Applied Bayesian Nonparametrics
2. Hierarchical Models

Tutorial at CVPR 2012

Erik Sudderth
Brown University

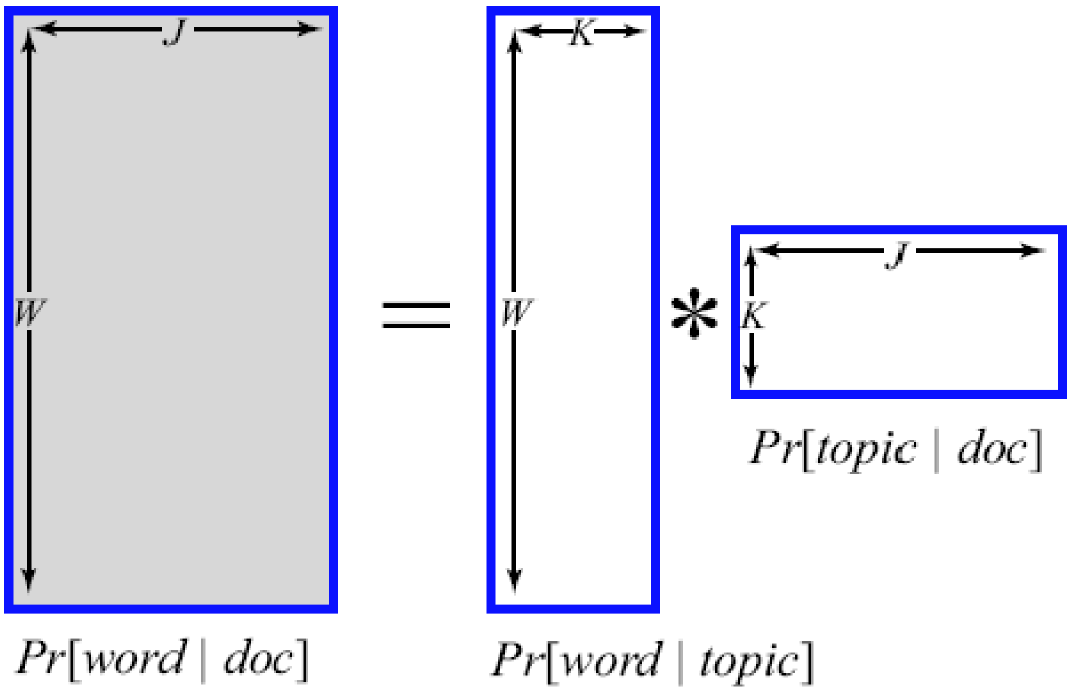
Work by E. Sudderth, A. Torralba, W. Freeman, & A. Willsky
IJCV 2008: *Describing Visual Scenes using Transformed Objects & Parts*
CVPR 2006: *Depth from Familiar Objects: A Hierarchical Model for 3D Scenes*
NIPS 2005: *Describing Visual Scenes using Transformed Dirichlet Processes*
Building on work by Y. W. Teh, M. Jordan, M. Beal, & D. Blei
JASA 2006: *Hierarchical Dirichlet Processes*

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
…

$W \approx 10,000$



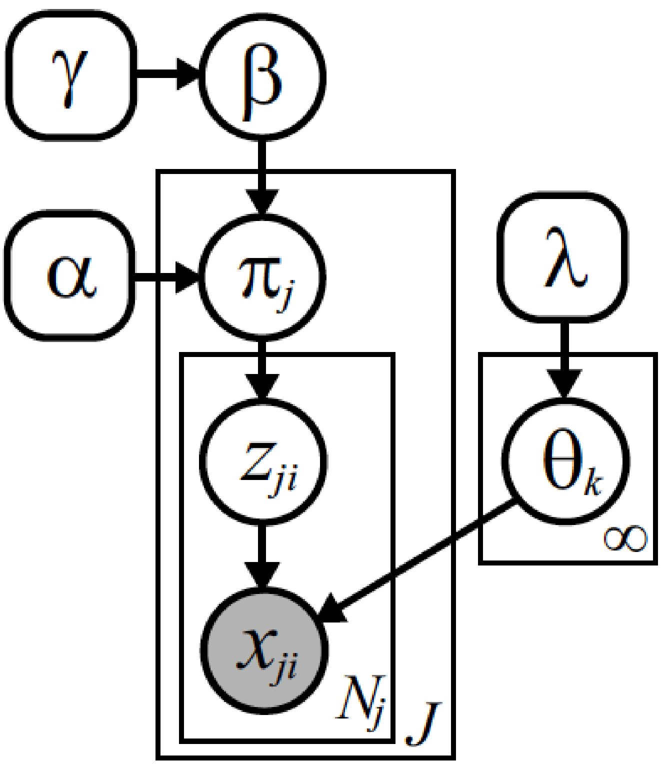
Algorithms
Neuroscience
Statistics
Vision
…

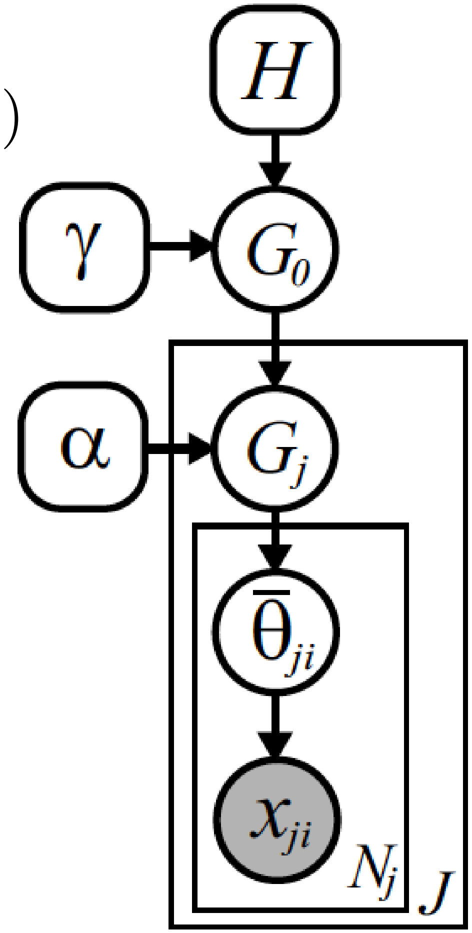
$K \approx 100$

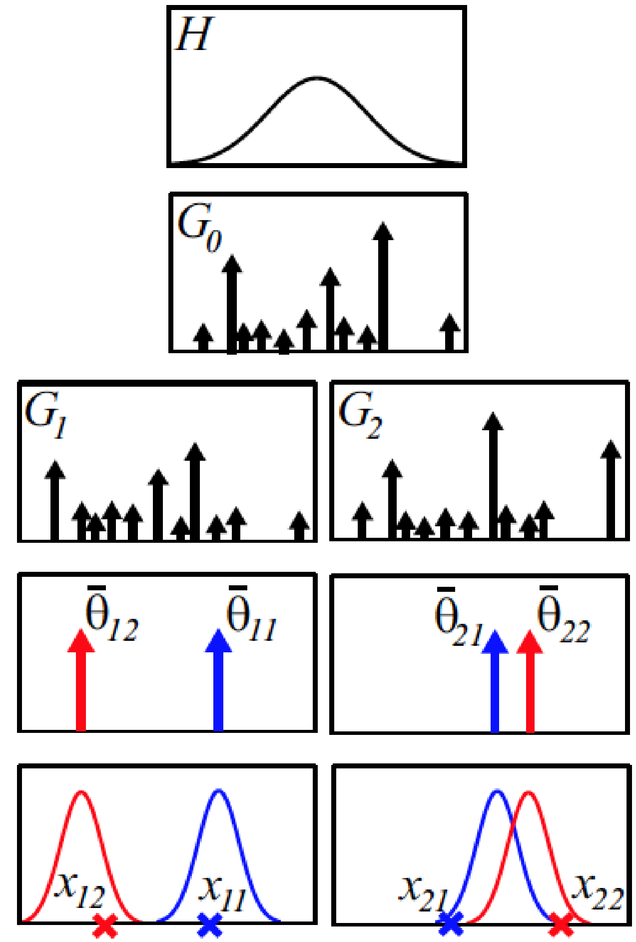
- **pLSA**: Probabilistic Latent Semantic Analysis *(Hofmann 2001)*
- **LDA**: Latent Dirichlet Allocation *(Blei, Ng, & Jordan 2003)*
- **HDP**: Hierarchical Dirichlet Processes *(Teh, Jordan, Beal, & Blei 2006)*

Hierarchical Dirichlet Process

$$G_0 \sim \mathrm{DP}(\gamma, H)$$
$$G_j \sim \mathrm{DP}(\alpha, G_0)$$







$$G_0(\theta) = \sum_{k=1}^{\infty} \beta_k \delta(\theta, \theta_k)$$

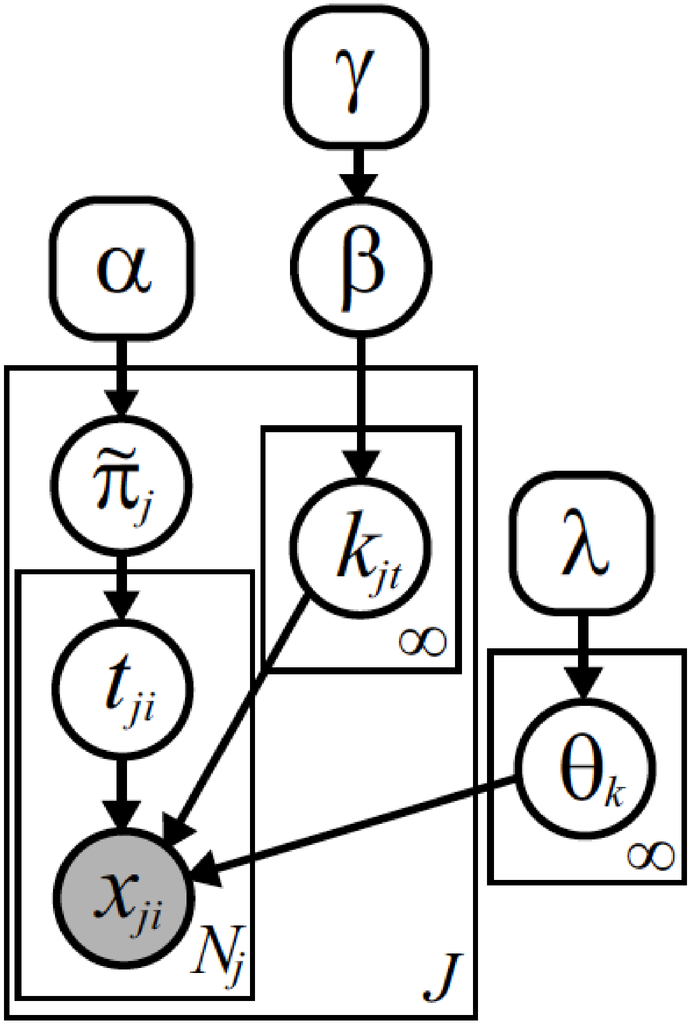
$$G_j(\theta) = \sum_{k=1}^{\infty} \pi_{jk} \delta(\theta, \theta_k)$$

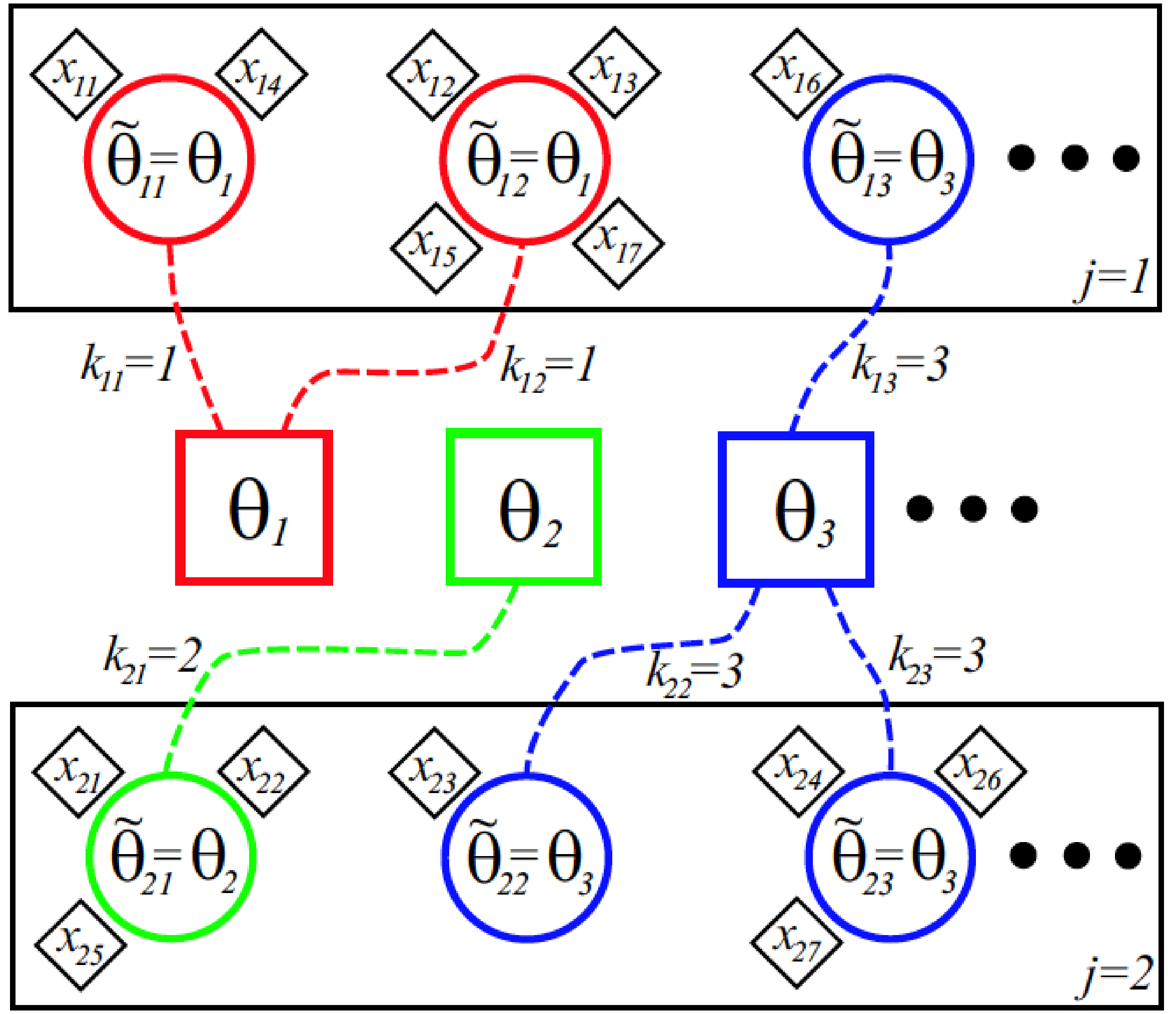
$$\boldsymbol{\beta} \sim \mathrm{GEM}(\gamma)$$
$$\theta_k \sim H(\lambda) \qquad k = 1, 2, \ldots$$

$$\mathbb{E}[\pi_j] = \beta$$

J groups of data: documents, images, …

Chinese Restaurant Franchise



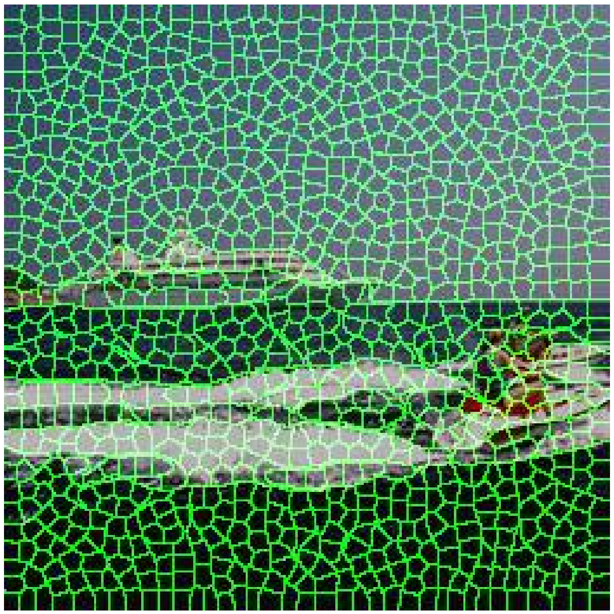


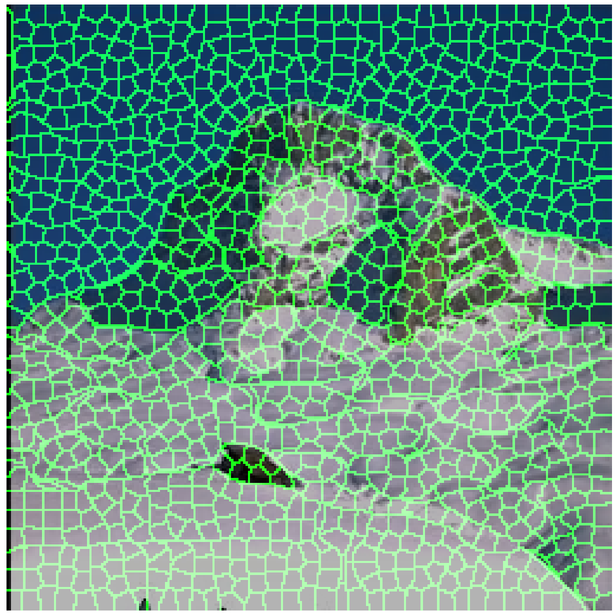
$$p(t_{ji} \mid t_{j1}, \ldots, t_{ji-1}, \alpha) \propto \sum_t N_{jt} \delta(t_{ji}, t) + \alpha \delta(t_{ji}, \bar{t})$$

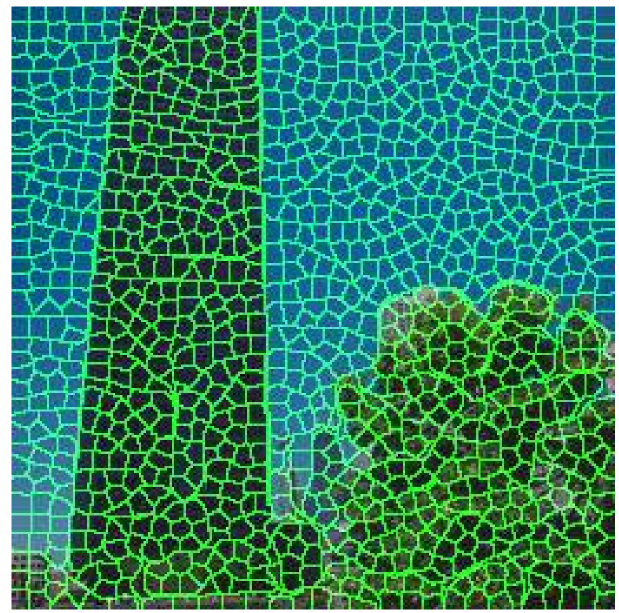
$$p(k_{jt} \mid \mathbf{k}_1, \ldots, \mathbf{k}_{j-1}, k_{j1}, \ldots, k_{jt-1}, \gamma) \propto \sum_k M_k \delta(k_{jt}, k) + \gamma \delta(k_{jt}, \bar{k})$$

Local Visual Features: Superpixels

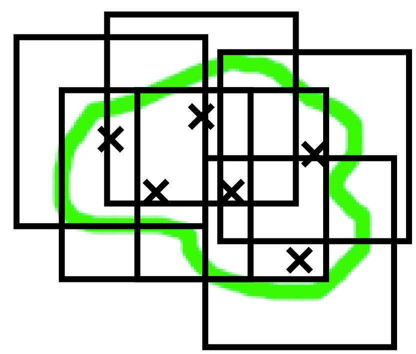
Inspired by the successes of *topic models* for text data, some have proposed learning from *local image features*



