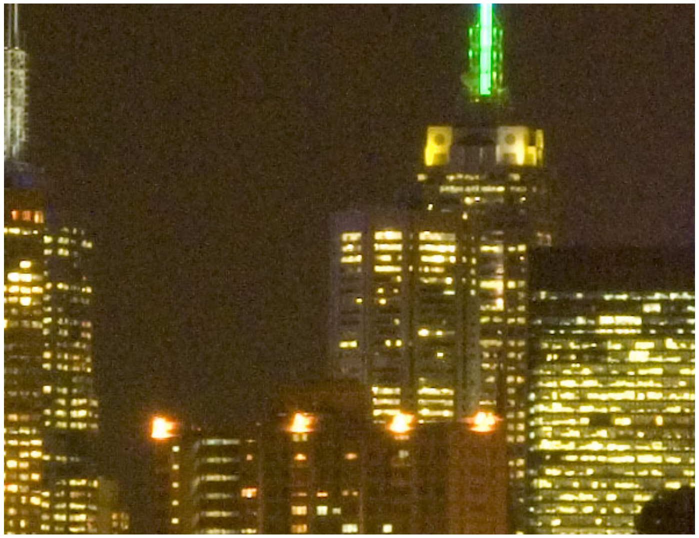
Visual Learning via Topics, Transformations, and Trees

Erik Sudderth

Department of Computer Science
Brown University

Joint work on
Transformations: Antonio Torralba, Bill Freeman, Alan Willsky
Trees: Jyri Kivinen, Michael Jordan

Low-level Image Analysis



Noise Removal

Deblurring

Inpainting & Restoration

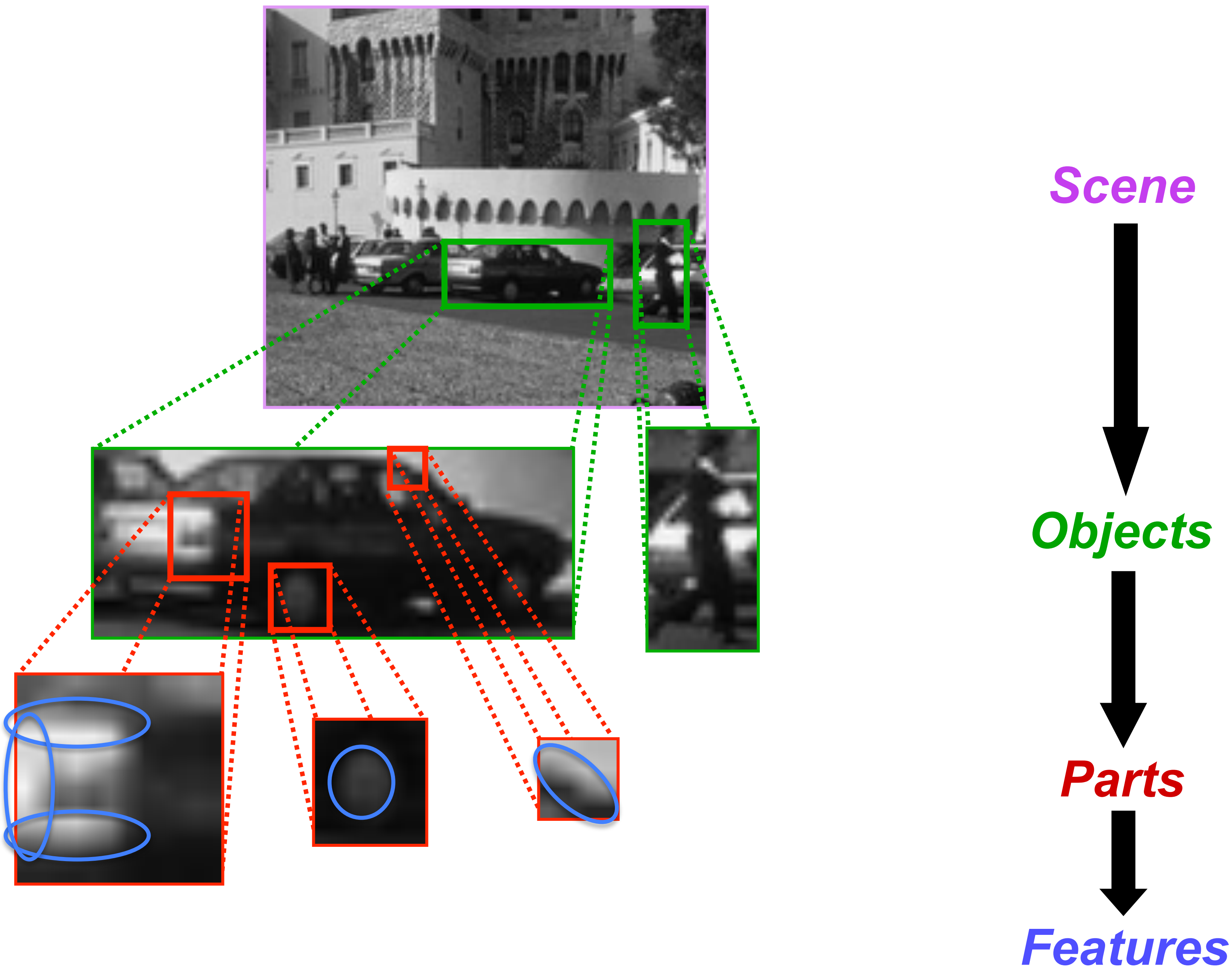
What are the statistical properties of natural images?

Natural Scene Categorization

Coast **Forest** **Open Country** **Street** **Tall Building**

How do semantic labels affect these properties?

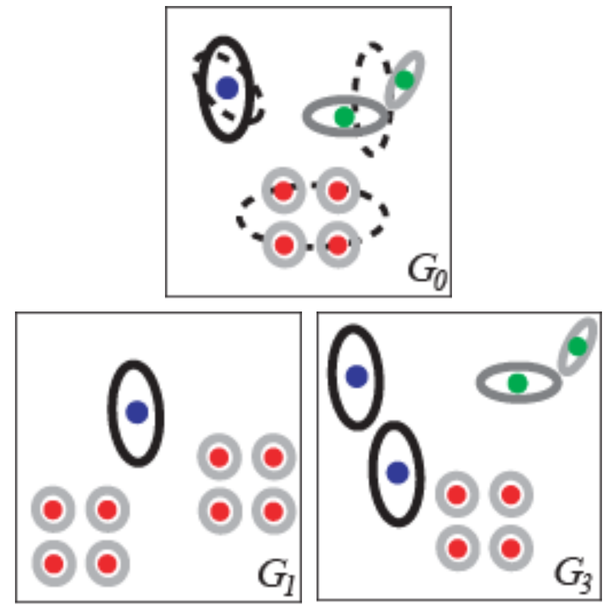
Scenes, Objects, and Parts



Outline

Topics

- Bag of feature image representations
- Hierarchical Bayesian modeling

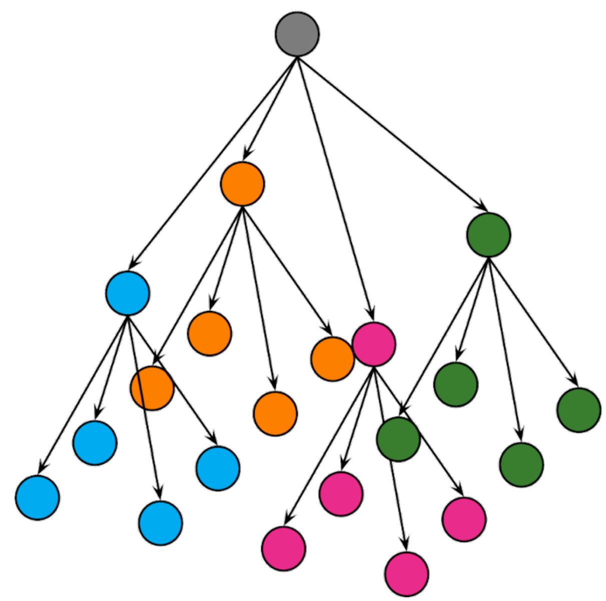


Transformations

- Sharing parts among object categories
- Spatial models for visual scenes

Trees

- Multiscale nonparametric Markov models
- Image denoising and scene categorization

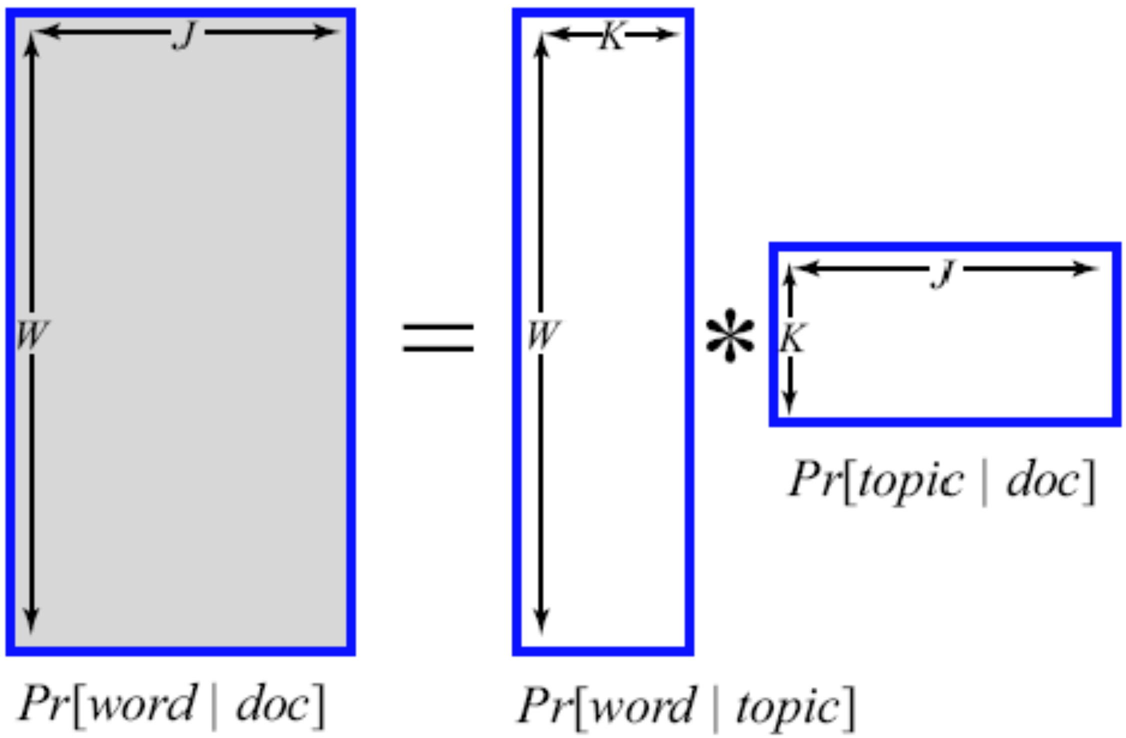


Learning with Topic Models

Framework for unsupervised discovery of *low-dimensional* latent structure from *bag of word* representations

model
neural
stochastic
recognition
nonparametric
gradient
dynamical
Bayesian
…

$W \approx 10,000$



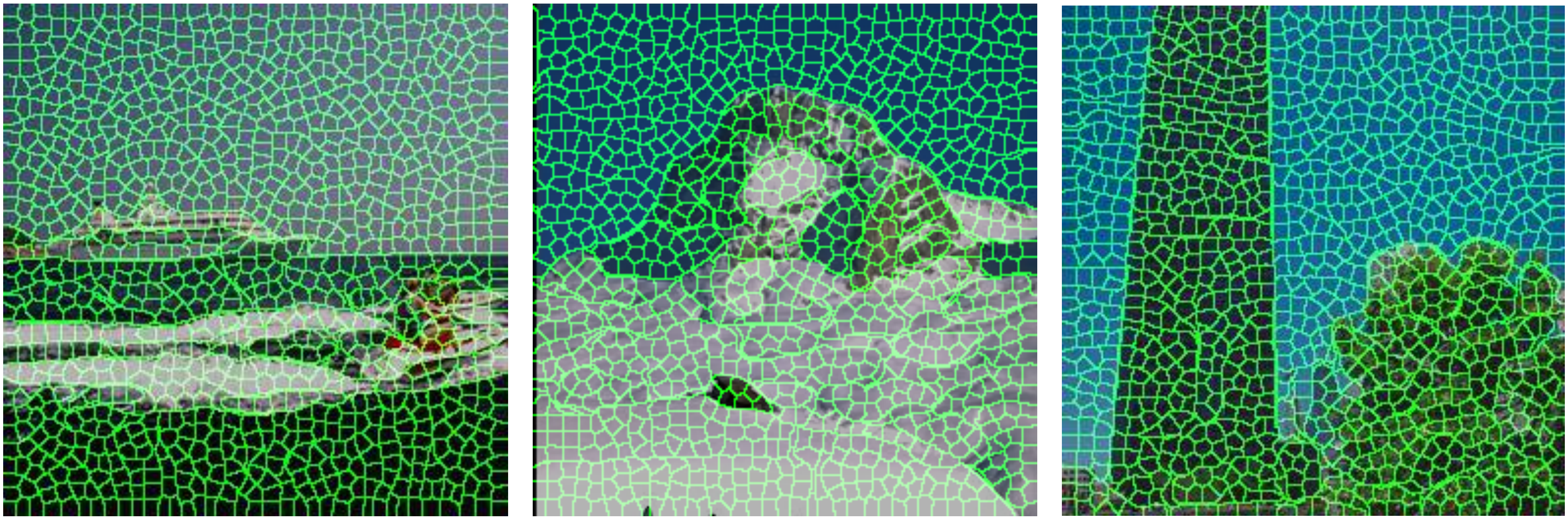
Algorithms
Neuroscience
Statistics
Vision
…

$K \approx 100$

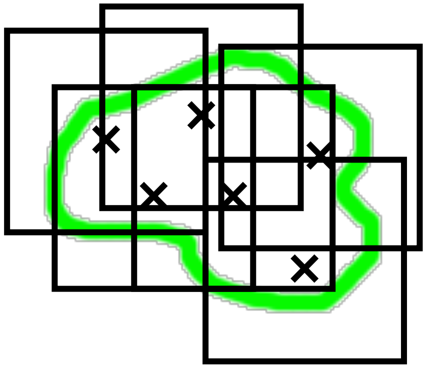
- **pLSA**: Probabilistic Latent Semantic Analysis *(Hofmann 2001)*
- **LDA**: Latent Dirichlet Allocation *(Blei, Ng, & Jordan 2003)*
- **HDP**: Hierarchical Dirichlet Processes *(Teh, Jordan, Beal, & Blei 2006)*

Local Visual Features: Superpixels

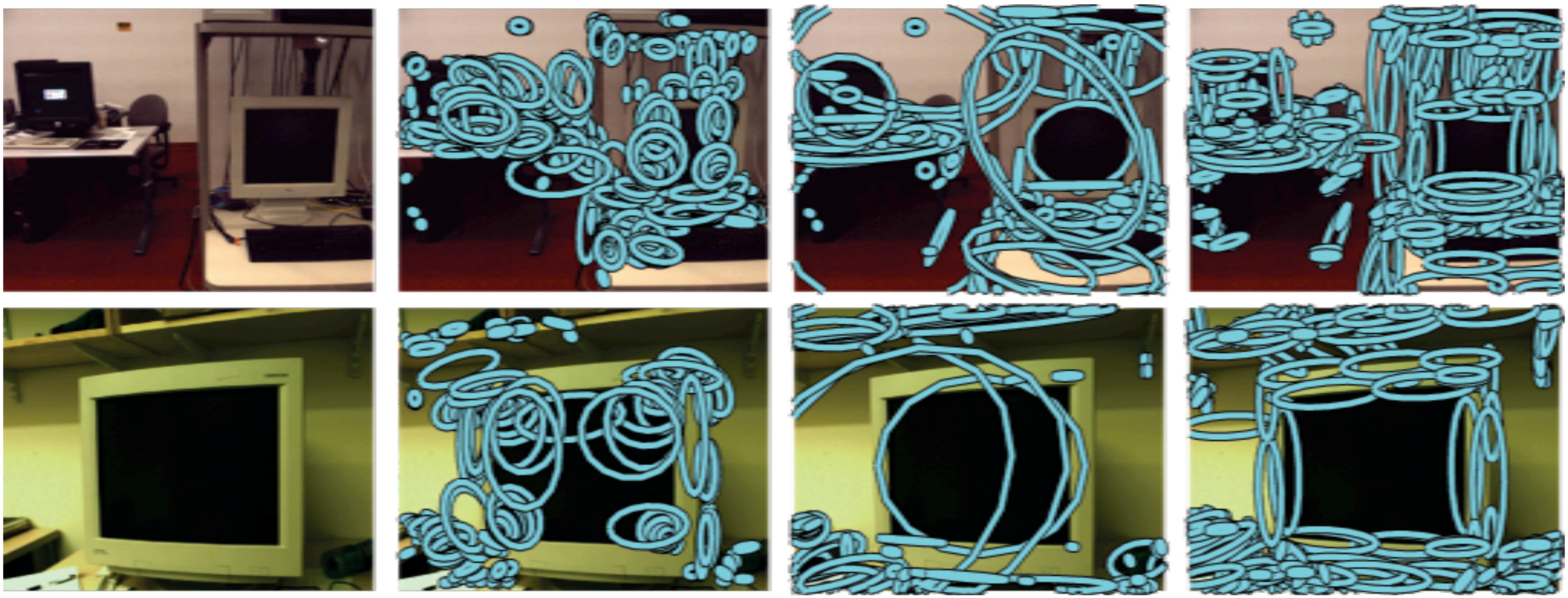
Inspired by the successes of *topic models* for text data, some have proposed learning from *local image features*



- Partition image into ~1,000 *superpixels*
- Goal: Reduce dimensionality, aggregate information spatially – *hopefully not across object boundaries!*



Local Visual Features: Interest Regions

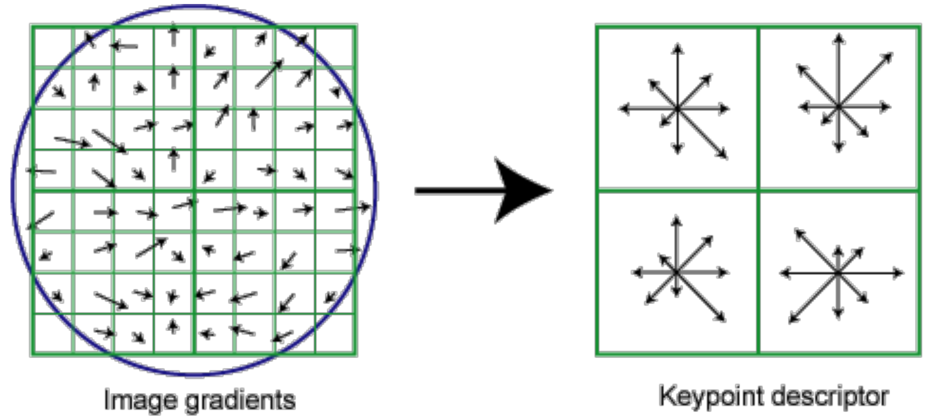


- Some invariance to lighting & pose variations
- Dense, multiscale *over-segmentation* of image

A Discrete Feature Vocabulary

SIFT Descriptors

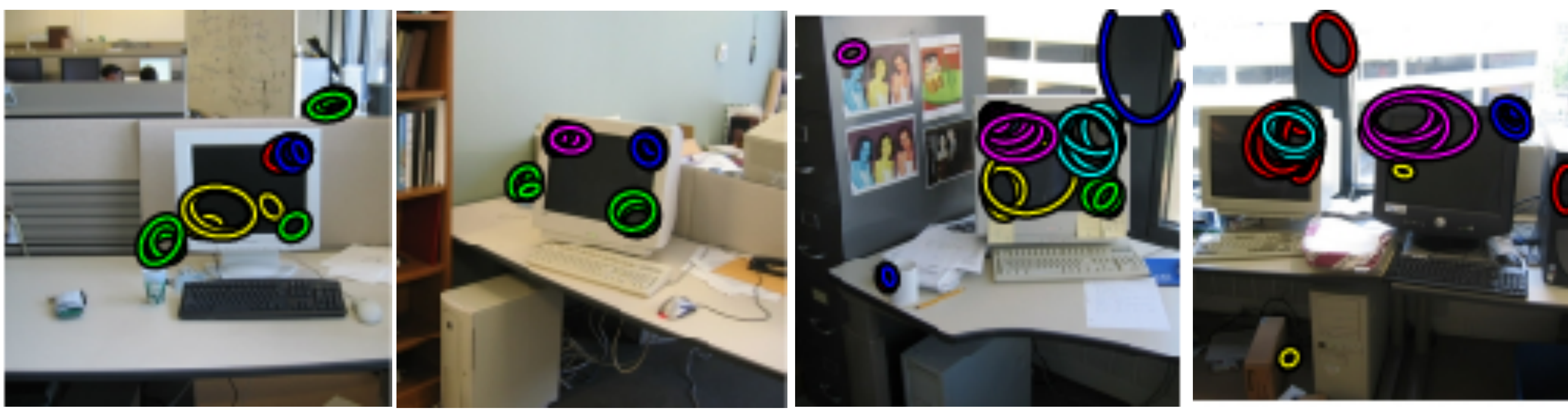
- Normalized histograms of orientation energy
- Compute ~1,000 word dictionary via K-means
- Map each feature to nearest *visual word*



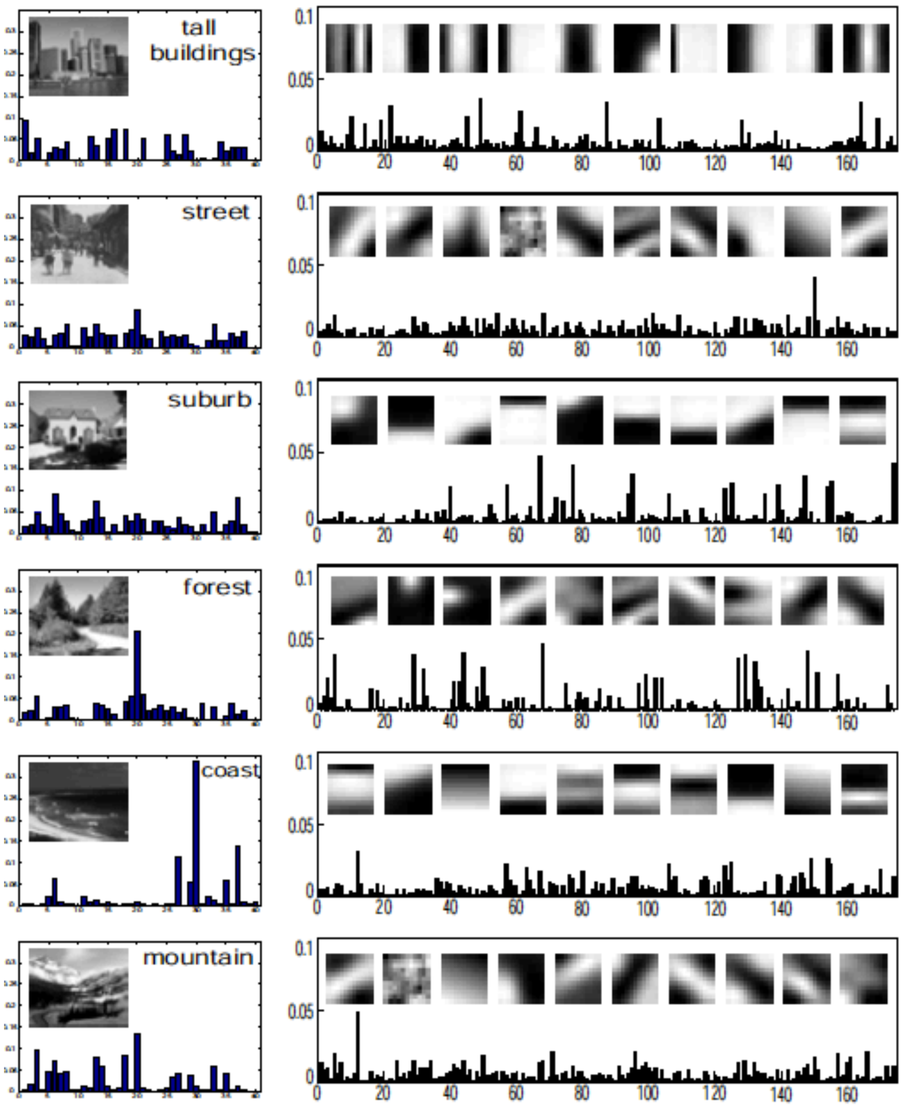
Lowe, IJCV 2004

$w_{ji} \longrightarrow$ appearance of feature i in image j

$v_{ji} \longrightarrow$ 2D position of feature i in image j

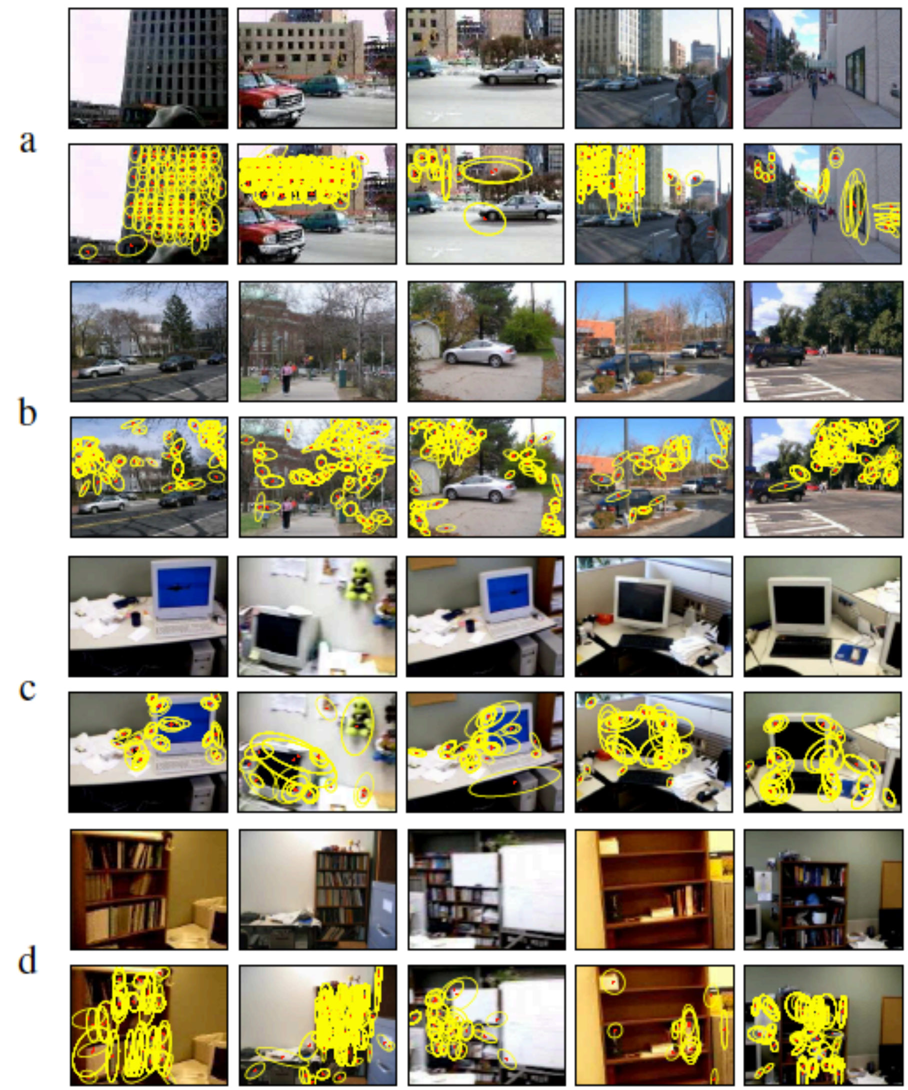


The World as a Bag of Visual Words



*Fei-Fei & Perona, **CVPR 2005***

Topics as *visual themes* composing a known set of scene categories



*Sivic, Russell, Efros, Zisserman, & Freeman, **ICCV 2005***

Topics as *visual object classes* within a (carefully chosen) image collection

Images as more than Bags of Features

- How do I know this is ocean beneath a clear sky?
- How many bicycles and tricycles am I looking at?

