VOC 2008: A Unified Approach for Detection, Classification and Segmentation

Derek Hoiem[1] Santosh Divvala[2] James Hays[2]

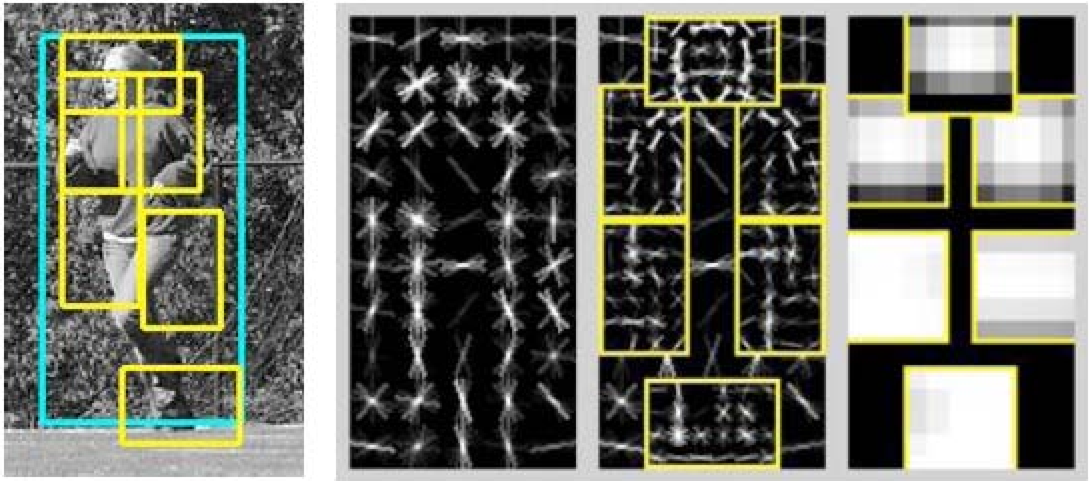
[1]University of Illinois at Urbana-Champaign, Beckman Institute

[2]Carnegie Mellon University, Robotics Institute

Take a Good Detector and Make It Better

- UofCTTI from VOC 2007 (CVPR 2008)
- Many thanks to Pedro Felzenszwalb, David McAllester, and Deva Ramanan!

Deformable Parts Model



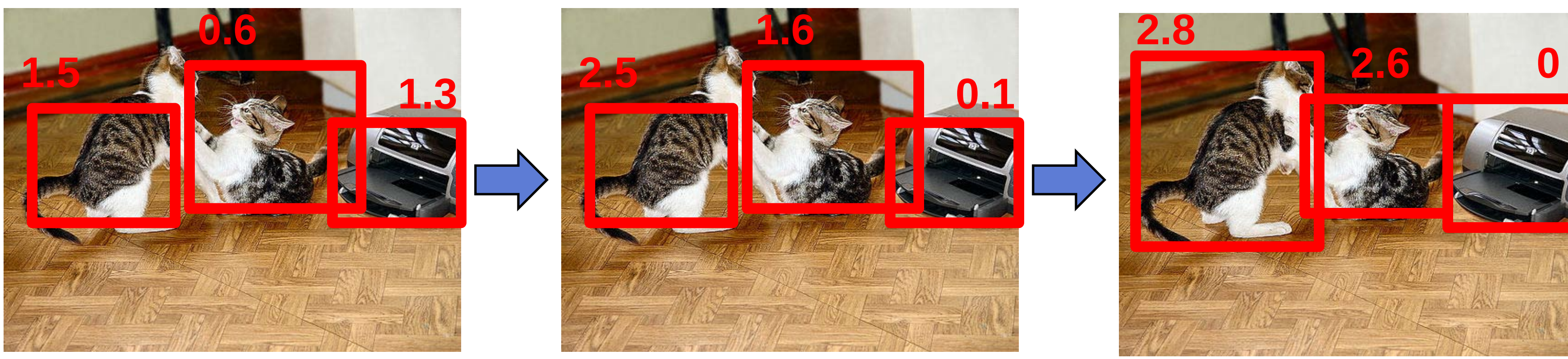
Felzenszwalb McAllester Ramanan 2008

Goal: Better Detection using Context and Segmentation

Local Detector Candidates

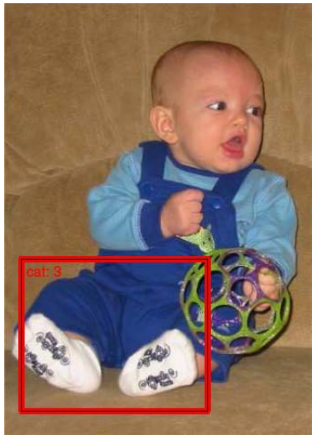
Improved Scores using Context Cues

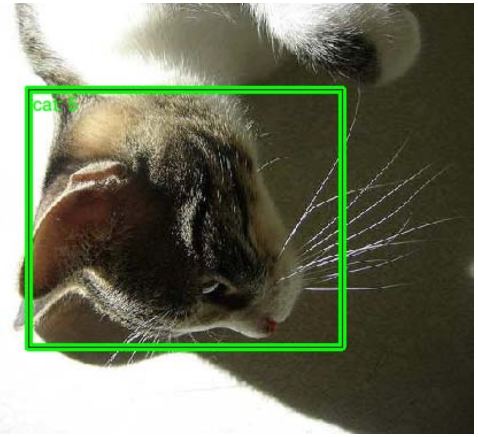
Improved Localization and Scores using Segmentation



