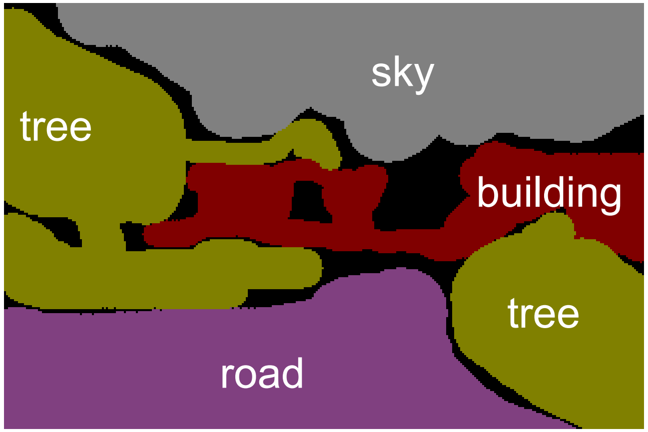
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Weakly Supervised Structured Output Learning for Semantic Segmentation

A. Vezhnevets, V. Ferrari, J. M. Buhmann
ETH Zurich

Semantic segmentation



- A task of *simultaneous* object *segmentation* and *recognition*
  - And we try to learn it weakly supervised – with seeing only image “tags” during training;

Weakly supervised training set

road
dog

road
cat

water
boat

water
boat
sky

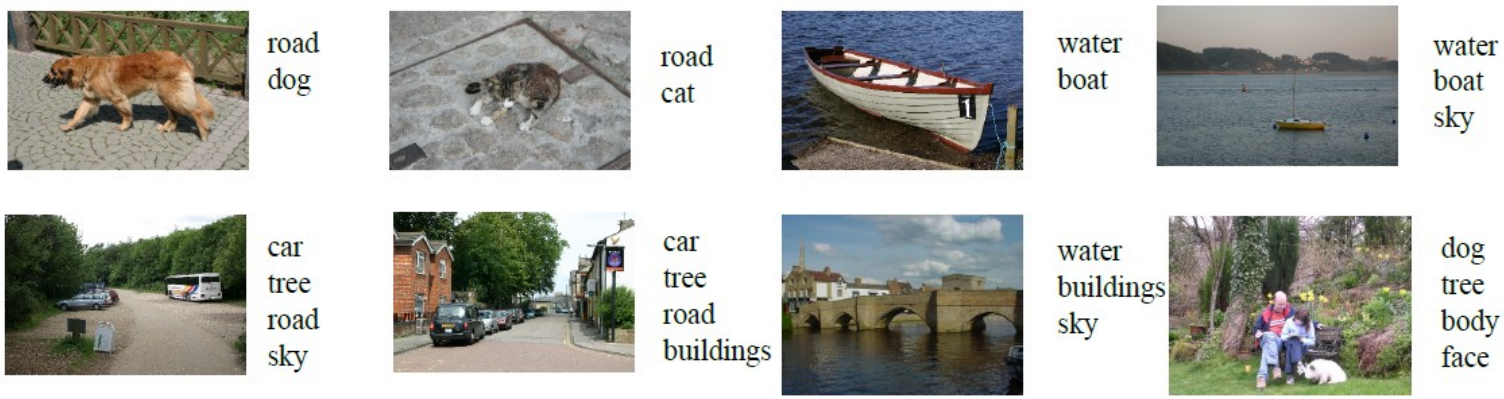
car
tree
road
sky

car
tree
road
buildings

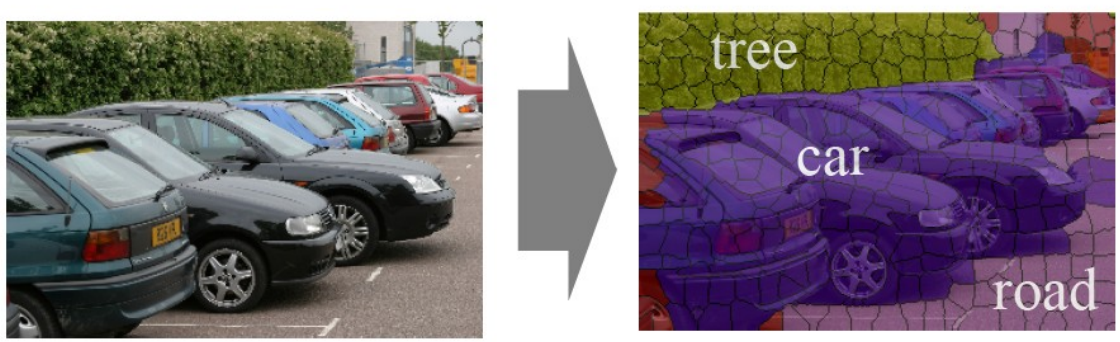
water
buildings
sky

dog
tree
body
face

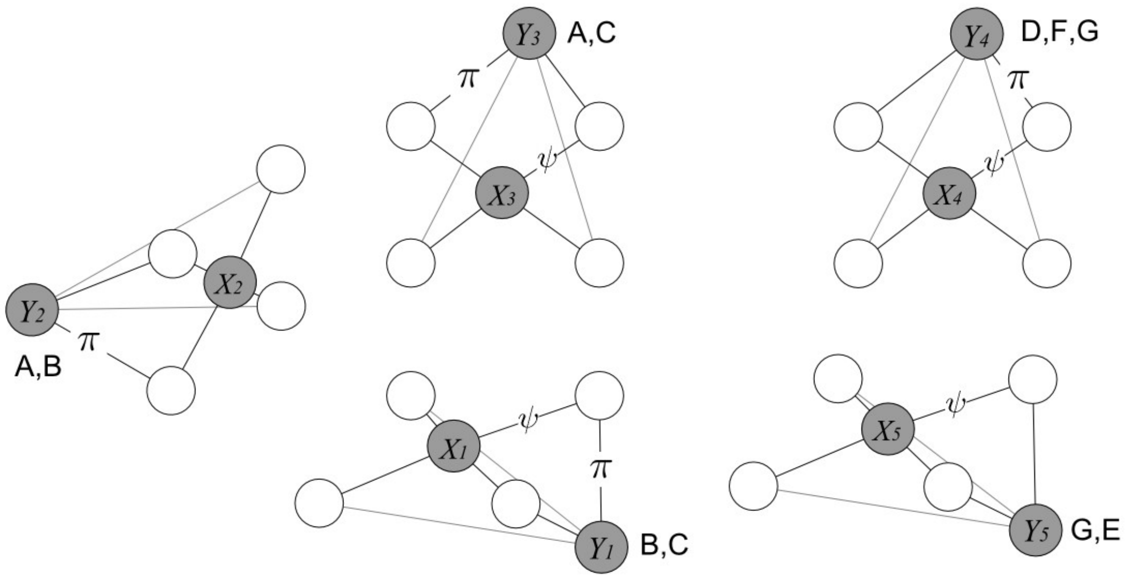
Weakly supervised training set



Semantic segmentation on test set

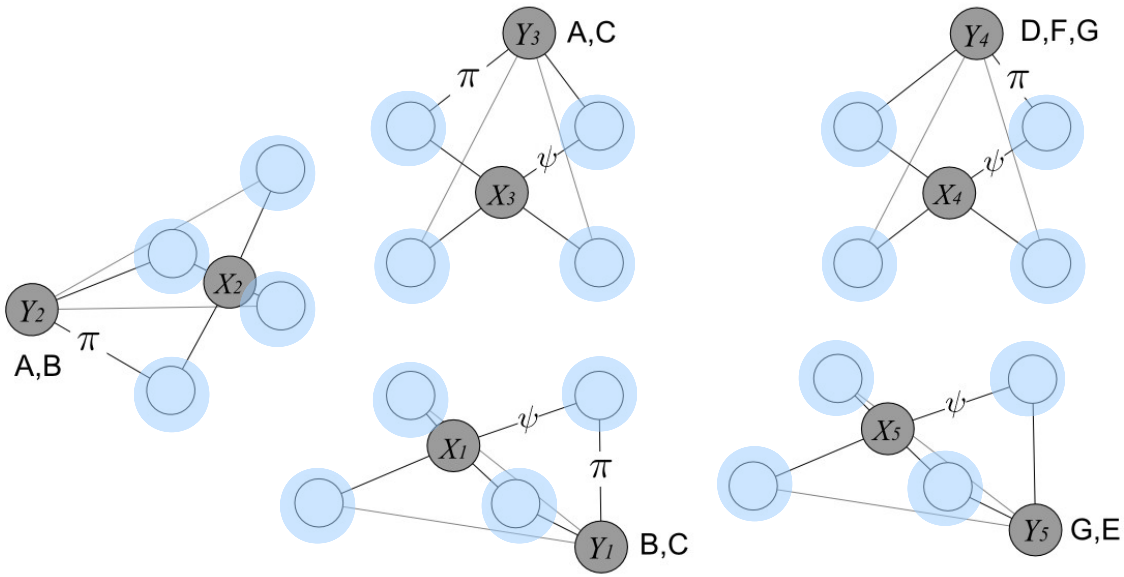


Constrained clustering



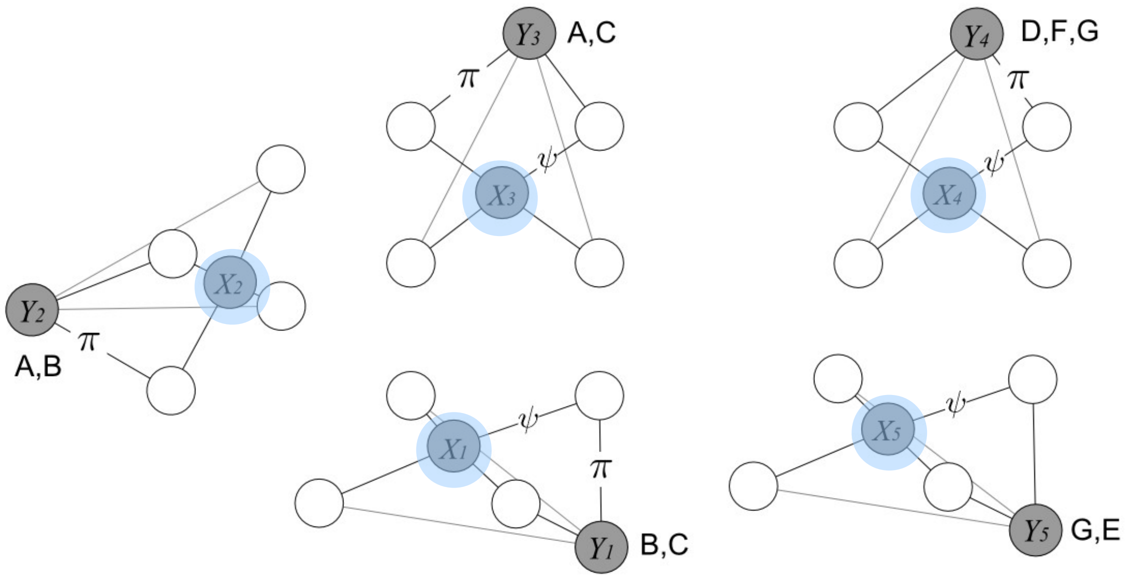
$$\mathcal{E}\left(\{y_i^j\}, \quad , \theta\right) = \sum_{x_i^j \in I^j ; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right)$$

