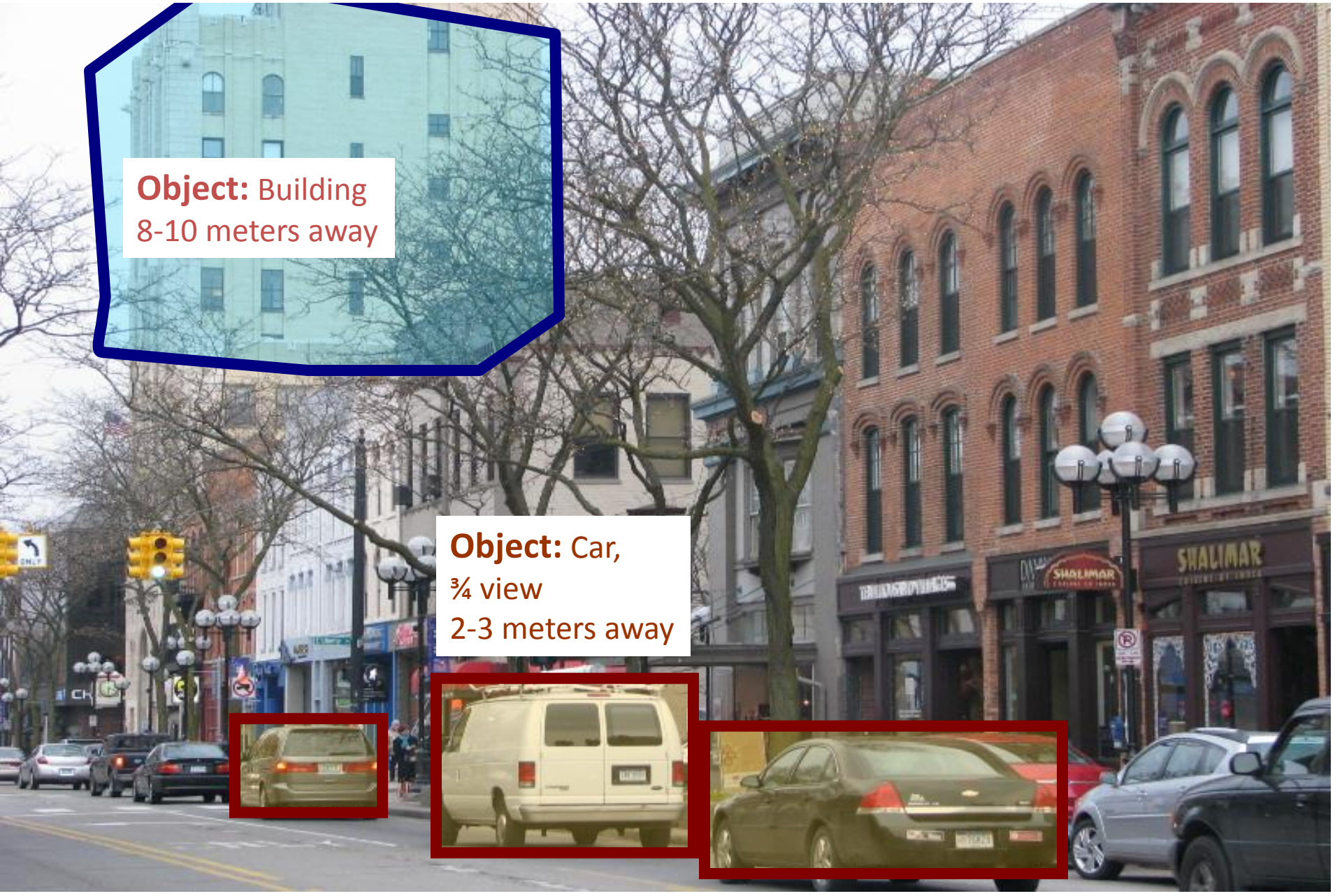
Lecture 15
Visual recognition

- 3D object detection
  - Introduction
  - Single instance 3D object detectors
  - Generic 3D object detectors

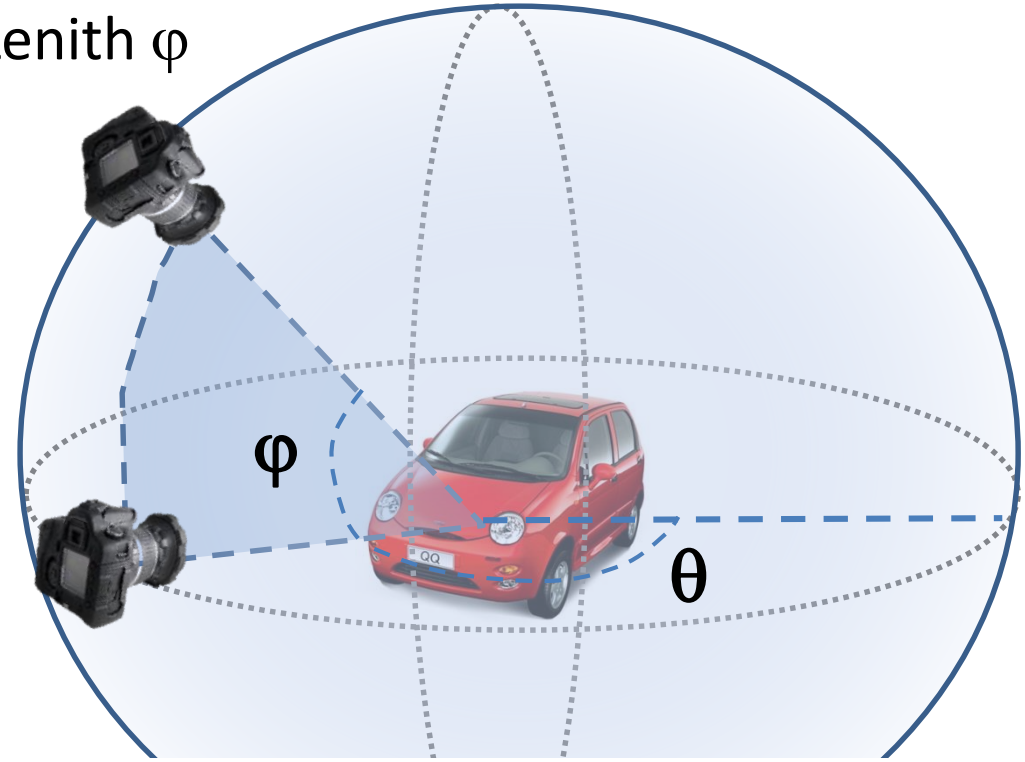
3D object detection



Properties of a 3D object detector

Viewing sphere

Azimuth θ, Zenith φ

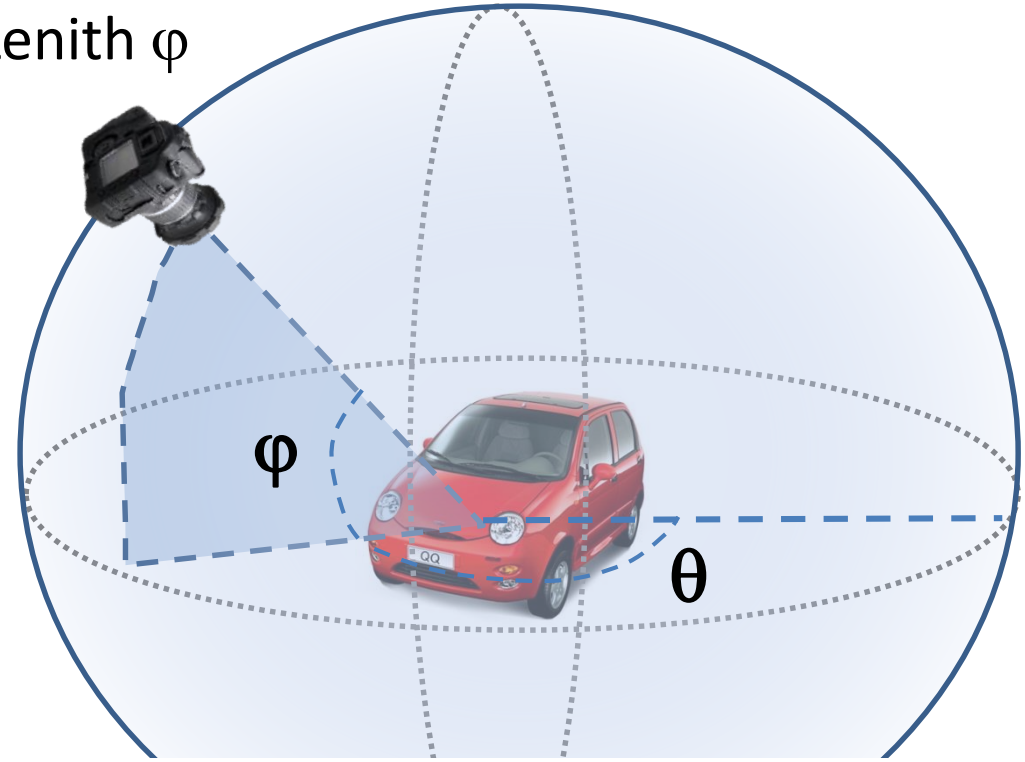


- Detect objects under generic view points
- Estimate object pose & 3D shape

Properties of a 3D object detector

Viewing sphere

Azimuth θ, Zenith φ

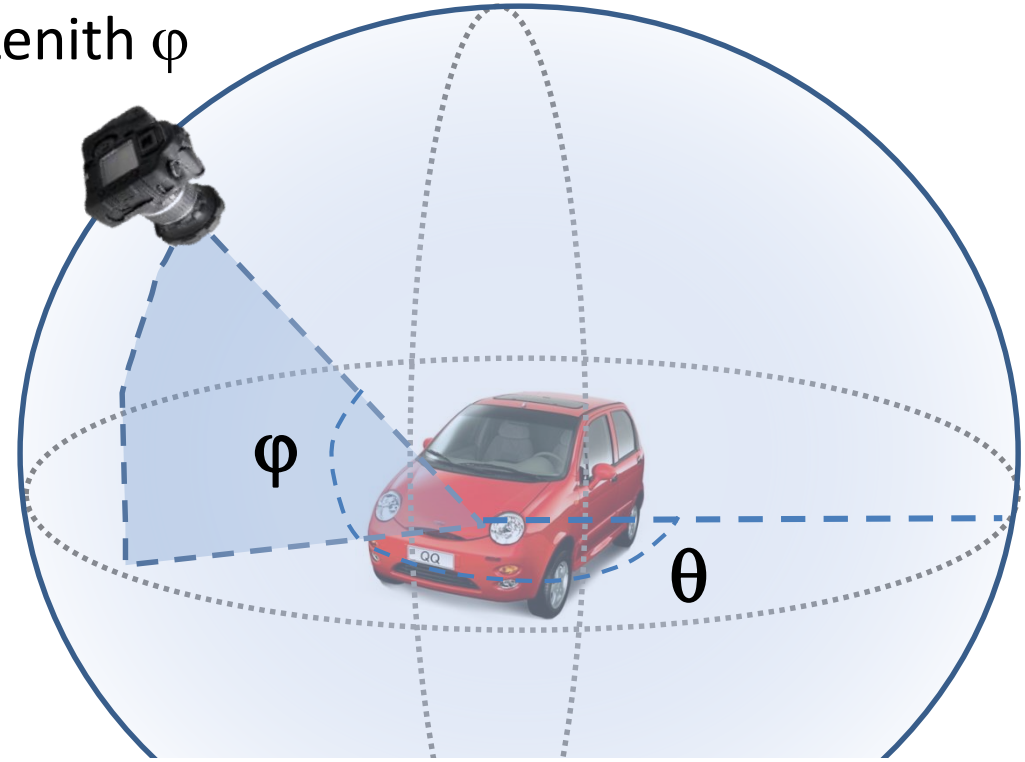


- Detect objects under generic view points
- Estimate object pose & 3D shape
- Work for object categories

Properties of a 3D object detector

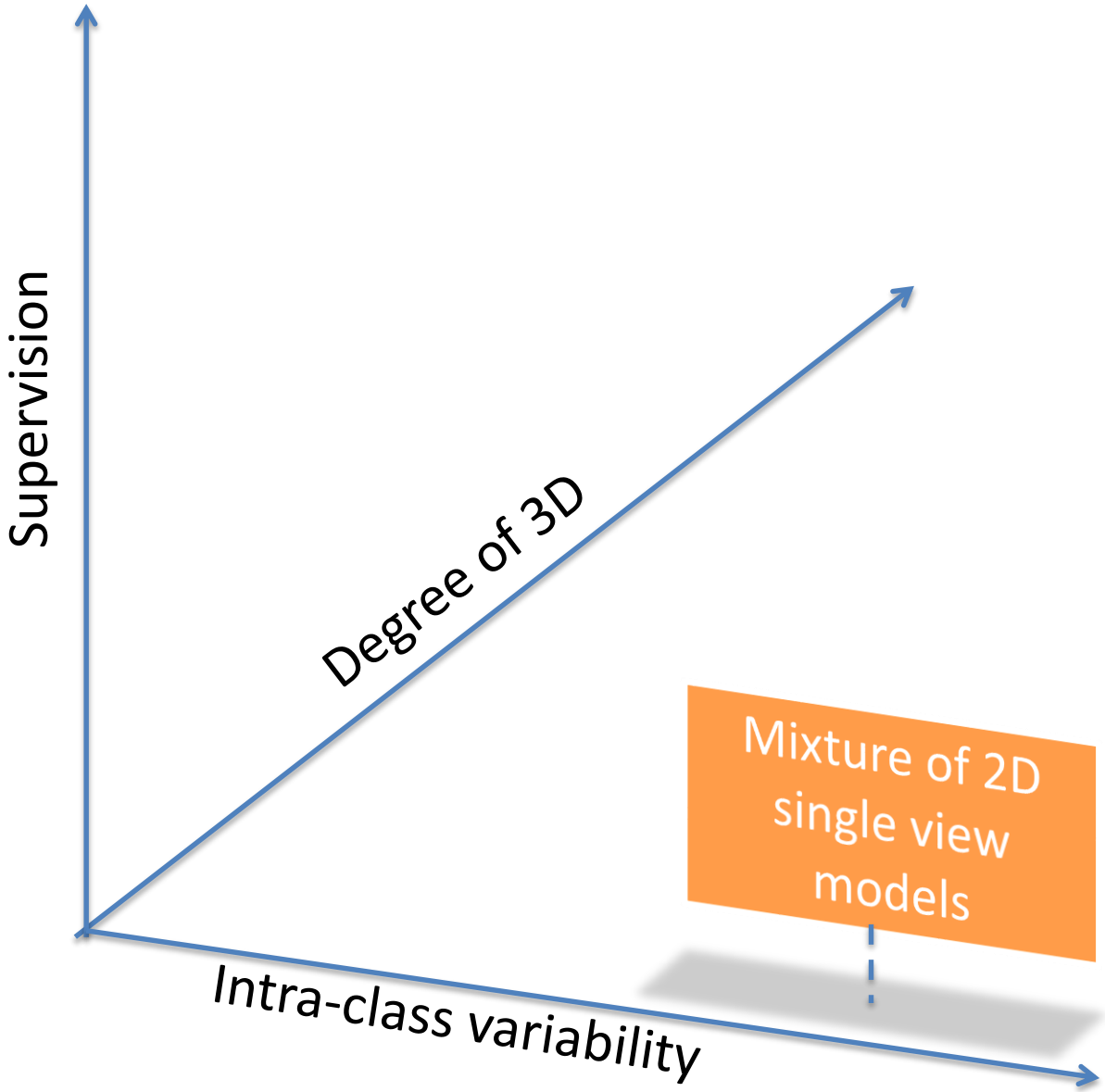
Viewing sphere

Azimuth θ, Zenith φ

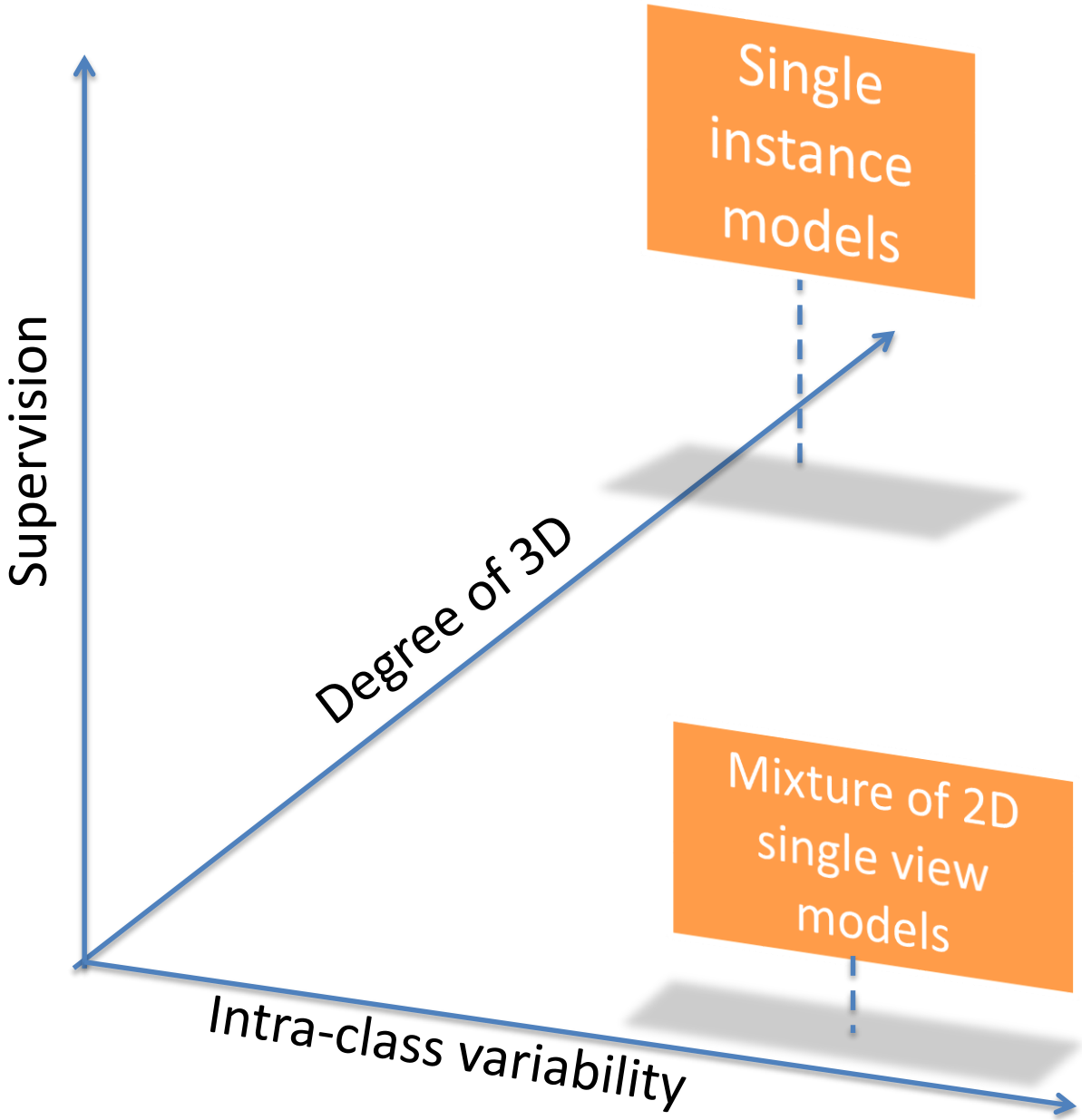


- Detect objects under generic view points
- Estimate object pose & 3D shape
- Work at different levels of specificity
- Limited amount of supervision