I. Need for Context

- Example: Top 5 Cat Detections





Global Context

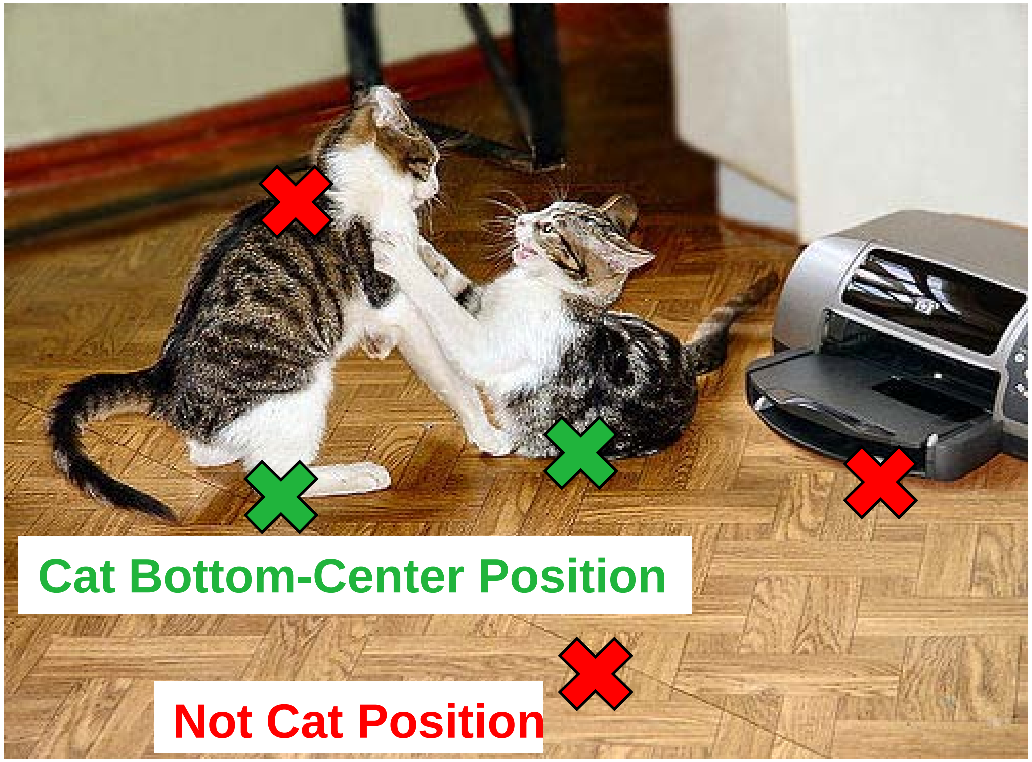
1. Object presence: P(object_present | image)

Contains Cat

No Cat

Global Context

1. Object presence: P(object_present | image)
2. Object position: P(object_xy | object_present, image)



Global Context

1. Object presence: P(object_present | image)
2. Object position: P(object_xy | object_present, image)
3. Object size: P(object_size | object_xy, object_present, image)

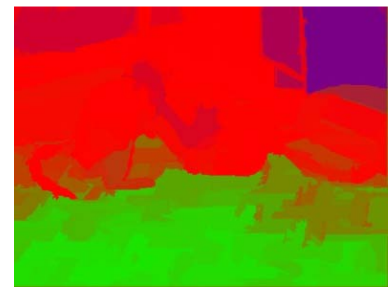
Likelihood of Object Presence

Image

Image Statistics

Gist

Geometric Context



Associated Data

High Population

Kitten

Puppy

Urban

gist: Torralba Oliva 2003
geom context: Hoiem et al. 2005
im2gps: Hays and Efros 2008

Classification by Association

Input Image

Sample of Nearest Neighbors

Associated **Tags**/**Geo**

Squirrel …
Seattle

Urban …
High Pop

Lion …
Zoo

Grassland …
Low Pop

…

River …
Waterfall

Water …
Low Pop

Associated Data for Decision

Keywords

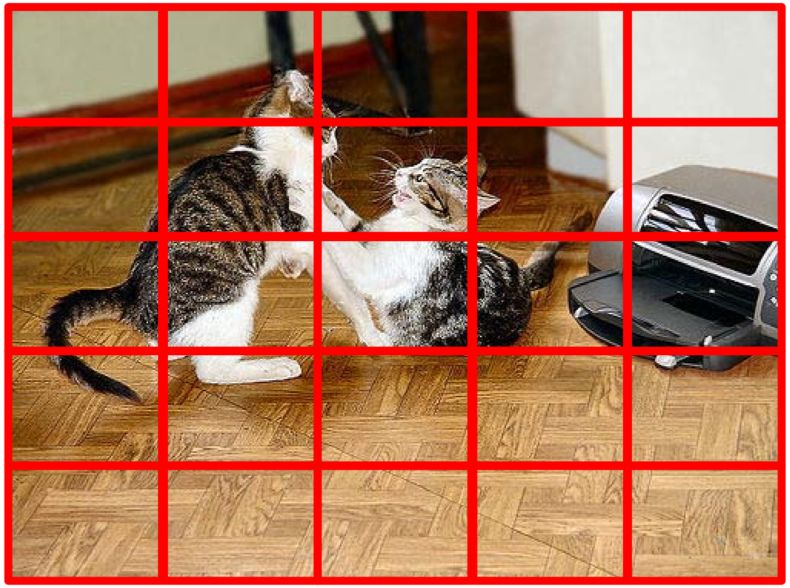
Animals: 4
Food: 1
Aquarium: 2
Sky: 0
Boat: 0
Race: 0
Verdict: Likely

Geographic Context

Photo Density: 0.2
Pop Density: 0.6
CropsGrass: 0.2
Boat: 0
Race: 0
Verdict: Not Sure

