LOCUS: Learning Object Classes with Unsupervised Segmentation

16-721: Advanced Perception

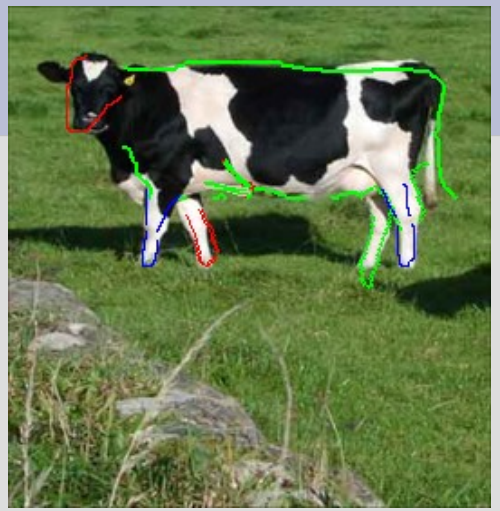
Nik Melchior

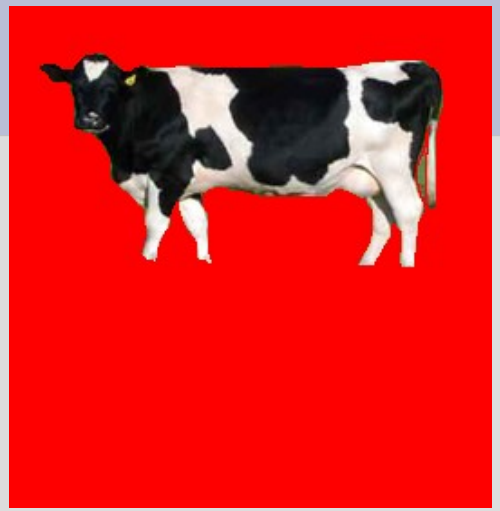
April 10, 2006

Outline

- Learning Flexible Sprites in Video Layers (Jojic & Frey 2001)
- LOCUS

Motivation





Object category recognition and segmentation in cluttered scenes

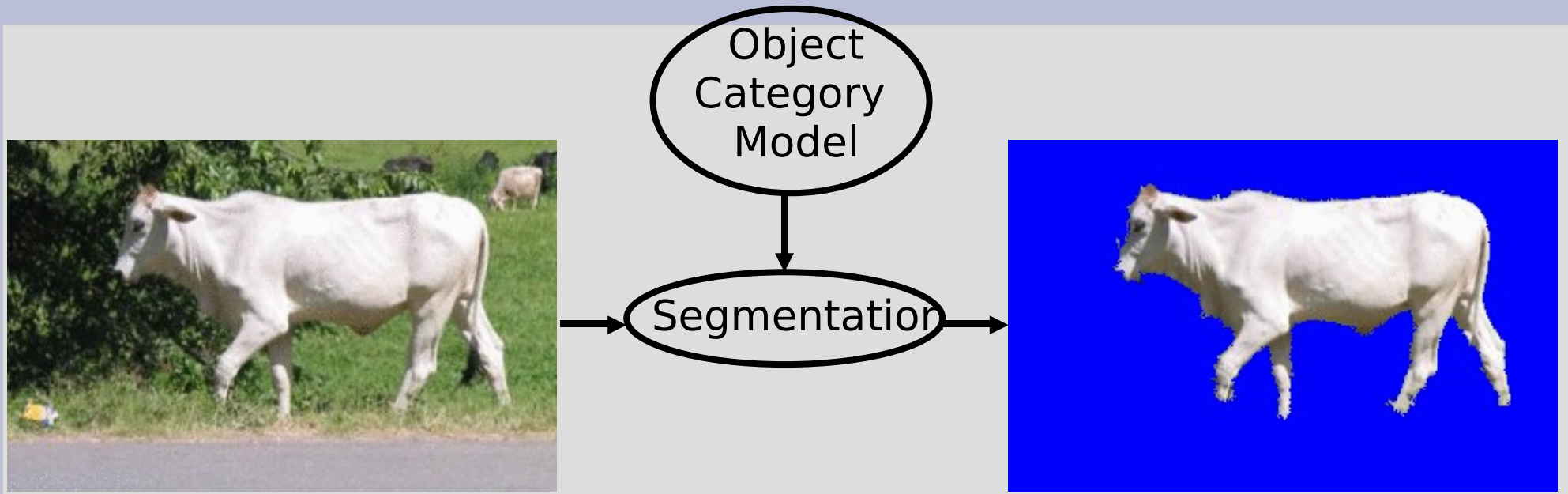
Challenges

- ✓ Spatial variability
- ✓ Texture variability
- ✓ Occlusion
- ✓ Pose
- ✓ Lighting