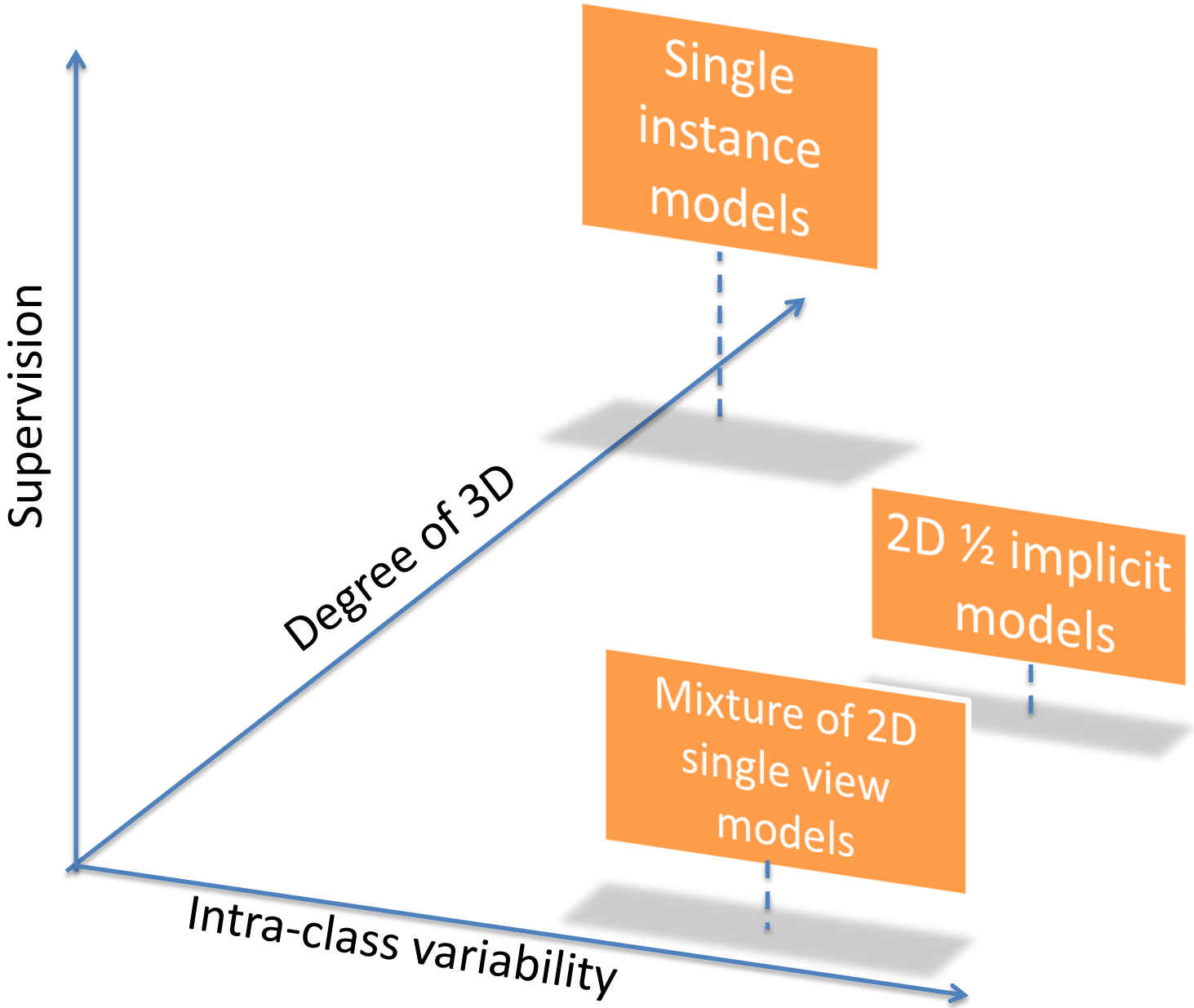
Models for 3d Object detection



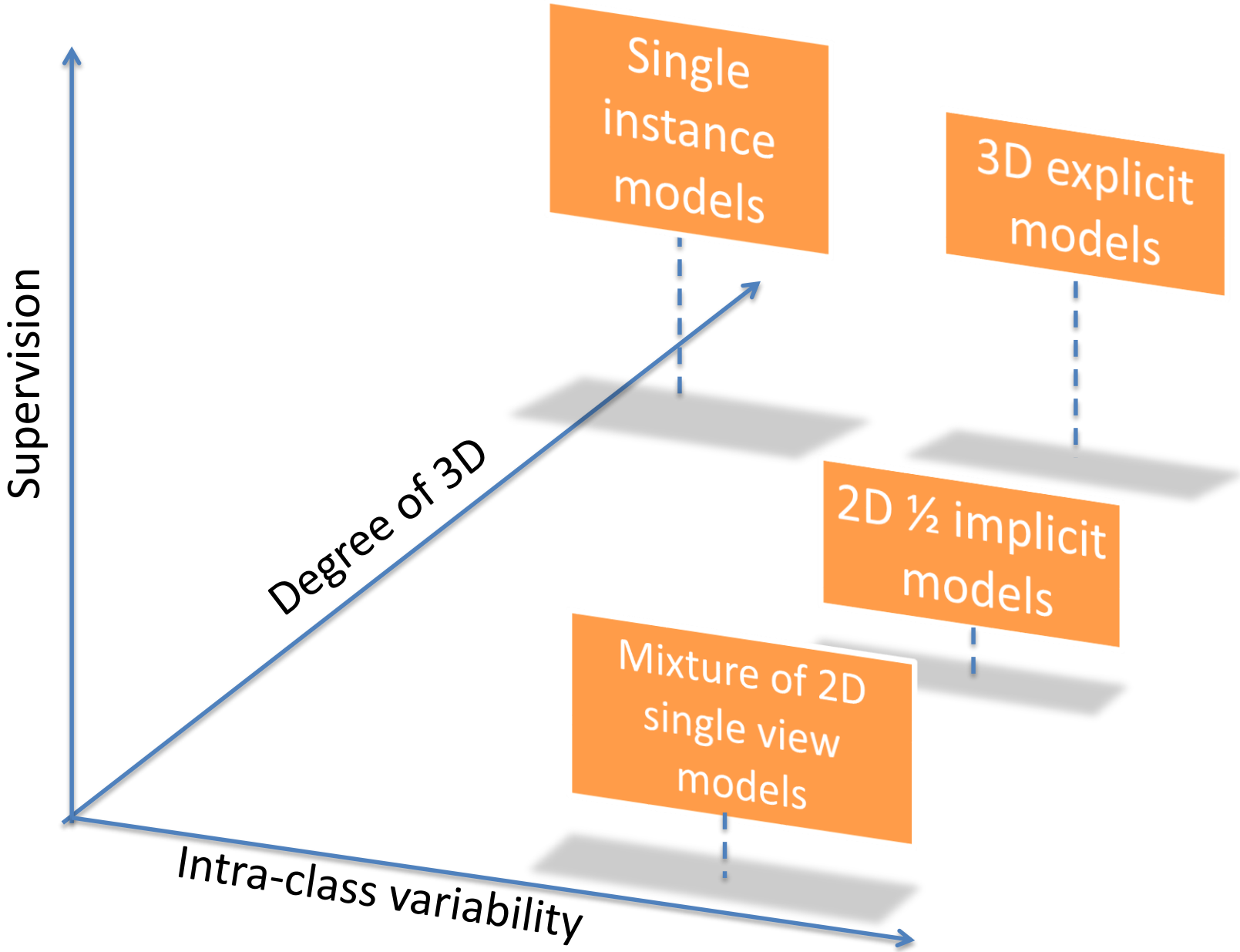
Models for 3d Object detection



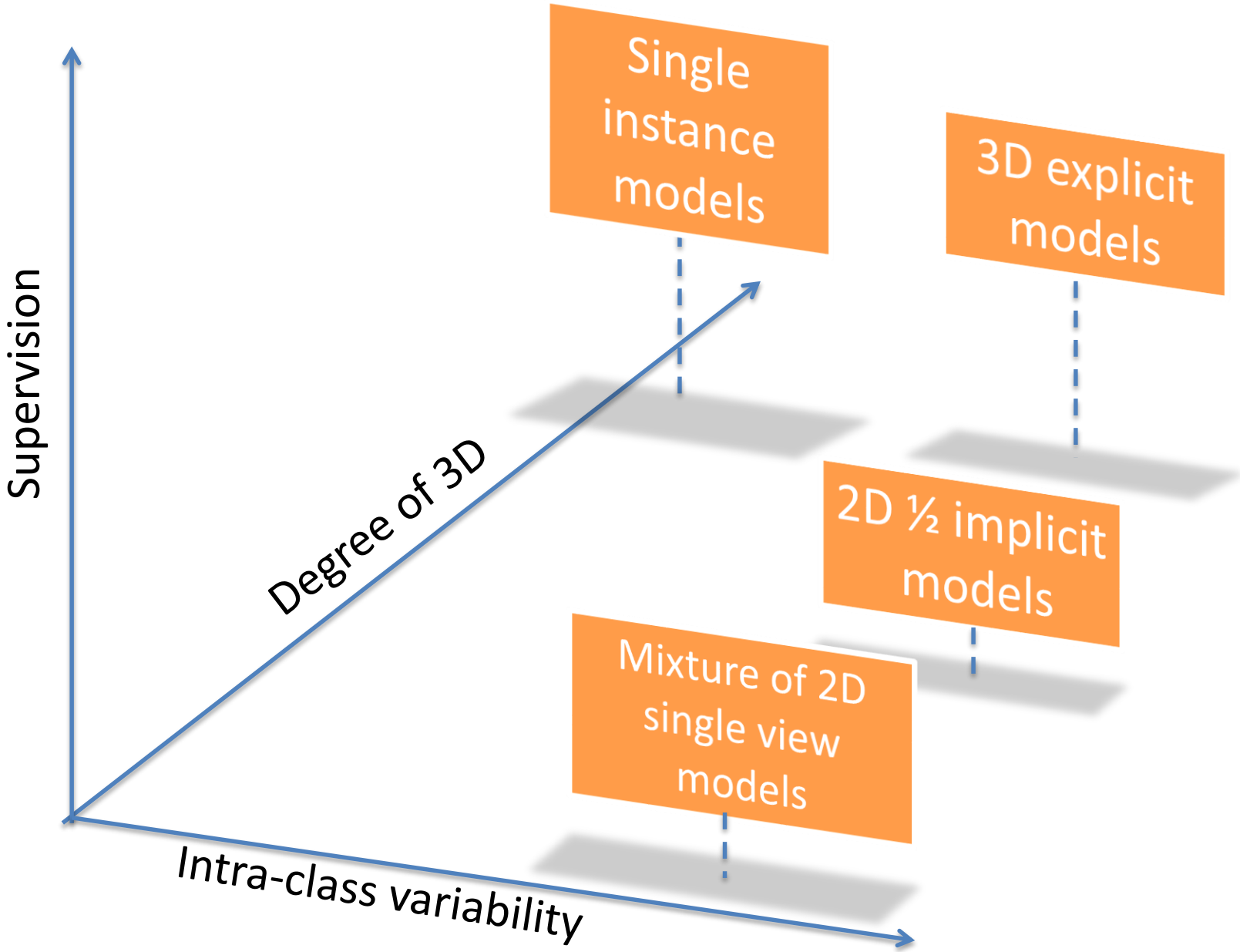
Models for 3d Object detection



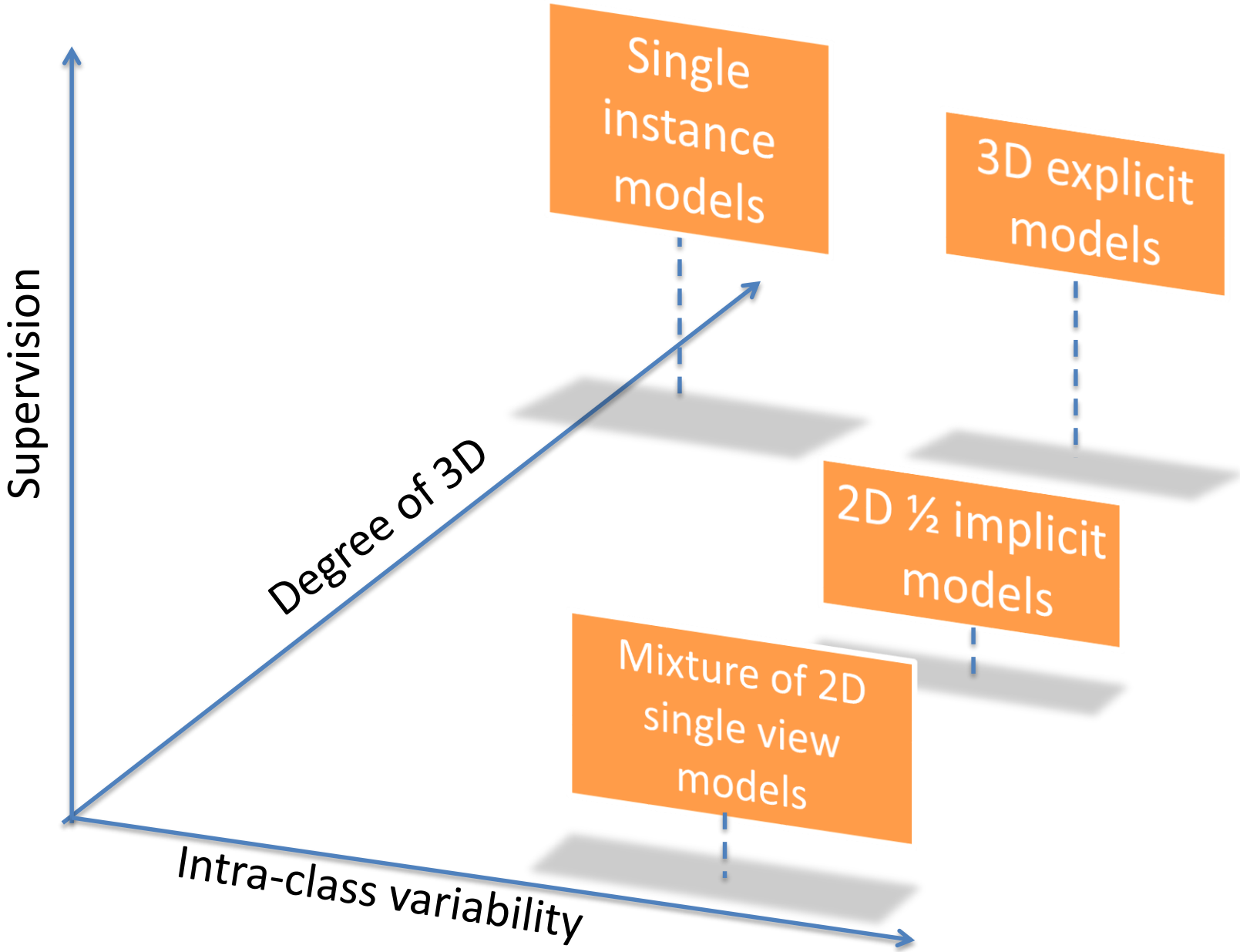
Models for 3d Object detection



Models for 3d Object detection



Models for 3d Object detection

















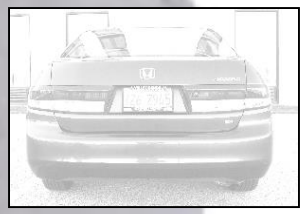




Single 3D object recognition









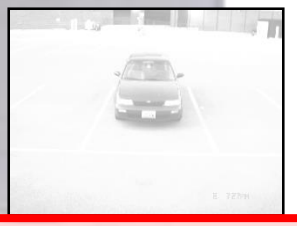










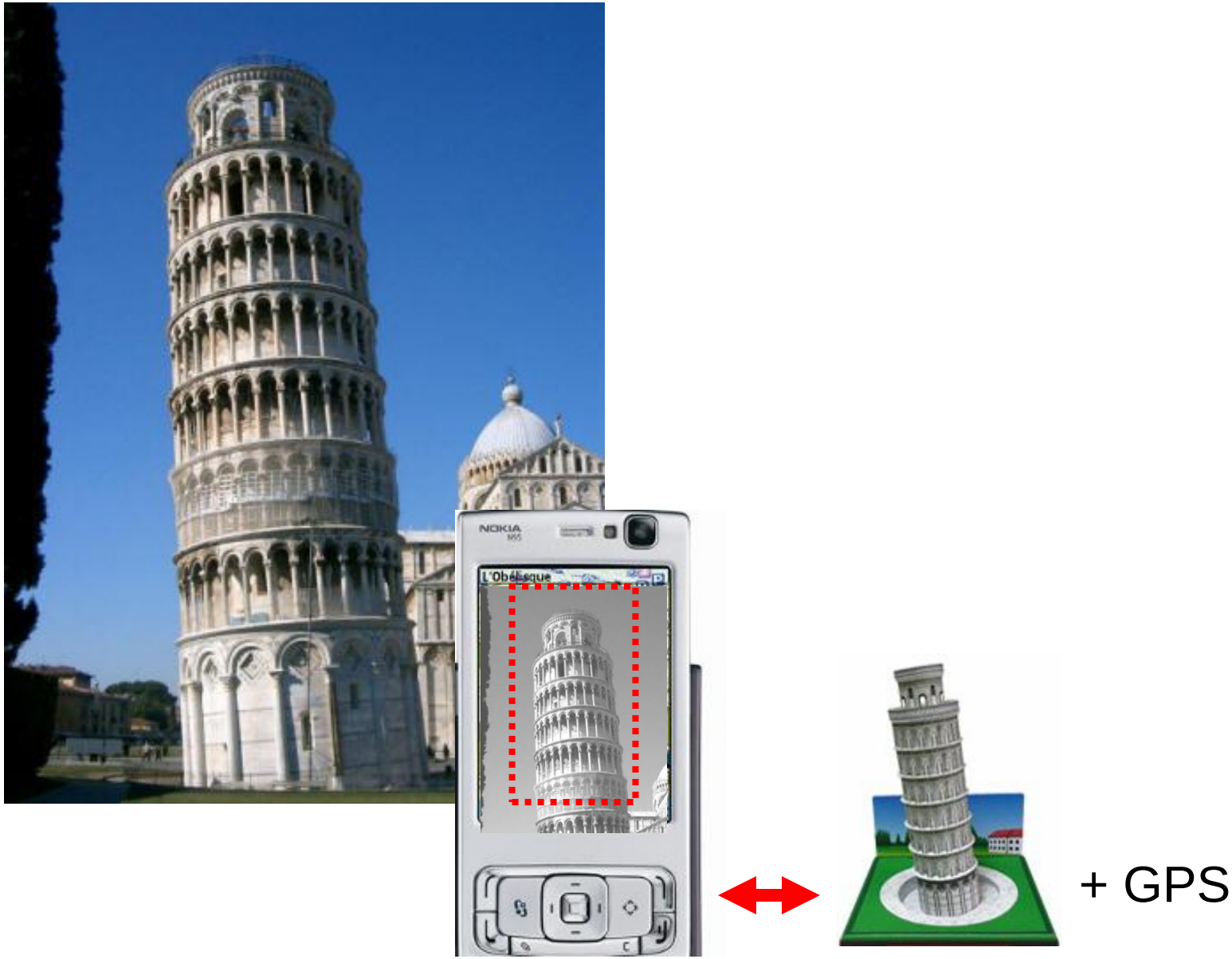


- Ballard, ‘81
- Grimson & L.-Perez, ‘87
- Lowe, ’87

- Edelman et al. ’91
- Ullman & Barsi, ’91
- Rothwell ‘92
- Linderberg, ’94
- Murase & Nayar ‘94

- Zhang et al ’95
- Schmid & Mohr, ‘96
- Schiele & Crowley, ’96
- **Lowe, ‘99**
- **Jacob & Barsi, ‘99**
- Mahamud and Herbert, 00

- **Rothganger et al., ‘04**
- **Ferrari et al, ’05**
- Moreels and Perona, 05
- **Brown & Lowe ’05**
- **Snavely et al ’06**
- **Yin & Collins, ‘07**


NOKIA
L'Obélisque
+ GPS

Where is the crunchy nut?

Usual Challenges:

Variability due to:

- View point
- Illumination
- Occlusions

But… no intra-class variability

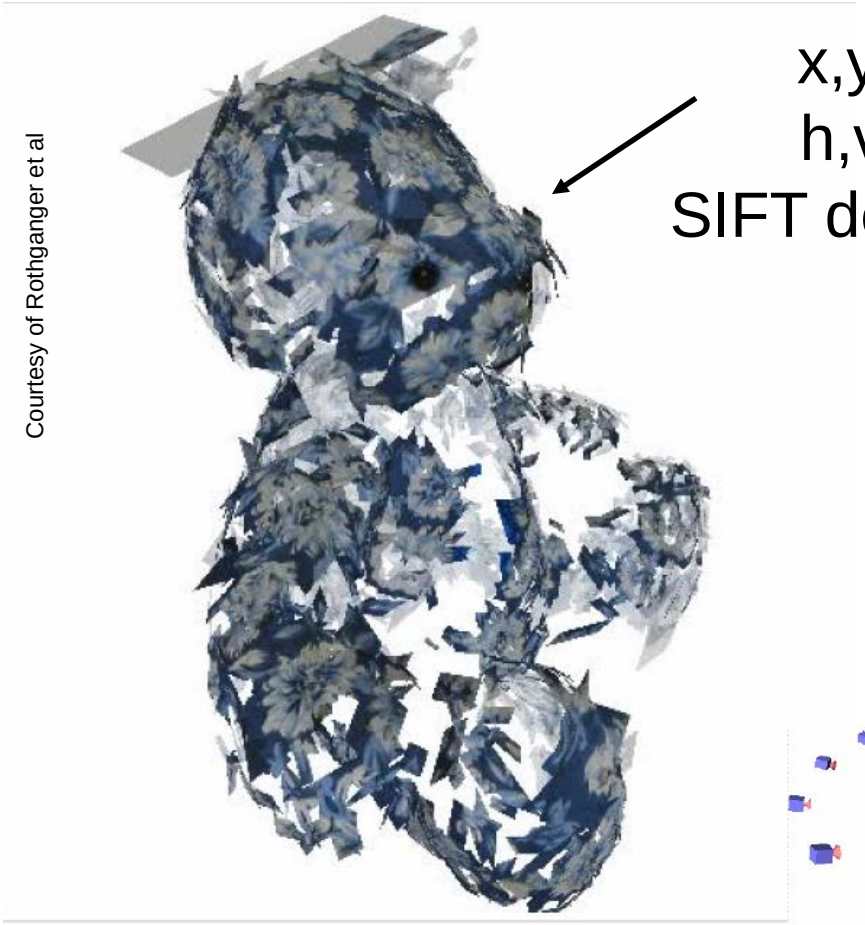
Recognition of single 3D objects

- **Representation**
  - **Features**
  - **2D/3D Geometrical constraints**
- **Model learning**
- **Recognition**
  - **Hypothesis generation**
  - **Validation**

- Marr ’78, ‘82
- Ballard, ‘81
- Grimson & L.-Perez, ‘87
- Lowe, ’87
- Forsyth et al. ‘91
- Edelman et al. ’91
- Ullman & Barsi, ’91
- Rothwell ‘92
- Linderberg, ’94
- Murase & Nayar ‘94

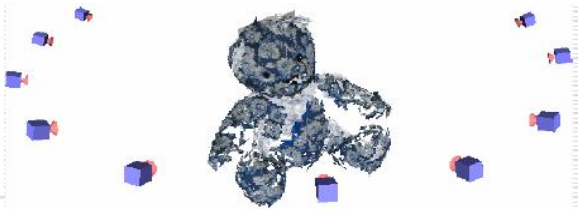
- Rothganger et al. ’04, ’06
- Brown et al, ‘05
- Lowe ‘99, ‘04
- Ferrari et al. ’04, ‘06
- Lazebnick et al ’04
- Hsiao et al., ‘10

Object representation:
2D or 3D location of key points



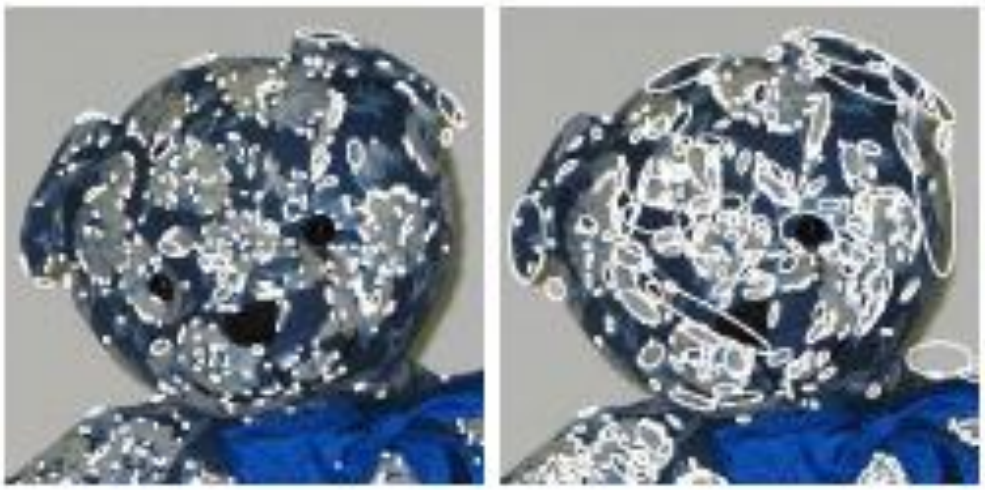
Courtesy of Rothganger et al

x,y,z +
h,v +
SIFT descriptor



Detection

Harris-Laplace: used in Rothganger et al. ‘06

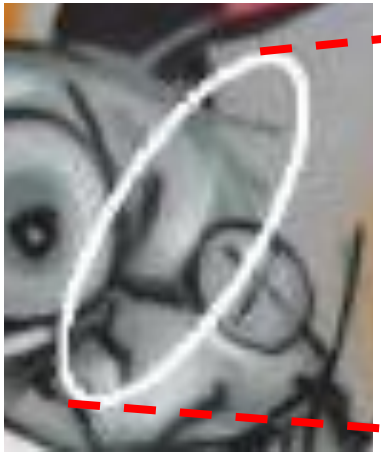


Courtesy of Rothganger et al.

- x,y
- Scale
- Orientation
- Affine structure

“View” invariant descriptions

View 1



Scale, rotation

SIFT

View 2

SIFT

Basic scheme

- Representation
  - Features
  - 2D/3D Geometrical constraints
- **Model learning**
- Recognition
  - hypothesis generation
  - validation

Model learning

Rothganger et al. ‘03 ’06

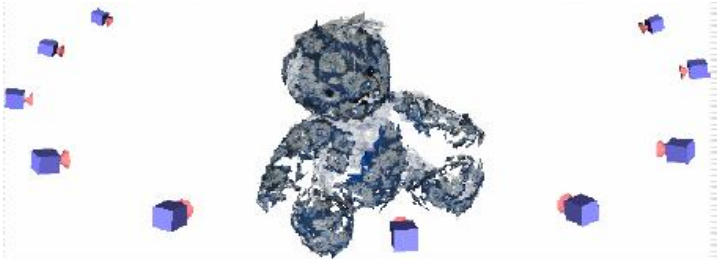
Build a 3D model:

- N images of object from N different views
- Extract key points from each view
- Match key points between 2 views

- Use affine structure from motion to compute 3D location and orientation + camera locations from 2 views
- Find connected components
- Use bundle adjustment to refine the model

$$E = \sum_{j=1}^{n} \sum_{i \in I_j} |\mathcal{S}_{ij} - \mathcal{M}_i \mathcal{N}_j|^2,$$

- Upgrade model to Euclidean assuming zero skew and square pixels



Learnt models

Rothganger et al. ‘03 ’06

Courtesy of Rothganger et al

Basic scheme

- **Representation**
  - **Features**
  - **2D/3D Geometrical constraints**
- **Model learning**
- **Recognition** **[object instance from object model]**
  - **hypothesis generation**
  - **Model verification**

Recognition

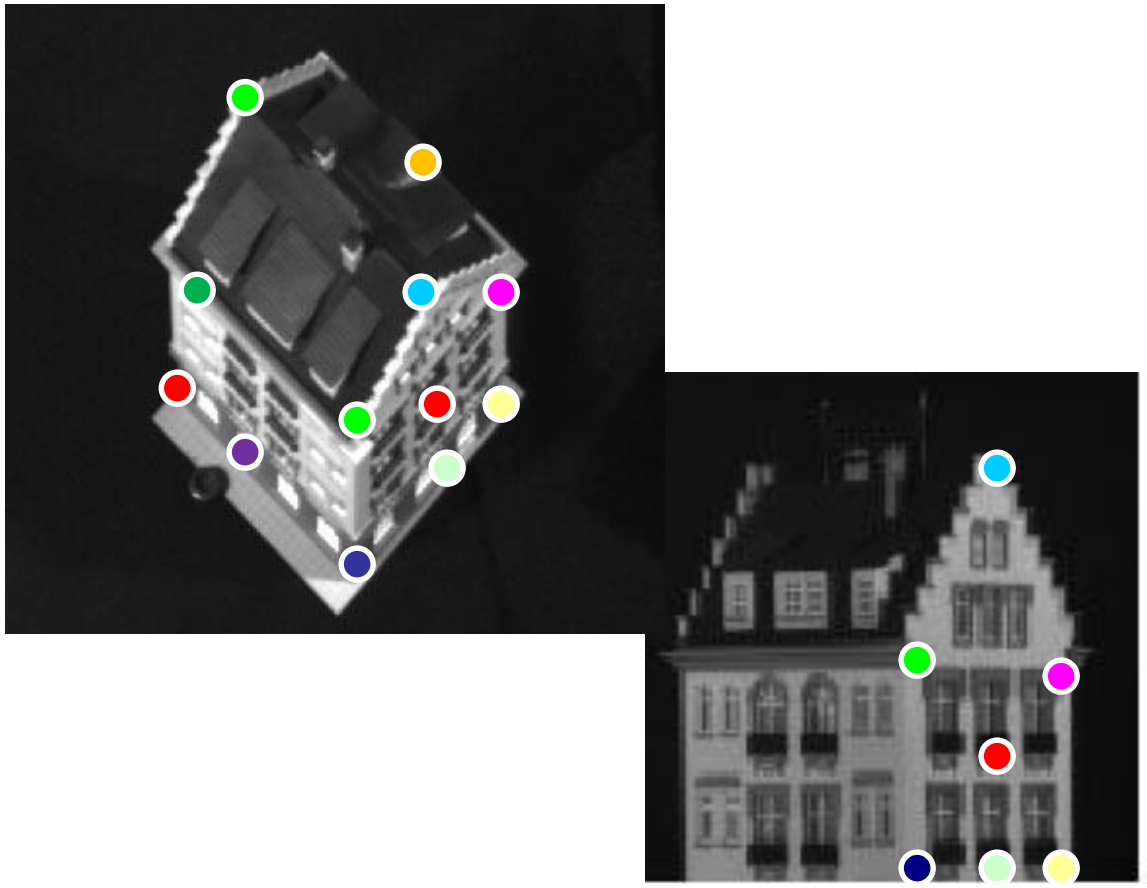
Goal: given a query image I, identify object model in the image I (match learned model to I)

- Generate hypothesis
- Verify hypothesis
- Select hypothesis with lowest fitting error
- Generate recognition results

Recognition

Goal: given a query image I, identify object model in the image I (match learned model to I)

query



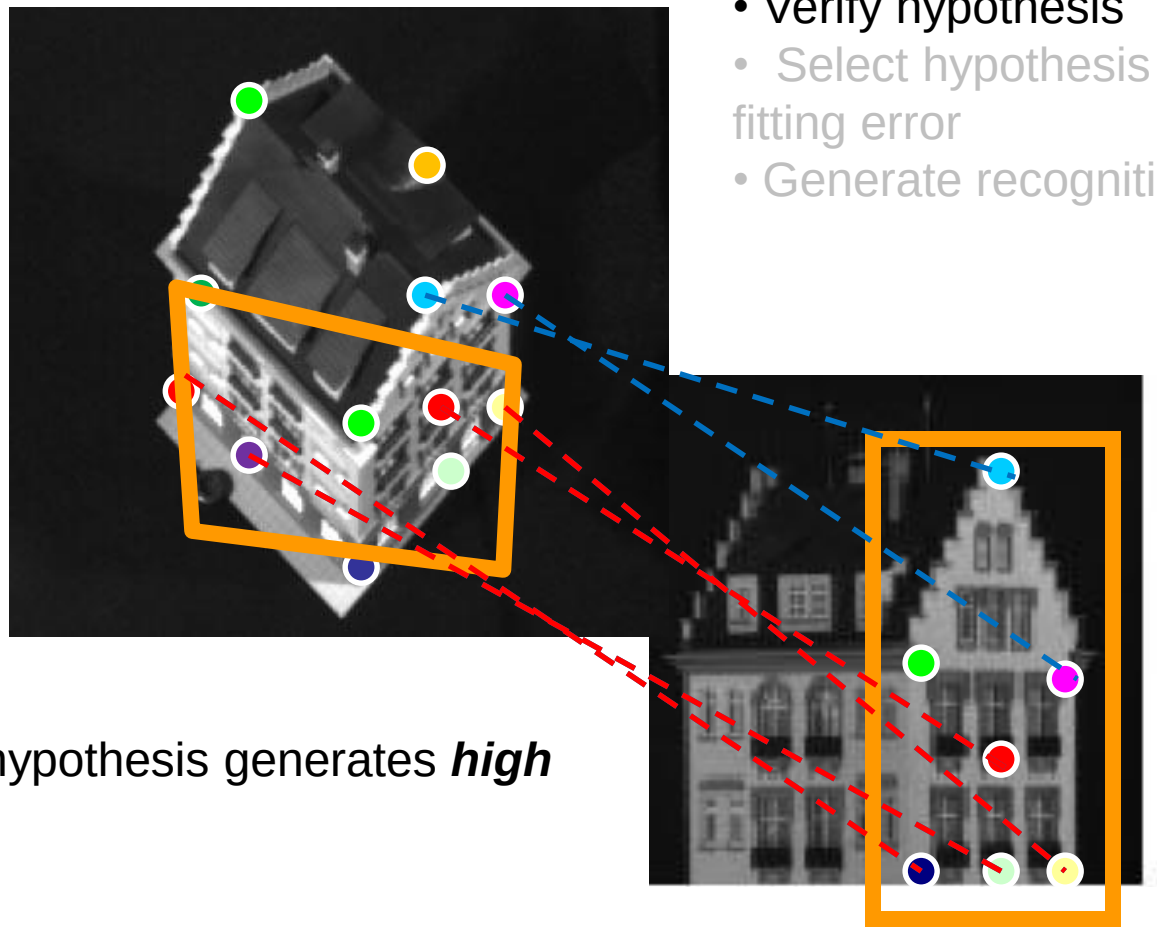
model

Recognition

Goal: given a query image I, identify object model in the image I (match learned model to I)