Why are we trying to squeeze images into topic models?

My work explores the larger space of nonparametric and hierarchical Bayesian models.

Dirichlet Process Mixtures

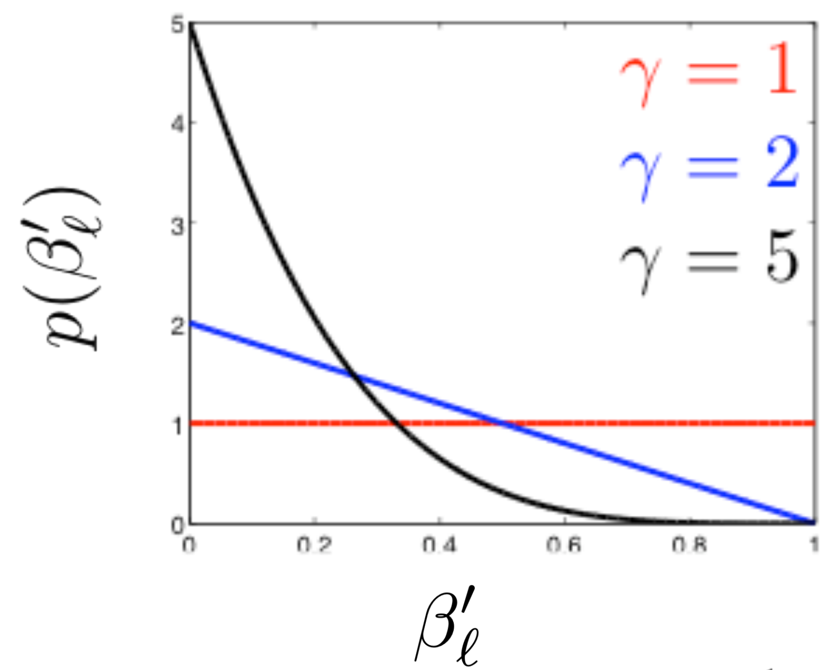
$$p(x_{ti} \mid \beta, \Lambda_1, \Lambda_2, \ldots) = \sum_{k=1}^{\infty} \beta_k \mathcal{N}(x_{ti}; 0, \Lambda_k)$$

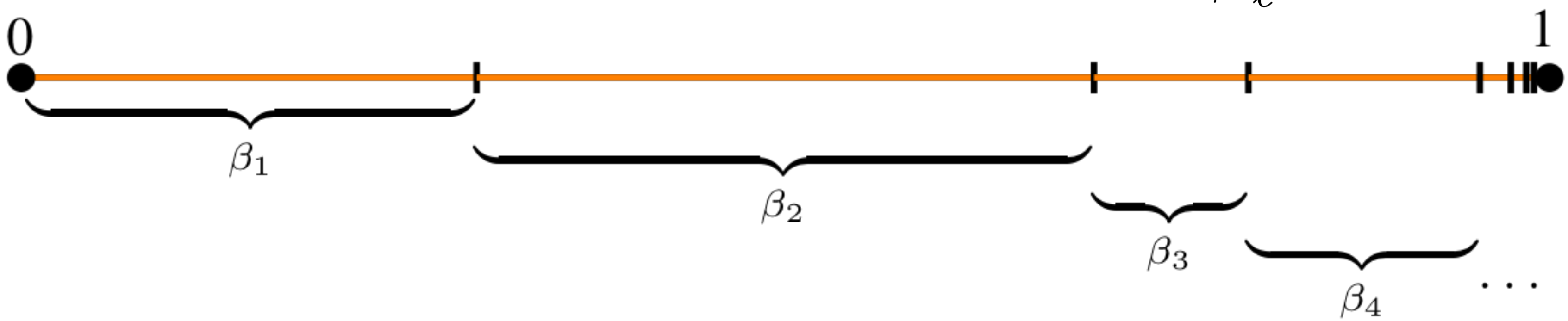
Stick-breaking prior for mixture weights controls complexity:

$$\beta_k = \beta'_k \prod_{\ell=1}^{k-1} (1 - \beta'_\ell)$$

$$\beta'_\ell \sim \mathrm{Beta}(1, \gamma)$$

$\gamma \longrightarrow$ Concentration parameter



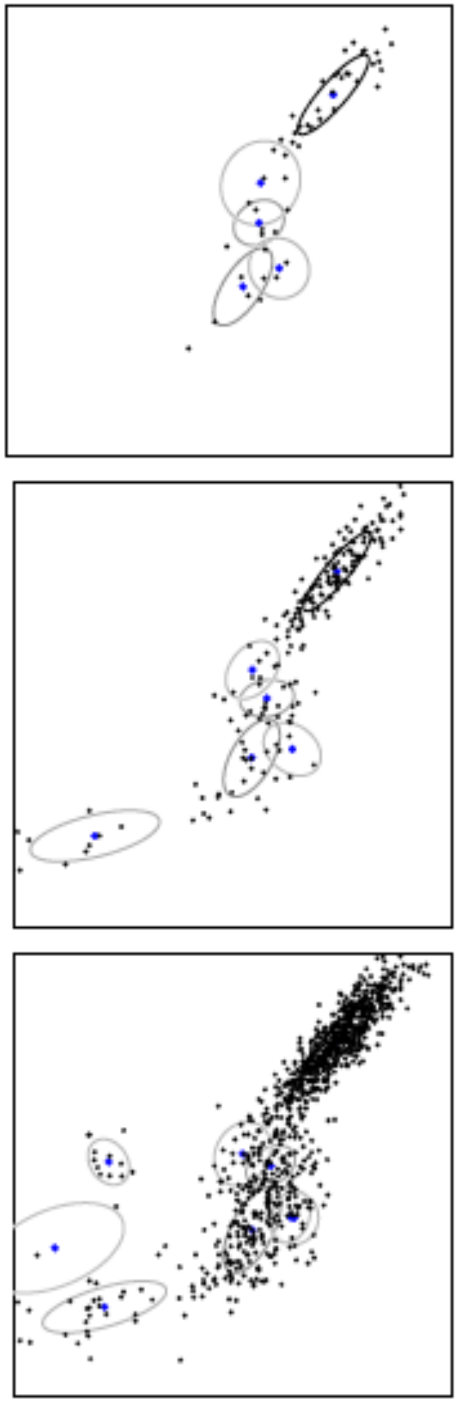


Why the Dirichlet Process ?

$$p(x) = \sum_{k=1}^{\infty} \beta_k f(x \mid \Lambda_k)$$

$$\beta \sim \mathrm{Stick}(\gamma)$$

$$\Lambda_k \sim H$$



- Basis for *nonparametric* models whose complexity grows as data is observed
- Attractive *asymptotic guarantees*
- Leads to simple, effective variational and MCMC *computational methods*

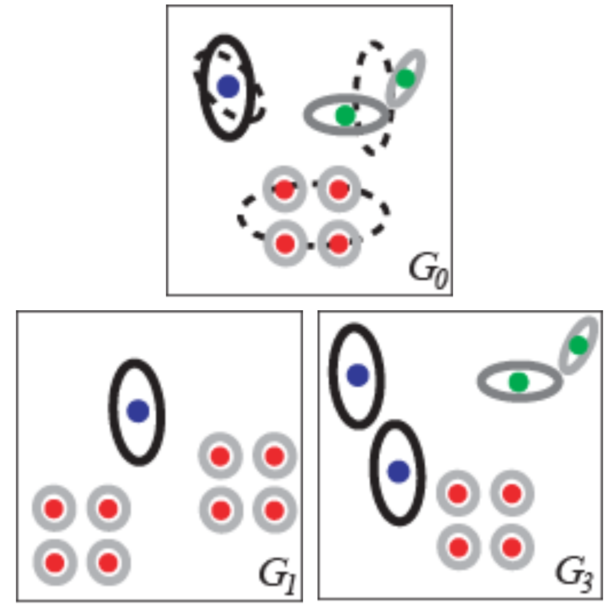
Outline

Topics

- Bag of feature image representations
- Hierarchical Bayesian modeling

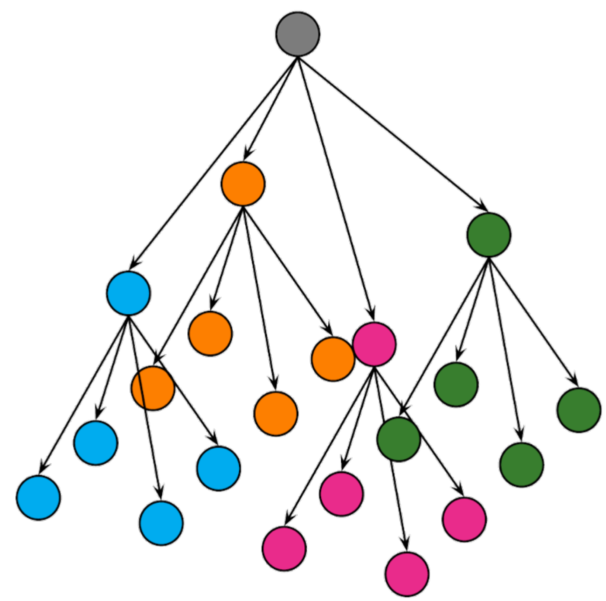
Transformations

- Sharing parts among object categories
- Spatial models for visual scenes

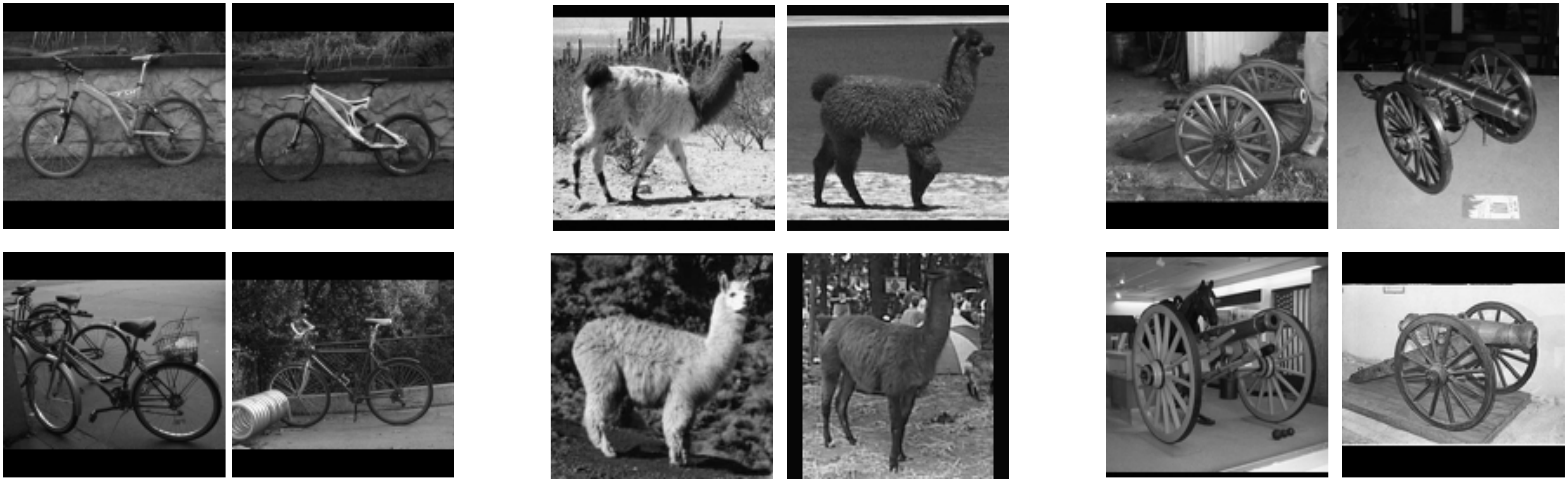


Trees

- Multiscale nonparametric Markov models
- Image denoising and scene categorization

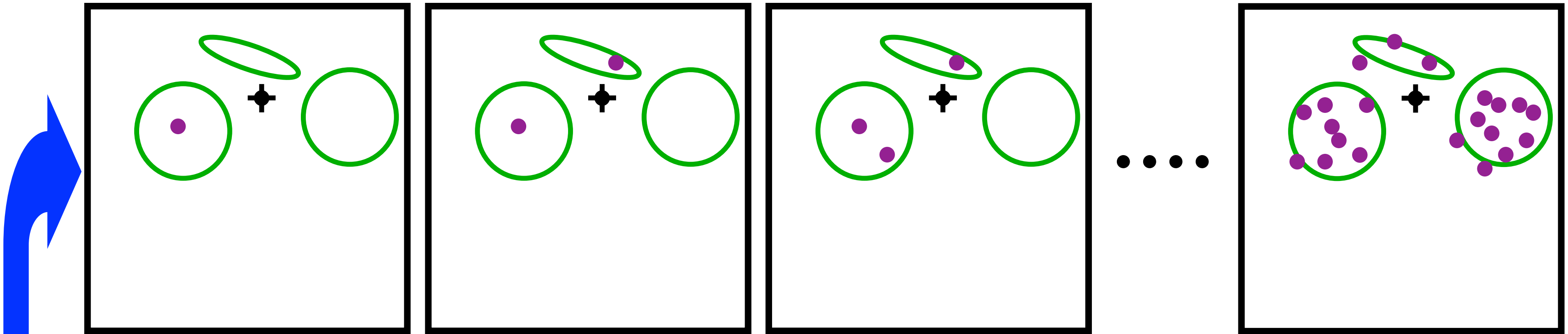


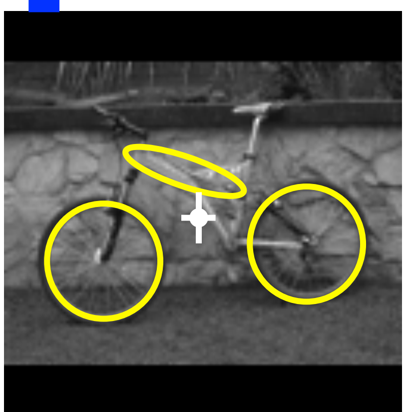
Visual Object Categorization



- **GOAL:** Visually *recognize* and *localize* object categories
- Robustly *learn* appearance models from few examples

Generative Model for Objects

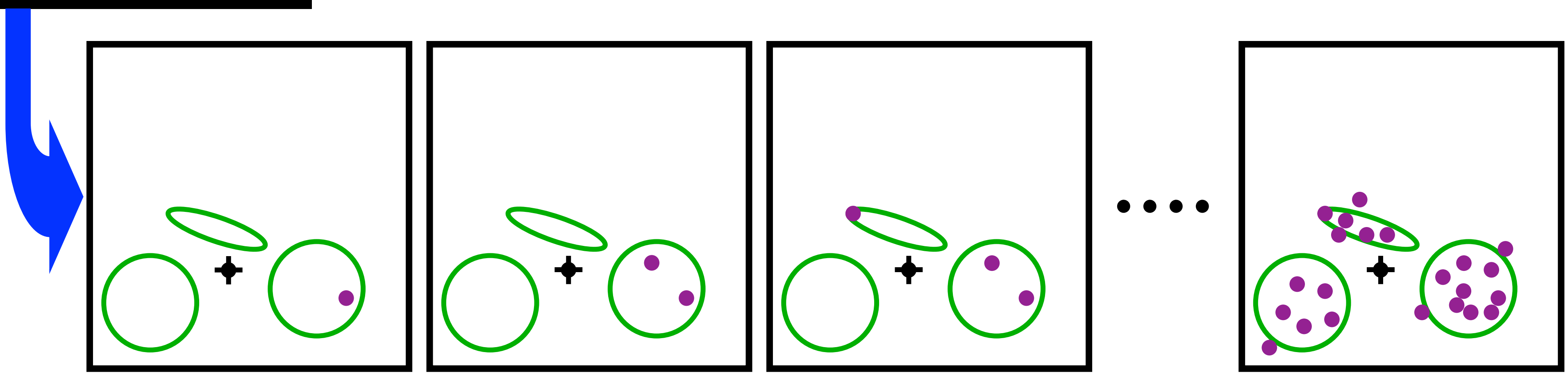




For each image: Sample a reference position

For each feature:

- ➢ Randomly choose one part
- ➢ Sample from that part's feature distribution



Objects as Distributions

$$p(w_{ji}, v_{ji}|\rho_j) = \sum_{k=1}^{\infty} \pi_k \underbrace{\eta_k(w_{ji})}_{\text{Pr(appearance | part)}} \underbrace{\mathcal{N}(v_{ji}; \mu_k + \rho_j, \Lambda_k)}_{\text{Pr(position | part)}}$$

Pr(part)

Feature appearance

Feature position

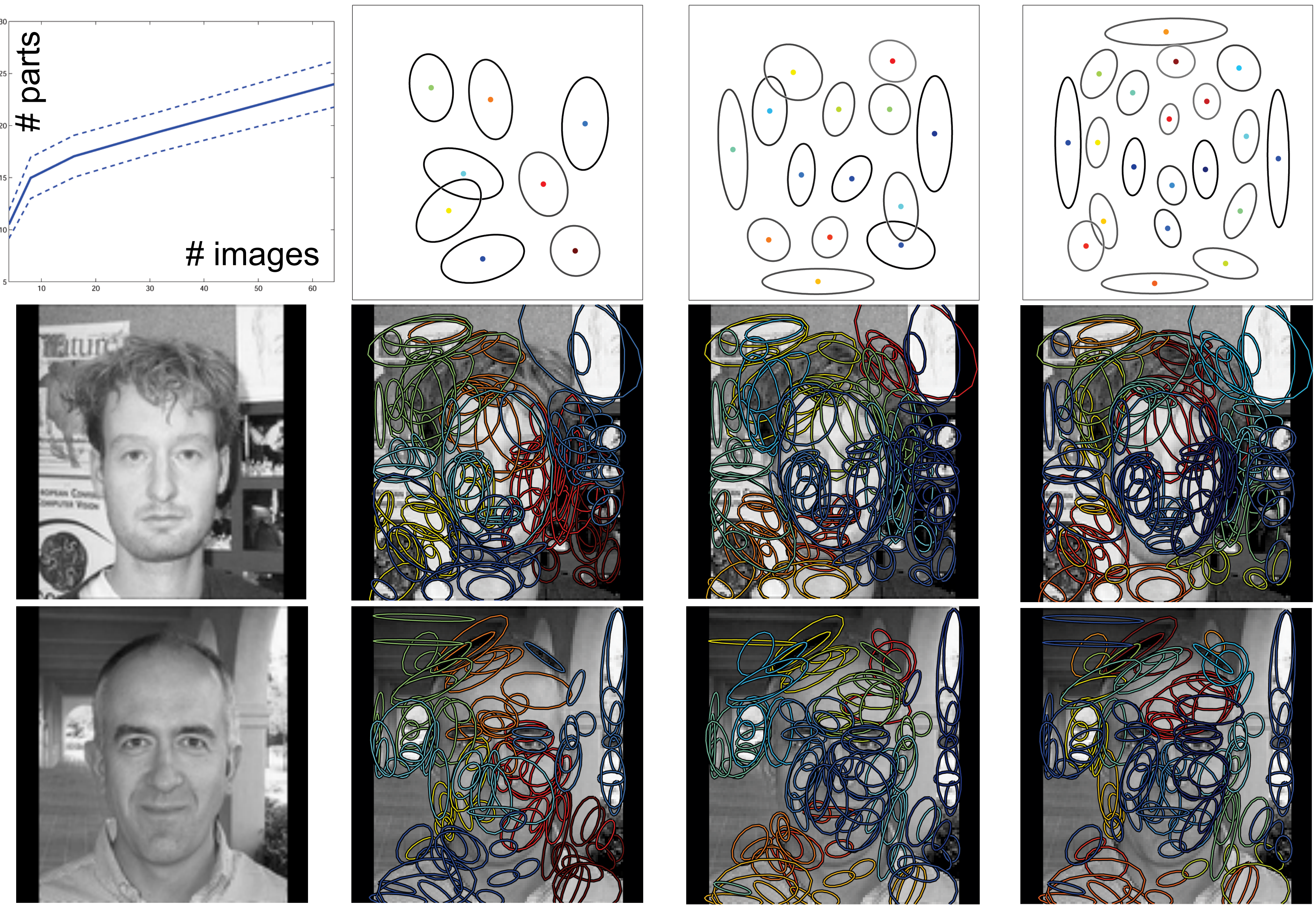
- Parts are defined by *parameters*, which encode distributions on visual features:

$$\theta_k = \{\eta_k, \mu_k, \Lambda_k\}$$

- Objects are defined by *distributions* on the infinitely many potential part parameters:

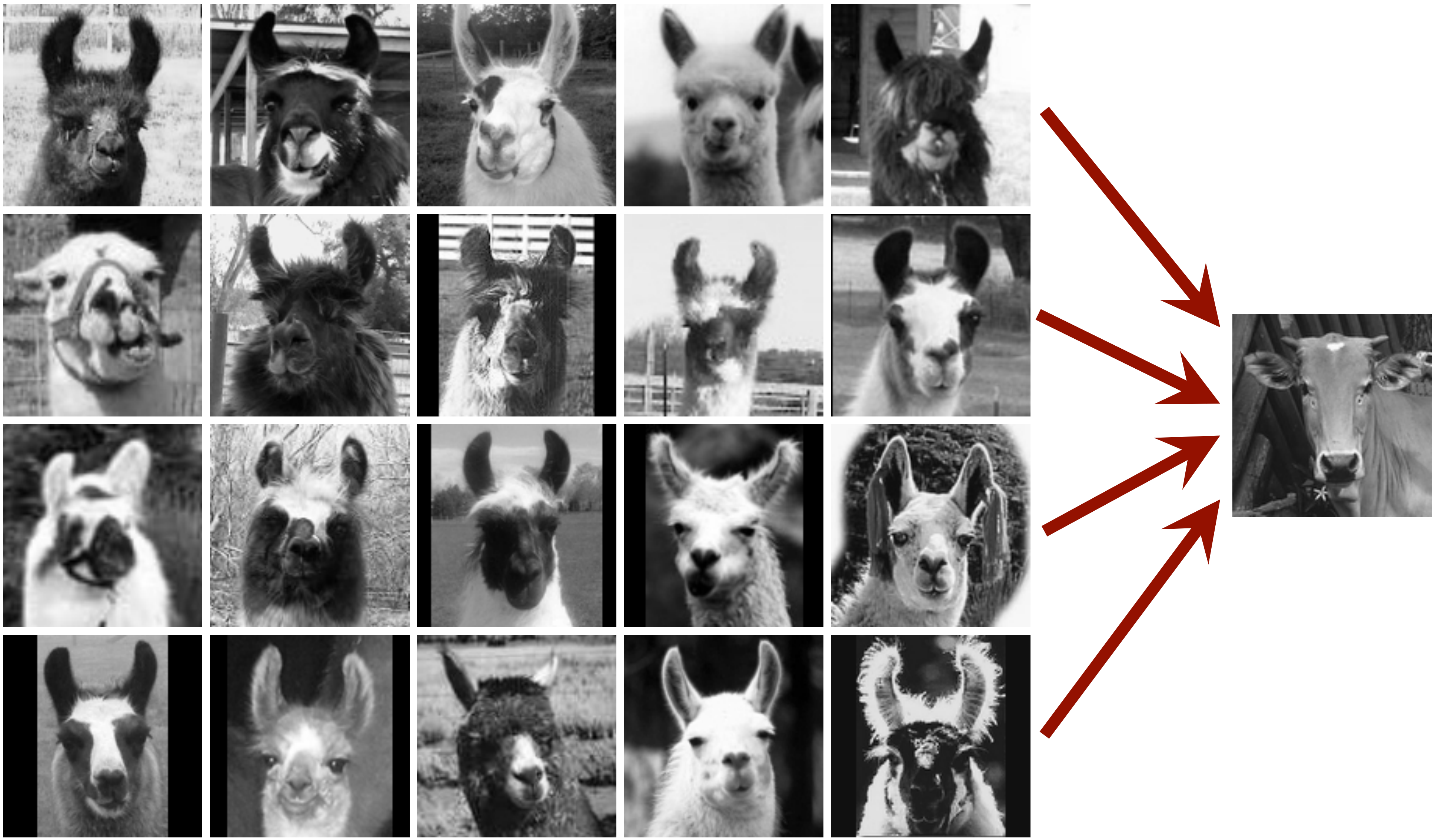
$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k) \qquad \pi \sim \text{Stick}(\alpha)$$

A Nonparmametric Part-Based Model



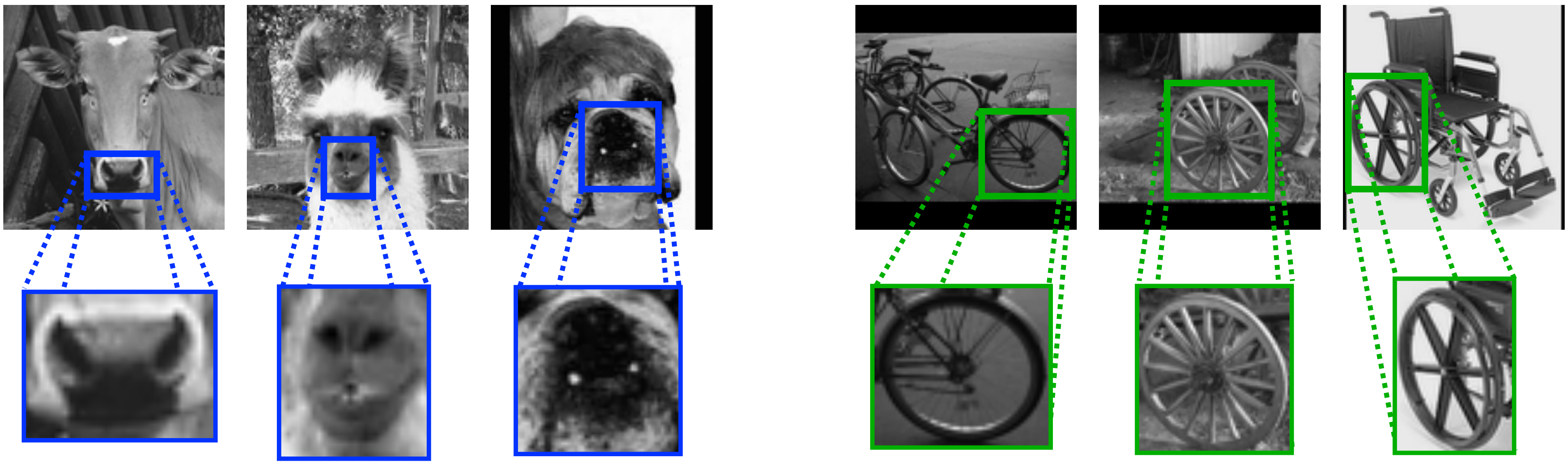
4 Images ***16 Images*** ***64 Images***

Generalizing Across Categories



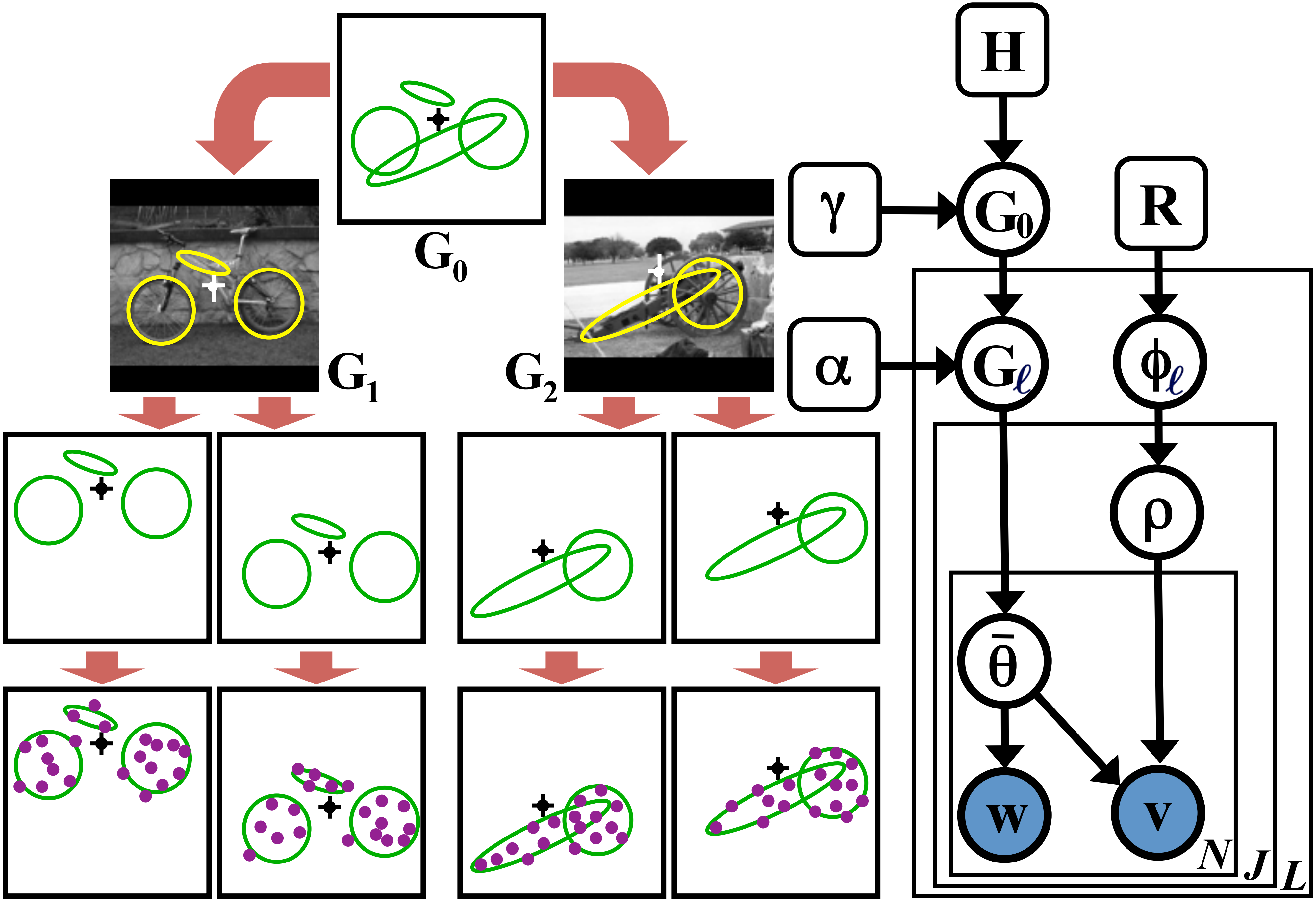
Can we transfer knowledge from one object category to another?

