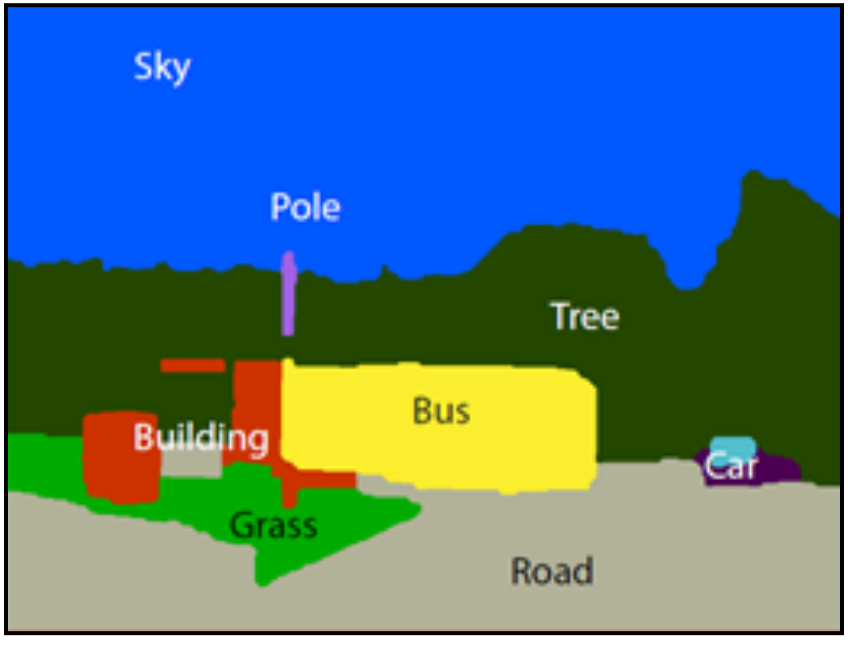
Finding Things: Image Parsing with Regions and Per-Exemplar Detector

Joseph Tighe and Svetlana Lazebnik

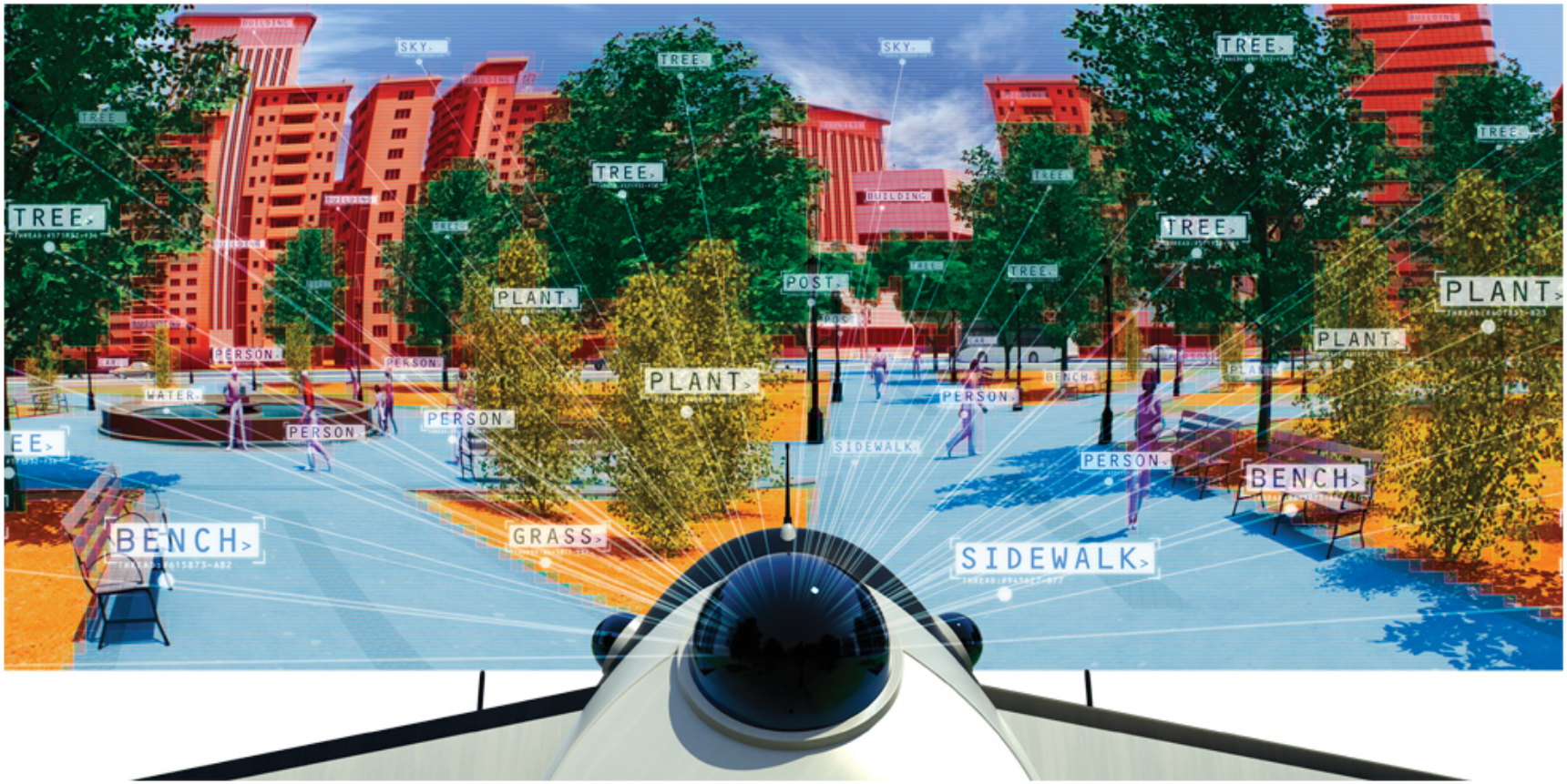
UNC Chapel Hill, University of Illinois at Urbana-Champaign



THE UNIVERSITY of NORTH CAROLINA at CHAPEL HILL

Semantic Segmentation

Definition: assign to each image pixel a label from a predefined set.



Closed-universe image parsing

Tens of classes, thousands of images, fixed datasets

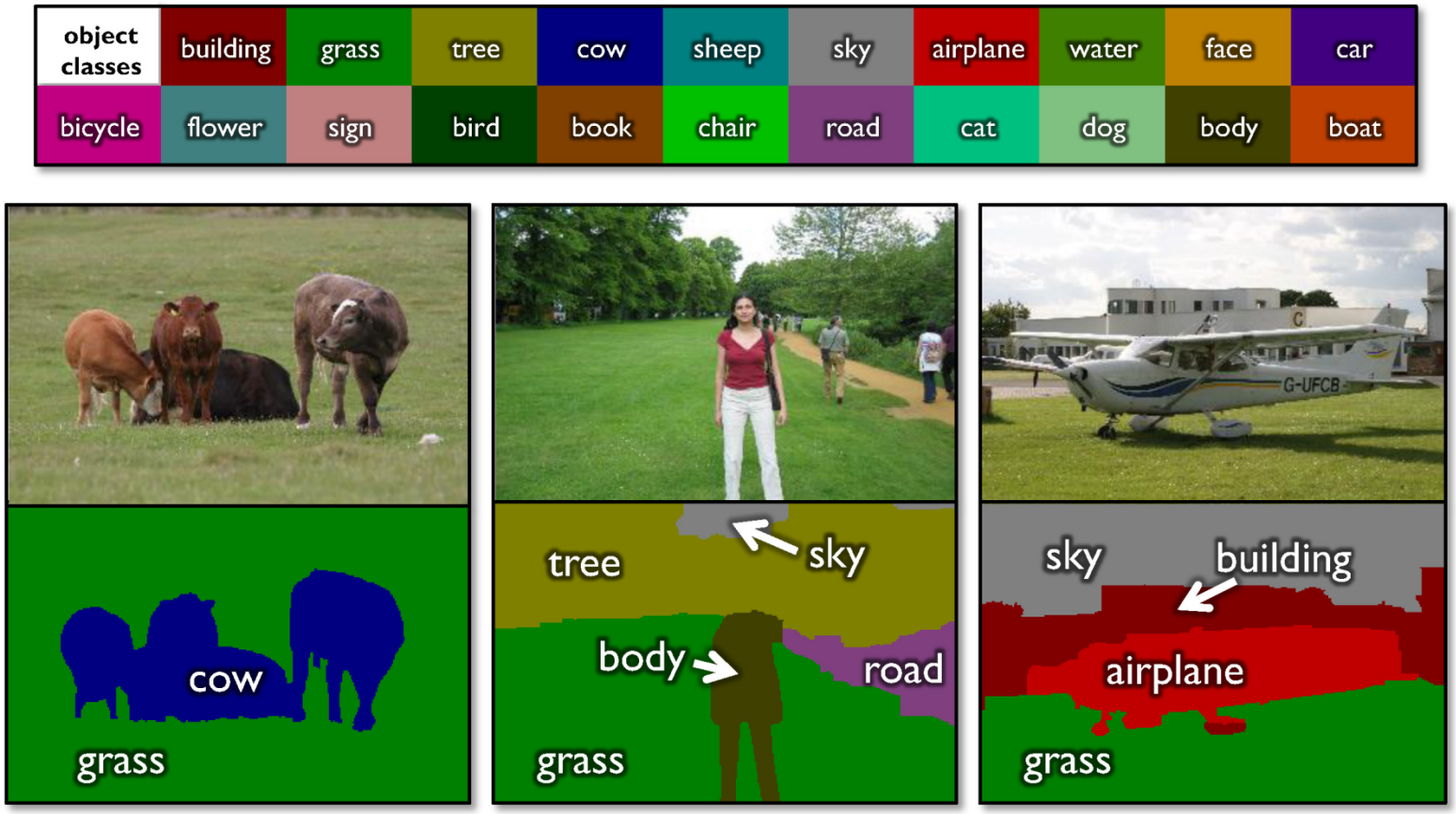
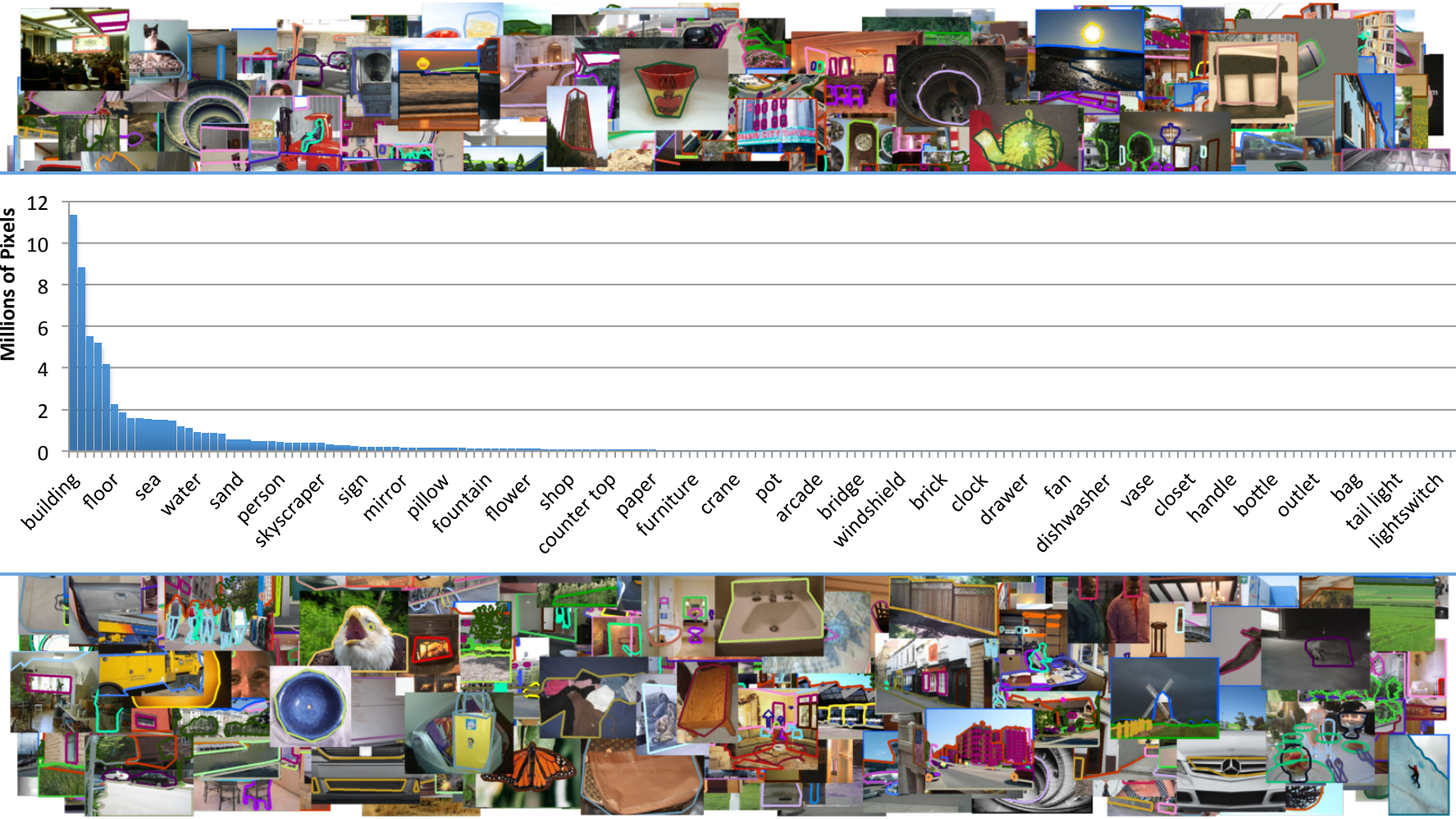


Figure from Shotton et al. (2009)

He et al. (2004), Hoiem et al. (2005), Shotton et al. (2006, 2008, 2009), Verbeek and Triggs (2007), Rabinovich et al. (2007), Galleguillos et al. (2008), Gould et al. (2009), etc.