- Generate hypothesis
- Verify hypothesis
- Select hypothesis with lowest fitting error
- Generate recognition results

query



model

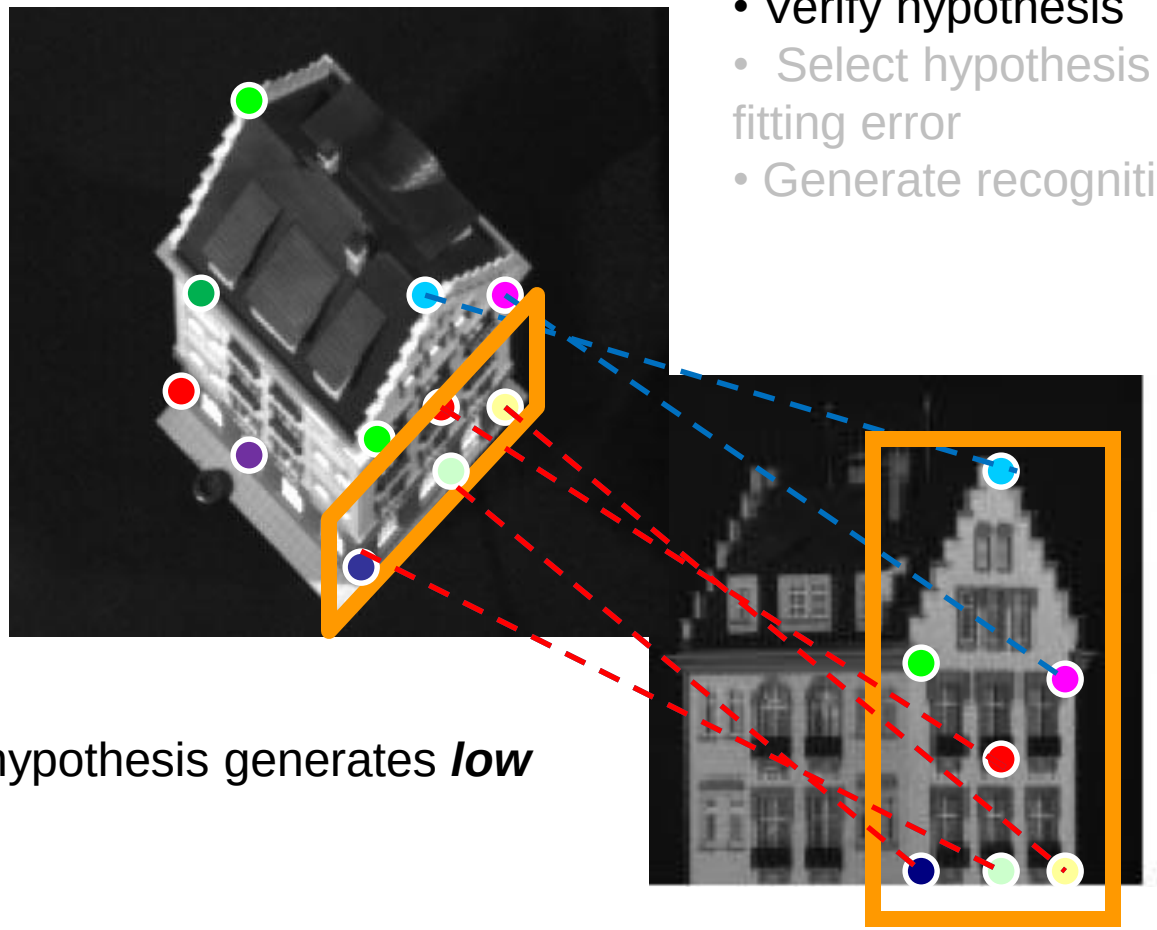
Verification: The hypothesis generates ***high*** fitting error

Recognition

Goal: given a query image I, identify object model in the image I (match learned model to I)

- Generate hypothesis
- Verify hypothesis
- Select hypothesis with lowest fitting error
- Generate recognition results

query



model

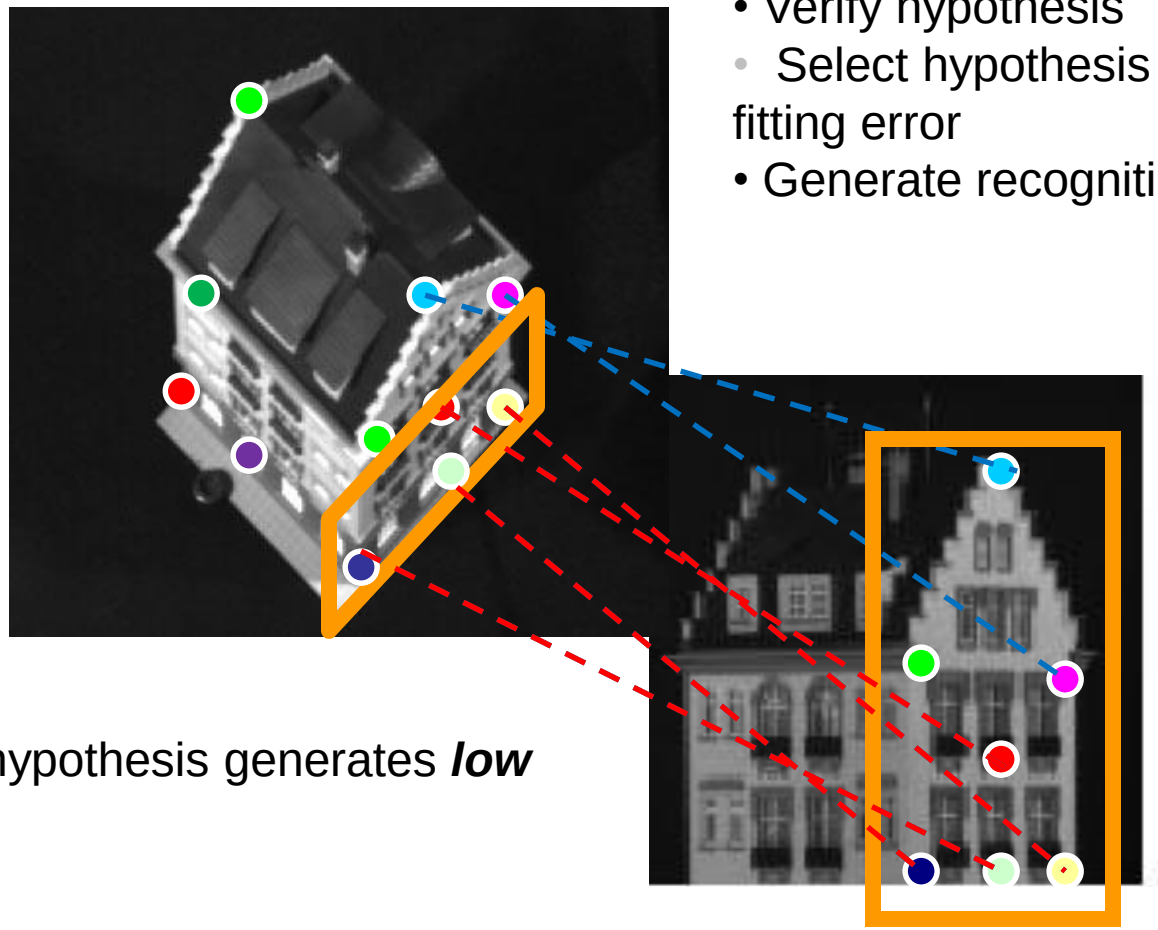
Verification: The hypothesis generates ***low*** fitting error

Recognition

Goal: given a query image I, identify object model in the image I (match learned model to I)

- Generate hypothesis
- Verify hypothesis
- Select hypothesis with lowest fitting error
- Generate recognition results

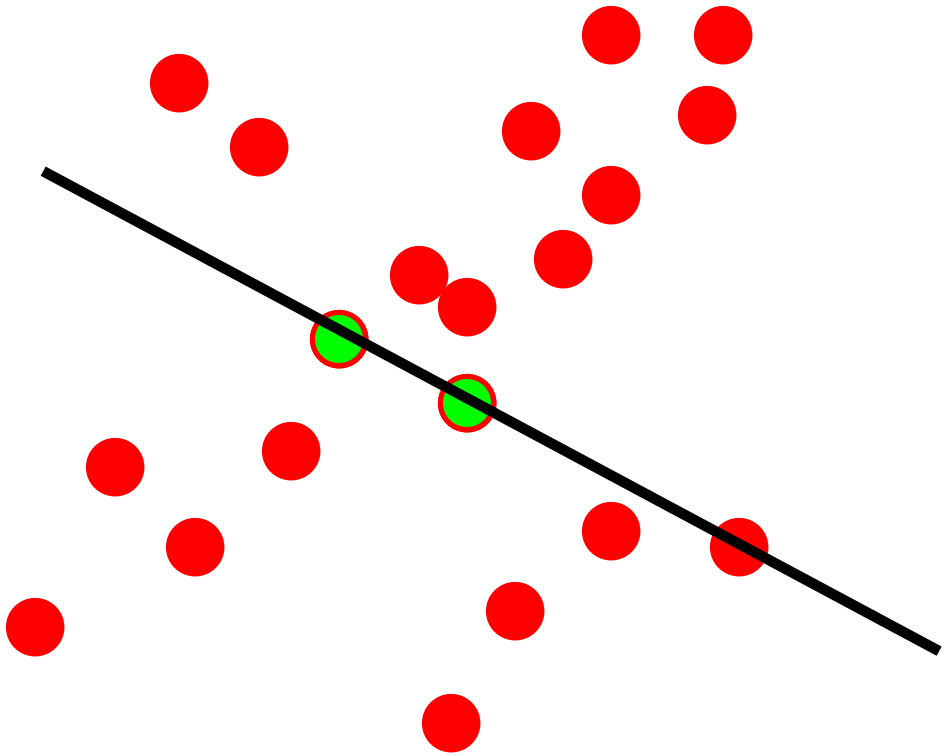
query



model

Verification: The hypothesis generates ***low*** fitting error

Line fitting with outliers



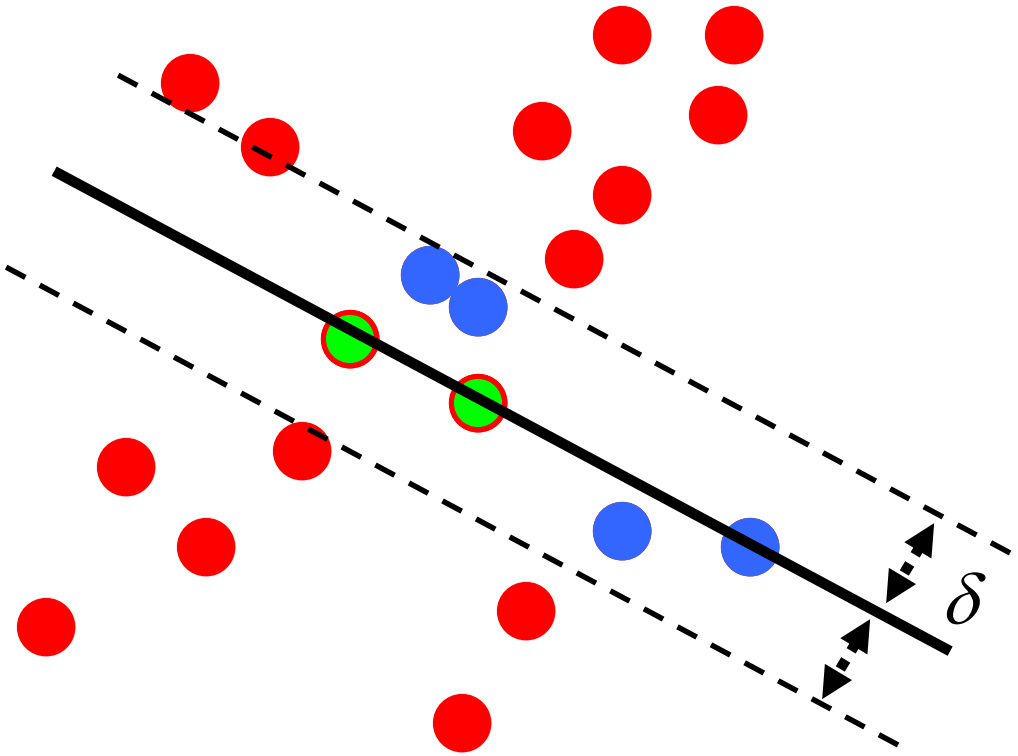
Sample set = set of points in 2D

Algorithm:

1. Select random sample of minimum required size to fit model [?] =[2]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1-3 until model with the most inliers over all samples is found

Line fitting with outliers



Sample set = set of points in 2D

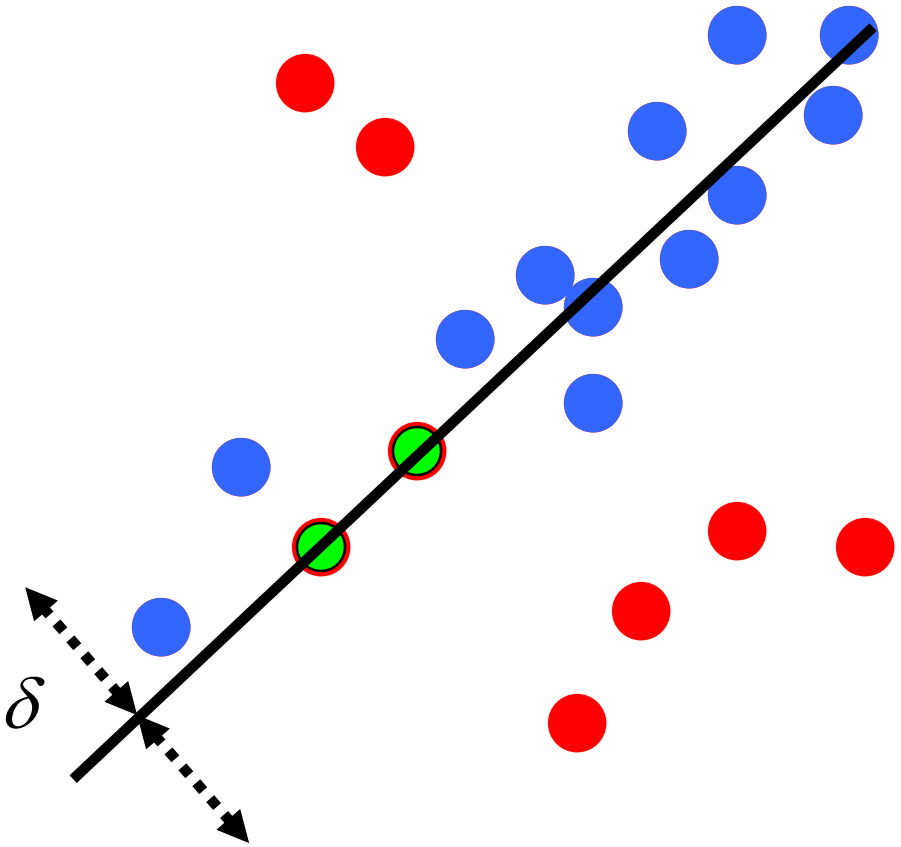
$|O| = 14$

Algorithm:

1. Select random sample of minimum required size to fit model [?] =[2]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1-3 until model with the most inliers over all samples is found

Line fitting with outliers



$|O| = 6$

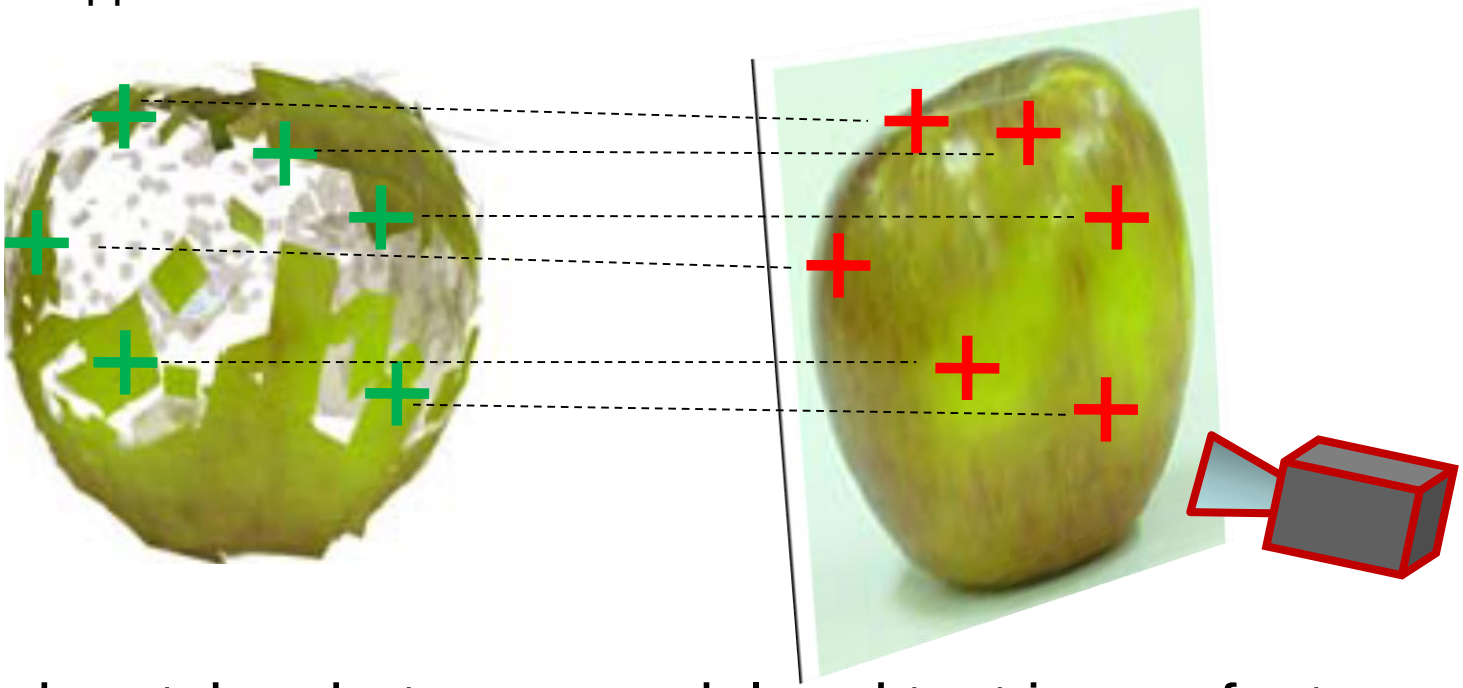
Algorithm:

1. Select random sample of minimum required size to fit model [?]
2. Compute a putative model from sample set
3. Compute the set of inliers to this model from whole data set

Repeat 1-3 until model with the most inliers over all samples is found

Recognition

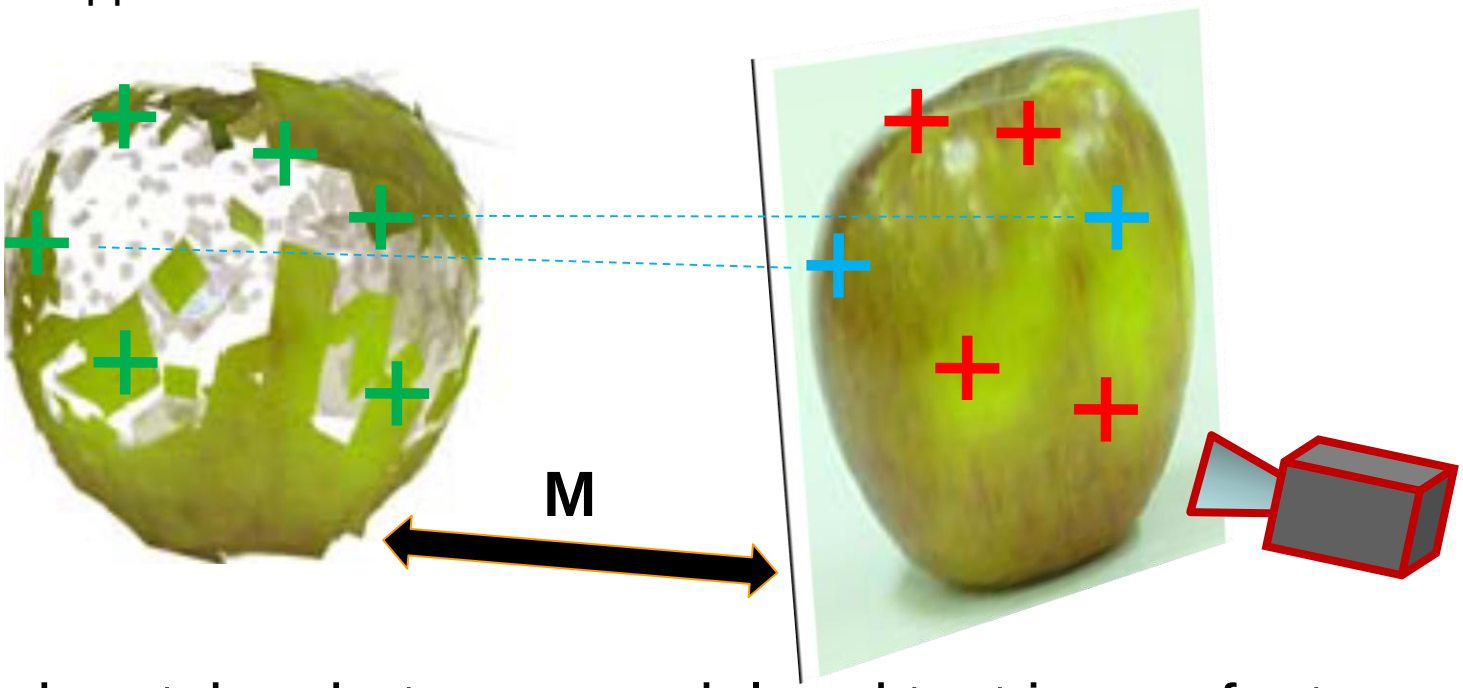
Class: apple



1. Find matches between model and test image features

Recognition

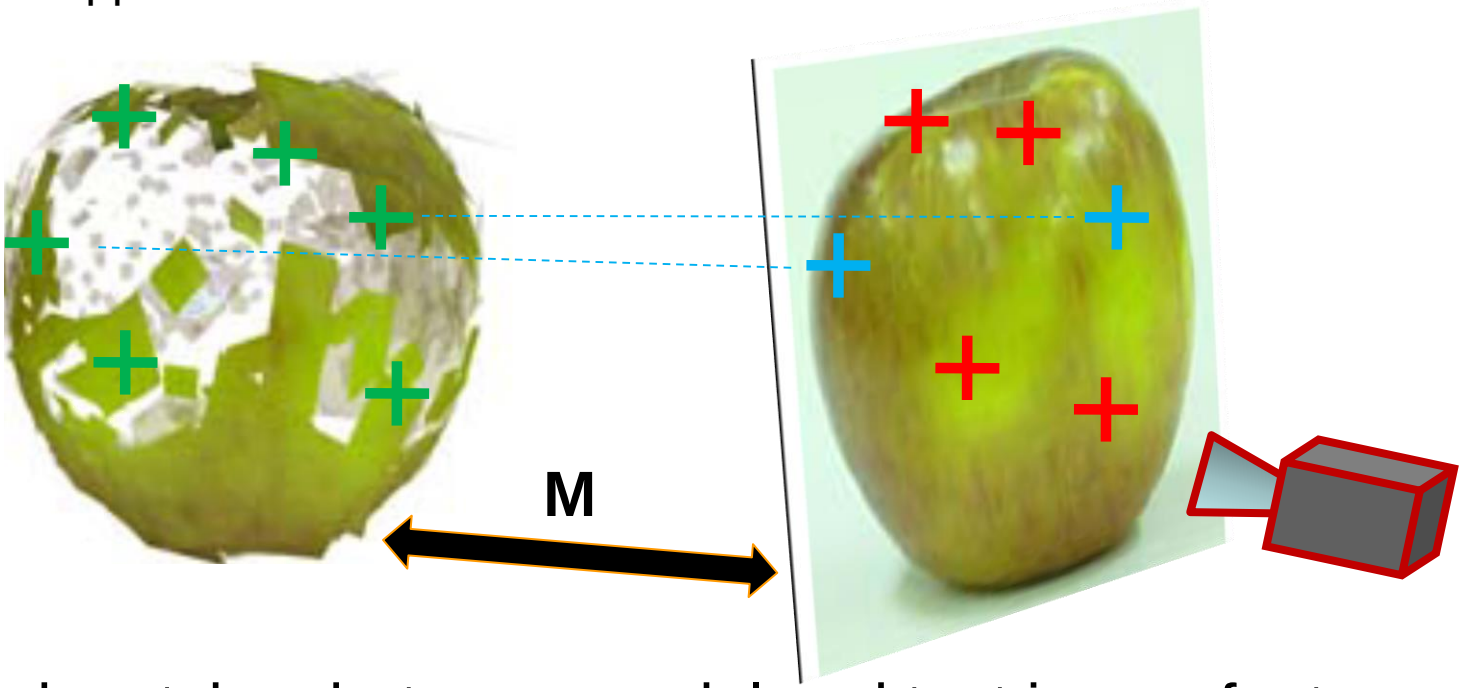
Class: apple



1. Find matches between model and test image features
2. Generate hypothesis:
   - Compute transformation M from N matches (N=2; affine camera; affine key points)

Recognition

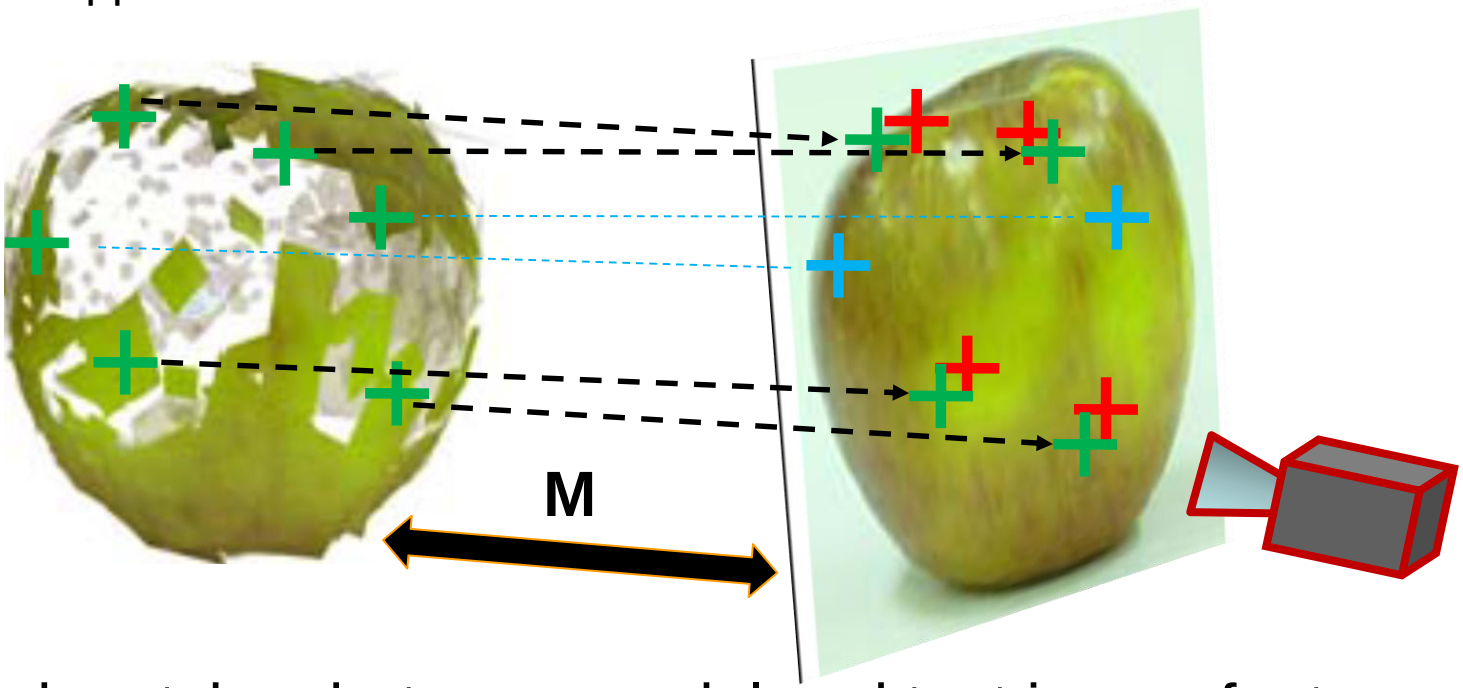
Class: apple



1. Find matches between model and test image features
2. Generate hypothesis:
   - Compute transformation M from N matches (N=2; affine camera; affine key points)
   - Generate hypothesis of object location and pose w.r.t. camera

