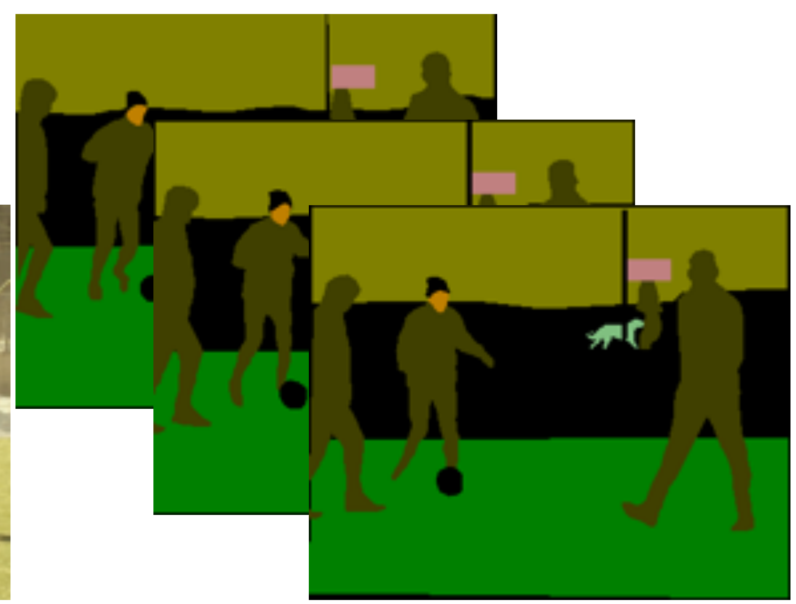
Bayesian Nonparametric Discovery of Layers and Parts from Scenes and Objects

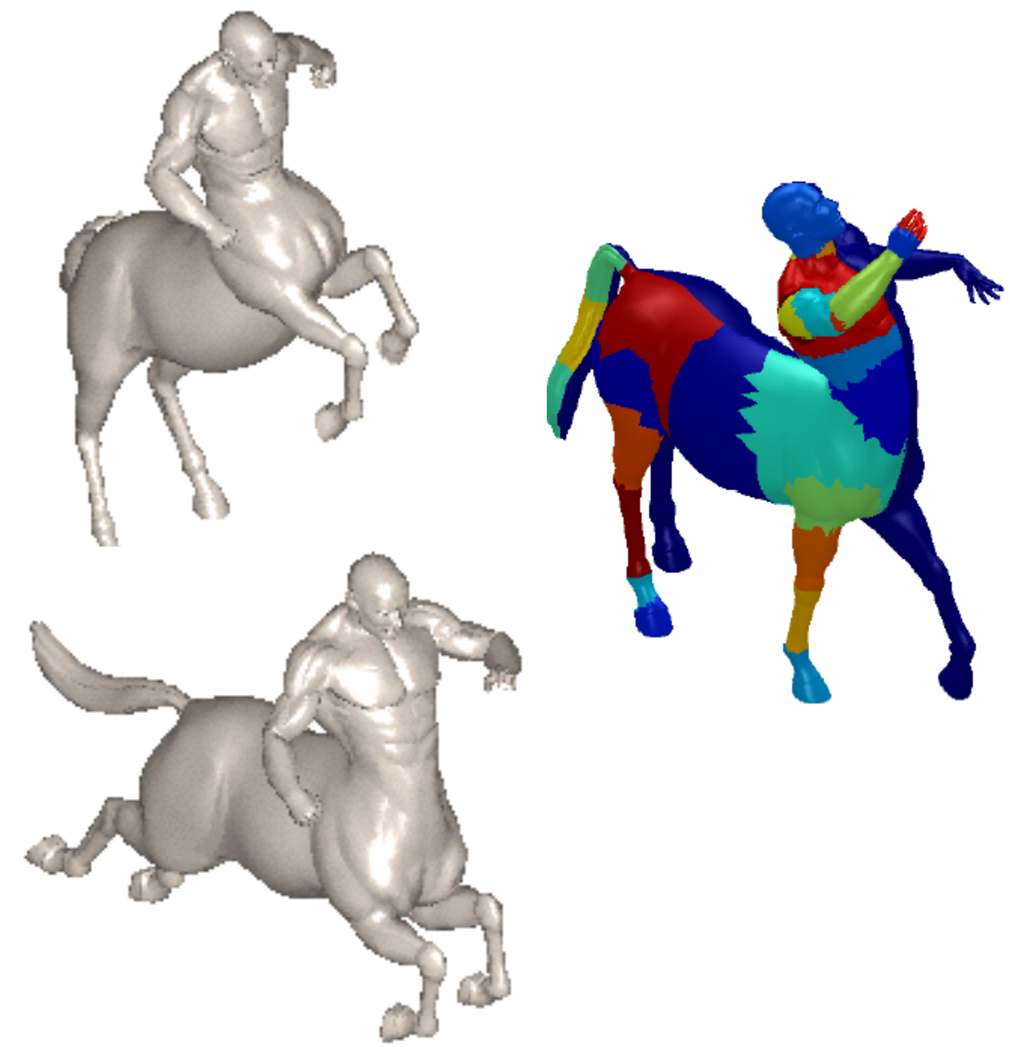
Soumya Ghosh

Advisor: Erik Sudderth

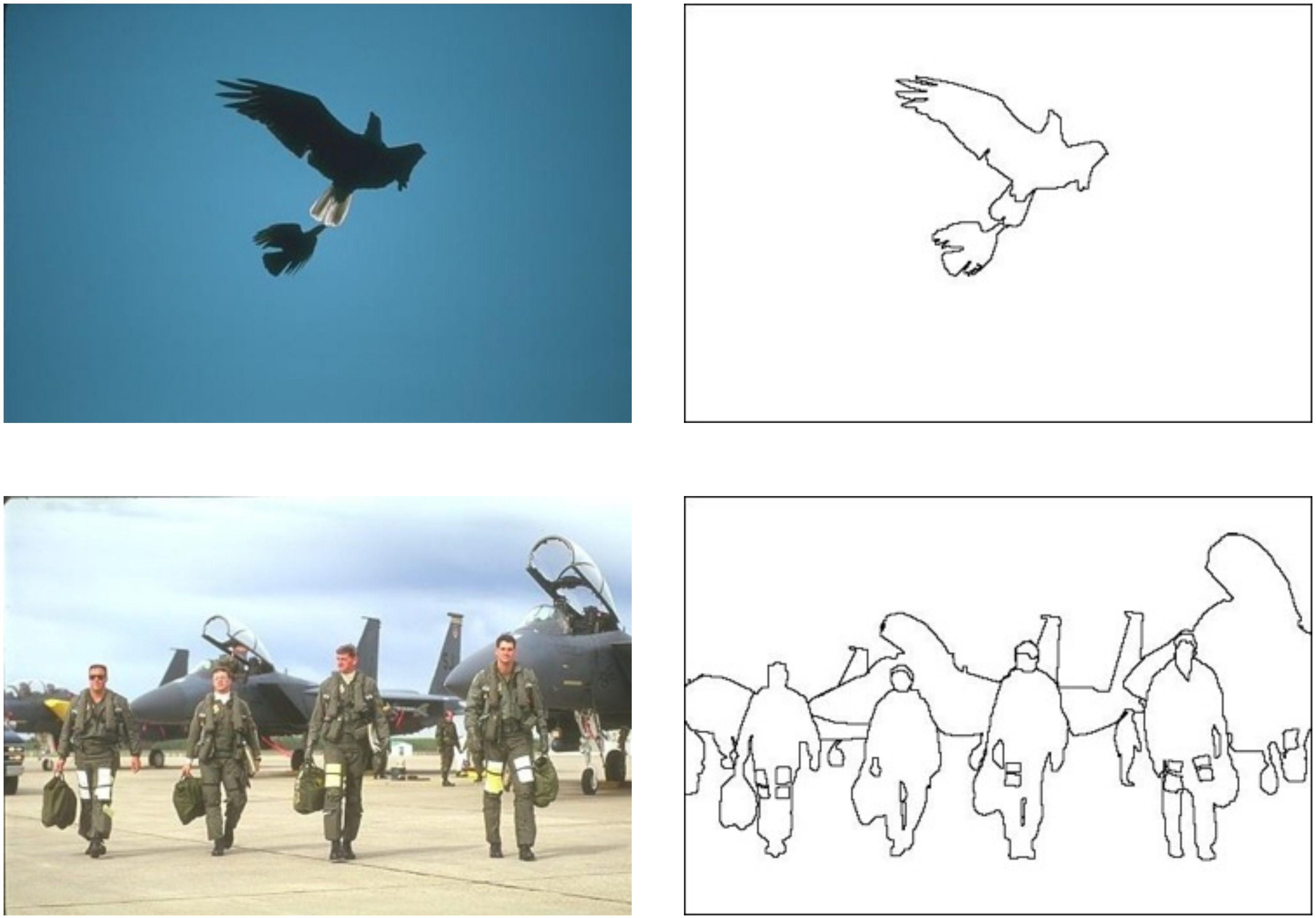
Committee: Michael Black and James Hays

Introduction

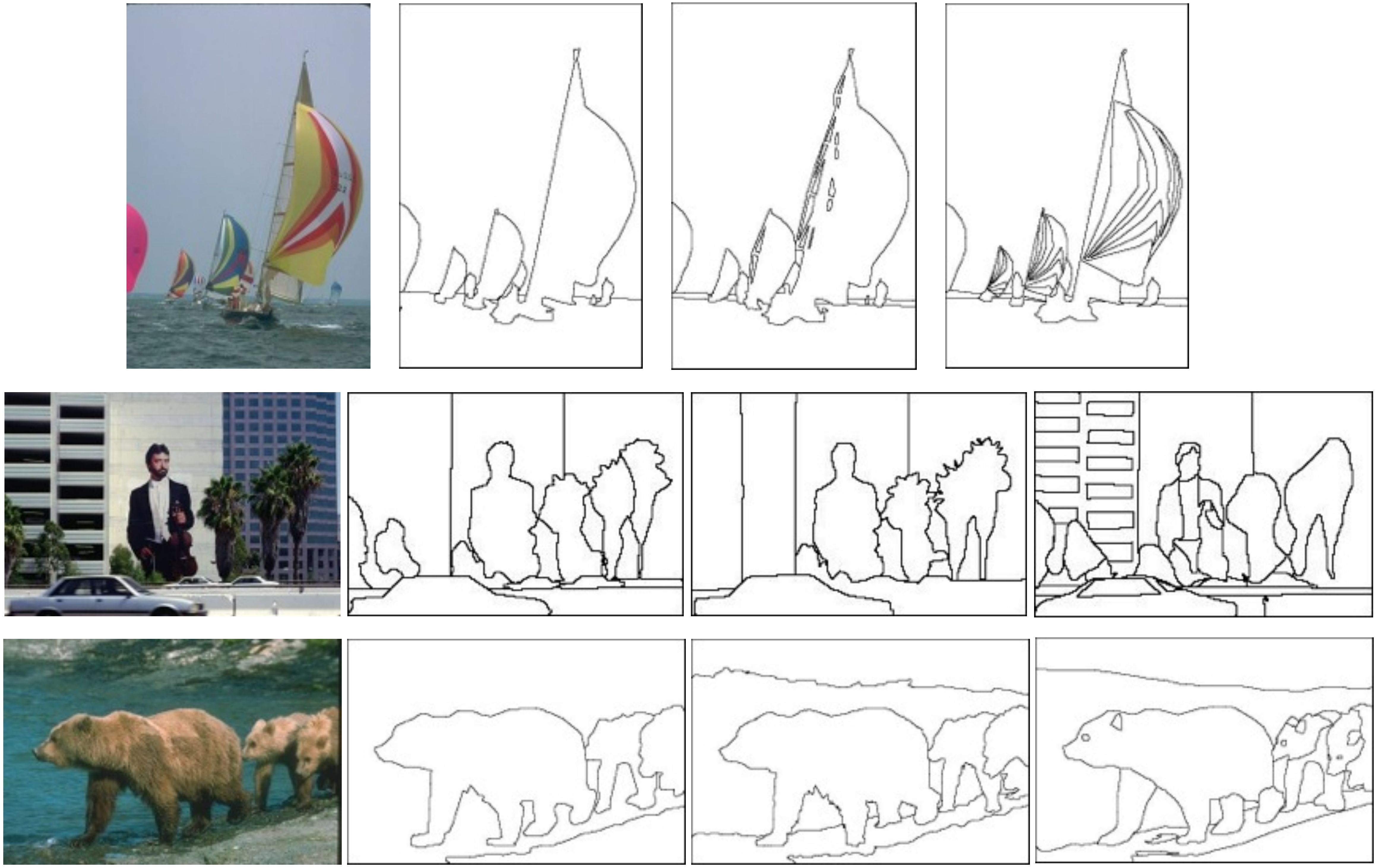




World is Complex



World is Complex

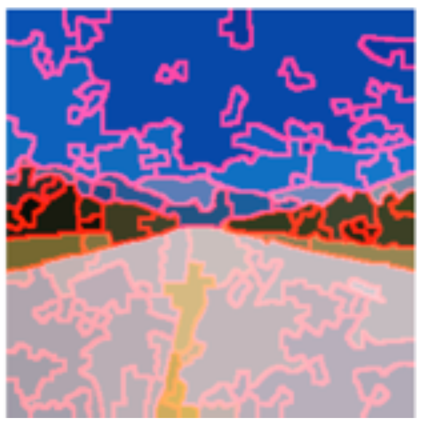


Human Segmentations

Model Desiderata

- Automatic model selection - adapt to variability in image/video/object complexity
- Manage uncertainty - retain a distribution over possible segmentations
- Model spatial and temporal correlations
- Match statistics of human segmentations

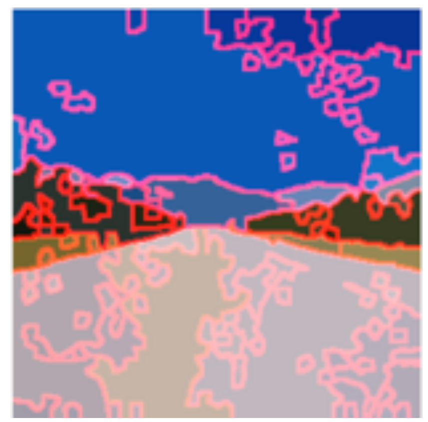
Adapting to complexity: Distributions over partitions



$$p(\mathbf{Z_1})$$

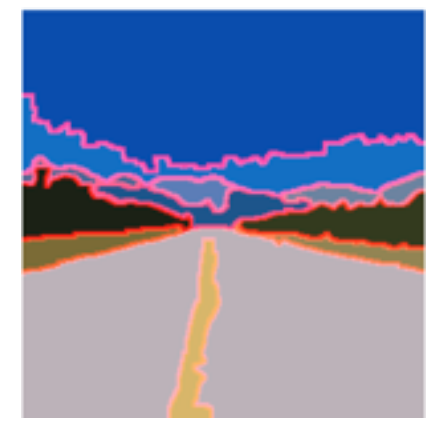
$$p(\mathbf{X}|\mathbf{Z_1})$$

Data Partition



$$p(\mathbf{Z_2})$$

$$p(\mathbf{X}|\mathbf{Z_2})$$

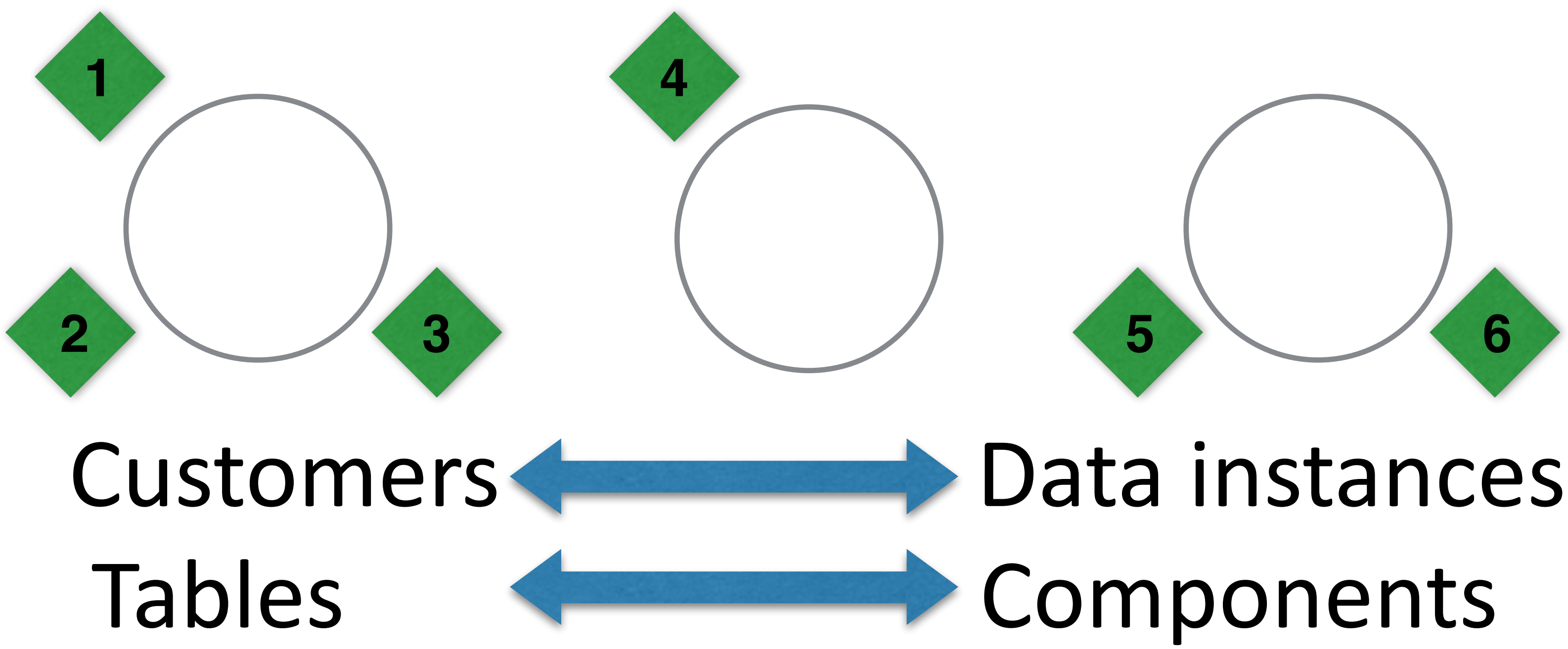


$$p(\mathbf{Z_3})$$

$$p(\mathbf{X}|\mathbf{Z_3})$$

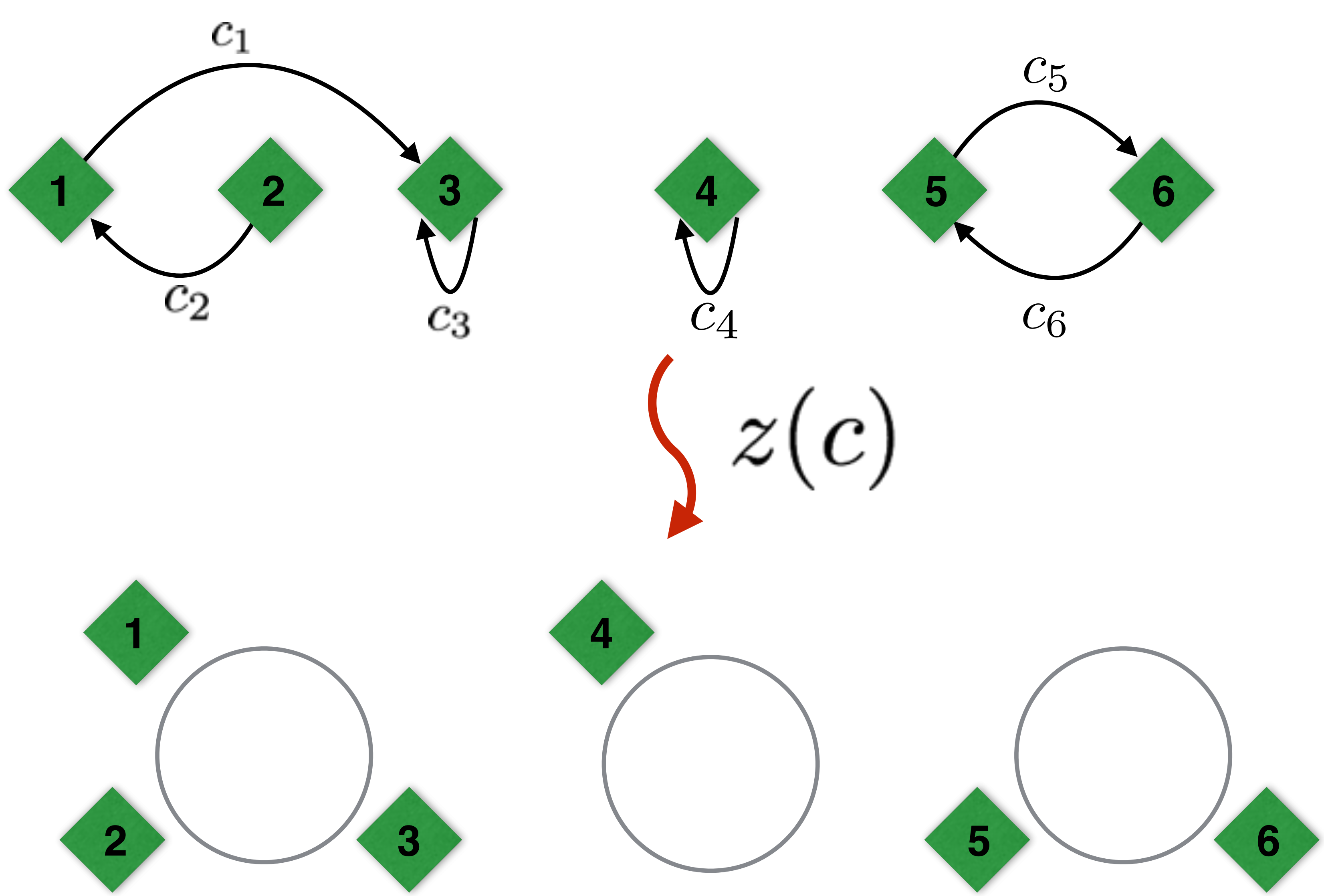
$$\mathbf{Z}^* = \arg\max_{\mathbf{Z_i}} p(\mathbf{Z}|\mathbf{X}) \propto p(\mathbf{X}|\mathbf{Z}=\mathbf{Z_i})\, p(\mathbf{Z}=\mathbf{Z_i})$$

A distribution over partitions: Chinese Restaurant Process



- A restaurant with infinitely many tables, each table with infinite capacity.
- The first customer sits at a table.
- Subsequent customers pick a table *k* w.p.: $\propto \begin{cases} n_k & \text{if } k \text{ is an existing table} \\ \alpha & k \text{ is a new table} \end{cases}$

Modeling Correlations: Distance dependent Chinese Restaurant Process (ddCRP)



$$p(c_m = n \mid D, \alpha) \propto \begin{cases} f(d_{mn}) & \text{if } m \neq n, \\ \alpha & \text{if } m = n. \end{cases}$$

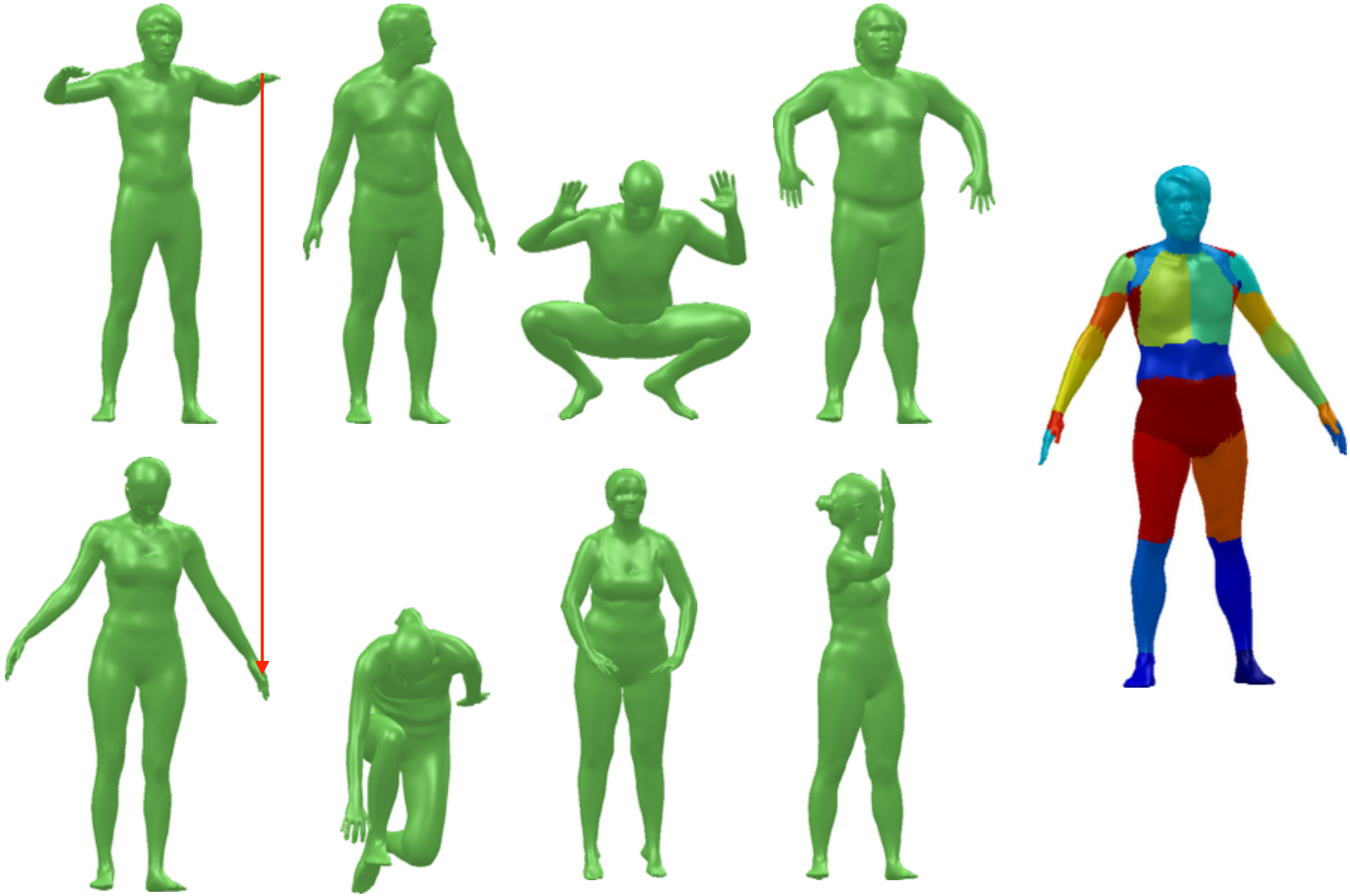
$$p(z(c) \mid D, \alpha) = \prod_n p(c_n \mid D, \alpha)$$

Blei & Frazier, 2011

Talk Outline

- Deformable 3D object segmentation
- Image and Video segmentation
- Layered decomposition of natural images
- Proposed Work

