# Guest Lecture in Computer Vision (CS 670)

## November 13, 2013

UMASS AMHERST

# Augmenting CRFs with Boltzmann Machine Shape Priors for Image Labeling

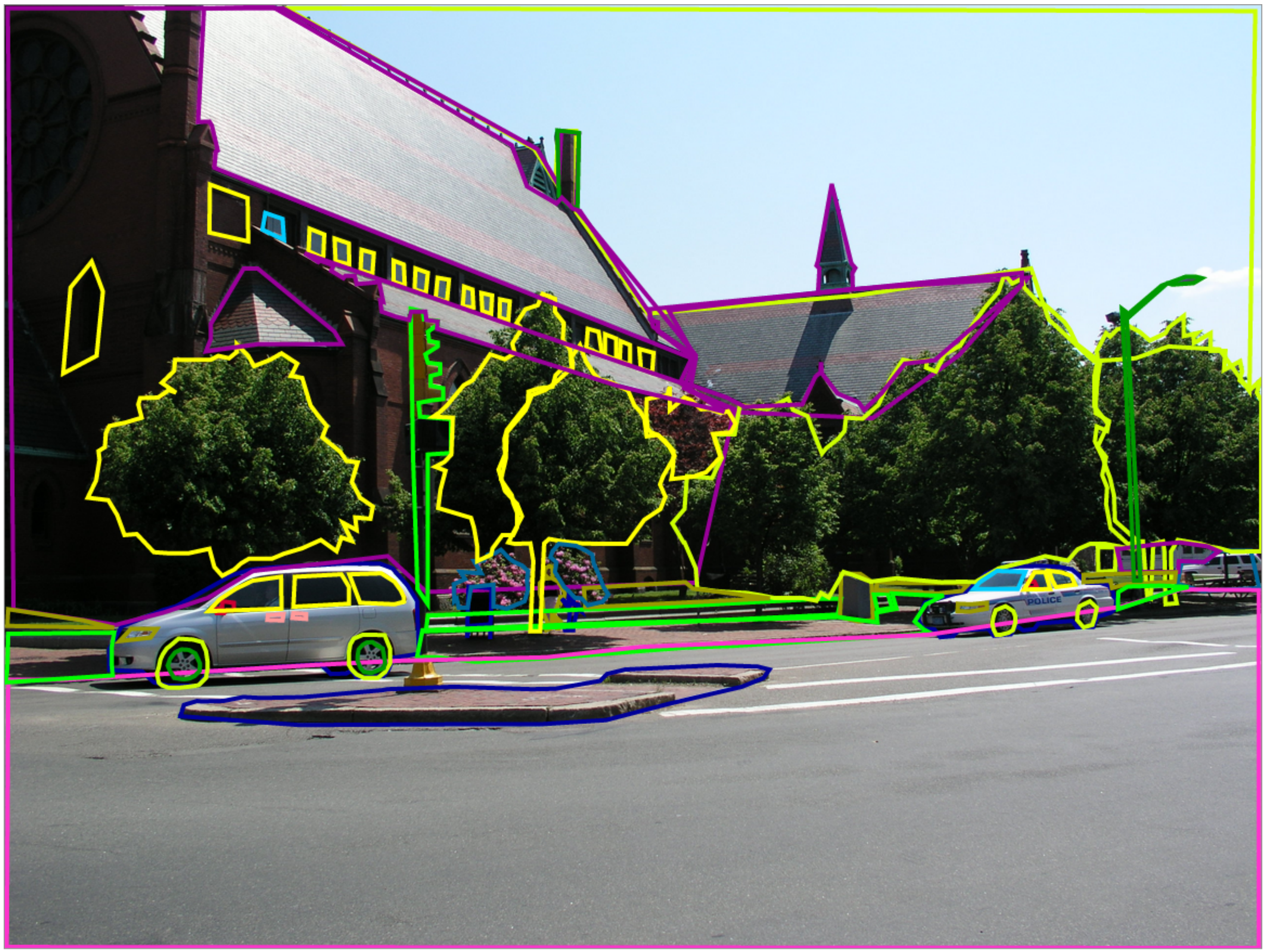
Andrew Kae
Erik Learned-Miller

Kihyuk Sohn
Honglak Lee

UMASS AMHERST

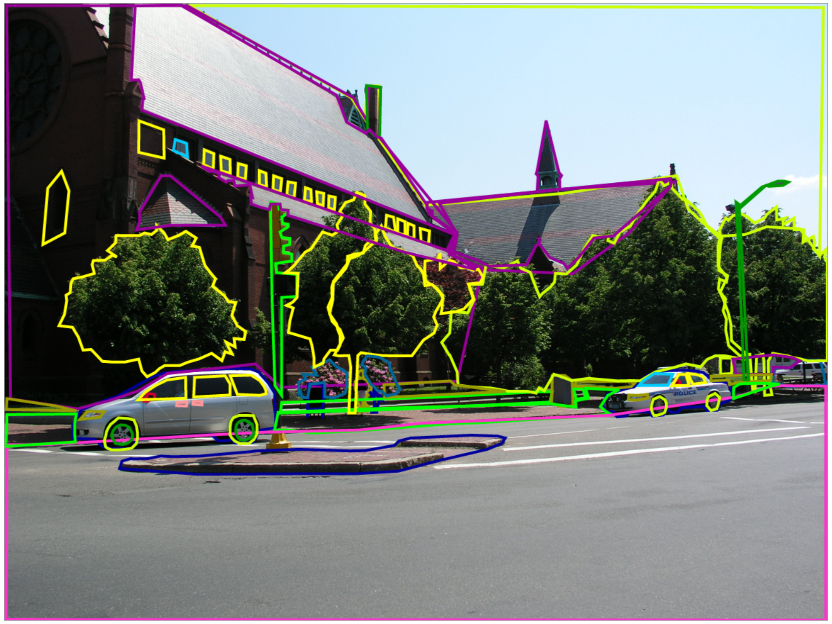
MICHIGAN

# Image Labeling [LabelMe]




carSide
carSide
car
sky
building
roof
traffic lights
litter bin
person walking
child walking
stairs
pedestal
stairs
grass
sidewalk
central reservation
trees
gate
trees
road
grass
sidewalk
tree
plants
tree
tree
tree
building
grass
sidewalk

# Image Labeling



Useful for

- object detection
- part analysis
- scene understanding

# Task

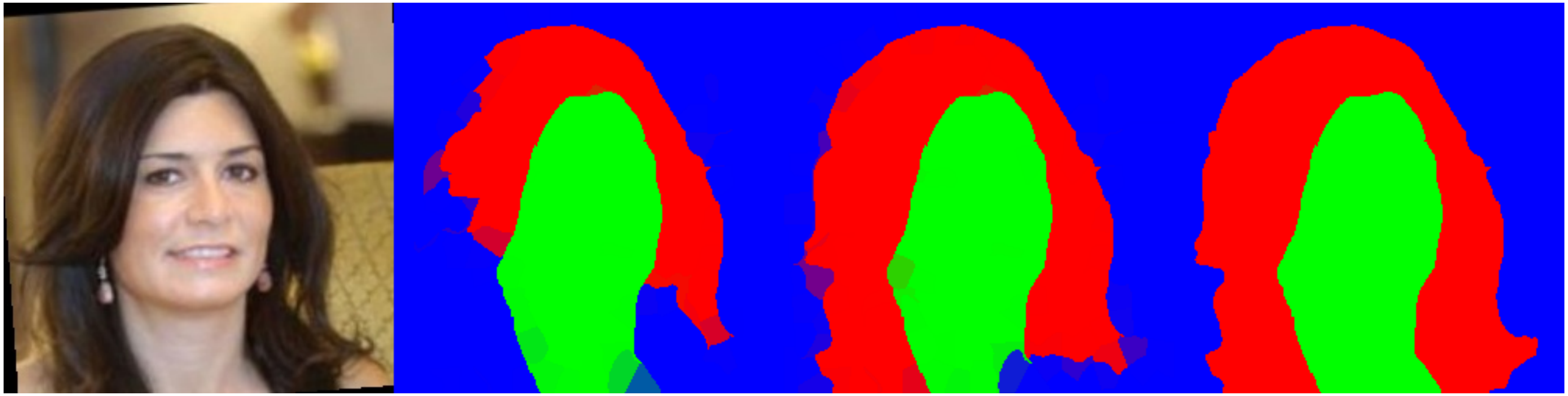
Aligned Image

Ground Truth

Why do we care?