Recognition

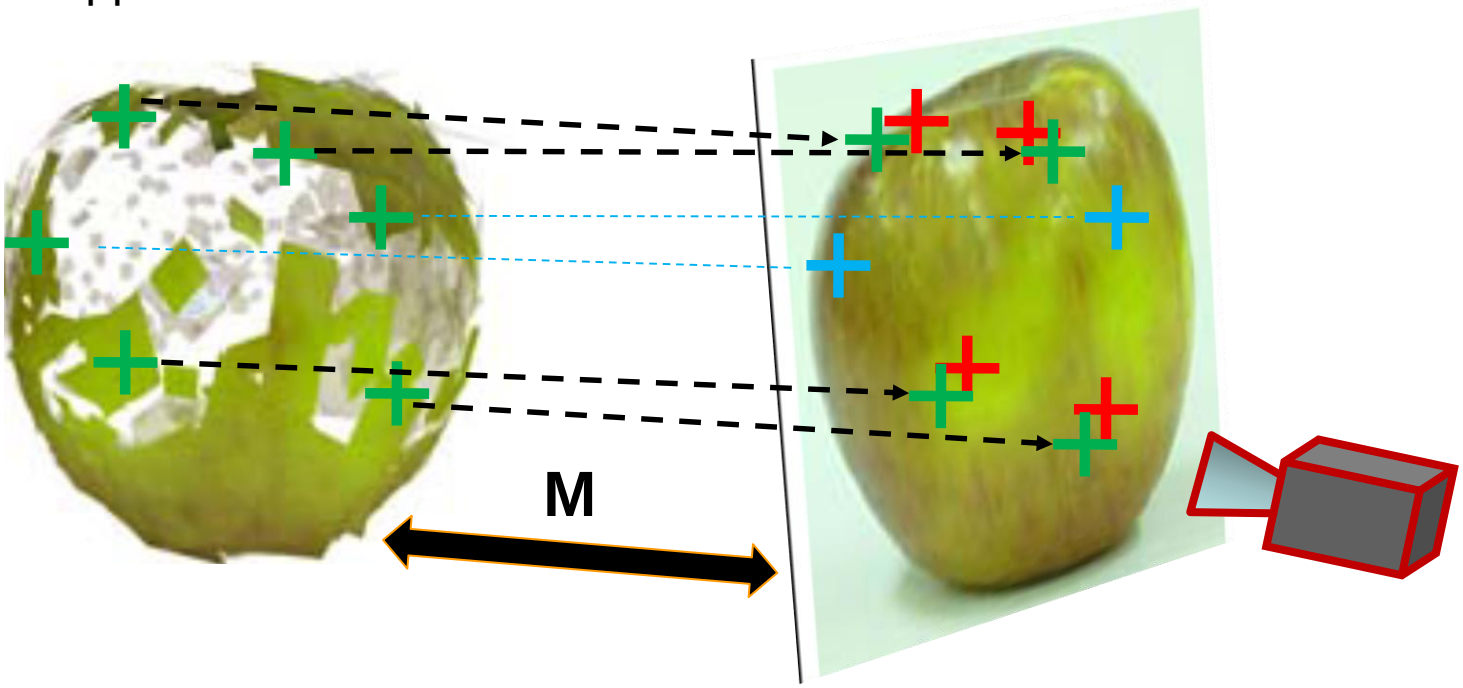
Class: apple



1. Find matches between model and test image features
2. Generate hypothesis:
   - Compute transformation M from N matches (N=2; affine camera; affine key points)
   - Generate hypothesis of object location and pose w.r.t. camera
3. Model verification
   - Use M to project other matched 3D model features into test image
   - Compute residual = D(projections, measurements)

Recognition

Class: apple



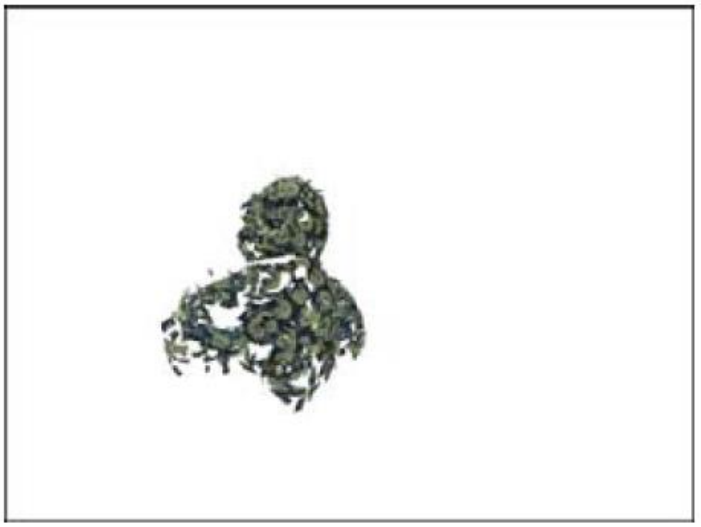
4. Repeat steps 2 and 3 until residual doesn't decrease anymore
5. Repeat steps 1-4 for different object instance C (apple, terry bear, etc…)
6. M and C corresponding to min residual return the estimated object pose and object instance

Object to recognize

Initial matches based on appearance

Matches verified with geometrical constraints

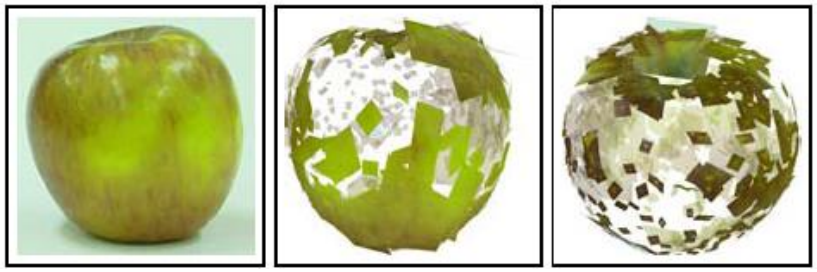
Recovered pose

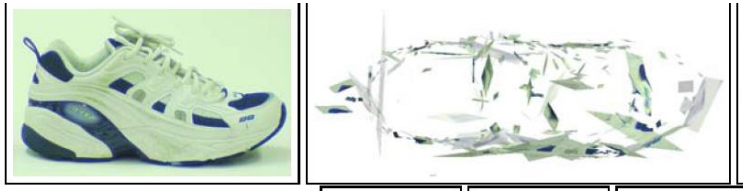


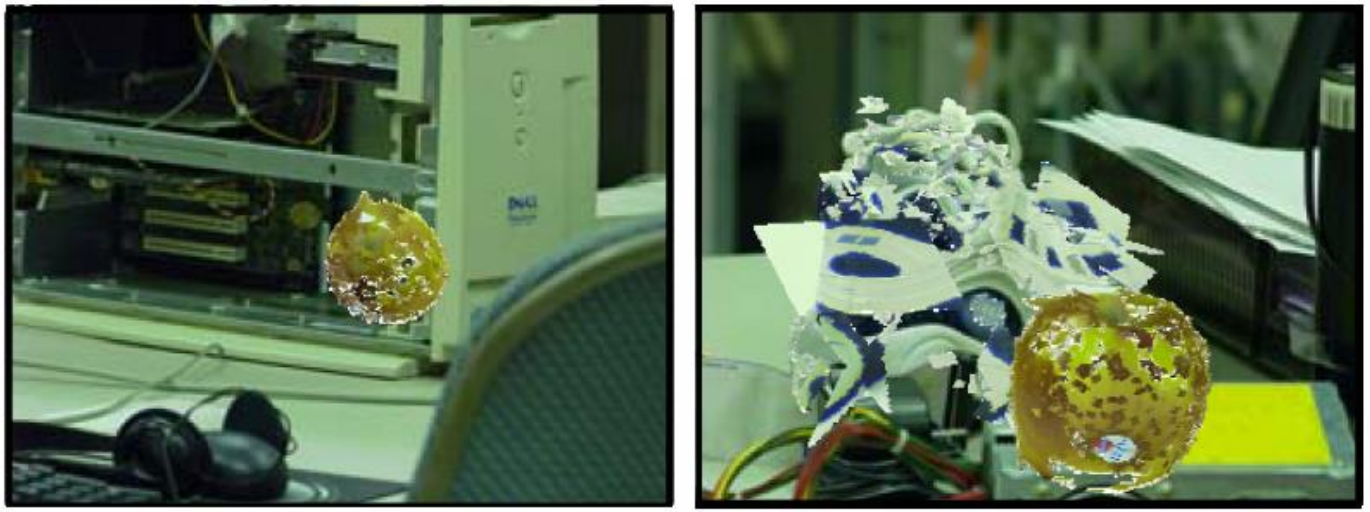
Courtesy of Rothganger et al

3D Object Recognition results

Rothganger et al. ‘03 ’06





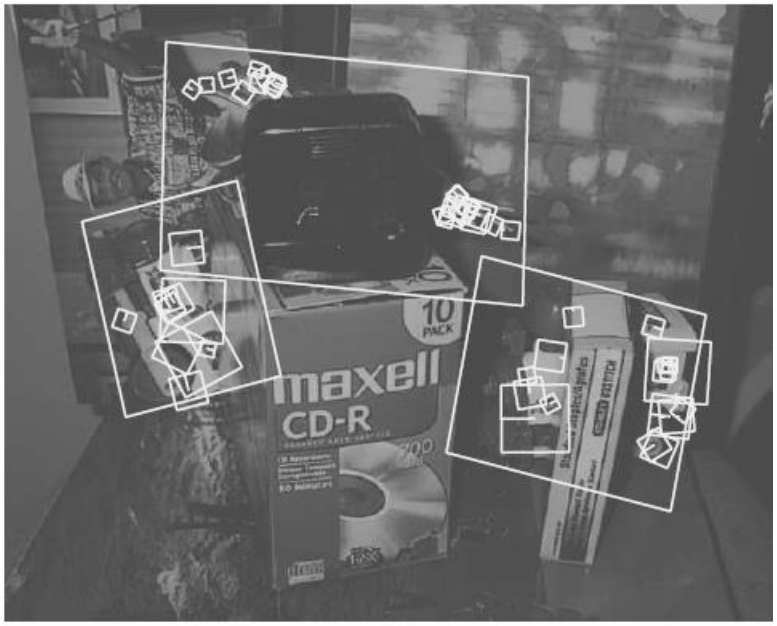


Courtesy of Rothganger et al

•Handle severe clutter

3D Object Recognition results

Lowe. ’99, ’04



Courtesy of D. Lowe

- Handle severe occlusions
- Fast!

3D Object Recognition results

[Ferrari et al '04]

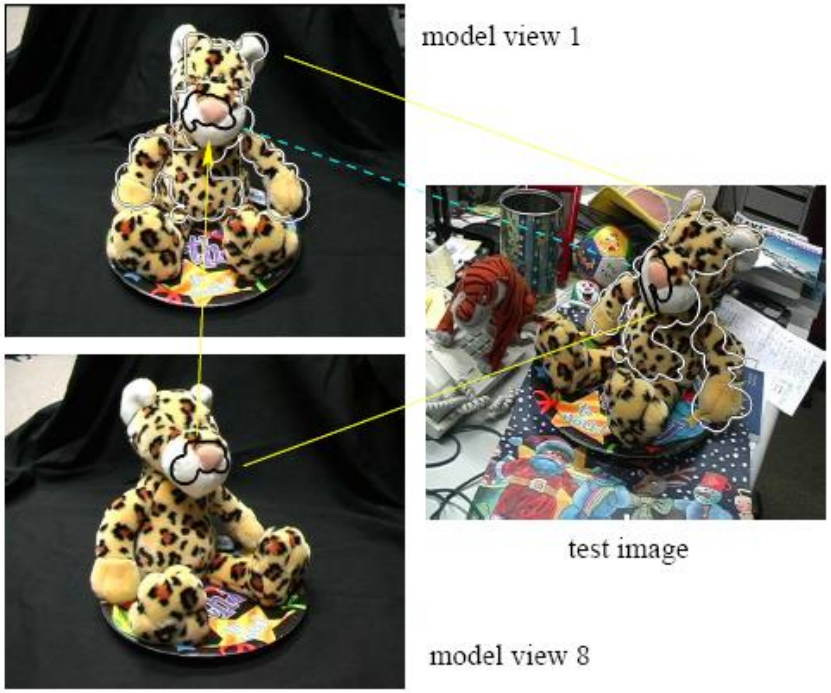


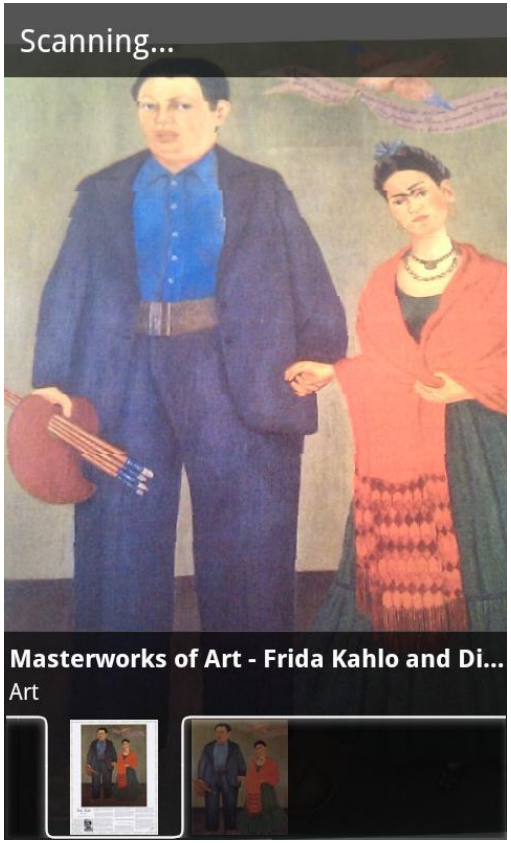
Figure 17: *Two compatible (and correct) GAMs. The nose GAM (black) is initially matched from model view 8, and is transferred to model view 1. Note how the other GAM (white) is very large and covers the head, arms and chest. A GAM can extend over multiple facets when the combination of viewpoints and surface orientations make the affine transformations of the region matches vary smoothly even across facet edges. In these cases, the resulting GAMs are larger and therefore more reliable and relevant.*

Courtesy of Ferrari et al

3D Object Recognition results

Edward Hsiao, Alvaro Collet and Martial Hebert. **Making specific features less discriminative to improve point-based 3D object recognition**. *IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, June, 2010.

Visual search and landmarks recognition



Google Goggles

Visual search and landmarks recognition

RICOH

A9

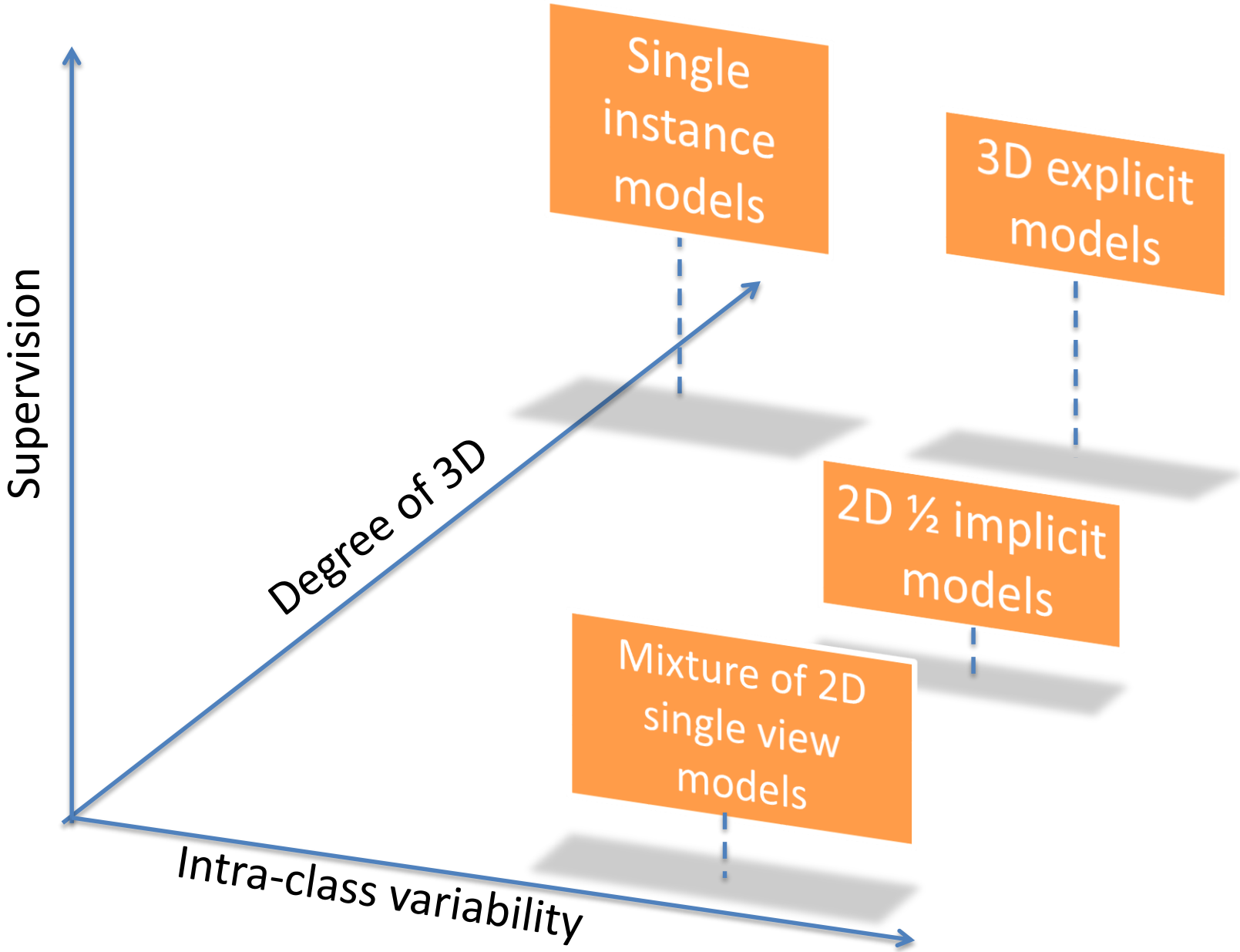
Limitations of single instance 3D object detectors

- Cannot handle intra-class variability.

Why?

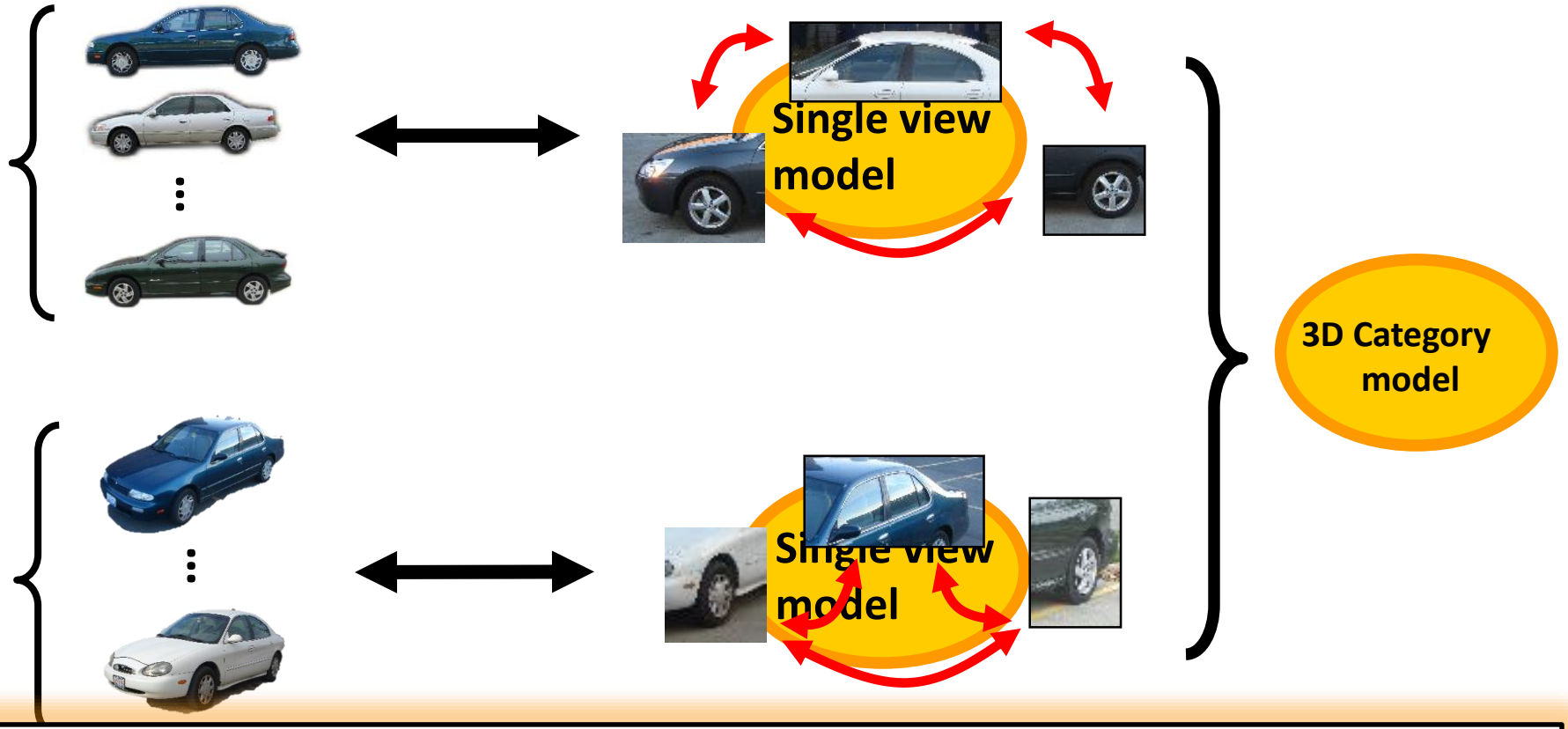
- Models capture fine-grained details of the object instance which are not shared across instances in the same class
- Hypothesis-generation and verification scheme is not designed to maximize discrimination power

Models for 3d Object detection



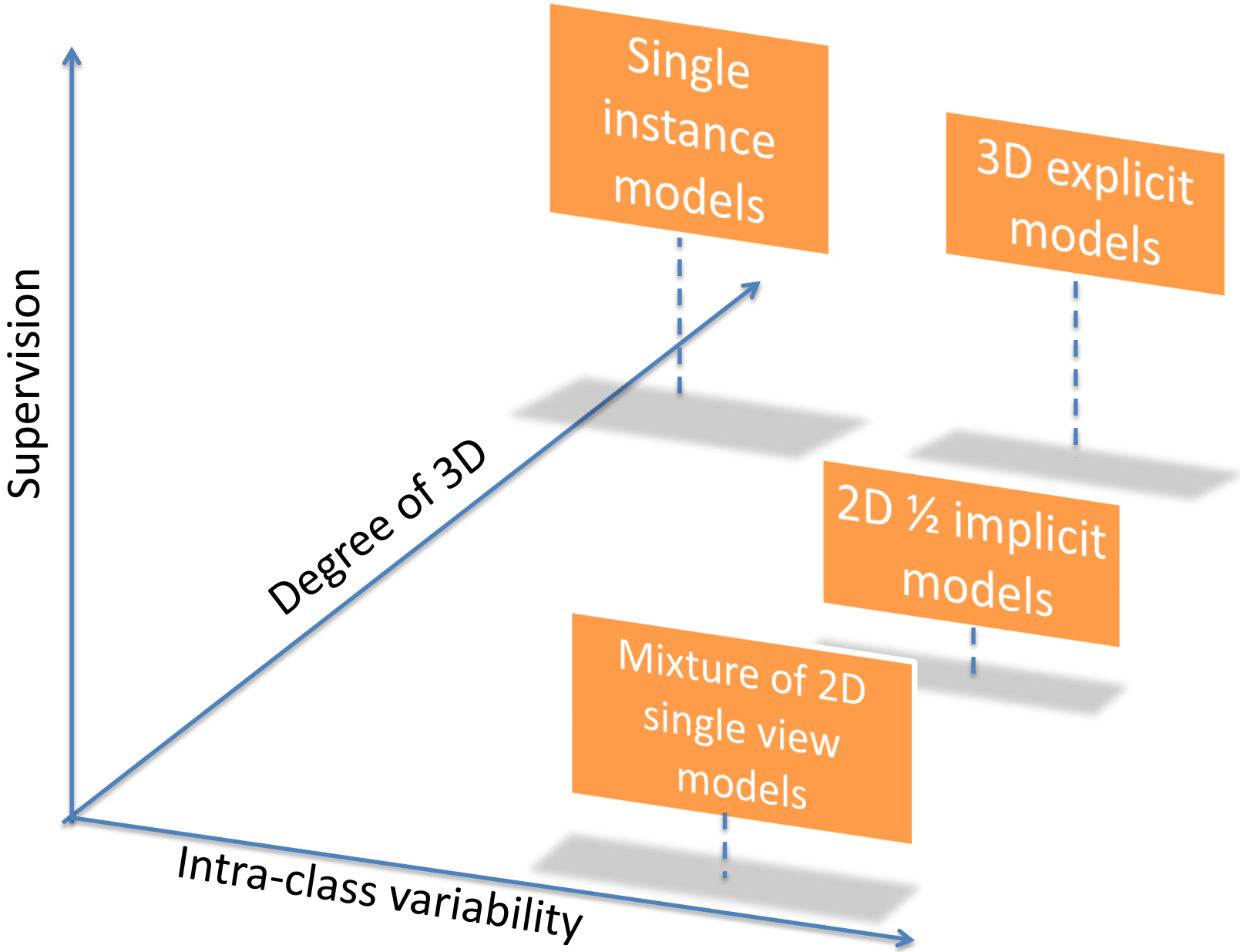
Mixture of 2D models

- Weber et al. '00
- Schneiderman et al. '01
- Ullman et al. 02
- Fergus et al. '03
- Torralba et al. '03
- Felzenszwalb & Huttenlocher '03
- Leibe et al. '04
- Shotton et al. '05
- Grauman et al. '05
- Savarese et al, '06
- Todorovic et al. '06
- Vedaldi & Soatto '08
- Zhu et al 08
- Gu & Ren, '10



