# Decomposing a Scene into Geometric and Semantically Consistent Regions


Stephen Gould
sgould@stanford.edu

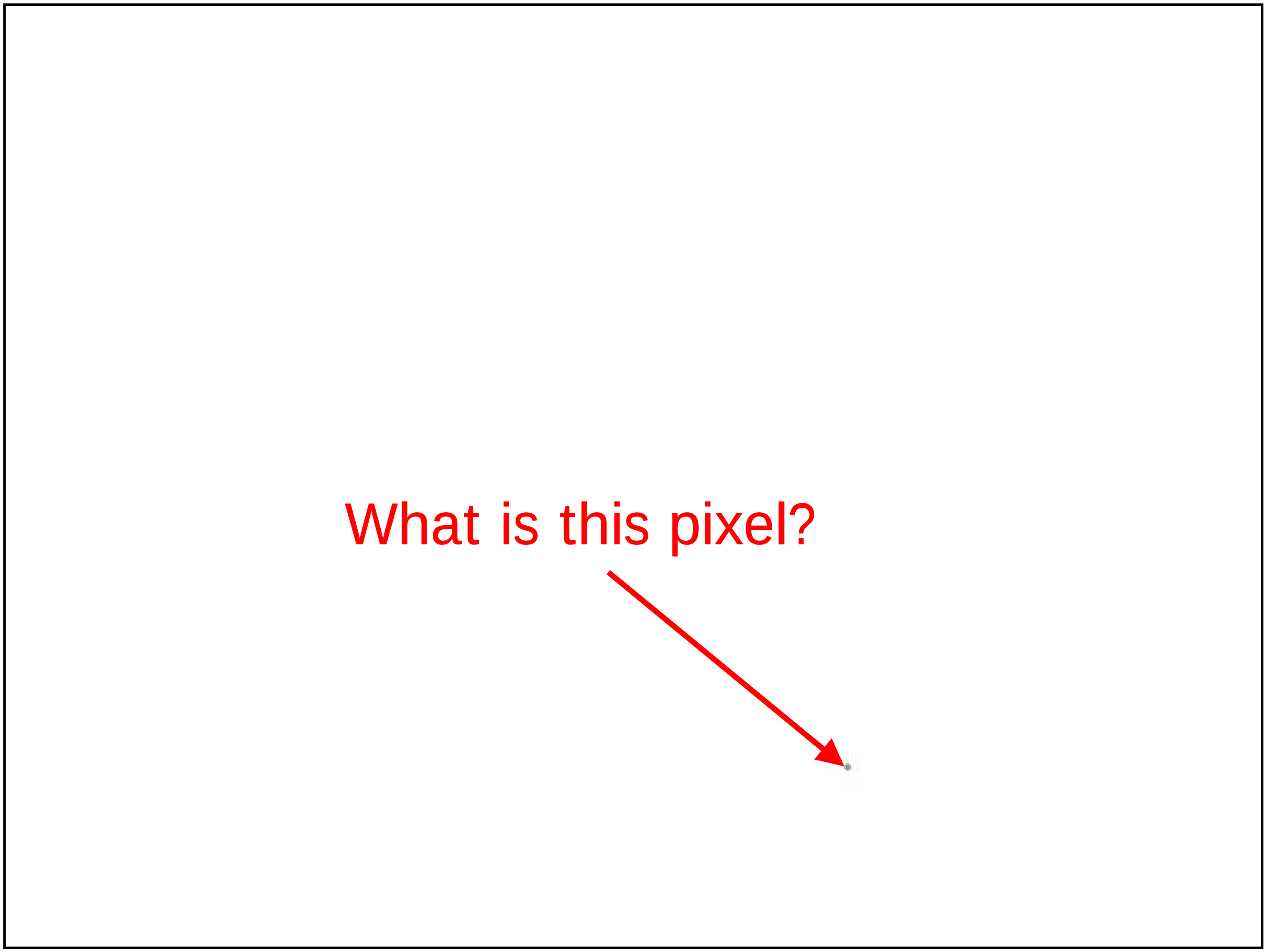
Richard Fulton
rafulton@cs.stanford.edu

Daphne Koller
koller@cs.stanford.edu



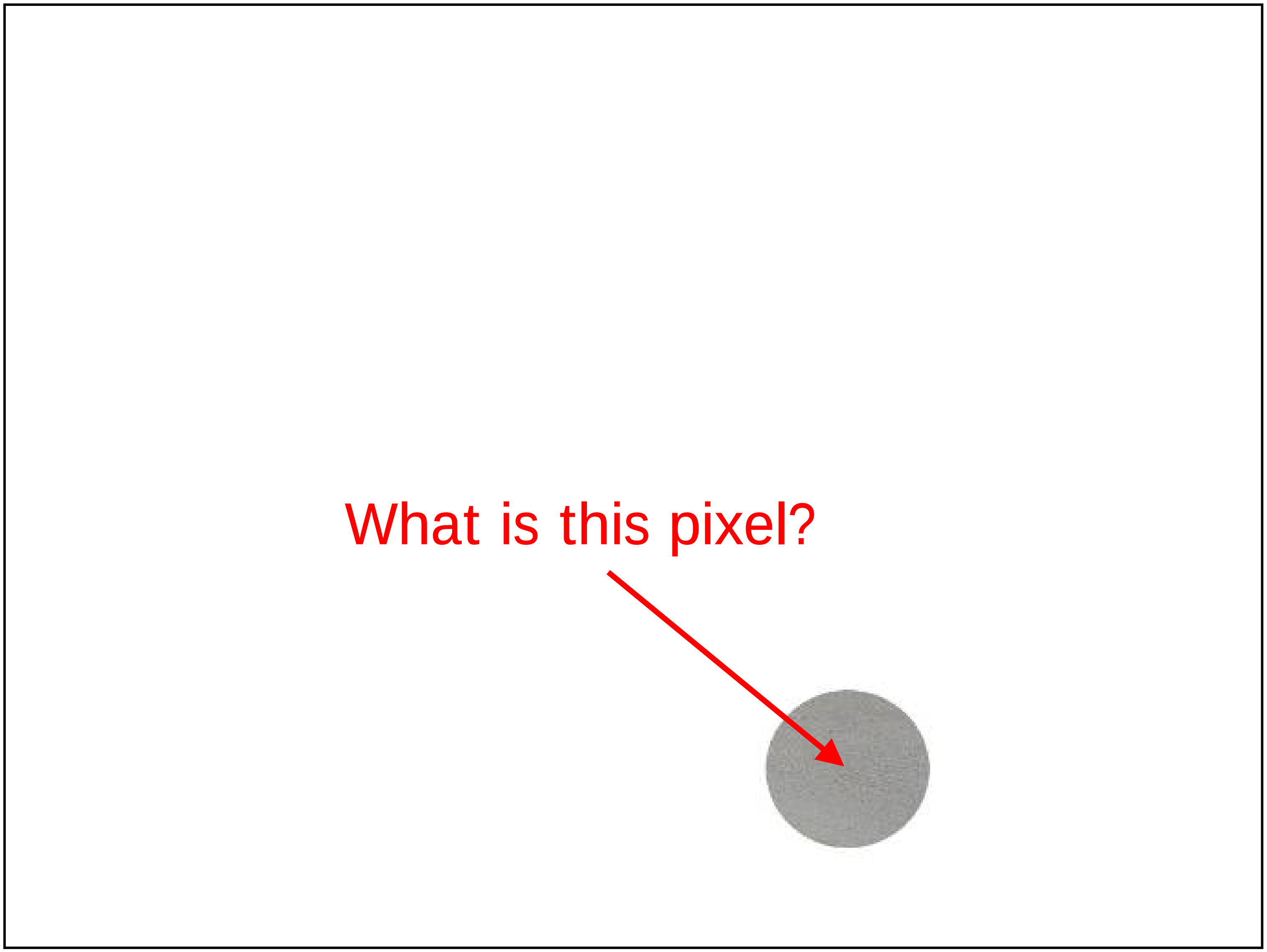
IEEE International Conference on Computer Vision
September 2009

