Toward Coherent Object Recognition and Scene Layout Understanding

VISION LAB

Silvio Savarese – University of Michigan at Ann Arbor

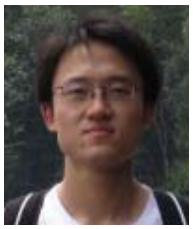

Byung Kim


Wongun Choi


Min Sun


Ying-ze Bao

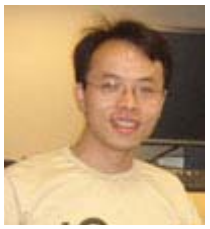

Ryan Tokola


Yu Xiang


Anush Mohan


Shyam Kumar

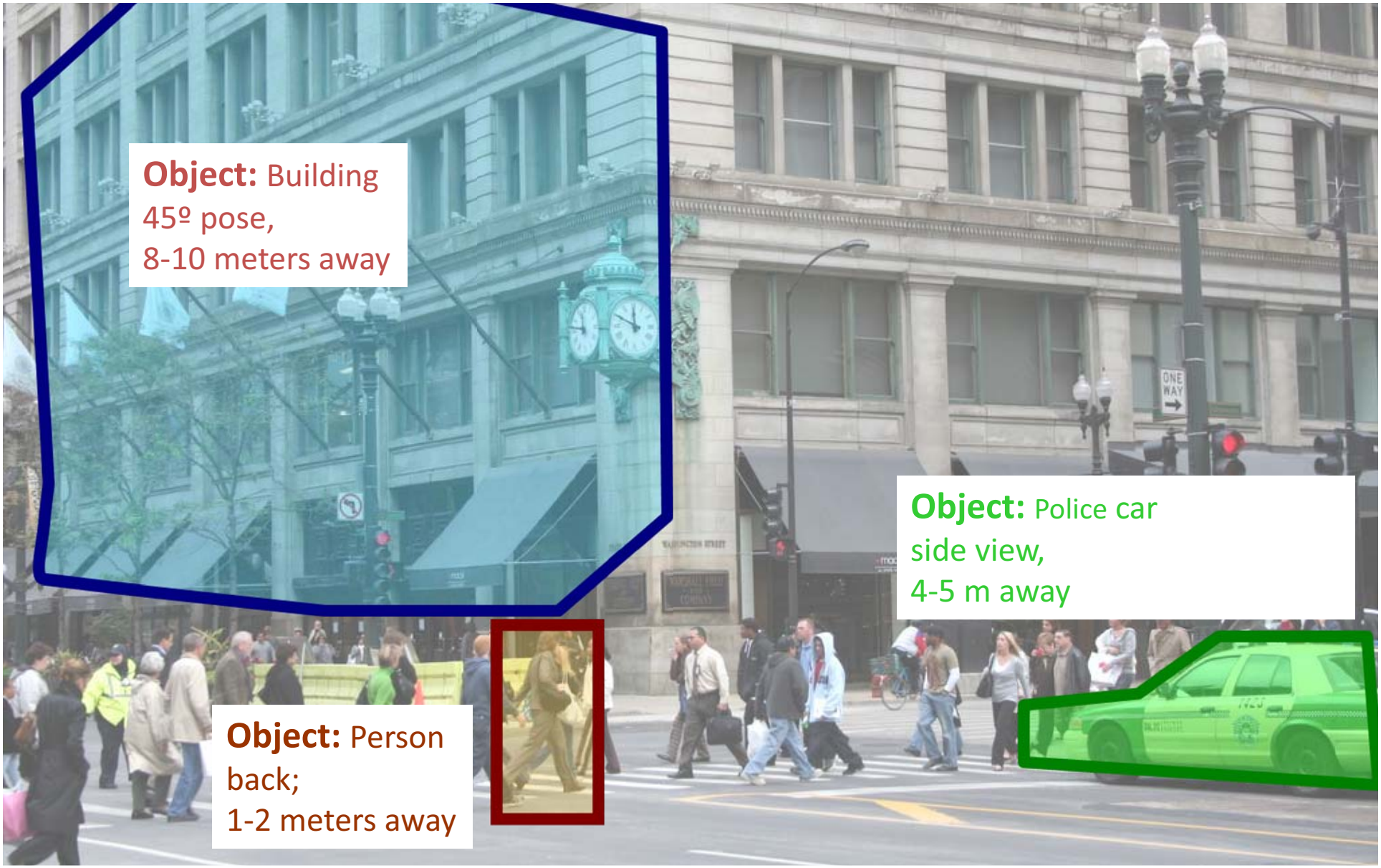

Fisher Yu

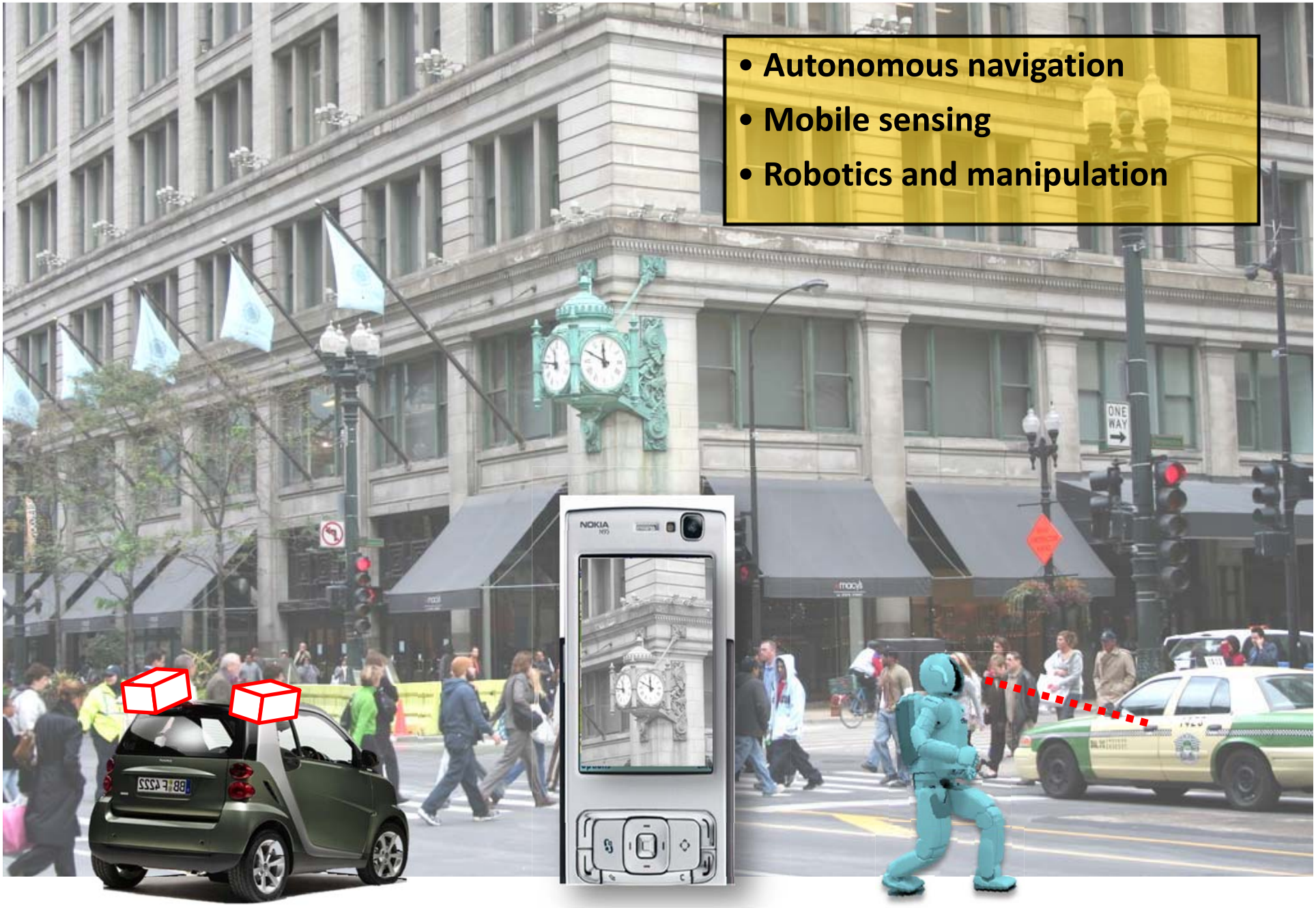

Jingen Liu

28 Dicembre 2010


ONE
WAY

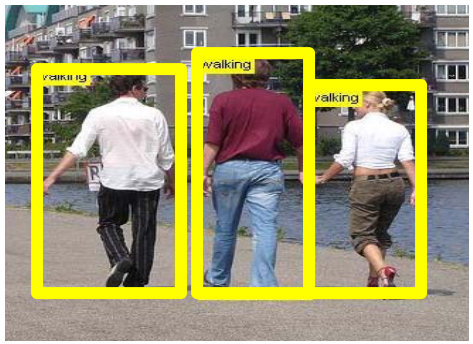
Geometrically “rich” scene understanding




• Autonomous navigation
• Mobile sensing
• Robotics and manipulation
NOKIA
ONE WAY

Semantic

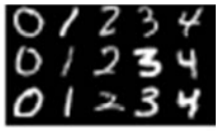
Object detection and tracking



Object categorization

Scene classification

Pattern recognition

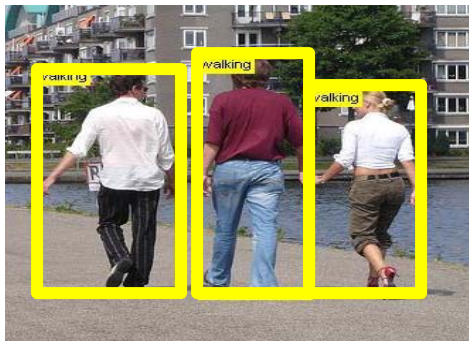


3D geometry

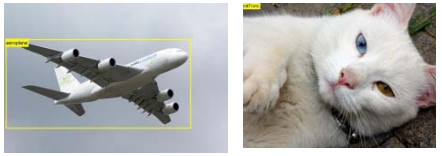
Joint semantic interpretation and geometric reconstruction of the scene

Semantic

Object detection and tracking

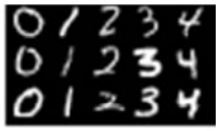


Object categorization

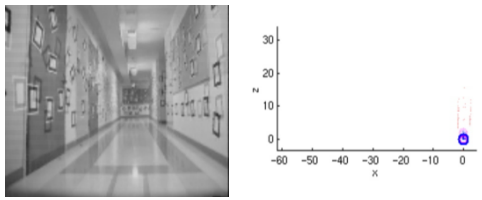


Scene classification

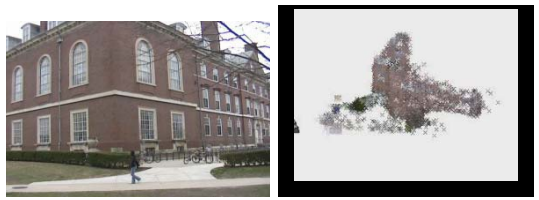
Pattern recognition



Camera localization



3D object & scene modeling

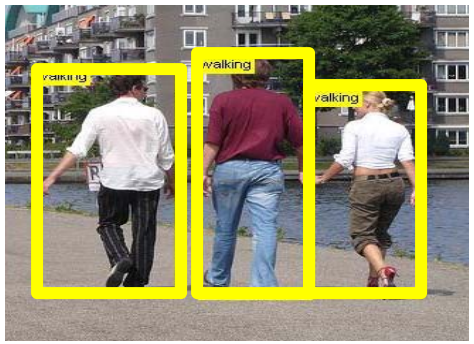


3D geometry

Joint semantic interpretation and geometric reconstruction of the scene

Semantic

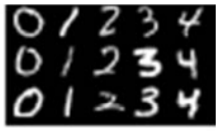
Object detection and tracking

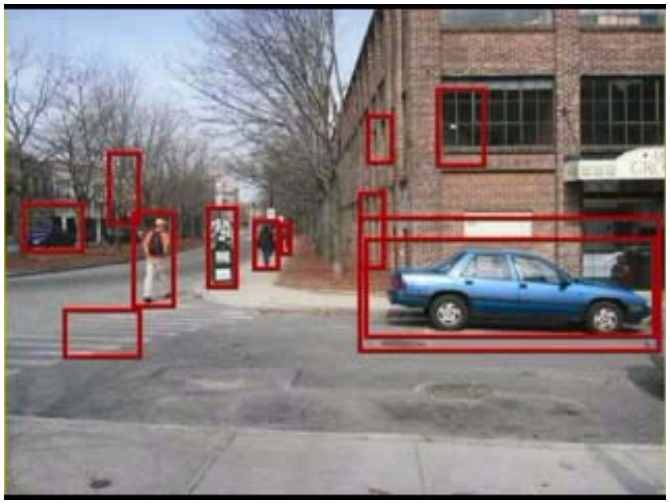


Object categorization

Scene classification

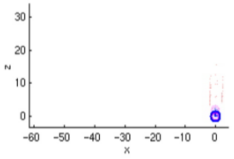
Pattern recognition



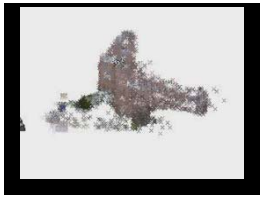


- Hoiem et al. 06-10
- Gould et al. 09
- Hedau et al. 09
- Gupta et al, 10
- Bao, Sun, Savarese 10
- Sun, Bao, Savarese 10

Camera localization



3D object & scene modeling



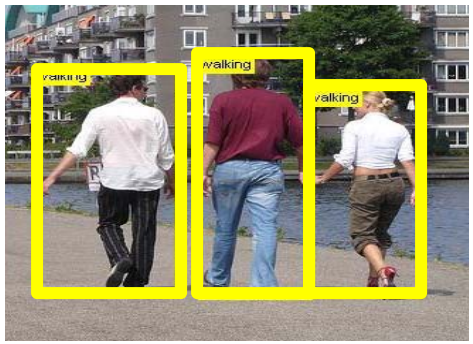
3D geometry

Joint semantic interpretation and geometric reconstruction of the scene

Semantic

3D geometry

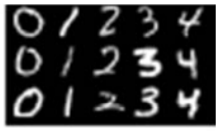
Object detection and tracking

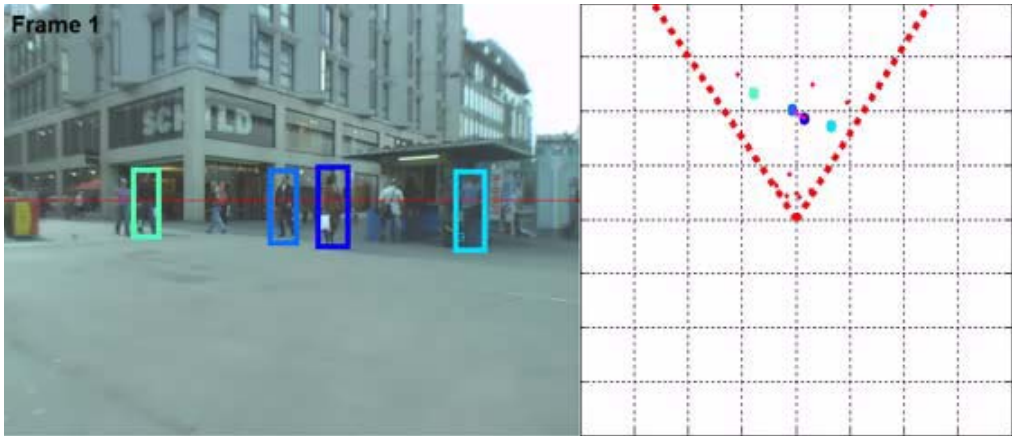


Object categorization

Scene classification

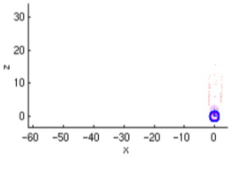
Pattern recognition



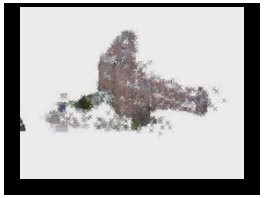


- Ess et al, 2009
- Pellegrini et al , 2010
- Choi , Shahid, Savarese , 2009
- Choi & Savarese, 2010

Camera localization

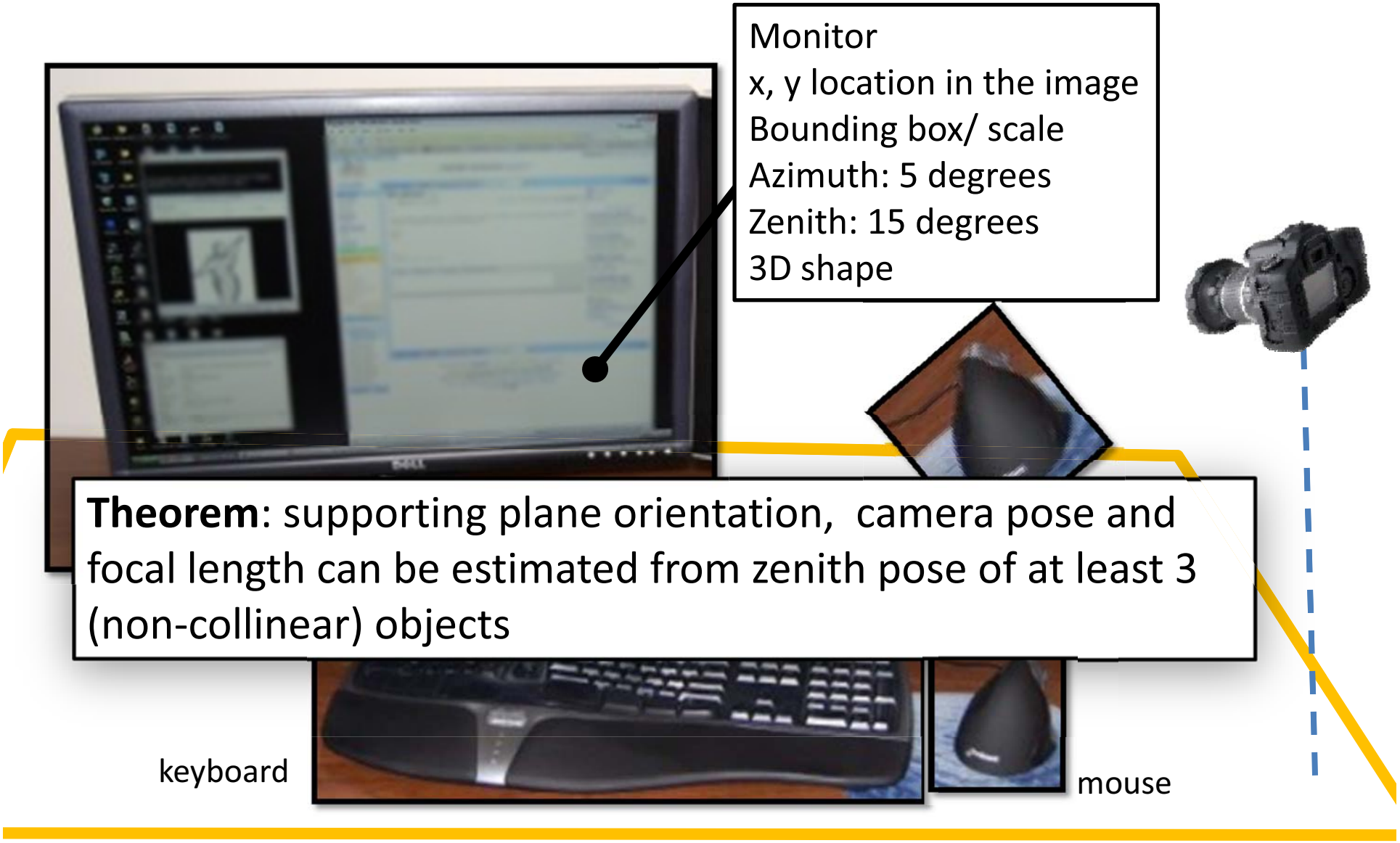


3D object & scene modeling



How can we achieve all of this?

Intuition: Objects in the 3D physical space show consistent geometrical properties within the same view and across views



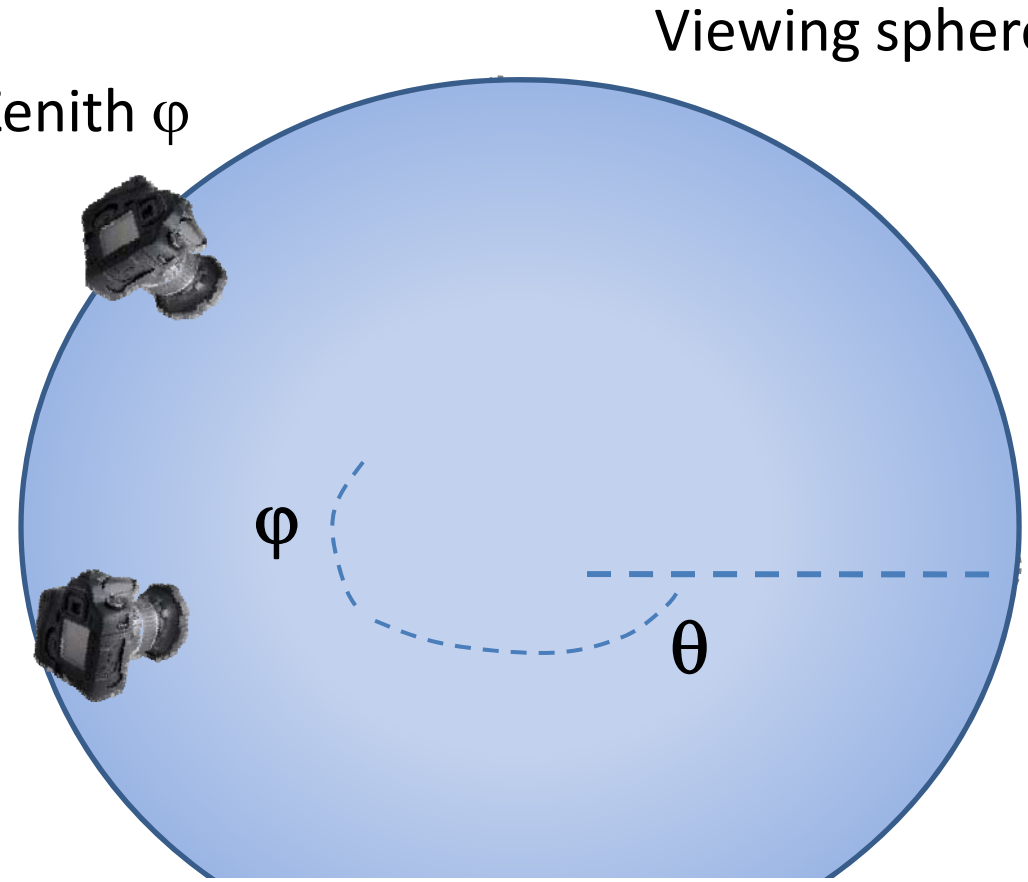
Intuition: Objects in the 3D physical space show consistent geometrical properties within the same view and across views

Our objectives

- Multi-view models for object categorization and 3D attribute estimation

- Coherent object detection and scene layout estimation

Our goals



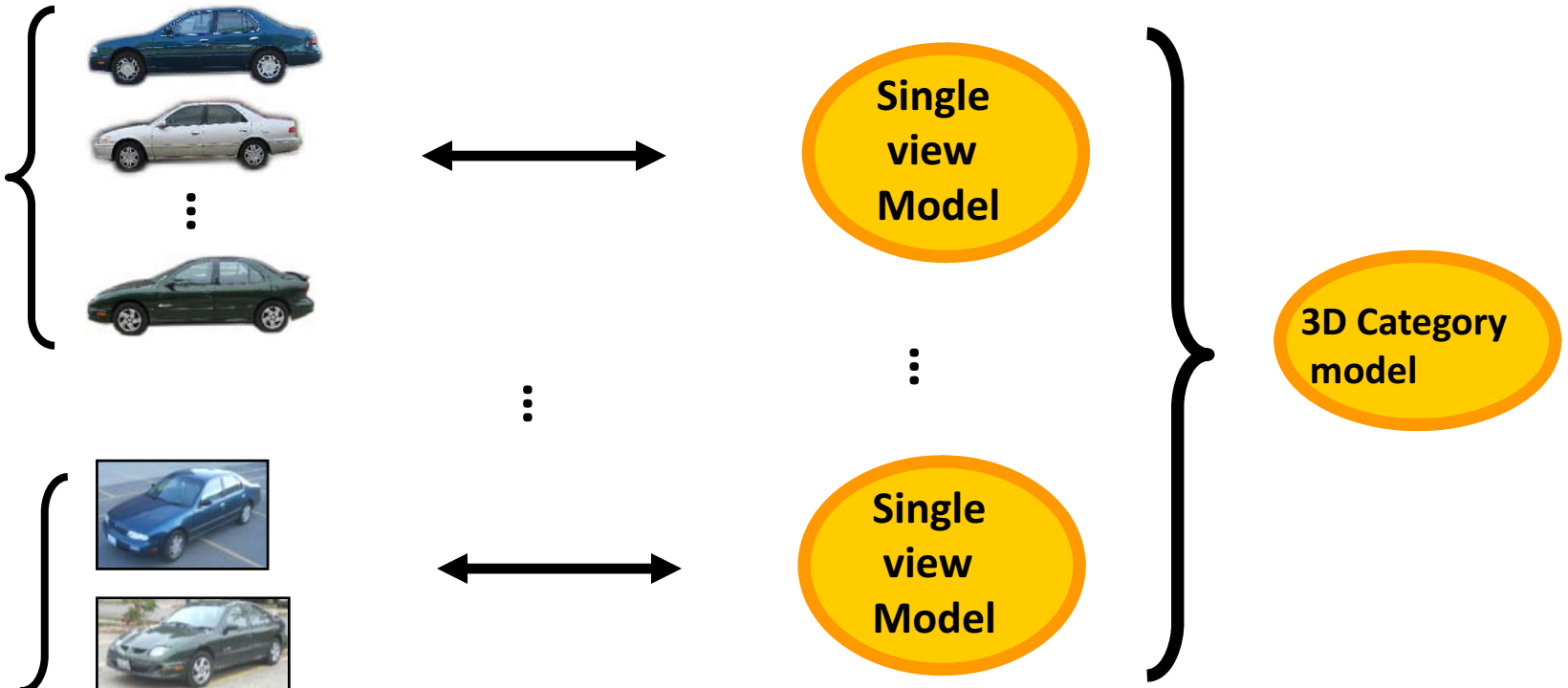
- Detect objects under generic view points
- Estimate object pose
- General and work for any object category

Our goals

- Detect objects under generic view points
- Estimate object pose
- General and work for any object 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
  - No sense of correspondences of parts under 3D transformations
- Non scalable to large number of categories/view-points

