C280, Computer Vision

Prof. Trevor Darrell
trevor@eecs.berkeley.edu

Lecture 15: Part-based models

Last Lecture: Discriminative Kernels

- SVM-BOW
- Pyramid and Spatial-Pyramid match
- Fast Intersection Kernels
- Latent-part SVM models

Recognition Lectures Summary

- Tues. 10/13: Introduction to Recognition
  - Scanning window paradigm
  - GIST
  - HOG
  - Boosted Face Detection
  - Local-feature Alignment; from Roberts to Lowe...
  - BOW Indexing
- Thur. 10/15: Topic models for Recognition
  - Topic models for category discovery [Sivic05]
  - Category discovery from web [Fergus05]
  - Bootstrapping a category model [Li07]
  - Using text in addition to image [Berg06]
  - Learning objects from a dictionary [Saenko08]
- Tues. 10/20: Discriminative Kernels
  - SVM-BOW
  - Pyramid and Spatial-Pyramid match
  - Fast Intersection Kernels
  - Latent-part SVM models
- Thurs. 10/22: Voting and Part Based Models
  - Naïve-Bayes Nearest Neighbor [Irani]
  - Implicit Shape Model (ISM)
  - Constellation Models
  - Transformed LDA Models [Sudderth]
  - 3-D view models [Saravese]

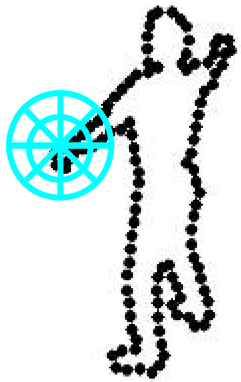
Today

- Naïve-Bayes Nearest Neighbor (Irani)
- ISM (Liebe)
- Constellation Models (Fergus)
- Transformed LDA Models (Sudderth)
- 3-D view models (Saravese)

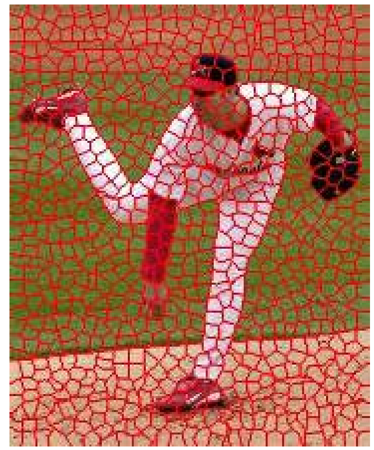
Multiple Features…

Wide variety of proposed local feature representations:

SIFT [Lowe]

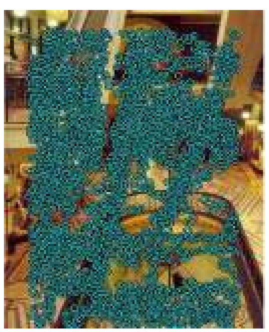


Shape context [Belongie et al.]



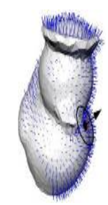
Superpixels [Ren et al.]

Maximally Stable Extremal Regions [Matas et al.]

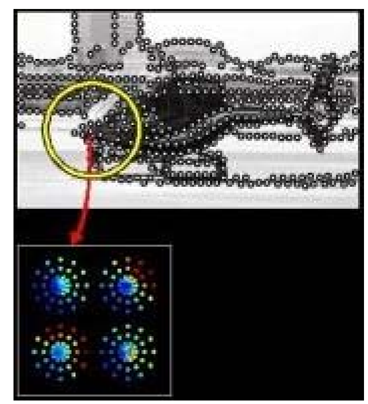


Salient regions [Kadir et al.]

Harris-Affine [Schmid et al.]



Spin images [Johnson and Hebert]



Geometric Blur [Berg et al.]

Discriminative Paradigm: Learning the Kernel

- Learn the kernel parameters
  - Improve accuracy and generalisation
  - Perform feature component selection
  - Perform dimensionality reduction
- Learn a linear combination of base kernels
  - $K(\mathbf{x}_i,\mathbf{x}_j) = \Sigma_k \, d_k \, K_k(\mathbf{x}_i,\mathbf{x}_j)$
  - Combine heterogeneous sources of data
  - Perform feature selection

[Varma]

Int J Comput Vis
DOI 10.1007/s11263-009-0268-3

Gaussian Processes for Object Categorization

Ashish Kapoor · Kristen Grauman · Raquel Urtasun · Trevor Darrell

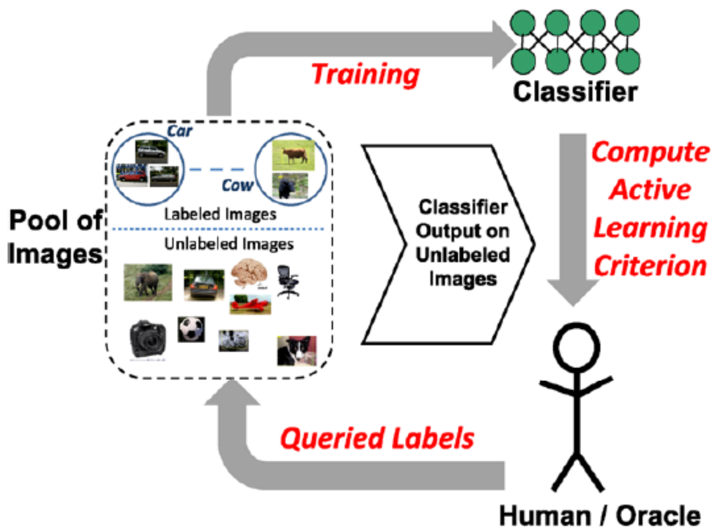


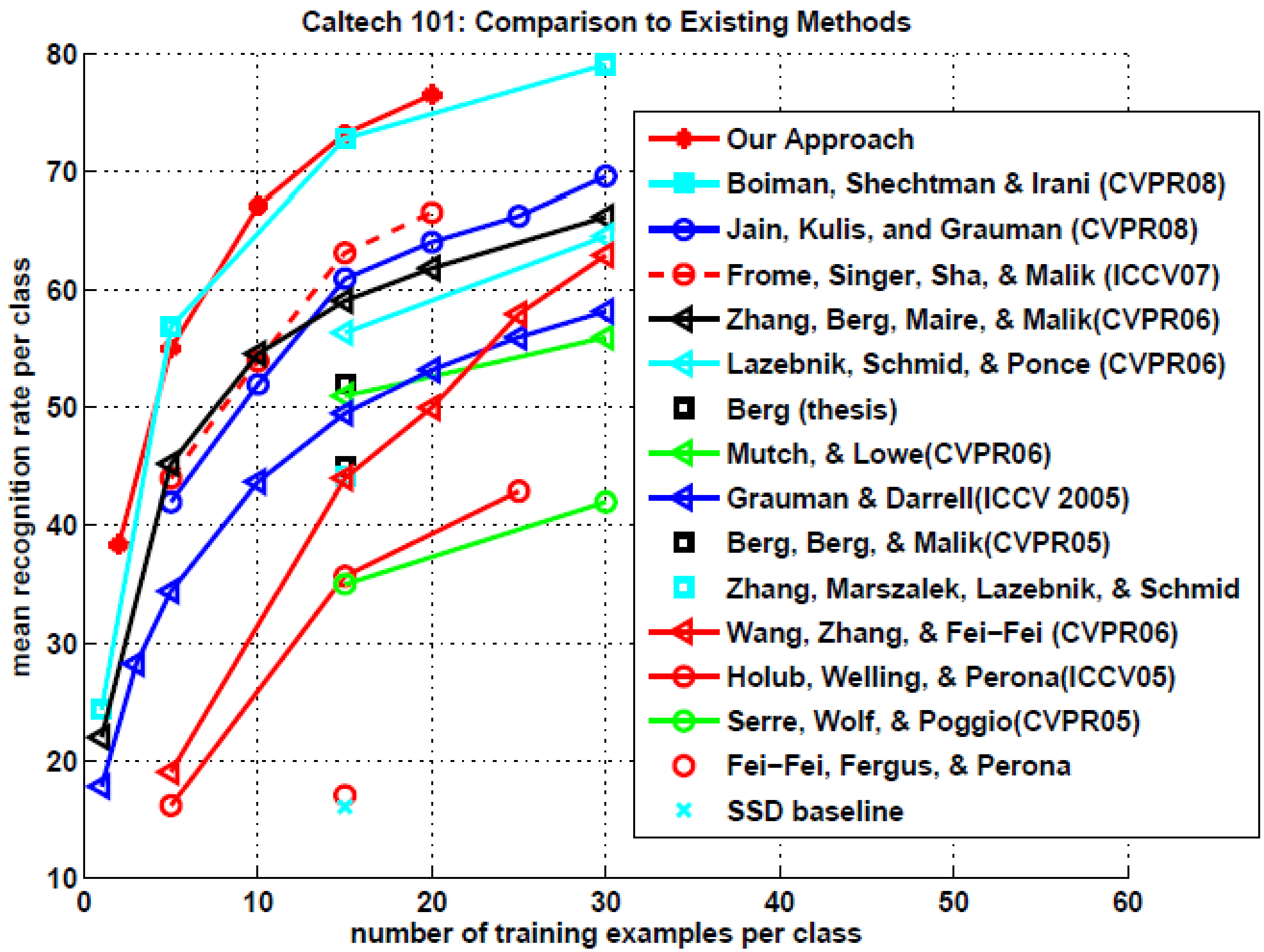
Fig. 1 The active learning framework. The goal of the system is to query labels for images that are most useful in training

$$\mathbf{K} = \sum_{i=1}^{k} \alpha_i \mathbf{K}^{(i)},$$

2. Kernel weights and kernel hyperparameters can be efficiently learned in a 1-vs-all setting (vs. SVM MKL)

1. GP uncertainty model facilitates active learning

The power of discriminative kernels?



In Defense of Nearest-Neighbor Based Image Classification

Oren Boiman
The Weizmann Institute of Science
Rehovot, ISRAEL

Eli Shechtman
Adobe Systems Inc. &
University of Washington

Michal Irani
The Weizmann Institute of Science
Rehovot, ISRAEL

Abstract

State-of-the-art image classification methods require an intensive learning/training stage (using SVM, Boosting, etc.) In contrast, non-parametric Nearest-Neighbor (NN) based image classifiers require no training time and have other favorable properties. However, the large performance gap between these two families of approaches rendered NN-based image classifiers useless.

We claim that the effectiveness of non-parametric NN-based image classification has been considerably under-valued. We argue that two practices commonly used in image classification methods, have led to the inferior performance of NN-based image classifiers: (i) Quantization of local image descriptors (used to generate "bags-of-words", codebooks). (ii) Computation of 'Image-to-Image' distance, instead of 'Image-to-Class' distance.

We propose a trivial NN-based classifier – NBNN, (Naive-Bayes Nearest-Neighbor), which employs NN-distances in the space of the local image descriptors (and not in the space of images). NBNN computes direct 'Image-to-Class' distances without descriptor quantization. We further show that under the Naive-Bayes assumption, the theoretically optimal image classifier can be accurately approximated by NBNN.

Although NBNN is extremely simple, efficient, and requires no learning/training phase, its performance ranks among the top leading learning-based image classifiers. Empirical comparisons are shown on several challenging databases (Caltech-101,Caltech-256 and Graz-01).

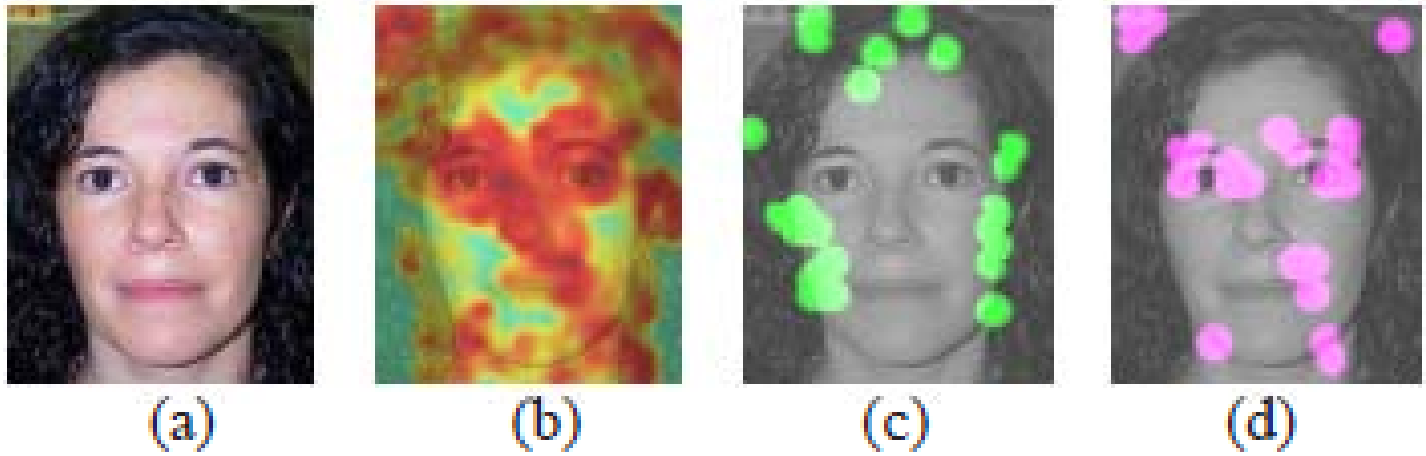


Figure 1. Effects of descriptor quantization – Informative descriptors have low database frequency, leading to high quantization error. *(a) An image from the Face class in Caltech101. (b) Quantization error of densely computed image descriptors (SIFT) using a large codebook (size 6,000) of Caltech-101 (generated using [14]). Red = high error; Blue = low error. The most informative descriptors (eye, nose, etc.) have the highest quantization error. (c) Green marks the 8% of the descriptors in the image that are most frequent in the database (simple edges). (d) Magenta marks the 8% of the descriptors in the image that are least frequent in the database (mostly facial features).*

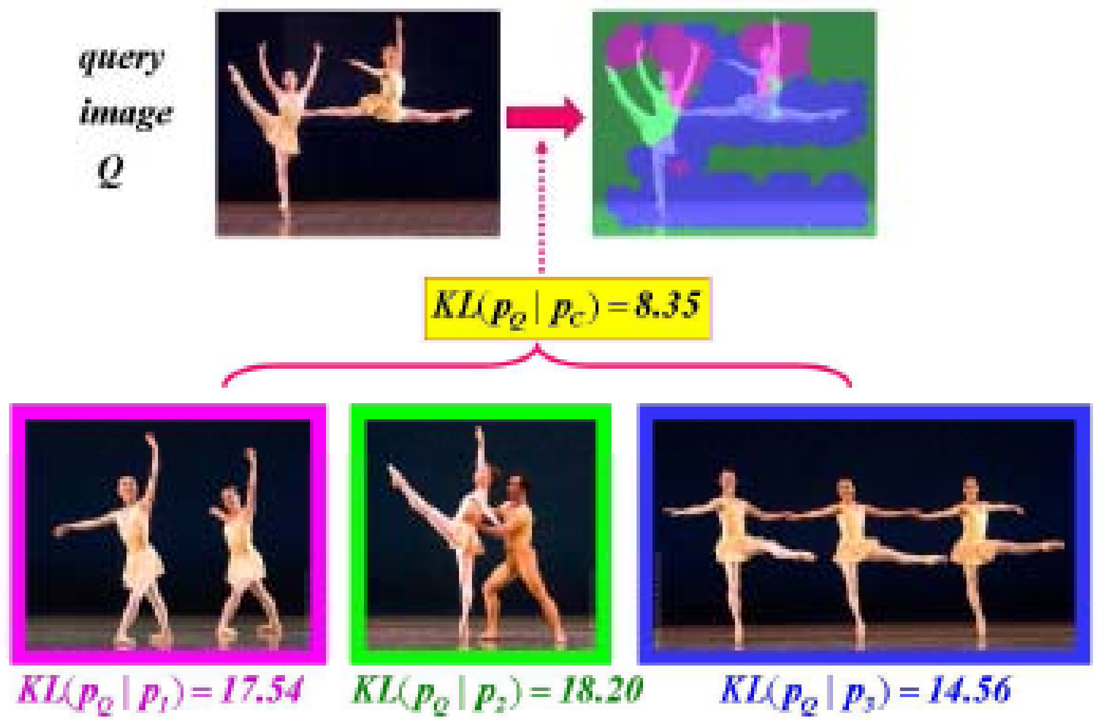


Figure 3. **"Image-to-Image" vs. "Image-to-Class" distance.** *A Ballet class with large variability and small number (three) of 'labelled' images (bottom row). Even though the "Query-to-Image" distance is large to each individual 'labelled' image, the "Query-to-Class" distance is small.* **Top right image:** *For each descriptor at each point in Q we show (in color) the 'labelled' image which gave it the highest descriptor likelihood. It is evident that the new query configuration is more likely given the three images, than each individual image seperately. (Images taken from [4].)*

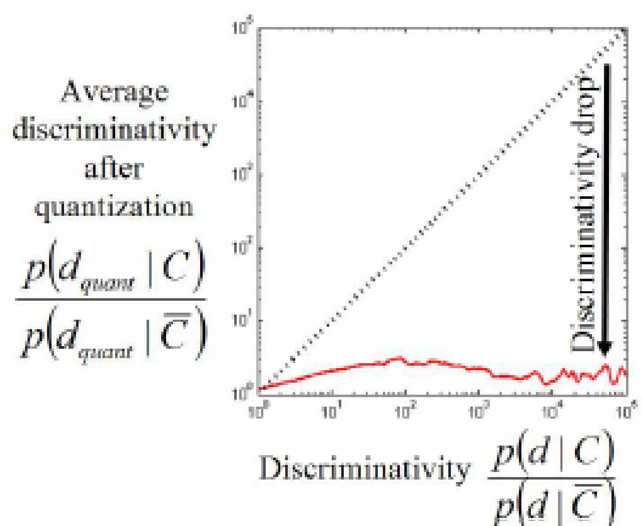


Figure 2. **Effects of descriptor quantization – Severe drop in descriptor discriminative power.** *We generated a scatter plot of descriptor discriminative power before and after quantization (for a very large sample set of SIFT descriptors* d *in Caltech-101, each for its respective class* C*). We then averaged this scatter plot along the y-axis. This yields the "Average discriminative power after quantization" (the RED graph). The display is in logarithmic scale in both axes. NOTE: The more informative (discriminative) a descriptor* d *is, the larger the drop in its discriminative power.*

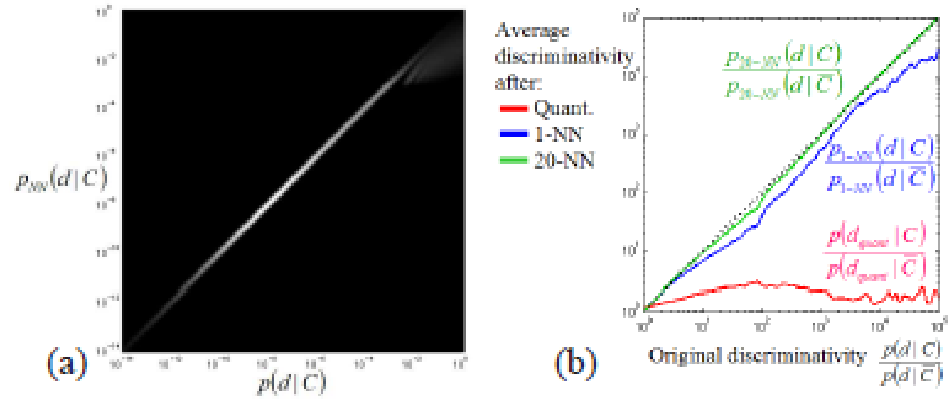


Figure 4. **NN descriptor estimation preserves descriptor density distribution and discriminativity.** **(a)** *A scatter plot of the 1-NN probability density distribution* $p_{NN}(d|C)$ *vs. the true distribution* $p(d|C)$*. Brightness corresponds to the concentration of points in the scatter plot. The plot shows that 1-NN distribution provides a very accurate approximation of the true distribution.* **(b)** *20-NN descriptor approximation (Green graph) and 1-NN descriptor approximation (Blue graph) preserve quite well the discriminative power of descriptors. In contrast, descriptor quantization (Red graph) severely reduces discriminative power of descriptors.* **Displays are in logarithmic scale in all axes.**

NBNN

The NBNN Algorithm:

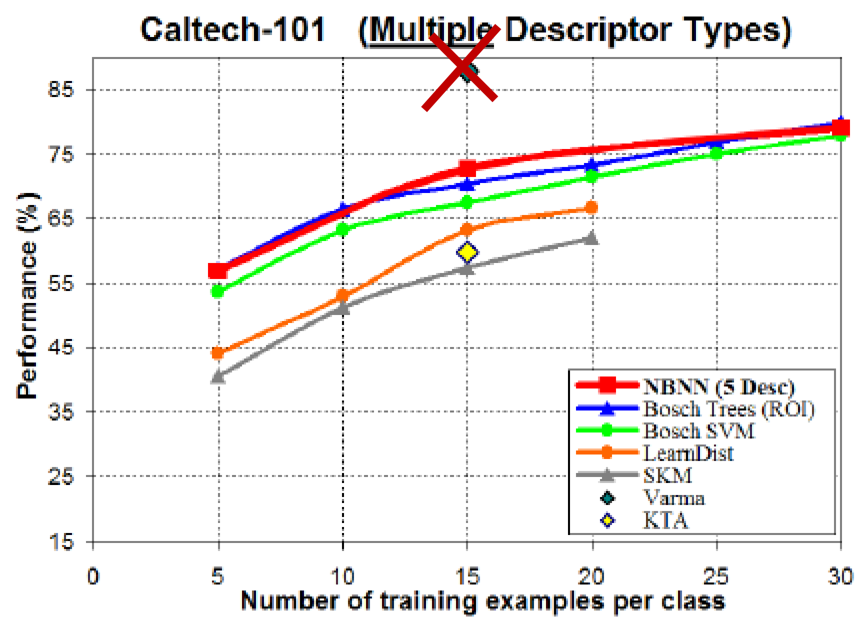
1. Compute descriptors $d_1, \ldots, d_n$ of the query image Q.
2. $\forall d_i \; \forall C$ compute the NN of d_i in C: $\mathrm{NN}_C(d_i)$.
3. $\hat{C} = arg \min_C \sum_{i=1}^{n} \| d_i - \mathrm{NN}_C(d_i) \|^2$.

with multiple feature types:

$$\hat{C} = arg \min_C \sum_{j=1}^{t} w_j \cdot \sum_{i=1}^{n} \| d_i^j - \mathrm{NN}_C(d_i^j) \|^2,$$

NN-based method	Performance
SPM NN Image [27]	42.1 ± 0.81%
GBDist NN Image [27]	45.2 ± 0.96%
GB Vote NN [3]	52%
SVM-KNN [30]	59.1 ± 0.56%
NBNN (1 Desc)	**65.0 ± 1.14%**
NBNN (5 Desc)	**72.8 ± 0.39%**

Table 1. *Comparing the performance of non-parametric NN-based approaches on the Caltech-101 dataset (n_{label} = 15). All the listed methods do not require a learning phase.*



Bosch Kernels used in original Varma paper have been withdrawn…

Back to shape: Parts-based Representation

- Object as set of parts
  - Generative representation

- Model:
  - Relative locations between parts
  - Appearance of part

- Issues:
  - How to model location
  - How to represent appearance
  - Sparse or dense (pixels or regions)
  - How to handle occlusion/clutter

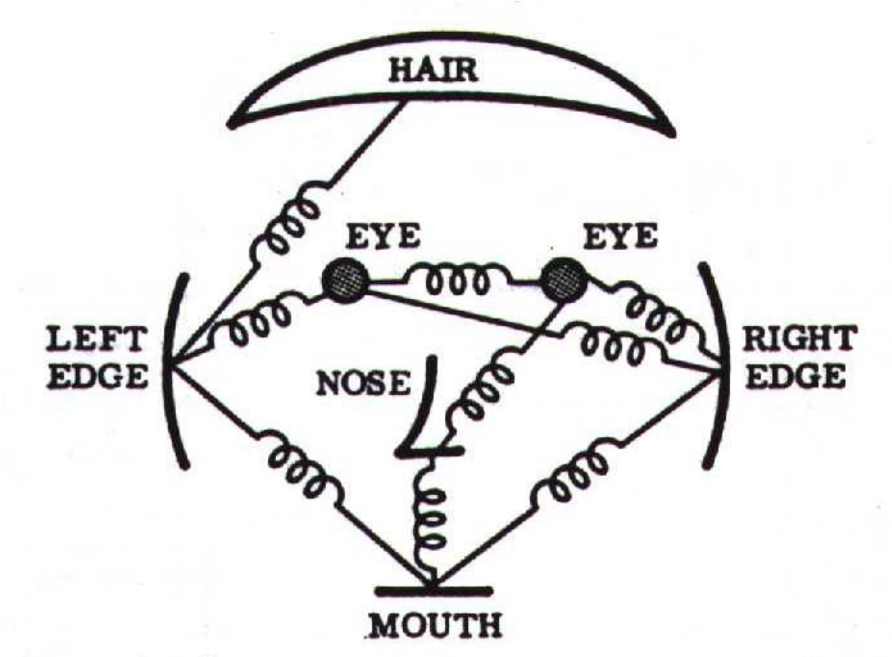
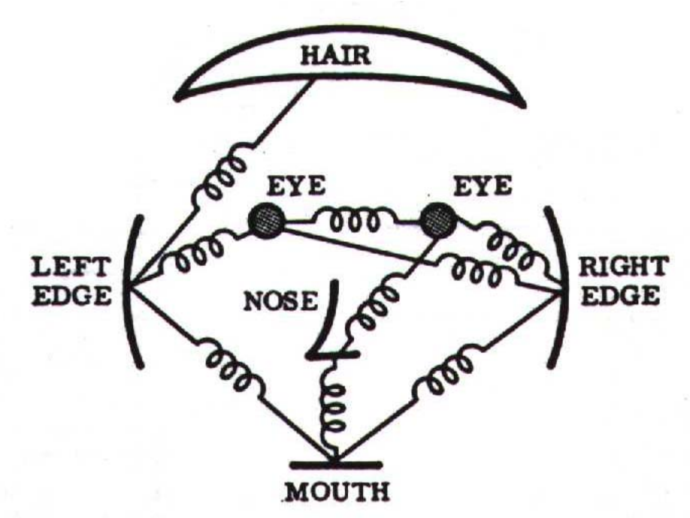


Figure from [Fischler & Elschlager 73]

Slide credit: Fergus

History of Parts and Structure approaches

- Fischler & Elschlager 1973



- Yuille '91
- Brunelli & Poggio '93
- Lades, v.d. Malsburg et al. '93
- Cootes, Lanitis, Taylor et al. '95
- Amit & Geman '95, '99
- Perona et al. '95, '96, '98, '00, '03, '04, '05
- Felzenszwalb & Huttenlocher '00, '04
- Crandall & Huttenlocher '05, '06
- Leibe & Schiele '03, '04

- Many papers since 2000

Slide credit: Fergus

Object class recognition using unsupervised scale-invariant learning

Rob Fergus
Pietro Perona
Andrew Zisserman

Oxford University
California Institute of Technology

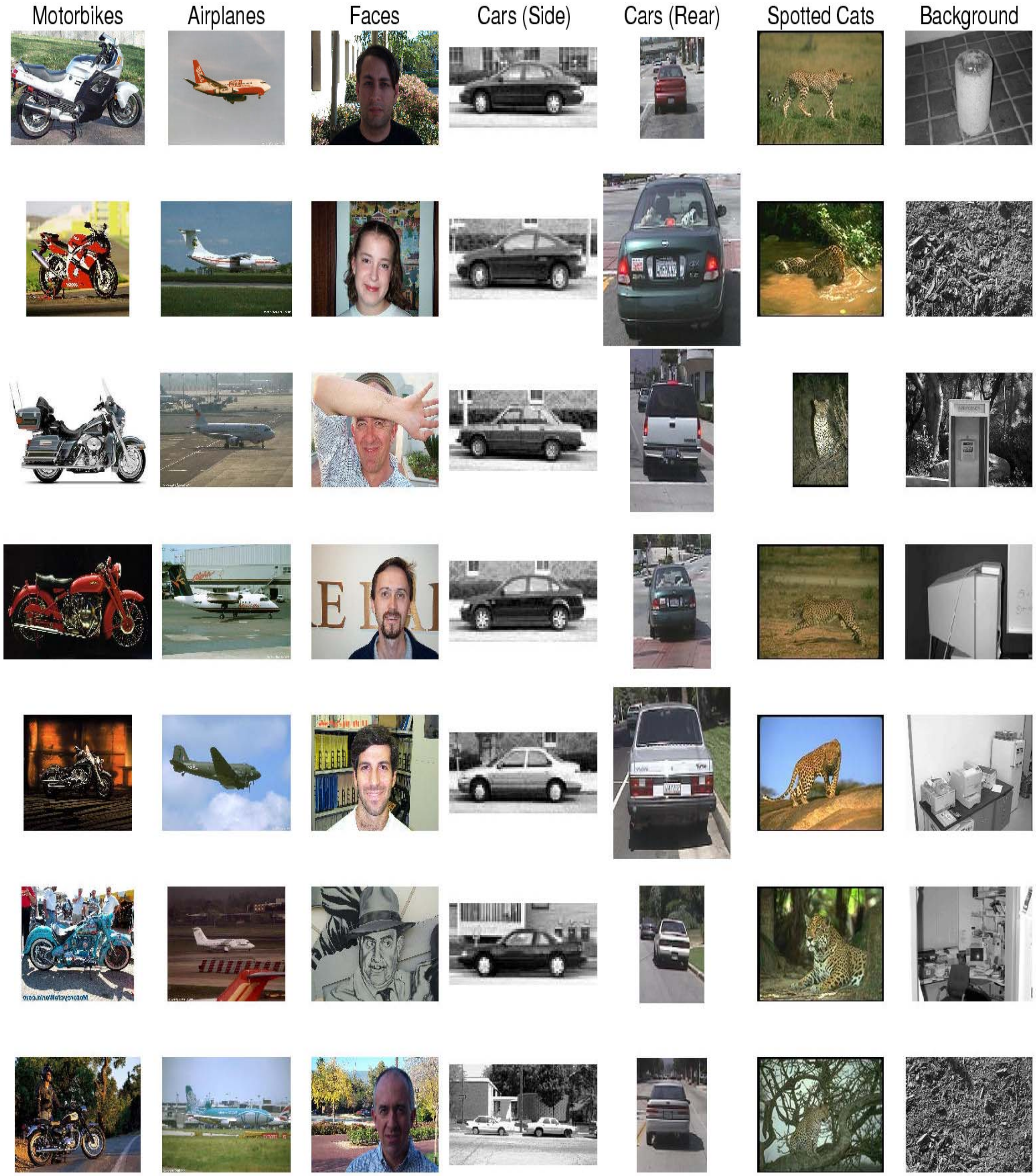
Slide credit: Fergus

Goal

- Recognition of object categories
- Unassisted learning

Slide credit: Fergus

Some object categories



Learn from examples

Difficulties:

- Size variation
- Background clutter
- Occlusion
- Intra-class variation

Main issues

- **Representation**
- Learning
- Recognition

Slide credit: Fergus

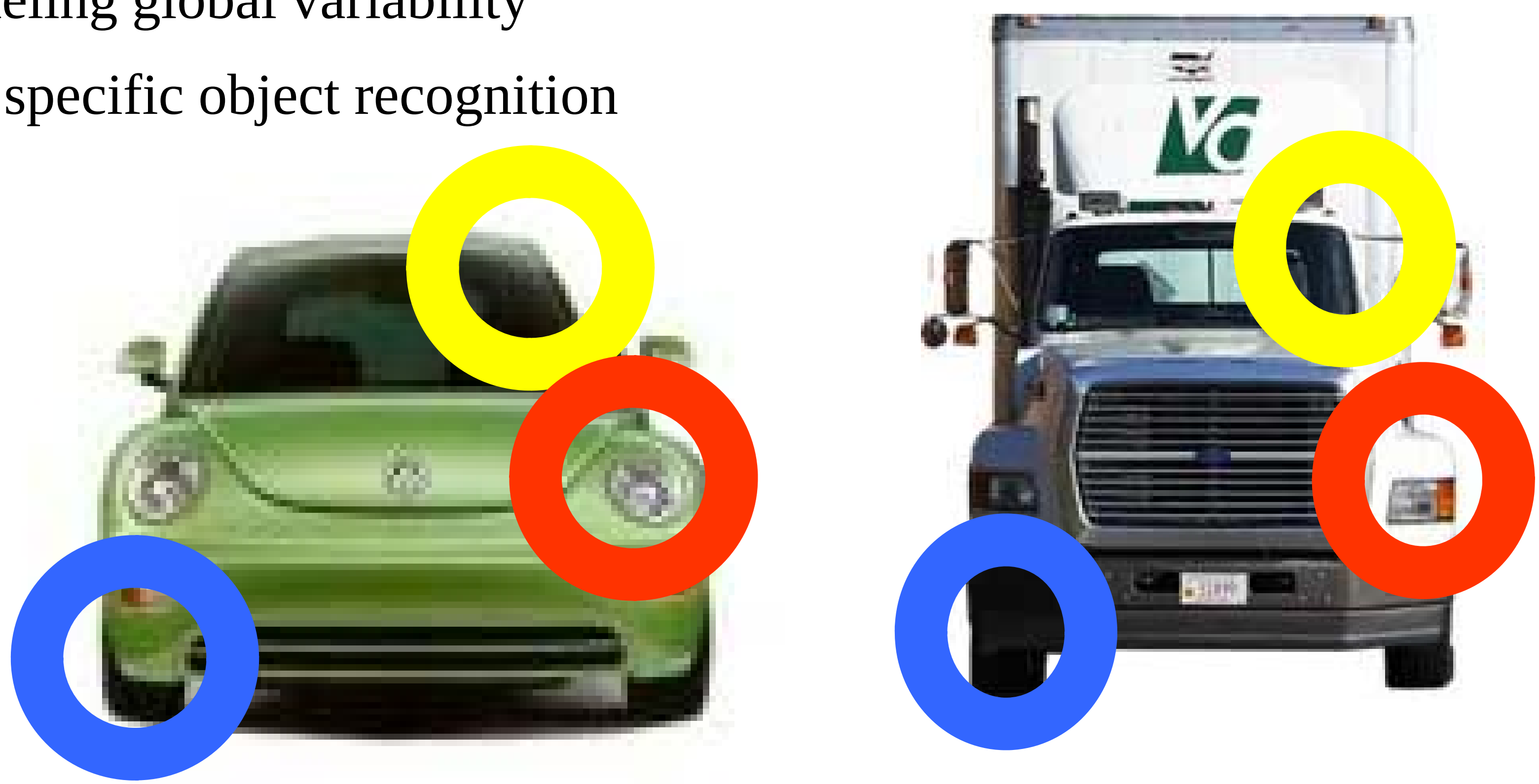
Sparse representation

+ Computationally tractable (10^5 pixels → 10^1 -- 10^2 parts)

+ Generative representation of class

+ Avoid modeling global variability

+ Success in specific object recognition

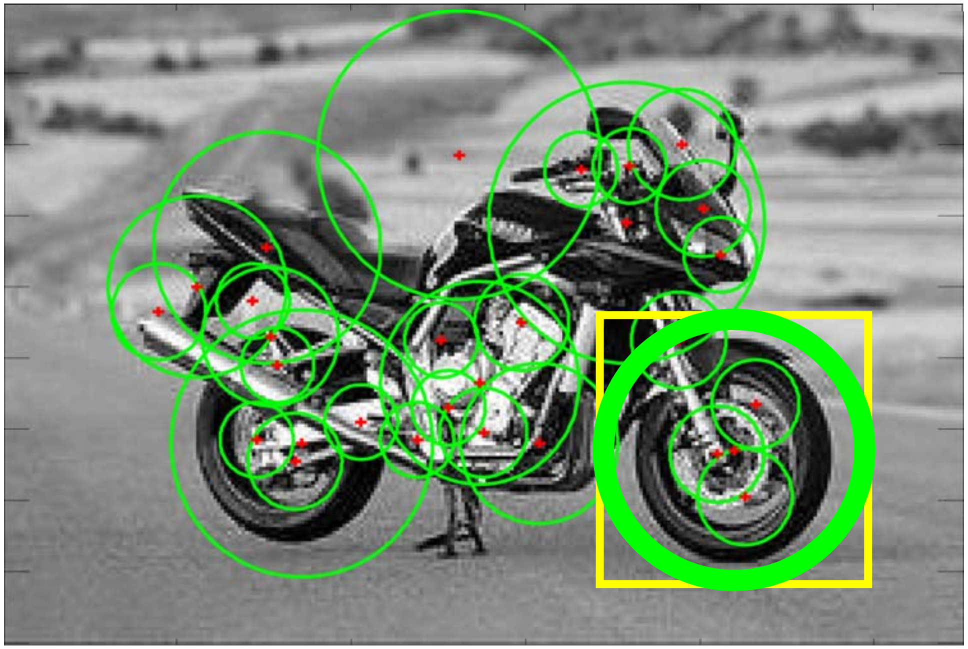


- Throw away most image information

- Parts need to be distinctive to separate from other classes

Slide credit: Fergus

Detection & Representation of regions



- Find regions within image
- Use Kadir and Brady's salient region operator [IJCV '01]

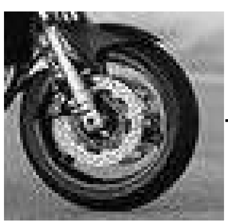
Location

(x,y) coords. of region center

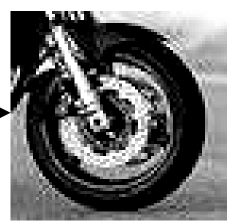
Scale

Diameter of region (pixels)

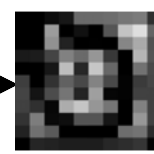
Appearance



Normalize →



11x11 patch →



Projection onto PCA basis →

$$\begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_{15} \end{pmatrix}$$

Gives representation of appearance in low-dimensional vector space

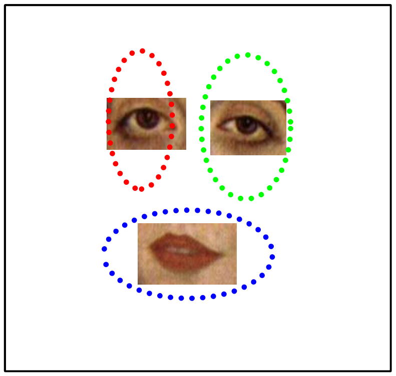
Slide credit: Fergus

Generative probabilistic model

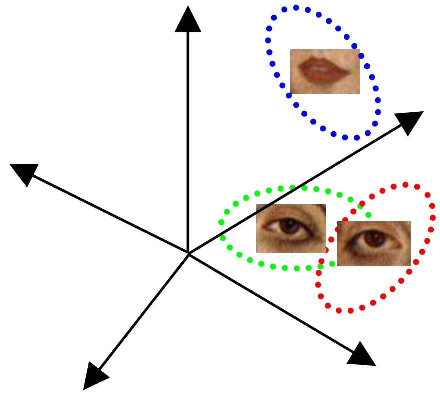
Foreground model

based on Burl, Weber et al. [ECCV '98, '00]

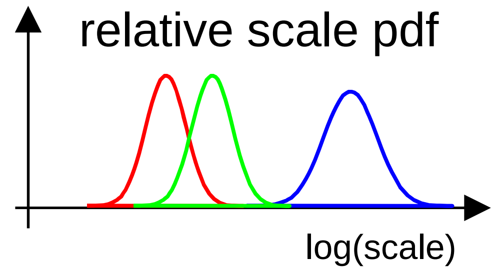
Gaussian shape pdf



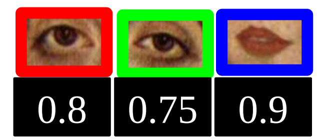
Gaussian part appearance pdf



Gaussian relative scale pdf

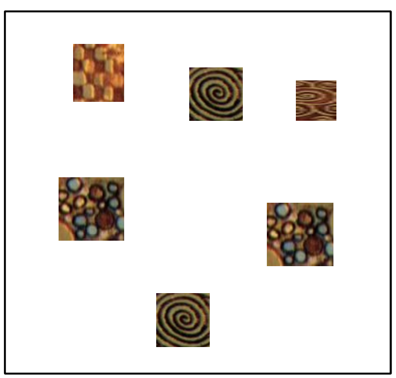


Prob. of detection

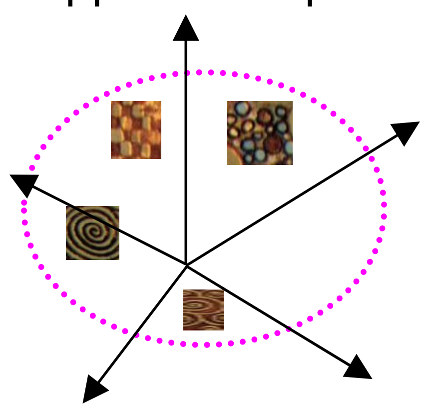


Clutter model

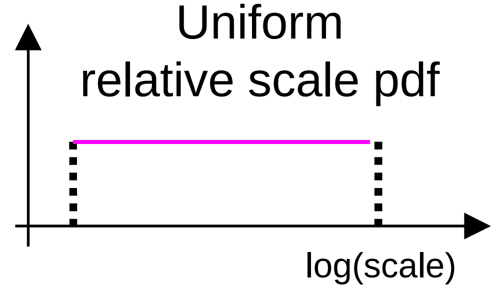
Uniform shape pdf



Gaussian background appearance pdf



Uniform relative scale pdf



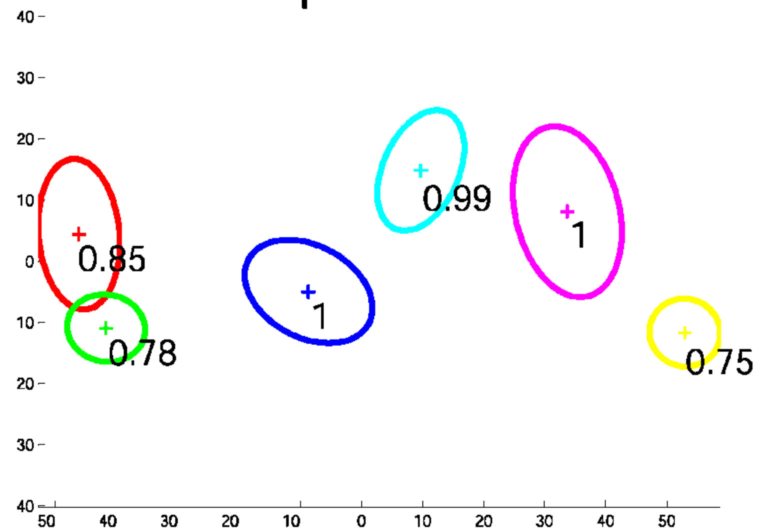
Poission pdf on # detections

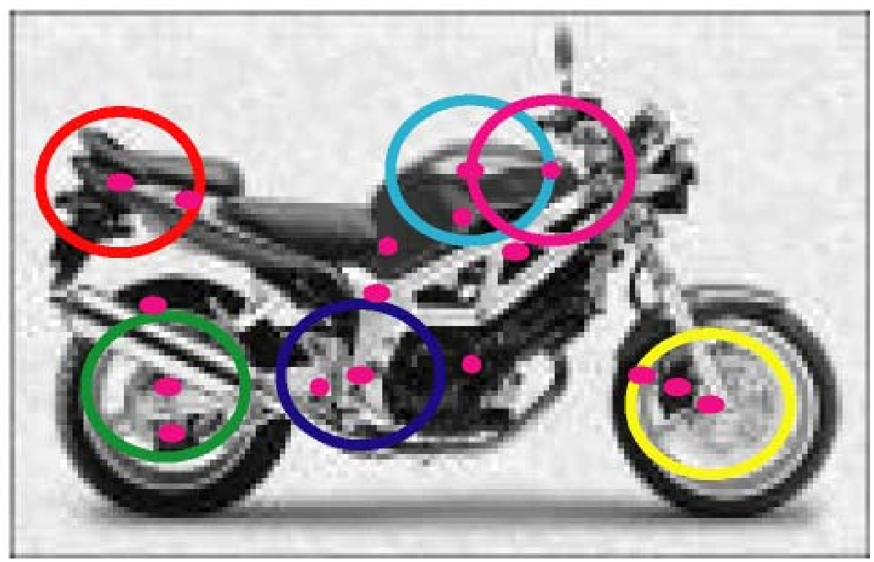
Slide credit: Fergus

Motorbikes

Samples from appearance model

Shape model

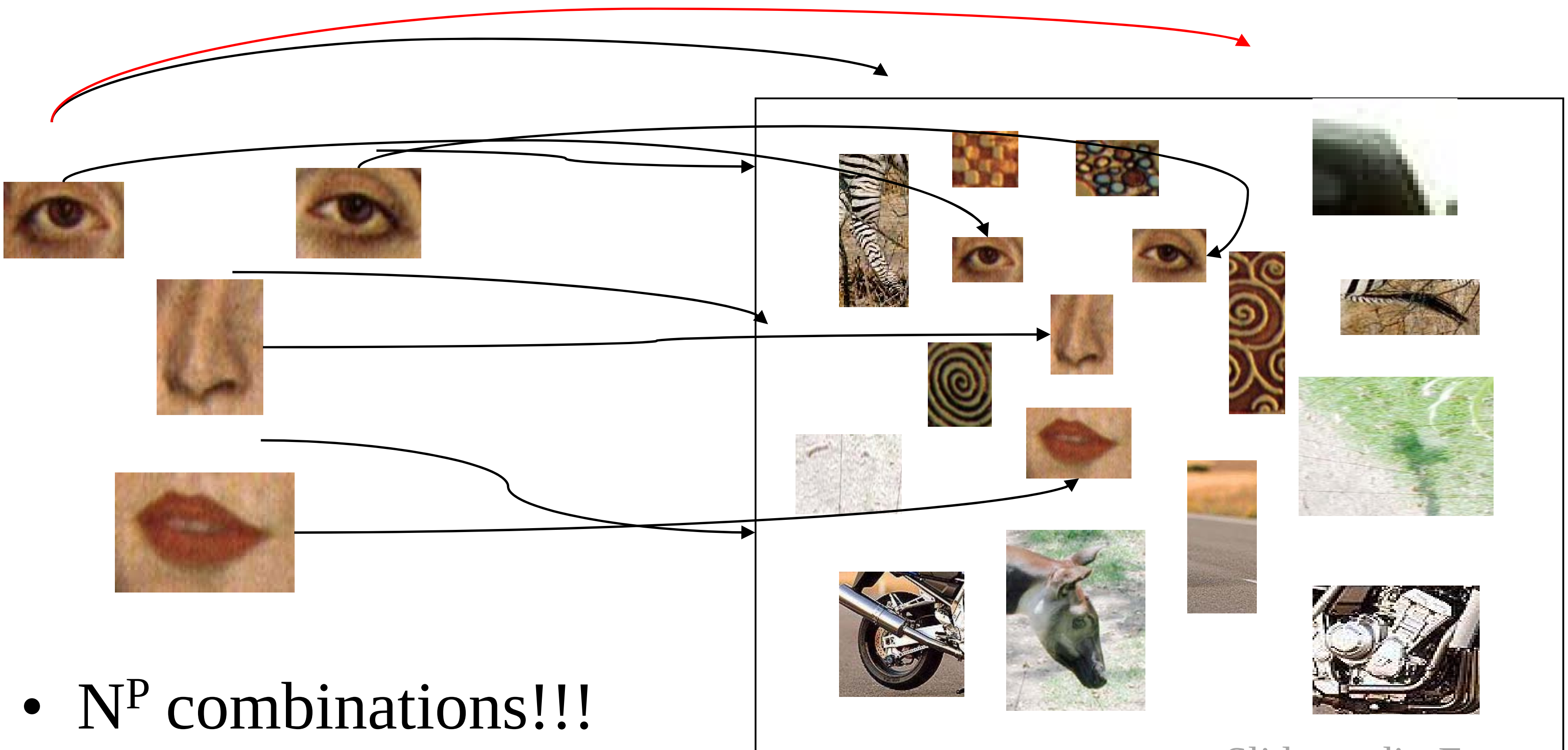




Slide credit: Fergus

The correspondence problem

- Model with P parts
- Image with N possible assignments for each part
- Consider mapping to be 1-1



- N^P combinations!!!

