MIT
COMPUTER
VISION

6.819 / 6.869: Advances in Computer Vision

High-level vision:

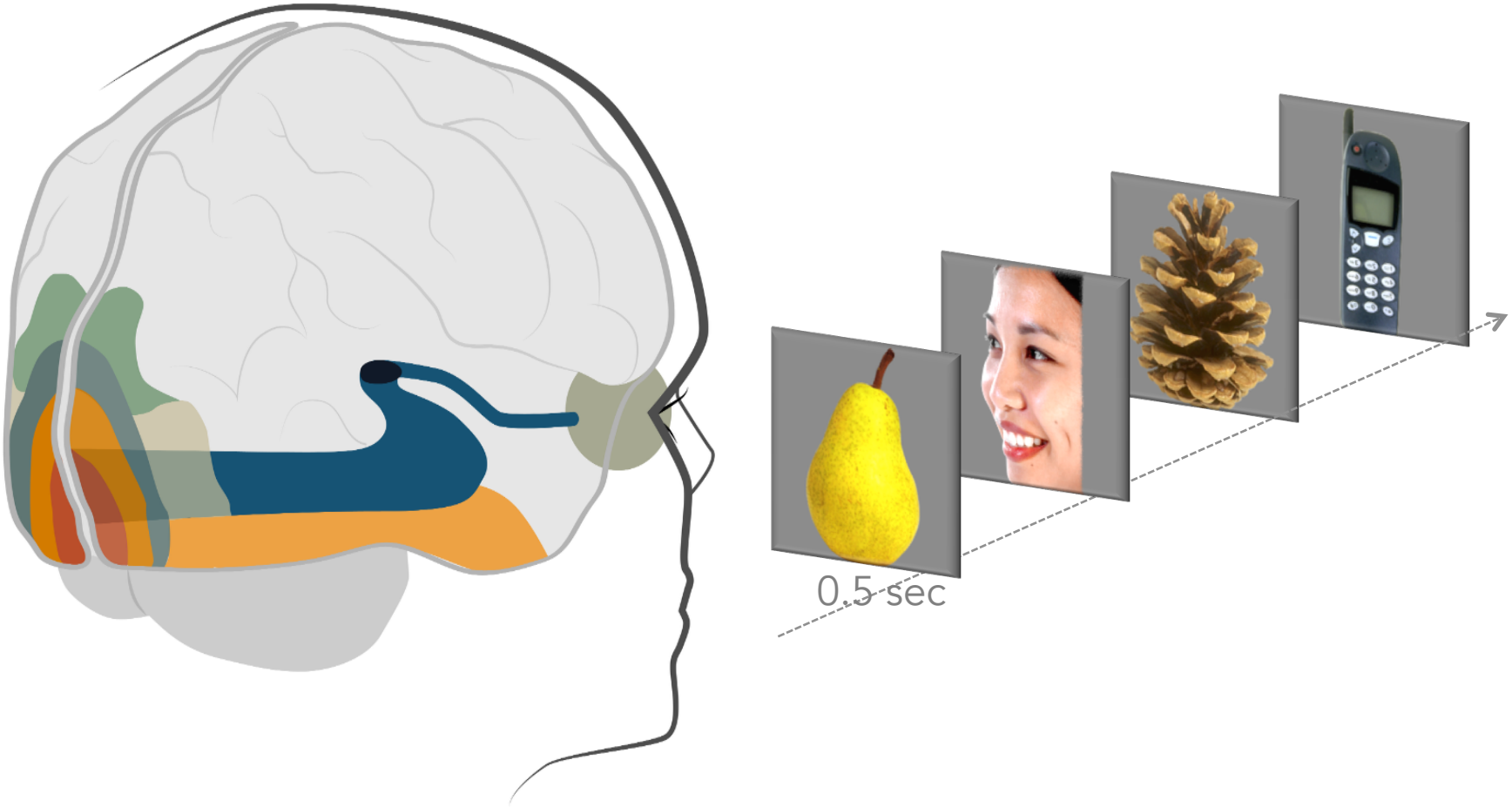
Object & Scene Recognition:
What are the next challenges? (cont)

Website:
http://6.869.csail.mit.edu/fa15/

Instructor: Aude Oliva

Lecture TR 9:30AM – 11:00AM
(Room 34-101)

High Computing Visual Engine: Object recognition



The object

The texture

The object

The texture

The object

The scene

Why do we care about recognition?

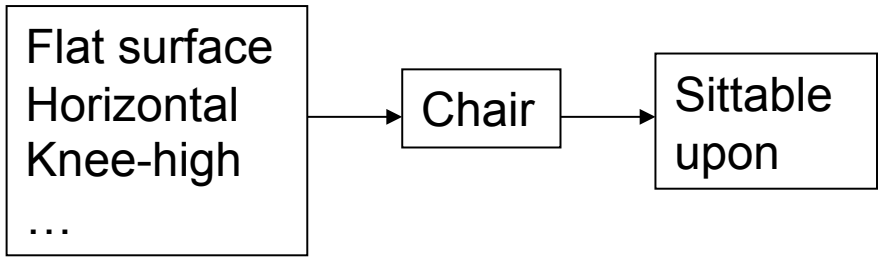
Perception of **function**: We can perceive the 3D shape, texture, material properties, without knowing about objects. But, the concept of category encapsulates also information about what can we do with those objects.

