Toward Geometrically Coherent Image Interpretation

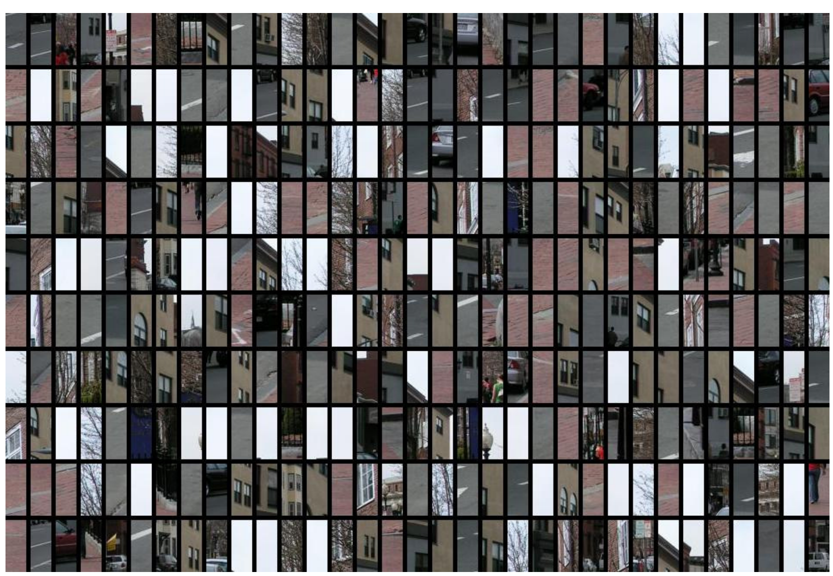
Alexei (Alyosha) Efros
CMU

Joint work with Derek Hoiem and Martial Hebert

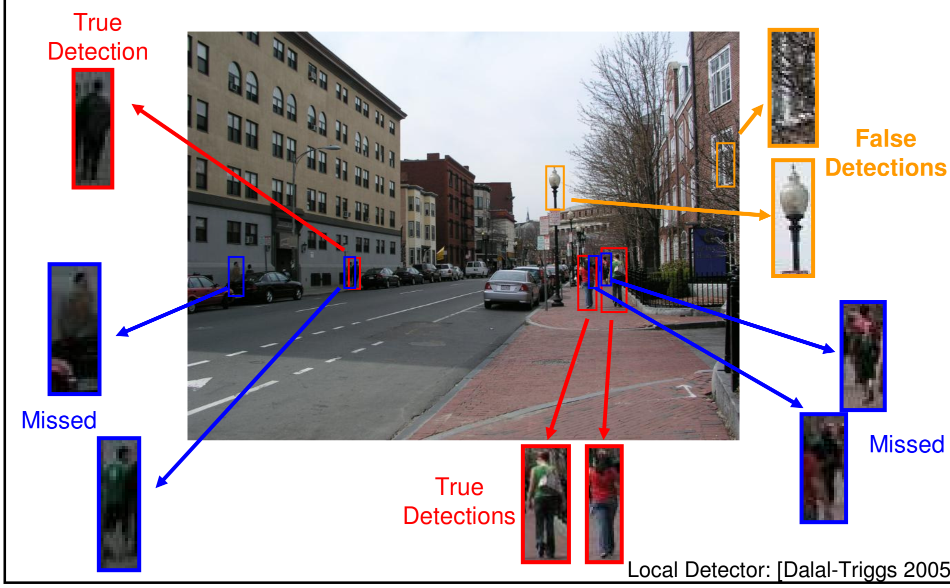
Understanding an Image

Today: Local and Independent

What the Detector Sees

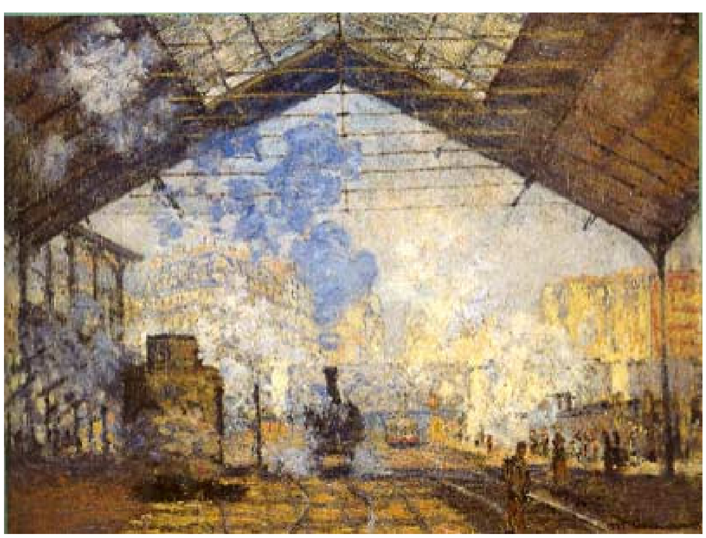


Local Object Detection



Local Detector: [Dalal-Triggs 2005]

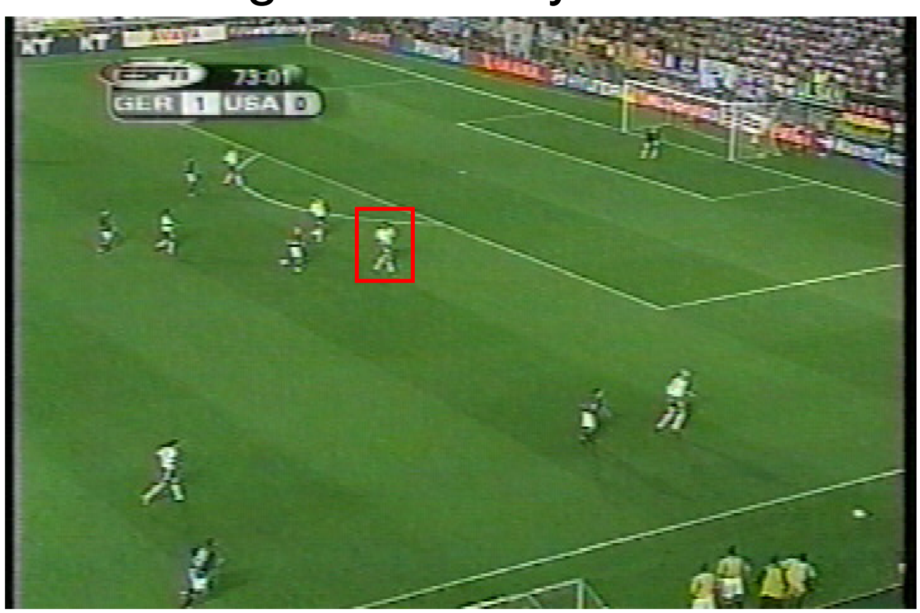
Importance of Context



Claude Monet
Gare St.Lazare
Paris, 1877

There is almost nothing <u>inside</u>!

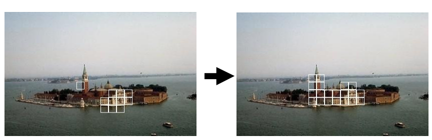
Seeing less than you think…



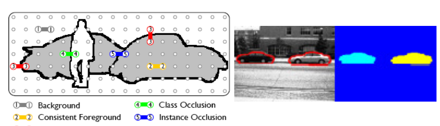
Seeing less than you think…

Need to think "outside the box"

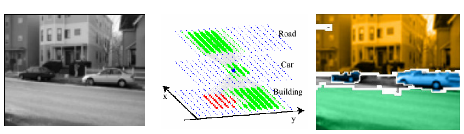
Recent Work on *2D* Spatial Context



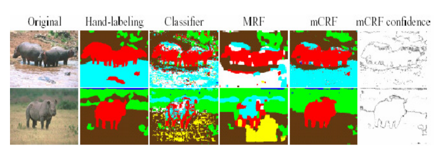
[Kumar & Hebert 2005]



[Winn & Shotton 2006]

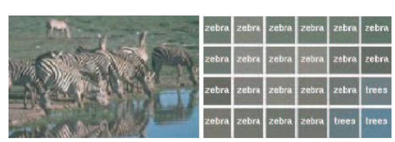


[Torralba, Murphy, Freeman 2004]



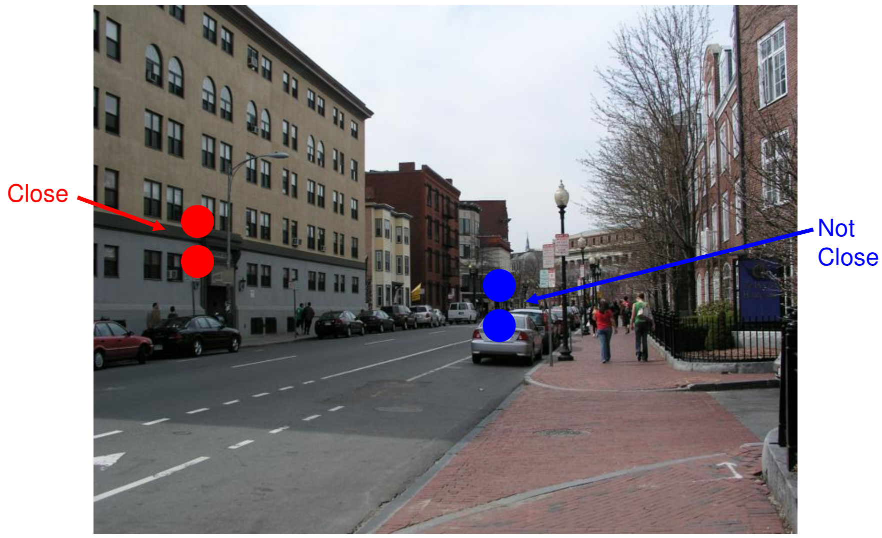
[He, Zemel, Cerreira-Perpiñán 2004]

[Fink & Perona 2003]

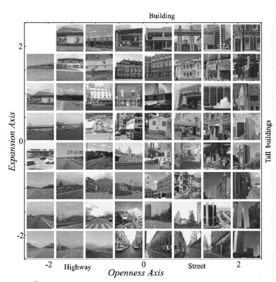


[Carbonetto, Freitas, Banard 2004]

