Computer Vision II – Scene Understanding

Michael Yang

Roadmap (last lecture)

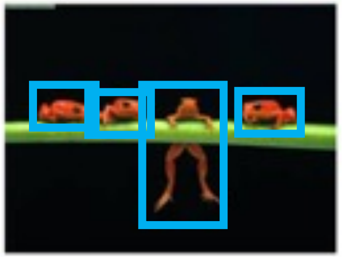
- Defining the Problem
- Rigid Template
  - HOG for human detection
  - Exemplar SVM detector
- Part Based Detector
  - Deformable Part Model
  - Poselets
- New development for object detection

Class-based recognition: Level of Detail

- Image Categorization (next lecture)
  - One or more categories per image

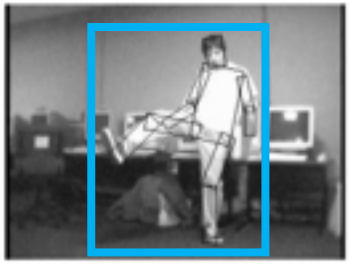
Frog, branch

- Object Class Detection
  - Also find bounding box

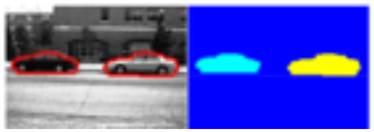


2D bounding box for each frog

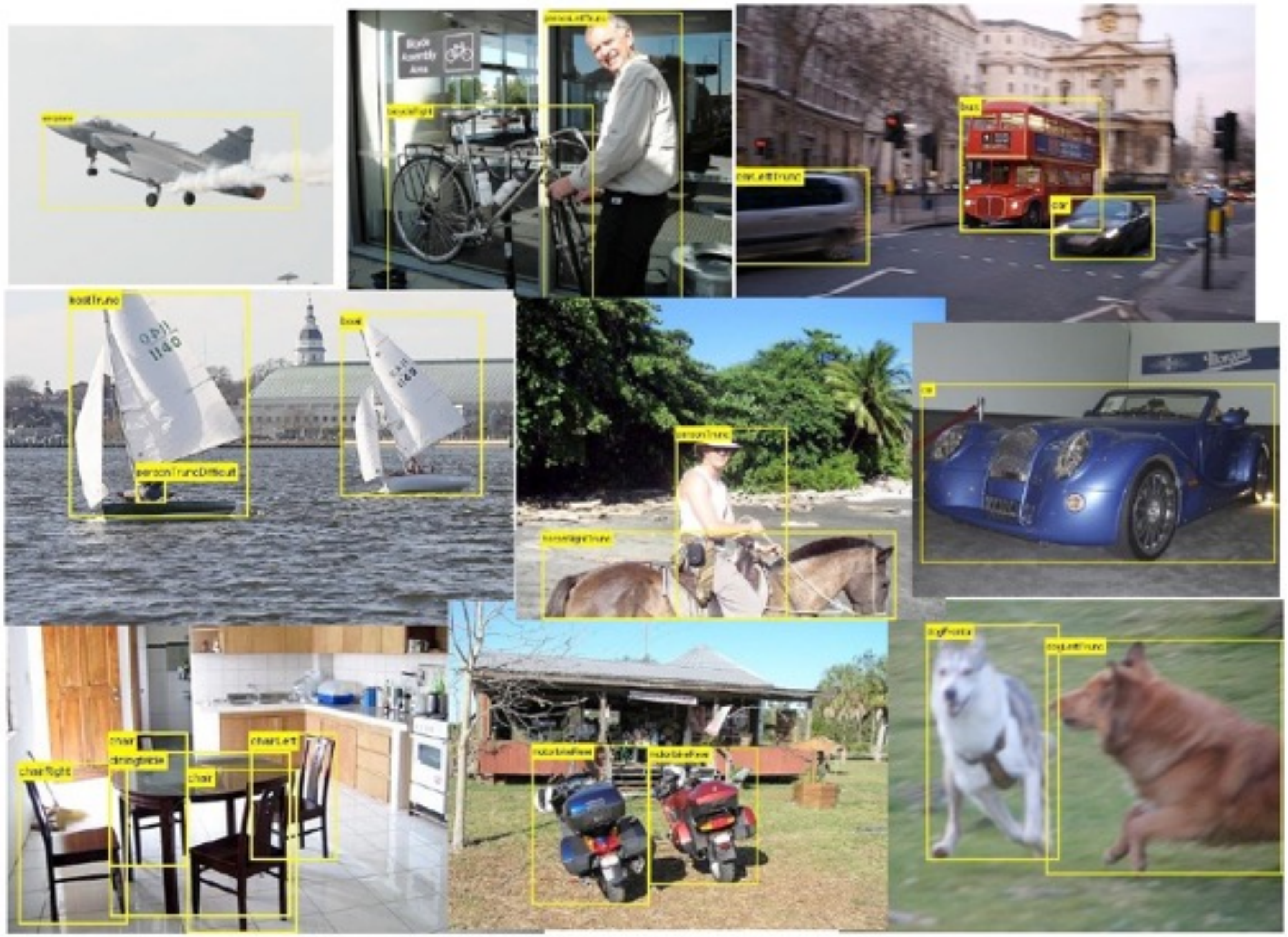
- Part-based Object Detection
  - Find parts of the object (and in this way the full object)



- Semantic Segmentation (segmentation implies pixel-wise accuracy)
  - Object-class segmentation



Task: Generic object detection



Summary of Basic object detection Steps

Training:

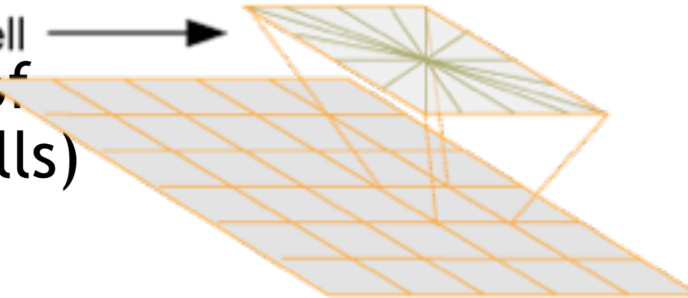
Train a classifier describe the detection target

Testing :

Detection by binary classification on all location

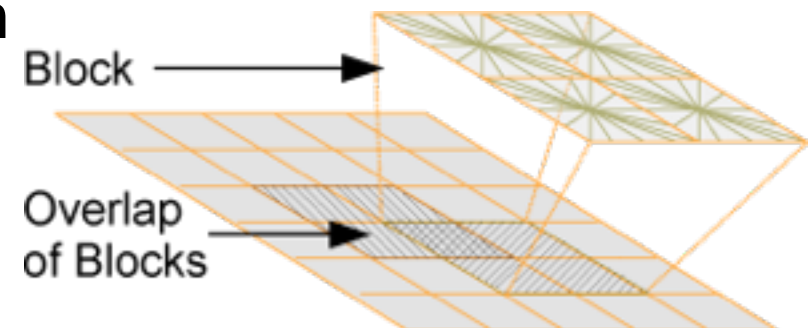
HOG Descriptor:

1. **Compute gradients** on an image region of 64x128 pixels



2. **Compute histograms** on 'cells' of typically 8x8 pixels (i.e. 8x16 cells)

3. **Normalize histograms** within overlapping blocks of cells

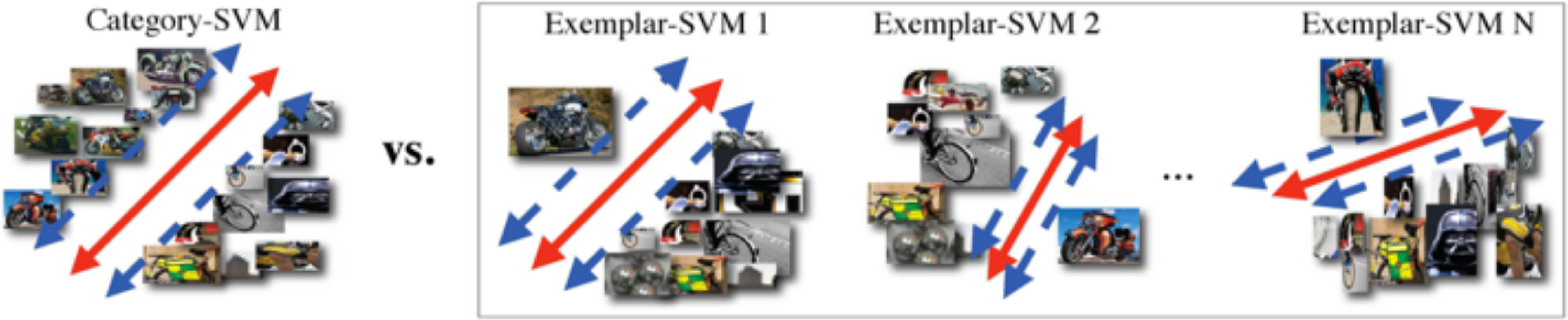


4. **Concatenate histograms**

It is a typical procedure of feature extraction !

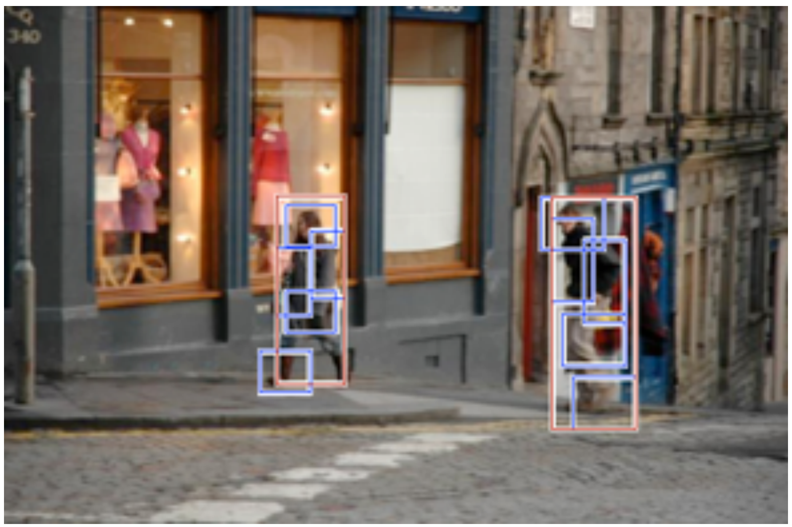
Exemplar-SVM

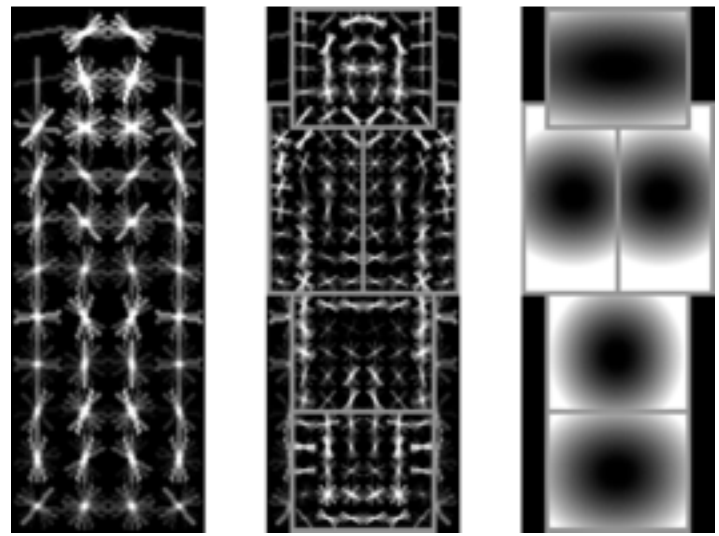
- Still a rigid template,but train a separate SVM for each positive instance



For each category it can has exemplar with different size aspect ratio

DPM : Object Detection with Discriminatively Trained Part Based Models





P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, [Object Detection with Discriminatively Trained Part Based Models](), PAMI 32(9), 2010

Roadmap (this lecture)

- Part Based Detector (cont. last lecture)
  - Deformable Part Model
  - Poselets
- Scene Understanding Problem
- Context
- Spatial Layout
- 3D Scene Understanding

Part Based Detector

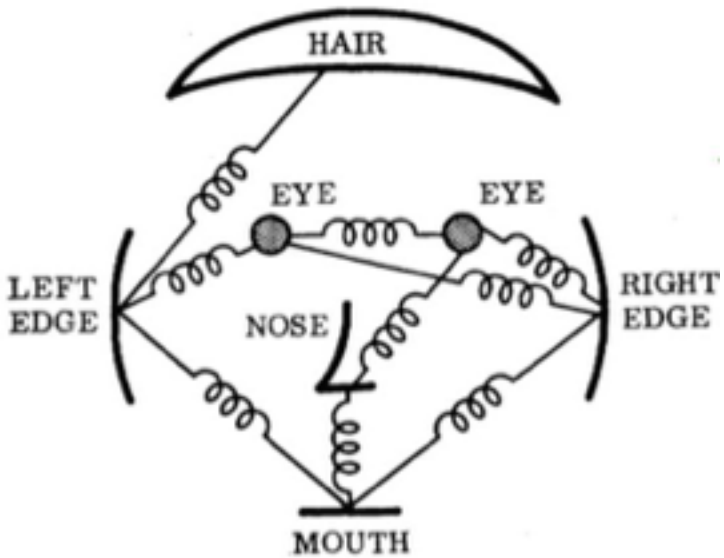
- Pictorial Structures

- Without part label
  - Deformable part model

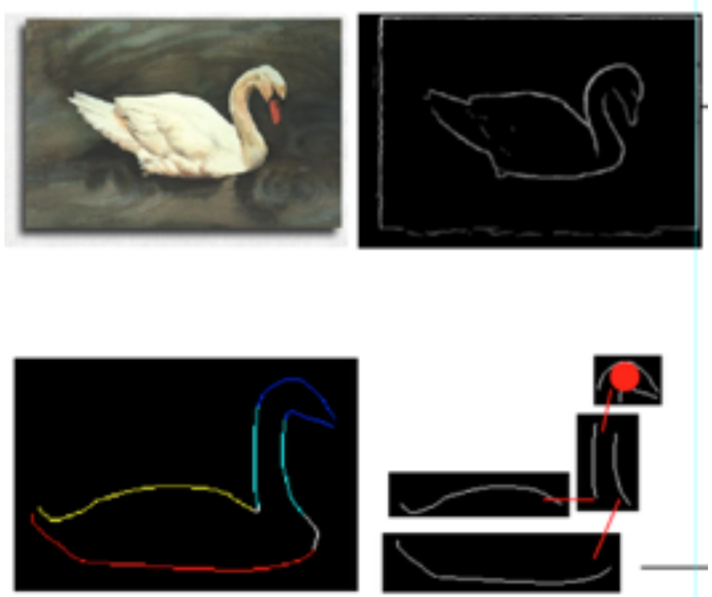
- With part labeled
  - Poselets

Part Based Detector

Objects are represented by features of parts and spatial relations between parts



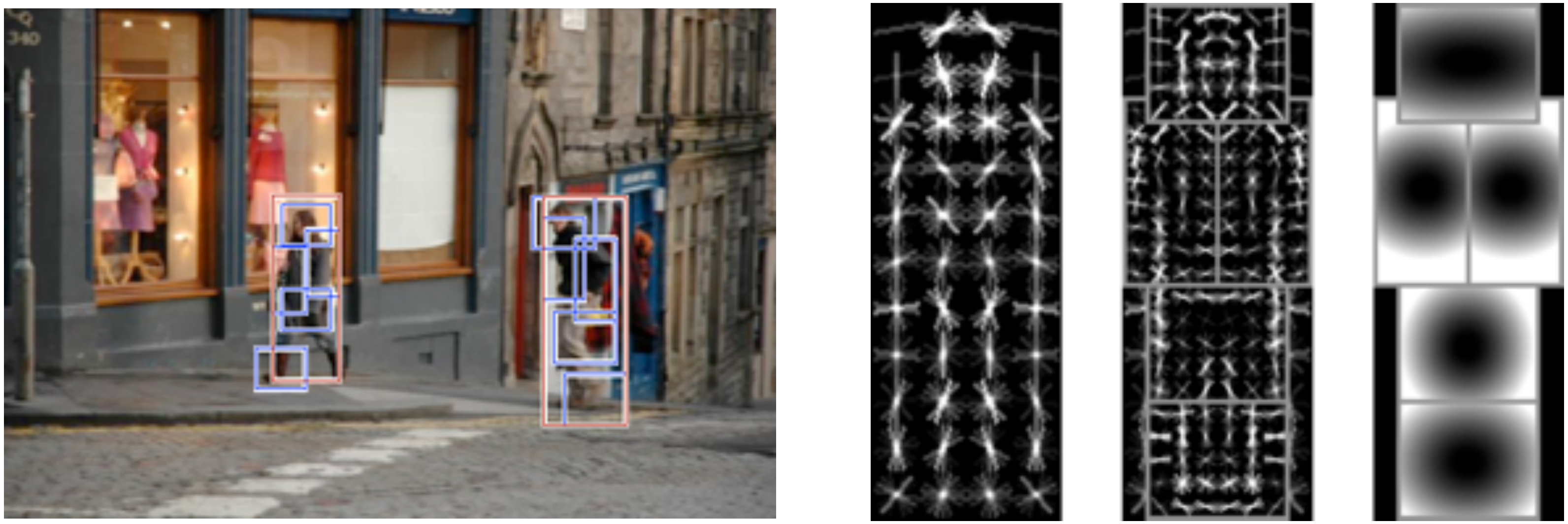
Face model by Fischler and Elschlager '73



Part Based Detector

- How to defined the parts for one object category
- How to represent their spatial relation shape
- How to combine parts detection and spatial relations to obtained the final detection

DPM : Object Detection with Discriminatively Trained Part Based Models



P. Felzenszwalb, R. Girshick, D. McAllester, D. Ramanan, [Object Detection with Discriminatively Trained Part Based Models](), PAMI 32(9), 2010

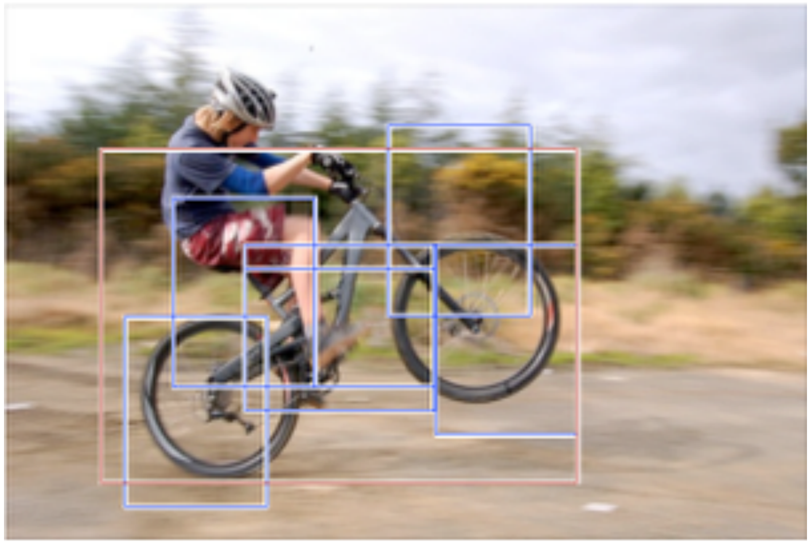
DPM: overview

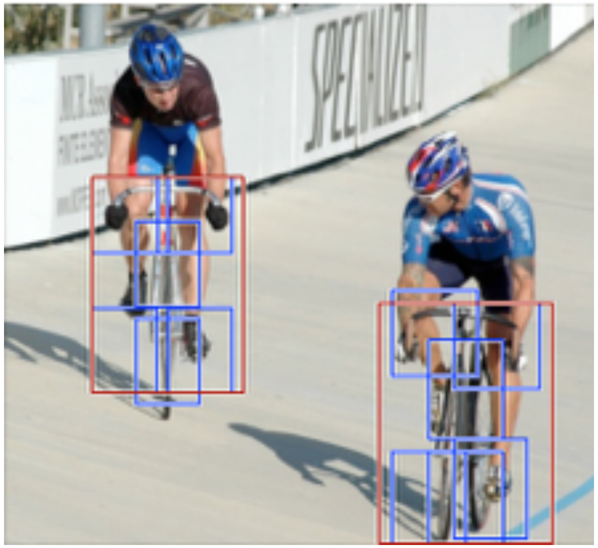
- Each category detector has mixture of deformable part models (components)

- Each component has global template + deformable parts

- Fully trained from bounding boxes alone (Latent SVM)