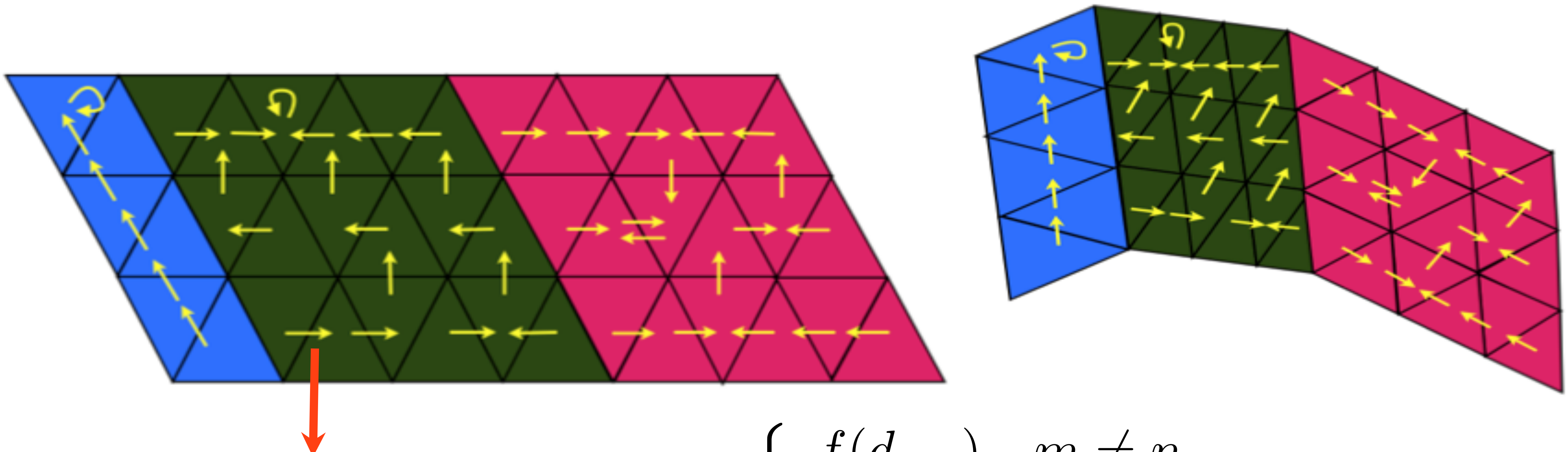
Discovering Parts from Deformations



Discovering Parts from Deformations: Big Picture

- **Cluster:** Mesh faces.
- **Prior:** over the space of plausible mesh partitions.
- **Likelihood:** Given segmentation into parts, model how multiple bodies deform across many poses.
- **Posterior:** Explored through MCMC.

ddCRP Prior over Mesh Partitions

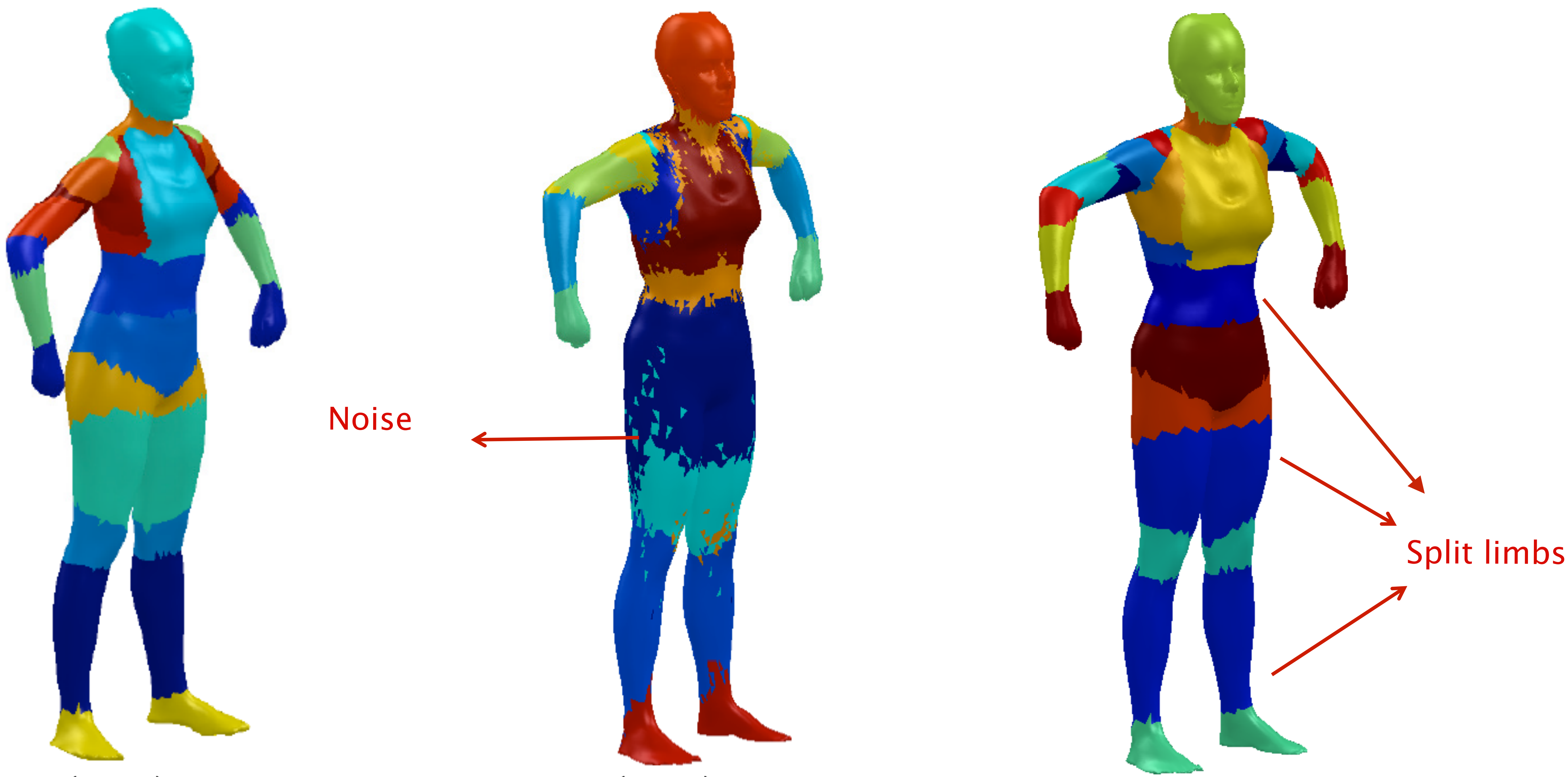


$$p\left(c_m = n \mid D, f, \alpha\right) \propto \begin{cases} f(d_{mn}) & m \neq n, \\ \alpha & m = n. \end{cases}$$

- Mesh faces are only allowed to link to neighboring faces:

$$f(d_{mn}) = \mathbf{1}[d_{mn} \leq 1]$$

Prior over plausible partitions



$p(Z_1) > 0$

Desirable

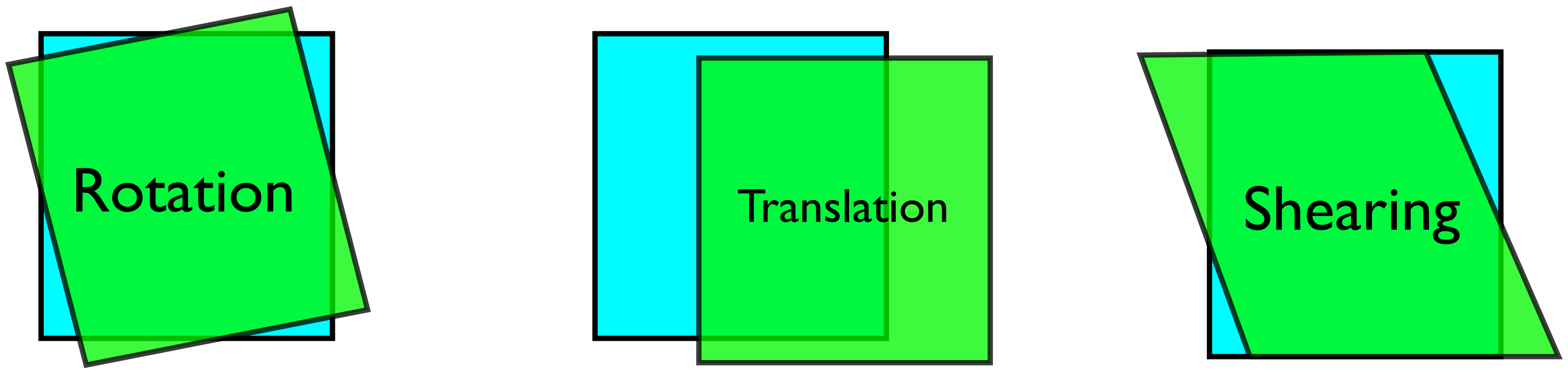
$p(Z_2) = 0$

Avoid: Noisy Parts

$p(Z_3) = 0$

Avoid: Disconnected Parts

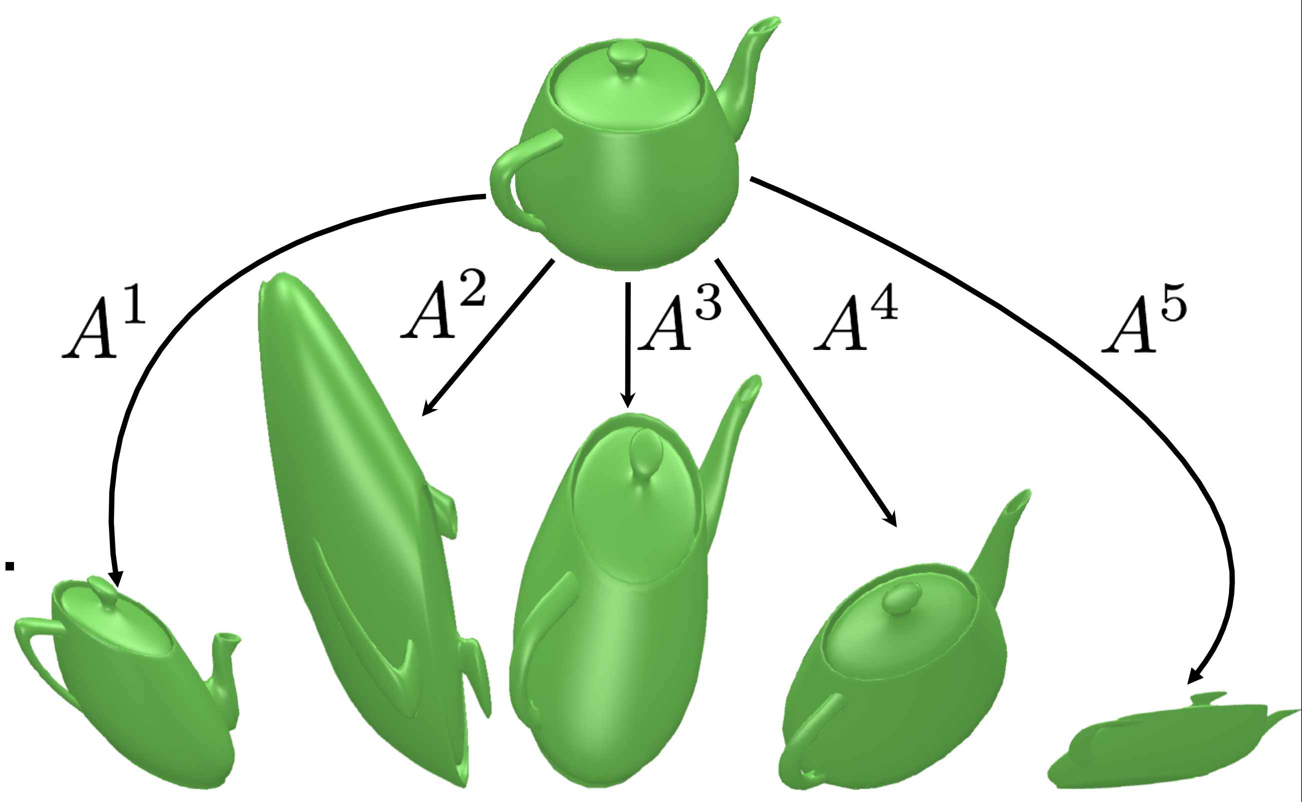
Distributions over Affine Transformations



Matrix Normal Inverse Wishart:

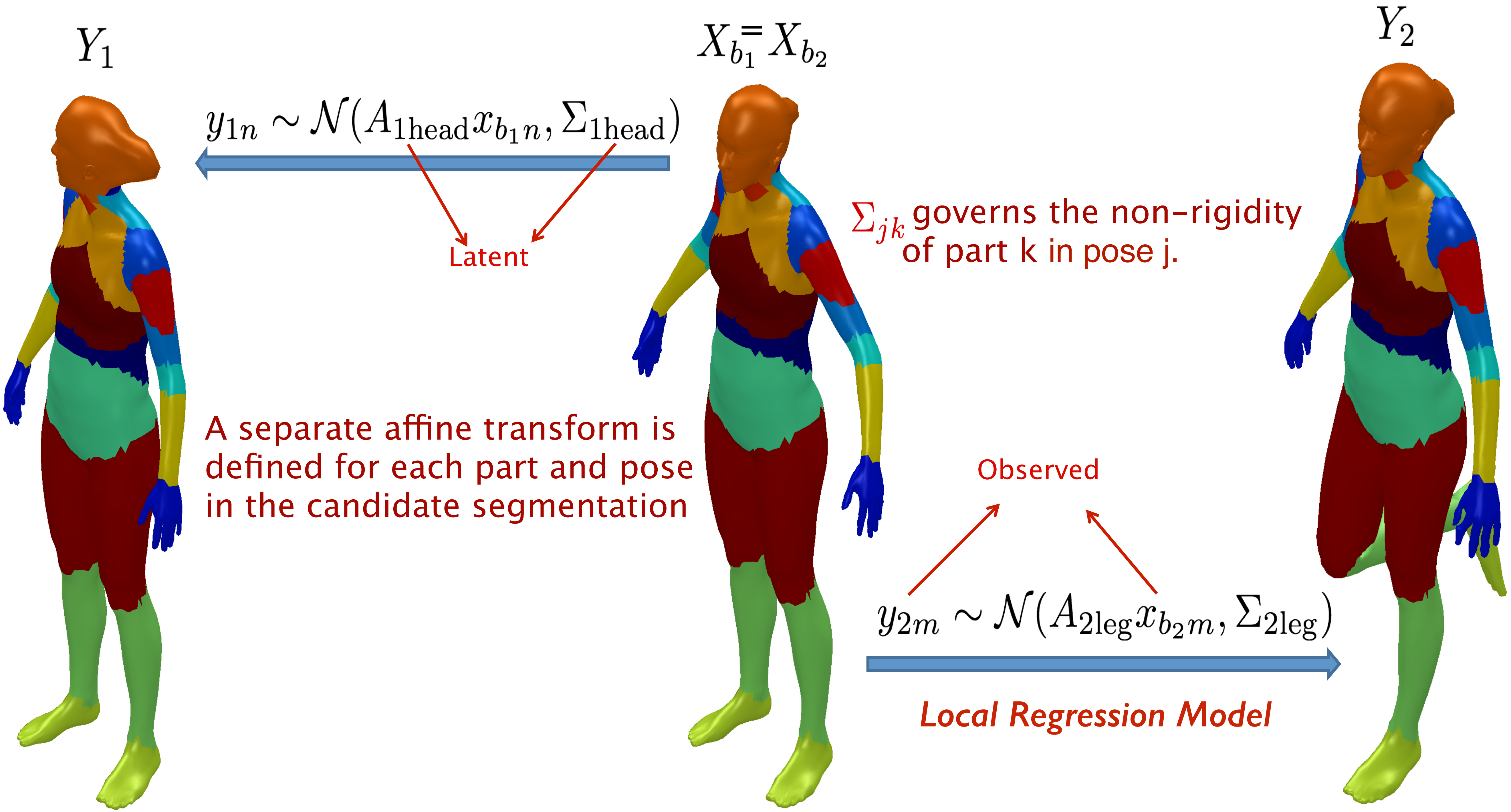
$$\Sigma \sim \mathcal{IW}(n_0, S_0)$$
$$A \mid \Sigma \sim \mathcal{MN}(M, \Sigma, K)$$

where $A \in \mathbb{R}^{3\times4}$ is an affine transformation.



$$A^1 \ldots A^5 \sim \mathcal{MNIW}(.)$$

Generative Affine Likelihoods



Marginal Affine Likelihoods

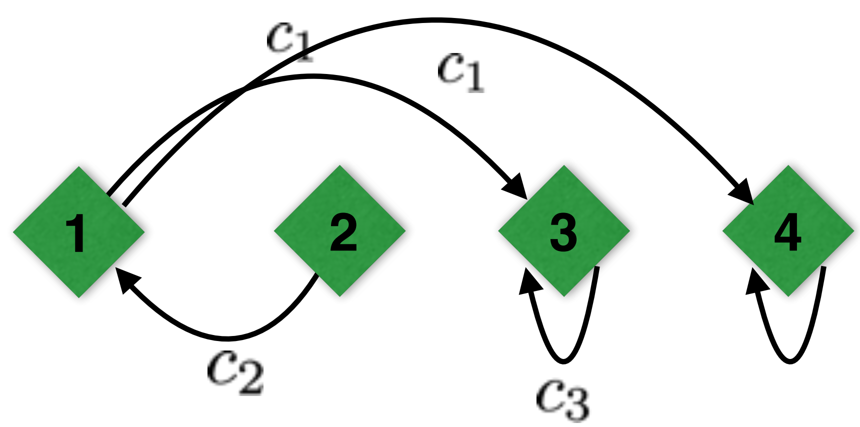
- We do not need to estimate the latent affine transformations. For each part and pose combination we can analytically marginalize them

$$p(Y_{jk} \mid X_{jk}) = \int p(Y_{jk}, A_{jk}, \Sigma_{jk} \mid X_{jk}) dA_{jk}, d\Sigma_{jk}$$

Marginal Likelihood

- Bayesian Model Selection:
  - Improper merges have low marginal likelihoods
  - Improper splits are "suspicious coincidences" and end up with lower marginal likelihoods

Inference through Gibbs Sampling



Customers = Mesh Faces
Tables = Object Parts

Collapsed Sampler:
Only need to sample links,
other random variables are
analytically marginalized out.

$$z(c)$$

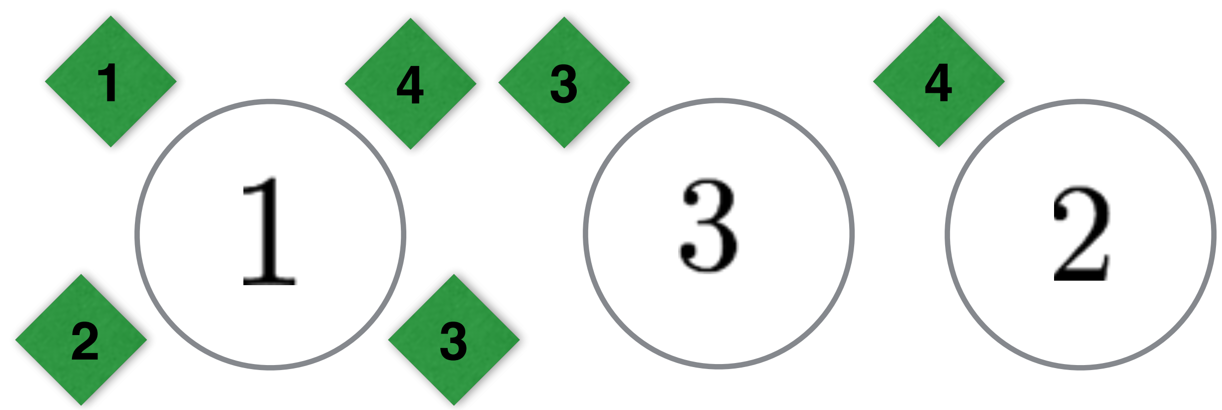


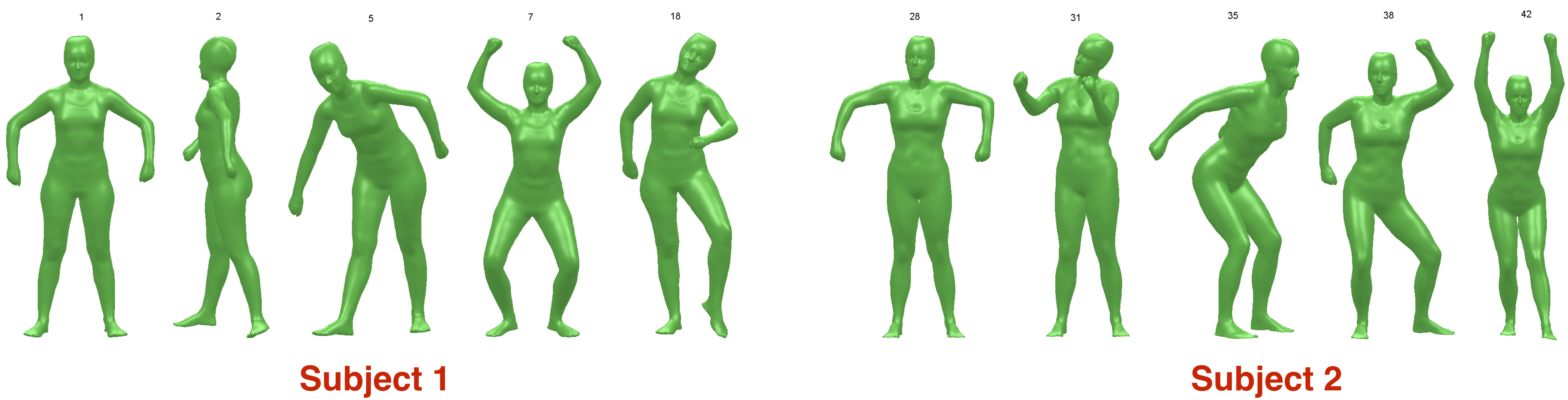
Table structure

Segmentation

Local changes in the link structure lead to large changes in the partition structure

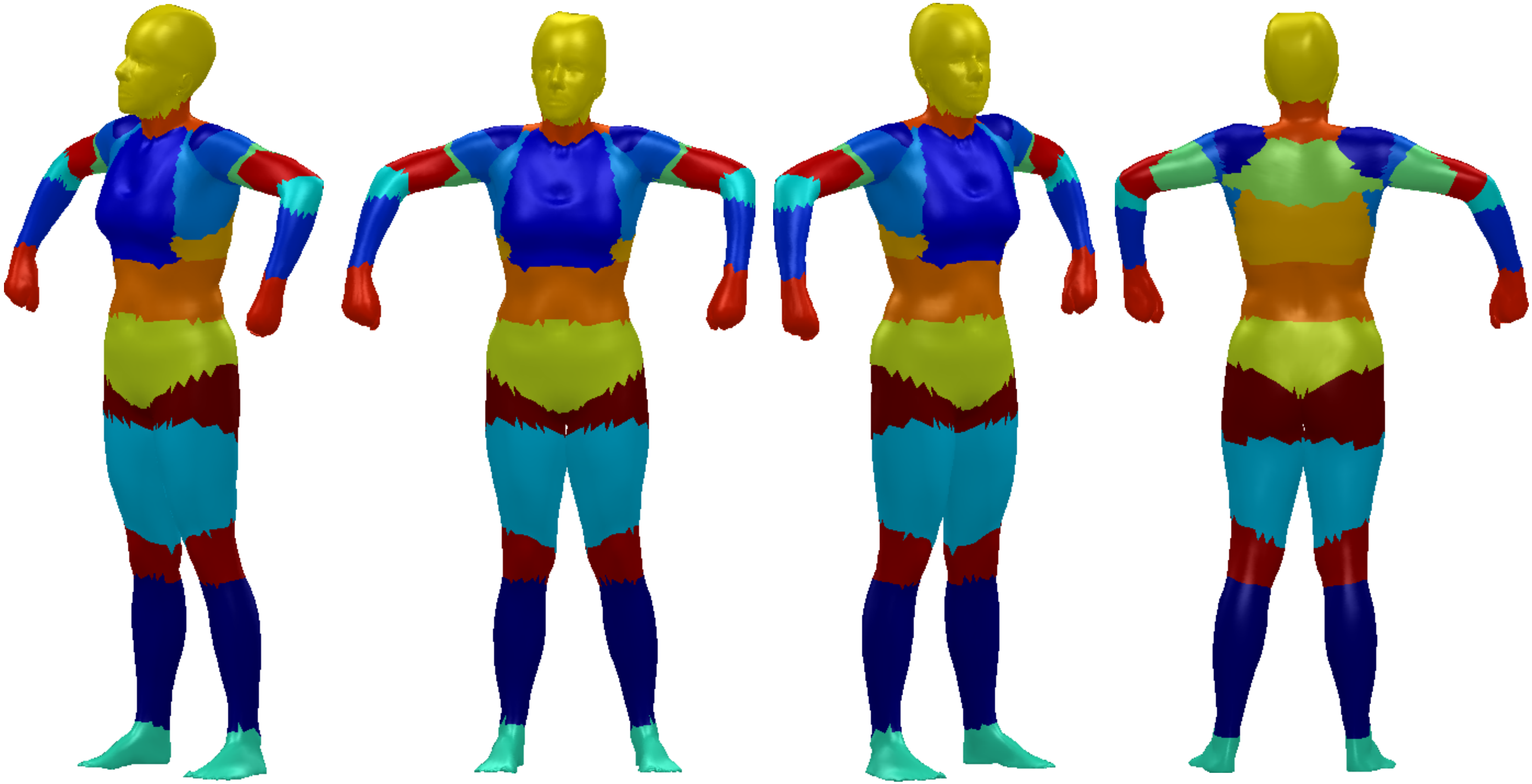
Analysis of Human Bodies in Motion

- 56 Aligned scans from two human subjects
- Wide variability in poses, limited variability in body shapes



Ghosh et al., NIPS 2012

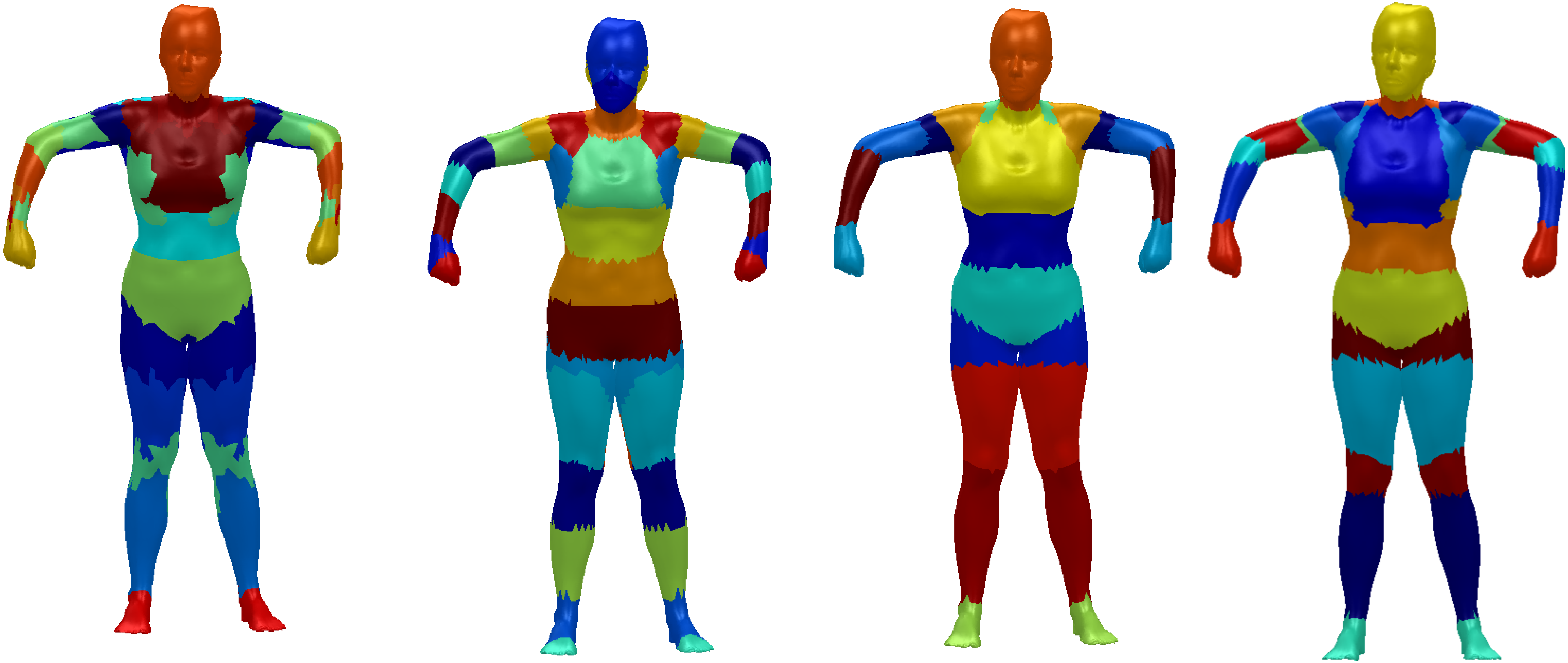
Inferred Segmentation



Discovered a segmentation with 20 Parts

Ghosh et al., NIPS 2012

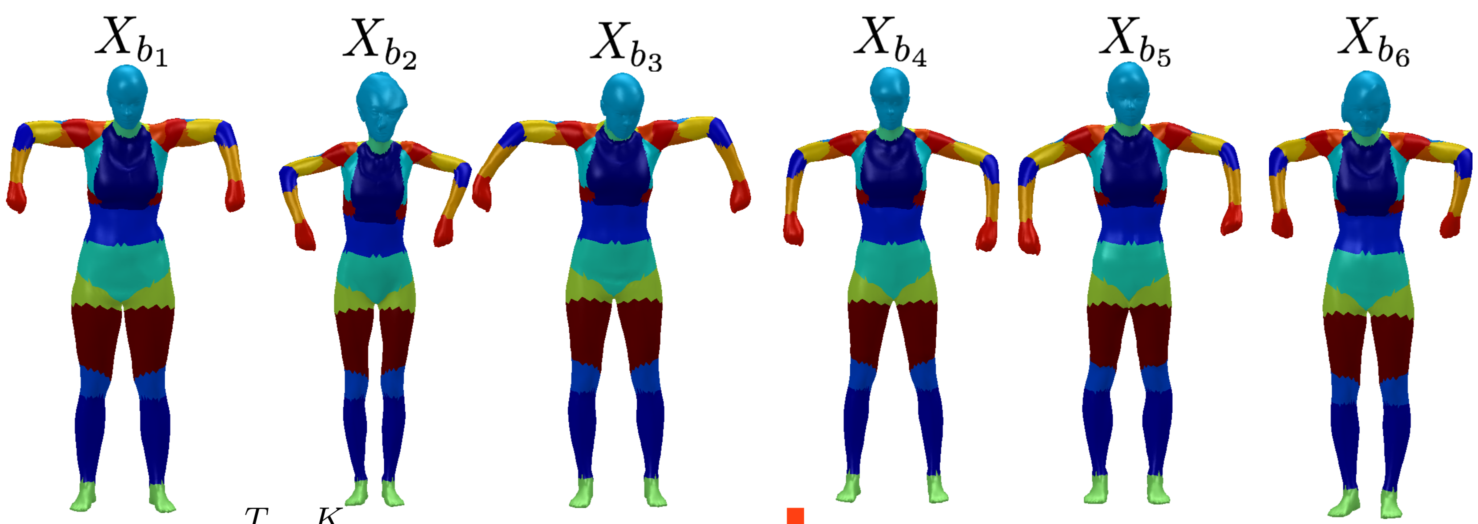
Visual Comparisons



Agglomerative | **Spectral Clustering** *Liu & Zhang, 2004* | **CRP** | **ddCRP**

