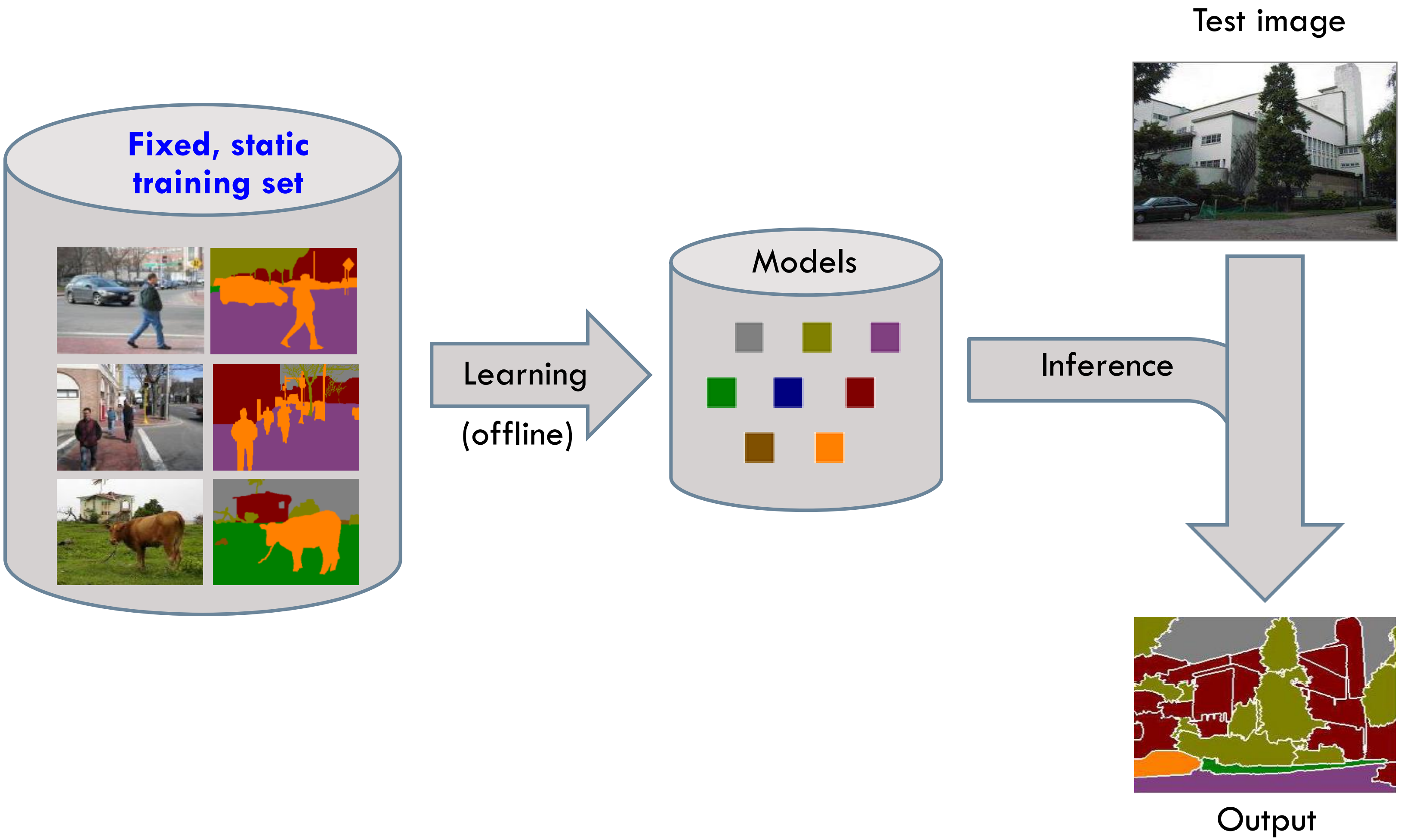
Recognition Methods for Open-Universe Datasets

Lana Lazebnik

Closed-universe recognition

Fixed, pre-defined set of classes

■ sky 🟨 tree 🟪 road 🟩 grass 🟦 water 🟥 bldg 🟫 mntn 🟧 fg obj.



Test image

Fixed, static training set

Learning

(offline)

Models

Inference

Output

Closed-universe datasets

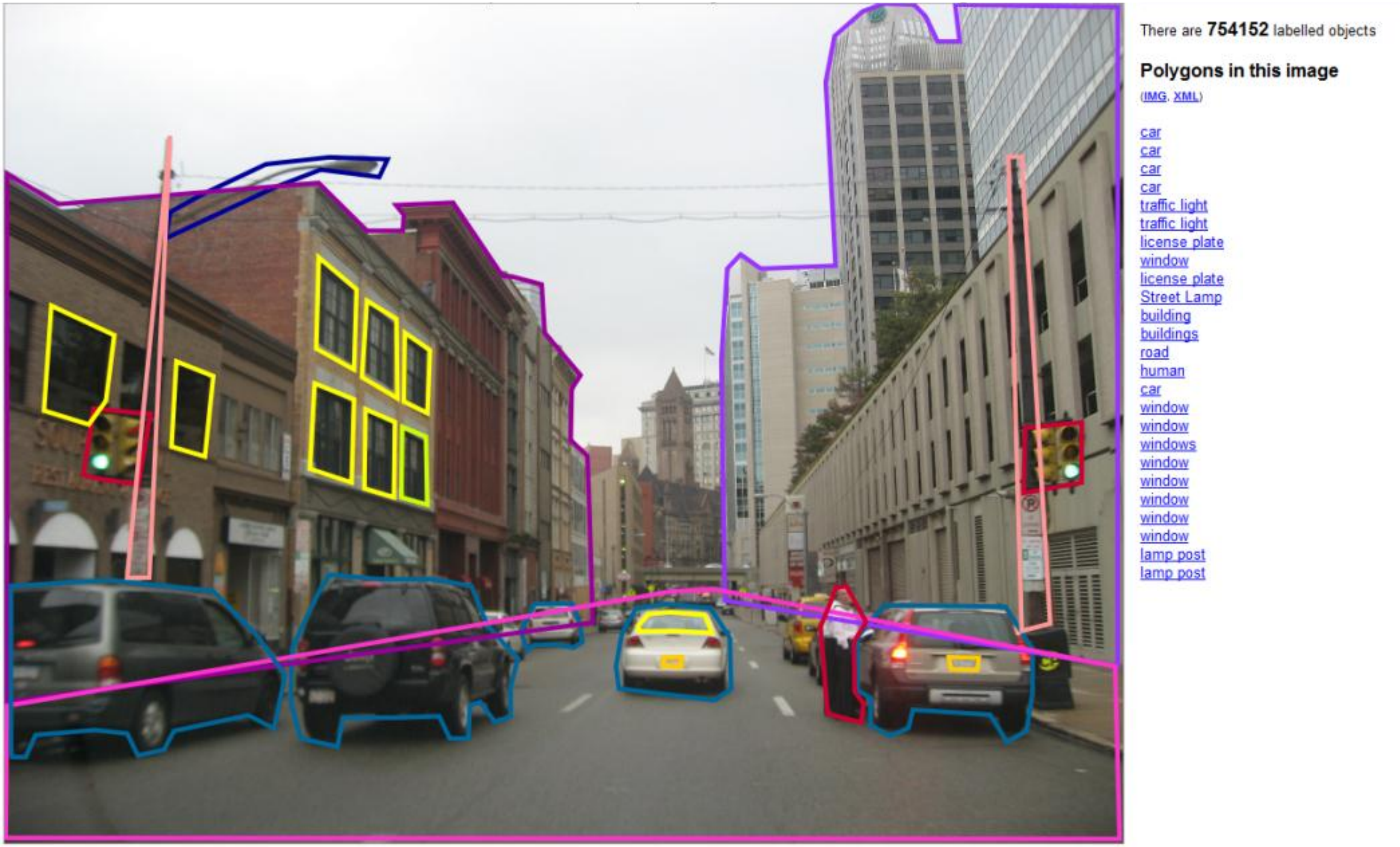
- Small amount of data
- Static datasets
- Limited variation
- Full annotation

Open-universe datasets

Google Image Search bing flickr facebook YouTube

- Large amount of data
- Evolving datasets
- Wide variation
- Incomplete annotation

Open-universe recognition



There are **754152** labelled objects

Polygons in this image

(IMG, XML)

car
car
car
car
traffic light
traffic light
license plate
window
license plate
Street Lamp
building
buildings
road
human
car
window
window
windows
window
window
window
window
window
lamp post
lamp post

Evolving training set

http://labelme.csail.mit.edu/

Open-universe recognition

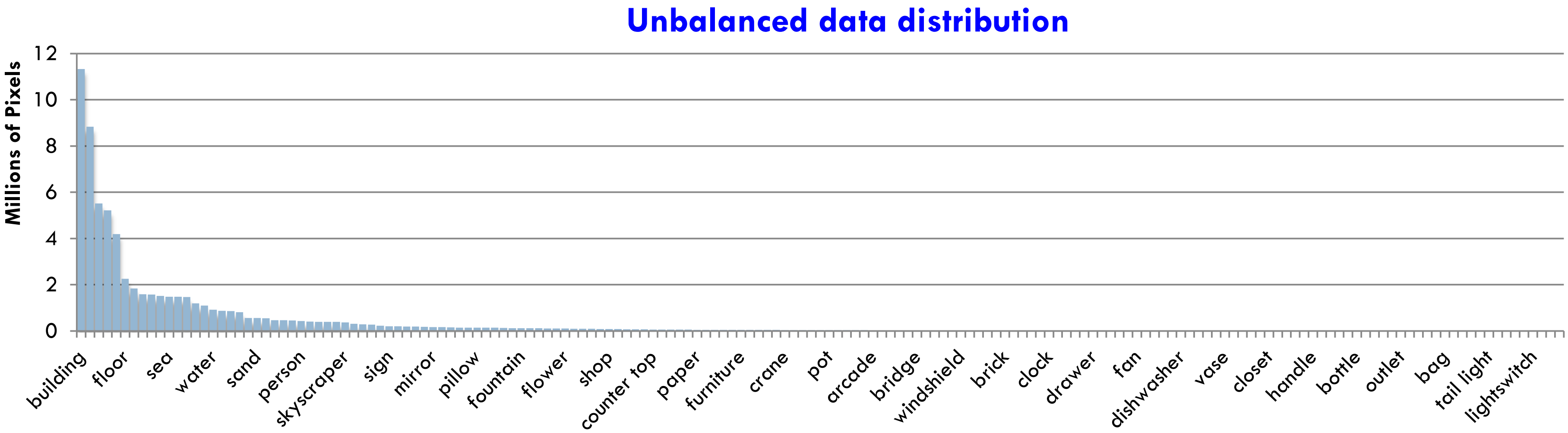
Very large/open-ended set of classes

building mountain tree sky road sea car field window plant river grass rock sidewalk sand door desert bridge person balcony fence staircase awning sign crosswalk streetlight boat bus pole sun cow bird moon • • • wall plate table chair ground desk bookshelf utensil glass screen floor bed placem

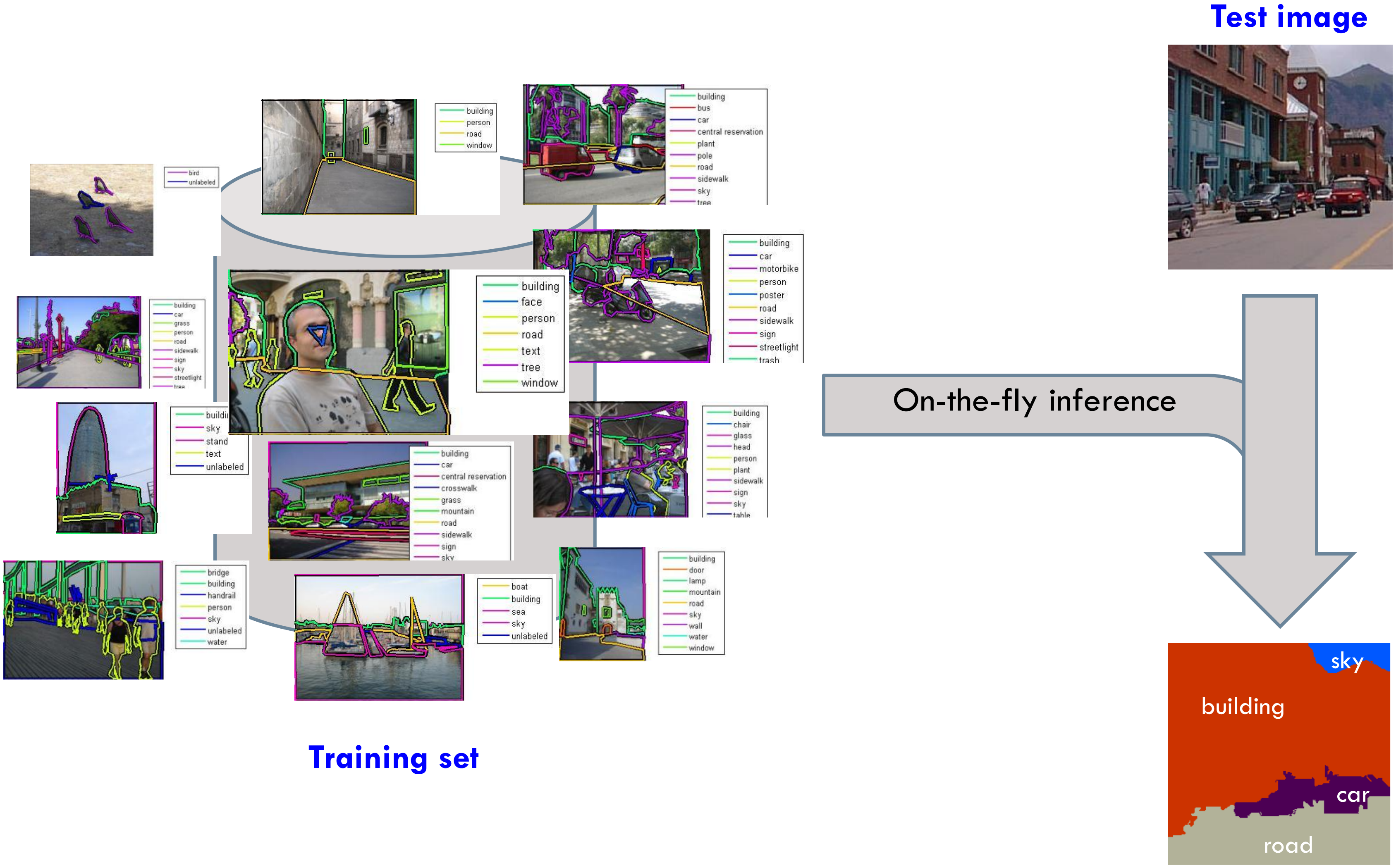
Open-universe recognition

Very large/open-ended set of classes

building mountain tree sky road sea car field window plant river grass rock sidewalk sand door desert bridge person balcony fence staircase awning sign crosswalk streetlight boat bus pole sun cow bird moon • • • wall plate table chair ground desk bookshelf utensil glass screen floor bed place