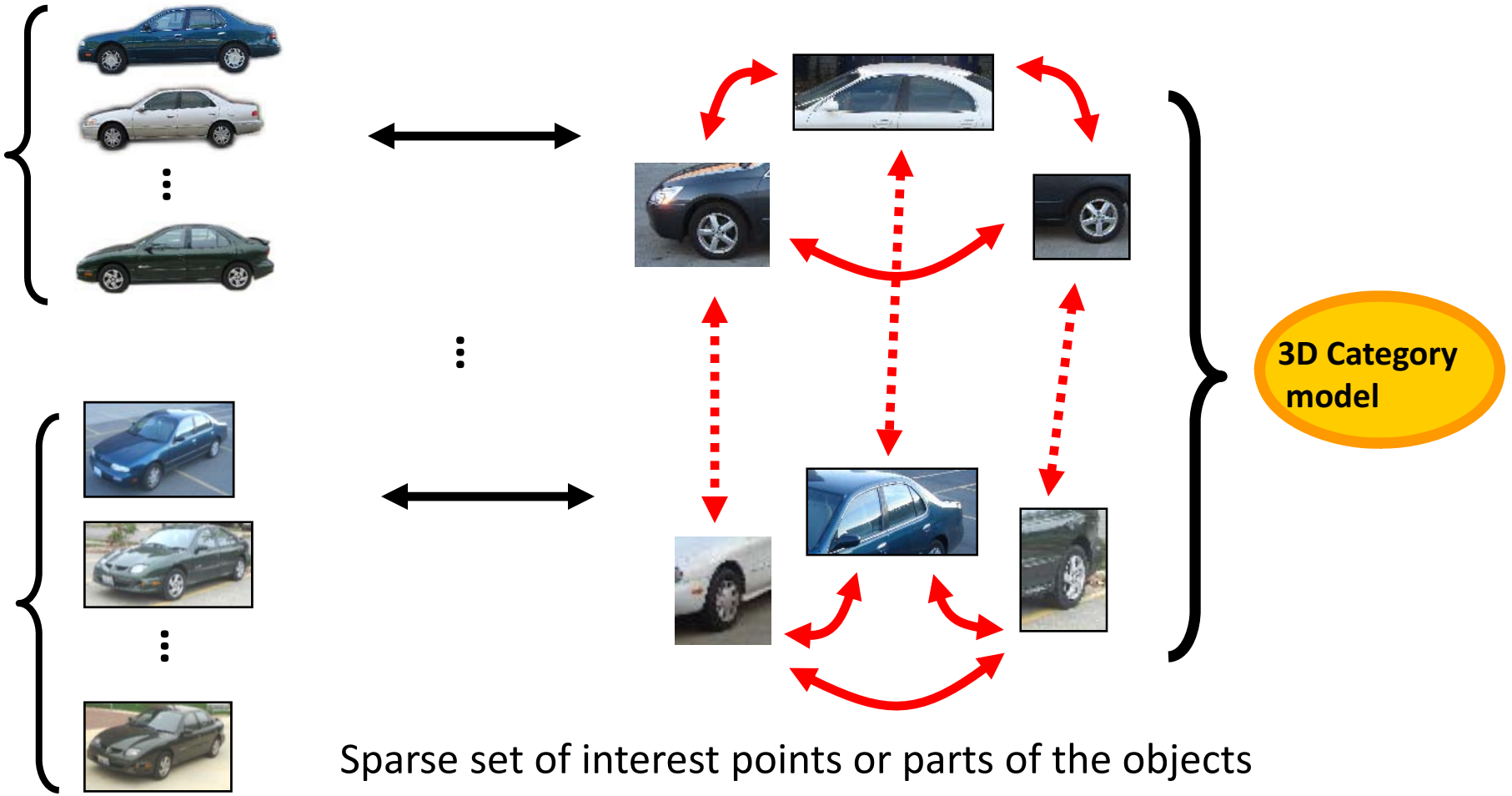
A new recent paradigm

- 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
- Sun et al 09

- Liebelt et al., '08
- Xiao et al.,'08
- Arie-Nachimson & Basri, '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

Su et al, ICCV 2009
Sun, et al, CVPR 2009

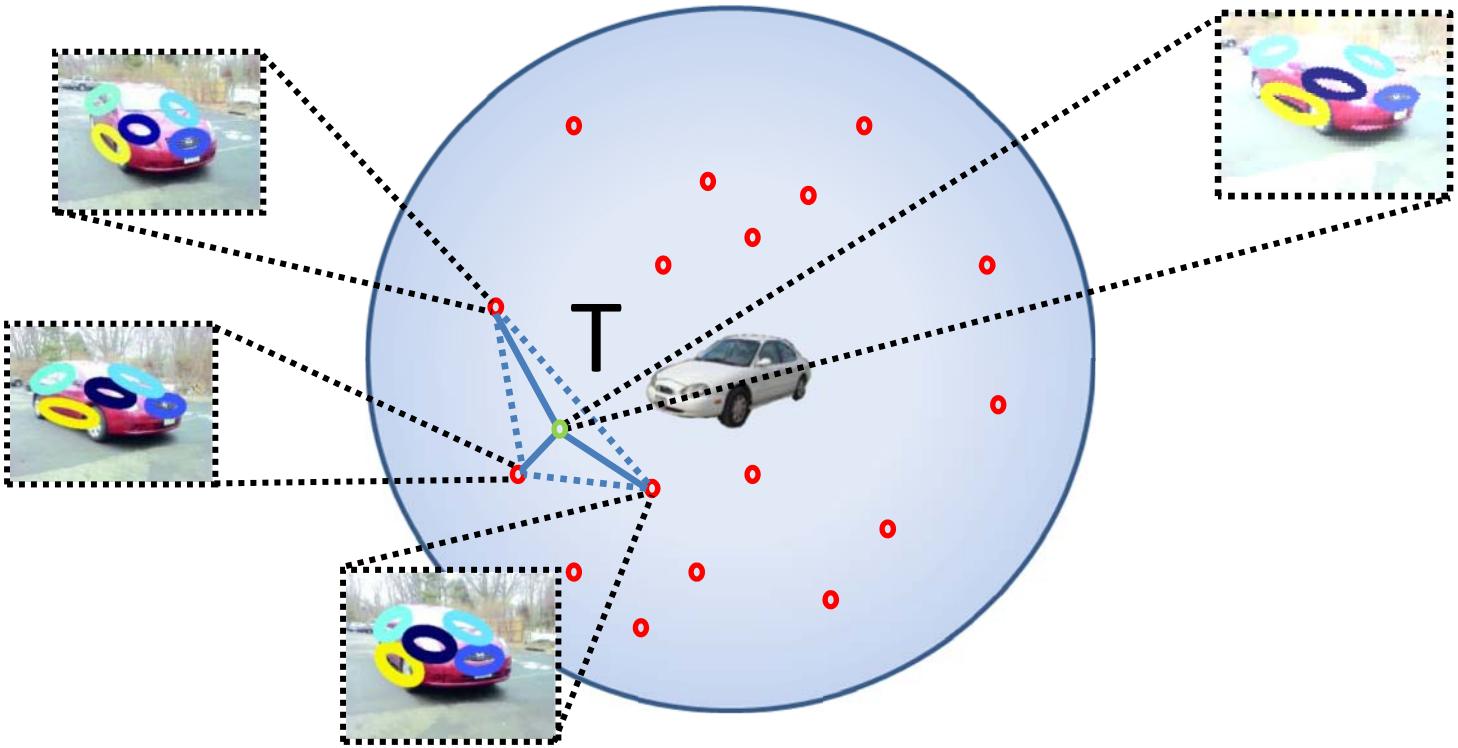
Min Sun

- 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

Key contributions

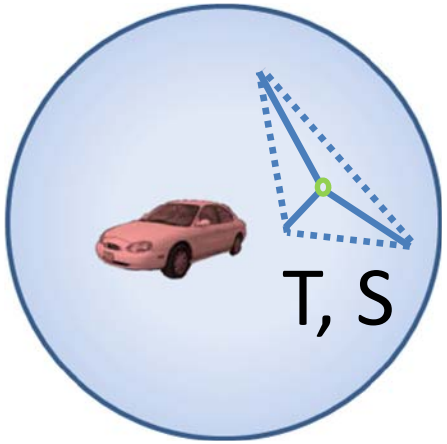
- **Representation:**
- **Learning:**
  - Dense representation on the viewing sphere:
    - Model object appearance and shape from any position on the viewing sphere
    - Enable view synthesis from novel view points
  - Multi-view generative part-based model
    - Object is represented by collections 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

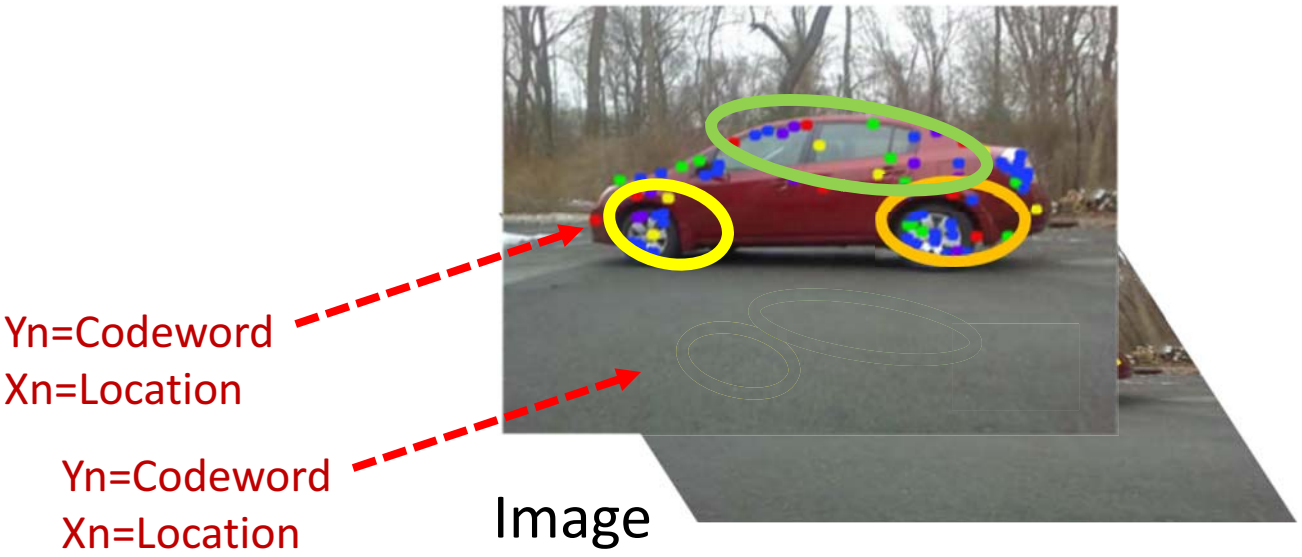
Dense representation on view-sphere



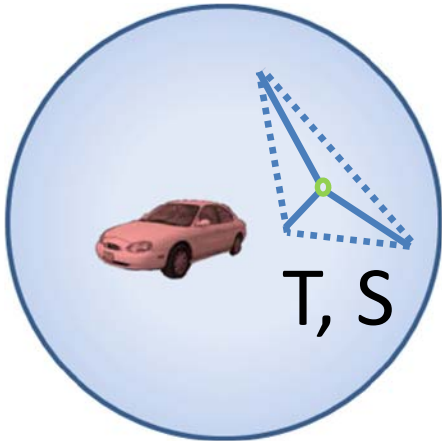
- Triangle T
- Parameter S

Multi-view generative part-based model





Multi-view generative part-based model



α = Part Prop. Prior

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

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

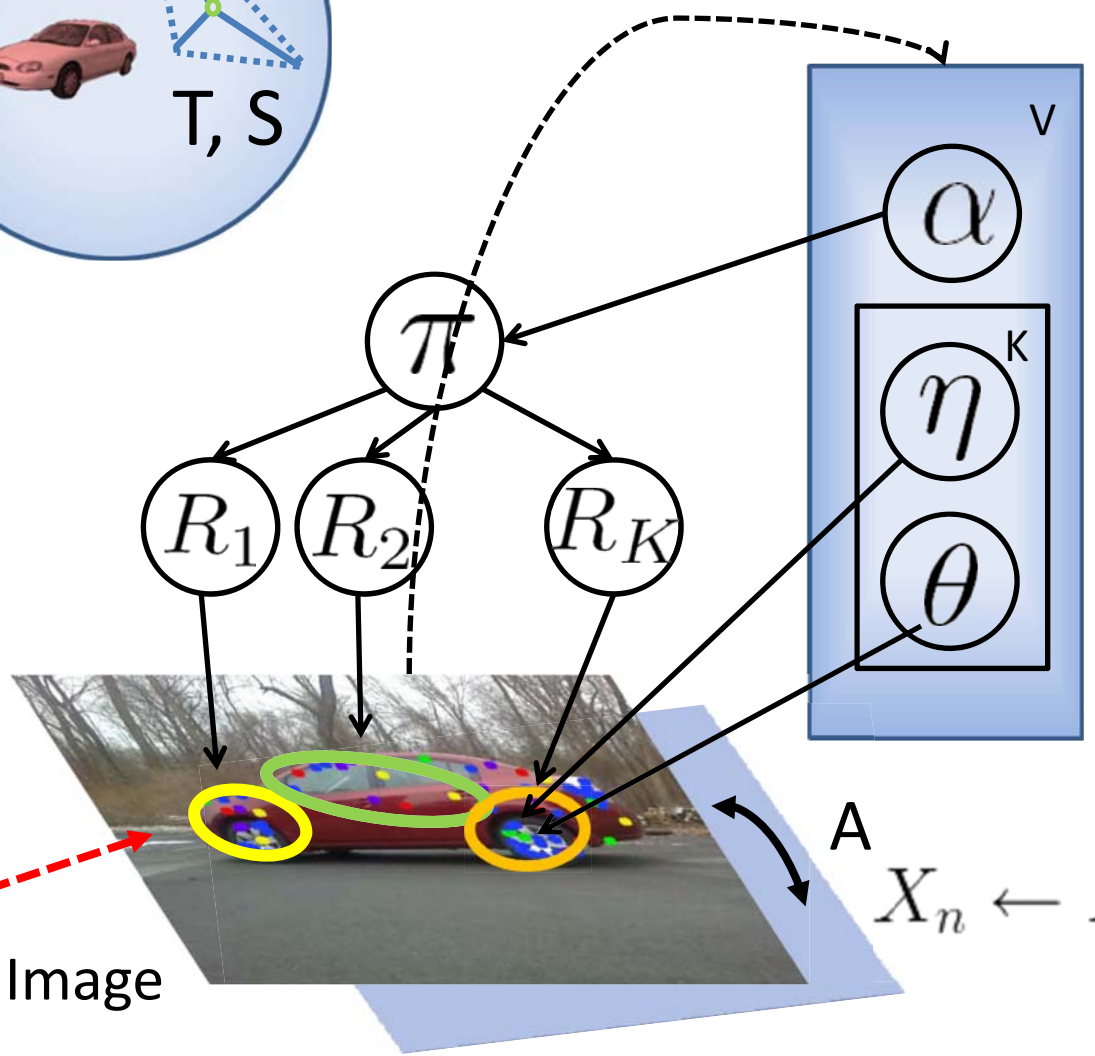
$$Y_n \sim Mult(\eta)$$

$$X_n \sim N(theta)$$

η = Part Appearance

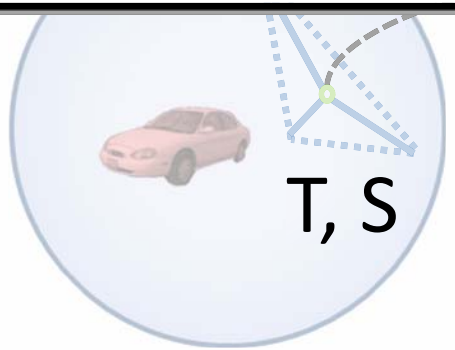
θ = Part Location/shape

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

$\pi \sim Dir(\alpha)$

$R \sim Mult(\pi)$

$Y_n \sim Mult(\eta)$

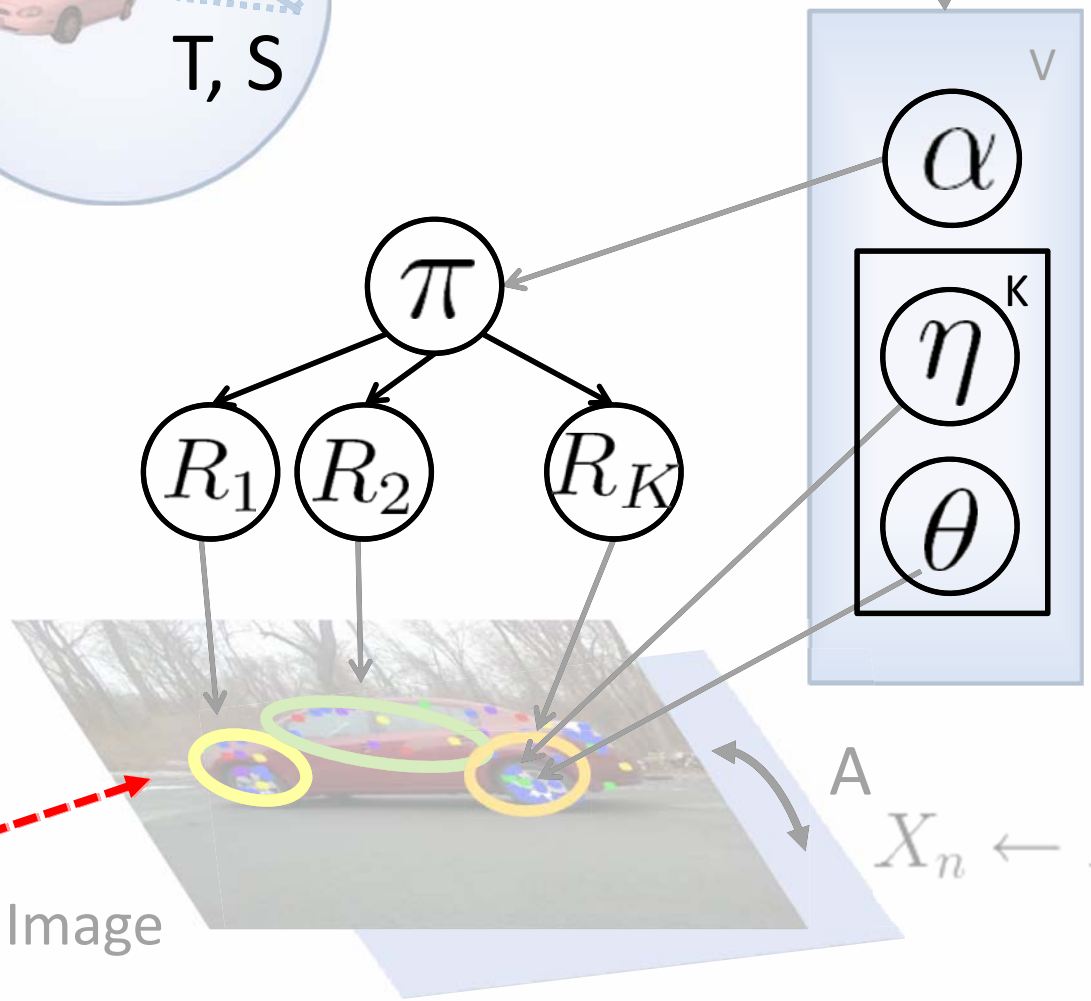
$X_n \sim N(theta)$

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:

α = Part Prop. Prior

$\pi \sim Dir(\alpha)$

$R \sim Mult(\pi)$

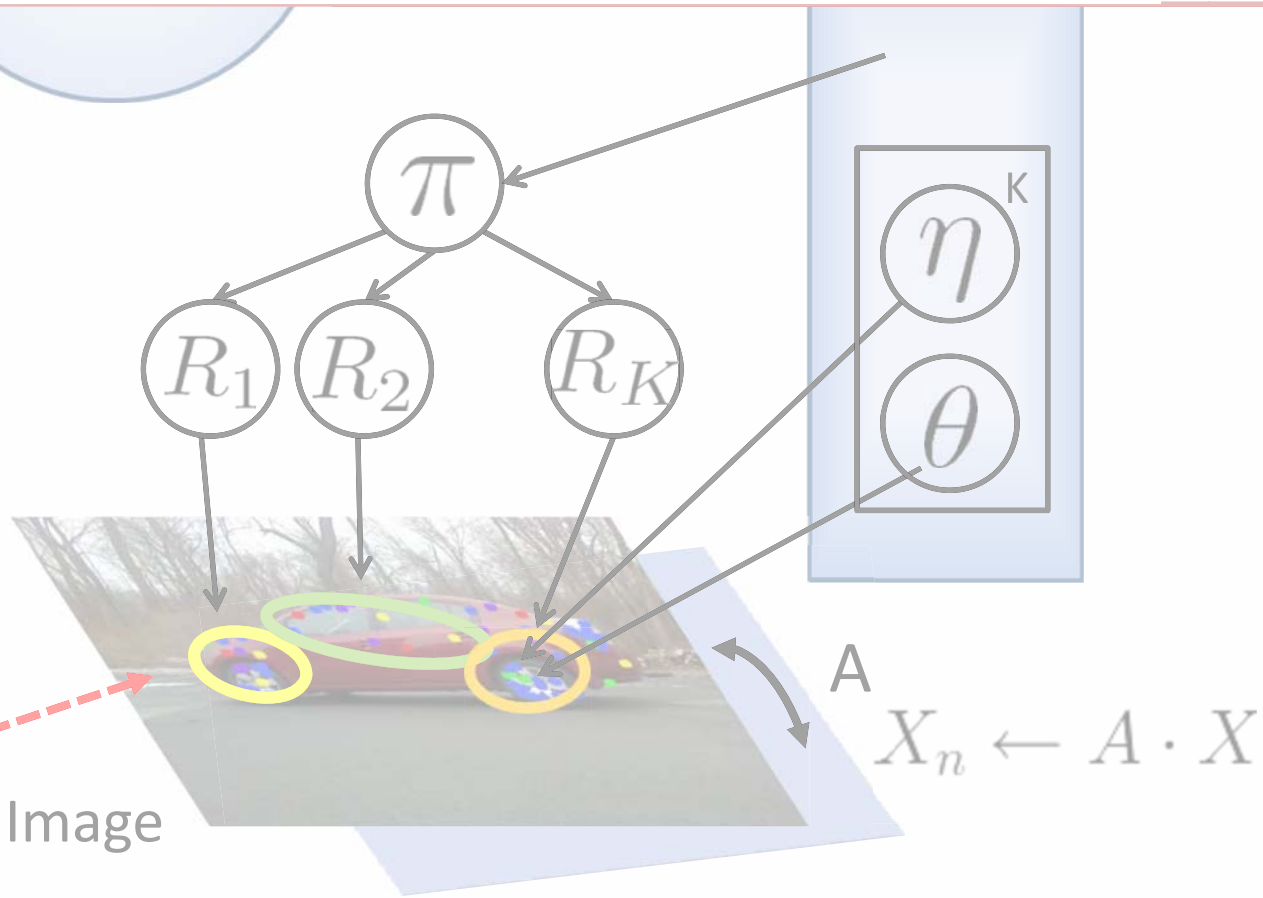
$Y_n \sim Mult(\eta)$

$X_n \sim N(theta)$

η = Part Appearance

θ = Part Location/shape

Yn=Codeword
Xn=Location

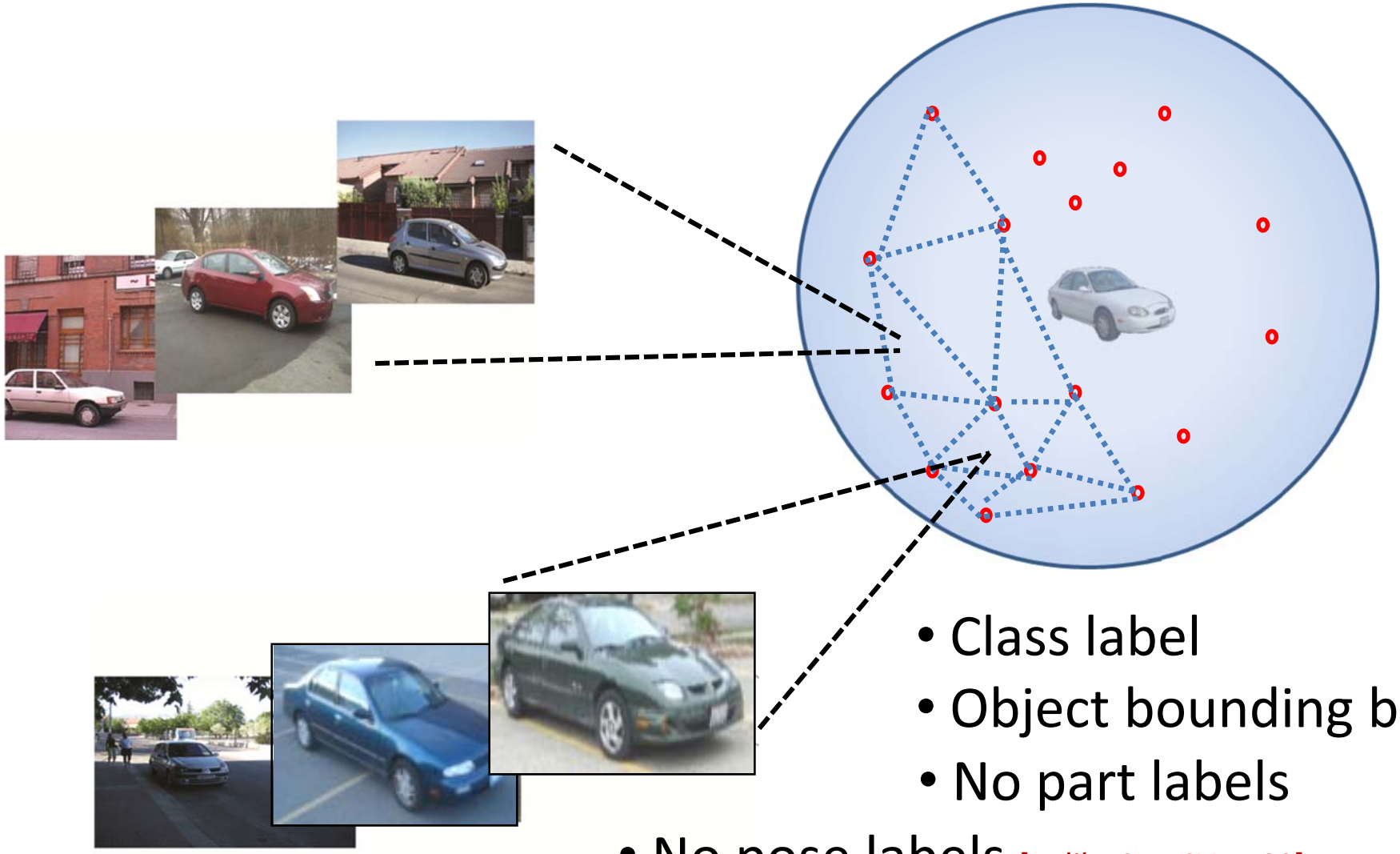


Key contributions

- **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

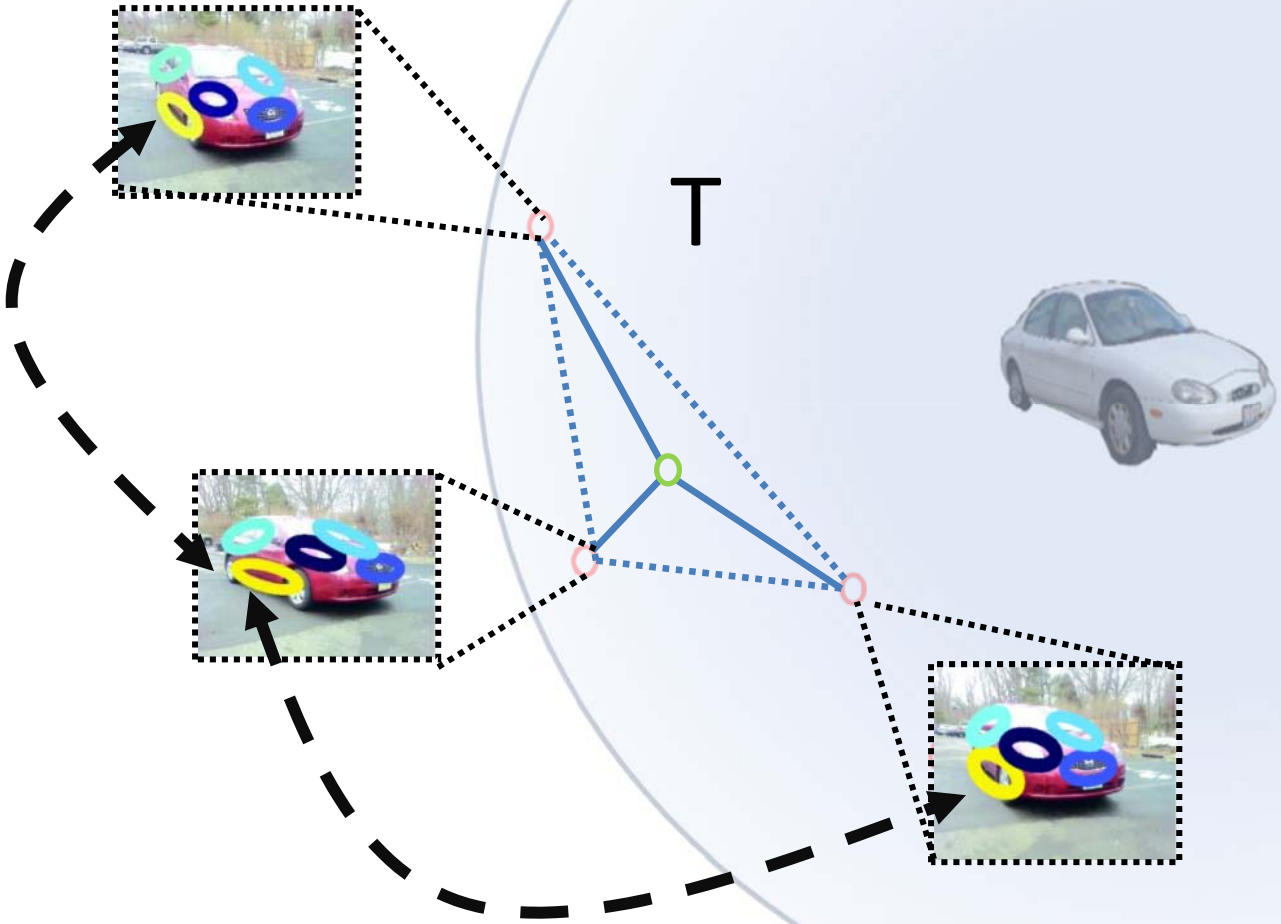
Semi-supervised



- Class label
- Object bounding box
- No part labels

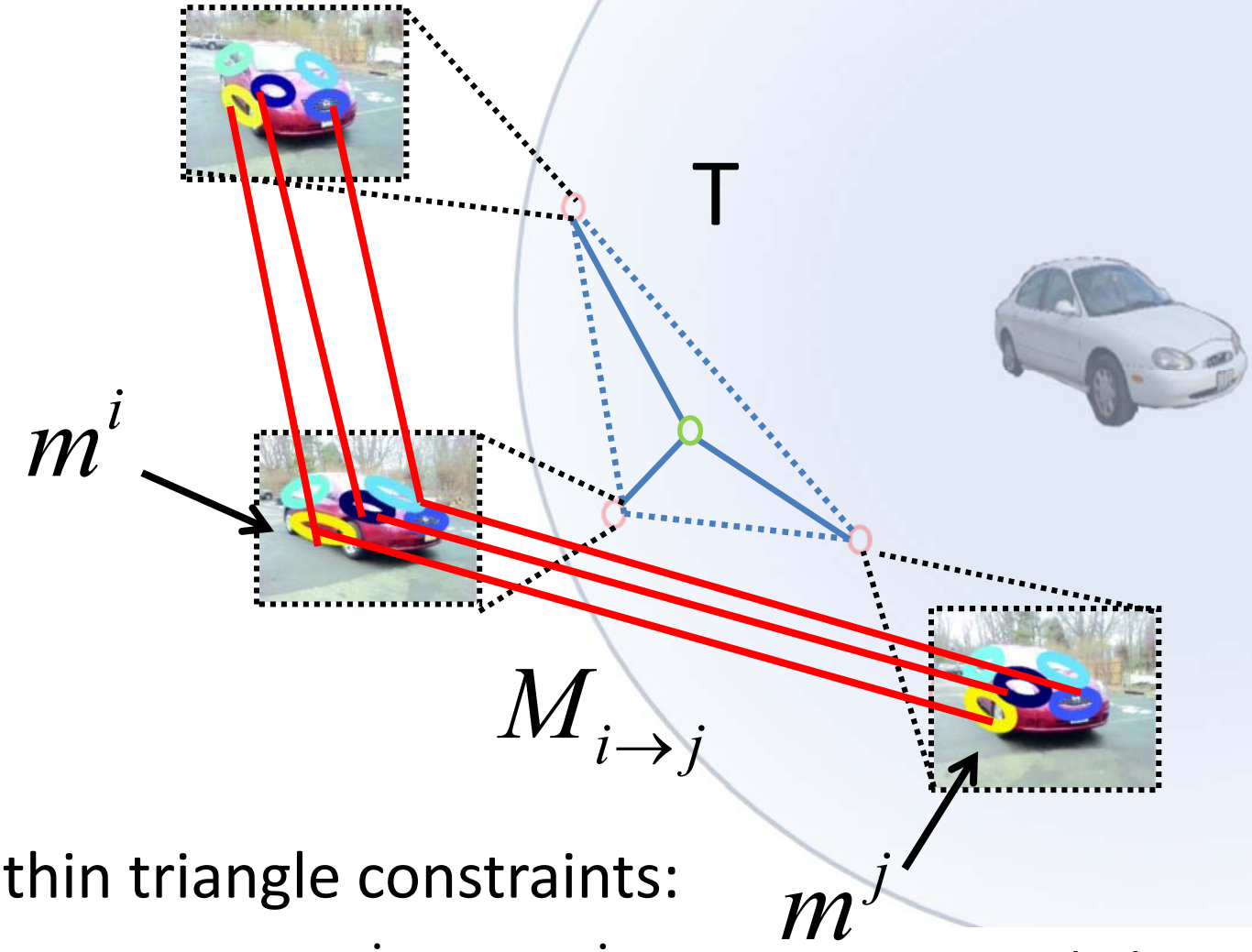
- No pose labels [unlike Sun CVPR 09]

Incorporating geometrical constraints



- Parts are linked across views
- Part topology is preserved under view point transformations

Incorporating geometrical constraints

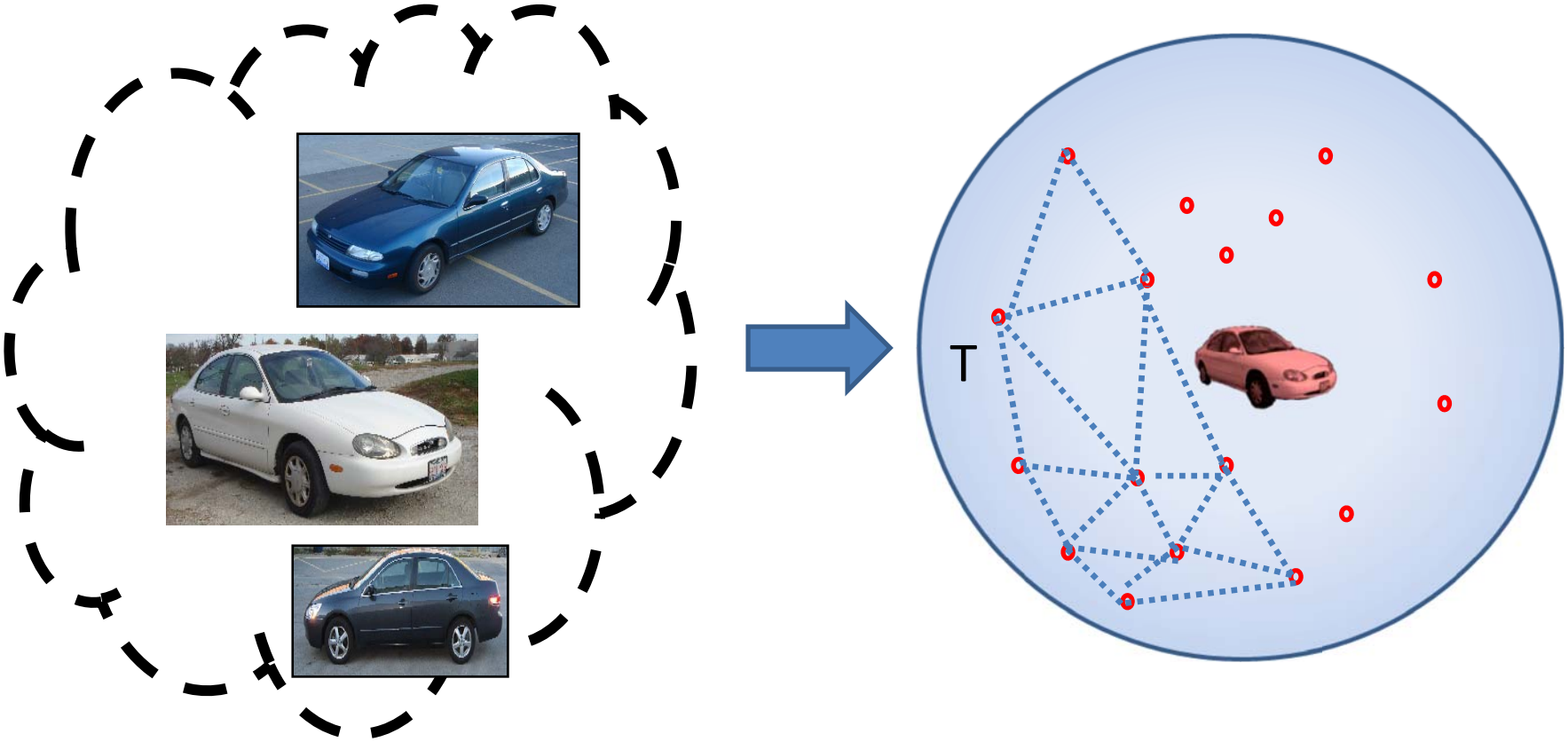


Within triangle constraints:

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

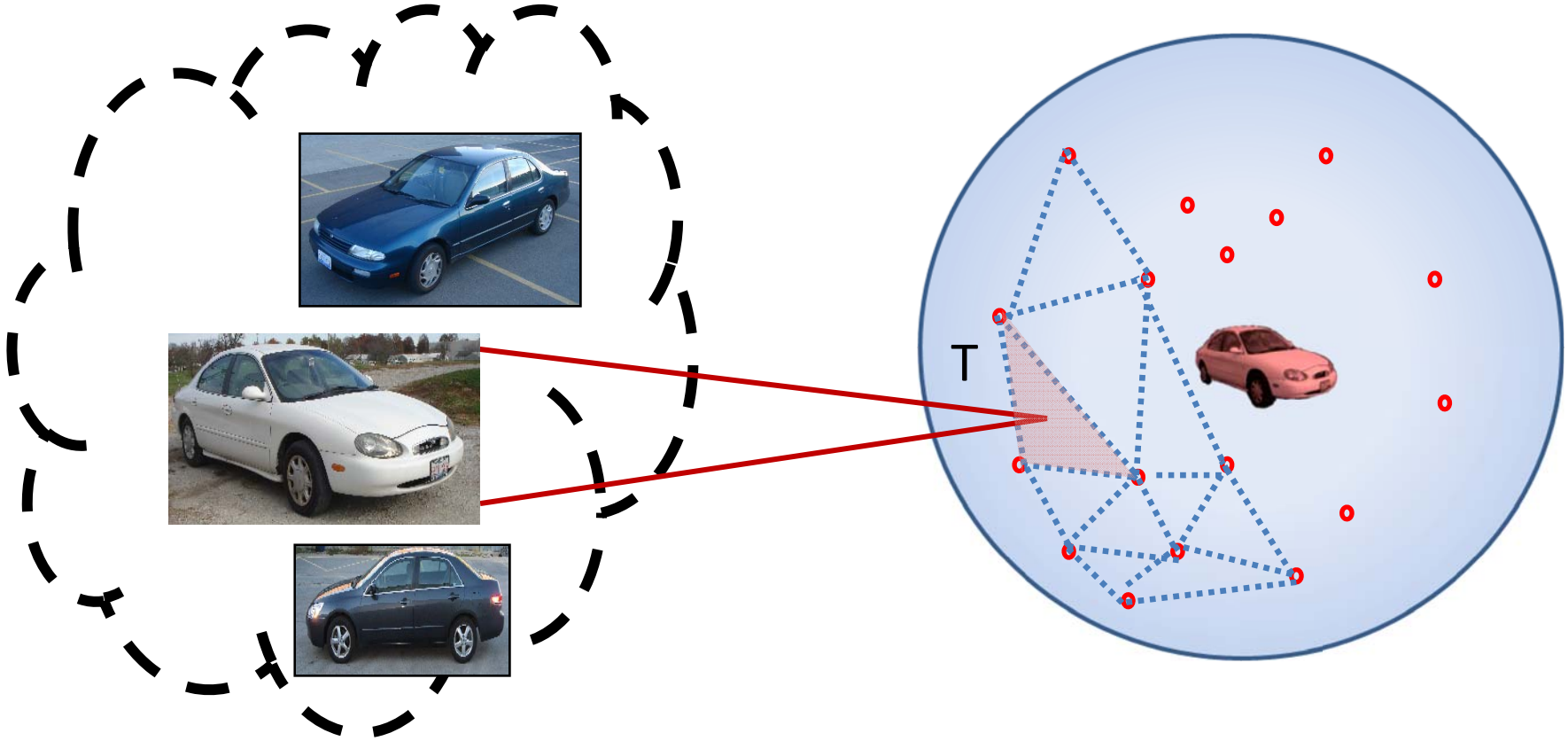
Encoded as a penalty term in variational *EM*

Incremental learning



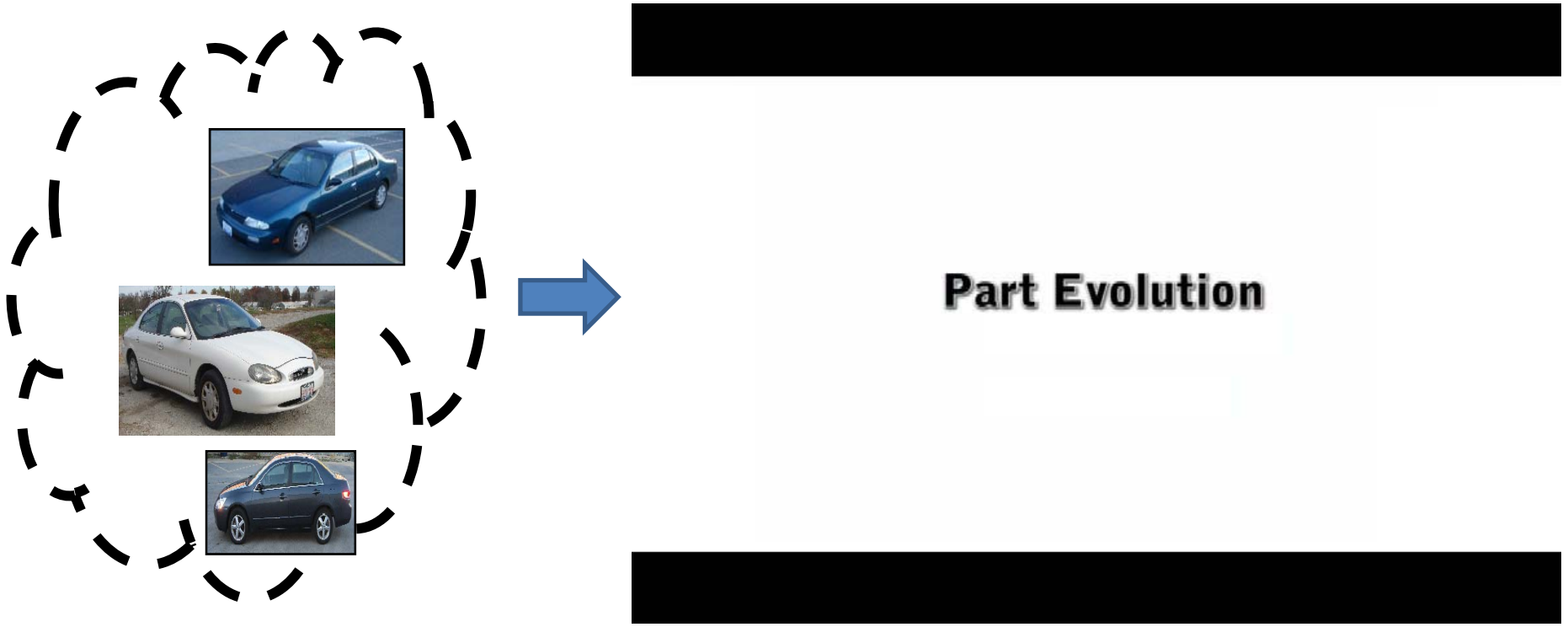
- Enable unorganized and on-line collection training images
- Increase efficiency in learning (no need large storage space)

Incremental learning

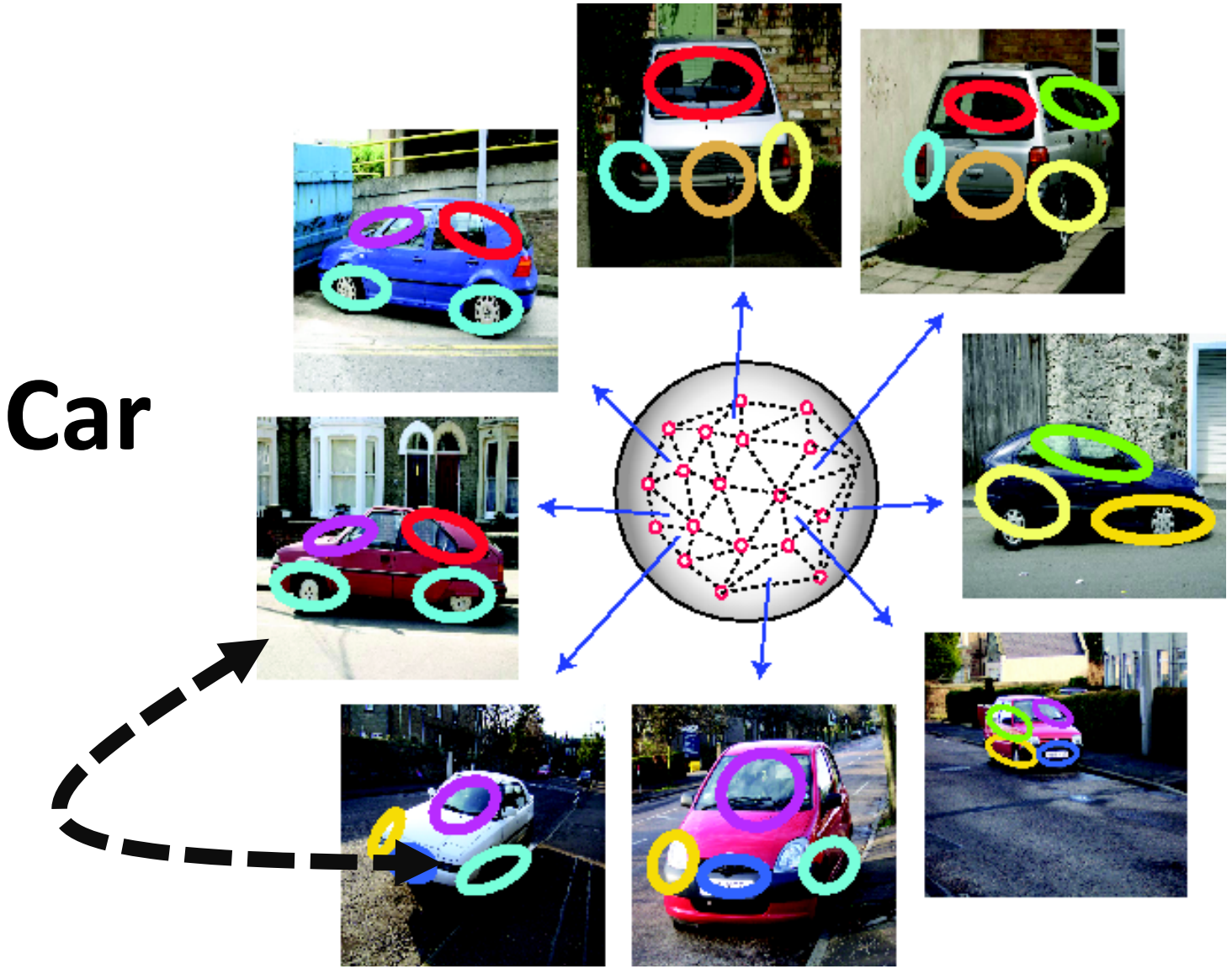


- Assign new training image to a triangle of the view sphere
- Evidence of training image is used to update model parameters
- Re-estimate sufficient statistics in a iterative fashion