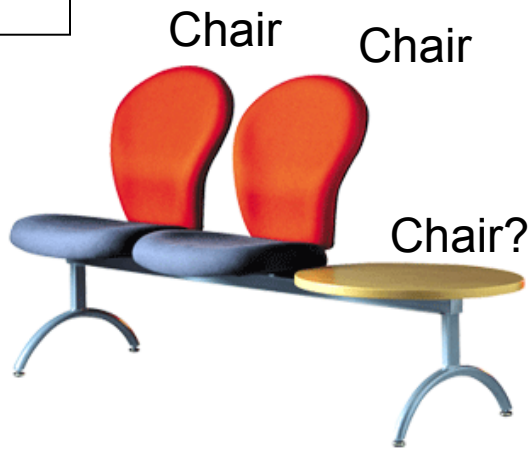
"We therefore include the perception of function as a proper –indeed, crucial- subject for vision science", *from Vision Science, chapter 9, Palmer*.

The perception of function

- Direct perception (affordances): Gibson

- Mediated perception (Categorization)

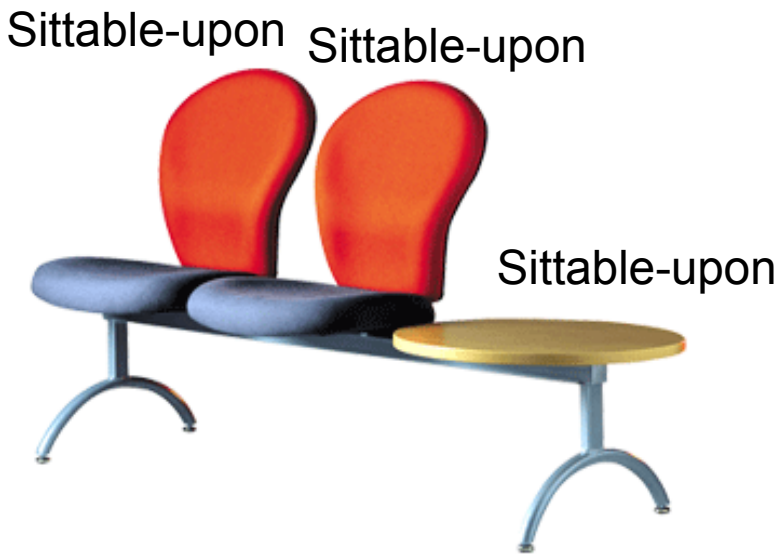




Direct perception

Some aspects of an object function can be perceived directly

- Functional form: Some forms clearly indicate to a function (“sittable-upon”, container, cutting device, …)



It does not seem easy to sit-upon this…

Limitations of Direct Perception

Objects of similar structure might have very different functions

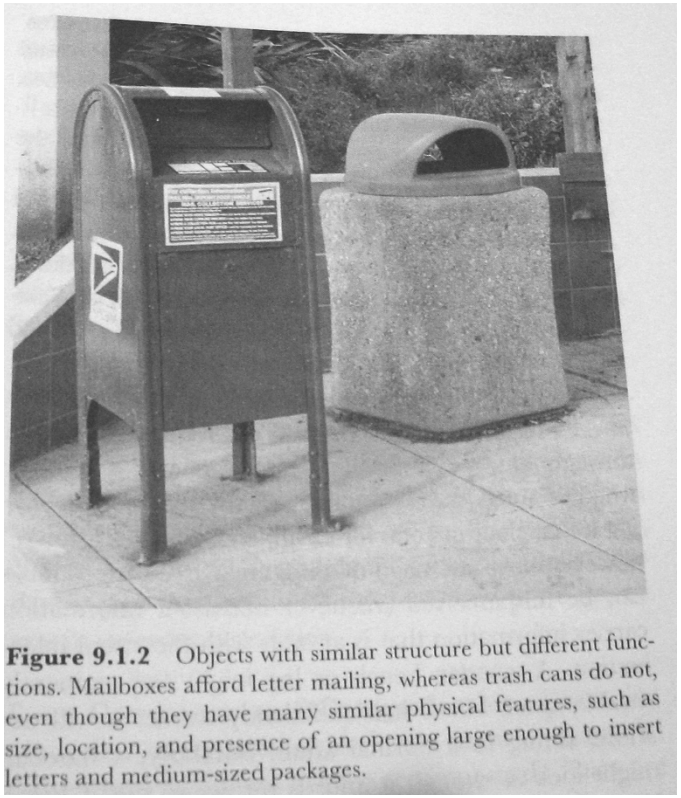
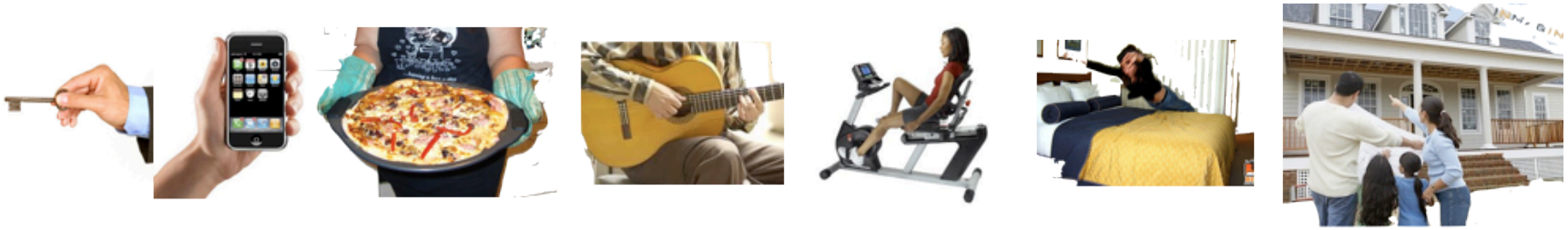


Figure 9.1.2 Objects with similar structure but different functions. Mailboxes afford letter mailing, whereas trash cans do not, even though they have many similar physical features, such as size, location, and presence of an opening large enough to insert letters and medium-sized packages.