Constrained clustering



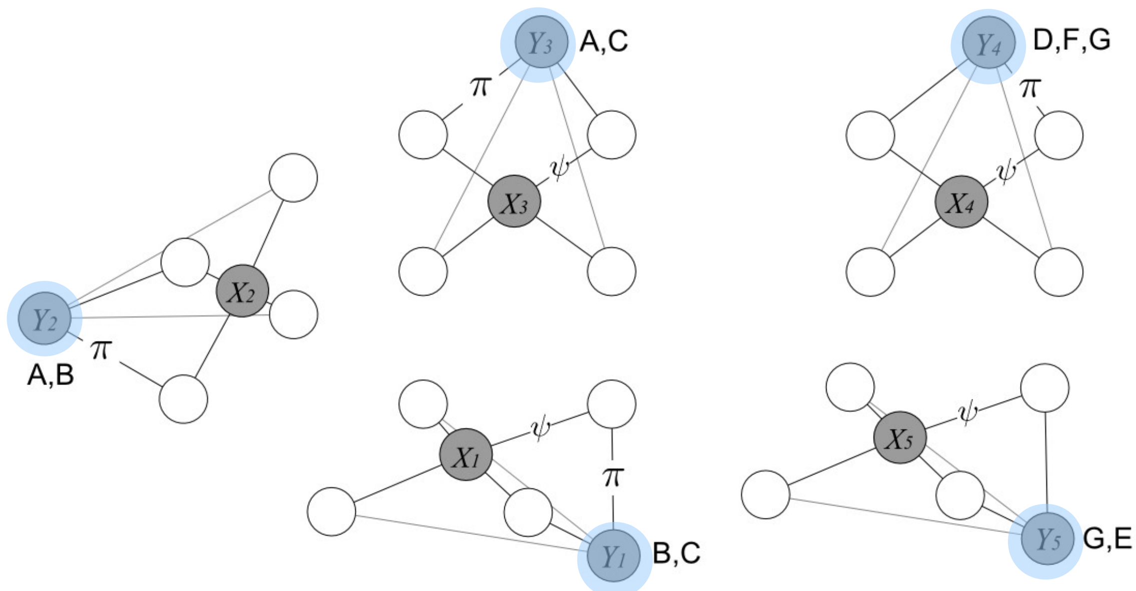
$$\mathcal{E}\left(\{y_i^j\}, \quad, \theta\right) = \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right)$$

Constrained clustering



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

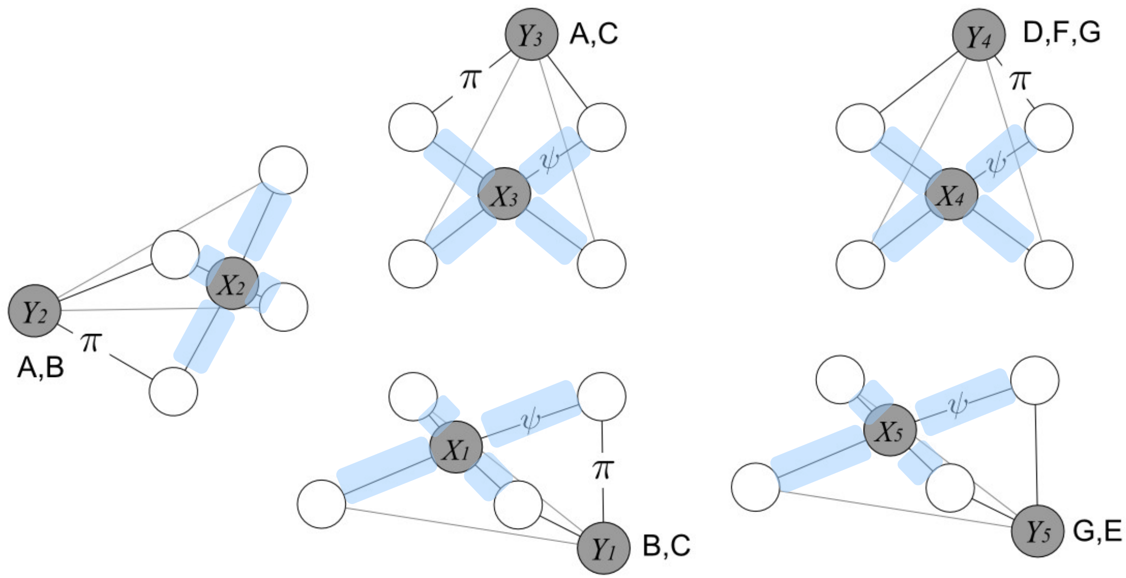
Constrained clustering



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

Unary potential: appearance likelihood for a superpixel

Constrained clustering

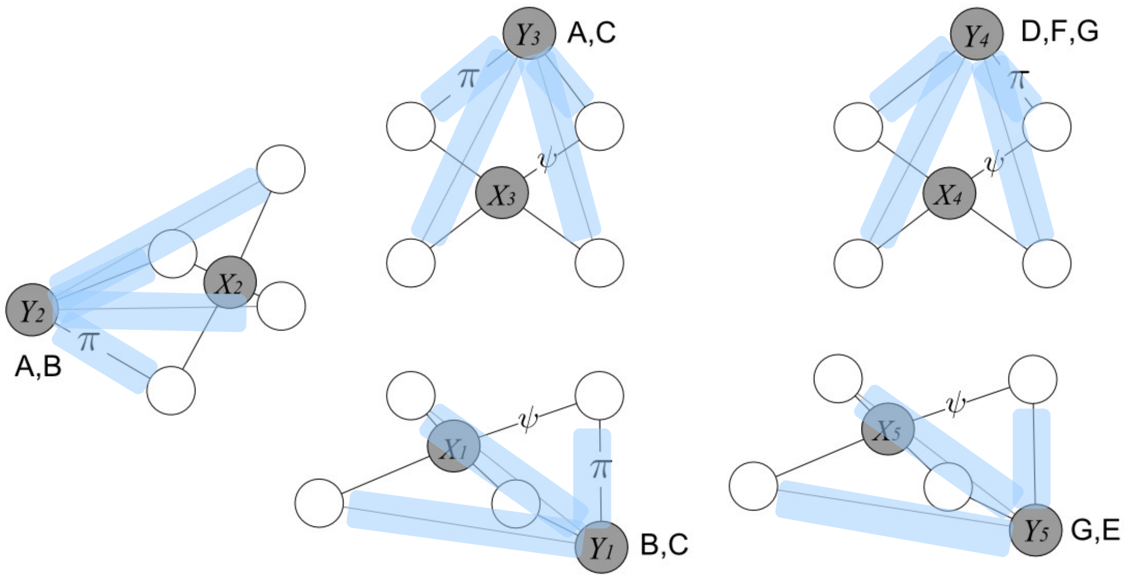


$$\mathcal{E}\left(\{y_i^j\}, \quad, \theta\right) = \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right)$$

$$\psi\left(y, x, \theta\right) = -\log f_y\left(x, \theta\right)$$

Unary potential: a superpixel can only take a label given to the image

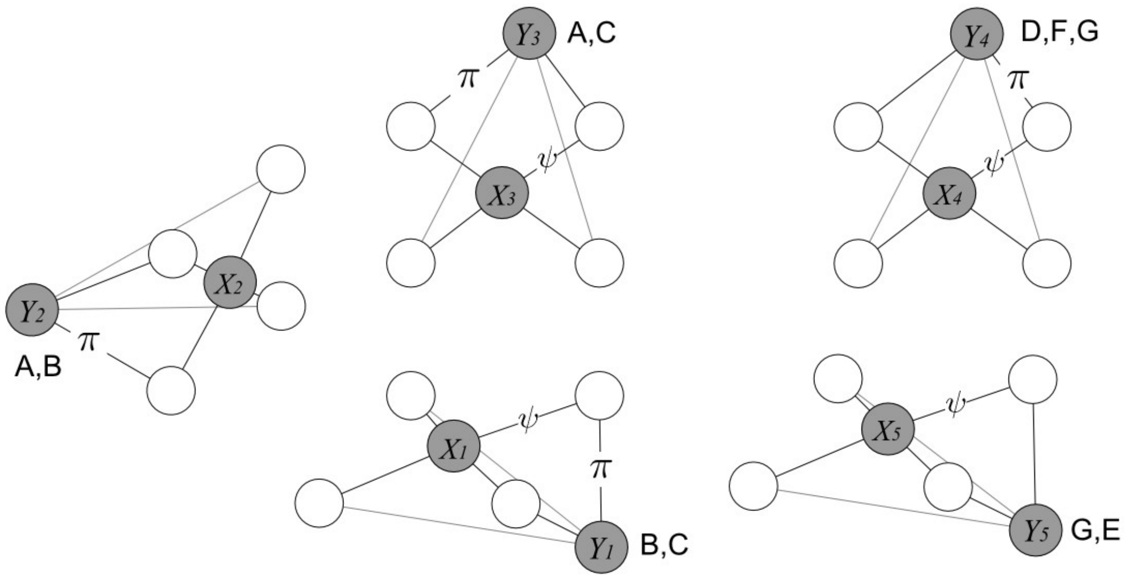
Constrained clustering



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

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

Constrained clustering



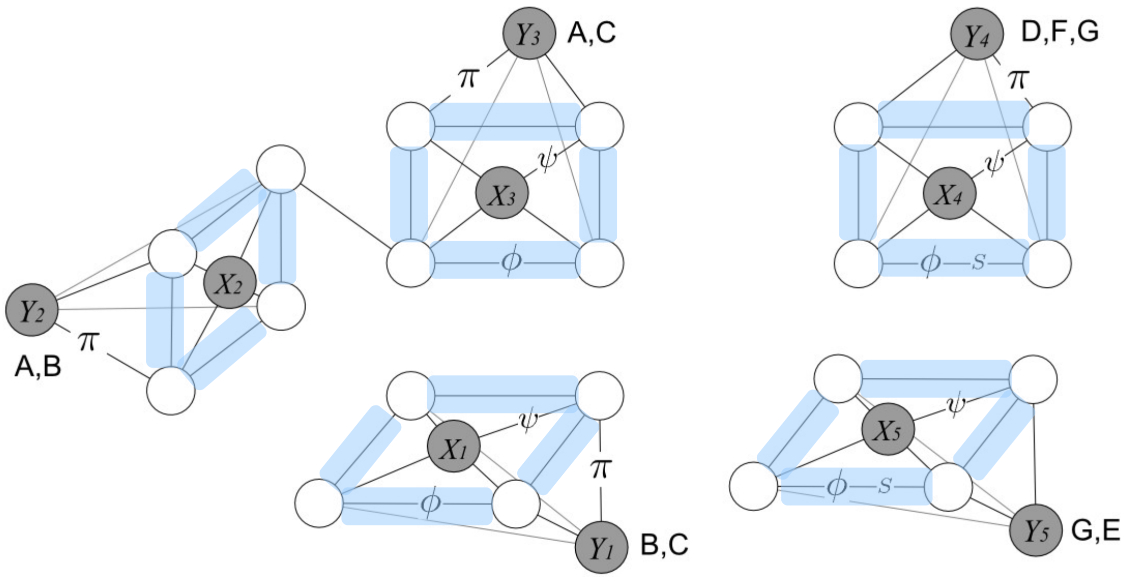
$$\mathcal{E}\left(\{y_i^j\}, \quad , \theta\right) = \sum_{x_i^j \in I^j ; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right)$$

- We can solve it using modified k-means (depending on an appearance model);
- Does not model all the dependencies in the data;
- Unregularized.

