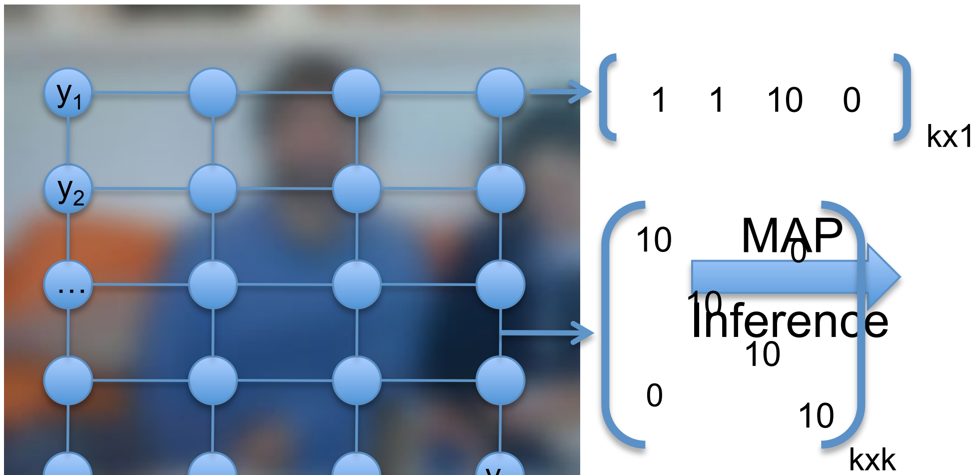
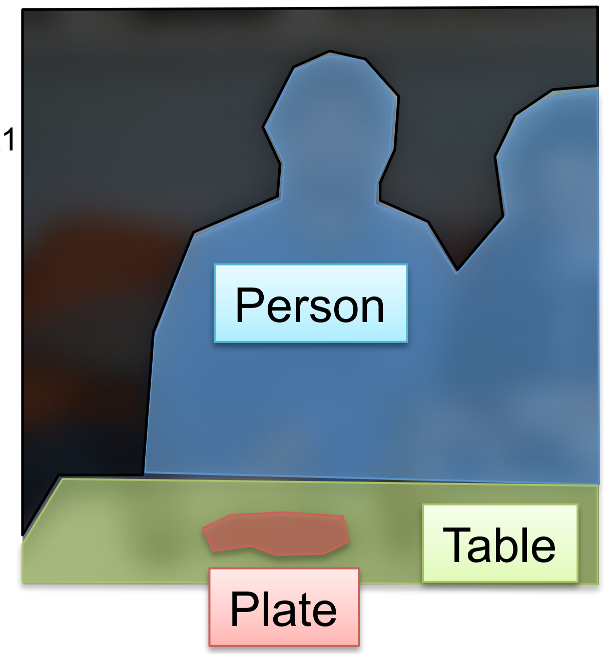
M-Best and Diverse M-Best MAP Inference

Dhruv Batra
Virginia Tech

[CVPR ‘14 Tutorial on Learning and Inference in Discrete Graphical Models]

MAP Inference

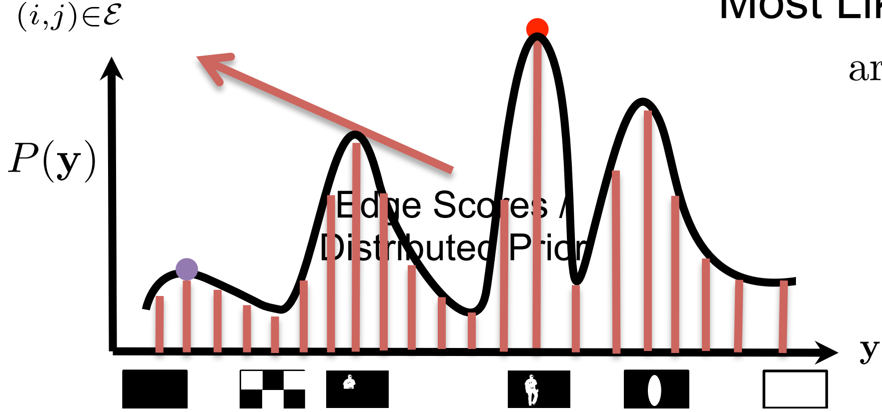




$$S(\mathbf{y}) = \sum_{i \in \mathcal{V}} \theta_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(y_i, y_j)$$

$$P(\mathbf{y}) = \frac{1}{Z} e^{S(\mathbf{y})}$$

Node Scores /
Local Rewards



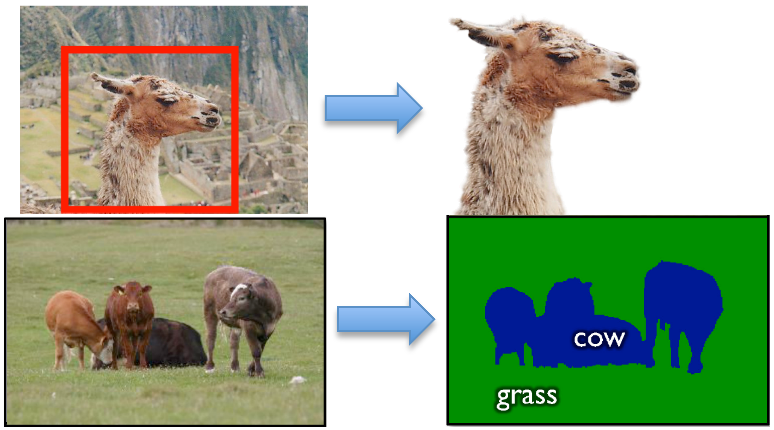
Most Likely Assignment

$$\underset{\mathbf{y}}{\text{argmax}} \; S(\mathbf{y})$$

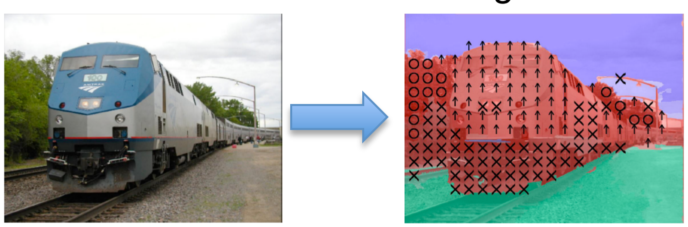
Vision in 2000s

Graphical Models in Vision

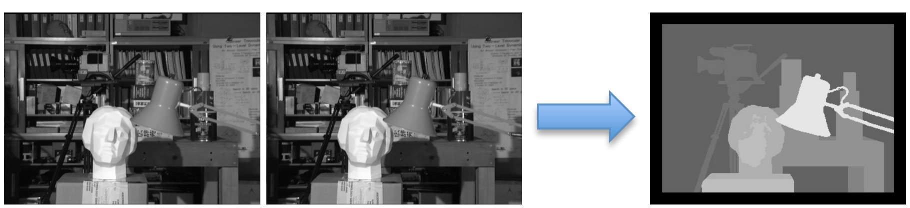
Segmentation



Geometric Labelling

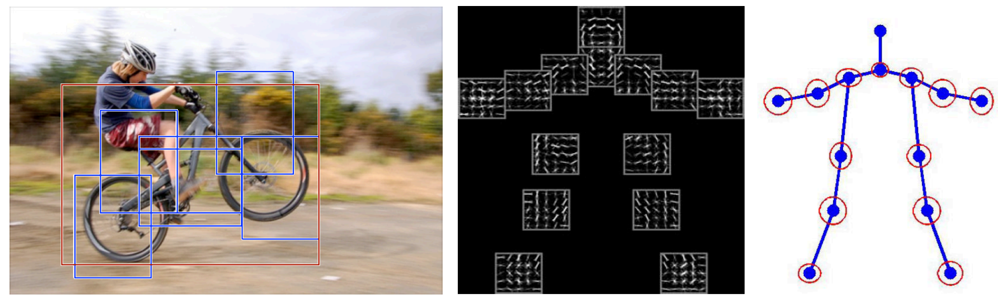


Stereo

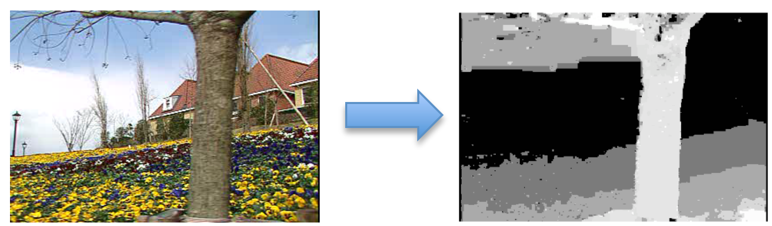


Left image Right image Disparity map

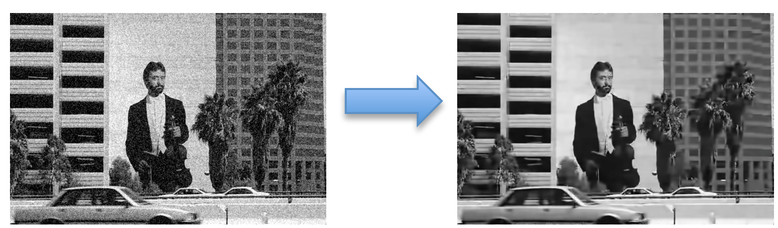
Object Recognition / Pose Estimation

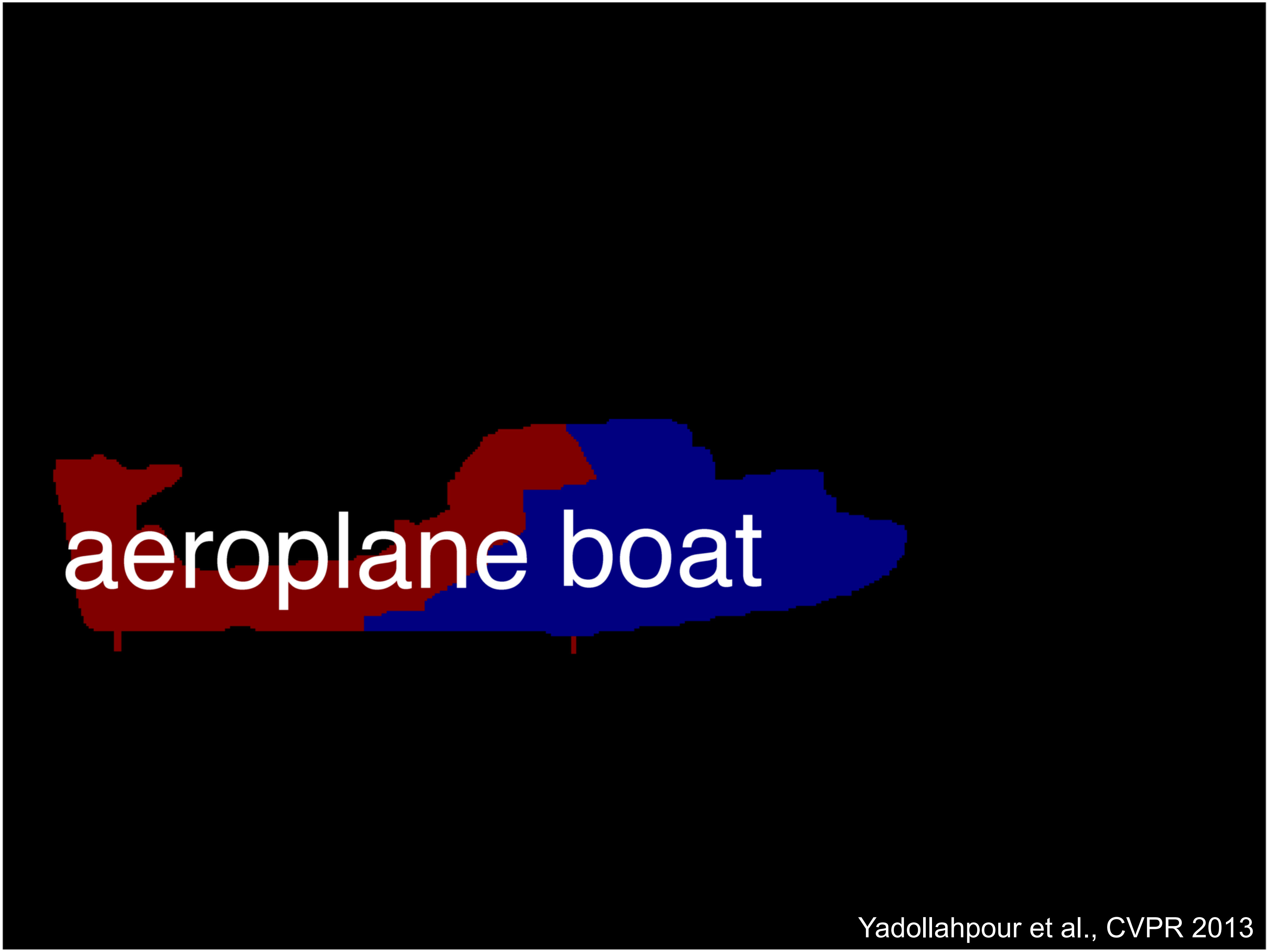


Motion Flow



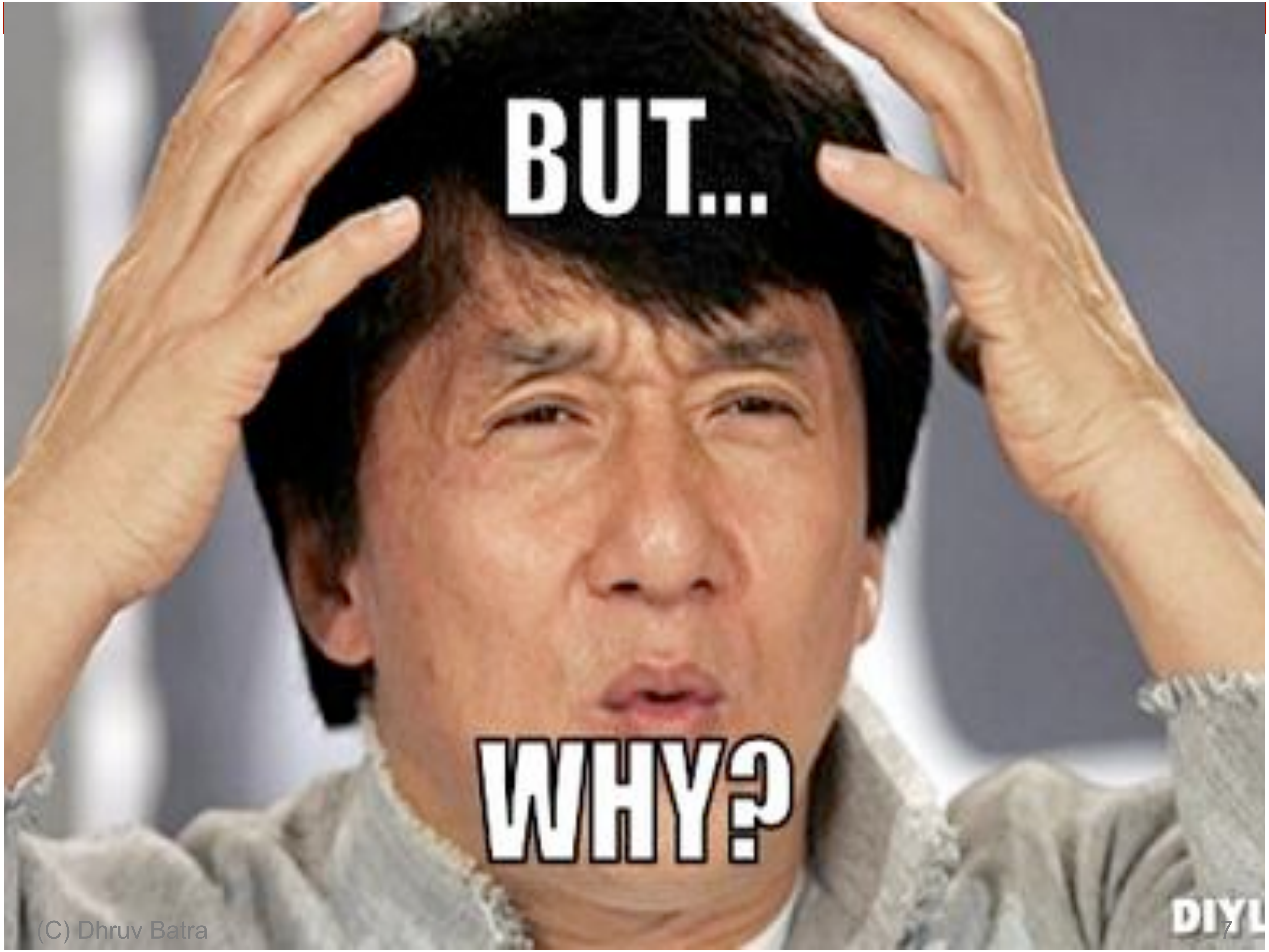
Denoising




aeroplane boat
Yadollahpour et al., CVPR 2013

Slide Credit: Larry Zitnick

Dollar et al., BMVC 2009


BUT...
WHY?

Problems

Model-Class is Wrong!

-- Approximation Error

Human Body ≠ Tree

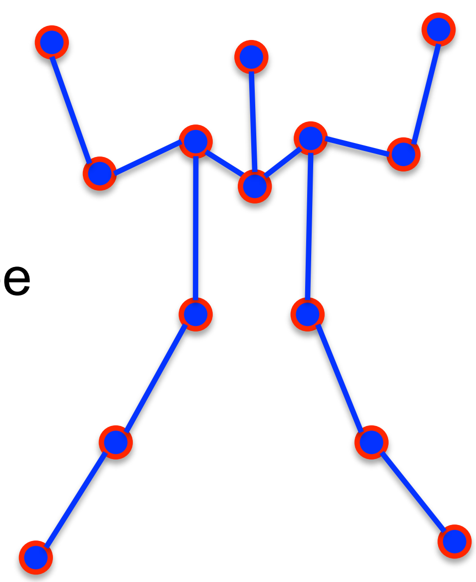
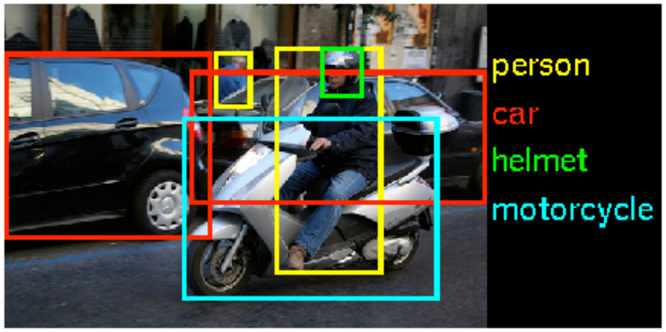


Figure Courtesy: [Yang & Ramanan ICCV '11]

Problems

Not Enough Training Data!

-- Estimation Error



Problems

MAP is NP-Hard

-- Optimization Error

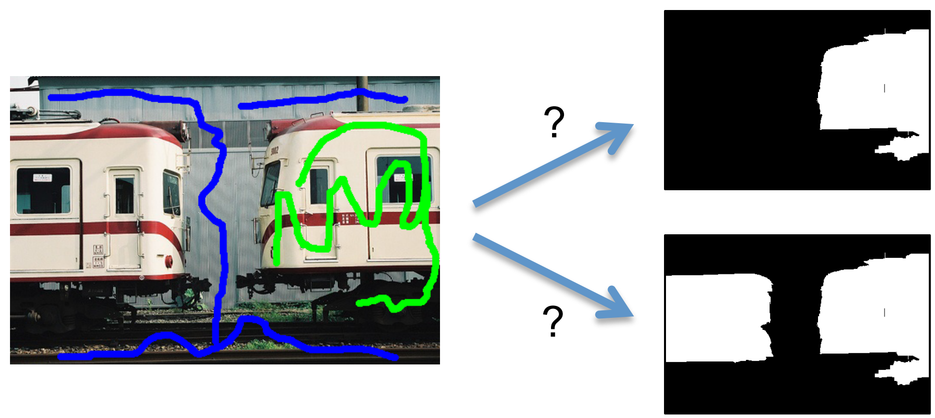
Biggest Problem

Inherent Ambiguity

-- Bayes Error

Rotating clockwise / anti-clockwise?

Old Lady looking left / Young woman looking away?



One instance / Two instances?

Problems

Inherent Ambiguity

-- Bayes Error

Problems

Single Prediction = Uncertainty Mismanagement

Need: Better Representation of Uncertainty

Representation of Uncertainty

MAP = Most Likely Soln.