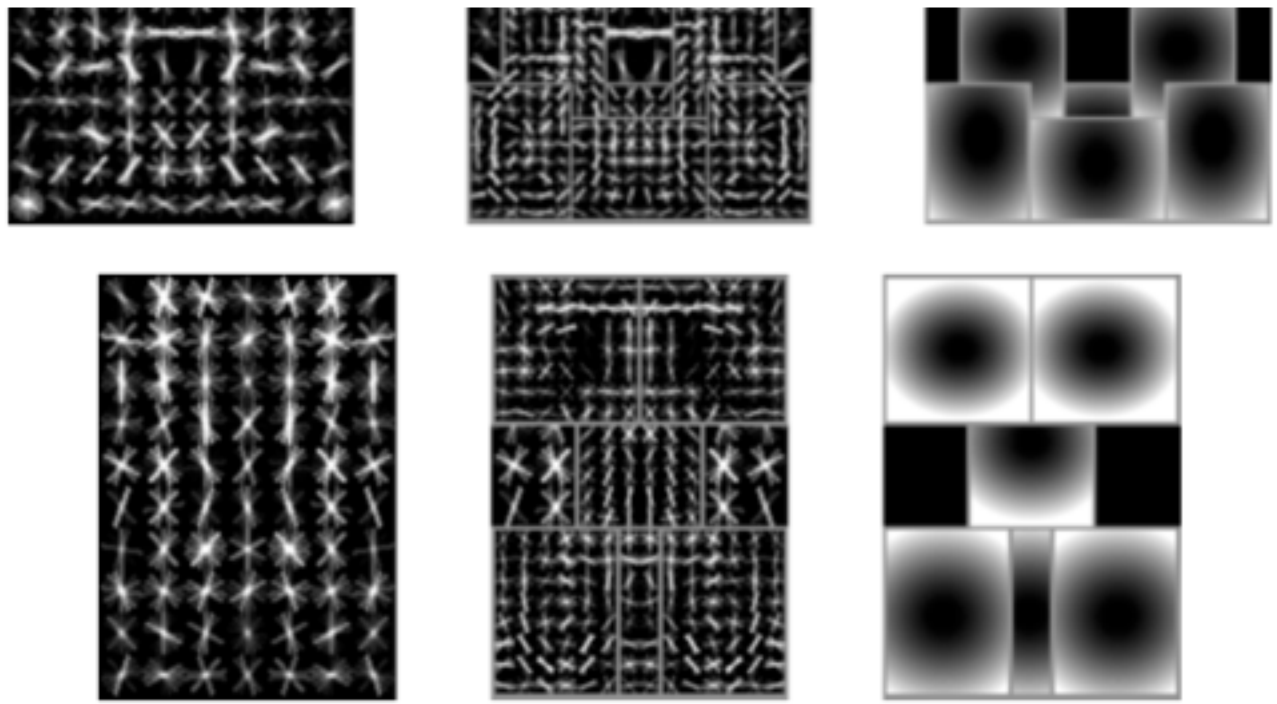
DPM: component





- Each category detector has mixture of component for different aspect ratio (handle intra-class variance)
- Each component has a it's own DPM model

DPM: component



root filters
coarse resolution

part filters
finer resolution

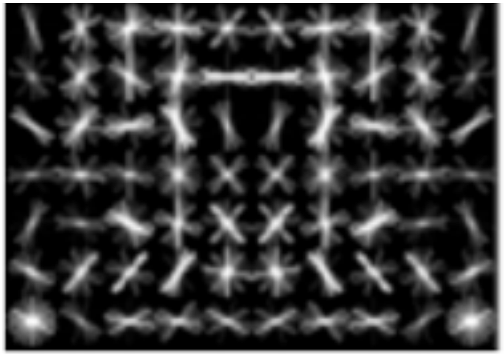
deformation
models

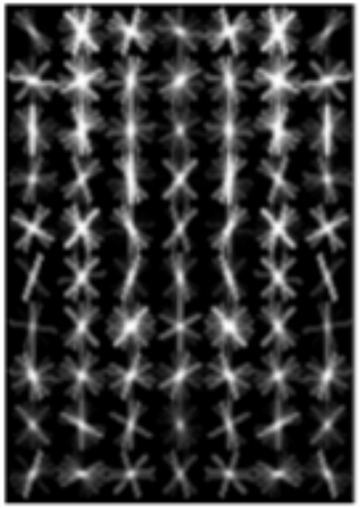
Each component has a root filter F_0
and n part models (F_i, v_i, d_i)

F: filter, v: 2D vector for anchor position, d: deformation parameter

DPM: Initialization

Root filter for each component

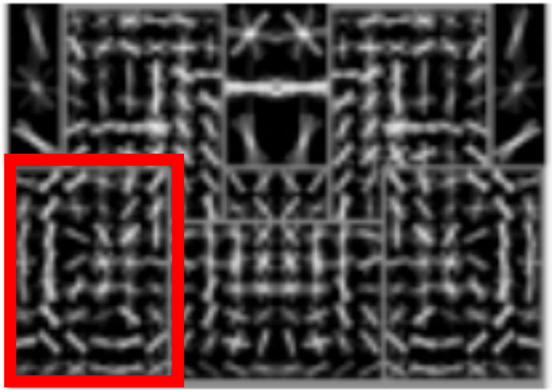


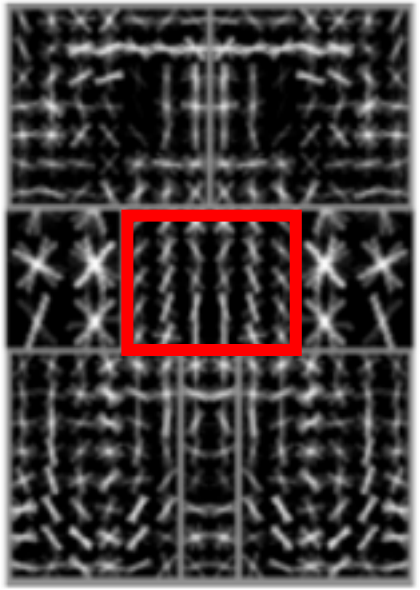


- For each component warp all positives to have same size
- Random pick negatives with same size
- Standard SVM no latent information

DPM: Initialization

Initializing Part Filter

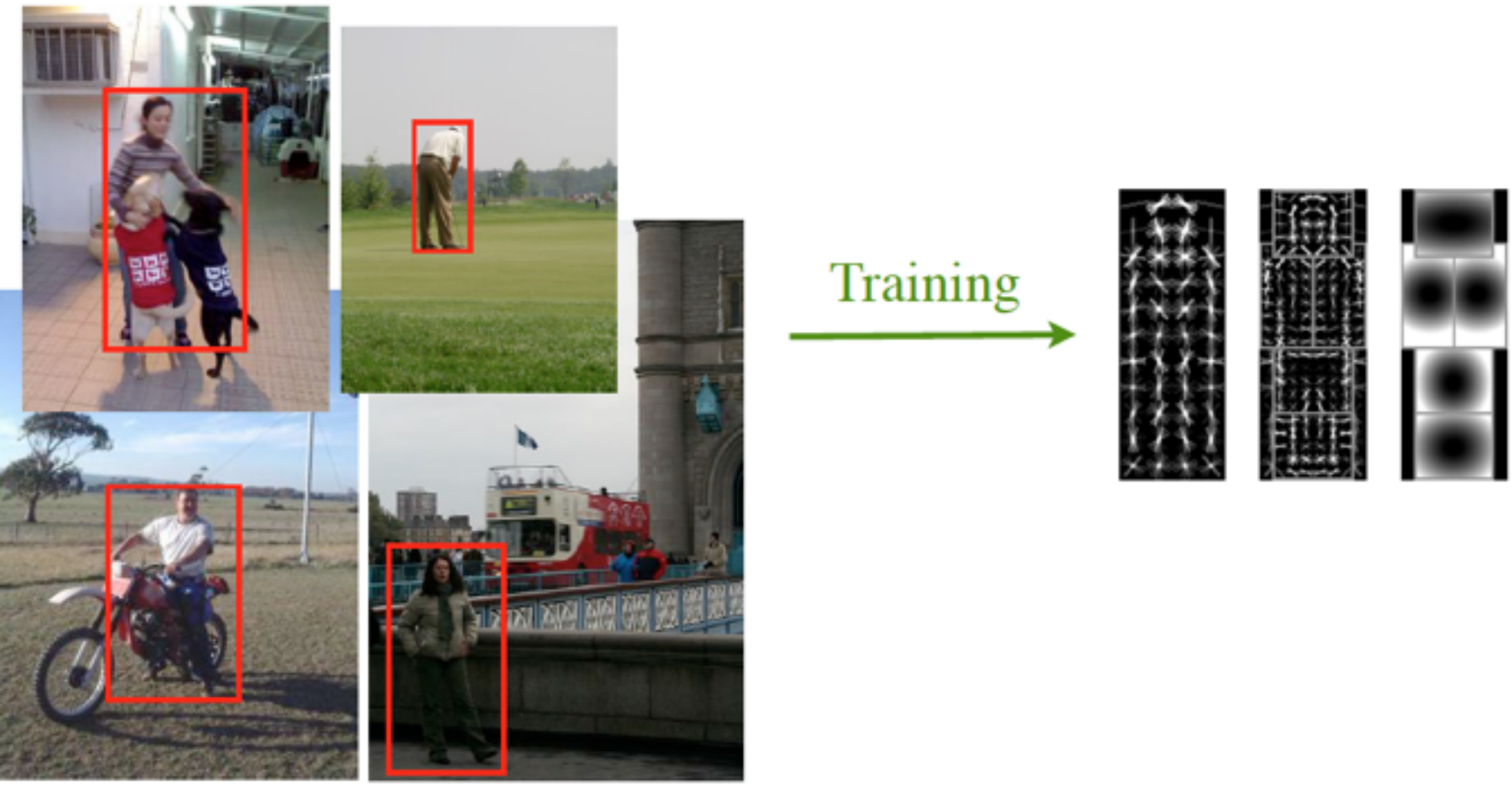




- Fixed number : 6 parts per component
- Choose the high-energy regions of the root filter
  (Energy : norm of positive weight in subwindow)
- Greedy approach: once part placed set to zero and find next high-energy part

DPM: Training

- Training data consists of images with labeled bounding boxes.
- Need to learn the model structure, filters and deformation costs.



DPM: Detection

$$\mathrm{score}(p_0,\ldots,p_n) = \overbrace{\sum_{i=0}^{n} \underbrace{F_i}_{\text{filters}} \cdot \phi(H,p_i)}^{\text{“data term”}} - \overbrace{\sum_{i=1}^{n} \underbrace{d_i}_{\text{deformation parameters}} \cdot \underbrace{(dx_i^2, dy_i^2)}_{\text{displacements}}}^{\text{“spatial prior”}}$$

Score for one part at certain location :
filter response score – deform cost relative to root

F: subwindow filter, F_i: vector by concatenating the weight vectors

phi: vector by concatenating the feature vectors in the subwindow

H: feature pyramid, p = (x, y, l) specify a position (x, y) in pyramid level l

DPM: Detection

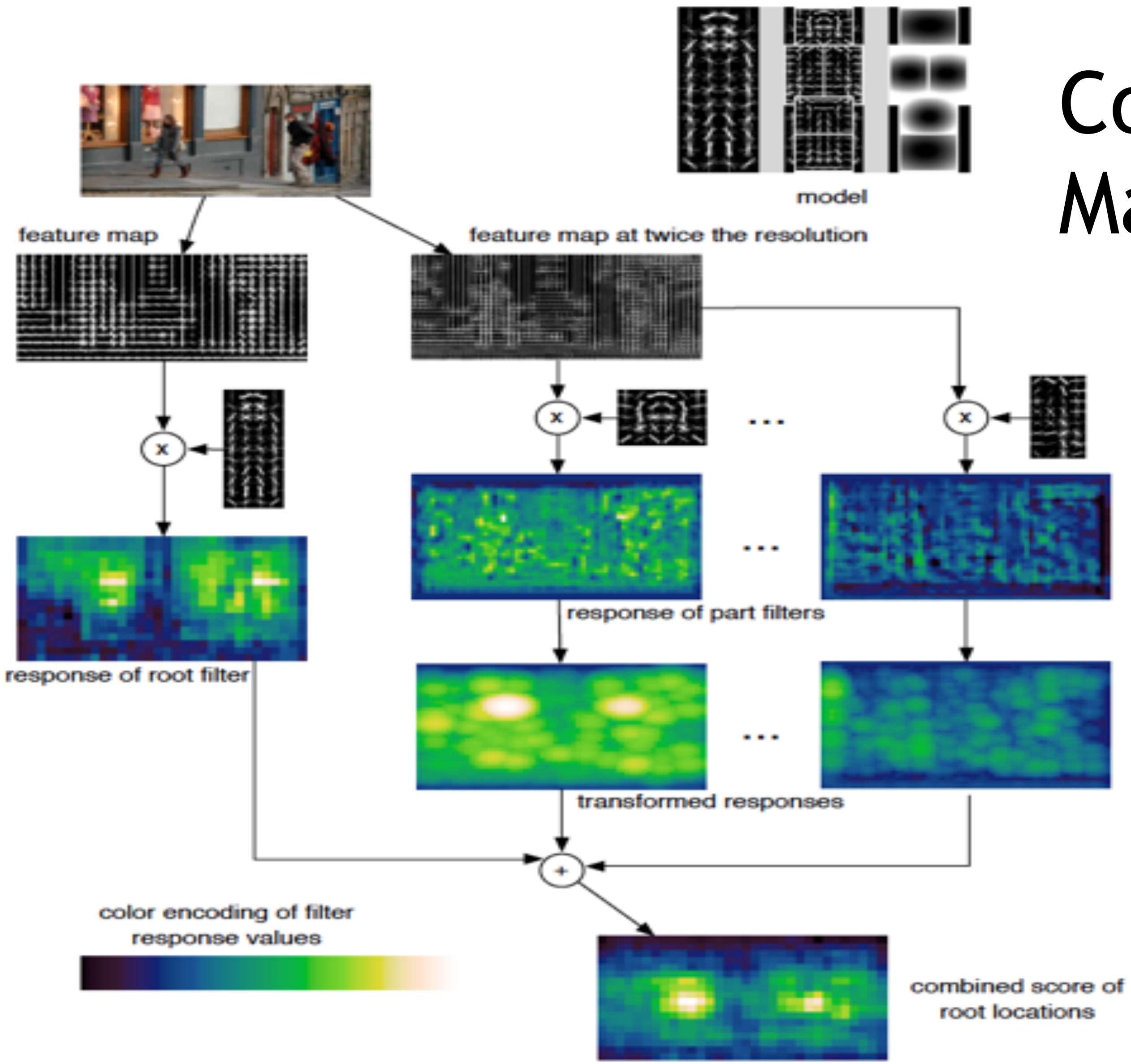
- Define an overall score for each root location
  - Based on best placement of parts

$$\text{score}(p_0) = \max_{p_1,\ldots,p_n} \text{score}(p_0,\ldots,p_n).$$

- High scoring root locations define detections

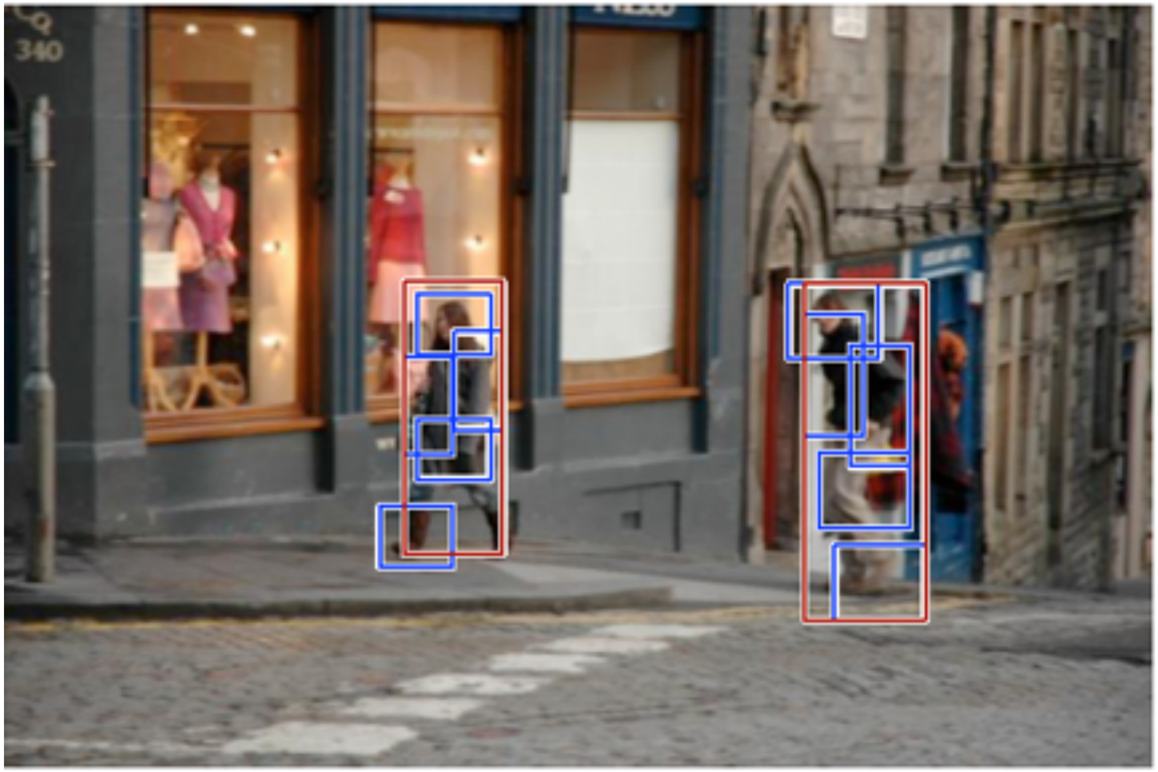
- Efficient computation: dynamic programming + generalized distance transforms

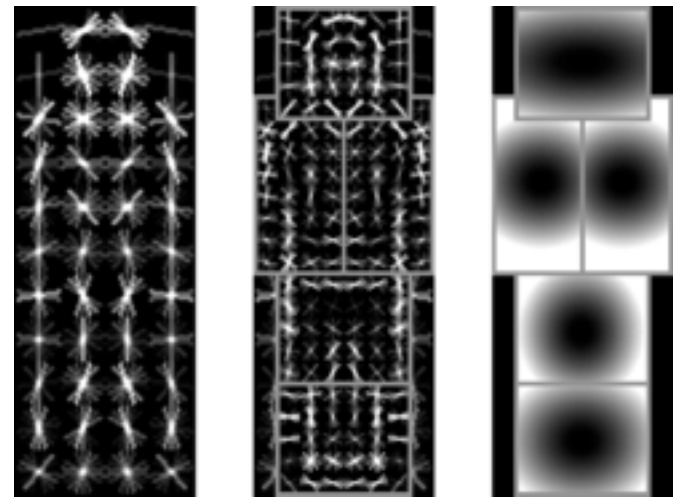
DPM: Detection



Combine Many Parts

DPM: Detection

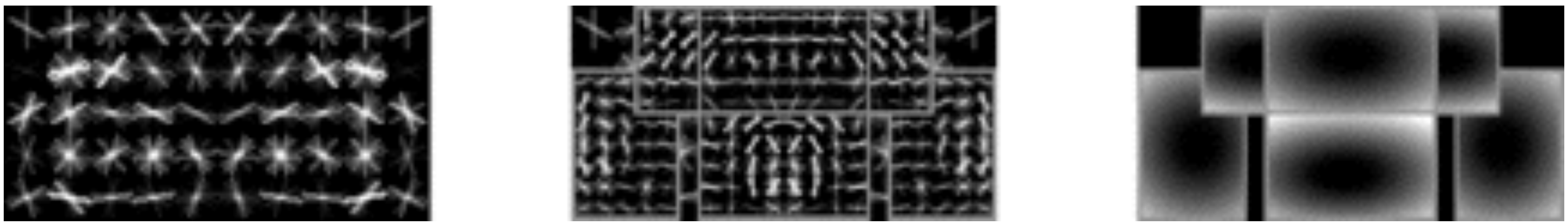




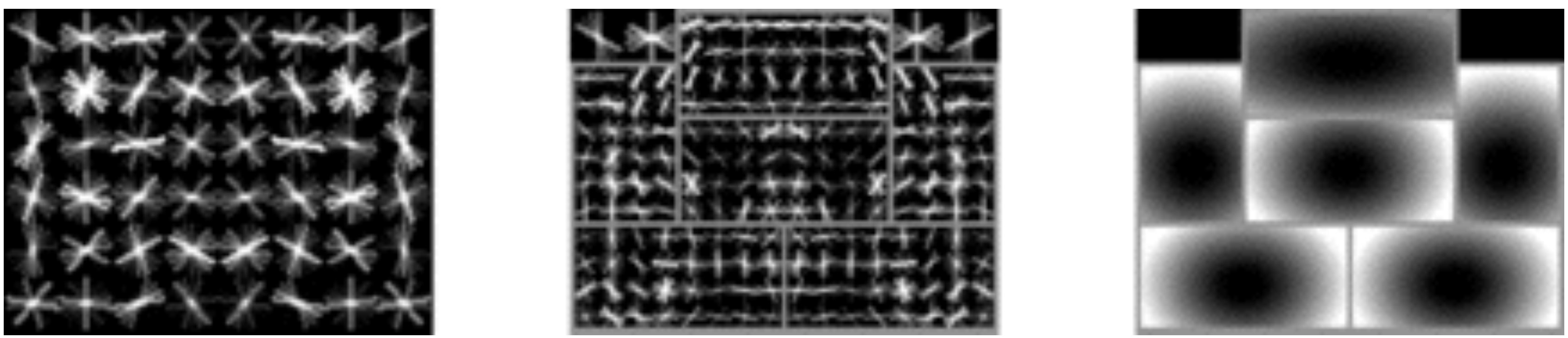
(after non-maximum suppression)
~1 second to search all scales

