max planck institut
informatik

UNIVERSITÄT
DES
SAARLANDES

Towards 3D Visual Scene “Understanding”

Bernt Schiele

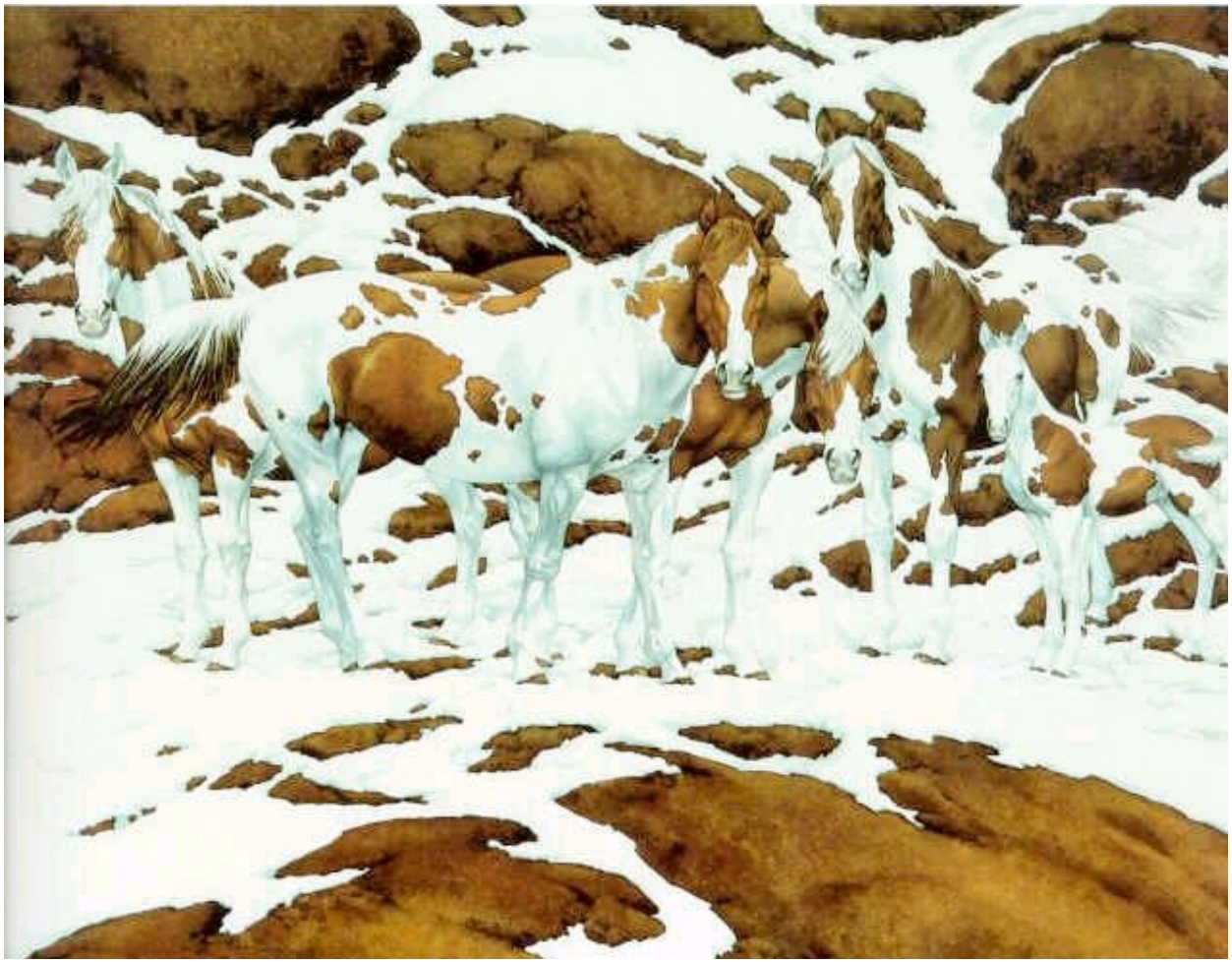
Max Planck Institute for Informatics, Saarbrücken

Saarland University, Saarbrücken

Complexity of 3D Visual Scene Understanding

- components for “understanding”
  - 3D object models
  - 3D scene layout models

Complexity of 3D Visual Scene Understanding



- components for “understanding”
  - 3D object models
  - 3D scene layout models
  - 3D occlusion reasoning

Complexity of 3D Visual Scene Understanding

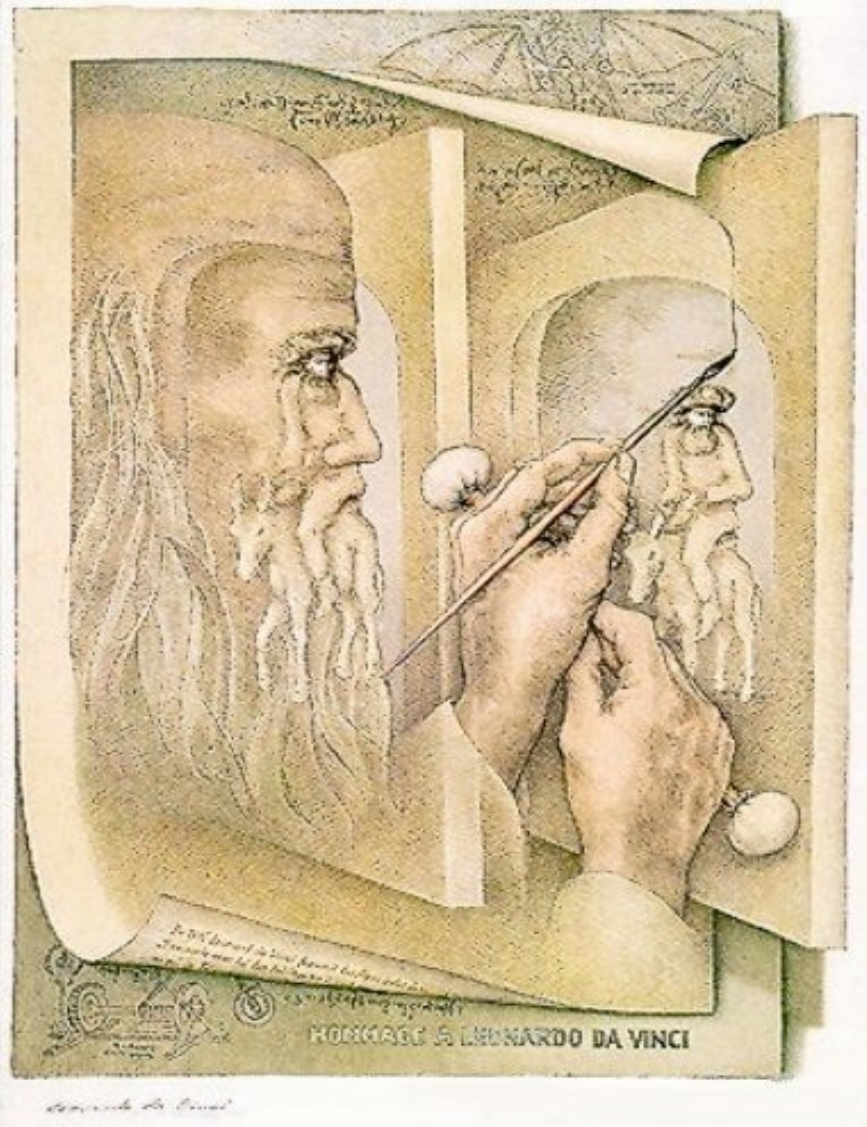
- components for "understanding"
  - 3D object models
  - 3D scene layout models
  - 3D occlusion reasoning
  - 3D motion & behavior models
  - prior information about 3D scenes
  - etc. ...

Advances Towards 3D Scene Understanding

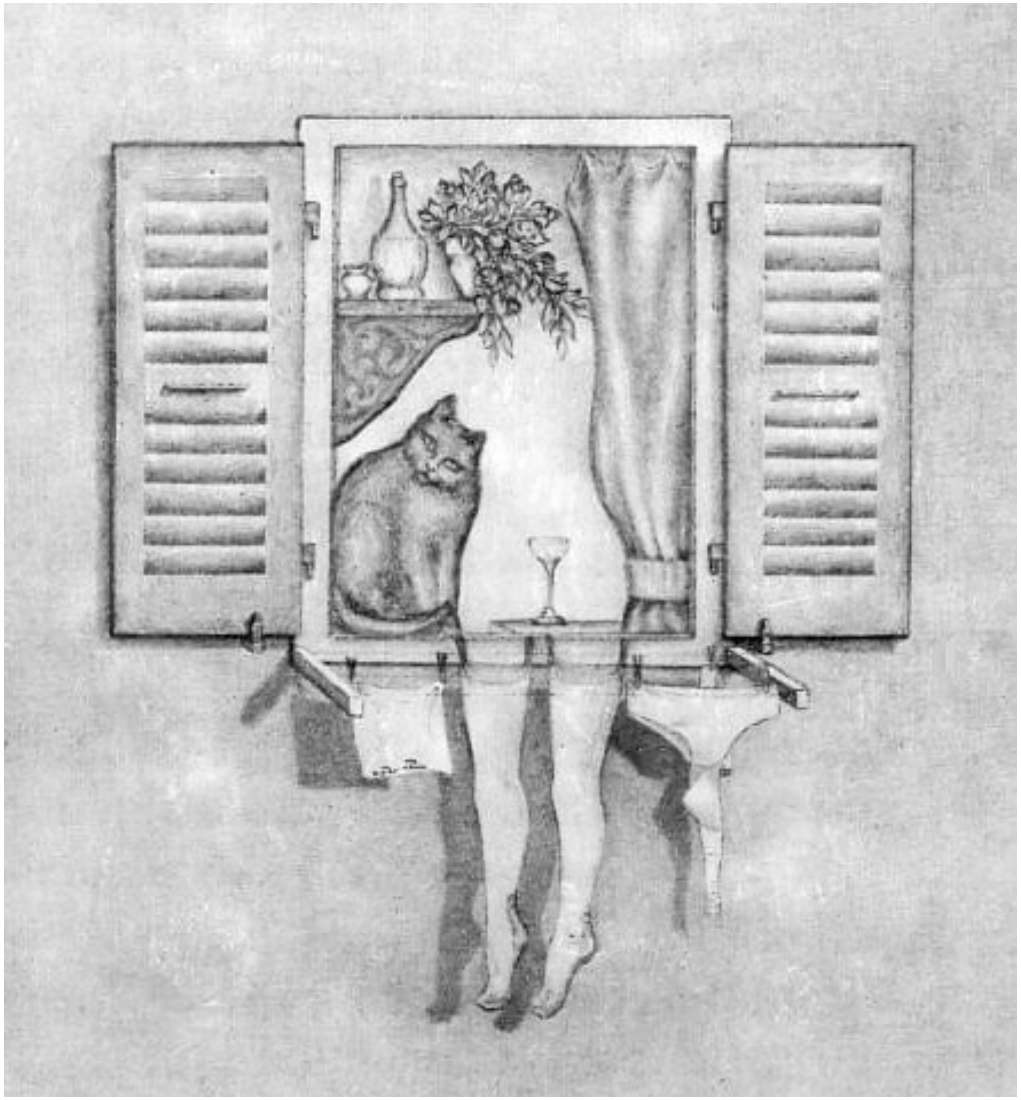
- **Component Research on…**
  - 3D Object Recognition and Segmentation
  - People Detection and Tracking in 3D

- **Beyond Component Research on…**
  (and Towards 3D Scene Understanding)
  - 3D Scene Understanding - traffic scene analysis as a case study
  - Knowledge Harvesting from Language
  - Video and Scene Descriptions

Complexity of Recognition

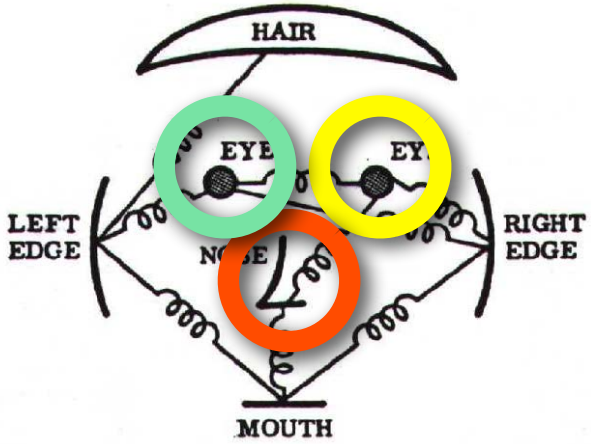


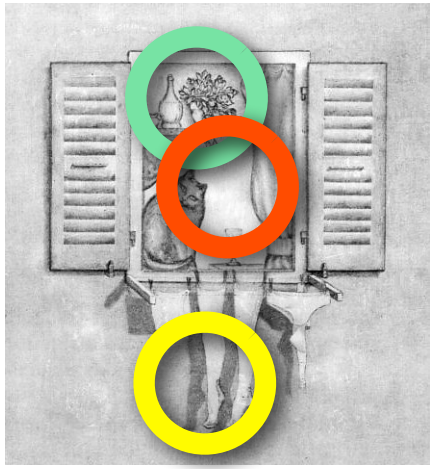
Complexity of Recognition



Class of Object Models for Recognition
Part-Based Models / Pictorial Structures

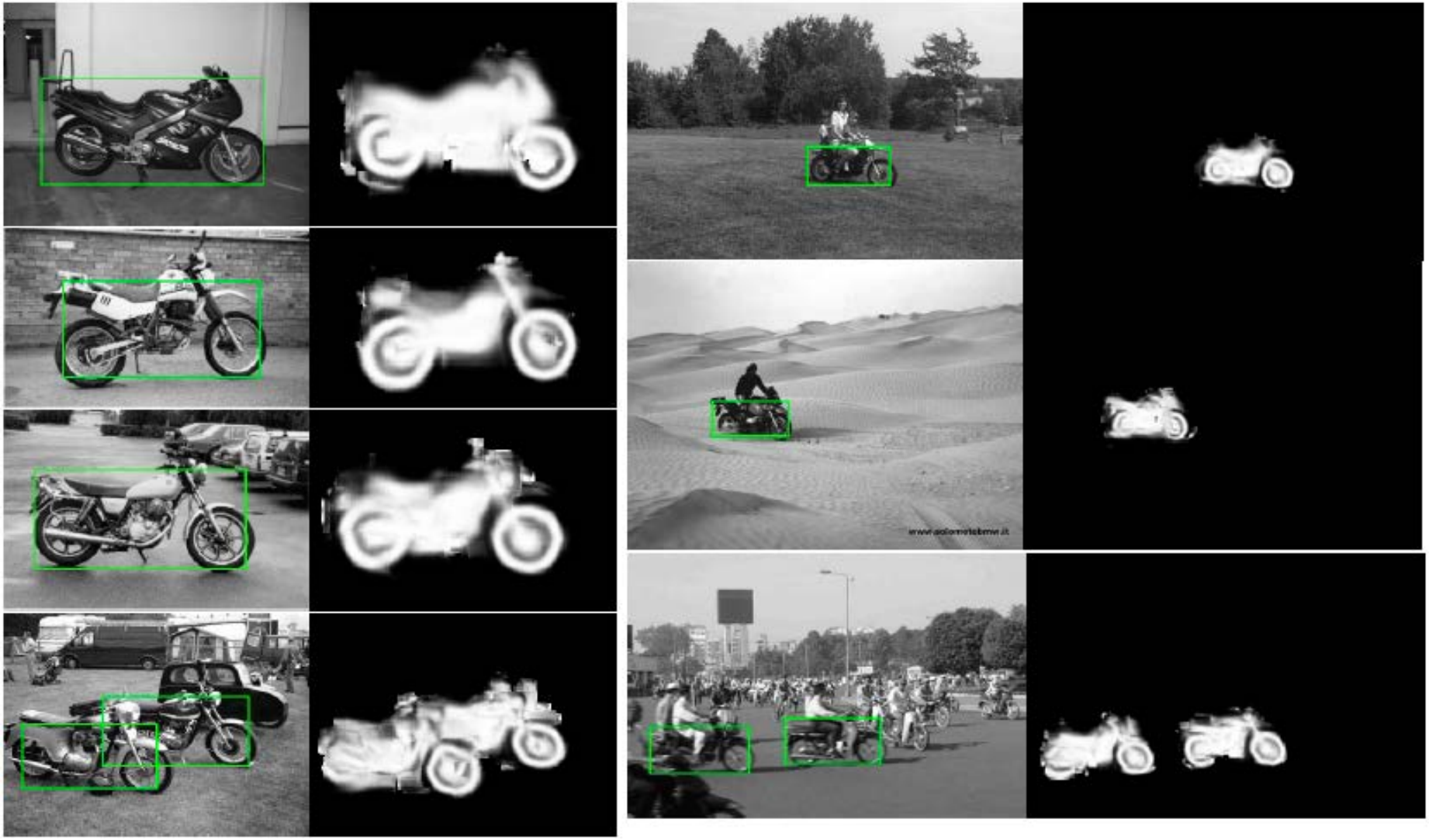
- Pictorial Structures [Fischler & Elschlager 1973]
  - Model has two components
    - parts (2D image fragments)
    - structure (configuration of parts)