- better understand face structure
- obtain useful face descriptions
- may be useful for other tasks such as recognition, retrieval

# Task




- Problem with model based only on local information.
    - Result doesn't *look* like hair/skin shape
- Useful to incorporate **global** shape information

# Goals

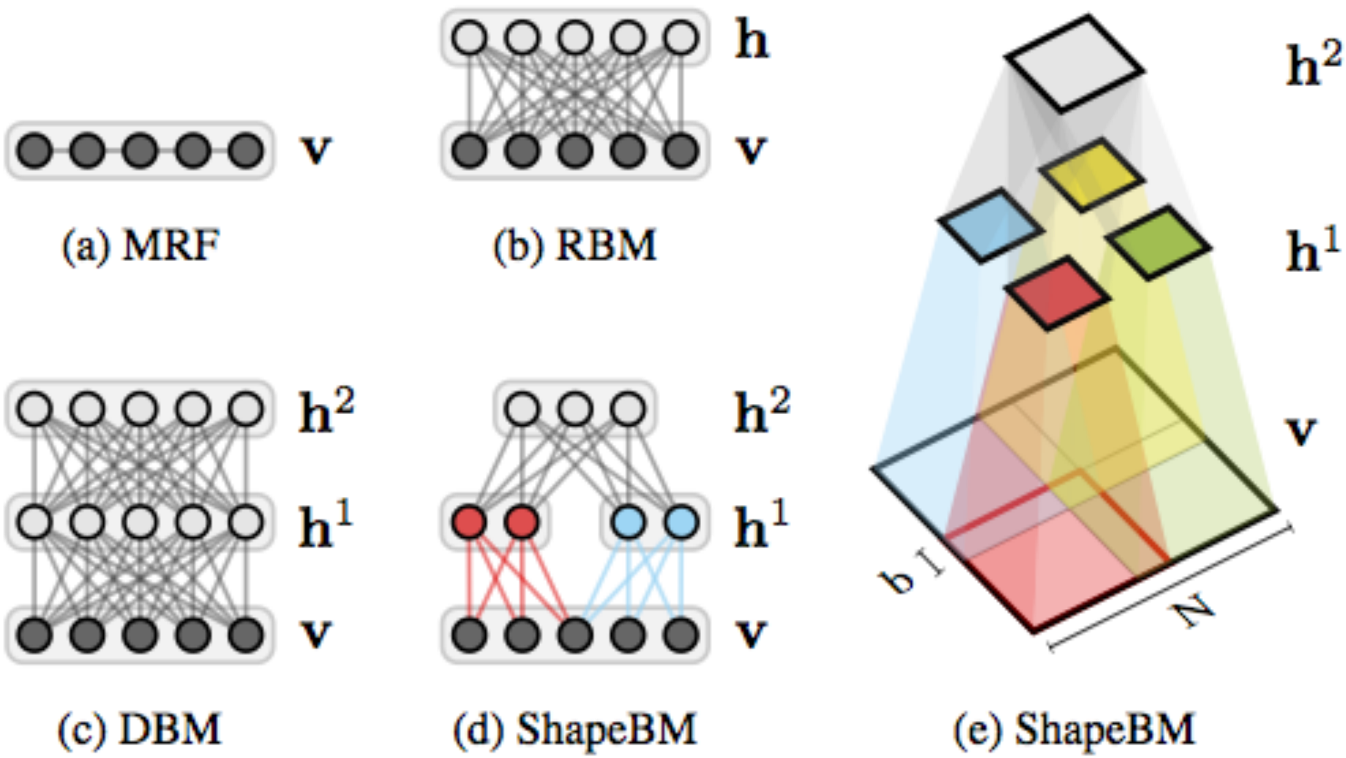
- Incorporate shape information to model **global** and **local** information together
- Demonstrate the improved performance of this hybrid (GLOC) model
- Learn efficient training/inference methods
- Learn face descriptions ("attributes")

# Contents

# Previous Work

# Shape Boltzmann Machine [Eslami et al. 2012]

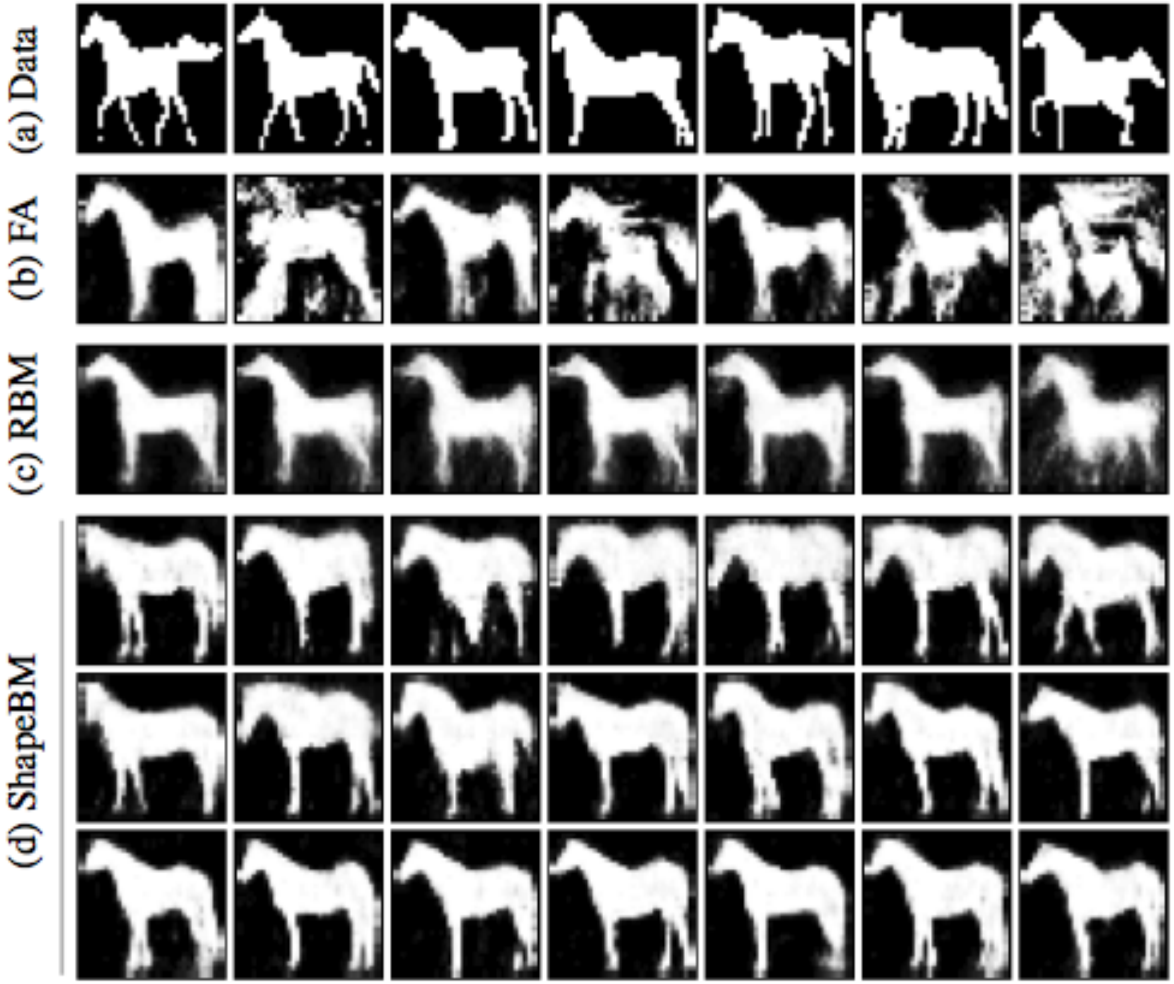
- Modified deep Boltzmann machine



# Shape Boltzmann Machine [Eslami et al. 2012]

- MRF
  - unary, pairwise potentials
- RBM
  - bipartite graph with hidden layer h
  - h can capture high order dependencies among v
  - inference is efficient due to conditional independences
- DBM
  - learn more complex structure
- SBM
  - fewer parameters due to parameter sharing, quadrant structures

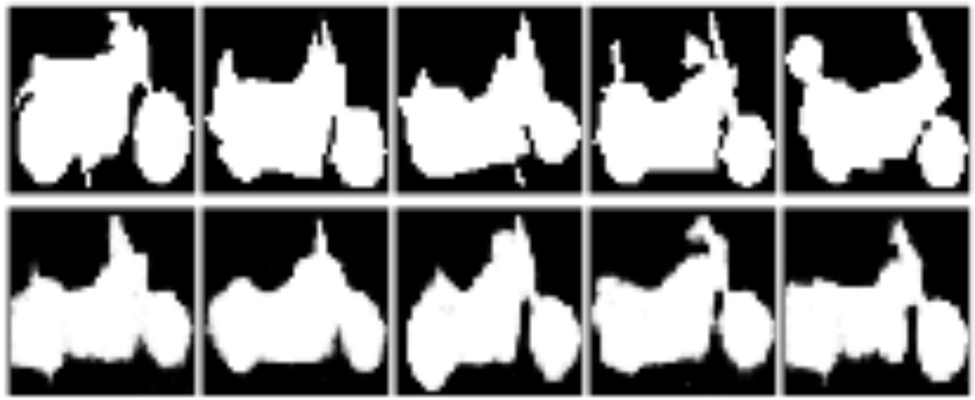
# Shape Boltzmann Machine [Eslami et al. 2012]