# Segmentation and Context



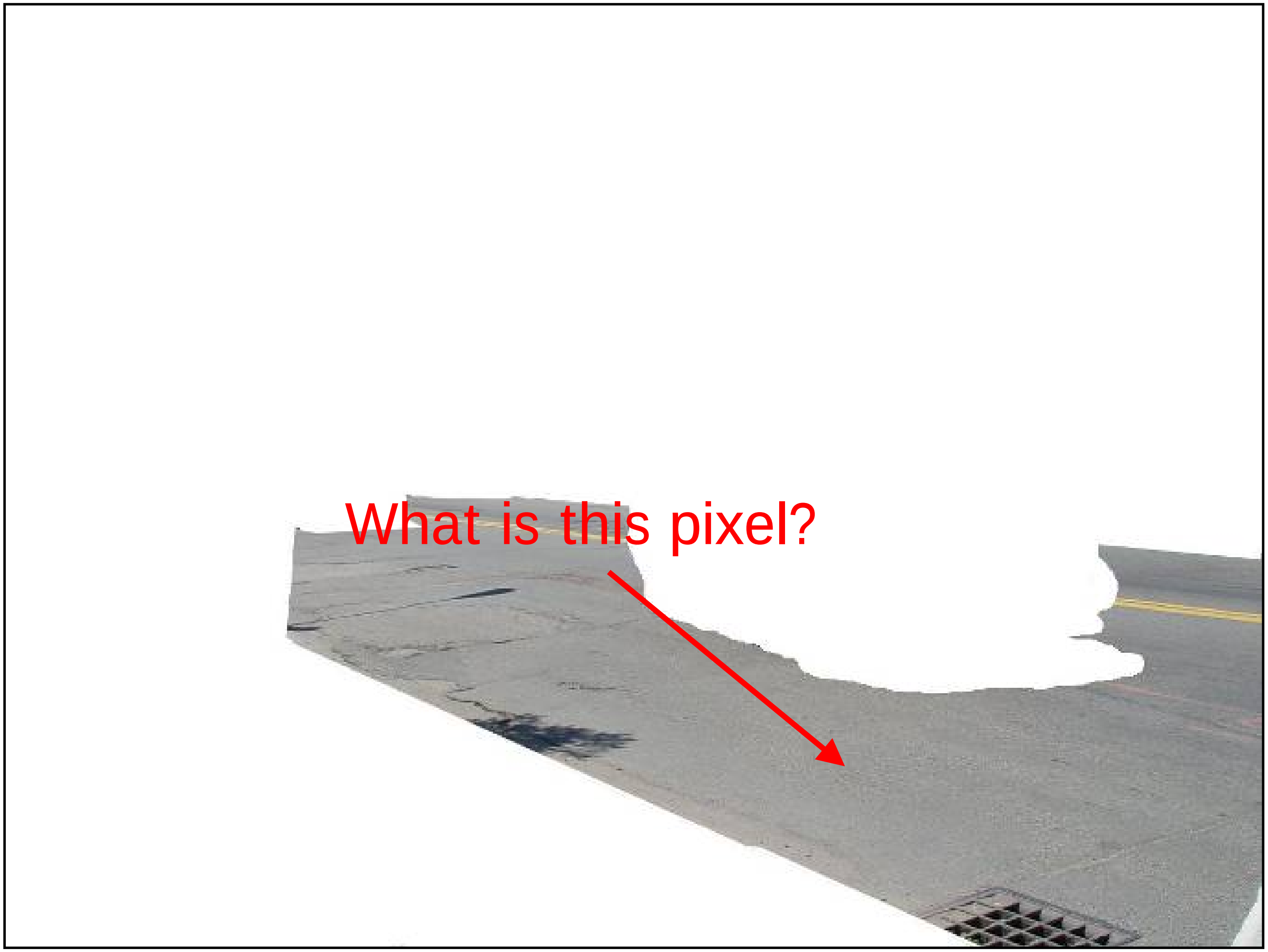

Gould, et al.

# Segmentation and Context



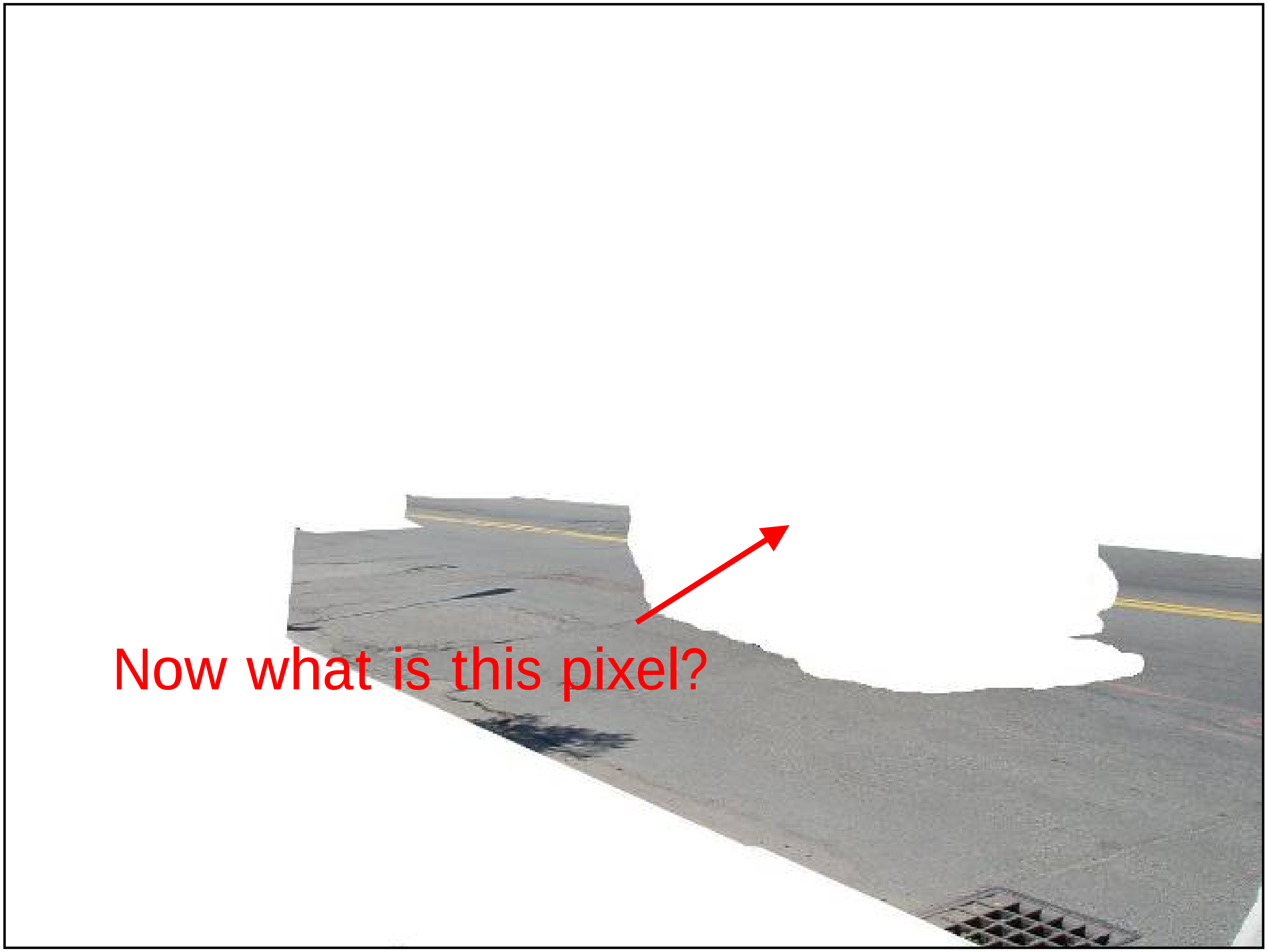

Gould, et al.