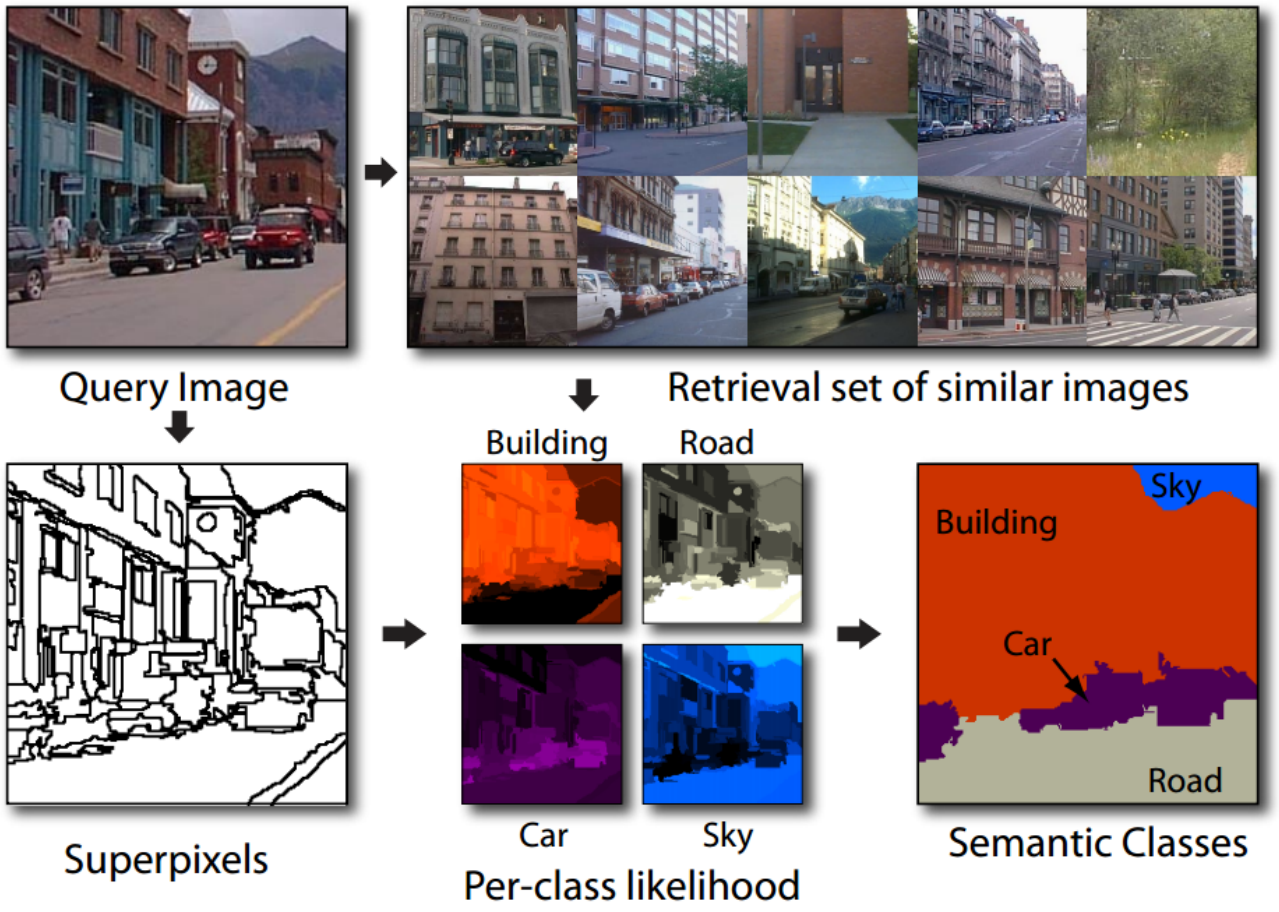
Large-scale open-universe parsing

Hundreds of classes, tens of thousands of images, evolving datasets



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

Nonparametric region-based approach

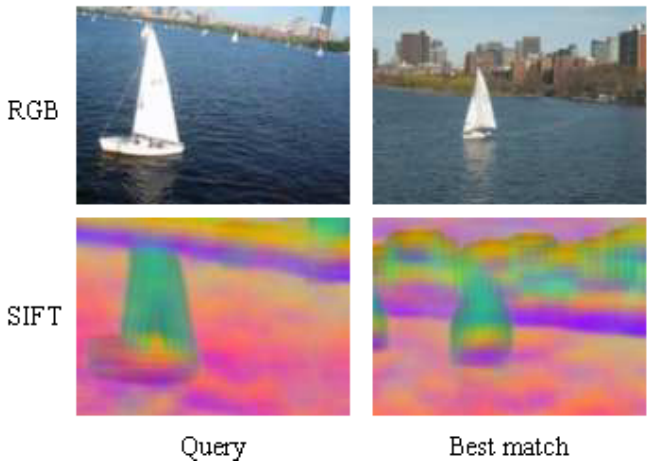


SuperParsing: Scalable Nonparametric Image Parsing with Superpixels
J. Tighe and S. Lazebnik, ECCV 2010, IJCV 2013

Nonparametric region-based approach

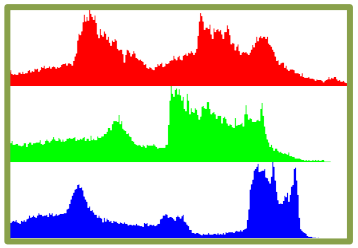
- **Lazy learning:** do (almost) nothing at training time
- At test time:
  - Find a *retrieval set* of similar images for each query image
  - Transfer labels from the retrieval set by matching segmentation regions (superpixels)

- **Related work:** SIFT Flow (Liu et al. 2008, 2009)

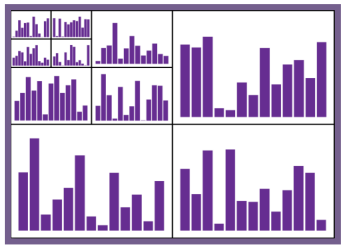


Step 1: Scene-level matching

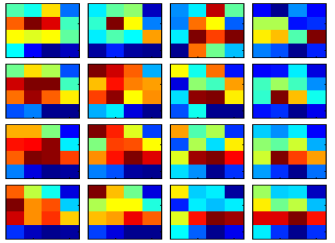
Color Histogram



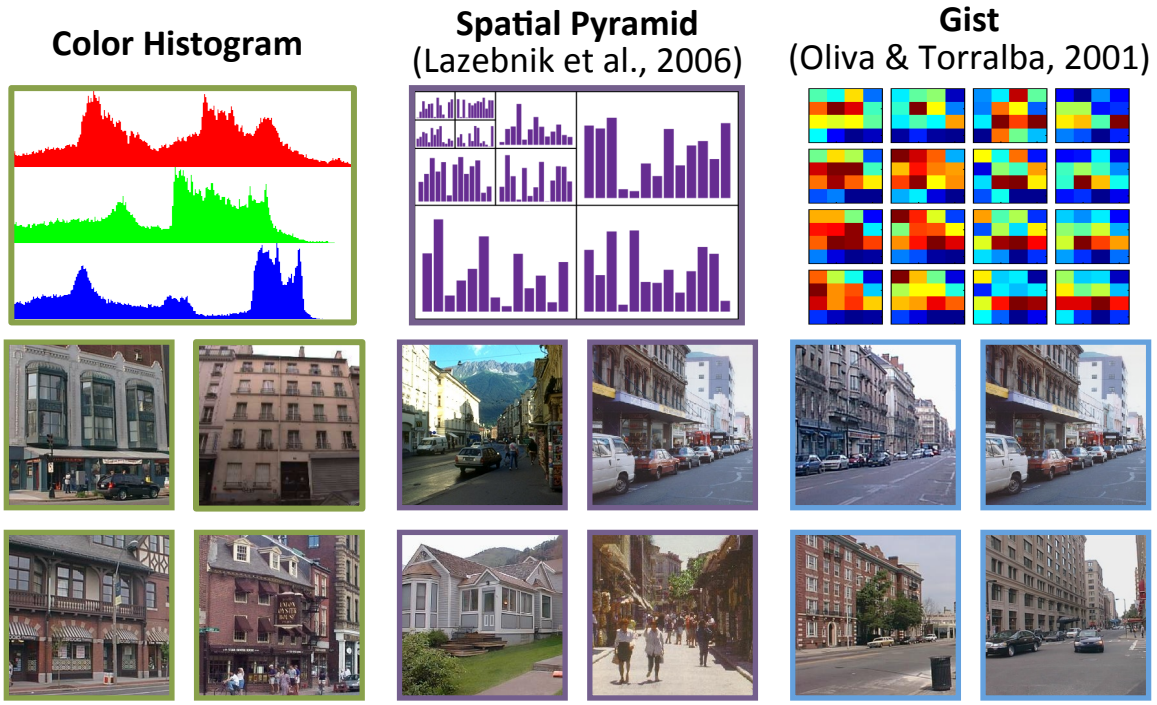
Spatial Pyramid
(Lazebnik et al., 2006)



Gist
(Oliva & Torralba, 2001)



Step 1: Scene-level matching



Compute using euclidean distance for each feature f

Retain the top K images (K=200)

Step 2: Region-level matching

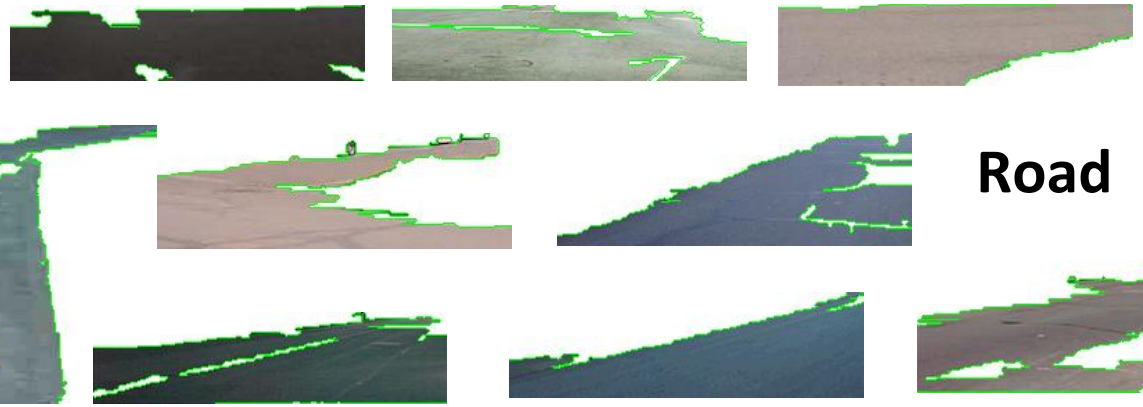
Superpixels
(Felzenszwalb & Huttenlocher, 2004)

Superpixel features

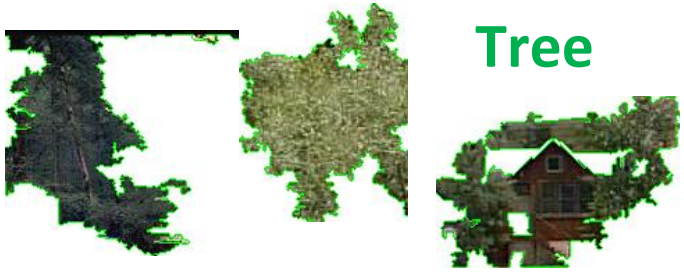
Shape	Mask of superpixel shape over its bounding box (8×8)	64
	Bounding box width/height relative to image width/height	2
	Superpixel area relative to the area of the image	1
Location	Mask of superpixel shape over the image	64
	Top height of bounding box relative to image height	1
Texture/SIFT	Texton 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)

