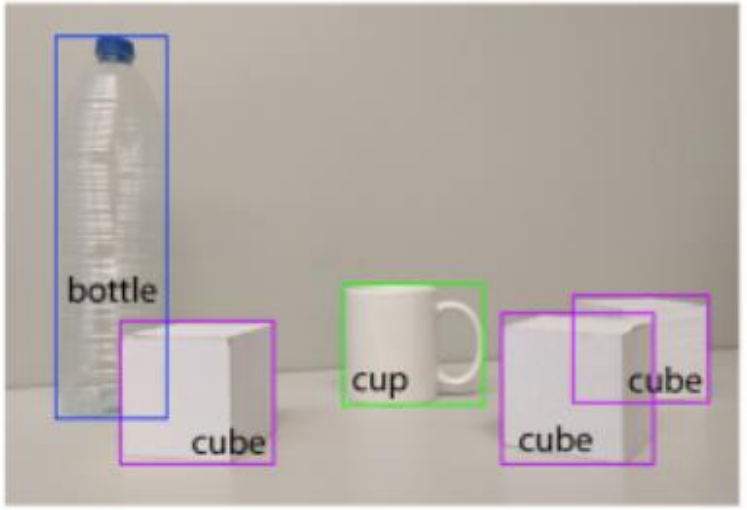
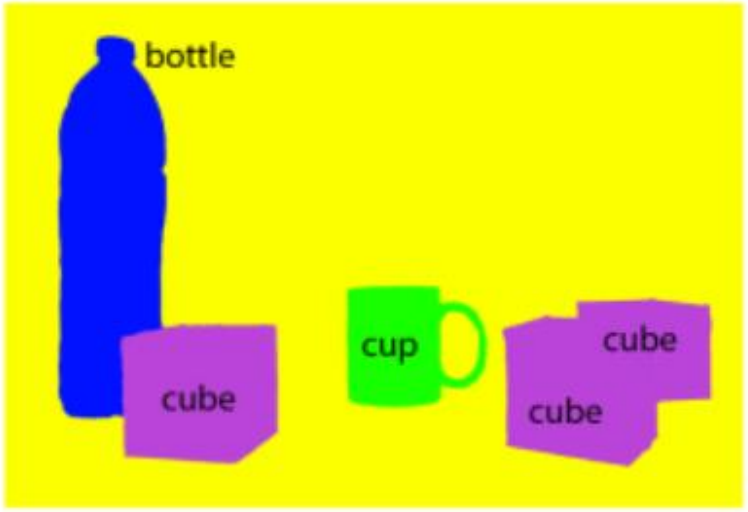
Active Learning in Semantic Segmentation

Problem Description

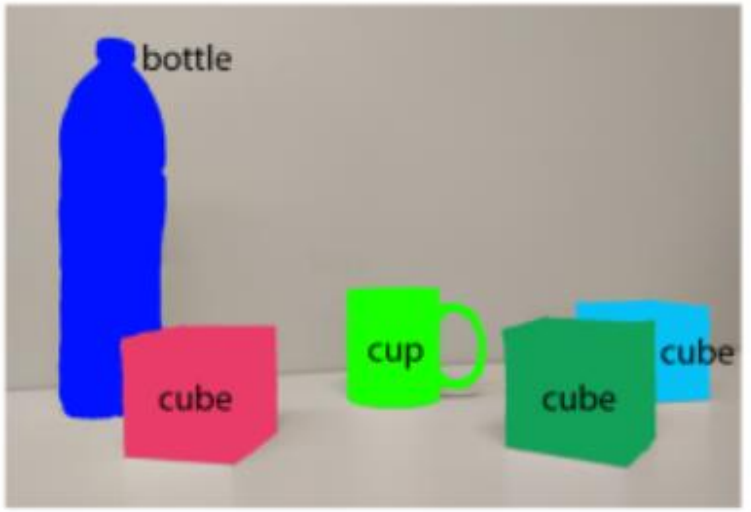
(a) Image classification



(b) Object localization



(c) Semantic segmentation



(d) Instance segmentation

Combining Generative and Discriminative Models for Semantic Segmentation of CT Scans via Active Learning

Juan Eugenio Iglesias[1,2], Ender Konukoglu[2], Albert Montillo[3,2], Zhuowen Tu[1], and Antonio Criminisi[2]

IPMI-2011

Intuition

- **A combined generative-discriminative model**

(a) a classifier that focuses on the appearance of the organs(object);

(b) a generative model, which captures organ(object) relative location and thus global, probabilistic shape information.

- **Unlabeled Selection**

QBC（query by committee）

Methods

■ Discriminative Voxel Classification

Following the work in [7], we have applied random forest classification to the task of assigning organ class probabilities to all voxels of a previously unseen CT scan.

The classifier is based on **box-shaped visual features**.

Problem:
Although random forests do capture some level of context, they fail at modeling the long-range spatial relationships between organs.

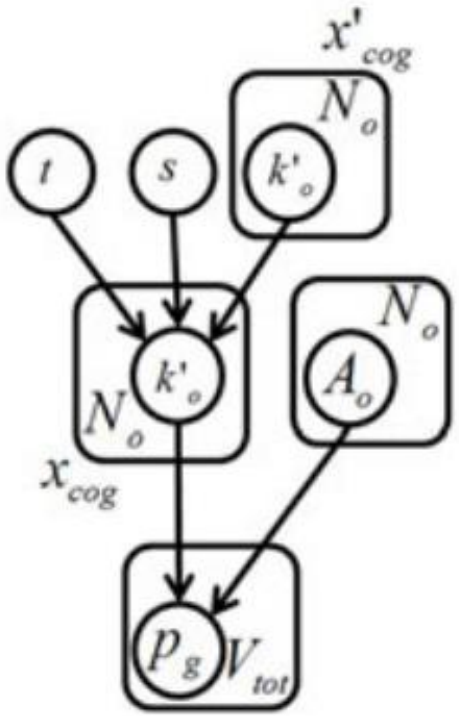
[7] Geremia, E., Menze, B.H., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N.:
Spatial decision forests for MS lesion segmentation in multi-channel MR images.
In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010.
LNCS, vol. 6361, pp. 111–118. Springer, Heidelberg (2010)

Methods

■ Generative Model of CT Scans

A generative graphical model which captures relative organ positions and organ shapes probabilistically. Each organ is represented by its centroid location k_o and a probabilistic atlas of shape $A_o(r)$ such that the probability that the voxel at location r is inside the organ is $A_o(r - k_o)$

$$\theta = \{s, b, t, \varepsilon\}$$



Methods

■ Generative Model of CT Scans

two coordinate systems in the model $\theta = \{s, b, t, \varepsilon\}$

A reference frame in which the sets of centroids from the training dataset are jointly aligned $\boldsymbol{x}'_{cog} = [\boldsymbol{k}'^{t}_{1}, \ldots, \boldsymbol{k}'^{t}_{N_o}]^t$ follow a multivariate Gaussian distribution: $\boldsymbol{x}'_{cog} \sim \mathcal{N}(\bar{\boldsymbol{x}}'_{cog}, \Sigma_{cog})$

A physical coordinate system in which all CT scans are defined. These coordinates are mapped to the physical frame by a simple rigid transform $\boldsymbol{x}_{cog} = [\boldsymbol{k}^{t}_{1}, \ldots, \boldsymbol{k}^{t}_{N_o}]^t = s\boldsymbol{x}'_{cog} + \boldsymbol{t}$

Use probabilistic principal component analysis (PPCA [17]) to deal with missing data $\boldsymbol{x}'_{cog} = \bar{\boldsymbol{x}}'_{cog} + \boldsymbol{\Phi b} + \boldsymbol{\varepsilon}$ where Φ is the matrix with the orthonormal principal components

$$p_{g,o}(\boldsymbol{r}) = \frac{1}{Z_g(\boldsymbol{r})} A_o(\boldsymbol{r} - \boldsymbol{k}_o) \prod_{o'=1, o' \neq o}^{N_o} [1 - A_{o'}(\boldsymbol{r} - \boldsymbol{k}_{o'})], \quad o \in [1, \ldots, N_o]$$

$$p_{g,Nc}(\boldsymbol{r}) = \frac{1}{Z_g(\boldsymbol{r})} \prod_{o=1}^{N_o} [1 - A_o(\boldsymbol{r} - \boldsymbol{k}_o)] \quad \text{(for the background)}$$

Methods

■ Segmenting Previously Unseen Scans

1. Fitting the generative model to the classifier output
2. Combining the models via Bayes` theorem

Obtain two models $p_g(r)$ and $p_d(r)$

Minimize the **Jensen Shannon divergence**

$J_S(P||Q) = (1/2)[KL(P||R) + (KL(Q||R)]$, where $R = (P + Q)$

$$\boldsymbol{\theta}^* = \{s^*, \boldsymbol{b}^*, \boldsymbol{t}^*, \boldsymbol{\varepsilon}^*\} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{\boldsymbol{r}} JS(\boldsymbol{r}, \boldsymbol{\theta}) =$$

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{2V_{tot}} \sum_{\boldsymbol{r}} \sum_{c=1}^{N_c} \left(p_{g,c}(\boldsymbol{r}) \log \frac{2p_{g,c}(\boldsymbol{r})}{p_{g,c}(\boldsymbol{r}) + p_{d,c}^{\boldsymbol{\theta}}(\boldsymbol{r})} + p_{d,c}^{\boldsymbol{\theta}}(\boldsymbol{r}) \log \frac{2p_{d,c}^{\boldsymbol{\theta}}(\boldsymbol{r})}{p_{g,c}(\boldsymbol{r}) + p_{d,c}^{\boldsymbol{\theta}}(\boldsymbol{r})} \right)$$

Methods

■ Segmenting Previously Unseen Scans

1. Fitting the generative model to the classifier output
2. Combining the models via Bayes` theorem

Bayesian semantic segmentation

The aligned generative model can be interpreted as a location prior in a Bayesian framework. The posterior probability of label L at location r is therefore given by Bayes' theorem

$$p[L(\boldsymbol{r}) = c] = \frac{p_{d,c}(\boldsymbol{r}) \cdot p_{g,c}^{\boldsymbol{\theta}^*}(\boldsymbol{r})}{Z_L(\boldsymbol{r})}$$

Methods

■ Training Set Construction through Active Learning

At each iteration an expert labels the scan which maximizes the disagreement between the **discriminative** and **generative** models

$$JS_w := \frac{1}{2}\sum_{c=1}^{N_C}\boxed{\frac{1}{V_c}}\sum_{\boldsymbol{r}}\left(p_{d,c}(\boldsymbol{r})\log\frac{2p_{d,c}(\boldsymbol{r})}{p_{d,c}(\boldsymbol{r})+p_{g,c}^{\boldsymbol{\theta}^*}(\boldsymbol{r})}+p_{g,c}^{\boldsymbol{\theta}^*}(\boldsymbol{r})\log\frac{2p_{g,c}^{\boldsymbol{\theta}^*}(\boldsymbol{r})}{p_{d,c}(\boldsymbol{r})+p_{g,c}^{\boldsymbol{\theta}^*}(\boldsymbol{r})}\right)$$

Making the utilities inversely proportional to the average volumes of the organs to remove the bias towards larger organs

$$V_c = \sum_{\boldsymbol{r}} A_c(\boldsymbol{r}), c \in [1, \ldots, N_o]$$

Algorithm

1. The generative and discriminative models are built starting with 2 labeled scans.
2. The remaining unlabeled scans are fed to the classifier, yielding multi-class probability maps for each voxel.
3. Align the generative model by minimizing JS in (1).
4. Compute disagreement via the weighted JS divergence as in (3).
5. Rejection of outlying scans via the local outlier factor (LOF) [4] on JS_w.
6. Select the unlabeled scan that maximizes JS_w and obtain its manual ground truth from a human expert.
7. Update the classifier and the generative model.
8. If the testing segmentation accuracy is satisfactory then stop. Otherwise, goto 2.

[4]Breunig, M., Kriegel, H., Ng, R., Sander, J.: **LOF: identifying density-based local outliers.** Sigmod Rec. 29(2), 93–104 (2000)

Experiment

Setup

1. the accuracy of our Bayesian segmentation approach versus the discriminative classifier alone
2. the validity of our database construction strategy as compared to alternative techniques

- Dataset

  196 clinical CT scans

  Two scans are randomly selected to form the initial training set

Experiment

Comparing algorithm

A1. The proposed active learning approach.

A2. Randomly selecting scans from the unlabeled pool.

A3. Uncertainty sampling [8], in which the scan that maximizes the mean voxel entropy is selected. Our generative model is thus not used for data selection, but it is still used in segmentation.

A4. Same as 2, but the generative model is replaced by a generic Markov Random Field (MRF) prior in segmentation.