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

Pairwise potential within an image: encourages label smoothness

Pairwise potentials!



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

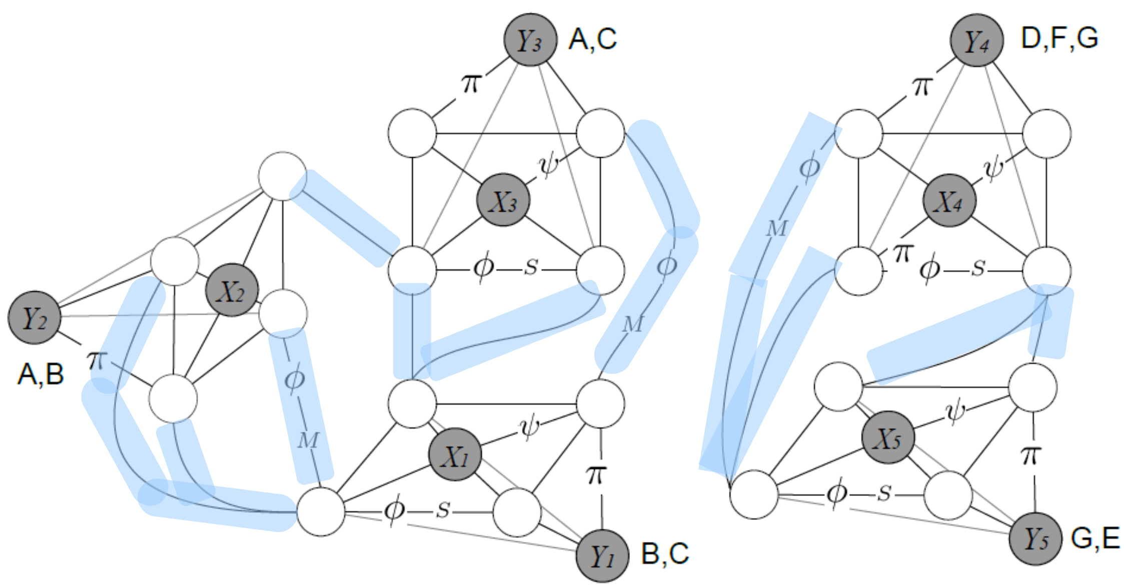
$$+ \sum_{(y_i^j, y_{i'}^{j'}) \in S} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right)$$

We can solve it using iterative minimization;

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

Pairwise potential **between** images:
similar superpixels → same label

Multi Image Potentials



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

$$+ \sum_{(y_i^j, y_{i'}^{j'}) \in S} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) + \sum_{(y_i^j, y_{i'}^{j'}) \in M} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right)$$

Building MIM: connect which images/superpixels ?

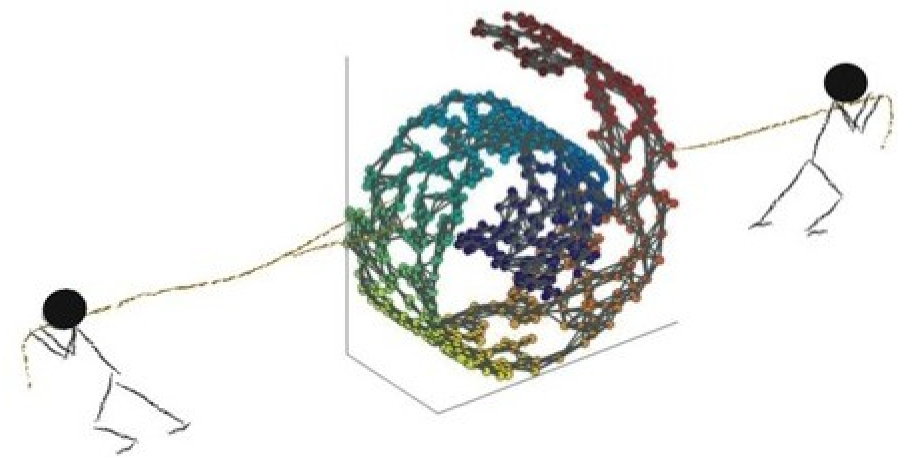
Data-driven construction of a *sparse* set of connections

Connect each superpixel to

- k nearest appearance neighbors in other images *sharing labels*
- <= p superpixels in one image (variety)

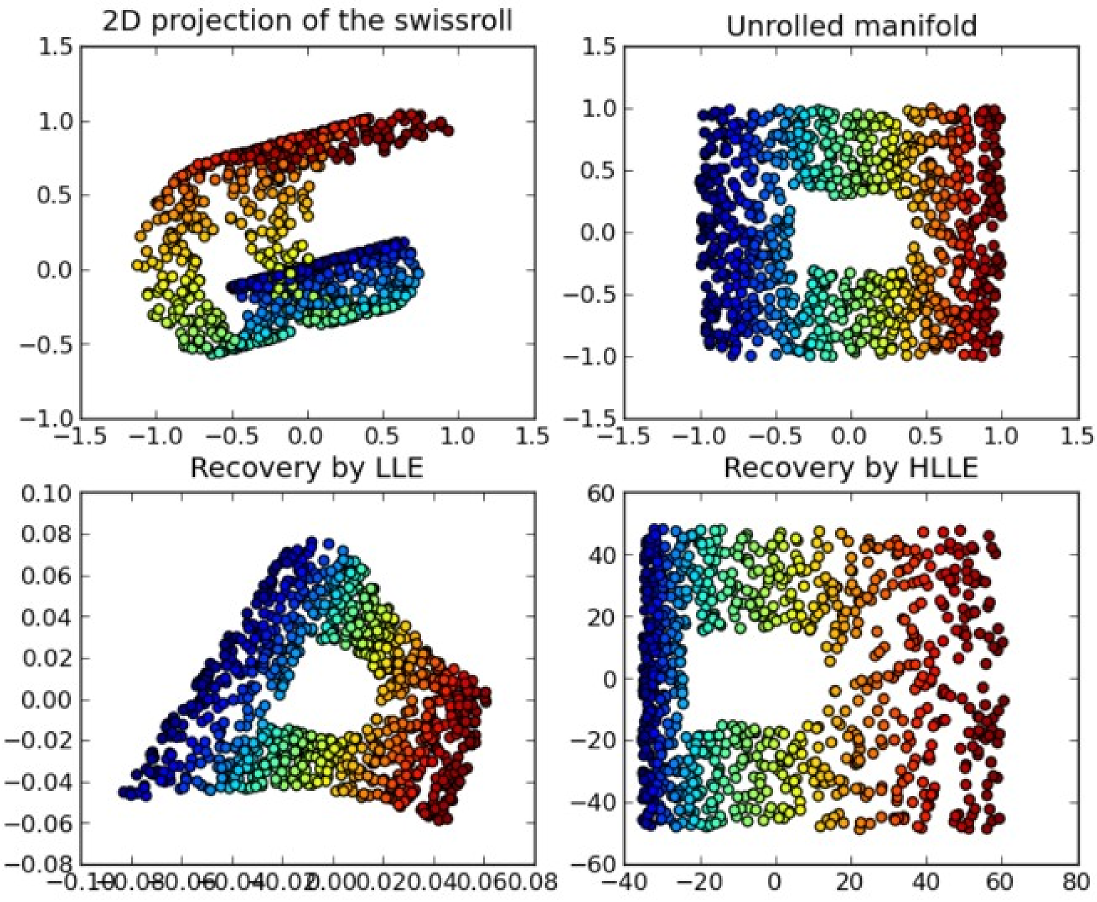
Why do MI potentials make any sense?

- A nearest neighbour graph can be interpreted as a model for a manifold, on which data lives;
- The pairwise potential penalizes cutting through the manifolds (areas of high density):
  - A very similar regularizer is used in semi-supervised learning and dimensionality reduction;
  - In a certain sense, it "unrolls" the manifolds;
  - There is a relation to graph Laplacians and Laplace-Beltrami operator;



In essence, we penalize labelling for changing on the manifolds formed by superpixels from images that share a label

*Image courtesy of K.Q. Weinberger



2D projection of the swissroll
Unrolled manifold
Recovery by LLE
Recovery by HLLE

Inference: get y given *Θ*

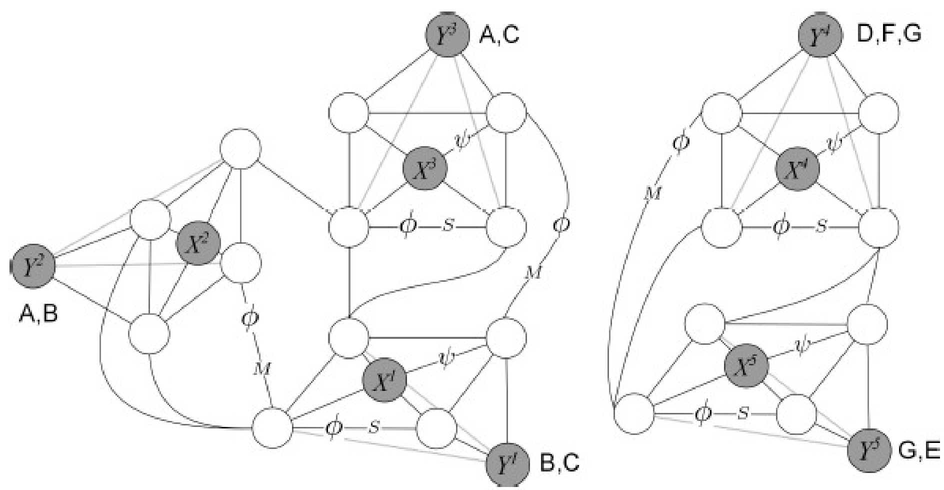
Energy minimization

MIM potentials are multi-label submodular

→ alpha-expansion
[Boykov et al PAMI 2011]

$$\mathcal{E}\left(\{y_i^j\},\theta\right) = \sum_{x_i^j \in I^j; I^j \in \tau} \psi\left(y_i^j, x_i^j, \theta\right) \cdot 1_{y_i^j \in Y_i^j} +$$

$$\sum_{(y_i^j, y_{i'}^{j'}) \in S} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) + \sum_{(y_i^j, y_{i'}^{j'}) \in M} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right)$$