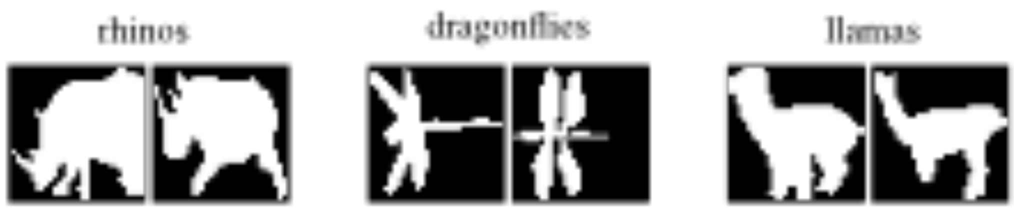
# Additional Examples

Original

SBM



Original




SBM

# Video of SBM

1. Show video of SBM

# Face Labeling

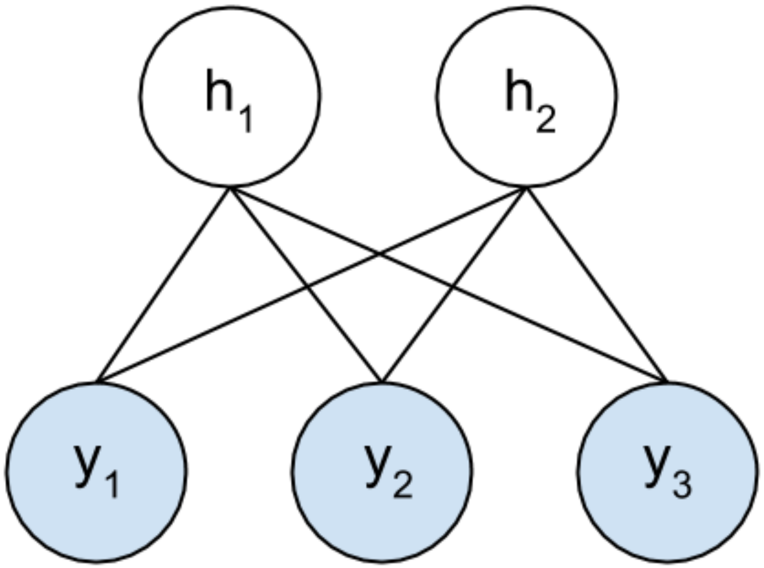
# RBM Shape Model

- Restricted Boltzmann Machine [Smolensky 1986]
  - multinomial visible units (L)
  - ~200 Hidden Units (K)
  - Labeled image 250 x 250 -> 24 x 24 (S)
  - trained with Contrastive Divergence

$$P(\mathbf{Y}, \mathbf{h}) = \frac{\exp(-E_{\text{rbm}}(\mathbf{Y}, \mathbf{h}))}{Z_{\text{rbm}}},$$

$$E_{\text{rbm}}(\mathbf{Y}, \mathbf{h}) = -\sum_{s=1}^{S}\sum_{l=1}^{L}\sum_{k=1}^{K} y_{sl} W_{slk} h_k - \sum_{k=1}^{K} b_k h_k - \sum_{s=1}^{S}\sum_{l=1}^{L} c_{sl} y_{sl},$$

# RBM Shape Model

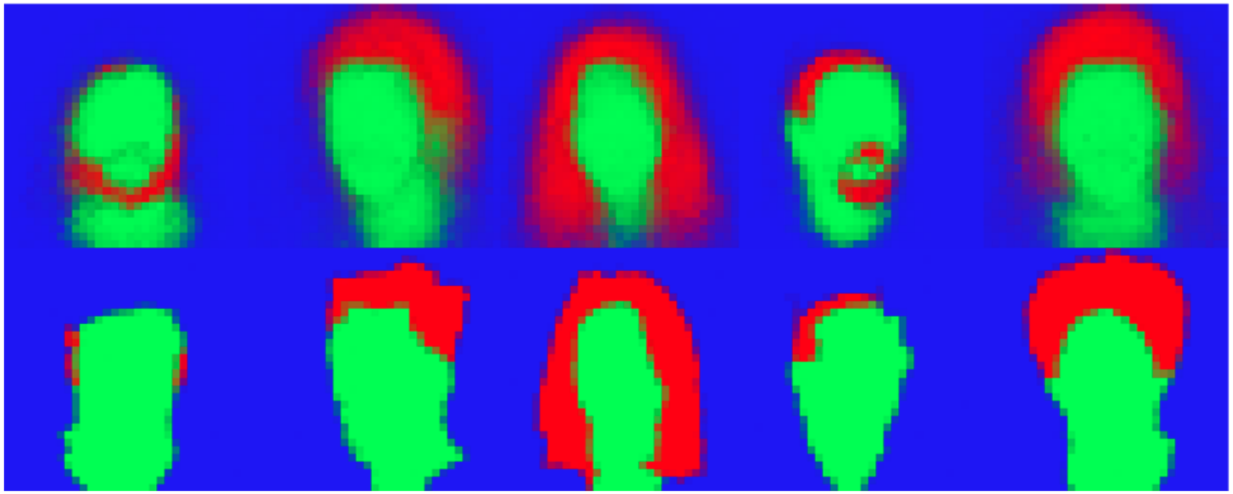



$$P(\mathbf{Y}, \mathbf{h}) = \frac{\exp(-E_{\text{rbm}}(\mathbf{Y}, \mathbf{h}))}{Z_{\text{rbm}}},$$

$$E_{\text{rbm}}(\mathbf{Y}, \mathbf{h}) = -\sum_{s=1}^{S}\sum_{l=1}^{L}\sum_{k=1}^{K} y_{sl} W_{slk} h_k - \sum_{k=1}^{K} b_k h_k - \sum_{s=1}^{S}\sum_{l=1}^{L} c_{sl} y_{sl},$$

