# * Semantic Segmentation

Scene Understanding

Bebe Shi

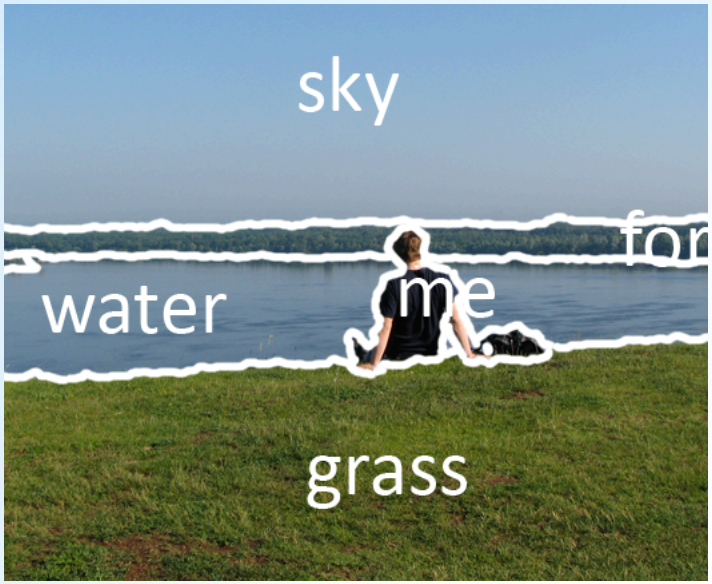
PRINCETON UNIVERSITY

* Assign a class label to each pixel of an image

# * Formalized Task

PRINCETON UNIVERSITY

Semantic Segmentation




Object Detection


# *Comparisons

PRINCETON UNIVERSITY

*http://people.csail.mit.edu/jxiao/StreetSeg/

* J. Xiao and L. Quan Multiple View Semantic Segmentation for Street View Images. Proceeding of 12th IEEE International Conference on Computer Vision (ICCV2009)

PRINCETON UNIVERSITY

# *Difficulties

* Ambiguities
* Arbitrary viewpoints
* Severe occlusion

PRINCETON UNIVERSITY

* Texton: feature that jointly models shape appearance and context
* Focus on speed and larger datasets

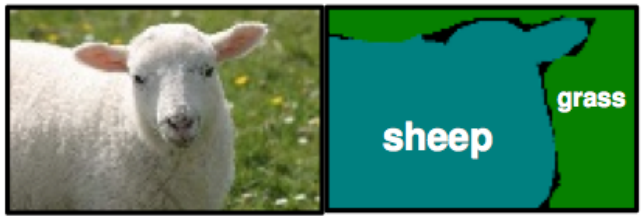
* *J.Shotton,J.Winn,C.Rother,andA.Criminisi.Text onBoost: Joint Appearance, Shape and Context Modeling for Multi-class Object Recognition and Segmentation. ECCV 2006*

# *TextonBoost

PRINCETON UNIVERSITY

* 591 photos of 21 object classes: building, grass, tree, cow, sheep, sky, aeroplane, water, face, car, bike, flower, sign, bird, book, chair, road, cat, dog, body, boat
* All hand-labeled
* All images are approximately 320 × 240 pixels.
* 45% training, 10% validation and 45% test sets






# * Datasets

PRINCETON UNIVERSITY

# *Datasets

- For comparison with previous work:
- 7-class Corel database (180×120 pixels)
- 7-class Sowerby database (96 × 64 pixels)

- LabelMe Outdoor (LMO) database
- SUN database

PRINCETON UNIVERSITY

* *J. Shotton, M. Johnson, R. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation*
* weight each training example by the inverse class frequency

$$w_i = \xi_{c_i} \text{ with } \xi_c = \left(\sum_{i \in I} [c = c_i]\right)^{-1}$$

# *Datasets

$$\log P(\mathbf{c}|\mathbf{x},\boldsymbol{\theta}) = \sum_i \overbrace{\psi_i(c_i,\mathbf{x};\boldsymbol{\theta}_\psi)}^{\text{shape-texture}} + \overbrace{\pi(c_i,\mathbf{x}_i;\boldsymbol{\theta}_\pi)}^{\text{color}} + \overbrace{\lambda(c_i,i;\boldsymbol{\theta}_\lambda)}^{\text{location}}$$

$$+ \sum_{(i,j)\in\mathcal{E}} \overbrace{\phi(c_i,c_j,\mathbf{g}_{ij}(\mathbf{x});\boldsymbol{\theta}_\phi)}^{\text{edge}} - \log Z(\boldsymbol{\theta},\mathbf{x})$$

# *Conditional Random Field (CRF)

PRINCETON UNIVERSITY

# Learning the CRF Parameters

* Ideal case: maximize the conditional likelihood by using gradient ascent
* Piecewise training

# *Learning the CRF Parameters

* *Sutton, C., McCallum, A.: Piecewise training of undirected models. In: 21st Conference on Uncertainty in Artificial Intelligence. (2005)*
* Overcounting
* Added power parameters for the location and color potentials

PRINCETON UNIVERSITY

*Boykov, Y., Jolly, M.P.: Interactive graph cuts for optimal boundary and region segmentation of objects in n-d images. Proc. of IEEE ICCV. (2001)

# *Finding c*

PRINCETON UNIVERSITY

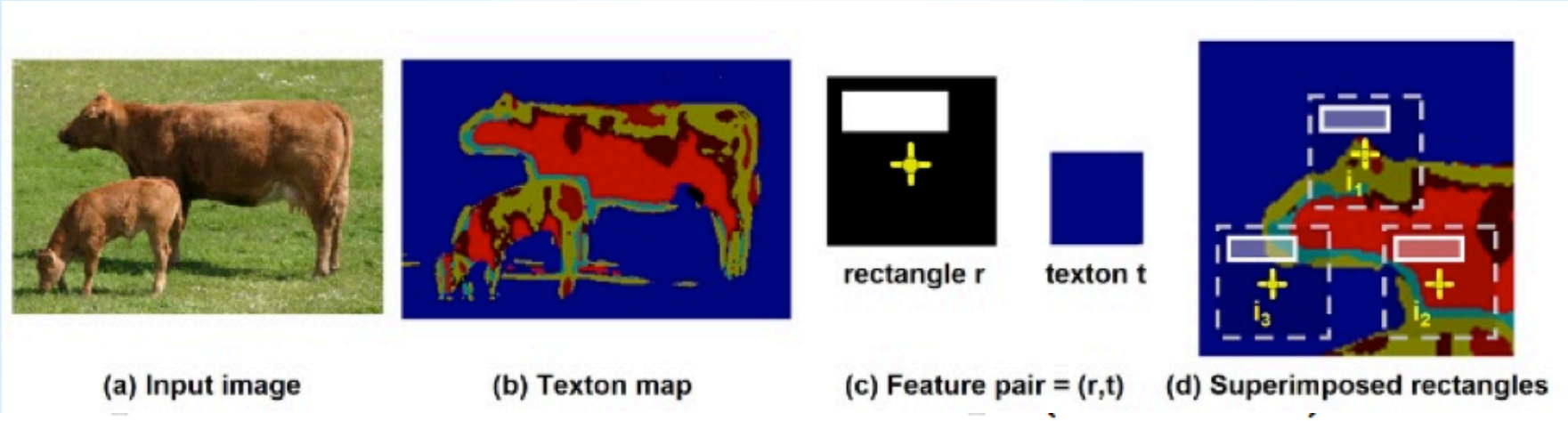
# *Boosted Learning of Shape, Texture and Context