Unbalanced data distribution

Potential solution: Lazy learning

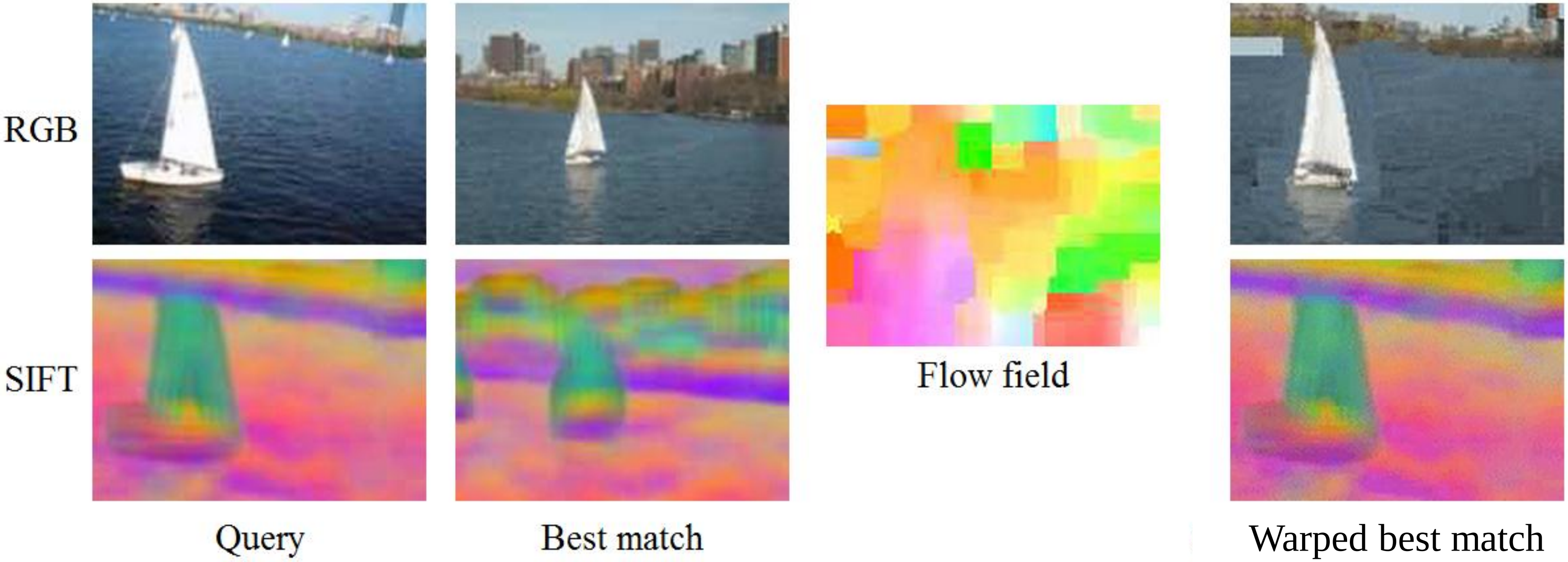


Test image

On-the-fly inference

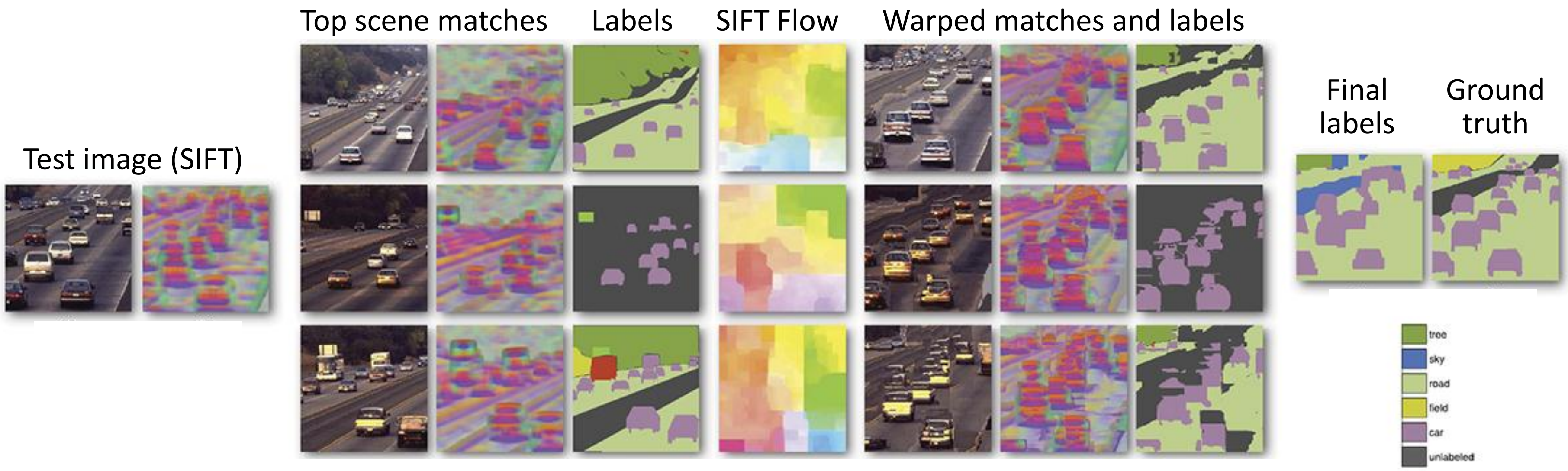
Training set

Dense Scene Alignment by SIFT Flow



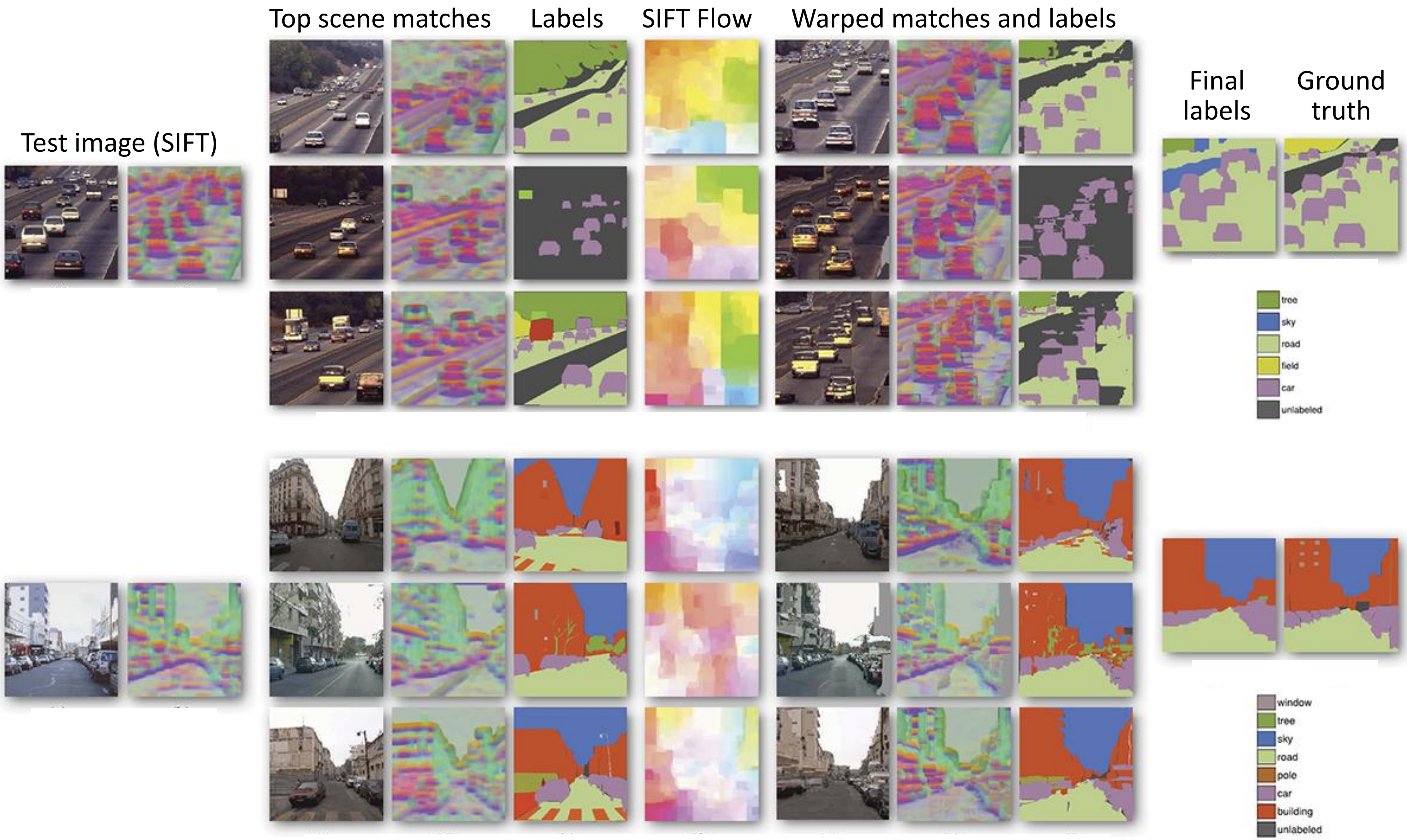
C. Liu, J. Yuen, and A. Torralba, “Nonparametric Scene Parsing via Label Transfer,” PAMI 2011

Dense Scene Alignment by SIFT Flow



C. Liu, J. Yuen, and A. Torralba, “Nonparametric Scene Parsing via Label Transfer,” PAMI 2011

Dense Scene Alignment by SIFT Flow



C. Liu, J. Yuen, and A. Torralba, “Nonparametric Scene Parsing via Label Transfer,” PAMI 2011

SIFT Flow: Pros and cons

- Advantages
  - Nonparametric method, can work with any number of labels, evolving training set
  - Initial global scene matching step improves efficiency, provides scene-level context
- Disadvantages
  - Computing SIFT flow is very computationally intensive
  - Warping model is not necessarily the most natural one for describing object-level correspondence

LARGE-SCALE NONPARAMETRIC IMAGE PARSING

Joseph Tighe and Svetlana Lazebnik

ECCV 2010, new stuff in preparation

Step 1: Scene-level matching

Color Histogram

Spatial Pyramid
(Lazebnik et al., 2006)

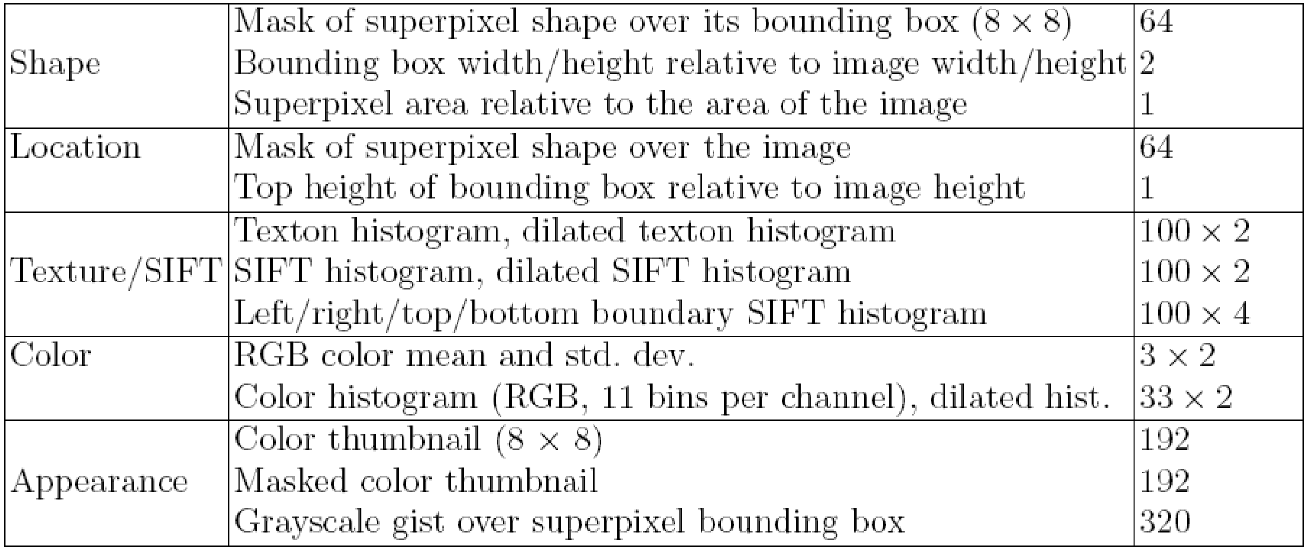
Gist
(Oliva & Torralba, 2001)

Step 2: Region-level matching

Superpixels

(Felzenszwalb & Huttenlocher, 2004)

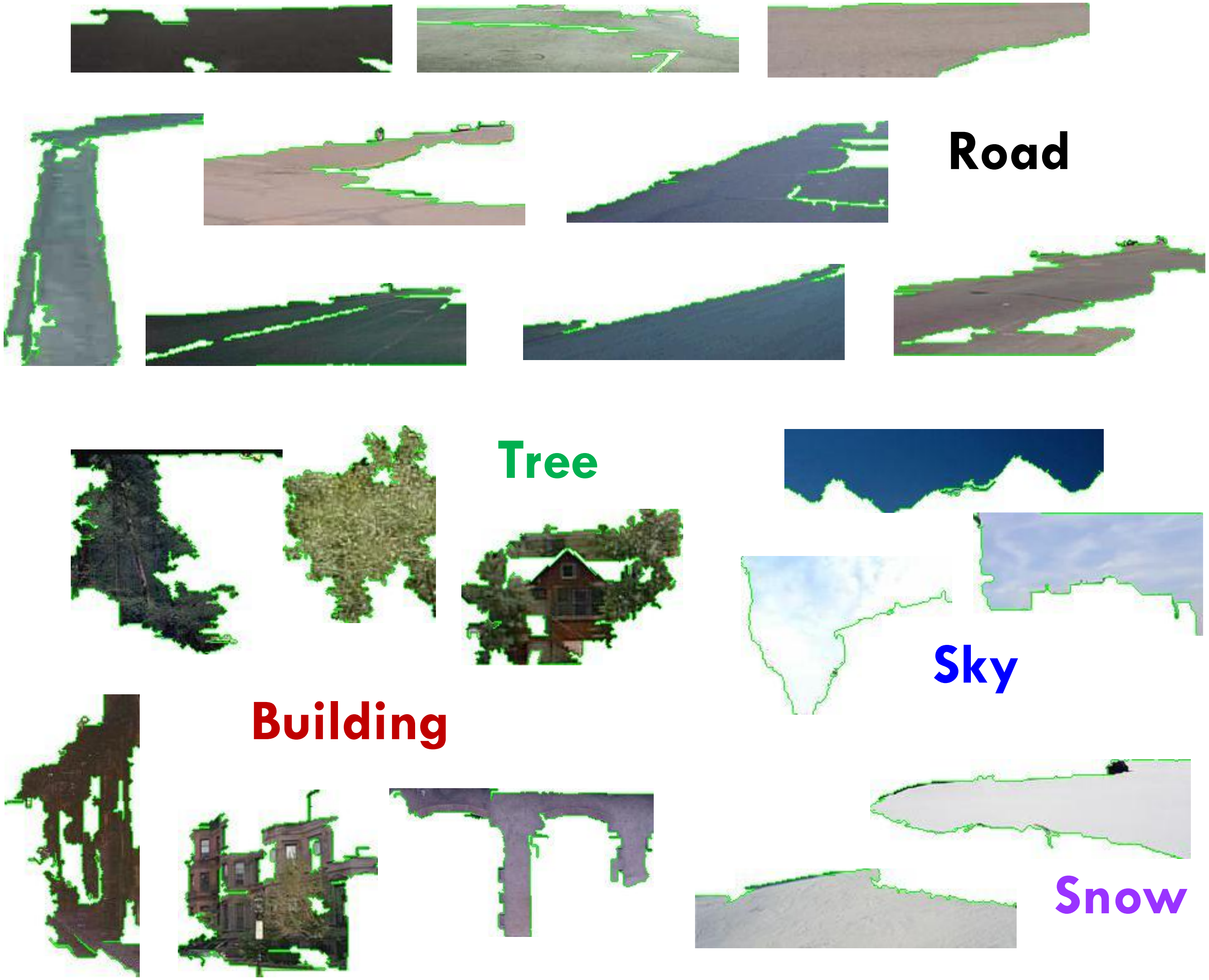
Superpixel features



Shape	Mask of superpixel shape over its bounding box (8×8)	64
	Bounding box width/height relative to image width/height	2
	Superpixel area relative to the area of the image	1
Location	Mask of superpixel shape over the image	64
	Top height of bounding box relative to image height	1
Texture/SIFT	Texton histogram, dilated texton histogram	100×2
	SIFT histogram, dilated SIFT histogram	100×2
	Left/right/top/bottom boundary SIFT histogram	100×4
Color	RGB color mean and std. dev.	3×2
	Color histogram (RGB, 11 bins per channel), dilated hist.	33×2
Appearance	Color thumbnail (8×8)	192
	Masked color thumbnail	192
	Grayscale gist over superpixel bounding box	320