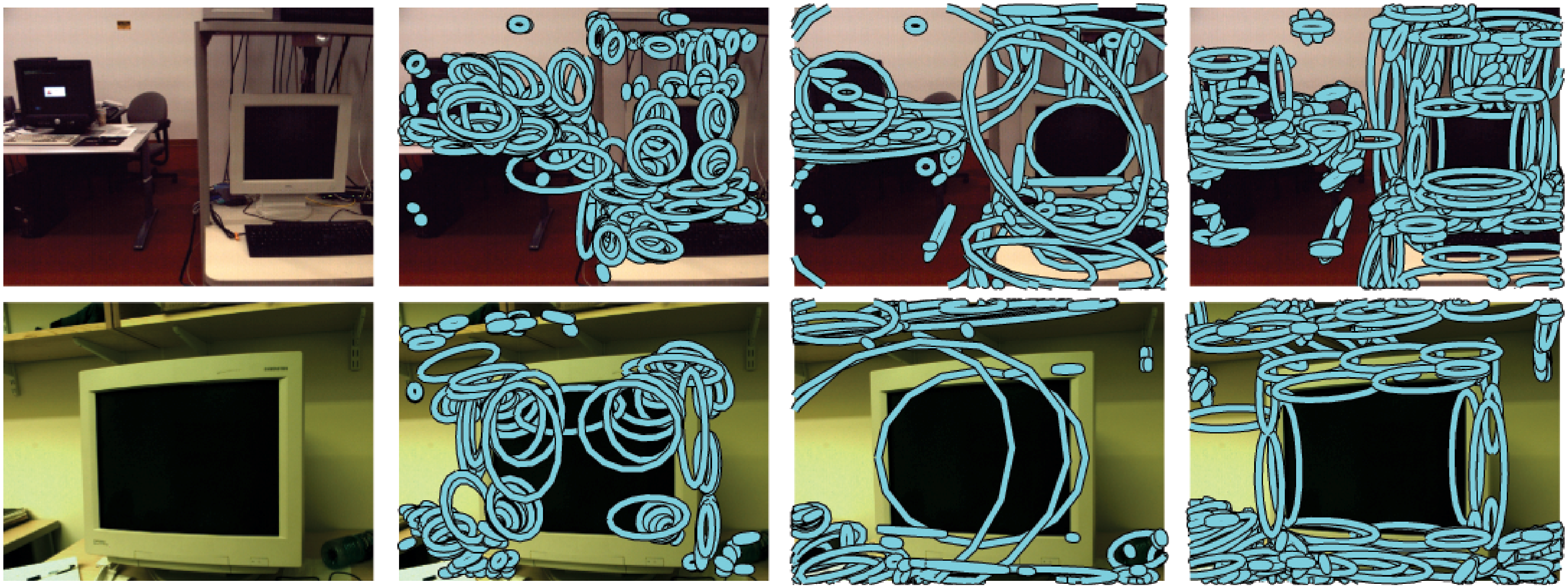




- Partition image into ~1,000 *superpixels*
- Goal: Reduce dimensionality, aggregate information spatially – *hopefully not across object boundaries!*



Local Visual Features: Interest Regions



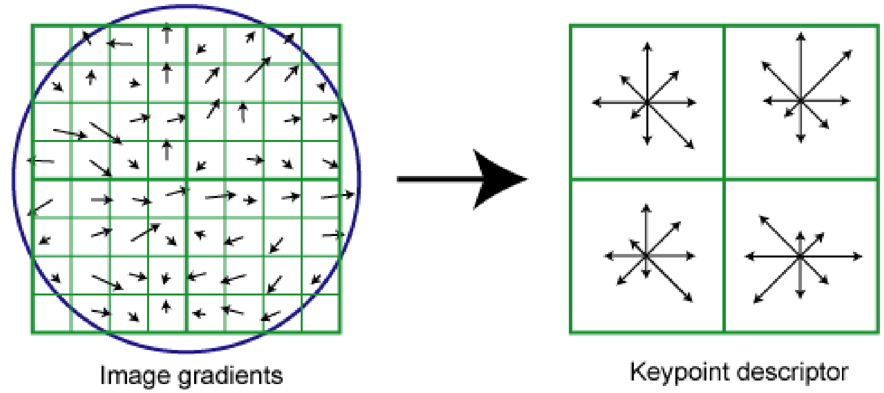
Affinely Adapted Harris Corners | Maximally Stable Extremal Regions | Linked Sequences of Canny Edges

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

A Discrete Feature Vocabulary

SIFT Descriptors

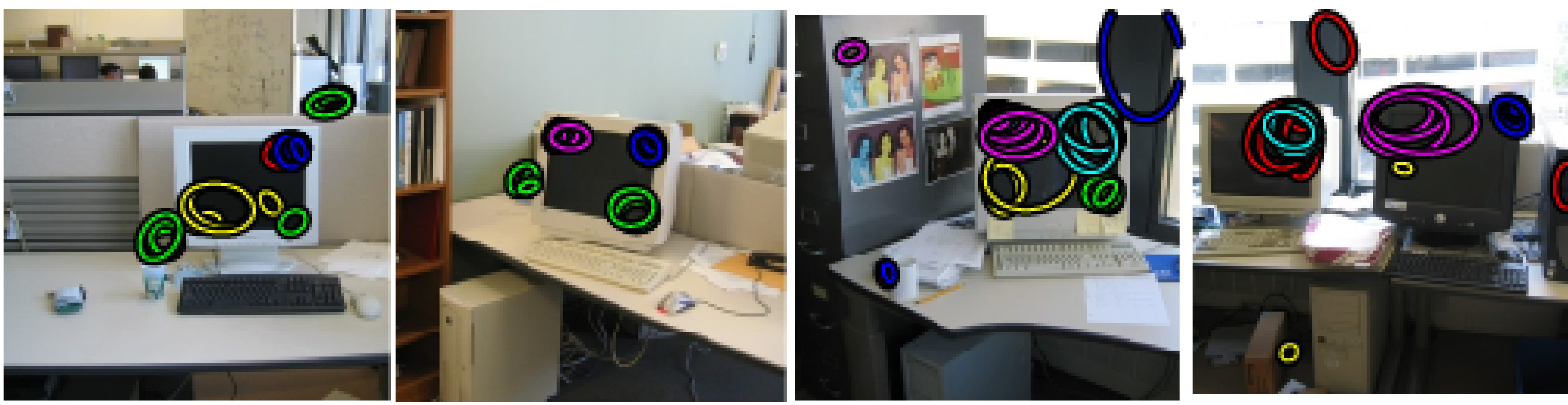
- Normalized histograms of orientation energy
- Compute ~1,000 word dictionary via K-means
- Map each feature to nearest *visual word*



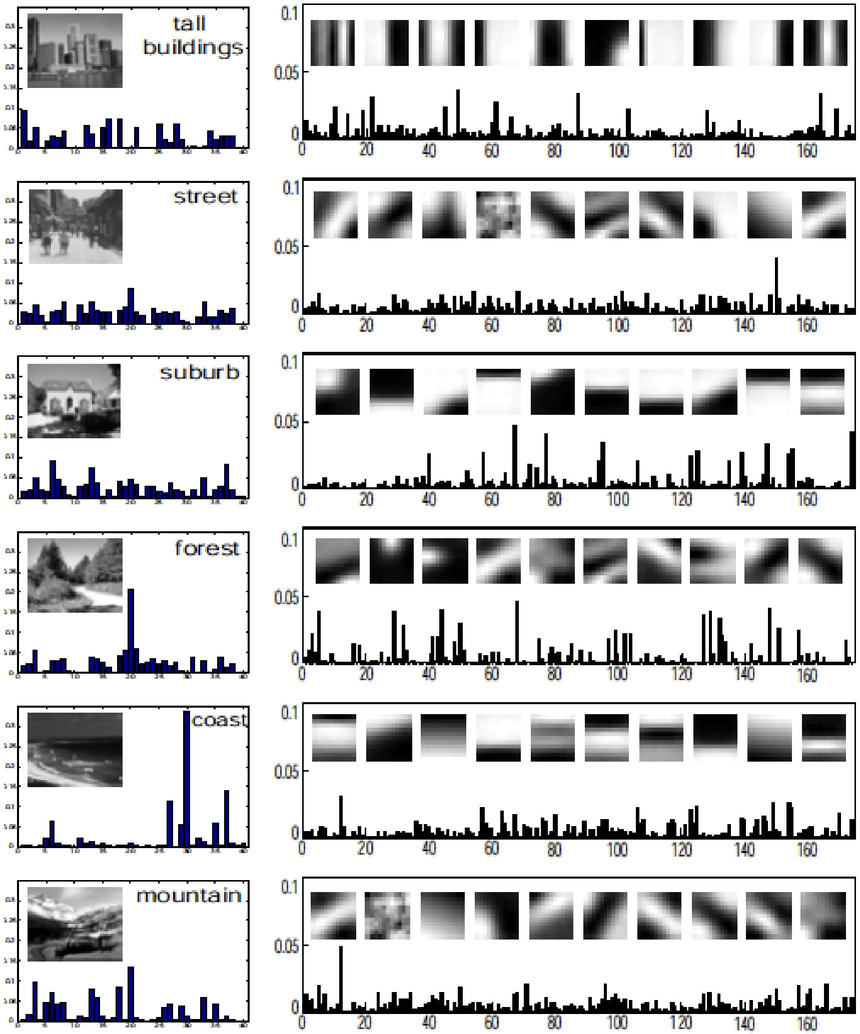
Lowe, IJCV 2004

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

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

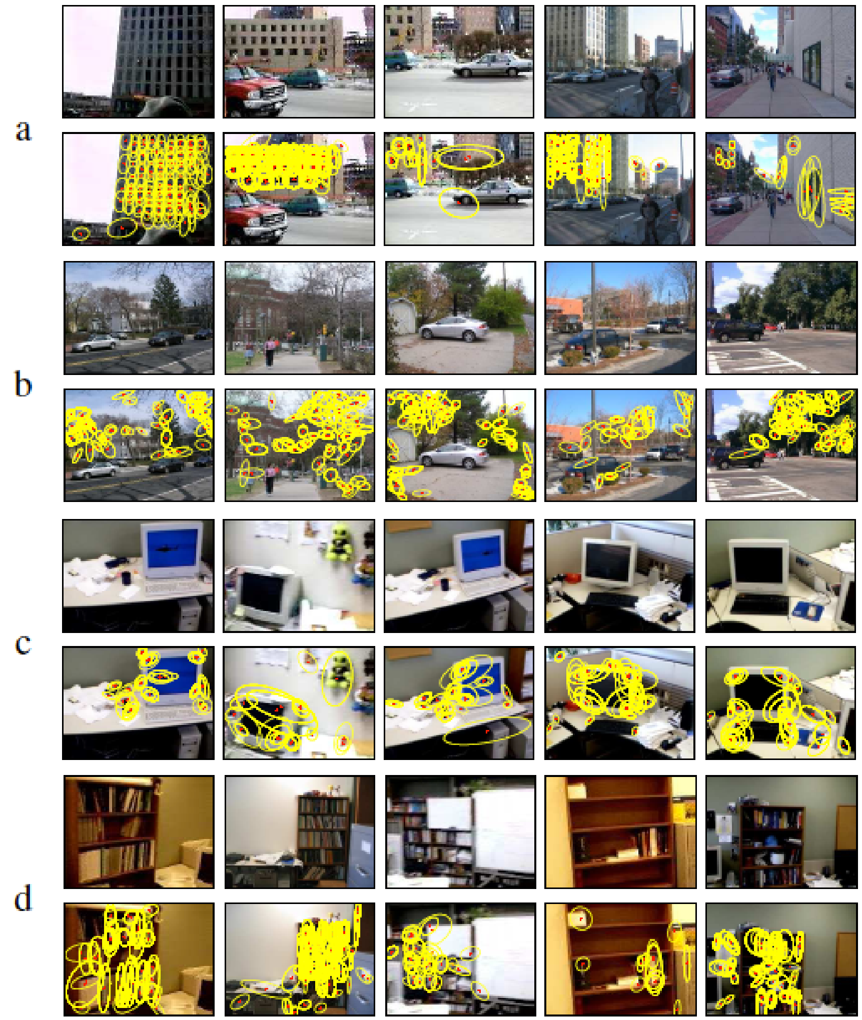


The World as a Bag of Visual Words



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

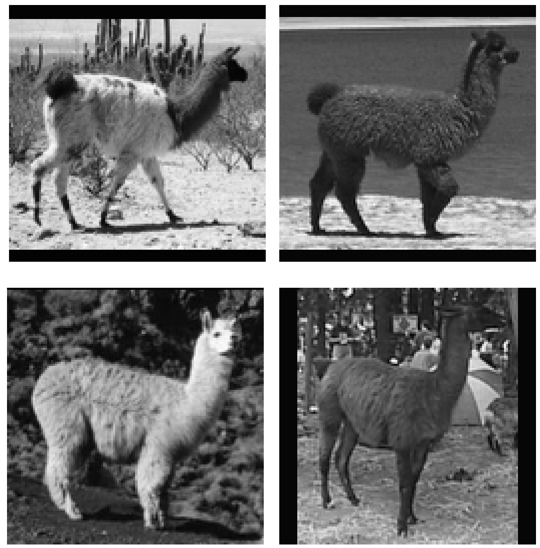
Images as more than Bags of Features

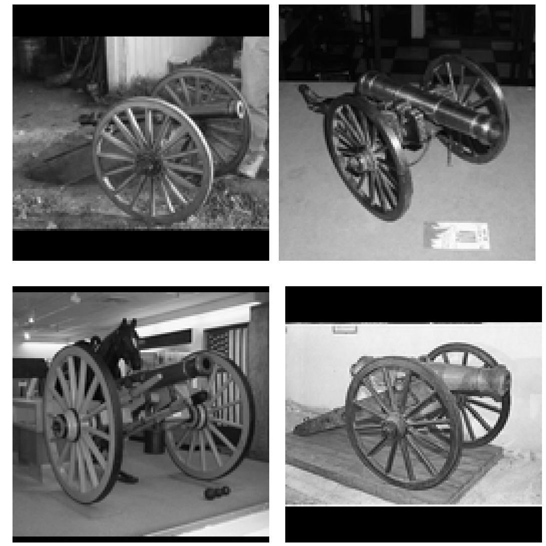
- How do I know this is ocean beneath a clear sky?
- How many bicycles and tricycles am I looking at?

Why are we trying to squeeze images into topic models?

There are many more tools available by adapting nonparametric and hierarchical Bayesian models.

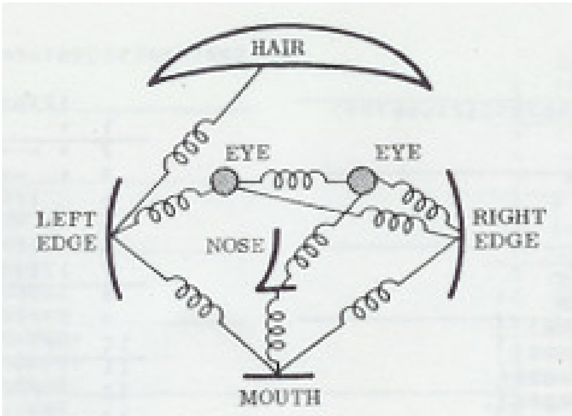
Visual Object Categorization



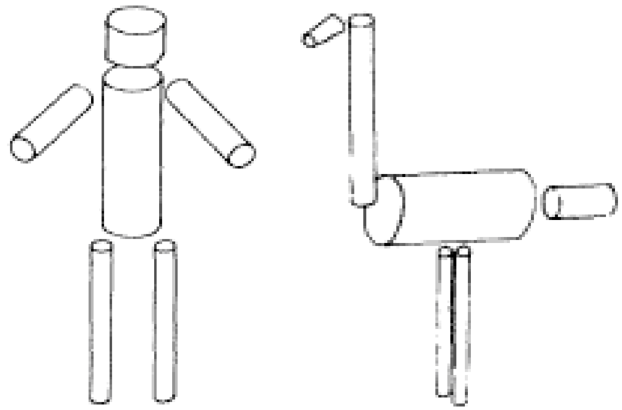


- **GOAL:** Visually *recognize* and *localize* object categories
- Robustly *learn* appearance models from few examples

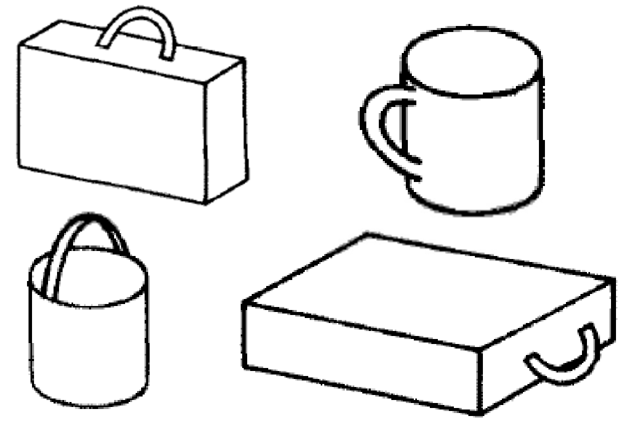
Part-Based Models for Objects



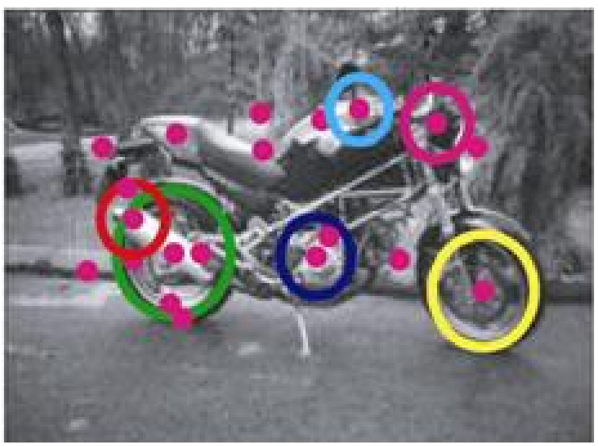
Pictorial Structures
Fischler & Elschlager, 1973



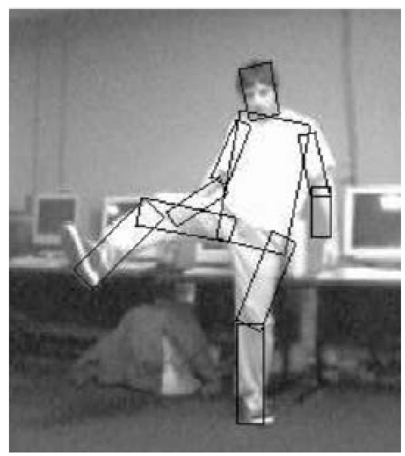
Generalized Cylinders
Marr & Nishihara, 1978



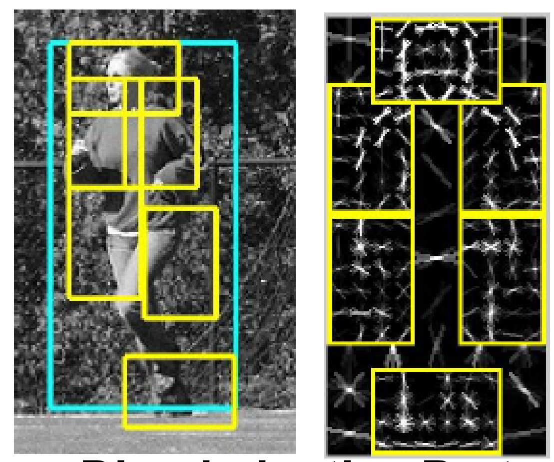
Recognition by Components
Biederman, 1987



Constellation Model
Perona, Weber, Welling, Fergus, Fei-Fei, 2000 to …

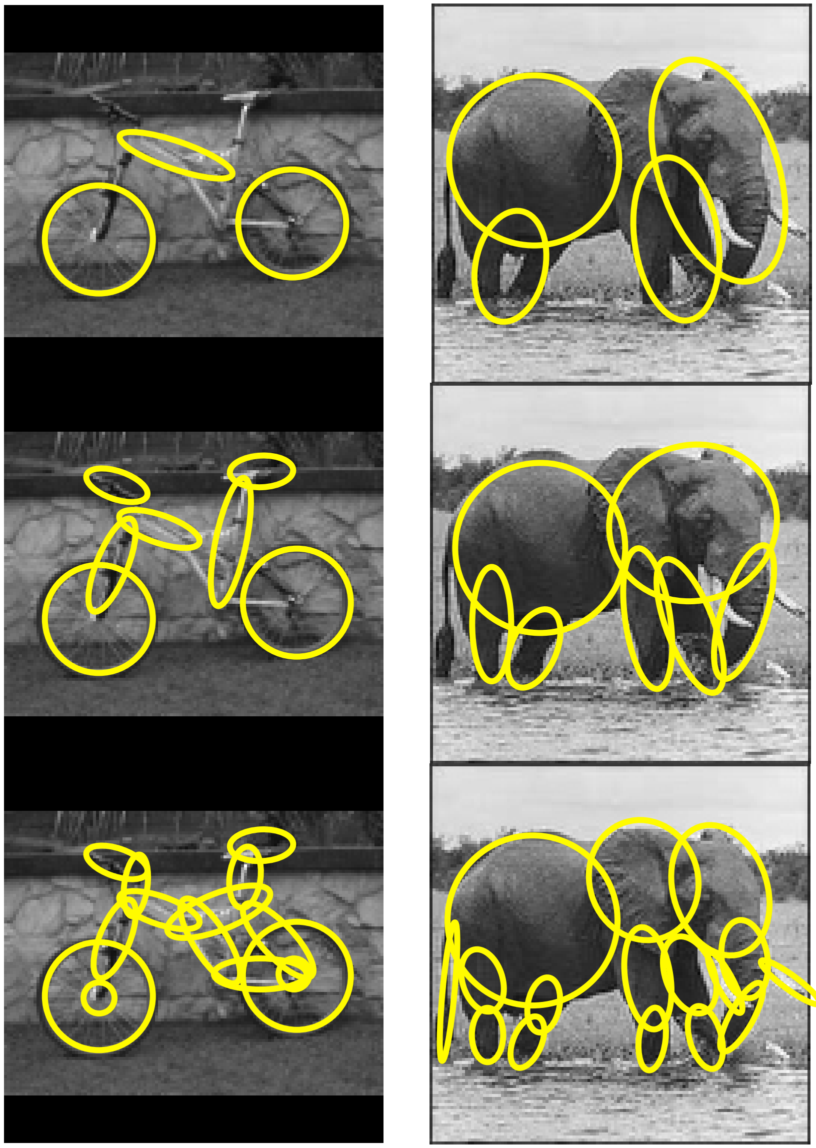


Efficient Matching
Felzenszwalb & Huttenlocher, 2005



Discriminative Parts
Felzenszwalb, McAllester, Ramanan, 2008 to …

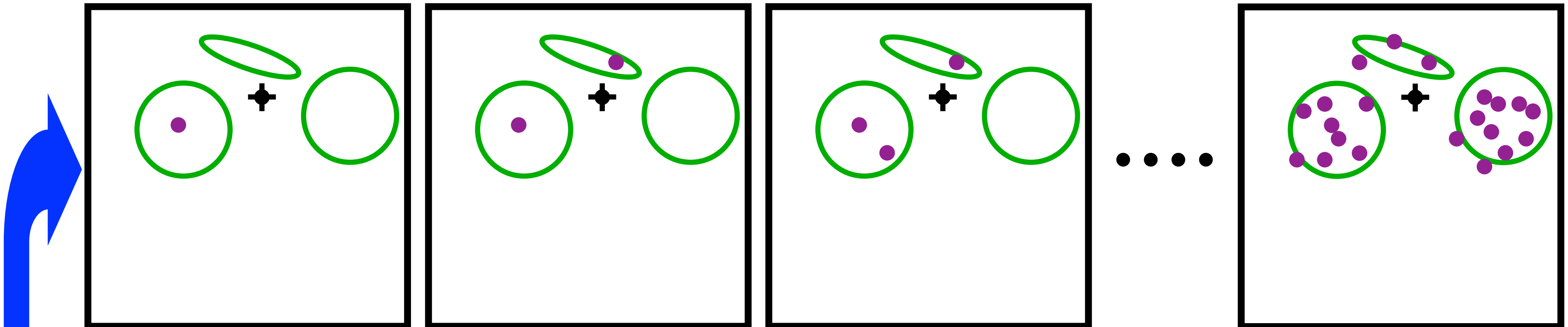
Counting Objects & Parts

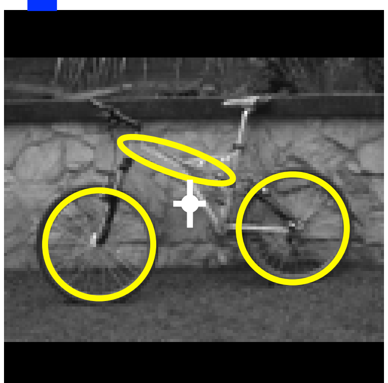


How many parts?