Car model

Component 1

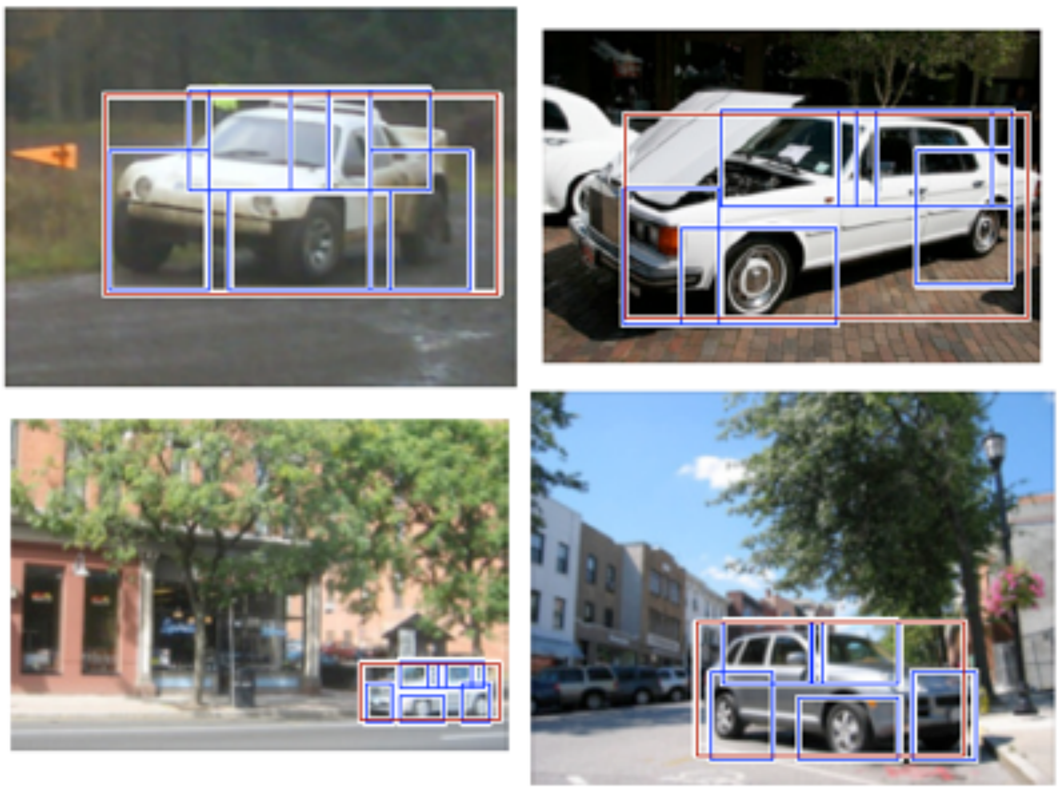


Component 2

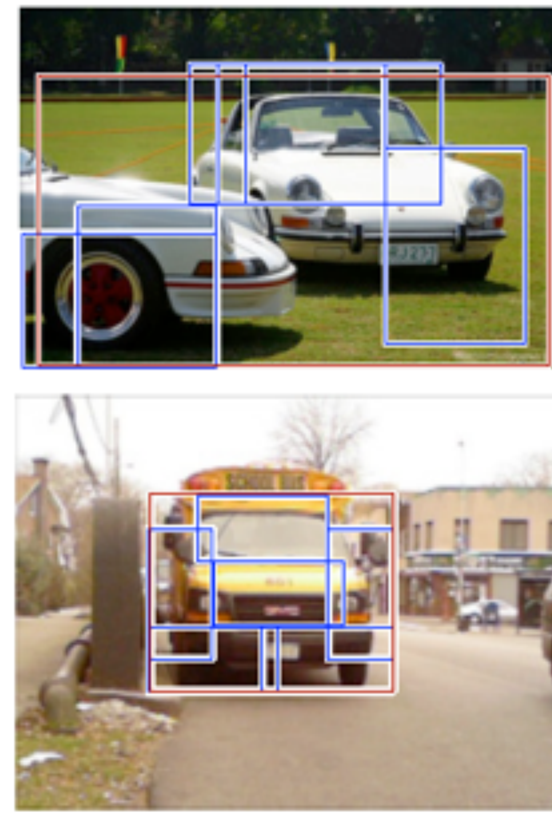


Car detections

high scoring true positives

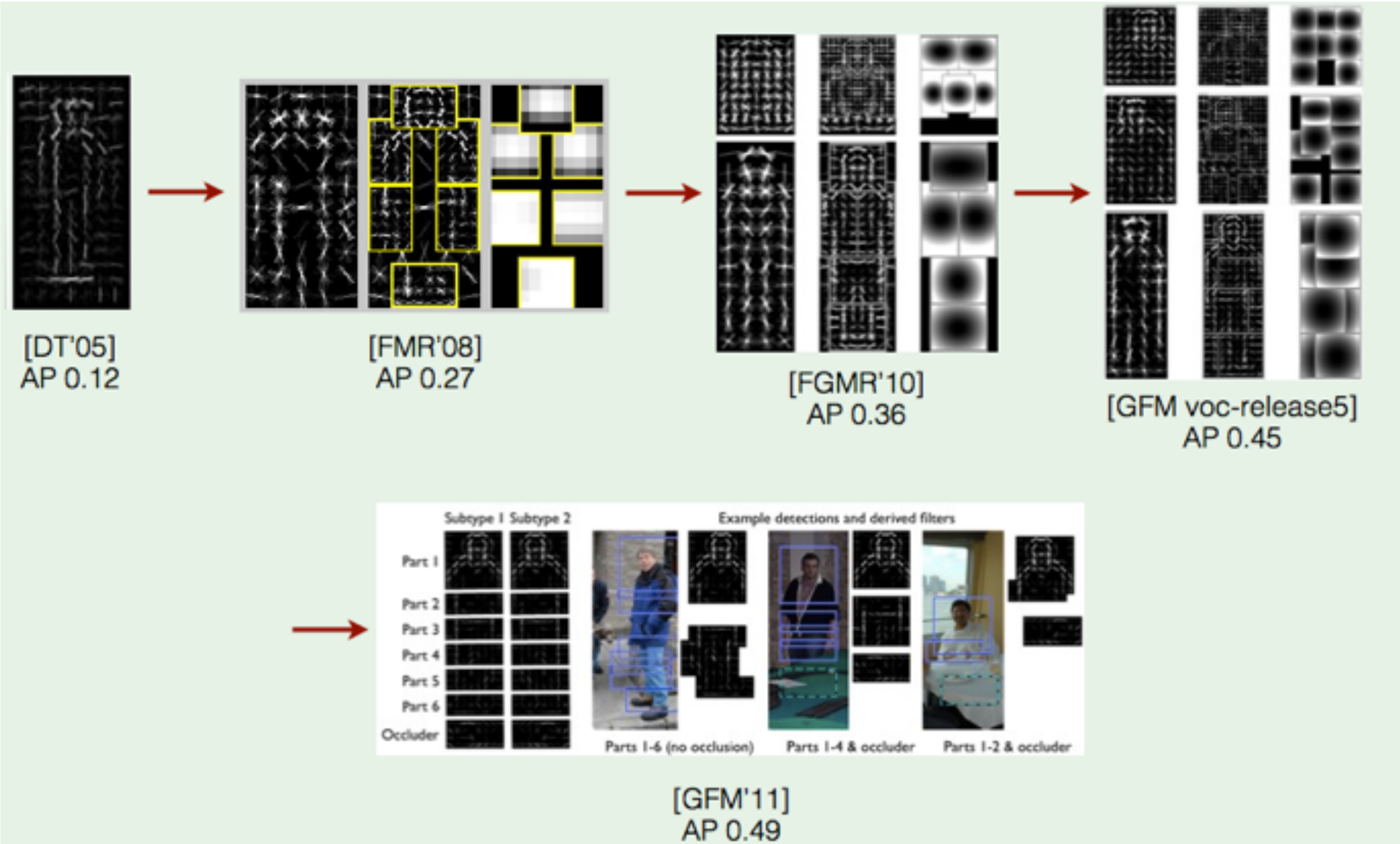


high scoring false positives



More detections

Summary of Results



Poselet

Poselet

Poselets capture part of the pose from a given viewpoint

[Bourdev & Malik, ICCV09]

Poselet

Examples may differ visually but have common semantics

[Bourdev & Malik, ICCV09]

Poselet

One poselet one classifier

not a model for whole human body

How do we train a poselet?

How do we train a poselet?

given pose configuration

How do we train a poselet?

Finding correspondences at training time

Given part of a human pose

How do we find a similar pose configuration in the training set?

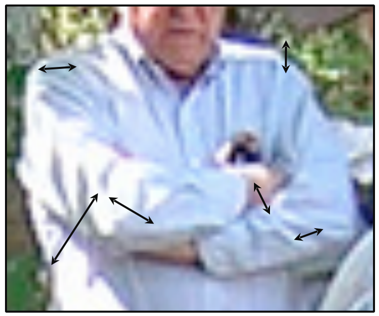
How do we train a poselet?

Finding correspondences at training time

We use key points to annotate the joints, eyes, nose, etc. of people

How do we train a poselet?

Finding correspondences at training time



Residual Error

Training poselet classifiers

Residual Error: 0.15 0.20 0.10 0.85 0.15 0.35

1. Given a seed patch
2. Find the closest patch for every other person
3. Sort them by residual error
4. Threshold them

Training poselet classifiers

1. Given a seed patch
2. Find the closest patch for every other person
3. Sort them by residual error
4. Threshold them
5. Use them as positive training examples to train a linear SVM with HOG features

Training poselet classifiers

One poselet one classifier not a model for whole human body

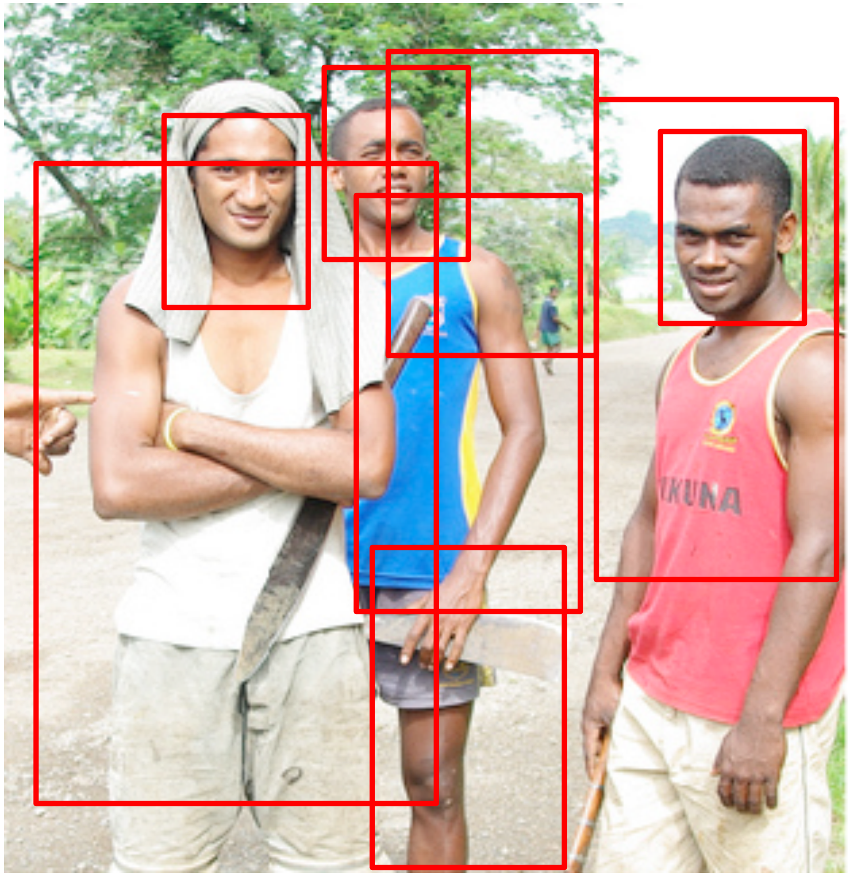
1. Given a seed patch
2. Find the closest patch for every other person
3. Sort them by residual error
4. Threshold them
5. Use them as positive training examples to train a linear SVM with HOG features