# Segmentation and Context



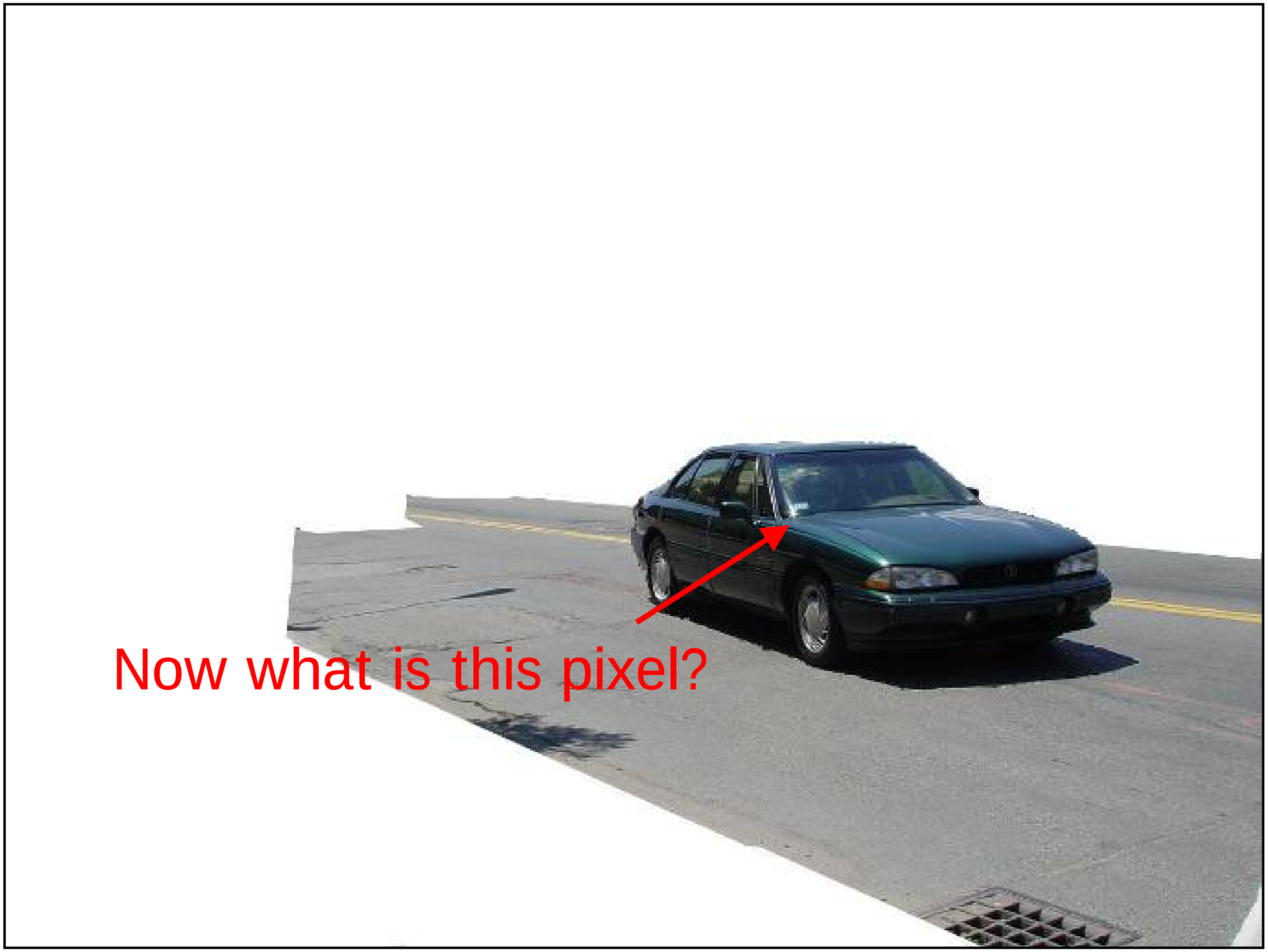

Gould, et al.

# Segmentation and Context




Gould, et al.

# Segmentation and Context



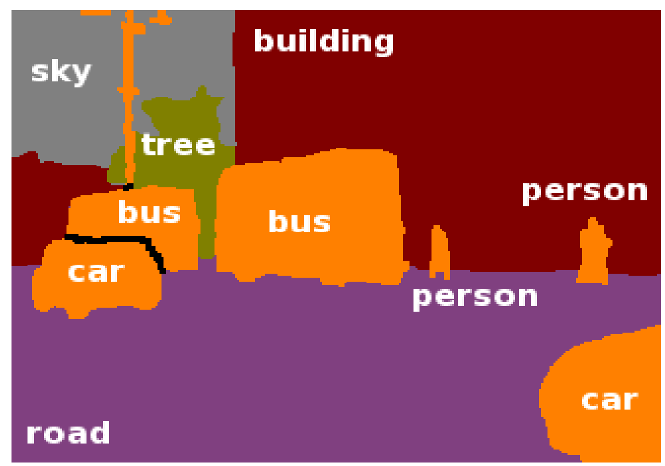

Gould, et al.

# Segmentation and Context

Gould, et al.

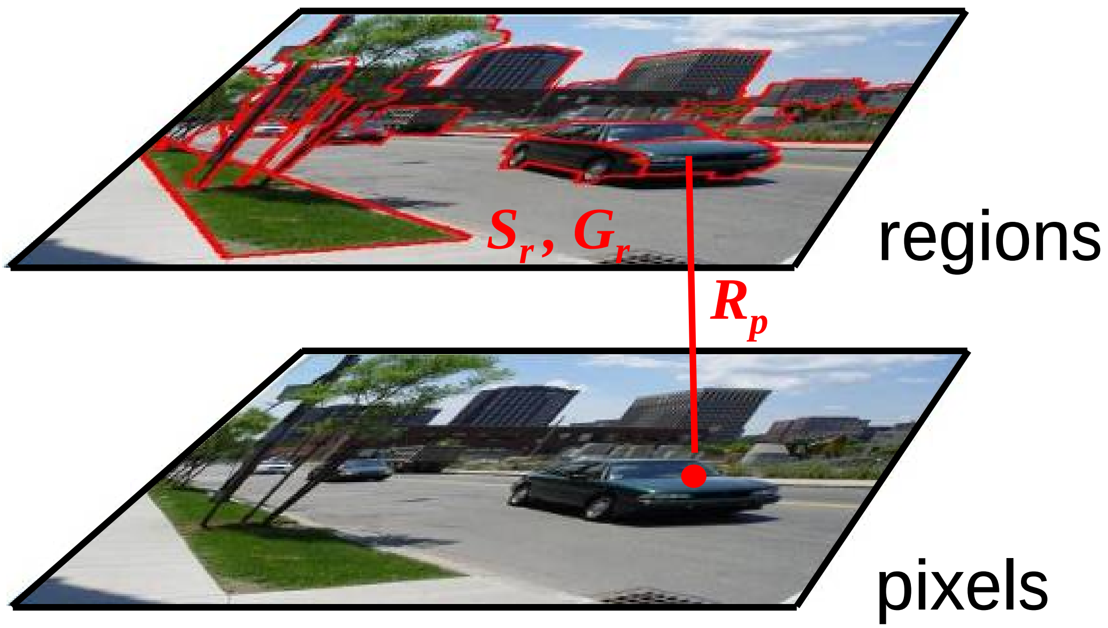
# Goal of Scene Decomposition

- Decompose the scene into **regions** with
  - semantic region labels (e.g., road, sky, building, etc.)
  - coherent geometric placement (orientation and location with respect to the horizon)


Gould, et al.

# Region-based Model




**Variables**

$\alpha_p$: pixel appearance
$\mathbf{R_p}$: pixel-to-region correspondence
$\mathbf{A_r}$: region appearance
$\mathbf{S_r}$: region semantic class
$\mathbf{G_r}$: region geometry
$\mathbf{v^{hz}}$: location of horizon

***Energy Function***

$$E(\mathbf{R}, \mathbf{A}, \mathbf{S}, \mathbf{G}, v^{hz}, K \mid I, \theta)$$

Gould, et al.

# Energy Function

$$E(\mathbf{R}, \mathbf{A}, \mathbf{S}, \mathbf{G}, v^{hz}, K \mid I, \theta)$$

$$=$$

$$\psi^{\text{horizon}}(v^{hz}) \quad \psi^{\text{region}}(S_r, G_r, A_r, v^{hz}) \quad \psi^{\text{boundary}}(A_r, A_s) \quad \psi^{\text{pair}}(S_r, S_s, G_r, G_s)$$

+

+

+

**Horizon Term**
e.g., vanishing lines

**Region Term**
e.g., consistent appearance and location

**Boundary Term**
e.g., difference in color/texture between regions

**Pairwise Term**
e.g., foreground on road

Gould, et al.

# Energy Function

$$E(\mathbf{R}, \mathbf{A}, \mathbf{S}, \mathbf{G}, v^{hz}, K \mid I, \theta)$$

$$=$$

$\psi^{horizon}(v^{hz})$ $\psi^{region}(S_r, G_r, A_r, v^{hz})$ $\psi^{boundary}(A_r, A_s)$ $\psi^{pair}(S_r,S_s,G_r,G_s)$




**Exact inference is intractable**

**H**
e.
lines

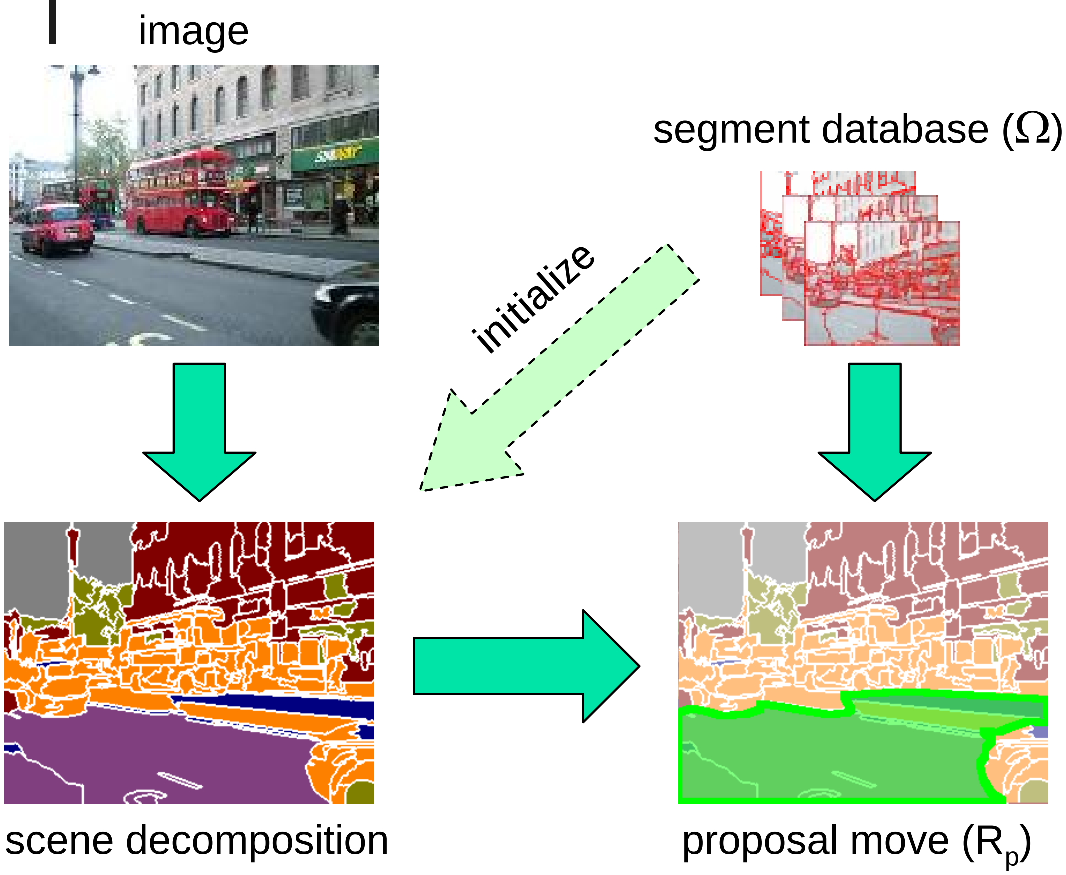
appearance and
location

in color/texture
between regions

**m**
nd
on road

Gould, et al.

