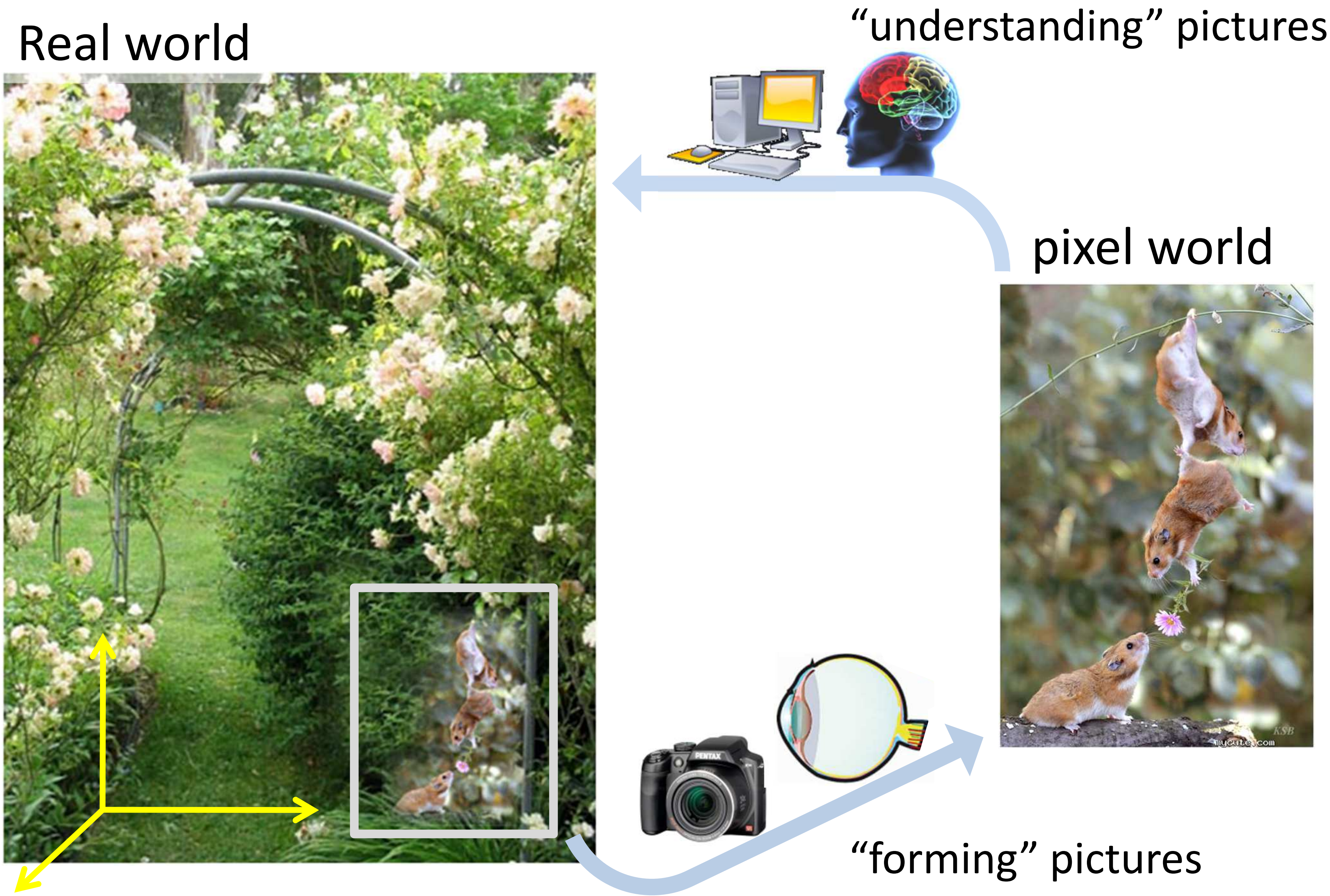
Lecture 19: Advanced Topics from Stanford Vision Lab

Professor Fei-Fei Li

Stanford Vision Lab

What is vision?



What is vision?

Real world

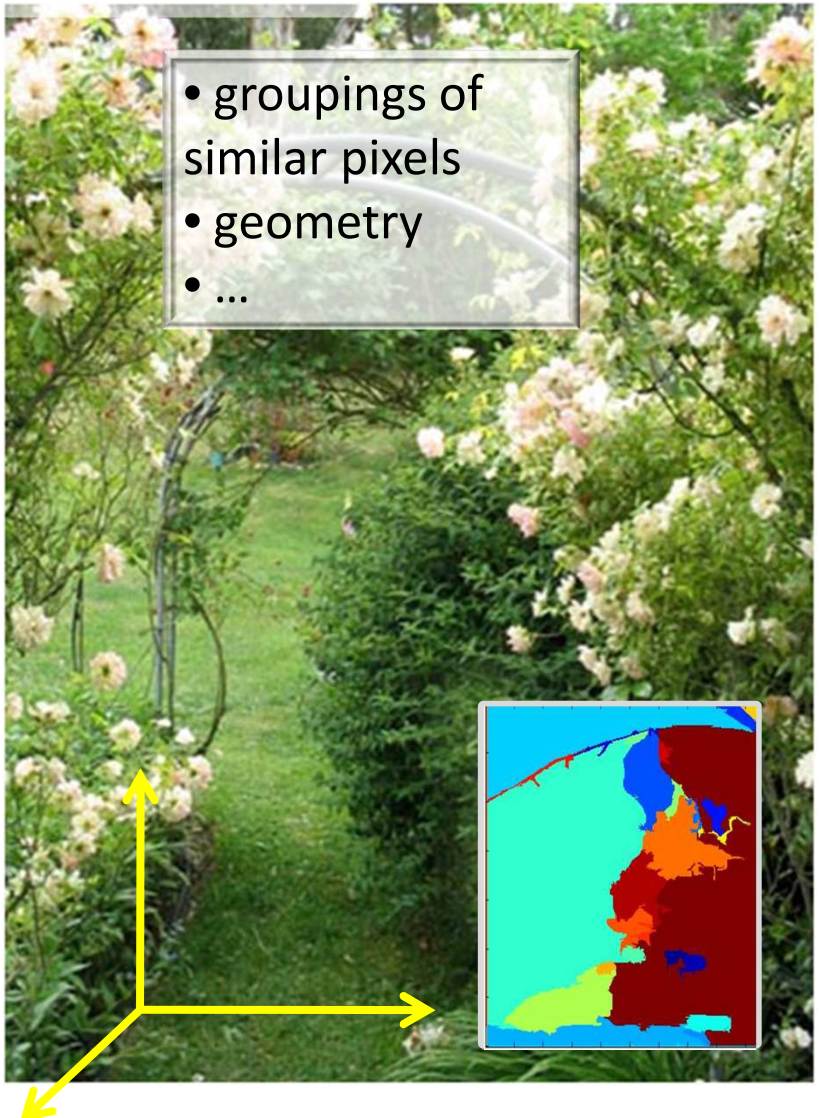
“understanding” pictures

pixel world

Low-Level Vision

What is vision?

Real world



“understanding” pictures

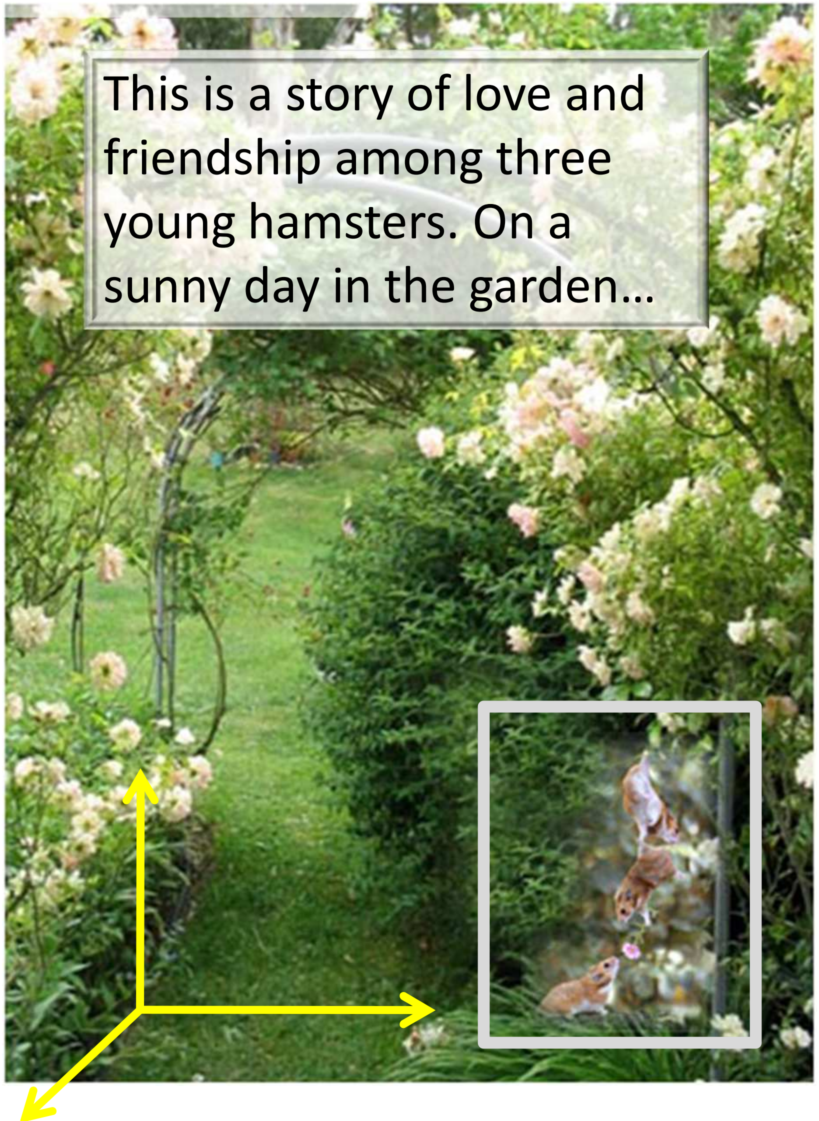
pixel world

Mid-Level Vision

Low-Level Vision

What is vision?

Real world



"understanding" pictures

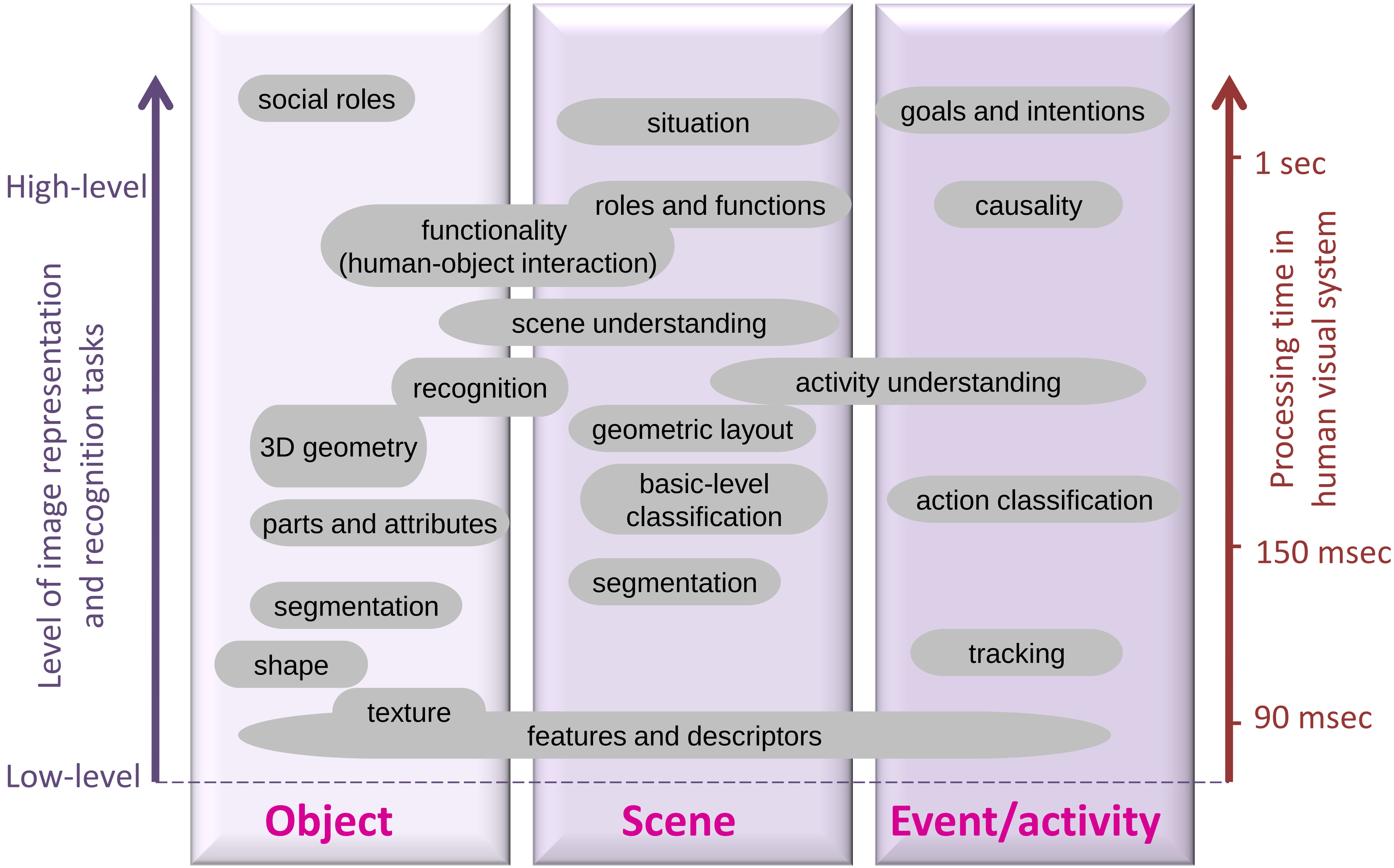
pixel world

High-Level Vision

Mid-Level Vision

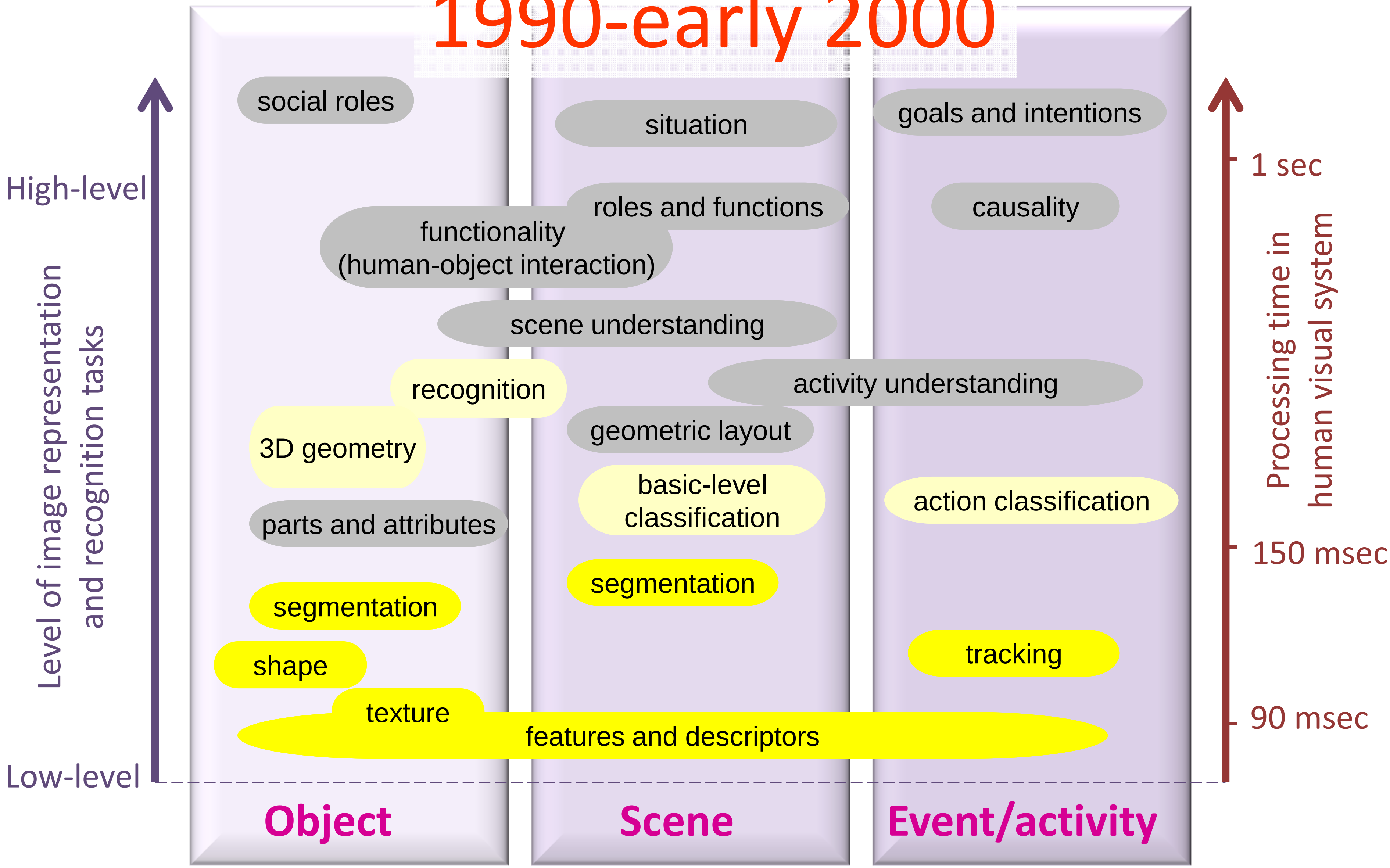
Low-Level Vision

Story telling in images

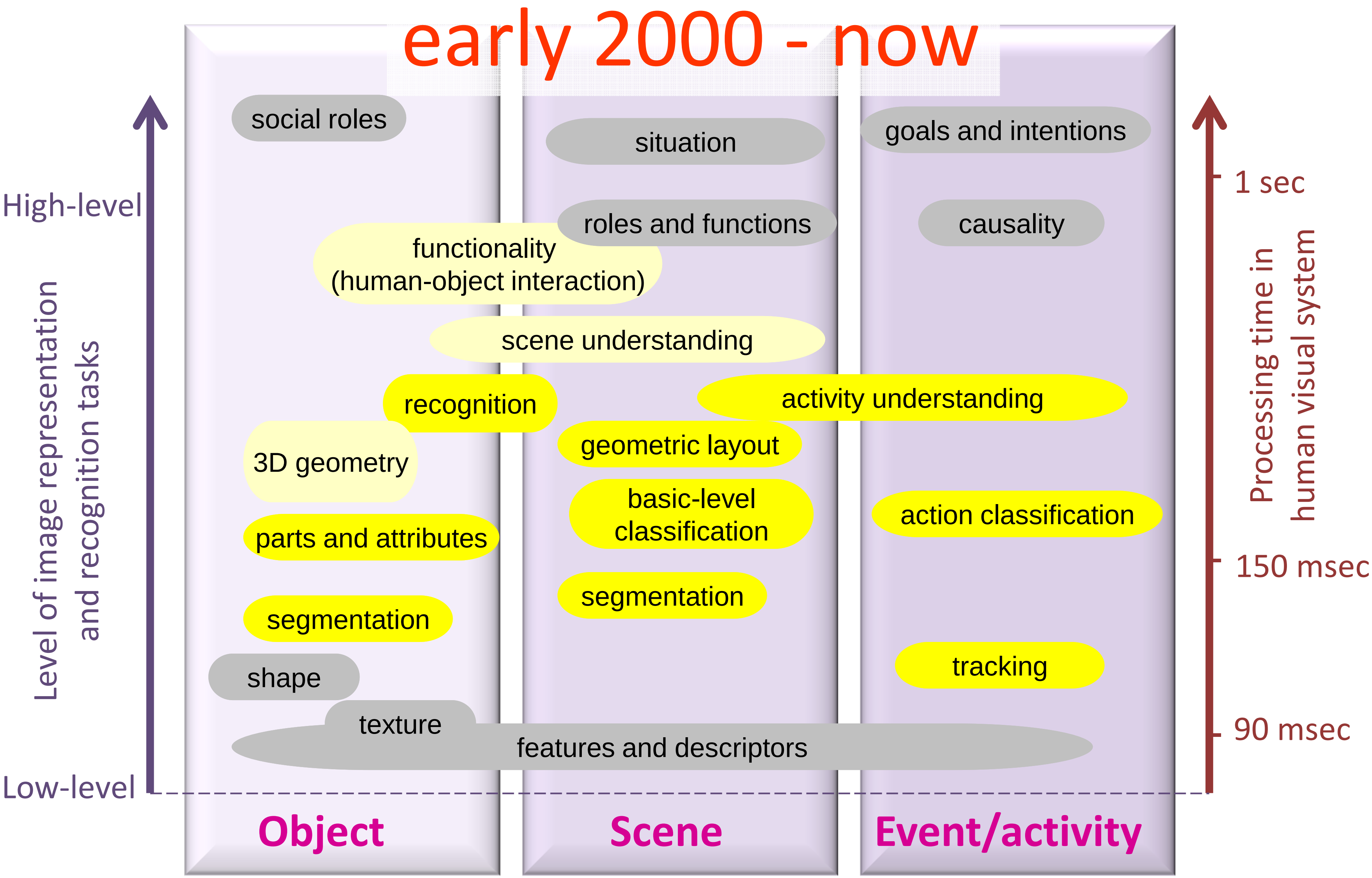


Story telling in images

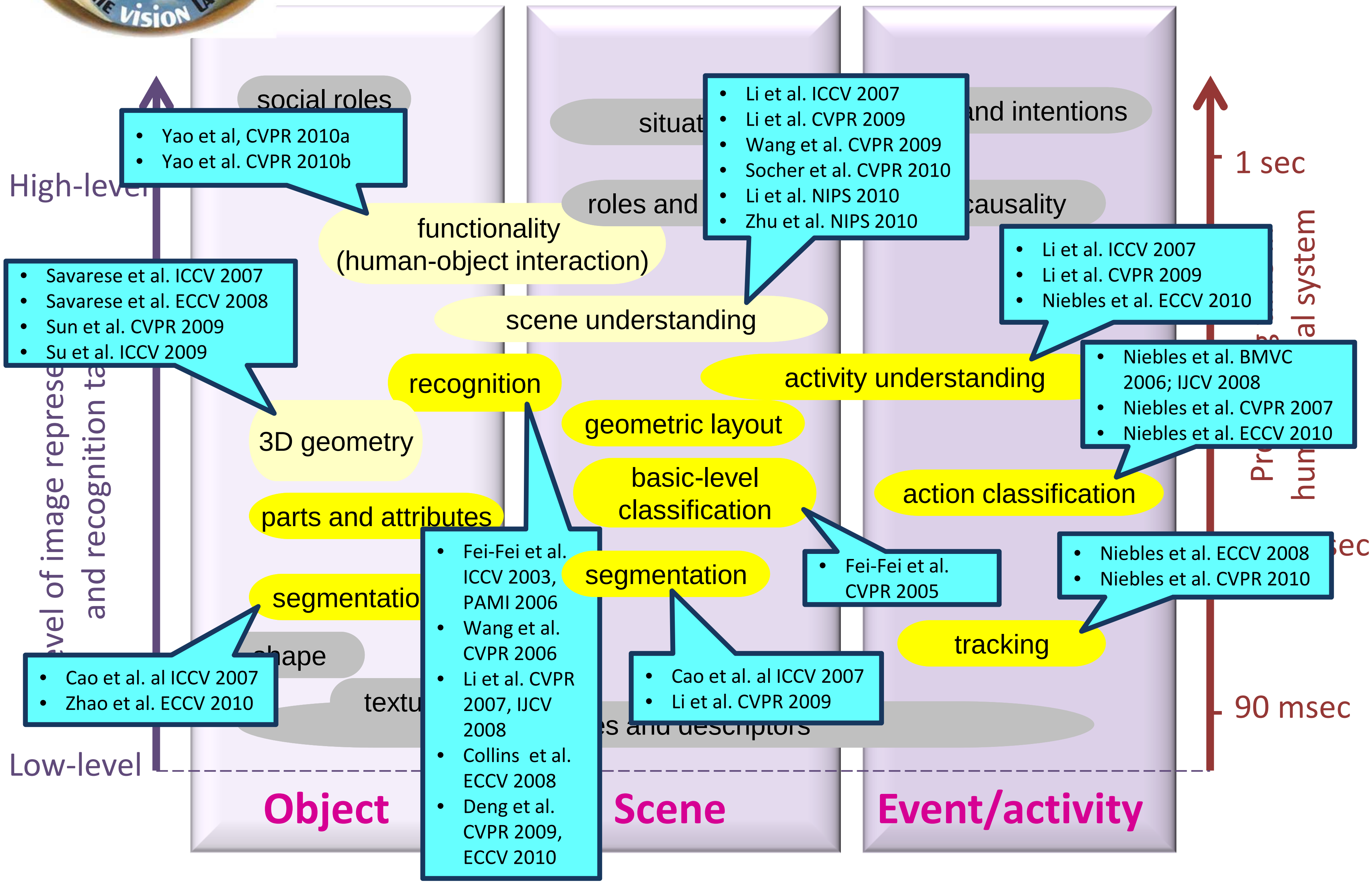
1990-early 2000



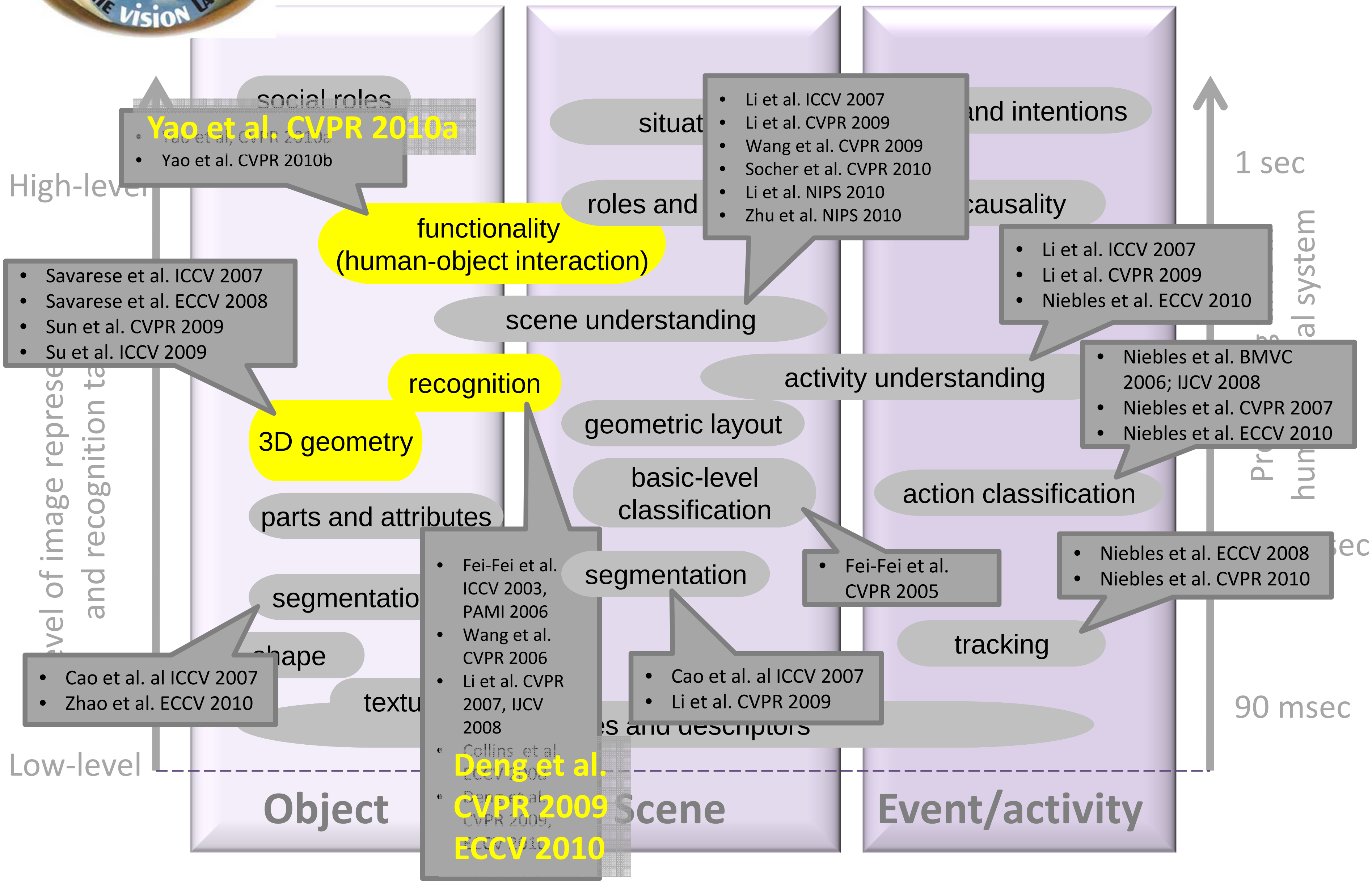
Story telling in images



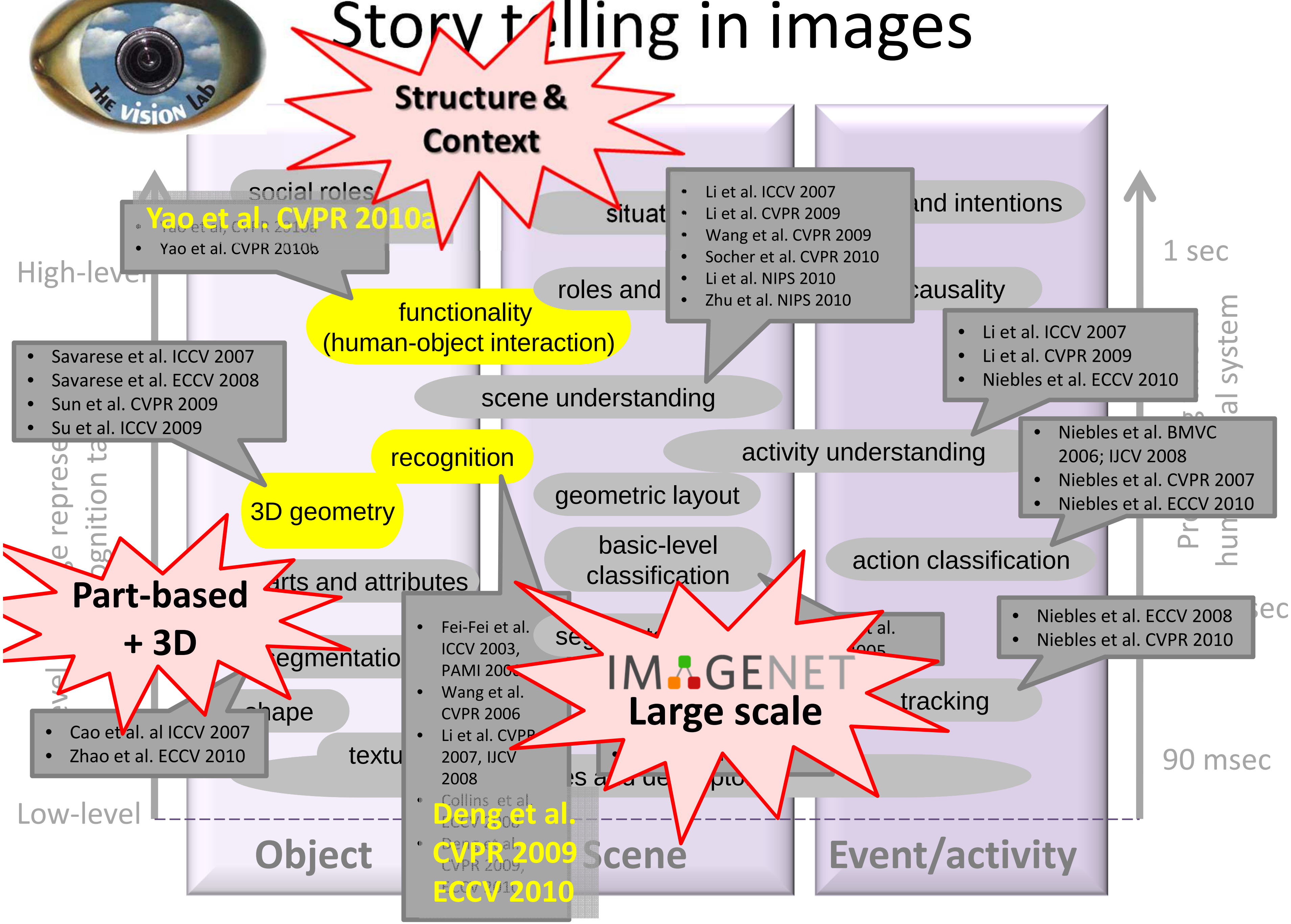
Story telling in images



Story telling in images

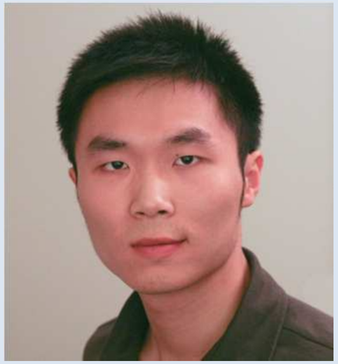


Story telling in images



J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei. **ImageNet: A Large-Scale Hierarchical Image Database.** *Computer Vision and Pattern Recognition (CVPR).* 2009.

J. Deng, A. Berg, K. Li and L. Fei-Fei. **What does classifying more than 10,000 image categories tell us?** *Proceedings of the 12th European Conference of Computer Vision (ECCV).* 2010.



Jia Deng
Princeton/Stanford

Prof. Kai Li
Princeton University

Prof. Alex Berg
SUNY-Stony Brook

- Cao et al. al ICCV 2007
- Zhao et al. ECCV 2010

- Fei-Fei et al. ICCV 2003, PAMI 2006
- Wang et al. CVPR 2006
- Li et al. CVPR 2007, IJCV 2008

Deng et al. CVPR 2009 ECCV 2010

Fei-Fei et al.

- Niebles et al. ECCV 2008
- Niebles et al. CVPR 2010

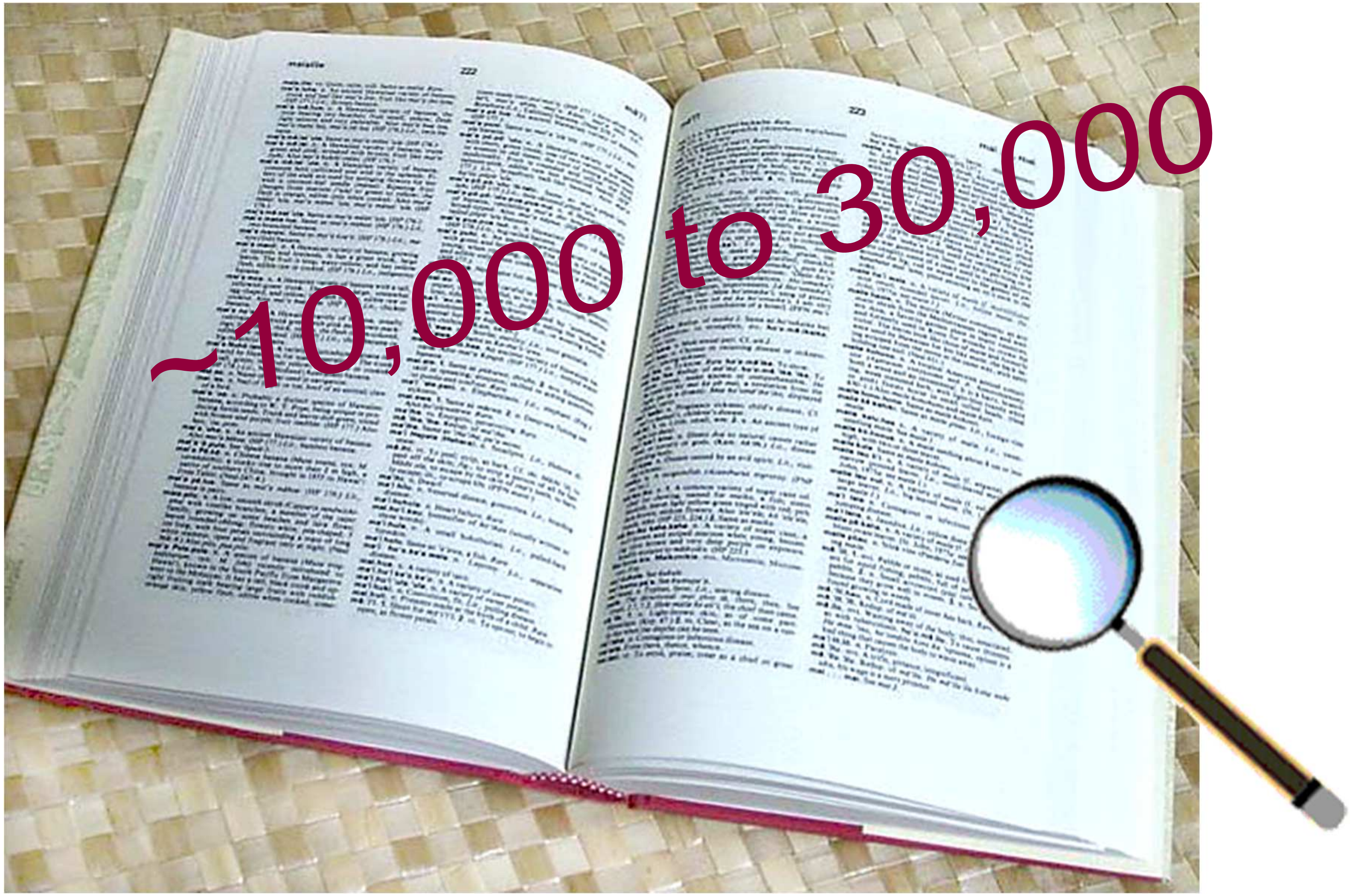
IMAGENET
Large scale

segmentation
shape
tracking
90 msec
Low-level
Object
Scene
Event/activity

Outline

- ImageNet dataset
  - Properties of ImageNet
- ECCV2010: a 10000-way classification benchmark experiment
  - Size matters
  - Density matters
  - Hierarchy matters
- An “infallible” classifier

How many object categories are there?



Biederman 1987

Datasets and computer vision

UIUC Cars (2004)
S. Agarwal, A. Awan, D. Roth

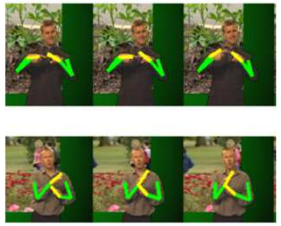
CMU/VASC Faces (1998)
H. Rowley, S. Baluja, T. Kanade

FERET Faces (1998)
P. Phillips, H. Wechsler, J. Huang, P. Raus

COIL Objects (1996)
S. Nene, S. Nayar, H. Murase

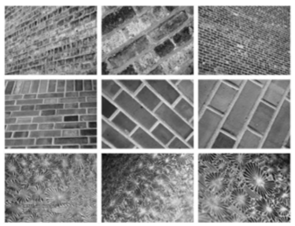
MNIST digits (1998-10)
Y LeCun & C. Cortes

KTH human action (2004)
I. Laptev & B. Caputo



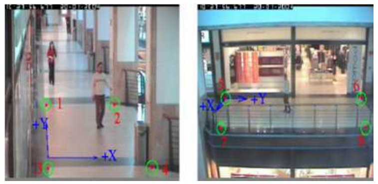
Sign Language (2008)
P. Buehler, M. Everingham, A. Zisserman

Segmentation (2001)
D. Martin, C. Fowlkes, D. Tal, J. Malik.

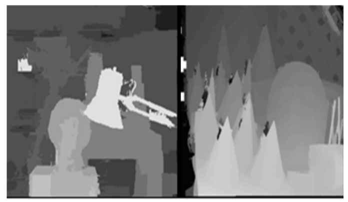


3D Textures (2005)
S. Lazebnik, C. Schmid, J. Ponce

CuRRET Textures (1999)
K. Dana B. Van Ginneken S. Nayar J. Koenderink



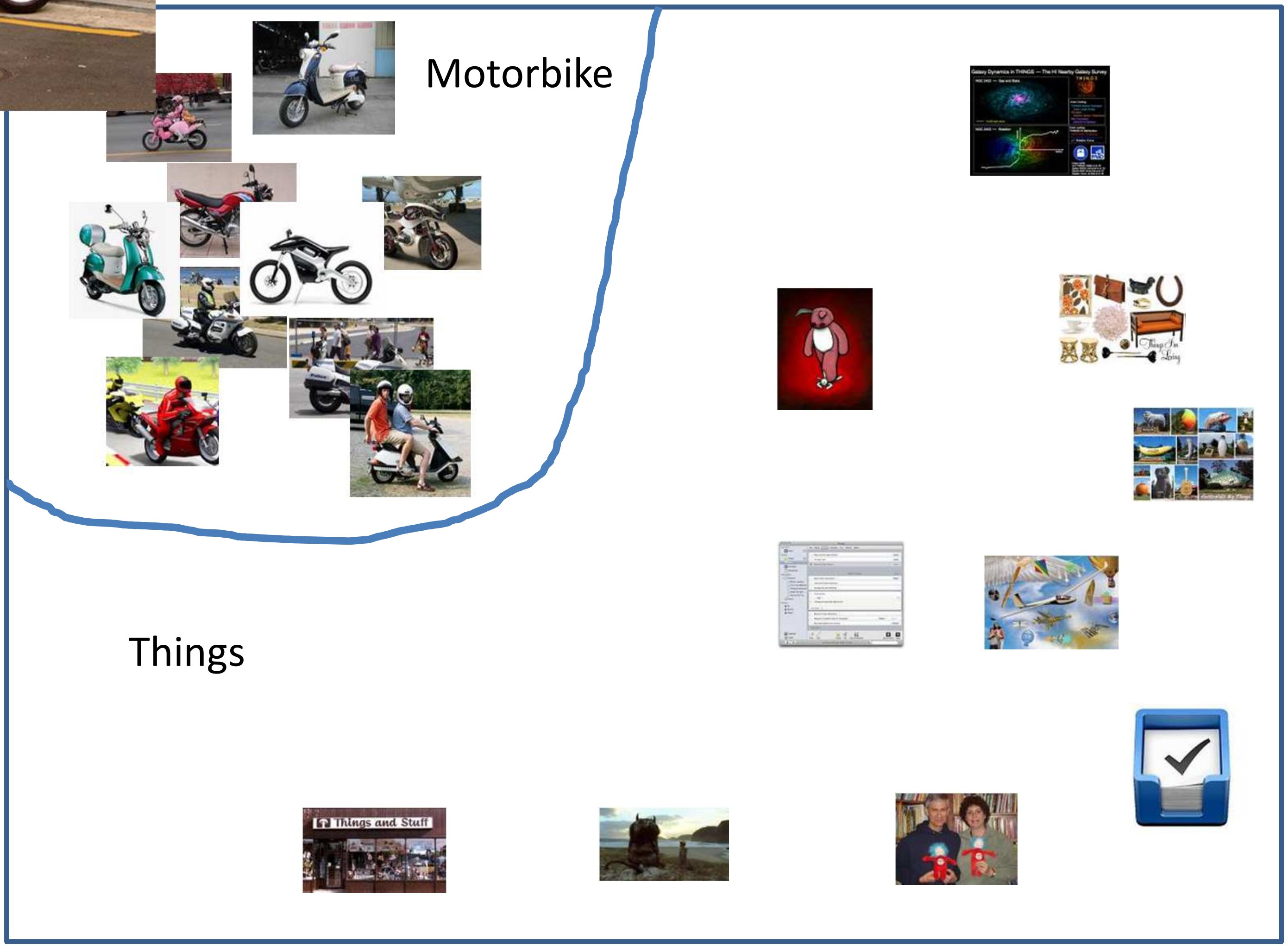
CAVIAR Tracking (2005)
R. Fisher, J. Santos-Victor J. Crowley



Middlebury Stereo (2002)
D. Scharstein R. Szeliski

Fergus, Perona, Zisserman, CVPR 2003

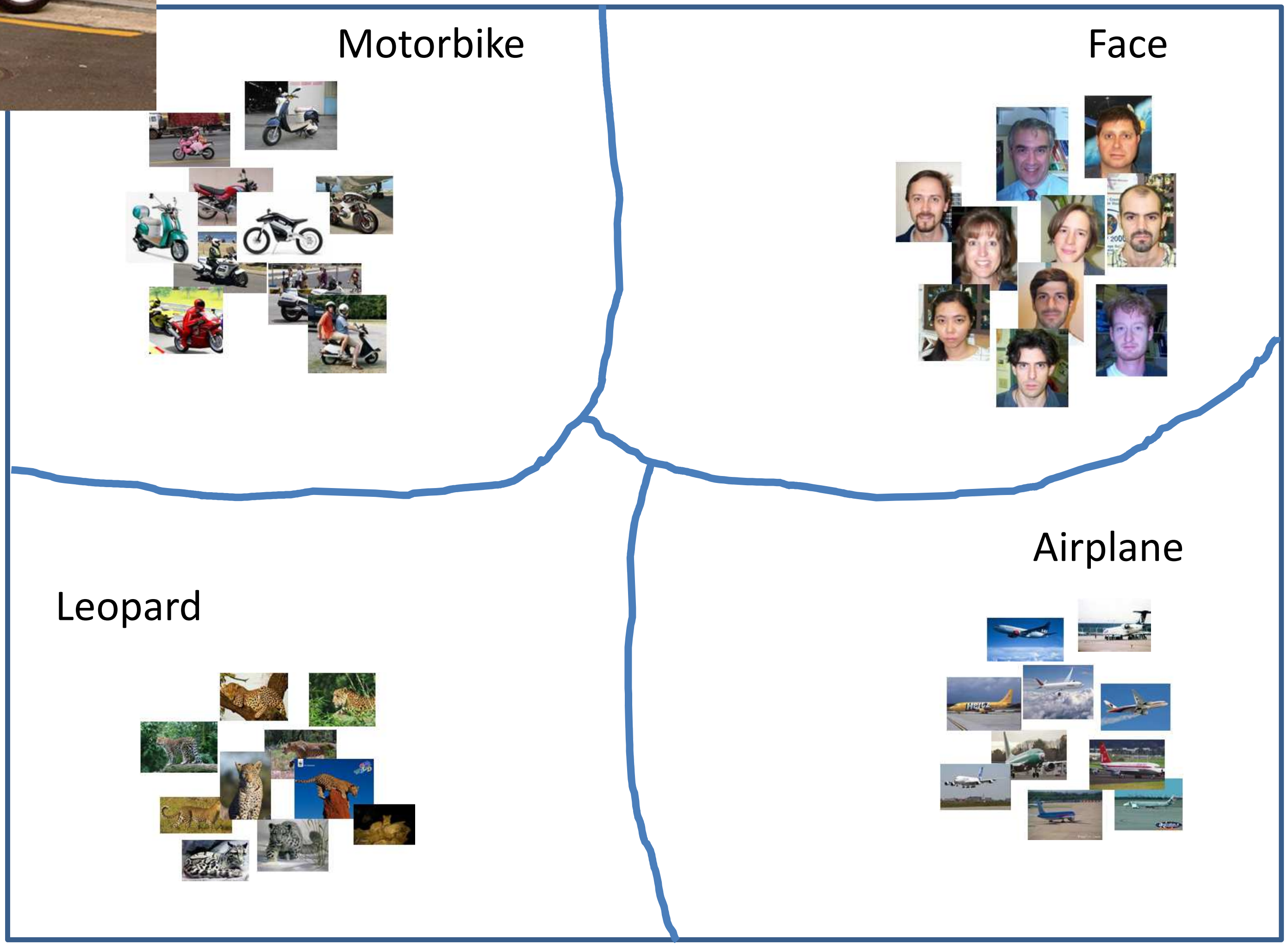
Object Recognition



Fergus, Perona, Zisserman, CVPR 2003

Holub, et al. ICCV 2005; Sivic et al. ICCV 2005

Object Recognition

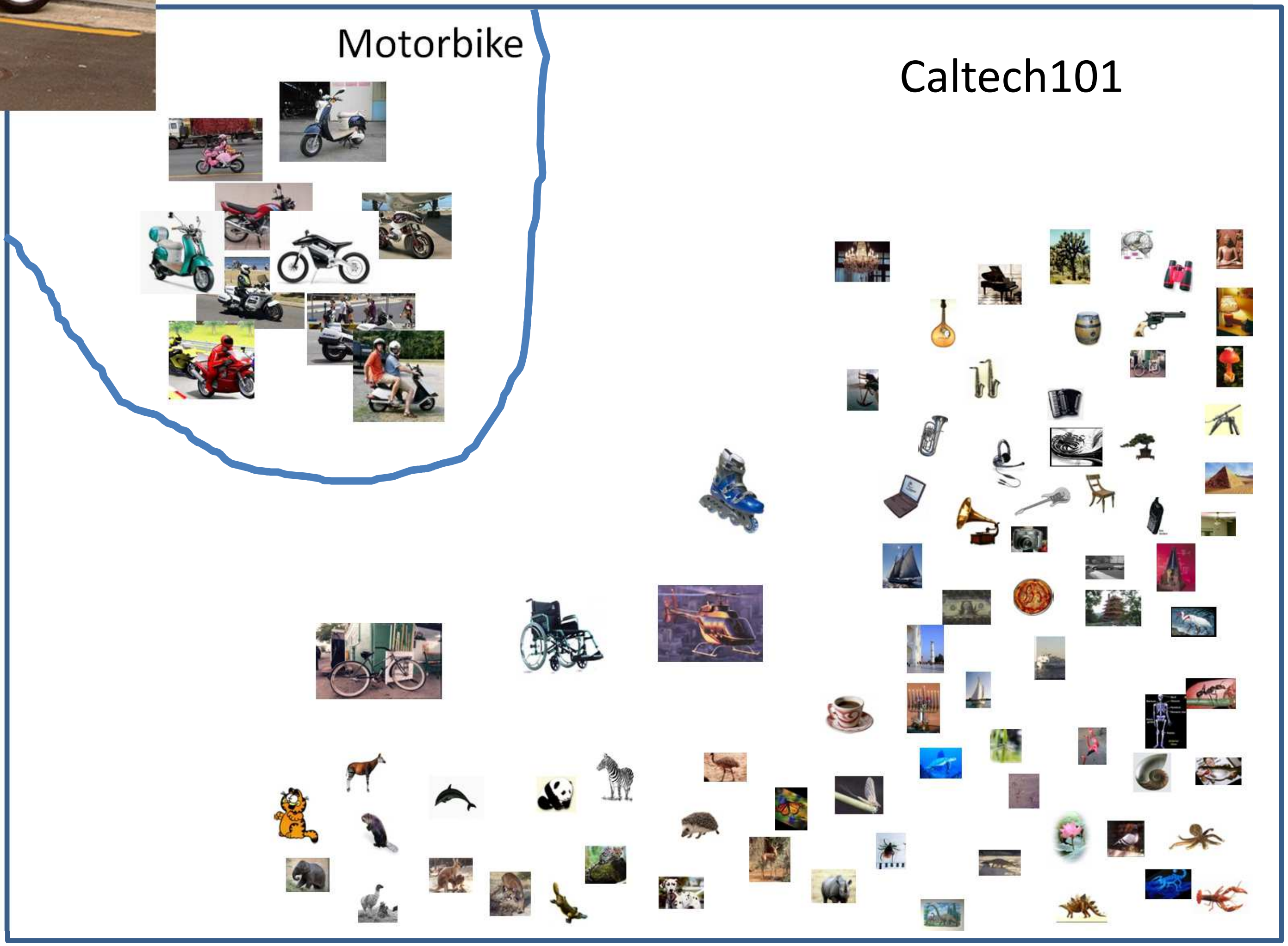


Fergus, Perona, Zisserman, CVPR 2003

Holub, et al. ICCV 2005; Sivic et al. ICCV 2005

Fei-Fei et al. CVPR 2004; Grauman et al. ICCV 2005; Lazebnik et al. CVPR 2006
Zhang & Malik, 2006; Varma & Sizzerman 2008; Wang et al. 2006; [....]