How many objects?

Generative Model for Objects

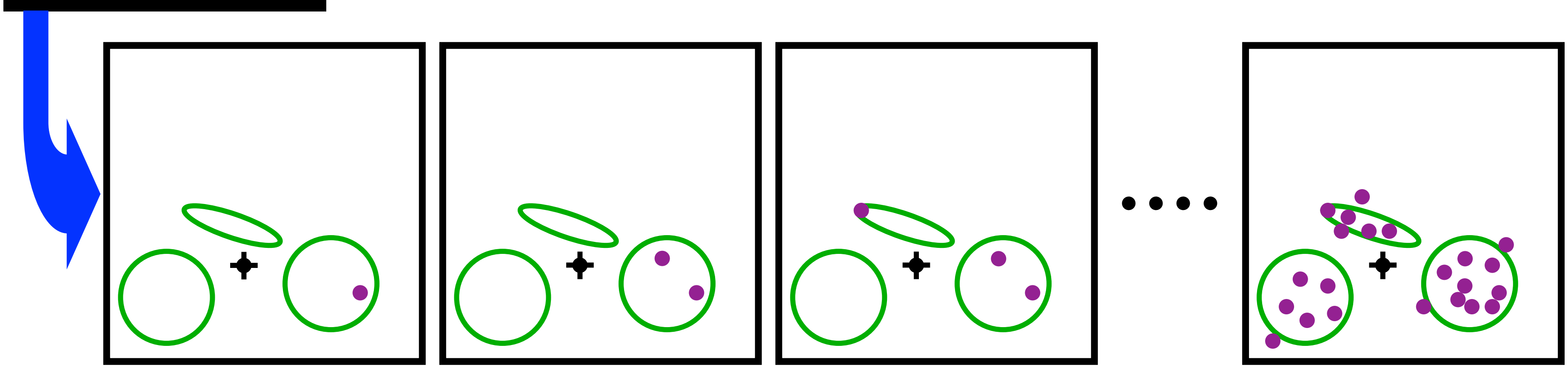




For each image: Sample a reference position

For each feature:

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



Objects as Distributions

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

Feature appearance

Feature position

Pr(part)

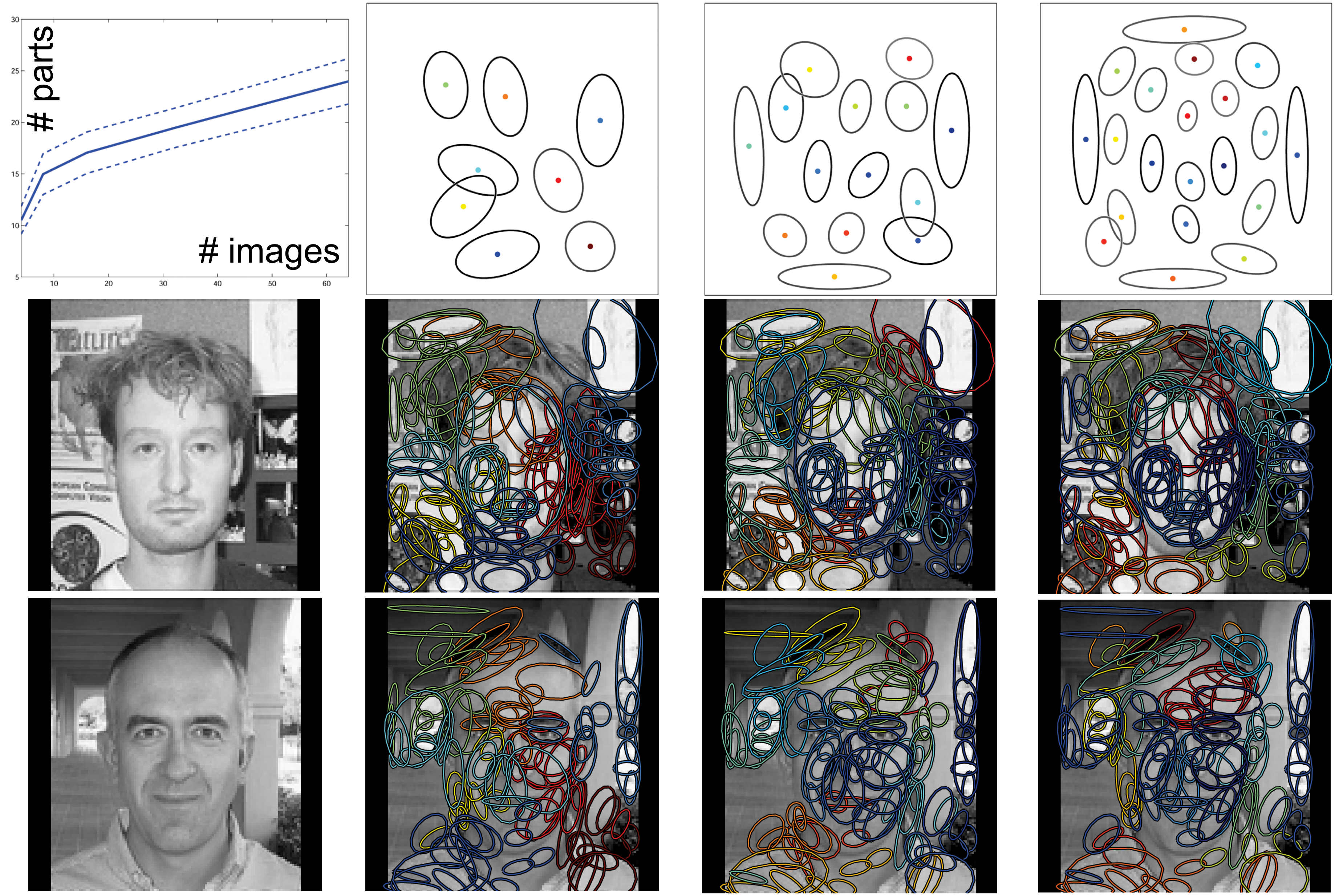
- Parts are defined by *parameters*, which encode distributions on visual features:

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

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

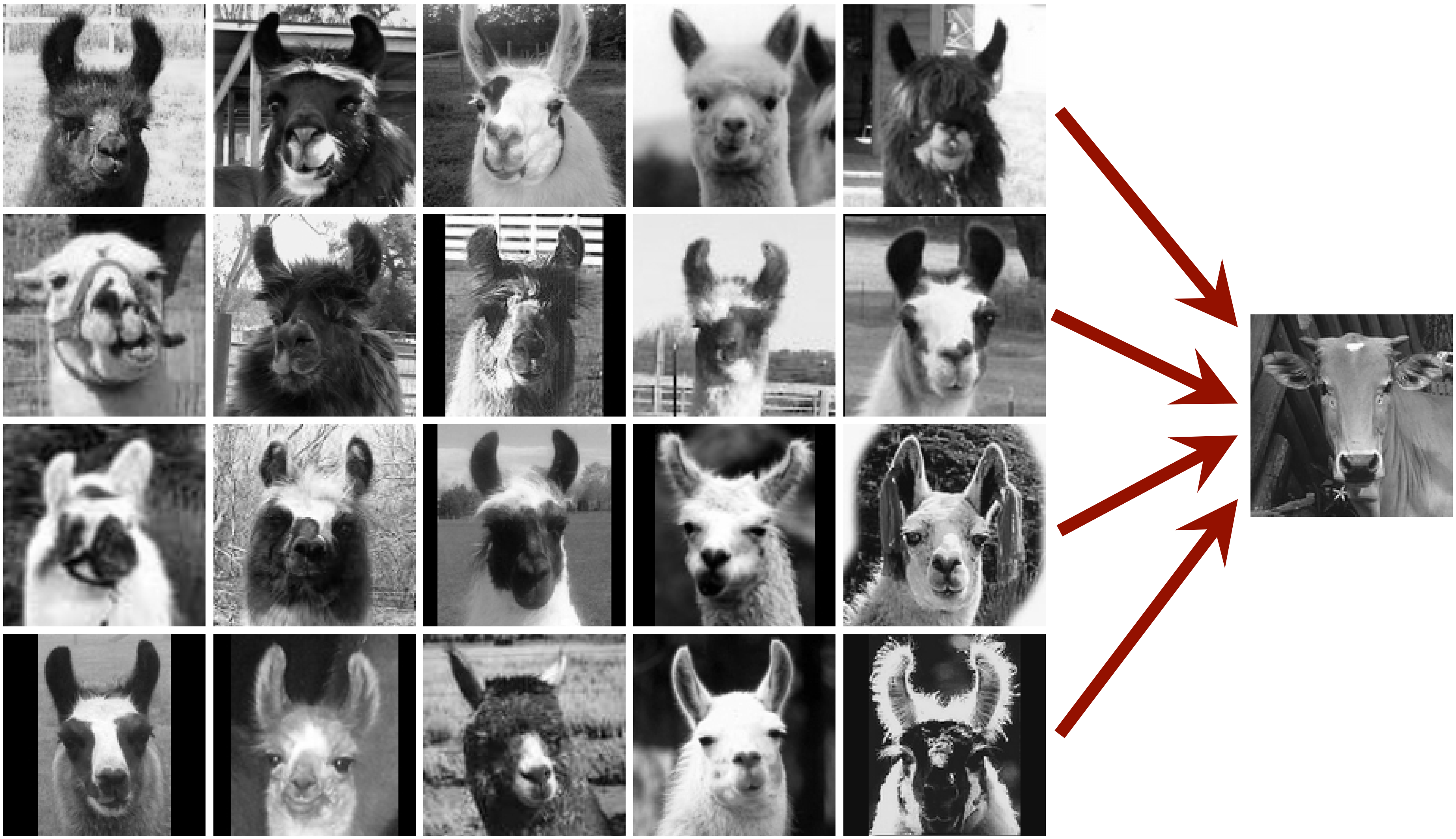
$$G(\theta) = \sum_{k=1}^{\infty} \pi_k \delta(\theta, \theta_k) \qquad \boldsymbol{\pi} \sim \text{Stick}(\alpha)$$

A Nonparametric Part-Based Model



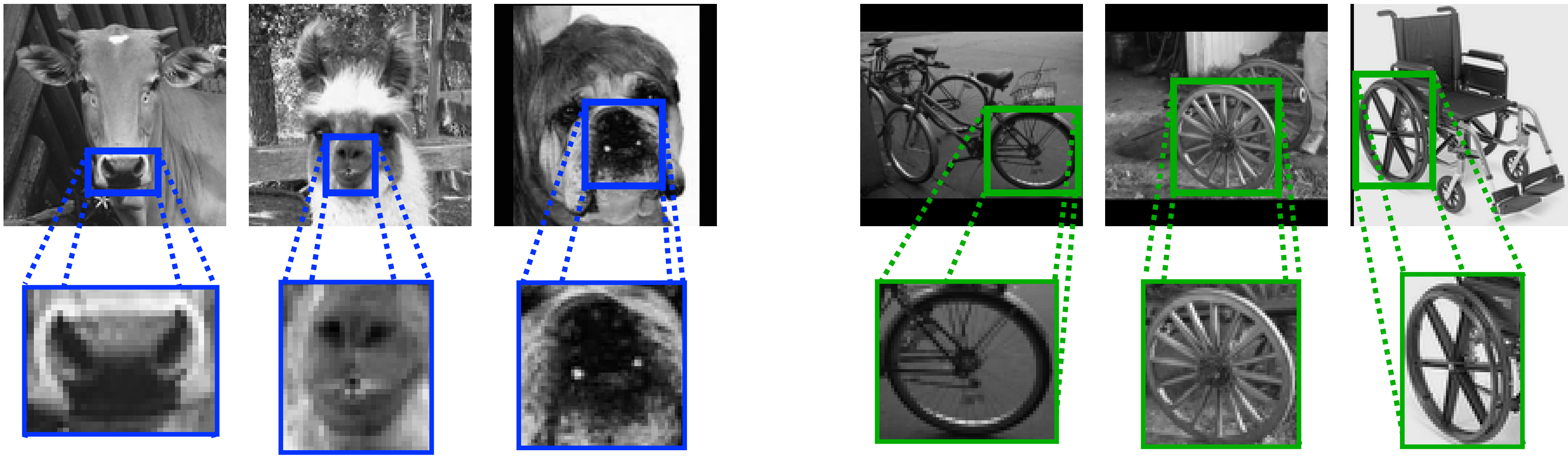
4 Images **16 Images** **64 Images**

Generalizing Across Categories



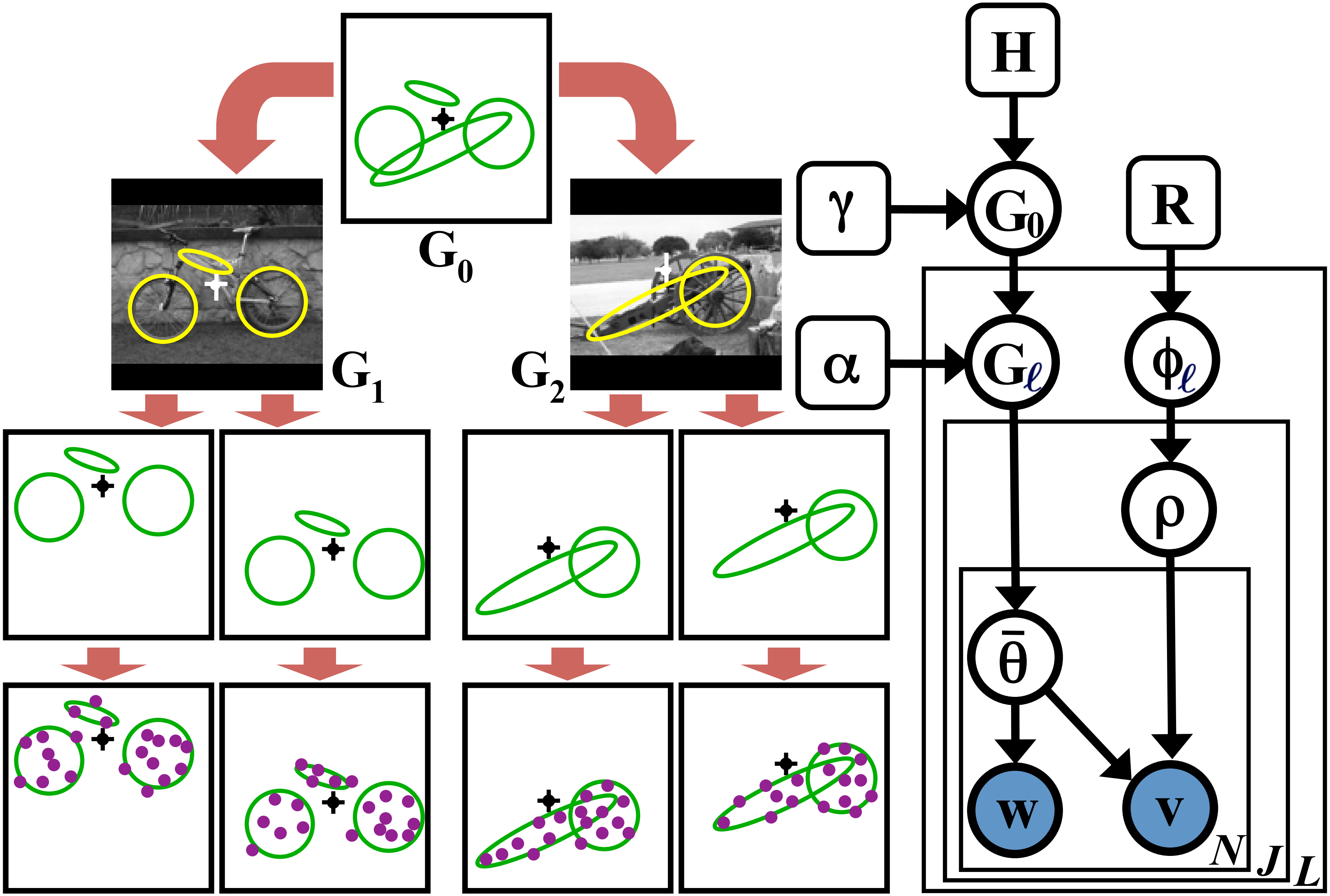
Can we transfer knowledge from one object category to another?

Learning Shared Parts

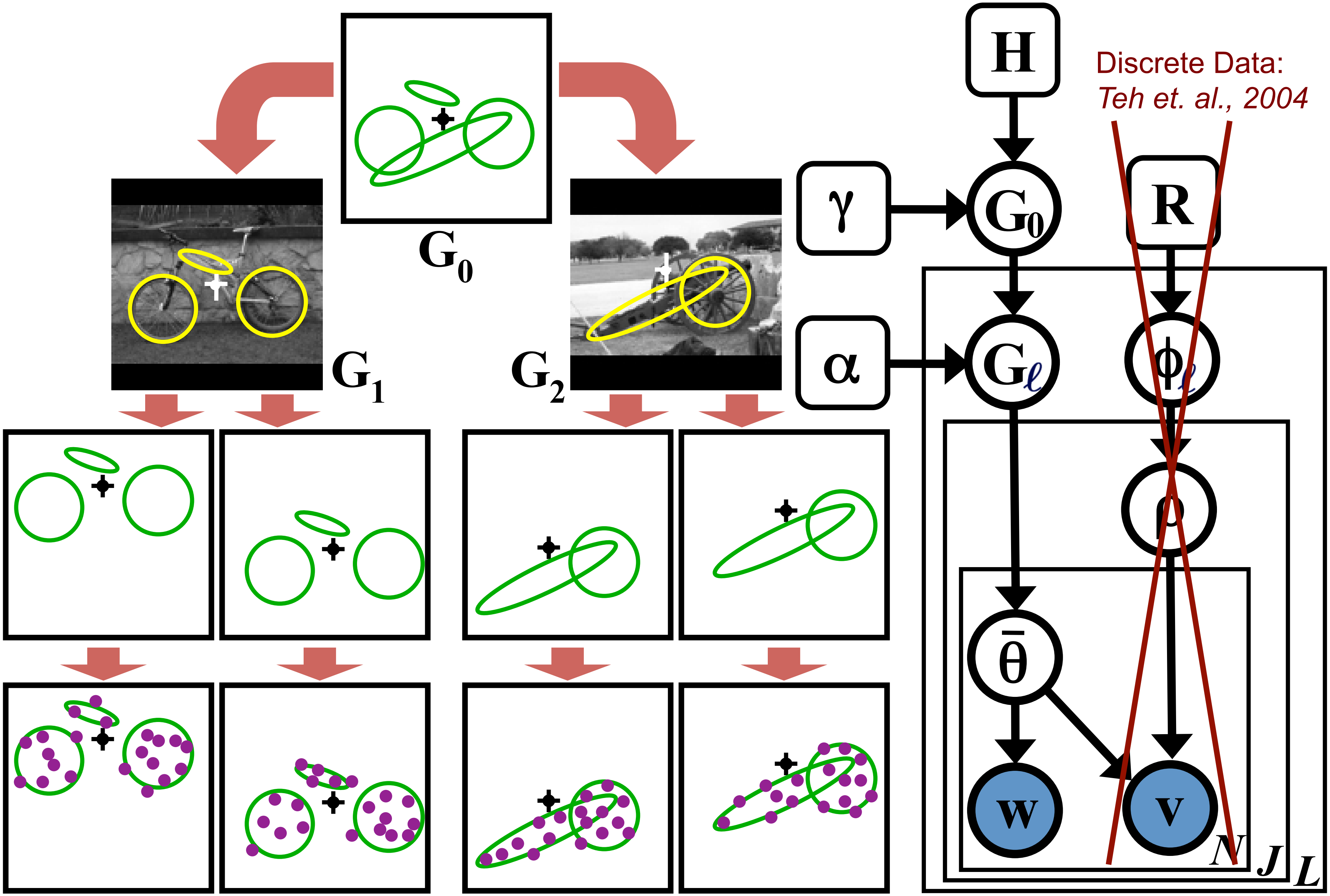


- Objects are often locally similar in appearance
- Discover *parts* shared across categories
  - How many total parts should we share?
  - How many parts should each category use?