[Leibe,Schiele@BMVC'03]
[Leibe,Leonardis,Schiele@IJCV'08]

Implicit Shape Model

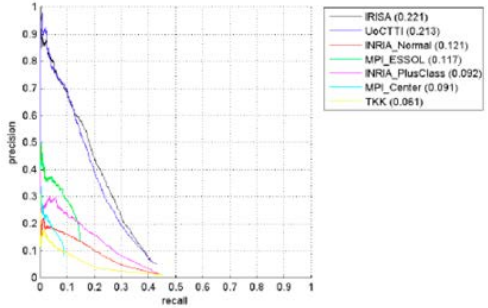


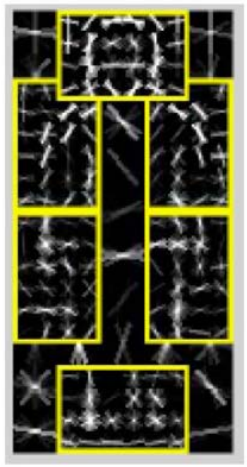
[Felzenszwalb,Girshick,McAllester,Ramanan@PAMI'10]

Deformable Parts Model (DPM)

PASCAL VOC 2007 Person Detection

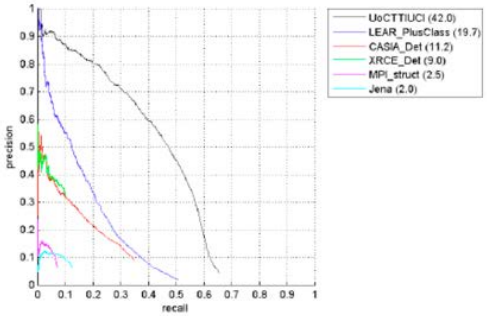
- Pictorial structure model
  - 45% precision at 20% recall

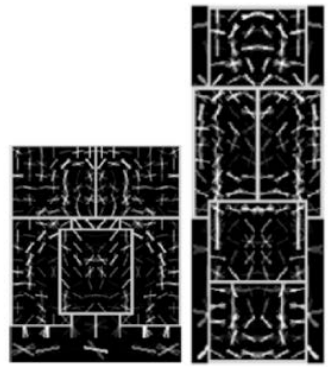




PASCAL VOC 2008 Person Detection

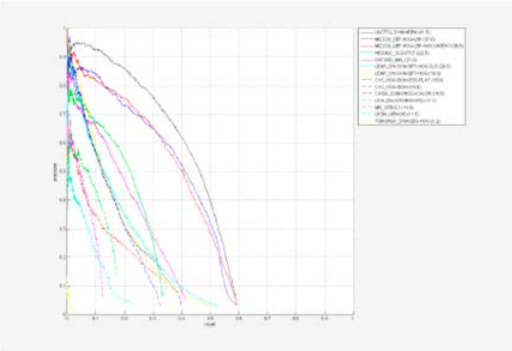
- Disjunction of two pictorial structures
  - 80% precision at 20% recall

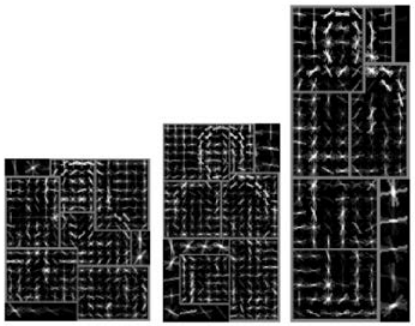




PASCAL VOC 2009 Person Detection

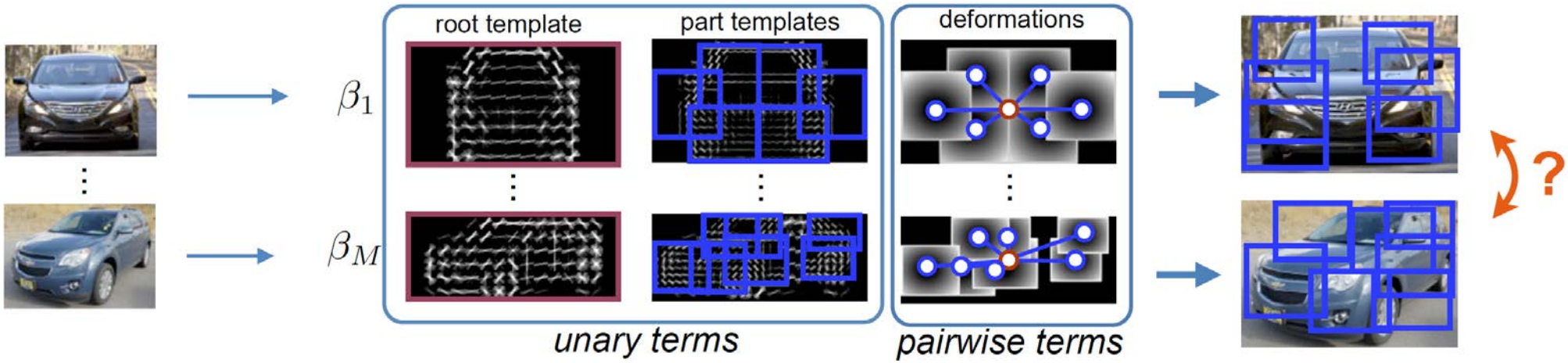
- Disjunction of three pictorial structures
  - 85% precision at 20% recall





Towards a “True” 3D Object Model

- Standard Deformable Parts Model:



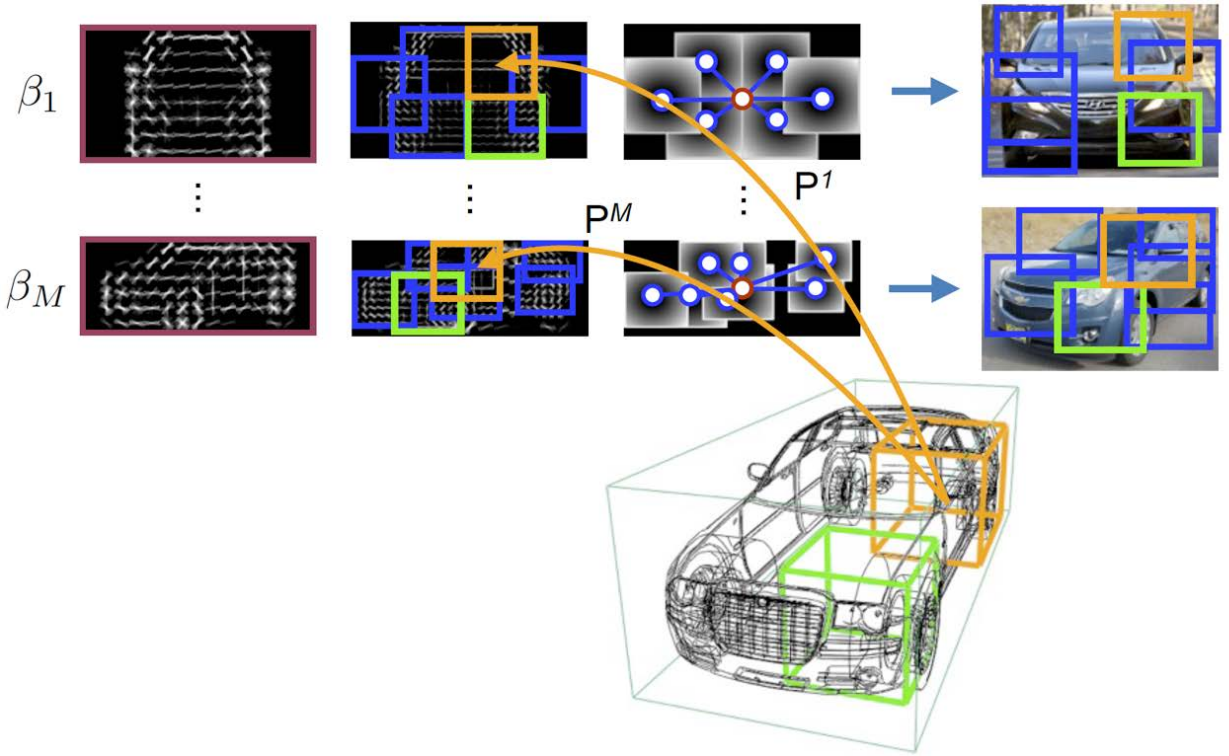
[Felzenszwalb@pami10]

- parts and deformations parametrized in 2D
- each viewpoint modeled independently

[Pepik,Stark,Gehler,Schiele@PAMI'15]

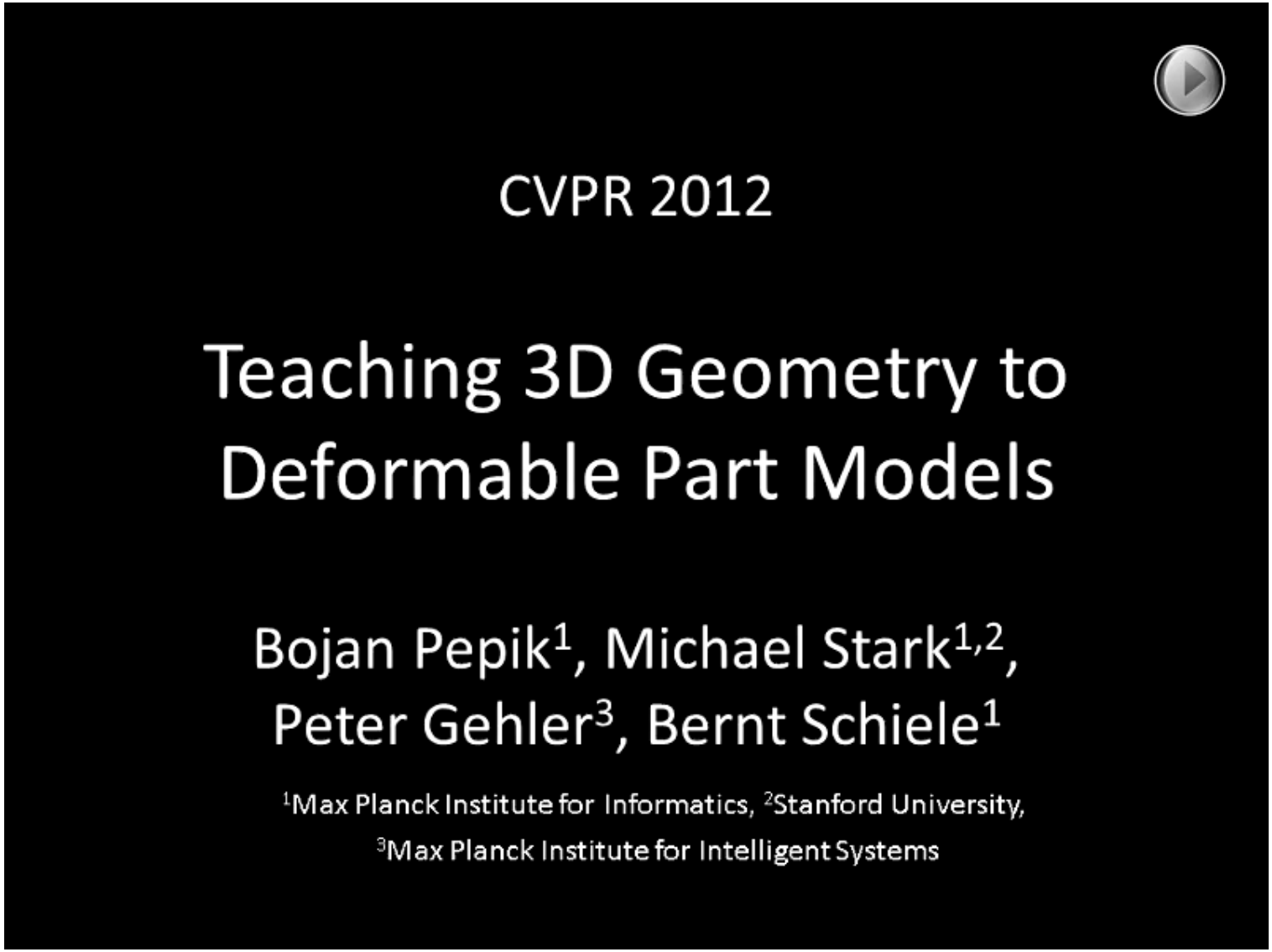
"True" 3D Object Model

- 3D^{2}PM (3D - DPM):
  - 3D parts and 3D deformations are parametrized in 3D object coordinates
  - therefore: parts are linked across different viewpoints !



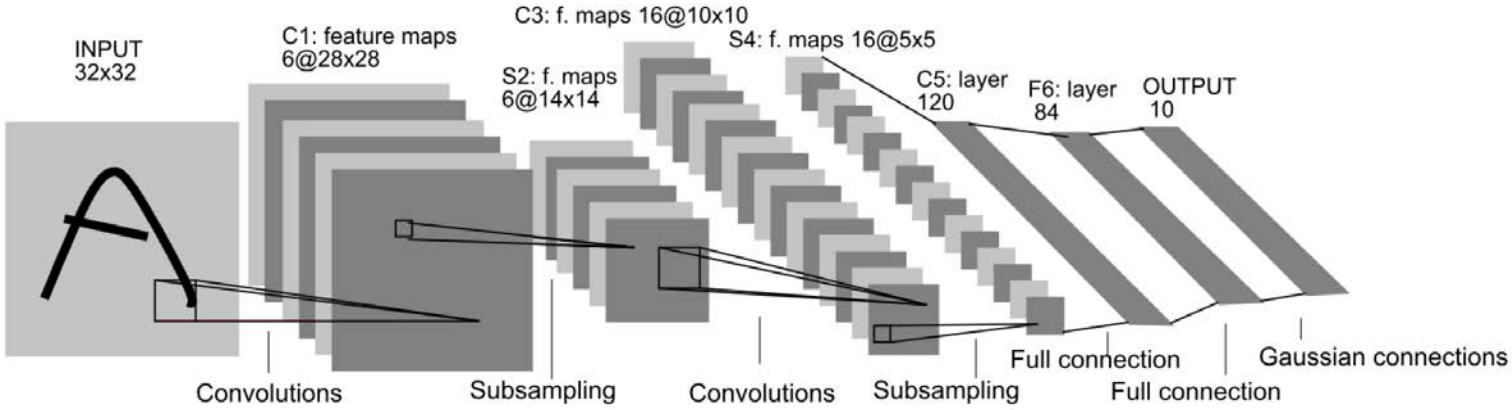
[Pepik,Stark,Gehler,Schiele@PAMI'15]

