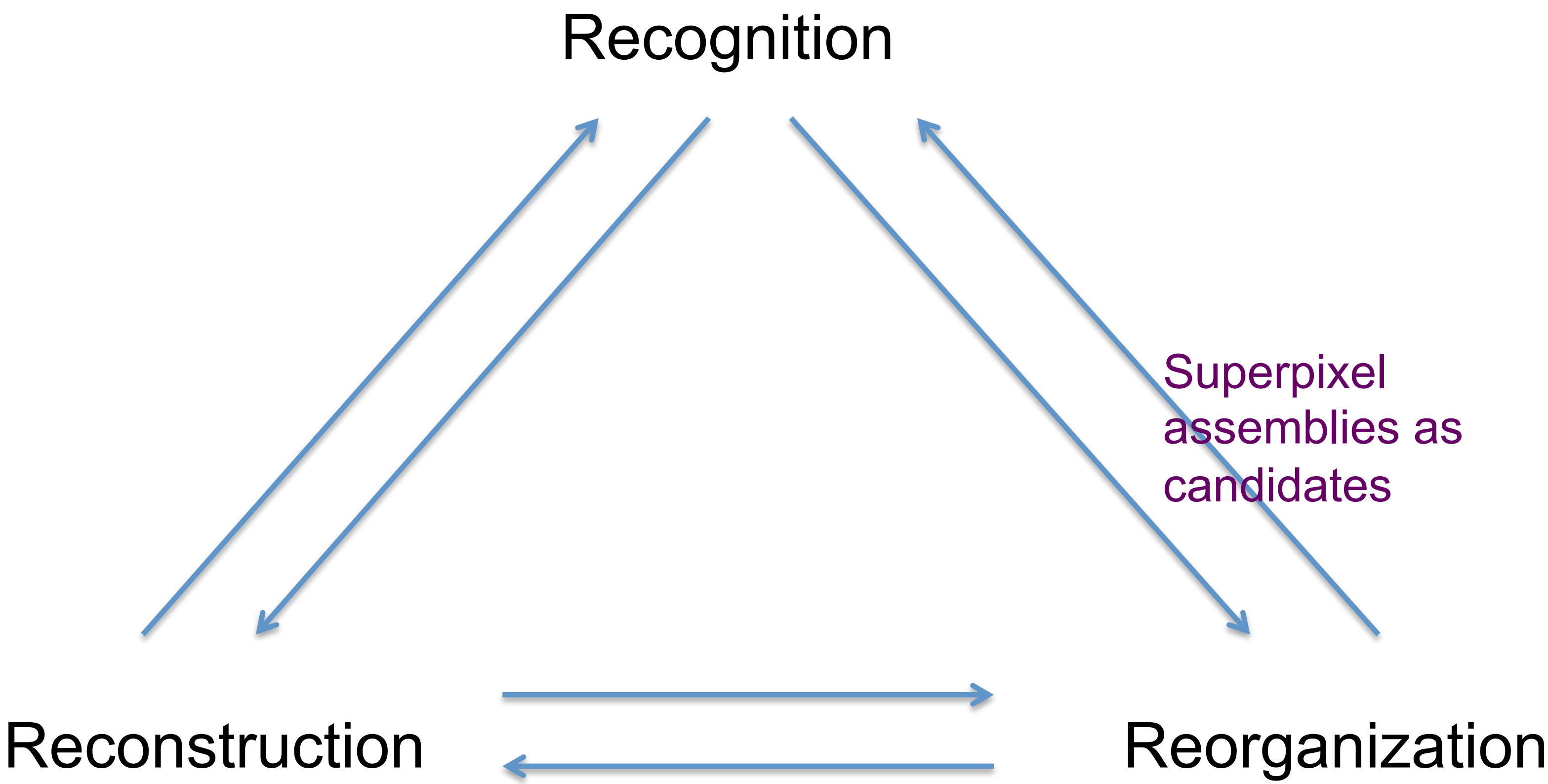


Modified orientation

Conclusions

- Unification shape-from-shading, intrinsic images, and color constancy
- Solving the unified problem > Solving any sub-problem
- Not a toy
- Not (and can never be?) *metrically* accurate