Evolution of learnt parts

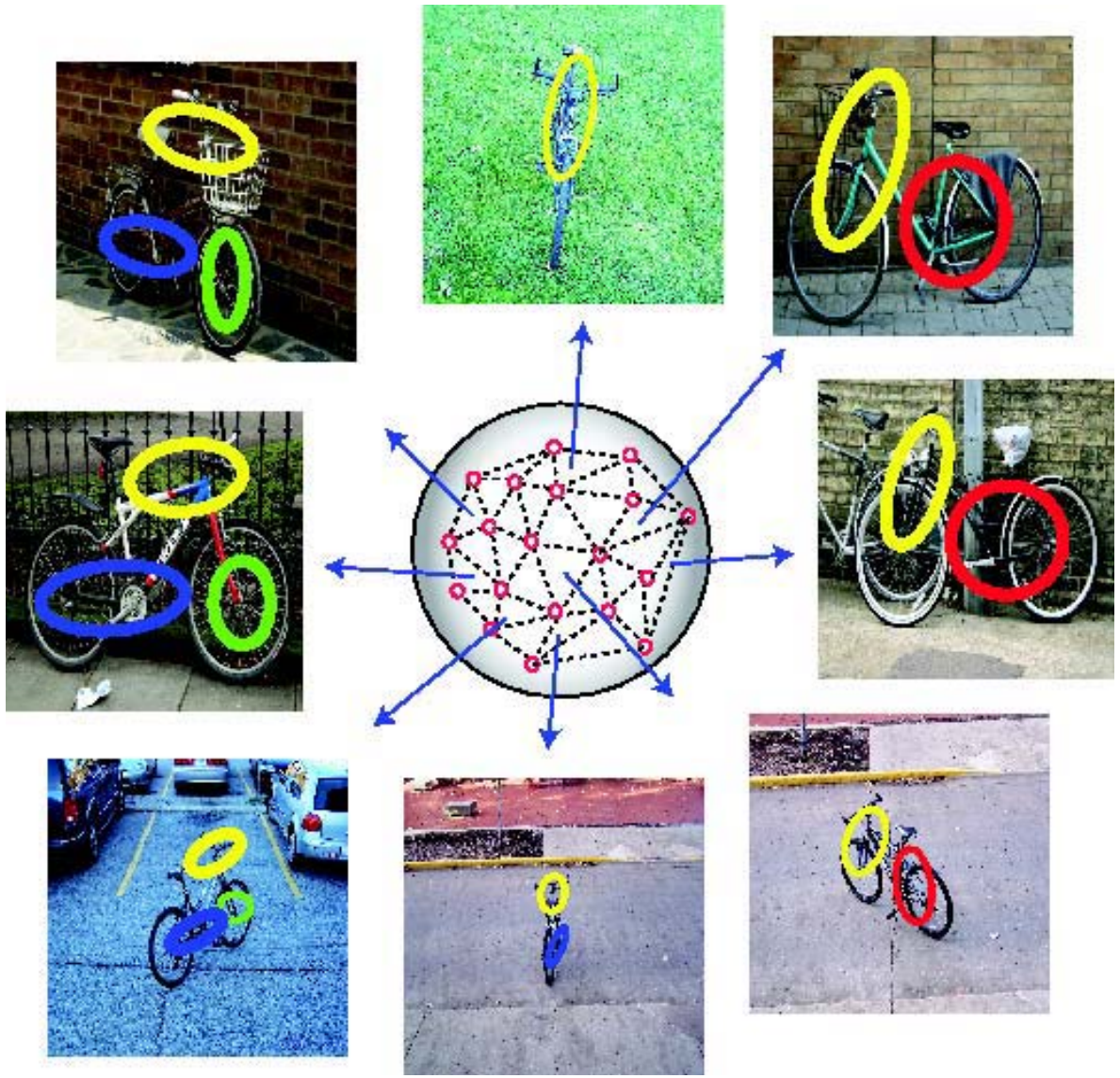


Examples of learnt part-based models



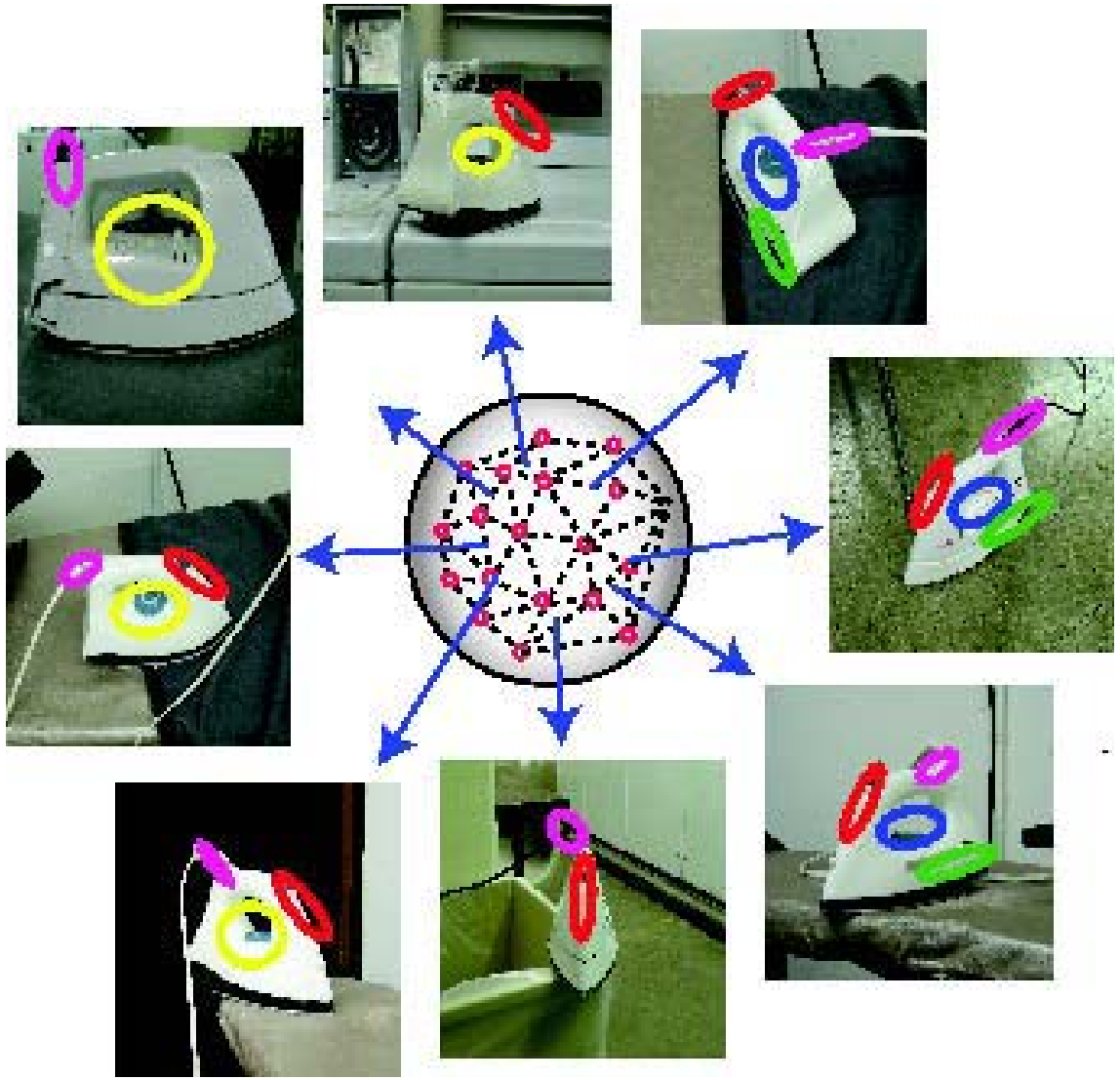
Examples of learnt part-based models

Bicycle



Examples of learnt part-based models

Travel iron

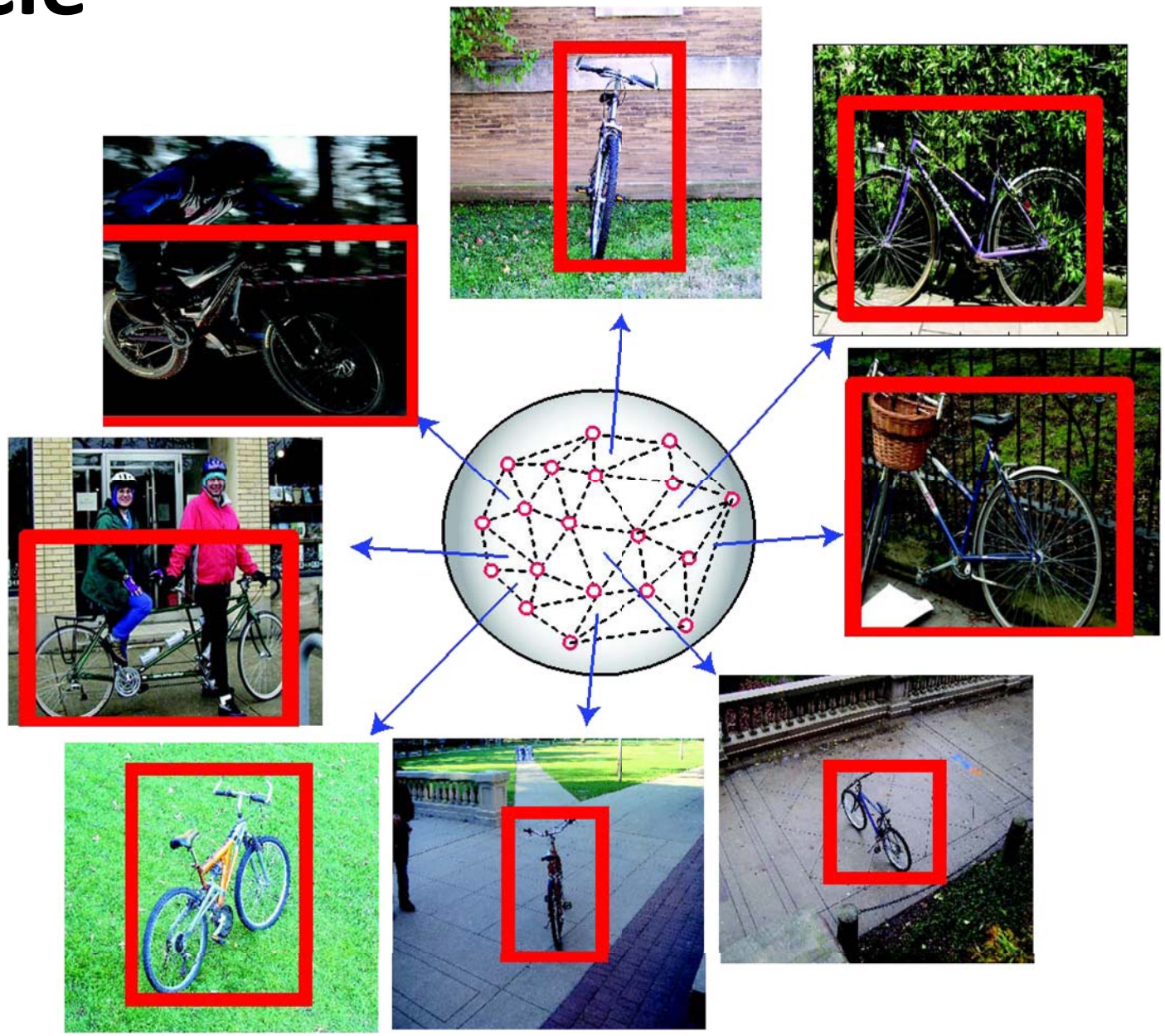


Detection and pose estimation

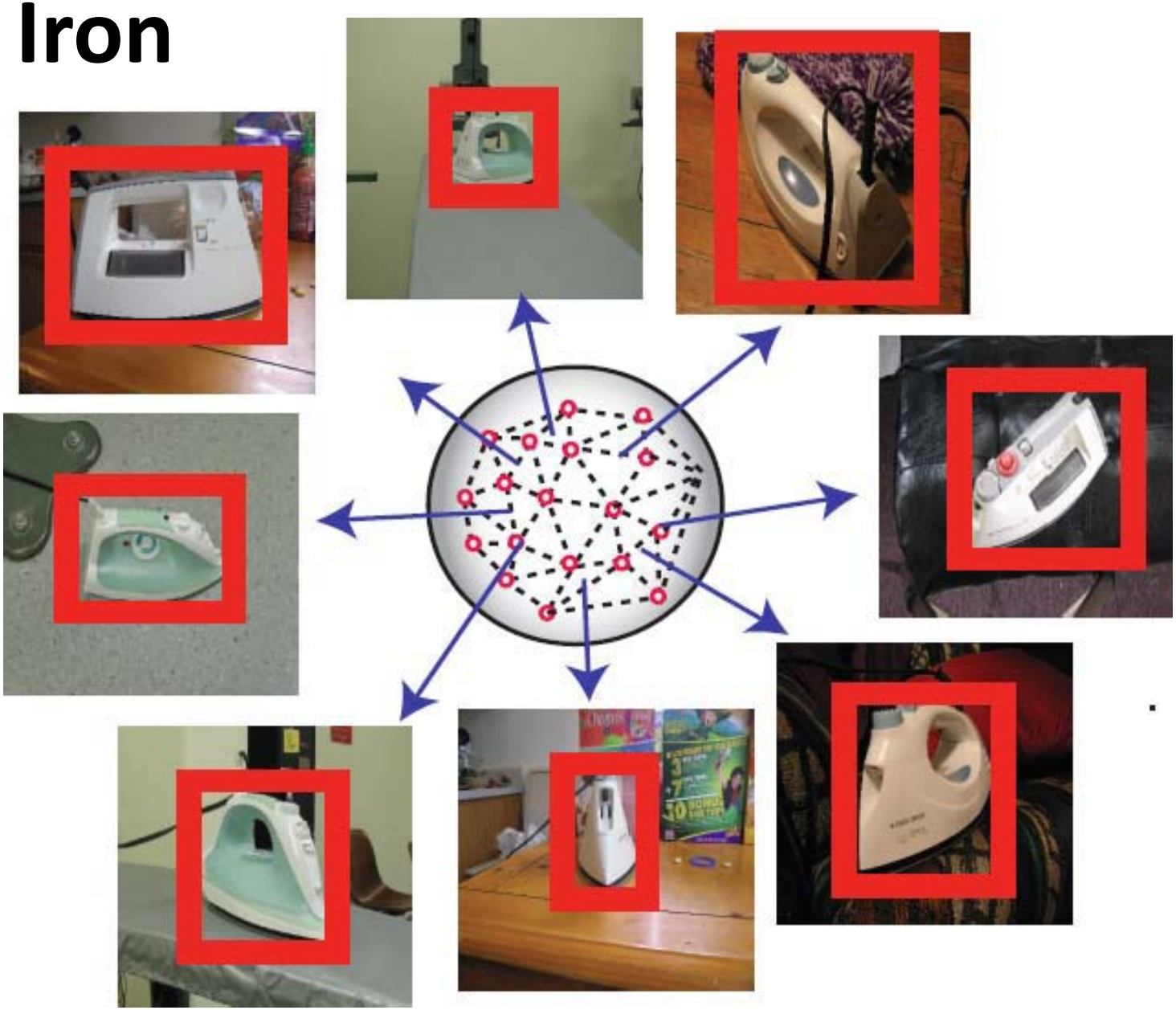
- **Detect** objects from any viewing angles
- Accurate **pose estimation**
- **Synthesize** (generate) object shape and appearance from novel views

- PASCAL 2006 dataset
- 3D Object Dataset

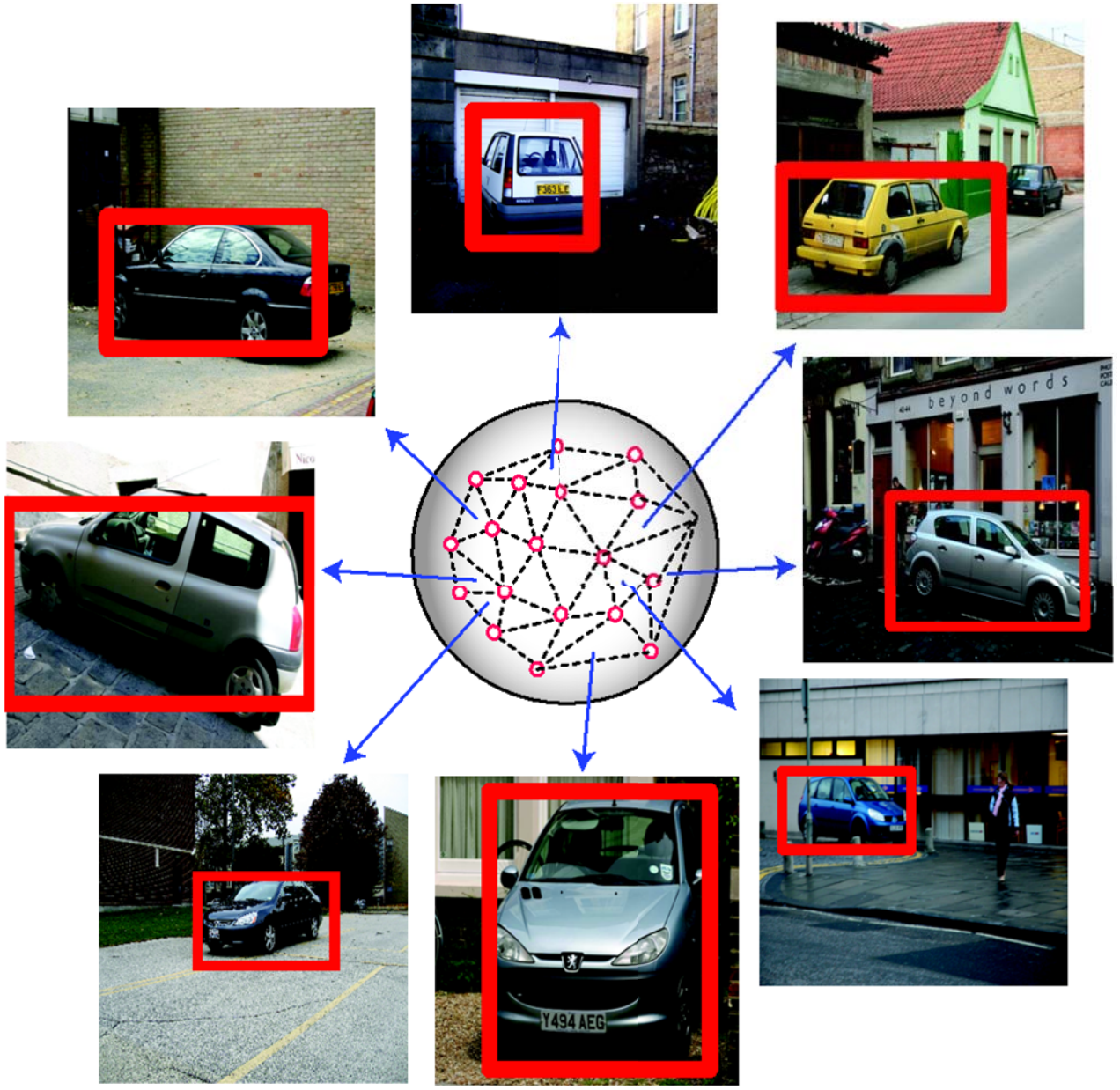
Bicycle



Travel Iron

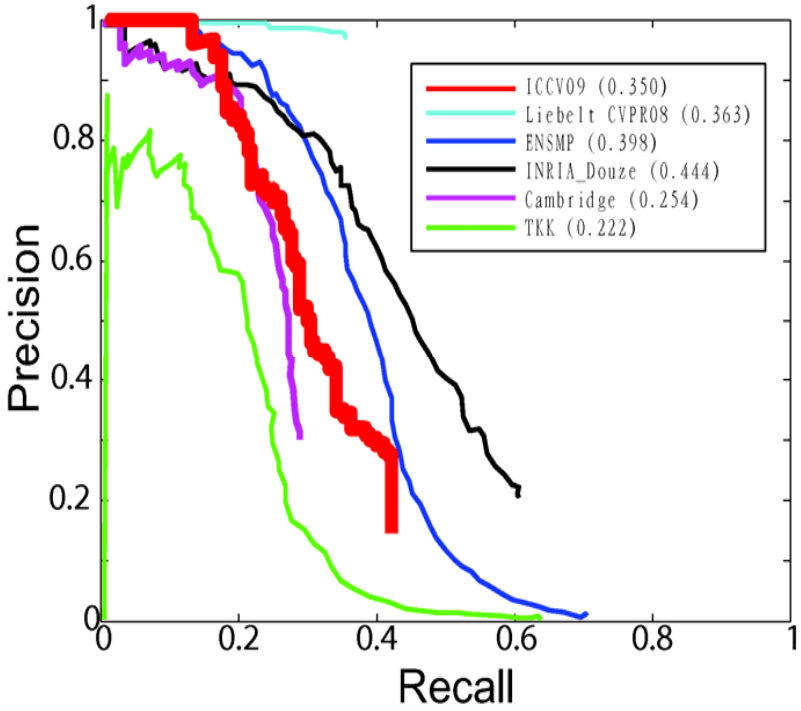


Car



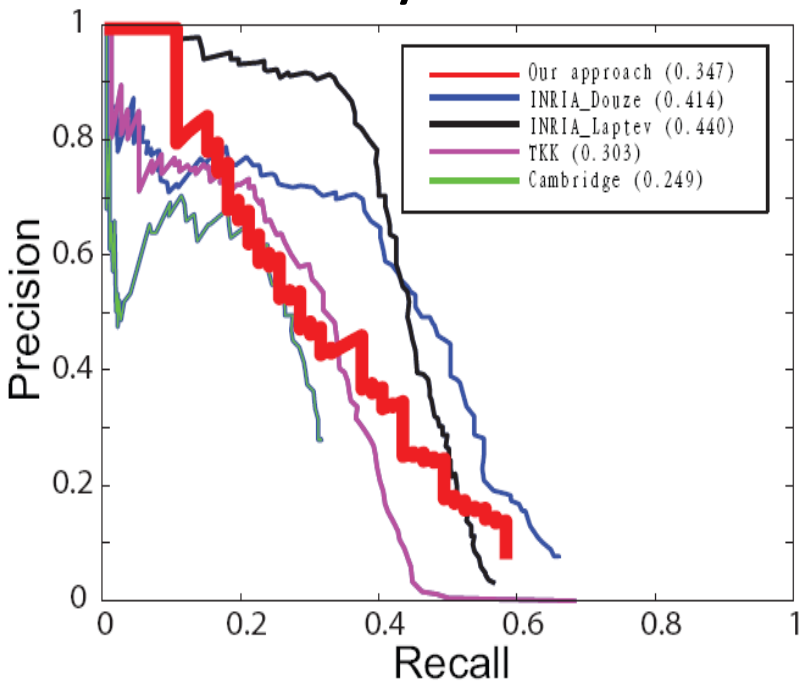
Detection - Pascal 2006 dataset

Car



0.35(average p)

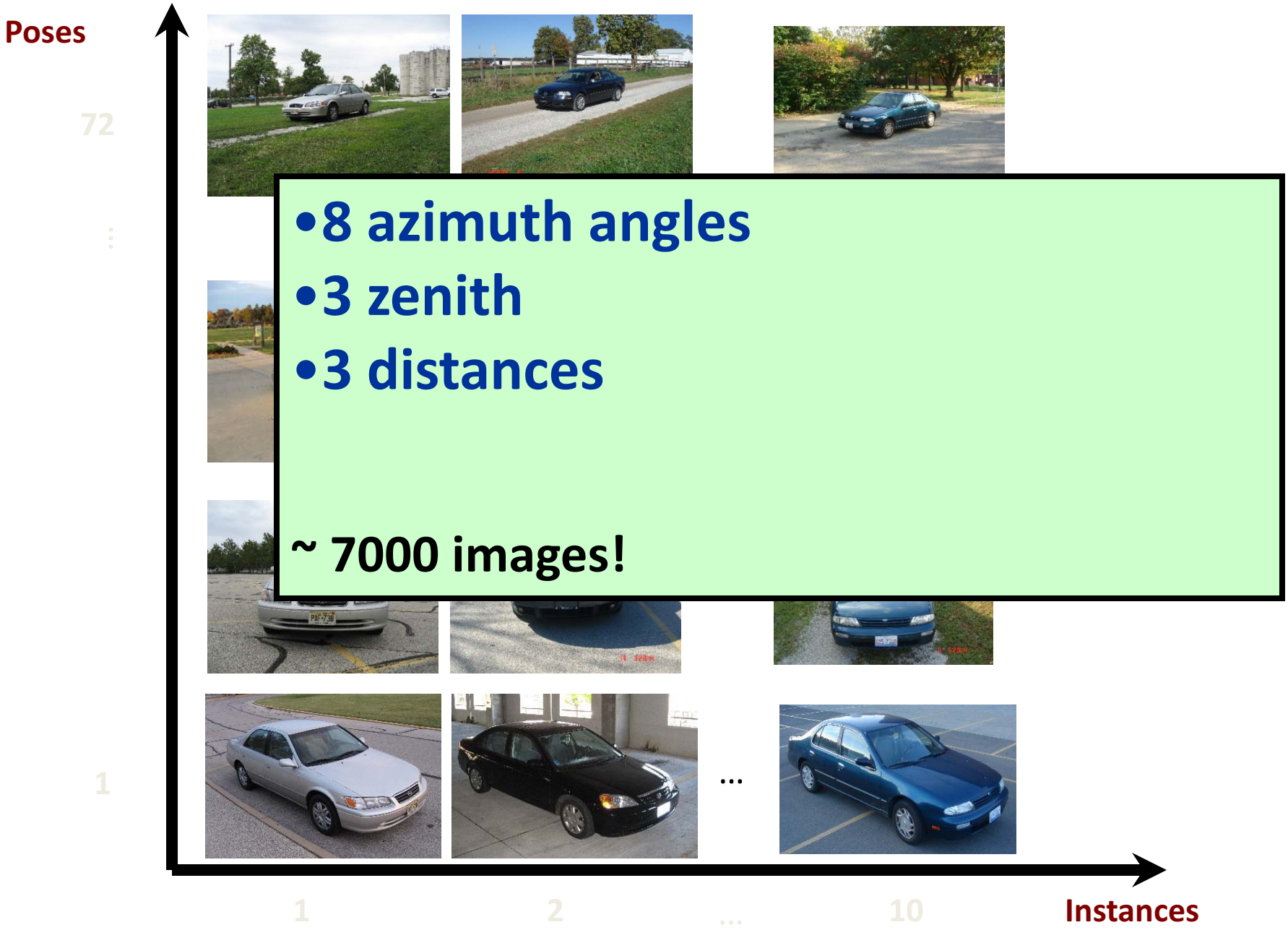
Bicycle



0.347(average p)

— Our model

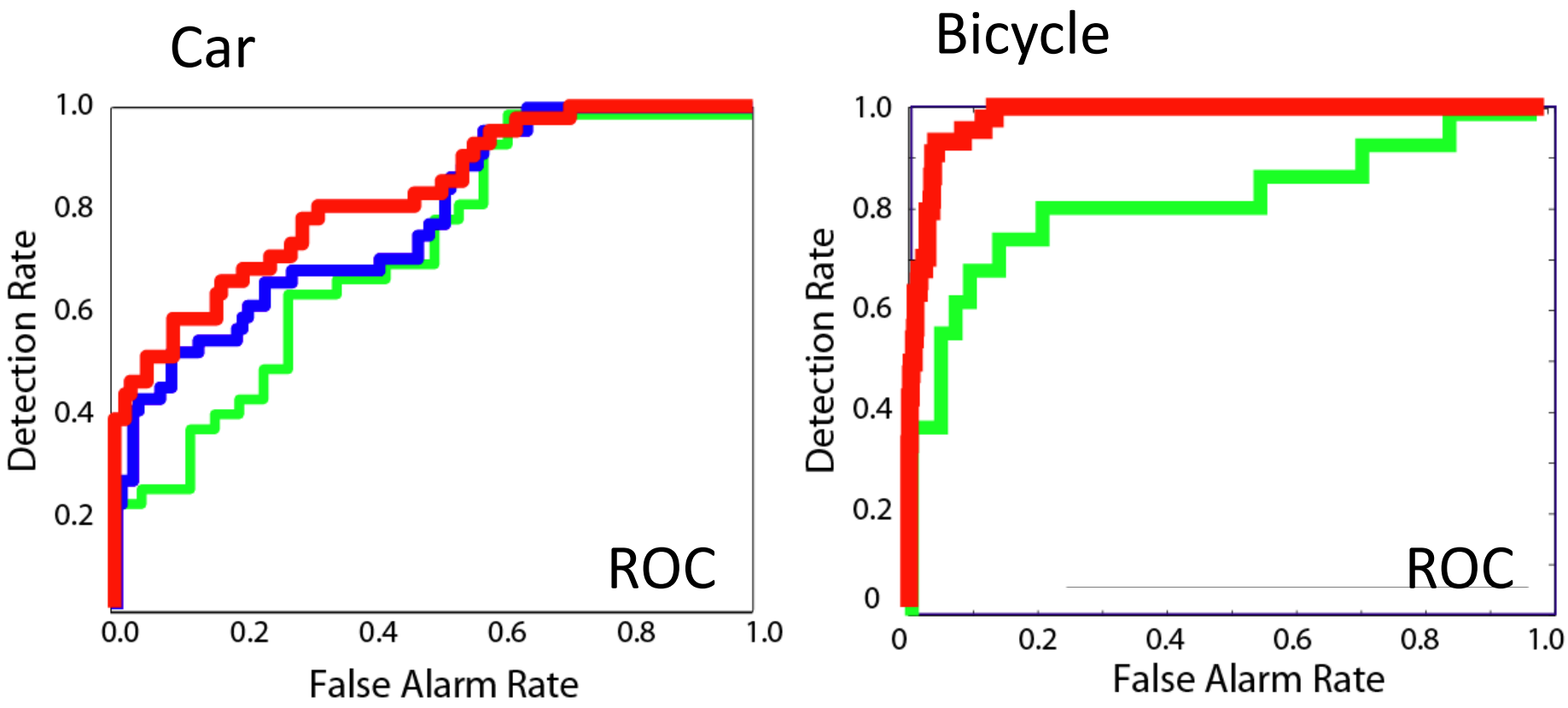
3D Object Dataset [Savarese et al 07]


Poses
72
1
1
2
...
10
Instances
•8 azimuth angles
•3 zenith
•3 distances
~ 7000 images!

Detection

3D object dataset

[Savarese et al 07]

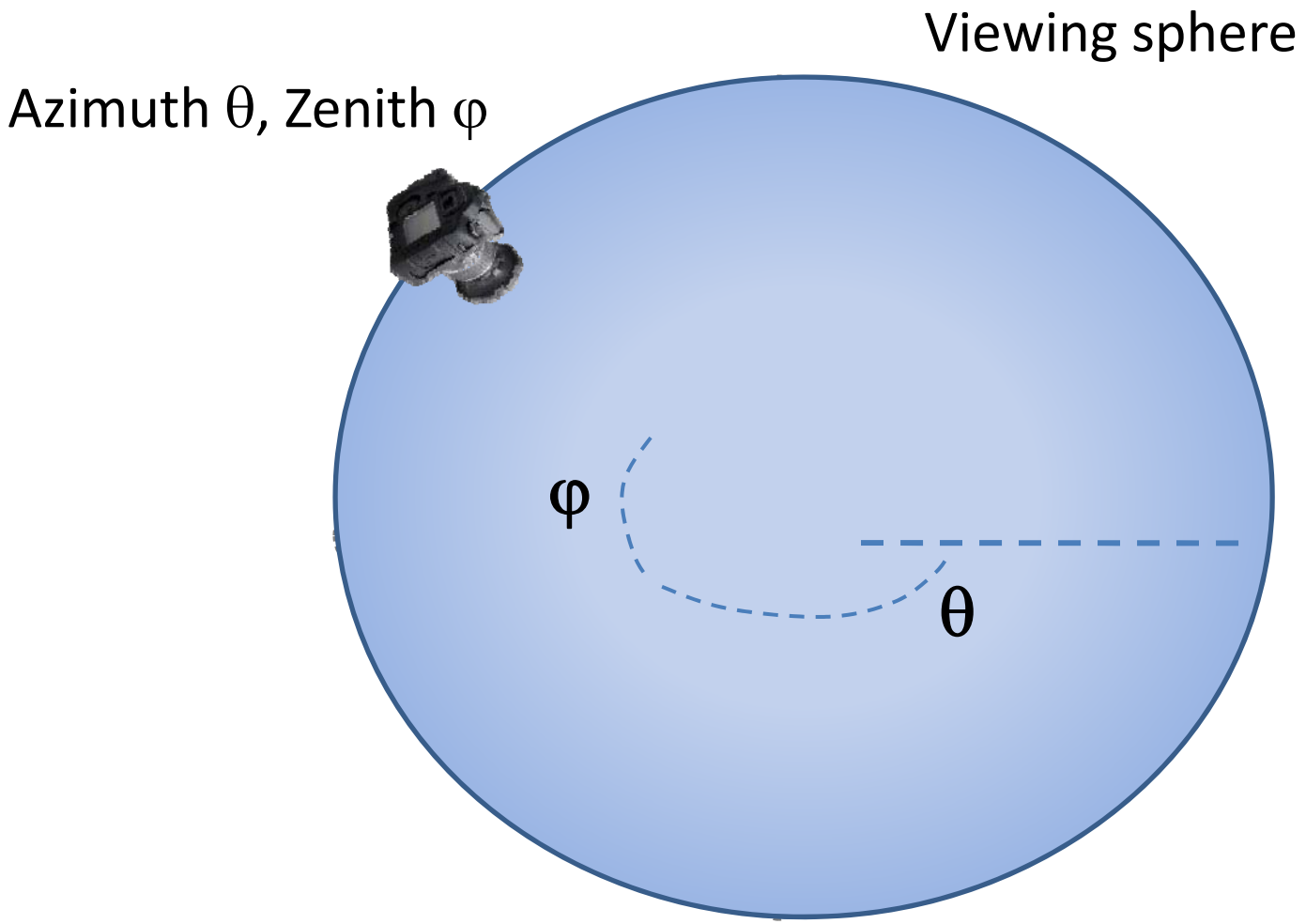


Sun et al, ICCV 2009

Su et al, CVPR

Savarese et al, ICCV '07

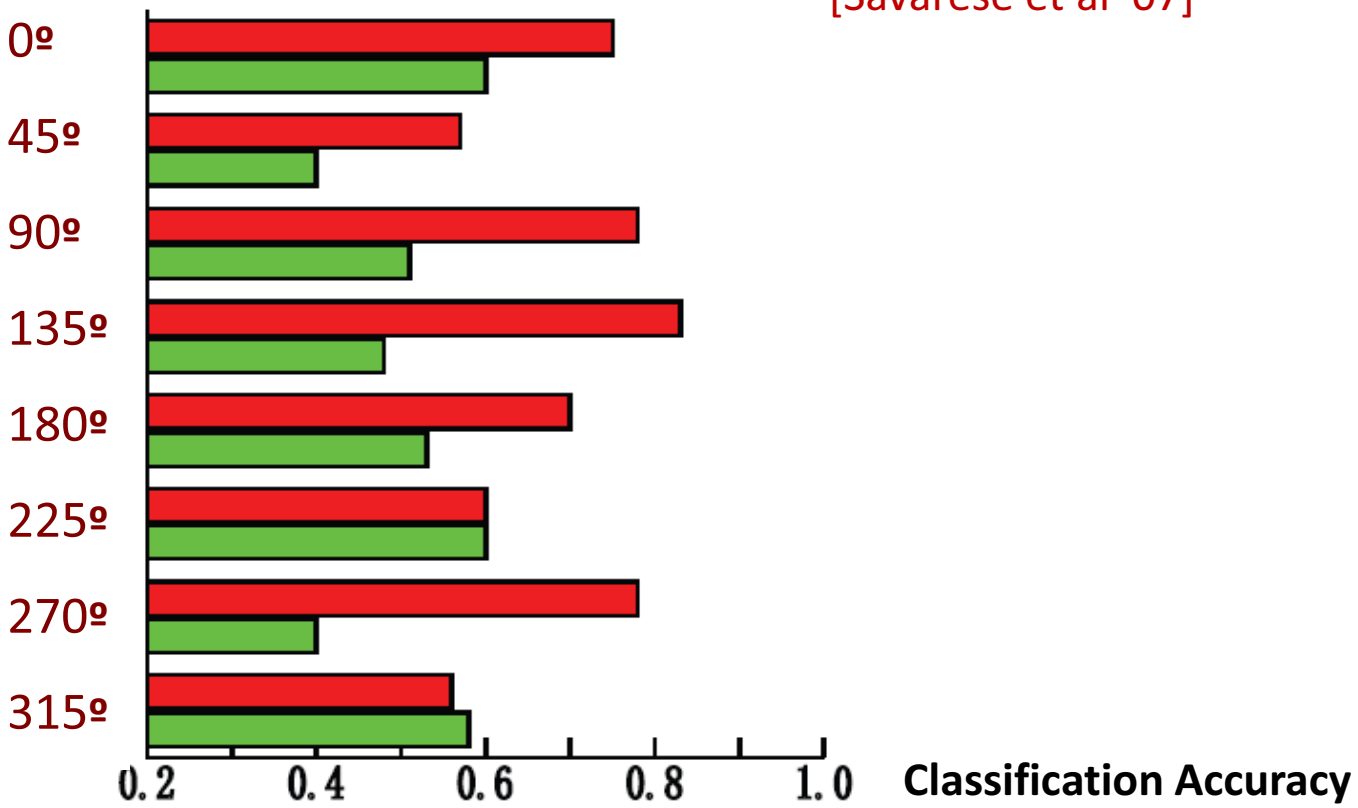
Viewpoint Classification



Viewpoint Classification

3D object dataset

[Savarese et al 07]