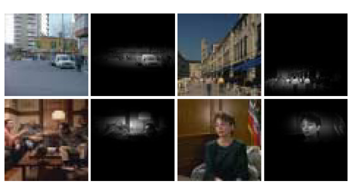
Real Relationships are 3D



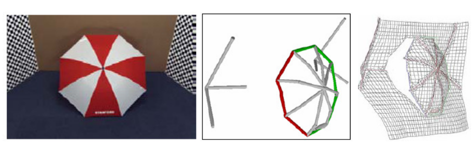
Recent Work in 3D



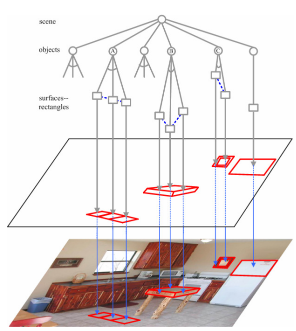
[Oliva & Torralba 2001]



[Torralba, Murphy & Freeman 2003]

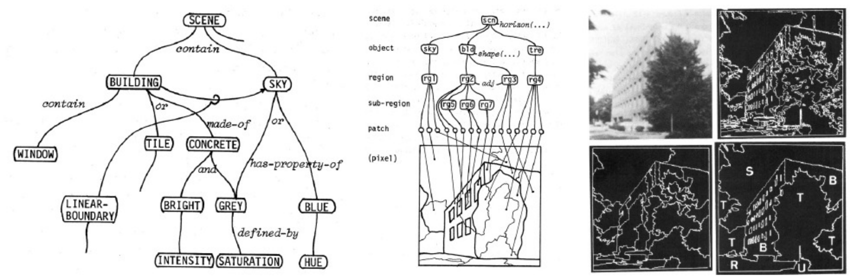


[Han & Zu 2003]



[Han & Zu 2005]

Scene Understanding in 1970s

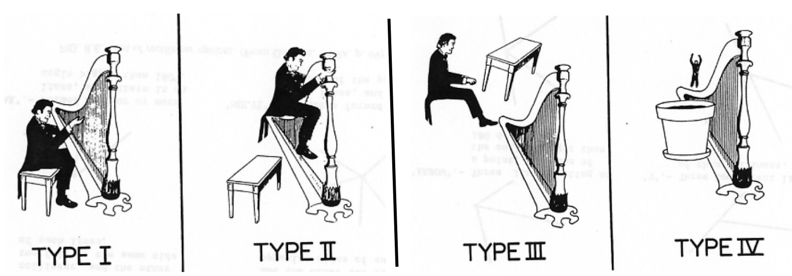


(a) Bottom-up process (b) Top-down process (c) Result

[Ohta & Kanade 1978]

- Guzman (*SEE*), 1968
- Hansen & Riseman (*VISIONS*), 1978
- Barrow & Tenenbaum 1978
- Brooks (*ACRONYM*), 1979
- Marr, 1982
- Ohta & Kanade, 1978
- Yakimovsky & Feldman, 1973

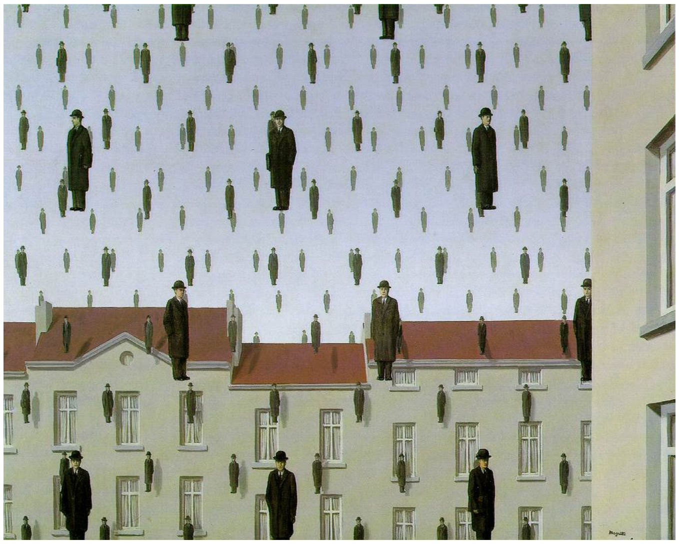
Objects and Scenes



Hock, Romanski, Galie, & Williams 1978

- Biederman's Relations among Objects in a Well-Formed Scene (1981):
  - Support
  - Size
  - Position
  - Interposition
  - Likelihood of Appearance

Support



Rene Magritte, *Golconde*

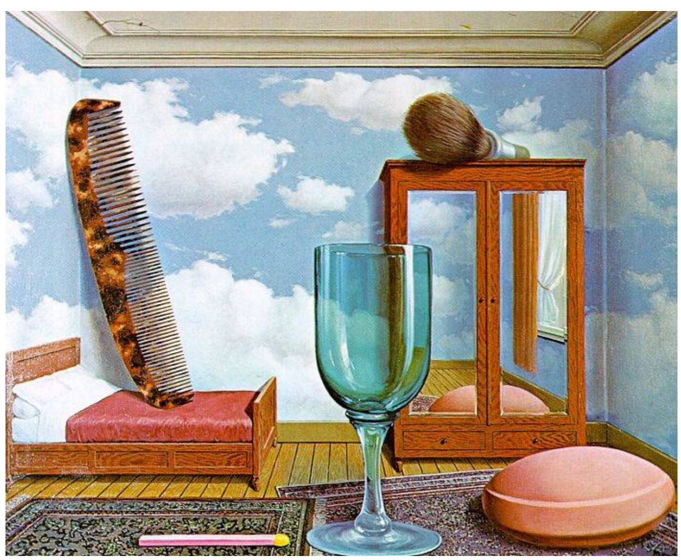
Size

Rene Magritte, *The Listening Room*

Interposition

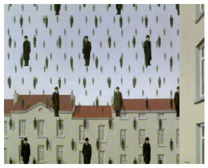
Rene Magritte, *Black Check*

Position, Probability, Size



Rene Magritte, *Personal Values*

Talk Outline



Estimating Surface Layout
[ICCV'05]

Putting Objects in Perspective
[CVPR'06]

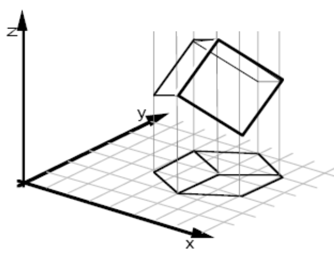
Automatic Photo Pop-up
[SIGGRAPH'05]

The World Behind the Image

Automatic Photo Pop-up, SIGGRAPH'05

The Problem

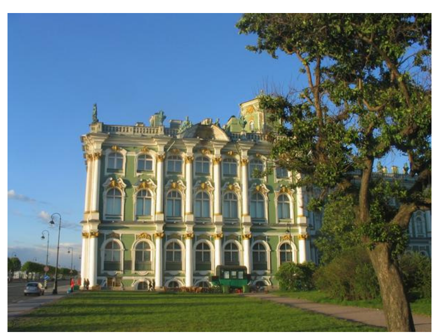
- Recovering 3D geometry from **single** 2D projection
- Infinite number of possible solutions!



from [Sinha and Adelson 1993]

Our World is Structured

Abstract World



Our World

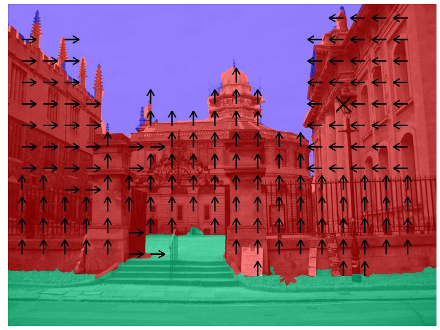
Image Credit (left): F. Cunin and M.J. Sailor, UCSD

Our Goals

- Simple, piecewise planar models
- Rough "Geometric Frame"
- Outdoor scenes

Rough "Geometric Frame"

Label Geometric Classes

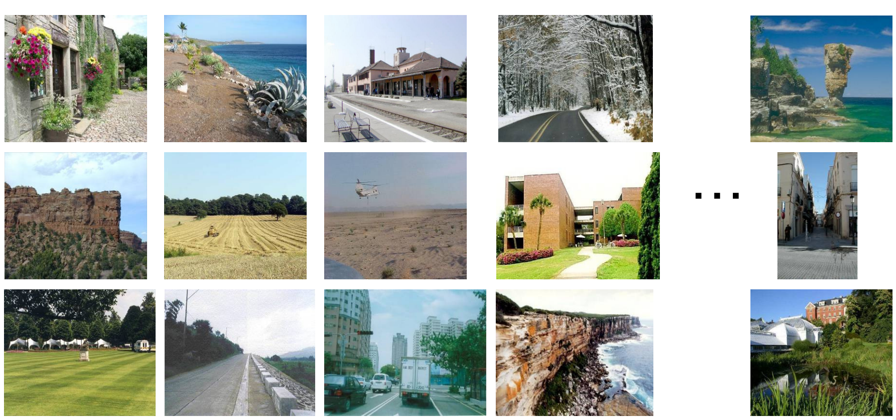



Goal: learn labeling of image into 7 Geometric Classes:

- **Support (ground)**
- **Vertical**
  - Planar: facing **Left** (**←**), **Center** (**↑**), **Right** (**→**)
  - Non-planar: **Solid** (**X**), **Porous** or wiry **(O)**
- **Sky**

Our Approach: Learning

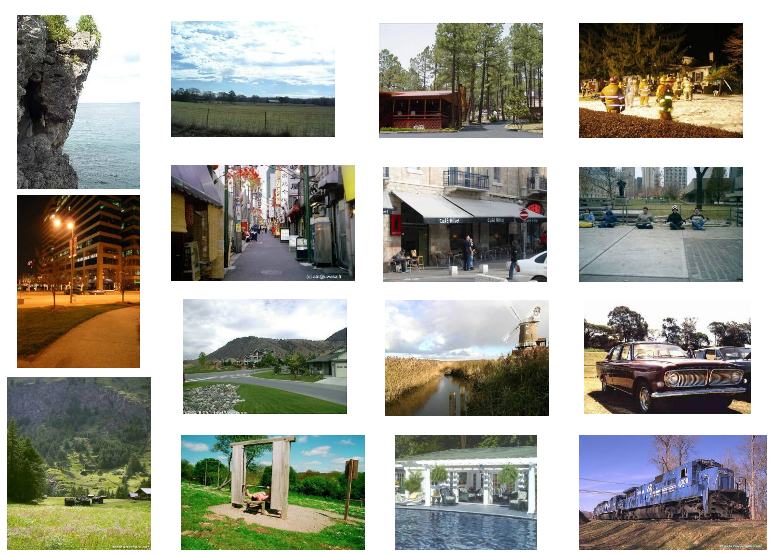
- Learn structure of the world from labeled examples