Slide credit: Fergus

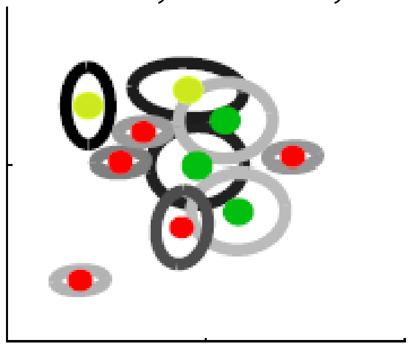
The correspondence problem

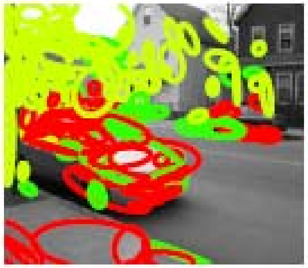
- 1 – 1 mapping
  - Each part assigned to unique feature

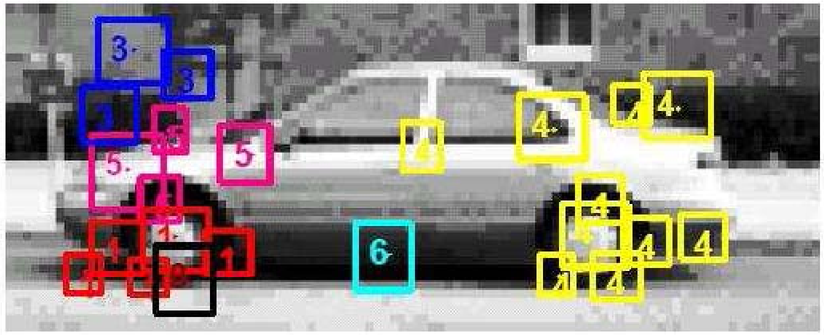
As opposed to:

- 1 – Many
  - Bag of words approaches
  - Sudderth, Torralba, Freeman '05
  - Loeff, Sorokin, Arora and Forsyth '05

- Many – 1
  - Quattoni, Collins and Darrell, 04







Learning

- Task: Estimation of model parameters
- Chicken and Egg type problem, since we initially know neither:
  - Model parameters
  - Assignment of regions to foreground / background
- Let the assignments be a hidden variable and use EM algorithm to learn them and the model parameters

Slide credit: Fergus

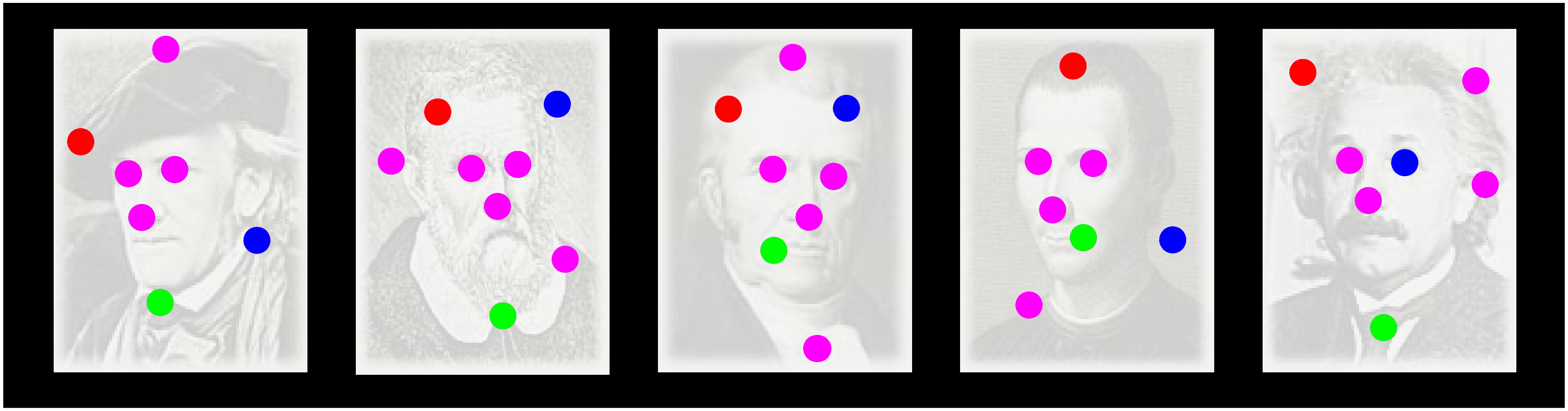
Learning procedure

- Find regions & their location, scale & appearance
- Initialize model parameters
- Use EM and iterate to convergence:

  E-step: Compute assignments for which regions are foreground / background

  M-step: Update model parameters

- Trying to maximize likelihood – consistency in shape & appearance



Slide credit: Fergus

Experimental procedure

Two series of experiments:

- Fixed-scale model - Objects the same size (manual normalization)
- Scale-invariant model - Objects between 100 and 550 pixels in width

Datasets

Training
- 50% images
- No identifcation of object within image

Motorbikes

Airplanes

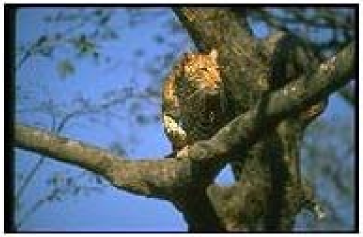
Frontal Faces

Testing
- 50% images
- Simple object present/absent test

Cars (Side)

Cars (Rear)

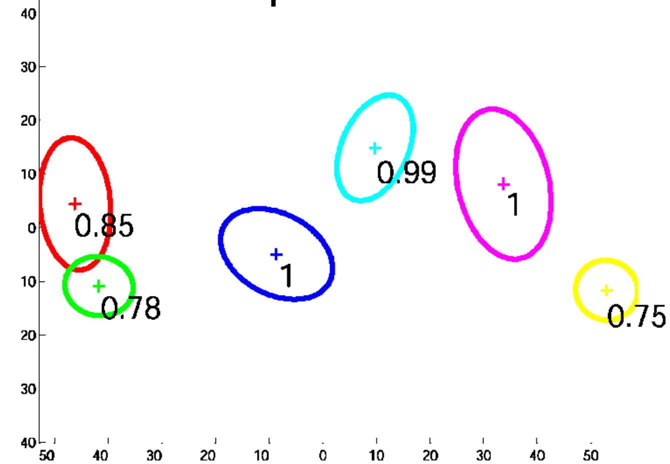
Spotted cats

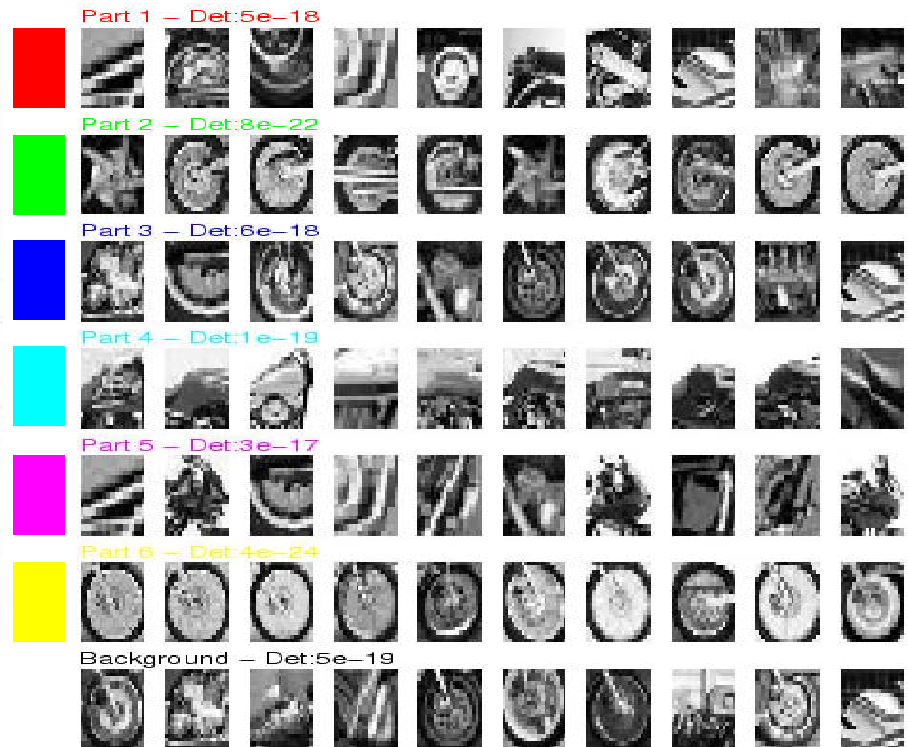


Slide credit: Fergus

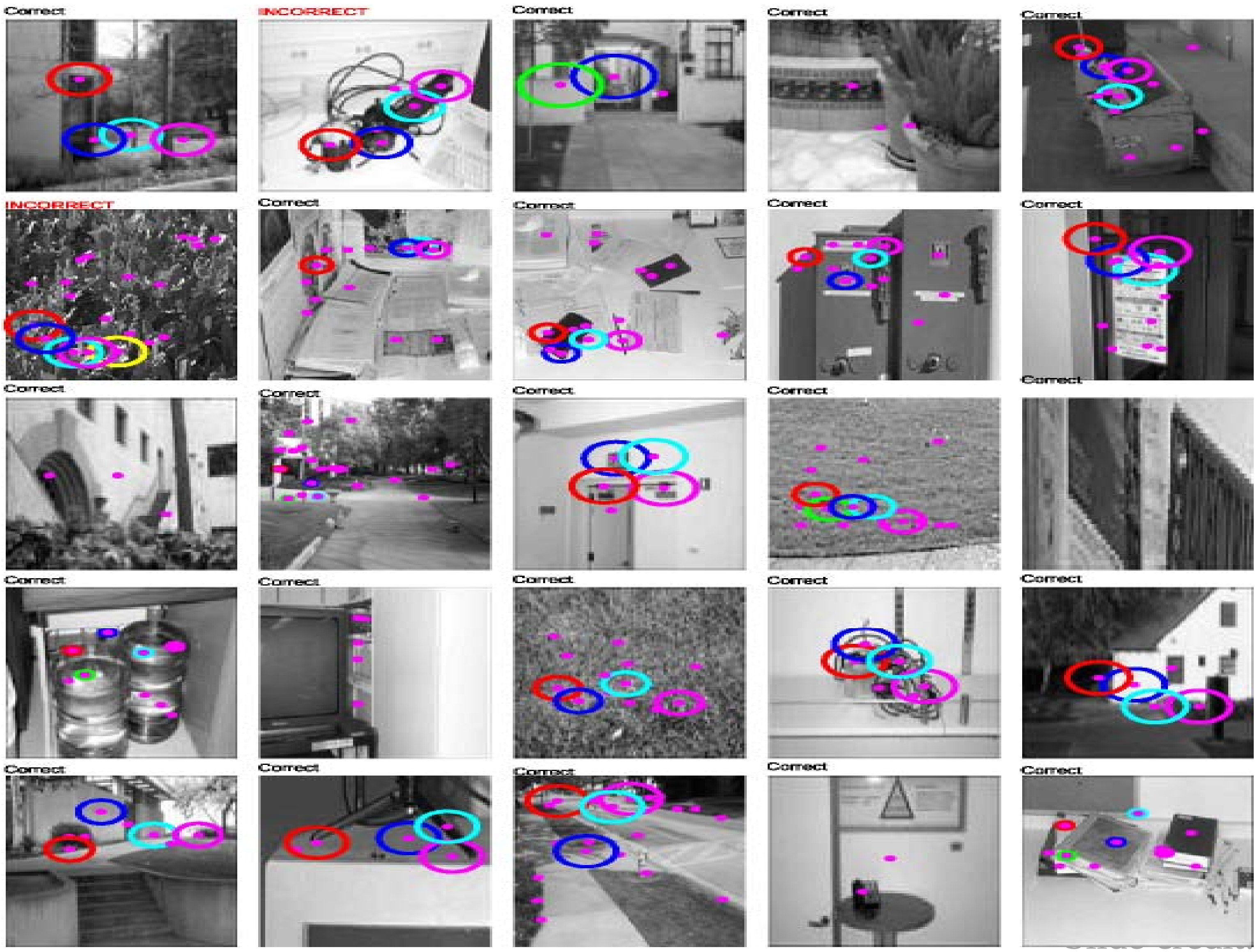
Motorbikes

Shape model



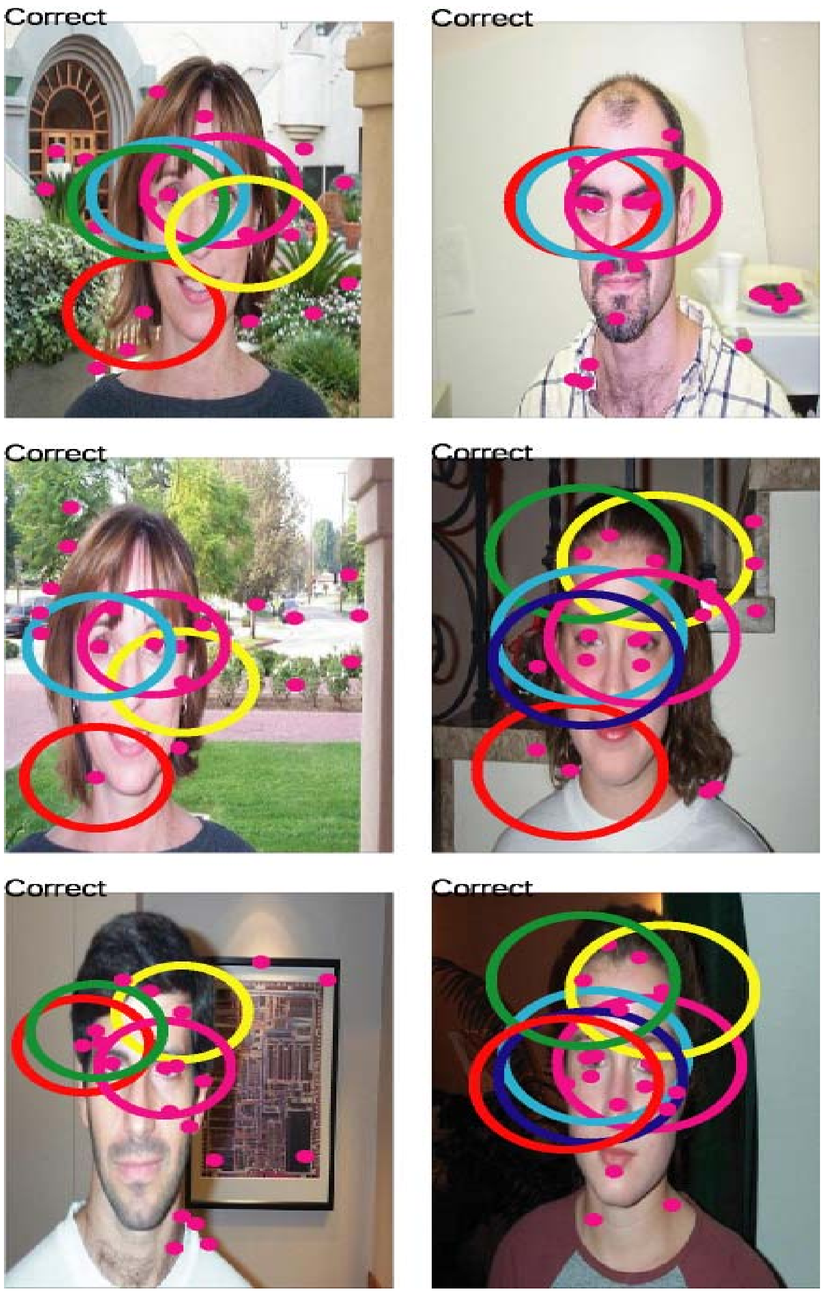


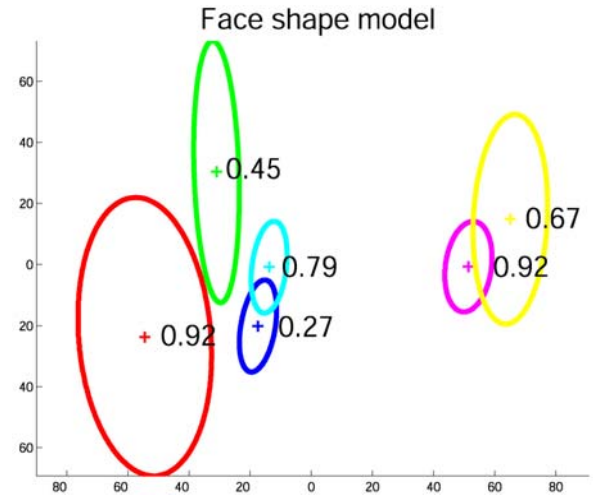
Background images evaluated with motorbike model

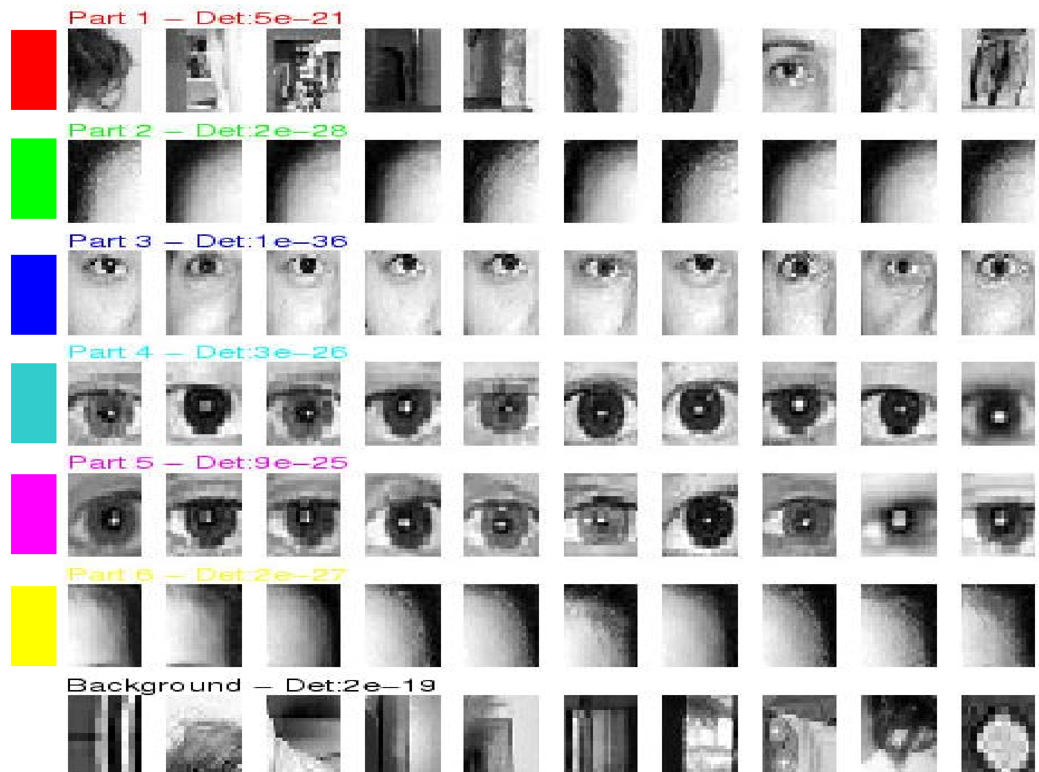


Fergus

Frontal faces

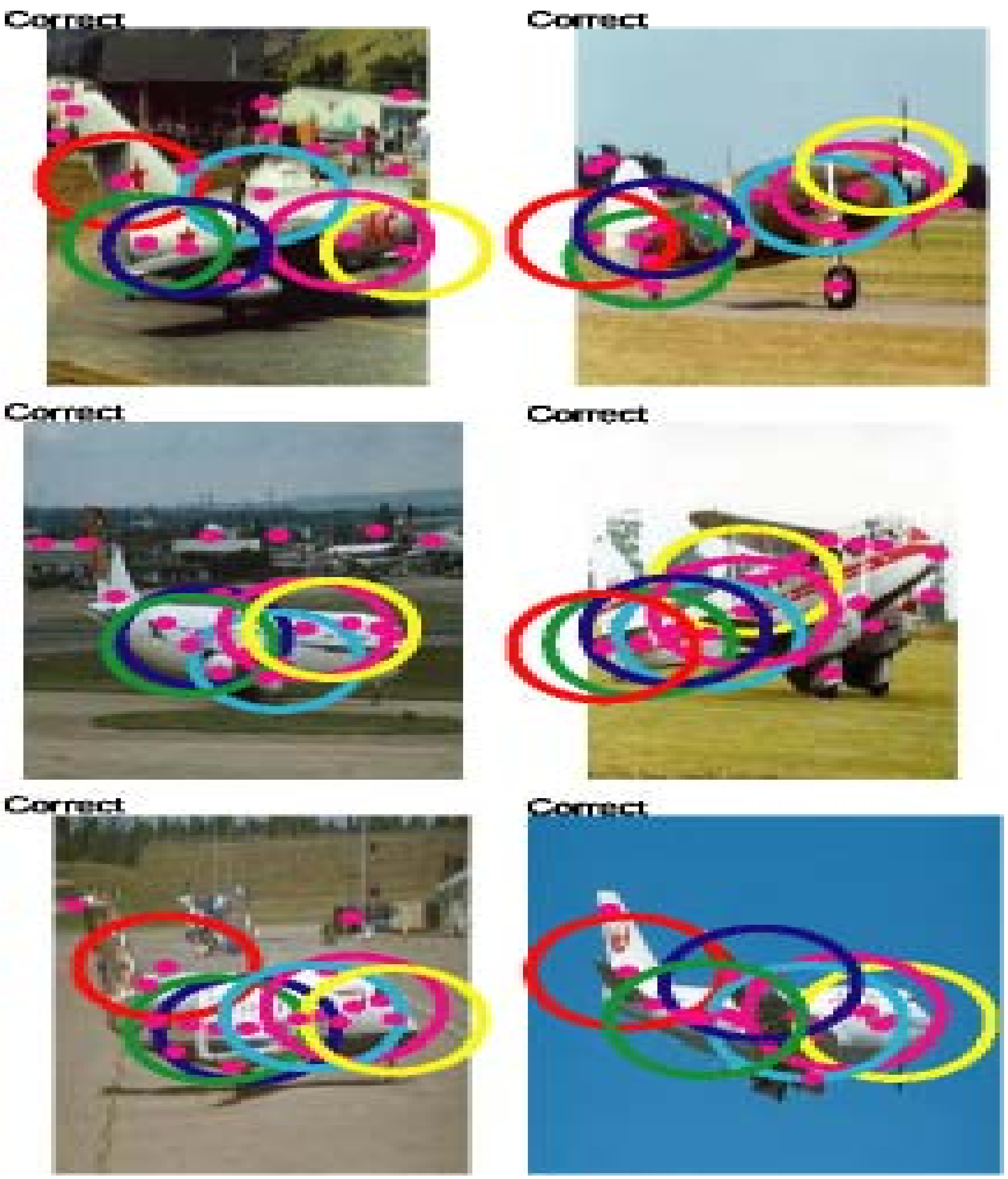


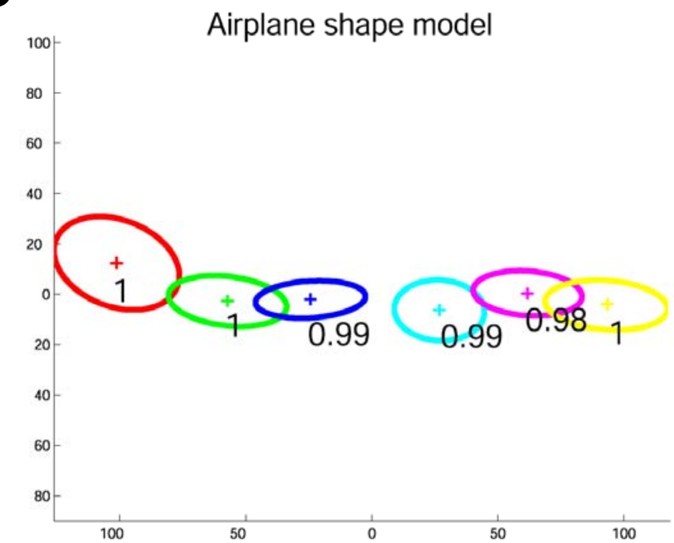




Slide credit: Fergus

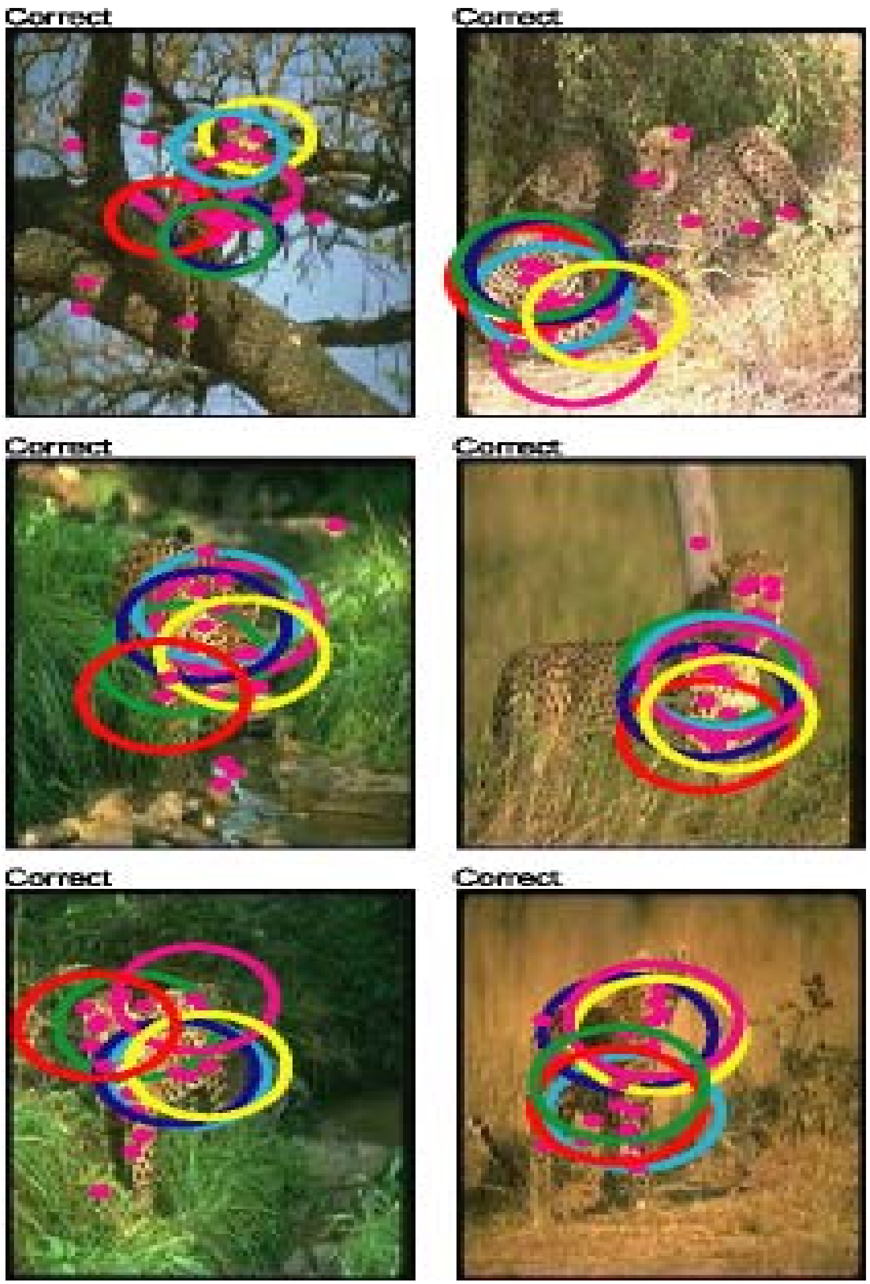
Airplanes

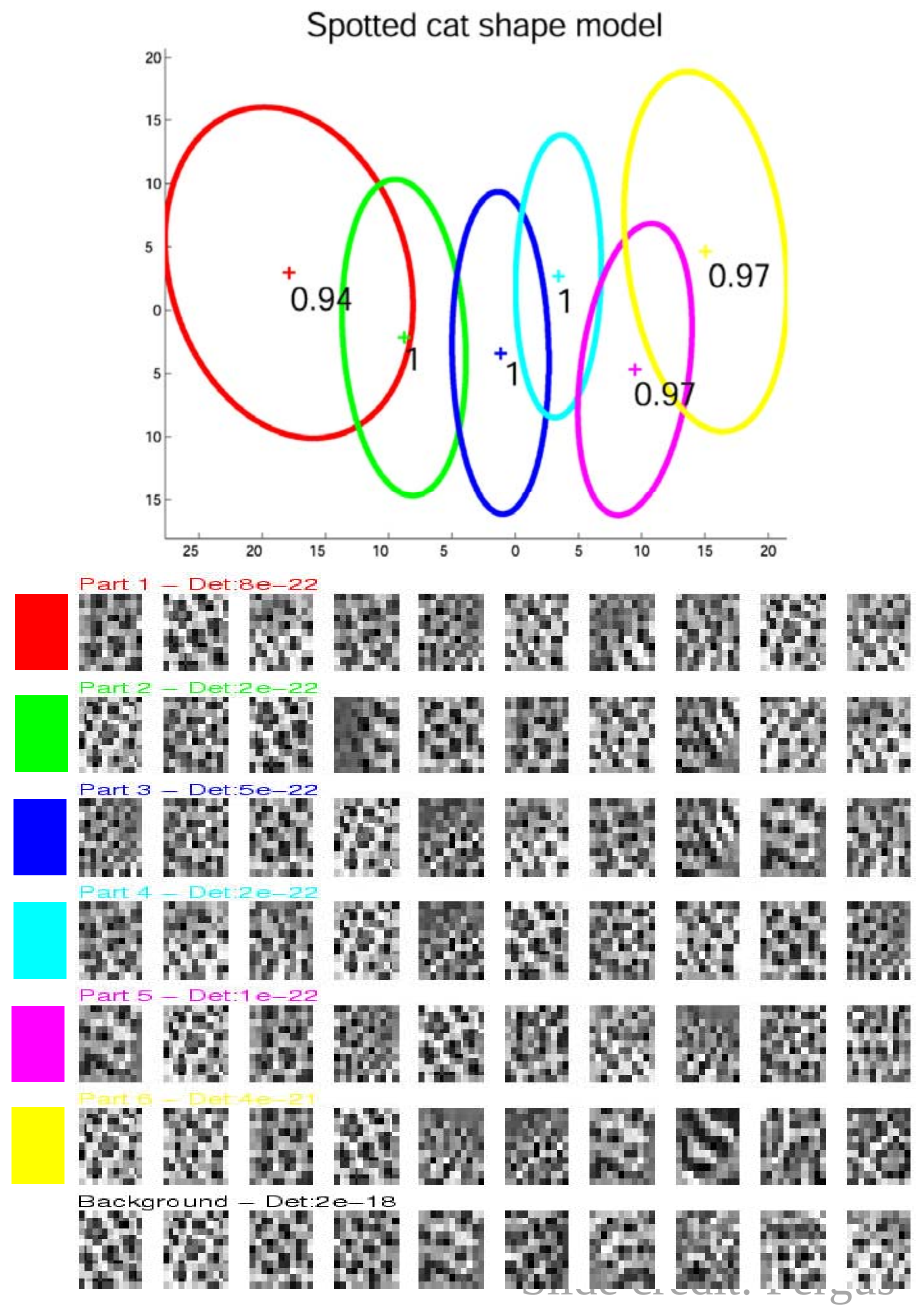

Correct
Correct
Correct
Correct
Correct
Correct


Airplane shape model
100
80
60
40
20
0
20
40
60
80
100
50
0
50
100
1
1
0.99
0.99
0.98
1


Part 1 – Det:3e–19
Part 2 – Det:9e–22
Part 3 – Det:1e–23
Part 4 – Det:2e–22
Part 5 – Det:7e–24
Part 6 – Det:5e–22
Background – Det:1e–20

Spotted cats





Slide credit: Fergus

Summary of results

Dataset	Fixed scale experiment	Scale invariant experiment
Motorbikes	7.5	6.7
Faces	4.6	4.6
Airplanes	9.8	7.0
Cars (Rear)	15.2	9.7
Spotted cats	10.0	10.0

% equal error rate

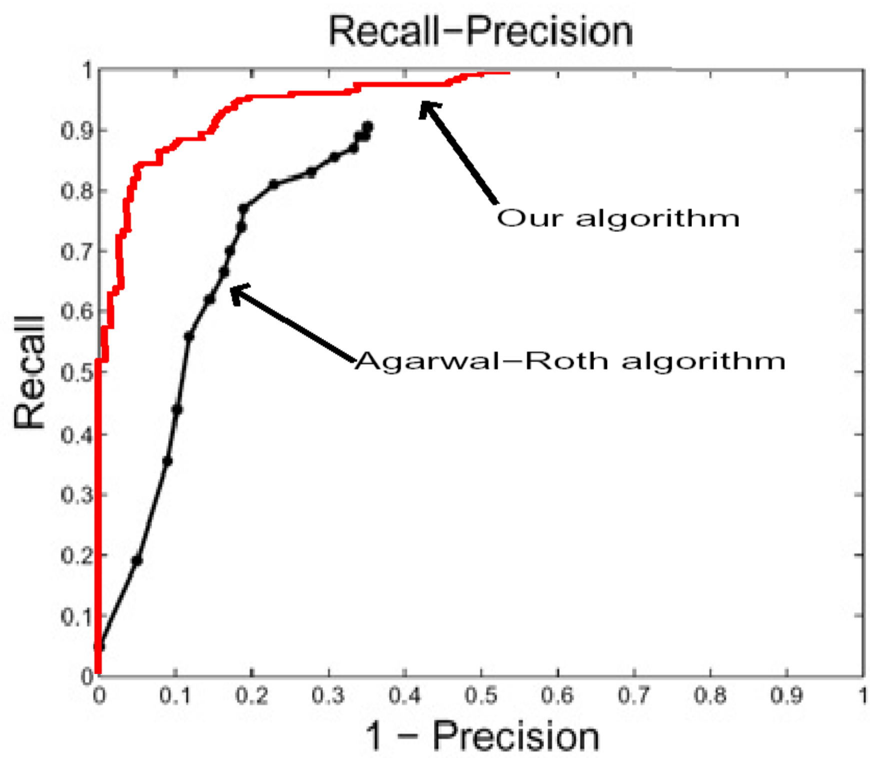
Note: Within each series, same settings used for all datasets

Slide credit: Fergus

Comparison to other methods

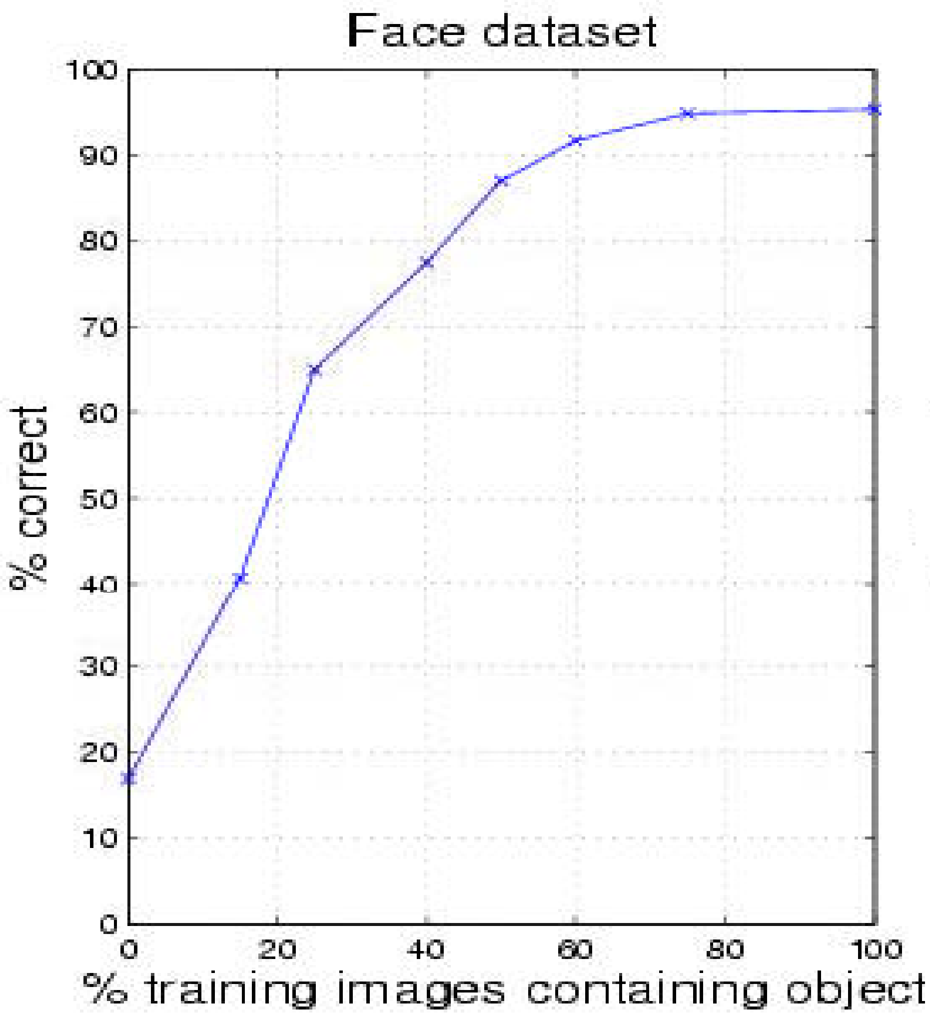
Dataset	Ours	Others	
Motorbikes	7.5	16.0	Weber et al. [ECCV '00]
Faces	4.6	6.0	Weber
Airplanes	9.8	32.0	Weber
Cars (Side)	11.5	21.0	Agarwal Roth [ECCV '02]

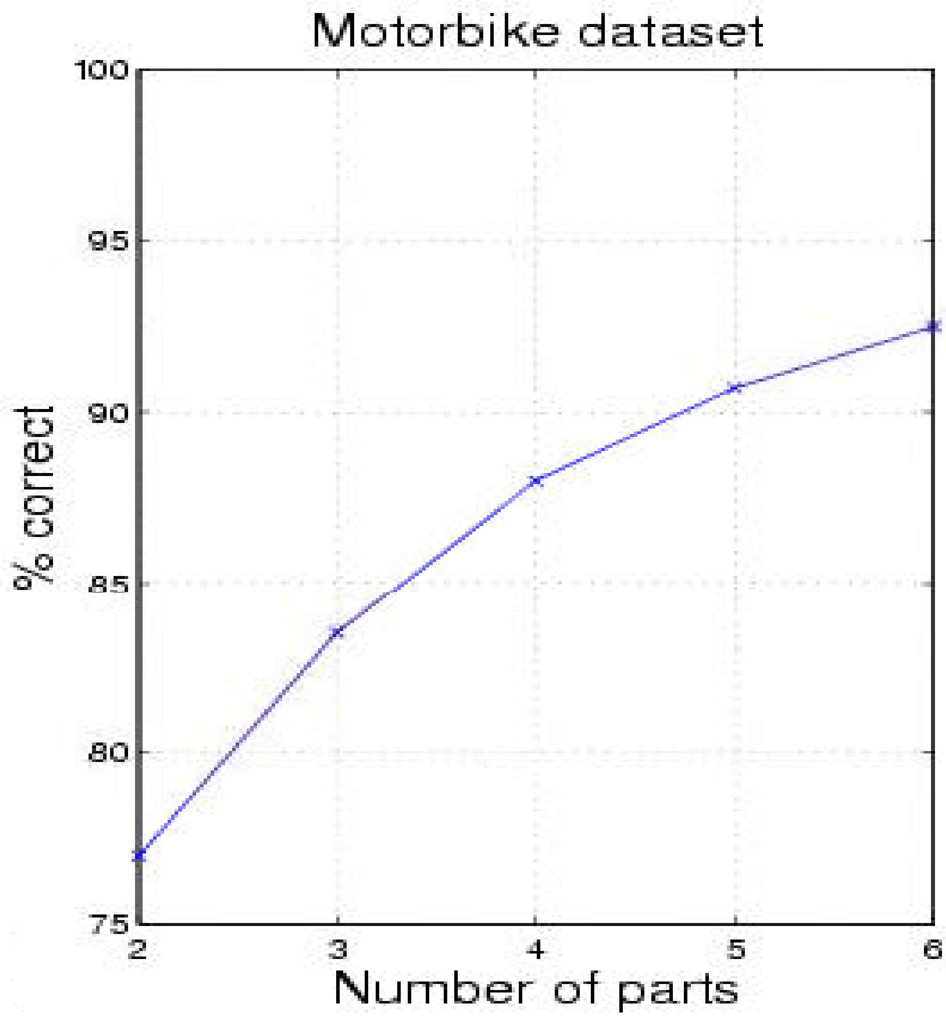
% equal error rate



Slide credit: Fergus

Robustness of Algorithm





Slide credit: Fergus

Summary -- Fergus

- Comprehensive probabilistic model for object classes
- Learn appearance, shape, relative scale, occlusion etc. simultaneously in scale and translation invariant manner
- Same algorithm gives <= 10% error across 5 diverse datasets with identical settings

Limitations $\rightarrow$ future work

- Very reliant on region detector
  - Different part types (e.g. edgel curves)
- Only learns a single viewpoint

Use mixture models

- Need lots of images to learn
  - Bayesian learning - fewer images [ICCV '03 (Fei Fei, Fergus, Perona)]
- Need more through testing
  - Looking towards testing 100's of datasets

Datasets available from:
http://www.robots.ox.ac.uk/~vgg/data

Slide credit: Fergus

Implicit Shape Model

[Leibe,Schiele04]

Mario Fritz

Learning Object Appearance Models
via
Transformed Dirichlet Processes

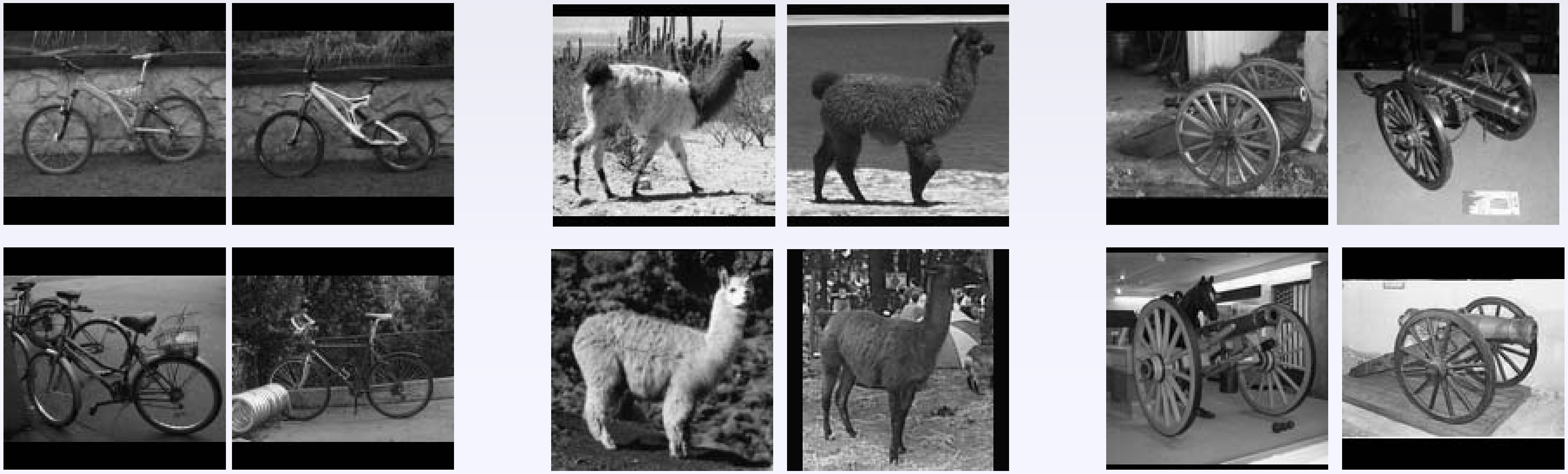
Erik Sudderth

University of California, Berkeley

Joint work with

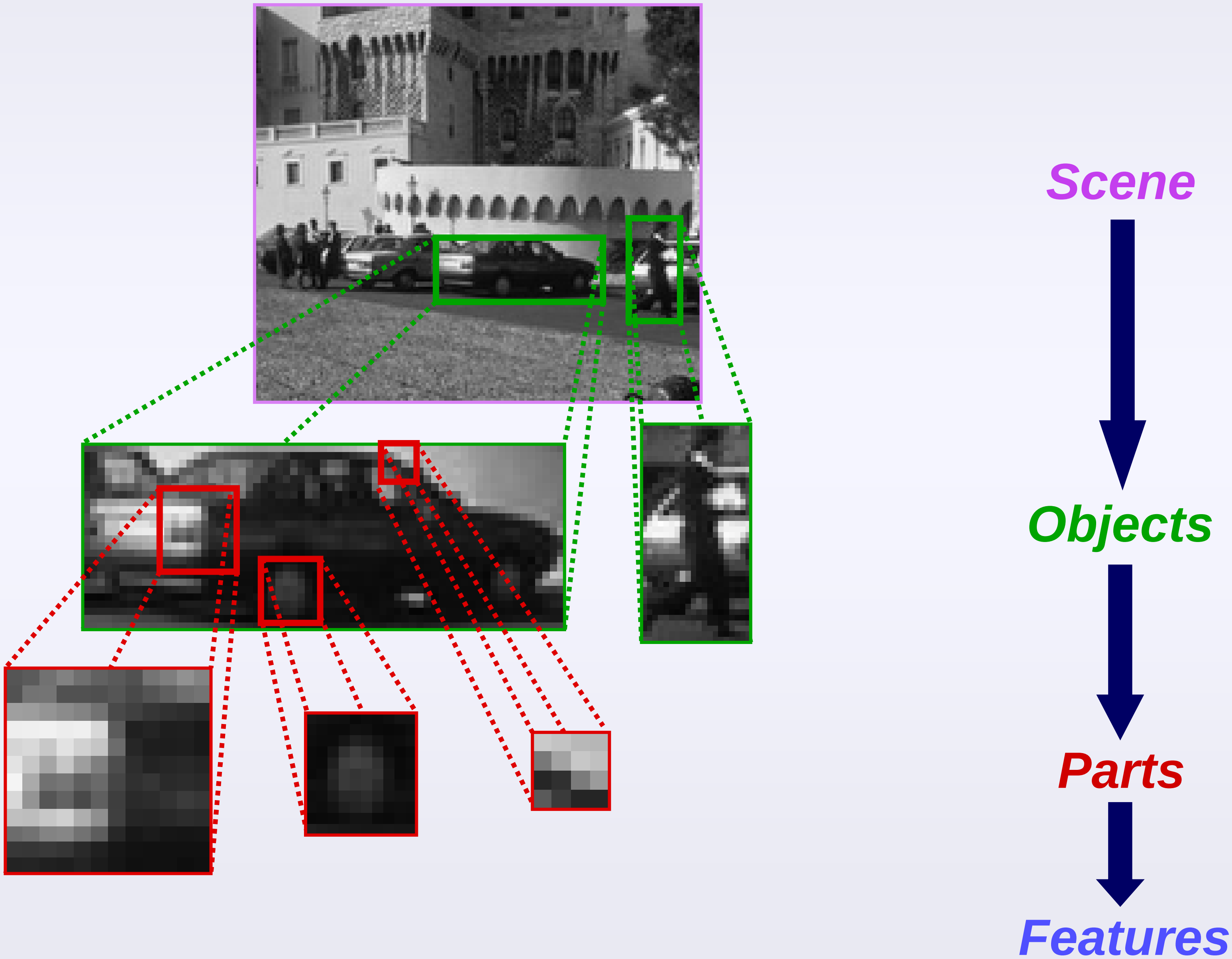
Antonio Torralba
William Freeman
Alan Willsky

Visual Object Categorization



- **GOAL:** Visually *recognize* and *localize* object categories
- Robustly *learn* appearance models from few examples
  - Hierarchical model *transfers* knowledge among categories
  - Nonparametric, *Dirichlet process* prior gives flexibility

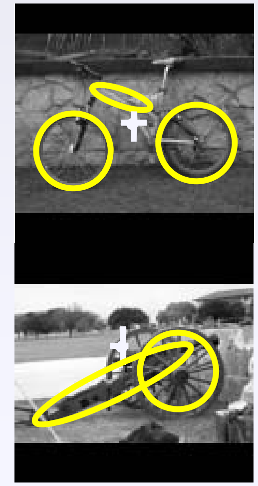
Scenes, Objects, and Parts



Outline

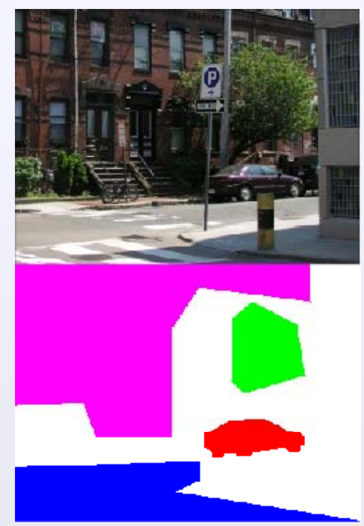
Object Recognition with Shared Parts

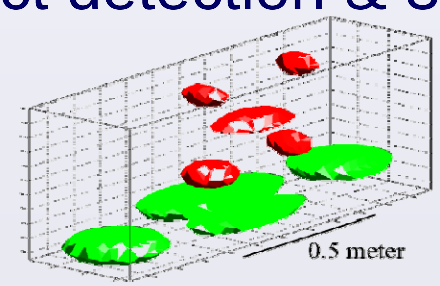
- Learning parts via Dirichlet processes
- Hierarchical DP model for 16 object categories



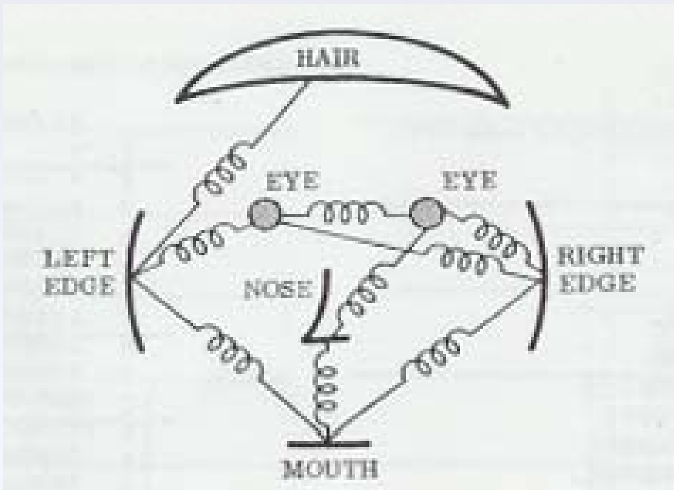
Multiple Object Scenes

- Transformed Dirichlet processes
- Part-based models for 2D scenes
- Joint object detection & 3D reconstruction

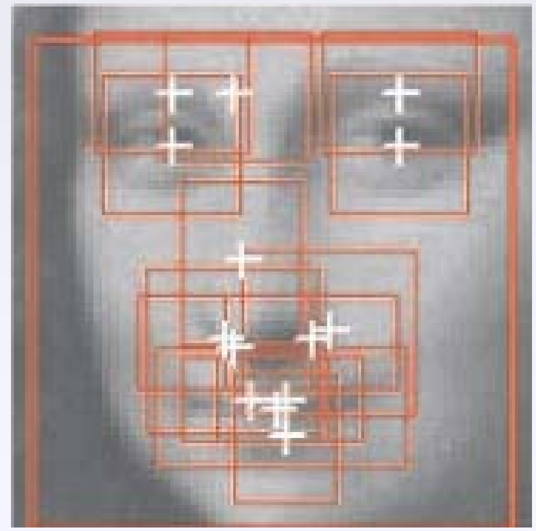




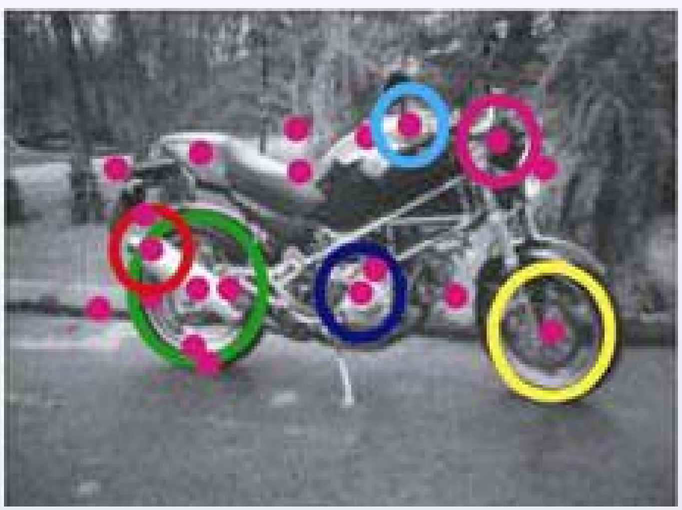
Describing Objects with Parts



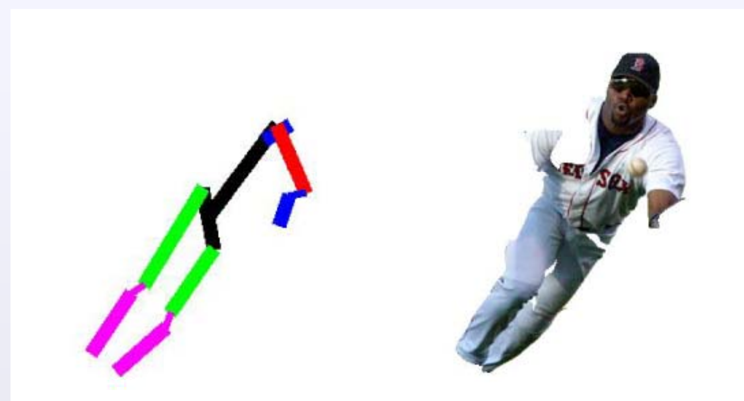
Pictorial Structures
Fischler & Elschlager, IEEE Trans. Comp. 1973



Cascaded SVM Detectors
Heisele, Poggio, et. al., NIPS 2001

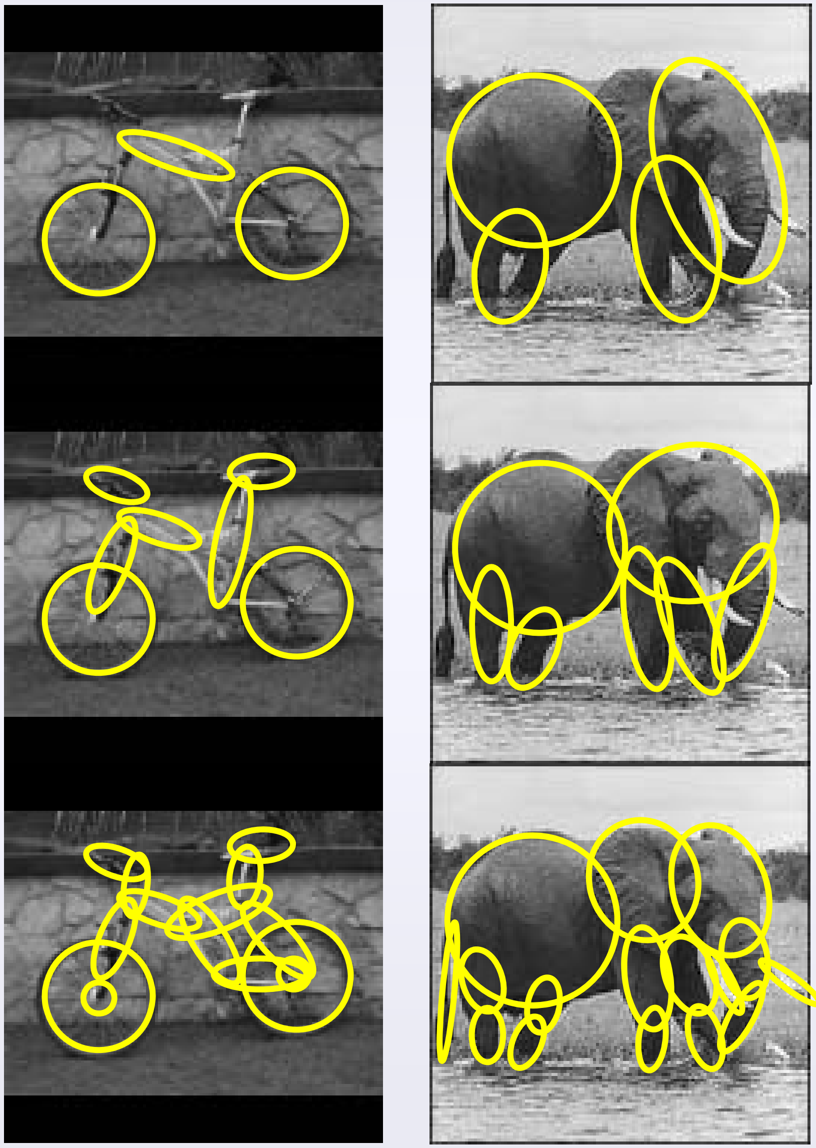


Constellation Model
Fergus, Perona, & Zisserman, CVPR 2003

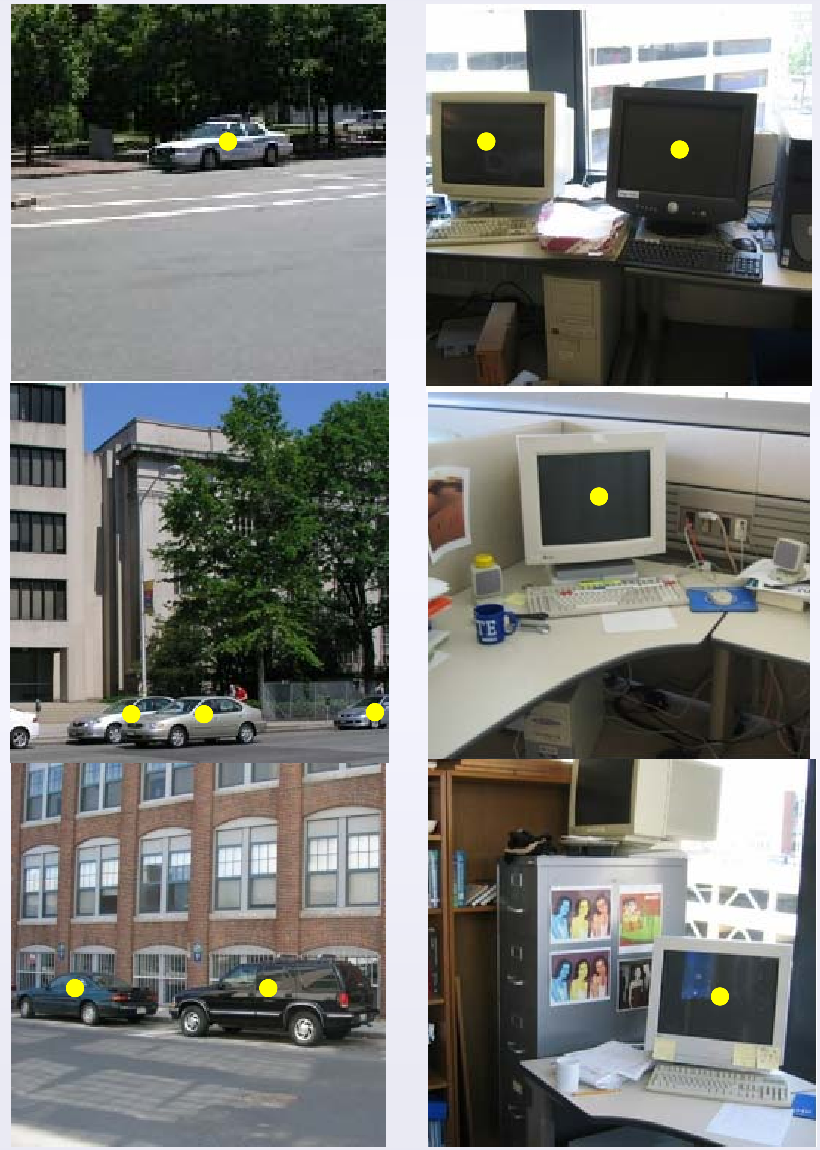


Model-Guided Segmentation
Mori, Ren, Efros, & Malik, CVPR 2004

Counting Objects & Parts

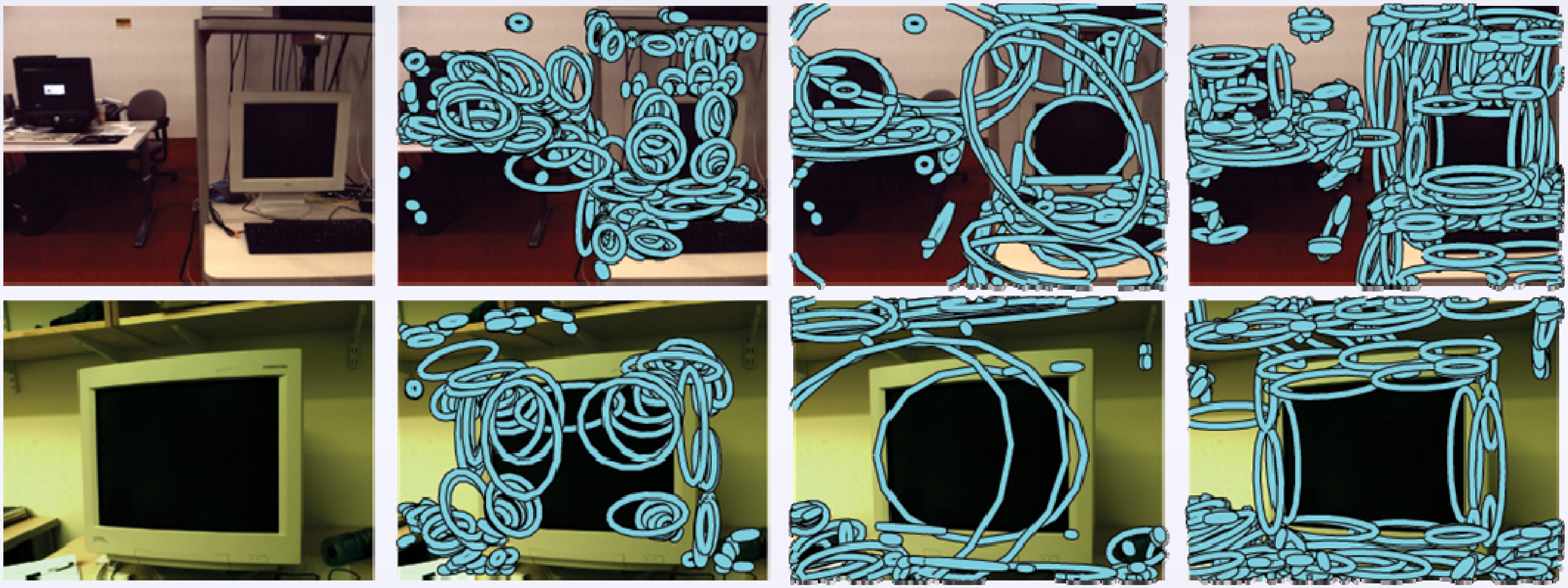


How many parts?