Learning: get y *and Θ*

$$\mathcal{E}\left(\{y_i^j\},\theta\right) = \sum_{x_i^j \in I^j; I^j \in \tau} \psi\left(y_i^j, x_i^j, \theta\right) \cdot 1_{y_i^j \in Y_i^j} +$$

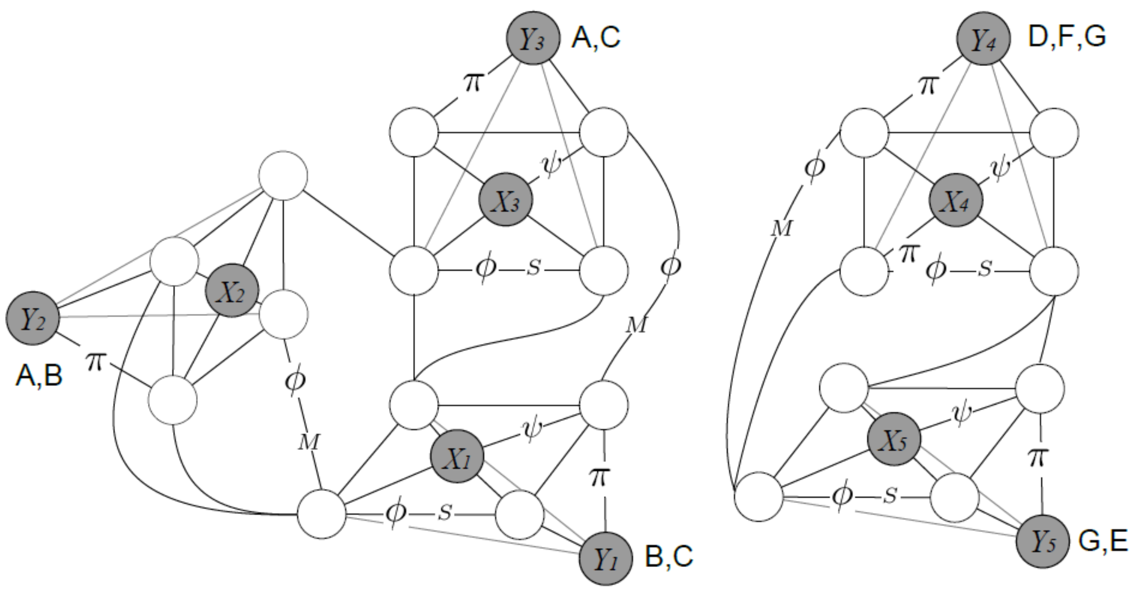
$$\sum_{(y_i^j, y_{i'}^{j'}) \in S} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) + \sum_{(y_i^j, y_{i'}^{j'}) \in M} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right)$$

- Problem of minimizing the energy is mixed continuous/discrete
- Energy is
  - Convex, if labeling y if fixed;
  - Metric, if θ is fixed;
- Iterative minimization:

Init: set y to random labels fulfilling image label constraints

1. Fix y, train θ in standard ‘supervised’ maximum likelihood
2. Fix θ, use alpha-expansion (previous slide);

MIM (as of ICCV'11)

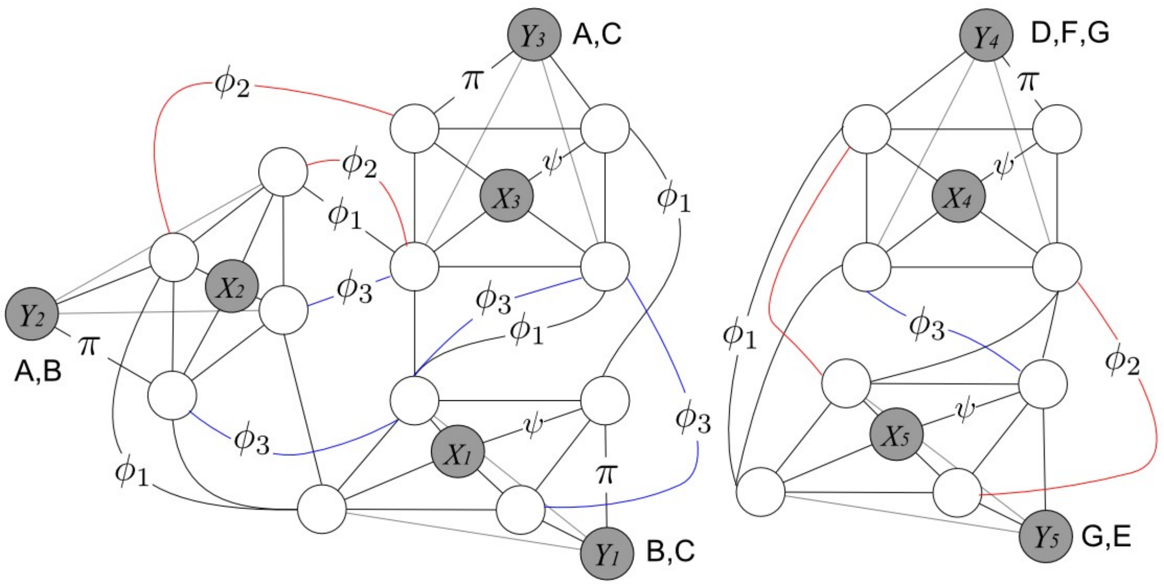


What is missing?

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

$$+ \sum_{(y_i^j, y_{i'}^{j'}) \in S} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) + \sum_{(y_i^j, y_{i'}^{j'}) \in M} \phi\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right)$$

Generalized MIM

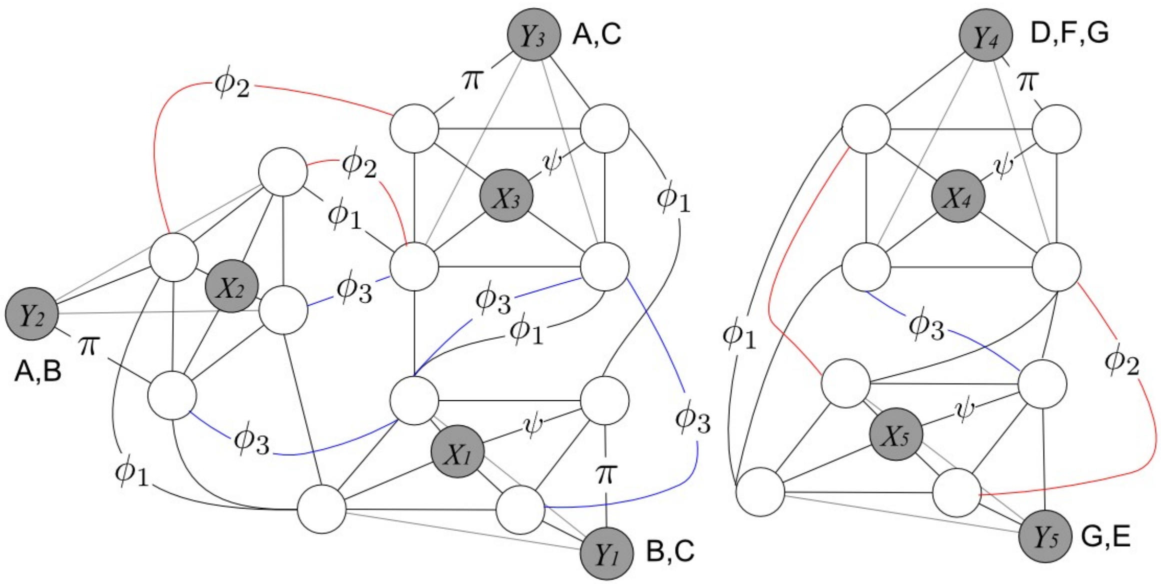


$$\mathcal{E}\left(\{y_i^j\}, \boldsymbol{\alpha}, \theta\right) = \alpha_0 \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right) +$$

$$(1-\alpha_0)\sum_{k=1}^{K} \alpha_k \left(\sum_{(y_i^j, y_{i'}^{j'}) \in E_k} \phi_k\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right)\right)$$

Each potential □ a different metric of similarity (colour, SIFT, texture, e.t.c.)

Generalized MIM

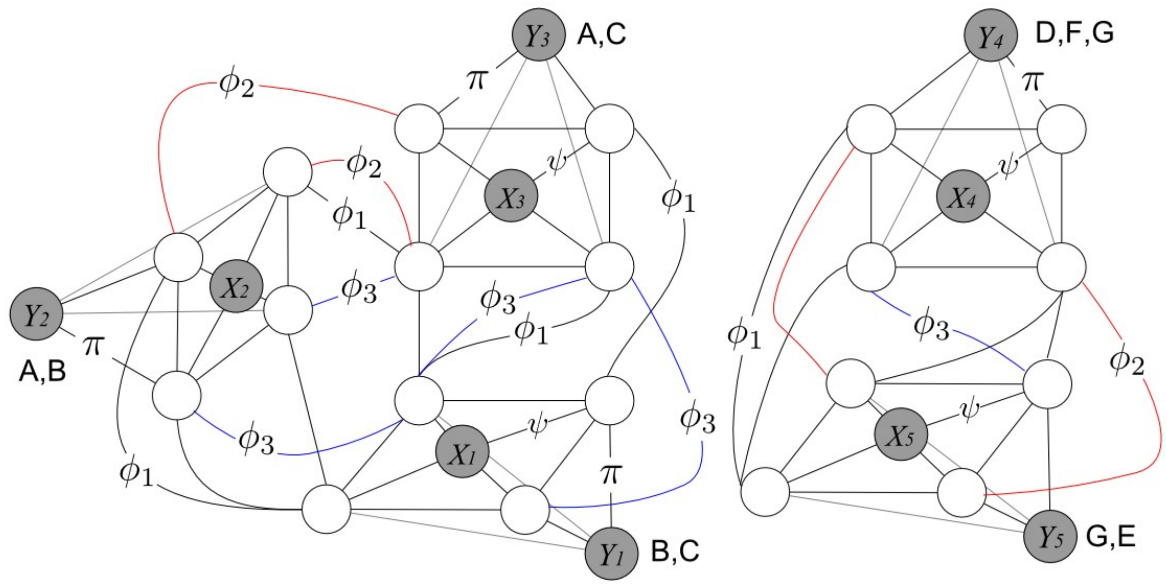


$$\mathcal{E}\left(\{y_i^j\},\boldsymbol{\alpha},\theta\right)=\alpha_0\sum_{x_i^j\in I^j;I^j\in\tau}\left(\psi\left(y_i^j,x_i^j,\theta\right)+\pi(y_i^j,Y_i^j)\right)+$$

$$(1-\alpha_0)\sum_{k=1}^{K}\alpha_k\left(\sum_{(y_i^j,y_{i'}^{j'})\in E_k}\phi_k\left(y_i^j,y_{i'}^{j'},x_i^j,x_{i'}^{j'}\right)\right)$$