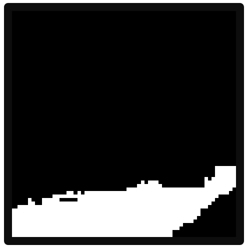
Step 2: Region-level matching

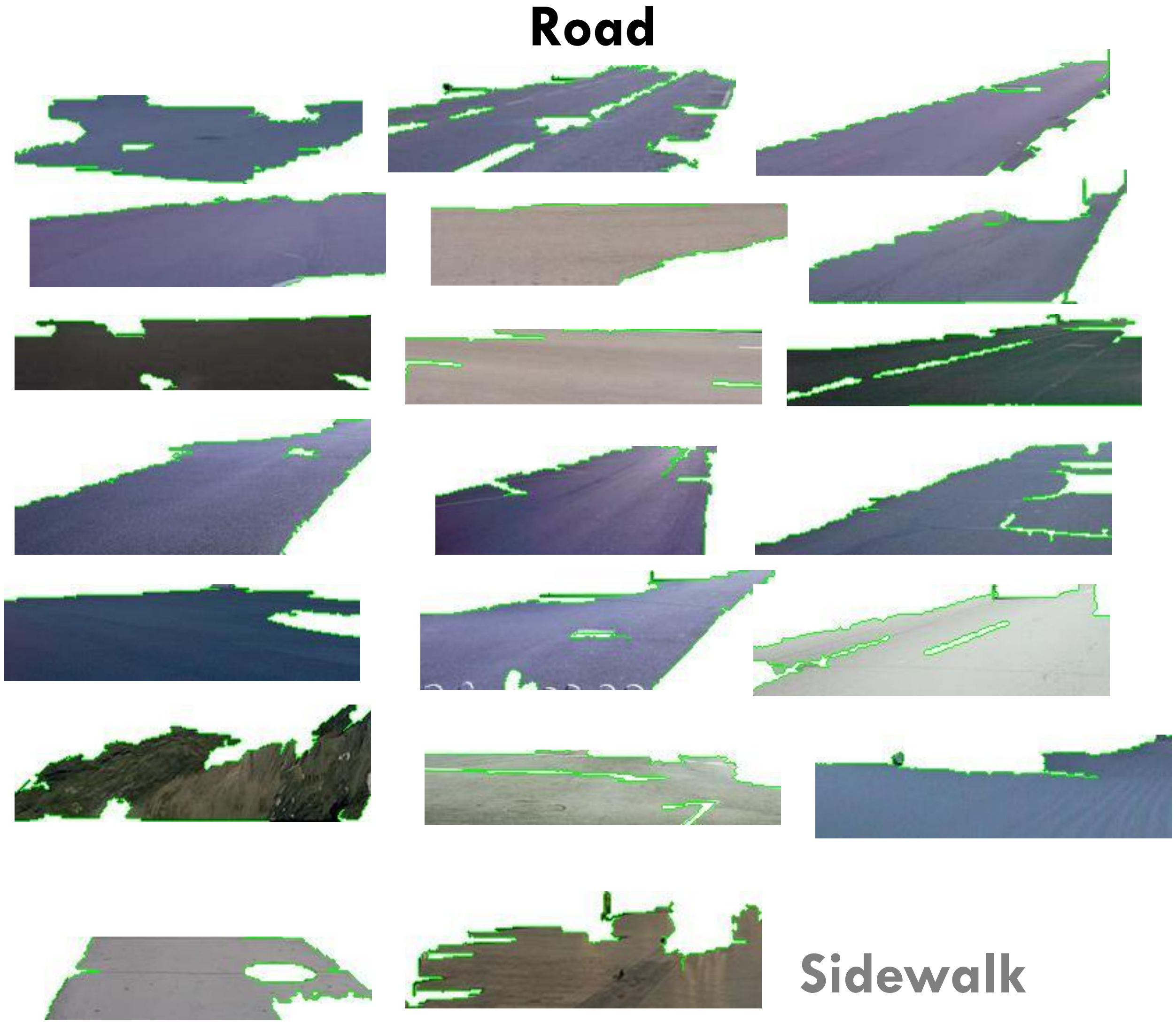
Pixel Area (size)



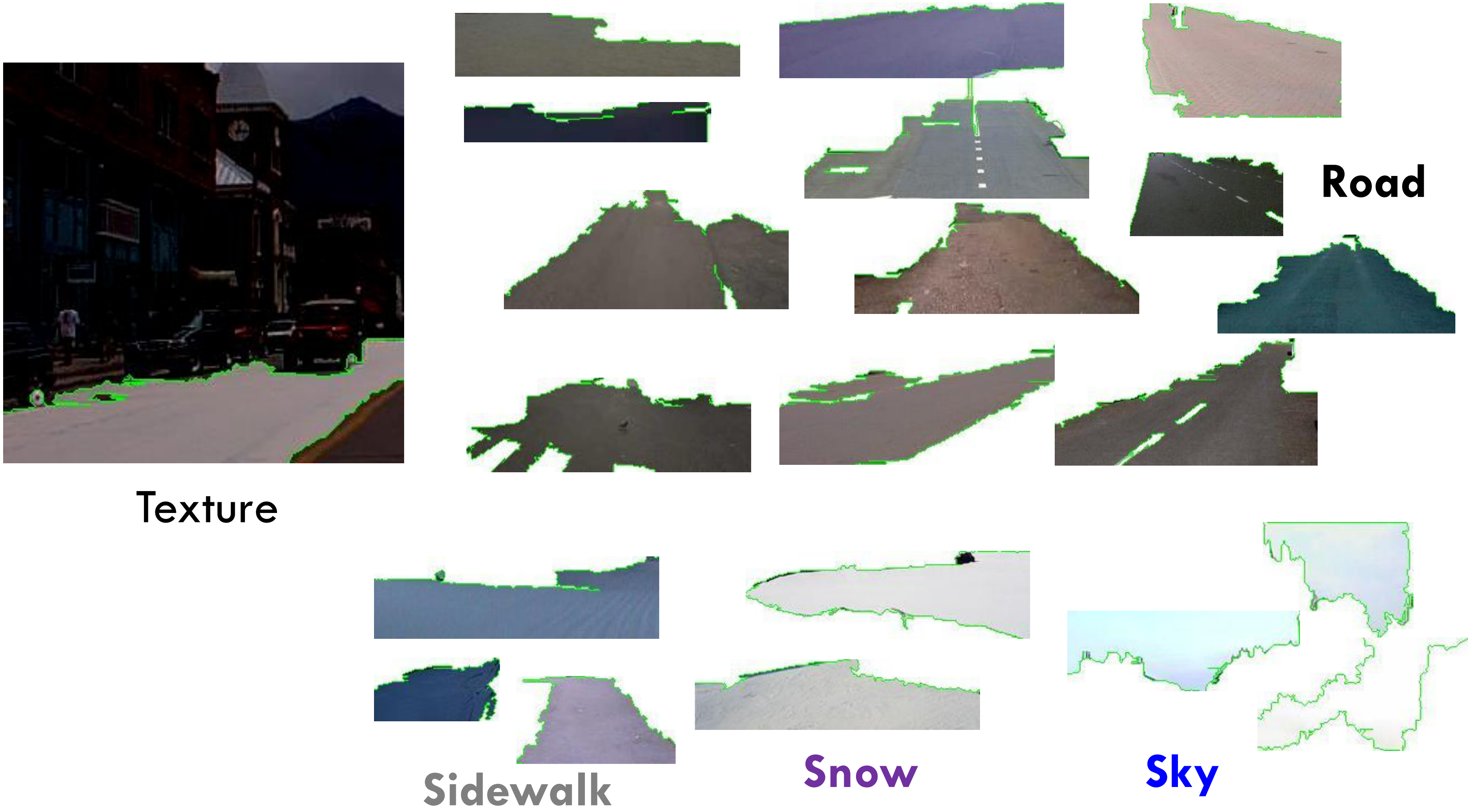
Step 2: Region-level matching

Absolute mask
(location)



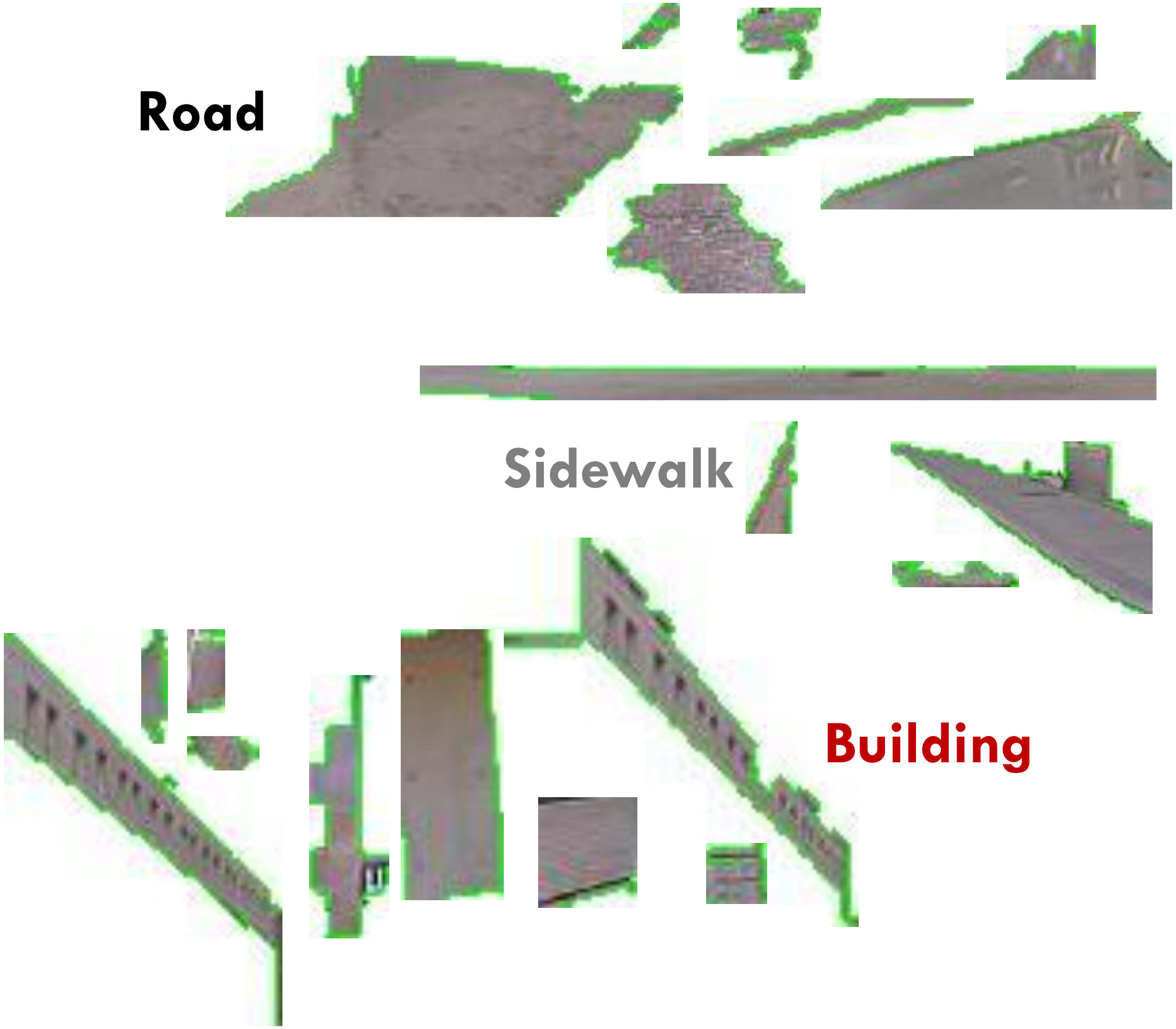


Step 2: Region-level matching



Step 2: Region-level matching

Color histogram



Region-level likelihoods

- Nonparametric estimate of class-conditional densities for each class c and feature type *k*:

$$\hat{P}(f_k(r_i)\,|\,c) = \frac{\#(N(f_k(r_i)),c)}{\#(D,c)}$$

Features of class c within some radius of r_i

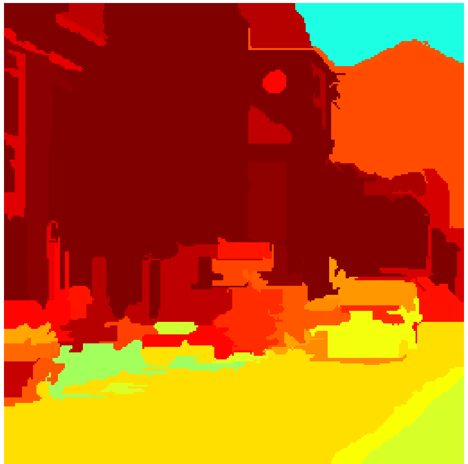
Total features of class c in the dataset

kth feature type of ith region

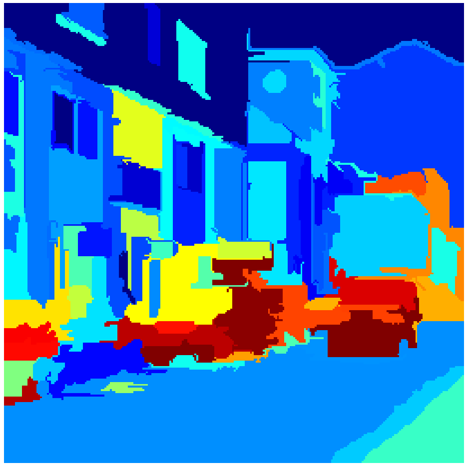
- Per-feature likelihoods combined via Naïve Bayes:

$$\hat{P}(r_i\,|\,c) = \prod_{\text{features } k} \hat{P}(f_k(r_i)\,|\,c)$$

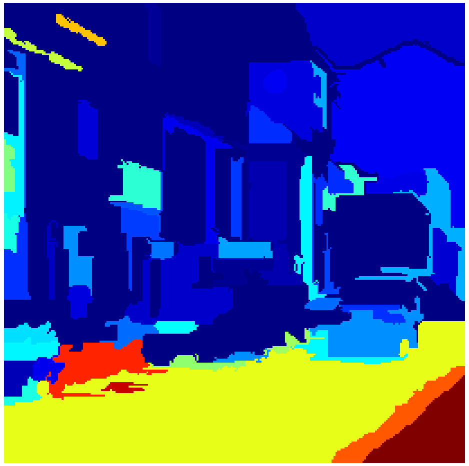
Region-level likelihoods



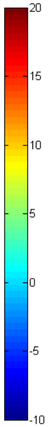
Building

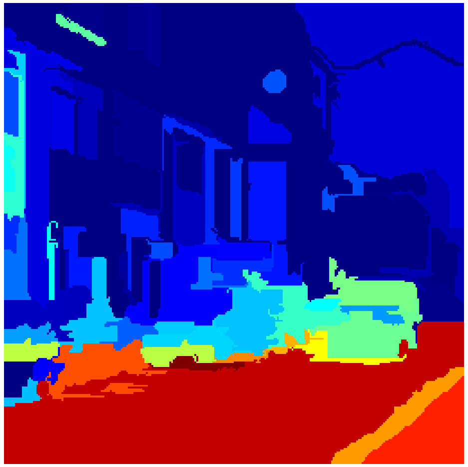


Car

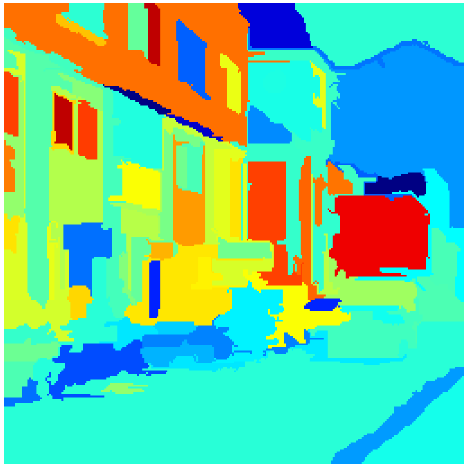


Crosswalk





Road



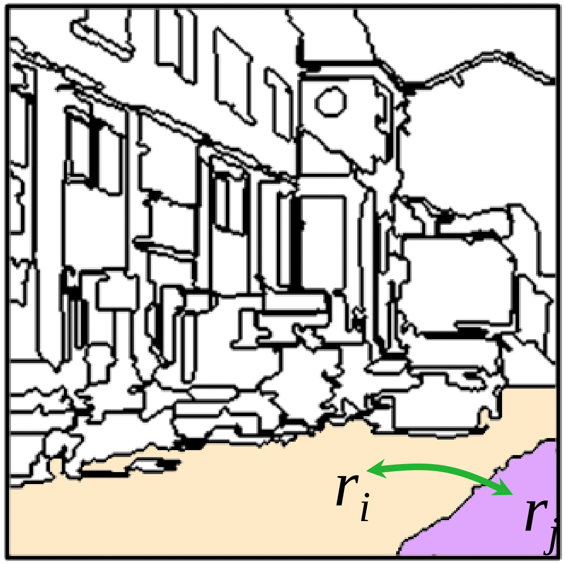
Window

Sky

Step 3: Global image labeling

- Compute a global image labeling by optimizing a Markov random field (MRF) energy function:

$$\underset{\substack{\uparrow \\ \text{Vector of region labels}}}{E(\mathbf{c})} = \underset{\text{Regions}}{\sum_i} -\log \underbrace{L(r_i, c_i)}_{\text{Likelihood score for region } r_i \text{ and label } c_i} + \lambda \underset{\text{Neighboring regions}}{\sum_{i,j}} \underbrace{\delta[c_i \neq c_j]}_{\text{Smoothing penalty}} \underbrace{\varphi(c_i, c_j)}_{\text{Co-occurrence penalty}}$$



Efficient approximate minimization using α-expansion (Boykov et al., 2002)