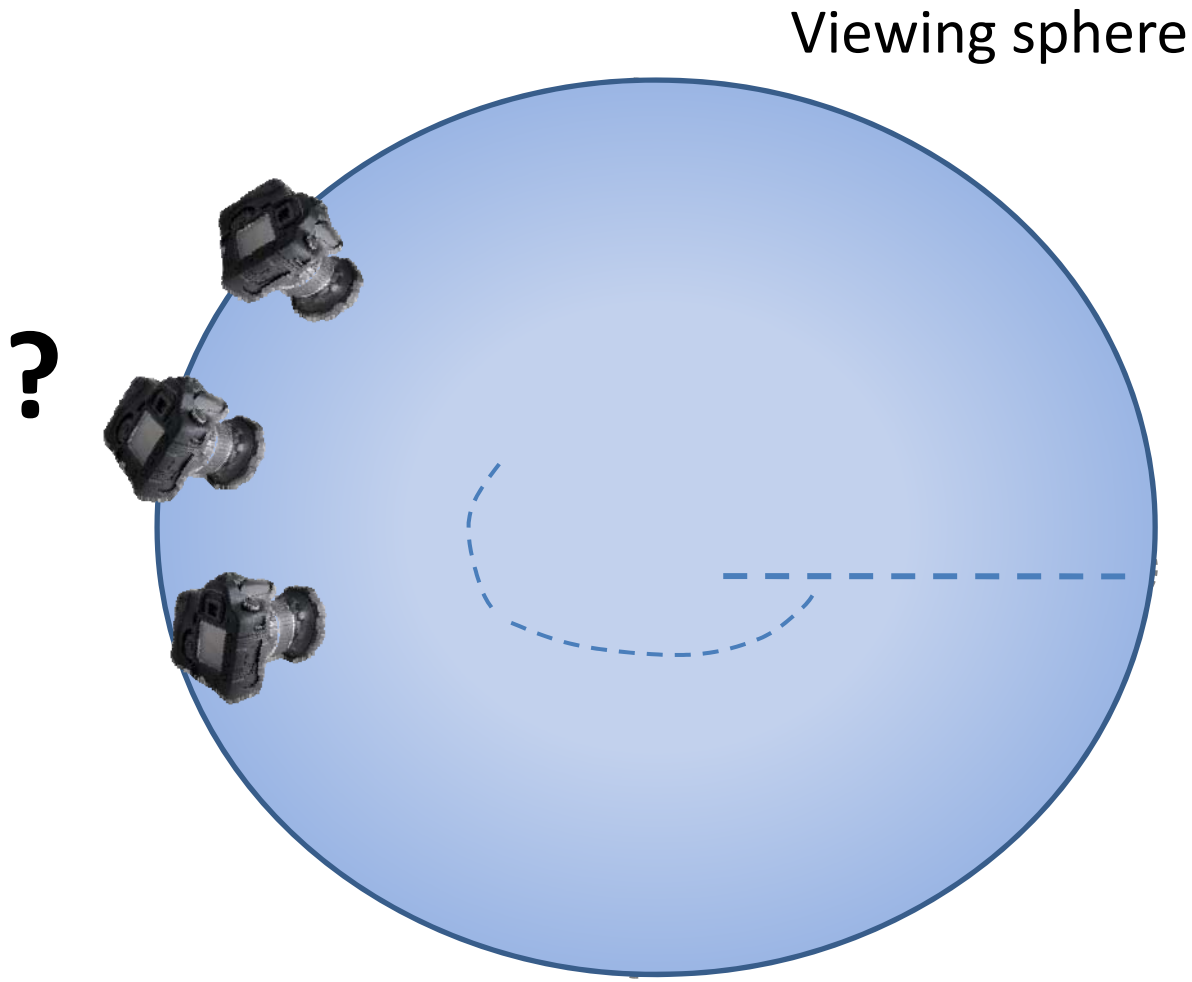
Sun et al ICCV 09

Savarese ICCV '07

Failure example

Category: car
Azimuth = 225º
Zenith = 30º

Predicting object appearance from novel views



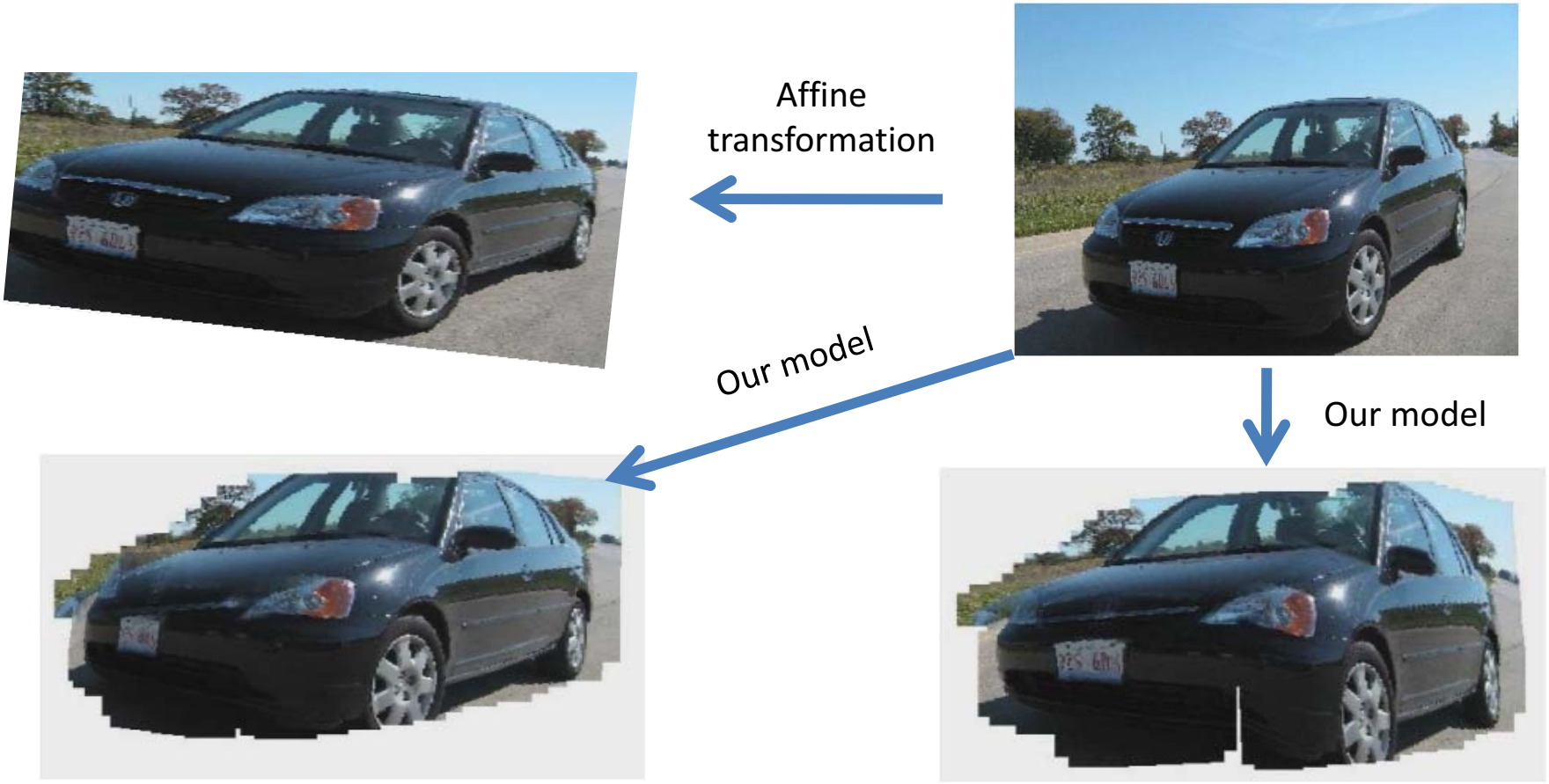
Predicting object appearance from novel views

[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

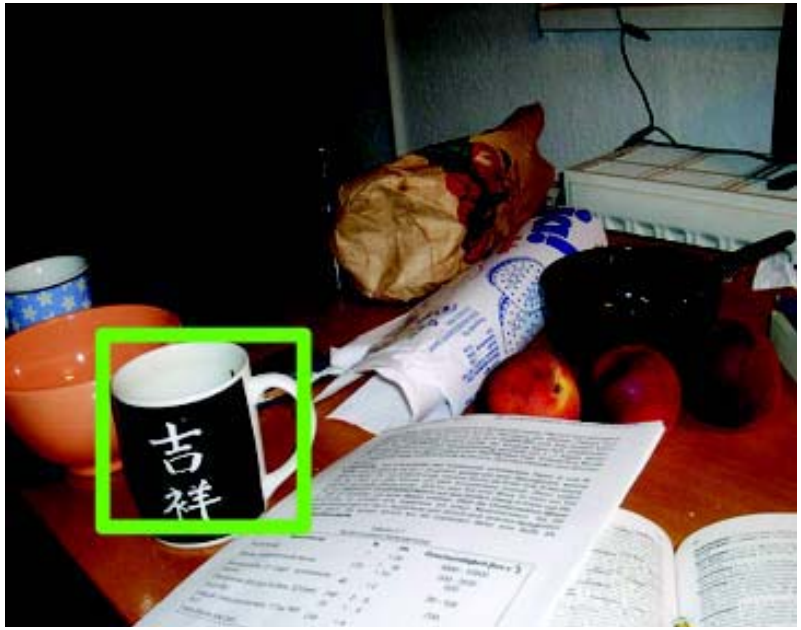


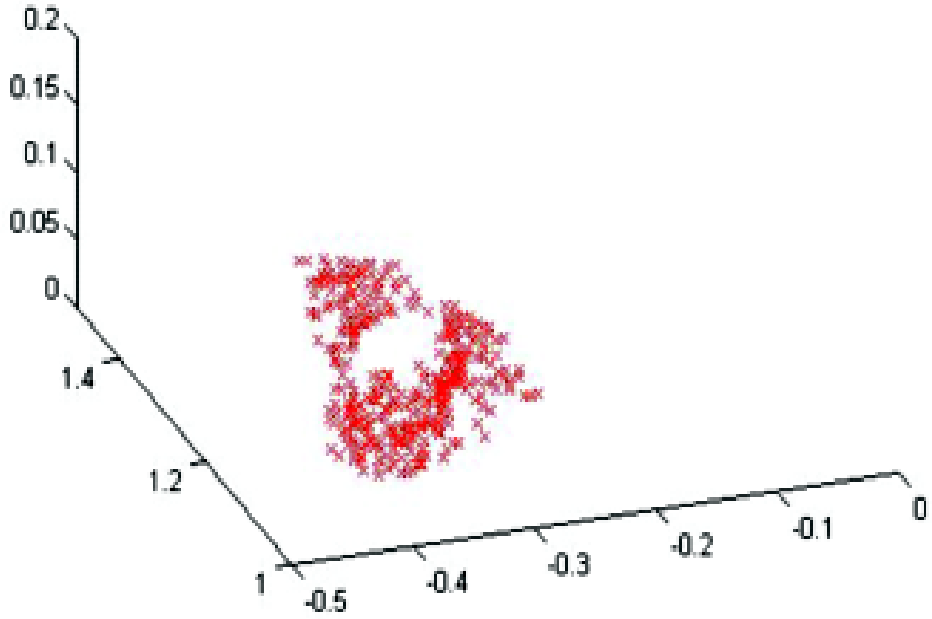
Outline

- Multi-view models for object categorization and 3D attribute estimation
  - 3D Pose
  - 3D shape
- Coherent object detection and scene layout estimation

Depth-Encoded Hough Voting for coherent object detection and shape recovery

M. Sun, B. Xu, G. Bradski, S. Savarese, ECCV 2010





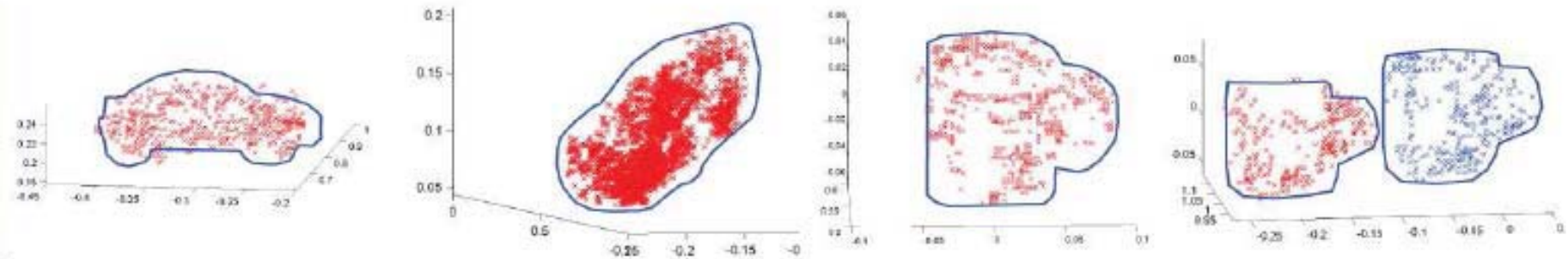
A new class of detectors:

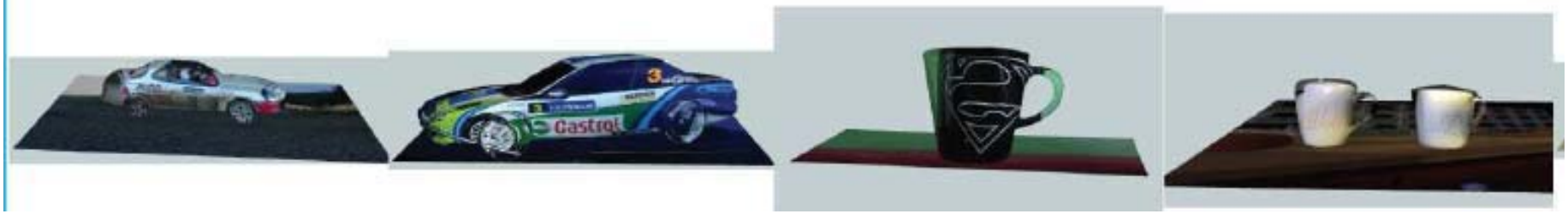
input:

- single image;
- rough 3D location (optional)

Output:

- localize object;
- 3D shape




Castrol

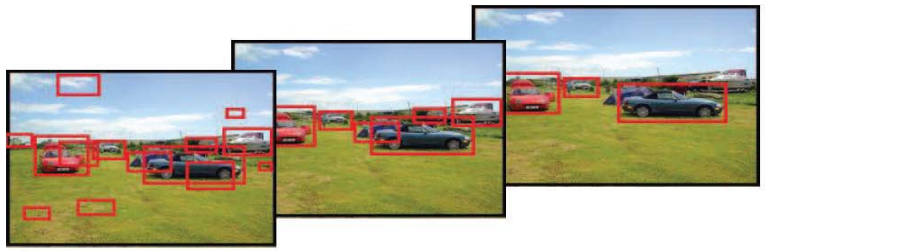
Outline

- Multi-view models for object categorization and 3D attribute estimation
  - 3D Pose
  - 3D shape

- Coherent object detection and scene layout estimation

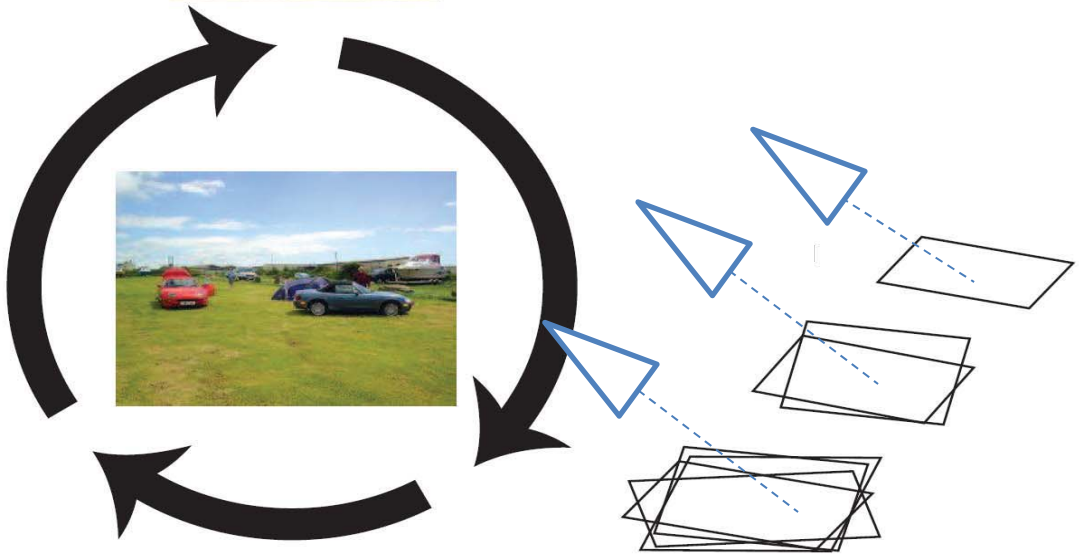
Toward coherent object detection and scene layout understanding

Y. Bao, M. Sun, S. Savarese, CVPR 2010



Min Sun

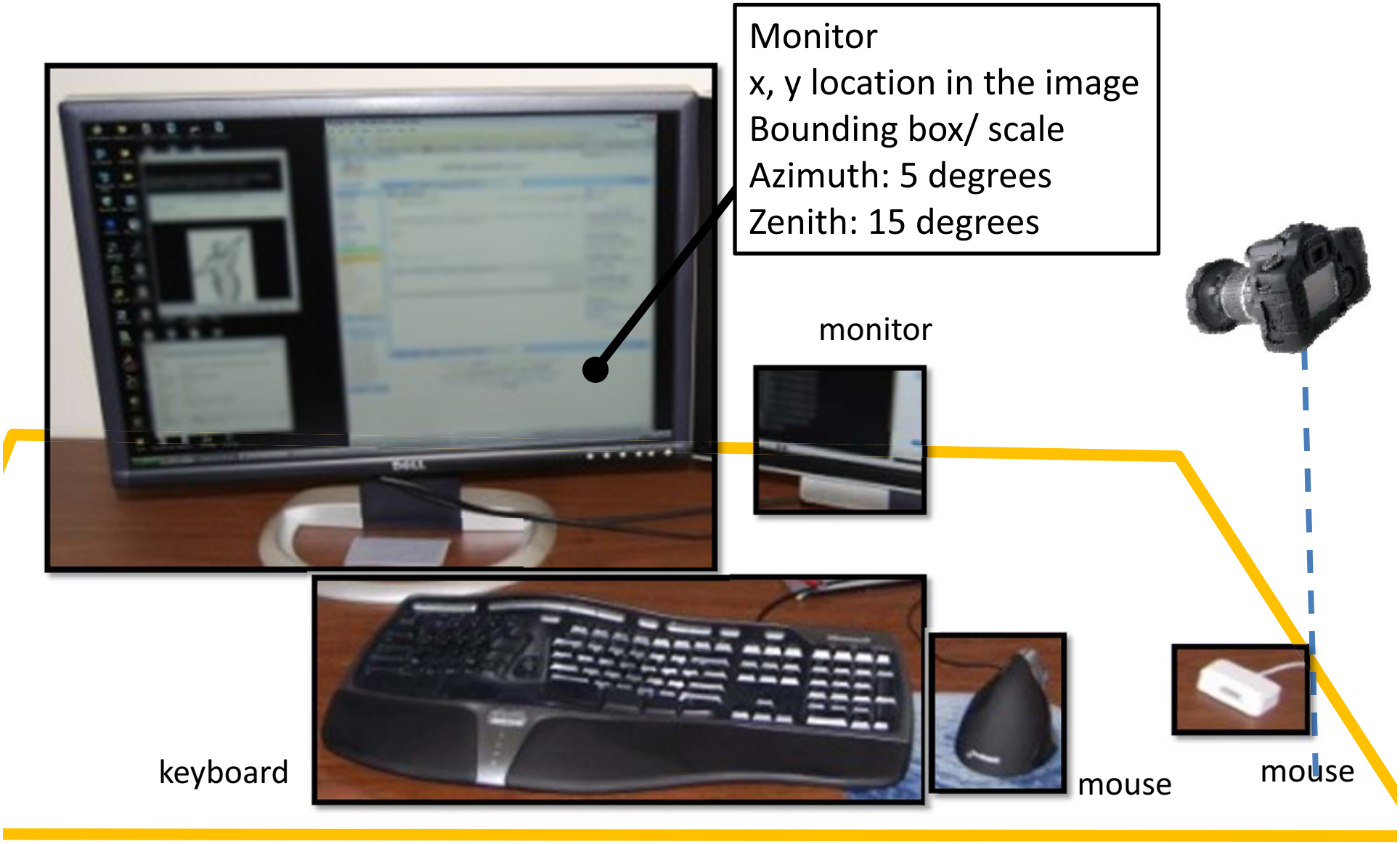
Ying-ze Bao



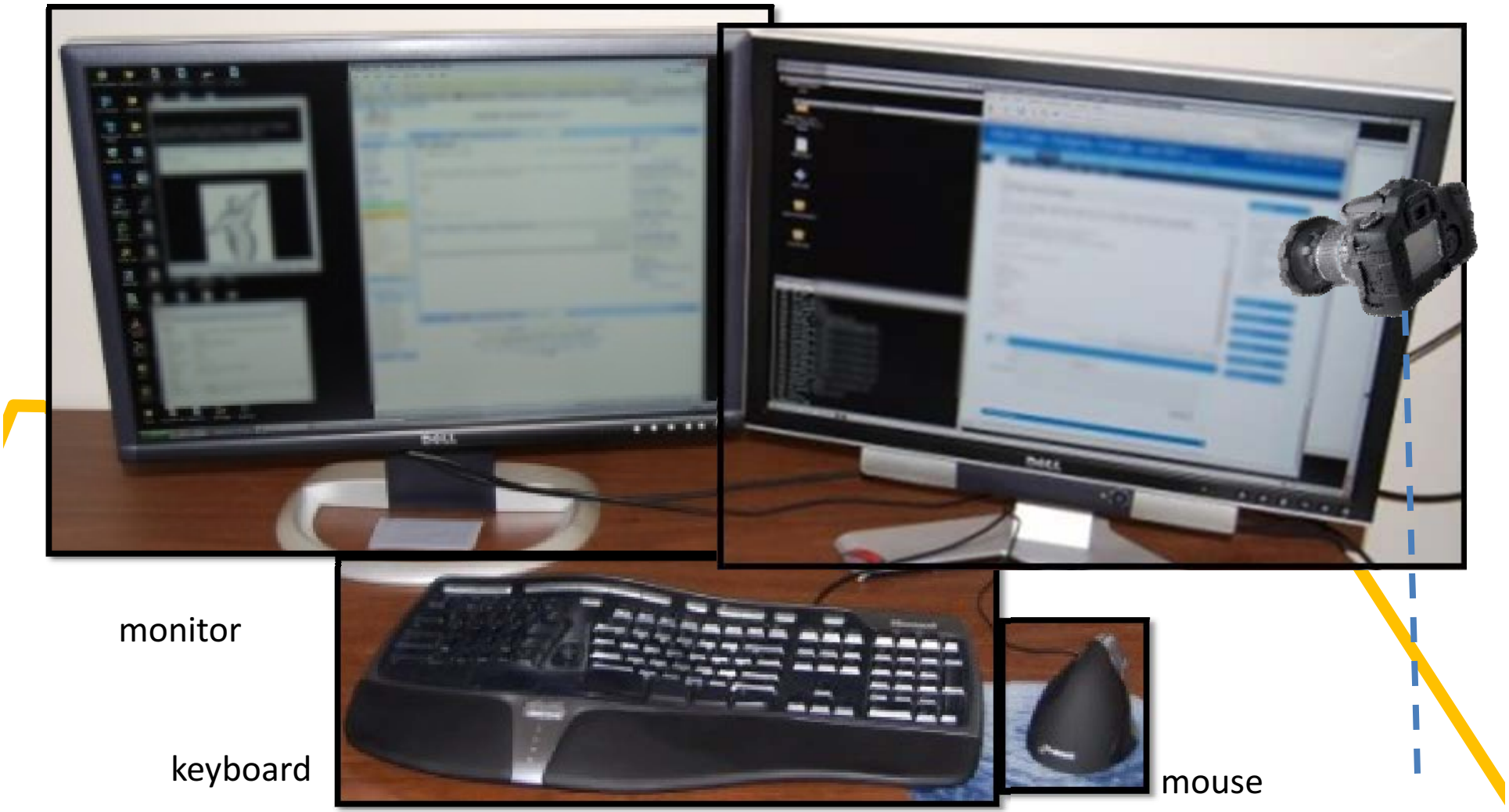
- Recognizing objects can help estimate the layout
- Estimating the layout can help reinforce the existence of the objects