A.5 Same setup as in 2, but without any generative model at all the MAP estimate

8. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: Proc. ACM SIGIR Conf. Res. and Dev. in Inf., pp. 3–12 (1994)

Experiment

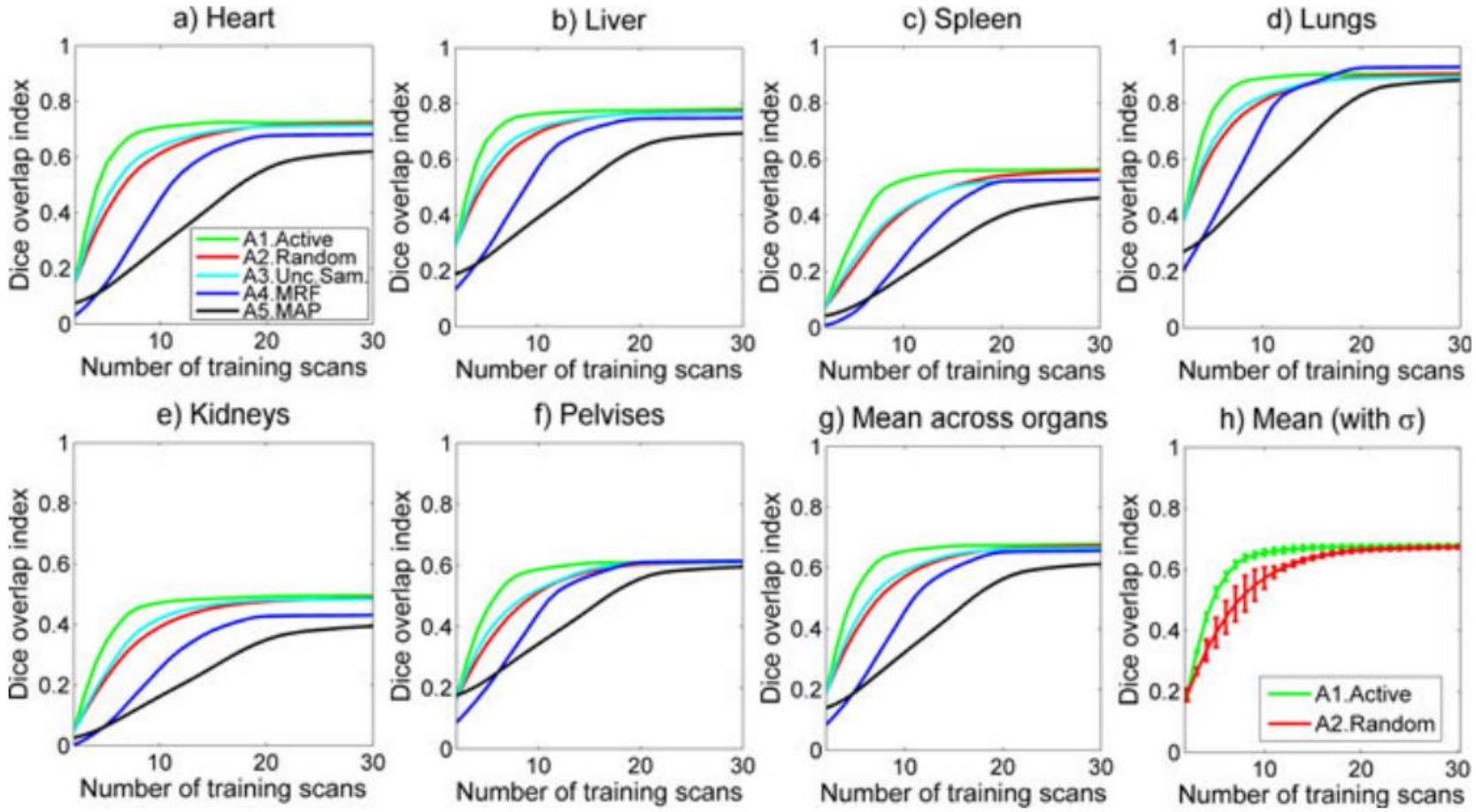


Fig. 4. Segmentation accuracy (Dice) vs. training set size for different database construction approaches. Plots a-f) are organ specific. Plot g) displays the mean for all organs. Plot h) is a zoom-in of f) that displays standard deviations (but only shows random selection and active learning for clarity).

Experiment

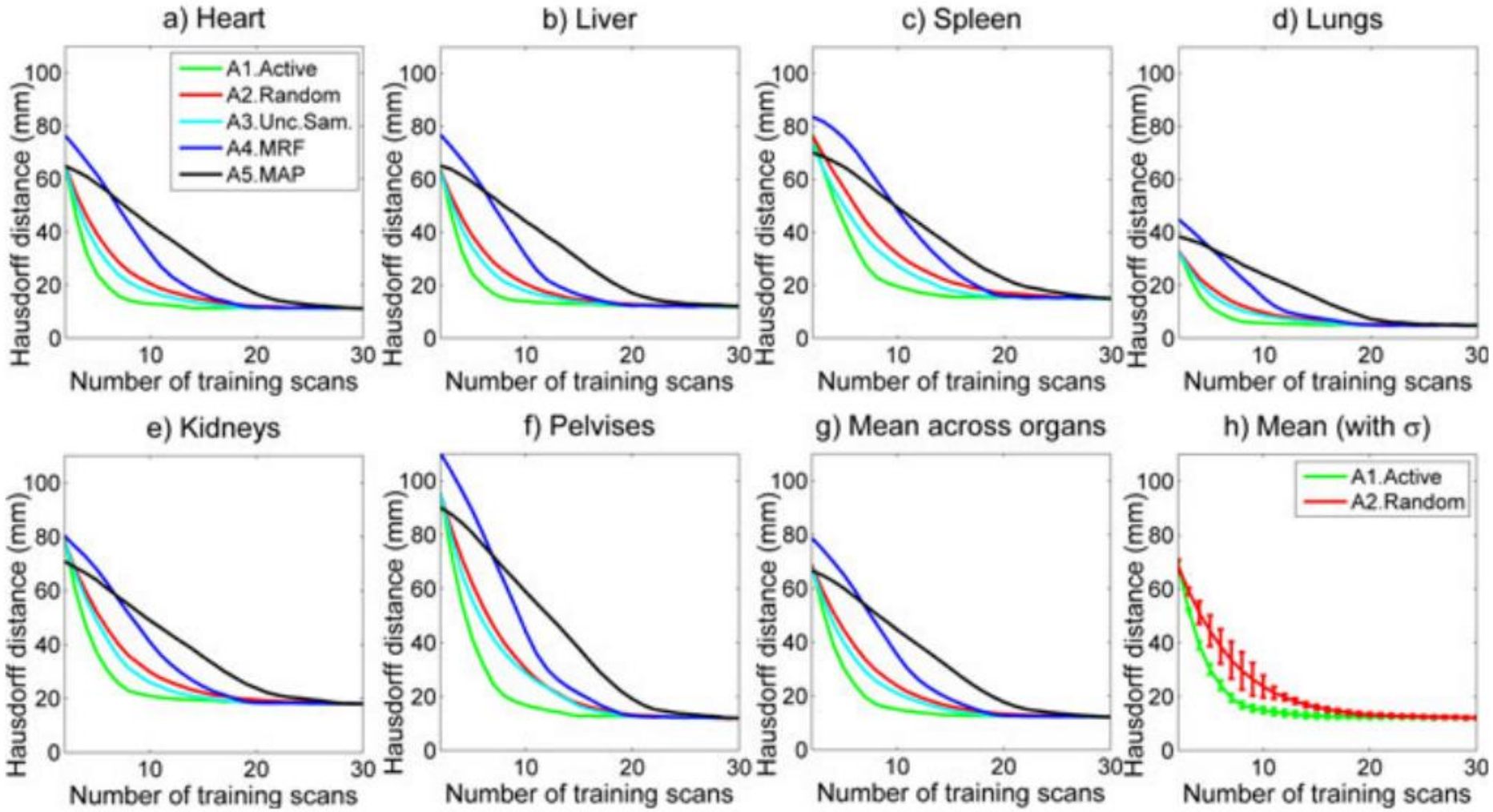


Fig. 5. Haussdorff distance vs. training set size for different database construction approaches. See caption of Figure 4 for the explanation of the plots.

Experiment

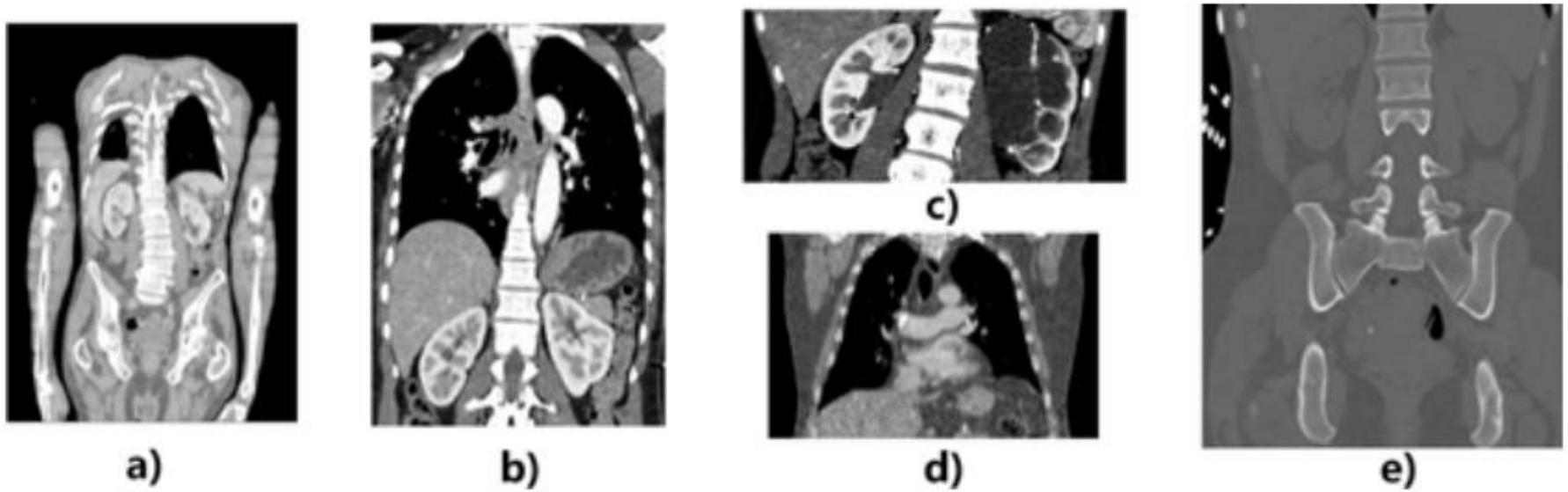


Fig. 7. Active selection of training scans. a-b) initial training set. c) Scan with minimal weighted JS score at the first iteration, which displays a kidney with a large cyst and is rejected by LOF. d-e) Scans actually selected in the first two iterations.

Because of the cyst in the left kidney and the lack of context due to reduced field of view. Adding it to the training set could negatively affect the performance of the system

Active Learning for Semantic Segmentation with Expected Change

Alexander Vezhnevets[1] Joachim M. Buhmann[1] Vittorio Ferrari[2]
[1]ETH Zurich
Zurich, Switzerland
[2]The University of Edinburgh
Edinburgh, UK

CVPR-2012

Problem Setting

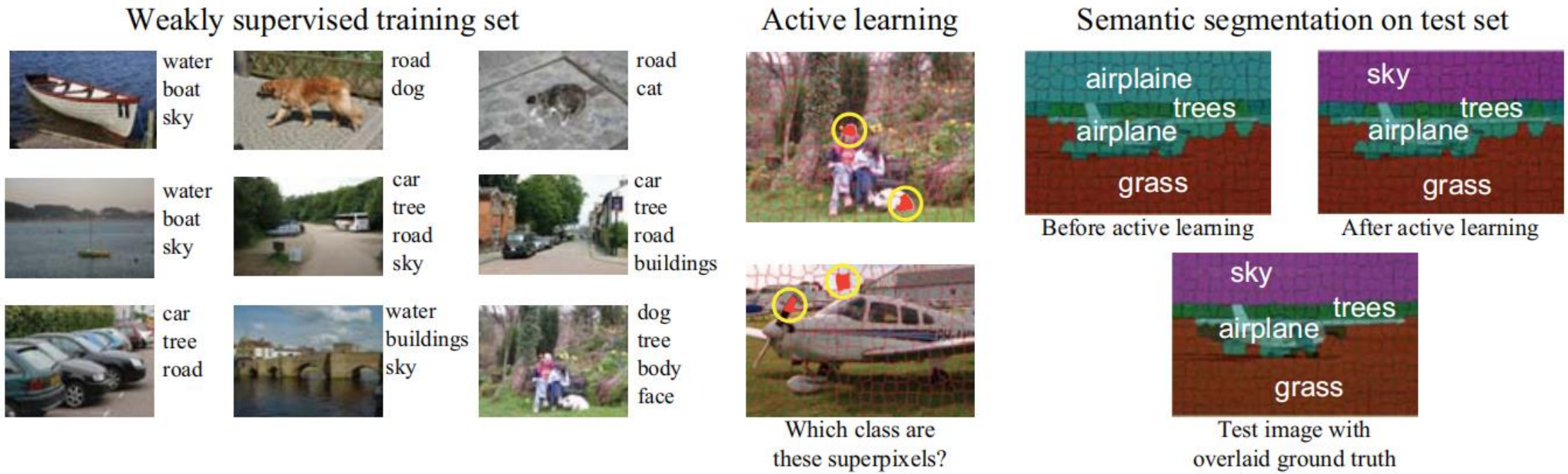


Figure 1. **weakly supervised semantic segmentation with active learning.** The input is a weakly supervised training set, where images are labeled by the classes they contain. The active learning proceeds by querying an oracle for the true label of superpixels selected by a specialized criterion. We use acquired information to classify superpixels in previously unseen test images.

Intuition

Model the problem as **a pairwise CRF** and cast active learning as finding its most informative nodes.

1. Train a weakly supervised model Multi-Image Model (MIM) [6] to recover a first approximation of these labels.
2. Run a active learning algorithm which queries the oracle for the true state of a few latent variables selected by a novel criterion.

When the true state of a variable is revealed, it induces change to the state of other variables as well. These changes propagate locally through the pairwise potentials of the CRF, as well as globally through the unary potential.

[6] A. Vezhnevets, V. Ferrari, and J. Buhmann, "Weakly supervised se
mantic segmentation with a multi image model," in *ICCV*, 2011.