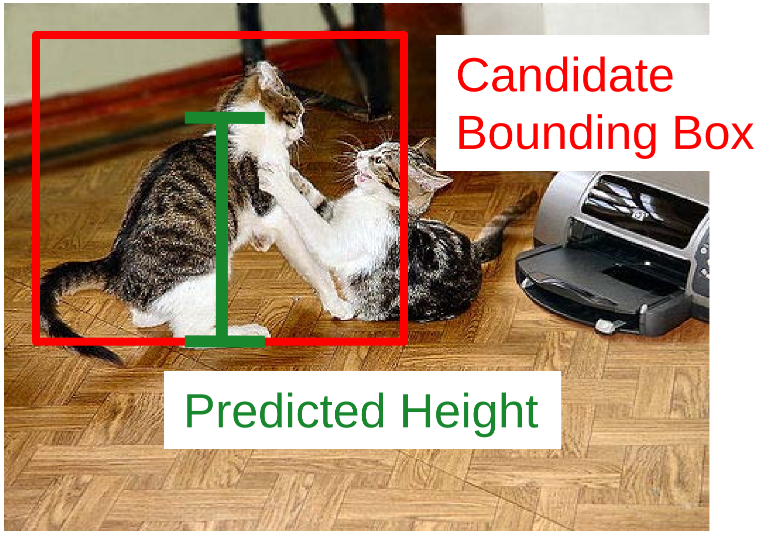
Likelihood of Object Position

- Build classifier for each cell based on whole image gist and geometric context



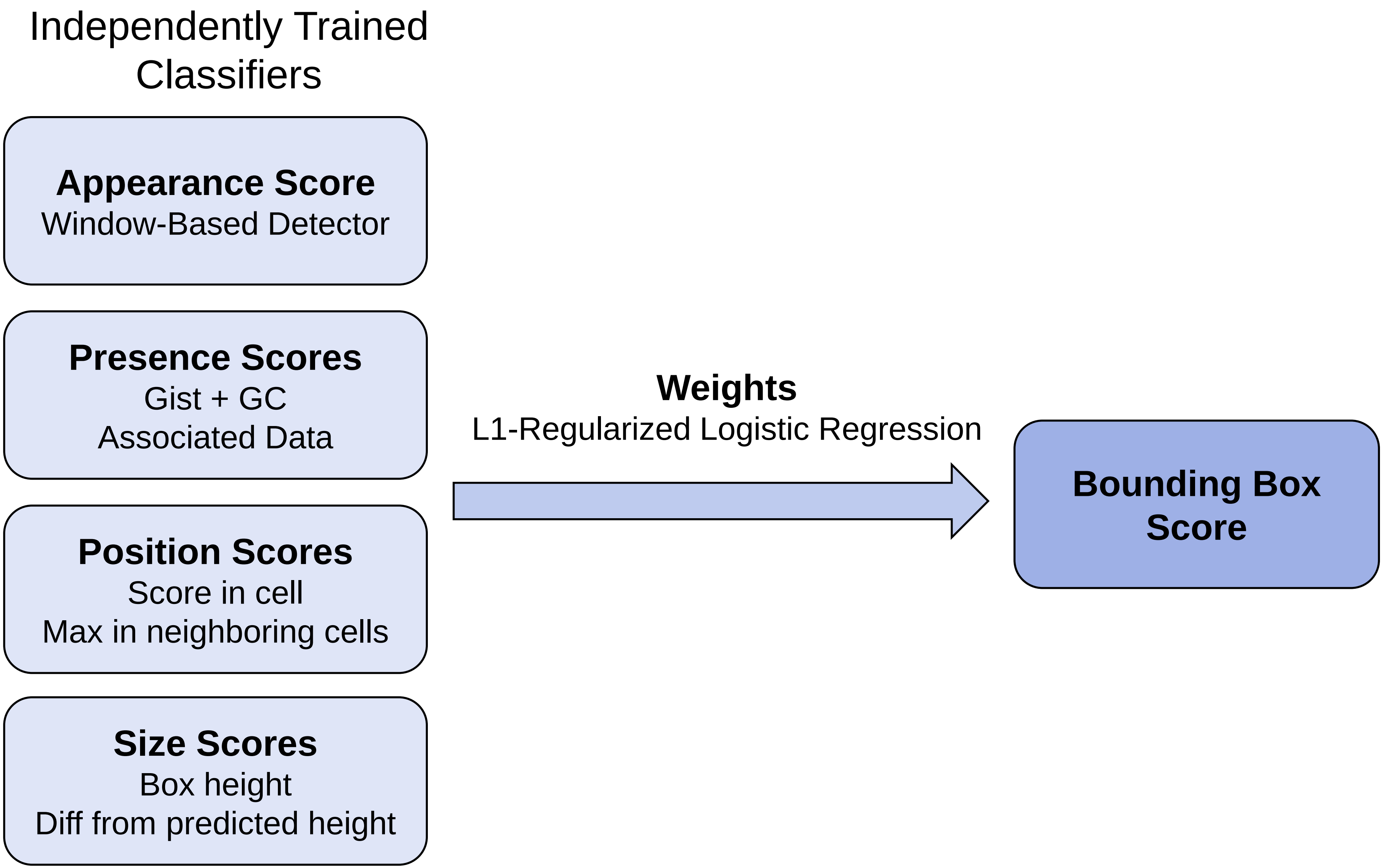
Likelihood of Object Size

- Predict bounding box height at given location
  - y-position
  - depth estimate at position
  - global gist and geometric context



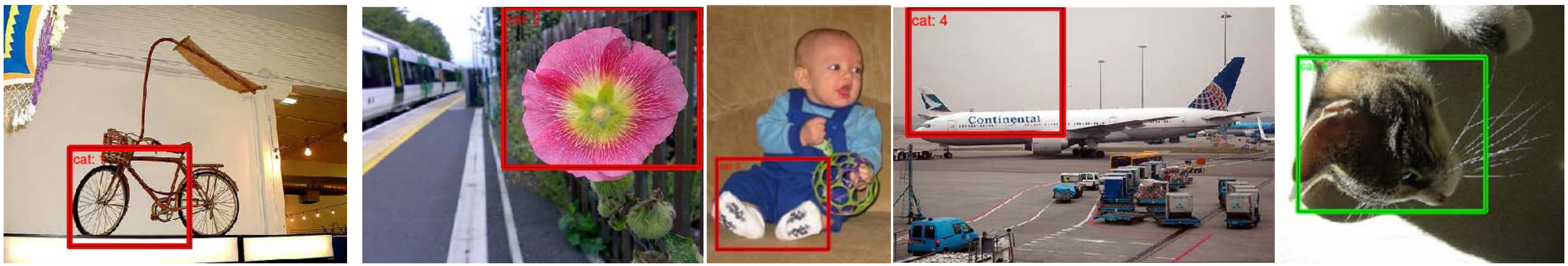
Depth: Hoiem et al. 2007
Size from Gist: Torralba Oliva 2003

Score Combination



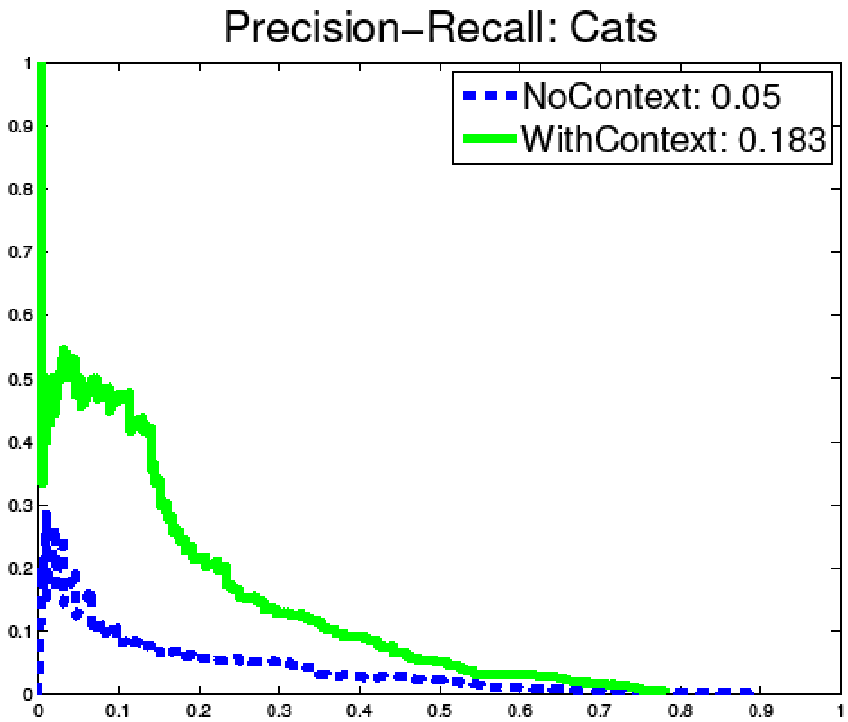
Top-Ranked Candidates Are More Reliable with Context

