CS 558: Computer Vision
13th Set of Notes

Instructor: Philippos Mordohai
Webpage: www.cs.stevens.edu/~mordohai
E-mail: Philippos.Mordohai@stevens.edu
Office: Lieb 215

Overview

- Context and Spatial Layout
- Relating Objects and Geometry
- Putting Objects in Perspective
- Interpretation of indoor scenes

  – Based on slides by D. Hoiem

Context in Recognition

- Objects usually are surrounded by a scene that can provide context in the form of nearby objects, surfaces, scene category, geometry, etc.

Context provides clues for function

- What is this?

Examples from Antonio Torralba

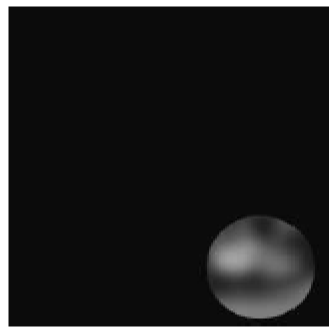
Context provides clues for function

- What is this?

- Now can you tell?

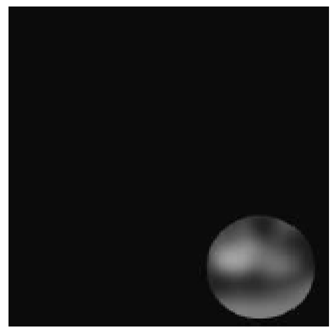
Sometimes context is *the* major component of recognition

- What is this?



Sometimes context is *the* major component of recognition

- What is this?



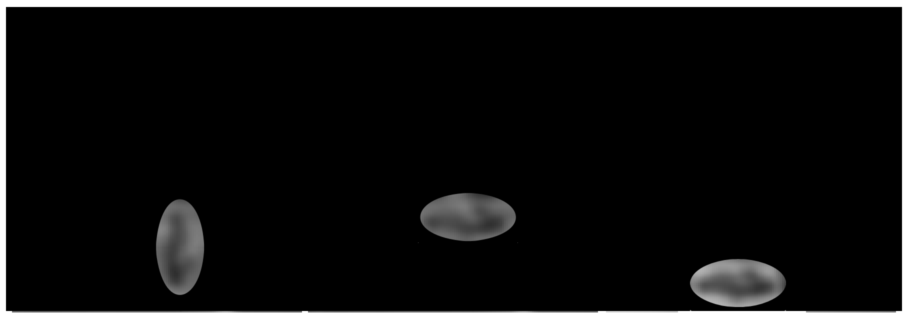
- Now can you tell?

More Low-Res

- What are these blobs?

More Low-Res

- The same pixels! (a car)

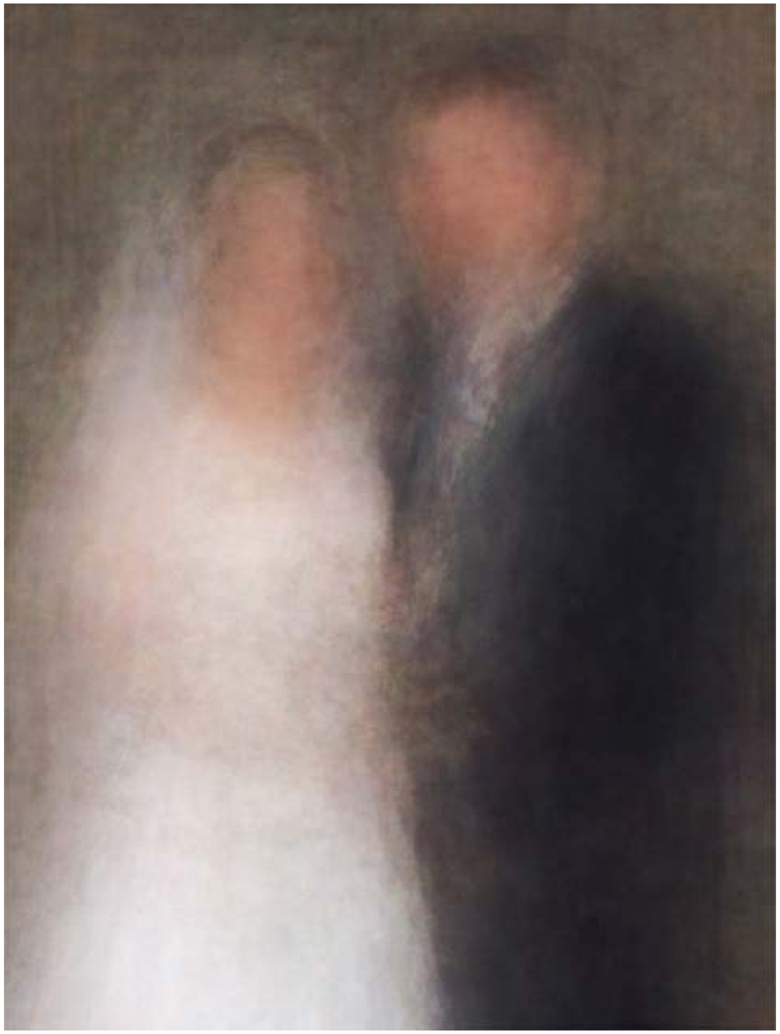


There are many types of context

- **Local pixels**
  - window, surround, image neighborhood, object boundary/shape, global image statistics
- **2D Scene Gist**
  - global image statistics
- **3D Geometric**
  - 3D scene layout, support surface, surface orientations, occlusions, contact points, etc.
- **Semantic**
  - event/activity depicted, scene category, objects present in the scene and their spatial extents, keywords
- **Photogrammetric**
  - camera height orientation, focal length, lens distortion, radiometric, response function
- **Illumination**
  - sun direction, sky color, cloud cover, shadow contrast, etc.
- **Geographic**
  - GPS location, terrain type, land use category, elevation, population density, etc.
- **Temporal**
  - nearby frames of video, photos taken at similar times, videos of similar scenes, time of capture
- **Cultural**
  - photographer bias, dataset selection bias, visual clichés, etc.

from Divvala et al. CVPR 2009

Cultural context



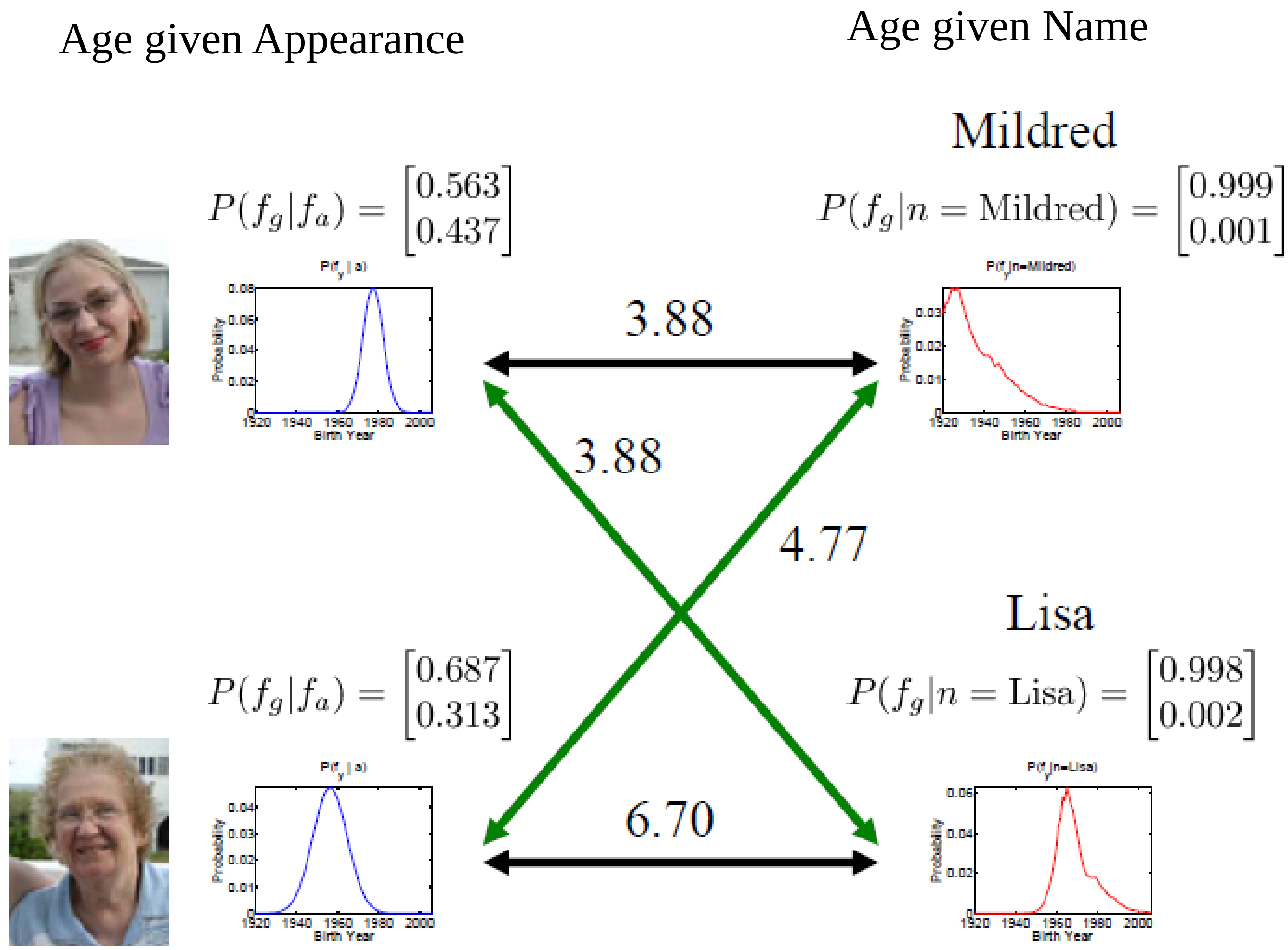
Jason Salavon: http://salavon.com/SpecialMoments/Newlyweds.shtml

Cultural context

"Mildred and Lisa": Who is Mildred? Who is Lisa?

Andrew Gallagher: http://chenlab.ece.cornell.edu/people/Andy/projectpage_names.html

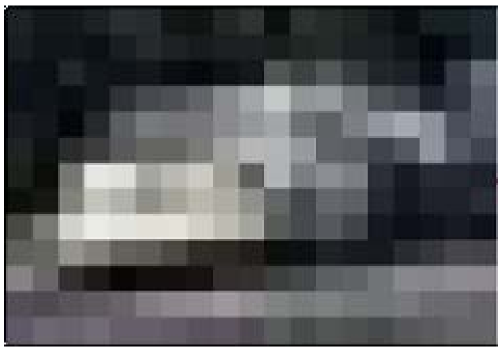
Cultural context

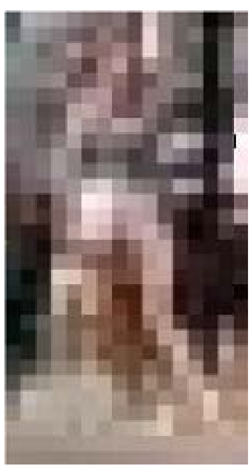


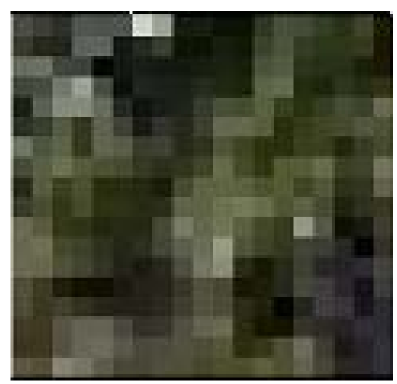
Andrew Gallagher: http://chenlab.ece.cornell.edu/people/Andy/projectpage_names.html