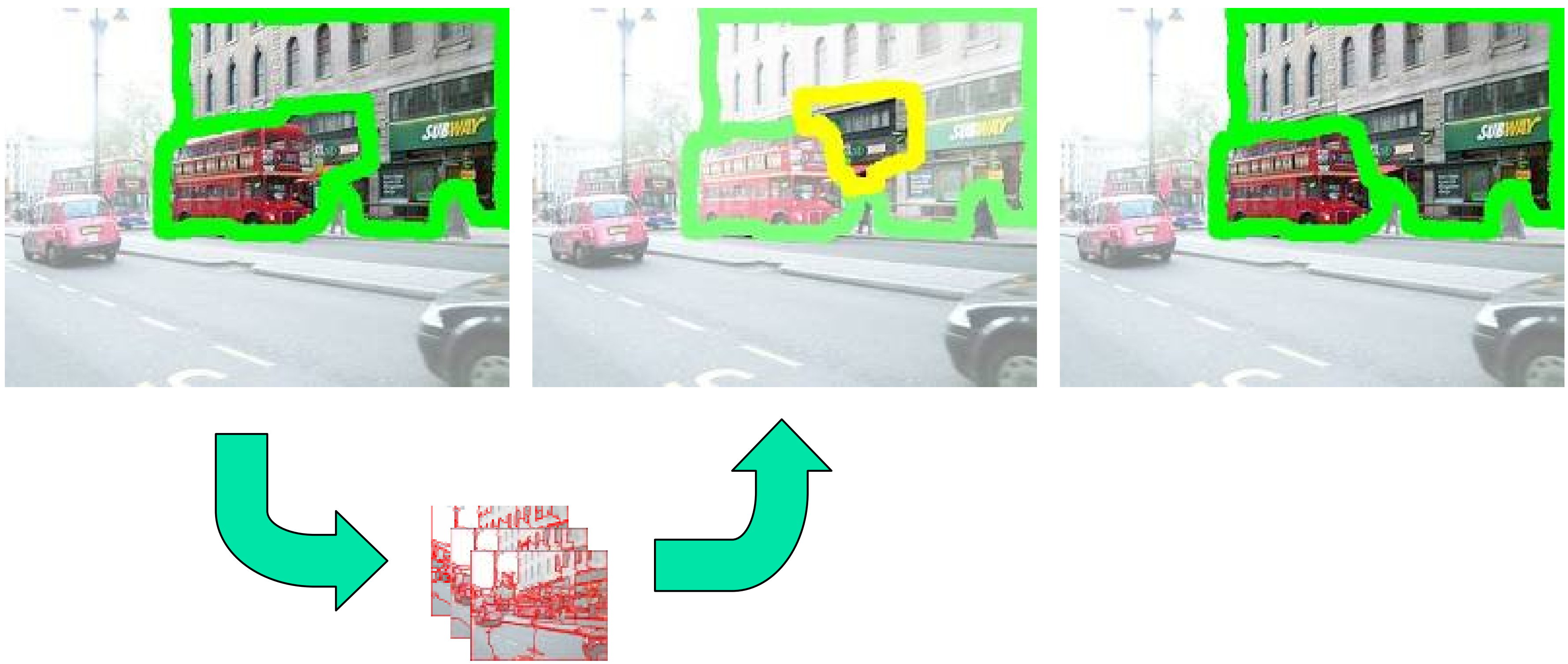
# Inference




Gould, et al.

# (Segment) Proposal Moves

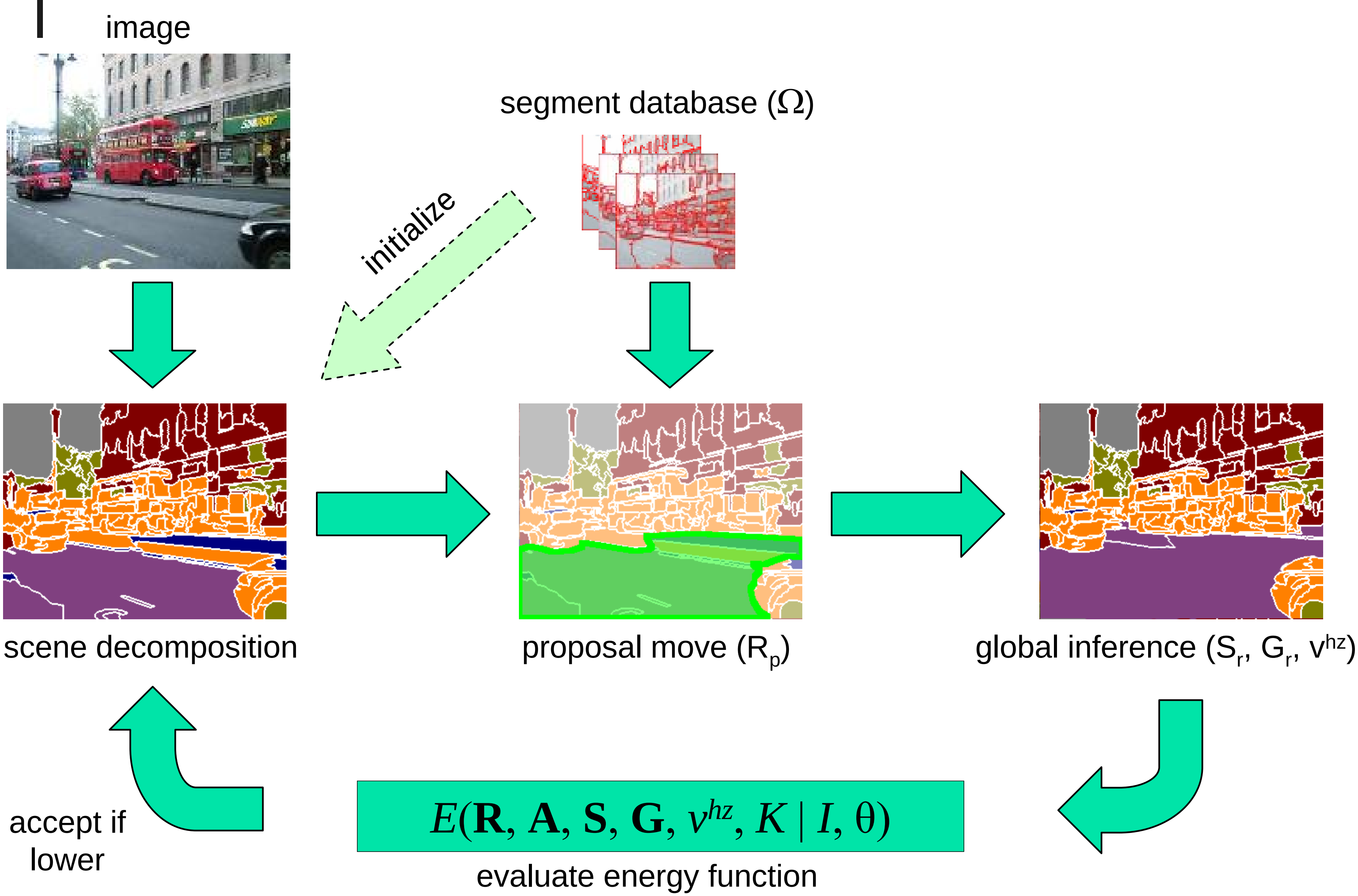
initial decomposition    proposal move    final decomposition



segment database ($\Omega$)

Gould, et al.

# Inference




Gould, et al.