PRINCETON UNIVERSITY

# Textons

* *Varma, M., Zisserman, A.: A statistical approach to texture classification from single images. International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis 62 (2005) 61-81*

* Texton Map




$$v(i_1, r, t) \approx A,\ v(i_2, r, t) \approx 0,\ \text{and}\ v(i_3, r, t) \approx A/2.$$

# *Shape Filters

PRINCETON UNIVERSITY

* *Torralba, A., Murphy, K., Freeman, W.: Sharing features: efficient boosting procedures for multiclass object detection. Proc. Of IEEE CVPR (2004) 762-769*

# * Joint Boosting Algorithm

$$h(c_i) = \begin{cases} a\delta(v(i,r,t) > \theta) + b & \text{if } c_i \in N \\ k_{c_i} & \text{otherwise} \end{cases}$$

Weak Classifier

$$H(c_i) = \sum_{m=1}^{M} h_m(c_i)$$

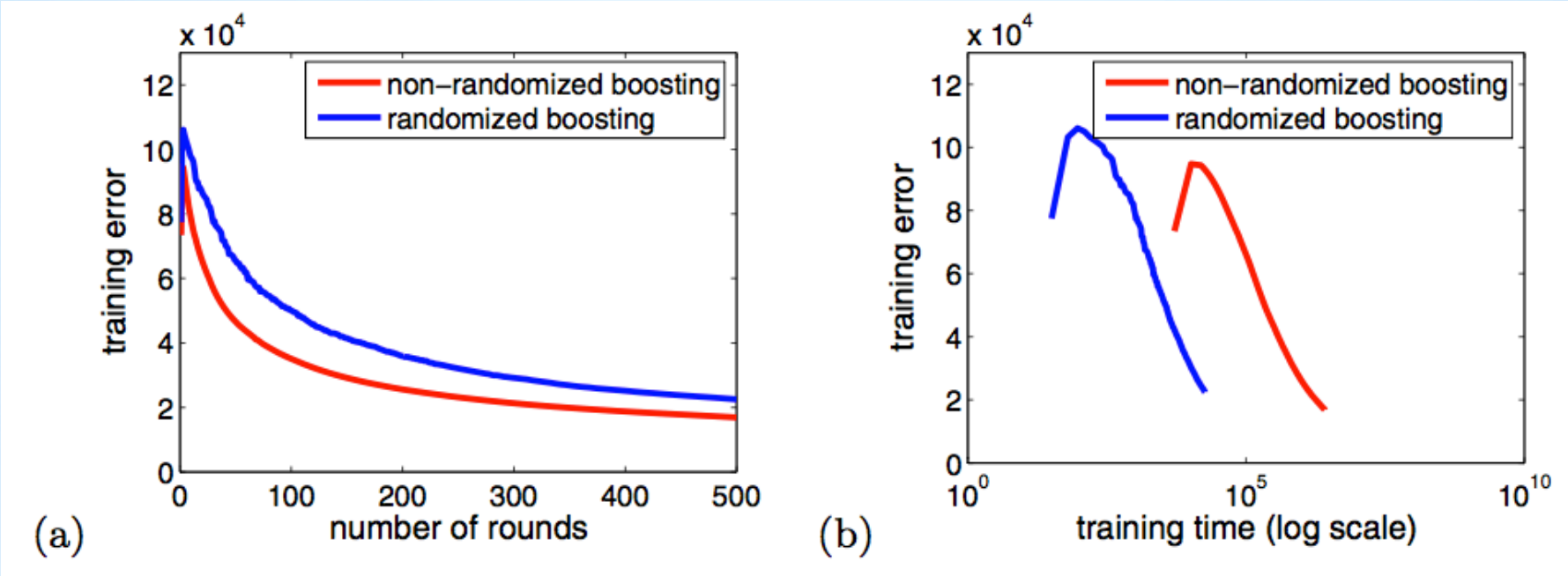
$$\widetilde{P}_i(c_i|\mathbf{x}) = \frac{\exp(H(c_i))}{\sum_{c'_i} \exp(H(c'_i))}$$

Softmax Transformation

# *Joint Boosting

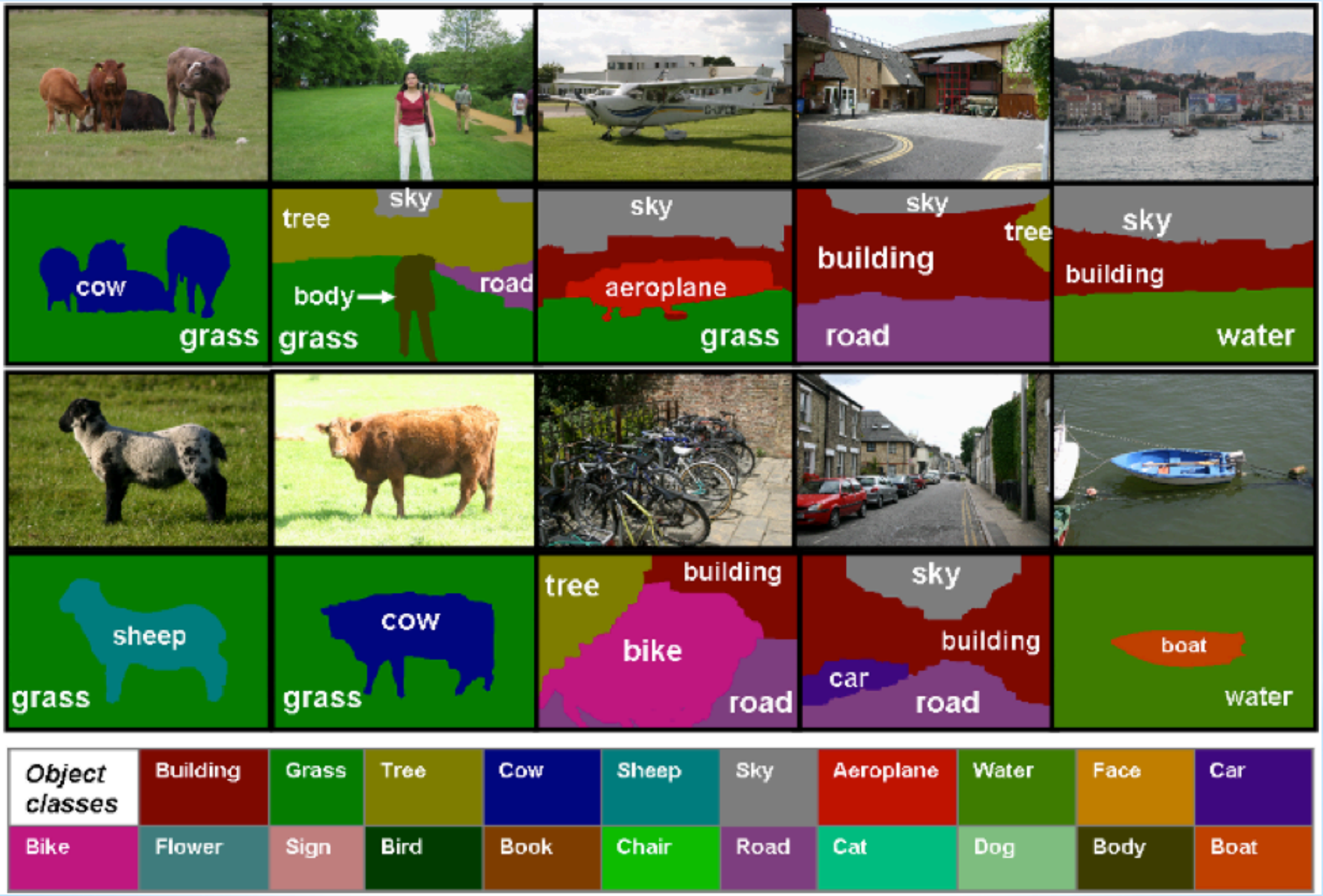
# Sub-Sampling and Random Feature Selection