Spatial layout is especially important

1. Context for recognition

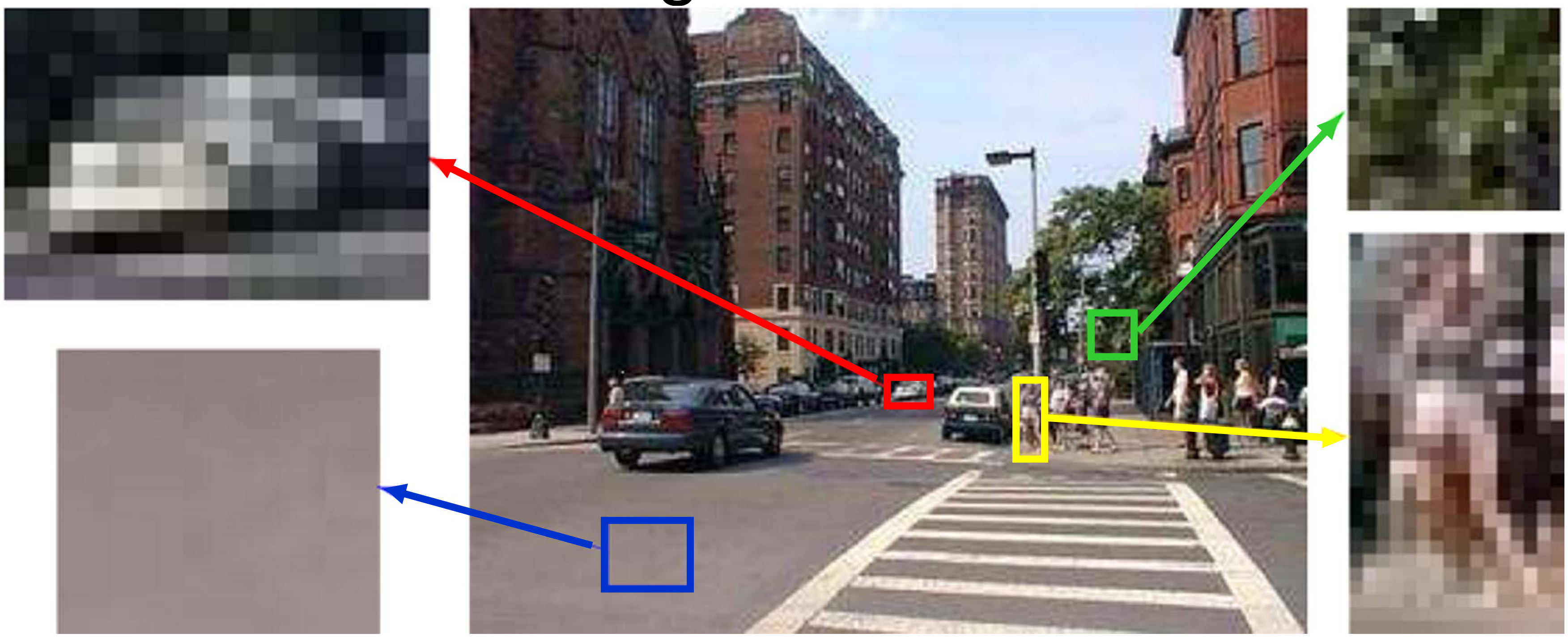






Spatial layout is especially important

1. Context for recognition



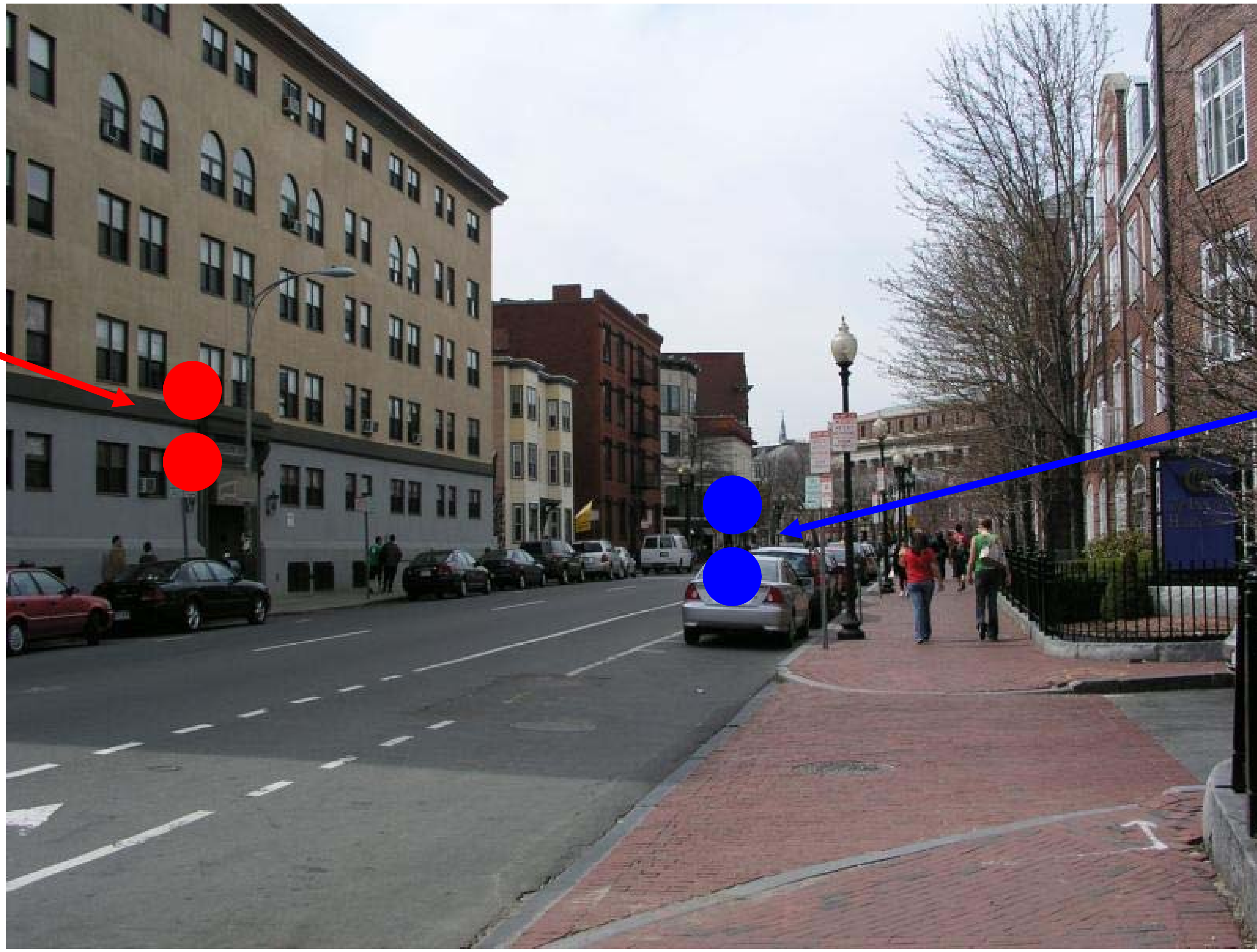
Spatial layout is especially important

1. Context for recognition
2. Scene understanding

Spatial Layout: 2D vs. 3D

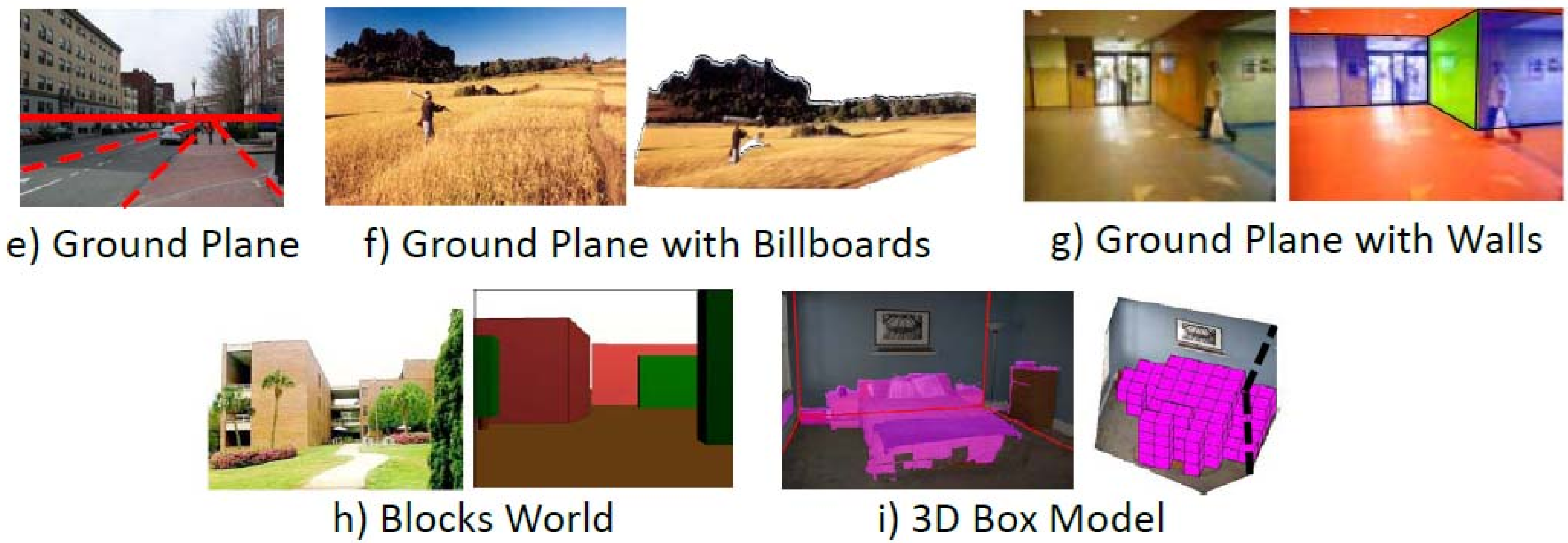
But object relations are in 3D…

Close



Not Close

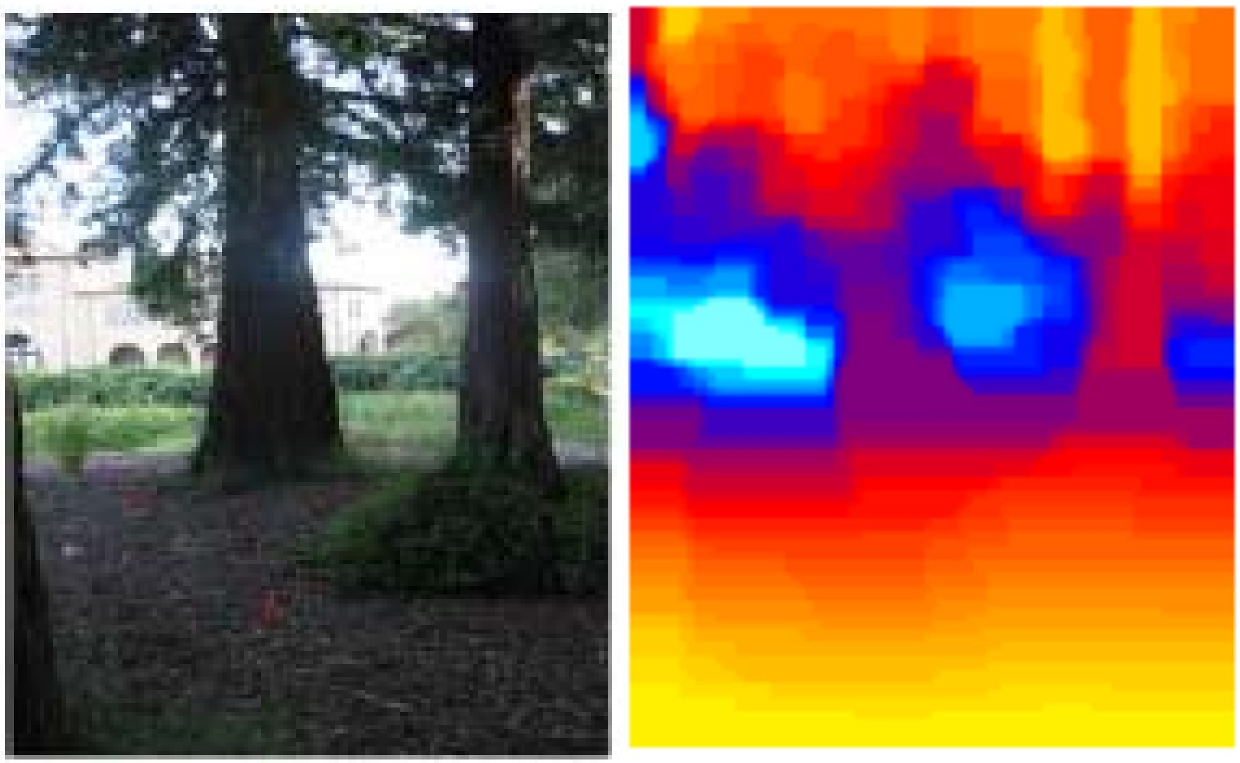
Highly Structured 3D Models



Figs from Hoiem - Savarese 2011 book

High detail, Low abstraction

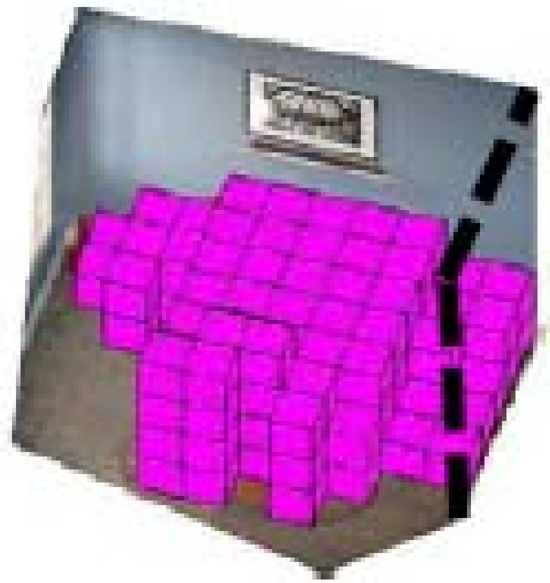
Depth Map



Saxena, Chung & Ng 2005, 2007

Medium detail, High abstraction

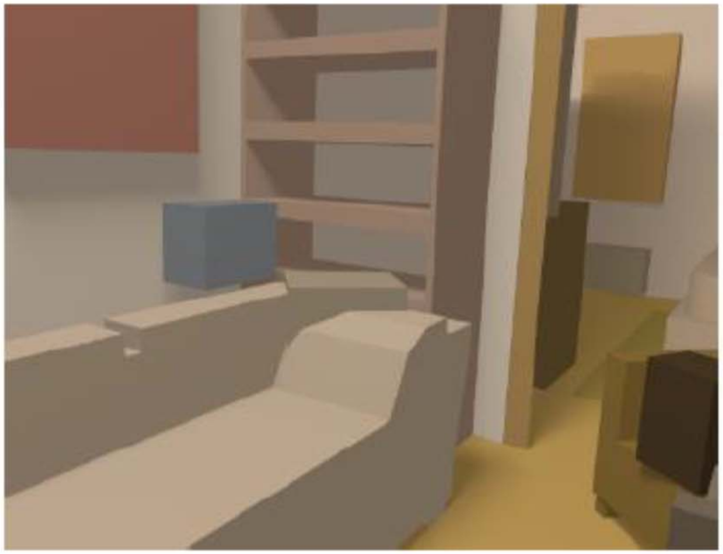
Room as a Box

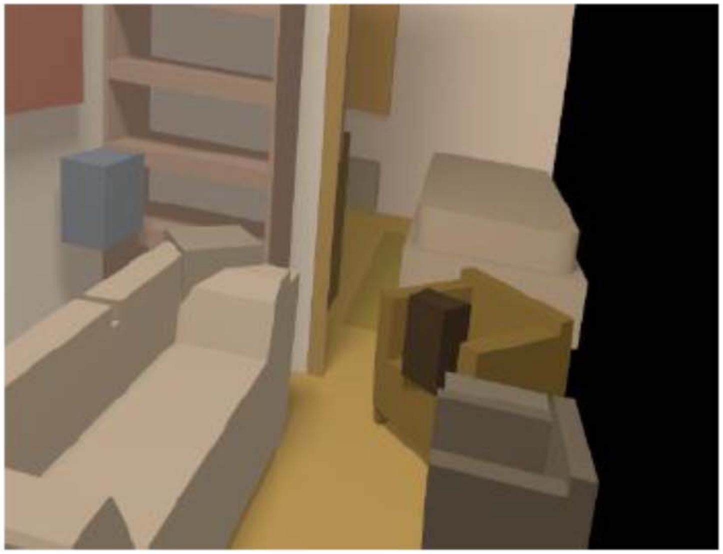


Hedau Hoiem Forsyth 2009

Med-High detail, High abstraction

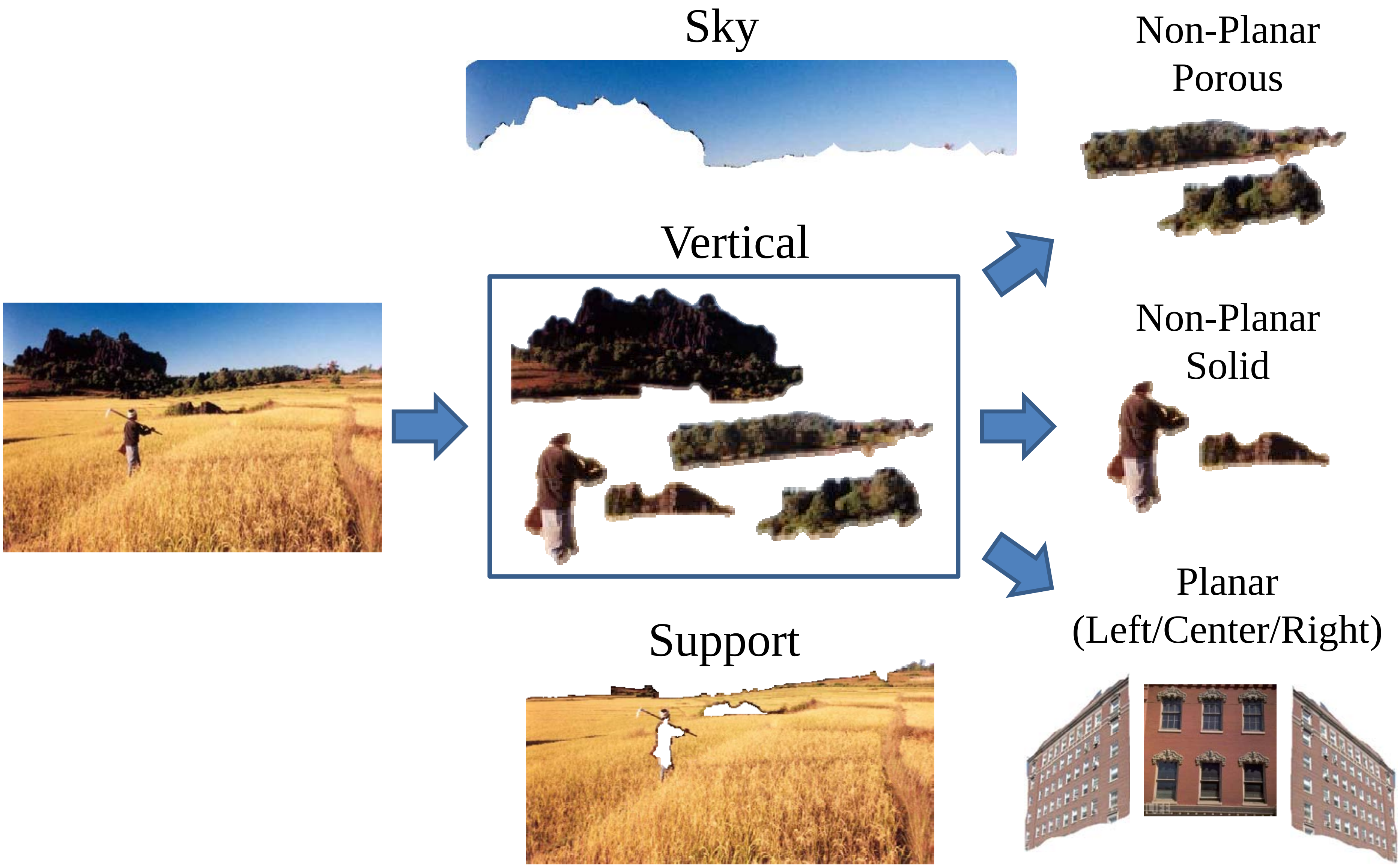
Complete 3D Layout



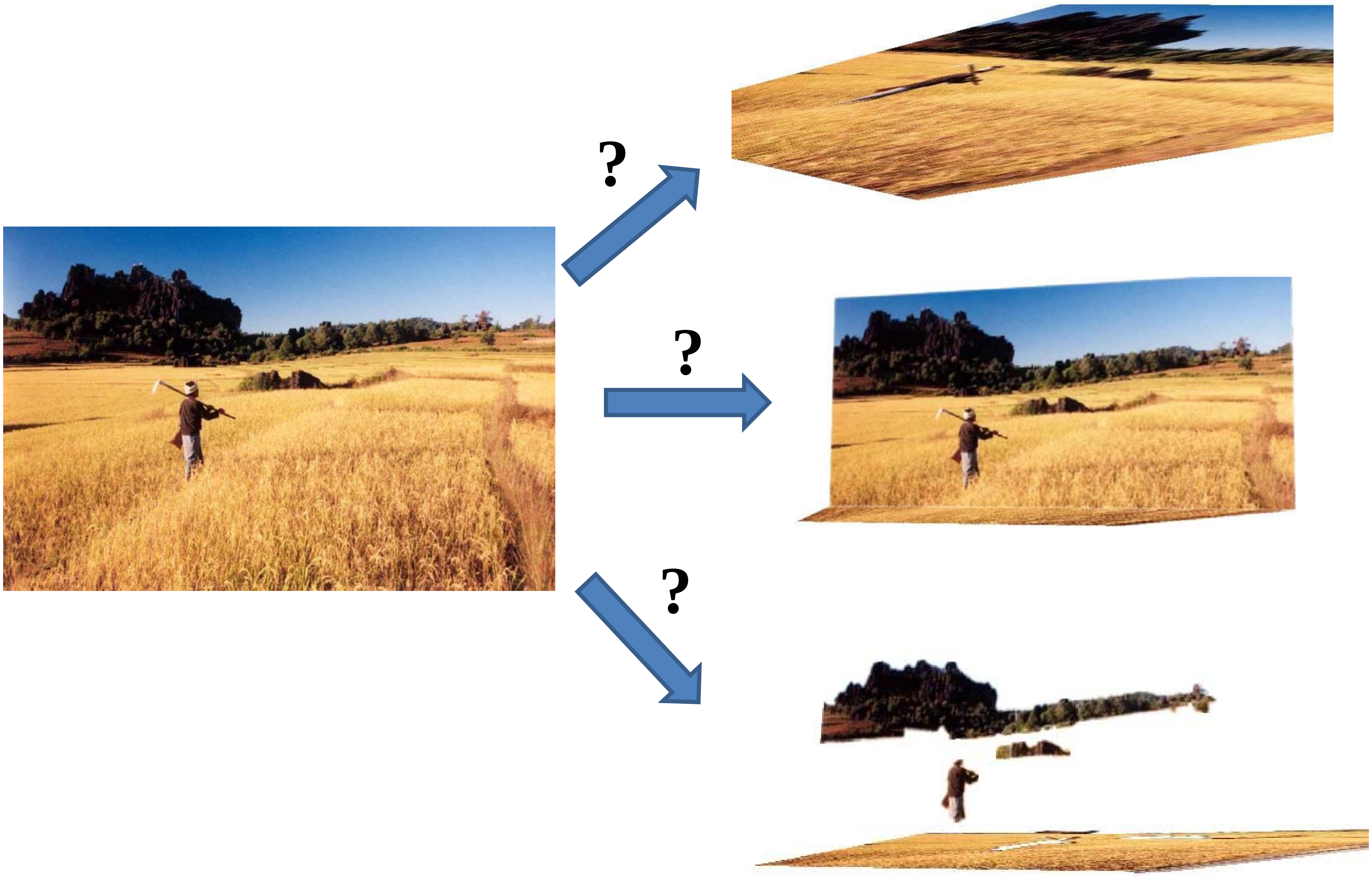


Guo Zou Hoiem 2015

Surface Layout: describe 3D surfaces with geometric classes



The challenge



Our World is Structured

Abstract World

Our World

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

Learn the Structure of the World

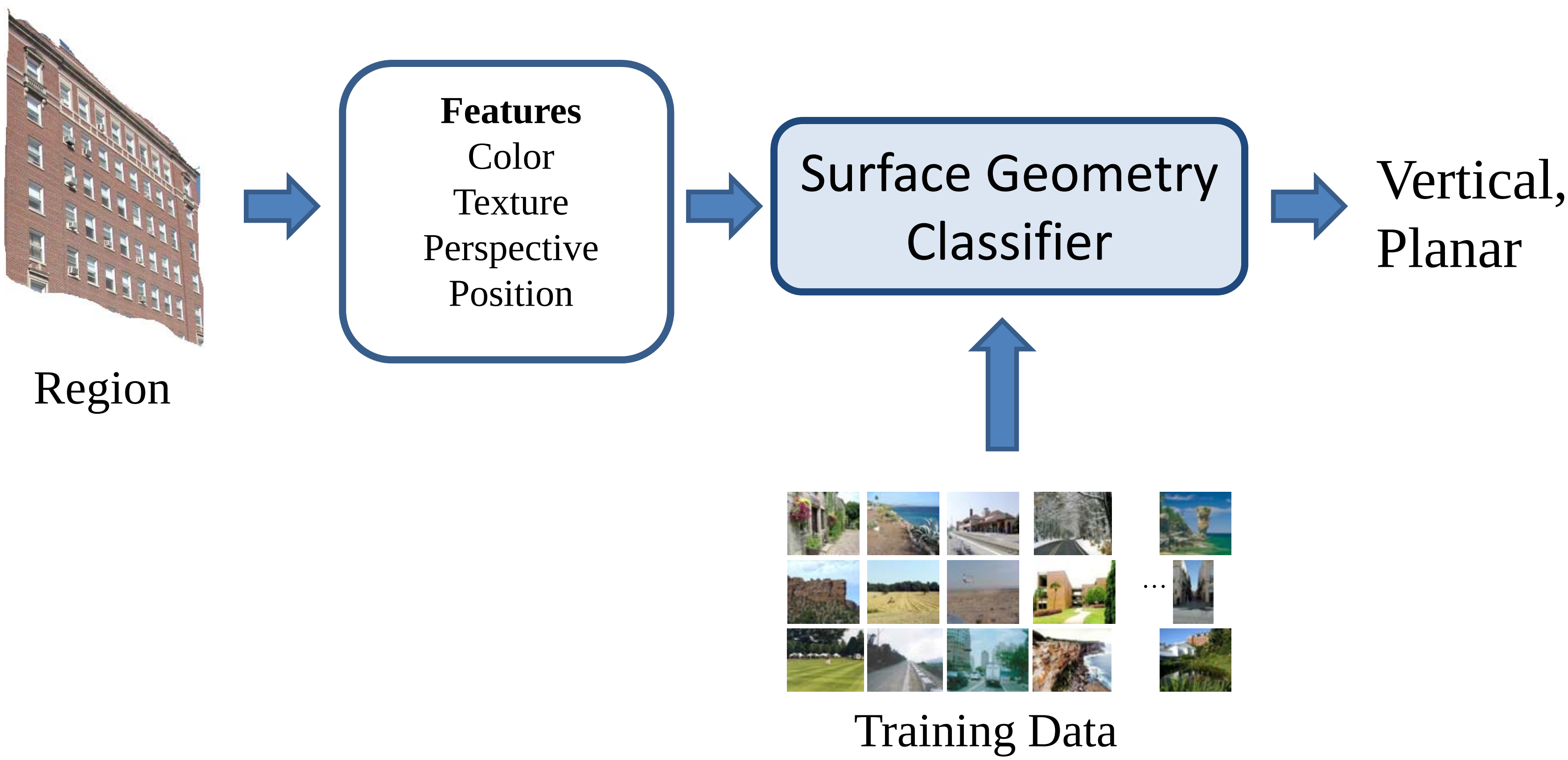
Training Images

Infer the most likely interpretation

Unlikely

Likely

Geometry estimation as recognition



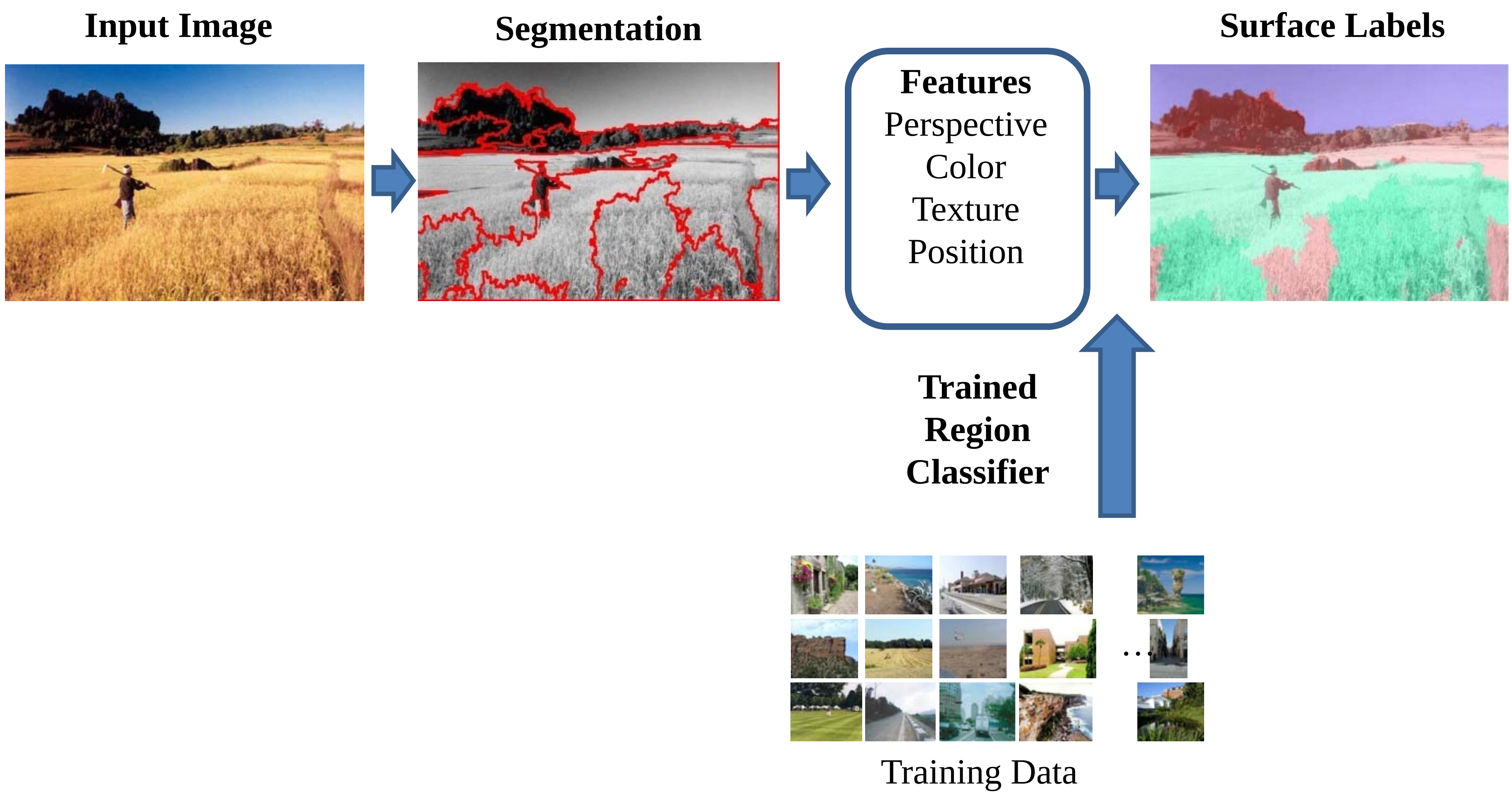
Use a variety of image cues

Vanishing points, lines

Color, texture, image location

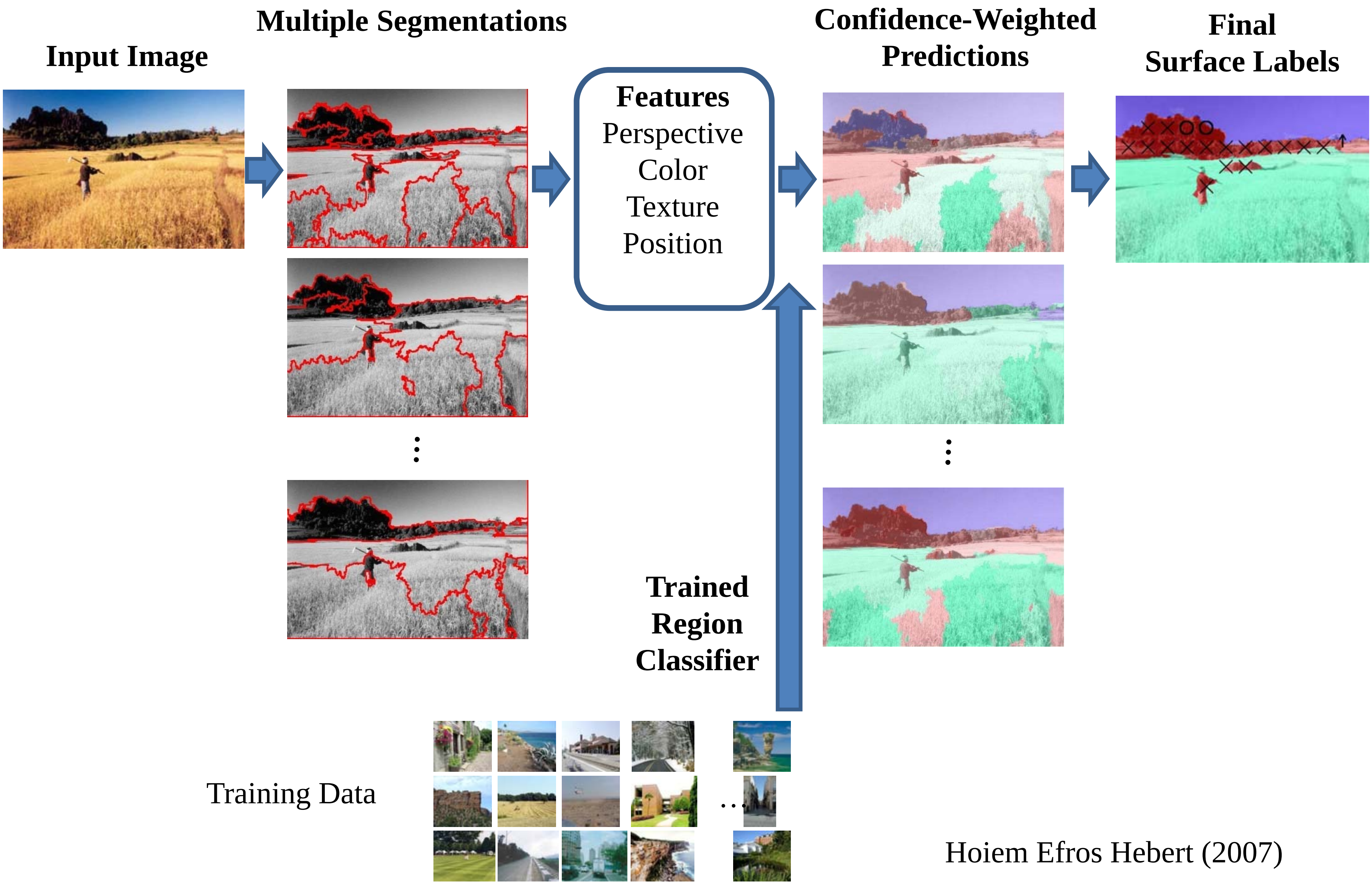
Texture gradient

Surface Layout Algorithm



Hoiem Efros Hebert (2007)

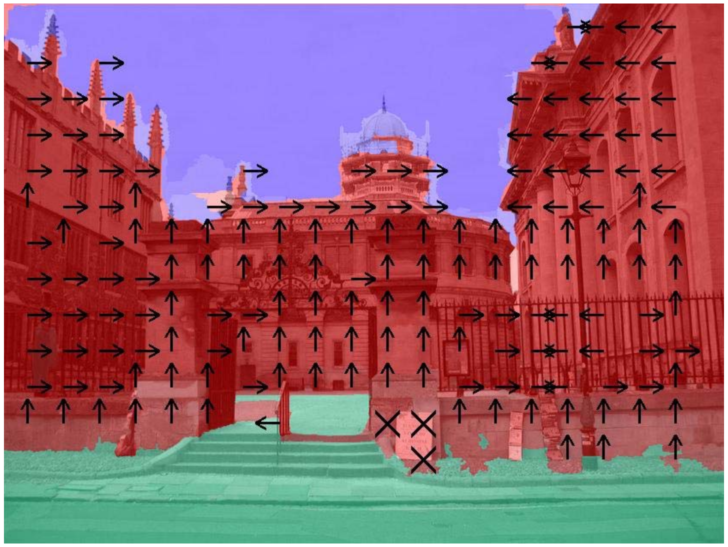
Surface Layout Algorithm



Hoiem Efros Hebert (2007)

Geometric Classes

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



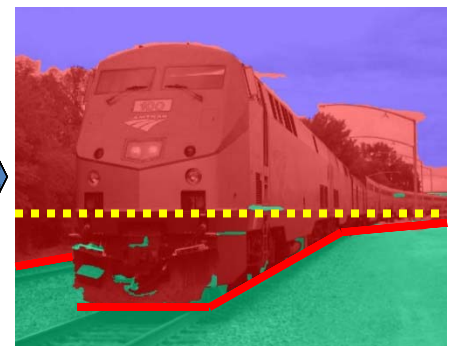
Surface Description Result

Automatic Photo Popup

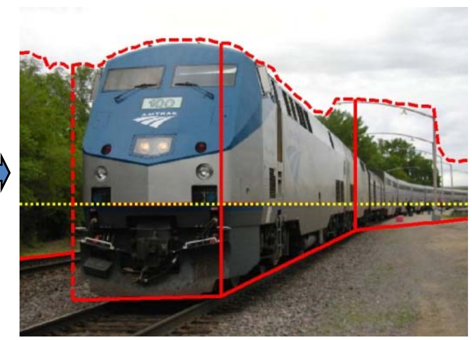
Labeled Image

Fit Ground-Vertical Boundary with Line Segments

Form Segments into Polylines



Cut and Fold

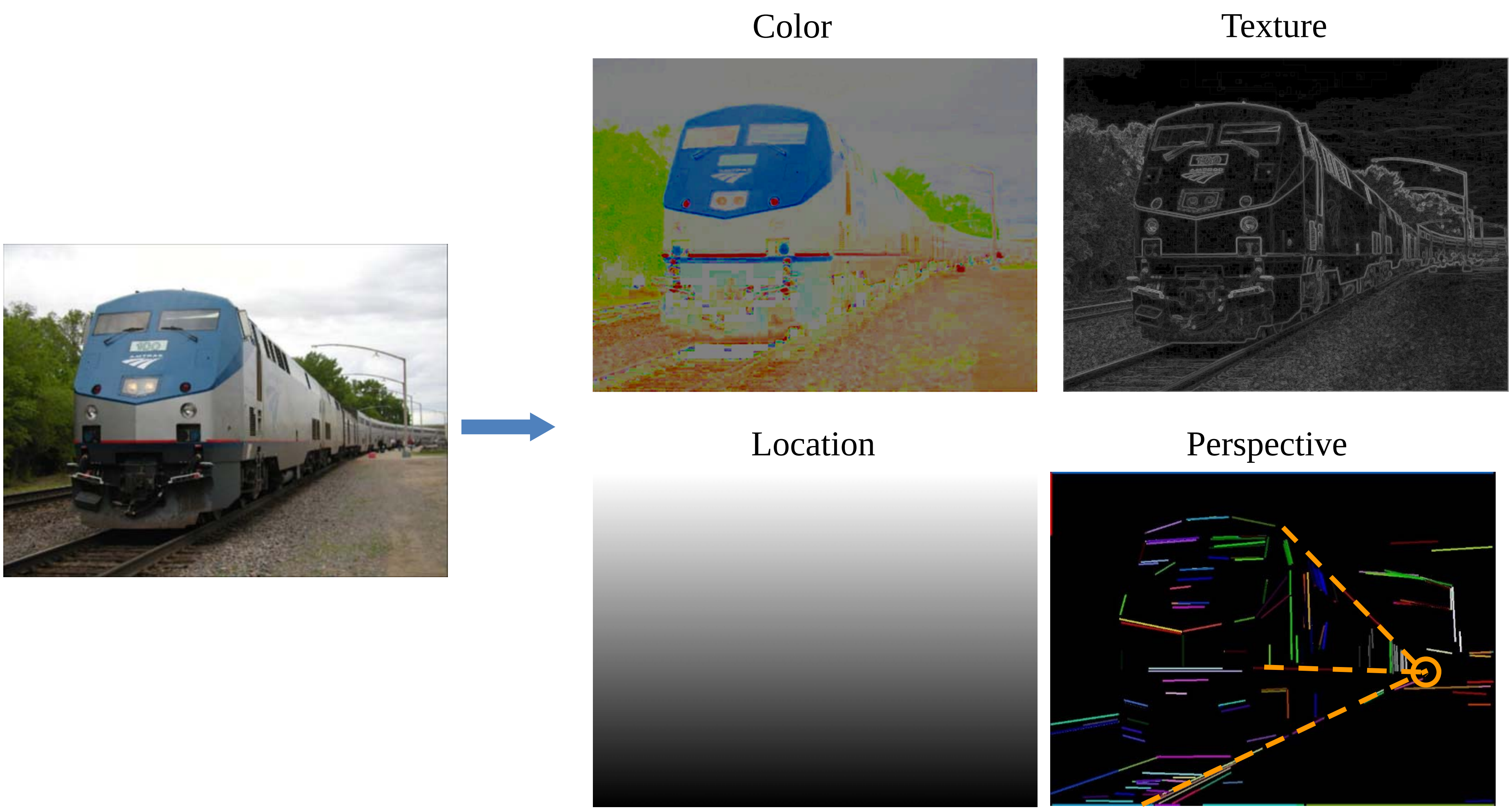


Final Pop-up Model

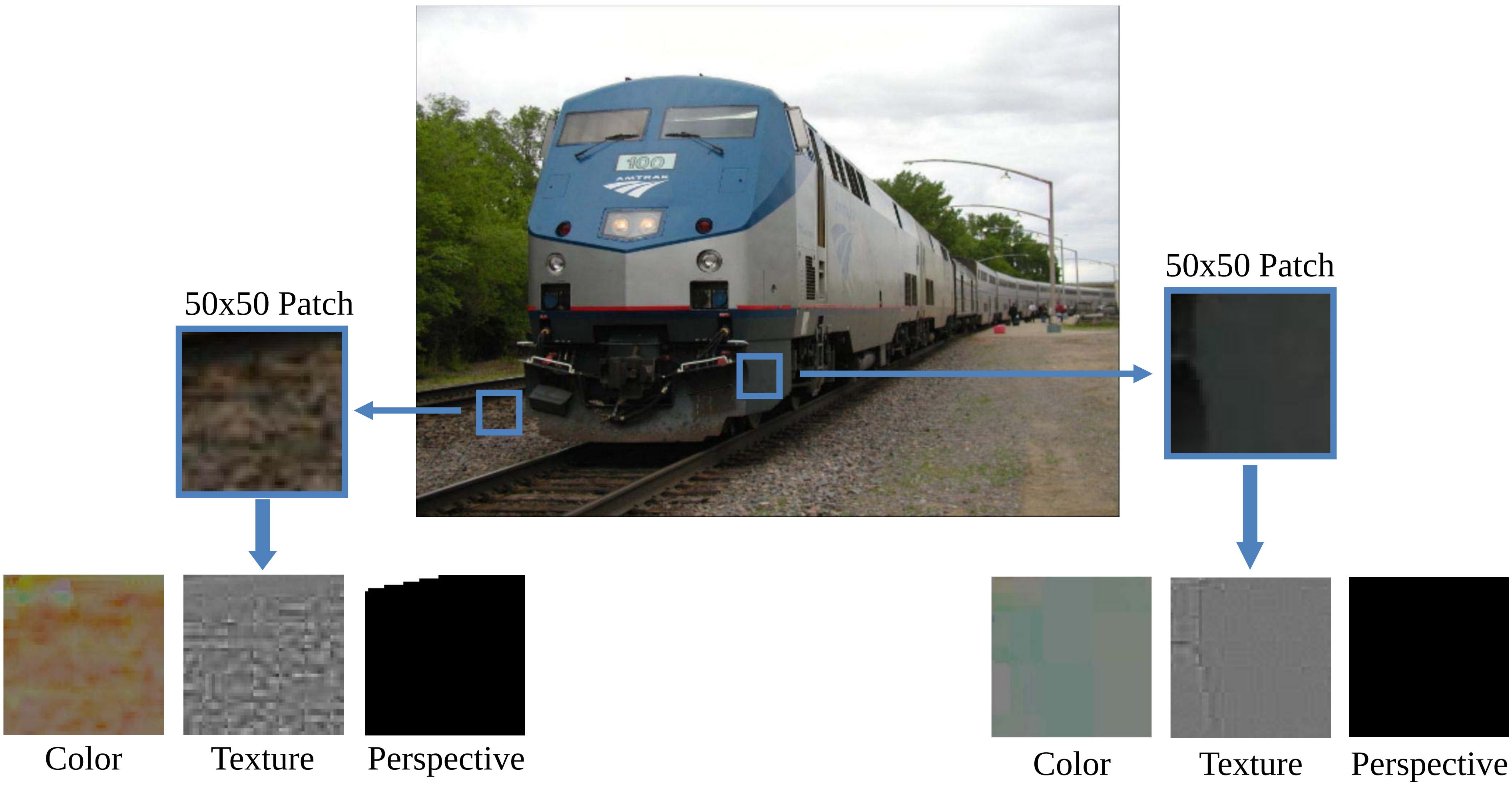
[Hoiem Efros Hebert 2005]