# Inference Animation

image

semantic overlay


regions

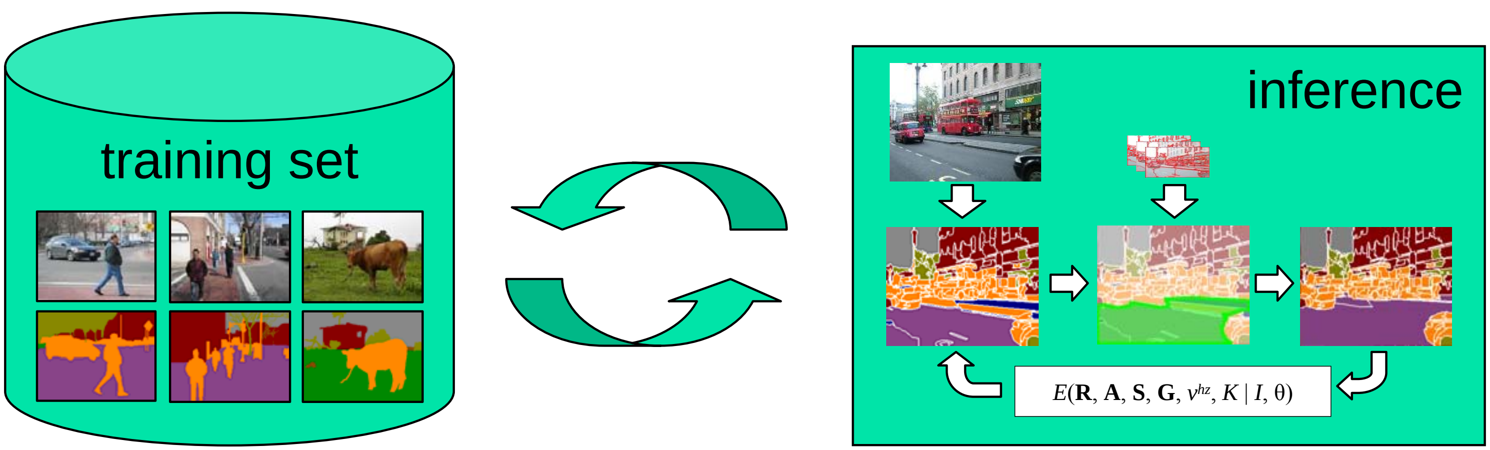
geometry overlay

sky tree road grass water bldg mntn fg obj.

Gould, et al.

# Parameter Learning

- Positive examples: all coherent regions and segments
- Negative examples: exponentially many
  - Most of them are ridiculously easy
- Closed-loop learning
  - Learn simple region and context models
  - Run inference (on training set) sampling errors
  - Re-train with augmented training set




Gould, et al.

# Results: 21-class MSRC

- Validate against state-of-the-art approaches
- Region/pixel class only
- Ground truth labels are approximate
- **No geometry** information

| **21 CLASS** | Mean |
|---|---|
| *Shotton et al.* | *72.2* |
| *Gould et al.* | ***76.5*** |
| Pixel CRF | 75.3 |
| Region-based | **76.4** |

hand labeled | image | pixel CRF | region-based

Gould, et al.

# High Quality Dataset

- MSRC dataset is limited
  - poorly labeled boundaries
  - many missing pixels (void)
  - no geometry information

- Collected images from MSRC, Hoiem et al., Pascal VOC

- 715 outdoor scenes with high-quality labels
  - region boundaries
  - region class and geometry
  - horizon

- Used Amazon's Mechanical Turk for labeling

- Available for download from: `http://www.stanford.edu/~sgould`

Gould, et al.

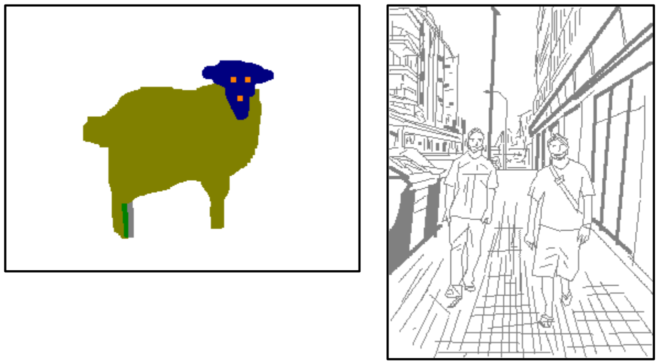
# Amazon Mechanical Turk (AMT)

- $0.10 per task (regions, classes, surface types)
- 5-10 minutes per task
- 24-48 hour turn-around time (for 715 images)
- Less than 10% of tasks needed rework

- **Total cost for labels:** under $250 (includes $40 textbook on Adobe Flash)
- **Saving me from having to label image:** priceless.


Gould, et al.

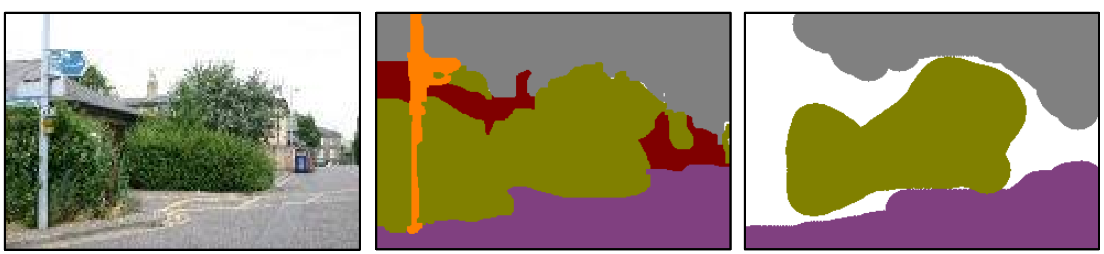
# AMT: Label Quality



You don’t always get
what you want

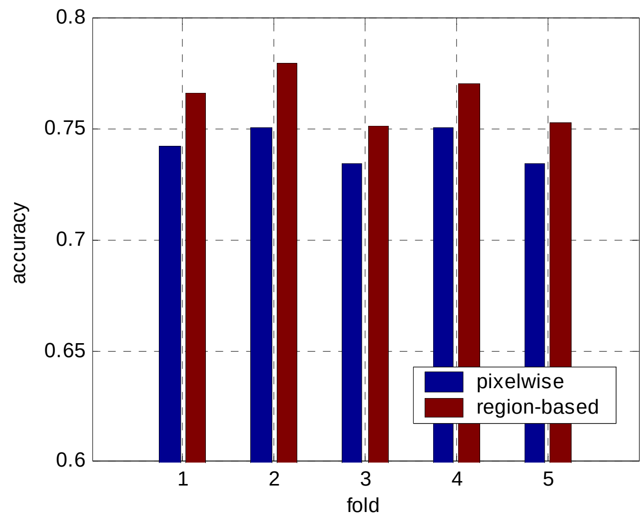
Typical quality (hand labeled)

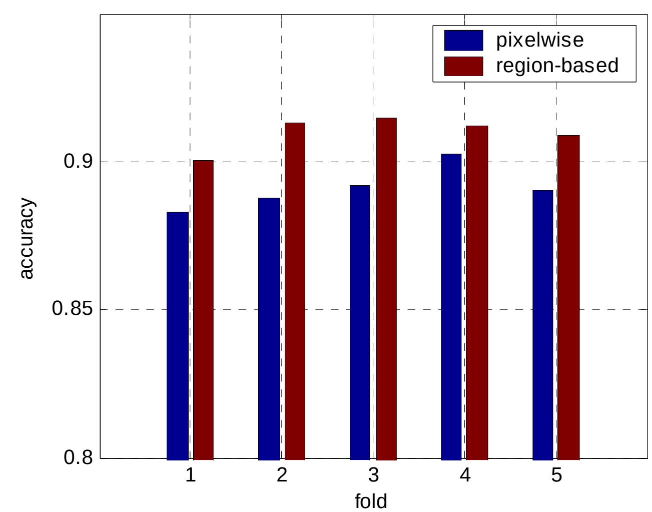
Comparison with MSRC labels



Gould, et al.

# Quantitative Results




| **CLASS** | Mean | Std |
|---|---|---|
| Pixel CRF | 74.3 | 0.80 |
| Region-based | **76.4** | 1.22 |