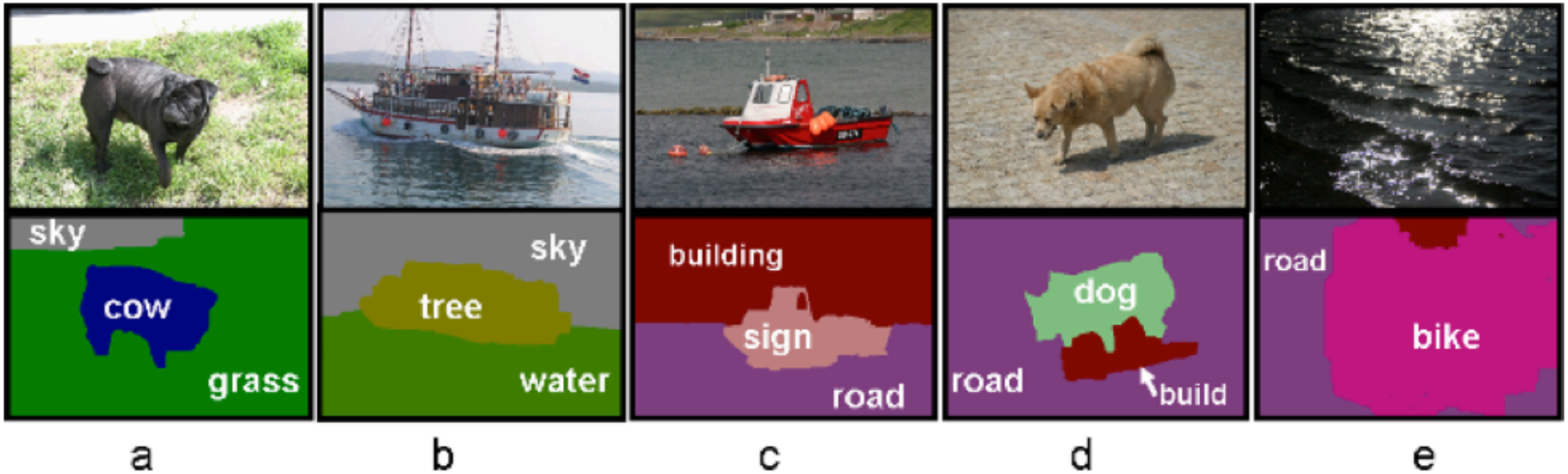
- Filter responses are calculated on a $\Delta \times \Delta$ grid
- Shape filters are still calculated at full resolution at test time
- At each round of boosting, a fraction of $\tau << 1$ of the features are chosen




# *Sub-Sampling and Random Feature Selection

PRINCETON UNIVERSITY




sky
cow
grass
sky
tree
water
building
sign
road
dog
road
build
road
bike
a
b
c
d
e

PRINCETON UNIVERSITY

# Semantic Texton Forests

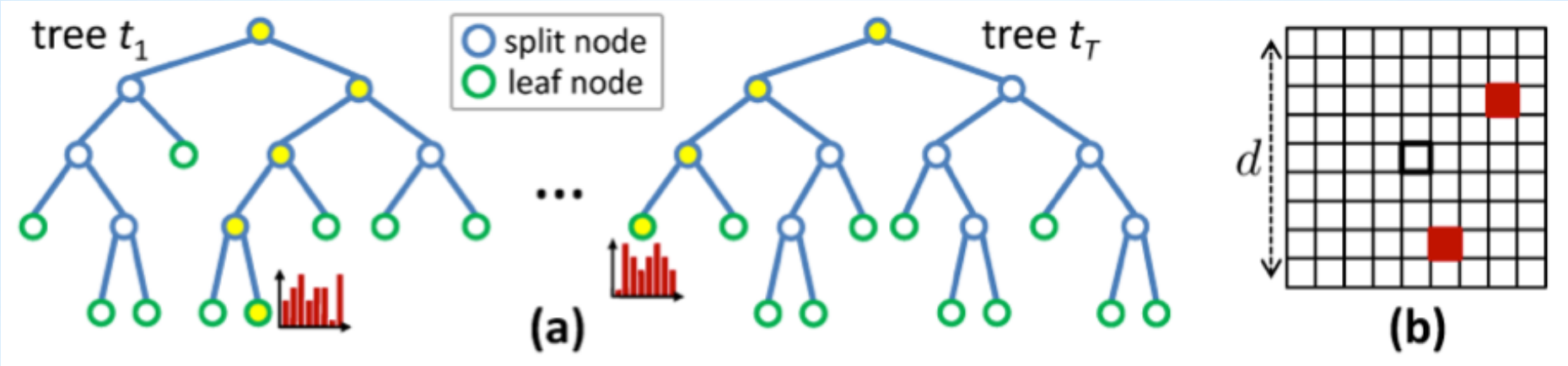
* *J. Shotton, M. Johnson, R. Cipolla. Semantic Texton Forests for Image Categorization and Segmentation. CVPR 2008*

* Extremely fast to both train and test
* Improves both quantitative performance and execution speed

# Randomized Decision Forests

* Ensemble of T decision trees
* Each node n is a learned class distribution P(c|n)
* Recursively branches left or right down the tree according to a learned binary function of the feature vector, until a leaf node l is reached.

$$P(c|L) = \frac{1}{T}\sum_{t=1}^{T} P(c|l_t)$$


tree $t_1$
split node
leaf node
tree $t_T$
...
(a)
$d$
(b)

# *Randomized Learning

- **Extremely Randomized Trees Algorithm**
- Each tree is trained separately on a small random subset of the training data