The Three R's of Vision

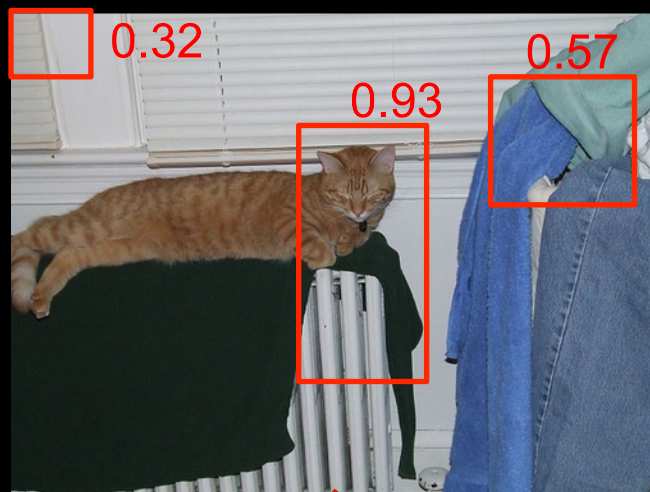


Semantic Segmentation using Regions and Parts

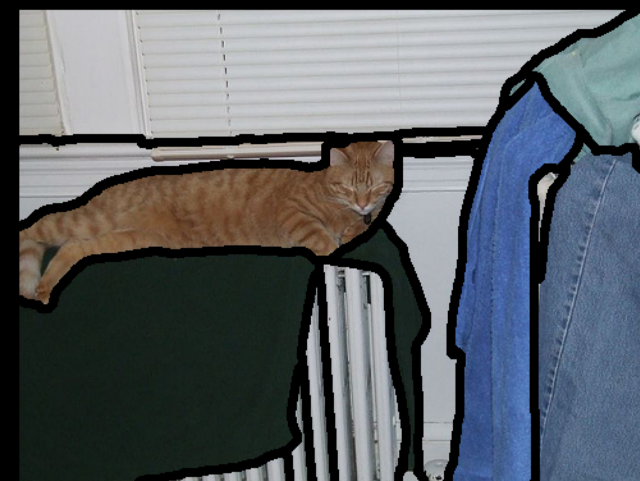
P. Arbeláez, B. Hariharan, S. Gupta, C. Gu, L. Bourdev and J. Malik