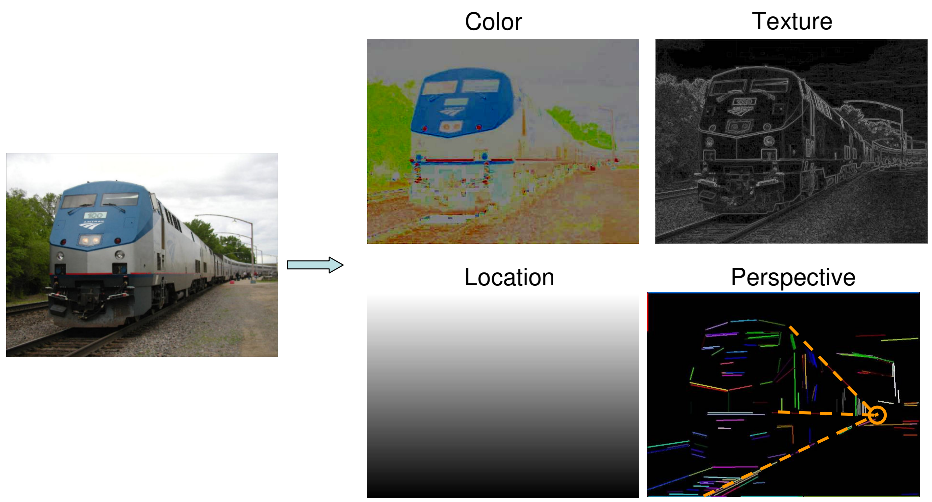
The General Case (outdoors)

- Typical outdoor photograph off the Web
  - Got 300 images using Google Image Search keyboards: “outdoor”, “scenery”, “urban”, etc.
- Certainly not random samples from world
  - 100% horizontal horizon
  - Camera axis usually parallel to ground plane
  - 97% pixels belong to 3 classes -- ground, sky, vertical (gravity)
- Still very general dataset!

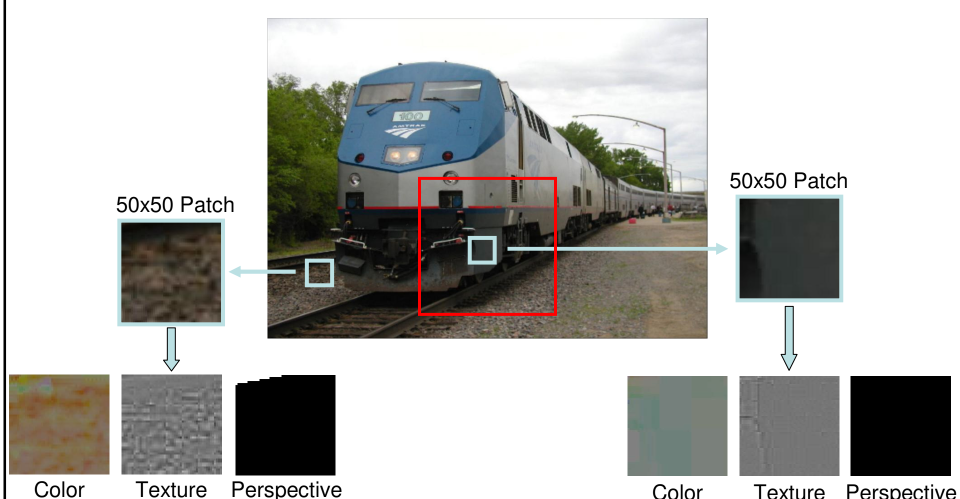
More samples from our dataset



Weak Geometric Cues



Need Spatial Support

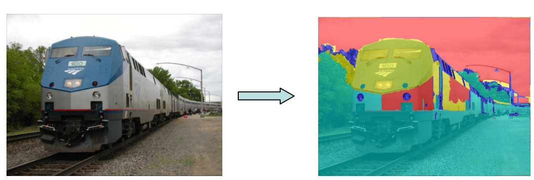


The Right Spatial Support

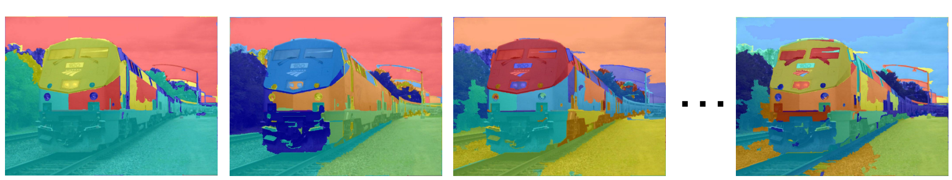
- Some features are (relatively) local
  - Color, location, texture
- But geometric features are more global
  - Long lines, vanishing points, texture gradients
- Need to find the right <u>spatial support</u> for computing features
- *Conjecture:* getting better spatial support would allow for simpler features

Image Segmentation

- Naïve Idea #1: segment the image



  - Chicken & Egg problem
- Naïve Idea #2: <u>multiple</u> segmentations



  - Decide later which segments are good

Learn from training images

Homogeneity Likelihood

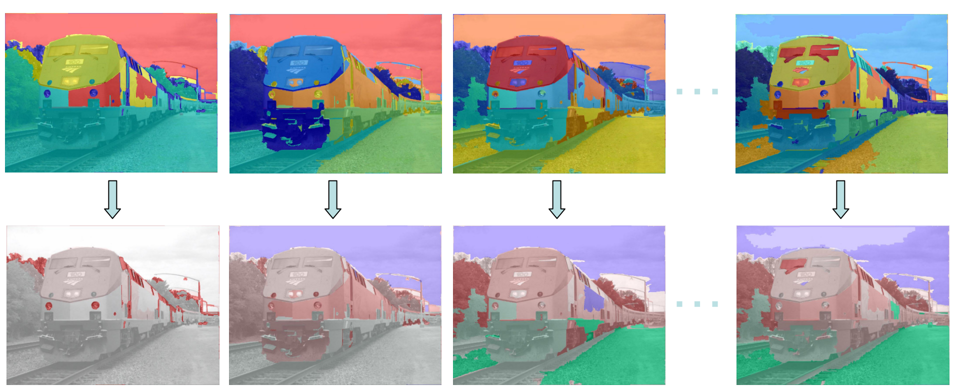
$P(\mathbf{h}_{ji}|\mathbf{x})$

Label Likelihood

$P(y_j = v|\mathbf{x}, \mathbf{h}_{ji})$

- Prepare training images
  - Create multiple segmentations of training images
  - Get segment labels from ground truth – ground, vertical, sky, or “mixed”
- Density estimation by boosted decision trees
  - 8 nodes per tree
  - Adaboost

Labeling Segments



For each segment:

- Get $P(y_j = v|\mathbf{x}, \mathbf{h}_{ji})P(\mathbf{h}_{ji}|\mathbf{x})$

Image Labeling

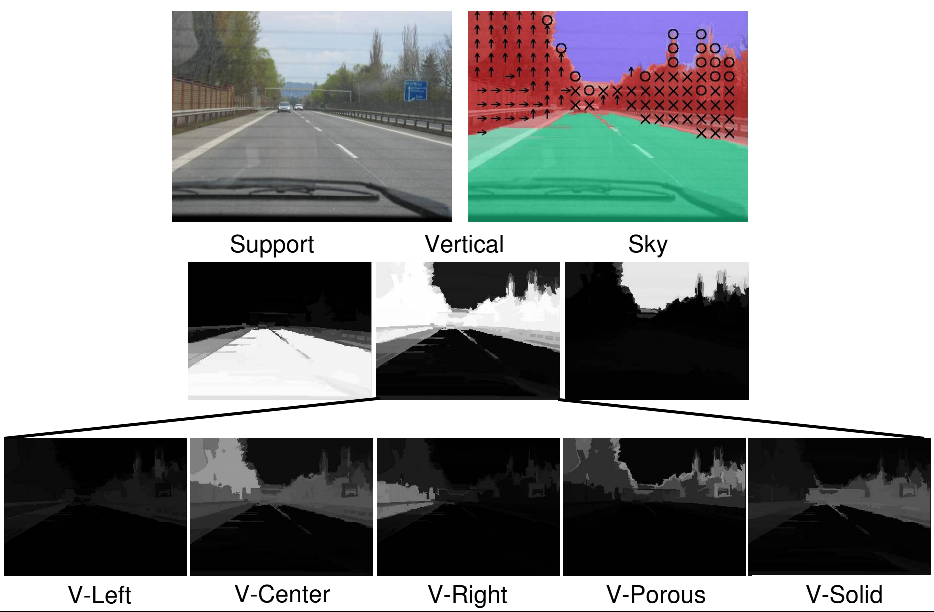
Labeled Segmentations

$$C(y_i = v|\mathbf{x}) = \sum_{j}^{n_h} P(y_j = v|\mathbf{x}, \mathbf{h}_{ji}) P(\mathbf{h}_{ji}|\mathbf{x})$$