Object Recognition



PASCAL

[Everingham et al, 2009]

MSRC

[Shotton et al. 2006]

Fergus, Perona, Zisserman, CVPR 2003

Holub, et al. ICCV 2005; Sivic et al. ICCV 2005

Fei-Fei et al. CVPR 2004; Grauman et al. ICCV 2005; Lazebnik et al. CVPR 2006
Zhang & Malik, 2006; Varma & Sizzerman 2008; Wang et al. 2006; [….]

Biederman 1987

Object Recognition

ESP

[Ahn et al, 2006]

LabelMe

[Russell et al, 2005]

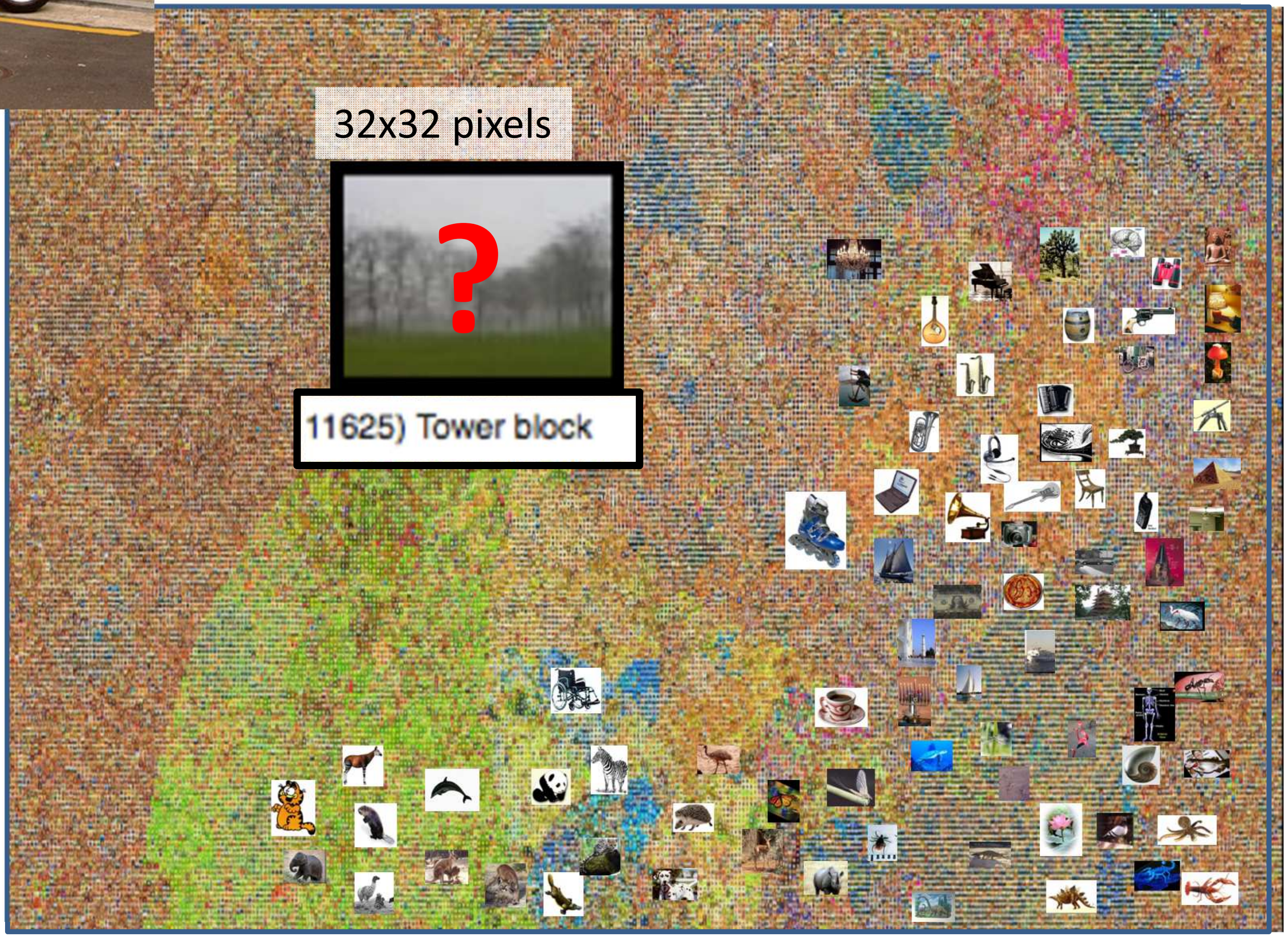
Lotus Hill

[Yao et al, 2007]

TinyImage

Torralba et al. 2007

Background image courtesy: Antonio Torralba

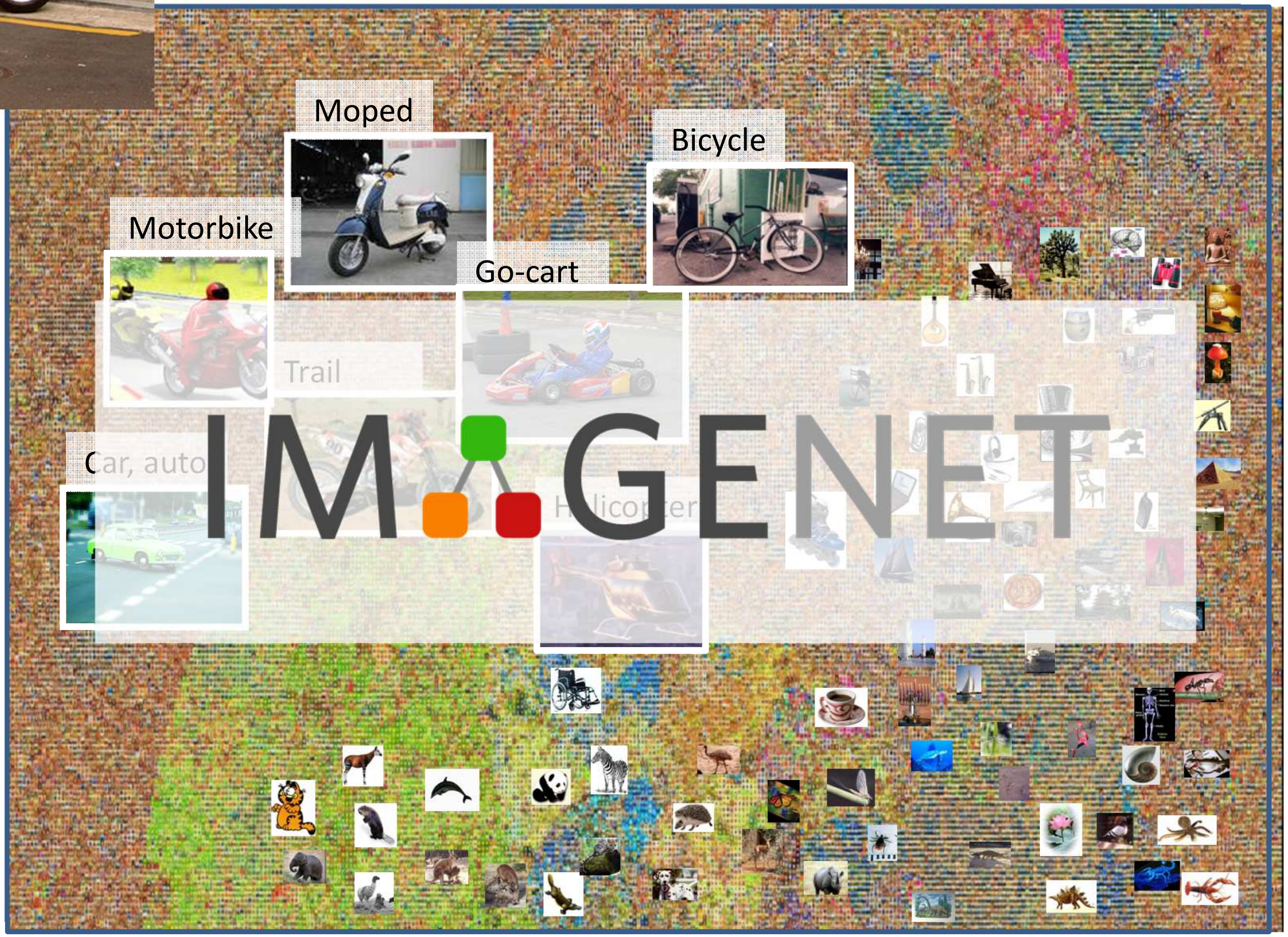


Deng, Dong, Socher, Li, Li, Fei-Fei, CVPR, 2009

- Full resolution images
- All human verified
- Ongoing: providing b.box, attributes, features, etc.

Object Recognition

amazon mechanical turk
Artificial Artificial Intelligence
beta



IMAGENET in a glance

15K categories; 11+million images; ~800im/categ; free to public at **www.image-net.org**

IMAGENET

11,231,732 images, 15589 synsets indexed

Explore New! Download New! **Challenge** People Publication About

Not logged in. Login | Signup

ImageNet is an image database organized according to the WordNet hierarchy (currently only the nouns), in which each node of the hierarchy is depicted by hundreds and thousands of images. Currently we have an average of over five hundred images per node. We hope ImageNet will become a useful resource for researchers, educators, students and all of you who share our passion for pictures.

Click here to learn more about ImageNet, Click here to join the ImageNet mailing list.

SEARCH

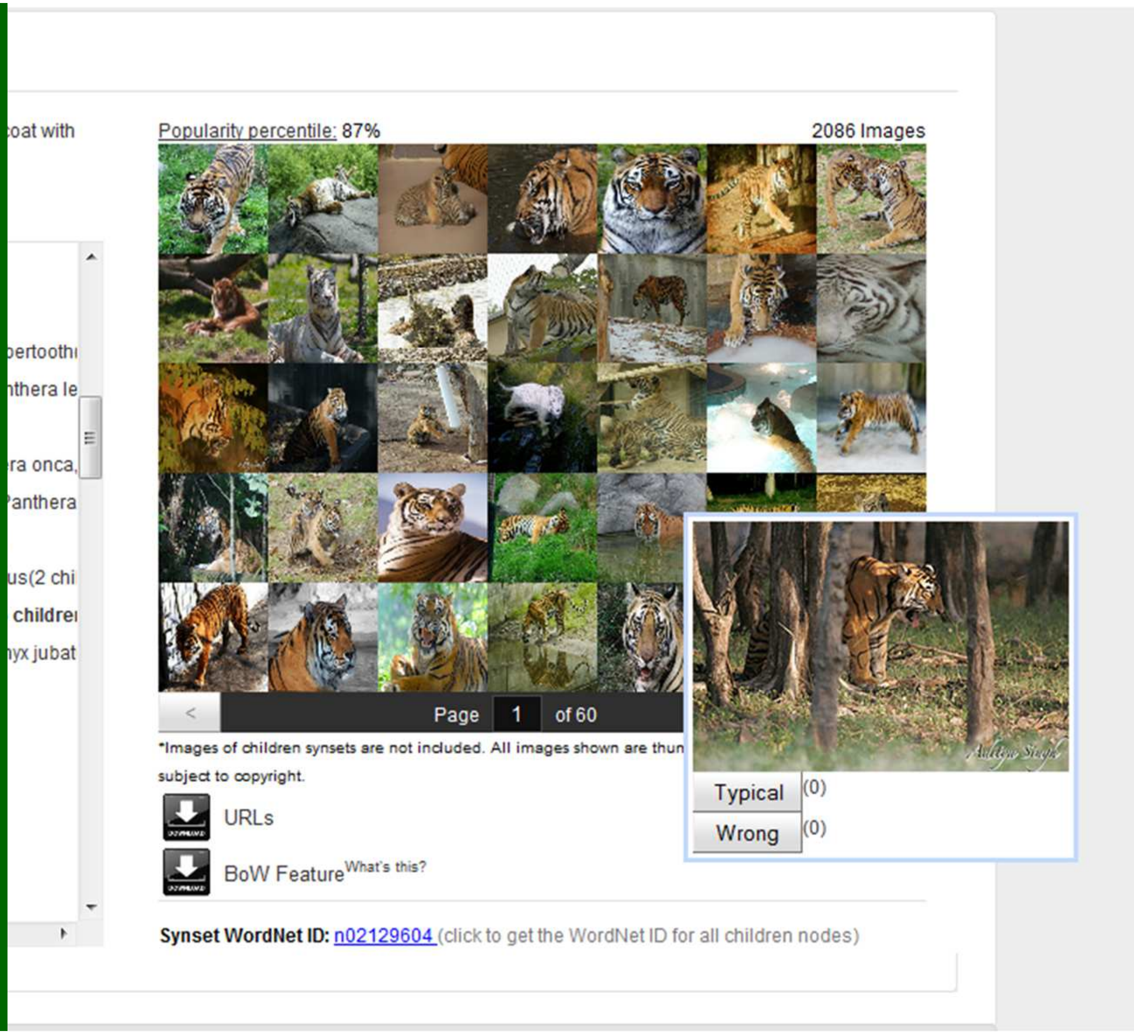
What do these images have in common? *Find out!*

ImageNet 2010 Spring Release is up! Click here to check out what's new!

IMAGENET in a glance

15K categories; 11+million images; ~800im/categ; free to public at **www.image-net.org**

- Animals
  - Birds
  - Fish
  - Mammal
  - Invertebrate
- Scenes
  - Indoor
  - Geological formations
- Sport activities
- Materials and fabric
- Instrumentation
  - Tools
  - Appliances
  - …
- Plants
  - …



IMAGENET is a knowledge ontology

- Taxonomy

- S: (n) Eskimo dog, **husky** (breed of heavy-coated Arctic sled dog)
  - *direct hypernym* / ***inherited hypernym*** / *sister term*
    - S: (n) working dog (any of several breeds of usually large powerful dogs bred to work as draft animals and guard and guide dogs)
      - S: (n) dog, domestic dog, Canis familiaris (a member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds) *"the dog barked all night"*
        - S: (n) canine, canid (any of various fissiped mammals with nonretractile claws and typically long muzzles)
          - S: (n) carnivore (a terrestrial or aquatic flesh-eating mammal) *"terrestrial carnivores have four or five clawed digits on each limb"*
            - S: (n) placental, placental mammal, eutherian, eutherian mammal (mammals having a placenta; all mammals except monotremes and marsupials)
              - S: (n) mammal, mammalian (any warm-blooded vertebrate having the skin more or less covered with hair; young are born alive except for the small subclass of monotremes and nourished with milk)
                - S: (n) vertebrate, craniate (animals having a bony or cartilaginous skeleton with a segmented spinal column and a large brain enclosed in a skull or cranium)
                  - S: (n) chordate (any animal of the phylum Chordata having a notochord or spinal column)
                    - S: (n) animal, animate being, beast, brute, creature, fauna (a living organism characterized by voluntary movement)
                      - S: (n) organism, being (a living thing that has (or can develop) the ability to act or function independently)
                        - S: (n) living thing, animate thing (a living (or once living) entity)
                          - S: (n) whole, unit (an assemblage of parts that is regarded as a single entity) *"how big is that part compared to the whole?"; "the team is a unit"*
                            - S: (n) object, physical object (a tangible and visible entity; an entity that can cast a shadow) *"it was full of rackets, balls and other objects"*
                              - S: (n) physical entity (an entity that has physical existence)
                                - S: (n) entity (that which is perceived or known or inferred to have its own distinct existence (living or nonliving))

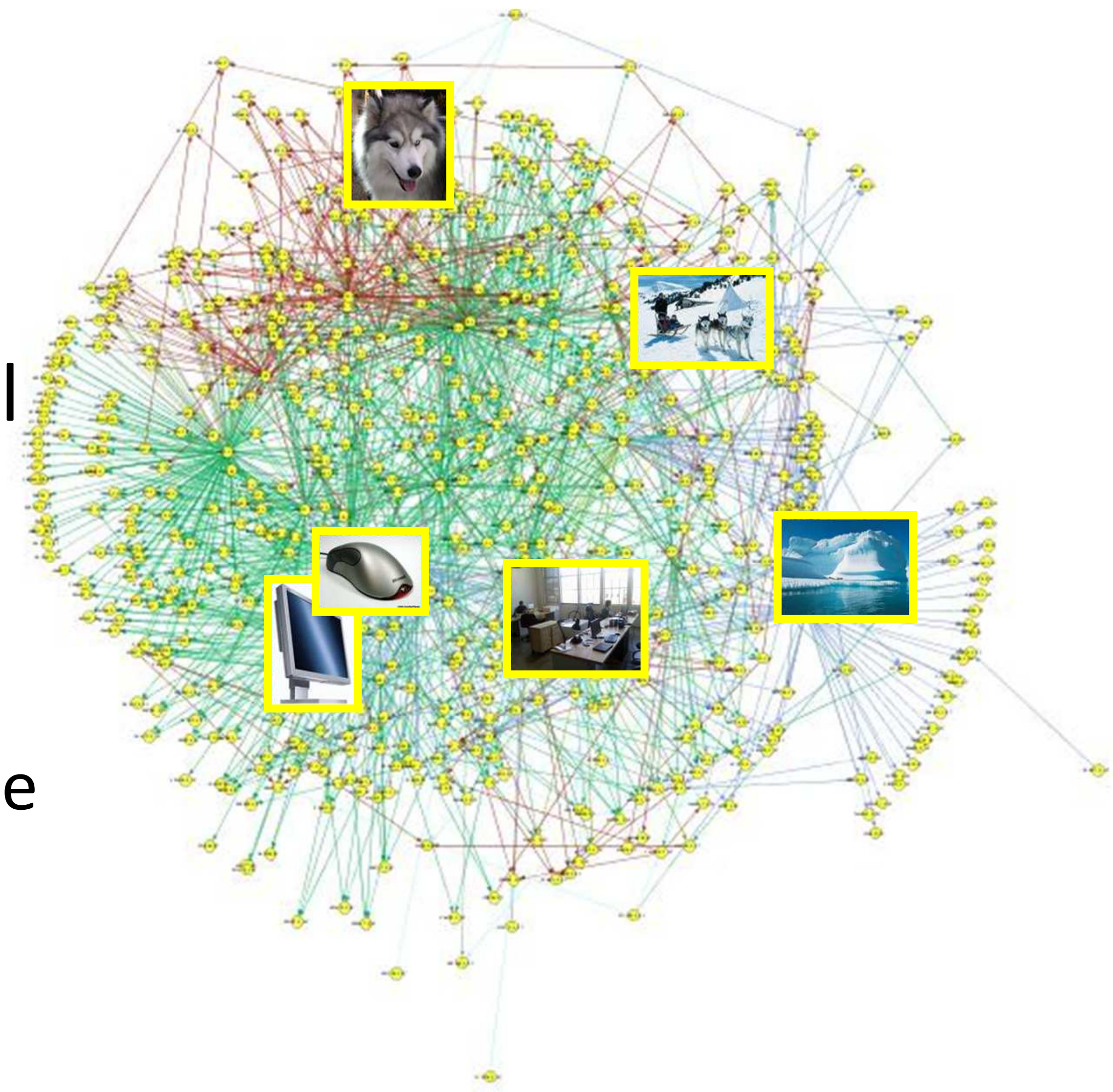
IMAGENET is a knowledge ontology

- Taxonomy
- Partonomy

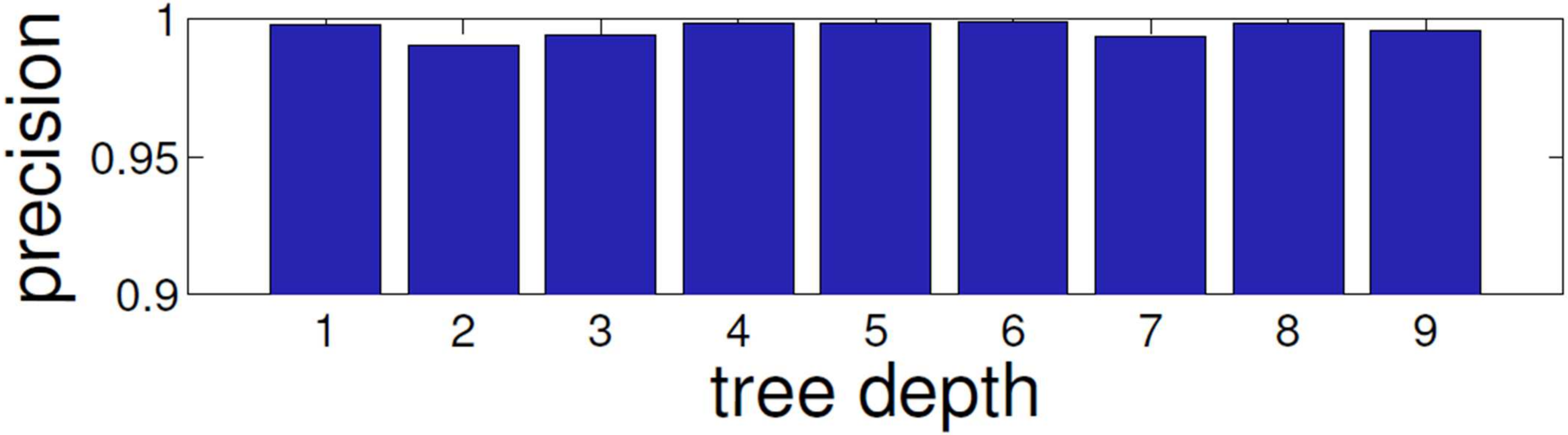
- S: (n) **car**, auto, automobile, machine, motorcar (a motor vehicle with four wheels; usually propelled by an internal combustion engine) *"he needs a car to get to work"*
  - *direct hyponym* / *full hyponym*
  - ***part meronym***
    - S: (n) accelerator, accelerator pedal, gas pedal, gas, throttle, gun (a pedal that controls the throttle valve) *"he stepped on the gas"*
    - S: (n) air bag (a safety restraint in an automobile; the bag inflates on collision and prevents the driver or passenger from being thrown forward)
    - S: (n) auto accessory (an accessory for an automobile)
    - S: (n) automobile engine (the engine that propels an automobile)
    - S: (n) automobile horn, car horn, motor horn, horn, hooter (a device on an automobile for making a warning noise)
    - S: (n) buffer, fender (a cushion-like device that reduces shock due to an impact)
    - S: (n) bumper (a mechanical device consisting of bars at either end of a vehicle to absorb shock and prevent serious damage)
    - S: (n) car door (the door of a car)
    - S: (n) car mirror (a mirror that the driver of a car can use)
    - S: (n) car seat (a seat in a car)
    - S: (n) car window (a window in a car)
    - S: (n) fender, wing (a barrier that surrounds the wheels of a vehicle to block splashing water or mud) *"in Britain they call a fender a wing"*
    - S: (n) first gear, first, low gear, low (the lowest forward gear ratio in the gear box of a motor vehicle; used to start a car moving)
    - S: (n) floorboard (the floor of an automobile)
    - S: (n) gasoline engine, petrol engine (an internal-combustion engine that burns gasoline; most automobiles are driven by gasoline engines)
    - S: (n) glove compartment (compartment on the dashboard of a car)
    - S: (n) grille, radiator grille (grating that admits cooling air to car's radiator)
    - S: (n) high gear, high (a forward gear with a gear ratio that gives the greatest vehicle velocity for a given engine speed)
    - S: (n) hood, bonnet, cowl, cowling (protective covering consisting of a metal part that covers the engine) *"there are powerful engines under the hoods of new cowling in order to repair the plane's engine"*
    - S: (n) luggage compartment, automobile trunk, trunk (compartment in an automobile that carries luggage or shopping or tools) *"he put his golf bag in the trunk*
    - S: (n) rear window (car window that allows vision out of the back of the car)
    - S: (n) reverse, reverse gear (the gears by which the motion of a machine can be reversed)
    - S: (n) roof (protective covering on top of a motor vehicle)
    - S: (n) running board (a narrow footboard serving as a step beneath the doors of some old cars)
    - S: (n) stabilizer bar, anti-sway bar (a rigid metal bar between the front suspensions and between the rear suspensions of cars and trucks; serves to stabilize the ch
    - S: (n) sunroof, sunshine-roof (an automobile roof having a sliding or raisable panel) *"'sunshine-roof' is a British term for 'sunroof'"*
    - S: (n) tail fin, tailfin, fin (one of a pair of decorations projecting above the rear fenders of an automobile)
    - S: (n) third gear, third (the third from the lowest forward ratio gear in the gear box of a motor vehicle) *"you shouldn't try to start in third gear"*
    - S: (n) window (a transparent opening in a vehicle that allow vision out of the sides or back; usually is capable of being opened)

IMAGENET is a knowledge ontology

- Taxonomy
- Partonomy
- The "social network" of visual concepts
  - Prior knowledge
  - Context
  - Hidden knowledge and structure among visual concepts



IMAGENET has high accuracy



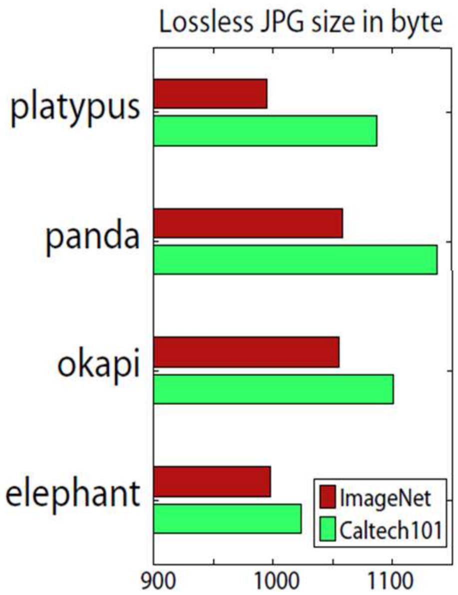
e.g. mammal

e.g. dog

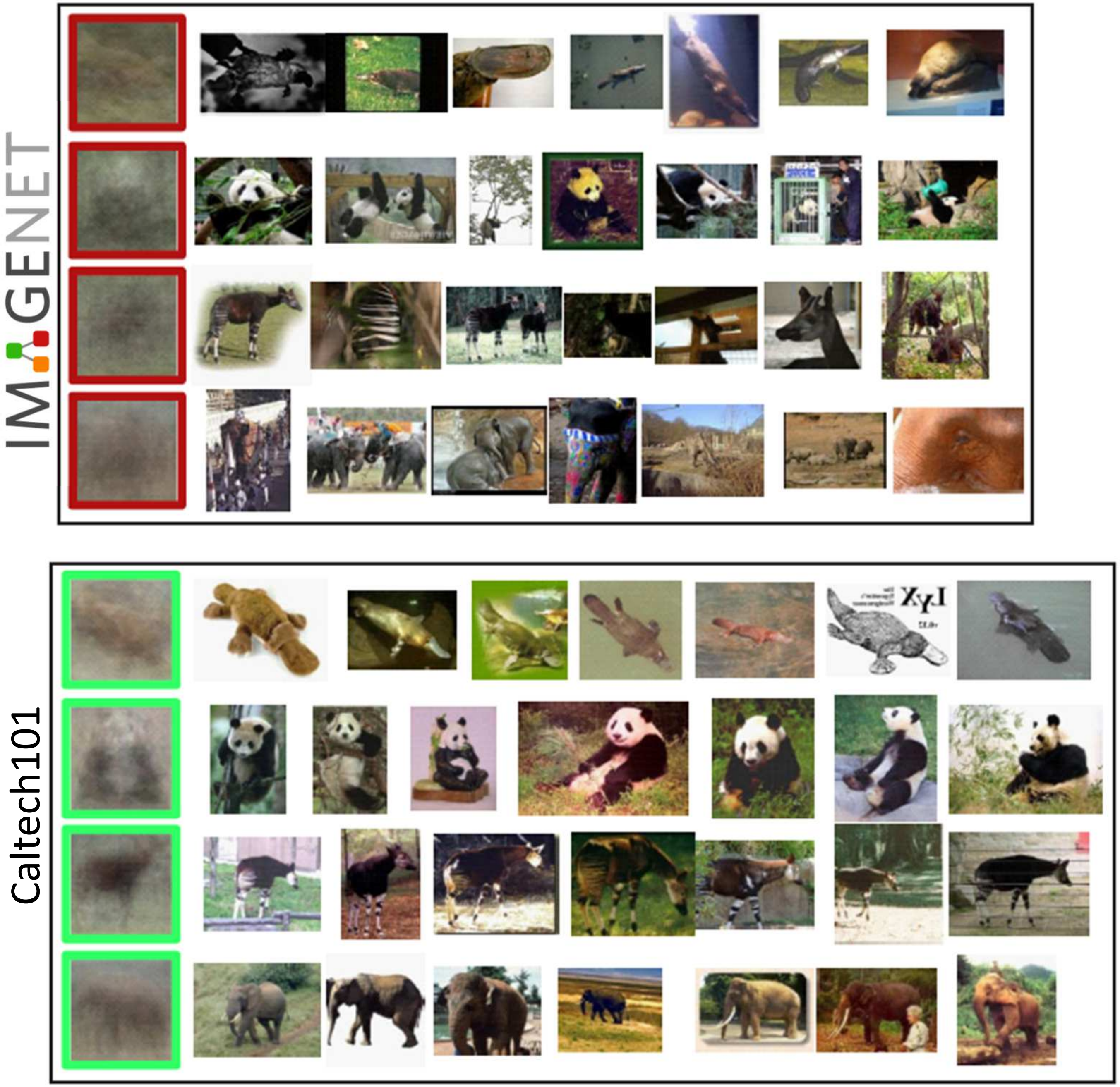
e.g. German Shepherd

Deng, Dong, Socher, Li, Li, & Fei-Fei, *CVPR*, 2009

Diverse

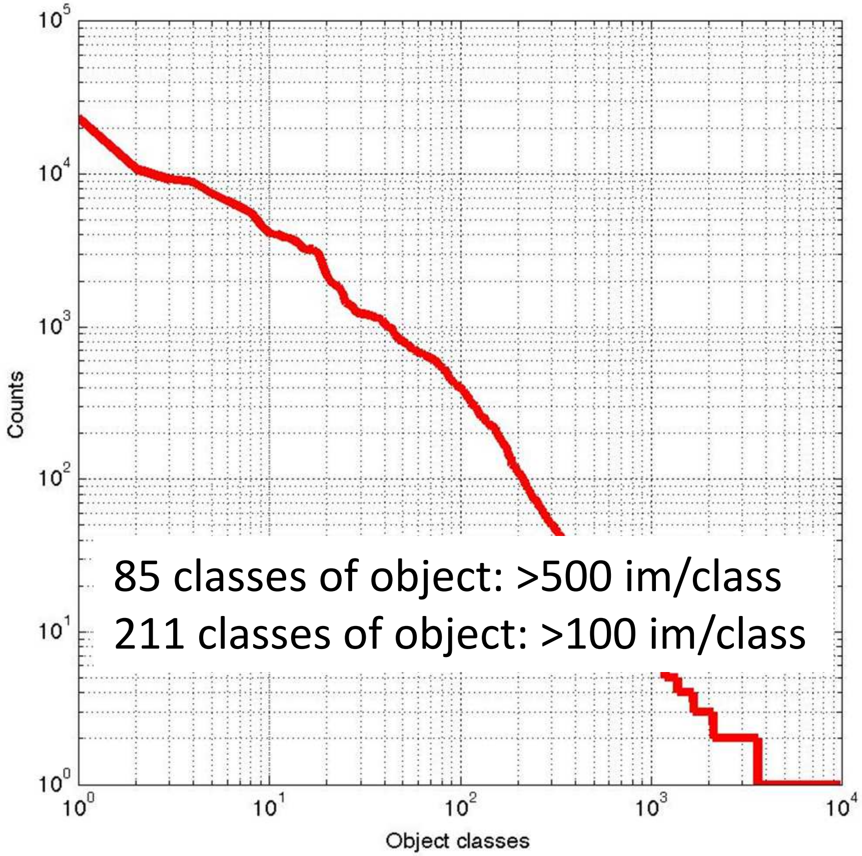


Lossless JPG size in byte
platypus
panda
okapi
elephant
ImageNet
Caltech101
900
1000
1100



IMAGENET
Caltech101

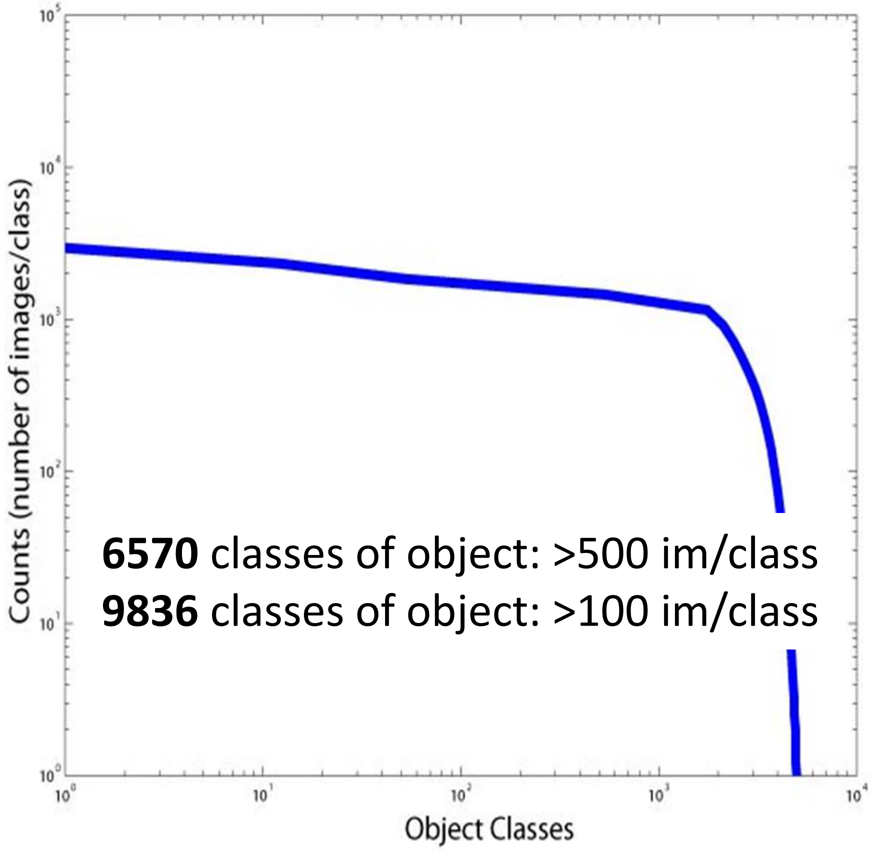
IMAGENET is large scale



85 classes of object: >500 im/class
211 classes of object: >100 im/class

LabelMe

Russell et al. 2005;
statistics obtained in 2009



6570 classes of object: >500 im/class
9836 classes of object: >100 im/class

IMAGENET

Deng et al. 2009
statistics obtained in 2009

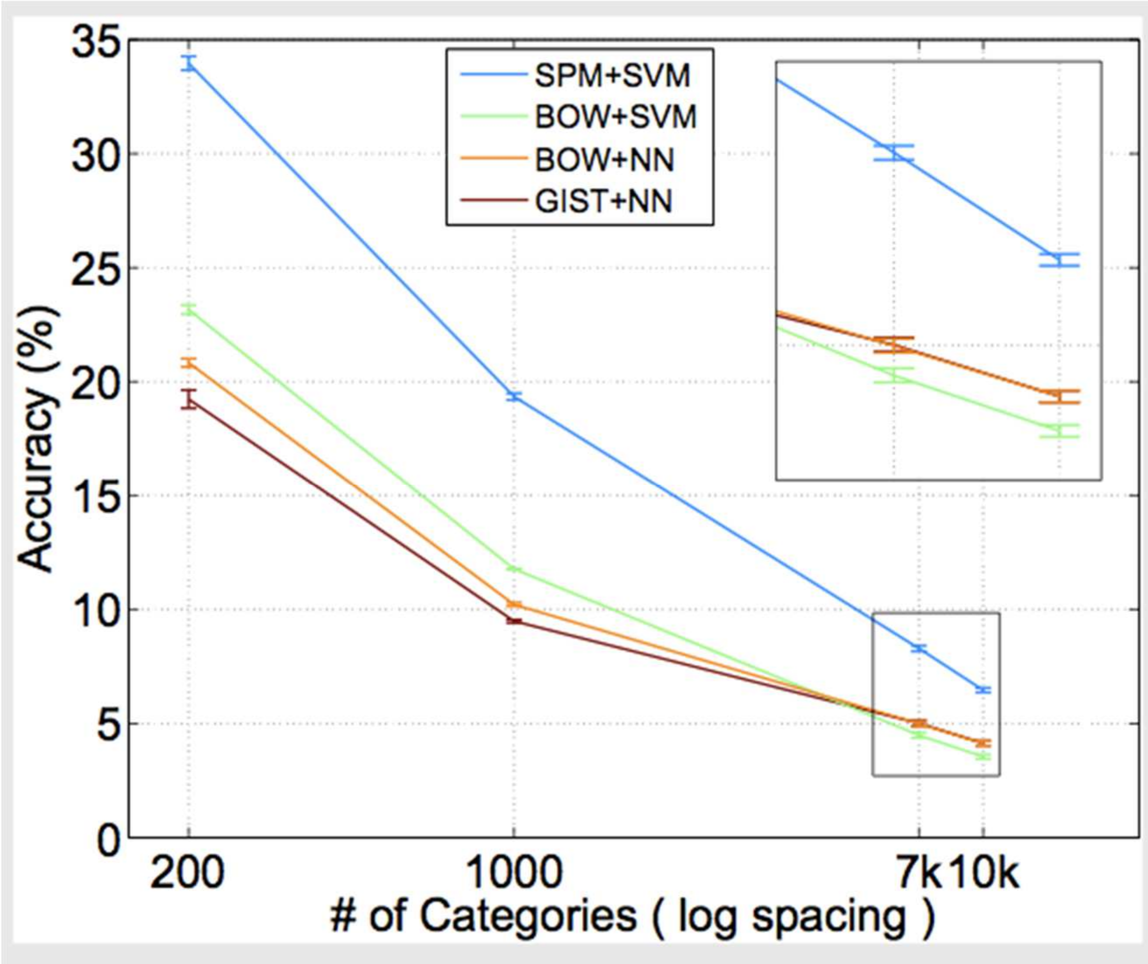
Outline

- ImageNet dataset
  - Properties of ImageNet
- ECCV2010: a 10000-way classification benchmark experiment
  - Size matters
  - Density matters
  - Hierarchy matters
- An “infallible” classifier

What does classifying more than 10,000 image categories tell us?