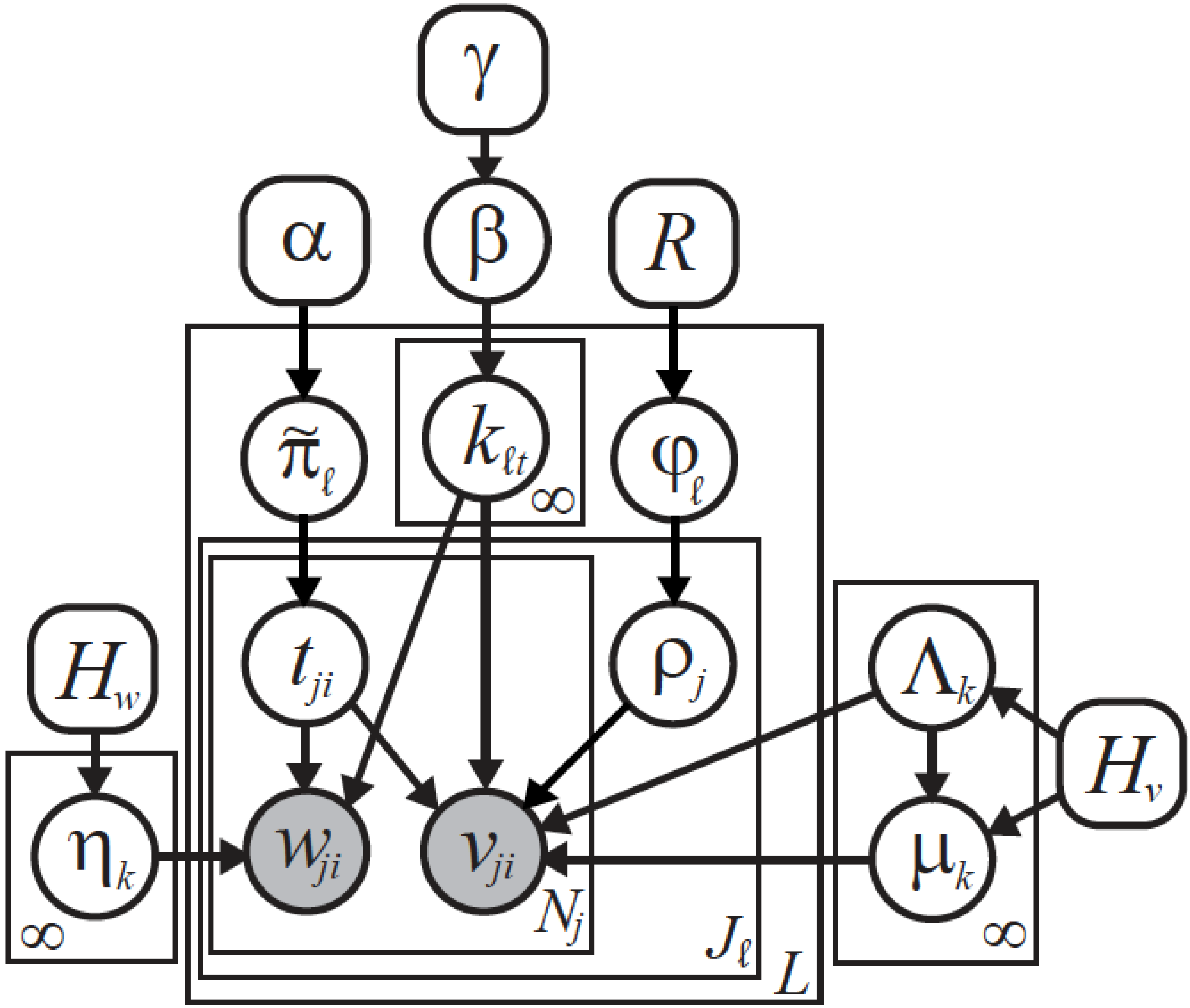
Hierarchical DP Object Model



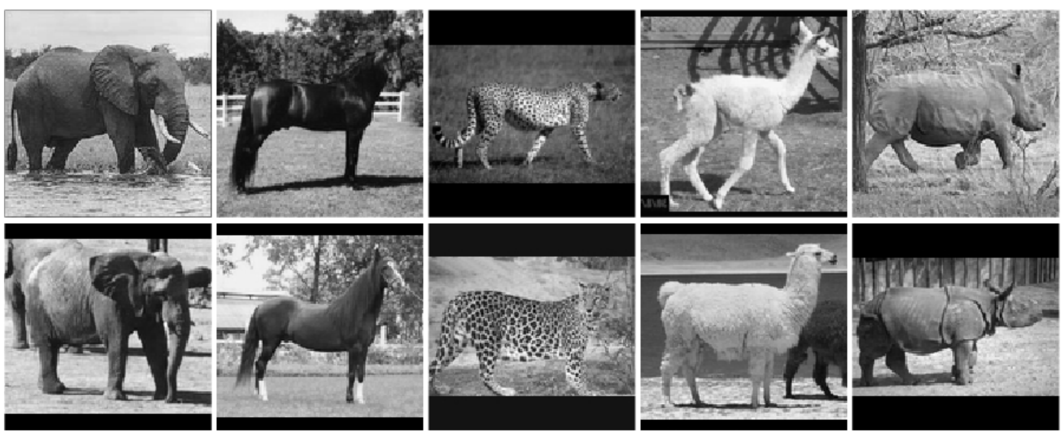
Hierarchical DP Object Model

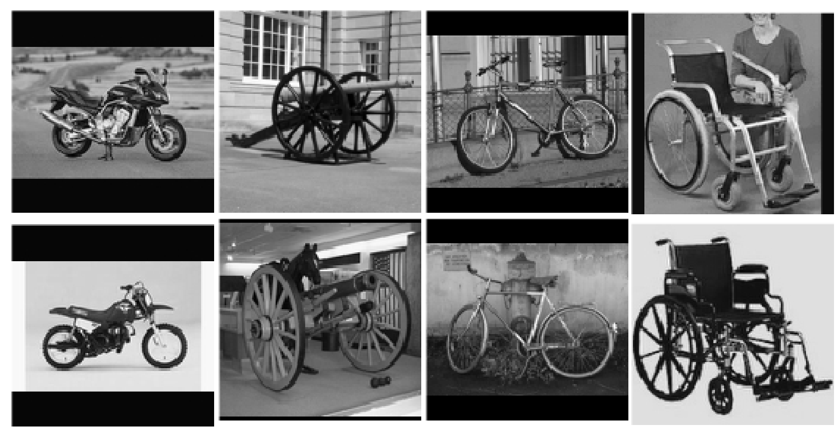


Chinese Restaurant Franchise



Sharing Parts: 16 Categories

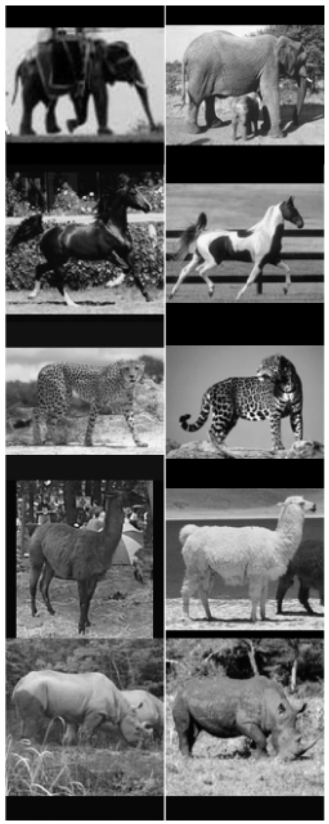




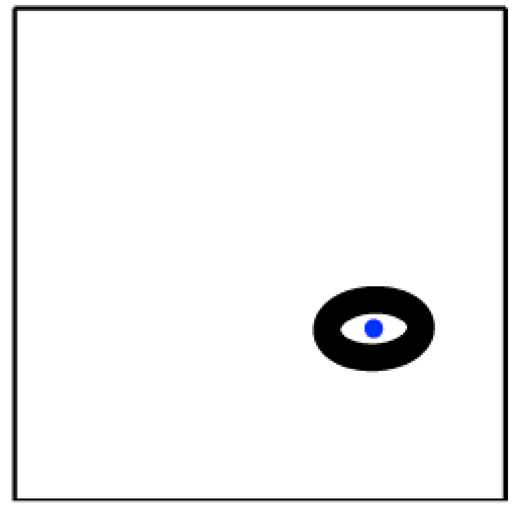
- Caltech 101 Dataset (Li & Perona)
- Horses (Borenstein & Ullman)
- Cat & dog faces (Vidal-Naquet & Ullman)

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

Visualization of Shared Parts



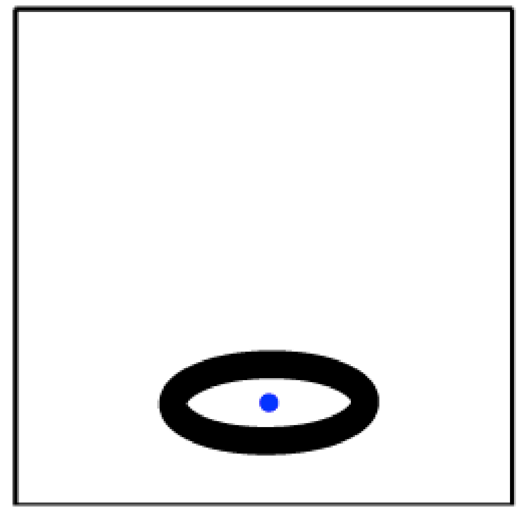
Pr(appearance | part)



Pr(position | part)

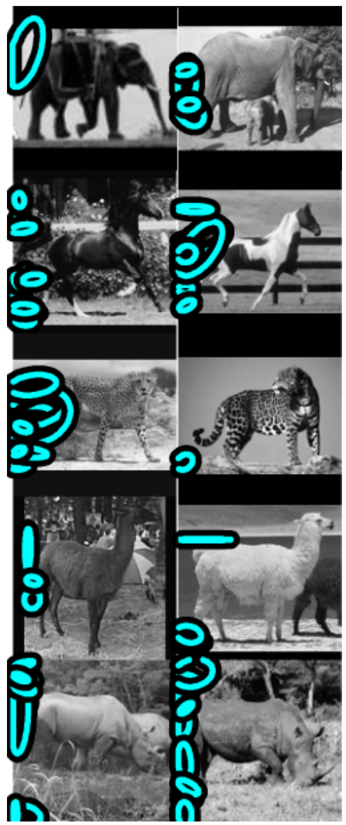
Visualization of Shared Parts

Pr(appearance | part)

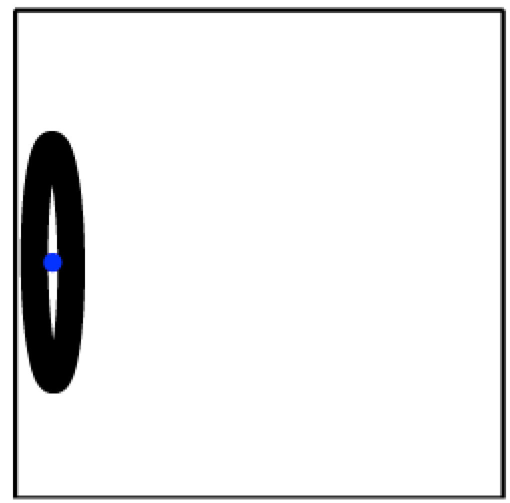


Pr(position | part)

Visualization of Shared Parts

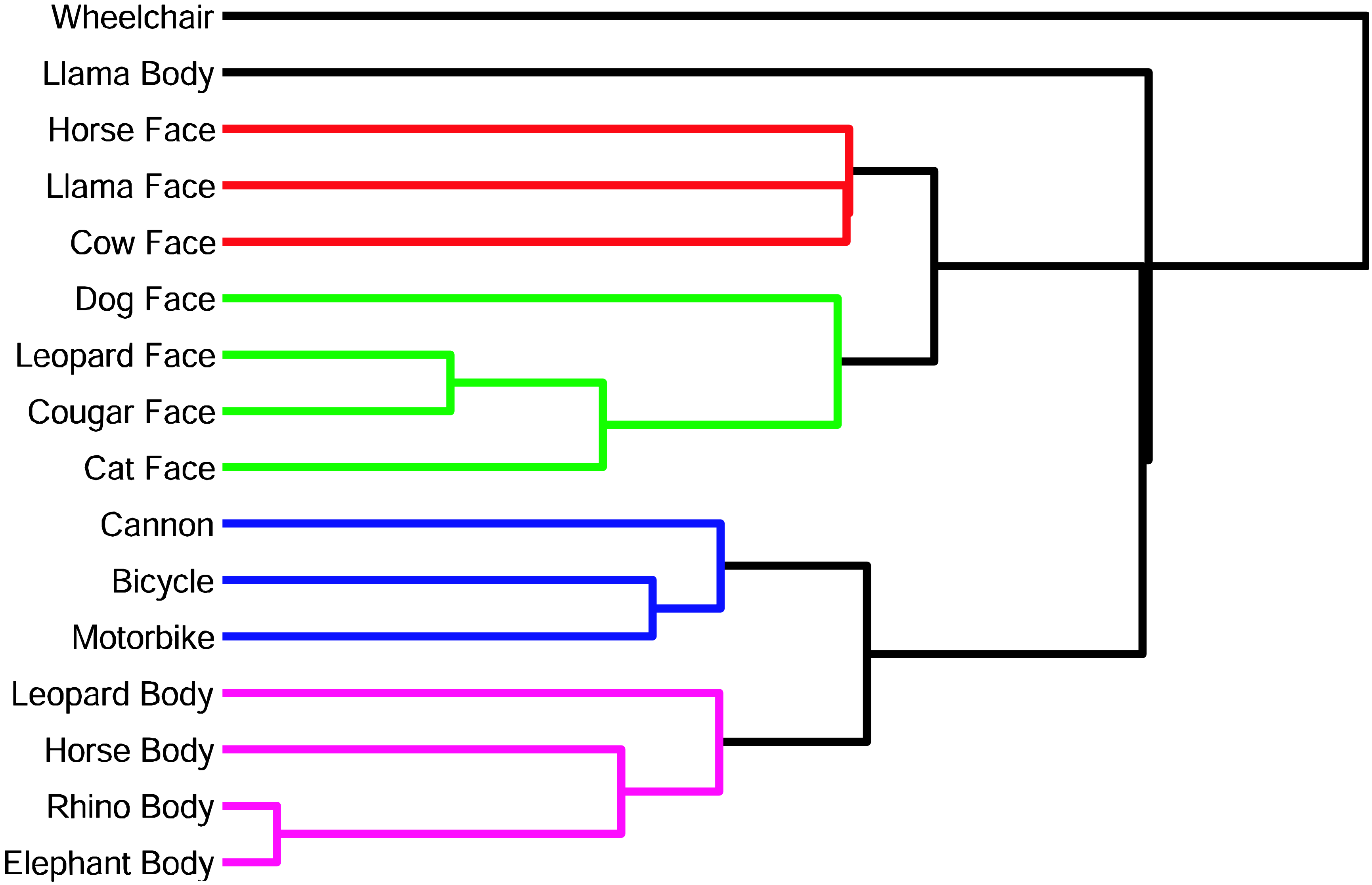


Pr(appearance | part)



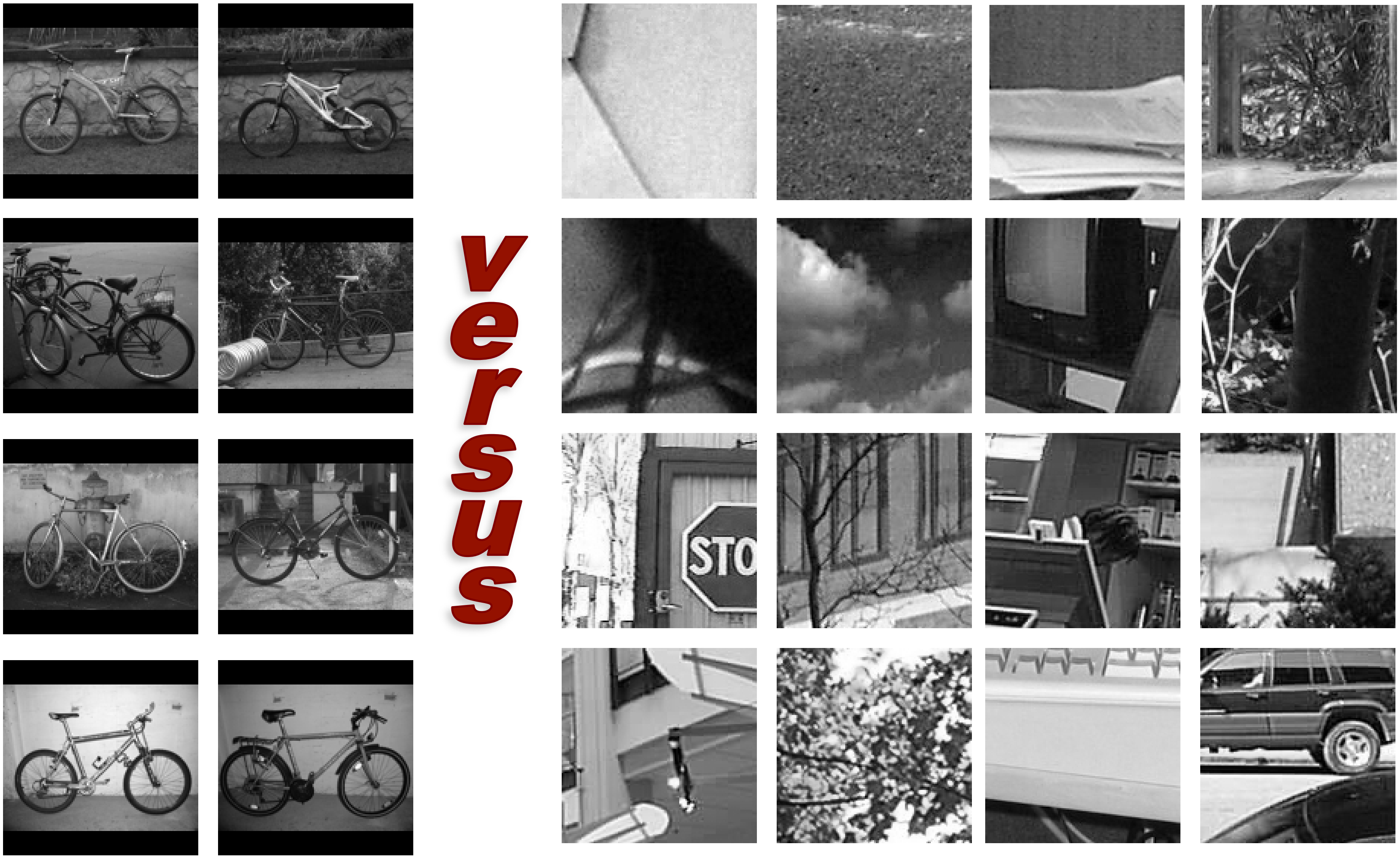
Pr(position | part)