Learning Shared Parts

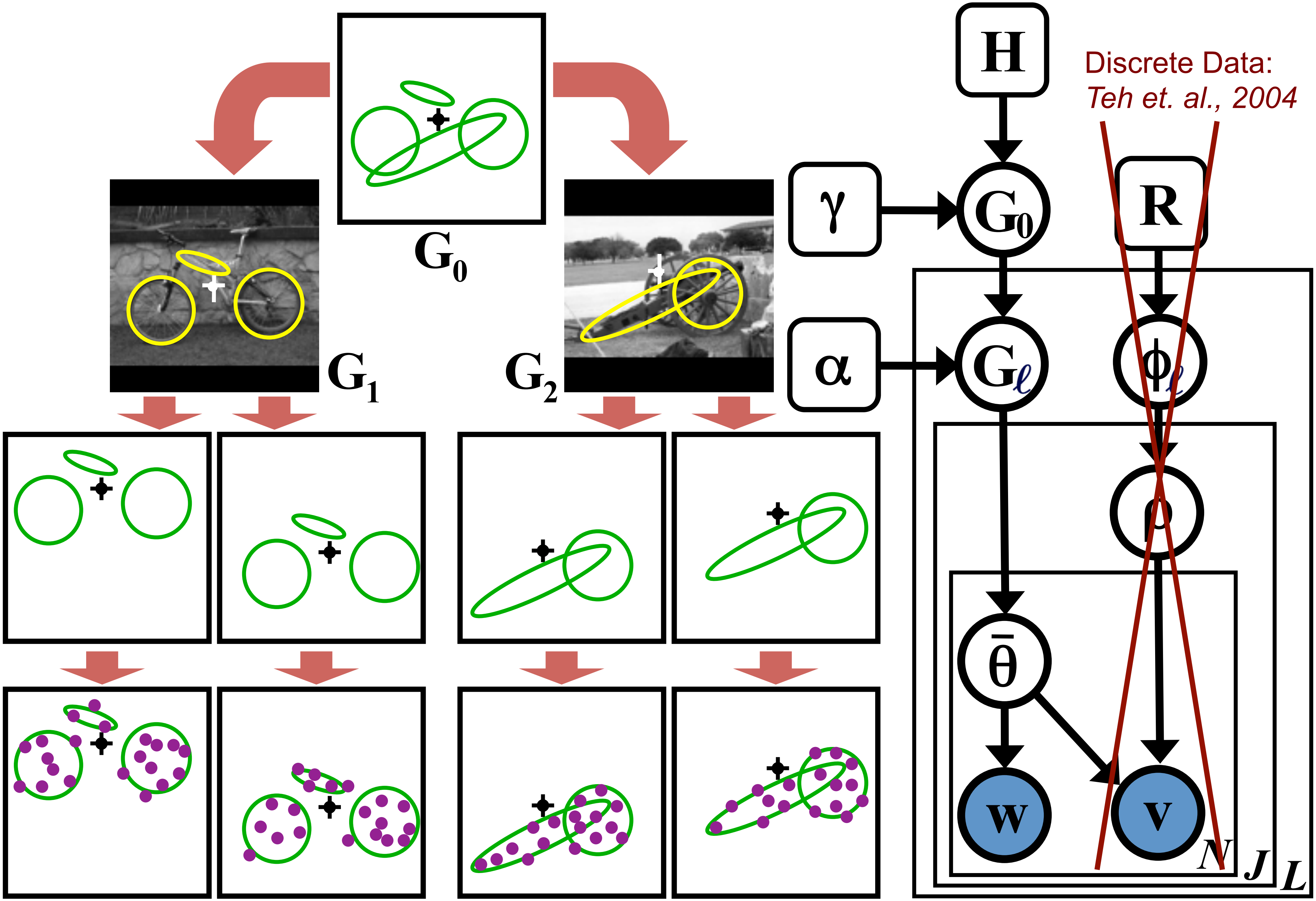


- Objects are often locally similar in appearance
- Discover *parts* shared across categories
  - How many total parts should we share?
  - How many parts should each category use?

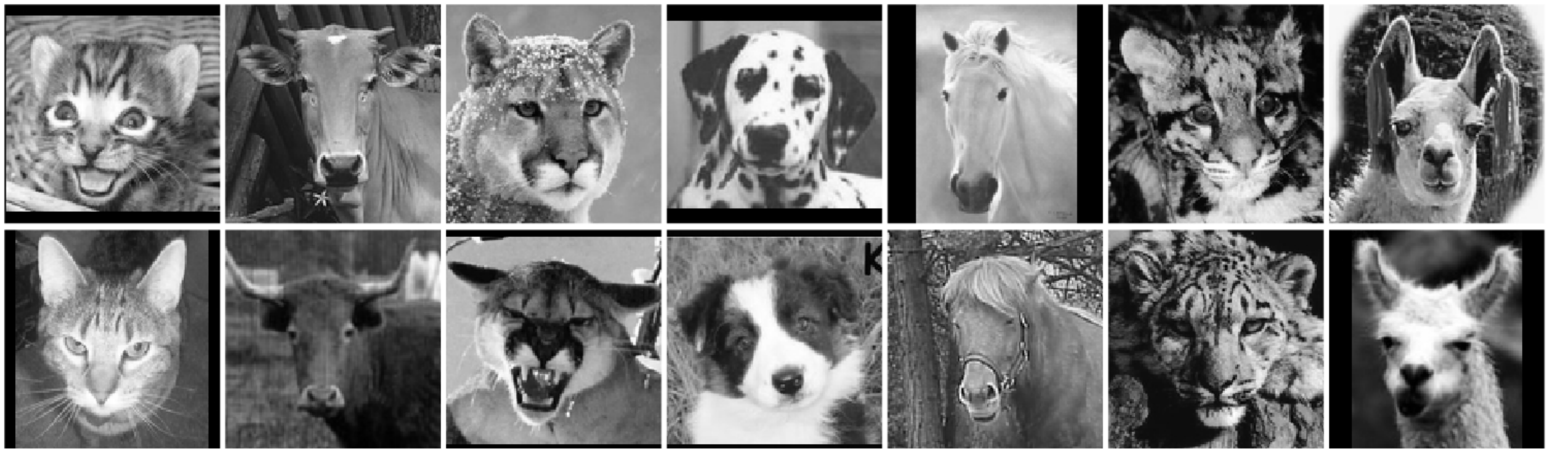
Hierarchical DP Object Model

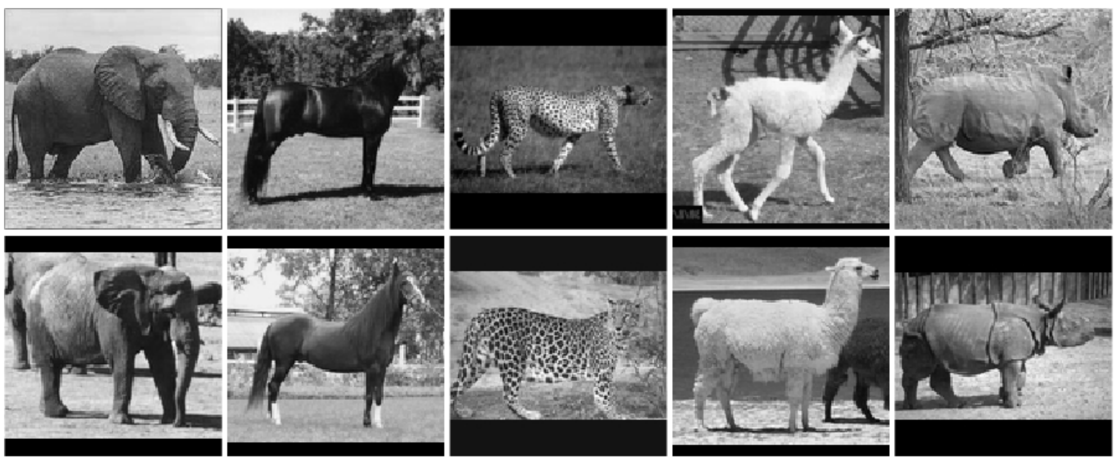


Hierarchical DP Object Model



Sharing Parts: 16 Categories

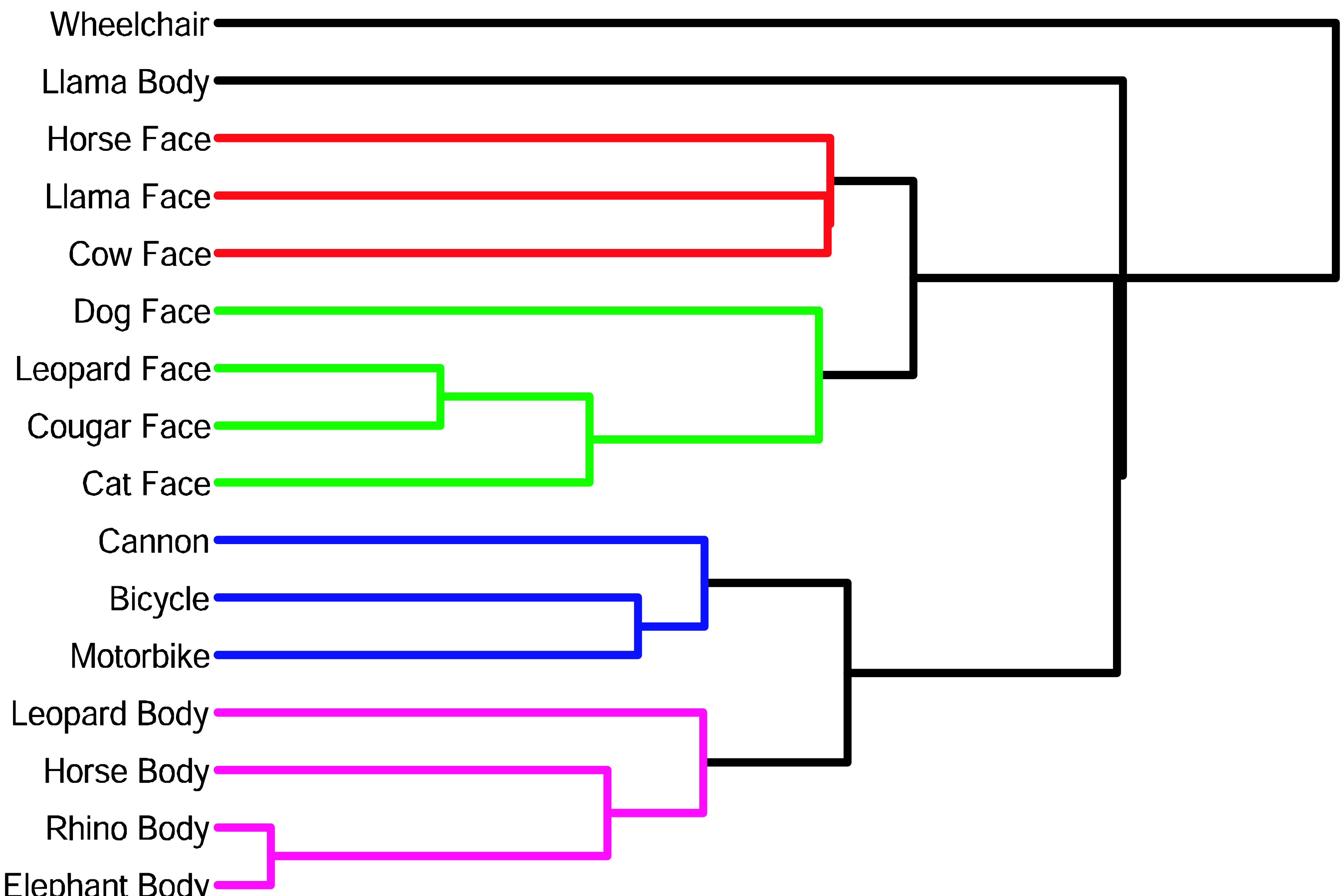




- Caltech 101 Dataset (Li & Perona)
- Horses (Borenstein & Ullman)
- Cat & dog faces (Vidal-Naquet & Ullman)

- Bikes from Graz-02 (Opelt & Pinz)
- Google…

Visualization of Part Densities

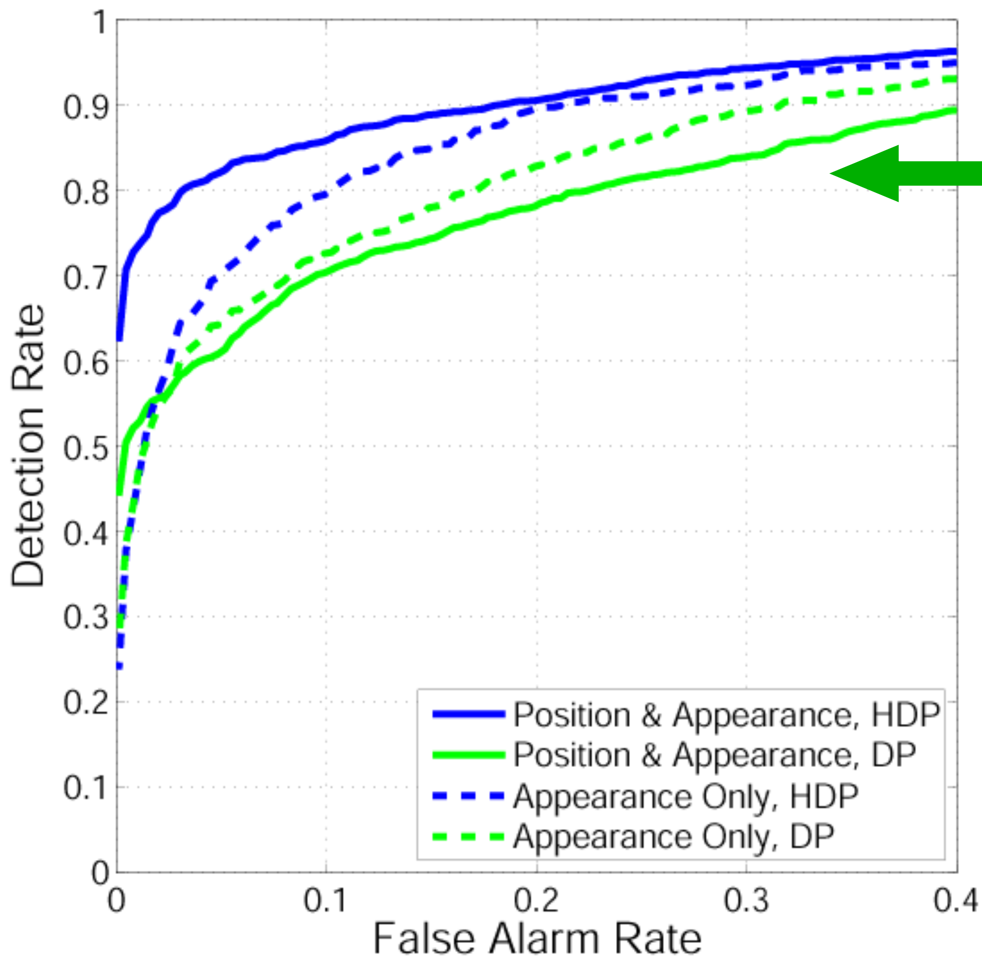


Hierarchical Clustering of Pr(part | object)

Detection Task

versus

Detection Results

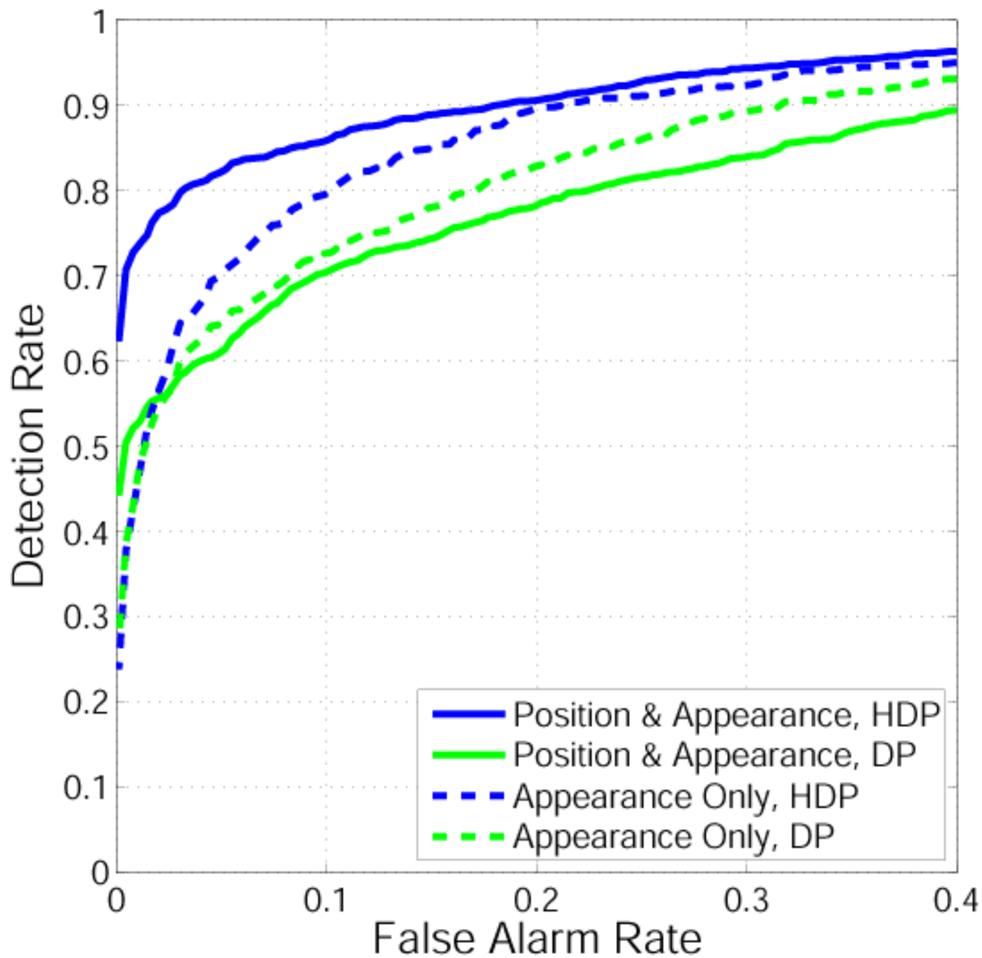


6 Training Images per Category
(ROC Curves)

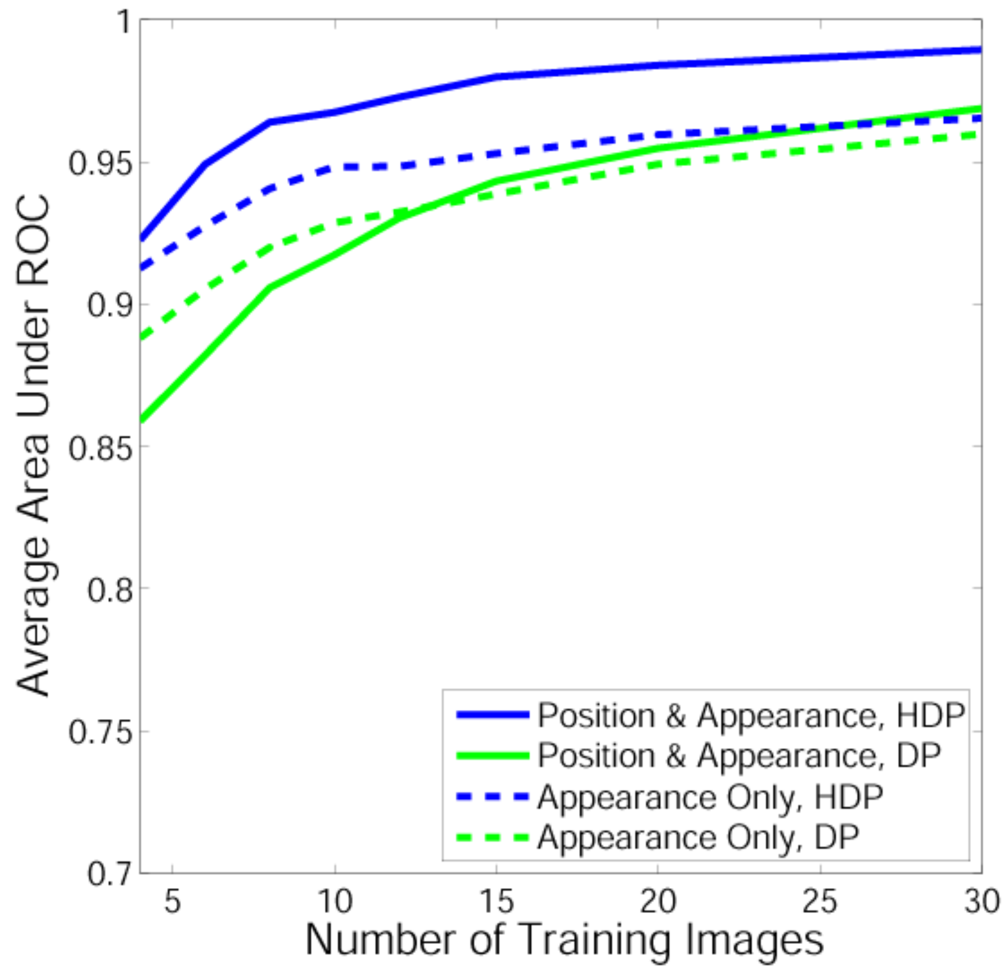
Shared Parts
more accurate than
Unshared Parts

Modeling feature positions ***improves shared*** detection, but ***hurts unshared*** detection