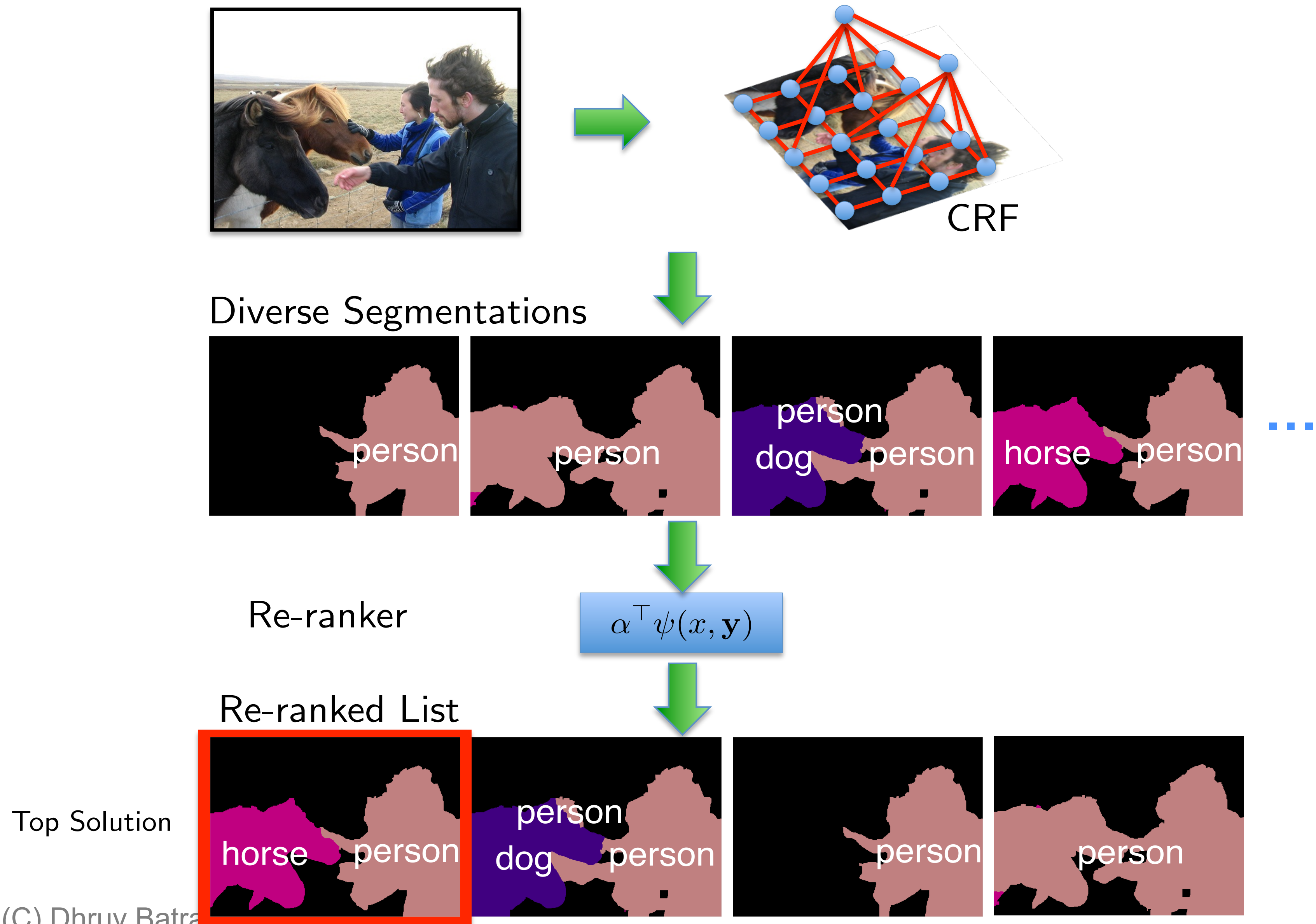
Marginal Beliefs

No Uncertainty

No Structure

This line of work

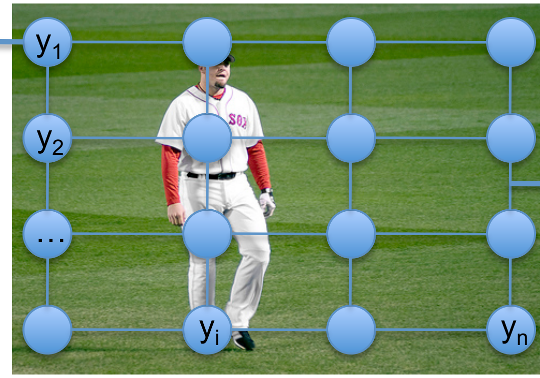
Example Result



MAP in Pairwise MRFs

$G = (\mathcal{V}, \mathcal{E})$

$$\begin{bmatrix} \theta_{ij}(1,1) & & \theta_{ij}(1,k) \\ & \cdots & \\ \theta_{ij}(k,1) & & \theta_{ij}(k,k) \end{bmatrix} \quad \begin{bmatrix} \theta_1(1) \\ \cdots \\ \theta_1(k) \end{bmatrix}_{kx1}$$



$$\begin{bmatrix} \theta_{ij}(1,1) & & \theta_{ij}(1,k) \\ & \cdots & \\ \theta_{ij}(k,1) & & \theta_{ij}(k,k) \end{bmatrix}_{kxk} \quad \begin{bmatrix} \theta_n(1) \\ \cdots \\ \theta_n(k) \end{bmatrix}$$

- Over-Complete Representation

y_1	$\cdots$	y_n	(y_1, y_2)	$\cdots$	(y_{n-1}, y_n)

$$\boldsymbol{\theta} = \Big(\theta_1(1)\ldots\theta_1(k) \quad \theta_n(1)\ldots\theta_n(k) \quad \theta_{12}(1,1)\ldots\theta_{12}(k,k) \quad \theta_{n-1,n}(1,1)\ldots\theta_{n-1,n}(k,k) \Big)$$

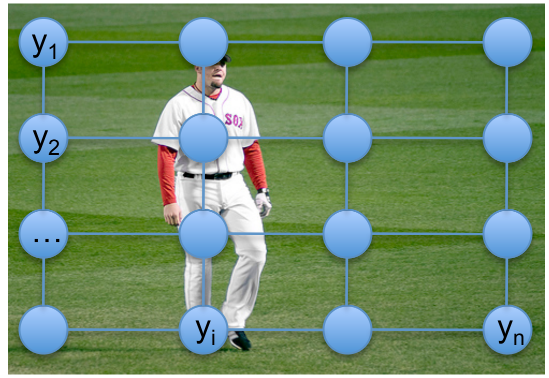
$\mu_1(1)\ldots\mu_1(k)$ $\quad$ $\mu_n(1)\ldots\mu_n(k)$

$y_i \Rightarrow \begin{bmatrix} \cdot \\ \mu_i(s) \\ \cdot \\ \cdot \end{bmatrix}_{kx1}$

$\begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}$ $y_i = 1$ $\quad$ $\begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ $y_i = 2$ $\quad$ ✗ $\begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}$ $\begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}$

MAP in Pairwise MRFs

$G = (\mathcal{V}, \mathcal{E})$

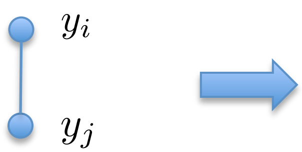


- Over-Complete Representation

$$\begin{array}{ccccccc} & y_1 & \cdots & y_n & (y_1, y_2) & \cdots & (y_{n-1}, y_n) \end{array}$$

$$\boldsymbol{\theta} = \big[\theta_1(1)\ldots\theta_1(k) \quad \theta_n(1)\ldots\theta_n(k) \quad \theta_{12}(1,1)\ldots\theta_{12}(k,k) \quad \theta_{n-1,n}(1,1)\ldots\theta_{n-1,n}(k,k)\big]$$

$$\boldsymbol{\mu}_{\mathbf{y}} = \big[\mu_1(1)\ldots\mu_1(k) \quad \mu_n(1)\ldots\mu_n(k) \quad \mu_{12}(1,1)\ldots\mu_{12}(k,k) \quad \mu_{n-1,n}(1,1)\ldots\mu_{n-1,n}(k,k)\big]$$



$\mu_{ij}(s,t)$: $\begin{bmatrix} \vdots \\ \vdots \end{bmatrix}_{k^2\text{x}1}$

$$S(\mathbf{y}) = \boldsymbol{\theta} \cdot \boldsymbol{\mu}_{\mathbf{y}}$$

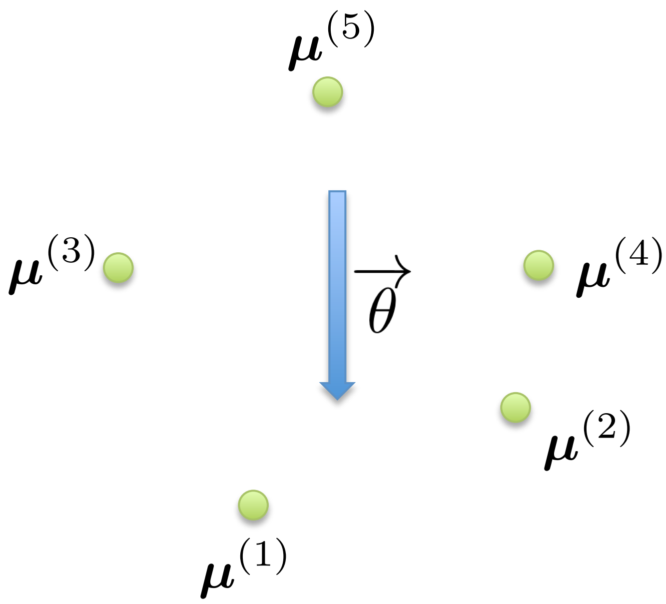
$\begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$ $y_i = 1$, $y_j = 1$

$\begin{bmatrix} 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}^T$ $y_i = 1$, $y_j = 2$

MAP in Pairwise MRFs

- MAP Integer Program

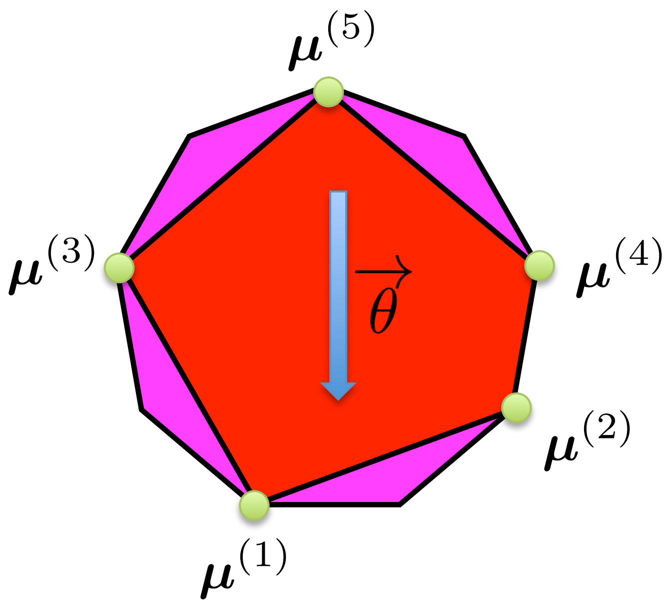
$$\max_{\boldsymbol{\mu}} \quad \boldsymbol{\theta}^T \boldsymbol{\mu}$$
$$s.t. \quad A\boldsymbol{\mu} = b$$
$$\boldsymbol{\mu}(\cdot) \in \{0, 1\}$$



MAP in Pairwise MRFs

- MAP Linear Program

$$\max_{\boldsymbol{\mu}} \quad \boldsymbol{\theta}^T \boldsymbol{\mu}$$
$$s.t. \quad A\boldsymbol{\mu} = b$$
$$\boldsymbol{\mu}(\cdot) \in [0, 1]$$

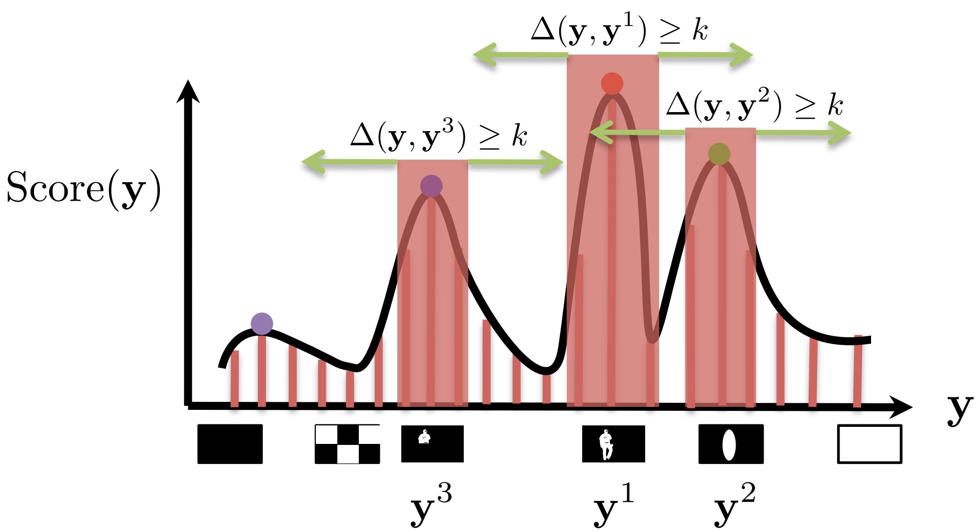


- Properties
  - If LP-opt is integral, MAP is found
  - LP always integral for trees
  - Efficient message-passing schemes for solving LP

Diverse M-Best

$$\max_{\mathbf{y}} \ \text{Score}(\mathbf{y})$$

$$
\begin{aligned}
s.t. \quad & \Delta(\mathbf{y}, \mathbf{y}^1) \geq k \\
& \Delta(\mathbf{y}, \mathbf{y}^2) \geq k \\
& \vdots \\
& \Delta(\mathbf{y}, \mathbf{y}^{M-1}) \geq k
\end{aligned}
$$



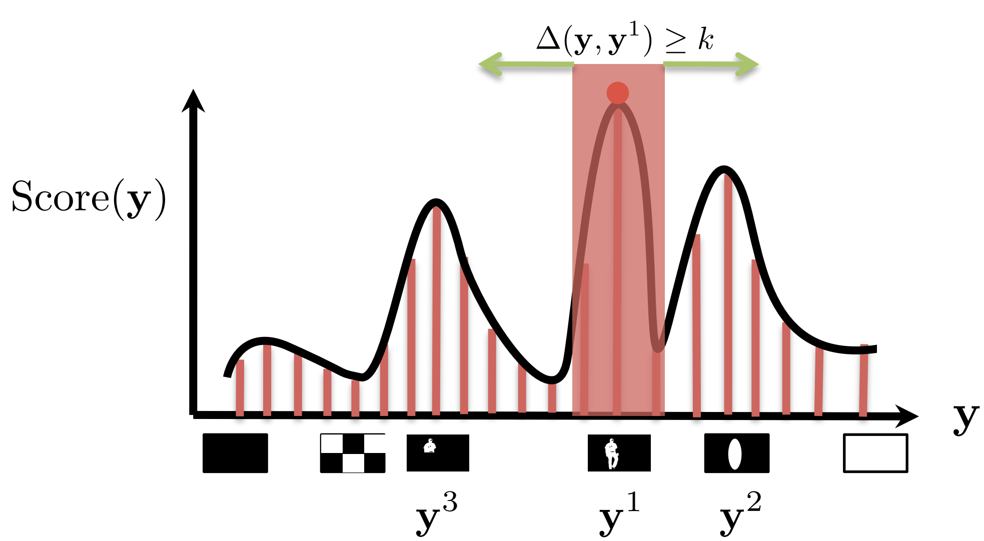
Diverse 2nd-Best

Diversity-Augmented Score

$$\max_{\mathbf{y}} \ \underbrace{\text{Score}(\mathbf{y}) \ +\lambda \cdot \left(\Delta(\mathbf{y}, \mathbf{y}^1) - k\right)}$$

Dualize

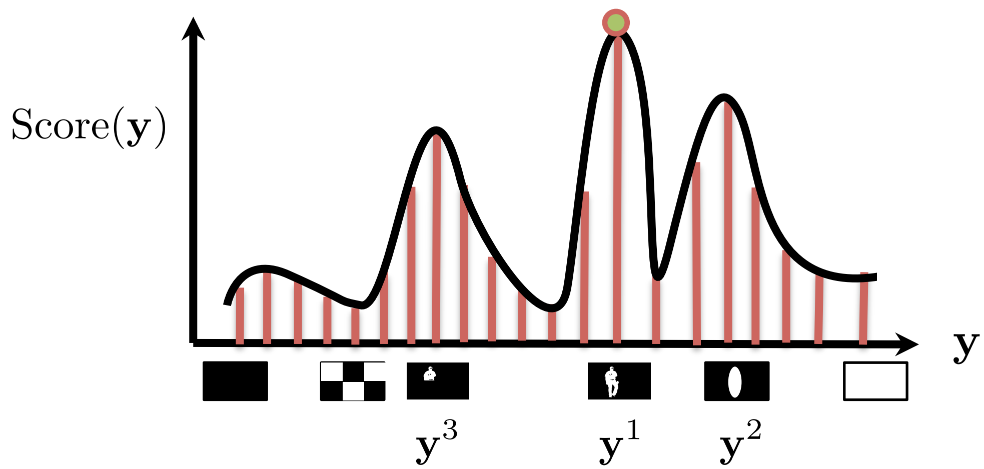
$$S(\mathbf{y}) \quad + \quad Div(\mathbf{y}, \mathbf{y}^{(1)})$$



Diverse 2nd-Best

$$\max_{\mathbf{y}} \; \underbrace{\text{Score}(\mathbf{y}) \; + \lambda \cdot \left(\Delta(\mathbf{y}, \mathbf{y}^1) - k \right)}_{\text{Diversity-Augmented Score}}$$

$$S(\mathbf{y}) \quad + \quad Div(\mathbf{y}, \mathbf{y}^{(1)})$$



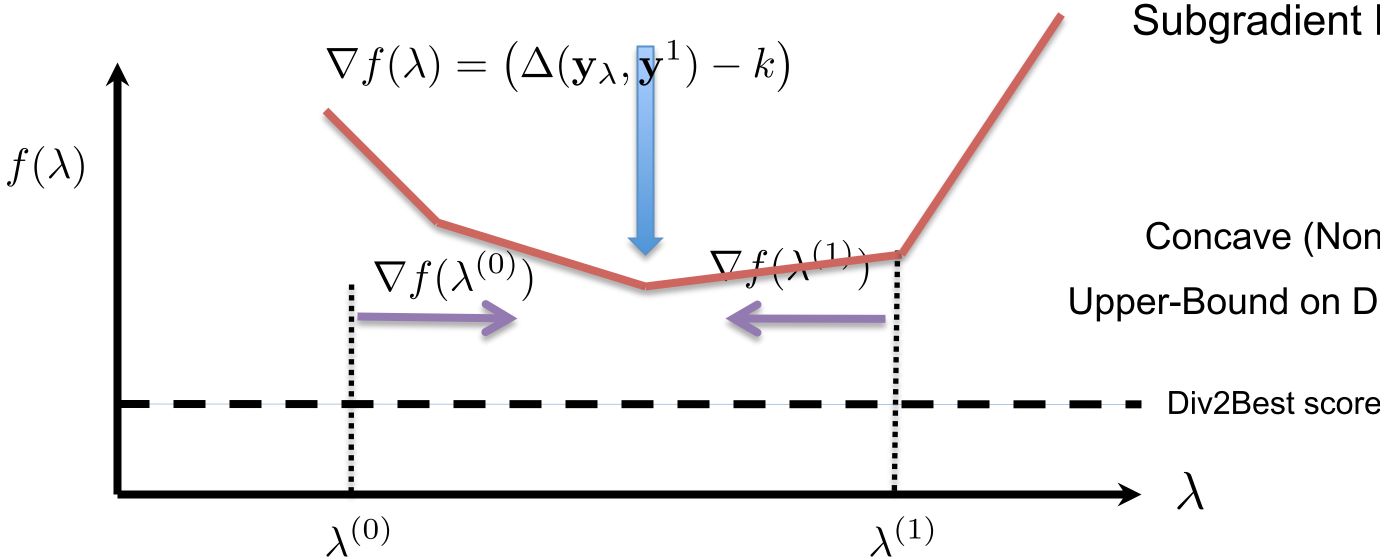
Diverse 2nd-Best

Diversity-Augmented Score

$$f(\lambda) = \max_{\mathbf{y}} \ \text{Score}(\mathbf{y}) \ + \lambda \cdot \left(\Delta(\mathbf{y}, \mathbf{y}^1) - k \right)$$

- Lagrangian Relaxation

Dual $\min_{\lambda \geq 0} \ f(\lambda)$



Subgradient Descent

Concave (Non-smooth)

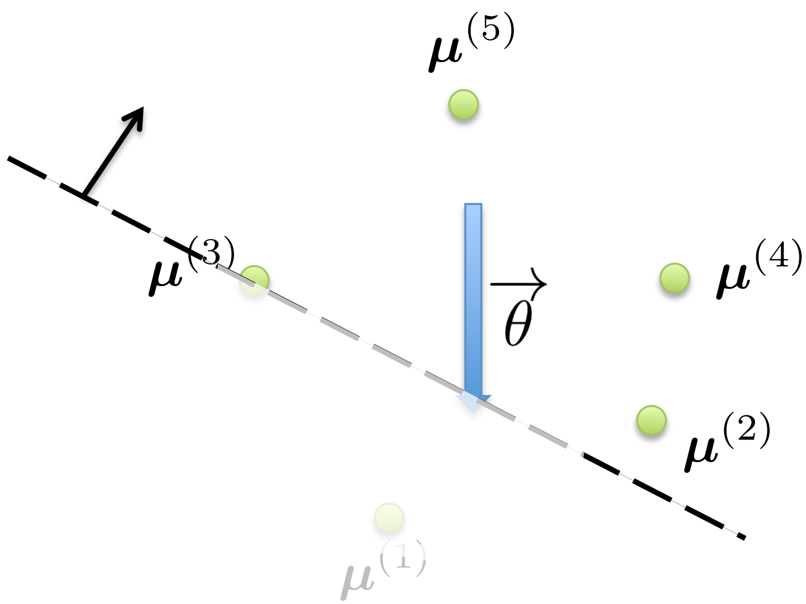
Upper-Bound on Div2Best Score

Diverse 2nd-Best

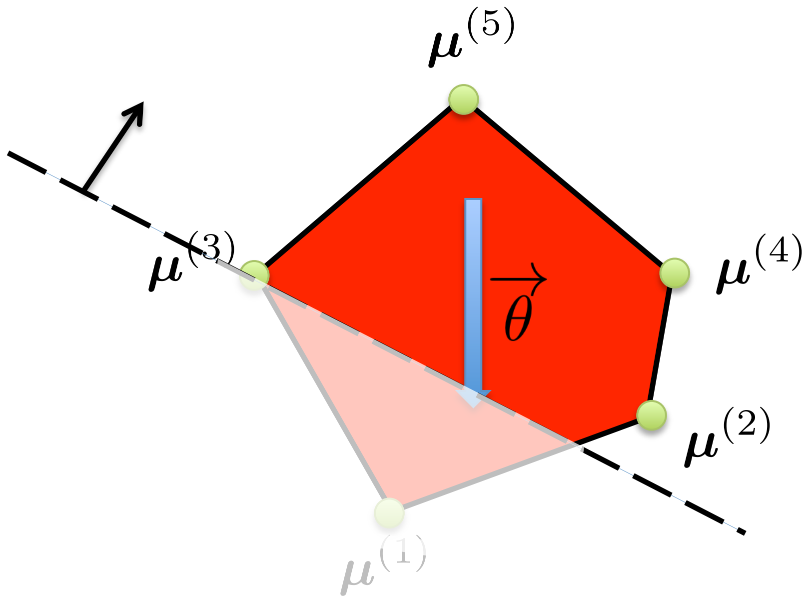
- Many ways to solve:

  1. Subgradient Ascent. Optimal. Slow.
  2. Binary Search. Optimal for M=2. Faster.
  3. Grid-search on lambda. Sub-optimal. Fastest.