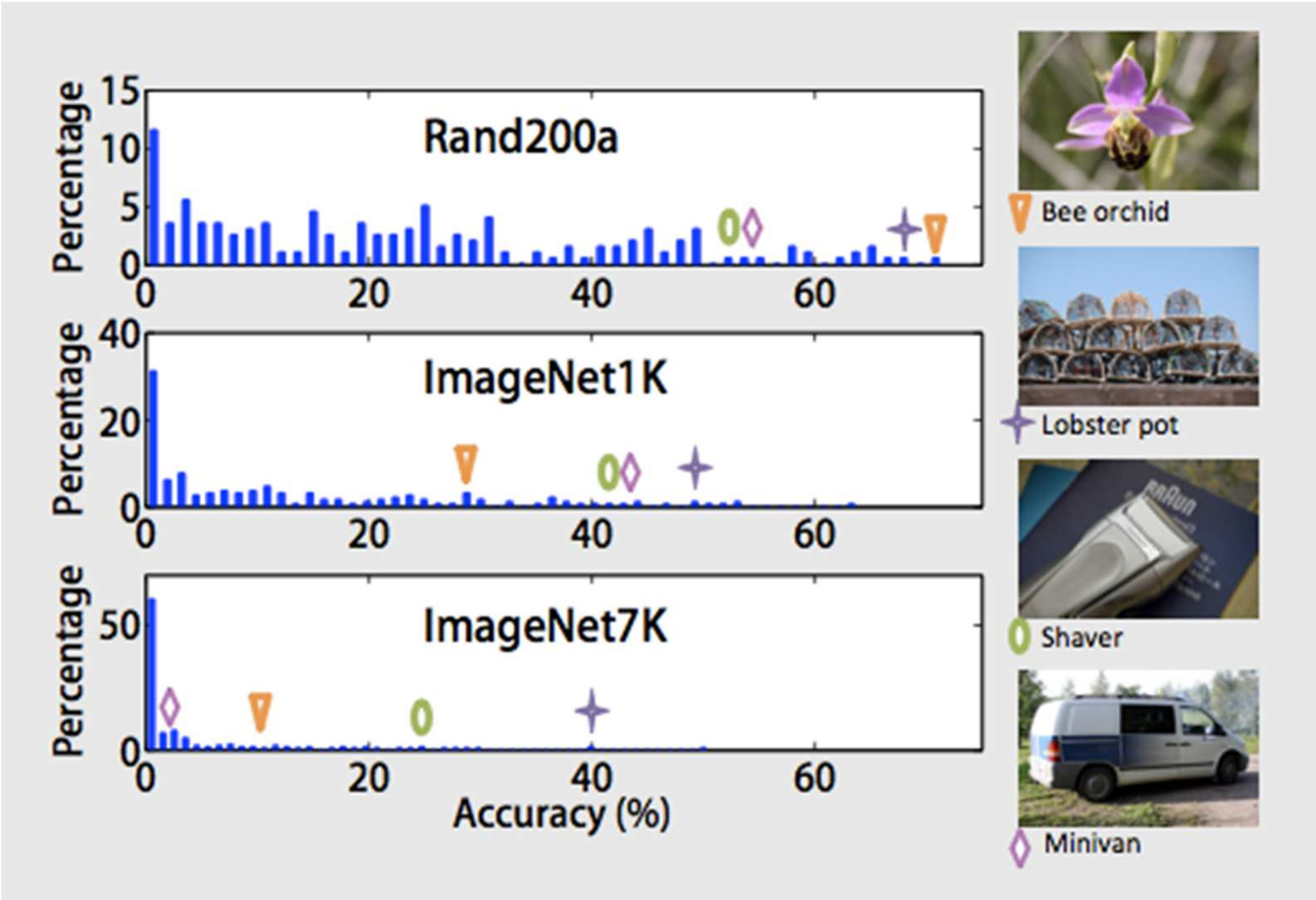
J. Deng, A.C. Berg, K. Li, L. Fei-Fei, ECCV 2010

Size matters.



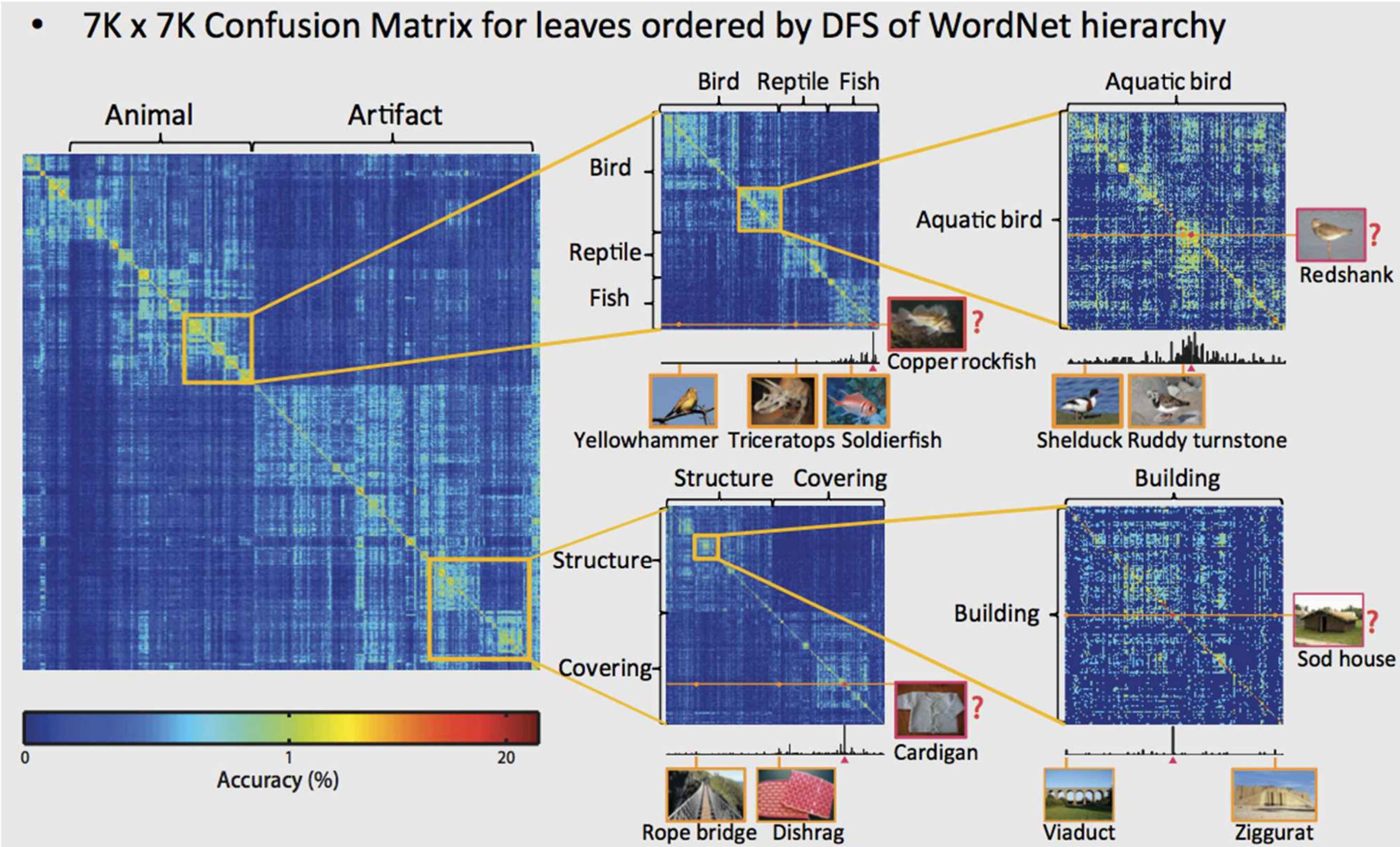
What does classifying more than 10,000 image categories tell us?

J. Deng, A.C. Berg, K. Li, L. Fei-Fei, ECCV 2010



What does classifying more than 10,000 image categories tell us?

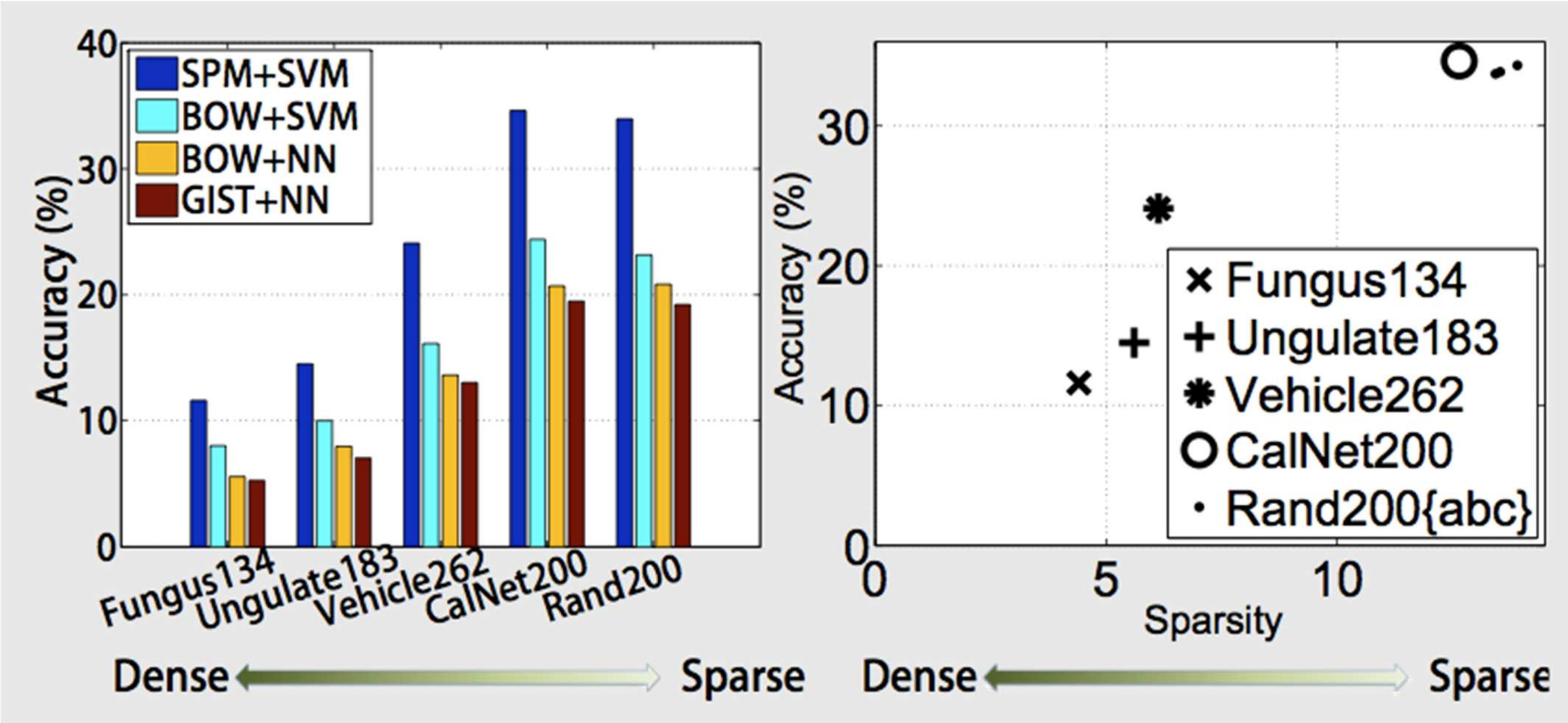
J. Deng, A.C. Berg, K. Li, L. Fei-Fei, ECCV 2010



What does classifying more than 10,000 image categories tell us?

J. Deng, A.C. Berg, K. Li, L. Fei-Fei, ECCV 2010

Density matters.



Evidence of correlation between semantic distance and difficulty of visual recognitioı

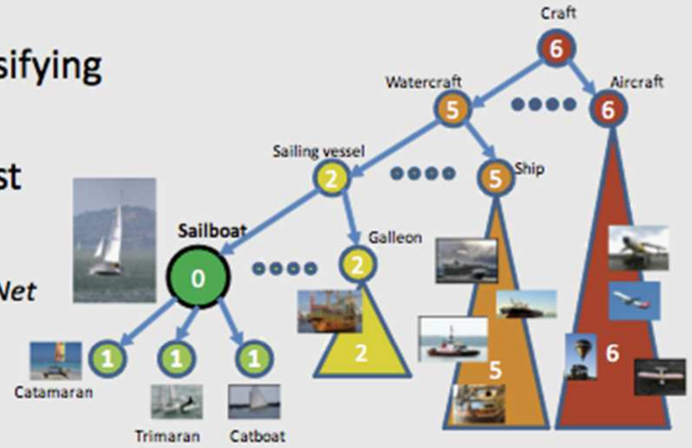
What does classifying more than 10,000 image categories tell us?

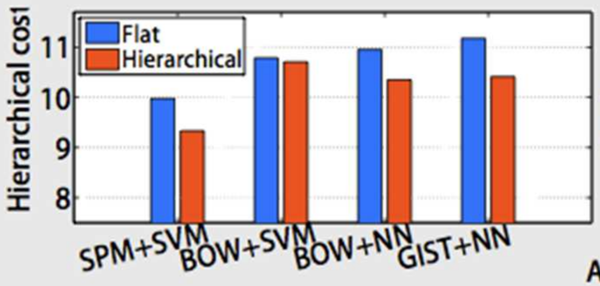
J. Deng, A.C. Berg, K. Li, L. Fei-Fei, ECCV 2010

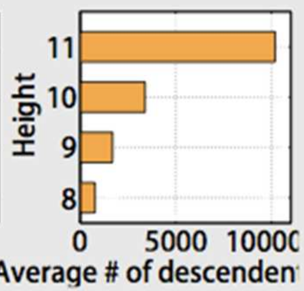
Hierarchy matters.

- Classifying a "dog" as "cat" is probably not as bad as classifying it as "microwave"
- A simple way to incorporate hierarchical classification cost

$$C_{i,j} = \begin{cases} 0 & i{=}j, \text{ or } i \text{ is a descendent of } j \\ h(i,j) & h \text{ is the height of the lowest common ancestor in WordNet} \end{cases}$$



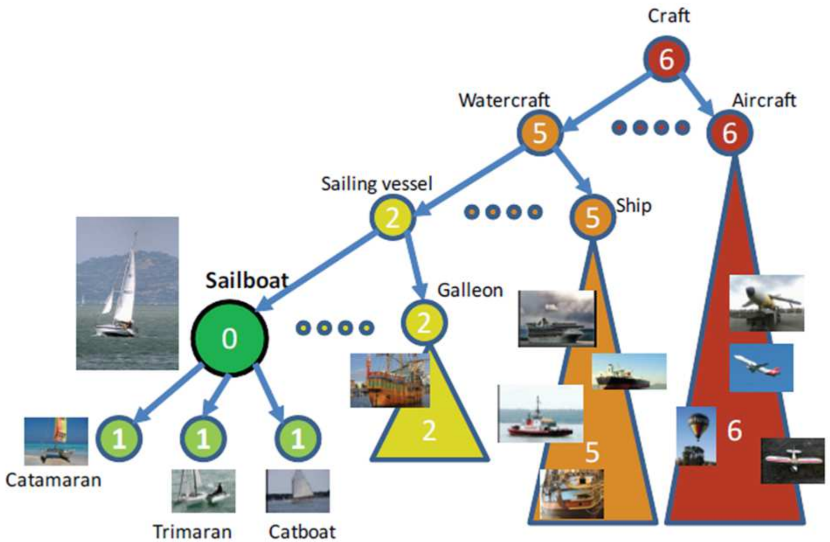


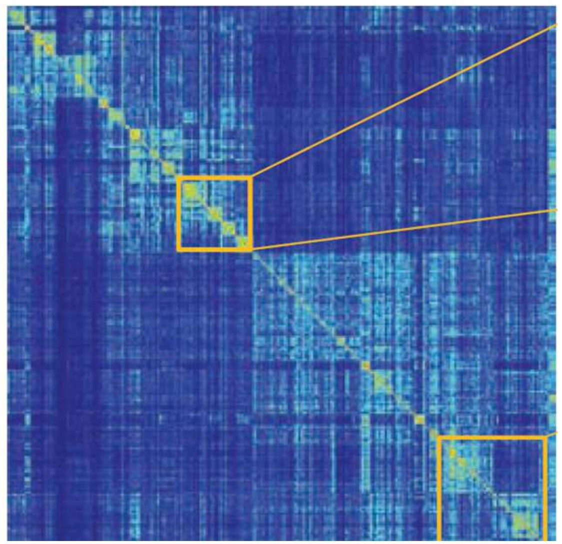


- Cost sensitive classification leads to more informative results

Take away messages from benchmarking

- Hierarchical structure is useful
- There is a long way to go before we achieve high accuracy recognition





Current 'state-of-the-art'

Google goggles labs

- ✓ unique landmarks
- ✓ book/CD/magazine covers
- ✓ paintings
- ✓ logos (in good conditions)

Current ‘state-of-the-art’

Google goggles labs

- ✓ u
- ✓ b s
- ✓ p
- ✓ l

✗ What animal is this?

Current 'state-of-the-art'

Google goggles labs

- ✓ uniq
- ✓ book
- ✓ paint
- ✓ logos

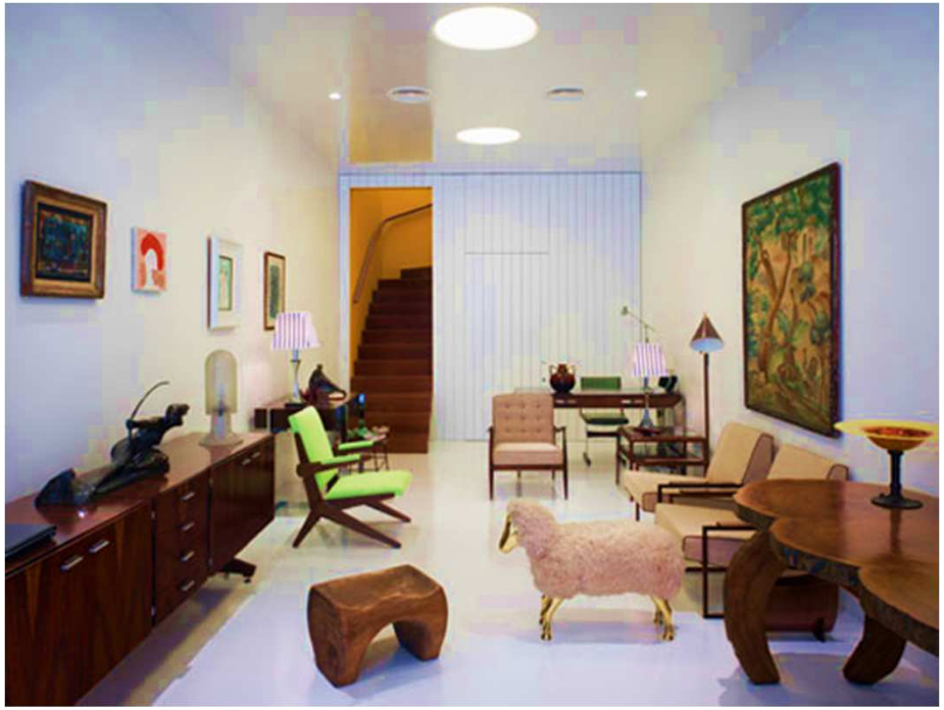
- ✗ What animal is this?
- ✗ Tell me the brand of her purse?

Current ‘state-of-the-art’

Google goggles labs

- ✘ What animal is this?
- ✘ Tell me the brand of this purse?
- ✘ Is this edible?

Current 'state-of-the-art'



- ✘ What animal is this?
- ✘ Tell me the brand of this purse?
- ✘ Is this edible?
- ✘ Point out all the chairs in this room?

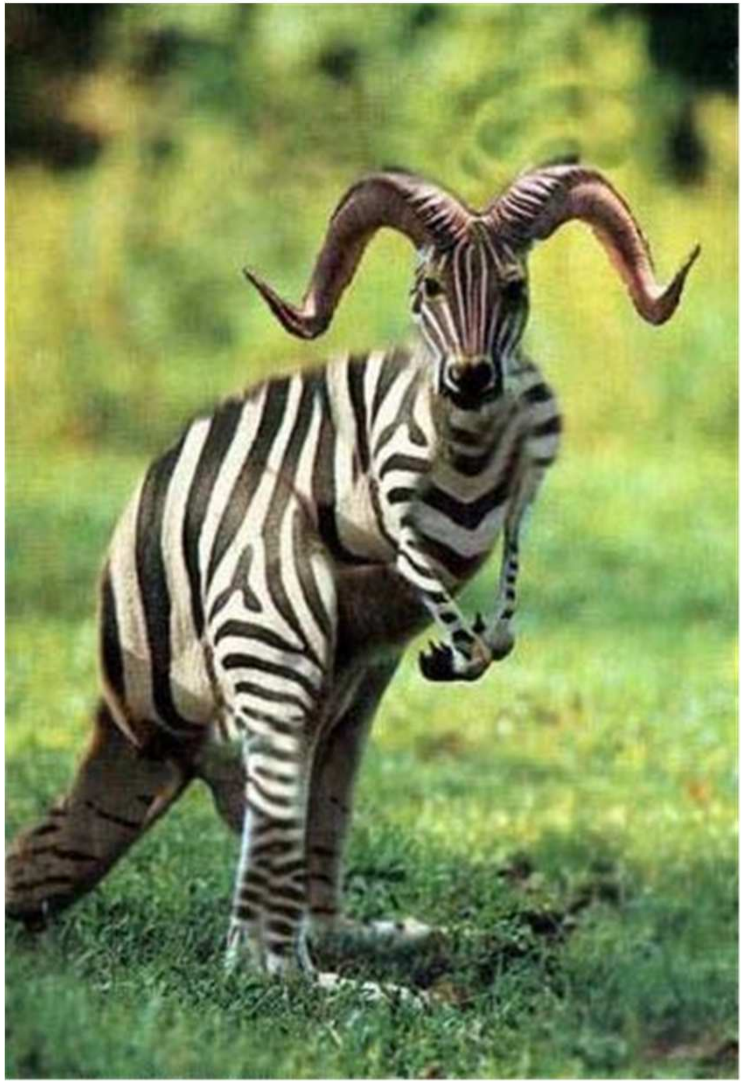
Outline

- ImageNet dataset
  - Properties of ImageNet
- ECCV2010: a 10000-way classification benchmark experiment
  - Size matters
  - Density matters
  - Hierarchy matters
- An “infallible” classifier

A Brain Teaser: What is this?

<u>Kangaroo</u> ✓

A Brain Teaser: What is this?



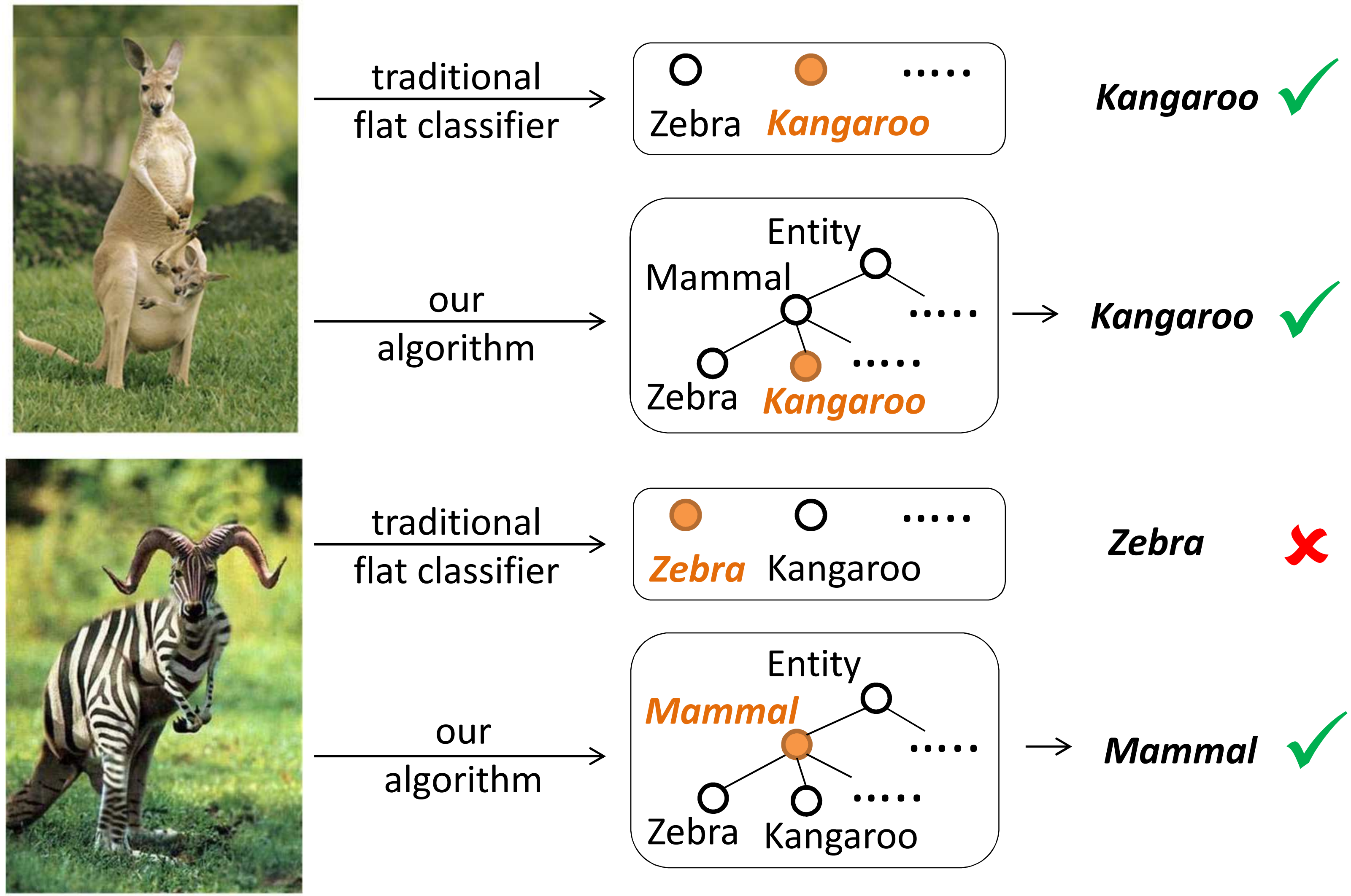
Kangaroo ✗

Zebra ✗

Mammal ✓

Entity ✓

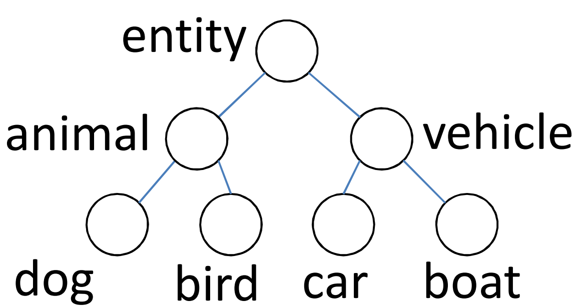
A more serious question:
How do we build a recognition system that almost never make mistakes, but is as informative as possible?



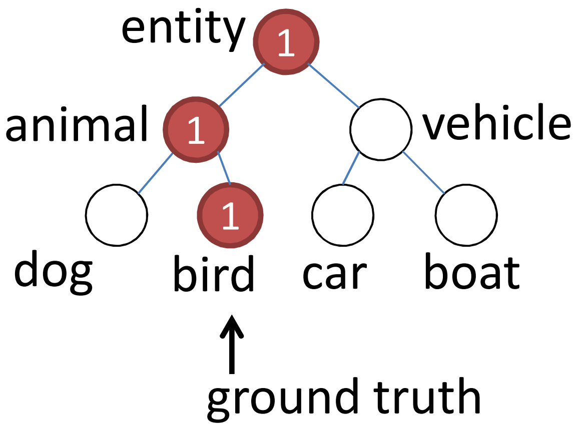
Deng, Krause, Berg, & Fei-Fei, *submitted*

Problem Setup

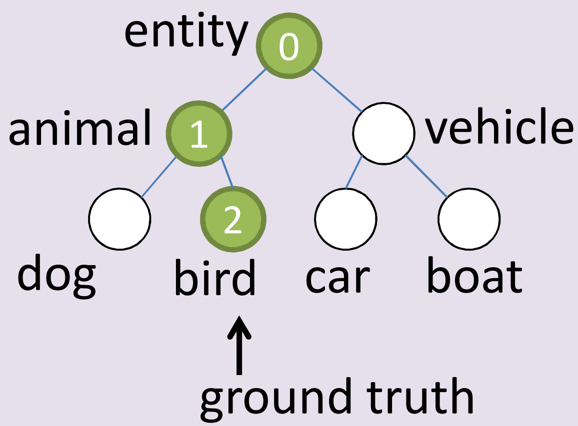
Semantic hierarchy



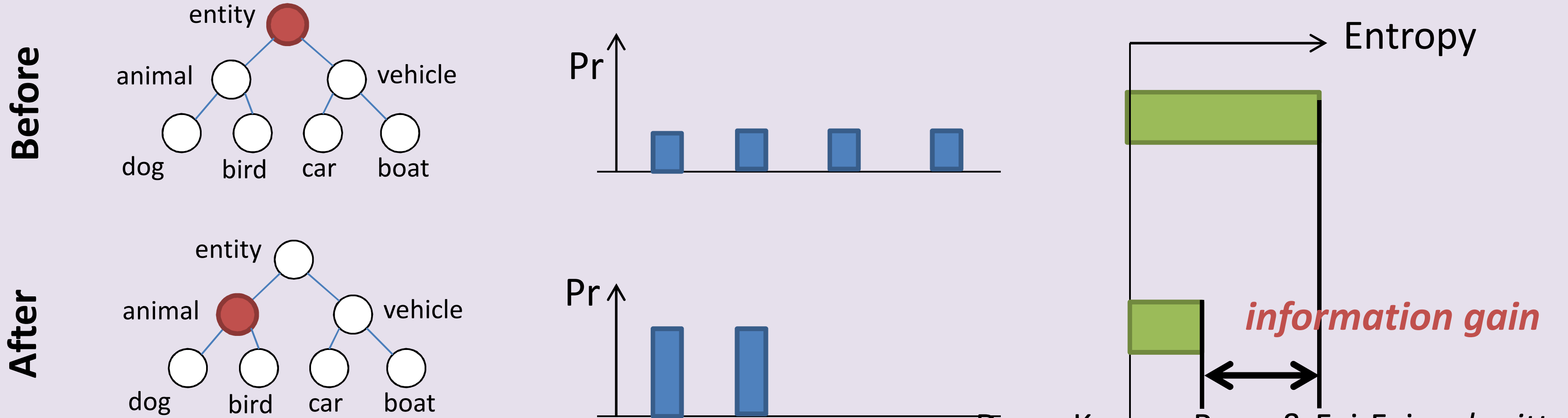
Accuracy



Reward



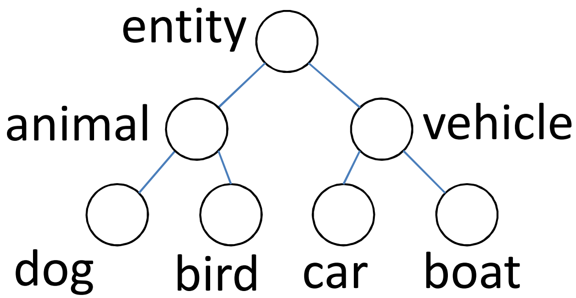
Reward: amount of correct *information gain* *(i.e. decrease of uncertainty)*



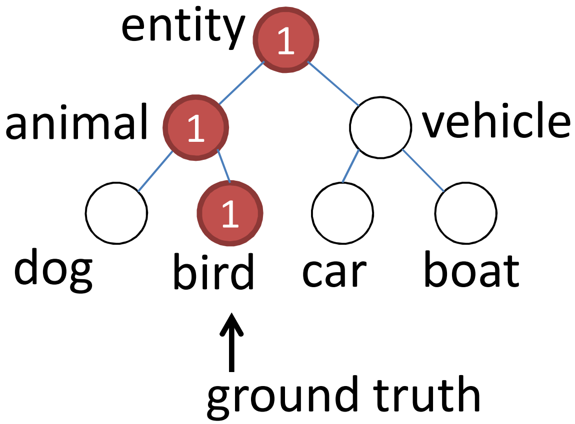
Deng, Krause, Berg, & Fei-Fei, *submitted*

Problem Setup

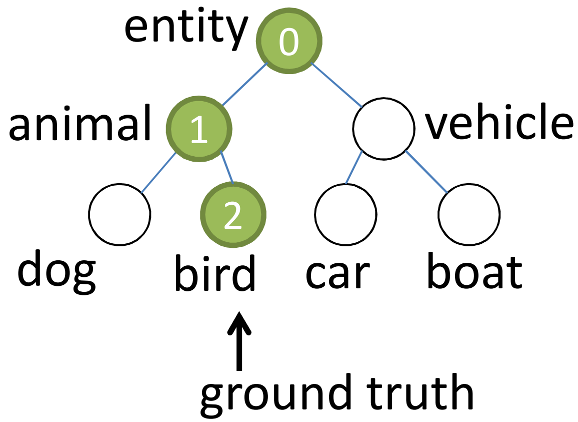
Semantic hierarchy



Accuracy



Reward



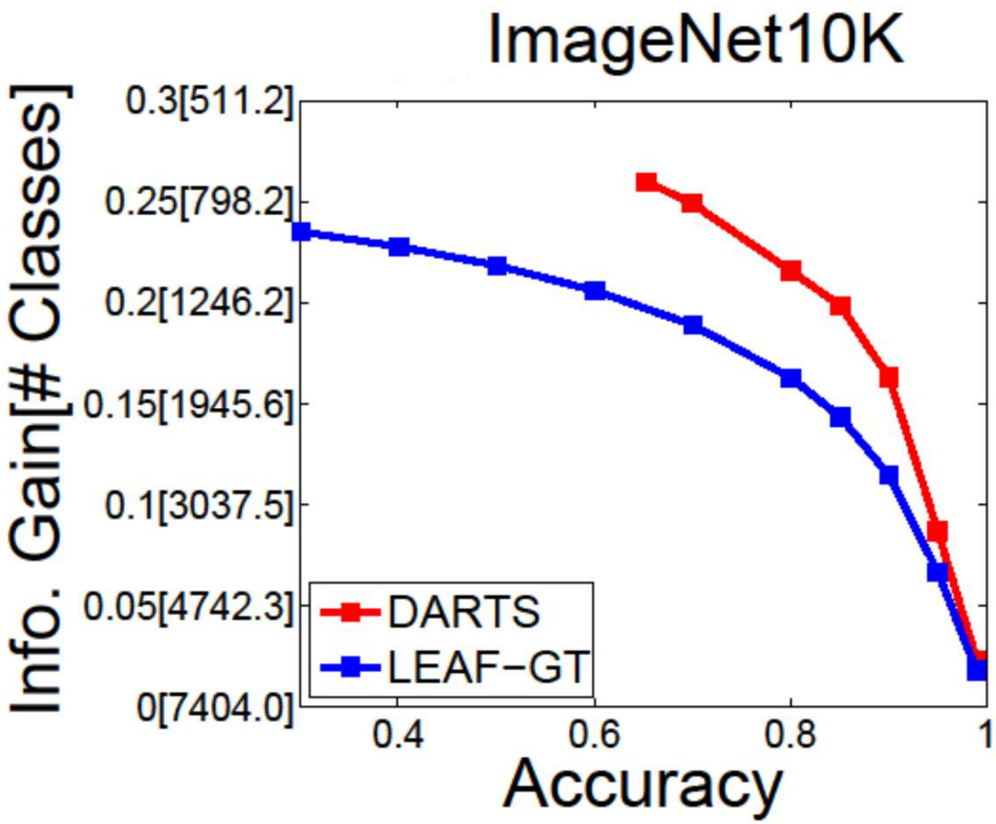
Classifier f, Accuracy $\Phi(f)$, Reward $R(f)$, Accuracy guarantee $1-\epsilon$

$$\begin{aligned} \underset{f}{\text{maximize}} \quad & R(f) \\ \text{Subject to} \quad & \Phi(f) \geq 1-\epsilon \end{aligned}$$

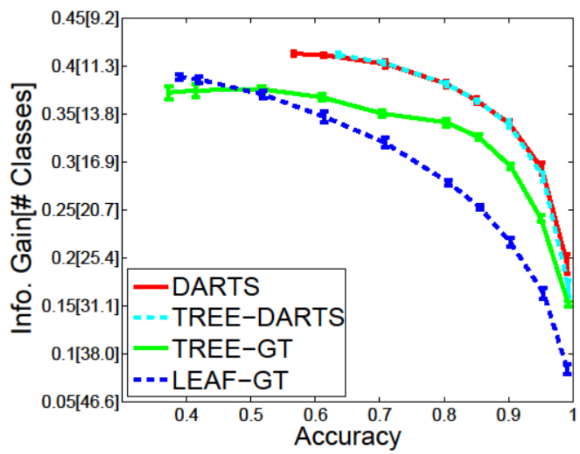
Deng, Krause, Berg, & Fei-Fei, *submitted*

Results

<u>Datasets:</u> 10,000 image classes from ImageNet (~9million images)
<u>Baselines:</u> Flat classifier and decision tree with thresholding



More comparisons (65 classes)



Deng, Krause, Berg, & Fei-Fei, *submitted*

Some examples: flat classifier vs. high fidelity

red fox

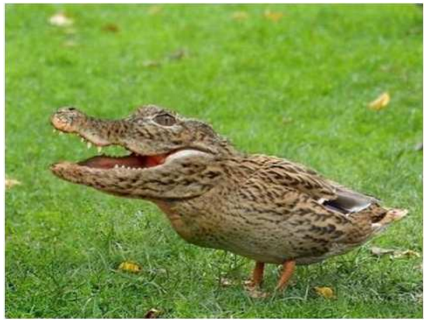
Flat	hyena	Egyptian cat	orangutan	mantis	jelly fungus
Ours	**canine**	**carnivore**	**mammal**	**animal**	**living thing**

trimaran

Flat	catamaran	submarine	airship	iron	electric guitar
Ours	**sailboat**	**watercraft**	**craft**	**artifact**	**artifact**

Deng, Krause, Berg, & Fei-Fei, *submitted*

Let's get extreme… what the heck is *this*?!

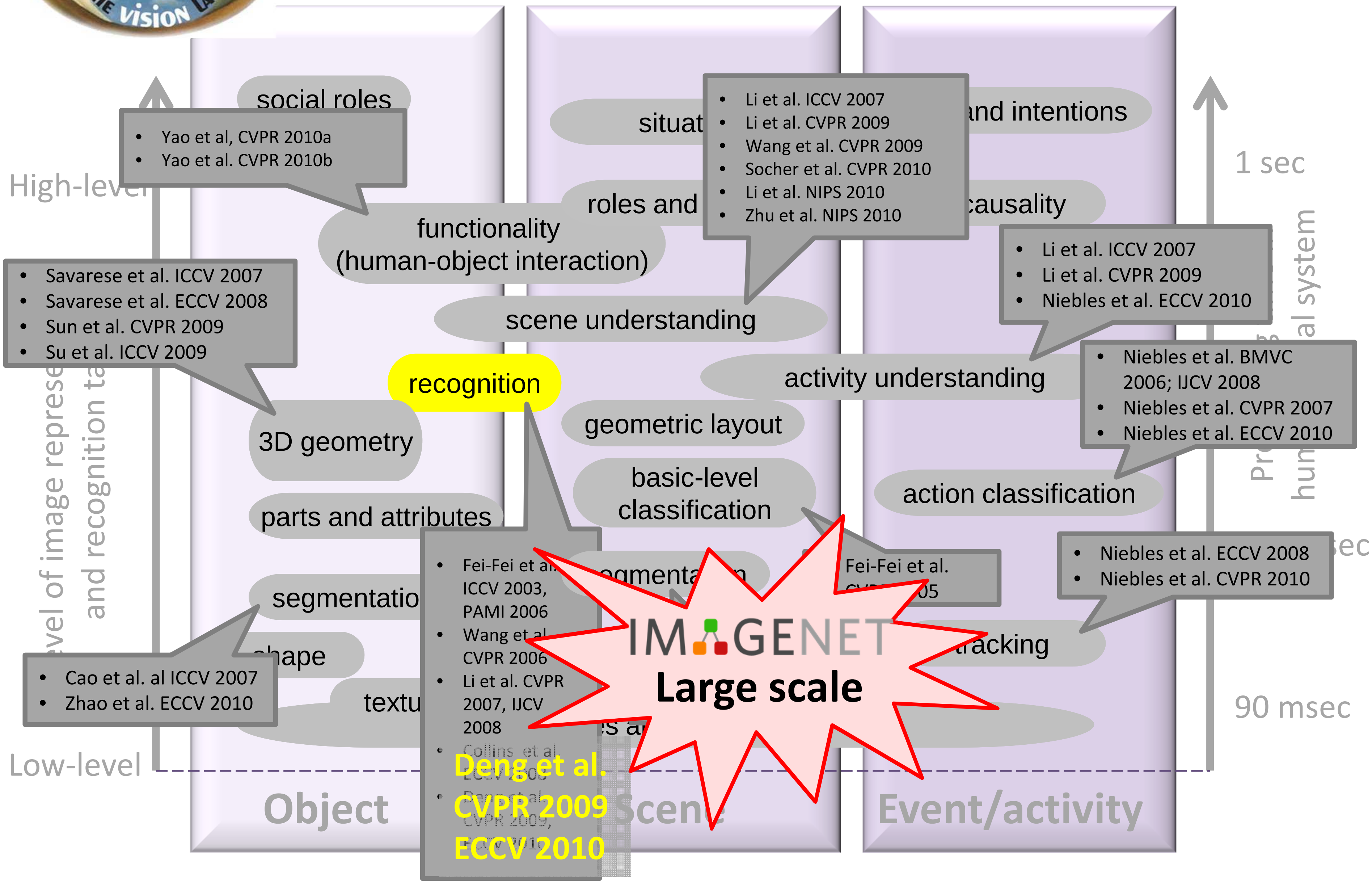
Flat	bobsled	pheasant	mortar	canoe
Ours	**vehicle**	**animal**	**edible fruit**	**watercraft**




Flat	loggerhead	cannon	Bouvier des Flandres	grapefruit
Ours	**animal**	**animal**	**living thing**	**citrus fruit**

Deng, Krause, Berg, & Fei-Fei, *submitted*

Story telling in images



Story

- Yao et al, CVPR 2010a
- Yao et al. CVPR 2010b

social roles

High-level

functiona
(human-object i

- Savarese et al. ICCV 2007
- Savarese et al. ECCV 2008
- Sun et al. CVPR 2009
- Su et al. ICCV 2009

recognitio

3D geometry

Part-based + 3D

arts and attributes

egmentatio

shape

textu

- Cao et al. al ICCV 2007
- Zhao et al. ECCV 2010

- Fei-Fei ICCV 20 PAMI 2
- Wang e CVPR 2
- Li et al. 2007, I. 2008
- Collins ECCV 2
- Deng e CVPR 2 ECCV 2

Low-level

Object

H. Su*, M. Sun*, L. Fei-Fei and S. Savarese. **Learning a dense multi-view representation for detection, viewpoint classification and synthesis of object categories.** *International Conference on Computer Vision (ICCV), 2009.*

S. Savarese and L. Fei-Fei. **3D generic object categorization, localization and pose estimation.** *IEEE Intern. Conf. in Computer Vision (ICCV).* 2007.

Hao Su
Stanford U.