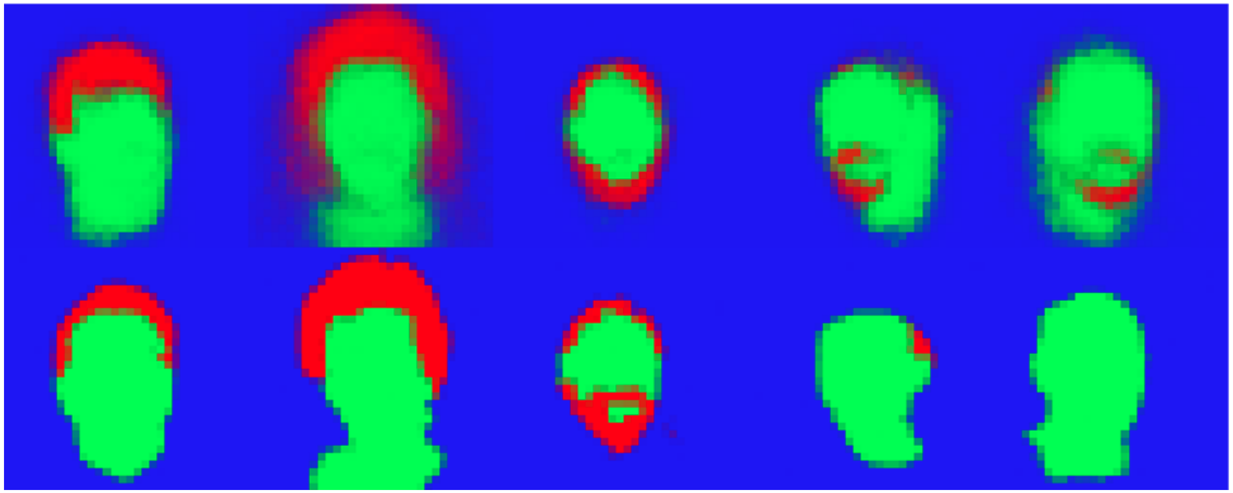
# Samples from RBM

**Samples**



**Closest Training Example**

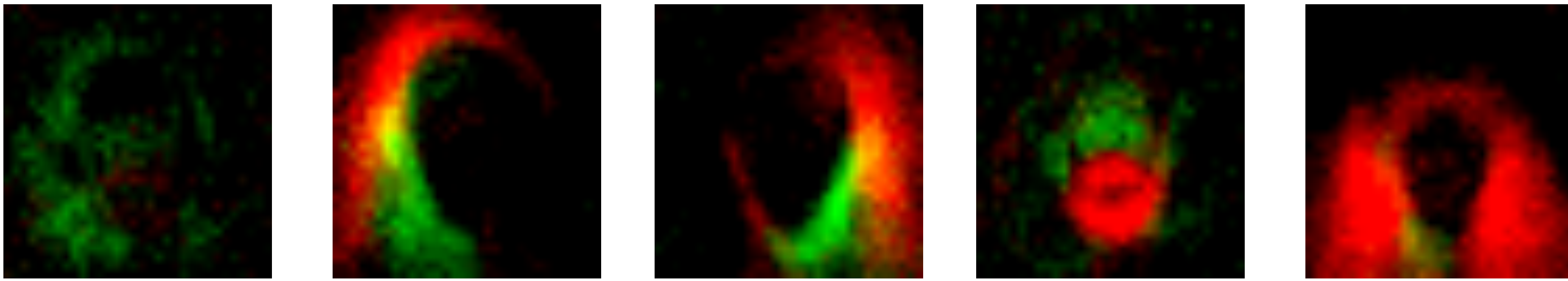
**Samples**



**Closest Training Example**

# RBM Hidden Units

No Hair | Looking Right | Looking Left | Beard | Big Hair



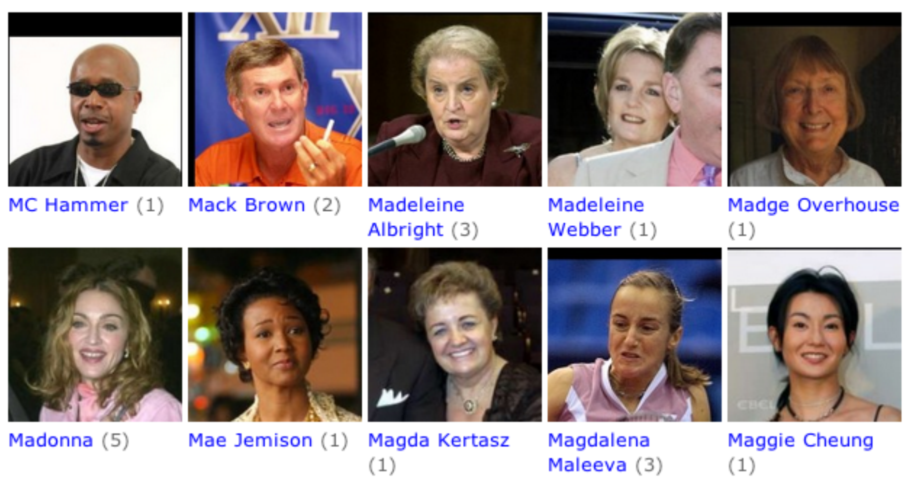
- Green : Skin, Red : Hair, Background : set to 0.
- RBM captures structure of face segmentations
- Some RBM hidden units can correspond to "attributes"

# Point to take home

- RBMs can learn the structure of simple object shapes

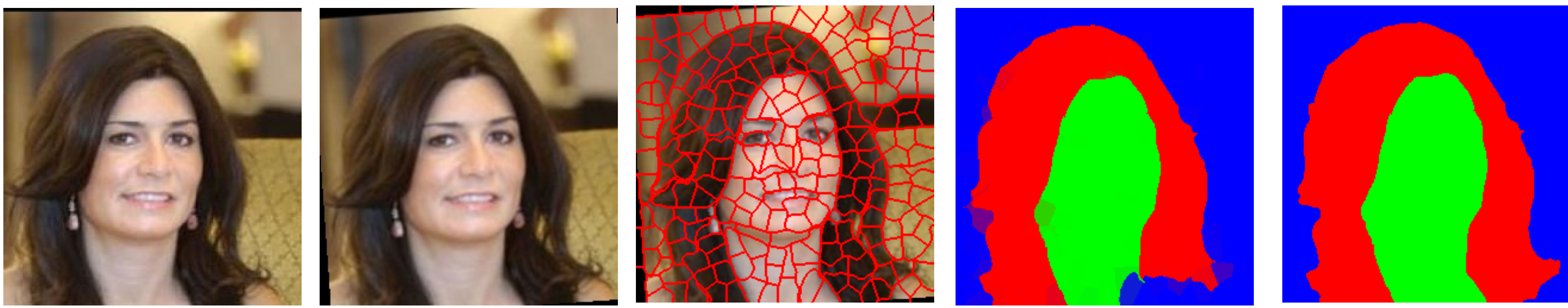
# Data

- Labeled Faces in the Wild (LFW)
  - 13,233 face images and their identities
  - taken from newswire (in the "wild") and automatically aligned
  - benchmark for face recognition
- Subset labeled for H/S/B
  - 2927 labeled images [http://vis-www.cs.umass.edu/lfw/part_labels/]



# Pipeline

LFW Image | Alignment | Segmentation | Model | Ground Truth



1. Perform automatic alignment [Huang et al. 2007]
2. Generate superpixels [http://www.cs.sfu.ca/~mori/research/superpixels/]
3. Generate features
4. Run GLOC model
5. Evaluate

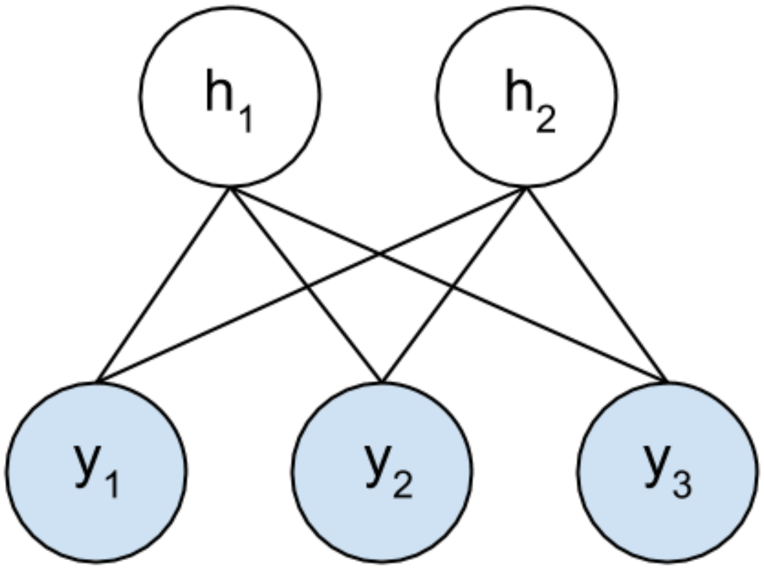
# Baseline

- CRF [Huang et al. 2008]
  - ~250 superpixels per image
  - Node features (128 dimensions)
    - Color : Normalized histogram over 64 bins generated by K-means over pixels in LAB space.
    - Texture : Normalized histogram over 64 textons.
    - Location : Normalized histogram of the proportion of a superpixel that falls within each of the 8 × 8 grid elements on the image.
  - Edge features (3 dimensions)
    - Sum of Pb (probability of boundary) values along border
    - Euclidean distance between mean color histograms
    - Chi-squared distance between texture histograms
  - Loopy BP inference
  - 93.23% superpixel accuracy

# Spatial CRF

- Small modification to CRF
- Node features may depend on position
  - N x N grid
- Initialize to CRF weights during training
- 93.95% superpixel accuracy
  - ~0.7% improvement over CRF

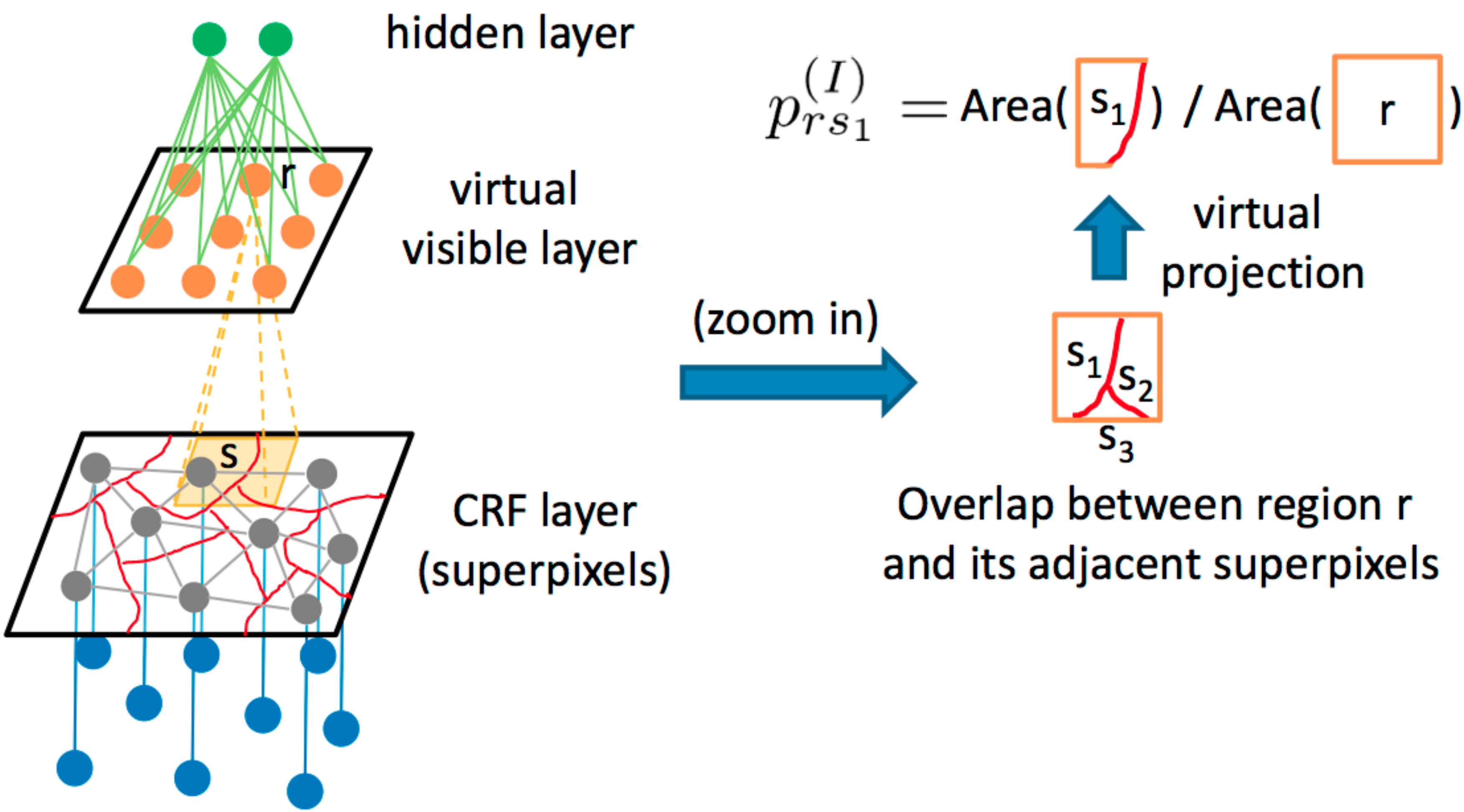
# RBM Shape Model




$$P(\mathbf{Y},\mathbf{h}) = \frac{\exp(-E_{\mathrm{rbm}}(\mathbf{Y},\mathbf{h}))}{Z_{\mathrm{rbm}}},$$