Qualitative Results

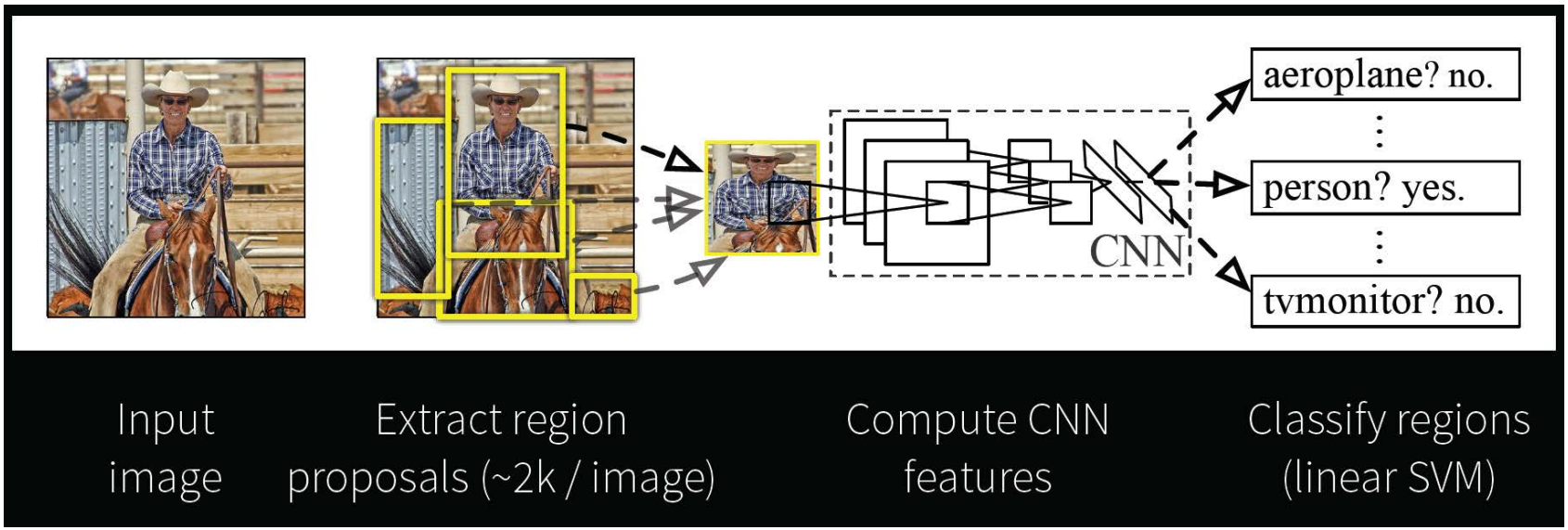


Deep Learning…
e.g. Convolutional Networks (CNN)

- LeNet5 - Yann LeCun 1998

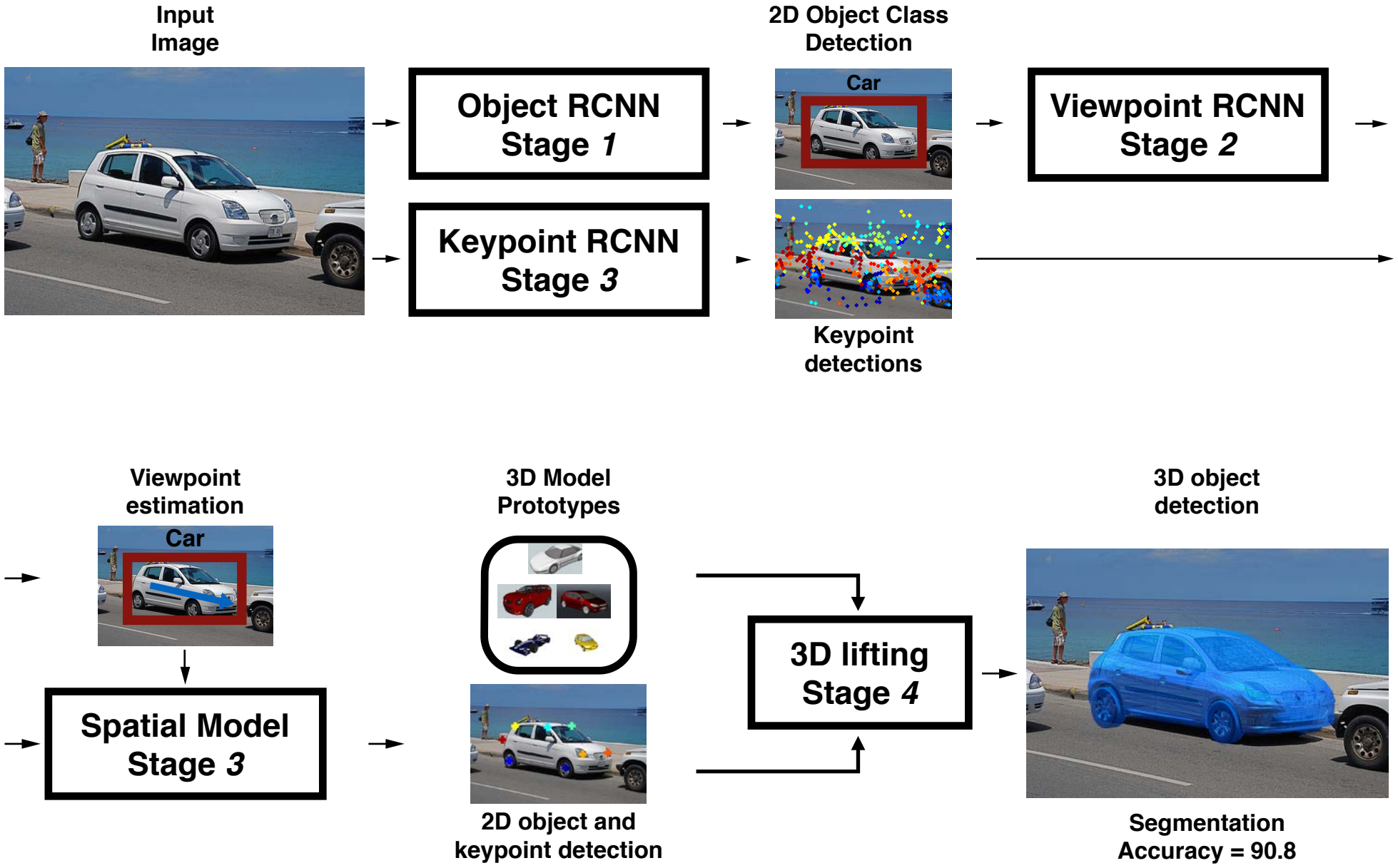


- Regions with CNN (R-CNN) - 2014



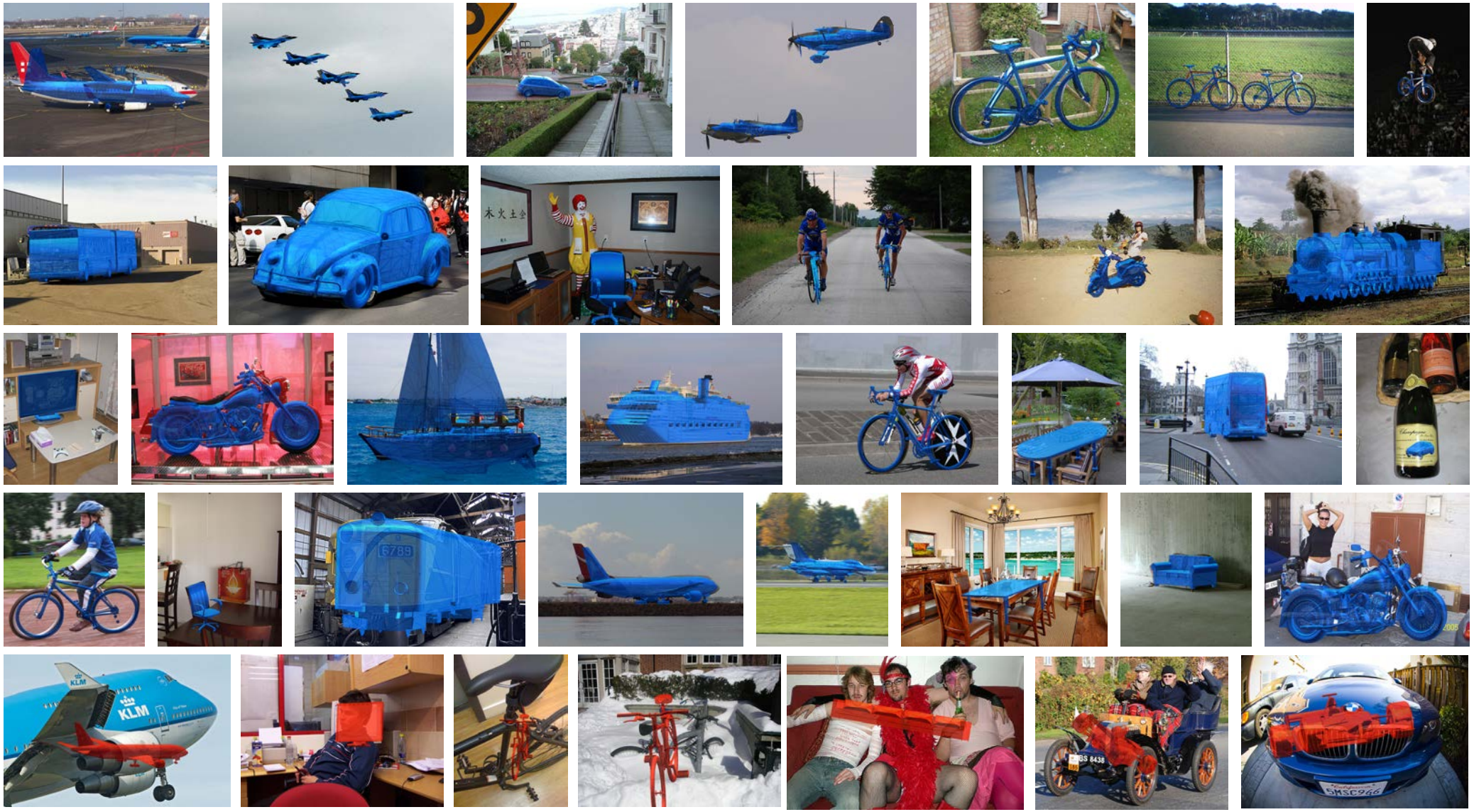
3D Object Class Detection and Segmentation based on CNNs

[Pepik,Stark,Gehler,Ritschel,Schiele@cvpr15-workshop]



[Pepik,Stark,Gehler,Ritschel,Schiele@cvpr15-workshop]

Qualitative results



Advances Towards 3D Scene Understanding

- **Component Research on…**
  - 3D Object Recognition and Segmentation
  - **People Detection and Tracking in 3D**

- **Beyond Component Research…**
  (and Towards 3D Scene Understanding)
  - 3D Scene Understanding - traffic scene analysis as a case study
  - Knowledge Harvesting from Language
  - Video and Scene Descriptions

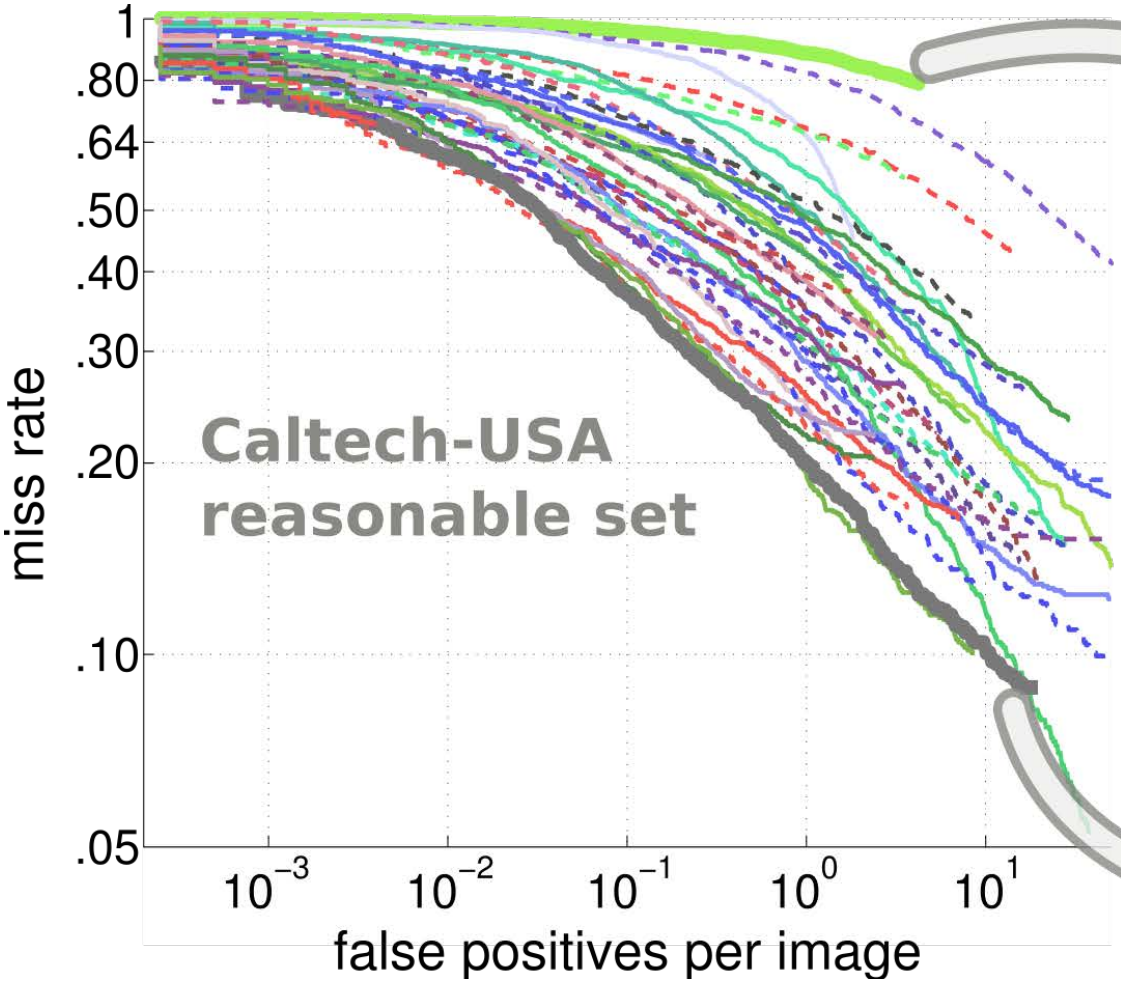
[Dollar,Wojek,Perona,Schiele@PAMI'12]

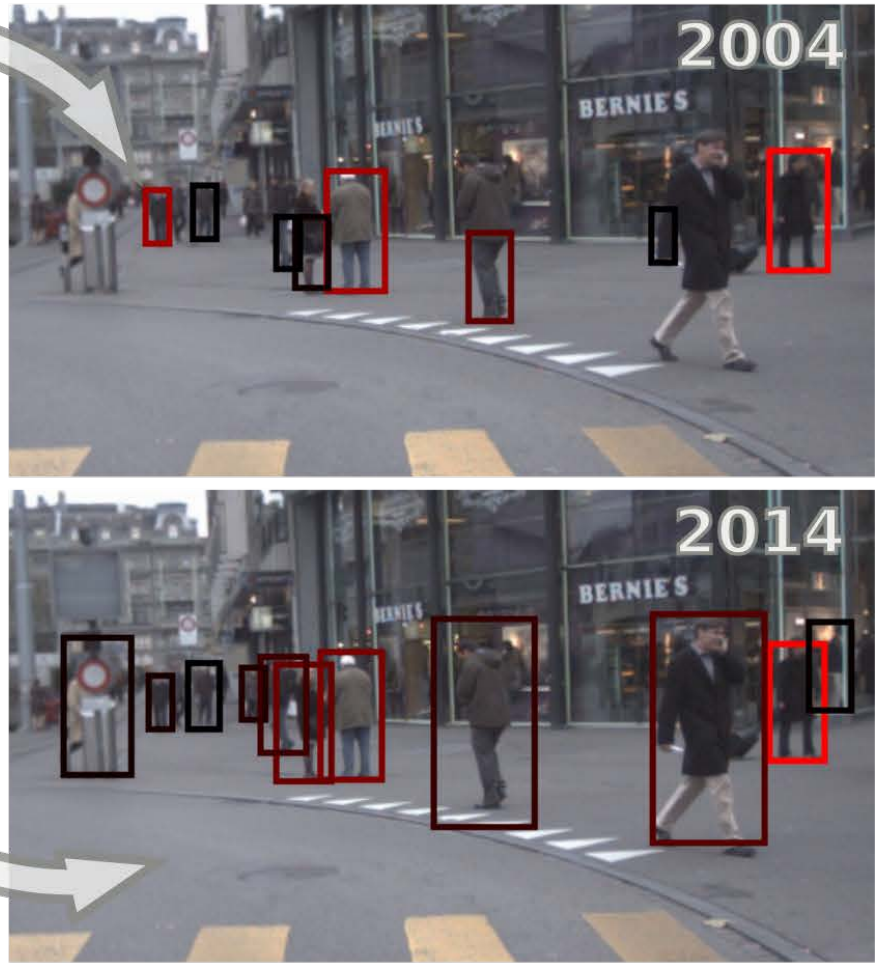
Caltech Pedestrian Benchmark

- Features of the Pedestrian Dataset:
  - 11h of 'normal' driving in urban environment (greater LA area)
  - annotation:
    - 250,000 frames (~137 min) annotated with 350,000 labeled bounding boxes of 2,300 unique pedestrians
    - occlusion annotation: 2 bounding boxes for entire pedestrian & visible region
    - difference between 'single person' and 'groups of people'