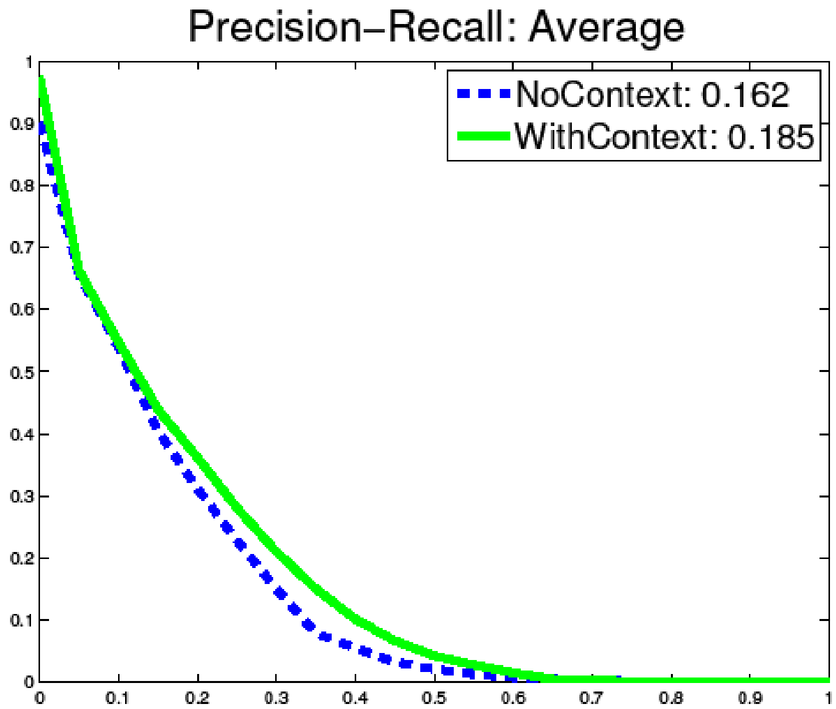
Top 5: Before Context



Top 5: After Context

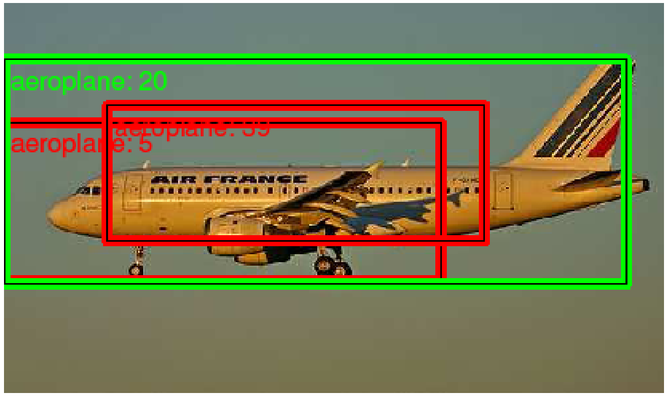
Quantitative Improvement with Context





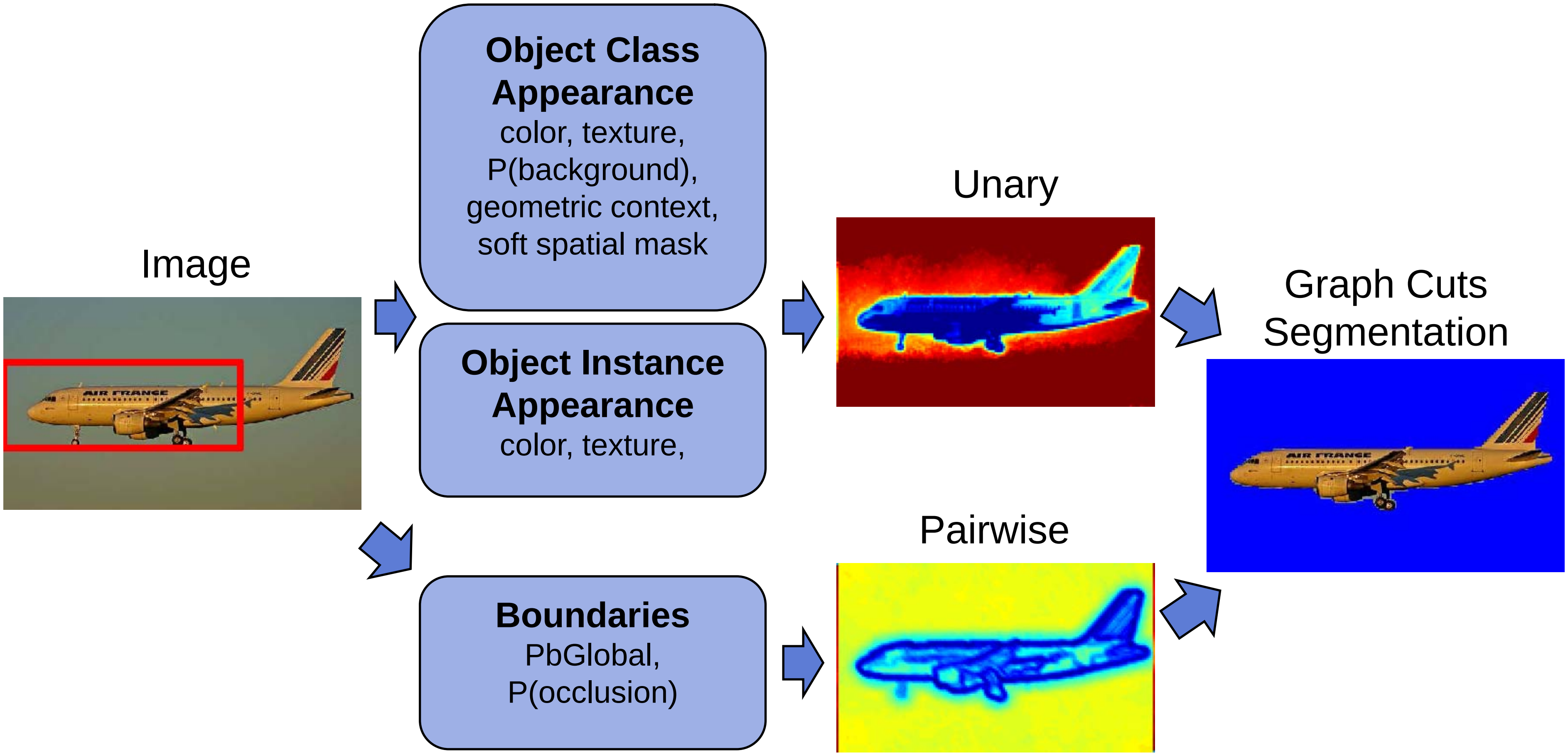
II. Need for Better Localization