Step 3: Global image labeling

- Compute a global image labeling by optimizing a Markov random field (MRF) energy function:

$$E(\underset{\substack{\uparrow \\ \text{Vector of} \\ \text{region} \\ \text{labels}}}{\mathbf{c}}) = \underset{\text{Regions}}{\sum_{i}} -\log \underbrace{L(r_i, c_i)}_{\substack{\text{Likelihood score for} \\ \text{region } r_i \text{ and label } c_i}} \quad + \quad \lambda \underset{\substack{\text{Neighboring} \\ \text{regions}}}{\sum_{i,j}} \underbrace{\delta[c_i \neq c_j]}_{\substack{\text{Smoothing} \\ \text{penalty}}} \, \underbrace{\varphi(c_i, c_j)}_{\substack{\text{Co-occurrence} \\ \text{penalty}}}$$

Step 3: Global image labeling

- Compute a global image labeling by optimizing a Markov random field (MRF) energy function:

$$E(\mathbf{c}) = \sum_{i} -\log \underbrace{L(r_i, c_i)}_{\text{Likelihood score for region } r_i \text{ and label } c_i} \quad + \quad \lambda \sum_{i,j} \underbrace{\delta[c_i \neq c_j]}_{\text{Smoothing penalty}} \underbrace{\varphi(c_i, c_j)}_{\text{Co-occurrence penalty}}$$

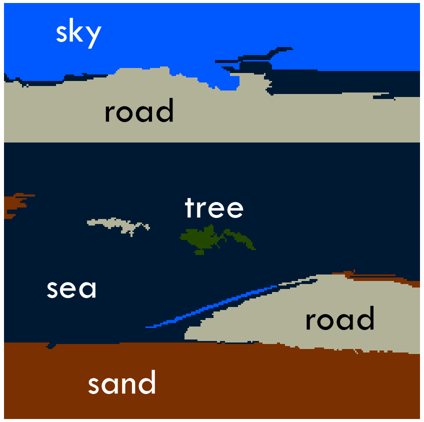
Vector of region labels

Regions

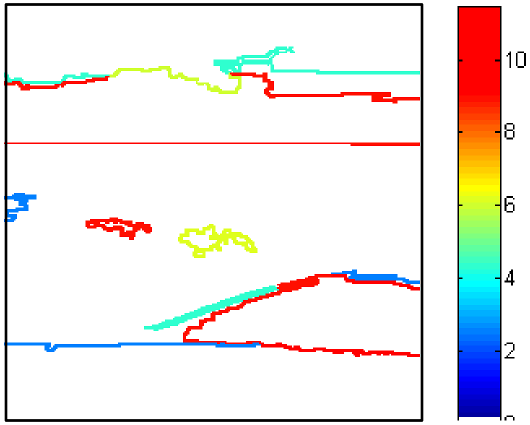
Neighboring regions

Original image

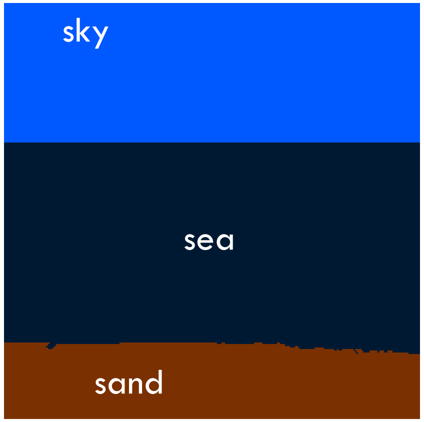
Maximum likelihood labeling



Edge penalties

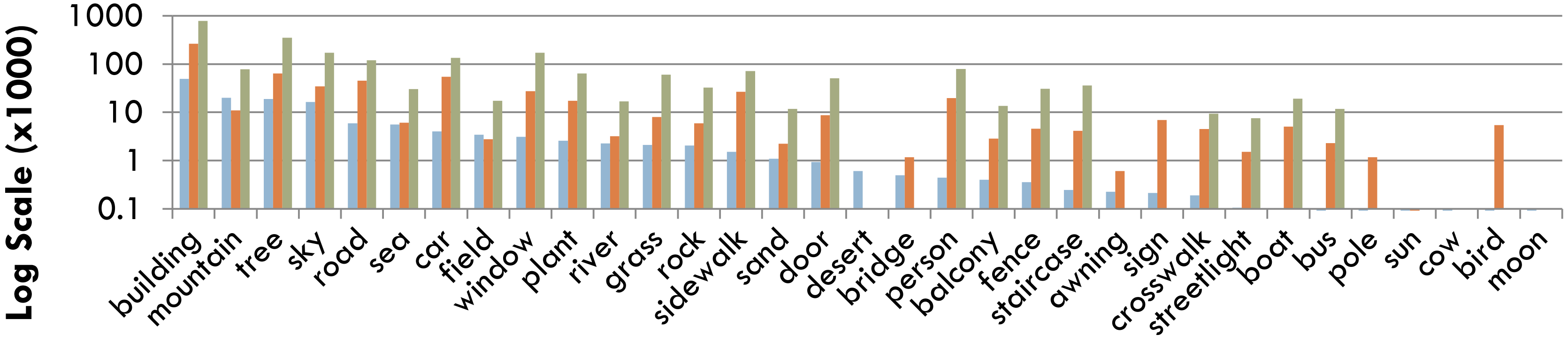


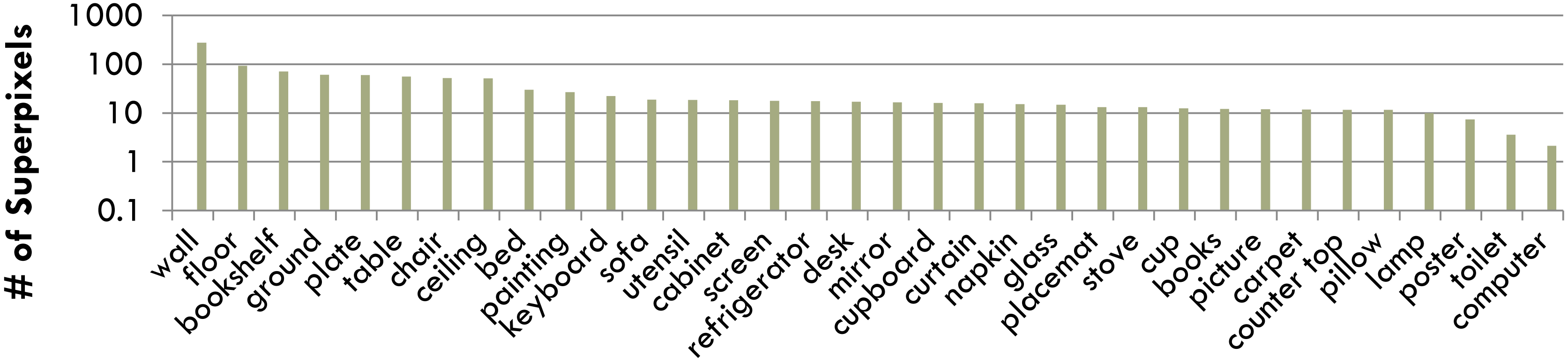
MRF labeling



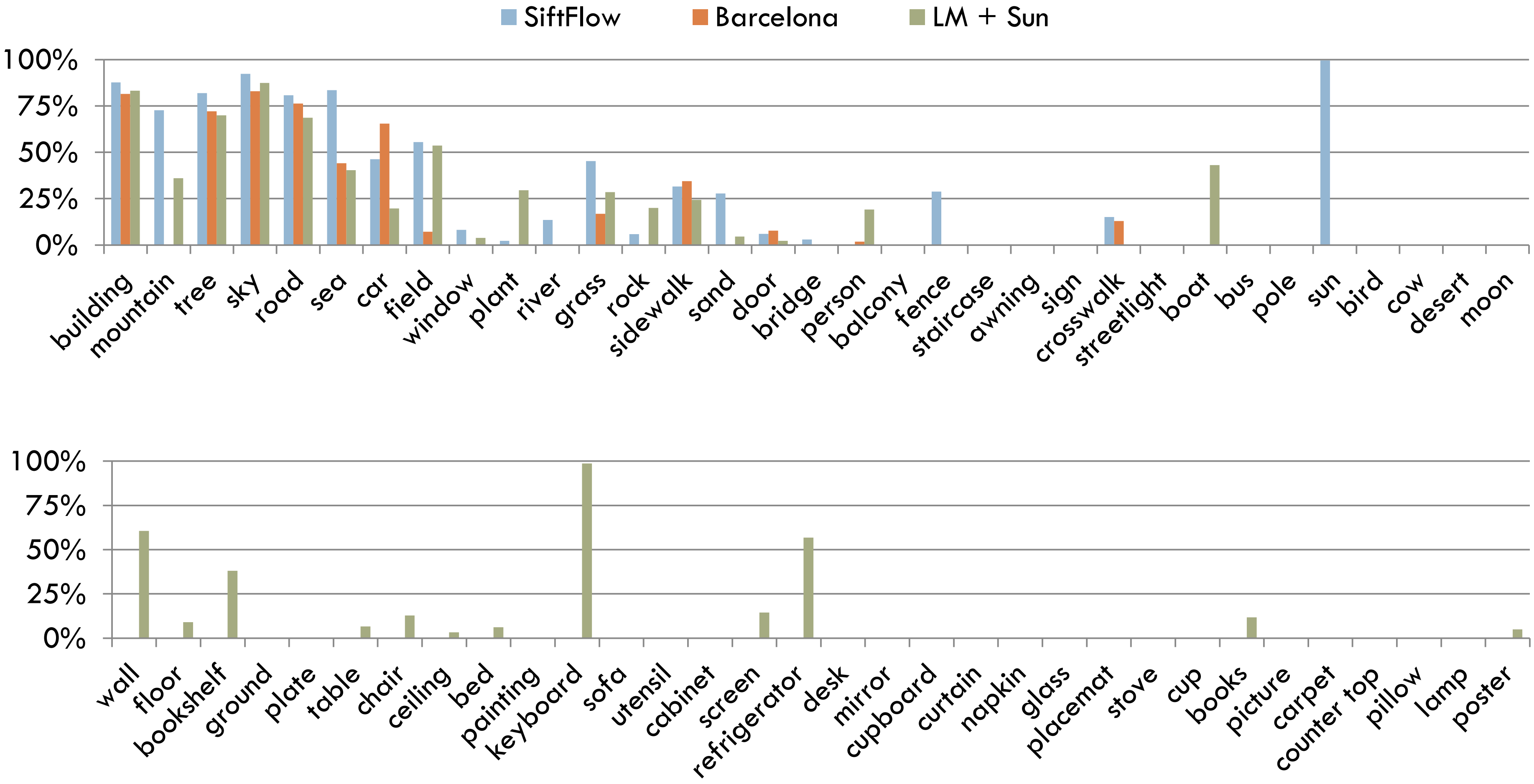
Datasets

	Training images	Test images	Labels
SIFT Flow (Liu et al., 2009)	2,488	200	33
Barcelona	14,871	279	170
LabelMe+SUN	50,424	300	232