slide: Kumar, Torr, Zisserman

Motivation

- Given an image and object category, to segment the object



Cow Image

Segmented Cow

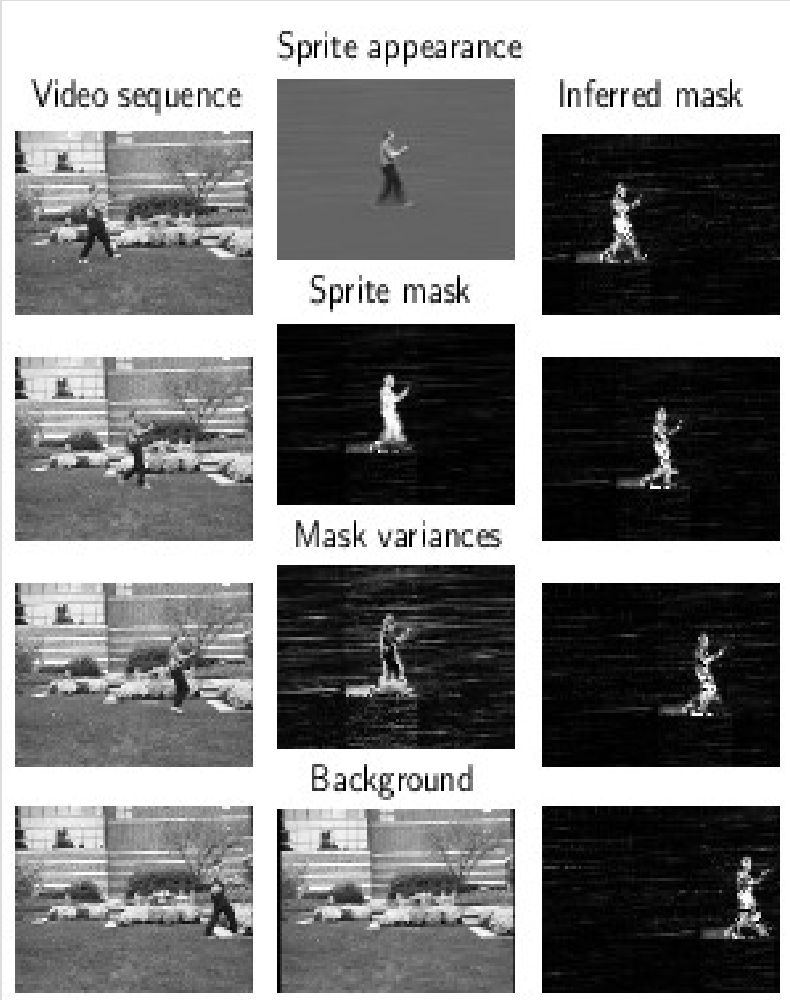
Segmentation should (ideally) be

- shaped like the object e.g. cow-like
- obtained efficiently in an unsupervised manner
- able to handle self-occlusion

slide: Fei-Fei

Flexible Sprites

- Provides a stable segmentation of whole moving objects in video
- Each flexible sprite is a separate layer containing the transformation of a rigid sprite