Multiple Detections



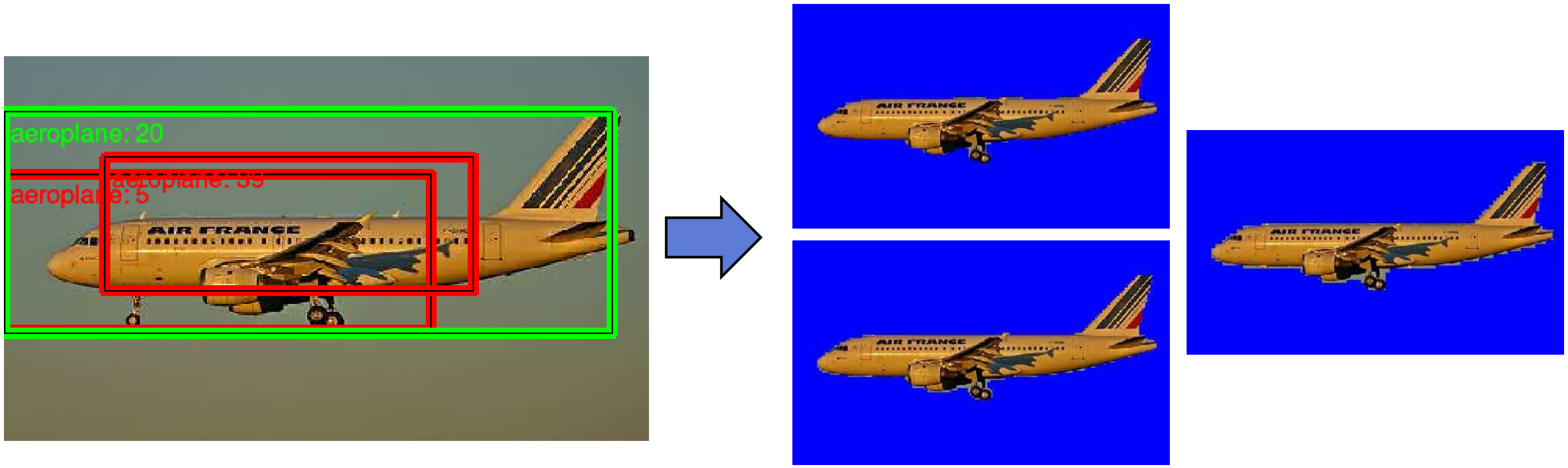
Poor Localization

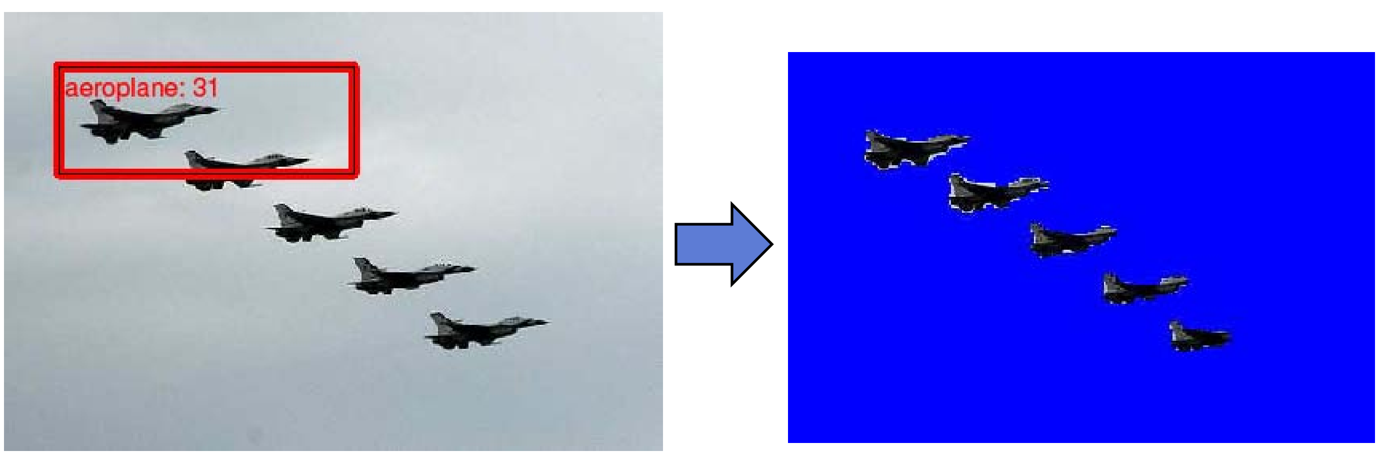
Segmentation



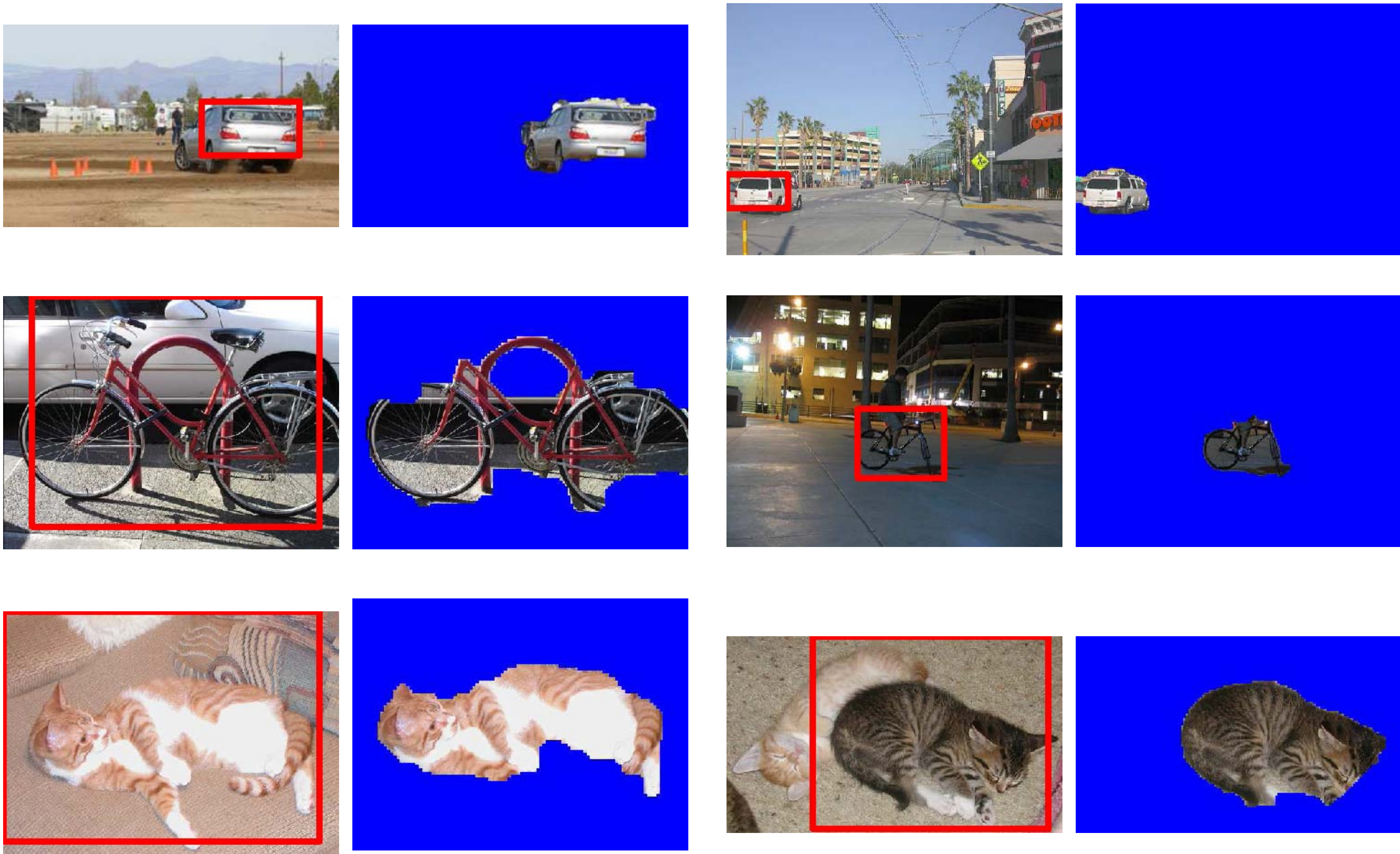
PbGlobal: Maire et al. 2008
Occlusion: Hoiem et al. 2007
GraphCuts: Boykov et al. 2001