Not all functions seem to be available from direct visual information only.

The functions are the same at some level of description: we can put things inside in both and somebody will come later to empty them. However, we are not expected to put inside the same kinds of things…

Interactions are driven by real-world size

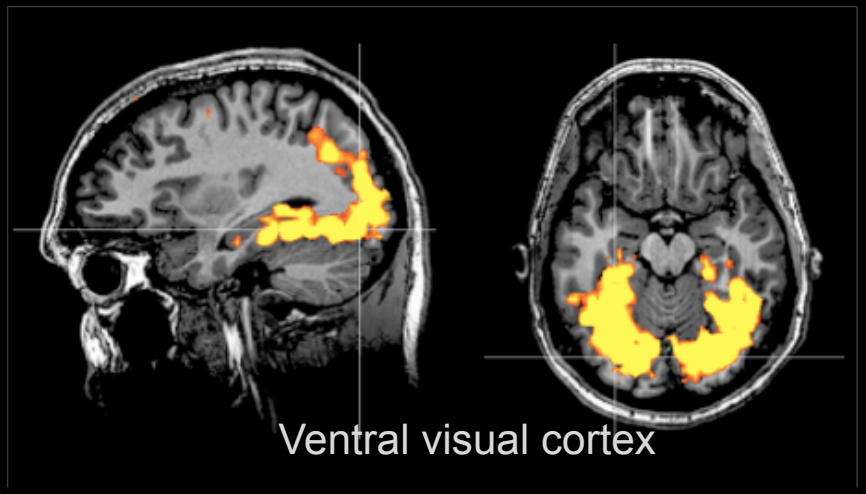


coordinate frames of action:

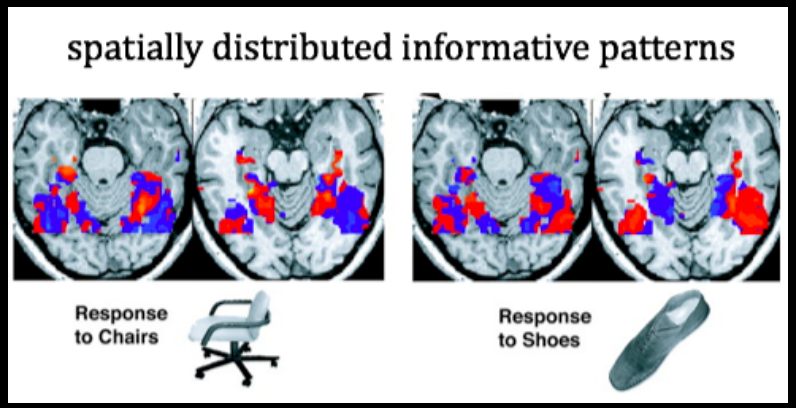
Finger/hand *hand/arm* *arm/body* *body* *body/space* *bodies/spaces*