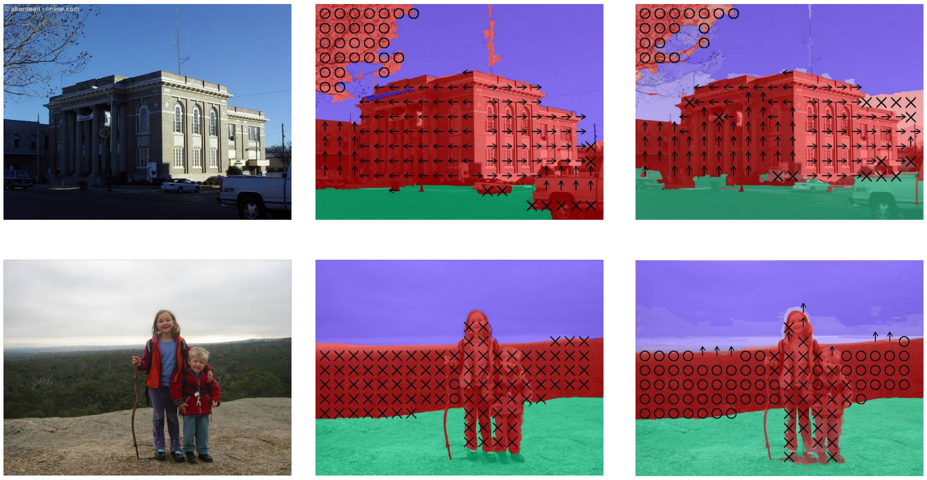
Learned from training images

Labeled Pixels

No Hard Decisions

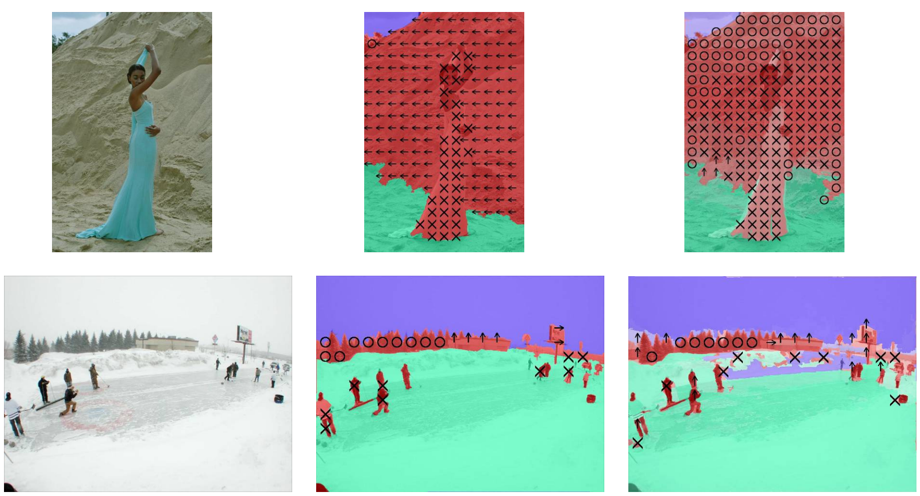


Labeling Results



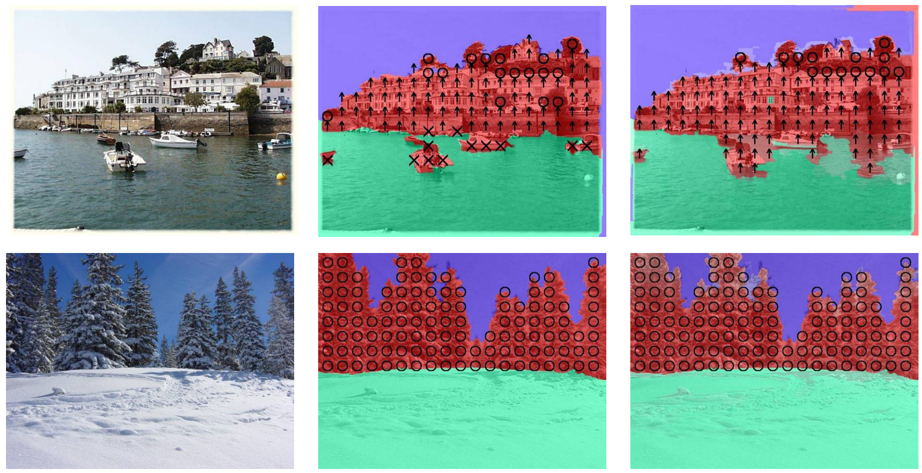
Input image Ground Truth Our Result

Labeling Results



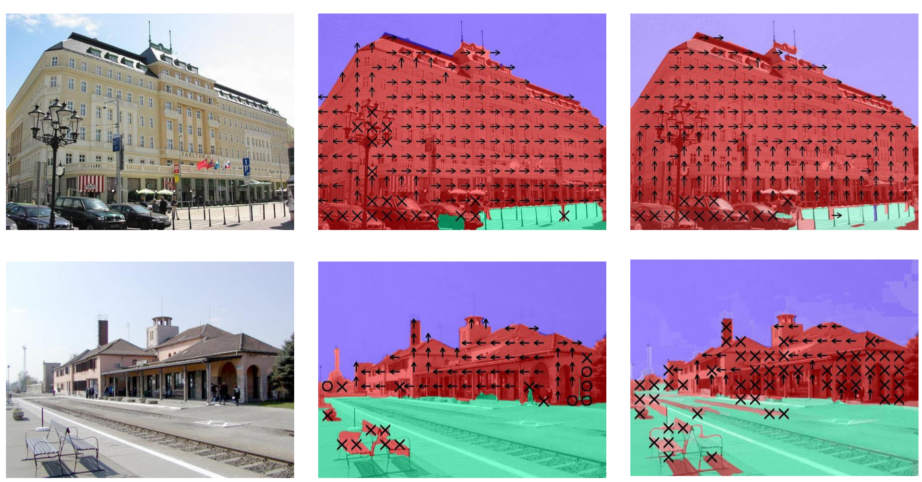
Input image Ground Truth Our Result

Labeling Results



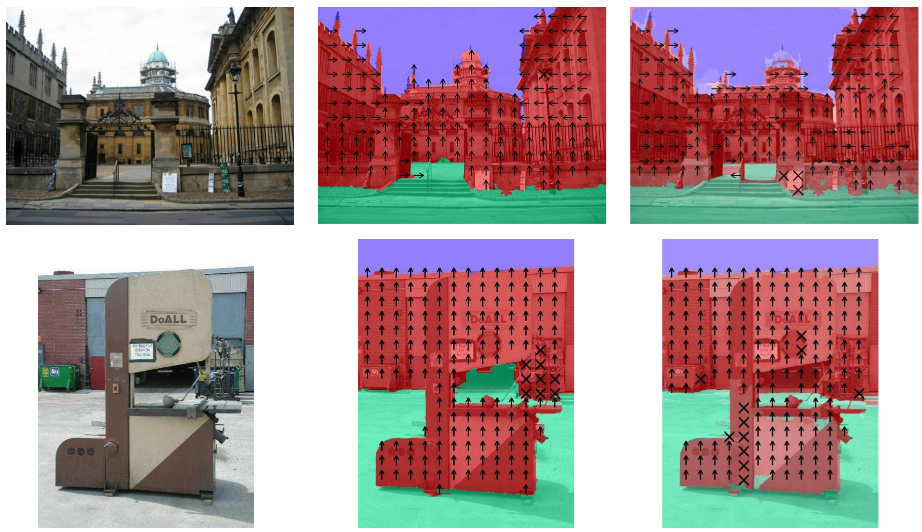
Input image | Ground Truth | Our Result

Labeling Results



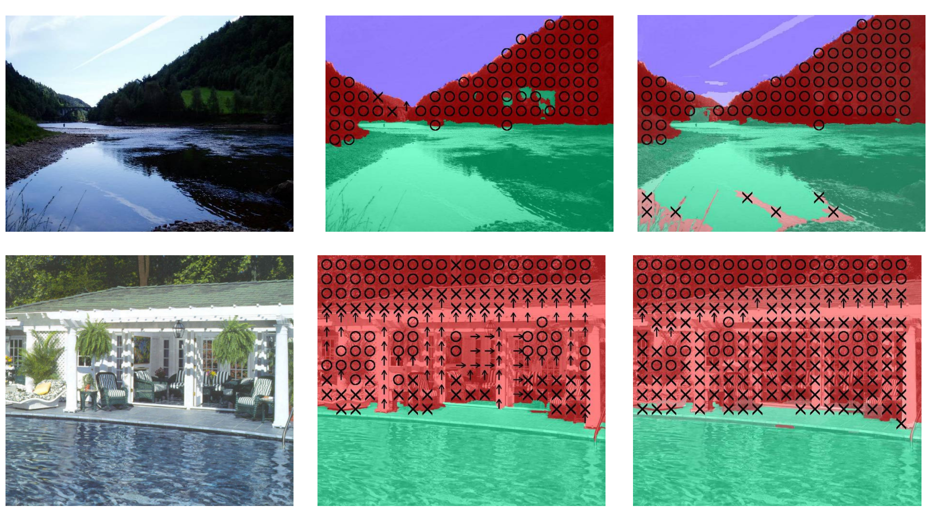
Input image | Ground Truth | Our Result

Labeling Results



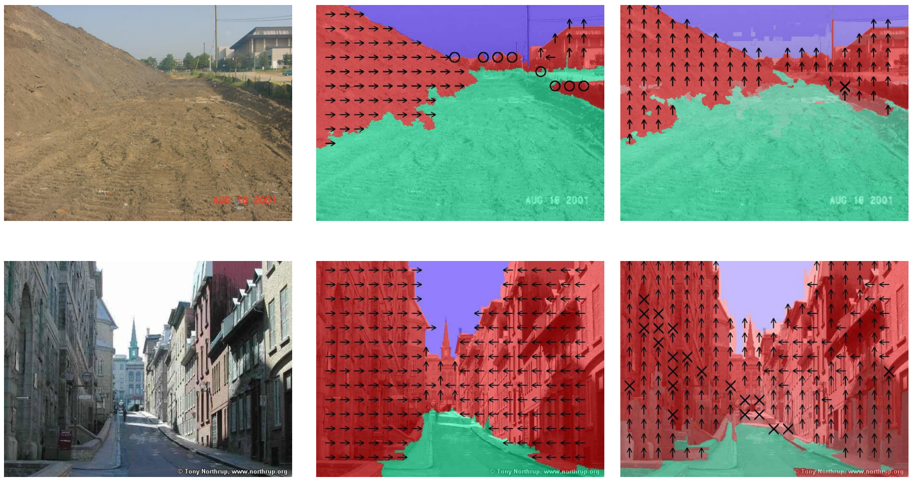
Input image | Ground Truth | Our Result

Labeling Results



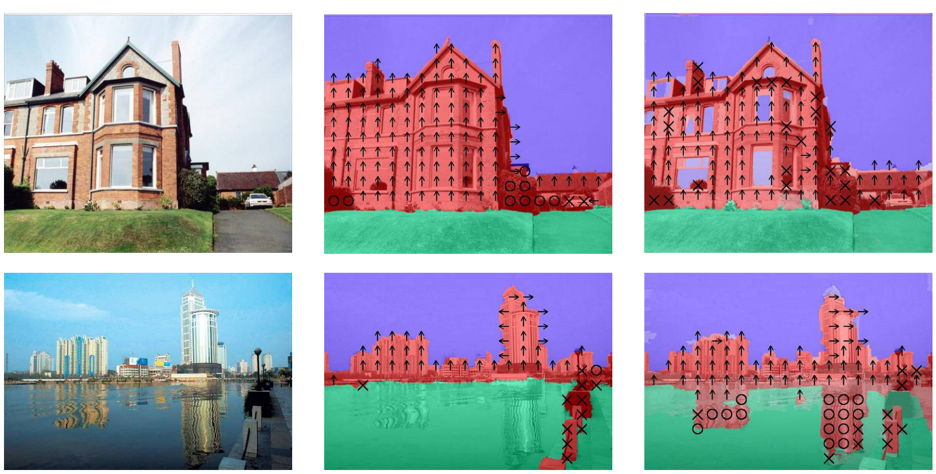
Input image | Ground Truth | Our Result