**Region Classes:** sky, tree, road, grass, water, building, mountain, fg. object

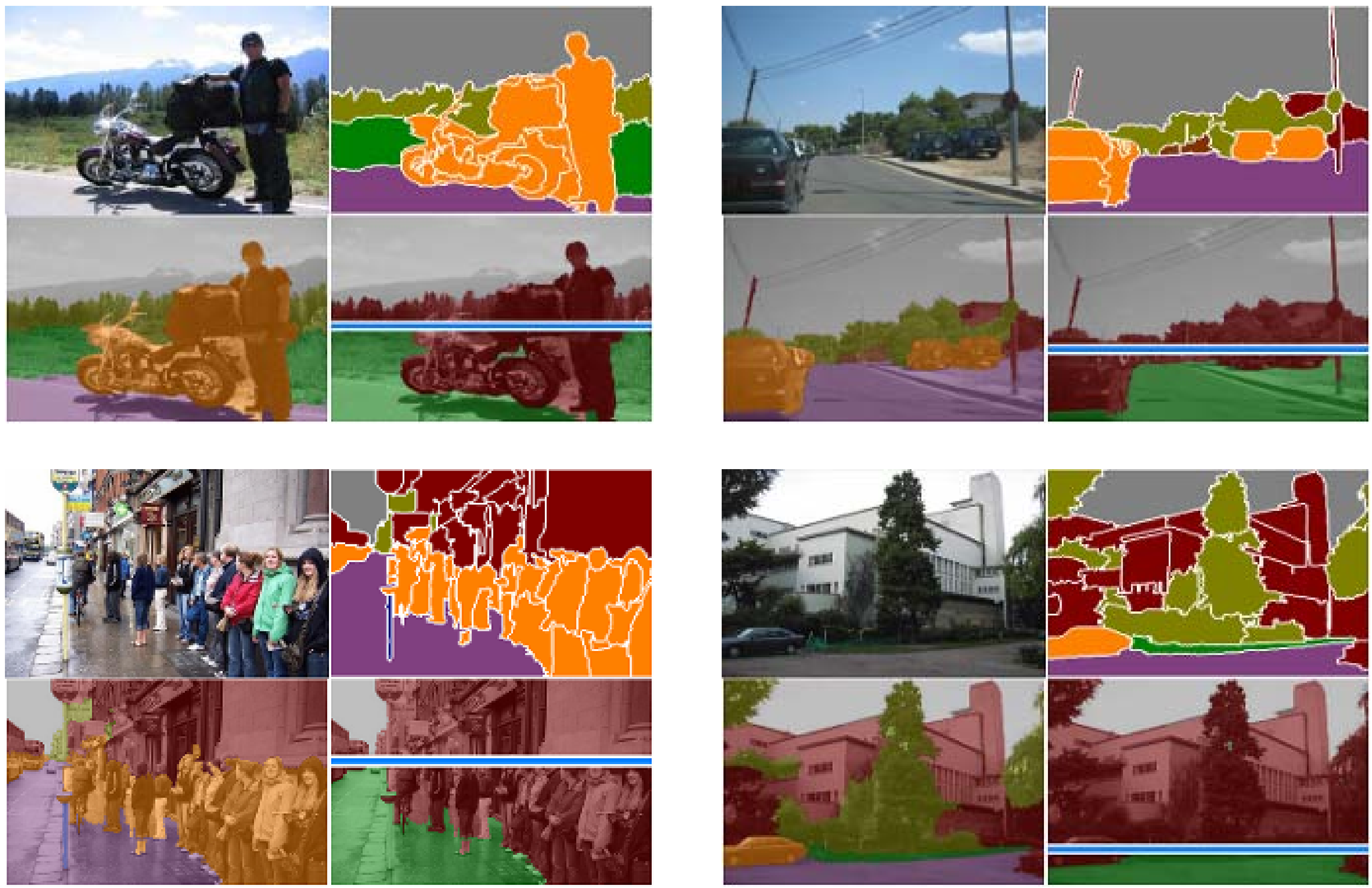
| **GEOMETRY** | Mean | Std |
|---|---|---|
| Pixel CRF | 89.1 | 0.73 |
| Region-based | **91.0** | 0.56 |

**Region Geometry:** sky, vertical, horizontal (support)

**Horizon error:** 6.9% (17 pixels)

Gould, et al.

# Example Results



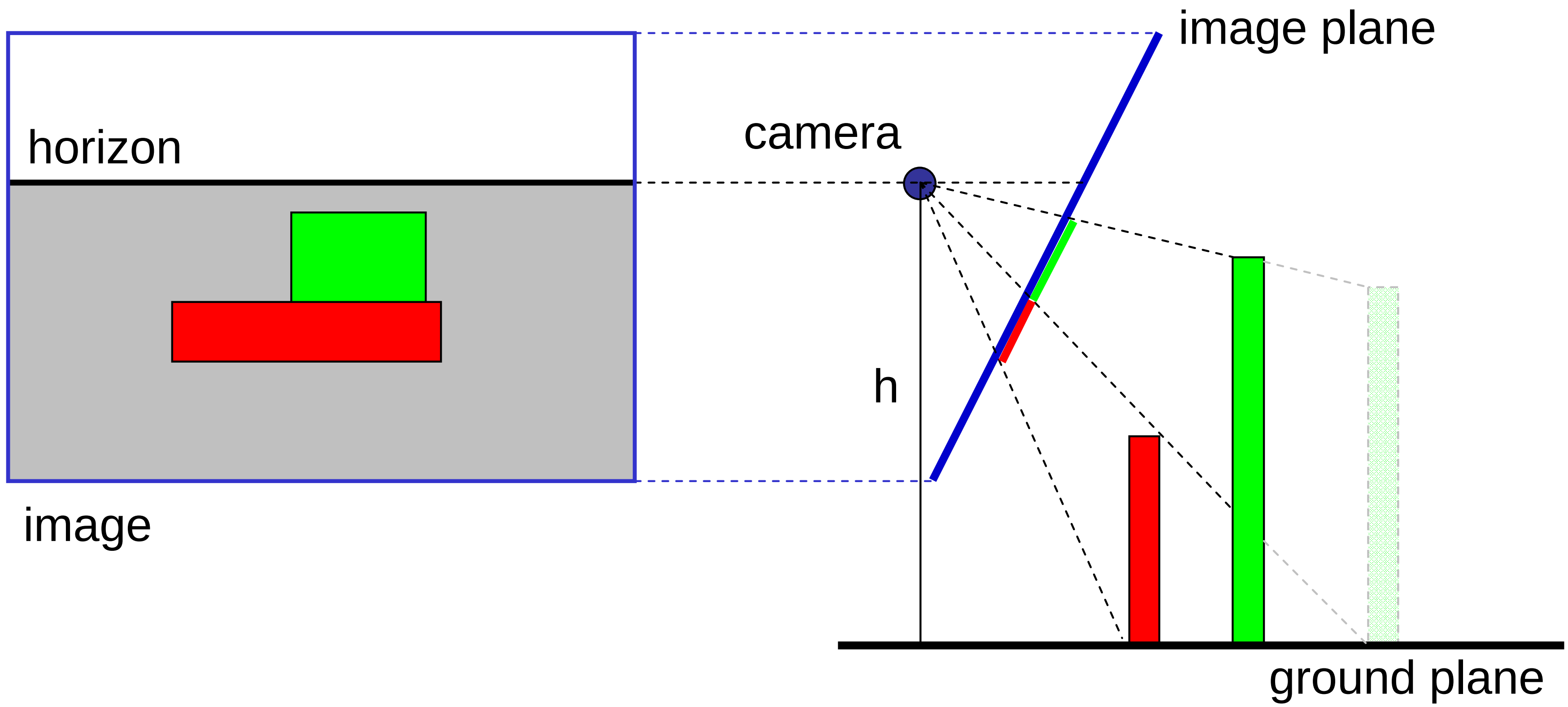
sky tree road grass water bldg mntn fg obj.

sky horz. vert.

Gould, et al.