Testing

Goal

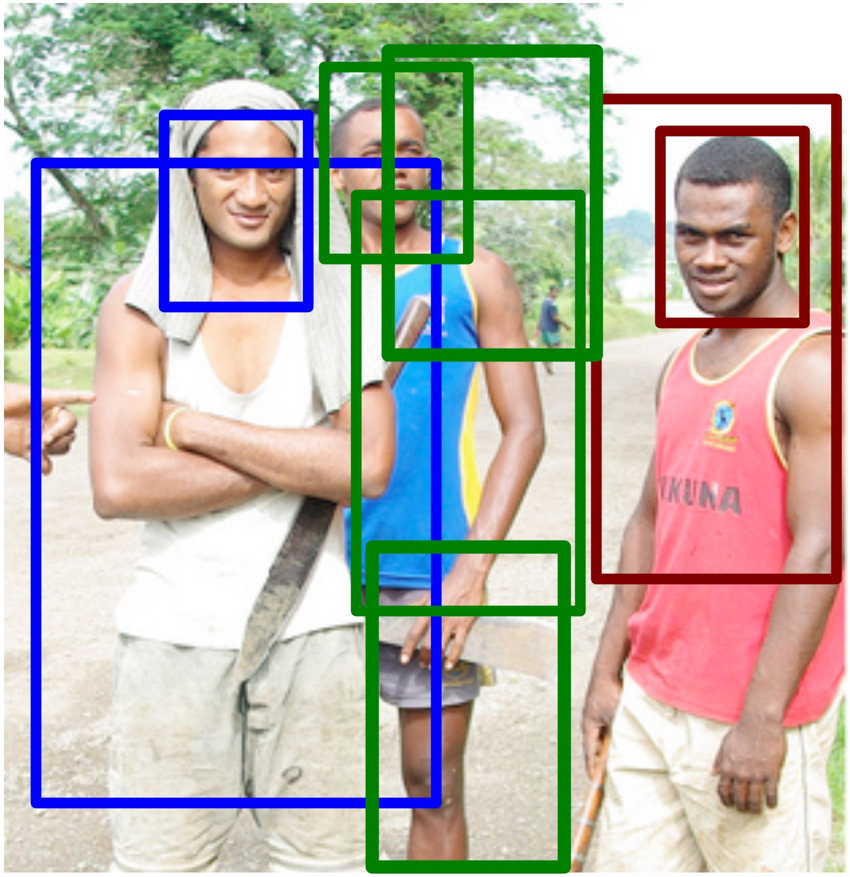
Testing

Step 1: Detect poselet activations



Testing

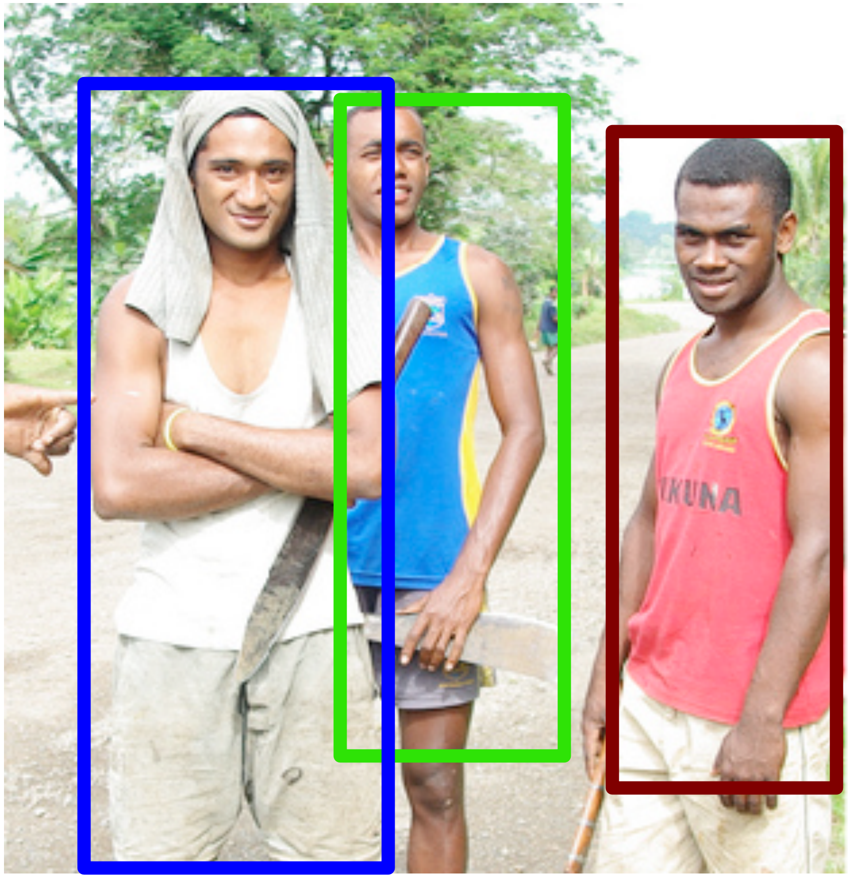
Step 2: Cluster the activations



Because we know the joint for each poselet

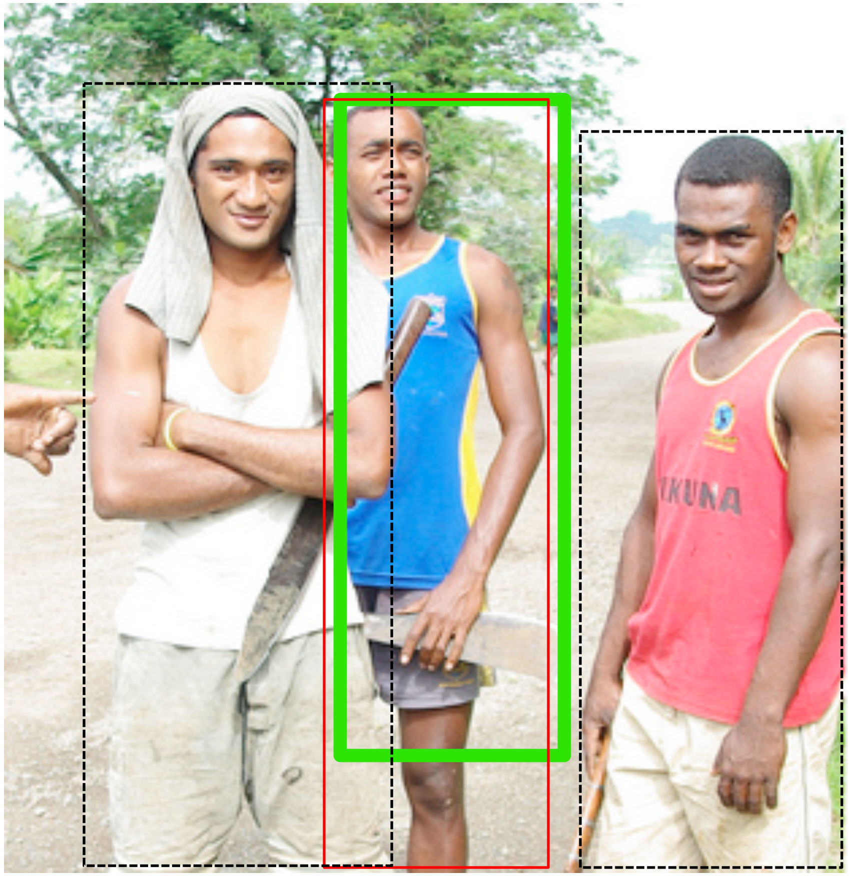
Testing

Step 3: Predict person bounds



Testing

Step 4: Identify the correct cluster



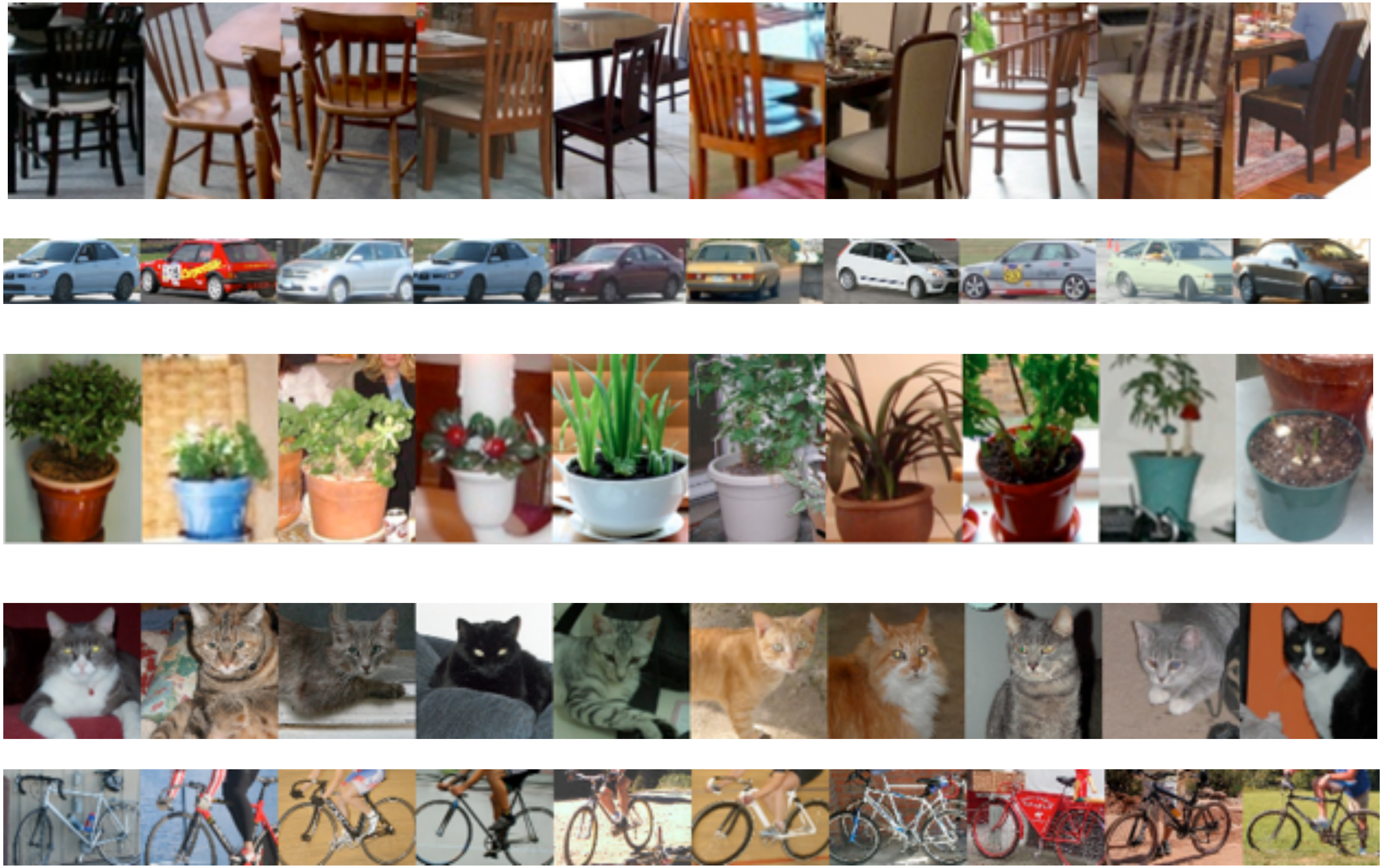
Max-flow in bipartite graph

Detection results on PASCAL

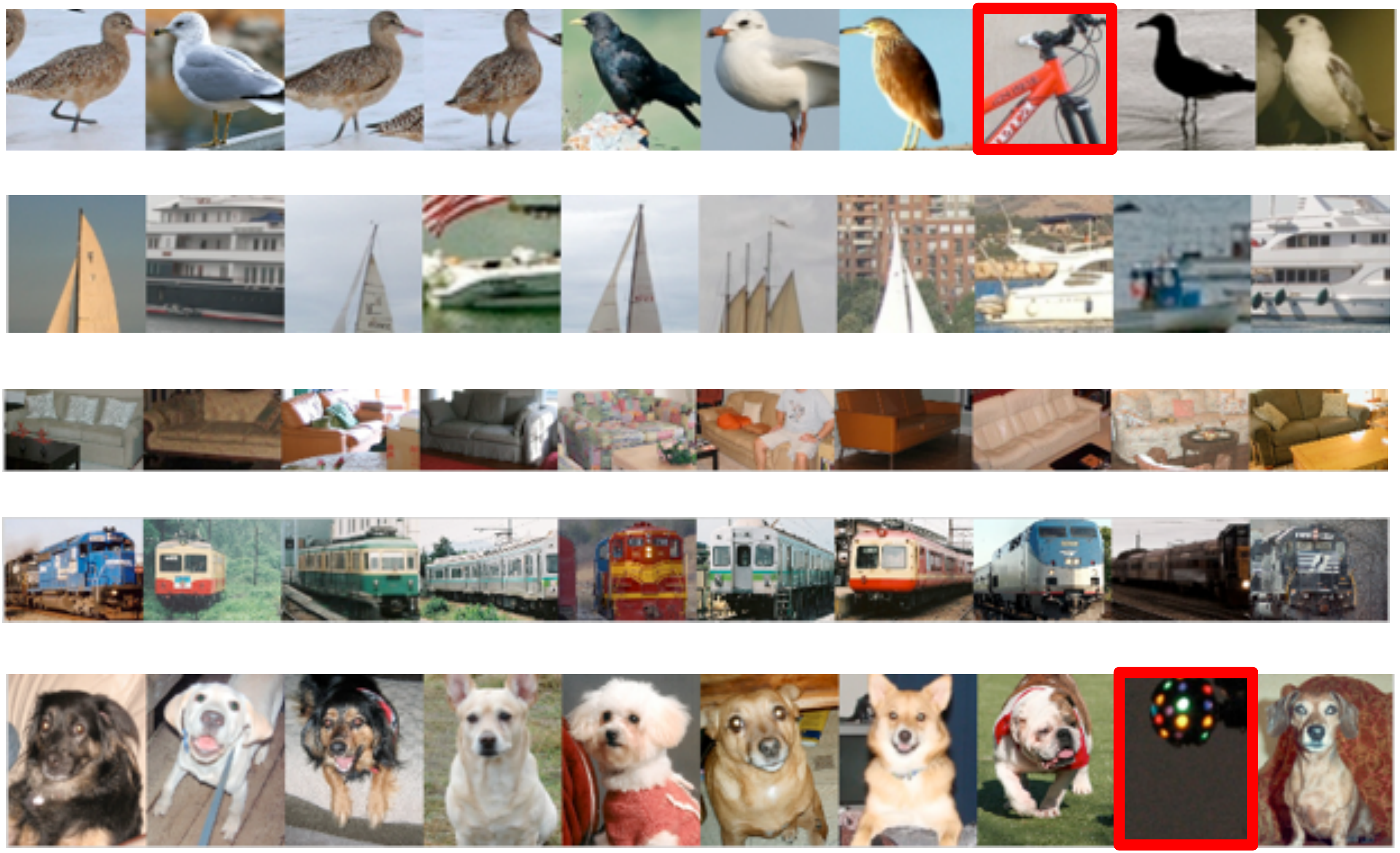
- Person recognition:

	Poselets	DPMs
2010	48.5%	47.7%
2009	48.3%	47.4%
2008	54.1%	43.1%
2007	46.9%	43.2%

Highest scoring hits on PASCAL test set



Highest scoring hits on PASCAL test set



Roadmap (this lecture)

- Part Based Detector (cont. last lecture)
  - Deformable Part Model
  - Poselets
- Scene Understanding Problem
- Context
- Spatial Layout
- 3D Scene Understanding

Scene Understanding

- **What is goal of scene understanding:**
  - Build machine that can see like humans to automatically interpret the content of the images

- **Comparing with traditional vision problem:**
  - Study on larger scale
  - Human vision related tasks

Larger Scale

More image information.
Context information.

Human vision related task

More similar as the way that human understand the image
Infer more useful information from image

How DO human learn?

- Bayesian Rules:

$$P(A|B) = P(B|A) \cdot P(A) / P(B)$$

- In practice: Infer abstract knowledge based on observation

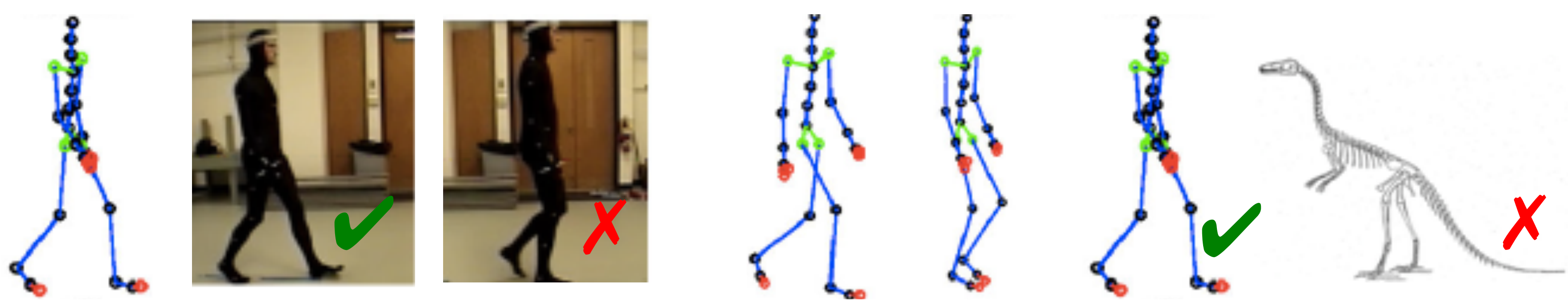
$$P(W|I) = P(I|W) \cdot P(W) / P(I)$$

$$\propto P(I|W) \cdot P(W)$$

Posterior probability

Likelihood: The probability of getting I given model W

Prior: The probability of W w/o seeing any observation



How DO human learn?

- To teach human baby what is "horse": show 3 pictures and let them learn by themselves.
- They can be very successful to learn the correct concept.

- But all the following concepts can explain the images:
  - "horse" = all horse
  - "horse" = all horse but not Clydesdales
  - "horse" = all animal

"horse"

$I =$

Roadmap (this lecture)

- Part Based Detector (cont. last lecture)
  - Deformable Part Model
  - Poselets
- Scene Understanding Problem
- Context
- Spatial Layout
- 3D Scene Understanding

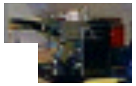
- Objects usually are surrounded by a scene that can provide context in the form of nearby objects, surfaces, scene category, geometry, etc.

Contextual Reasoning

- Definition: Making a decision based on more than *local* image evidence.

Context provides clues for function

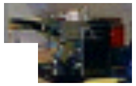
- What is this?



Context provides clues for function

- What is this?

- Now can you tell?



Context provides clues for function

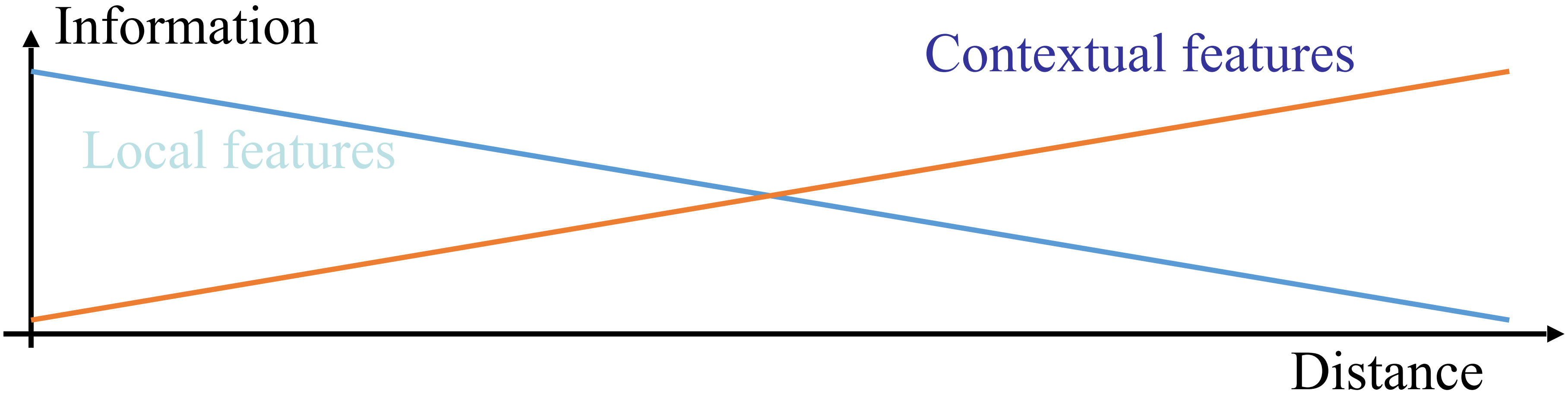
- once more how amazing is the visual system

Context provides clues for function

- once more how amazing is the visual system

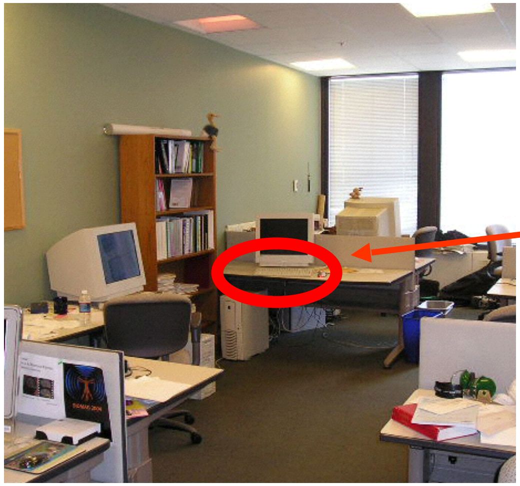
Is local information enough?

Is local information enough?



Context in Recognition

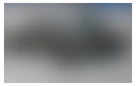
We know there is a keyboard present in this scene even if we cannot see it clearly.

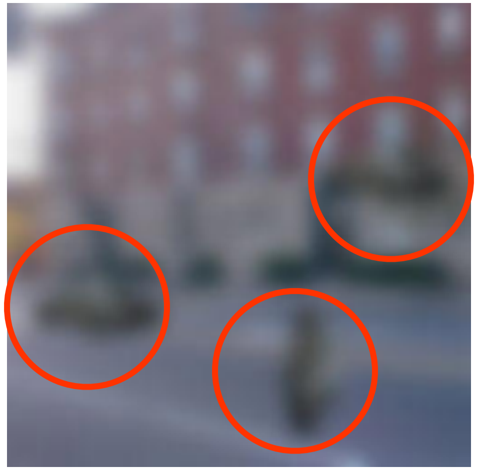


We know there is no keyboard present in this scene

… even if there is one indeed.

Context in Recognition



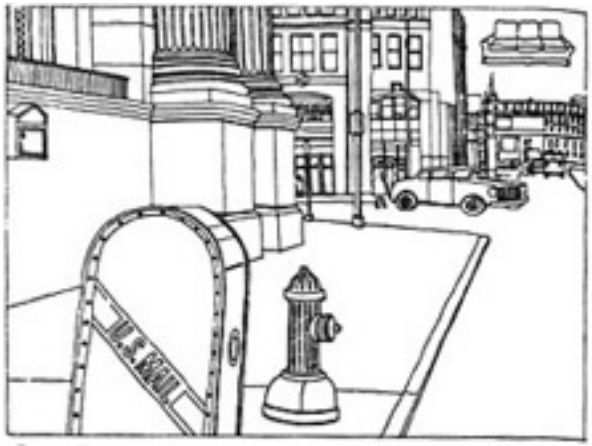


Context in Recognition

Look-Alikes by Joan Steiner

Context in Recognition

- Pictures shown for 150 ms
- Objects in appropriate context were detected more accurately than objects in an inappropriate context
- Scene consistency affects object detection



Biederman 1982

Why is context important?

- Changes the interpretation of an object (or its function)

- Context defines what an unexpected event is

There are many types of context

- **Local pixels**
  - window, surround, image neighborhood, object boundary/shape, global image statistics
- **2D Scene Gist**
  - global image statistics
- **3D Geometric**
  - 3D scene layout, support surface, surface orientations, occlusions, contact points, etc.
- **Semantic**
  - event/activity depicted, scene category, objects present in the scene and their spatial extents, keywords
- **Photogrammetric**
  - camera height orientation, focal length, lens distortion, radiometric, response function
- **Illumination**
  - sun direction, sky color, cloud cover, shadow contrast, etc.
- **Geographic**
  - GPS location, terrain type, land use category, elevation, population density, etc.
- **Temporal**
  - nearby frames of video, photos taken at similar times, videos of similar scenes, time of capture
- **Cultural**
  - photographer bias, dataset selection bias, visual cliches, etc.

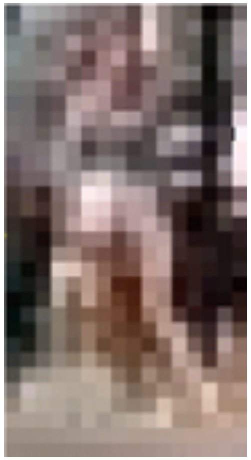
from Divvala et al. CVPR 2009

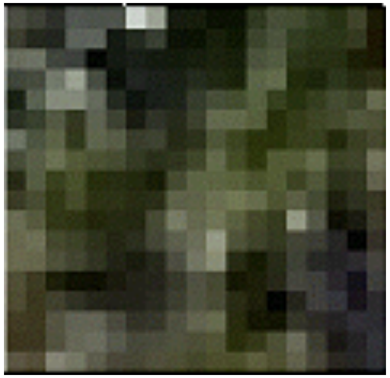
Roadmap (this lecture)

- Part Based Detector (cont. last lecture)
  - Deformable Part Model
  - Poselets
- Scene Understanding Problem
- Context
- Spatial Layout
- 3D Scene Understanding

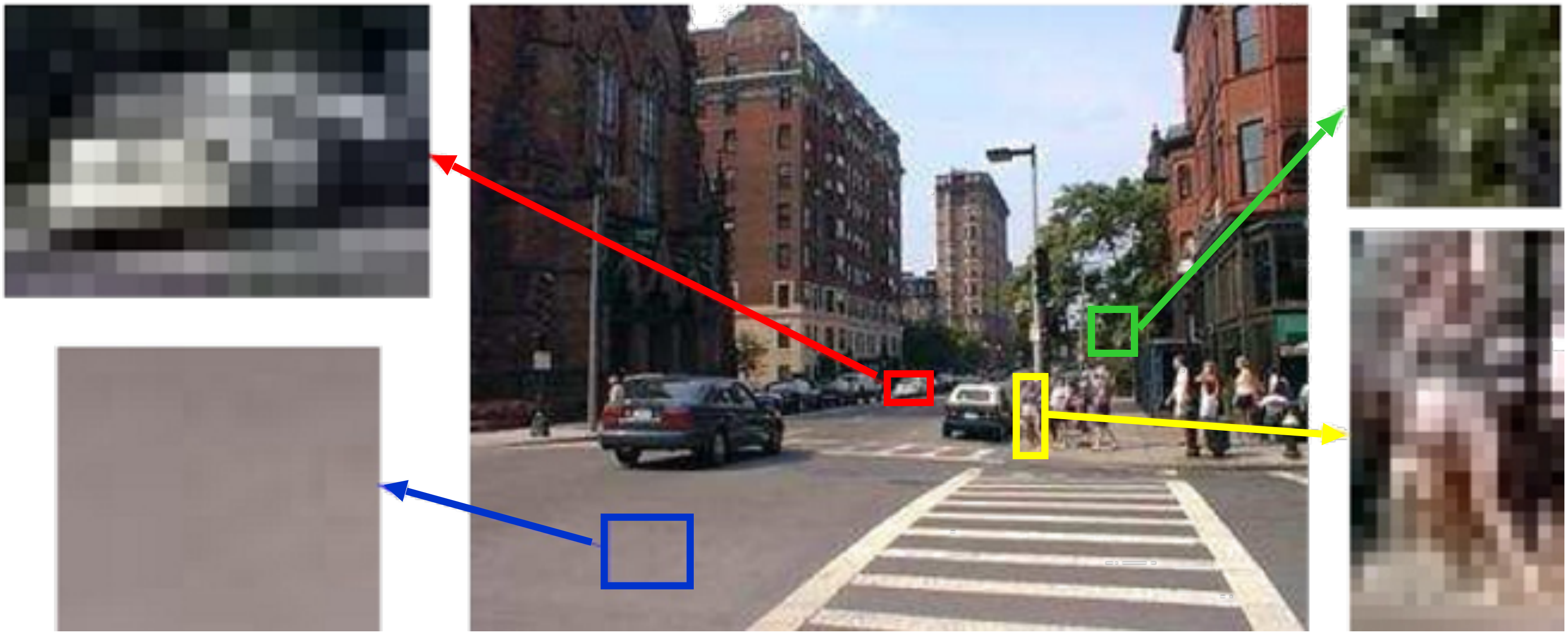
Spatial layout is especially important

1. Context for recognition





1. Context for recognition

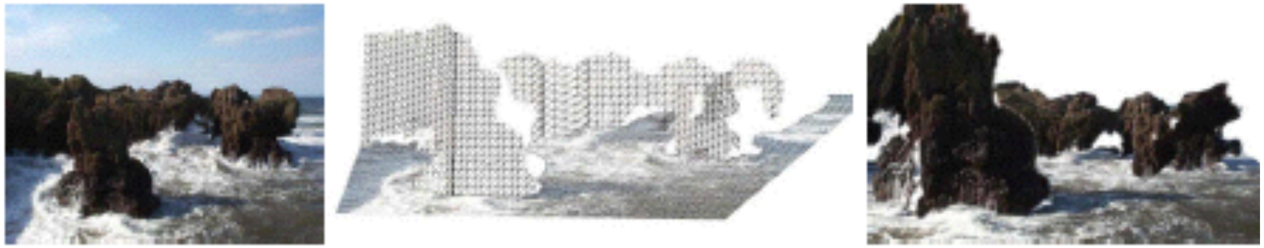


Spatial layout is especially important

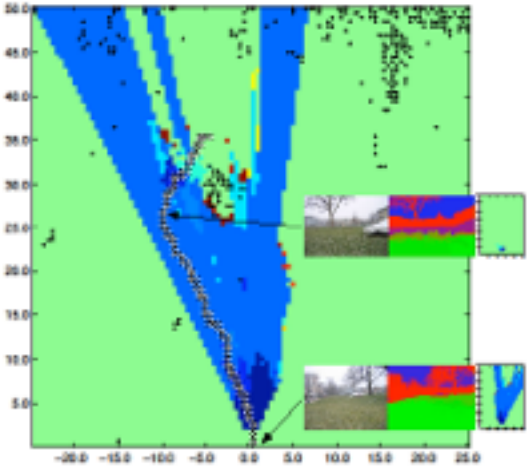
1. Context for recognition
2. Scene understanding