Visualization of Part Densities

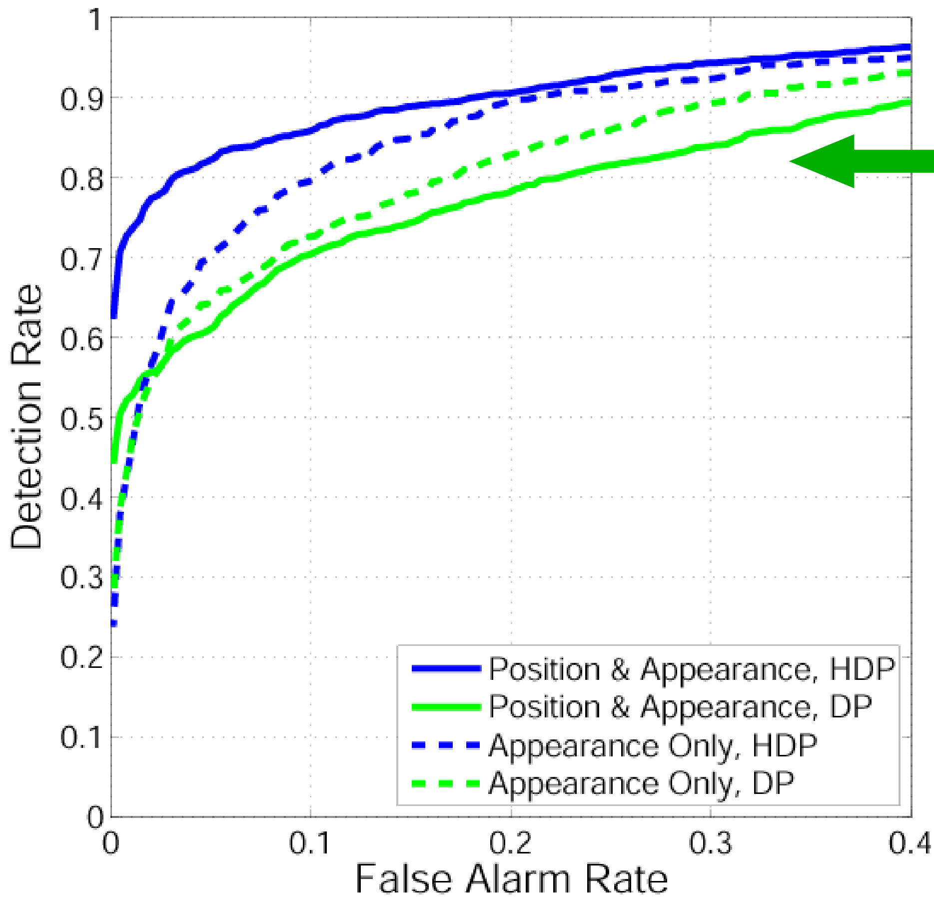


Hierarchical Clustering of Pr(part | object)

Detection Task



Detection Results

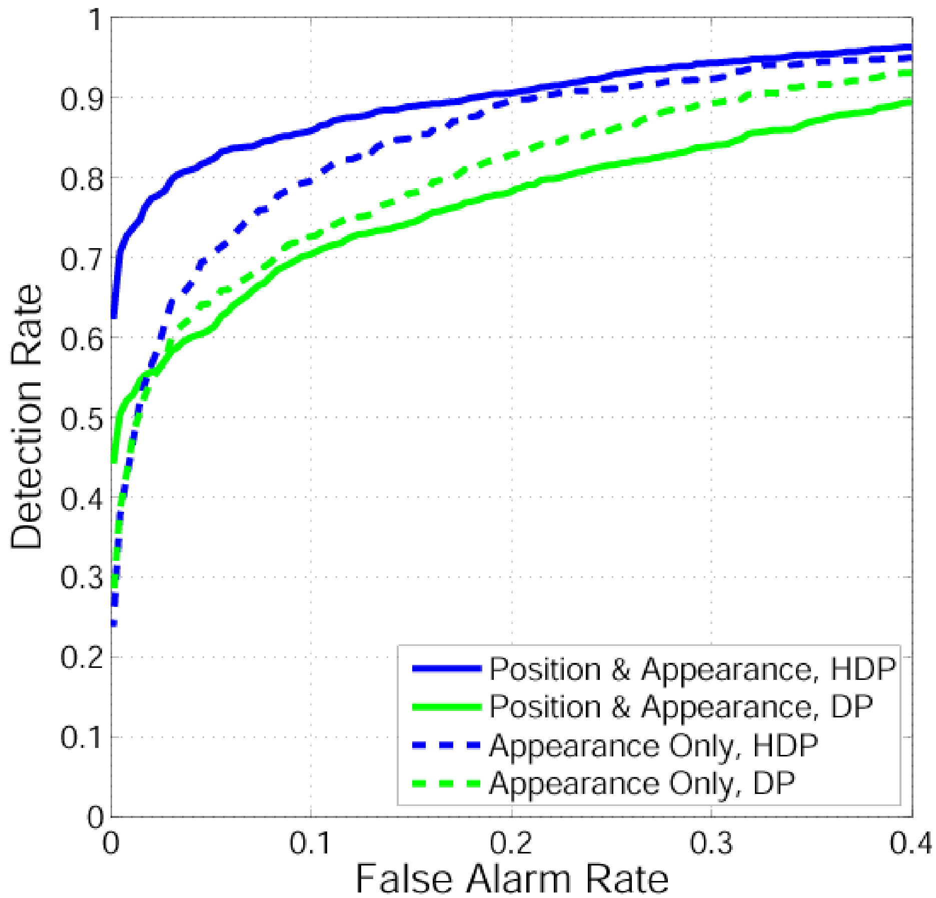


6 Training Images per Category
(ROC Curves)

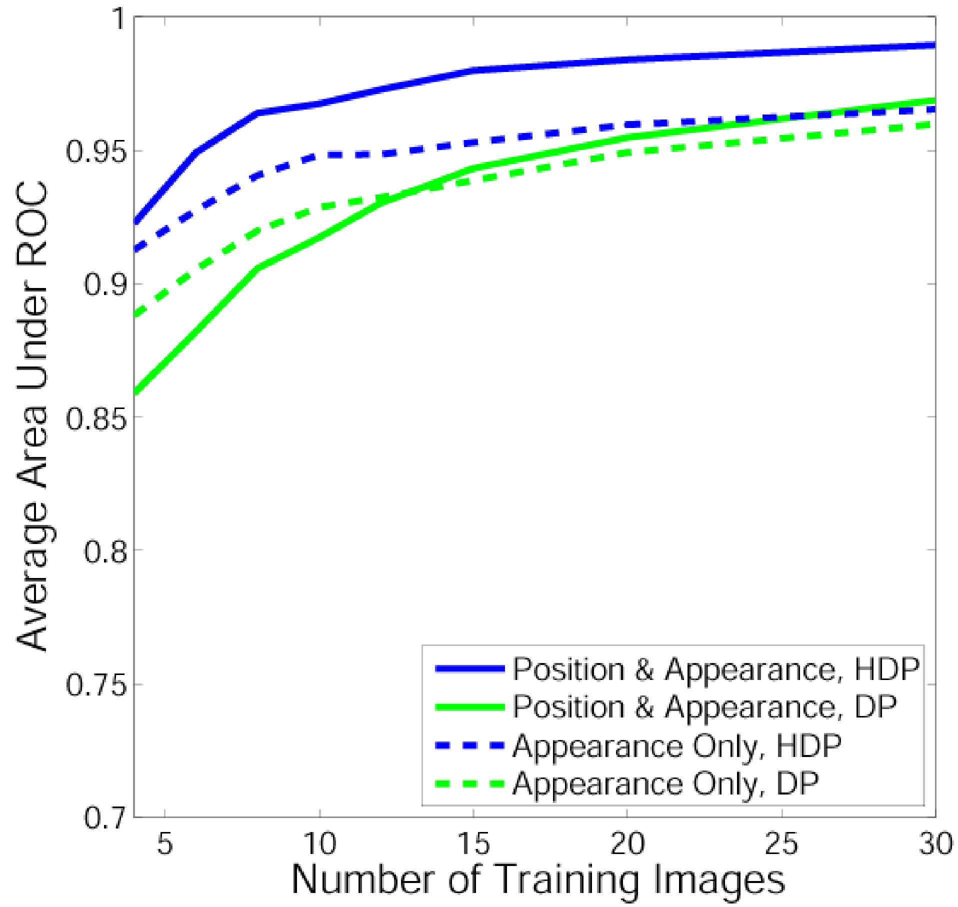
Shared Parts
more accurate than
Unshared Parts

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

Detection Results

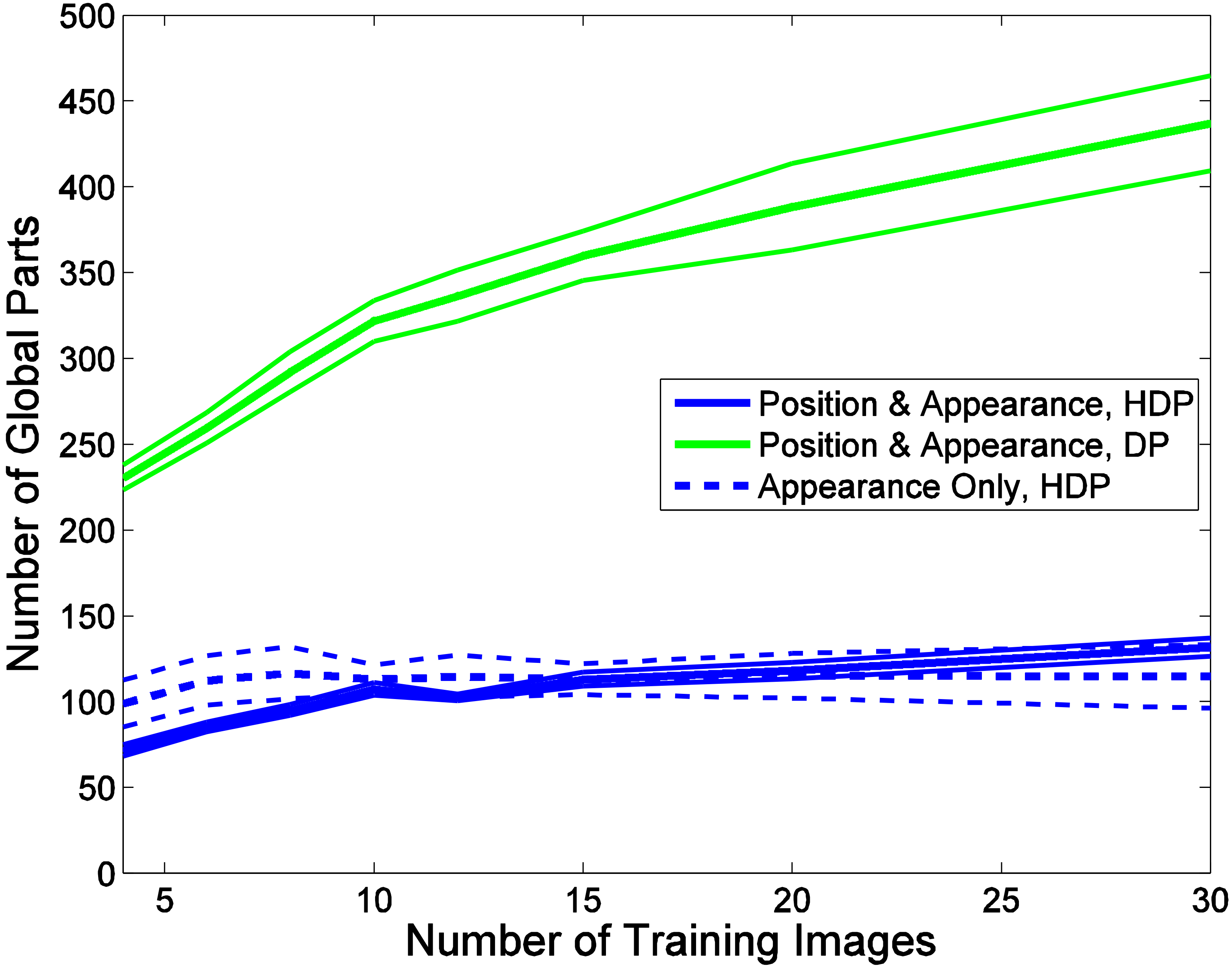


6 Training Images per Category
(ROC Curves)

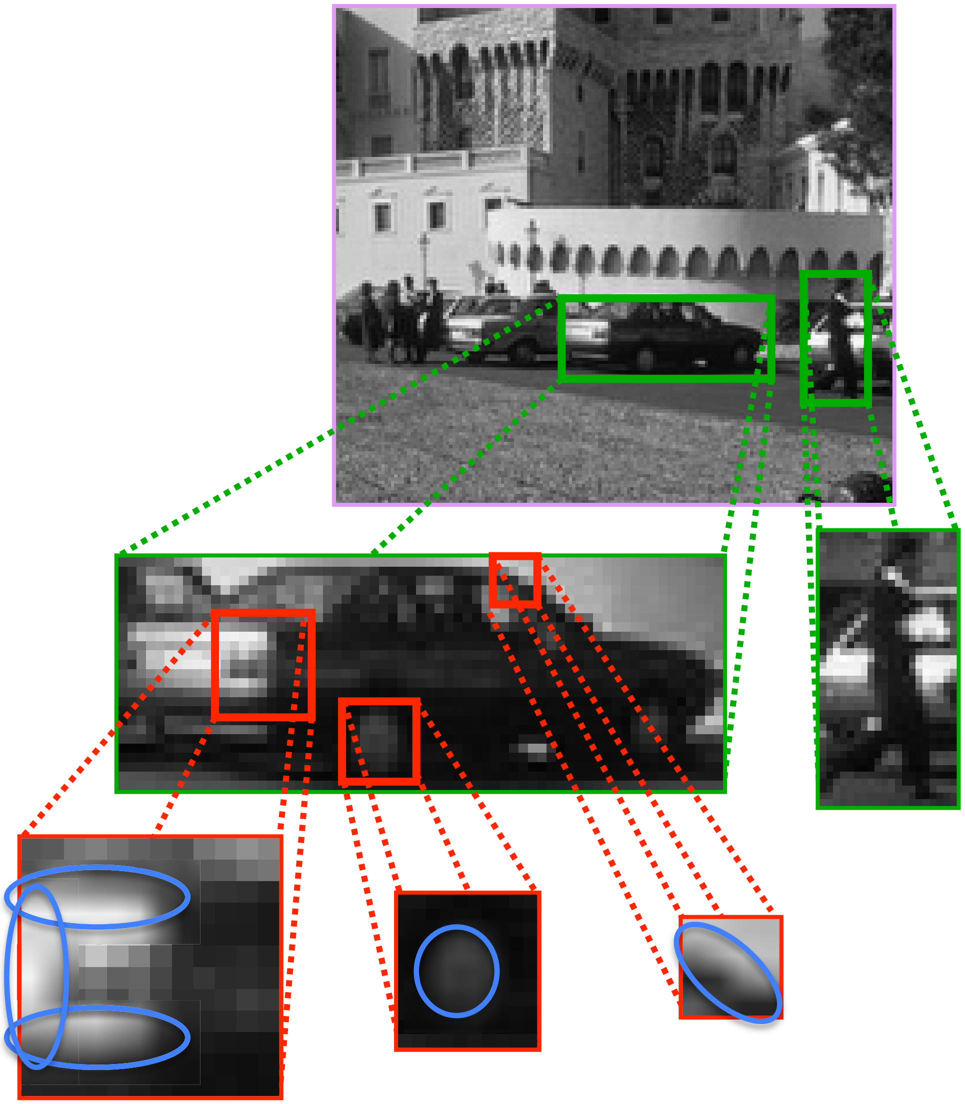


Detection vs. Training Set Size
(Area Under ROC)