How many objects?

From Images to Features



Affinely Adapted Harris Corners

Maximally Stable Extremal Regions

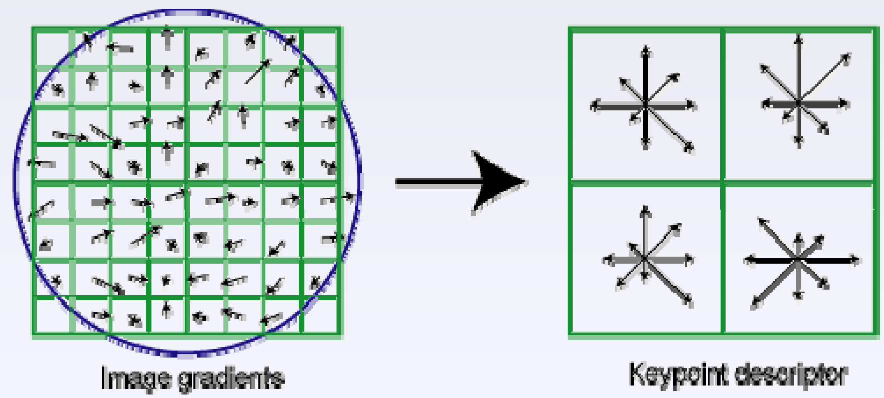
Linked Sequences of Canny Edges

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

A Discrete Feature Vocabulary

SIFT Descriptors

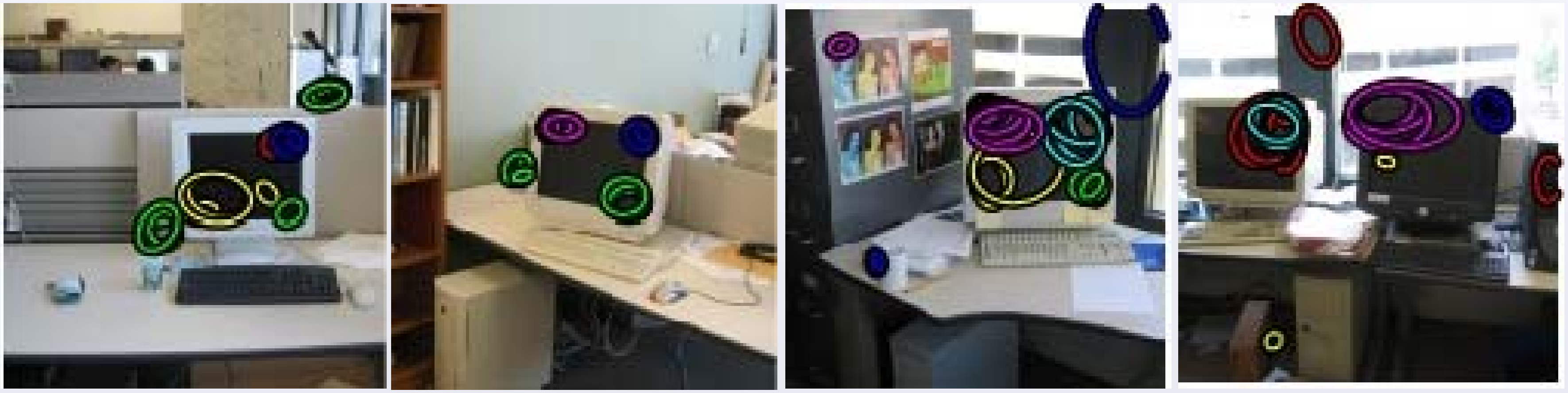
- Normalized histograms of orientation energy
- Compute ~1,000 word dictionary via K-means
- Map each feature to nearest *visual word*



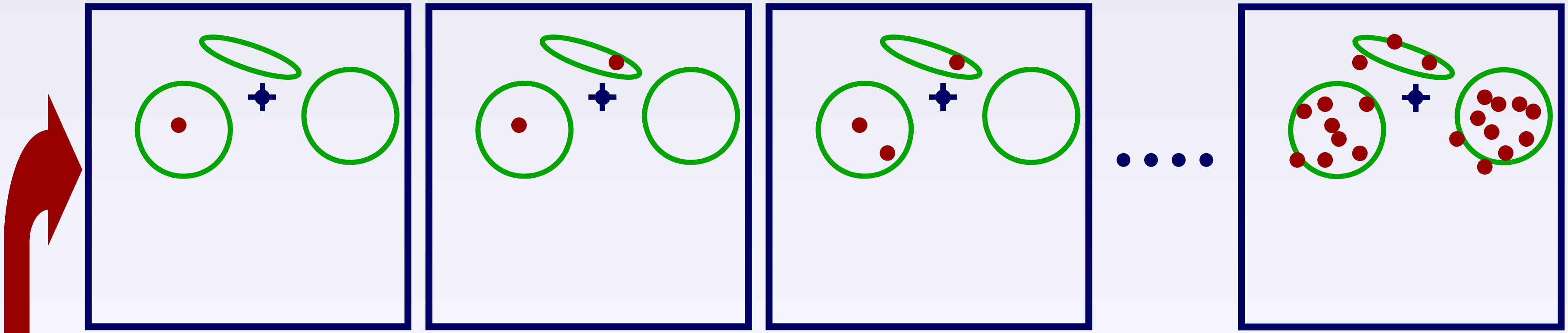
Lowe, IJCV 2004

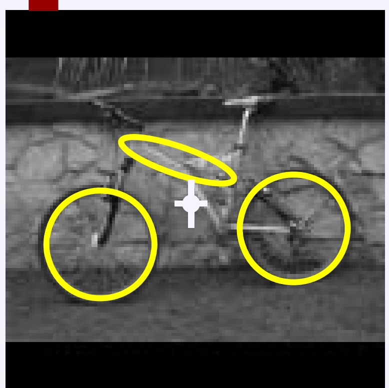
$w_{ji} \longrightarrow$ appearance of feature i in image j

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



Generative Model for Objects

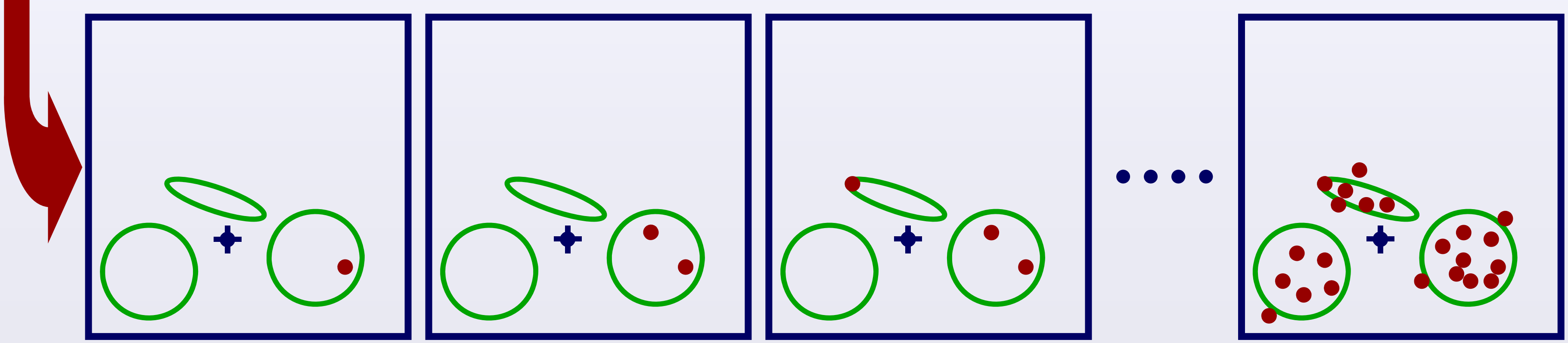




For each image: Sample a reference position

For each feature:

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



Objects as Mixture Models

- For a fixed reference position, our generative model is equivalent to a finite mixture model:

$$p(w_{ji}, v_{ji} \mid \rho_j) = \sum_{k=1}^{K} \pi_k \eta_k(w_{ji}) \mathcal{N}(v_{ji}; \mu_k + \rho_j, \Lambda_k)$$

Feature appearance

Feature position

Pr(part)

Pr(appearance | part)

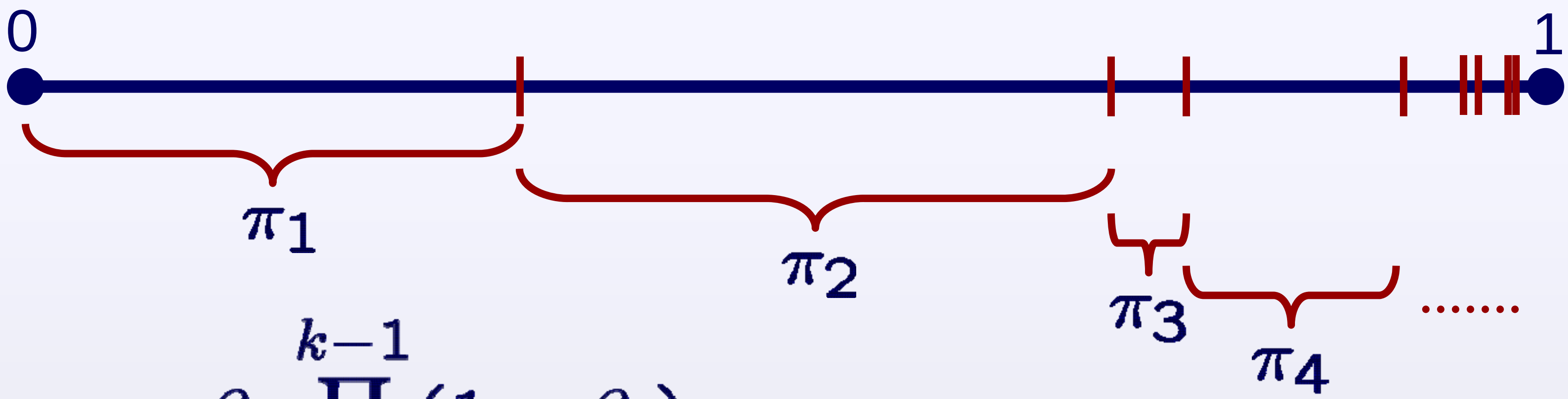
Pr(position | part)

- How many parts should we choose?
  - ➢ Too few reduces model accuracy
  - ➢ Too many causes overfitting & poor generalization

Dirichlet Process Mixtures

$$p(x) = \sum_{k=1}^{\infty} \pi_k f\left(x \mid \theta_k\right)$$

- *Dirichlet processes* define a prior distribution on weights assigned to mixture components:



$$\pi_k = \beta_k \prod_{\ell=1}^{k-1} (1 - \beta_\ell)$$

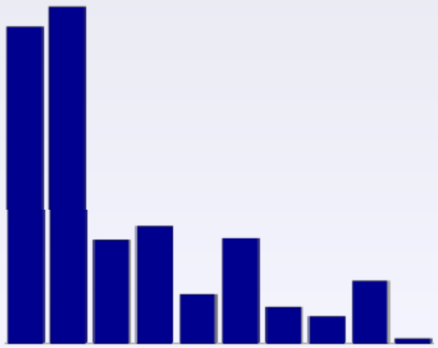
$$\beta_k \sim \text{Beta}(1, \alpha)$$

α → concentration parameter

Stick-Breaking Construction: *Sethuraman, 1994*

Why the Dirichlet Process?

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



Nonparametric $\neq$ No Parameters

- Model complexity grows as data observed:
  - Small training sets give *simple, robust* predictions
  - Reduced sensitivity to prior assumptions

Flexible but Tractable

- Literature showing attractive *asymptotic properties*
- Leads to simple, effective *computational methods*
  - Avoids challenging model selection issues

Objects as Distributions

$$p(w_{ji}, v_{ji} | \rho_j) = \sum_{k=1}^{\infty} \pi_k \eta_k(w_{ji}) \mathcal{N}(v_{ji}; \mu_k + \rho_j, \Lambda_k)$$

Feature appearance

Feature position

Pr(appearance | part)

Pr(position | 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:

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

Dirichlet Process Object Model

Part-based object model sampled from DP prior:

$$G \sim \mathrm{DP}(\alpha, H)$$

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

$$\boldsymbol{\pi} \sim \mathrm{Stick}(\alpha) \qquad \theta_k \sim H$$

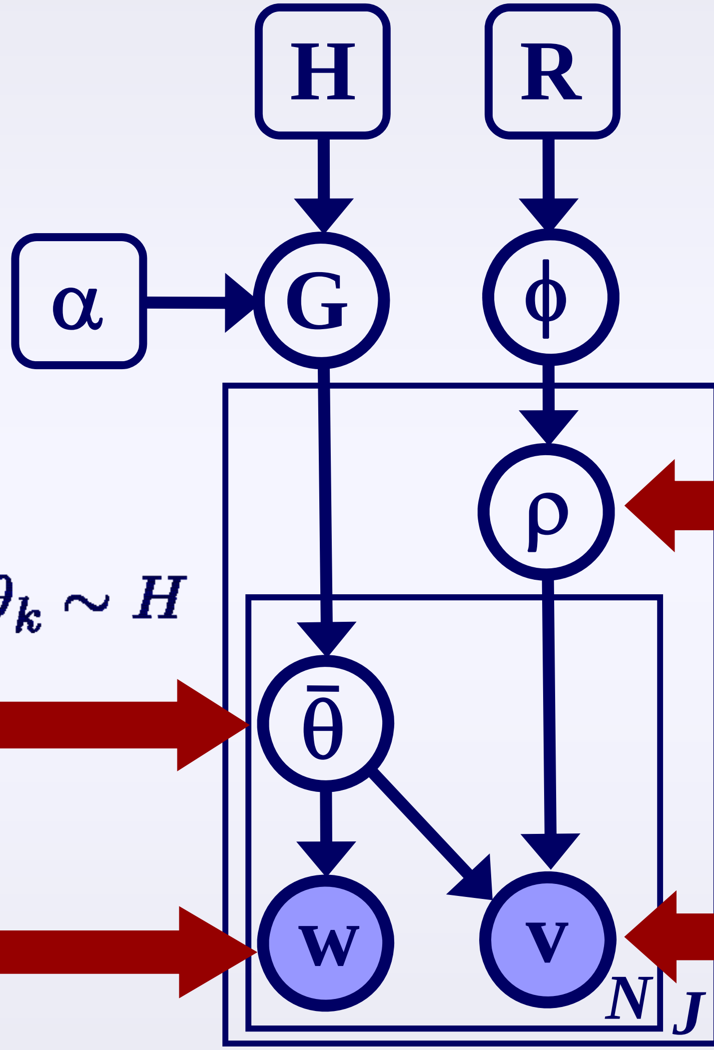
For each of N features, sample *part parameters:*

$$\bar{\theta}_{ji} \sim G(\theta)$$

Sample multinomial *feature appearance:*

$$w_{ji} \sim \bar{\eta}_{ji}(w)$$

$$\bar{\theta}_{ji} = \{\bar{\eta}_{ji}, \bar{\mu}_{ji}, \bar{\Lambda}_{ji}\}$$



For each of J images, sample a *reference position:*

$$\rho_j \sim \mathcal{N}(\rho; \phi)$$

Sample Gaussian *feature position:*

$$v_{ji} \sim \mathcal{N}(v; \bar{\mu}_{ji} + \rho_j, \bar{\Lambda}_{ji})$$

Learning DPs: Gibbs Sampling

Part Scale &
Visual Sparsity
(insensitive)

Image Scale
(insensitive)

Sample

Integrate

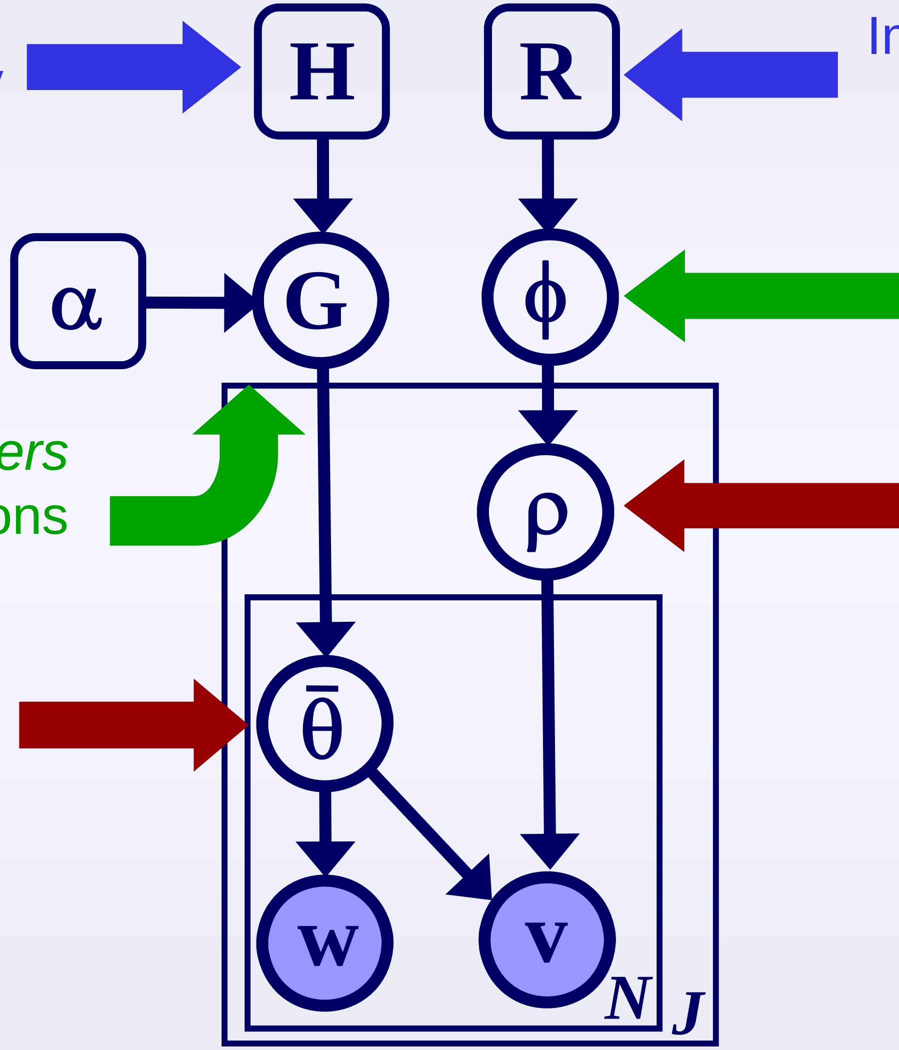
Integrate *parameters*
defining feature distributions

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

Sample

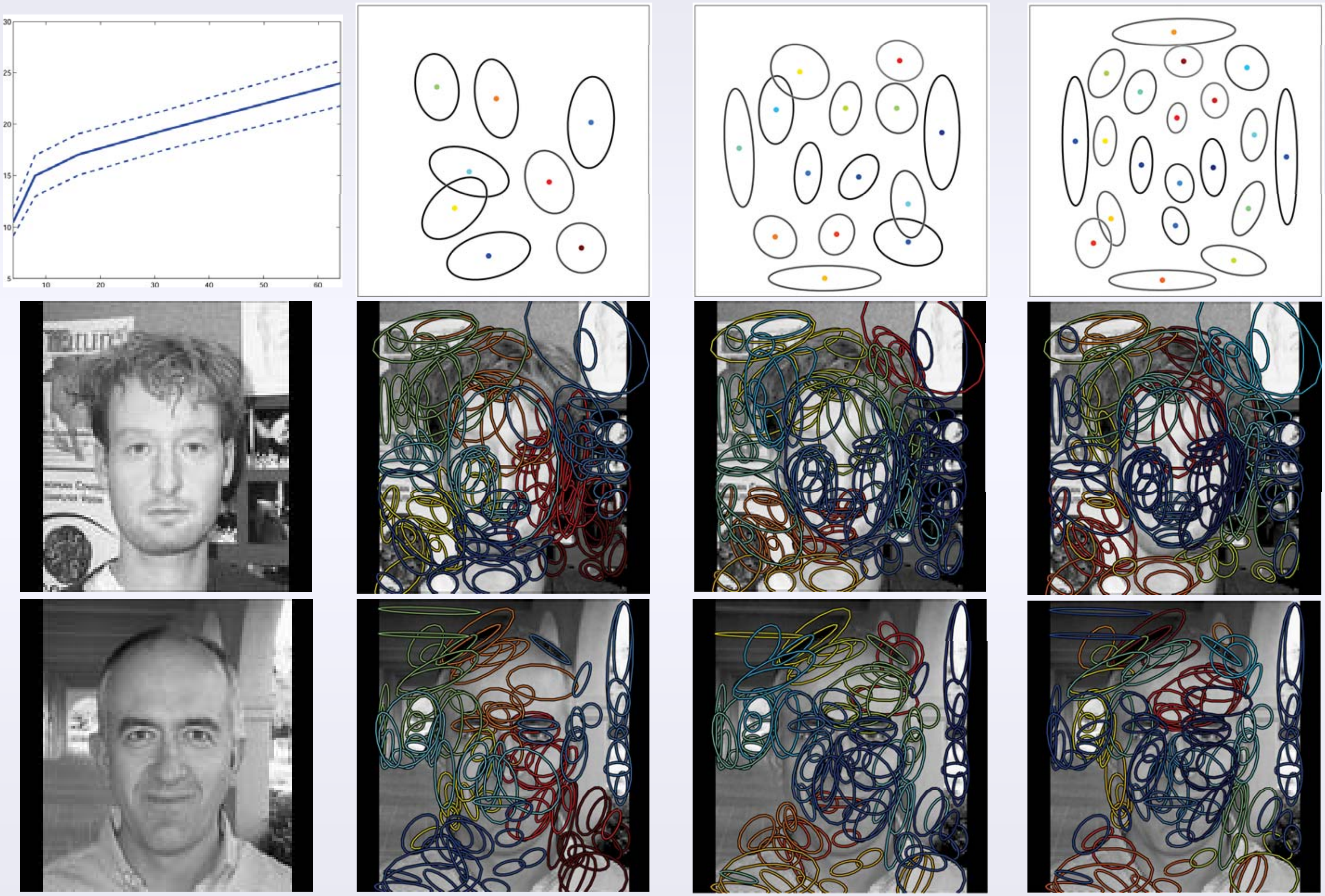
Sample *assignments*
clustering features to parts

$$(w_{ji}, v_{ji}) \rightarrow k_{ji}$$



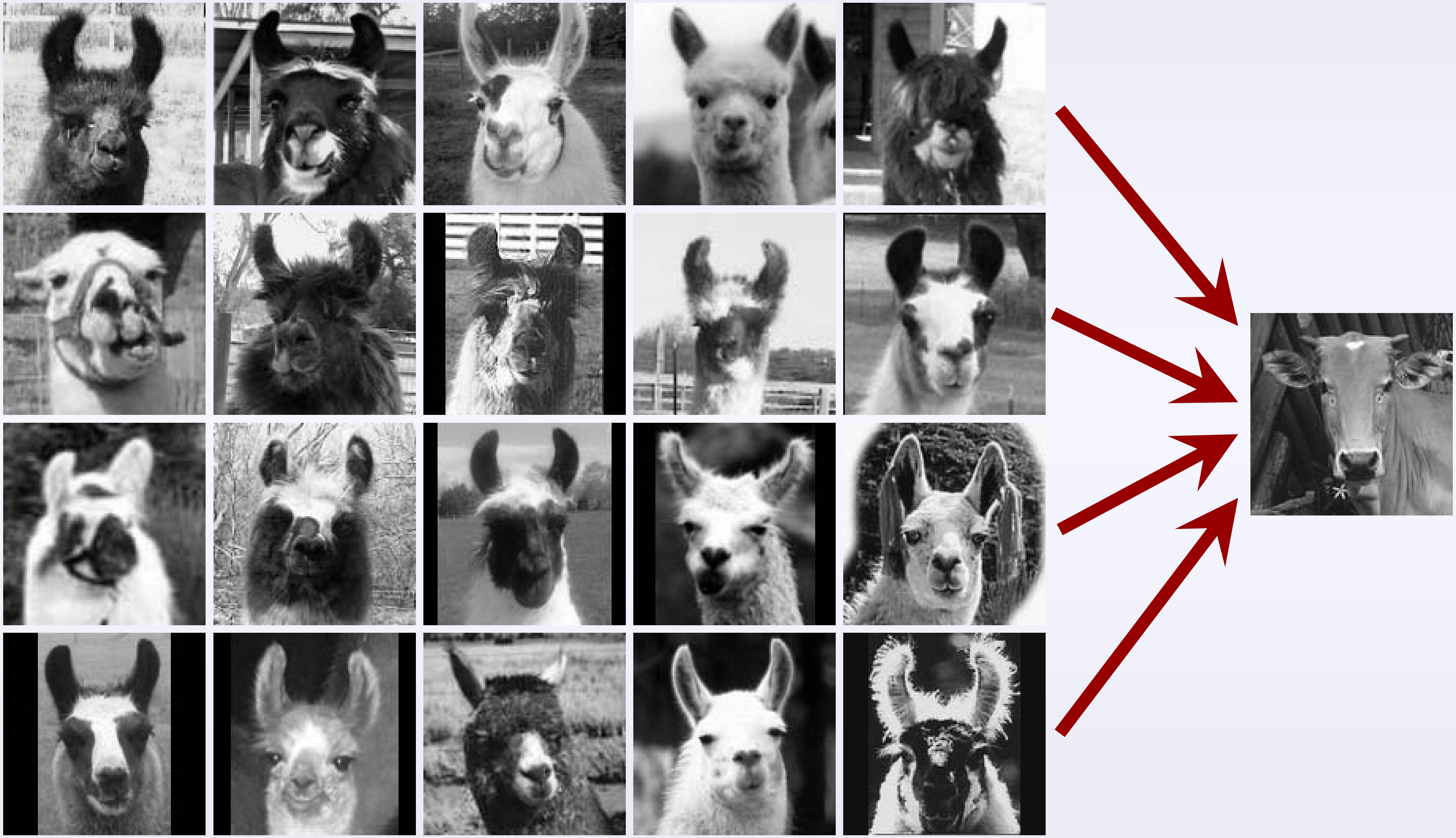
Dirichlet processes have many desirable analytic properties, which lead to efficient *Rao-Blackwellized* learning algorithms

Decomposing Faces into Parts



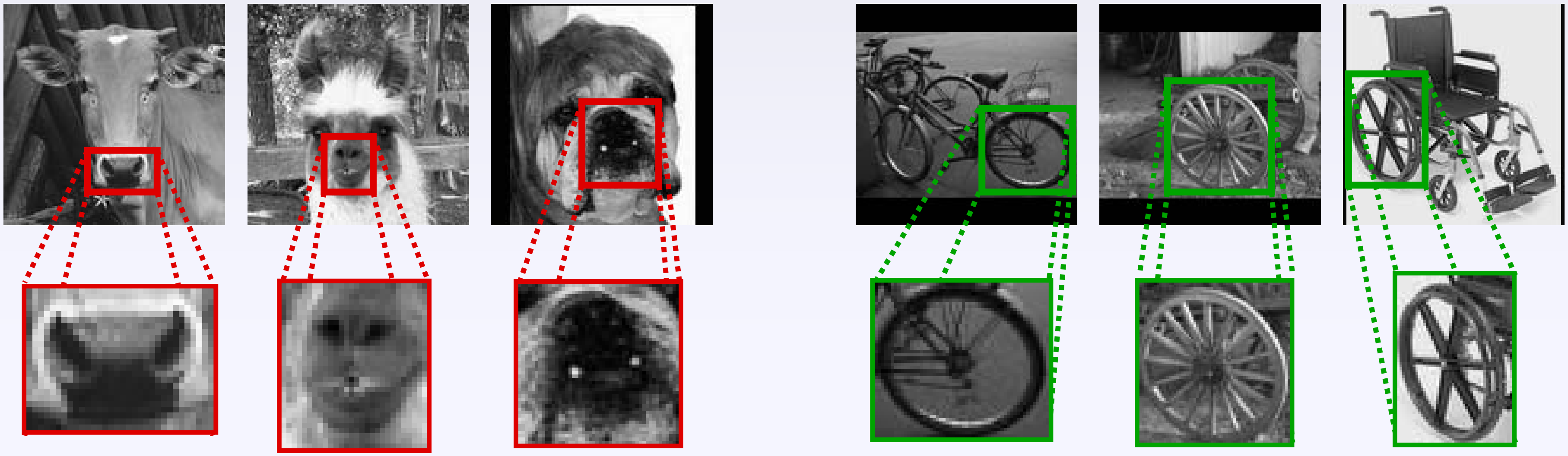
4 Images *16 Images* *64 Images*

Generalizing Across Categories



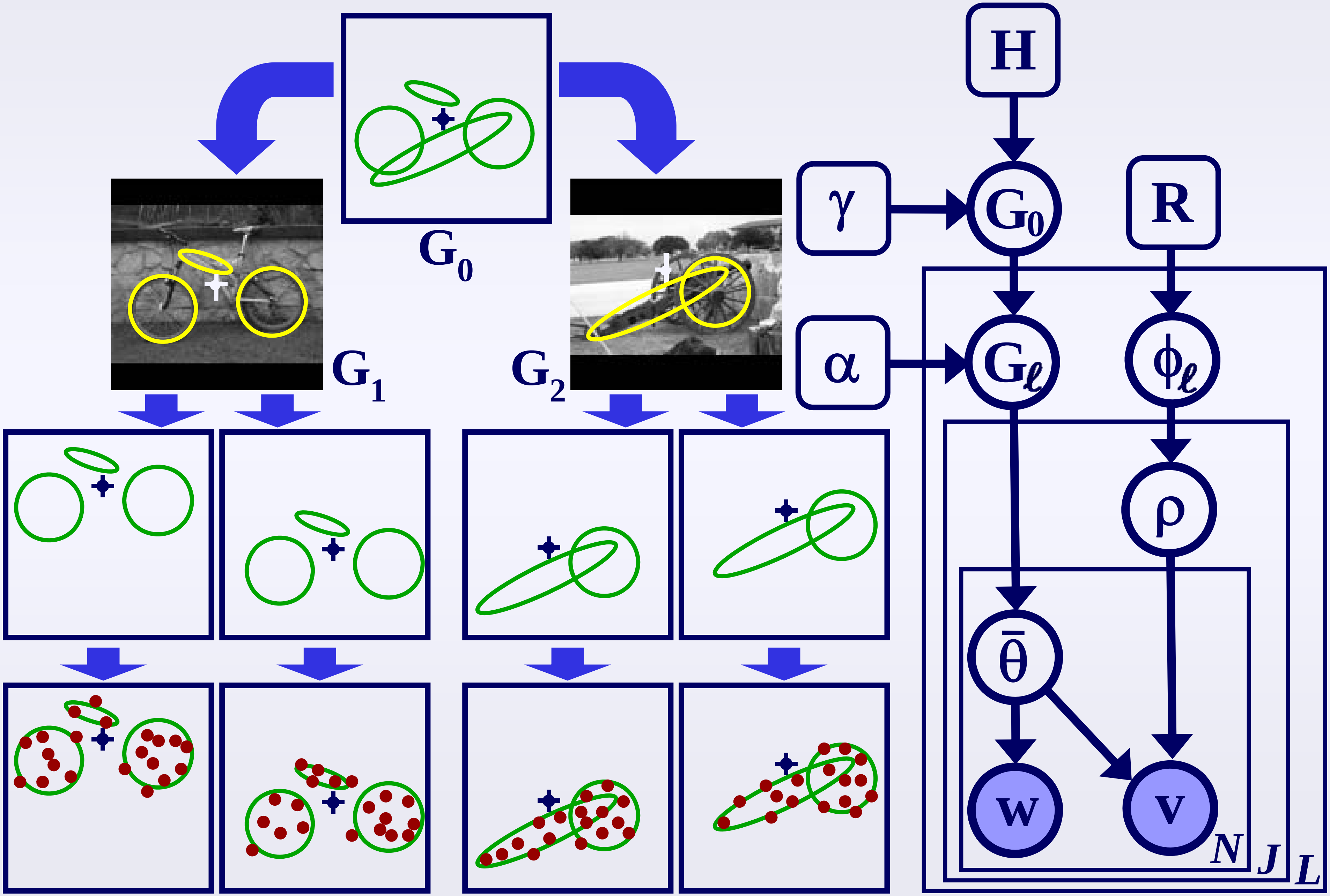
Can we transfer knowledge from one object category to another?

Learning Shared Parts

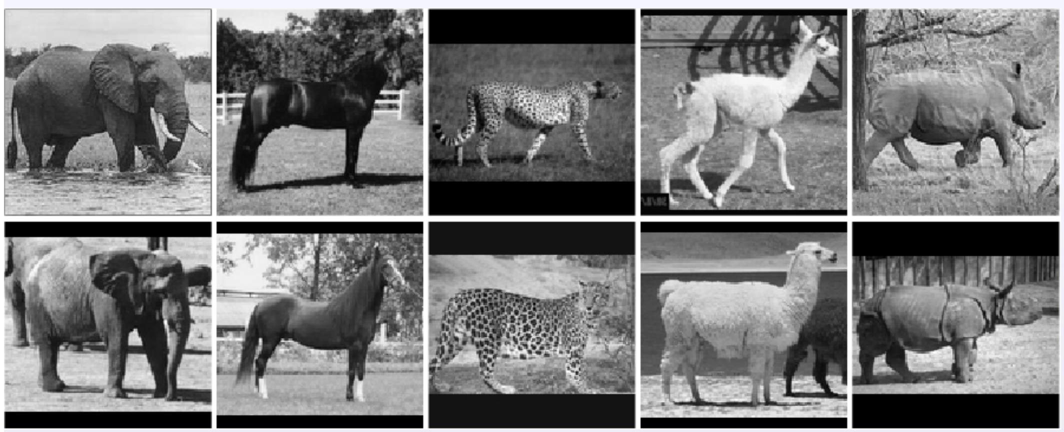


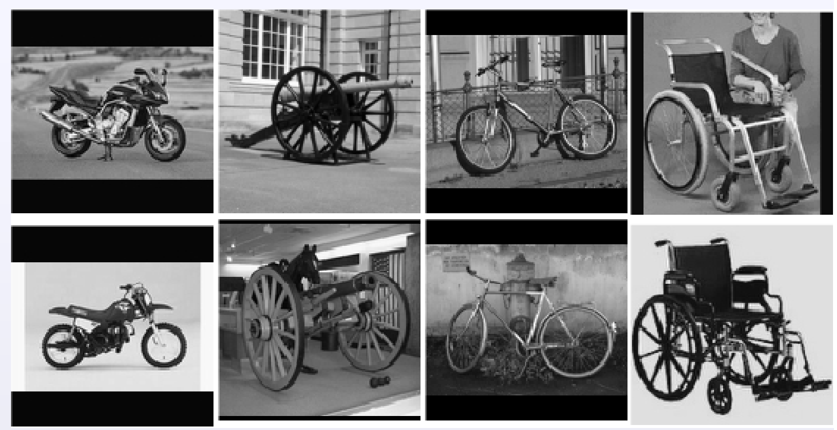
- Objects are often locally similar in appearance
- Discover *parts* shared across categories
  - ➢ How many total parts should we share?
  - ➢ How many parts should each category use?

Hierarchical DP Object Model



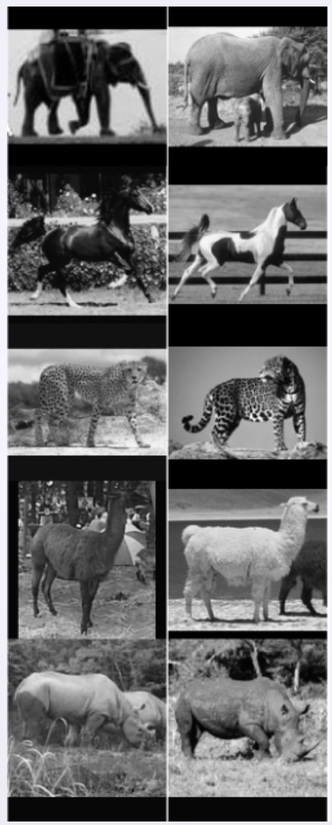
Sharing Parts: 16 Categories



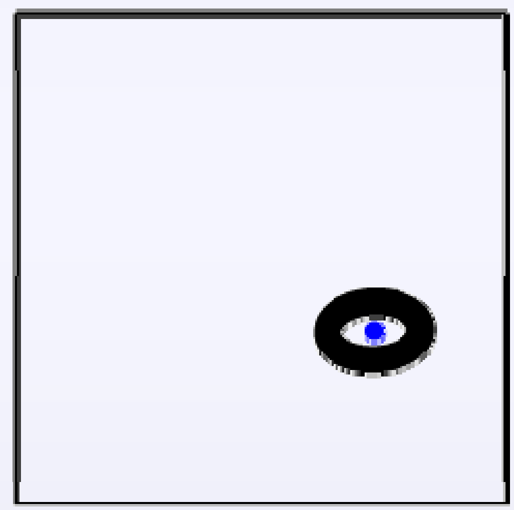


- Caltech 101 Dataset (Li & Perona)
- Horses (Borenstein & Ullman)
- Cat & dog faces (Vidal-Naquet & Ullman)
- Bikes from Graz-02 (Opelt & Pinz)
- Google…

Visualization of Shared Parts

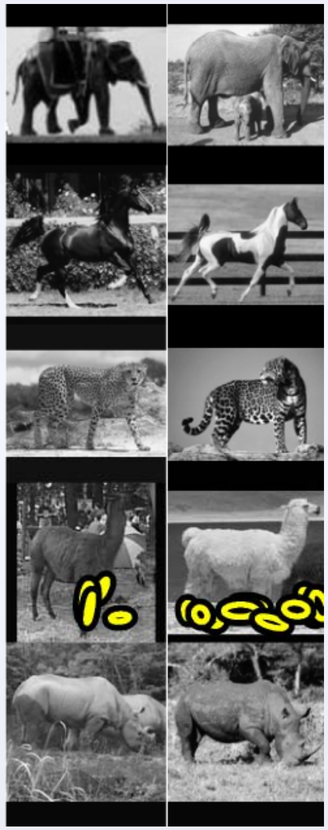


Pr(appearance | part)

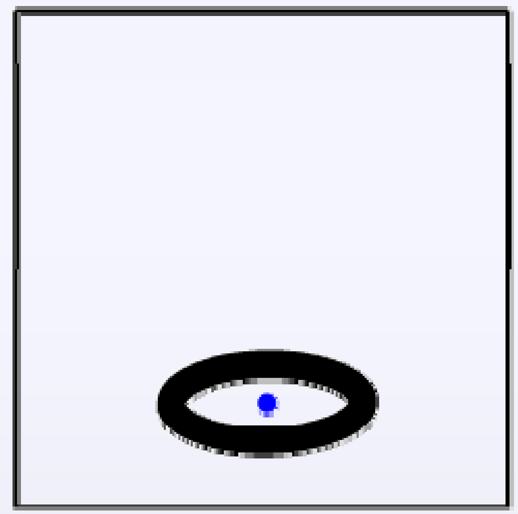


Pr(position | part)

Visualization of Shared Parts

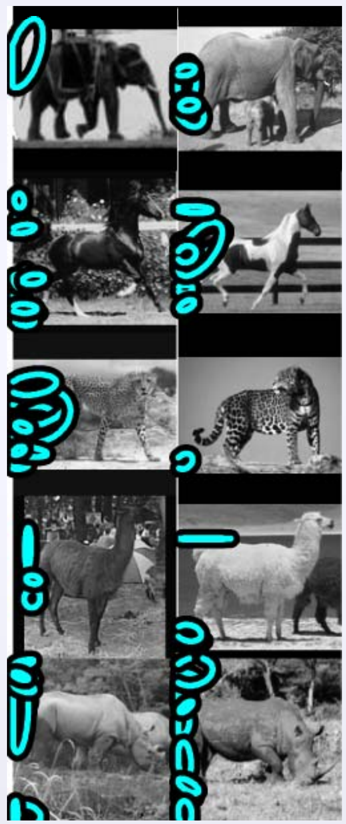


Pr(appearance | part)

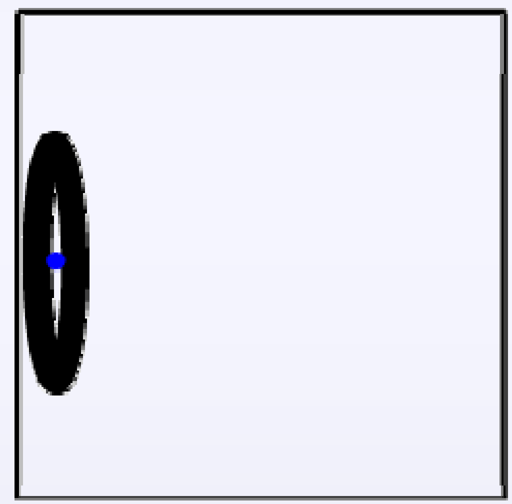


Pr(position | part)

Visualization of Shared Parts

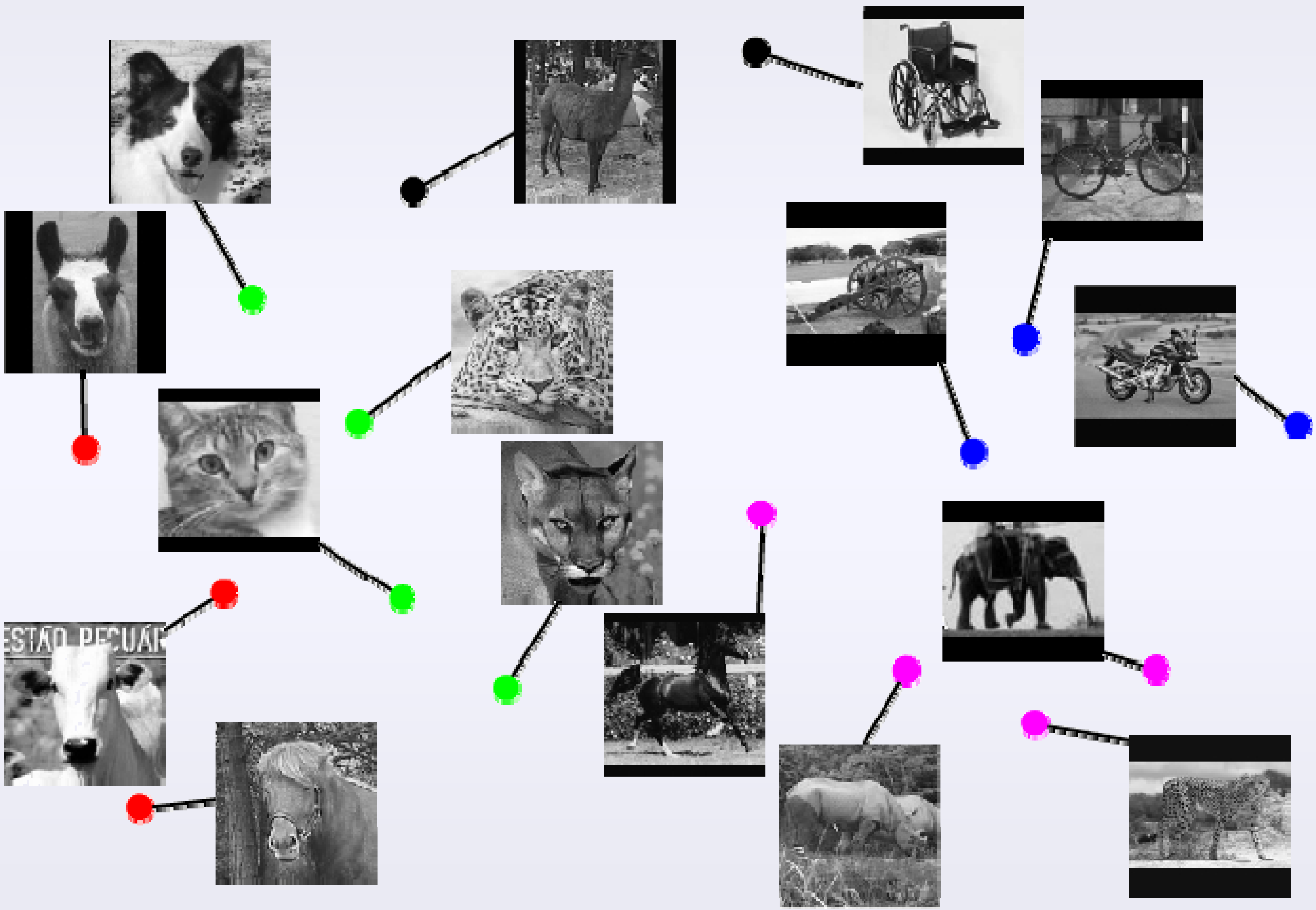


Pr(appearance | part)



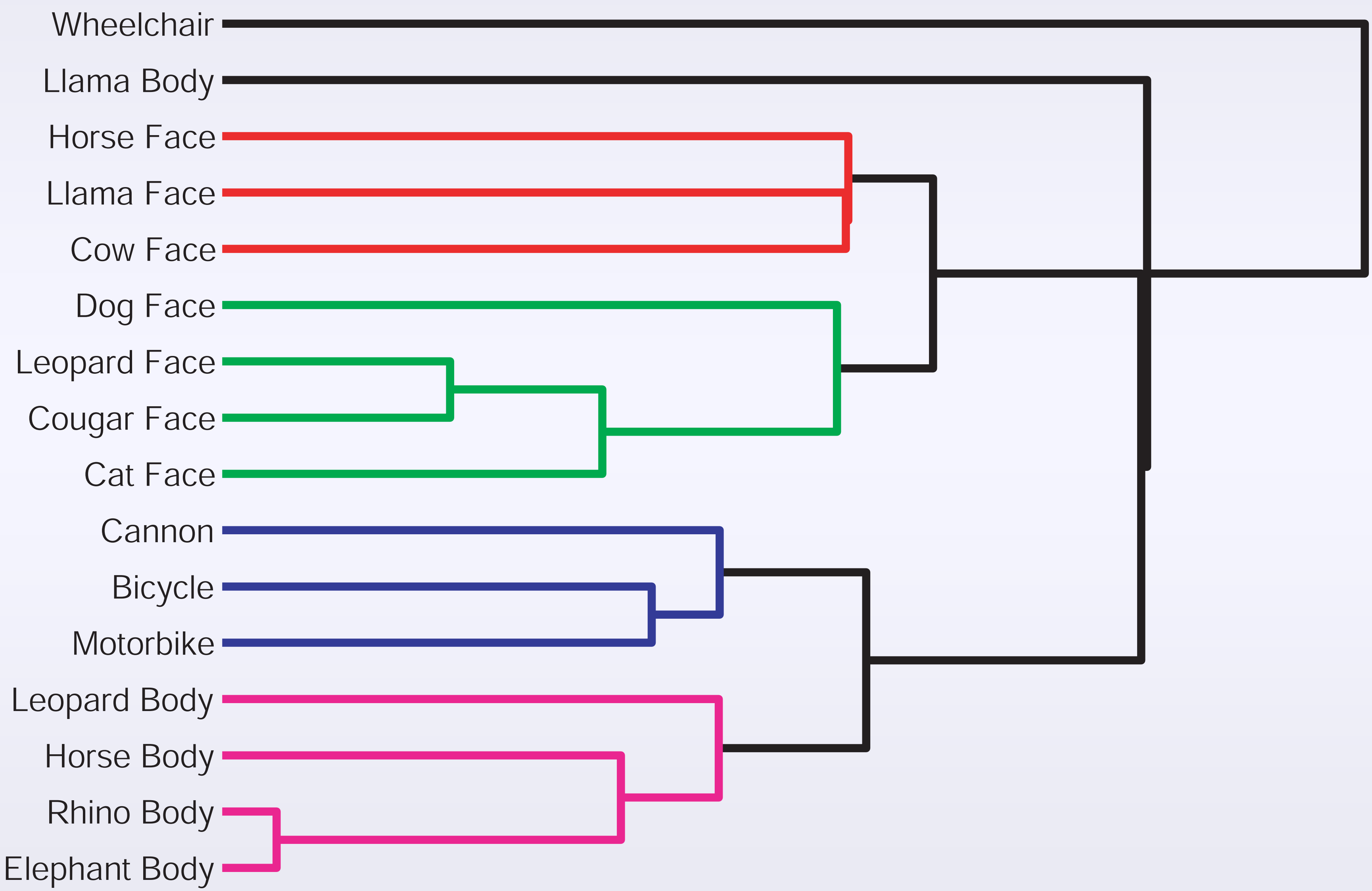
Pr(position | part)

Visualization of Part Densities



MDS Embedding of Pr(part | object)