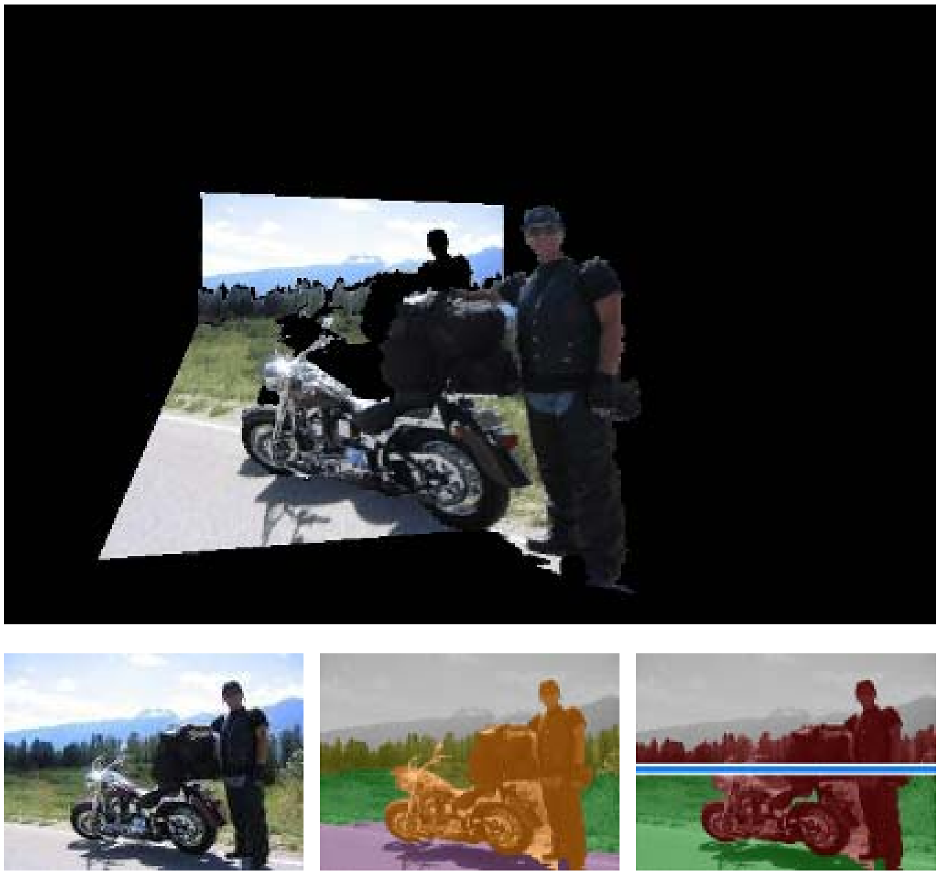
# Application: 3d Reconstruction

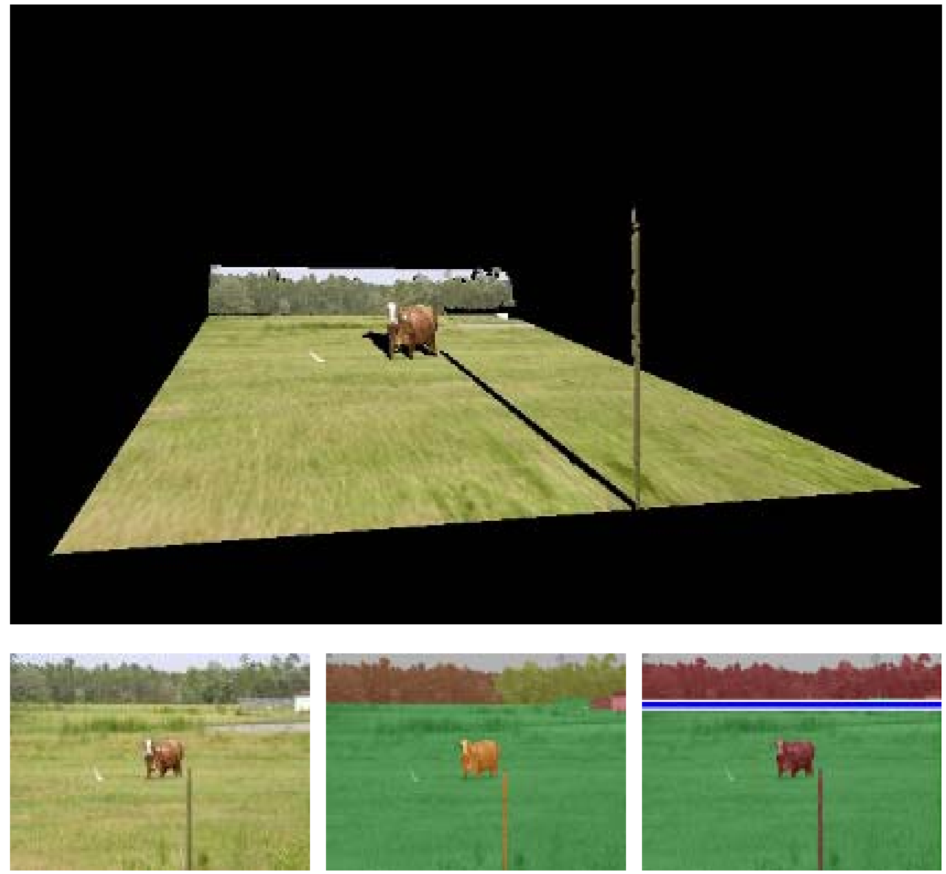
- Estimate camera tilt from location of horizon
- Predict region 3d position using ray projected through camera plane




Gould, et al.

# Example 3d Reconstructions

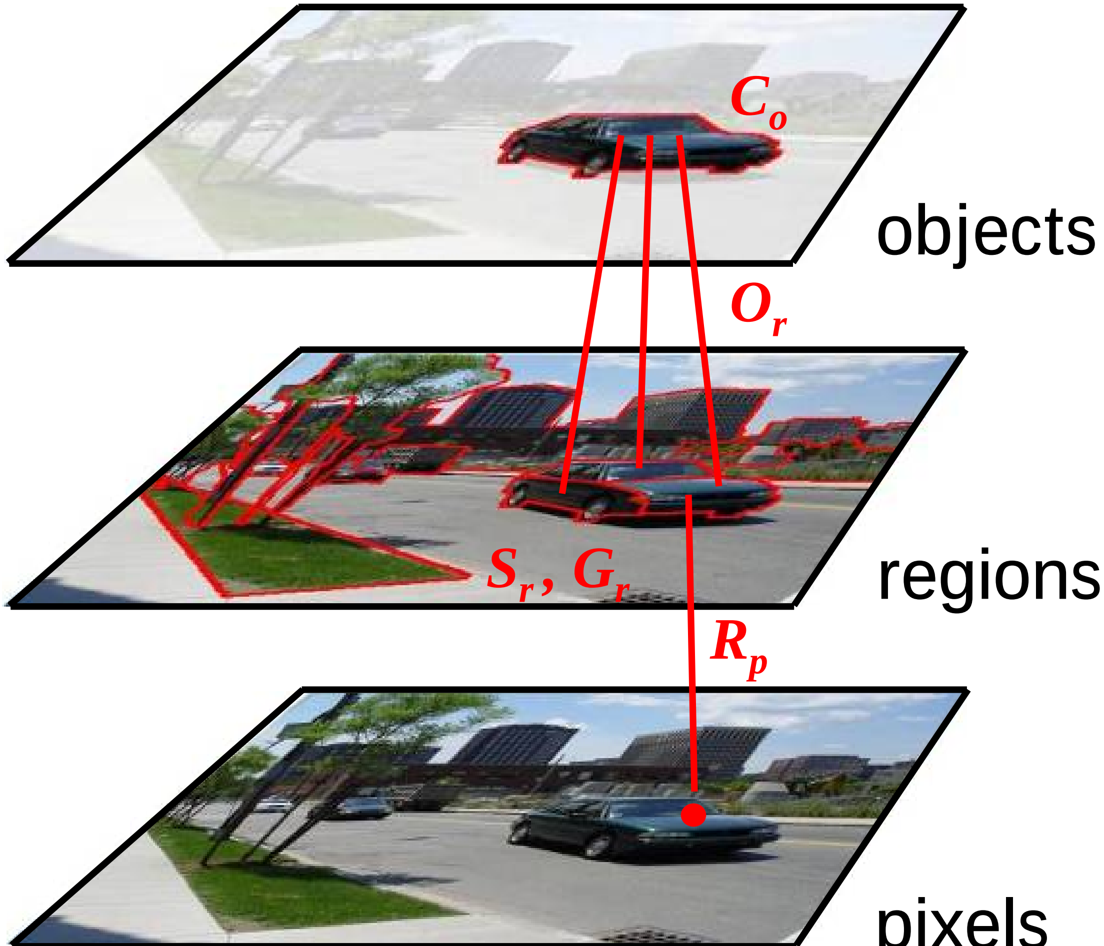




**Related work:** [Saxena et al., PAMI 08], [Hoiem et al., IJCV 07], [Russell and Torralba, CVPR 09]

# NIPS 2009 Sneak Peak

# Hierarchical Scene Model




**Variables**
$\alpha_p$: pixel appearance
$R_p$: pixel-to-region correspondence
$A_r$: region appearance
$S_r$: region semantic class
$G_r$: region geometry
$O_r$: region-to-object correspondence
$C_o$: object class
$v^{hz}$: location of horizon

energy function

$$E(\mathbf{R}, \mathbf{A}, \mathbf{S}, \mathbf{G}, \mathbf{O}, \mathbf{C}, v^{hz}, K)$$

[Gould, Gao, Koller NIPS 09]

# Energy Function

$$E(\mathbf{R}, \mathbf{A}, \mathbf{S}, \mathbf{G}, \mathbf{O}, \mathbf{C}, v^{hz}, K \mid I, \theta)$$