Methods

- **weakly supervised semantic segmentation with a pairwise CRF**

$\tau = \left\{ I^j = \left(\{x_i^j\}_{i=1}^{N_j}, Y^j \right) \right\}_{j=1}^{N}$ be the training set, where image I_j consists of **superpixels** x_i^j. $Y^j \subset Y$, all possible labels $Y = \{1, \ldots, C\}$

Model:

Model the weakly labeled training set as a CRF, where nodes correspond to the latent superpixel labels. The total energy E of the model is a function of these labels y_i^j and appearance model parameters θ

$$\mathcal{E}\left(\{y_i^j\}, \theta\right) = \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j) \right) + \sum_{(y_i^j, y_{i'}^{j'}) \in E} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) \quad (1)$$

Methods

- **weakly supervised semantic segmentation with a pairwise CRF**

Model as CRF:

$$\mathcal{E}\left(\{y_i^j\},\theta\right) = \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right) + \sum_{(y_i^j, y_{i'}^{j'}) \in E} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) \quad (1)$$

The first unary potential $\psi\left(y_i^j, x_i^j, \theta\right)$ measures how well the appearance of x_i^j matches the appearance model $\theta_{y_i^j}$ of class y_i^j

$$\pi(y_i^j, Y_i^j) = \begin{cases} \infty & y_i^j \notin Y^j \\ 0 & y_i^j \in Y^j \end{cases}$$

makes sure that a superpixel can only take a label from the label set Y_j of the image

Methods

- **weakly supervised semantic segmentation with a pairwise CRF**

Model as CRF:

$$\mathcal{E}\left(\{y_i^j\},\theta\right) = \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right) +$$

$$\sum_{(y_i^j, y_{i'}^{j'}) \in E} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) \qquad (1)$$

The pairwise potential φ encourages connected superpixels to take the same label if their appearance similarity is high

$$\phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) = \begin{cases} 1 - D\left(x_i^j, x_{i'}^{j'}\right) & y_i^j \neq y_{i'}^{j'} \\ 0 & y_i^j = y_{i'}^{j'} \end{cases}$$

$D()$ is a similarity metric, $[0,1]$

Methods

■ Active Learning

The **Expected Change**(EC) score of y_i^j is defined as

$$EC(i,j) =$$

$$\frac{1}{|Y^j|}\sum_{l\in Y^j}\sum_{i'\neq i, j'\neq j} w_i^j \left[F\left(x_{i'}^{j'}, \theta^t | y_i^j = l\right) \neq F\left(x_{i'}^{j'}, \theta^t\right)\right]$$

$F(x,\theta)$ be the output (a label) of the CRF for a training superpixel x, with the appearance model parameter vector θ.

Here w_i^j is the importance weight of y_i^j , i.e. the number of pixels in superpixel x_i^j

Algorithm

Algorithm 1 Generic active learning procedure

Input: Training set τ, initial parameters θ^0, initial labeling $L^0 = \{y_i^j = F(x_i^j, \theta^0)\}$, maximum number M of queries to the oracle O, query scoring function S, query selection rule U.
Output: updated labeling and parameters θ^*

1. $t = 0$ and $m = 0$
2. **while** $m < M$
   (a) **for each** unknown latent variable y_i^j, evaluate $S(i, j)$
   (b) Select query set Q with selection rule U
   (c) Query the oracle for the labels $l_i^j = O(i, j),\ \forall (i, j) \in Q$
   (d) Set $y_i^j = l_i^j\ \forall (i, j) \in Q$
   (e) Retrain appearance models θ^{t+1} and infer latent variable labels
   (f) $m = m + |Q|$ and $t = t + 1$
3. return $\theta^* = \theta^T$ and latest labeling of the training set $L^T = \{y_i^j = F(x_i^j, \theta^T)\}$

Algorithm 2 Evaluating expected change

Input: Training set τ, current parameters θ, current labeling $L = \{y_i^j = F(x_i^j, \theta)\}$
Output: EC scores for each latent variable

1. **for each** latent variable y_i^j
   (a) **for each** admissible label $l \in Y^j$
      i. retrain appearance model parameters $\theta' = (\theta^t | y_i^j = l)$
      ii. infer MAP labeling with unary potentials $\psi\left(y_i^j, x_i^j, \theta'\right)$ and y_i^j clamped to l
      iii. record change $C(y_i^j, l) = \sum_{i' \neq i, j' \neq j} w_{i'}^{j'} \left[F\left(x_{i'}^{j'}, \theta^t | y_i^j = l\right) \neq F\left(_{i'}^{j'}, \theta^t\right)\right]$
   (b) set $EC(i, j) = \frac{1}{|Y^j|} \sum_{l \in Y^j} C(y_i^j, l)$
2. return EC

Experiment

- **Dataset**

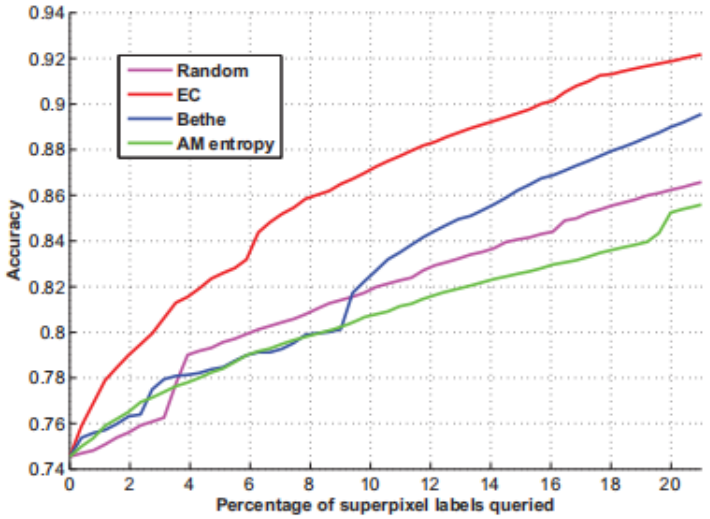
  MSRC-21 LabelMe

- **Baselines**

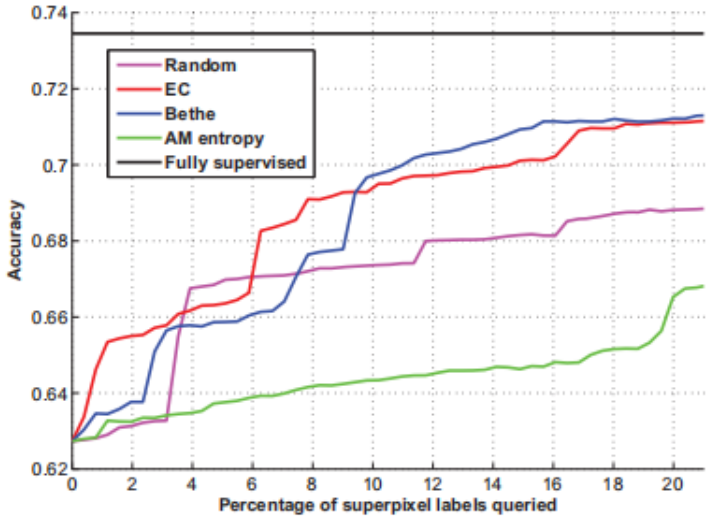
1. random sampling
2. the uncertainty sampling by the entropy of the unary potential(this score is calculated from the outputs of Naive Bayes appearance models)
3. the third criterion samples according to maximal uncertainty measured by the Bethe entropy [28] of the full CRF

[28] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algo rithms.," *IEEE Transactions on Information Theory*, pp. 2282–2312, 2005

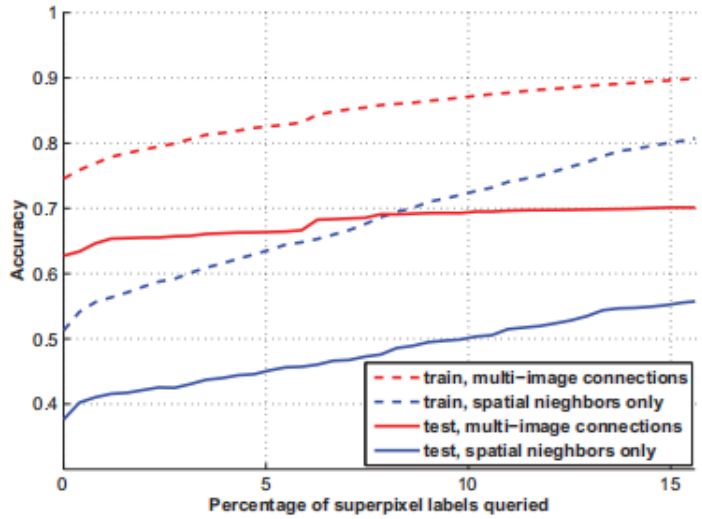
Experiment



(a) Weakly labeled training set



(b) Test set



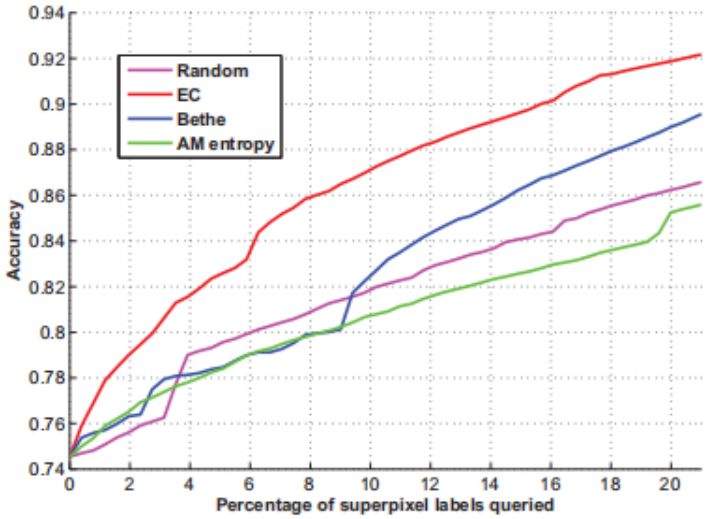
(c) With and without multi-image connections

Figure 3. Results on MSRC data-set. Plotted is the accuracy over share of queries asked: a) accuracy on weakly supervised training set; b) accuracy on test set; c) comparison between a CRF with only pairwise connections between spatial neighbours within one image vs also including multi-image connections.

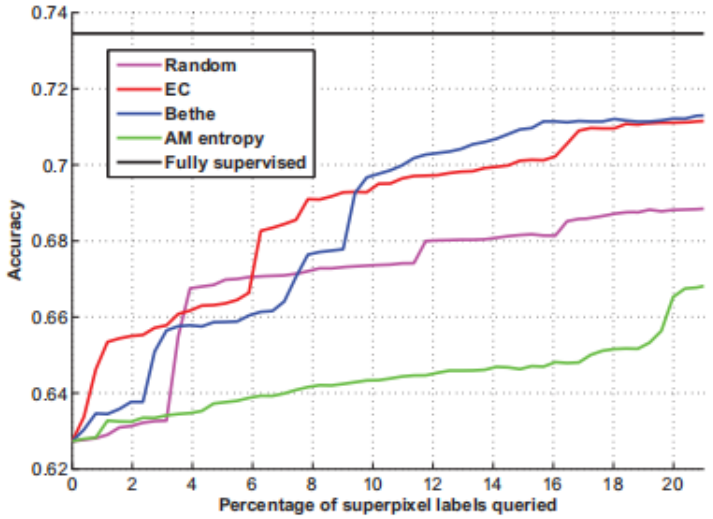
$$\mathcal{E}\left(\{y_i^t\}\right) = \sum_i \left(\psi\left(y_i^t, x_i^t, \theta^*\right) + \mu\left(y_i^t, I^t\right)\right) +$$

$$\sum_{(y_i^t, y_{i'}^j) \in S} \phi\left(y_i^t, y_{i'}^t, x_i^t, x_{i'}^t\right) + \sum_{(y_i^t, y_{i'}^j) \in M^t} \phi\left(y_i^t, y_{i'}^j, x_i^t, x_{i'}^j\right)$$

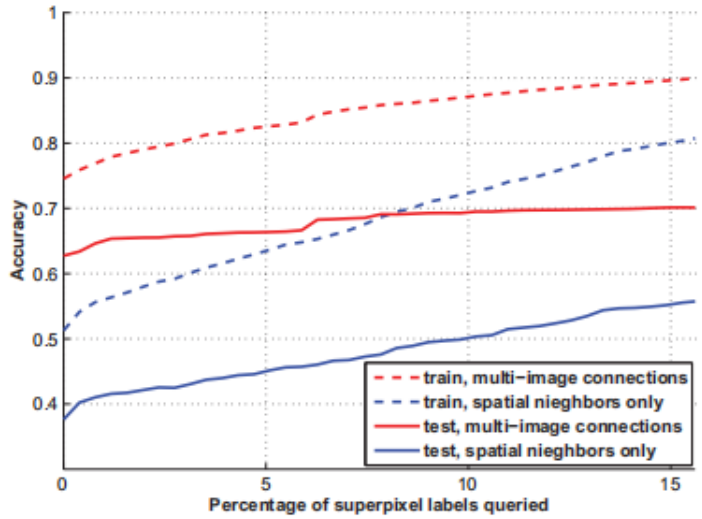
Experiment



(a) Weakly labeled training set



(b) Test set

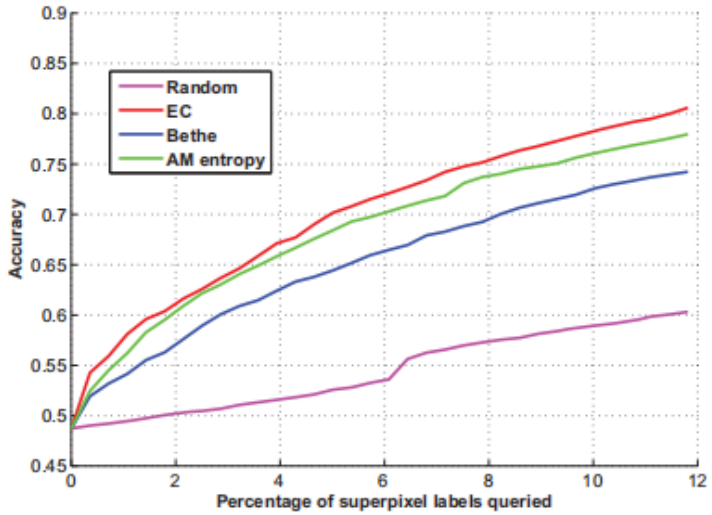


(c) With and without multi-image connections

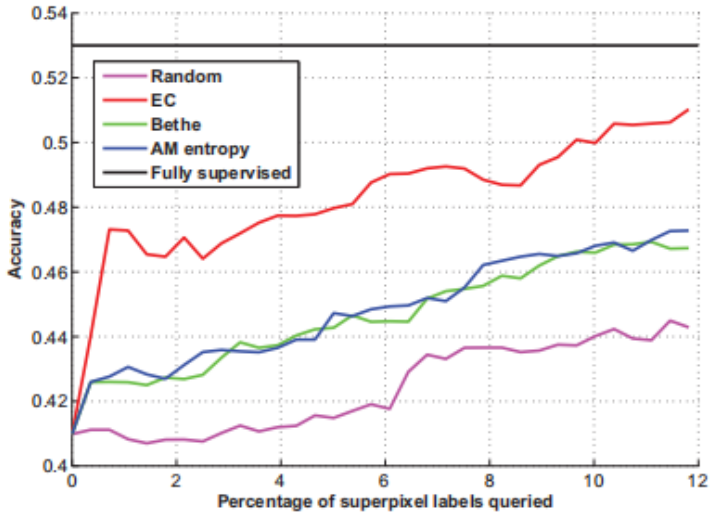
Figure 3. Results on MSRC data-set. Plotted is the accuracy over share of queries asked: a) accuracy on weakly supervised training set; b) accuracy on test set; c) comparison between a CRF with only pairwise connections between spatial neighbours within one image vs also including multi-image connections.