1 person *1000+ people*

Object representation in the brain



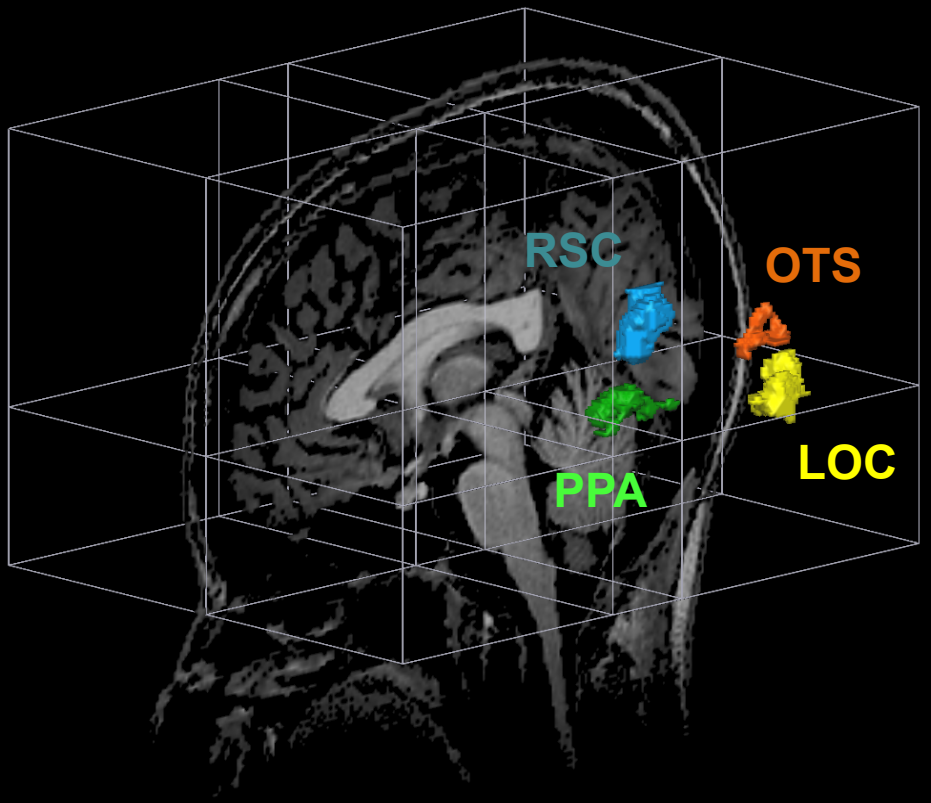
Ventral visual cortex



Haxby, 2001

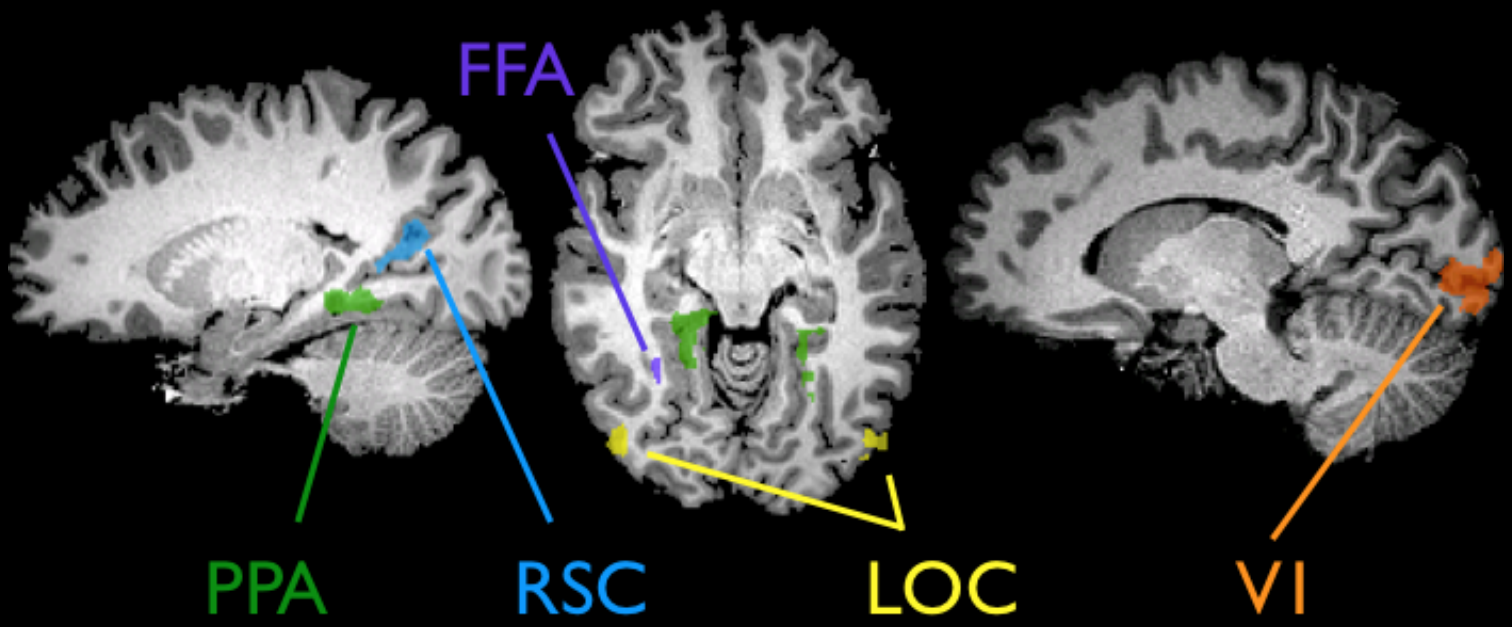
Regions of Interests (ROIs)

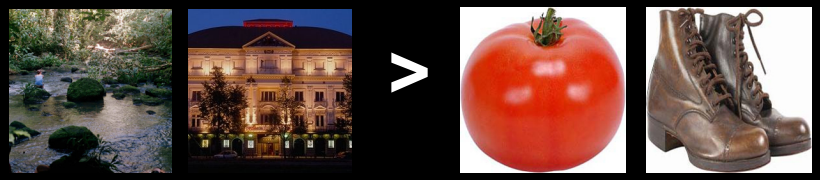
Patches of cortex with similar location and function in everyone



Regions of Interest

Functionally defined for each individual

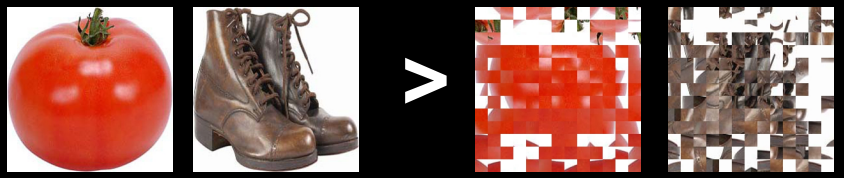




Epstein & Kanwisher (1998)

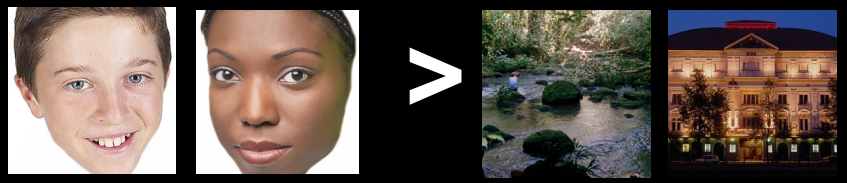
Parahippocampal place area (PPA)