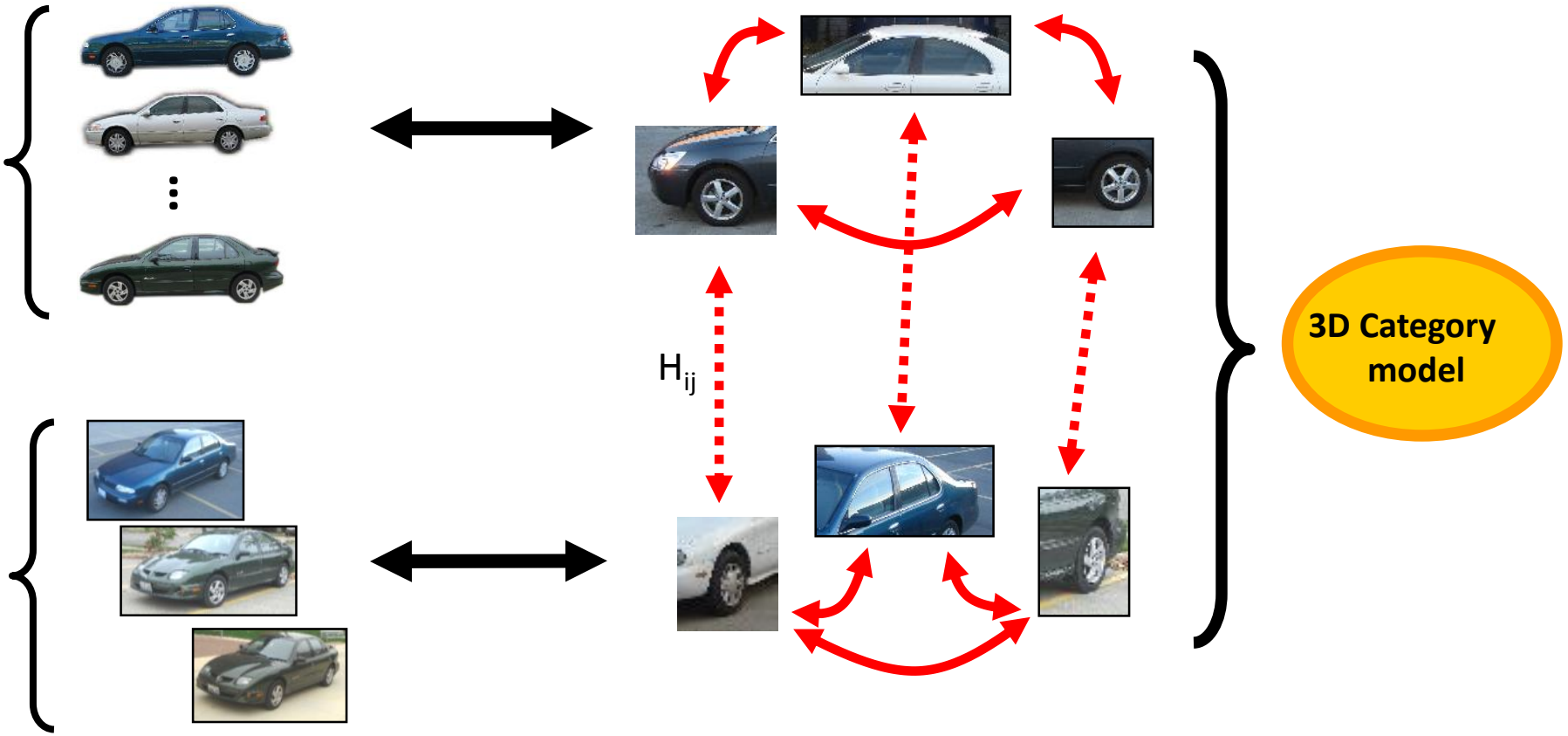
CONS: Single view models are independent • Non scalable to large number of categories/view-points • Just b. boxes • Cannot estimate 3D pose or 3D layout

Models for 3d Object detection



2D ½ implicit models

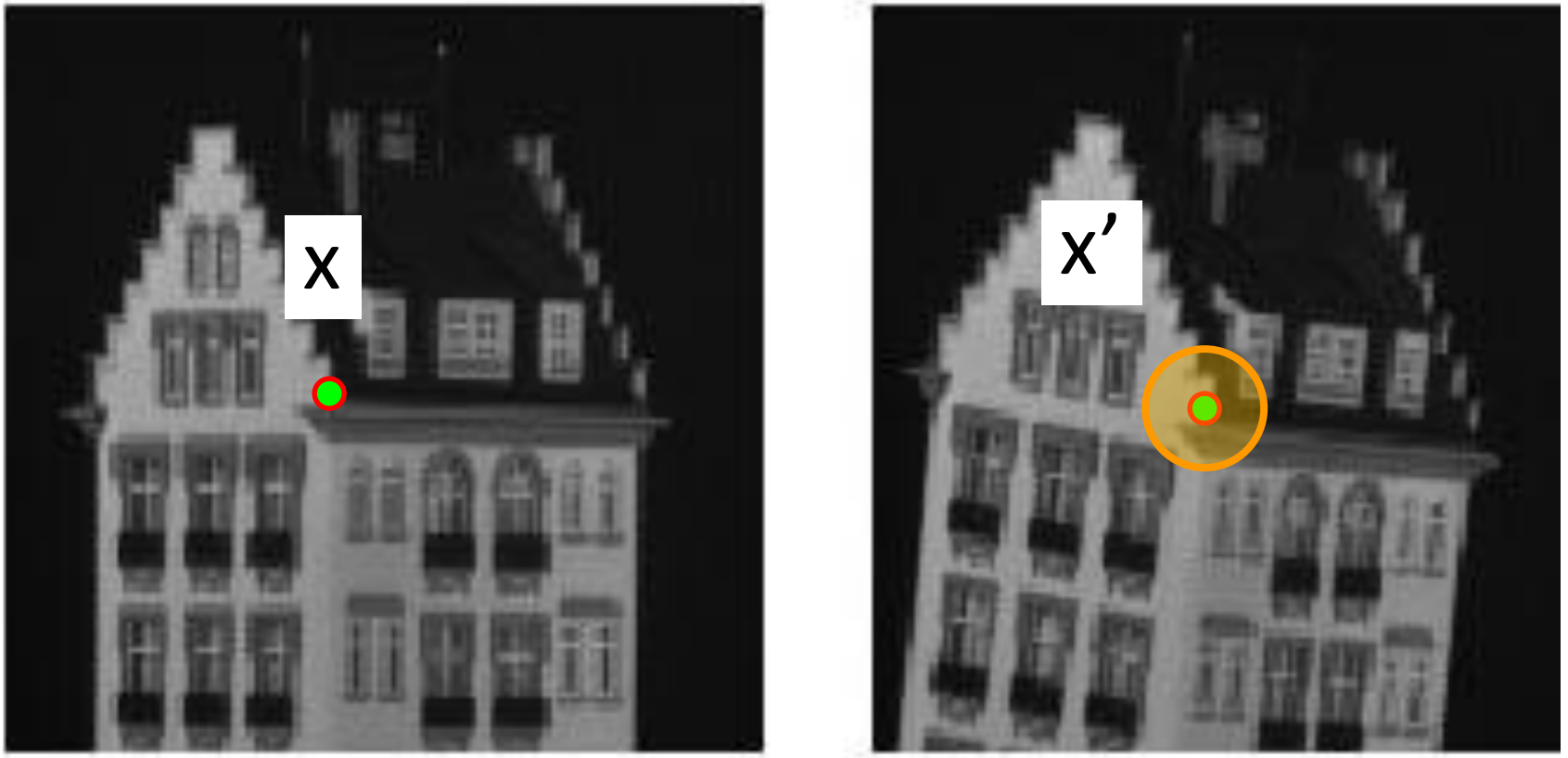
- Savarese & Fei-Fei, ICCV 07
- Savarese & Fei-Fei, ICCV 07
- Su, Sun, Fei-Fei, Savarese., CVPR 2009
- Sun, Su, Fei-Fei, Savarese, ICCV 2009
- Thomas et al. '06-09
- Kushal, et al., '07
- Farhadi '09
- Zhu et al. '09
- Ozuysal et al. '10
- Stark et al.'10
- Payet & Todorovic, 11
- Glasner et al., '11



- Parts relationship are probabilistic and learnt automatically

Linking features or parts across views:

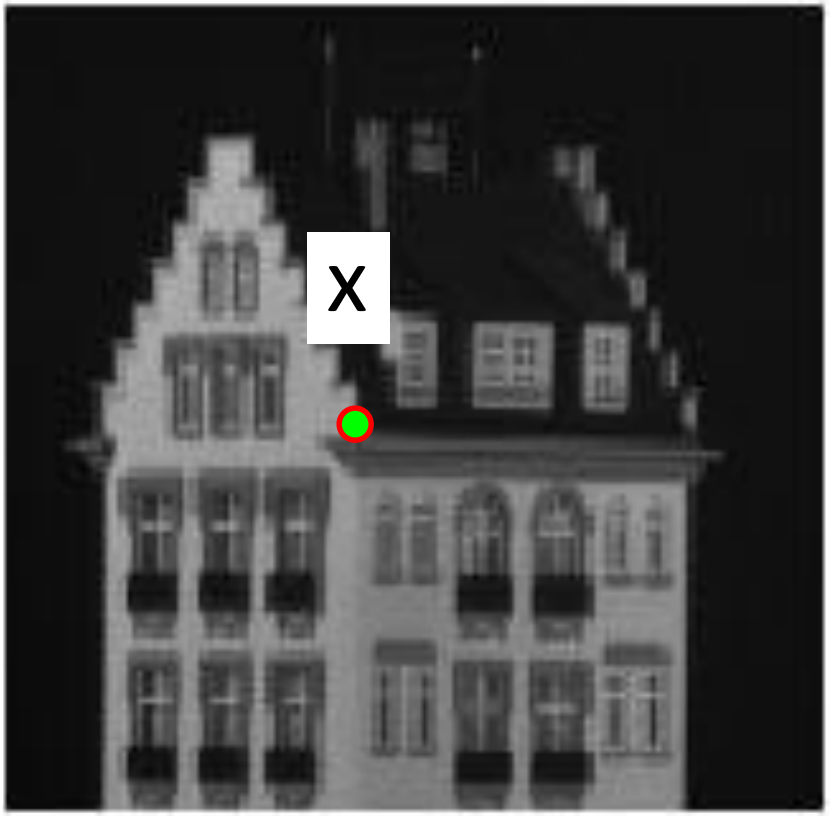
Perspective or affine transformation constraints

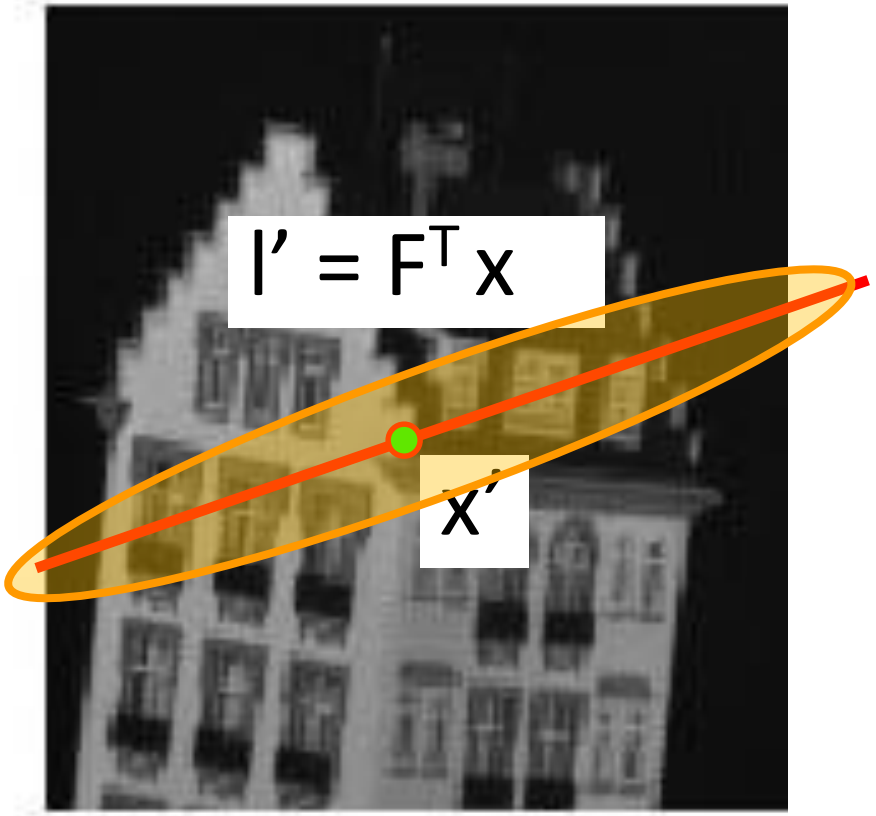


$$x' = H\,x$$

Linking features or parts across views:

Epipolar Transformation Constraints



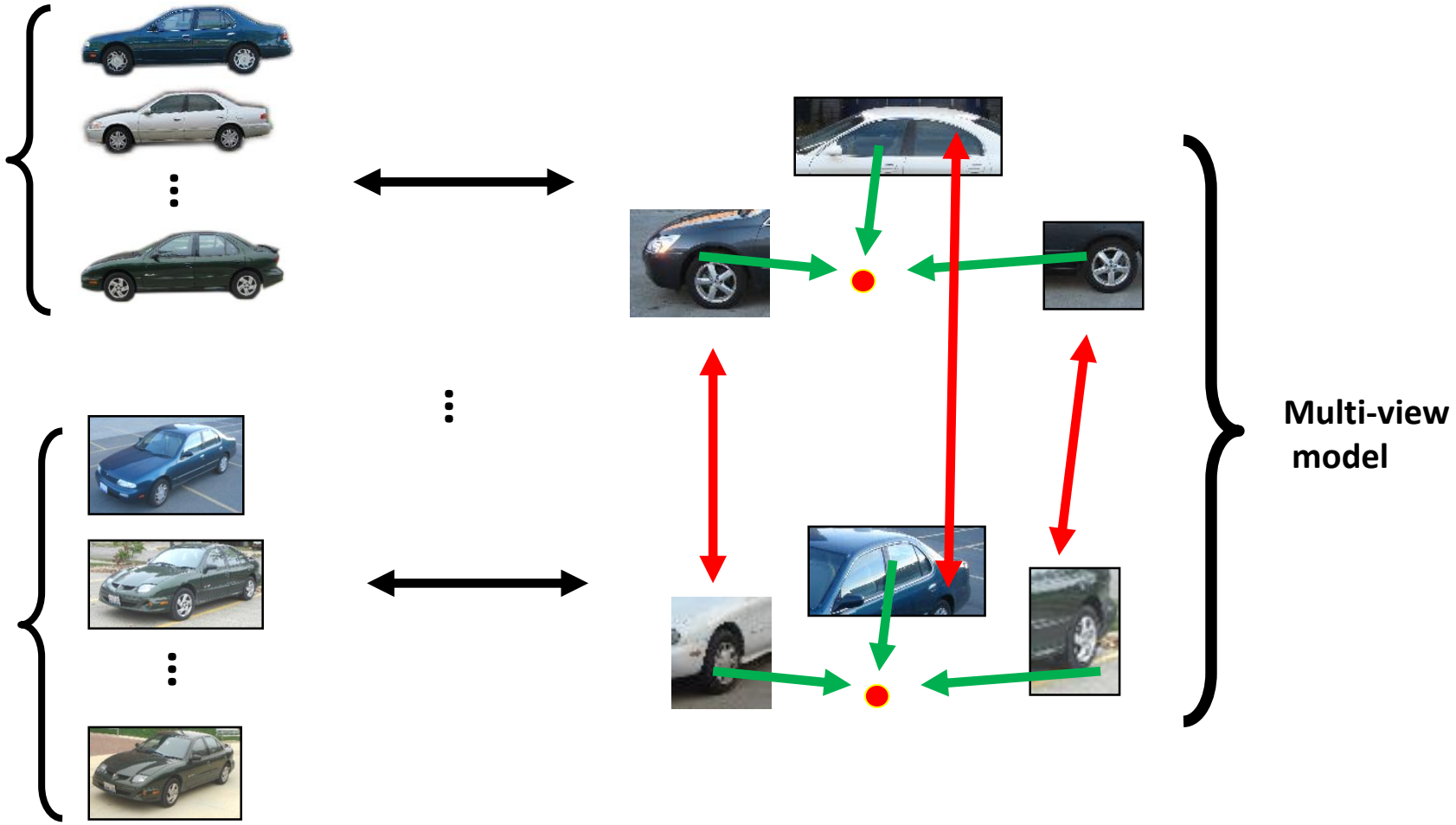


$$l' = F^T x$$

$$x' \in l'$$

Implicit 3D models by ISM representations

- Thomas et al. ’06
- Leibe et al. ‘04

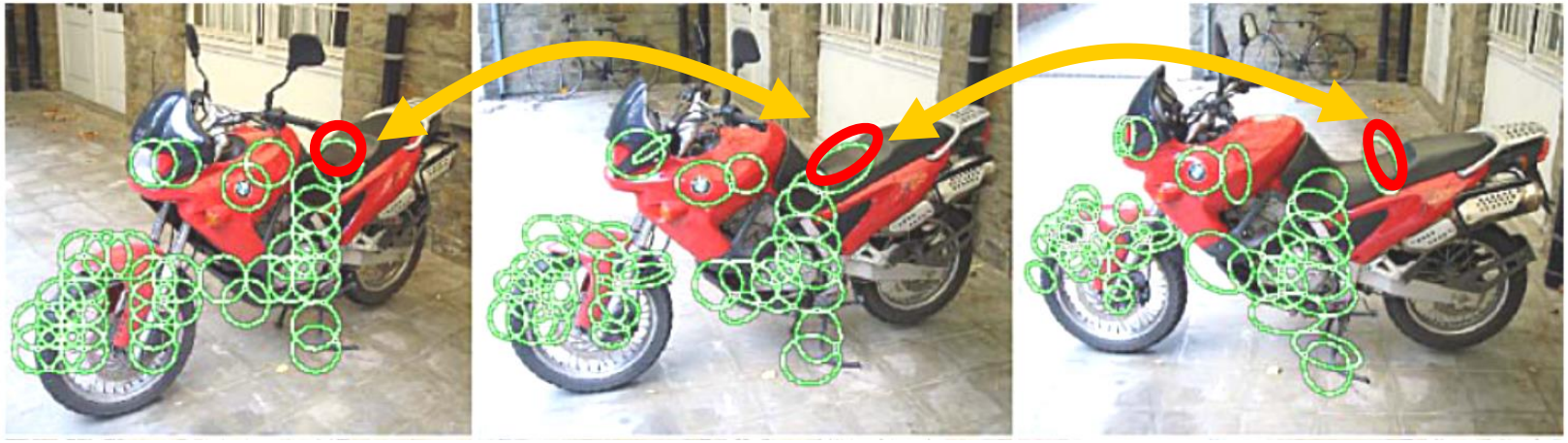


Sparse set of interest points or parts of the objects are linked across views.

Implicit 3D models by ISM representations

Region tracks

[Ferrari et al. ’04, ‘06]

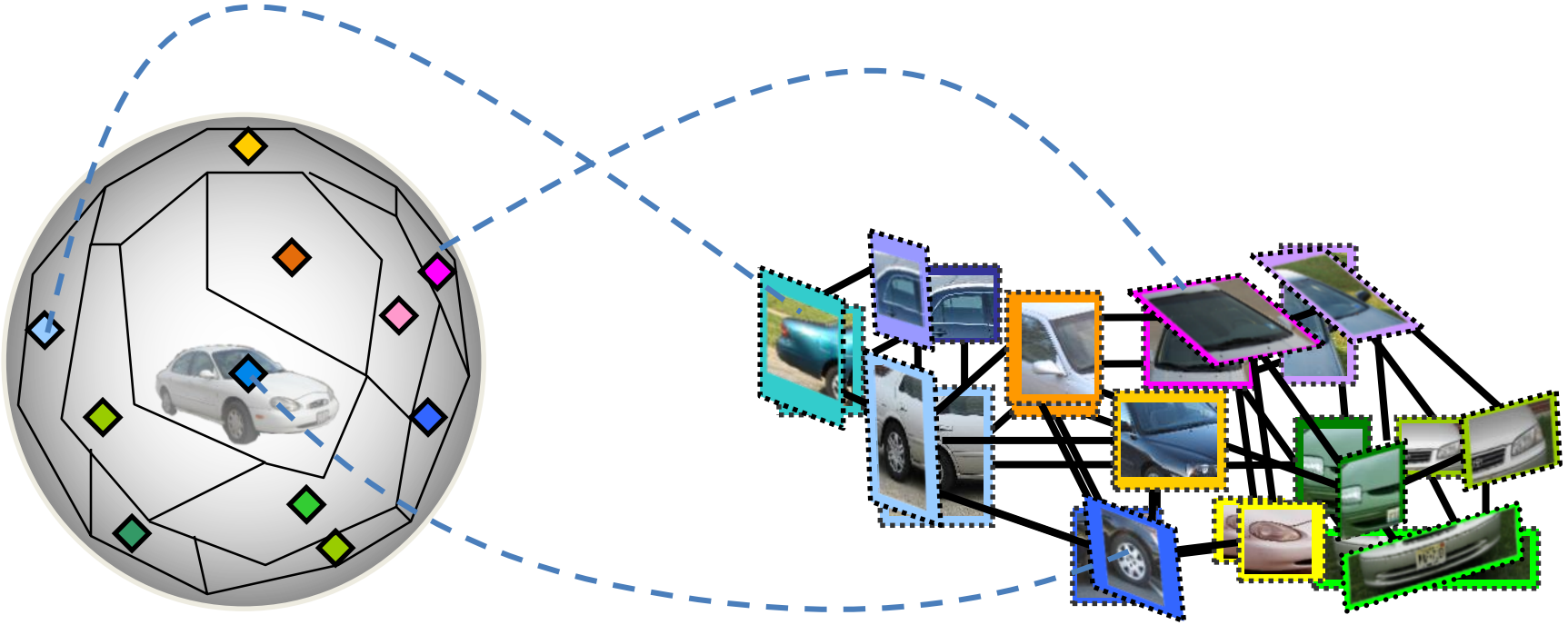


Courtesy of Thomas et al. 06

Set of *region-tracks* connecting model views
Each track is composed of image regions of a single physical surface patch along the model views in which it is visible.

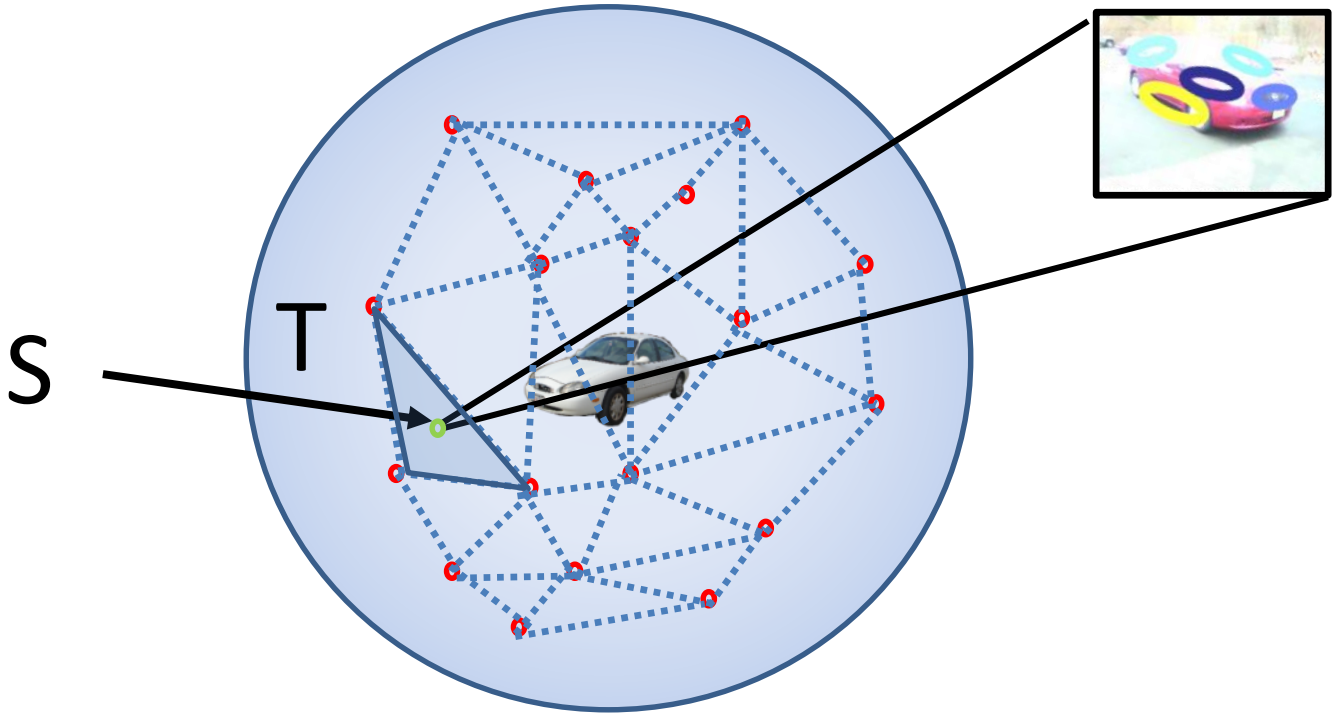
Implicit 3D models by graph-based representations

Savarese, Fei-Fei, ICCV 07
Sun, et al, CVPR 2009, ICCV 09



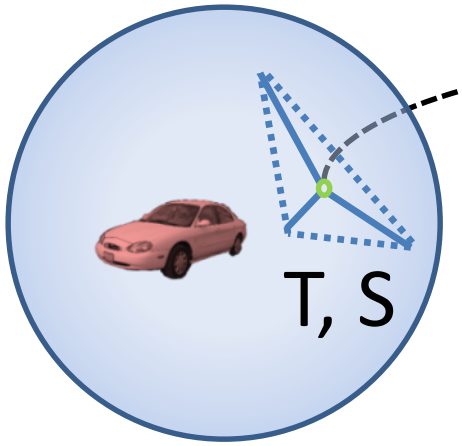
- Canonical parts captures view invariant diagnostic appearance information
- 2d ½ structure linking parts via weak geometry
- Parts and relationship are modeled in a probabilistic fashion
  - Parameters are learnt so as to maximize detection accuracy

Parameterization on view-sphere



- Model the object as collection of parts for any T and S on the viewing sphere

Multi-view generative part-based model



α = Part Prop. Prior

$$\pi \sim Dir(\alpha)$$

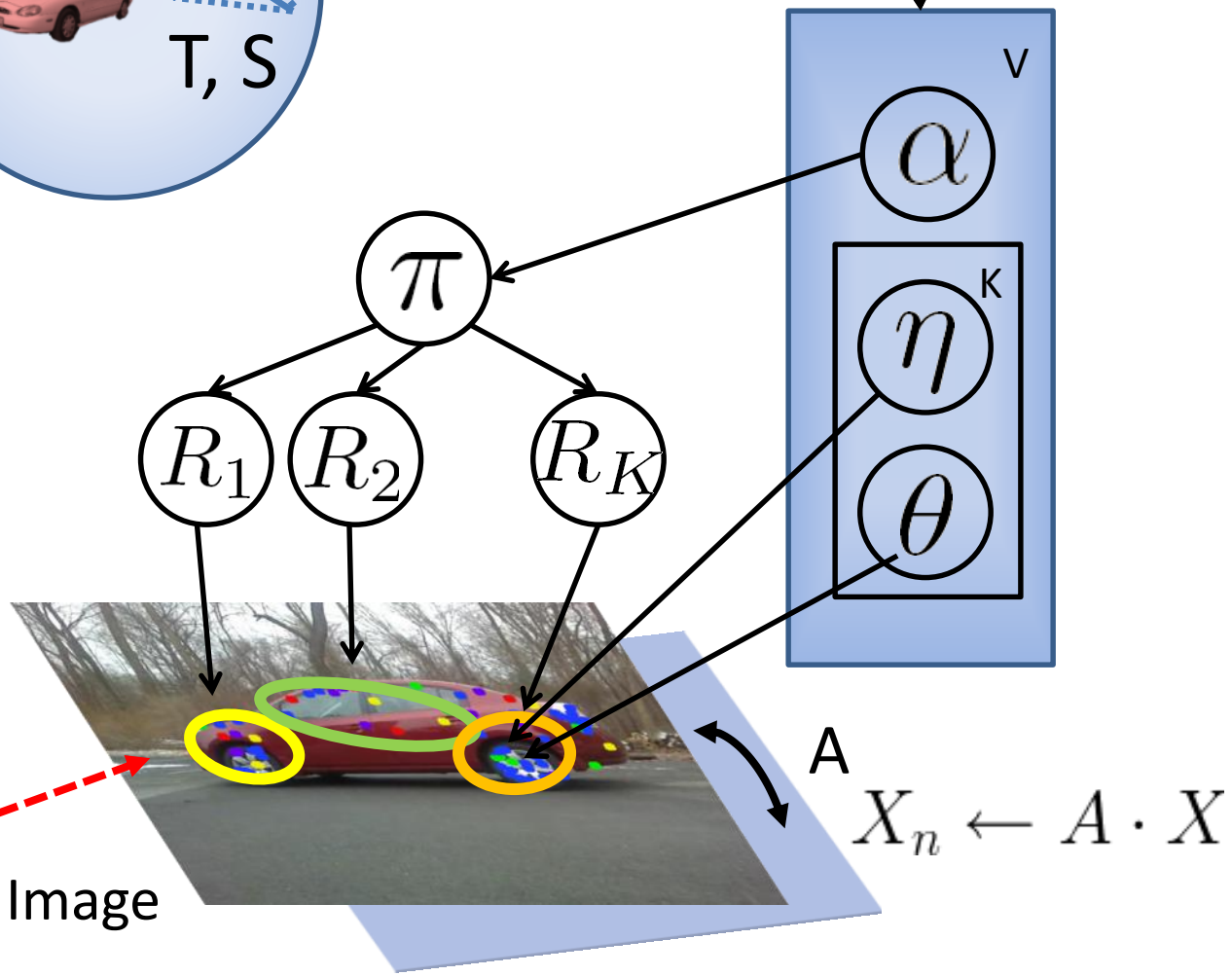
$$R \sim Mult(\pi)$$

$$Y_n \sim Mult(\eta)$$

η = Part Appearance

$$X_n \sim N(theta)$$

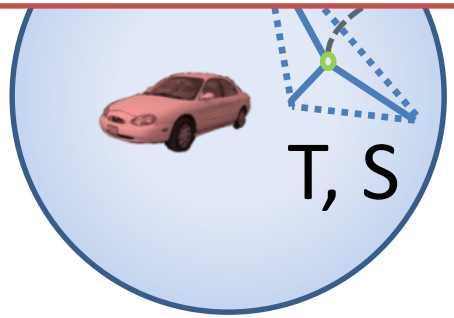
θ = Part Location/shape



Yn=Codeword

Xn=Location

- Learning: estimate the latent variables and relevant parameters, given the observations
- Variational EM can be used Blei, ICML 2004.