This effect might be due to the latter method neglecting the CRF connections, thereby over-fitting to unary potential uncertainty. Moreover, the good performance of the random strategy suggests this data-set to be fairly simple.

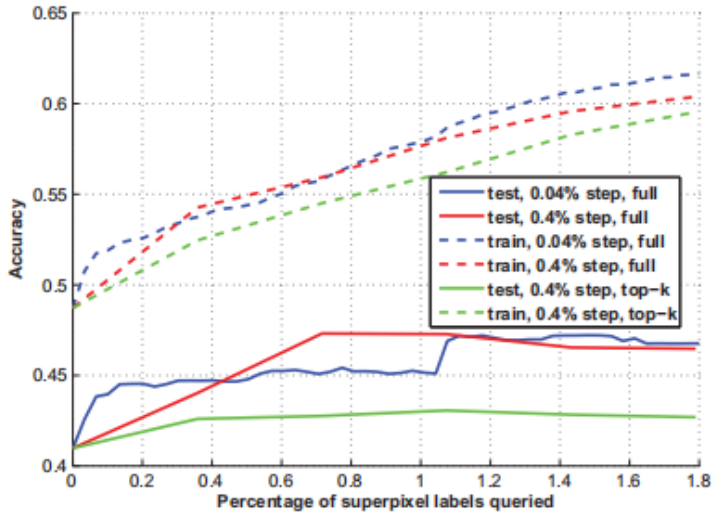
Experiment



(a) Weakly labeled training set



(b) Test set



(c) Batch query variations

Figure 4. Results on LabelMe data-set. Plotted is the accuracy over share of queries asked: a) accuracy on weakly supervised training set; b) accuracy on test set; c) variations of the batch query scheme: our full scheme from sec. 4.2 (0.4%, full), a scheme with a 10× smaller batch size (0.04%), and a simplified scheme where the top-k superpixel labels are queried, without taking into account their influence areas.

Random sampling performs much worse then other criteria, confirming our judgement that this data-set is more challenging.

Active MAP Inference in CRFs for Efficient Semantic Segmentation

Gemma Roig[1] * Xavier Boix[1] *

Roderick de Nijs[2] Sebastian Ramos[3] Kolja Kühnlenz[2] Luc Van Gool[1,4]

[1]ETH Zürich, Switzerland [2]TU Munchen, Germany [3]CVC Barcelona, Spain [4]KU Leuven, Belgium

* *Both first authors contributed equally.* {boxavier,gemmar}@vision.ee.ethz.ch

ICCV-2013

Intuition

■ Problem Description (Applicable Scene)

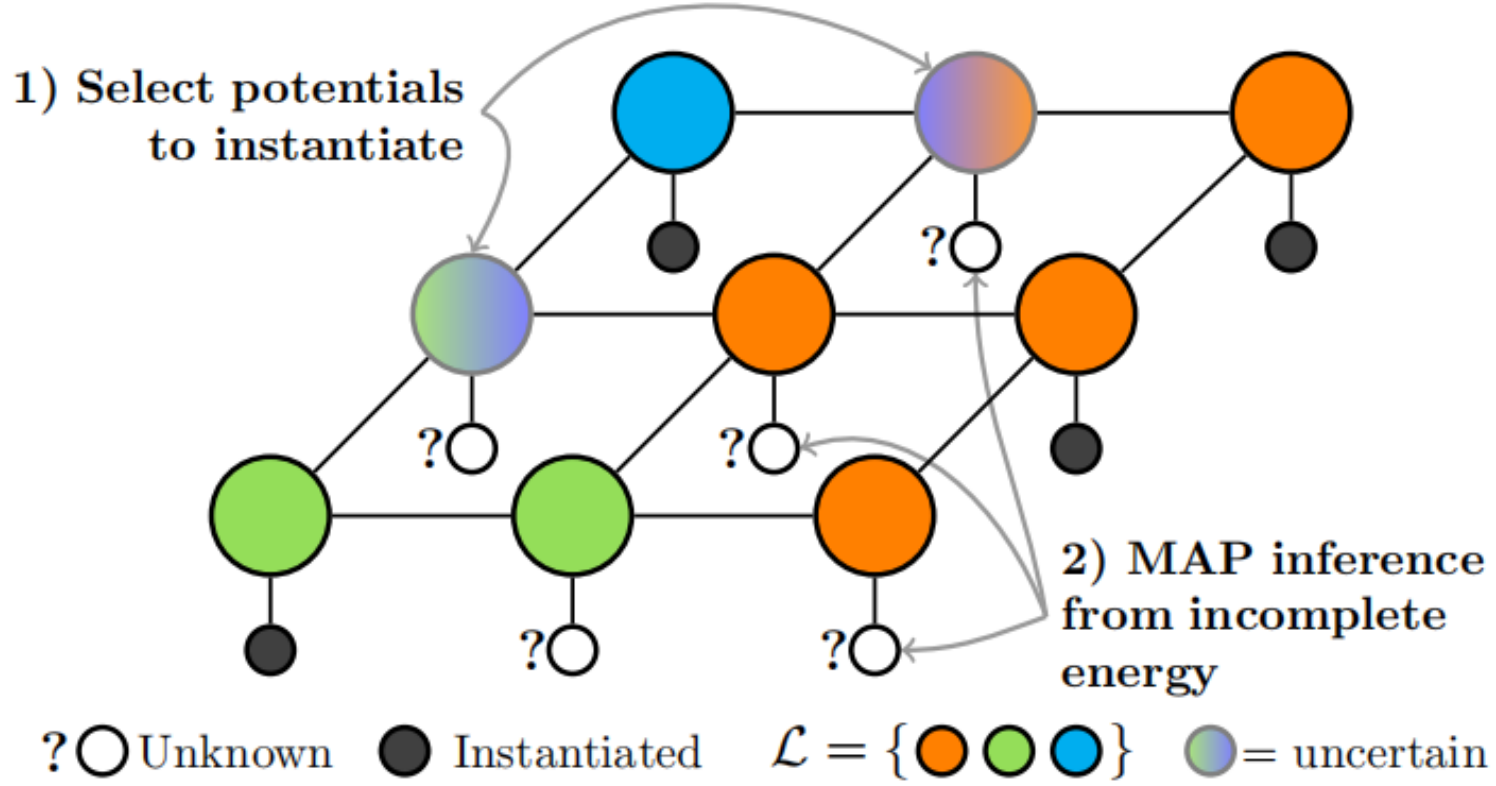
Most MAP inference algorithms for CRFs optimize an energy function knowing all the potentials.

In this paper focus on CRFs where the computational cost of instantiating the potentials is orders of magnitude higher than MAP inference. This is often the case in semantic image segmentation, where most potentials are instantiated by slow classifiers fed with costly features.

Intuition

■ Active MAP inference

1. **Select a subset of potentials** to be instantiated in the energy function, leaving the rest of the parameters of the potentials unknown (**Active**)
2. Estimate the MAP labeling from such incomplete energy function



Preliminaries

■ Active MAP Inference in CRFs

$G = (V, E)$ be the graph that represents the distribution

Denote $P(x|\theta)$ as the probability density distribution of a labeling modeled with the graph G

$E_\theta(x) = \theta^T \varphi(x)$, in which $\varphi(x) = (\varphi_1(x), \dots, \varphi_M(x))T$ is the vector of potentials

The most probable state x^* is obtained by inferring the Maximum a Posteriori (MAP) of $P(x|\theta)$,

or equivalently by minimizing the energy, i.e. $x^* = \arg\min_{x \in L^N} \theta^T \varphi(x)$

The parameters of θ related to the unary potentials are typically the result of evaluating classifiers fed with features extracted from the image.

Active MAP inference aims at estimating x^* with only a subset of the elements of $\theta, \{\theta_j\}$

$$\theta^j_{\delta^j} = \begin{cases} \theta^j & \text{if } \delta^j = 1 \\ \text{unknown} & \text{otherwise} \end{cases}$$

Preliminaries

■ Perturb-and-MAP [16]

PM is based on injecting noise in the energy function to perturb it, $\tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + \boldsymbol{\epsilon}$.

The different $\tilde{\theta}$s that yield the same MAP labeling x, $\mathcal{P}_{\mathbf{x}} = \left\{ \tilde{\boldsymbol{\theta}} \in \mathbb{R}^M | \mathbf{x} = \arg \min_{\mathbf{x}' \in \mathcal{L}^N} \tilde{\boldsymbol{\theta}}^T \boldsymbol{\phi}(\mathbf{x}') \right\}$

Define the set of perturbations ϵ, that yields the labeling x: $\left\{ \boldsymbol{\epsilon} \in \mathbb{R}^M | \mathbf{x} = \arg \min_{\mathbf{x}' \in \mathcal{L}^N} (\boldsymbol{\theta} + \boldsymbol{\epsilon})^T \boldsymbol{\phi}(\mathbf{x}') \right\}$

PM assigns a probability to x equal to the probability of drawing a perturbation ϵ that belongs to the set $P_x - \theta$. Thus, the PM distribution is $f_{PM}(\mathbf{x}|\boldsymbol{\theta}) = \int_{\mathcal{P}_{\mathbf{x}} - \boldsymbol{\theta}} f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon}) d\boldsymbol{\epsilon}$,

We can easily draw samples from a PM distribution by simply doing MAP inference on a perturbed energy

[16] G. Papandreou and A. Yuille. Perturb-and-map random fifields: Using discrete optimization to learn and sample from energy models. In *ICCV*, 2011.

Methods

■ MAP Inference for Incomplete Energies

$P(\theta|\theta_\delta)$ is the probability that the parameters of the potentials take the values θ given θ_δ.

Use a model to approximate it, referred to as $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\delta}, \boldsymbol{\pi})$

Use $P(\mathbf{X}^\star = \mathbf{x}|\boldsymbol{\theta}_{\boldsymbol{\delta}})$ to define probability on x^*, i.e. the probability that x is the MAP labeling.

$$\int_{\mathbb{R}^M} \mathbf{I}\left[\mathbf{x} = \arg\min_{\mathbf{x}' \in \mathcal{L}^N} E_{\boldsymbol{\theta}}(\mathbf{x}')\right] P(\boldsymbol{\theta}|\boldsymbol{\theta}_{\boldsymbol{\delta}}) d\boldsymbol{\theta}$$

It can be shown that $P(\mathbf{X}^\star = \mathbf{x}|\boldsymbol{\theta}_{\boldsymbol{\delta}})$ is indeed a PM random field, from which we can easily draw samples.

$$P(\boldsymbol{\theta}|\boldsymbol{\theta}_{\boldsymbol{\delta}}) \qquad \boldsymbol{\theta}|\boldsymbol{\theta}_{\boldsymbol{\delta}} \Longleftrightarrow \tilde{\boldsymbol{\theta}} = \boldsymbol{\theta} + \boldsymbol{\epsilon} \qquad f_{\boldsymbol{\epsilon}}(\boldsymbol{\epsilon})$$

Methods

■ Relation to Perturb-and-MAP

Proposition 1. $P(\mathbf{X}^\star = \mathbf{x}|\boldsymbol{\theta}_{\boldsymbol{\delta}}) = f_{PM}(\mathbf{x}|\boldsymbol{\delta}, \boldsymbol{\mu})$.

$$f_{PM}(\mathbf{x}|\boldsymbol{\delta}, \boldsymbol{\mu}) = \int_{\mathcal{P}_{\mathbf{x}} - \boldsymbol{\mu}} f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon} + \boldsymbol{\mu}|\boldsymbol{\delta}, \boldsymbol{\pi}) d\boldsymbol{\epsilon},$$

where $\mathcal{P}_{\mathbf{x}} - \boldsymbol{\mu}$ is the set of $\boldsymbol{\epsilon} \in \mathbb{R}^M$ such that x minimizes the energy function $E_{(\boldsymbol{\mu}+\boldsymbol{\epsilon})}$

Note also that we draw ϵ from $f_{\boldsymbol{\theta}}(\boldsymbol{\epsilon} + \boldsymbol{\mu}|\boldsymbol{\delta}, \boldsymbol{\pi})$ which is f_θ centered at 0.

To obtain samples of x^* in practice, we simply perturb μ using ϵ, and then, we apply MAP inference to $E_{(\boldsymbol{\mu}+\boldsymbol{\epsilon})}$

Methods

■ Model of the Missing Parameters

Use a simple collection of independent Gaussian variables to define $f_{\boldsymbol{\theta}}(\boldsymbol{\theta}|\boldsymbol{\delta}, \boldsymbol{\pi})$

1. If the parameter of the potential is unknown ($\delta_i = 0$), it is a univariate Gaussian distribution, centered at μ_i and deviation σ_i
2. Otherwise it is consistent with the instantiate potential, $f_\theta(\theta^i|\delta^i = 1, \pi^i) = \mathbf{I}[\theta^i = \theta^i_{\delta^i}]$

Set π to a fixed value that we learn by cross-validation. Thus, all $f_\theta(\theta^i|\delta^i = 0, \pi^i)$ are a Gaussian distribution with the same parameters.

Methods

■ Selection of δ

The algorithm starts from $\delta = 0$, and it sequentially determines which potential to compute next, until the time budget expires.