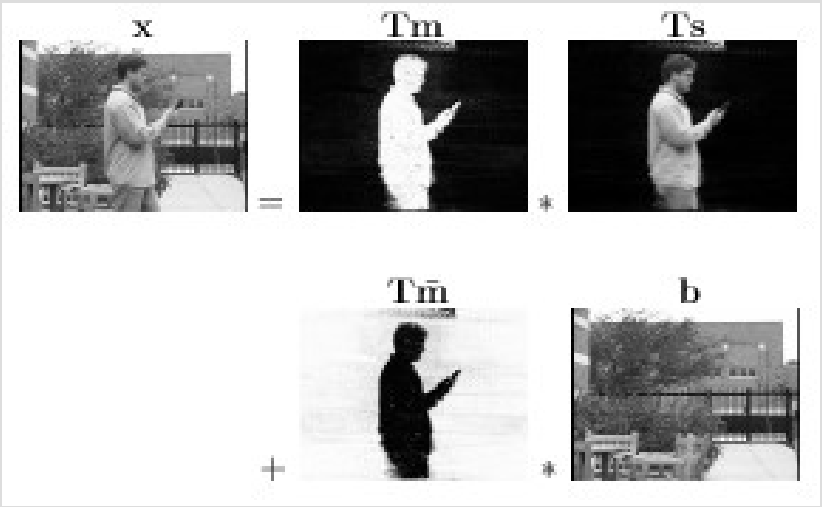
Big Picture

- Generative model
- Seeks the best explanation for the pixels of all images, constrained by number of layers
- Operates directly on pixels rather than low-level features
- Rigorous formulation similar to DDMCMC

Flexible Sprites

- User specifies number of layers
- Each layer contains:
  - Sprite appearance `'s'`
  - Real-valued mask `'m'`
  - Discretized simple transformation `'T'` “permutation matrix that rearranges the pixels in s”

$$x = Tm * Ts + T\bar{m} * b + noise$$



Fo

Appearance, mask, and transform

Recursive formula for image appearance

Zero-mean Gaussian noise

$$p(x|\{s_l, m_l, T_l\}) = N\left(x; \sum_L \left(\left(\prod_L T_i \bar{m}_i\right) * T_l m_l * T_l s_l\right), \beta\right)$$

- Assuming independence of each class, appearance, mask, and transformation ac

Transformation prior (uniform)

Appearance

Mask

Class prior (to be learned)

$$p(x, \{s_l, m_l, T_l\}) = N\left(x; \sum_L \left(\left(\prod_L T_i \bar{m}_i\right) * T_l m_l * T_l s_l\right), \beta\right)$$
$$* \prod_L \left(N(s_l; \mu_{c_l}, \phi_{c_l}) N(m_l; \eta_{c_l}, \psi_{c_l}) \pi_{c_l} \rho_{T_l}\right)$$

Formulation

- Use EM to maximize the posterior

$$p\left(\{c_l, T_l, s_l, m_l\}|X\right)=\frac{p\left(x,\{c_l, T_l, s_l, m_l\}\right)}{\sum\int p\left(x,\{c_l, T_l, s_l, m_l\}\right)}$$

$$\approx\prod_L q(c_l, T_l)\,q(s_l)\,q(m_l)$$

$$D=\sum\int\left(\prod_L q(c_l, T_l)\,q(s_l)\,q(m_l)\right)=\ln\frac{p\left(\{c_l, T_l, s_l, m_l\}|x\right)}{\prod_L q(c_l, T_l)\,q(s_l)\,q(m_l)}$$

Formulation

$$F = D + \ln p(x)$$

$$= \sum \int \left(\prod_L q(c_l, T_l)\, q(s_l)\, q(m_l) \right) = \ln \frac{p\left(x, \{c_l, T_l, s_l, m_l\}\right)}{\prod_L q(c_l, T_l)\, q(s_l)\, q(m_l)}$$

- Maximize F
  - log of numerator is a sum of Mahalanobis distances
  - q ln 1/q is the entropy of q

E step

- For each single image:
  - the mean should stay close to the overall mean, but large where image does not match background
  - variance should be high where the transformed image is similar to the background

$$\boldsymbol{\nu} \leftarrow \left(\boldsymbol{\psi}^{-1} + \sum_{\mathbf{T}} \xi_{\mathbf{T}} \mathbf{T}^{-1} \left(\boldsymbol{\beta}^{-1} * (\mathbf{T}\boldsymbol{\mu} - \mathbf{b})\right)\right)^{-1}$$

and

$$\boldsymbol{\gamma} \leftarrow \boldsymbol{\nu} * \left(\boldsymbol{\psi}^{-1}\boldsymbol{\eta} + \sum_{\mathbf{T}} \xi_{\mathbf{T}} \mathbf{T}^{-1} \left(\boldsymbol{\beta}^{-1} * (\mathbf{T}\boldsymbol{\mu} - \mathbf{b}) * (\mathbf{x} - \mathbf{b})\right)\right),$$