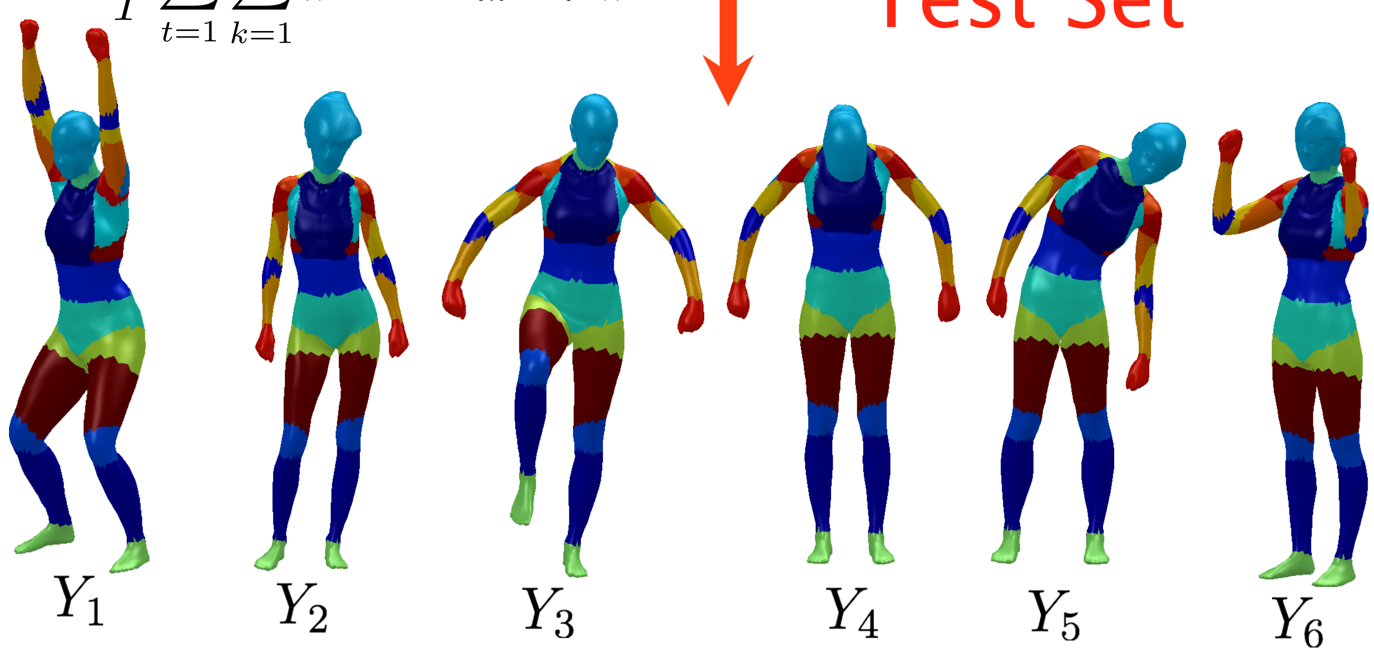
Ghosh et al., NIPS 2012

Quantitative Evaluation

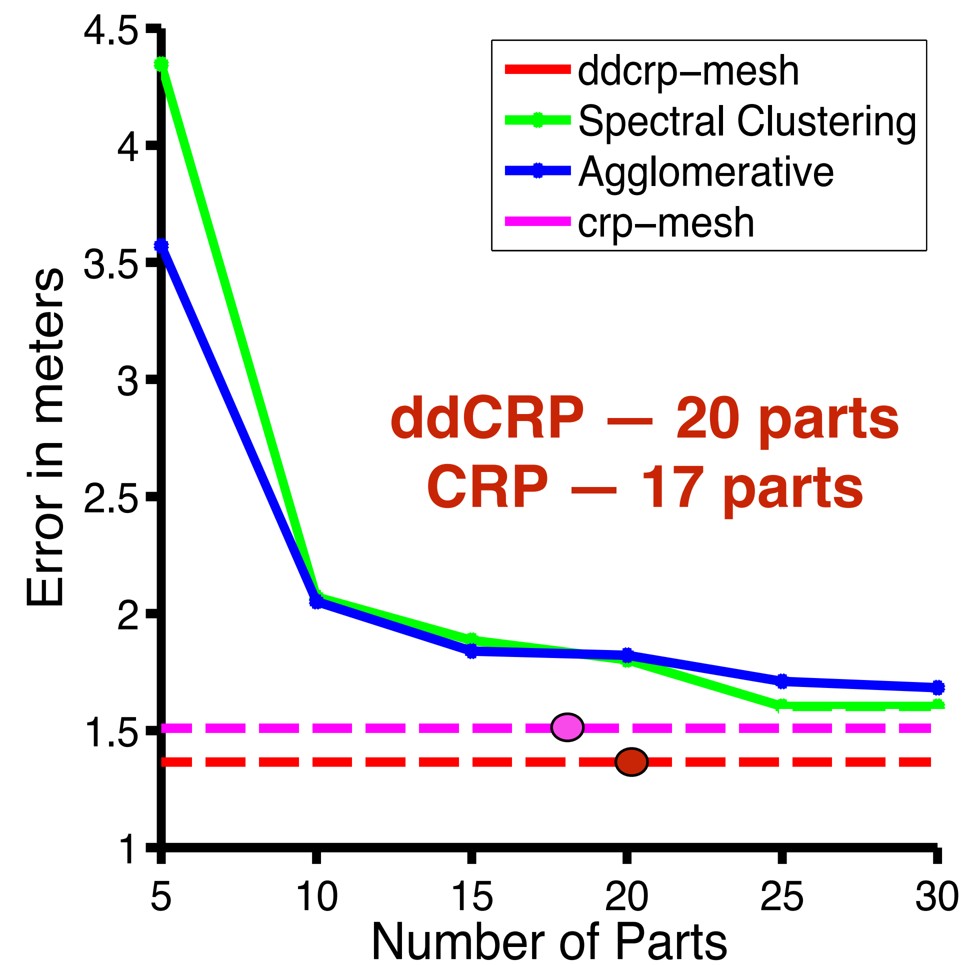


$$\mathcal{E} = \frac{1}{T}\sum_{t=1}^{T}\sum_{k=1}^{K}||Y_{tk} - A^*_{tk}X_{b_t k}||_2$$

Test Set



12 Test meshes from six subjects.

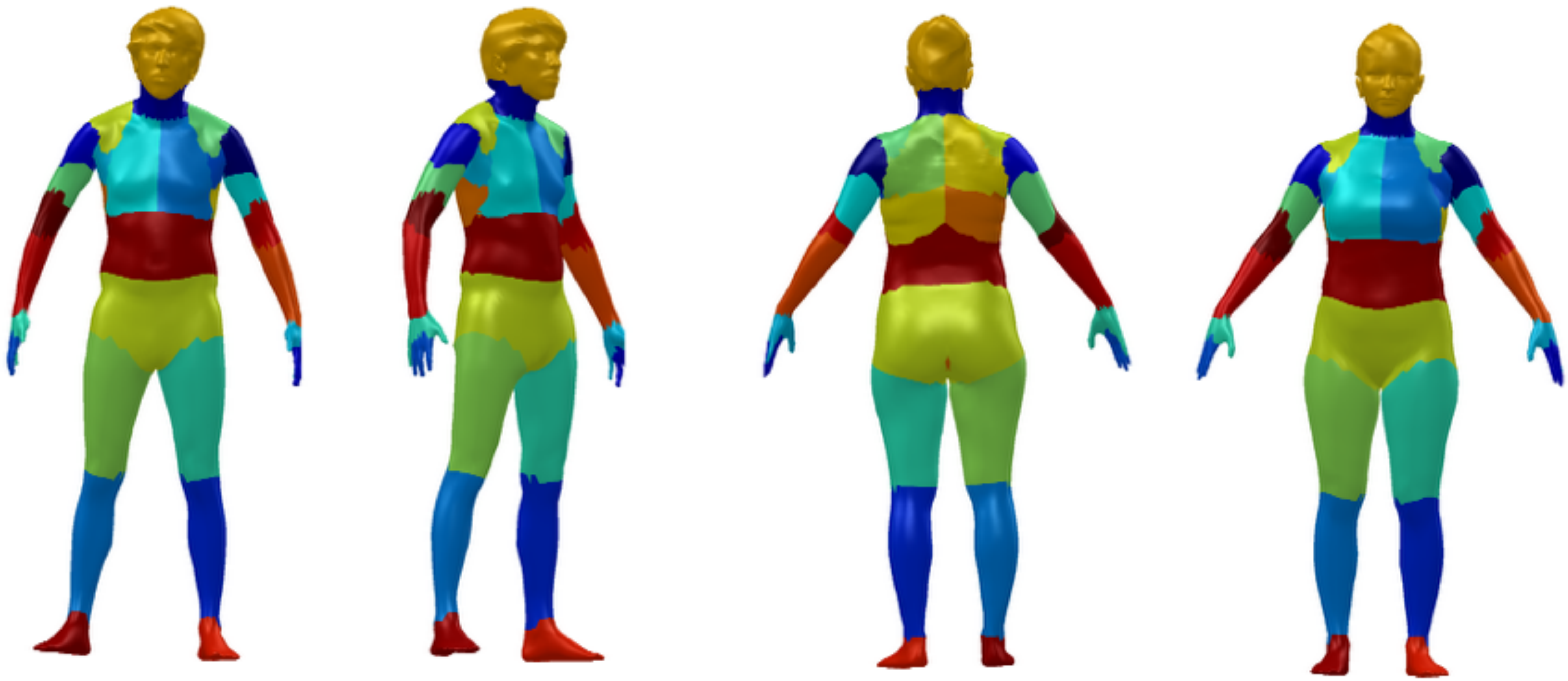


- Measure error in predicted motion for the candidate segmentations

Ghosh et al., NIPS 2012

Larger Studies

- 1732 meshes, 78 subjects.
- Wide variability in both body shapes and poses.
- Similar results — part boundaries are cleaner.



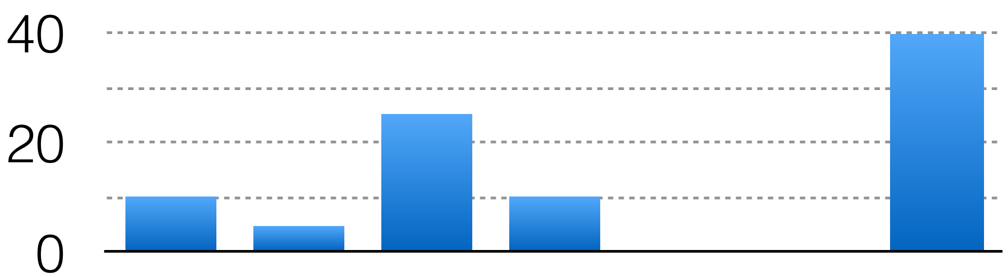
Computer generated meshes

Talk Outline

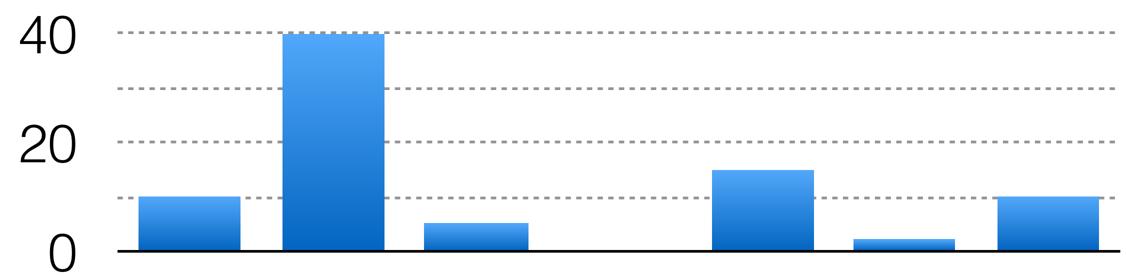
- Deformable 3D object segmentation
- Image and Video segmentation
- Layered decomposition of natural images
- Proposed Work

Image Representation

- ~1000 super-pixels
- Each super-pixel is described through a histogram (~120 bin) of color and texture

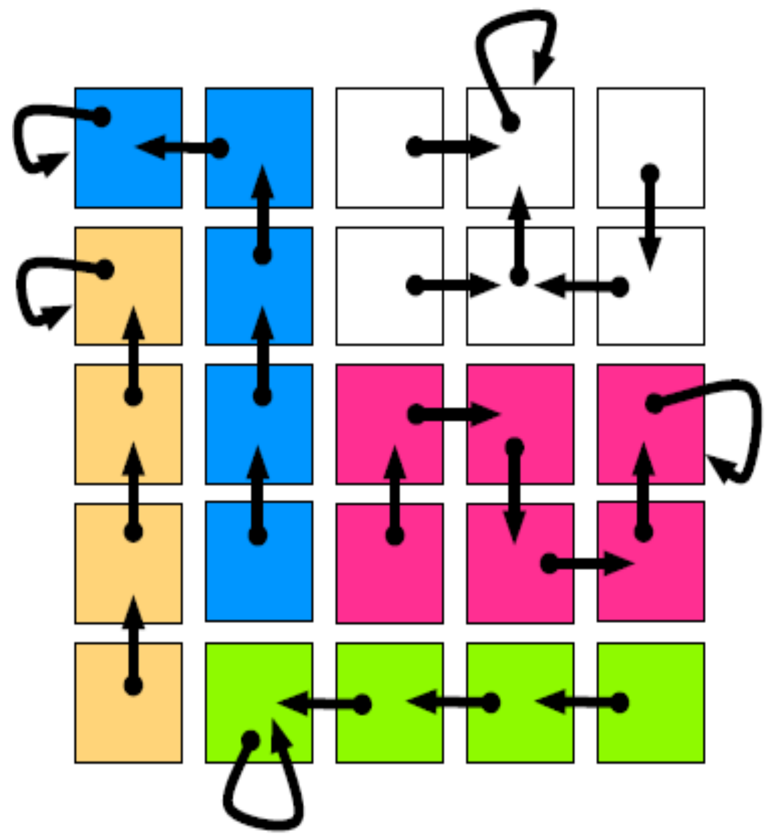


$$x_i^{color} \sim \mathrm{Mult}(\theta^{color})$$



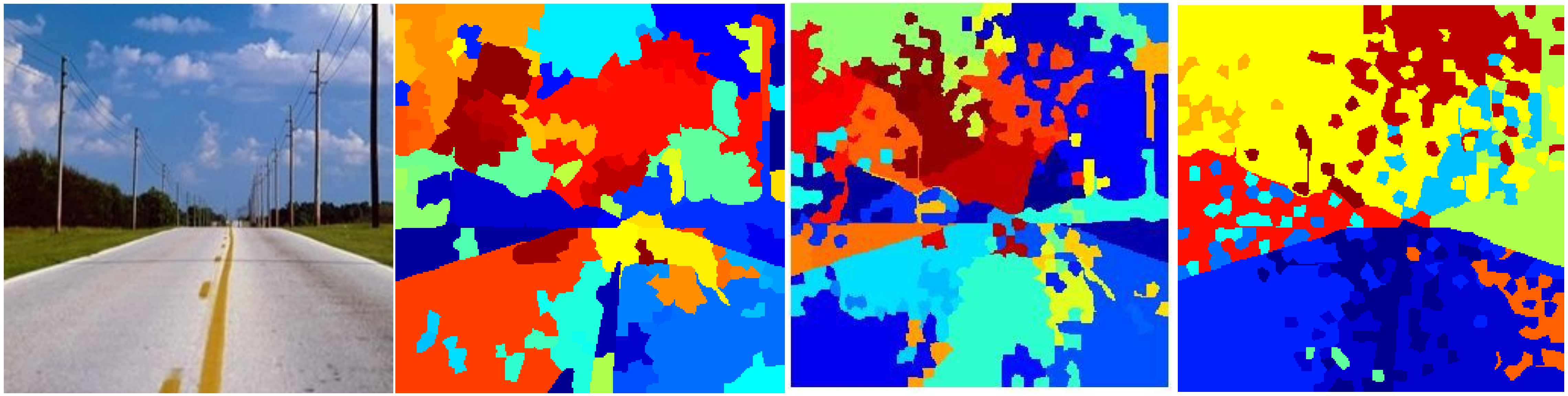
$$x_i^{texture} \sim \mathrm{Mult}(\theta^{texture})$$

Spatial ddCRP prior on Image Segmentations



$$p(c_m = n \mid D, \alpha) \propto \begin{cases} f(d_{mn}) & \text{if } m \neq n, \\ \alpha & \text{if } m = n. \end{cases}$$

ddCRP1 $f(d_{mn}) = \mathbf{1}[d_{mn} \leq 1]$ **ddCRP2** $f(d_{mn}) = \mathbf{1}[d_{mn} \leq 2]$ **ddCRP5** $f(d_{mn}) = \mathbf{1}[d_{mn} \leq 5]$



Hierarchical Variant - “region” ddCRP

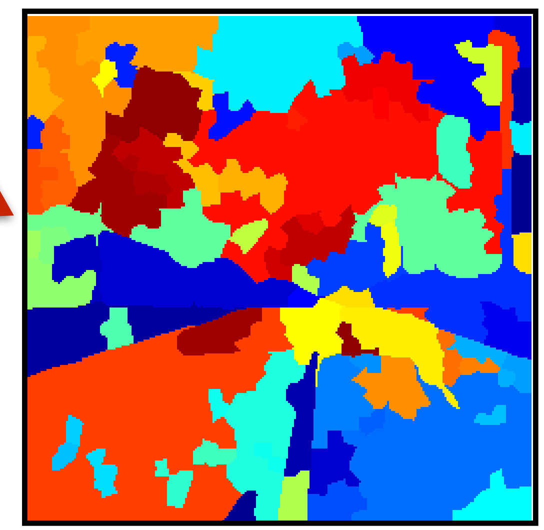
- Each super-pixel connects to a spatial neighbor. This produces an oversegmentation of the image.

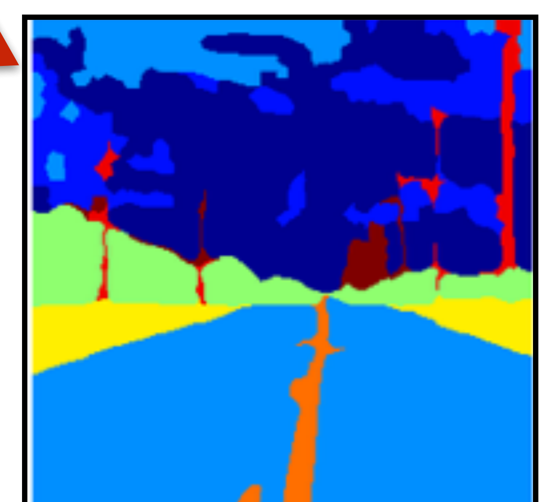
$$c_i \sim \mathrm{ddCRP}(\alpha, f, D)$$

- Group the segments into regions according to a Chinese restaurant process.

- Each region k maintains an appearance distribution parametrized by ϕ_k

- Color and texture histograms at super-pixels are then generated by sampling the appropriate region appearance distributions: $x_i \sim p(x_i \mid \phi_{z_i})$





Ghosh et al., NIPS 2011

Inference - Gibbs Sampling

- Repeat till exhausted:
    - Sample super-pixel links (identical to ddCRP):

$$c_i \sim p(c_i \mid c_{-i}, k, x)$$

    - These define an over segmentation of the image.
    - *Extra wrinkle*: sample region assignments for each segment.

$$k_t \sim p(k_t \mid c, k_{-t}, x)$$

Ghosh et al., NIPS 2011

Eight Natural Scene Category Dataset (LabelMe)

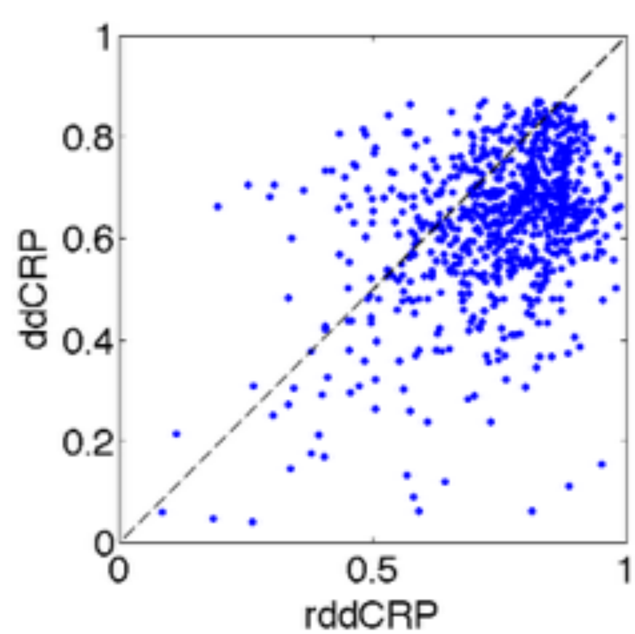
800 images in all

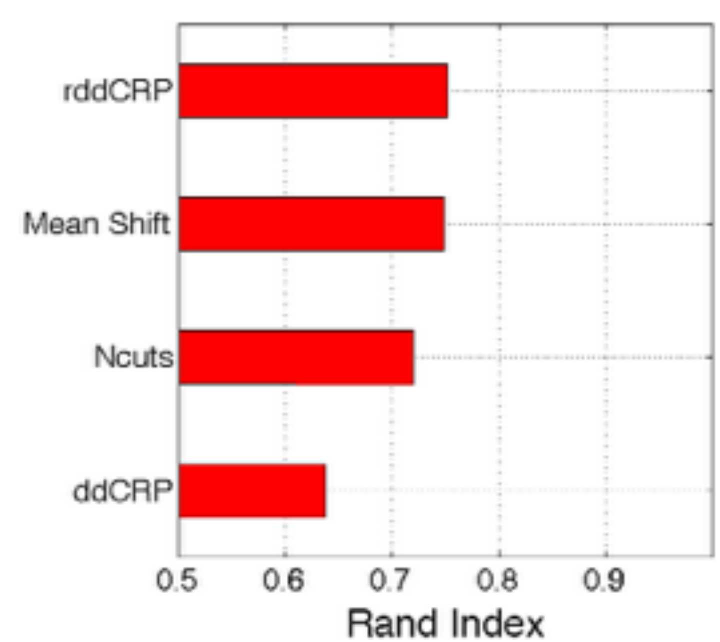
Oliva and Torralba 2001

Segmentation Results

Rand Index

Measures agreement
with human segmentations





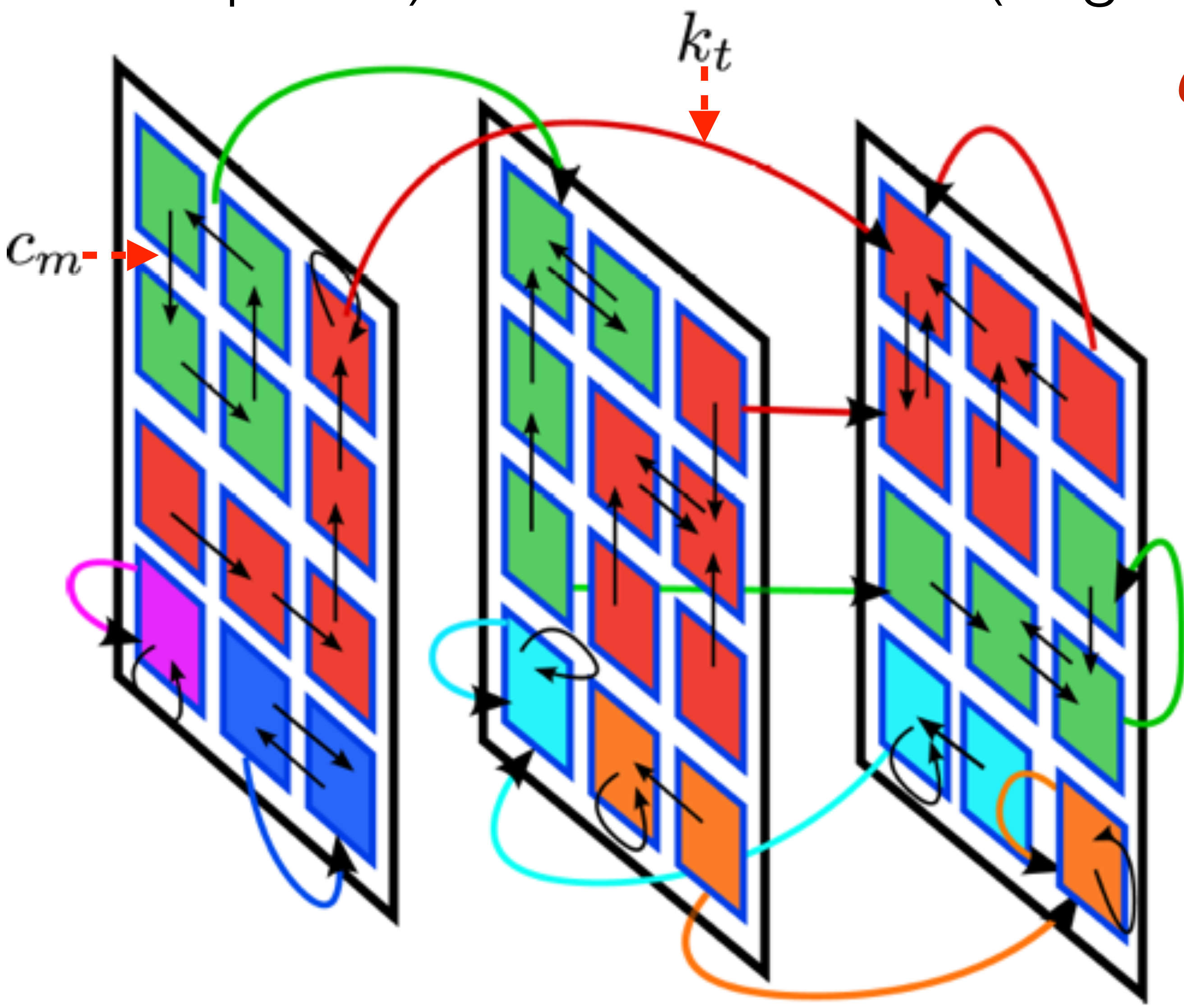
Video Segmentation

- “region” ddCRP models spatial correlations between pixels

- But tables are grouped with a CRP— Inter table correlations are not modeled

Generalized Hierarchical ddCRPs

- Model for grouped data.
- Key Idea: Allow distances between both customers (super-pixels) and *latent* tables (segments).