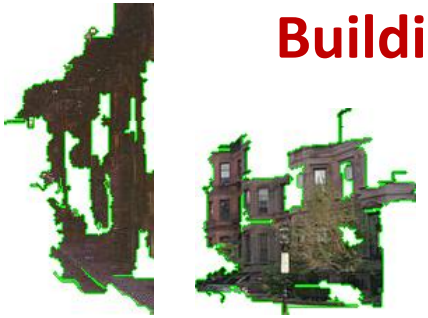


Road

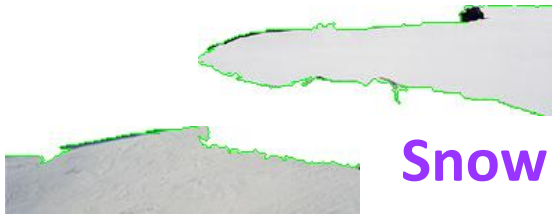


Tree

Sky

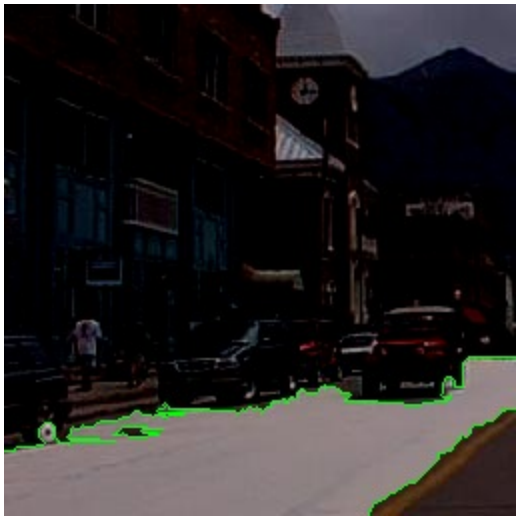


Building



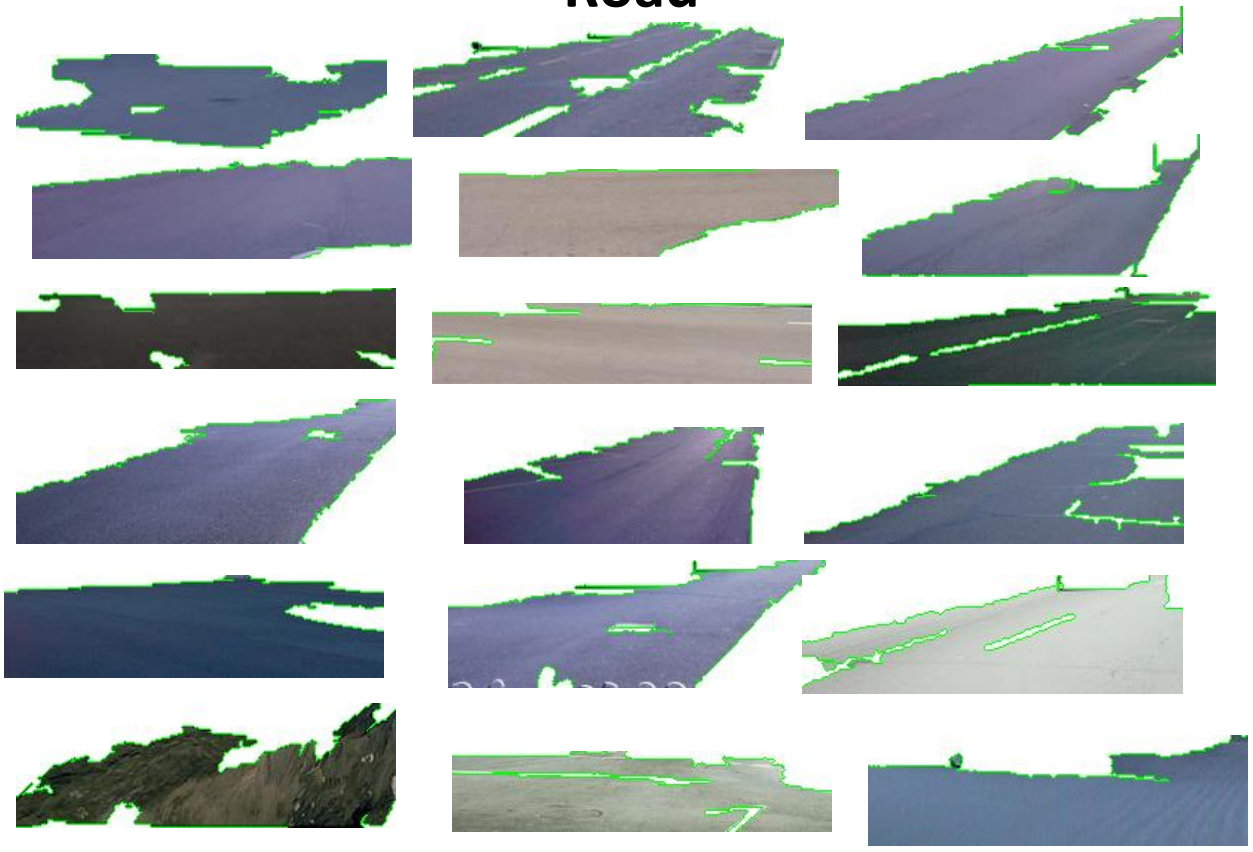
Snow

Step 2: Region-level matching



Absolute mask (location)

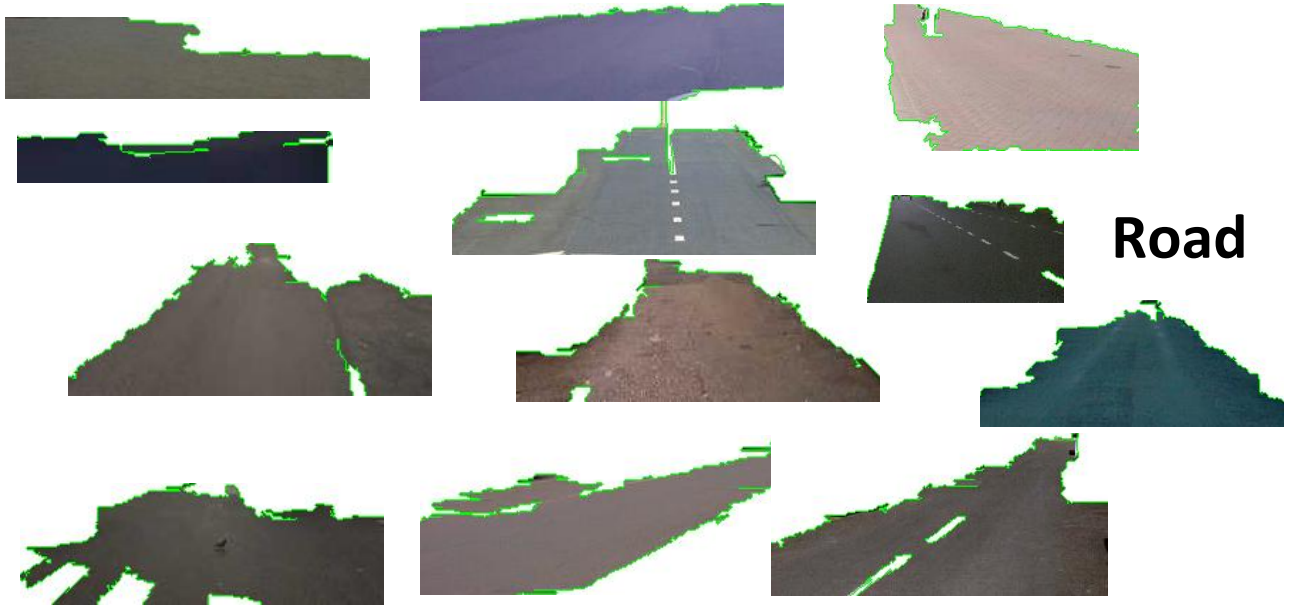
Road



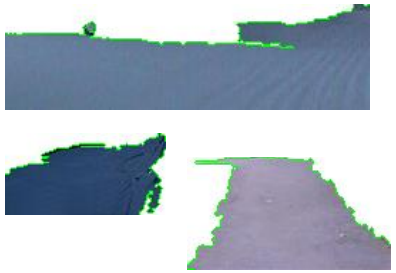
Sidewalk

Step 2: Region-level matching

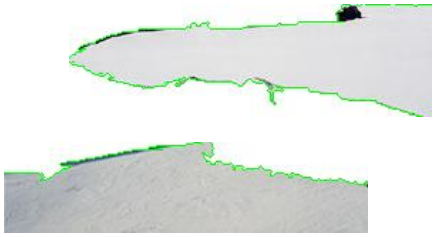
Texture



Road



Sidewalk

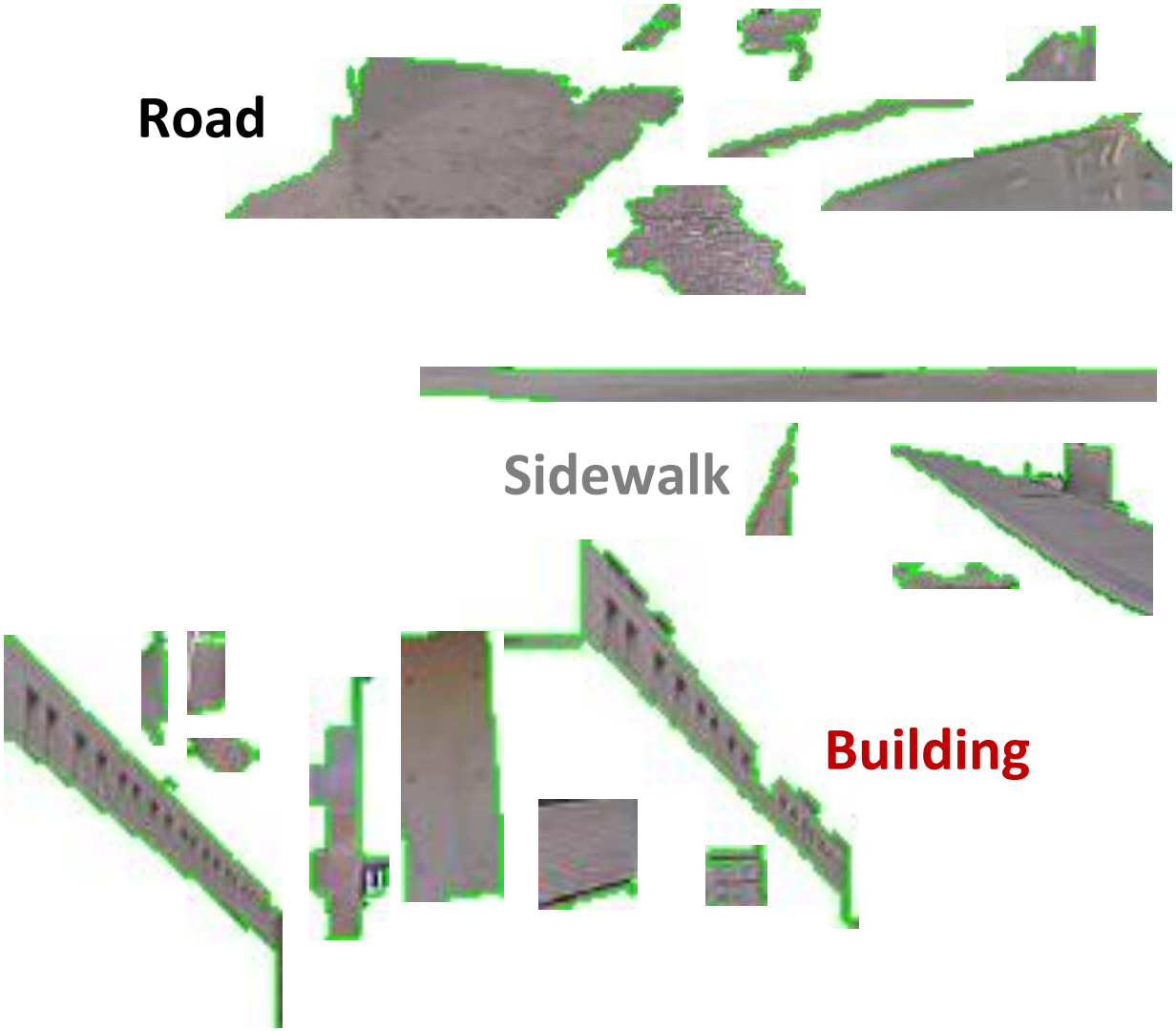


Snow

Sky

Step 2: Region-level matching

Color histogram



Region-level likelihoods

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

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

kth feature type of ith region

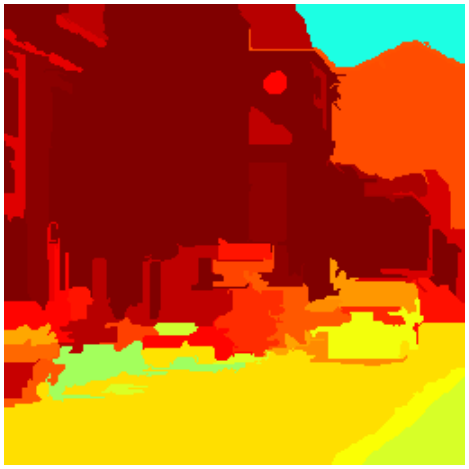
Regions of class c within some radius of r_i

