High-order potentials, high-order losses

Anton Osokin
Pushmeet Kohli
Stefanie Jegelka

Interactive image segmentation

Input:
image,
user-defined “seeds”

Output:
segmentation

MRFs for Image Labelling

$$P(Y \mid X) \propto \prod_i \Psi(x_i \mid y_i) \prod_{ij} \Psi_{ij}(y_i, y_j)$$

Posterior Probability

Likelihood Term

Pairwise Consistency Prior

Negative log

$$E(Y) = \sum_i \varphi_i(y_i) + \sum_{ij} \varphi_{ij}(y_i, y_j)$$

Energy

Maximum a Posteriori (MAP) Inference

$$Y^* = \arg\max_Y P(Y \mid X)$$

Energy Minimization

$$Y^* = \arg\min_Y E(Y)$$

Image Segmentation

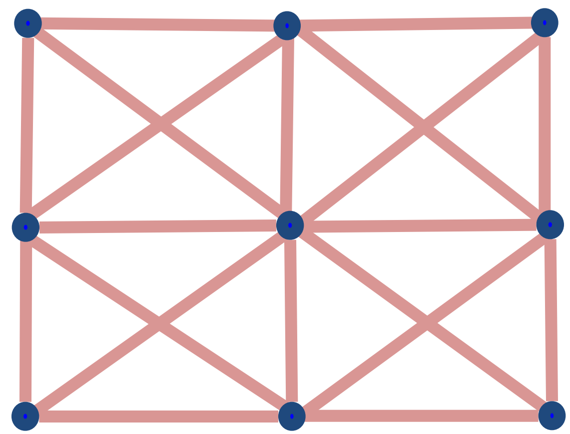
$$E(Y) = \sum_i \varphi_i(y_i) + \sum_{ij} \varphi_{ij}(y_i, y_j)$$

$$E : \{0,1\}^N \to \mathbb{R}$$

$$0 \to \text{bg}$$

$$1 \to \text{fg}$$

N = number of pixels



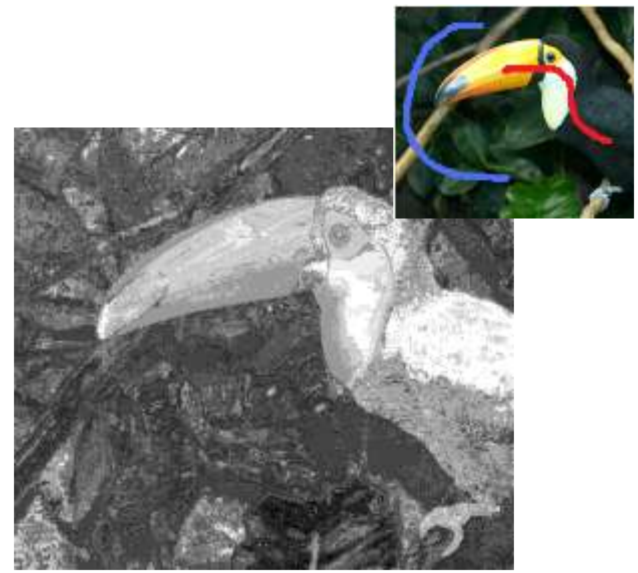
[Boykov and Jolly ' 01] [Blake et al. '04] [Rother, Kolmogorov and Blake `04]

Unary potentials

$$E(Y) = \sum_i \varphi_i(y_i) + \sum_{ij} \varphi_{ij}(y_i, y_j)$$

$$\sum_i c_i(1 - y_i)$$

Pixel color



Unary Cost c_i

Dark (negative) Bright (positive)

[Boykov and Jolly ' 01] [Blake et al. '04] [Rother, Kolmogorov and Blake `04]

Pairwise potentials

$$E(Y) = \sum_i \varphi_i(y_i) + \sum_{ij} \varphi_{ij}(y_i, y_j)$$

Discontinuity Cost d_{ij}

$$\sum_{ij} d_{ij}|y_i - y_j|$$

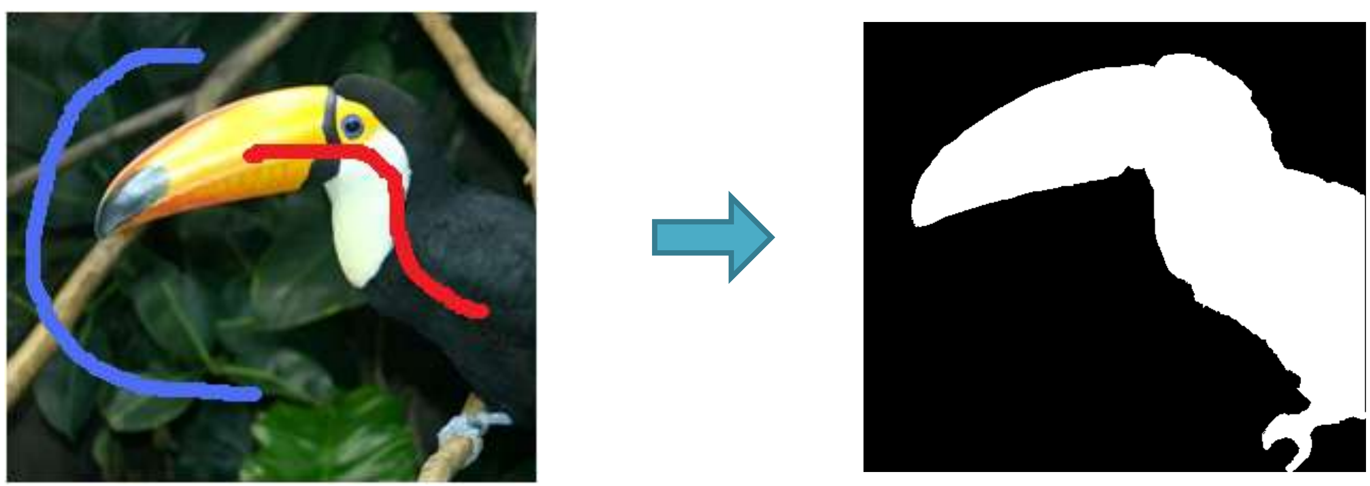
Smoothness Prior

[Boykov and Jolly ‘ 01] [Blake et al. ‘04] [Rother, Kolmogorov and Blake `04]

Energy minimization

$$E(Y) = \sum_i c_i(1 - y_i) + \sum_{ij} d_{ij}|y_i - y_j|$$

If all $d_{ij} \geq 0$ then the energy is submodular
and can be minimized with max-flow/min-cut algorithm.