Visualization of Part Densities

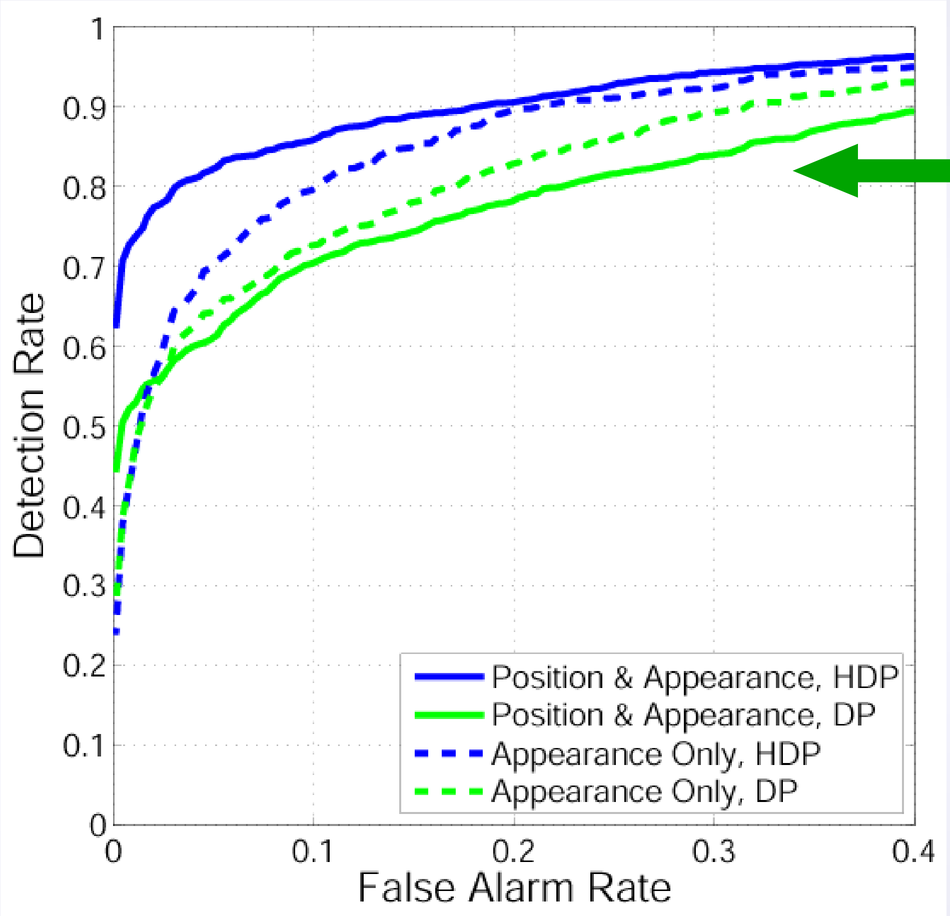


Hierarchical Clustering of Pr(part | object)

Detection Task

versus

Detection Results

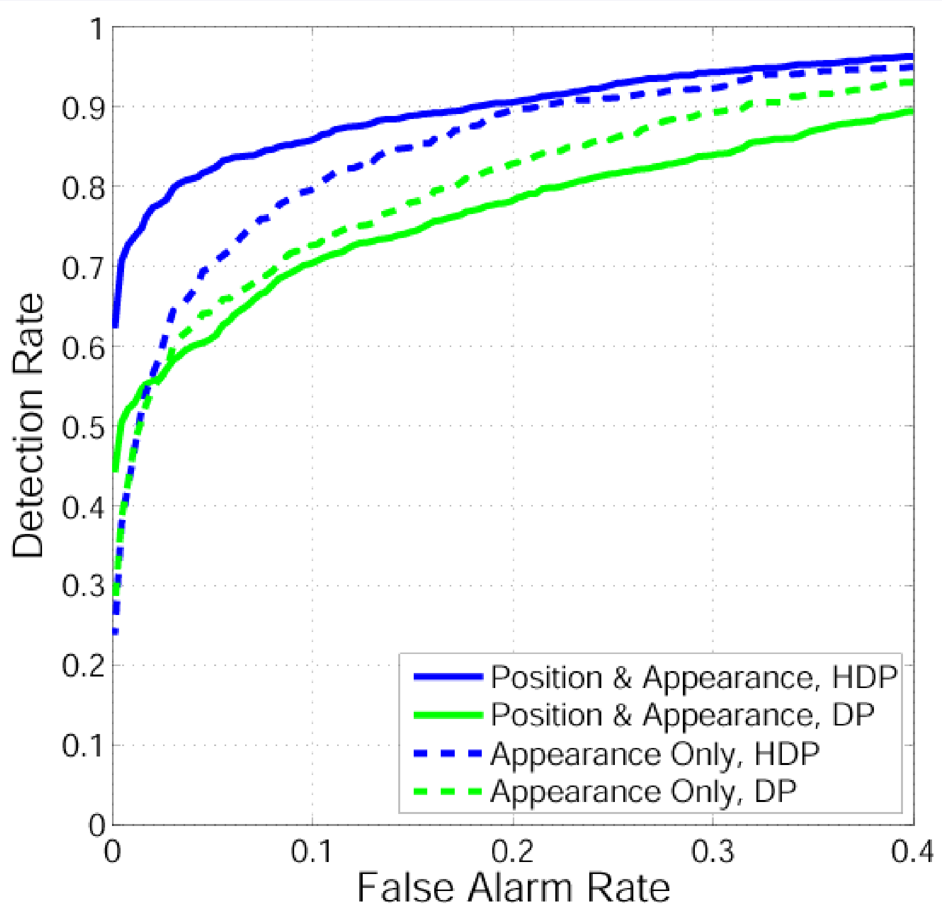


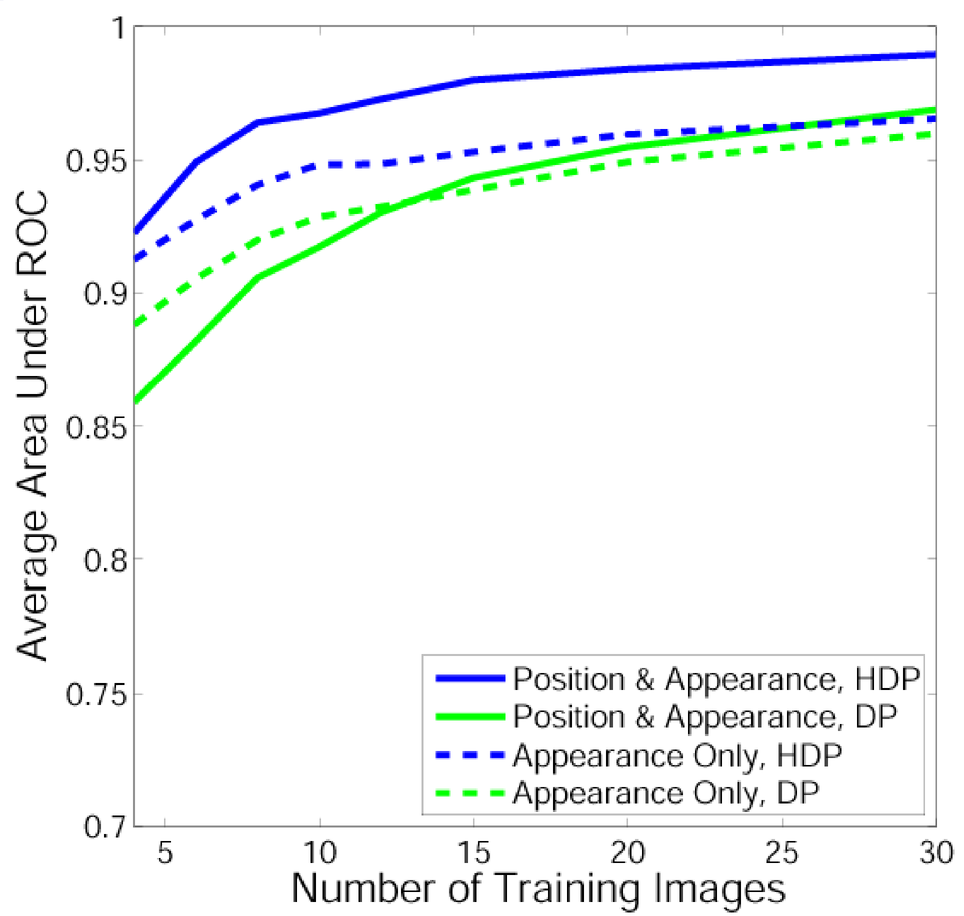
6 Training Images per Category
(ROC Curves)

Shared Parts
more accurate than
Unshared Parts

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

Detection Results

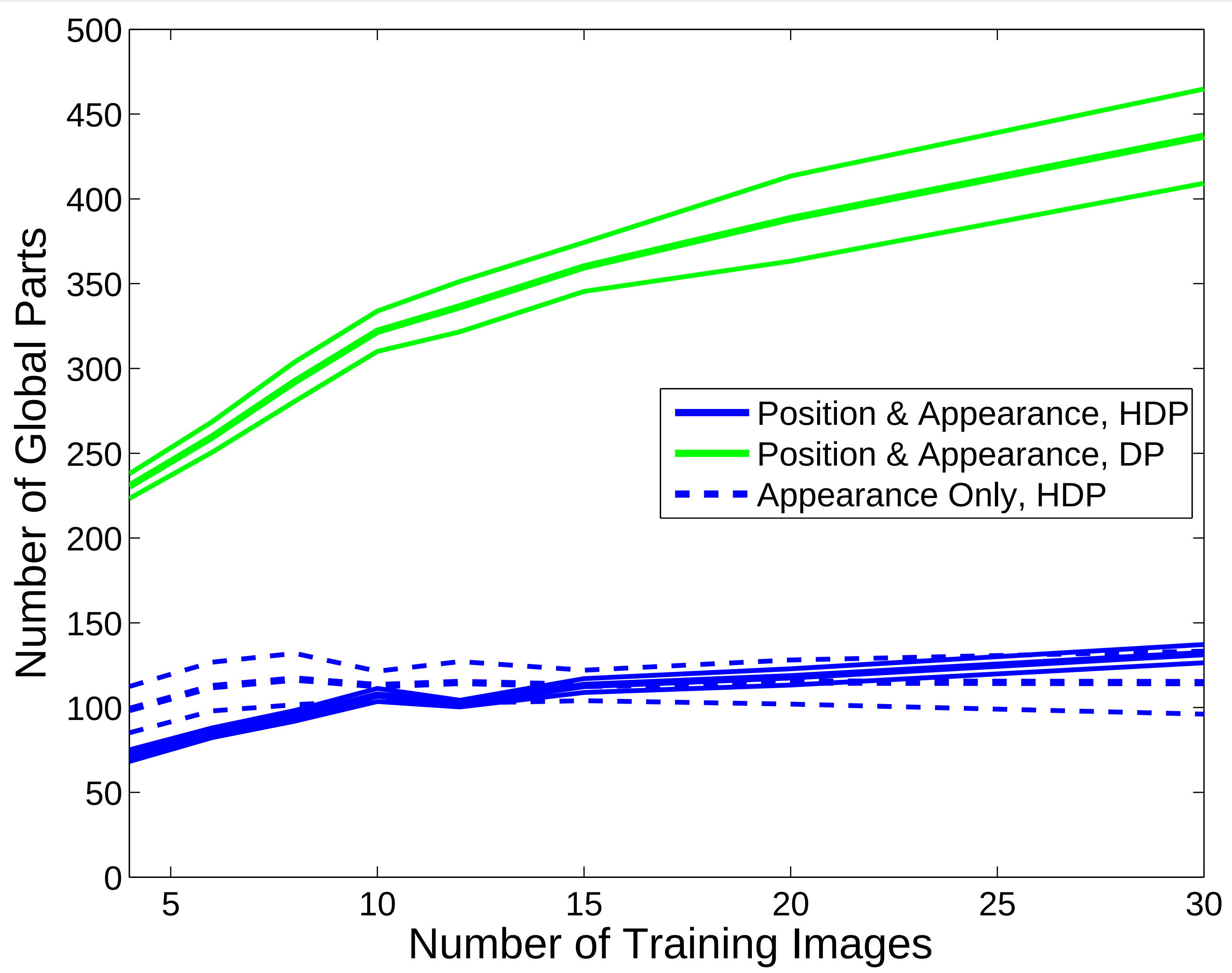




6 Training Images per Category
(ROC Curves)

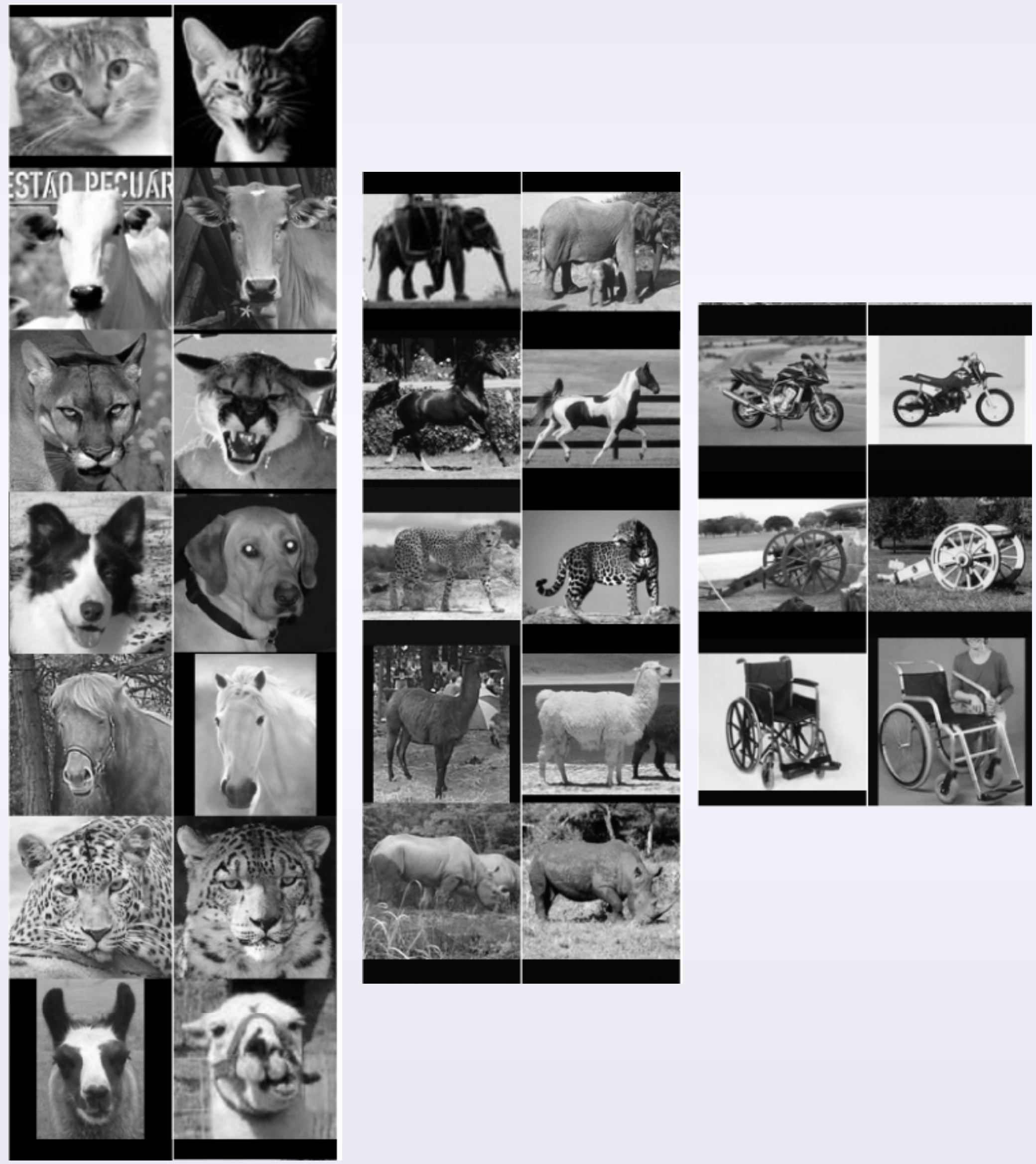
Detection vs. Training Set Size
(Area Under ROC)

Sharing Simplifies Models

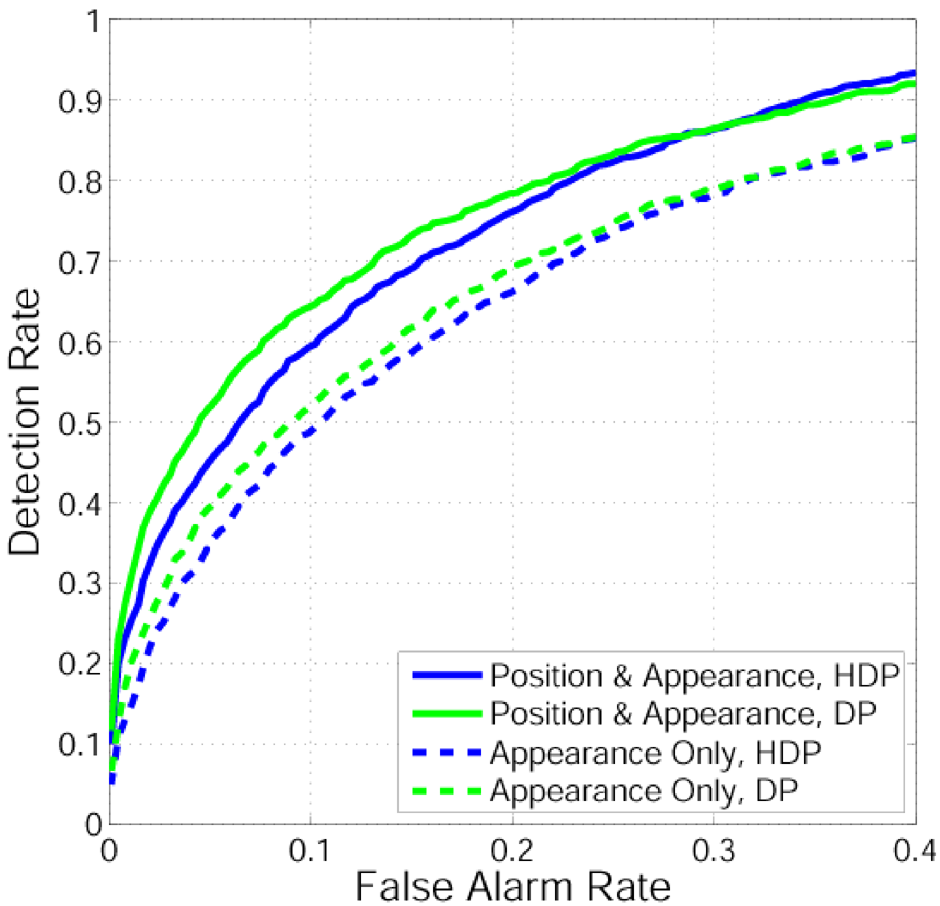


Recognition Task

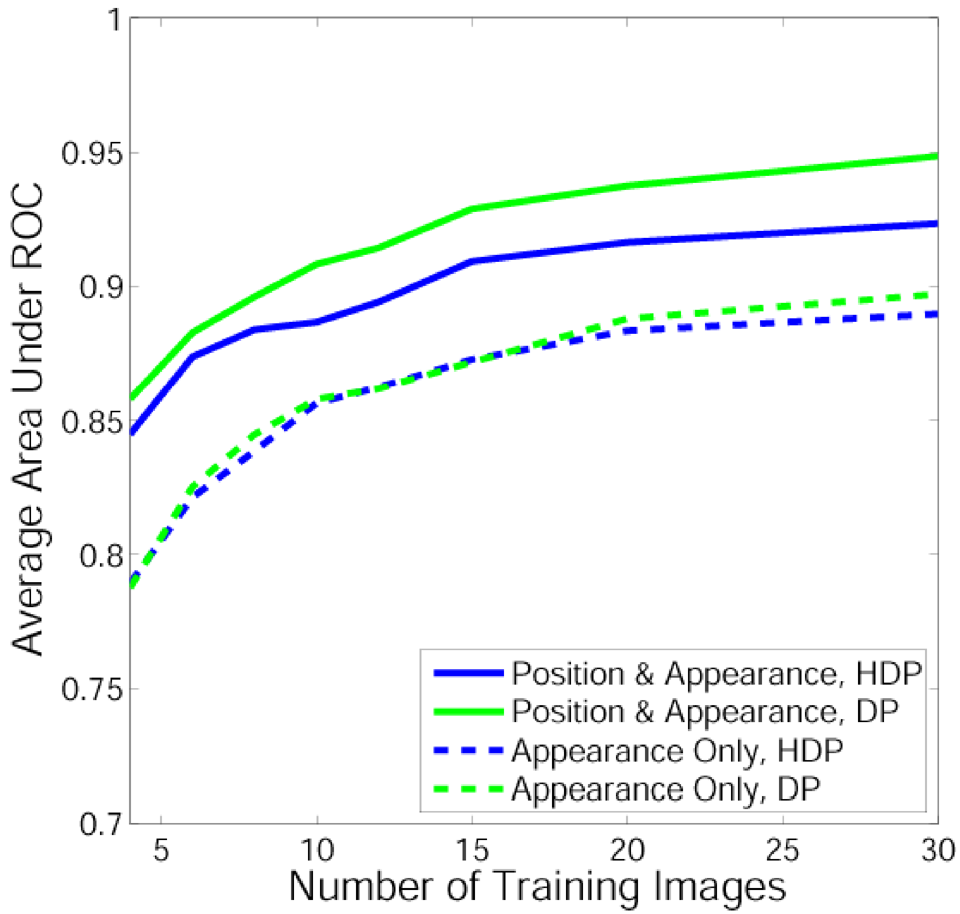
versus



Recognition Results



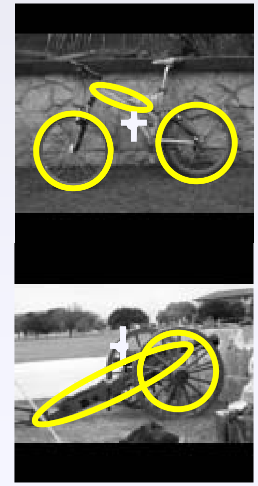
6 Training Images per Category
(ROC Curves)



Detection vs. Training Set Size
(Area Under ROC)

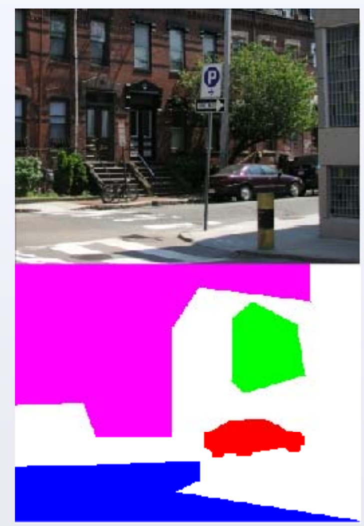
Outline

Object Recognition with Shared Parts

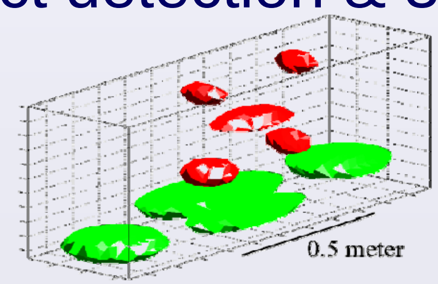


- Learning parts via Dirichlet processes
- Hierarchical DP model for 16 object categories

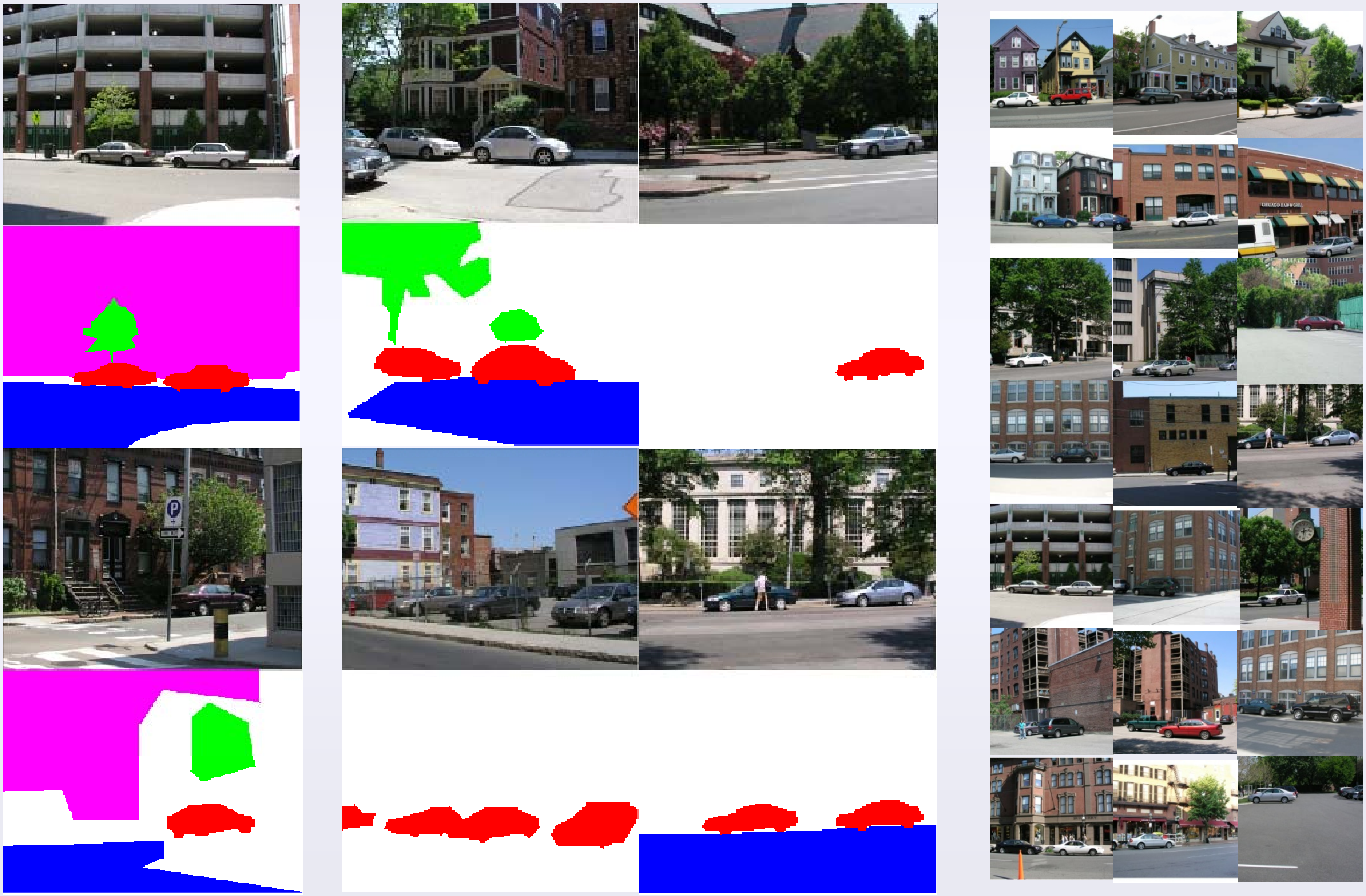
Multiple Object Scenes



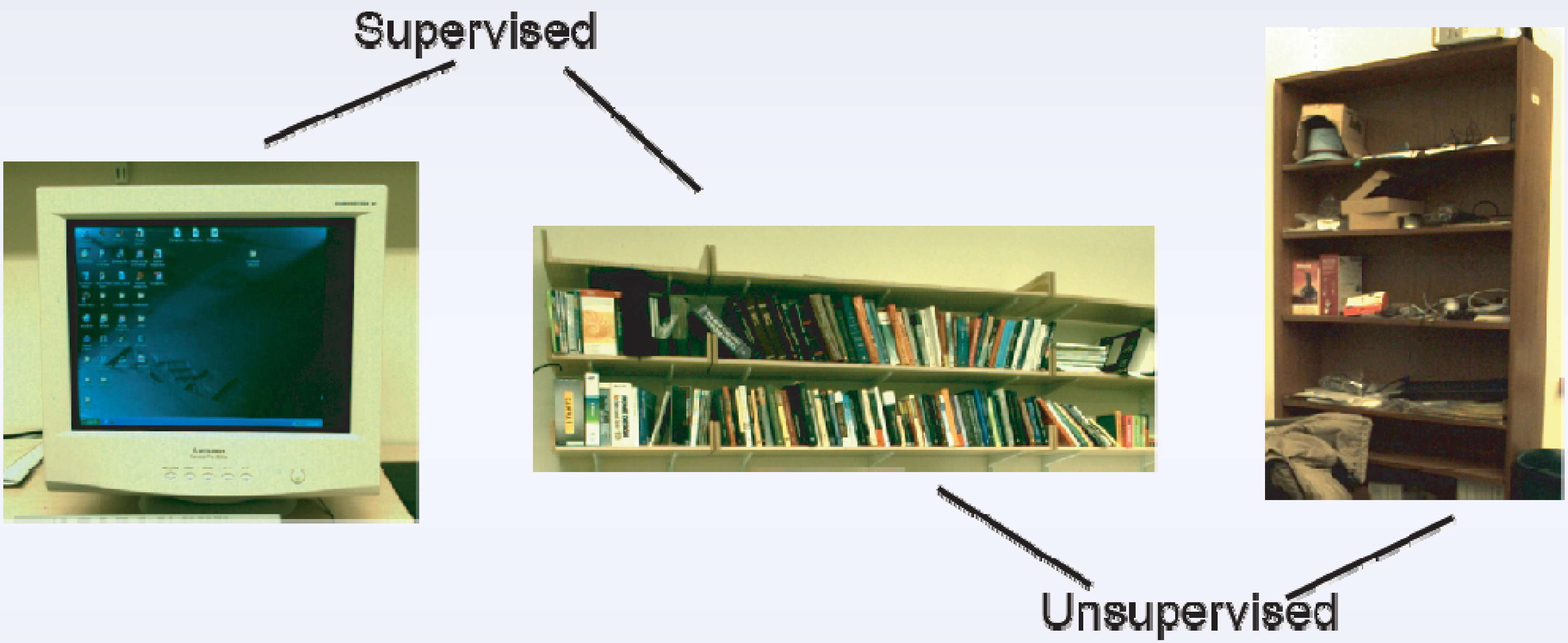
- Transformed Dirichlet processes
- Part-based models for 2D scenes
- Joint object detection & 3D reconstruction



Semi-supervised Learning

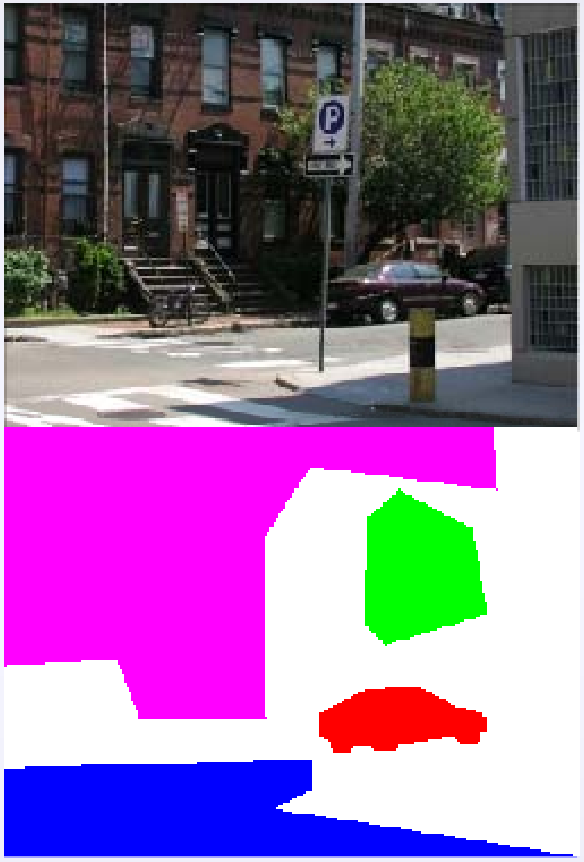


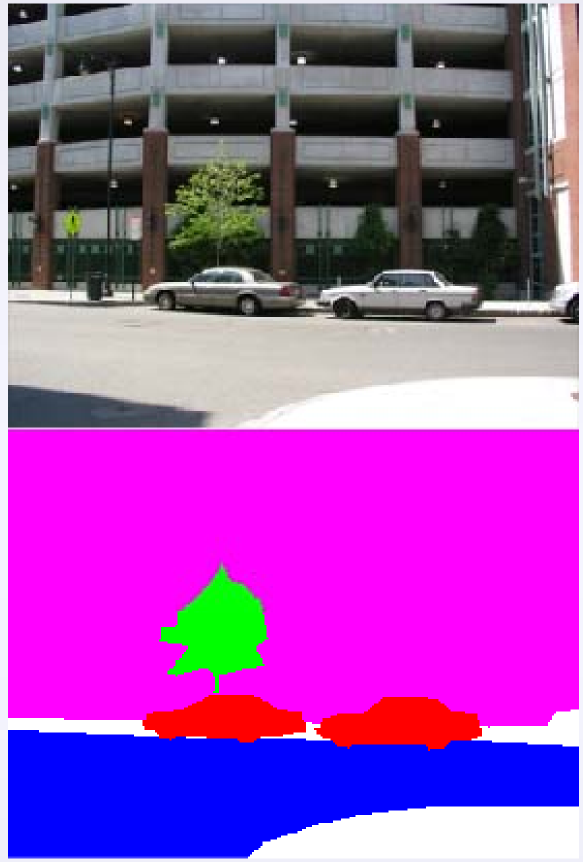
Object vs. Visual Categories

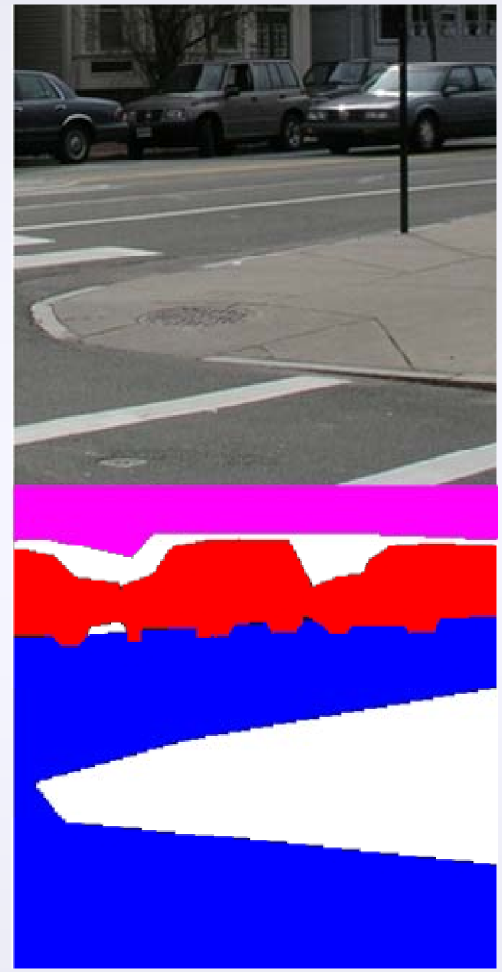


- Assume training data contains object category labels
- Discover underlying visual categories automatically

Multiple Object Scenes





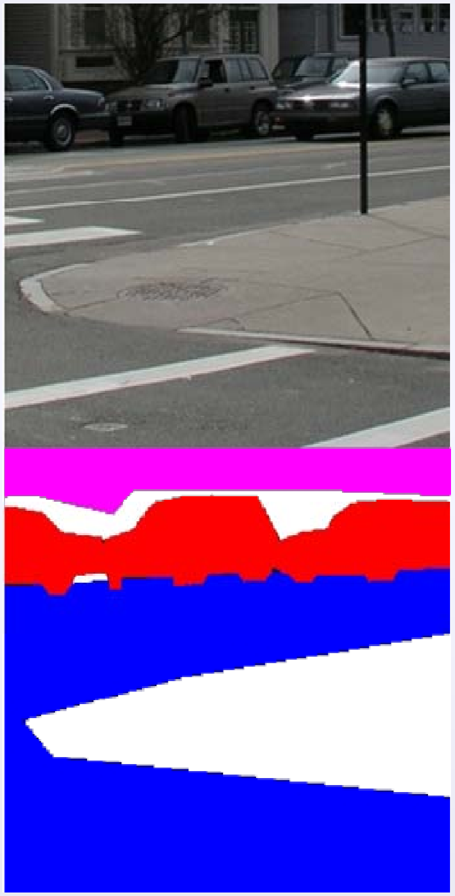


- How many cars are there?
- Where are those cars in the scene?

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:*



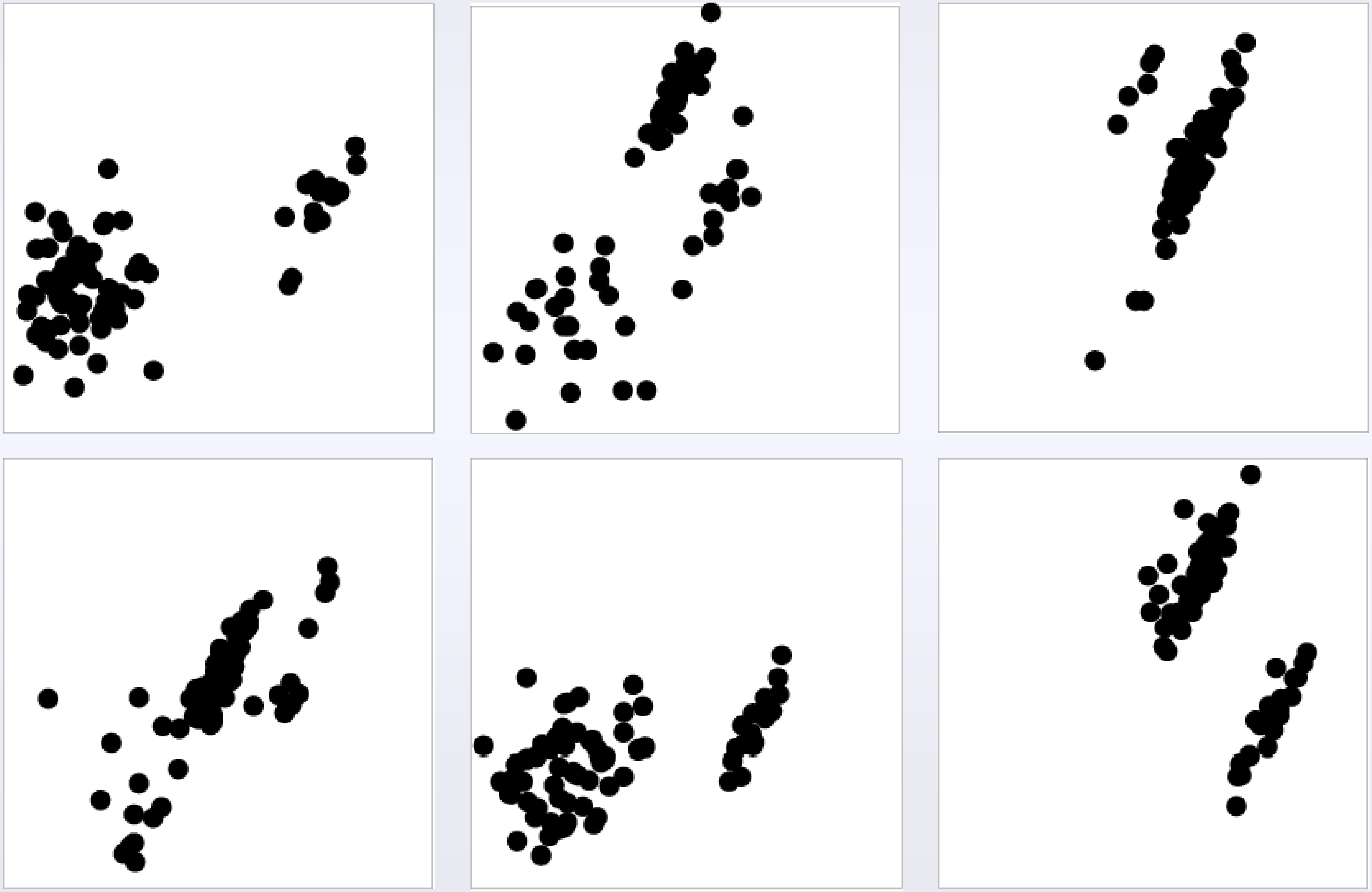
$$\tau((\mu,\Lambda);\rho) = (\mu + \rho, \Lambda)$$

Parameterized family of transformations

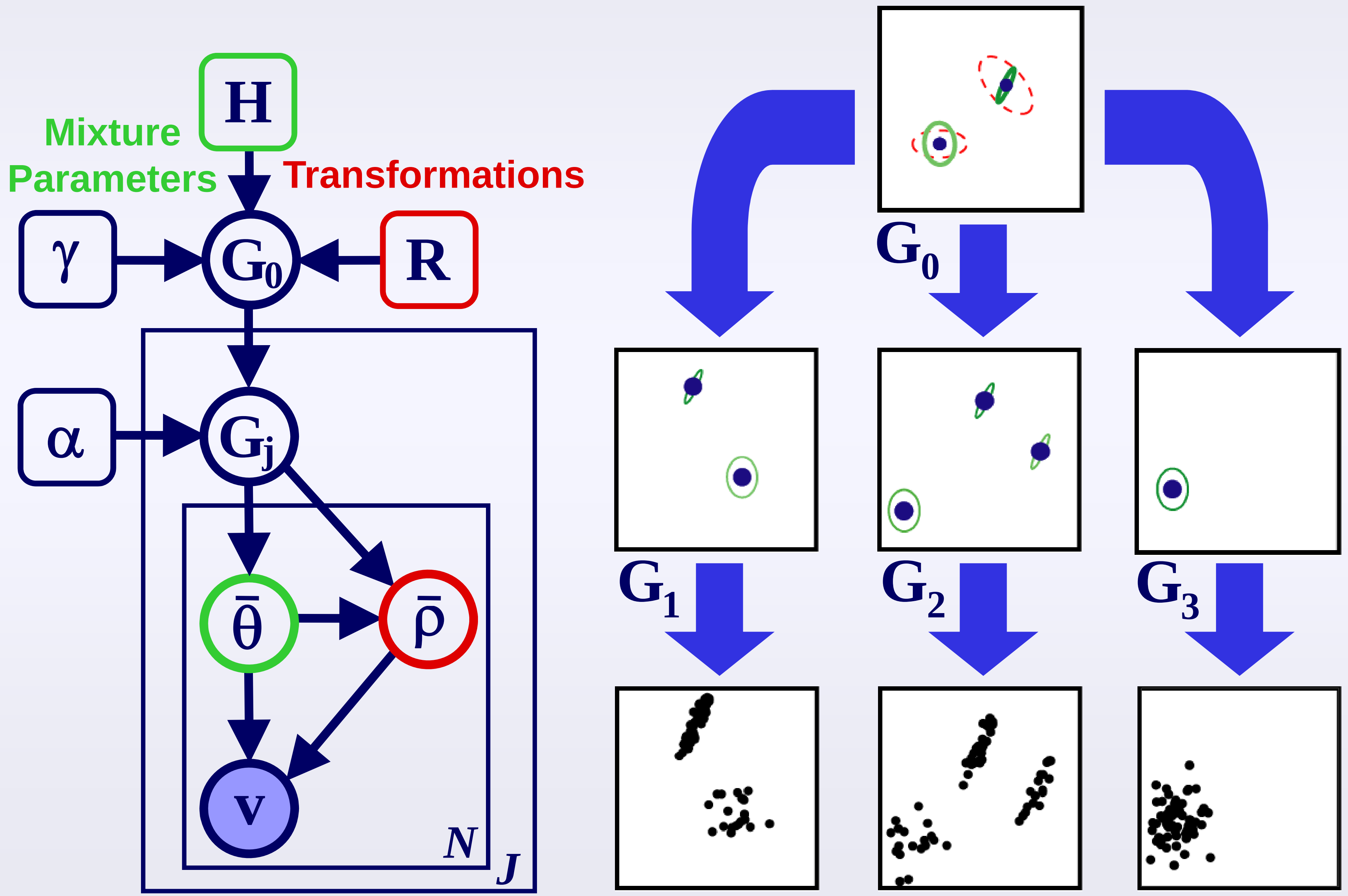
Shift cluster from canonical coordinate frame to object location in a given image

Layered Motion Models *(Wang & Adelson, Jojic & Frey)*
Nonparametric Transformation Densities *(Learned-Miller & Viola)*

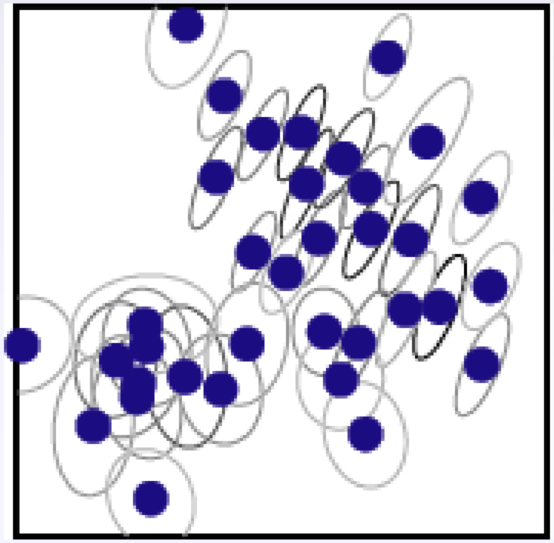
A Toy World: Bars & Blobs



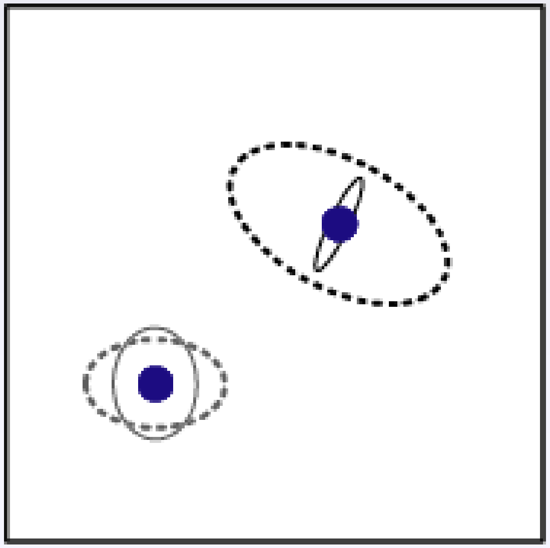
Transformed Dirichlet Process



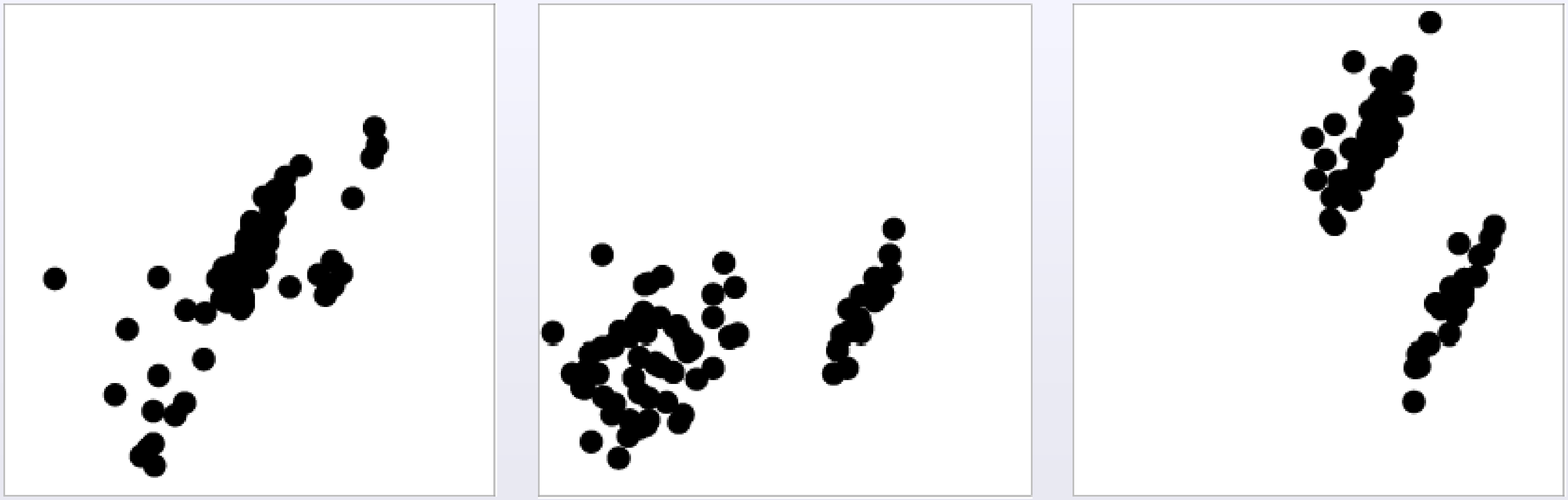
Importance of Transformations



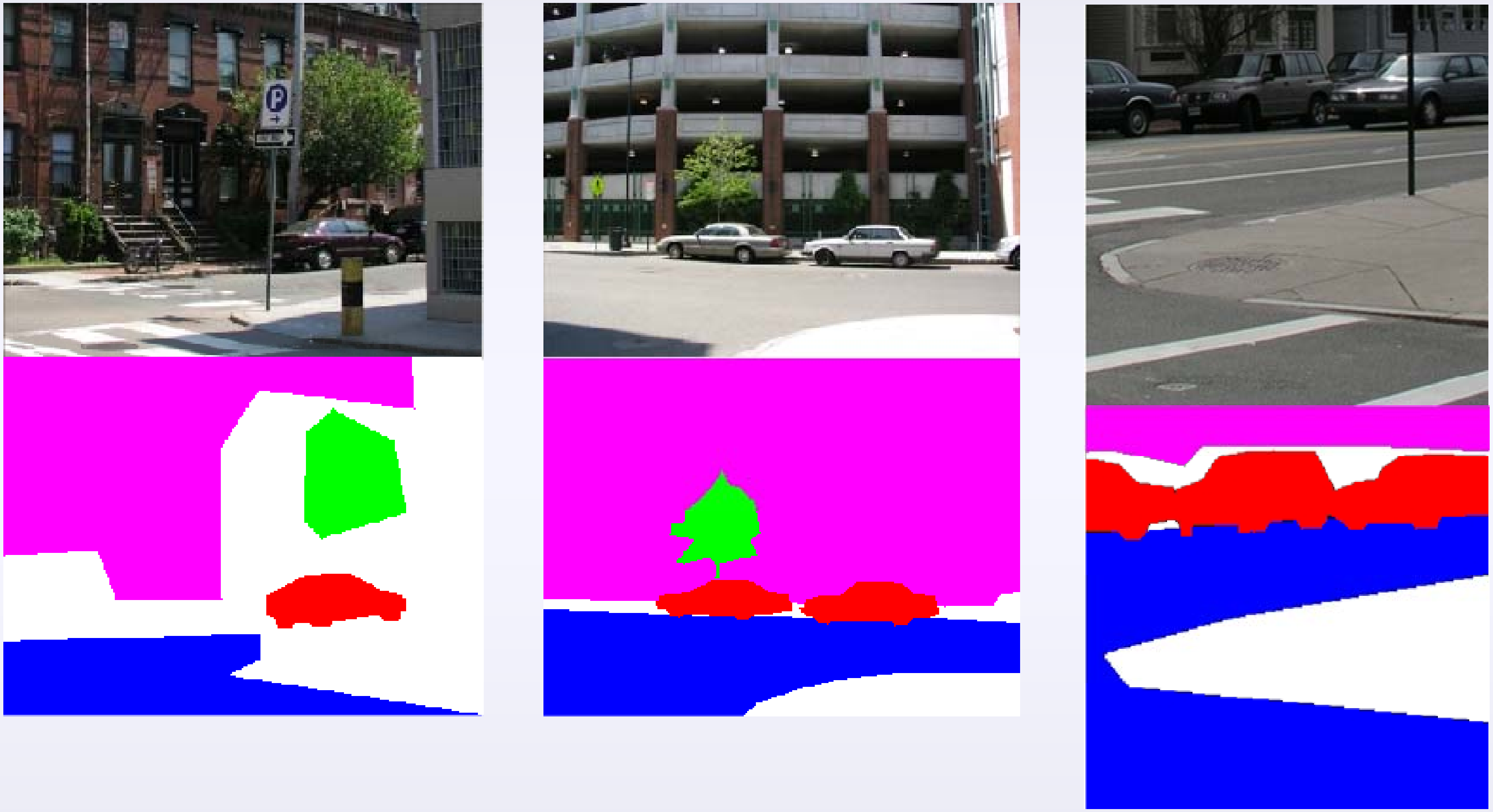
HDP



TDP



Counting & Locating Objects

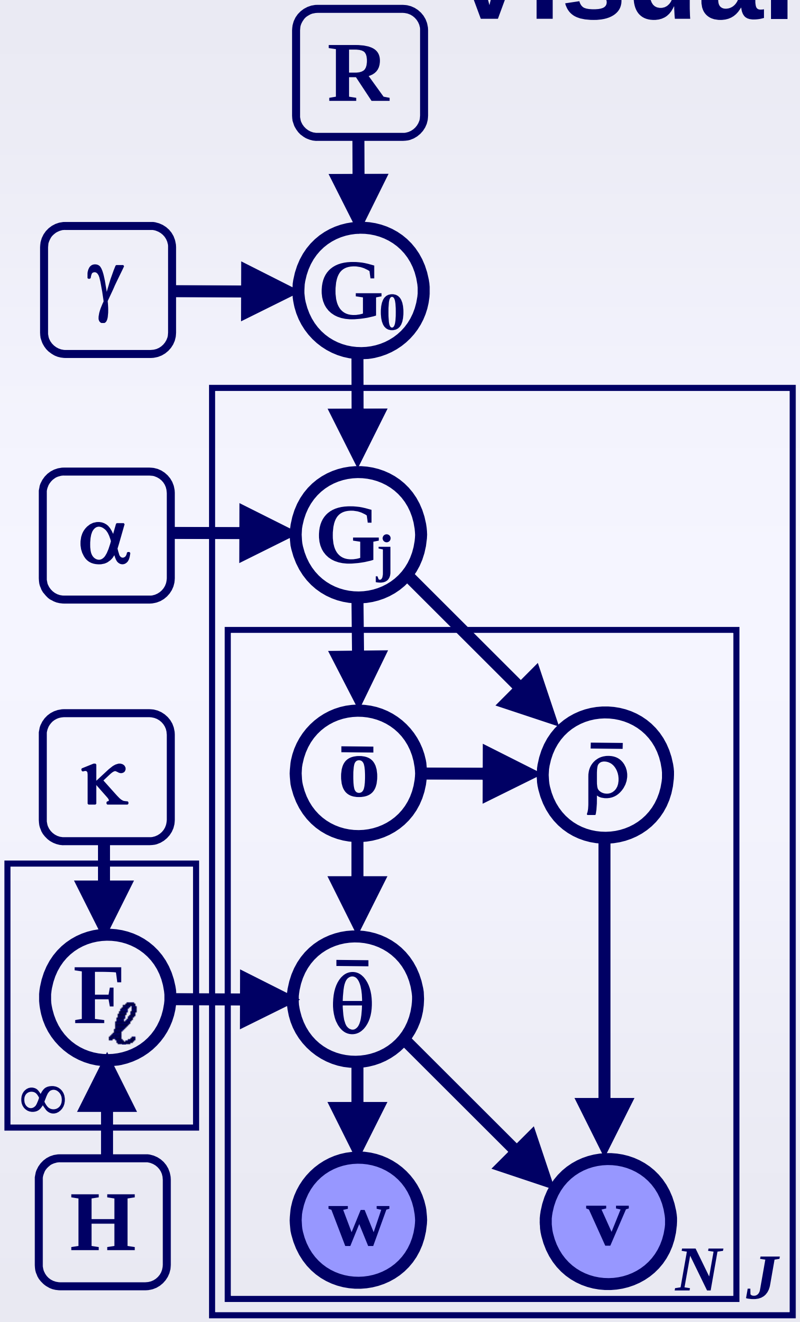


- How many cars are there?
- Where are those cars in the scene?

