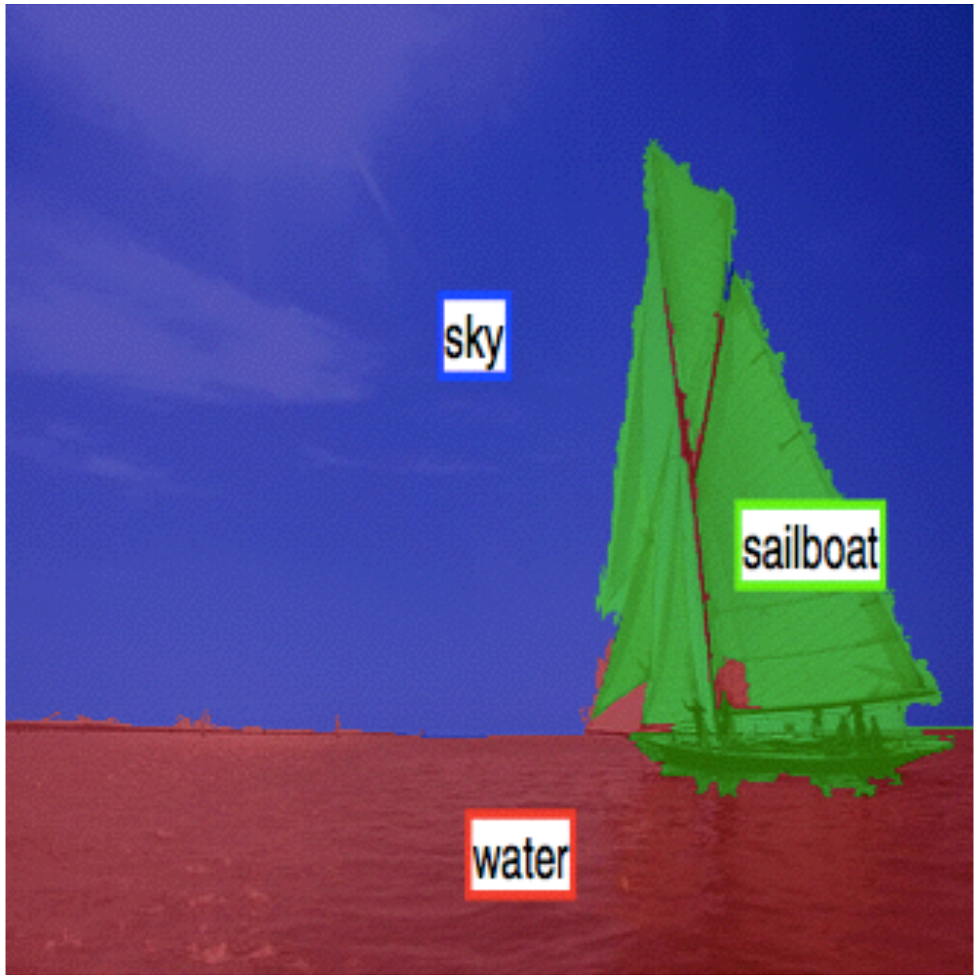
Connecting Modalities: Semi-supervised Segmentation and Annotation of Images Using Unaligned Text Corpora

Richard Socher & Li Fei-Fei

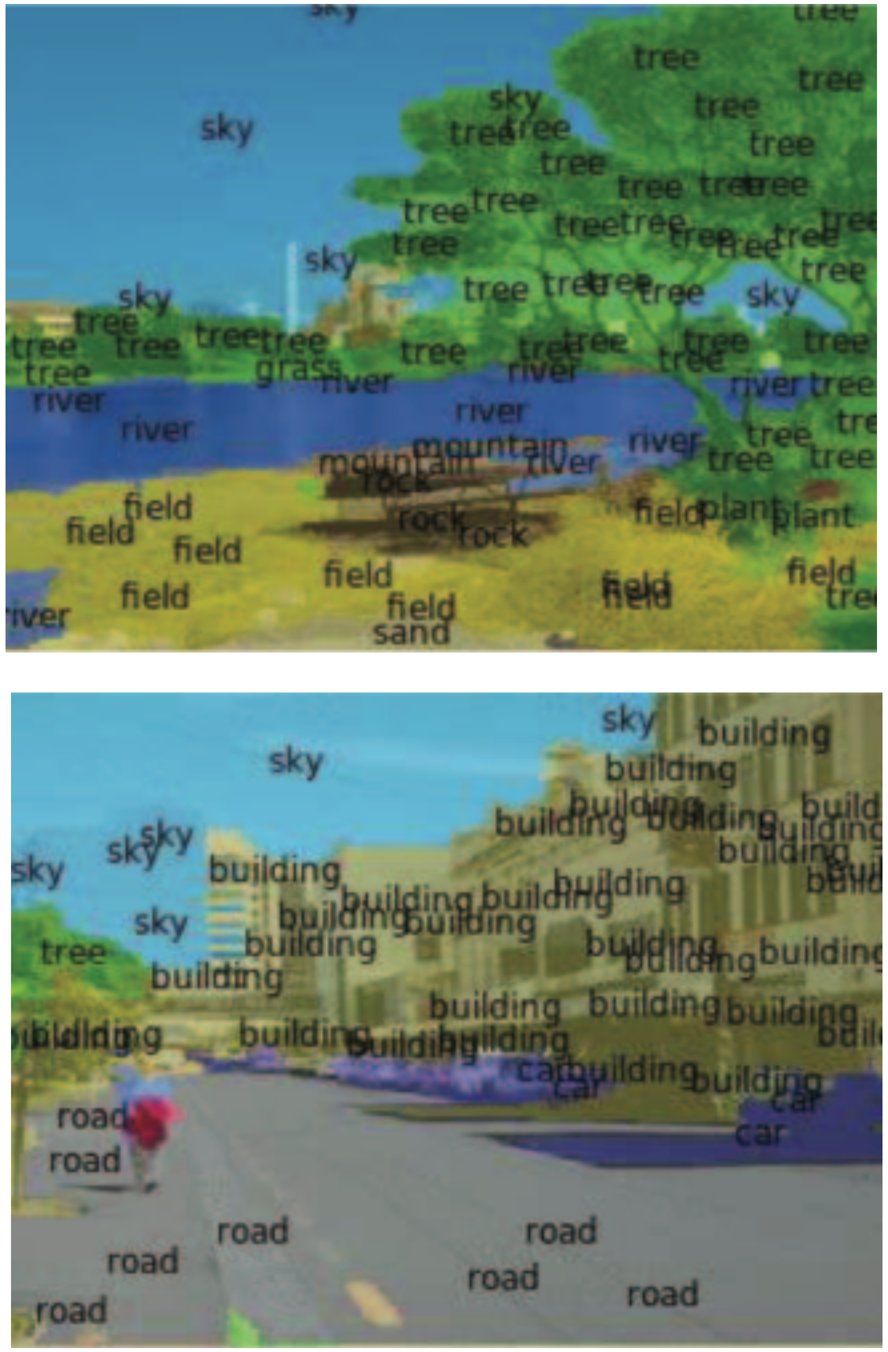
Presented by Jake Snell
CSC 2523
Feb 25, 2015

Overview



- A method for exploiting unaligned text corpora to build a **segmentation** and **annotation** model from a few labeled images.
- Novel use of kCCA to model similarity between visual words and corresponding text words.
- Achieved state-of-the-art performance in annotation and reasonable performance in segmentation

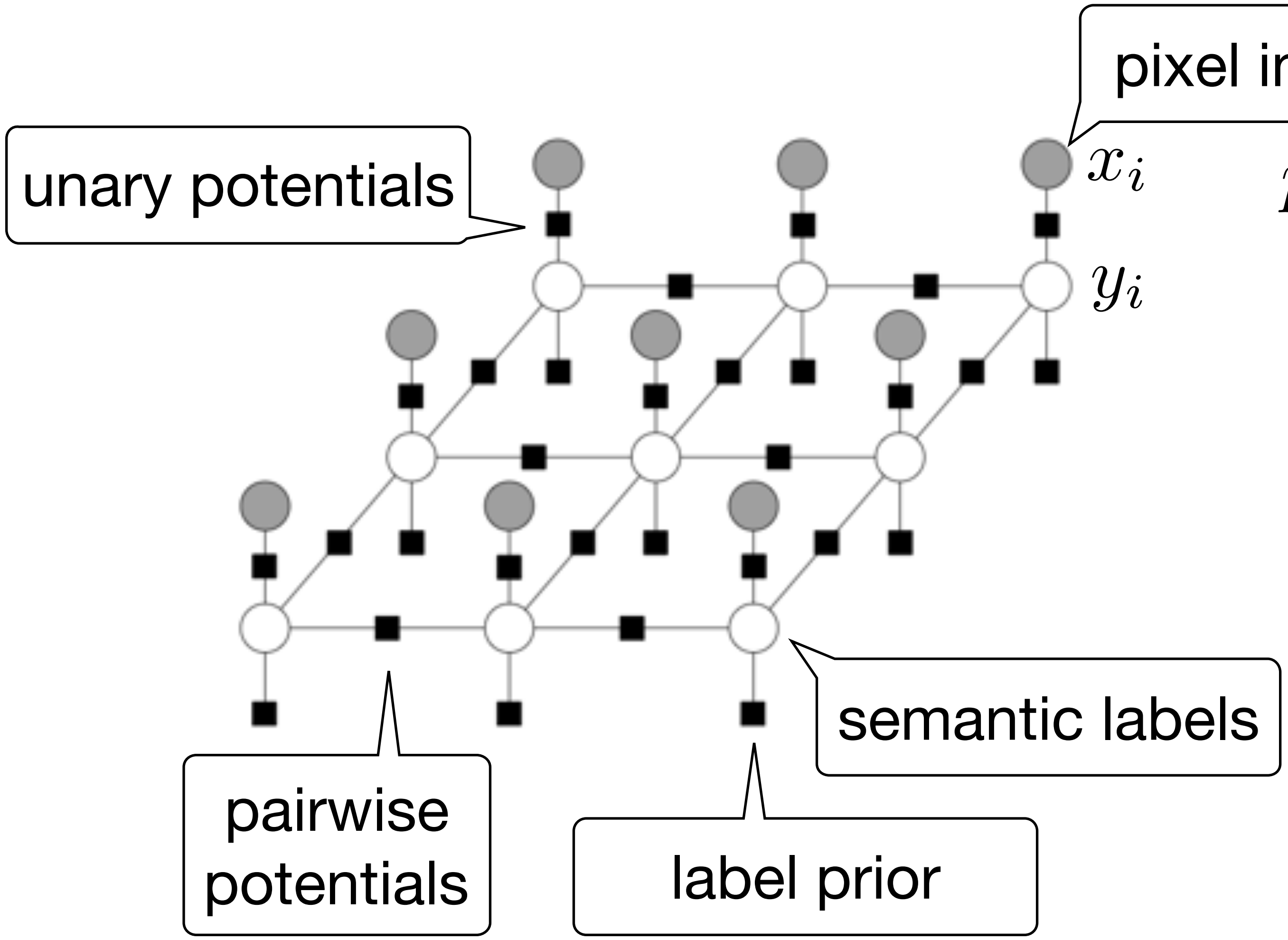
Semantic Image Segmentation



- **Goal:** Assign each pixel in an image to its semantic label.
- Requires more fine-grained level of understanding than object detection.
- **Challenge:** Fully-labeled training data is expensive to collect
  - VOC2012: 2,913 trainval images over 20 categories
  - ILSVRC 2012: 1.2 million images over 1,000 categories

C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning Hierarchical Features for Scene Labeling,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.