[Boykov and Jolly ‘ 01] [Blake et al. ‘04] [Rother, Kolmogorov and Blake `04]

Problems

- Encourages Short Boundaries

[Jegelka & Bilmes, 11]

Image

Segmentation

- Does not enforce connectivity

[Vincente et al.] [Nowozin and Lampert] [Rhemann et al.]

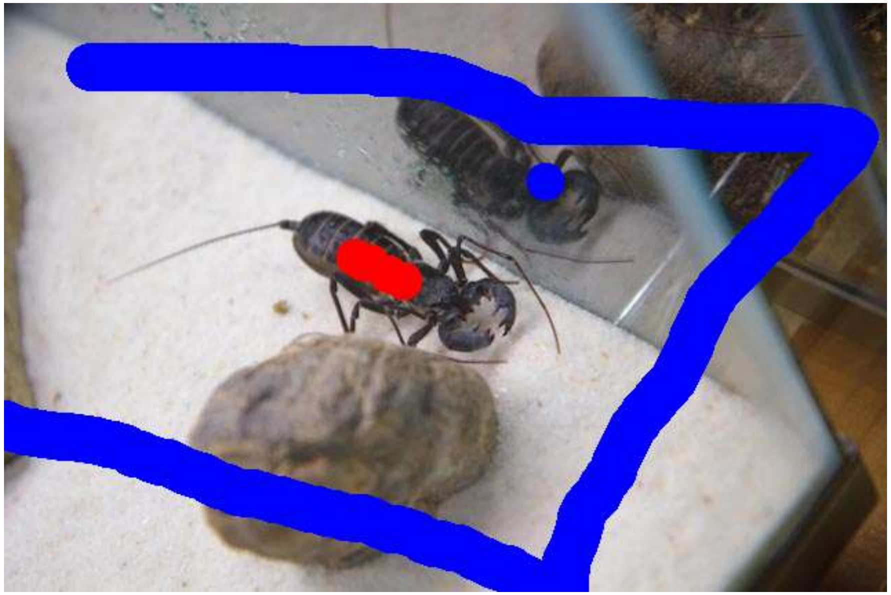
- Inconsistent labeling of similar pixels

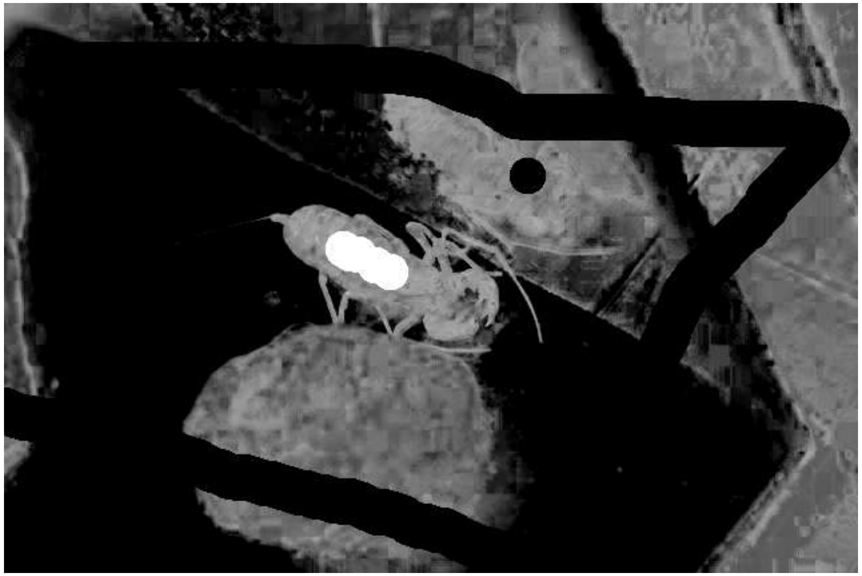
[Kohli et al. 07, 08]

- Consistency with Area/Boundary length

[Boykov et al, 07][Lim et al, 08]

Example





Example: GraphCut results

Ground truth:

Pairwise weight = 1.5

Pairwise weight = 1.0

Pairwise weight = 0.6

Pairwise weight = 0.5

Part 1: Cooperative cut model

Overcoming short-boundary bias

$$E(Y) = \sum \varphi_i(y_i) \quad + \boxed{\sum d_{ij}|y_i - y_j|}$$

Encourages short boundaries

Image

Segmentation

Penalize types of boundaries not the actual number of boundaries!