M step

$$\psi \leftarrow \frac{1}{N}\sum_n \left(\boldsymbol{\nu}^{(n)} + (\boldsymbol{\gamma}^{(n)} - \boldsymbol{\eta})^2\right),$$

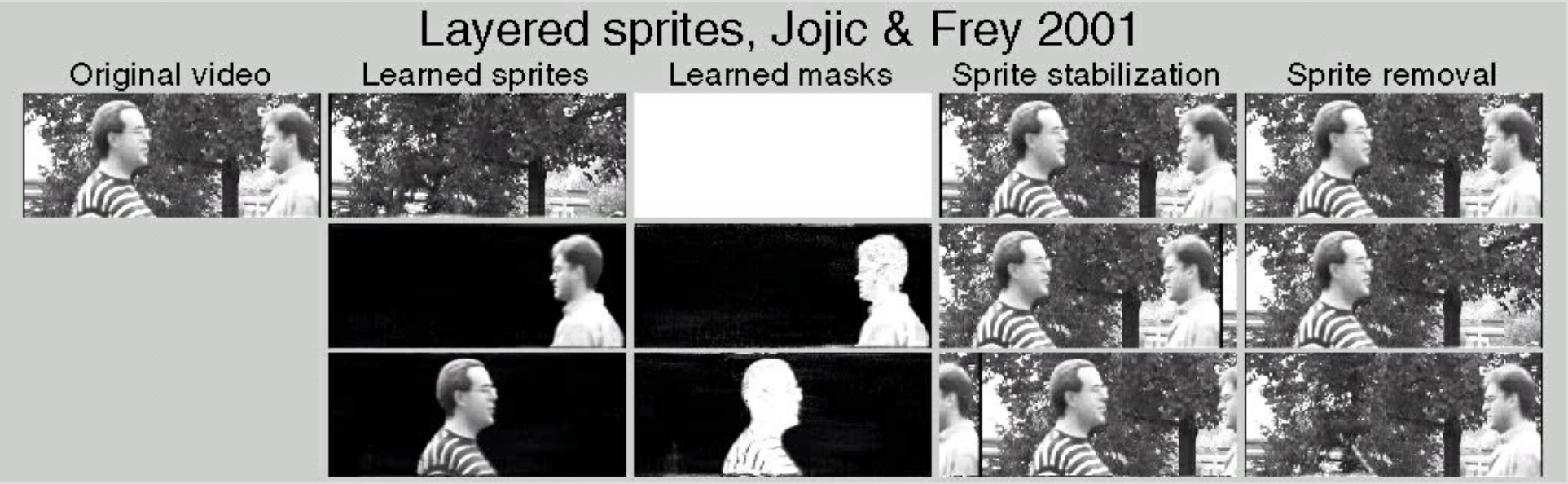
$$\boldsymbol{\eta} \leftarrow \frac{1}{N}\sum_n \boldsymbol{\gamma}^{(n)},$$

$$\boldsymbol{\mu} \leftarrow \left(\sum_n \sum_{\mathbf{T}} \xi_{\mathbf{T}}^{(n)} \left(\boldsymbol{\gamma}^{(n)^2} + \boldsymbol{\nu}^{(n)}\right)\right)^{-1} * \left(\sum_n \sum_{\mathbf{T}} \xi_{\mathbf{T}}^{(n)} \cdot \left(\boldsymbol{\gamma}^{(n)} * (\mathbf{T}^{-1}\mathbf{x}^{(n)} - \bar{\boldsymbol{\gamma}}^{(n)} * \mathbf{T}^{-1}\mathbf{b}) + \boldsymbol{\nu}^{(n)} * \mathbf{T}^{-1}\mathbf{b}\right)\right).$$