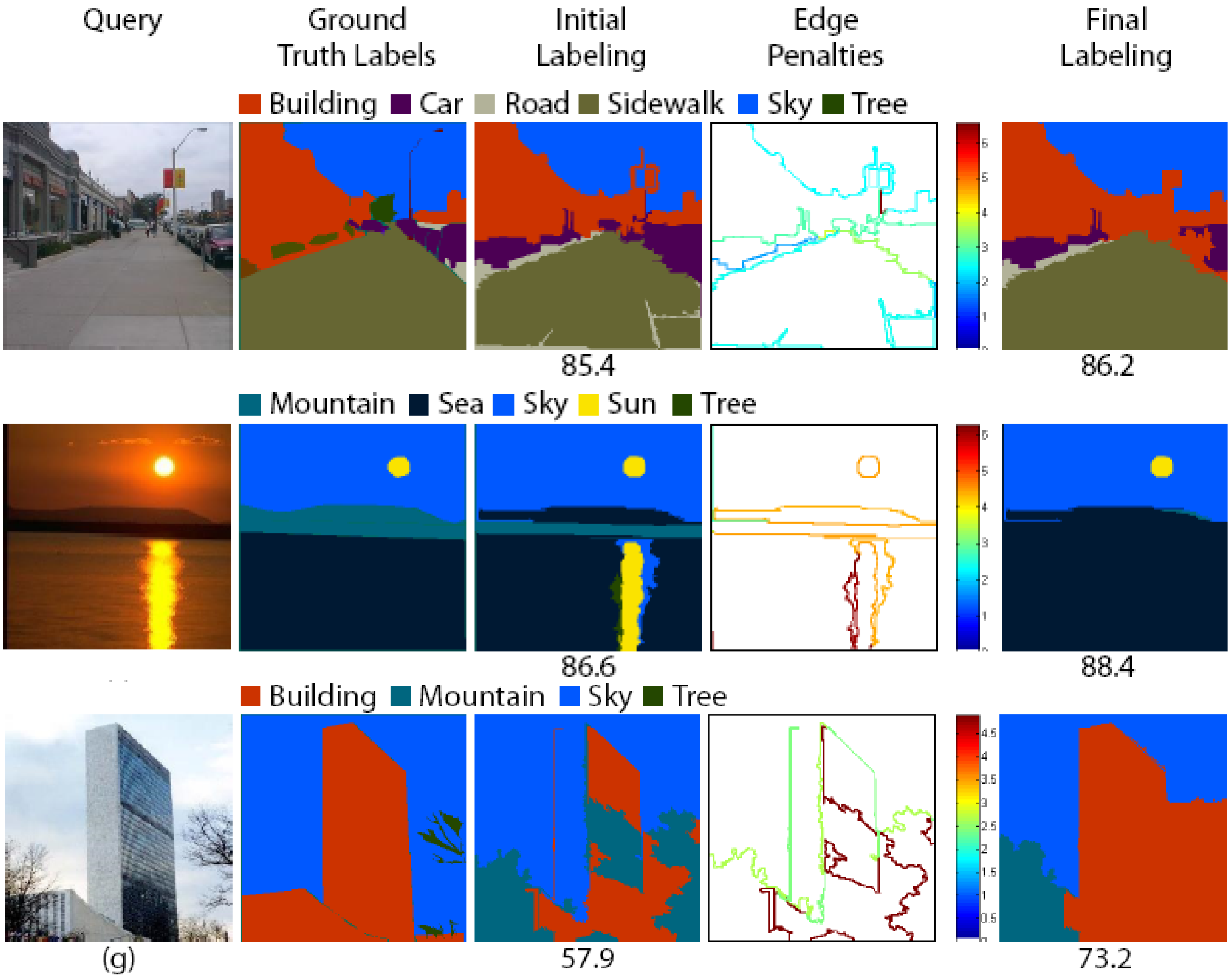




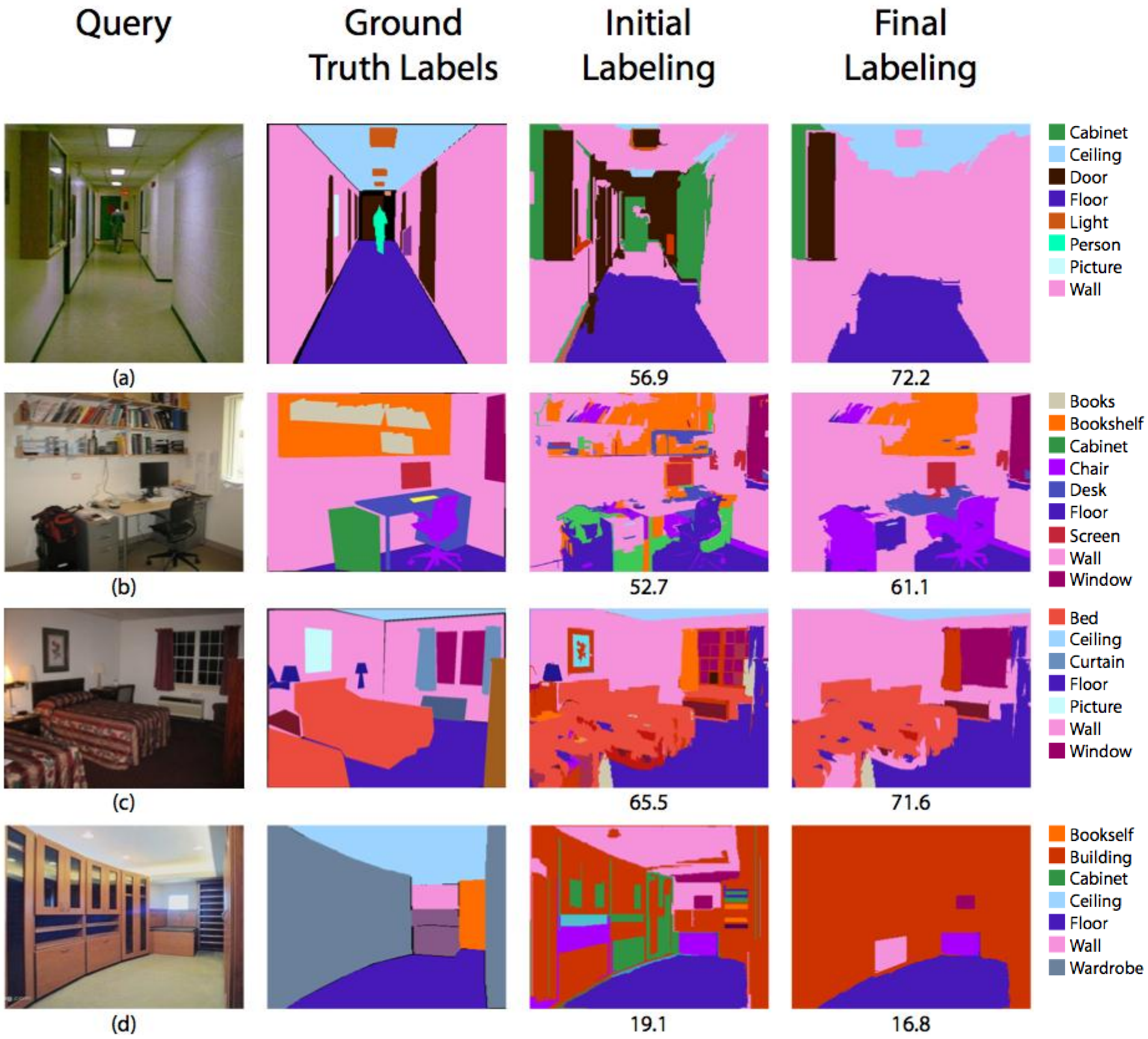
Per-class classification rates



Results on SIFT Flow dataset



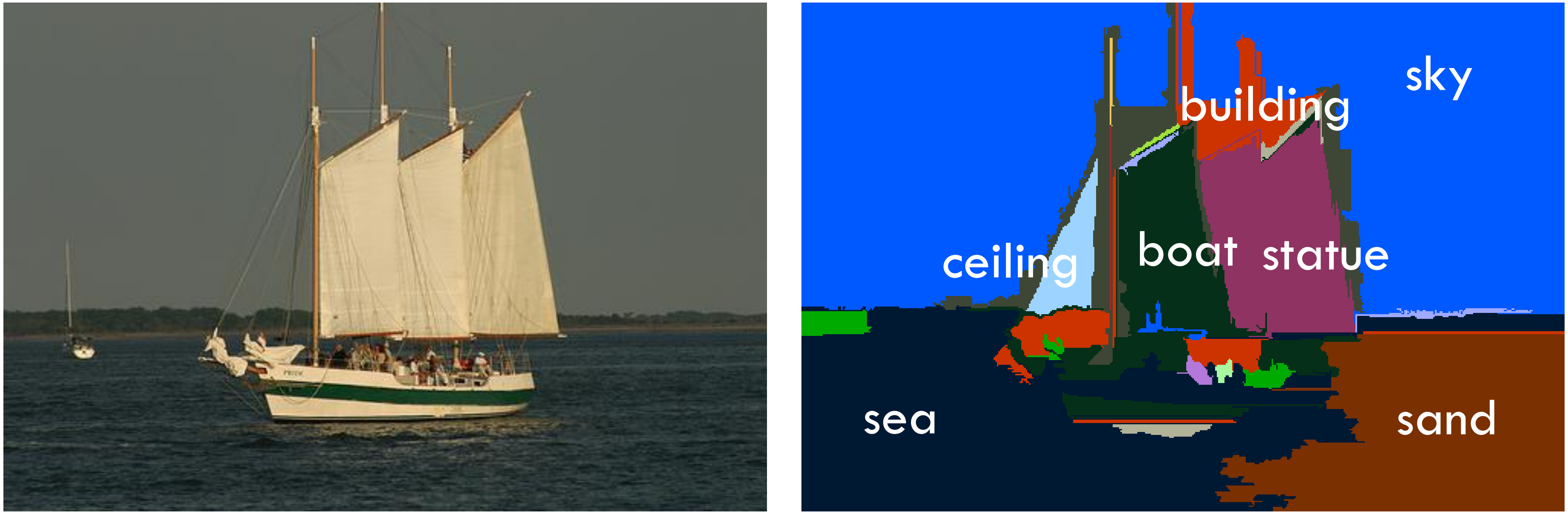
Results on LM+SUN dataset



Summary so far

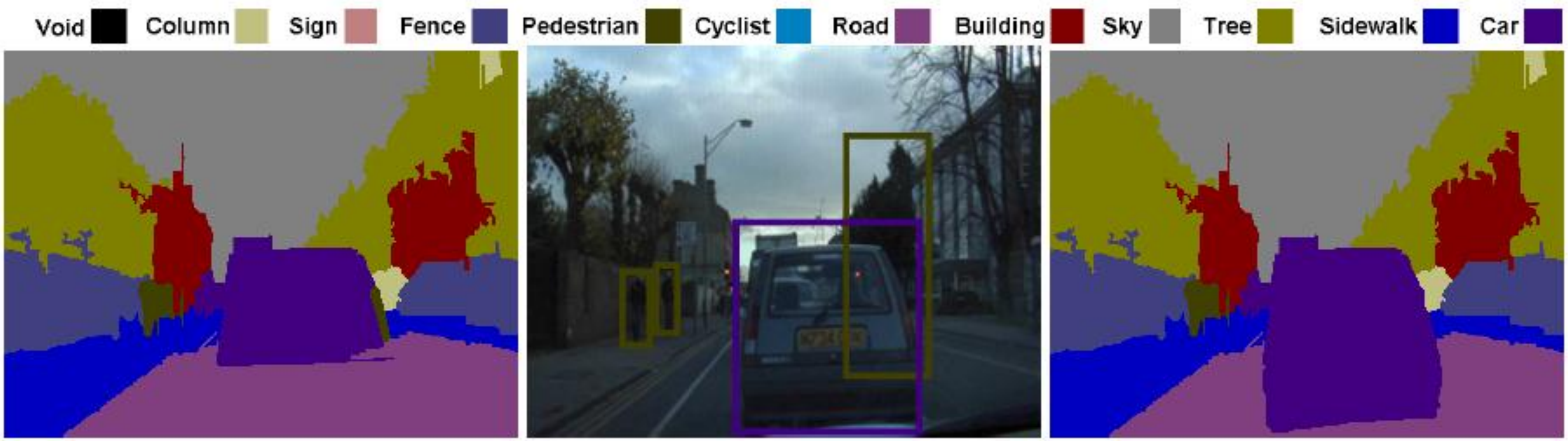
- A lazy learning method for image parsing:
  - Global scene matching
  - Superpixel-level matching
  - MRF optimization
- Challenges
  - Indoor images are hard!
  - We do well on “stuff” but not on “things”

We get the “stuff” but not the “things”



To get the "things" use detectors

- Ladicky et al. used detector output coupled with bounding box based foreground/background segmentation to improve performance on things



Result without detections

Set of detections

Final Result

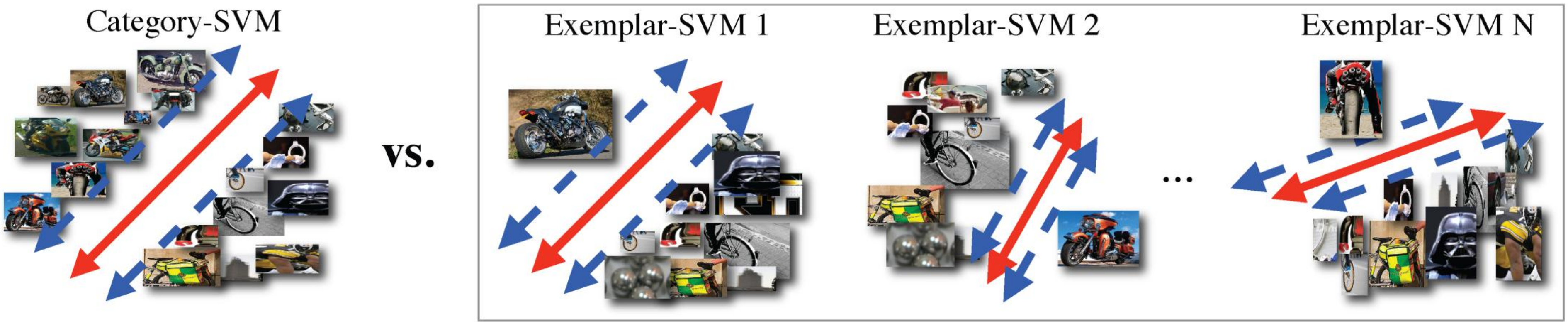
Ľubor Ladický, Paul Sturgess, Karteek Alahari, Chris Russell, Philip H.S. Torr
What, Where & How Many? Combining Object Detectors and CRFs, ECCV 2010

Problems with this approach

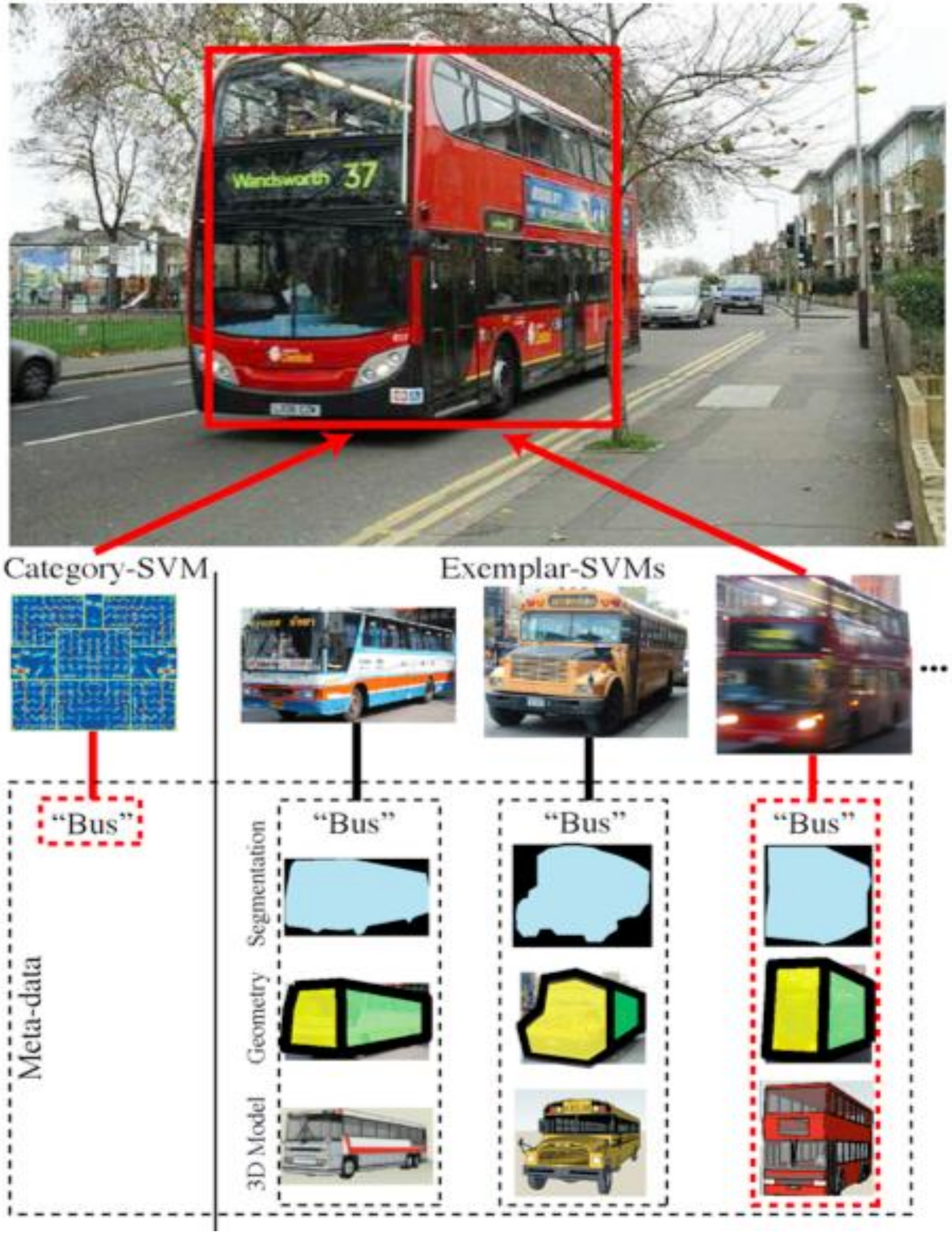
- The mask for bounding boxes is obtained by an automatic segmentation, which can fail
- The models must be pre-trained and cannot adapt to new data easily
- There is little flexibility for objects that take many forms

Per-exemplar detectors

- For each instance of a class: train SVM based on HOG features
- Negative examples are taken from all images that do not contain the class



Tomasz Malisiewicz, Abhinav Gupta, Alexei A. Efros. Ensemble of Exemplar-SVMs for Object Detection and Beyond. In ICCV, 2011

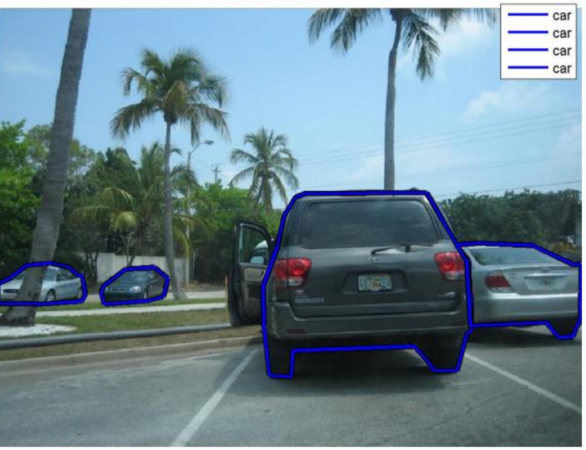


Tomasz Malisiewicz, Abhinav Gupta, Alexei A. Efros. Ensemble of Exemplar-SVMs for Object Detection and Beyond . In ICCV, 2011

Per-exemplar detectors for parsing

- Retrieve a set of similar images using global image descriptors
- Train per-exemplar detectors for "things" in retrieval set
- Run trained detectors on query and transfer weighted masked for all positive detections