Group indexed (j) ddCRPs :

$$p(c_{jm} \mid D_j, f_j, \alpha_j) \propto \begin{cases} f_j(d_j(mn)) & \text{if } jm \neq jn \\ \alpha_j & \text{if } jm = jn \end{cases}$$

Tables are coupled using a inter-table ddCRP :

$$p(k_t = t' \mid D_0(c), f_0, \alpha_0) \propto \begin{cases} f_0(d_0(t, t', c)) & \text{if } t \neq t' \\ \alpha_0 & \text{if } t = t' \end{cases}$$

Table distances could be functions of:

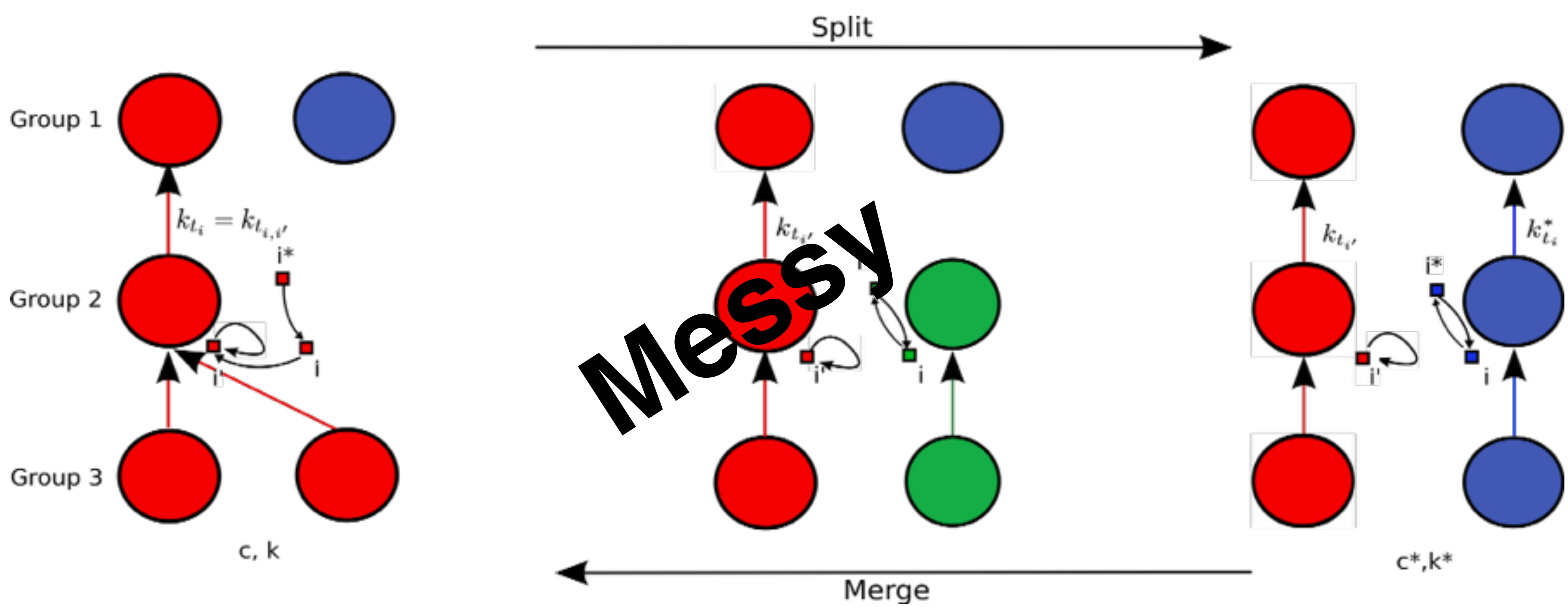
i) table sizes
ii) table shapes
iii) table locations
iv) table neighborhoods

Inference - Metropolis Hastings

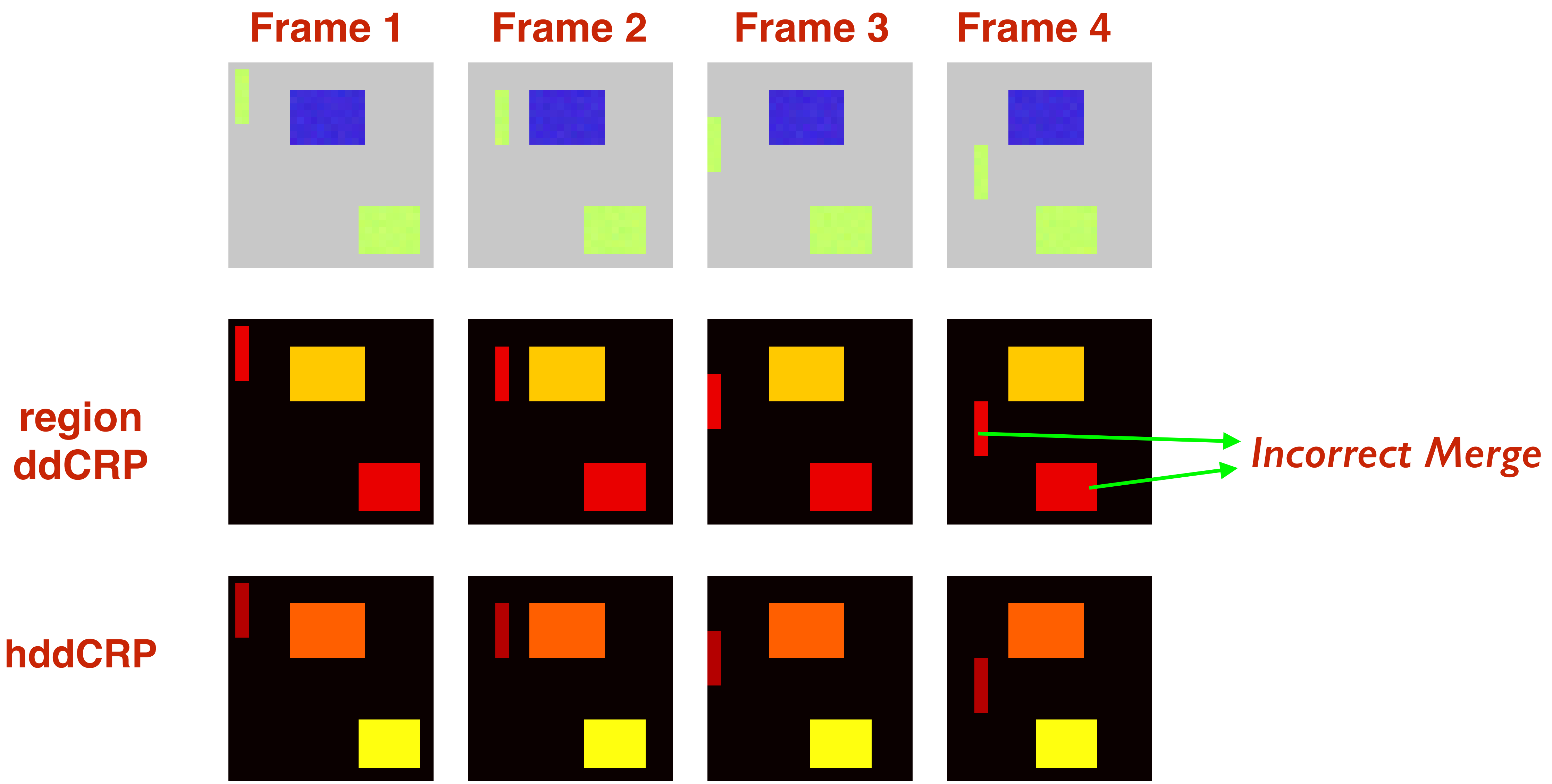
- Composite moves — make coordinated changes to customer and table links.

$$q(c_i^*) \propto p(c_i^* \mid \alpha, D)\Gamma(\mathbf{X}, \mathbf{z}, \lambda),$$

$$\Gamma(\mathbf{X}, \mathbf{z}, \lambda) = \begin{cases} \dfrac{p(\mathbf{X}_{\mathbf{z}(\Delta)=m_a} \cup \mathbf{X}_{\mathbf{z}(\Delta)=m_b} \mid \lambda)}{p(\mathbf{X}_{\mathbf{z}(\Delta)=m_a} \mid \lambda) p(\mathbf{X}_{\mathbf{z}(\Delta)=m_b} \mid \lambda)} & \text{if } c_i^* \text{ merges dishes } m_a \text{ and } m_b \\ 1 & \text{otherwise.} \end{cases}$$



Video Segmentation-Toy Data

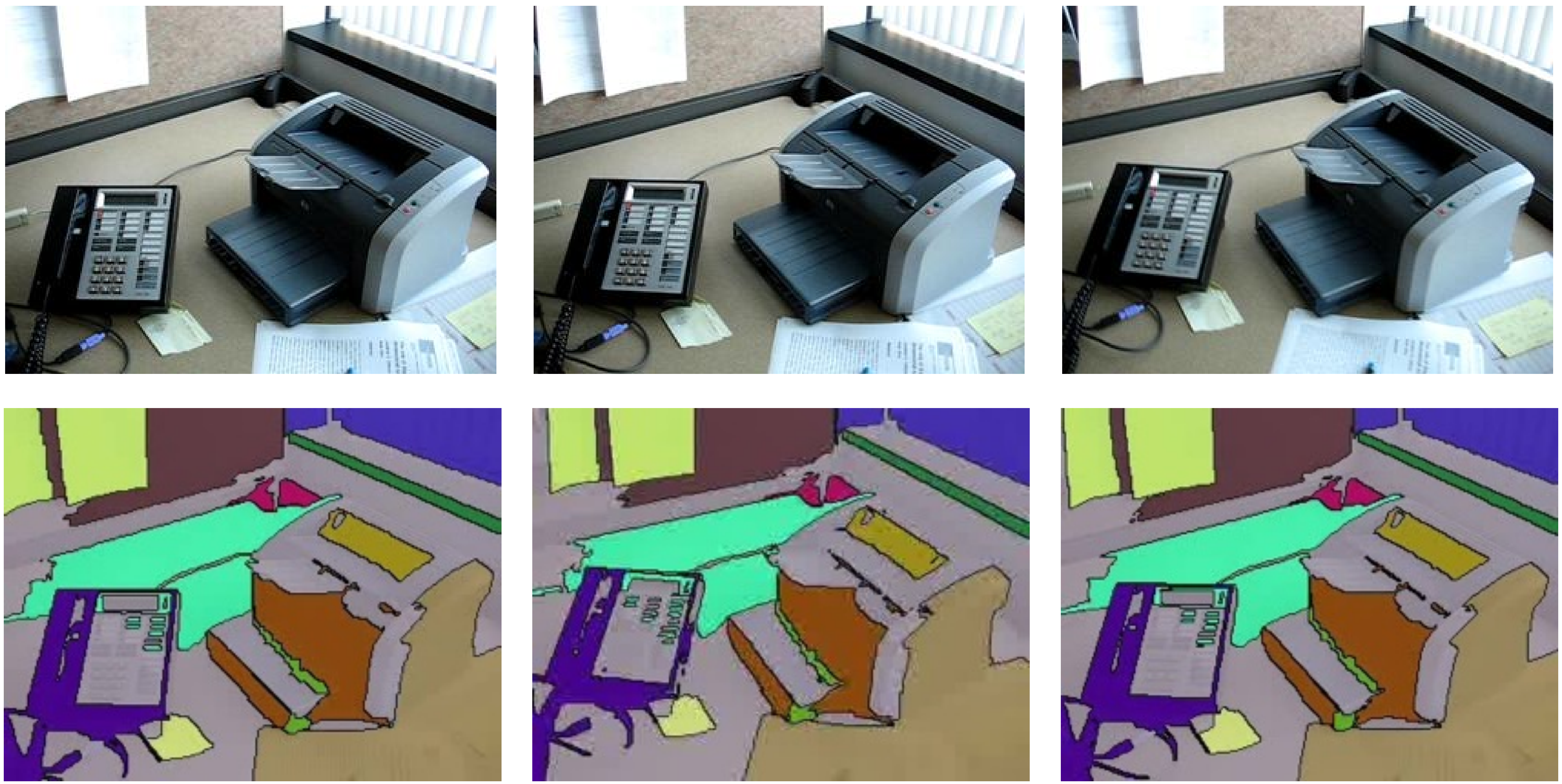


Size based inter-table distances.

Groups those tables (segments) together which in addition to having similar appearance have similar sizes.

Video Segmentation Results

Frame 1 **Frame 5** **Frame 10**

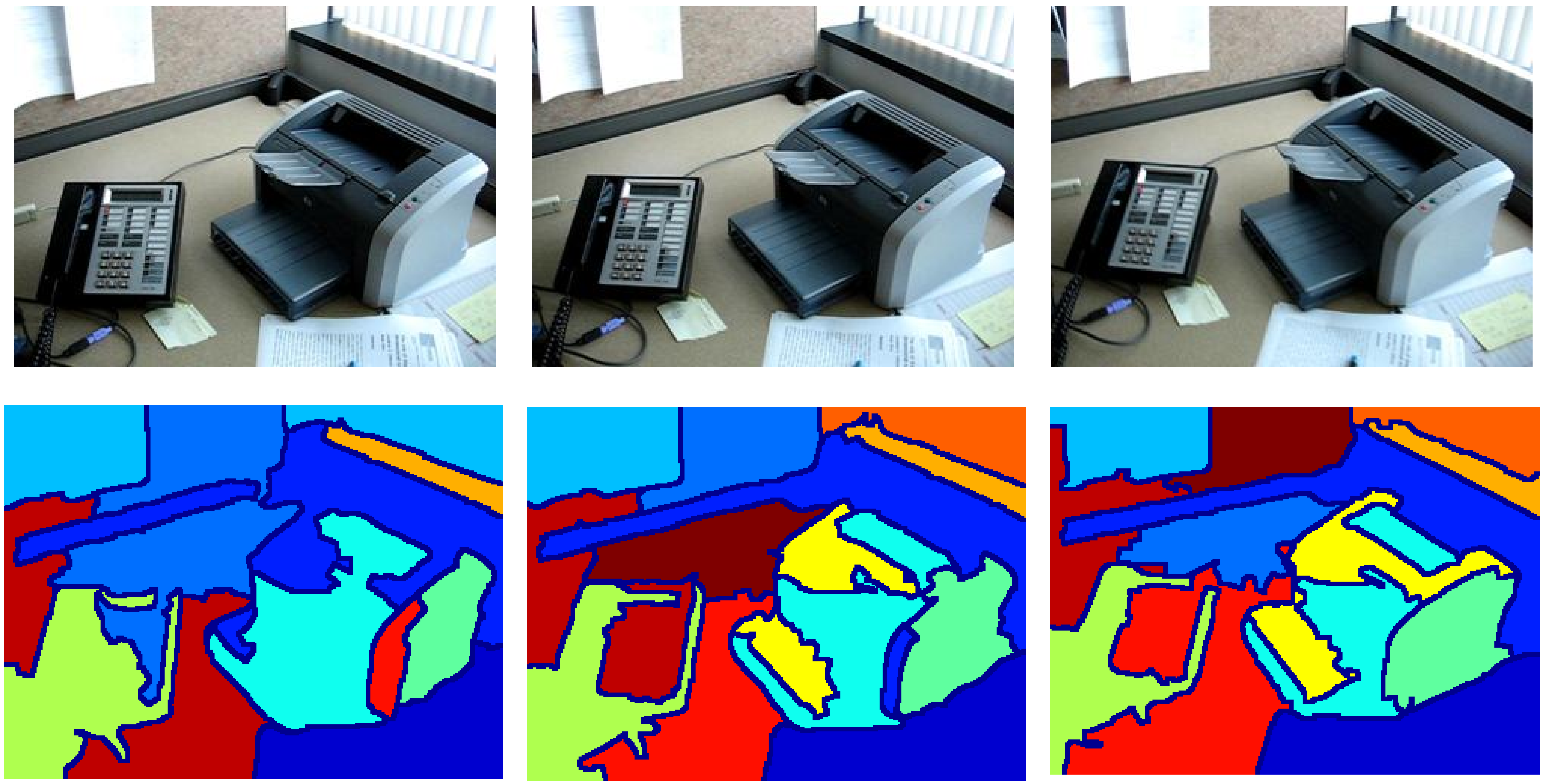


HGVS - Grundman’10

Rand Index - 0.49

Video Segmentation Results

Frame 1 **Frame 5** **Frame 10**

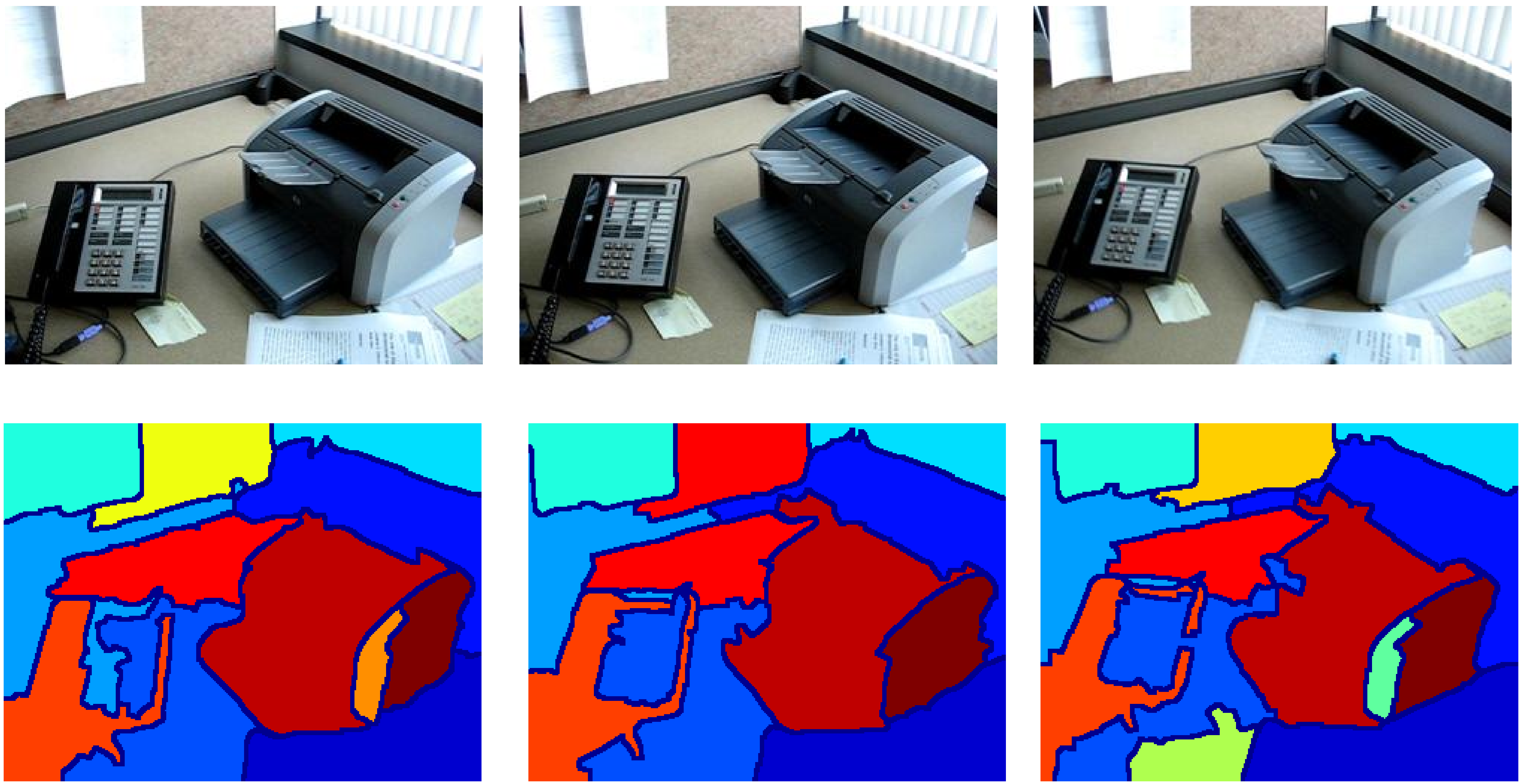


region ddCRP

Rand Index - 0.53

Video Segmentation Results

Frame 1 **Frame 5** **Frame 10**

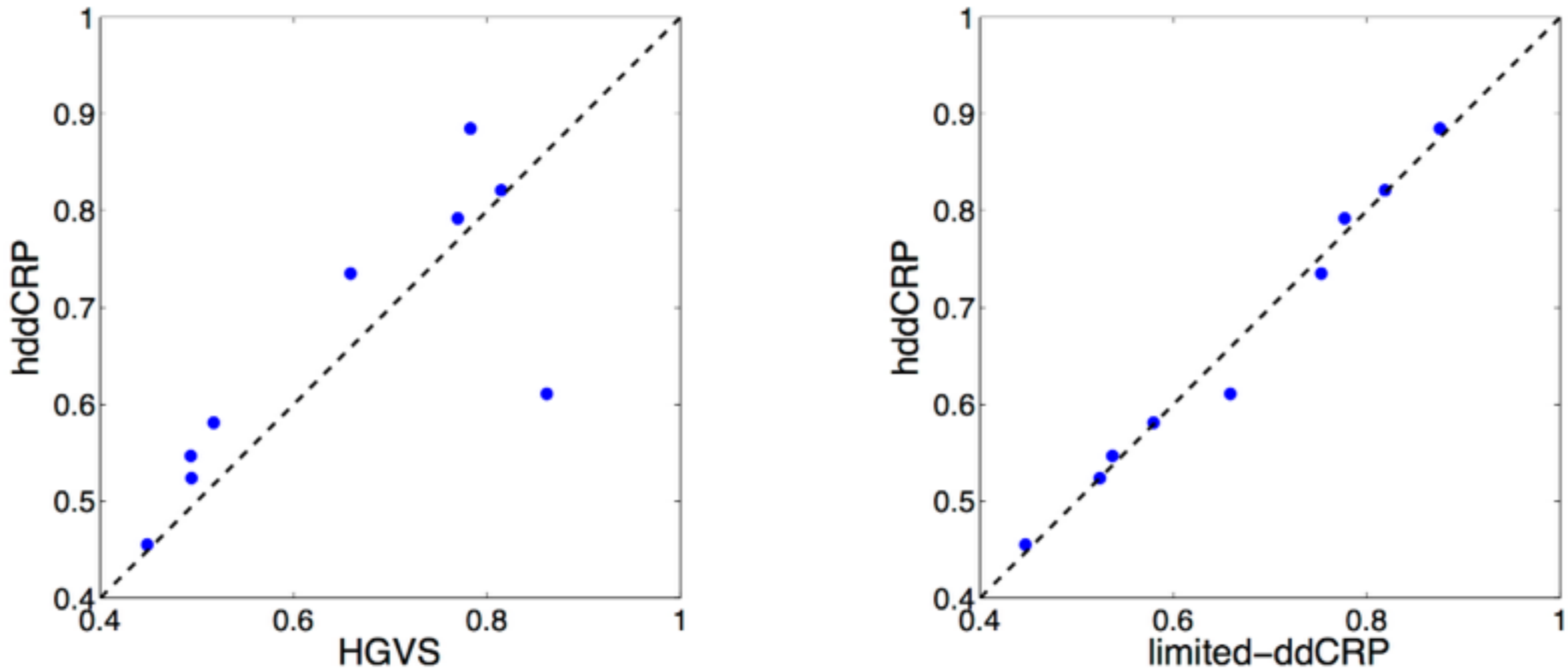


hddCRP

Rand Index - 0.55

Quantitative Performance

- Nine human segmented video sequences. (Liu et al., 08)



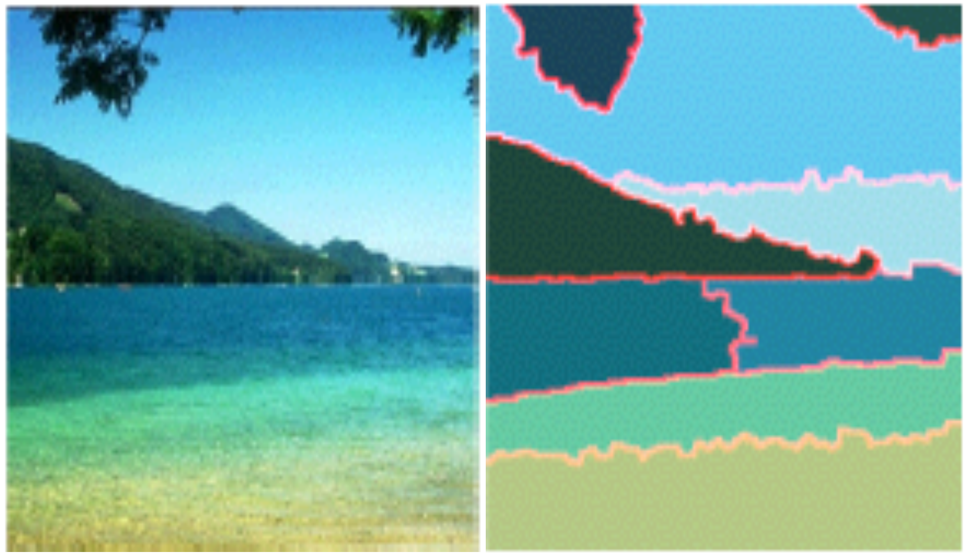
Rand Index — Higher = Better

- **Proposed Work:**
  - Explore whether the modest improvement (over rddCRP) is an inference problem or a modeling issue.
  - Use fancier inter-table distances.