$$\mathbf{b} \leftarrow \left(\sum_n \sum_{\mathbf{T}} \xi_{\mathbf{T}}^{(n)} \mathbf{T}(\bar{\boldsymbol{\gamma}}^{(n)} + \boldsymbol{\nu}^{(n)})\right)^{-1} * \left(\sum_n \sum_{\mathbf{T}} \xi_{\mathbf{T}}^{(n)} \cdot \left(\mathbf{T}\bar{\boldsymbol{\gamma}}^{(n)} * (\mathbf{x}^{(n)} - \mathbf{T}\boldsymbol{\gamma}^{(n)} * \mathbf{T}\boldsymbol{\mu}) + \mathbf{T}\boldsymbol{\nu}^{(n)} * \mathbf{T}\boldsymbol{\mu})\right)\right).$$

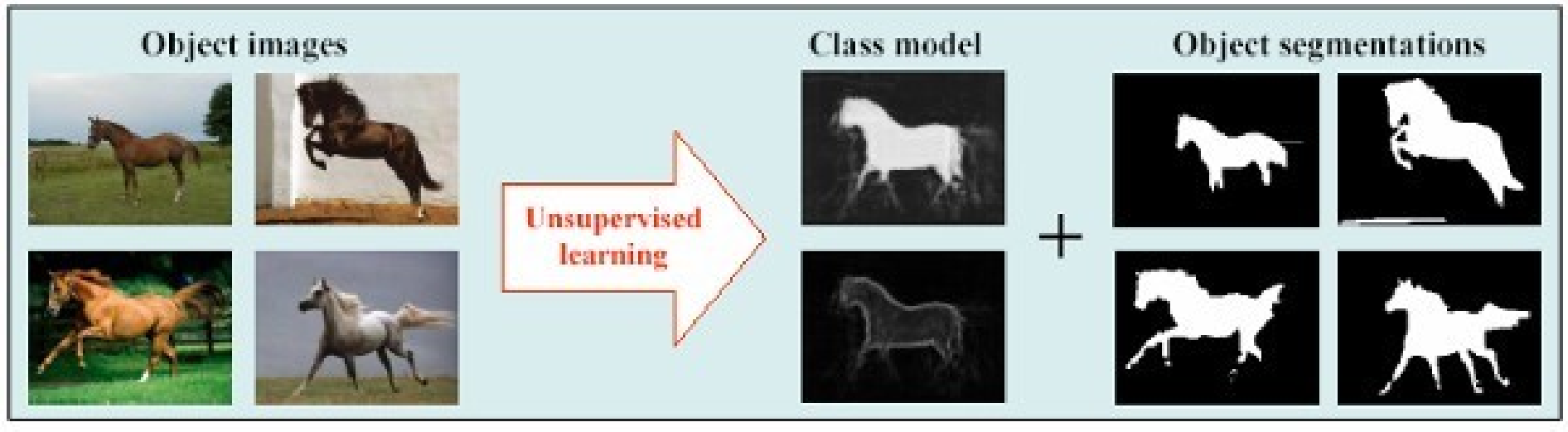
Results

- 1 fps at 320x240



LOCUS

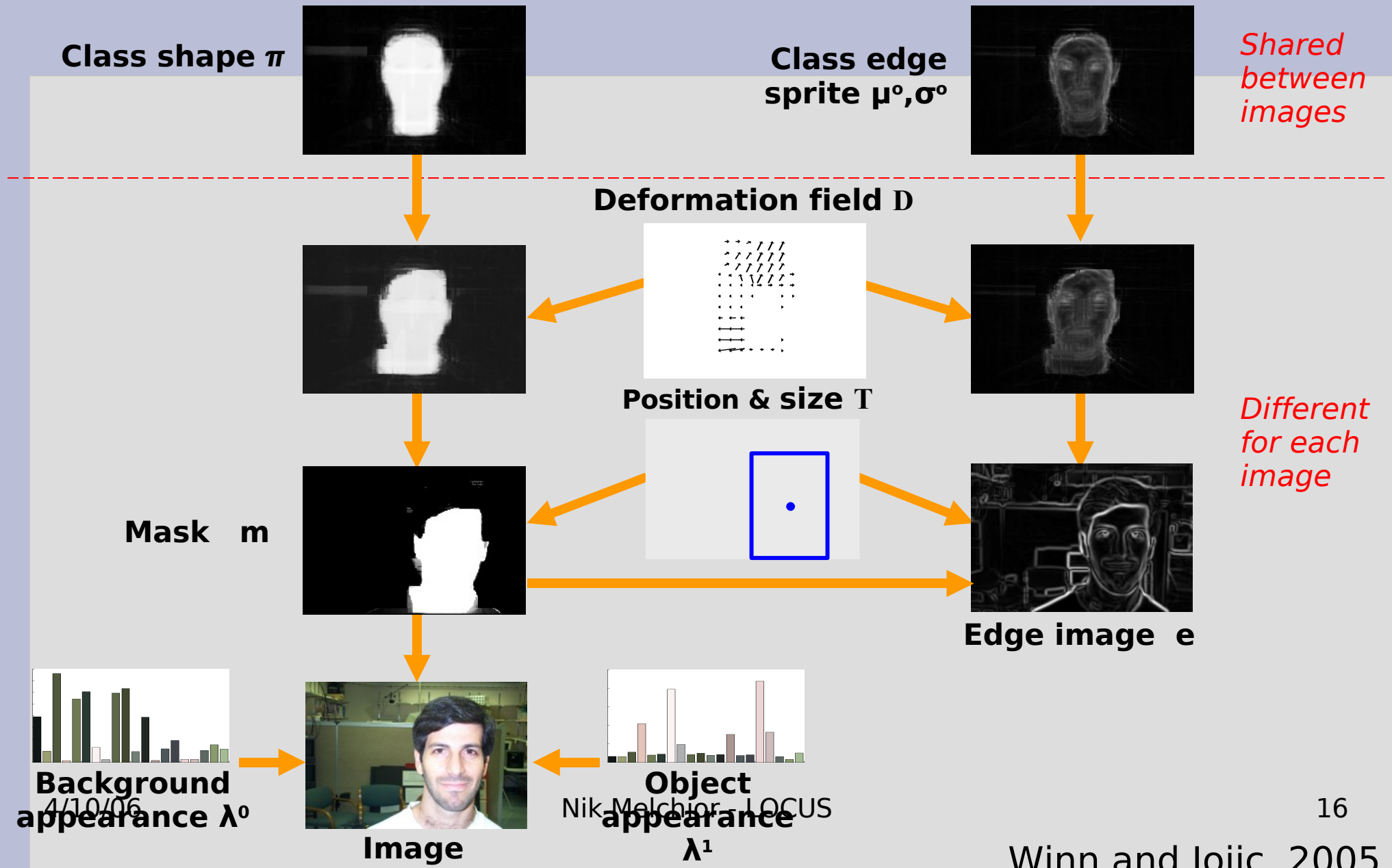
- Learning Object Classes with Unsupervised Segmentation



Comparison to Flexible Sprites

- Classes instead of single sprites
  - more complex appearance model
  - addition of edge model
- 2 layers: object and model
- weak prior of palettes (appearance)
  - one for object, one for background

LOCUS model



Winn and Jojic, 2005

LOCUS model

- Deformation field D
  - Markov Random Field
- Transform T
  - scaling and translation
  - rotation or full affine transform possible
- Mask m
- Edge model e
- Appearance model
  - Gaussian mixture model (histogram of colors or textures)

$$P(\mathbf{D}) = \frac{1}{Z} \exp \sum_{(i,j)\in\bar{E}} -\alpha |\mathbf{d}_i - \mathbf{d}_j|^2$$

$$P_{\mathrm{mrf}}(\mathbf{m} \mid \gamma, \beta) = \frac{1}{Z} \exp \sum_{(i,j)\in\bar{E}} -\delta(m_i \neq m_j) \gamma e^{-\beta ||\mathbf{z}_i - \mathbf{z}_j||^2}$$

$$P(\mathbf{z}_i \mid m_i, \boldsymbol{\theta}) = \sum_{k=1}^{K} \lambda_k^{m_i} \mathcal{N}(\mathbf{z}_i \mid \eta_k, \Sigma_k).$$