Define $S^i_{\boldsymbol{\delta}_t}$ as the expected reward of instantiating the potential i.

$$S^i_{\boldsymbol{\delta}_t} = \mathbb{E}_\theta \left[R \left(P \left(\mathbf{X}^\star = \mathbf{x} | \boldsymbol{\theta}_{\boldsymbol{\delta}_t} : \theta^i = \theta \right) \right) \right]$$

where the expected value is over,

$$\theta \sim f_\theta \left(\theta^i | \delta^i = 0, \pi^i \right)$$

which is the Gaussian model of the posterior $P(\theta^i | \boldsymbol{\theta}_{\boldsymbol{\delta}_t})$

Algorithm 1: Active MAP

$\boldsymbol{\delta}_0 = \mathbf{0}$;
while $t < t_{total}$ **do**
  ▷ *Compute the score for the Unkown Unary Potentials:*
  forall the $\delta^i_t = 0$ **do**
    $S^i_{\boldsymbol{\delta}_t} = \mathbb{E}_\theta \left[R \left(f_{PM} \left(\mathbf{x} | \boldsymbol{\delta}_t : \delta^i = 1, \boldsymbol{\pi}_t : \mu^i = \theta \right) \right) \right]$
  end
  ▷ *Instantiate the Unary Potential with higher* $S^i_{\boldsymbol{\delta}_t}$*:*
  $i^\star = \arg\max_i S^i_{\boldsymbol{\delta}_t}$
  $\delta^{i^\star} = 1$, Compute $\theta^{i^\star}$
end
$\mathbf{x}^\star = \arg\max_{\mathbf{x}} f_{PM}(\mathbf{x} | \boldsymbol{\delta}, \boldsymbol{\mu})$

Methods

■ Expected Reward

Expected Residual Entropy (ERE)

$$-H\left(f_{PM}\left(\mathbf{x}|\boldsymbol{\delta}_t : \delta^i = 1, \boldsymbol{\pi}_t : \mu^i = \theta\right)\right)$$

can be computed by drawing samples from the PM

Expected Labeling Change (ELC)

$$\Delta\left(\mathbf{x}_t^\star, \arg\max_{\mathbf{x}} f_{PM}(\mathbf{x}|\boldsymbol{\delta}_t : \delta^i = 1, \boldsymbol{\pi}_t : \mu^i = \theta)\right)$$

where x_t^* is the MAP labeling at iteration t, and $\Delta(\cdot,\cdot)$ is a function that counts how many labels of x_t^* differ from the labeling that we obtain with the PM when instantiating $\theta_i = \theta$.

Methods

■ Efficient Computation of the Reward

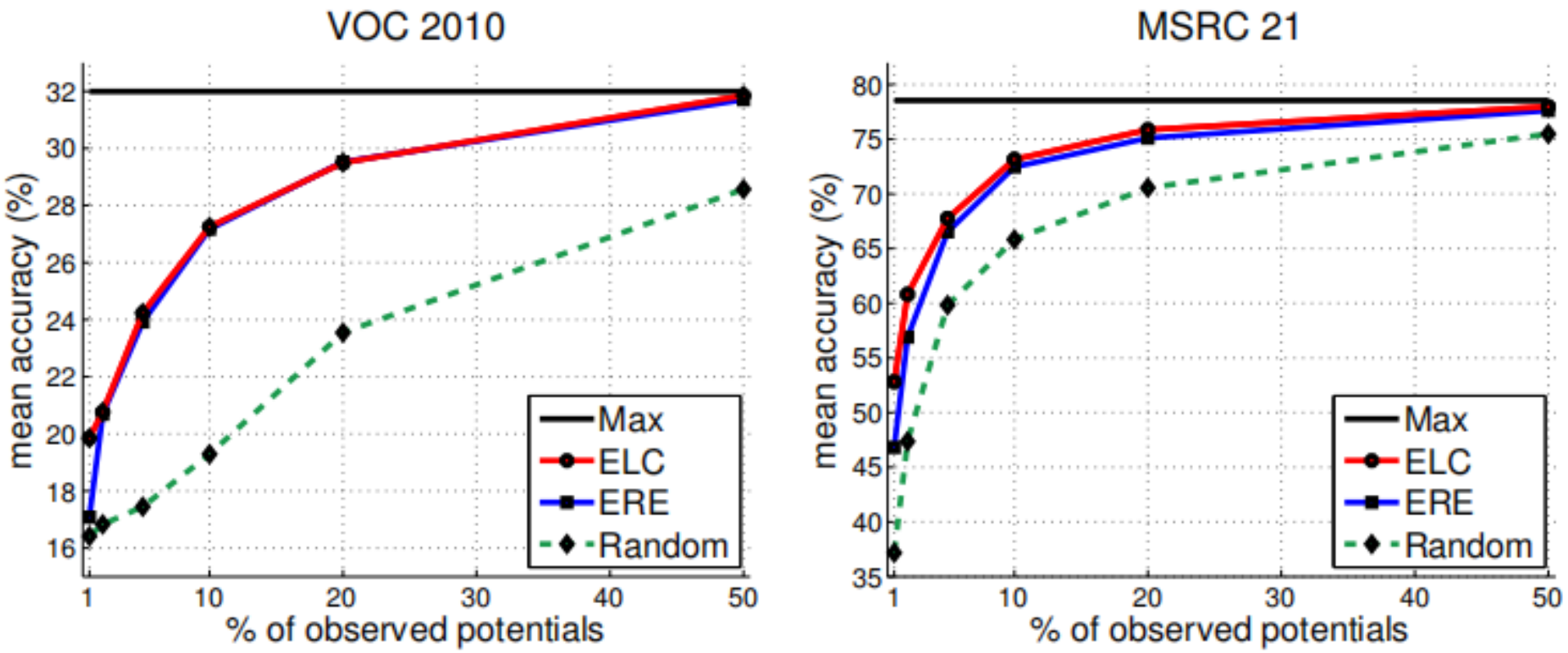
Efficient computation of the expected reward

the PM draws the unknown parameters of the energy from $f_{\boldsymbol{\theta}}\left(\theta^i|\delta^i=0,\pi^i\right)$, which is the same probability distribution that we use to generate the θs for the expected value.

Area of Influence

It is assumed that instantiating a potential reduces the score of the potentials that are in its "area of influence", while the rest remain unchanged.

Experiment



(a) (b)

Figure 7. *Results on (a) VOC10 and (b) MSRC-21.* Accuracy when varying the percentage of instantiated potentials.

Experiment

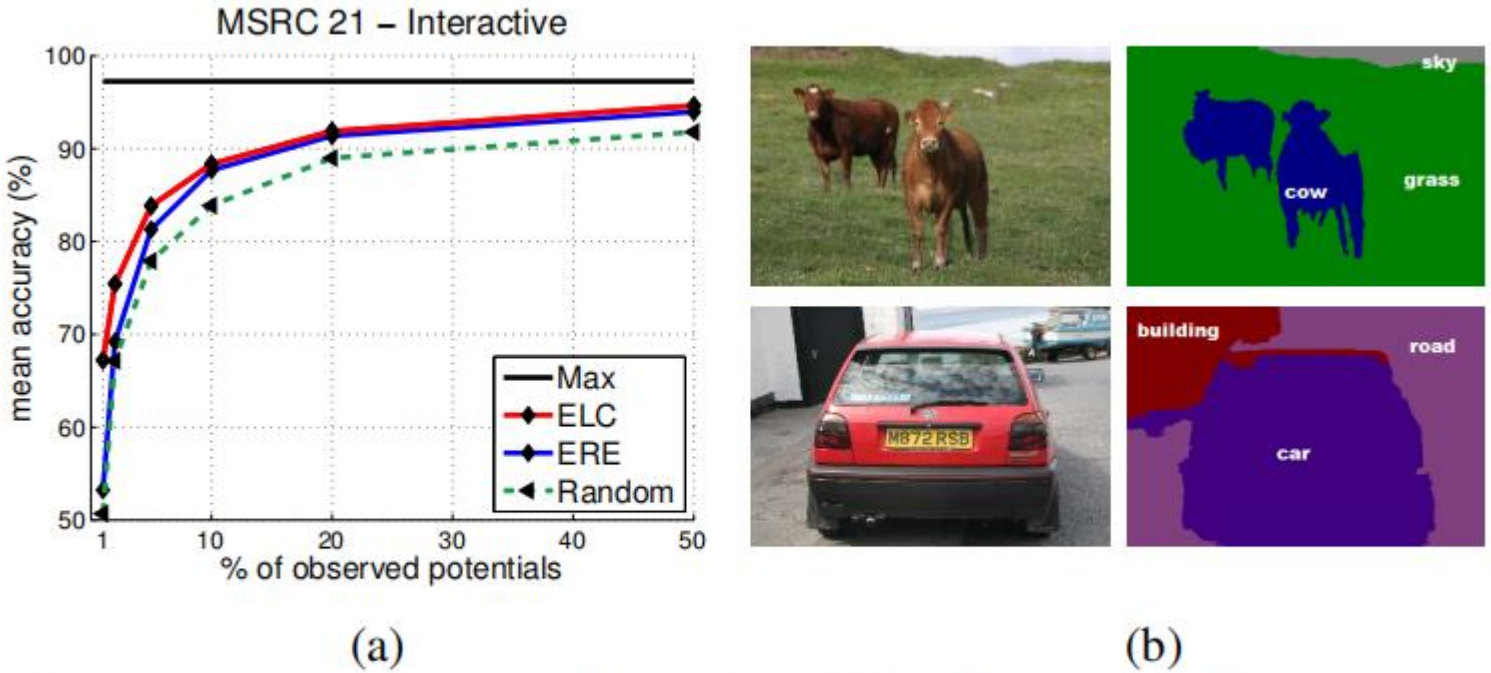
Method	Global	Average	Features	Inference	Total	Average	Features	Inference	Total
	MSRC-21: Test Set					VOC10: Validation Set			
All CRF	78	**78**	3s	0.02s	3.02s	**32.9**	16s	0.03	16.03s
All max	78	**78**	3s	–	3s	32.0	16s	–	16s
ELC 20%	76	**76**	0.6s	0.34s	0.94s	**29.5**	3.2s	0.37s	3.57s
ERE 20%	76	75	0.6s	0.34s	0.94s	29.5	3.2s	0.37s	3.57s
Random 20%	72	70	0.6s	0.1s	0.7s	23.5	3.2s	0.12s	3.32s
ELC 5%	70	**68**	0.15s	0.34s	0.49s	**24.2**	0.8s	0.12s	0.92s
ERE 5%	69	67	0.15s	0.34s	0.49s	23.9	0.8s	0.12s	0.92s
Random 5%	65	60	0.15s	0.1s	0.25s	17.4	0.8s	0.03s	0.83s
	MSRC-21: Human-in-the-loop					VOC10: Test Set			
All	98	**97**	–	–	300 clicks	33.5	16s	–	16s
ELC 20%	94	**92**	–	0.34s	60 clicks	30.4	3.2s	0.37s	3.57s
ELC 5%	86	84	–	0.34s	15 clicks	24.8	0.8s	0.12s	1.17s
ELC 1%	67	67	–	0.34s	3 clicks	–	–	–	–

Table 1. *Summary of all the results in MSCR-21 and VOC10.* The average score provides the per-class average. The time measurements are for one image.

Experiment

Active MAP for human-in-the-loop segmentation

Set the instantiated unary potentials to add high penalties for the labels different from the ground-truth, or 0



(a) (b)

Figure 8. *Results on MSRC-21 with a human in the loop.* (a) Average accuracy, and (b) example of resulting images.

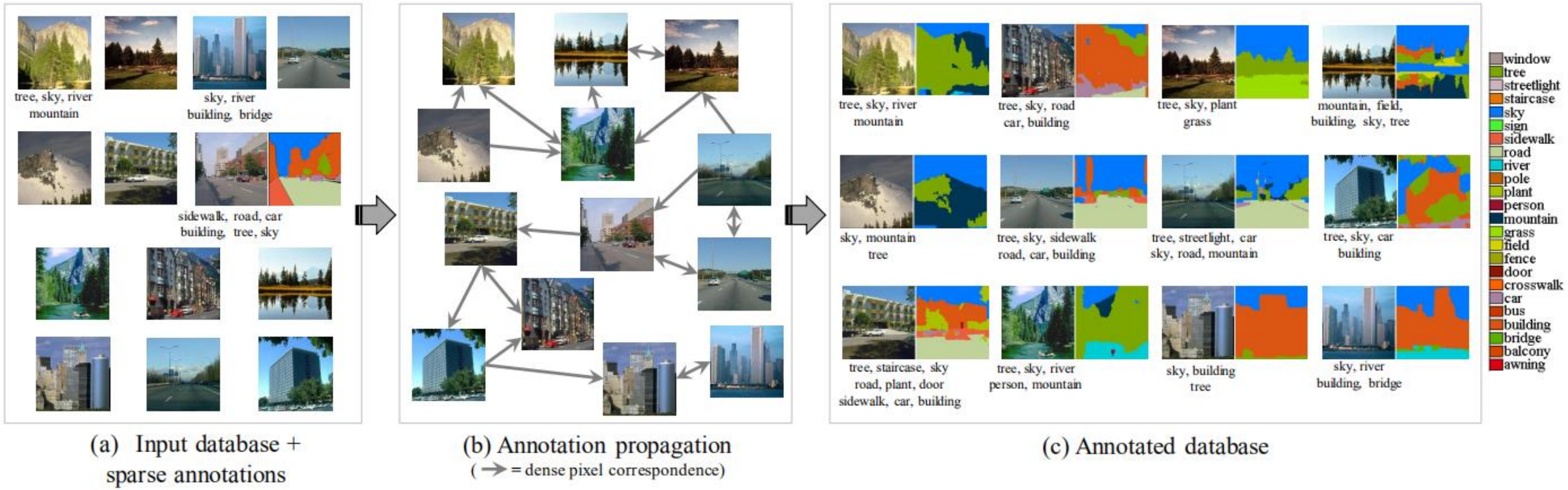
Annotation Propagation in Large Image Databases via Dense Image Correspondence

Michael Rubinstein[1,2], Ce Liu[2], and William T. Freeman[1]

[1]MIT CSAIL [2]Microsoft Research New England

ECCV-2012

System overview



(a) Input database + sparse annotations

(b) Annotation propagation
(→ = dense pixel correspondence)

(c) Annotated database

Fig. 1. Input and output of our system. (a) The input image database, with scarce image tags and scarcer pixel labels. (b) Annotation propagation over the image graph, connecting similar images and corresponding pixels. (c) The fully annotated database by our system (more results can be found in Fig. 5 and the supplemental material).