Our Contributions

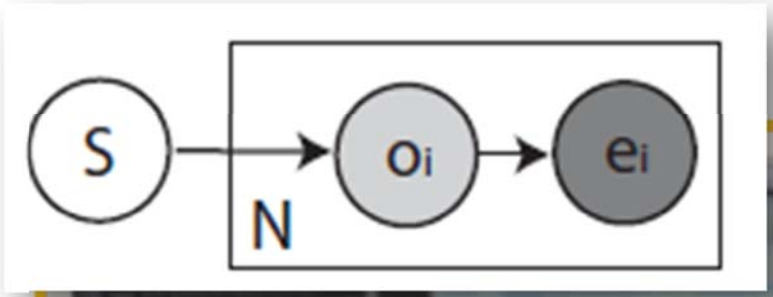
- Only use object detections to estimate 3D surface
  - No scene appearance model needed
    - Hoiem et al. 06-10
    - Gould et al. 09
- Estimate focal length from single image
- General camera pose model
  - camera tilt, camera in-planar rotation
    - Hoiem et al. 06-10
    - Gould et al. 09
    - Hedau et al. 09
  - No assumptions on camera height



1. Layout: most likely supporting surface 3D locations & orientations • camera pose & focal length• 3D objects location



2. Improve detection: remove false alarms, discover missed detections



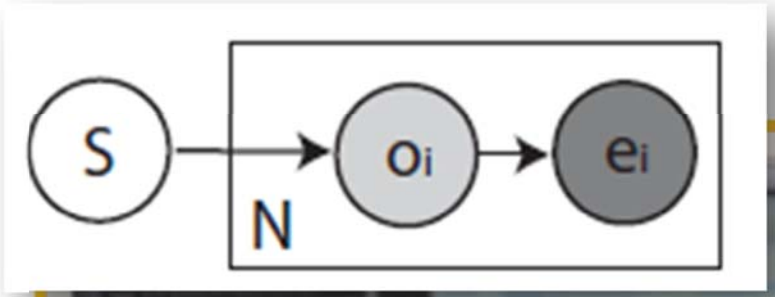
o_i = detection candidate

n

h

$$p(o, E, S) = p(S) \prod_{i=1}^{N} p(o_i|S)p(e_i|o_i)$$

- S = scene parameters (f, n_1… n_N, h_1… h_N)
- o_i = object detection candidates
- E = {e}: image evidence seen by the detector



o_i = detection candidate

t = detection is true positive

n

h

$$\{S, t\} = \arg\max_{S,\{t_i\}} \log p(o, E, S)$$

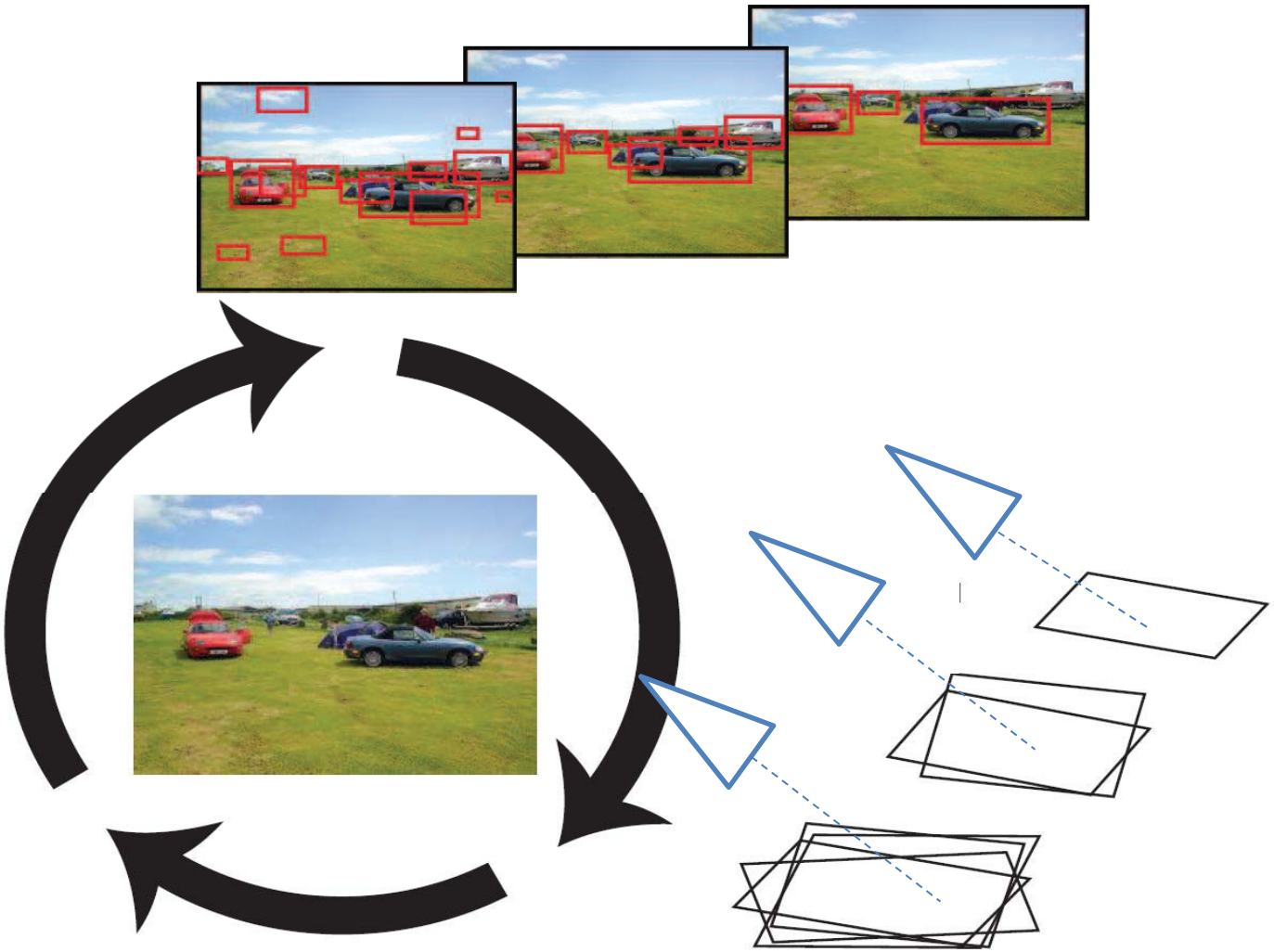
Layout

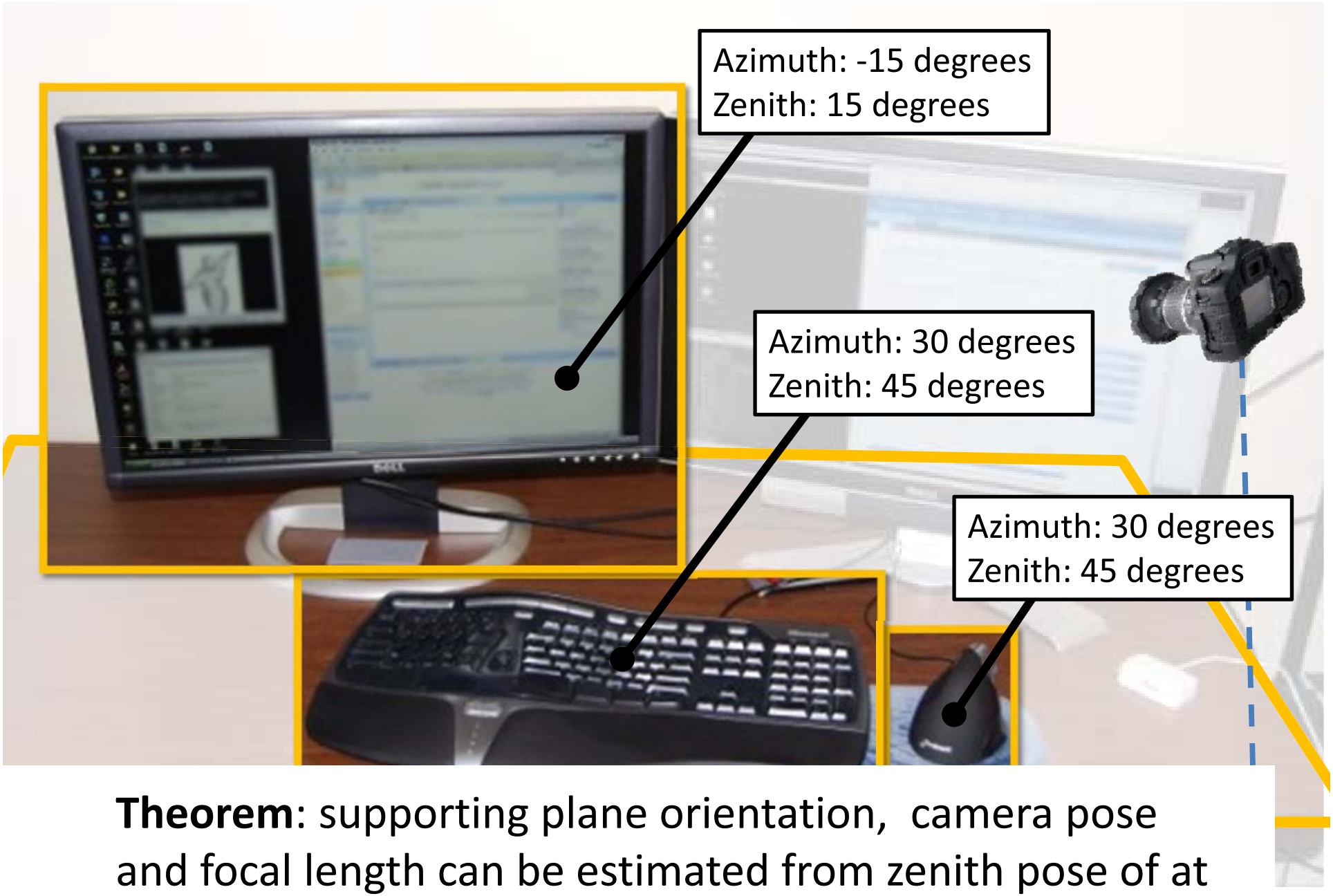
Object

GPU parallel programming:

640x480; 0.2 sec

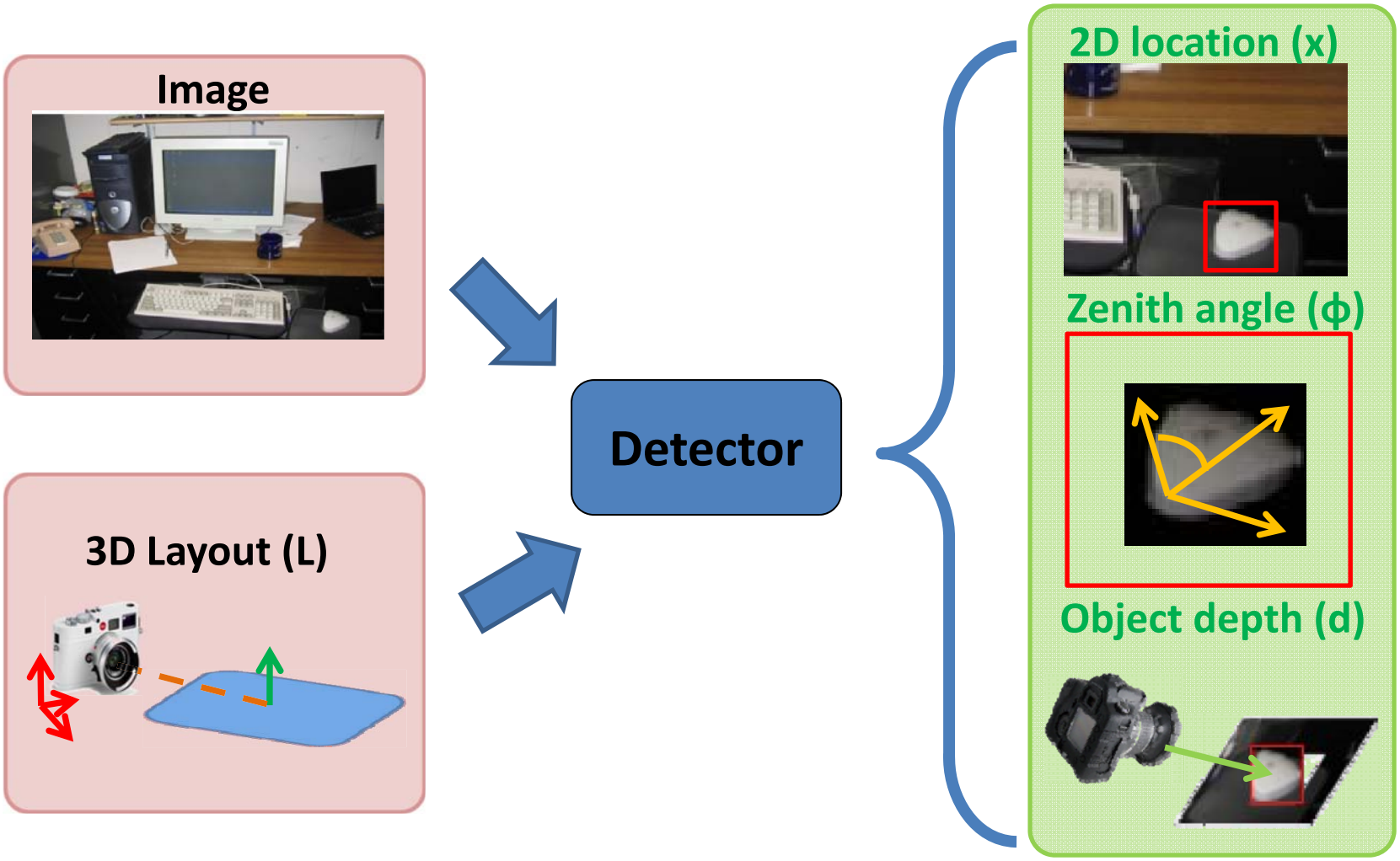
Joint inference process





Theorem: supporting plane orientation, camera pose and focal length can be estimated from zenith pose of at least 3 (non-collinear) objects