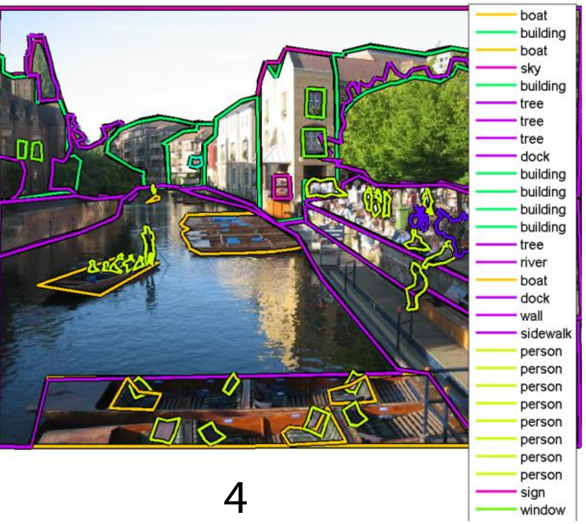
Retrieval set for



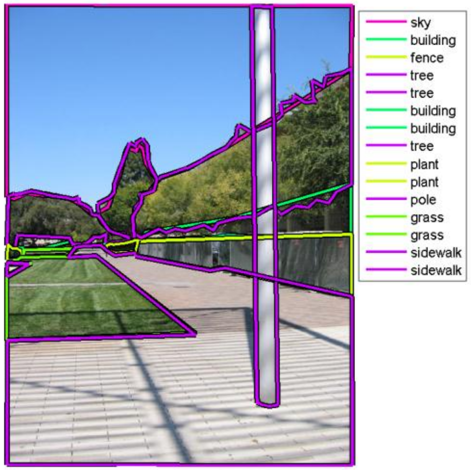
1

2

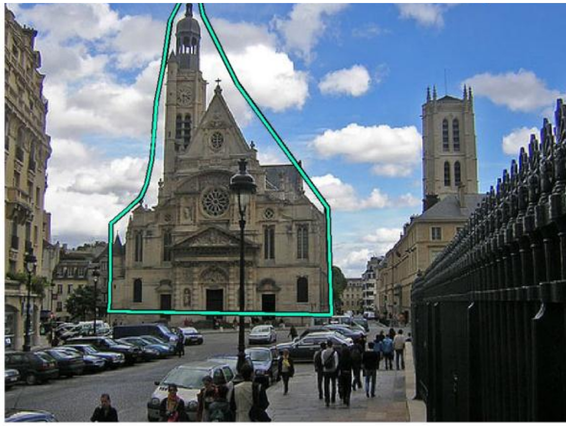
3



4

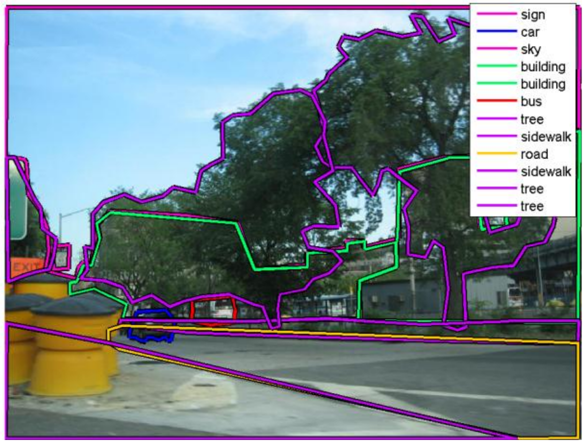


5

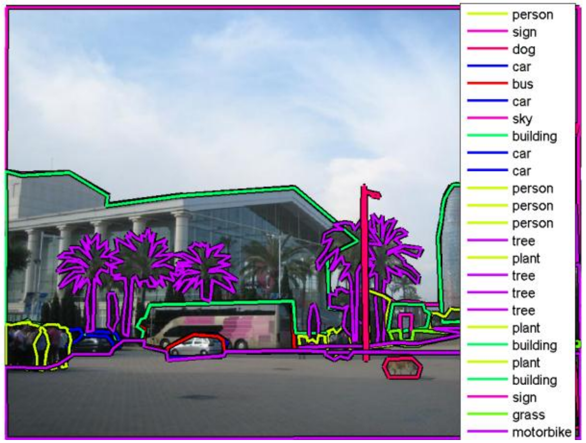


6

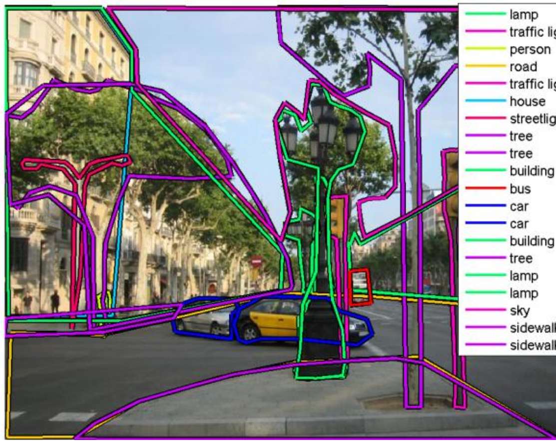
Retrieval set for



16



26



59

342

410

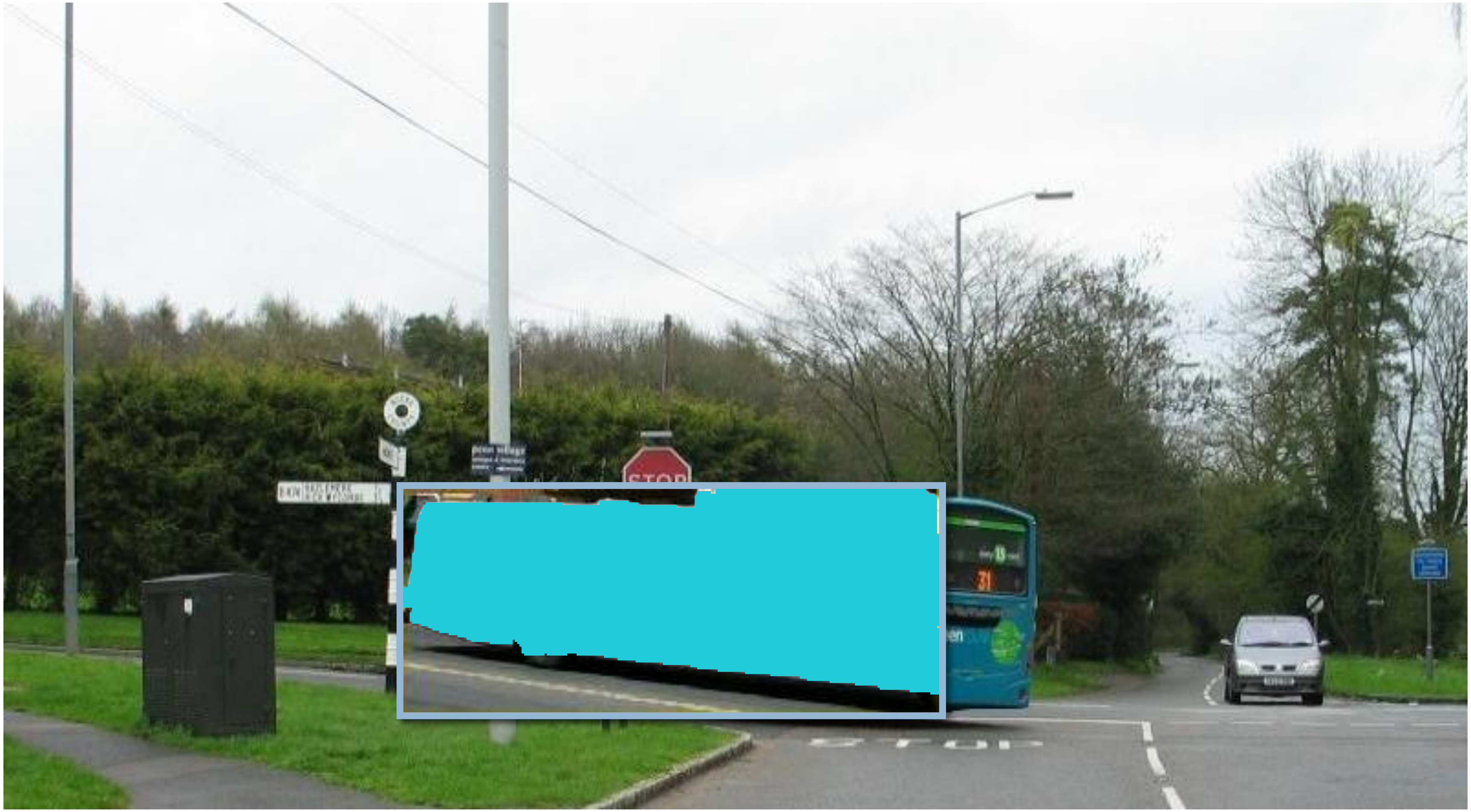
491

Per-exemplar detectors for parsing

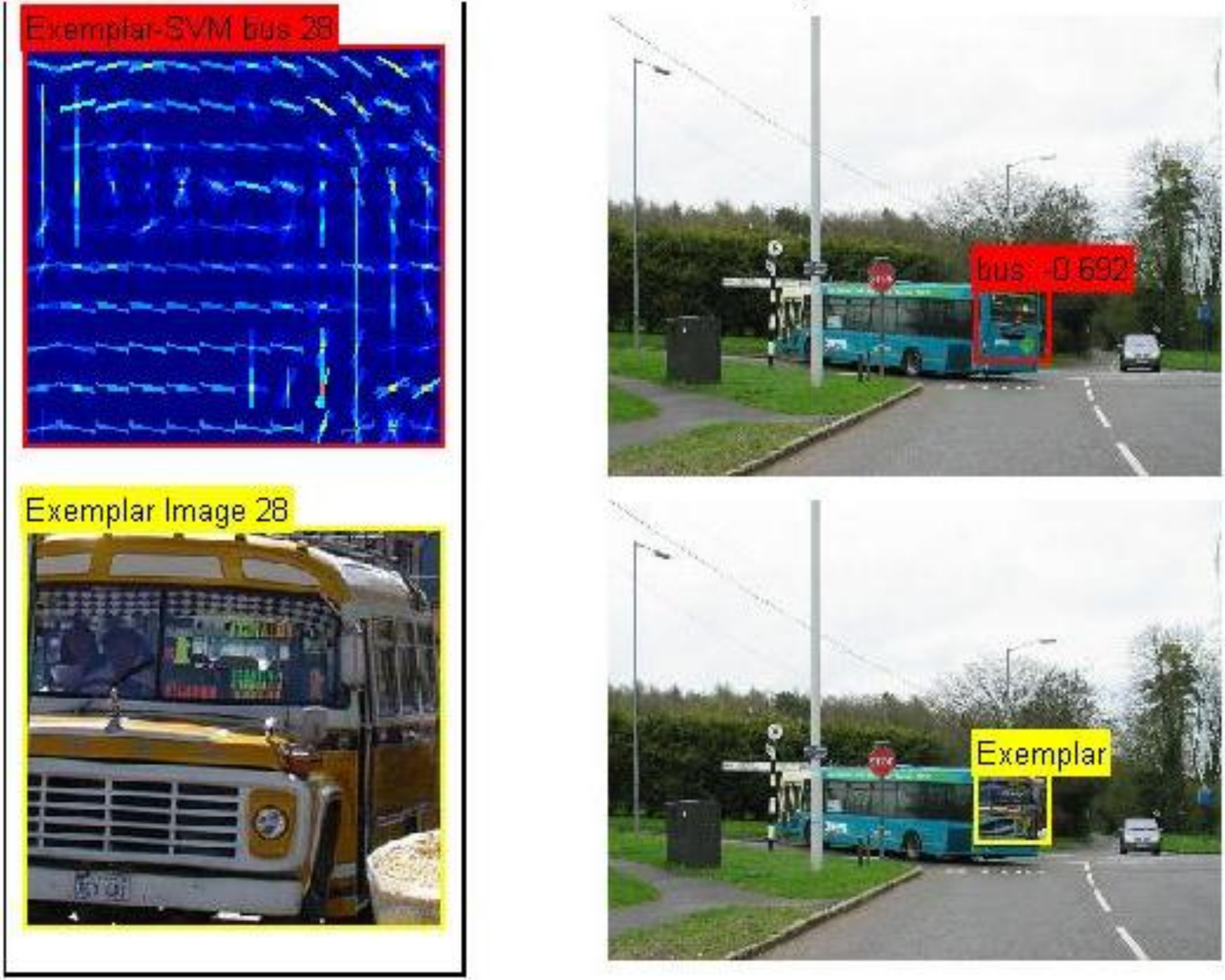
- Retrieve a set of similar images using global image descriptors
- Train per-exemplar detectors for each object in retrieval set
- Run trained detectors on query and transfer weighted masked for all positive detections

Per-exemplar detectors for parsing

Per-exemplar detectors for parsing



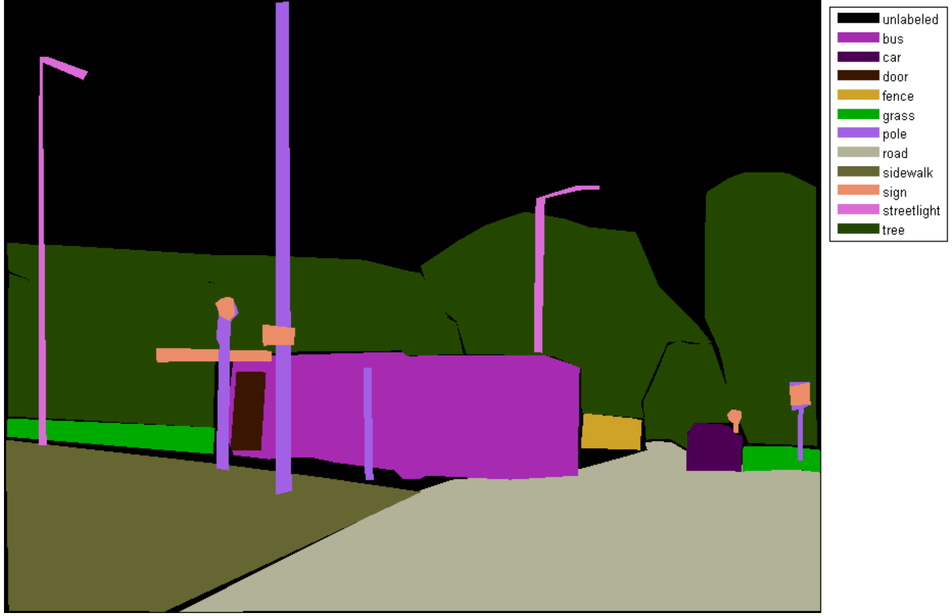

Exemplar SVM bus 20
Exemplar Image 20
bus: -0.625
Exemplar


Exemplar-SVM bus 28
Exemplar Image 28
bus -0.692
Exemplar

Per-exemplar detectors for parsing

- Retrieve a set of similar images using global image descriptors
- Train per-exemplar detectors for “things” in retrieval set
- Run trained detectors on query and transfer weighted masks for all positive detections

Per-exemplar detectors for parsing



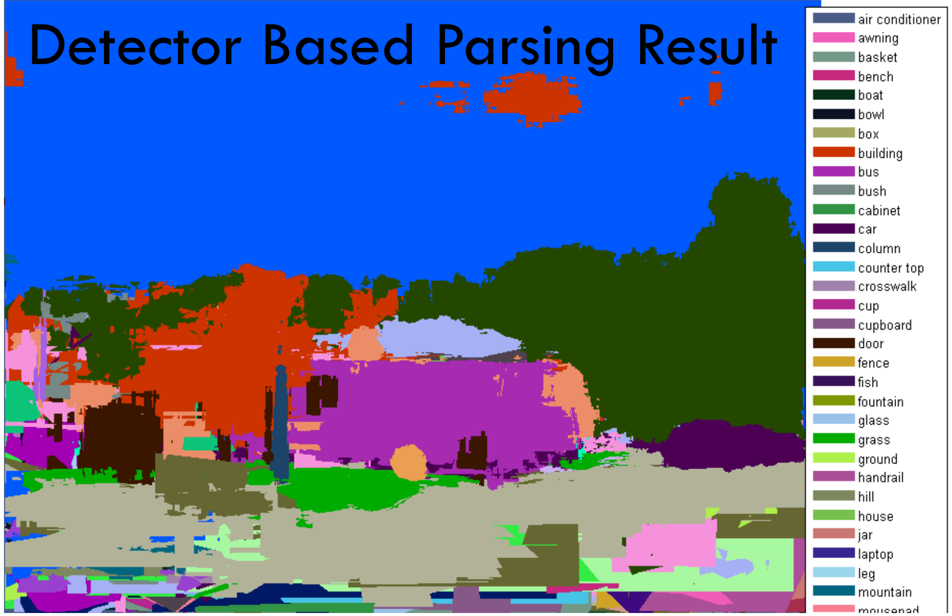
Superparsing Result

55% (23%)

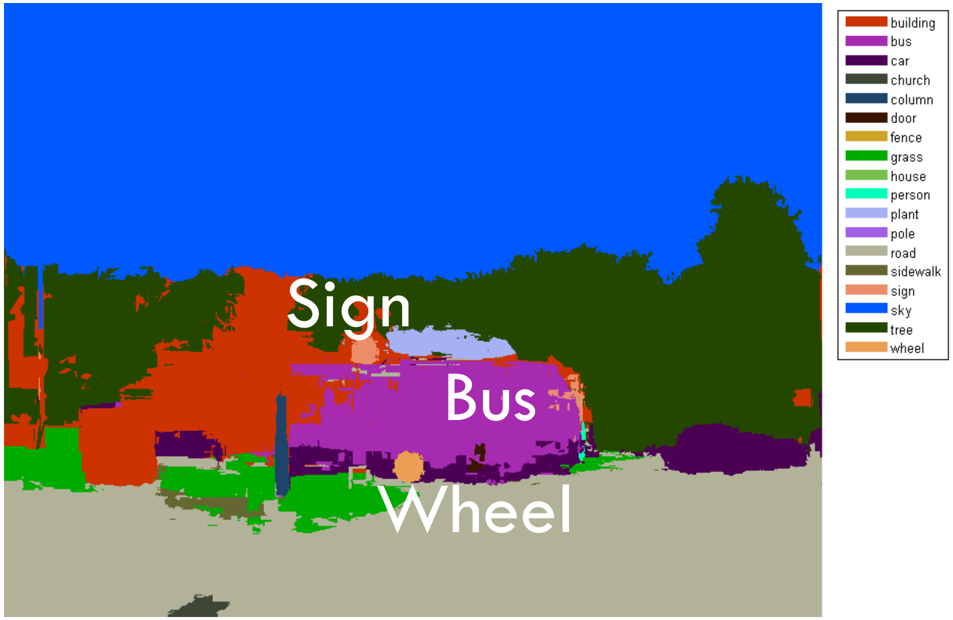
Detector-based Parsing Result

45% (26%)