Retrosplenial complex (RCS)



Malach et al (1995)

Lateral Occipital Complex (LOC)



Kanwisher et al (1997)

Fusiform Face Area (FFA)







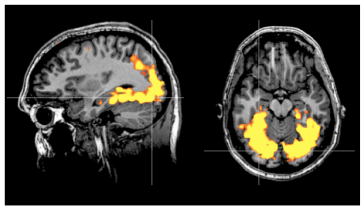










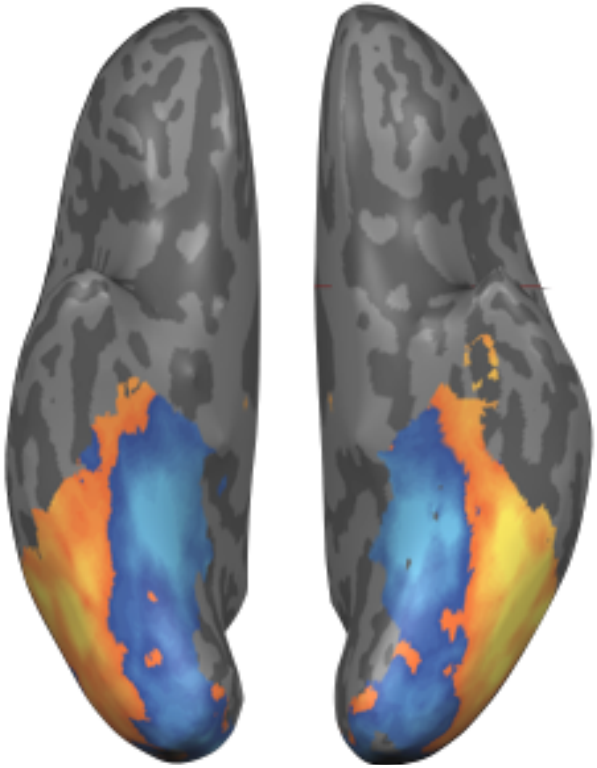



Different cortical regions for small and big objects

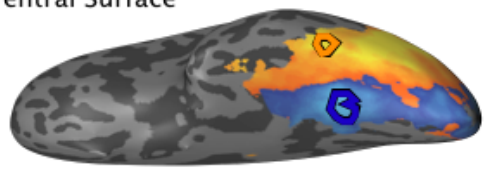
Small Objects

Big Objects

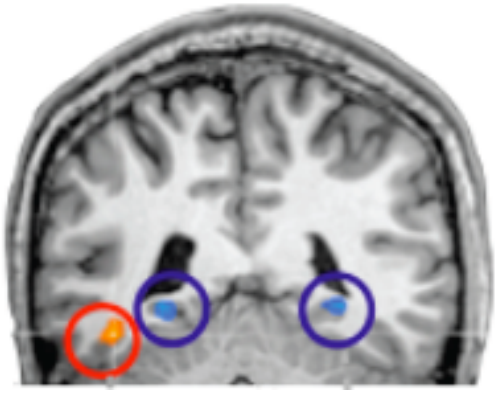
View of the bottom surface of an "inflated" brain



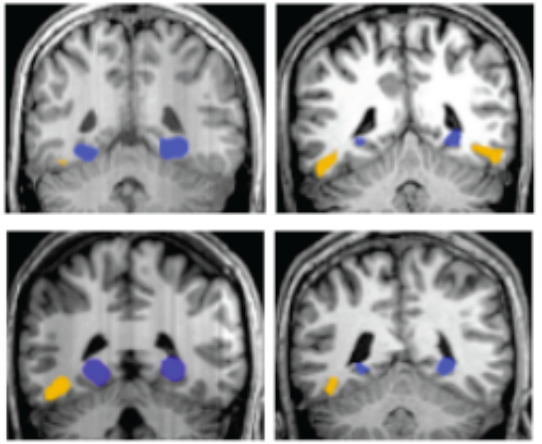
Ventral Surface



Whole brain analysis (n=12)