3D Encoded Detector



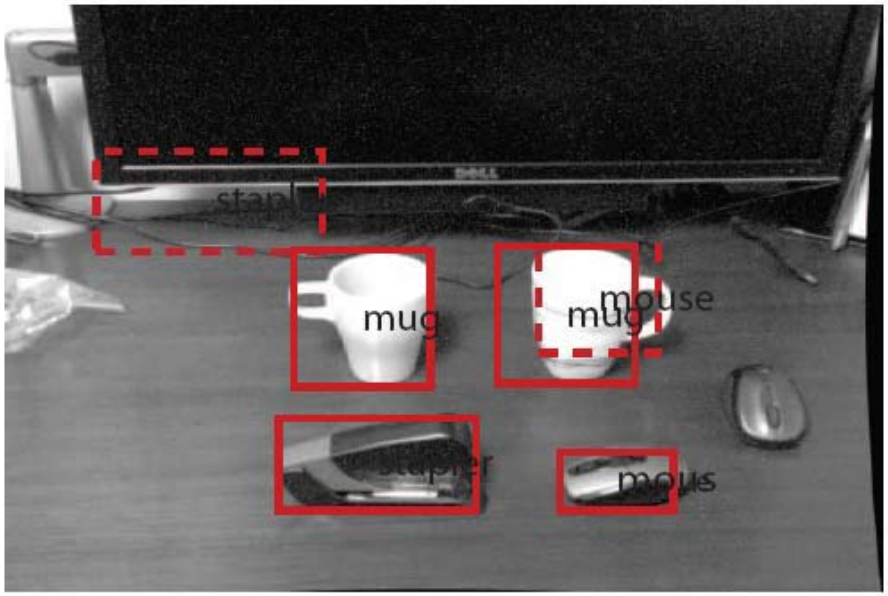
Modified from Sun et. al. ECCV'10

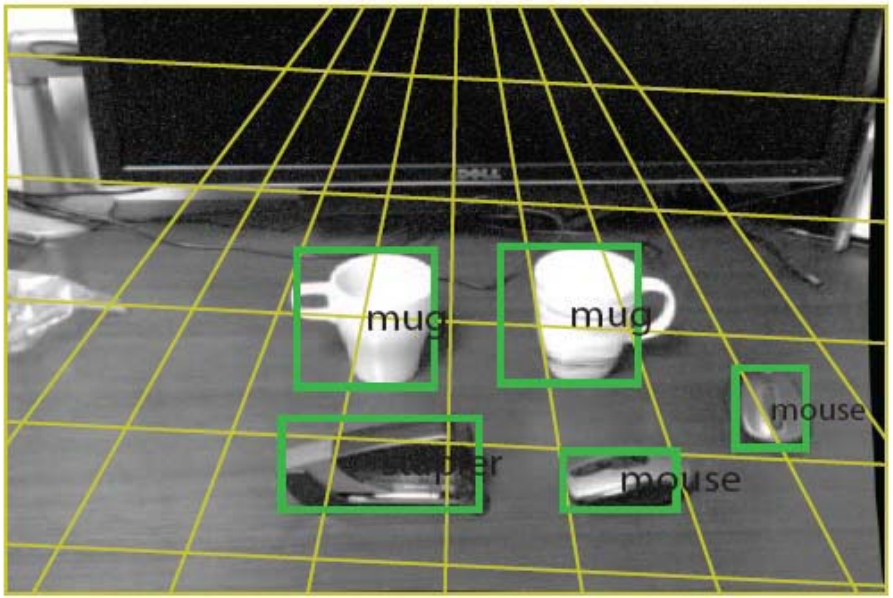
Experimental Results

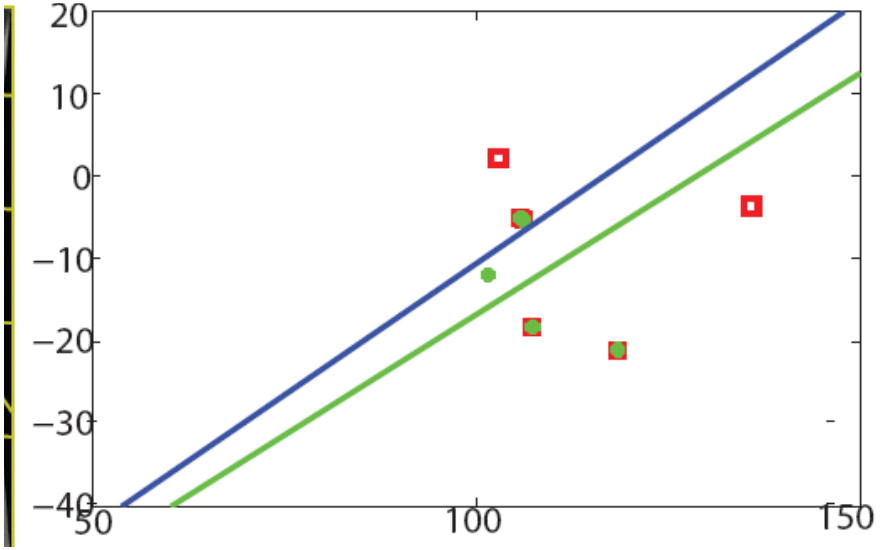
indoor dataset

Experimental Results

indoor dataset

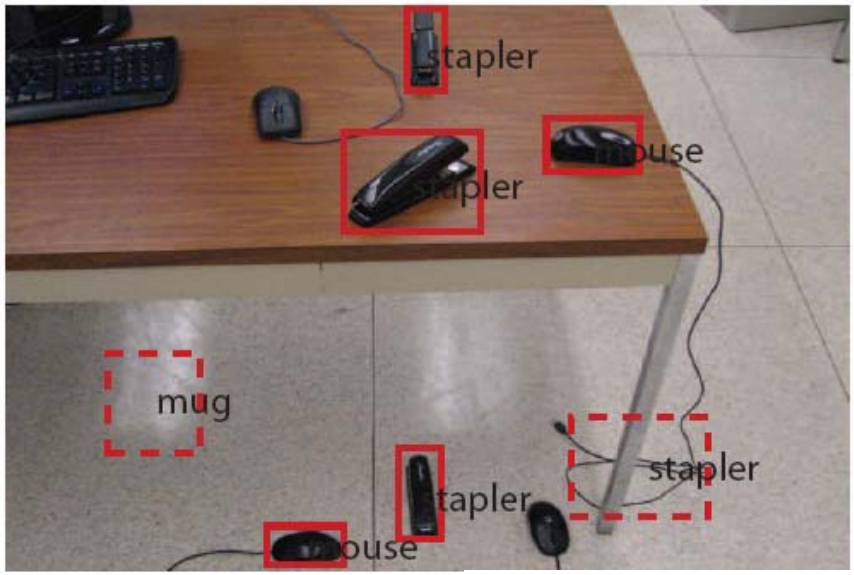


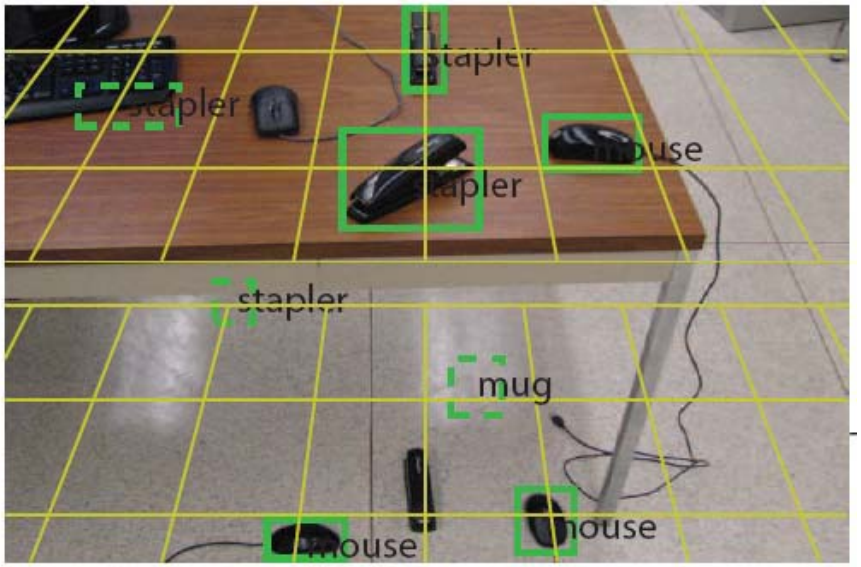


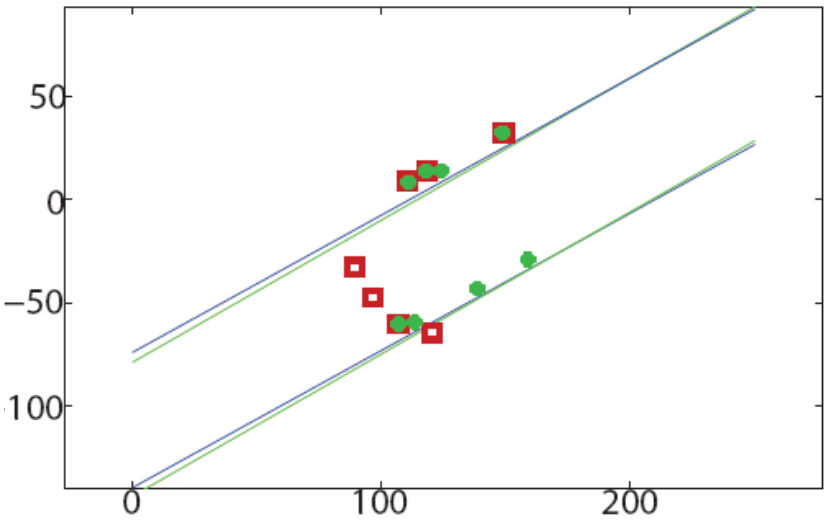


Experimental Results

indoor dataset

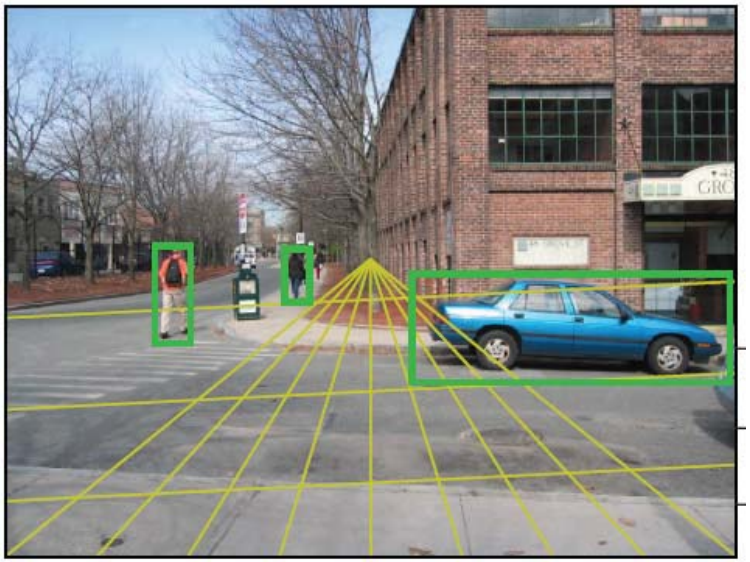


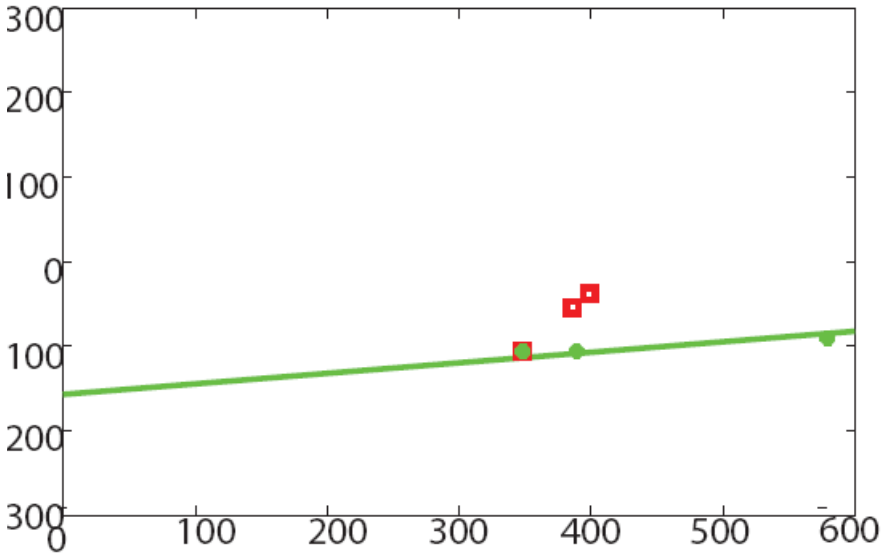




Experimental Results

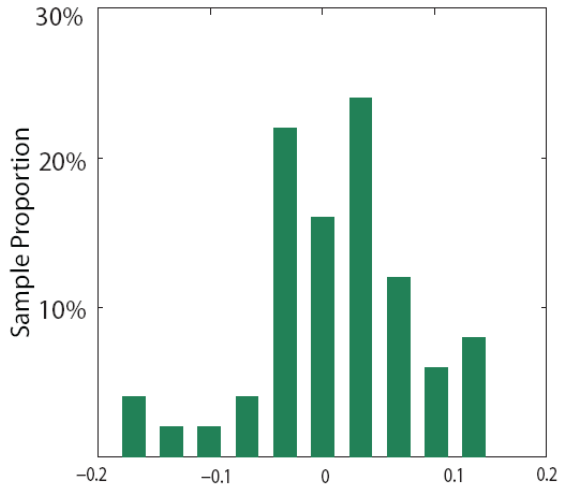
Labelme dataset



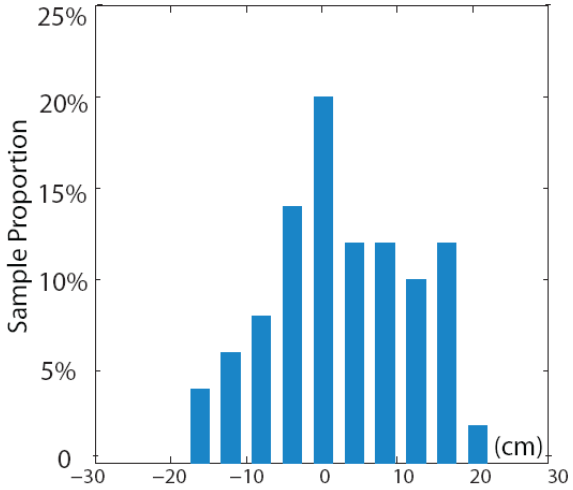


Experimental Results

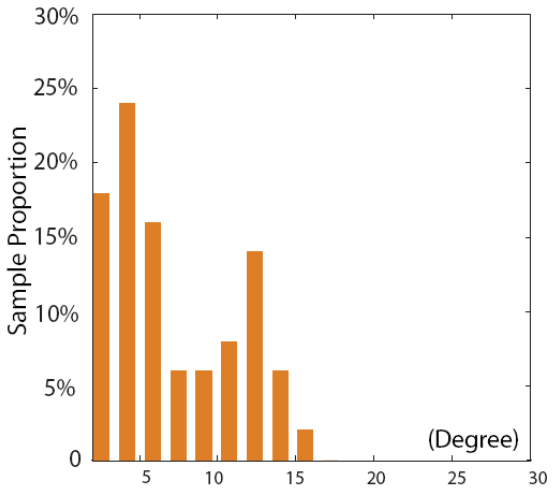
indoor dataset



(a) Focal Length Est. Error



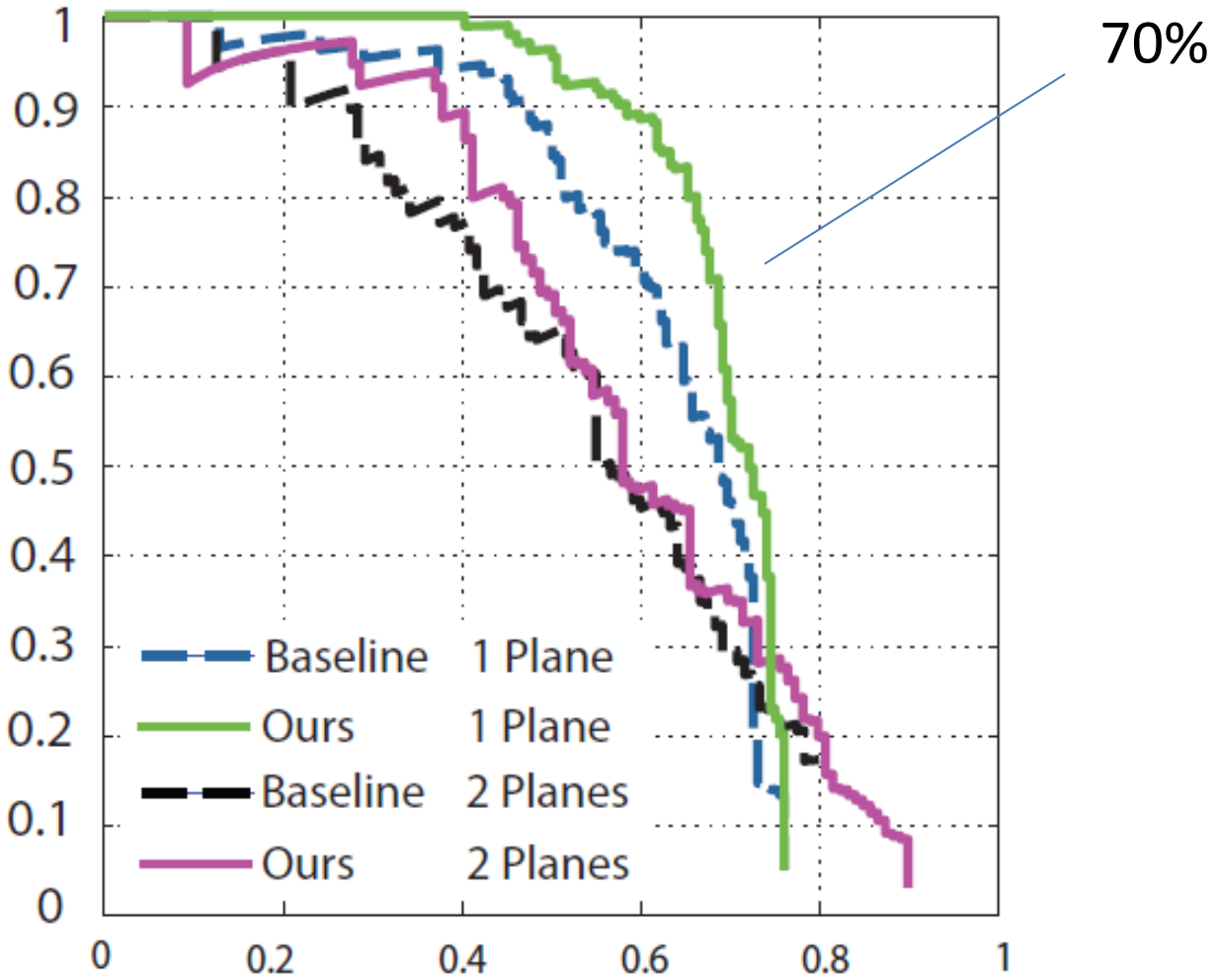
(b) Height Est. Error



(c) Plane Normal Est. Error

Experimental Results

indoor dataset



Outline

- Multi-view models for object categorization and 3D attribute estimation

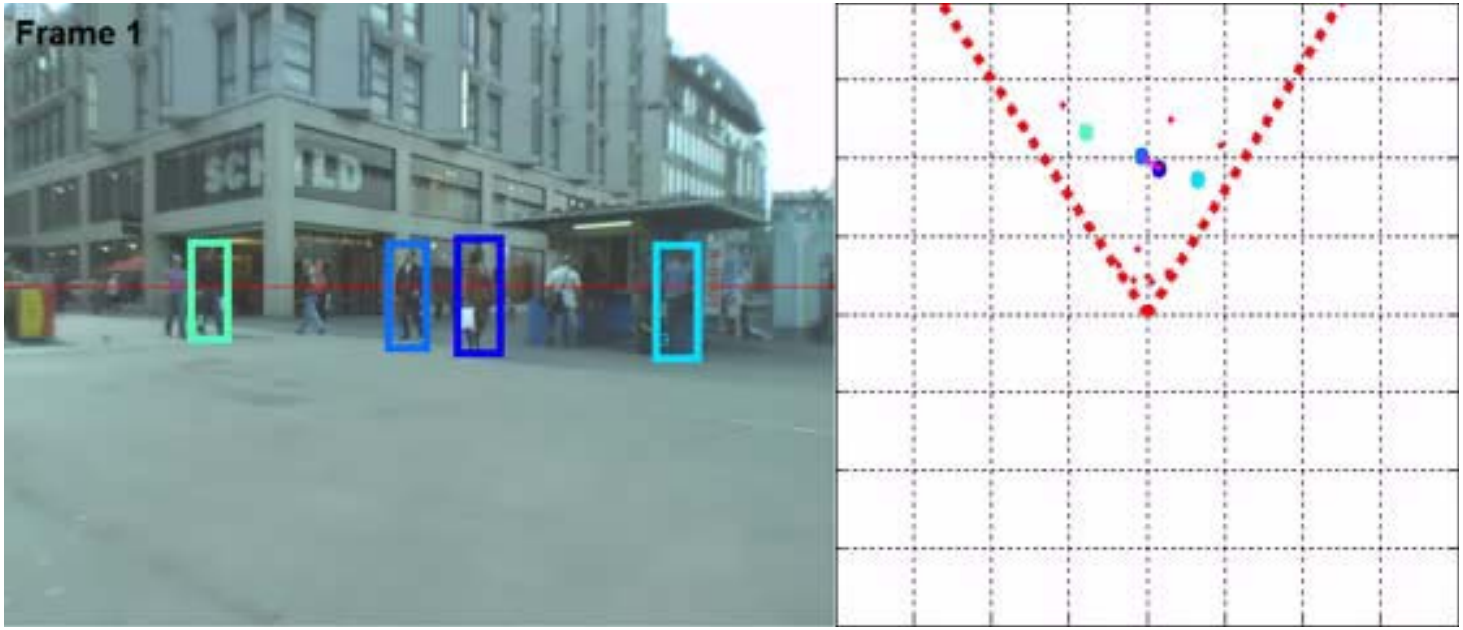
- Coherent object detection and scene layout estimation
  - From single image
  - From videos

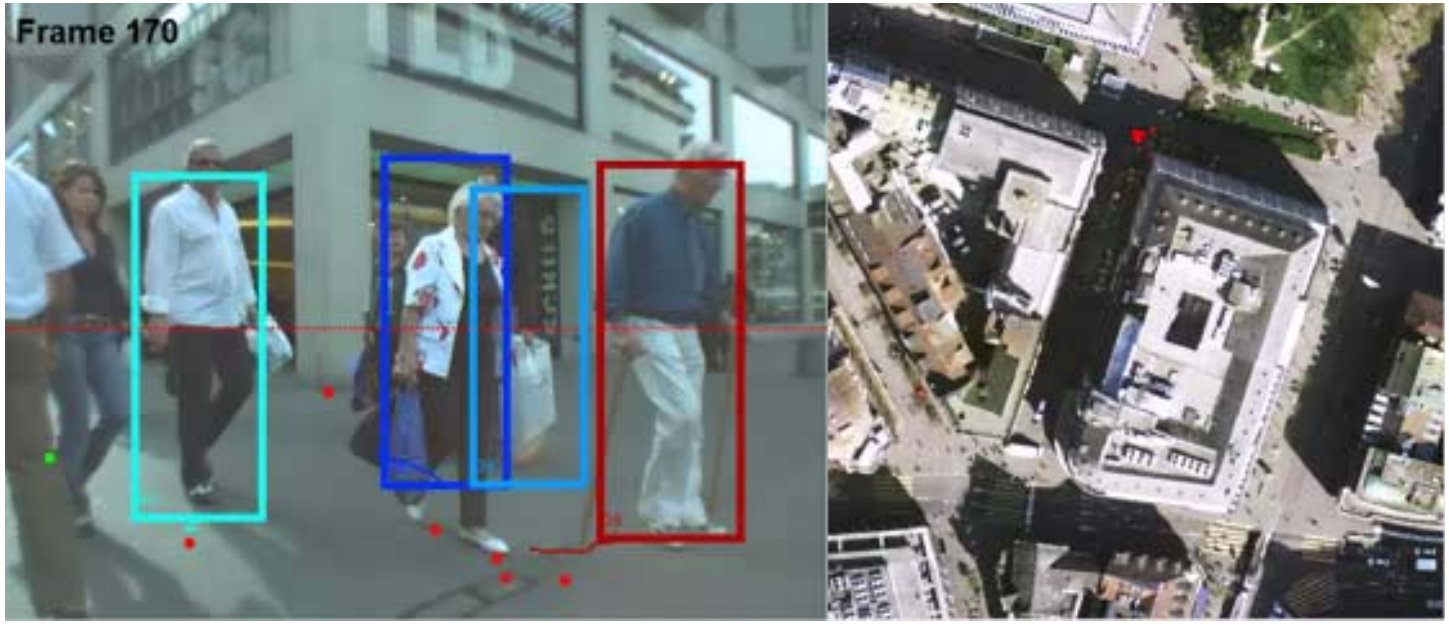
Joint multi-target tracking and camera motion estimation from videos

Wongun Choi

W. Choi & K. Shahid & S. Savarese WMC 2010
W. Choi & S. Savarese , ECCV 2010

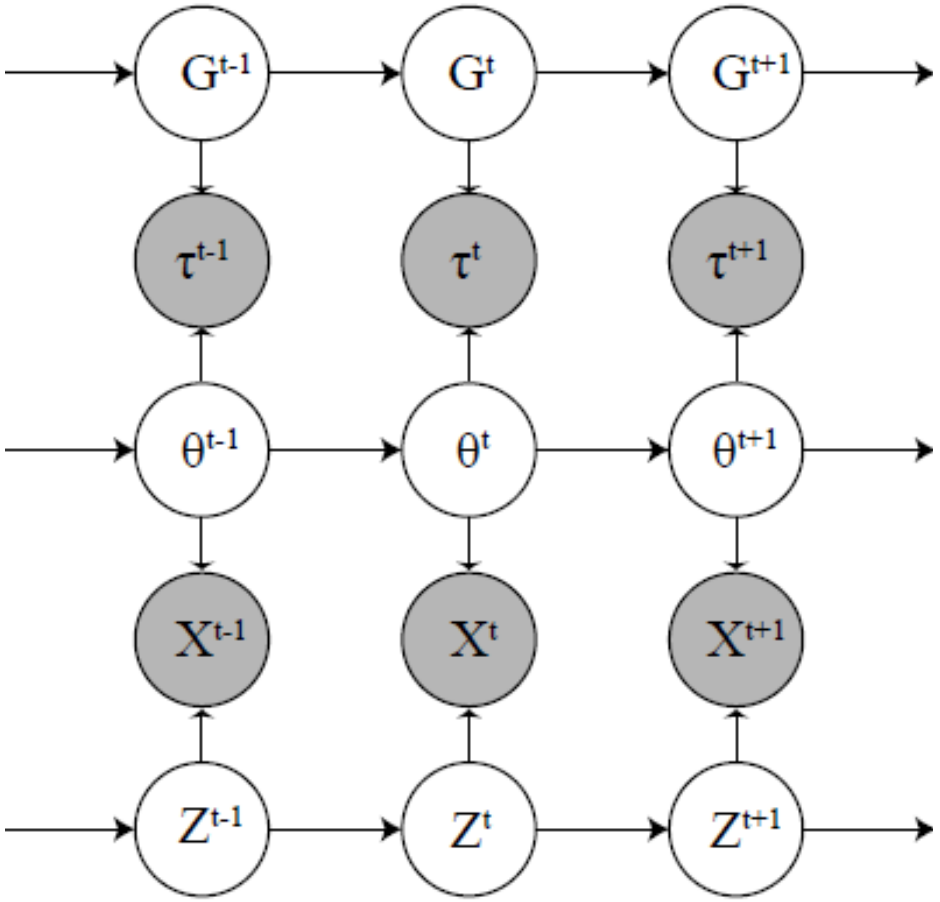
- Monocular cameras
- Un-calibrated cameras
- Arbitrary motion
- Highly cluttered scenes
  - Occlusion
  - Background clutter


Frame 1


Frame 170

Joint camera and object track estimation

•Choi & Savarese, ECCV 2010



Interest Points in 3D

Tracked Interest Points

Camera Parameters

Pedestrian Detections

Target Location in 3D

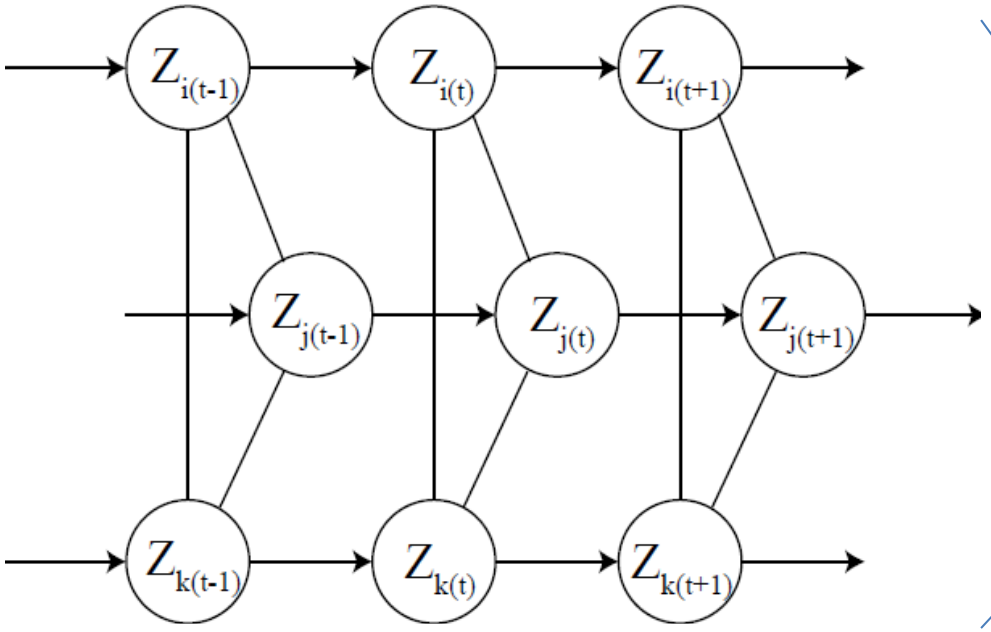
$$P(\Omega_t|\chi^t) \propto P(\Omega_t, \chi_t|\chi^{t-1}) = P(\chi_t|\Omega_t) \int P(\Omega_t|\Omega_{t-1}) P(\Omega_{t-1}|\chi^{t-1}) d\Omega_{t-1}$$

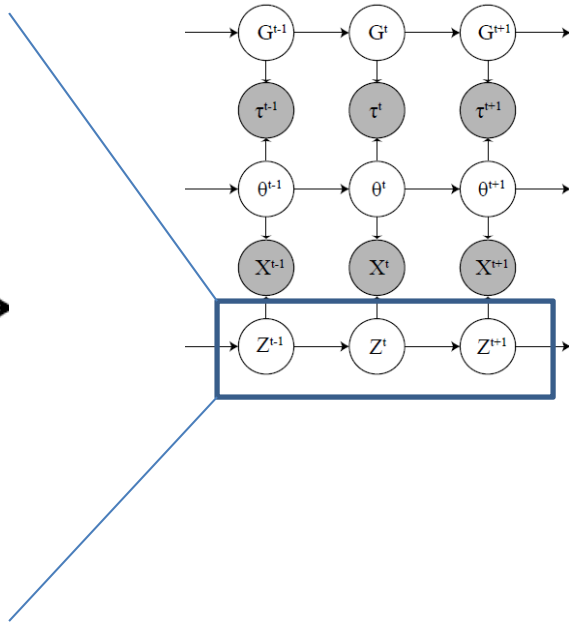
Ω : set of state variables

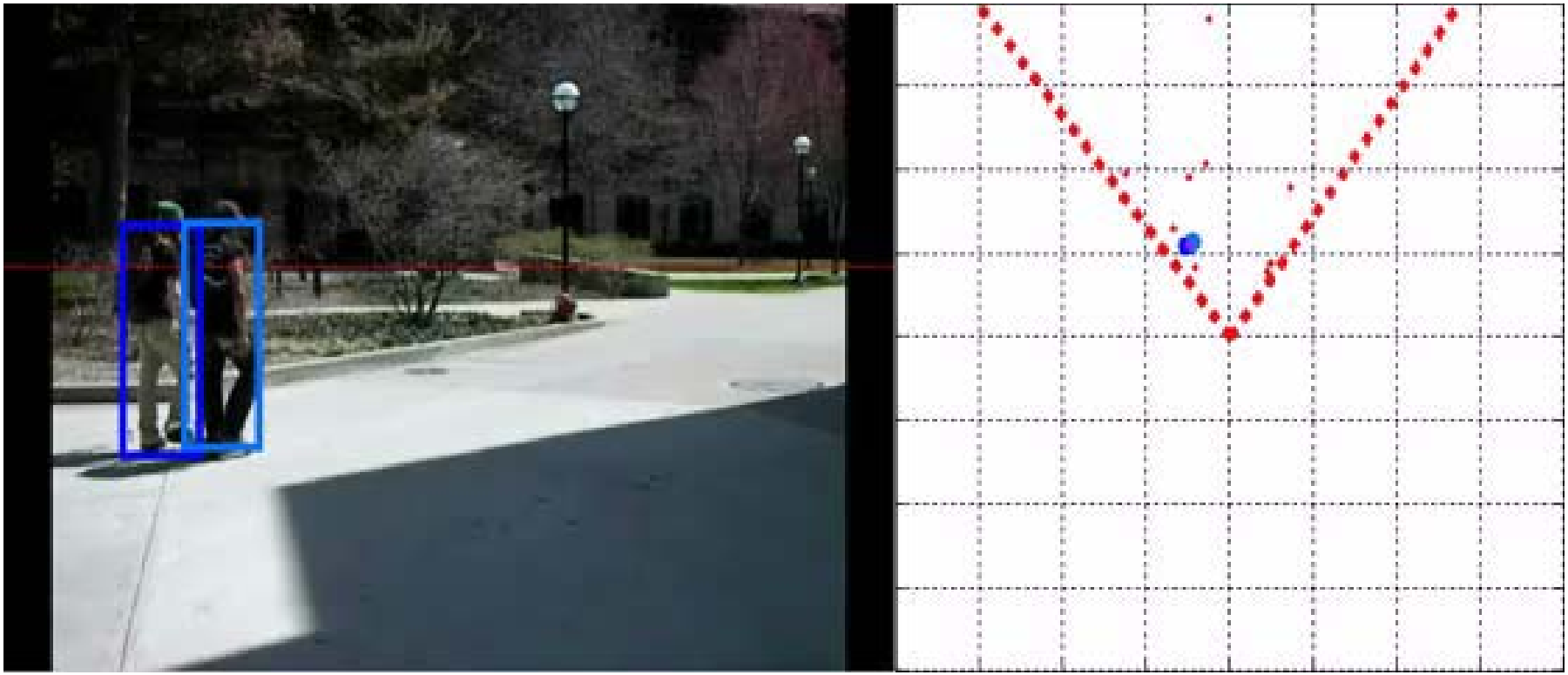
X : set of observationse

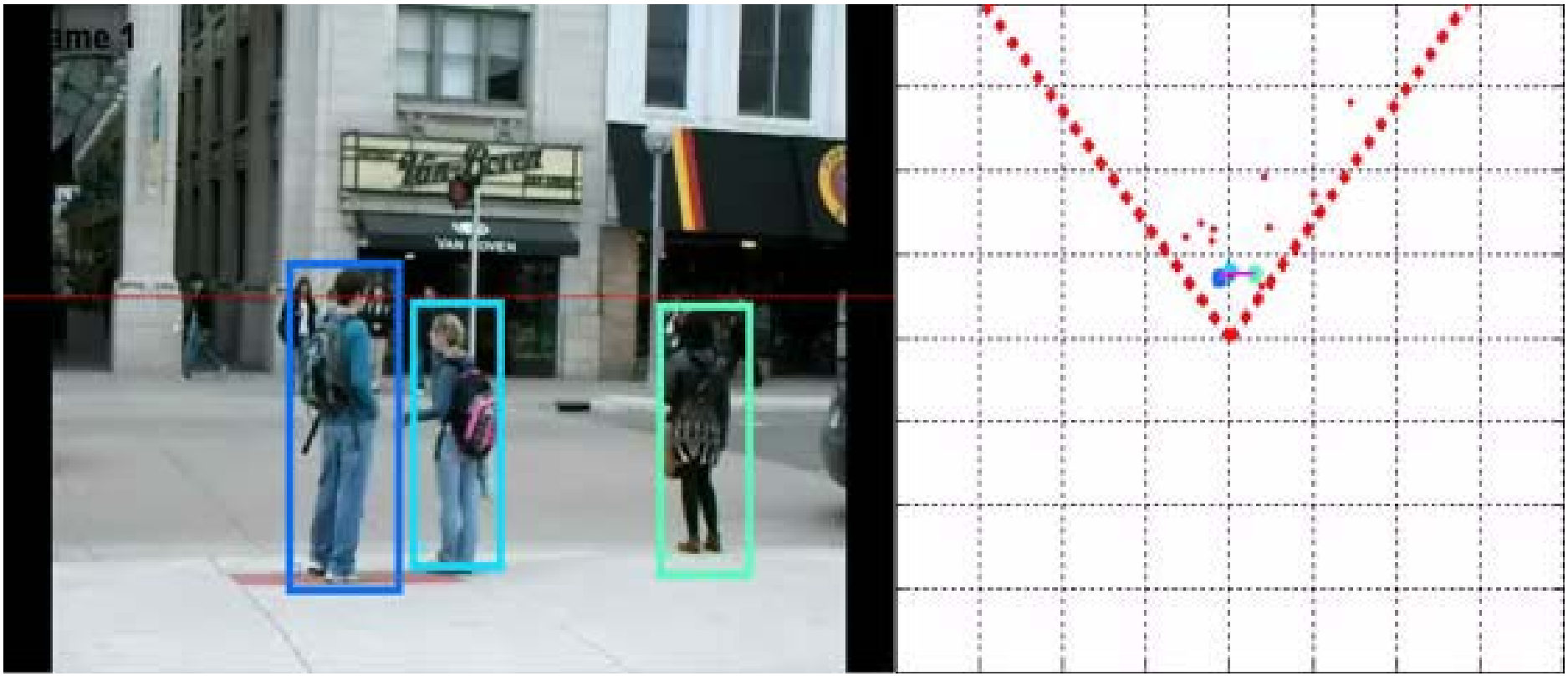
Interaction between Targets

- Interaction is modeled as a pair-wise MRF.
- Two exclusive models
  - Group motion vs. repulsion
  - Hidden variable to find the “mode” (hypothesis for the underlying interaction)





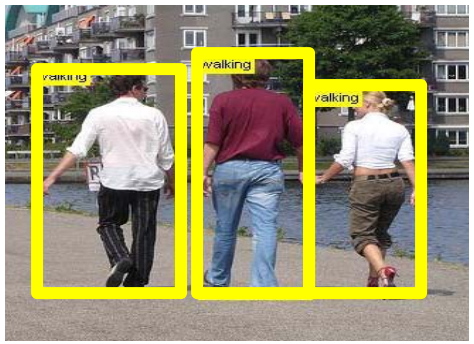





Activity:
pushing a stroll
Activity: walking
Activity:
crossing street
ONE WAY

Modeling human-human & human-object interaction

Object detection and tracking



Object categorization

- Gupta et al 2009
- Yao and Fei-Fei 2010
- Desai et al 2010
- Lan et al 2010
- Choi et al , 2009
- Choi & Savarese, 2010