Effect of Lagrangian Relaxation



Effect of Lagrangian Relaxation



Theorem Statement

- Theorem [Batra et al ’12]: Lagrangian Dual corresponds to solving the Relaxed Primal:
    - Based on result from [Geoffrion ‘74]

Dual

$$\min_{\lambda \geq 0} \quad \text{LagrangianDual}(\lambda)$$

Relaxed Primal

$$\max_{\boldsymbol{\mu}} \quad \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{ij} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$$

$$s.t. \quad \boldsymbol{\mu} \in Co\big\{\boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0,1\} \mid \boldsymbol{\mu} \in \mathcal{C}\big\}$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(\mathbf{1})}) \geq k$$

Diversity

- [Special Case] 0-1 Diversity $\Longrightarrow$ M-Best MAP
  - [Yanover NIPS03; Fromer NIPS09; Flerova Soft11]
- [Special Case] Max Diversity $\Longrightarrow$ [Park & Ramanan ICCV11]
- Hamming Diversity
- Cardinality Diversity
- Any Diversity $\max_{\mathbf{y}} \; S(\mathbf{y}) + \lambda \cdot \Delta(\mathbf{y}, \mathbf{y}^1)$

Hamming Diversity

$$\Delta(\mathbf{y}, \mathbf{y}^1) = \sum_{i \in \mathcal{V}} [\![y_i \neq y_i^1]\!]$$

- Diversity Augmented Inference:

$$\max_{\mathbf{y}} \quad S(\mathbf{y}) + \lambda \cdot \Delta(\mathbf{y}, \mathbf{y}^1)$$

$$= \sum_{i \in \mathcal{V}} \theta_i(y_i) + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(y_i, y_j) + \lambda\, \Delta(\mathbf{y}, \mathbf{y}^1)$$

$$= \sum_{i \in \mathcal{V}} \underbrace{\left(\theta_i(y_i) + \lambda [\![y_i \neq y_i^1]\!]\right)}_{\tilde{\boldsymbol{\theta}}_i} + \sum_{(i,j) \in \mathcal{E}} \theta_{ij}(y_i, y_j)$$

Hamming Diversity

$$\Delta(\mathbf{y}, \mathbf{y}^1) = \sum_{i \in \mathcal{V}} [[y_i \neq y_i^1]]$$

- Diversity Augmented Inference:

```
for i = 1,2,...,n
    θ_i[y_i^1] -= λ
endfor
y^2 = Find_MAP(θ_i, θ_ij)
```

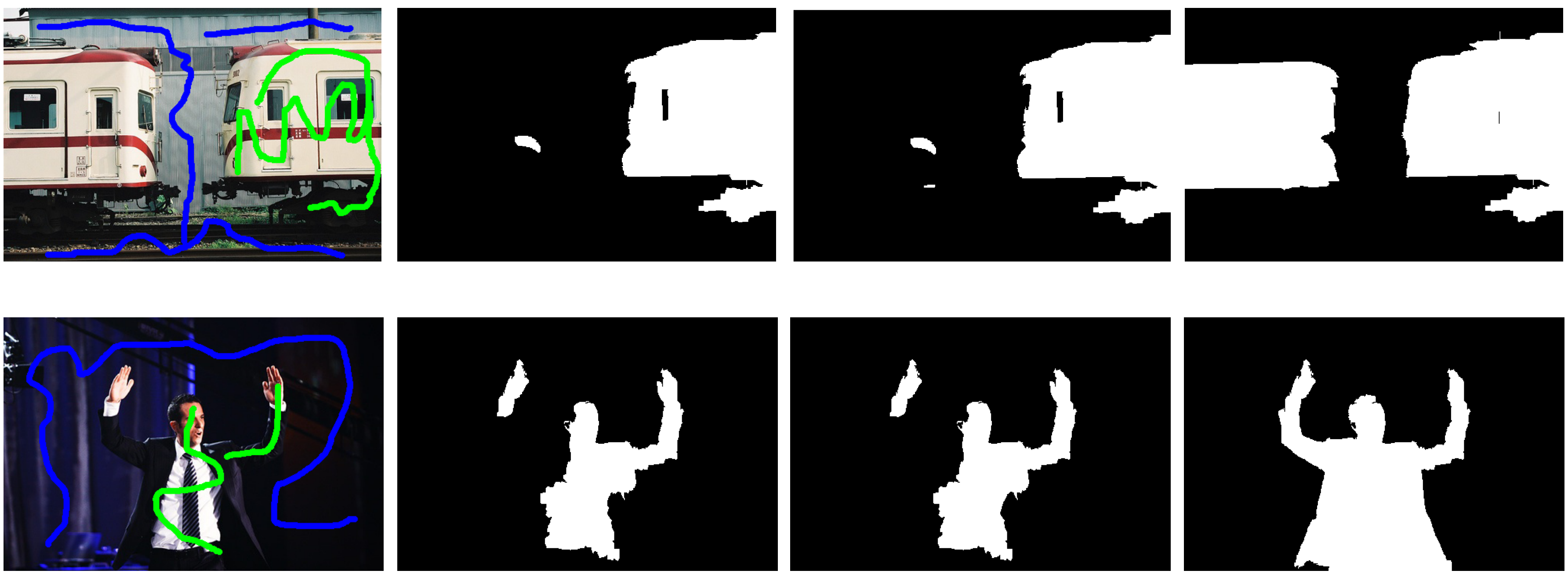
Unchanged.
Can still use graph-cuts!

Simply edit node-terms. Reuse MAP machinery!

Interactive Segmentation

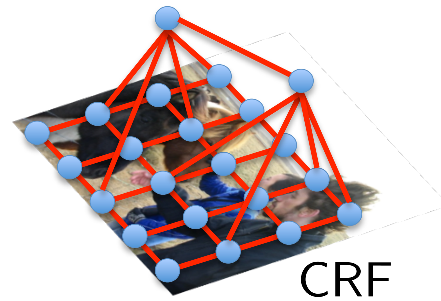
- Setup
  - Model: Color/Texture + Potts Grid CRF
  - Inference: Graph-cuts
  - Dataset: 50 train/val/test images from PASCAL

Image + Scribbles | MAP | 2nd Best MAP | Diverse 2nd Best



1-2 Nodes Flipped | 100-500 Nodes Flipped

Example Result



Diverse Segmentations

Now what?

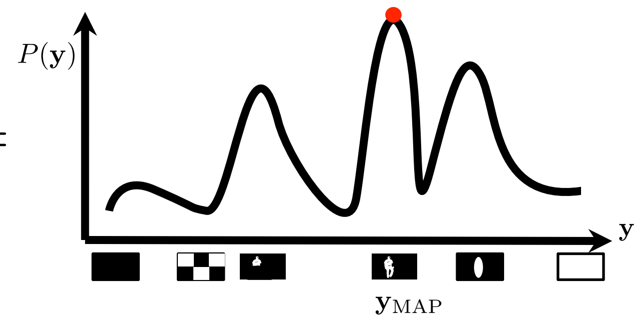
Your Options

- Nothing
  - Additional Information: None
  - User in the loop
- (Approximate) Min Bayes Risk
  - Additional Information: Loss function
  - Approximate P(y|x); Optimize Bayes Risk
- Re-ranking
  - Additional Information: higher-order or temporal features
  - Pick a good solution from the list

Increasing Side Information

Statistics 101

- Loss $\mathcal{L}(\mathbf{y}^{gt}, \hat{\mathbf{y}})$
  - Hamming, Pascal, …

- "True" Distribution $P(\mathbf{y}^{gt} \mid \mathbf{x}) =$ 

  Fit model (CRF, etc) to mimic

- Expected Loss: $BR(\hat{\mathbf{y}}) = \mathbb{E}_{P(\mathbf{y}|\mathbf{x})}\left[\mathcal{L}(\mathbf{y}^{gt}, \hat{\mathbf{y}})\right]$

- Min Bayes Risk $$\min_{\mathbf{y} \in \mathcal{Y}} \sum_{\mathbf{y}^{gt} \in \mathcal{Y}} \mathcal{L}(\mathbf{y}^{gt}, \hat{\mathbf{y}})\, P(\mathbf{y}^{gt}|\mathbf{x})$$

DivMBest for MBR

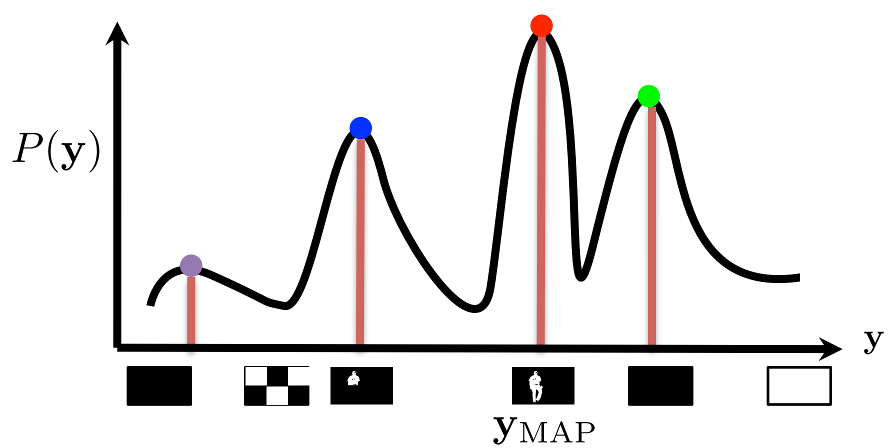
- Min Bayes Risk $$\min_{\mathbf{y} \in \mathcal{Y}} \sum_{\mathbf{y}^{gt} \in \mathcal{Y}} \mathcal{L}(\mathbf{y}^{gt}, \hat{\mathbf{y}}) \, P(\mathbf{y}^{gt} | \mathbf{x})$$

- Two Problems

Intractable

Intractable

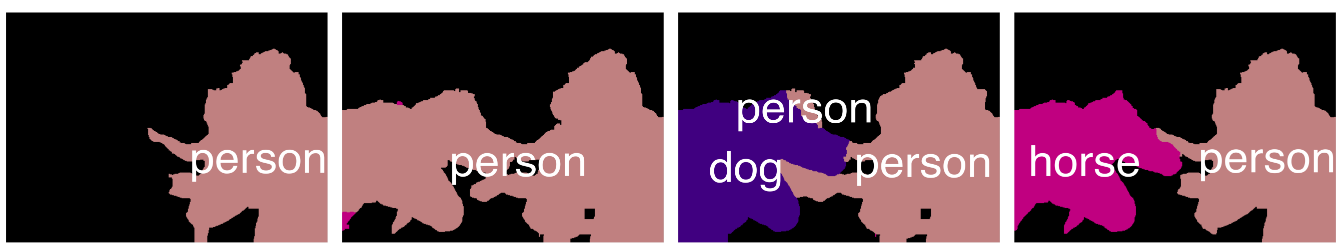
- Approximate MBR: $$\min_{\hat{\mathbf{y}} \in Div\text{MBest}} \sum_{\mathbf{y}^{gt} \in Div\text{MBest}} \mathcal{L}(\mathbf{y}^{gt}, \hat{\mathbf{y}}) \, P(\mathbf{y}^{gt} | \mathbf{x})$$



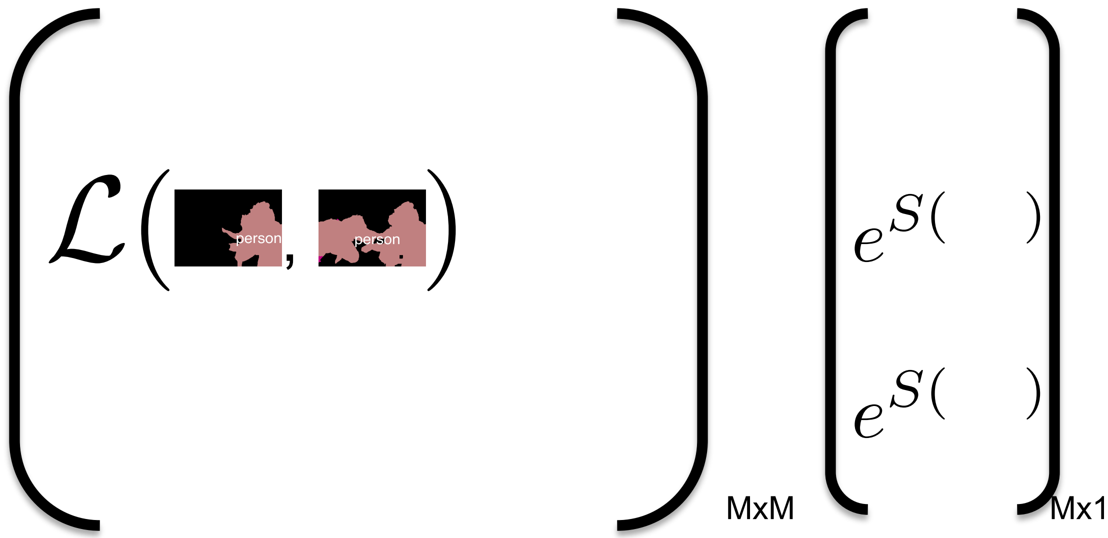
DivMBest for MBR

$$\begin{pmatrix} \mathcal{L}(\qquad) & & \\ & & \\ & & \end{pmatrix}_{\text{MxM}}$$

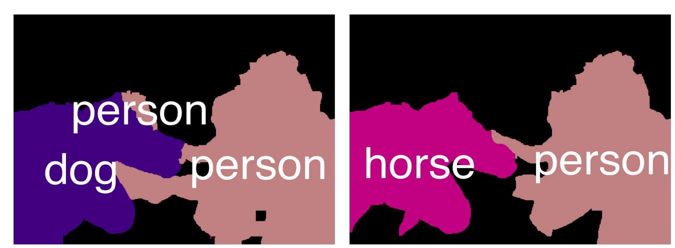
$$\min_{\hat{\mathbf{y}} \in Div\text{MBest}} \sum_{\mathbf{y}^{gt} \in Div\text{MBest}} \mathcal{L}(\mathbf{y}^{gt}, \hat{\mathbf{y}})\, P(\mathbf{y}^{gt}|\mathbf{x})$$



DivMBest for MBR



$$\min_{\hat{\mathbf{y}} \in DivMBest} \sum_{\mathbf{y}^{gt} \in DivMBest} \mathcal{L}(\mathbf{y}^{gt}, \hat{\mathbf{y}})\, P(\mathbf{y}^{gt} | \mathbf{x})$$



DivMBest for MBR

- Diversity Augmented Inference with Hamming:

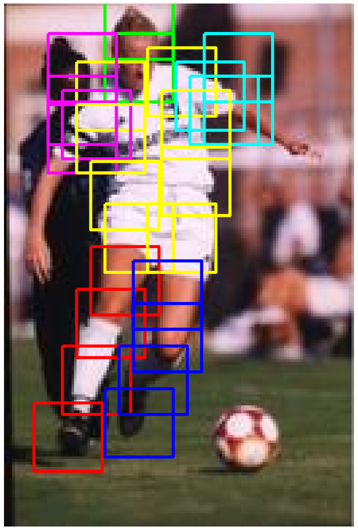
```
for i = 1,2,...,n
    θ_i[y_i^1] -= λ
endfor
y^2 = Find_MAP(θ_i, θ_ij)
```

- Empirical MBR prediction [PBT CVPR14]

$$\min_{\hat{\mathbf{y}} \in Div\text{MBest}} \sum_{\mathbf{y}^{gt} \in Div\text{MBest}} \mathcal{L}(\mathbf{y}^{gt}, \hat{\mathbf{y}})\, P(\mathbf{y}^{gt}|\mathbf{x})$$

Pose Estimation

- Setup
  - Model: Mixture of Parts Tree [Park & Ramanan, ICCV '11]
  - Inference: Dynamic Programming
  - Dataset: PARSE



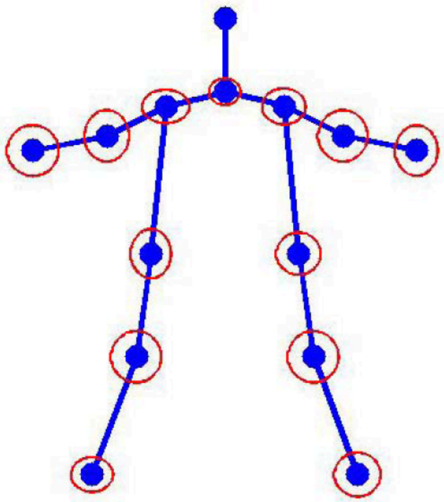
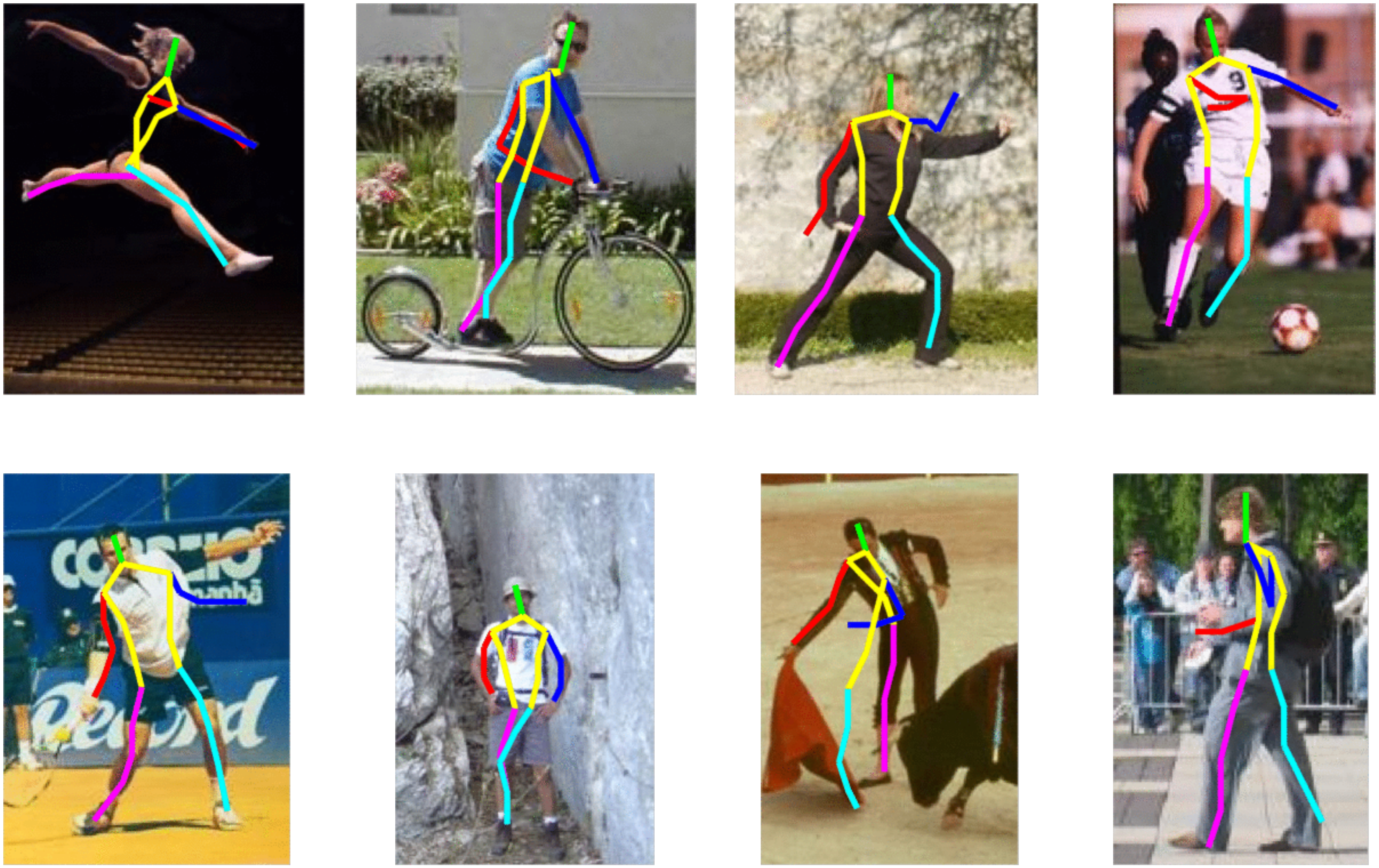


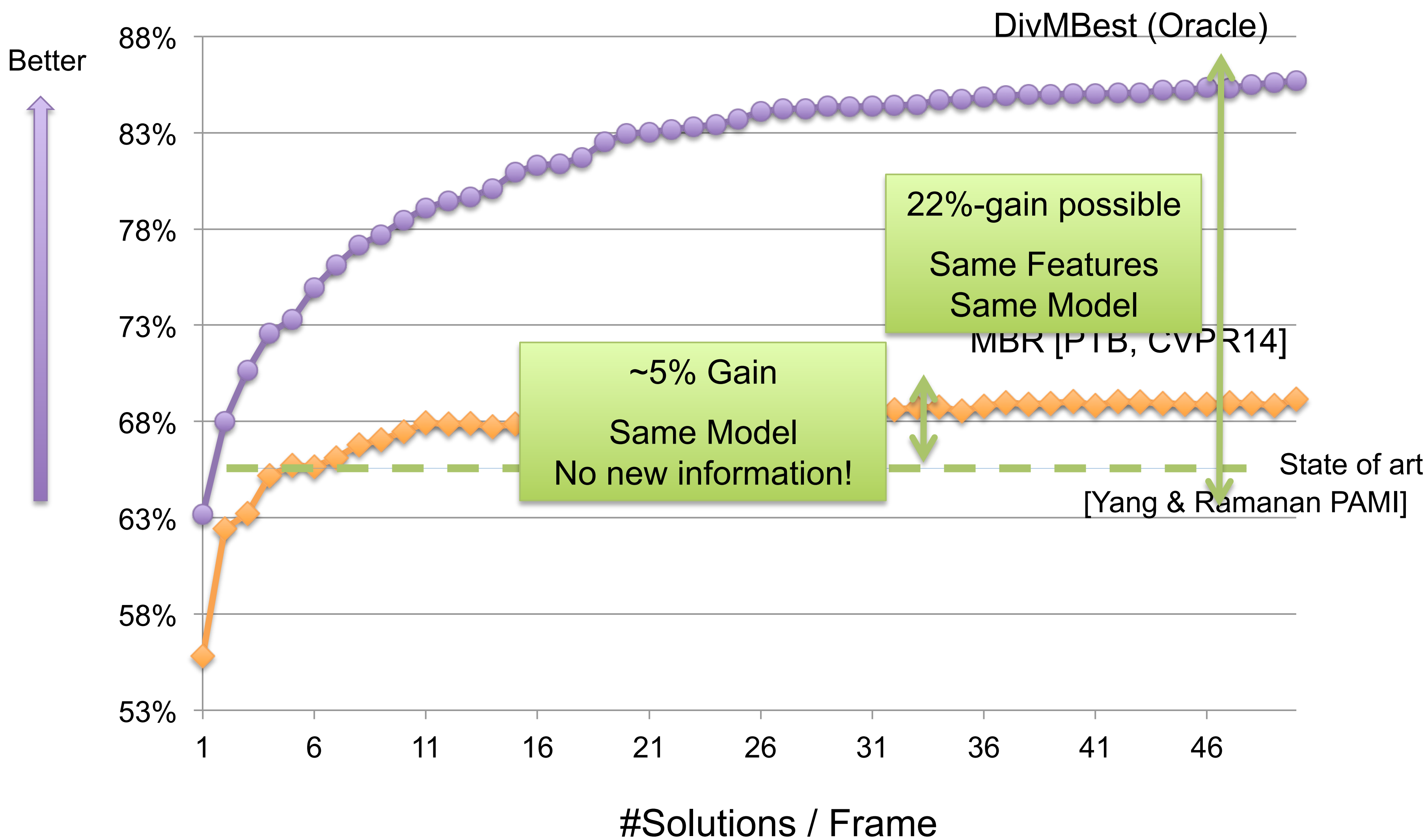
Image Credit: [Yang & Ramanan, ICCV '11]