Labeling Results



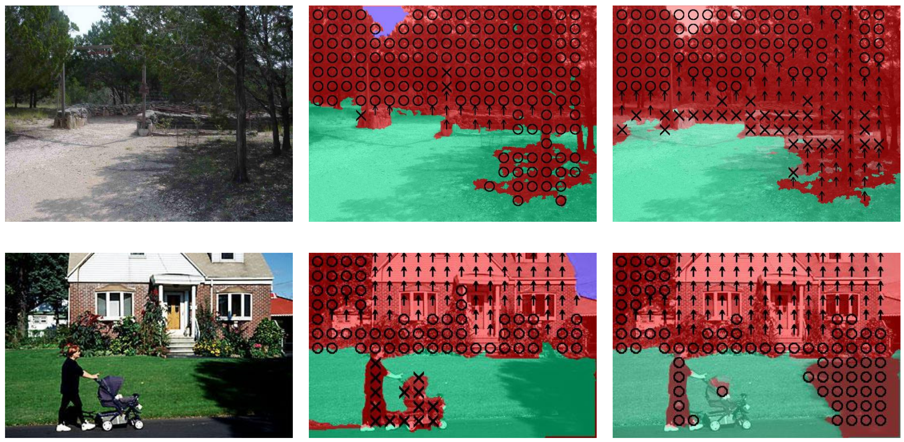
Input image | Ground Truth | Our Result

Reflection Failures



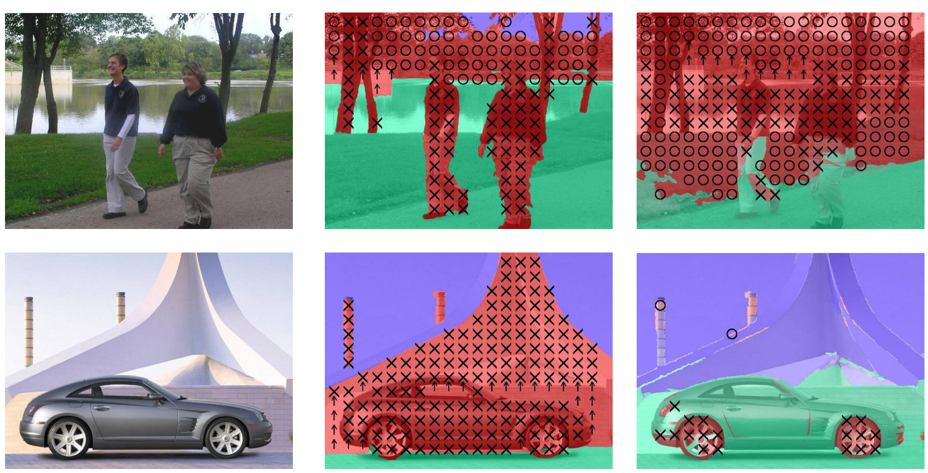
Input image | Ground Truth | Our Result

Shadows Failures



Input image Ground Truth Our Result

Catastrophic Failures



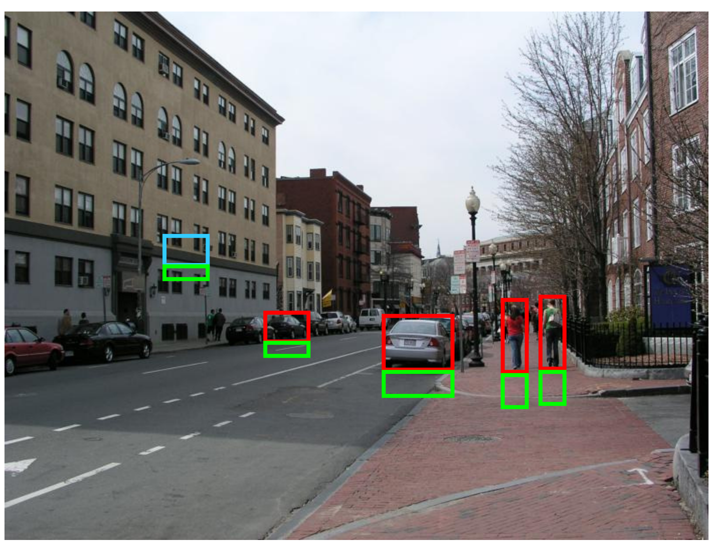
Input image Ground Truth Our Result

Quantitative Results

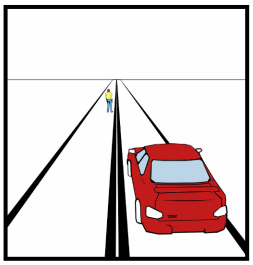
Geometric Class			
	Ground	Vertical	Sky
Ground	0.78	0.22	0.00
Vertical	0.09	0.89	0.02
Sky	0.00	0.10	0.90

Vertical Subclass					
	Left	Center	Right	Porous	Solid
Left	0.15	0.46	0.04	0.15	0.21
Center	0.02	0.55	0.06	0.19	0.18
Right	0.03	0.38	0.21	0.17	0.21
Porous	0.01	0.14	0.02	0.76	0.08
Solid	0.02	0.20	0.03	0.26	0.50