Conditional Random Fields


pixel intensity

unary potentials

semantic labels

pairwise potentials

label prior

$$p(y|x,w) = \frac{1}{Z(x,w)} \exp(-\langle w, \phi(x,y) \rangle)$$

$$\mathcal{L}(w) = -\sum_{n=1}^{N} \log p(y^n | x^n, w)$$

$$= \sum_{n=1}^{N} \langle w, \phi(x^n, y^n) \rangle + \sum_{n=1}^{N} \log Z(x^n, w)$$

- Alternatively, use SSVM which optimizes a margin-based criteria
- Simplistic model if graph is only 4-connected
- Strength depends to a large extent on unary potentials

S. Nowozin and C. H. Lampert, “Structured Learning and Prediction in Computer Vision,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3, pp. 185–365, Mar. 2011.

Effect of Unary & Pairwise Potentials

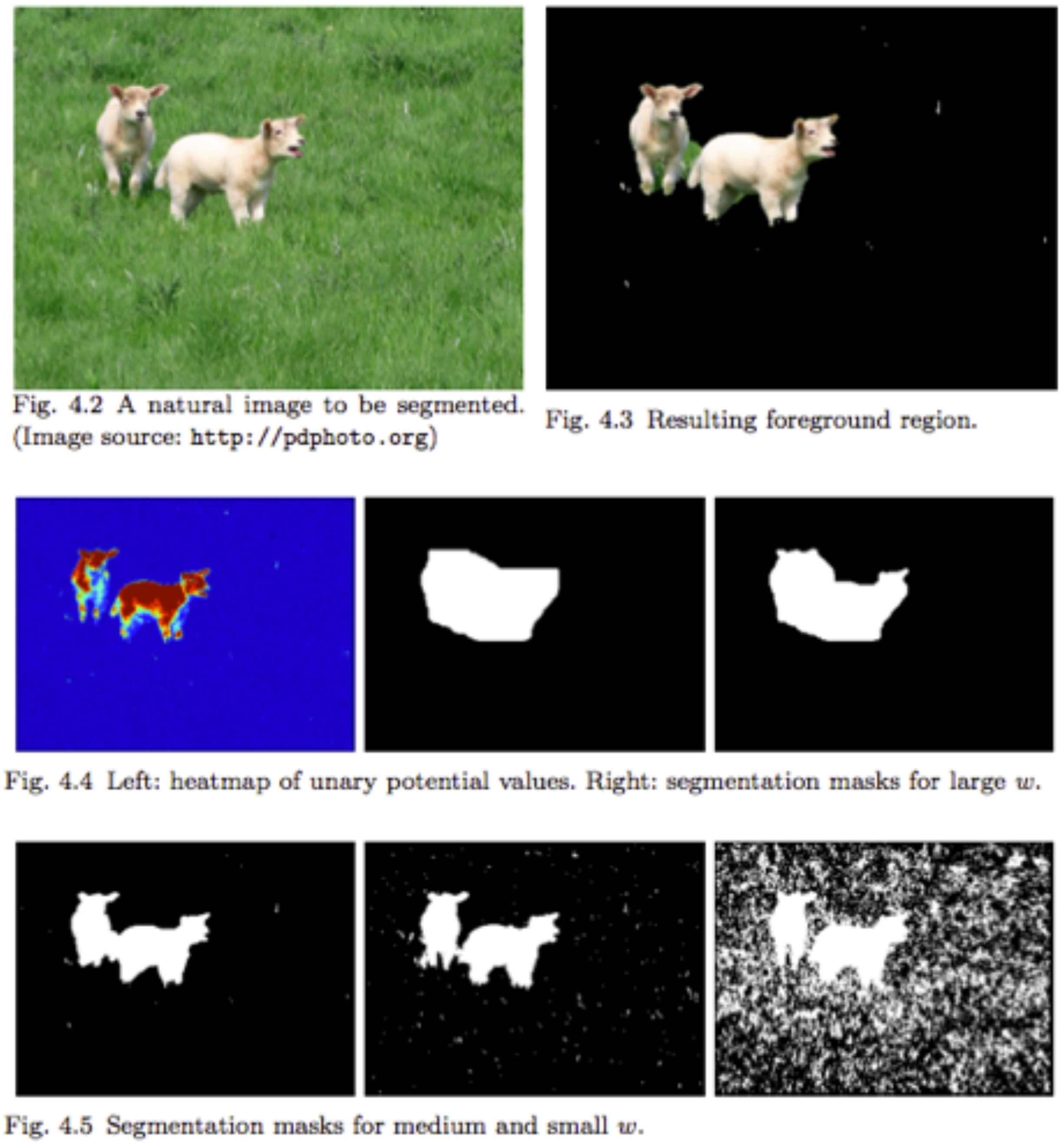


Fig. 4.2 A natural image to be segmented. (Image source: http://pdphoto.org)

Fig. 4.3 Resulting foreground region.

Fig. 4.4 Left: heatmap of unary potential values. Right: segmentation masks for large w.

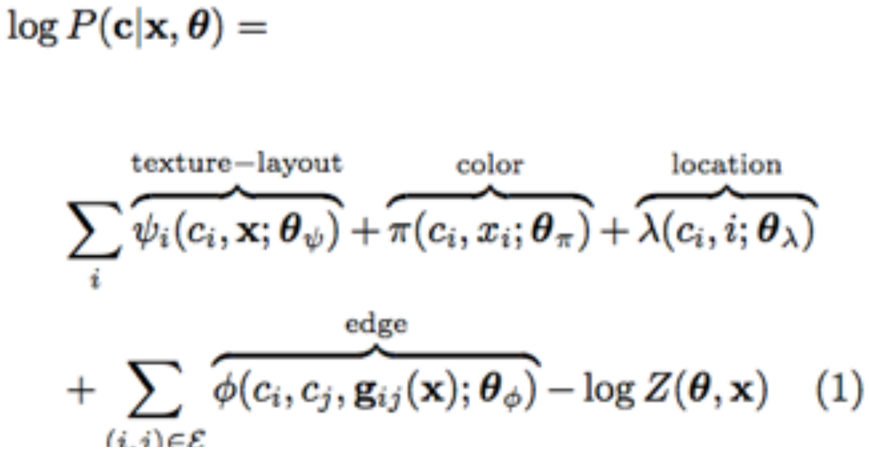
Fig. 4.5 Segmentation masks for medium and small w.

S. Nowozin and C. H. Lampert, “Structured Learning and Prediction in Computer Vision,” *Foundations and Trends® in Computer Graphics and Vision*, vol. 6, no. 3, pp. 185–365, Mar. 2011.

TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context

Jamie Shotton*
Machine Intelligence Laboratory, University of Cambridge
jamie@shotton.org

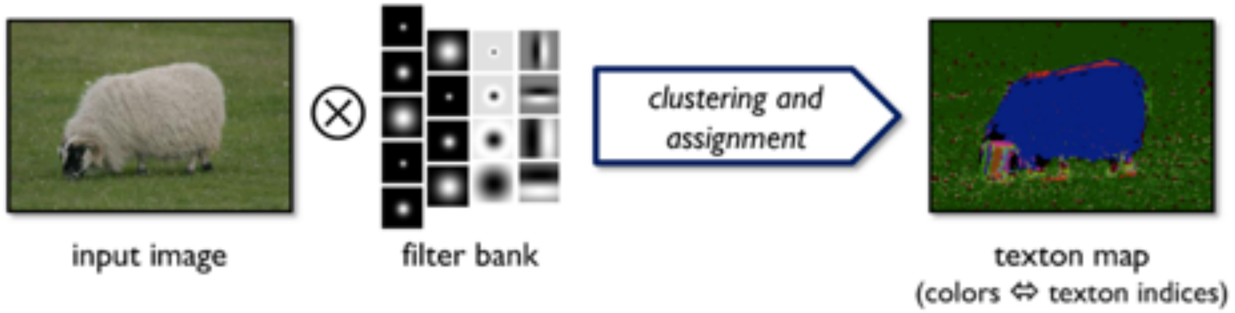
John Winn, Carsten Rother, Antonio Criminisi
Microsoft Research Cambridge, UK
[jwinn,carrot,antcrim]@microsoft.com

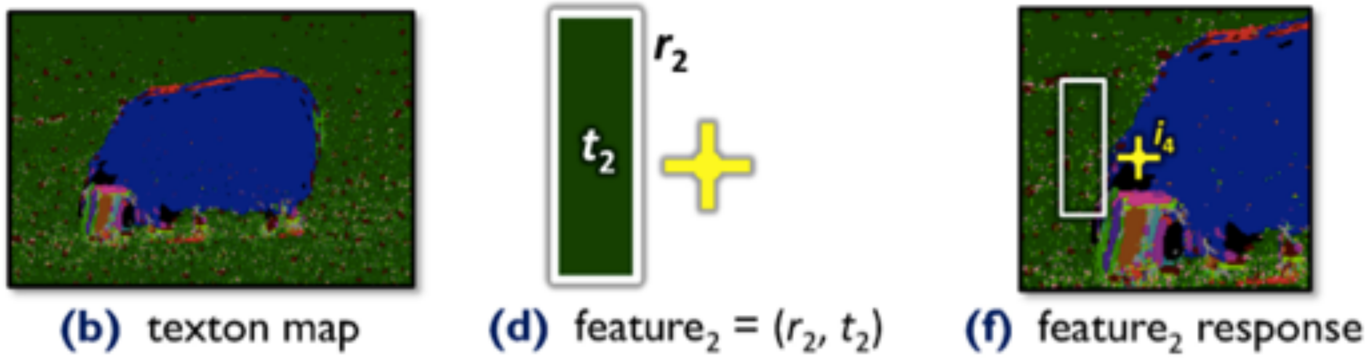


$$\log P(\mathbf{c}|\mathbf{x},\boldsymbol{\theta}) =$$

$$\sum_i \overbrace{\psi_i(c_i,\mathbf{x};\boldsymbol{\theta}_\psi)}^{\text{texture-layout}} + \overbrace{\pi(c_i,x_i;\boldsymbol{\theta}_\pi)}^{\text{color}} + \overbrace{\lambda(c_i,i;\boldsymbol{\theta}_\lambda)}^{\text{location}}$$

$$+ \sum_{(i,j)\in\mathcal{E}} \overbrace{\phi(c_i,c_j,\mathbf{g}_{ij}(\mathbf{x});\boldsymbol{\theta}_\phi)}^{\text{edge}} - \log Z(\boldsymbol{\theta},\mathbf{x}) \quad (1)$$