New *structure parameters* vector **α** controls the regularization

Generalized MIM



$$\mathcal{E}\left(\{y_i^j\}, \boldsymbol{\alpha}, \theta\right) = \alpha_0 \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right) +$$

Balance unary vs pairwise

Balance different pairwise potentials

$$(1-\alpha_0)\sum_{k=1}^{K} \alpha_k \left(\sum_{(y_i^j, y_{i'}^{j'}) \in E_k} \phi_k\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right)\right)$$

GMIM - questions

$$\mathcal{E}\left(\{y_i^j\}, \boldsymbol{\alpha}, \theta\right) = \alpha_0 \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right) +$$

$$(1-\alpha_0) \sum_{k=1}^{K} \alpha_k \left(\sum_{(y_i^j, y_{i'}^{j'}) \in E_k} \phi_k\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) \right)$$

- If alpha is fixed, we know what to do
  - Iterative minimization from ICCV'11

- But how to choose **α**?
  - Since it controls the strength and the form of regularization, we cannot use energy itself to select it;
    - Trivial solution = minimal regularization:
      - **α**_0 = 1;
  - Our problem is weakly supervised, therefore we can't use cross-validation;

Model selection for GMIM

$$\mathcal{E}\left(\{y_i^j\}, \boldsymbol{\alpha}, \theta\right) = \alpha_0 \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right) +$$

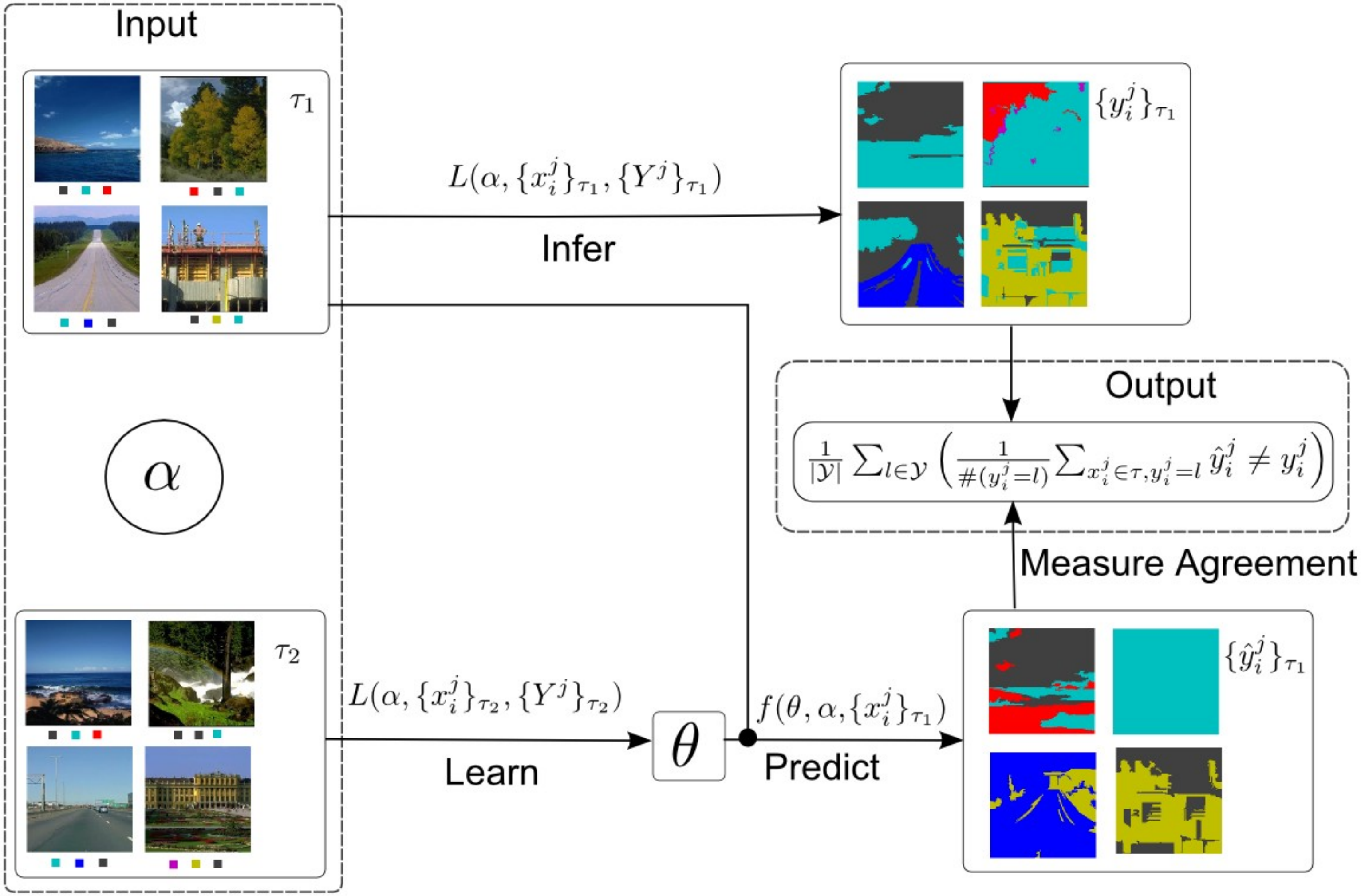
$$(1-\alpha_0) \sum_{k=1}^{K} \alpha_k \left(\sum_{(y_i^j, y_{i'}^{j'}) \in E_k} \phi_k\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right) \right)$$

- Model selection view:
  - Every value of **α** defines a model;
  - Space of all **α** span a family of models;

- We wish to select a model out the family:
  - We need a meta-principle to score models;
  - We need a practical search algorithm to find the one with the best score;

Expected agreement $\mathcal{A}(\boldsymbol{\alpha})$

Labeling *inferred* and labelling *predicted* should *agree!*



Learning and inference algorithm:

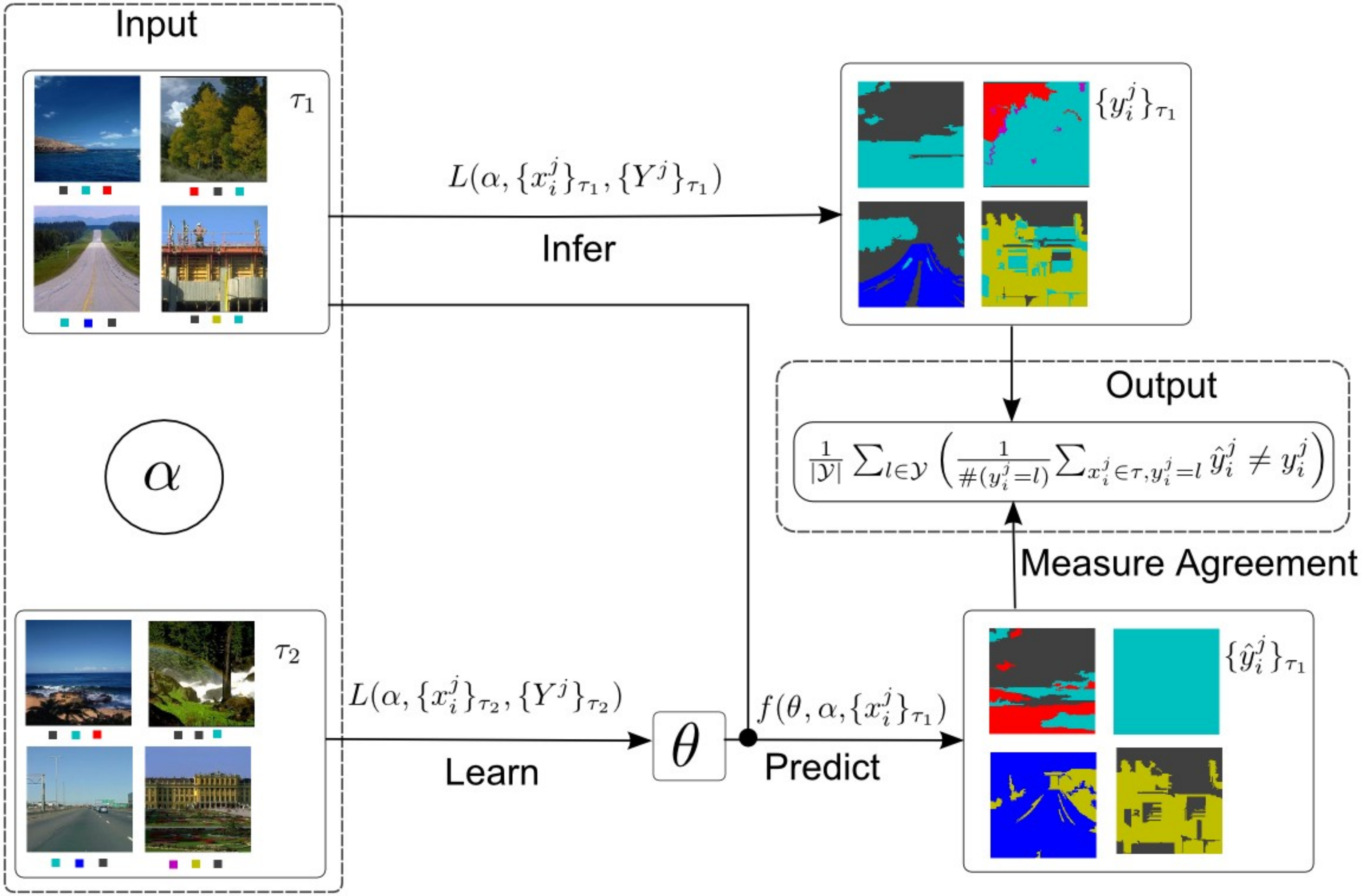
$$L : \left(\boldsymbol{\alpha}, \{x_i^j\}_{\tau_1}, \{Y^j\}_{\tau_1}\right) \to \left(\theta, \{y_i^j\}_{\tau_1}\right)$$

Prediction algorithm:

$$f : \left(\theta, \boldsymbol{\alpha}, \{x_i^j\}_{\tau_2}\right) \to \{\hat{y}_i^j\}_{\tau_2}$$