Total features of class c in the dataset

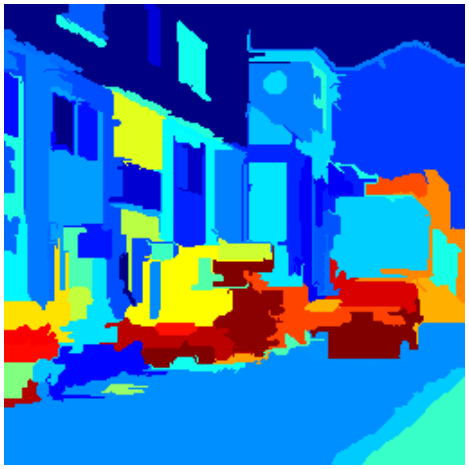
- Per-feature likelihoods combined via Naïve Bayes:

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

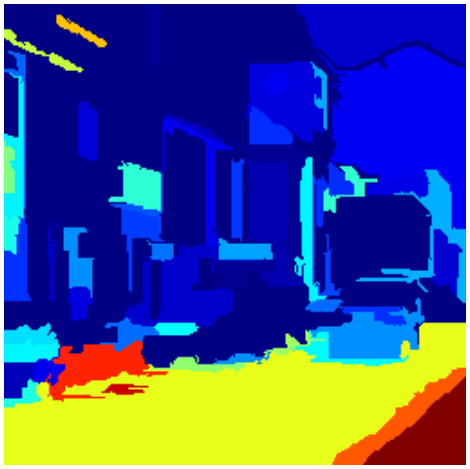
Region-level likelihoods



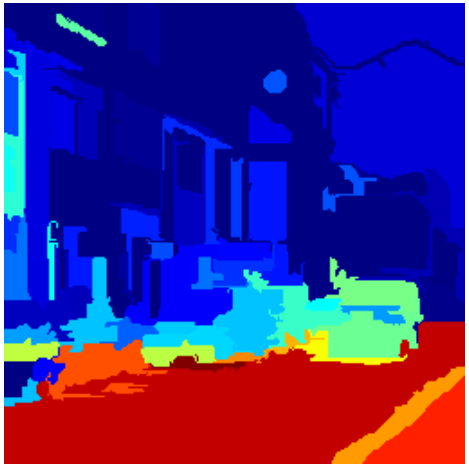
Building



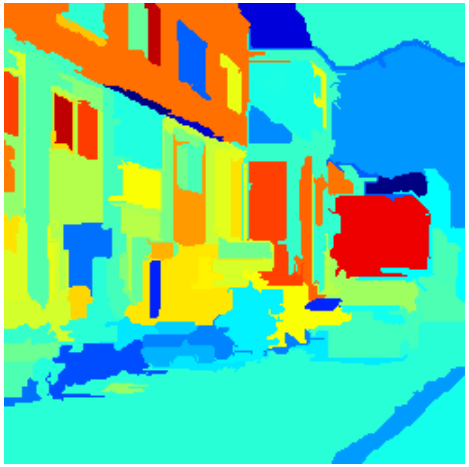
Car



Crosswalk

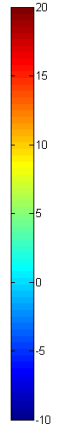


Road



Window

Sky



Step 3: Global image labeling

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

$$E(\boldsymbol{c}) = \sum_i -\log L(r_i, c_i) \quad + \quad \lambda \sum_{i,j} \delta[c_i \neq c_j]\, \varphi(c_i, c_j)$$

Vector of region labels

Regions

Likelihood score for region r_i and label c_i

Neighboring regions

Smoothing penalty

Co-occurrence penalty

Step 3: Global image labeling

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

$$E(\boldsymbol{c}) = \sum_i -\log L(r_i, c_i) + \lambda \sum_{i,j} \delta[c_i \neq c_j]\, \varphi(c_i, c_j)$$

Vector of region labels — $\boldsymbol{c}$

Regions — i

Likelihood score for region r_i and label c_i — $-\log L(r_i, c_i)$

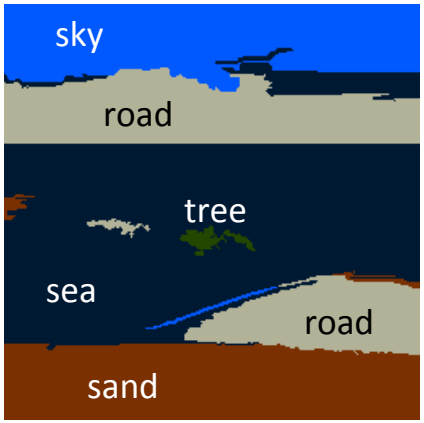
Neighboring regions — i, j

Smoothing penalty — $\delta[c_i \neq c_j]$

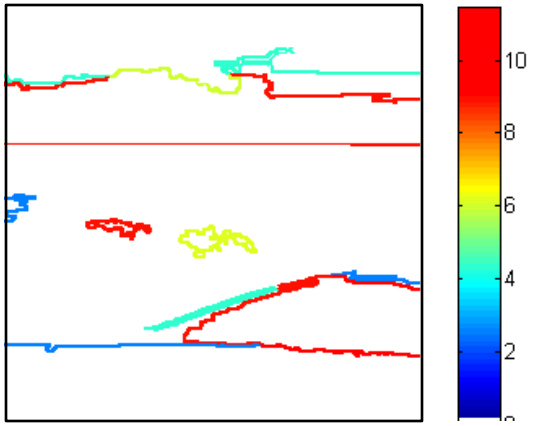
Co-occurrence penalty — $\varphi(c_i, c_j)$

Original image

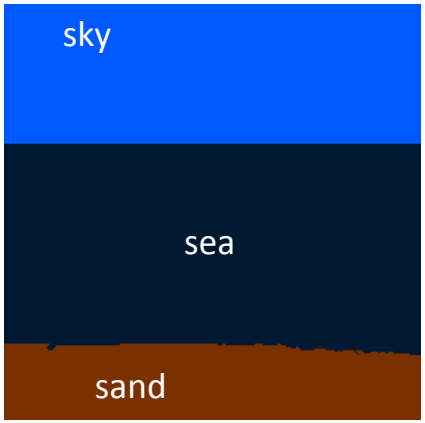
Maximum likelihood labeling



Edge penalties

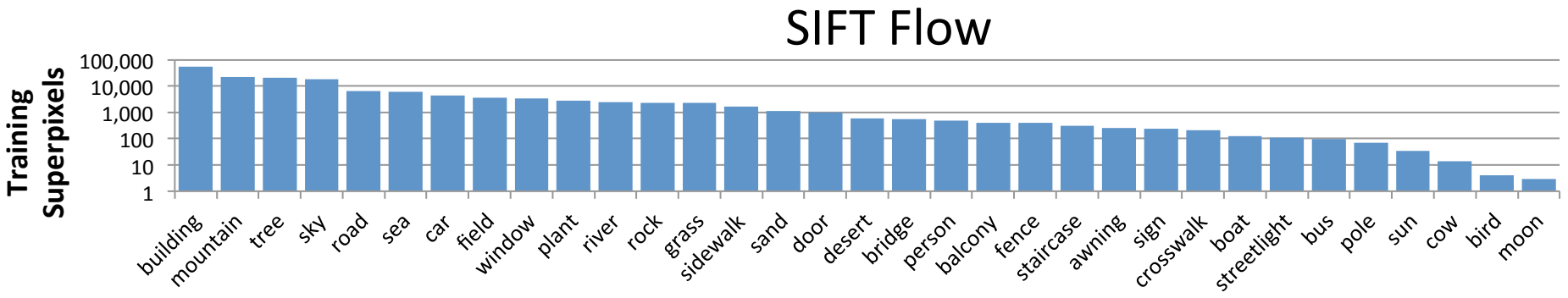


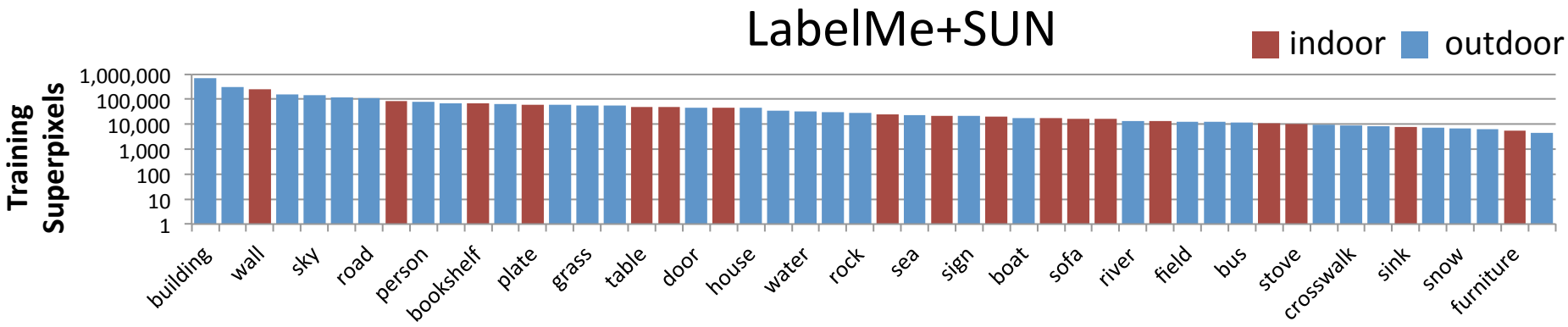
MRF labeling



Evaluation: Datasets

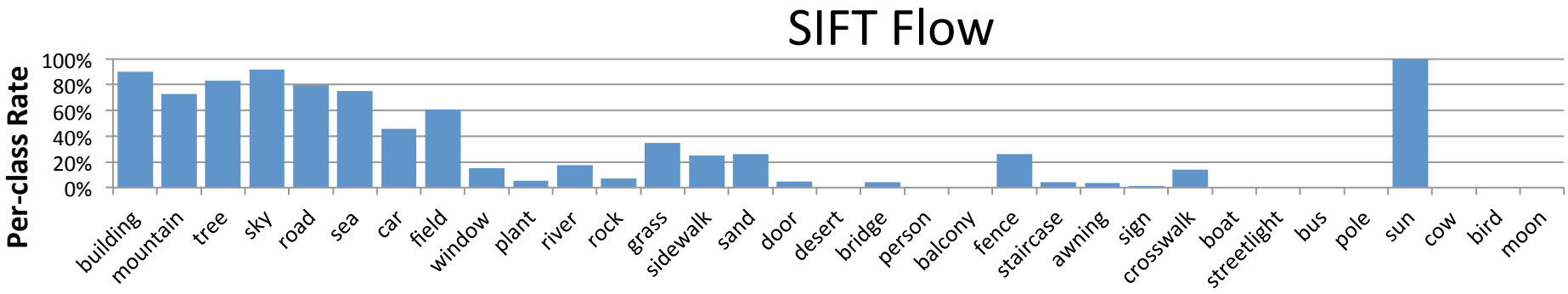
	Training Images	Test Images	Labels
SIFT Flow (Liu et al., 2009)	2,488	200	33
LabelMe+SUN	45,176	500	232