Caltech-USA currently the most active dataset

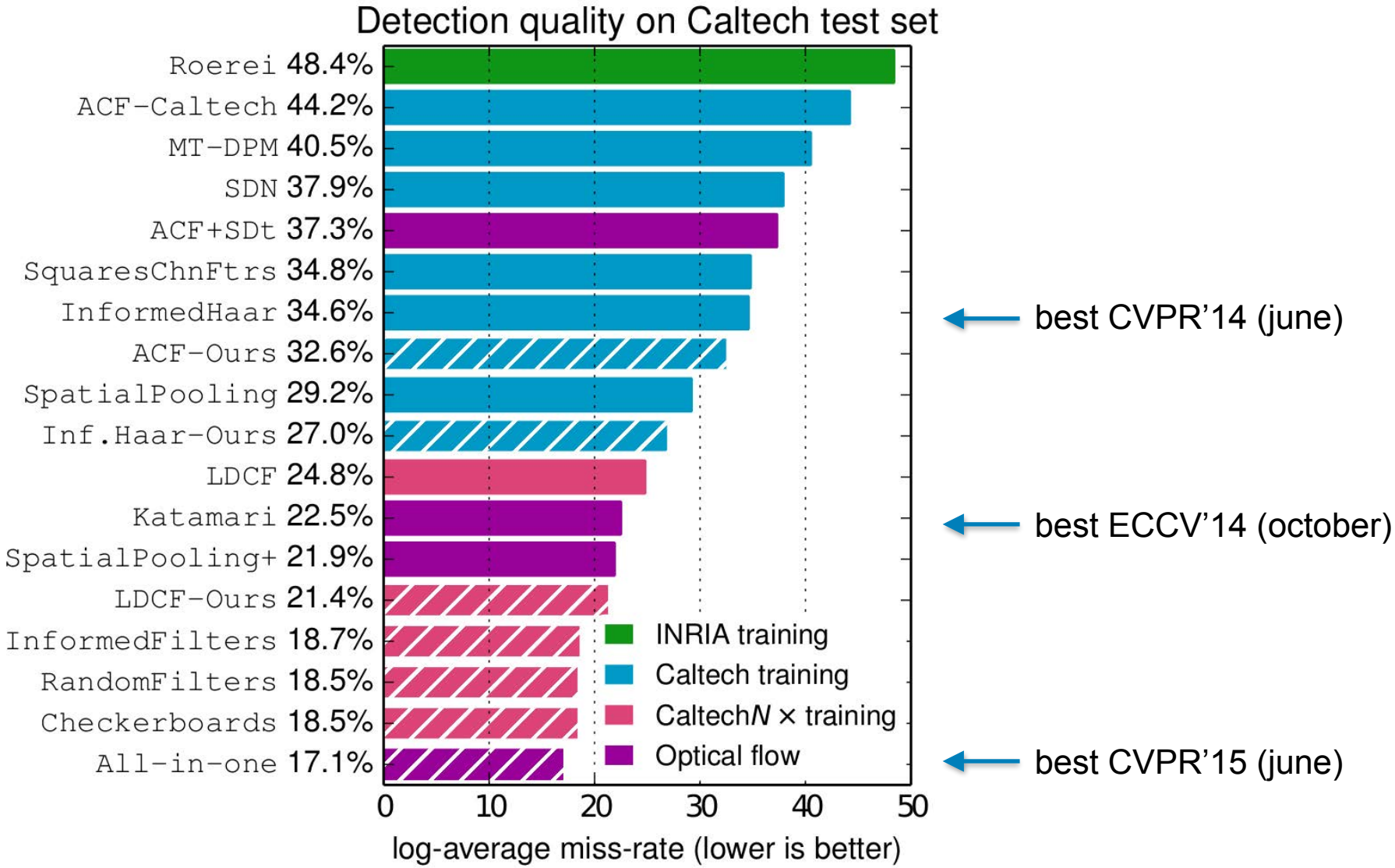
Great Progress in People / Pedestrian Detection During last 10 Years





[Zhang,Benenson,Schiele@CVPR'15]

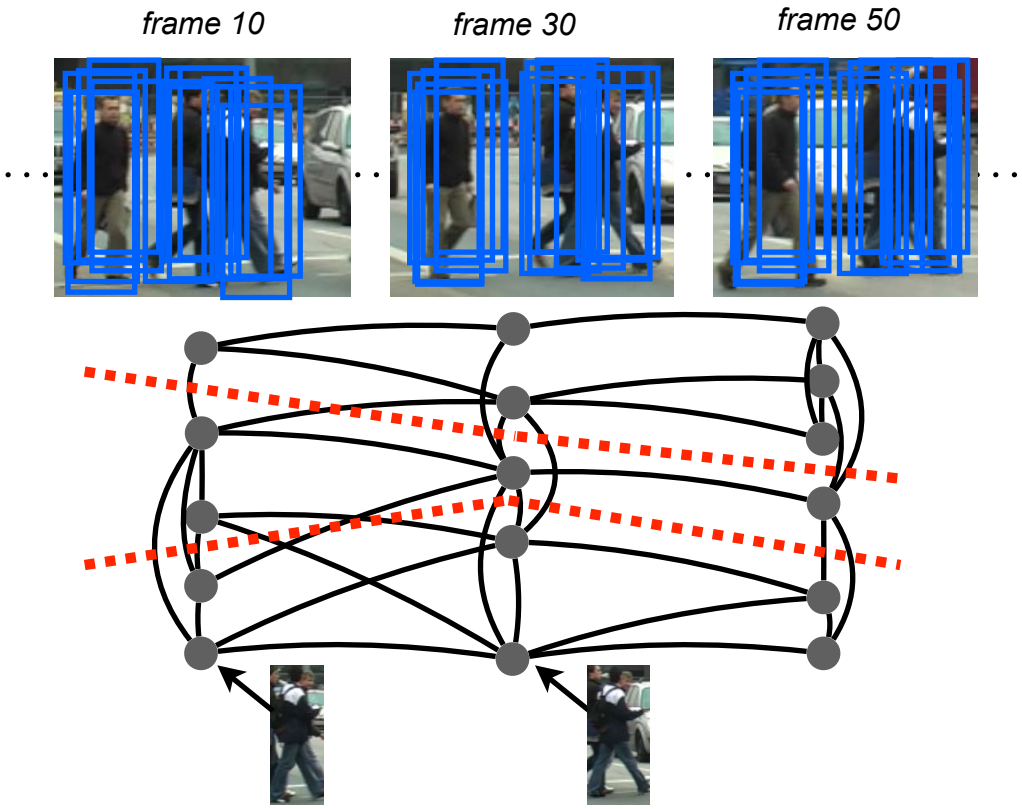
Performance is Still Improving



[Tang,Andriluka,Andres,Schiele@CVPR'15]

Multipeople Tracking using Multi Cut Subgraph Partitioning

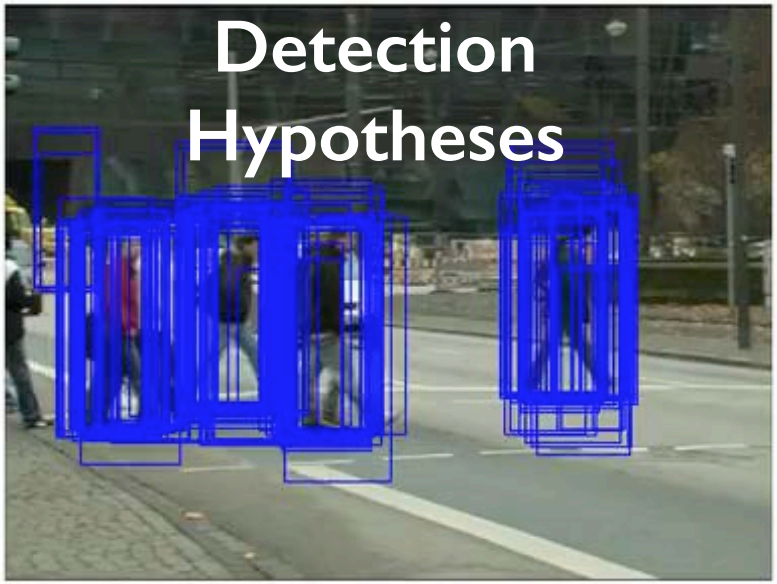
- Subgraph decomposition for multi-object tracking

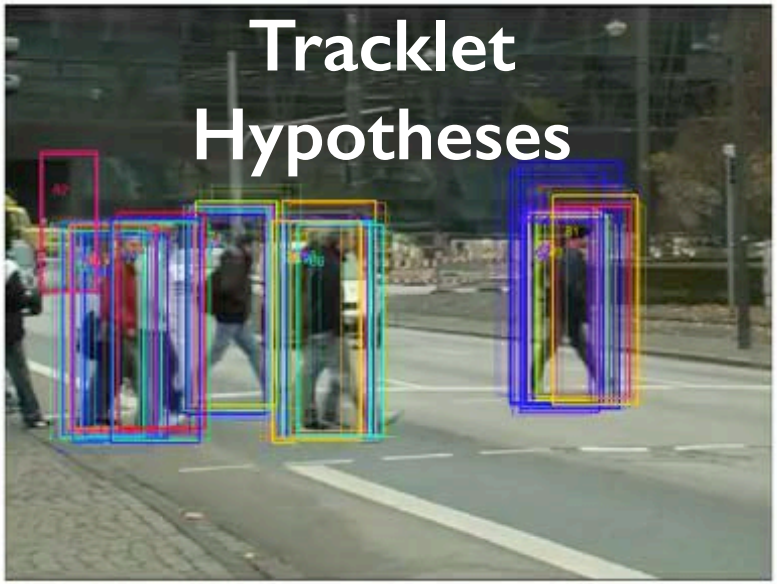


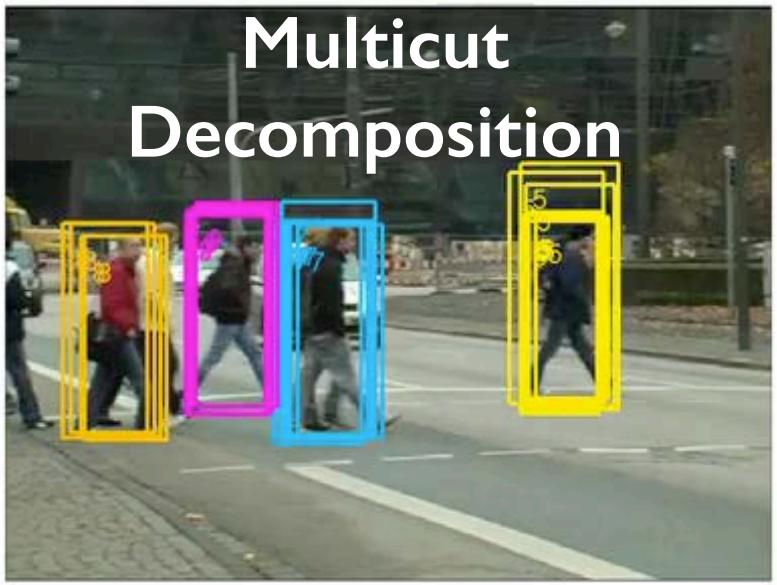
- Desired property of "tracking by graph decomposition"
  - joint spatial-temporal association
  - resulting in robust tracking results

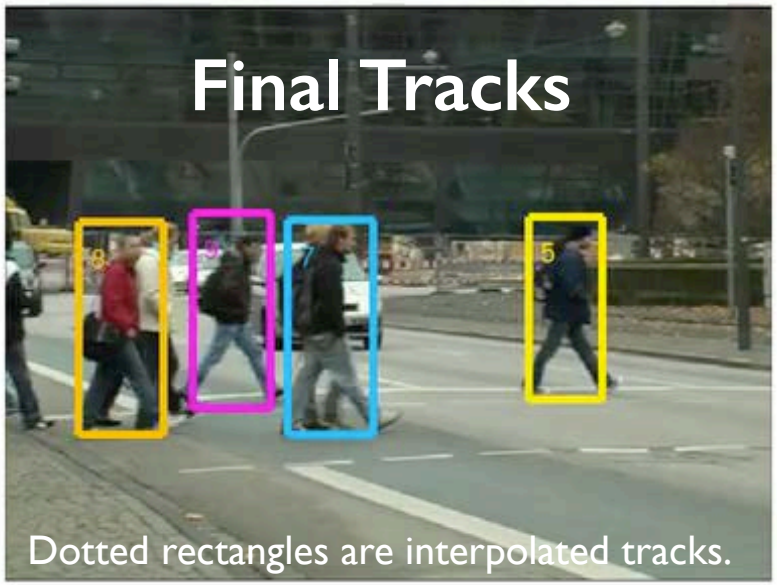
Multipeople Tracking Qualitative Results

[Tang,Andriluka,Andres,Schiele@CVPR'15]









Body Pose Estimation: Pictorial Structures (PS) Model

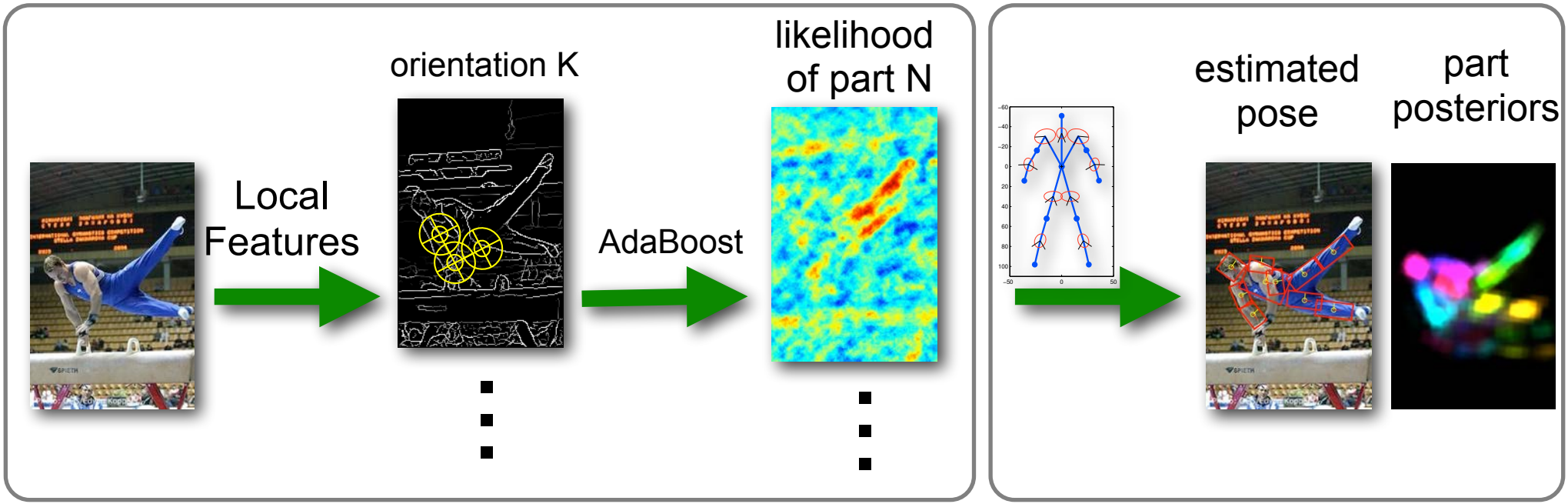
[Andriluka,Roth,Schiele@IJCV'12]
[Felzenszwalb,Huttenlocher@IJCV'05]

Posterior over body poses

$$p(L|D) \propto p(D|L)p(L)$$

likelihood of part observations (appearance model)

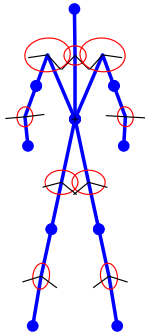
prior on body poses

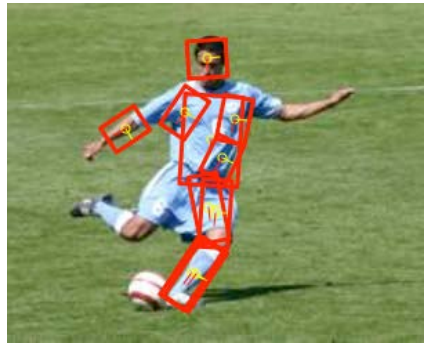


[Pishculin,Andriluka,Gehler,Schiele@CVPR'13]