Detection Results

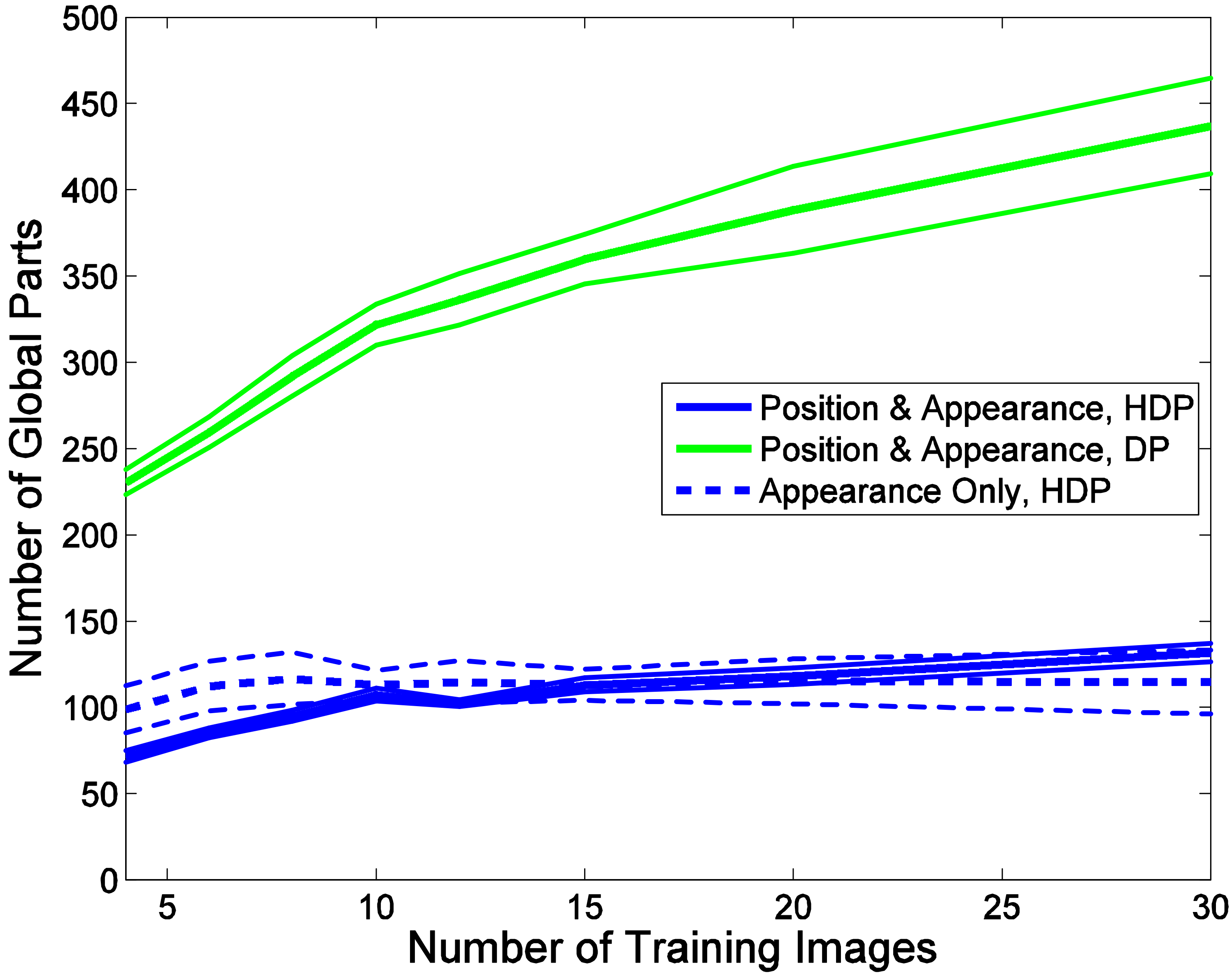


6 Training Images per Category
(ROC Curves)

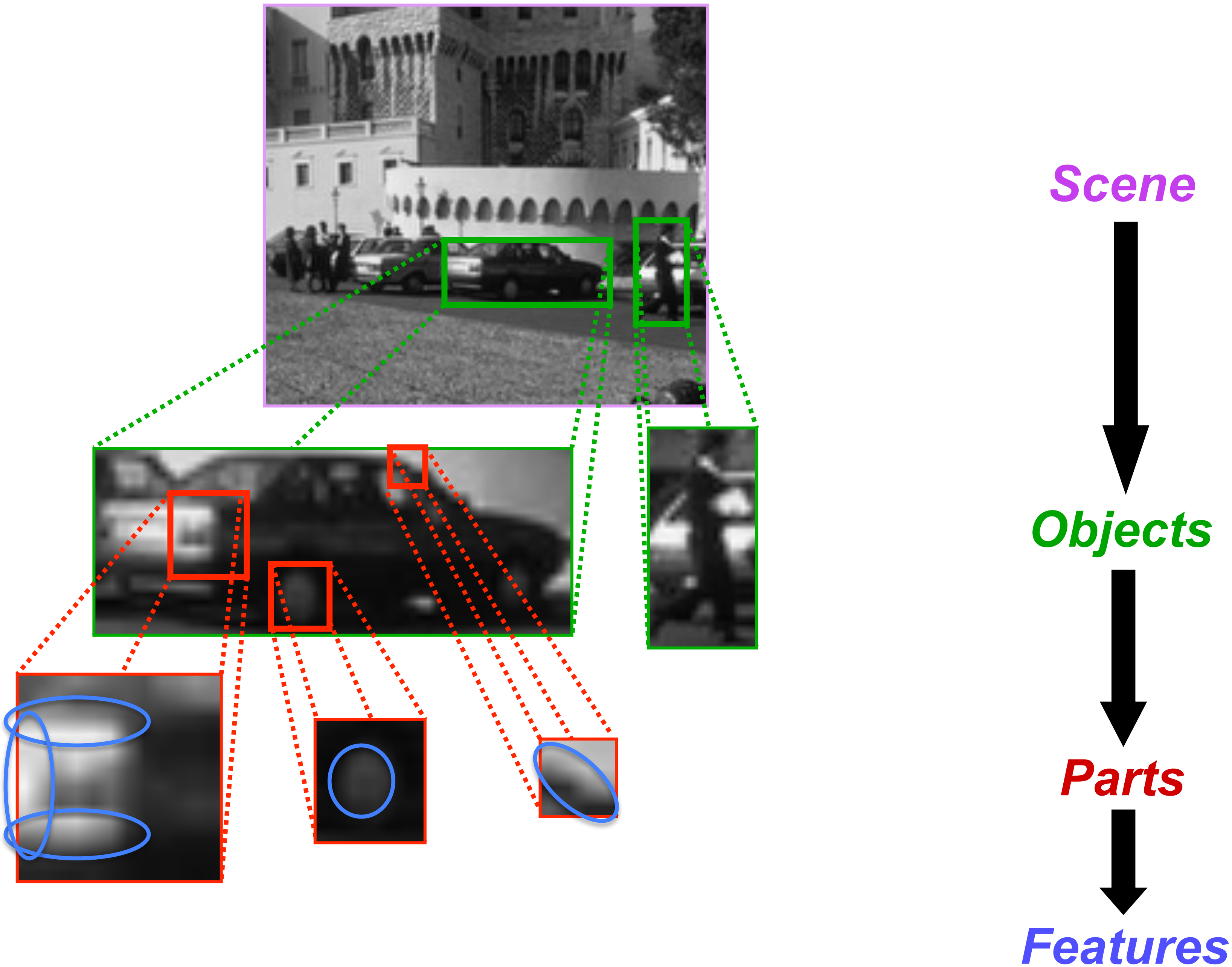


Detection vs. Training Set Size
(Area Under ROC)

Sharing Simplifies Models

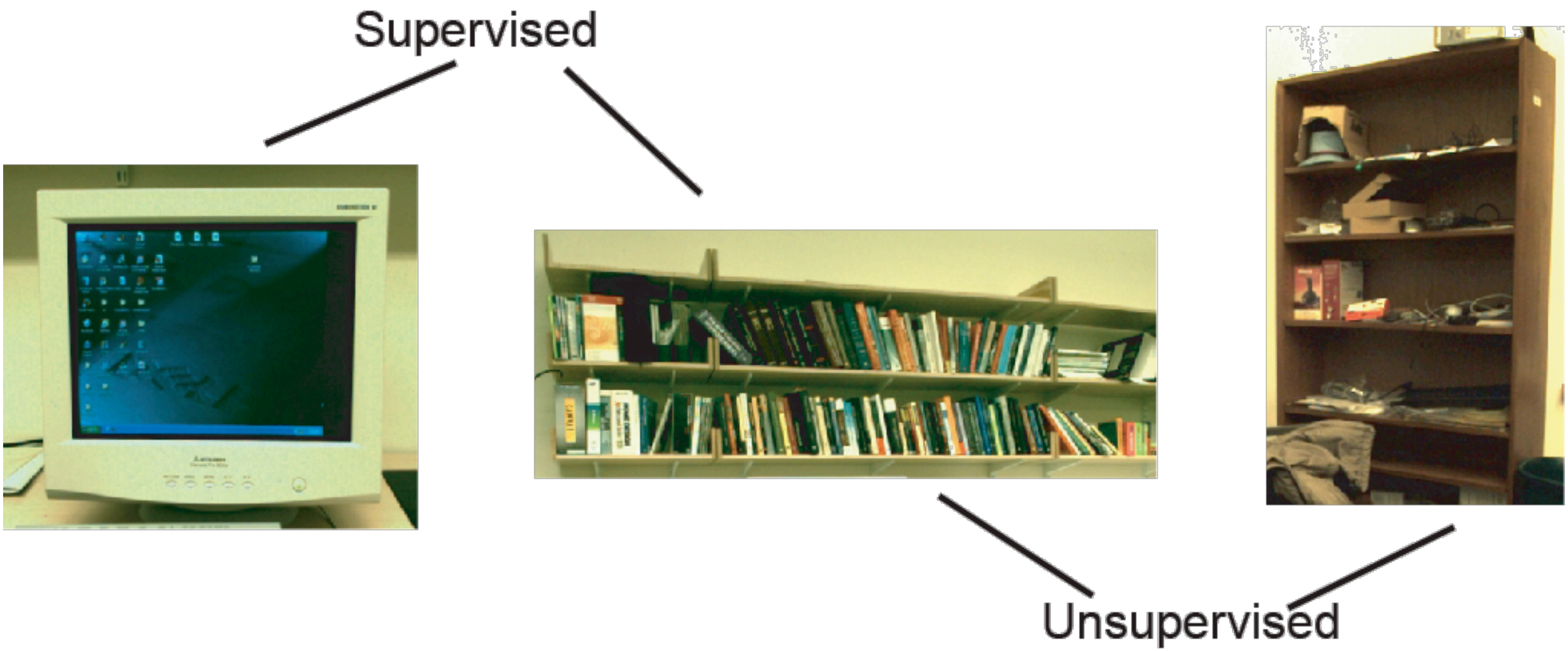


Scenes, Objects, and Parts



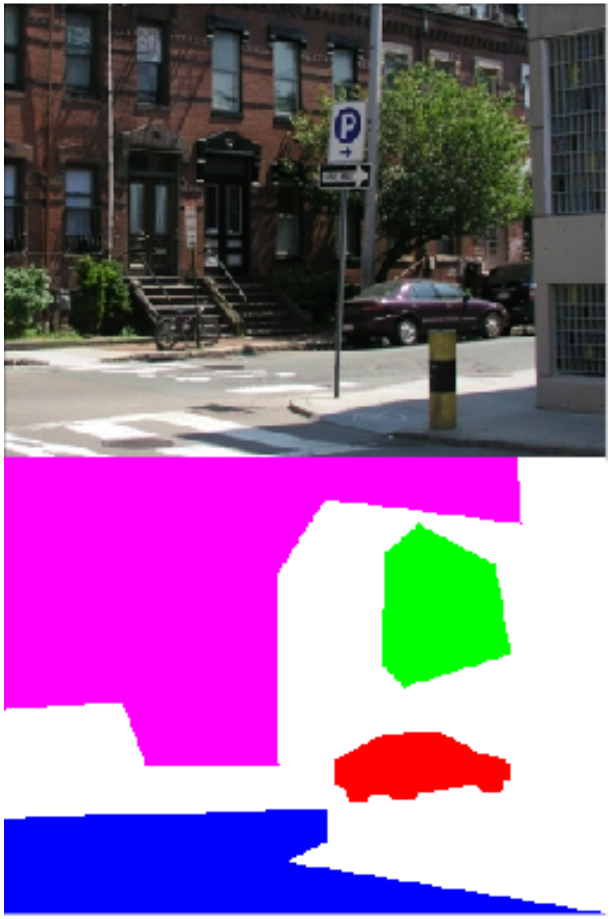
Contextual Transfer Learning

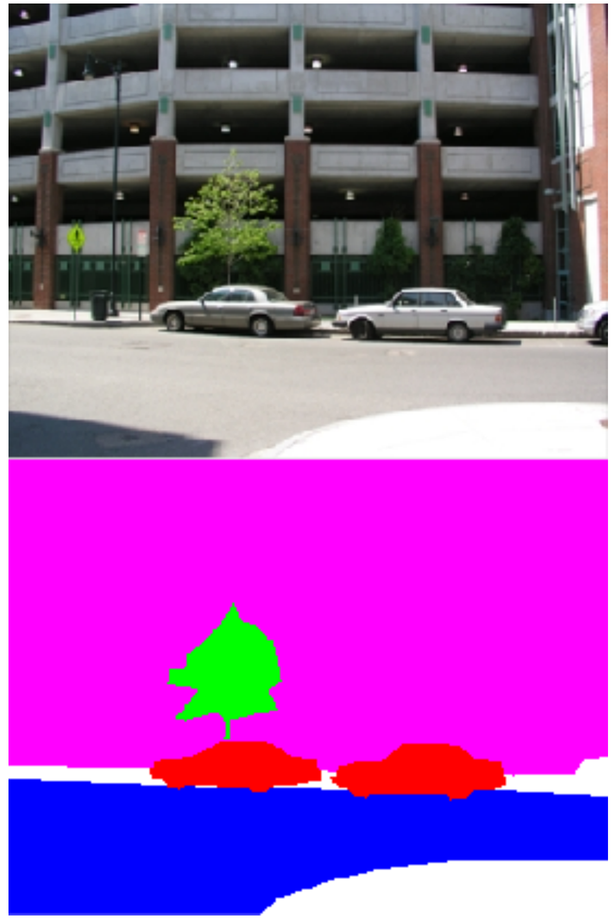
Object vs. Visual Categories

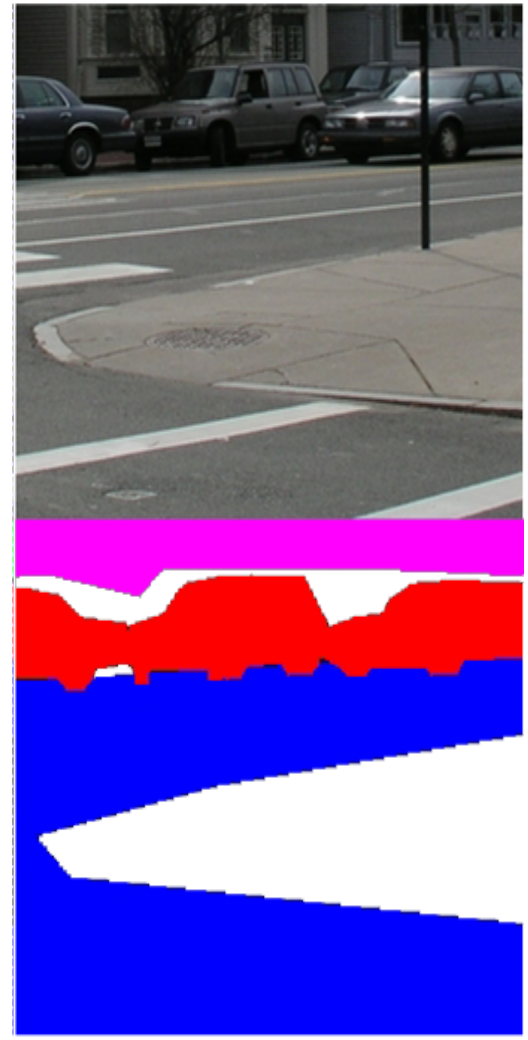


- Assume training data contains object category labels
- Discover underlying visual categories automatically

Multiple Object Scenes







- How many cars are there?
- Where are those cars in the scene?

Standard dependent Dirichlet process models (Gelfand et. al., 2005) inappropriate

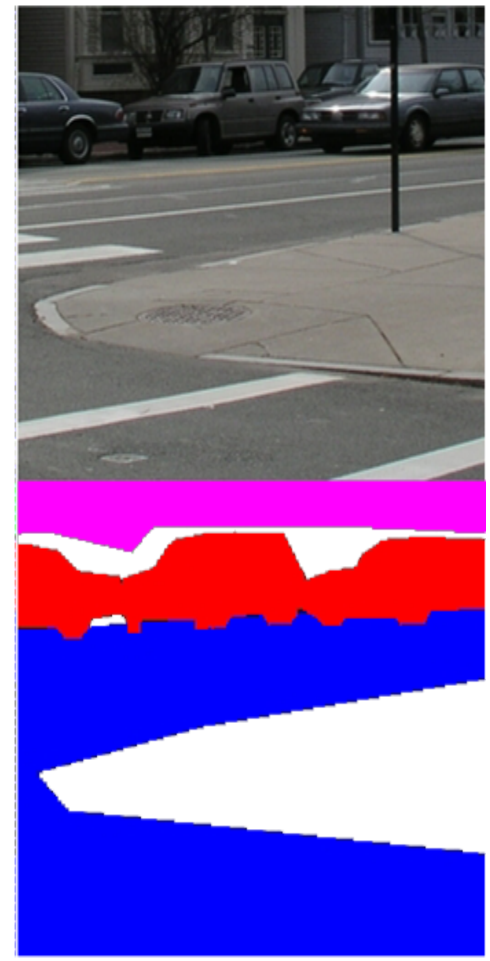
Spatial Transformations

- Let global DP clusters model objects in a *canonical* coordinate frame
- Generate images via a random *set of transformations:*

$$\tau((\mu, \Lambda); \rho) = (\mu + \rho, \Lambda)$$

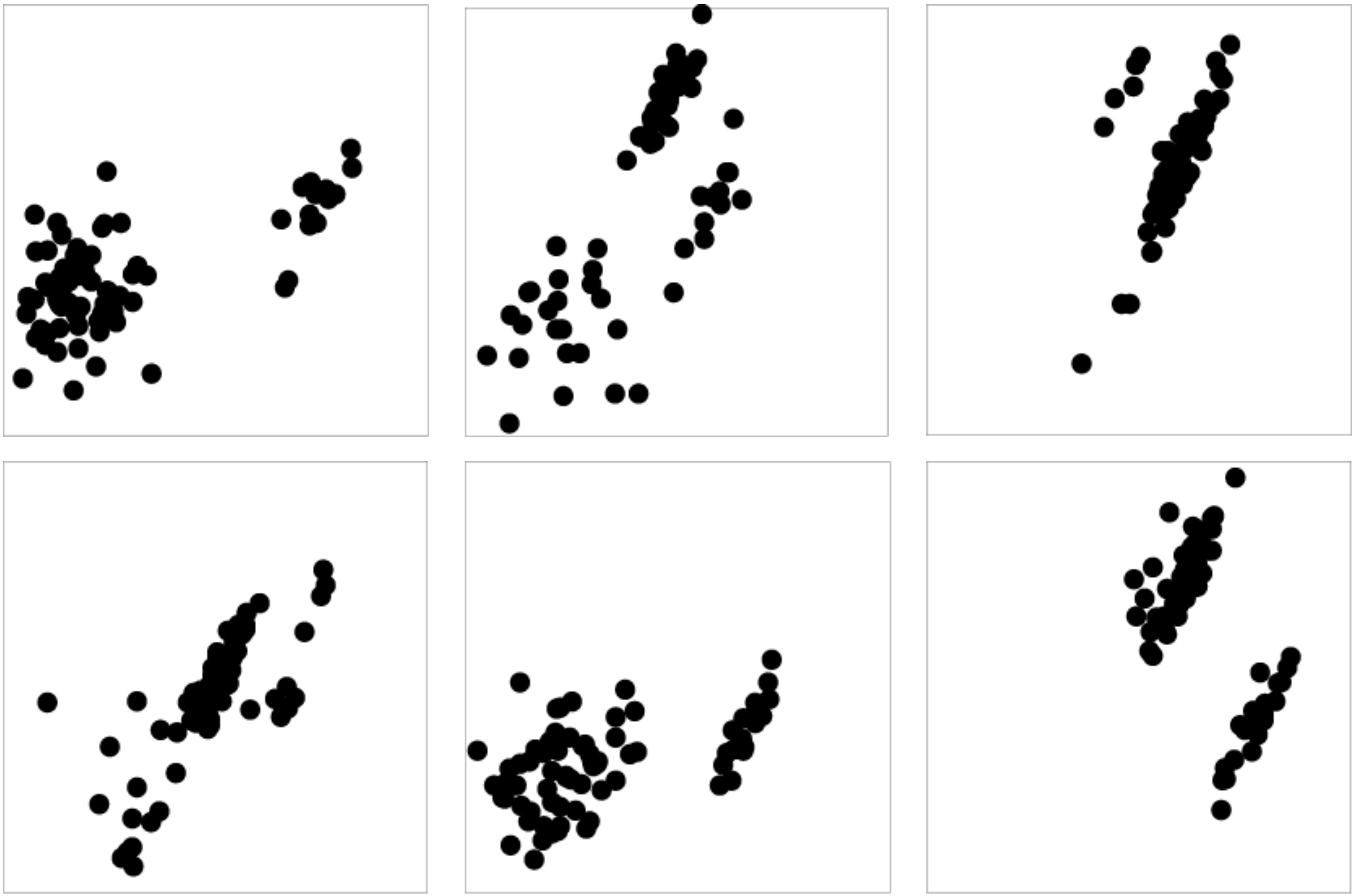
Parameterized family of transformations

Shift cluster from canonical coordinate frame to object location in a given image

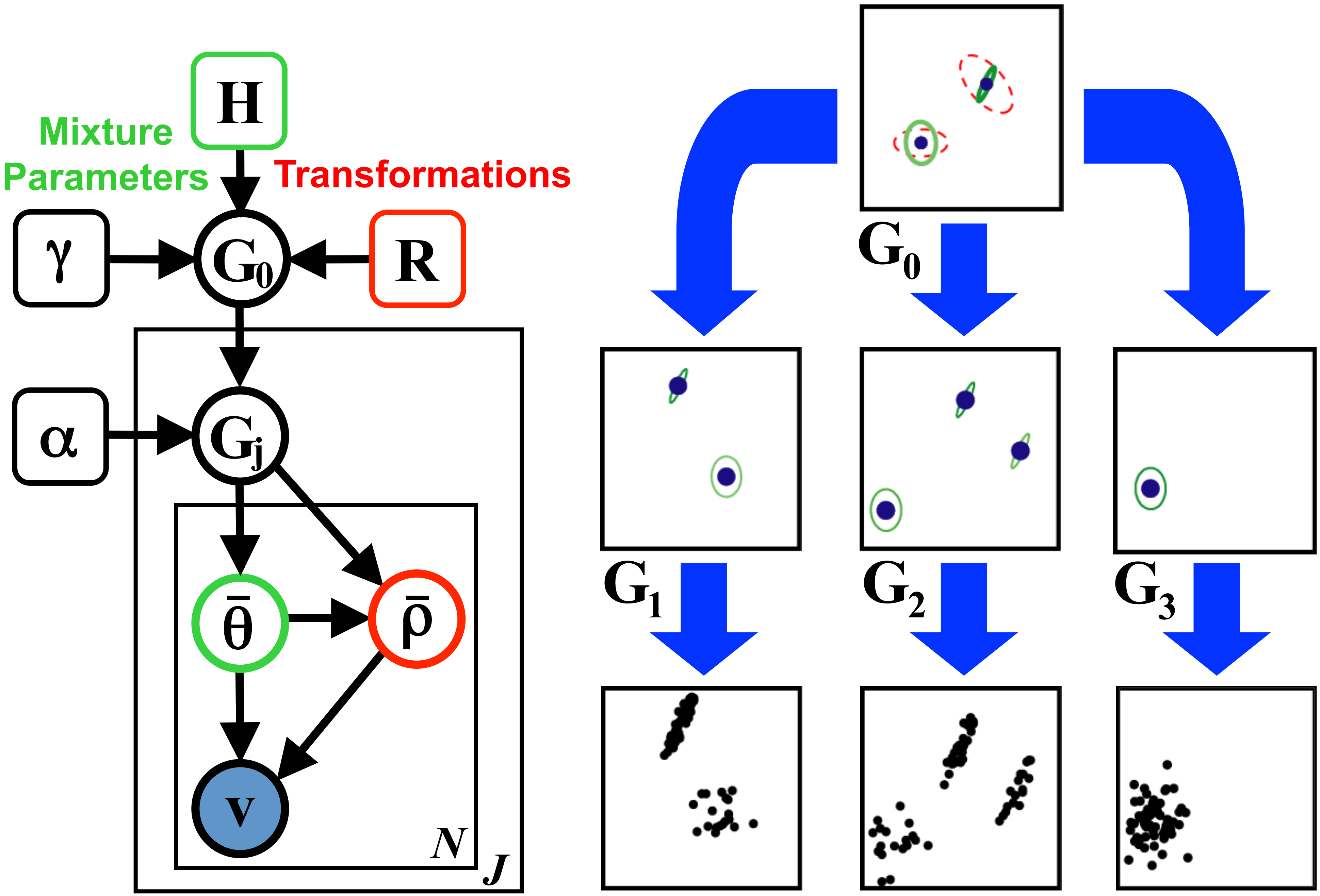


Layered Motion Models *(Darrell & Pentland 1991, Wang & Adelson 1994, Jojic & Frey 2001)*
Nonparametric Transformation Densities *(Learned-Miller & Viola 2000)*