$$=$$

$\psi^{\text{horizon}}(\mathrm{v}^{\text{hz}})$

**Horizon Term**
e.g., vanishing lines

$+$

$\psi^{\text{region}}(\mathrm{S}_r, \mathrm{G}_r, \mathrm{v}^{\text{hz}})$

**Region Term**
e.g., consistent appearance and location

$+$

$\psi^{\text{boundary}}(\mathrm{A}_r, \mathrm{A}_s)$

**Boundary Term**
e.g., difference in color/texture between regions

$+$

$\psi^{\text{object}}(\mathrm{C}_o, \mathrm{v}^{\text{hz}})$

**Object Term**
e.g. wheel-like appearance in bottom corner

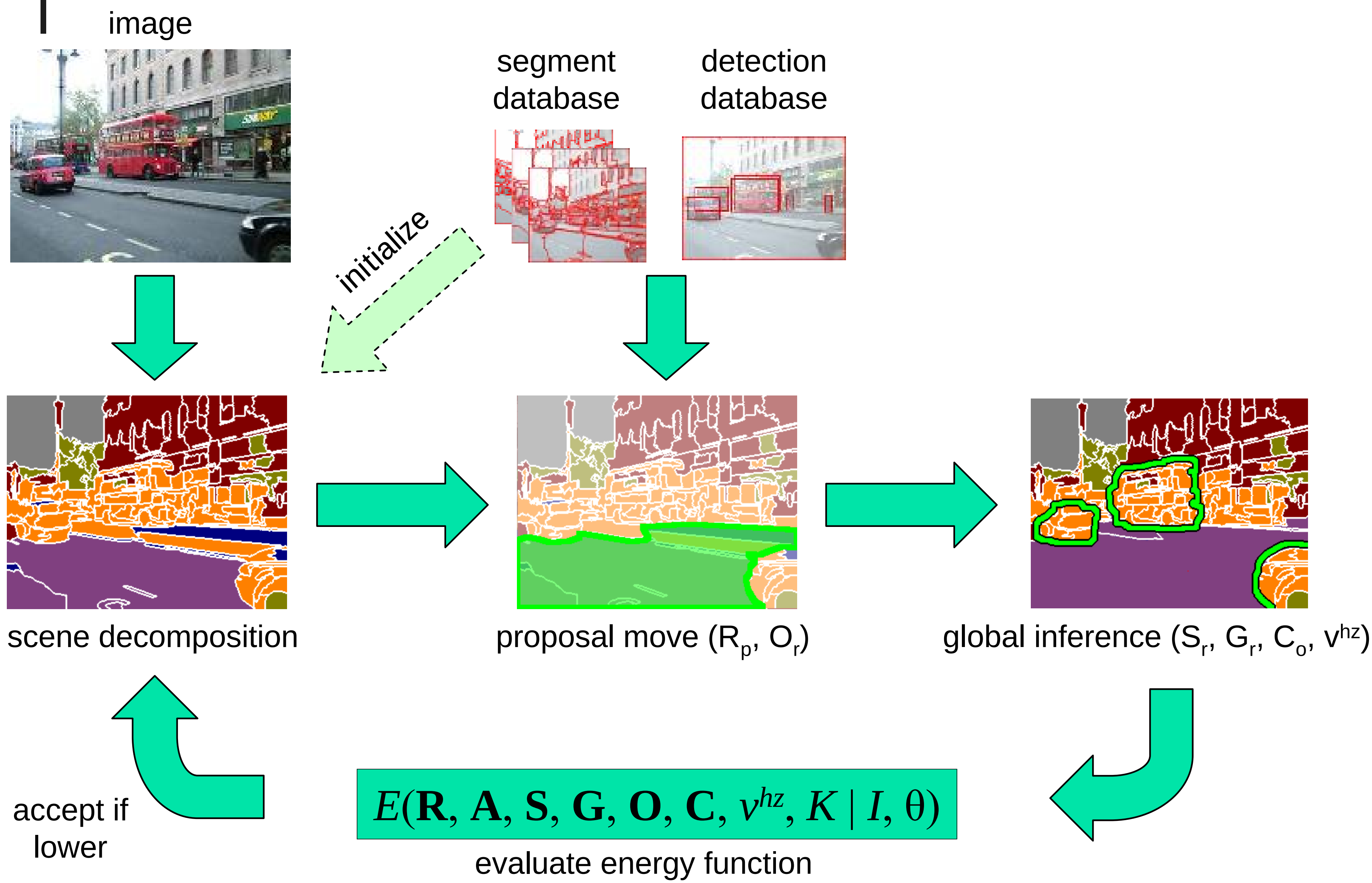
$+$

$\psi^{\text{context}}(\mathrm{C}_o, \mathrm{S}_r)$

**Context Term**
e.g., cars on road

Gould, et al.

# Top-down Proposal Moves

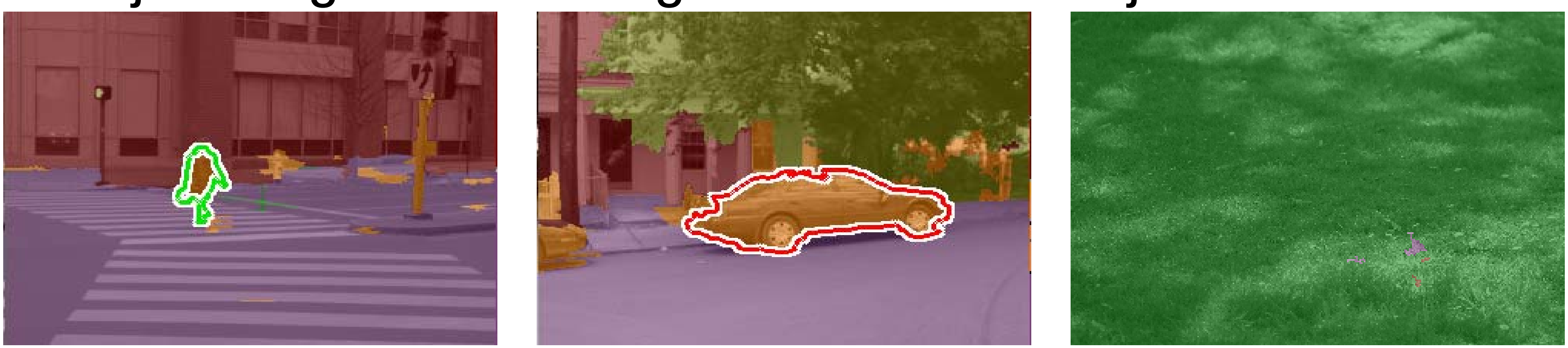


[Gould, Gao, Koller NIPS 09]

# Object Detection Results

Sliding-window detector's top two detections per image

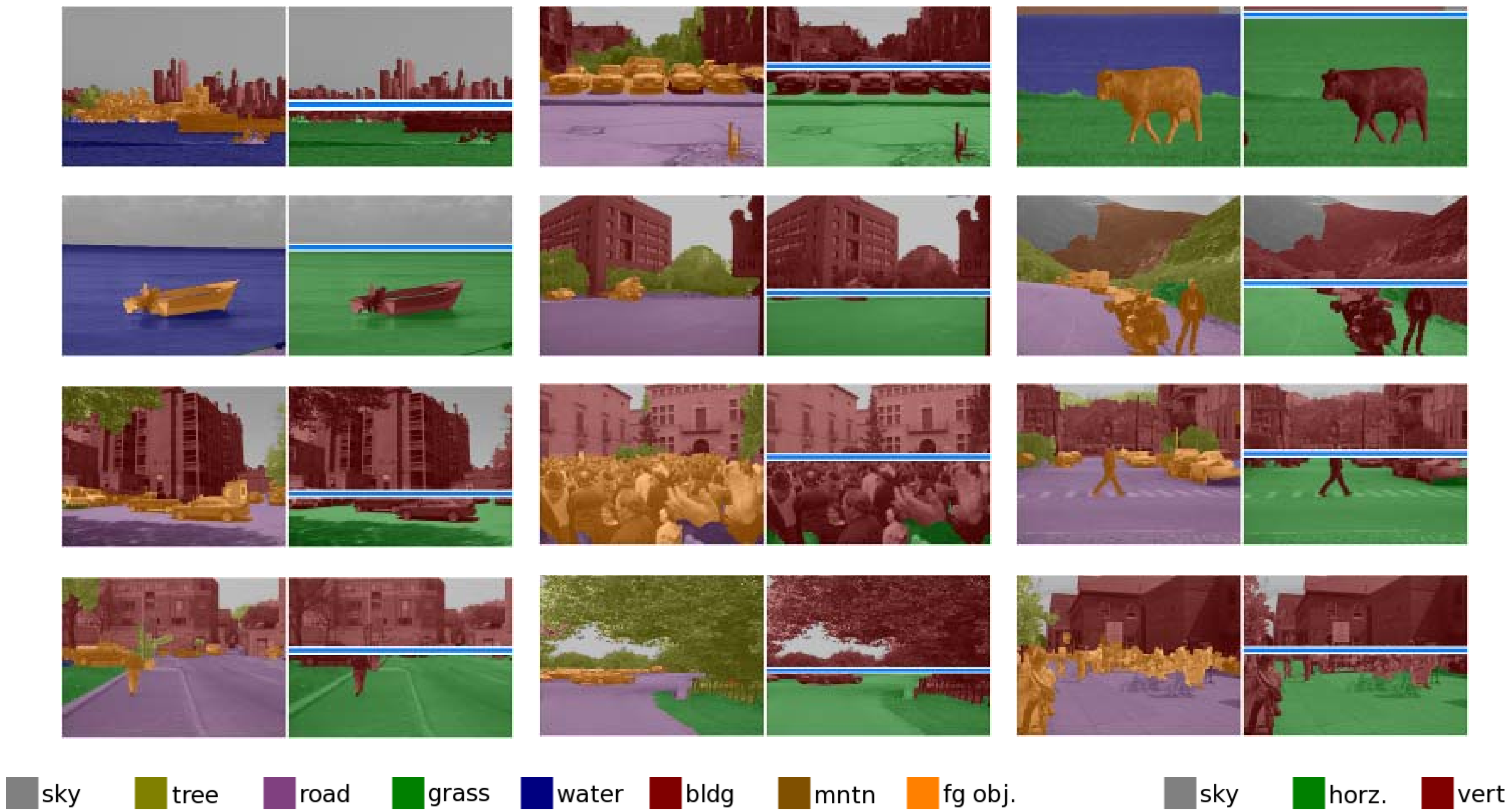

Our joint region-based segmentation and object detection



# Summary

- Our model decomposes a scene into geometric and semantically consistent regions using a **unified energy function** over pixels and regions
- By classifying large regions rather than individual pixels we can compute more **robust features** and reduce inference complexity
- **Multiple over-segmentations** allow us to refine region boundaries and make large moves in energy space
- **Context** can be easily captured using a pairwise term between adjacent regions
- **Our model provides a base for integrating many other vision tasks (e.g., 3D reconstruction and object detection)**

# Thank You　ありがとうございます。




Gould, et al.