α = Part Prop. Prior

$$\pi \sim Dir(\alpha)$$

$$R \sim Mult(\pi)$$

$$Y_n \sim Mult(\eta)$$

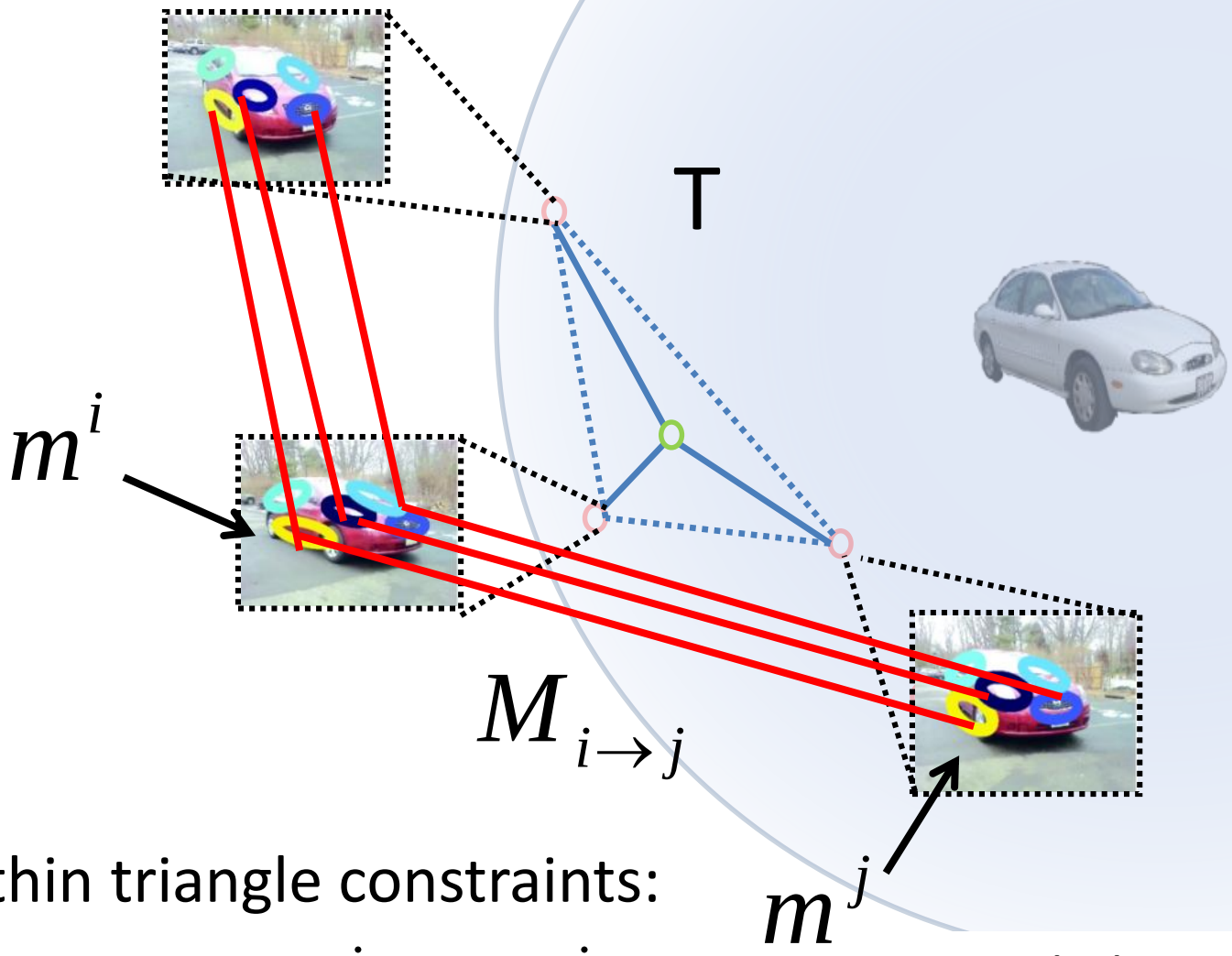
η = Part Appearance

$$X_n \sim N(theta)$$

θ = Part Location/shape

Yn=Codeword
Xn=Location

Incorporating geometrical constraints

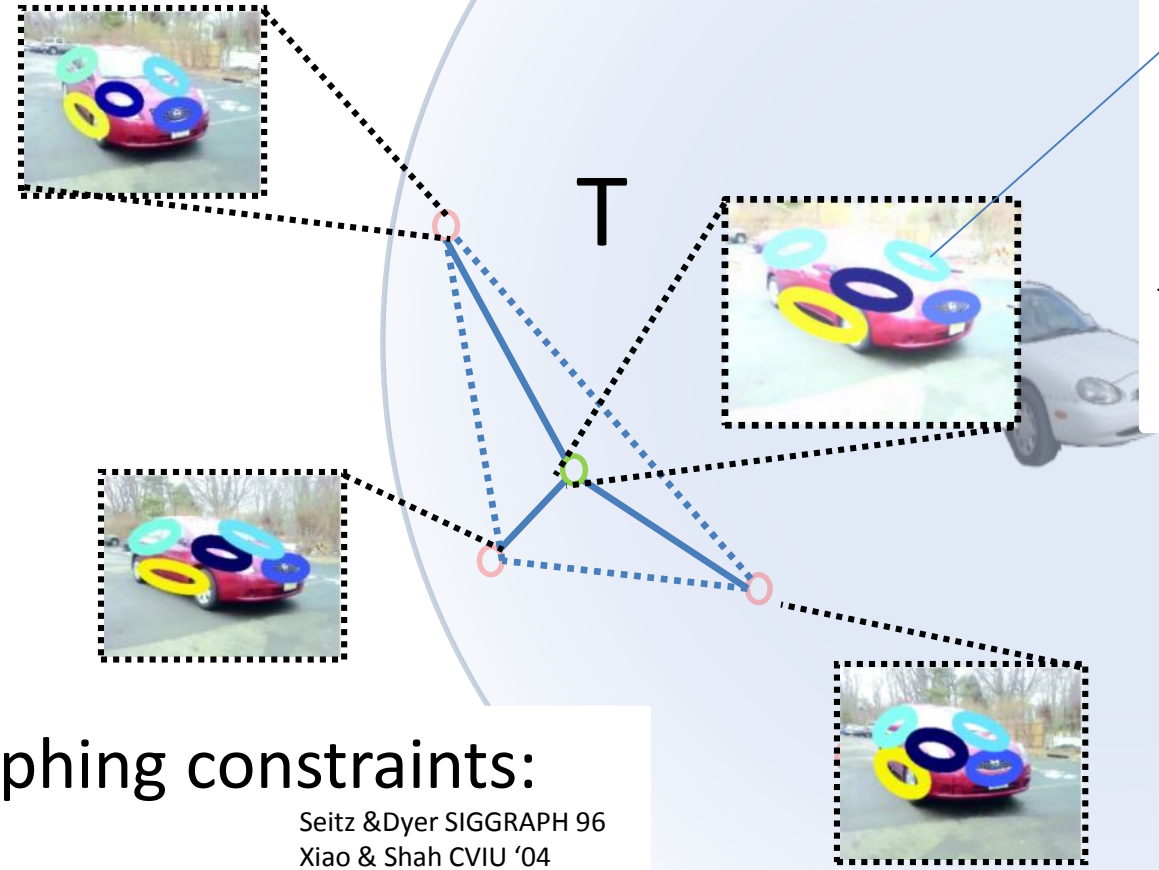


Within triangle constraints:

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

Encoded as a penalty term in variational *EM*

Incorporating geometrical constraints



$m(S)$ = Center

$W(S)$ = Shape

$$\begin{cases} \Sigma = WW^T \\ \theta = (m, \Sigma) \end{cases}$$

View morphing constraints:

Seitz &Dyer SIGGRAPH 96
Xiao & Shah CVIU '04

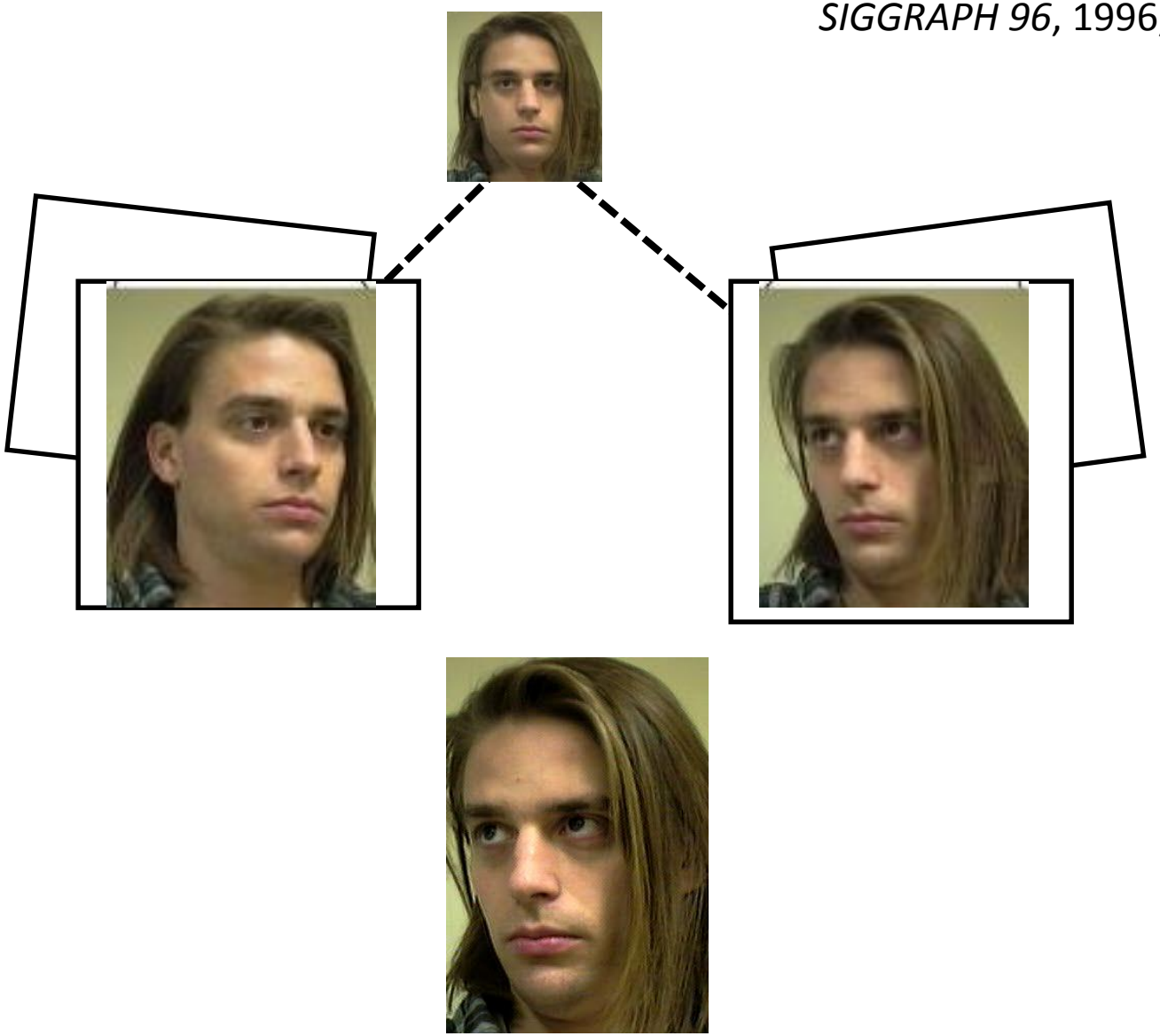
$$m(S) = \sum_{g=1}^{3} m_T^g \cdot s_g$$

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

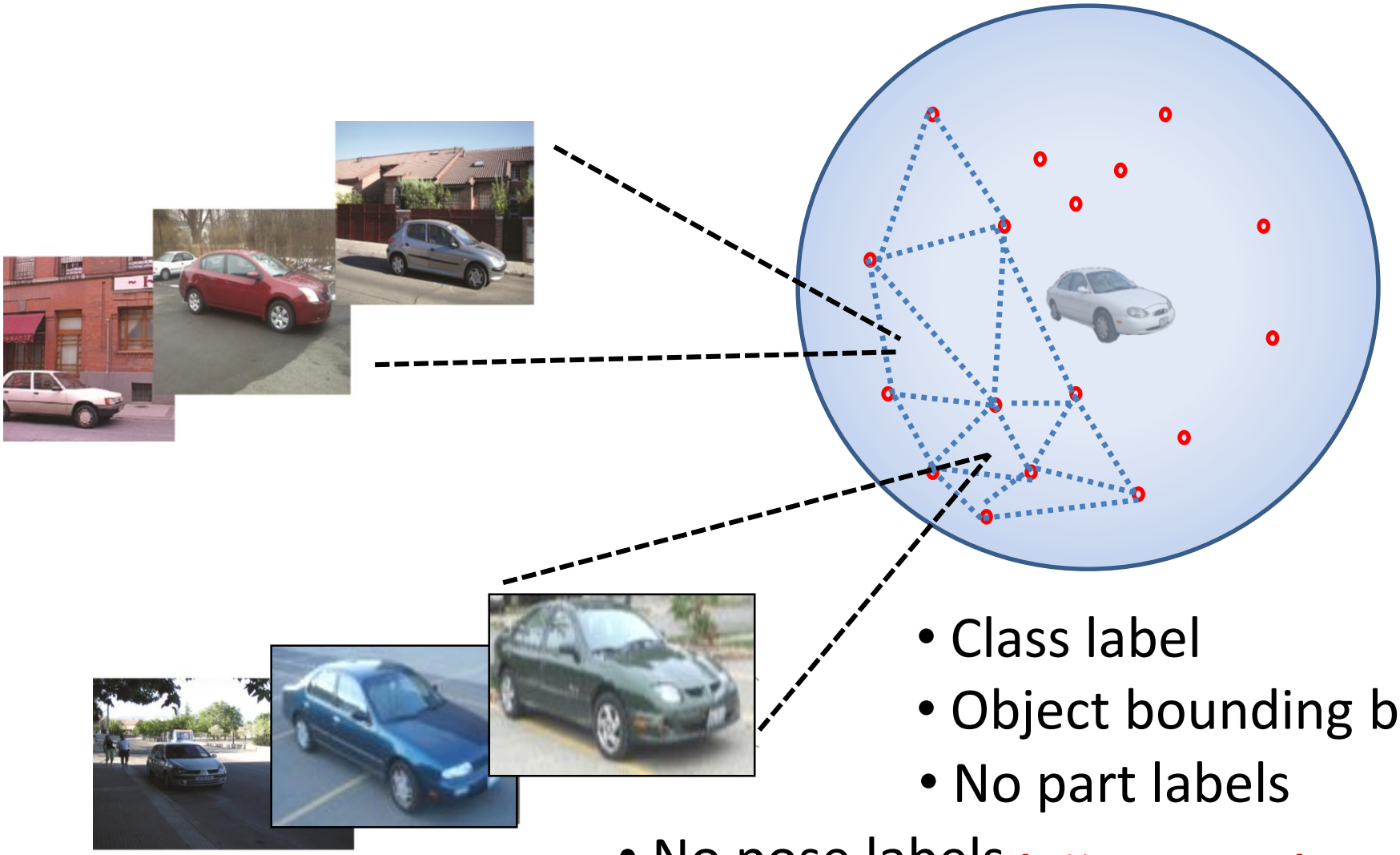
Encoded as a penalty term in variational *EM*

Incorporating geometrical constraints

S. M. Seitz and C. R. Dyer, *Proc. SIGGRAPH 96*, 1996, 21-30



Semi-supervised

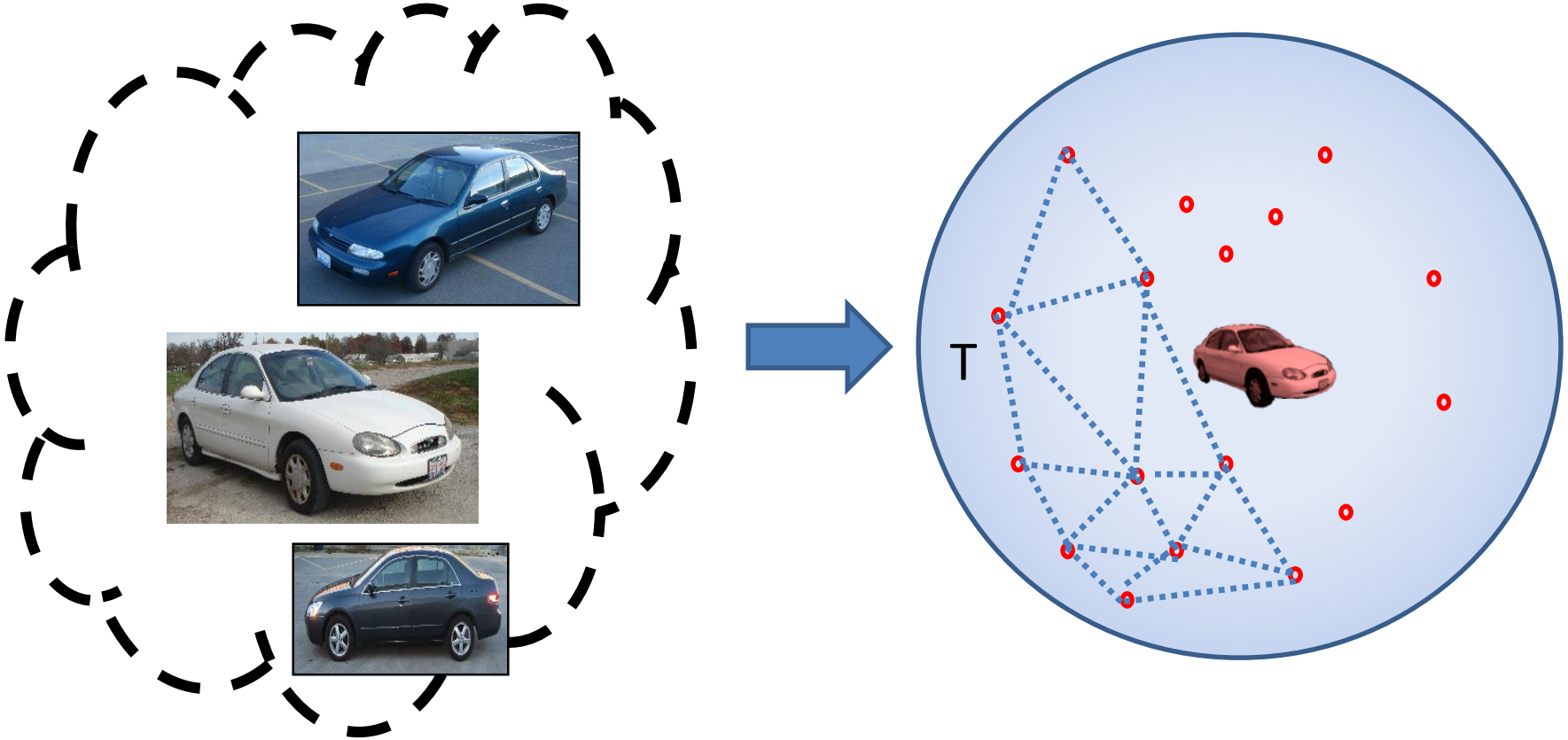


- Class label
- Object bounding box
- No part labels

- No pose labels [unlike Sun CVPR 09]

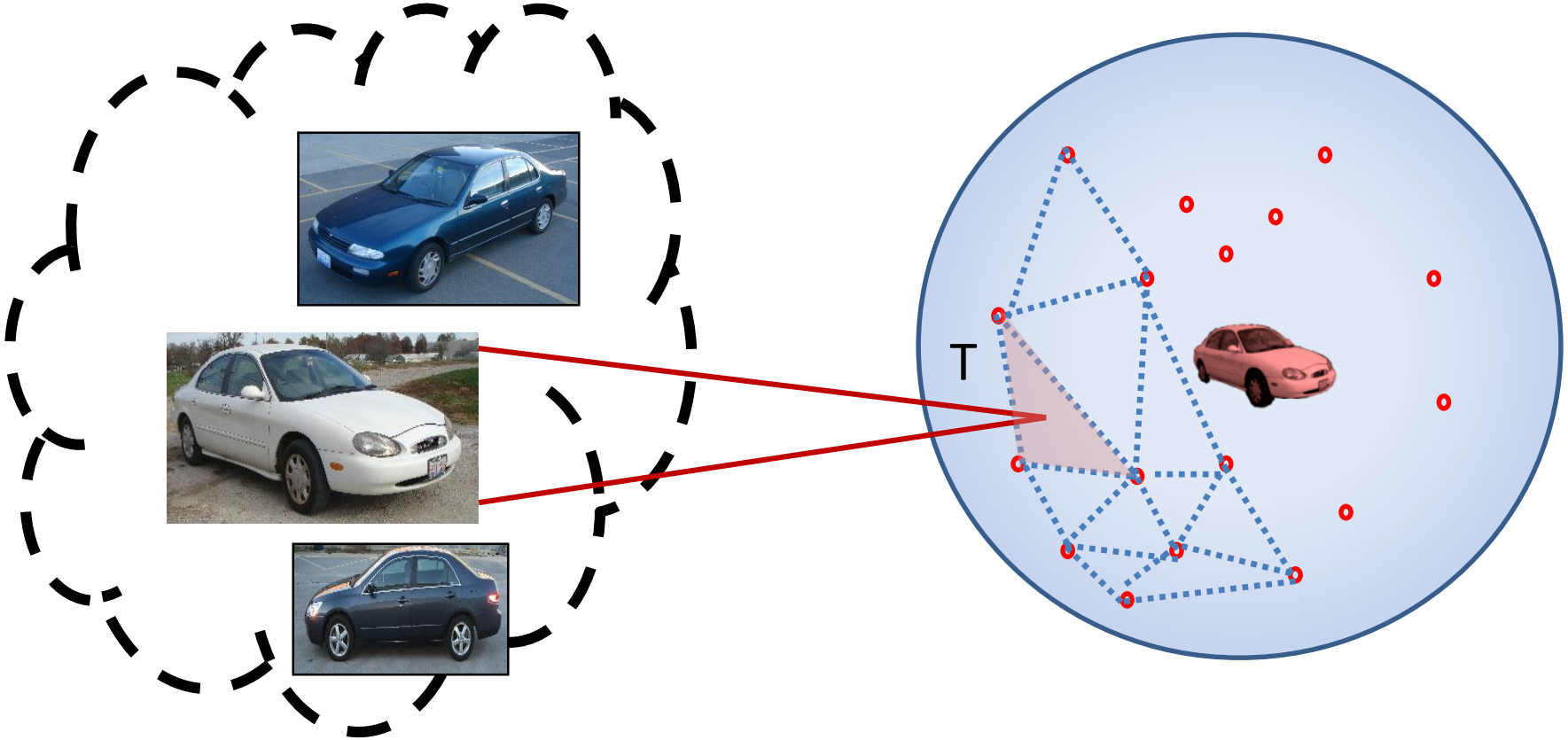
- No need to observe same object instance from multiple views
[unlike Savarese & Fei-Fei, 07, 08]

Incremental learning



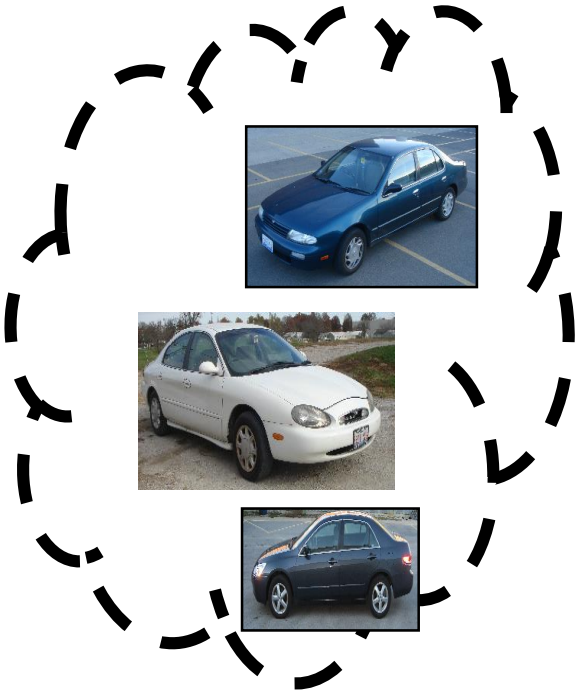
- Enable unorganized and on-line collection training images
- Increase efficiency in learning (no need large storage space)