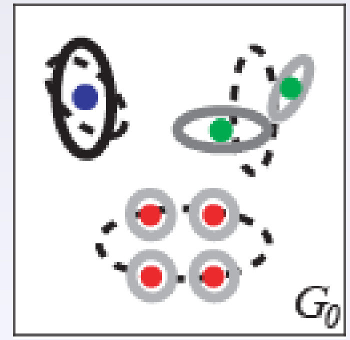
Dirichlet Processes

Transformations

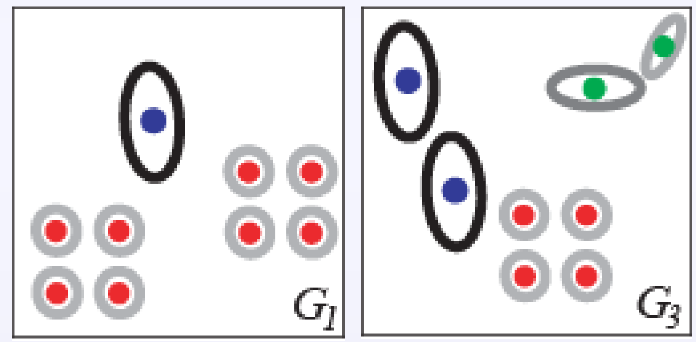
Visual Scene TDP



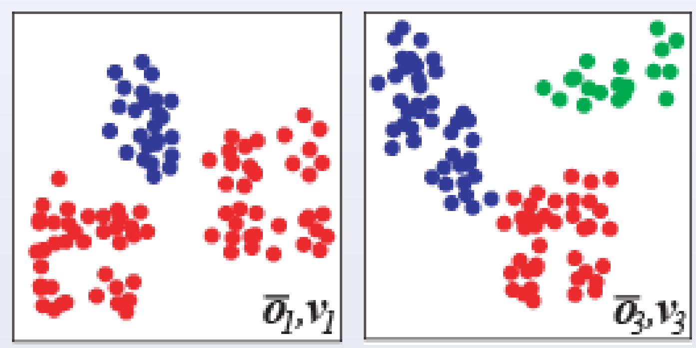
Global Density
Object category
Part size & shape
Transformation prior



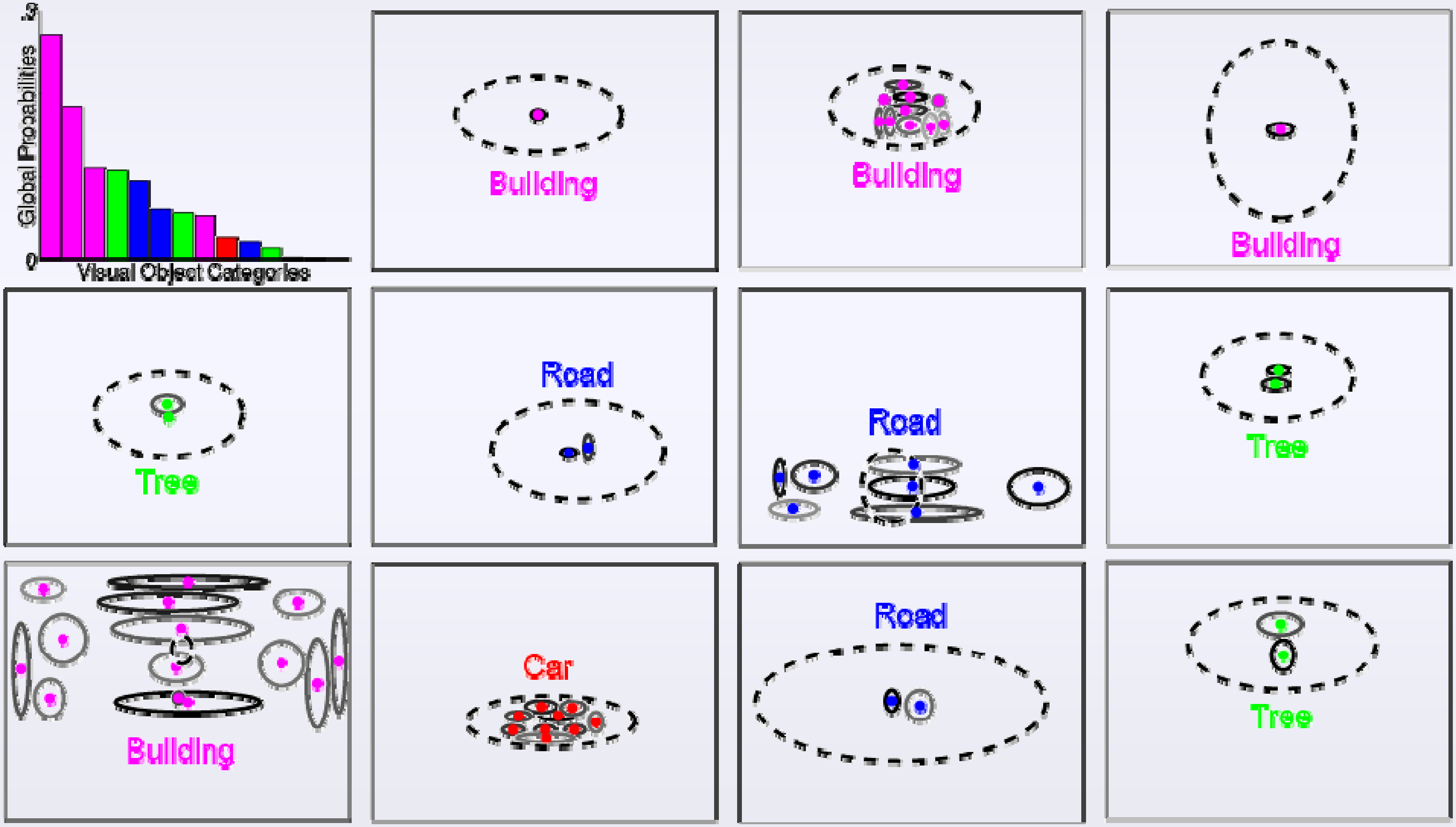
Transformed Densities
Object category
Part size & shape
Instance locations



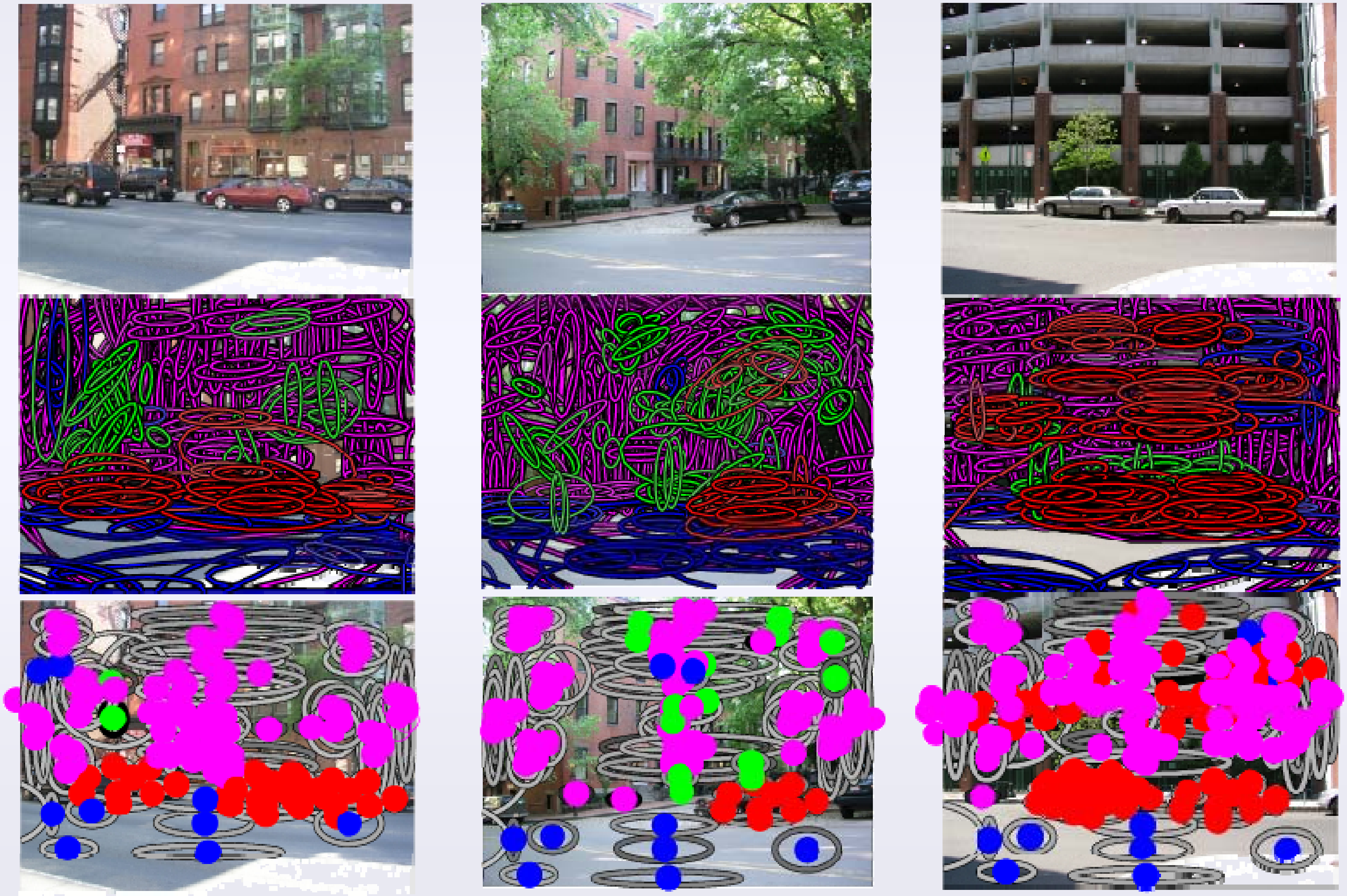
2D Image Features
Appearance
Location



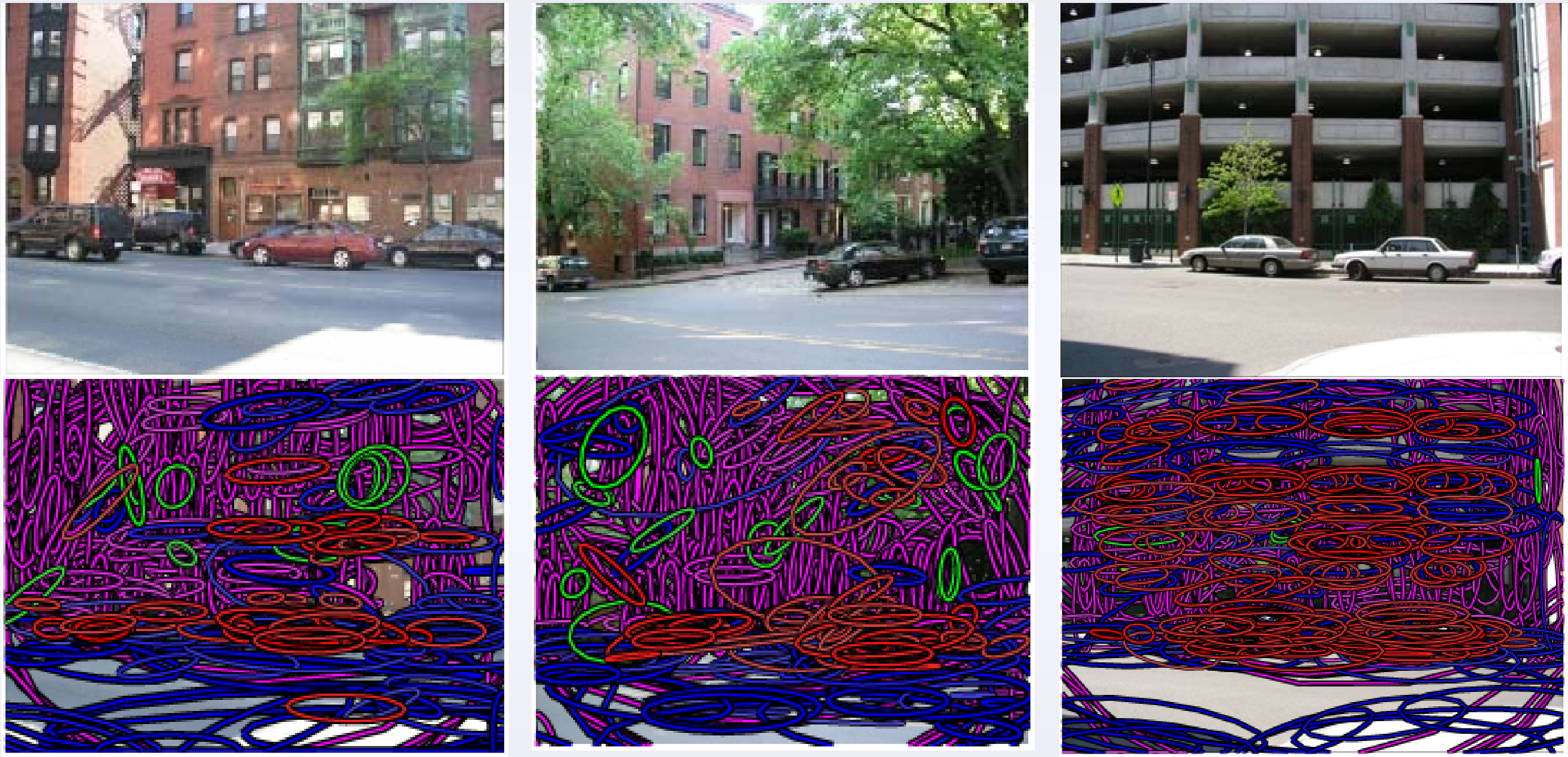
Street Scene Visual Categories



Street Scene Segmentations

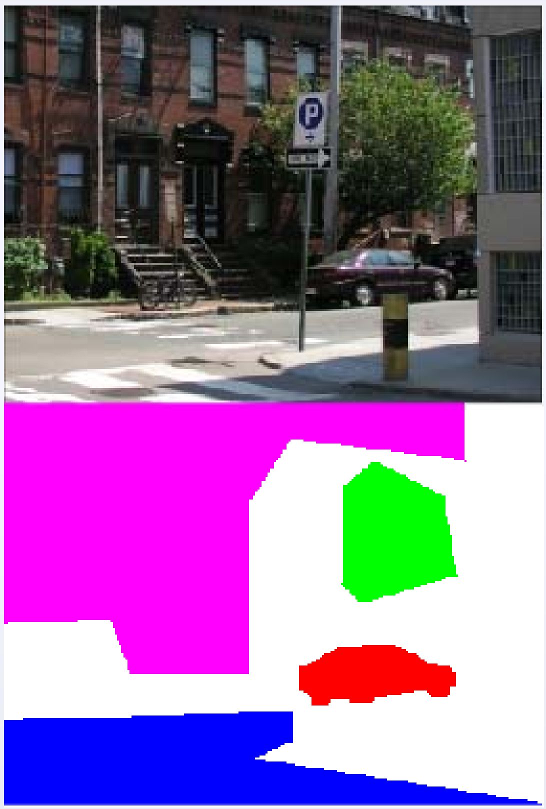


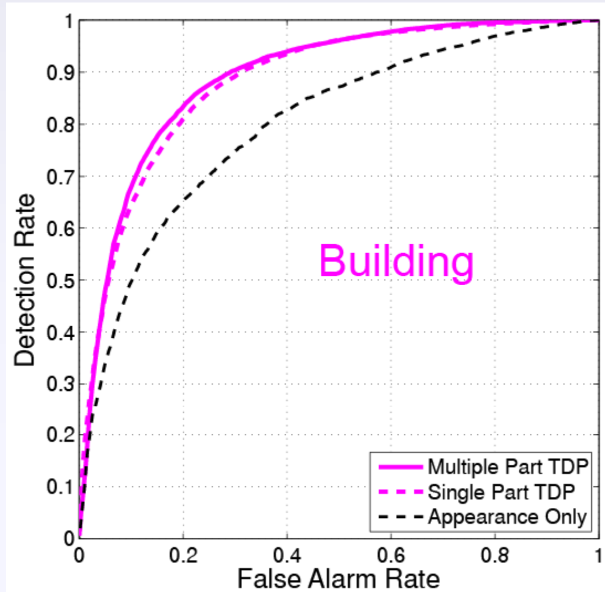
Appearance Only

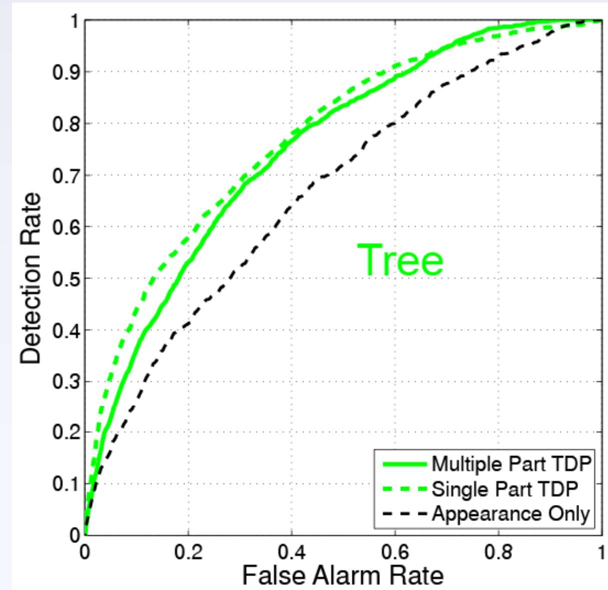


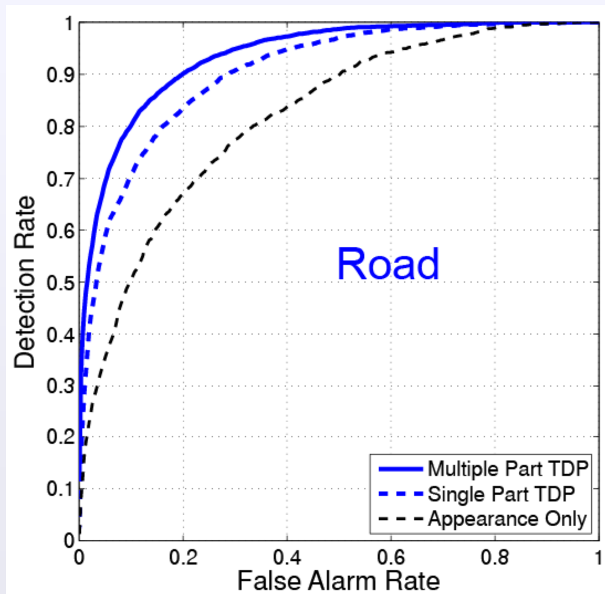
- “Bag of features” model, ignores feature positions
- Inferior segmentations, cannot count objects

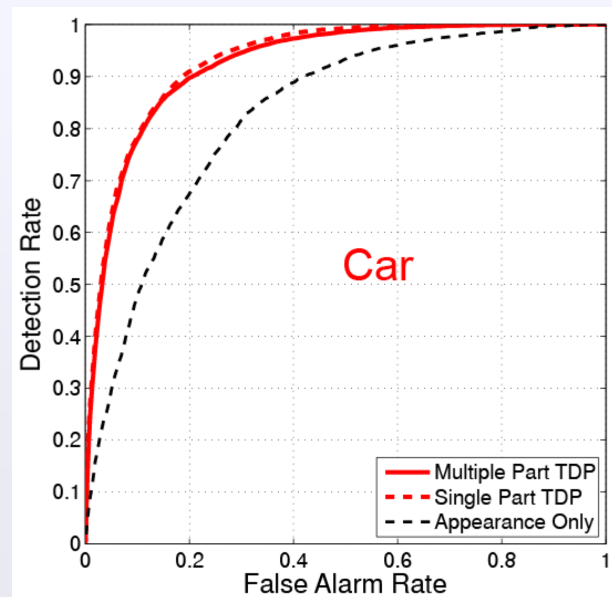
Segmentation Performance











Objects & 3D Reconstruction

An Office Scene

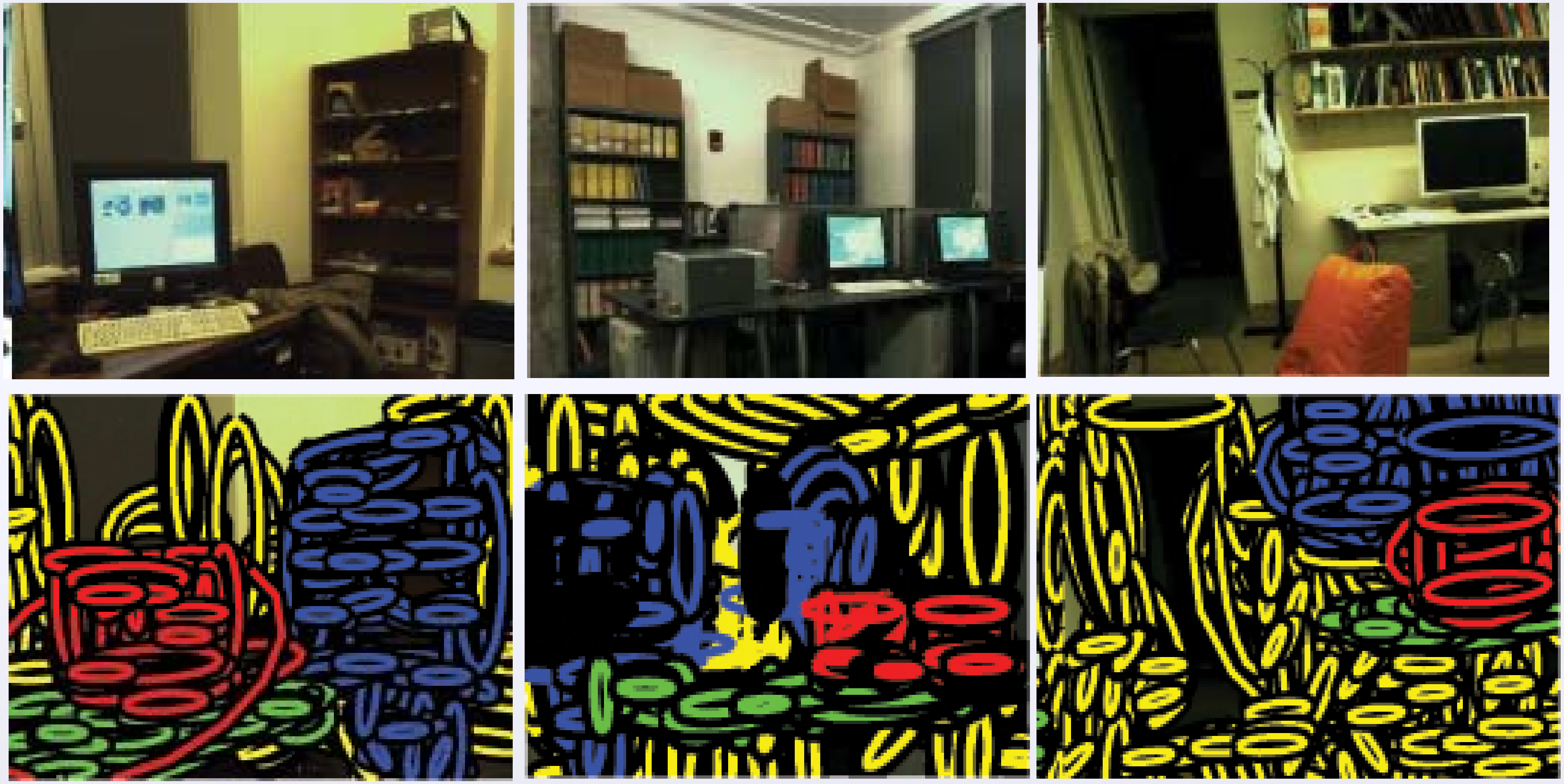
Green ⟷ *Near*

Red ⟷ *Far*

- Given 3D structure, segmentation is easier
- Identifying objects regularizes depth estimation

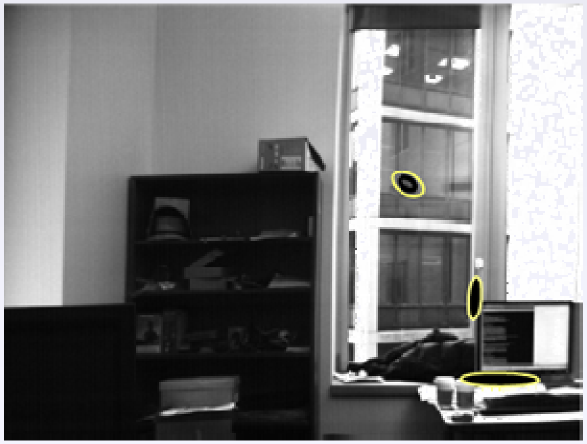
Office Scene Training Images

Objects at Multiple Scales

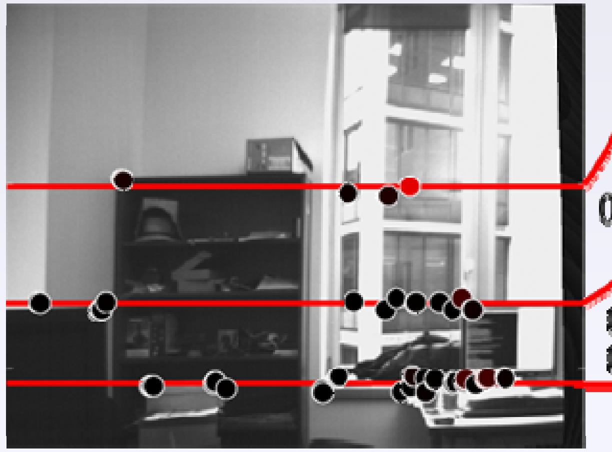


Computer Screens
Desks
Bookshelves

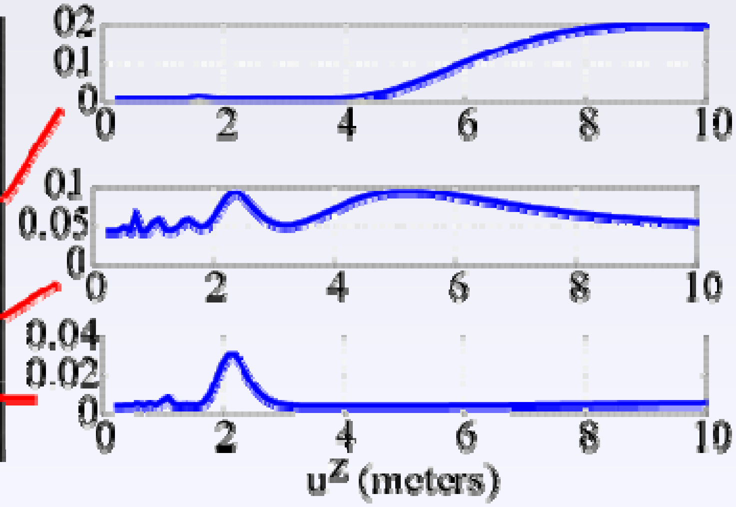
3D Structure from Stereo



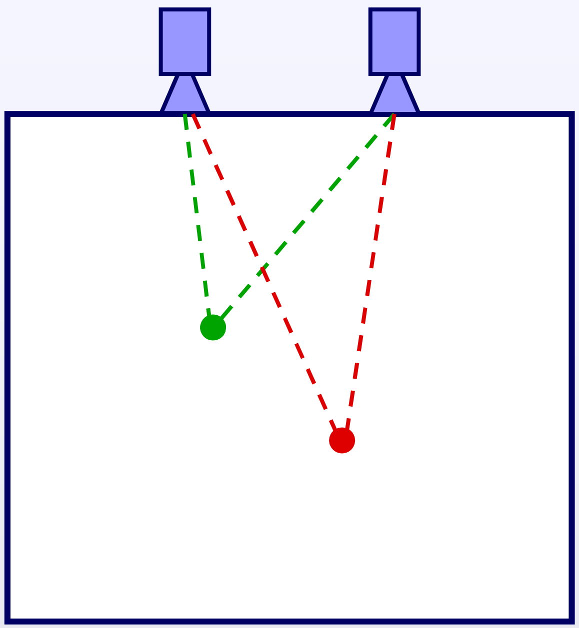
Reference (left) Image



Potential Matches



Depth Densities

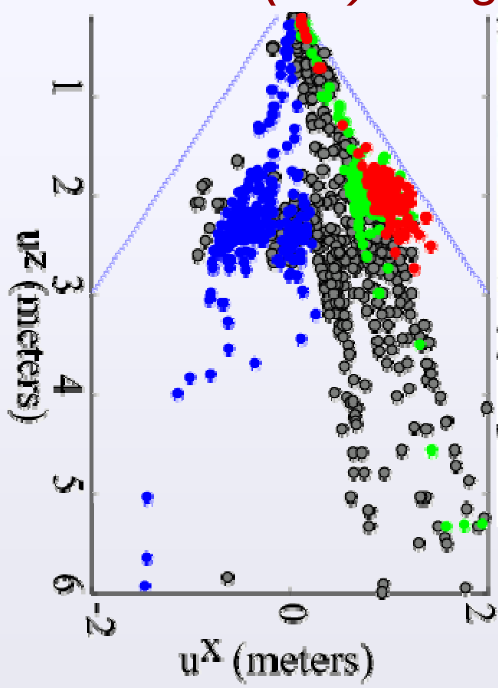


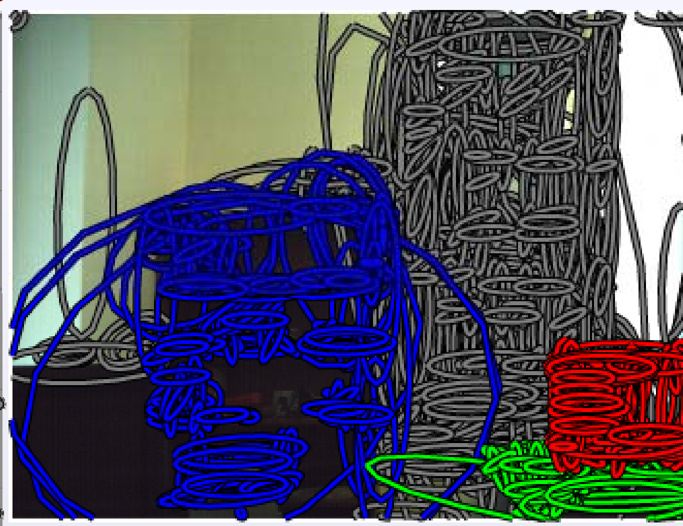
Overhead View

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

Greedy Depth Estimates

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

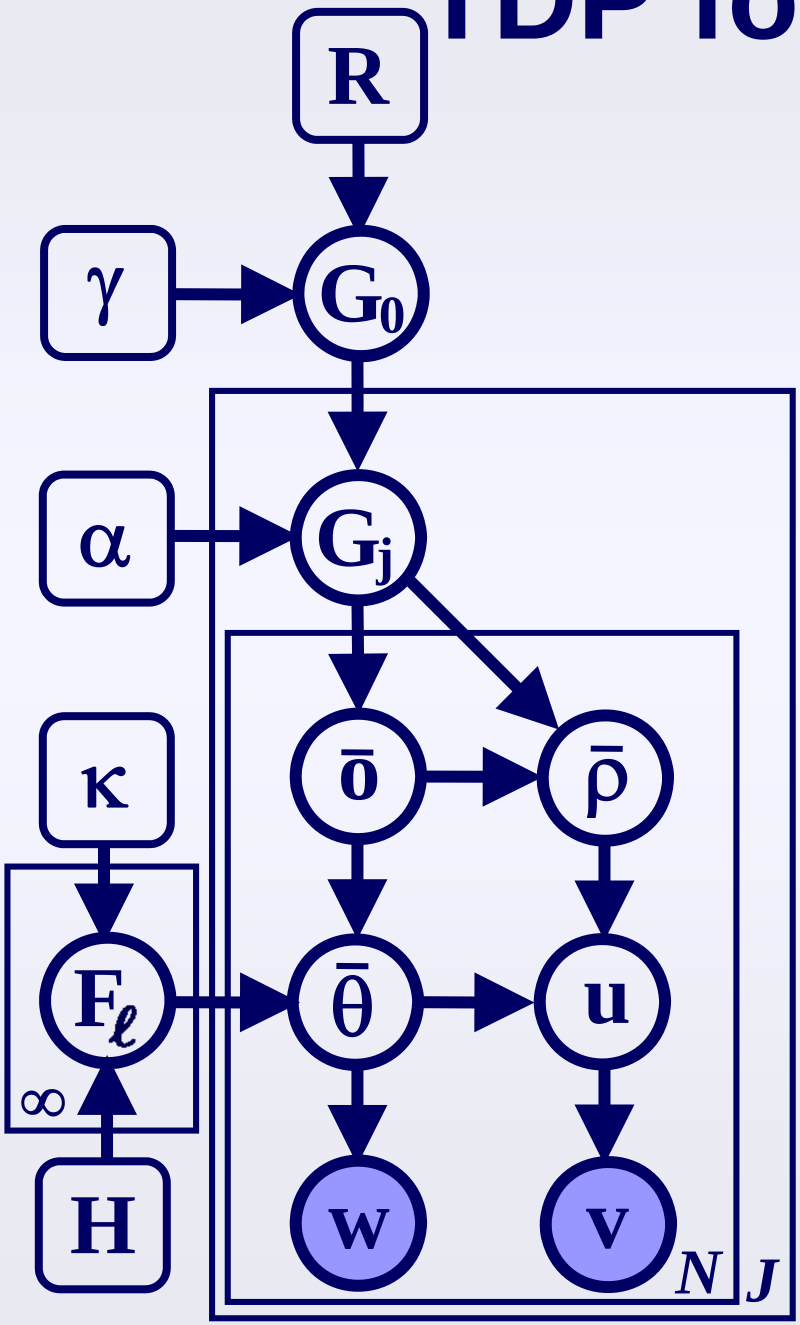




Green ⟷ Near

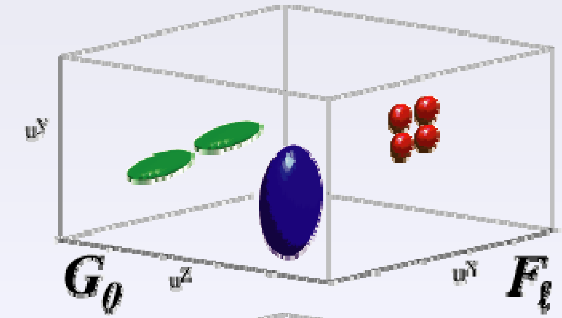
Red ⟷ Far

TDP for 3D Scenes



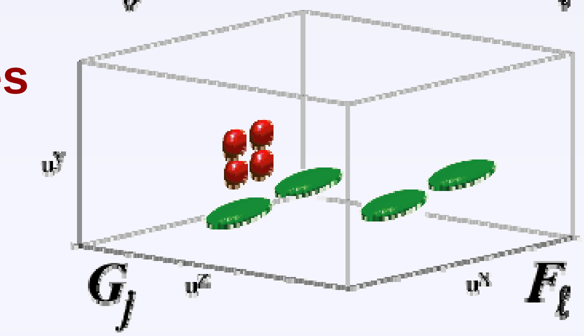
Global Density

Object category
Part size & shape
Transformation prior



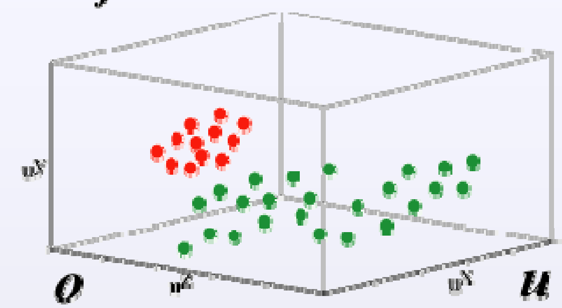
Transformed Densities

Object category
Part size & shape
Transformed locations



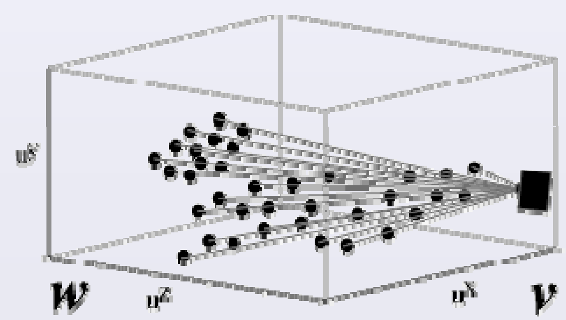
3D Scene Features

Object category
3D Location

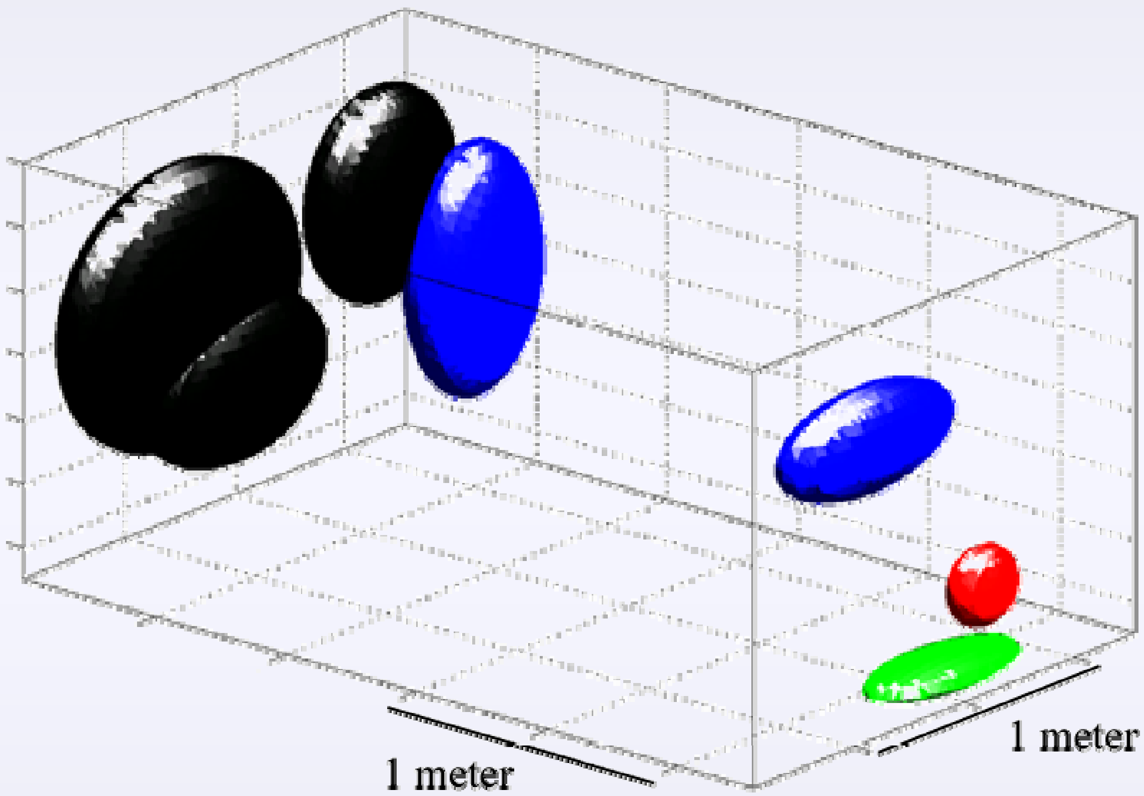


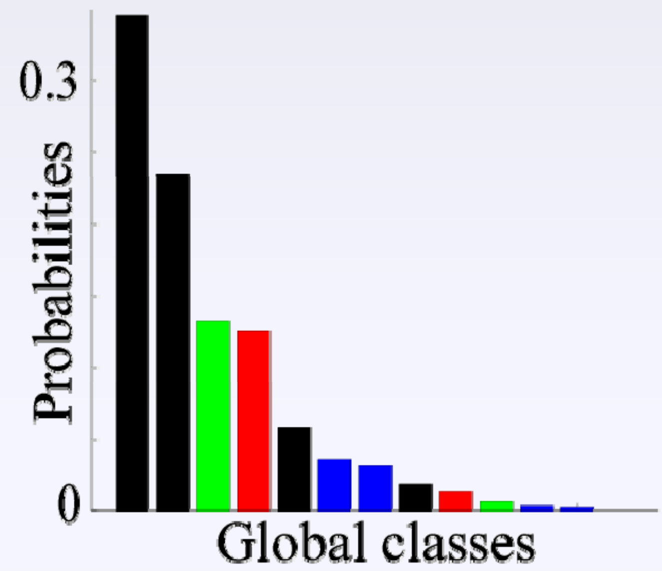
2D Image Features

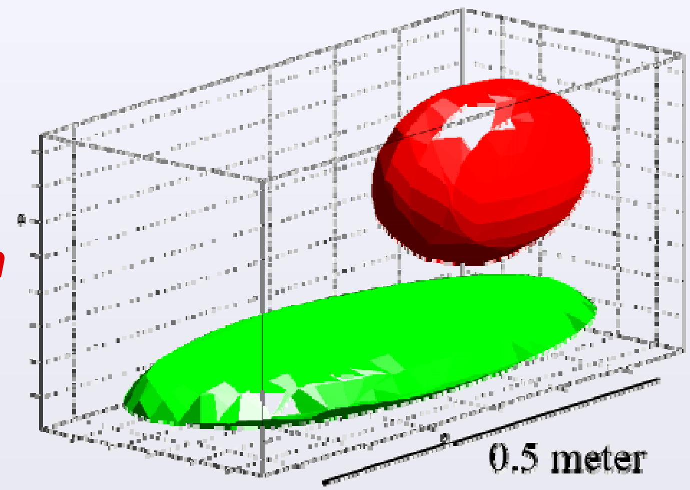
Appearance Descriptors
2D Pixel Coordinates



Single-Part Office Scene Model

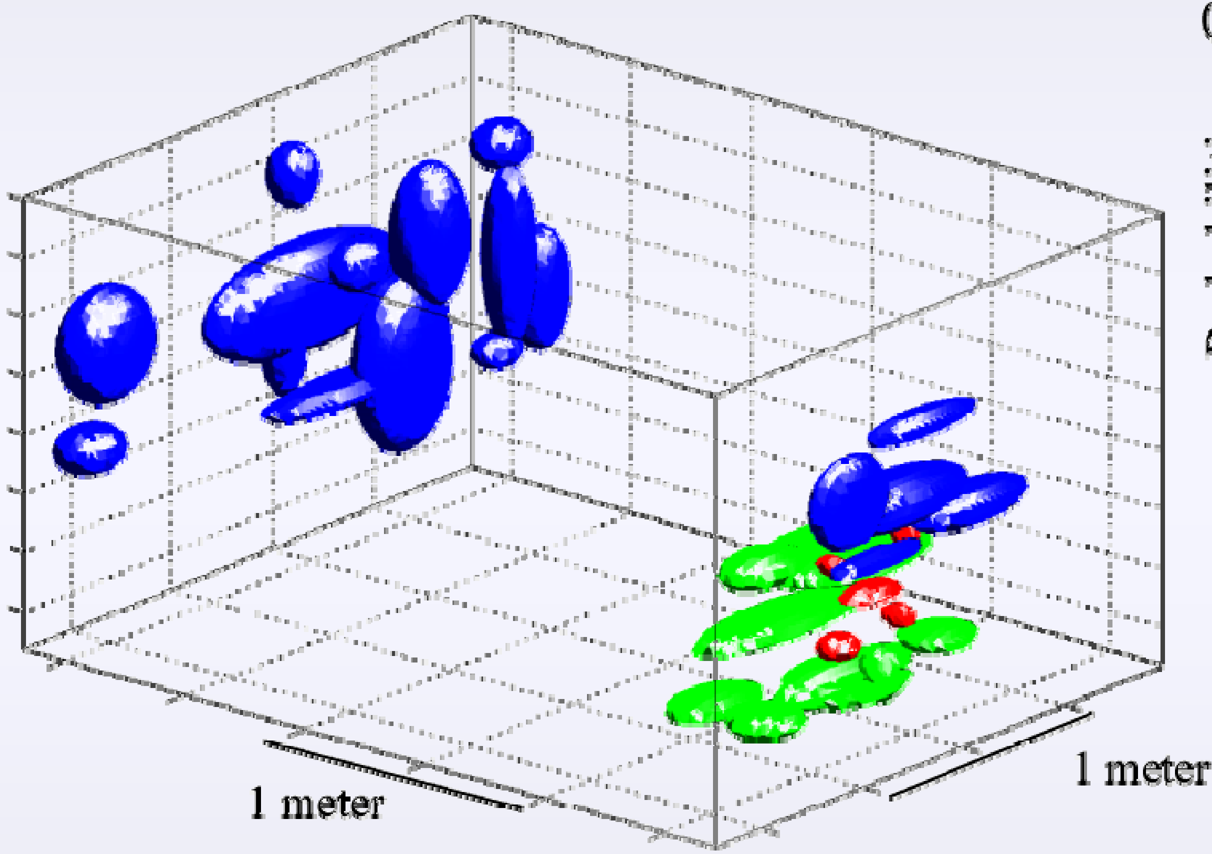


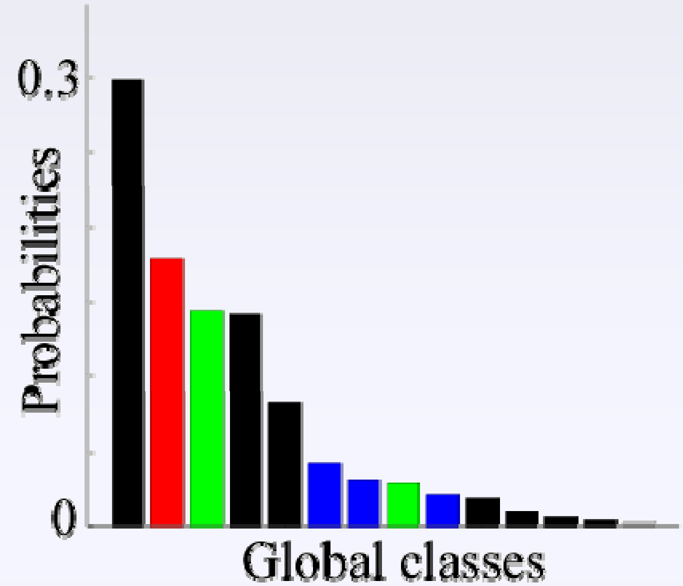




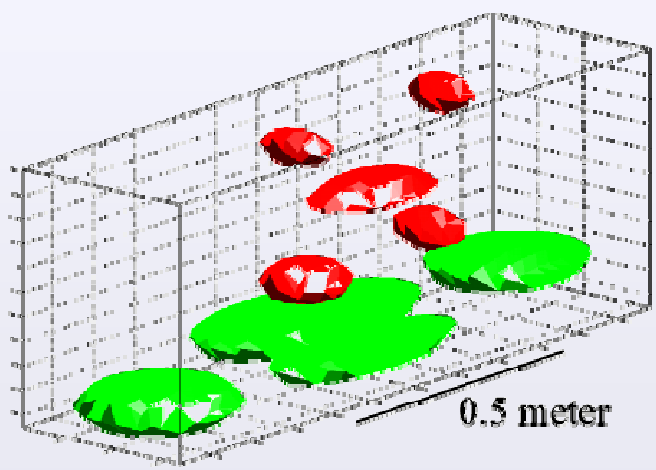
Background ***Bookshelves*** ***Computer Screen*** ***Desk***

Multi-Part Office Scene Model

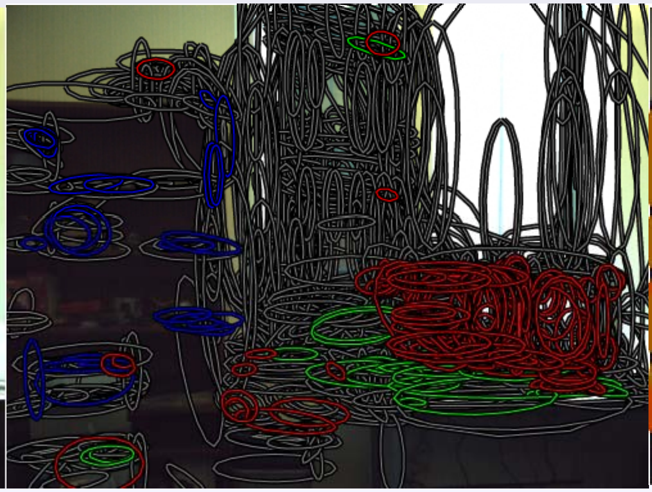


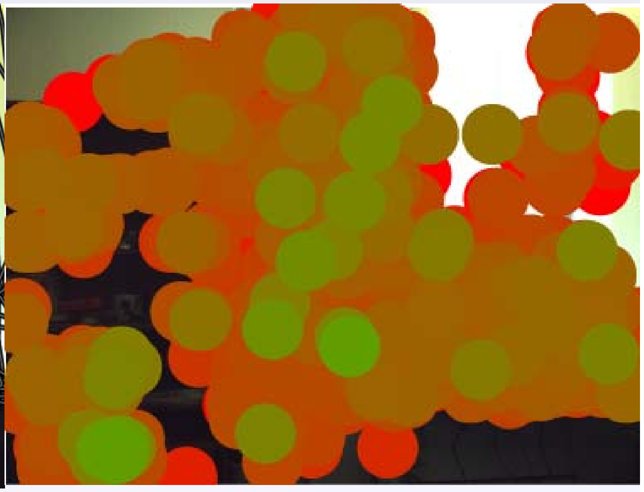


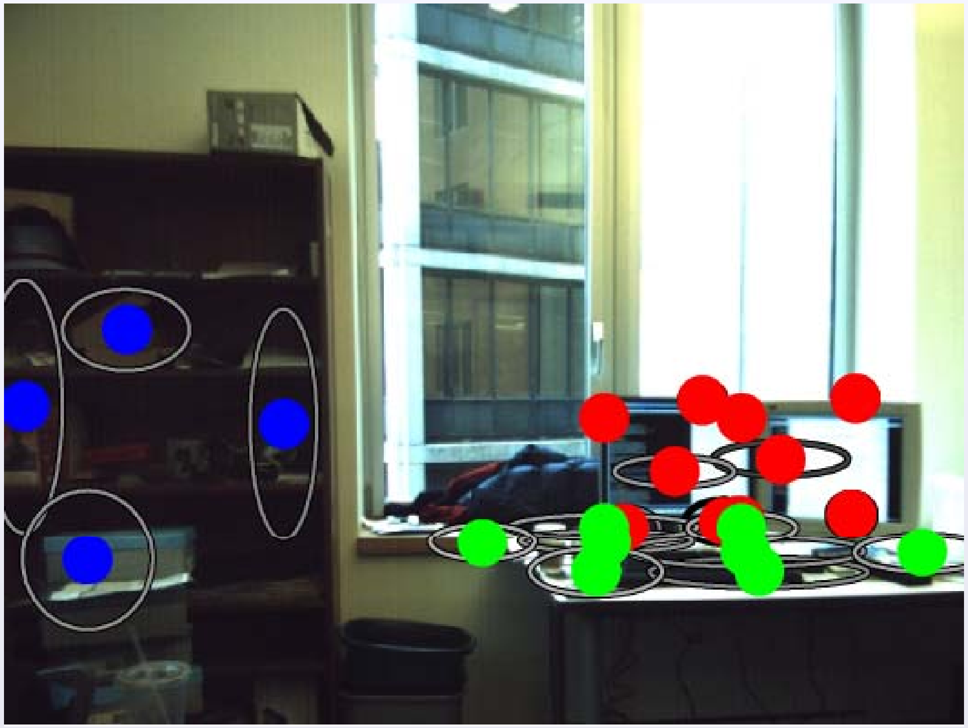
Background ***Bookshelves*** ***Computer Screen*** ***Desk***