The World Behind the Image

Geometric Cues



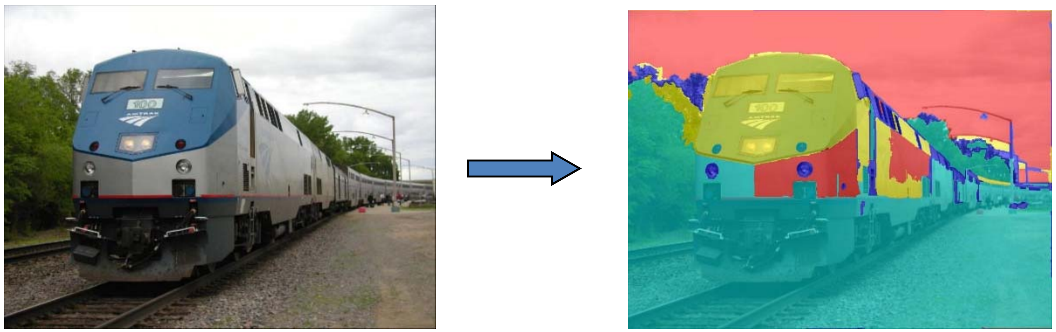
Need Good Spatial Support



Need Good Spatial Support

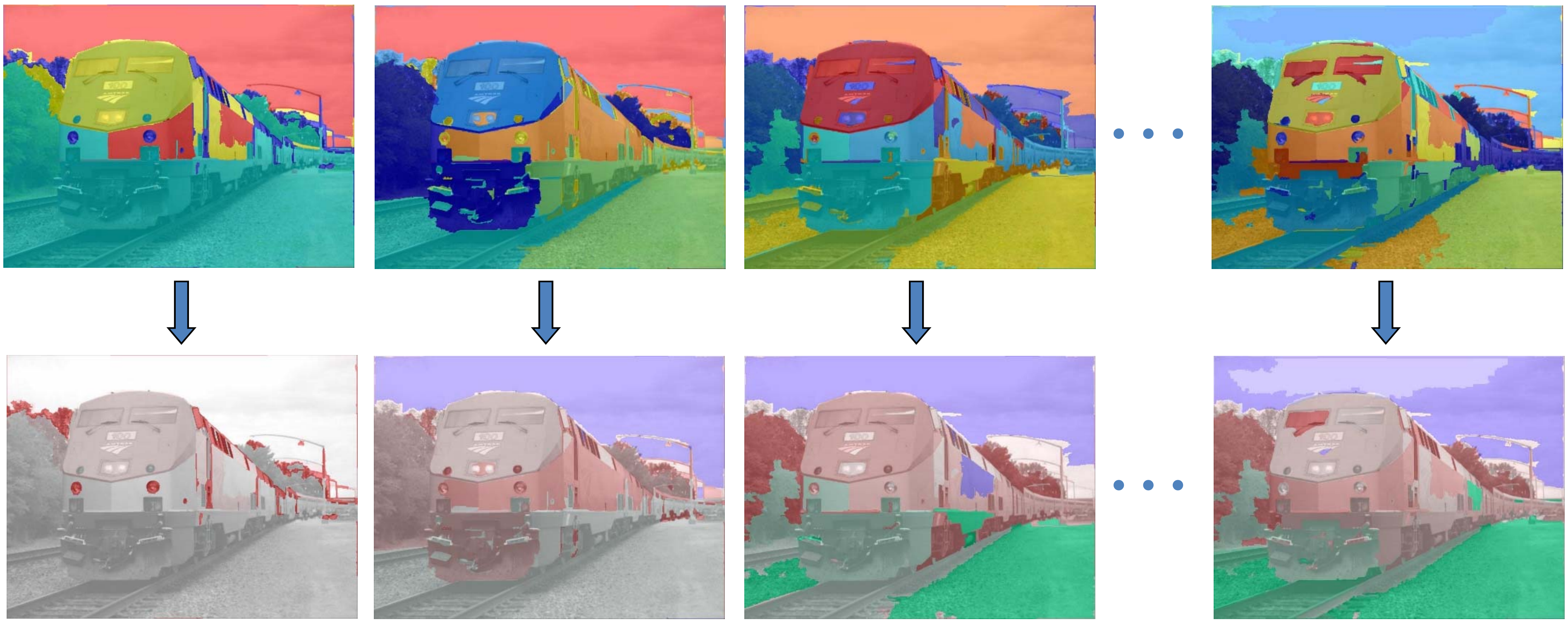
Image Segmentation

- Single segmentation won't work



- Solution: multiple segmentations

Labeling Segments



For each segment:

- Get P(*good segment* | *data*) P(*label* | *good segment*, *data*)

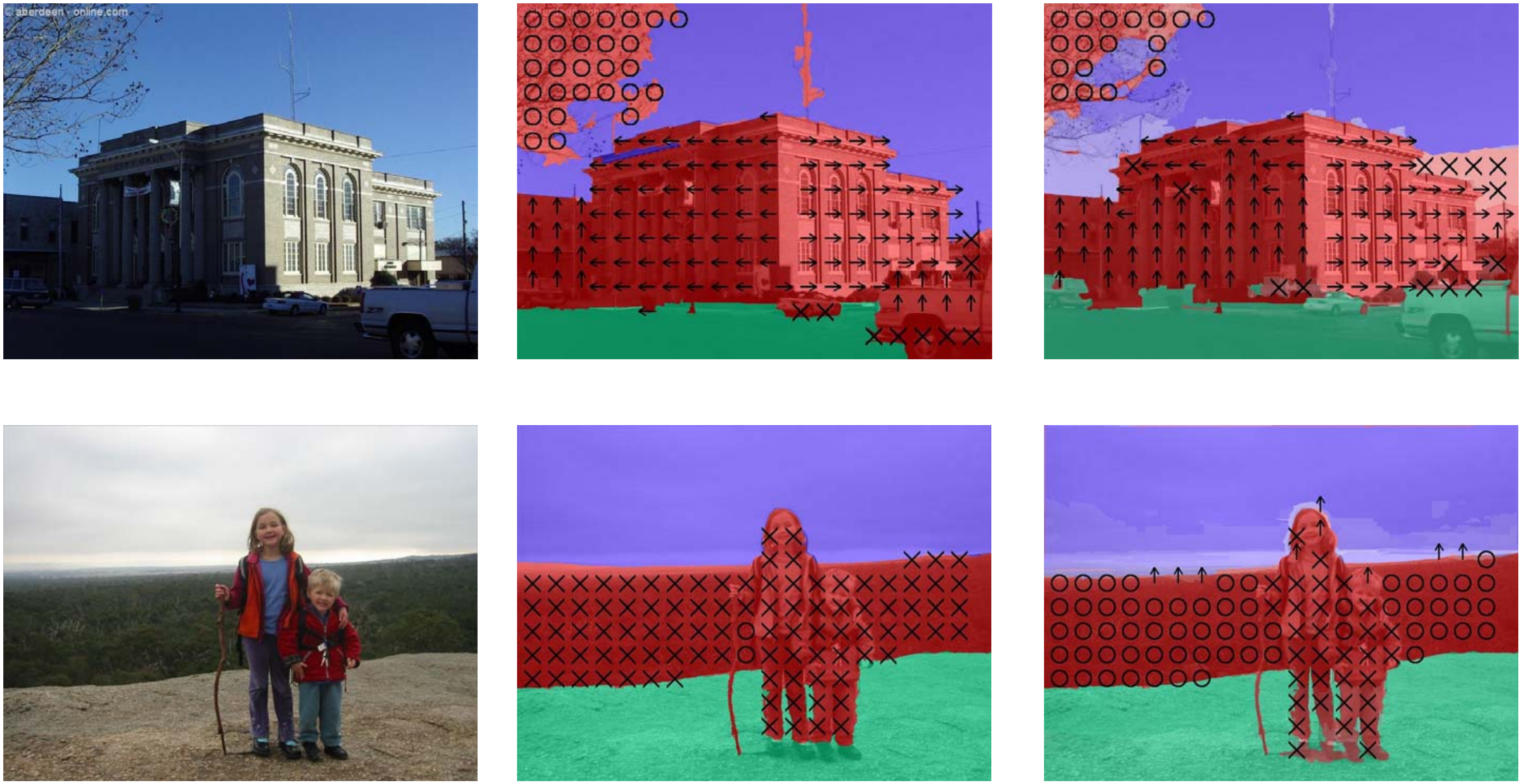
Image Labeling

Labeled Segmentations

Labeled Pixels

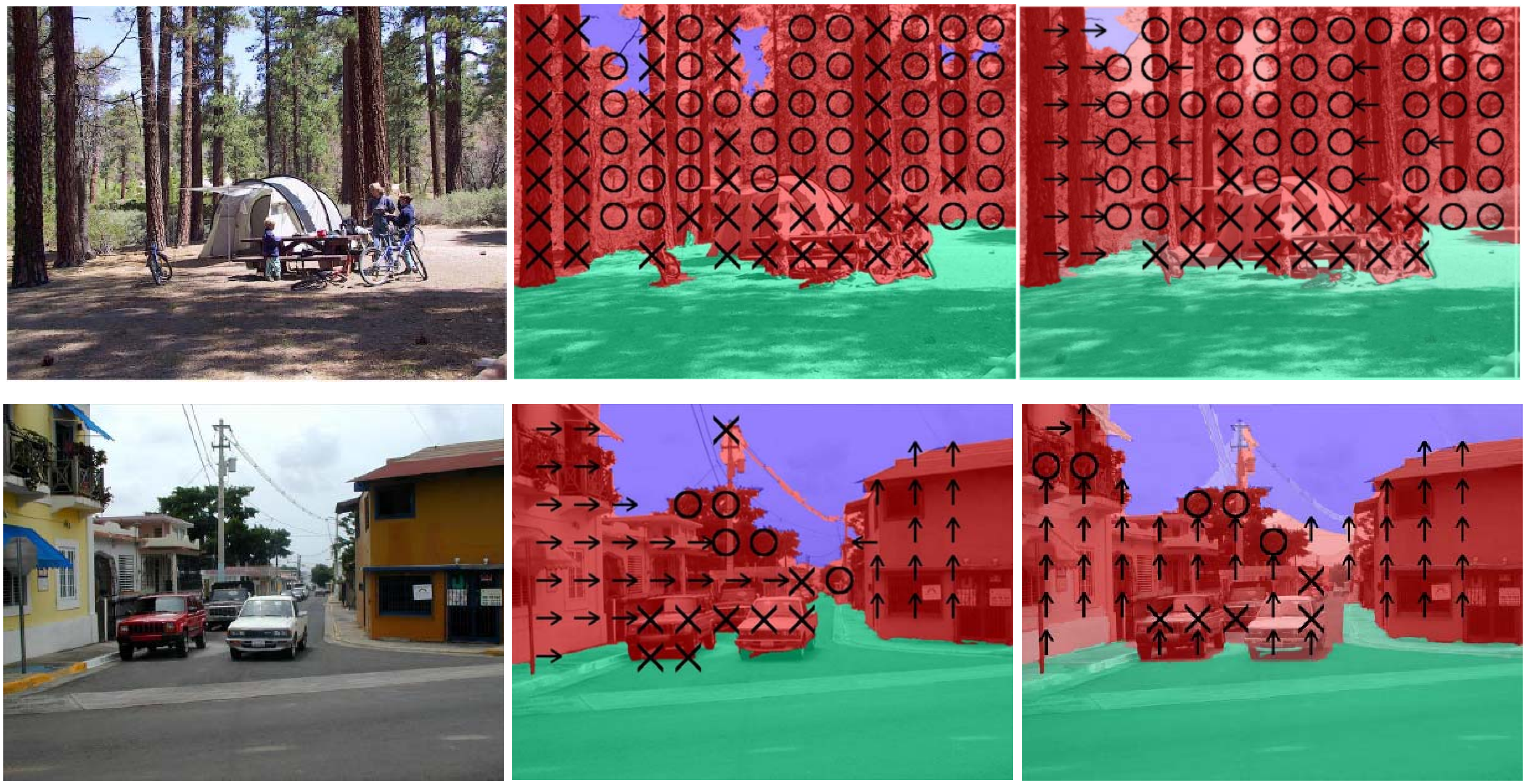
$$P(label \mid data) \propto \sum_{segments} P(good\ segment \mid data) P(label \mid good\ segment, data)$$

Labeling Results



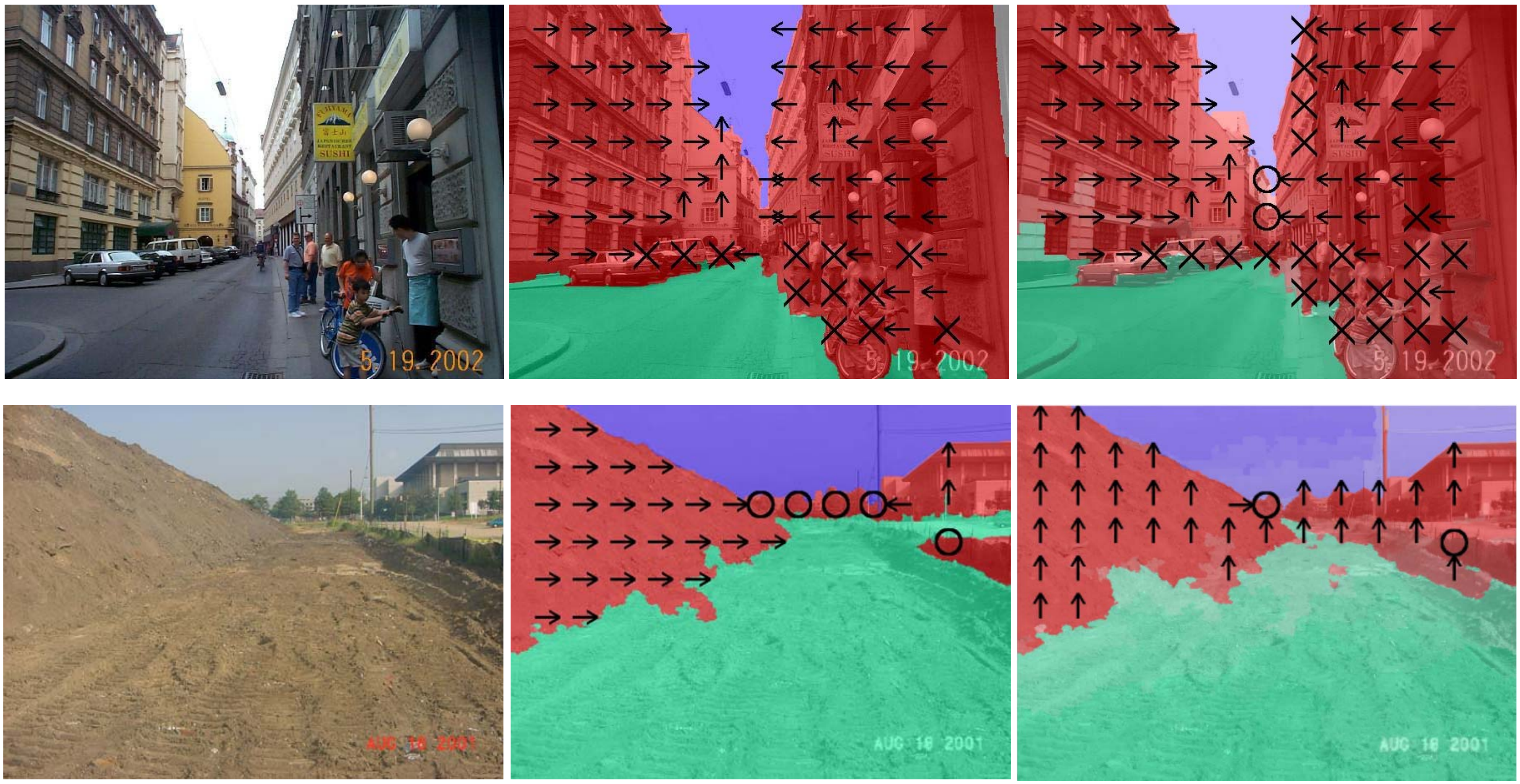
Input image　　Ground Truth　　Hoiem et al.

Results



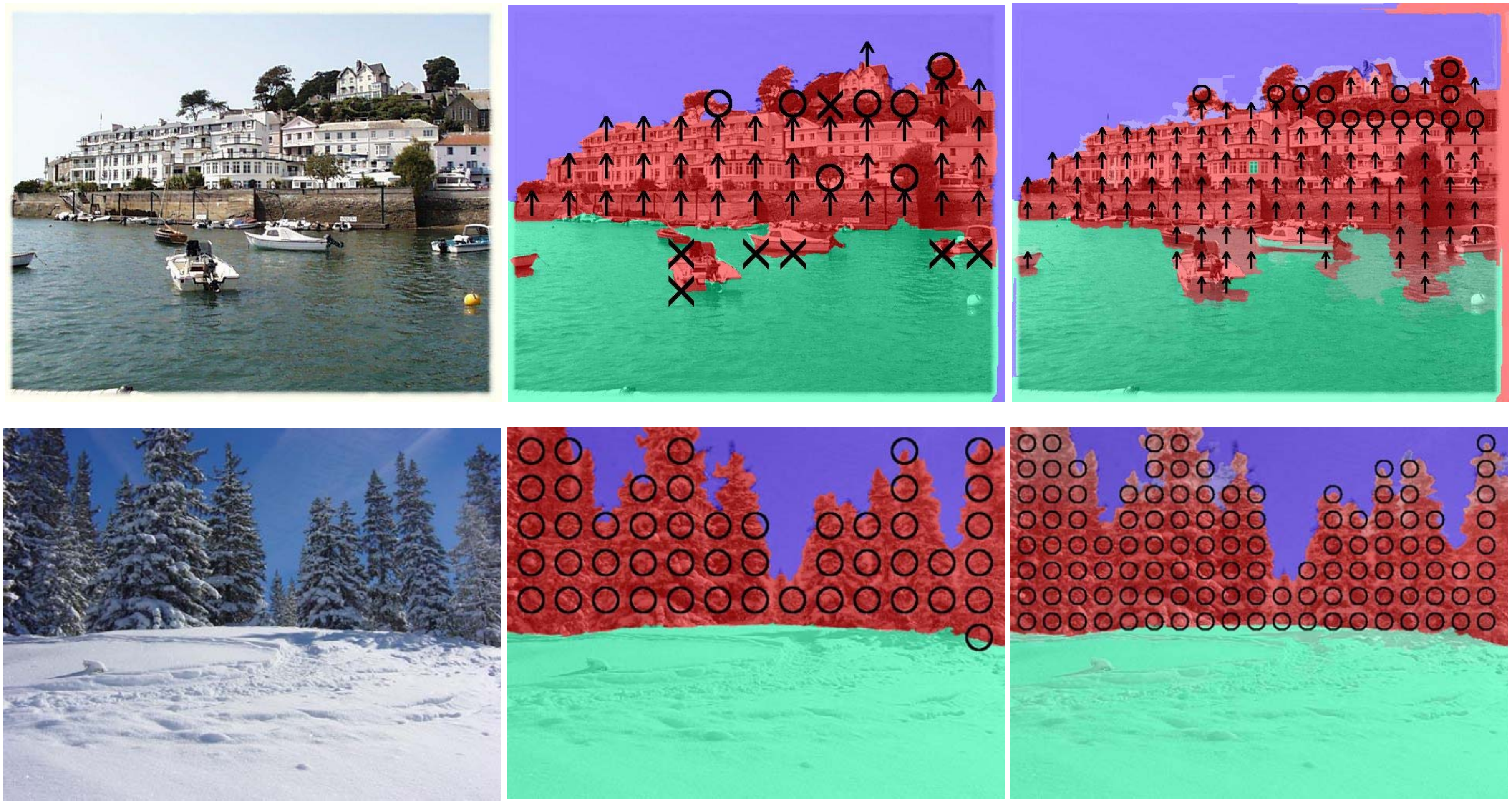
Input Image | Ground Truth | Hoiem et al.

Results



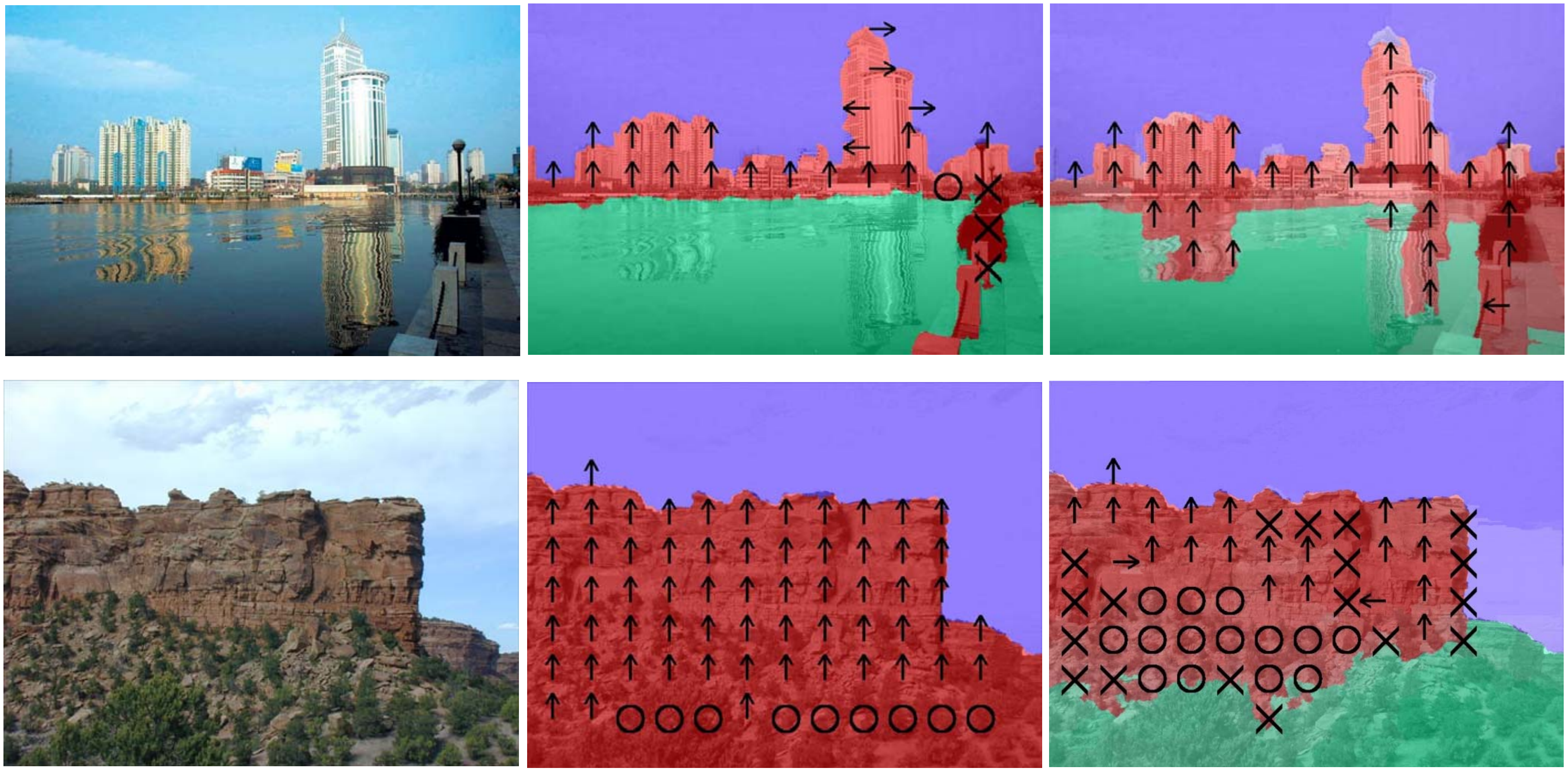
Input Image | Ground Truth | Hoiem et al.

Results



Input Image Ground Truth Hoiem et al.

Failures: Reflections, Rare Viewpoint



Input Image Ground Truth Hoiem et al.

Average Accuracy

Main Class: 88%

Subclasses: 61%

Main Class			
	Support	Vertical	Sky
Support	**0.84**	0.15	0.00
Vertical	0.09	**0.90**	0.02
Sky	0.00	0.10	**0.90**

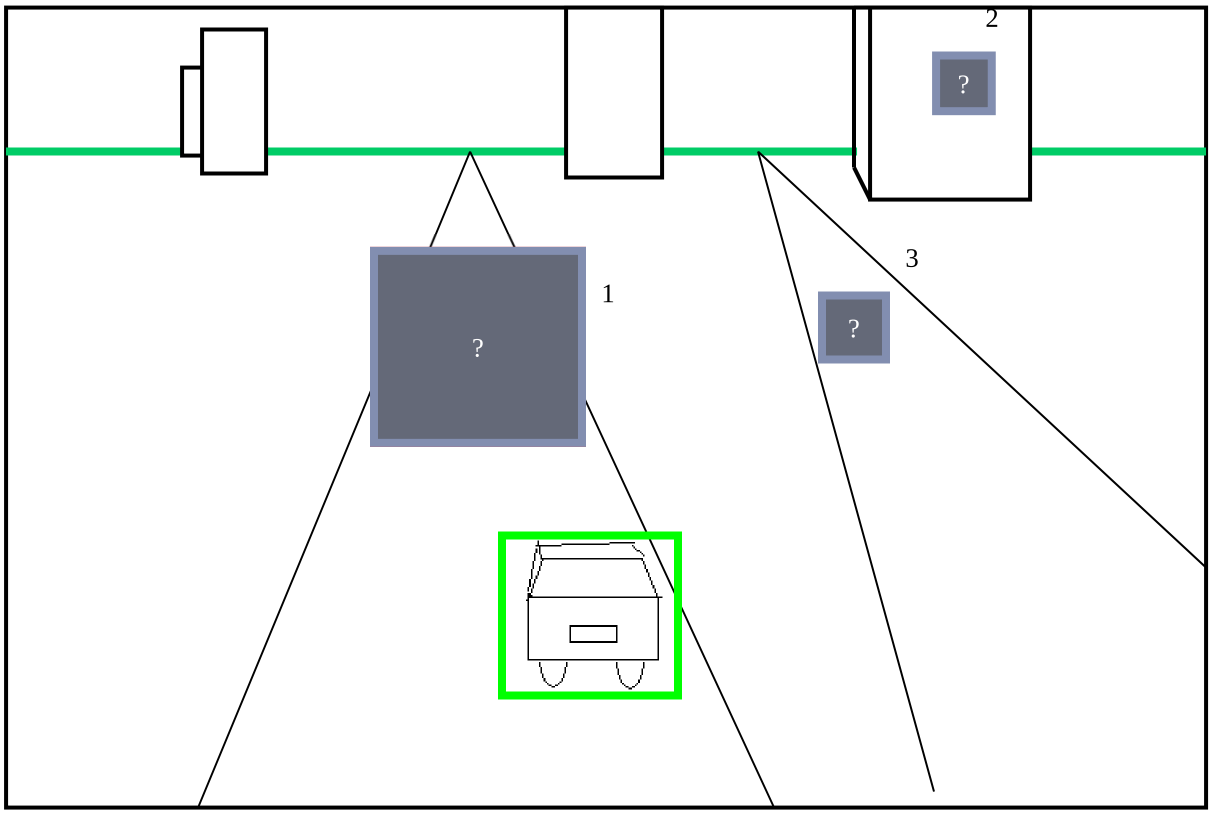
Vertical Subclass					
	Left	Center	Right	Porous	Solid
Left	**0.37**	0.32	0.08	0.09	0.13
Center	0.05	**0.56**	0.12	0.16	0.12
Right	0.02	0.28	**0.47**	0.13	0.10
Porous	0.01	0.07	0.03	**0.84**	0.06
Solid	0.04	0.20	0.04	0.17	**0.55**