Min Sun
U. Michigan

Prof. Silvio Savarese
U. Michigan

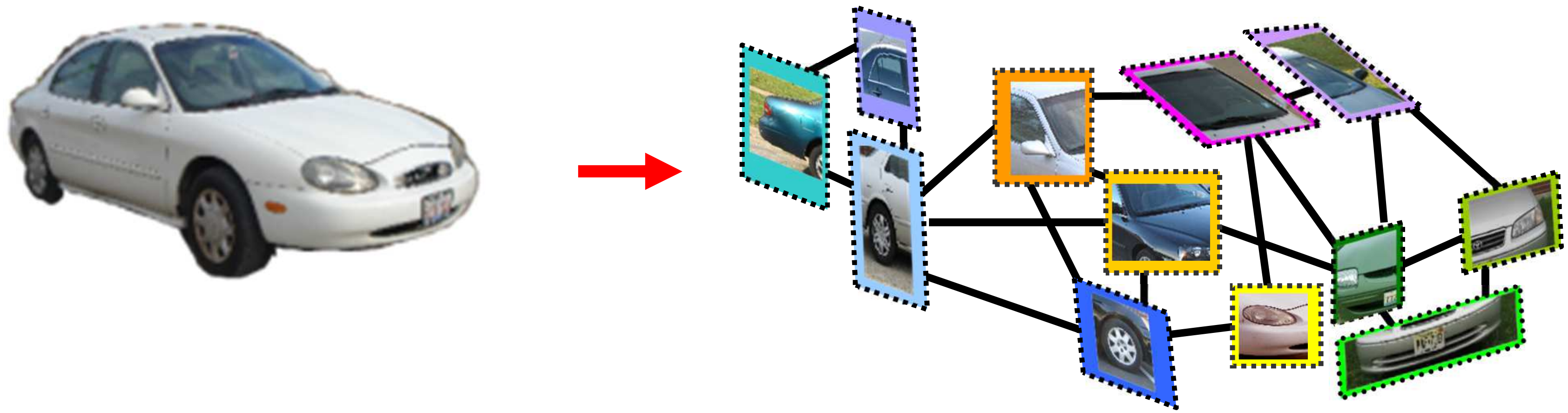
A **unified framework** for 3D object detection, pose classification, pose synthesis

Savarese, Fei-Fei, ICCV 07

Savarese, Fei-Fei, ECCV 08

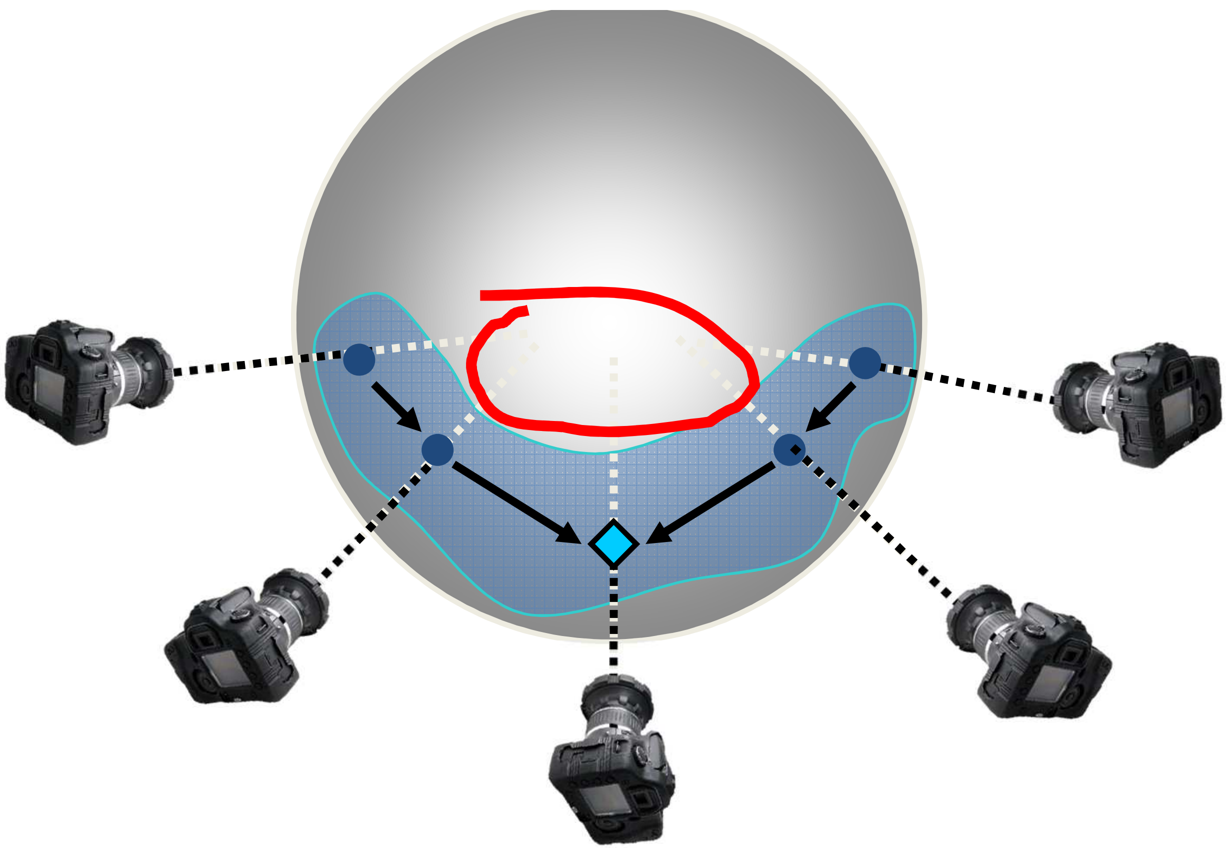
Sun, Su, Savarese, Fei-Fei, CVPR 09

Su, Sun, Fei-Fei, Savarese, ICCV 09



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

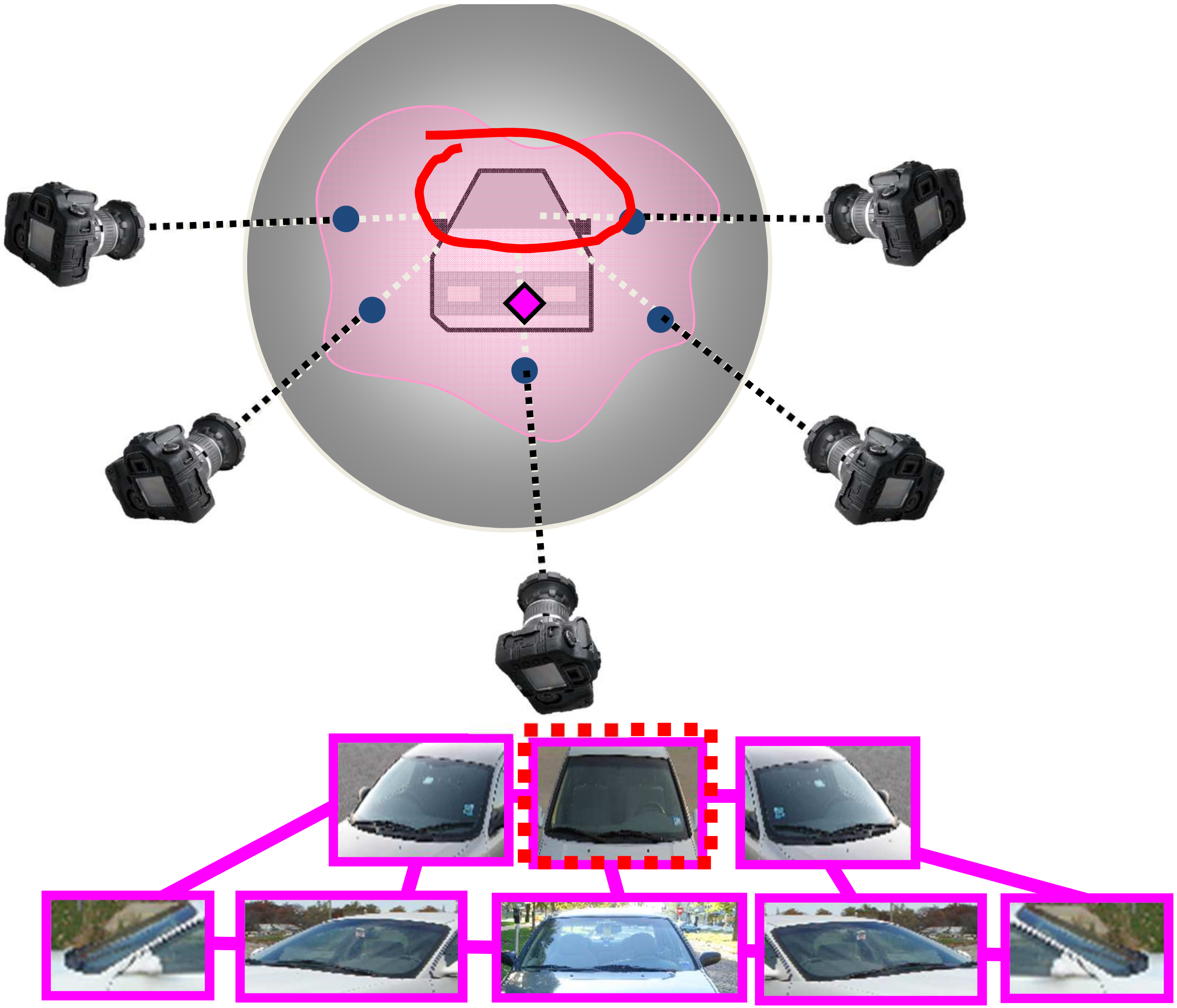
Canonical parts



- If physical part is planar, canonical part is stable point on the manifold
- Canonical part can be computed from connected component of parts

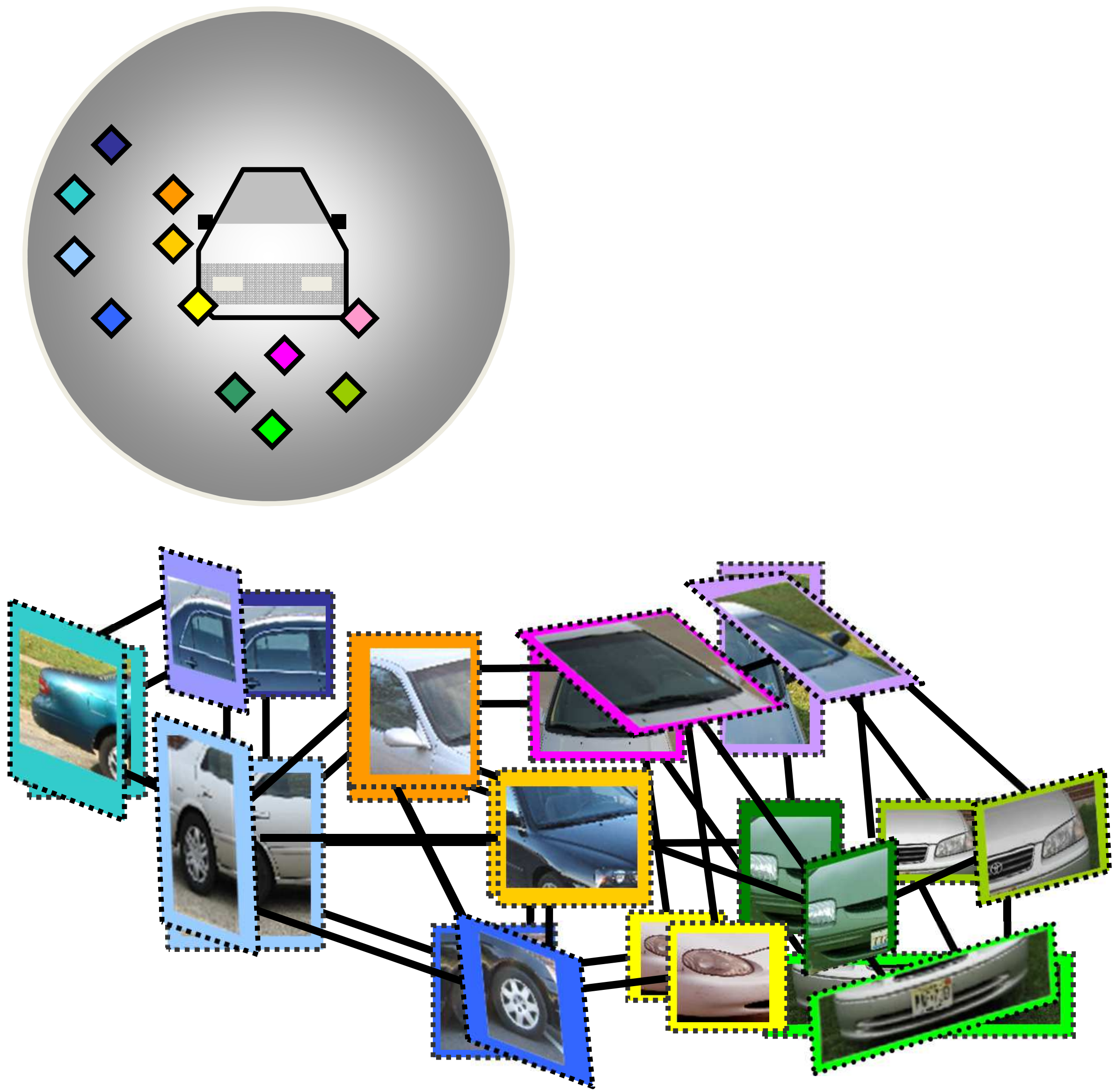
connected component of parts

Canonical parts



connected
component of parts

Linkage structure



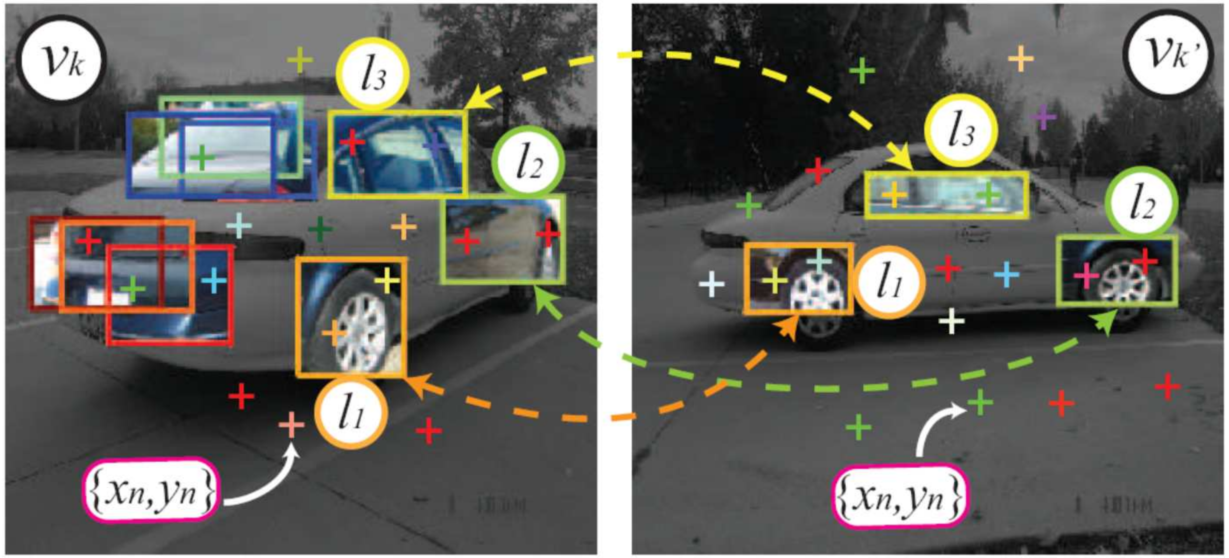
A **unified framework** for 3D object detection, pose classification, pose synthesis

Savarese, Fei-Fei, ICCV 07

Savarese, Fei-Fei, ECCV 08

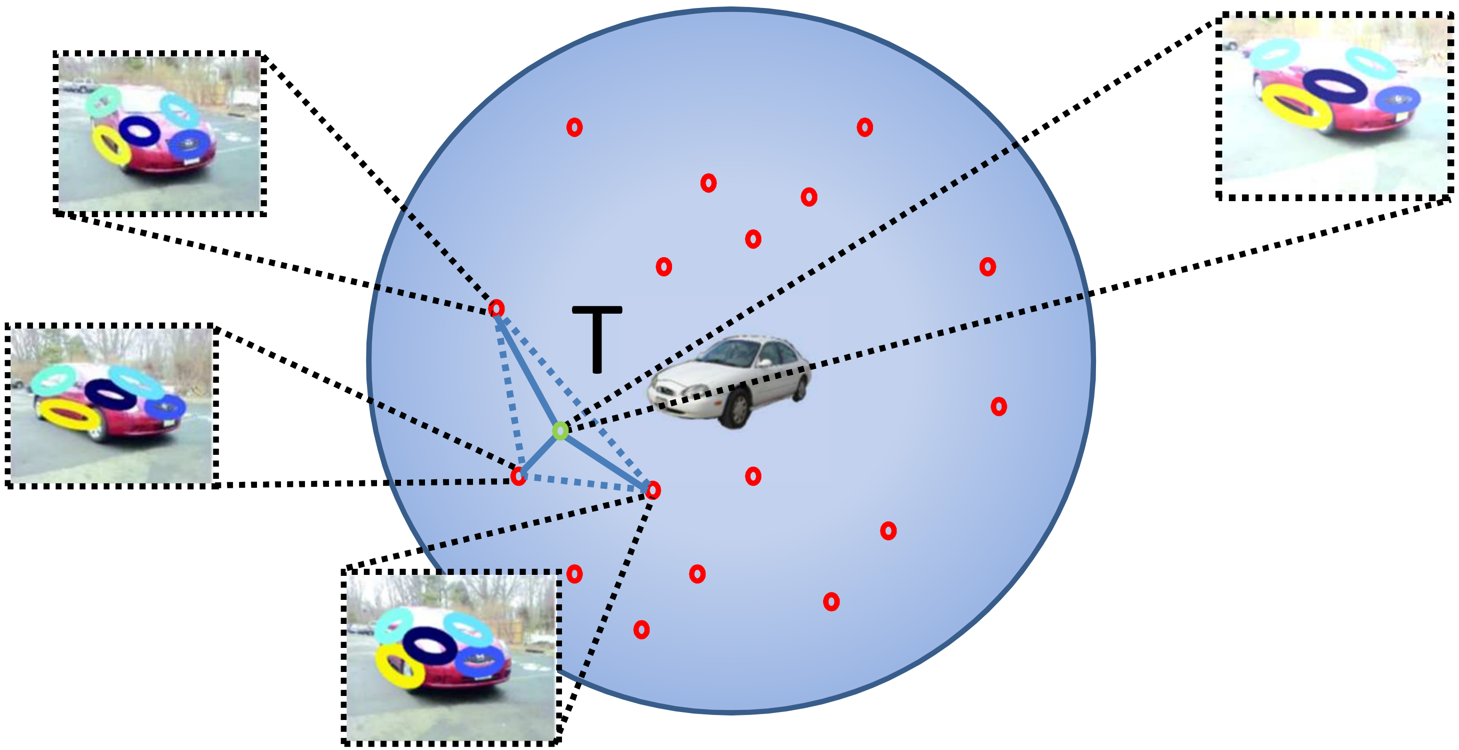
Sun, Su, Savarese, Fei-Fei, CVPR 09

Su, Sun, Fei-Fei, Savarese, ICCV 09

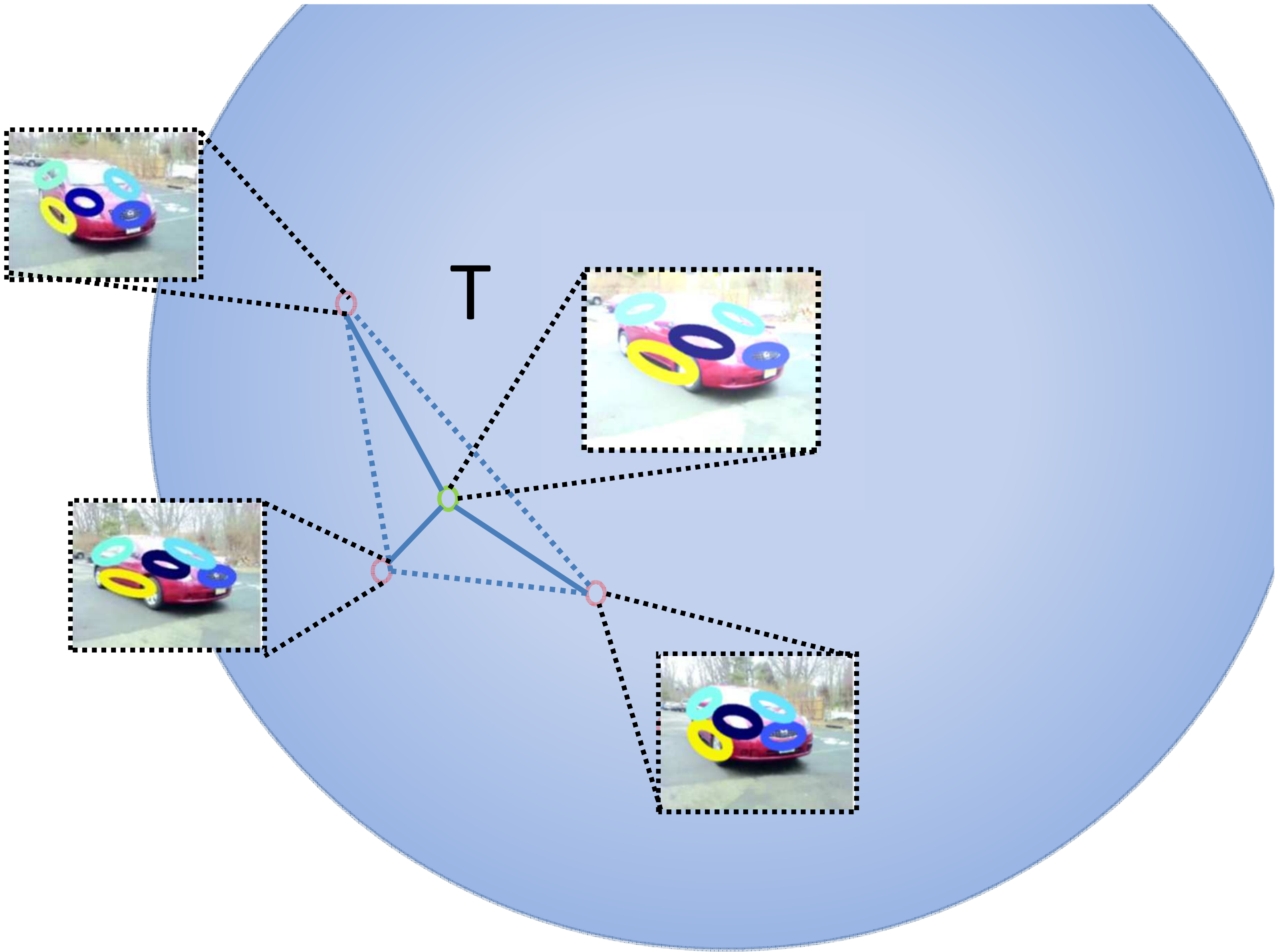


- **Probabilistic generative part-based model**
- **Dense Multi-view representation on the viewing sphere**

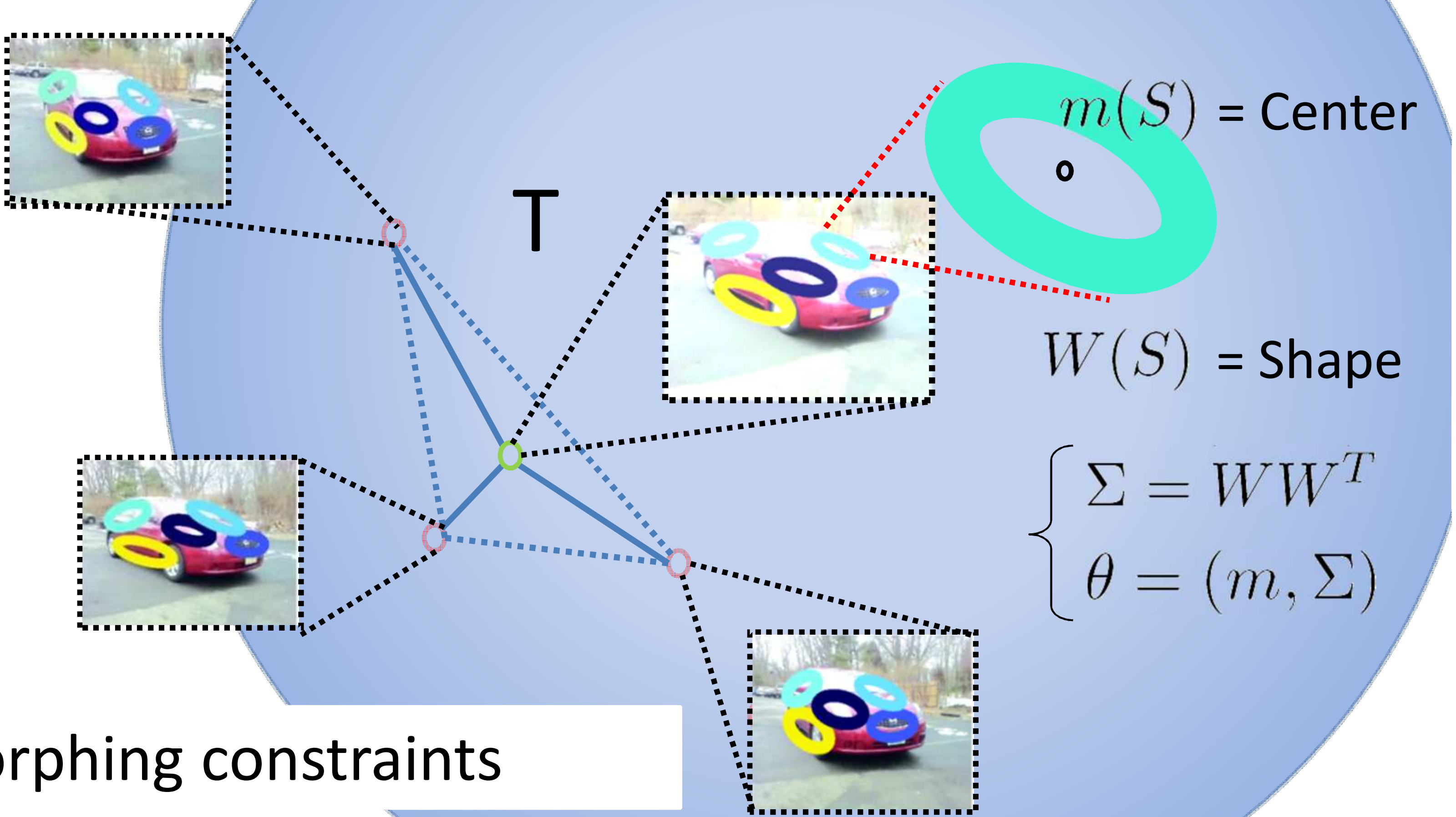
Dense representation on view-sphere



- Triangle T
- Morphing parameter S


T

View Morphing Constraints



View morphing constraints

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

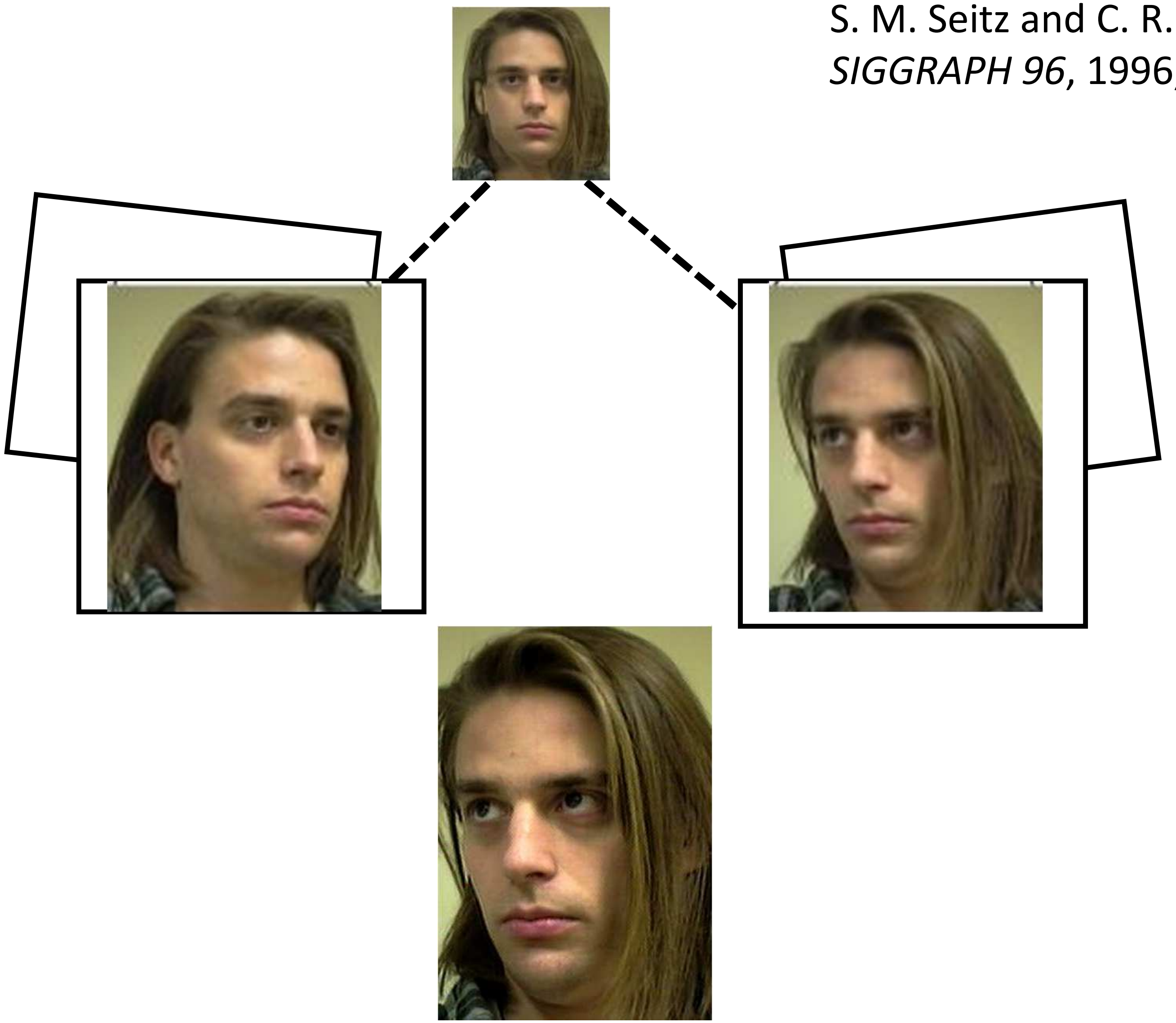
$$W(S) = \sum_{g=1}^{3} W_T^g \cdot s_g$$

Seitz &Dyer SIGGRAPH 96
Xiao & Shah CVIU '04

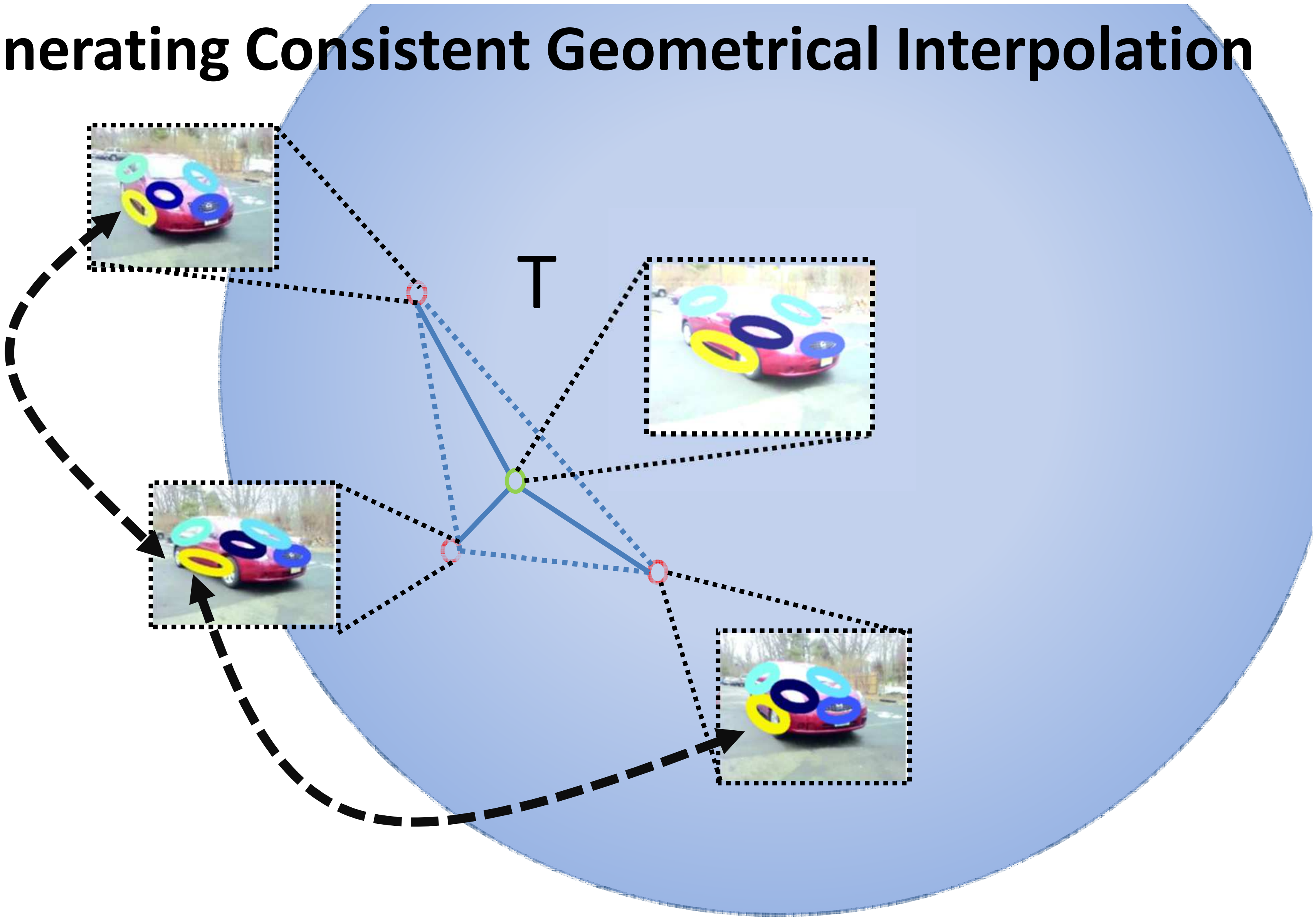
For first time used for modeling object categories!

View Morphing

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

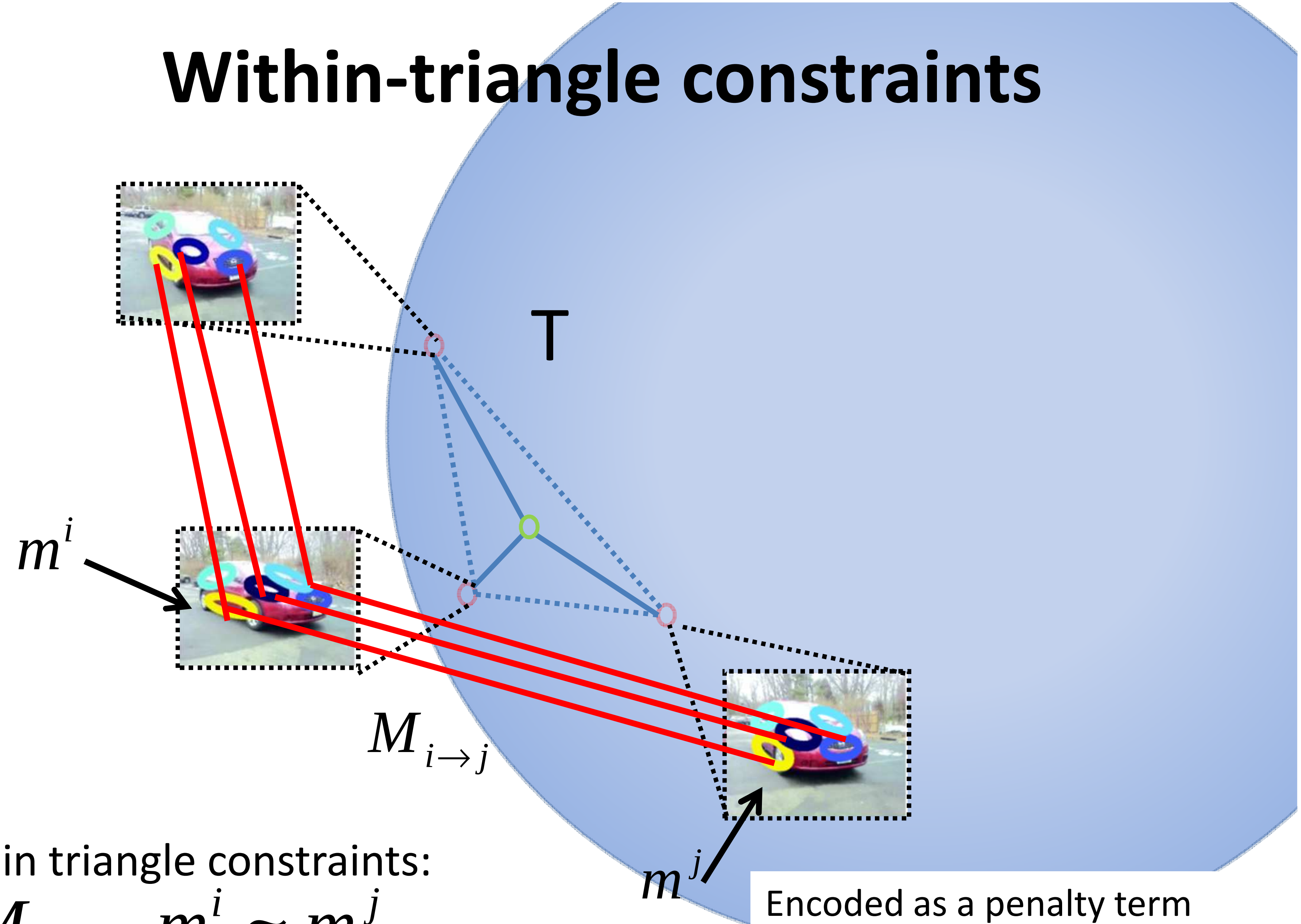


Generating Consistent Geometrical Interpolation



1. Correct correspondence between parts must be established

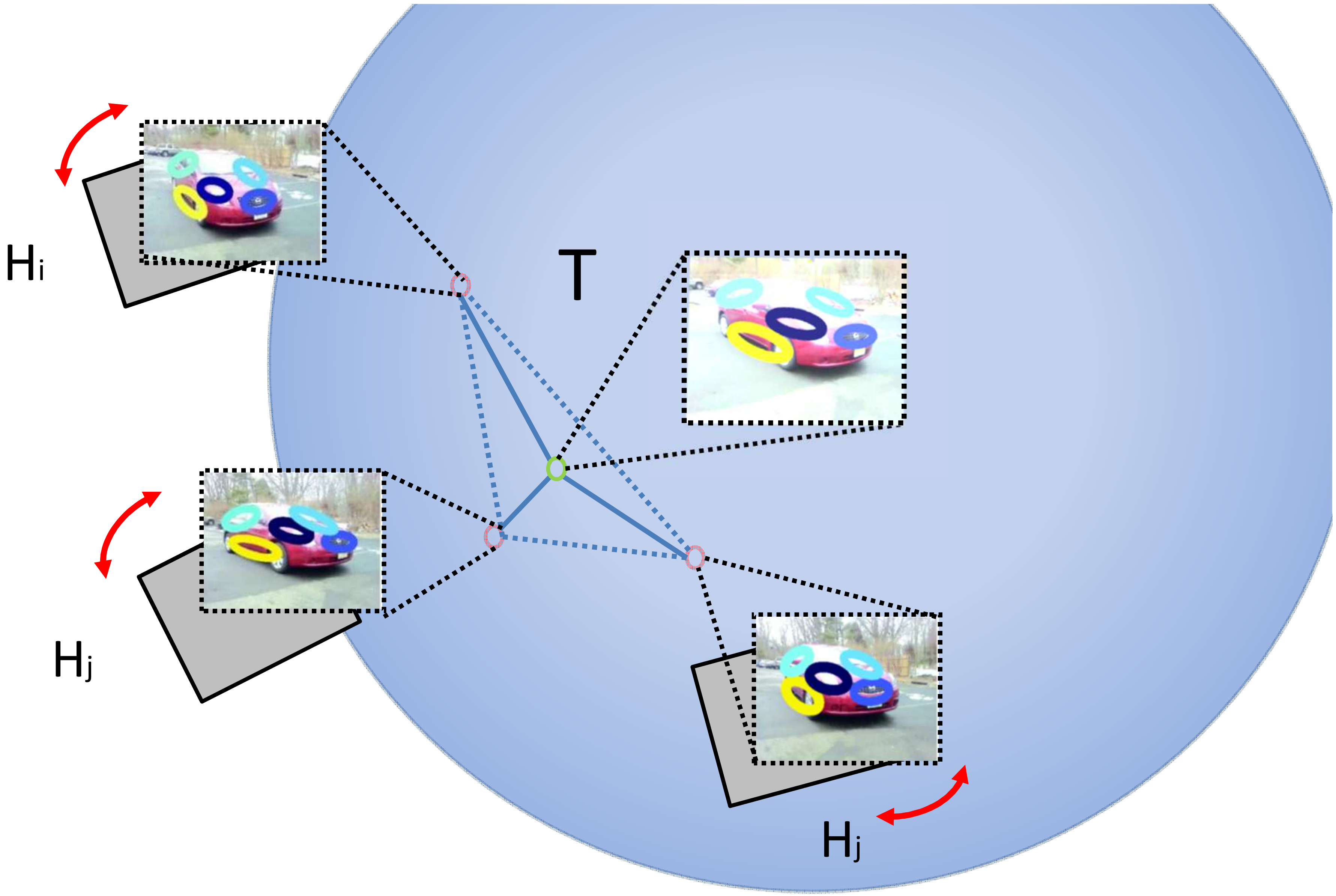
Within-triangle constraints



Within triangle constraints:

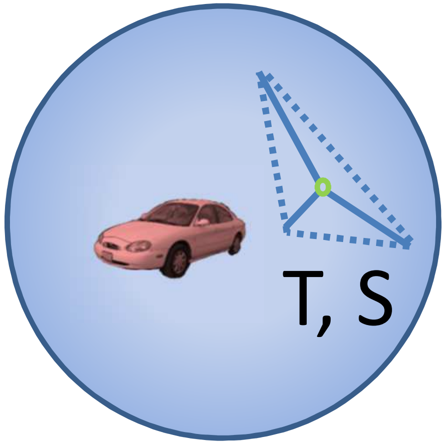
$$M_{i \to j} \cdot m^i \approx m^j$$

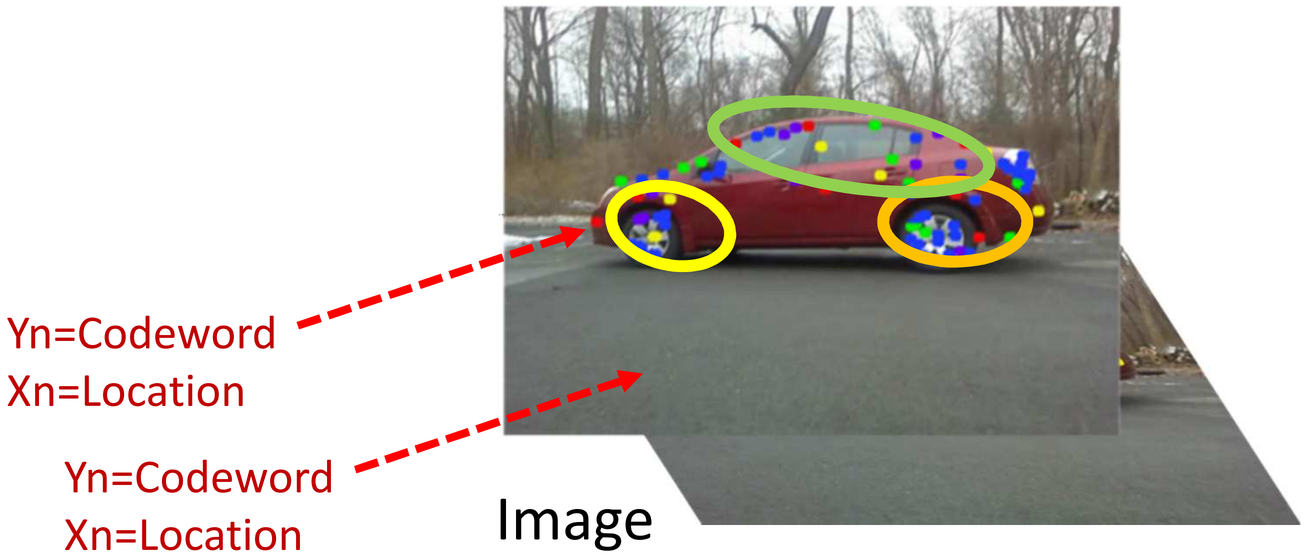
Encoded as a penalty term in variational *EM*



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

Multi-view generative part-based model





Multi-view generative part-based model

α = Part Prop. Prior

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

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

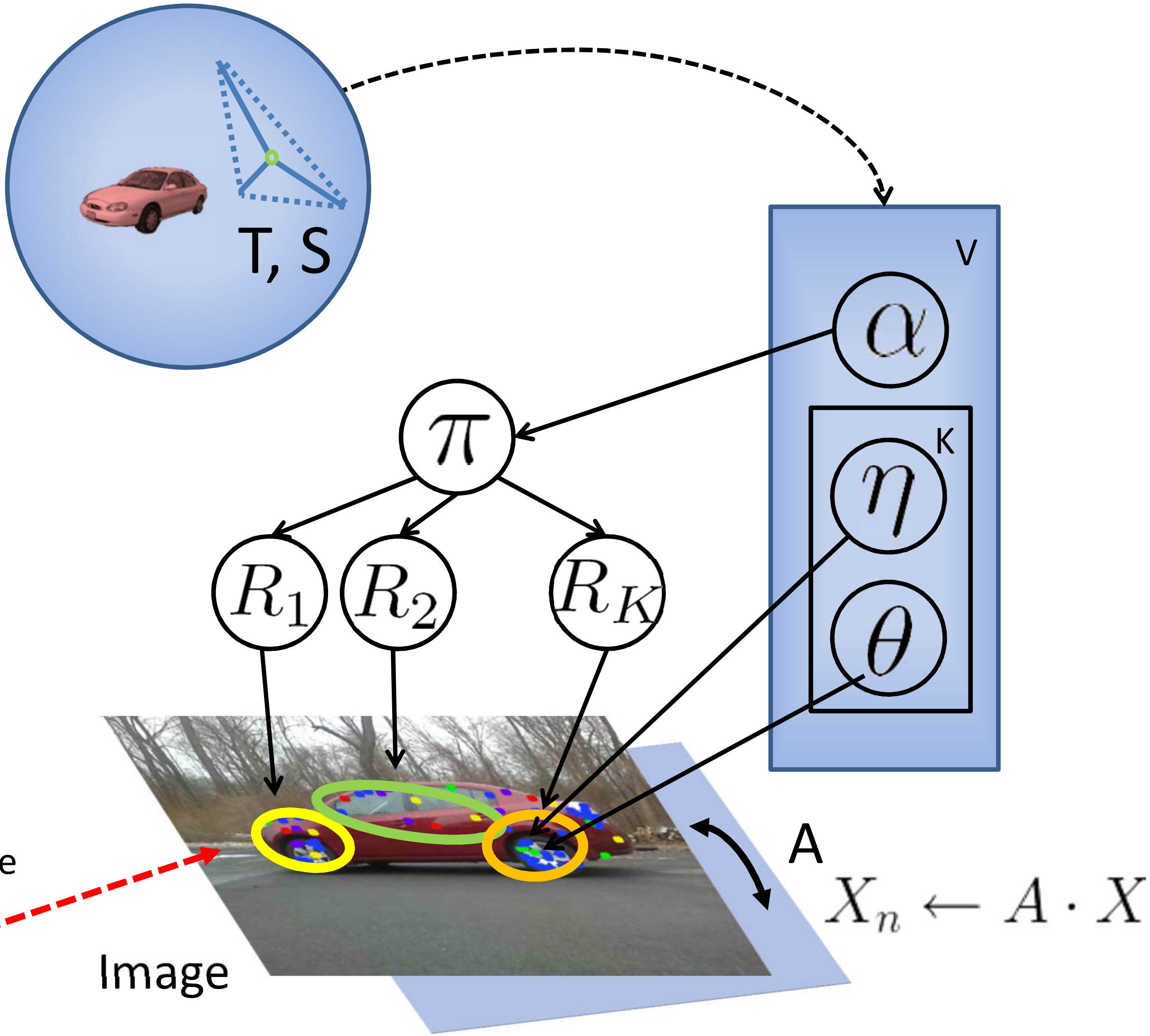
$$Y_n \sim Mult(\eta)$$

$$X_n \sim N(theta)$$

η = Part Appearance

θ = Part Location/shape

Yn=Codeword
Xn=Location



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

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

α = Part Prop. Prior

$\pi \sim Dir(\alpha)$

$R \sim Mult(\pi)$

$Y_n \sim Mult(\eta)$

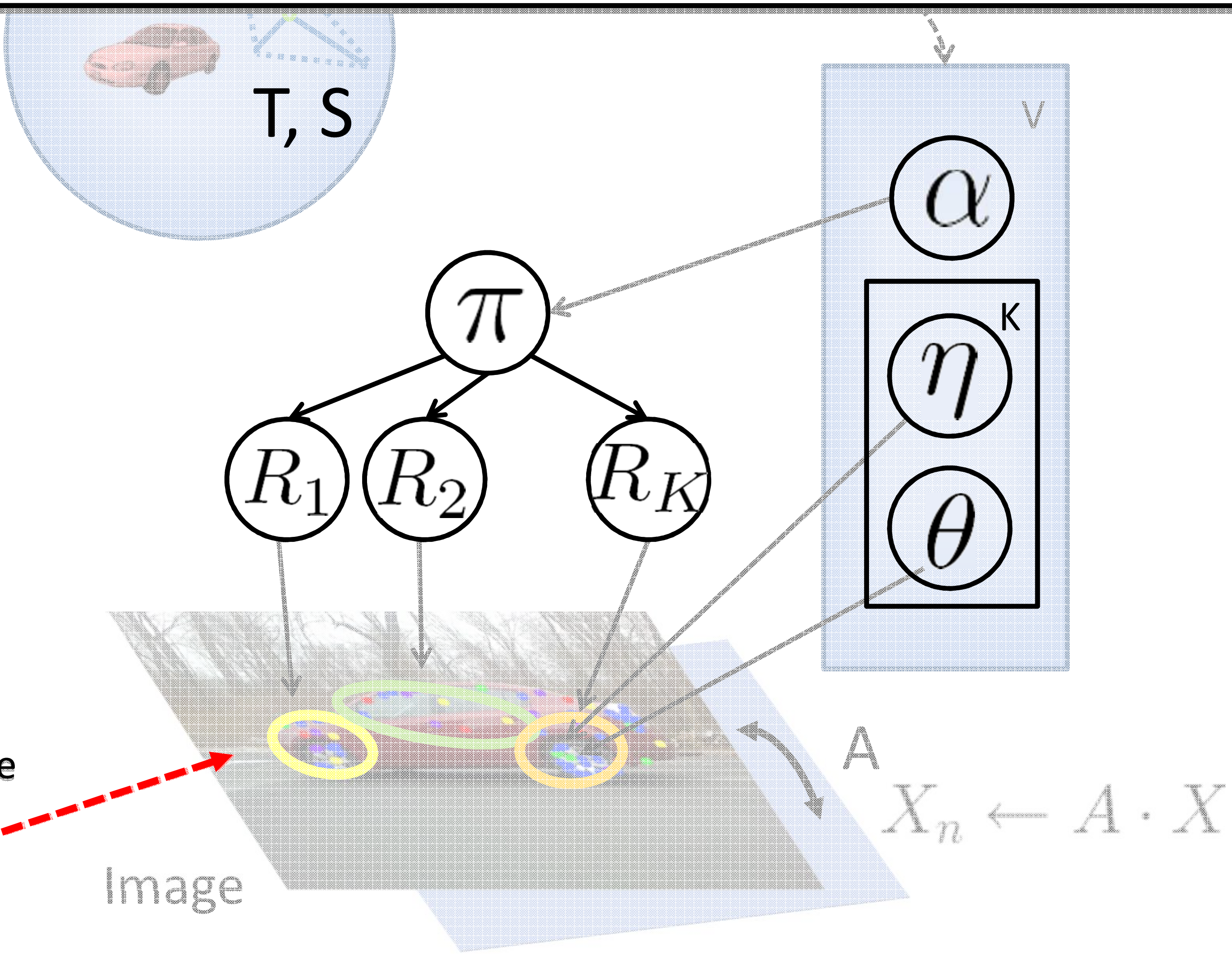
$X_n \sim N(theta)$

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:

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

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

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

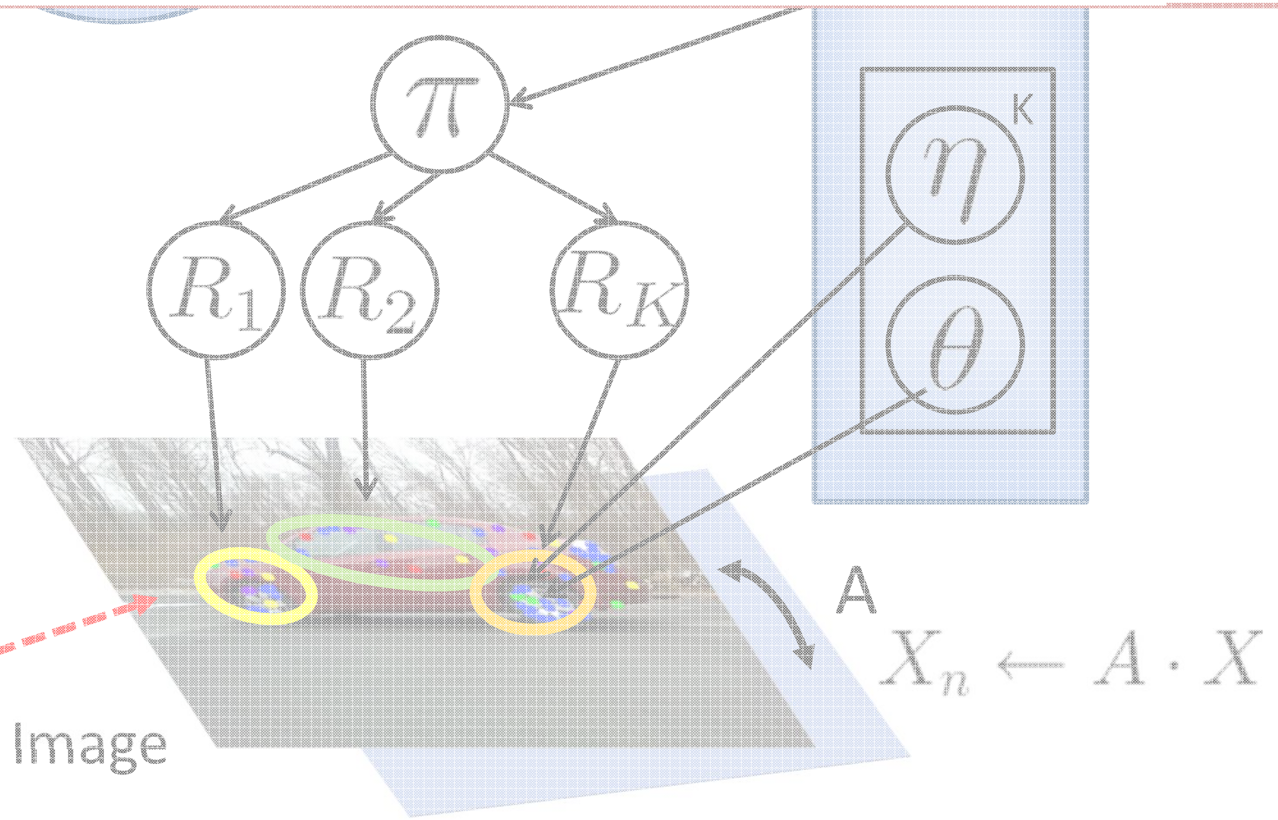
$$X_n \sim N(theta)$$

η = Part Appearance

θ = Part Location/shape

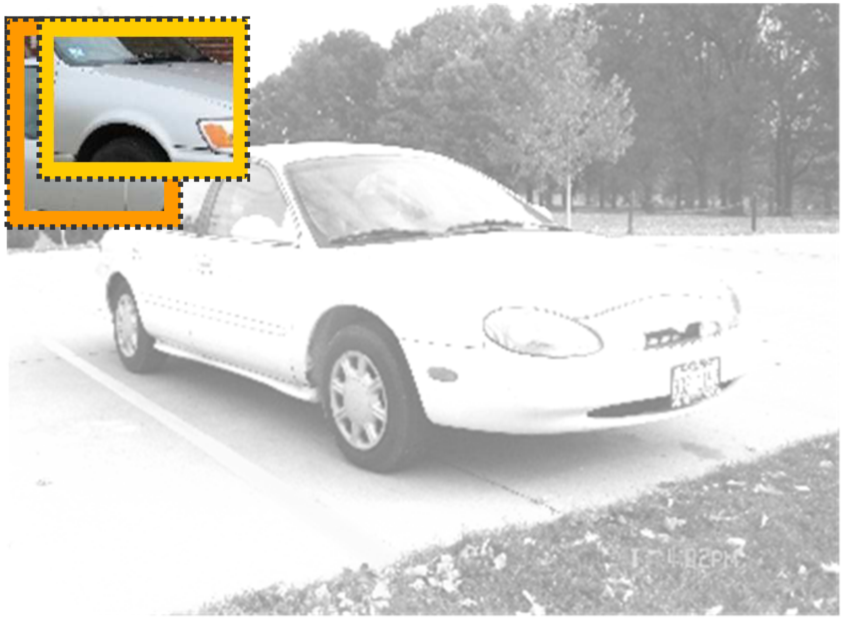
Yn=Codeword

Xn=Location



Object Recognition

Query image



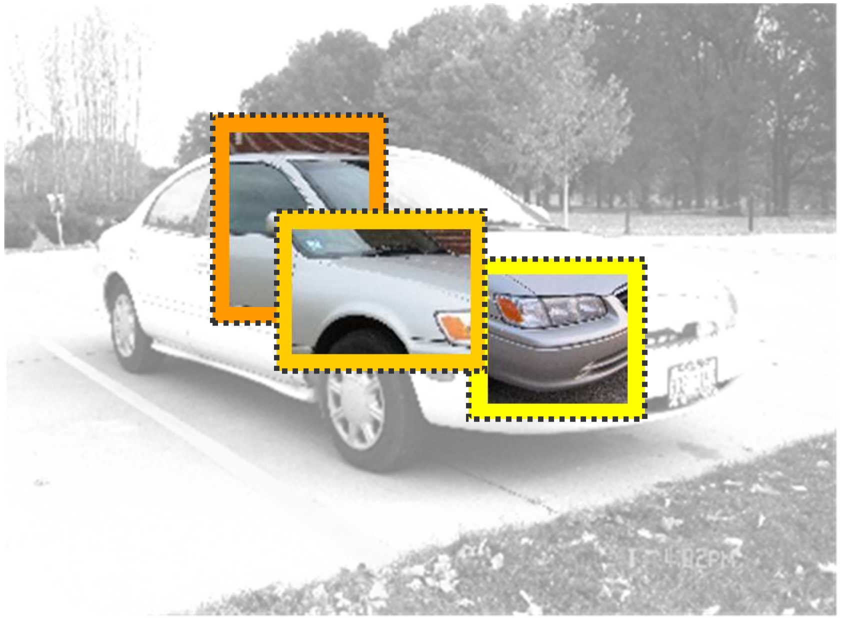
model

Algorithm

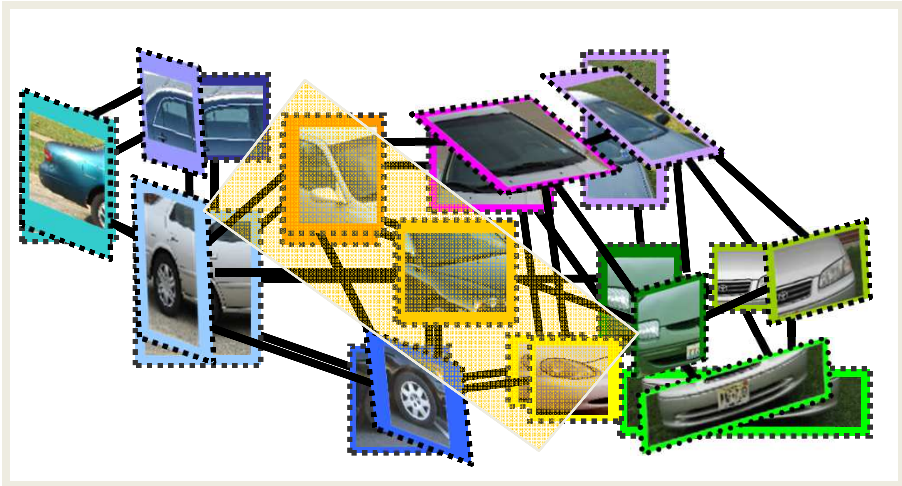
1. Find hypotheses of canonical parts consistent with a given pose

Object Recognition

Query image



model



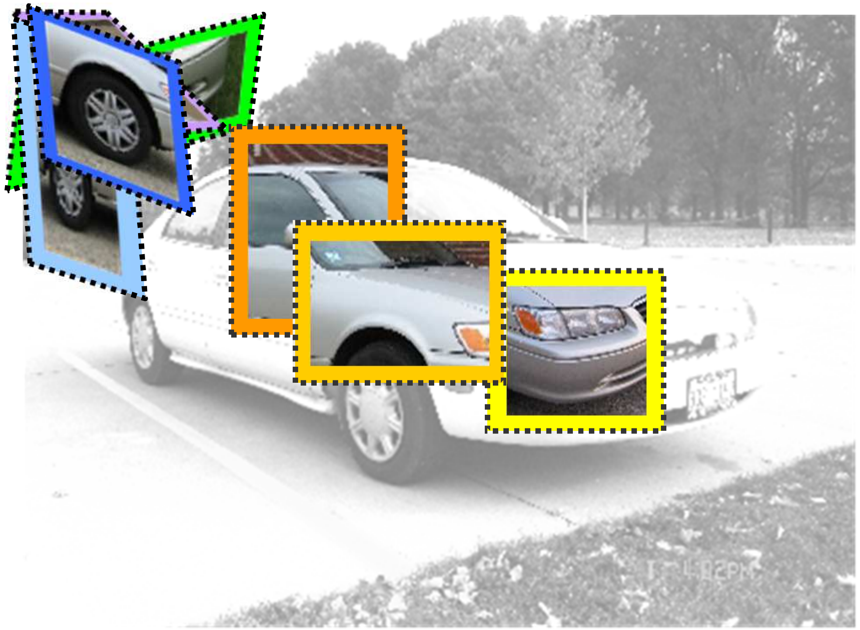
Algorithm

1. Find hypotheses of canonical parts consistent with a given pose

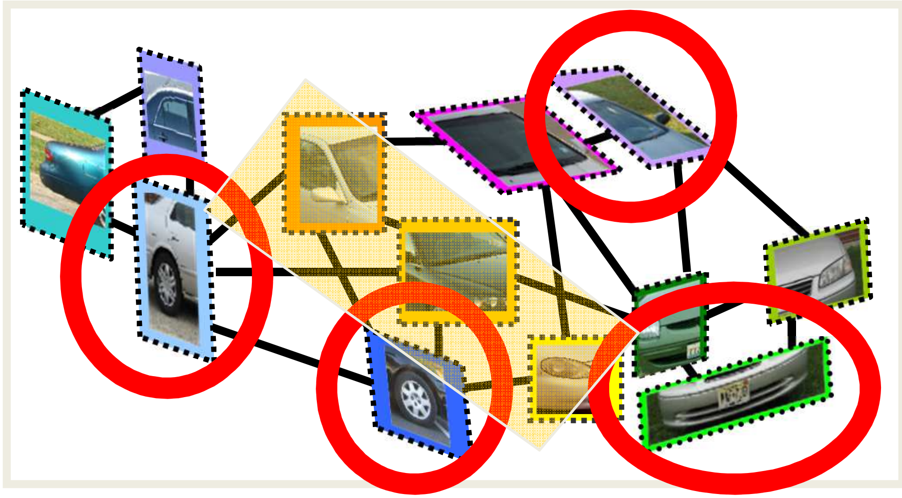
2. Infer position and pose of other canonical parts

Object Recognition

Query image



model



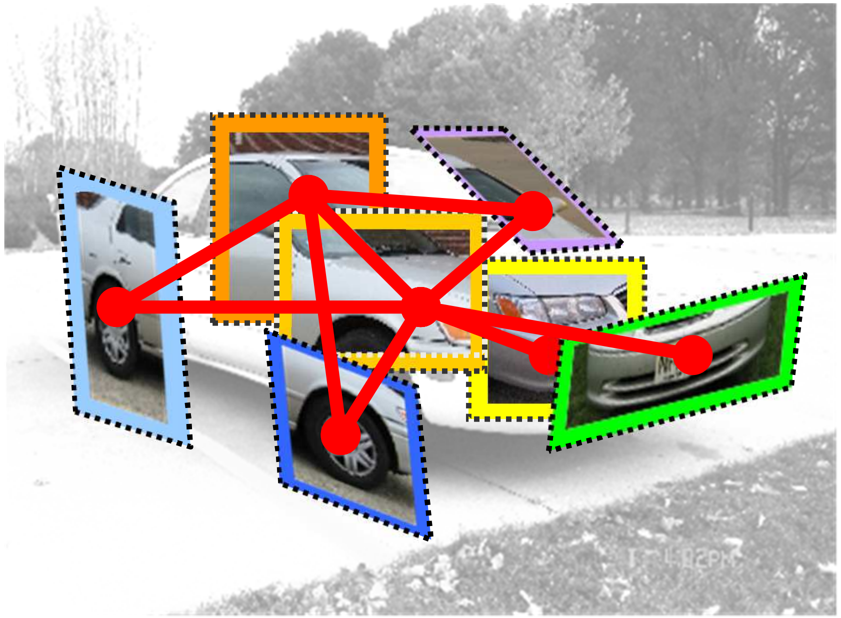
Algorithm

1. Find hypotheses of canonical parts consistent with a given pose

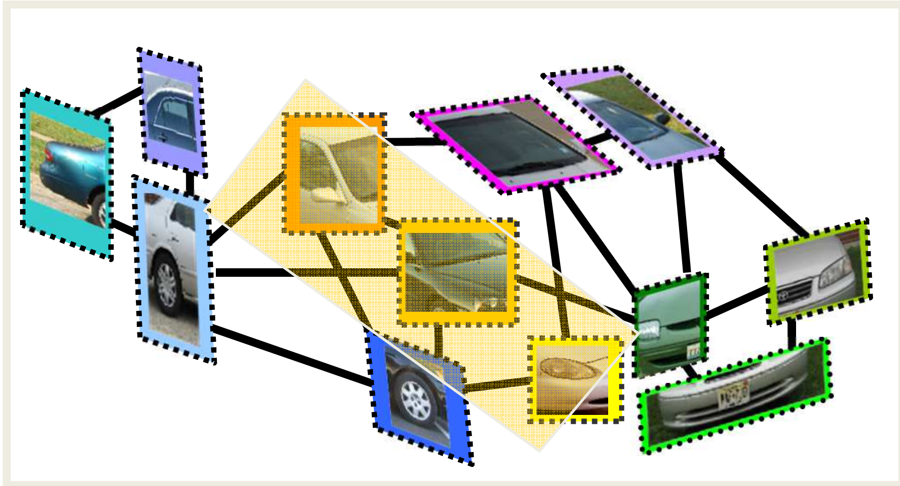
2. Infer position and pose of other canonical parts

Object Recognition

Query image



model



Algorithm

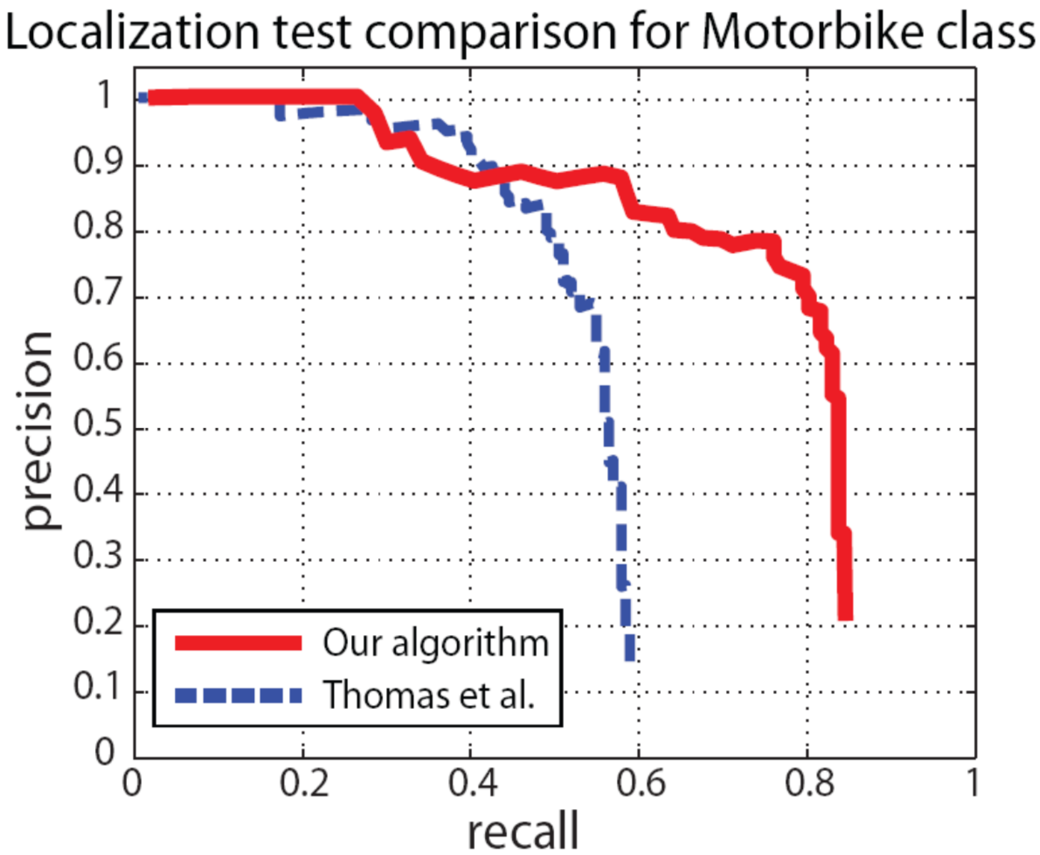
1. Find hypotheses of canonical parts consistent with a given pose

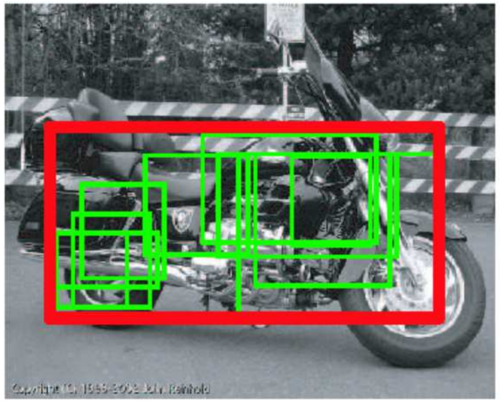
2. Infer position and pose of other canonical parts

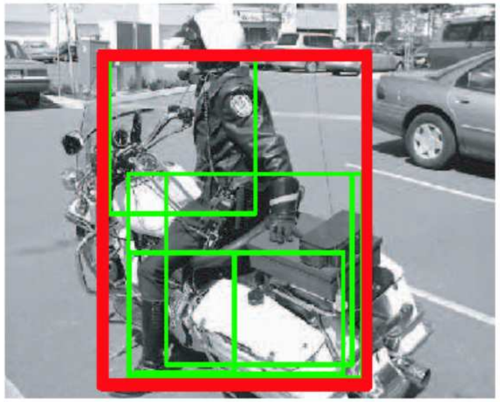
3. Optimize over $\mathbf{E}$, $\mathbf{G}$ and $\mathbf{s}$ to find best combination of hypothesis

$\rightarrow$ **error**

A unified framework for 3D object **detection**, pose classification, pose synthesis





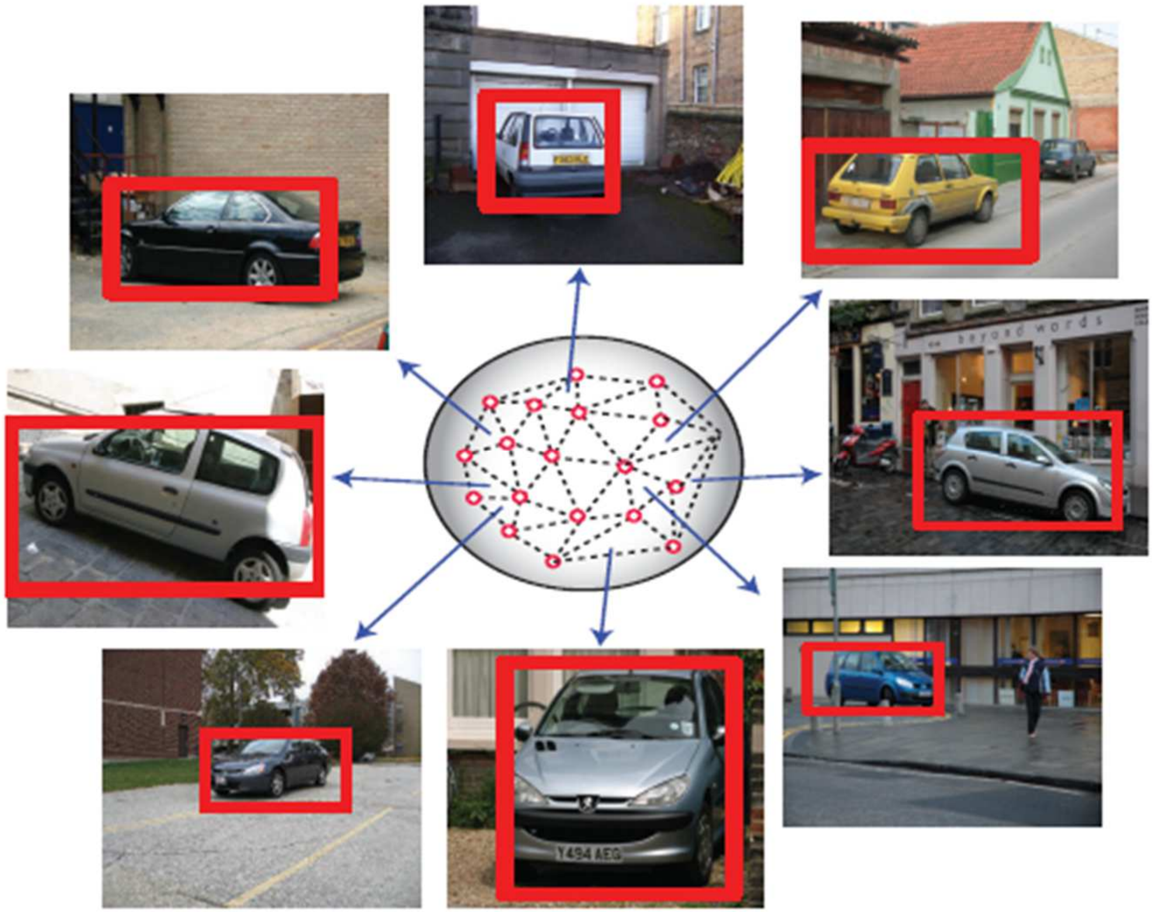


Su, Sun, Fei-Fei, Savarese, ICCV 09

A unified framework for 3D object detection, **pose classification**, pose synthesis

Pose estimation (by angle):
av. of 8 categ (56%)

	f	fl	l	lb	b	br	r	rf
front	.70		.10		.20			
front–left	.03	.37	.05	.08		.32	.08	.08
left			.53	.15	.03	.05	.25	
left–back		.04	.13	.58		.02	.08	.15
back	.35		.04	.04	.50	.04	.04	
back–right		.14			.03	.73	.08	.03
right			.26	.03		.19	.48	.03
right–front		.03	.09	.09		.03	.12	.64

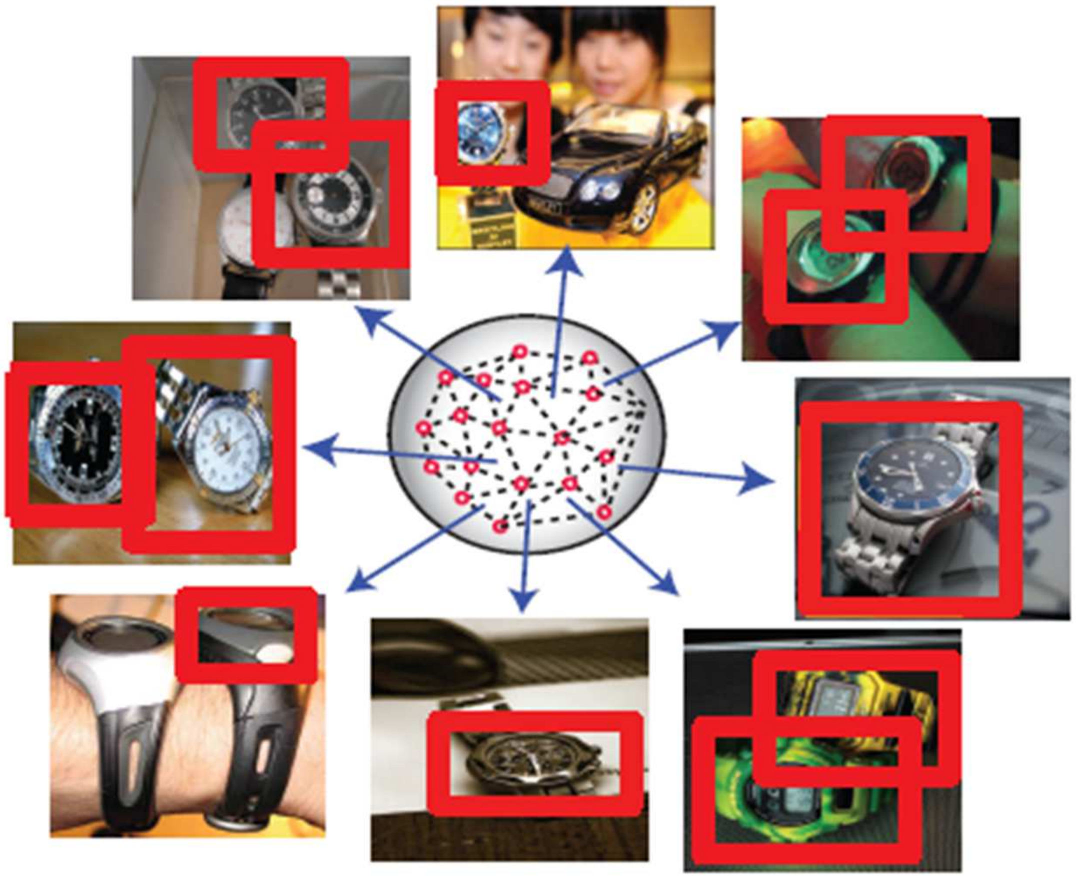


Su, Sun, Fei-Fei, Savarese, ICCV 09

A unified framework for 3D object detection, **pose classification**, pose synthesis

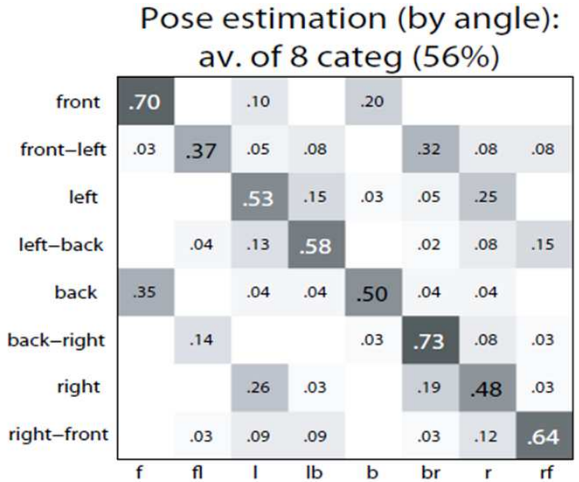
Pose estimation (by angle):
av. of 8 categ (56%)

	f	fl	l	lb	b	br	r	rf
front	.70		.10		.20			
front–left	.03	.37	.05	.08		.32	.08	.08
left			.53	.15	.03	.05	.25	
left–back		.04	.13	.58		.02	.08	.15
back	.35		.04	.04	.50	.04	.04	
back–right		.14			.03	.73	.08	.03
right			.26	.03		.19	.48	.03
right–front		.03	.09	.09		.03	.12	.64



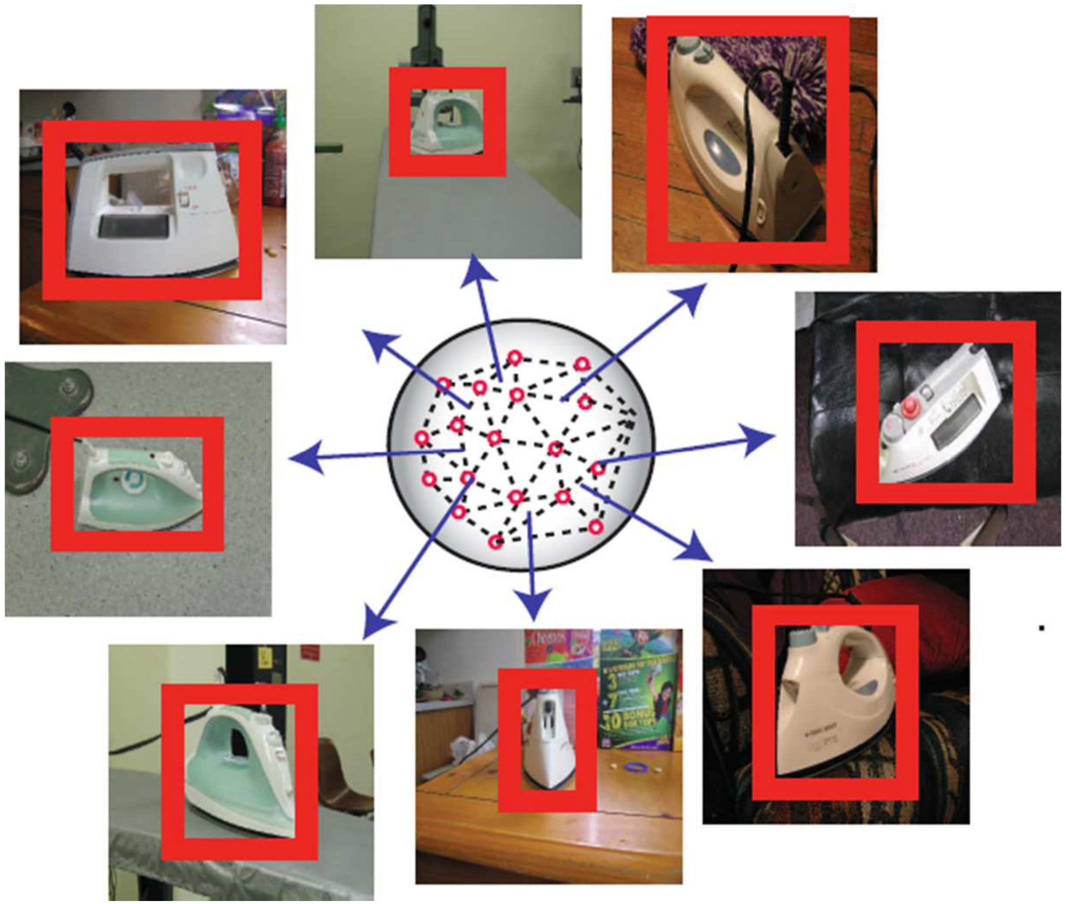
Su, Sun, Fei-Fei, Savarese, ICCV 09

A unified framework for 3D object detection, **pose classification**, pose synthesis



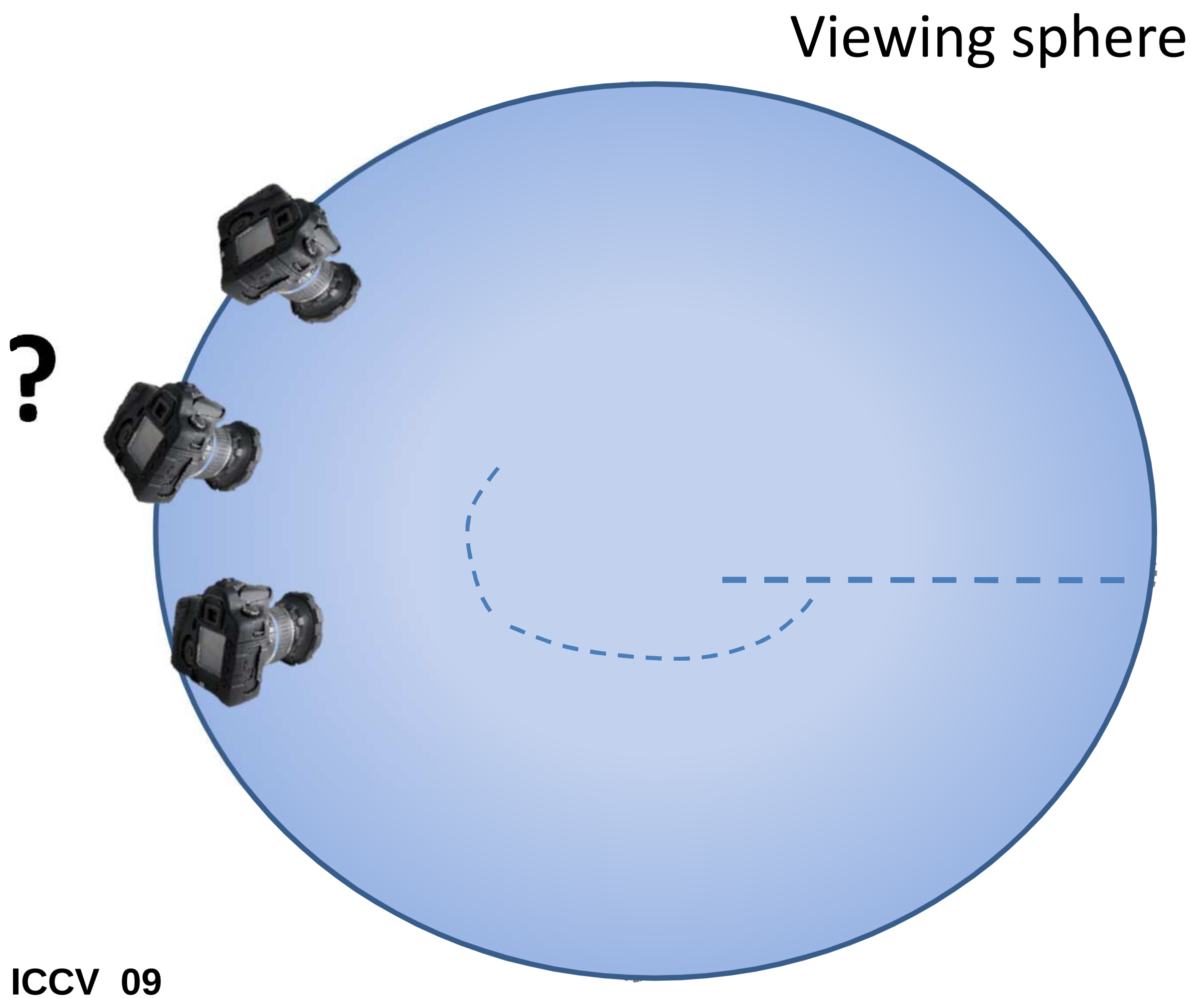
Pose estimation (by angle):
av. of 8 categ (56%)

	f	fl	l	lb	b	br	r	rf
front	.70		.10		.20			
front–left	.03	.37	.05	.08		.32	.08	.08
left			.53	.15	.03	.05	.25	
left–back		.04	.13	.58		.02	.08	.15
back	.35		.04	.04	.50	.04	.04	
back–right		.14			.03	.73	.08	.03
right			.26	.03		.19	.48	.03
right–front		.03	.09	.09		.03	.12	.64



Su, Sun, Fei-Fei, Savarese, ICCV 09

Predicting object appearance from novel views



Su, Sun, Fei-Fei, Savarese, ICCV 09

Novel view object synthesis from a single image

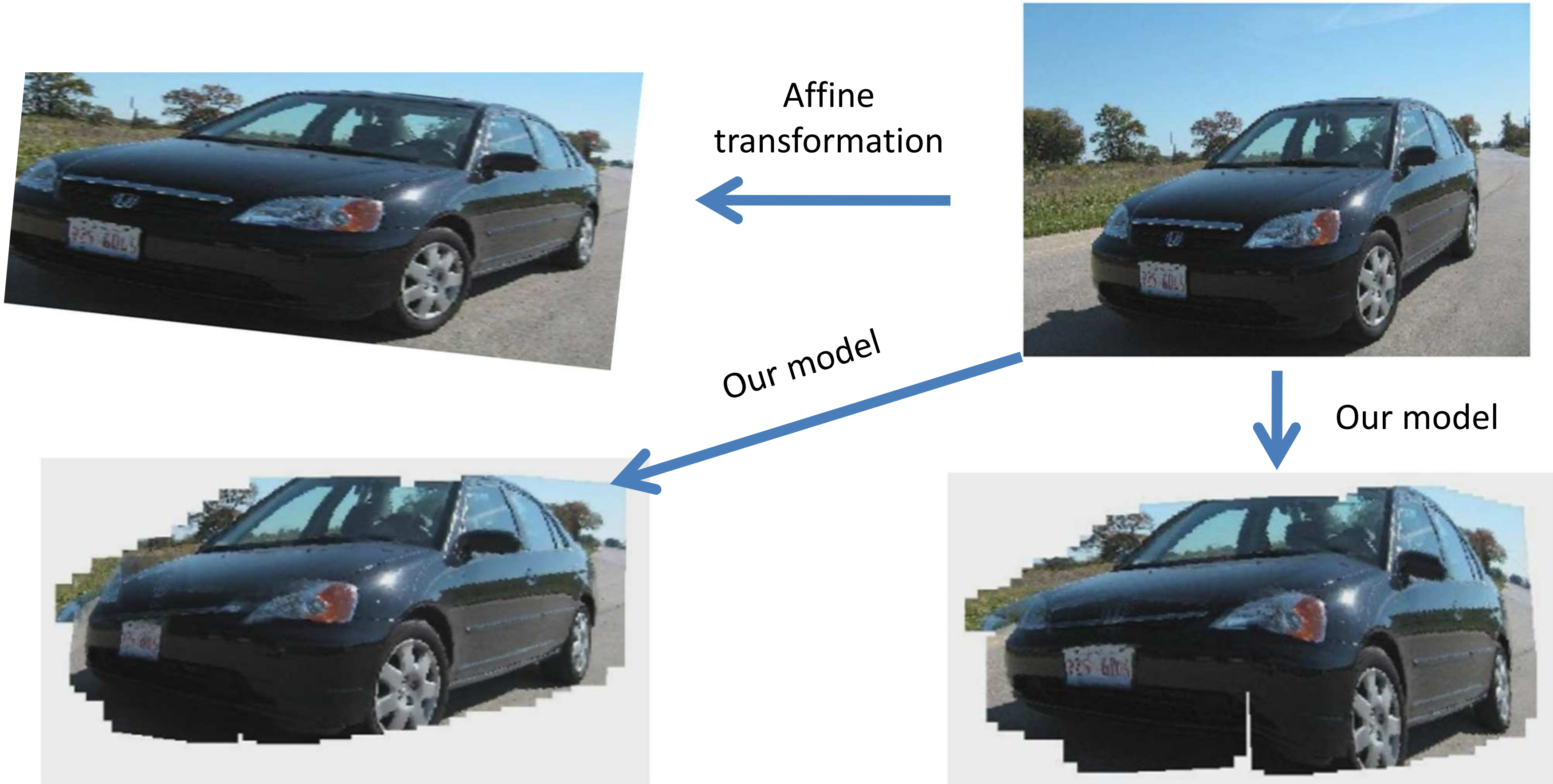
[For natural scenes, see Hoiem et al 07;
Saxena et al 07]

Thomas et al 08
Cremer et al 09

Su, Sun, Fei-Fei, Savarese, ICCV 09

Su, Sun, Fei-Fei, Savarese, ICCV 09

Novel view object synthesis from a single image For the first time!