Spatial layout is especially important

1. Context for recognition
2. Scene understanding
3. Many direct applications
    a) Assisted driving
    b) Robot navigation/interaction
    c) 2D to 3D conversion for 3D TV
    d) Object insertion



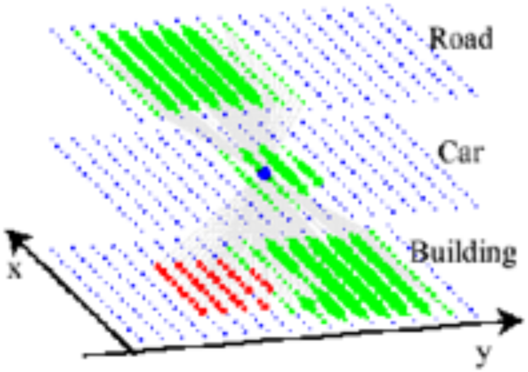
3D Reconstruction: Input, Mesh, Novel View



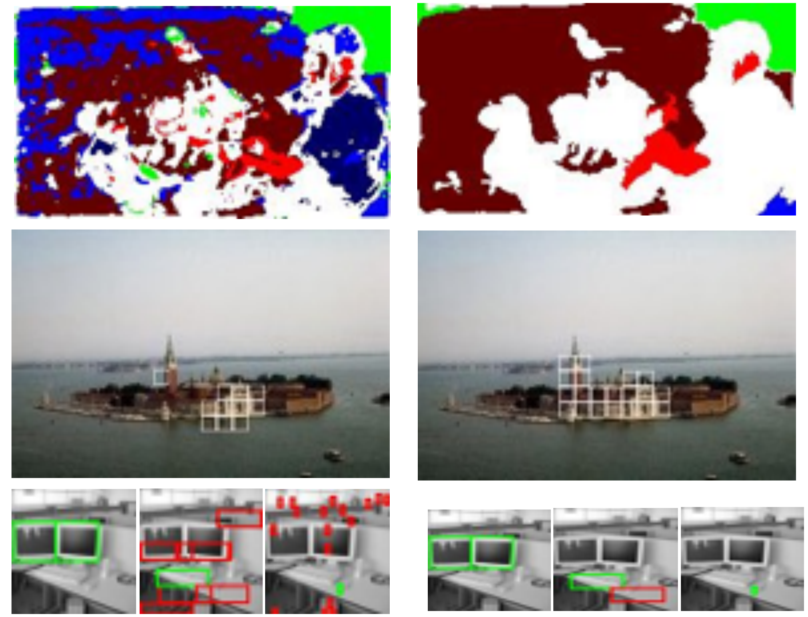
Robot Navigation: Path Planning

Spatial Layout: 2D vs. 3D

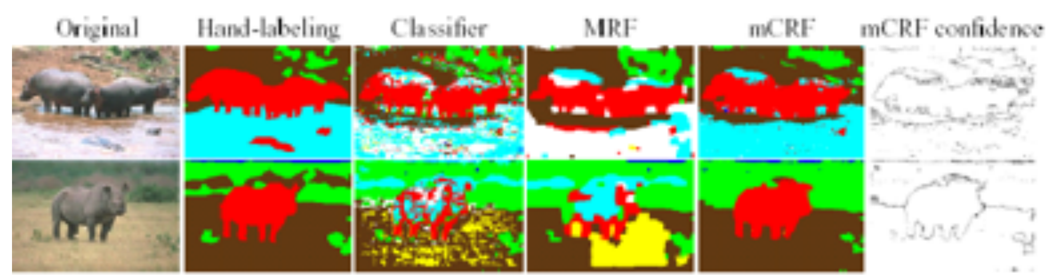
Context in Image Space



[Torralba Murphy Freeman 2004]

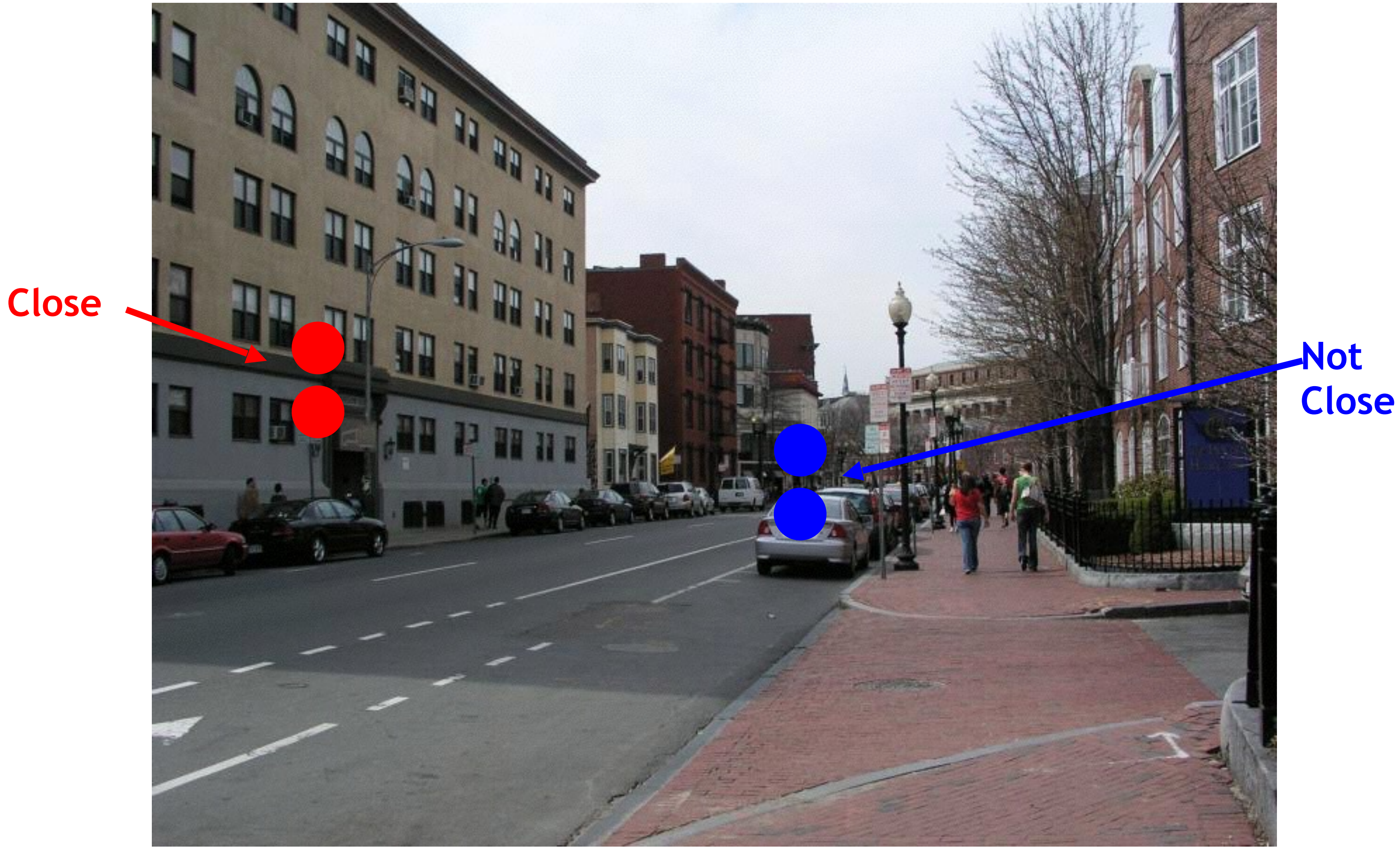


[Kumar Hebert 2005]



[He Zemel Cerreira-Perpiñán 2004]

But object relations are in 3D…



How to represent scene space?

Wide variety of possible representations

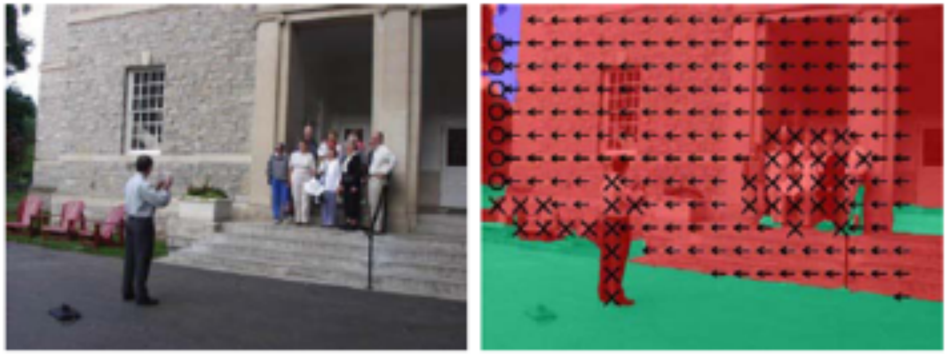
Scene-Level Geometric Description

a) Gist, Spatial Envelope

b) Stages

Wide variety of possible representations

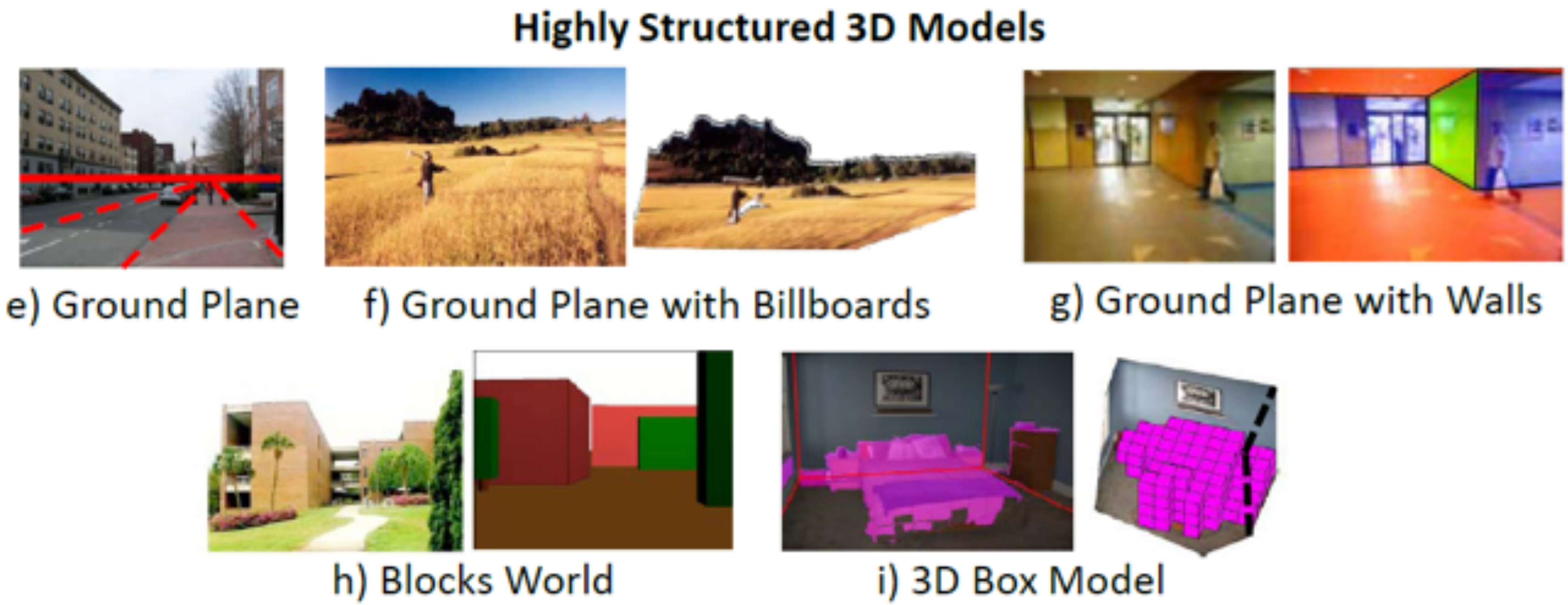
Retinotopic Maps



c) Geometric Context

d) Depth Maps

Wide variety of possible representations



Key Trade-offs

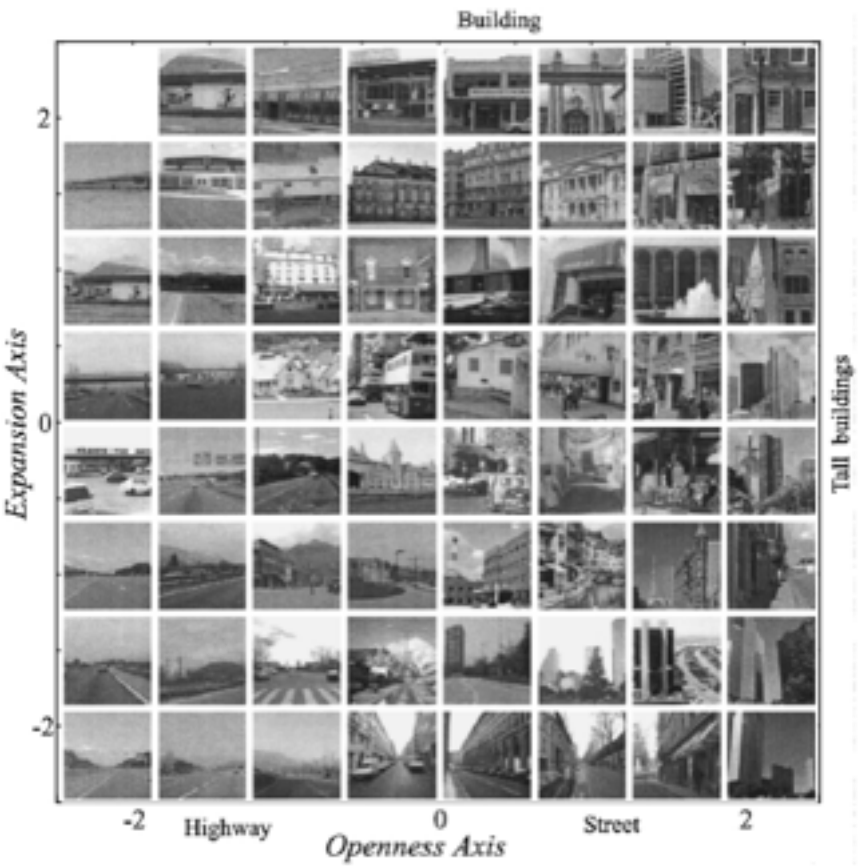
- Level of detail: rough “gist”, or detailed point cloud?
  - Precision vs. accuracy
  - Difficulty of inference

- Abstraction: depth at each pixel, or ground planes and walls?
  - What is it for: e.g., metric reconstruction vs. navigation

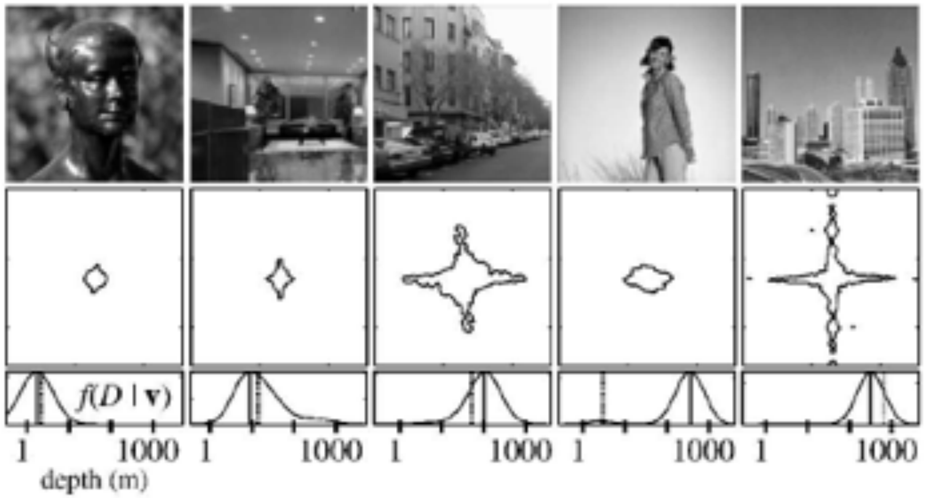
Slide: Derek Hoiem

Low detail, Low abstraction

Holistic Scene Space: "Gist"