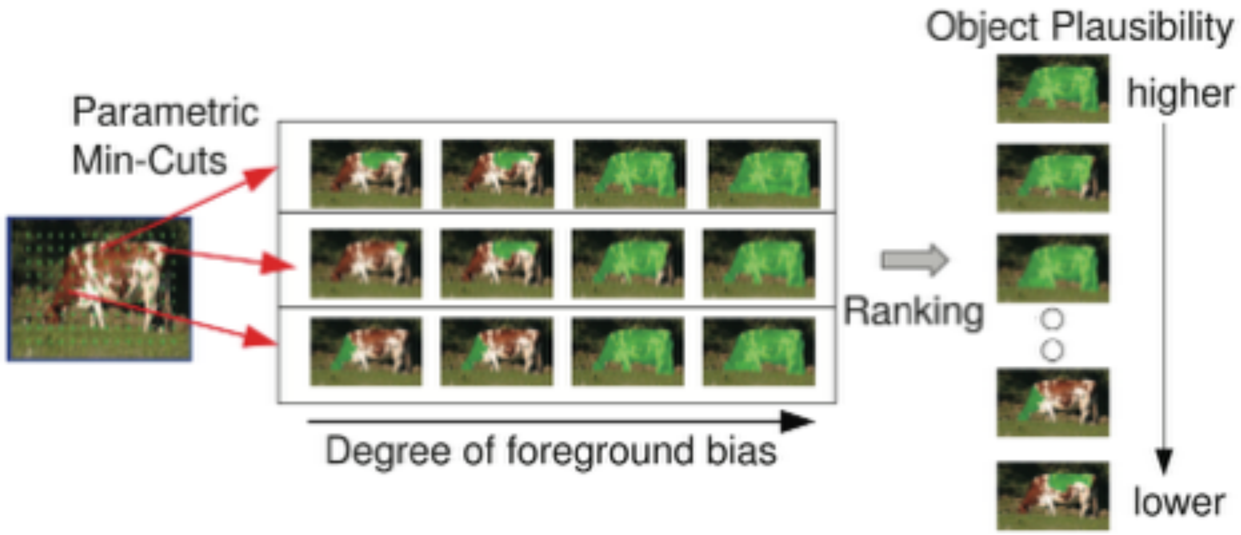


- CRF with piecewise training
- Unary potentials from boosted classifier on top of texture-layout filters
- Context is important!

J. Shotton, J. Winn, C. Rother, and A. Criminisi, “TextonBoost for Image Understanding: Multi-Class Object Recognition and Segmentation by Jointly Modeling Texture, Layout, and Context,” *Int J Comput Vis*, vol. 81, no. 1, pp. 2–23, 2009.

CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts

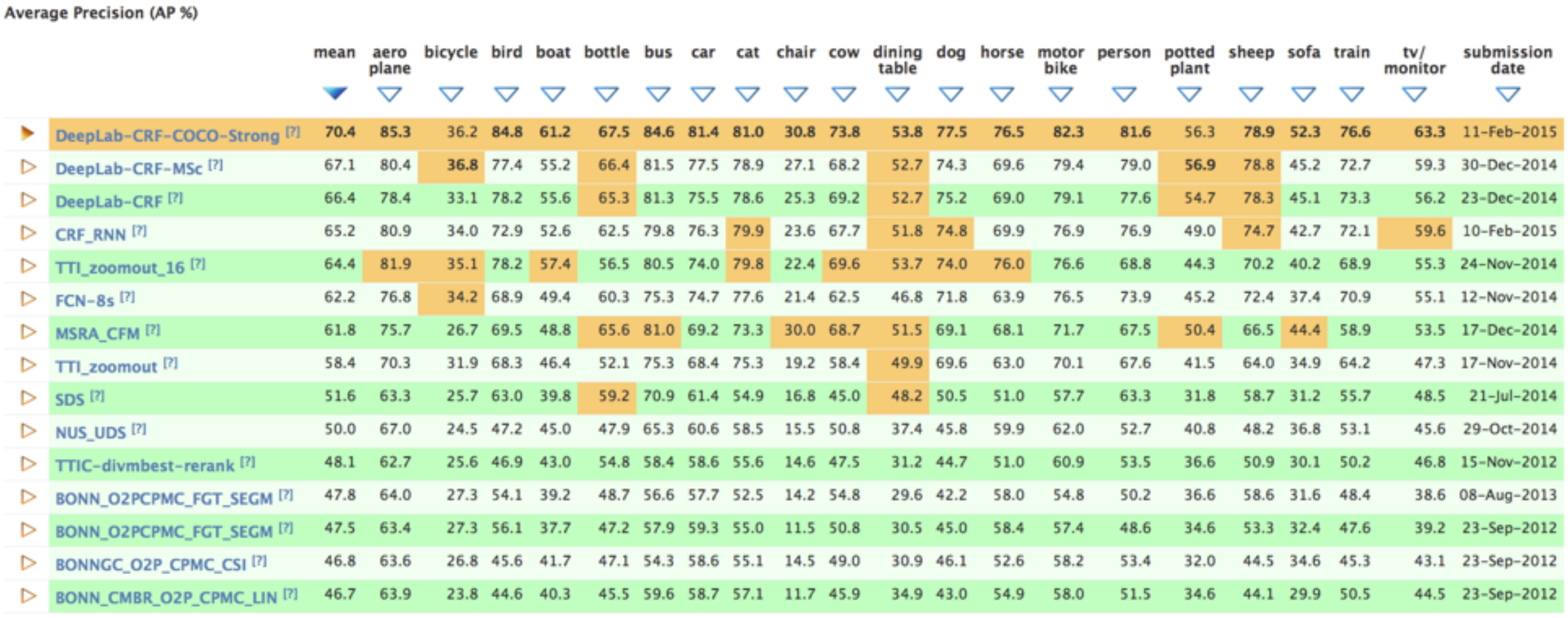
João Carreira, *Student Member, IEEE*, and Cristian Sminchisescu, *Member, IEEE*



- Winner of VOC2009 & 2010
- Use simple graph cut algorithm to make segment proposals
- Rerank proposed segments based on mid-level region properties
- Combine ranked regions to obtain final segmentation

J. Carreira and C. Sminchisescu, "CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 7, pp. 1312–1328, Jul. 2012.

PASCAL VOC2012 Segmentation Leaderboard



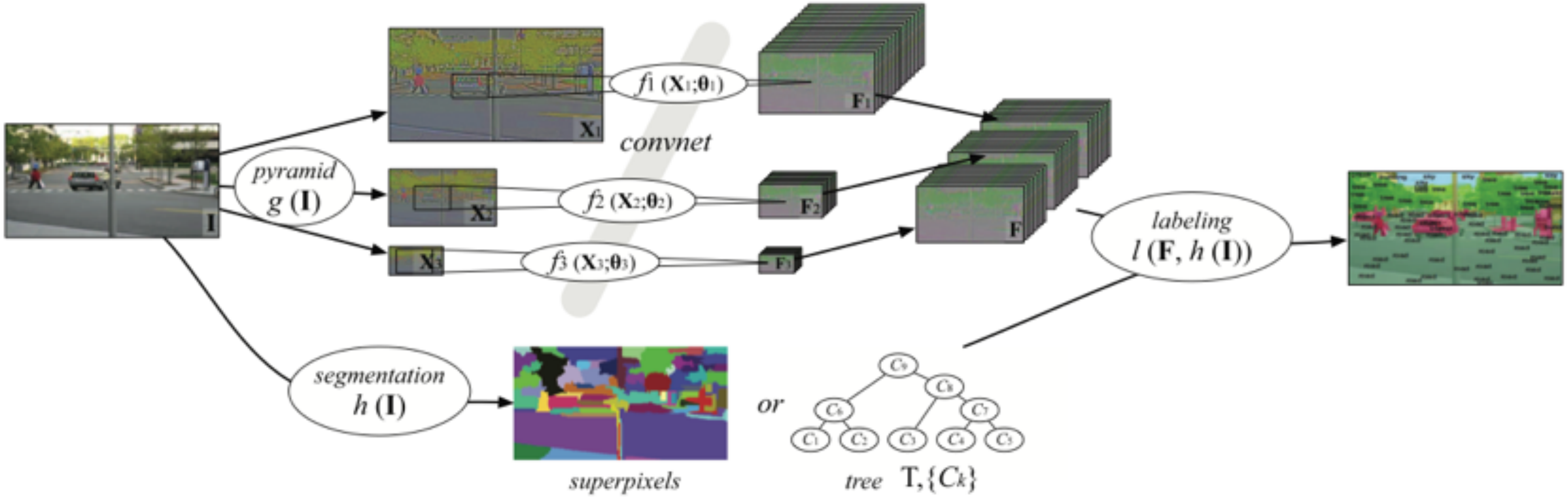
Average Precision (AP %)

	mean	aero plane	bicycle	bird	boat	bottle	bus	car	cat	chair	cow	dining table	dog	horse	motor bike	person	potted plant	sheep	sofa	train	tv/ monitor	submission date
DeepLab-CRF-COCO-Strong [?]	70.4	85.3	36.2	84.8	61.2	67.5	84.6	81.4	81.0	30.8	73.8	53.8	77.5	76.5	82.3	81.6	56.3	78.9	52.3	76.6	63.3	11-Feb-2015
DeepLab-CRF-MSc [?]	67.1	80.4	36.8	77.4	55.2	66.4	81.5	77.5	78.9	27.1	68.2	52.7	74.3	69.6	79.4	79.0	56.9	78.8	45.2	72.7	59.3	30-Dec-2014
DeepLab-CRF [?]	66.4	78.4	33.1	78.2	55.6	65.3	81.3	75.5	78.6	25.3	69.2	52.7	75.2	69.0	79.1	77.6	54.7	78.3	45.1	73.3	56.2	23-Dec-2014
CRF_RNN [?]	65.2	80.9	34.0	72.9	52.6	62.5	79.8	76.3	79.9	23.6	67.7	51.8	74.8	69.9	76.9	76.9	49.0	74.7	42.7	72.1	59.6	10-Feb-2015
TTI_zoomout_16 [?]	64.4	81.9	35.1	78.2	57.4	56.5	80.5	74.0	79.8	22.4	69.6	53.7	74.0	76.0	76.6	68.8	44.3	70.2	40.2	68.9	55.3	24-Nov-2014
FCN-8s [?]	62.2	76.8	34.2	68.9	49.4	60.3	75.3	74.7	77.6	21.4	62.5	46.8	71.8	63.9	76.5	73.9	45.2	72.4	37.4	70.9	55.1	12-Nov-2014
MSRA_CFM [?]	61.8	75.7	26.7	69.5	48.8	65.6	81.0	69.2	73.3	30.0	68.7	51.5	69.1	68.1	71.7	67.5	50.4	66.5	44.4	58.9	53.5	17-Dec-2014
TTI_zoomout [?]	58.4	70.3	31.9	68.3	46.4	52.1	75.3	68.4	75.3	19.2	58.4	49.9	69.6	63.0	70.1	67.6	41.5	64.0	34.9	64.2	47.3	17-Nov-2014
SDS [?]	51.6	63.3	25.7	63.0	39.8	59.2	70.9	61.4	54.9	16.8	45.0	48.2	50.5	51.0	57.7	63.3	31.8	58.7	31.2	55.7	48.5	21-Jul-2014
NUS_UDS [?]	50.0	67.0	24.5	47.2	45.0	47.9	65.3	60.6	58.5	15.5	50.8	37.4	45.8	59.9	62.0	52.7	40.8	48.2	36.8	53.1	45.6	29-Oct-2014
TTIC-divmbest-rerank [?]	48.1	62.7	25.6	46.9	43.0	54.8	58.4	58.6	55.6	14.6	47.5	31.2	44.7	51.0	60.9	53.5	36.6	50.9	30.1	50.2	46.8	15-Nov-2012
BONN_O2PCPMC_FGT_SEGM [?]	47.8	64.0	27.3	54.1	39.2	48.7	56.6	57.7	52.5	14.2	54.8	29.6	42.2	58.0	54.8	50.2	36.6	58.6	31.6	48.4	38.6	08-Aug-2013
BONN_O2PCPMC_FGT_SEGM [?]	47.5	63.4	27.3	56.1	37.7	47.2	57.9	59.3	55.0	11.5	50.8	30.5	45.0	58.4	57.4	48.6	34.6	53.3	32.4	47.6	39.2	23-Sep-2012
BONNGC_O2P_CPMC_CSI [?]	46.8	63.6	26.8	45.6	41.7	47.1	54.3	58.6	55.1	14.5	49.0	30.9	46.1	52.6	58.2	53.4	32.0	44.5	34.6	45.3	43.1	23-Sep-2012
BONN_CMBR_O2P_CPMC_LIN [?]	46.7	63.9	23.8	44.6	40.3	45.5	59.6	58.7	57.1	11.7	45.9	34.9	43.0	54.9	58.0	51.5	34.6	44.1	29.9	50.5	44.5	23-Sep-2012

http://host.robots.ox.ac.uk:8080/leaderboard/displaylb.php?challengeid=11&compid=6 (Accessed Feb 24, 2015)