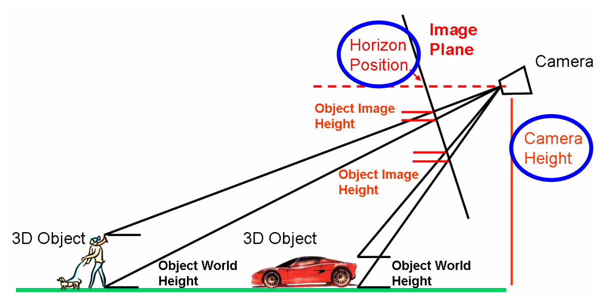
Object Support



Object Size in the Image



Image

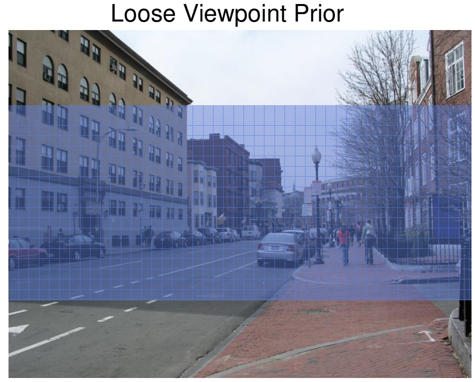


World

Object Size ↔ Camera Viewpoint

Input Image

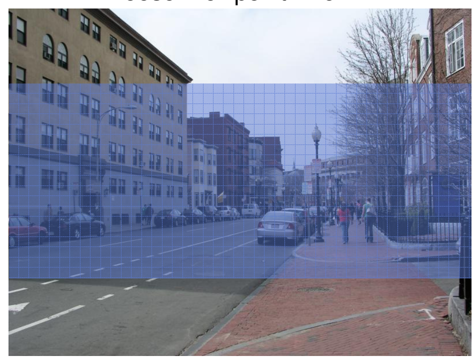
Loose Viewpoint Prior



Object Size ↔ Camera Viewpoint

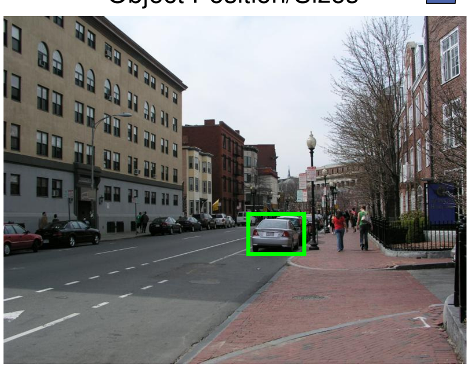
Input Image

Loose Viewpoint Prior



Object Size ↔ Camera Viewpoint

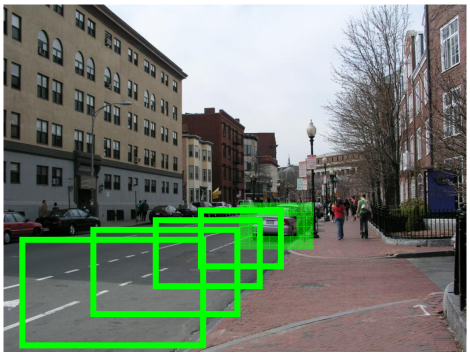
Object Position/Sizes



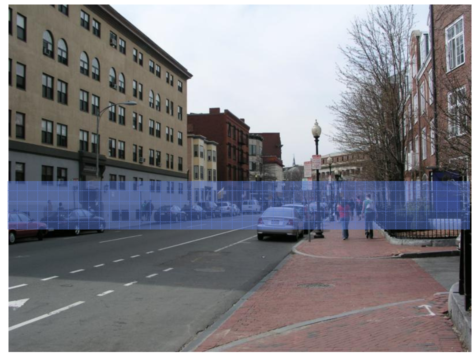
Viewpoint

Object Size ↔ Camera Viewpoint

Object Position/Sizes

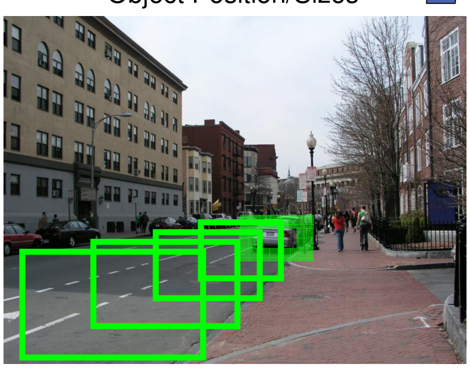


Viewpoint



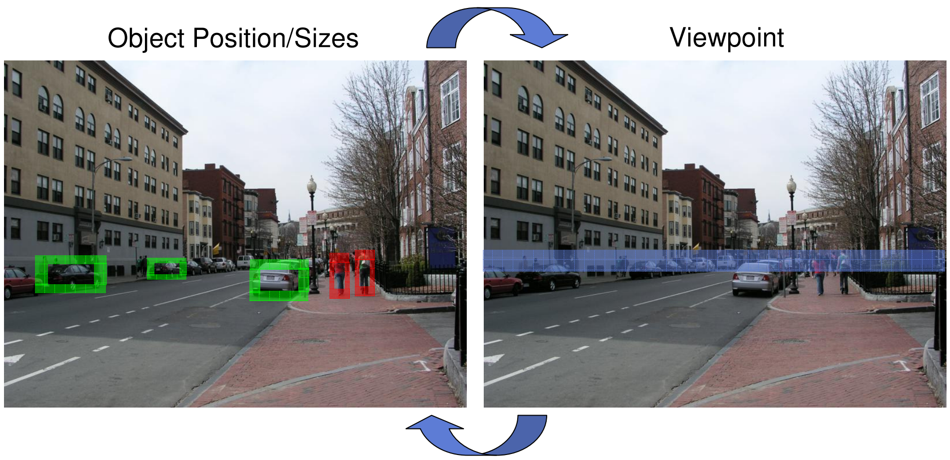
Object Size ↔ Camera Viewpoint

Object Position/Sizes

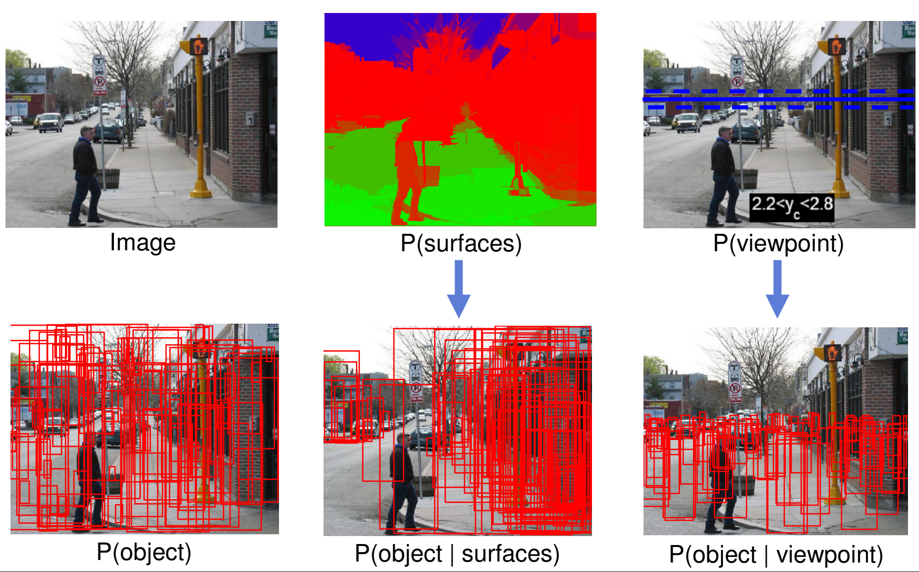


Viewpoint

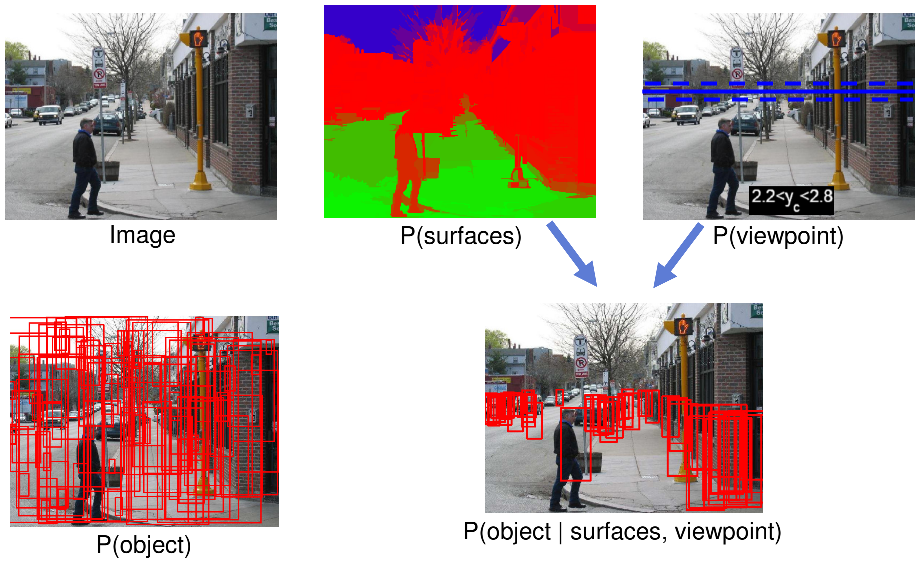
Object Size ↔ Camera Viewpoint



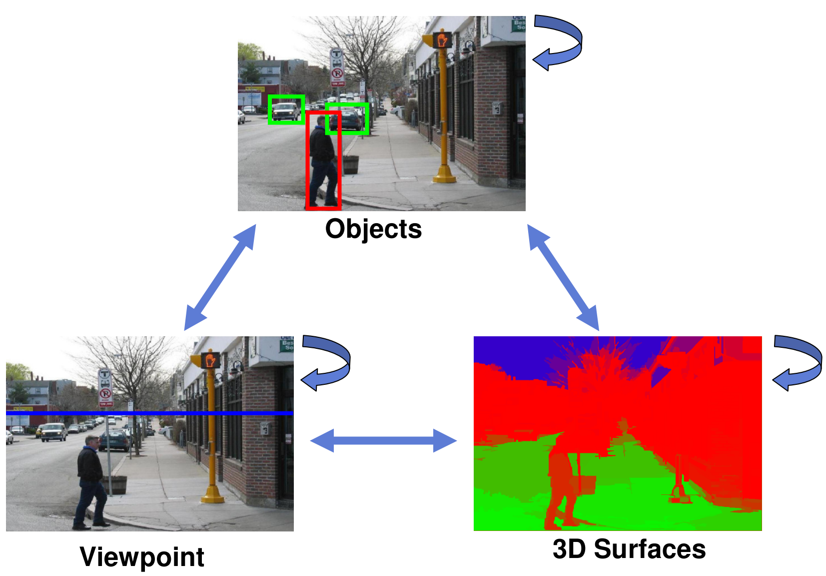
What does surface and viewpoint say about objects?



What does surface and viewpoint say about objects?

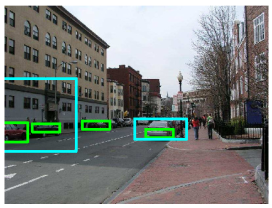


Scene Parts Are All Interconnected

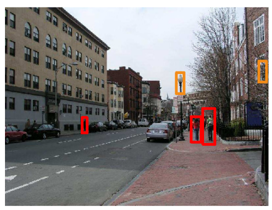


Input to Our Algorithm

Object Detection

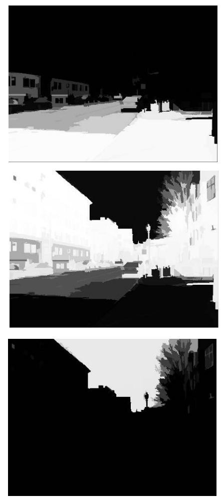


Local Car Detector

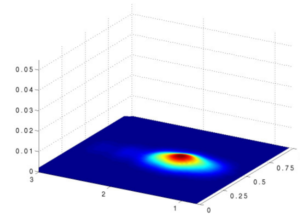


Local Ped Detector

Surface Estimates



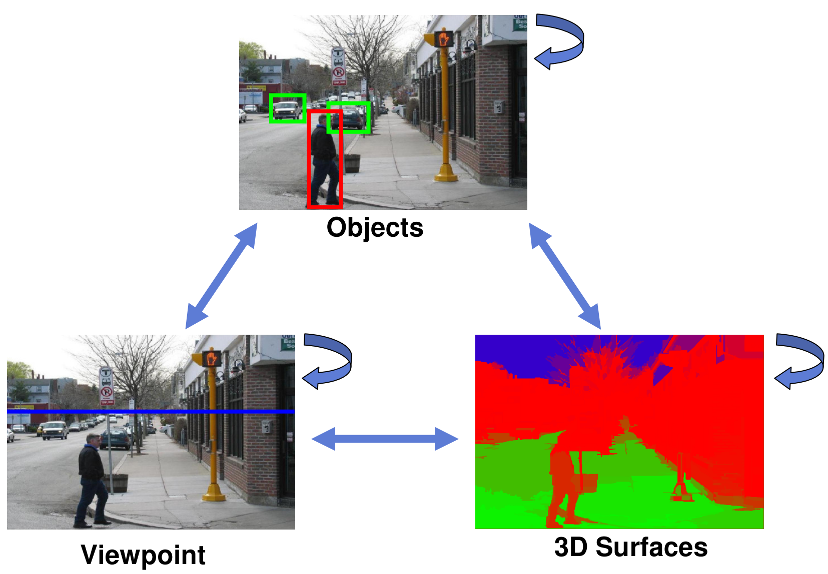
Viewpoint Prior



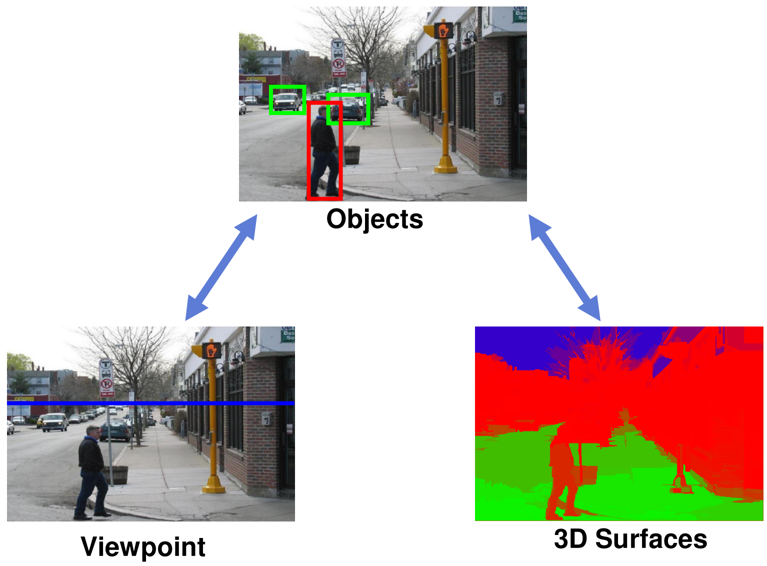
Local Detector: [Dalal-Triggs 2005]