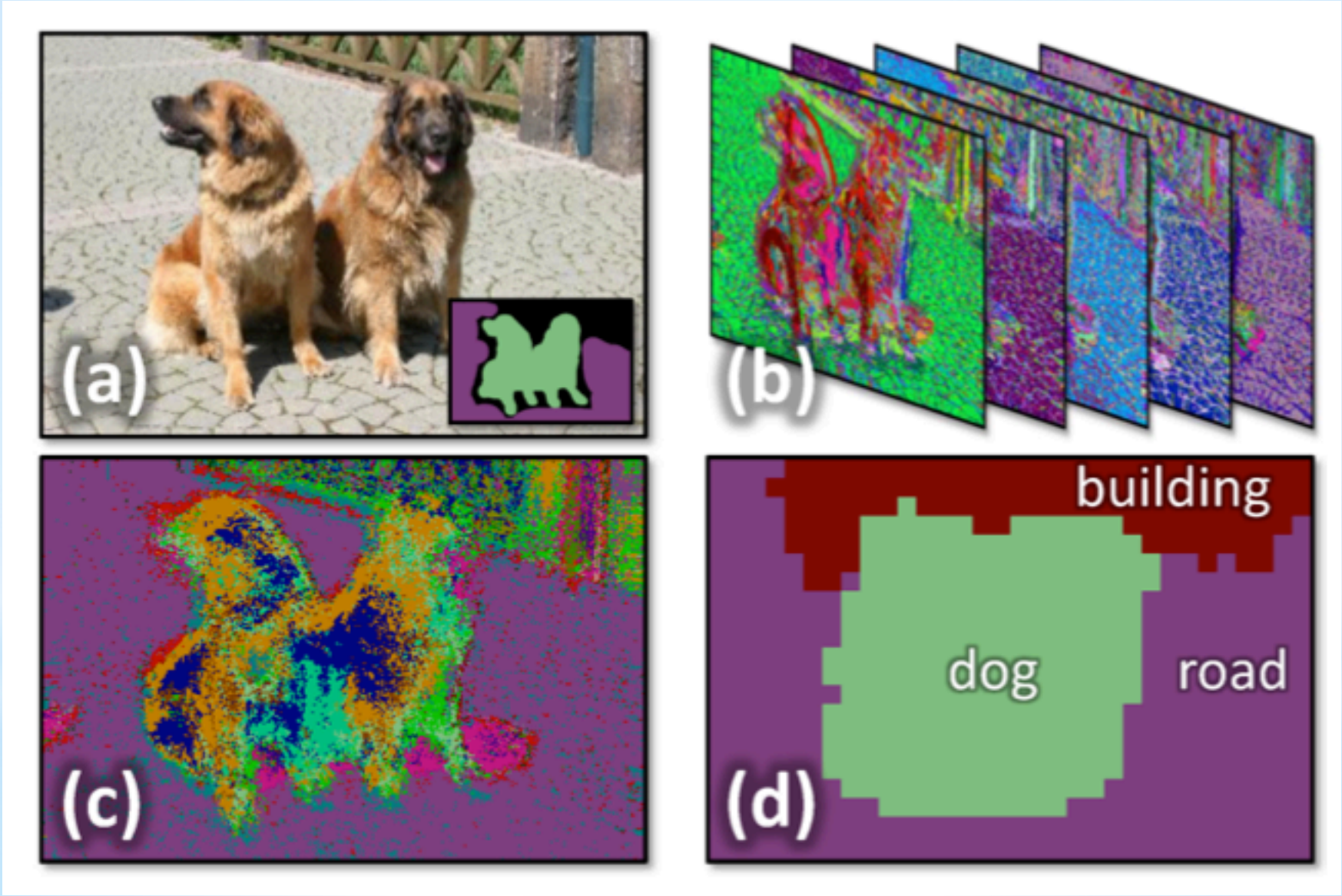
$$
\begin{aligned}
I_{\mathrm{l}} &= \{i \in I_n \mid f(\mathbf{v}_i) < t\} \\
I_{\mathrm{r}} &= I_n \setminus I_{\mathrm{l}} \,.
\end{aligned}
$$

$$
\Delta E = -\frac{|I_{\mathrm{l}}|}{|I_n|} E(I_{\mathrm{l}}) - \frac{|I_{\mathrm{r}}|}{|I_n|} E(I_{\mathrm{r}})
$$

Shannon Entropy

# Advantages

- Small random subsets of the data are used for training
- Only a small portion of the tree is traversed for each data point

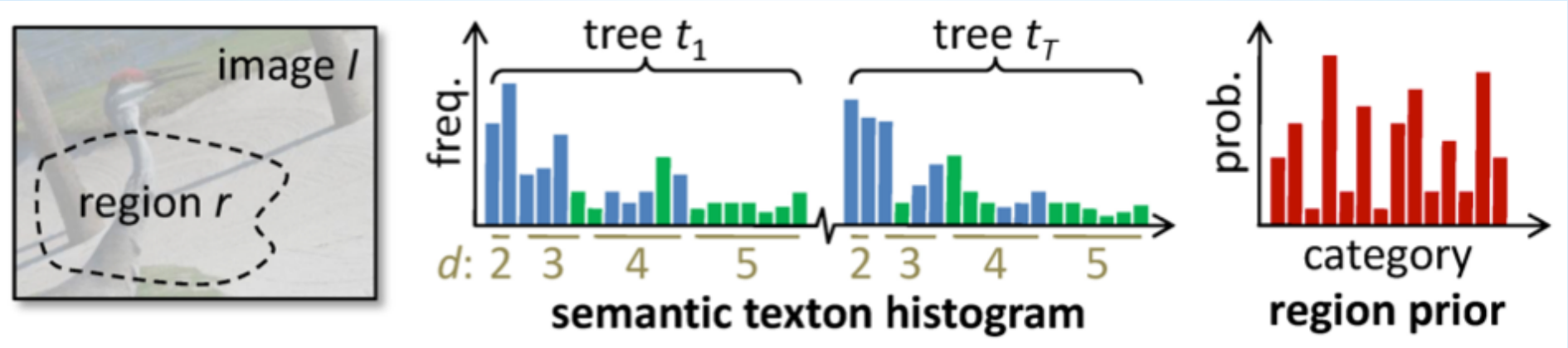

(a)
(b)
(c)
(d)
building
dog
road

PRINCETON UNIVERSITY

# *Learning Invariances

* Augment the training data with image copies that are artificially transformed geometrically and photometrically.

* Explored rotation, scaling and left-right flipping as geometric transformations and affine photometric transformations.