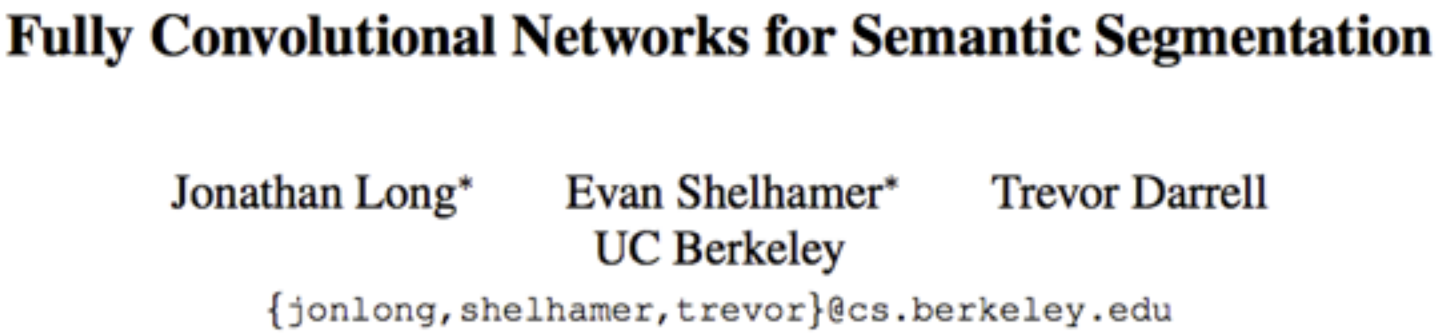
Learning Hierarchical Features for Scene Labeling

Clément Farabet, Camille Couprie, Laurent Najman, and Yann LeCun



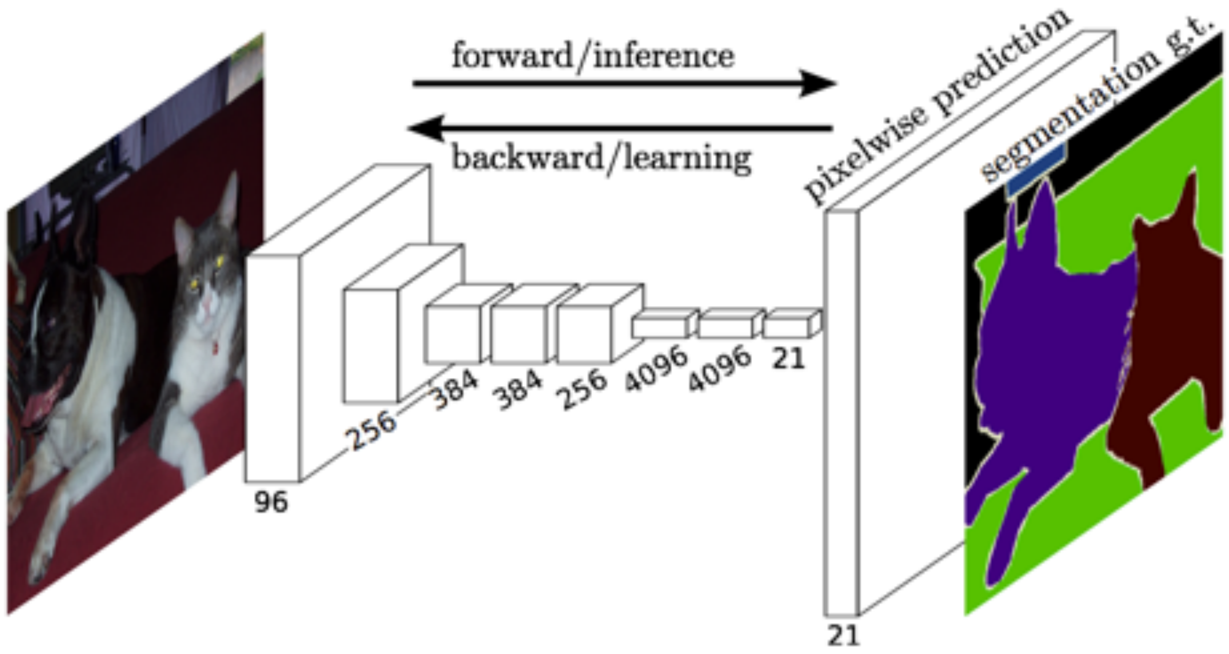
- Train multiscale convnet to get strong unary potentials
- Use tree to explain each superpixel by the ancestor with the lowest impurity (entropy over categories)

C. Farabet, C. Couprie, L. Najman, and Y. LeCun, “Learning Hierarchical Features for Scene Labeling,” *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 8, pp. 1915–1929, Aug. 2013.



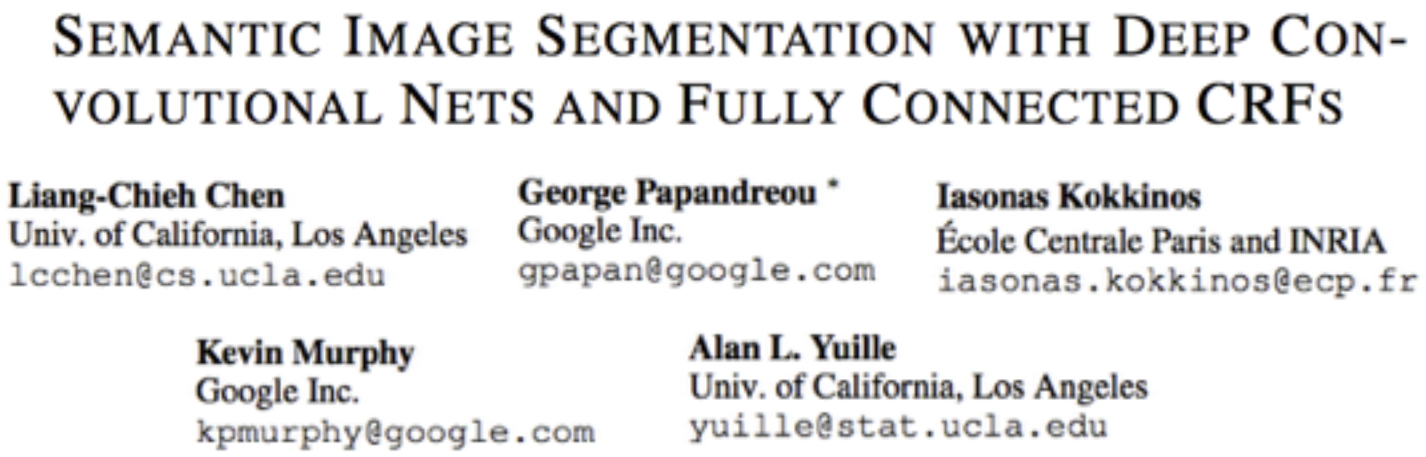
Fully Convolutional Networks for Semantic Segmentation

Jonathan Long* Evan Shelhamer* Trevor Darrell
UC Berkeley
{jonlong,shelhamer,trevor}@cs.berkeley.edu



- Currently sixth on VOC2012 leaderboard
- Leverage classification convnets to obtain a coarse heatmap over semantic labels
- Deconvolutional layer to scale the heatmap up to full size
- Fine-tune network by backpropagating per-pixel multinomial logistic loss

J. Long, E. Shelhamer, and T. Darrell, "Fully Convolutional Networks for Semantic Segmentation," *arXiv.org*, vol. cs.CV. 14-Nov-2014.



Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs

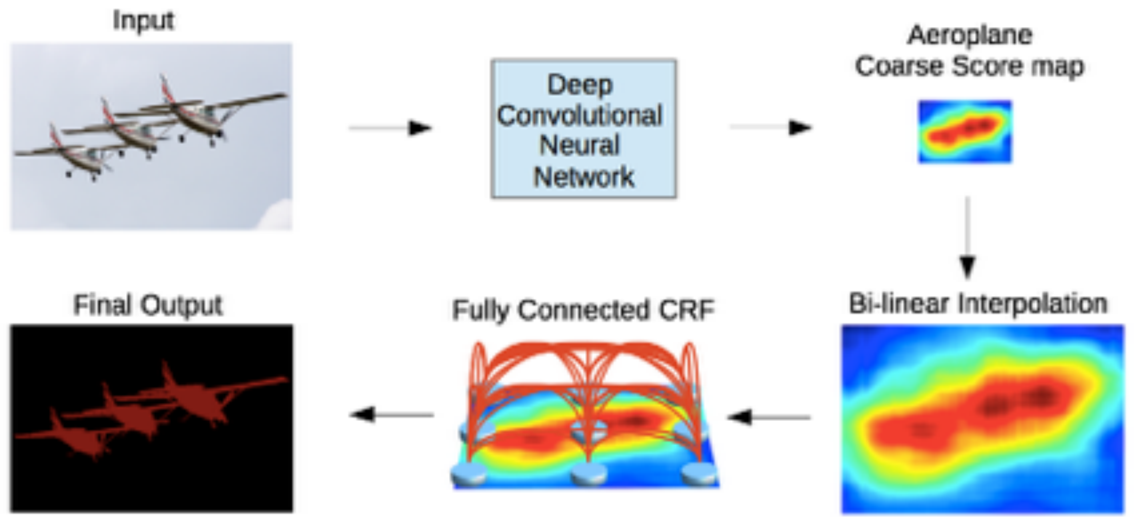
Liang-Chieh Chen
Univ. of California, Los Angeles
lcchen@cs.ucla.edu