Surfaces: [Hoiem-Efros-Hebert 2005]

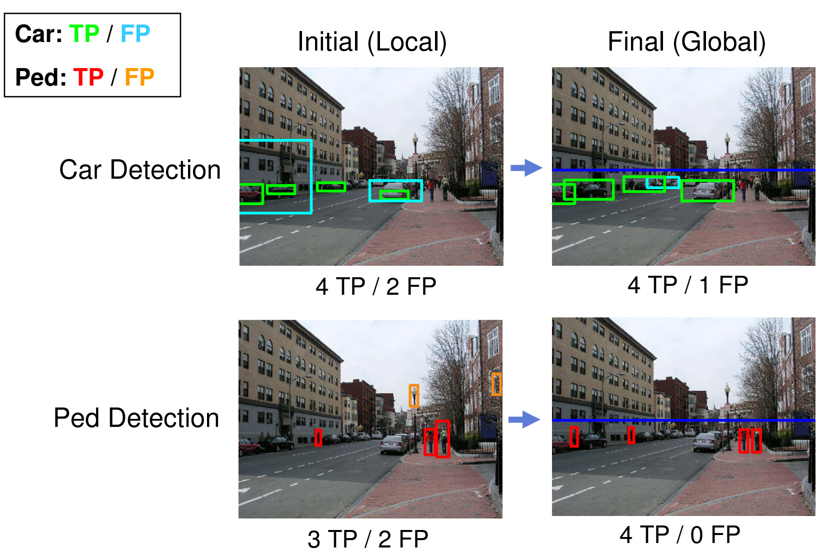
Scene Parts Are All Interconnected



Our Approximate Model (solve by BP)

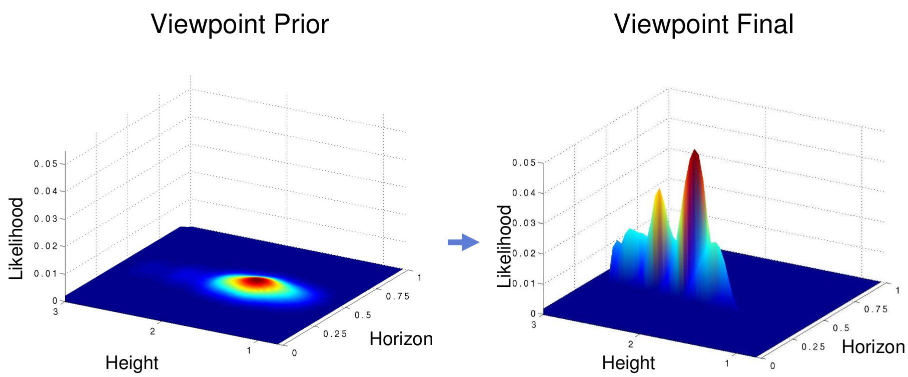


After Inference



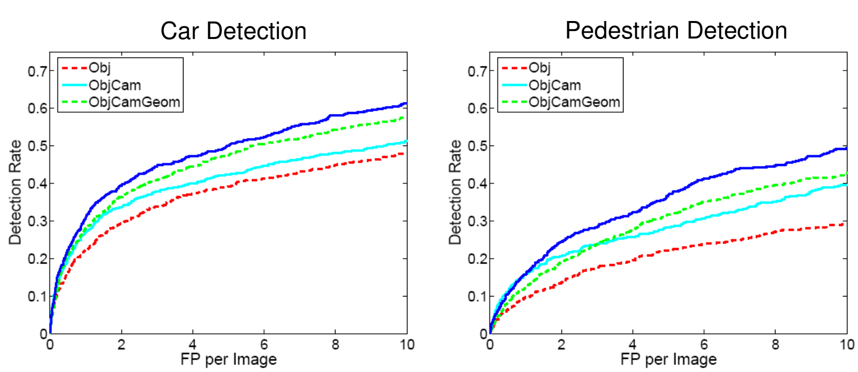
Local Detector: [Dalal-Triggs 2005]

After Inference



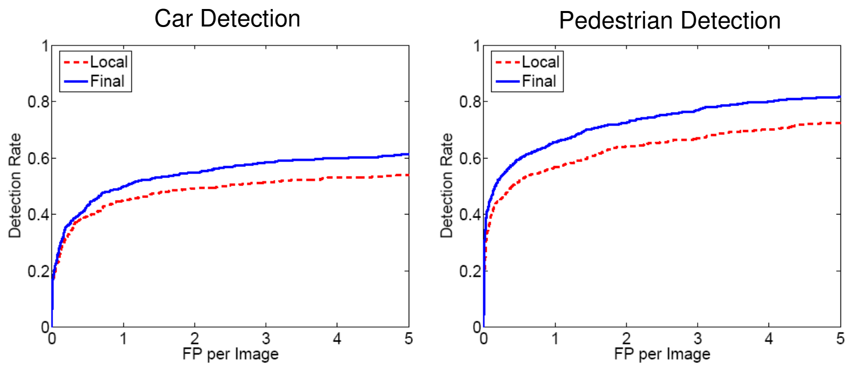
Each piece of evidence improves performance

- Testing with LabelMe dataset: 422 images
  - 923 Cars at least 14 pixels tall
  - 720 Peds at least 36 pixels tall



Local Detector from [Murphy-Torralba-Freeman 2003]

Can be used with any detector that outputs confidences

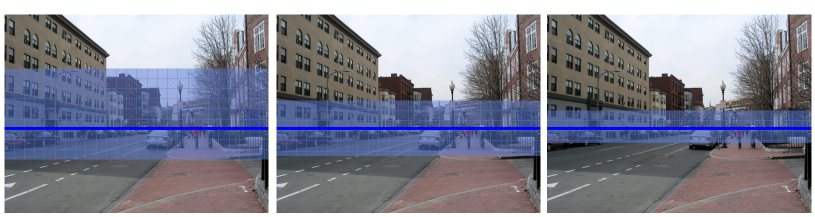


Local Detector: [Dalal-Triggs 2005] (SVM-based)

Accurate Horizon Estimation

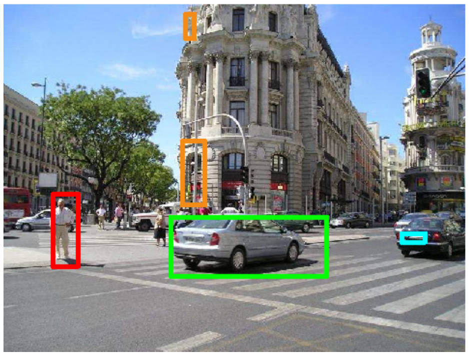
	Horizon Prior	[Murphy-Torralba-Freeman 2003]	[Dalal-Triggs 2005]
Median Error:	8.5%	4.5%	3.0%

90% Bound:

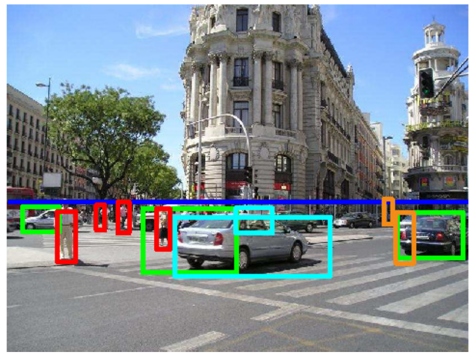


Qualitative Results

Car: TP / FP Ped: TP / FP



Initial: 2 TP / 3 FP

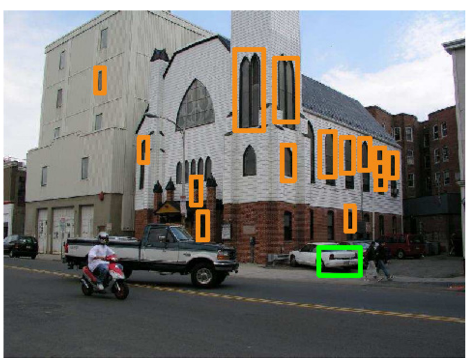


Final: 7 TP / 4 FP

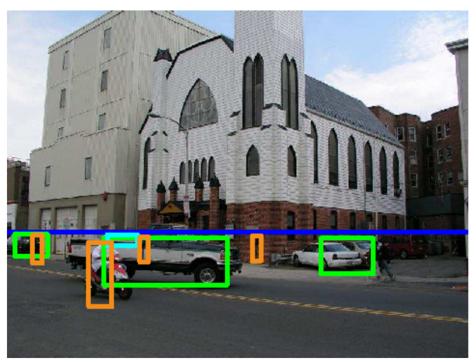
Local Detector from [Murphy-Torralba-Freeman 2003]

Qualitative Results

Car: TP / FP Ped: TP / FP



Initial: 1 TP / 14 FP

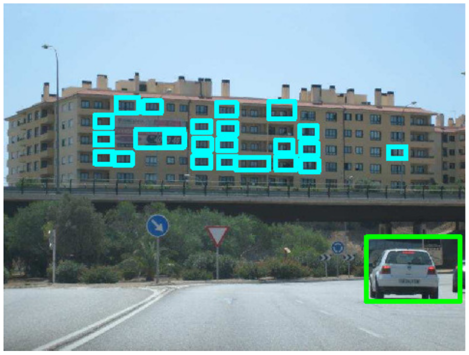


Final: 3 TP / 5 FP

Local Detector from [Murphy-Torralba-Freeman 2003]