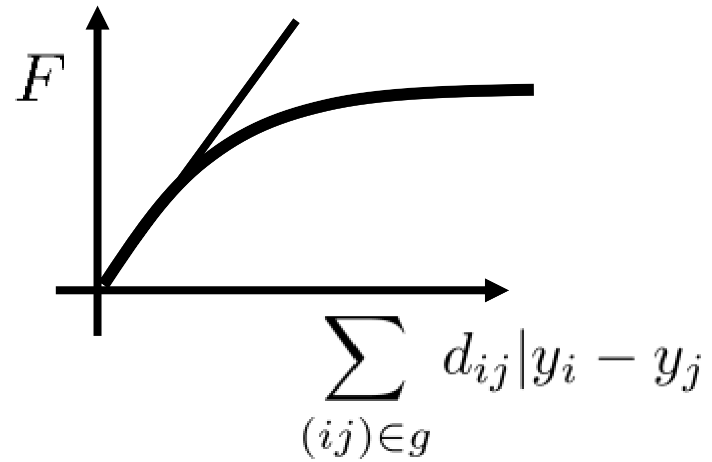
Cooperative cut model [Jegelka and Bilmes, 11]

Overcoming short-boundary bias

$$E(Y) = \sum \varphi_i(y_i) \; + \; \sum d_{ij}|y_i - y_j| + \sum h_g(Y)$$

$$h_g(Y) = F\left(\sum_{(ij)\in g} d_{ij}|y_i - y_j|\right)$$

- Divide edges into different types
- Incorporate a higher order consistency potential over the edges



[Jegelka and Bilmes, 11]

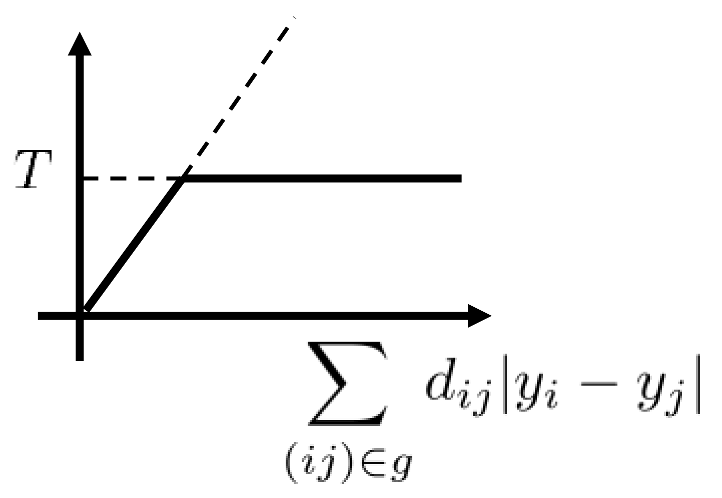
Constructing groups

For each edge construct a feature vector:
absolute difference between two nodes.

Exclude edges with no difference.

Cluster all remaining edges into 10 clusters using K-means.

For each cluster use a truncated linear
Function *F*



Example

Ground truth:

Pairwise weight = 1.0

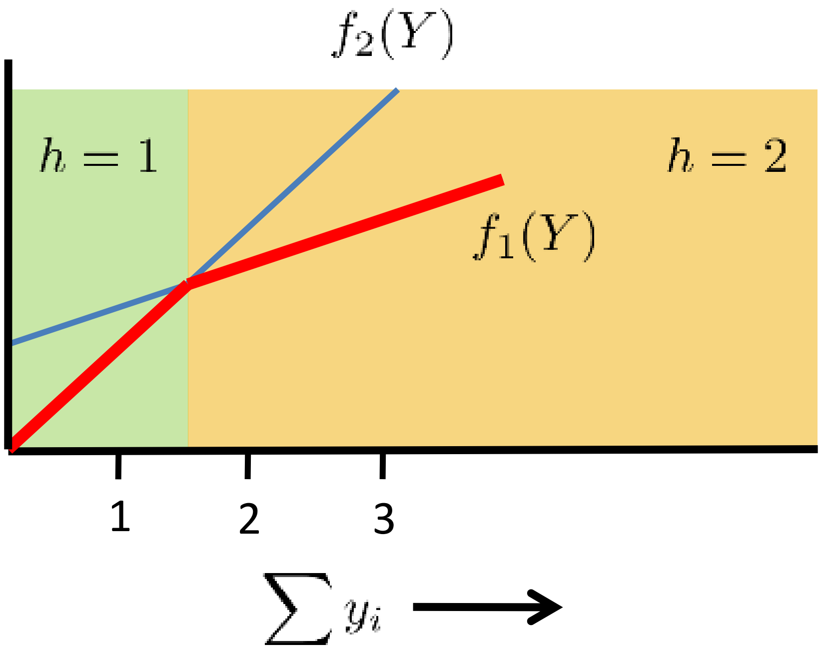
Cooperation

Example

Ground truth:

Cooperation

Transformation to pairwise energy



$$f(y) = \min(f_1(Y), f_2(Y))$$

Lower envelop of concave functions is concave

$$\min_{Y} f(Y) \quad = \quad \min_{Y, h \in \{0,1\}} h f_1(Y) + (1-h) f_2(Y)$$

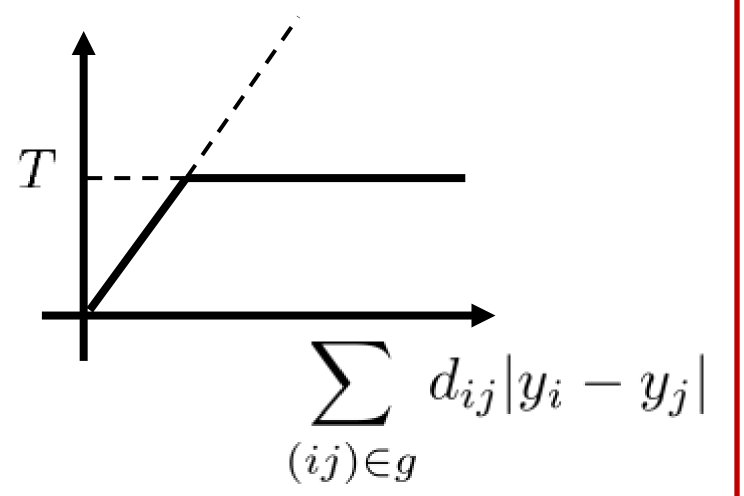
Higher Order Submodular Function

Quadratic Submodular Function

[Kohli et al. 08] [Kohli and Kumar, 10]