Segmentation Examples





Segmentation Examples

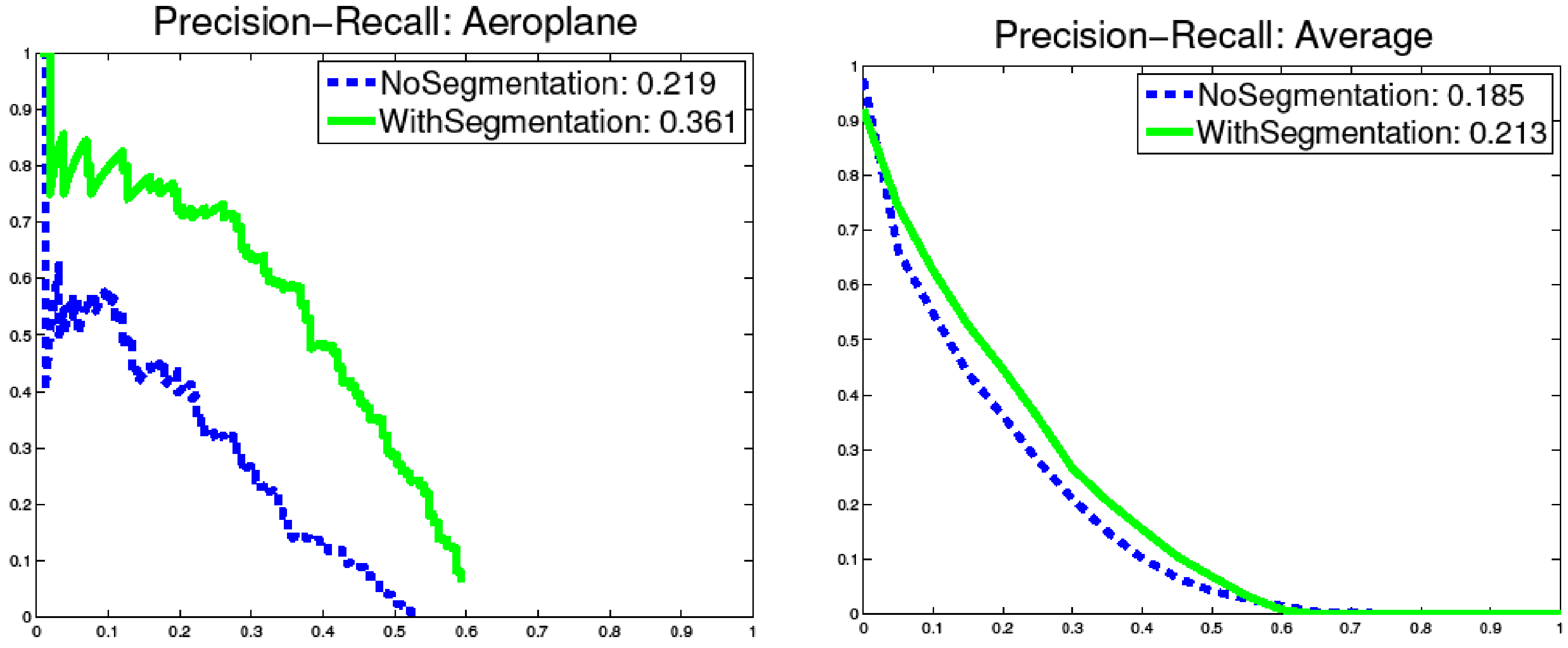


Segment Appearance

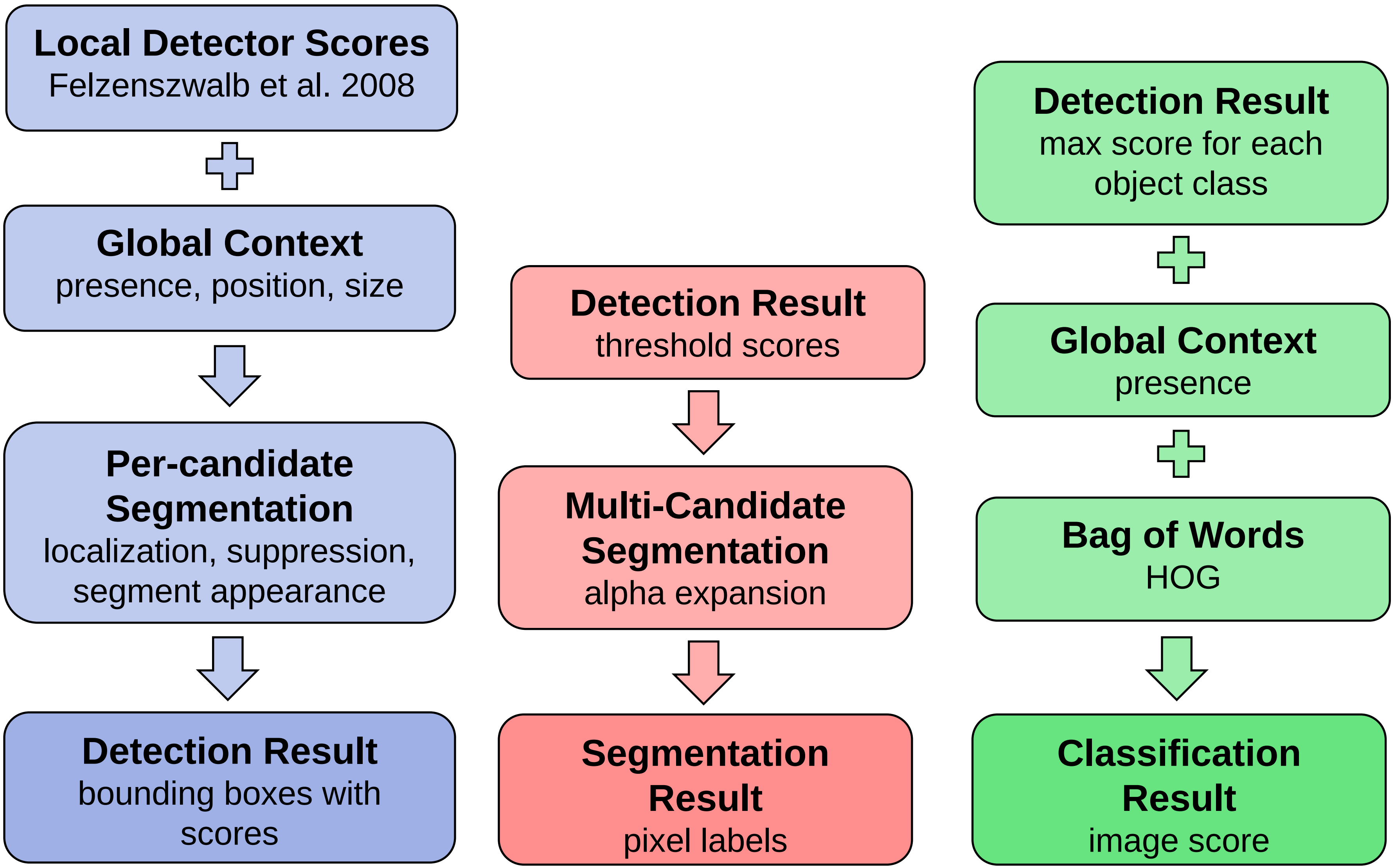
- Histogram (normalized bin count + entropy)
  - Quantized color
  - Textons
  - Quantized HOG

- Final score = w_b bbox_score + w_s segment_score

Quantitative Improvement with Segmentation



Detection, Segmentation, Classification



Overall VOC'08 Challenge Results

	UIUC_CMU	Top	Second
Classification (comp2)	44.3	58.6[1]	54.2[2]
Detection (comp4)	22.0	22.9[3]	22.6[4]
Segmentation (comp6)	19.5	25.4[5]	20.1[6]

1. UvA_0708Soft5ColorSift
2. UvA_AdapTagRelDom
3. LEAR_PlusClass (comp3)
4. UoCTTIUCI (comp3)
5. XRCE_Seg (comp5)
6. BrookesMSRC (comp5)

Detection Results

= First = Second

	LEAR (Comp3)	UoCTTI (Comp3)	UIUC_CMU (Comp4)
AEROPLANE	36.5	32.6	34.5
BICYCLE	34.3	42.0	32.7
BIRD	10.7	11.3	12.3
BOAT	11.4	11.0	11.0
BOTTLE	22.1	28.2	22.4
BUS	23.8	23.2	18.5
CAR	36.6	32.0	27.8
CAT	16.6	17.9	21.6
CHAIR	11.1	14.6	8.8