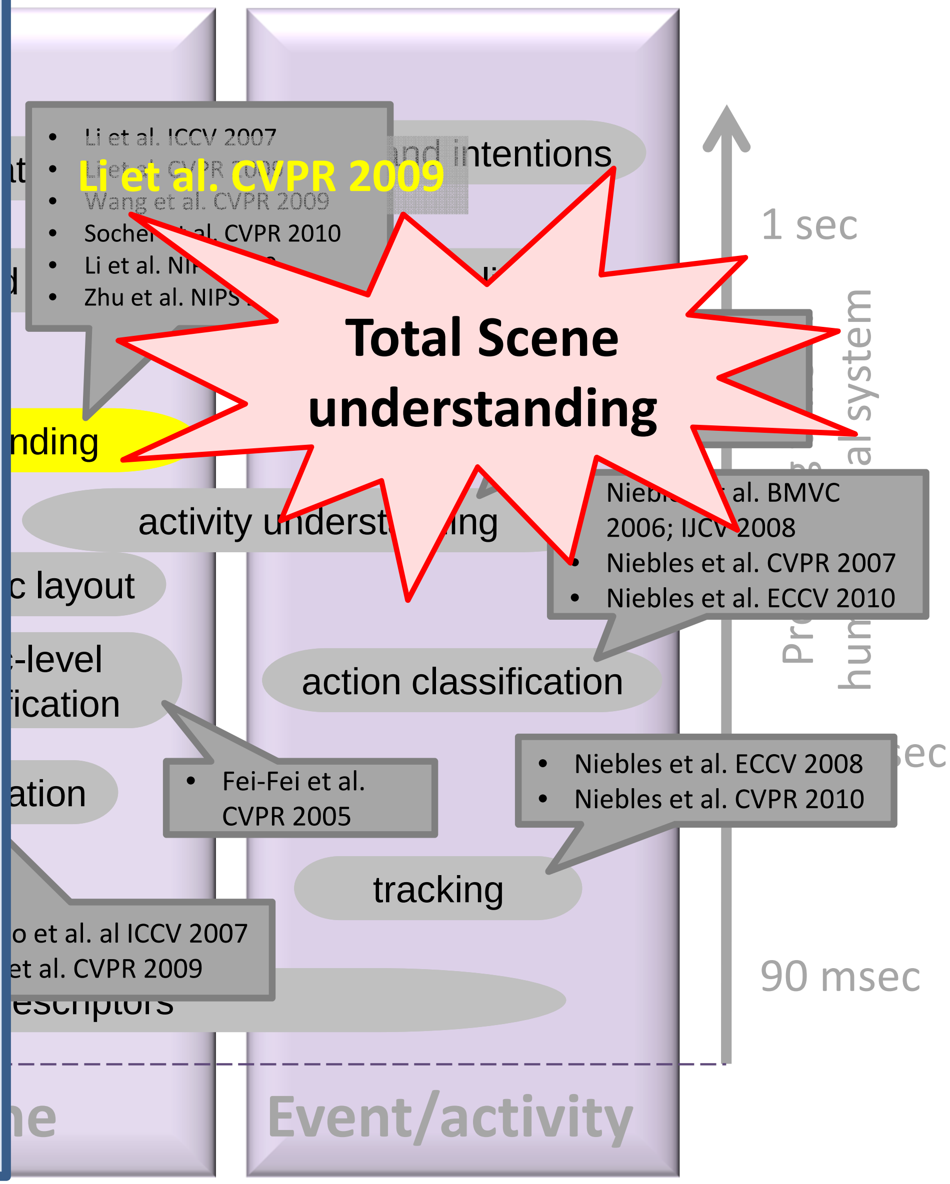


Su, Sun, Fei-Fei, Savarese, ICCV 09

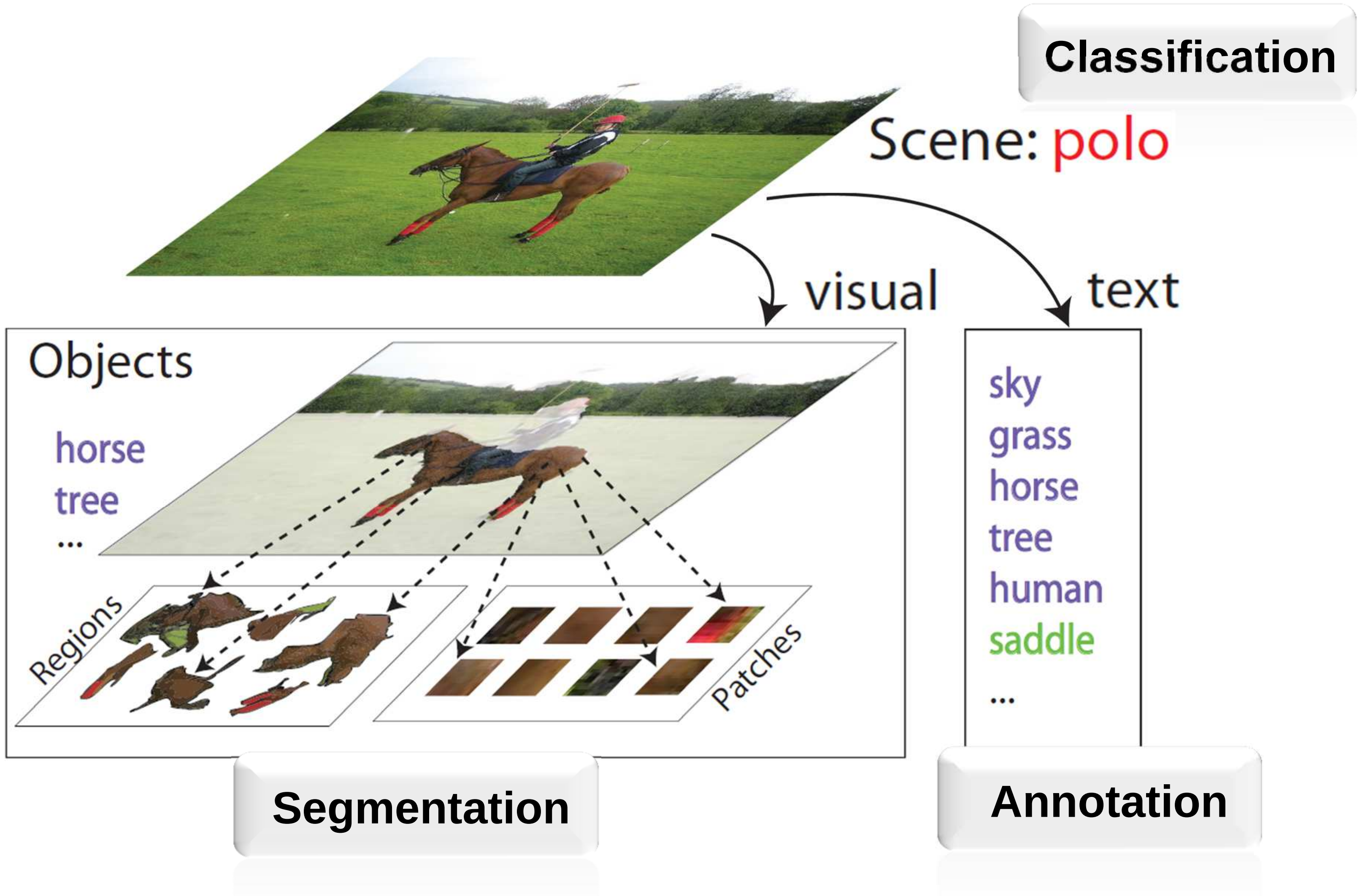
g in images

L.-J. Li, R. Socher and L. Fei-Fei. **Towards Total Scene Understanding: Classification, Annotation and Segmentation in an Automatic Framework.** *Computer Vision and Pattern Recognition (CVPR) 2009.*

Li-Jia Li
Stanford University



Towards total scene understanding



Li, Socher, & Fei-Fei, *CVPR,* 2009

Classification

Annotation

Segmentation

class: Polo

Related Work:

Weber et al. 00
Fergus et al 03
Fei-Fei et al 03
Felzenswalb et al 04

Fei-Fei et al 05
Sivic et al 05
Bosch et al. 06

Oliva et al 01
Lazebnik et al 06

Classification

Annotation

Segmentation

Athlete
Horse
Grass
Trees
Sky
Saddle

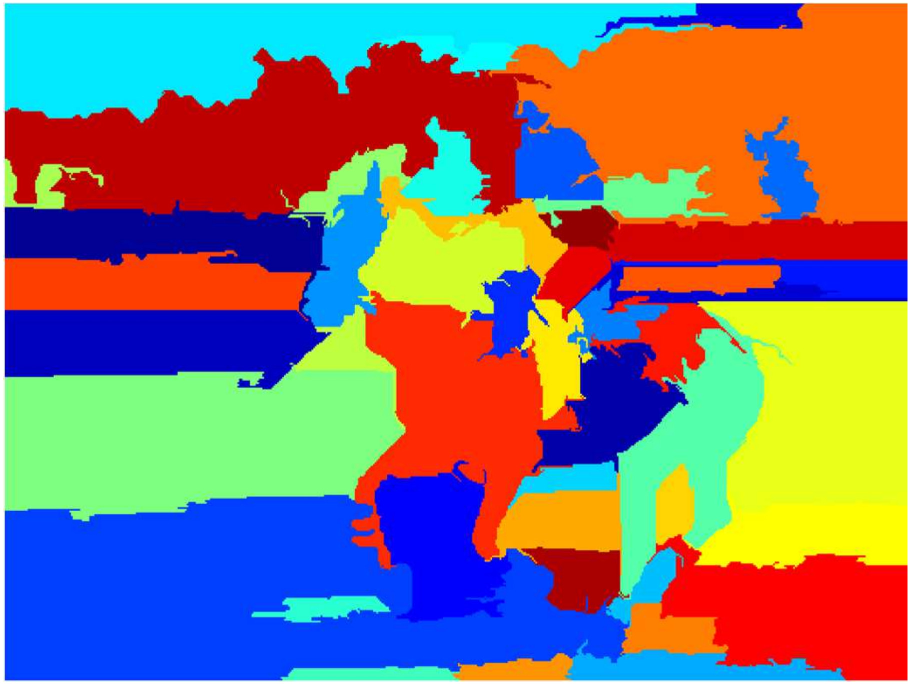
Related Work:

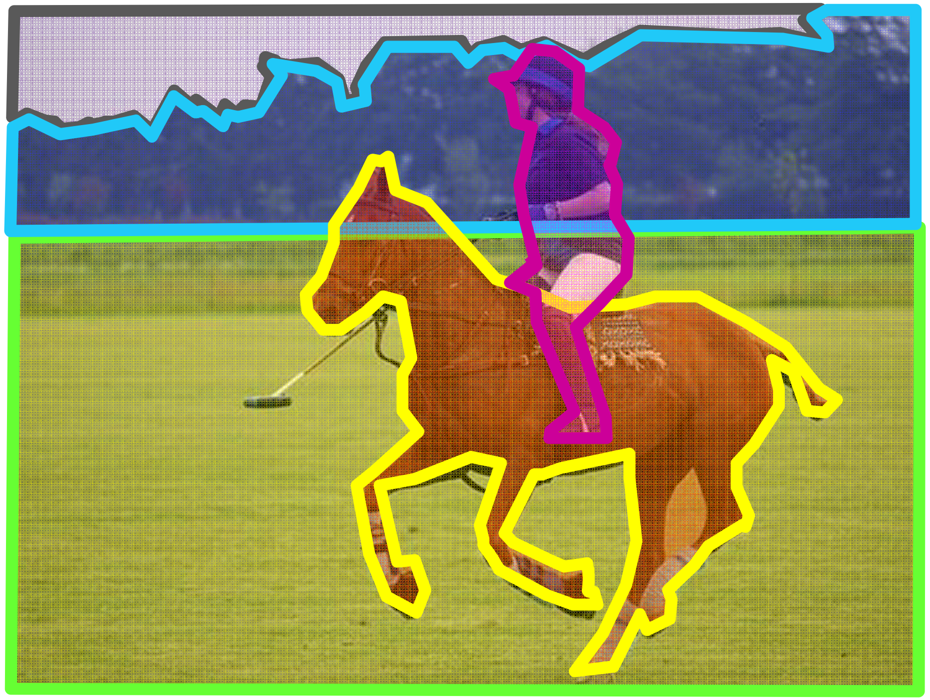
Duygulu et al 02 Blei et al 03

Barnard et al 03 Gupta et al 08

Alipr (Li et al 03)

Classification | **Annotation** | **Segmentation**





Related Work:

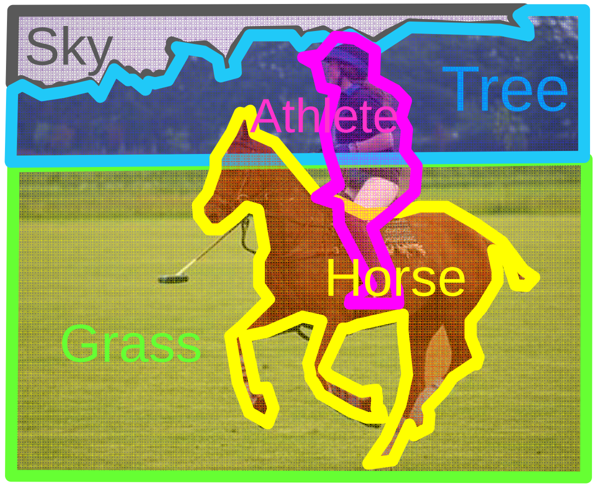
Shi & Malik 00

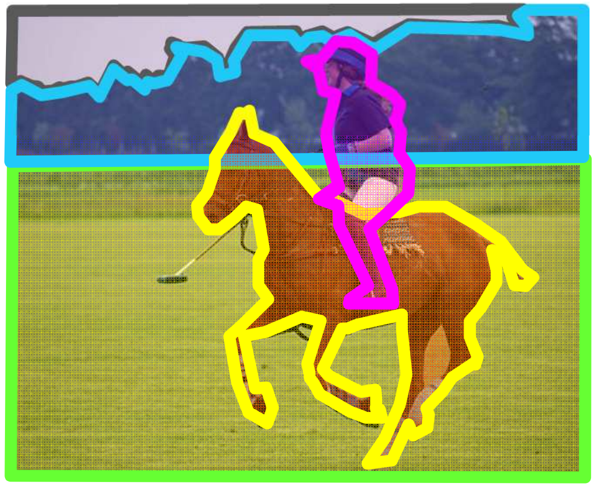
Felzenszwalb & Huttenlocher 04

Sali et al. 99
Winn et al. 05
Kumar et al. 05

Cao & Fei-Fei 07
Russell et al. 06
Wang et al. 07
Todorovic et al. 06

Annotation Segmentation	Classification Segmentation	Classification Annotation





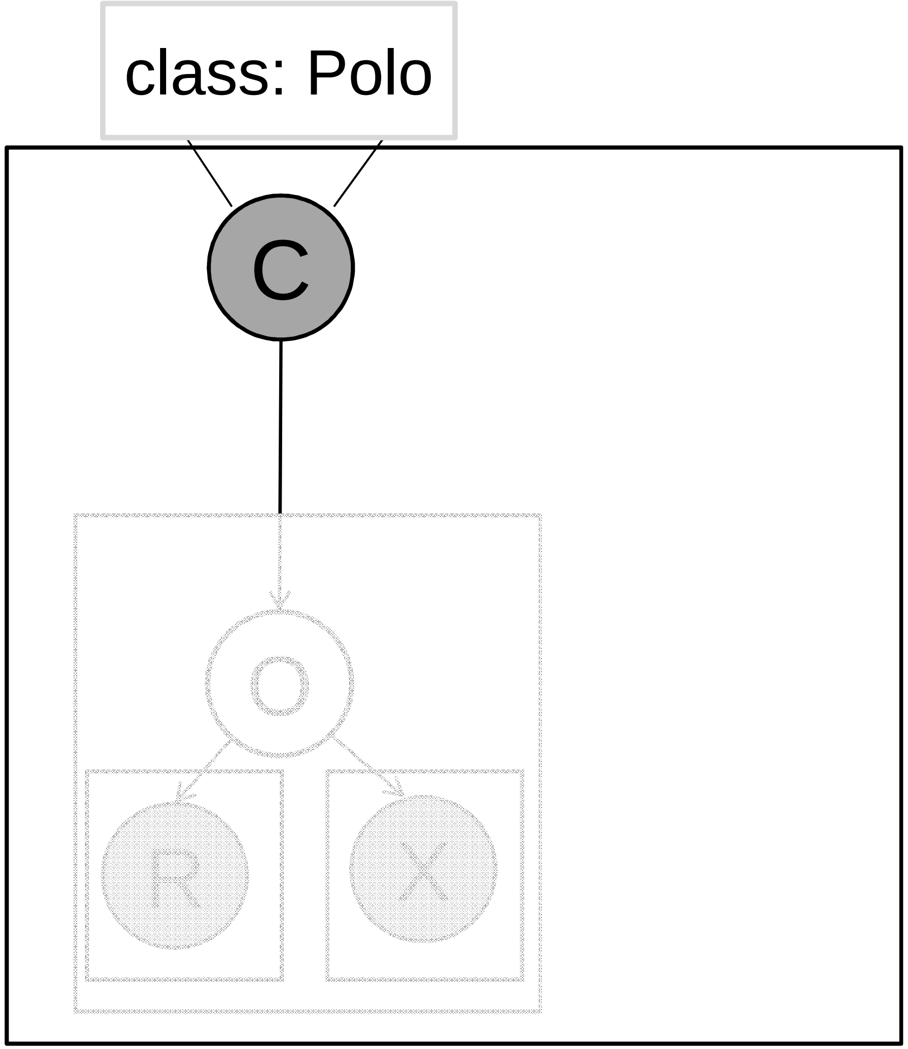
Class: Polo

Class: Polo

Related Work:

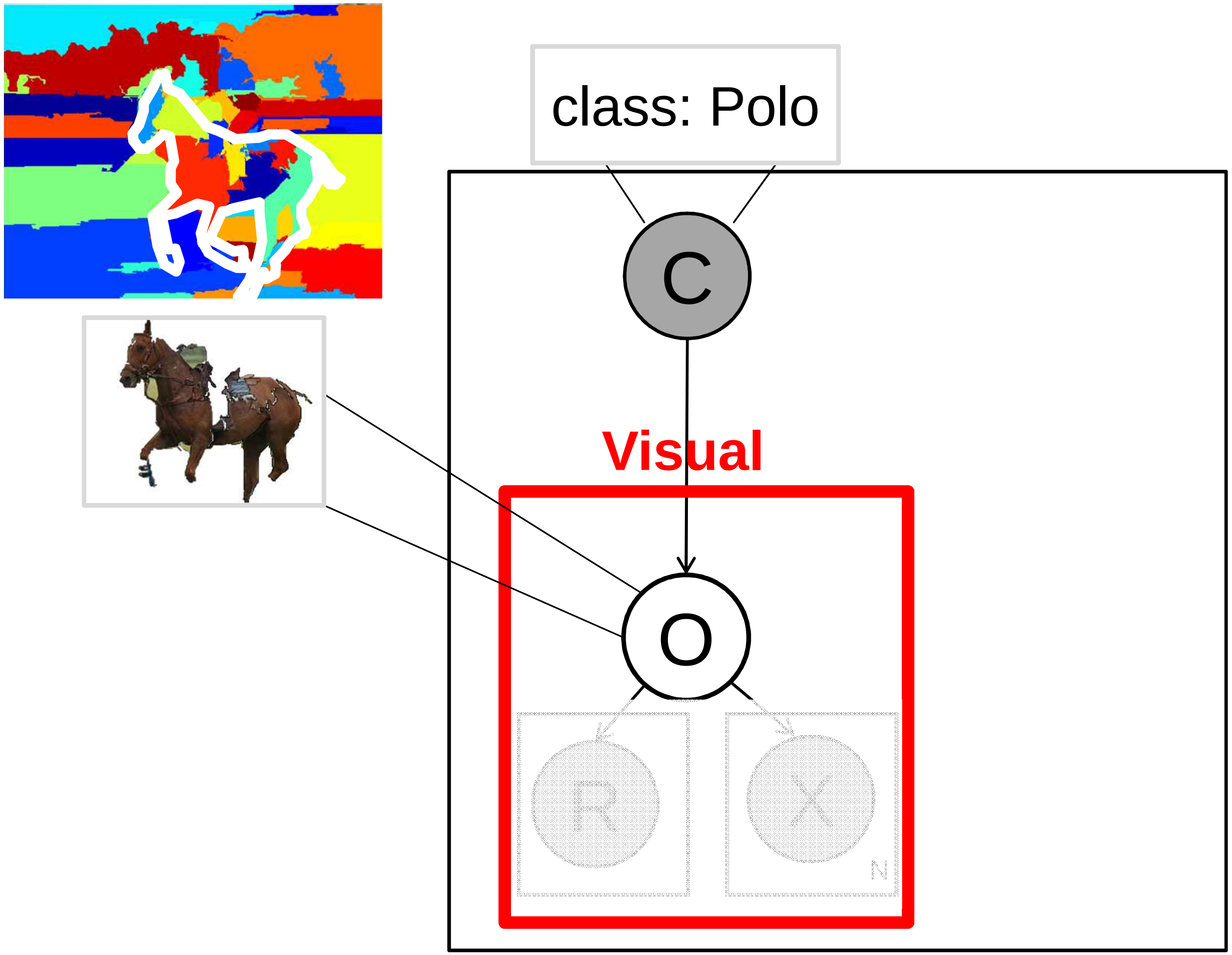
Annotation Segmentation	Classification Segmentation	Classification Annotation
Tu et al 03; Gupta et al. 2008	Heitz et al 08	Li & Fei-Fei 07

A joint model for image classification, annotation & segmentation



$$p(C, \mathbf{O}, \mathbf{R}, \mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{Z} | \eta, \alpha, \beta, \gamma, \theta, \varphi) = p(C) \cdot \text{Visual Component} \cdot \text{Text Component}$$

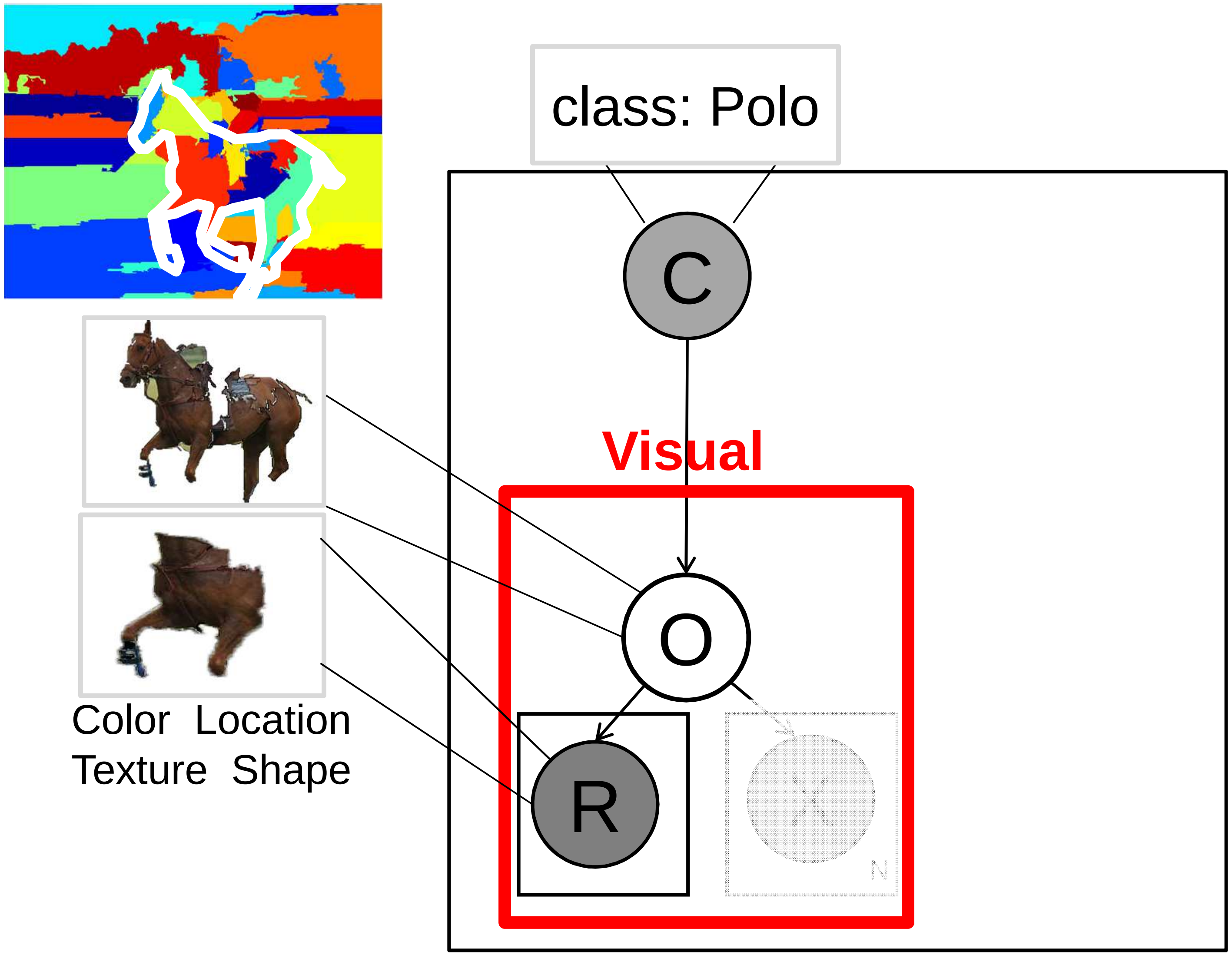
A joint model for image classification, annotation & segmentation



$$p(C,\mathbf{O},\mathbf{R},\mathbf{X},\mathbf{S},\mathbf{T},\mathbf{Z}|\eta,\alpha,\beta,\gamma,\theta,\varphi)= p(C)\cdot\left(\prod_{n=1}^{N_r} p(O_n|\eta,C)\right)$$

$\cdot$ **Text Component**

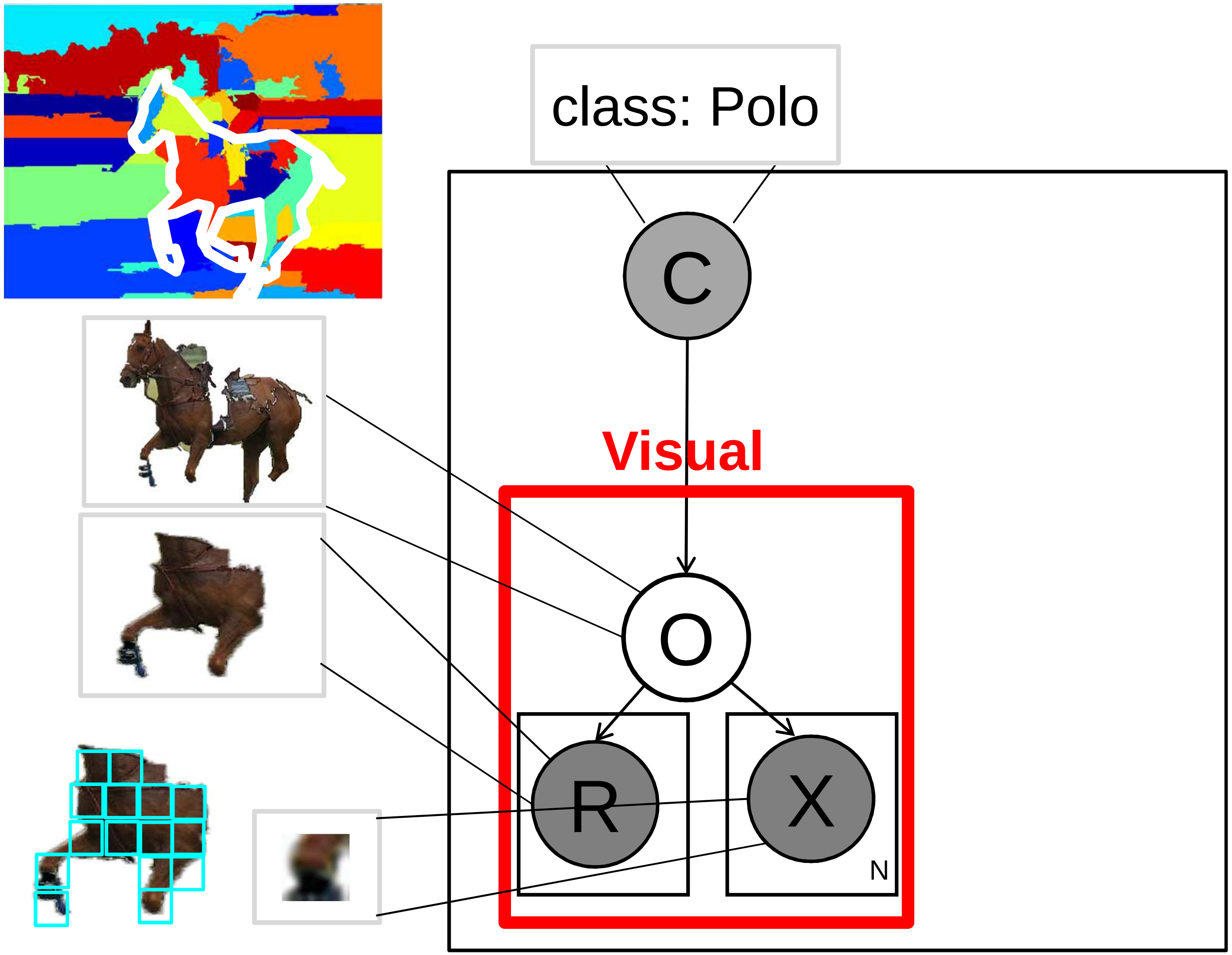
A joint model for image classification, annotation & segmentation



$$p(C,\mathbf{O},\mathbf{R},\mathbf{X},\mathbf{S},\mathbf{T},\mathbf{Z}|\eta,\alpha,\beta,\gamma,\theta,\varphi)= p(C)\cdot\left(\prod_{n=1}^{N_r} p(O_n|\eta,C)\right)\prod_{n=1}^{N_r}\left(\left(\prod_{i=1}^{N_F} p(R_{ni}|O_n,\alpha_i)\right)\right.$$

$\cdot$ **Text Component**

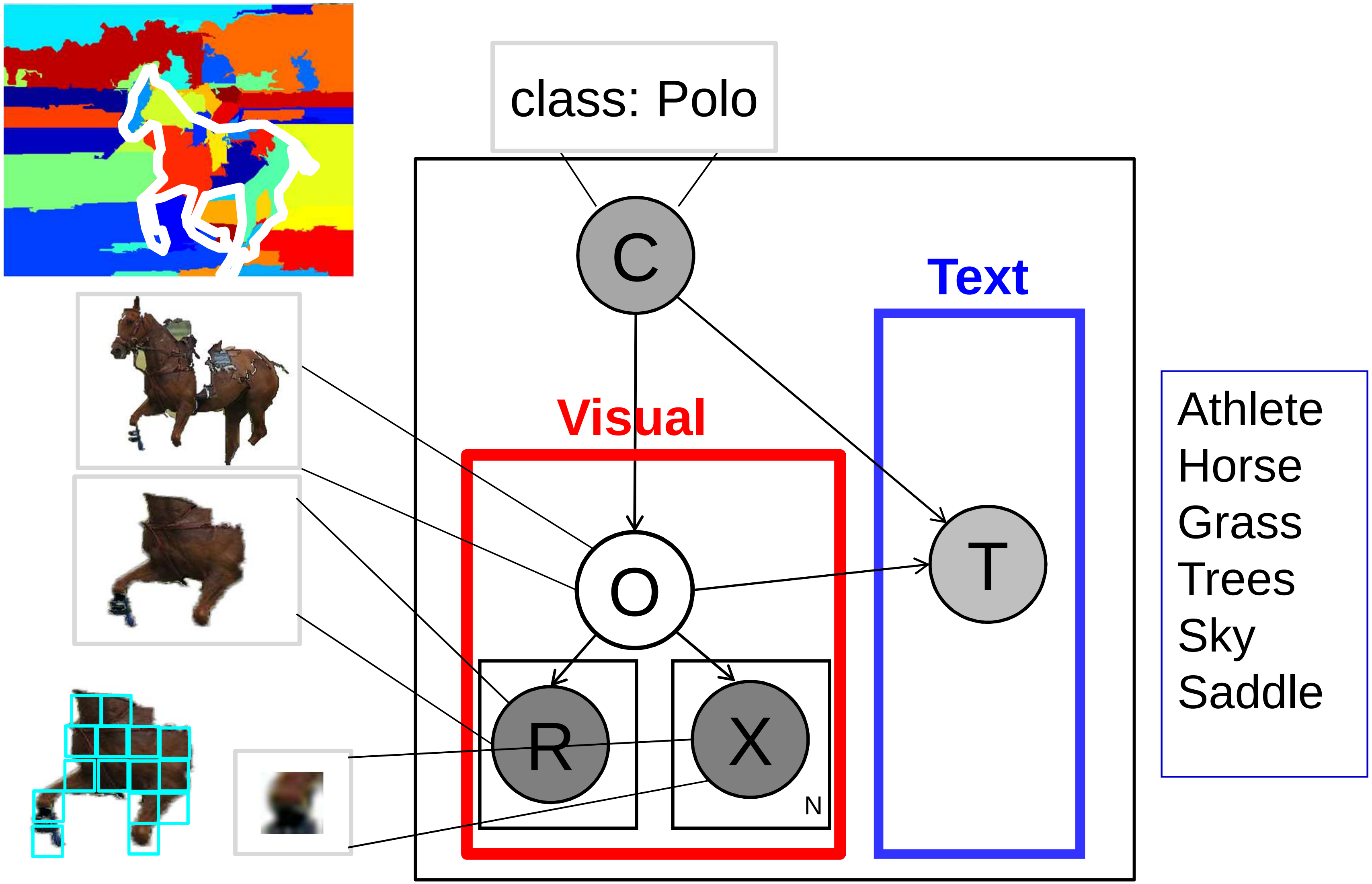
A joint model for image classification, annotation & segmentation



$$p(C,\mathbf{O},\mathbf{R},\mathbf{X},\mathbf{S},\mathbf{T},\mathbf{Z}|\eta,\alpha,\beta,\gamma,\theta,\varphi)= p(C)\cdot\left(\prod_{n=1}^{N_r} p(O_n|\eta,C)\right)\prod_{n=1}^{N_r}\left(\left(\prod_{i=1}^{N_F} p(R_{ni}|O_n,\alpha_i)\right)\cdot\prod_{r=1}^{A_r} p(X_{nr}|O_n,\beta)\right)$$

$\cdot$ **Text Component**

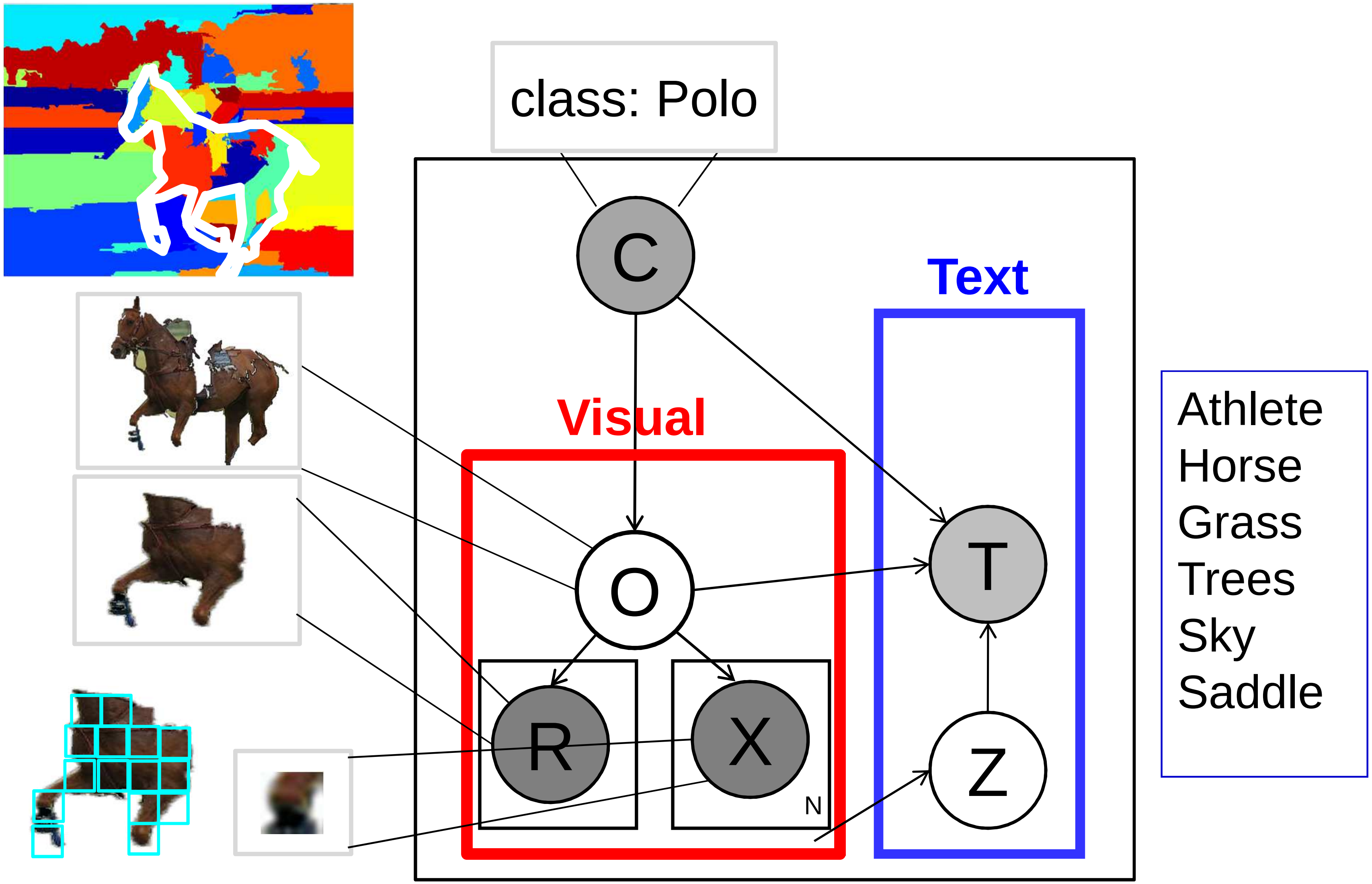
A joint model for image classification, annotation & segmentation



$$p(C, \mathbf{O}, \mathbf{R}, \mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{Z} | \eta, \alpha, \beta, \gamma, \theta, \varphi) = p(C) \cdot \boxed{\left(\prod_{n=1}^{N_r} p(O_n|\eta, C)\right) \prod_{n=1}^{N_r} \left(\left(\prod_{i=1}^{N_F} p(R_{ni}|O_n, \alpha_i)\right) \cdot \prod_{r=1}^{A_r} p(X_{nr}|O_n, \beta)\right.}$$

$$\cdot \boxed{p(T_m|O_{Z_m}, S_m, \theta, C, \varphi)}$$

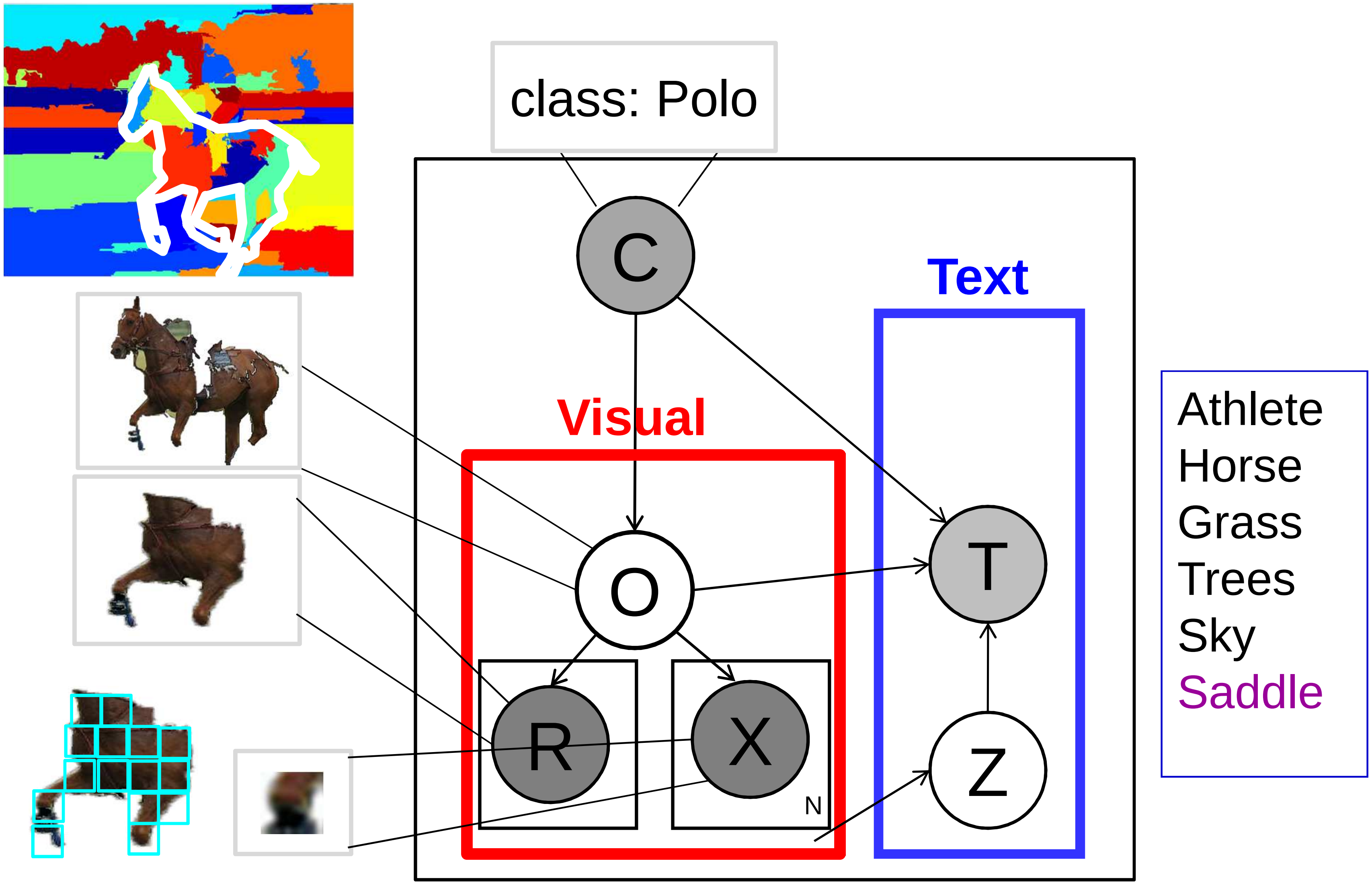
A joint model for image classification, annotation & segmentation



$$p(C, \mathbf{O}, \mathbf{R}, \mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{Z} | \eta, \alpha, \beta, \gamma, \theta, \varphi) = p(C) \cdot \left(\prod_{n=1}^{N_r} p(O_n|\eta, C)\right) \prod_{n=1}^{N_r} \left(\left(\prod_{i=1}^{N_F} p(R_{ni}|O_n, \alpha_i)\right) \cdot \prod_{r=1}^{A_r} p(X_{nr}|O_n, \beta\right)$$

$$\cdot \prod_{m=1}^{N_t} p(Z_m|N_r) \quad p(T_m|O_{Z_m}, S_m, \theta, C, \varphi)$$

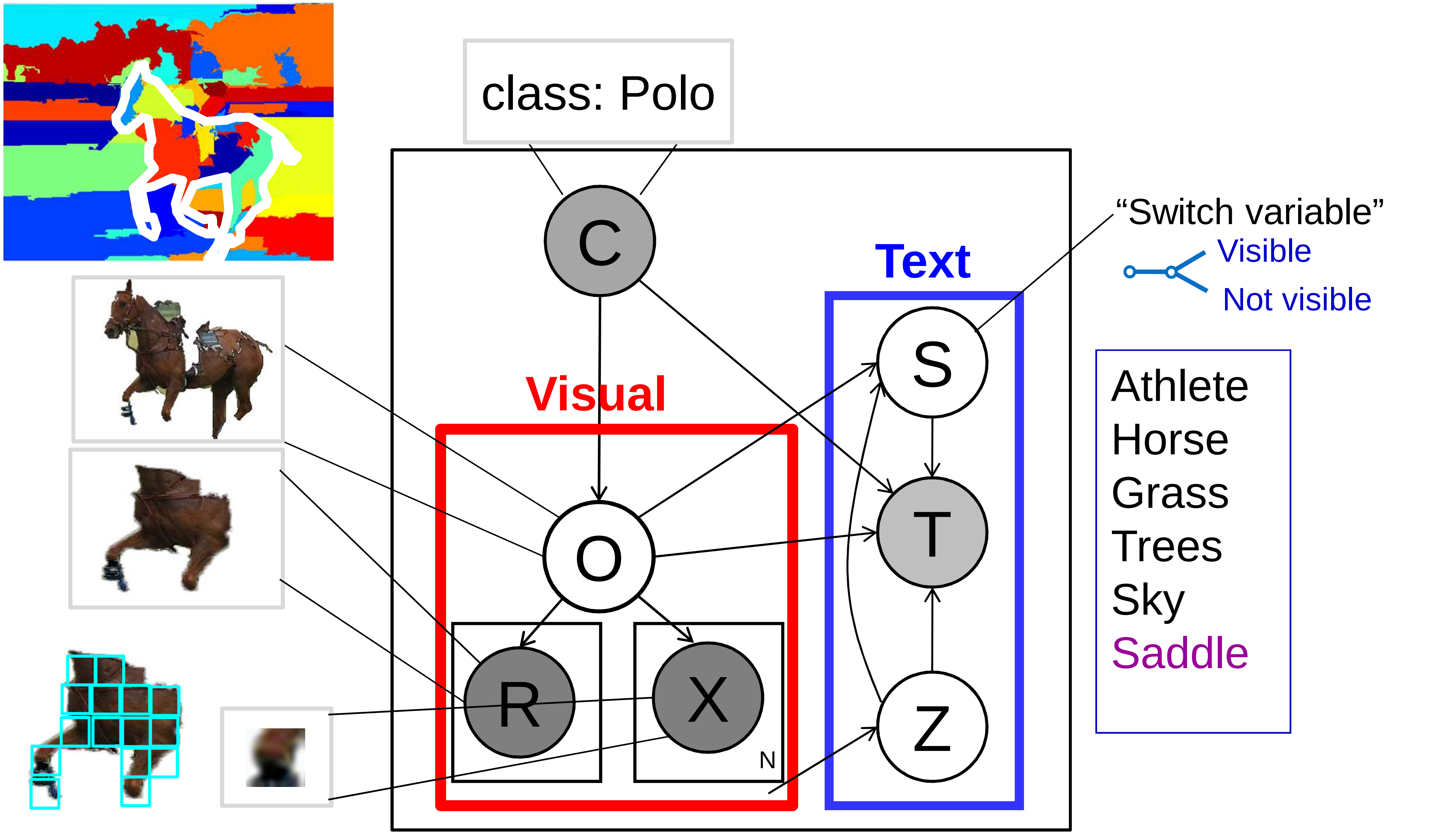
A joint model for image classification, annotation & segmentation



$$p(C,\mathbf{O},\mathbf{R},\mathbf{X},\mathbf{S},\mathbf{T},\mathbf{Z}|\eta,\alpha,\beta,\gamma,\theta,\varphi)= p(C)\cdot\Big(\prod_{n=1}^{N_r} p(O_n|\eta,C)\Big)\prod_{n=1}^{N_r}\Big(\Big(\prod_{i=1}^{N_F} p(R_{ni}|O_n,\alpha_i)\Big)\cdot\prod_{r=1}^{A_r} p(X_{nr}|O_n,\beta)\Big)$$

$$\cdot\prod_{m=1}^{N_t} p(Z_m|N_r) \quad p(T_m|O_{Z_m},S_m,\theta,C,\varphi)$$

A joint model for image classification, annotation & segmentation



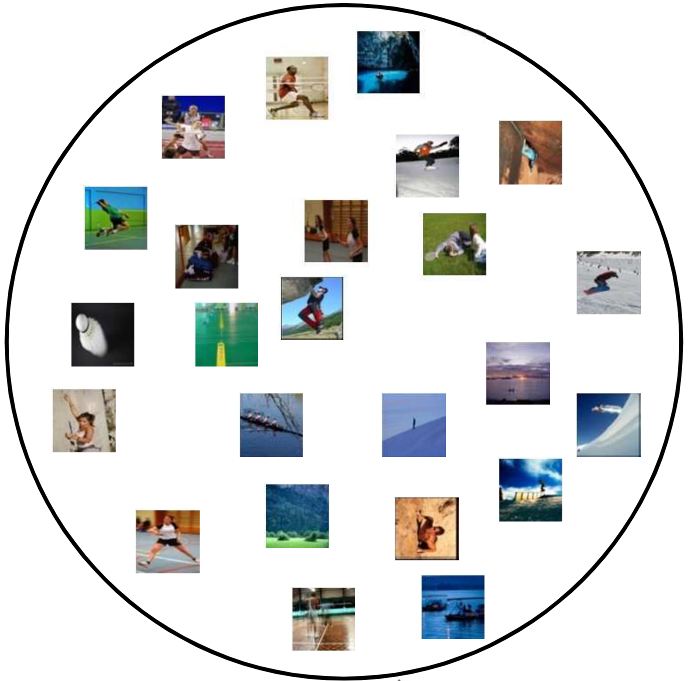
$$p(C, \mathbf{O}, \mathbf{R}, \mathbf{X}, \mathbf{S}, \mathbf{T}, \mathbf{Z} | \eta, \alpha, \beta, \gamma, \theta, \varphi) = p(C) \cdot \left(\prod_{n=1}^{N_r} p(O_n|\eta, C)\right) \prod_{n=1}^{N_r} \left(\left(\prod_{i=1}^{N_F} p(R_{ni}|O_n, \alpha_i)\right) \cdot \prod_{r=1}^{A_r} p(X_{nr}|O_n, \beta)\right)$$

$$\cdot \prod_{m=1}^{N_t} p(Z_m|N_r) \quad p(S_m|O_{Z_m}, \gamma) \quad p(T_m|O_{Z_m}, S_m, \theta, C, \varphi)$$