Motivation

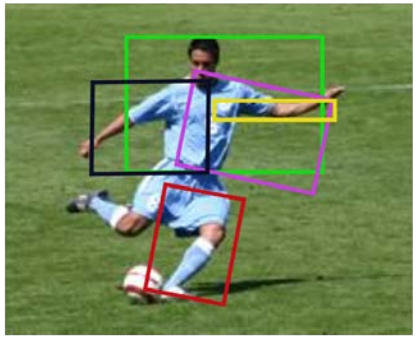
Classic: tree-structured models

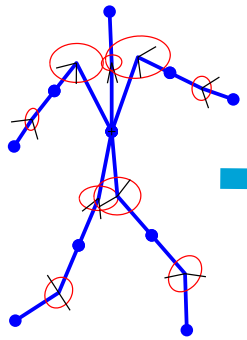


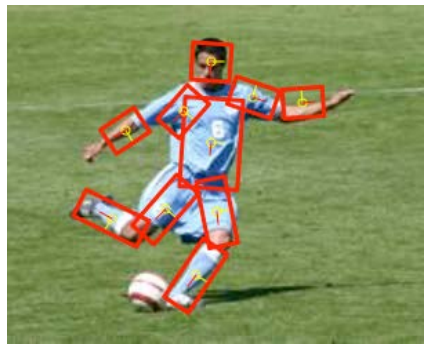


- ***generic*** kinematic tree
- capture ***adjacent*** part dependencies ***only***

[cvpr13]





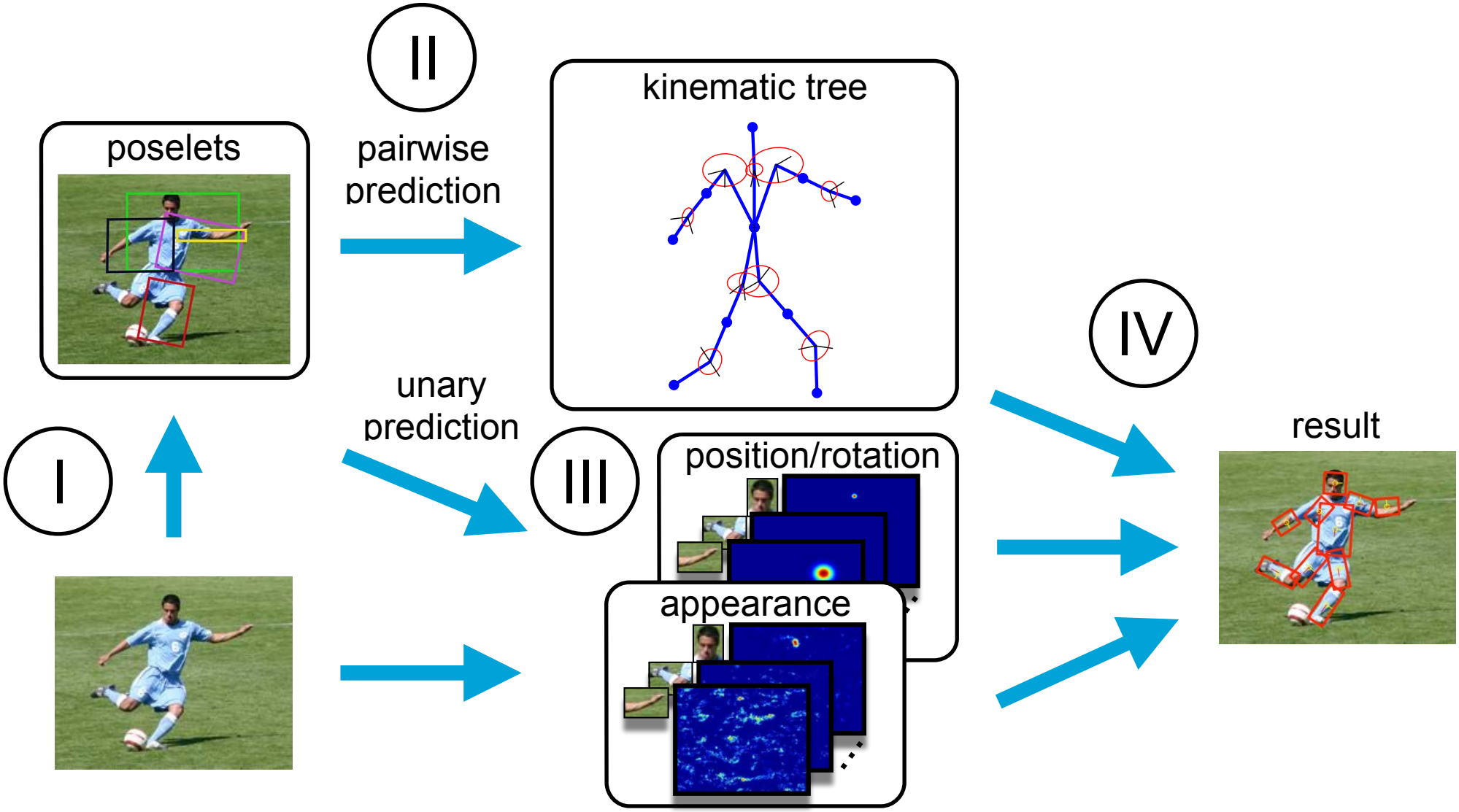


✓ ***poselet conditioned*** kinematic tree

✓ ***poselets*** capture ***non-adjacent*** part dependencies

[Pishculin,Andriluka,Gehler,Schiele@CVPR'13]

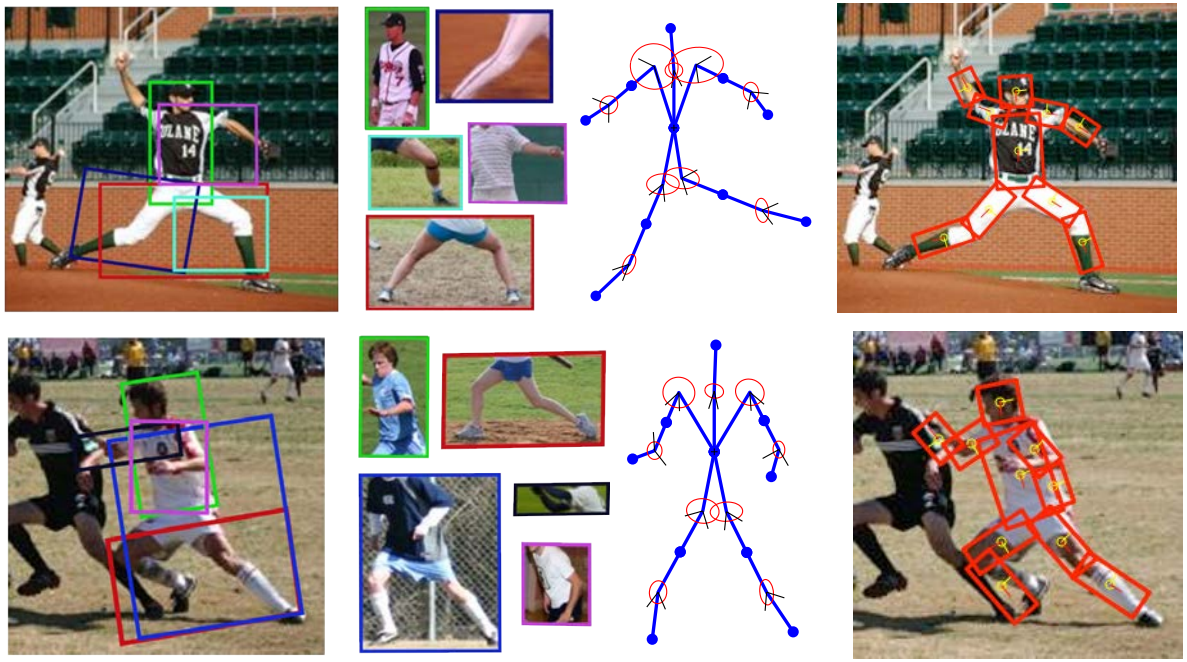
Poselet Conditioned Pictorial Structures Model



[Pishculin,Andriluka,Gehler,Schiele@CVPR'13]

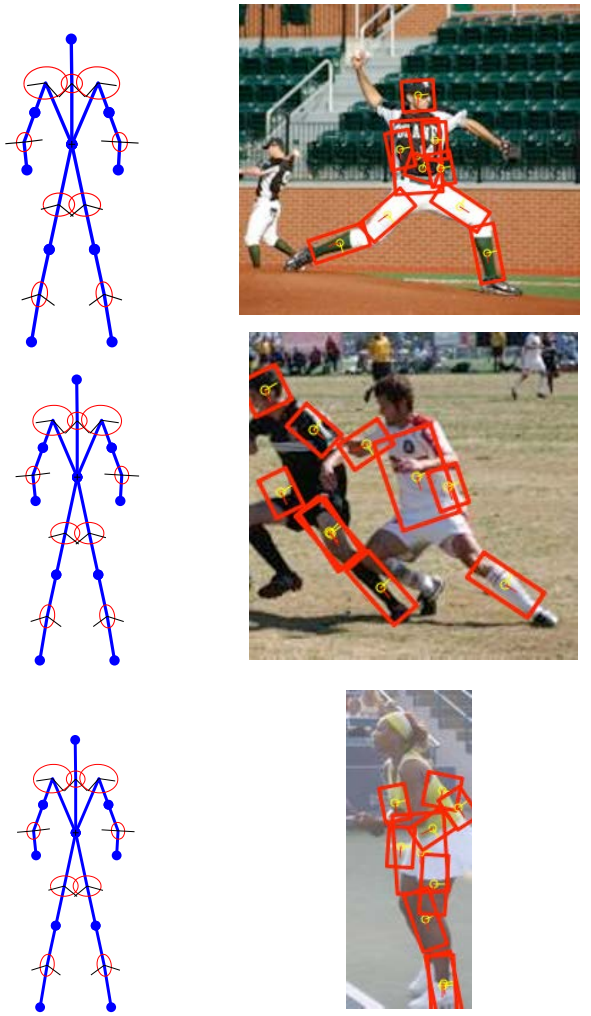
Poselet Conditioned Model - Qualitative Results

Poselet Conditioned PS



Top poselet detections | Cluster medoids | Prediction | Result using prediction

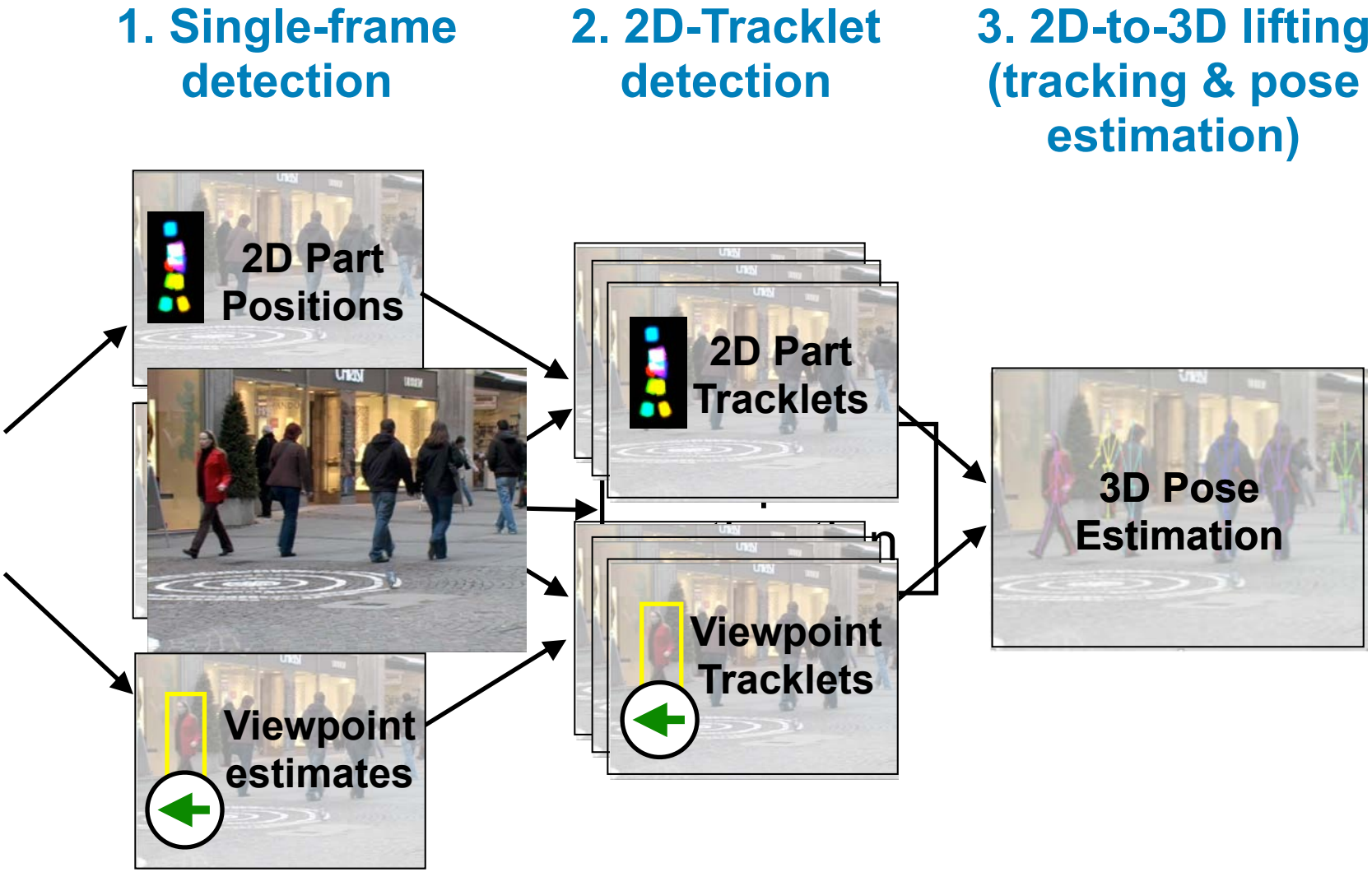
Classic PS



Generic tree | Result using generic tree

[Andriluka,Roth,Schiele@CVPR'10]

3D Pose Estimation and Tracking

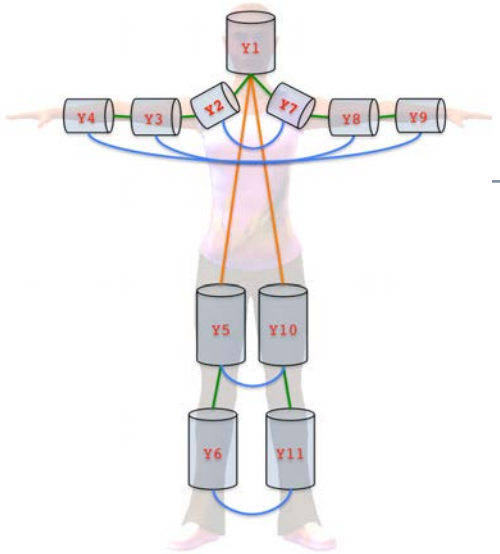


[Andriluka,Roth,Schiele@CVPR'10]

Integrated Detection & Tracking in 3D

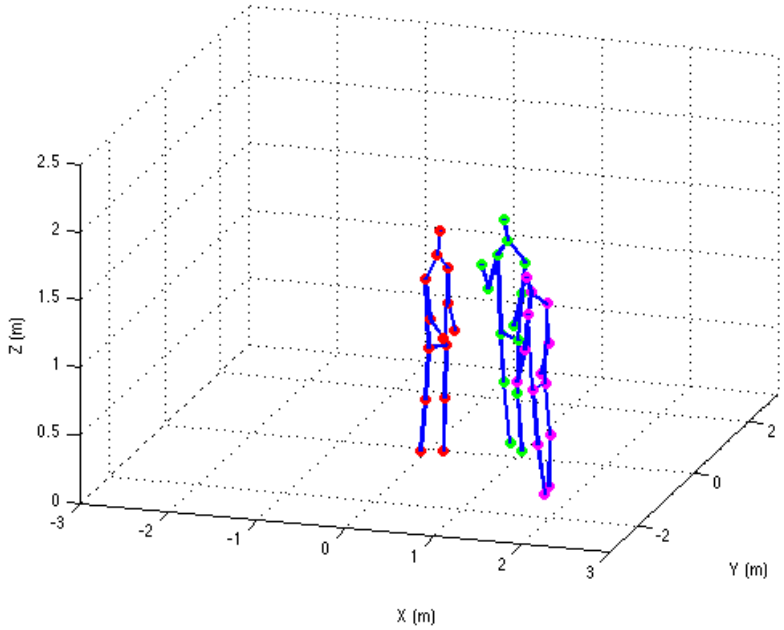
- example of multi-viewpoint detection & tracking
  - note: monocular camera, no static camera assumption is used