This Work

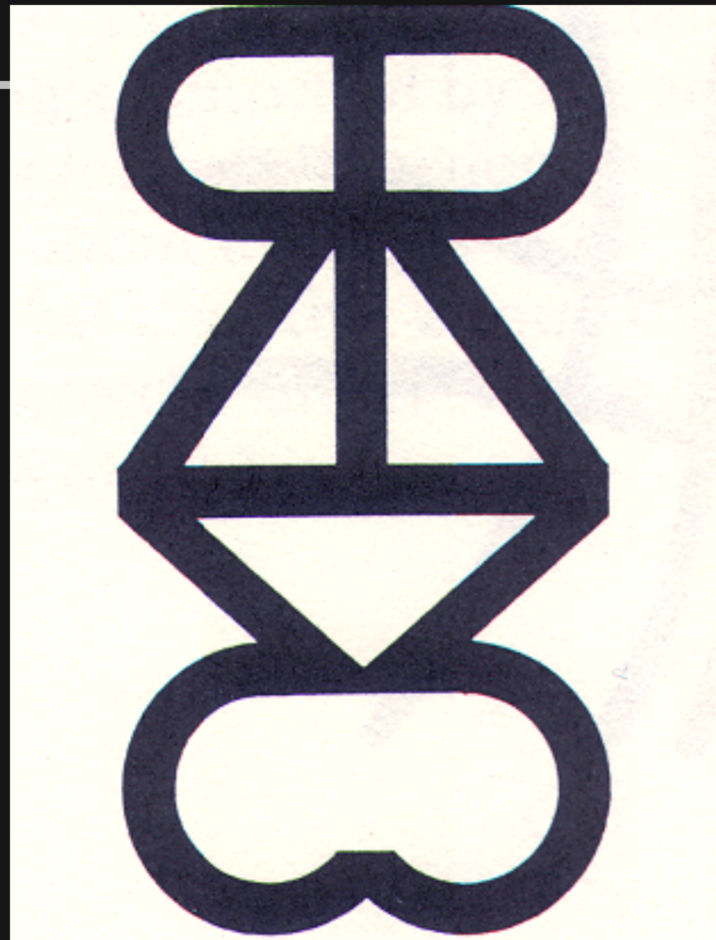
Top-down Part/Object Detectors



Bottom-up Region Segmentation



Cat Segmenter




R
3


WITHOUT MOTIVATION

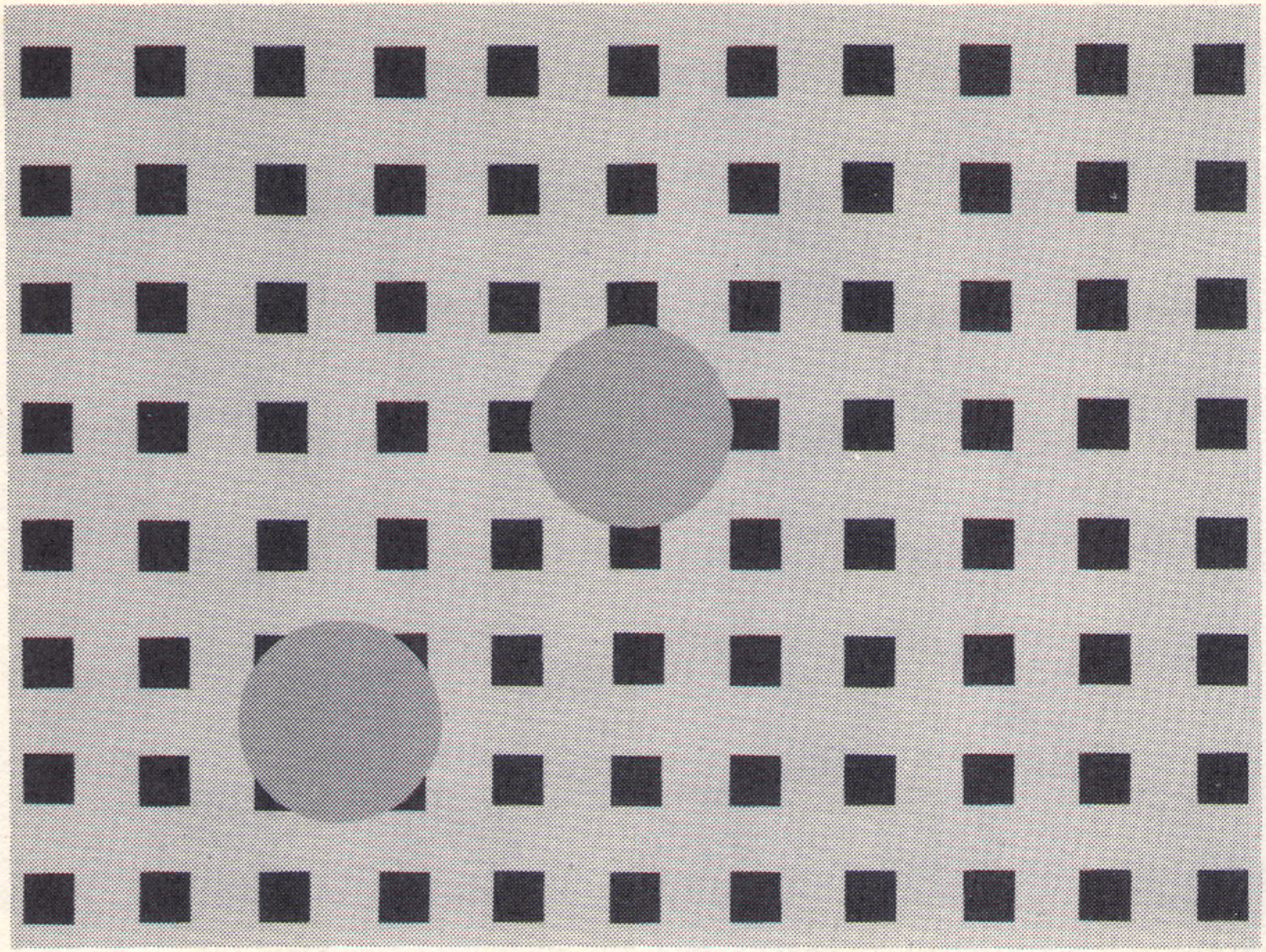


Figure 4.15 The "law of the whole" does not impose upon the parts. Behind the disk there is a cross or a large square, but not the squares that are the elements of the regular array.