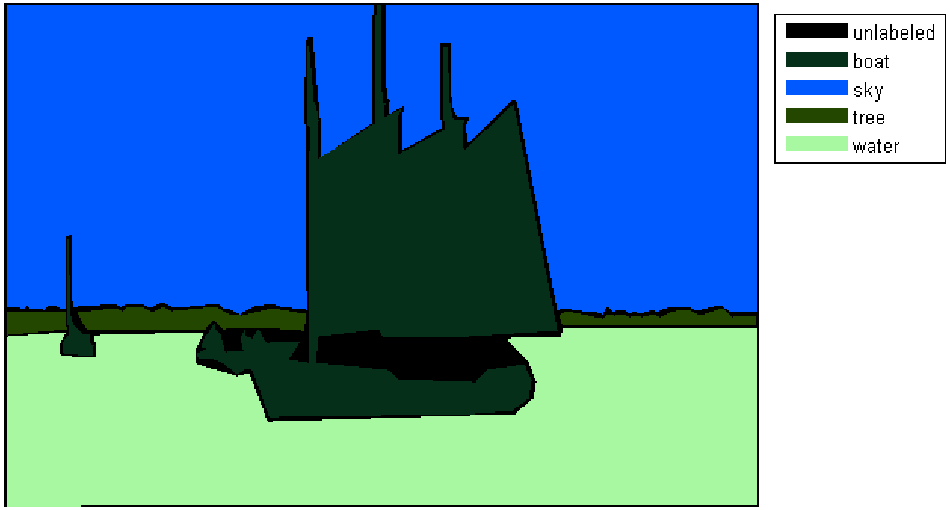
55% (23%)



45% (26%)

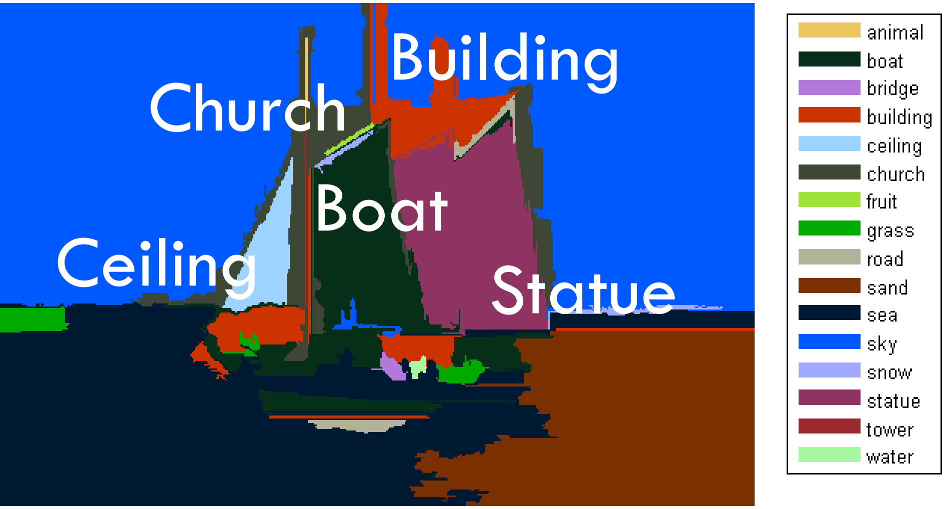


61% (31%)



Superparsing Result

Detector Based Parsing Result



52% (31%)

19% (25%)

Superparsing Result

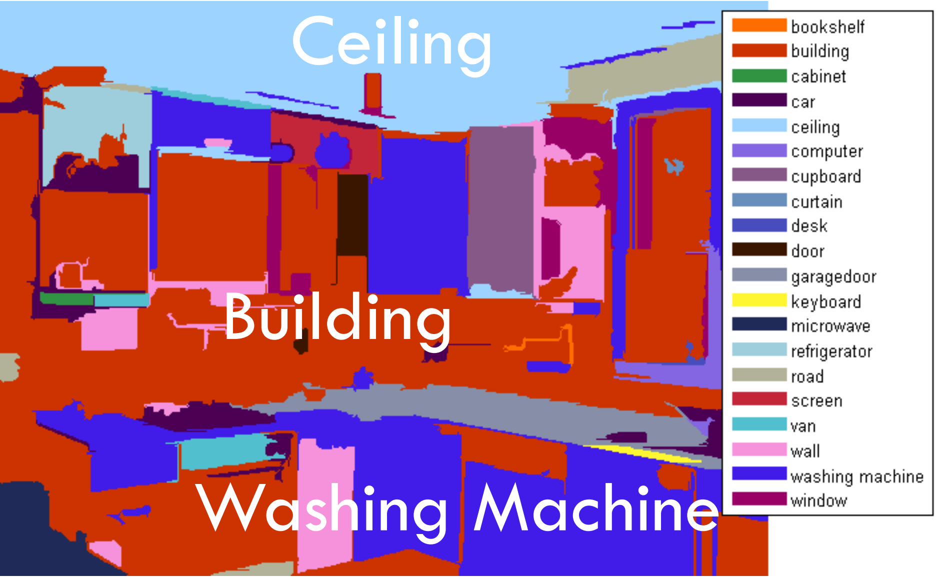
52% (31%)

Detector Based Parsing Result

19% (25%)

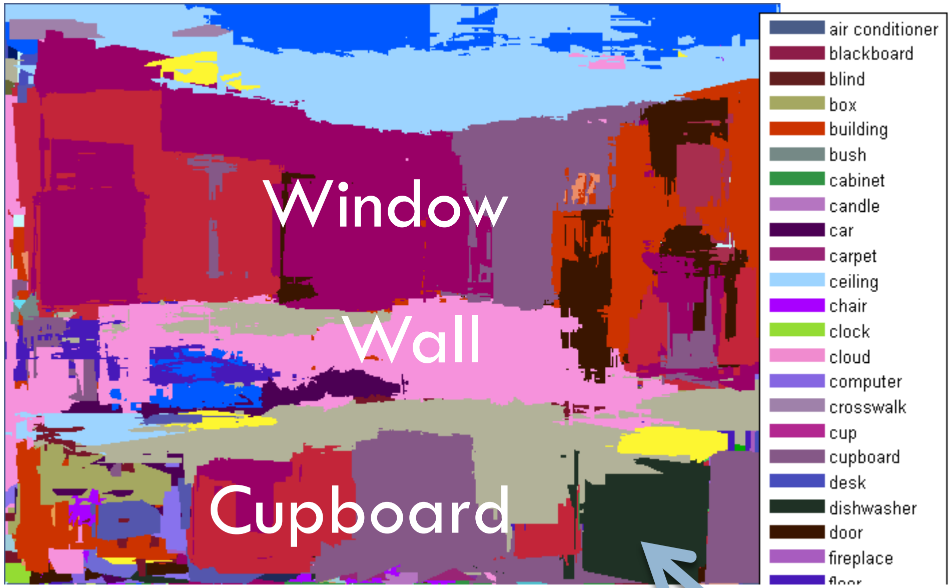
62% (46%)

Superparsing Result



12% (7%)

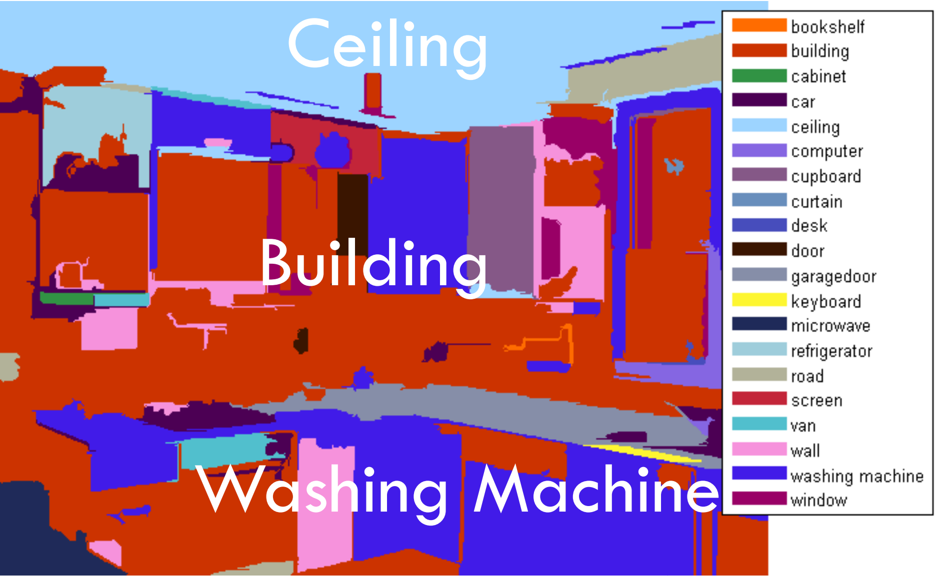
Detector Based Parsing Result



20% (9%)

Dishwasher

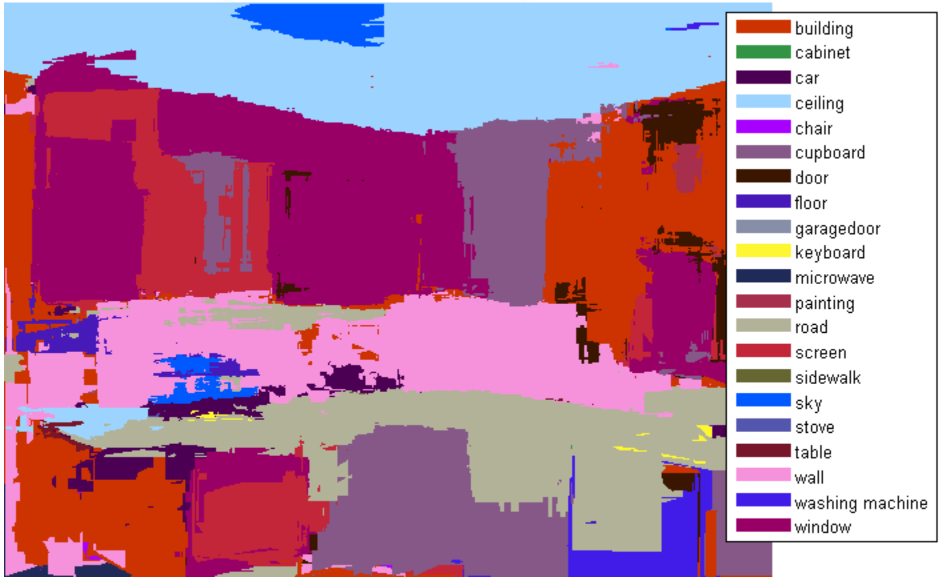
Superparsing Result



12% (7%)

Detector Based Parsing Result

20% (9%)



24% (10%)

Next Steps

- Determine which detectors to run
  - Manually select the “thing” classes
  - Use context from parsing
- Find better methods for integrating image parsing and detectors

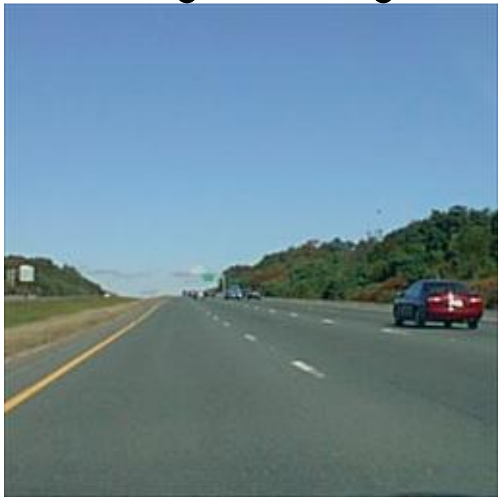
Review so far

- Image parsing with superpixels
  - Scene-level matching
  - Superpixel-level matching
  - MRF optimization
- Getting “things” with detectors
  - Use per-exemplar detectors of Malisiewicz et al.

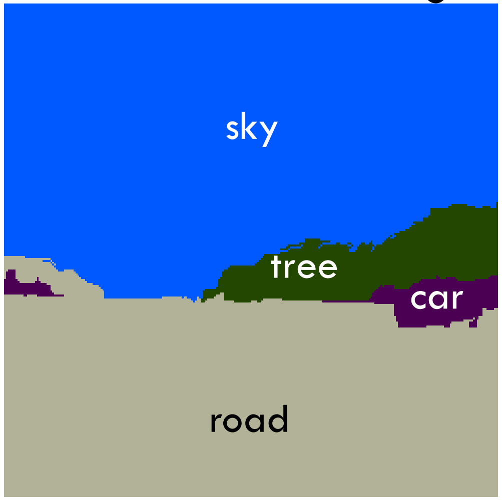
Joint geometric/semantic labeling

- **Semantic labels:** road, grass, building, car, etc.
- **Geometric labels:** sky, vertical, horizontal
  - Gould et al. (ICCV 2009)

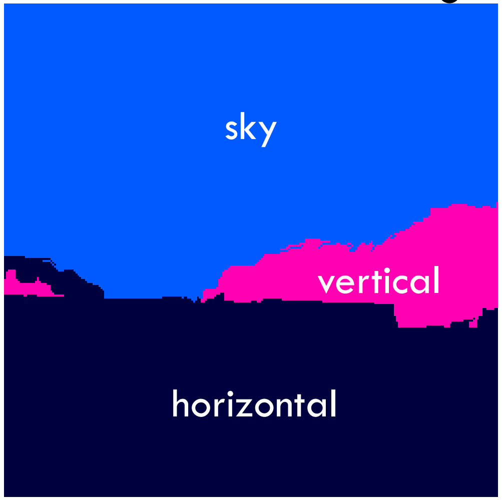
Original image



Semantic labeling



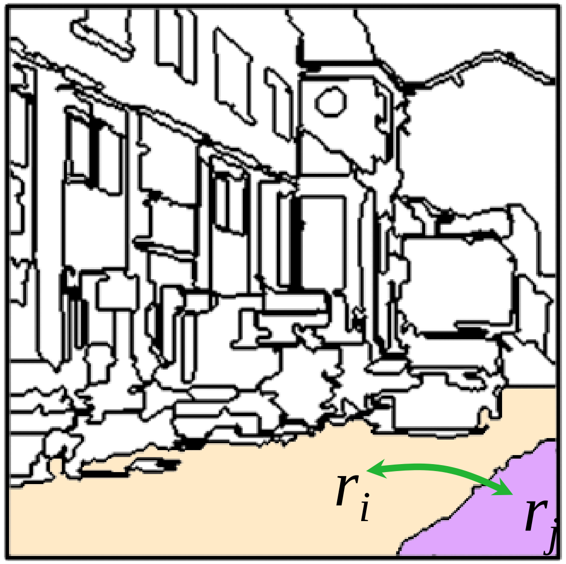
Geometric labeling



Recall: Global image labeling

- Compute a global image labeling by optimizing a Markov random field (MRF) energy function:

$$E(\underset{\substack{\uparrow\\ \textit{Vector of}\\ \textit{region}\\ \textit{labels}}}{\mathbf{c}}) = \underset{\textit{Regions}}{\sum_i} -\log \underbrace{L(r_i, c_i)}_{\substack{\textit{Likelihood score for}\\ \textit{region } r_i \textit{ and label } c_i}} \quad + \quad \lambda \underset{\substack{\textit{Neighboring}\\ \textit{regions}}}{\sum_{i,j}} \underbrace{\delta[c_i \neq c_j]}_{\substack{\textit{Smoothing}\\ \textit{penalty}}} \underbrace{\varphi(c_i, c_j)}_{\substack{\textit{Co-occurrence}\\ \textit{penalty}}}$$



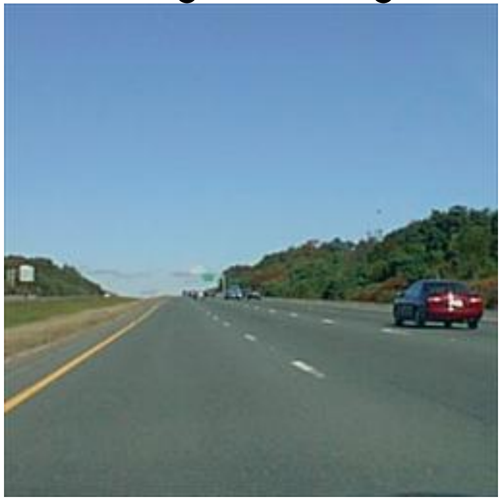
Efficient approximate minimization using α-expansion (Boykov et al., 2002)