# *Bags of Semantic Textons (BoSTs)

PRINCETON UNIVERSITY

# Nonparametric Scene Parsing

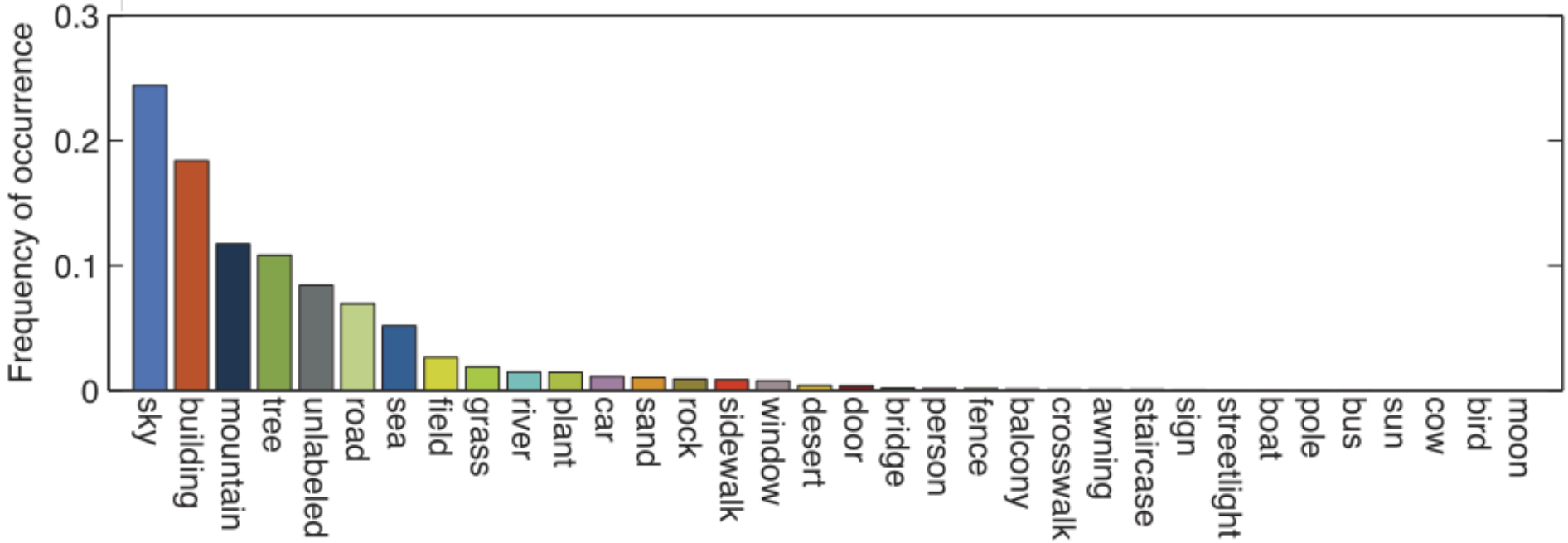
- *C. Liu, J. Yuen, A. Torralba. Nonparametric Scene Parsing via Label Transfer. IEEE VOL 33, NO. 12 (2011)*
- Easy to implement
- Has few parameters
- Embeds contexual information naturally in the retrieval/alignment procedure
- Transfer the labels from existing samples in a large database to annotate an image

# *Datasets

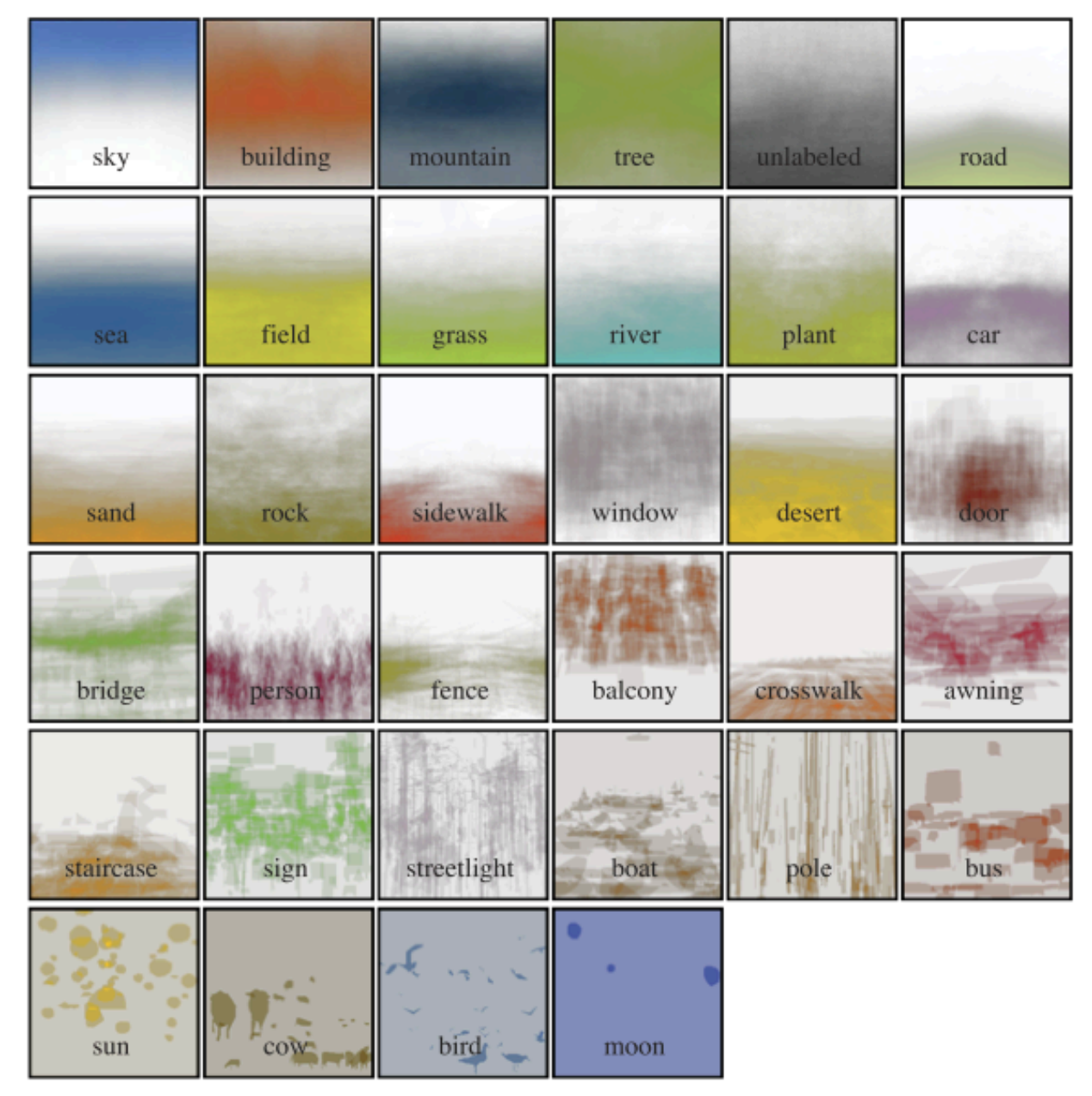
* LabelMe Ourdoor (LMO) Database (2,688 fully annotated images, mostly outdoor scenes)
* Sun Database (9,566 fully annotated images, covers both indoor and outdoor scenes)
* Both annotated using the LabelMe online annotation tool http://new-labelme.csail.mit.edu/Release3.0/index.php