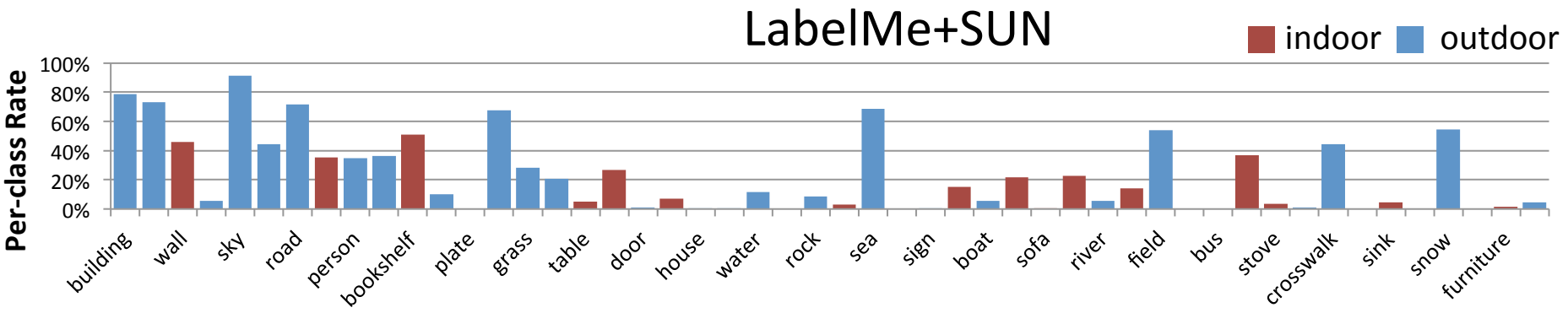




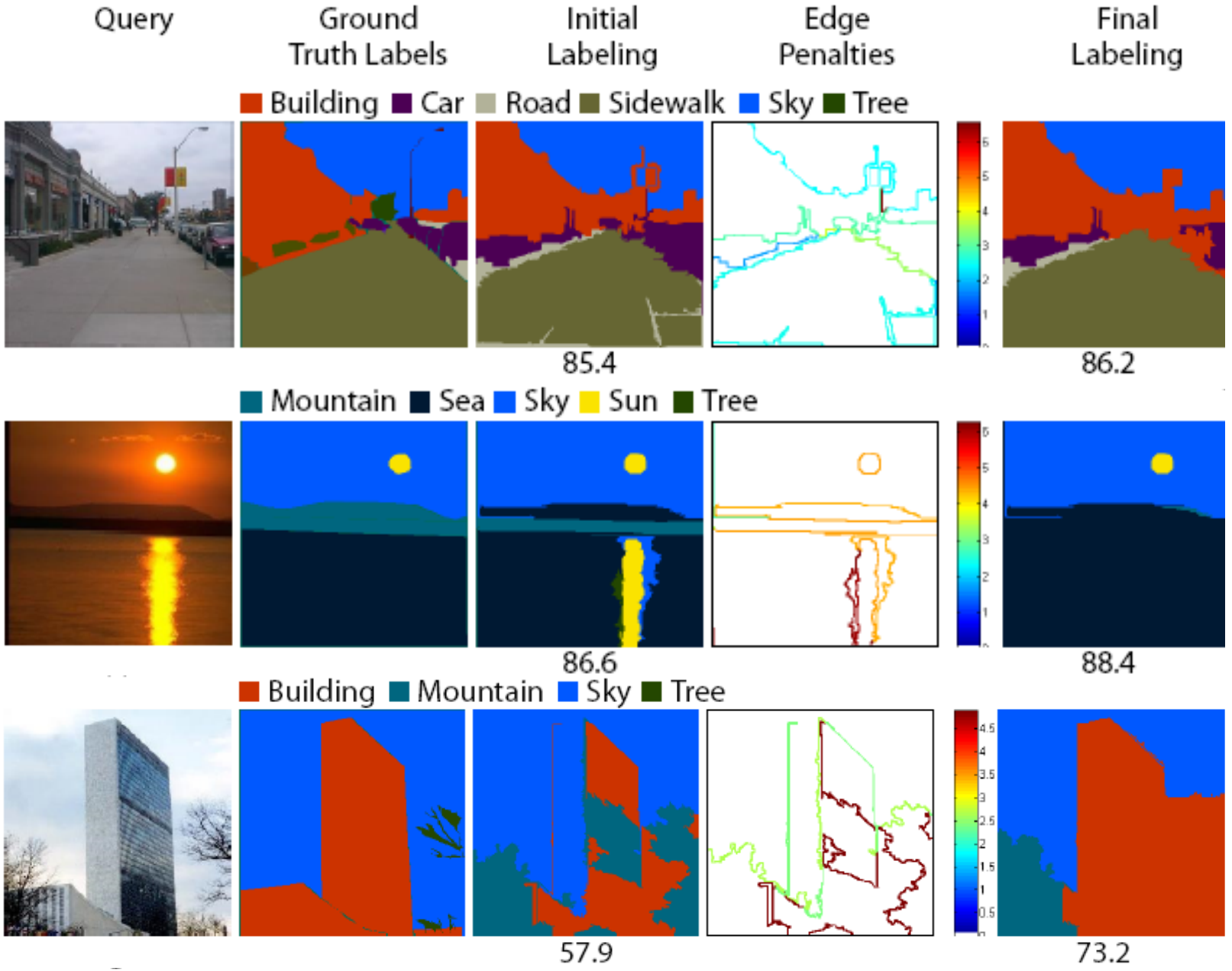
Evaluation: Performance

	Per-pixel (per-class)	
SIFT Flow (Liu et al., 2009)	77.0%	(30.1%)
LabelMe+SUN	54.9%	(7.1%)

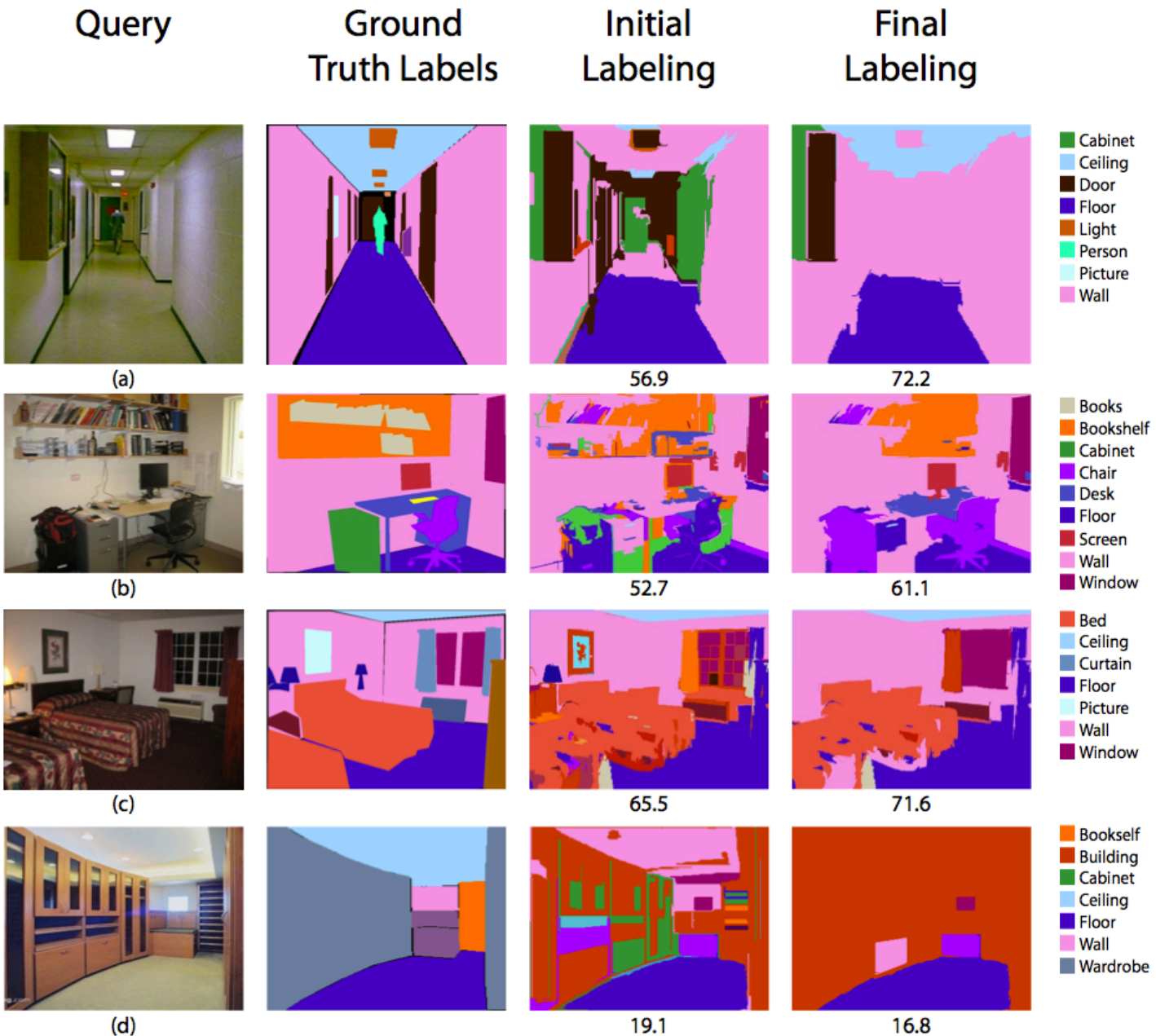




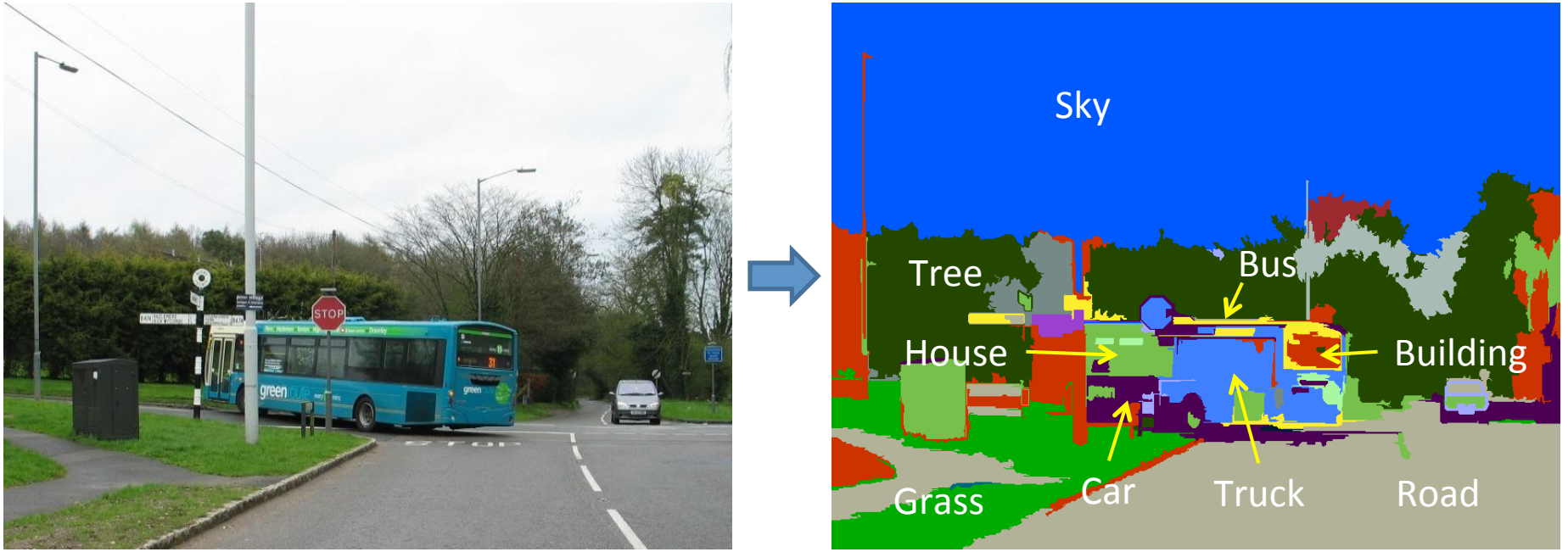
SIFT Flow Examples



LabelMe+SUN Examples

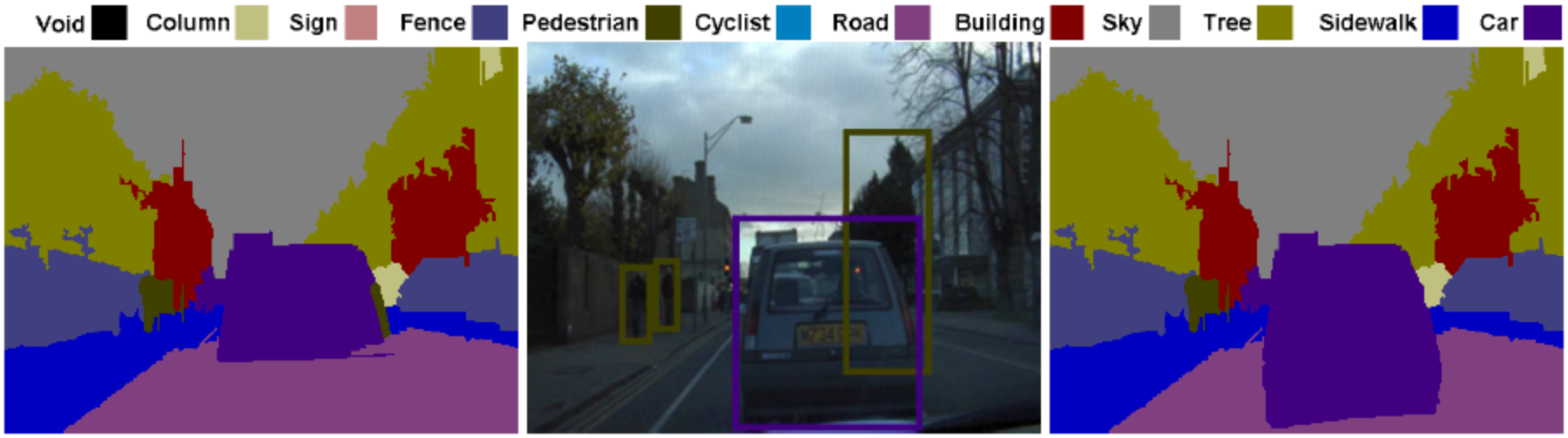


This work: Finding things



Finding Things: Image Parsing with Regions and Per-Exemplar Detectors
J. Tighe and S. Lazebnik, CVPR 2013