3D Pictorial Structures Model for 3D Human Pose Estimation



- Multi-View – Multiple Human

  [Belagiannis,Amin,Andriluka,Schiele,Navab,Ilic@CVPR'14]



Interim Discussion: Computer Vision Components…

- Significant progress in the last 10+ years
  - has also let to increased industry interest :)
- Machine Learning
  - played a prominent role in the last decade (deep neural networks particularly in the last few years)
  - will remain instrumental
- 3D information is important and becomes more accessible by
  - 3D modeling & inference
  - 3D sensors

Advances Towards 3D Scene Understanding

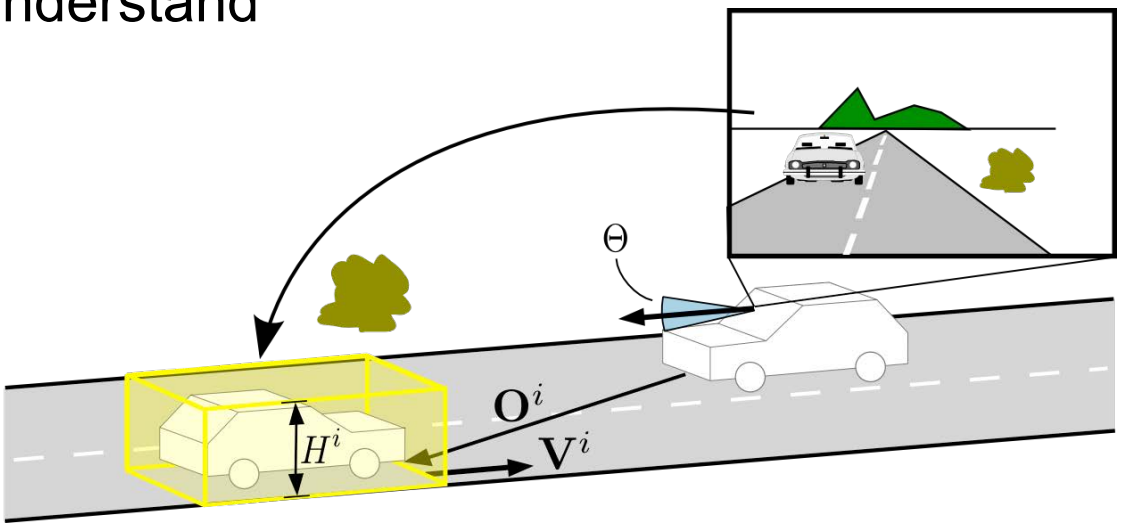
- Component Research on…
  - 3D Object Recognition and Segmentation
  - People Detection and Tracking in 3D

- **Beyond Component Research on… (and Towards 3D Scene Understanding)**
  - **3D Scene Understanding - traffic scene analysis as a case study**
  - Knowledge Harvesting from Language
  - Video and Scene Descriptions

[Wojek,Roth,Schindler,Schiele@PAMI'13]

3D Scene Understanding

- 3D scene analysis for mobile platforms (i.e. robots, cars)
  - mobile observer aims to "understand"
    its 3D mobile environment
    i.e. traffic, people, etc



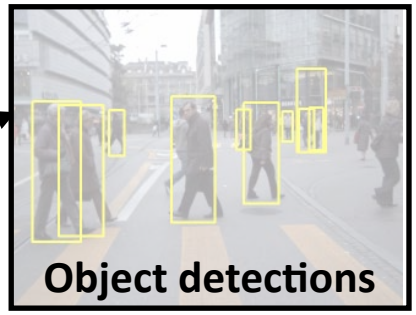
- Application scenarios
  - Traffic safety and driver assistance
  - Autonomous vehicles
  - Robotics

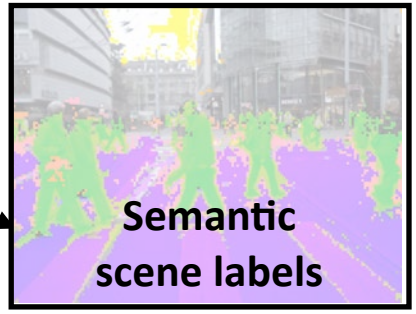
[Wojek,Roth,Schindler,Schiele@PAMI'13]

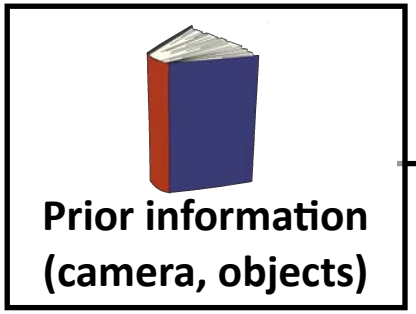
A state-of-the-art Approach (monocular camera)

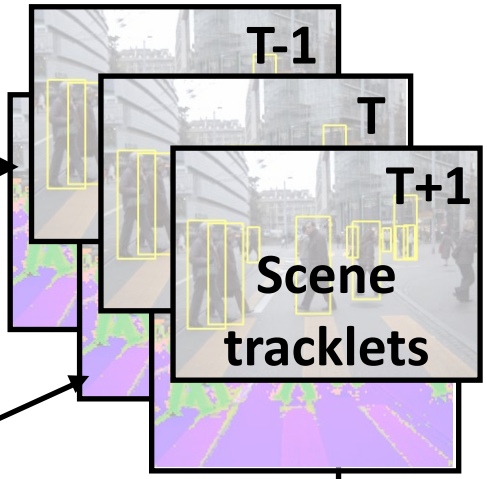
Image sequence

Bayesian 3D scene model



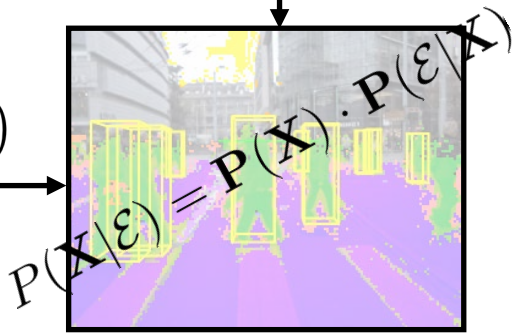






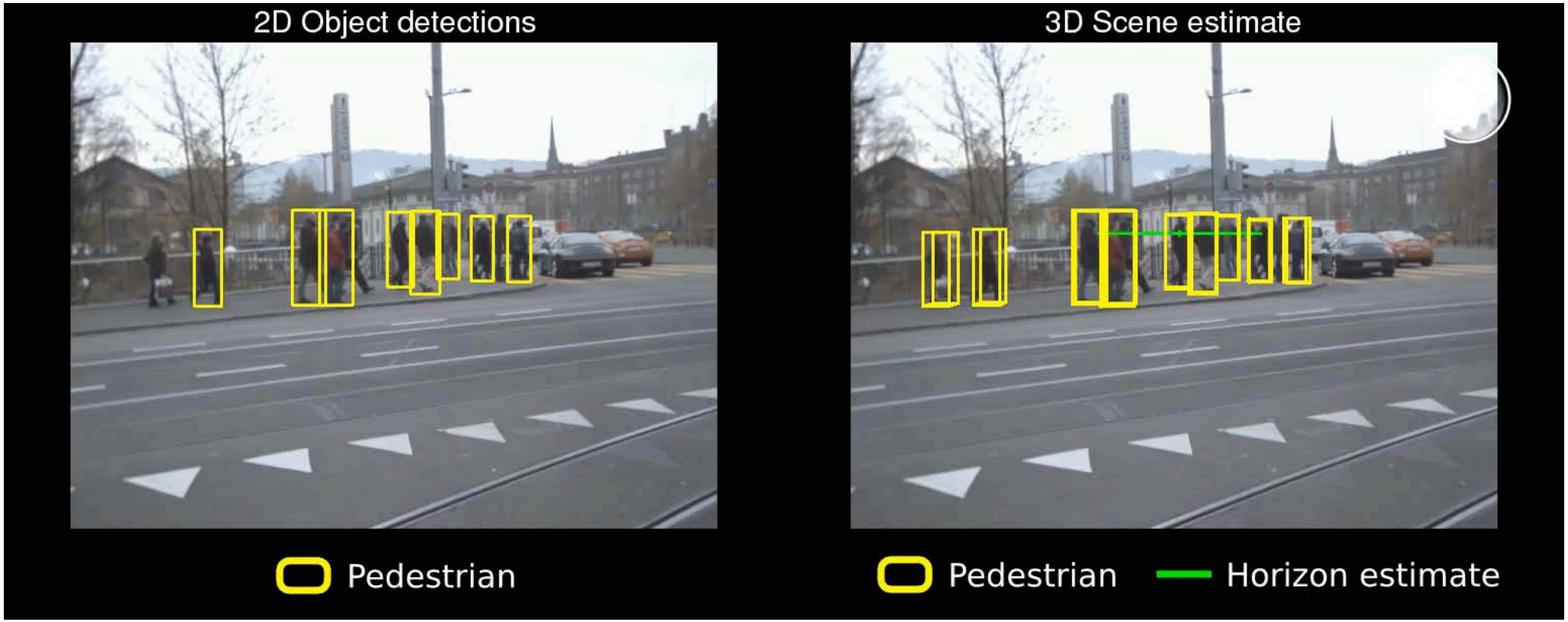
$\mathbf{P}(\mathcal{E}|\mathbf{X})$

$\mathbf{P}(\mathbf{X})$



[Wojek,Roth,Schindler,Schiele@PAMI'13]

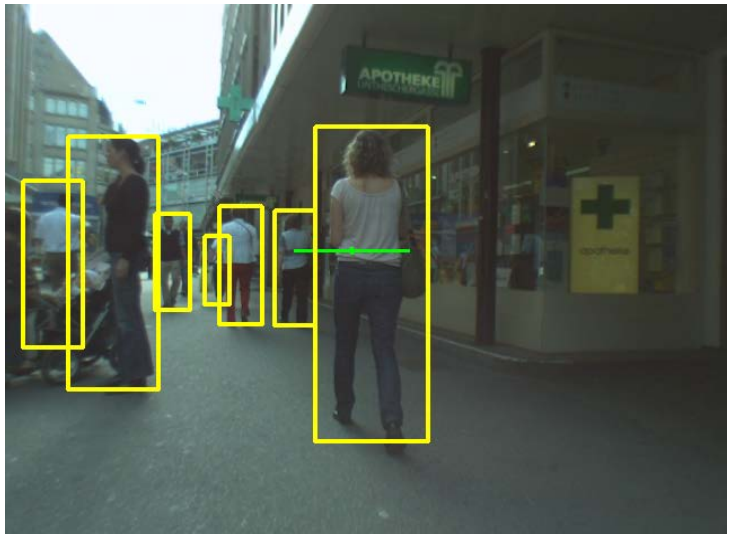
System sample video (pedestrians)

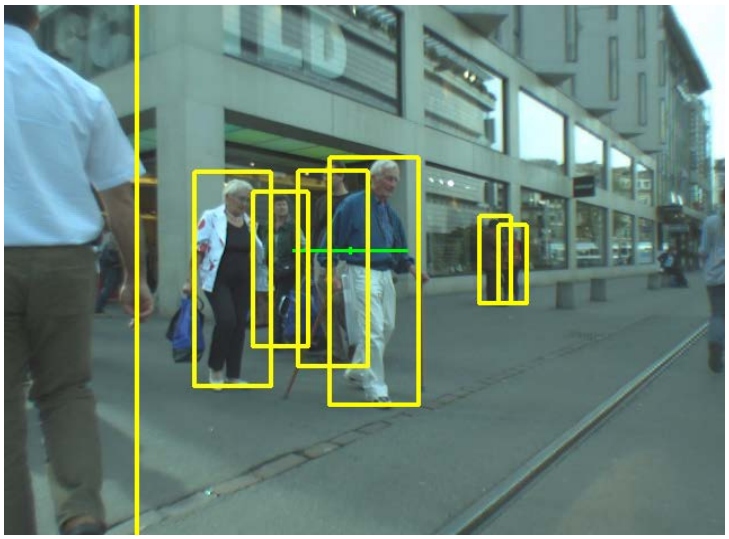


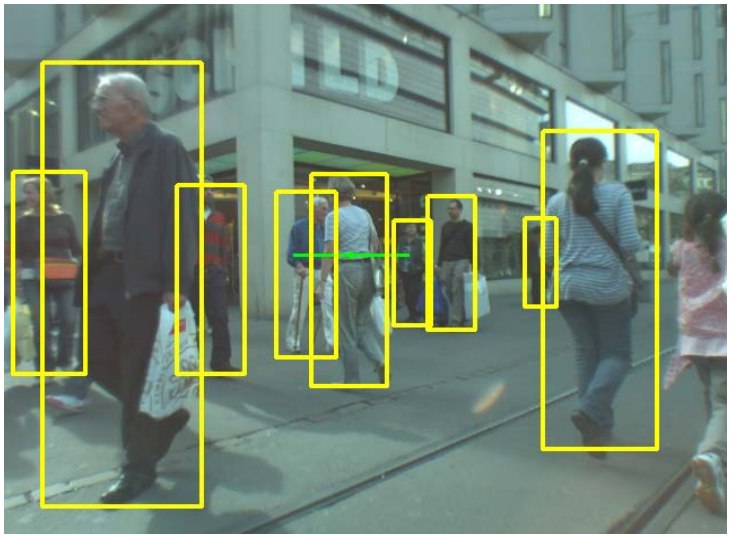
ETH-Loewenplatz sequence: By courtesy of ETH Zürich [Ess et al., PAMI '09]

[Wojek,Roth,Schindler,Schiele@CVPR'11]

Sample Result including Occlusion Reasoning

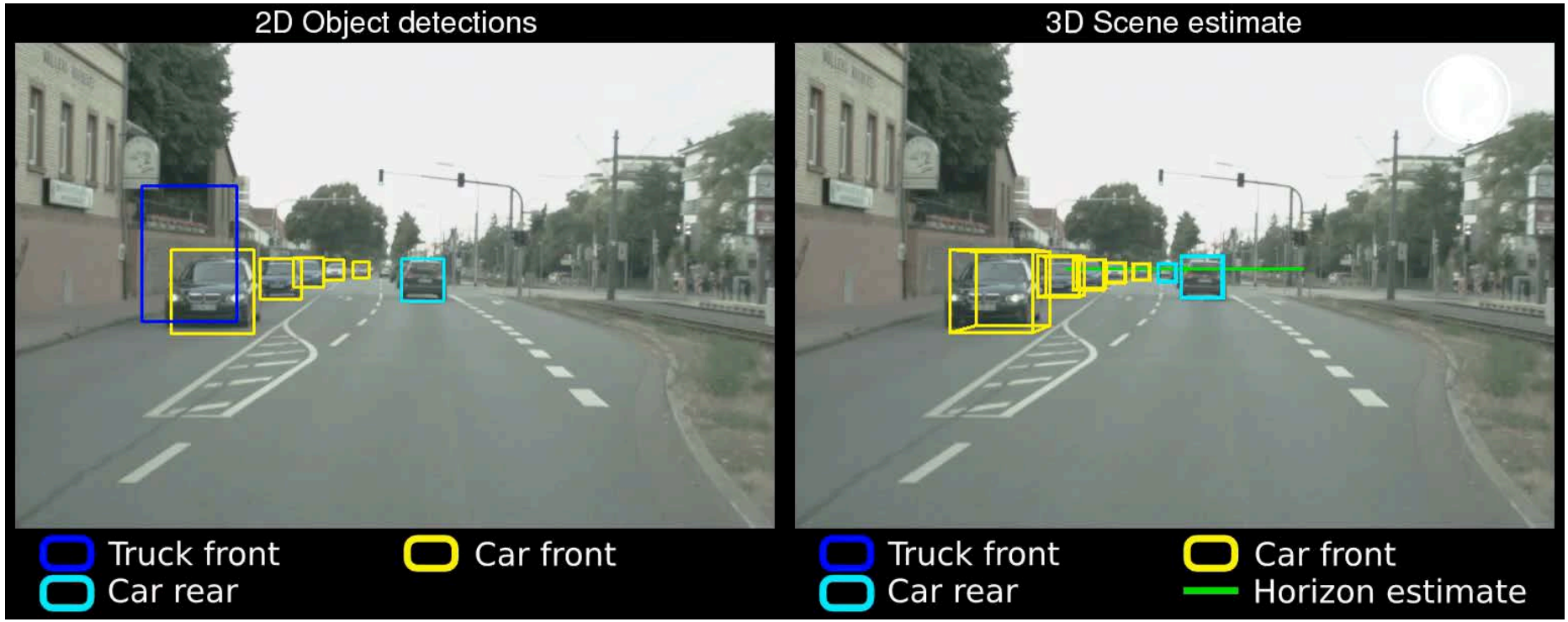






[Wojek,Roth,Schindler,Schiele@PAMI'13]

System sample video (different types of vehicles)



- **Message**
  - **modeling and reasoning in 3D is powerful:**
    e.g. for occlusion reasoning, inclusion of prior information

Advances Towards 3D Scene Understanding

- **Component Research on…**
  - 3D Object Recognition and Segmentation
  - People Detection and Tracking in 3D

- **Beyond Component Research on… (and Towards 3D Scene Understanding)**
  - 3D Scene Understanding - traffic scene analysis as a case study
  - **Knowledge Harvesting from Language**
  - Video and Scene Descriptions

Motivation: Knowledge Harvesting from Language

Log in / create account

WIKIPEDIA
The Free Encyclopedia

Article | Discussion | View source | View history | Search

Main page
Contents
Featured content
Current events
Random article
Donate to Wikipedia

Interaction

Giant Panda

The **giant panda** is a bear which is native to central-western and south western China.
It is easily recognized by its large, distinctive black patches around the eyes, over the ears, and across its round body.