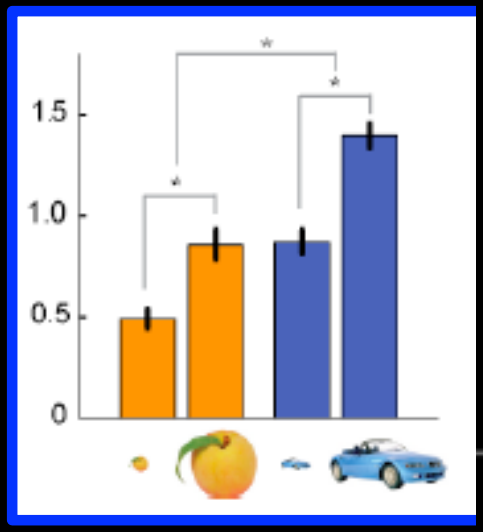


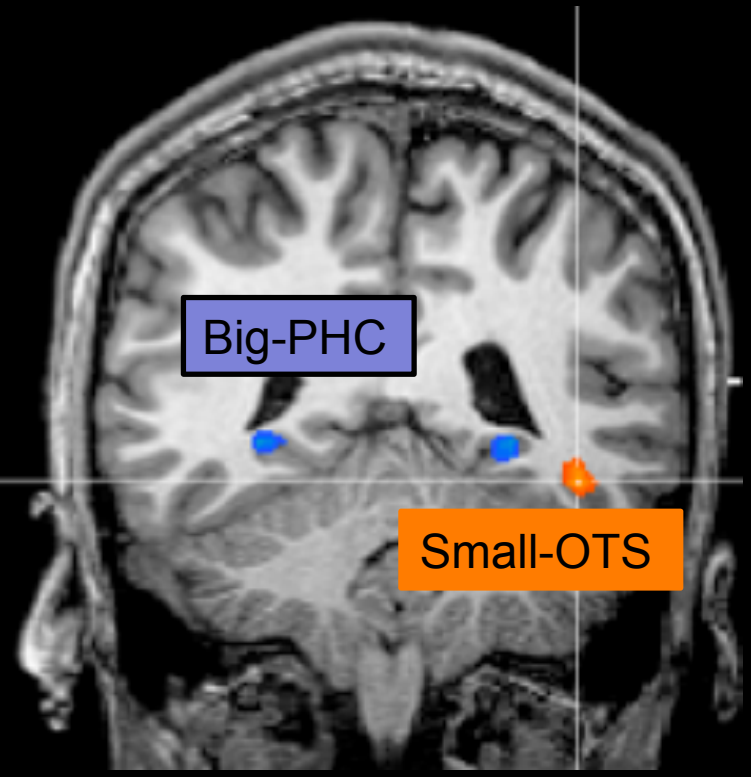
Individual brains

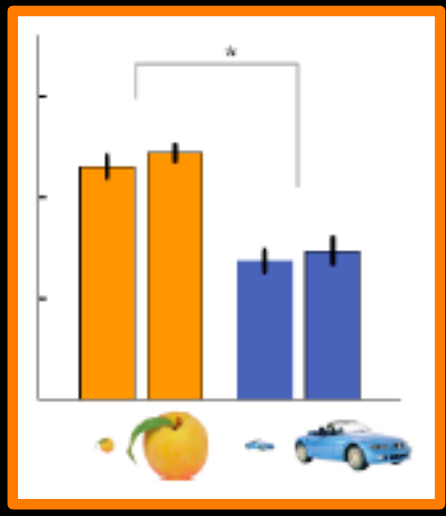


Konkle & Oliva (2012). *Neuron*

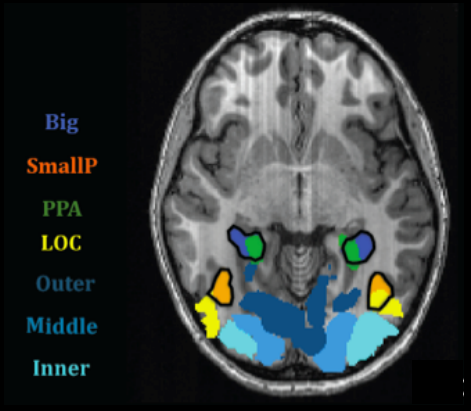
Invariant to retinal (image) size



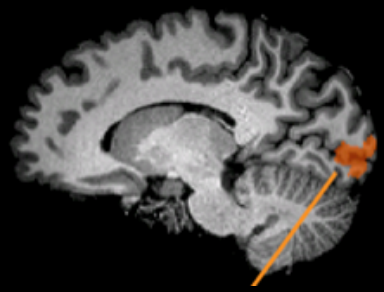


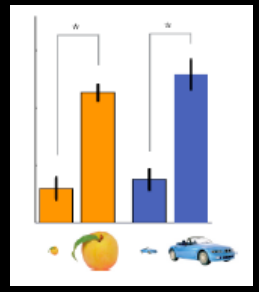


Big-PHC overlaps with PPA