$$E_{\mathrm{rbm}}(\mathbf{Y},\mathbf{h}) = -\sum_{s=1}^{S}\sum_{l=1}^{L}\sum_{k=1}^{K} y_{sl}W_{slk}h_k - \sum_{k=1}^{K} b_k h_k - \sum_{s=1}^{S}\sum_{l=1}^{L} c_{sl}y_{sl},$$

# GLOC (Global + Local)



# GLOC (Global + Local)

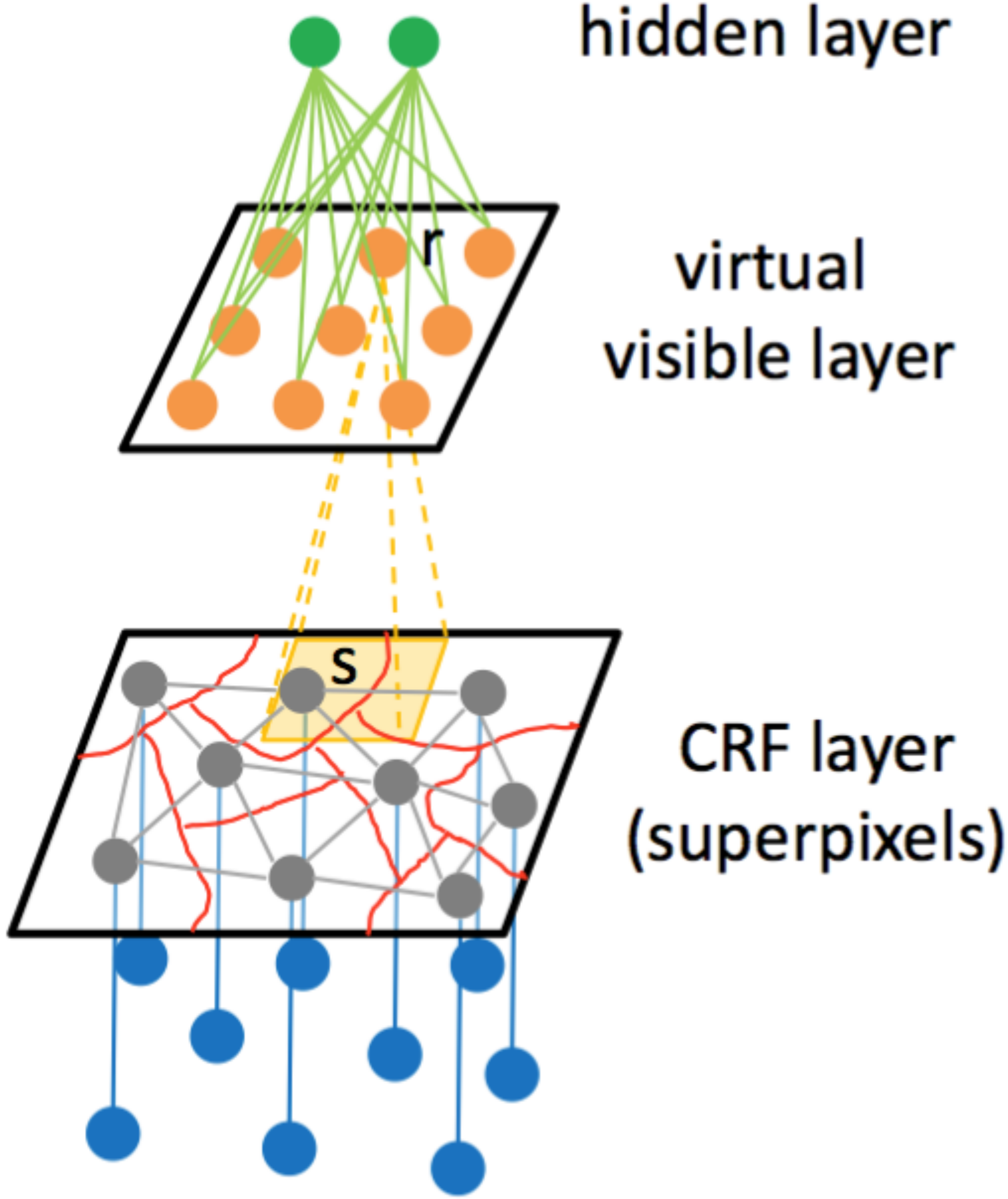
- Virtual visible layer computed deterministically from CRF labels
- Projection Matrix : Num Grid x Num SP
  - Rows sum to 1
- RBM Grid : 24x24
- CRF Grid : 16x16
- (slight complication) actually 2 projection matrices
  - RBM
  - CRF

# GLOC formulation

- X : visible
- Y : superpixel labels
- h : hidden units

$$P(\mathbf{Y}|\mathbf{X}) = \frac{1}{Z} \sum_{\mathbf{h}} \exp\left(-E(\mathbf{X}, \mathbf{Y}, \mathbf{h})\right)$$

$$E\left(\mathbf{X}, \mathbf{Y}, \mathbf{h}\right) = E_{\mathrm{crf}}\left(\mathbf{X}, \mathbf{Y}\right) + E_{\mathrm{rbm}}\left(\mathbf{Y}, \mathbf{h}\right)$$


hidden layer
r
virtual
visible layer
s
CRF layer
(superpixels)

# GLOC (RBM component)

- R : RBM Grid Dimension (24)
- S : Number of superpixels
- p : Projection Matrix between RBM Grid and superpixels

$$E_{\text{rbm}}(\mathbf{Y}, \mathbf{h}; I) = -\sum_{r=1}^{R^2} \sum_{l=1}^{L} \sum_{k=1}^{K} \bar{y}_{rl} W_{rlk} h_k - \sum_{k=1}^{K} b_k h_k - \sum_{r=1}^{R^2} \sum_{l=1}^{L} c_{rl} \bar{y}_{rl}$$

$$\bar{y}_{rl} = \sum_{s=1}^{S} p_{rs} y_{sl}$$

# GLOC (CRF component)

- N : CRF Grid Dimension (16)
- q : Projection Matrix between CRF Grid and superpixels

$$E_{\mathrm{crf}}(\mathbf{Y},\mathbf{X}) = E_{\mathrm{node}}(\mathbf{X},\mathbf{Y}) + E_{\mathrm{edge}}(\mathbf{X},\mathbf{Y})$$

$$E_{\mathrm{node}}(\mathbf{X},\mathbf{Y}) = -\sum_{n=1}^{N^2}\sum_{s=1}^{S} q_{sn} \sum_{l=1}^{L}\sum_{d=1}^{D_n} x_{sd}y_{sl}\Gamma_{ndl}$$

$$E_{\mathrm{edge}}(\mathbf{X},\mathbf{Y}) = -\sum_{(i,j)\in\mathcal{E}}\sum_{l,l'=1}^{L}\sum_{e=1}^{D_e} y_{il}y_{jl'}\Psi_{ll'e}x_{ije}$$

# Inference

- Exact inference of $P(Y|X)$ is intractable
- Approximate $P(Y|X)$ by alternating between manageable $P(Y|X,H)$ and $P(H|Y)$
- Sample $P(H|Y)$
  - $$P\left(h_k = 1|\mathbf{Y}\right) = \sigma\left\{\sum_{r=1}^{R^2}\sum_{l=1}^{L}\bar{y}_{rl}w_{rlk} + b_k\right\}$$
- Sample $P(Y|X,H)$ using mean-field
  - RBM augments node potential of CRF

$$\sum_{s=1}^{S}\sum_{l=1}^{L} y_{sl}\left(-\sum_{r=1}^{R^2} p_{rs}\sum_{k=1}^{K} w_{rlk}h_k - \sum_{n=1}^{N^2} q_{sn}\sum_{d=1}^{D}\Gamma_{ndl}x_{sd}\right)$$

# Learning

- Train model parameters : $\{\mathbf{W}, \mathbf{b}, \mathbf{C}, \Gamma, \Psi\}$
- Piecewise
  - scalar parameter $\lambda$ weights the contribution of RBM component during CRF inference
  - pretrain RBM, CRF
  - $\lambda$ learned through validation (~0.1 works well)
- Joint
  - Contrastive Divergence
  - CD-PercLoss [Mnih et al. 2011]
    - alternative to Contrastive Divergence
    - may be better suited for a Conditional RBM
  - Sequence of pre-training steps
    - pre-train weights for CRF, CRBM, then all weights together.