Pose Estimation



[Premachandran, Tarlow, Batra, CVPR14]

Pose Estimation



[Premachandran, Tarlow, Batra, CVPR14]

Your Options

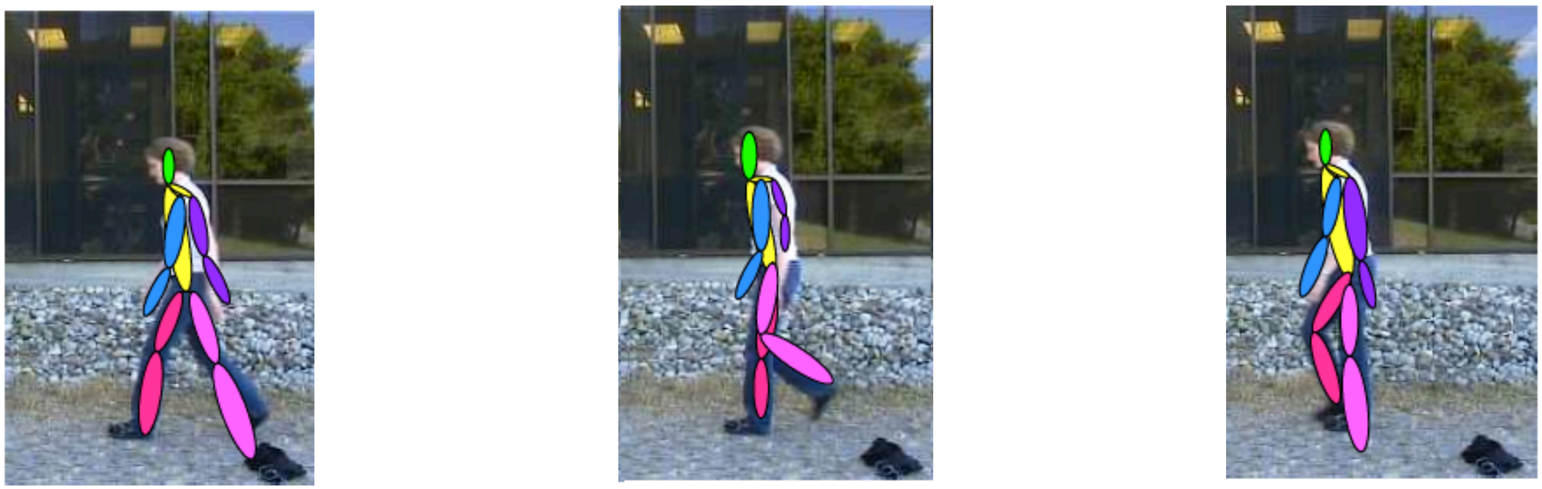
- Nothing
  - Additional Information: None
  - User in the loop

- (Approximate) Min Bayes Risk
  - Additional Information: Loss function
  - Approximate P(y|x); Optimize Bayes Risk

- Re-ranking
  - Additional Information: higher-order or temporal features
  - Pick a good solution from the list

Pose Tracking

- Chain CRF with M states at each frame



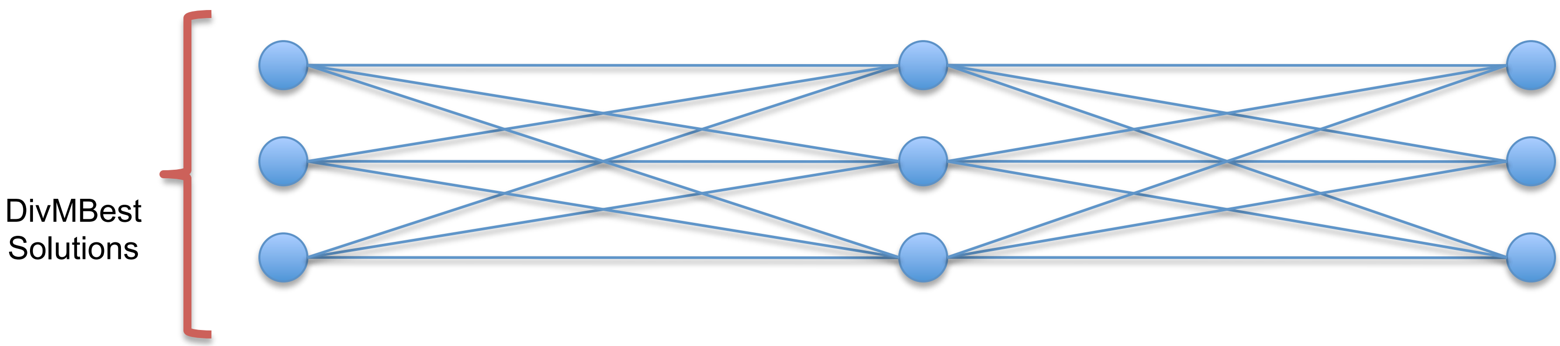
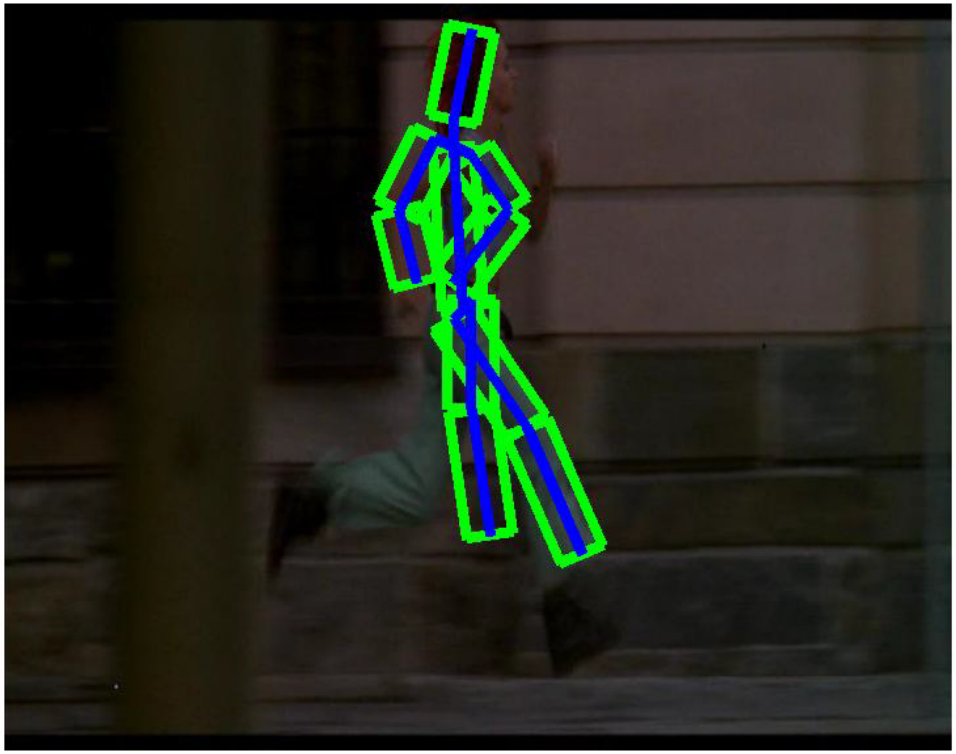


Image Credit: [Yang & Ramanan, ICCV '11]

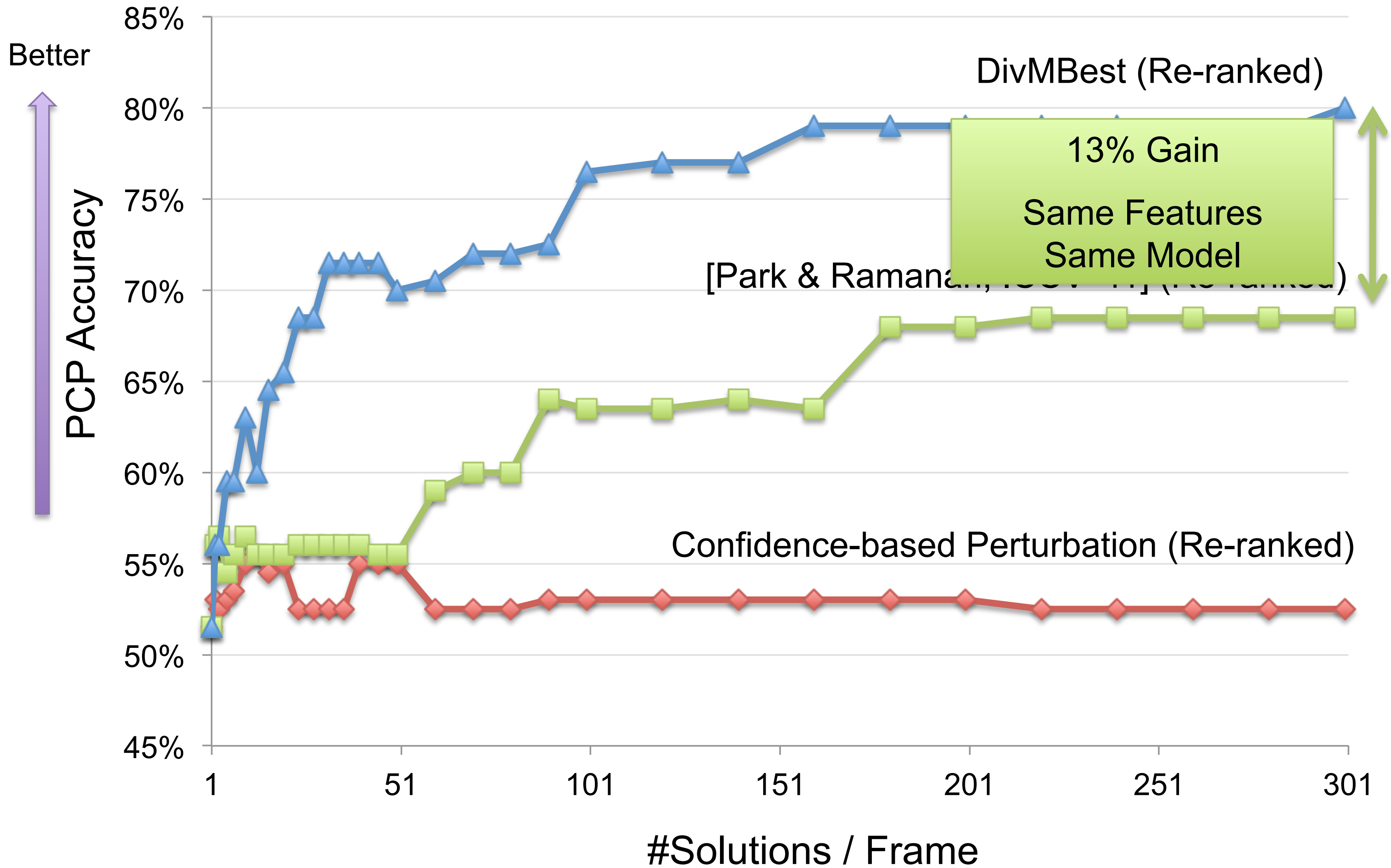
Pose Tracking



MAP

DivMBest + Viterbi

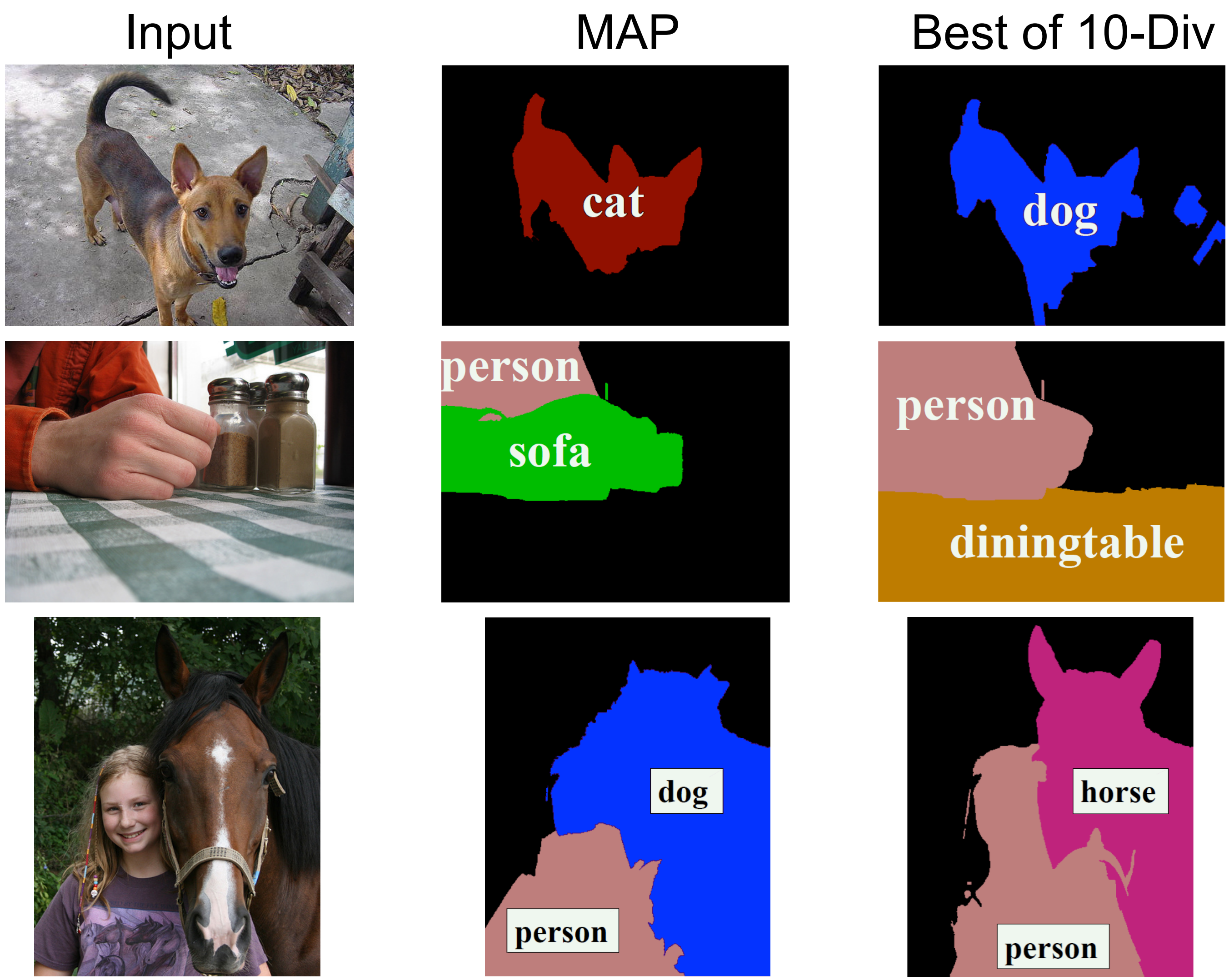
Pose Tracking



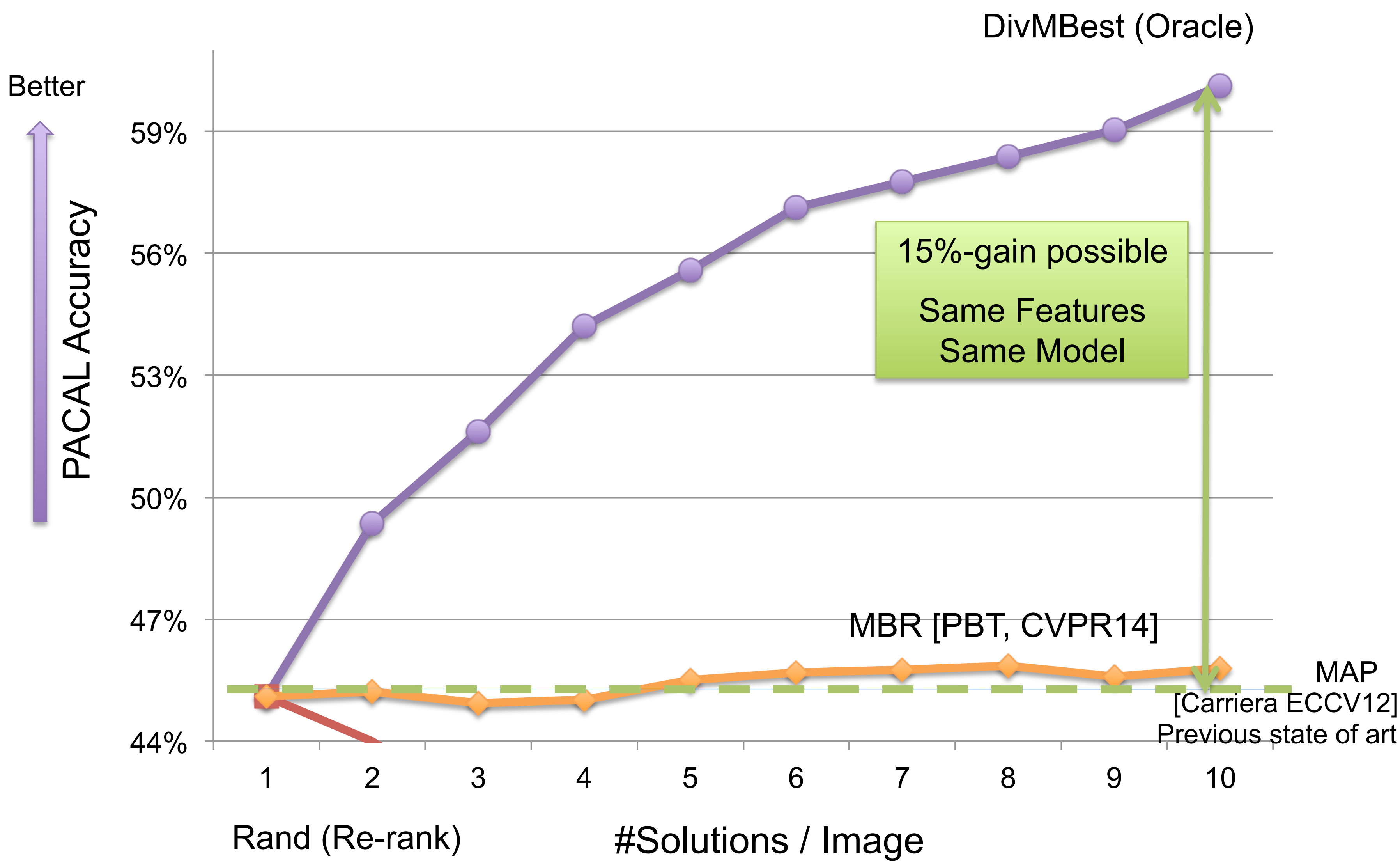
Semantic Segmentation

- Setup
  - Models: Second-Order Pooling Flat CRF [Carreira ECCV12]
  - Inference: Greedy Pasting
  - Dataset: Pascal Segmentation Challenge (VOC 2012)
    - 20 categories + background; ~1500 train/val/test images

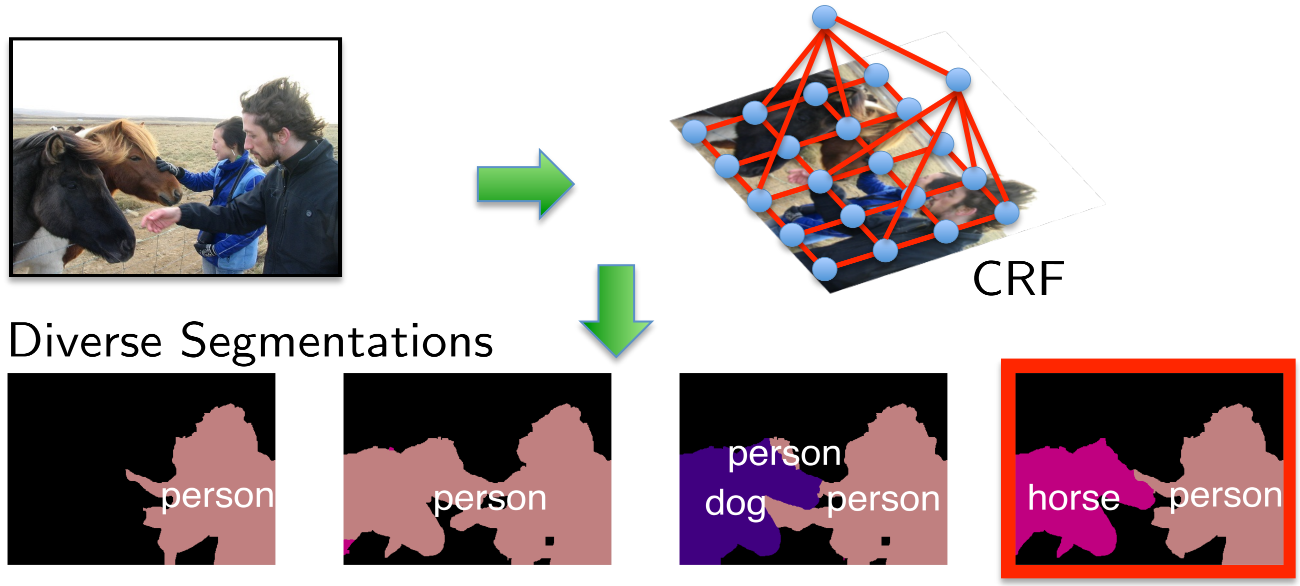
Semantic Segmentation



Semantic Segmentation

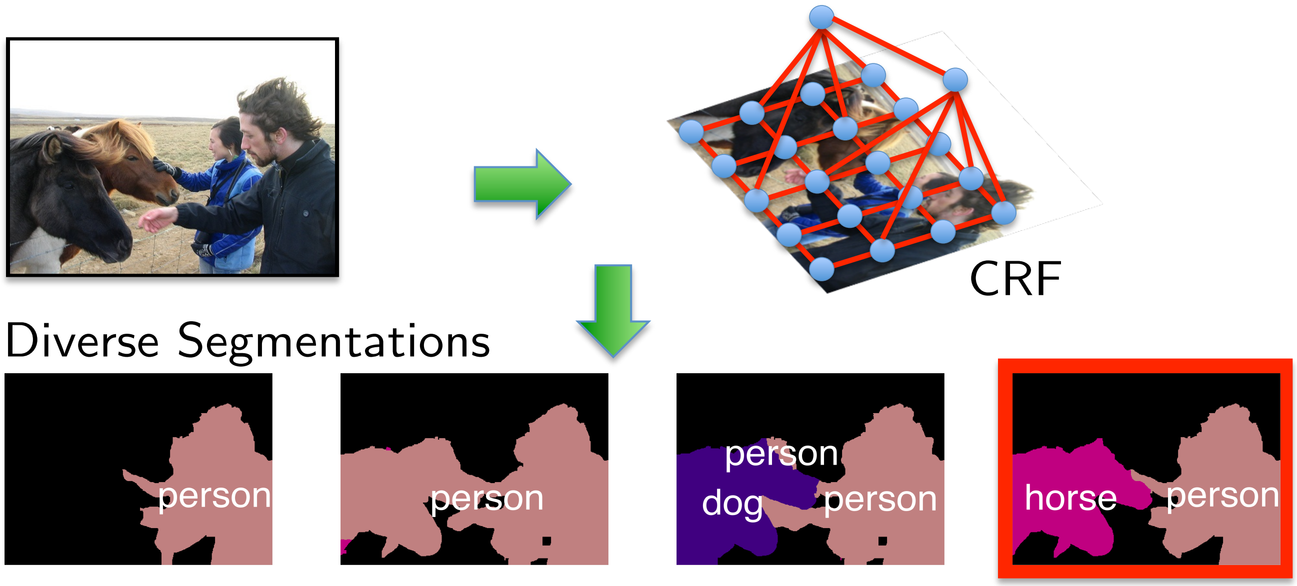


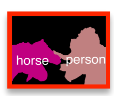
Large-Margin Re-ranking



[Yadollahpour, Batra, Shakhnarovich, CVPR13]