Our transformation

$$H_g(Y) = \min \left\{ \sum_{(ij)\in g} d_{ij}|y_i - y_j|, \quad T \right\}$$



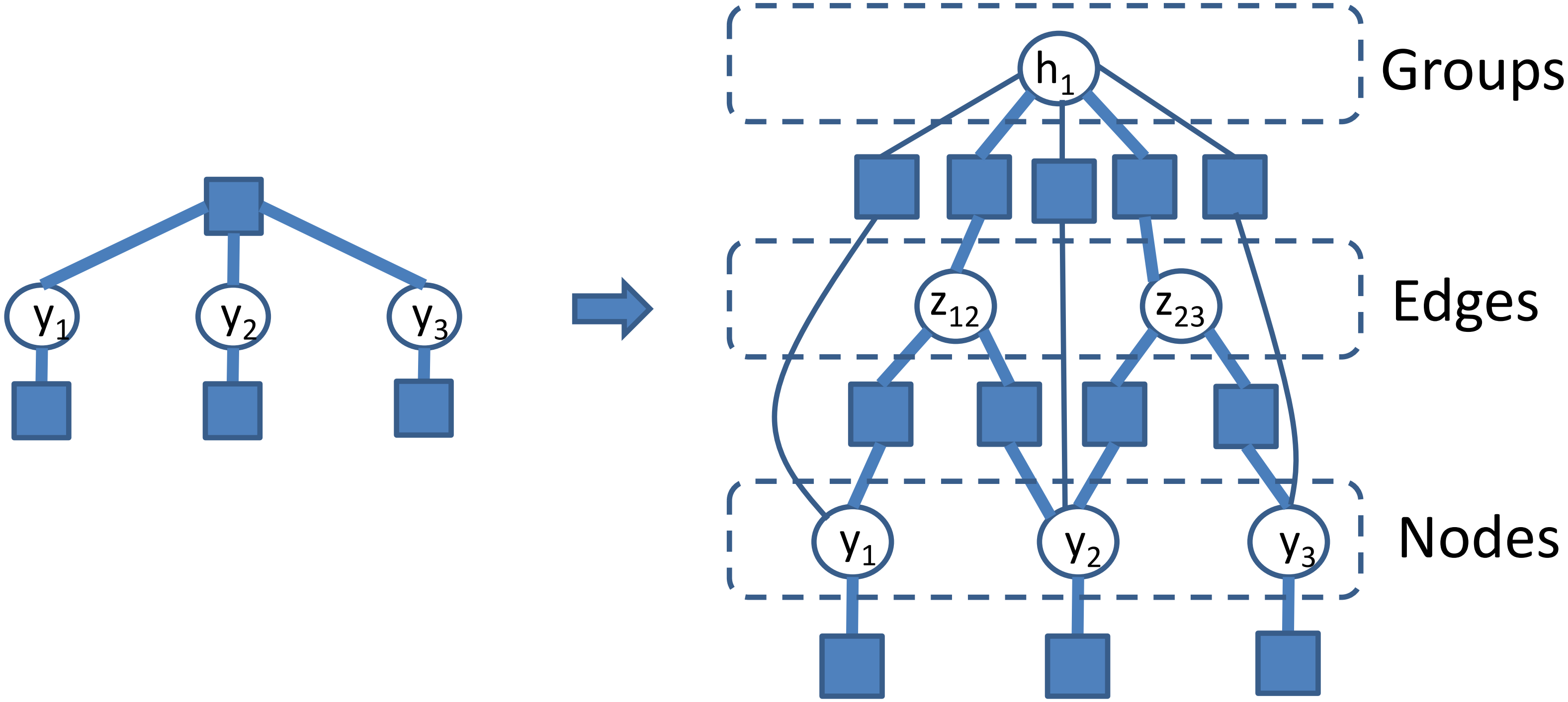
+ Switching variables h_g

$$H_g(Y) = \min_{h_g \in \{0,1\}} h_g \sum d_{ij}(y_i + y_j - 2y_i y_j) + T(1 - h_g)$$

+ Standard reduction for $-h_g y_i y_j$ using variable z_{ij}

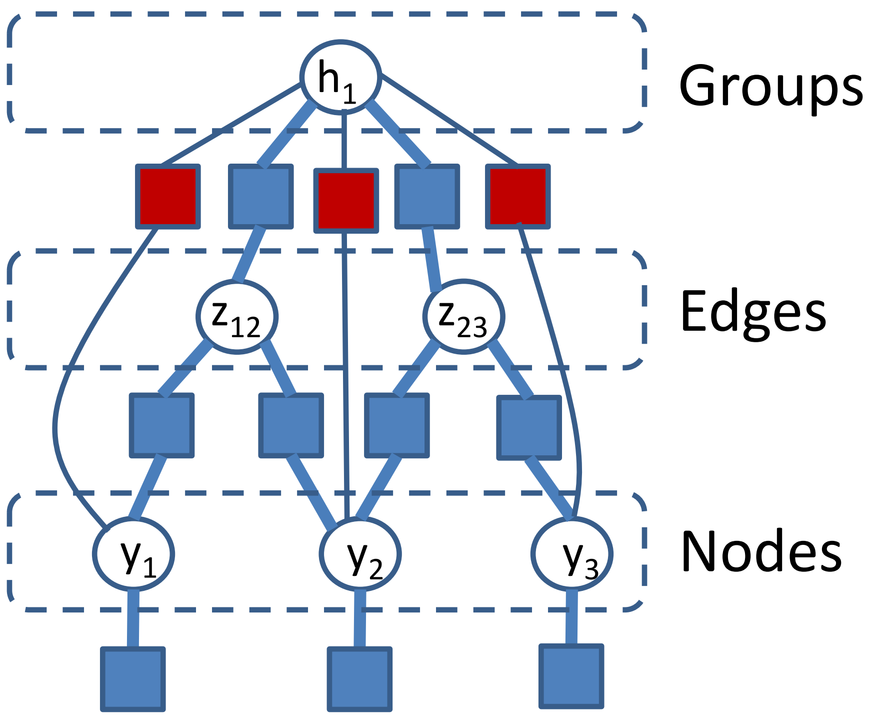
$$-h_g y_i y_j = \min_{z_{ij} \in \{0,1\}} (-z_{ij}(y_i + y_j + h_g - 2))$$