COW	17.7	11.1	14.1
DINING TABLE	15.1	6.6	15.2
DOG	9.0	10.2	17.8
HORSE	36.1	32.7	27.4
MOTORBIKE	40.3	38.6	40.9
PERSON	19.7	42.0	37.4
POTTED PLANT	11.5	12.6	11.2
SHEEP	19.4	16.1	7.0
SOFA	17.3	13.6	13.5
TRAIN	29.6	24.4	28.2
TV MONITOR	34.0	37.1	38.5

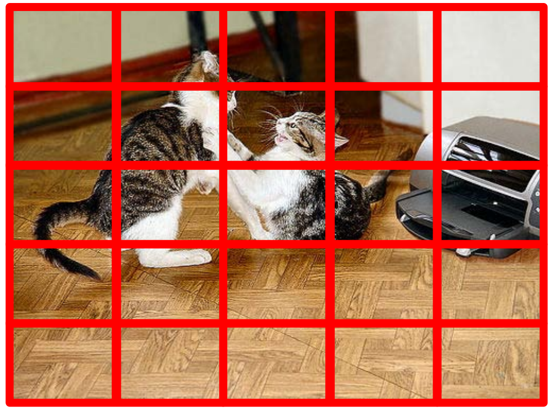
Importance of Context & Segmentation for Detection

	Mean A.P.*	Classes most benefitted
Local Detector (UoCTTI'07)	18.1	
+ Context	20.5	Dining table, Motorbike, Cat, Dog, Person
+ Segmentation	21.3	Airplane
Final (UIUC_CMU'08)	22.6	TV monitor, Train

(*on VOC Val'08)

Relative Importance of Contextual Features

P(object_present | image)



P(object_xy | object_present, image)

P(object_size | object_xy, object_present, image)

	Mean A.P.*
Local Detector (UoCTTI'07)	18.1
+ Scene, Location, Size	20.5
except Scene	19.1
except Location	19.9
except Size	18.9

(*on VOC Val'08)

Qualitative Observations

- ☺ Classes helped: Airplane, bird, cat, cow, dog, dining table, person, sofa, tv monitor, train