Automatic Photo Popup

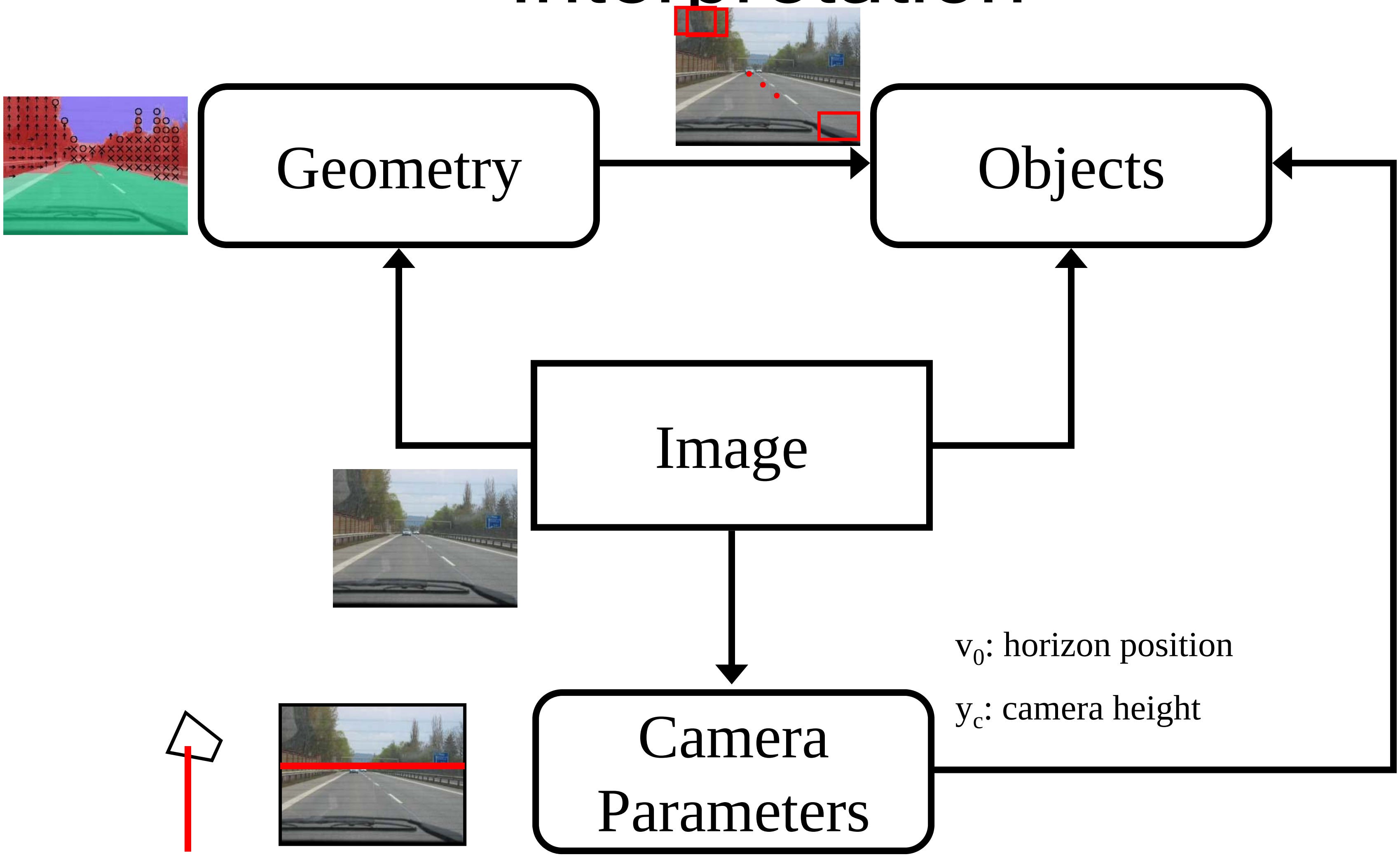
[SIGGRAPH 2005]

Relating Objects and Geometry

Knowledge of Geometry Critical for Object Detection



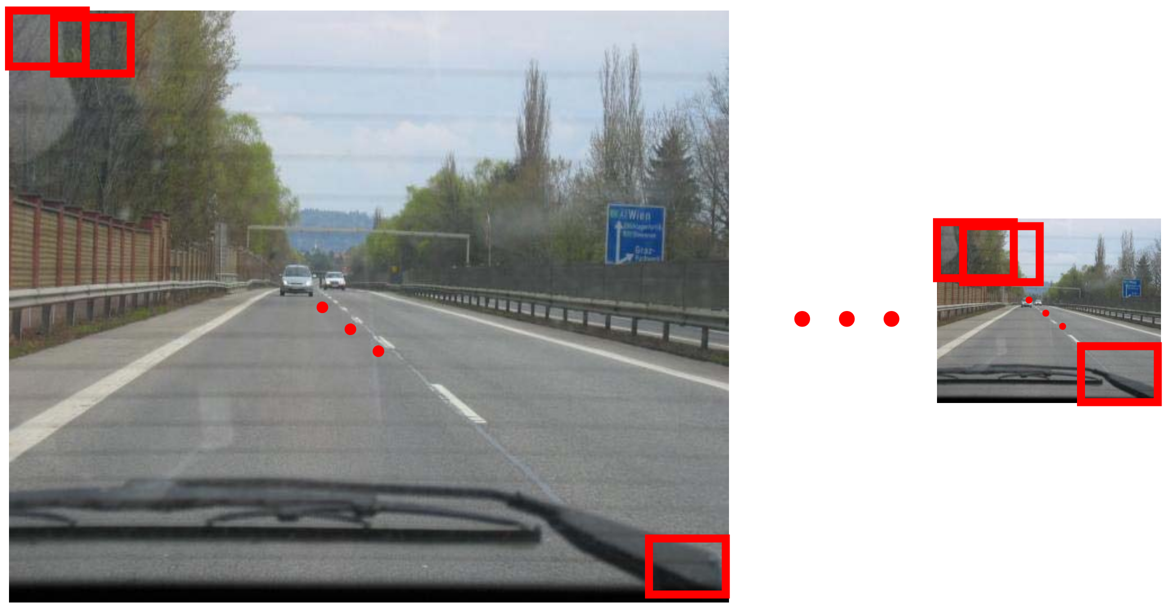
Geometrically Coherent Interpretation



What can we do with these models?

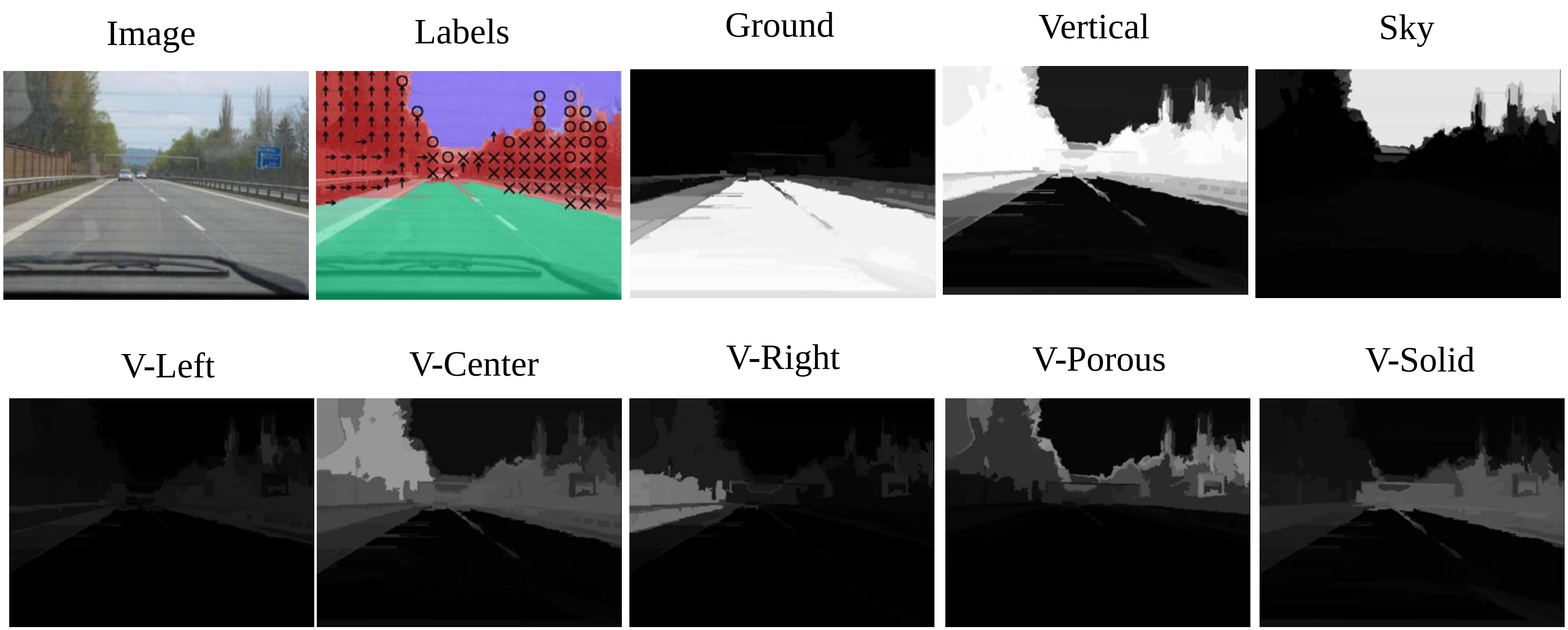
- Local Queries (marginalization)
  - What is the likelihood that this is a car?
  - What is the distribution of the camera height given the image?
- Global Queries (maximization)
  - What is the most likely complete hypothesis of objects, geometry, parameters?

Objects

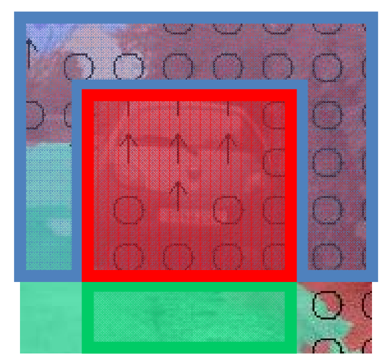


- Object detector:
  - Defines a set of objects
  - Estimate likelihood of object identities at possible locations/scales
- Learn distribution of object heights *in the 3D world*
  - E.g. consumerreports.com for cars

Geometry

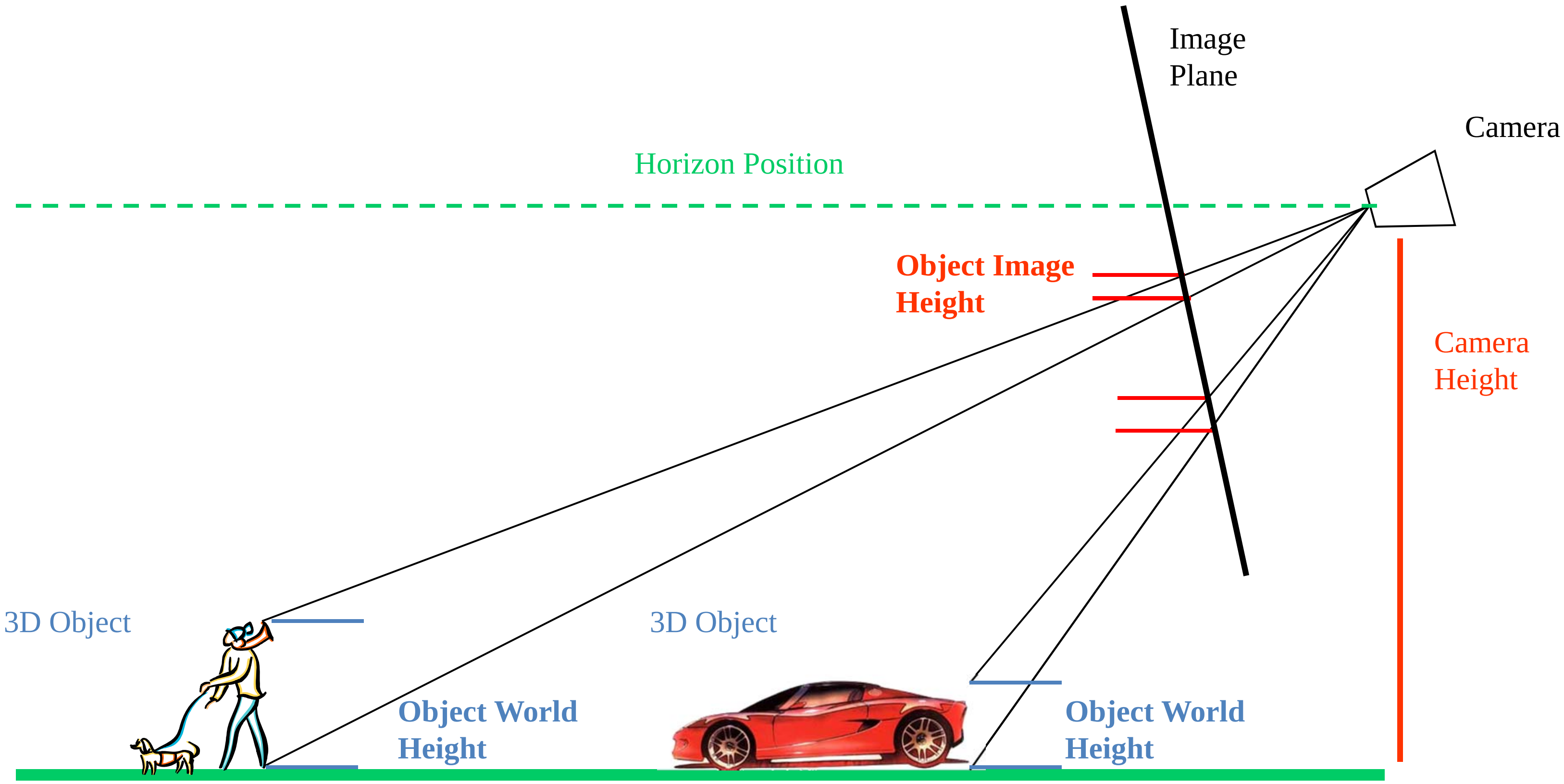


- Geometry estimates: produce probability maps for each label
- Compute:
  - Likelihoods of object identities at each position given the geometry and image data
  - Likelihood of geometry given image data
- Initial estimates (top of ground, bottom of sky) help horizon estimate



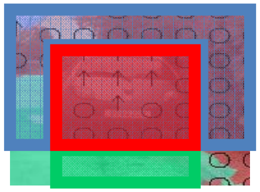
Camera Parameters

- Camera Height – Estimate prior from training images
- Horizon Position – Estimate based on prior, estimated geometry, and potential vanishing points
- Identified objects of known height distribution help refine camera parameter estimates



Local Queries (Marginalization)

- Exact marginalization not tractable
- Assumptions
  - Objects depend on local geometry
  - Local geometry independent
  - Objects independent given camera parameters

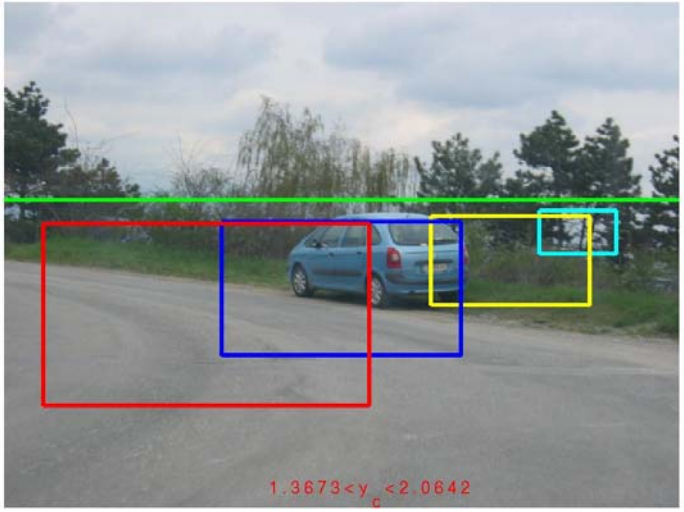


- Approximations
  - Marginalize only over hypotheses that have at most one other object
  - Discard extremely unlikely objects/camera parameters early

How far can camera parameters get us?

- What the system knows:
  - Estimated horizon position
  - Camera height prior (in meters)
  - Distribution of car heights in the world (in meters)
  - **No other image data!**

- What the system tells you:
  - 50% confidence interval for size of car (in the image) given bottom-center position
  - 50% confidence intervals for camera parameters



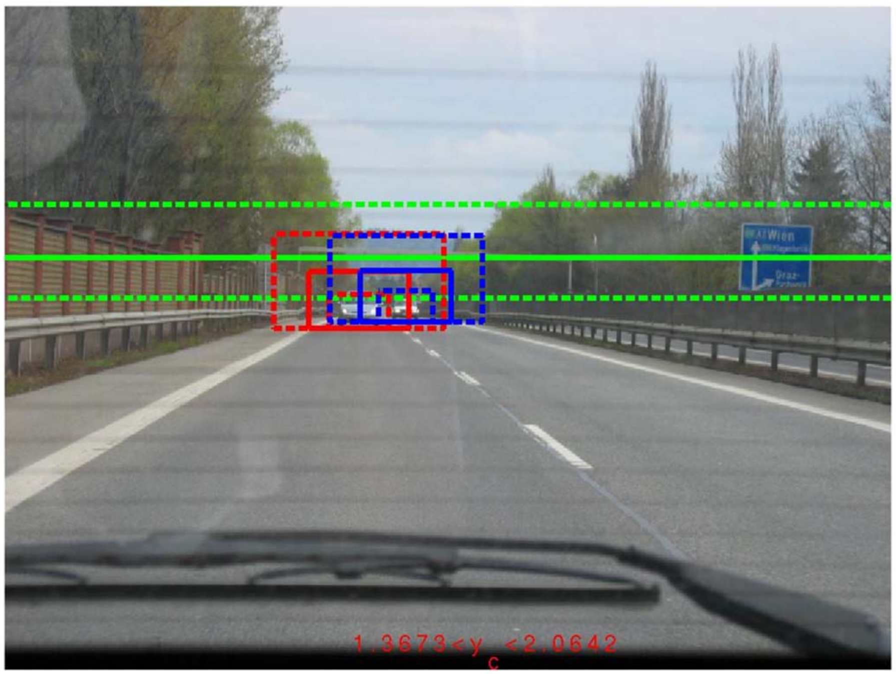
Hallucinations: Camera Parameters

No Objects Given

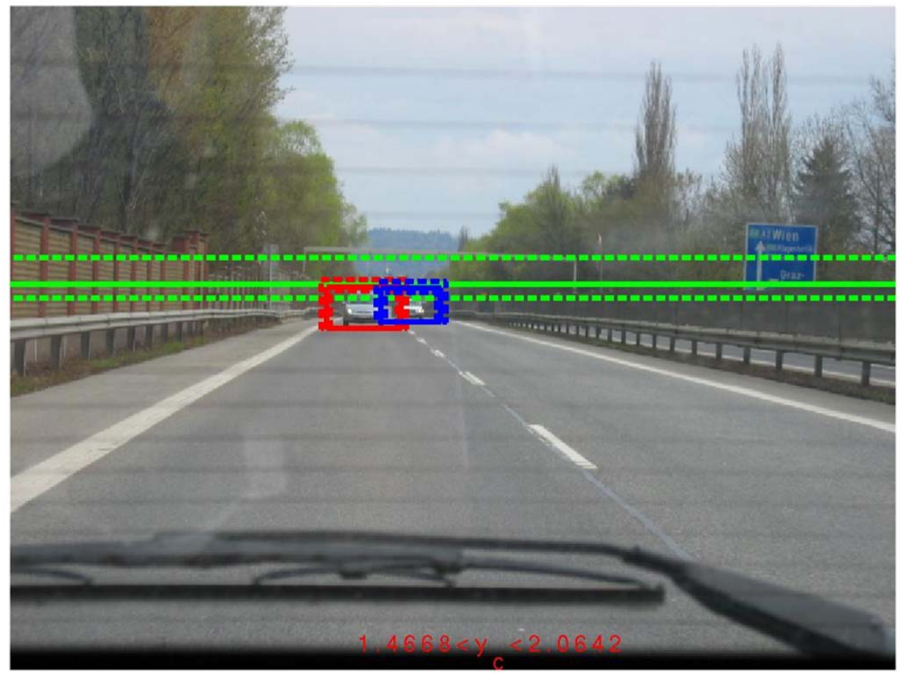
1 Object Given

Hallucinations: Camera Parameters

No Objects Given

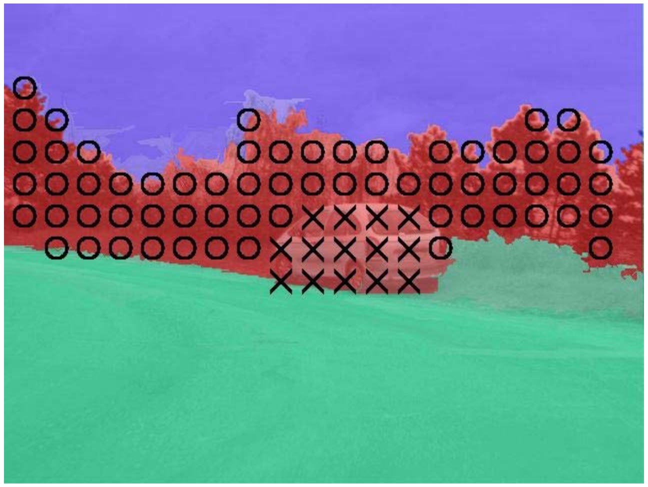


1 Object Given

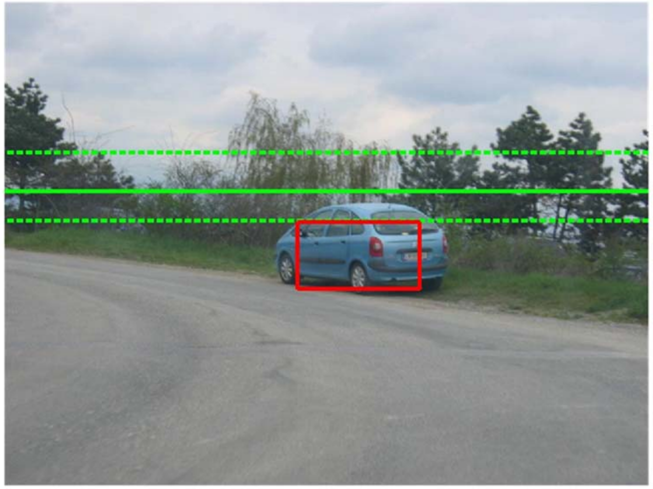


How far can camera parameters *and* geometry get us?

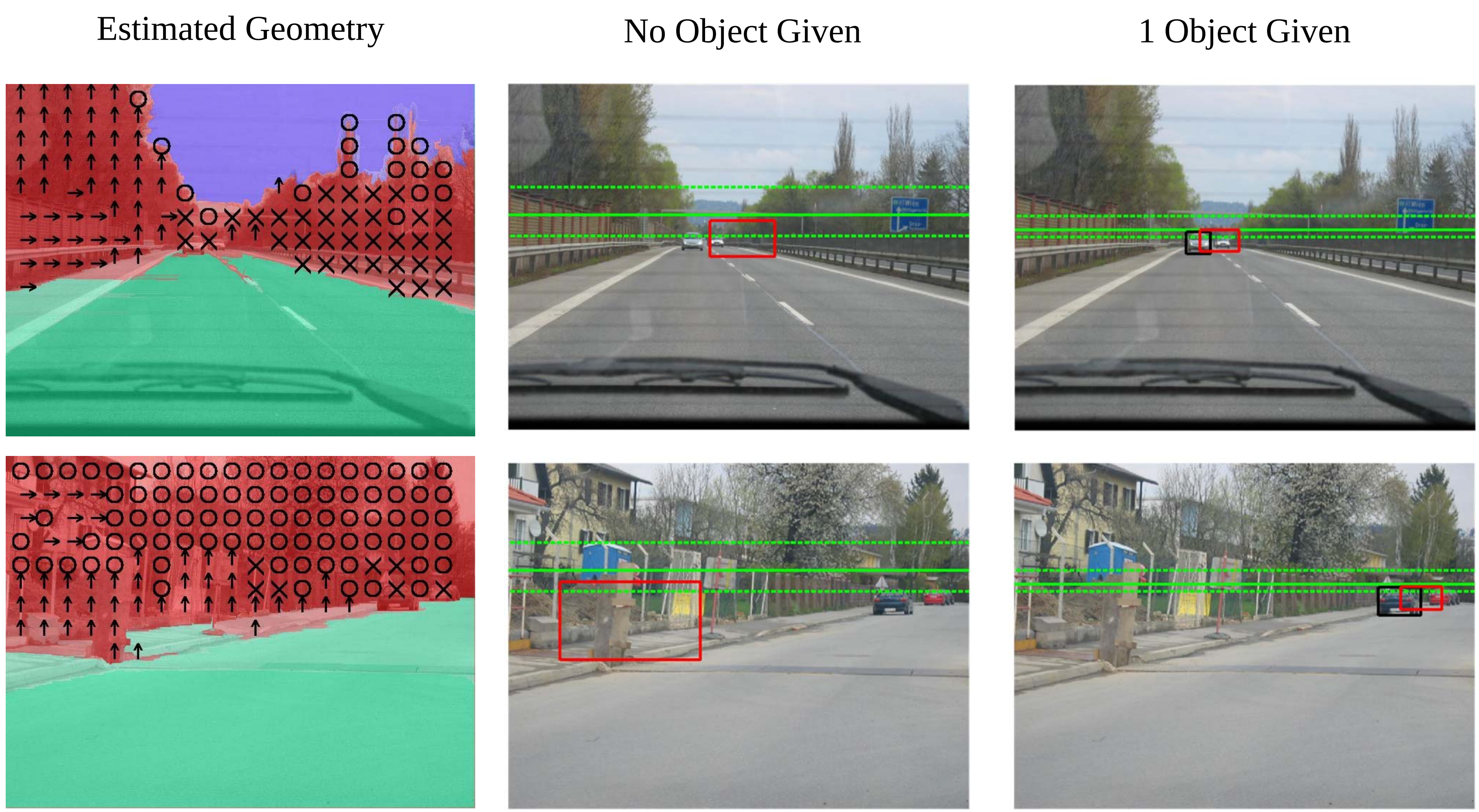
- What the system knows:
  - Same as before but with geometry estimates
  - **No other image data!**



- What the system tells you:
  - Most likely position/size of car in image
  - 50% confidence intervals for camera parameters

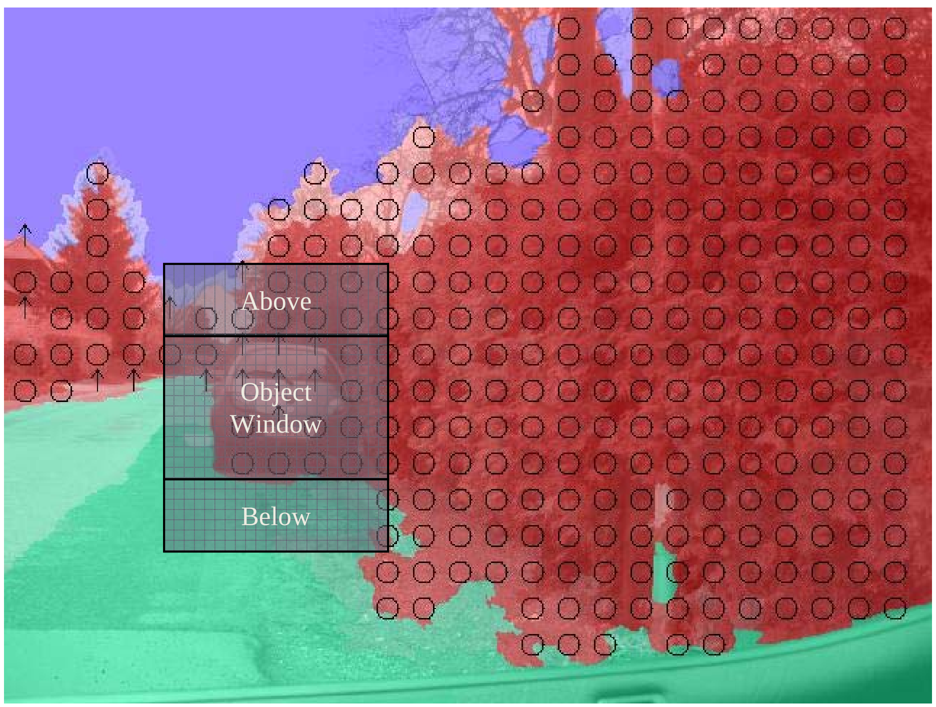


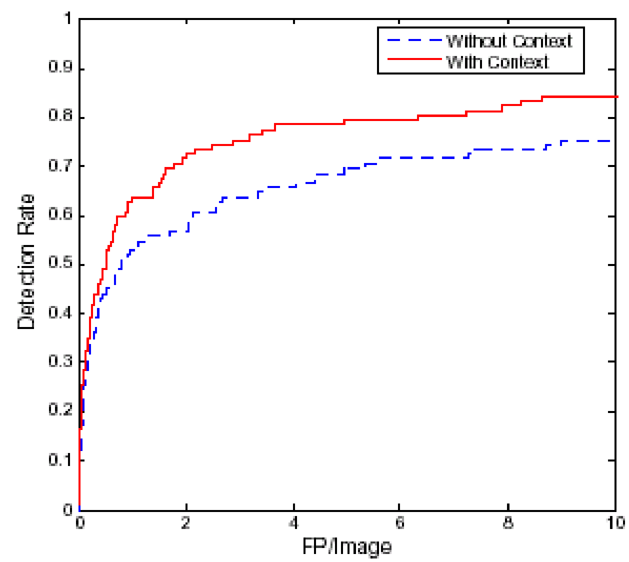
Hallucinations: Geometry and Camera Parameters



How much do geometry estimates help (with local image data)?