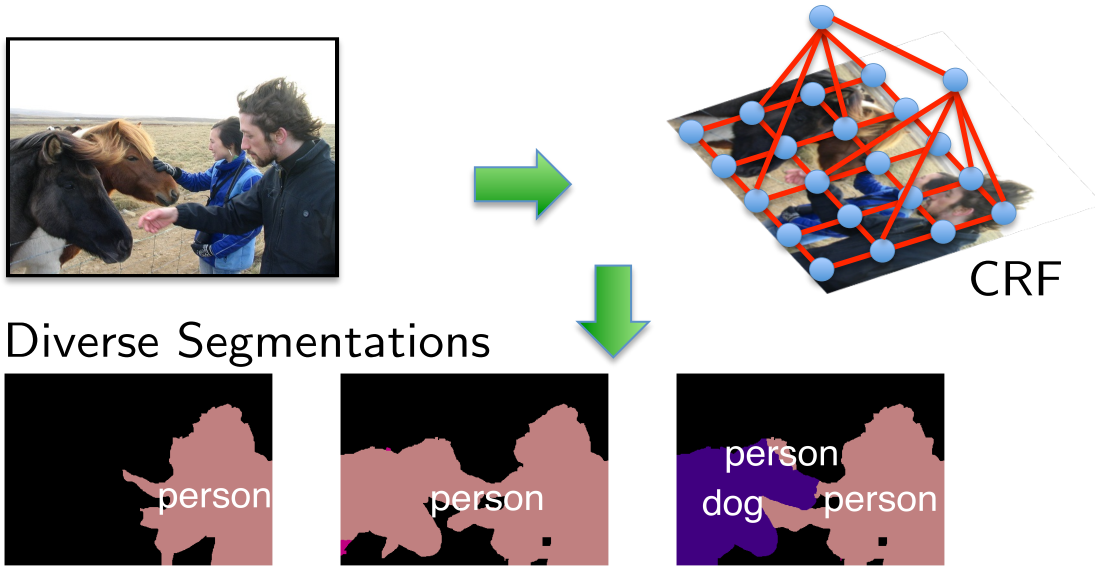
Large-Margin Re-ranking

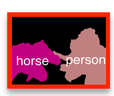
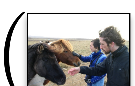


$\psi($,  $)$

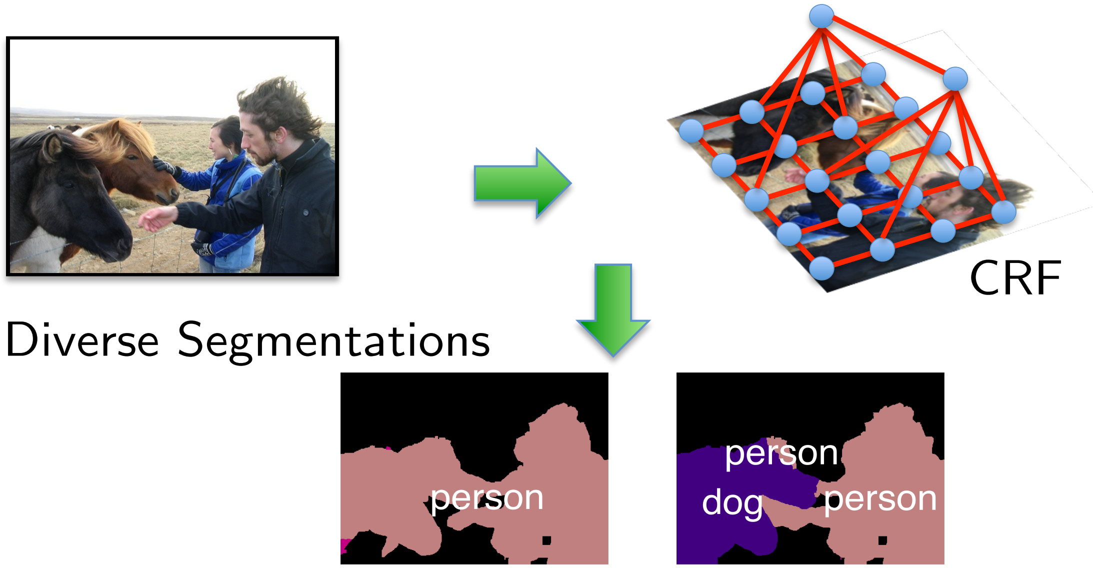
[Yadollahpour, Batra, Shakhnarovich, CVPR13]

Large-Margin Re-ranking

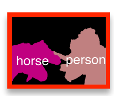


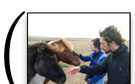
$\boldsymbol{\alpha}^{\intercal}\psi($, $)$ $-\boldsymbol{\alpha}^{\intercal}\psi($, $)$

Large-Margin Re-ranking



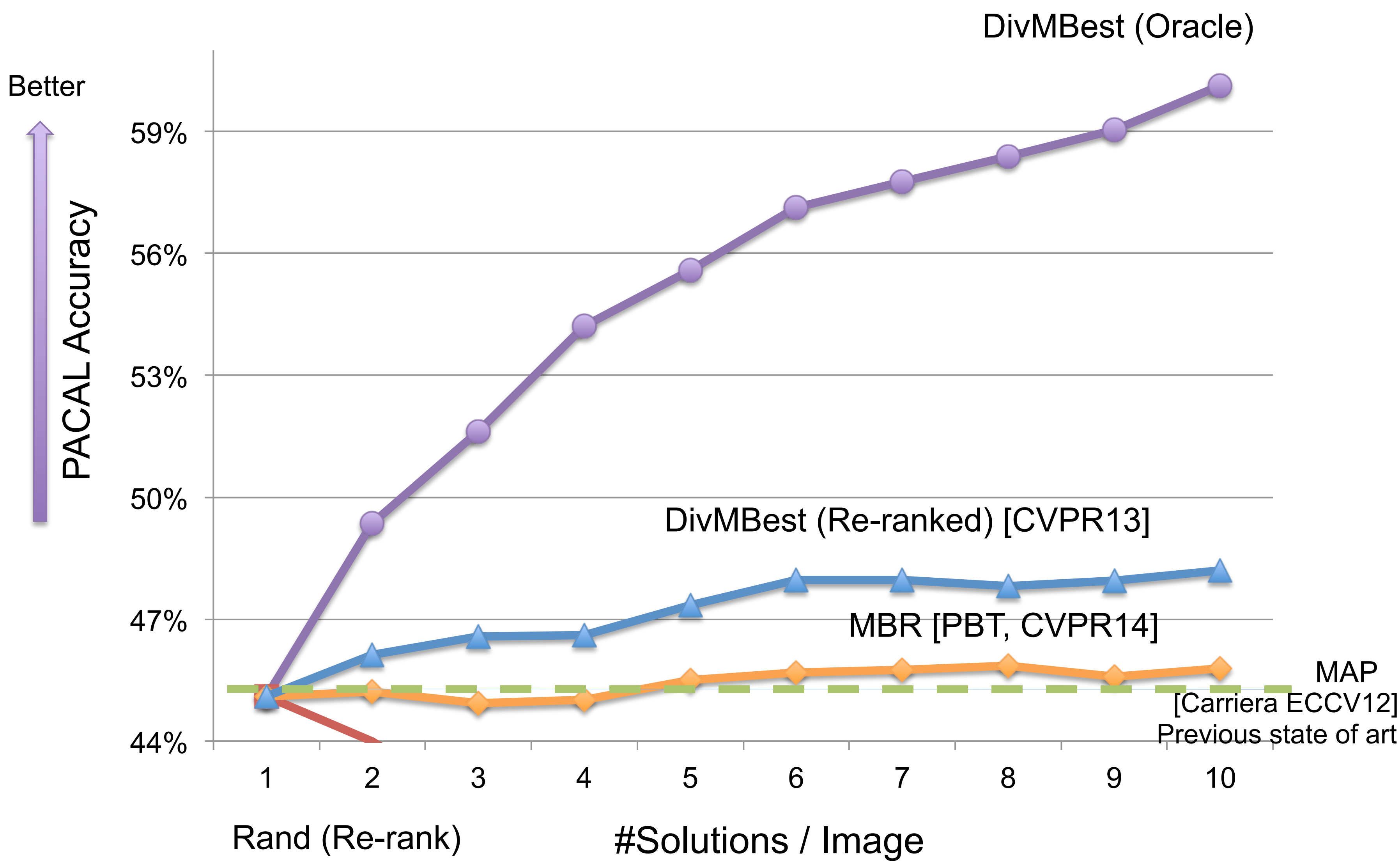
$$\min_{\boldsymbol{\alpha},\xi_i} \quad \|\boldsymbol{\alpha}\|^2 + C\sum_i \xi_i$$

$\boldsymbol{\alpha}^{\mathsf{T}}\psi($  , 

$) - \boldsymbol{\alpha}^{\mathsf{T}}\psi($  , 

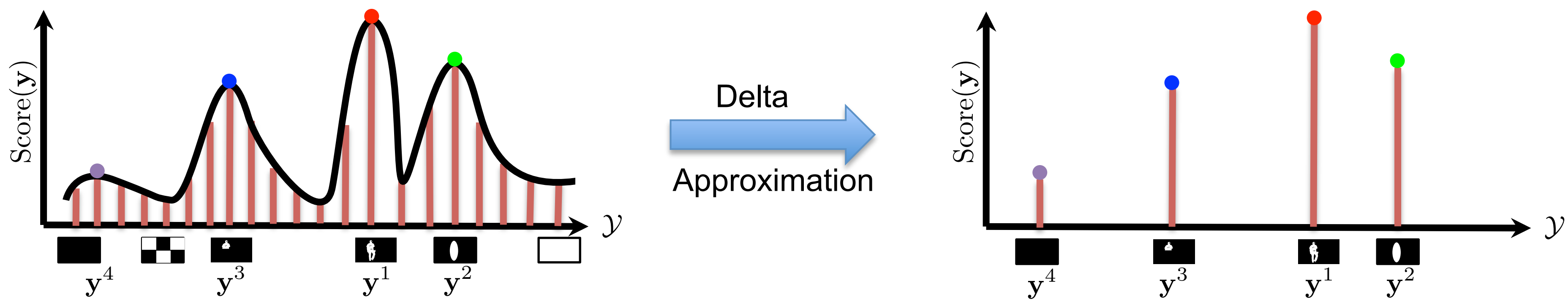
$) \geq 1 - \dfrac{\xi_i}{\text{loss}_i}$

Semantic Segmentation



[Yadollahpour, Batra, Shakhnarovich, CVPR13]

MAP as a Sampler!



- Intractable Quantities
  - Bayes Risk
  - Partition Function
  - Entropy
  - …

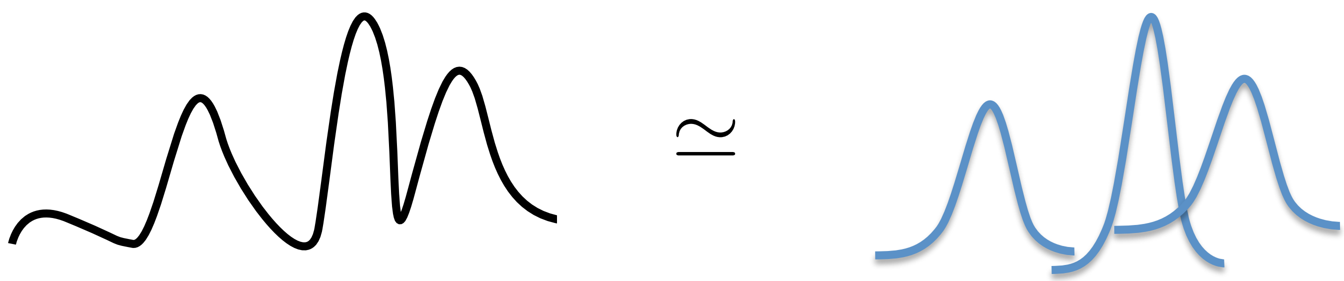
Summary

- All models are wrong
- Some beliefs are useful
- Focus on optimization error is unhelpful
  - The goal is low generalization error
- Diverse Multiple Solutions
  - A way to get useful beliefs out.
- DivMBest +
  MBR / Reranking / Active-Learning / Faster Training
  - Big impact on many applications!

Summary

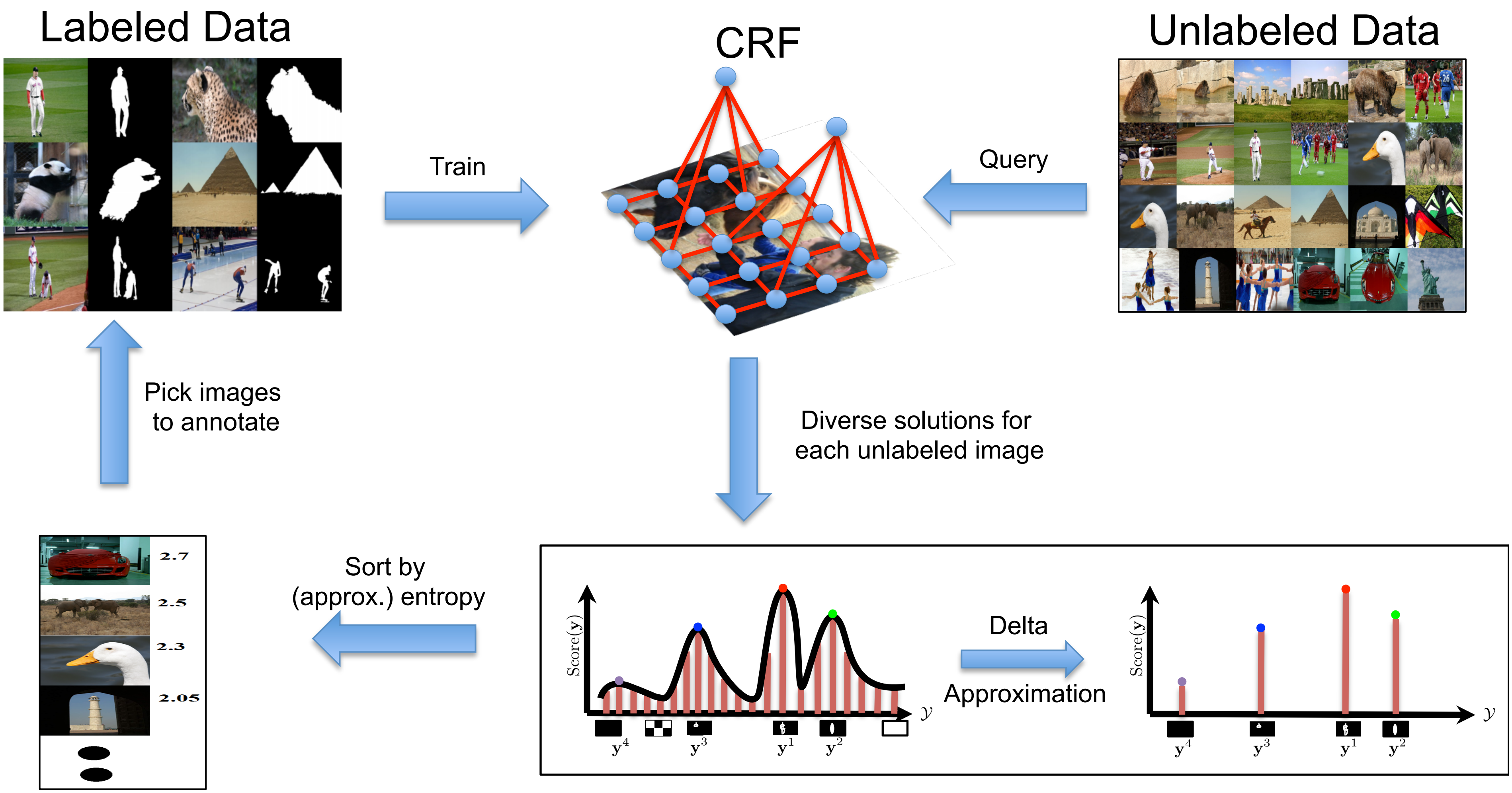
- What does my model *believe*?

Posterior Summary



Thanks!