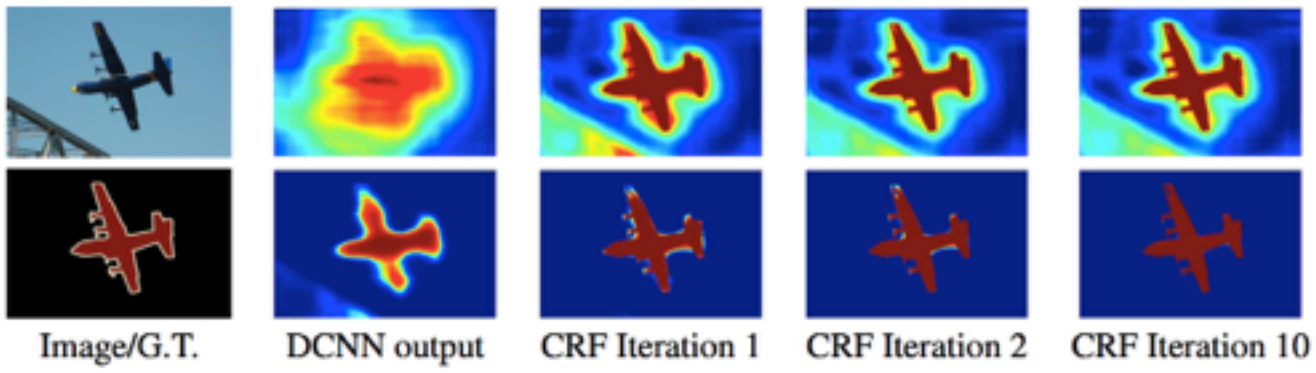
George Papandreou *
Google Inc.
gpapan@google.com

Iasonas Kokkinos
École Centrale Paris and INRIA
iasonas.kokkinos@ecp.fr

Kevin Murphy
Google Inc.
kpmurphy@google.com

Alan L. Yuille
Univ. of California, Los Angeles
yuille@stat.ucla.edu





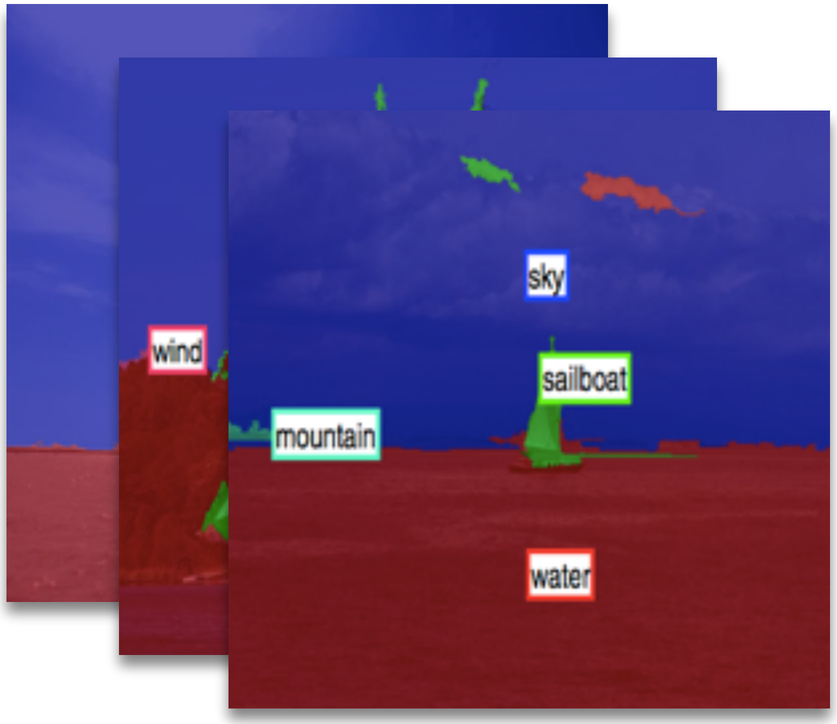
- Currently second on VOC2012 leaderboard
- Also based on classification convnets
- Use bi-linear interpolation to upscale coarse heatmap
- Fully connected CRF on top to clean up output
- Piecewise training to decouple unary potentials from CRF parameters

L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, “Semantic Image Segmentation with Deep Convolutional Nets and Fully Connected CRFs,” *arXiv.org*, vol. cs.CV. 22-Dec-2014.

Takeaways

- Lack of data is a challenge
  - Semi-supervised learning with auxiliary data
  - Base off of classification models trained with lots of data
- Best approaches have both:
  - strong unary potentials (convnets are a boon)
  - way to incorporate context (structured model to help squeeze out extra few %)

Motivation



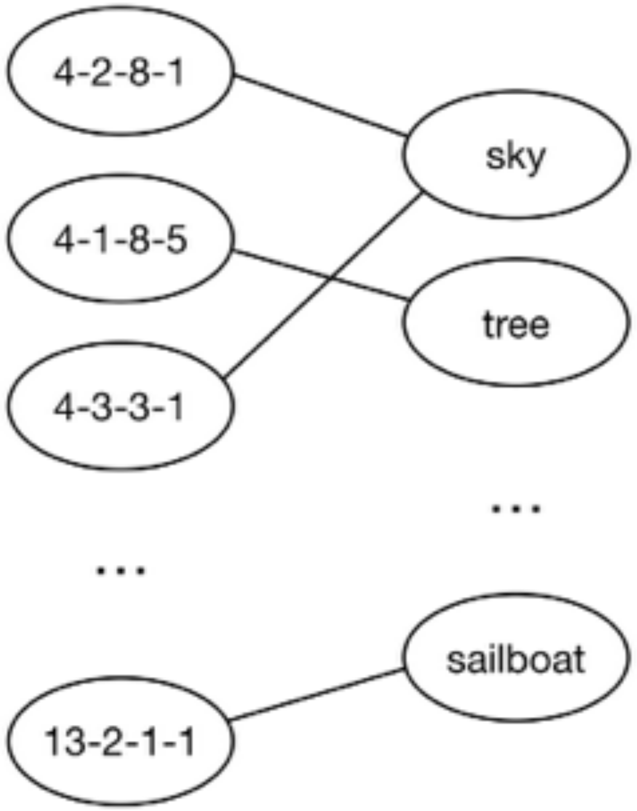
1-5 Labeled Images

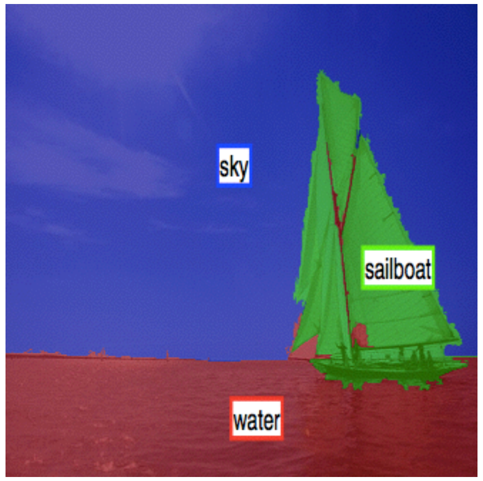
*The halyard released, hands almost numb with cold already, squirmed around to crawl back and froze as he felt the **sailboat** rise awkwardly to a huge wave. As far as the eye could see the black **ocean** was slashed with white streaks where **waves** were breaking. The … **sea** was angry and the **sky** screamed at it …*

Unlabeled Text Corpus

- Building strong models for segmentation is hard due to scarcity of labeled data.
- Unaligned text is relatively plentiful
- Can we apply co-occurences observed in text articles on the same topic to the image model itself?
- Key assumptions:
  - Concepts in the text have visual counterparts in the image.
  - Neighboring concept pairs in the text are more likely to also be neighbors in the image.

Problem





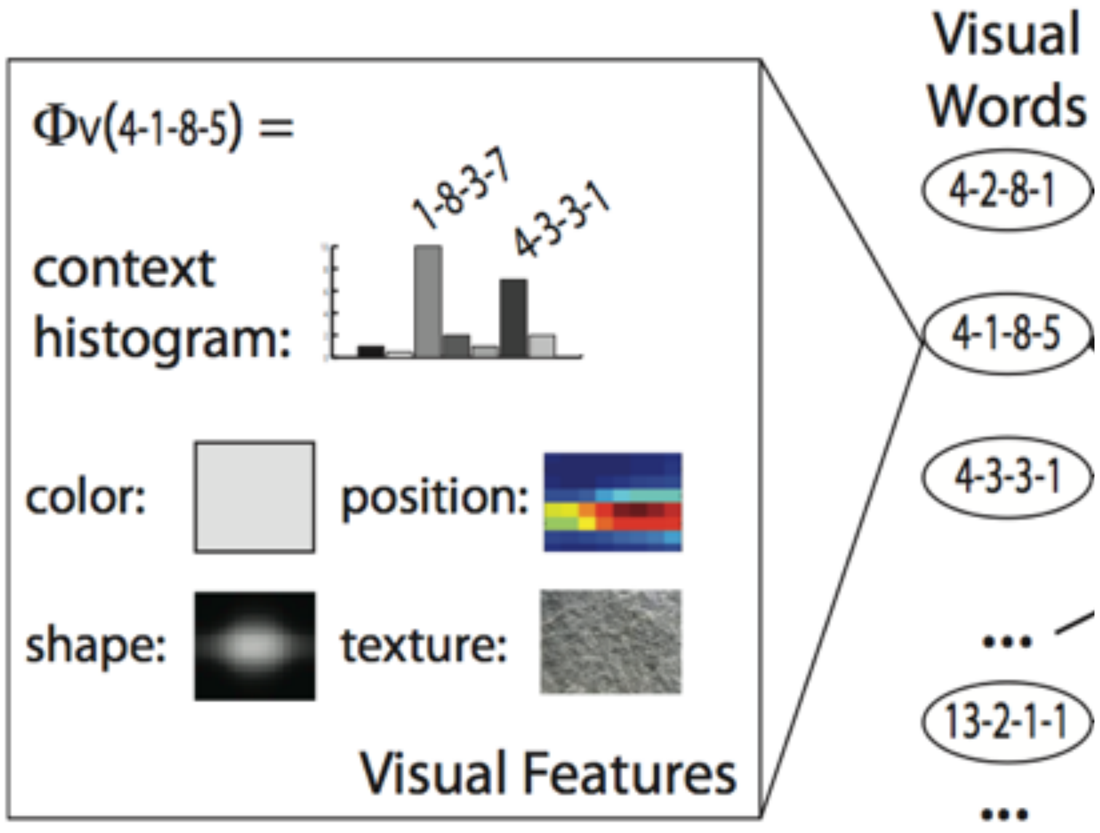
{sky, water, sailboat}

- Learn a mapping between region-level image features and text labels.
- Given a test image, use this mapping to predict text labels for the image at both a global level (**annotation**) and at the pixel level (**segmentation**).

Approach

- Use a superpixel algorithm to break images down into a set of non-overlapping regions.
- Extract visual features for each region, and assign each region to a visual word by clustering the features.
- Extract textual features for each text label by computing context and adjective histograms.
- Learn a generative model of visual and textual features consisting of:
  - A set of mappings between visual words and textual words, where many visual words may map to a single textual word.
  - A latent "concept" variable associated with each mapping which is responsible for explaining all associated visual and textual features.
  - A background model responsible for explaining all visual and textual left out of the mapping.
- Use the learned mapping to perform annotation and segmentation on unseen images.