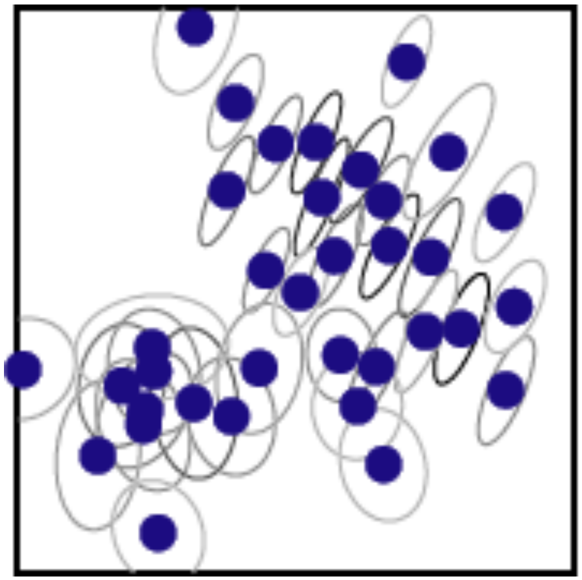
A Toy World: Bars & Blobs



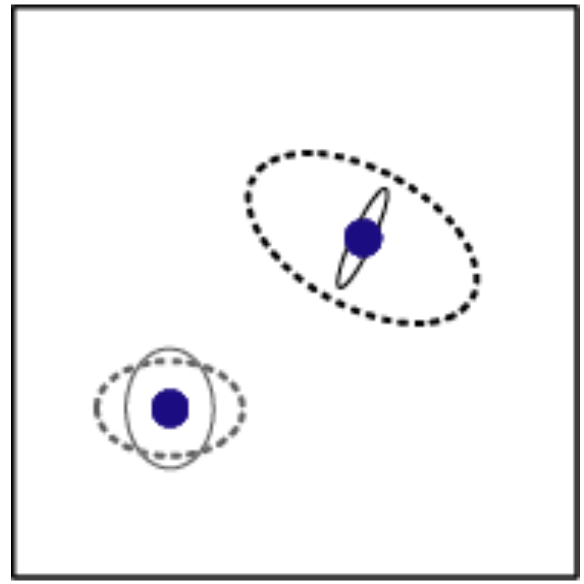
Transformed Dirichlet Process



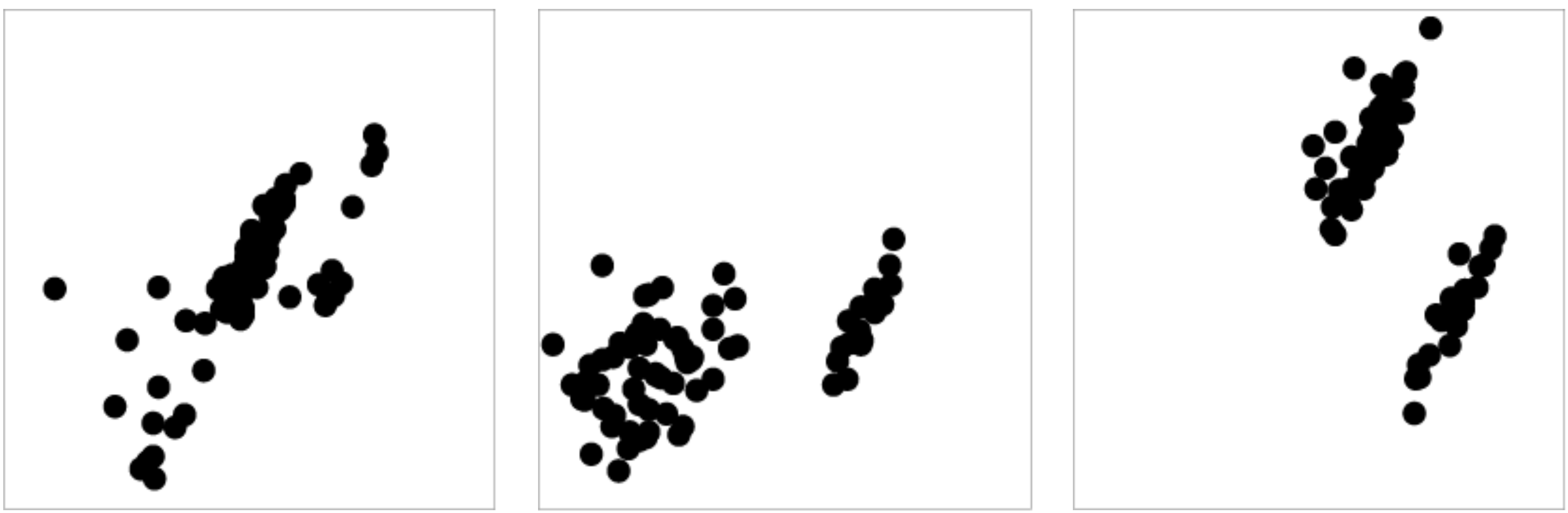
Importance of Transformations



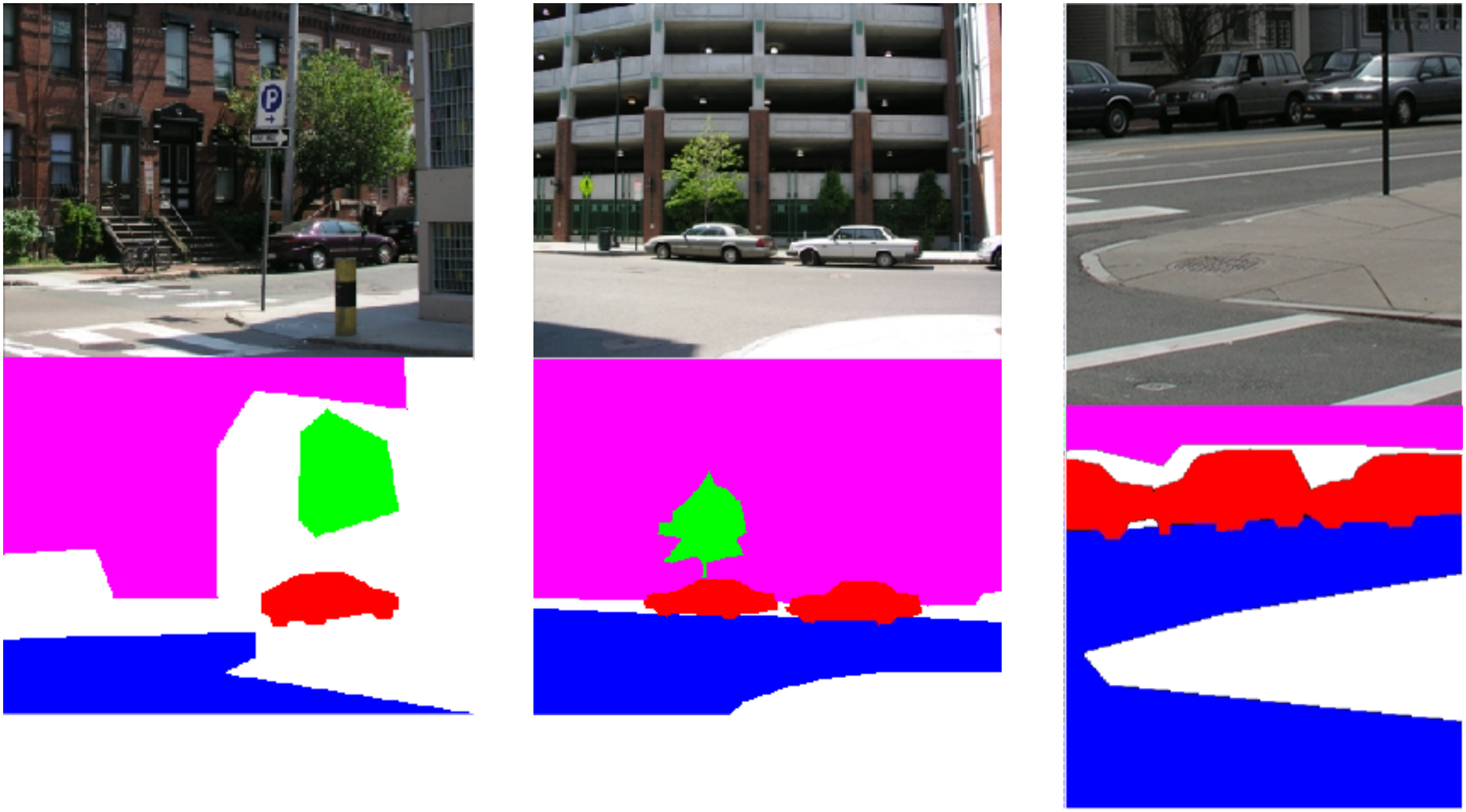
HDP



TDP



Counting & Locating Objects

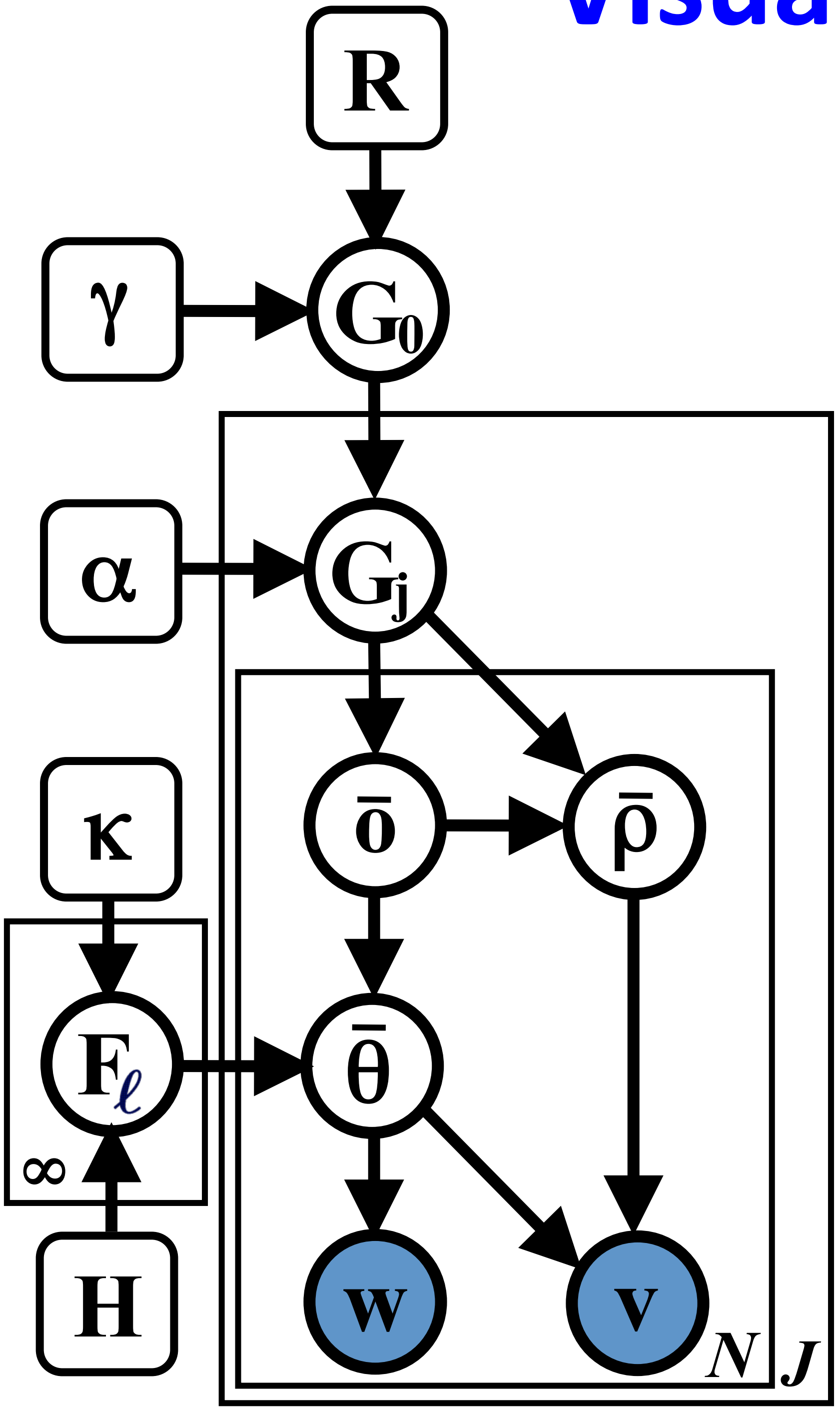


- How many cars are there?
- Where are those cars in the scene?

Dirichlet Processes

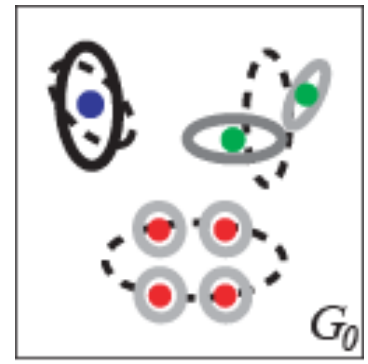
Transformations

Visual Scene TDP



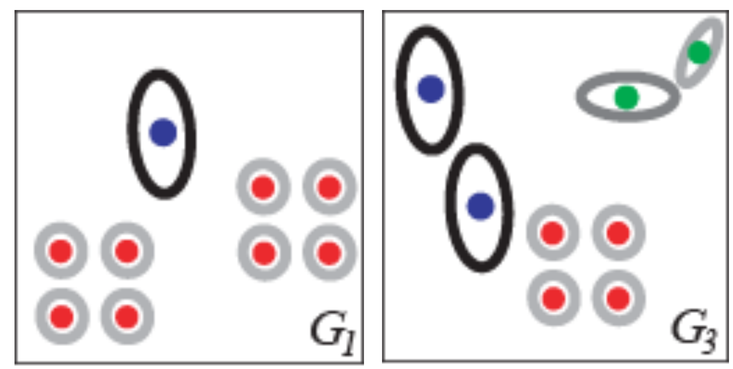
Global Density

Object category
Part size & shape
Transformation prior



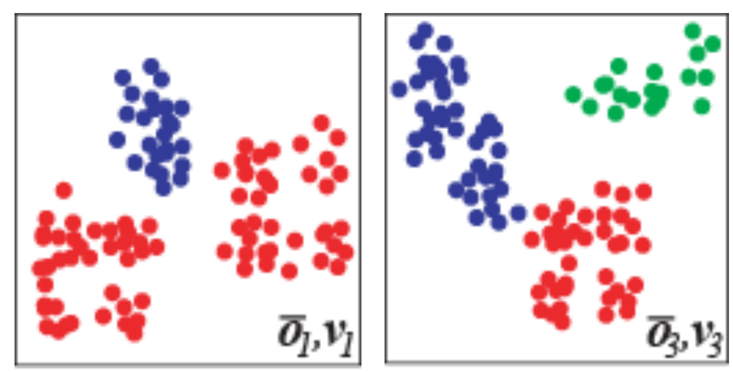
Transformed Densities

Object category
Part size & shape
Instance locations

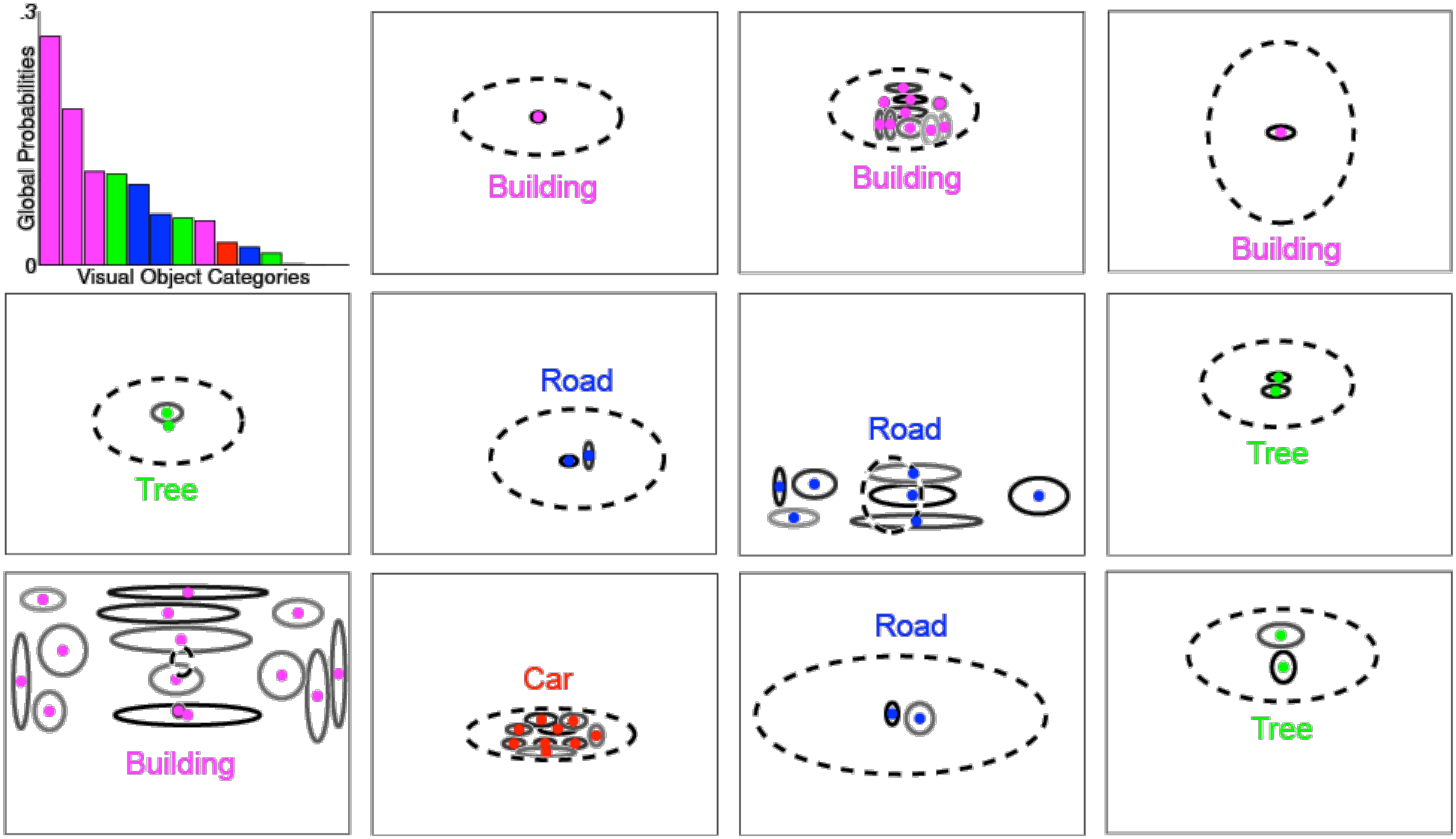


2D Image Features

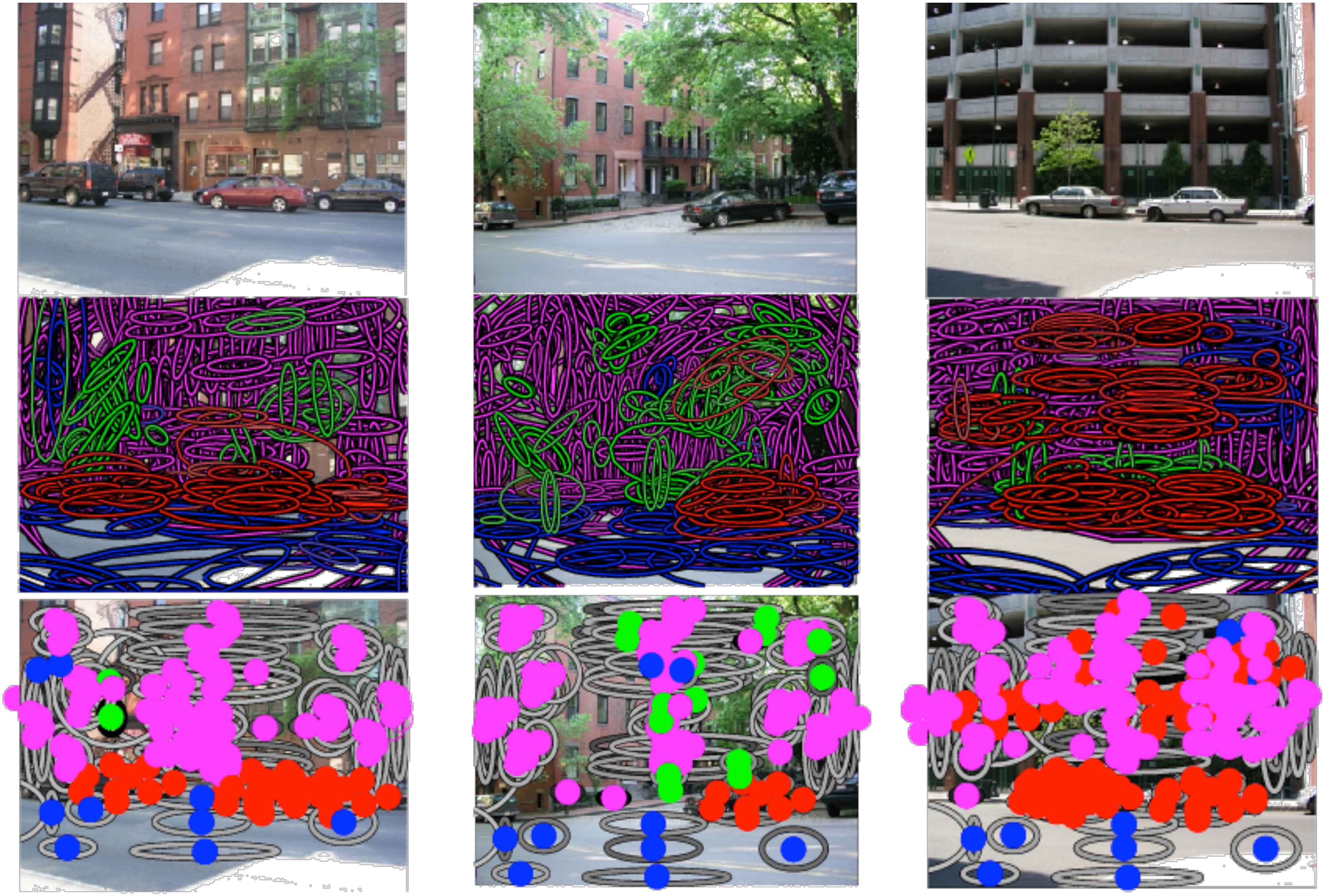
Appearance
Location



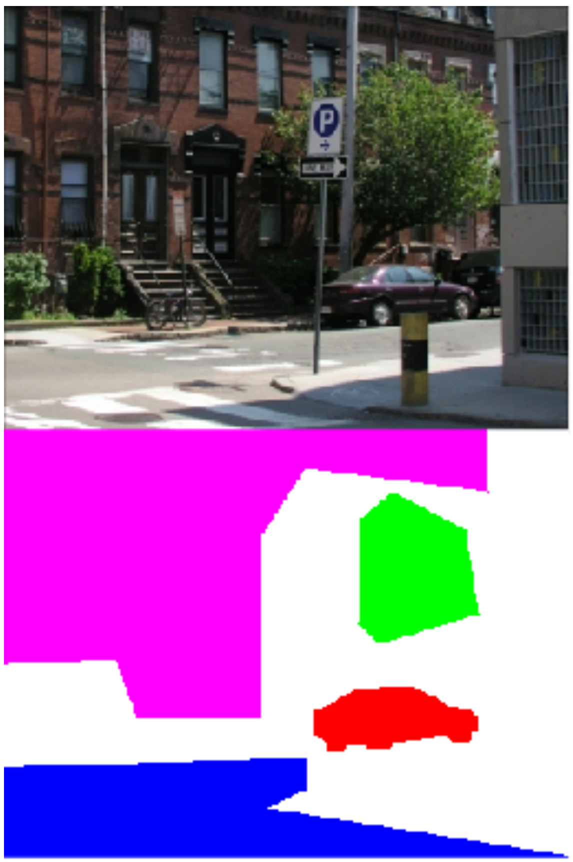
Street Scene Visual Categories

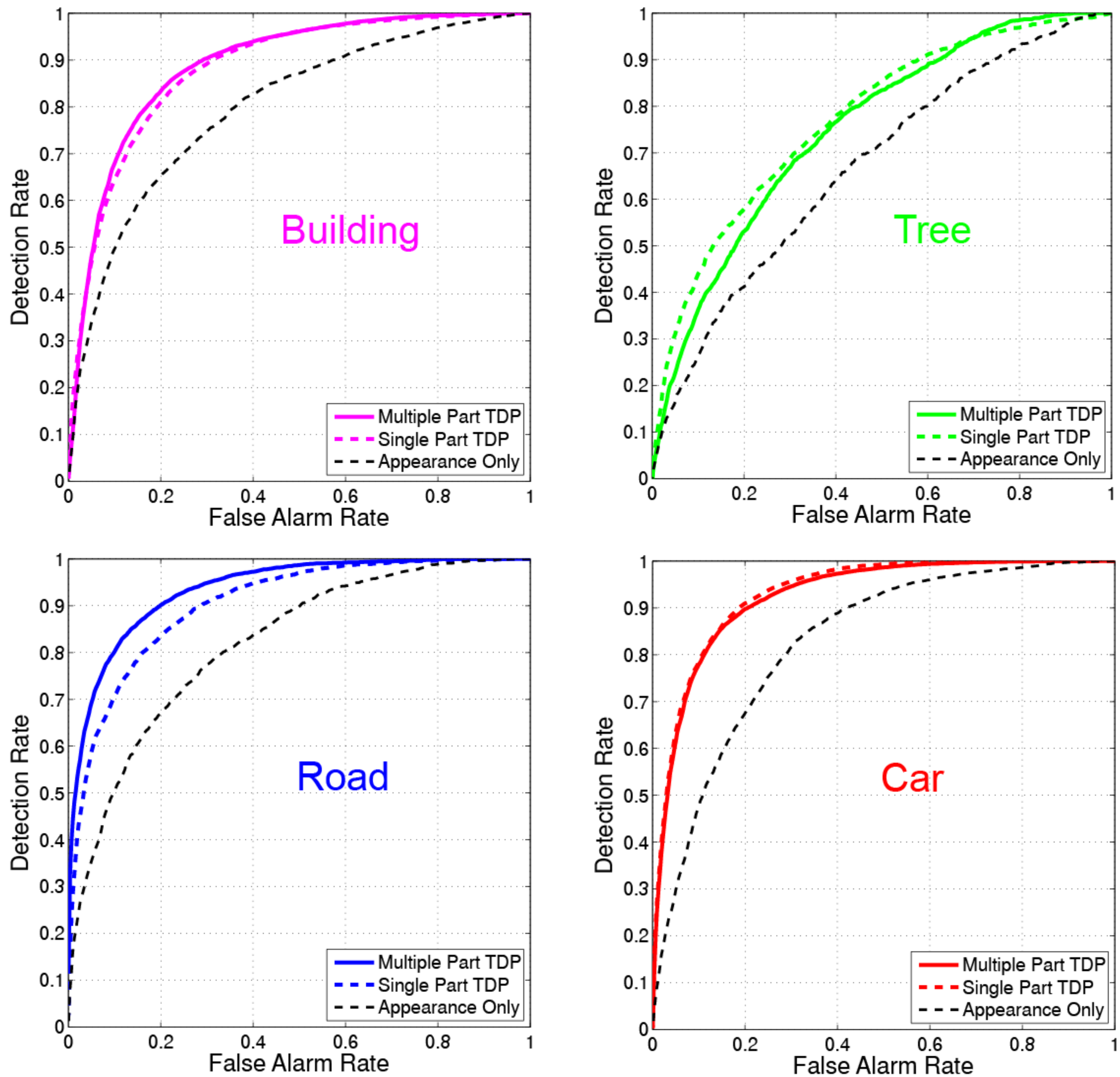


Street Scene Segmentations



Segmentation Performance

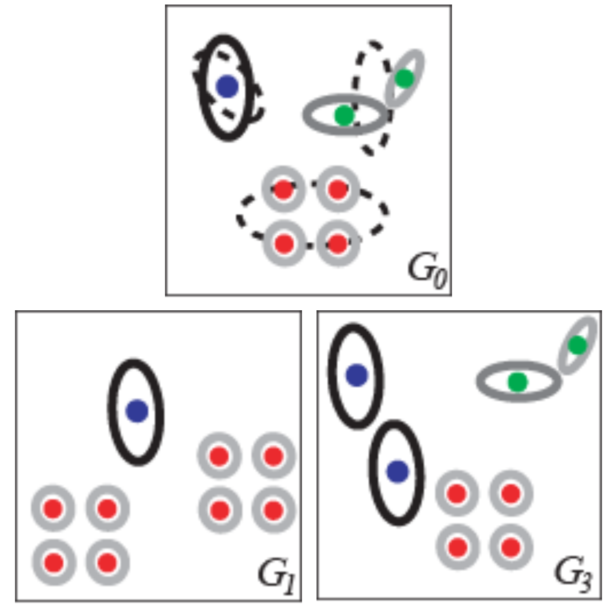




Outline

Topics

- Bag of feature image representations
- Hierarchical Bayesian modeling

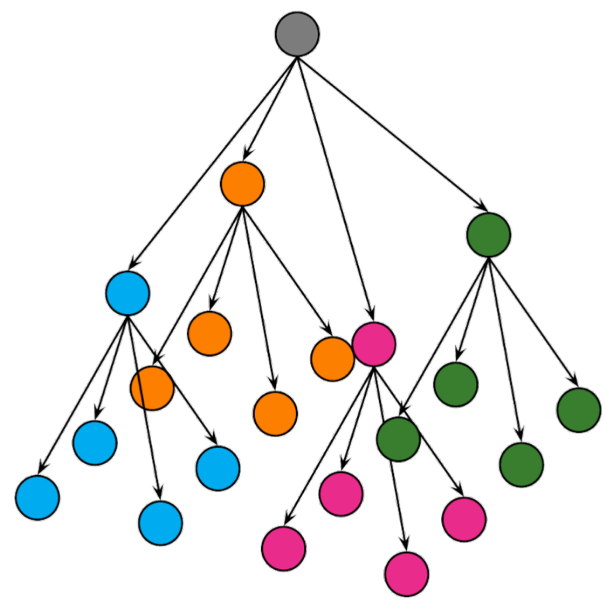


Transformations

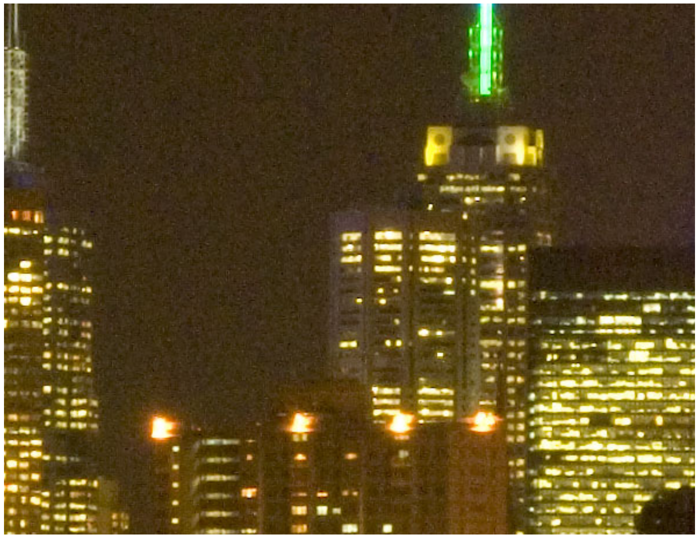
- Sharing parts among object categories
- Spatial models for visual scenes

Trees

- Multiscale nonparametric Markov models
- Image denoising and scene categorization