3D object & scene modeling

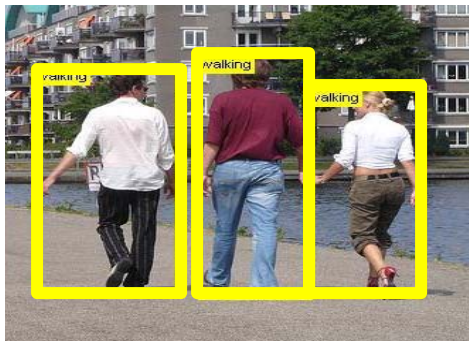
Object semantic

Activity semantic

3D geometry

Modeling human-human & human-object interaction

Object detection and tracking



Object categorization

- Choi et al , 2009
- Choi & Savarese, 2010

3D object & scene modeling

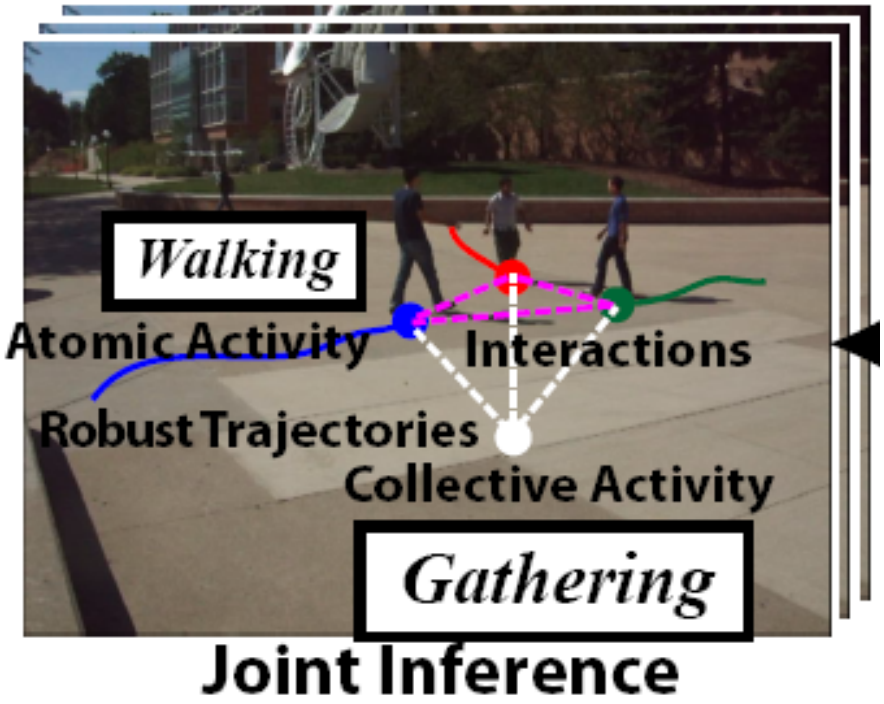
Object semantic

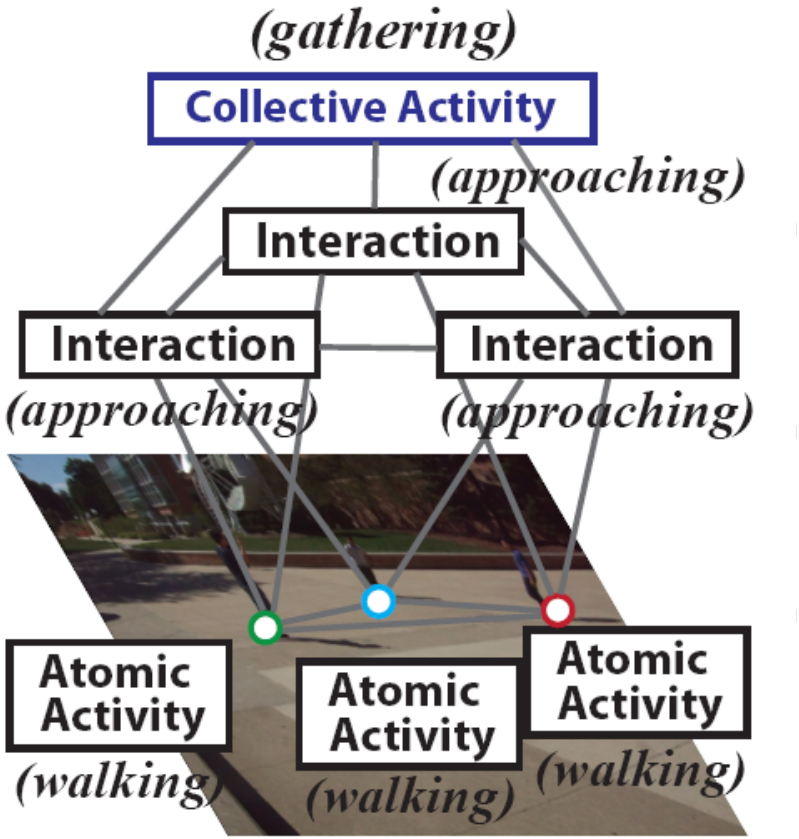
Activity semantic

3D geometry

Learning spatial-temporal relationship among humans

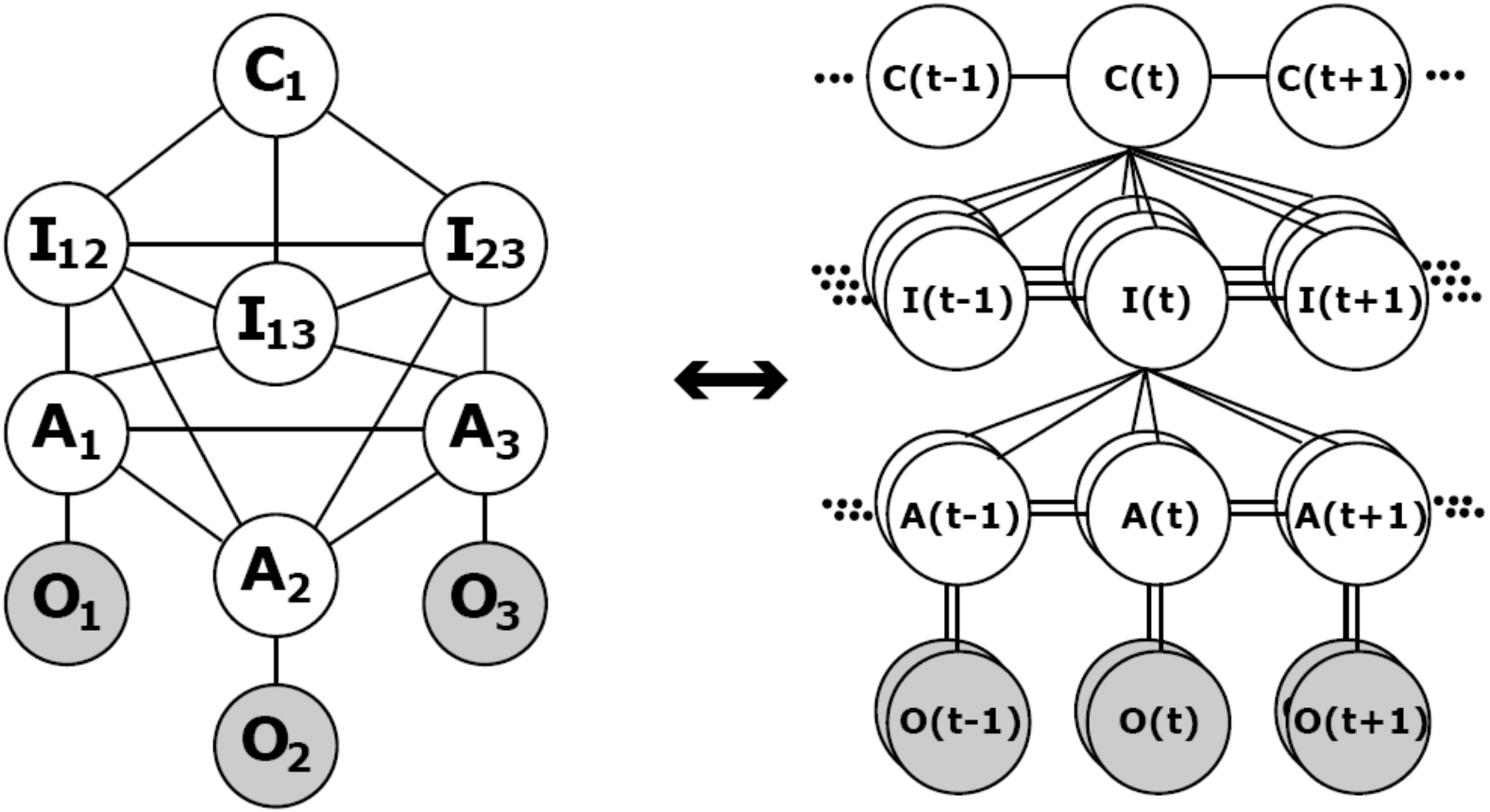
W. Choi & K. Shahid & S. Savarese , CVPR 2011, under review
W. Choi & S. Savarese, under preparation, 2011





Learning spatial-temporal relationship among humans

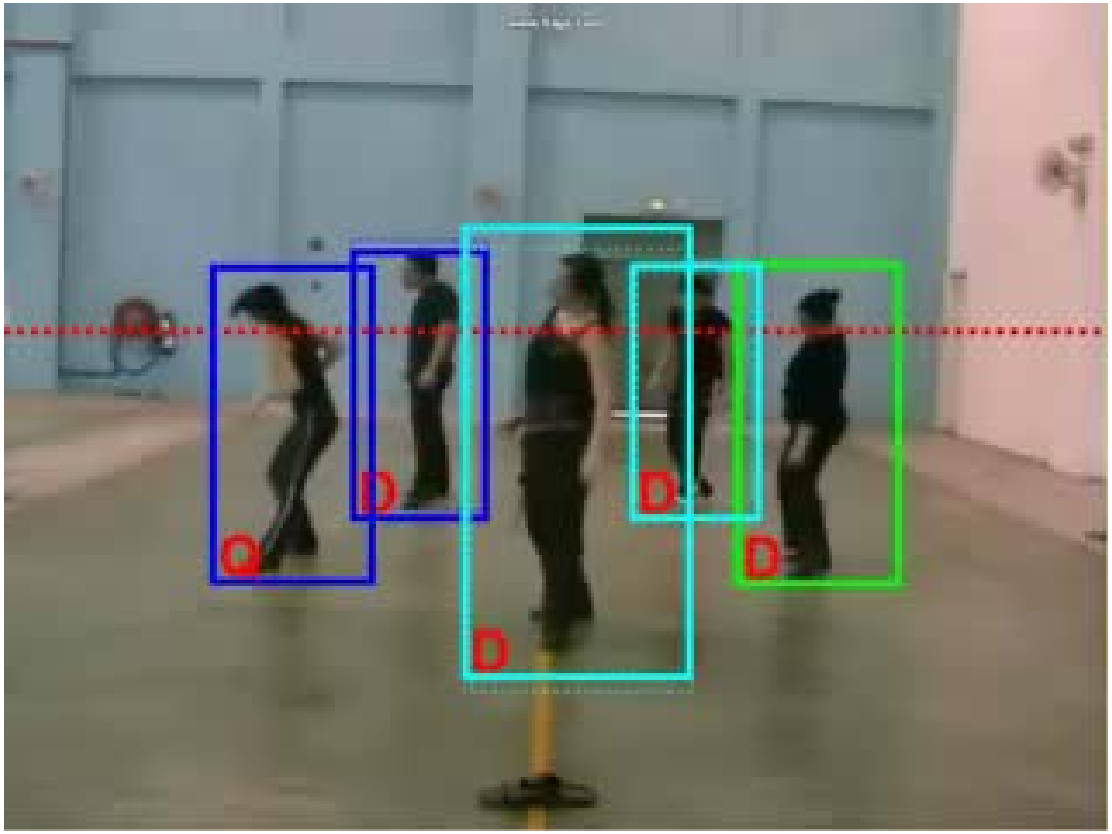
W. Choi & K. Shahid & S. Savarese , CVPR 2011, under review
W. Choi & S. Savarese, under preparation, 2011



Learning spatial-temporal relationship among humans

W. Choi & K. Shahid & S. Savarese WMC 2009
W. Choi & K. Shahid & S. Savarese , CVPR 2011, under review

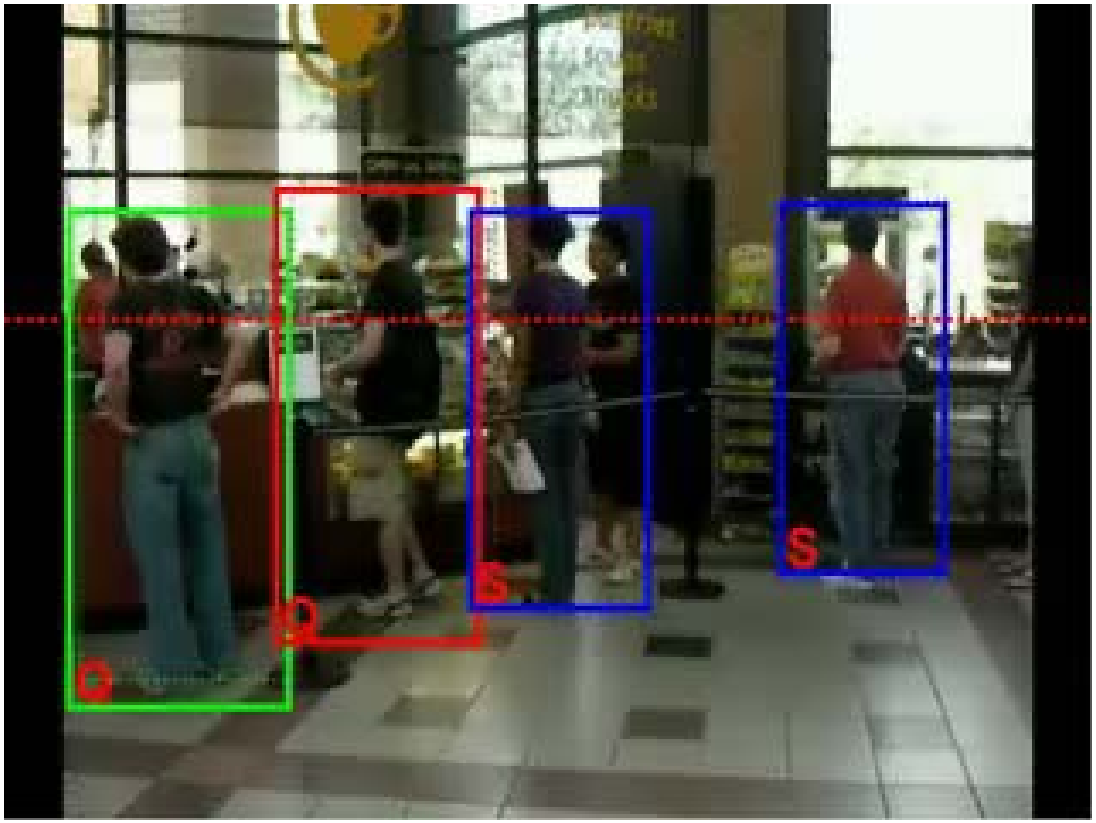
Crossing – Talking – Queuing – Dancing – jogging



Learning spatial-temporal relationship among humans

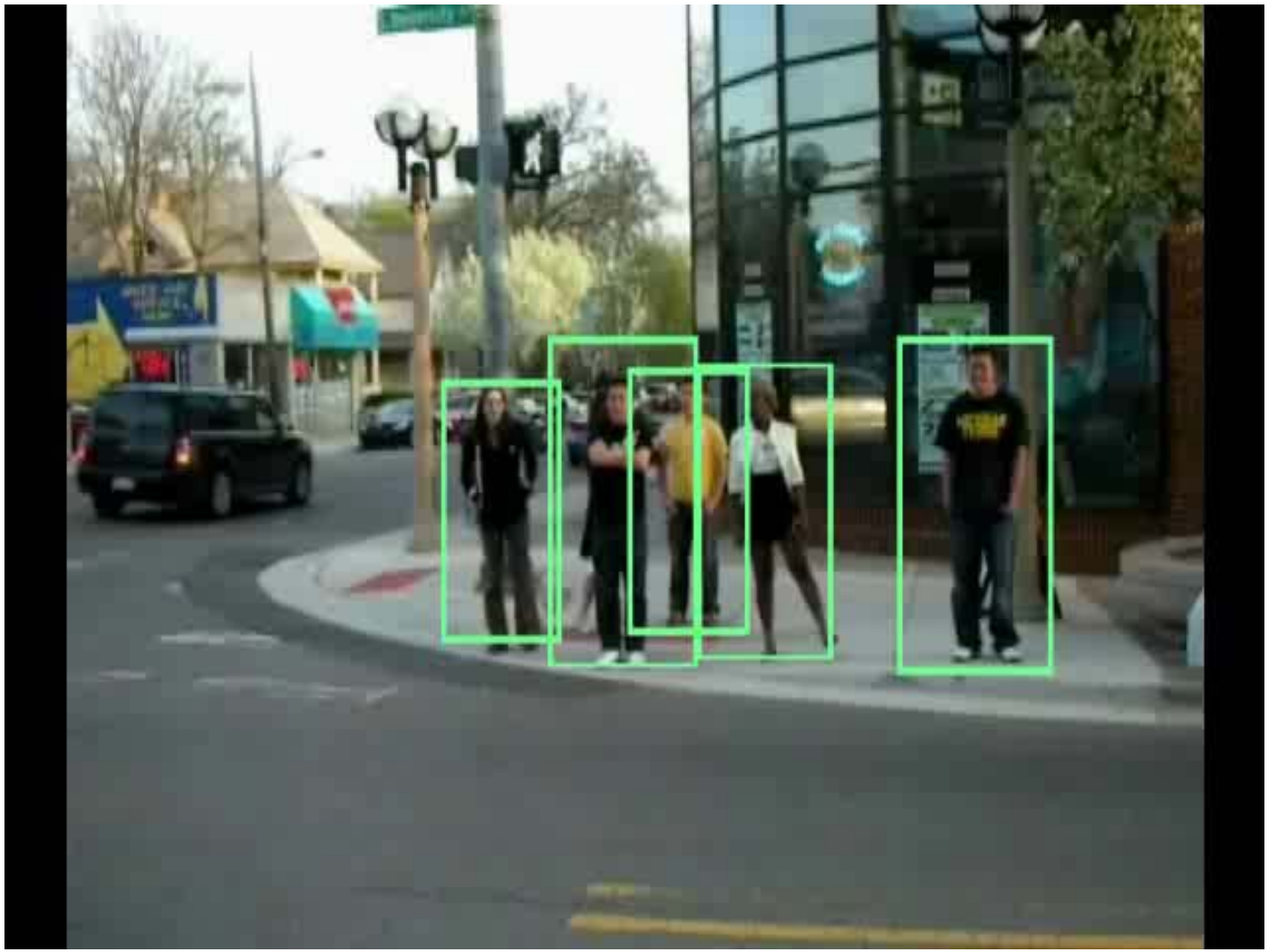
W. Choi & K. Shahid & S. Savarese WMC 2009
W. Choi & K. Shahid & S. Savarese , CVPR 2011, under review

Crossing – Talking – Queuing – Dancing – jogging



Learning spatial-temporal relationship among humans

W. Choi & K. Shahid & S. Savarese WMC 2009
W. Choi & K. Shahid & S. Savarese , CVPR 2011, under review
W. Choi & S. Savarese, under preparation, 2011



Conclusions

- Joint recognition and recognition is critical
- Leverage multi-view models for object categorization and 3D attribute estimation.
  - Pose estimation from a single view
  - 3D shape from a single view
- Models for coherent object detection/ tracking and scene 3D layout estimation
    - From single images
    - From videos

Thank you

FORD

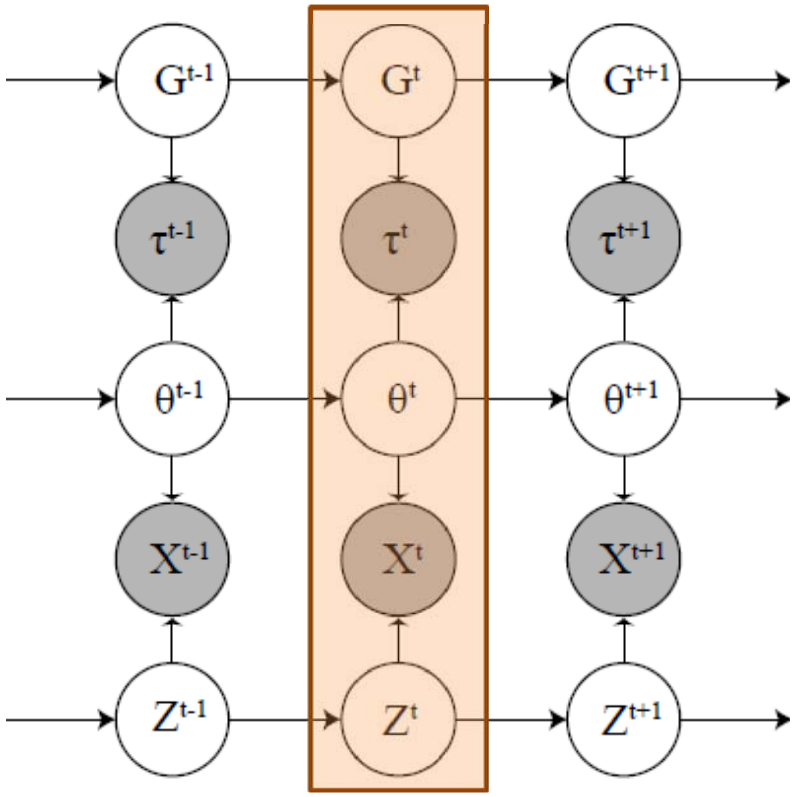
GigaScale Systems
Research Center (GSRC)

NSF

ARO ARMY

Probabilistic Formulation

$$P(\Omega_t|\chi^t) \propto P(\Omega_t, \chi_t|\chi^{t-1}) = P(\chi_t|\Omega_t) \int P(\Omega_t|\Omega_{t-1}) P(\Omega_{t-1}|\chi^{t-1}) d\Omega_{t-1}$$



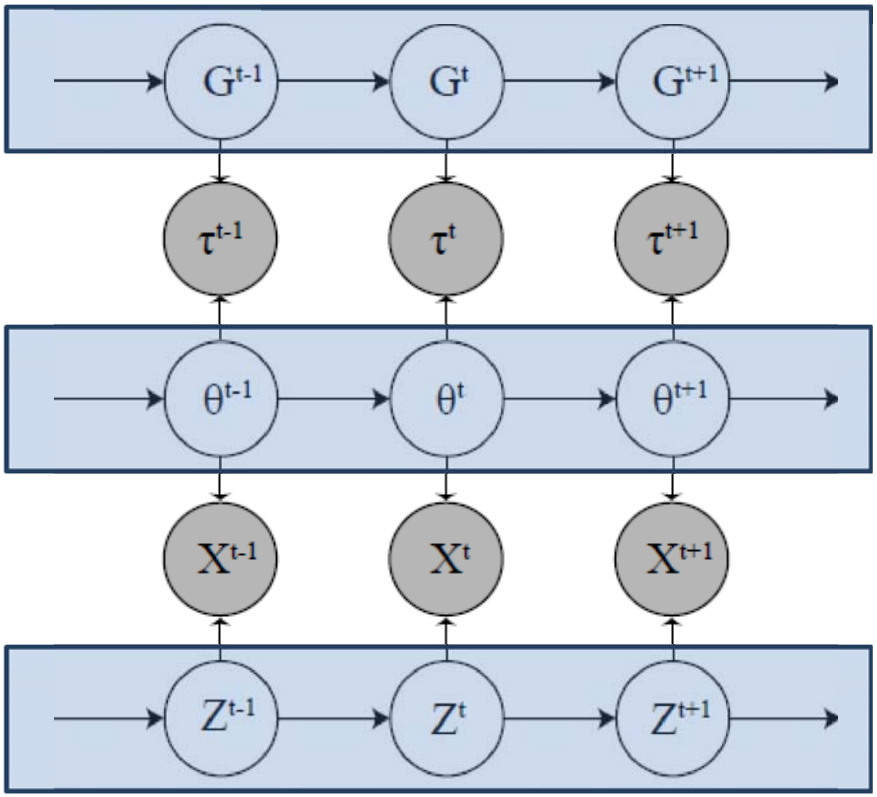
Measurement Model

$$P(\chi_t|\Omega_t) = P(X_t, Y_t|Z_t, \Theta_t) P(\tau_t|G_t, \Theta_t)$$

$$P(\Omega_t|\Omega_{t-1}) = P(Z_t|Z_{t-1}) P(\Theta_t|\Theta_{t-1}) P(G_t|G_{t-1})$$

Probabilistic Formulation

$$P(\Omega_t|\chi^t) \propto P(\Omega_t, \chi_t|\chi^{t-1}) = P(\chi_t|\Omega_t) \int P(\Omega_t|\Omega_{t-1}) P(\Omega_{t-1}|\chi^{t-1}) d\Omega_{t-1}$$



$$P(\chi_t|\Omega_t) = P(X_t, Y_t|Z_t, \Theta_t) P(\tau_t|G_t, \Theta_t)$$

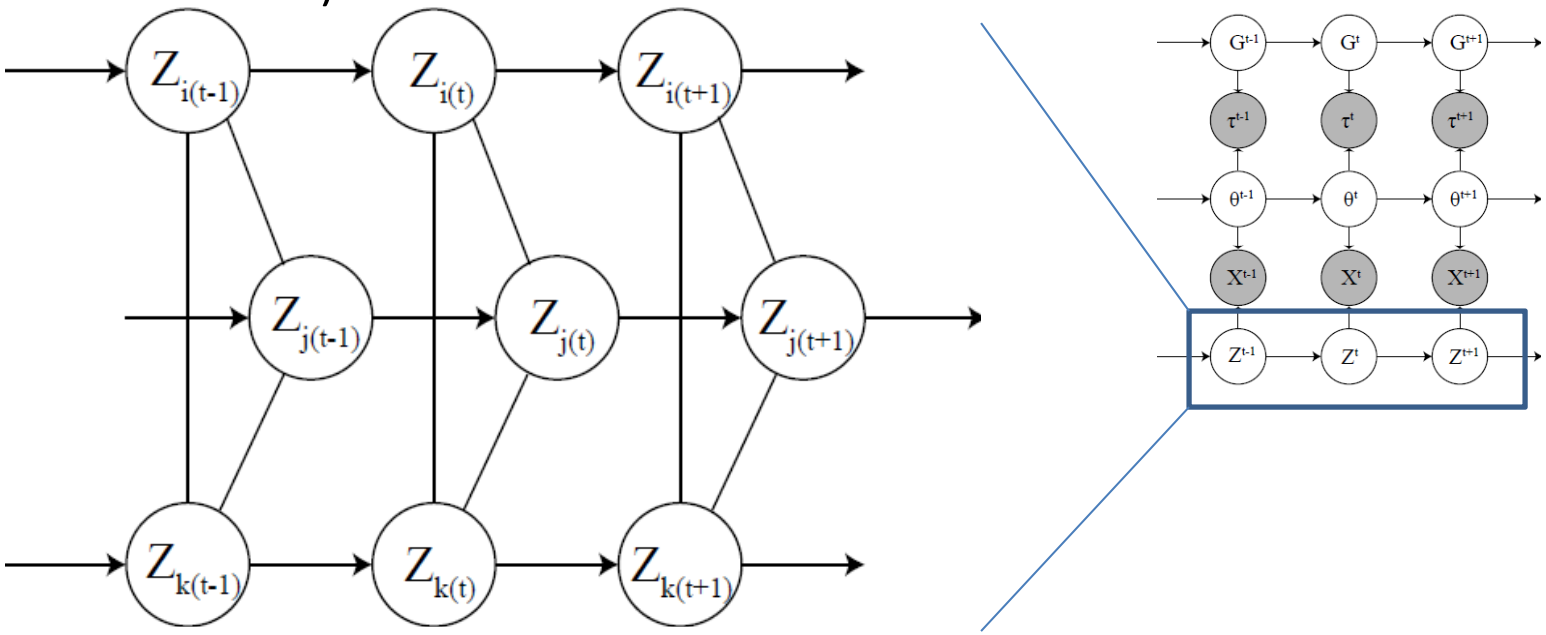
$$P(\Omega_t|\Omega_{t-1}) = P(Z_t|Z_{t-1}) P(\Theta_t|\Theta_{t-1}) P(G_t|G_{t-1})$$

Motion Model

Independent motion?

Interaction between Targets

- Interaction is modeled as a pair-wise MRF.
- Two exclusive models
  - Group motion vs. repulsion
  - Hidden variable to find the “mode”. (hypothesis for the underlying interaction)



Posterior Probability

$$P(\Omega_t|\chi^t) \propto P(\Omega_t, \chi_t|\chi^{t-1}) = P(\chi_t|\Omega_t) \int P(\Omega_t|\Omega_{t-1}) P(\Omega_{t-1}|\chi^{t-1}) d\Omega_{t-1}$$

$$P(\chi_t|\Omega_t) = P(X_t, Y_t|Z_t, \Theta_t) P(\tau_t|G_t, \Theta_t)$$

$$P(\Omega_t|\Omega_{t-1}) = P(Z_t|Z_{t-1}) P(\Theta_t|\Theta_{t-1}) P(G_t|G_{t-1})$$

- Challenges
  - Nonlinear projection.
  - Pair-wise MRF for interaction. (non-gaussian motion model)
  - Hypothesis variables for detection and ground features
- Not able to apply analytical inference algorithm.