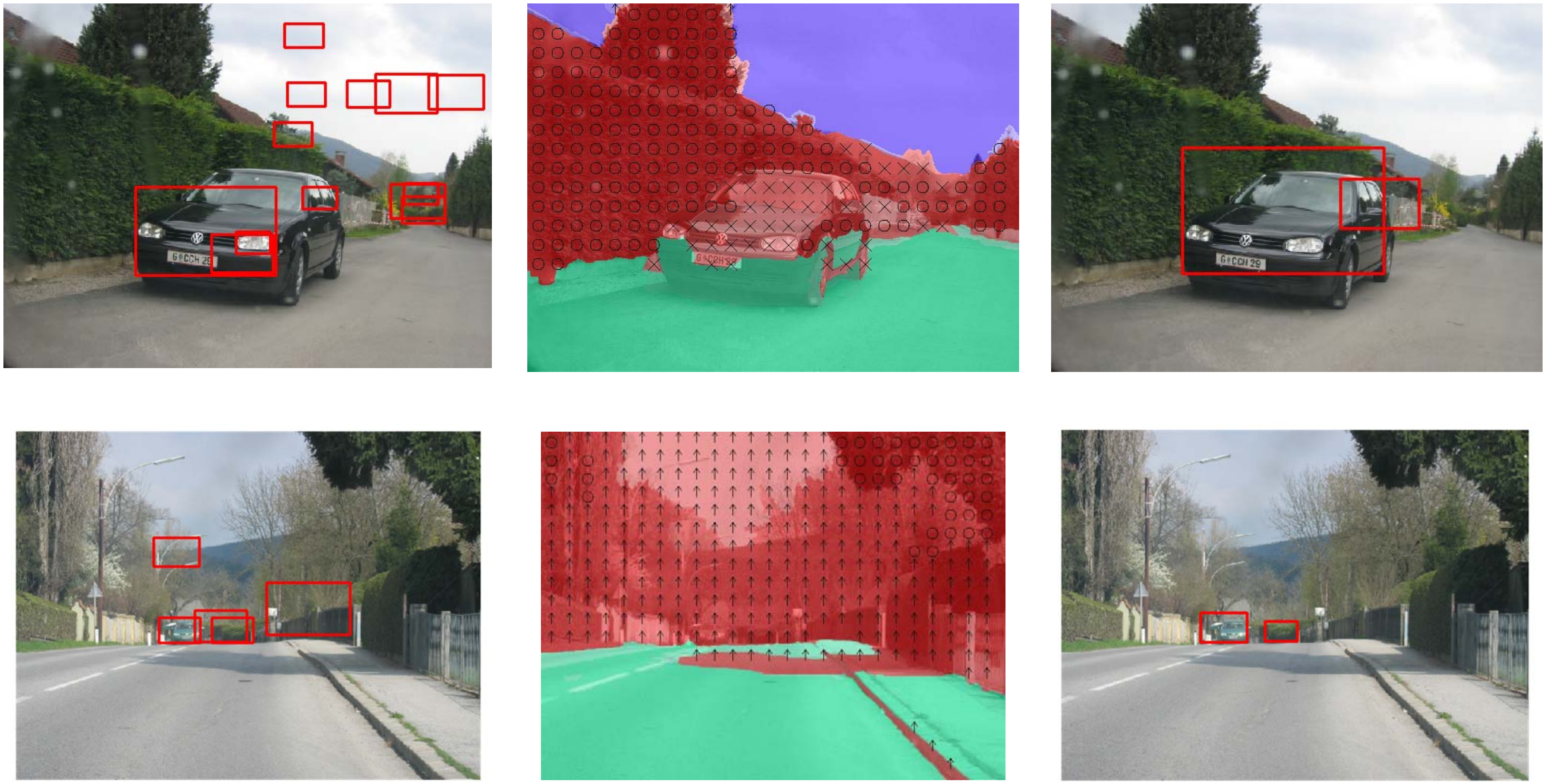
- 40 contextual features based on average confidence values of geometric labels within windows





[ICCV 2005]

Example Results



[Murphy et al., 2003] [Hoiem et al., 2005] + 40 context features

[ICCV 2005]

Global Queries (Maximization)

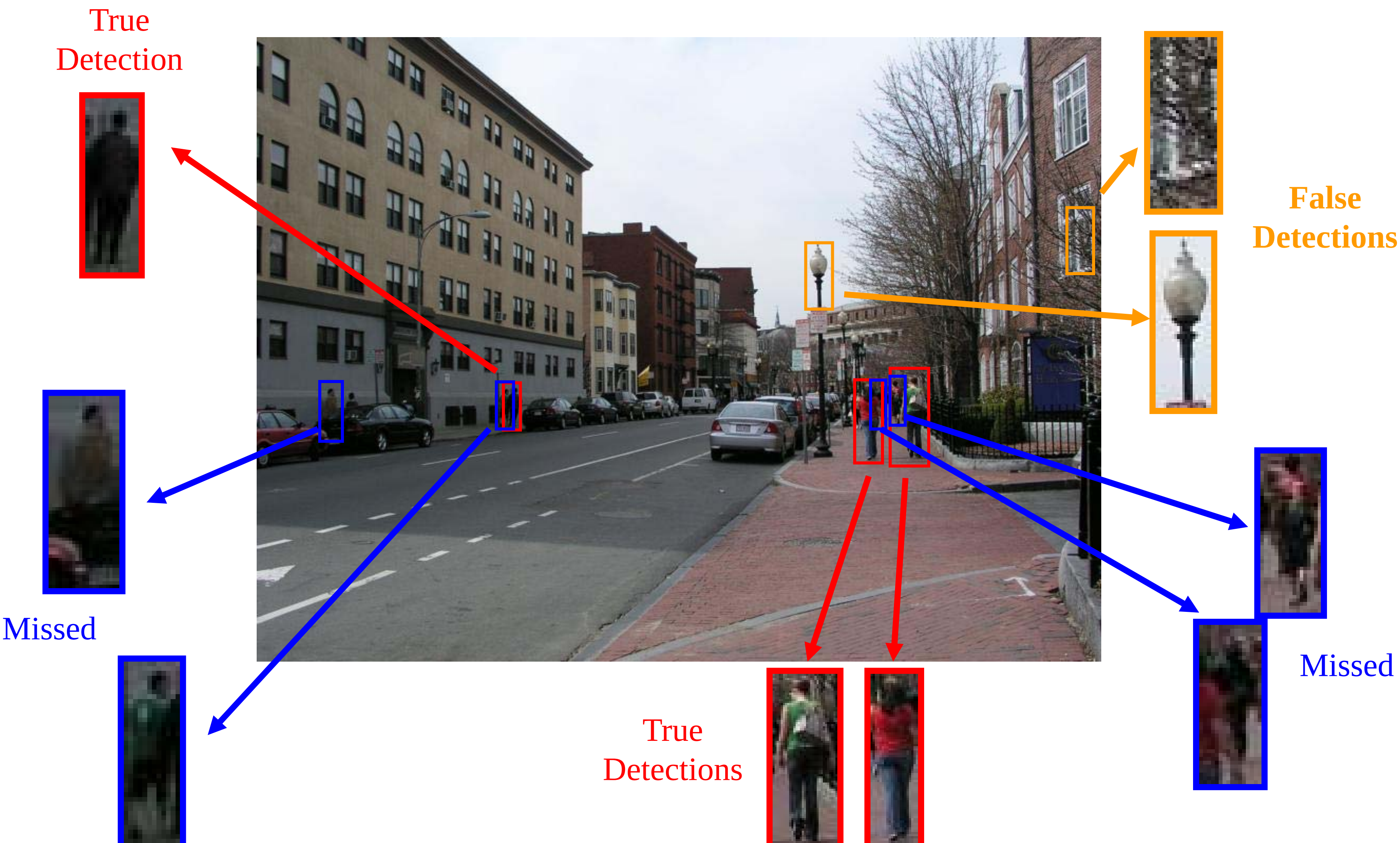
- Provides full image interpretation (for modeled aspects of scene)
- Finding optimal solution is intractable
  - Branch and bound algorithm
  - Greedy algorithms
- Usefulness depends on the peakedness of the joint distribution

Putting Objects in Perspective

Derek Hoiem
Alexei A. Efros
Martial Hebert

Carnegie Mellon University
Robotics Institute

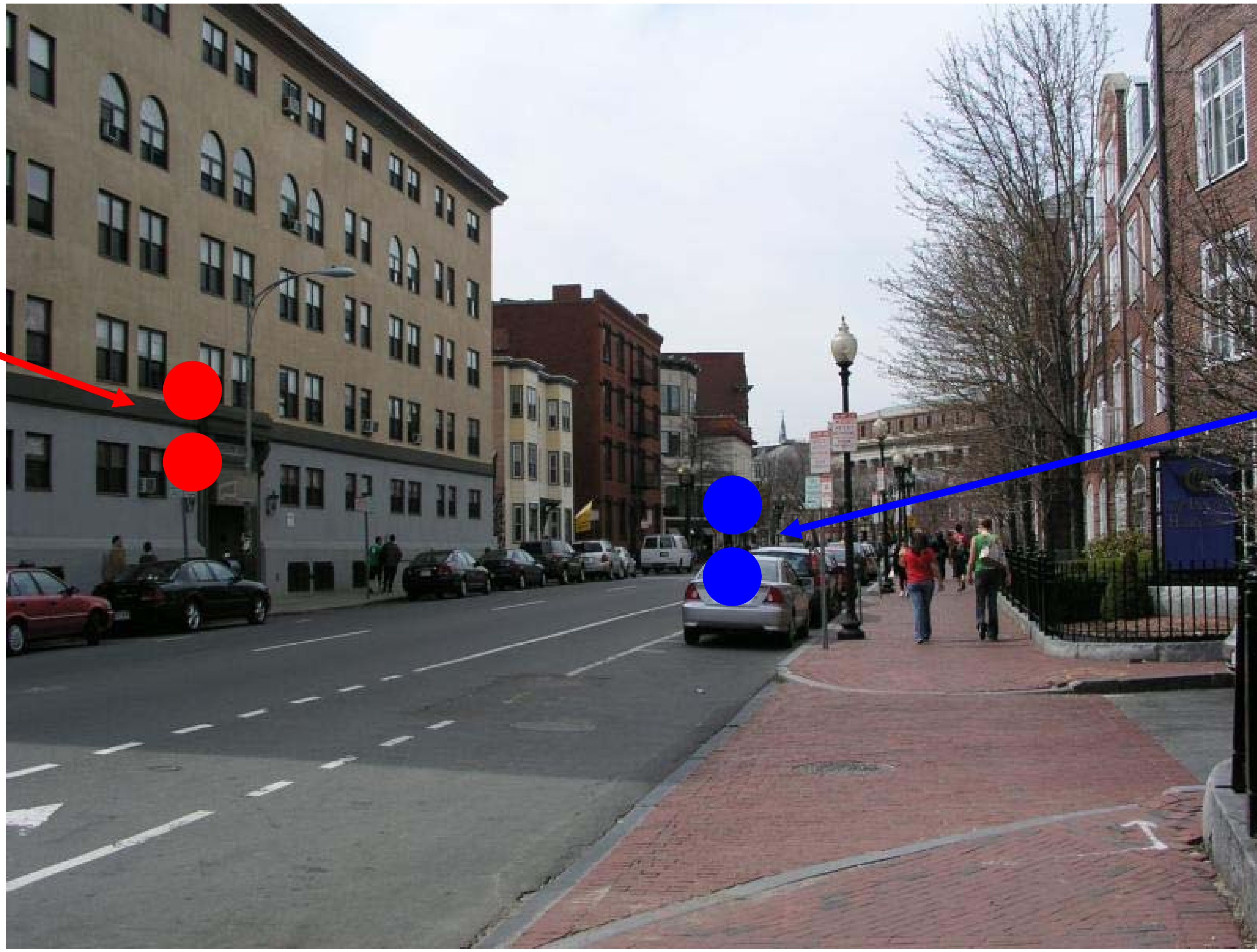
Local Object Detection



Local Detector: [Dalal-Triggs 2005]

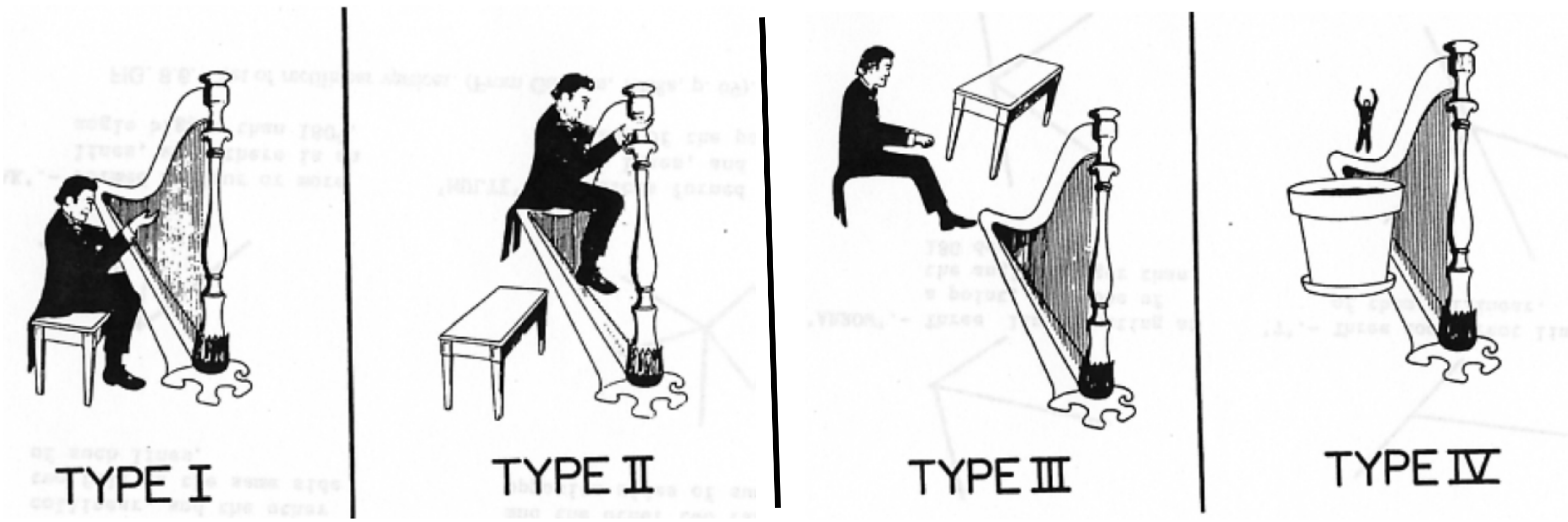
Real Relationships are 3D

Close



Not Close

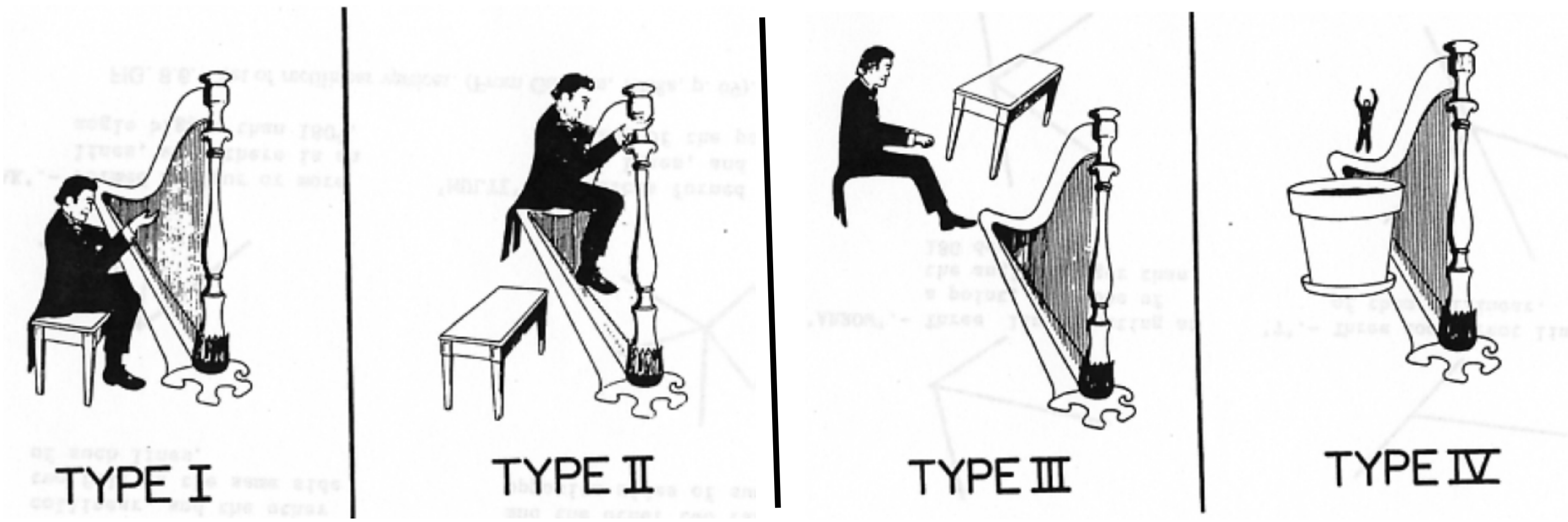
Objects and Scenes



Hock, Romanski, Galie, & Williams 1978

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

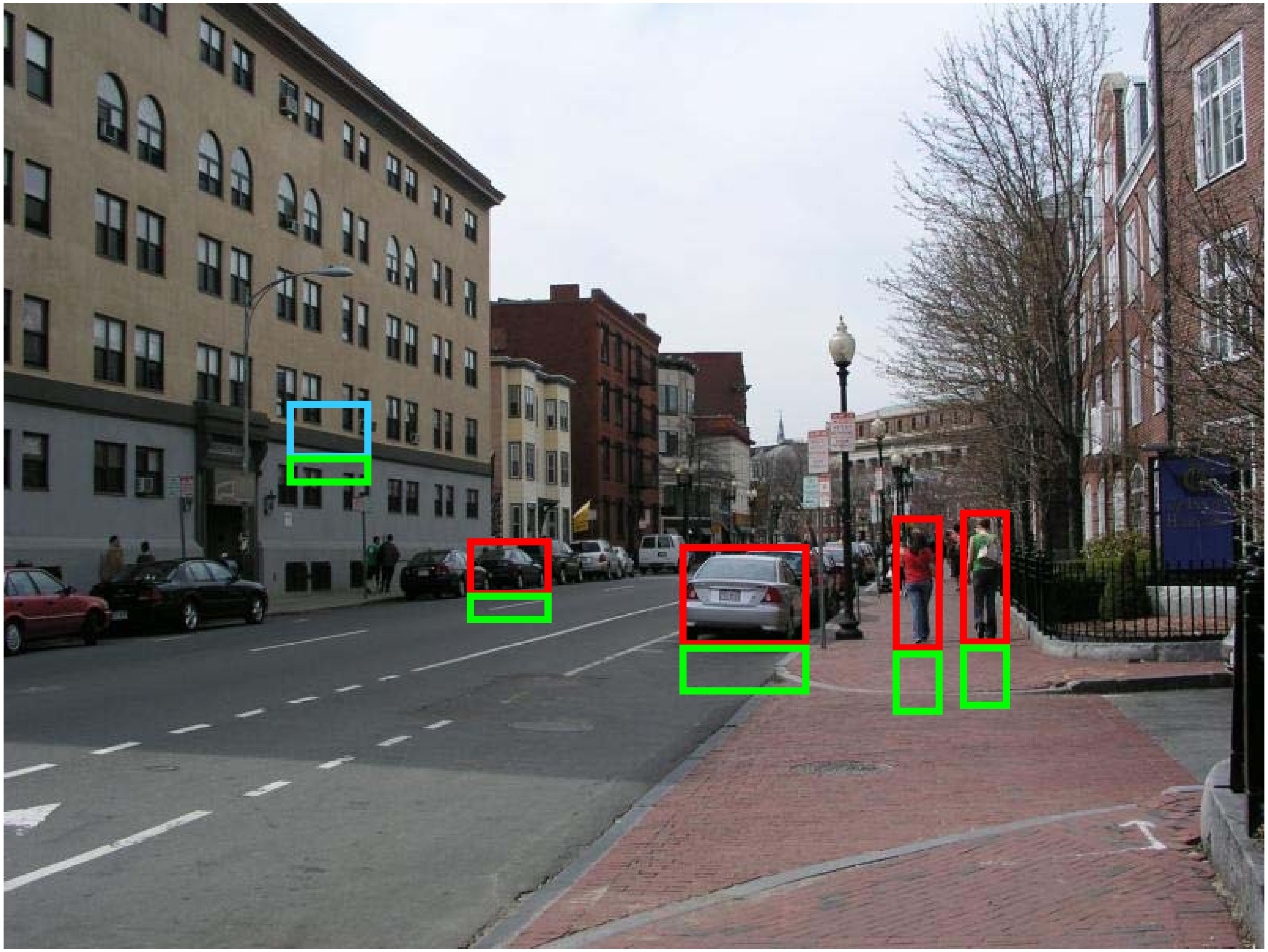
Contribution of this Research



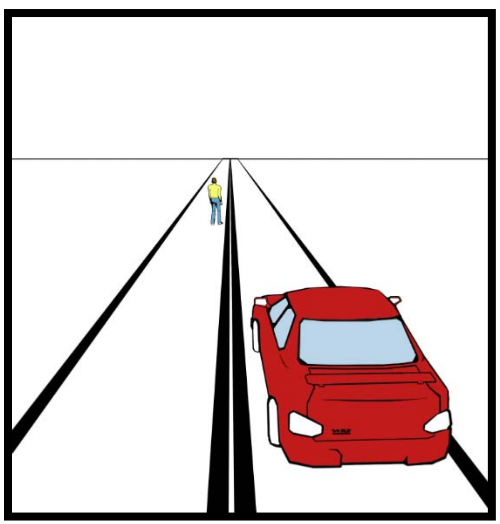
Hock, Romanski, Galie, & Williams 1978

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

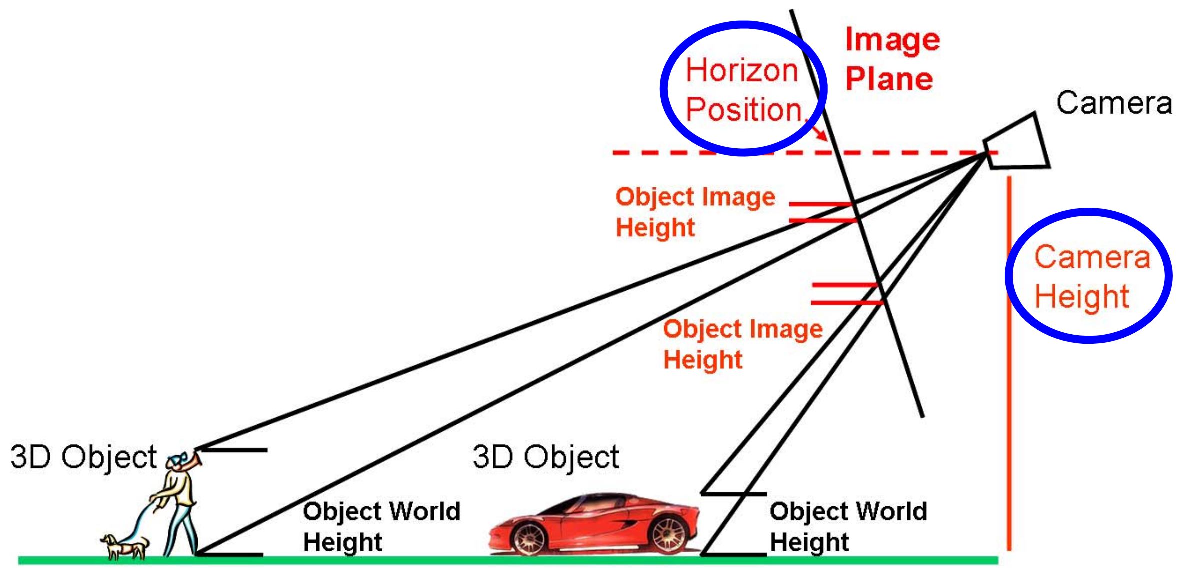
Object Support



Object Size in the Image



Is the person or the car taller?