Summary

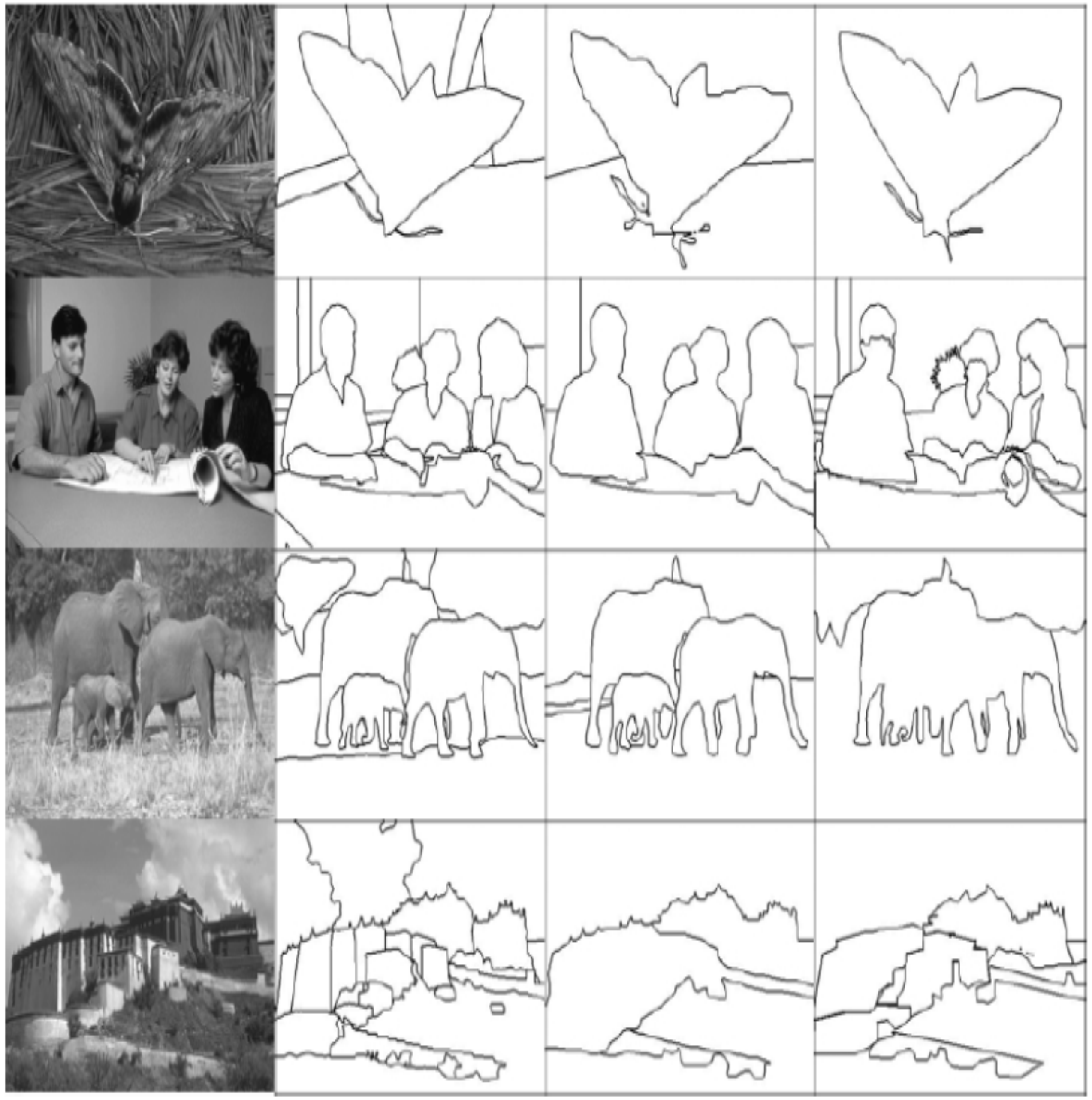
- ddCRP mixture of local regression models for articulated object segmentation

- "region" ddCRP and the generalized hierarchical ddCRP for image and video segmentation



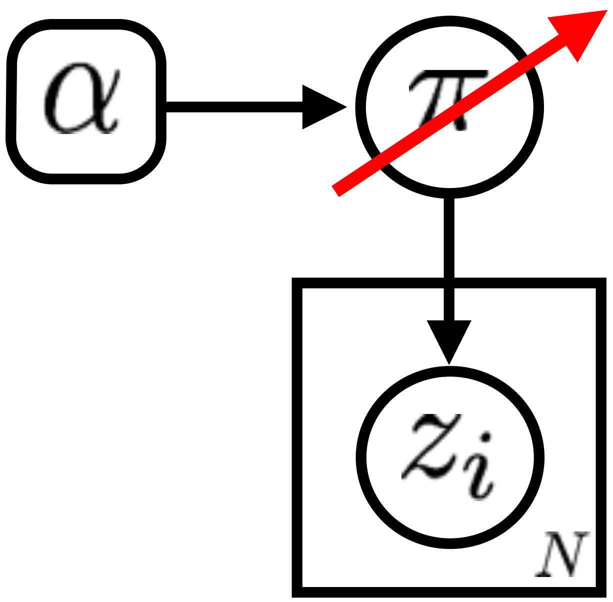
Statistics of Human Segments

- Human segment sizes follow power law behavior.



Sudderth & Jordan, 2008

Influencing Segment Sizes



$$\pi = \{\pi_1, \pi_2 \ldots \pi_\infty\}$$

$$\sum_{k=1}^{\infty} \pi_k = 1 \quad 0 \leq \pi_k \leq 1$$

$$z_i \sim \pi$$

Distribution over partitions - Generalized CRP

$$p(Z \mid \alpha) = \int p(\pi \mid \alpha) \prod_i p(z_i \mid \pi) d\pi$$

Pitman-Yor Process

- The Pitman-Yor process defines a distribution on infinite discrete measures, or partitions

$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \sim \mathrm{PY}(\alpha, H)$$

$$\theta_k \sim \mathrm{H}(\lambda)$$

Power Law Behavior:

Expected size of sorted component k:

$$E[\pi_k] \propto O(k^{-\frac{1}{\alpha_a}})$$

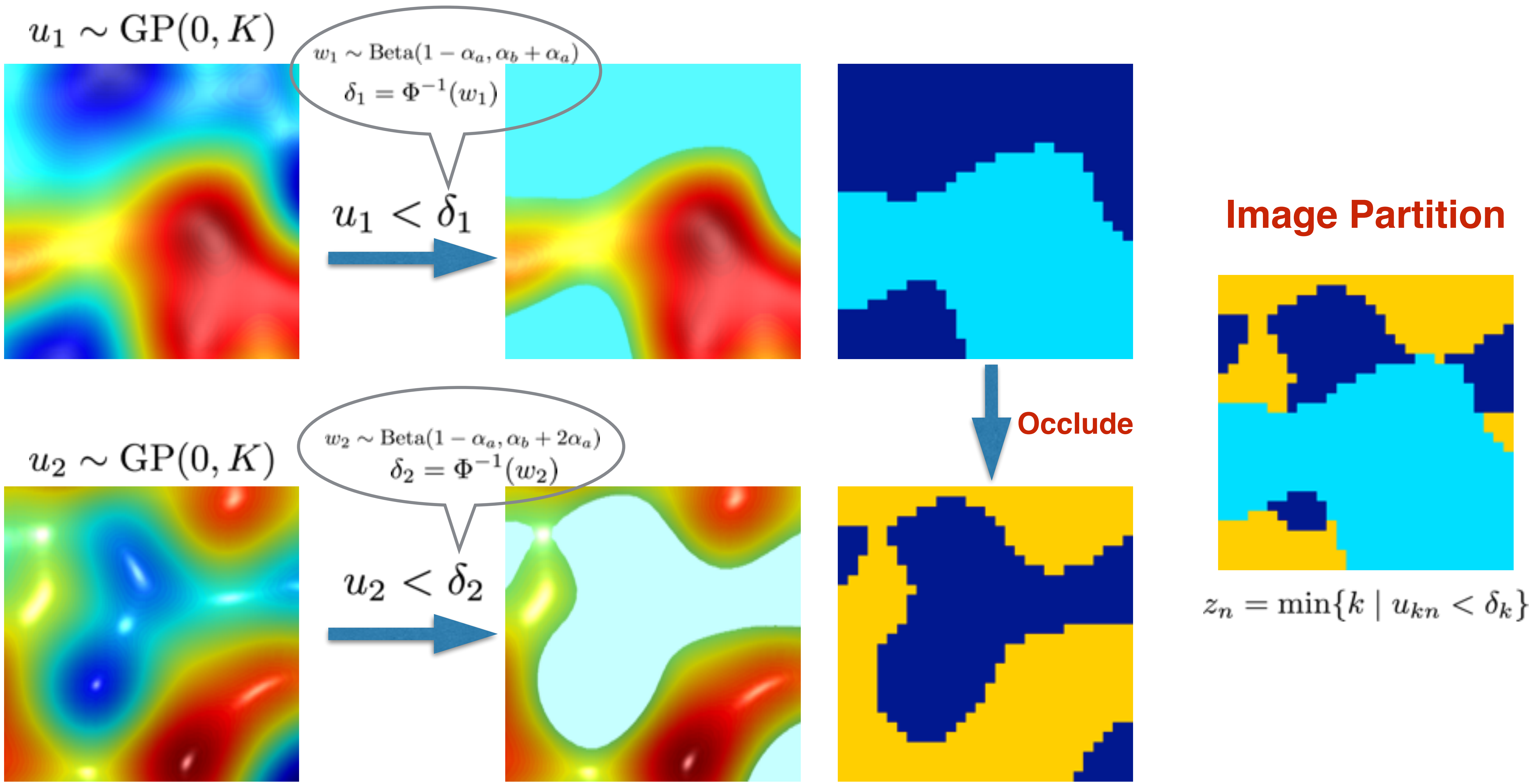
- Use as a distribution over segment sizes.

Talk Outline

- Deformable 3D object segmentation
- Image and Video segmentation
- Layered decomposition of natural images
- Proposed Work

Spatial Coupling through Layers

Smooth Layers **Thresholded layer support**

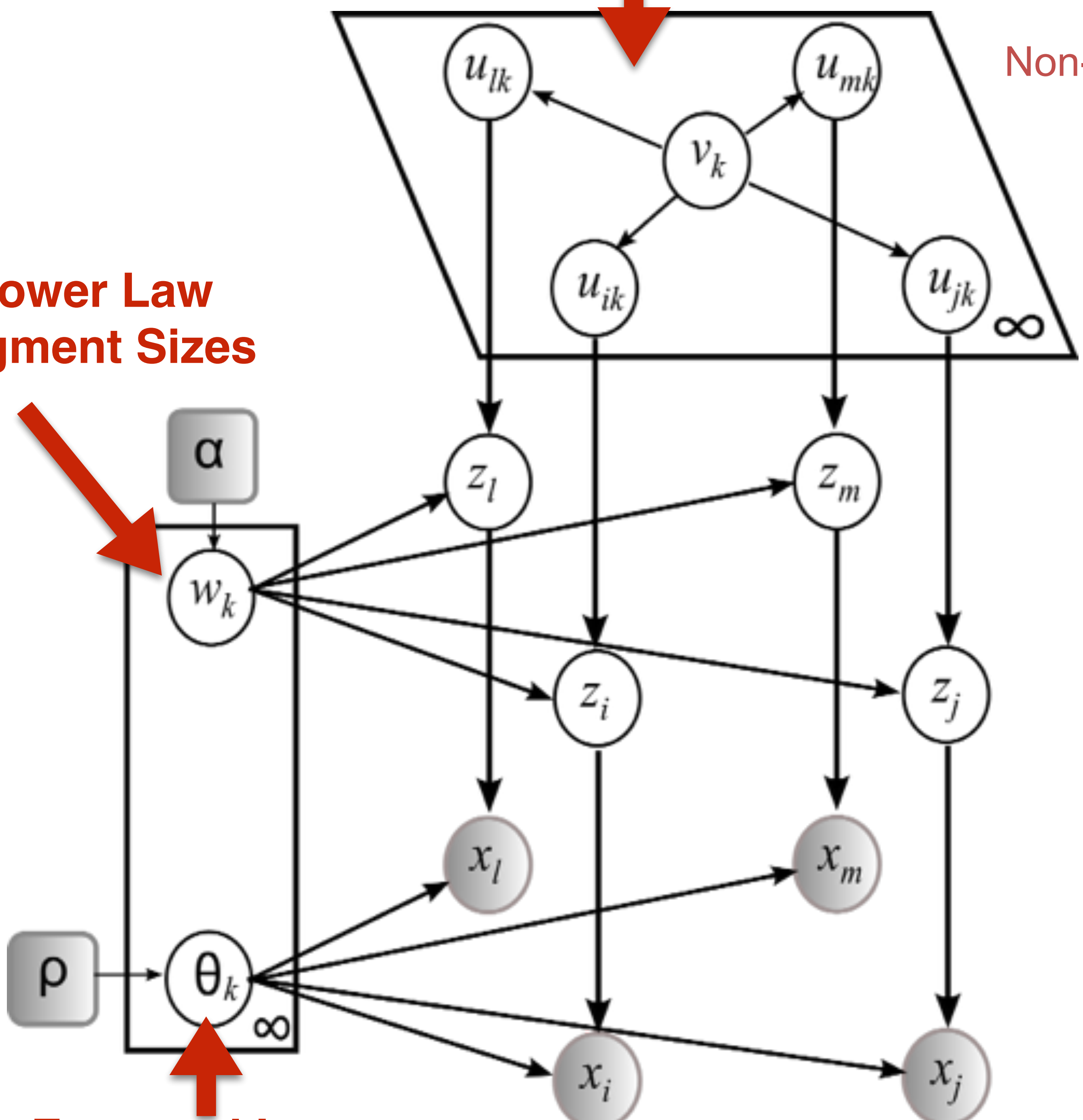


Sudderth & Jordan, 2008
Ghosh & Sudderth, 2012

Spatially Coupled PY Processes

Model Long Range Spatial Correlations

Power Law Segment Sizes



Expected Layer Appearance

Non-Markov Gaussian Processes: Spatial Dependence

$$\mathbf{u}_k = A\mathbf{v}_k + \epsilon_k \quad \mathbf{v}_k \sim \mathcal{N}(0, \mathbf{I}_D)$$

$$u_{kn} \sim \mathcal{N}(0,1) \quad \epsilon_k \sim \mathcal{N}(0, \Psi)$$

Pitman-Yor prior: Segment sizes follow a *power law* distribution

$$\Pr[z_n = k] = w_k \prod_{\ell=1}^{k-1} (1 - w_\ell)$$

$$w_k \sim \mathrm{Beta}(1 - \alpha_a, \alpha_b + k\alpha_a)$$

Segment Assignments: Layers

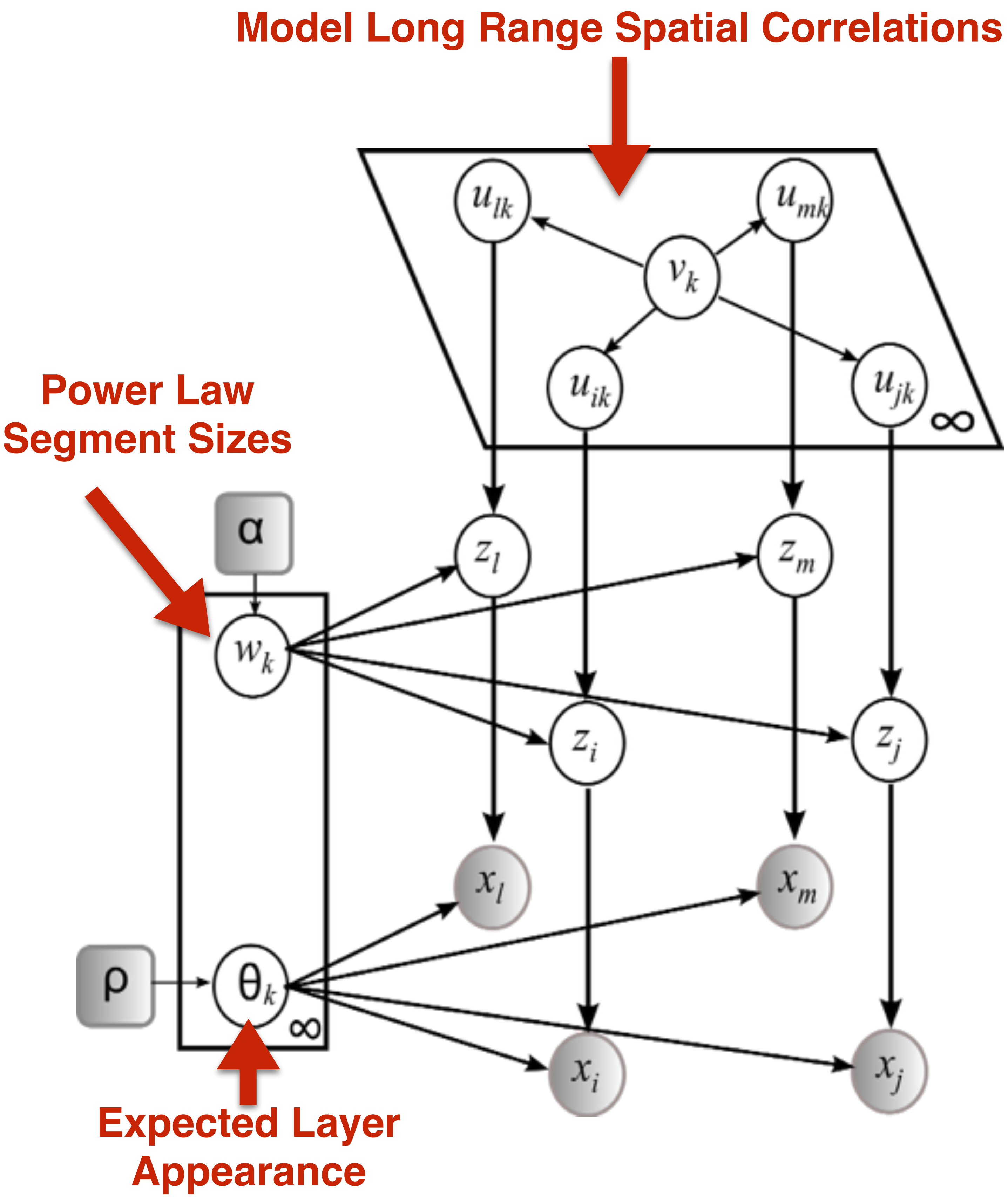
$$z_n = \min\{k \mid u_{kn} < \Phi^{-1}(w_k)\}$$

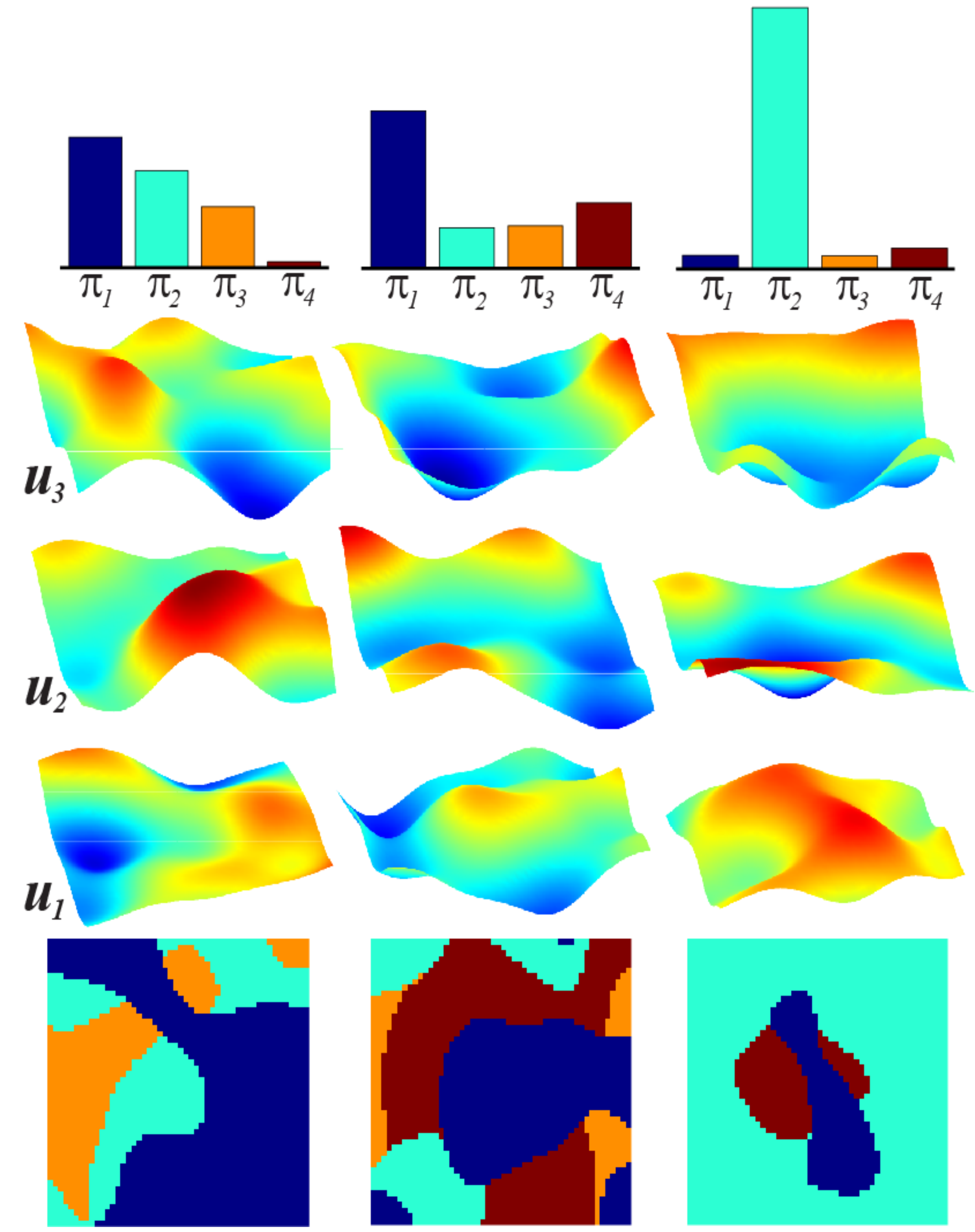
Features: Color & Texture

$$x_n^c \sim \mathrm{Mult}(\theta_{z_n}^c)$$

$$x_n^t \sim \mathrm{Mult}(\theta_{z_n}^t)$$

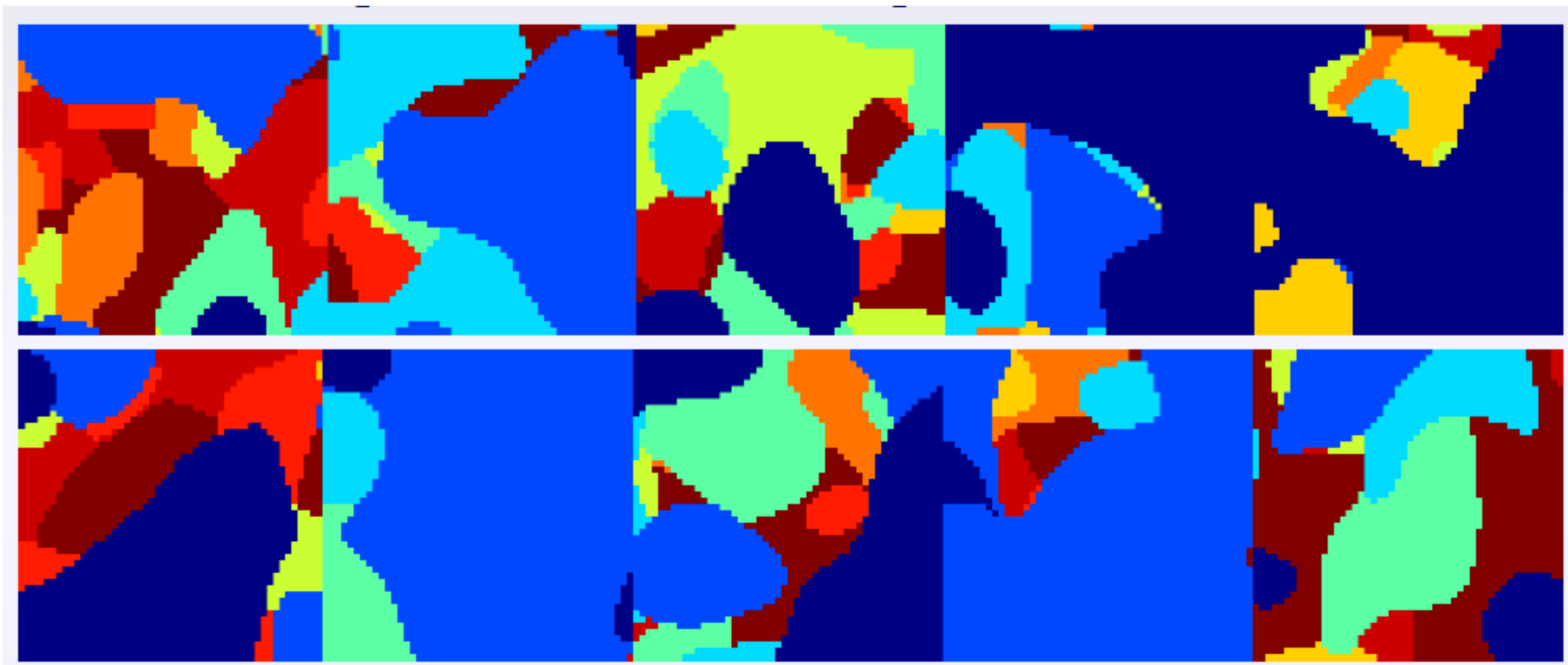
Spatially Coupled PY Processes



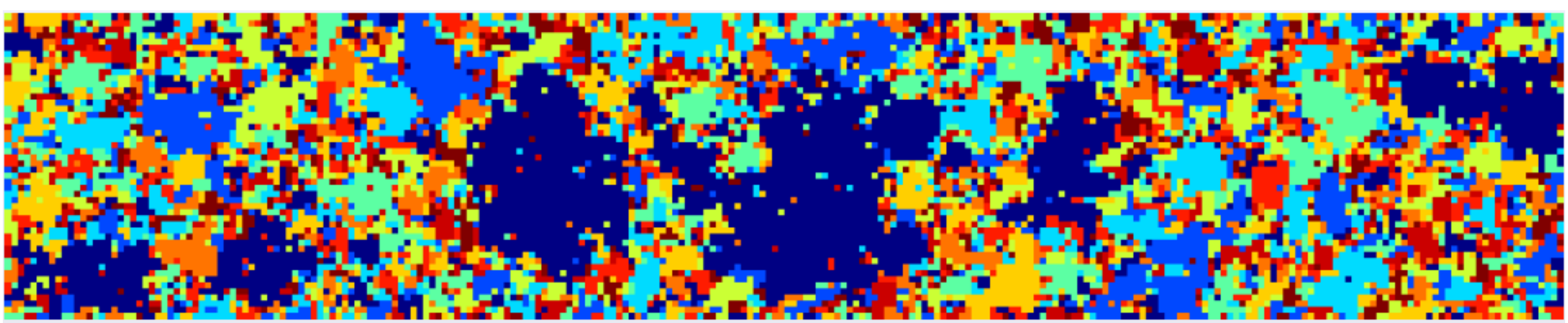


Sudderth & Jordan, 2008

Generative Samples from the Model

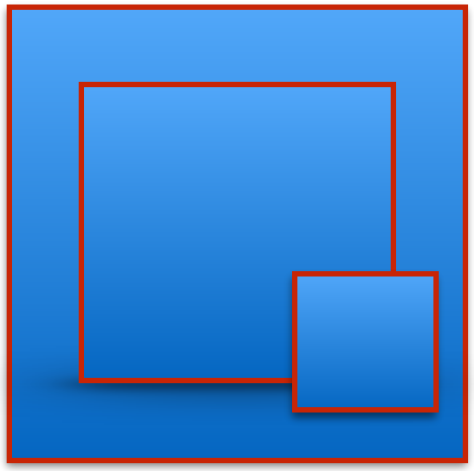


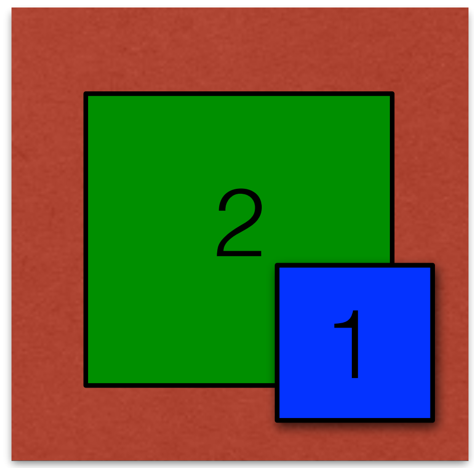
Samples from a Potts model:

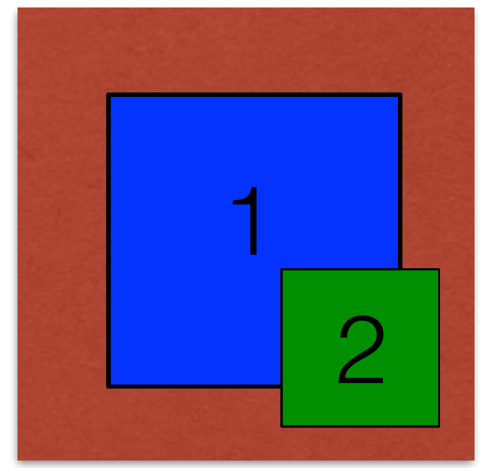


Depth-ordered Layers

Depth Order





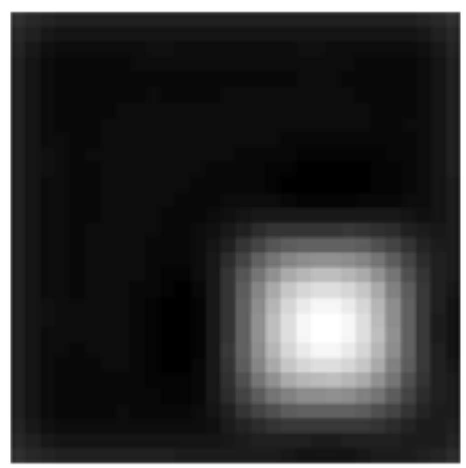


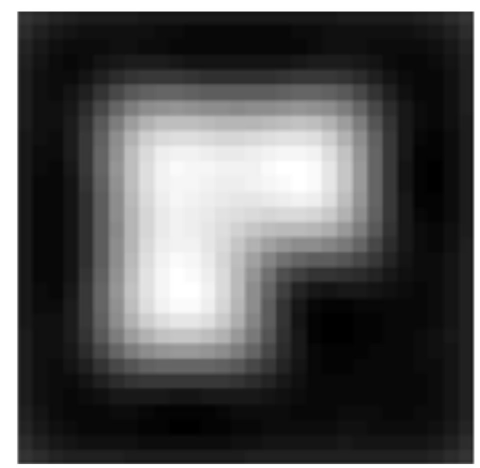
Far

Intermediate

Near

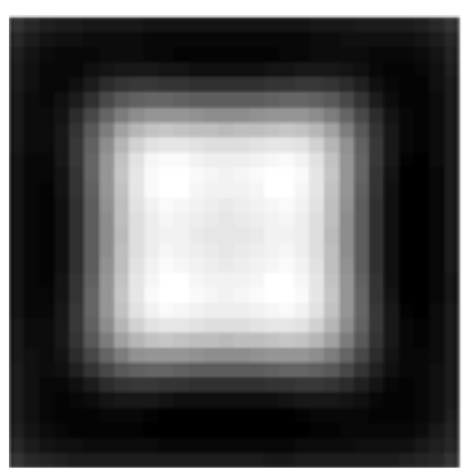
Expected Near Layer:

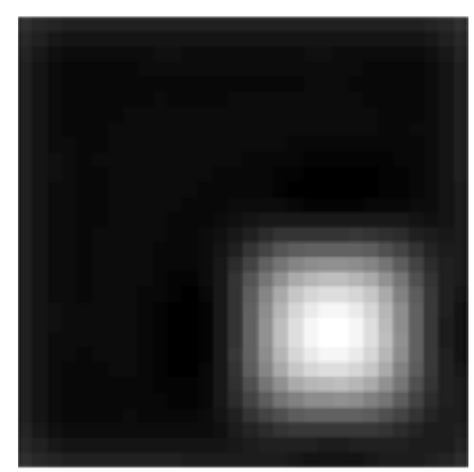




Lighter Shades = Higher values

Expected Int. Layer:

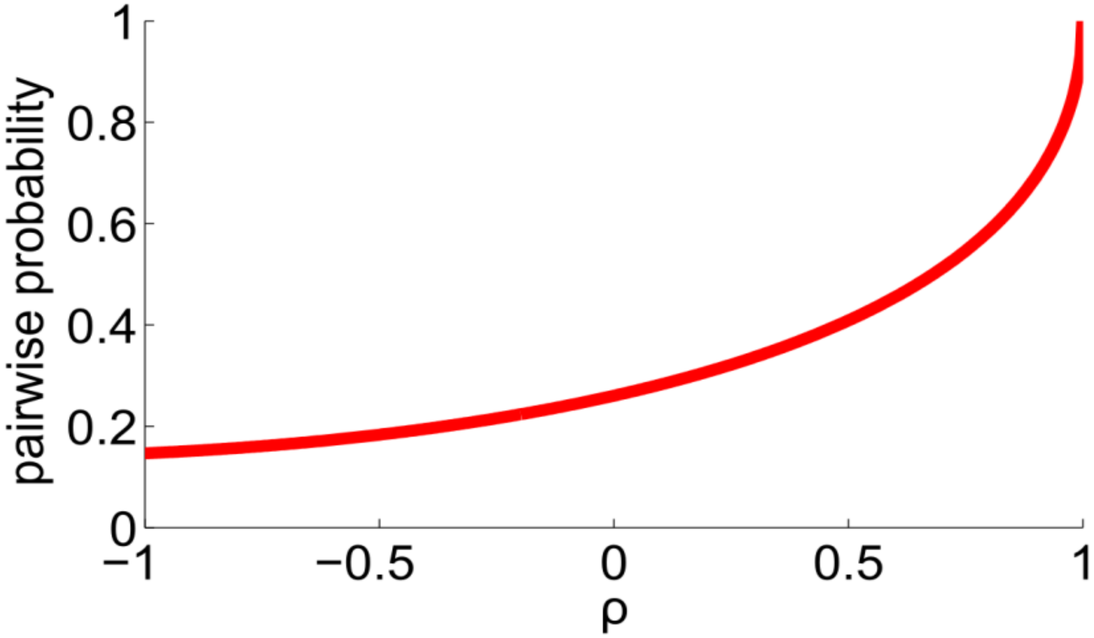




Lower Probability Partition

Calibrating Gaussian Processes

- *Binary Classification:* Probability of two super-pixels belonging to the same segment, learned via regularized logistic regression.
- *Co-occurrence to covariance:* Injective mapping between probabilities and GP correlation
- *Valid Model:* Project independently learned covariance matrix to the closest PSD matrix.



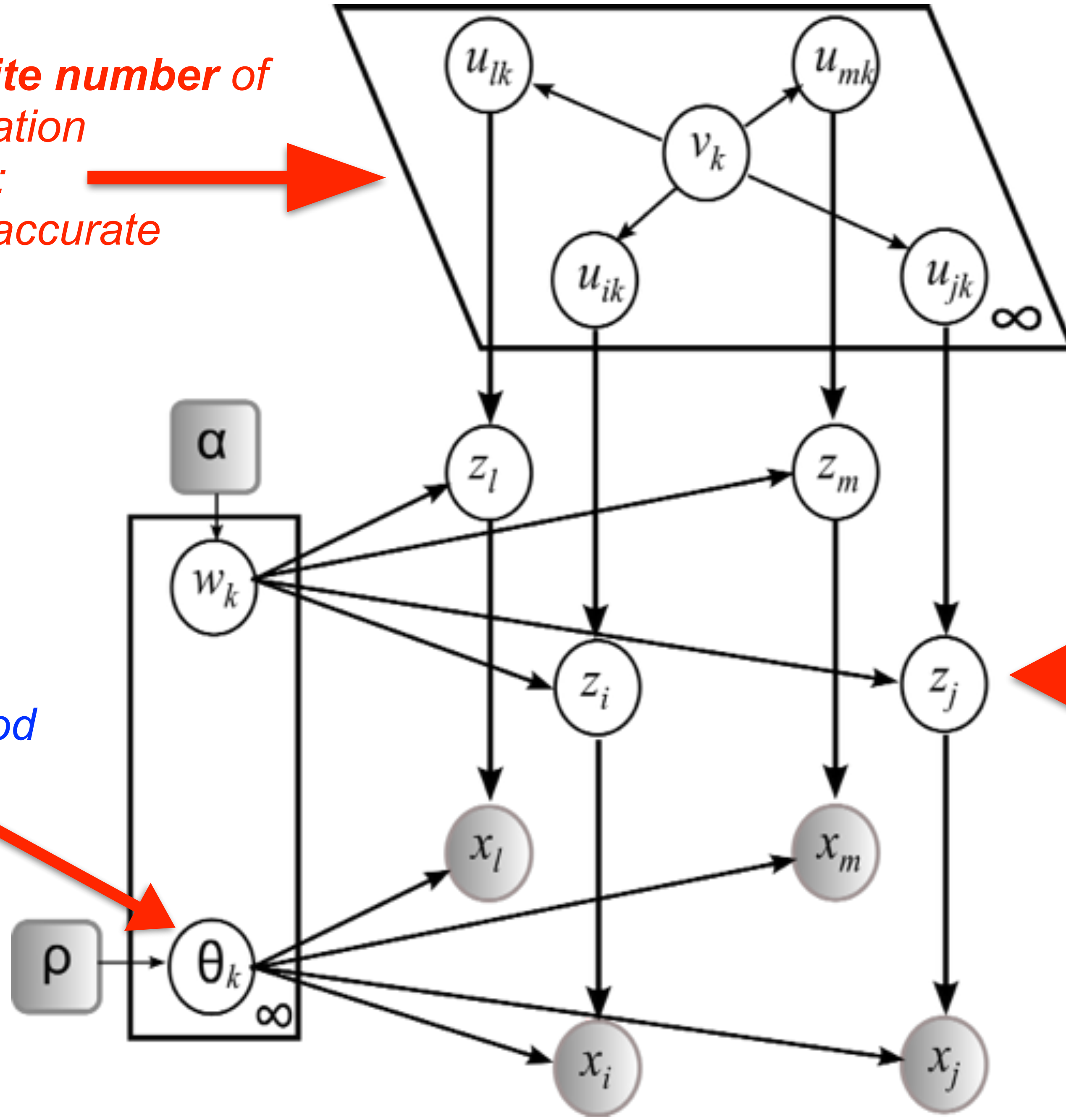
Ghosh & Sudderth, 2012

Inference via Stochastic Search

*Marginalize a **finite number** of layers via expectation propagation (EP): approximate but accurate*

Consider hard assignments of super-pixels to layers

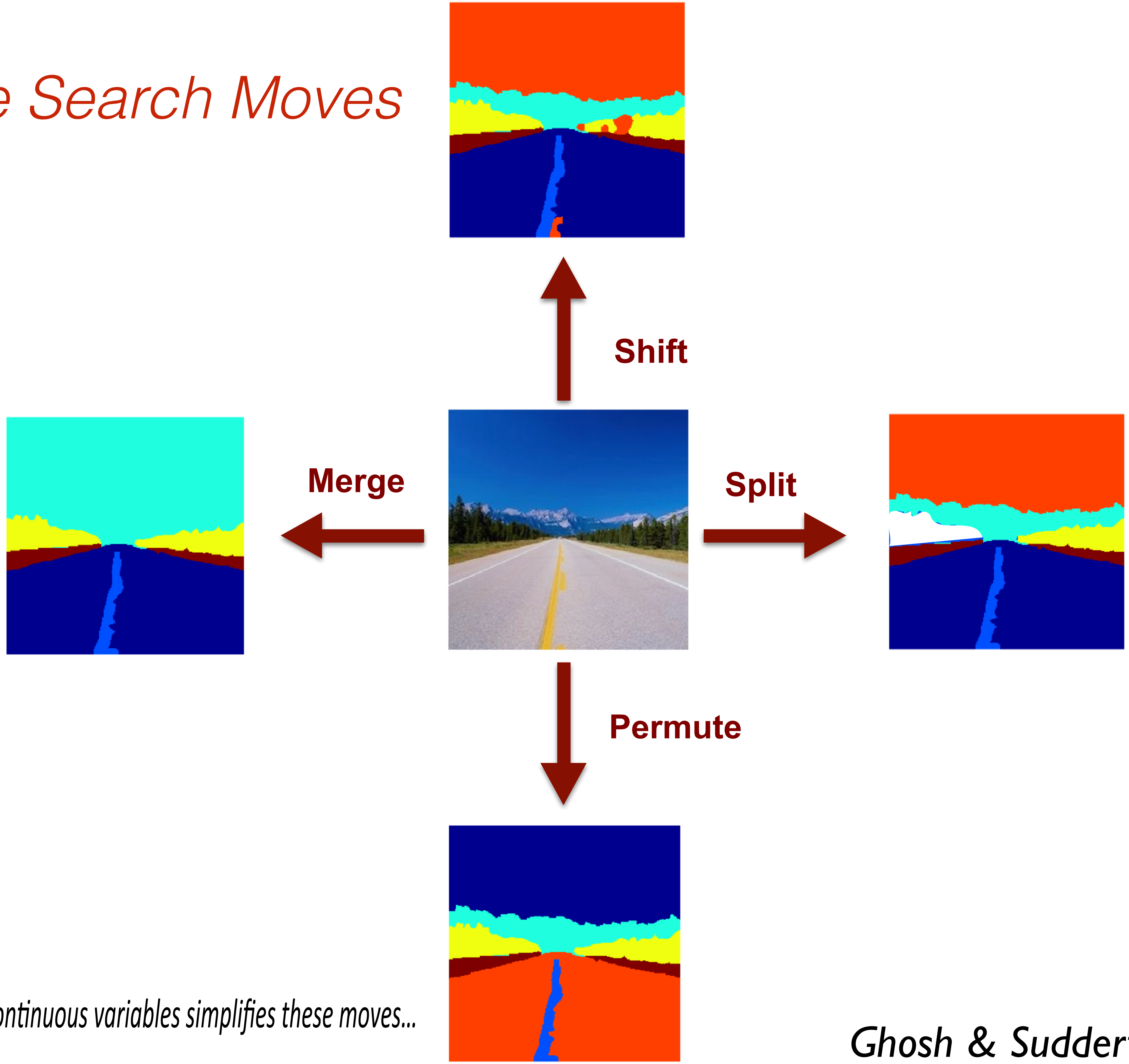
Integrate likelihood parameters analytically (conjugacy)



$$p(Z \mid X, \alpha, \rho) \propto p(X \mid Z, \rho) \underbrace{p(Z \mid \alpha)}_{\text{EP}}$$

Ghosh & Sudderth, 2012

Discrete Search Moves

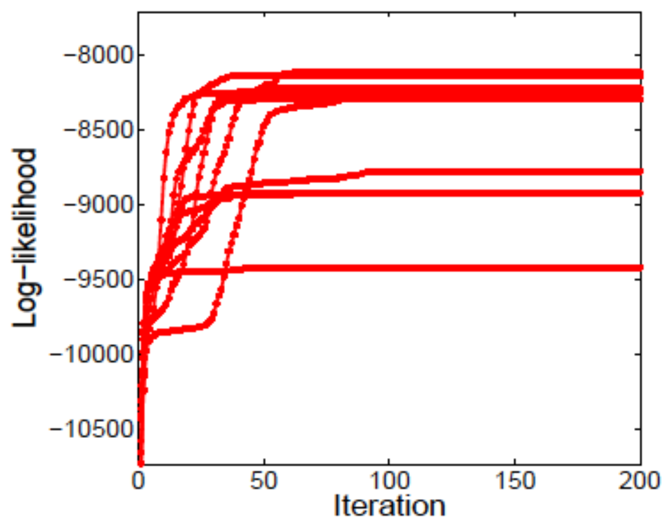


Marginalization of continuous variables simplifies these moves...

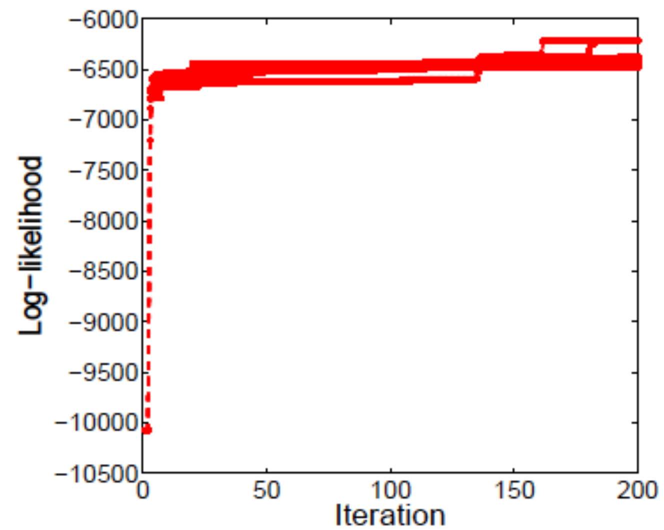
Ghosh & Sudderth, 2012

Robust, Reliable Inference

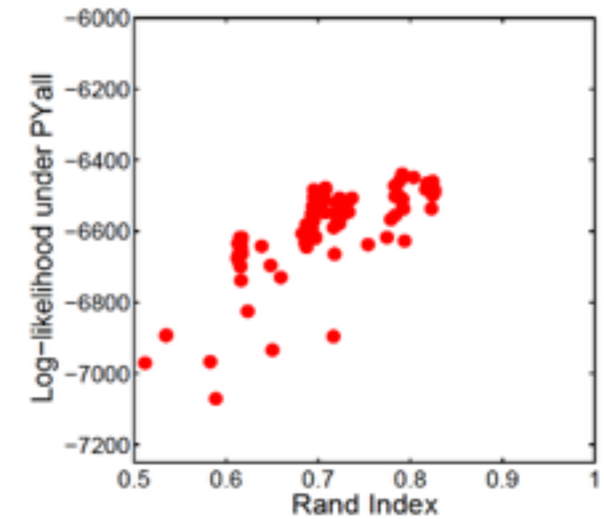
Mean Field Variational



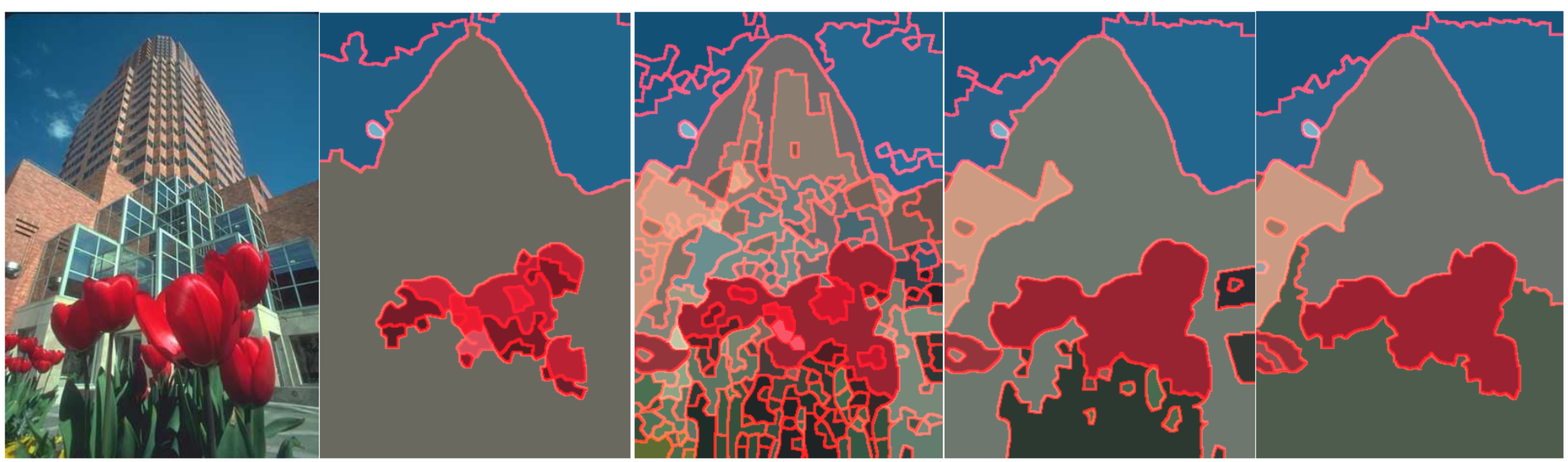
Stochastic Search EP



Rand Index vs Model Likelihood



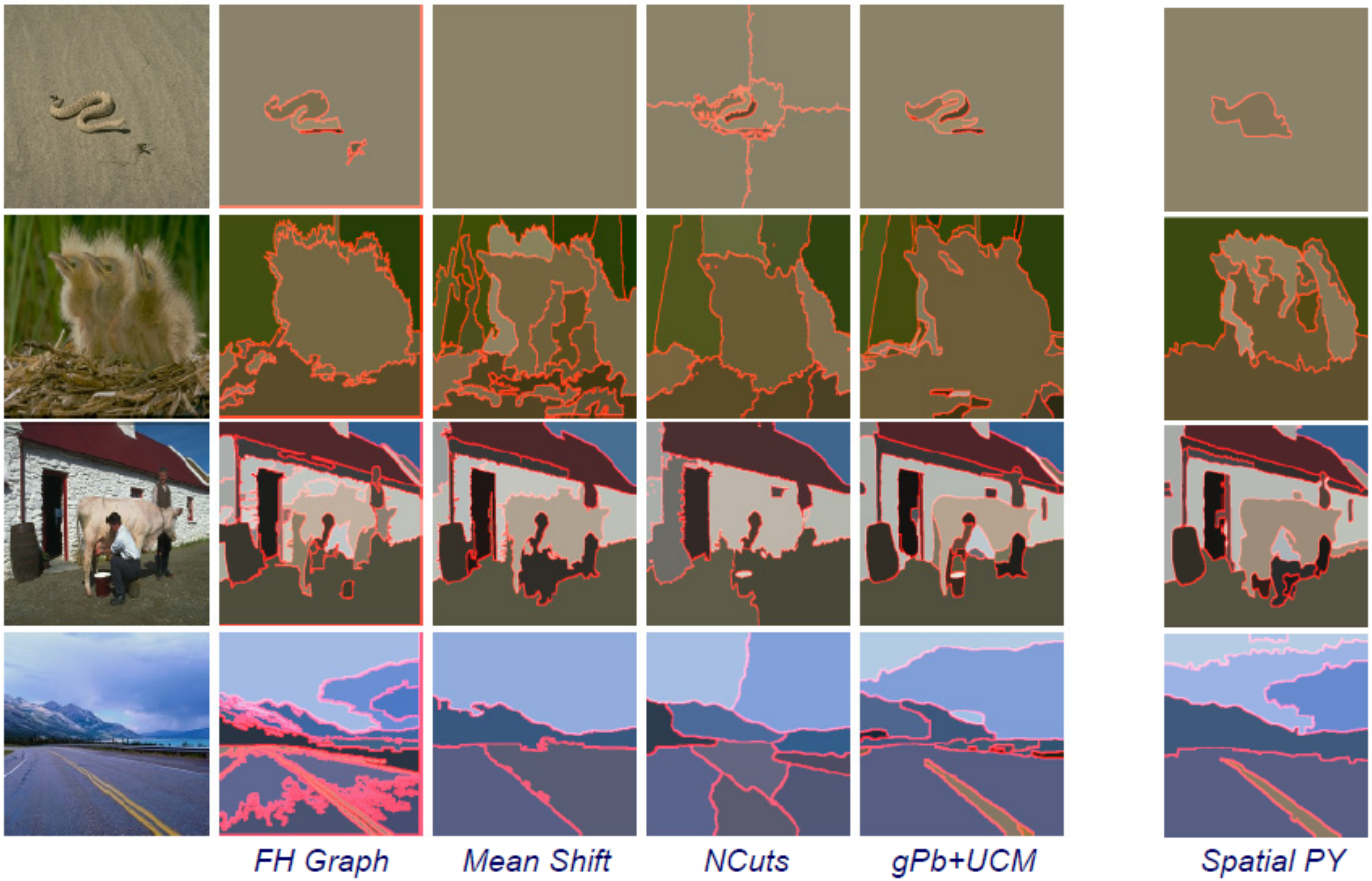
External Model Validation



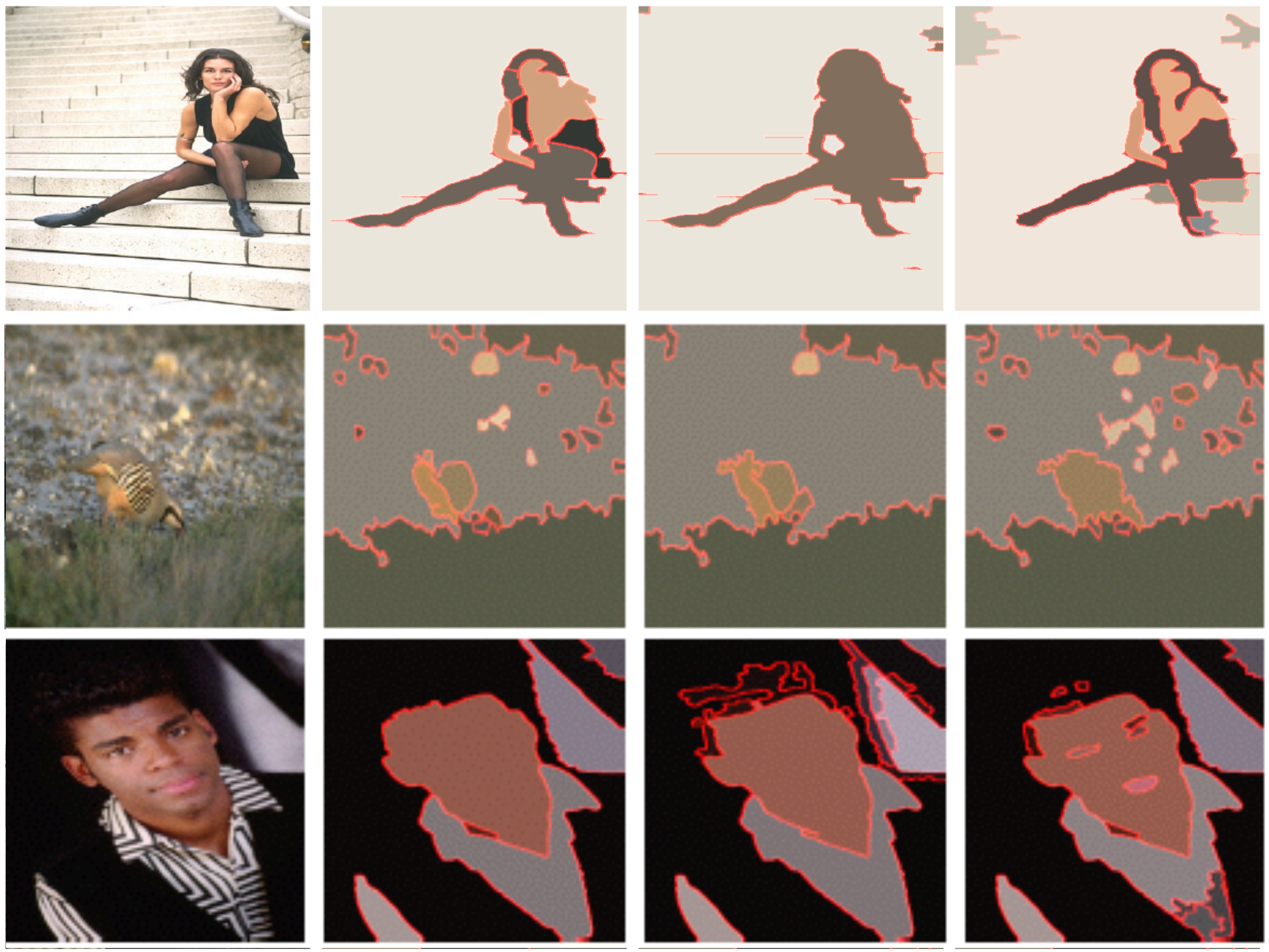
Best *Worst*
Mean Field Variational

Best *Worst*
Stochastic Search EP

Qualitative Results



Posterior distribution over segmentations



Most Probable

Quantitative Performance

Algorithms	PRI	VI	SegCover
Ncuts	0.74	2.5	0.38
MS	0.77	2.5	0.44
FH	0.77	2.1	0.52
gPb	0.81	2.0	0.58
PYdist	0.72	2.1	0.51
PYall	0.76	2.1	0.52