Incremental learning

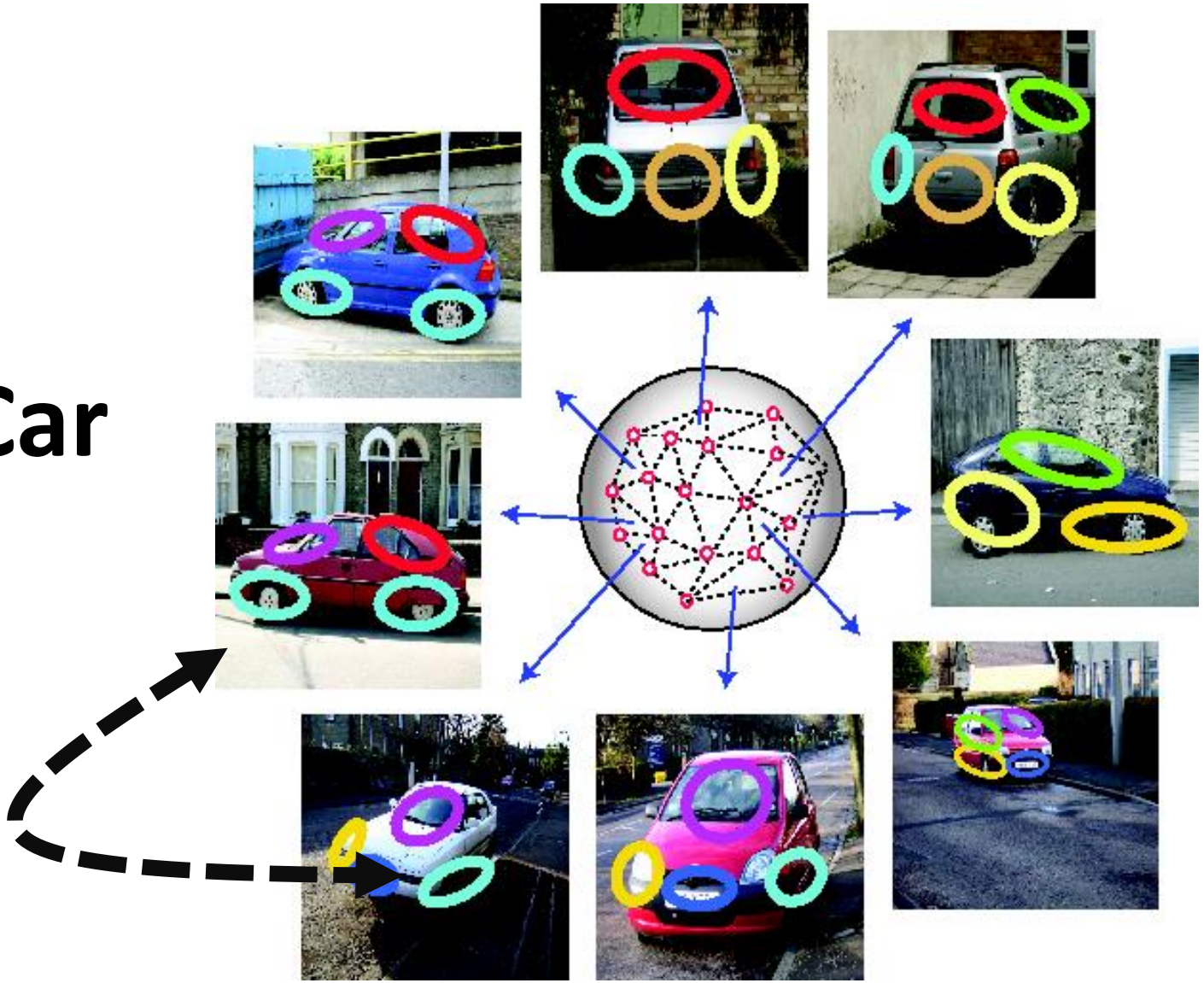


- Assign new training image to a triangle of the view sphere
- Evidence of training image is used to update model parameters
- Re-estimate sufficient statistics in a iterative fashion

Evolution of learnt parts

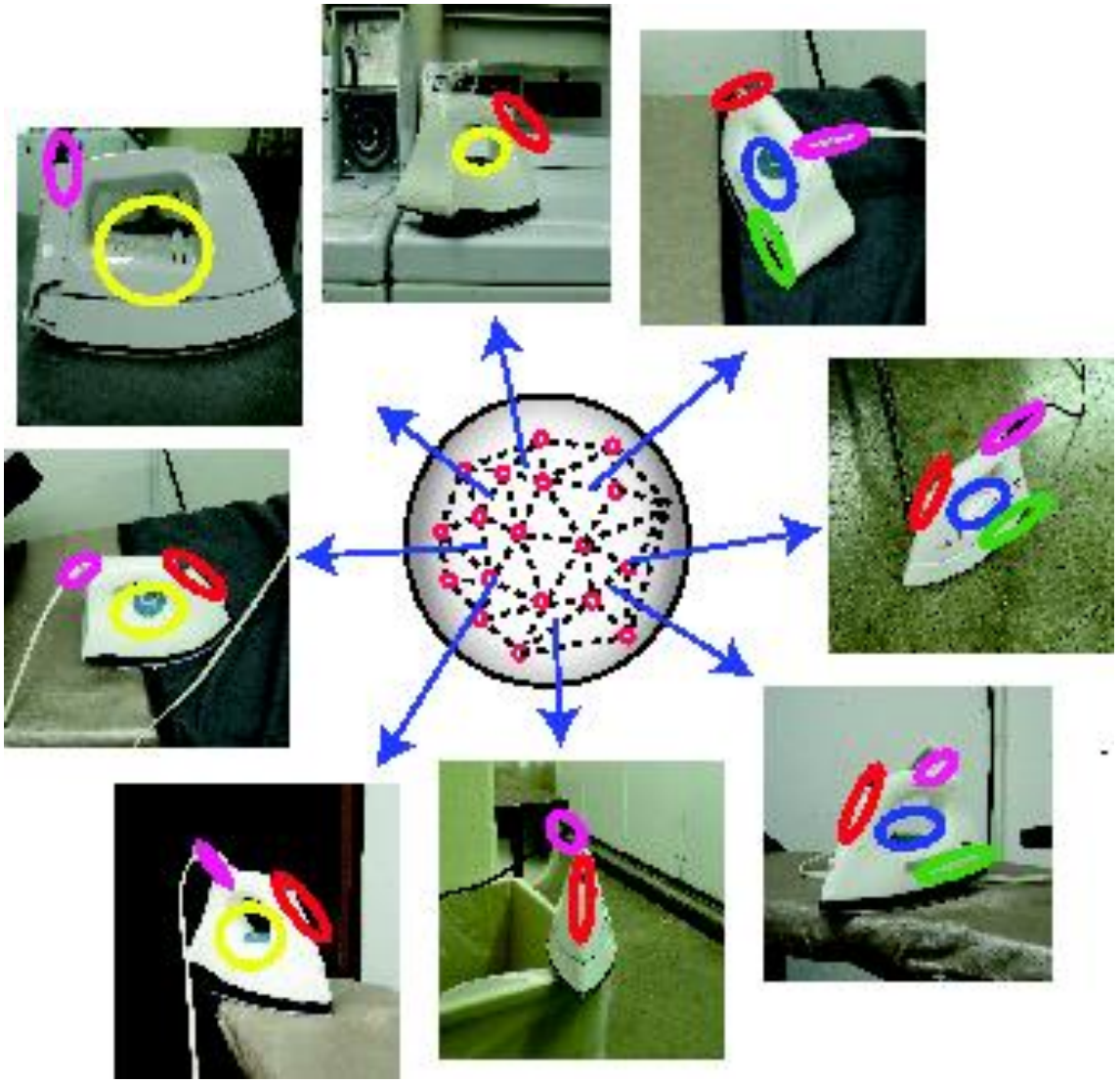


Examples of learnt part-based models



Examples of learnt part-based models

Travel iron

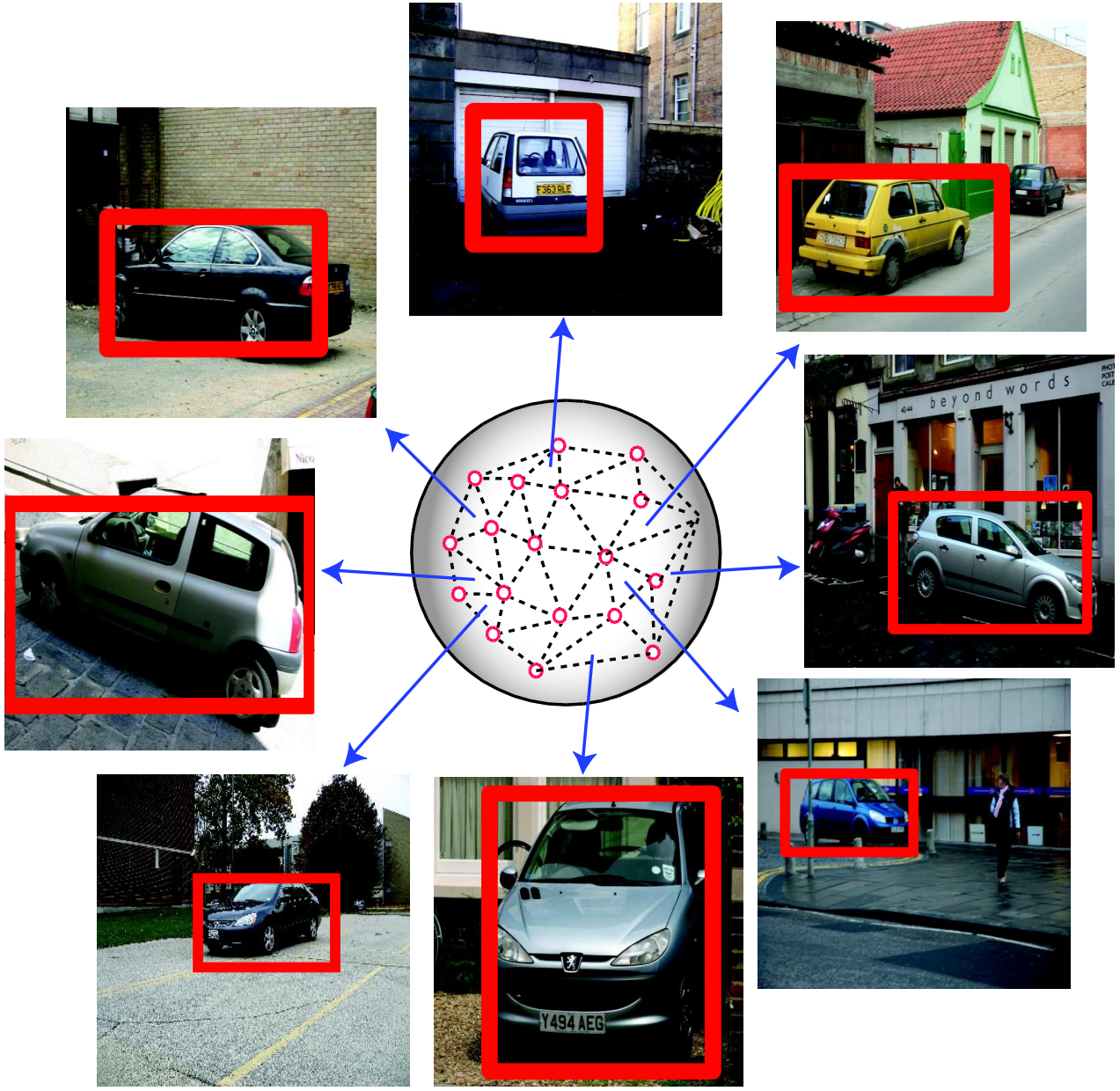


Experimental results

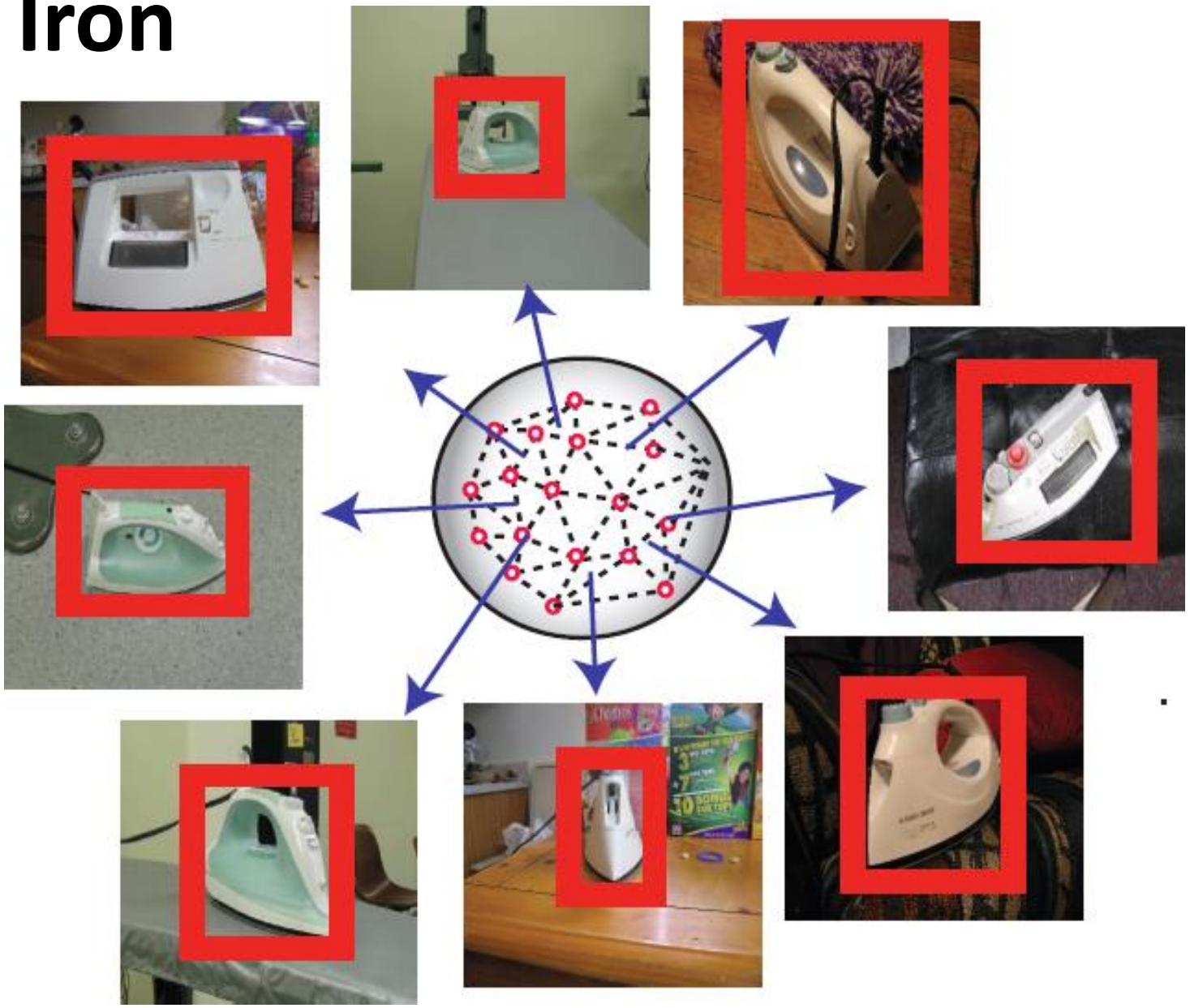
- Object detection from any viewing angles
- Accurate estimation of the object pose

- PASCAL 2006 dataset
- 3D Object Dataset

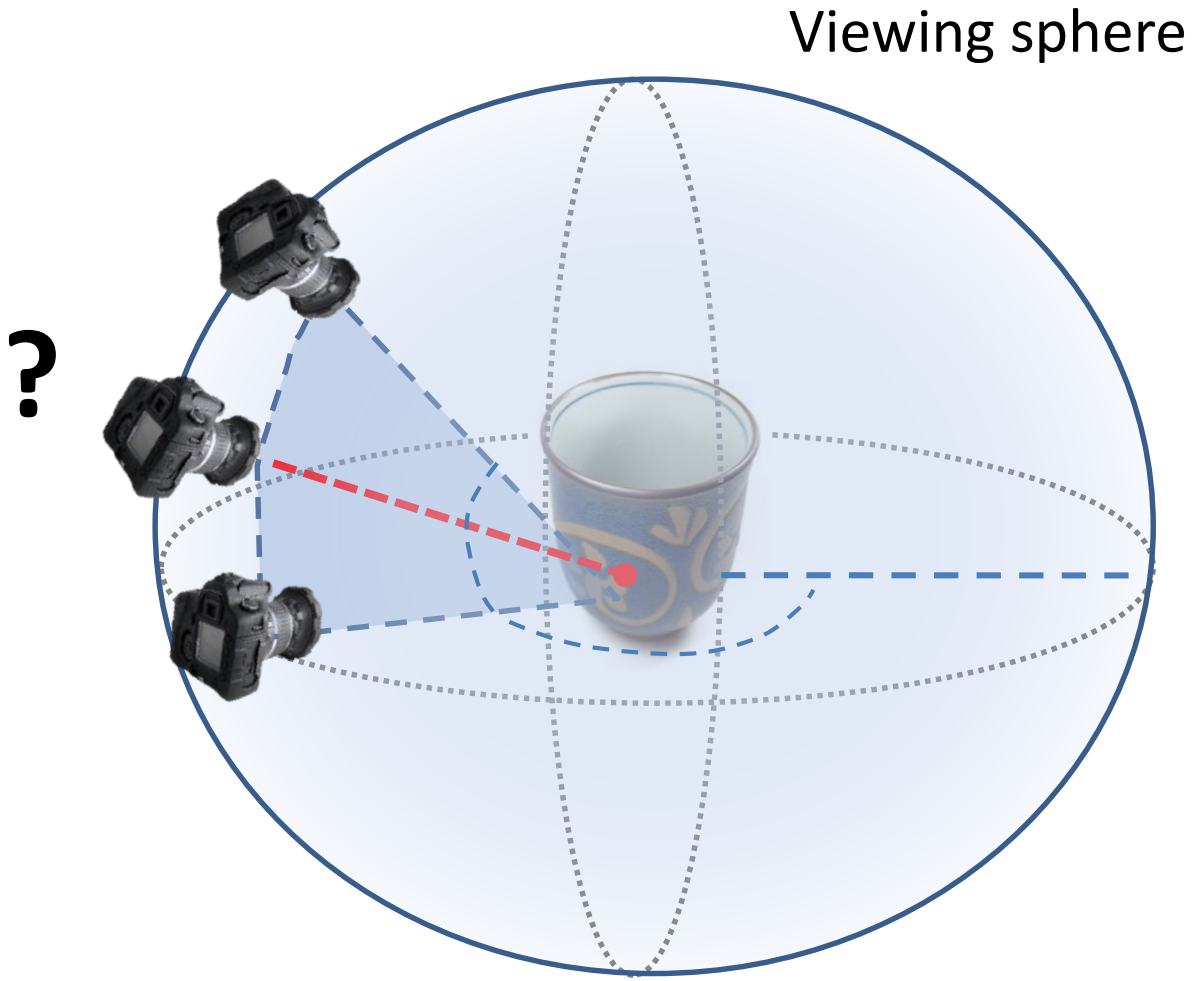
Car



Travel Iron



Predicting object appearance from novel views



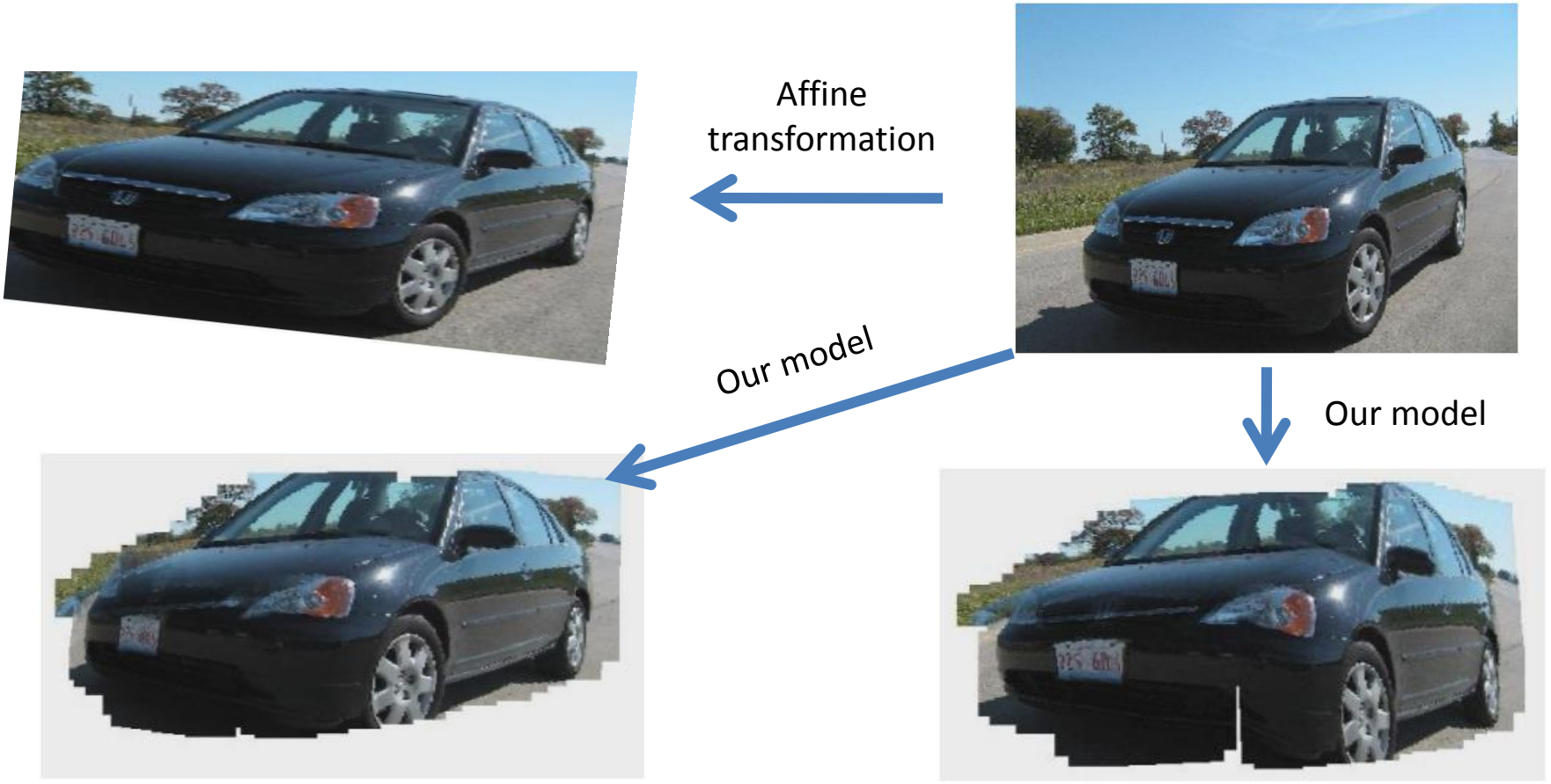
Predicting object appearance from novel views

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

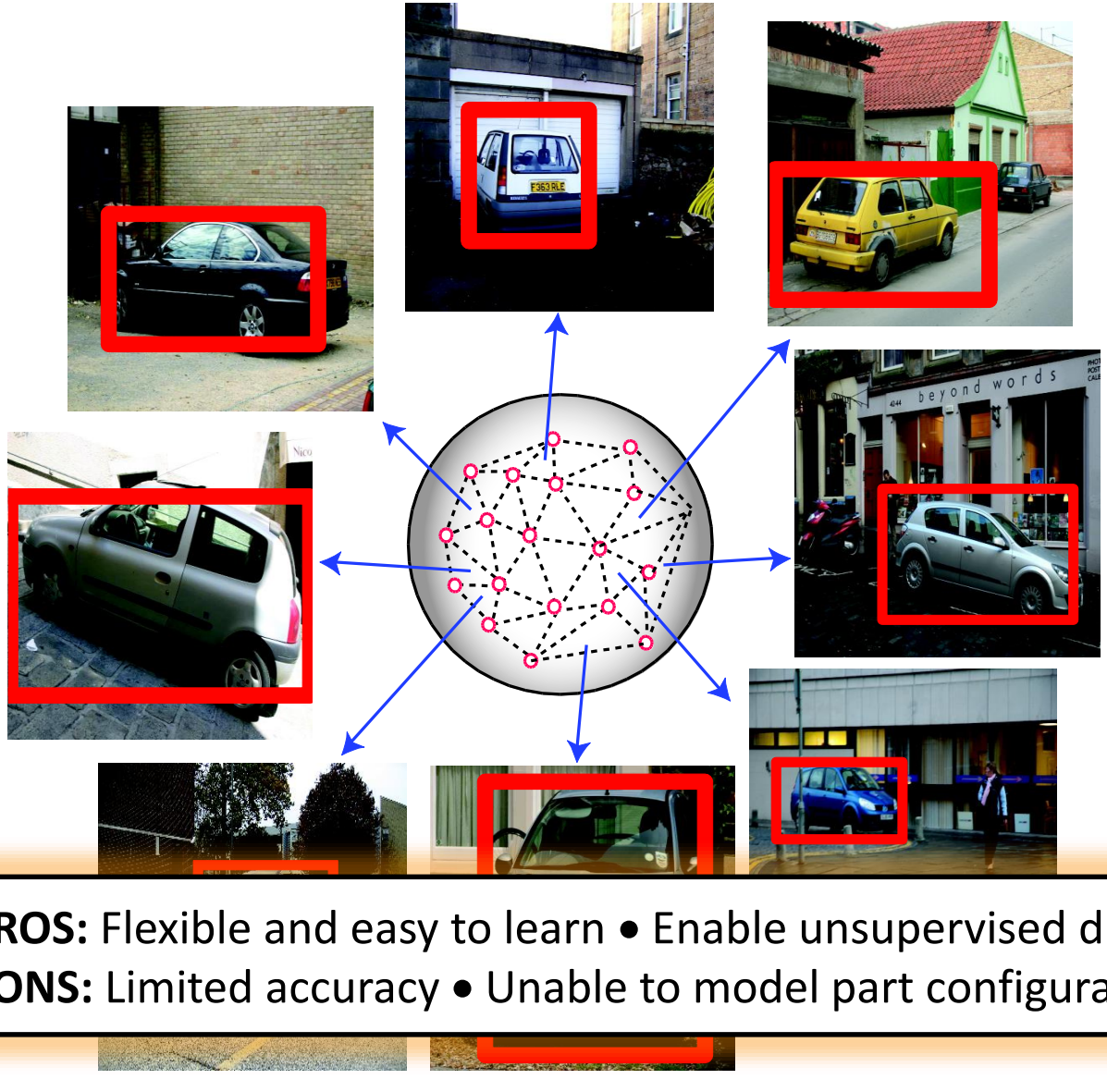
Thomas et al 08
Cremer et al 09


Cars

Predicting object appearance from novel views



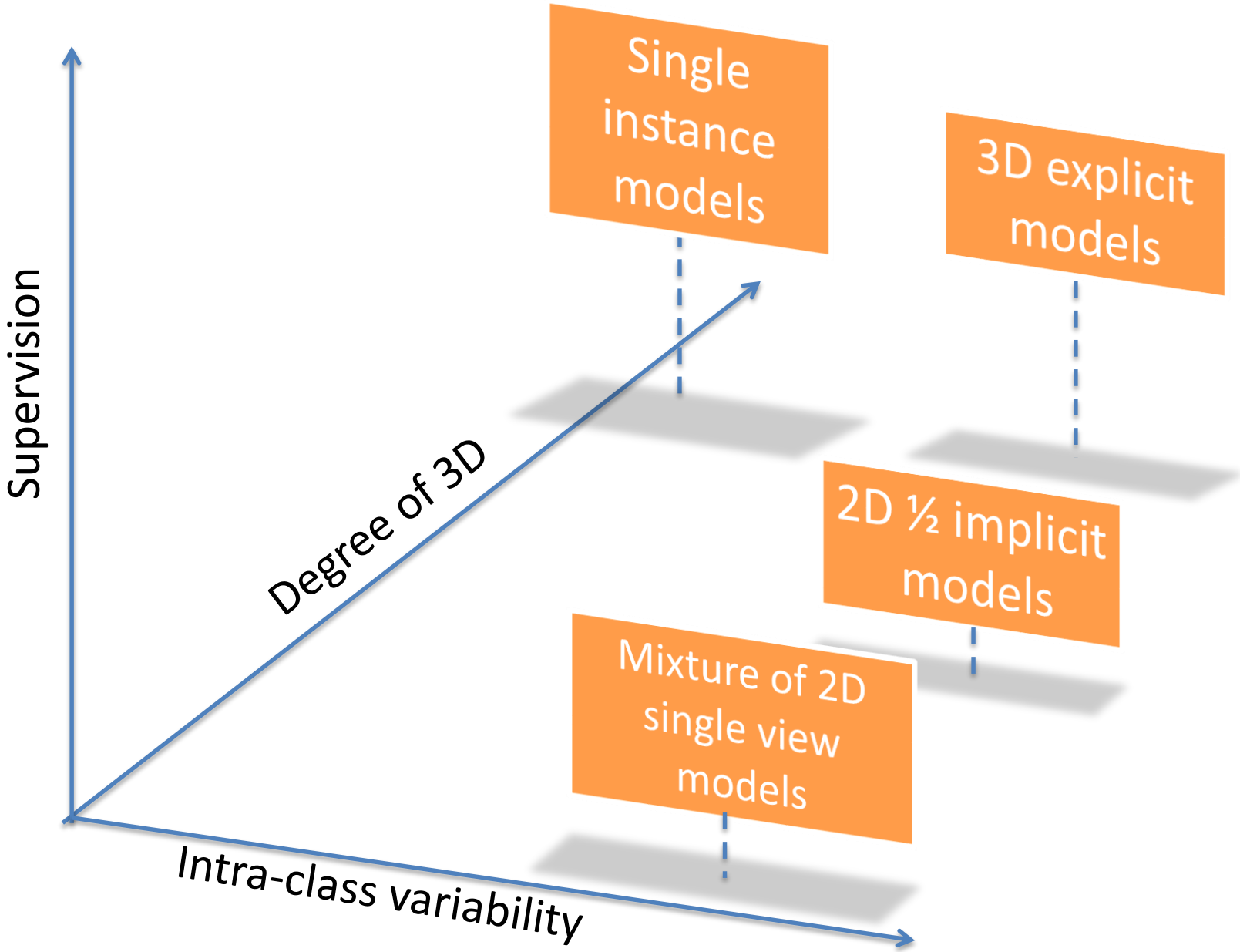
2D ½ implicit models



- Enable detection and pose estimation results
- Datasets:
  - 3D object dataset
  - VOC 06-08 Pascal
- Tested on up to 20 categories