Transformation



Observations

1) non-submodular factors are concentrated around h_g
2) if we fix h_g the energy becomes submodular



Algorithms

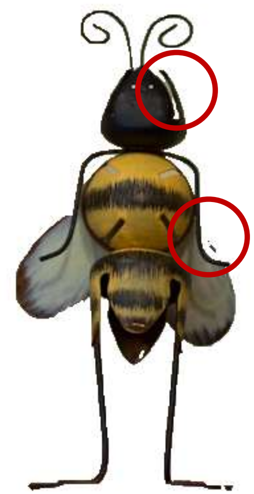
1) Exhaustive search over h_g
   a) dynamic graph cuts [Kohli and Torr, 05]
   b) special order of search

2) Different greedy strategies
   a) descent till convergence
   b) 1 pass over variables

3) Iterative bound minimization [Jegelka and Bilmes, 11]

Qualitative results

Graph cut



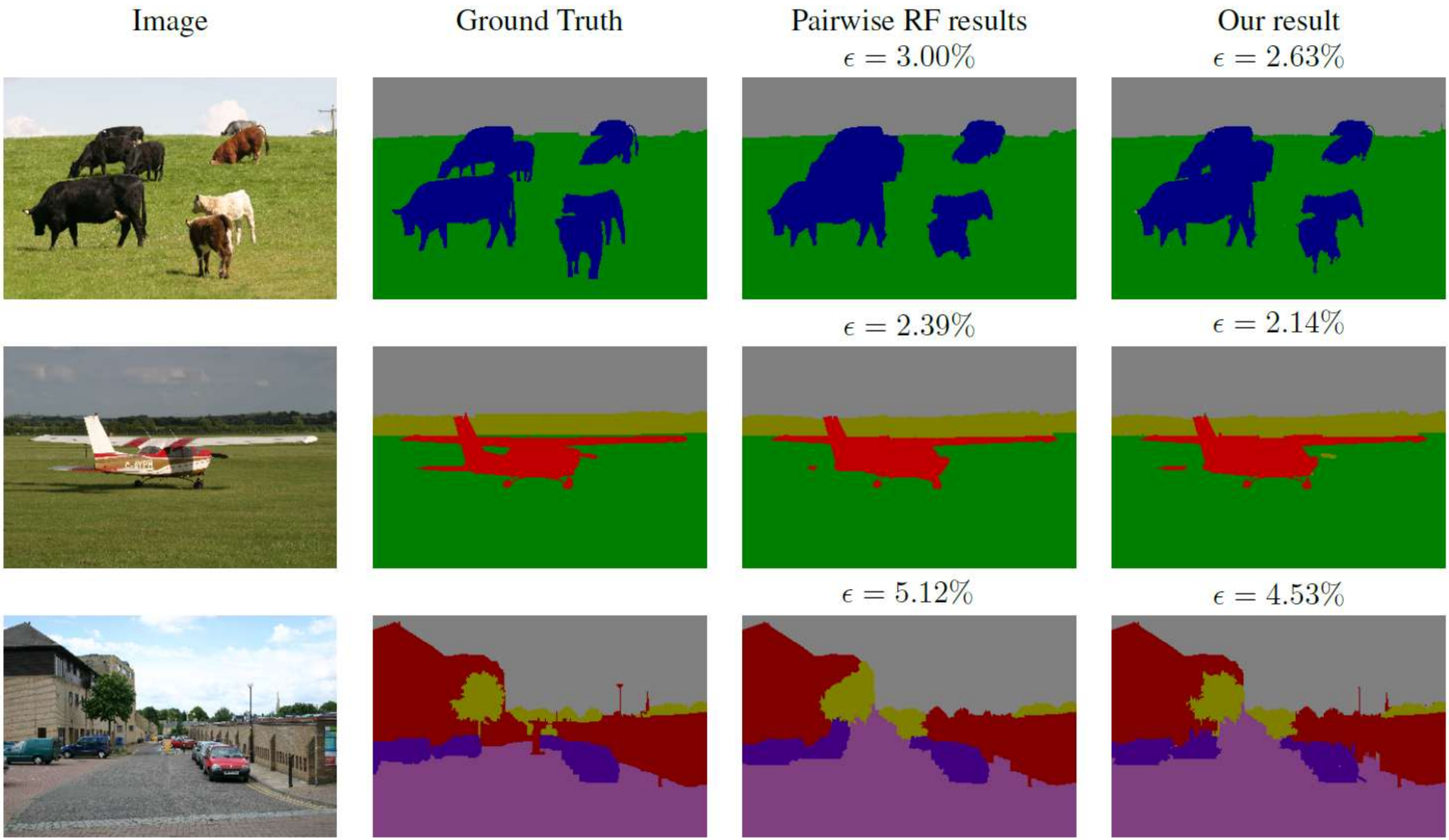
[Jegelka&Bilmes, 11]