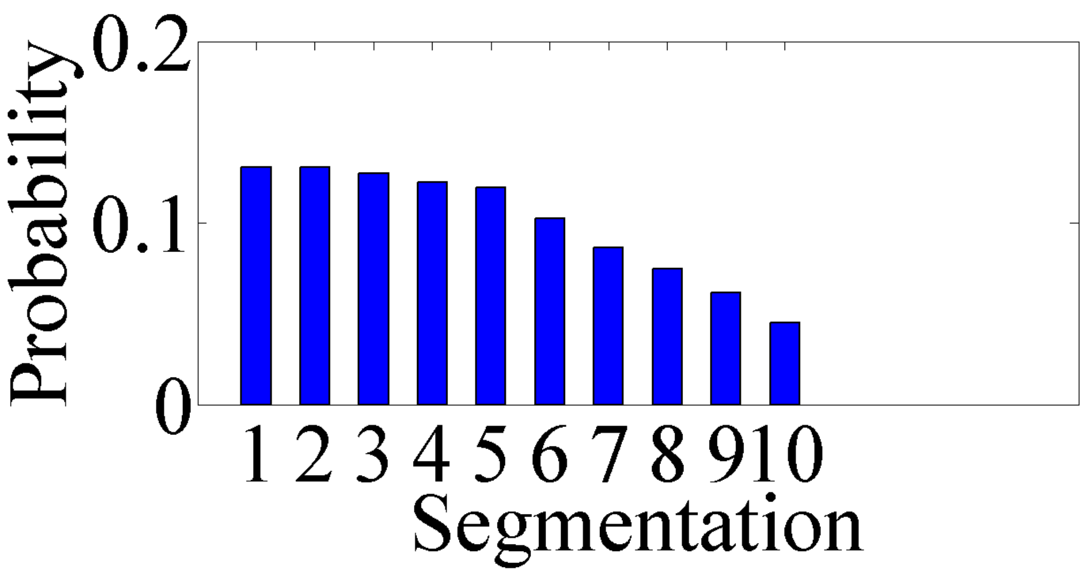
Active Learning



Active Learning

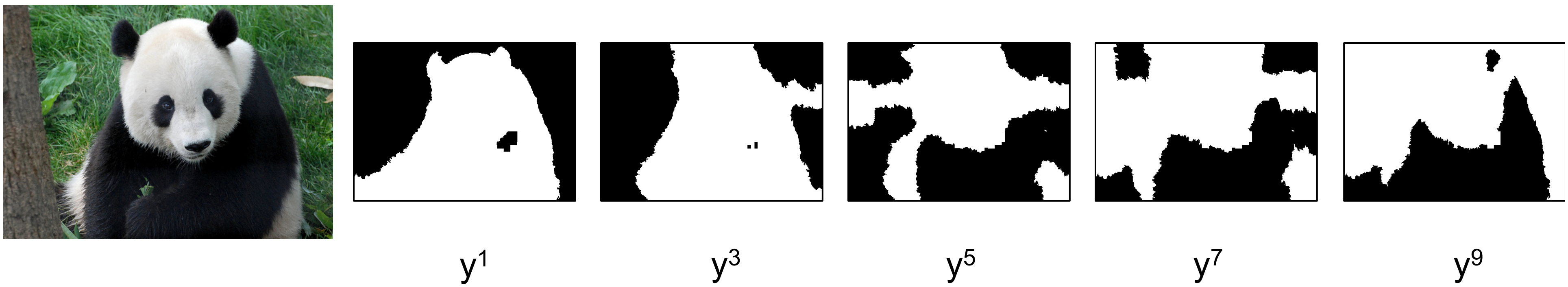
- High Entropy, Accurate MAP

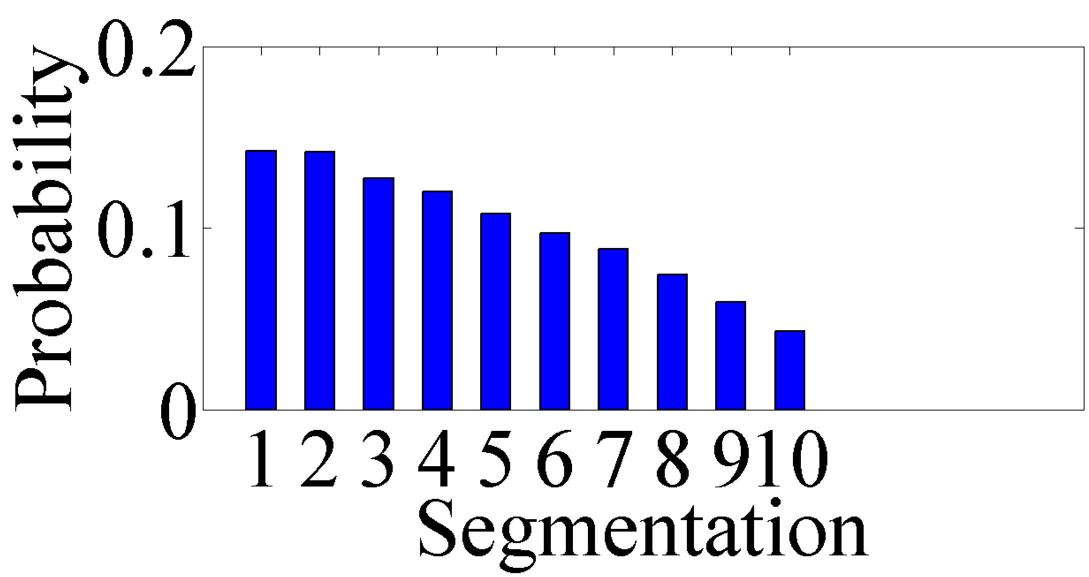
y^1 y^3 y^5 y^7 y^9



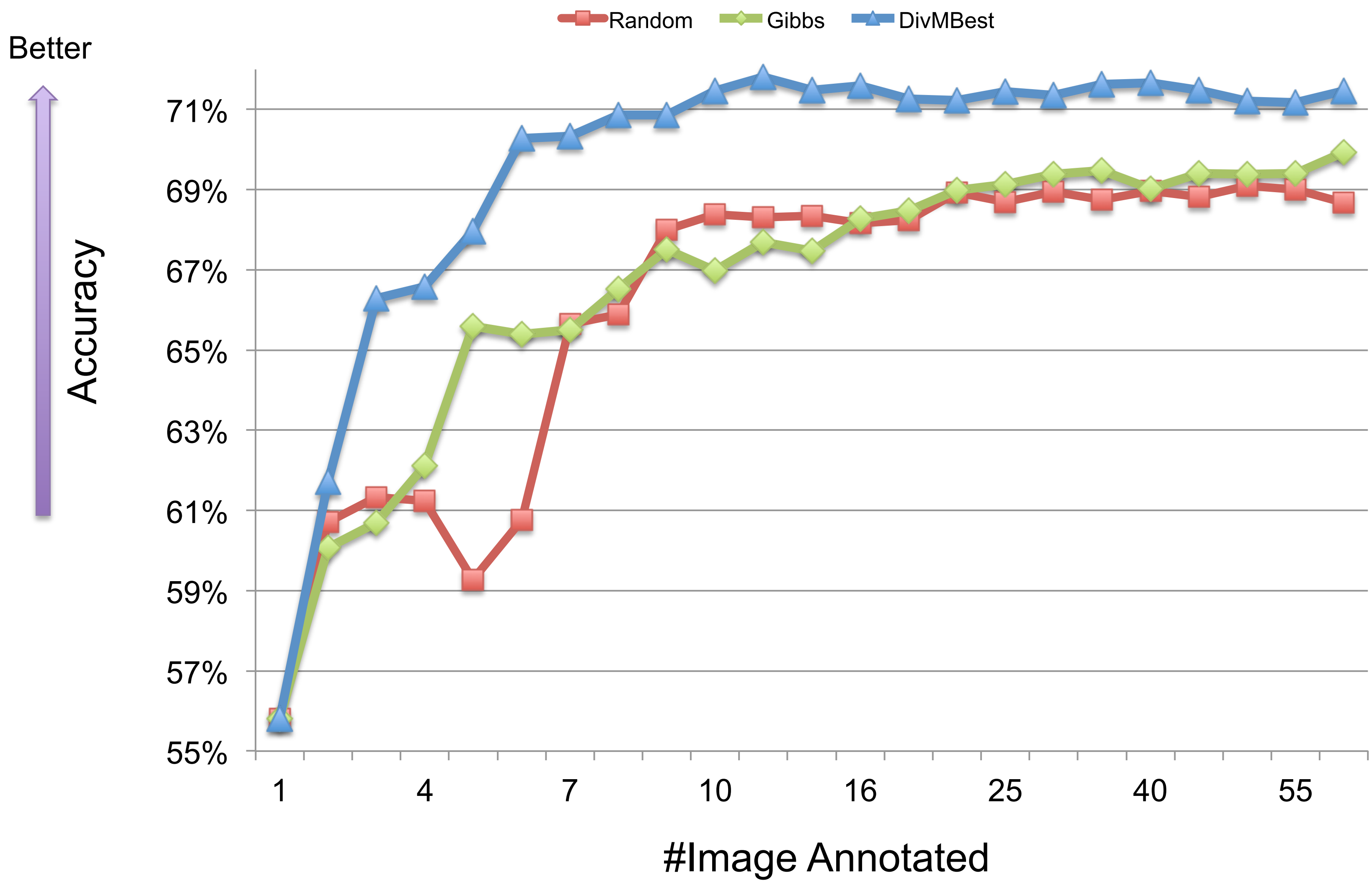
Active Learning

- Low Entropy, Accurate MAP





Active Learning



[Sun, Laddha, Batra, Under Review]

Improved Learning

- Classical Setup

```
for t = 1,...,converged
```

$$\mathbf{y}^1 = \texttt{Find\_MAP}(\theta_i,\ \theta_{ij})$$

$$\boldsymbol{\theta} = \texttt{Update}(\boldsymbol{\theta}, \phi(\mathbf{x}, \mathbf{y}^{gt}), \phi(\mathbf{x}, \mathbf{y}^1))$$

```
endfor
```

How to update?

$$\boldsymbol{\theta} \texttt{ = Update}(\boldsymbol{\theta}, \phi(\mathbf{x}, \mathbf{y}^{gt}), \phi(\mathbf{x}, \mathbf{y}^{1}))$$

- Structured SVMs
  - Subgradient of Hinge Loss

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \eta\Big[\phi(\mathbf{x}, \mathbf{y}^{gt}) - \phi(\mathbf{x}, \mathbf{y}^{1})\Big]$$

- CRFs
  - Gradient of log-liklihood

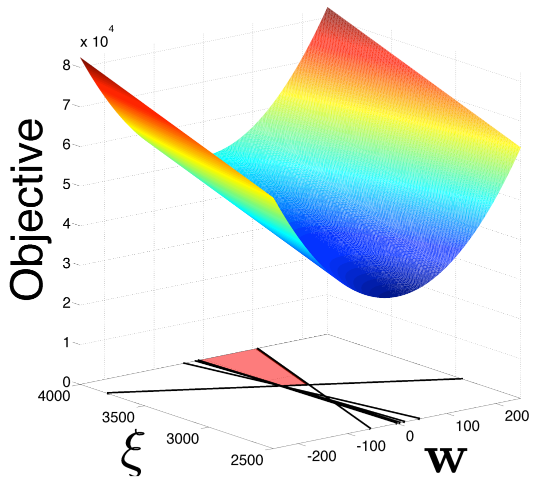
$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \eta\Big[\phi(\mathbf{x}, \mathbf{y}^{gt}) - \mathbb{E}_{\boldsymbol{\theta}^{(t)}}[\phi(\mathbf{x}, \mathbf{y})]\Big]$$

$$\sum_{\mathbf{y} \in \mathcal{Y}} \phi(\mathbf{x}, \mathbf{y})\, P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}^{(t)})$$

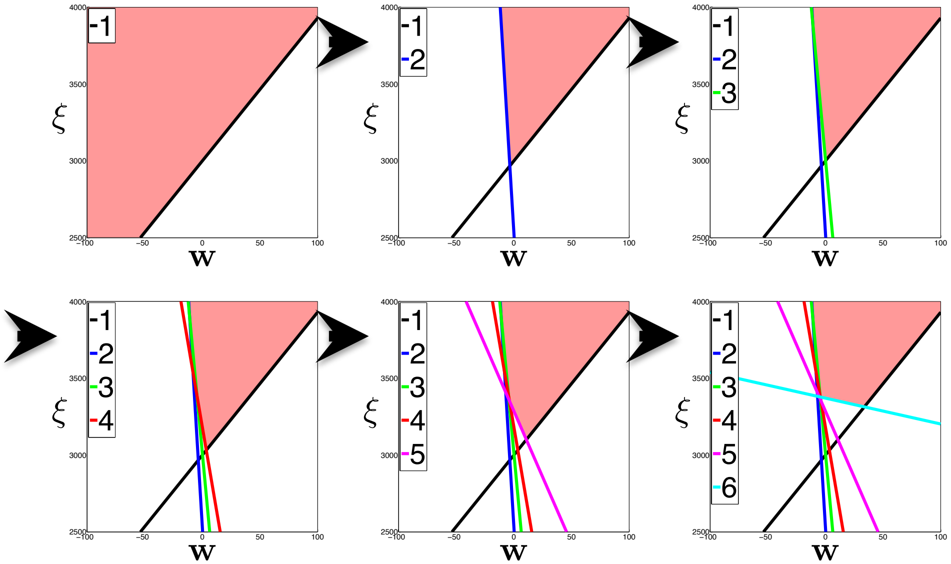
- [Meier, Sha, Globerson] $\displaystyle\sum_{\mathbf{y} \in Div\text{MBest}} \phi(\mathbf{x}, \mathbf{y})\, P(\mathbf{y}|\mathbf{x}; \boldsymbol{\theta}^{(t)})$

Structured SVM

Training SSVMs



Standard Approach, *M=1*

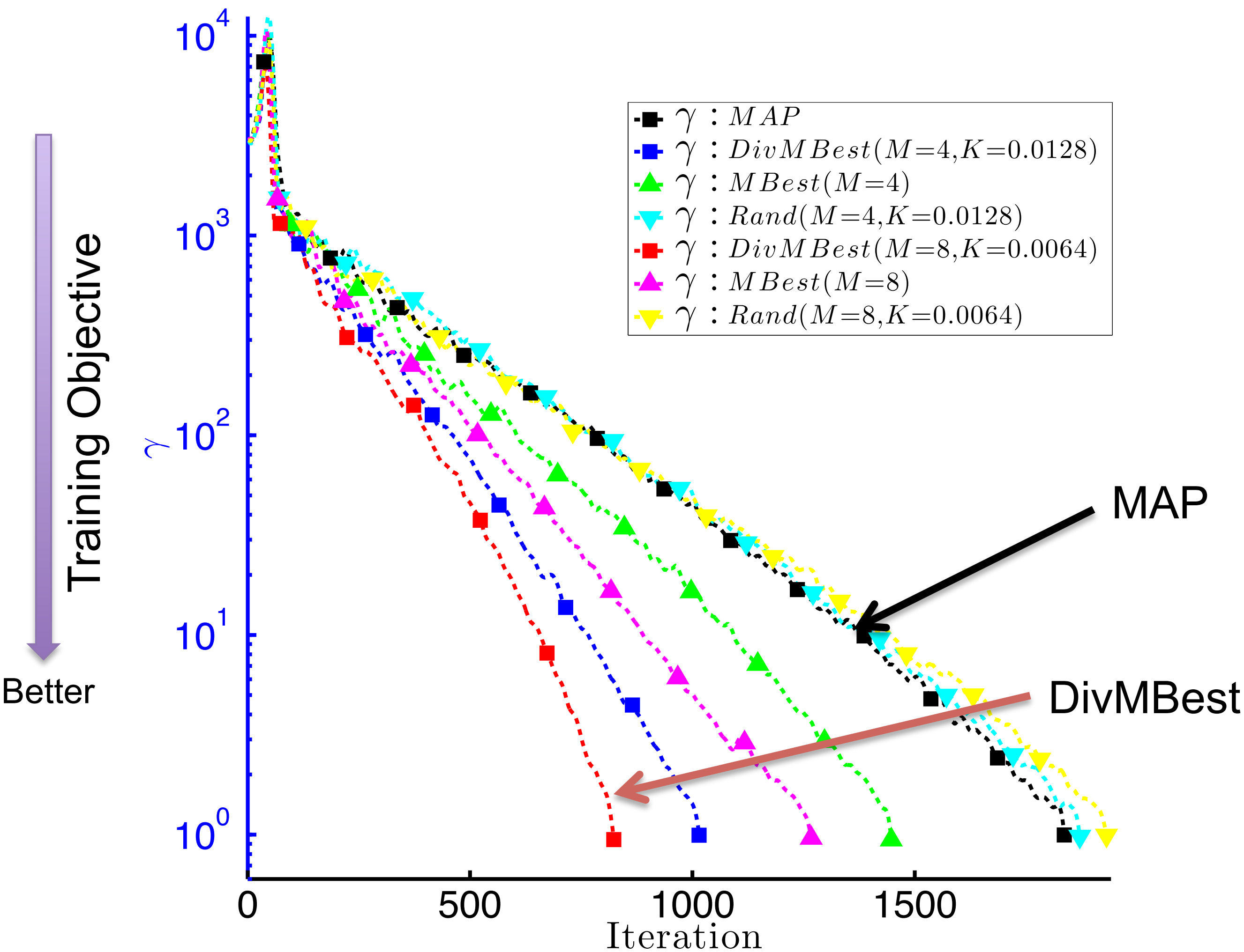


Exponential number of constraints.
Cutting-Plane!

[Guzman-Rivera, Kohli, Batra, AISTATS '12]

Structured SVM

- Training of SSVM with Diverse M-Best Cutting Planes



[Guzman-Rivera, Kohli, Batra, AISTATS '12]

Machine Translation

Input:

Die Regierung will die Folter von "Hexen" unterbinden und gab eine Broschüre heraus

5-Best Translations:

The government wants the torture of 'witch' and gave out a booklet
The government wants the torture of "witch" and gave out a booklet
The government wants the torture of 'witch' and gave out a brochure
The government wants the torture of 'witch' and gave out a leaflet
The government wants the torture of "witch" and gave out a brochure

[Gimpel, Batra, Dyer, Shakhnarovich, EMNLP13]

Machine Translation

Input:

Die Regierung will die Folter von "Hexen" unterbinden und gab eine Broschüre heraus

Diverse 5-Best Translations:

The government wants the torture of 'witch' and gave out a booklet
The government wants to stop torture of "witch" and issued a leaflet issued
The government wants to "stop the torture of" witches and gave out a brochure
The government intends to the torture of "witchcraft" and were issued a leaflet
The government is the torture of "witches" stamp out and gave a brochure

[Gimpel, Batra, Dyer, Shakhnarovich, EMNLP13]

Machine Translation

Input:

Die Regierung will die Folter von "Hexen" unterbinden und gab eine Broschüre heraus

Diverse 5-Best Translations:

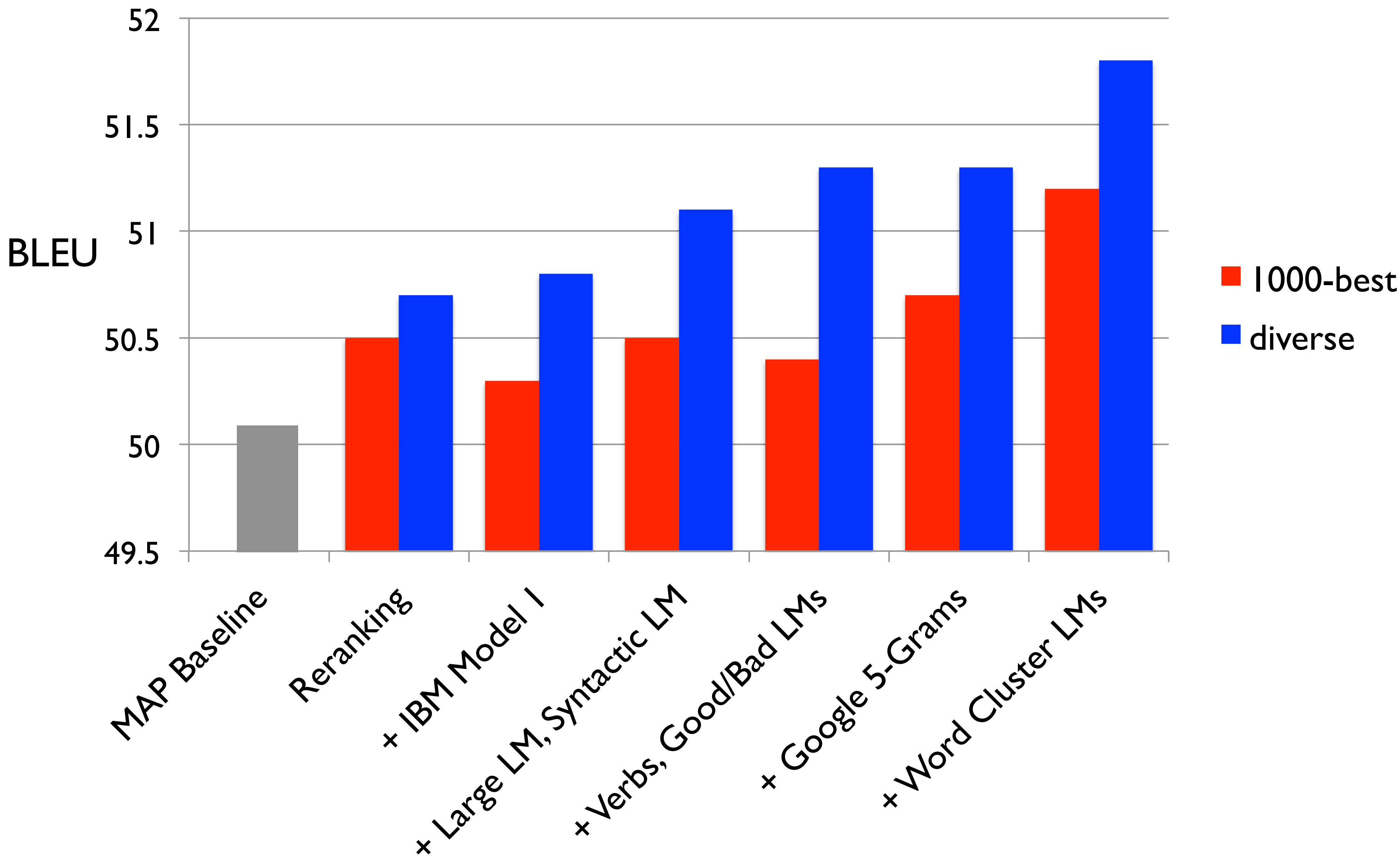
The government wants the torture of 'witch' and gave out a booklet
The government wants to stop torture of "witch" and issued a leaflet issued
The government wants to "stop the torture of" witches and gave out a brochure
The government intends to the torture of "witchcraft" and were issued a leaflet
The government is the torture of "witches" stamp out and gave a brochure

Correct Translation:

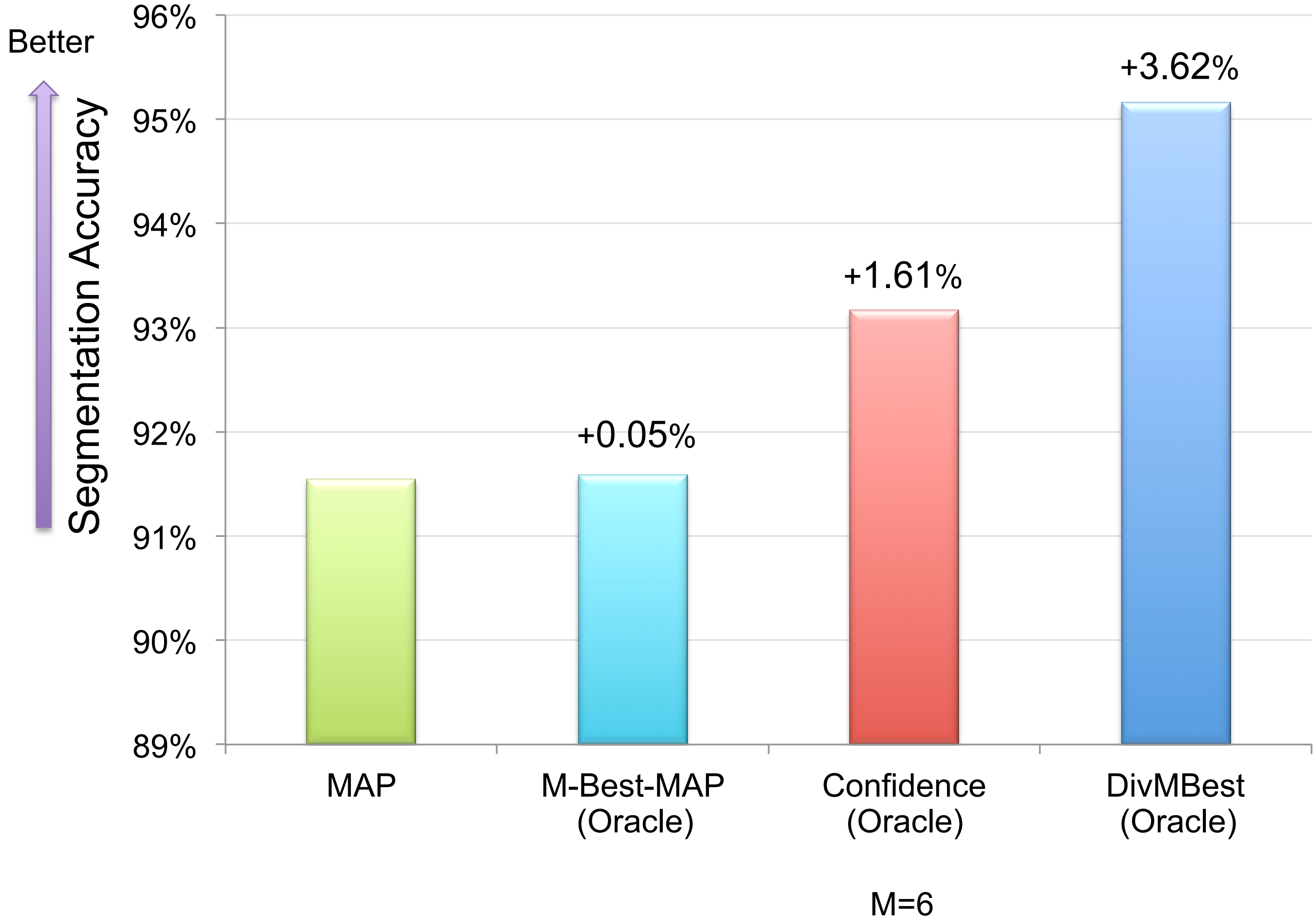
The government wants to limit the torture of "witches," a brochure was released

[Gimpel, Batra, Dyer, Shakhnarovich, EMNLP13]

Arabic-English Reranking Results



Interactive Segmentation



MAP Integer Program

$$\max_{\mathbf{y}} \; S(\mathbf{y}) = \sum_i \theta_i(y_i) + \sum_{(i,j)} \theta_{ij}(y_i, y_j)$$

$$\left[\ \cdot \;\; \theta_i(s) \;\; \cdot \;\; \cdot\ \right] \begin{bmatrix} \cdot \\ \mu_i(s) \\ \cdot \\ \cdot \end{bmatrix}$$

kx1

MAP Integer Program

$$\max_{\mathbf{y}} \; S(\mathbf{y}) = \sum_i \left[\cdot \;\; \theta_i(s) \;\; \cdot \;\; \cdot \right] \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \end{bmatrix}_{k\text{x}1} + \sum_{(i,j)} \theta_{ij}(y_i, y_j)$$

MAP Integer Program

$$\max_{\mathbf{y}} \; S(\mathbf{y}) = \sum_i \begin{bmatrix} \cdot & \theta_i(s) & \cdot & \cdot \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix}_{k\text{x}1} + \sum_{(i,j)} \theta_{ij}(y_i, y_j)$$

MAP Integer Program

$$\max_{\mathbf{y}} \; S(\mathbf{y}) = \sum_i \begin{bmatrix} \cdot & \theta_i(s) & \cdot & \cdot \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}_{\text{kx1}} + \sum_{(i,j)} \theta_{ij}(y_i, y_j)$$