# Evaluation

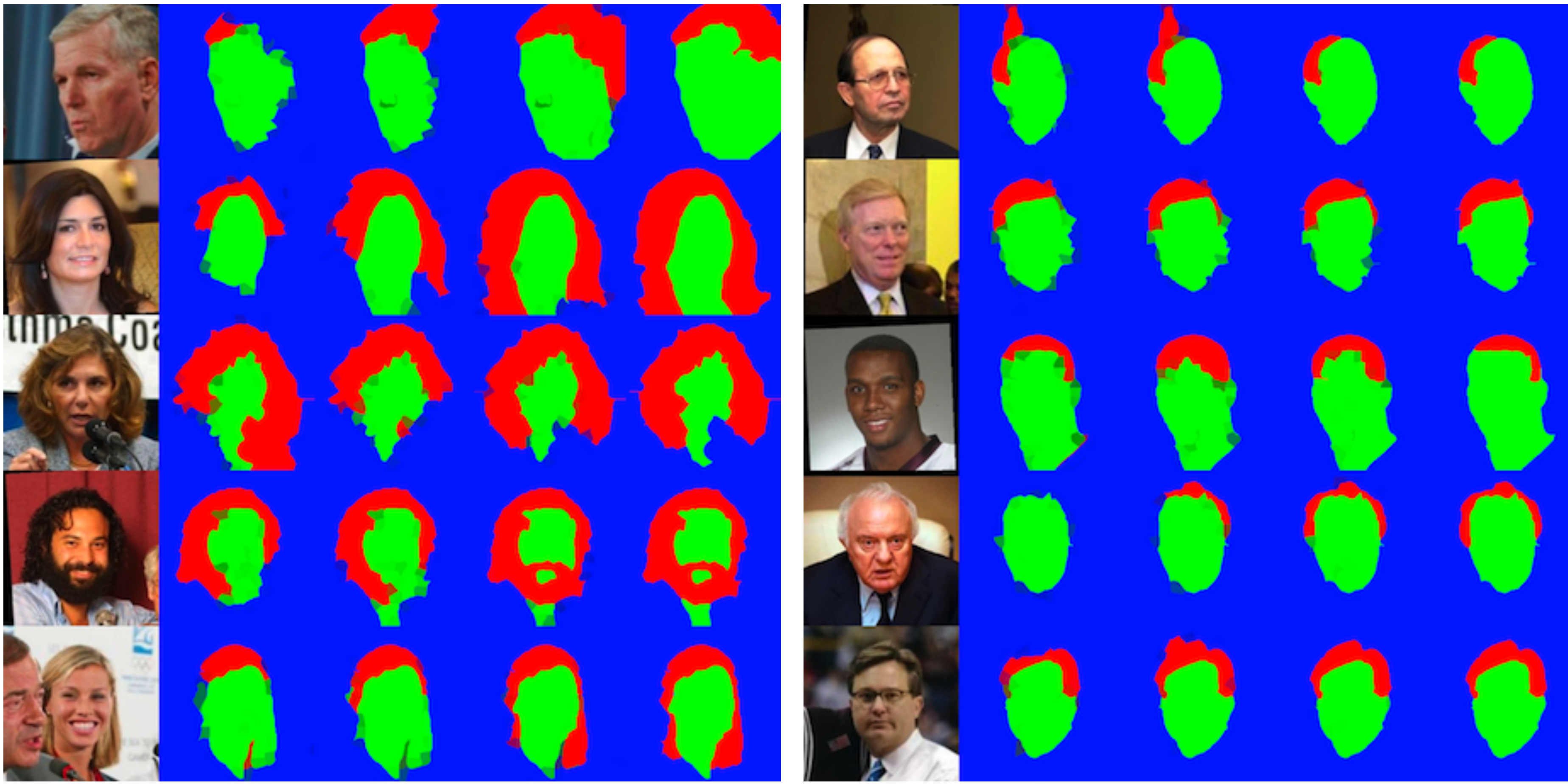
# Evaluation

| Approach | Superpixel Accuracy | Error Reduction over CRF |
|---|---|---|
| CRF | 93.23% | 0% |
| Spatial CRF | 93.95% | 10.64% |
| CRBM | 94.10% | 12.85% |
| GLOC (piecewise) | 94.34% | 16.40% |
| GLOC (joint) | **94.95%** | **25.41%** |

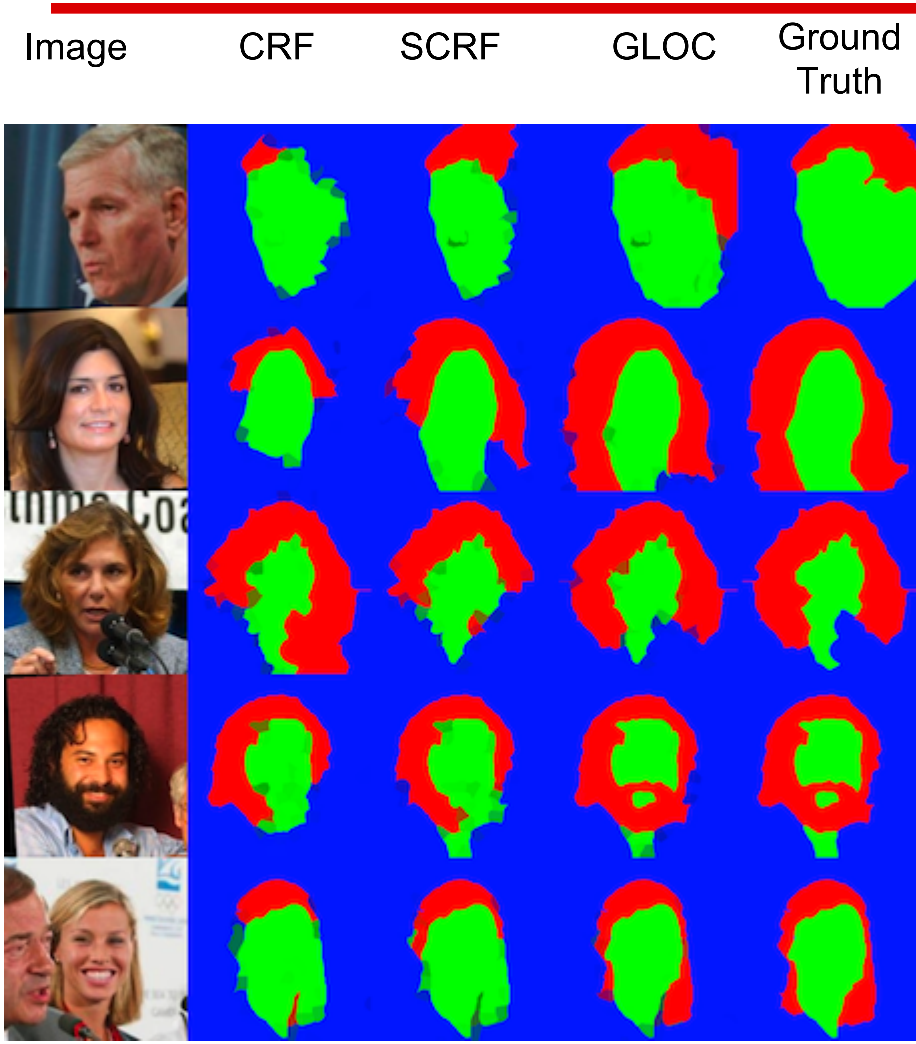
- 1500 training / 500 validation / 927 test
- Improvement over SCRF is small
  - subtle improvement
  - we believe it is significant