To get the things, use detectors



Result without detections

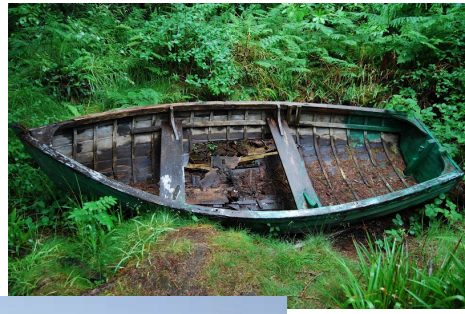
Set of detections

Final Result

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

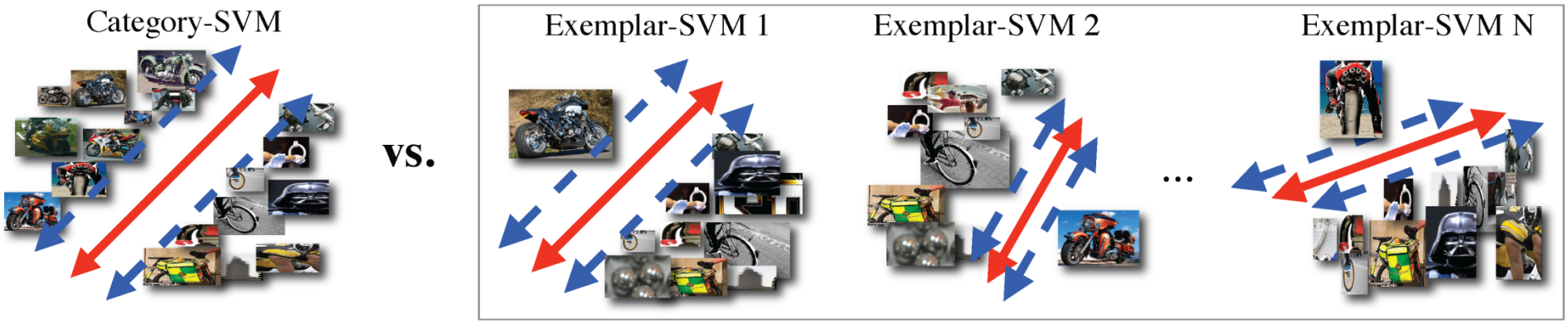
Problems with standard sliding window detectors

- They return only bounding box hypotheses, and obtaining segmentation hypotheses from them is challenging
- They do not work well for classes with few training examples and large intra-class variation

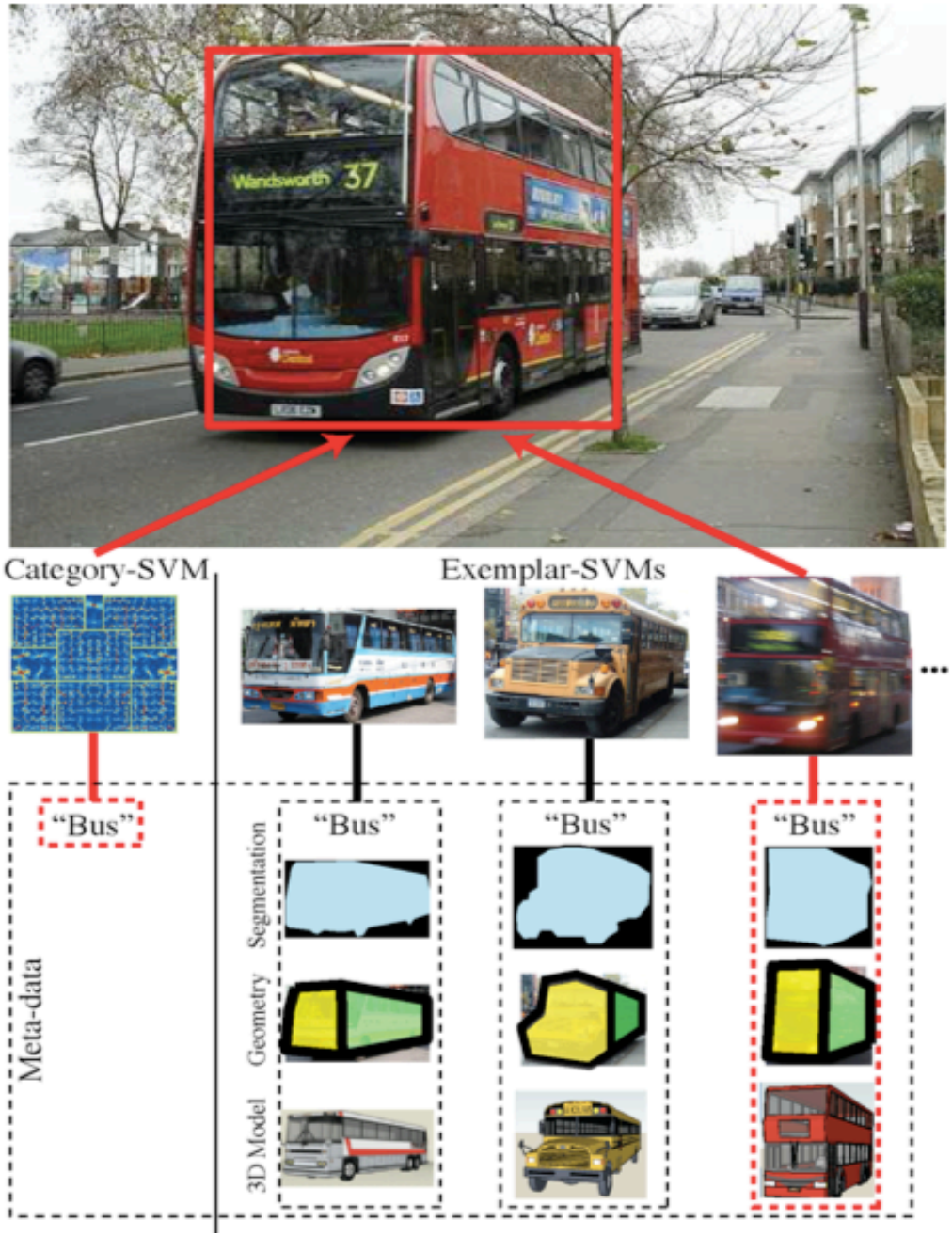




Per-exemplar detectors

- For each instance of a class: train a support vector machine based on HOG features
- Negative examples: all image windows 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

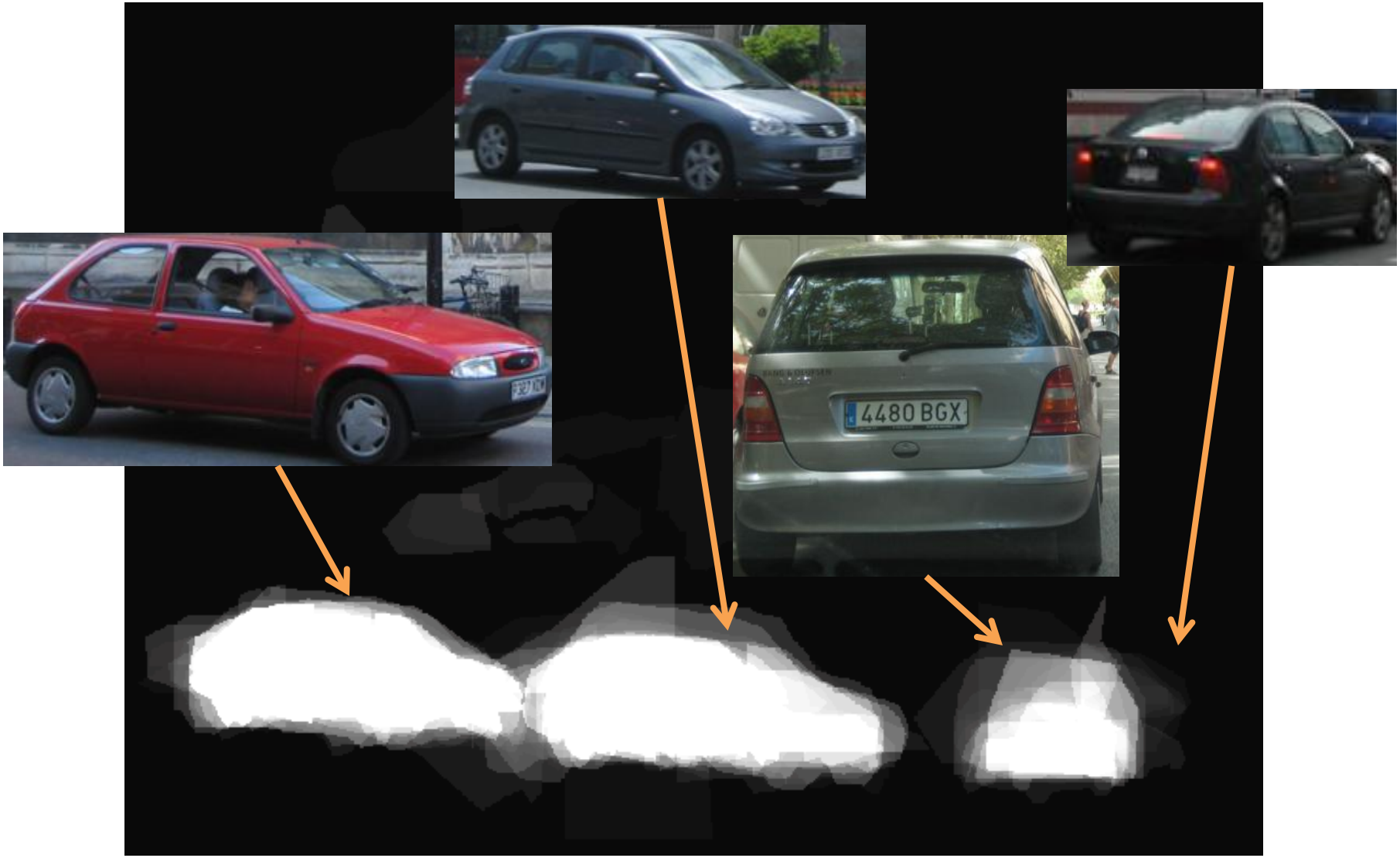


Tomasz Malisiewicz, Abhinav Gupta, Alexei A. Efros
Ensemble of Exemplar-SVMs for Object Detection and Beyond. In ICCV 2011

Our approach

Test image

Detector-based data term



From region to pixel terms

- $E\downarrow R\ (p\downarrow i\ ,c)=log\prod k\uparrow\blacksquare Pf\downarrow i\uparrow k\ c\ =\sum k\uparrow\blacksquare \log\ (Pf\downarrow i\uparrow k\ c\ )$
- $Pf\downarrow i\uparrow k\ c\ =\#(N(f\downarrow k\ (r\downarrow i\ )),c)/\#(D,c)$

- $E\downarrow D\ (p\downarrow i\ ,c)=\ \sum d\in D\uparrow\blacksquare (w\downarrow d-t\downarrow d\ )$
- $w\downarrow d$ is the detector score, D is the set of overlapping detection mask for pixel i and $t\downarrow d$ is the detection threshold (-1).

Detector and region term fusion

- After region term $E\downarrow R\ (p\downarrow i\ ,c)$ and detector term $E\downarrow D\ (p\downarrow i\ ,c)$ are computed for a dataset with C classes we have 2C values for each pixel.
- Predict final class training C 1-vs-rest SVM on the 2C values as features.
- Subsample dataset to make training feasible:
  - 67% of data by uniform sampling [may kill the long tail]
  - 33% of data by per class sampling [bias towards rare classes]
- Training is performed on 250,000 data points using linear SVM on approximate RBF embeddings.