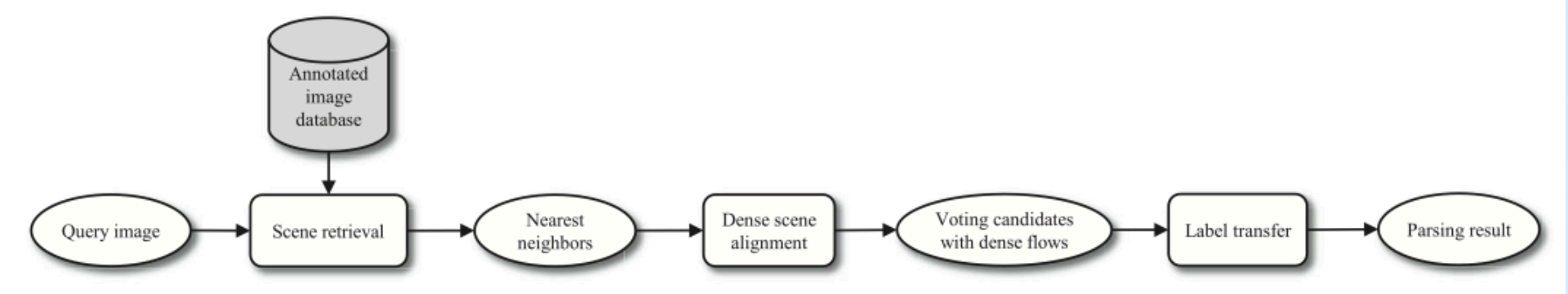
PRINCETON UNIVERSITY

- 2,488 for training and 200 for testing
- Top 33 object categories with the most labeled pixels are chosen




# LMO database


sky
building
mountain
tree
unlabeled
road
sea
field
grass
river
plant
car
sand
rock
sidewalk
window
desert
door
bridge
person
fence
balcony
crosswalk
awning
staircase
sign
streetlight
boat
pole
bus
sun
cow
bird
moon




# *System Pipeline

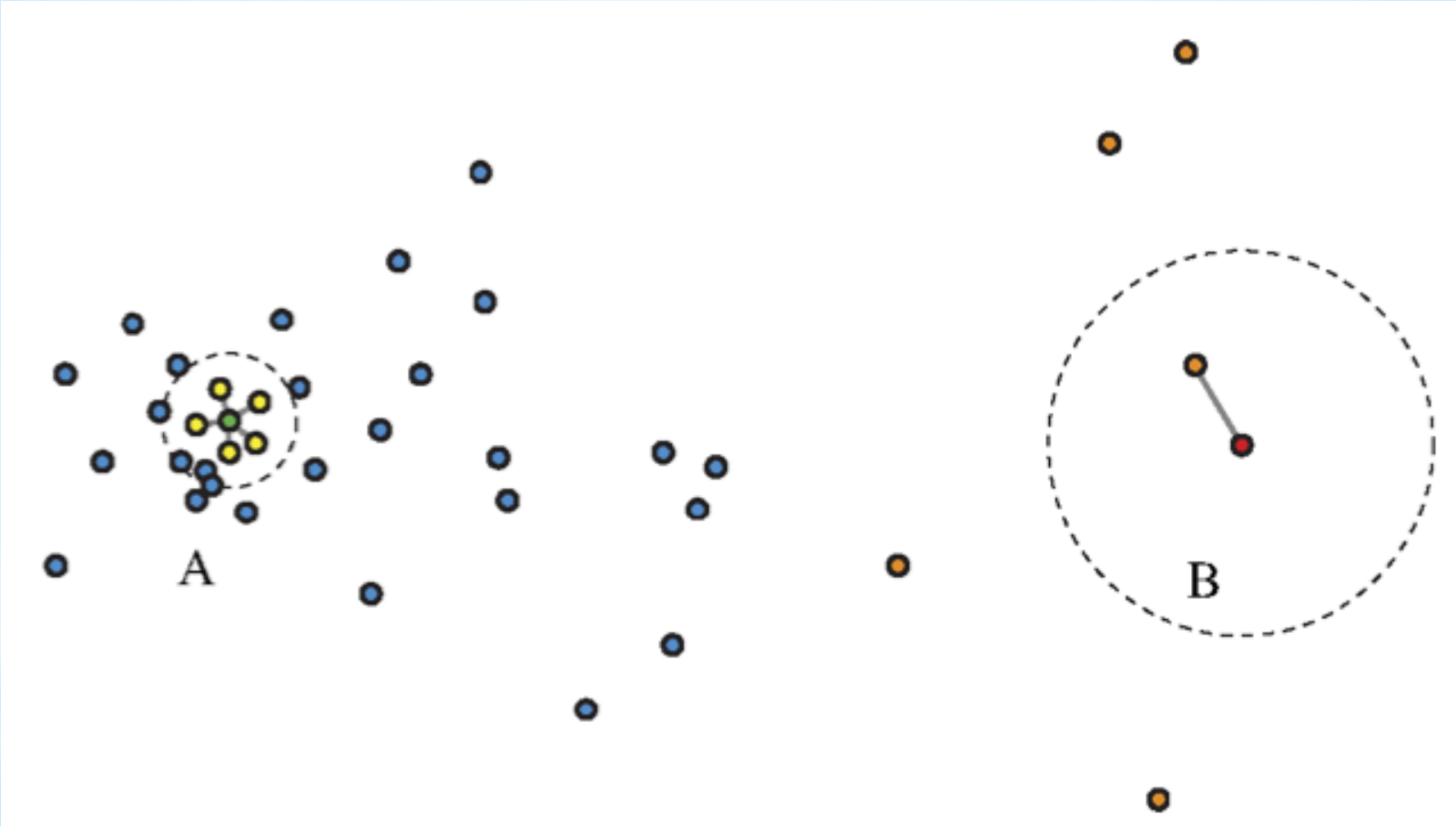
PRINCETON UNIVERSITY

* Given a query image, find a set of nearest neighbors that share similar scene configuration with the query.

# * Scene Retrieval

$$\mathcal{N}(x) = \{y_i \mid \mathrm{dist}(x, y_i) \leq (1+\epsilon)\mathrm{dist}(x, y_1),$$
$$y_1 = \arg\min\ \mathrm{dist}(x, y_i), i \leq K\}.$$

*<K, ε>-NN




# *<K, ε>-NN

PRINCETON UNIVERSITY

# Similarity/distance Metric

* Euclidean distance of GIST
* Spatial pyramid histogram intersection of HOG visual words
* Spatial pyramid histogram intersection of the ground-truth annotation