World

Object Size ↔ Camera Viewpoint

Input Image

Loose Viewpoint Prior

Object Size ↔ Camera Viewpoint

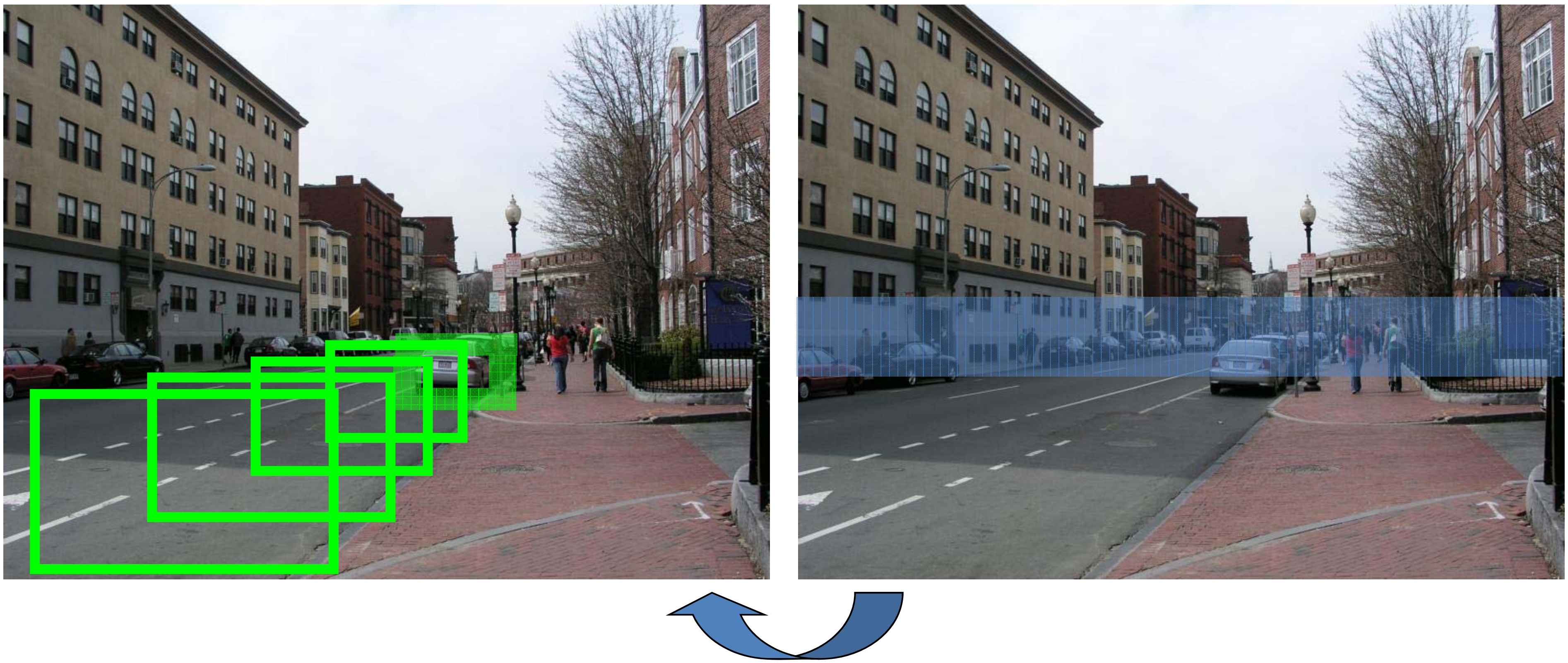
Object Position/Sizes

Viewpoint

Object Size ↔ Camera Viewpoint

Object Position/Sizes

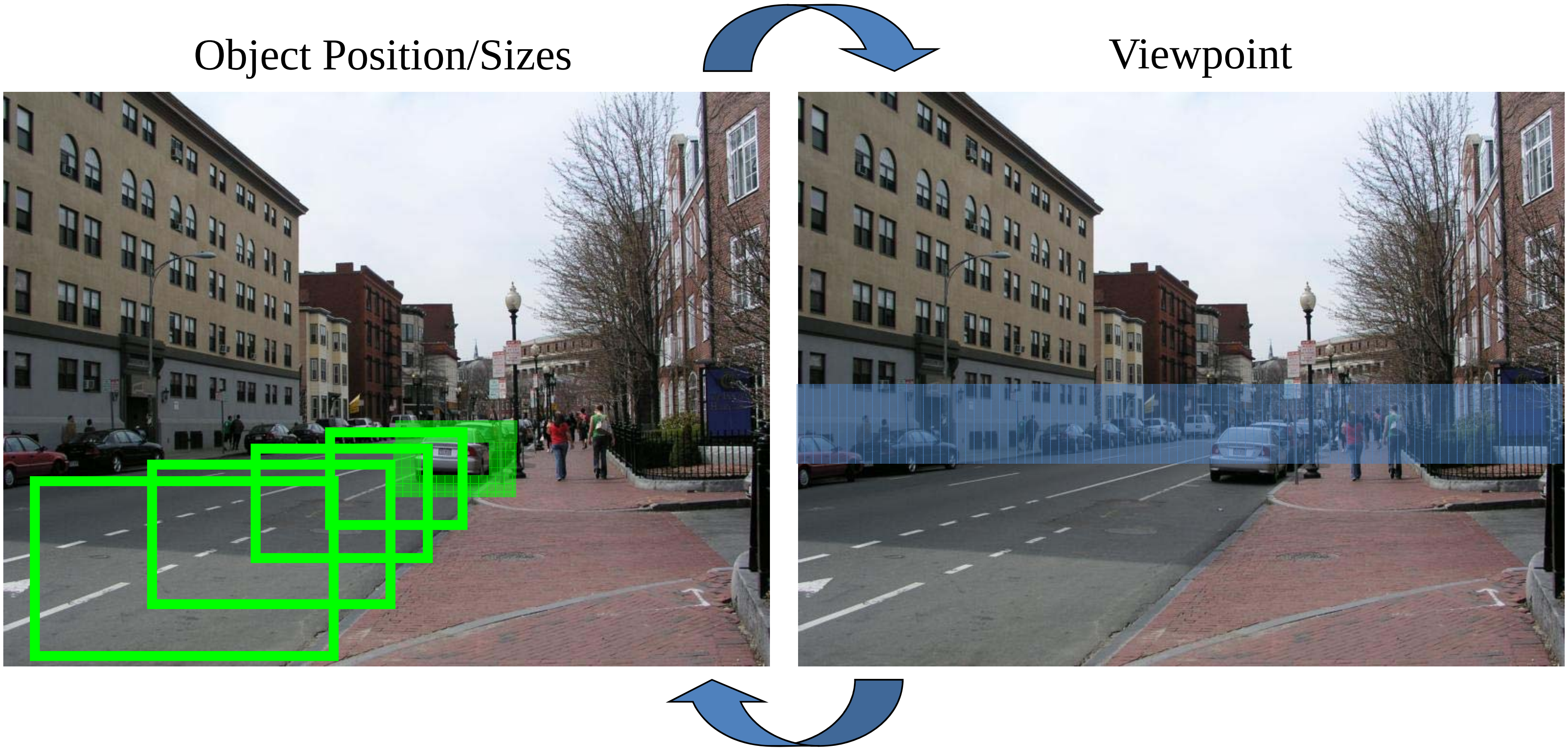
Viewpoint



Object Size ↔ Camera Viewpoint

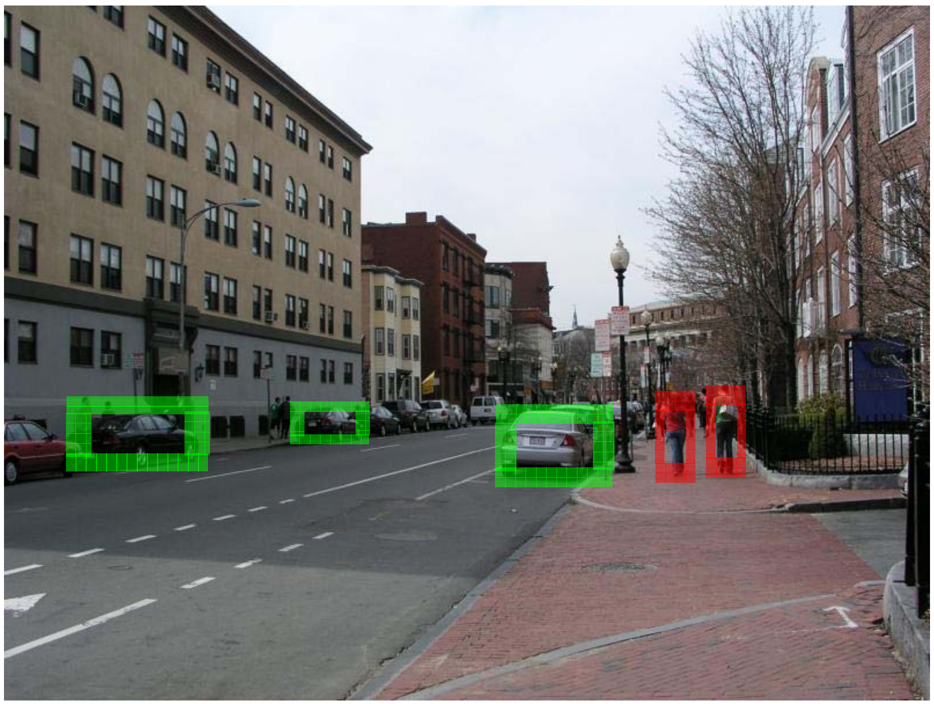
Object Position/Sizes

Viewpoint



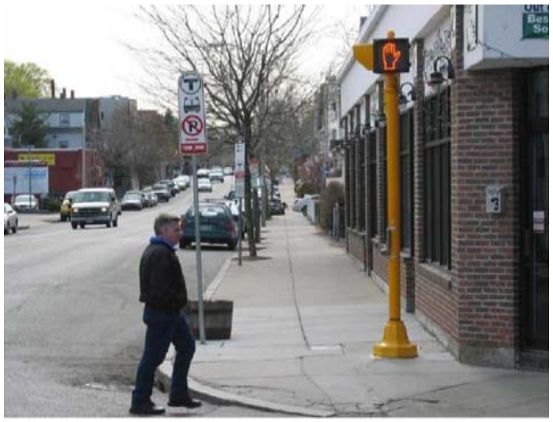
Object Size ↔ Camera Viewpoint

Object Position/Sizes

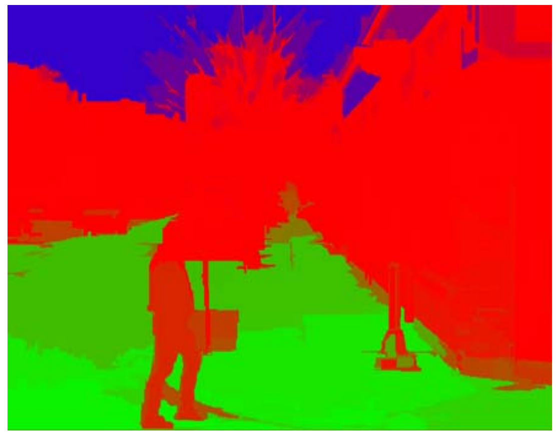


Viewpoint

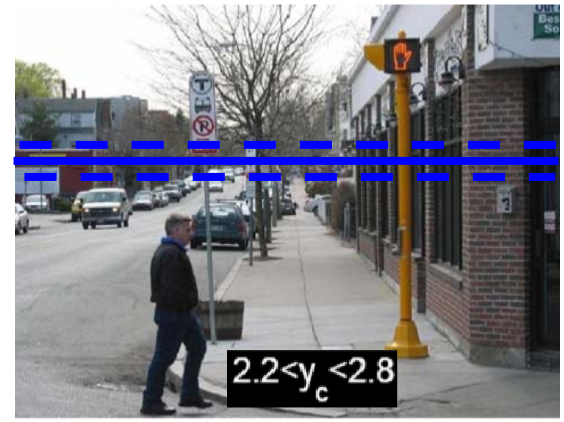
What does surface and viewpoint say about objects?



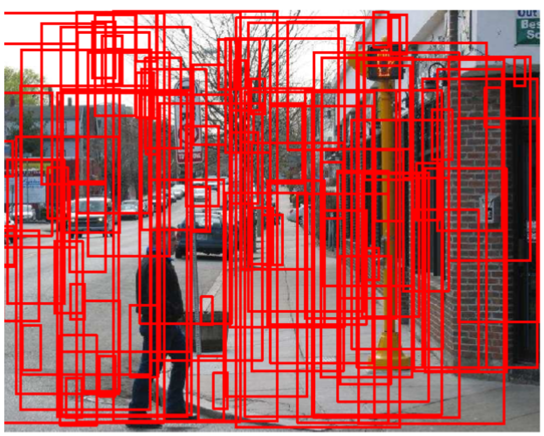
Image



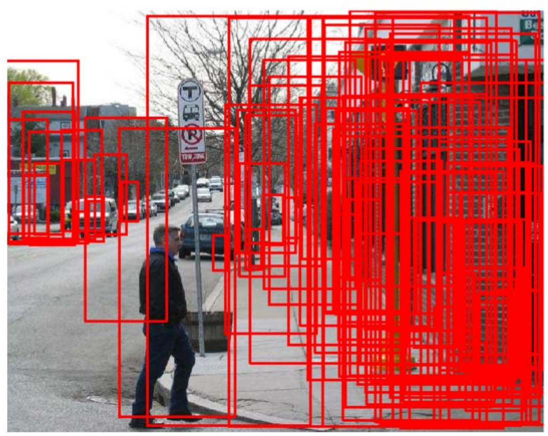
P(surfaces)



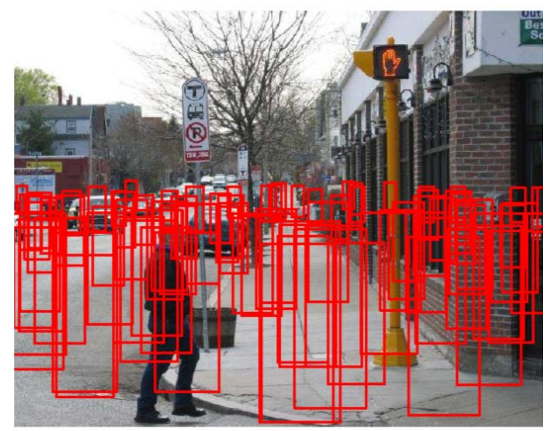
P(viewpoint)



P(object)

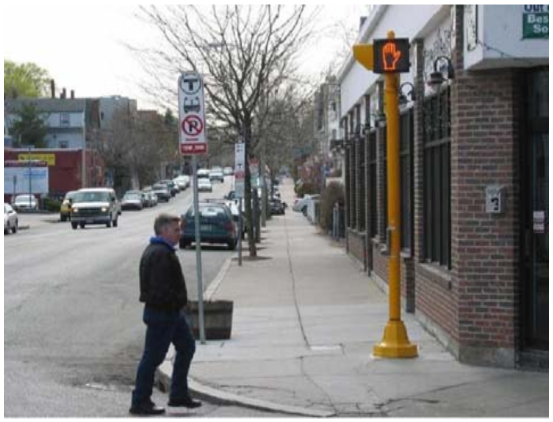


P(object | surfaces)

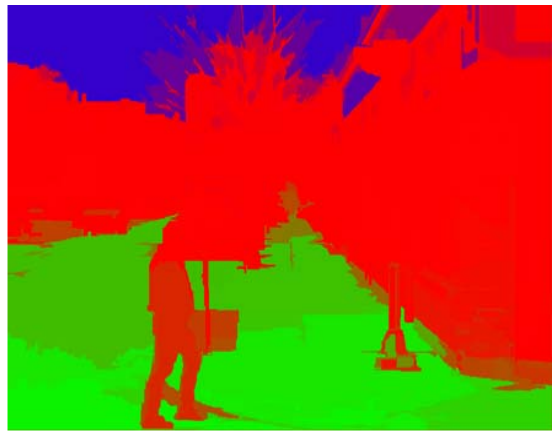


P(object | viewpoint)

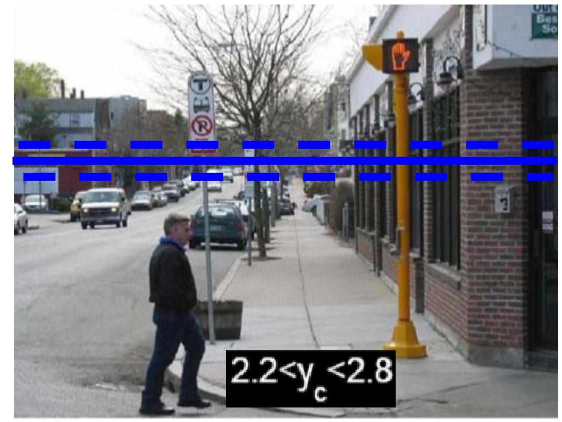
What does surface and viewpoint say about objects?



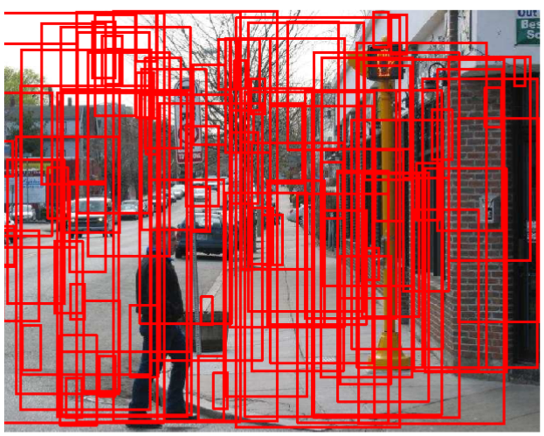
Image



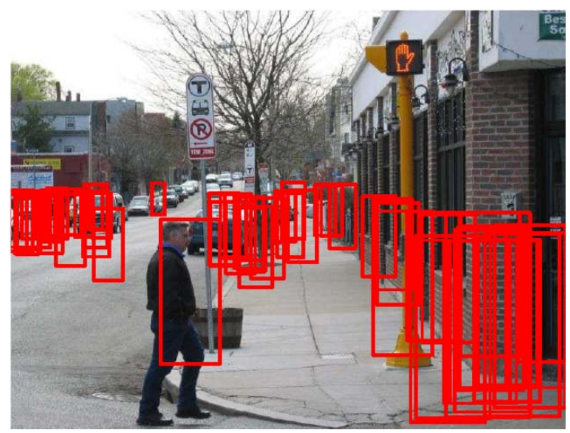
P(surfaces)



P(viewpoint)

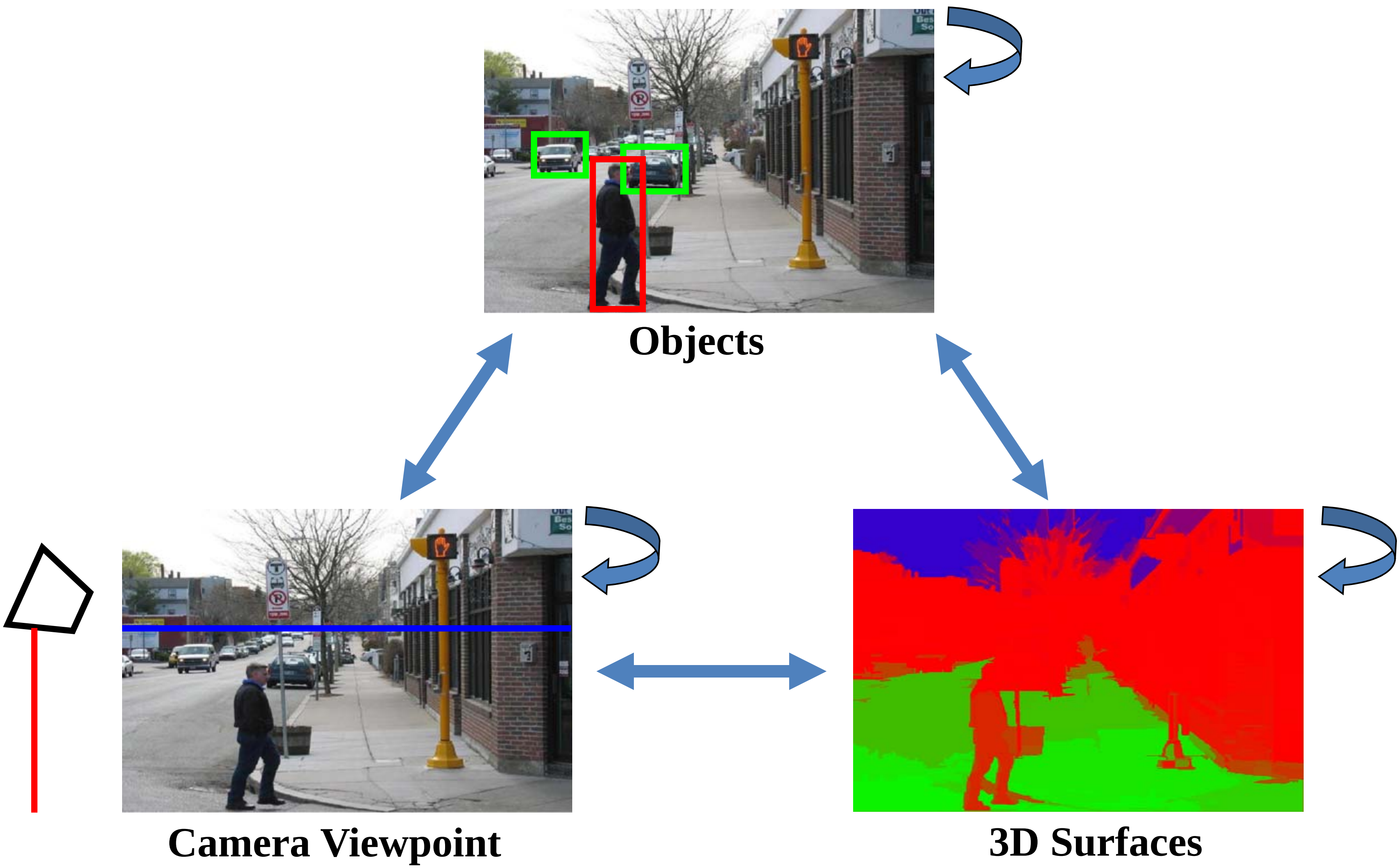


P(object)



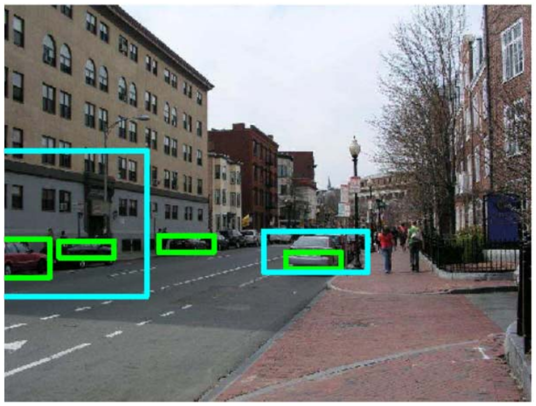
P(object | surfaces, viewpoint)

Scene Parts Are All Interconnected

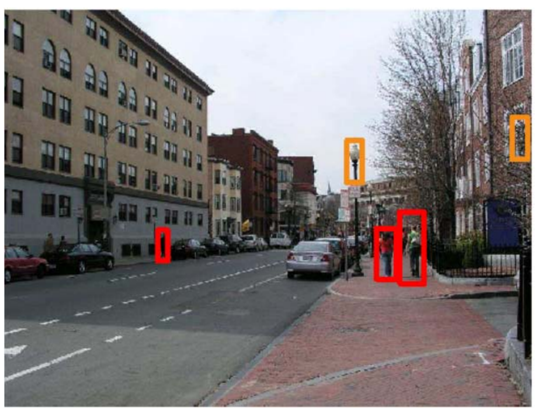


Input to Algorithm

Object Detection



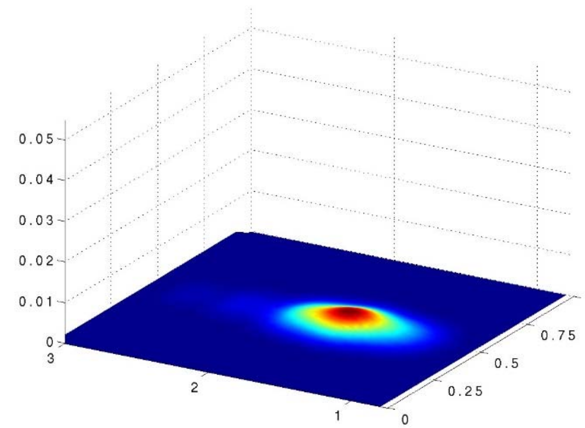
Local Car Detector



Local Ped Detector

Surface Estimates

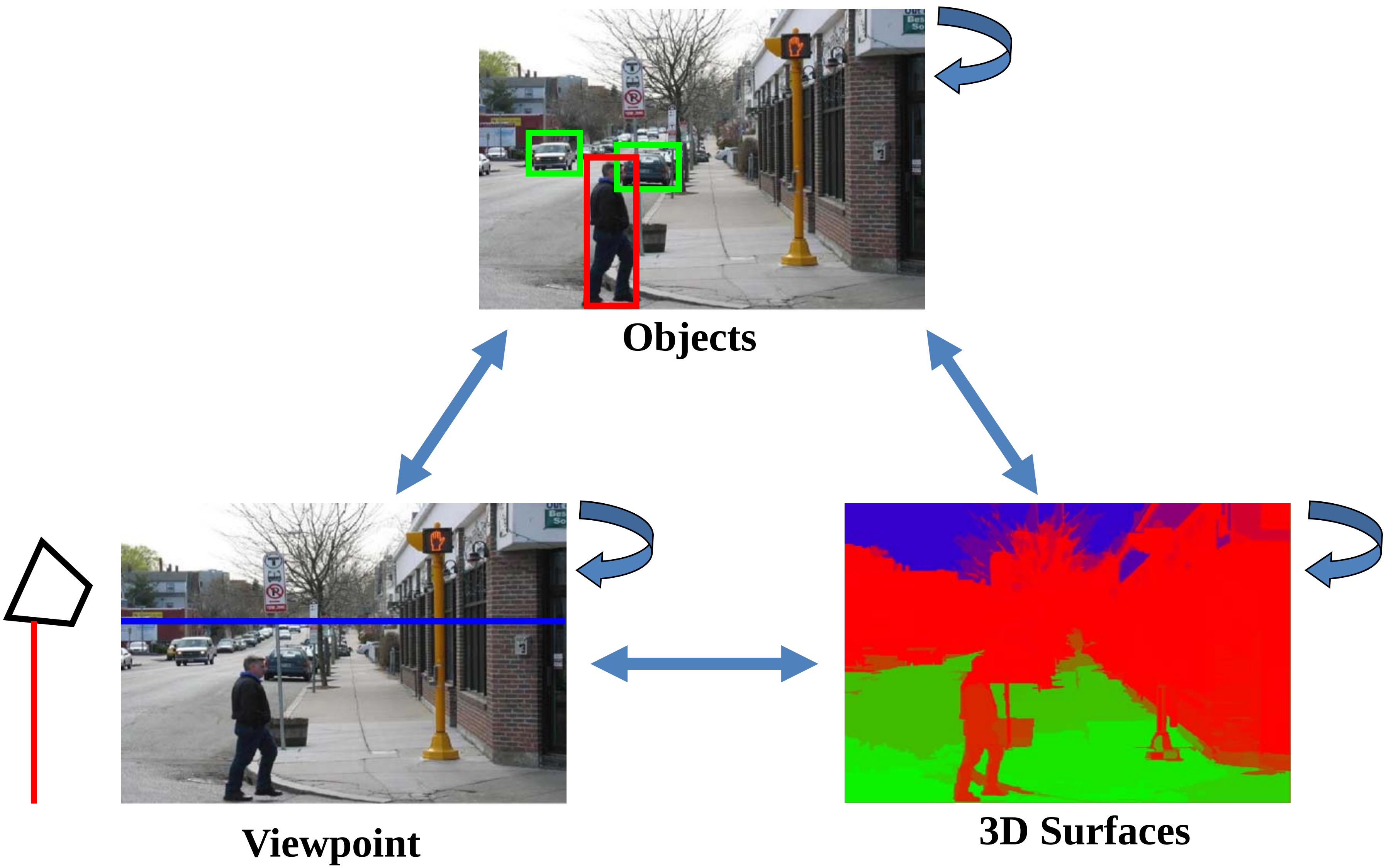
Viewpoint Prior



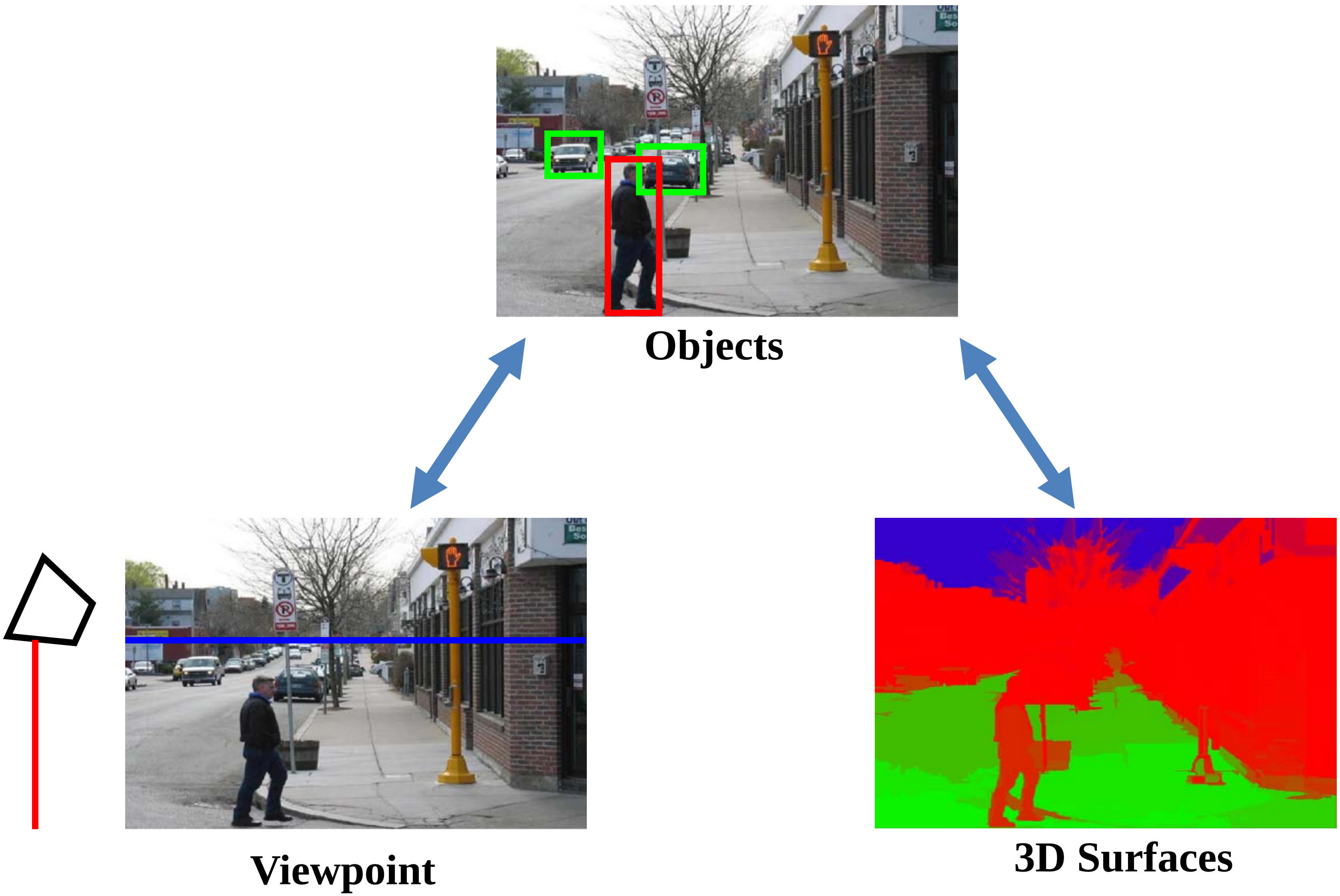
Local Detector: [Dalal-Triggs 2005]

Surfaces: [Hoiem-Efros-Hebert 2005]

Scene Parts Are All Interconnected



Approximate Model



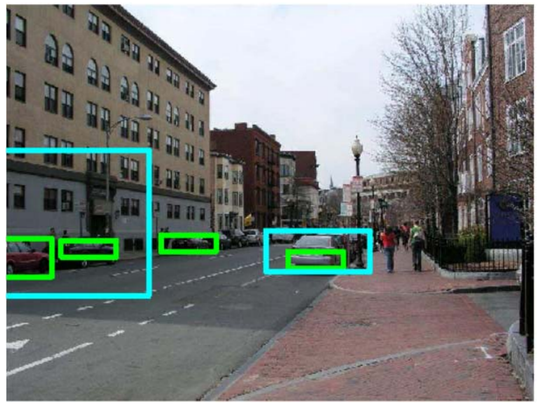
Object detection

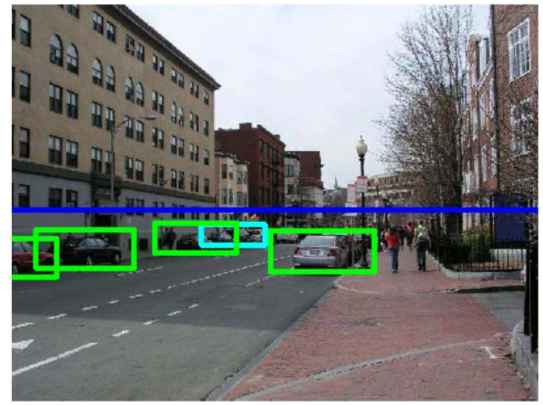
Car: TP / FP

Ped: TP / FP

Initial (Local) | Final (Global)