Sharing Simplifies Models



Scenes, Objects, and Parts



Scene

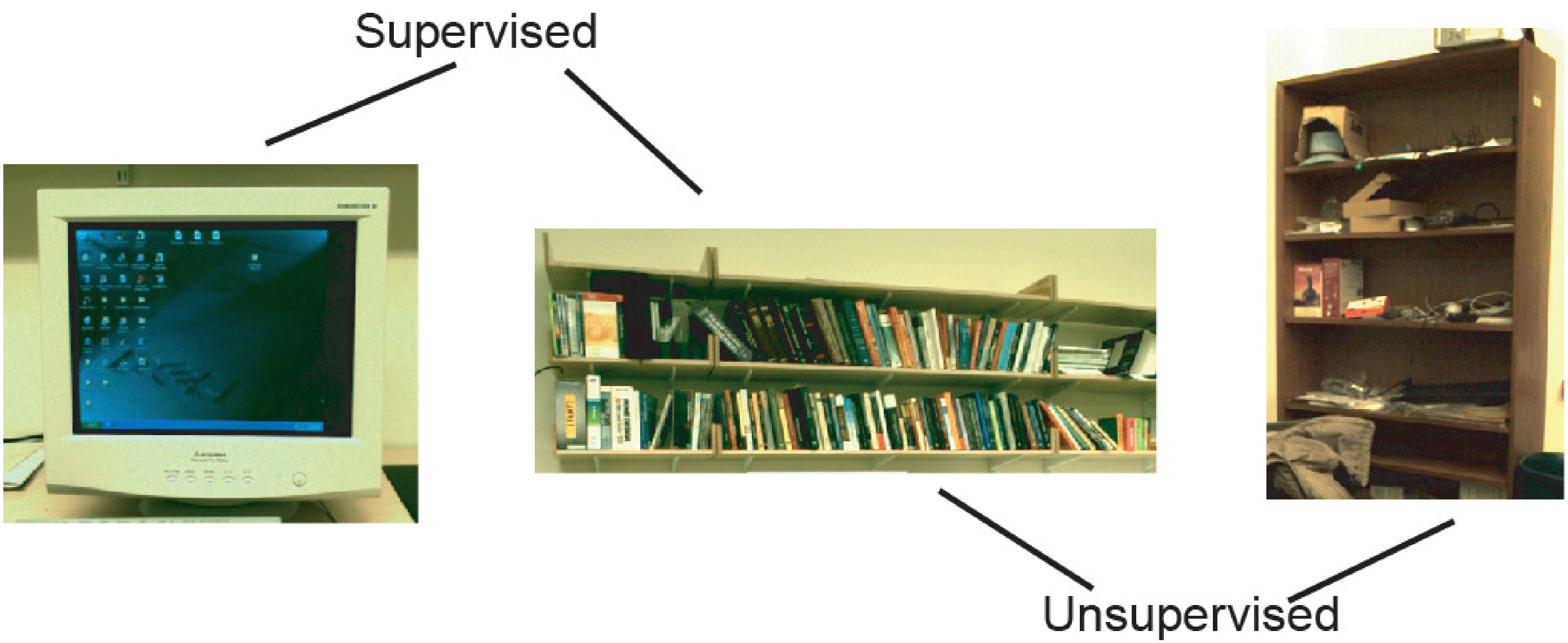
Objects

Parts

Features

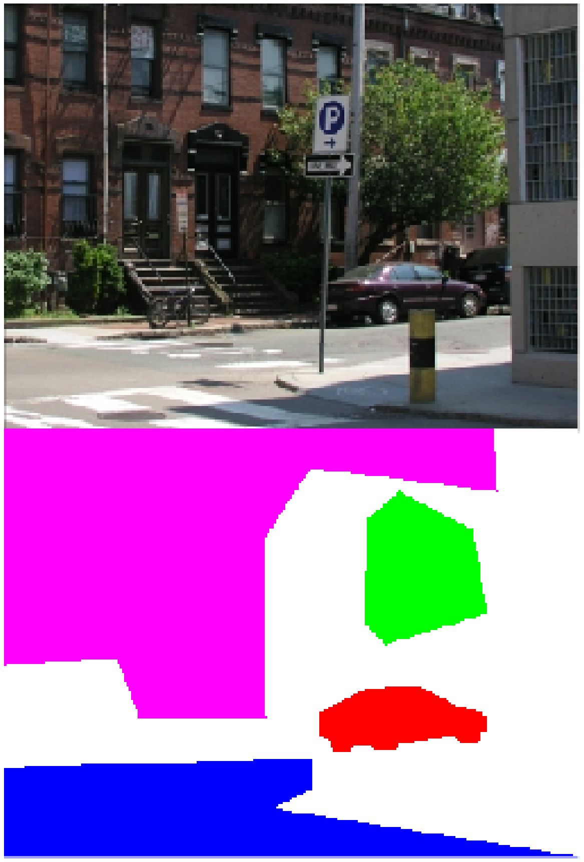
Contextual Transfer Learning

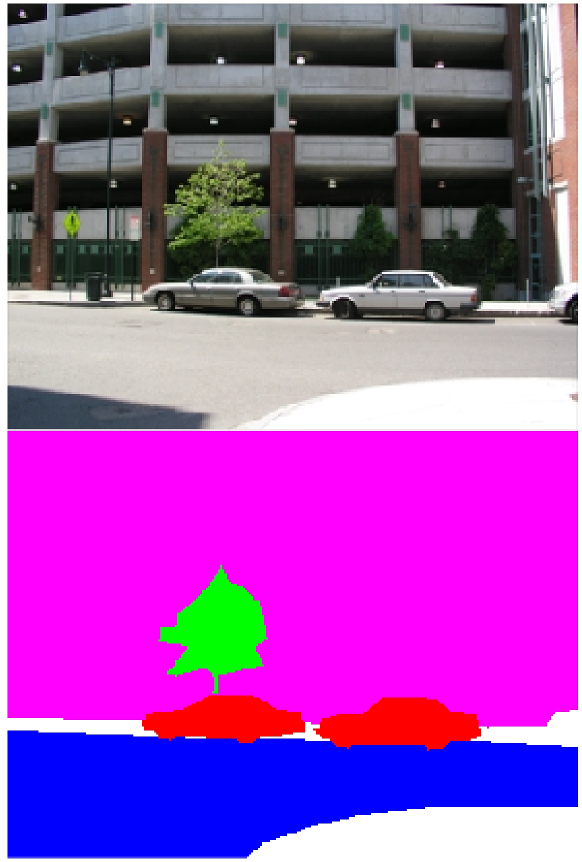
Object vs. Visual Categories



- 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

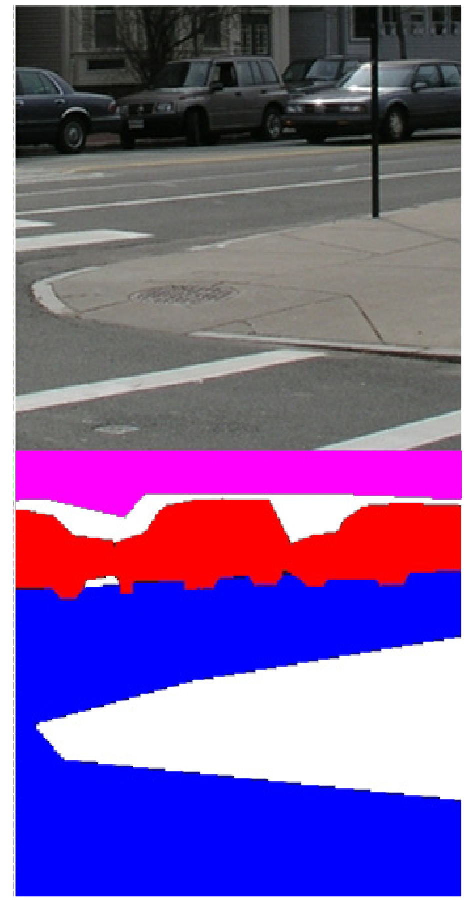
Spatial Transformations

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

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

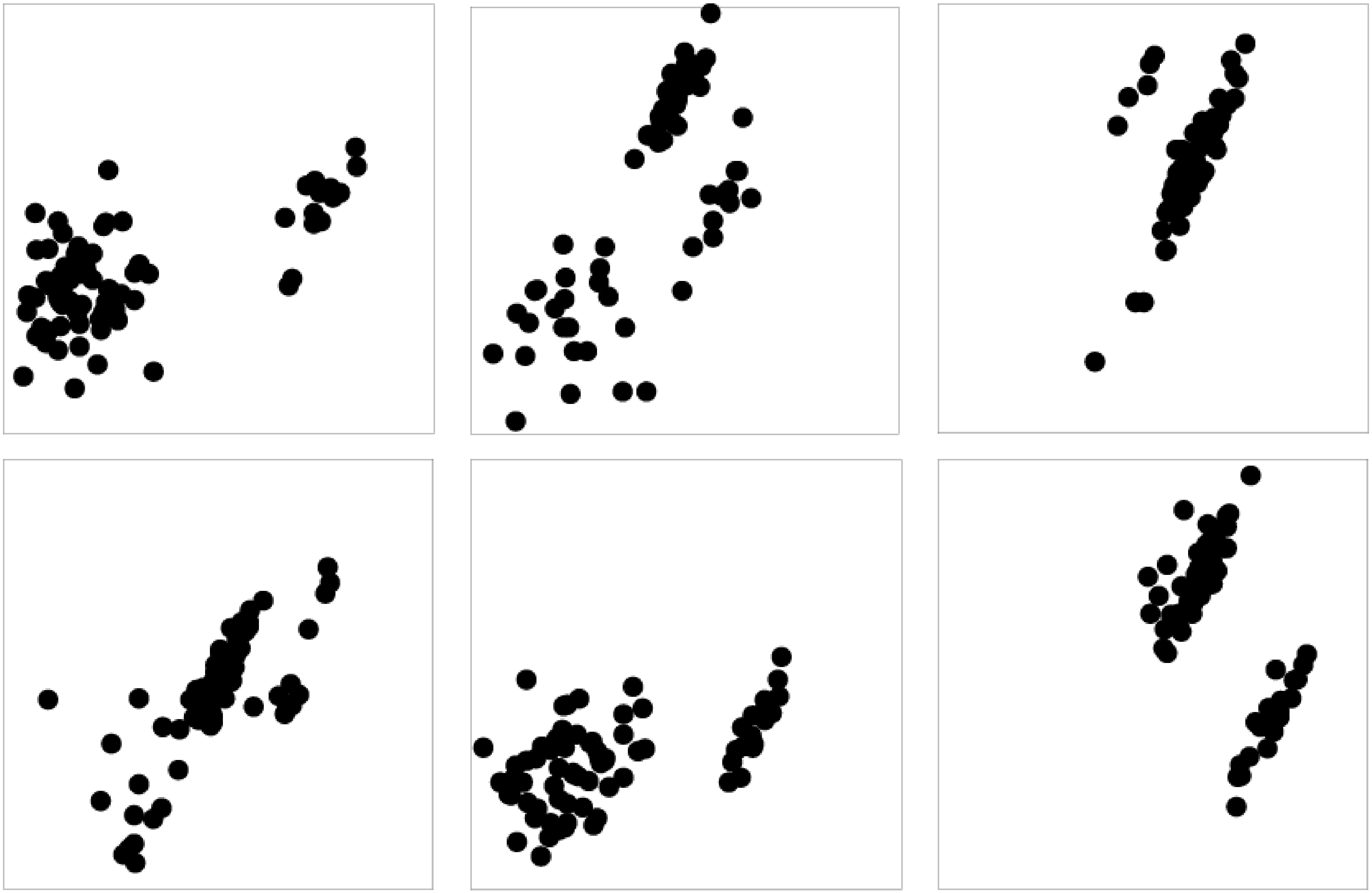
Parameterized family of transformations

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

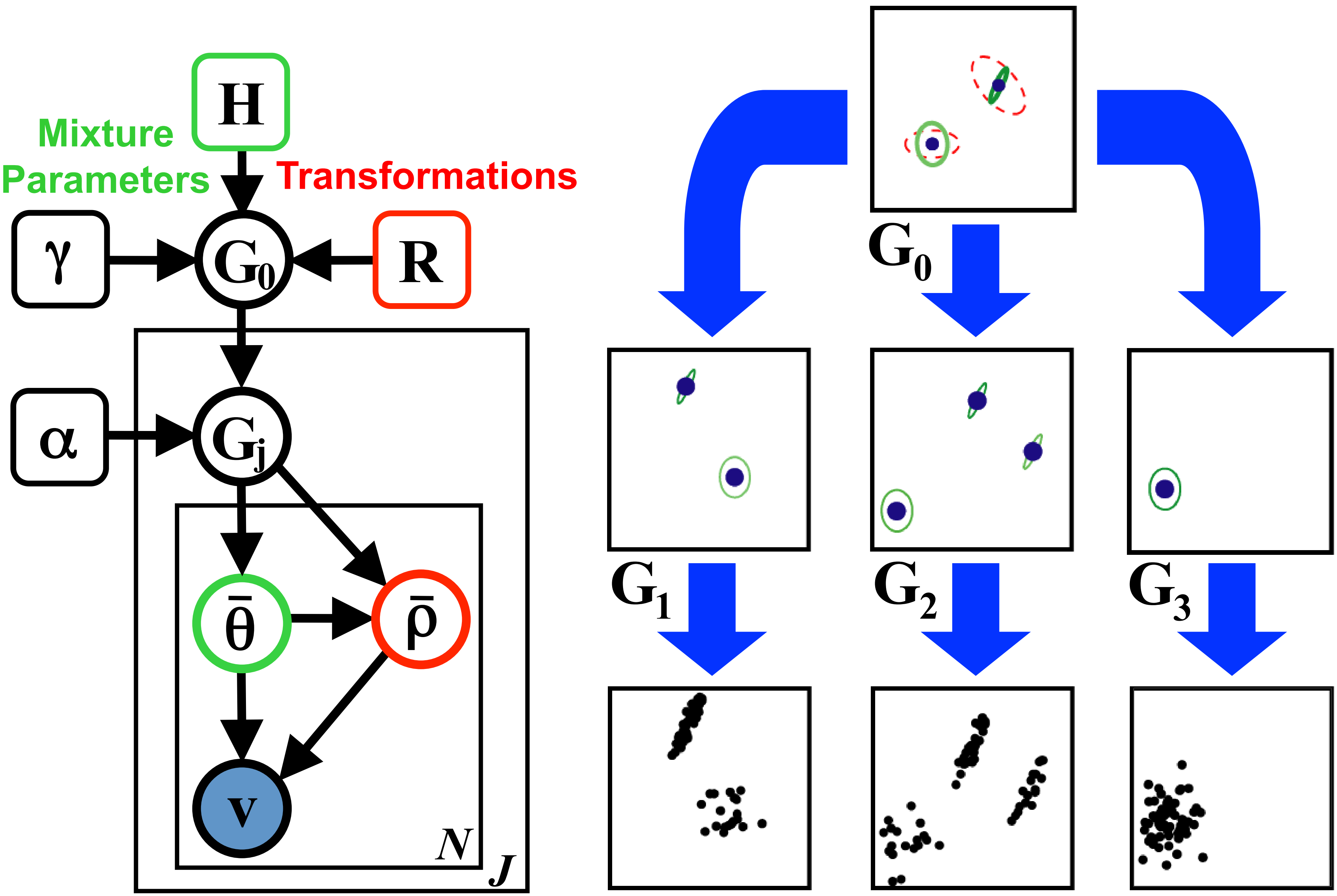


Layered Motion Models *(Darrell & Pentland 1991, Wang & Adelson 1994, Jojic & Frey 2001)*
Nonparametric Transformation Densities *(Learned-Miller & Viola 2000)*

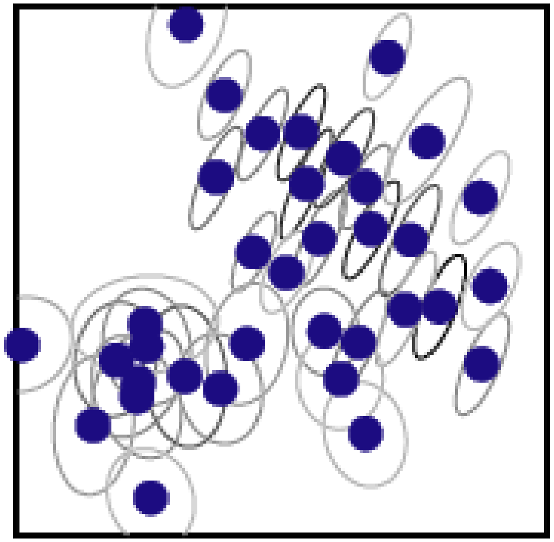
A Toy World: Bars & Blobs



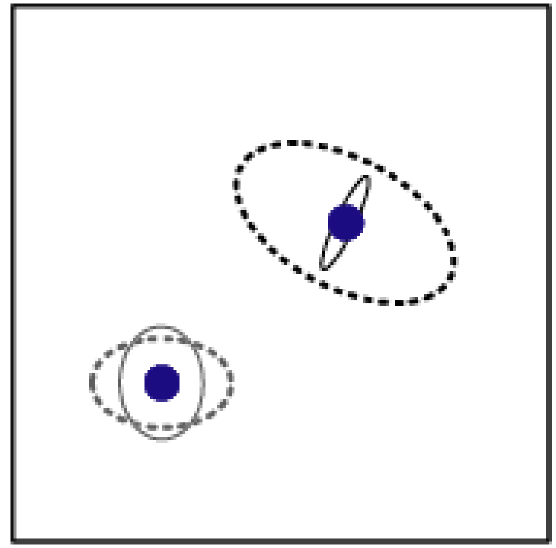
Transformed Dirichlet Process



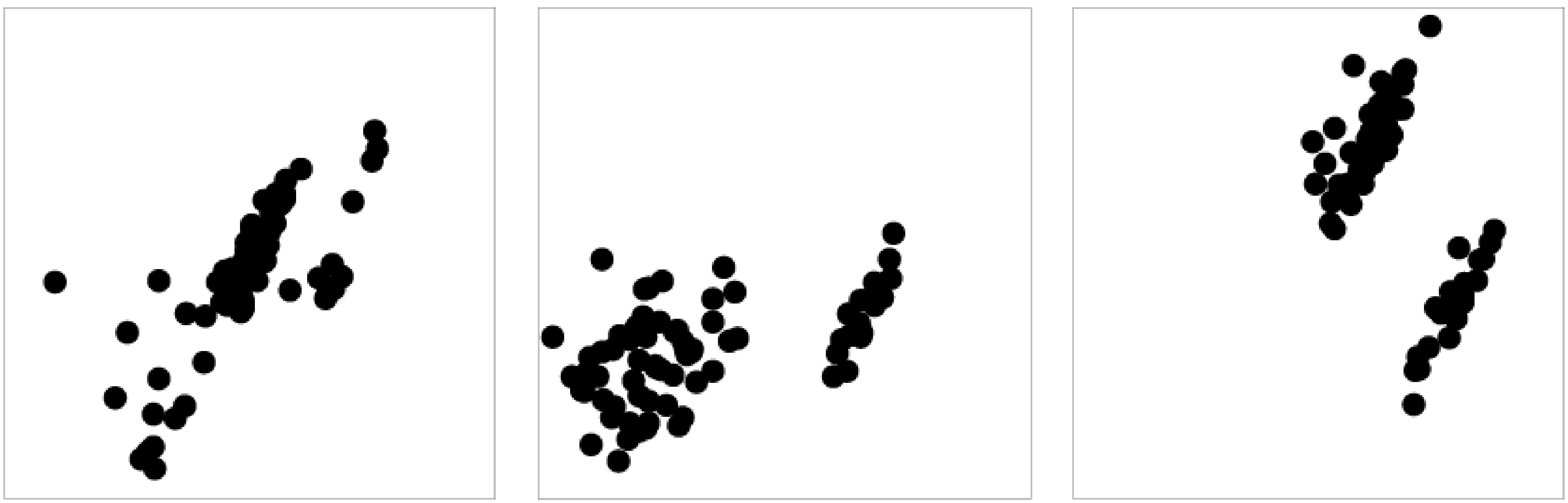
Importance of Transformations



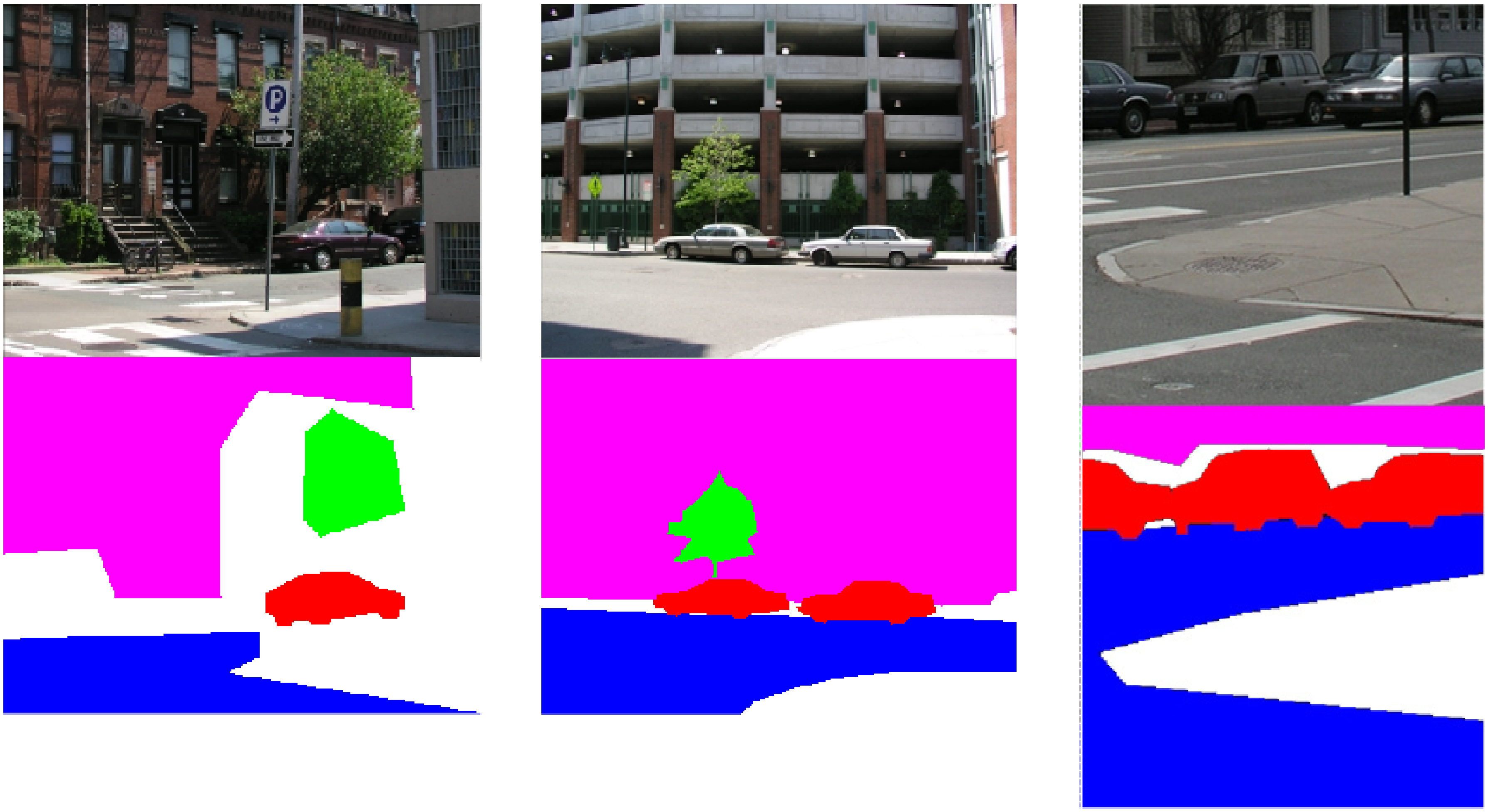
HDP



TDP



Counting & Locating Objects

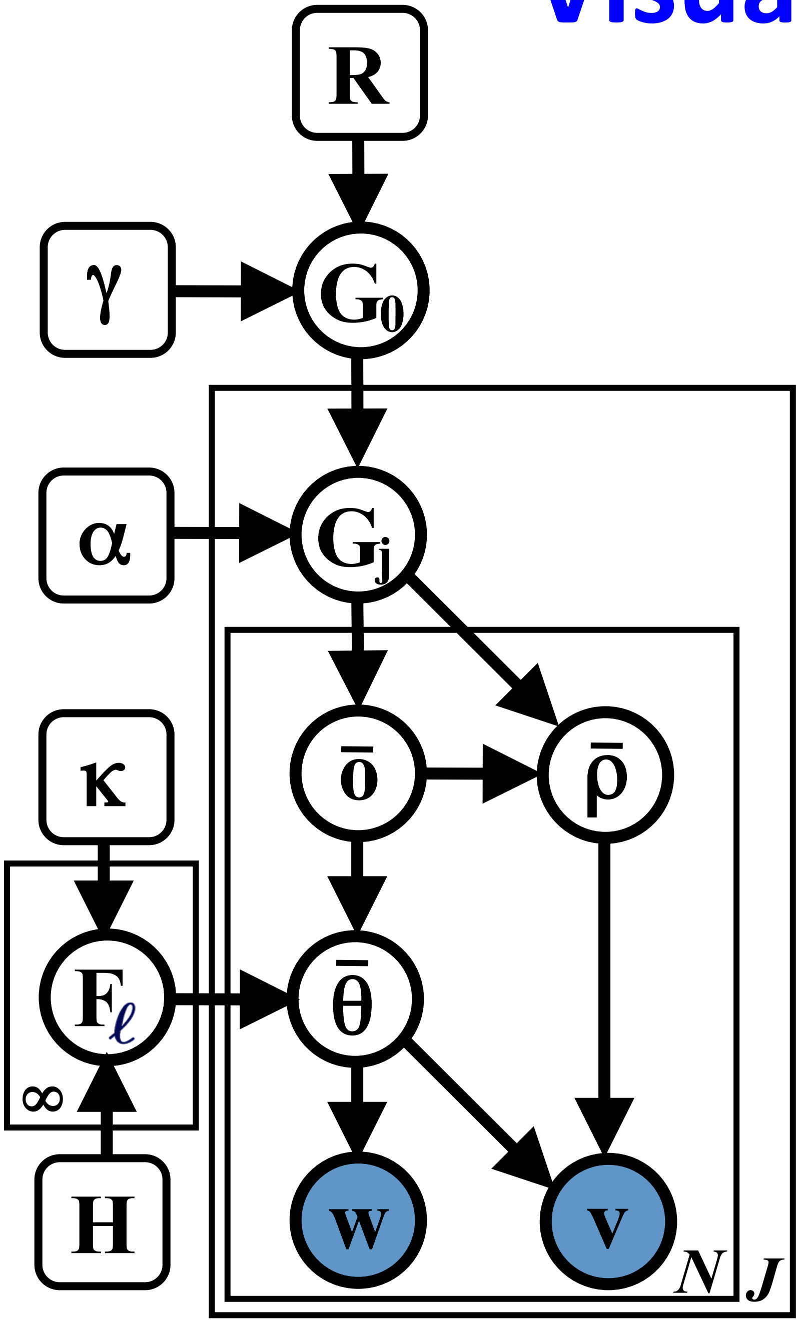


- How many cars are there?
- Where are those cars in the scene?