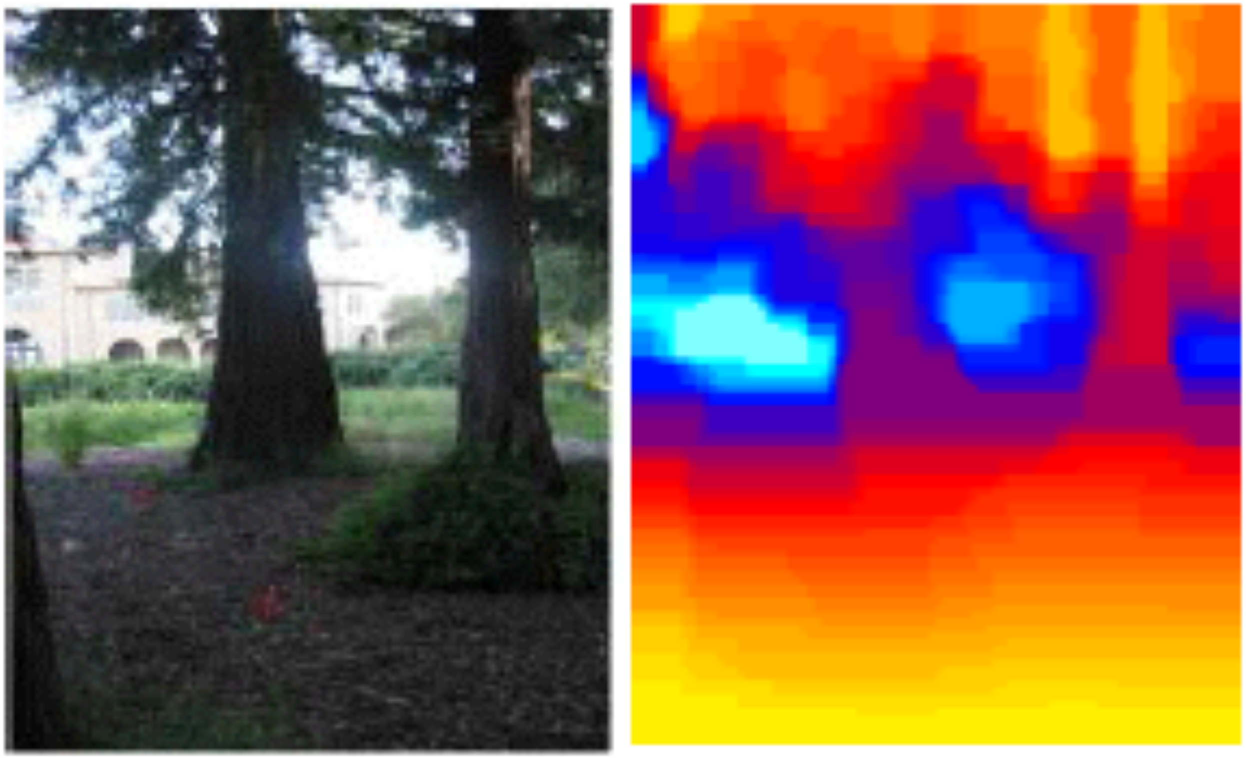
Oliva & Torralba 2001



Torralba & Oliva 2002

High detail, Low abstraction

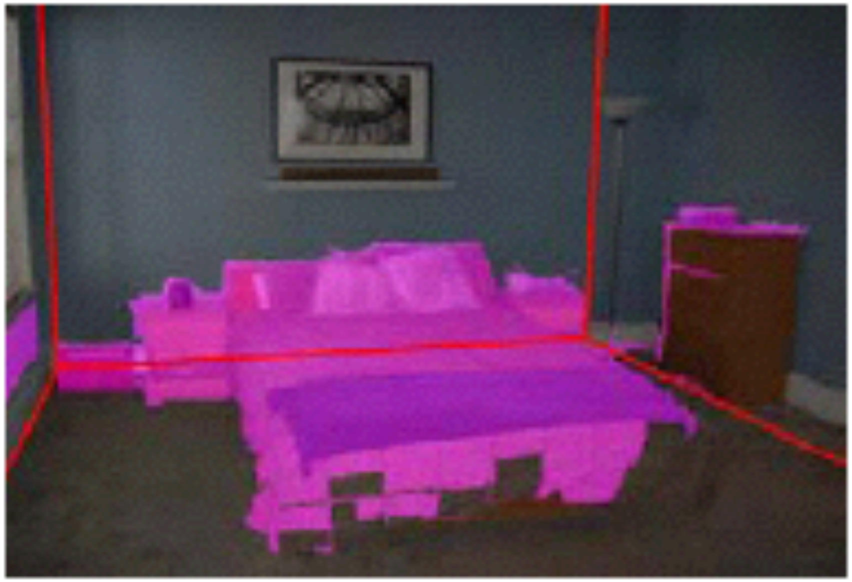
Depth Map

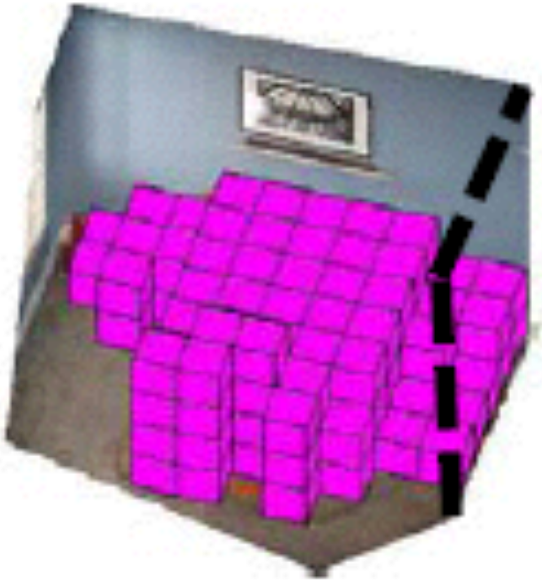


Saxena, Chung & Ng 2005, 2007

Medium detail, High abstraction

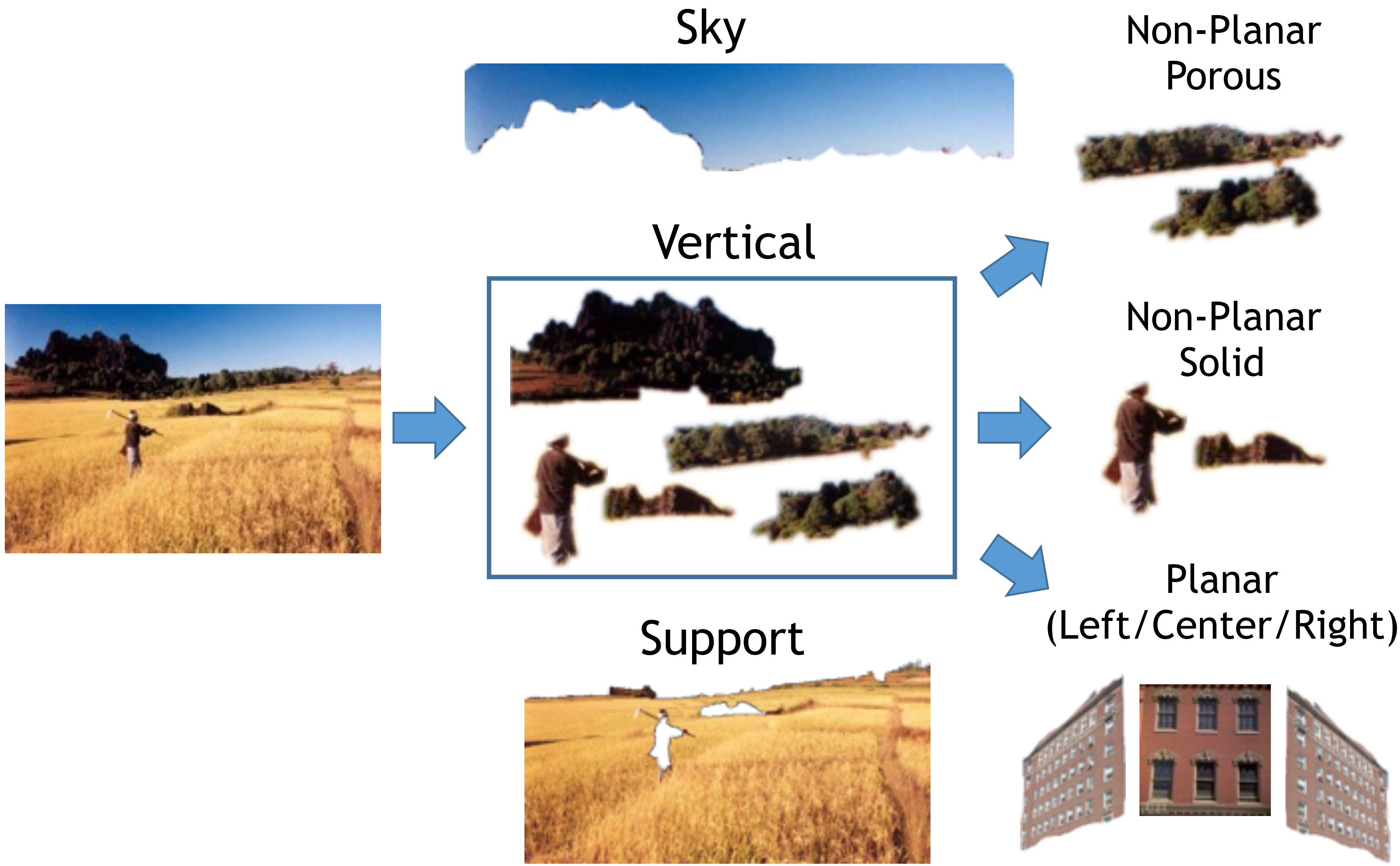
Room as a Box





[Hedau Hoiem Forsyth 2009]

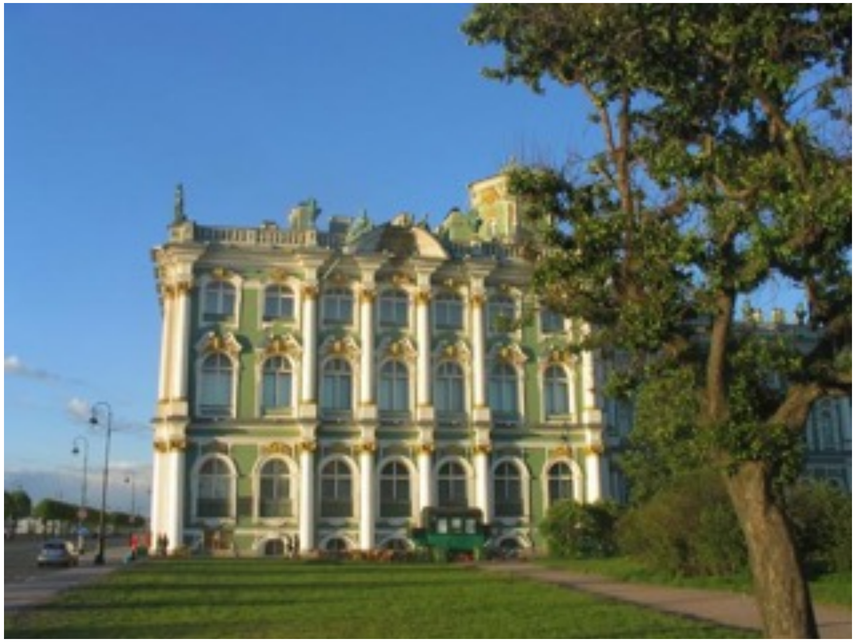
Surface Layout



The challenge

Our World is Structured

Abstract World

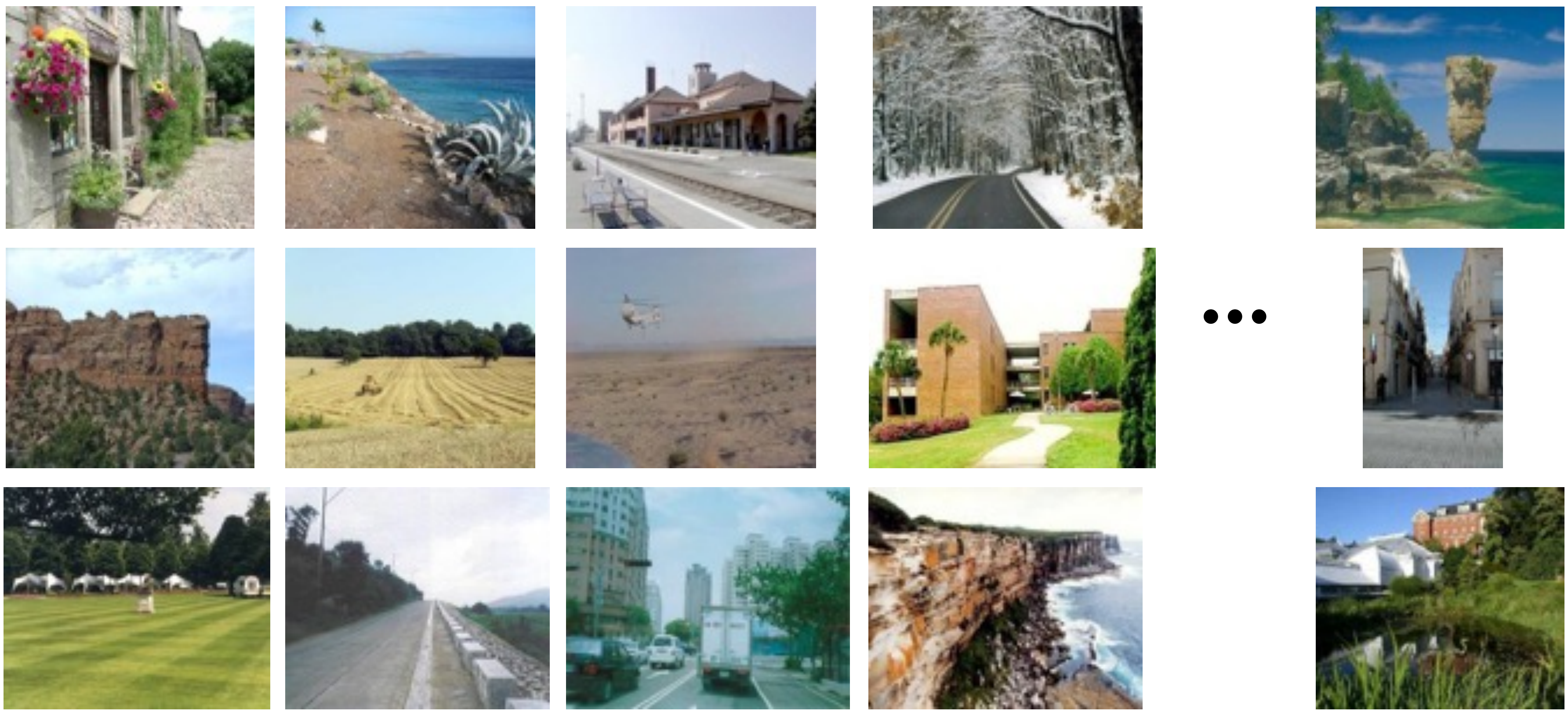


Our World

Image Credit (left): F. Cunin
and M.J. Sailor, UCSD

Learn the Structure of the World

Training Images



Infer the most likely interpretation

Unlikely

Likely

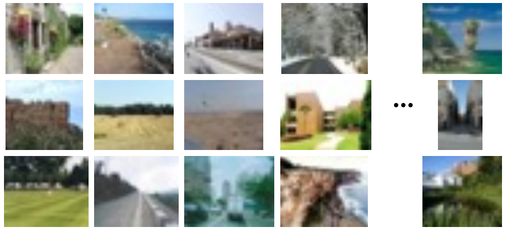
Geometry estimation as recognition

Region

Features
Color
Texture
Perspective
Position

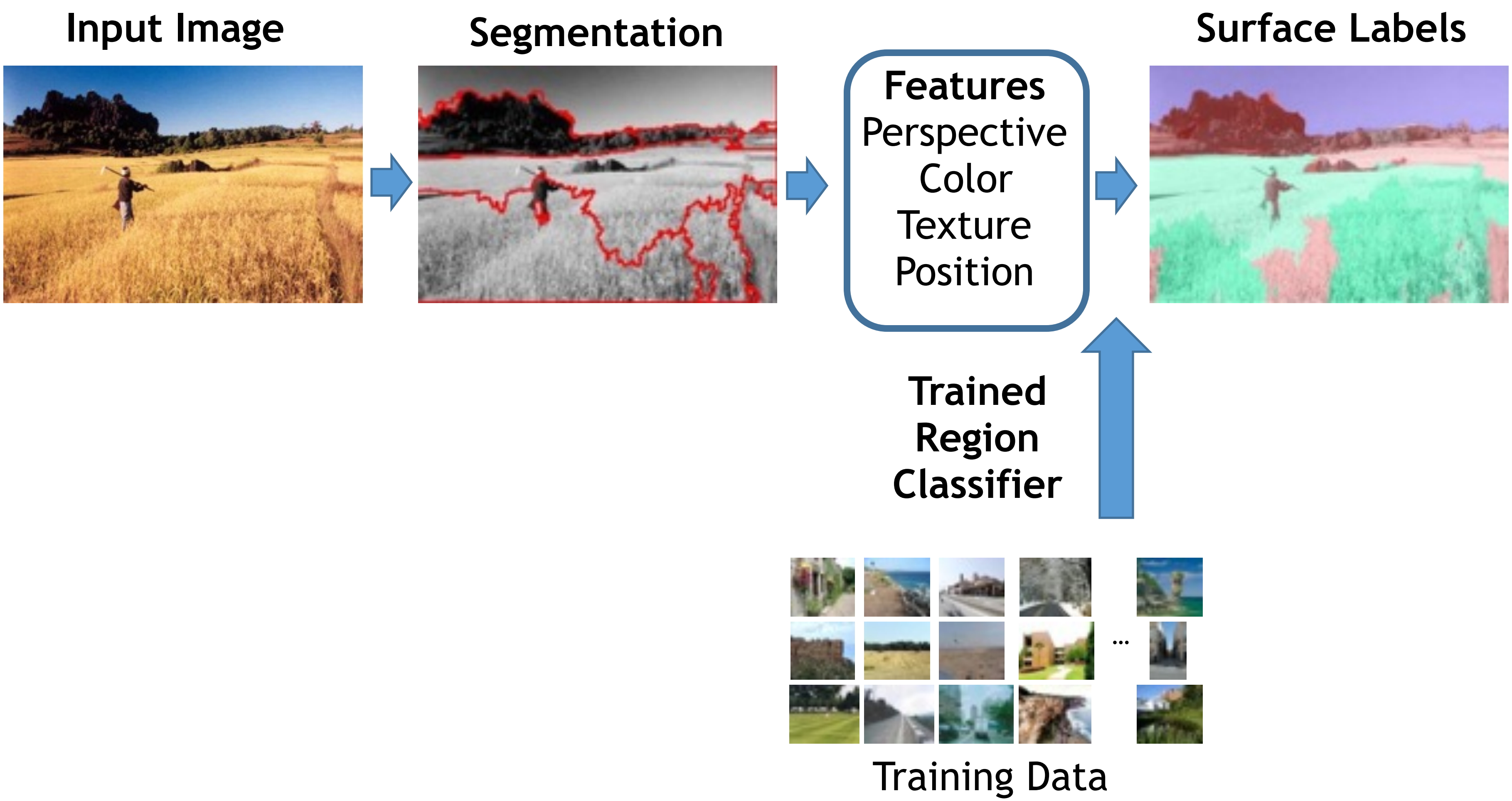
Surface Geometry Classifier

Vertical, Planar



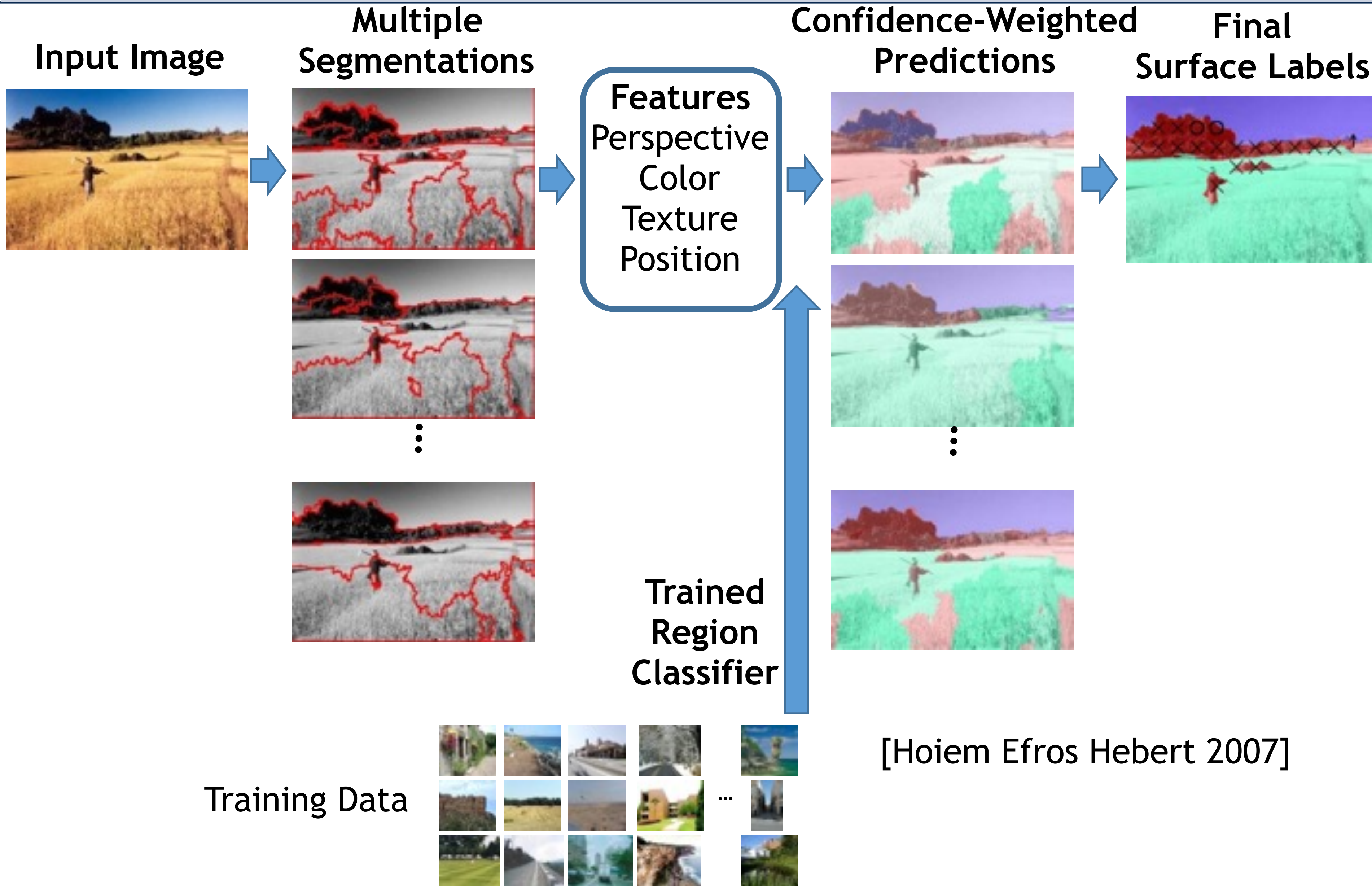
Training Data

Surface Layout Algorithm



[Hoiem Efros Hebert 2007]