# Successful Examples

Image CRF SCRF GLOC Ground Truth Image CRF SCRF GLOC Ground Truth

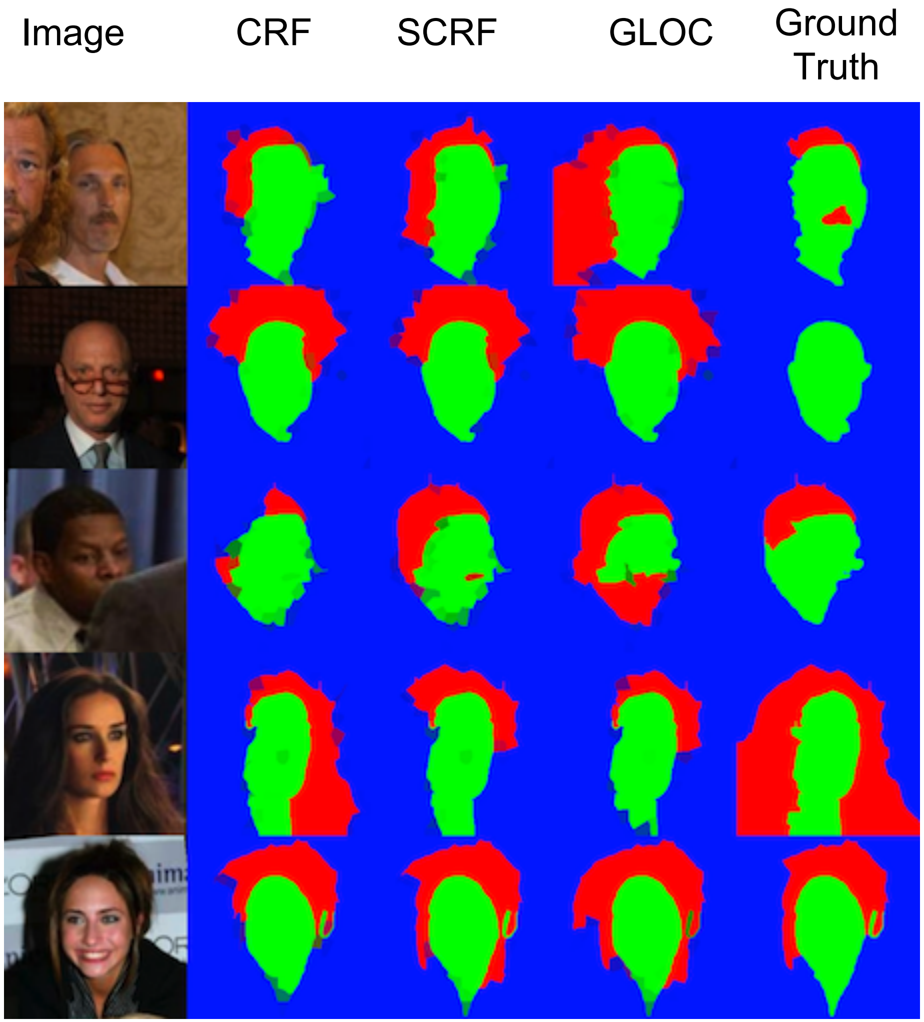


# Successful Examples




- Encourages a more realistic labeling by filling in or removing parts of hair/skin.
- More robust to multiple faces in close proximity.

# Unsuccessful Examples




- Heavy occlusion
- Background matches hair color
- Disparity in hair color
- Shape prior perhaps too strong

# Point to Take Home

- Can improve local modeling of CRF by using the RBM as a global shape prior
  - GLObal + LOCal = GLOC modeling

# Retrieval

- We can interpret some hidden units as attributes
- Run GLOC inference for all LFW images (except training set), rank the images in terms of hidden unit activations
- Obtain meaningful clusters