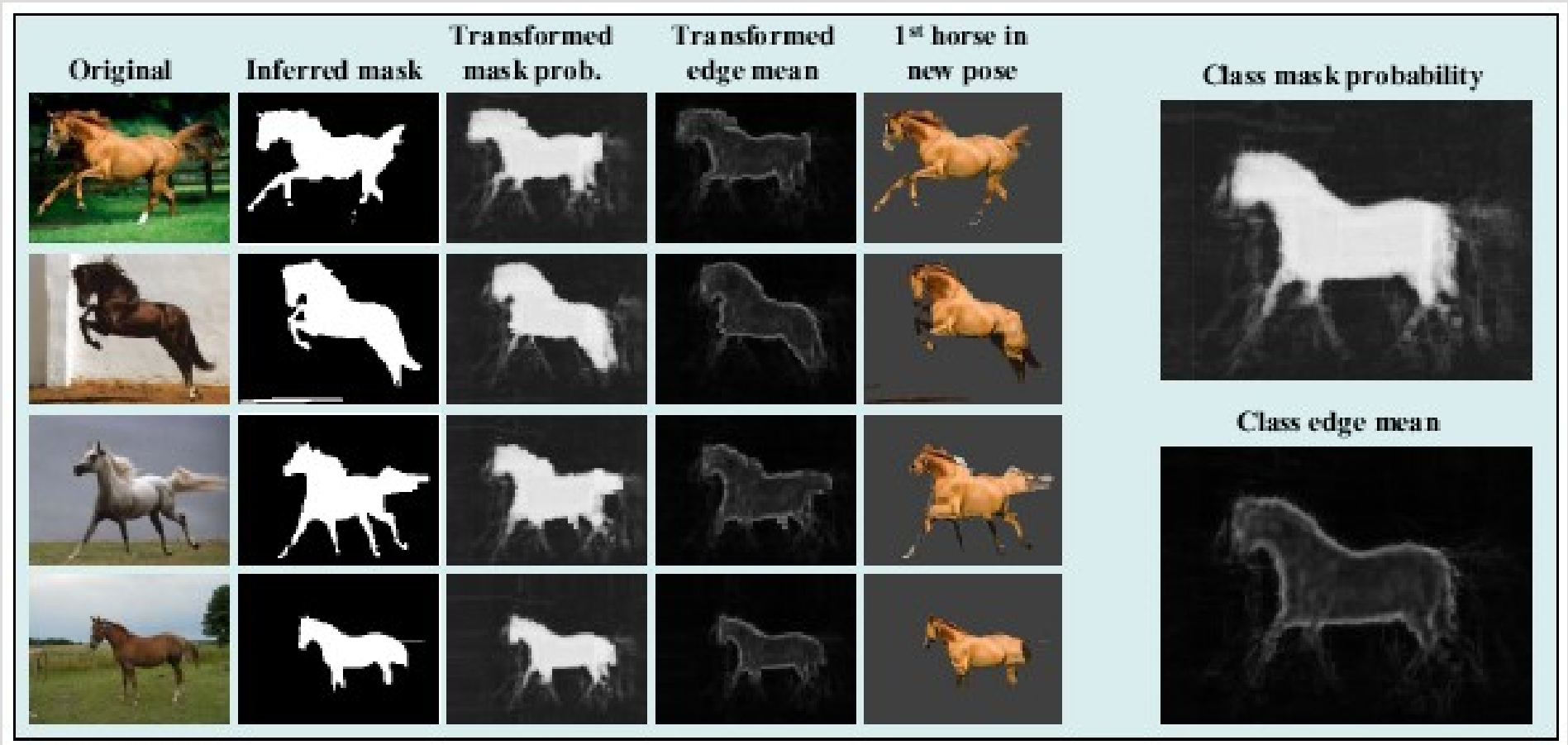
EM formulation

$$P(\pi, \mu^o, \sigma^o, \mu^b, \sigma^b | I) \approx \prod_i Q(\pi_i) Q(\mu_i^o, \sigma_i^o) Q(\mu_i^b, \sigma_i^b)$$