Global minimum

Quantitative results

Method	**Energy**	**Time**	**Error**
GraphCut	1.0	0.19	1.61
It. bound min.	0.39	0.47	0.77
Global minimum	0.0	14.32	0.73
Greedy	0.0	2.37	0.73
1 pass	0.03	1.23	0.87

Multilabel results



Part 2: High-order losses

Training setup

Data: $\{X^i, Y^i\},\ i = 1, \ldots, N$

X

Y

Energy parameterization: $E(X, Y, w) = -w^\intercal \psi(X, Y)$

Goal: find parameters such that model produces "good" results

"Good" is defined by loss function $\Delta(Y, Y^i)$

Large-margin approach

$$\min_{w,\xi\geq 0} \quad \frac{1}{2}\|w\|^2 + \frac{C}{N}\sum_{i=1}^{N}\xi_i,$$

$$\text{s.t.} \quad w^\intercal\left(\psi(X^i,Y) - \psi(X^i,Y^i)\right) + \Delta(Y,Y^i) \leq \xi_i, \quad \forall Y$$

The problem can be solved with cutting-plane method.

Key step: finding the most violated constraint

$$\arg\min\left(-w^\intercal\psi(X^i,Y) - \Delta(Y,Y^i)\right)$$

[Tsochantaridis et al, 05] [Taskar et al 04,05] [Szummer et al 08]……

Loss-related problems

- Correspondence to human perception

- We are not optimizing the loss, but its hinge-bound

- Biased loss estimates
  - Certain losses have higher Generalization error

Simple (decomposable) losses

Hamming distance: $\Delta(Y, Y^k) = \sum_i [y_i \neq y_i^k]$

Weighted Hamming distance:

$$\Delta(Y, Y^k) = \sum_i c_i(Y_k)[y_i \neq y_i^k]$$

Hamming distance averaged over classes (HAC):

$$\Delta(Y, Y^k) = \frac{1}{2} \sum_{n \in \{0,1\}} \frac{\sum_i [y_i \neq n][y_i^k = n]}{\sum_i [y_i^k = n]}.$$

High-order losses

[Tarlow and Zemel, 2012]

PASCAL VOC loss (Jaccard distance):

$$\frac{\#(\text{true positives})}{\#(\text{true positives}) + \#(\text{false positives}) + \#(\text{false negatives})}$$

Loss augmented inference by message-passing

High-order losses

[Pletscher and Kohli, 2012]

Observation:

the loss enters the energy minimization with the negative sign, so supermodular losses are good

Example: "count" loss

$$\Delta(Y, Y^k) = \left| \sum_i y_i - \sum_i y_i^k \right|$$

More generally, any upper envelope of linear functions can be done

[Kohli and Kumar, 2010]

Family of losses

S – arbitrary sets of pixels

$$\Delta(Y, Y^k) = \sum_{S \in \mathcal{S}} c_S \left| \sum_{i \in S} y_i - \sum_{i \in S} y_i^k \right|$$

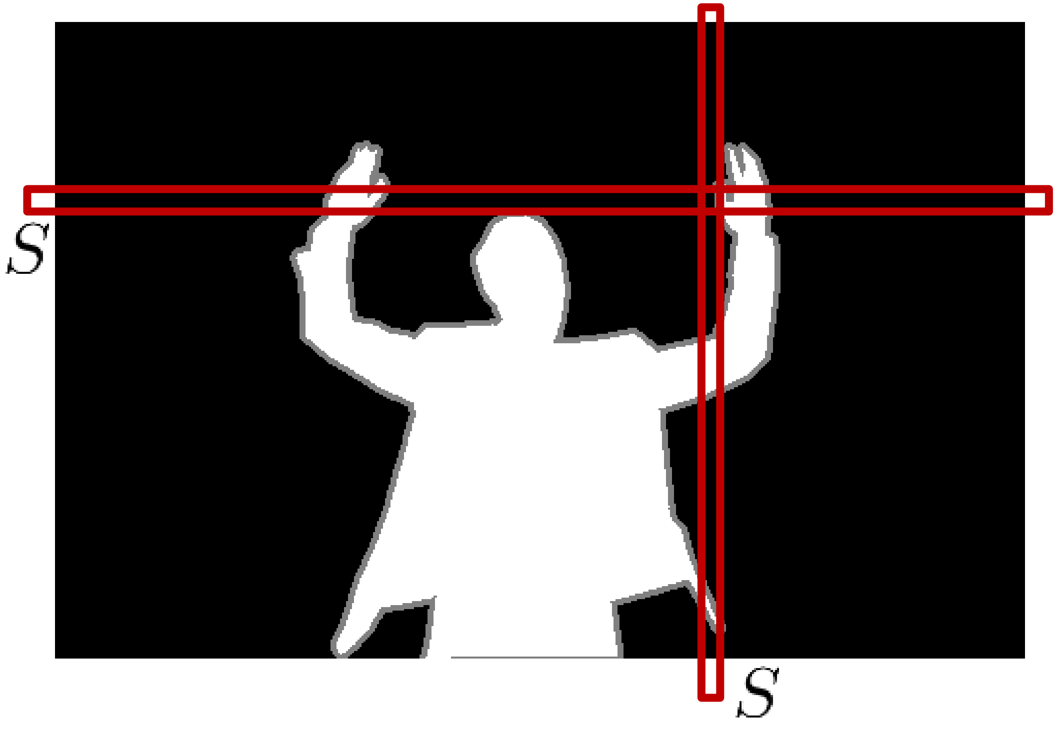
Special cases:

- Hamming (weighted, averaged) loss
- “Count” loss

Silhouette loss

S – rows and columns

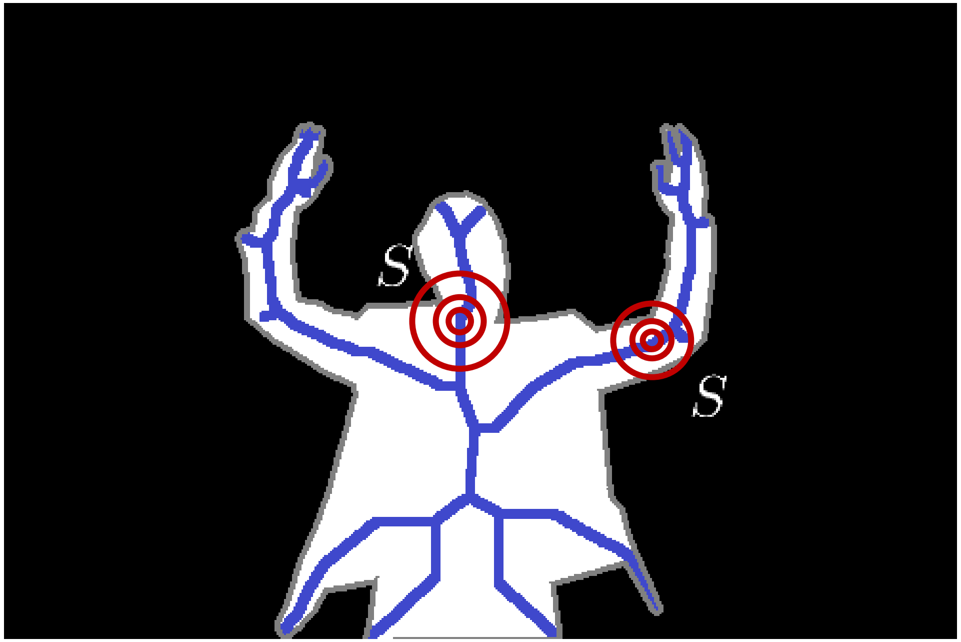
$$\Delta(Y, Y^k) = \sum_{S \in \mathcal{S}} c_S \left| \sum_{i \in S} y_i - \sum_{i \in S} y_i^k \right|$$



Skeleton based loss

S can depend on the groundtruth

$$\Delta(Y, Y^k) = \sum_{S \in \mathcal{S}} c_S \left| \sum_{i \in S} y_i - \sum_{i \in S} y_i^k \right|$$



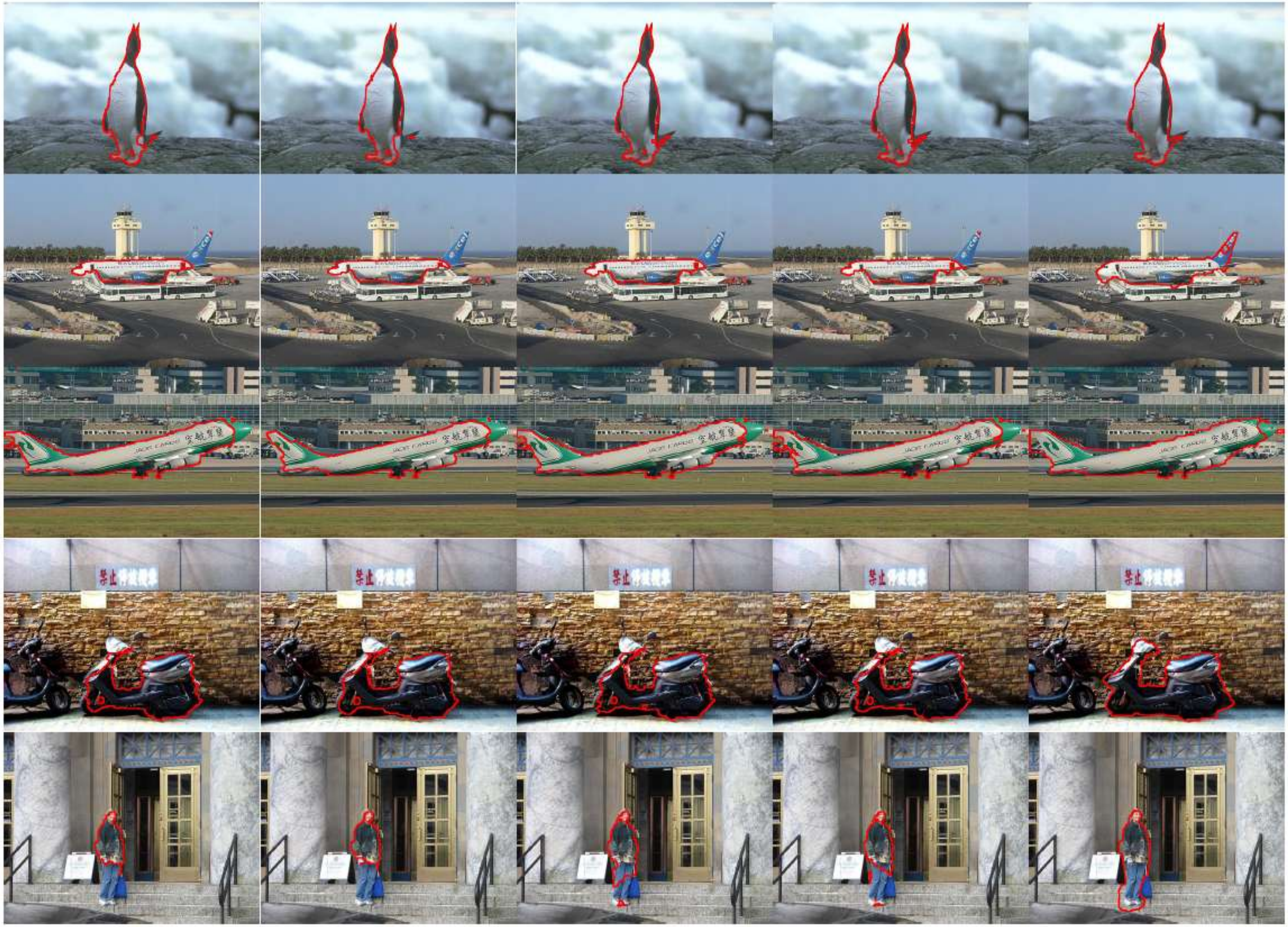
Experimental setup

- 60 images that are “more difficult”