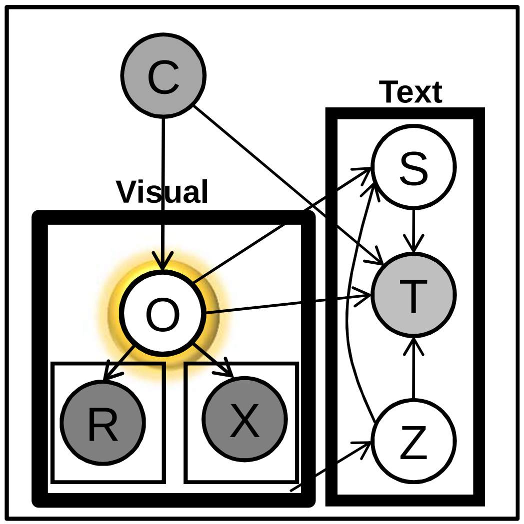
Auto-semi-supervised learning:

Small # of initialized images + Large # of uninitialized images

Scene/Event images

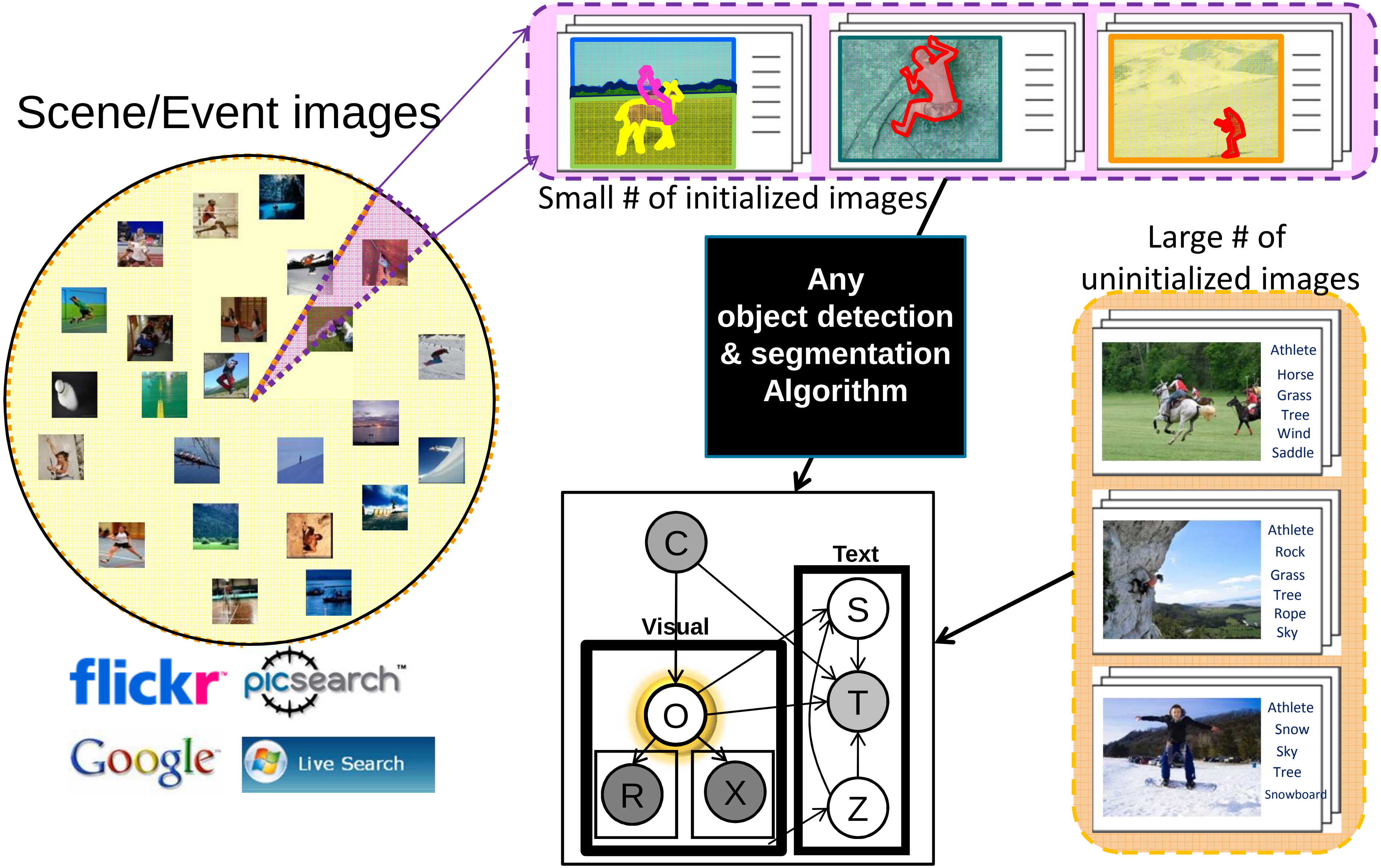


flickr picsearch Google Live Search



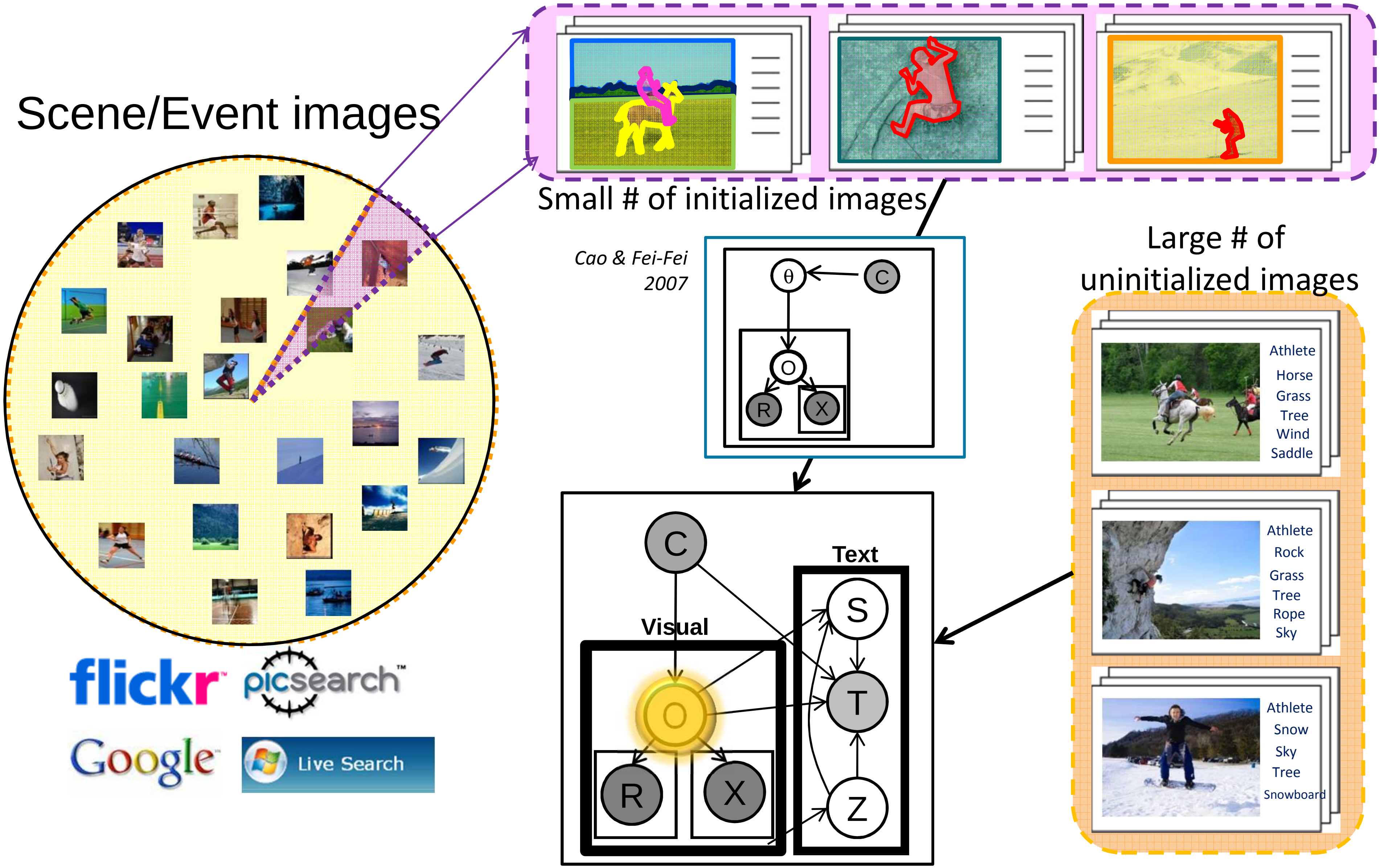
Auto-semi-supervised learning:

Small # of initialized images + Large # of uninitialized images



Auto-semi-supervised learning:

Small # of initialized images + Large # of uninitialized images



flickr 8 Event/Scene Classes

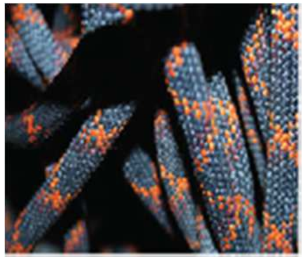
Badminton


Bocce


Croquet


Polo


Rock climbing



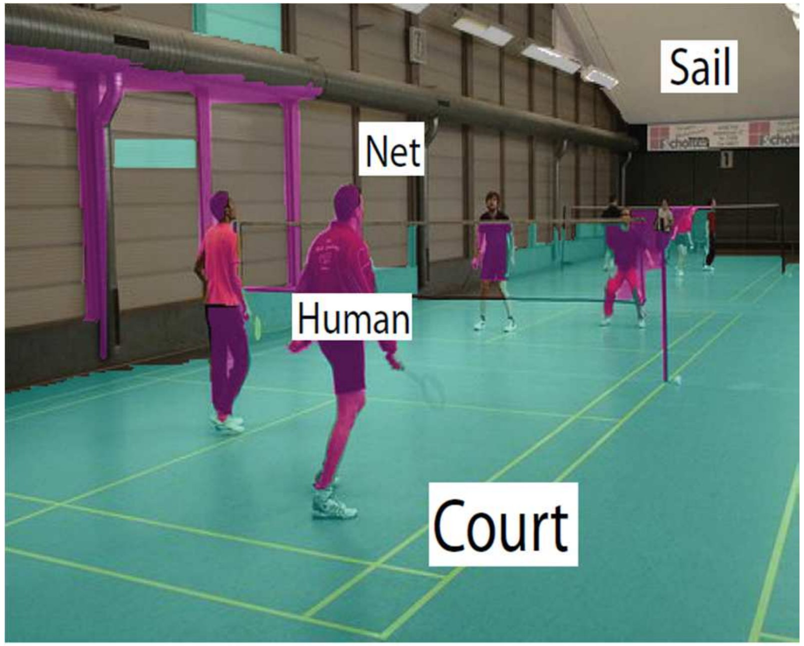

Rowing


Sailing

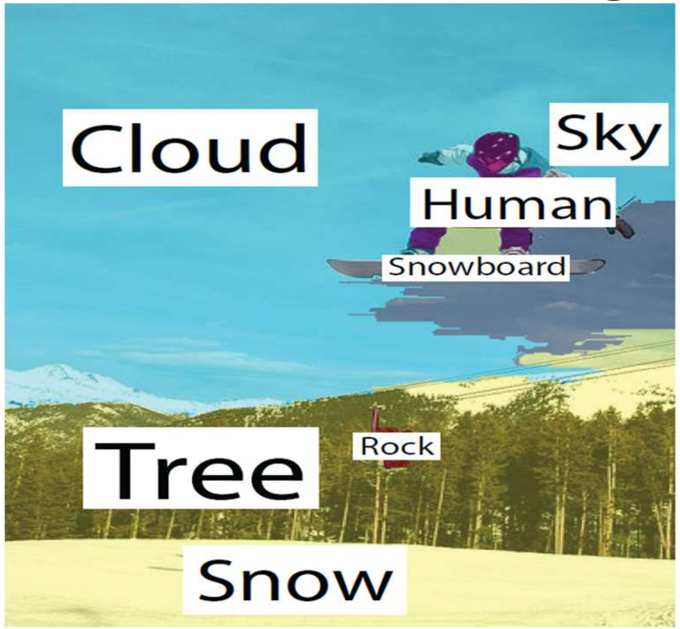

Snow boarding


Remark: 200 images per event classes; 40 visual objects; 1200 concept tags

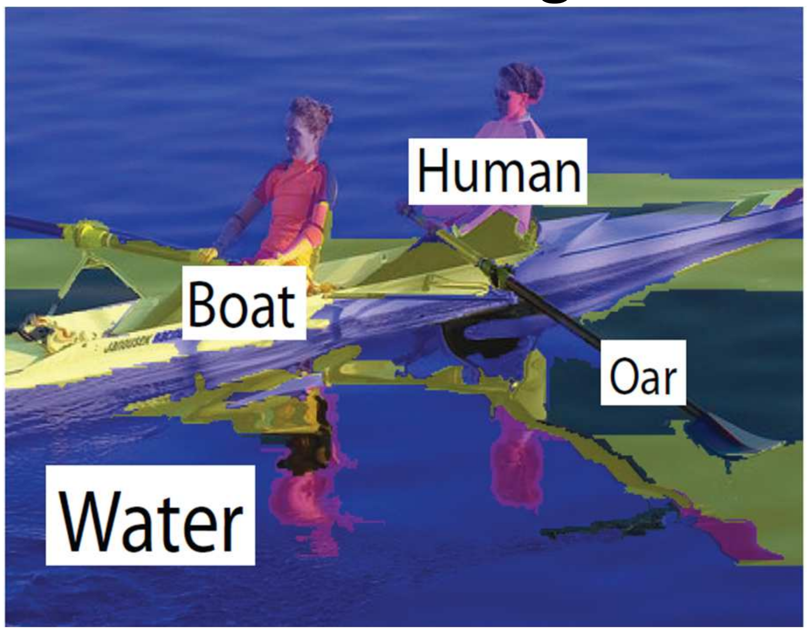
Class: Badminton



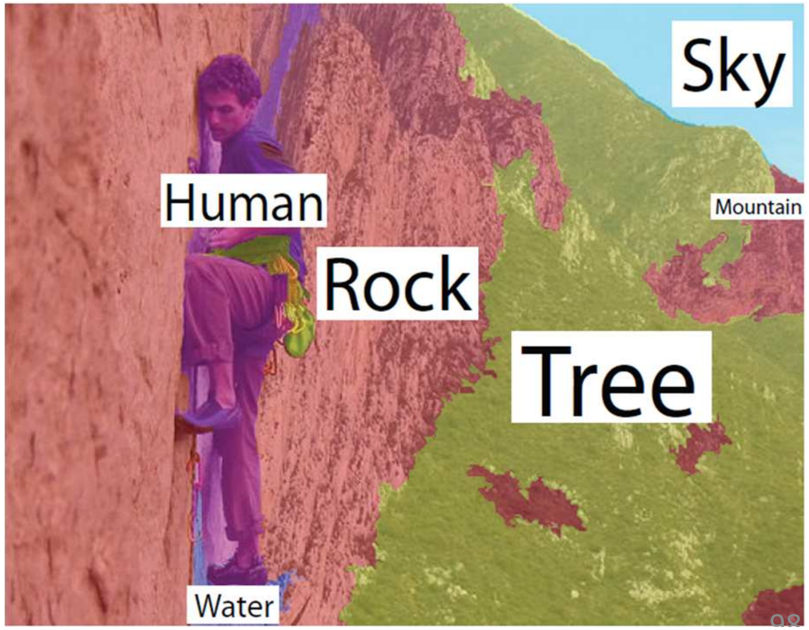
Class: Snowboarding



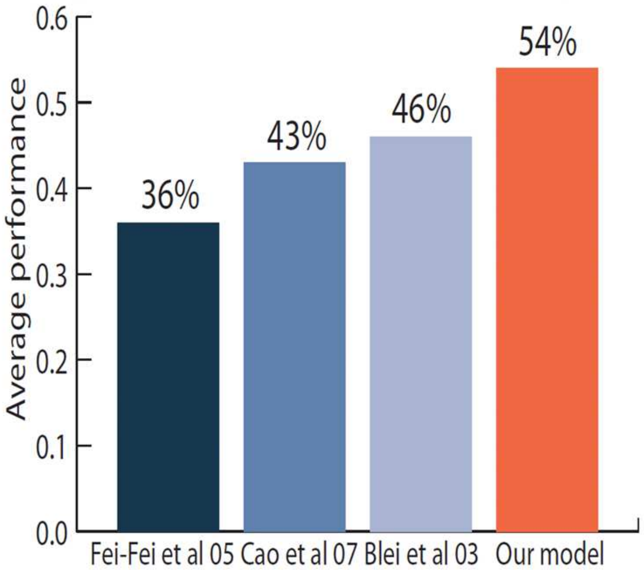
Class: Rowing



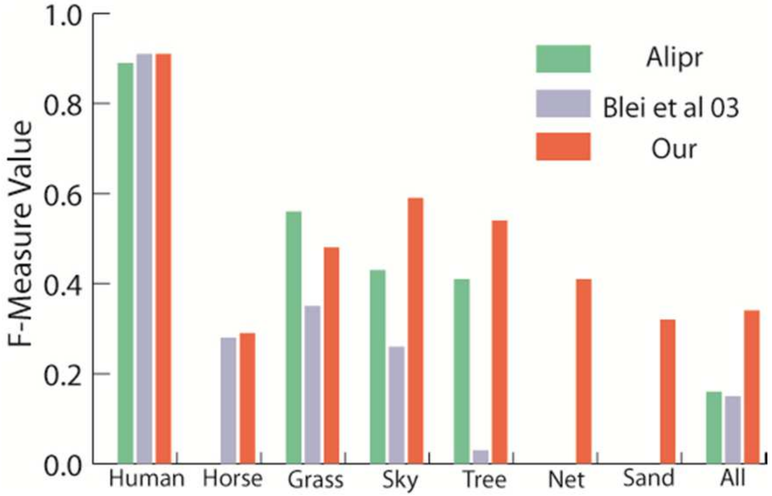
Class: Rock Climbing



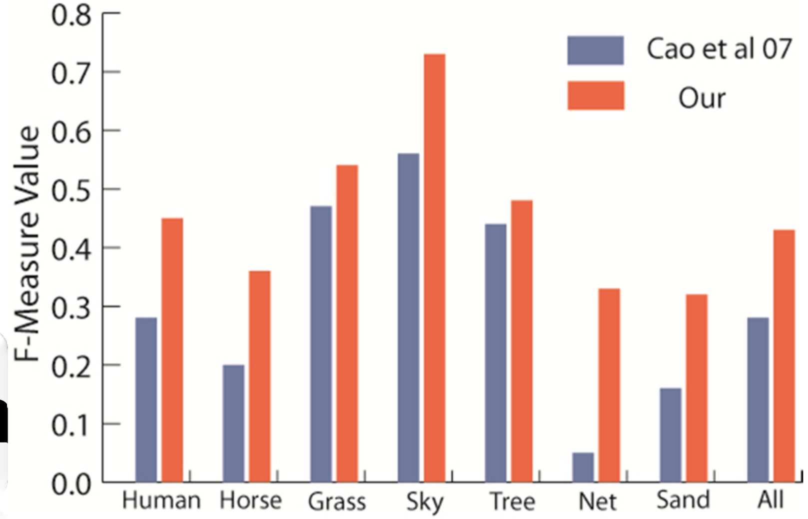
Classification



Annotation

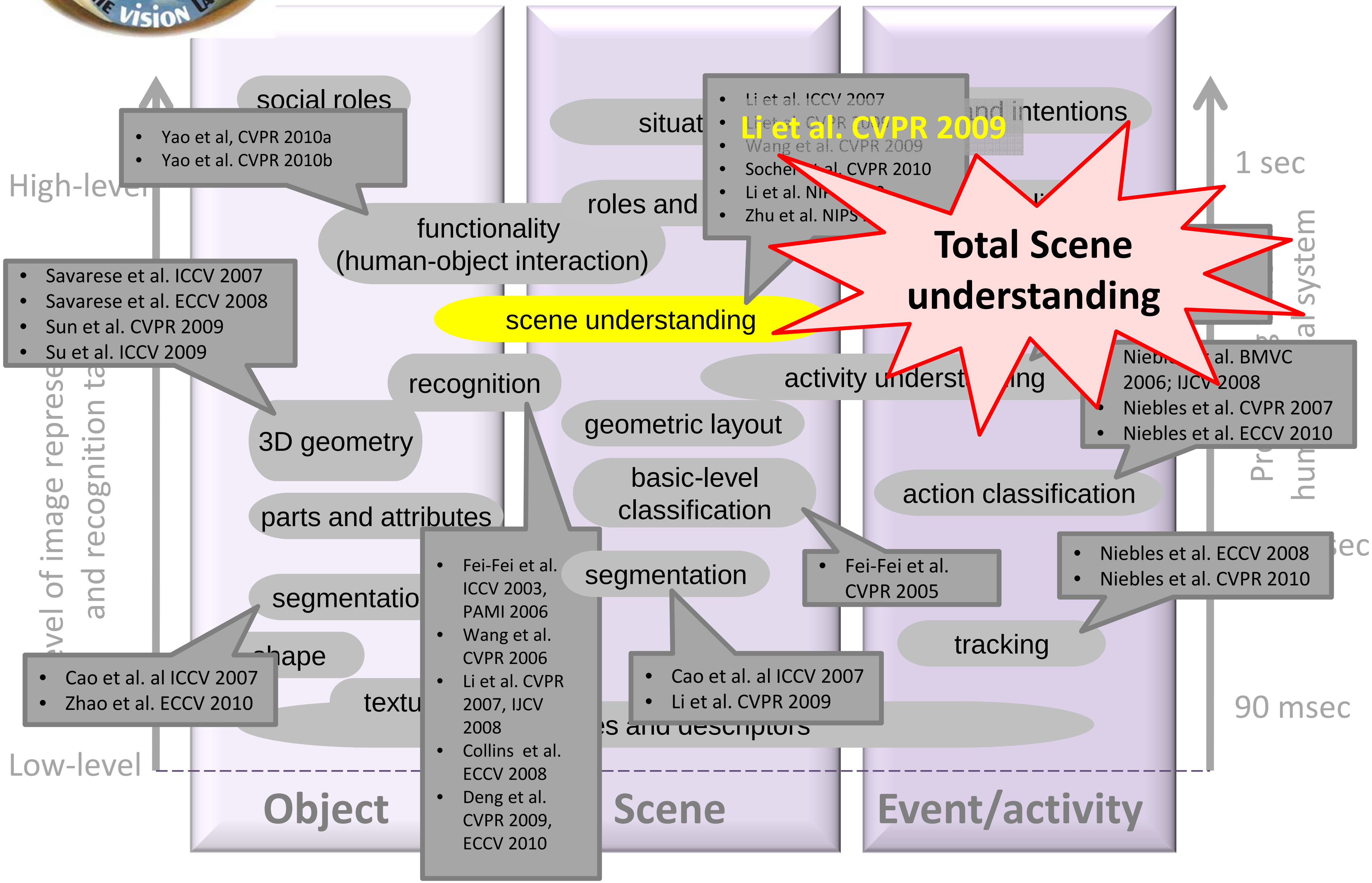


Segmentation



L.-J. Li, R. Socher & L. Fei-Fei, CVPR 2009

Story telling in images



Story telling

Yao et al. CVPR 2010a

- Yao et al. CVPR 2010a
- Yao et al. CVPR 2010b

social roles

High-level

functionality
(human-object interaction)

- Savarese et al. ICCV 2007
- Savarese et al. ECCV 2008
- Sun et al. CVPR 2009
- Su et al. ICCV 2009

Structure & Context

3D geometry

parts and attributes

segmentation

shape

texture

- Cao et al. al ICCV 2007
- Zhao et al. ECCV 2010

Level of image representation and recognition tasks

Low-level

Object

- Fei-Fei et al. ICCV 2003, PAMI 2006
- Wang et al. CVPR 2006
- Li et al. CVPR 2007, IJCV 2008
- Collins et al. ECCV 2008
- Deng et al. CVPR 2009, ECCV 2010

situa

roles and

sc understa

basic classi

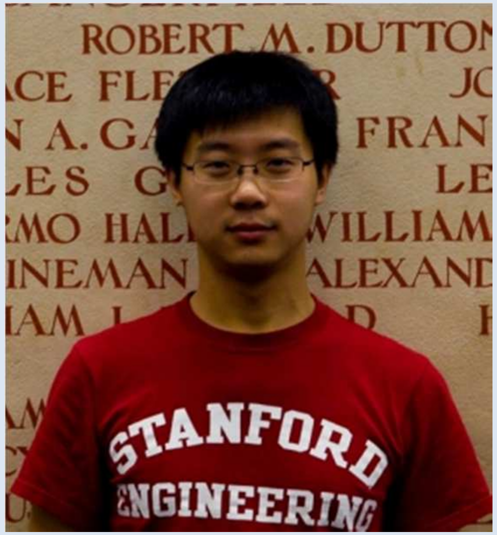
segment

- Ca
- Li

es and

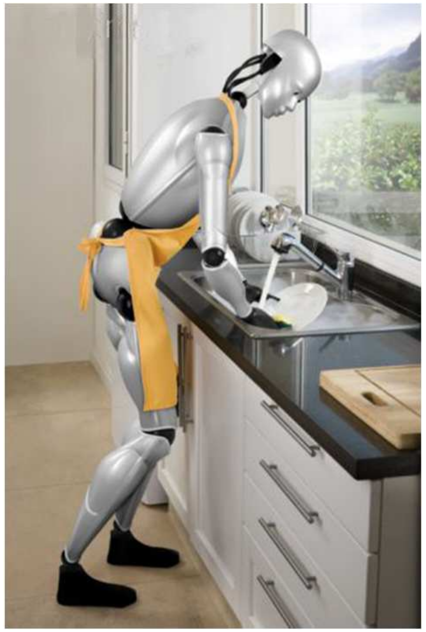
Scen

B. Yao and L. Fei-Fei. **Modeling Mutual Context of Object and Human Pose in Human-Object Interaction Activities.** *IEEE Computer Vision and Pattern Recognition (CVPR).* 2010.



Bangpeng Yao
Stanford University

Human-Object Interaction



Robots interact with objects

Automatic sports commentary

"Kobe is dunking the ball."

Medical care

Human-Object Interaction

Holistic image based classification (Grouplet (Yao & Fei-Fei, CVPR 2010b), Gupta et al. 2007, Yang et al. 2010, Delaitre et al. 2010)

Detailed understanding and reasoning

HOI activity: Tennis Forehand

Human-Object Interaction

Holistic image based classification

Detailed understanding and reasoning

- **Human pose estimation**

Human-Object Interaction

Holistic image based classification

Detailed understanding and reasoning

- Human pose estimation
- **Object detection**

Human-Object Interaction

Holistic image based classification

Detailed understanding and reasoning

- **Human pose estimation**
- **Object detection**

HOI activity: Tennis Forehand

Outline

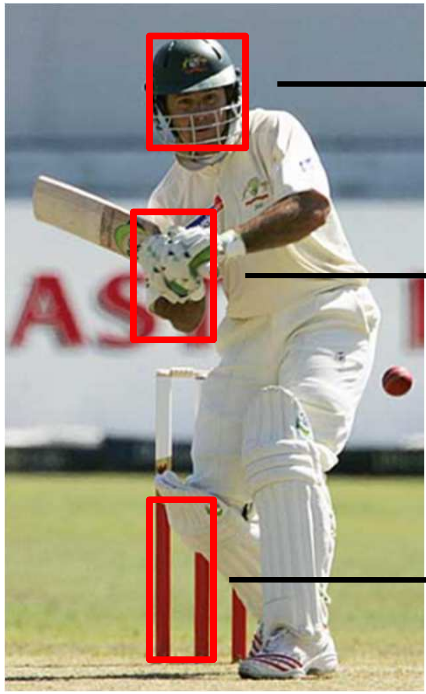
- Background and Intuition
- Mutual Context of Object and Human Pose
    - Model Representation
    - Model Learning
    - Model Inference
- Experiments
- Conclusion

Outline

- Background and Intuition
- Mutual Context of Object and Human Pose
  - Model Representation
  - Model Learning
  - Model Inference
- Experiments
- Conclusion

Human pose estimation & Object detection

Human pose estimation is challenging.



Difficult part appearance

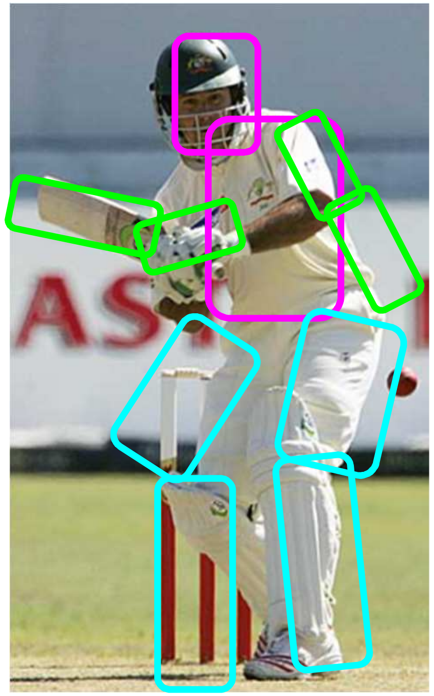
Self-occlusion

Image region looks like a body part

- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

Human pose estimation & Object detection

Human pose estimation is challenging.

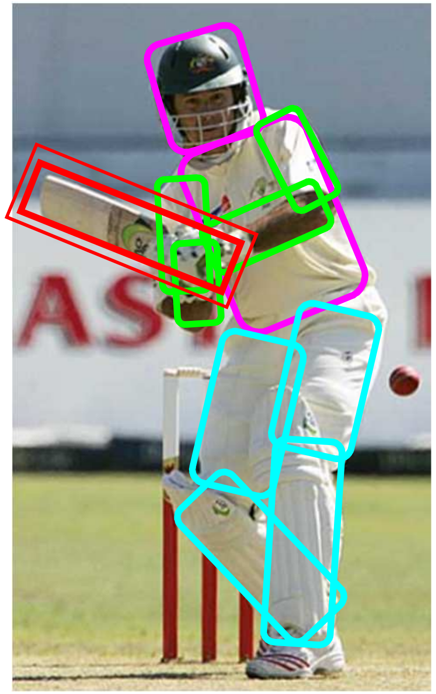


- Felzenszwalb & Huttenlocher, 2005
- Ren et al, 2005
- Ramanan, 2006
- Ferrari et al, 2008
- Yang & Mori, 2008
- Andriluka et al, 2009
- Eichner & Ferrari, 2009

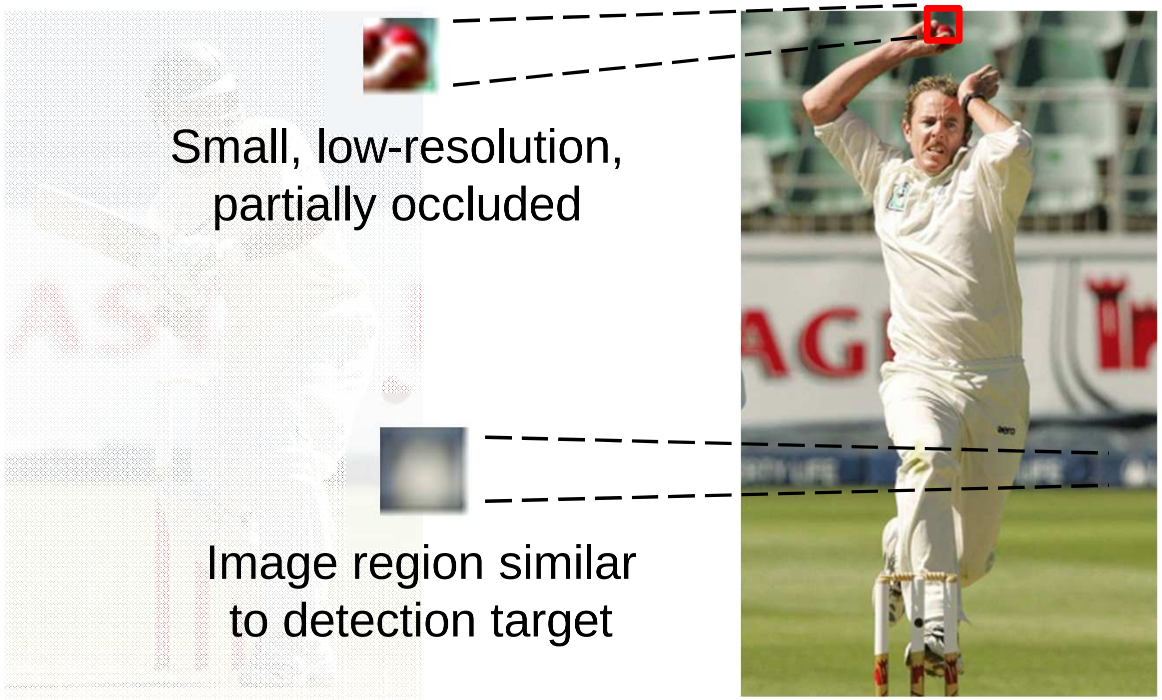
Human pose estimation & Object detection

Facilitate

Given the object is detected.



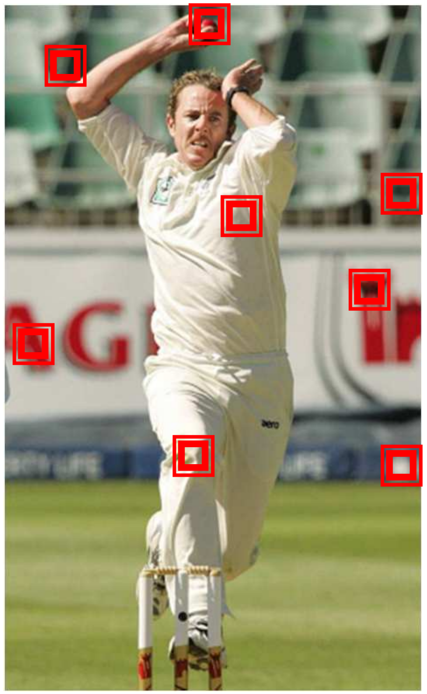
Human pose estimation & Object detection



Object detection is challenging

- Viola & Jones, 2001
- Lampert et al, 2008
- Divvala et al, 2009
- Vedaldi et al, 2009

Human pose estimation & Object detection

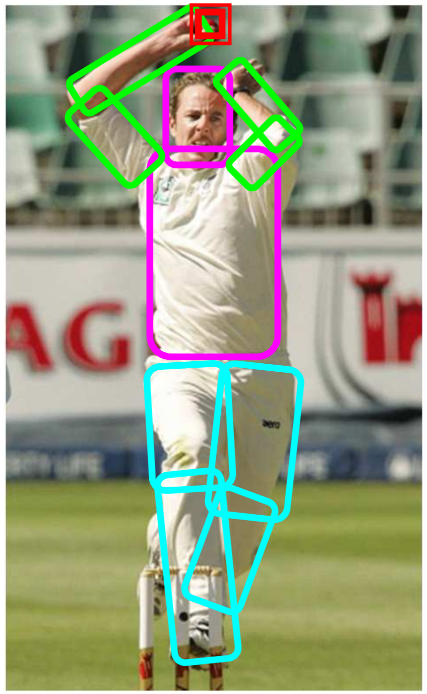


Object detection is challenging

- Viola & Jones, 2001
- Lampert et al, 2008
- Divvala et al, 2009
- Vedaldi et al, 2009

Human pose estimation & Object detection

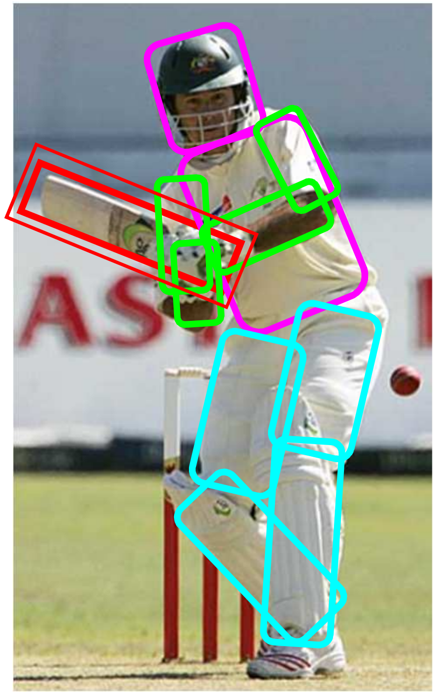
Facilitate

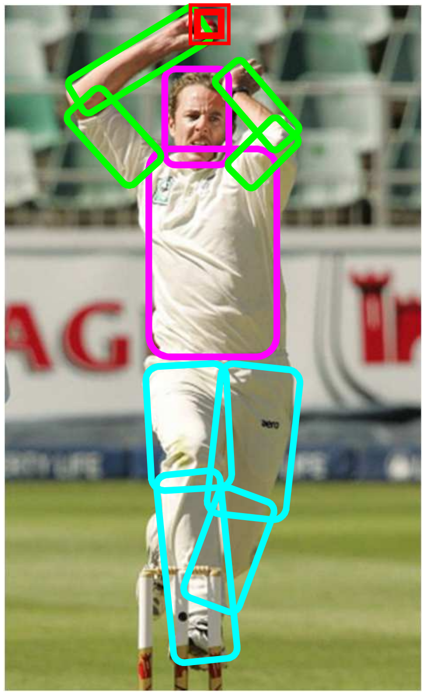


Given the pose is estimated.

Human pose estimation & Object detection

Mutual Context





Outline

- Background and Intuition
- Mutual Context of Object and Human Pose
  - Model Representation
  - Model Learning
  - Model Inference
- Experiments
- Conclusion

Mutual Context Model Representation

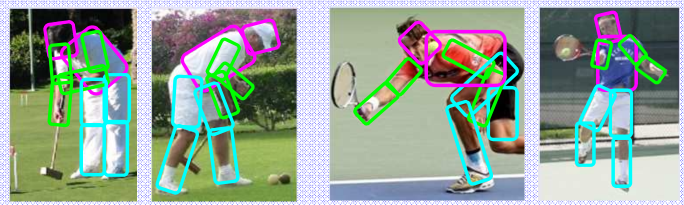
A:

Tennis forehand | Croquet shot | Volleyball smash | …

O:

Tennis racket | Croquet mallet | Volleyball | …

H:

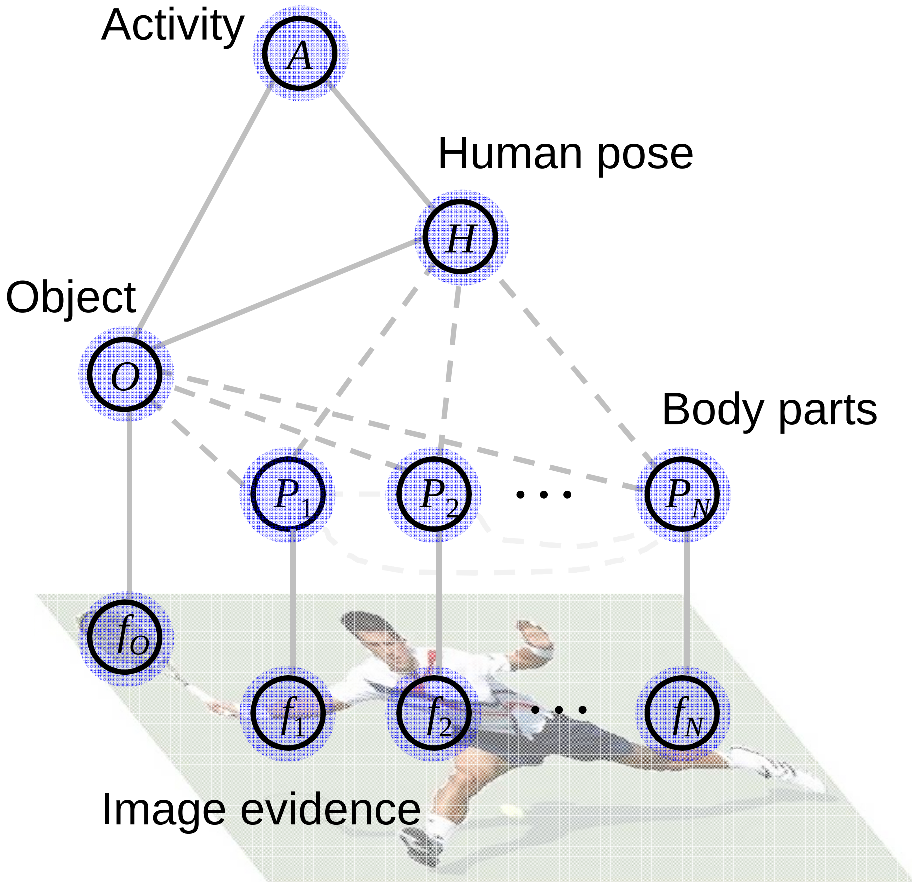


Intra-class variations

- More than one H for each A;
- **Unobserved** during training.

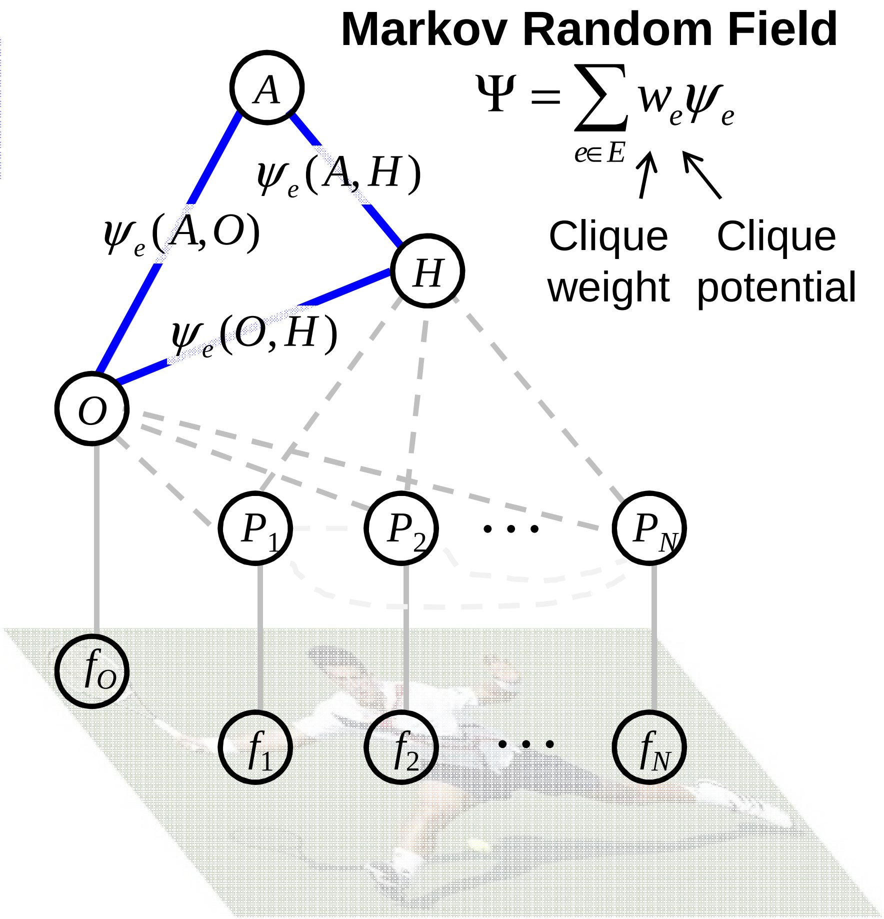
P: l_P: location; θ_P: orientation; s_P: scale.

f: Shape context. [Belongie et al, 2002]



Mutual Context Model Representation

- $\psi_e(A,O)$, $\psi_e(A,H)$, $\psi_e(O,H)$: Frequency of co-occurrence between A, O, and H.



Mutual Context Model Representation

- $\psi_e(A,O)$, $\psi_e(A,H)$, $\psi_e(O,H)$: Frequency of co-occurrence between A, O, and H.