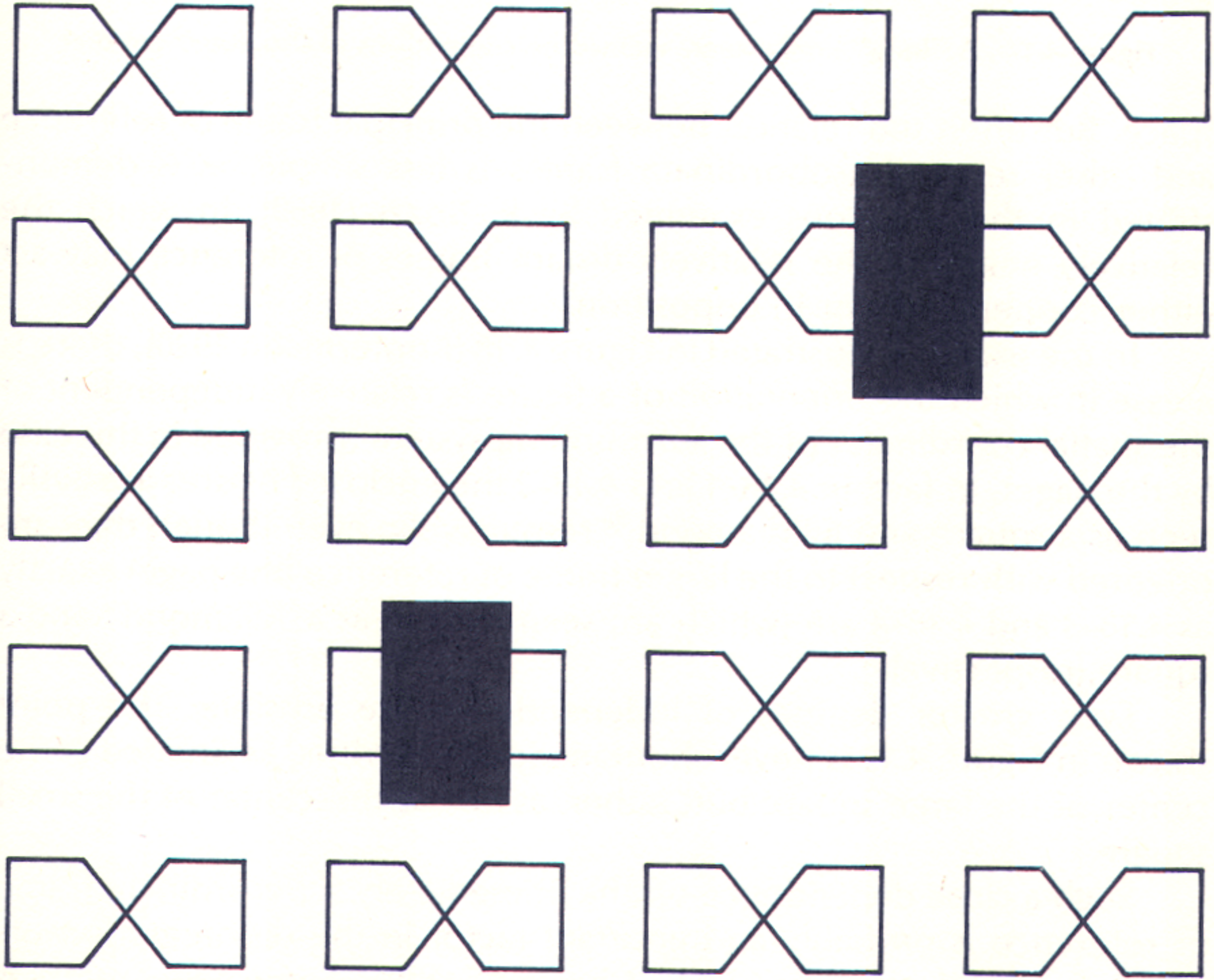


Figure 4.16 The totalization conforms to "local" conditions.