Motivation: Knowledge Harvesting from Language

- Text can provide
  - **Attributes**

Motivation: Knowledge Harvesting from Language

- Text can provide
  - **Attributes**
  - **(Direct) similarities between classes / objects**

Motivation: Knowledge Harvesting from Language

- Text can provide
  - **Attributes**
  - **(Direct) similarities between classes / objects**
  - **Context information (might be visible or not)**

Motivation: Knowledge Harvesting from Language

- Visual Side
  - Videos of Activities

- Linguistic Side
  - Textual Descriptions (scripts)

First wash carrot with water. Then peel skin of the carrot with peeler. Finally, cut off ends with knife and slice carrot.

Motivation: Knowledge Harvesting from Language

- Visual Side
  - Videos of Activities

- Linguistic Side
  - Textual Descriptions (scripts)

First wash carrot with water. **Then** peel skin of the carrot with peeler. **Finally**, cut off ends with knife and slice carrot.

  - Shared information
    - **Structure**

Motivation: Knowledge Harvesting from Language

- Visual Side
  - Videos of Activities

- Linguistic Side
  - Textual Descriptions (scripts)

First wash carrot with **water**. **Then** peel skin of the **carrot** with **peeler**. **Finally**, cut off ends with **knife** and slice carrot.

  - Shared information
    - **Structure**
    - **Ingredients & Tools**

Motivation: Knowledge Harvesting from Language

- Visual Side
  - Videos of Activities

- Linguistic Side
  - Textual Descriptions (scripts)

First wash carrot with **water**. **Then peel** skin of the **carrot** with **peeler**. **Finally**, **cut off** ends with **knife** and **slice** carrot.

  - Shared information
    - **Structure**
    - **Ingredients & Tools**
    - **Activities**

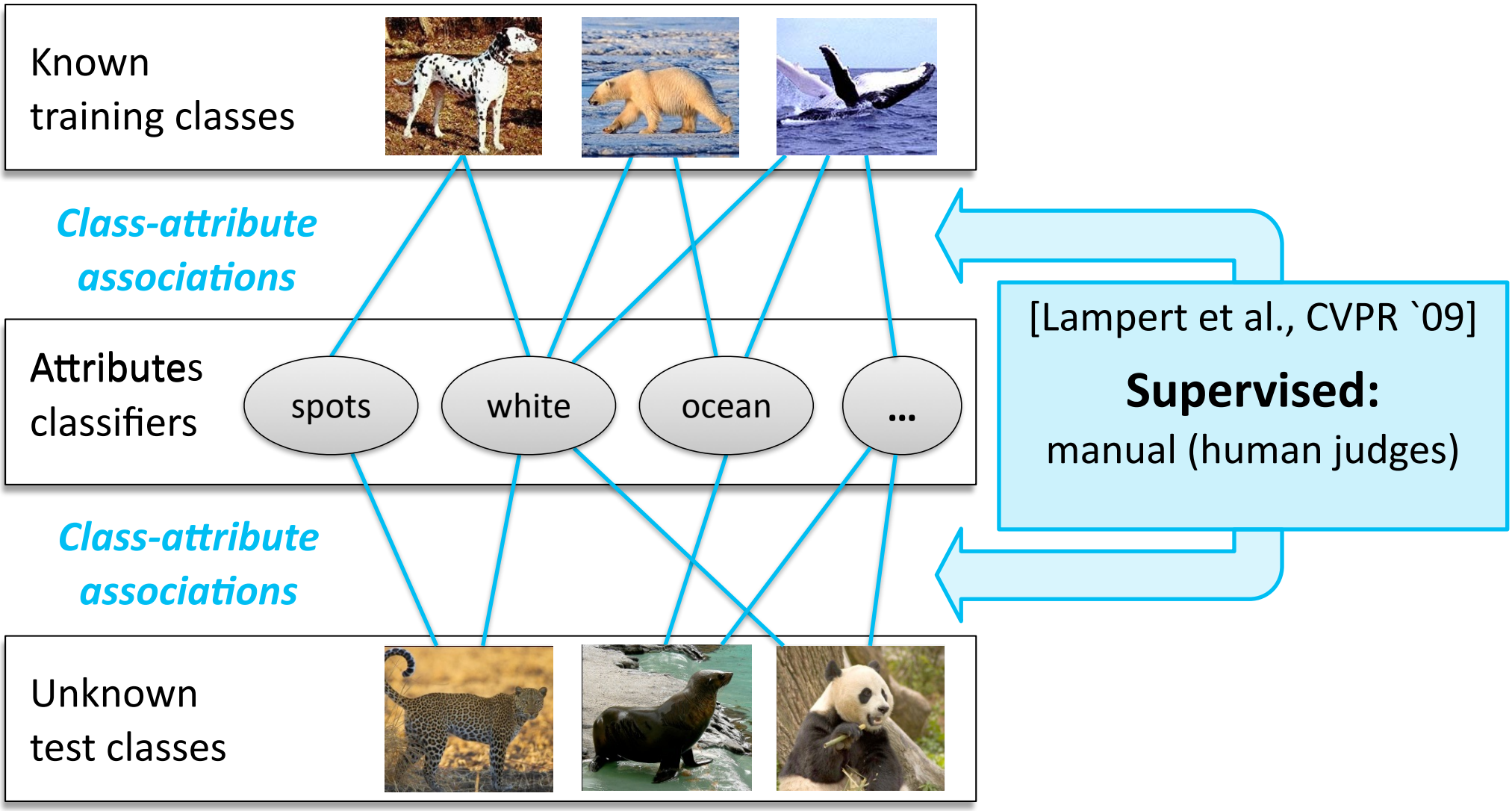
Knowledge Harvesting from Language

- **Why?**
  - **We won't have Sufficient Training Data to learn EVERYTHING from Data**

- Text is currently the largest source of knowledge
  - tapping into large text corpora (e.g. wikipedia)
  - leverage textual descriptions of images & videos
  - …
- Examples of interesting knowledge
  - Objects: relation to other objects, appearance, context, …
  - Activities: information about structure, objects involved, sub-activities, …
  - Prior knowledge (e.g. about traffic scenes)
  - …
- Examples from our work
  - Zero-Shot Learning of Objects and Activities

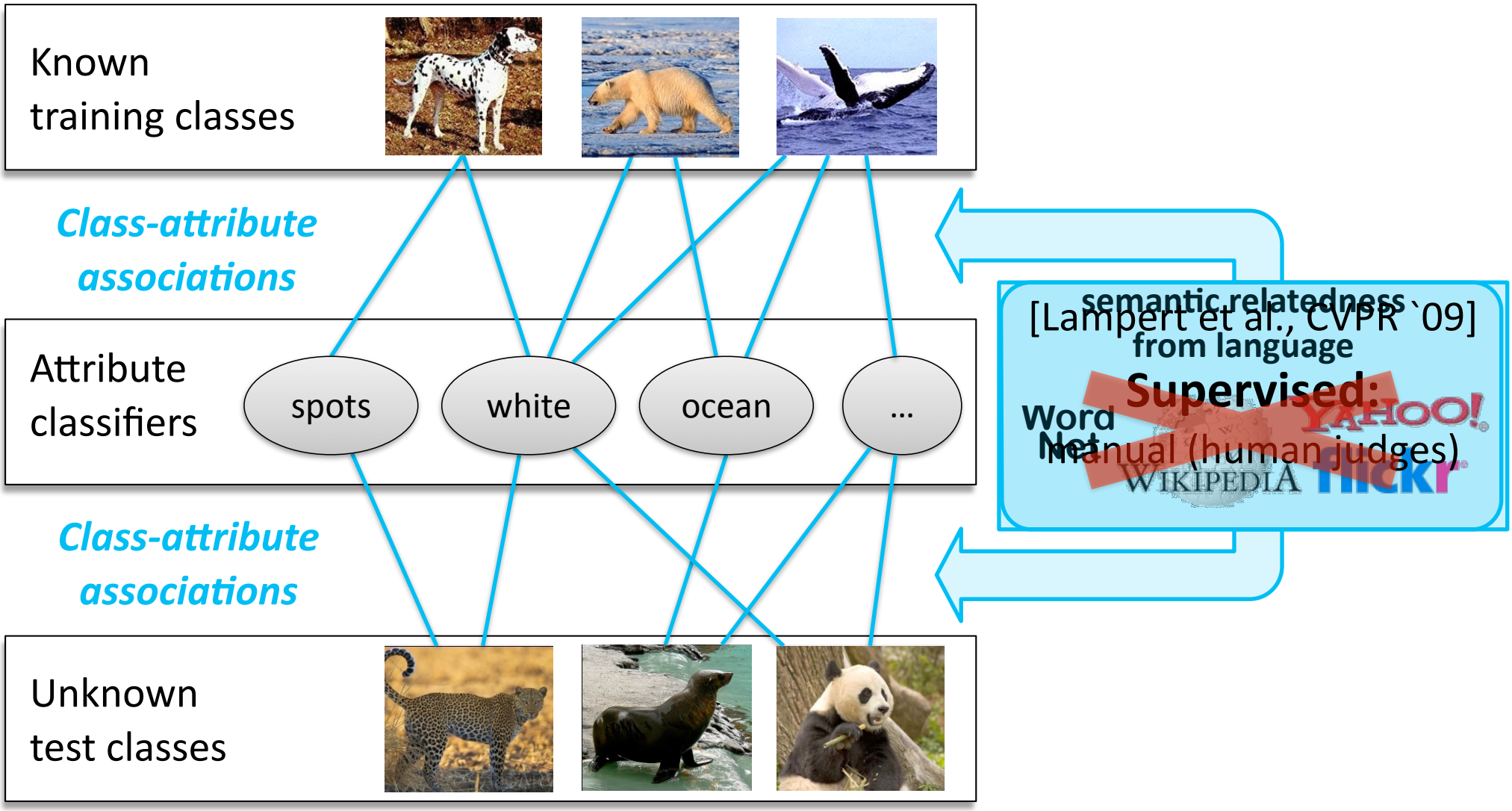
[Rohrbach,Stark,Szarvas,Gurevych,Schiele@CVPR'10]

Attribute-based Model for Object Class Recognition



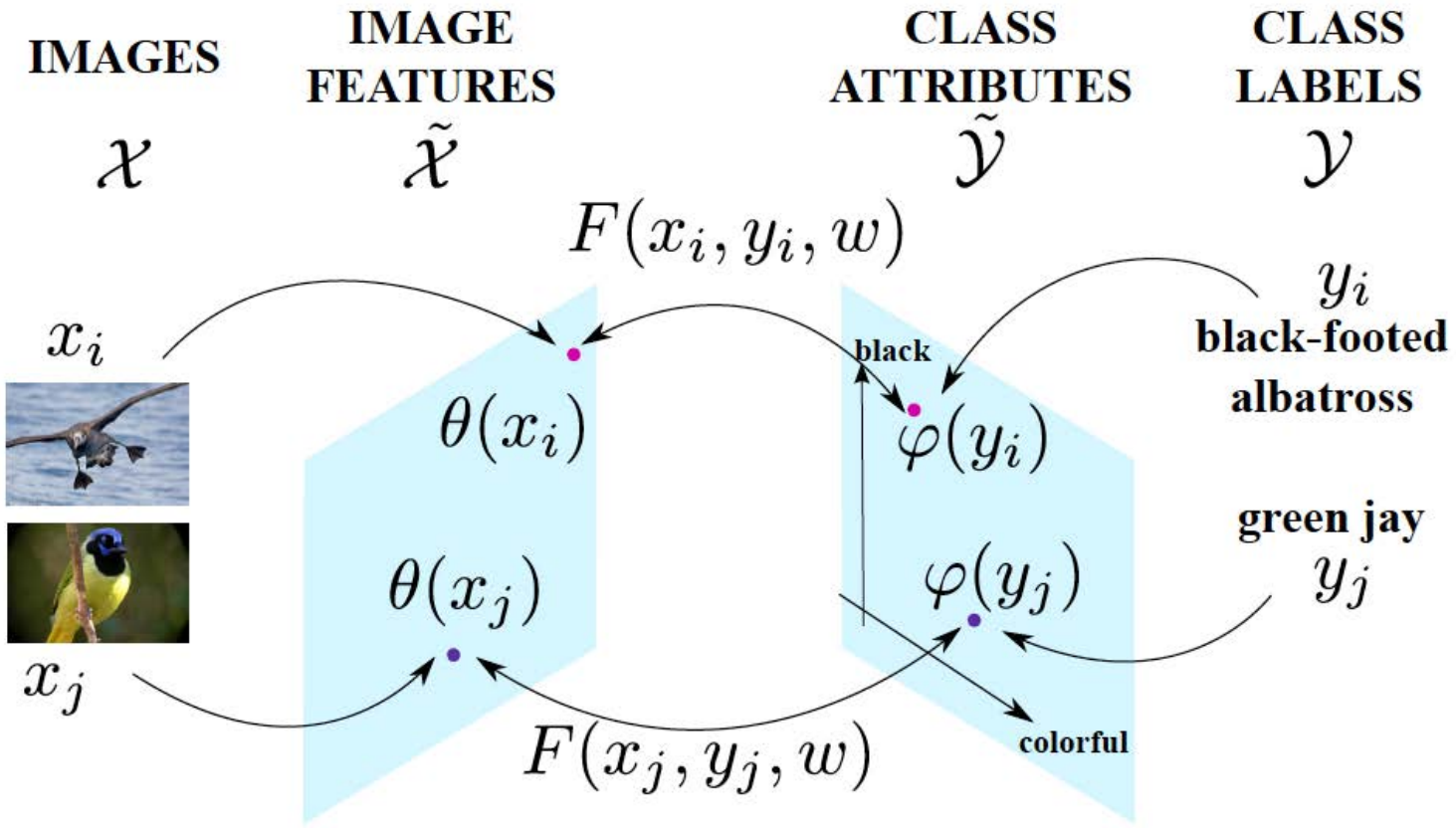
[Rohrbach,Stark,Szarvas,Gurevych,Schiele@CVPR'10]

Attribute-based Model for Object Class Recognition



[Akata,Reed,Walter,Lee,Schiele@CVPR'15]

Zero Shot Structured Embedding



$$f(x; W) = \arg\max_{y \in \mathcal{Y}} F(x, y; W) = \arg\max_{y \in \mathcal{Y}} \underbrace{\theta(x)^\top}_{\text{img feat}} W \underbrace{\varphi(y)}_{\text{class att}}$$

[Akata,Reed,Walter,Lee,Schiele@CVPR'15]

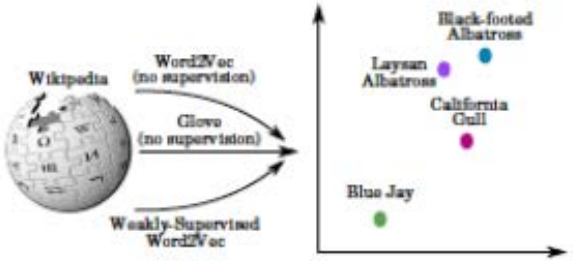
Output Embeddings beyond Attributes…

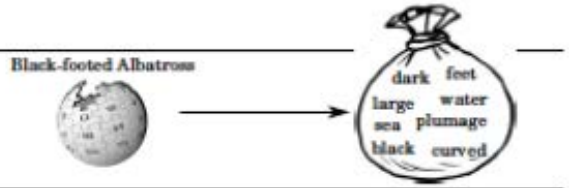
- **Attributes** - require human **supervision** !