MAP Integer Program

$$\max_{\mathbf{y}} \; S(\mathbf{y}) = \sum_i \begin{bmatrix} \cdot & \theta_i(s) & \cdot & \cdot \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}_{\text{kx1}} + \sum_{(i,j)} \theta_{ij}(y_i, y_j)$$

MAP Integer Program

$$\max_{\mathbf{y}} \; S(\mathbf{y}) = \sum_i \begin{bmatrix} \cdot & \theta_i(s) & \cdot & \cdot \end{bmatrix} \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}_{k\text{x}1} + \sum_{(i,j)} \begin{bmatrix} \cdot & \cdot & \theta_{ij}(s,t) & \cdot & \cdot \end{bmatrix} \begin{bmatrix} \cdot \\ \cdot \\ \cdot \\ \cdot \\ \mu_{ij}(s,t) \\ \cdot \\ \cdot \\ \cdot \\ \cdot \end{bmatrix}_{k^2\text{x}1}$$

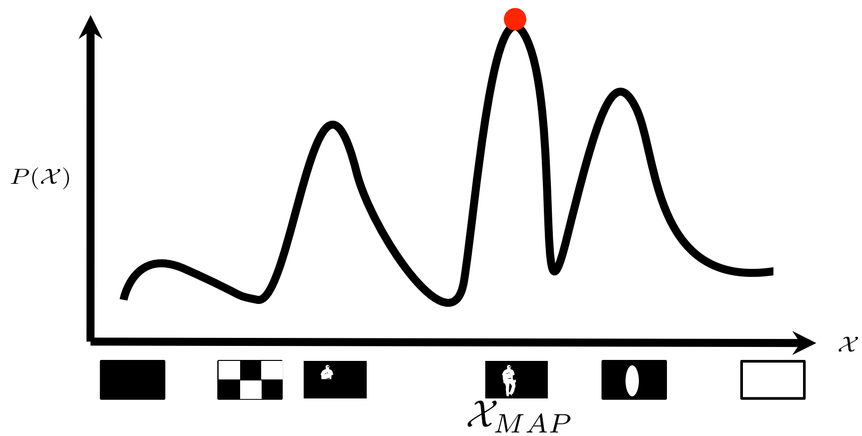
MAP Integer Program

$$\max_{\mathbf{y}} \; S(\mathbf{y}) = \sum_i \max_{\boldsymbol{\mu}\in\mathcal{C}} \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$$

$$s.t. \quad \boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0,1\}$$

$$\boldsymbol{\theta}_i = \begin{bmatrix} \cdots & \theta_i(s) & \cdots \end{bmatrix} \qquad \boldsymbol{\mu}_i = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}_{k \times 1}$$

$$\boldsymbol{\theta}_{ij} = \begin{bmatrix} \cdots & \theta_{ij}(s,t) & \cdots \end{bmatrix} \qquad \boldsymbol{\mu}_{ij} = \begin{bmatrix} \vdots \\ \mu_{ij}(s,t) \\ \vdots \end{bmatrix}_{k^2 \times 1}$$

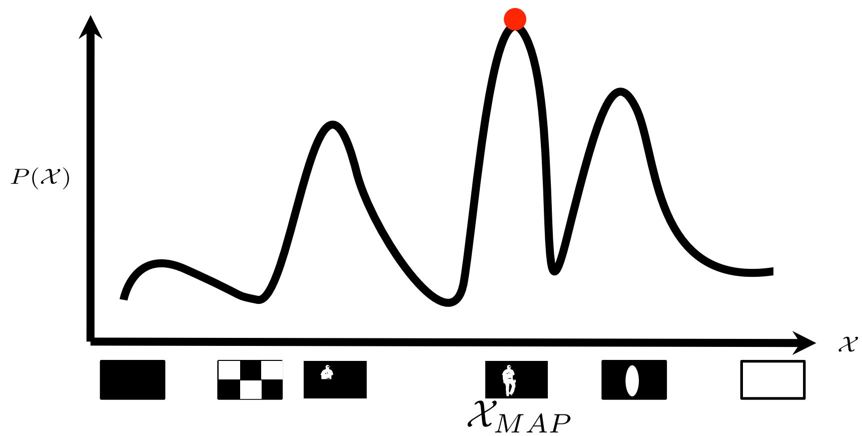


MAP Integer Program

$$\max_{\boldsymbol{\mu}\in\mathcal{C}} \quad \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$$

$$s.t. \quad \boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0,1\}$$

Graphcuts, BP, Expansion, etc



Diverse 2nd-Best

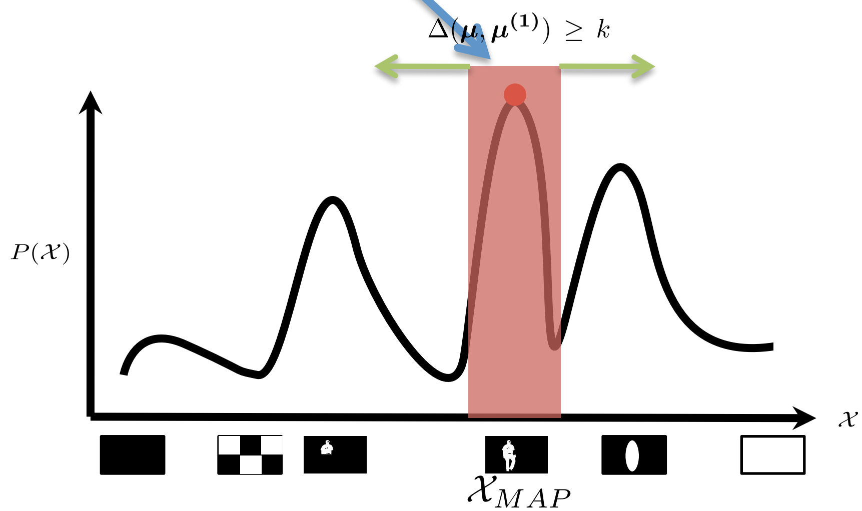
$$\max_{\boldsymbol{\mu} \in \mathcal{C}} \quad \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$$

$$s.t. \quad \boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0, 1\}$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}) \geq k$$

Diversity

MAP



Diverse M-Best

$$\max_{\boldsymbol{\mu} \in \mathcal{C}} \quad \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$$

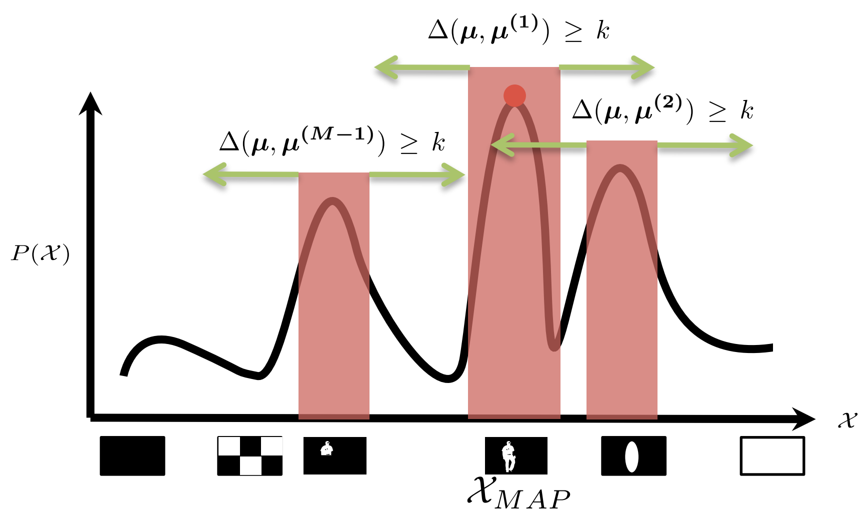
$$s.t. \quad \boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0,1\}$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}) \geq k$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(2)}) \geq k$$

$$\vdots$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(M-1)}) \geq k$$



Diverse 2nd-Best

$$\max_{\boldsymbol{\mu} \in \mathcal{C}} \quad \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$$

$$s.t. \quad \boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0,1\}$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(\mathbf{1})}) \geq k$$

Q1: How do we solve DivMBest?

Q2: What kind of diversity functions are allowed?

Q3: How much diversity?

Diverse 2nd-Best

Diversity-Augmented Score

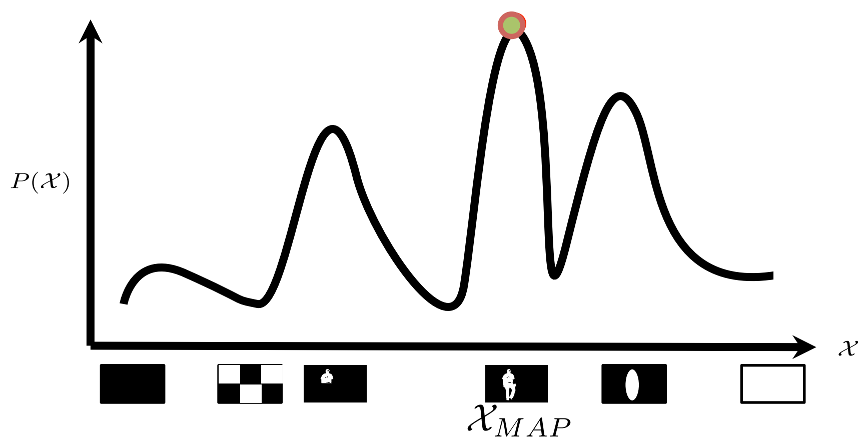
Primal

$$\max_{\boldsymbol{\mu} \in \mathcal{C}} \quad \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij} \quad + \lambda \cdot \Big(\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}) - k \Big)$$

$$s.t. \quad \boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0, 1\}$$

Dualize

$$S(\mathbf{y}) \quad + \quad Div(\mathbf{y}, \mathbf{y}^{(1)})$$



Diverse 2nd-Best

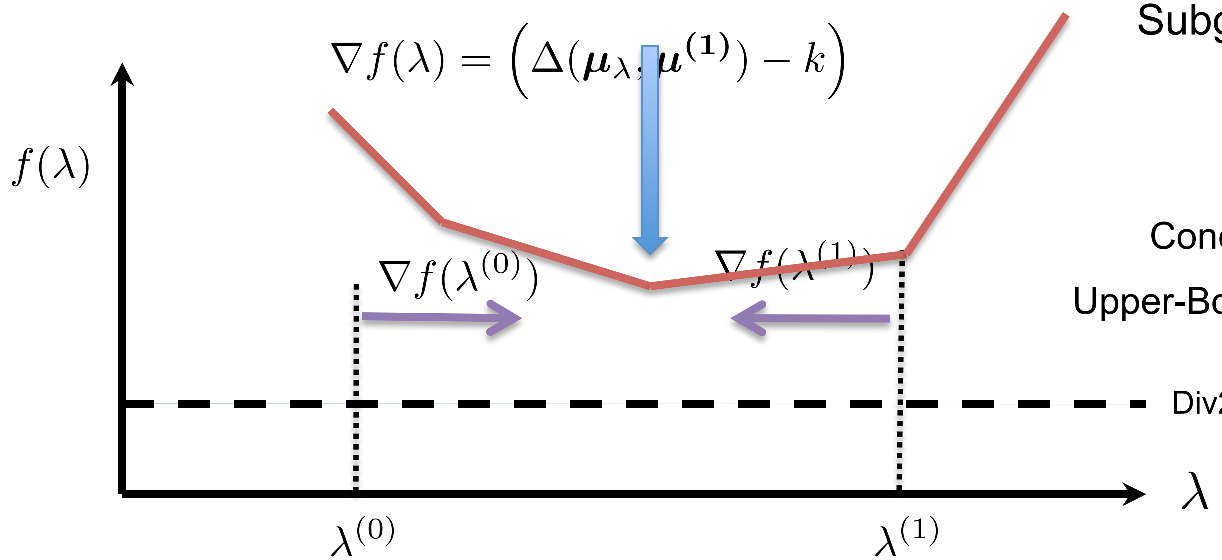
- Lagrangian Relaxation

Diversity-Augmented Score

$$f(\lambda) = \max_{\boldsymbol{\mu} \in \mathcal{C}} \underbrace{\sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij} + \lambda \cdot \left(\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}) - k \right)}_{\text{Diversity-Augmented Score}}$$

$$s.t. \quad \boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0, 1\}$$

Dual $\min_{\lambda \geq 0} \; f(\lambda)$



Subgradient Descent

Concave (Non-smooth)

Upper-Bound on Div2Best Score

Diverse 2nd-Best

- Many ways to solve:
  1. Subgradient Ascent. Optimal. Slow.
  2. Binary Search. Optimal for M=2. Faster.
  3. Grid-search on lambda. Sub-optimal. Fastest.

Theorem Statement

- Theorem [Batra et al '12]: Lagrangian Dual corresponds to solving the Relaxed Primal:
  - Based on result from [Geoffrion '74]

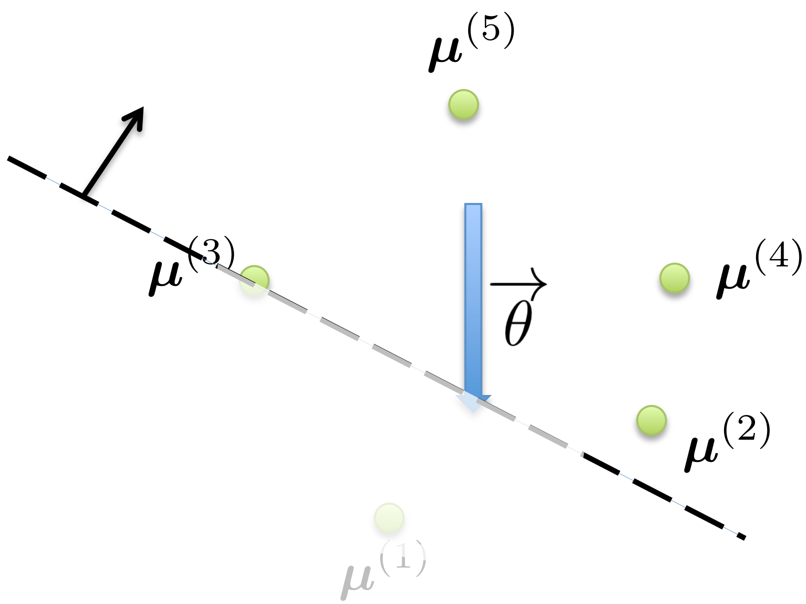
Dual

$$\min_{\lambda \geq 0} \quad \text{LagrangianDual}(\lambda)$$

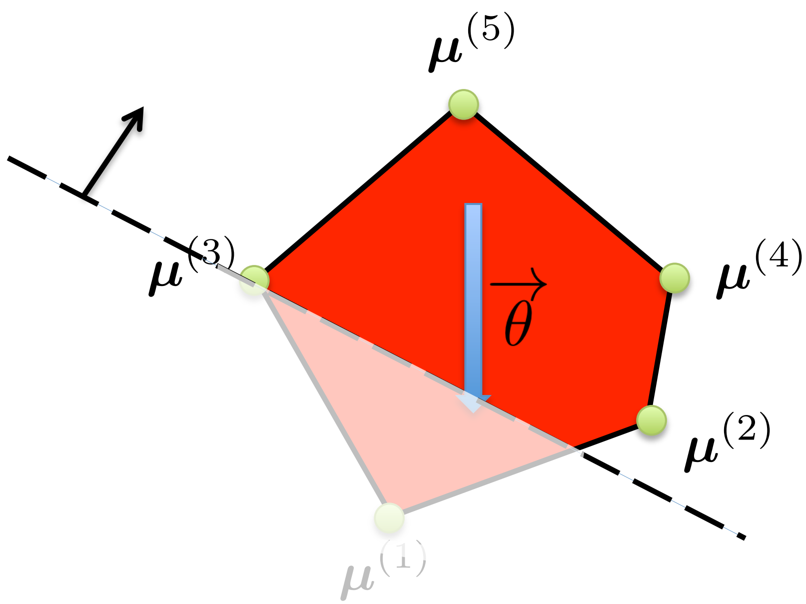
Relaxed Primal

$$\begin{aligned} \max_{\boldsymbol{\mu}} \quad & \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{ij} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij} \\ s.t. \quad & \boldsymbol{\mu} \in Co\big\{\boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0,1\} \mid \boldsymbol{\mu} \in \mathcal{C}\big\} \\ & \Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}) \geq k \end{aligned}$$

Effect of Lagrangian Relaxation

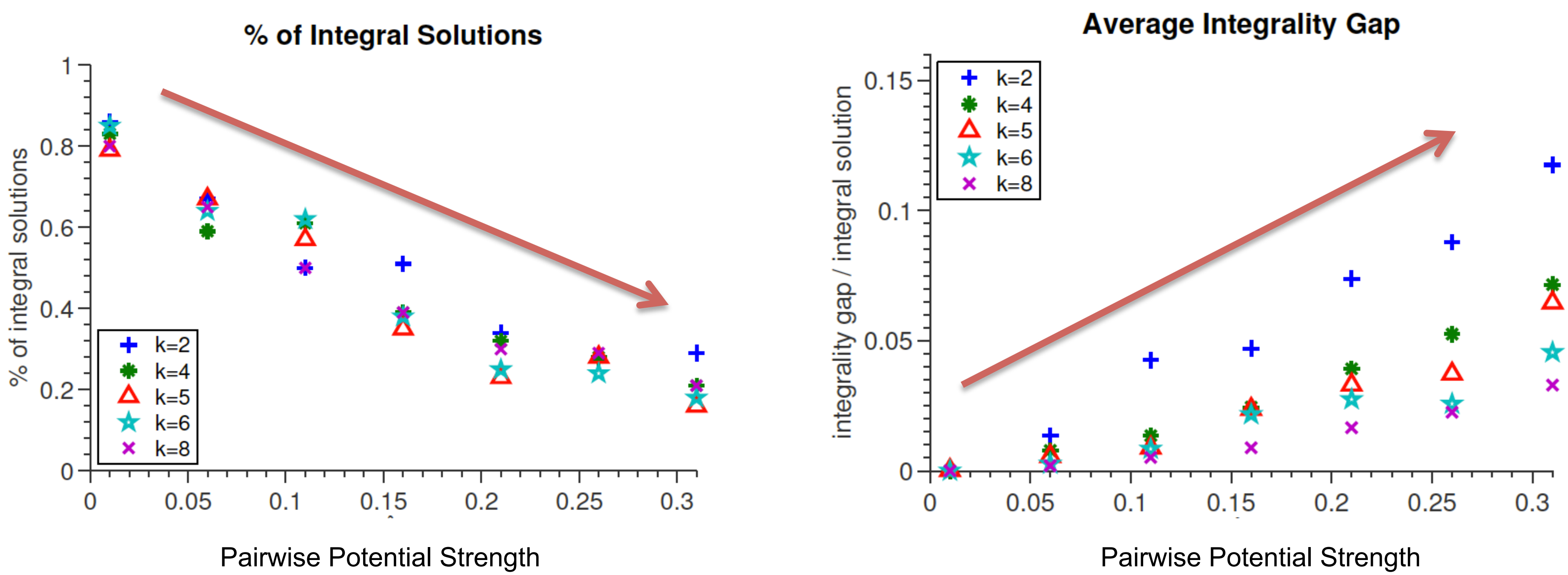


Effect of Lagrangian Relaxation



Effect of Lagrangian Relaxation

- [Mezuman et al. UAI13]



Diverse 2nd-Best

$$\min_{\boldsymbol{\mu} \in \mathcal{C}} \quad \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$$

$$s.t. \quad \boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0, 1\}$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(\mathbf{1})}) \geq k$$

Q1: How do we solve DivMBest?

Q2: What kind of diversity functions are allowed?

Q3: How much diversity?

Diversity

- [Special Case] 0-1 Diversity $\Longrightarrow$ M-Best MAP
  - [Yanover NIPS03; Fromer NIPS09; Flerova Soft11]
- [Special Case] Max Diversity $\Longrightarrow$ [Park & Ramanan ICCV11]
- Hamming Diversity
- Cardinality Diversity
- Any Diversity $\max_{\mu \in \mathcal{C}} \quad S(\boldsymbol{\mu}) + \lambda \, \Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)})$

Hamming Diversity

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}) = -\sum_{i \in \mathcal{V}} \underbrace{\boldsymbol{\mu}_i \cdot \boldsymbol{\mu}_i^{(1)}}$$

$$\begin{bmatrix} 0 & 1 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = 1$$

$$\begin{bmatrix} 1 & 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \\ 0 \\ 0 \end{bmatrix} = 0$$

Hamming Diversity

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}) = -\sum_{i \in \mathcal{V}} \boldsymbol{\mu}_i \cdot \boldsymbol{\mu}_i^{(1)}$$

- Diversity Augmented Inference:

$$\max_{\boldsymbol{\mu} \in \mathcal{C}} \quad \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij} + \lambda \, \Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)})$$

$$= \sum_i \underbrace{\left(\boldsymbol{\theta}_i - \lambda \boldsymbol{\mu}_i^{(1)}\right)}_{\tilde{\boldsymbol{\theta}}_i} \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$$

Hamming Diversity

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}) = -\sum_{i \in \mathcal{V}} \boldsymbol{\mu}_i \cdot \boldsymbol{\mu}_i^{(1)}$$

- Diversity Augmented Inference:

$\texttt{for i = 1,2,...,n}$

$\quad \theta_i\texttt{[}y_i^1\texttt{] -=}\ \lambda$

$\texttt{endfor}$

$\mathbf{y}^2\ \texttt{= Find\_MAP(}\theta_i\texttt{,}\ \theta_{ij}\texttt{)}$

Unchanged.
Can still use graph-cuts!

Simply edit node-terms. Reuse MAP machinery!

Diverse 2nd-Best

$$\min_{\boldsymbol{\mu} \in \mathcal{C}} \quad \sum_i \boldsymbol{\theta}_i \cdot \boldsymbol{\mu}_i + \sum_{(i,j)} \boldsymbol{\theta}_{ij} \cdot \boldsymbol{\mu}_{ij}$$

$$s.t. \quad \boldsymbol{\mu}_i(\cdot), \boldsymbol{\mu}_{ij}(\cdot) \in \{0, 1\}$$

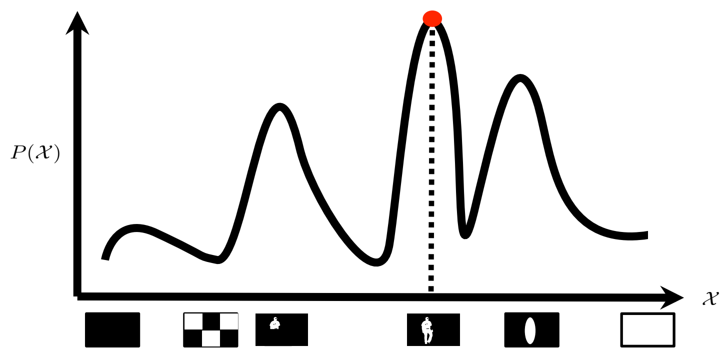
$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(\mathbf{1})}) \geq k$$

Q1: How do we solve DivMBest?