Aeroplane

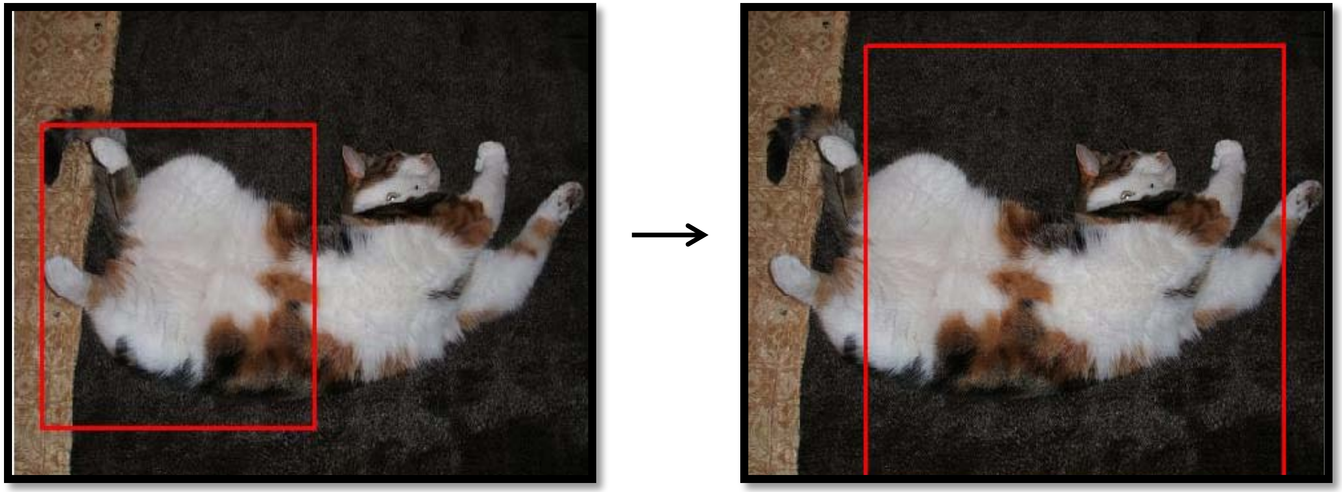

Two of the top 10 detections by only using UoCTTI’07

→

Segmentation: Improves Localization

Cat

Two of the top 10 detections by only using UoCTTI'07



Segmentation: Improves Localization

Qualitative Observations

- ☺ Classes helped: Airplane, bird, cat, cow, dog, dining table, person, sofa, tv monitor, train

- 😐 Classes not helped: Bottle, potted plant, horse, bus, car, bicycle, motorbike

What *context* should be used?

Potted Plant

Bottle

Qualitative Observations

- ☺ Classes helped: Airplane, bird, cat, cow, dog, dining table, person, sofa, tv monitor, train

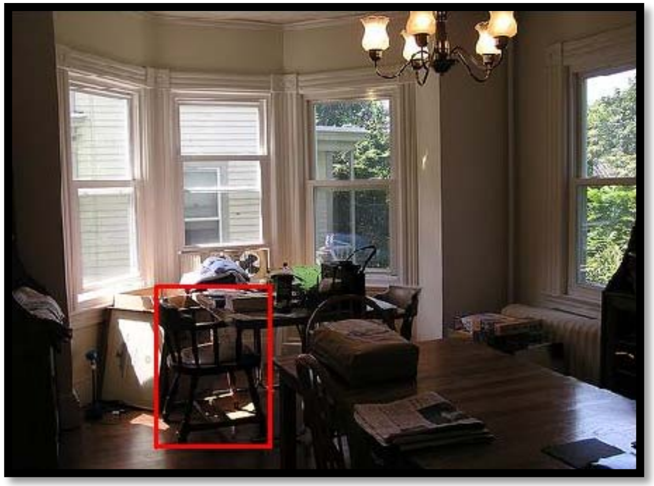
- 😐 Classes not helped: Bottle, potted plant, bus, car, bicycle, motorbike

- ☹ Classes hurt: Chair, sheep, boat

Poor Segmentation can misguide the detector

Before Segmentation

After Segmentation

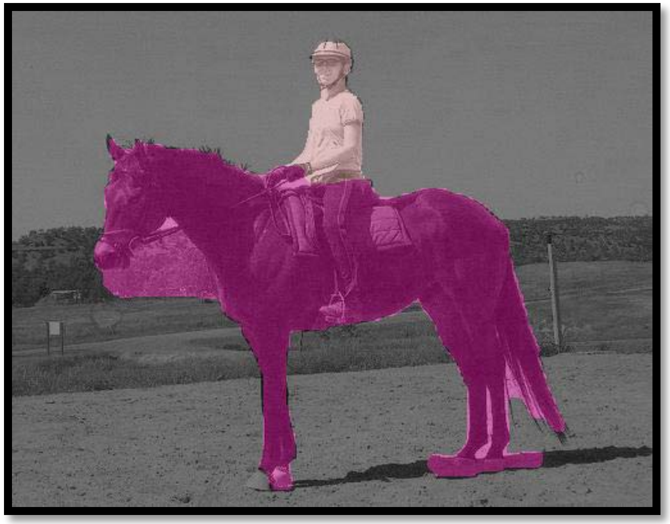


Segmentation Results

= First (yellow) = Second (orange)

	UIUC_CMU (comp6)	*XRCE_Seg (comp5)*	*Brookes_MSRC (comp5)*
AEROPLANE	31.9	25.8	36.9
BICYCLE	21.0	15.7	4.8
BIRD	8.3	19.2	22.2
BOAT	6.5	21.6	11.2
BOTTLE	34.3	17.2	13.7
BUS	15.8	27.3	13.8
CAR	22.7	25.5	20.4
CAT	10.4	24.2	10.0
CHAIR	1.2	7.9	8.7
COW	6.8	25.4	3.6
DINING TABLE	8.0	9.9	28.3
DOG	10.2	17.8	6.6
HORSE	22.7	23.3	17.1
MOTORBIKE	24.9	34.0	22.6
PERSON	27.7	28.8	30.6
POTTED PLANT	15.9	23.2	13.5
SHEEP	4.3	32.1	26.8
SOFA	5.5	14.9	12.1
TRAIN	19.0	25.9	20.1
TV MONITOR	32.1	37.3	24.8

Segmentation Results



Conclusions

- Common framework for classification, detection and segmentation
- Use of context and segmentation to improve object detection

Thank You

Local Detector Candidates

Improved Scores using Context Cues

Improved Localization and Scores using Segmentation