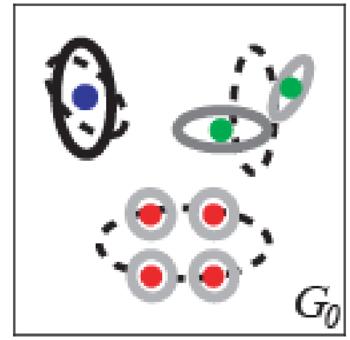
Dirichlet Processes

Transformations

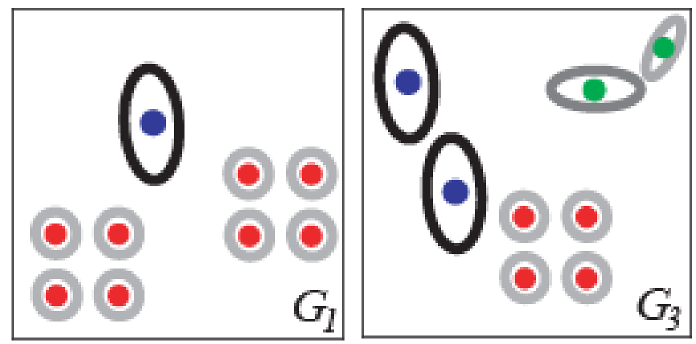
Visual Scene TDP



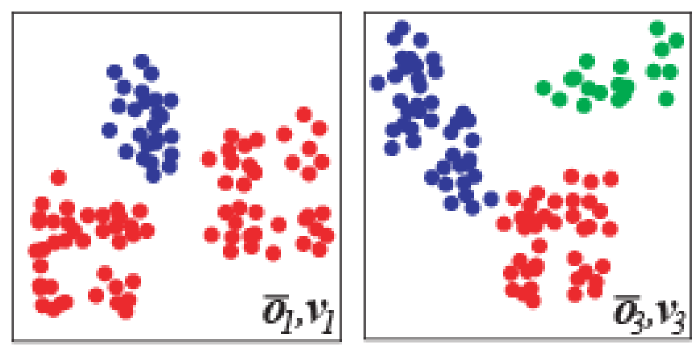
Global Density
Object category
Part size & shape
Transformation prior



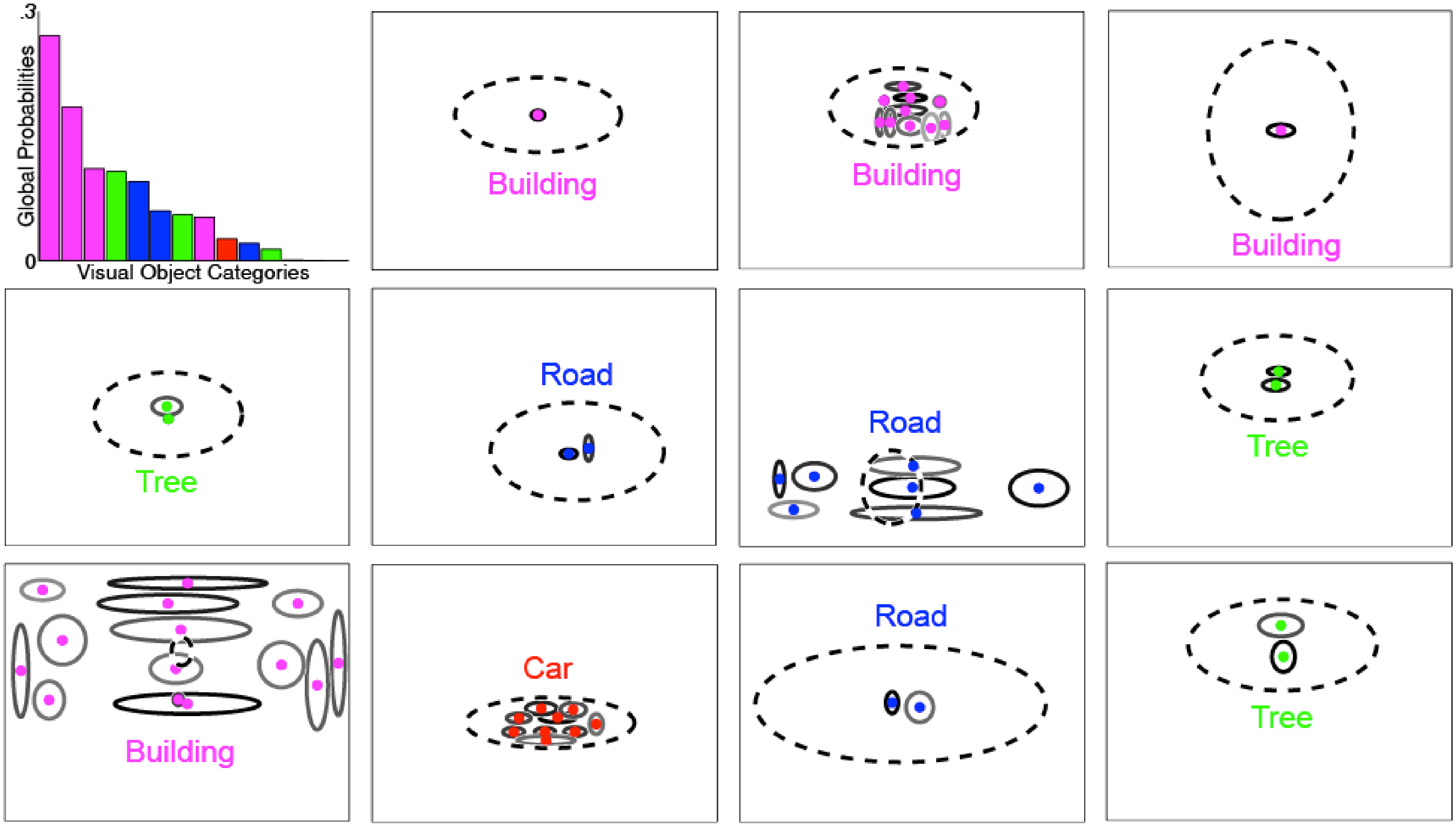
Transformed Densities
Object category
Part size & shape
Instance locations



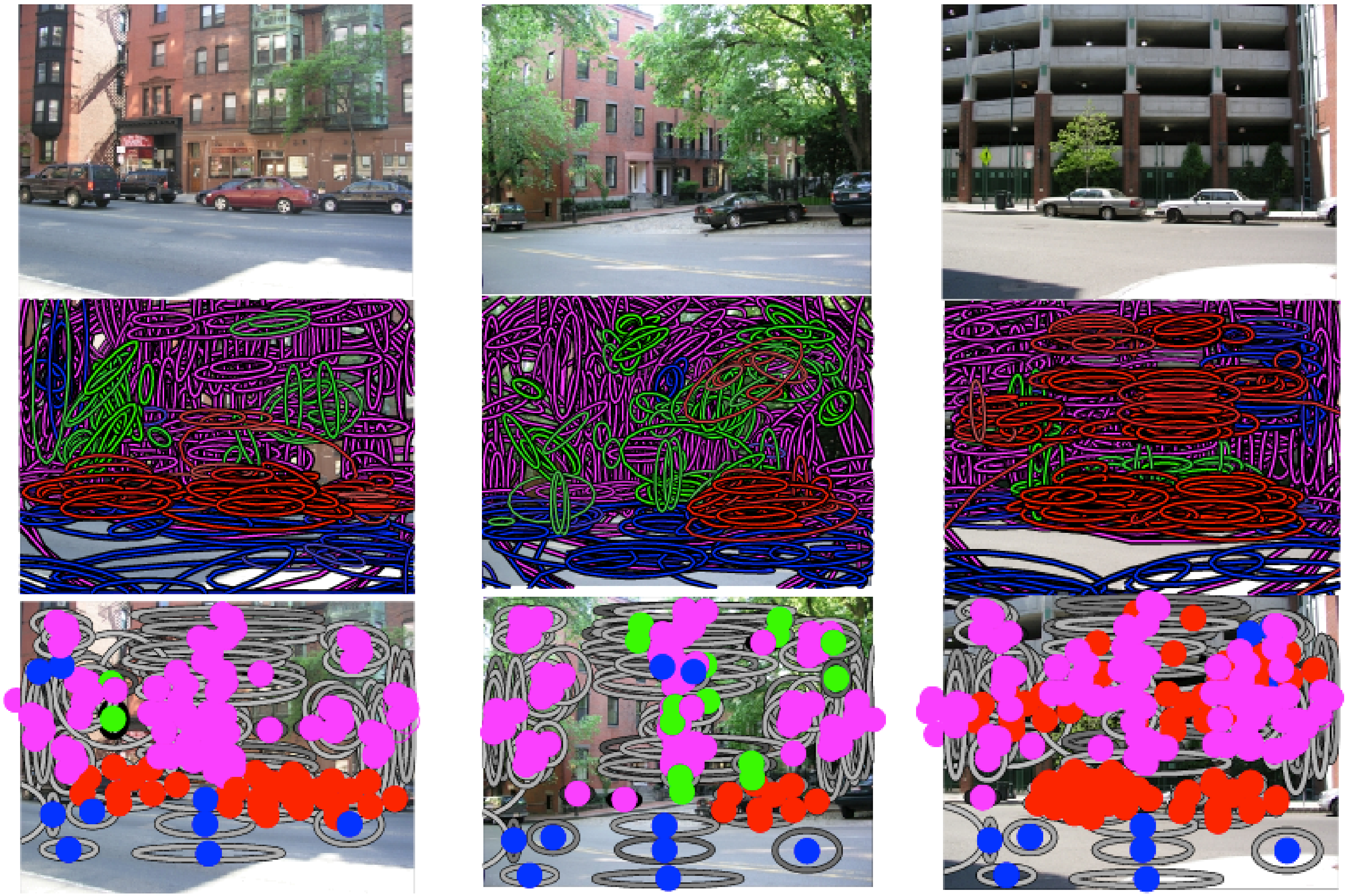
2D Image Features
Appearance
Location



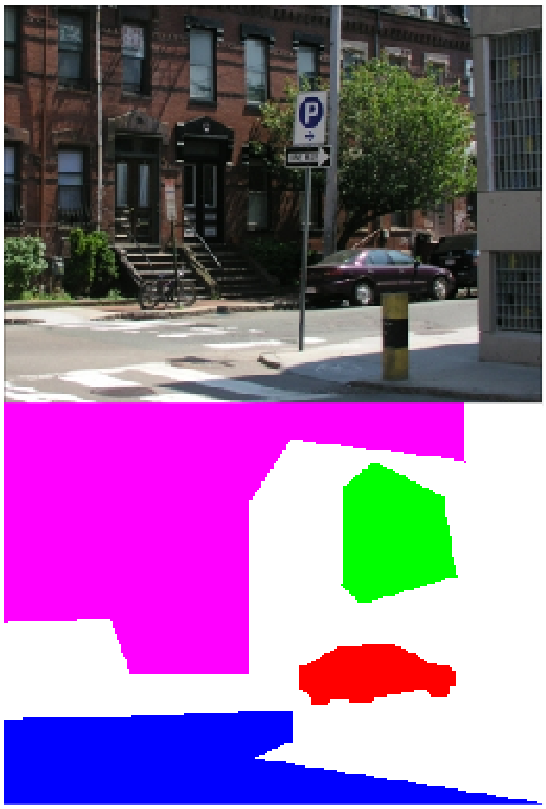
Street Scene Visual Categories

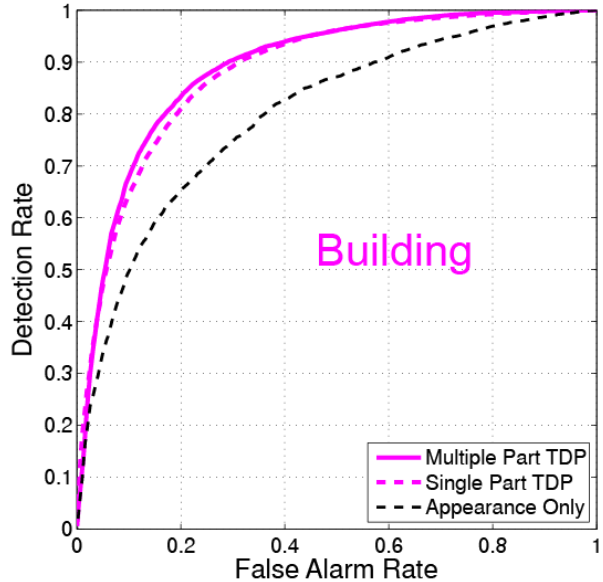


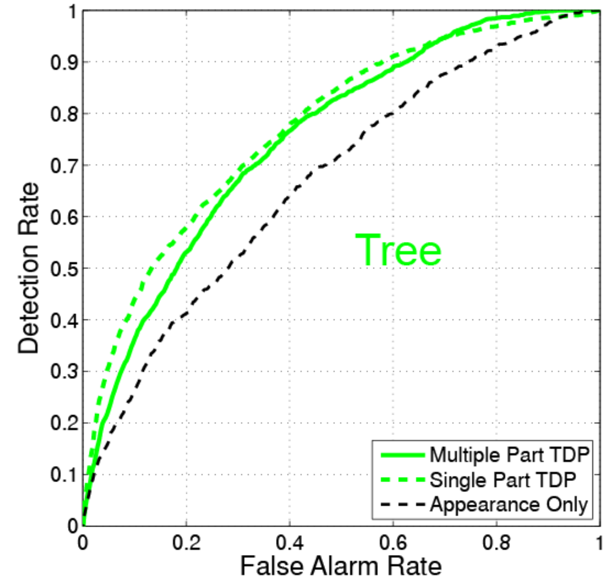
Street Scene Segmentations

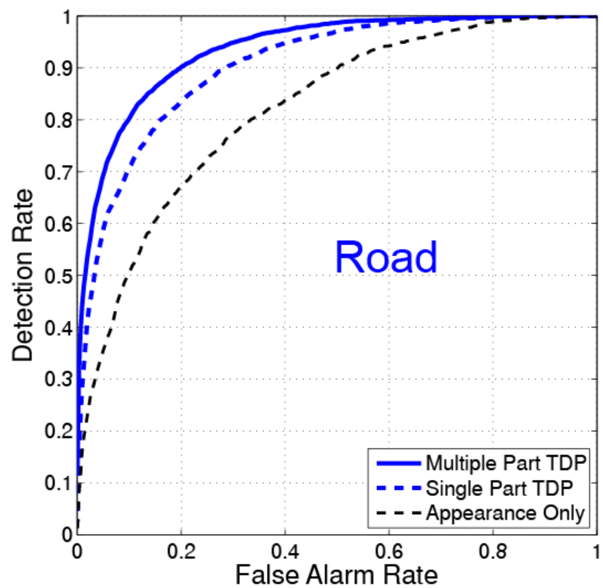


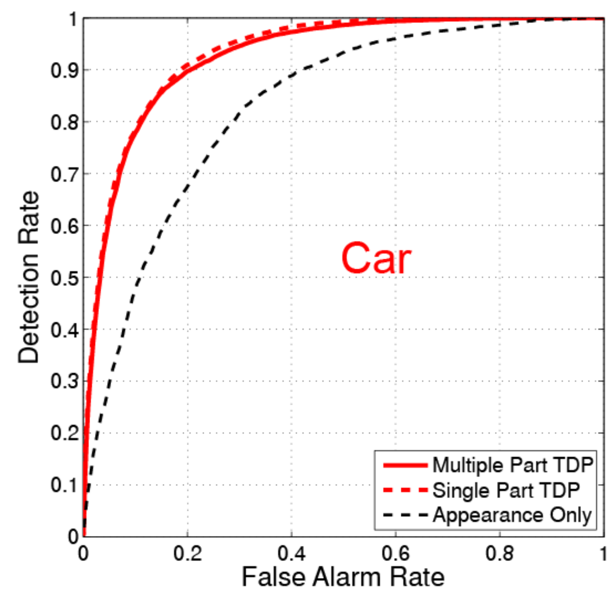
Segmentation Performance





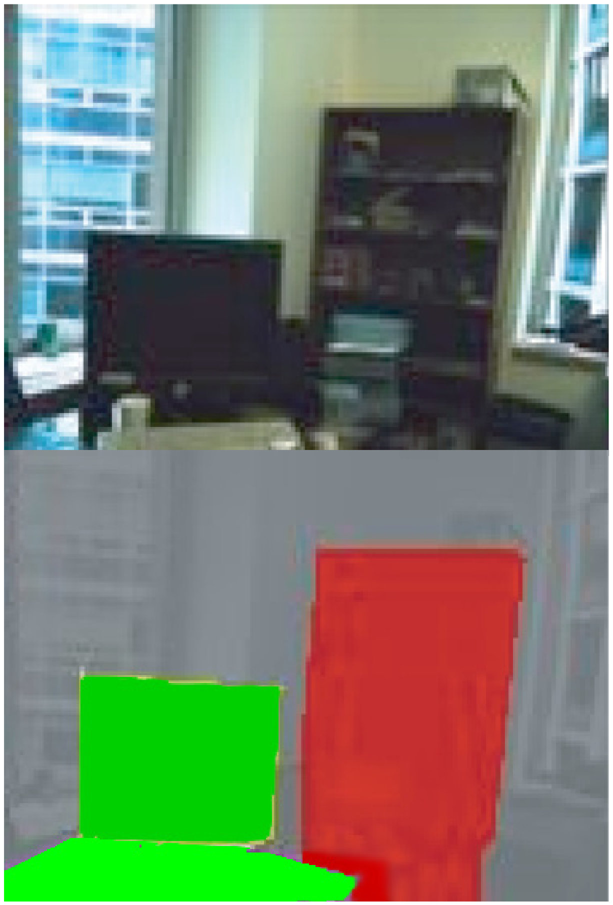






Extension: 3D Scenes

Office Scene

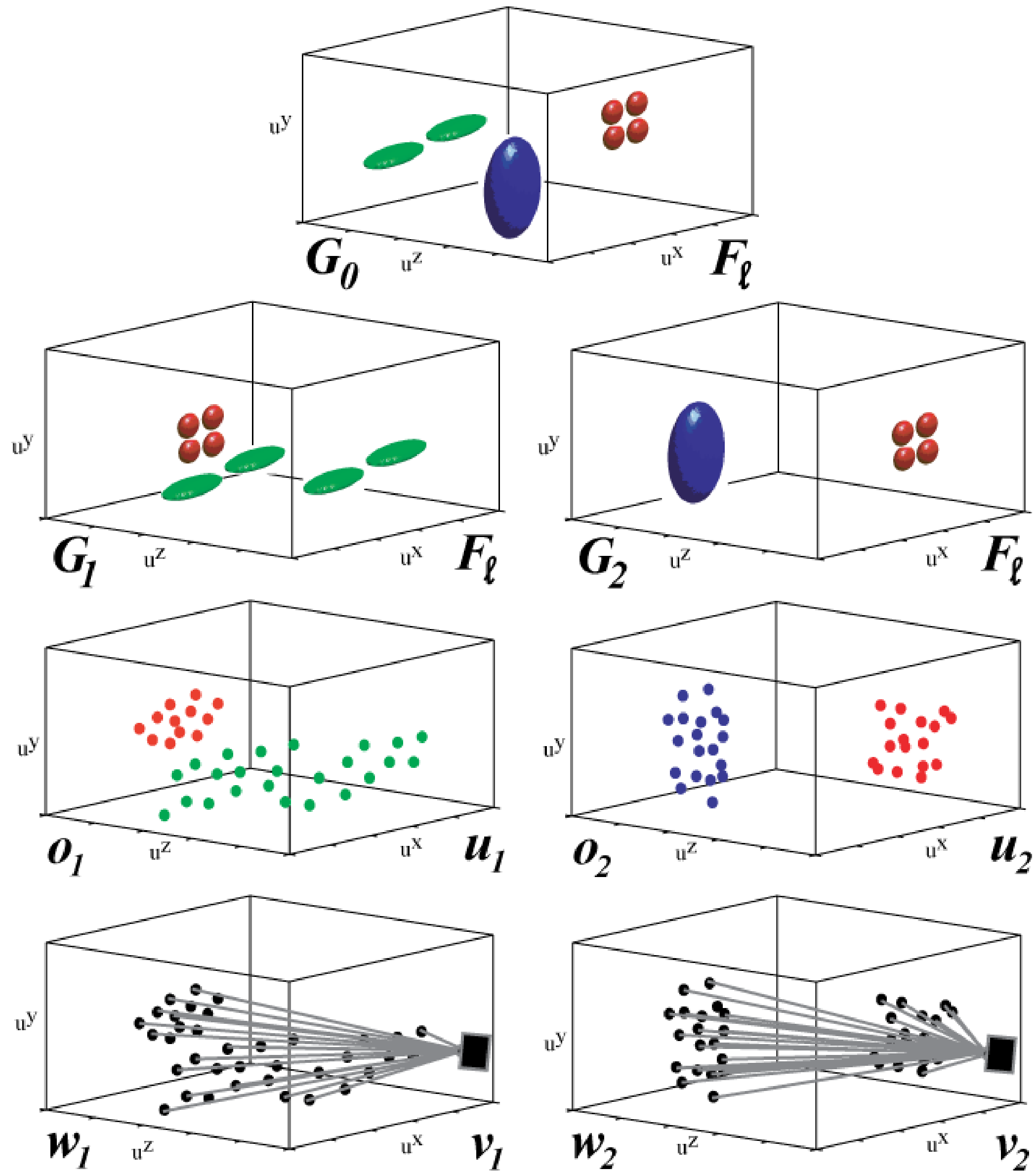


Red

Far

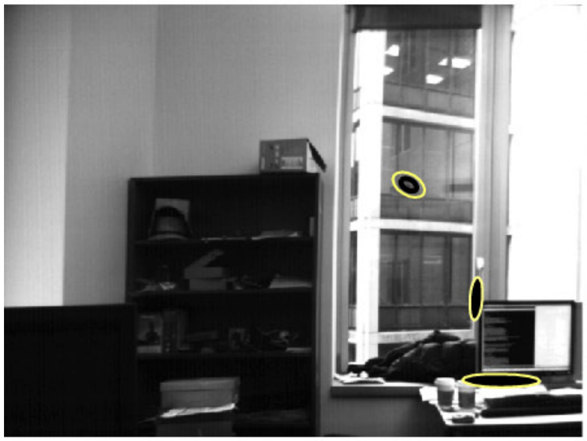
Green

Near

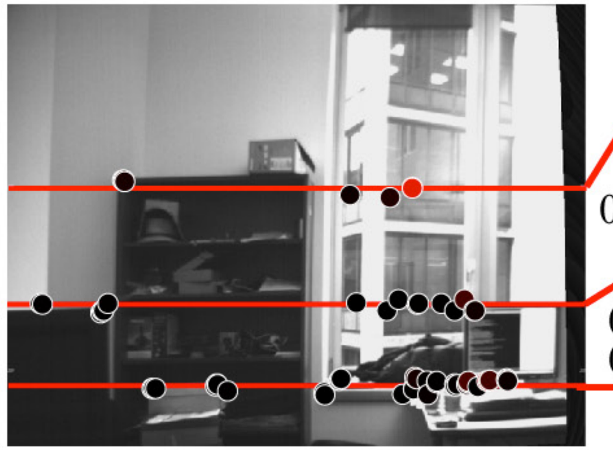


- Segmentation easier in 3D
- Identifying known objects regularizes depth estimation