Intuition

■ **Four Assumptions**

(a) regions corresponding to each tag have distinct visual appearances,

(b) neighboring pixels in one image tend to have the same annotation,

(c) similar patterns across different images should have similar annotation,

(d) tags and tag co-occurrences which are more frequent in the database are also more probable.

Terminology and Formulation

■ Terminology

pixel-level labeling (or semantic segmentation) as "**labeling**"
textual image-level annotation as "**tags**"

input database $\Omega = (\boldsymbol{I}, \boldsymbol{V})$ is comprised of RGB images $\boldsymbol{I} = \{I_1, \ldots, I_N\}$ and

vocabulary $\boldsymbol{V} = \{l_1, \ldots, l_L\}$

for images label $\boldsymbol{C} = \{c_1, \ldots, c_N\}$

for pixel $\boldsymbol{p} = (x, y)$, $ci(\boldsymbol{p}) \in \{1, \ldots, L, \emptyset\}$ ($\emptyset$ denoting unlabeled)

The tags associated with the images, $\boldsymbol{T} = \{\boldsymbol{t}_1, \ldots, \boldsymbol{t}_N : \boldsymbol{t}_i \subseteq \{1, \ldots, L\}\}$

Define directly as the set union of the pixel labels $\boldsymbol{t}_i = \cup_{\boldsymbol{p} \in \Lambda_i} c_i(\boldsymbol{p})$, where

Λ_i is image and I_i is lattice

Terminology and Formulation

■ Formulation

Formulate this discrete labeling problem in a probabilistic framework, where our goal is to construct the joint posterior distribution of pixel labels given the input database and available annotations, $P(\boldsymbol{C}|\Omega, \boldsymbol{T}_t, \boldsymbol{C}_l)$.

$\boldsymbol{I}_t$ and $\boldsymbol{T}_t$, tagged images with their tags

$\boldsymbol{I}_l$ and $\boldsymbol{C}_l$, labeled images with their labels

Terminology and Formulation

■ Formulation

Written in energy form $(-\log P(C|\Omega, T_t, C_l))$, the cost of global assignment C of labels to pixels is given by

$$E(\mathbf{C}) = \sum_{i=1}^{N} \sum_{\mathbf{p} \in \Lambda_i} \left[\Phi_p(c_i(\mathbf{p})) + \sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} \Psi_{int}\left(c_i(\mathbf{p}), c_i(\mathbf{q})\right) + \sum_{j \in \mathcal{N}_i} \Psi_{ext}\left(c_i(\mathbf{p}), c_j(\mathbf{p} + \mathbf{w}_{ij}(\mathbf{p}))\right) \right]$$

where N_p are the spatial neighbors of pixel p within its image,
and N_i are other images similar to image I_i.
For each such image $j \in N_i$, w_{ij} defines the (dense) correspondence between pixels in I_i and pixels in I_j

Terminology and Formulation

■ Formulation

Written in energy form $(-\log P(C|\Omega, T_t, C_l))$, the cost of global assignment C of labels to pixels is given by

$$E(\mathbf{C}) = \sum_{i=1}^{N} \sum_{\mathbf{p} \in \Lambda_i} \left[\Phi_p(c_i(\mathbf{p})) + \sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} \Psi_{int}\left(c_i(\mathbf{p}), c_i(\mathbf{q})\right) + \sum_{j \in \mathcal{N}_i} \Psi_{ext}\left(c_i(\mathbf{p}), c_j(\mathbf{p} + \mathbf{w}_{ij}(\mathbf{p}))\right) \right]$$

- The **data term**, Φ_p, by means of local visual properties of image I_i at pixel p
- The **intra-image** smoothness term, ψ_{int}, is used to regularize the labeling with respect to the image structures
- The **inter-image** smoothness term, ψ_{ext}, is used for regularization across corresponding pixels in similar images.

Text-to-Image Correspondence

■ Local Image Descriptors

Select features used prevalently in object and scene recognition to characterize local image structures and color features

Using both **SIFT** and **HOG** features

yield a 527-dimensional descriptor $D_i(\boldsymbol{p})$ for every pixel $\boldsymbol{p}$ in image I_i

Using PCA to reduce the descriptor to $d = 50$ dimensions

Text-to-Image Correspondence

■ Learning Appearance Models

Use a generative model based on **Gaussian mixtures** to represent the distribution of the above continuous features

$$P(D_i(\mathbf{p}); \mathbf{\Theta}) = \sum_{l=1}^{L} \left(\rho_l \sum_{k=1}^{M} \pi_{l,k} \mathcal{N}\left(D_i(\mathbf{p}); \boldsymbol{\mu}_{l,k}, \mathbf{\Sigma}_{l,k}\right) \right) + \rho_\epsilon \mathcal{N}\left(D_i(\mathbf{p}); \boldsymbol{\mu}_\epsilon, \mathbf{\Sigma}_\epsilon\right)$$

where ρ_l is the weight of model (word) l in generating the feature $D_i(p)$,
M is the number of components in each model (M = 5),
$\theta_l = (\pi_{l,k}, \mu_{l,k}, \Sigma_{l,k})$is the mixture weight, mean and covariance of component k in model l
use a Gaussian outlier model with parameters $\theta_\epsilon = (\mu_\epsilon, \Sigma_\epsilon)$ and weight ρ_ϵ

using a standard EM algorithm to optimize

Text-to-Image Correspondence

■ Learning Appearance Models

Use a generative model based on Gaussian mixtures to represent the distribution of the above continuous features

$$P(D_i(\mathbf{p});\boldsymbol{\Theta}) = \sum_{l=1}^{L}\left(\rho_l \sum_{k=1}^{M} \pi_{l,k}\mathcal{N}\left(D_i(\mathbf{p});\boldsymbol{\mu}_{l,k},\boldsymbol{\Sigma}_{l,k}\right)\right) + \rho_\epsilon \mathcal{N}\left(D_i(\mathbf{p});\boldsymbol{\mu}_\epsilon,\boldsymbol{\Sigma}_\epsilon\right)$$

Given the learned model parameters, Θ, and an observed descriptor, $D_i(p)$, the probability of the pixel belonging to word l is computed by

$$P_a(c_i(\mathbf{p}) = l; D_i(\mathbf{p}),\boldsymbol{\Theta}) = \frac{\rho_l \sum_{k=1}^{M} \pi_{l,k}\mathcal{N}\left(D_i(\mathbf{p});\boldsymbol{\mu}_{l,k},\boldsymbol{\Sigma}_{l,k}\right)}{\sum_{j=1}^{L}\left(\rho_j \sum_{k=1}^{M} \pi_{j,k}\mathcal{N}\left(D_i(\mathbf{p});\boldsymbol{\mu}_{j,k},\boldsymbol{\Sigma}_{j,k}\right)\right)+\rho_\epsilon \mathcal{N}\left(D_i(\mathbf{p});\boldsymbol{\mu}_\epsilon,\boldsymbol{\Sigma}_\epsilon\right)}$$

Annotation Propagation

■ Dense Image Correspondence

First fetch the top (K, ϵ) nearest neighbors, $\boldsymbol{N_i}$, for every image I_i using the GIST descriptor, where K is the maximum number of neighbors, and ϵ is a threshold on the distance between the images

Use SIFT-flow to align the image with each of its neighbors, which results in an integer warp field $\boldsymbol{w}_{ij}$ mapping each pixel in I_i to a pixel in I_j

Annotation Propagation

■ Large-scale Inference

$$E(\mathbf{C}) = \sum_{i=1}^{N} \sum_{\mathbf{p} \in \Lambda_i} \left[\Phi_p(c_i(\mathbf{p})) + \sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} \Psi_{int}\left(c_i(\mathbf{p}), c_i(\mathbf{q})\right) + \sum_{j \in \mathcal{N}_i} \Psi_{ext}\left(c_i(\mathbf{p}), c_j(\mathbf{p} + \mathbf{w}_{ij}(\mathbf{p}))\right) \right]$$

Data term

$$\Phi_p\Big(c_i(\mathbf{p}) = l\Big) = -\log P_a(c_i(\mathbf{p})) - \log P_t^i(l) - \lambda_s \log P_s(c_i(\mathbf{p})) - \lambda_c \log P_c^i(c_i(\mathbf{p}))$$

where $P_a(c_i(p) = l; D_i(p), \Theta)$ was defined in the previous section,

$P_t^i(l)$ is the estimated probability of image I_i having the tag l,

$P_s\ (c_i(p))$ and $P_c^i(c_i(p))$ capture the probability of the pixel p having the word l based on its relative spatial position and color

Annotation Propagation

Data term

$$\Phi_p\Big(c_i(\mathbf{p}) = l\Big) = -\log P_a(c_i(\mathbf{p})) - \boxed{\log P_t^i(l)} - \lambda_s \log P_s(c_i(\mathbf{p})) - \lambda_c \log P_c^i(c_i(\mathbf{p}))$$

The term $P_t^i(l)$ is used to bias the tags of image I_i towards ones with higher frequency and co-occurrence among its neighbors.

$$P_t^i(l) = \frac{\beta}{|\mathcal{N}_i|} \sum_{j \in \mathcal{N}_i} \delta\left[l \in \mathbf{t}_j\right] + \frac{1-\beta}{Z} \sum_{j \in \mathcal{N}_i} \sum_{m \in \mathbf{t}_j} \mathbf{h}^o(l, m)$$

$\boldsymbol{h}^o$ is the $L \times L$ row-normalized tag co-occurrence matrix, computed from the current tag estimates and initialized from the known tags, $Z = \sum_{j \in N_i} |t_j|$

The first term measures the frequency of word l among image I_i's neighbors

The second term is the mean co-occurrence rate of word l within its neighbors' tags

Annotation Propagation

Data term

$$\Phi_p\Big(c_i(\mathbf{p}) = l\Big) = -\log P_a(c_i(\mathbf{p})) - \log P_t^i(l) - \lambda_s \log P_s(c_i(\mathbf{p})) - \lambda_c \log P_c^i(c_i(\mathbf{p}))$$

spatial location term $P_s\Big(c_i(\mathbf{p}) = l\Big) = \mathbf{h}_l^s(\mathbf{p})$

where $h_l^s(p)$ is the normalized spatial histogram of word l across all images in the database.

This term will assist in places where the appearance and pixel correspondence might not be as reliable

color term $P_c^i\Big(c_i(\mathbf{p}) = l\Big) = \mathbf{h}_c^{i,l}(I_i(\mathbf{p}))$

where $h_c^{i,l}$ is the color histogram of word l in image I_i

Annotation Propagation

■ Large-scale Inference

$$E(\mathbf{C}) = \sum_{i=1}^{N} \sum_{\mathbf{p} \in \Lambda_i} \left[\Phi_p(c_i(\mathbf{p})) + \sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} \Psi_{int}\left(c_i(\mathbf{p}), c_i(\mathbf{q})\right) + \sum_{j \in \mathcal{N}_i} \Psi_{ext}\left(c_i(\mathbf{p}), c_j(\mathbf{p} + \mathbf{w}_{ij}(\mathbf{p}))\right) \right]$$

Smoothness terms

inter-image compatibility between corresponding pixels p and $r = p + w_{ij}(p)$ in images I_i and I_j

$$\Psi_{ext}\Big(c_i(\mathbf{p}) = l_p, c_j(\mathbf{r}) = l_r\Big) = \delta\left[l_p \neq l_r\right] \frac{\alpha_j}{\alpha_i} \lambda_{ext} \exp\Big(- |S_i(\mathbf{p}) - S_j(\mathbf{r})| \Big)$$

where α_i, α_j are the image weights as defined in Section 3.2, and S_i is the SIFT descriptor for image I_i

Annotation Propagation

■ Large-scale Inference

$$E(\mathbf{C}) = \sum_{i=1}^{N} \sum_{\mathbf{p} \in \Lambda_i} \left[\Phi_p(c_i(\mathbf{p})) + \sum_{\mathbf{q} \in \mathcal{N}_\mathbf{p}} \Psi_{int}\left(c_i(\mathbf{p}), c_i(\mathbf{q})\right) + \sum_{j \in \mathcal{N}_i} \Psi_{ext}\left(c_i(\mathbf{p}), c_j(\mathbf{p} + \mathbf{w}_{ij}(\mathbf{p}))\right) \right]$$

Smoothness terms

intra-image compatibility

$$\Psi_{int}\left(c_i(\mathbf{p}) = l_p, c_i(\mathbf{q}) = l_q\right) = -\lambda_o \log \mathbf{h}^o\left(l_p, l_q\right) + \delta\left[l_p \neq l_q\right] \lambda_{int} \exp\left(-\|I_i(\mathbf{p}) - I_i(\mathbf{q})\|\right)$$

$\boldsymbol{h}^o$ is the $L \times L$ row-normalized tag co-occurrence matrix, computed from the current tag estimates and initialized from the known tags

Annotation Propagation

■ Choosing Images to Annotate

Want to strategically choose "**image hubs**" that have many similar images, as they will have many direct neighbors in the image graph to which they can propagate labels efficiently