Car Detection

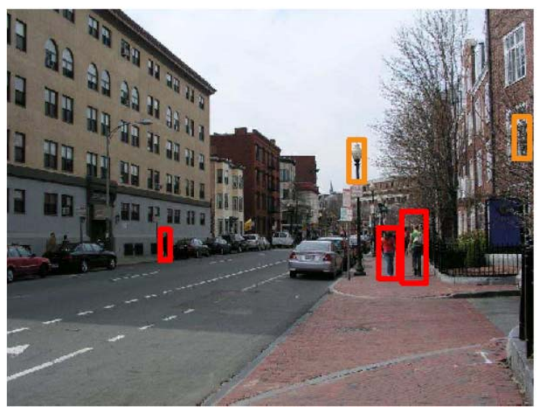


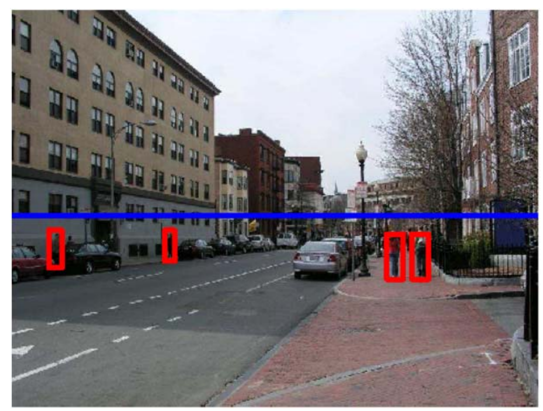


4 TP / 2 FP

4 TP / 1 FP

Ped Detection





3 TP / 2 FP

4 TP / 0 FP

Local Detector: [Dalal-Triggs 2005]

Experiments on LabelMe Dataset

- Testing with LabelMe dataset:
  - Cars as small as 14 pixels
  - Peds as small as 36 pixels



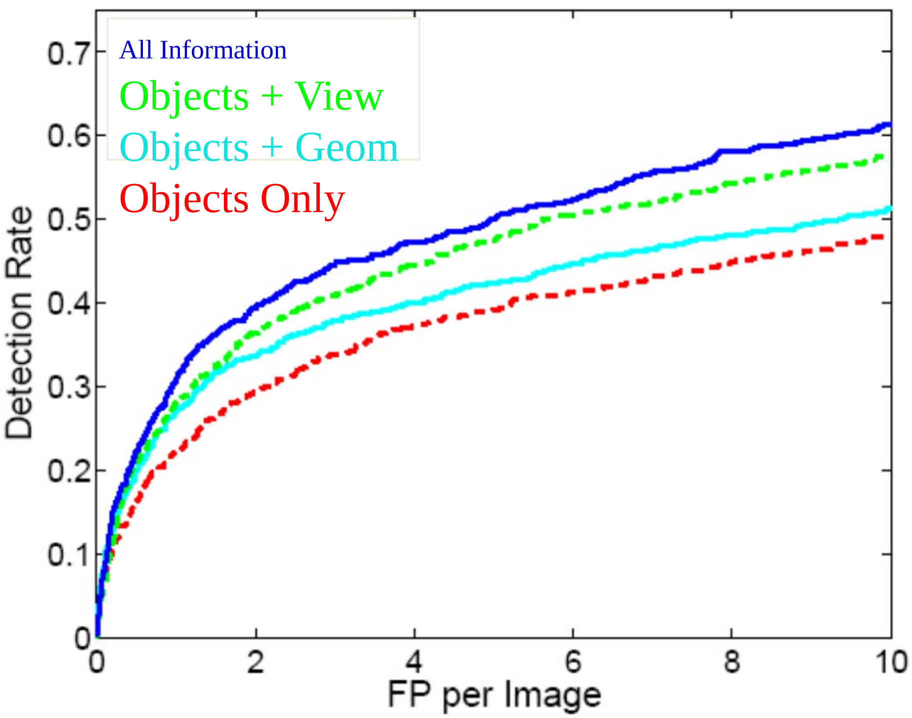




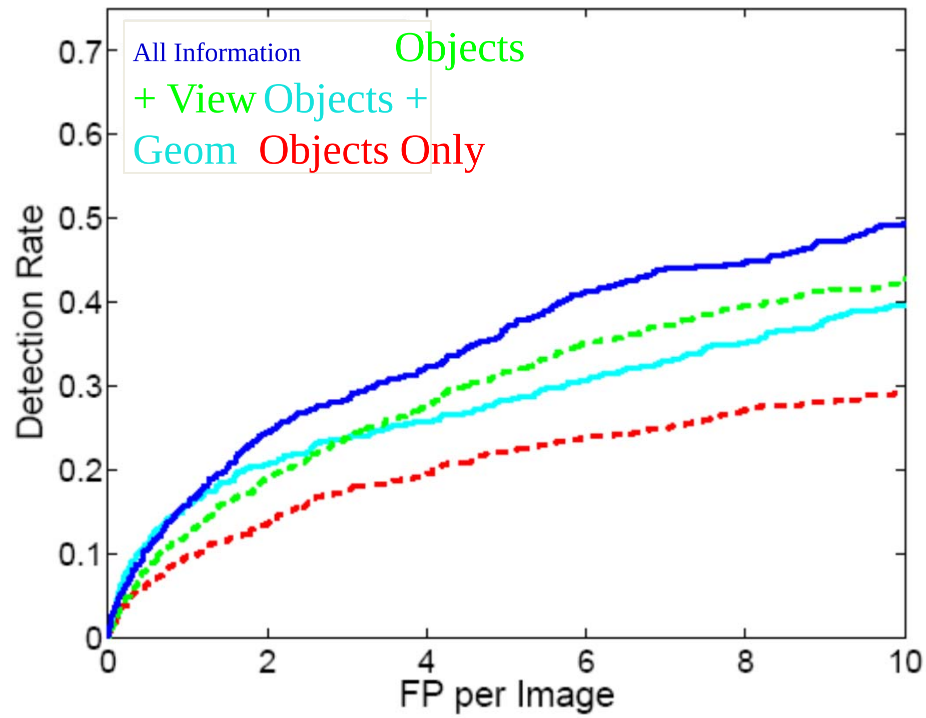

More Tasks → Better Detection

Local Detector from Murphy et al. 2003

Car Detection



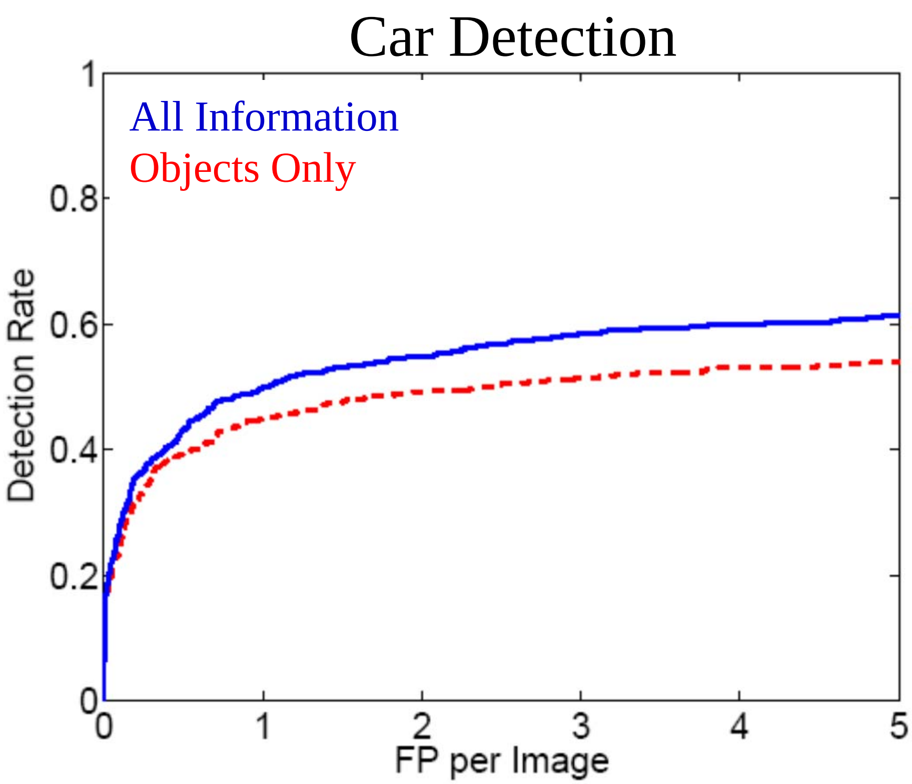
Pedestrian Detection

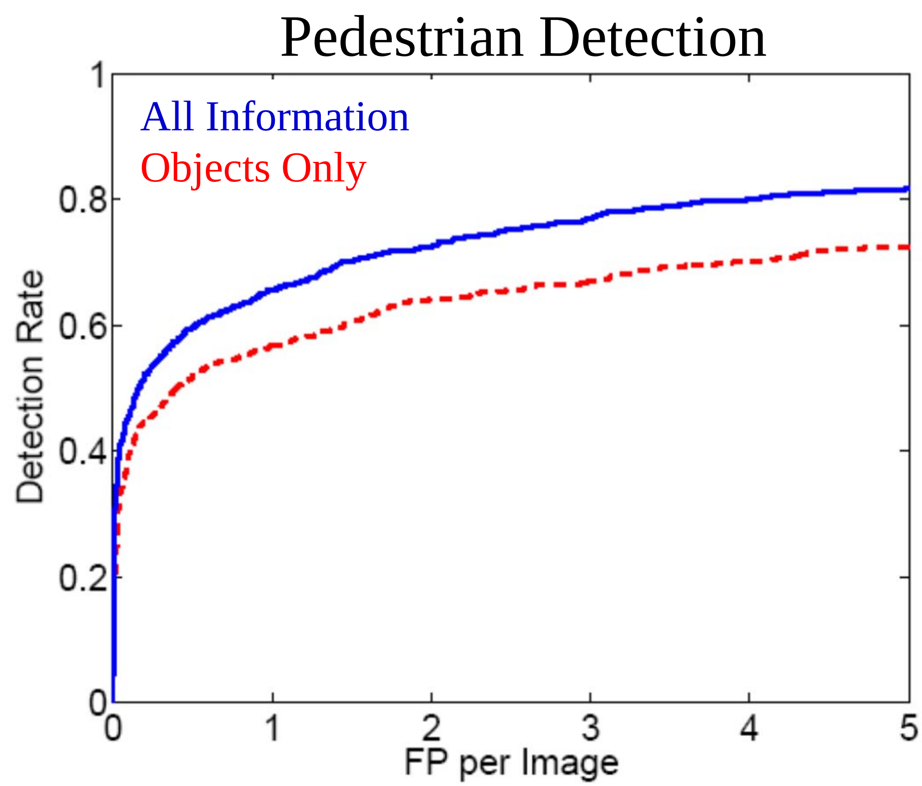


[Hoiem Efros Hebert 2006]

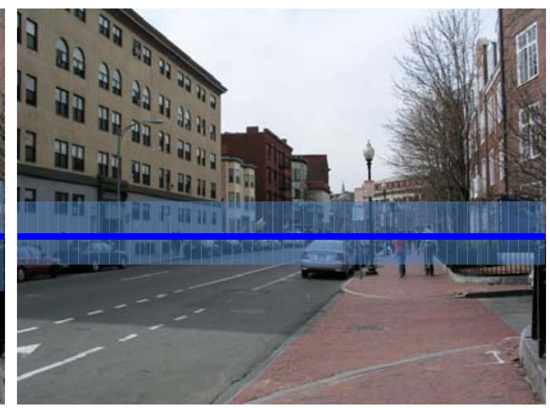
Good Detectors Become Better

Local Detector from Dalal-Triggs 2005





Better Detectors → Better Viewpoint

	Horizon Prior	Using 2003 Local Detector	Using 2005 Local Detector
Median Error:	8.5%	3.8%	3.0%
90% Bound:			

More is Better

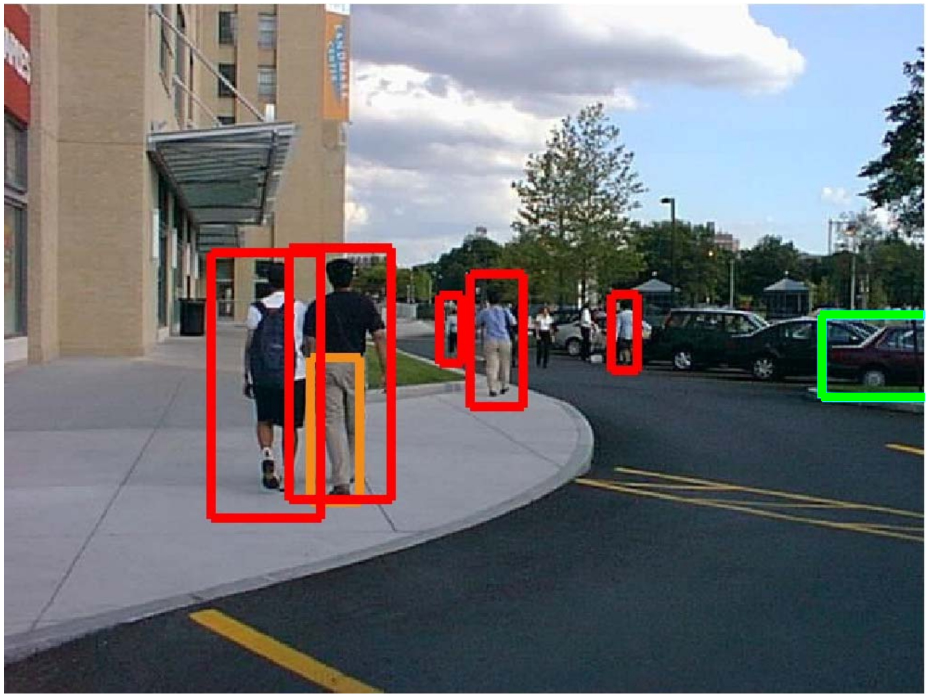
More objects →	Better viewpoint estimates
Detect Cars Only	7.3% Error
Detect Peds Only	5.0% Error
Detect Both	**3.8% Error**

Better viewpoint → Better object detection

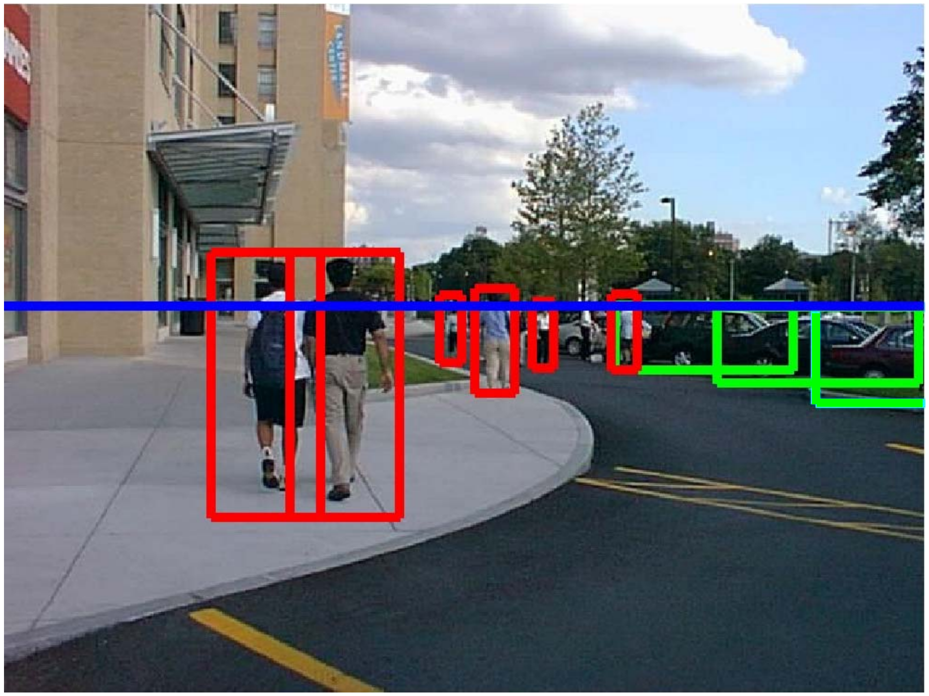
10% fewer false positives at same detection rate