Berkeley Segmentation

gPb	0.74	2.1	0.53
PYall	0.73	1.9	0.55

LabelMe Scenes

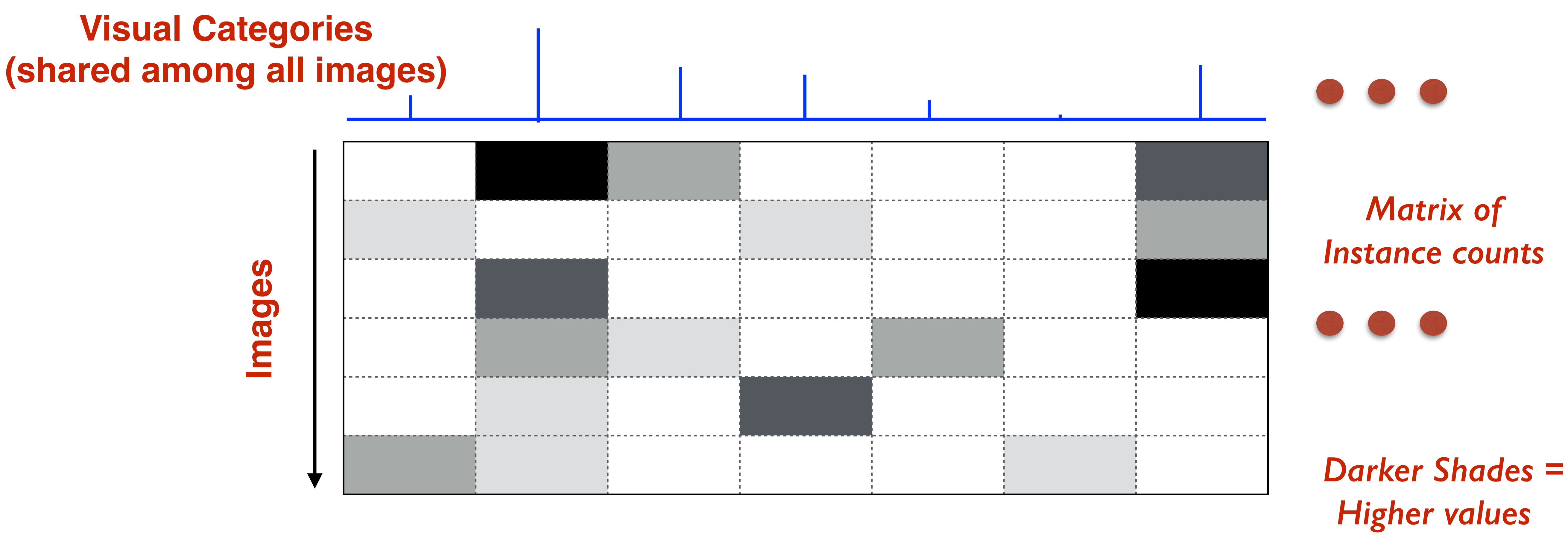
- On BSDS, similar or better than all methods except gPb
- On LabelMe, performance of Spatial PY is better than gPb

Talk Outline

- Deformable 3D object segmentation
- Image and Video segmentation
- Layered decomposition of natural images
- Proposed Work

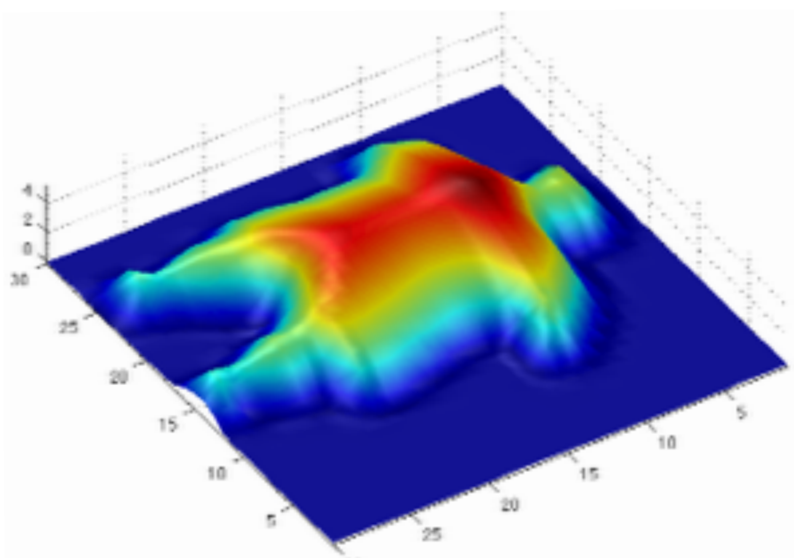
Shared Analysis of Image Corpuses

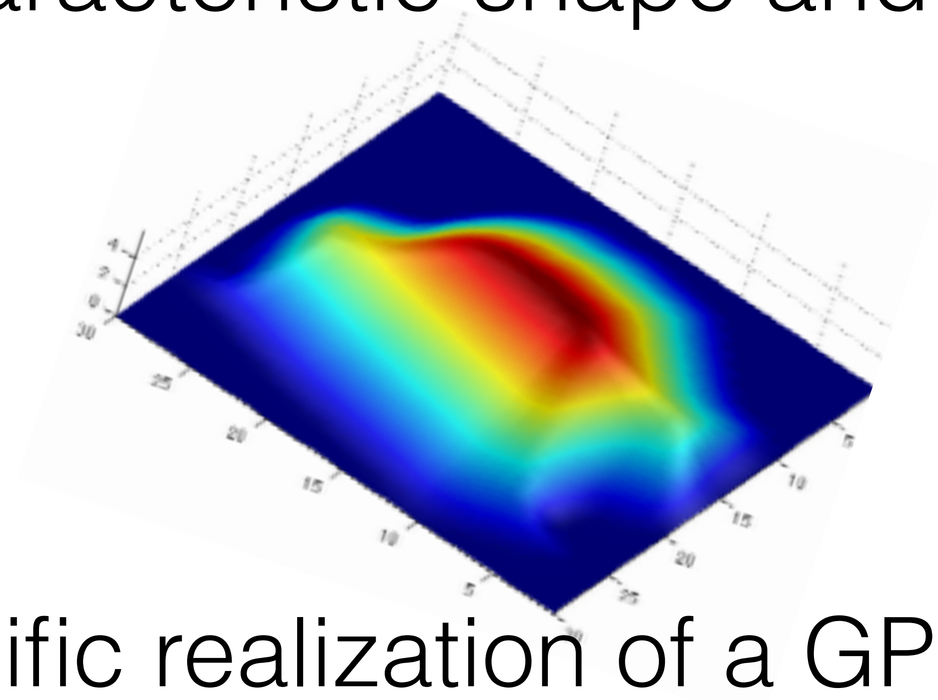
- Joint image parsing and labeling
- Simultaneous inference of coarse 3D structure — placement of layers in the 3D scene
- Explicit modeling of object and instance occurrence through the Beta-Geometric process:



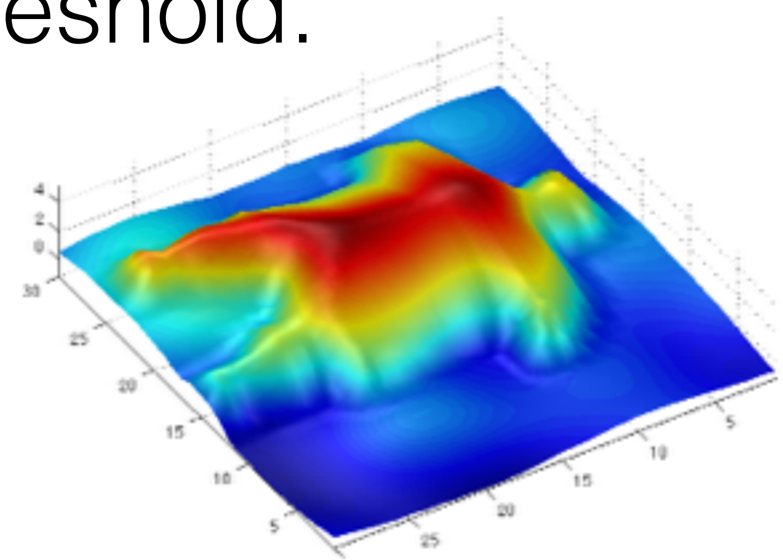
Category specific shapes and sizes

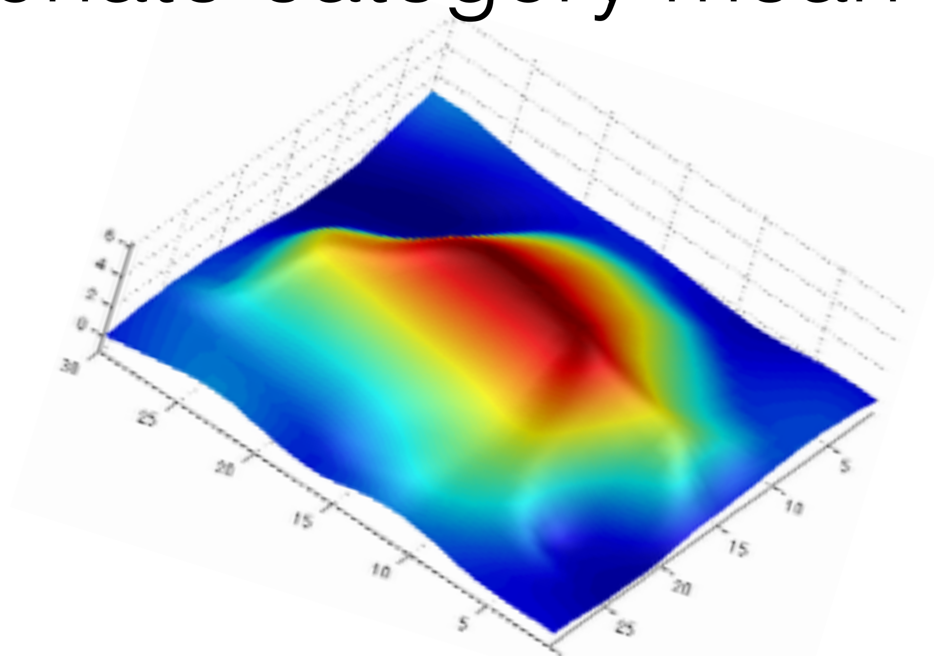
- Each visual category is embellished with a mean function modeling its characteristic shape and size





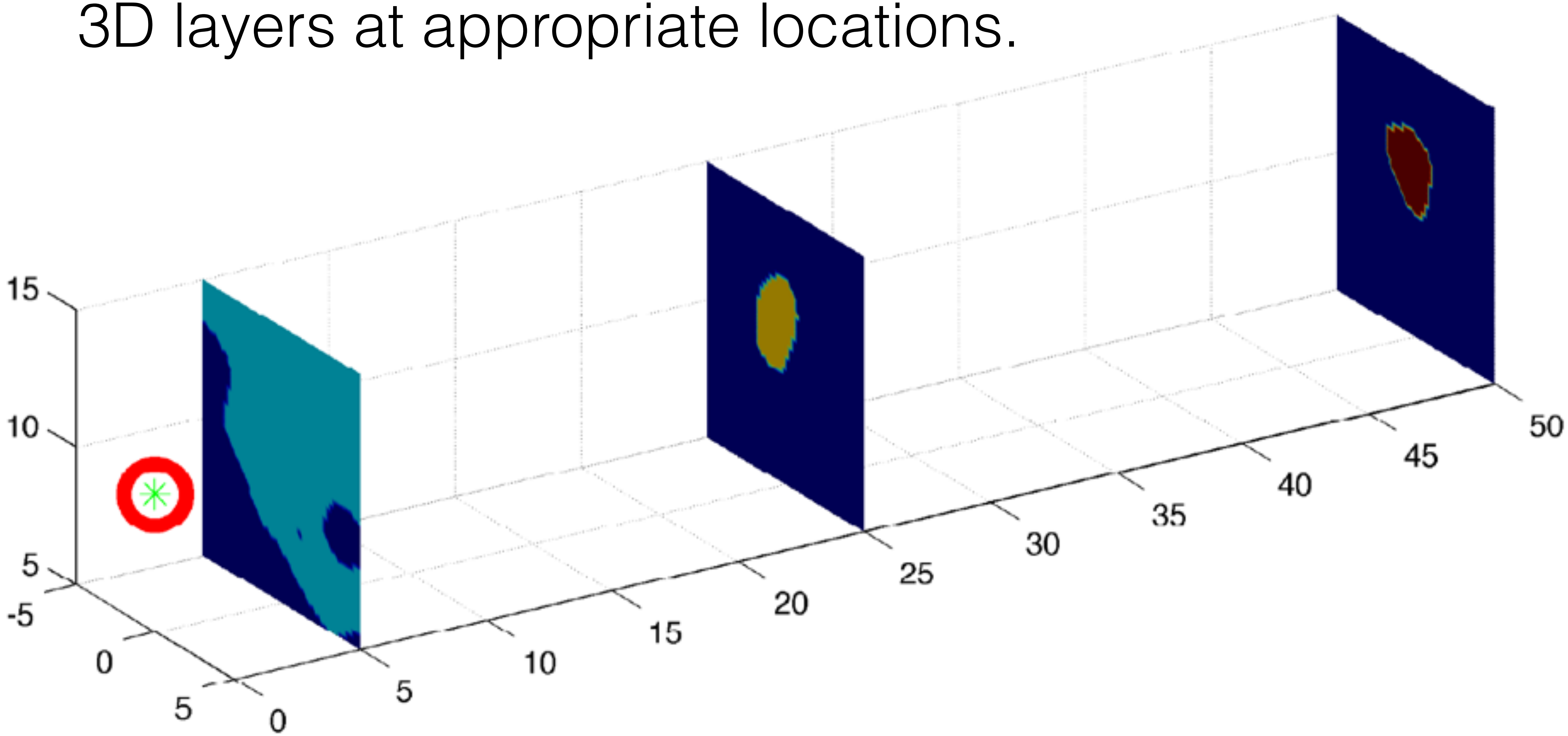
- Sample an instance specific realization of a GP parametrized with the appropriate category mean and threshold.





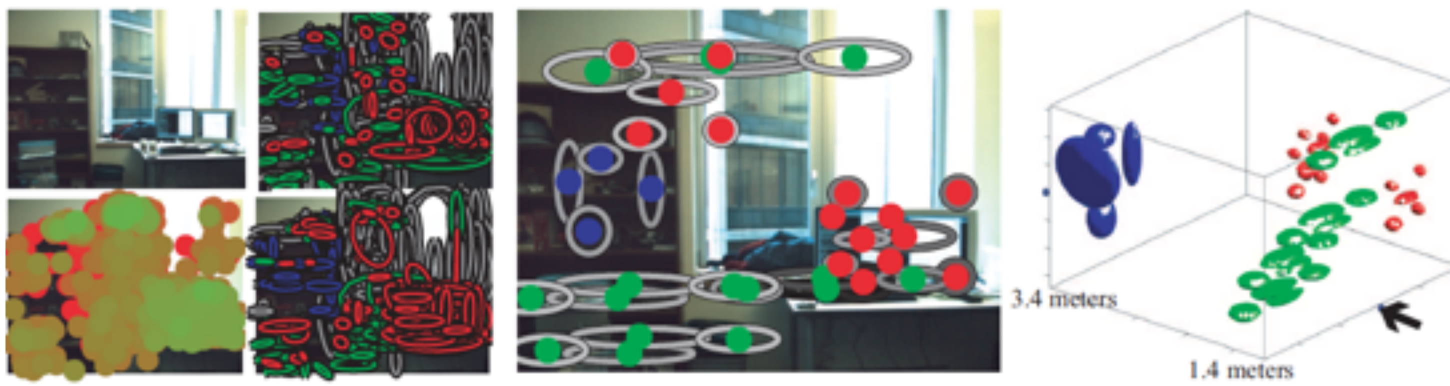
2D Images from 3D structure

- Sample a 3D location and place instance specific 3D layers at appropriate locations.

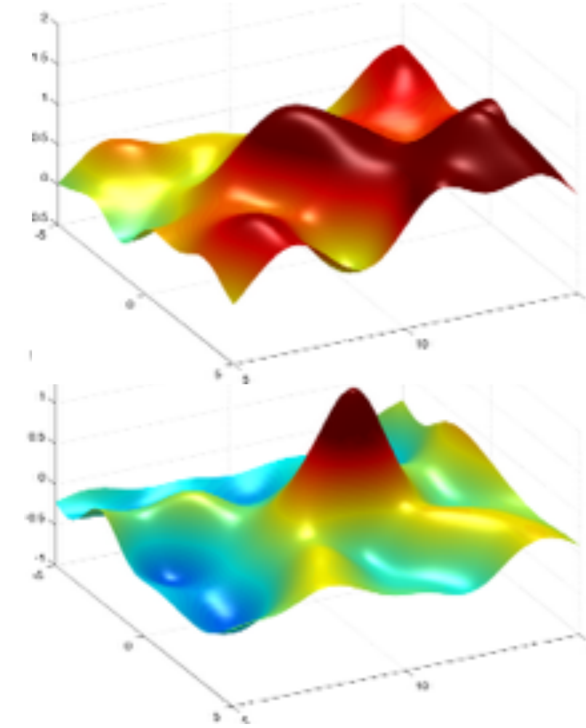


- Project 3D layers into the image through perspective projection.

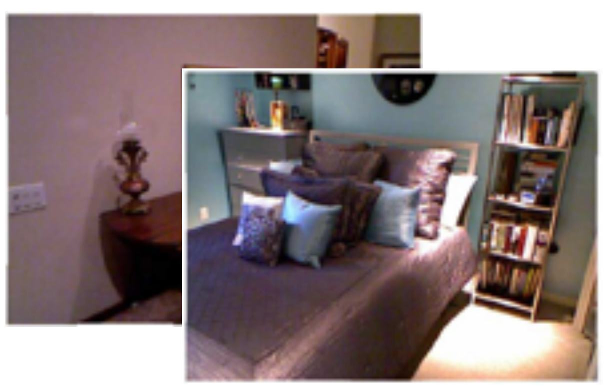
Building Blocks

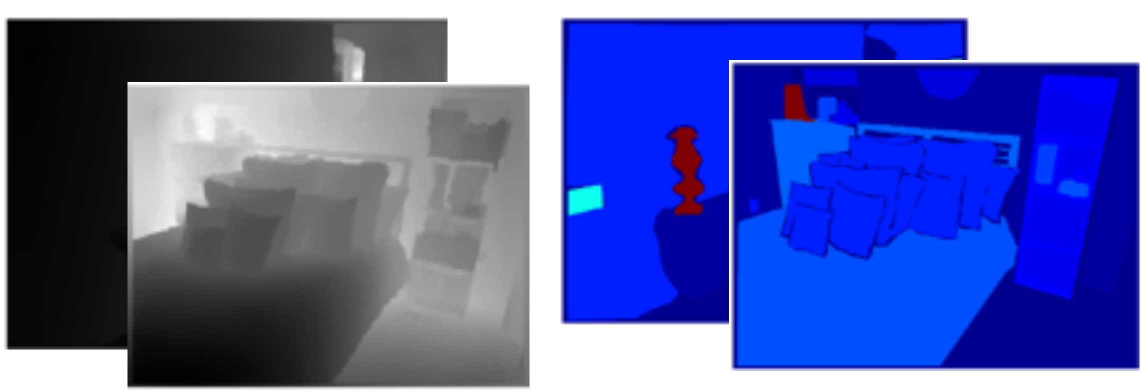


Depth from familiar objects
Sudderth, Torralba, Freeman & Willsky. CVPR'06



BNP Layered models
Ghosh and Sudderth. CVPR'12
Sudderth and Jordan. NIPS'08





Color-depth datasets
Silberman, Kohli, Hoiem & Fergus. ECCV'12

Proposed Work

Timeline

Publications

- Ghosh et al., *Spatial distance dependent Chinese restaurant processes for image segmentation*, **NIPS'11.**
- Ghosh & Sudderth., *Nonparametric Learning for Layered Segmentation of Natural Images*, **CVPR'12.**
- Ghosh et al., *From Deformations to Parts:Motion-based Segmentation of 3D Objects*, **NIPS'12.**

TO DO

- Clean up the work on Generalized Hierarchical ddCRP and submit to ICML 2014
- A parametric version of the shared segmentation model to ECCV 2014
- The Beta-Geometric nonparametric version will go to NIPS 2014
- Summer'2014 tie up loose ends, defend and graduate at the end of summer.

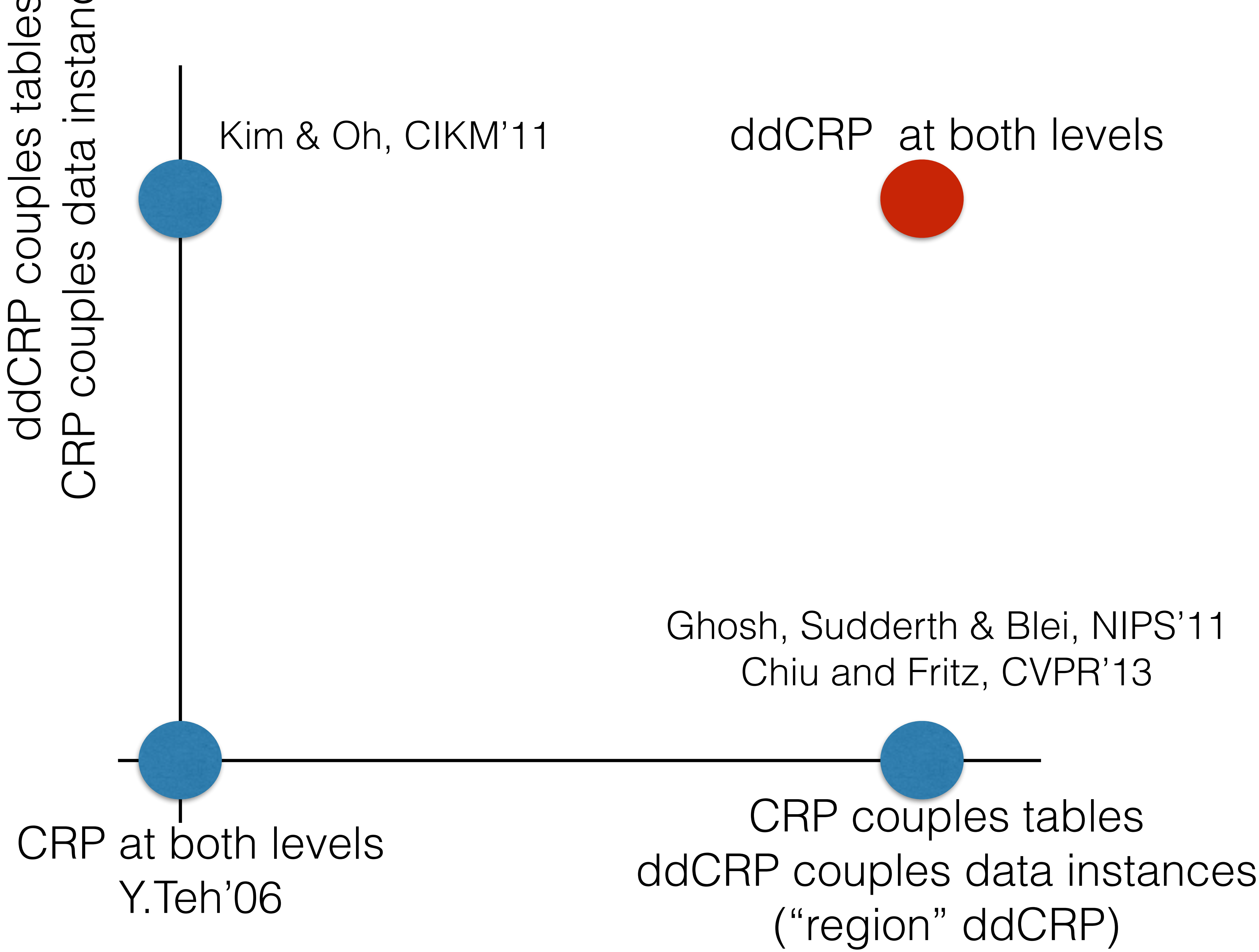
Questions?

Mixture of Regressions

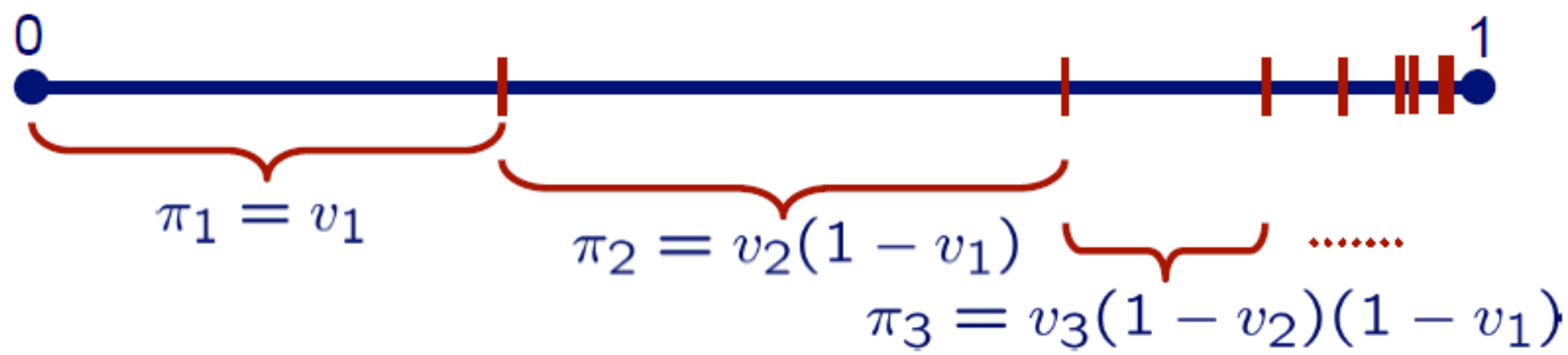
- Model is more general.
- Groups (feature,response) pairs with similar relationships.
- Show examples.