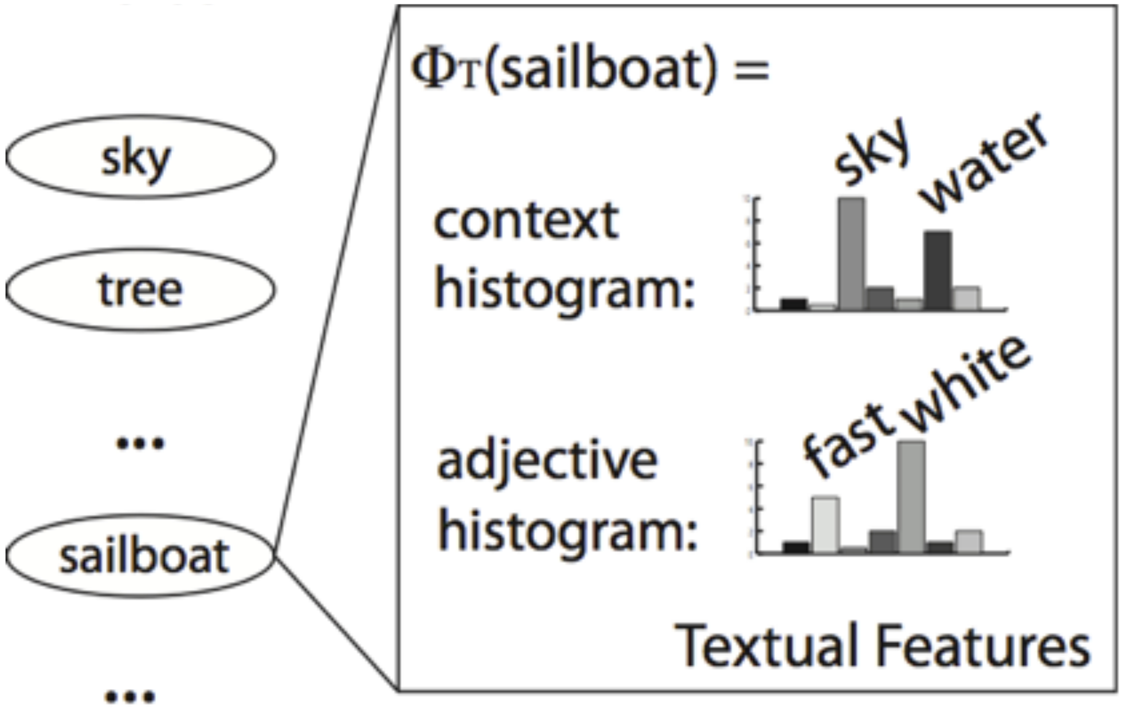
Visual Features



$$C_{\texttt{color}} - C_{\texttt{position}} - C_{\texttt{texture}} - C_{\texttt{shape}}$$

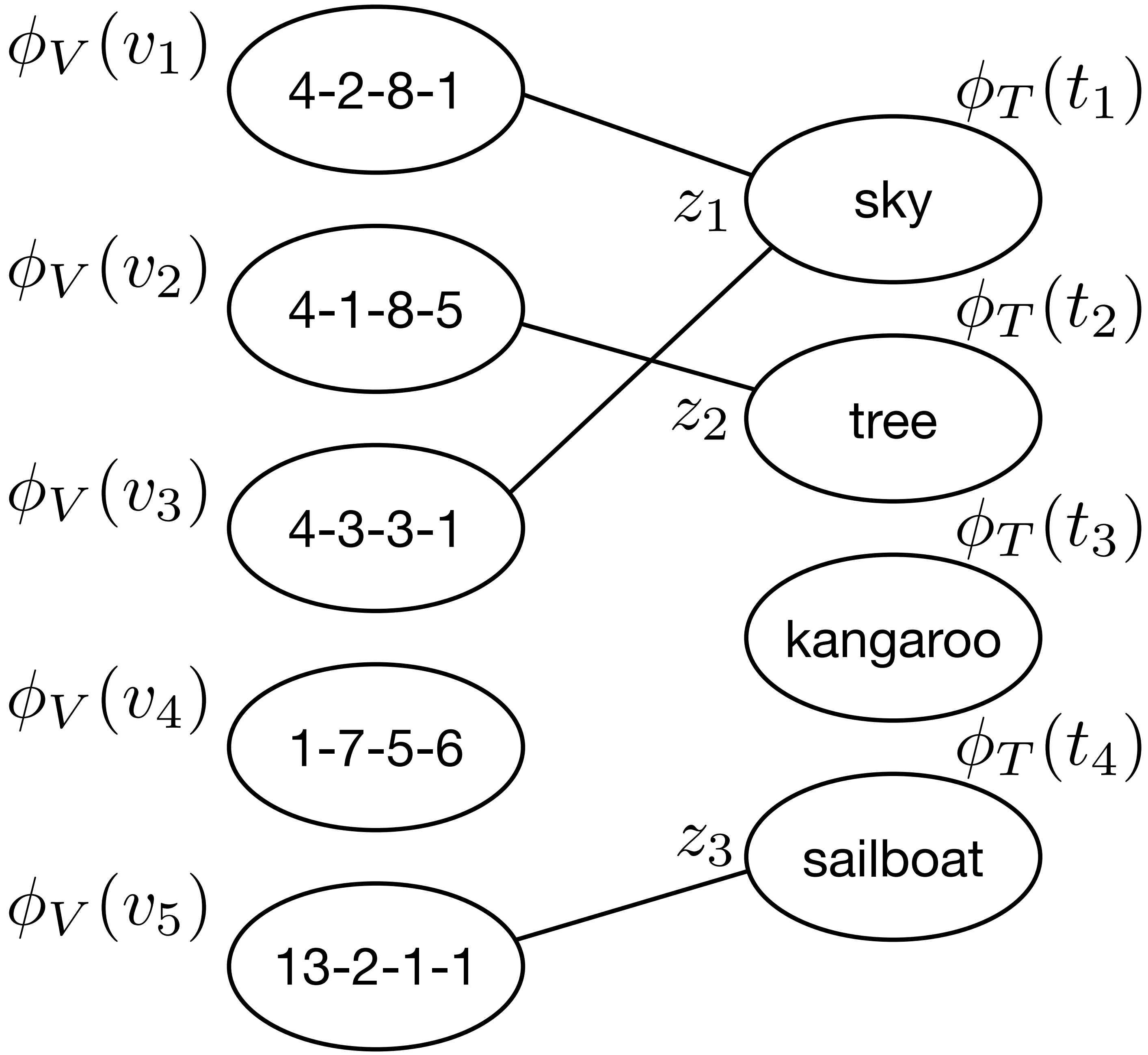
- For each region, extract the following features:
  - **Color** - RGB histogram
  - **Texture** - Mean responses of filterbanks
  - **Position** - location in an 8x8 grid
  - **Shape** - binary histogram of the segment mask downscaled to 32 x 32
- Cluster each feature independently
- Assign each region to a visual word by concatenating the assigned cluster for each of the four features

Textual Features



- **Context histogram**: normalized frequency of words within window of size four (only counting nouns)
- **Adjective histogram**: Normalized frequencies of co-occurring adjectives

Generative Process



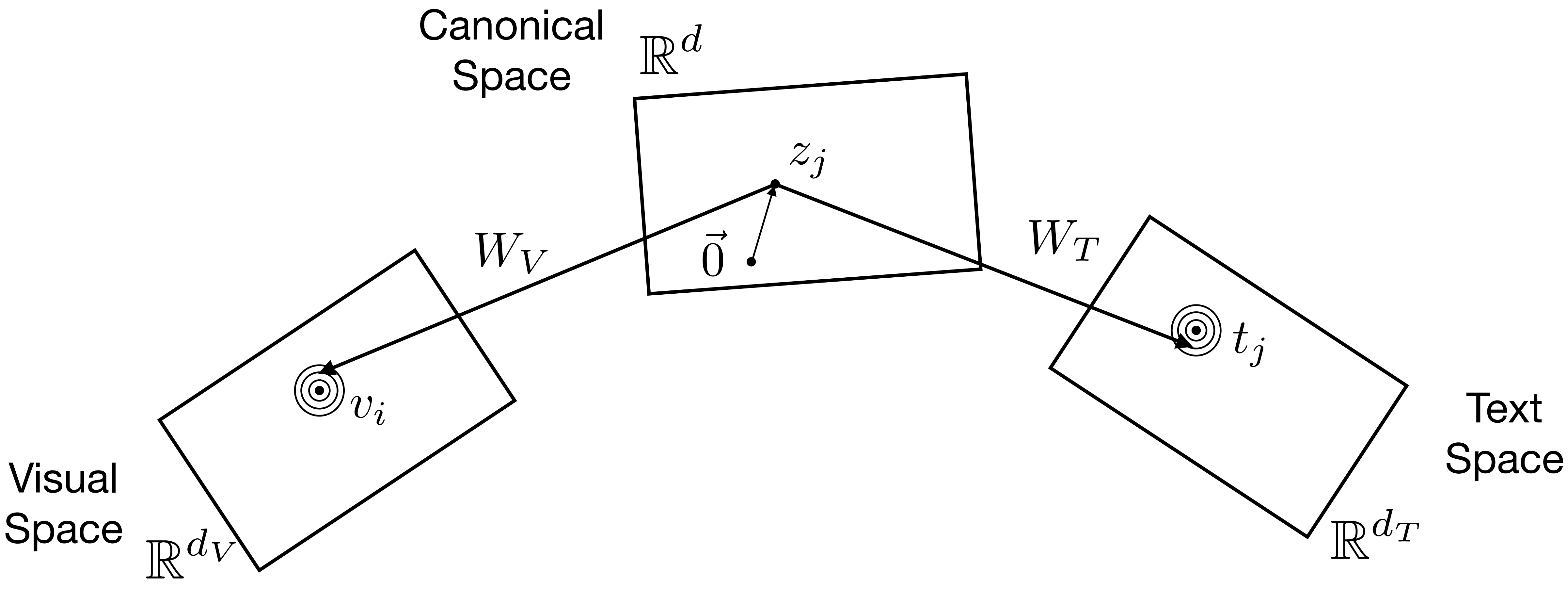
EM Algorithm

- M-Step
  - Given a mapping, update projection matrices by maximizing log likelihood:

$$\max_{\xi} \sum_{(i,j)\in M} \log p(v_i, t_j, M_{ij}; \xi) \qquad \xi = (W_V, \Psi_V, W_T, \Psi_T)$$

- E-Step
  - Approximate the posterior distribution over all possible mappings by a single weighted mapping *M.*