Low-level Image Analysis



Noise Removal

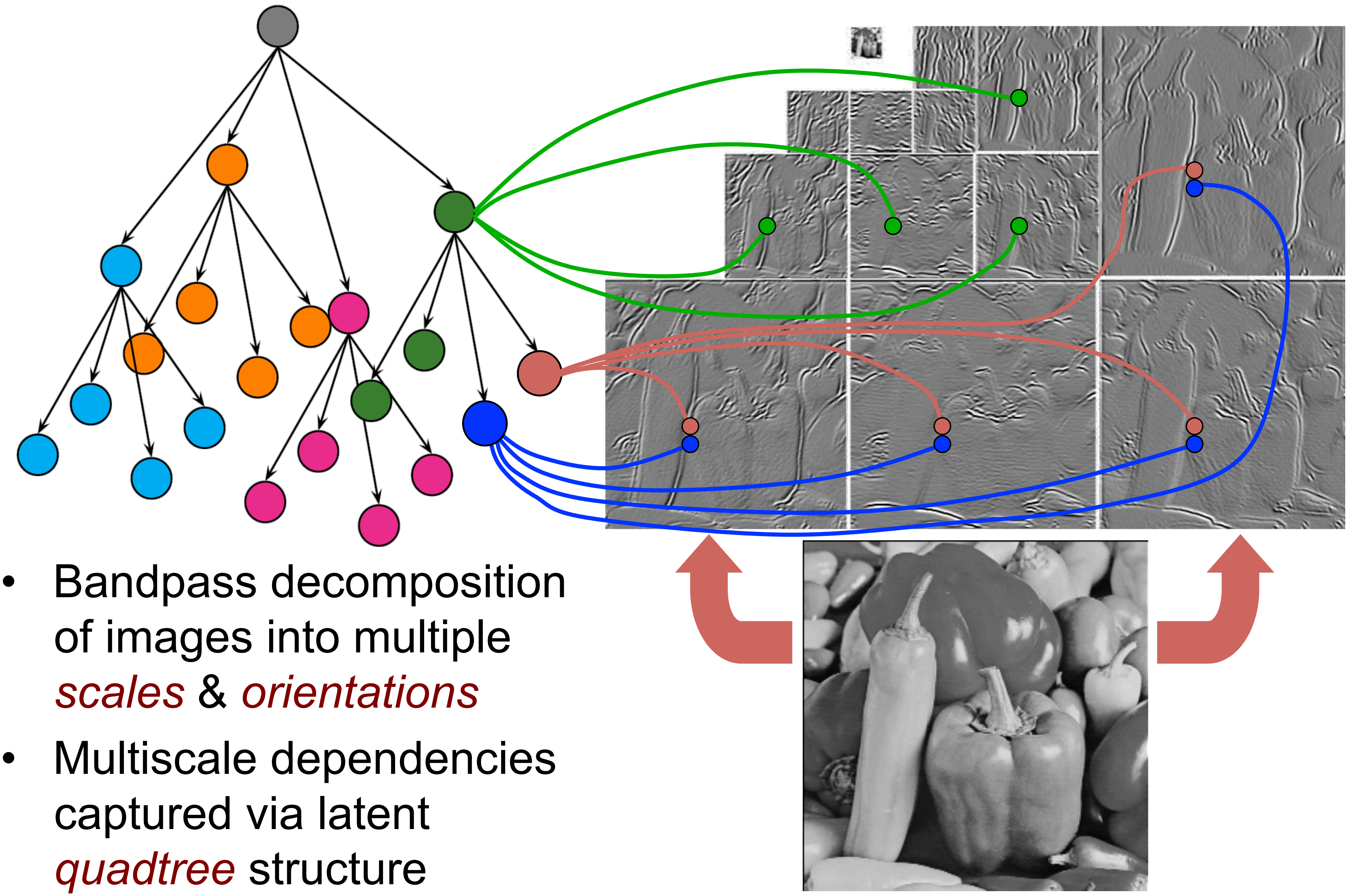
Deblurring

Inpainting & Restoration

Goals:

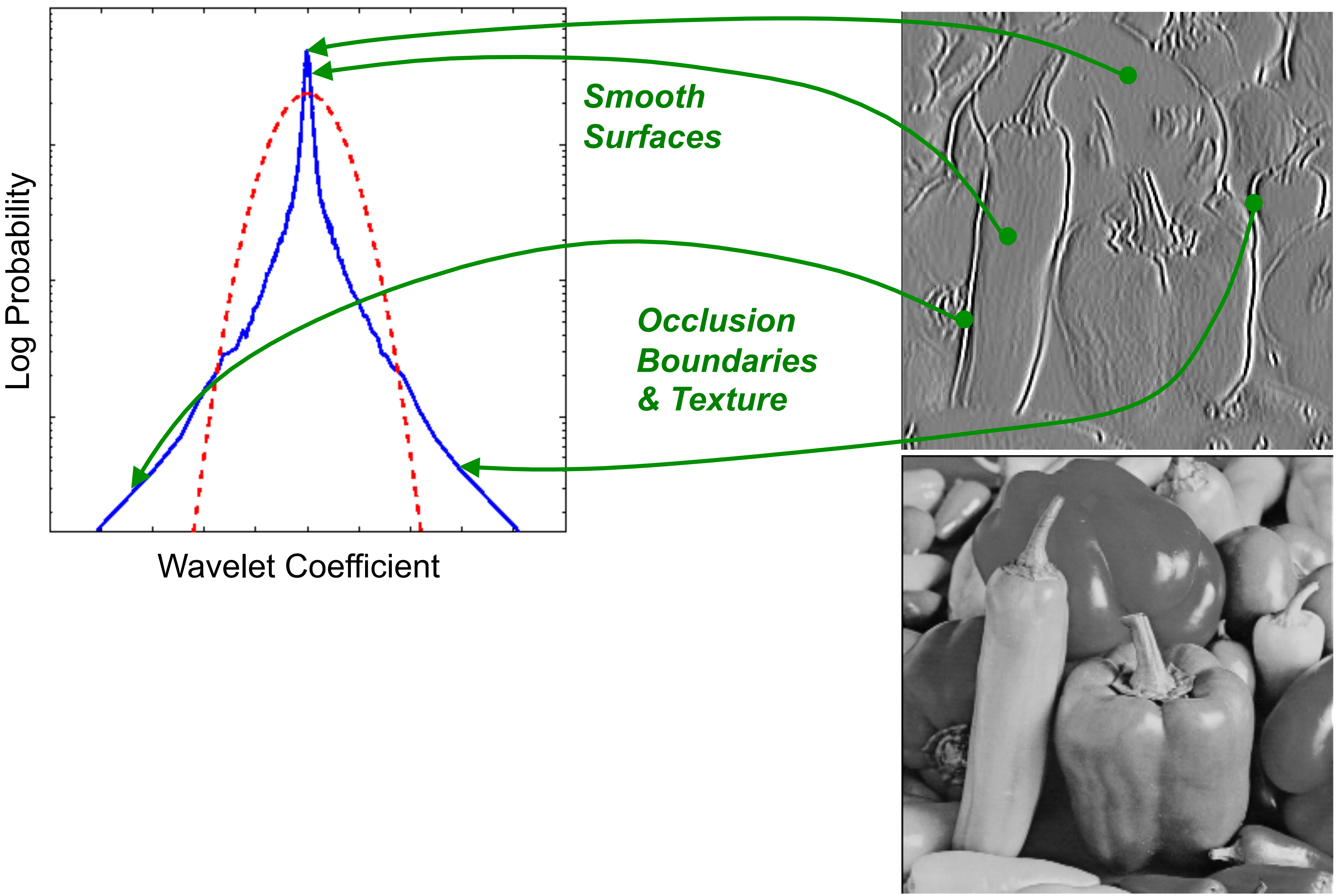
- Accurately model the statistics of *natural images*
- Exploit the availability of large digital *image collections*

Wavelet Decompositions

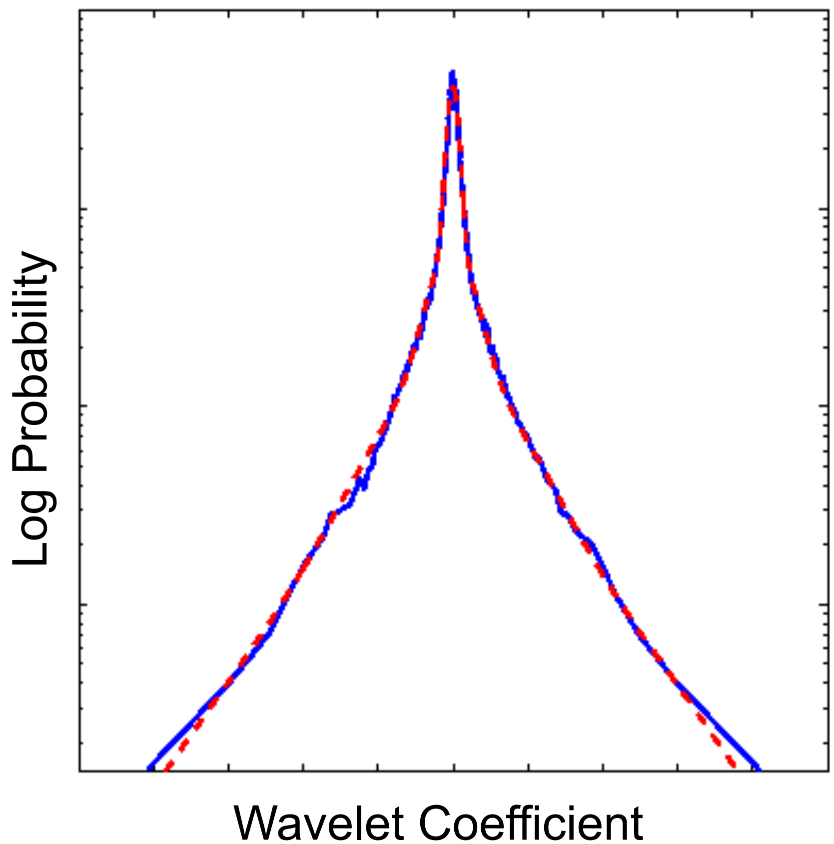


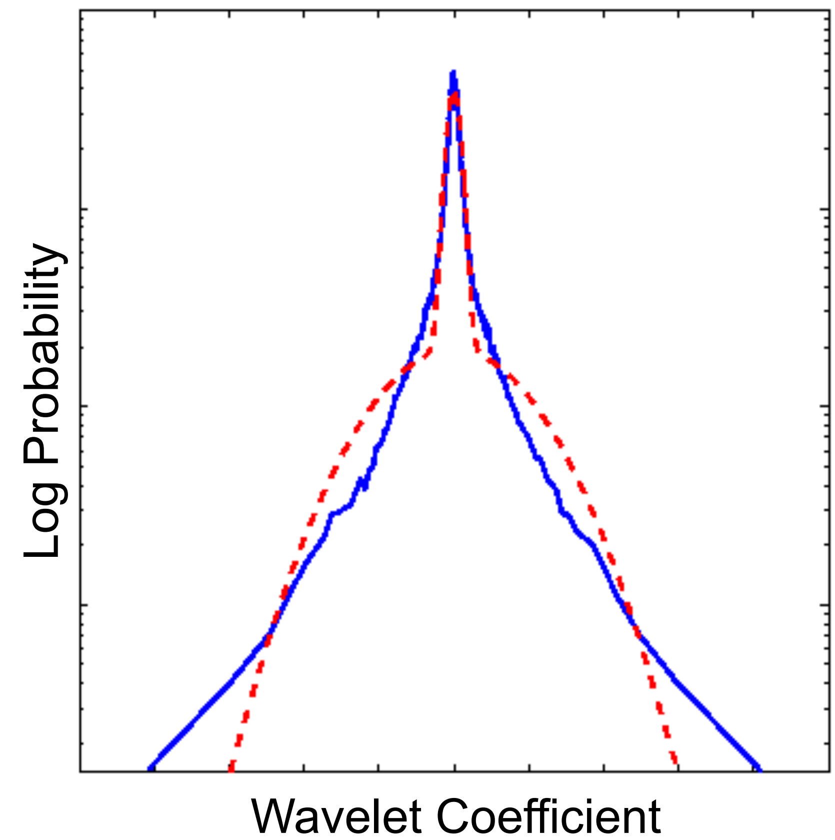
- Bandpass decomposition of images into multiple *scales* & *orientations*
- Multiscale dependencies captured via latent *quadtree* structure

Wavelets: Marginal Statistics



Gaussian Mixture Models





$$x_i = v_i u_i$$

$$v_i \geq 0 \qquad u_i \sim \mathcal{N}(0, \Lambda)$$

Gaussian Scale Mixture (GSM)

Wainwright & Simoncelli, 2000

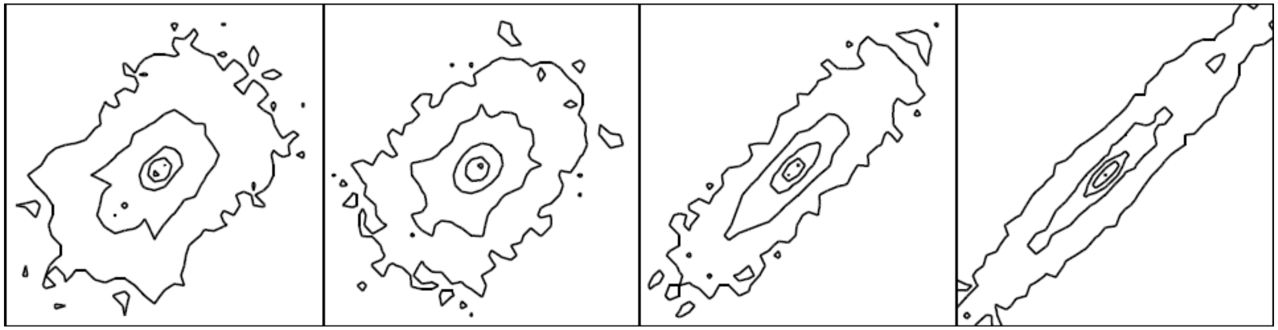
$$x_i \sim \pi\,\mathcal{N}(0, \Lambda_0) + (1-\pi)\mathcal{N}(0, \Lambda_1)$$

Binary Gaussian Mixture

Computational advantages…

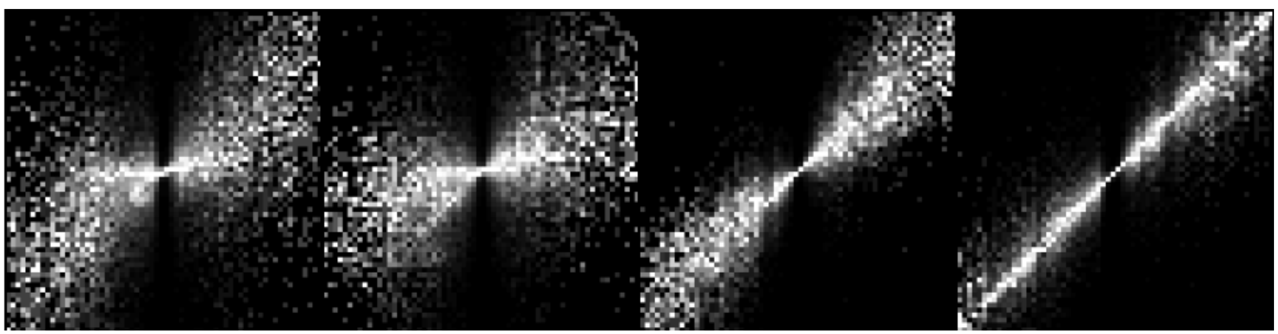
Wavelets: Joint Statistics

Pairwise Joint Histograms:

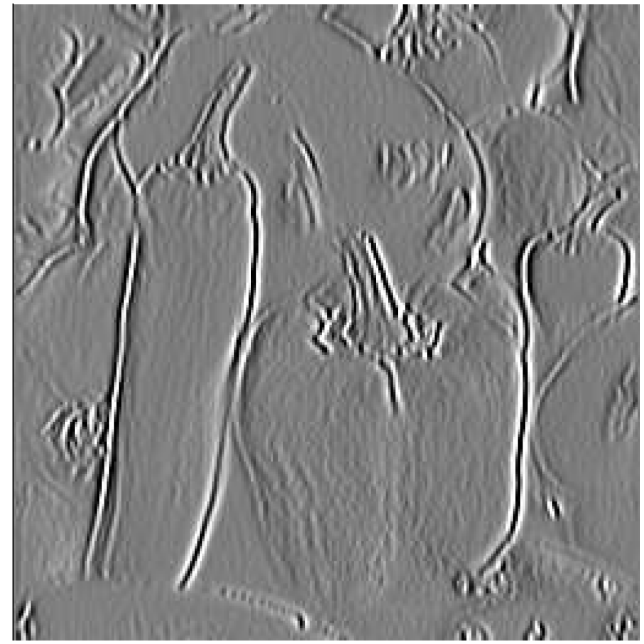


Orientation *Scale* *Vertical* *Horizontal*

Pairwise Conditional Histograms:



Orientation *Scale* *Vertical* *Horizontal*

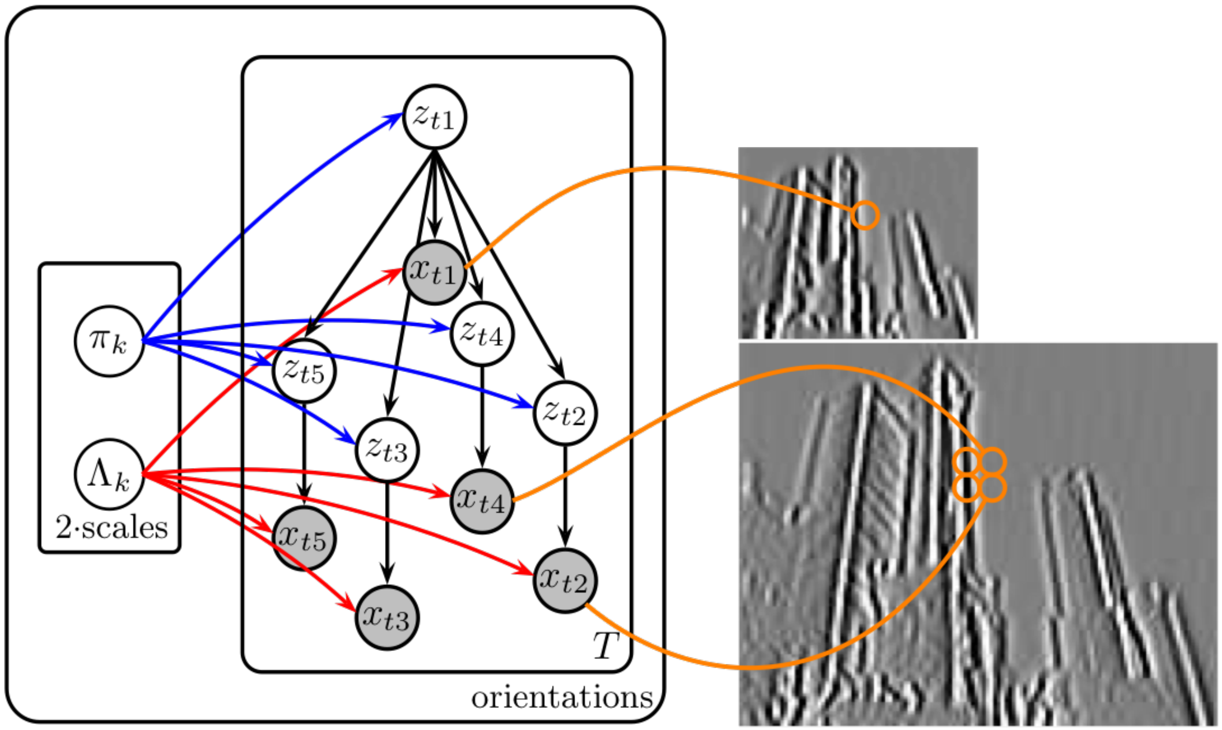


Large magnitude wavelet coefficients…

- *Persist* across multiple scales
- *Cluster* at adjacent spatial locations

Binary Hidden Markov Trees

Crouse, Nowak, & Baraniuk, 1998



π_k → state *transition* distributions

$$z_{ti} \sim \pi_{z_{\mathrm{Pa}(ti)}}$$

Λ_k → state-specific *emission* covariances

$$x_{ti} \sim \mathcal{N}\left(0, \Lambda_{z_{ti}}\right)$$

z_{ti} → hidden *state* or cluster assignment

$$z_{ti} \in \{0, 1\}$$

x_{ti} → *observed* wavelet coefficient

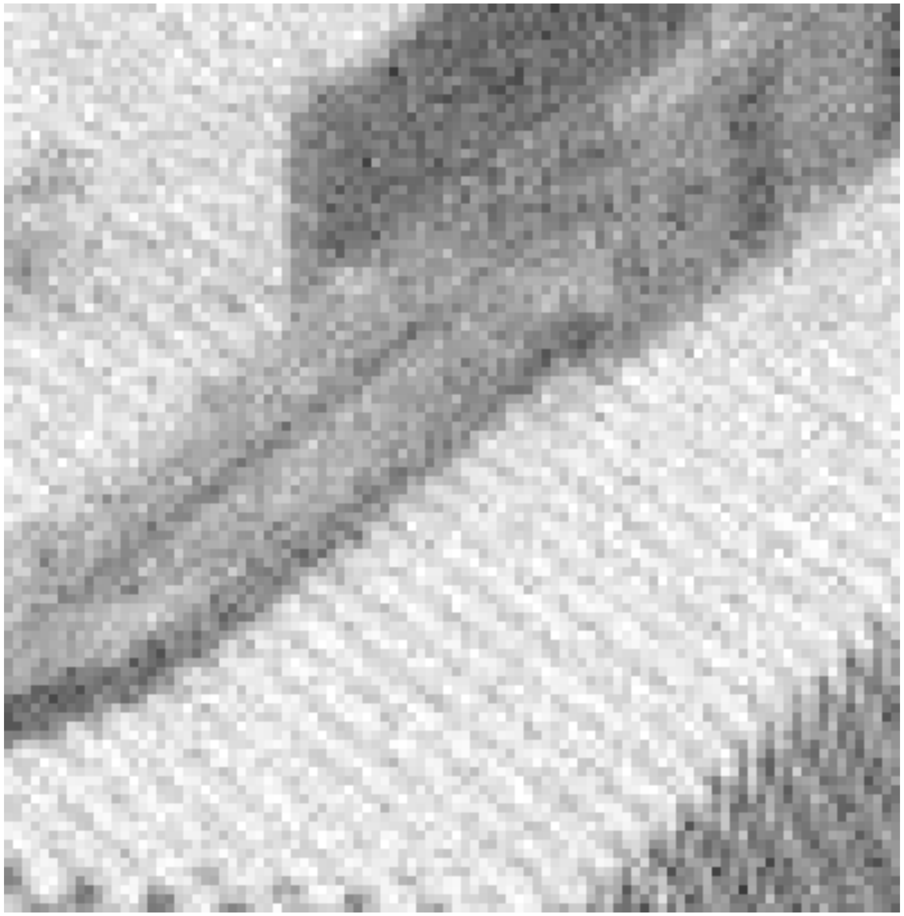
- Coefficients marginally distributed as mixtures of two Gaussians
- Markov dependencies between hidden states capture persistence of image contours across locations and scales
- Each orientation is modeled independently

Validation : Image Denoising

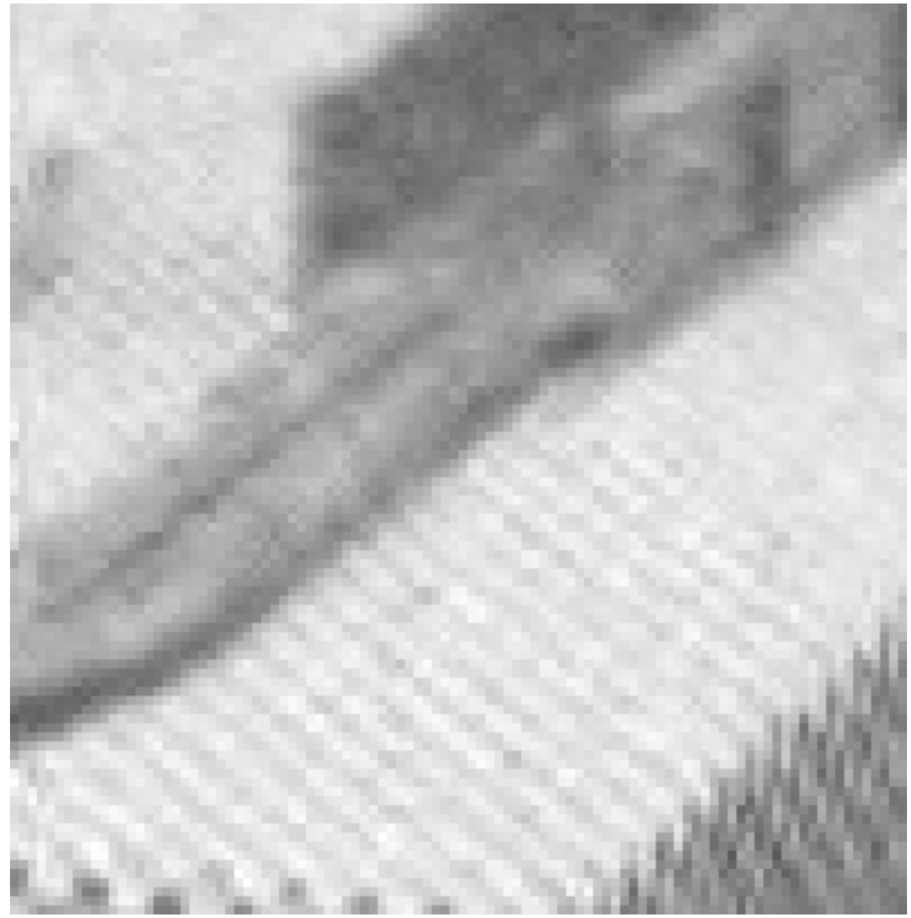
Original

Corrupted by Additive White Gaussian Noise (PSNR = 24.61 dB)

Denoising with Binary HMTs



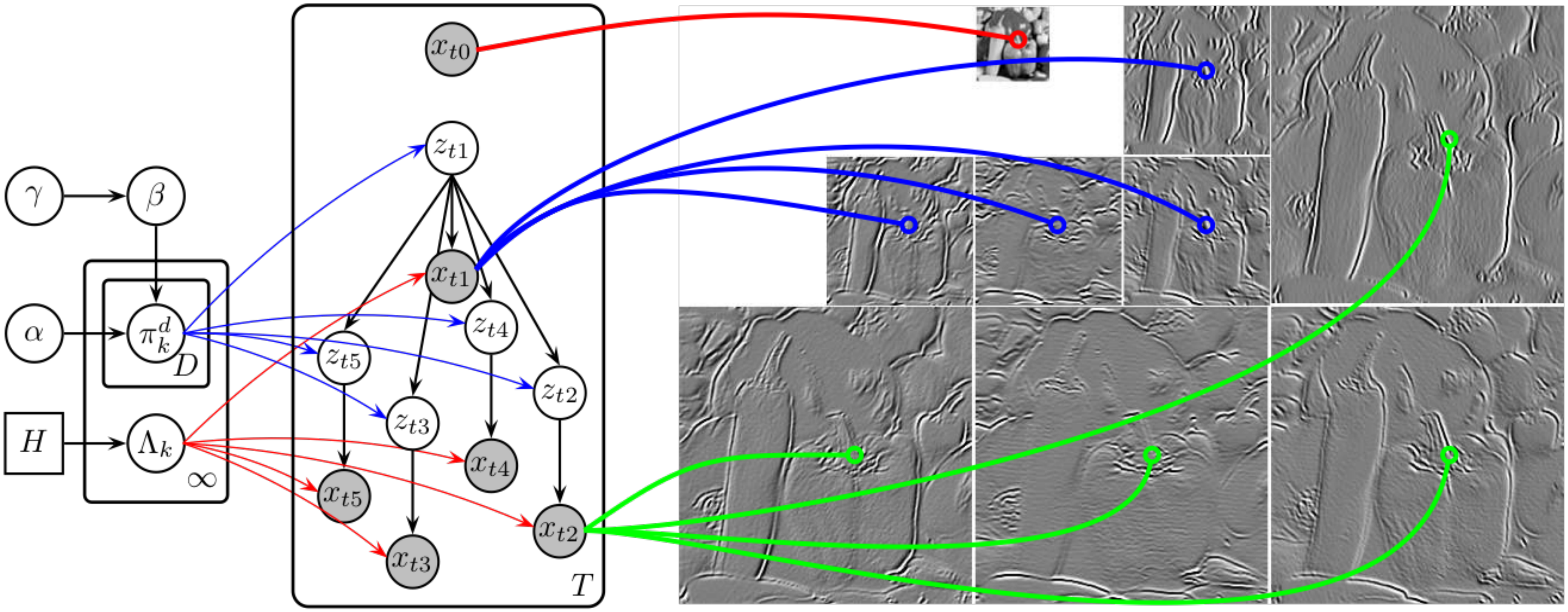
Noisy Input



Denoised (EM algorithm)

- Is two states per scale sufficient? How many is enough?
- Should states be shared the same way for all images, or for all wavelet decompositions?