Moderate robustness to alignment errors

Generalized Hierarchical ddCRPs



Stick Breaking to Layers



$$v_k = \mathbb{P}(z_i = k \mid z_i \neq k-1, \ldots, 1)$$

Sequential Binary Sampler:

$$b_{ki} \sim \text{Bernoulli}(v_k)$$

$$z_i = \min\{k \mid b_{ki} = 1\}$$

- *For each data instance i, go through the bins in order 1 through infinity.*
- *Toss a biased coin (with the probability of heads = v_k) for each bin .*
- *Pick the bin if the coin turns up heads*

Stick Breaking to Layers

Sequential Binary Sampler:

$$b_{ki} \sim \text{Bernoulli}(v_k)$$

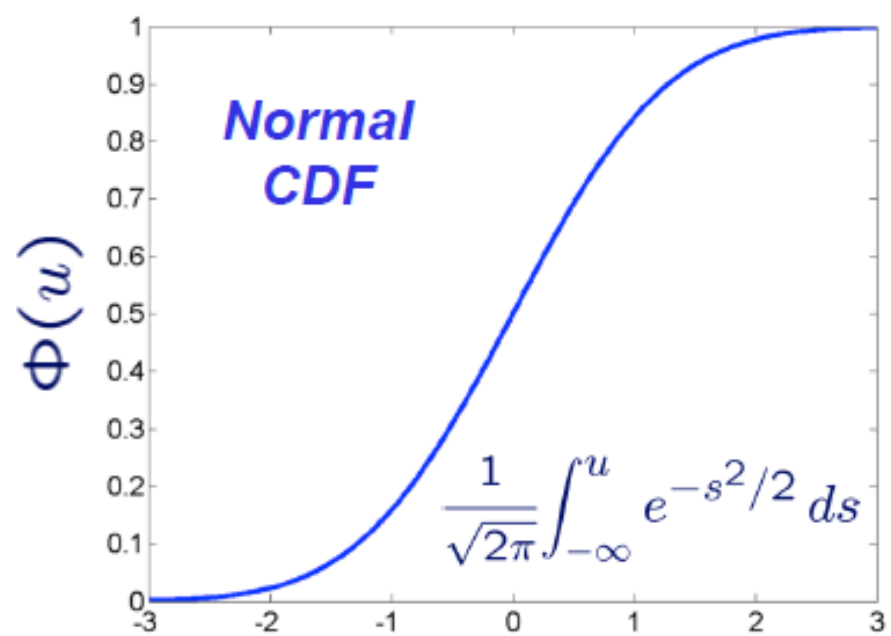
$$z_i = \min\{k \mid b_{ki} = 1\}$$

Gaussian Sampler:

$$u_{ki} \sim \mathcal{N}(0,1)$$

$$z_i = \min\{k \mid u_{ki} < \Phi^{-1}(v_k)\}$$

- *For each data instance, go through the bins in order 1 through infinity.*
- *Sample a unit normal distribution for each bin.*
- *Pick the bin if the sample is less than the threshold* $\Phi^{-1}(v_k)$



$$\mathbb{P}[\Phi(u_{ki}) < v_k] = v_k$$

because

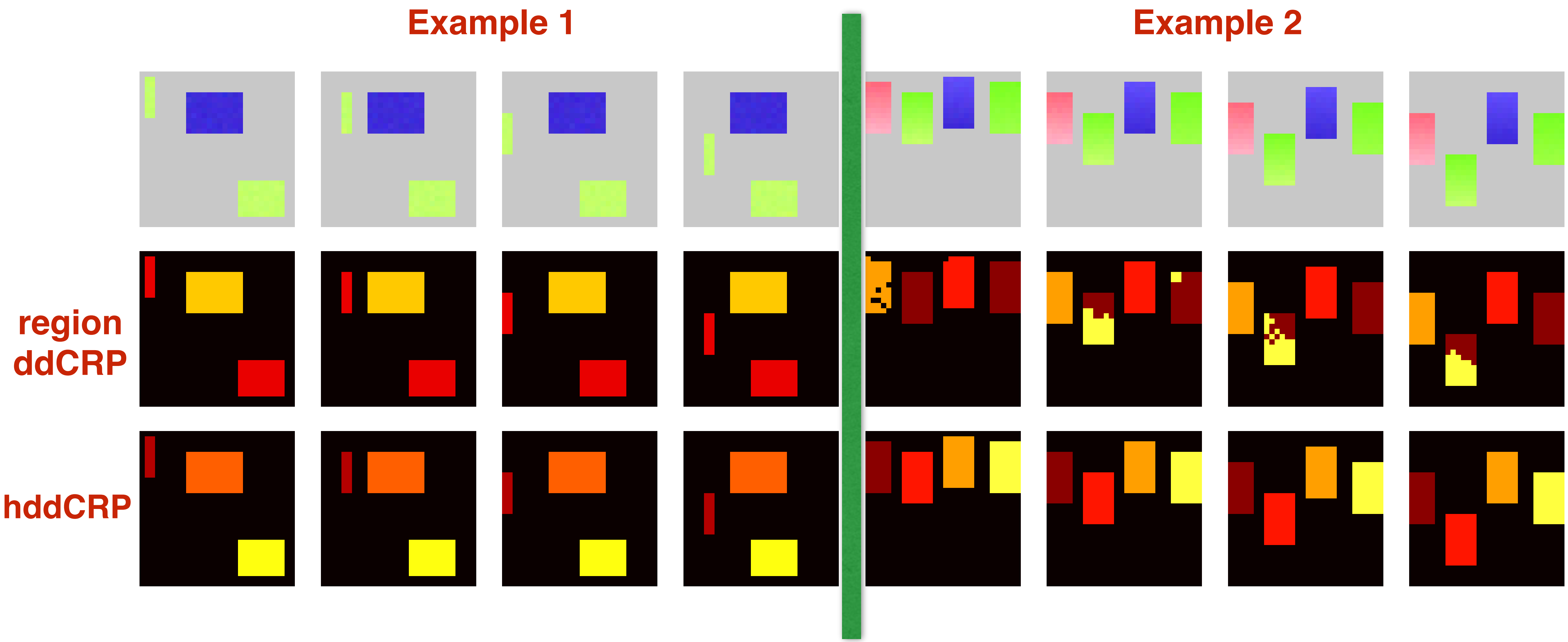
$$\Phi(u_{ki}) \sim \text{Unif}(0,1)$$

Calibrating Covariance Kernels

- Gaussian Process covariance determine the layer shape.

$$K(i, j) \longleftrightarrow \text{Probability that features at i and j are in the same layer}$$

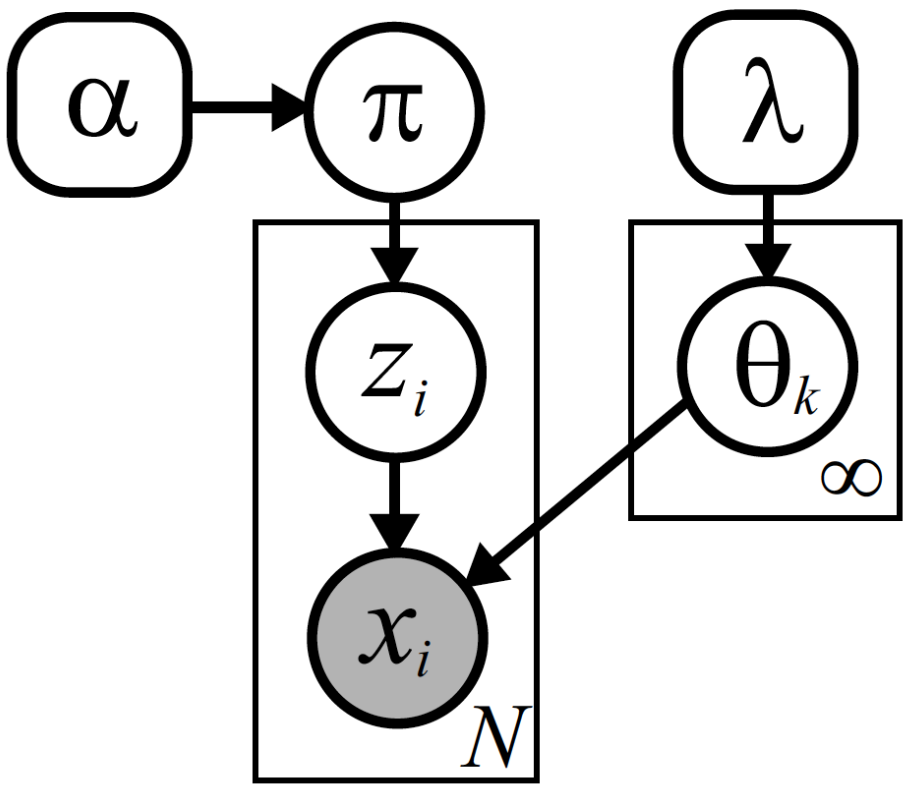
Video Segmentation-Toy Data



Size based inter-table distances

Motion based inter-table distances

Bayesian Nonparametric Priors



$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta)$$

$$\sum_{k=1}^{\infty} \pi_k = 1 \quad 0 \leq \pi_k \leq 1$$

$$\theta_k \sim \mathrm{H}(\lambda)$$

Pitman-Yor Process

Power Law Behavior

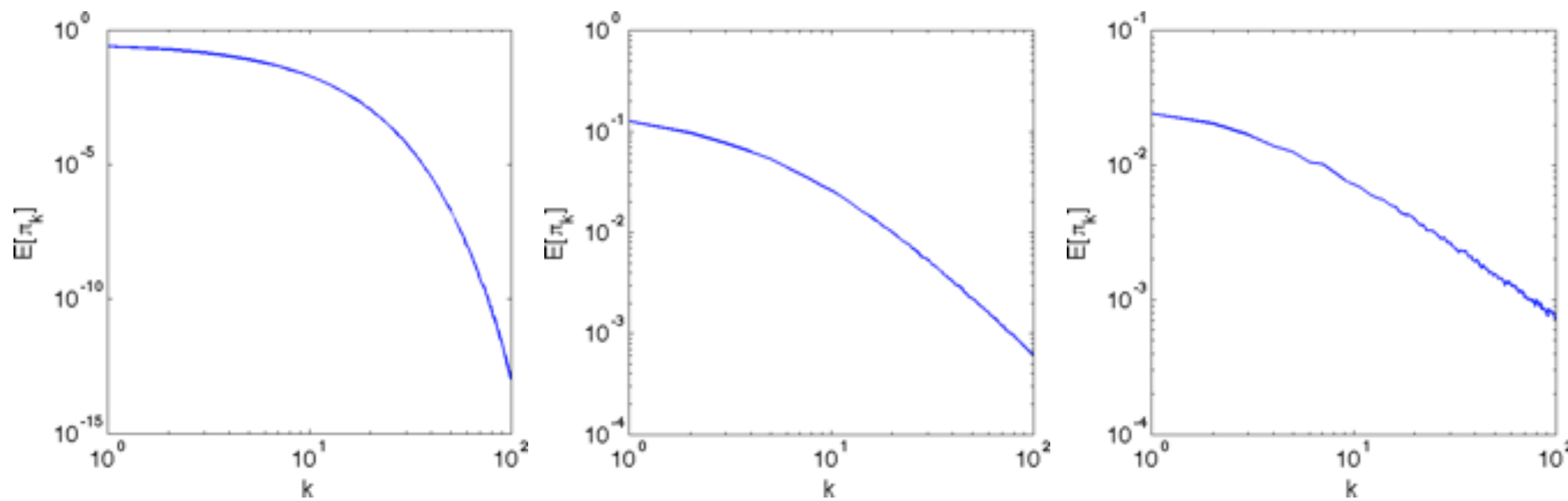
$$E[w_k] = \frac{1-\alpha_a}{(1+\alpha_b+(k-1)\alpha_a)}$$

$$\pi_k = w_k \prod_{l=1}^{k-1}(1-w_l)$$

$$w_k \sim \mathrm{Beta}(a_k, b_k)$$

$$a_k = 1-\alpha_a$$

$$b_k = \alpha_b + k\alpha_a$$

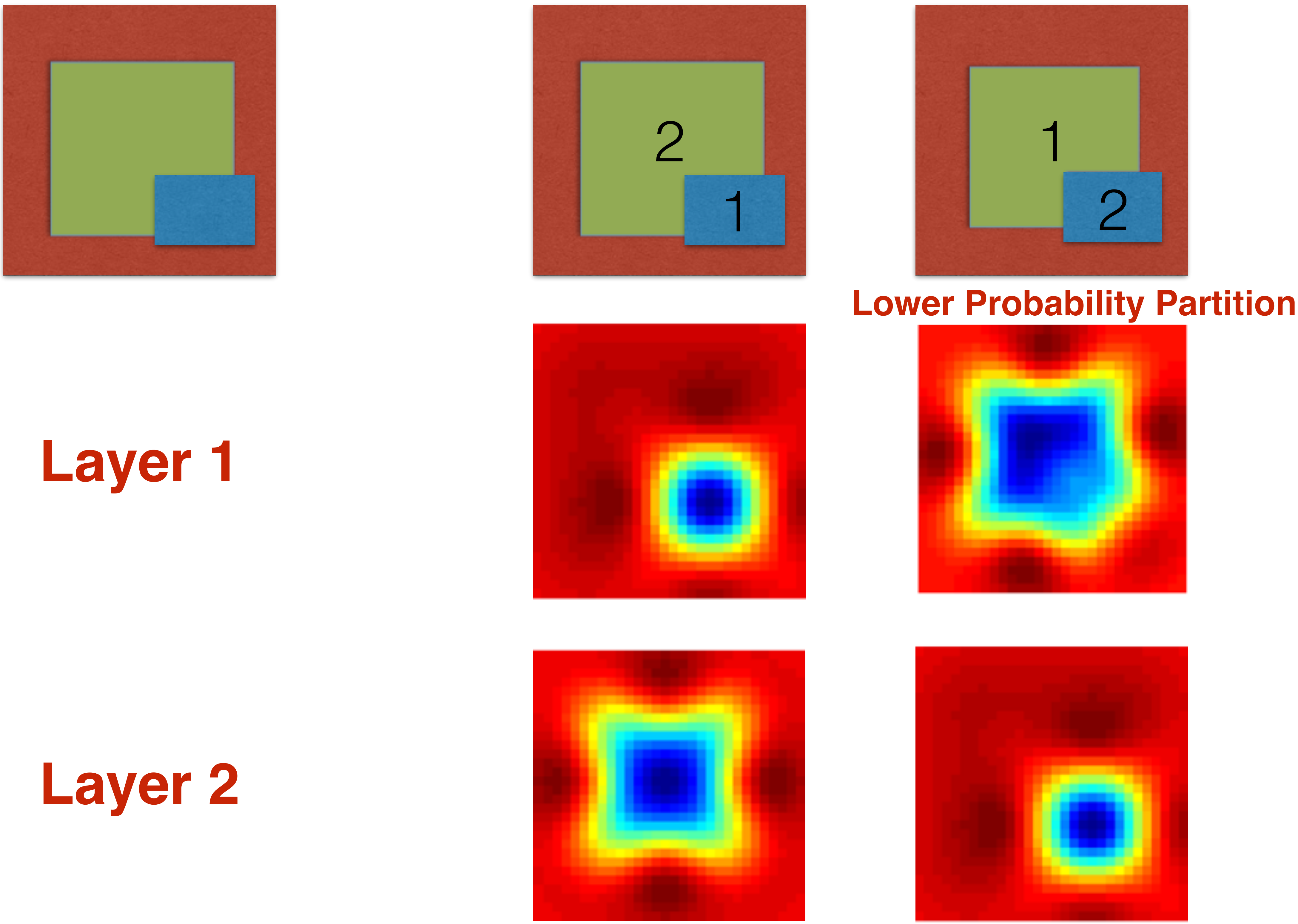


$\alpha_a = 0, \alpha_b = 3$ $\quad$ $\alpha_a = 0.5, \alpha_b = 3$ $\quad$ $\alpha_a = 0.9, \alpha_b = 3$

Number of unique clusters in N observations: $O(\alpha_b N^{\alpha_a})$

Expected size of sorted component k: $O(k^{-\frac{1}{\alpha_a}})$

Depth-ordered Layers

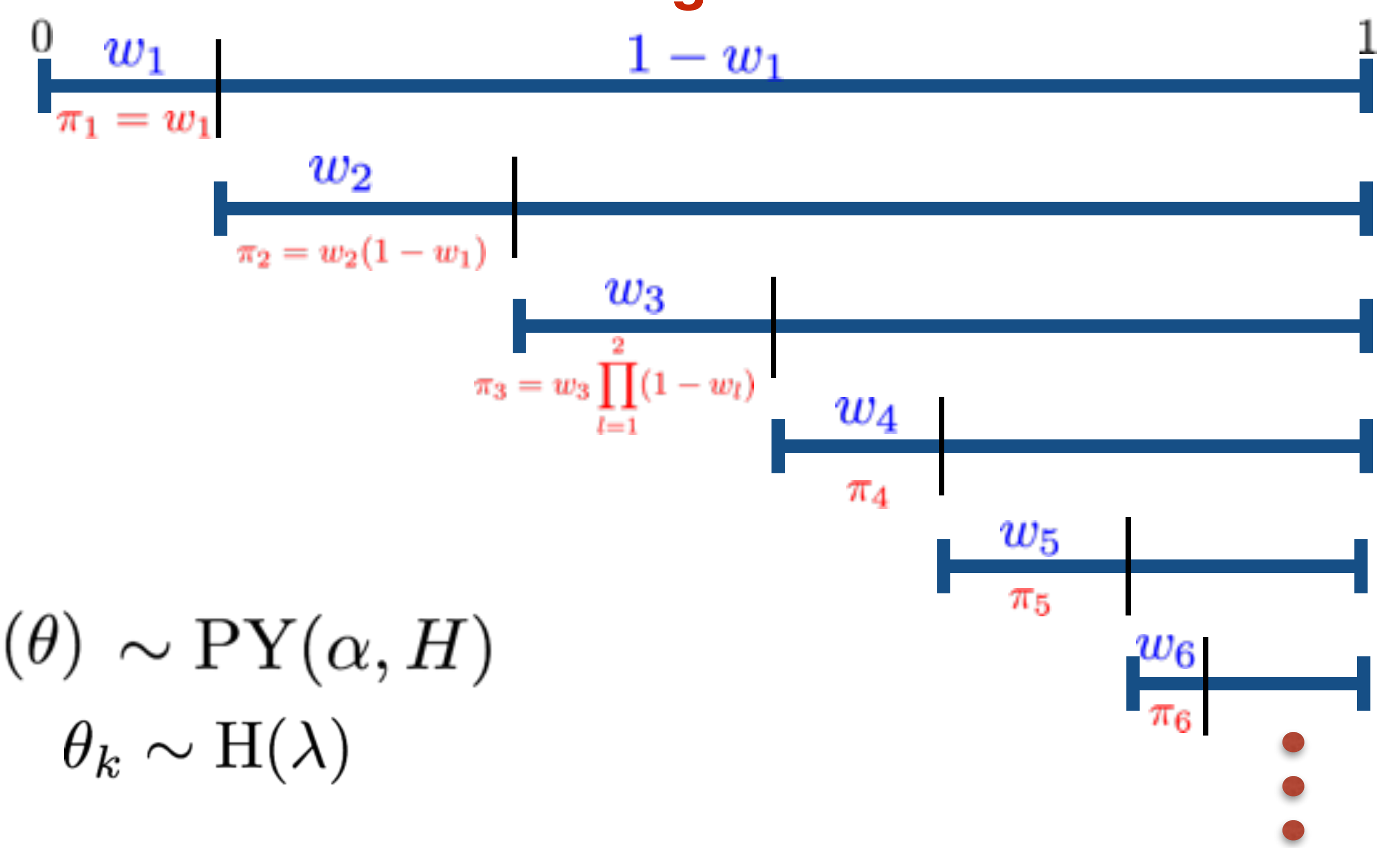


Pitman-Yor Process

- The Pitman-Yor process defines a distribution on infinite discrete measures, or partitions

$$\pi_k = w_k \prod_{l=1}^{k-1}(1 - w_l) \qquad w_k \sim \mathrm{Beta}(1 - \alpha_a, \alpha_b + k\alpha_a)$$

Stick Breaking Construction:



$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta_{\theta_k}(\theta) \sim \mathrm{PY}(\alpha, H)$$

$$\theta_k \sim \mathrm{H}(\lambda)$$

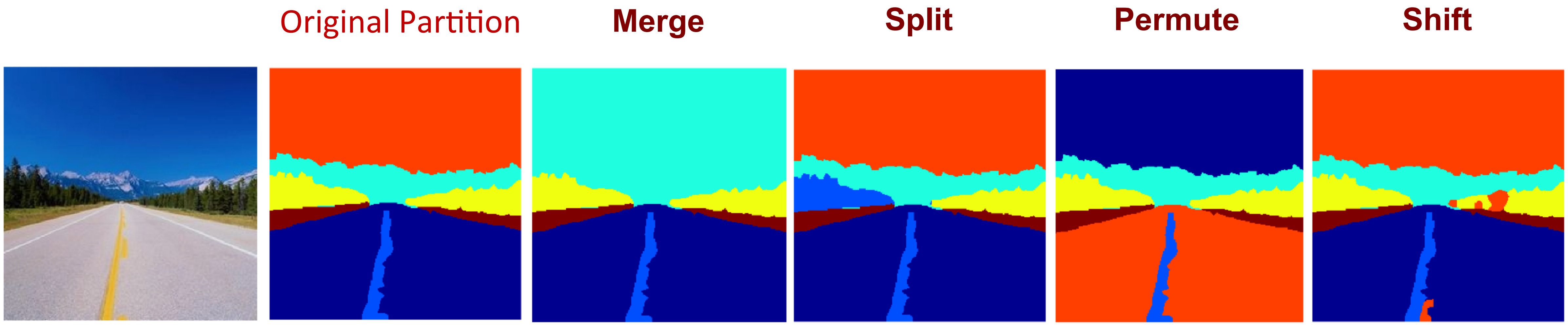
Sethuraman, 1994
Ishwaran and James, 2001

Proposed Work

- Shared segmentation of image corpuses.
- Recover coarse scene geometry.
- Discover visual categories.
- Novel, scalable inference.

Discrete Search Moves

- Merge: Combine a pair of regions into a single region
- Split: Break a single region into a pair of regions.
- Shift: Sequentially move super-pixels to most probable layers.
- Permute: Swap depth-order of two layers.



Marginalization of continuous variables simplifies these moves…

Ghosh & Sudderth, 2012

Pitman-Yor Process

$$\pi_k = w_k \prod_{l=1}^{k-1}(1 - w_l)$$

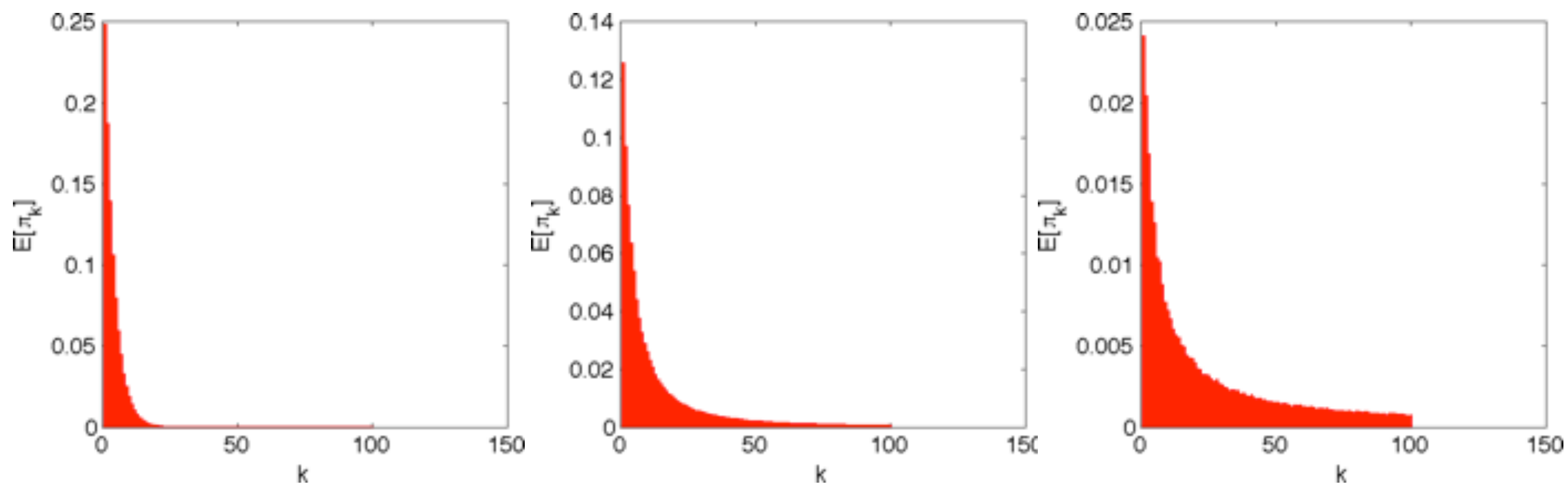
$$w_k \sim \mathrm{Beta}(a_k, b_k)$$

$$a_k = 1 - \alpha_a$$

$$b_k = \alpha_b + k\alpha_a$$

Power Law Behavior

$$E[w_k] = \frac{1 - \alpha_a}{(1 + \alpha_b + (k-1)\alpha_a)}$$



$\alpha_a = 0, \alpha_b = 3$ $\alpha_a = 0.5, \alpha_b = 3$ $\alpha_a = 0.9, \alpha_b = 3$

Dirichlet Process:

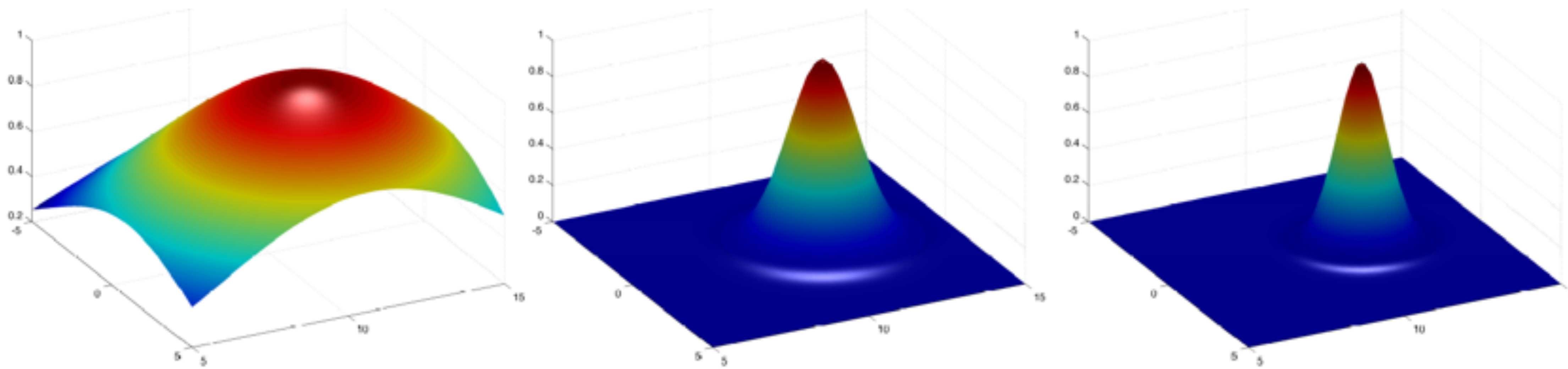
$\alpha_a = 0$

Number of unique clusters in N observations: $O(\alpha_b N^{\alpha_a})$

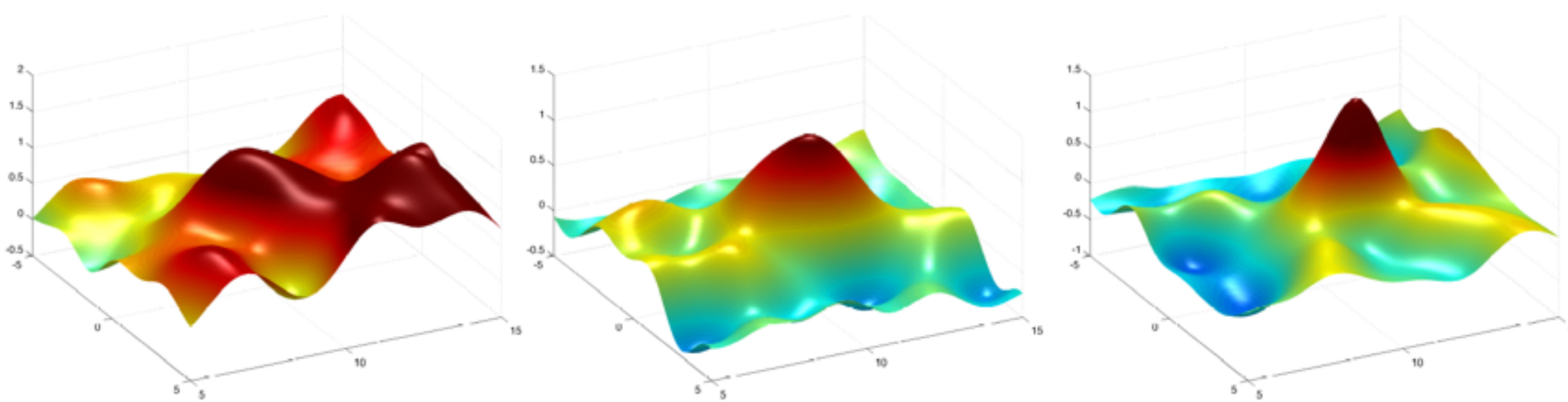
Expected size of sorted component k: $O(k^{-\frac{1}{\alpha_a}})$

Category specific shapes and sizes

- Each visual category is embellished with a mean function modeling its characteristic shape and size



- Sample an instance specific realization of a GP parametrized with the appropriate category mean and threshold.



Probabilities to Correlations

- Two super-pixels belong to the same layer if they are both below the threshold. Thus w.p.:

$$p_-|\delta_k = \int_{-\infty}^{\delta_k}\int_{-\infty}^{\delta_k} \mathcal{N}\left(\begin{bmatrix} u_i \\ u_j \end{bmatrix} \mid \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix}\right) du_i du_j$$

- Marginalizing over random threshold we have:

$$q_-^k(\alpha,\rho) = \int_{-\infty}^{\infty}\int_{-\infty}^{\delta_k}\int_{-\infty}^{\delta_k} \mathcal{N}\left(\begin{bmatrix} u_i \\ u_j \end{bmatrix} \mid \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix}\right) p(\delta_k|\alpha) du_i du_j d\delta_k$$

$$q_+^k(\alpha,\rho) = \int_{-\infty}^{\infty}\int_{\delta_k}^{\infty}\int_{\delta_k}^{\infty} \mathcal{N}\left(\begin{bmatrix} u_i \\ u_j \end{bmatrix} \mid \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & c \\ c & 1 \end{bmatrix}\right) p(\delta_k|\alpha) du_i du_j d\delta_k$$

- Probability of belonging to the same layer then is:

$$\begin{aligned} q_{ij} &= q_-^1(\alpha,\rho) + q_-^2(\alpha,\rho)q_+^1(\alpha,\rho) + q_-^3(\alpha,\rho)q_+^1(\alpha,\rho)q_+^2(\alpha,\rho) + \dots \\ &\approx \sum_{k=1}^{K} q_-^k(\alpha,\rho) \prod_{l=1}^{K-1} q_+^l(\alpha,\rho) \end{aligned}$$