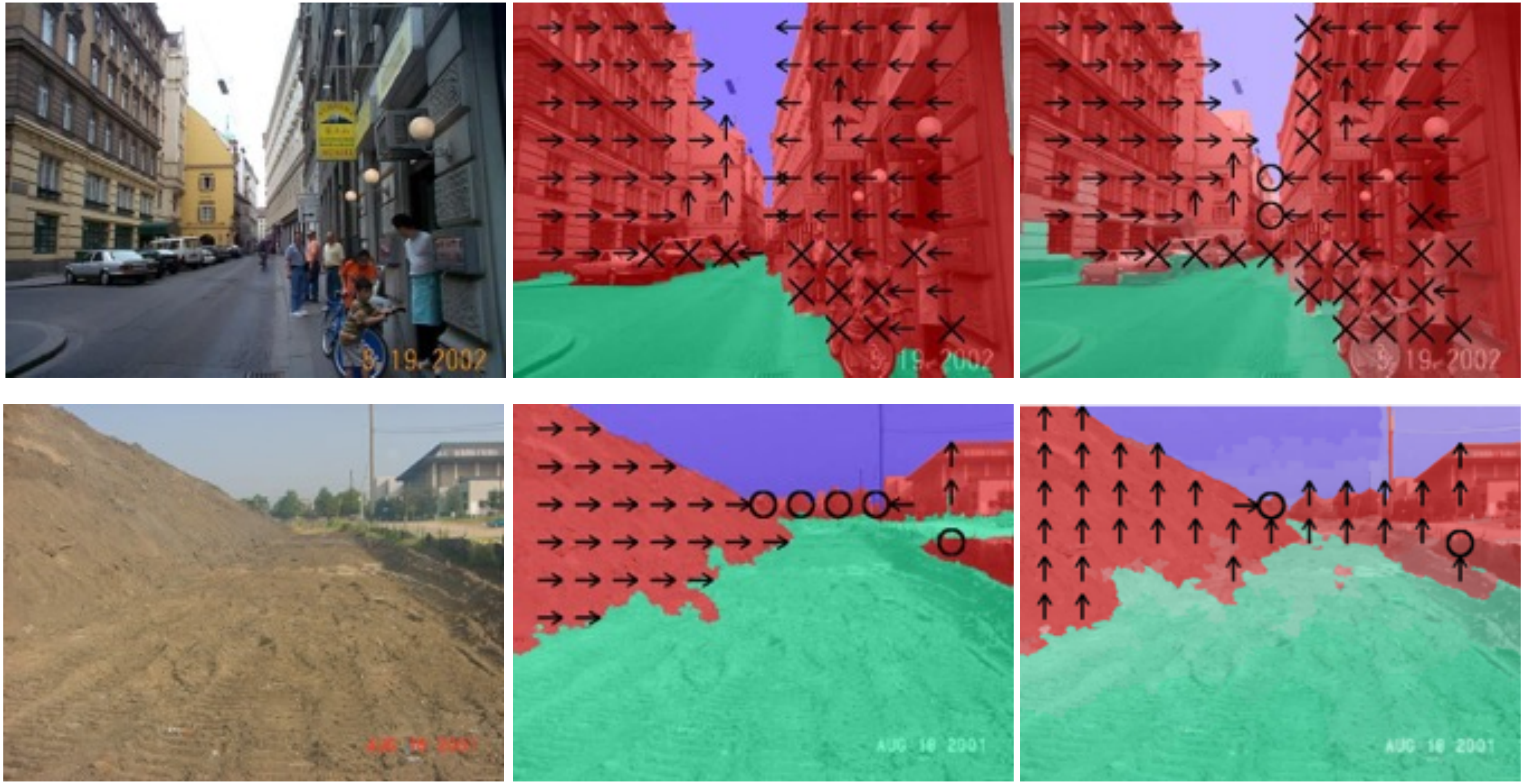
Surface Layout Algorithm



[Hoiem Efros Hebert 2007]

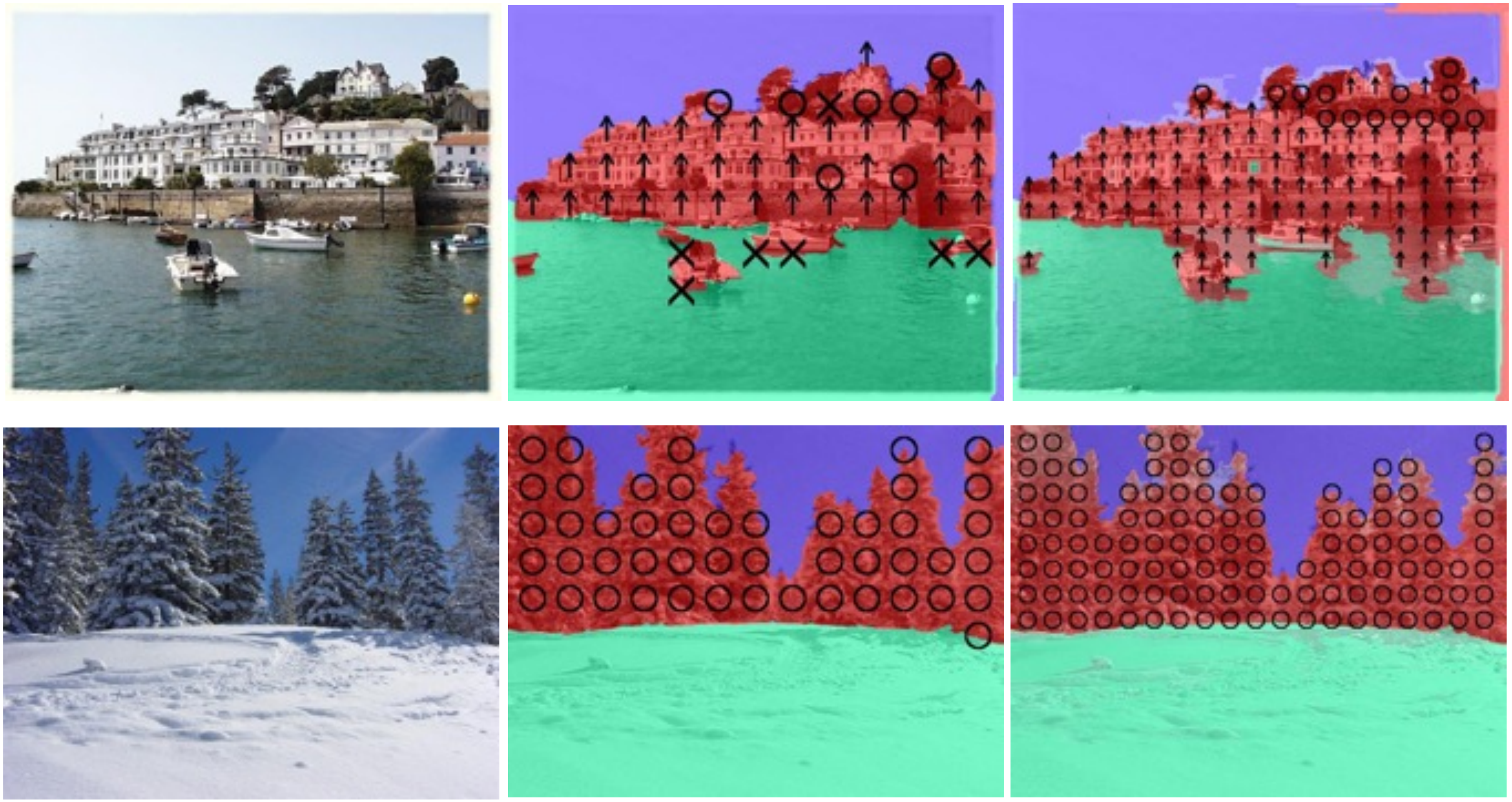
Surface Description Result

Results



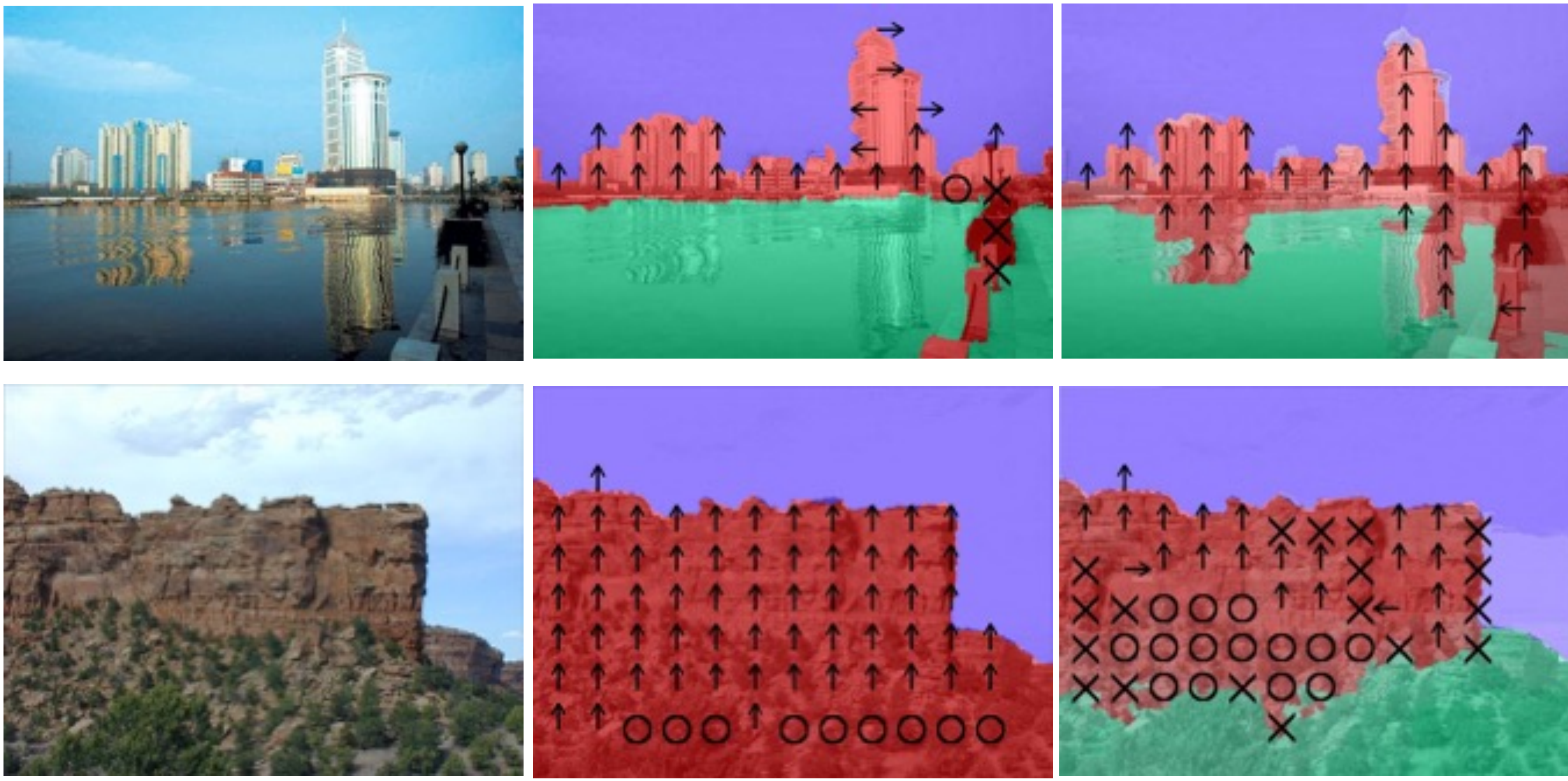
Input Image | Ground Truth | Result

Results



Input Image Ground Truth Result

Failures: Reflections, Rare Viewpoint



Input Image | Ground Truth | Result

Average Accuracy

Main Class: 88%

Subclasses: 61%

Main Class			
	Support	Vertical	Sky
Support	**0.84**	0.15	0.00
Vertical	0.09	**0.90**	0.02
Sky	0.00	0.10	**0.90**

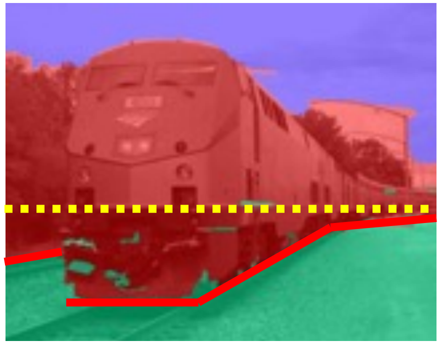
Vertical Subclass					
	Left	Center	Right	Porous	Solid
Left	**0.37**	0.32	0.08	0.09	0.13
Center	0.05	**0.56**	0.12	0.16	0.12
Right	0.02	0.28	**0.47**	0.13	0.10
Porous	0.01	0.07	0.03	**0.84**	0.06
Solid	0.04	0.20	0.04	0.17	**0.55**

Automatic Photo Popup

Labeled Image

Fit Ground-Vertical Boundary with Line Segments

Form Segments into Polylines



Cut and Fold

Final Pop-up Model

[Hoiem Efros Hebert 2005]

Mini-conclusions

- Can learn to predict surface geometry from a single image
- Very rough models, much room for improvement

Things to remember

- Objects should be interpreted in the context of the surrounding scene
  - Many types of context to consider
- Spatial layout is an important part of scene interpretation, but many open problems
  - How to represent space?
  - How to learn and infer spatial models?
- Consider trade-offs of detail vs. accuracy and abstraction vs. quantification

Roadmap (this lecture)

- Part Based Detector (cont. last lecture)
  - Deformable Part Model
  - Poselets
- Scene Understanding Problem
- Context
- Spatial Layout
- 3D Scene Understanding

Complete Scene Understanding

Involves

- Localization of all instances of foreground objects (“things”)
- Localization of all background classes (“stuff”)
- Pixel-wise segmentation
- 3D reconstruction
- Pose detection
- Action recognition
- Event recognition
- …..

KITTI (video)

3D Traffic Scene Understanding from Movable Platforms

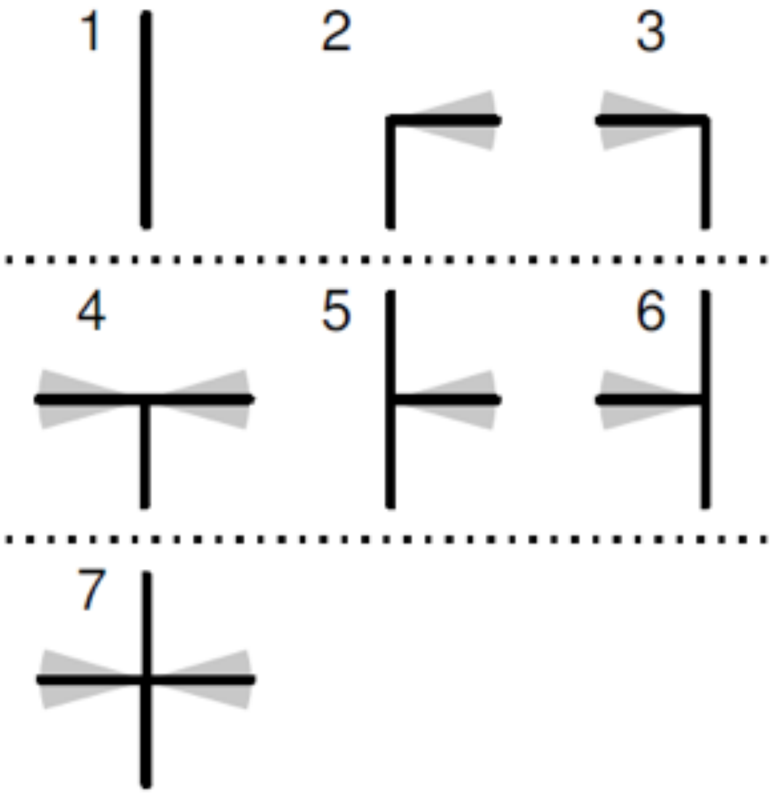
Andreas Geiger

3D Traffic Scene Understanding

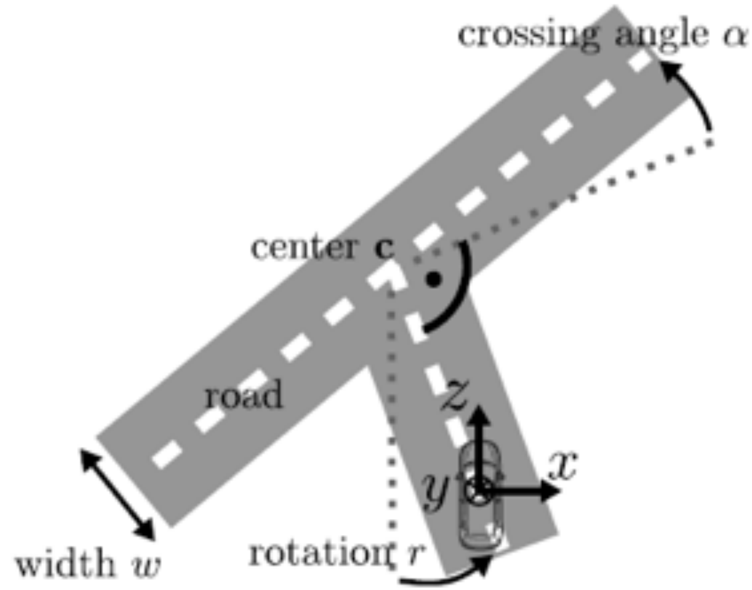
- Goal: Infer from short video sequences (moving observer)
    - Topology and geometry of the scene
    - Semantic information (traffic situation)

- **Probabilistic generative model of 3D urban scenes**

Topology and Geometry Model



Topology Model (κ)