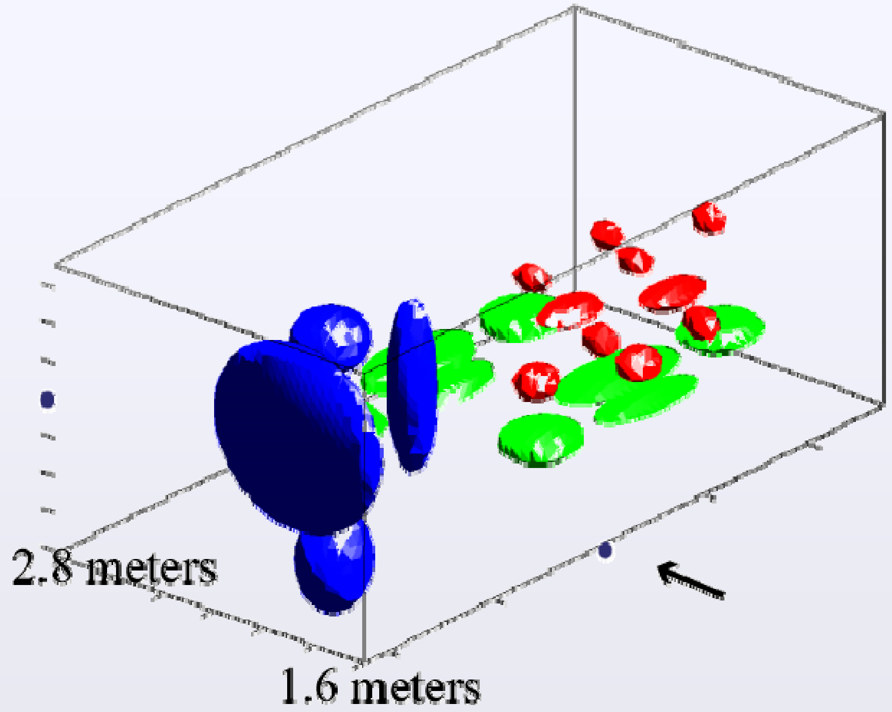


Stereo Test Image I

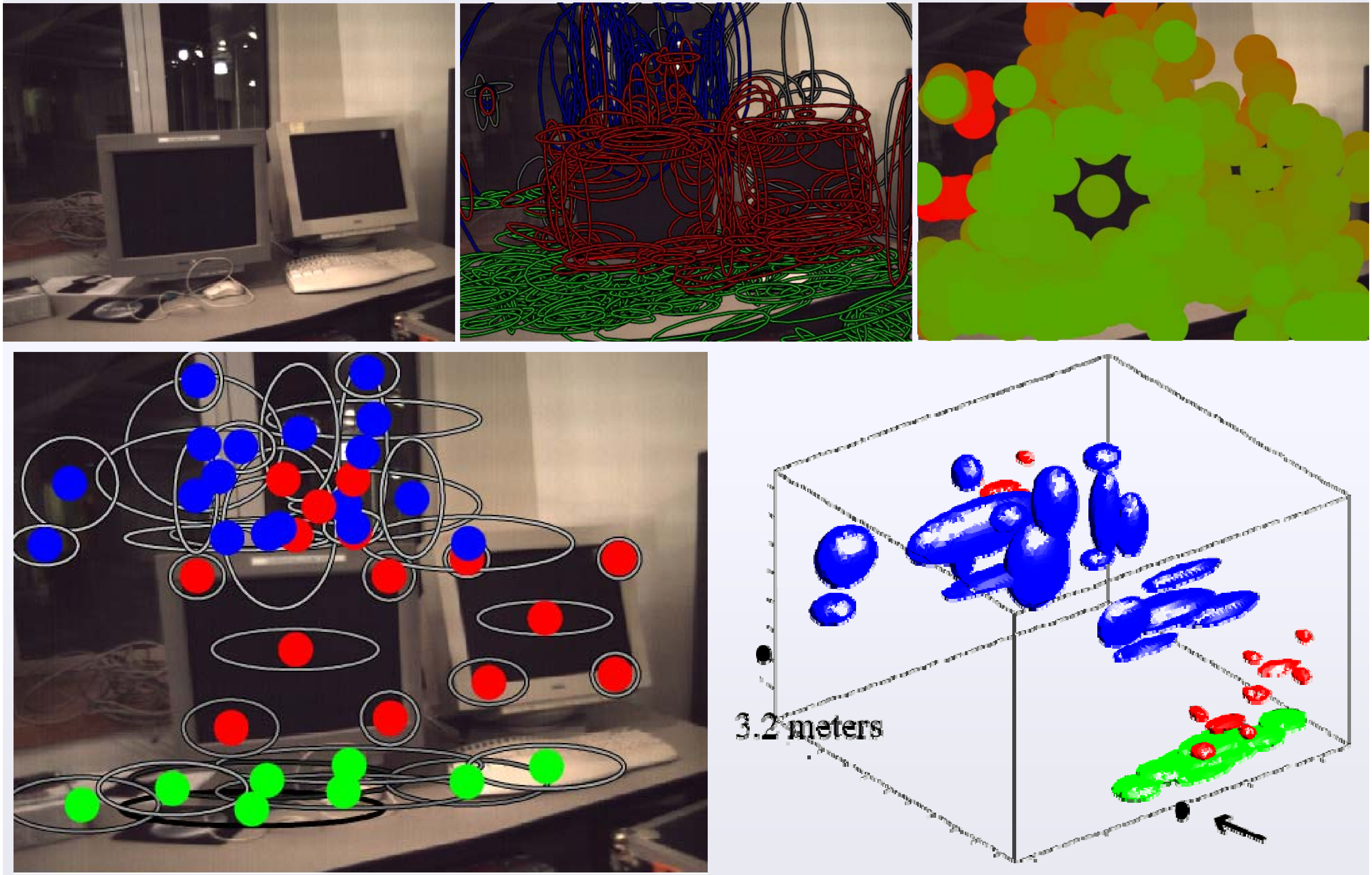




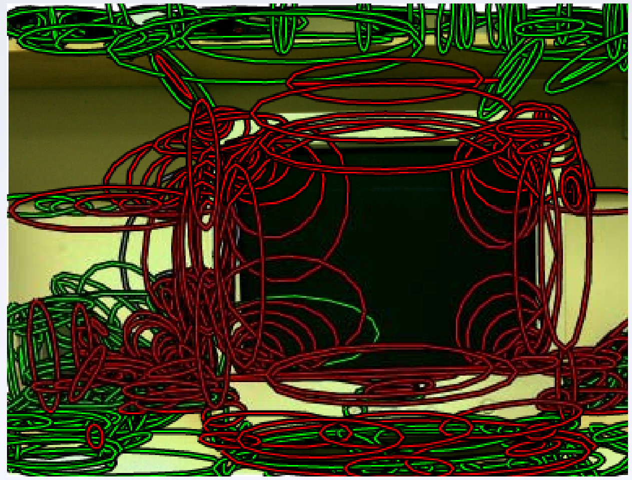


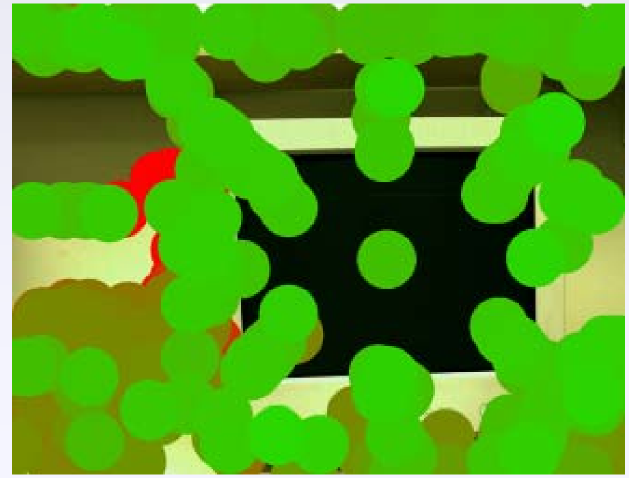


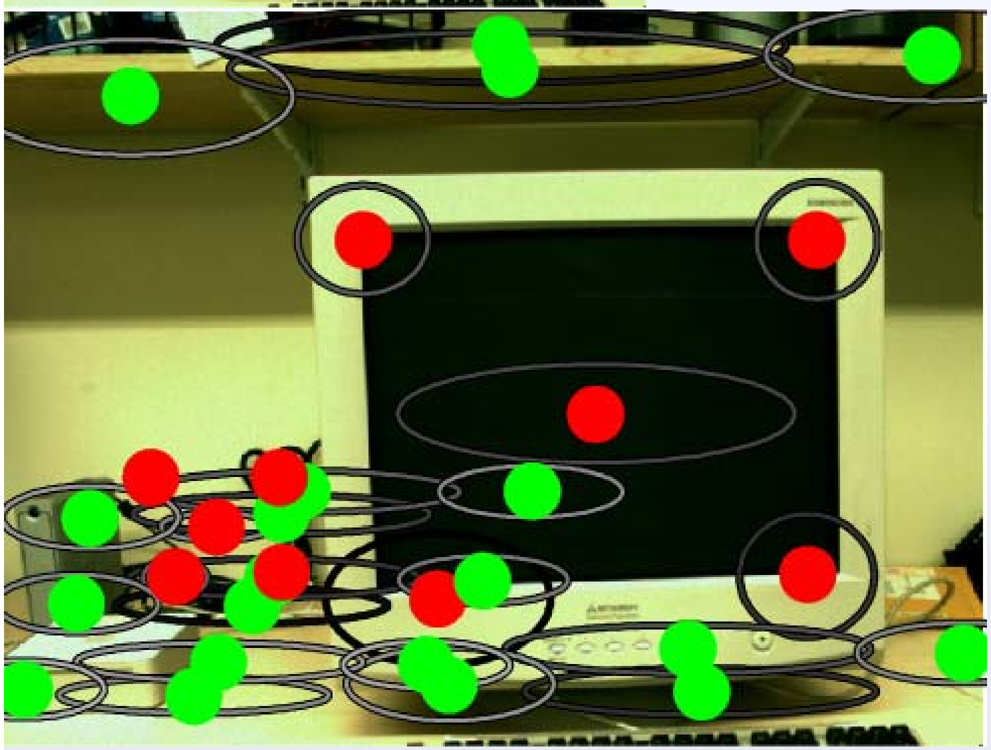
Stereo Test Image II

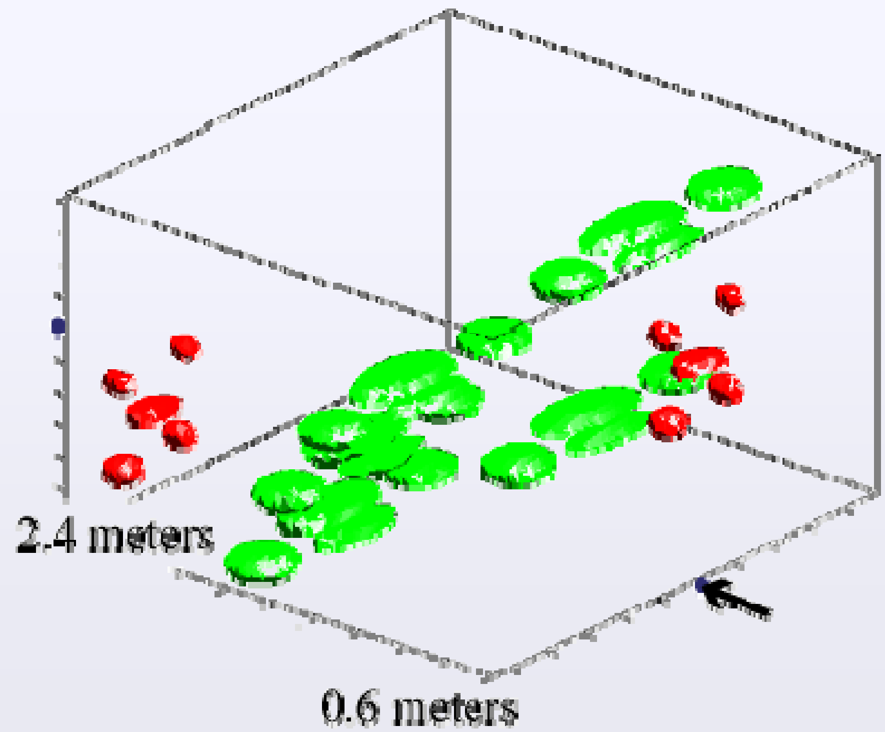


Ongoing Work: Monocular Test

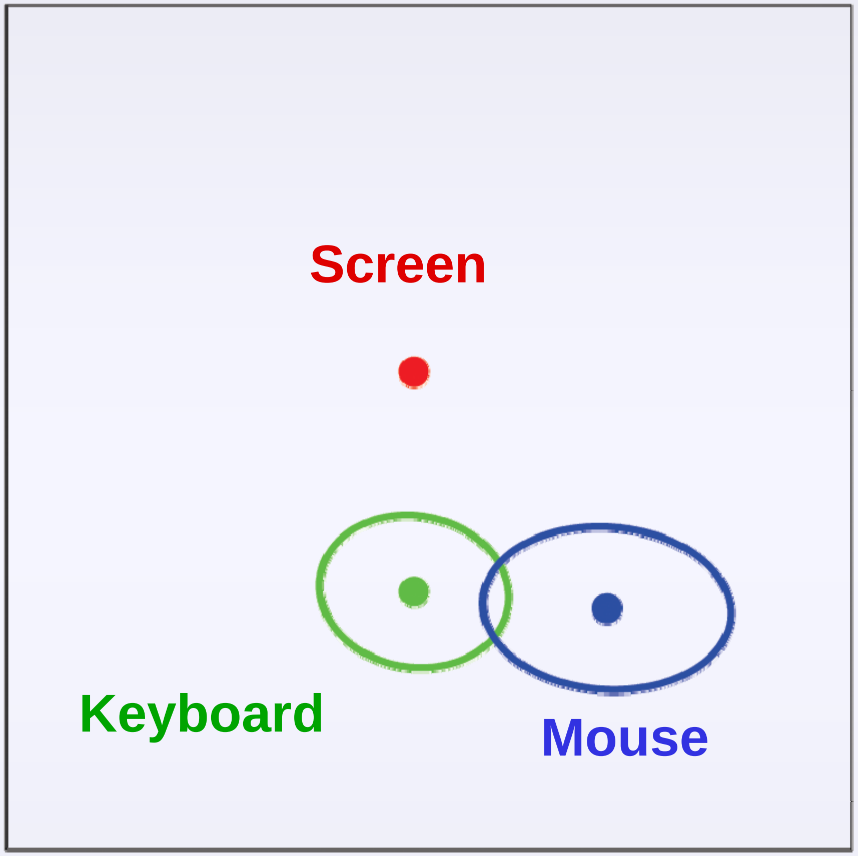








Ongoing Work: Context

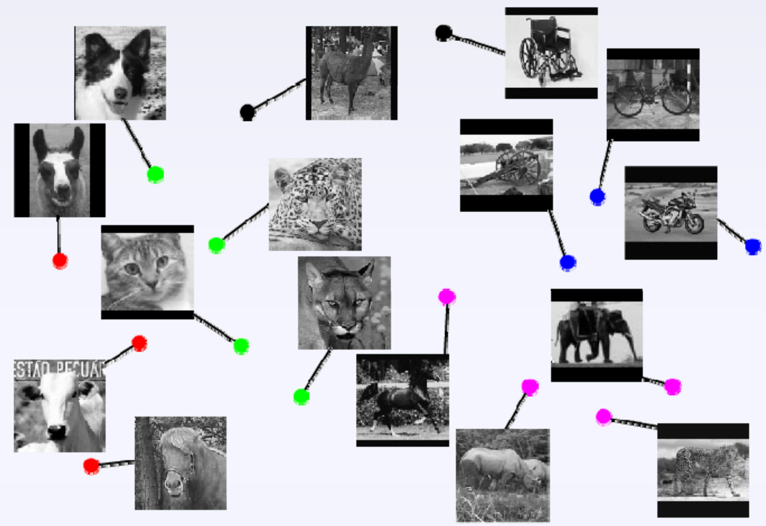


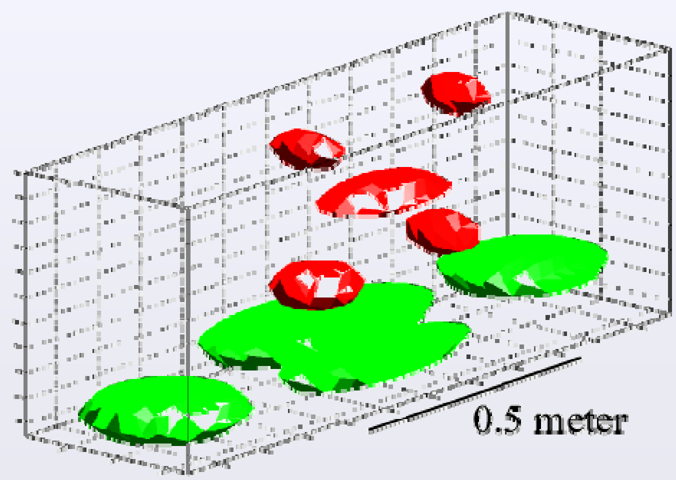
- Developed *fixed-order* contextual scene model
- Extension to Transformed DP model is an open problem
- Needed: Richer models for *background* scene structure

Sudderth Conclusions

Transformed Dirichlet Processes allow…

- flexible *transfer* of knowledge among related object categories
- robust learning from small, *partially labeled* datasets
- *an integrated* view of object recognition & 3D reconstruction
- potential *scaling* of nonparametric methods to complex domains





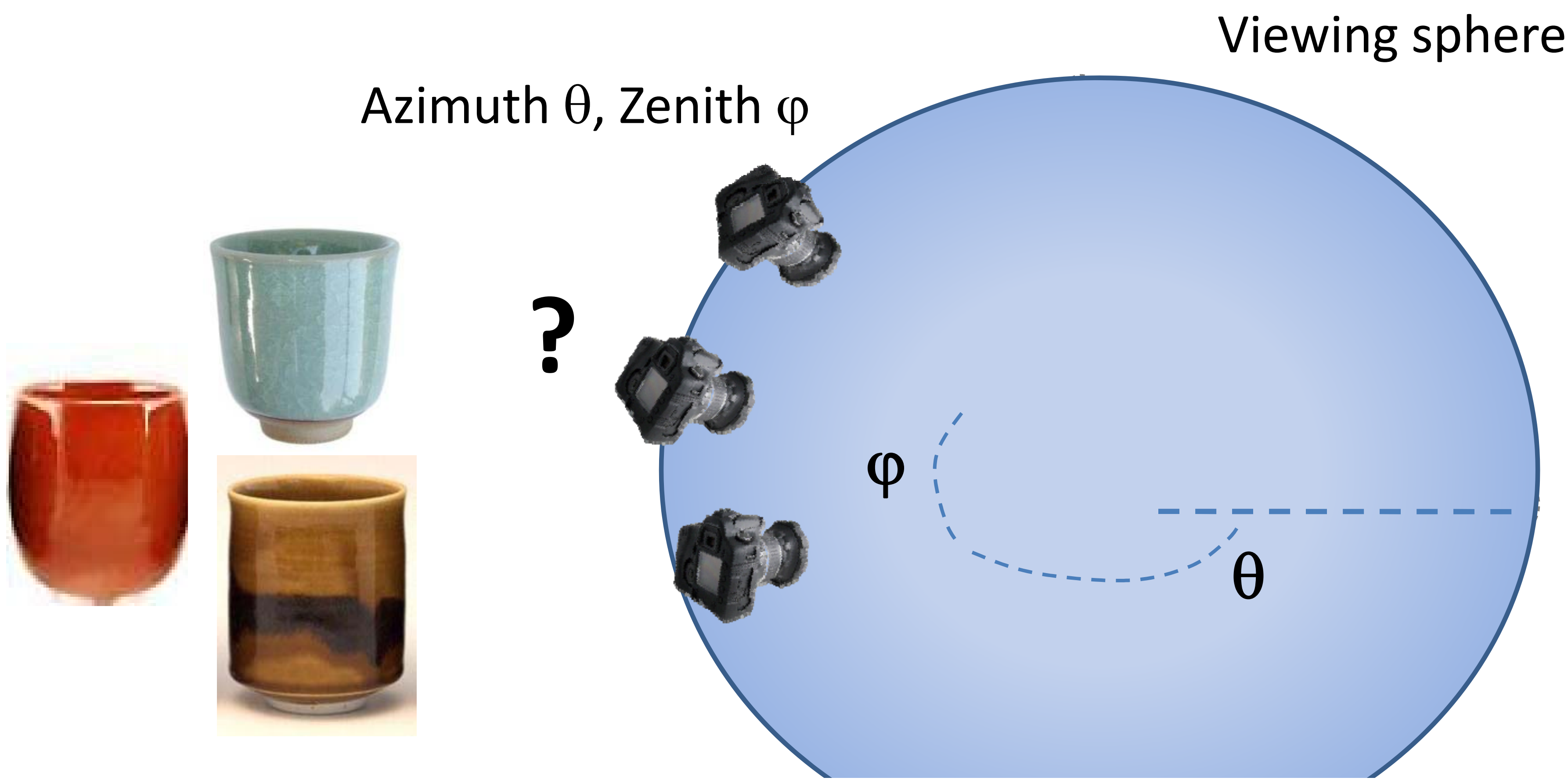
Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories

*H. Su *M. Sun L. Fei-Fei and S. Savarese.

Min Sun
University of Michigan, USA

Hao Su
Beihang University, China
Stanford, USA

Our goals



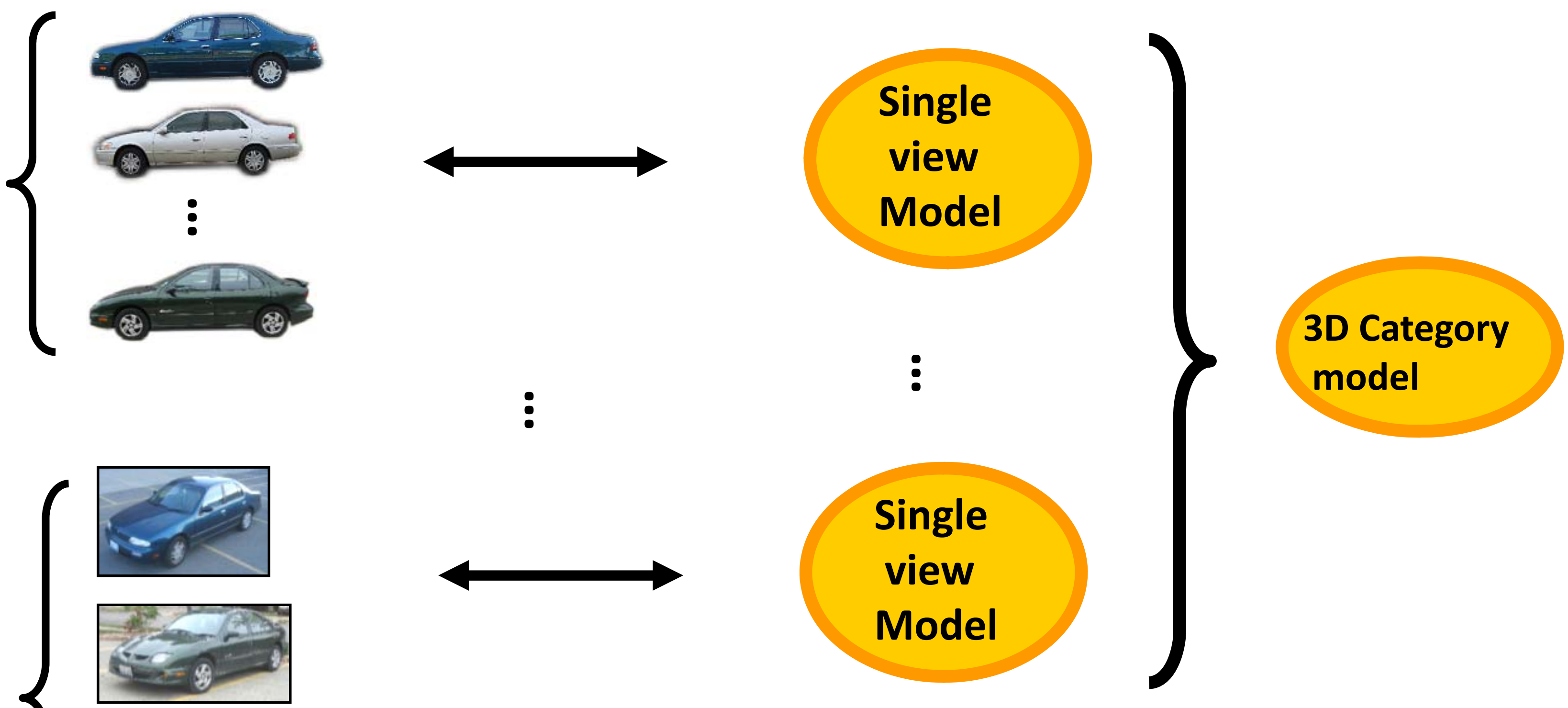
- Detect objects under generic view points
- Estimate object pose
- Predict object appearance from novel views

Our goals

- Detect objects under generic view points
- Estimate object pose
- Predict object appearance from novel views
- Generic and work for any category

Current paradigm

- Leung et al '99
- Weber et al. '00
- Schneiderman et al. '01
- Ullman et al. 02
- Fergus et al. '03
- Torralba et al. '03
- Felzenszwalb & Huttenlocher '03
- Fei-Fei et al. '04
- Bart et al '04
- Leibe et al. '04
- Kumar & Hebert '04
- Sivic et al. '05
- Shotton et al '05
- Grauman et al. '05
- Sudderth et al '05
- Torralba et al. '05
- Lazebnik et al. '06
- Todorovic et al. '06
- Bosh et al '07
- Vedaldi & Soatto '08
- Zhu et al 08



- Single view models are independent
  - No information is shared [except Torralba et al. '03]
  - No sense of correspondences of parts under 3D transformations
- Non scalable to large number of categories/view-points

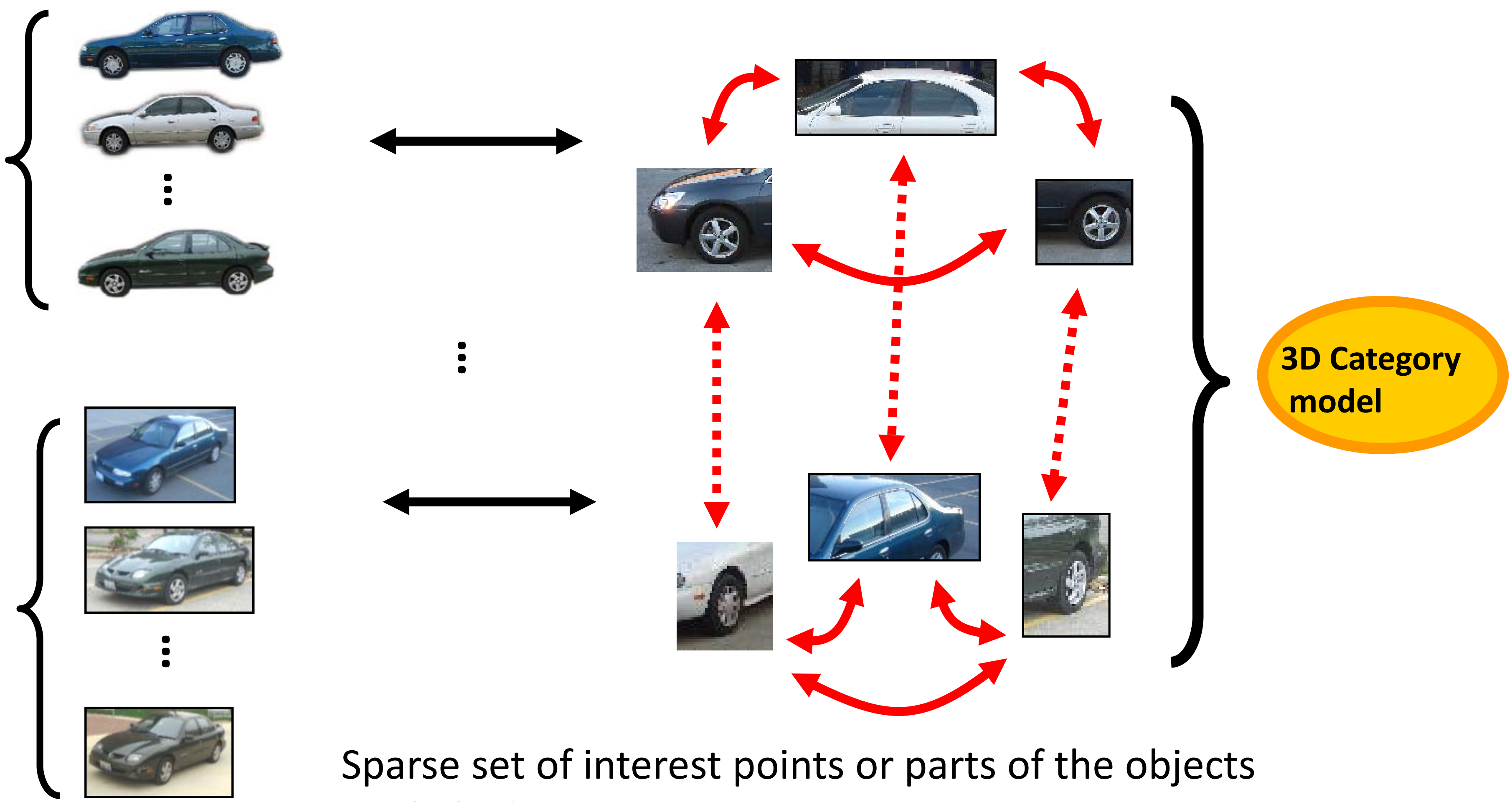
A new recent paradigm

- Thomas et al. '06
- Kushal, et al., '07
- Savarese et al, 07, 08

- Chiu et al. '07
- Hoiem, et al., '07
- Yan, et al. '07

- Liebelt et al., '08
- Xiao et al.,'08
- Liebelt et al., '08
- Xiao et al.,'08

- Sun et al 09
- Farhadi let al ICCV 09
- Arie-Nachimson & Basri, 'ICCV 09

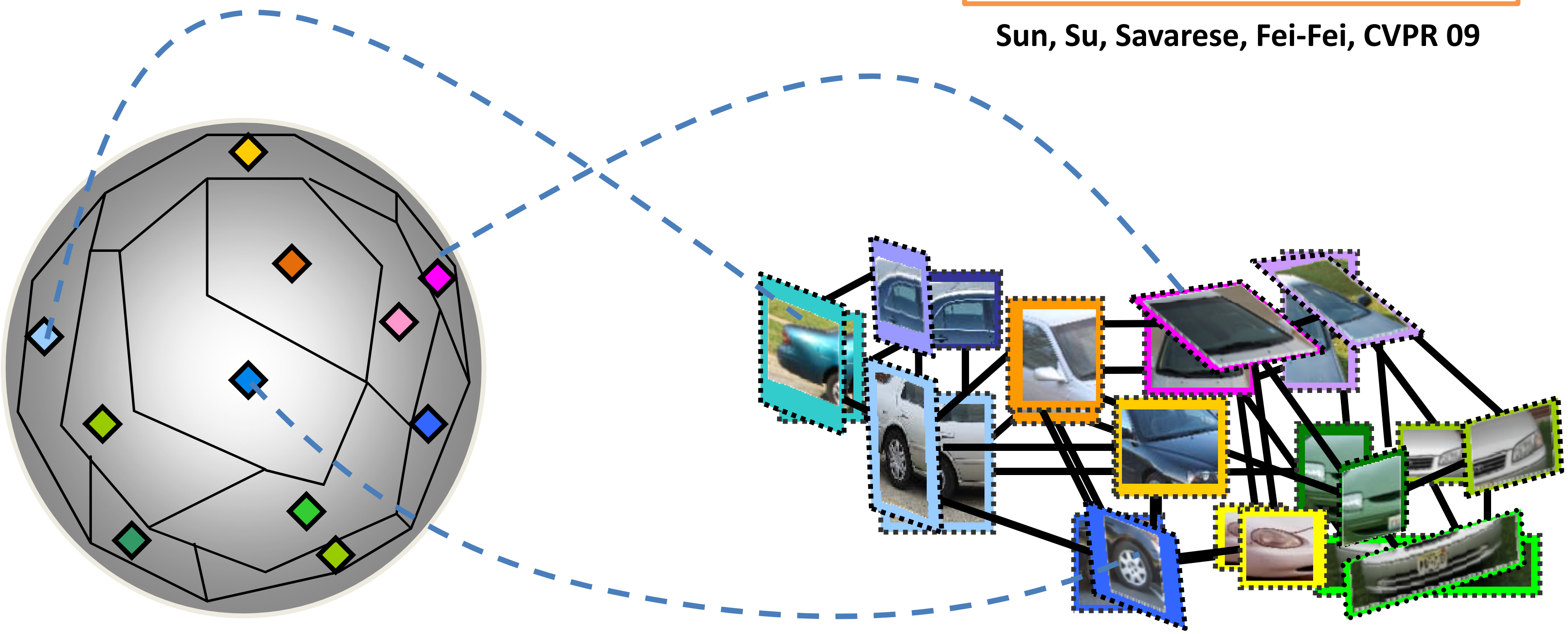


Sparse set of interest points or parts of the objects are linked across views.

A new recent paradigm

Savarese, Fei-Fei, ICCV 07
Savarese, Fei-Fei, ECCV 08

Sun, Su, Savarese, Fei-Fei, CVPR 09



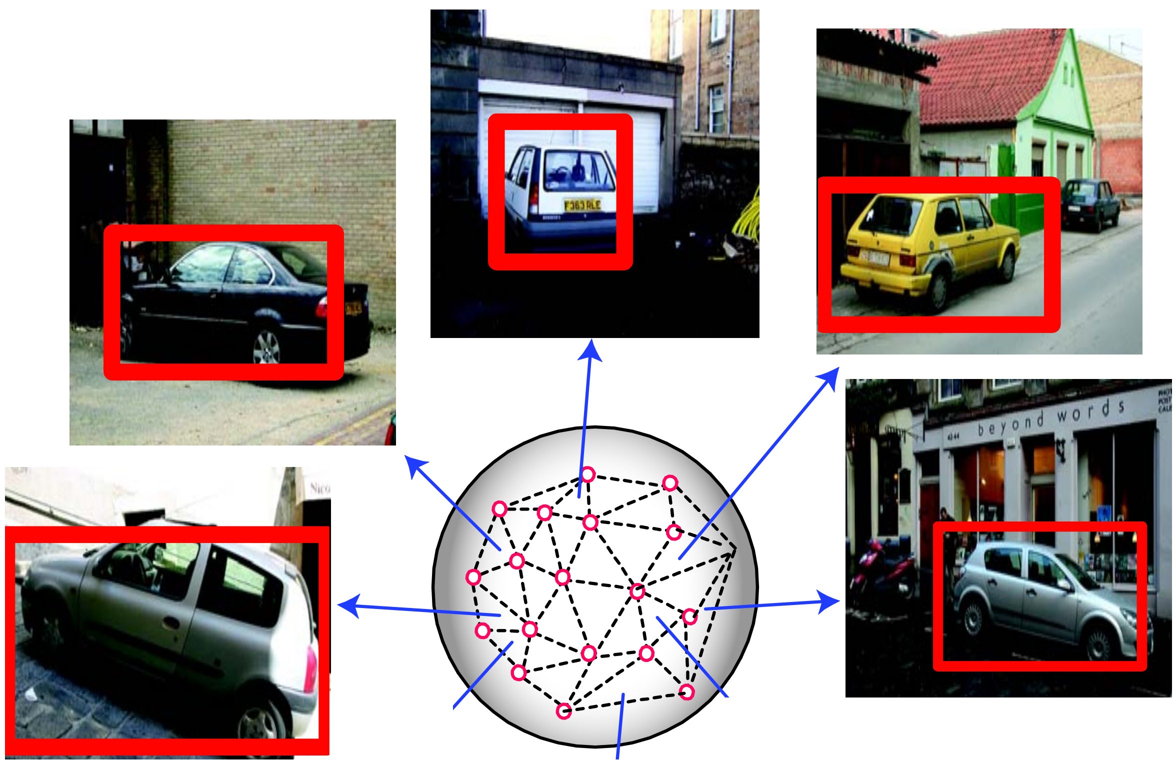
- Canonical parts captures view invariant diagnostic appearance information
- 2d ½ structure linking parts via weak geometry

Drawbacks

- Supervision
  - Part labels required
  - Pose labels required
    - Except [Savarese & Fei-Fei 07, 08] , but… [Arie-Nachimson & Basri ICCV 09]
- No pose estimation
  - Except [Savarese & Fei-Fei 07, 08] [Sun et al 09] [Liebelt 08] [Arie-Nachimson & Basri ICCV 09] [Farhadi ICCV 09]
  - Few poses (at most 8 azimuth, 3 zenith)
  - Still inaccurate
- No or limited ability to synthesize novel views
- Tested on few categories
  - Usually 1-2, but no more than 8 [Savarese & Fei-Fei 07, 08] [Sun et al 09]

Propose a new multi-view model to:

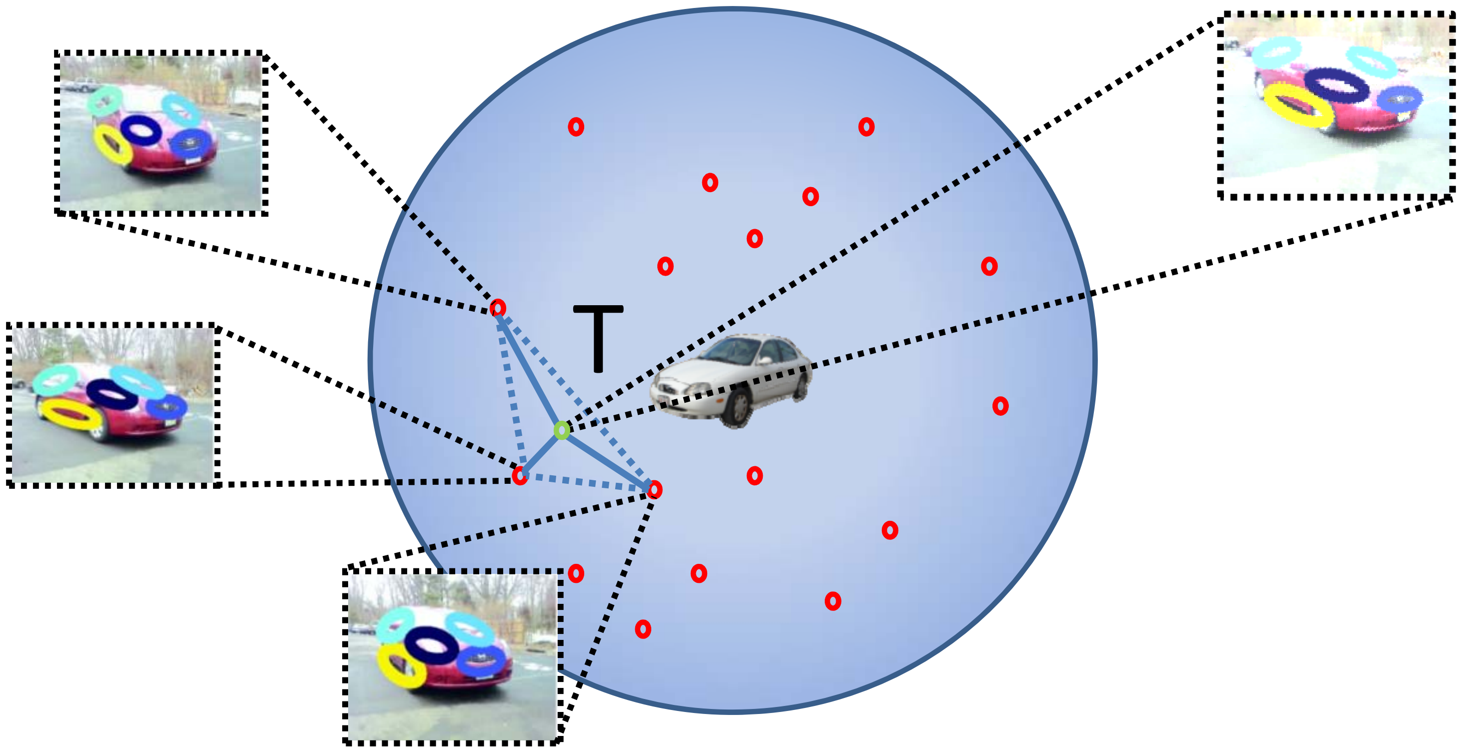
- Detect objects from any viewing angles
- Accurately estimate object pose
- Synthesize object appearance from novel views



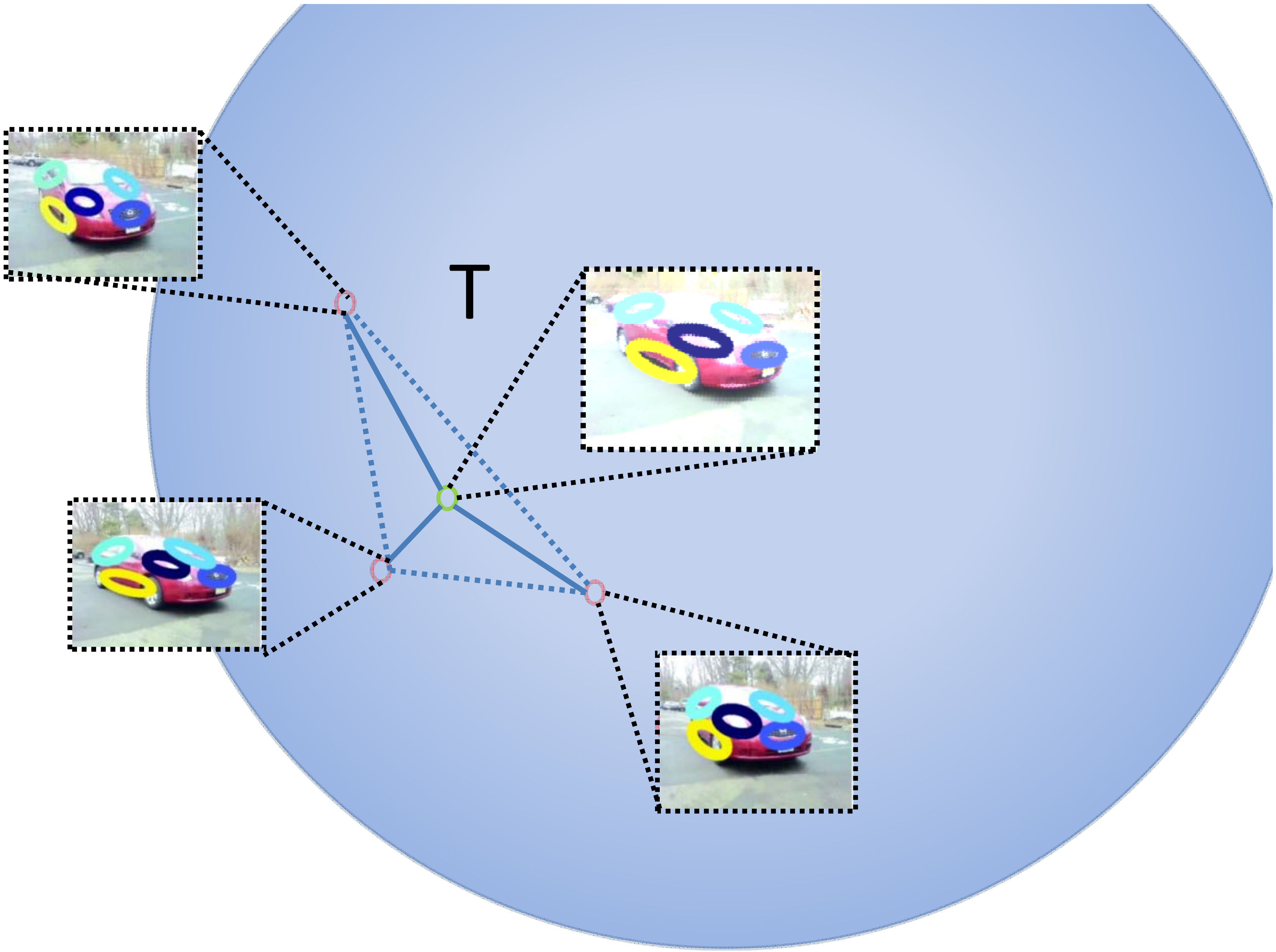
Key contributions

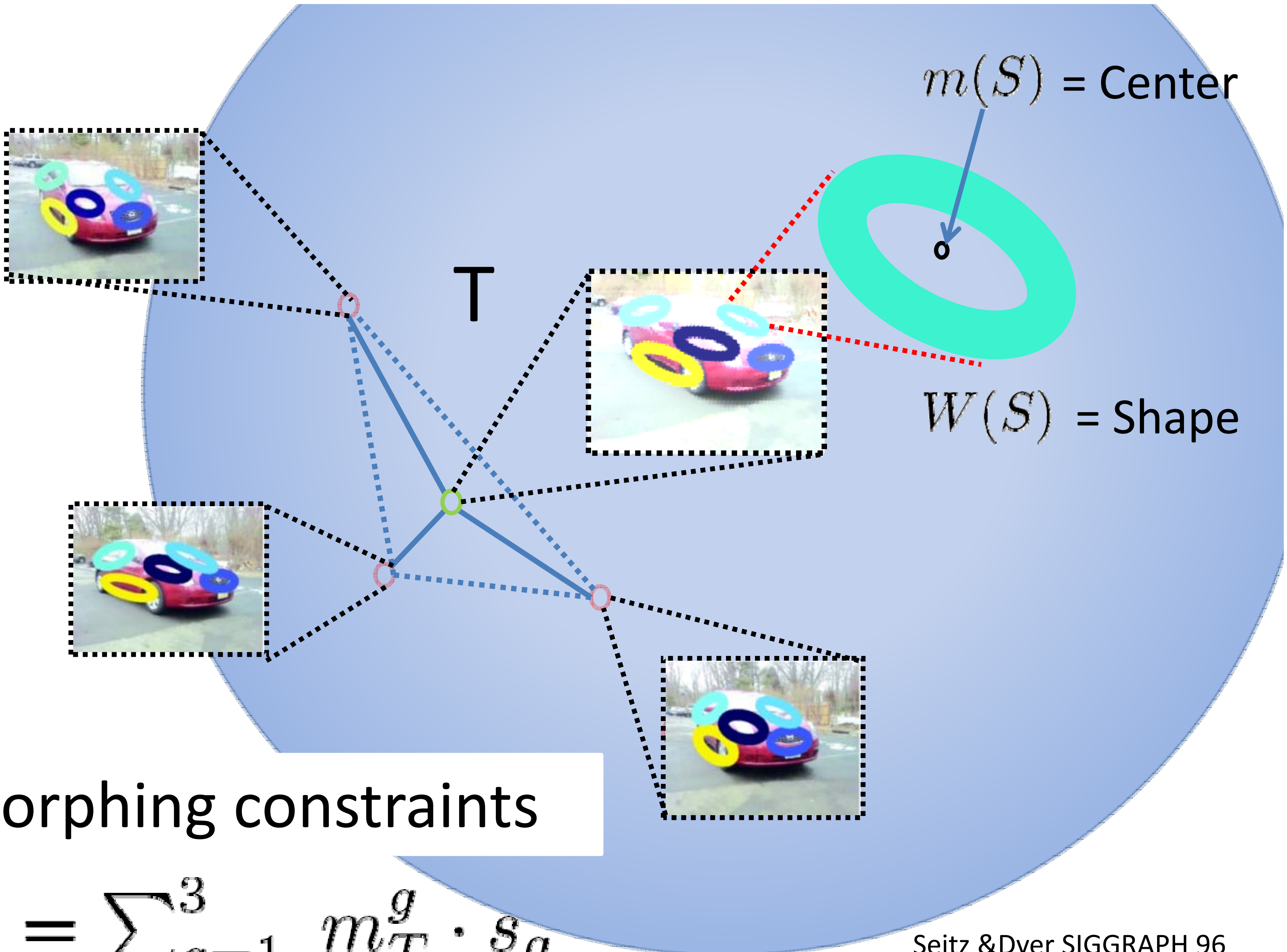
- **Representation**
- **Learning**
    - Dense representation on the viewing sphere:
        - Model object appearance and shape from any position on the viewing sphere
        - Enable view synthesis from novel view points
    - Multi-view generative part-based model [Sun et al. CVPR 09]
        - Object is represented by collections of parts
        - Parts are linked across views
        - Parts and relationships are probabilistic
    - Semi-supervised learning
        - No part or pose labels are required
    - Incremental:
        - Training images can be provided sequentially

Dense representation on view-sphere



- Triangle T
- Morphing parameter S


T



View morphing constraints

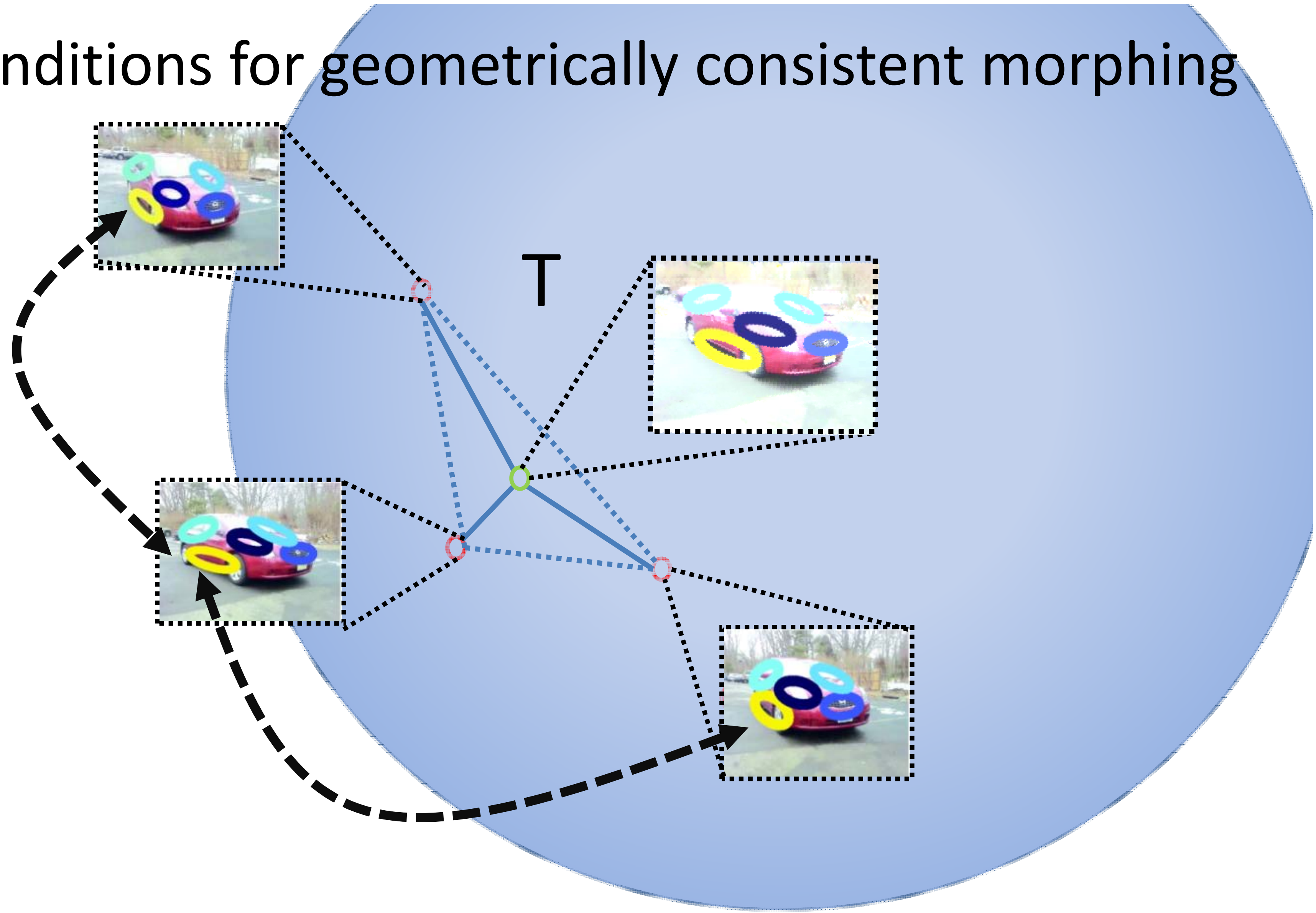
$$m(S) = \sum_{g=1}^{3} m_T^g \cdot s_g$$

$$W(S) = \sum_{g=1}^{3} W_T^g \cdot s_g$$

Seitz &Dyer SIGGRAPH 96
Xiao & Shah CVIU '04

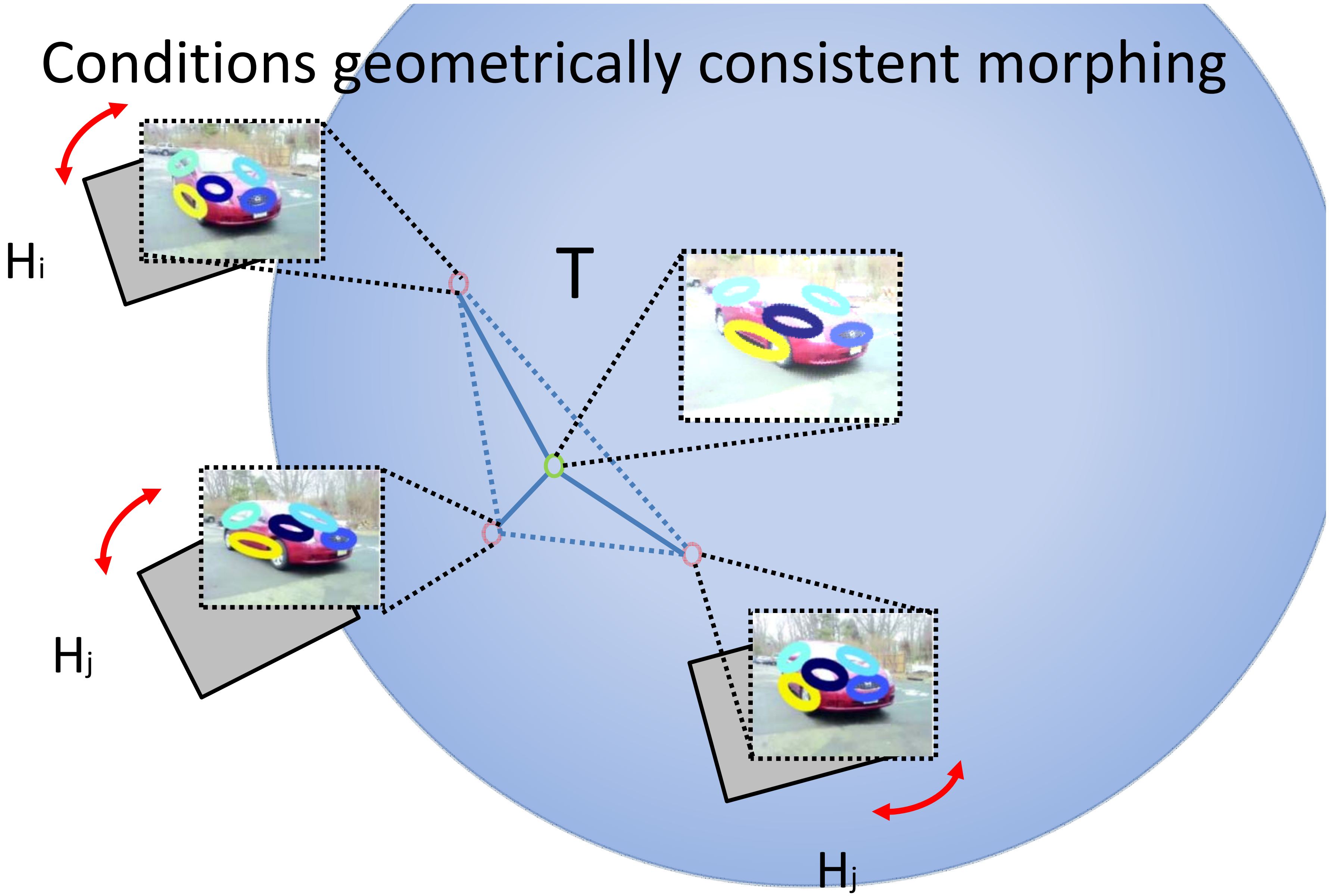
For first time used for modeling object categories!