[Gulshan et al., 2010; Pletscher and Kohli, 2012]

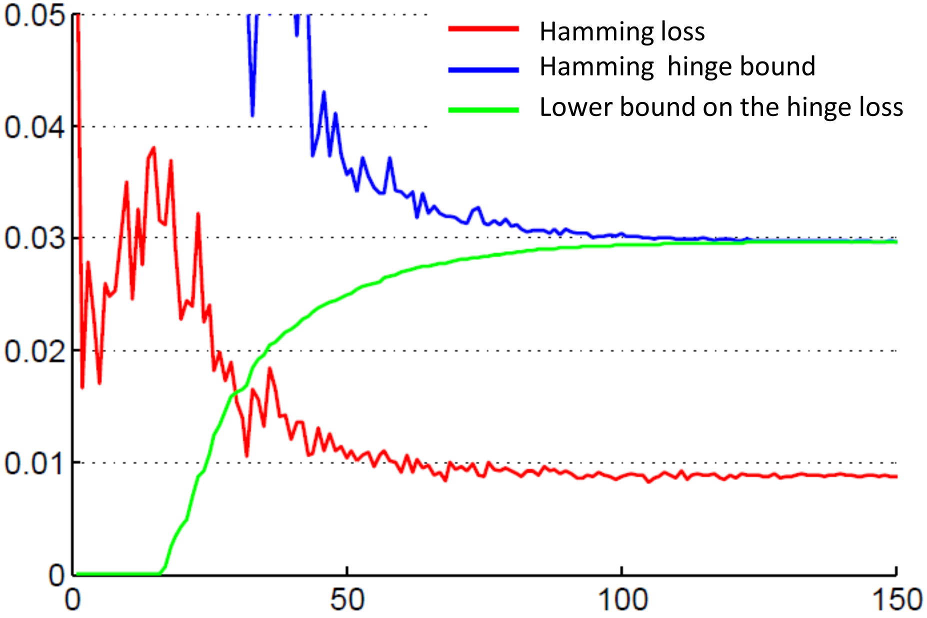
- 8-neighborhood
- 51 unary features:
  - color (GMM)
  - geodesic distances from seeds
- 6 pairwise features: (contrast sensitive) Potts
- Large enough seeds to make the problem easier

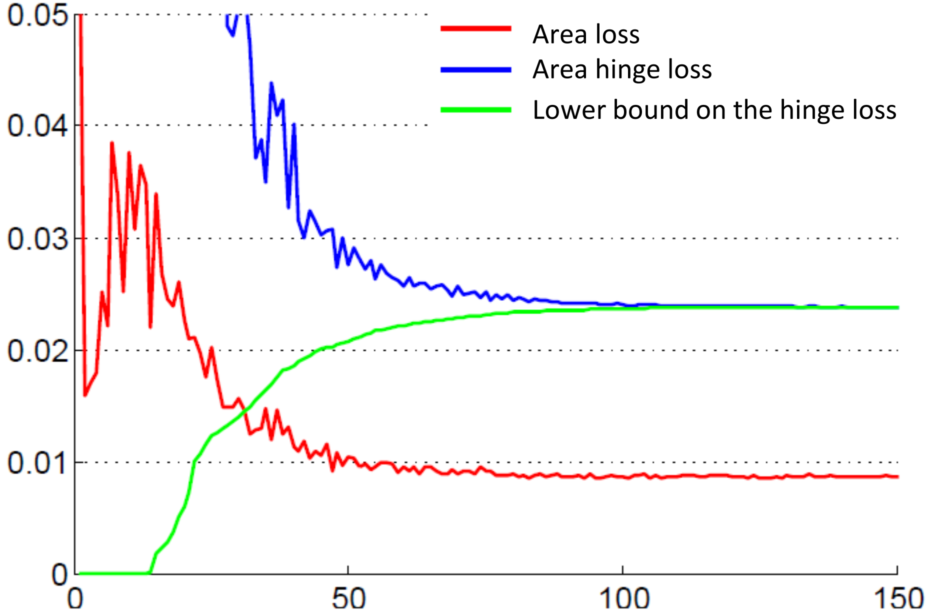
Results



Hamming loss | Weighted Hamming loss | HAC | Silhouette | Skeleton

Hinge bound is not tight!





Correlation between a loss and a hinge bound on the training set

Loss	H hinge	wH hinge	HAC hinge	area hinge
H	0.11	0.28	0.28	0.54
wH	0.04	0.27	0.31	0.63
HAC	-0.34	0.23	0.42	0.51
Area	0.20	0.16	0.23	0.74
J	-0.25	0.27	0.44	0.49

Training with one loss, testing with the other

Training Loss	Hamming		weighted Hamming		HAC		area	
	Train	Test	Train	Test	Train	Test	Train	Test
Hamming	1.1953	1.3319	12.222	12.613	4.4621	4.8987	0.5841	0.6743
HammingW	1.2014	1.4316	12.278	12.886	4.7536	5.2949	0.6924	0.8452
HAC	1.3431	1.4834	11.987	12.303	2.1226	2.3265	0.8563	0.8916
Area	1.1324	1.3091	11.901	12.408	3.7944	4.3215	0.3466	0.6807

Thank you!