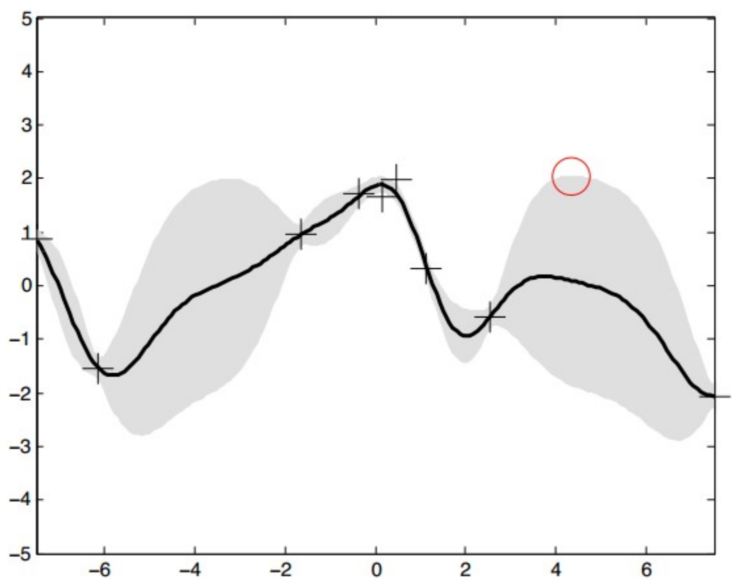
Expected agreement $\mathcal{A}(\alpha)$

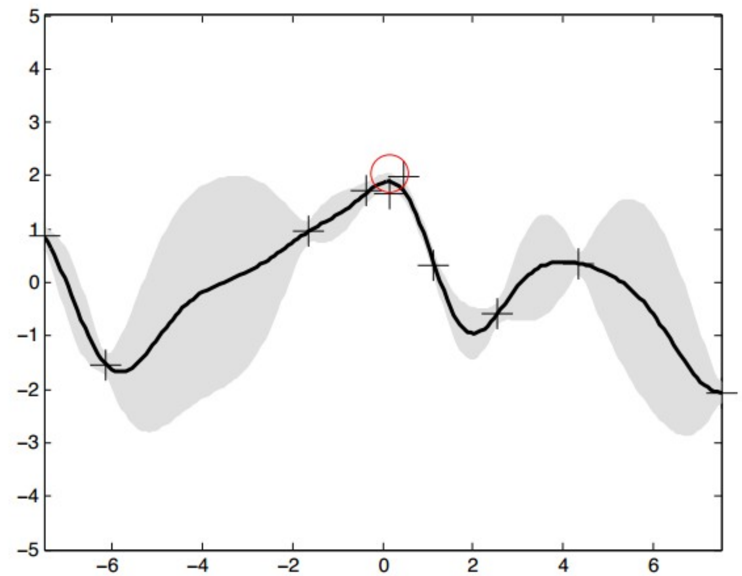
Labeling *inferred* and labelling *predicted* should *agree!*



- **α** lives in a multidimensional space (~7);
- No gradients available;
- How do we search for best **α**?

Bayesian optimization with GP





Model expected agreement as GP:

$$\mathcal{A}(\boldsymbol{\alpha}) \sim \mathcal{GP}\left(m(\boldsymbol{\alpha}), k(\boldsymbol{\alpha}, \boldsymbol{\alpha}')\right)$$

With a Gaussian kernel:

$$k(\boldsymbol{\alpha}, \boldsymbol{\alpha}') = \gamma \exp\left(-\frac{1}{2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}')^T \mathrm{diag}(\boldsymbol{v})^{-2}(\boldsymbol{\alpha} - \boldsymbol{\alpha}')\right)$$

Next point □ Upper Confidence Bound:

$$\boldsymbol{\alpha}_{t+1} := \mu_t(\boldsymbol{\alpha}_{t+1}) + \beta\sigma_t^2(\boldsymbol{\alpha}_{t+1})$$