Qualitative Results

Car: TP / FP Ped: TP / FP



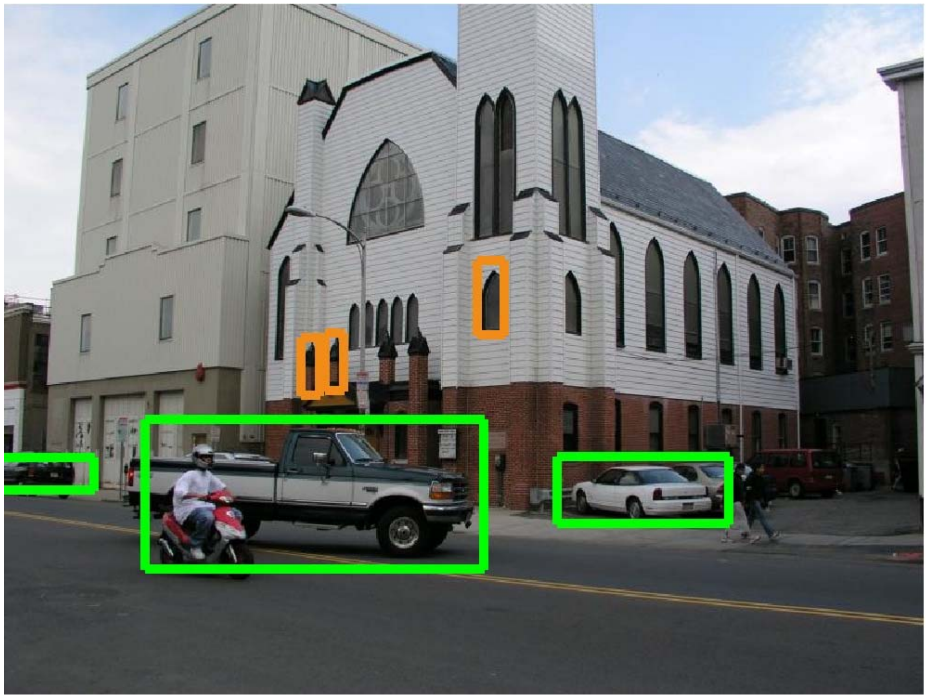
Initial: 6 TP / 1 FP



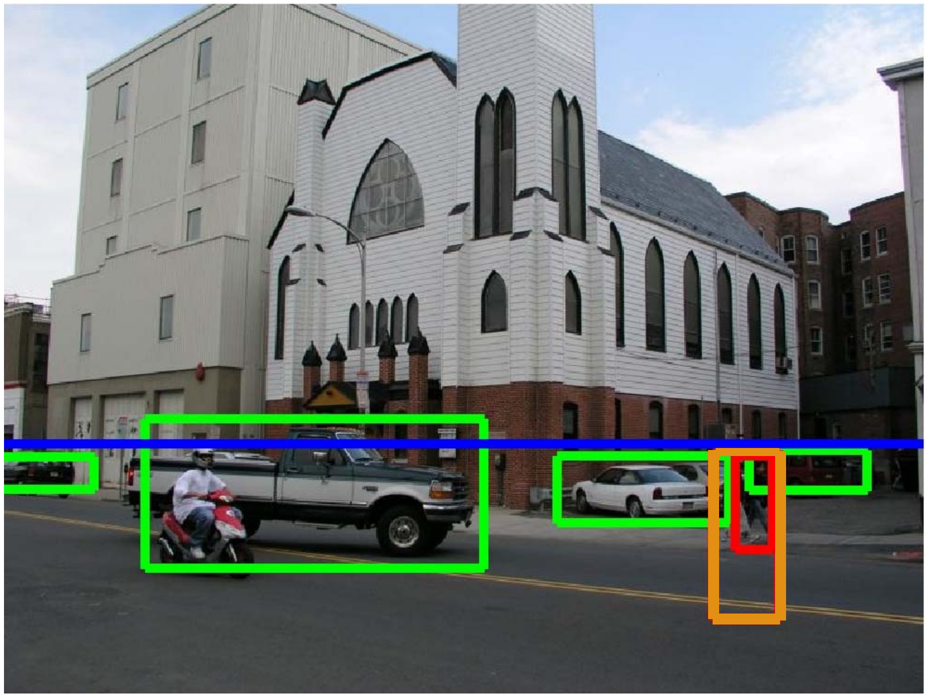
Final: 9 TP / 0 FP

Qualitative Results

Car: TP / FP Ped: TP / FP

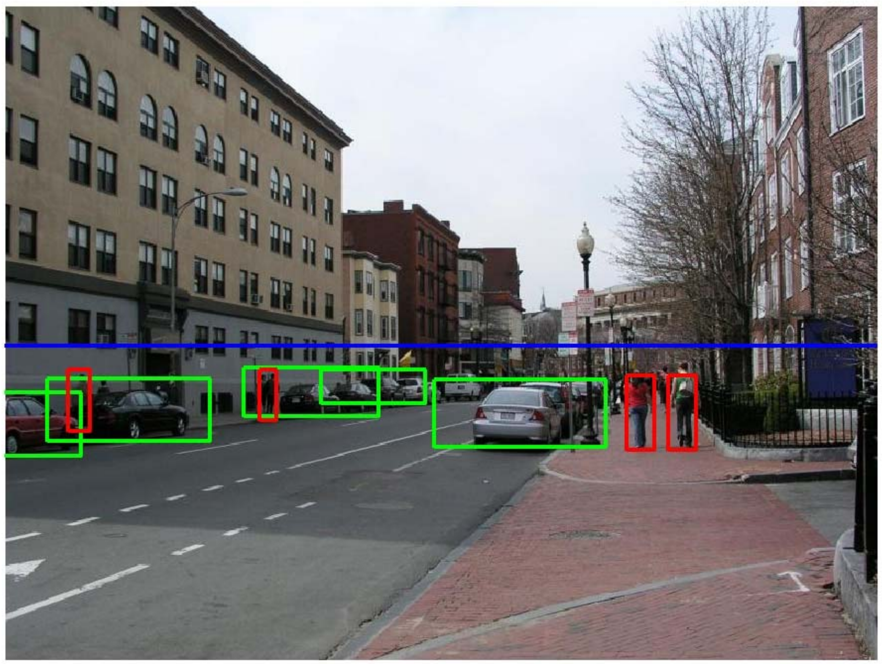


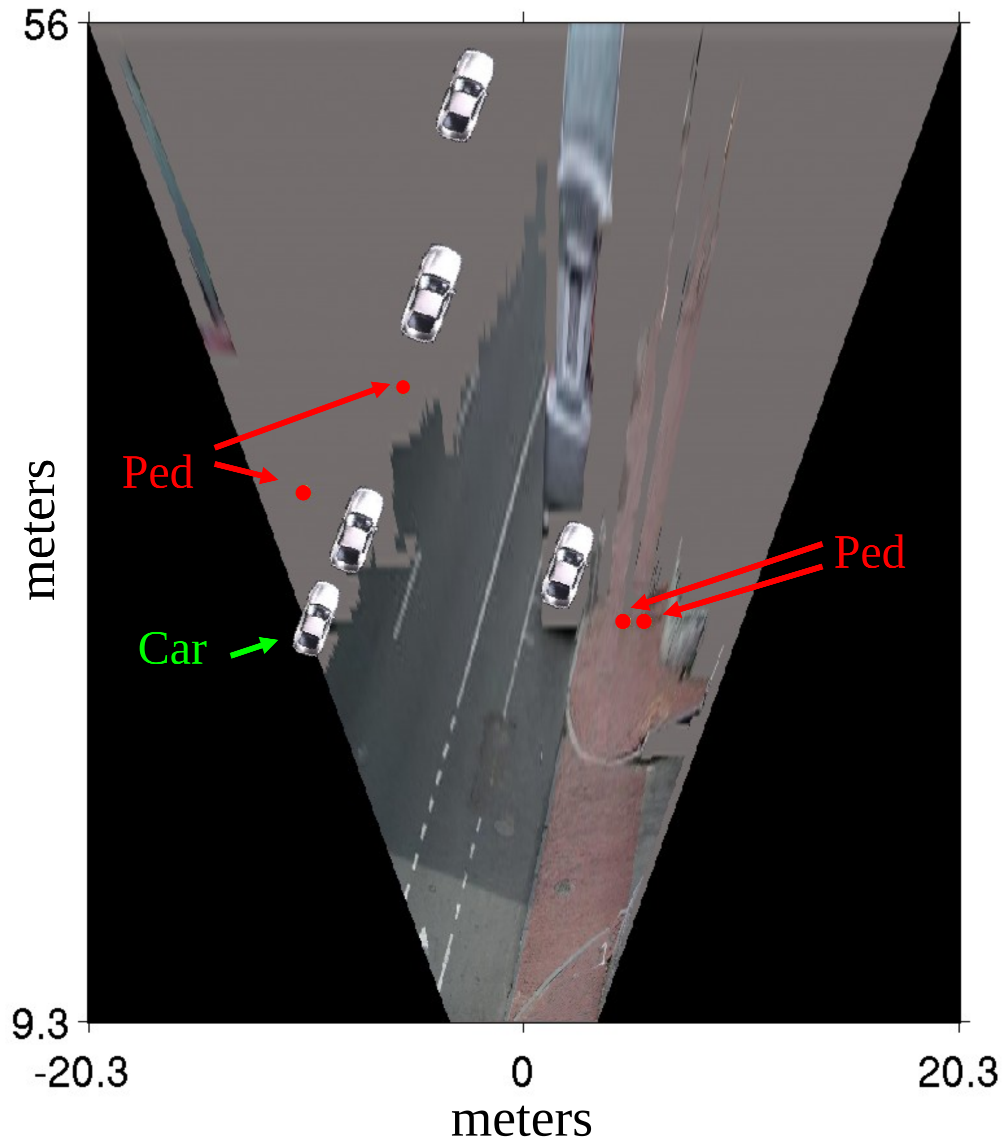
Initial: 3 TP / 3 FP



Final: 5 TP / 1 FP

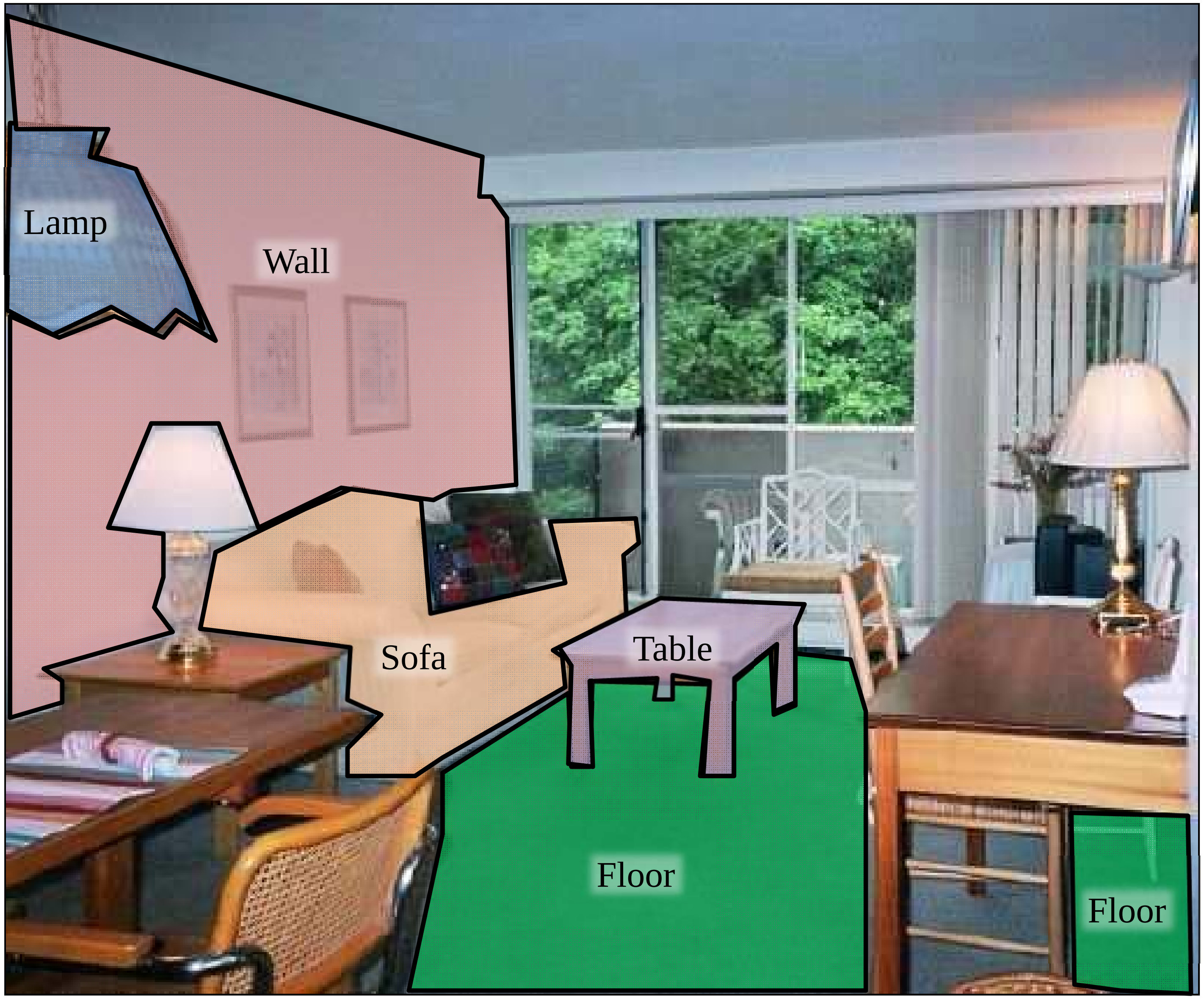
Putting Objects in Perspective



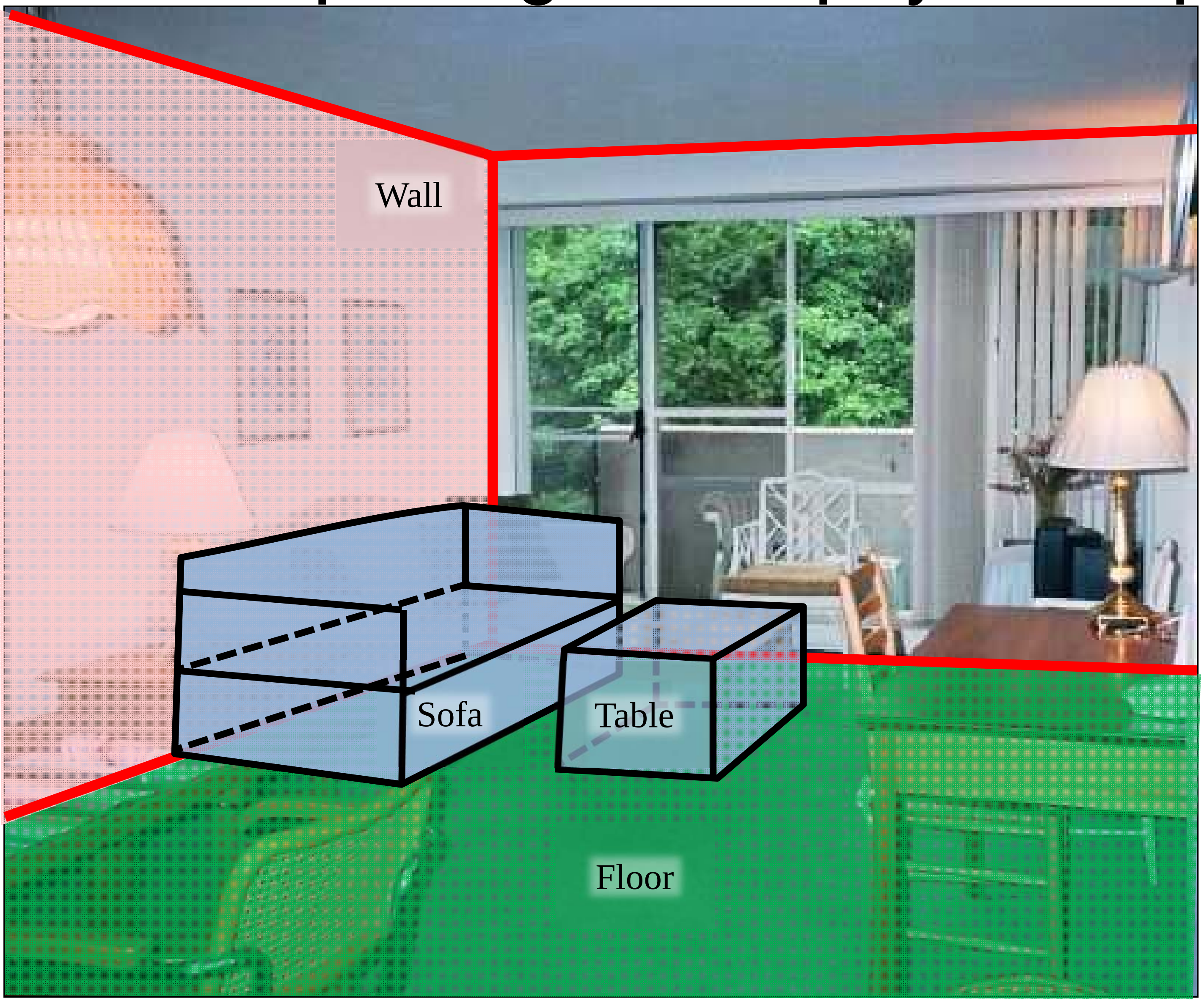


Interpretation of indoor scenes

Vision = assigning labels to pixels?



Vision = interpreting within physical space



Physical space needed for affordance

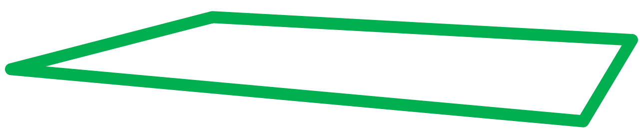
Is this a good place to sit?

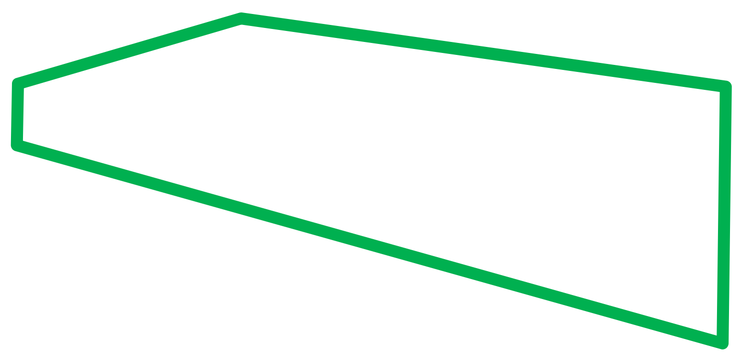
Could I stand over here?

Can I put my cup here?

Walkable path

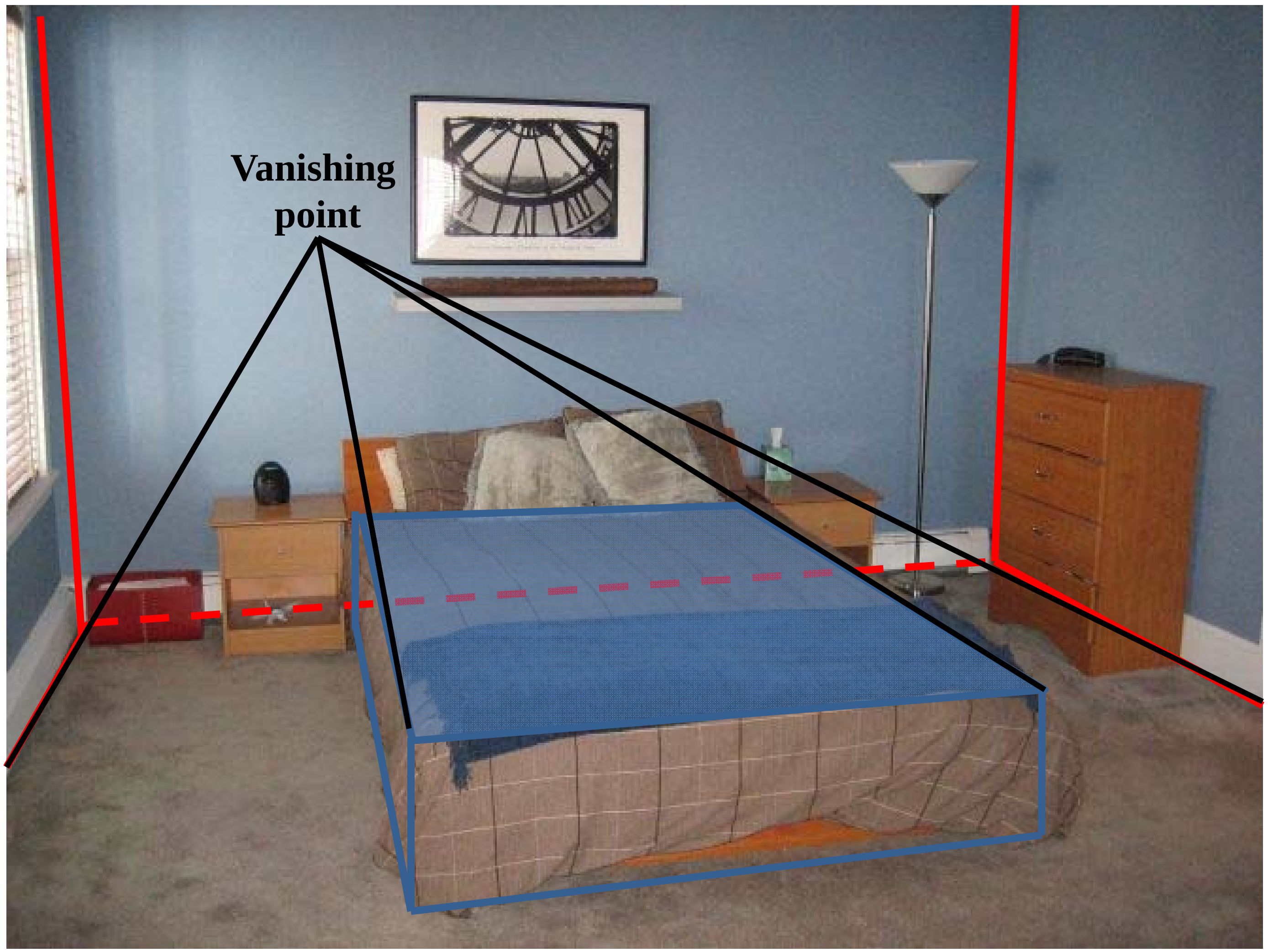
Physical space needed for recognition





Apparent shape depends strongly on viewpoint

Physical space needed for recognition



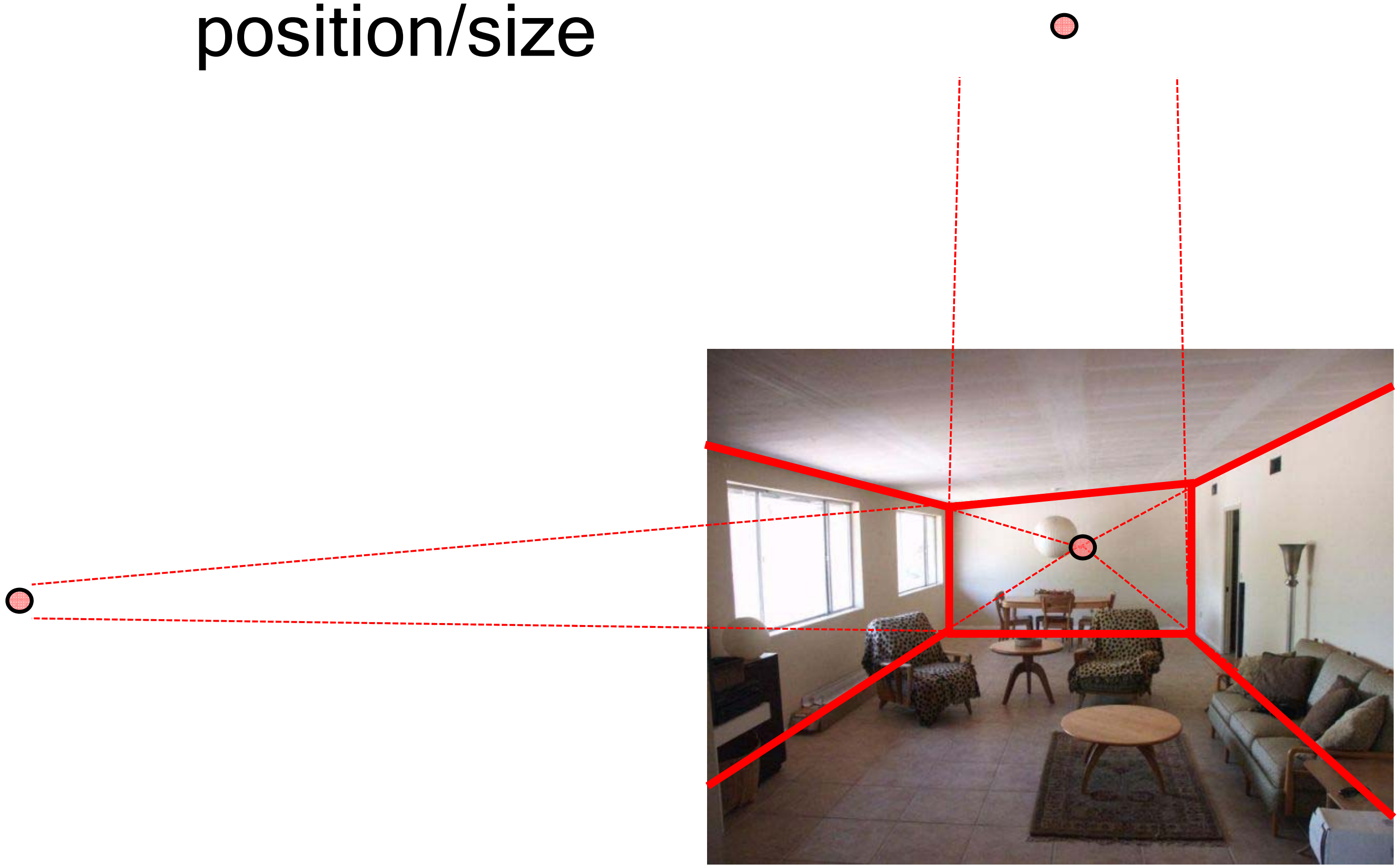
Key challenges

- How to represent the physical space?
  - *Requires seeing beyond the visible*
- How to estimate the physical space?
  - Requires simplified models
  - Requires learning from examples

Hedau Hoiem Forsyth, ICCV 2009

Box Layout

- Room is an oriented 3D box
  - Three vanishing points specify orientation
  - Two pairs of sampled rays specify position/size



Box Layout

Another box consistent with the same vanishing points

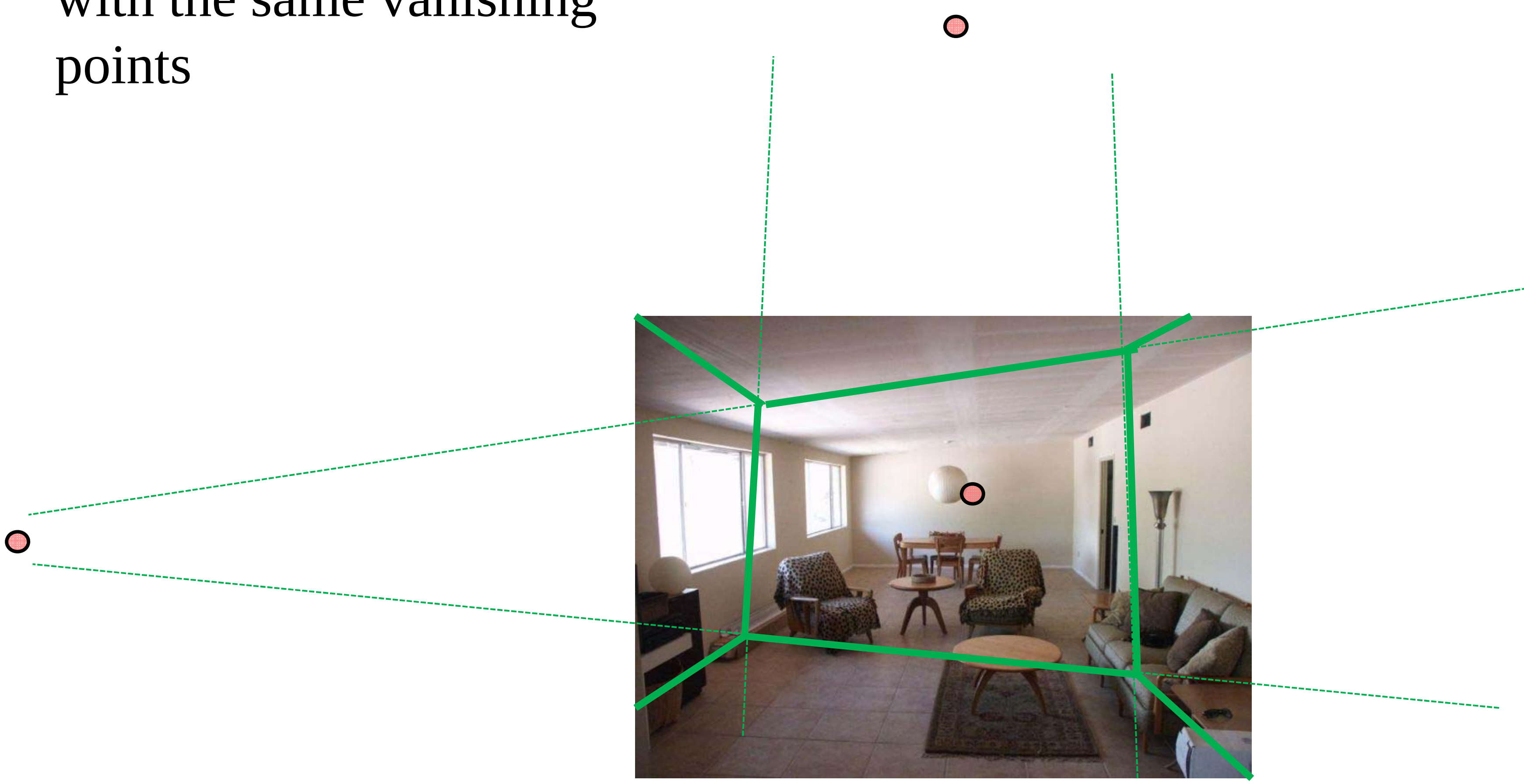
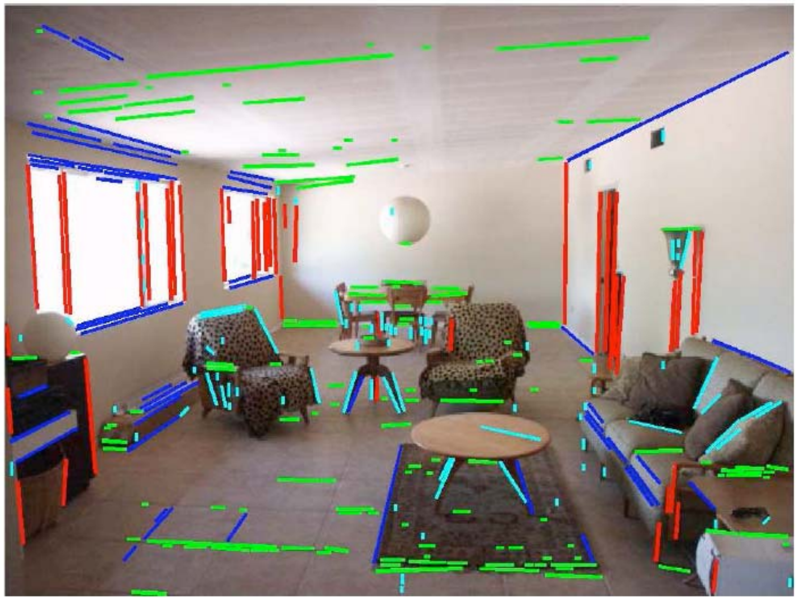
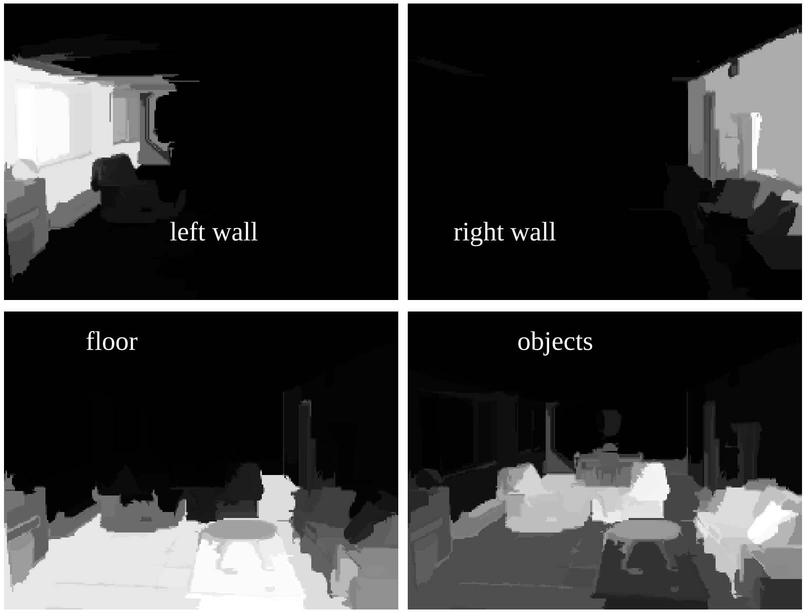


Image Cues for Box Layout

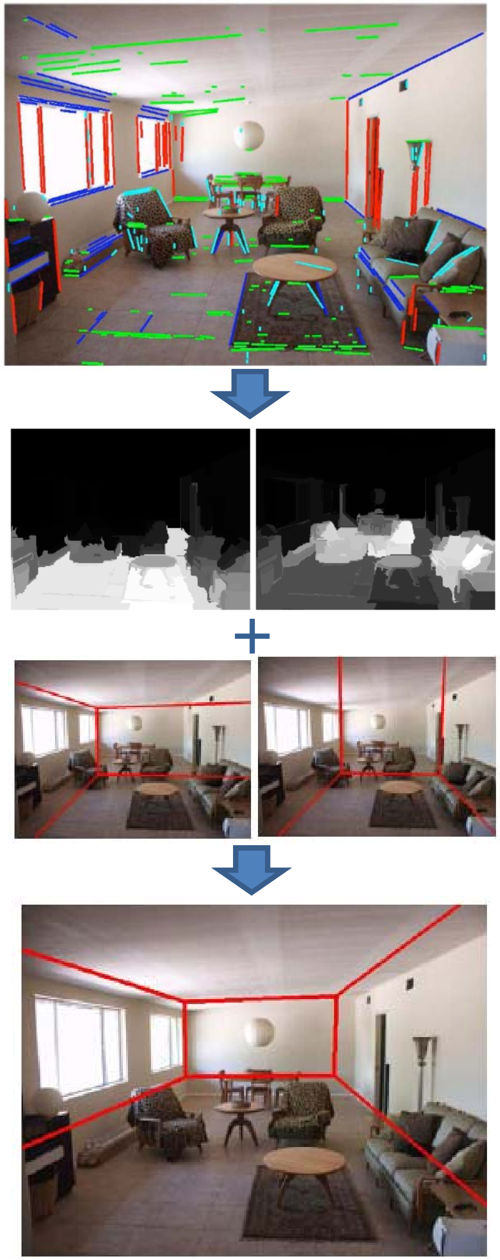
- Straight edges
  - Edges on floor/wall surfaces are usually oriented towards VPs
  - Edges on objects might mislead



- Appearance of visible surfaces
  - Floor, wall, ceiling, object labels should be consistent with box



Box Layout Algorithm



1. Detect edges

2. Estimate 3 orthogonal vanishing points

3. Apply region classifier to label pixels with visible surfaces
   – Boosted decision trees on region based on color, texture, edges, position

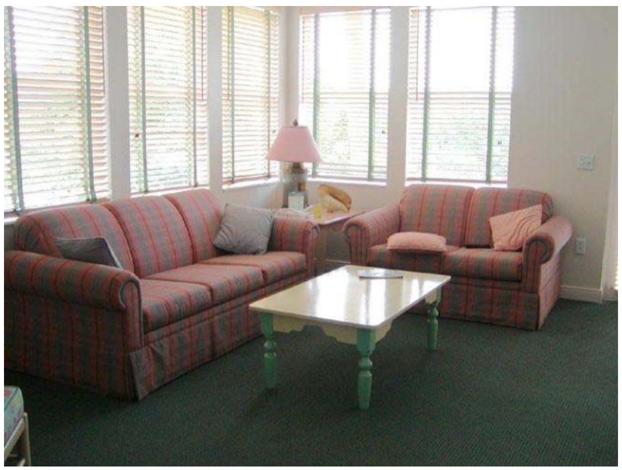
4. Generate box candidates by sampling pairs of rays from VPs

5. Score each box based on edges and pixel labels
   – Learn score via structured learning

6. Jointly refine box layout and pixel labels to get final estimate

Evaluation

- Dataset: 308 indoor images
  - Train with 204 images, test with 104 images






Experimental results

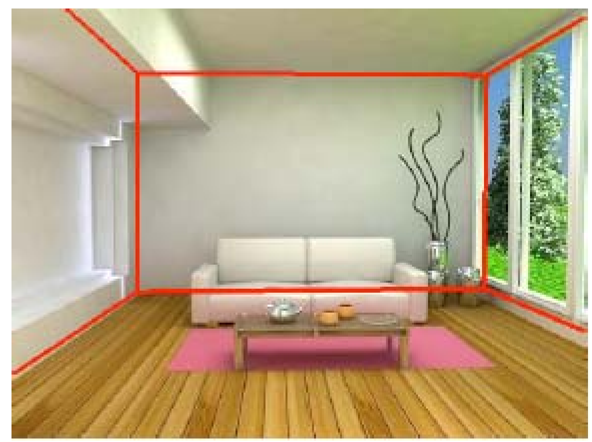

Detected Edges


Surface Labels

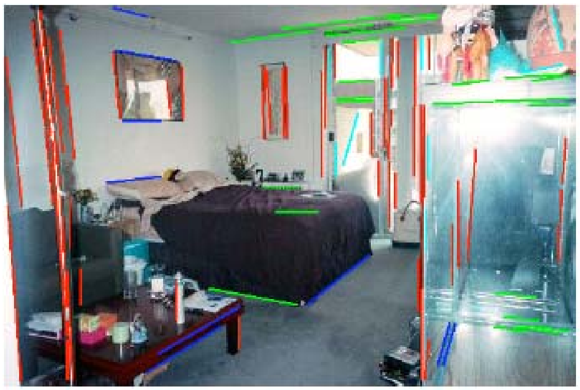

Box Layout


Detected Edges

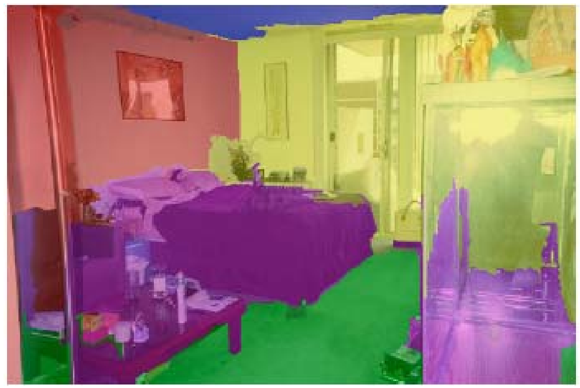

Surface Labels


Box Layout

Experimental results



Detected Edges



Surface Labels

Box Layout

Detected Edges

Surface Labels

Box Layout

Experimental results

- Joint reasoning of surface label / box layout helps
  - Pixel error: 26.5% → 21.2%
  - Corner error: 7.4% → 6.3%
- Similar performance for cluttered and uncluttered rooms

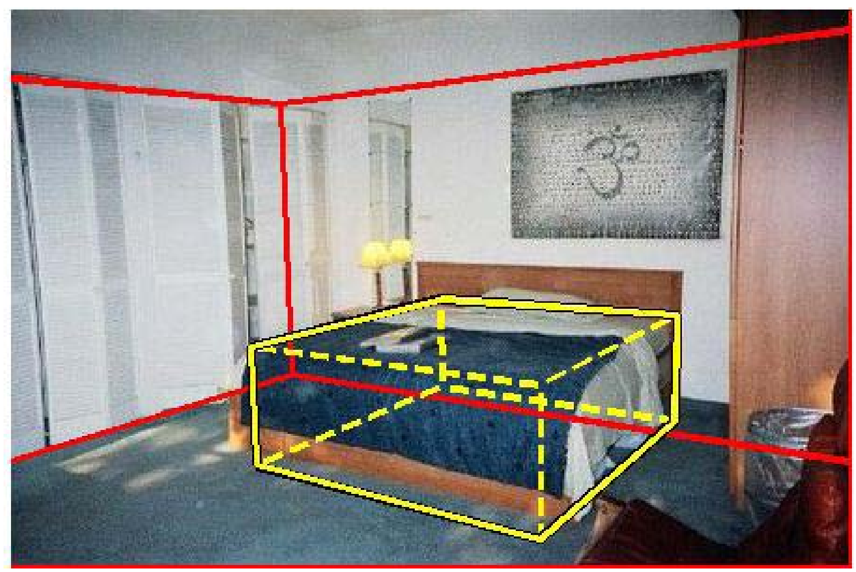
Using room layout to improve object detection

Box layout helps

1. Predict the appearance of objects, because they are often aligned with the room
2. Predict the position and size of objects, due to physical constraints and size consistency

2D Bed Detection

3D Bed Detection with Scene Geometry



Hedau, Hoiem, Forsyth, ECCV 2010, CVPR 2012