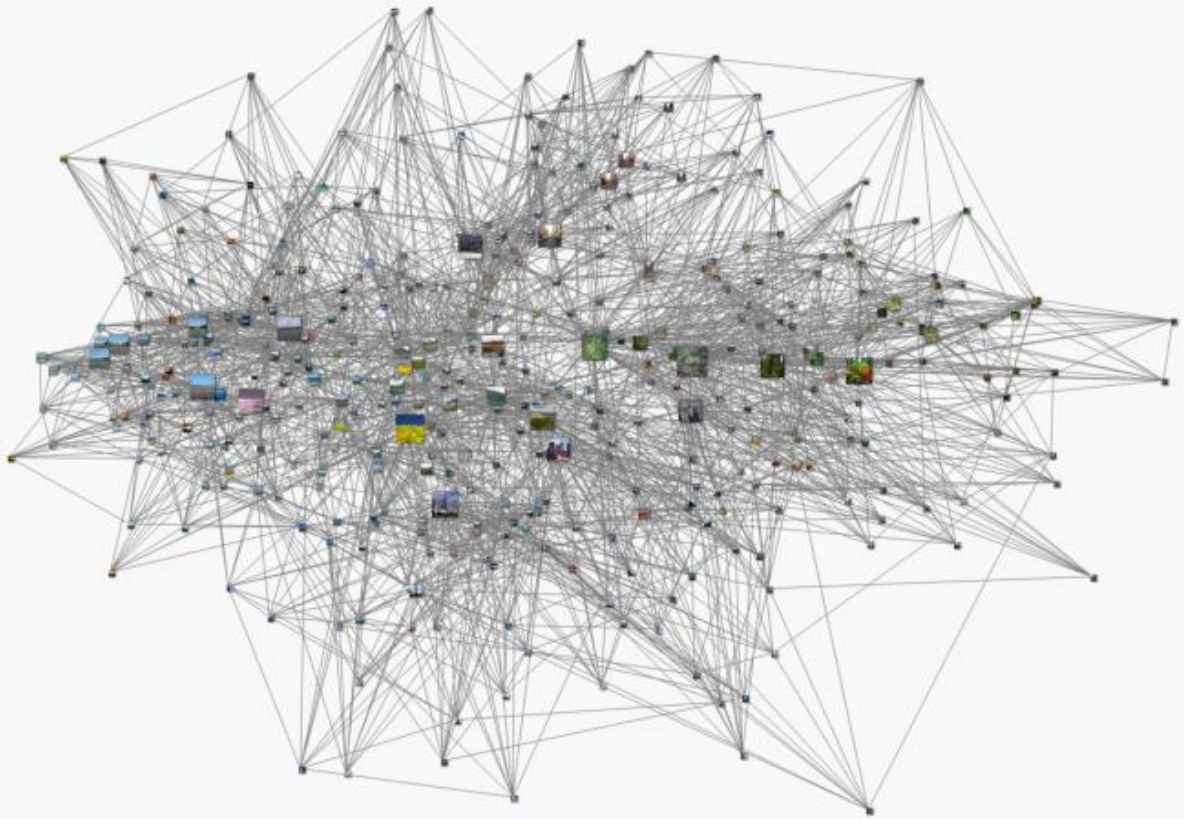
■ Method

Using the GIST descriptor as the image similarity measure.

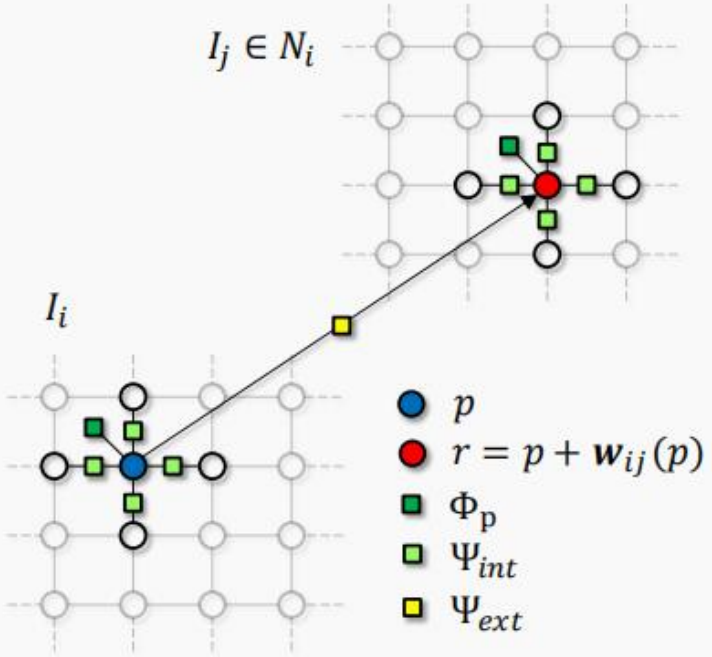
Use visual **pagerank** to find good images to label

VisualRank: Applying **PageRank** to Large-Scale Image Search TPAMI-2008

Annotation Propagation



(a) Image graph



(b) Graphical model

(c) Image hubs used for human labeling and tagging

System overview

The Architecture

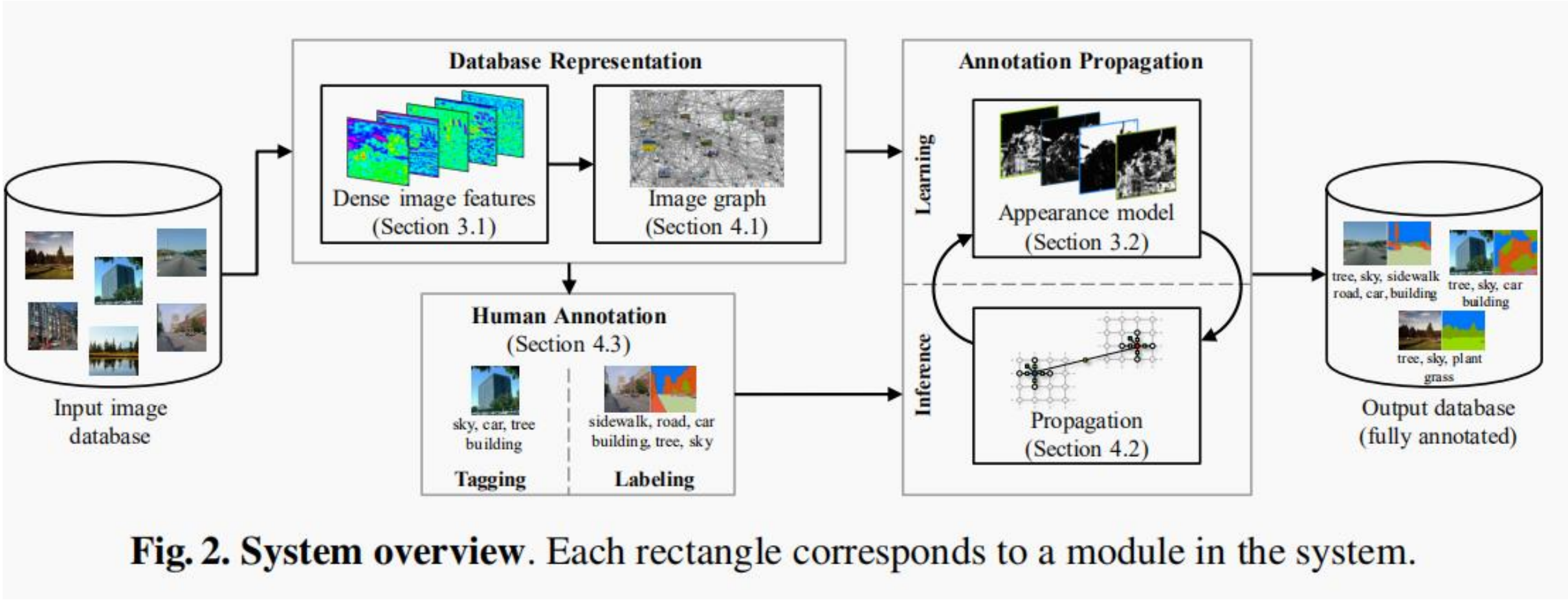


Fig. 2. System overview. Each rectangle corresponds to a module in the system.

Experiment

■ Active segmentation propagation

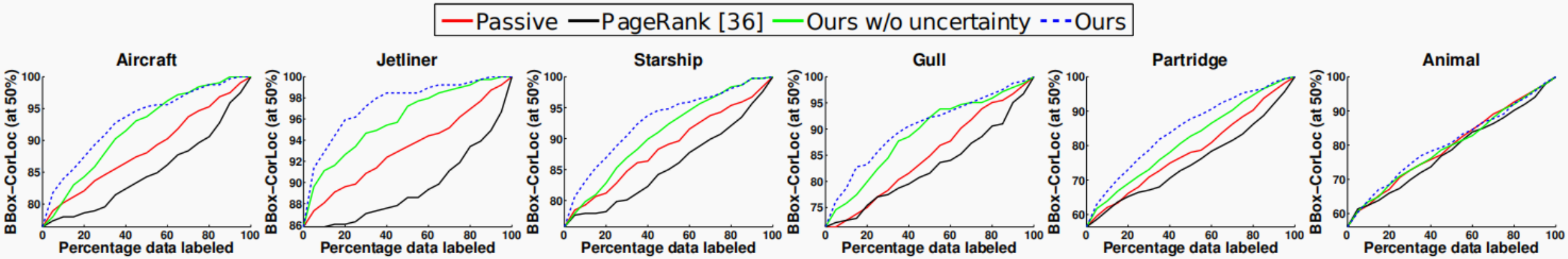


Figure 3: Active propagation for varying amounts of human annotation on a subset of the 3,624 ImageNet total synsets we tested (more in Supp.). Since only bounding box ground truth is available, we show bounding-box localization (BBox-CorLoc) accuracy (see Supp. for bounding-box Jaccard plots). Last plot (Animal) shows a failure case. Best viewed in color.

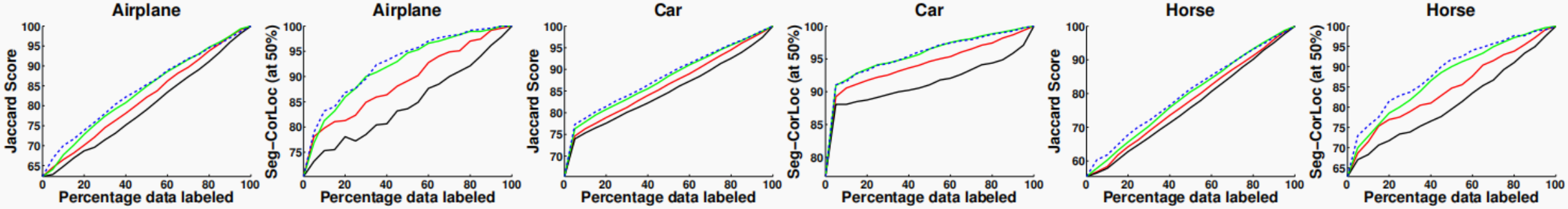


Figure 4: Active propagation results for varying amounts of human annotation for MIT Object Discovery dataset. We show both segmentation overlap (Jaccard) and segmentation localization (Seg-CorLoc) accuracy for each of the three classes. Best viewed in color.

Active Image Segmentation Propagation

Suyog Dutt Jain Kristen Grauman
University of Texas at Austin
suyog@cs.utexas.edu, grauman@cs.utexas.edu

CVPR-2016

Intuition

■ weakly supervised segmentation algorithms

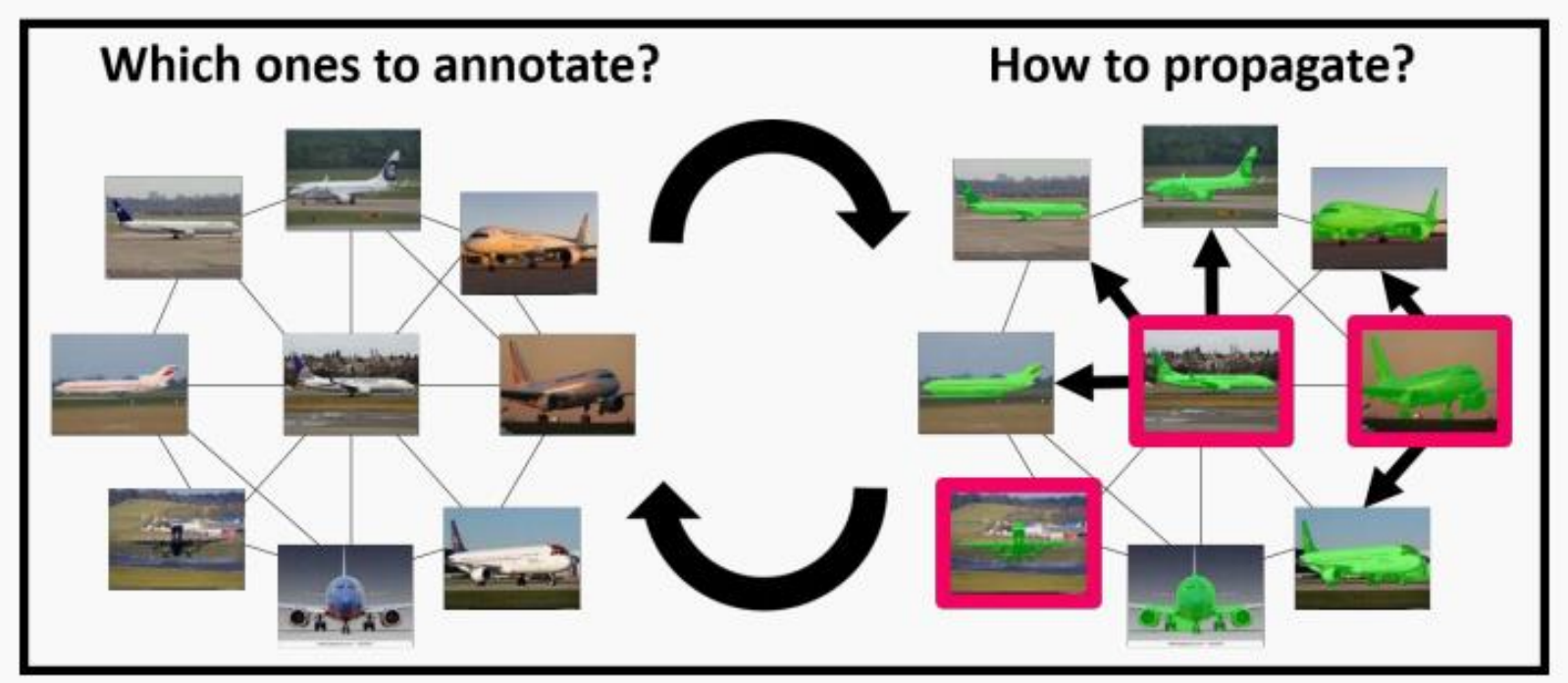
Take a pool of images known to contain the same object category, and exploit the repeated patterns to jointly segment out the foreground per image