Q2: What kind of diversity functions are allowed?

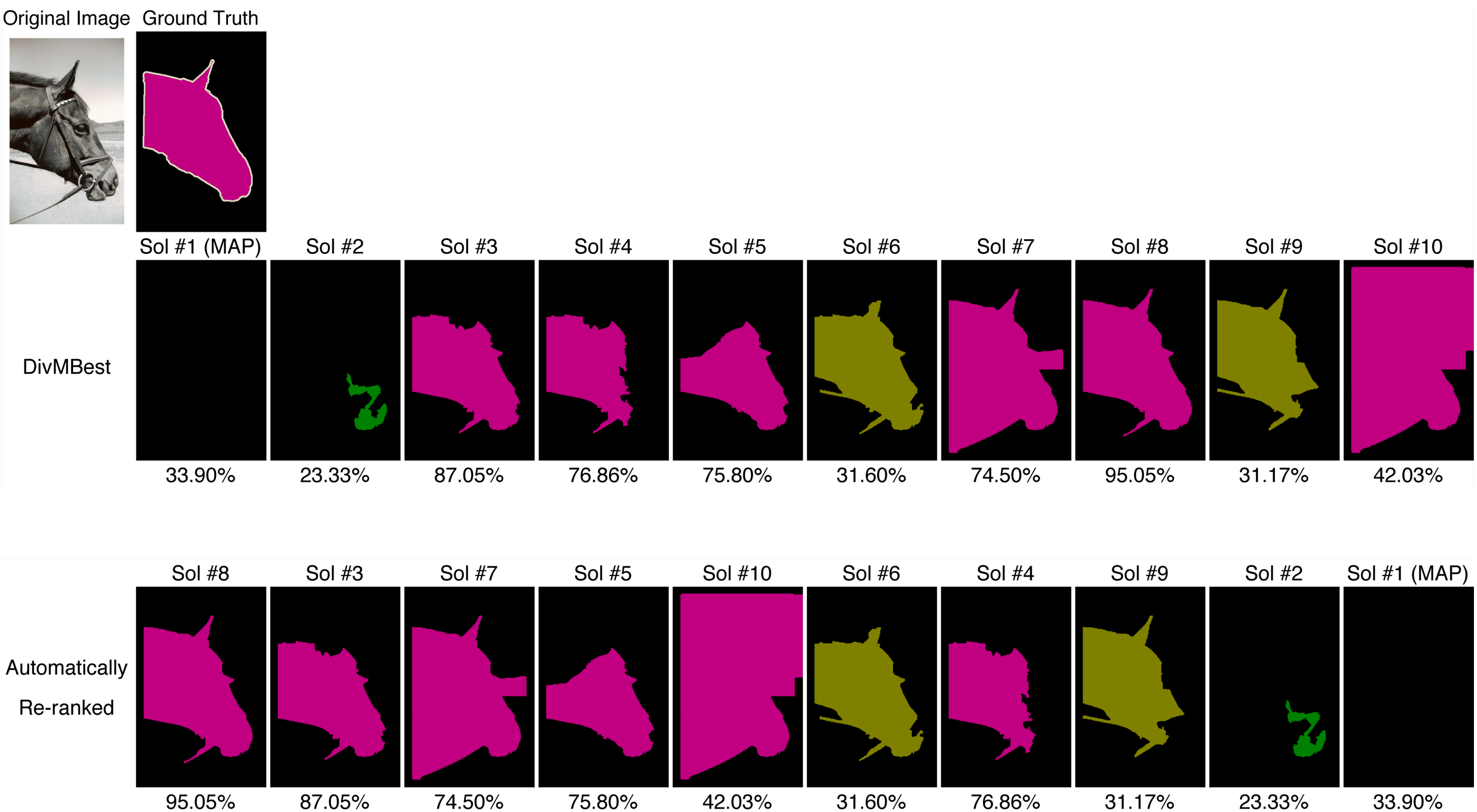
Q3: How much diversity?

How Much Diversity?

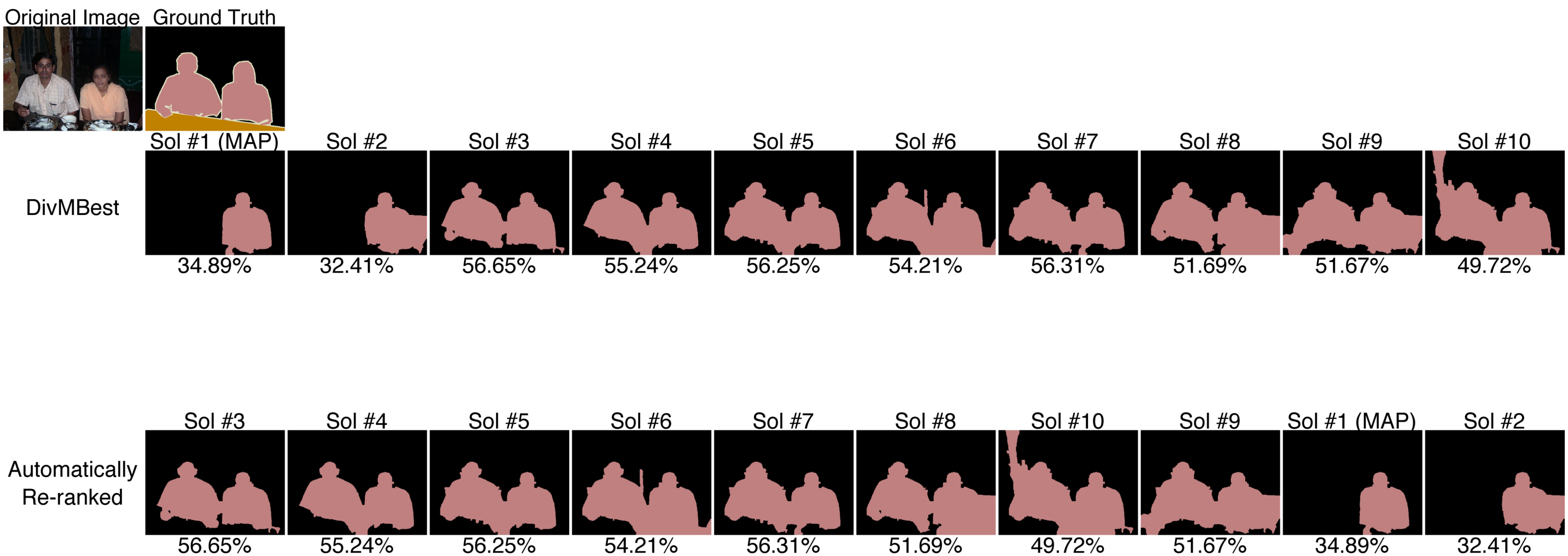


- Empirical Solution: Cross-Val for k
- More Efficient: Cross-Val for λ

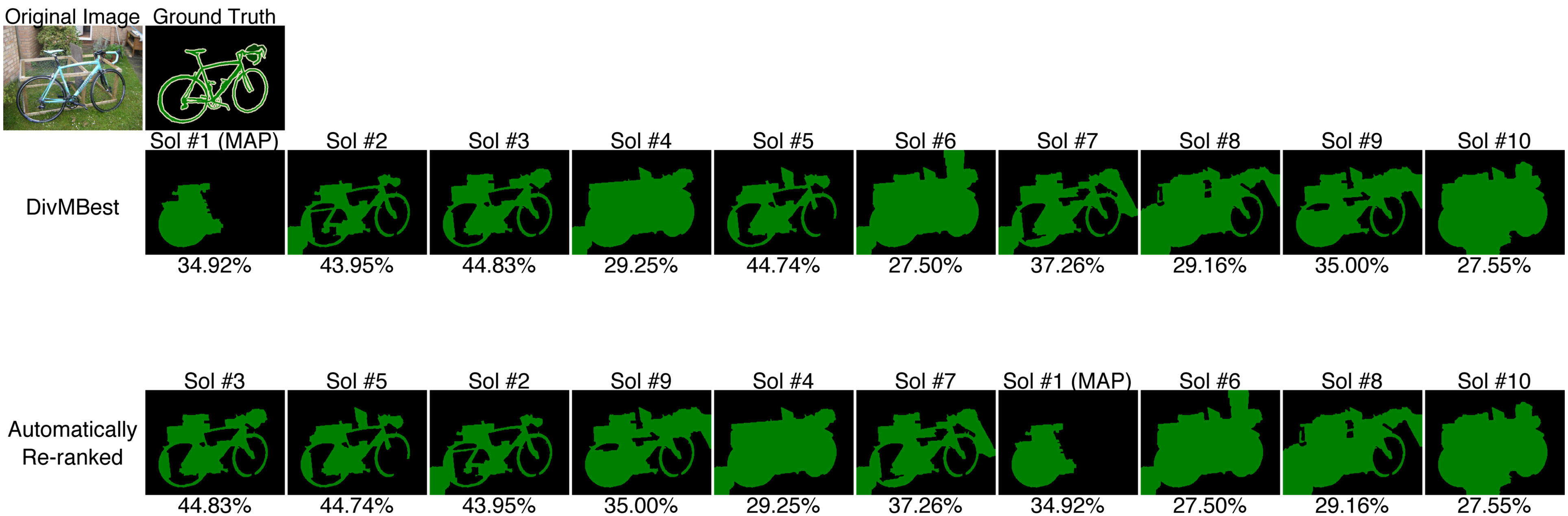
Qualitative Results: Success



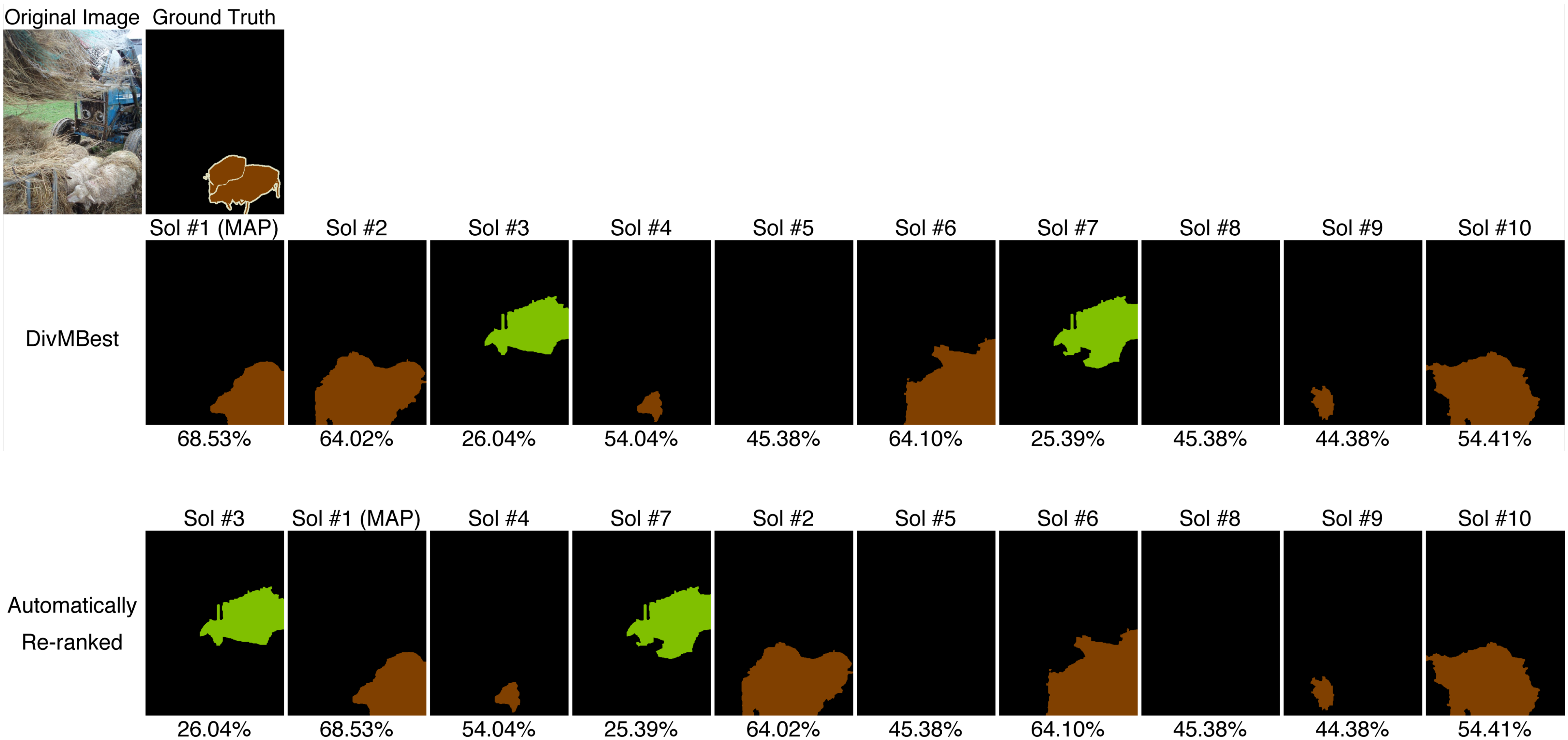
Qualitative Results: Success



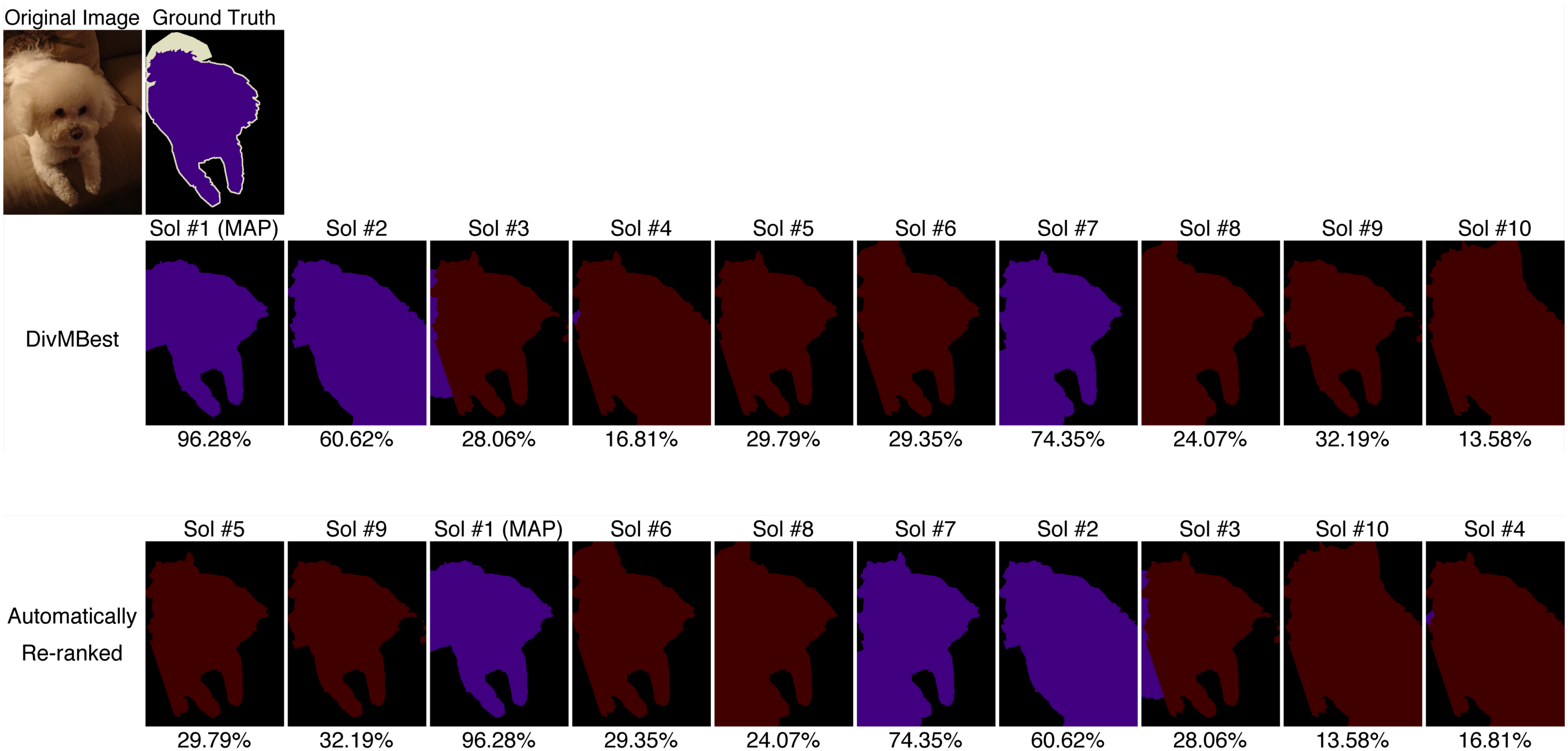
Qualitative Results: Success



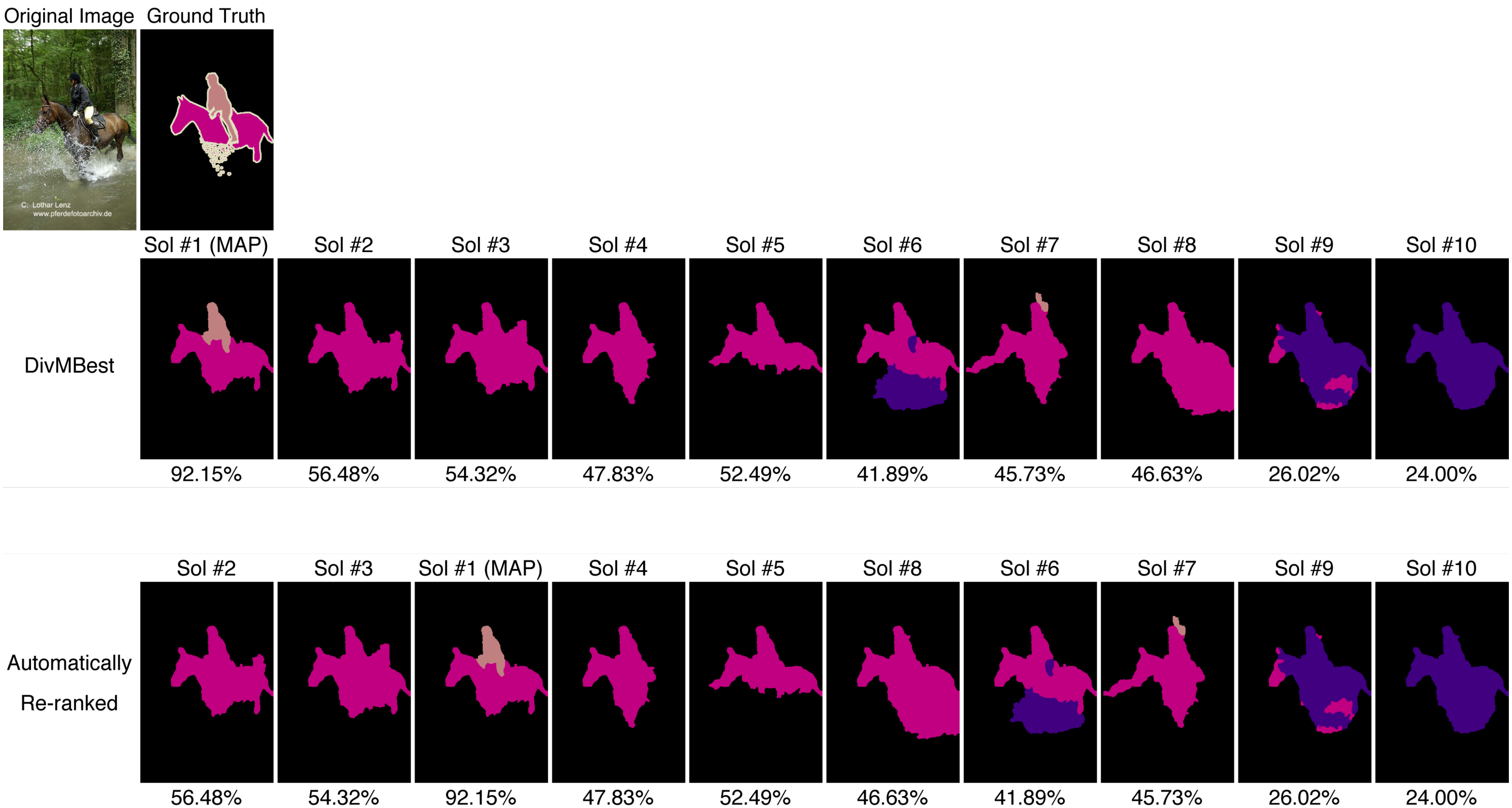
Qualitative Results: Failures



Qualitative Results: Failures



Qualitative Results: Failures



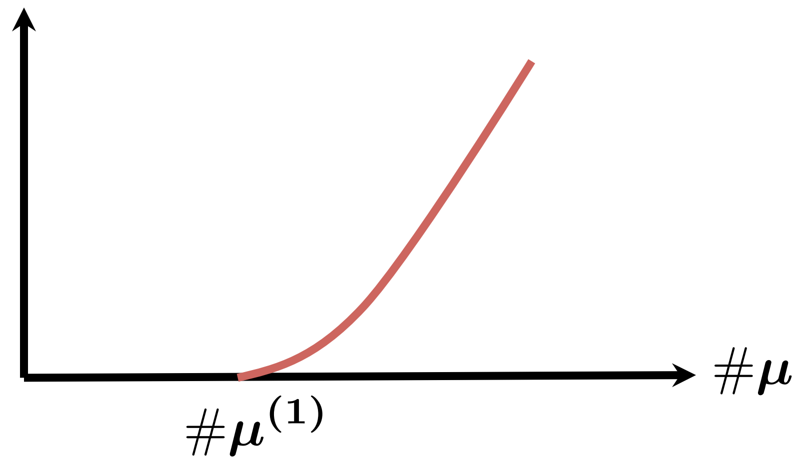
Higher-Order Dissimilarity

- Cardinality Potential

$$\#\boldsymbol{\mu} \triangleq \sum_i \mu_i(1)$$

$$\Delta(\boldsymbol{\mu}, \boldsymbol{\mu}^{(1)}) = \begin{cases} \left(\#\boldsymbol{\mu} - \#\boldsymbol{\mu}^{(1)}\right)^2 & \text{if } \#\boldsymbol{\mu} \geq \#\boldsymbol{\mu}^{(1)} \\ 0 & \text{else} \end{cases}$$

- Efficient Inference
  - Cardinality [Tarlow ‘10]
  - Lower Linear envelop [Kohli ‘10]
  - Pattern Potentials [Rother ‘10]



Example Results