Overview

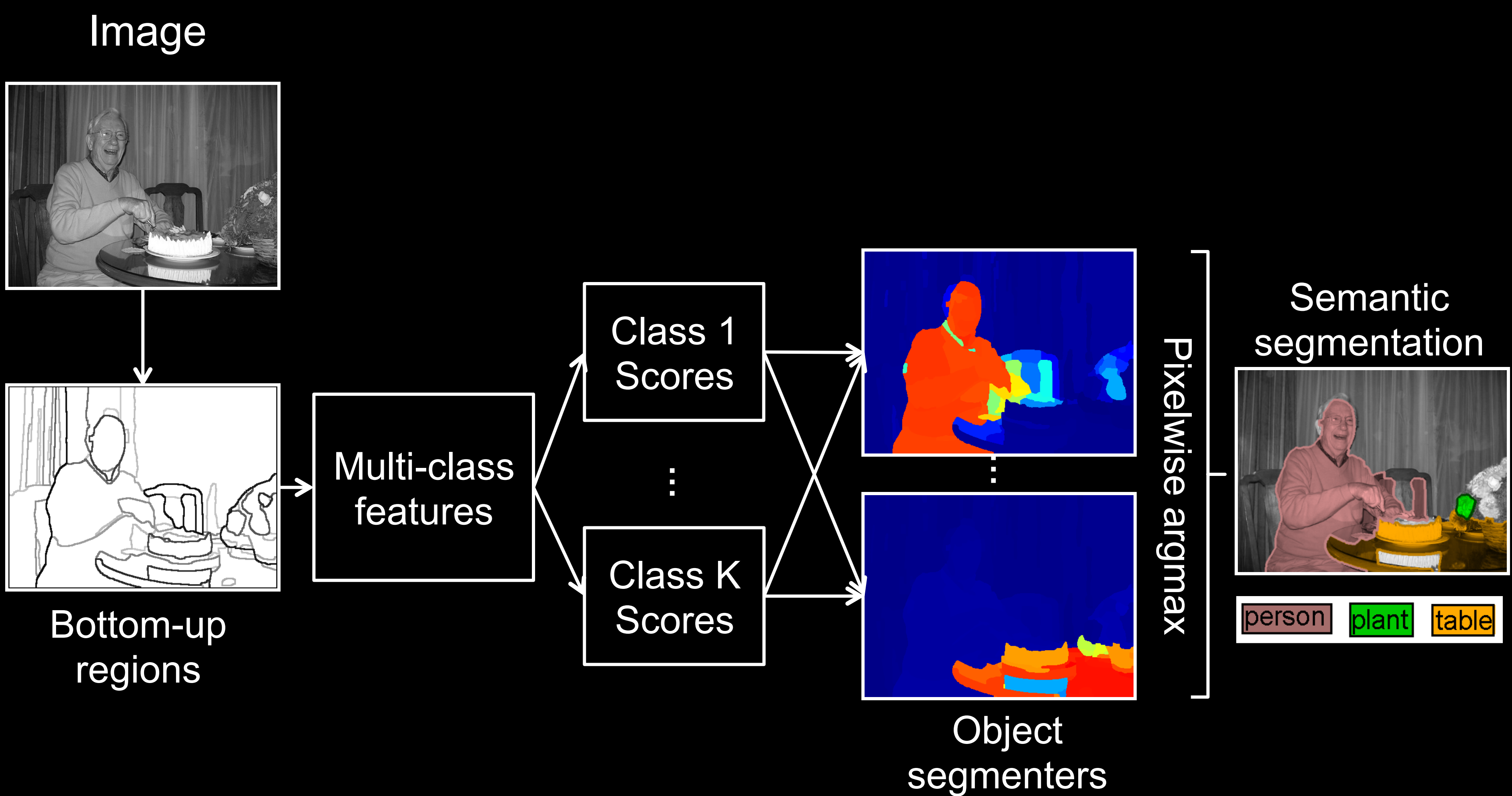


Image
Bottom-up
regions
Multi-class
features
Class 1
Scores
⋮
Class K
Scores
⋮
Object
segmenters
Pixelwise argmax
Semantic
segmentation
person
plant
table

Region Generation

- Hierarchical segmentation tree based on contrast
- Hierarchy computed at three image resolutions
- Nodes of the trees are object candidates, and also pairs and triplets of adjacent nodes

- **Output:** For each image, a set of ~1000 object candidates (connected regions)

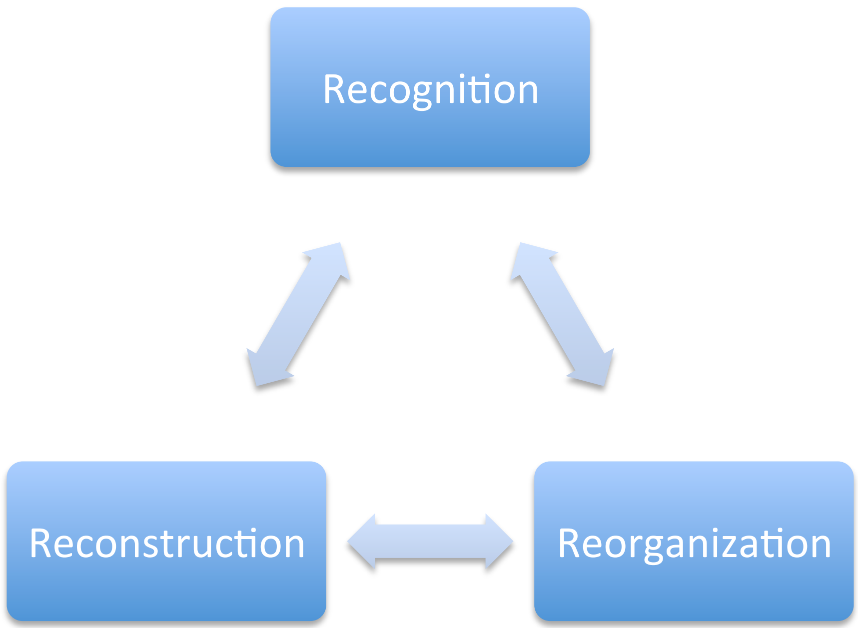
	plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	table	dog	horse	mbike	person	plant	sheep	sofa	train	tv	artic	transp	indoors	all
gPb-owt-ucm [4]	59.3	32.9	70.3	51.1	61.3	51.2	57.6	74.3	58.0	68.6	67.4	67.5	64.3	48.5	53.6	53.5	72.6	71.2	55.1	73.1	67.3	50.8	64.1	60.6
Our regions	**76.7**	**41.6**	**84.0**	**74.2**	**77.2**	**75.8**	**74.9**	**85.2**	**69.6**	**79.1**	**82.9**	**82.4**	**75.9**	**69.6**	**74.4**	**70.4**	**80.3**	**83.2**	**76.5**	**85.1**	**80.2**	**69.9**	**78.1**	**76.0**

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

How to think about Vision

- “Theory”



- Models
  - Feature Histograms
  - Support Vector machines
  - Randomized decision trees
  - Spectral partitioning
  - L1 minimization
  - Stochastic Grammars
  - Deep Learning
  - Markov Random Fields
  - …

Thanks!