■ semi-automatic segmentation propagation

The idea is to actively request human annotations for select images that, once labeled with their foreground, are most expected to help co-segment the remaining unlabeled images

Propagation engine proceeds

1. Actively choosing images which once annotated by humans will likely be most useful in propagating segmentations to other images
2. Given human annotations on actively chosen images (marked in pink), propagating them (dark arrows) to generate segmentations for other unlabeled images



Method

■ Problem Description

$I = \{I_1, I_2, \ldots, I_N\}$ be a collection of weakly supervised images, all of which contain instances of the **same object category**

Our goal is to jointly segment these images, yielding a foreground object mask $M = \{M_1, M_2, \ldots, M_N\}$ for each one

$R = \{R_{ij}\}$ denote the set of all region proposals in all N images, where R_{ij} denotes the j-th region for image I_i

Method

■ Region proposals and descriptors

1. Generate the generic object proposals and compute a saliency map using [Saliency Detection via Absorbing Markov Chain ICCV-2013]
2. Obtain two ranked lists of these proposals using **saliency** and **objectiveness** scores using[CPMC: Automatic Object Segmentation Using Constrained Parametric Min-Cuts]. retain the union of the top 30% from each list.
3. Cluster the reduced set into r clusters using the regions' HOG similarity and spatial overlap (IoU metric)

For each image I_i, extract a global appearance descriptor denoted I_i^c

For each region R_{ij} , extract two features: a saliency rating R_{ij}^s , and a region appearance descriptor R_{ij}^c

Method

■ Semi-automatic joint foreground segmentation

Define a Markov Random Field (MRF) joint segmentation graph $G = (R, E)$ based on the filtered region proposals across all images in the collection

Keep a sparse set of edges E by only connecting regions whose similarity exceeds a threshold τ

$$Y_{ij} = \begin{cases} 1 & \text{if proposal } R_{ij} \text{ is a good segmentation for } I_i \\ 0 & \text{otherwise.} \end{cases}$$

$S \subseteq I$ denote the current subset of images labeled with foreground masks by human annotators

1. Replace all nodes R_j^s by the single mask region given by the human annotator, denoted R^s, and we clamp its label $Y_s = 1$.
2. Modify the edge set E appropriately, such that in image s, only the mask $\hat{R}_s$ has edges to similar regions in unlabeled images

Method

■ Semi-automatic joint foreground segmentation

The intuition is that a good region proposal (i.e., one close to the actual foreground object segment) will strongly match a human-labeled ground truth region

Define energy function $E(Y)$ for jointly segmenting the image collection I

$$E(\mathcal{Y}) = \sum_{R_{ij}} -\log \Phi(Y_{ij}) + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij}).$$

The unary term is defined as

$$\Phi(Y_{ij}) = \begin{cases} Y_{ij} & \text{if } i \in \mathcal{S} \\ \alpha^{\mathrm{s}} \Phi^{\mathrm{s}}(Y_{ij}) + \alpha^{\mathrm{m}} \Phi^{\mathrm{m}}(Y_{ij}) & \text{if } i \in \mathcal{I} \backslash \mathcal{S}. \end{cases}$$

$$\Phi^{\mathrm{s}}(Y_{ij}) = Y_{ij} R^{\mathrm{s}}_{ij} + (1 - Y_{ij})(1 - R^{\mathrm{s}}_{ij}), \qquad \Phi^{\mathrm{m}}(Y_{ij}) = Y_{ij} R^{\mathrm{m}}_{ij} + (1 - Y_{ij})(1 - R^{\mathrm{m}}_{ij}), \text{ where}$$

so that we favor assigning $Y_{ij} = 1$ if R_{ij} is very salient. $\qquad R^{\mathrm{m}}_{ij} = \max_{p \in \mathcal{N}(\mathcal{I}_i, \mathcal{S})} sim(R^{\mathrm{c}}_{ij}, \bar{R}^{\mathrm{c}}_{p})$

Method

■ Semi-automatic joint foreground segmentation

The intuition is that a good region proposal (i.e., one close to the actual foreground object segment) will strongly match a human-labeled ground truth region

Define energy function $E(Y)$ for jointly segmenting the image collection I

$$E(\mathcal{Y}) = \sum_{R_{ij}} -\log \Phi(Y_{ij}) + \sum_{R_{ij}, R'_{ij} \in \mathcal{E}} \Psi(Y_{ij}, Y'_{ij}).$$

$$\Psi(Y_{ij}, Y'_{ij}) = \delta(Y_{ij} \neq Y'_{ij})\ sim(R^{c}_{ij}, R'^{c}_{ij})$$

The pairwise term encourages similar-looking regions to take the same label

The minimum energy solution $Y^* = \arg\min_Y E(Y)$ yields a set of good region proposals for each image in the collection.

Method

■ Active selection for propagation

Active selection algorithm accounts for three criteria—**influence**, **diversity**, and **uncertainty**

$$\textsc{Influence}(\mathcal{S}_t) = \frac{1}{|\mathcal{S}'_t|} \sum_{I_i \in \mathcal{S}_t} \sum_{I_j \in \mathcal{S}'_t} sim(I_i^c, I_j^c)$$

$$\textsc{Diversity}(\mathcal{S}_t) = -\frac{1}{|\mathcal{S}_t|} \sum_{I_i \in \mathcal{S}_t} \sum_{I_j \in \mathcal{S}_t} sim(I_i^c, I_j^c)$$

where S'_t denotes all unlabeled images not in the candidate batch S_t and sim is the cosine similarity

$$\textsc{Uncertainty}(\mathcal{S}_t) = \frac{1}{|\mathcal{S}_t|} \sum_{I_i \in \mathcal{S}_t} D(M_i)$$

where $D(\cdot)$ is a learned predictor of image difficulty

Pseudocode

First, extract the K > k most uncertain unlabeled images

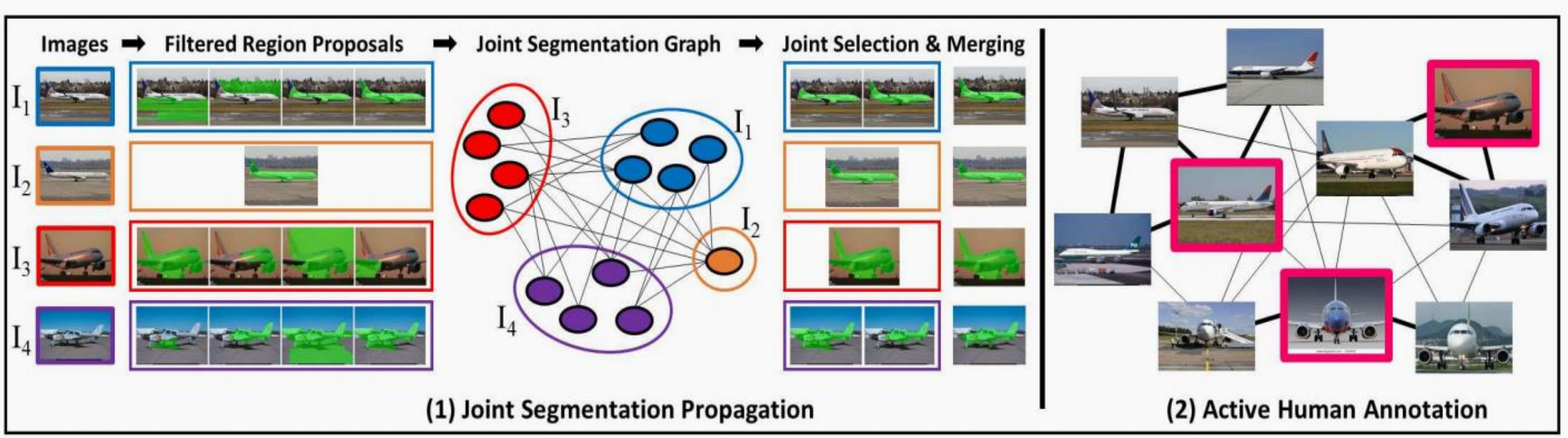
$$F(S_t) = INFLUENCE(S_t) + DIVERSITY(S_t)$$

Algorithm 1 Active Selection Algorithm

procedure ACTIVESELECTION
  Input: $\mathcal{I}, \mathcal{I}^u = \mathcal{I}, \mathcal{I}^l = \phi$;
  Define: $\mathcal{F}(\mathcal{S}) = \text{INFLUENCE}(\mathcal{S}) + \text{DIVERSITY}(\mathcal{S}), \mathcal{S} \subseteq \mathcal{I}$;
  for each stage t = 1,2, ..., T **do**
    Candidate set: $\mathcal{I}_t^u = \phi$;
    for i = 1,2, ..., K **do**
      $s^* = \arg\max_{s \in \mathcal{I}^u \setminus \mathcal{I}_t^u} D(M_s^{t-1})$; $\mathcal{I}_t^u = \mathcal{I}_t^u \cup s^*$;
    end for
    $\mathcal{S}_t = \phi, \mathcal{S}_t' = \mathcal{I}_t^u$;
    for i = 1,2, ..., k **do**
      $s^* = \arg\max_{s \in \mathcal{S}_t'} \mathcal{F}(\mathcal{S}_t \cup s) - \mathcal{F}(\mathcal{S}_t)$;
      $\mathcal{S}_t = \mathcal{S}_t \cup s^*$; $\mathcal{S}_t' = \mathcal{S}_t' \setminus s^*$;
    end for
    $\mathcal{I}^l = \mathcal{I}^l \cup \mathcal{S}_t$; $\mathcal{I}^u = \mathcal{I}^u \setminus \mathcal{S}_t$;
  end for
end procedure

R-CNN

■ Architecture and Process



Experiment

■ Active segmentation propagation

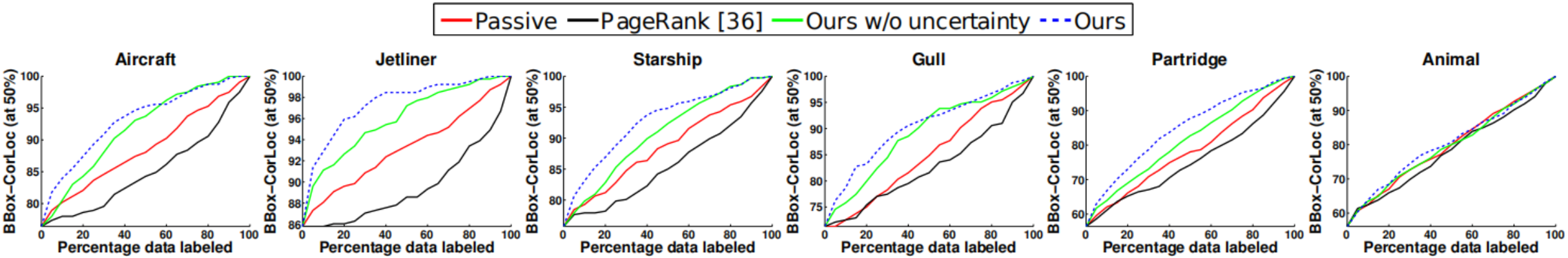


Figure 3: Active propagation for varying amounts of human annotation on a subset of the 3,624 ImageNet total synsets we tested (more in Supp.). Since only bounding box ground truth is available, we show bounding-box localization (BBox-CorLoc) accuracy (see Supp. for bounding-box Jaccard plots). Last plot (Animal) shows a failure case. Best viewed in color.

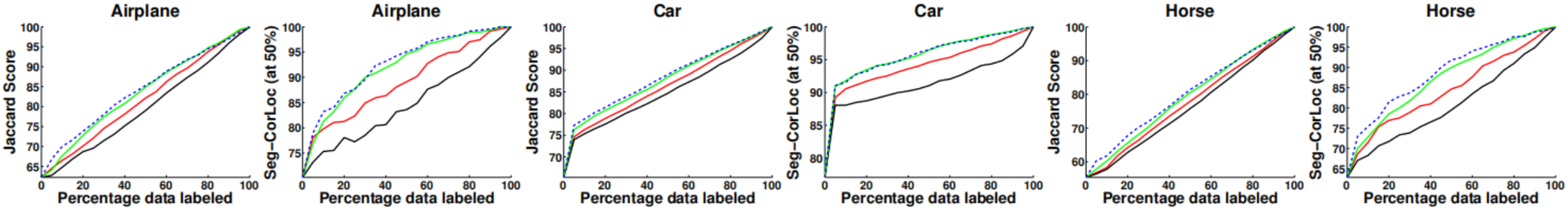


Figure 4: Active propagation results for varying amounts of human annotation for MIT Object Discovery dataset. We show both segmentation overlap (Jaccard) and segmentation localization (Seg-CorLoc) accuracy for each of the three classes. Best viewed in color.

Experiment

■ Weakly supervised foreground segmentation

Weak supervision (i.e., all images have an object from the same category) is the only information available. No additional human annotation is requested.

Methods	MIT dataset (subset)			MIT dataset (full)		
	Airplane	Car	Horse	Airplane	Car	Horse
# Images	82	89	93	470	1208	810
Joulin et al. [19]	15.36	37.15	30.16	n/a	n/a	n/a
Joulin et al. [20]	11.72	35.15	29.53	n/a	n/a	n/a
Kim et al. [21]	7.9	0.04	6.43	n/a	n/a	n/a
Rubinstein et al. [35]	55.81	64.42	51.65	55.62	63.35	53.88
Chen et al. [9]	54.62	**69.2**	44.46	60.87	62.74	**60.23**
Ours	**58.65**	66.47	**53.57**	**62.27**	**65.3**	55.41

Table 1: Comparison with state-of-the-art methods on MIT dataset for weakly supervised joint foreground segmentation (Metric: Jaccard score).

Experiment

■ Weakly supervised foreground segmentation

Weak supervision (i.e., all images have an object from the same category) is the only information available. No additional human annotation is requested.

ImageNet dataset	
# Classes	**# Images**
3,624	939,516

Methods	BBox-CorLoc
Top obj. box [3]	37.42
Tang et al. [42]	53.20
Ours	**57.64**

Table 2: Comparison with state-of-the-art methods on ImageNet for weakly supervised joint foreground segmentation (Metric: Avg. BBox-CorLoc).