M-Step



$$v_i \sim \mathcal{N}(W_V z_j + \mu_V, \Psi_V)$$

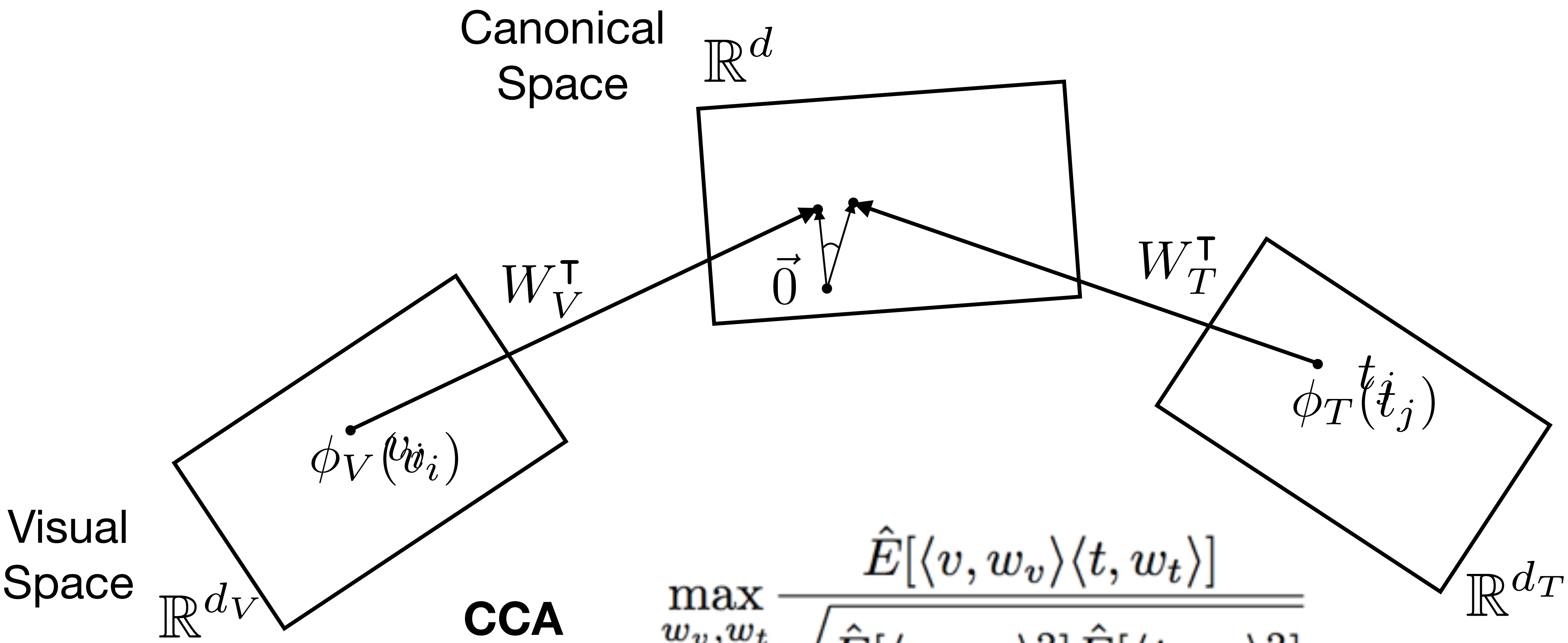
$$t_j \sim \mathcal{N}(W_T z_j + \mu_T, \Psi_T)$$

$$\max_{\xi} \sum_{(i,j)\in M} \log p(v_i, t_j, M_{ij}; \xi)$$

$$\xi = (W_V, \Psi_V, W_T, \Psi_T)$$

Adapted from: A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein, “Learning Bilingual Lexicons from Monolingual Corpora.,” *ACL*, 2008.

An Alternate View



CCA

$$\max_{w_v,w_t} \frac{\hat{E}[\langle v, w_v\rangle\langle t, w_t\rangle]}{\sqrt{\hat{E}[\langle v, w_v\rangle^2]\hat{E}[\langle t, w_t\rangle^2]}}$$

$$K_V(v_i, v_j) = \langle \phi_V(v_i), \phi_V(v_j)\rangle$$

$$K_T(t_i, t_j) = \langle \phi_T(t_i), \phi_T(t_j)\rangle$$

kCCA

$$\max_{w_v,w_t} \frac{w_v^T K_V K_T w_t}{\sqrt{w_v^T K_V^2 w_v \cdot w_t^T K_T^2 w_t}}$$

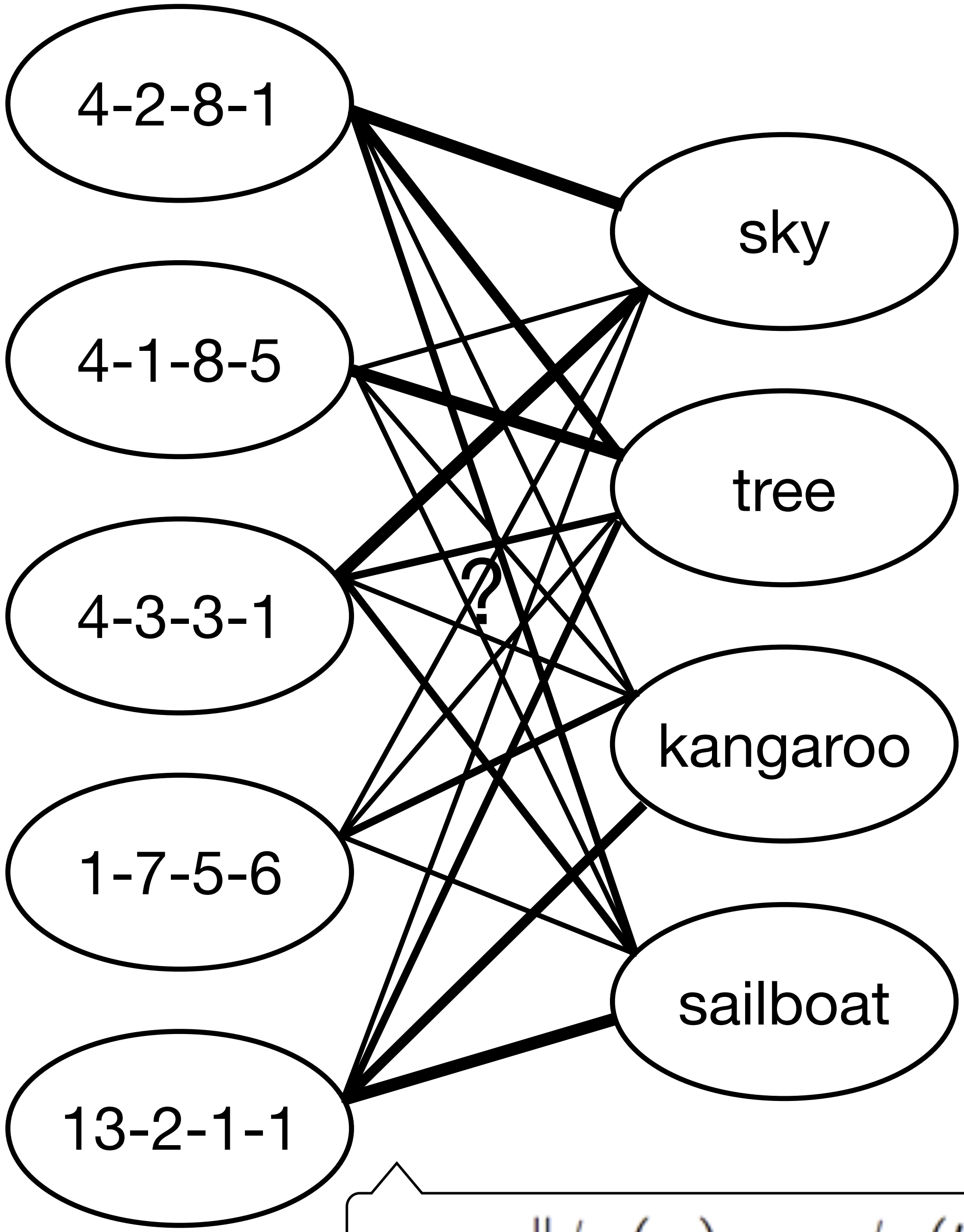
Can be cast as eigenvalue problem

Adapted from: A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein, “Learning Bilingual Lexicons from Monolingual Corpora.,” *ACL*, 2008.

Kernels

- Visual features:
  - Product of linear context kernel and chi-squared kernels for each the color, position, texture, and shape features.
- Textual features:
  - Product of linear context kernel and linear adjective kernel.

E-step



$$c_{i,j} = \|\phi_V(v_i)w_v - \phi_T(t_j)w_t\|_2$$

- Computing expected value over all mapping pairs is intractable
- Instead, do hard EM and take *k* best mapping pairs

$$M_{\mathbf{new}} = \underset{M_{1:k}}{\mathbf{argmax}} \log p(V, T, M; \xi)$$

- Approximate with weighted matching of bipartite graph
- Add new mapping pairs to kCCA training set and repeat

Strengths/Weaknesses of Approach

Strengths

- Little reliance on labeled image
- Bootstraps visual-text mapping starting with only the initial seed set
- Probabilistic model

Weaknesses

- Visual features are relatively simple; spatial relationships not preserved
- Sensitive to choices about visual word clustering
- May not generalize to infrequent visual words
- Many approximations in E-step

Evaluation

- Three components:
  1. Justification of method for selecting visual word clusters by balancing purity and frequency
  2. Experimental comparison of annotation and segmentation performance against several other models.
  3. Exploration of performance of the model under various settings of training set size and text label size.

Visual Word Clustering

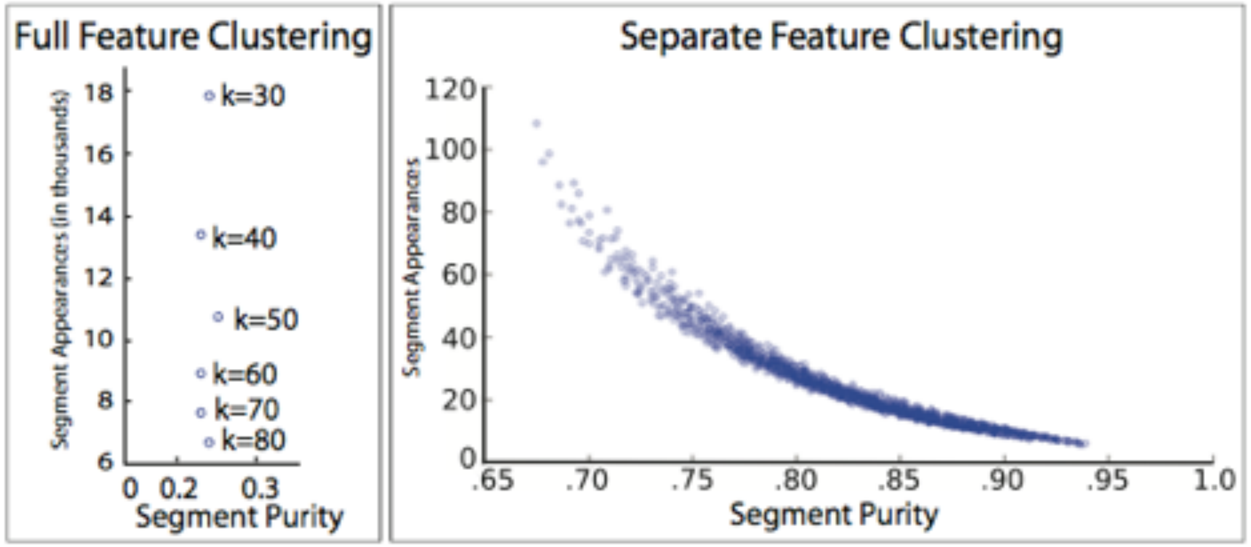
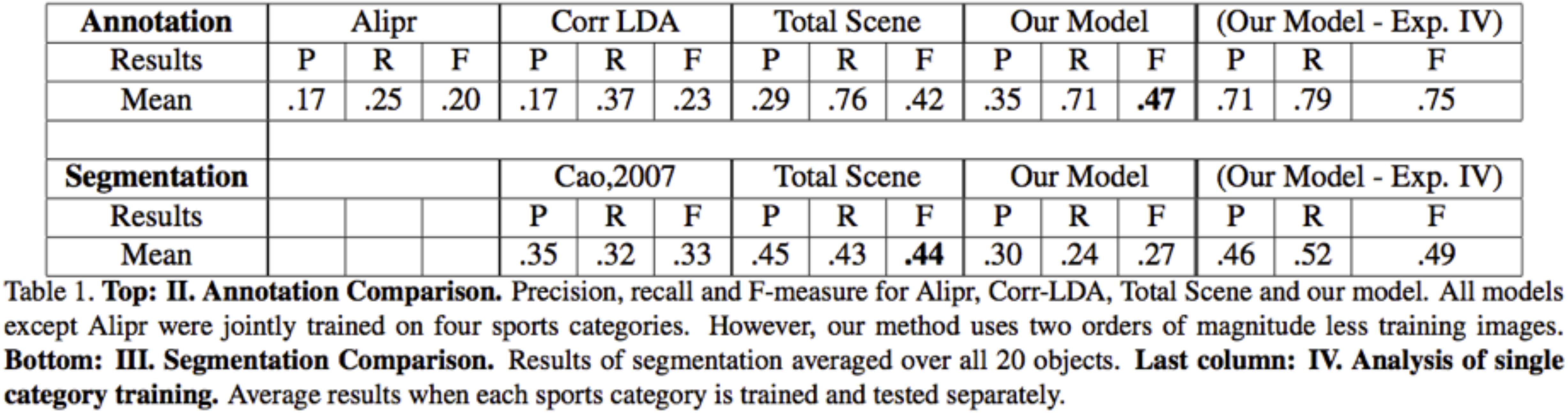