Geometry Model $(\mathbf{c}, w, r, \alpha)$

Road Layout $\mathcal{R} = \{\kappa, \mathbf{c}, w, r, \alpha\}$

Image Evidence

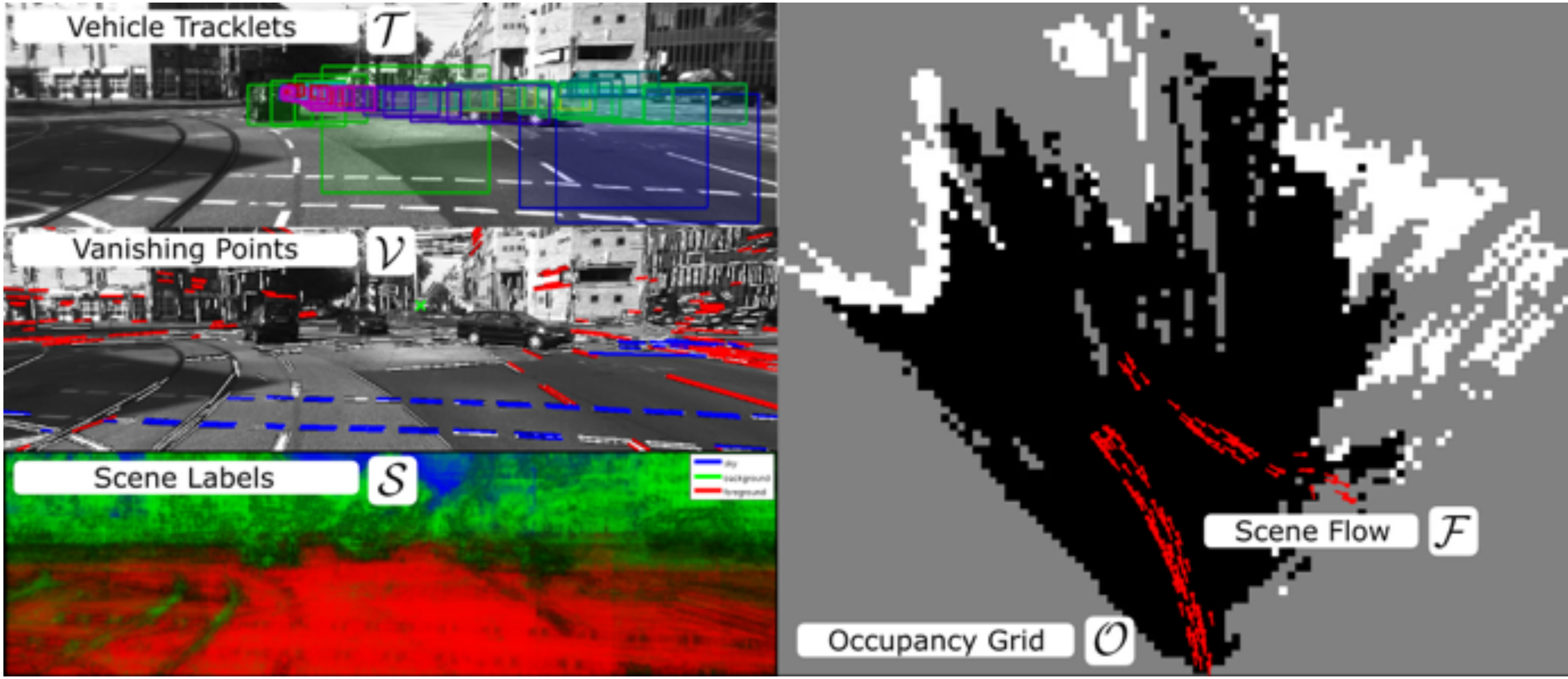
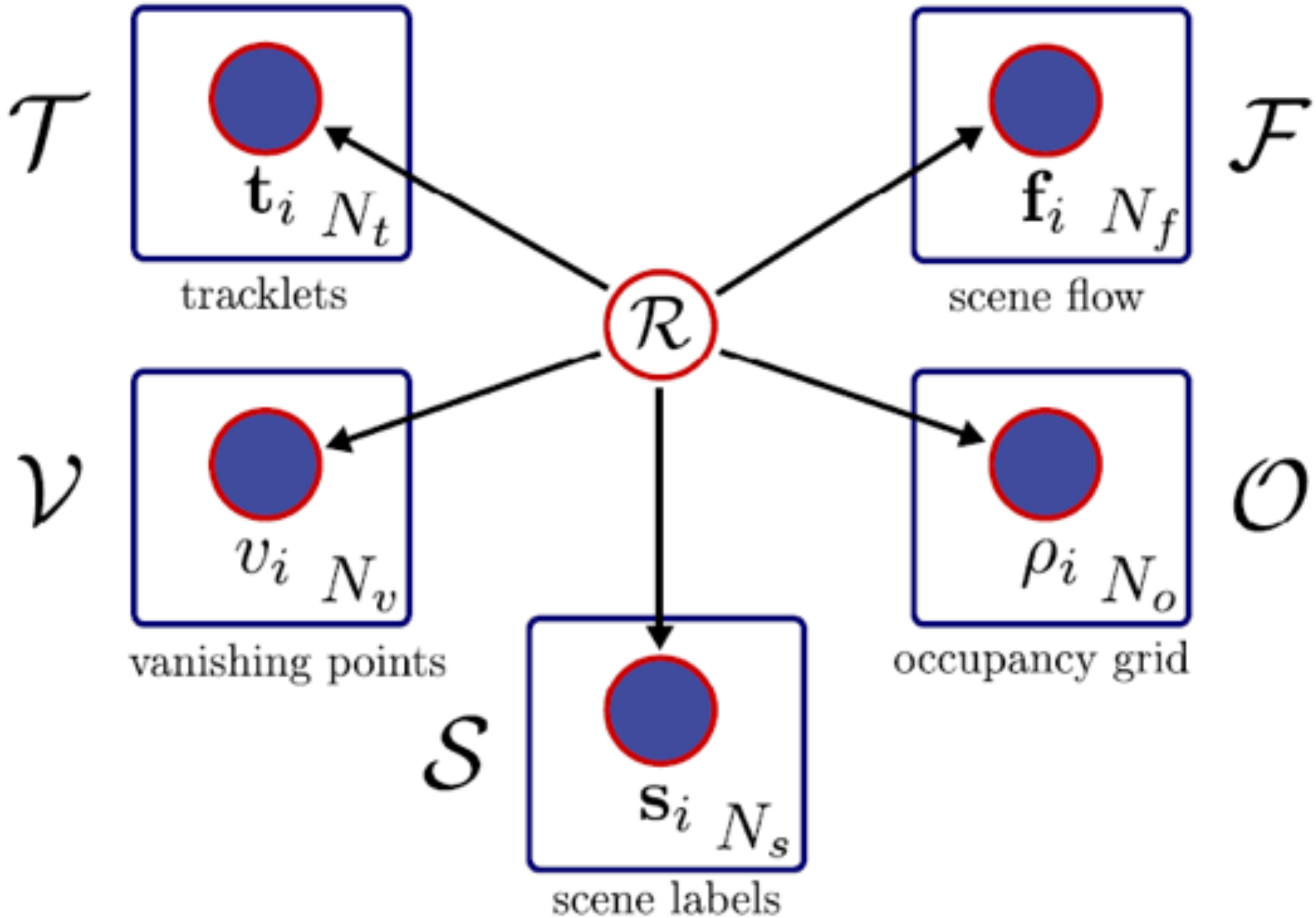


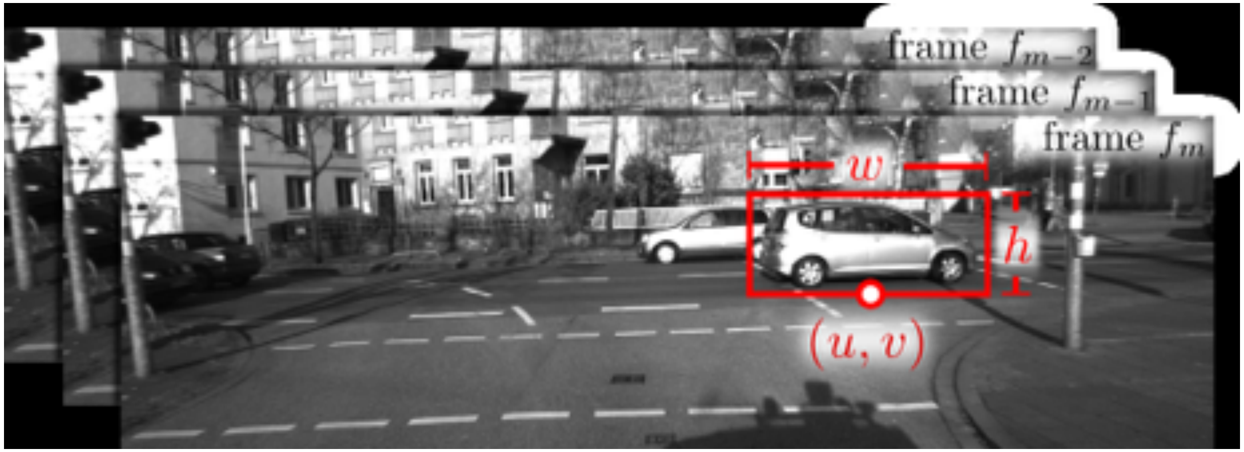
Image Evidence E = {T ; V; S; F;O}

Probabilistic Graphical Model



$$p(\mathcal{E}, \mathcal{R}) = p(\mathcal{R}) \prod_i p(\mathbf{t}_i|\mathcal{R}) \prod_i p(v_i|\mathcal{R}) \prod_i p(\mathbf{s}_i|\mathcal{R}) \prod_i p(\rho_i|\mathcal{R}) \prod_i p(\mathbf{f}_i|\mathcal{R})$$

Vehicle Tracklets

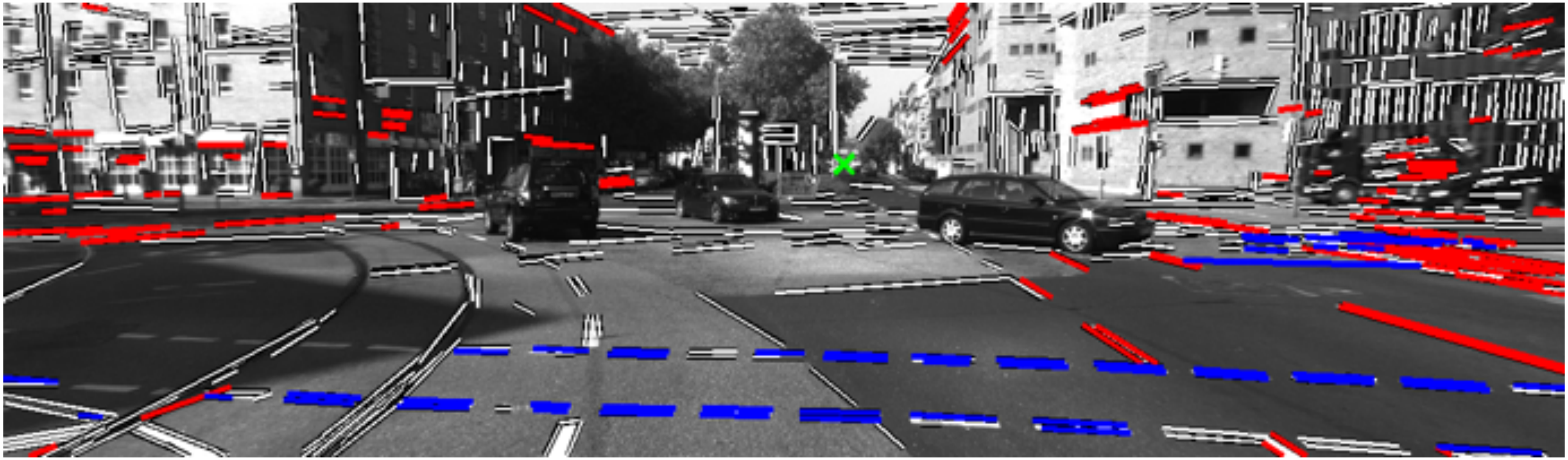


- Object detection [Felzenszwalb et al. 2010]
- Associate objects over time (tracking by detection)

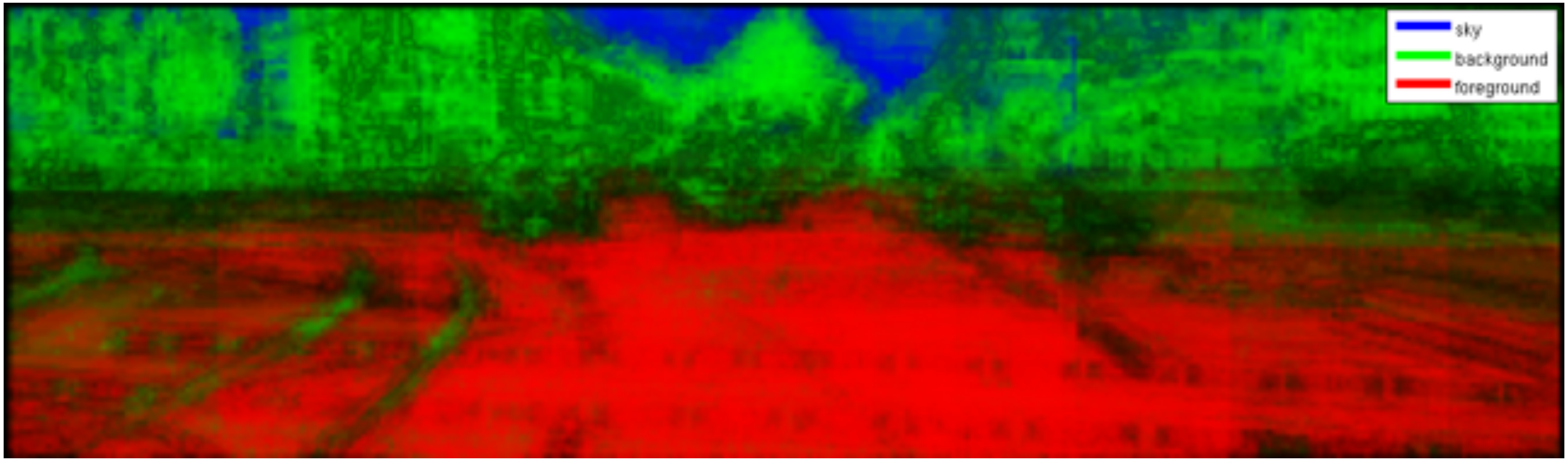
- Projection to 3D object tracklet **t = {d1, … , d}**
  (**d** captures the object location and orientation)

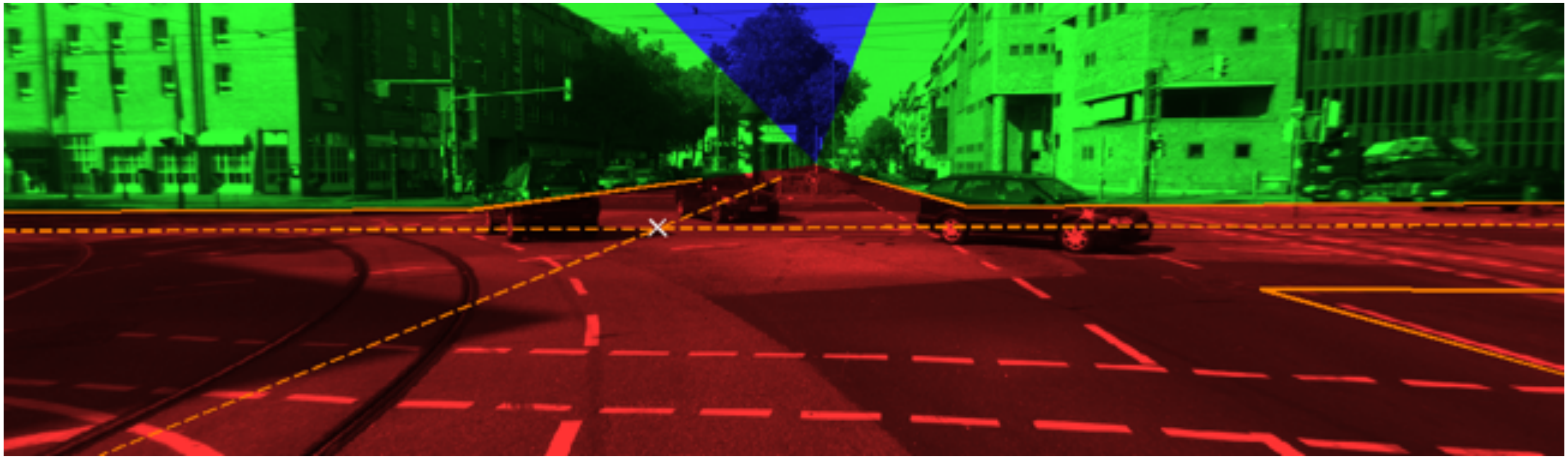
Probabilistic Graphical Model

Vanishing Points



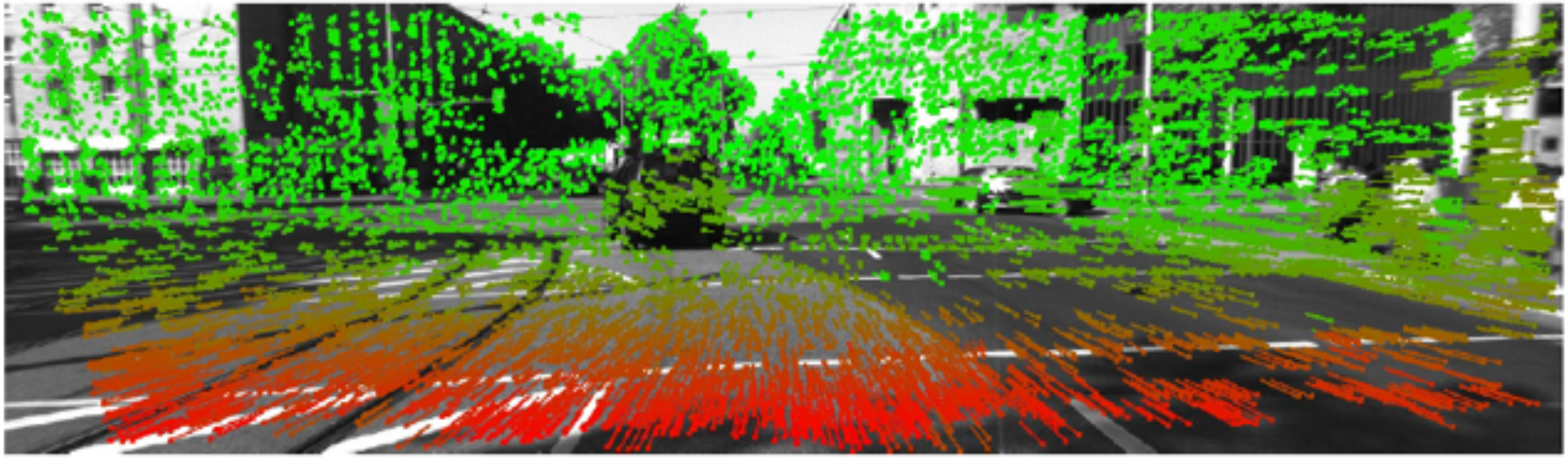
Semantic Labels





Probabilistic Graphical Model

Occupancy, Scene Flow



Inference

Denote

- $\mathcal{E}$ the image evidence
- $\mathcal{R}$ the road layout
- $\mathcal{C}$ the location of cars in the scene

Given $\mathcal{E}$, inference of $\mathcal{R}$ and $\mathcal{C}$ is solved in two steps:

- Infer road layout $\mathcal{R}$ while marginalizing $\mathcal{C}$

$$\hat{\mathcal{R}} = \underset{\mathcal{R}}{\operatorname{argmax}}\ p(\mathcal{R}|\mathcal{E}) \qquad \text{(Metropolis-Hastings)}$$

- Infer car locations $\mathcal{C}$ using MAP road layout $\mathcal{R}$

$$\hat{\mathcal{C}} = \underset{\mathcal{C}}{\operatorname{argmax}}\ p(\mathcal{C}|\mathcal{E}, \mathcal{R}) \qquad \text{(Dynamic programming)}$$

Experimental Results

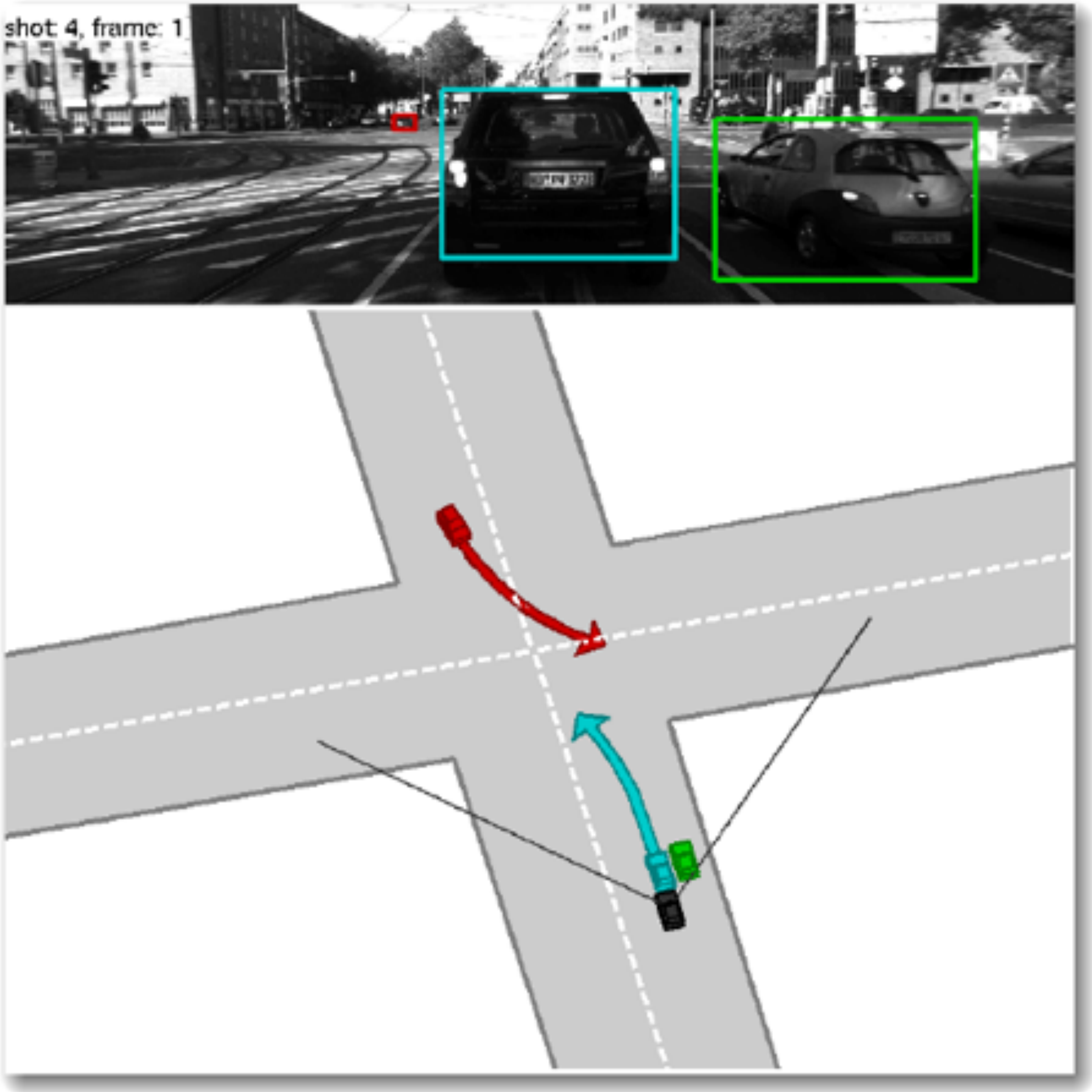
Experiments

- 113 sequences 5-30 seconds (9438 frames)
- Best results when combining all feature cues
- Most important: Occupancy grid, tracklets, 3D scene flow
- Less important: Semantic labels, vanishing points

Metrics

- Topology Accuracy: 92.0%
- Location Error: 3.0 m
- Street Orientation Error: 3.0
- Tracklet-to-Lane Accuracy: 82.0%
- Vehicle Orientation Error: 14.0

Experimental Results



3D Scene Understanding

- Defining the Problem
- Context
- Spatial Layout
- 3D Scene Understanding