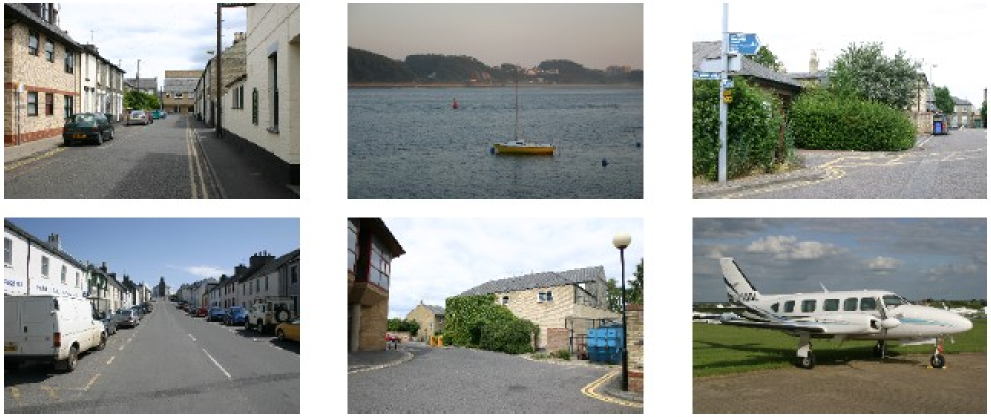
Srinivas et al. ICML 10

*Image courtesy of A. Krause

Prediction with MIM

Test image

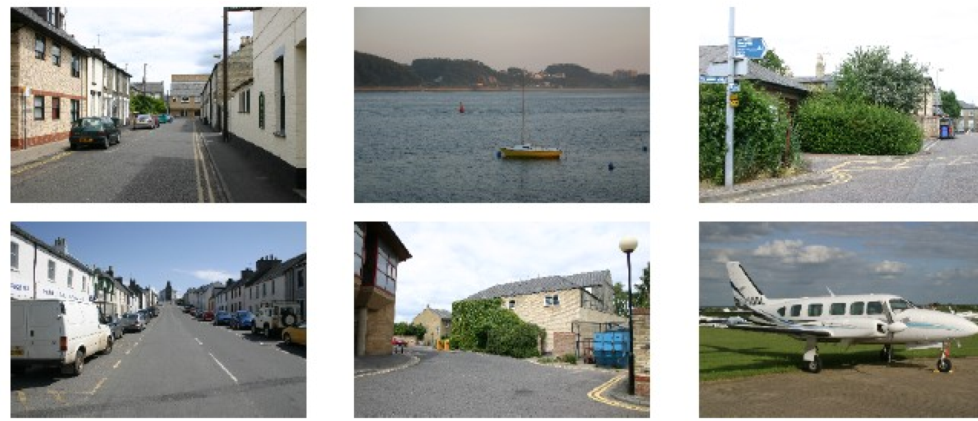
Retrieved, similar training images



Prediction with MIM

Test image

Retrieved, similar training images



Y_1 B,C Y_2 A,B Y_3 A,C

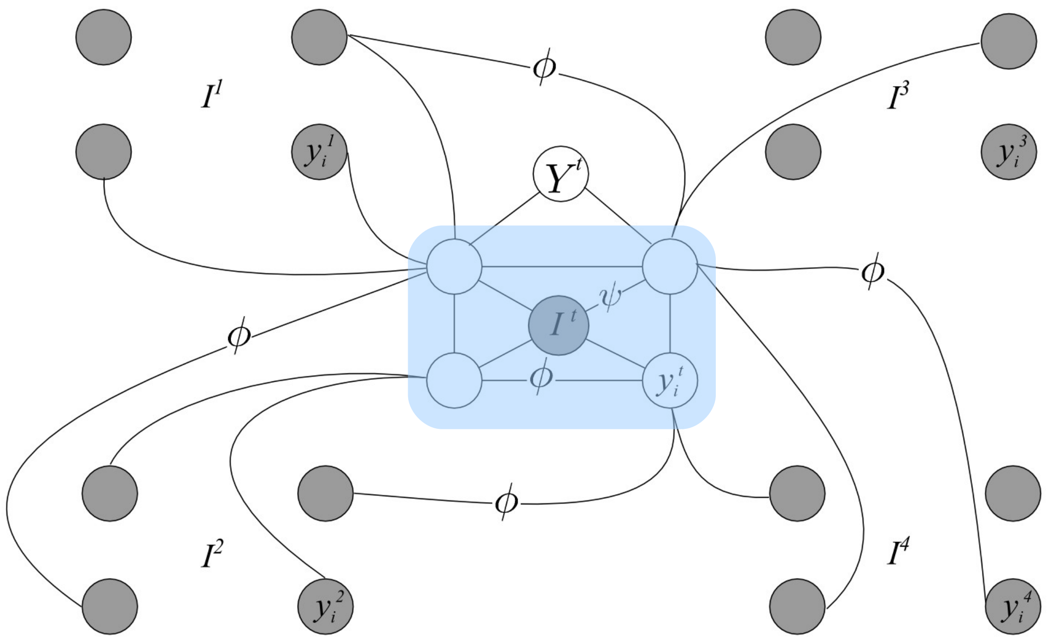
Y_4 D,F,G Y_5 G,E Y_6 G,E,A

$$\mu\left(y_i^t\right) = -\log P\left(y_i^t \in Y^t\right)$$

Shotton et al. CVPR 08; Guillaumin et al ICCV 09

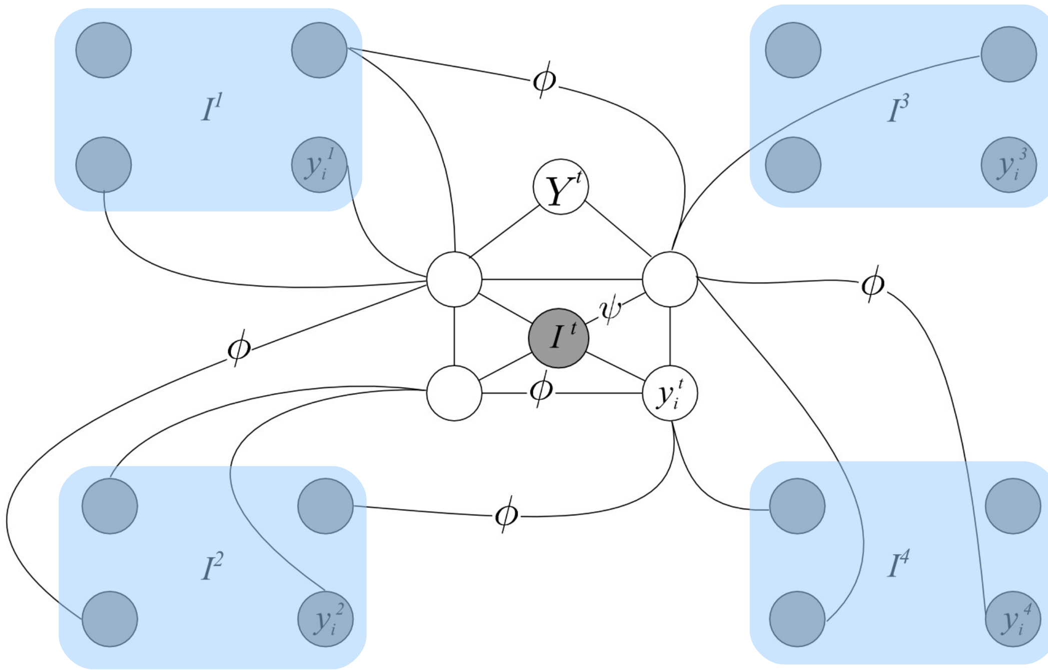
Prediction with MIM

Test image

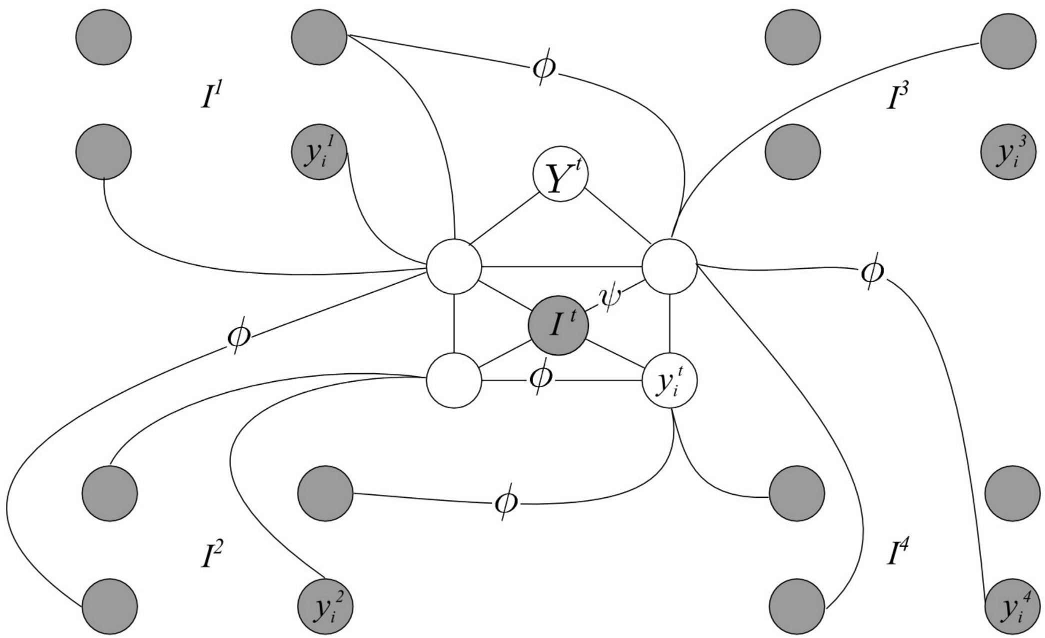


Prediction with MIM

Training images



Prediction with MIM



$$\mathcal{E}\left(\{y_i^t\}\right) = \alpha_0^* \sum_i \left(\psi\left(y_i^t, x_i^t, \theta^*\right) + \mu\left(y_i^t, I^t\right)\right) +$$

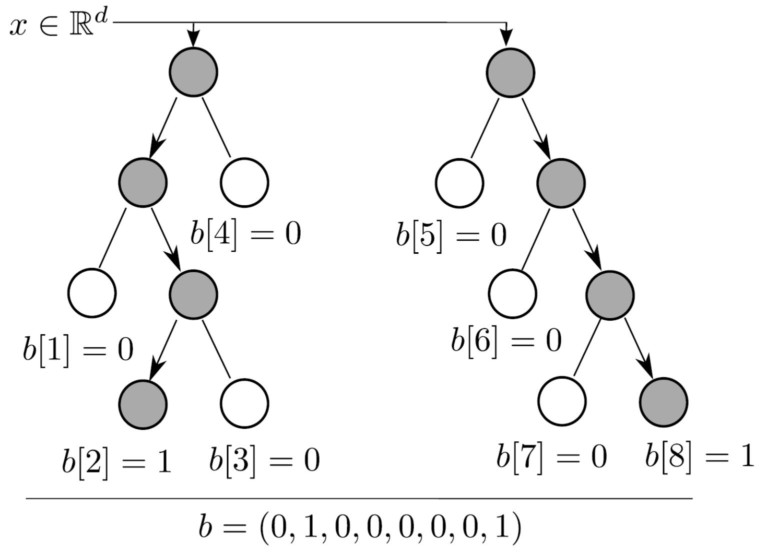
$$+ \left(1 - \alpha_0^*\right) \sum_{k=1}^{K} \alpha_k^* \left(\sum_{(y_i^t, y_{i'}^j) \in E_k^t} \phi_k\left(y_i^t, y_{i'}^j, x_i^t, x_{i'}^j\right) \right)$$

Appearance models via ERHF

Extremely Randomized Hashing Forest

- Requirements:
  - Fast in training – we re-estimate them iteratively many-many times during training;
  - Leverage diverse features – visual classes are very varied in appearance;

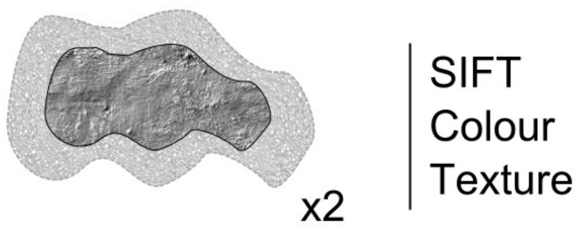
- Extremely Randomized Hashing Forest representation
  - A forest of decision trees;
  - Built upon <u>any</u> feature set;
  - Every predicate in a tree is a hashing function;



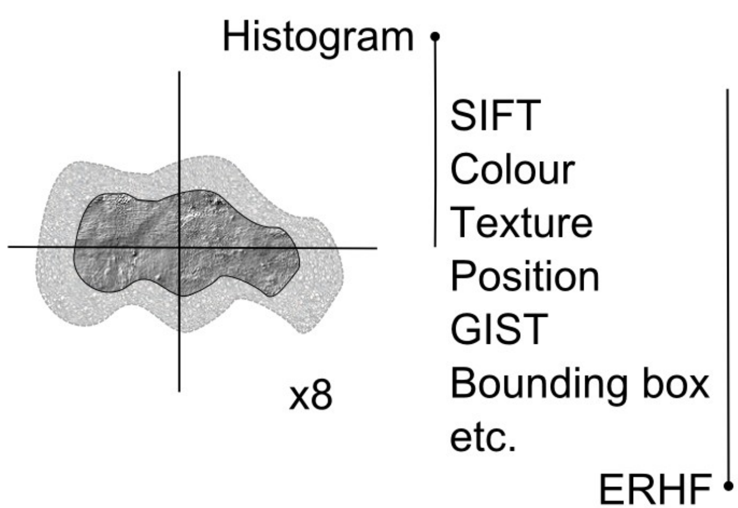
- Naive Bayes:
  - 3 (sparse) matrix multiplications to retrain the model;
  - Easy to weight data (according to amount of pixels in the superpixel;

Implementation details

- Features for similarity metrics
  - 3 different features;
  - Over superpixel and dilated area;
  - Chi-square distance;
  - 2x3=6 pairwise potentials.



- Superpixel features for appearance models
  - Only histogram features (baseline)
    - 1248
  - ERHF with a full set
    - 3115



J. Tighe and S. Lazebnik ECCV 10

GMIM vs state of the art

Method	[1]	[2]	[3]	[4]	GMIM
supervision	full	full	full	weak	weak
average acc.	13	24	29	14	21

- Data - LabelMe subset of [2]:
  - 2.5K images, 33 classes;
- Quality metric:
  - Average per class accuracy

1) J. Shotton, J.Winn, C. Rother, and A. Criminisi, “Textonboost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation,” in ECCV, 2006.
2) C. Liu, J. Yuen, and A. Torralba, “Nonparametric scene parsing: label transfer via dense scene alignment.,” in CVPR, 2009.
3) J. Tighe and S. Lazebnik, “Superparsing: Scalable nonparametric image parsing with superpixels,” in ECCV, 2010.
4) MIM of ICCV'11

Evaluation of components

ERHF		Histograms	
Setup	Av. acc.	Setup	Av. acc.
MEA	21	MEA	19
average	6	average	5
best* α	21	best* α	20
best* α_0 + average	17	best* α_0 + average	17

- MEA – Maximum Expected Agreement
  - Full framework
- Baselines:
  - average – set α to [0.5 1/k … 1/k];
  - *best α – grid search, looking at training set pixel labels;
  - *best α_0 + average – grid search for best α_0 looking at training set pixel labels and average for the rest;
  - ERHF vs set of histogram features;

Pictures!