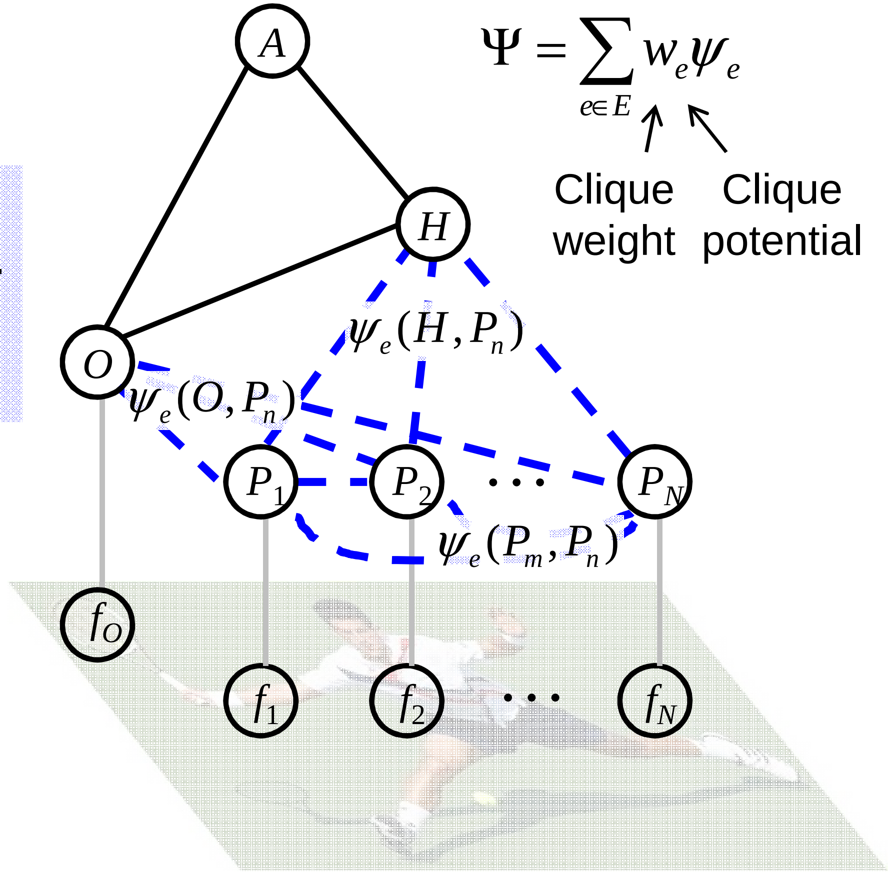
- $\psi_e(O,P_n)$, $\psi_e(H,P_n)$, $\psi_e(P_m,P_n)$: Spatial relationship among object and body parts.

$$\underbrace{\mathrm{bin}\left(l_O - l_{P_n}\right)}_{\textbf{location}} \cdot \underbrace{\mathrm{bin}\left(\theta_O - \theta_{P_n}\right)}_{\textbf{orientation}} \cdot \underbrace{\mathrm{N}\left(s_O / s_{P_n}\right)}_{\textbf{size}}$$

Markov Random Field

$$\Psi = \sum_{e \in E} w_e \psi_e$$

Clique weight Clique potential



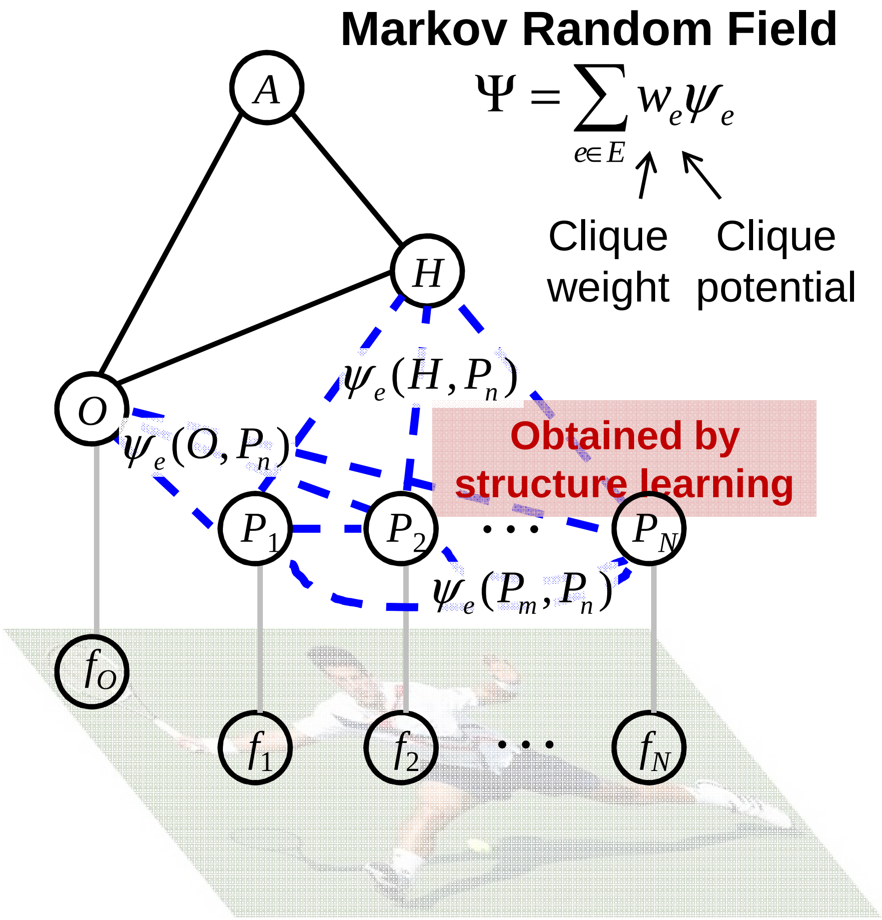
Mutual Context Model Representation

- $\psi_e(A,O)$, $\psi_e(A,H)$, $\psi_e(O,H)$: Frequency of co-occurrence between A, O, and H.

- $\psi_e(O,P_n)$, $\psi_e(H,P_n)$, $\psi_e(P_m,P_n)$: Spatial relationship among object and body parts.

$$\underbrace{\mathrm{bin}\left(l_O - l_{P_n}\right)}_{\textbf{location}} \cdot \underbrace{\mathrm{bin}\left(\theta_O - \theta_{P_n}\right)}_{\textbf{orientation}} \cdot \underbrace{\mathrm{N}\left(s_O / s_{P_n}\right)}_{\textbf{size}}$$

- Learn structural connectivity among the body parts and the object.



Mutual Context Model Representation

- $\psi_e(A,O)$, $\psi_e(A,H)$, $\psi_e(O,H)$: Frequency of co-occurrence between A, O, and H.

- $\psi_e(O,P_n)$, $\psi_e(H,P_n)$, $\psi_e(P_m,P_n)$: Spatial relationship among object and body parts.

$$\underbrace{\mathrm{bin}\left(l_O - l_{P_n}\right)}_{\textbf{location}} \cdot \underbrace{\mathrm{bin}\left(\theta_O - \theta_{P_n}\right)}_{\textbf{orientation}} \cdot \underbrace{\mathrm{N}\left(s_O / s_{P_n}\right)}_{\textbf{size}}$$

- Learn structural connectivity among the body parts and the object.

- $\psi_e(O, f_O)$ and $\psi_e(P_n, f_{P_n})$: Discriminative part detection scores.

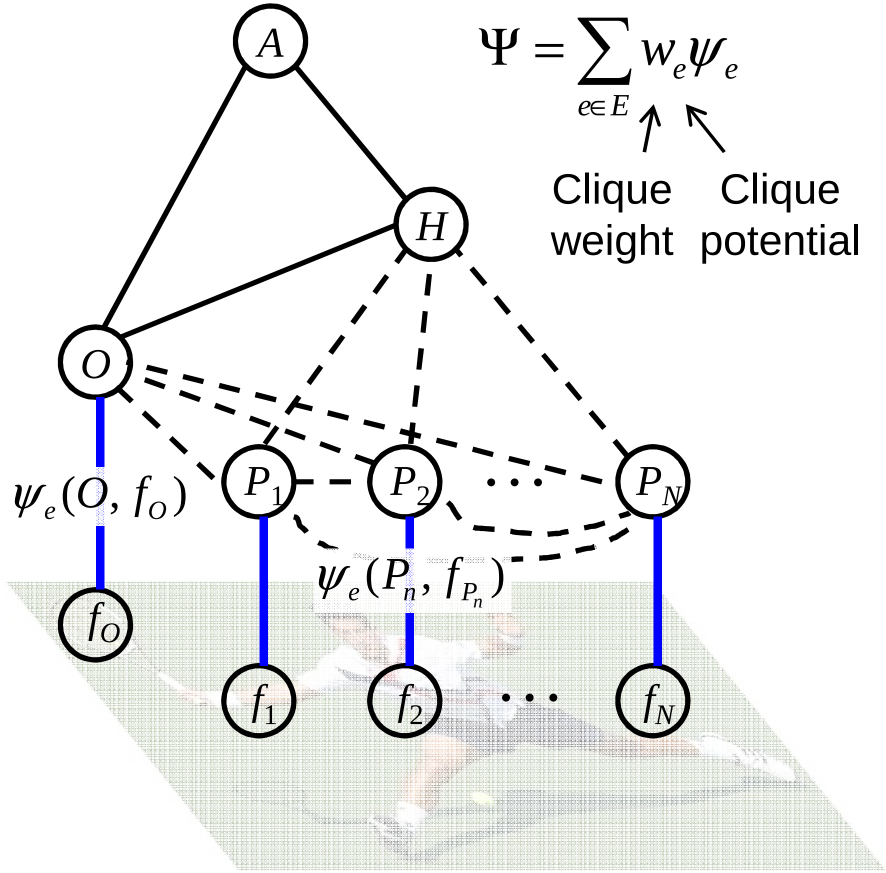
  Shape context + AdaBoost

  [Andriluka et al, 2009]
  [Belongie et al, 2002]
  [Viola & Jones, 2001]

Markov Random Field

$$\Psi = \sum_{e \in E} w_e \psi_e$$

Clique weight — Clique potential

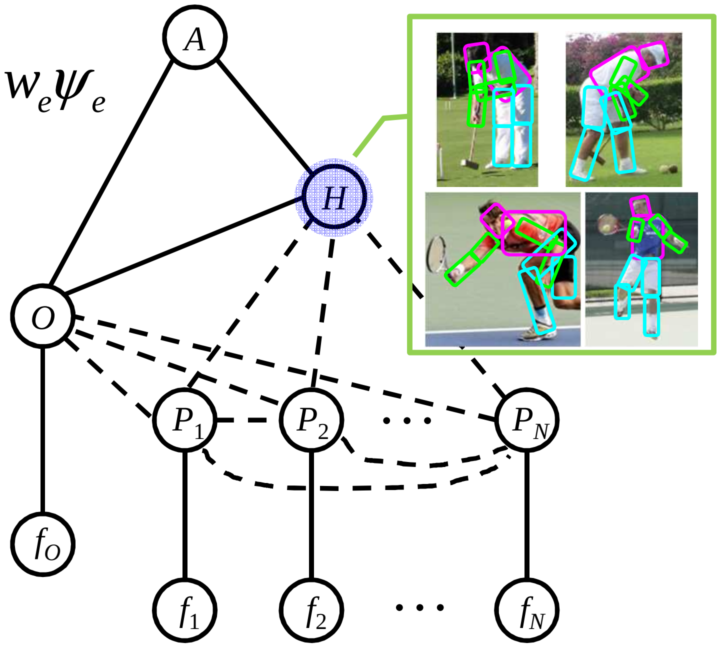


Outline

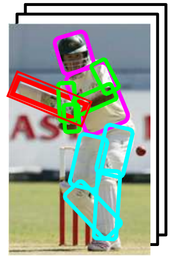
- Background and Intuition
- Mutual Context of Object and Human Pose
  - Model Representation
  - Model Learning
  - Model Inference
- Experiments
- Conclusion

Model Learning

$$\Psi = \sum_{e \in E} w_e \psi_e$$



Input:



cricket shot

cricket bowling

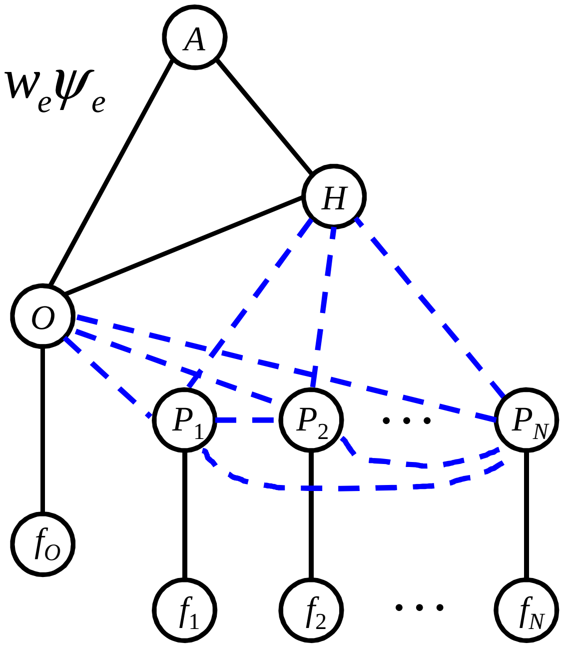
. . .

Goals:

Hidden human poses

Model Learning

$$\Psi = \sum_{e \in E} w_e \psi_e$$



Input:

cricket shot

cricket bowling

. . .

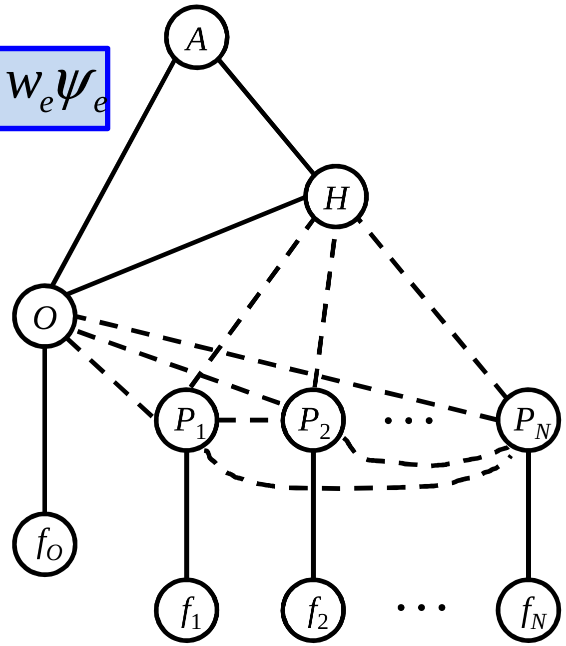
Goals:

Hidden human poses

Structural connectivity

Model Learning

$$\Psi = \sum_{e \in E} \boxed{w_e \psi_e}$$



Input:

cricket shot

cricket bowling

…

Goals:

Hidden human poses

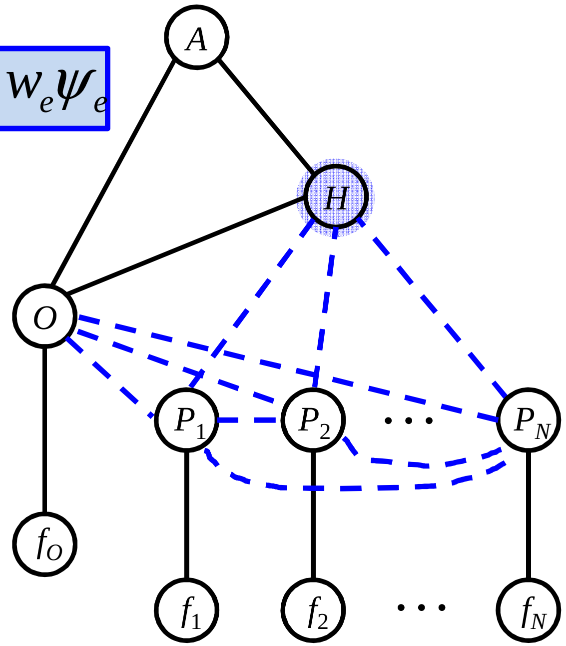
Structural connectivity

Potential parameters

Potential weights

Model Learning

$$\Psi = \sum_{e \in E} \boxed{w_e \psi_e}$$



Input:

cricket shot   cricket bowling   . . .

Goals:

Hidden human poses $\longrightarrow$ **Hidden variables**

Structural connectivity $\longrightarrow$ **Structure learning**

Potential parameters
Potential weights } **Parameter estimation**

Learning Results

Cricket defensive shot

Cricket bowling

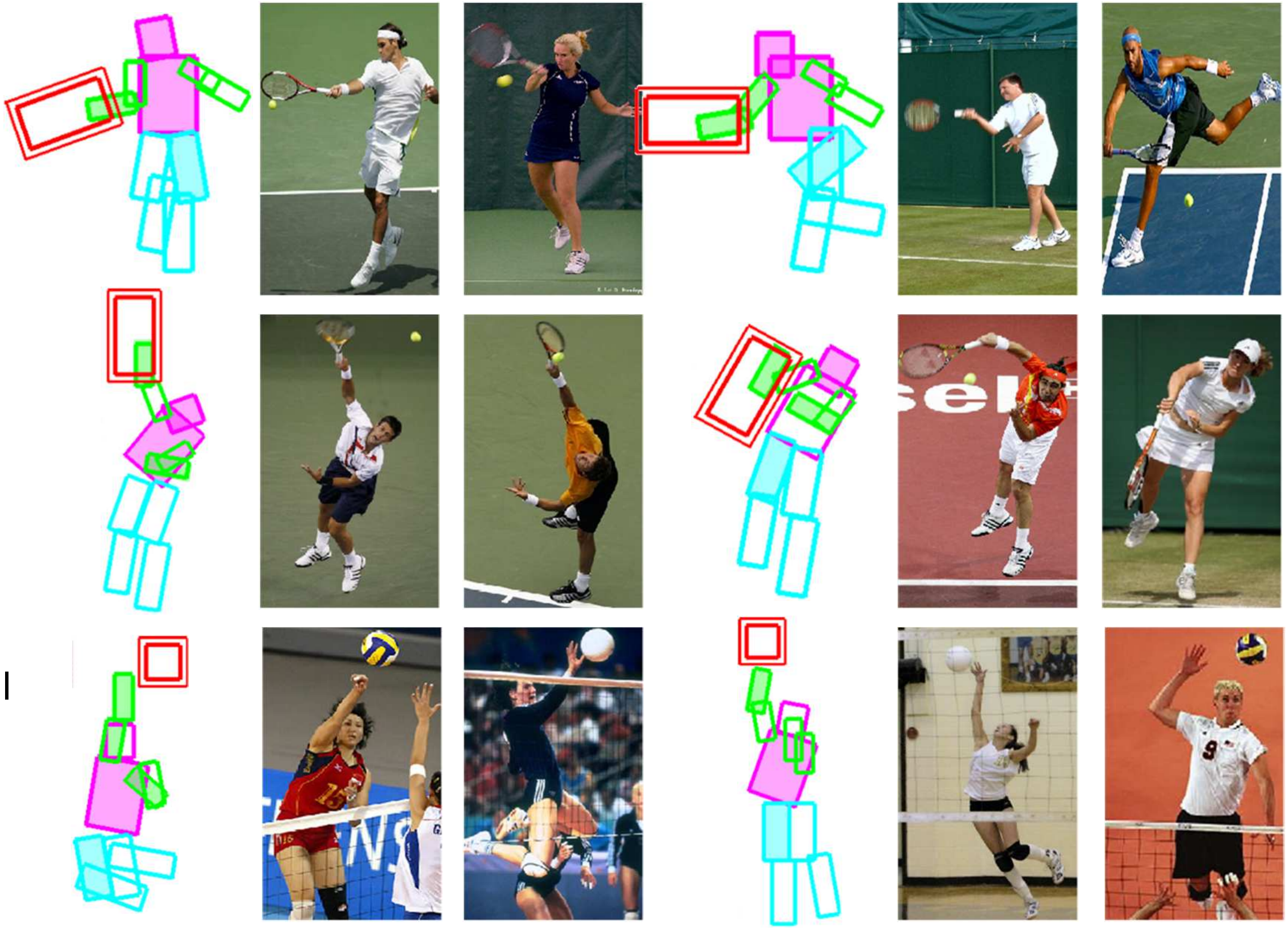
Croquet shot

Learning Results

Tennis forehand

Tennis serve

Volleyball smash



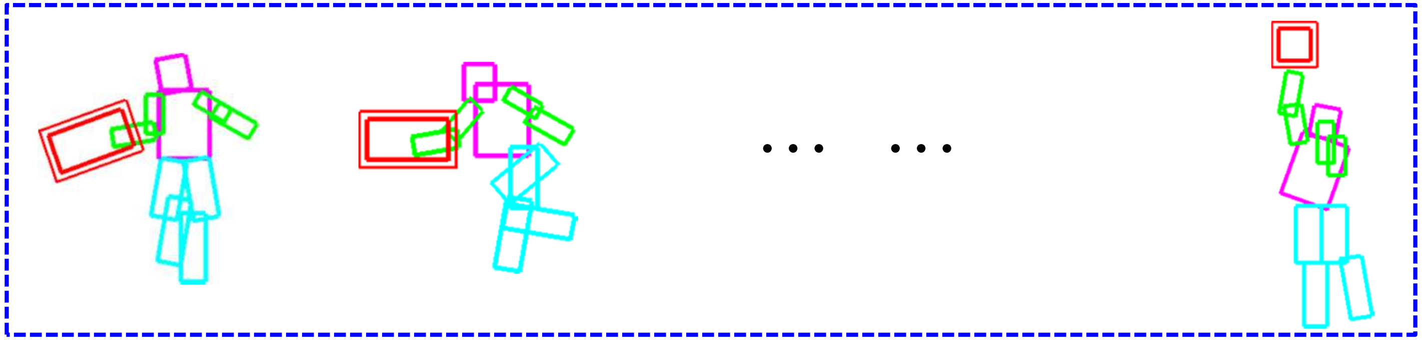
Outline

- Background and Intuition
- Mutual Context of Object and Human Pose
  - Model Representation
  - Model Learning
  - Model Inference
- Experiments
- Conclusion

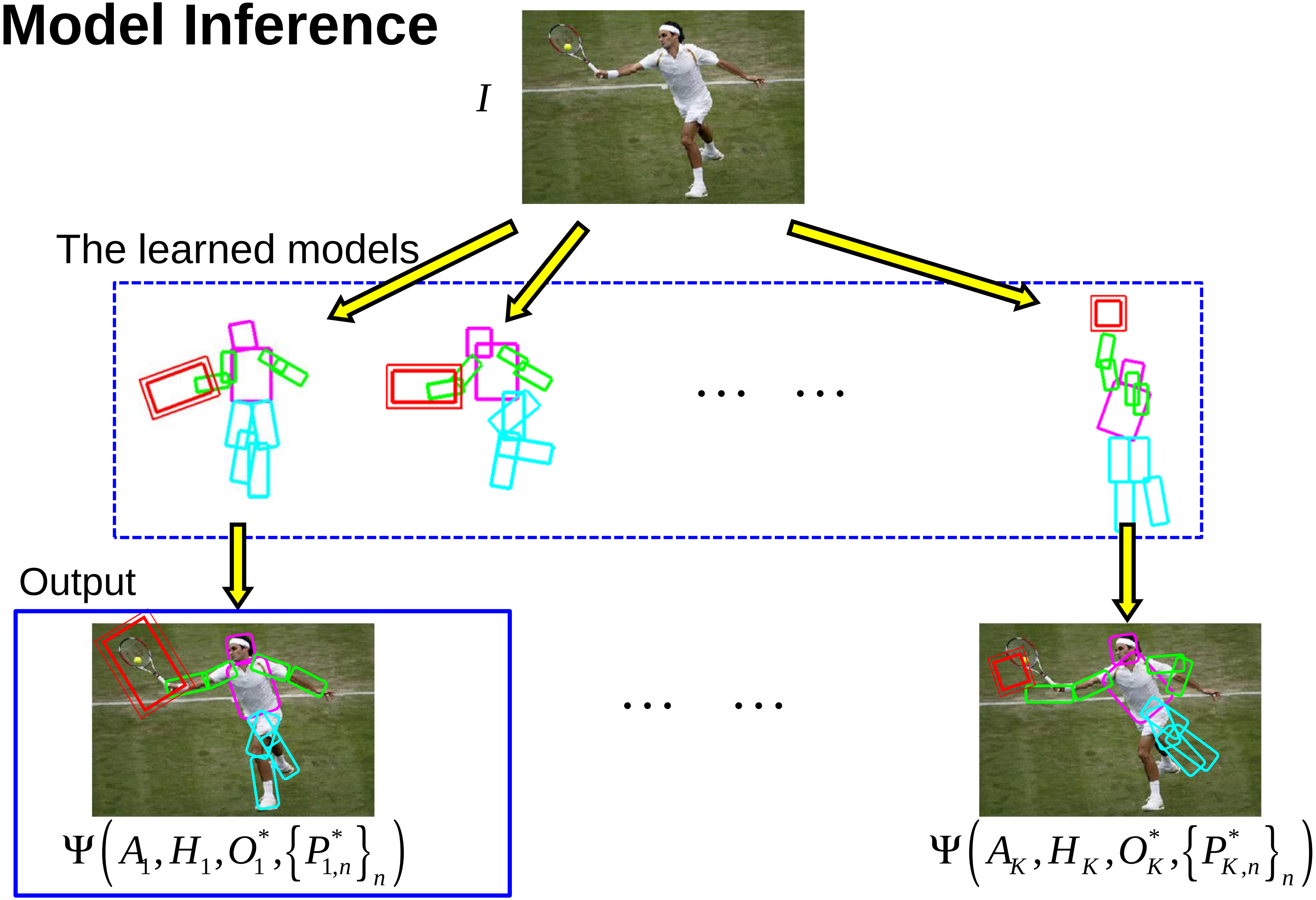
Model Inference

I

The learned models



Model Inference



Outline

- Background and Intuition
- Mutual Context of Object and Human Pose
  - Model Representation
  - Model Learning
  - Model Inference
- Experiments
- Conclusion

Dataset and Experiment Setup

Sport data set: 6 classes

180 training (supervised with object and part locations) & 120 testing images

Cricket defensive shot

Cricket bowling

Croquet shot

Tennis forehand

Tennis serve

Volleyball smash

[Gupta et al, 2009]

Tasks:

- Object detection;
- Pose estimation;
- Activity classification.

Dataset and Experiment Setup

Sport data set: 6 classes

180 training (supervised with object and part locations) & 120 testing images

Cricket defensive shot

Cricket bowling

Croquet shot

Tennis forehand

Tennis serve

Volleyball smash

[Gupta et al, 2009]

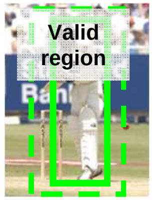
Tasks:

- **Object detection;**
- Pose estimation;
- Activity classification.

Object Detection Results

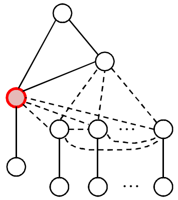
Sliding window

[Andriluka et al, 2009]



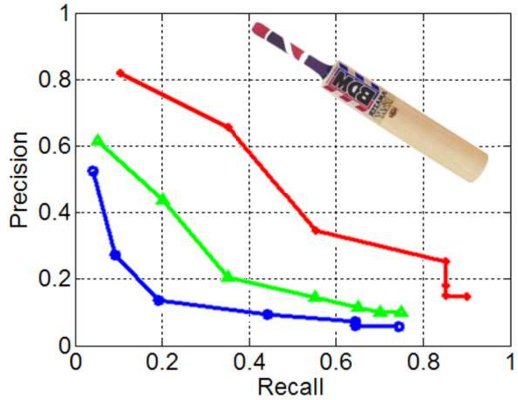
Pedestrian context

[Dalal & Triggs, 2006]

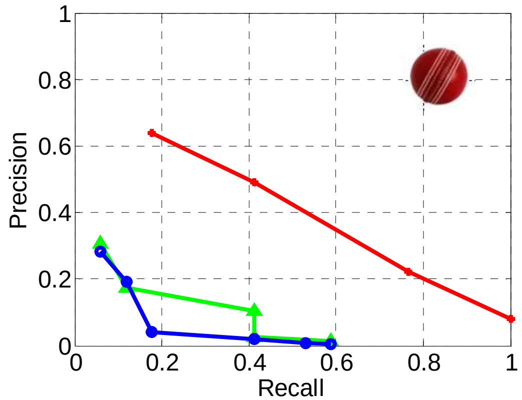


Our Method

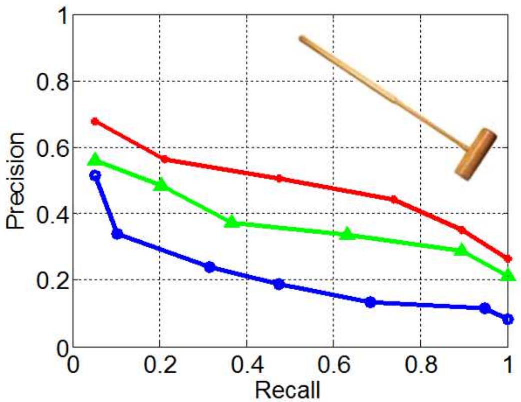
Cricket bat



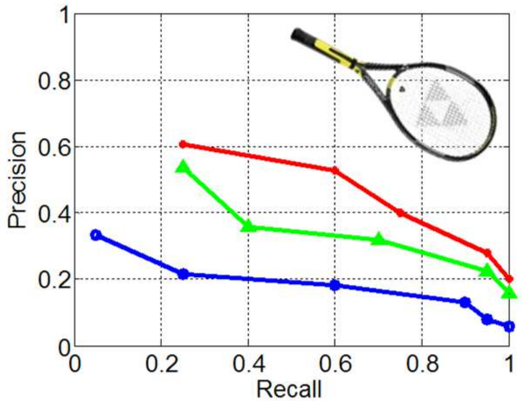
Cricket ball



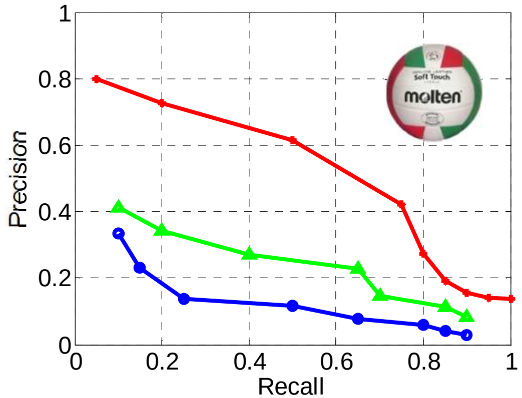
Croquet mallet



Tennis racket



Volleyball



Dataset and Experiment Setup

Sport data set: 6 classes

180 training & 120 testing images

Cricket
defensive shot

Cricket
bowling

Croquet
shot

Tennis
forehand

Tennis
serve

Volleyball
smash

[Gupta et al, 2009]

<u>Tasks:</u>

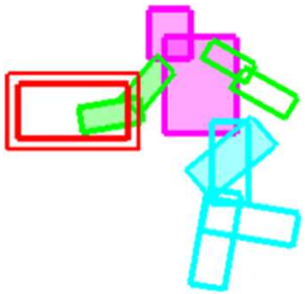
- Object detection;
- **Pose estimation;**
- Activity classification.

Human Pose Estimation Results

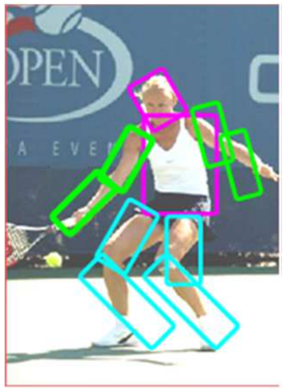
Method	Torso	Upper Leg		Lower Leg		Upper Arm		Lower Arm		Head
Ramanan, 2006	.52	.22	.22	.21	.28	.24	.28	.17	.14	.42
Andriluka et al, 2009	.50	.31	.30	.31	.27	.18	.19	.11	.11	.45
Our full model	**.66**	**.43**	**.39**	**.44**	**.34**	**.44**	**.40**	**.27**	**.29**	**.58**

Human Pose Estimation Results

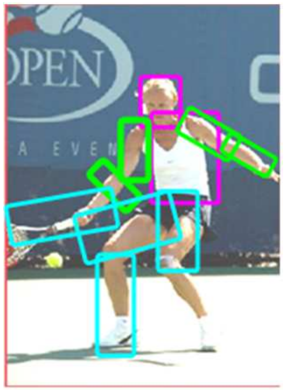
Method	Torso	Upper Leg		Lower Leg		Upper Arm		Lower Arm		Head
Ramanan, 2006	.52	.22	.22	.21	.28	.24	.28	.17	.14	.42
Andriluka et al, 2009	.50	.31	.30	.31	.27	.18	.19	.11	.11	.45
Our full model	**.66**	**.43**	**.39**	**.44**	**.34**	**.44**	**.40**	**.27**	**.29**	**.58**



Tennis serve model

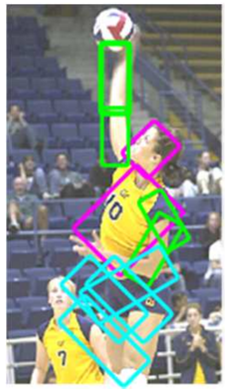


Our estimation result

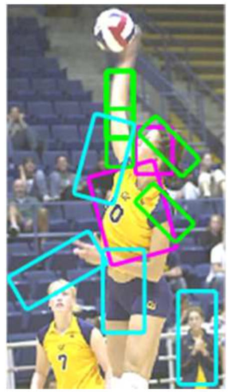


Andriluka et al, 2009

Volleyball smash model



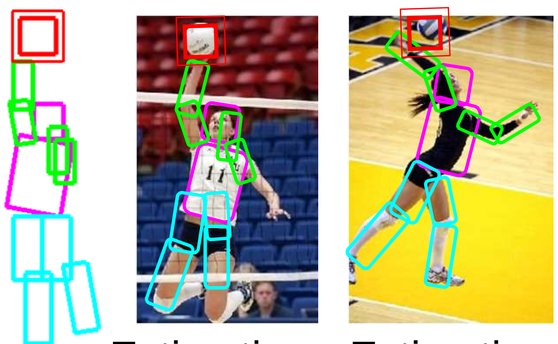
Our estimation result



Andriluka et al, 2009

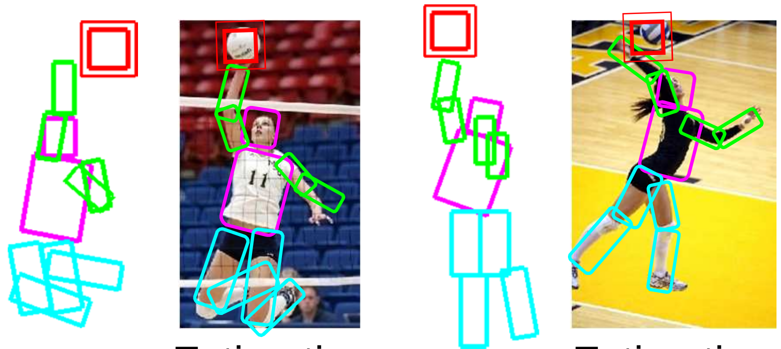
Human Pose Estimation Results

Method	Torso	Upper Leg		Lower Leg		Upper Arm		Lower Arm		Head
Ramanan, 2006	.52	.22	.22	.21	.28	.24	.28	.17	.14	.42
Andriluka et al, 2009	.50	.31	.30	.31	.27	.18	.19	.11	.11	.45
Our full model	**.66**	**.43**	**.39**	**.44**	**.34**	**.44**	**.40**	**.27**	**.29**	**.58**
One pose per class	.63	.40	.36	.41	.31	.38	.35	.21	.23	.52



Estimation result

Estimation result



Estimation result

Estimation result

Dataset and Experiment Setup

Sport data set: 6 classes

180 training & 120 testing images

Cricket
defensive shot

Cricket
bowling

Croquet
shot

Tennis
forehand

Tennis
serve

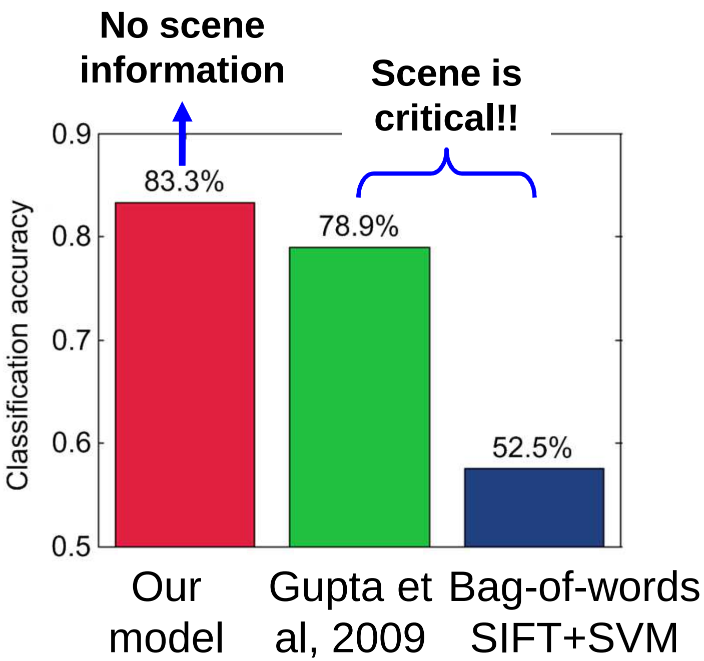
Volleyball
smash

[Gupta et al, 2009]

Tasks:

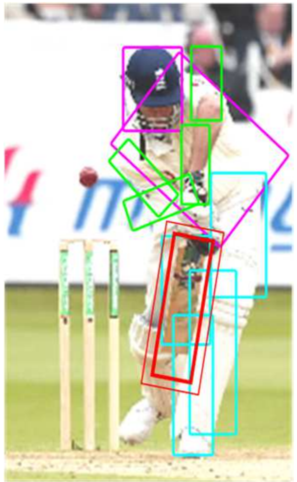
- Object detection;
- Pose estimation;
- **Activity classification.**

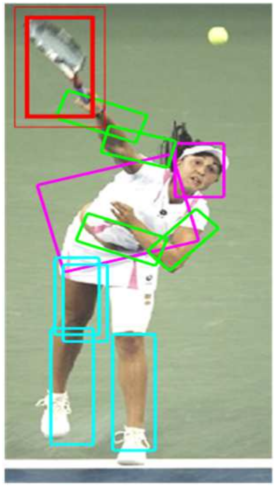
Activity Classification Results

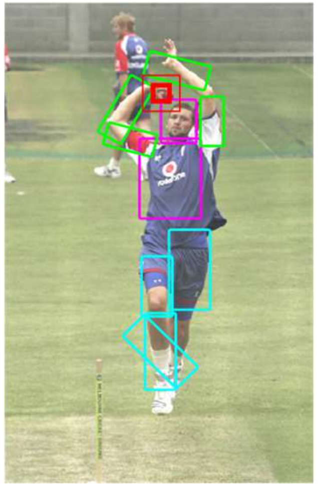


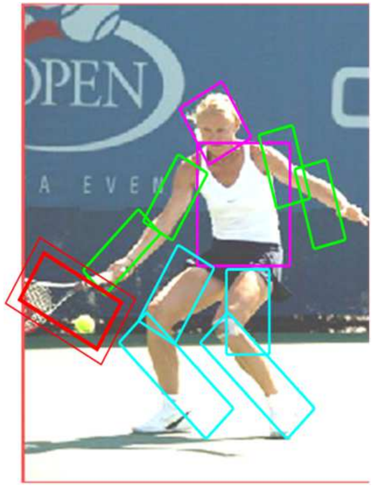
Cricket shot

Tennis forehand

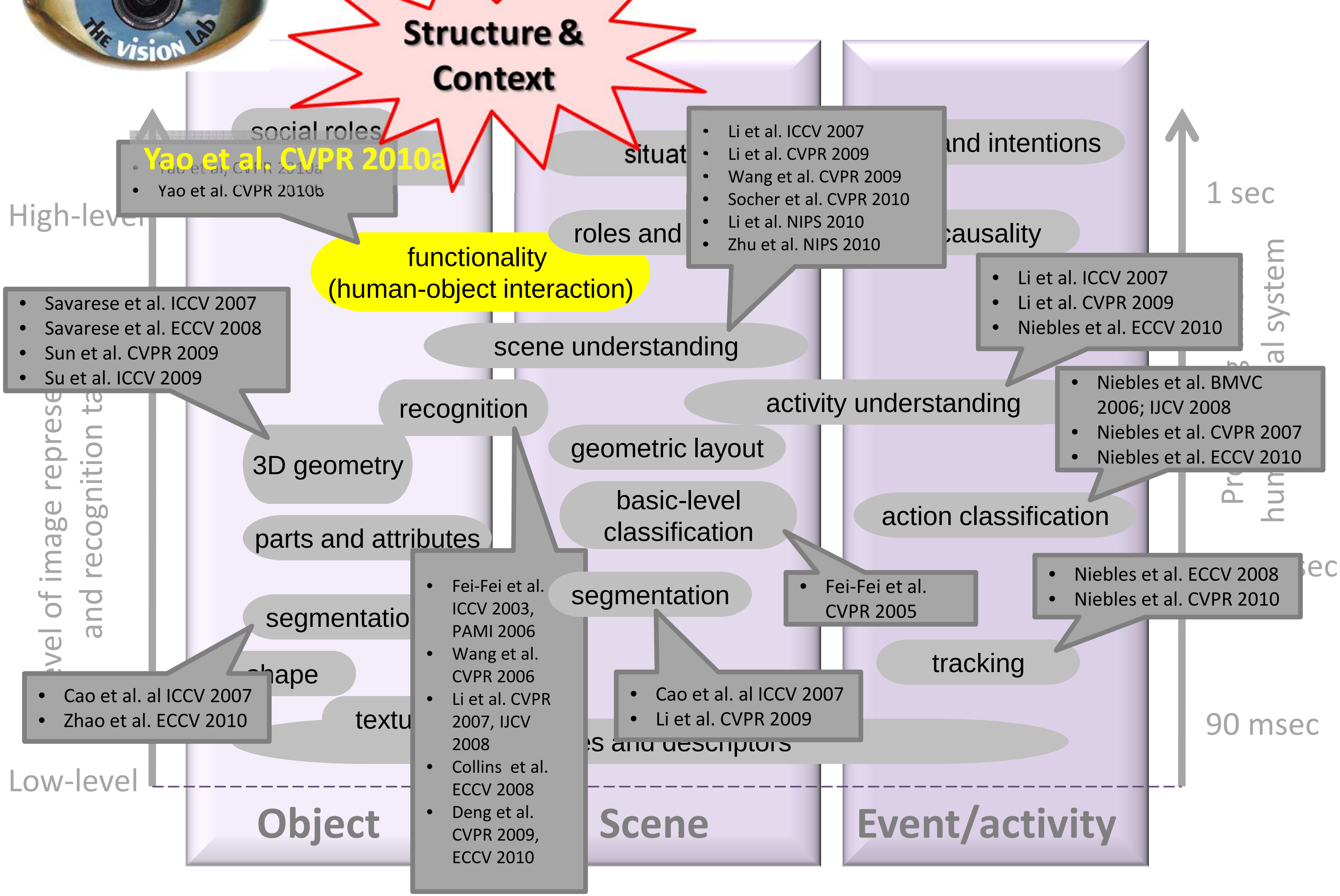




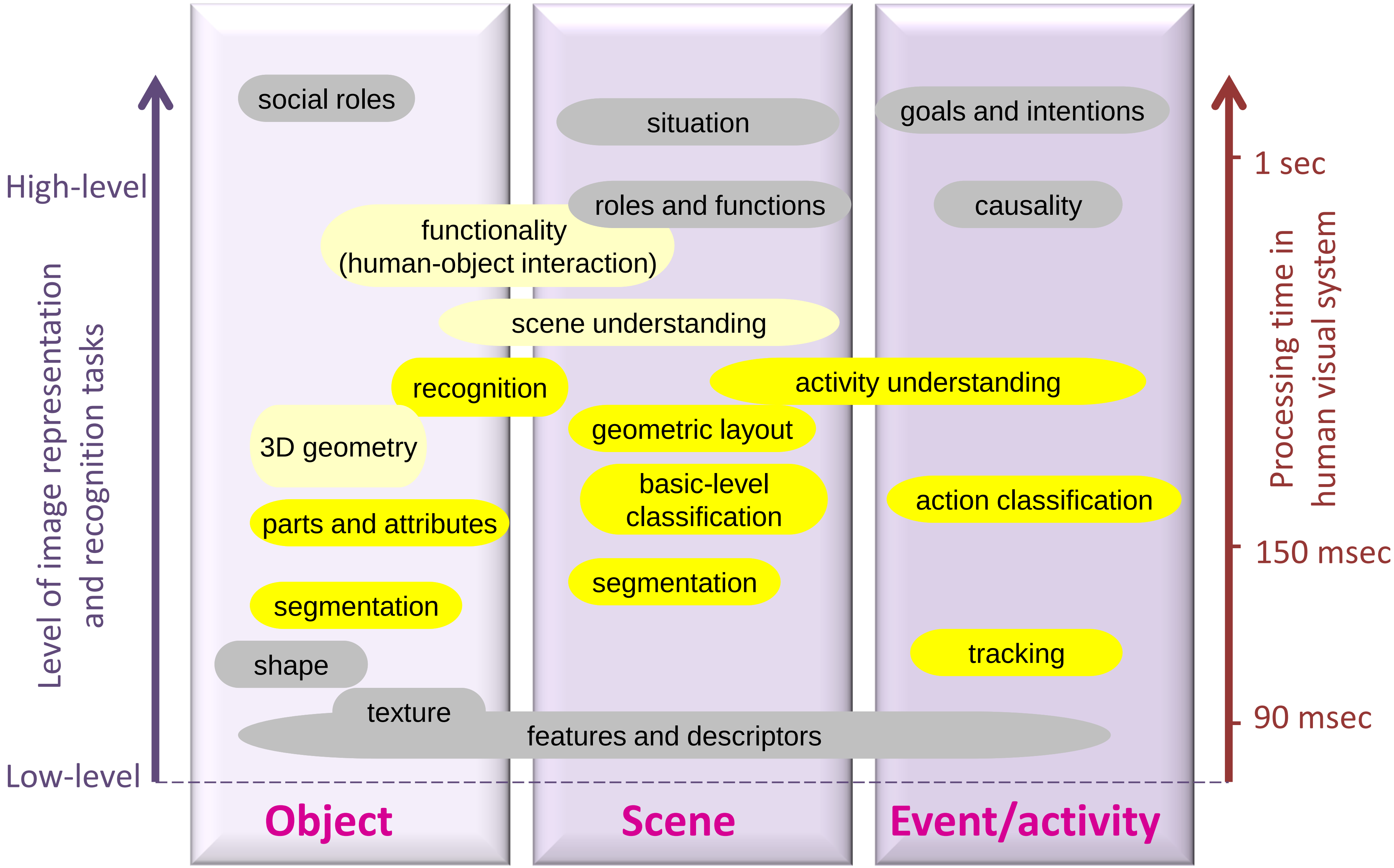




Story telling in images



Story telling in images



Course Project

- No late days
- Tues, Dec 13 (11:59pm): report and codes due
- Wed, Dec 14 (14:00pm): presentation due
  - Use PPT templates
  - 2 pages ONLY
- Thurs, Dec 15 (10am): final presentation
  - Mandatory
- What happens on Dec 15:
  - Every team gets 2min presentation time + 1min Q&A (very strictly enforced)
  - Class vote for “Best Project” and “Most Innovative”

Thank you, class!