Hierarchical Dirichlet Process Hidden Markov Trees



$z_{ti} \longrightarrow$ indexes *infinite* set of hidden states

$$z_{ti} \in \{1, 2, 3, \ldots\}$$

$\pi_k \longrightarrow$ infinite set of state *transition* distributions

$$z_{ti} \sim \pi^{d_{ti}}_{z_{\mathrm{Pa(ti)}}}$$

$x_{ti} \longrightarrow$ observed *vector* of wavelet coefficients

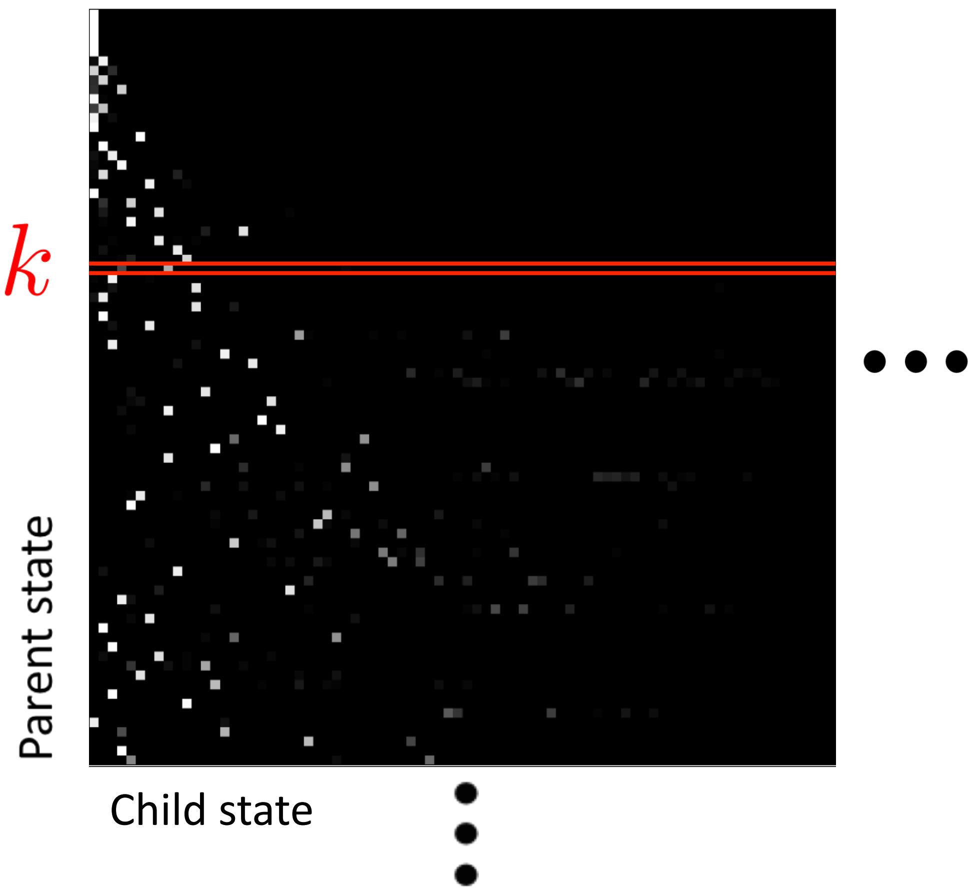
$\Lambda_k \longrightarrow$ state-specific *emission* covariances

$$x_{ti} \sim \mathcal{N}\left(0, \Lambda_{z_{ti}}\right)$$
$$\Lambda_k \sim H$$

Why a Hierarchical DP ? *(Teh et. al. 2004)*

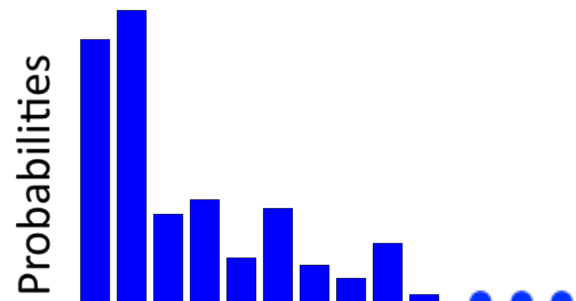
- Hierarchical DP prior allows us to learn a potentially infinite set of *appearance patterns* from natural images
- Hierarchical coupling ensures, with high probability, that a common set of *child* states are reachable from each *parent*

$$\pi_k^{d_{ti}}(\ell) = \Pr\left[z_{ti} = \ell \mid z_{\mathrm{Pa}(ti)}\right]$$



$$\beta \sim \mathrm{Stick}(\gamma)$$

Average state frequencies



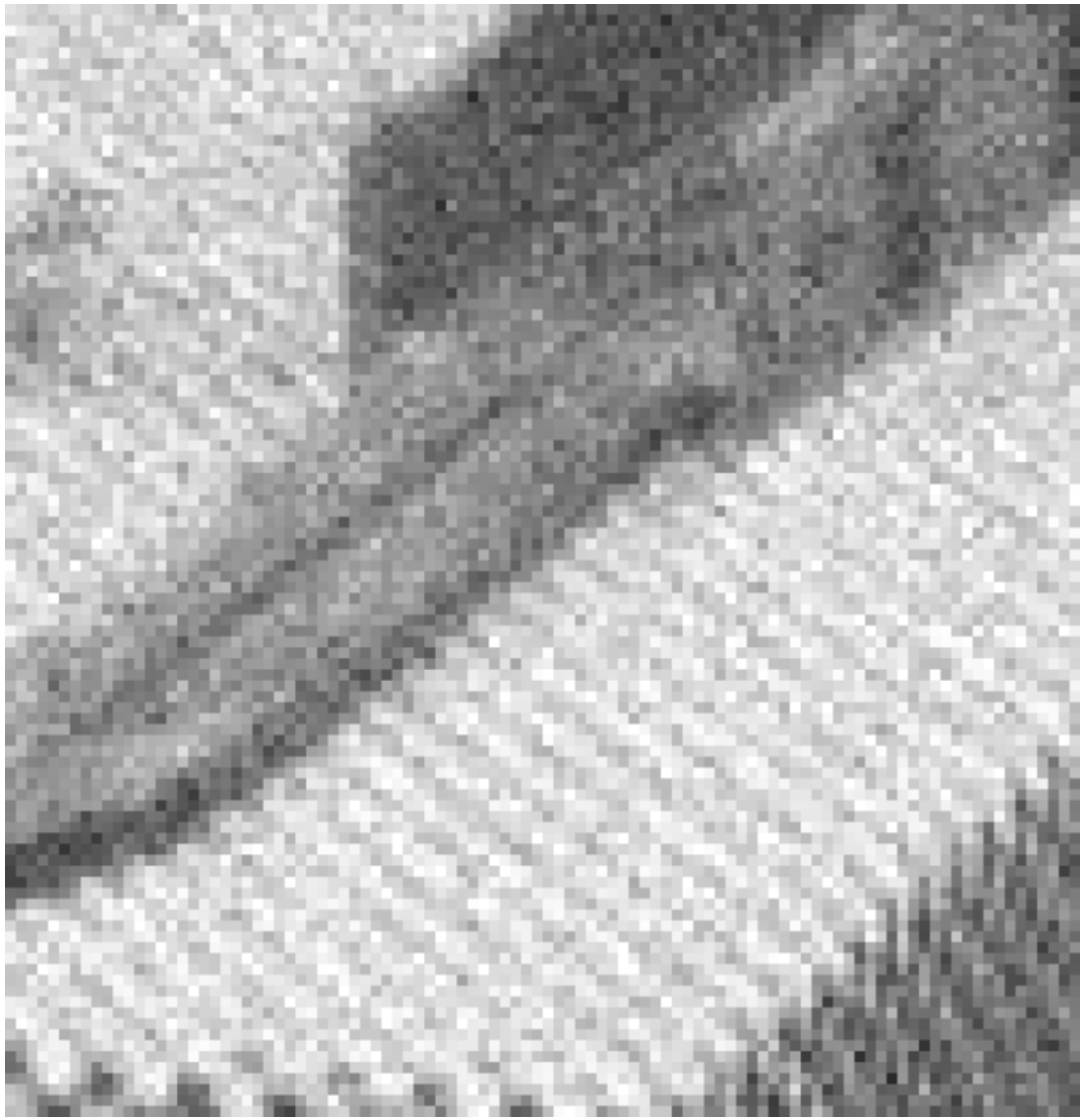
$$\pi_k^d \sim \mathrm{DP}(\alpha, \beta)$$

Transition distributions

$$\mathbb{E}\left[\pi_k^d\right] = \beta$$

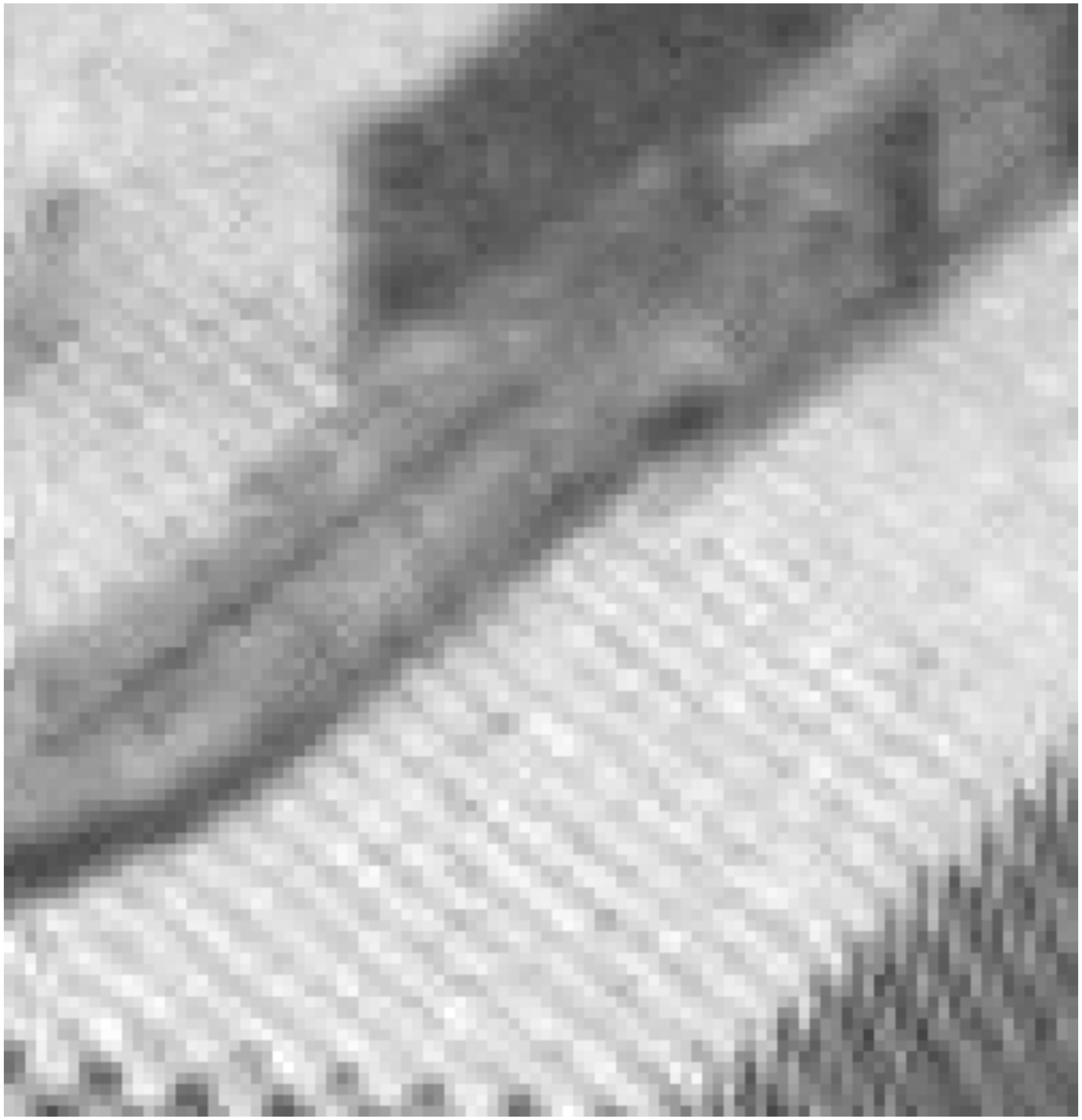
$\alpha \longrightarrow$ *Sparsity & variability of transition distributions*

Denoising: Input



24.61 dB

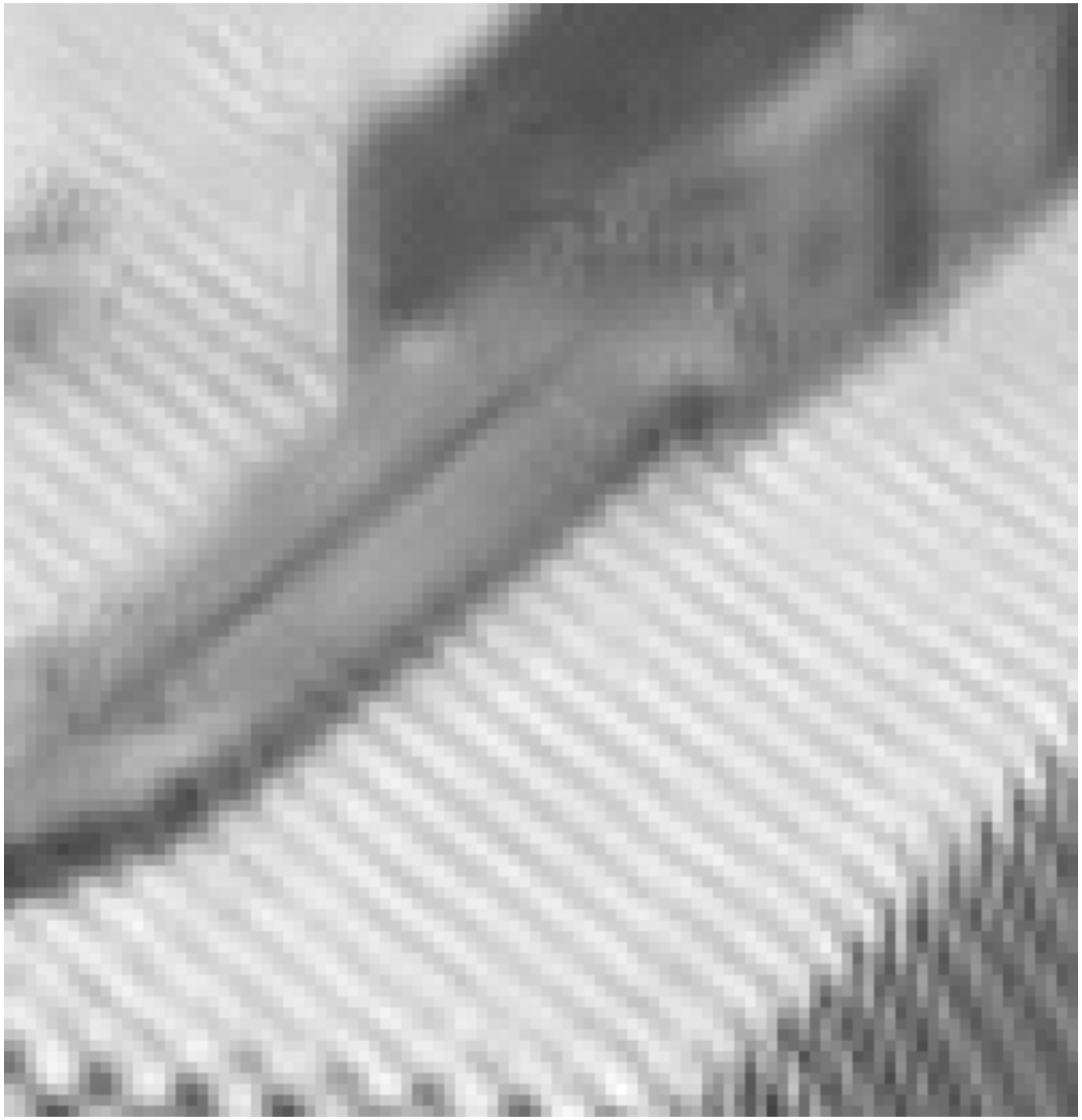
Denoising: Binary HMT



29.35 dB

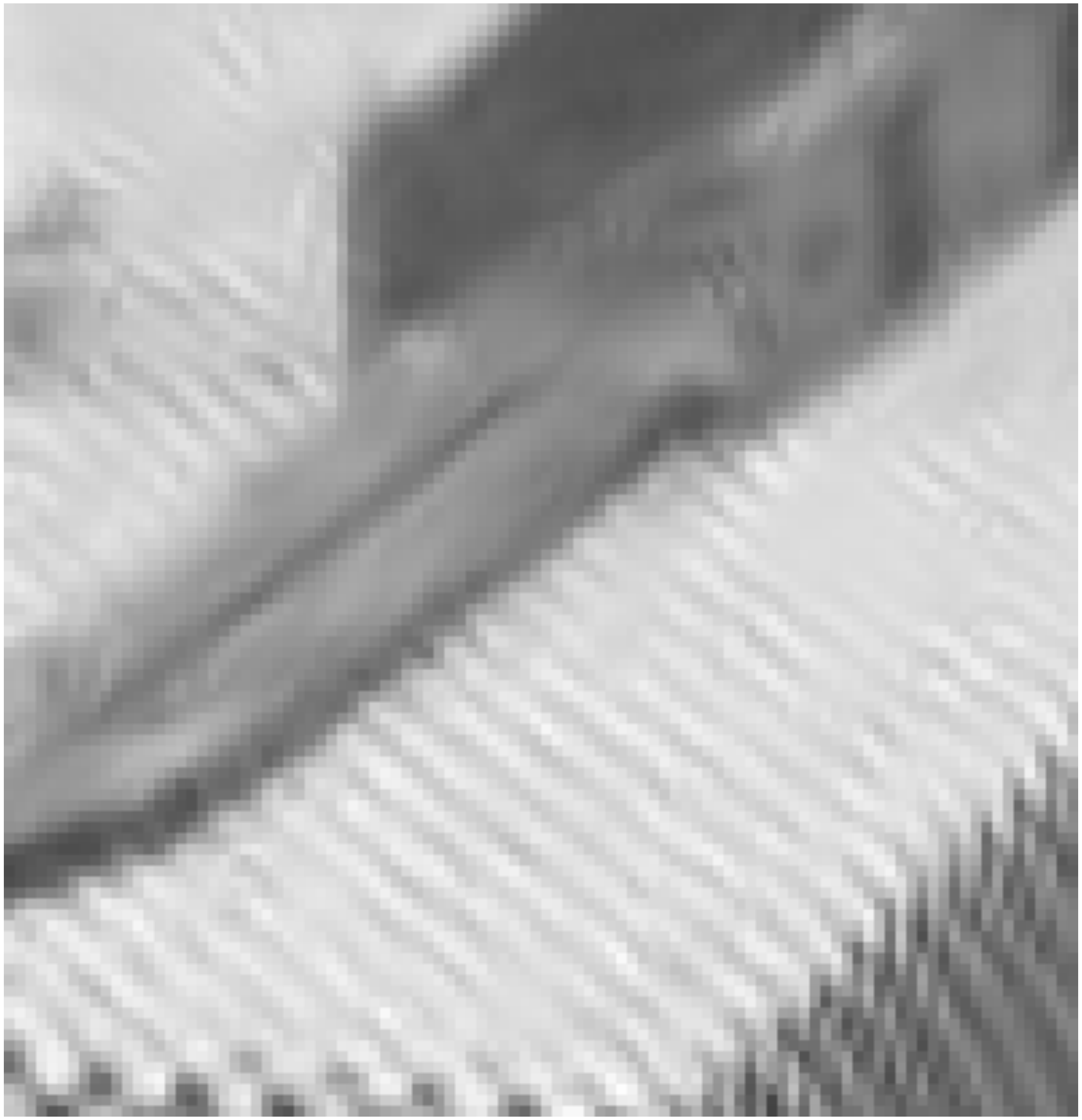
Crouse, Nowak, & Baraniuk, 1998

Denoising: HDP-HMT



32.10 dB

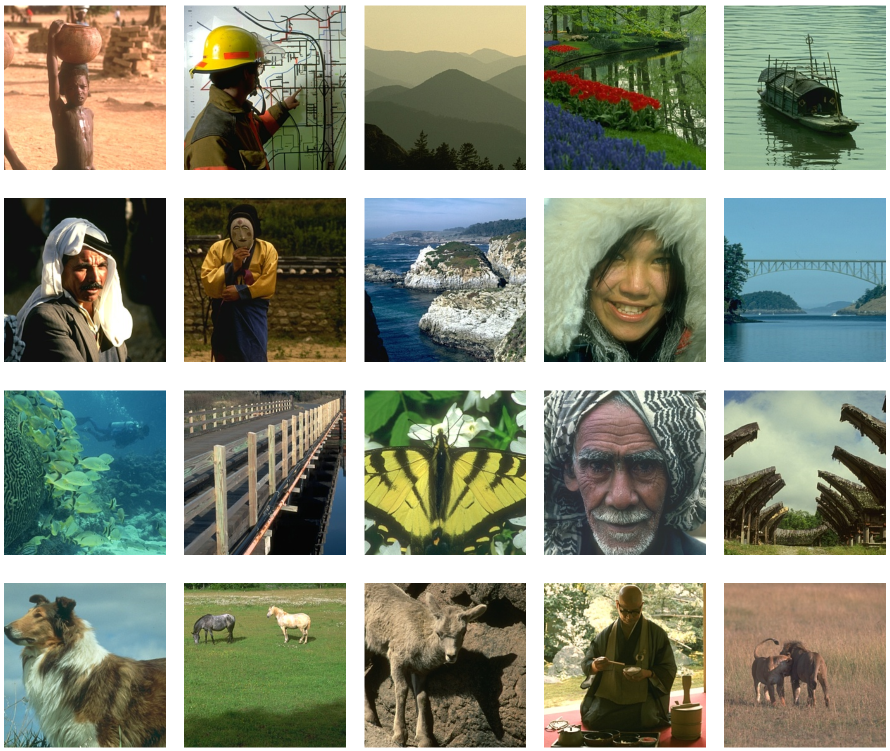
Denoising: Local GSM



31.84 dB

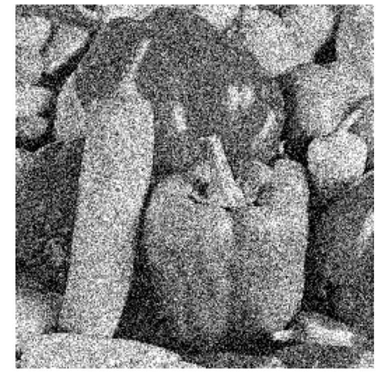
Portilla et. al., 2003

Estimating Clean Images

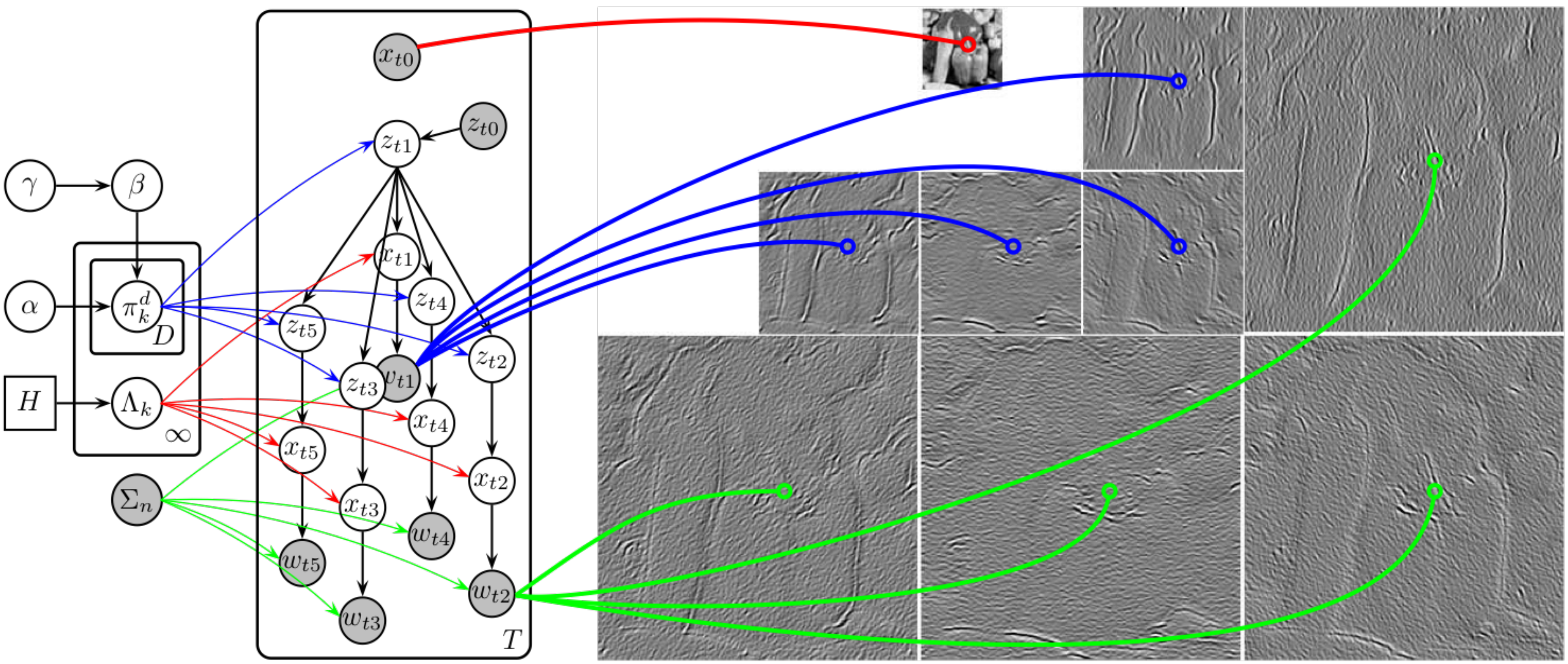


Empirical Bayesian approach estimates model parameters from the noisy image

Transfer denoising approach reuses multiscale hidden state patterns of clean images for making robust predictions