Joint geometric/semantic labeling

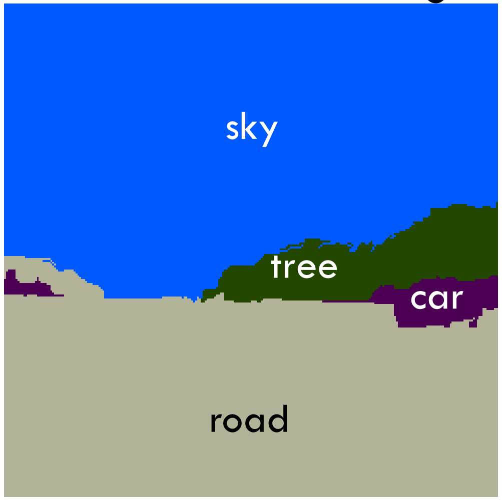
- Objective function for joint labeling:

$$F(\underset{\substack{\uparrow\\ \text{Semantic labels}}}{\mathbf{c}},\underset{\substack{\uparrow\\ \text{Geometric labels}}}{\mathbf{g}}) = \underbrace{E(\mathbf{c})}_{\text{Cost of semantic labeling}} + \underbrace{E(\mathbf{g})}_{\text{Cost of geometric labeling}} + \mu \sum_{\text{regions} r_i} \underbrace{\psi(c_i, g_i)}_{\text{Geometric/semantic consistency penalty}}$$

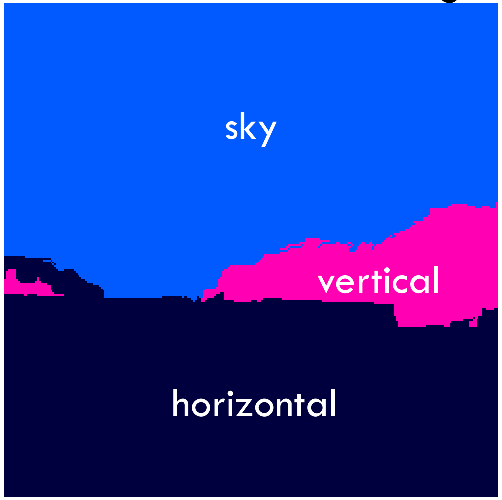
Original image



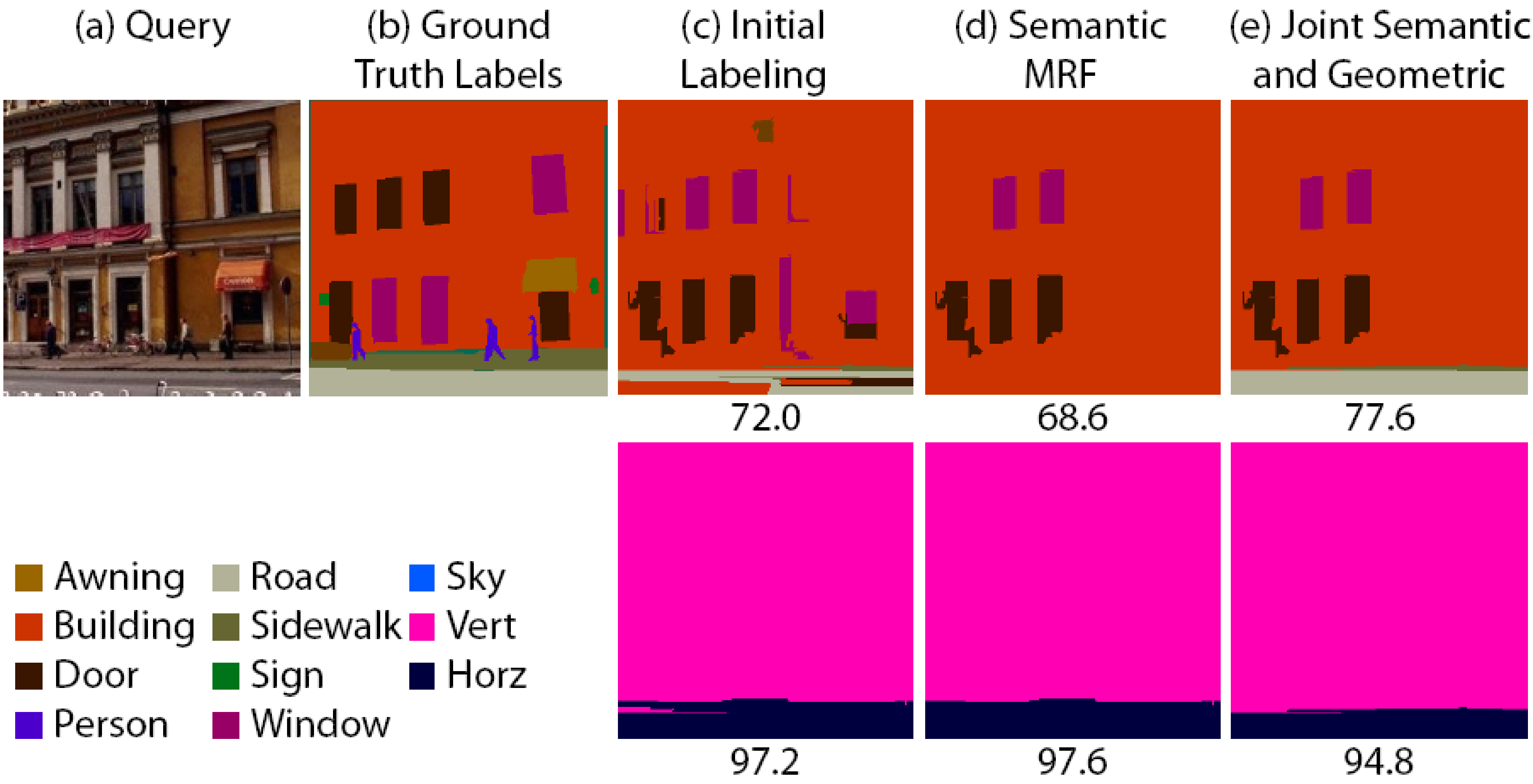
Semantic labeling



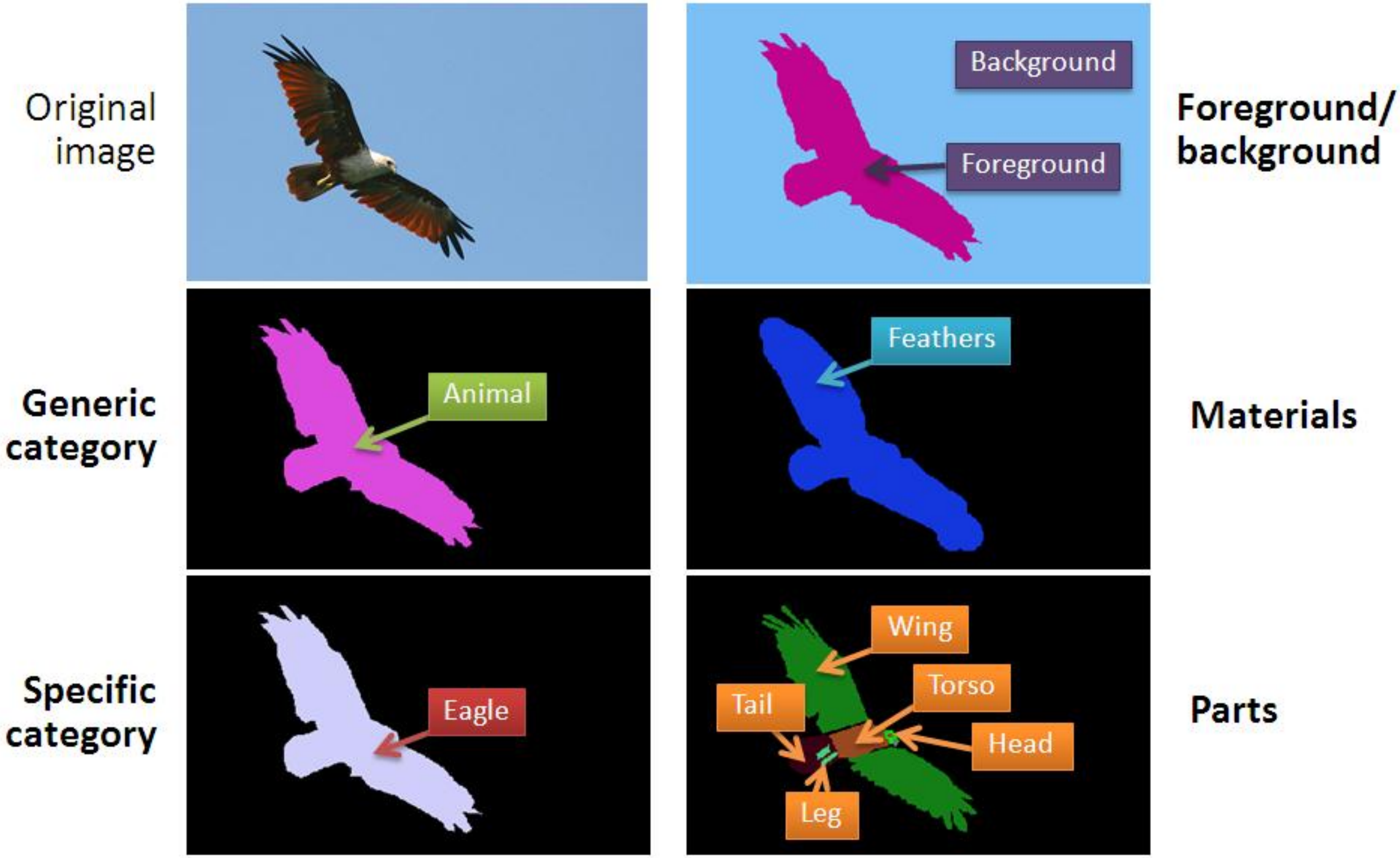
Geometric labeling



Example of joint labeling



Understanding scenes on many levels



Tighe and Lazebnik, ICCV 2011

Understanding scenes on many levels

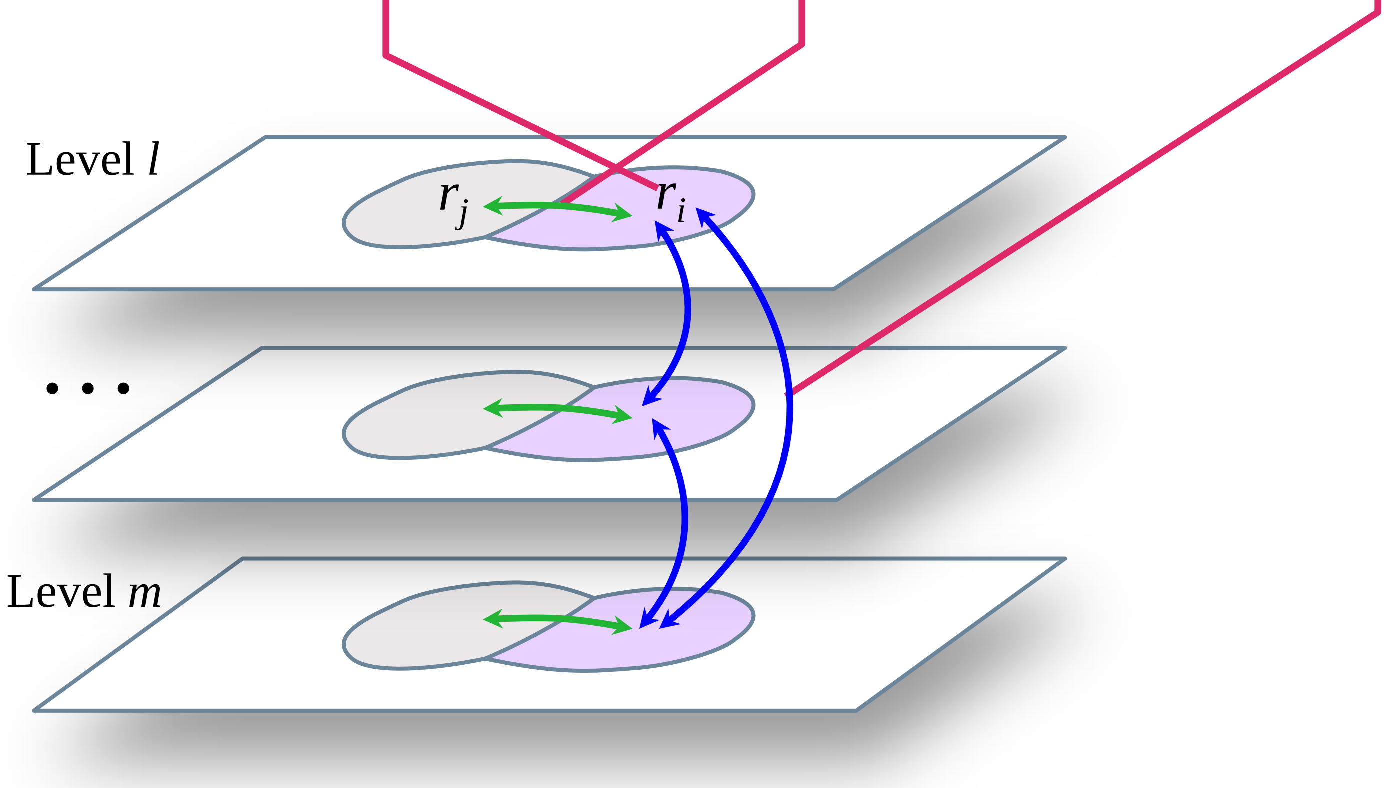
$$E(\mathbf{c}^1,\ldots,\mathbf{c}^L) = \sum_l \left(\underbrace{\sum_i E^l_{\text{data}}(r_i, c^l_i)} + \underbrace{\sum_{i,j} E^l_{\text{horiz}}(c^l_i, c^l_j)} \right) + \sum_{l \neq m} \sum_i \underbrace{E^{l,m}_{\text{vert}}(c^l_i, c^m_i)}$$

Region labelings for every level

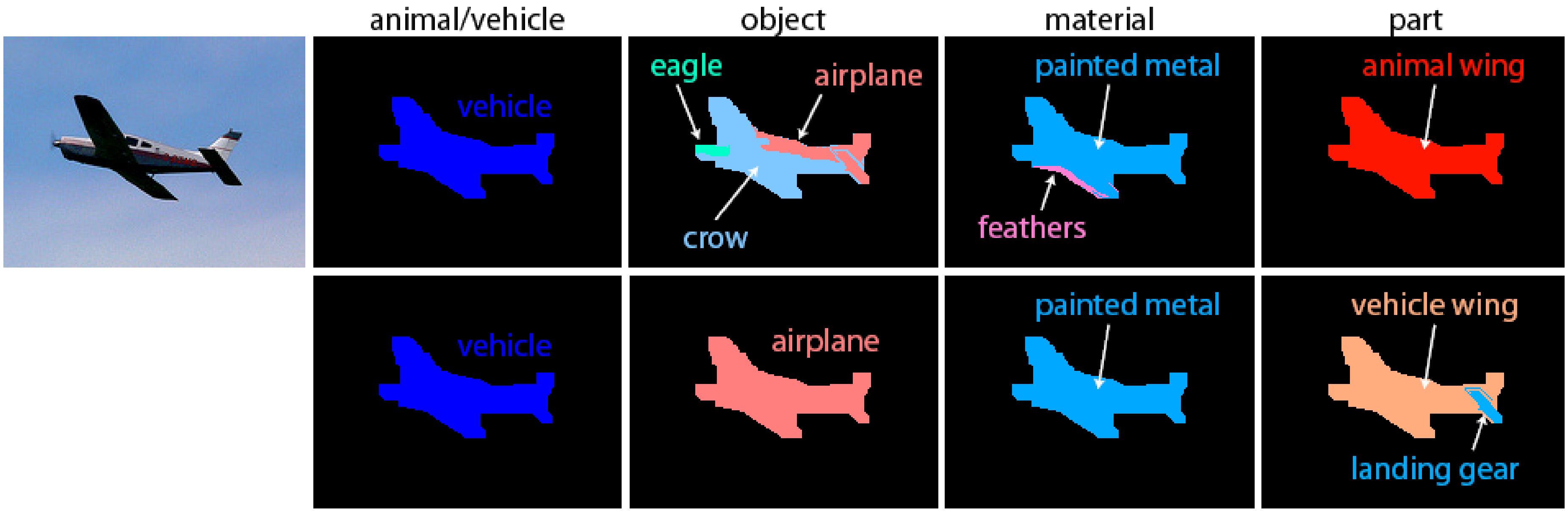
Likelihood score for level l, region r_i and label c_i

Single-level MRF Consistency penalty for neighbors on the same level

Consistency penalty for labels of region i on levels l and m



Understanding scenes on many levels



Review

- Nonparametric image parsing
- Beyond superpixels
- Beyond unique labels

Review so far

- Image parsing with superpixels
  - Scene-level matching
  - Superpixel-level matching
  - MRF optimization
- Getting “things” with detectors
  - Use per-exemplar detectors of Malisiewicz et al.
- Better scene understanding with multi-level labelings

Beyond labels: Attributes