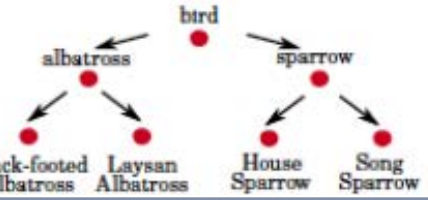
Attributes [2]	$\varphi^{0,1}$ $\varphi^{\mathcal{A}}$	$[0,0,1] \rightarrow hummingbird = sparrow < albatross$ $[2,10,90] \rightarrow hummingbird < sparrow << albatross$

- **Unsupervised** Embeddings (any Vectorized Representation)

Word2Vec [3]	$\varphi^{\mathcal{W}}$	(1) look-up table to retrieve word vector in vocabulary. (2) predict target word from context via hie. soft-max.
GloVe [4]	$\varphi^{\mathcal{G}}$	Objective: $\varphi^{\mathcal{G}}(v1).\varphi^{\mathcal{G}}(v2) = co-occurrence(v1,v2)$
Weakly-supervised Word2Vec	$\varphi^{\mathcal{W}_{ws}}$	Pre-train Word2vec layer (1) weights on `wikipedia`, Fine-tune layer (2) weights on fine-grained text corpus: $L = \sum_{w,c \in D_+} \log \sigma(v_c^\top v_w) + \sum_{w',c \in D_-} \log \sigma(-v_c^\top v_{w'})$ $v_c = \sum_{i \in \text{context(w)}} v_i / \lvert\text{context(w)}\rvert$
Bag-of-Words	$\varphi^{\mathcal{B}}$	Histograms of frequent words in `wikipedia` document.
Hierarchies	$\varphi^{\mathcal{H}}$	Distance between classes (synsets) in `WordNet` is calculated using different hierarchical distance metrics.







[Akata,Reed,Walter,Lee,Schiele@CVPR'15]

Zero Shot Structured Embedding - Results

- Dataset (more in the paper)
  - AWA - animals with attributes
- Input Embedding
  - GoogleLeNet 1024-dimensional output of last layer
- Main Results
  - unsupervised zero-shot learning shows promising results
  - complementary to human supervision (attributes)

supervision	source	φ	AWA
unsupervised	text	φ^{W}	51.2
	text	φ^{G}	58.8
	text	φ^{B}	44.9
	WordNet	φ^{H}	51.2
	text + WordNet	cmb	**60.1**
supervised	human	$\varphi^{0,1}$	52.0
	human	φ^{A}	66.7
	human + text	cmb	**73.9**
SoA [1]	human	φ^{A}	49.4

Advances Towards 3D Scene Understanding

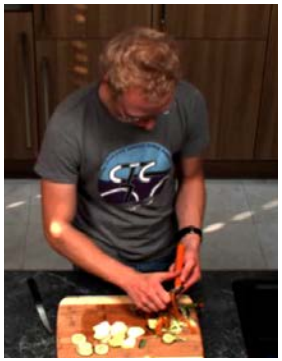
- **Component Research on…**
  - 3D Object Recognition and Segmentation
  - People Detection and Tracking in 3D

- **Beyond Component Research on… (and Towards 3D Scene Understanding)**
  - 3D Scene Understanding - traffic scene analysis as a case study
  - Knowledge Harvesting from Language
  - **Video and Scene Descriptions**

[Rohrbach,Qiu,Titov,Thater,Pinkal,Schiele@iccv13]

Describing Video With Natural Language

- Parallel Corpus
  - videos + semantic annotations + language descriptions



ACTIVITY
TOOL
OBJECT
SOURCE
TARGET

Peel skin of the carrot with peeler.

- Idea: 2-Step Approach:
  - 1. Extract Semantic Representation:
    - Video ➽ Semantic Representation
  - 2. Learn Language Translation Model:
    - Semantic Representation ➽ Language Descriptions

[Rohrbach@iccv13] & [Rohrbach@gcpr14]

Example: Multi-sentence video description

A woman got out a knife from the drawer. Then, she peeled the onion. She took out a cutting board and a cutting board from the drawer. The woman chopped the onion on the cutting board. She put the onion in the pan. Next, she diced and cut the onion on the cutting board. She threw away the peel. She added the onions in the pan. ...

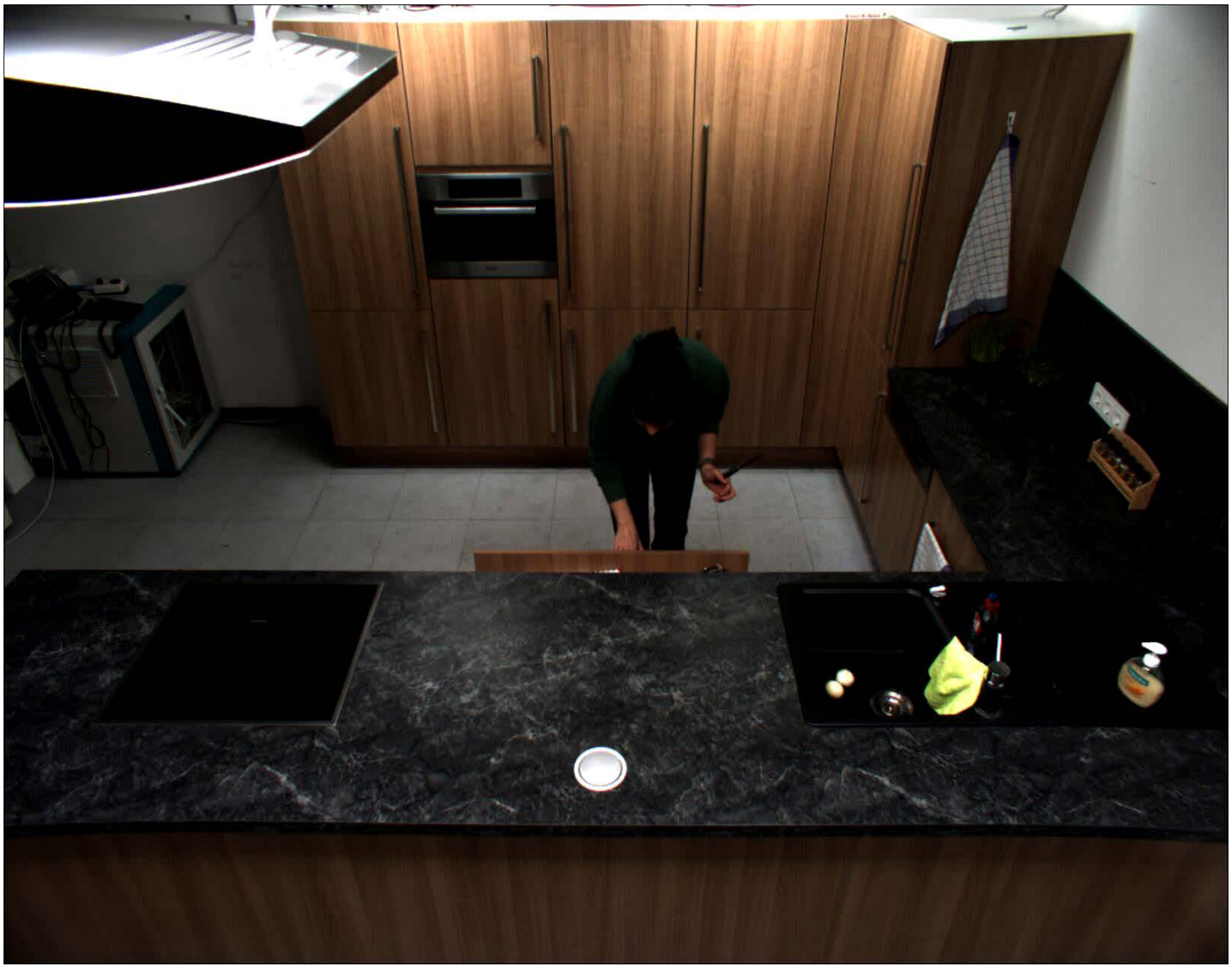
[Rohrbach@iccv13] & [Rohrbach@gcpr14]

Example: Multi-sentence video description

A woman got out a knife from the drawer. Then, she peeled the onion. She took out a cutting board and a cutting board from the drawer. The woman chopped the onion on the cutting board. She put the onion in the pan. Next, she diced and cut the onion on the cutting board. She threw away the peel. She added the onions in the pan. ...

[Rohrbach@iccv13] & [Rohrbach@gcpr14]

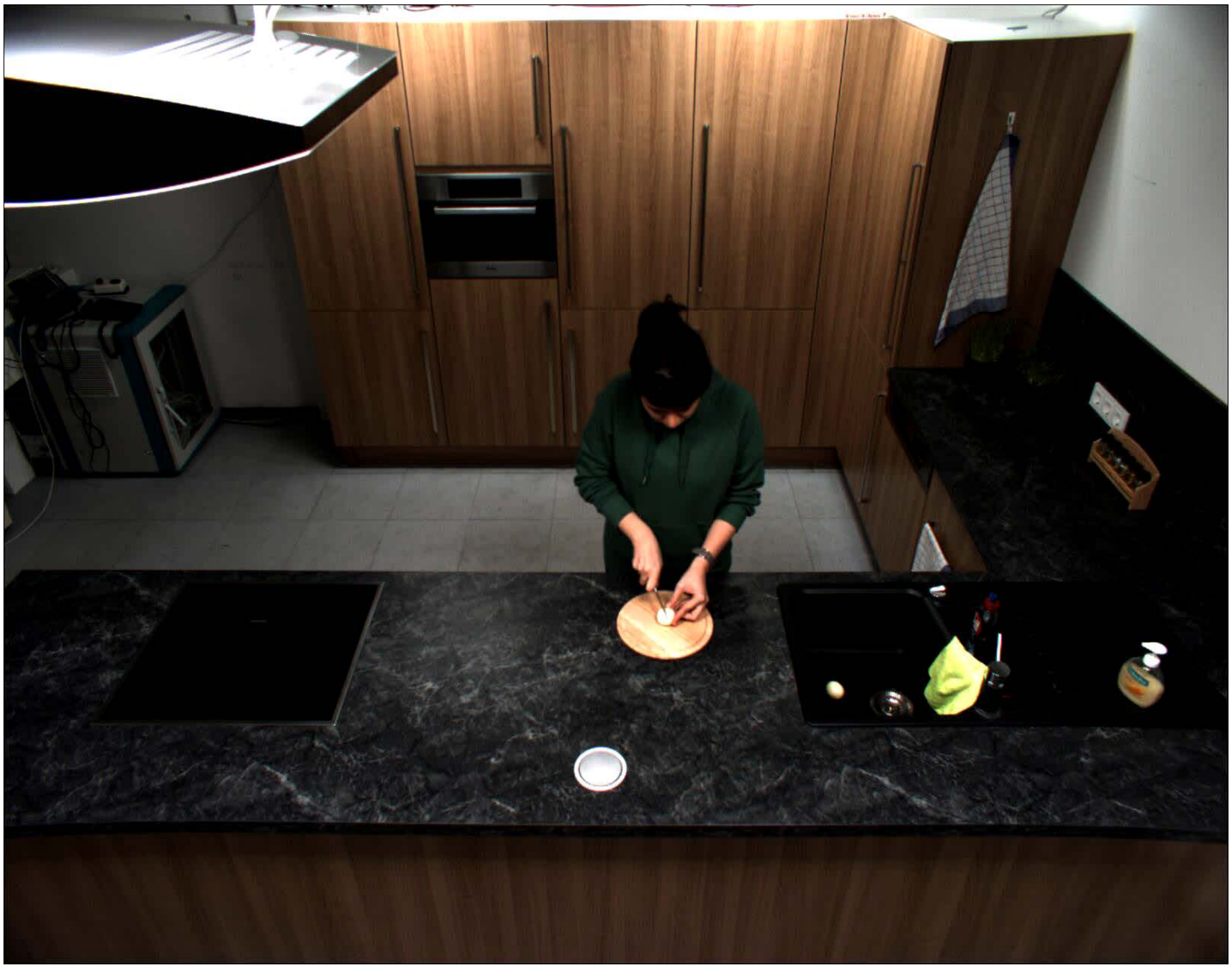
Example: Multi-sentence video description



A woman got out a knife from the drawer. Then, she peeled the onion. She took out a cutting board and a cutting board from the drawer. The woman chopped the onion on the cutting board. She put the onion in the pan. Next, she diced and cut the onion on the cutting board. She threw away the peel. She added the onions in the pan. ...

[Rohrbach@iccv13] & [Rohrbach@gcpr14]

Example: Multi-sentence video description



A woman got out a knife from the drawer. Then, she peeled the onion. She took out a cutting board and a cutting board from the drawer. The woman chopped the onion on the cutting board. She put the onion in the pan. Next, she diced and cut the onion on the cutting board. She threw away the peel. She added the onions in the pan. ...

[Rohrbach@iccv13] & [Rohrbach@gcpr14]

Example: Multi-sentence video description

A woman got out a knife from the drawer. Then, she peeled the onion. She took out a cutting board and a cutting board from the drawer. The woman chopped the onion on the cutting board. She put the onion in the pan. Next, she diced and cut the onion on the cutting board. She threw away the peel. She added the onions in the pan. ...

[Rohrbach@iccv13] & [Rohrbach@gcpr14]

Example: Multi-sentence video description

A woman got out a knife from the drawer. Then, she peeled the onion. She took out a cutting board and a cutting board from the drawer. The woman chopped the onion on the cutting board. She put the onion in the pan. Next, she diced and cut the onion on the cutting board. She threw away the peel. She added the onions in the pan. ...

[Rohrbach@iccv13] & [Rohrbach@gcpr14]

Example: Multi-sentence video description

A woman got out a knife from the drawer. Then, she peeled the onion. She took out a cutting board and a cutting board from the drawer. The woman chopped the onion on the cutting board. She put the onion in the pan. Next, she diced and cut the onion on the cutting board. She threw away the peel. She added the onions in the pan. ...

[Rohrbach@iccv13] & [Rohrbach@gcpr14]

Example: Multi-sentence video description

A woman got out a knife from the drawer. Then, she peeled the onion. She took out a cutting board and a cutting board from the drawer. The woman chopped the onion on the cutting board. She put the onion in the pan. Next, she diced and cut the onion on the cutting board. She threw away the peel. She added the onions in the pan. ...

Take Home Messages

- Machine learning has been and will continue to be a driver
  - lots of data (internet, storage, …)
  - fast processing (CPU and GPU clusters, …)
  - powerful machine learning models (deep learning, graphical models, …)
- 3D information is essential and becomes viable
  - through 3D sensors
  - through 3D modeling & inference
- Combining knowledge and computer vision is important
  - text is currently the largest source of knowledge
  - knowledge mining in all kinds of data (youtube, flickr, social networks, …)

- We are always looking for great people (PostDocs, Students, …:-)
  - send email to schiele@mpi-inf.mpg.de

max planck institut
informatik

UNIVERSITÄT
DES
SAARLANDES

Towards 3D Visual Scene “Understanding”

Bernt Schiele

Max Planck Institute for Informatics, Saarbrücken

Saarland University, Saarbrücken

schiele@mpi-inf.mpg.de