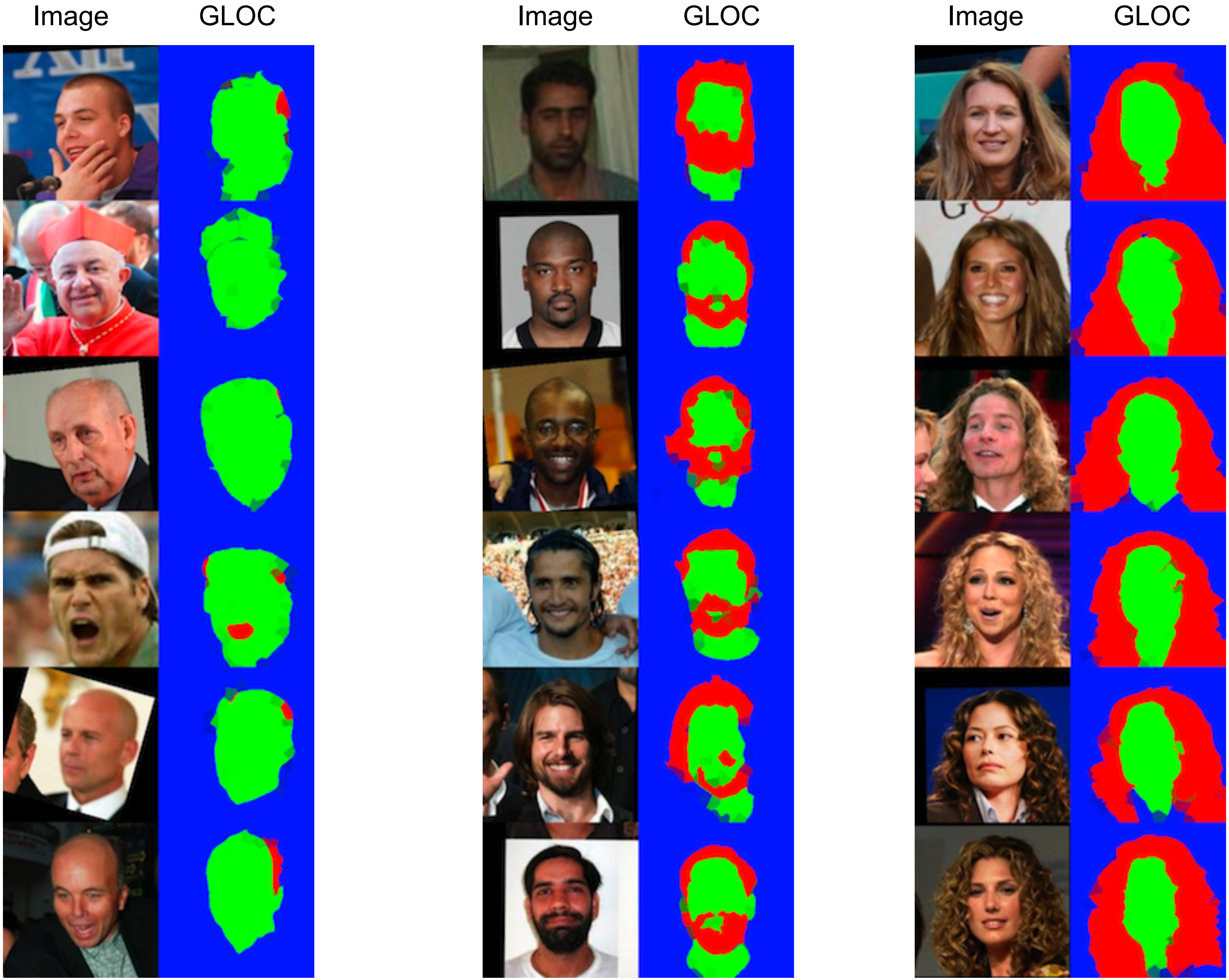


Image
GLOC
Image
GLOC
Image
GLOC

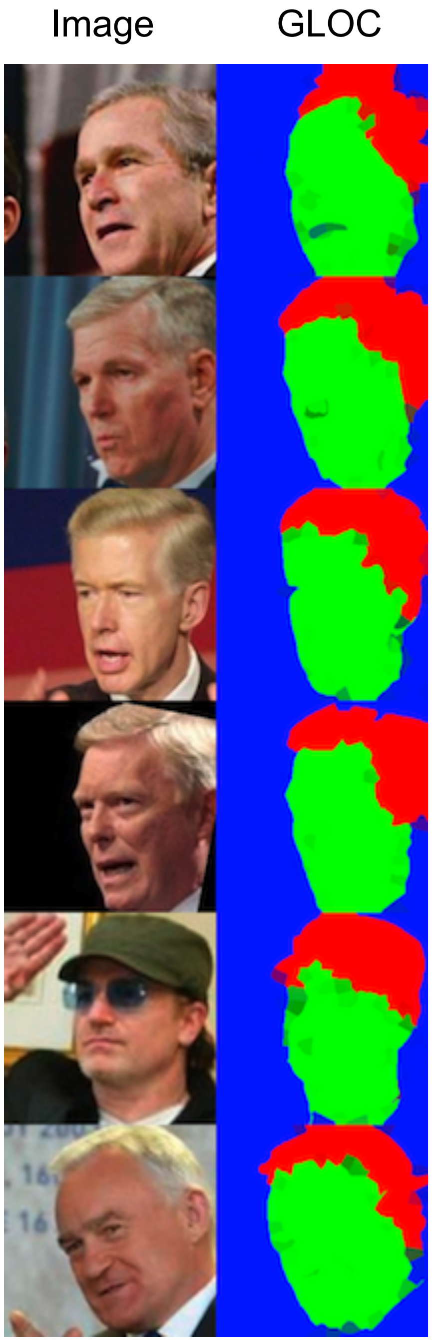


Image
GLOC

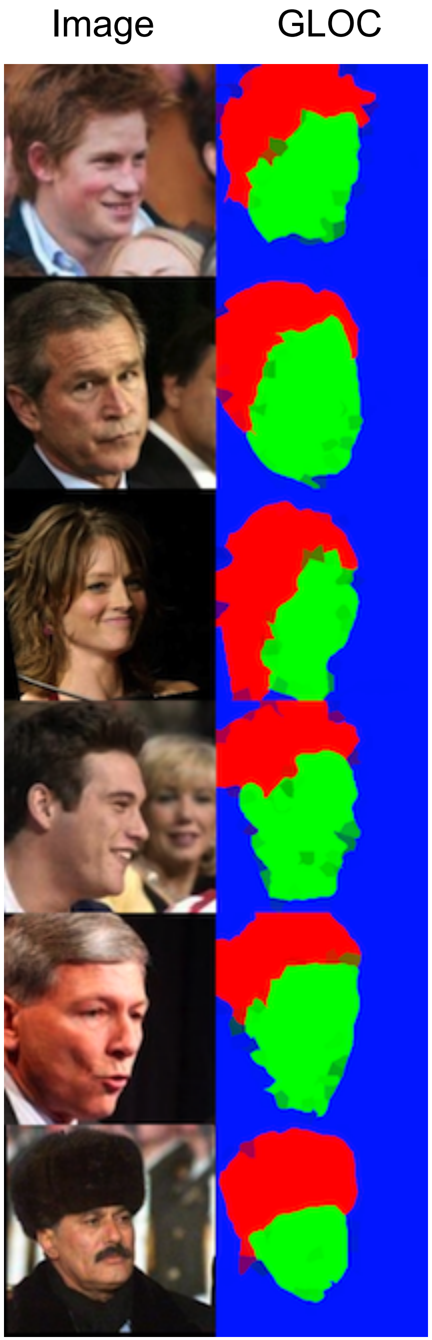


Image
GLOC

# Point to Take Home

- Can interpret RBM hidden units as attributes
    - Obtain meaningful clusters when the GLOC model is used to rank through hidden unit activations

# Practical Challenges

- Multiple hyperparameters
  - both RBM and CRF
  - number of hidden units, learning rate, regularization, number of CD steps
- Joint training
  - pre-training important
- Training time
  - about 1 day

# Points to take home

- RBMs can learn the structure of simple object shapes
- Can improve local modeling of CRF by using the RBM as a global shape prior
  - GLObal + LOCal = GLOC modeling
- Can interpret RBM hidden units as attributes
  - Obtain meaningful clusters when the GLOC model is used to rank through hidden unit activations

# Thank you!

Questions?

# Appendix

# Image Labeling

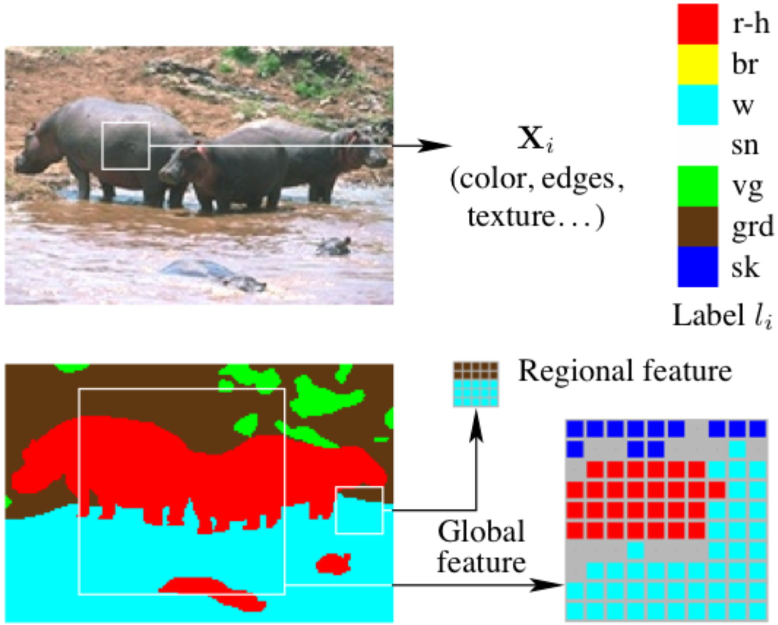
- Multiscale CRF [He et al. 2004]
  - natural scenes
- Face labeling [Wang et al. 2012]
  - closest related work in problem domain
- Boltzmann machine prior [Eslami et al. 2012]
  - ShapeBM (similar object shape prior)

# Multiscale CRF [He et al. 2004]

- Natural scenes (labels such as bear, water, sky)
- RBM at multiple scales, combined with models at local and regional scales multiplicatively.

<u>Drawbacks</u>

- no edge potentials
- pixel representation
- computation time (from pixel representation)



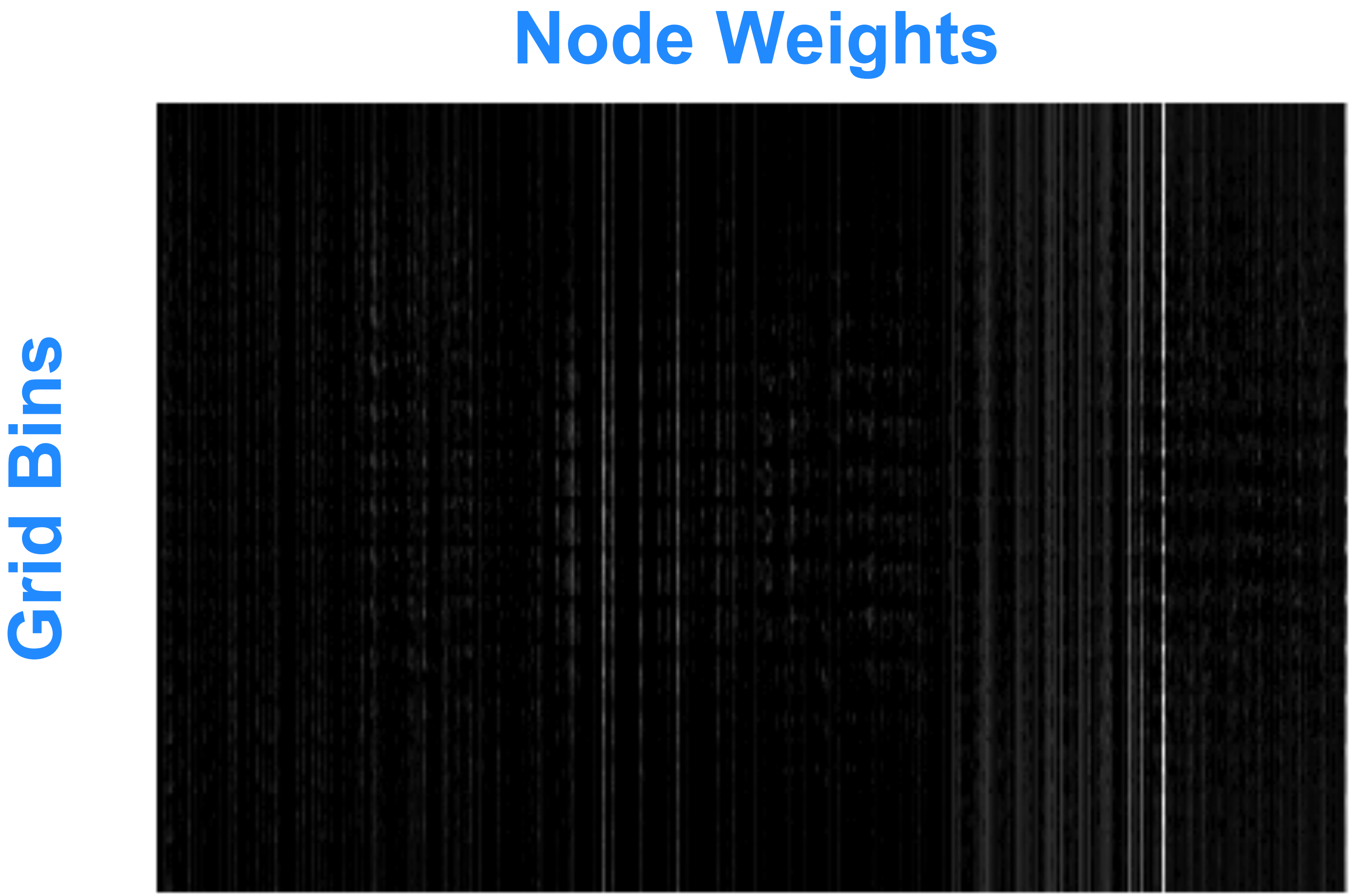
# Face Labeling [Wang et al. 2012]

- Hair/Skin/Background/Clothing
- Models a configuration of local hair parts

<u>Drawbacks</u>

- Lacks global shape model

# SCRF weights



# Ongoing Work

- Occlusion
- Better representation
  - inherent error in superpixels
- Better retrieval
- Finer grained labeling (parts of face)
- More structure (DBM or SBM)

# ShapeBM for Labeling [Eslami et al. 2012]

- Use ShapeBM within a parts-based generative model
- Label images of pedestrians, cars (competitive but not state-of-the-art)

Drawbacks

- No local modeling
- Modeled over pixels

# Face Labeling [Wang et al. 2012]

- Hair/Skin/Background/Clothing
- Models a configuration of local parts
- 90% reported accuracy, 90.7% GLOC (~3% superpixel error)

Drawbacks

- Lacks global shape model