- Posterior similar to flexible sprites
- Iteratively optimize over 9 latent variables

$$\{\lambda, m, T, D, \pi, \mu^o, \sigma^o, \mu^b, \sigma^b\}$$

Results

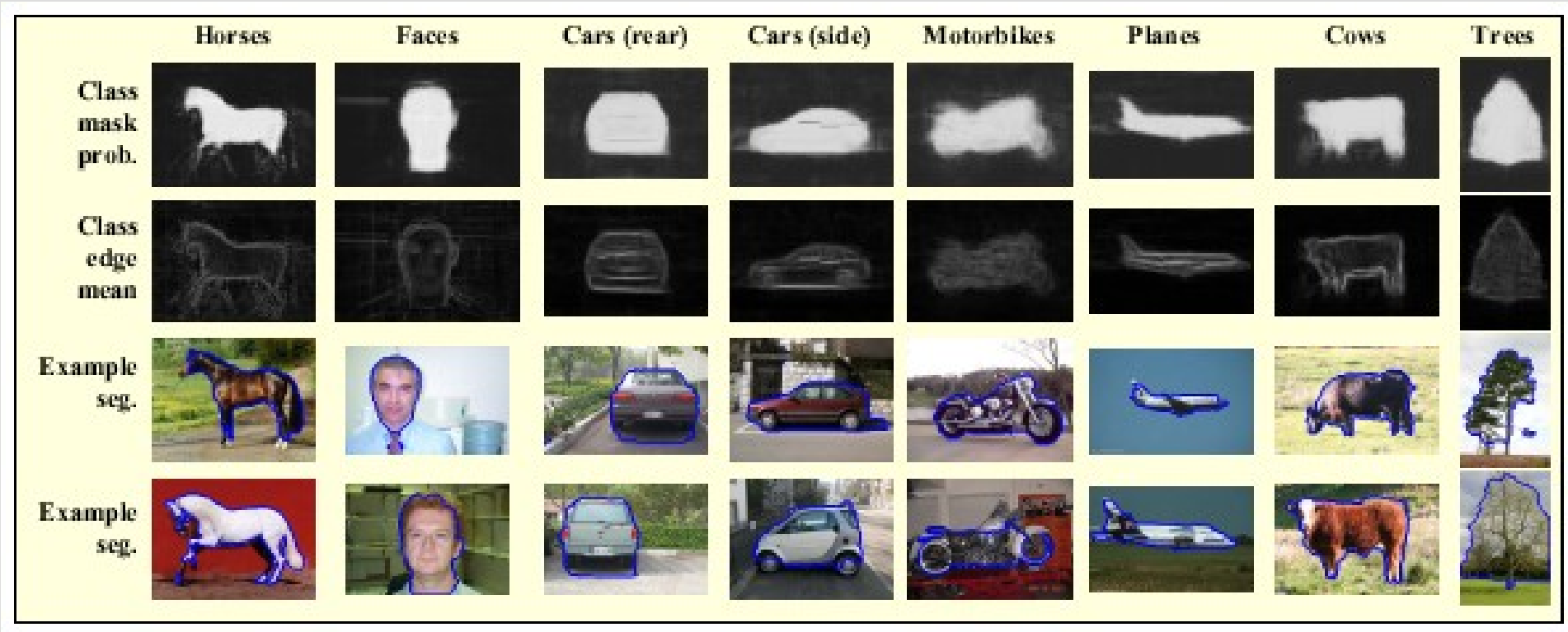

Original
Inferred mask
Transformed mask prob.
Transformed edge mean
1st horse in new pose
Class mask probability
Class edge mean

Results

- Comparison to Borenstein fragments
- Percentages indicate correctly labeled pixels across all images
- Last row removes class shape and edge information

	Segmentation accuracy	
	Horses	Cars (side)
LOCUS (color)	93.1%	91.4%
LOCUS (texture)	**93.0%**	**94.0%**
unannotated training images		
Borenstein et al	93.6%	-
hand-segmented training images		
LOCUS - no class model	88.6%	82.1%

Results


Horses
Faces
Cars (rear)
Cars (side)
Motorbikes
Planes
Cows
Trees
Class mask prob.
Class edge mean
Example seg.
Example seg.