HDP-HMT for noisy data



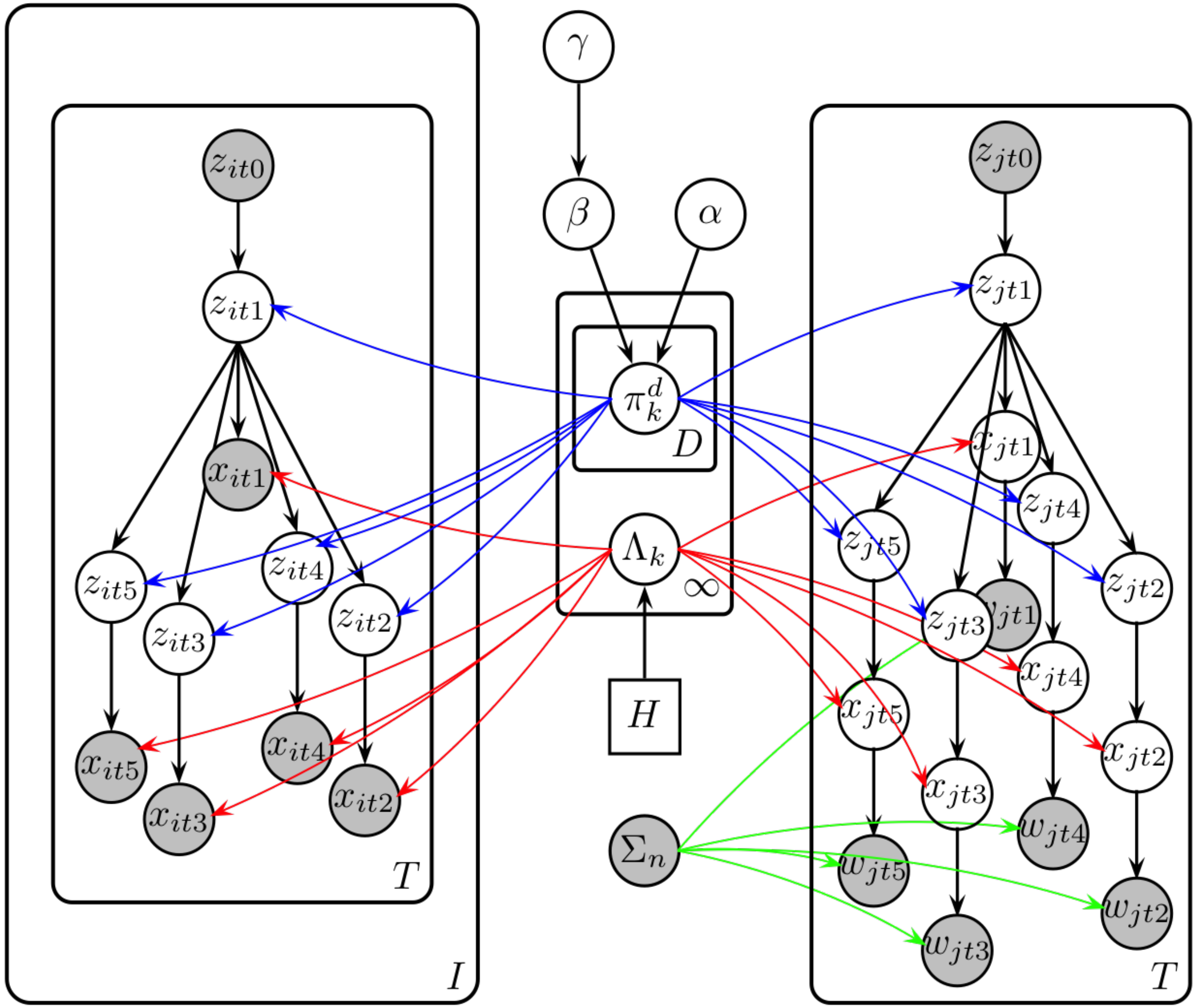
$x_{ti} \longrightarrow$ unobserved vector of *clean* wavelet coefficients

$\Sigma_n \longrightarrow$ noise variance

$w_{ti} \longrightarrow$ observed vector of *noisy* wavelet coefficients

$$w_{ti} \sim \mathcal{N}\left(x_{ti}, \Sigma_n\right)$$

… and for clean data as well



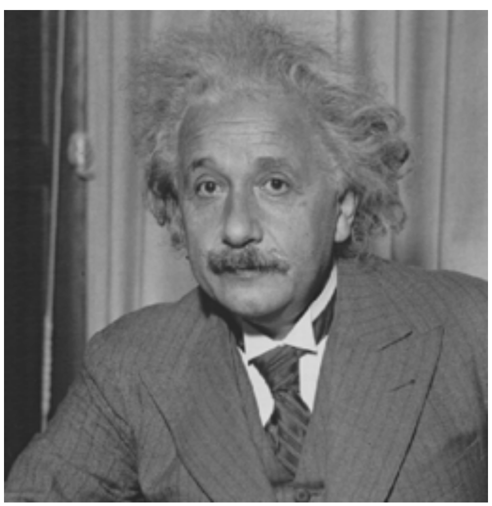
Denoising Einstein

Noisy
10.60 dB, 0.057

HDP-HMT
(Emp. Bayes)
25.64 dB, 0.564

HDP-HMT
(Transfer)
26.80 dB, 0.664

Original



BLS-GSM
26.38 dB, 0.647

BM3D
26.49 dB, 0.659

Natural Scene Denoising

Noisy
8.14 dB, 0.033

HDP-HMT
(Emp. Bayes)
24.24 dB, 0.519

HDP-HMT
(Transfer)
26.50 dB, 0.794

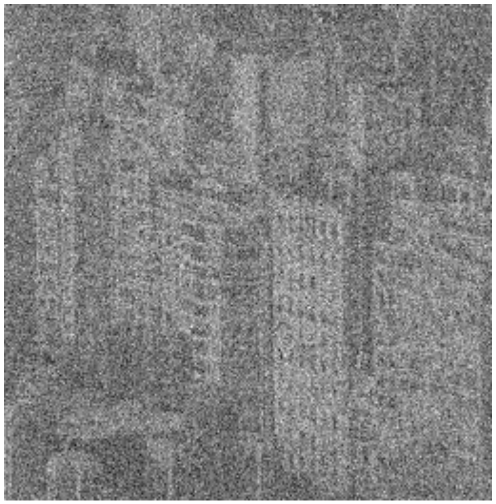
Original

BLS-GSM
25.59 dB, 0.726

BM3D
25.74 dB, 0.751

Natural Scene Denoising

Noisy
8.14 dB, 0.177



HDP-HMT
(Emp. Bayes)
18.55 dB, 0.484

HDP-HMT
(Transfer)
18.77 dB, 0.486

Original

BLS-GSM
18.59 dB, 0.454

BM3D
18.65 dB, 0.470

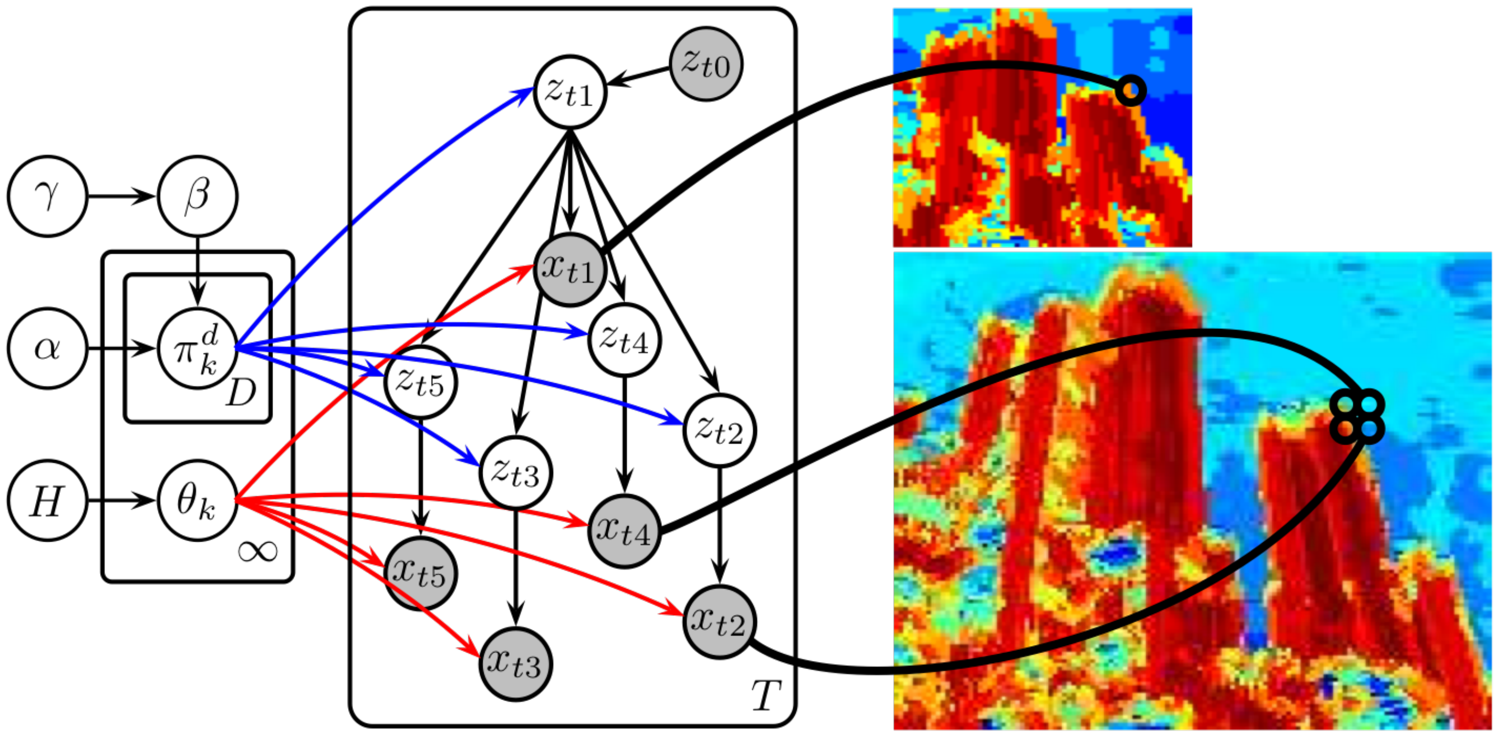
Natural Scene Categorization

Coast **Forest** **Open Country** **Street** **Tall Building**

Goals:

- Visually *recognize* natural scene categories
- Accurately model the statistics of *natural scene categories*

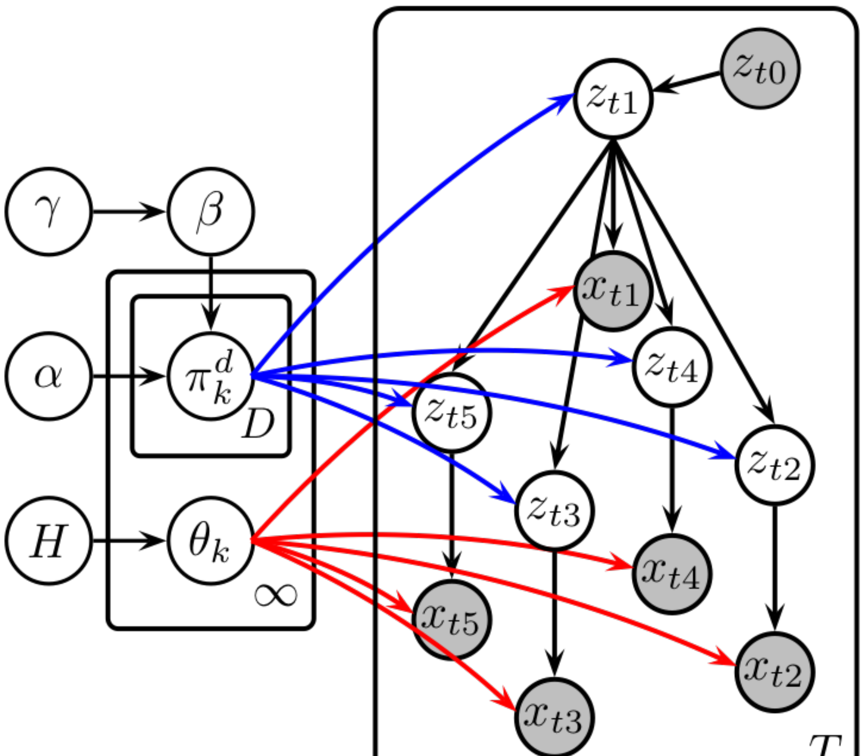
HDP-HMT Scene Model



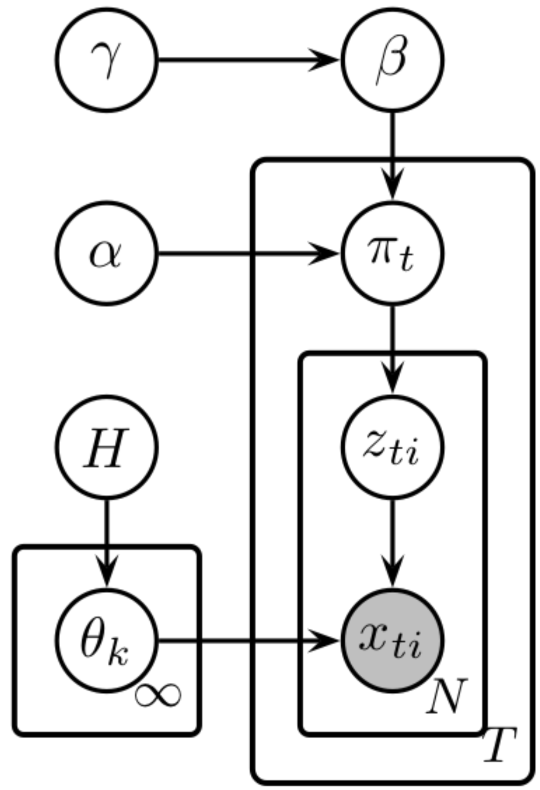
- Hidden states z_{ti} generate vectors of clean wavelet coefficients x_{ti} at multiple orientations, or dense multiscale SIFT descriptors

… versus baseline HDP-BOF

HDP-HMT



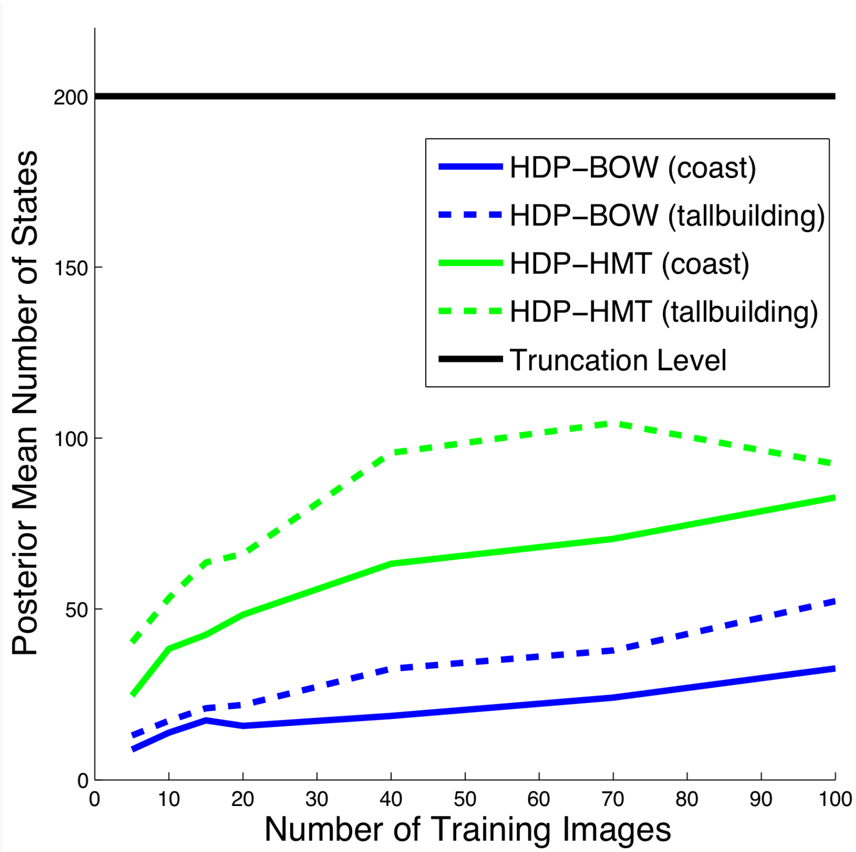
HDP-BOF



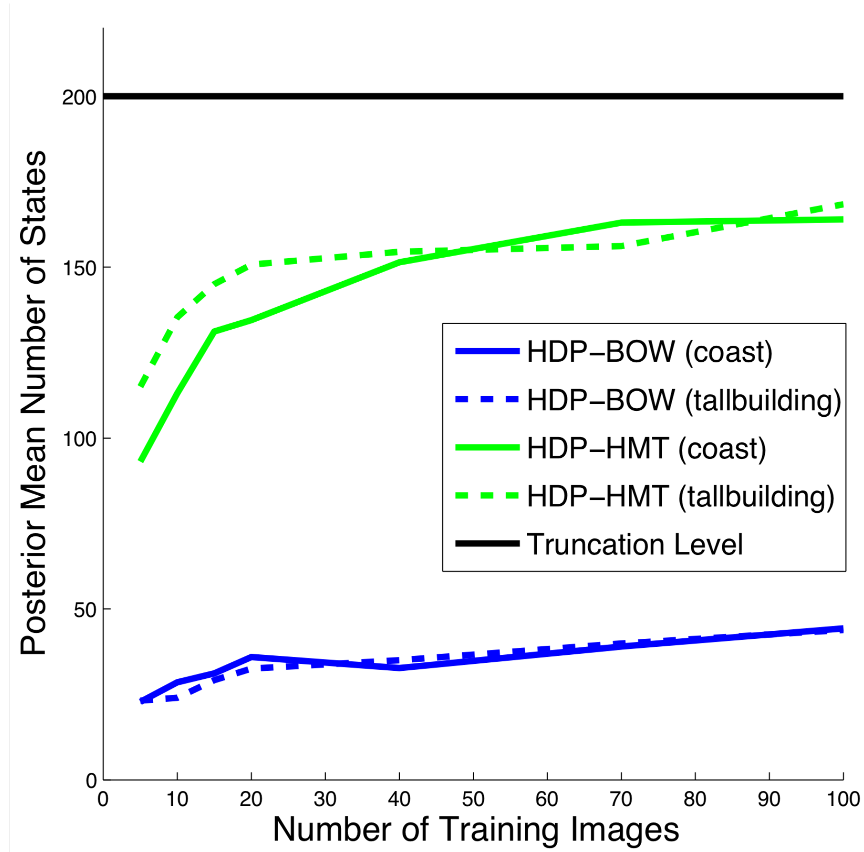
Nonparametric Bayesian extension of LDA scene models (Fei-Fei & Perona, 2005)
which ignore spatial locations of locally extracted image features

Number of States

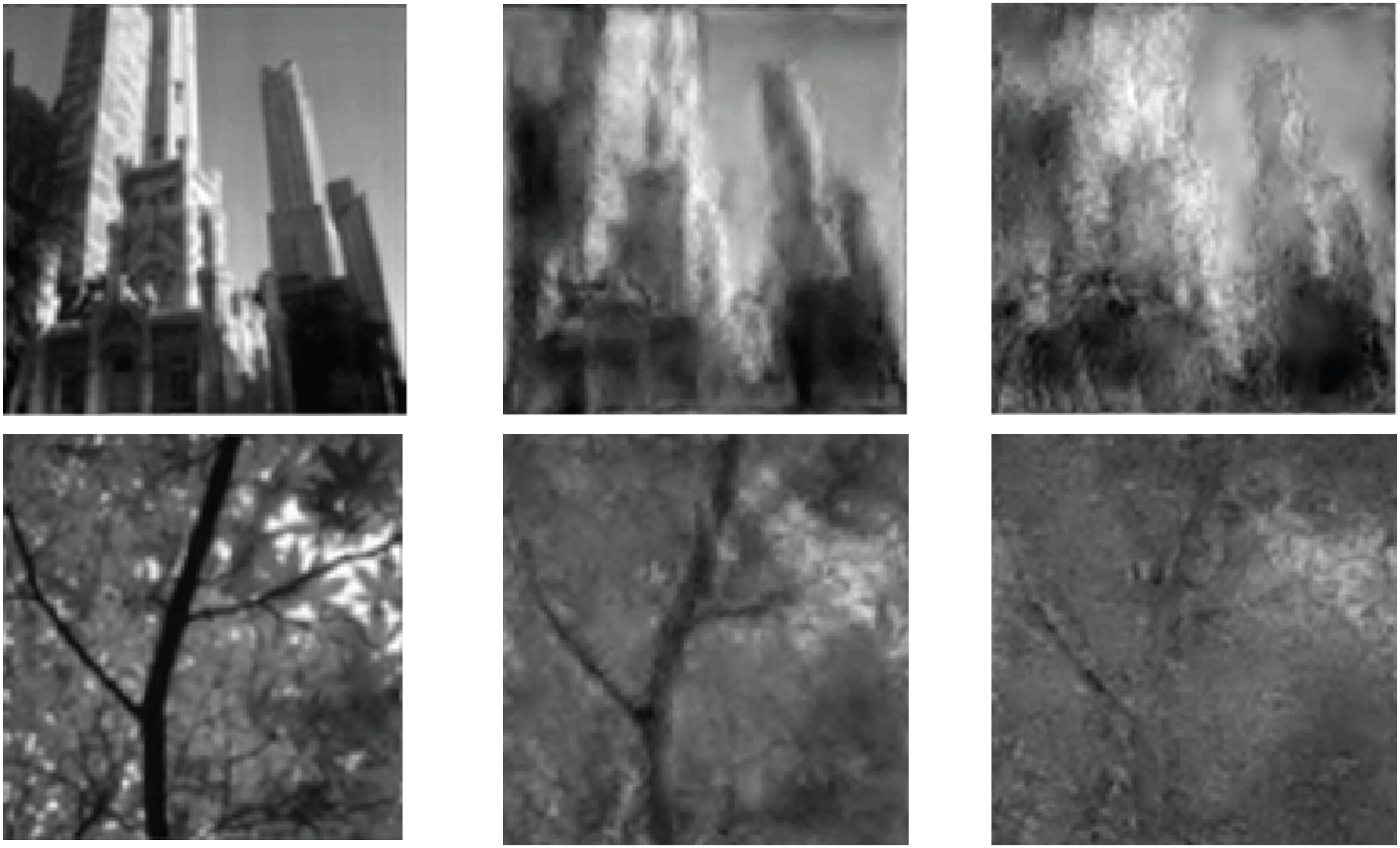
Wavelet (sp5)



SIFT



Samples given MAP states



Input Image

HDP Hidden Markov Tree

HDP Bag of Features

Categorizing Natural Scenes

Wavelet (sfp7)

coast	77.5	0.6	10.0	0.0	0.6	10.6	0.6	0.0
forest	0.0	91.4	0.0	0.0	5.5	0.8	2.3	0.0
highway	3.3	0.0	75.0	0.0	10.0	10.0	1.7	0.0
inside city	0.9	0.9	2.8	77.8	0.0	3.7	9.3	4.6
mountain	0.6	13.8	4.6	0.6	63.2	9.2	8.0	0.0
open country	8.6	10.0	3.3	0.5	11.0	61.9	4.8	0.0
street	0.0	1.1	5.4	2.2	7.6	0.0	81.5	2.2
tall building	0.0	0.0	2.6	13.5	0.6	0.6	8.3	74.4

HDP-BOF [75.3 %]

coast	90.6	0.6	4.4	0.0	1.9	2.5	0.0	0.0
forest	0.0	85.2	0.8	0.0	8.6	3.1	2.3	0.0
highway	8.3	0.0	80.0	0.0	6.7	0.0	3.3	1.7
inside city	1.9	0.9	5.6	75.0	0.9	0.9	10.2	4.6
mountain	1.7	1.1	2.9	0.0	91.4	1.1	1.7	0.0
open country	18.1	4.3	3.3	1.0	13.3	59.5	0.5	0.0
street	0.0	0.0	8.7	1.1	7.6	0.0	81.5	1.1
tall building	0.0	0.0	1.9	12.2	0.0	0.0	3.8	82.1

HDP-HMT [80.7 %]

SIFT

coast	90.0	0.6	1.2	0.0	1.9	6.2	0.0	0.0
forest	0.0	87.5	0.0	0.0	7.8	4.7	0.0	0.0
highway	6.7	0.0	80.0	1.7	1.7	5.0	5.0	0.0
inside city	0.0	0.0	1.9	87.0	0.0	0.0	9.3	1.9
mountain	1.1	0.6	0.6	0.0	90.2	5.7	0.6	1.1
open country	11.0	1.9	1.0	0.0	5.7	80.0	0.5	0.0
street	0.0	0.0	4.3	2.2	2.2	0.0	91.3	0.0
tall building	0.0	0.0	0.0	9.0	0.6	0.0	4.5	85.9

HDP-BOF [82.4 %]

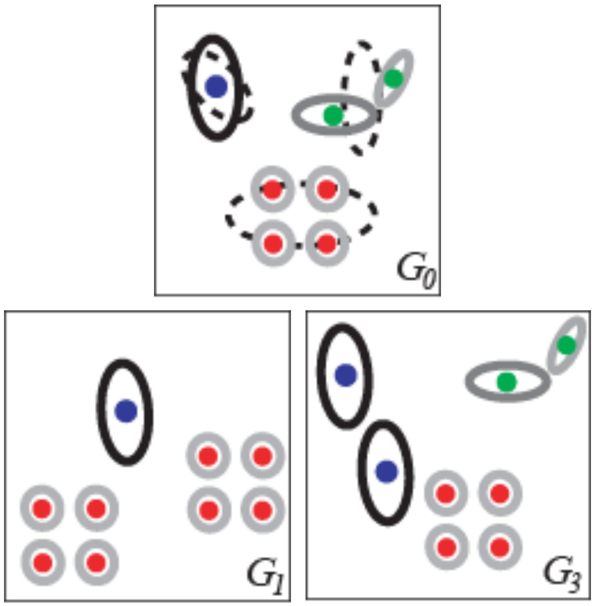
coast	86.2	1.2	4.4	0.0	0.6	7.5	0.0	0.0
forest	0.0	91.4	0.0	0.0	4.7	3.1	0.8	0.0
highway	6.7	0.0	75.0	1.7	3.3	6.7	6.7	0.0
inside city	0.0	0.9	3.7	82.4	0.0	0.9	10.2	1.9
mountain	0.6	4.0	3.4	0.0	81.0	8.0	2.3	0.6
open country	11.0	5.2	2.9	0.0	7.6	72.9	0.5	0.0
street	0.0	0.0	6.5	2.2	1.1	0.0	89.1	1.1
tall building	0.0	0.0	0.6	7.1	1.3	0.0	10.3	80.8

HDP-HMT [86.5 %]

Conclusions

Why move beyond topic models?

- Even with huge datasets, parametric (and nonparametric) models are constrained by their parameterizations
- Geometry and spatial relationships are more than entries in a feature vector



Lots to be done…

- Other geometric relationships: context, occlusion, composition, …
- Efficient, robust inference algorithms
- How should we balance design and learning of transferred representations?