Figure 3. **I. Analysis of visual word clustering.** Each datapoint corresponds to the average statistics of a segment clustering. **Left: I.(a).** All segment features are concatenated and clustered into k clusters. The figure shows the number of times a word of that cluster appears vs how pure the corresponding labels of this cluster are. Notice that no such clustering provides very pure words. **Right: I.(b).** Results from using our method of clustering different feature types separately and concatenating them (see text for details). This region-based representation gives much flexibility in the trade-off between frequency and purity.

- Strike balance between
  - Purity: a visual word should map to a single text label
  - Frequency: each visual word should be observed multiple times in the data.
- Concatenating and then clustering features yields low purity.
- Clustering first then concatenating provides a continuum between purity and frequency.

Annotation & Segmentation

- Dataset of 4 sports categories (badminton, rowing, sailing and snowboarding)
  - Images from searching flickr.com
  - Articles from the New York Times corpus
- Restrict set of text labels to those used in previous work
- Train with 4 x 5 images and test with 4 x 25
- Segmentation: precision computed on pixelwise per class level



Annotation	Alipr			Corr LDA			Total Scene			Our Model			(Our Model - Exp. IV)		
Results	P	R	F	P	R	F	P	R	F	P	R	F	P	R	F
Mean	.17	.25	.20	.17	.37	.23	.29	.76	.42	.35	.71	**.47**	.71	.79	.75
Segmentation				Cao,2007			Total Scene			Our Model			(Our Model - Exp. IV)		
Results				P	R	F	P	R	F	P	R	F	P	R	F
Mean				.35	.32	.33	.45	.43	**.44**	.30	.24	.27	.46	.52	.49

Table 1. **Top: II. Annotation Comparison.** Precision, recall and F-measure for Alipr, Corr-LDA, Total Scene and our model. All models except Alipr were jointly trained on four sports categories. However, our method uses two orders of magnitude less training images. **Bottom: III. Segmentation Comparison.** Results of segmentation averaged over all 20 objects. **Last column: IV. Analysis of single category training.** Average results when each sports category is trained and tested separately.

Influence of Training Set Size and Text Labels

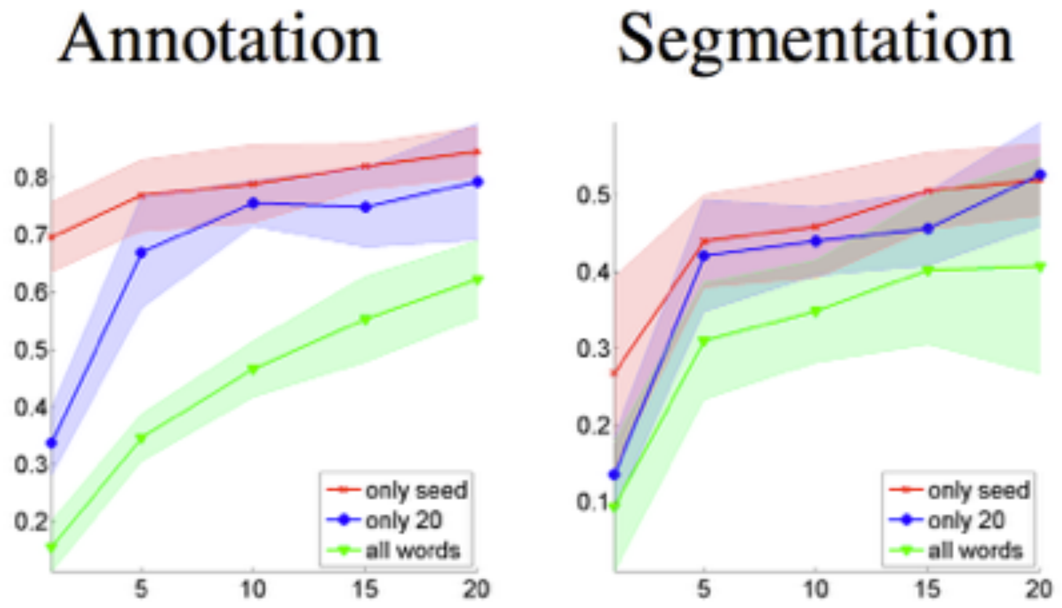


Figure 5. **V. Influence of the number of training images and possible text labels.** Average F-measures and standard deviation for different numbers of training images (x-axis) and different pools of textual words that may participate in the mapping. 5 sets of randomly chosen training images were used for each setting.

- More training images leads to better performance
- Better to restrict text labels if possible, but this can be overcome by adding more training images

Sample Segmentations

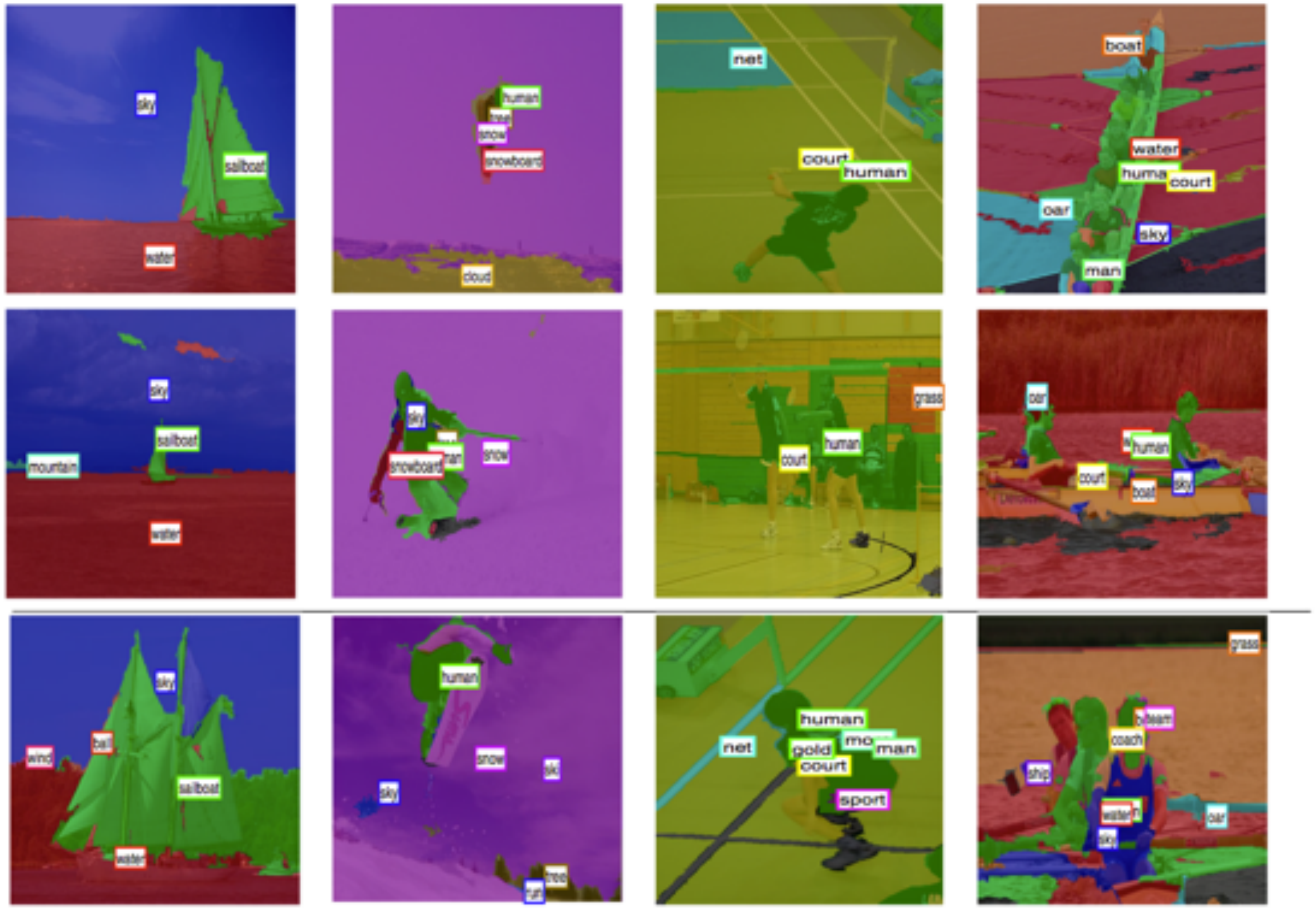


Figure 6. **Top two rows: IV. Analysis of single category training.** Results of annotation and segmentation of the test set. Labels are shown in boxes and the corresponding regions are overlayed with the same color as the boundary box. **Bottom row: V. Results with mappings from all words of the text corpus.** If all words of the text corpus are allowed in mappings the evaluation becomes very hard. *Man* might replace the *human* label in badminton images. *Wind* might show up in front of a *sailboat* etc.

Strengths/Weaknesses of Evaluation

Strengths

- Justification of visual word selection
- Exploration of behavior of model under various training settings.

Weaknesses

- No evaluation on standard segmentation benchmark
- Training settings are not comparable across models
- Single category training gets good results but other models are not evaluated under this setting.

Discussion

- How can we improve the visual and text features in this model?
- Some other multi-modal approaches dispense with discrete mappings and instead focus on a ranking loss in the latent space. Is the discrete mapping a feature or a weakness of this model?
- Current state-of-the-art approaches for segmentation get around the problem of small labeled data by leveraging convnets trained for image classification. Does this solve the problem or is there still more to be gained by exploring the relationship between images and text?