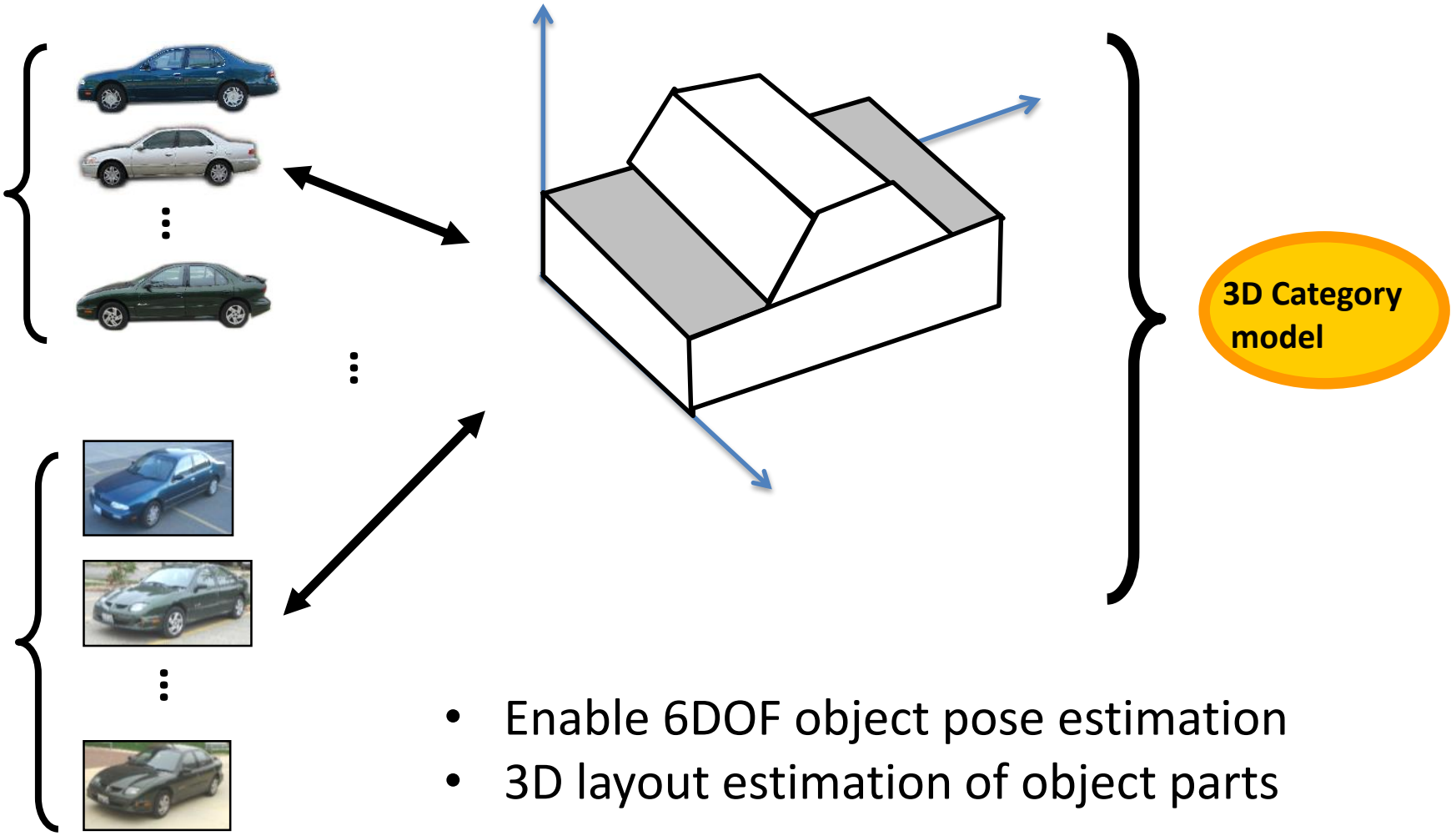
PROS: Flexible and easy to learn • Enable unsupervised discovery of parts
CONS: Limited accuracy • Unable to model part configurations in 3D

Models for 3d Object detection



3D explicit models

- Sun, Xu, Bradski, Savarese, ECCV 2010
- Sun, Kumar, Bradski, Savarese, 3DIM-PVT 2011
- Kumar, Sun, Savarese, CVPR 12
- Xiang & Savarese, CVPR 12
- Hoiem, et al. , '07
- Chiu et al . '07
- Liebelt et al. '08, 10
- Xiao et al . '08
- Yi et al. 09
- Arie-Nachimson & Barsi '09
- Sandhu et al . '09
- Hu & Zhu '10



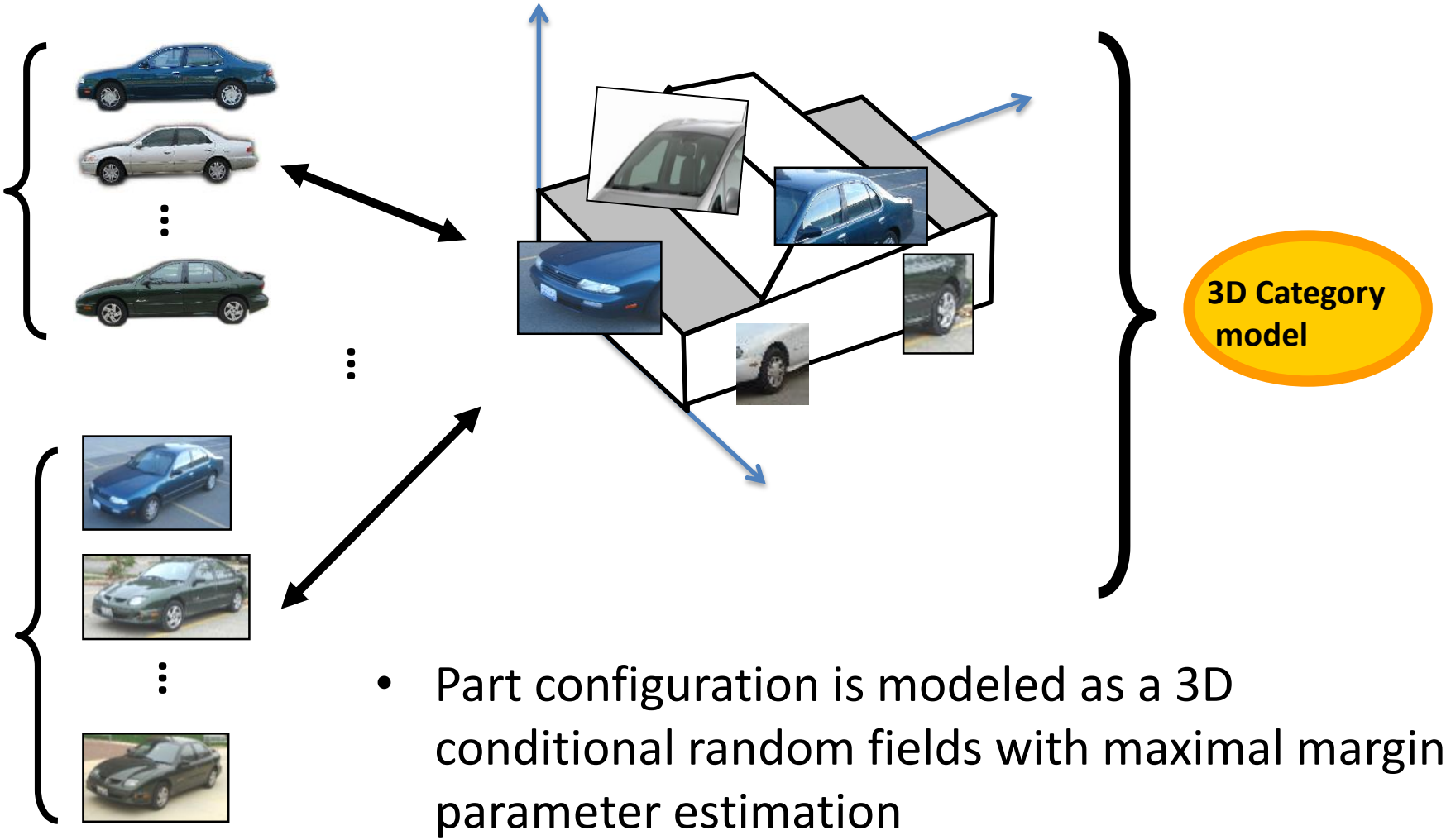
- Enable 6DOF object pose estimation
- 3D layout estimation of object parts

3D explicit models

Yan, et al. ’07

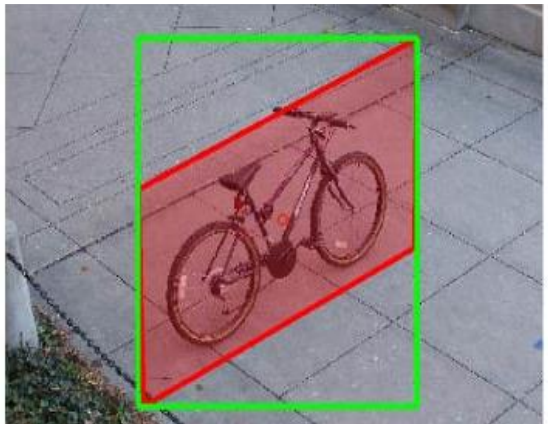
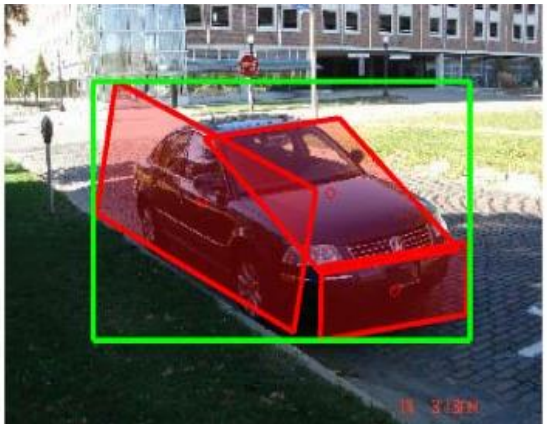
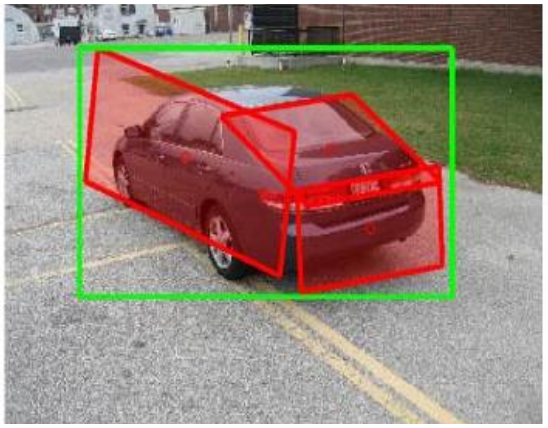
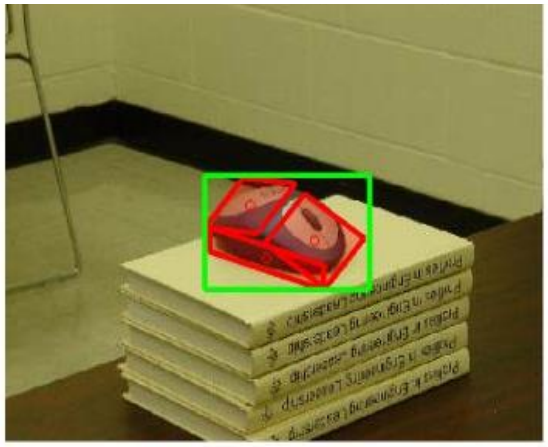
3D explicit models

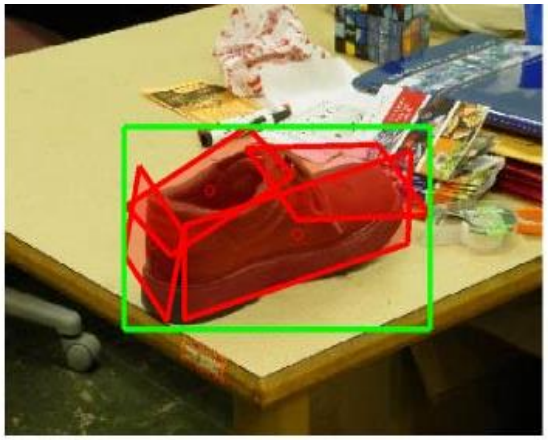
Xiang & Savarese, CVPR 12



- Part configuration is modeled as a 3D conditional random fields with maximal margin parameter estimation

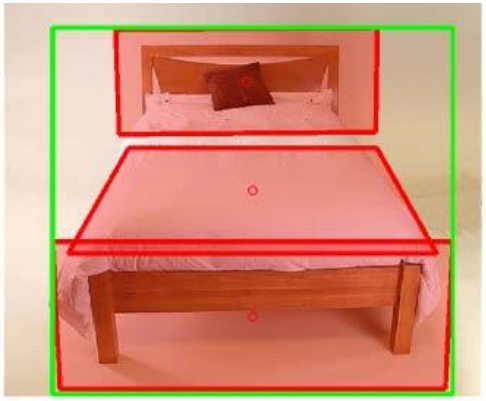
3D object detectors

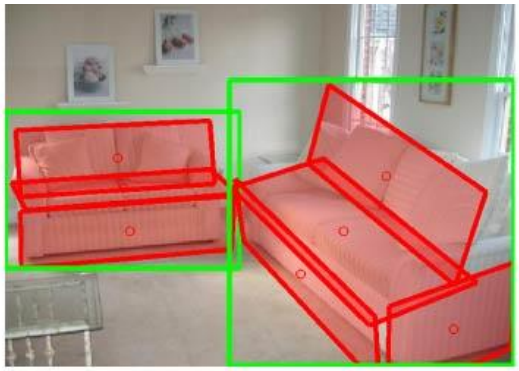





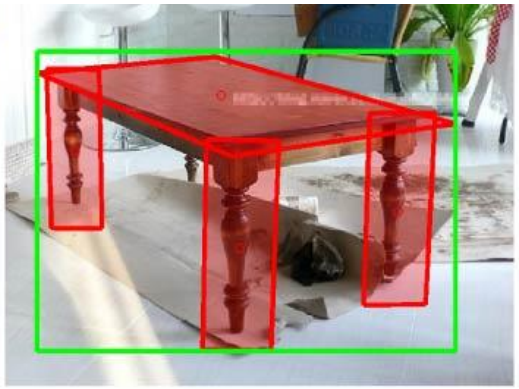


3D object dataset [Savarese & Fei-Fei, ICCV 07]

3D object detectors





ImageNet dataset [Deng et al. 2010]

3D object detectors

- Best results up-to-date in pose estimation and 3D part estimation

Cars from 3D Object dataset [Savarese 07]

Method	ours	[1]	[2]	[3]	[4]	[5]	[6]
Viewpoint (cars)	**93.4%**	85.4	85.3	81	70	67	48.5

Cars from EPFL dataset [Ozuysal 09]

Method	ours	Ours - baseline	DPM [7]	[8]
Viewpoint (cars)	**64.8%**	58.1	56.6	41.6

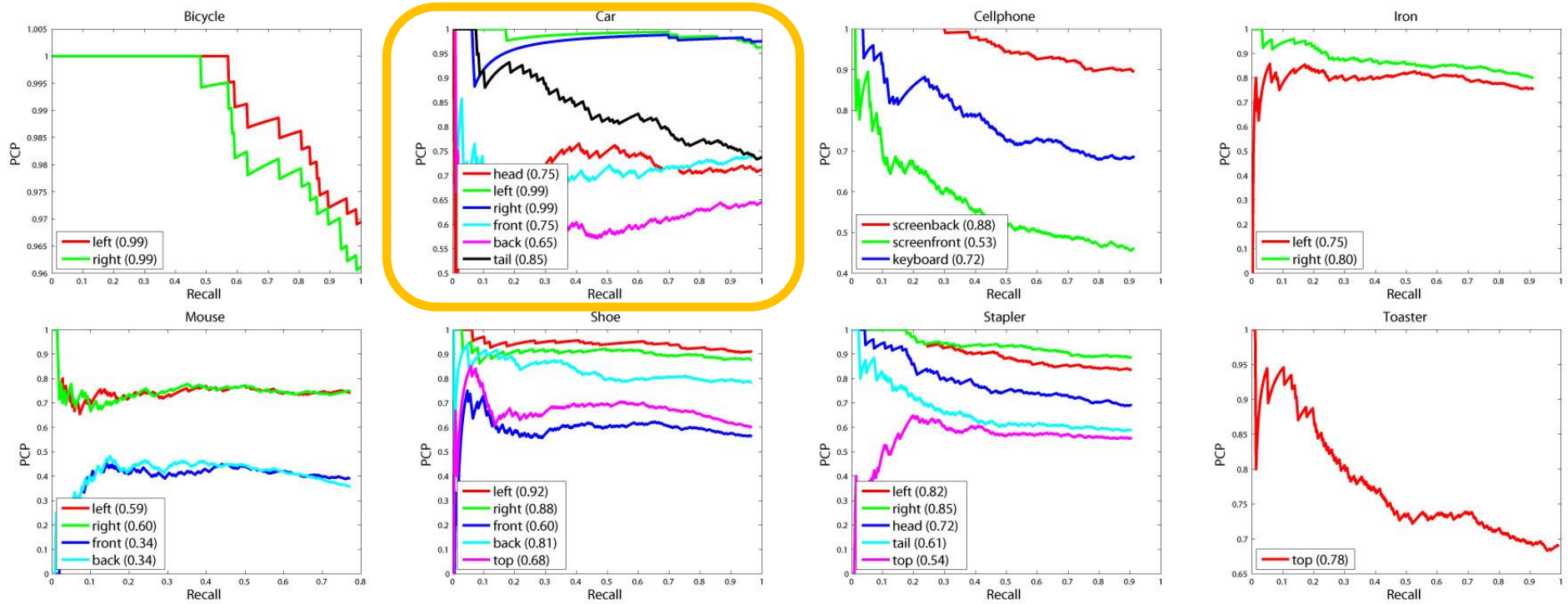
Chairs, tables and beds from IMAGE NET [Deng et al. CVPR09]

Method	ours	Ours - baseline	DPM [7]
Viewpoint	**63.4%**	34.0	49.5

[1] N. Payet and S. Todorovic. From contours to 3d object detection and pose estimation. In ICCV, 2011.
[2] D. Glasner, M. Galun, S. Alpert, R. Basri, and G. Shakhnarovich. Viewpoint-aware object detection and pose estimation. In ICCV, 2011.
[3] M. Stark, M. Goesele, and B. Schiele. Back to the future: Learning shape models from 3d cad data. In BMVC, 2010.
[4] J. Liebelt and C. Schmid. Multi-view object class detection with a 3D geometric model. In CVPR, 2010.
[5] H. Su, M. Sun, L. Fei-Fei, and S. Savarese. Learning a dense multiview representation for detection, viewpoint classification. In ICCV, 2009.
[6] M. Arie-Nachimson and R. Basri. Constructing implicit 3d shape models for pose estimation. In ICCV, 2009.
[7] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. TPAMI, 2010.
[8] M. Ozuysal, V. Lepetit, and P. Fua. Pose estimation for category specific multiview object localization. In CVPR, 2009.

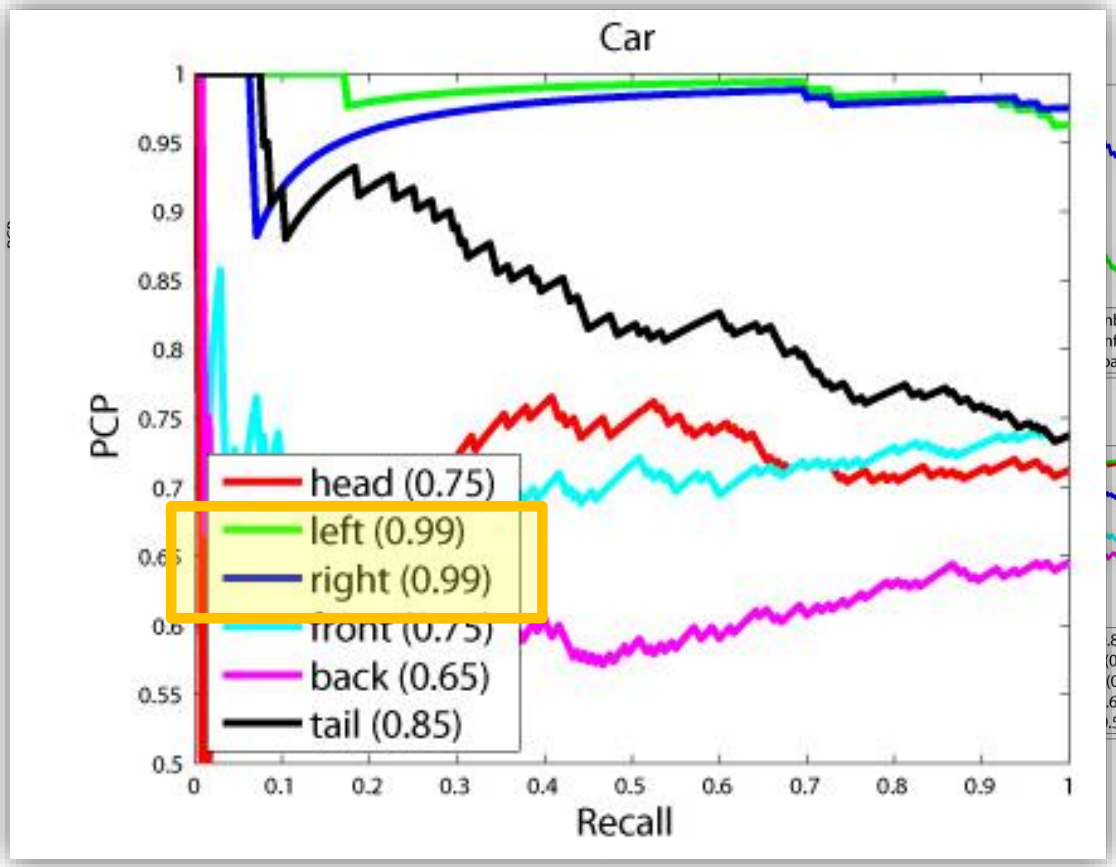
3D object detectors

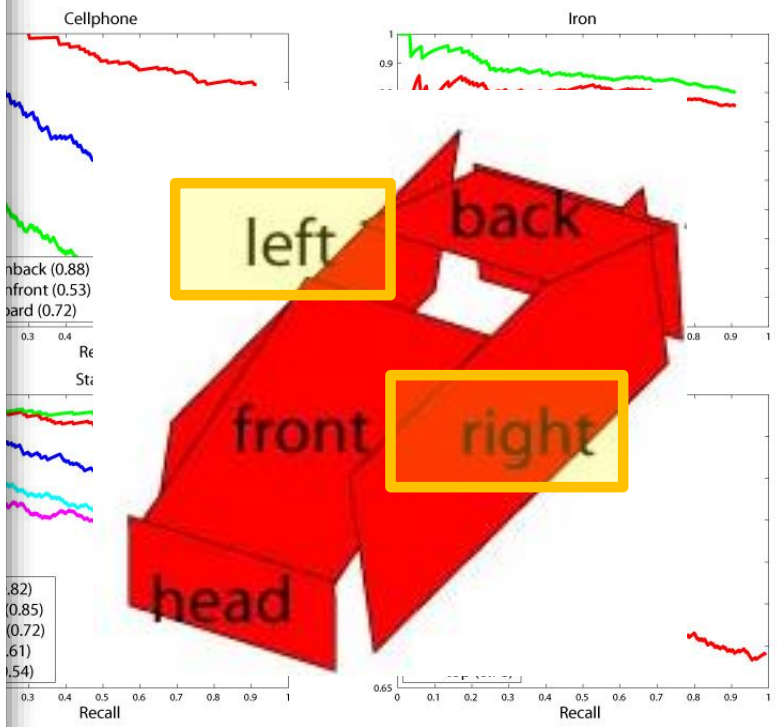
- Part localization on the 3DObject dataset



3D object detectors

- Part localization on the 3DObject dataset





Source code available! Please visit: http://www.eecs.umich.edu/vision/research.html

Agenda on recognition

Classification (images; areas)
- bag of words
- Pyramid matching

Detection (use slides with 3 axis: category; supervision; 3D info. Use intro job talk)
- Template based (holistic; part based)
- Multi-view (single instance; categories; 3D pose)

Scene understanding
- Segmentation (bottom up; semantic)
- 3D scene understanding

Activity understanding