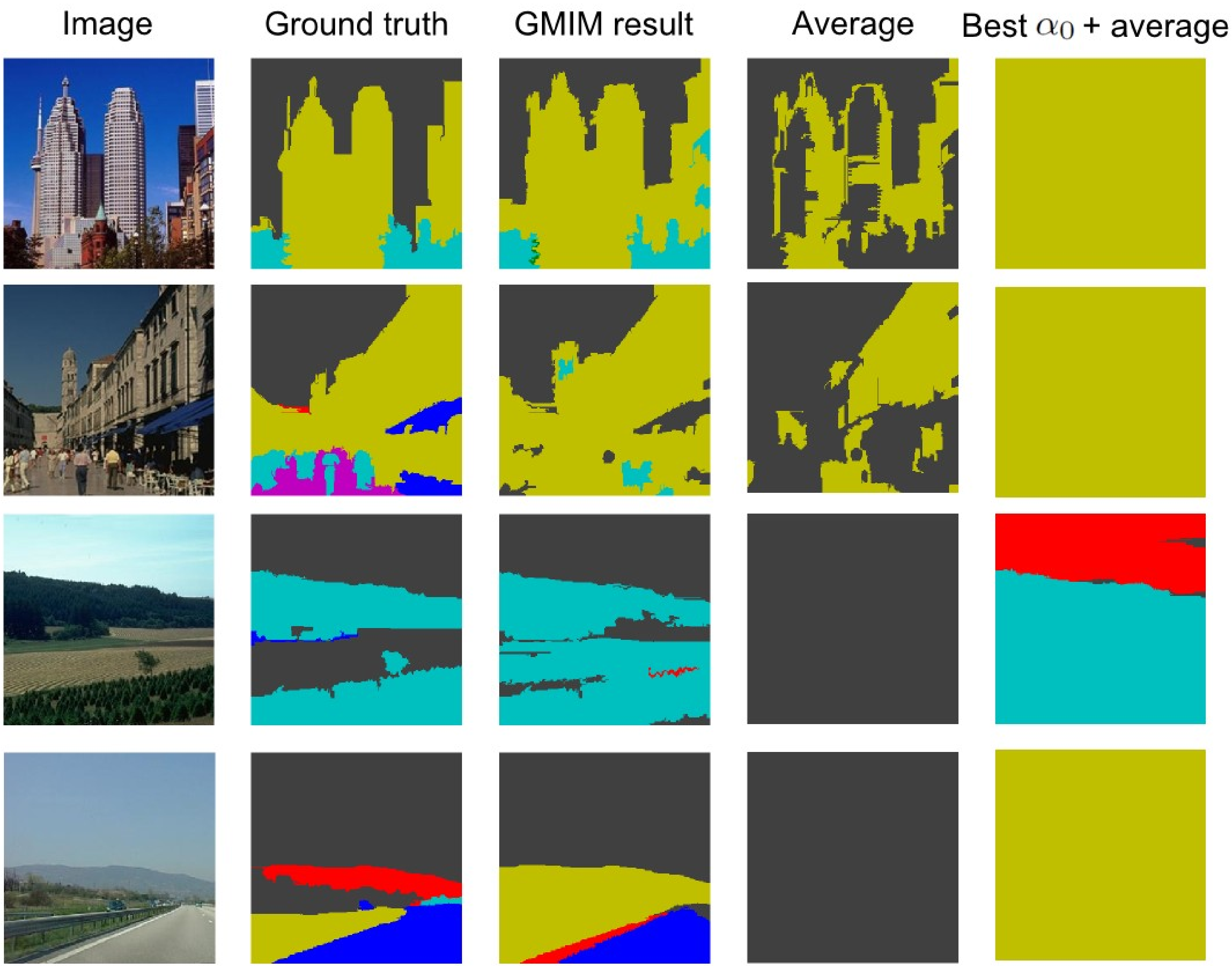


Image
Ground truth
GMIM result
Average
Best α_0 + average

Discussion: Framework

$$\mathcal{E}\left(\{y_i^j\}, \boldsymbol{\alpha}, \theta\right) = \alpha_0 \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right) +$$

$$(1-\alpha_0) \sum_{k=1}^{K} \alpha_k \left(\sum_{(y_i^j, y_{i'}^{j'}) \in E_k} \phi_k\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right)\right)$$

- Ingredients:
  - Regularizer form;
  - Criterion to select regularizer's strength and structure;
  - Way to search for the best one;
  - Fast and rich appearance models;

Discussion: Framework extensions

$$\mathcal{E}\left(\{y_i^j\}, \boldsymbol{\alpha}, \theta\right) = \alpha_0 \sum_{x_i^j \in I^j; I^j \in \tau} \left(\psi\left(y_i^j, x_i^j, \theta\right) + \pi(y_i^j, Y_i^j)\right) +$$

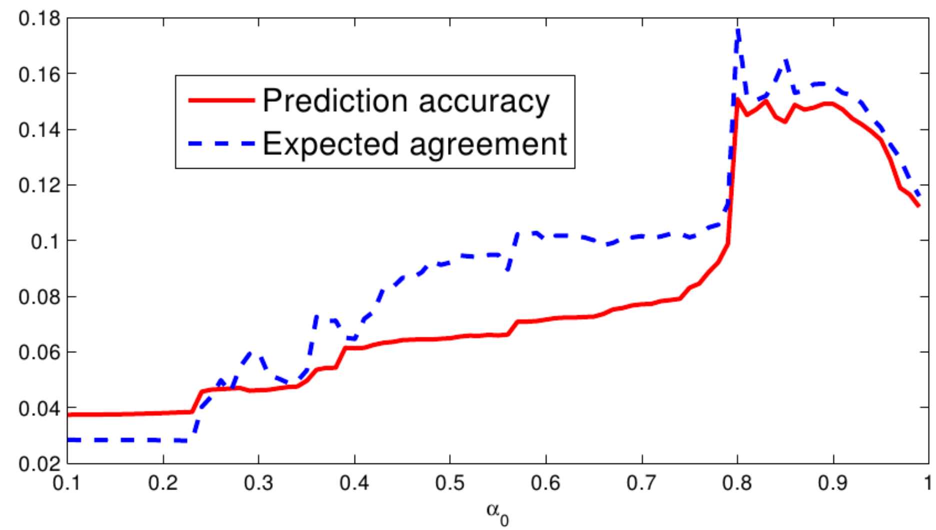
$$(1-\alpha_0)\sum_{k=1}^{K} \alpha_k \left(\sum_{(y_i^j, y_{i'}^{j'}) \in E_k} \phi_k\left(y_i^j, y_{i'}^{j'}, x_i^j, x_{i'}^{j'}\right)\right)$$

- Higher order potentials (or whichever buzzword you like)
  - More constraints from bag labels;
  - Hierarchical structure of superpixels/regions;
  - ...
  - → Can do, if you can do the inference;

- Pimp up appearance models
  - As long as you can train them;

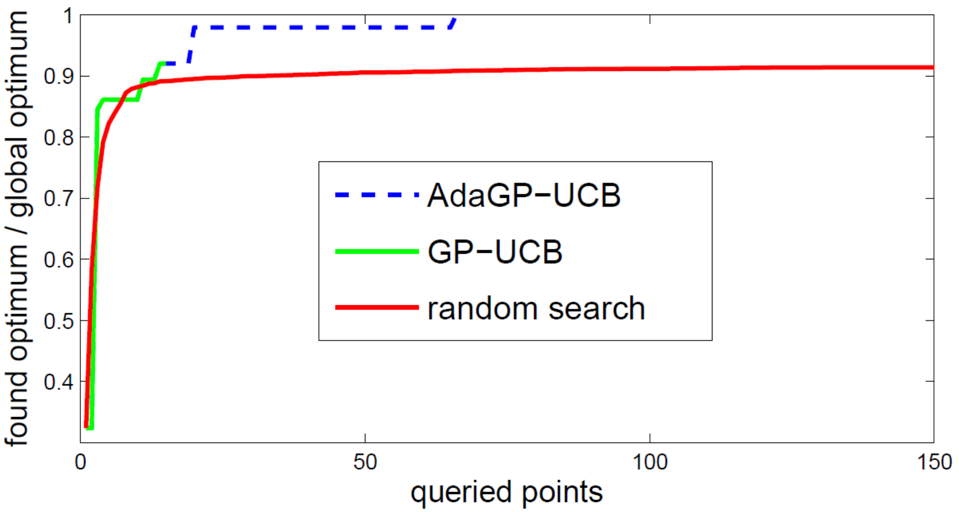
Discussion: Model selection

- Fast to evaluate
  - I have run it for ~100 parameter values now, but more in the future;
- Model-agnostic (better be)
  - Calculating probabilities or anything like it is a computational nightmare;
  - Changing model shouldn't change the validation;
  - Best – just work with strings of labels inferred/predicted;
  - VMI by Alberto could work!
- No i.i.d. Assumptions
  - Removing i.i.d. is in the basis of our model



Discussion: Optimization

- Should be black-box
  - We have no gradient information;
- Should be non myopic
  - Local extremums are there and they are plenty;
- Should be parallelizable
  - We have clusters, why not use them?



- GPs work fine. Nice to haves:
  - Show consistency;
  - Maybe "better than grid search" bound like this:
    - $\forall \delta > 0 \exists n(\delta, f) : \forall x' : m^n \geq min_{x'-\delta<x<x'+\delta} f(x) \wedge n = \ldots$
    - Given some reasonable assumptions of function smoothness

Future work

- Largestest scale
  - Millions of images, hundreds of classes;
  - Ideas – scaling by abstraction, peace-wise optimization;
- Transfer learning
  - If we know how the bike looks, learning how the motorbike looks should be easier;
- Integrate <u>any</u> source of supervision
  - Bounding boxes, some pixel labels, unlabelled data – all good!

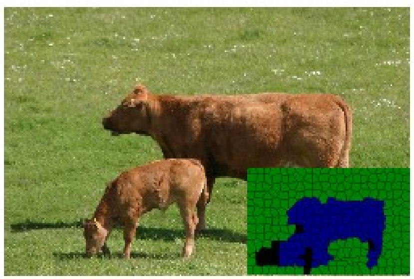
Results

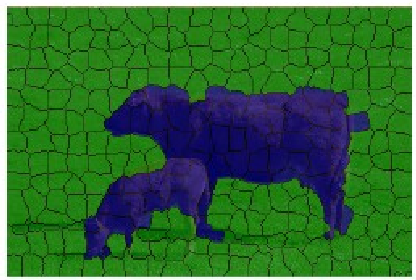
Image / ground truth

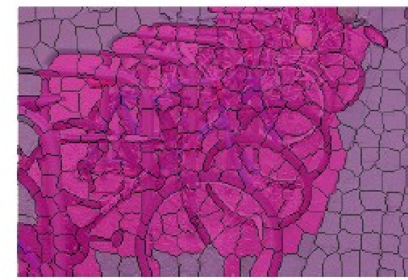
MIM results

Image / ground truth

MIM results







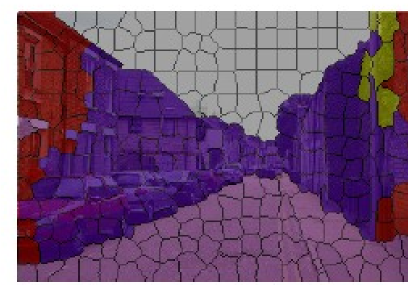


Image / ground truth

MIM results

Image / ground truth

MIM results

ICCV'11 results

LabelMe

Supervision	Average per class accuracy	Method
FS	13	J. Shotton, J. Winn, C. Rother, and A. Criminisi. *“TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation”.* In ECCV 2006.
FS	24	C. Liu, J. Yuen, A. Torralba. “*Nonparametric scene parsing: label transfer via dense scene alignment*”. In CVPR, 2009.
WS	**14**	MIM
FS	20	MIM

MSRC21

Supervision	Average per class accuracy	Method
FS	58	J. Shotton, J. Winn, C. Rother, and A. Criminisi. *“TextonBoost: Joint appearance, shape and context modeling for multi-class object recognition and segmentation”.* In ECCV 2006.
FS	67	J. Shotton, M. Johnson, and R. Cipolla. *“Semantic texton forests for image categorization and segmentation”.* In CVPR, 2008.
FS	75	L’ubor Ladick´y, Chris Russell and Pushmeet Kohli “*Associative Hierarchical CRFs for Object Class Image Segmentation*” In CVPR 2009.
WS	50	J. Verbeek and B. Triggs. “*Region classification with markov field aspect models*”. In CVPR, 2007
WS	**67**	MIM
FS	72	MIM