(a) Graph structure revealed by GIST distance

(b) The same as (a), showing ground-truth annotation

(c) The same as (d), showing RGB images

(d) Graph structure revealed by the distance on ground-truth annotation

PRINCETON UNIVERSITY

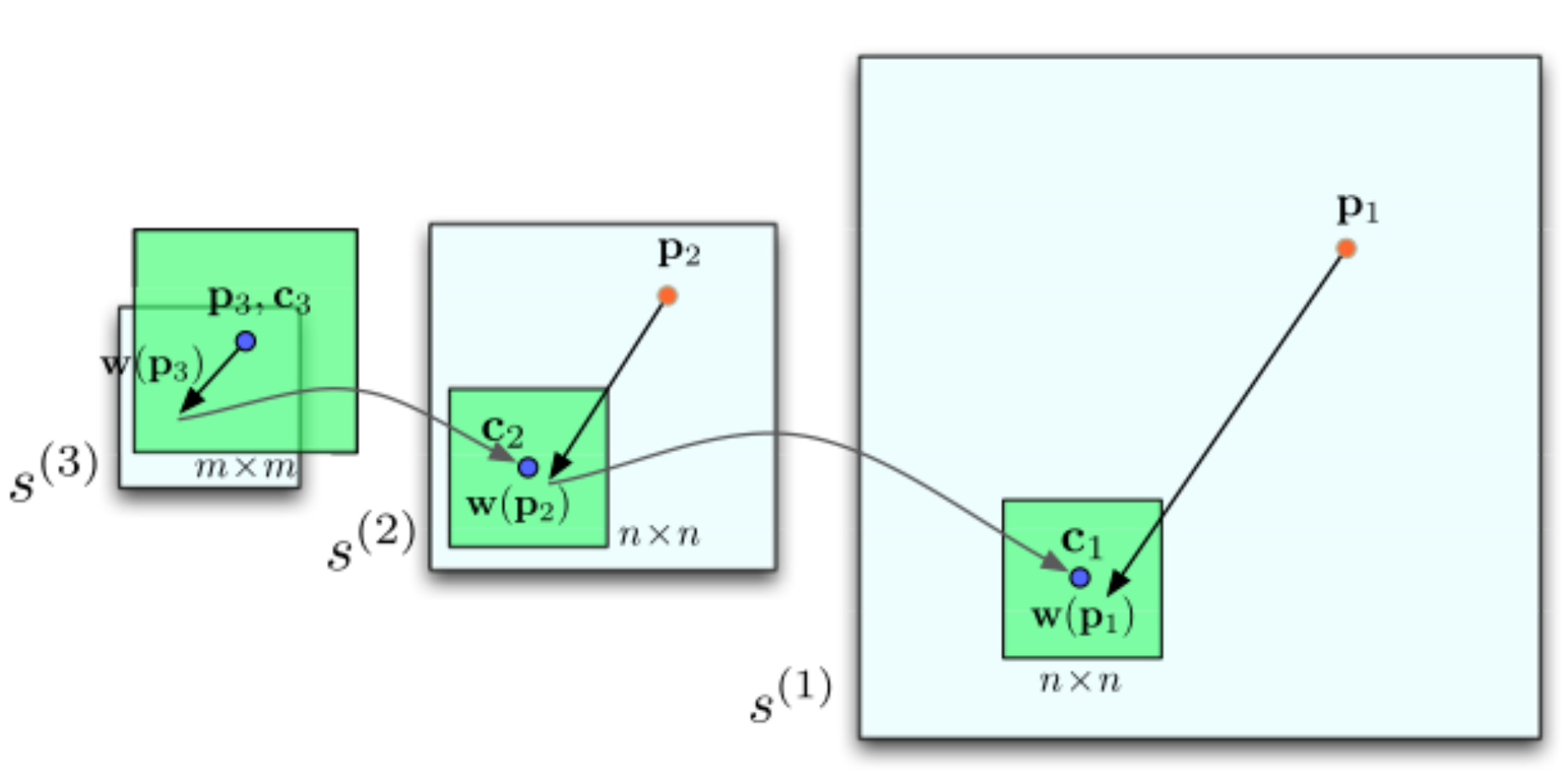
$$E(\mathbf{w}) = \sum_{\mathbf{p}} \min(\|s_1(\mathbf{p}) - s_2(\mathbf{p} + \mathbf{w}(\mathbf{p}))\|_1, t) +$$

$$\sum_{\mathbf{p}} \eta(|u(\mathbf{p})| + |v(\mathbf{p})|) +$$

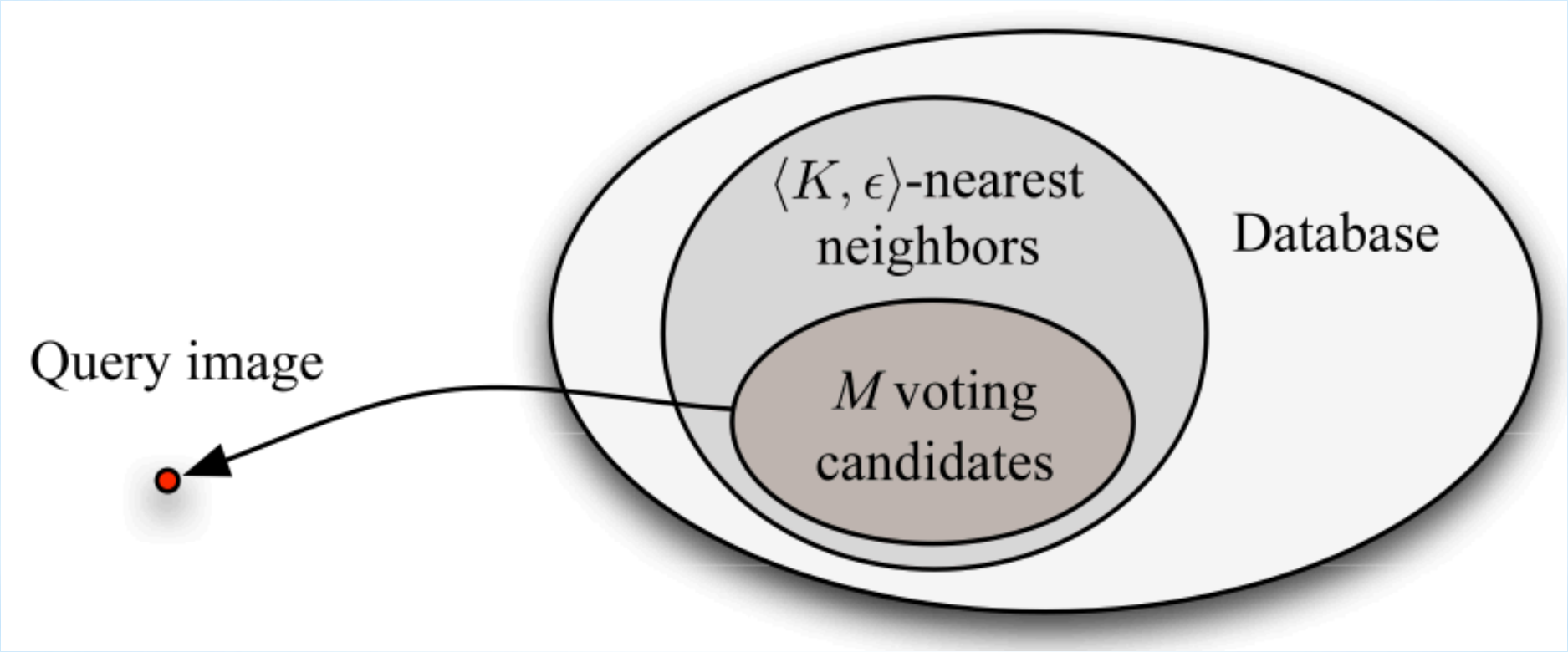
$$\sum_{(\mathbf{p},\mathbf{q})\in\varepsilon} \min(\lambda|u(\mathbf{p}) - u(\mathbf{q})|, d) +$$