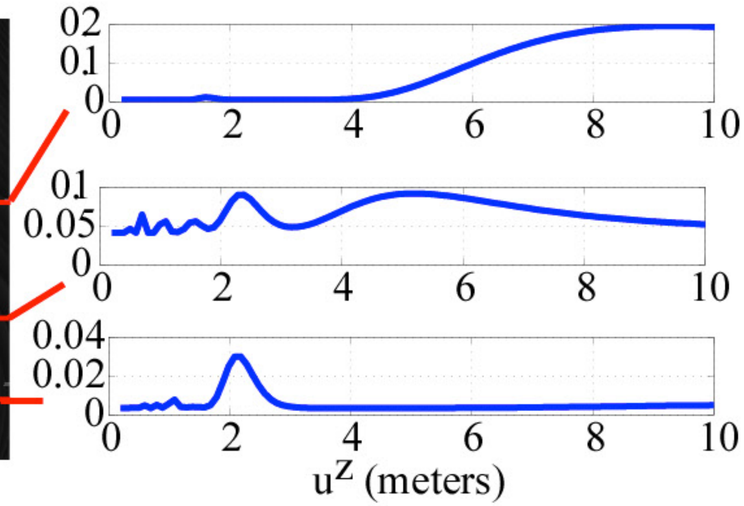
3D Structure from Stereo



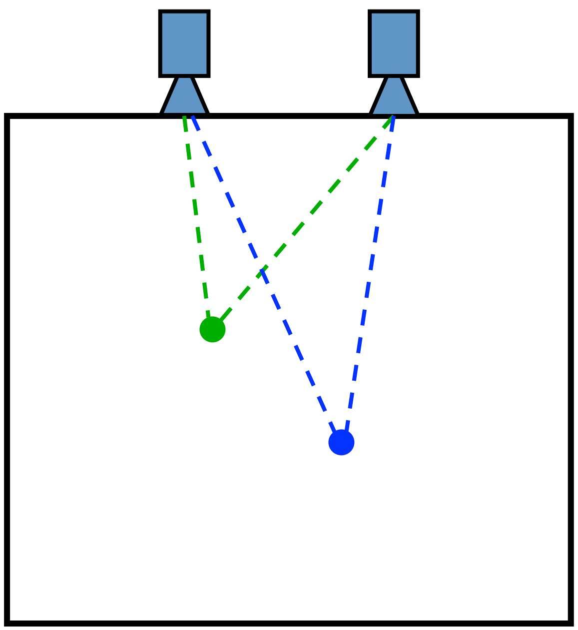
Reference (left) Image



Potential Matches



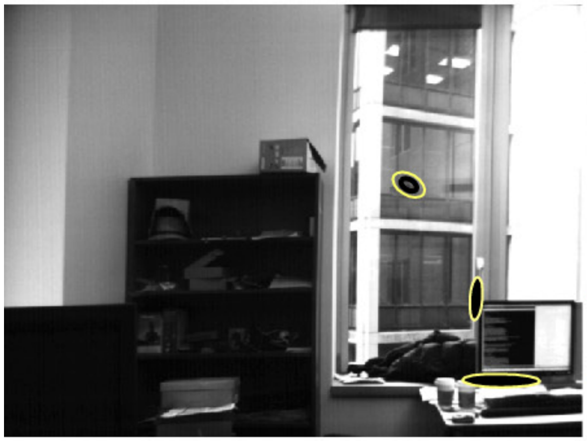
Depth Densities

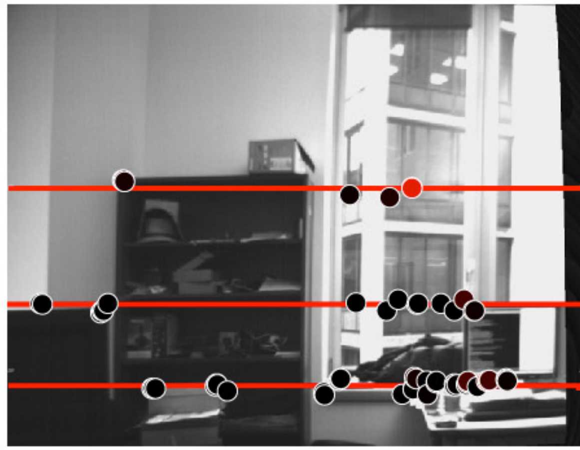


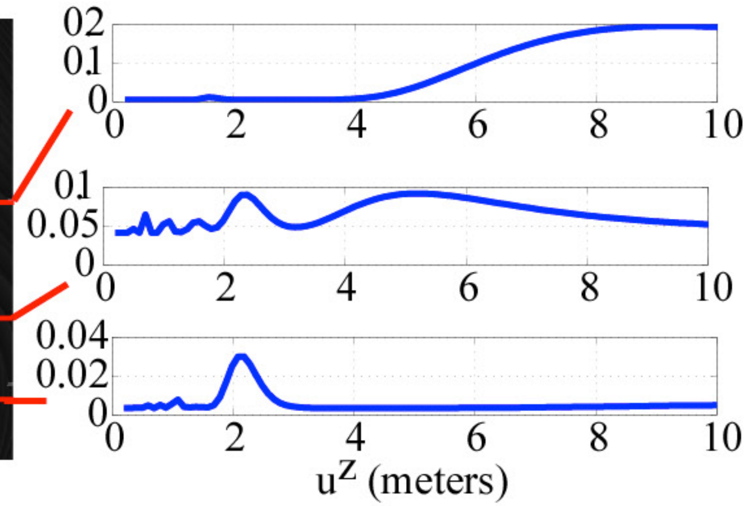
Overhead View

$$\textit{Depth} = \frac{\delta}{\textit{Disparity}}$$

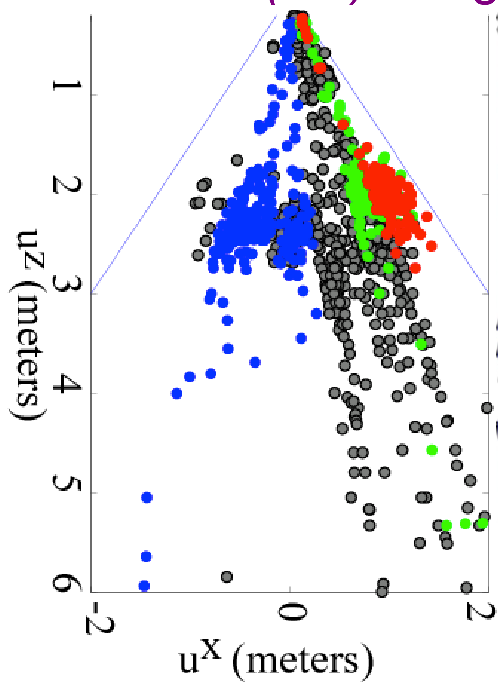
Greedy Depth Estimates

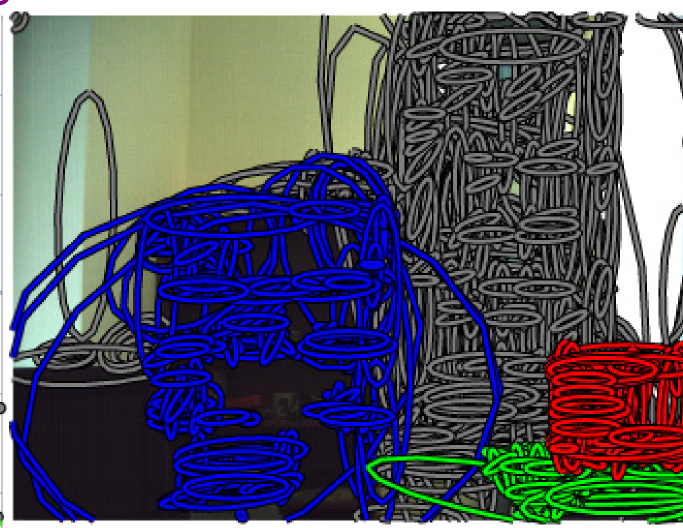


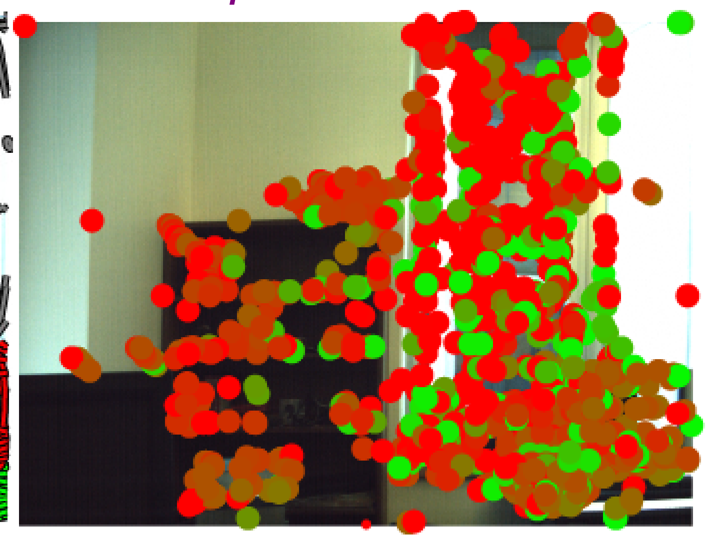




Reference (left) Image *Potential Matches* *Depth Densities*



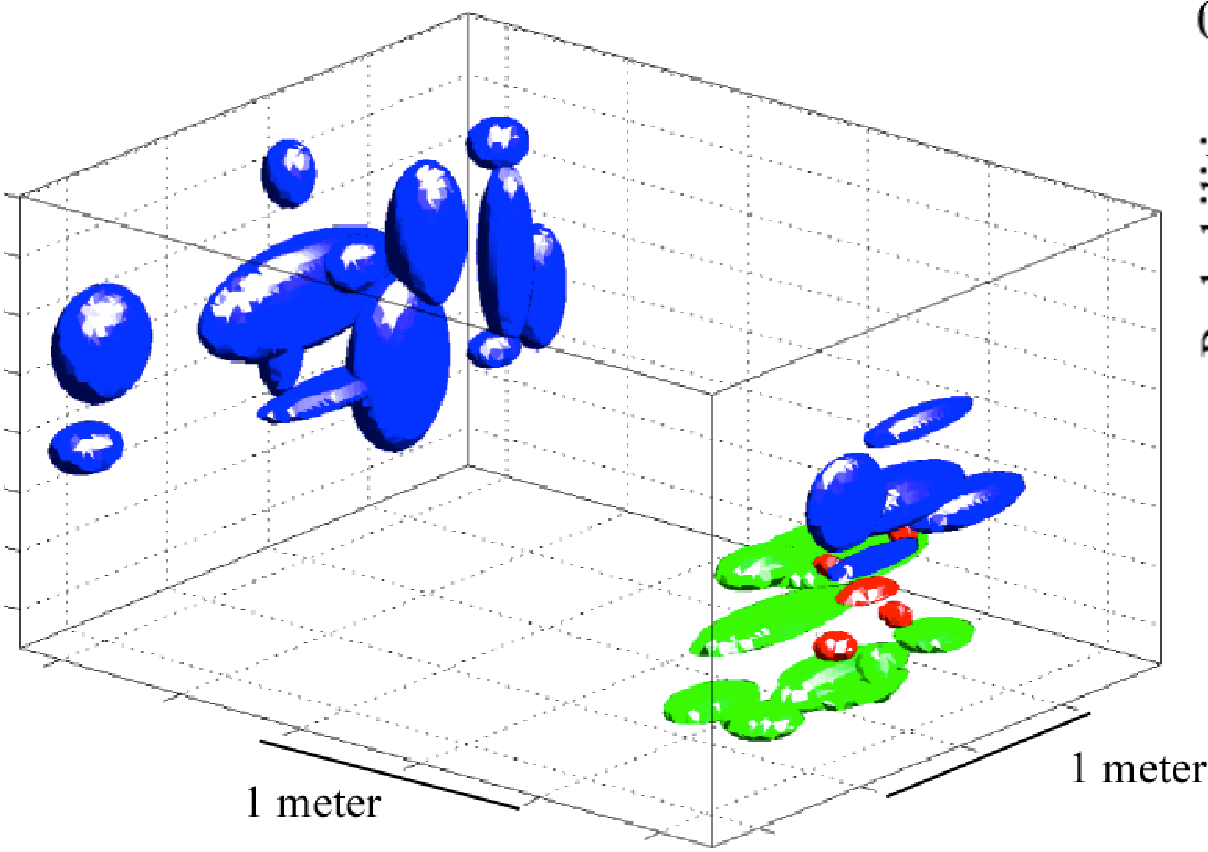


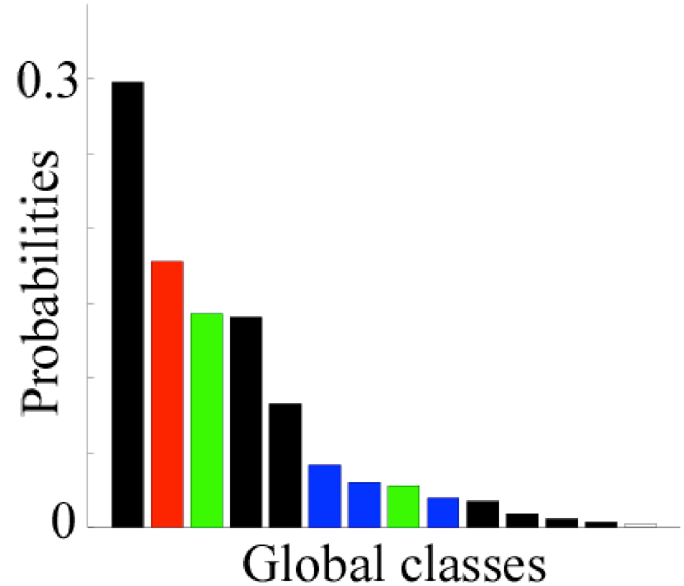


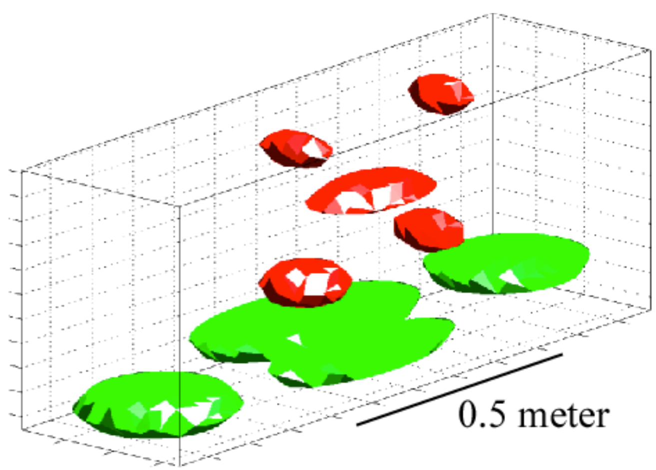
Green ⟷ Near

Red ⟷ Far

3D Transformed DP: Office Scenes

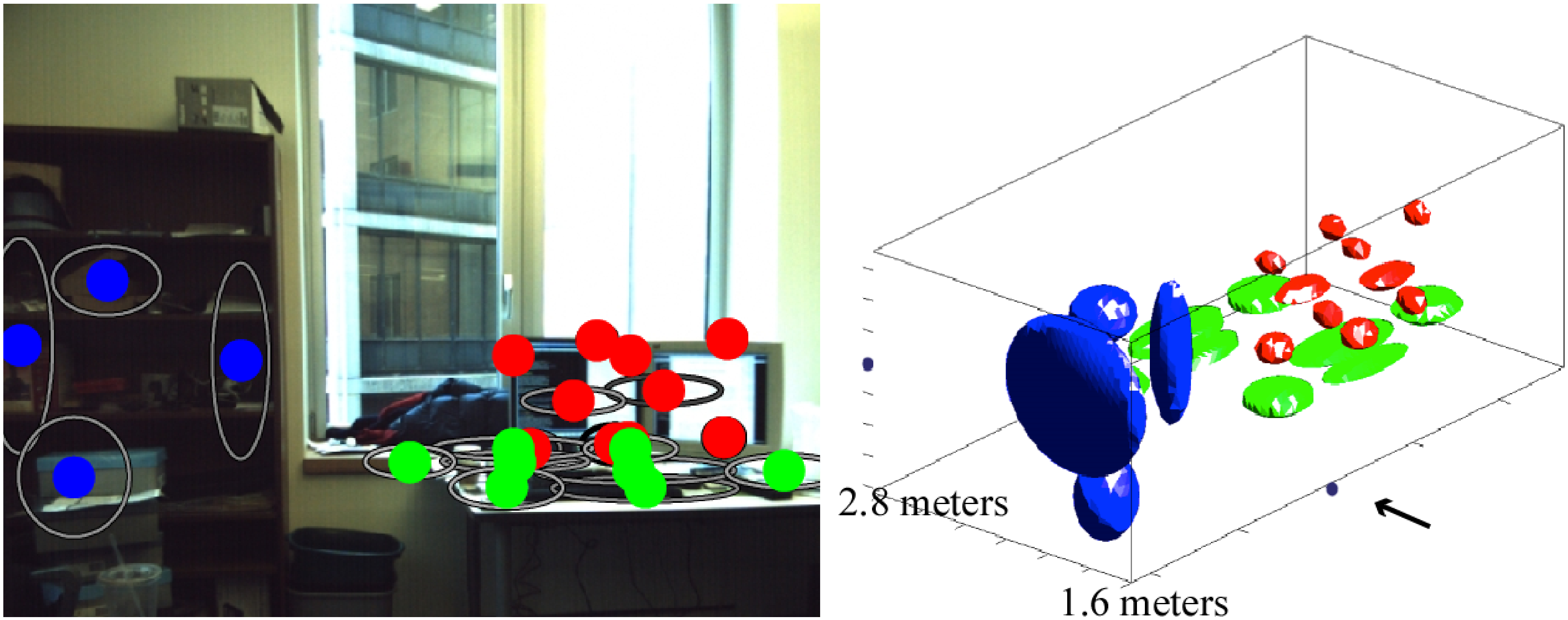






Background **Bookshelves** **Computer Screen** **Desk**

Stereo Test Image



Simultaneous *object recognition*
& coarse *3D reconstruction*