Global image labeling

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

Likelihood score for region r_i and label c_i

$$E(\mathbf{c}) = \sum_i \max[0, M - SVM(p_i, c_i)]$$

$$+ \quad \lambda \sum_{i,j} \delta[c_i \neq c_j] \varphi(p_i, p_j)$$

Neighboring regions

Smoothing penalty

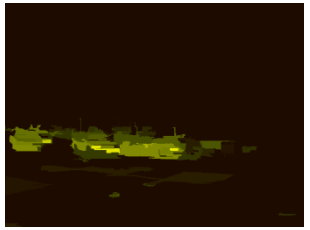
Query image

Ground truth

Query image

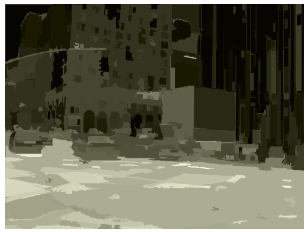
Ground truth

taxi
car
building
road
sky
fence
sidewalk
streetlight
truck
person
mailbox
van
window
trash can
manhole
traffic light



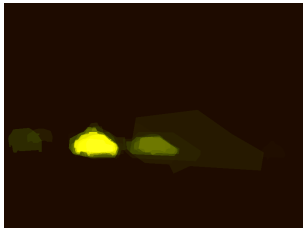
taxi

car



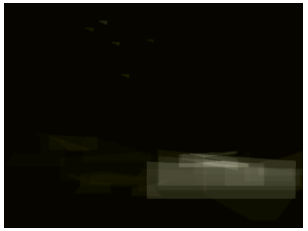
road

building

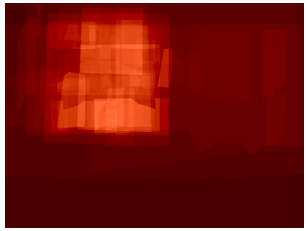


taxi

car

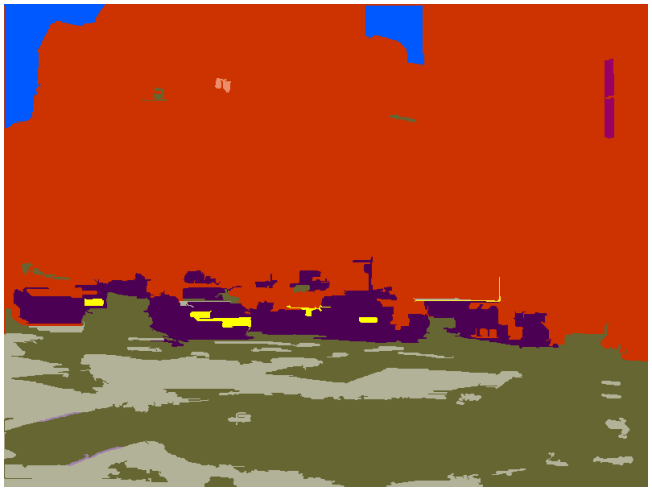


road



building

Region-based parsing result (67.2%)

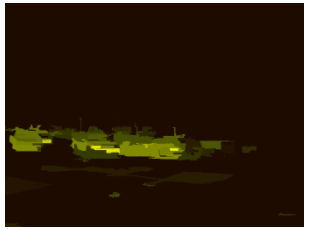


Detector-based parsing result (50.8%)

Query image

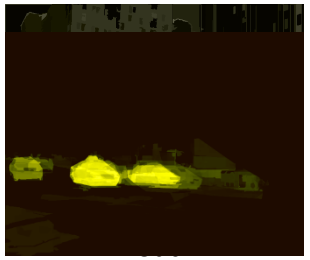
Ground truth

taxi
car
building
road
sky
fence
sidewalk
streetlight
truck
person
mailbox
van
window
trash can
manhole
traffic light



taxi

car



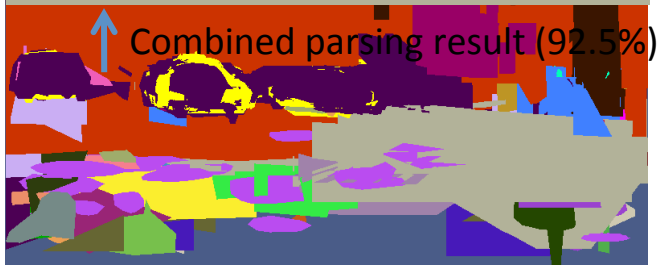
taxi

car

road

building

Region-based parsing result (67.2%)



Combined parsing result (92.5%)

Detector-based parsing result (50.8%)

Query image

Ground truth

Query image

Ground truth

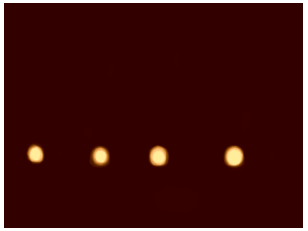
car
window
wheel
tree
building
road
sky
sidewalk
tail light
parking meter
headlight
door
fence
column
wall
sign
windshield

wheel

car

window

tree



wheel

car

window

tree

Region-based parsing result (59.7%)

Detector-based parsing result (31.6%)

Query image

Ground truth

wheel

car

wheel

car

window

tree

window

tree

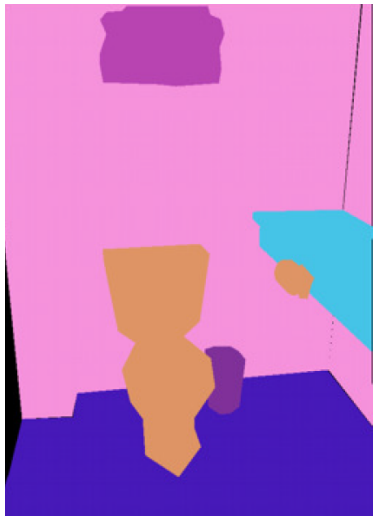
Region-based parsing result (59.7%)

Combined parsing result (77.3%)

Detector-based parsing result (31.6%)

car
window
wheel
tree
building
road
sky
sidewalk
tail light
parking meter
headlight
door
fence
column
wall
sign
windshield

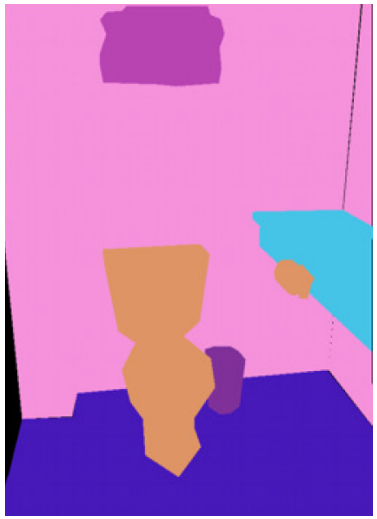
Query image



Ground truth

toilet
plate
wall
counter top
floor
mirror
person
pot
glass
cup
tree
painting
towel
trash can

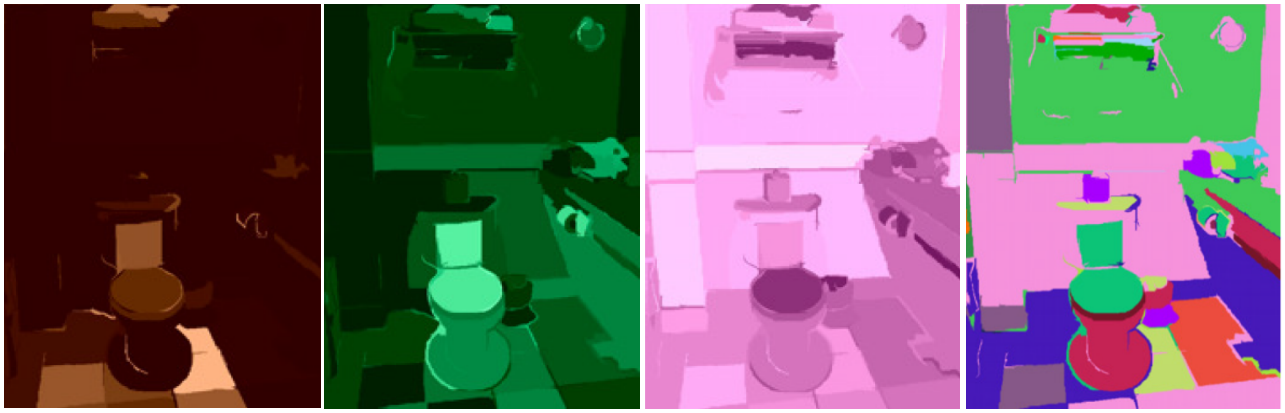
Query image



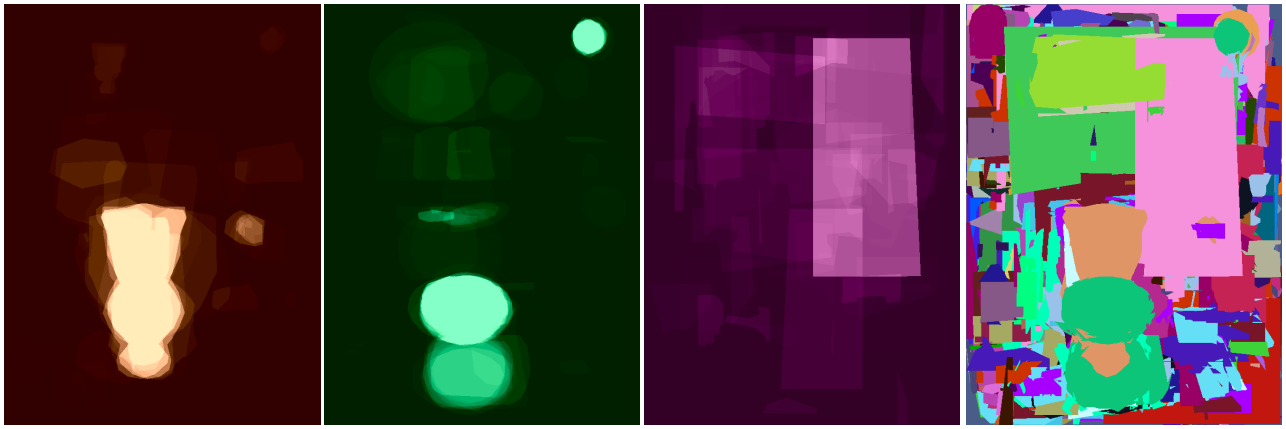
Ground truth

- toilet
- plate
- wall
- counter top
- floor
- mirror
- person
- pot
- glass
- cup
- tree
- painting
- towel
- trash can

Region-based parsing result (30.9%)



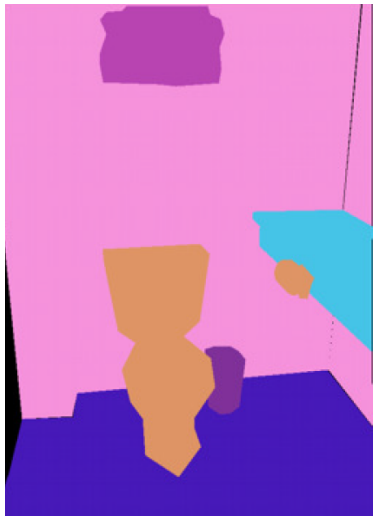
toilet plate wall



toilet plate wall

Detector-based parsing result (24.8%)

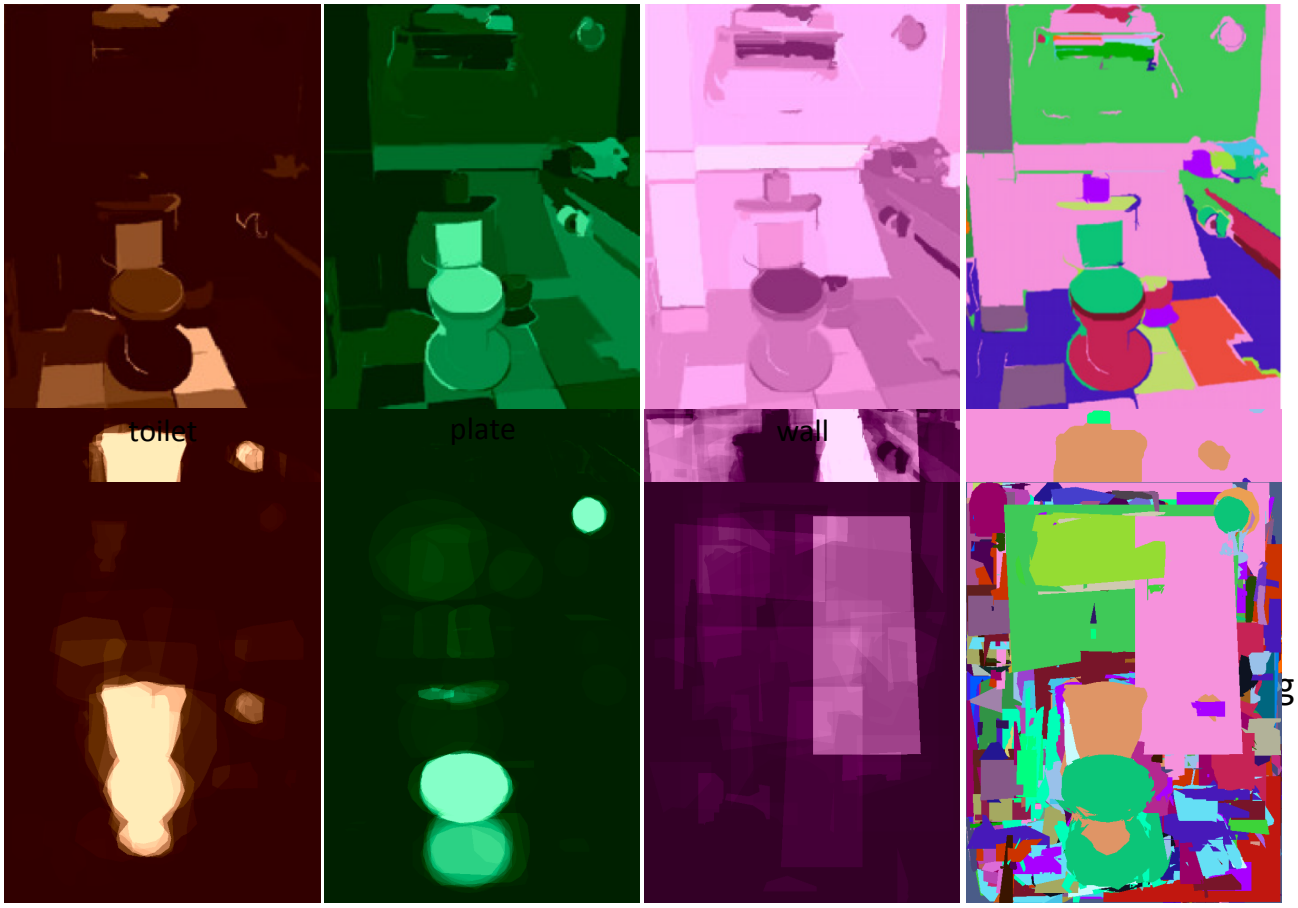
Query image



Ground truth

- toilet
- plate
- wall
- counter top
- floor
- mirror
- person
- pot
- glass
- cup
- tree
- painting
- towel
- trash can