Qualitative Results

Car: TP / FP Ped: TP / FP



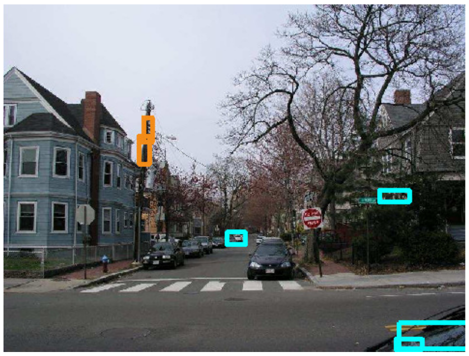
Initial: 1 TP / 23 FP

Final: 0 TP / 10 FP

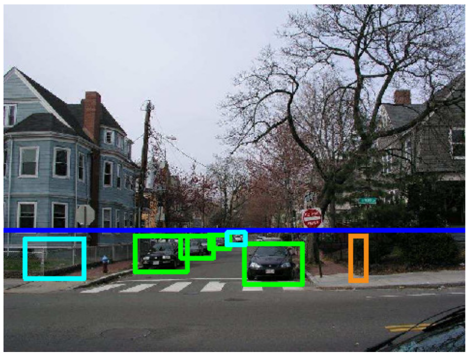
Local Detector from [Murphy-Torralba-Freeman 2003]

Qualitative Results

Car: TP / FP Ped: TP / FP



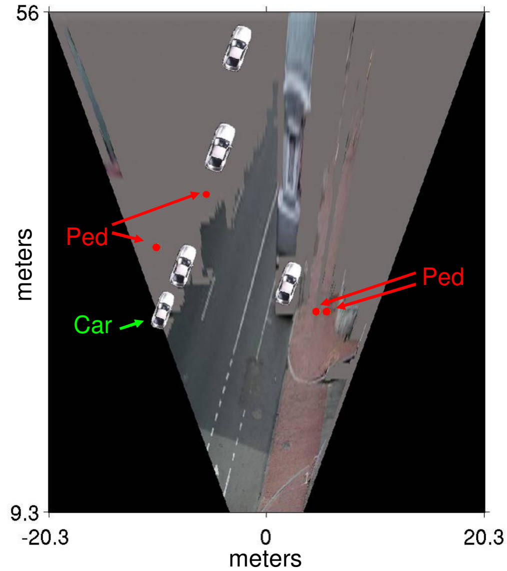
Initial: 0 TP / 6 FP



Final: 4 TP / 3 FP

Local Detector from [Murphy-Torralba-Freeman 2003]

Reasoning in 3D



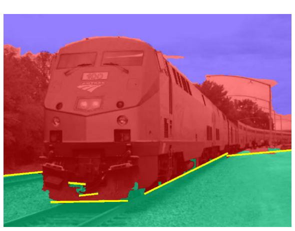
Future Work:

- Object to object
- Scene label
- Object segmentation

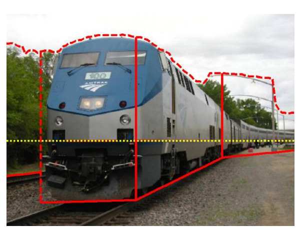
Automatic Photo Pop-up

Original Image

Geometric Labels



Fit Segments



Cut and Fold

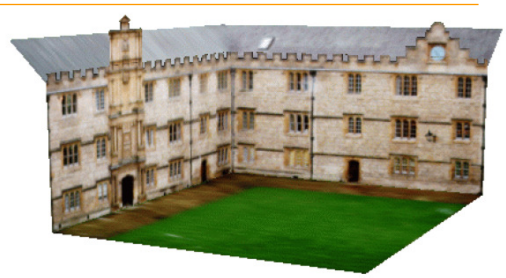
Novel View

More Pop-ups

More Pop-ups

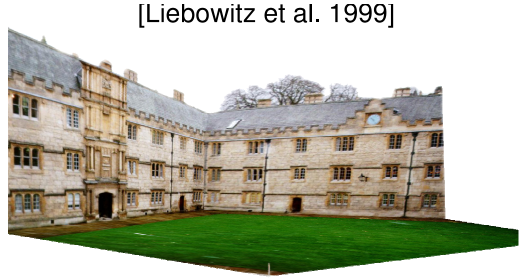
More Pop-ups

Comparison with Manual Method



[Liebowitz et al. 1999]

Input Image



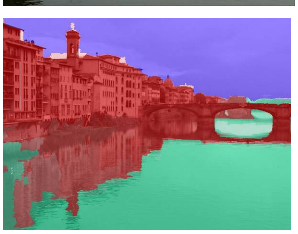
Automatic Photo Pop-up (30 sec)!

Disclaimer

- Gives reasonable model about 25-35% of the time
- Failures due to:
  - Labeling error
  - Bad ground-fitting
  - Modeling assumptions
  - Occlusions in image
  - Bad horizon estimates

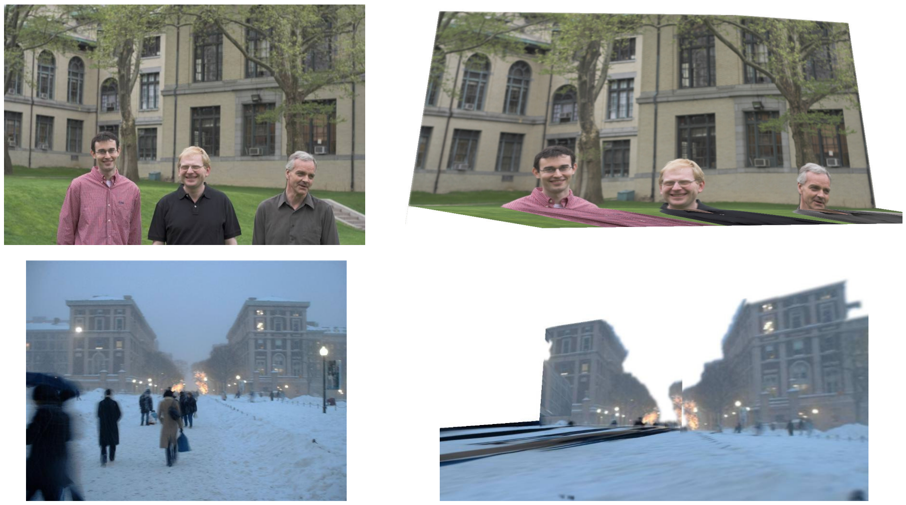
Failures

Labeling Errors

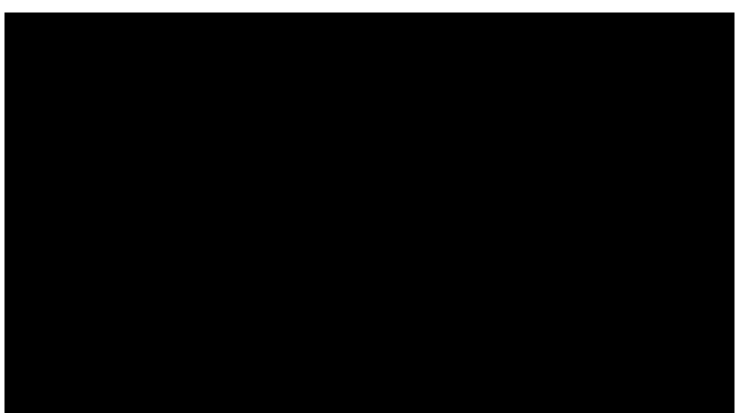


Failures

Foreground Objects



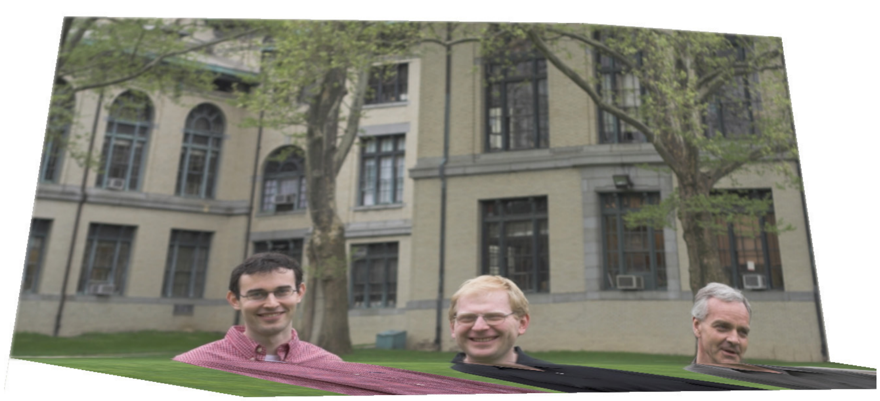
The Music Video



Conclusions

- Our ultimate goal is to understand the whole image
- We use data: explaining each image segment with something we have seen before
- Better understanding of the scene helps to recognize objects.

Thank you

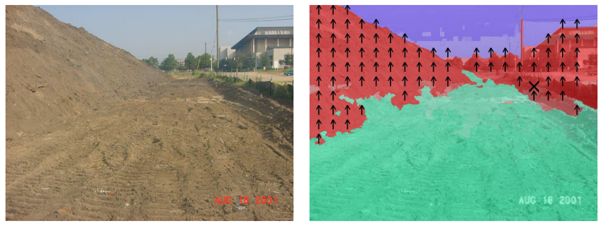


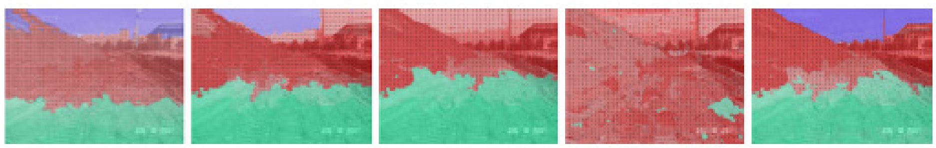
Questions?

Do all features help?

Importance of Different Feature Types				
	Color	Texture	Loc/Shape	Geometry
Main	6%	2%	16%	2%
Sub	6%	2%	8%	7%

Drop in accuracy due to remove of each type of feature





(c) Loc Only (d) No Color (e) No Texture (f) No Loc/Shp (g) No Geom

Does Better Spatial Support Help?

Intermediate Structure Estimation						
	CPrior	Loc	Pixel	SPixel	OneH	MultiH
Main	49%	66%	80%	83%	83%	86%
Sub	34%	36%	43%	45%	44%	52%

With "perfect" structure estimation:

- 95% accuracy for main classes
- 66% accuracy for subclasses