Conditions for geometrically consistent morphing



1. Correct correspondence between parts must be established

Conditions geometrically consistent morphing

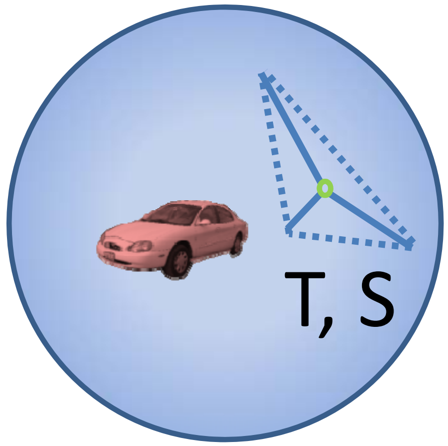


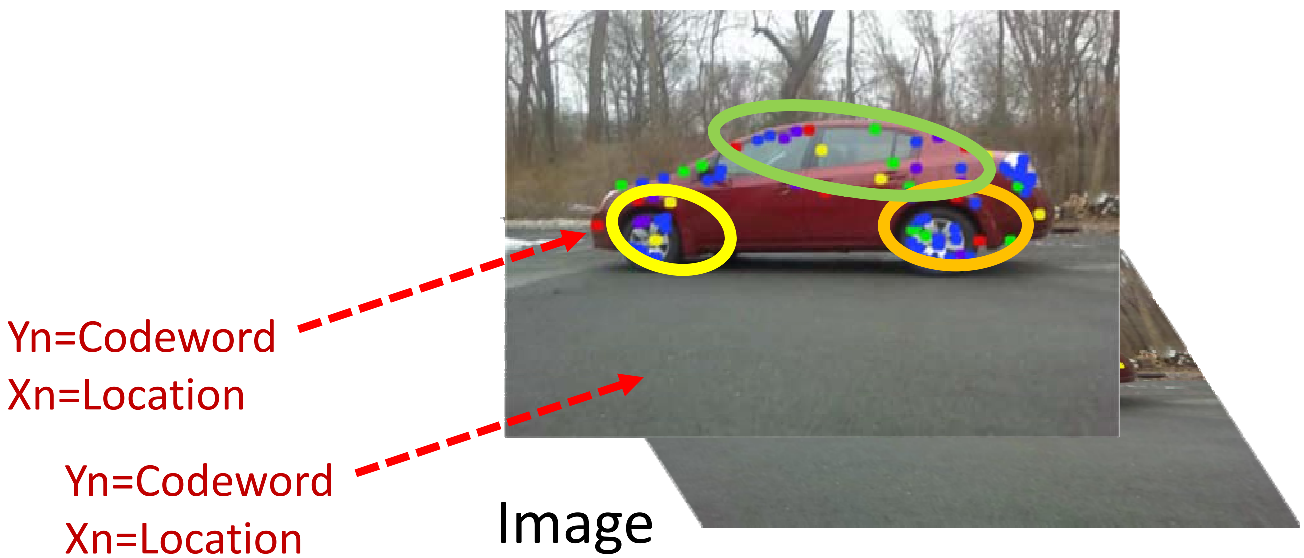
2. Key views are rectified
by a pre-warping transformations H

Key contributions

- **Representation:**
  - Dense representation on the viewing sphere:
    - Model object appearance and shape from any position on the viewing sphere
    - Enable view synthesis
  - Multi-view generative part-based model [Sun et al cvpr 09]
    - Object is represented by collections of parts
    - Parts are linked across views
    - Parts and relationships are probabilistic
- **Learning:**
  - Semi-supervised learning
    - no part or pose labels are required
  - Incremental:
    - Training images can be provided sequentially

Multi-view generative part-based model





Multi-view generative part-based model

α = Part Prop. Prior

$\pi \sim Dir(\alpha)$

$R \sim Mult(\pi)$

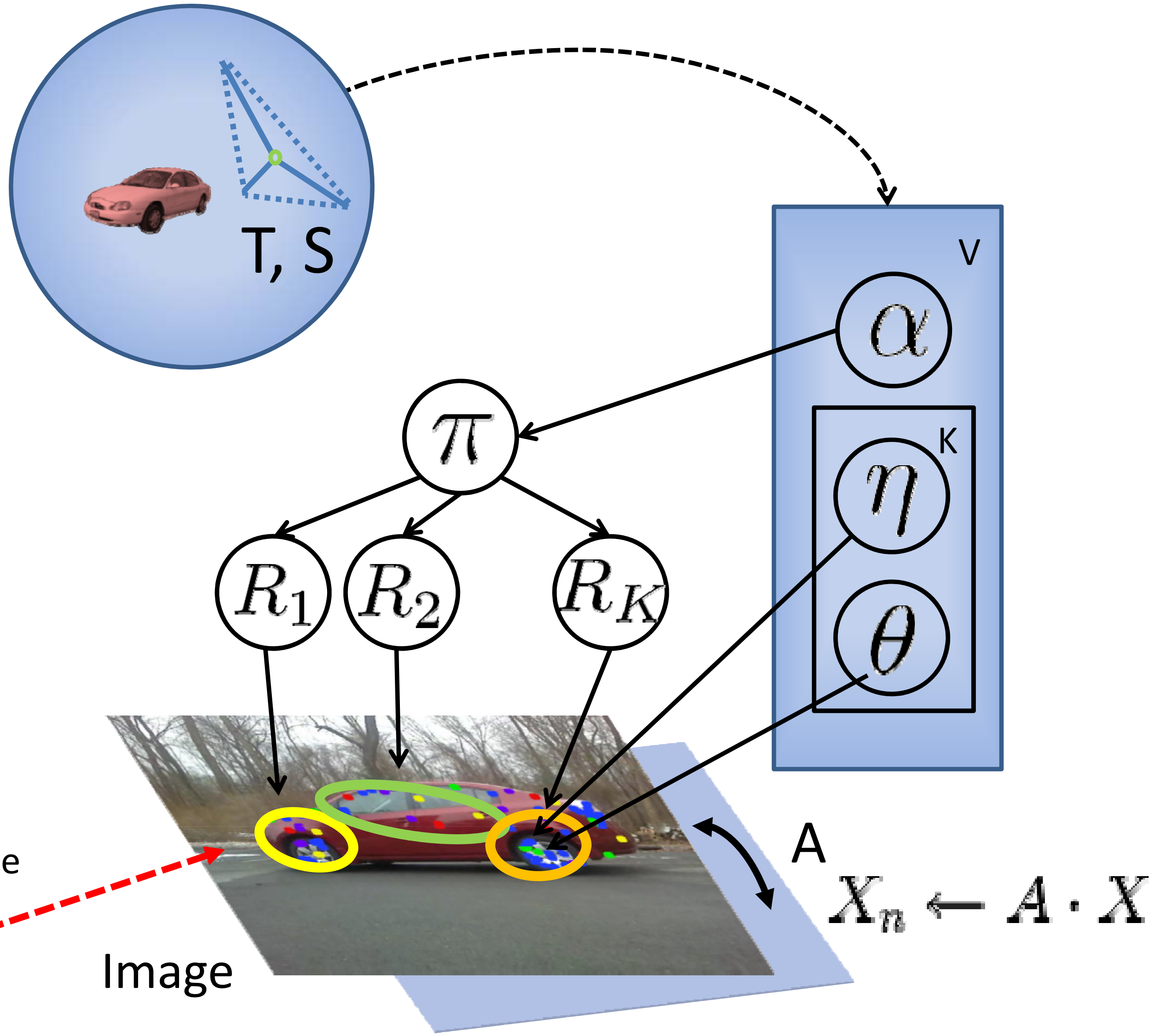
$Y_n \sim Mult(\eta)$

$X_n \sim N(theta)$

η = Part Appearance

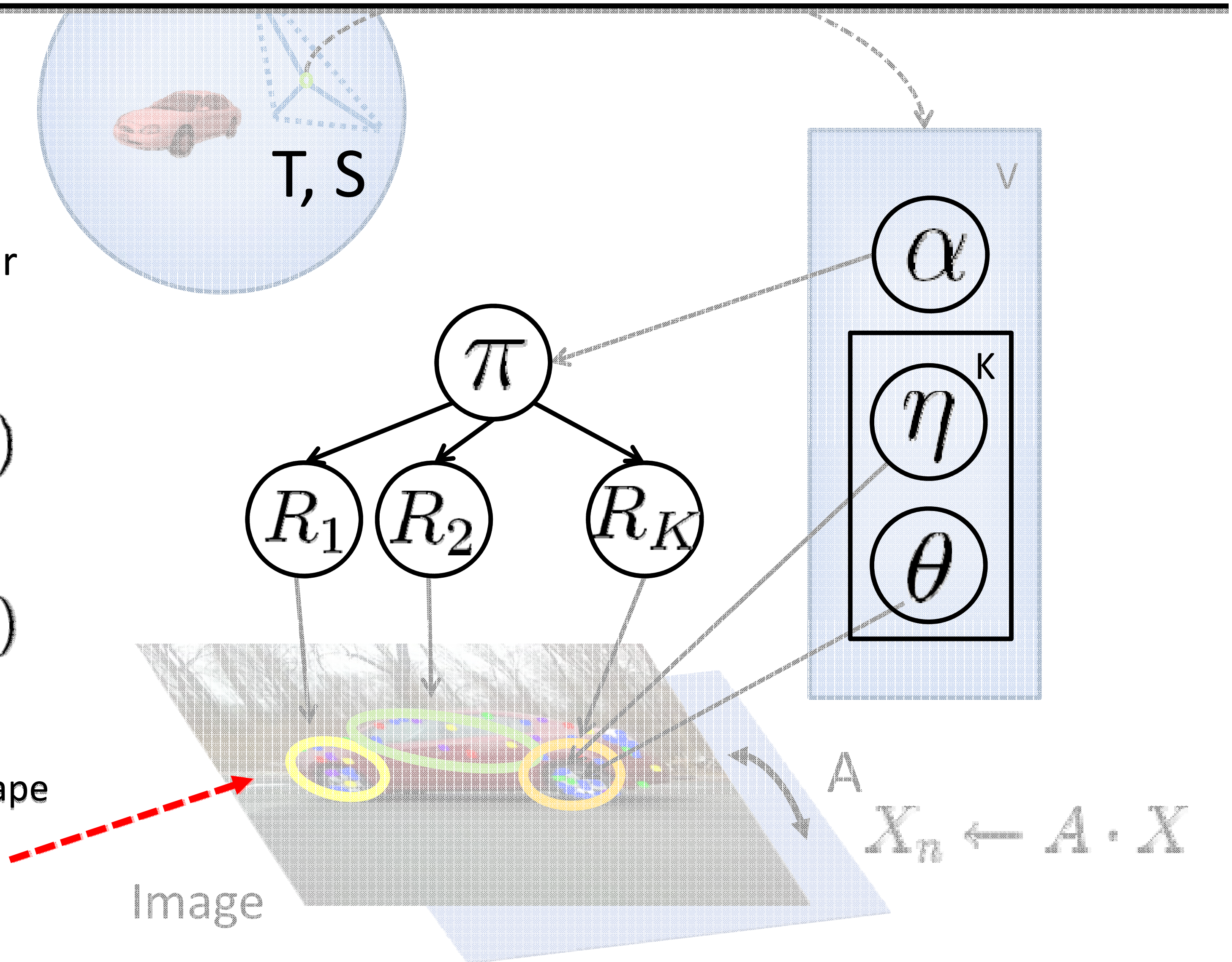
θ = Part Location/shape

Yn=Codeword
Xn=Location



$$P(X, Y, T, S, R, \pi) \propto P(\pi | \alpha_T)$$

$$\prod_n \{P(X_n | \theta_{TR_n}(S), A) P(Y_n | \eta_{TR_n}(S)) P(R_n | \pi)\}$$



α = Part Prop. Prior

$\pi \sim Dir(\alpha)$

$R \sim Mult(\pi)$

$Y_n \sim Mult(\eta)$

$X_n \sim N(theta)$

η = Part Appearance

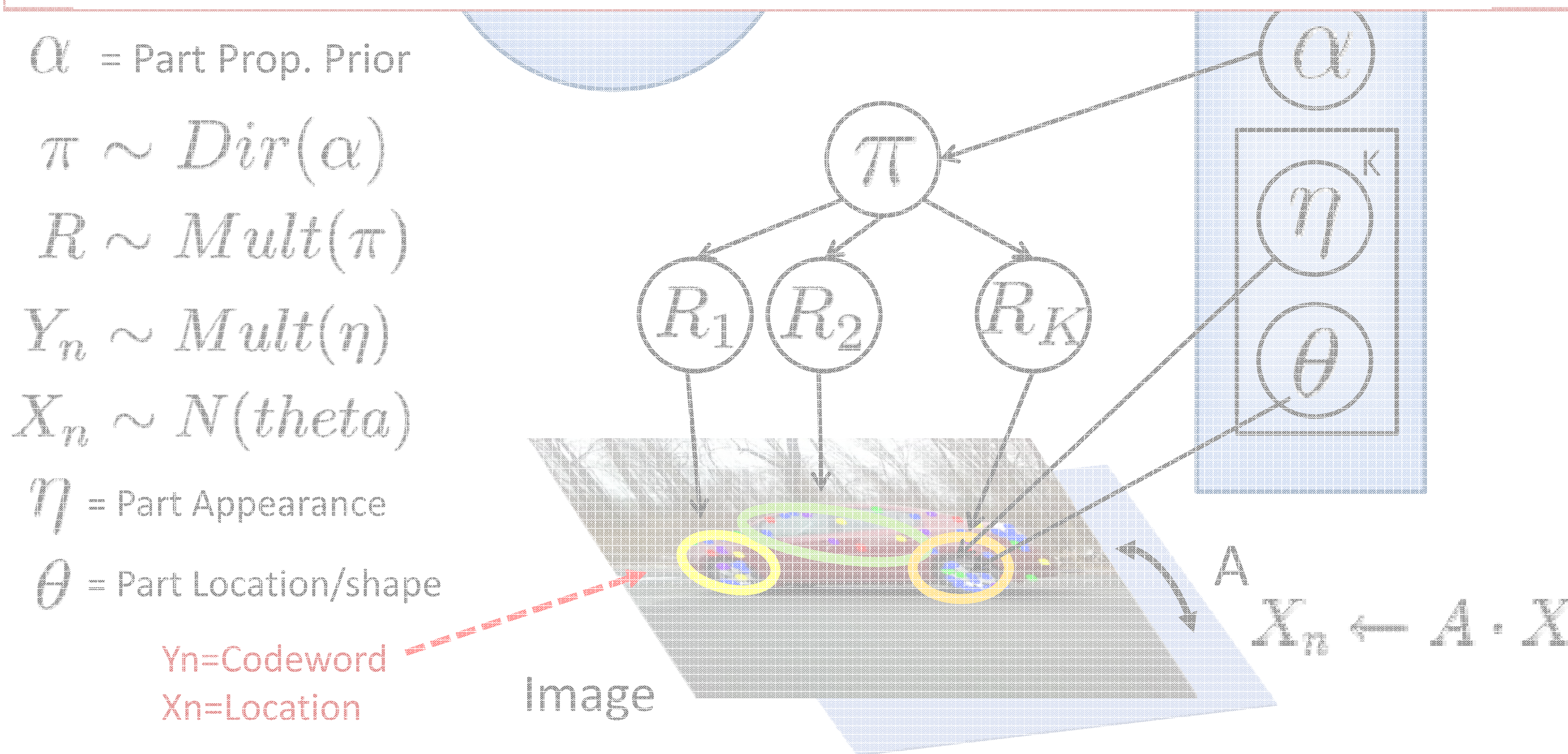
θ = Part Location/shape

Yn=Codeword
Xn=Location

$$P(X,Y,T,S,R,\pi) \propto P(\pi|\alpha_T)$$
$$\prod_n \{P(X_n|\theta_{TR_n}(S),A)P(Y_n|\eta_{TR_n}(S))P(R_n|\pi)\}$$

Exact Inference is intractable!

We use Variational EM

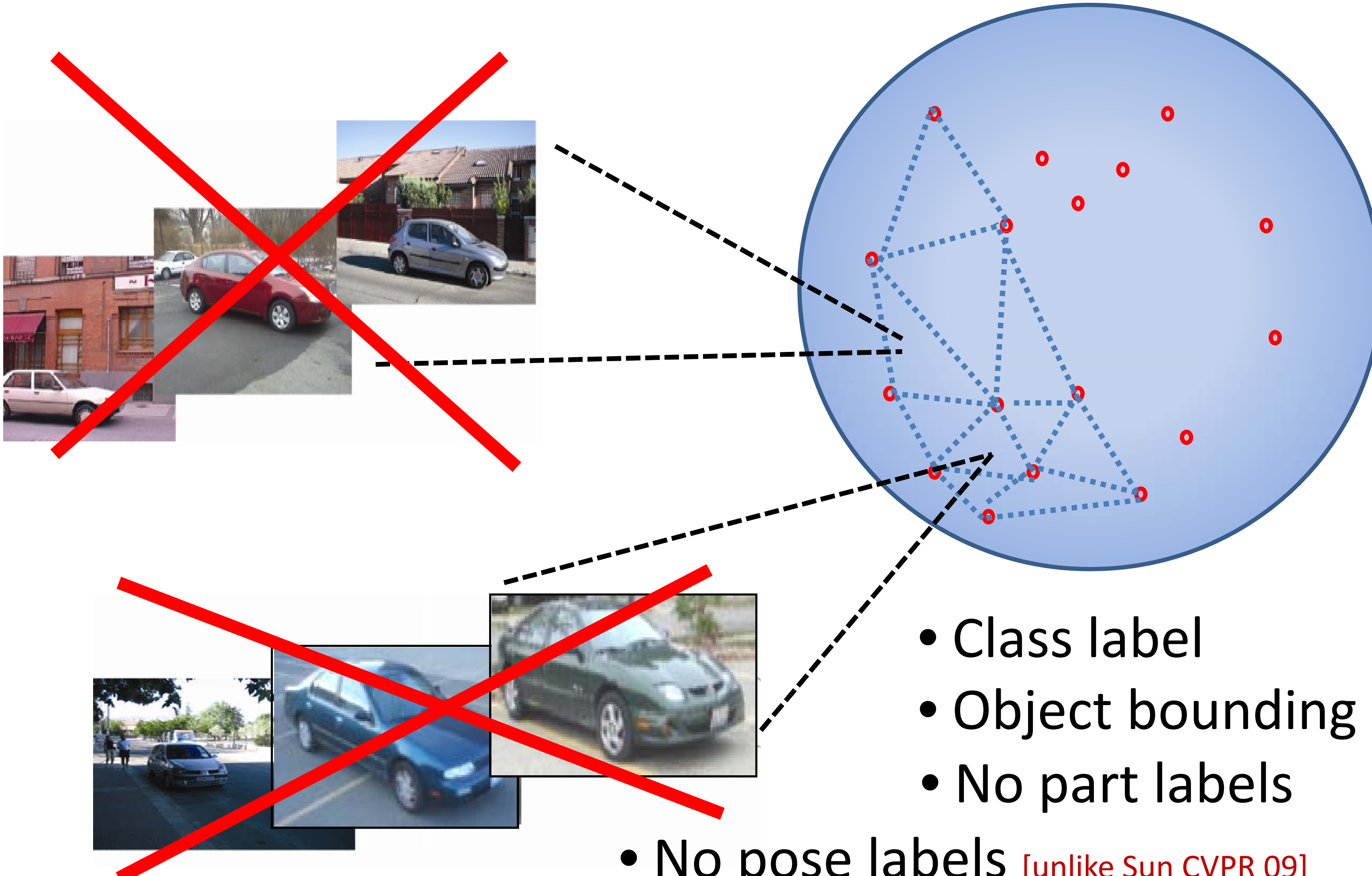


Key contributions

- **Representation:**
  - Dense representation on the viewing sphere:
    - Model object appearance and shape from any position on the viewing sphere
    - Enable view synthesis
  - Multi-view generative part-based model [Sun et al cvpr 09]
    - Object is represented by collections of parts
    - Parts are linked across views
    - Parts and relationships are probabilistic

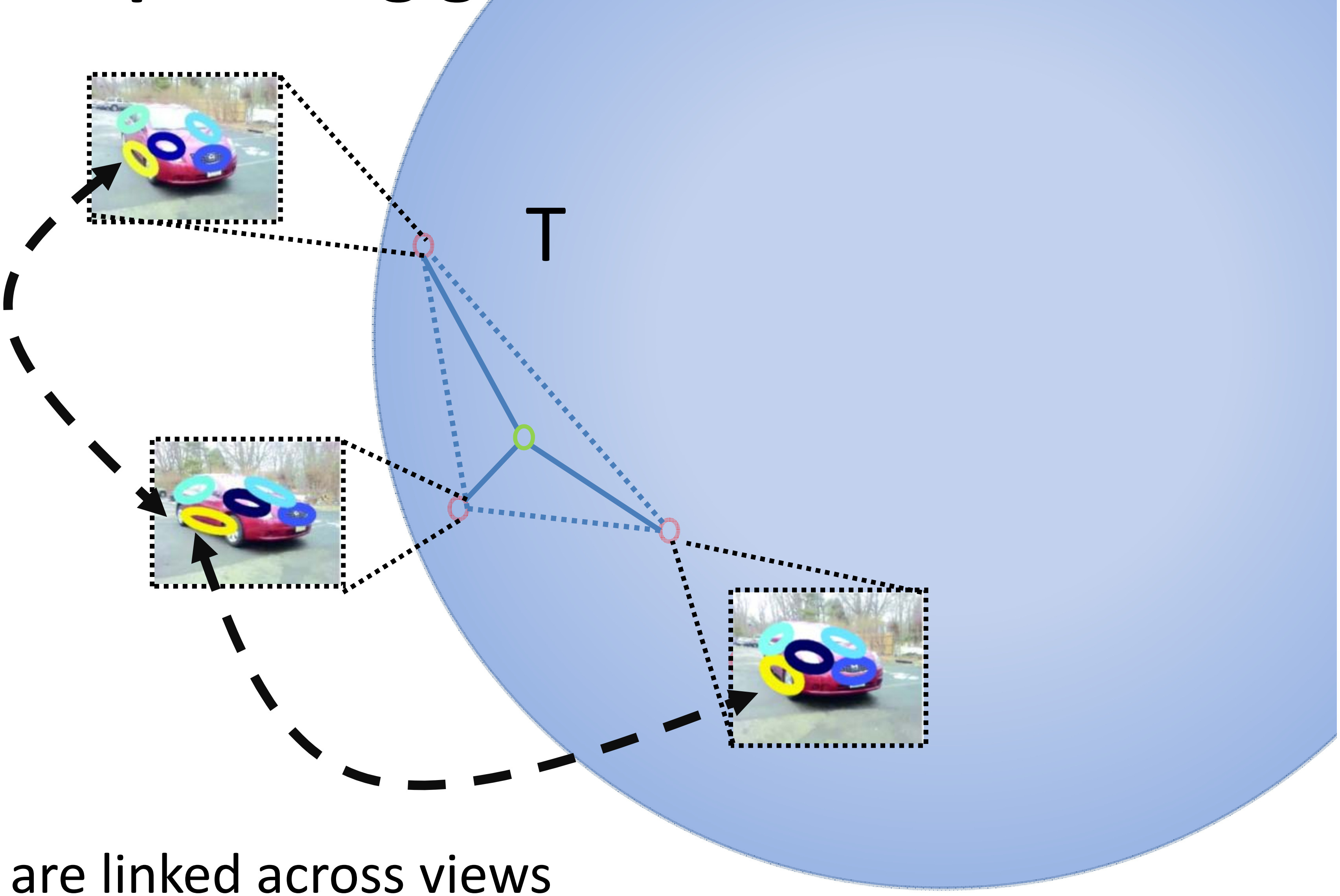
- **Learning:**
  - Semi-supervised learning
    - no part or pose labels are required
  - Incremental:
    - Training images can be provided sequentially

Semi-supervised



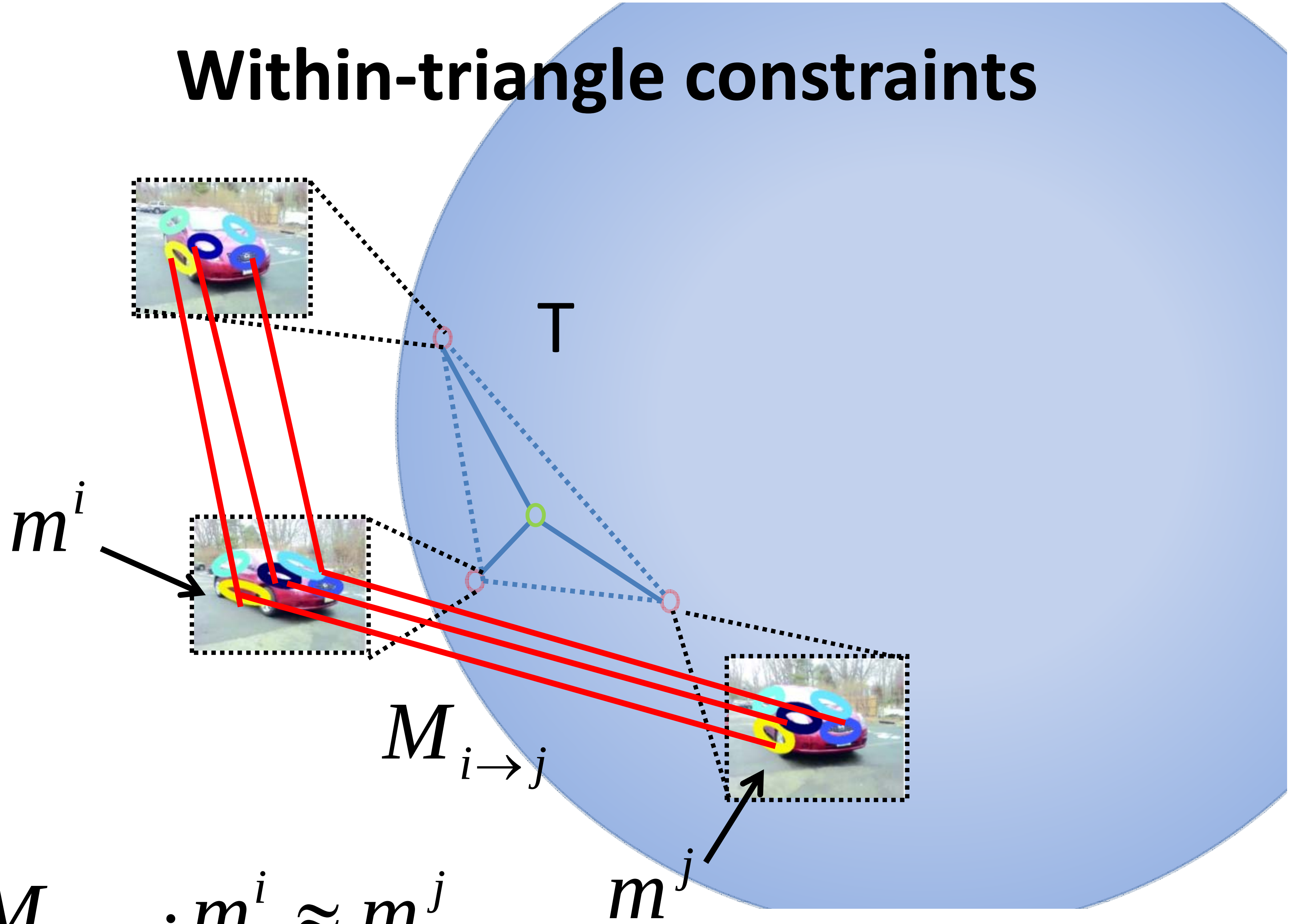
- Class label
- Object bounding box
- No part labels
- No pose labels [unlike Sun CVPR 09]
- No need to observe same object instance from multiple views [unlike Savarese & Fei-Fei, 07, 08]

Incorporating geometrical constraints



- Parts are linked across views
- Part topology is preserved under morphing transformation

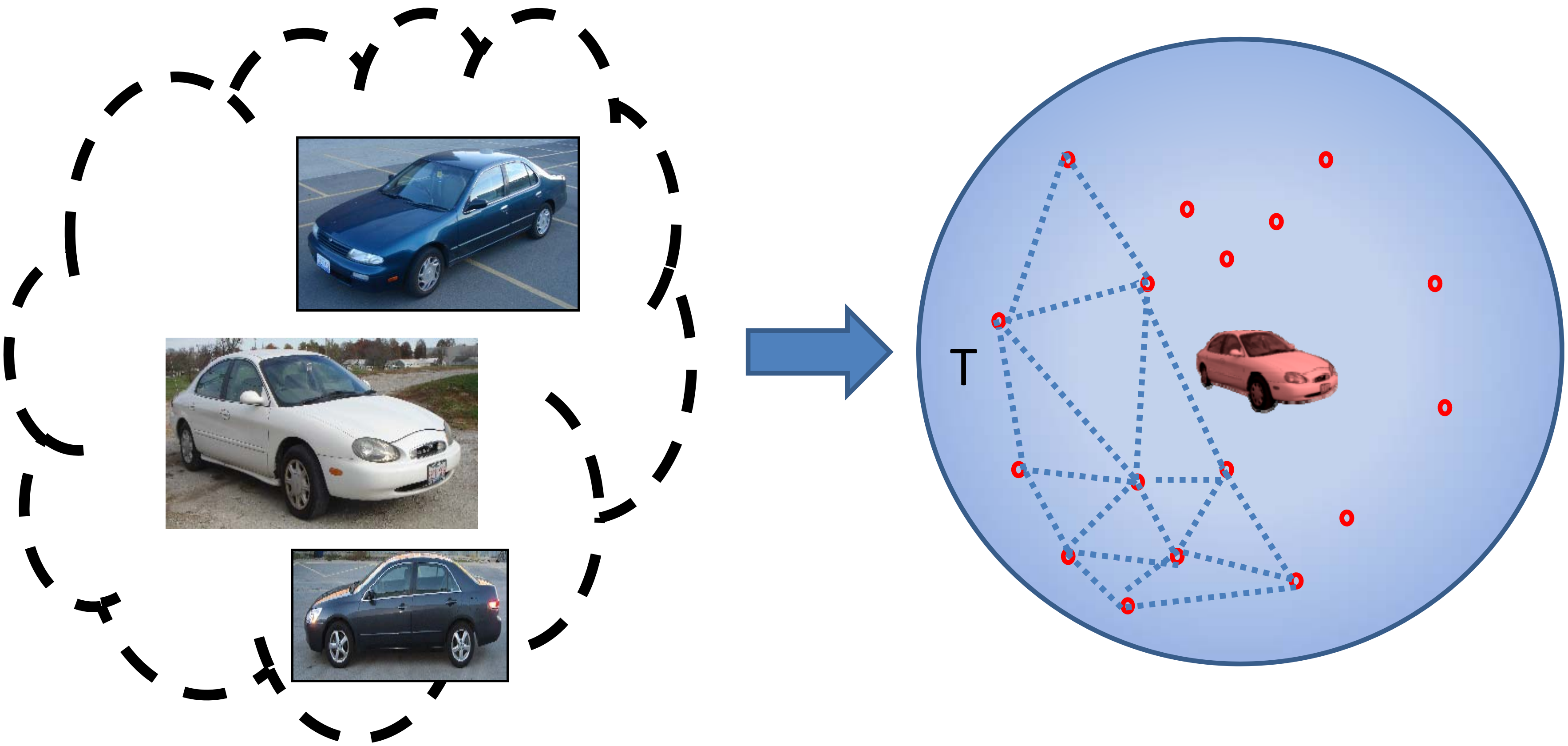
Within-triangle constraints



$$M_{i \to j} \cdot m^i \approx m^j$$

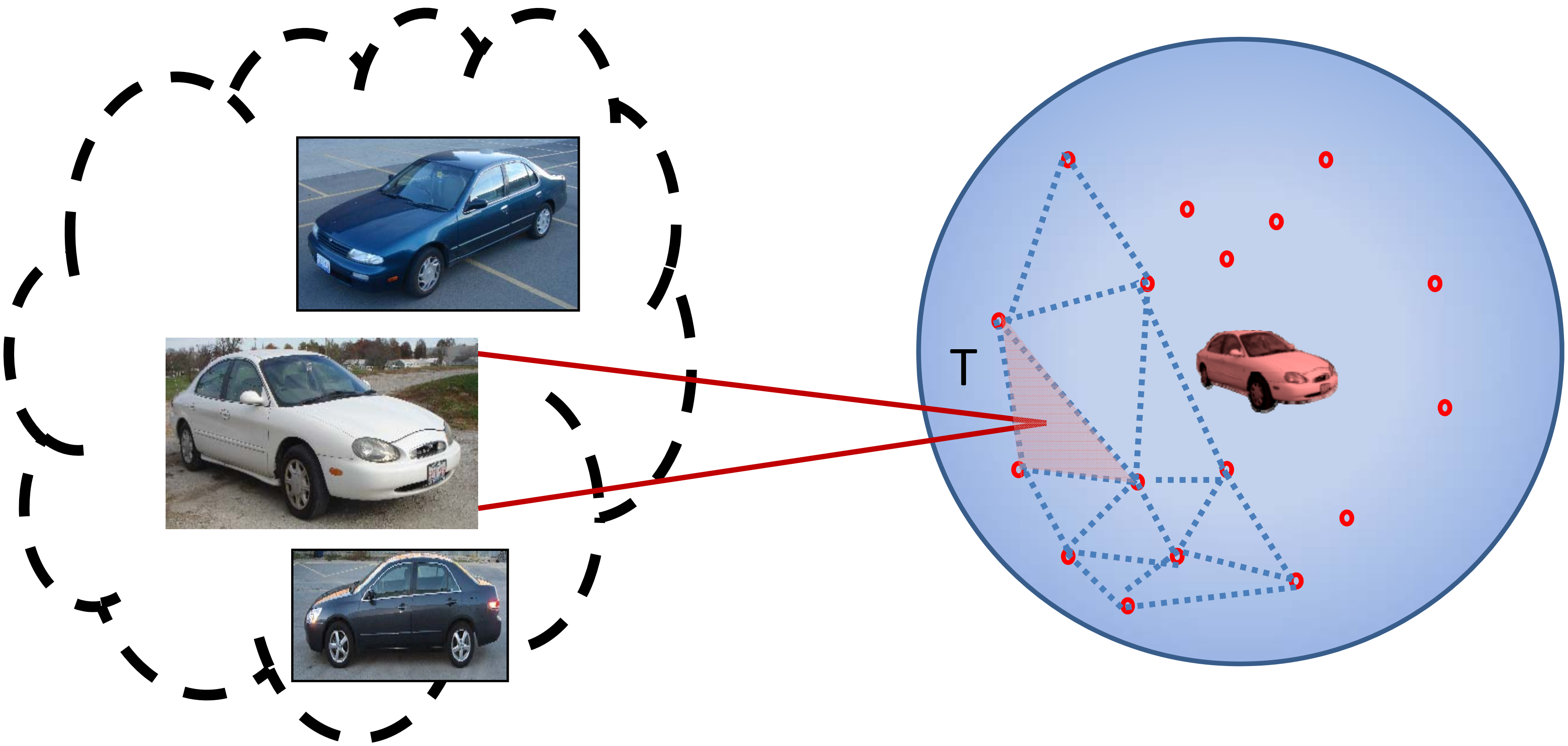
Encoded as a penalty term in variational *EM*

Incremental learning



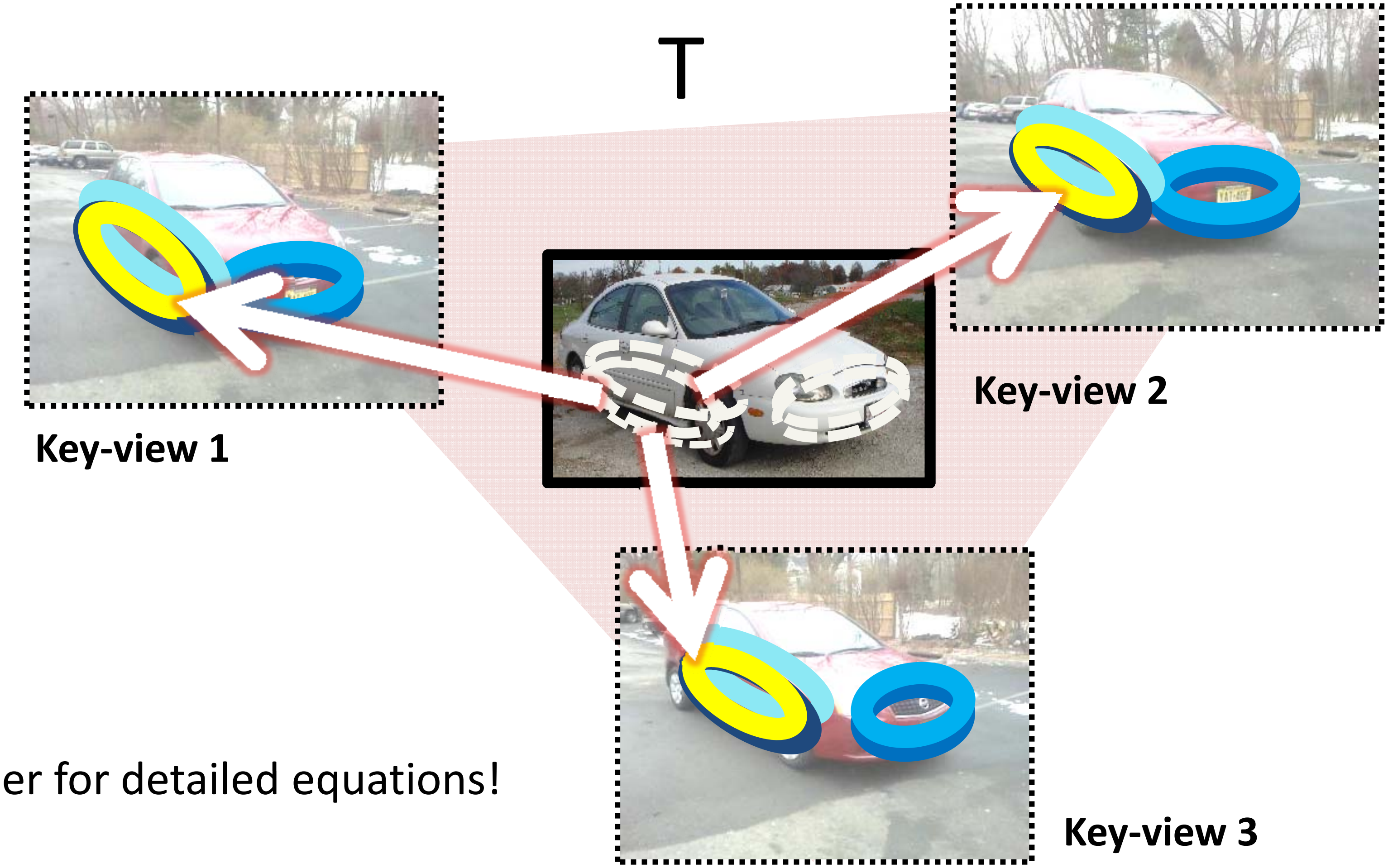
- Enable unorganized and on-line collection training images
- Increase efficiency in learning (no need large storage space)

Incremental learning



- Sequentially assign new training images to triangles on view sphere

Incremental learning



See paper for detailed equations!

- Sequentially assign new training images to triangles on view sphere
- Evidence of training image used to update model parameters

Initializing the model

- Estimating key views and triangles
- Defining initial parts

$$\pi : \boldsymbol{I}^h \rightarrow \left\{ \boldsymbol{P}_1^h,\ \boldsymbol{P}_2^h,\ \boldsymbol{P}_3^h, \boldsymbol{O}^h \right\}$$

$$\tau : \boldsymbol{I}^k \rightarrow \left\{ \boldsymbol{P}_1^k,\ \boldsymbol{P}_2^k,\ \boldsymbol{P}_3^k,\ \boldsymbol{O}^k \right\}$$