Region-based parsing result (30.9%)

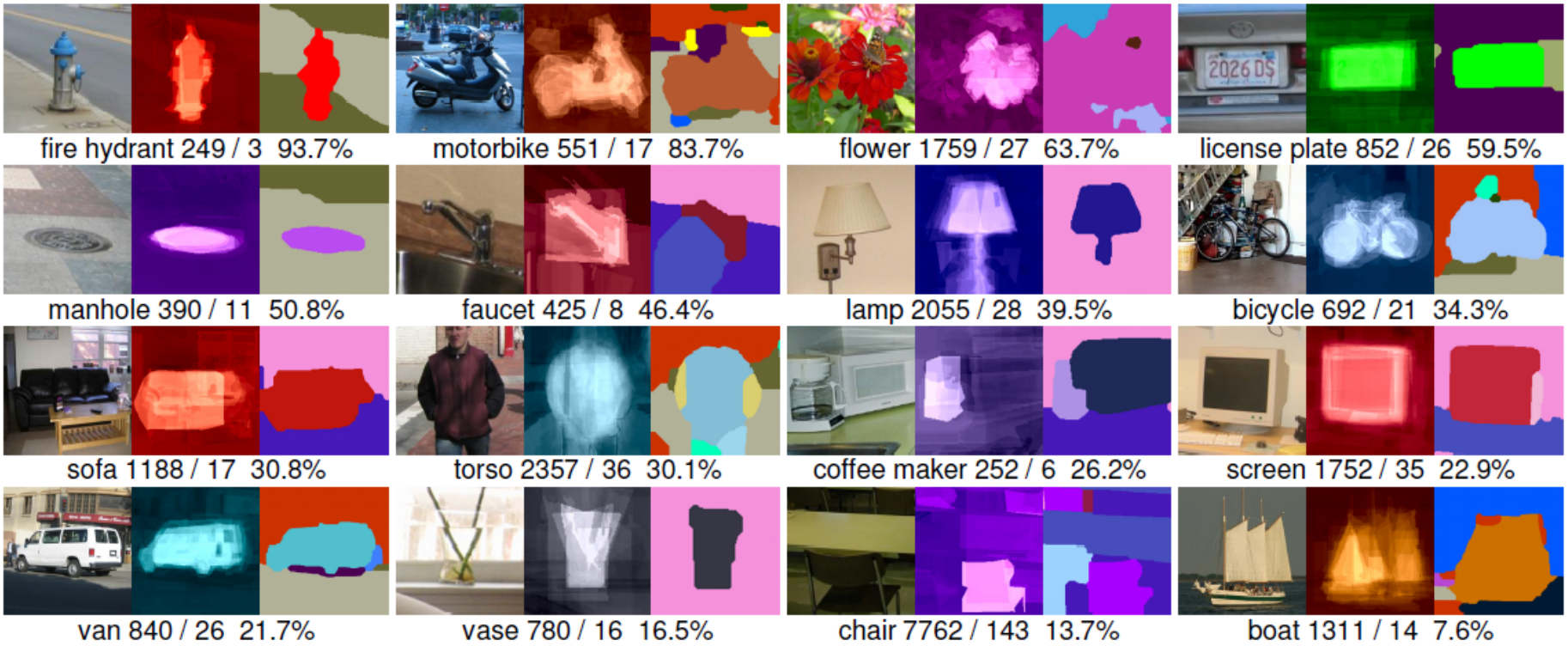


toilet plate wall

Detector-based parsing result (24.8%)

Quantitative evaluation

	Region-based	Detector-based	Region + Detector Combined
SIFT Flow (Liu et al., 2009)	77.7 (32.8)	71.1 (26.7)	**78.6 (39.2)**
LabelMe+SUN	58.3 (5.9)	52.5 (11.3)	**61.4 (15.2)**

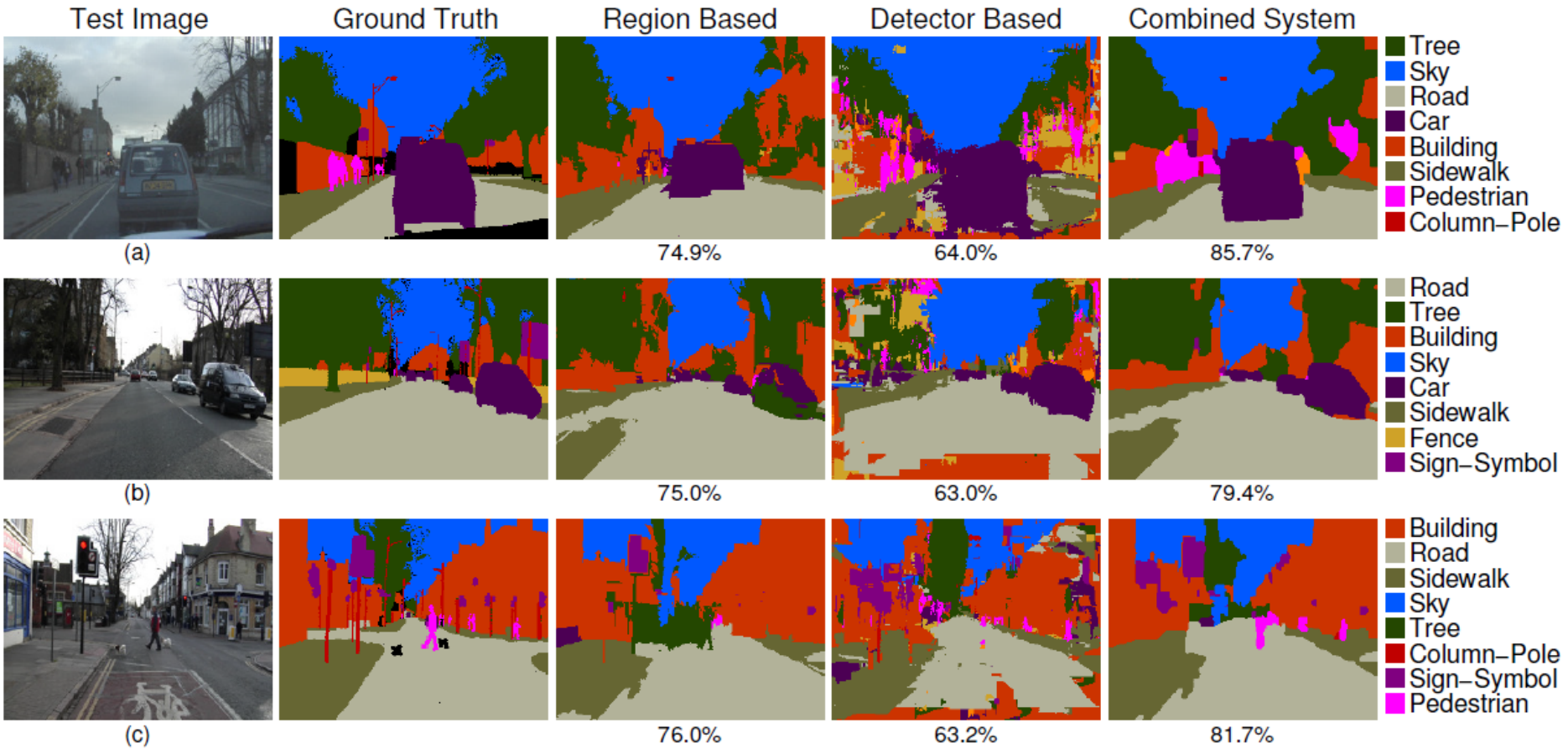


Comparison to state of the art

SIFT Flow	Per-Pixel	Per-Class
Our approach	78.6	39.2
Tighe and Lazebnik (2013)	77.0	30.1
Liu et al. (2011)	76.7	N/A
Farabet et al. (2012)	78.5	29.6
Farabet et al. balanced (2012)	74.2	46.0
Eigen and Fergus (2012)	77.1	32.5
Myeong et al. (2012)	77.1	32.3

LabelMe+SUN	Per-Pixel	Per-Class
Our approach	61.4	15.2
Outdoor	65.5	15.3
Indoor	46.3	12.2
Tighe and Lazebnik (2013)	54.9	7.1
Outdoor	60.8	7.7
Indoor	32.1	4.8

Video Parsing: CamVid dataset



Video Parsing: CamVid dataset

	Building	Tree	Sky	Car	Sign	Road	Pedestrian	Fence	Pole	Sidewalk	Bicyclist	Per-class	Per-pixel
Our approach	83.1	73.5	94.6	78.1	**48.0**	96.0	**58.6**	32.8	5.3	71.2	**45.9**	**62.5**	**83.9**
Tighe and Lazebnik (2013)	**87.0**	67.1	96.9	62.7	30.1	95.9	14.7	17.9	1.7	70.0	19.4	51.2	83.3
Brostow et al. (2008)	46.2	61.9	89.7	68.6	42.9	89.5	53.6	46.6	0.7	60.5	22.5	53.0	69.1
Sturgess et al. (2009)	84.5	72.6	**97.5**	72.7	34.1	95.3	34.2	45.7	8.1	77.6	28.5	59.2	83.8
Zhang et al. (2010)	85.3	57.3	95.4	69.2	46.5	**98.5**	23.8	44.3	**22.0**	38.1	28.7	55.4	82.1
Floros et al. (2011)	80.4	76.1	96.1	86.7	20.4	95.1	47.1	47.3	8.3	79.1	19.5	59.6	83.2
Ladicky et al. (2010)	81.5	**76.6**	96.2	**78.7**	40.2	93.9	43.0	**47.6**	14.3	**81.5**	33.9	**62.5**	83.8

Comparison of different data terms

Is it just the SVM over the responses to improve results?

	SIFT Flow		LM+Sun		CamVid	
	Per-Pixel	Per-Class	Per-Pixel	Per-Class	Per-Pixel	Per-Class
Detector ML	65.1	25.8	33.0	14.1	61.2	45.5
Detector SVM	62.5	25.4	46.1	12.0	61.4	47.0
Detector SVM MRF	71.1	26.7	52.5	11.3	63.8	47.3
Region ML	74.1	30.2	51.5	7.5	82.7	51.2
Region SVM	75.0	35.9	56.3	6.7	81.4	55.7
Region SVM MRF	77.7	32.8	58.3	5.9	83.5	55.7
Region + Thing SVM	74.4	36.9	58.5	14.1	82.4	60.0
Region + Thing SVM MRF	77.5	35.7	60.0	12.9	84.2	59.5
Combined	75.6	**41.1**	59.6	**15.5**	82.3	62.1
Combined MRF	**78.6**	39.2	**61.4**	15.2	**84.0**	**62.2**

- ML: assign class with maximum nonparametric likelihood
- SVM: predict class given class likelihoods
- Things SVM: exemplars trained only on THINGS: car,boat,person

Now what?

- Code and data publicly available on author websites
  - Other researchers should push for bigger datasets, broader coverage
- Lots more work to do
  - Improve computational efficiency of exemplar SVM training: try whitened HOG approach of Hariharan et al. (ECCV 2012)
  - Use adaptive context to decide which exemplar SVMs to run on a given test image
  - Make approach completely *open-universe*: eliminate reliance on batch offline training of SVM

Future work: Image description

A girl on rollerskates is talking on her cell phone while standing in a parking lot.

Closest matches based on global features

Future work: Image description

Two men, one in a gray shirt, one in a black shirt, are standing near a stove.

Closest matches based on global features

Comparison of different SVM kernels

	SIFT Flow		LM+Sun		CamVid	
	Per-Pixel	Per-Class	Per-Pixel	Per-Class	Per-Pixel	Per-Class
Linear	75.4	40.0	57.2	16.6	82.4	60.7
Linear MRF	77.5	40.2	59.5	15.9	83.8	60.7
Approx. RBF	75.6	41.1	59.6	15.5	82.3	62.1
Approx. RBF MRF	78.6	39.2	61.4	15.2	83.9	62.5
Exact RBF	75.4	41.6	N/A	N/A	82.3	61.9
Exact RBF MRF	77.6	42.0	N/A	N/A	84.0	62.2

Comparison to state of the art

SIFT Flow	Per-Pixel	Per-Class
Ours: Combined MRF	78.6	39.2
Tighe and Lazebnik [27]	77.0	30.1
Liu et al. [17]	76.7	
Farabet et al. [6]	78.5	29.6
Farabet et al. [6] balanced	74.2	46.0
Eigen and Fergus [5]	77.1	32.5
Myeong et al. [21]	77.1	32.3

LM+SUN	Per-Pixel	Per-Class
Ours: Combined MRF	61.4	15.2
Outdoor Images	65.5	15.3
Indoor Images	46.3	12.2
Tighe and Lazebnik [27]	54.9	7.1
Outdoor Images	60.8	7.7
Indoor Images	32.1	4.8