$$\min(\lambda|v(\mathbf{p}) - v(\mathbf{q})|, d),$$

# *SIFT Flow for Dense Scene Alignment




# *Coarse-to-Fine pyramid SIFT Flow Matching


$\langle K, \epsilon \rangle$-nearest neighbors
Database
Query image
$M$ voting candidates

$$\{s_i, c_i, \mathbf{w}_i\}_{i=1:M}$$

SIFT image, annotation and SIFT flow field

$$c_i(\mathbf{p}) \in \{1, \ldots, L\}$$

# *Label transfer

$$-\log P\big(c|I, s, \{s_i, c_i, \mathbf{w}_i\}\big) = \sum_{\mathbf{p}} \psi\big(c(\mathbf{p}); s, \{s_i'\}\big)$$
$$+ \alpha \sum_{\mathbf{p}} \lambda\big(c(\mathbf{p})\big) + \beta \sum_{\{\mathbf{p},\mathbf{q}\}\in\varepsilon} \phi\big(c(\mathbf{p}), c(\mathbf{q}); I\big) + \log Z$$

Three components: likelihood, prior, spatial smoothness

# *Markov Random Field Model

PRINCETON UNIVERSITY

$$\psi(c(\mathbf{p}) = l) = \begin{cases} \min\limits_{i \in \Omega_{\mathbf{p},l}} \|s(\mathbf{p}) - s_i(\mathbf{p} + \mathbf{w}(\mathbf{p}))\|, & \Omega_{\mathbf{p},l} \neq \emptyset, \\ \tau, & \Omega_{\mathbf{p},l} = \emptyset, \end{cases}$$

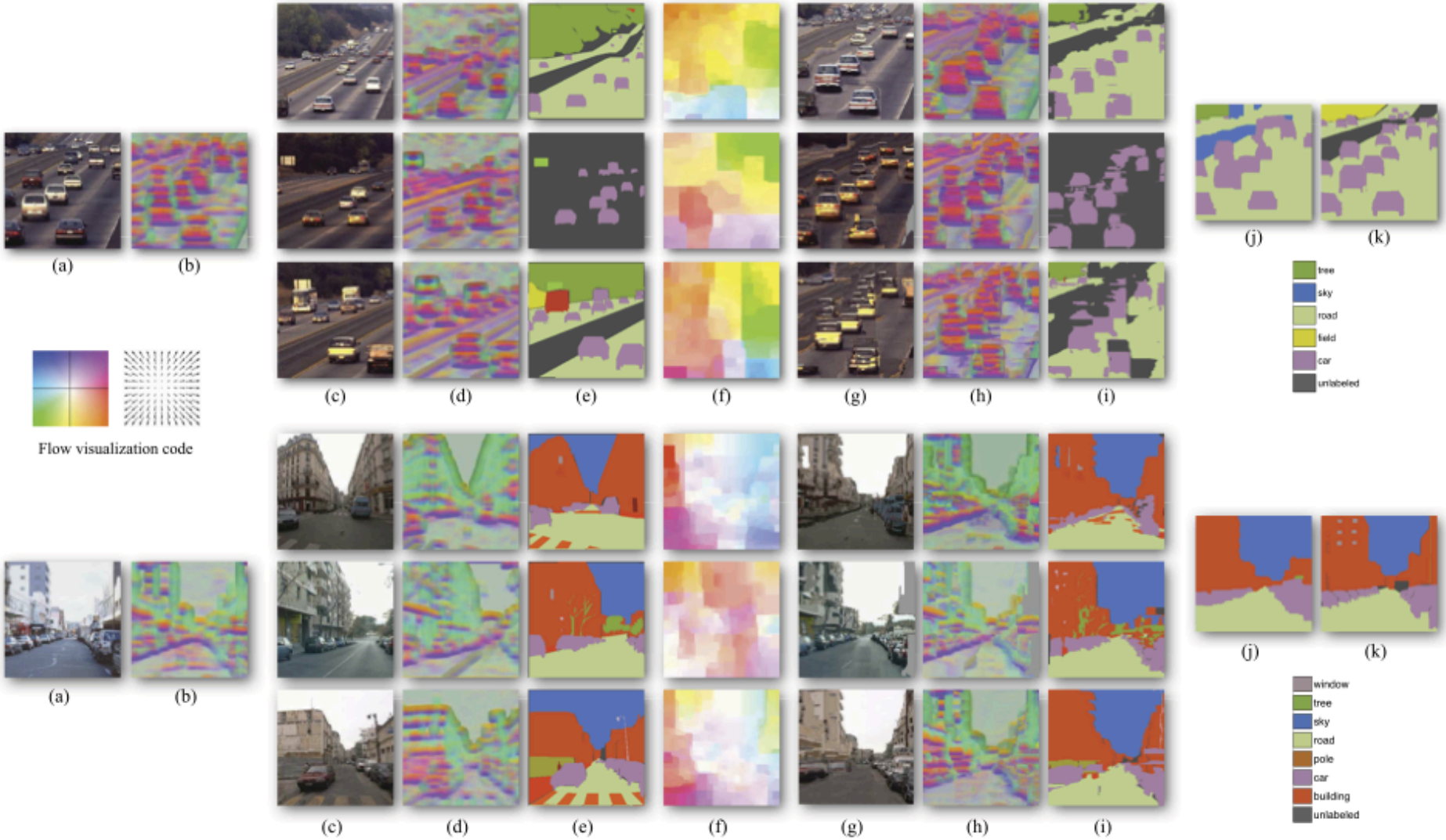
$$\Omega_{\mathbf{p},l} = \{i; c_i(\mathbf{p} + \mathbf{w}(\mathbf{p})) = l\}, \; l = 1, \ldots, L$$

$$\tau = \max_{s_1, s_2, \mathbf{p}} \|s_1(\mathbf{p}) - s_2(\mathbf{p})\|$$

# *Likelihood

$$\lambda\big(c(\mathbf{p}) = l\big) = -\log \mathrm{hist}_l(\mathbf{p})$$

*Prior


(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
tree
sky
road
field
car
unlabeled
Flow visualization code
(a)
(b)
(c)
(d)
(e)
(f)
(g)
(h)
(i)
(j)
(k)
window
tree
sky
road
pole
car
building
unlabeled

PRINCETON UNIVERSITY

# *Discussion

* The nonparametric scene parsing system outperforms TextonBoost in terms of both overall and per-class recognition rate.
* Color information is not included in the model.
* Similar performance to TextonBoost is achieved by matching color instead of dense SIFT descriptors.

PRINCETON UNIVERSITY

# *Discussion

* Smaller objects are better handled by detectors.
* Pixel-wise classifiers like TextonBoost can also be useful when good matching cannot be established or good nearest neighbors can hardly be retrieved.