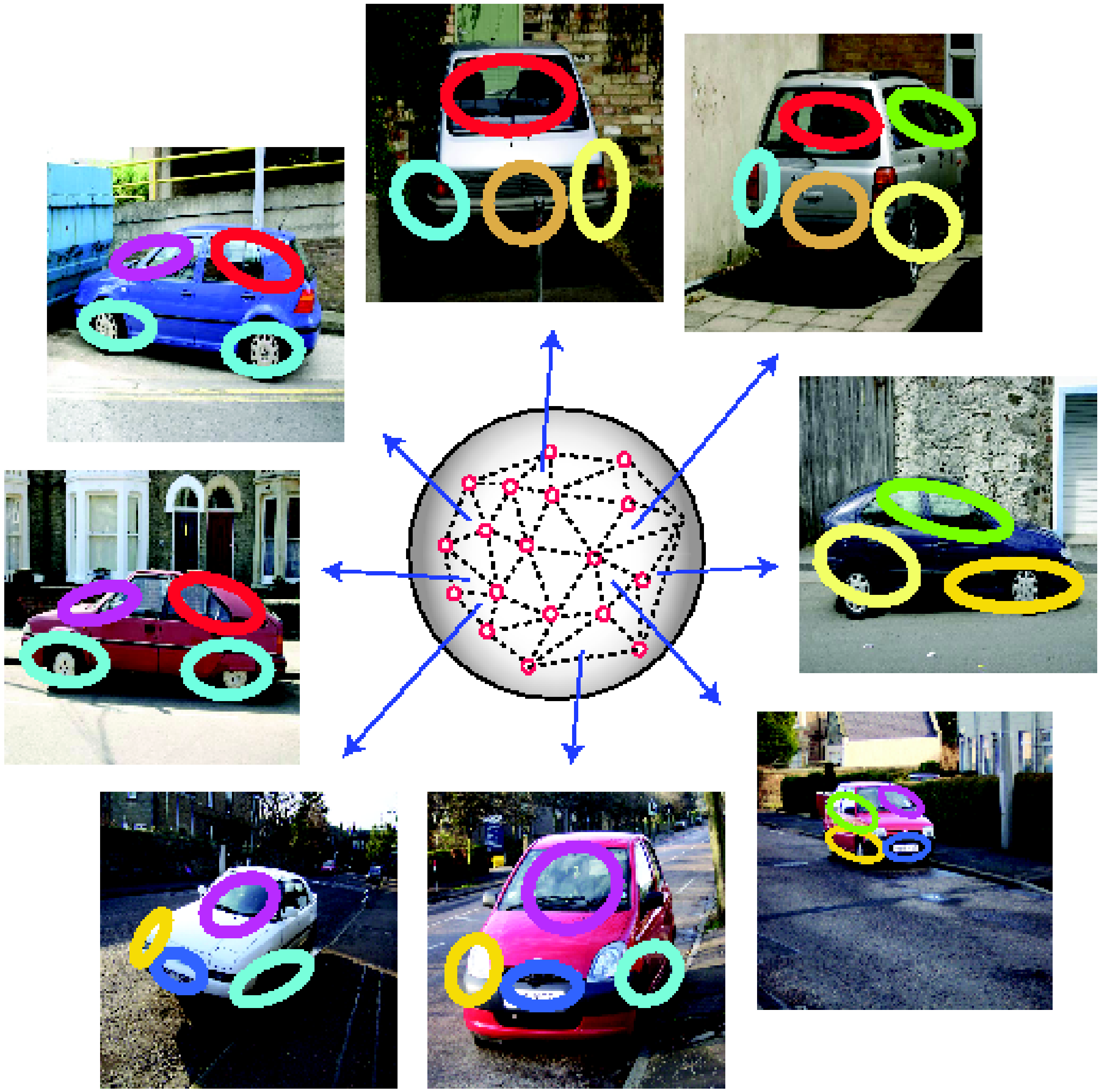
Sequential ransac

J-linkage

Example of part learning

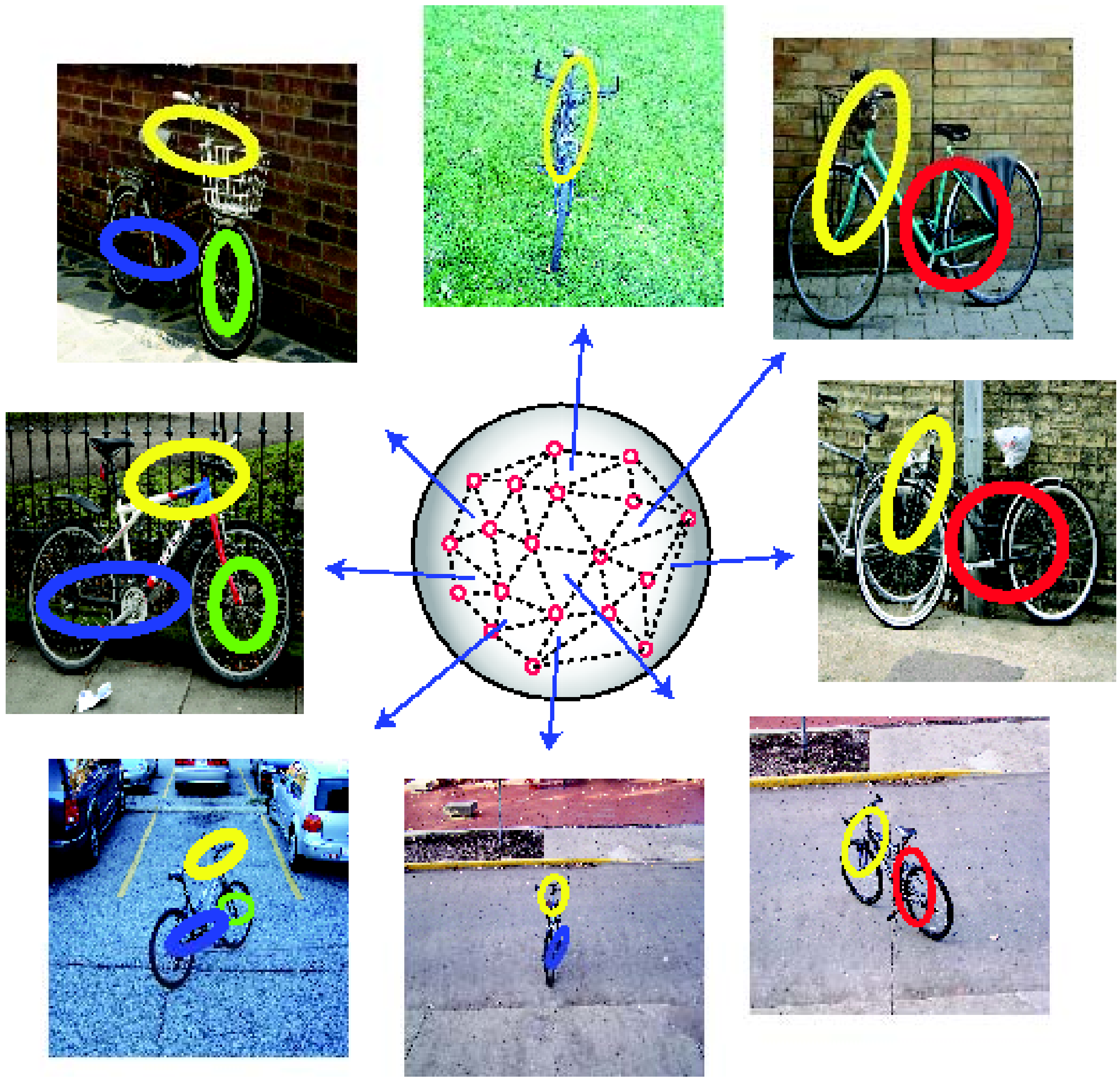
Examples of learnt part-based models

Car



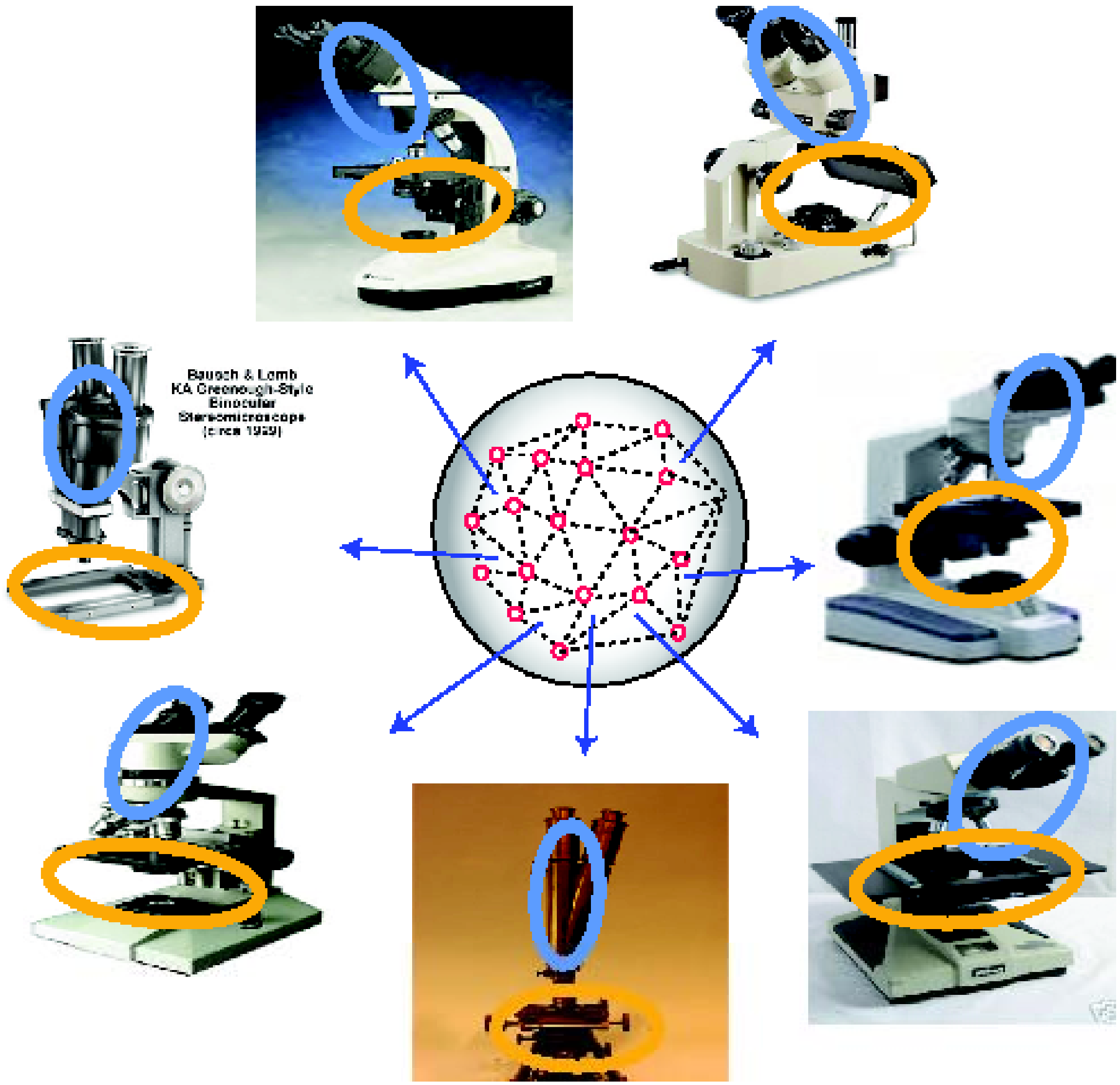
Examples of learnt part-based models

Bicycle



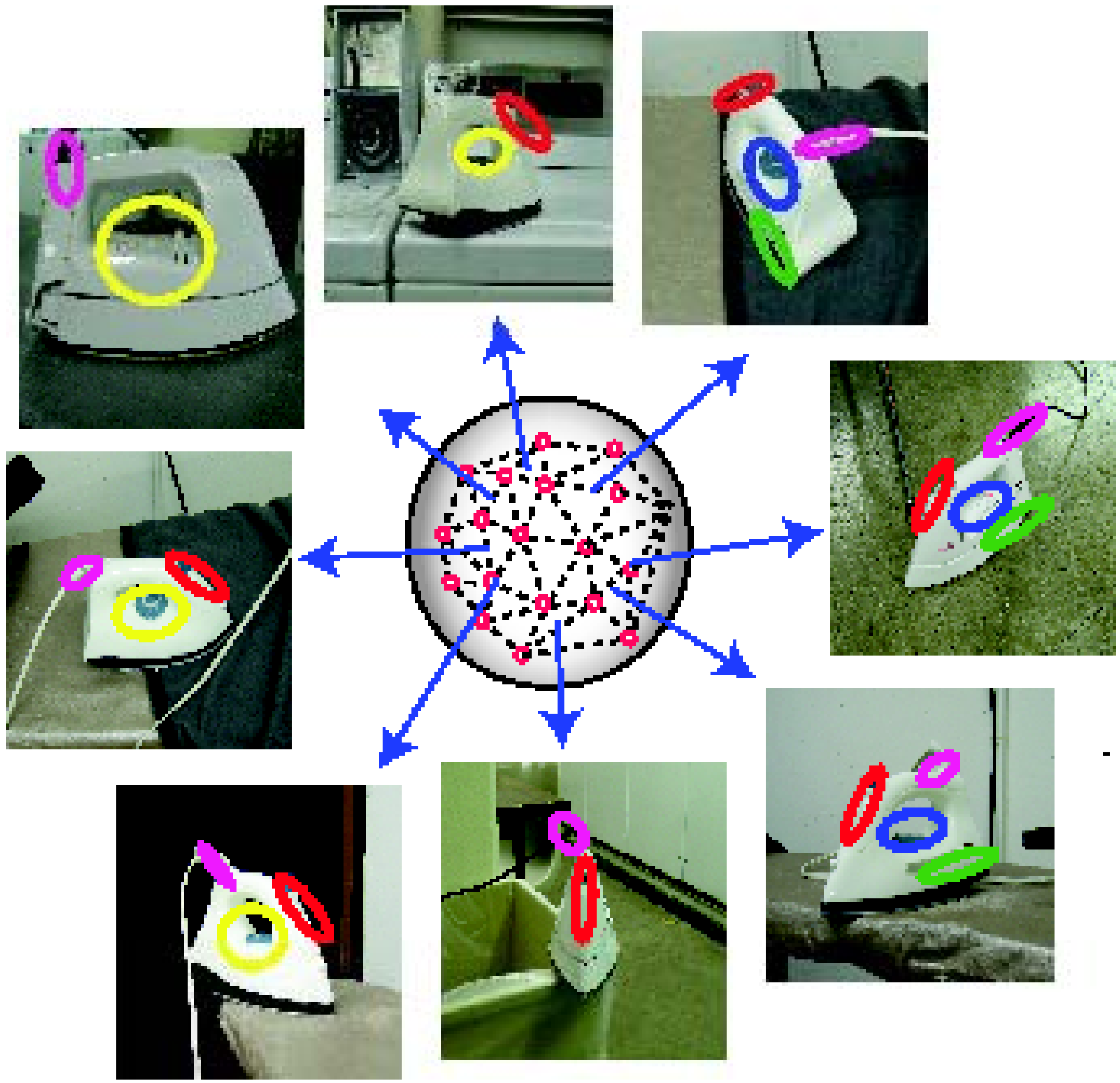
Examples of learnt part-based models

Binocular micro-scope



Examples of learnt part-based models

Travel iron

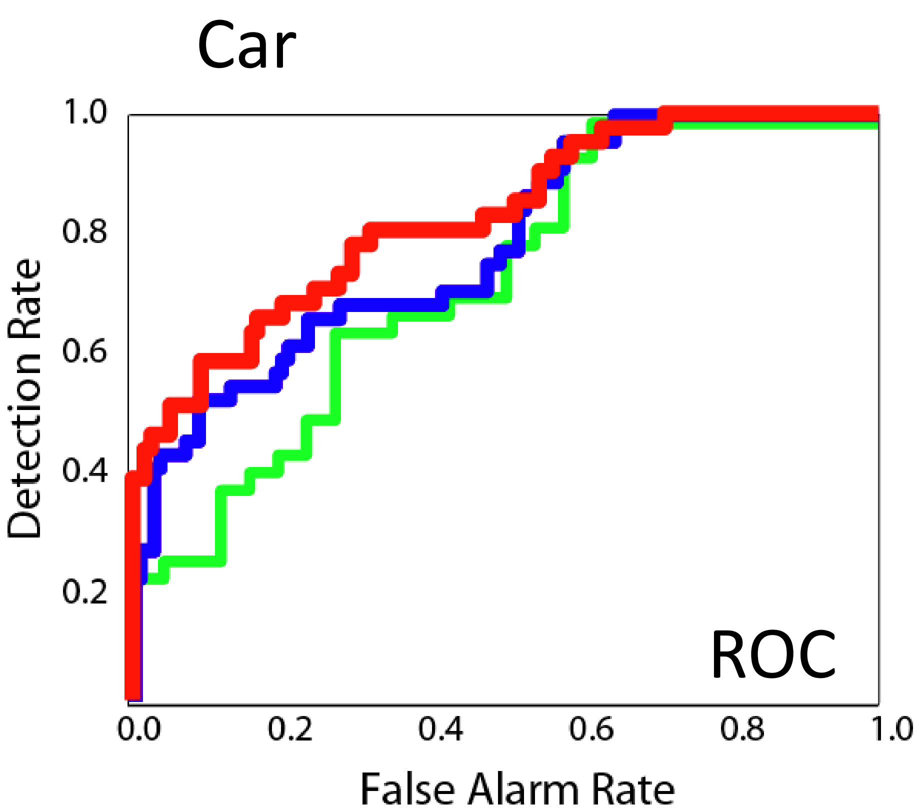


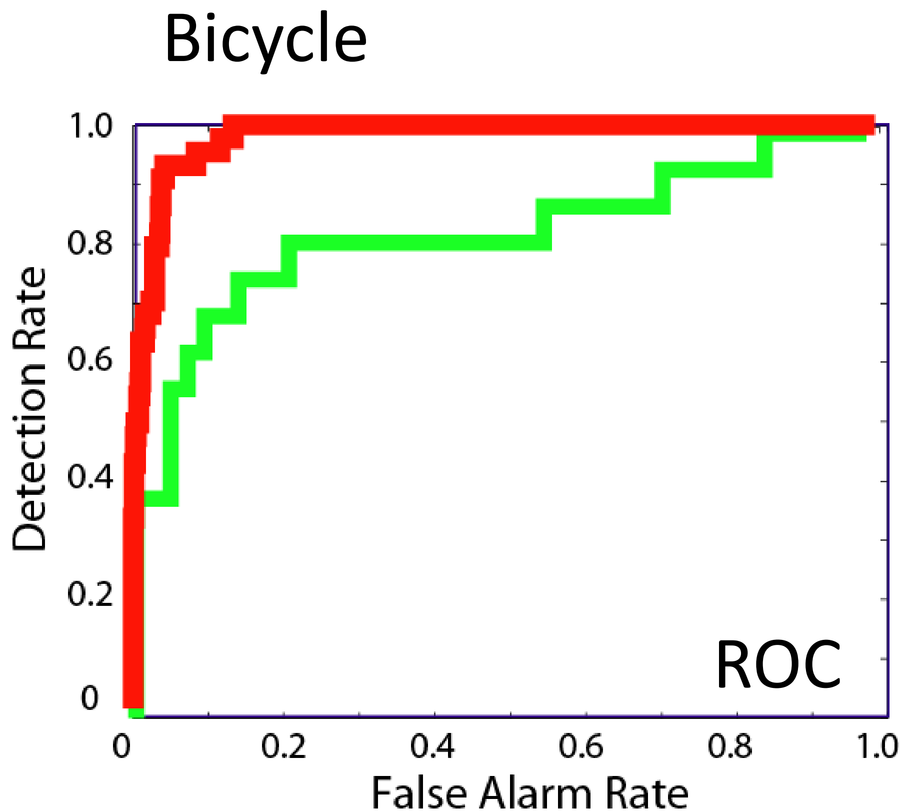
Let’s use our model!

- **Detect** objects from any viewing angles
- Accurate **pose estimation**
- **Synthesize** object shape and appearance from novel views

Detection – UIUC 3D dataset [Savarese & Fei-Fei 07]

Detection - UIUC 3D dataset





- Our model
- Min et al, CVPR 09
- Savarese & Fei-Fei ICCV '07

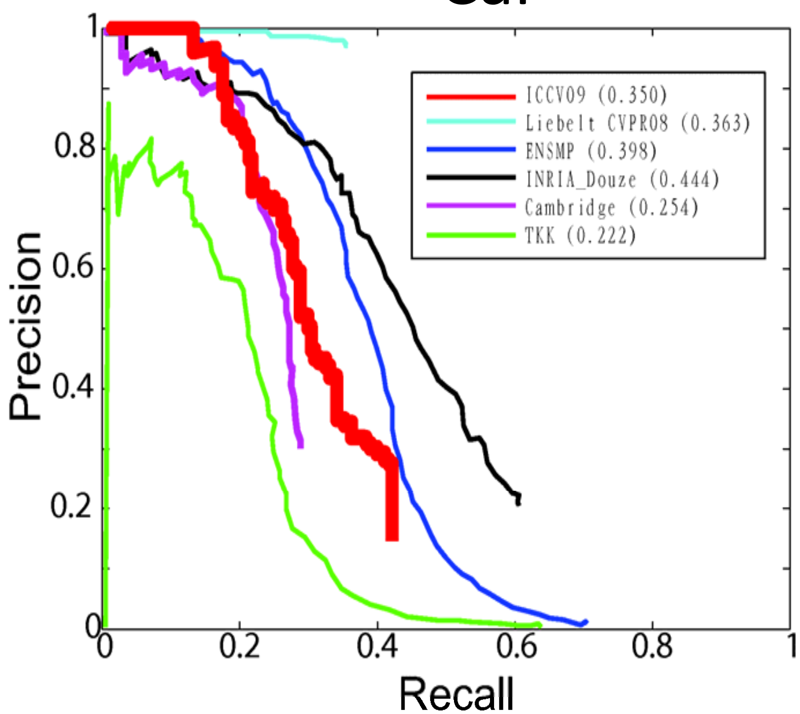
Detection - Pascal 2006 dataset

Bicycle

Car

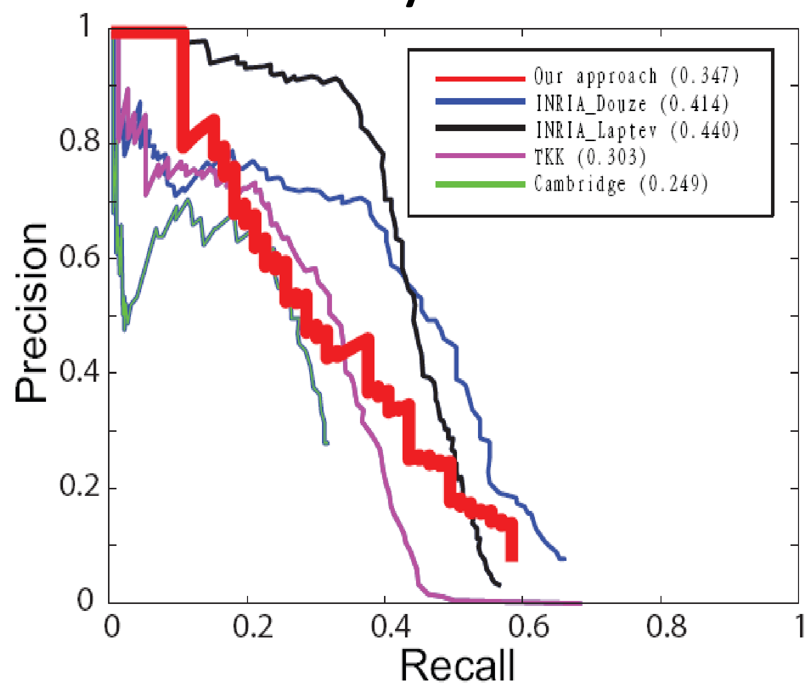
Detection - Pascal 2006 dataset

Car



0.35(average p)

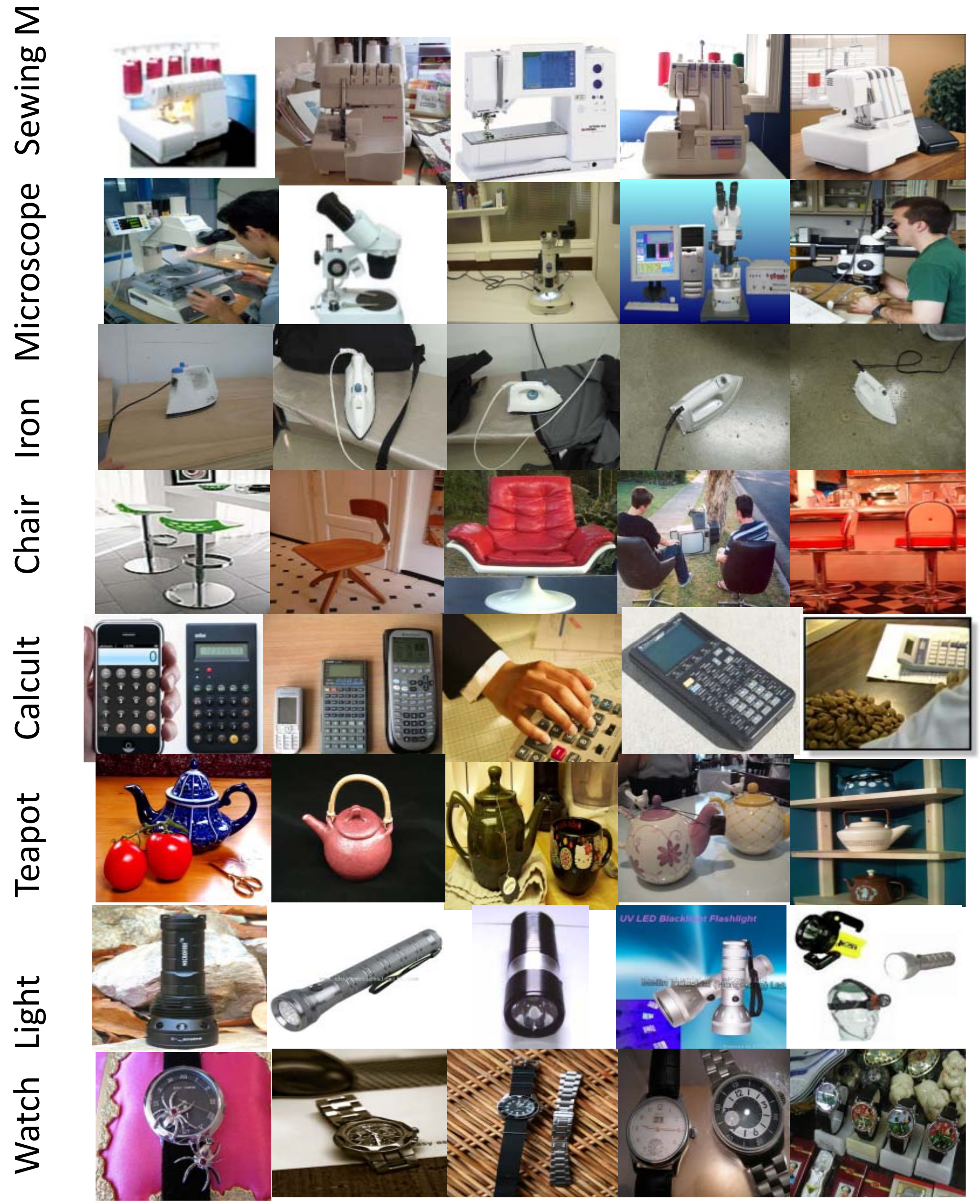
Bicycle



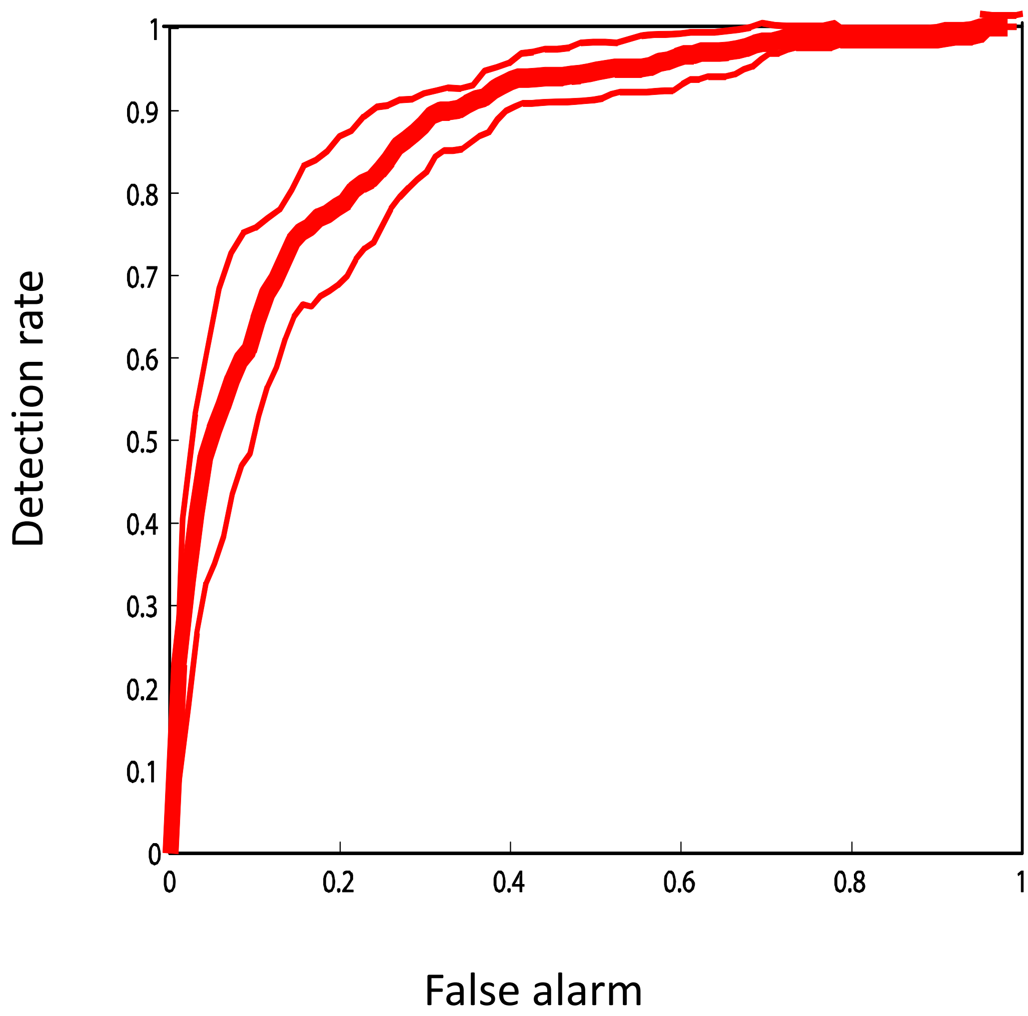
0.347(average p)

— Our model

Detection - Household Item Dataset

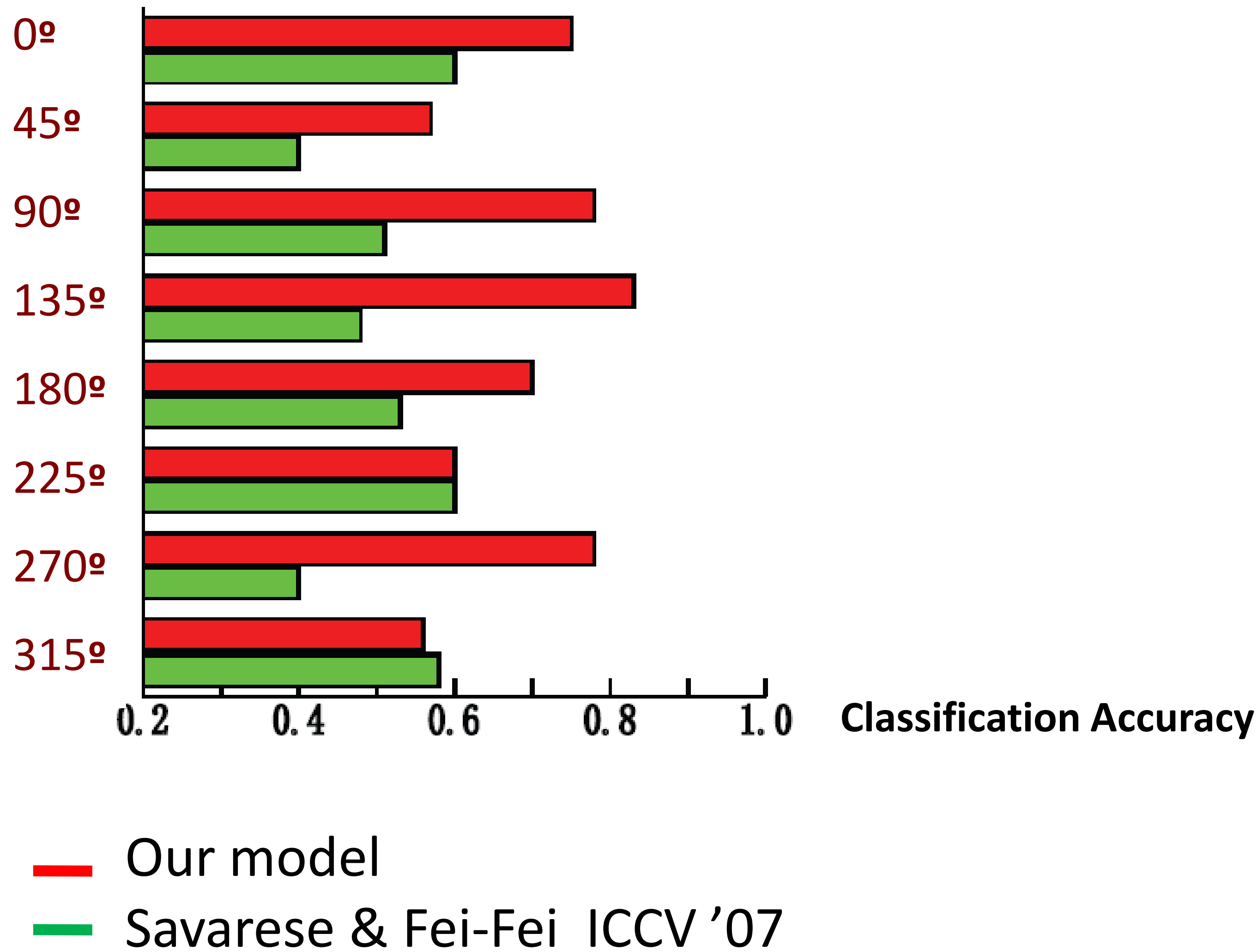


Detection - Household Item Dataset



Viewpoint Classification

Car from UIUC 3D Dataset

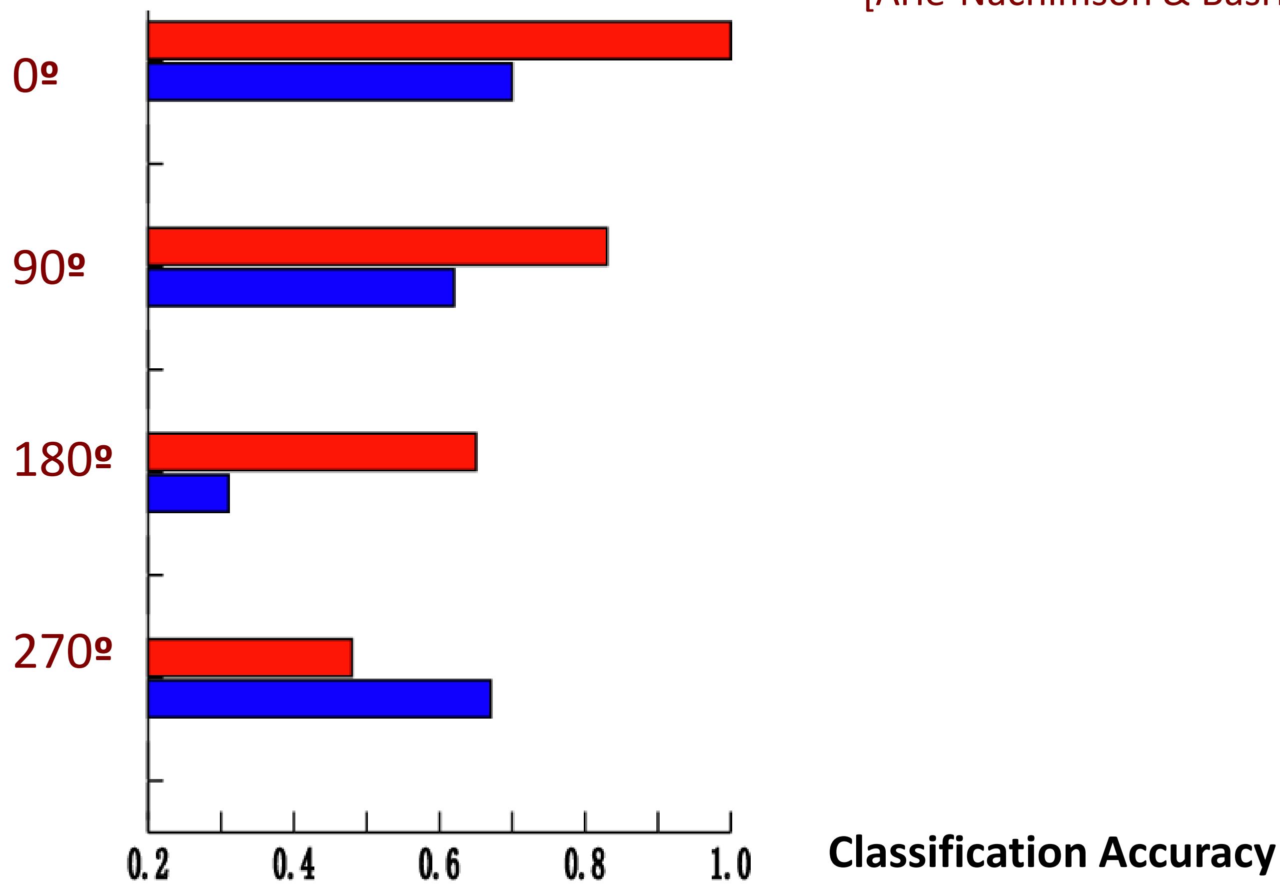


Viewpoint Classification

Car- Pascal 2006 dataset

First the time!

[Arie-Nachimson & Basri '09]

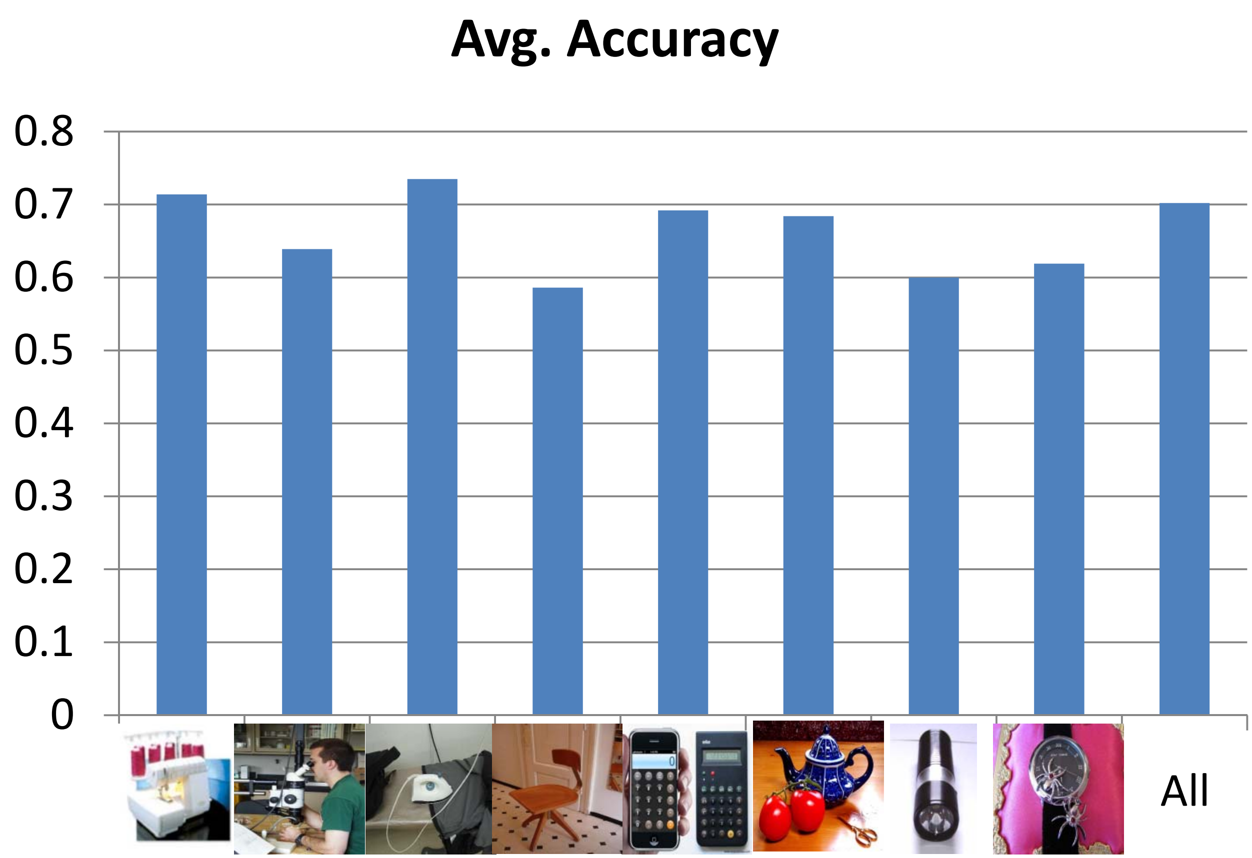


Our model

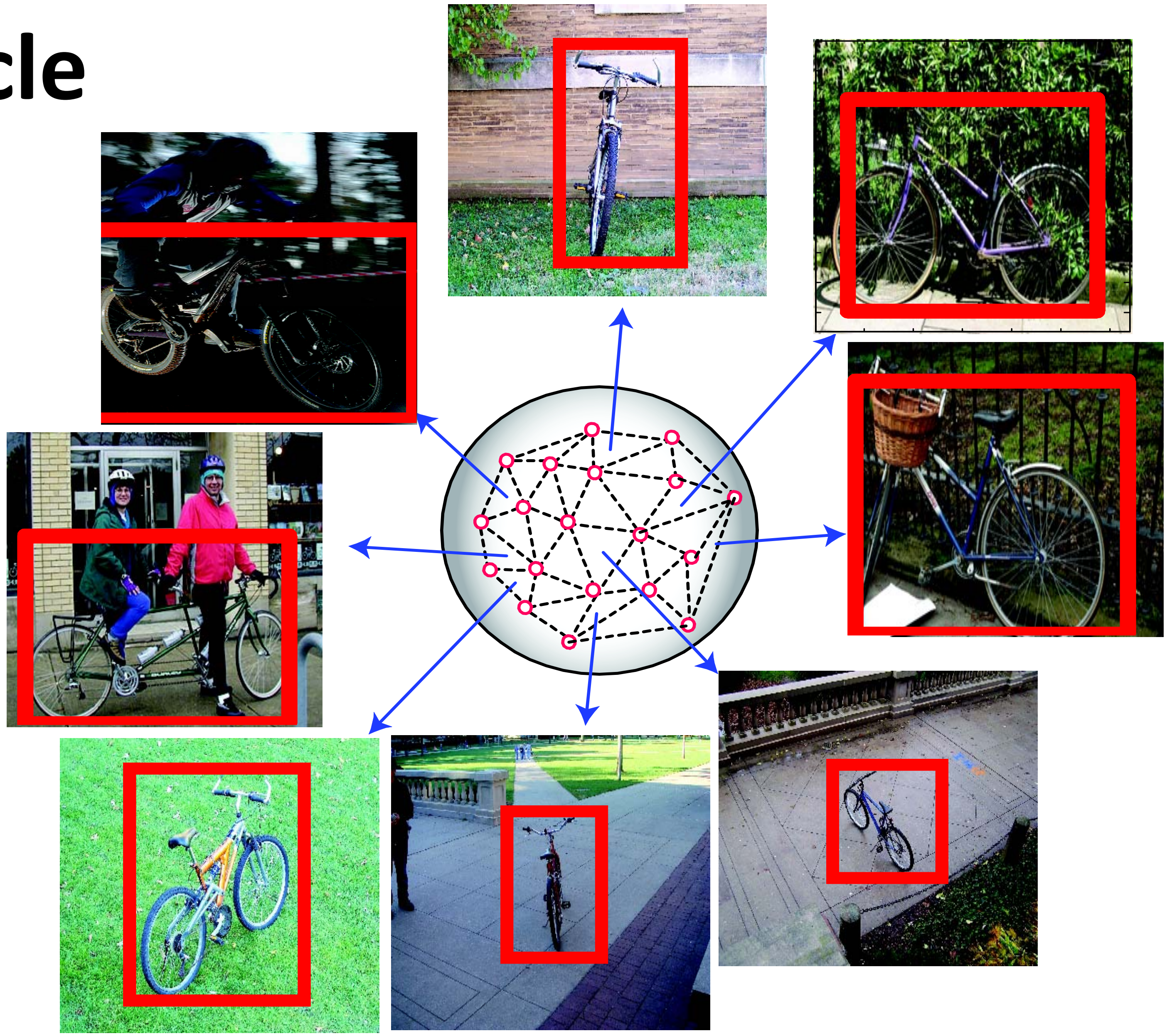
Min et al, CVPR 09

Viewpoint Classification

Household Item Dataset

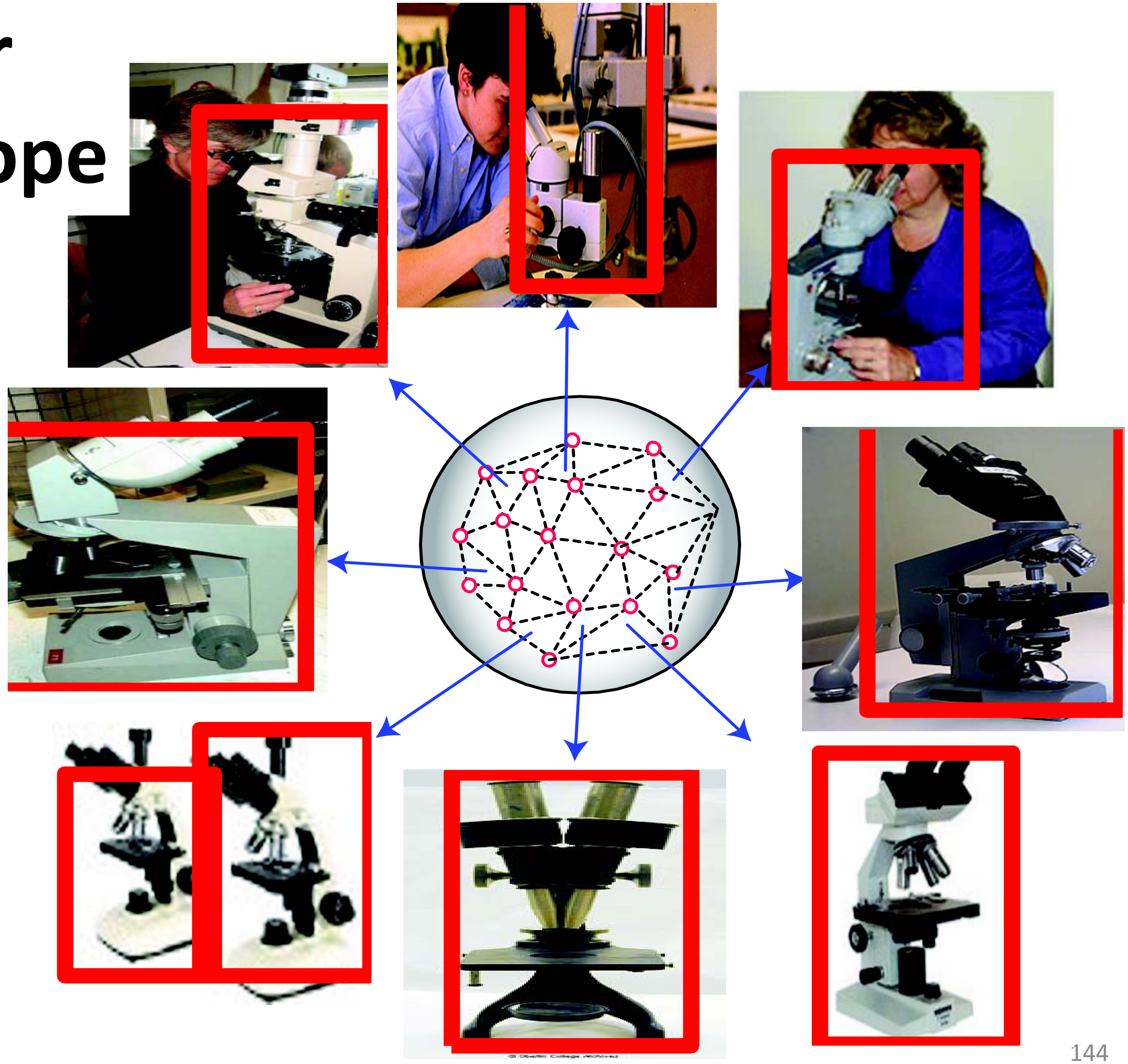


Bicycle

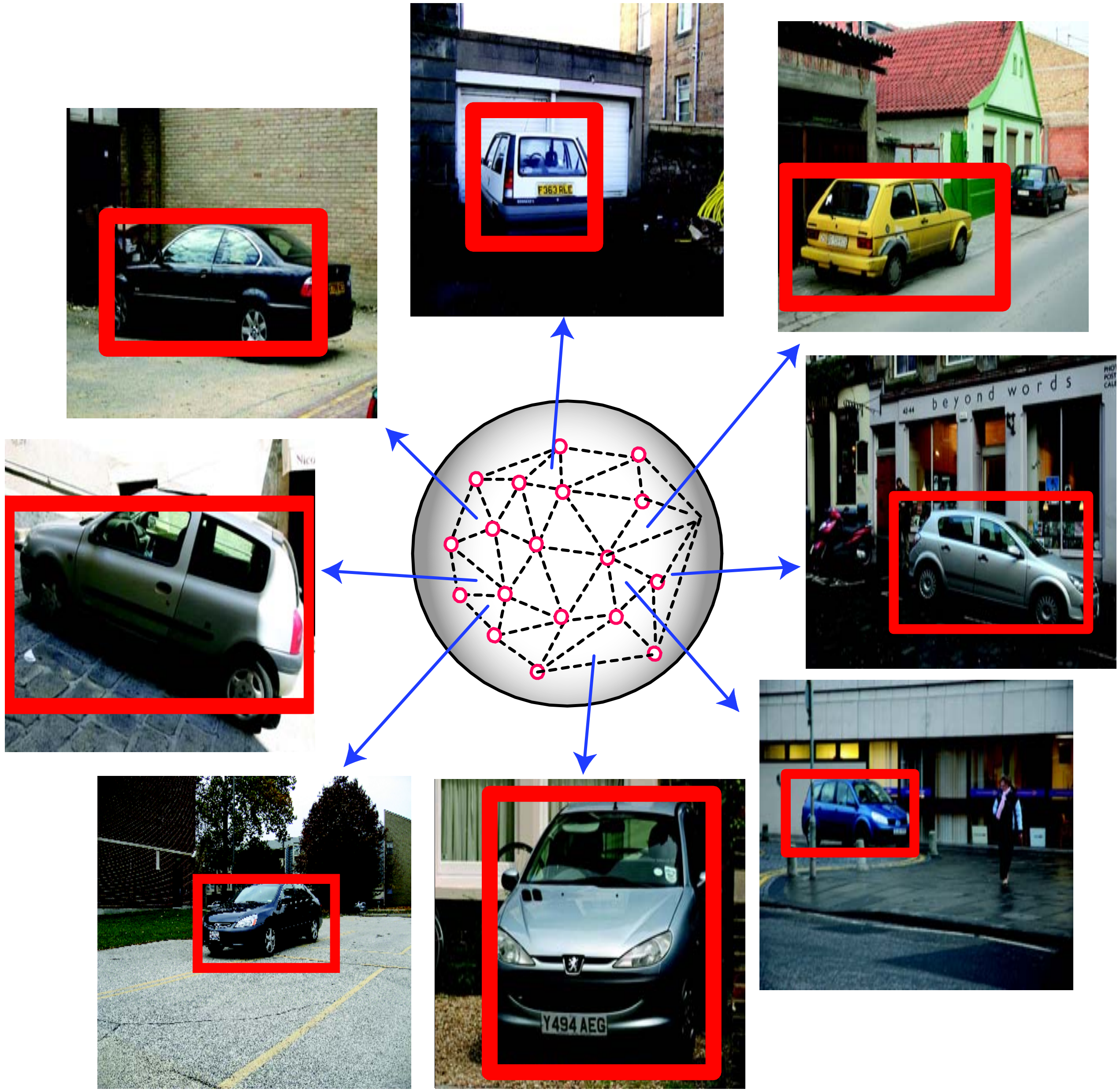


Notice the viewpoint variability in the dataset!

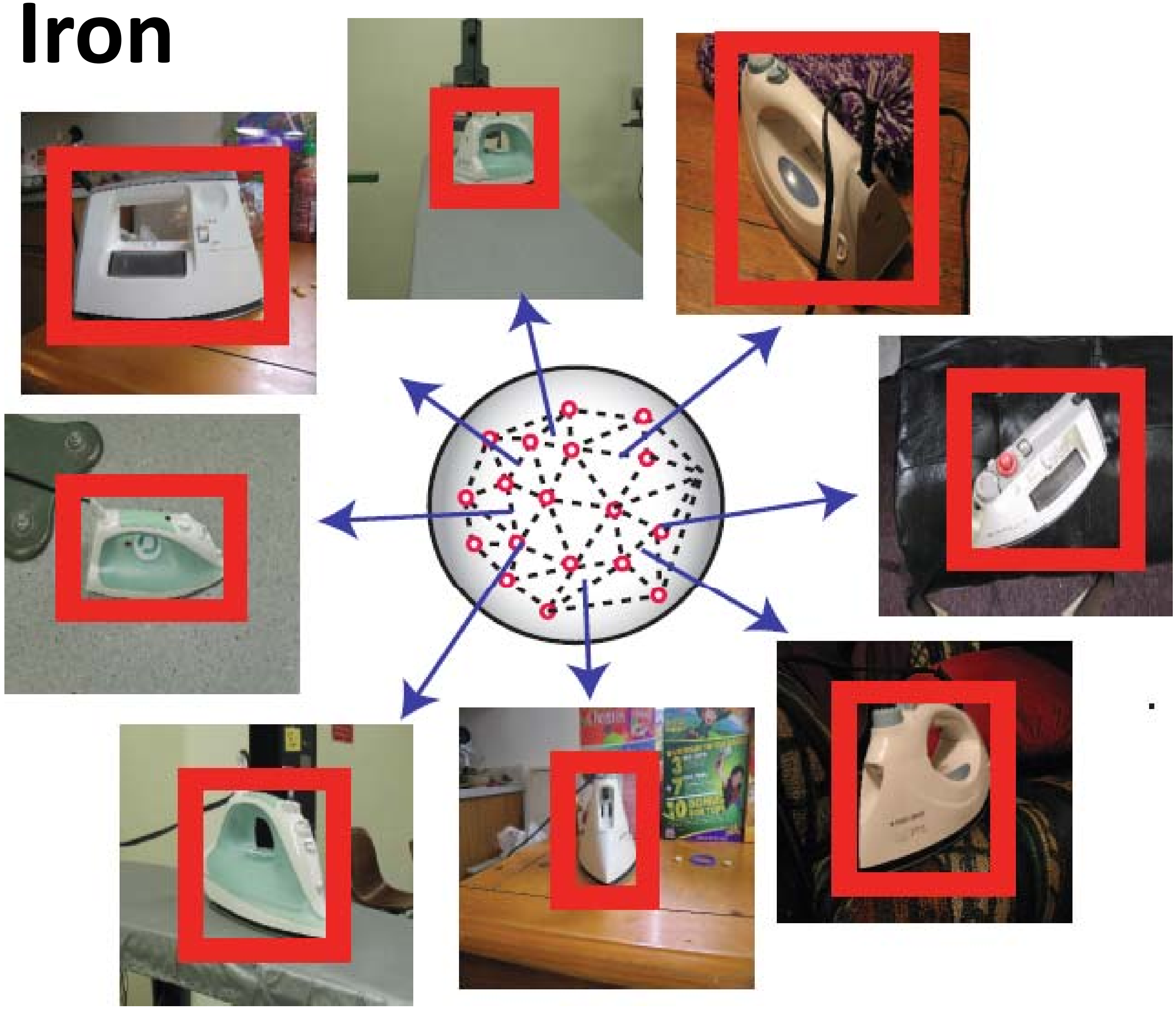
Binocular Microscope



Car



Travel Iron



Novel view object synthesis from a single image

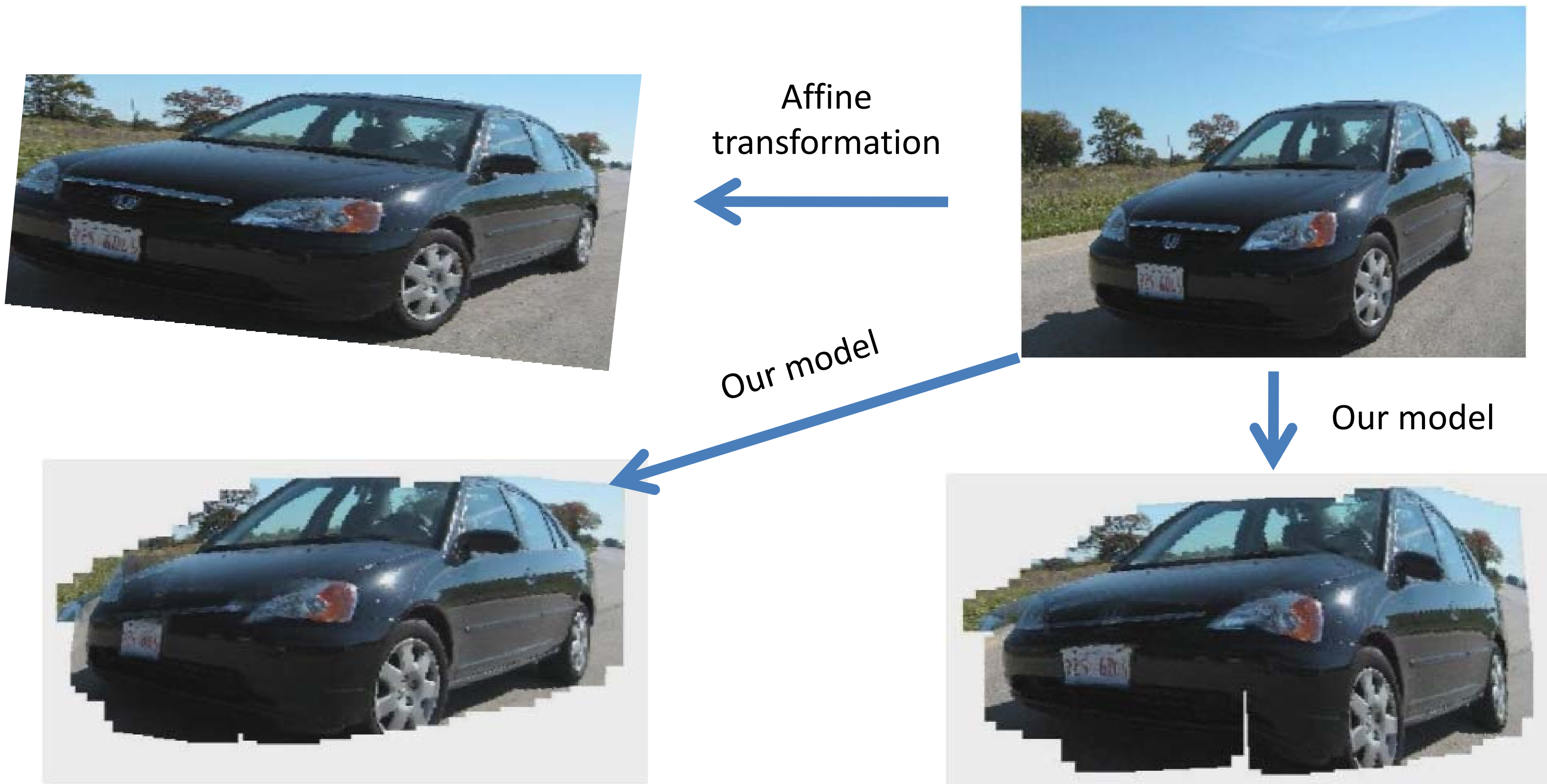
For the first time!

[For natural scenes, see Hoiem et al 07; Saxena et al 07]

Novel view object synthesis from a single image

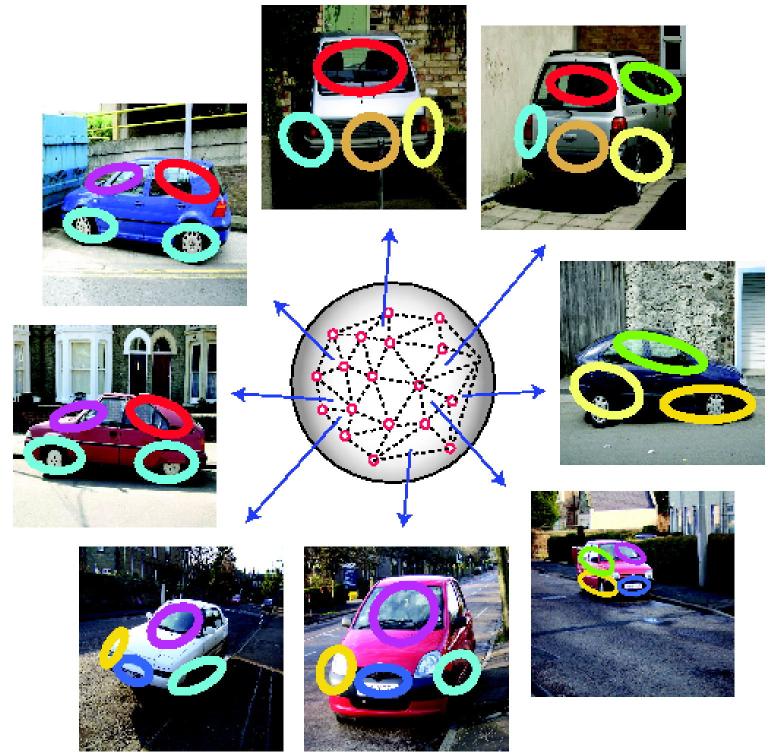
For the first time!

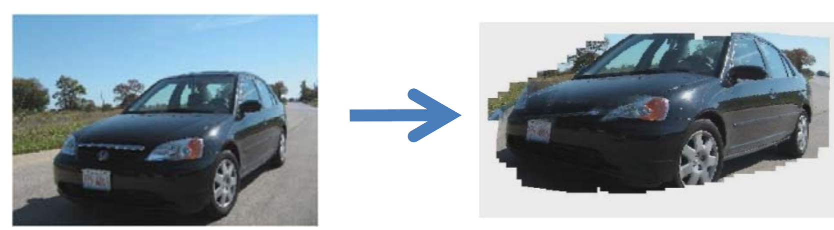
[For natural scenes, see Hoiem et al 07; Saxena et al 07]



Saravese Conclusions

- A **new part-based multi-view** representation for object categories
- **Incremental learning** scheme with **little supervision**
- Achieve **accurate pose estimation** tested on up to **16 categories**
- **Image based rendering from just one single image**!





Today

- Naïve-Bayes Nearest Neighbor (Irani)
- ISM (Liebe)
- Constellation Models (Fergus)
- Transformed LDA Models (Sudderth)
- 3-D view models (Saravese)