Early visual areas





Konkle & Oliva (2012). *Neuron*

Challenges 1: view point variation

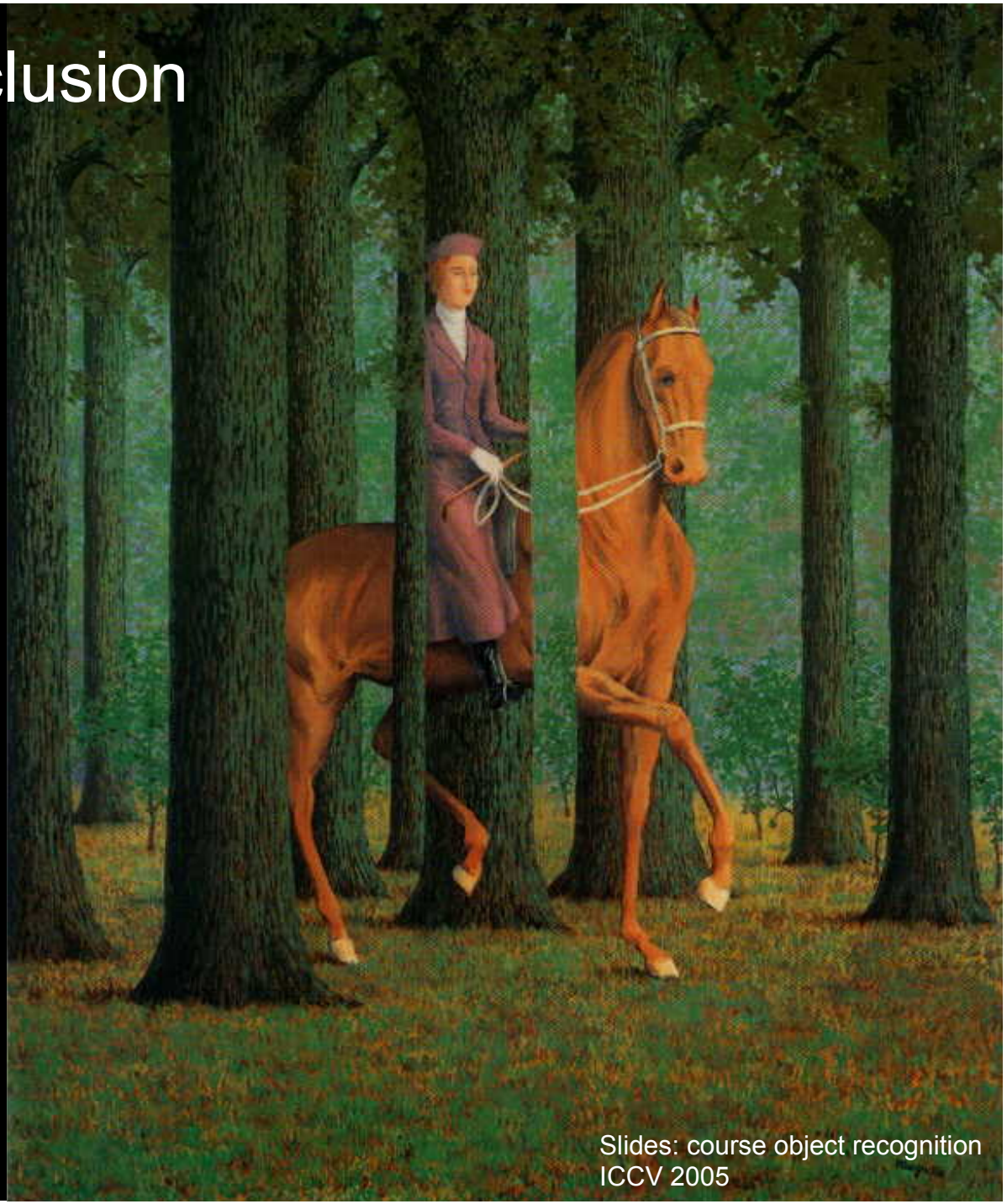
Michelangelo 1475-1564

Slides: course object recognition
ICCV 2005

Challenges 2: illumination

slide credit: S. Ullman

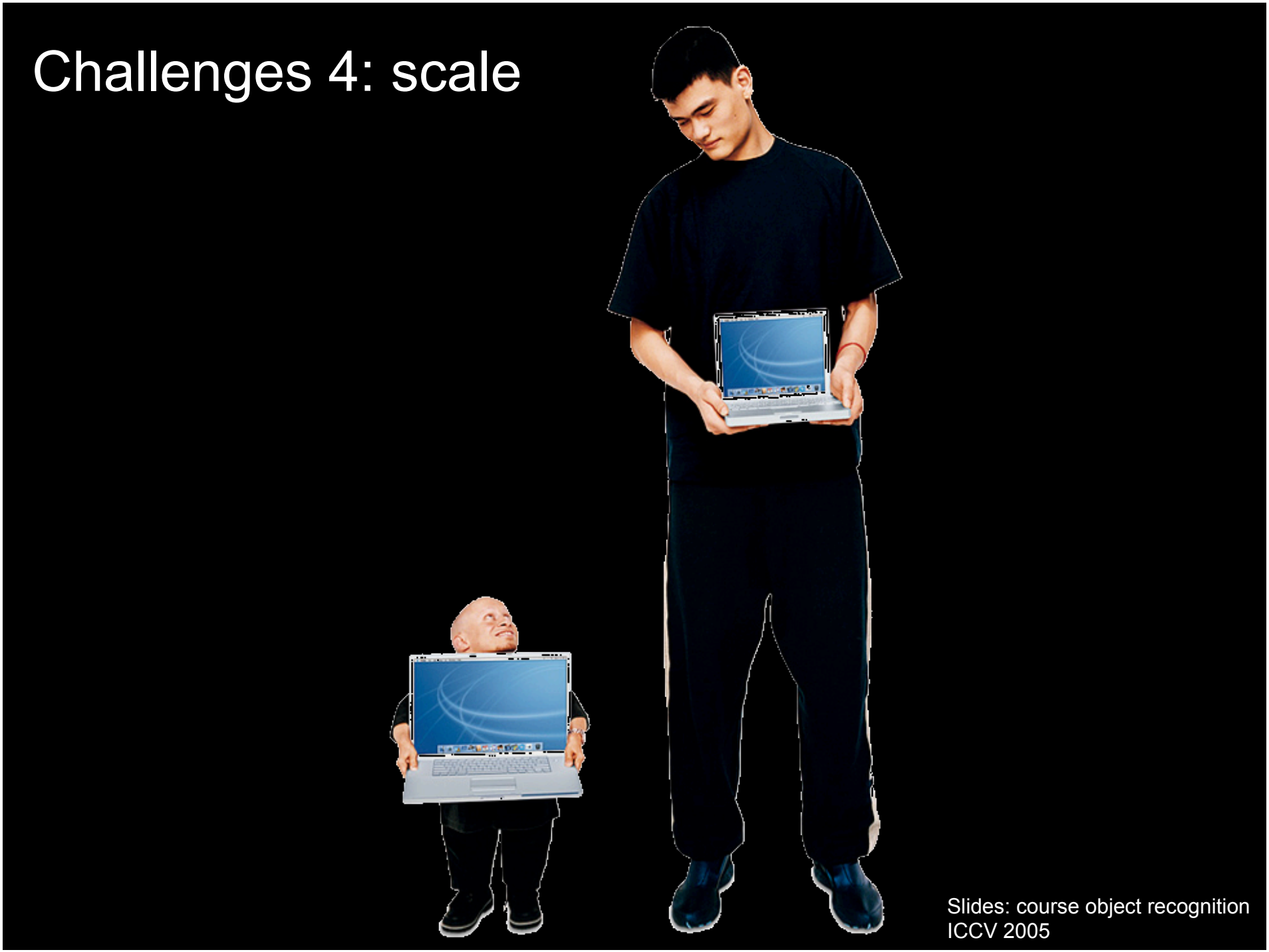
Challenges 3: occlusion



Magritte, 1957

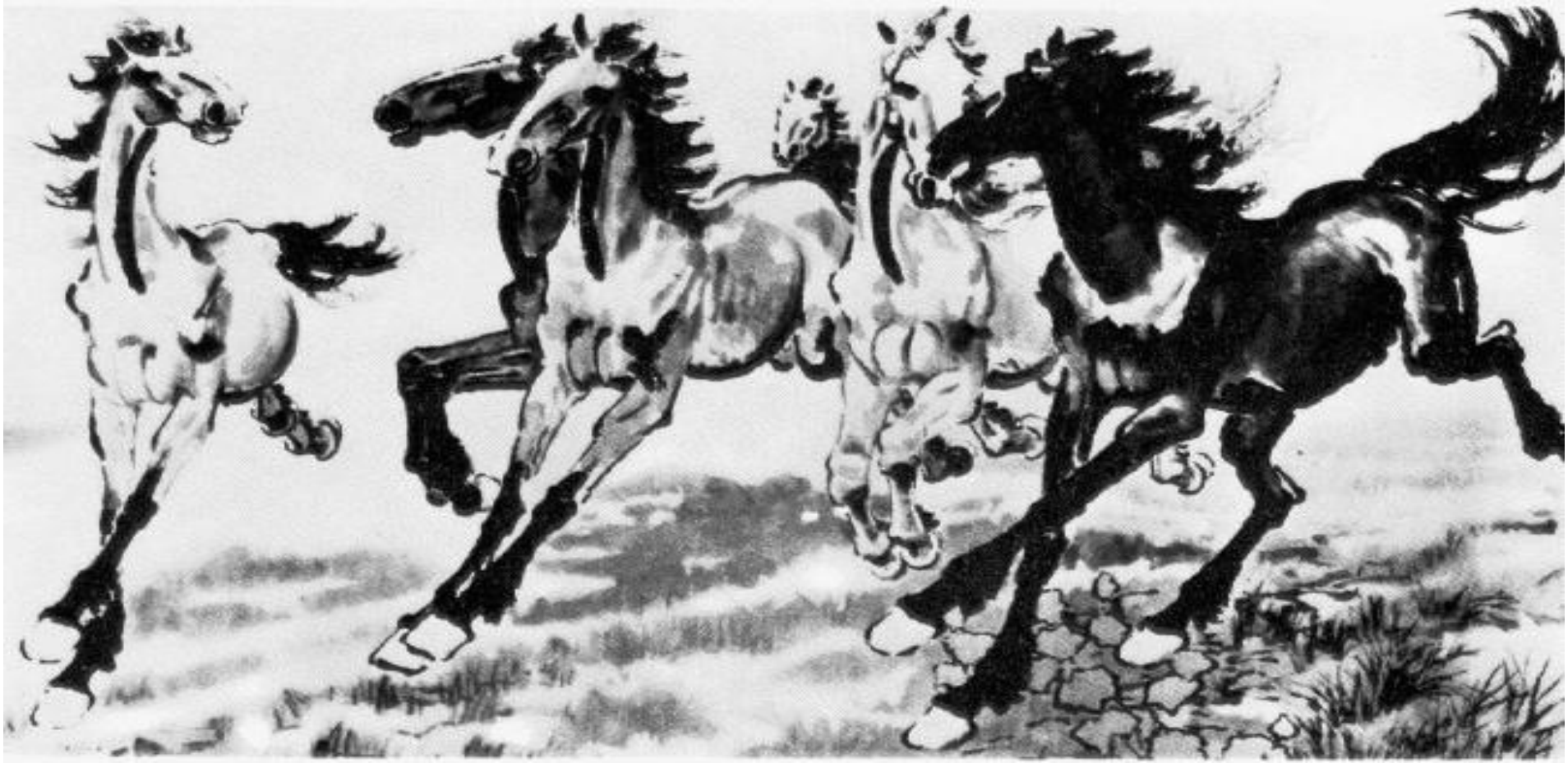
Slides: course object recognition
ICCV 2005

Challenges 4: scale



Slides: course object recognition
ICCV 2005

Challenges 5: deformation



Slides: course object recognition
ICCV 2005

Xu, Beihong 1943

Challenges 6: intra-class variation

Slides: course object recognition
ICCV 2005

Challenges 7: background clutter

Brady, M. J., & Kersten, D. (2003). Bootstrapped learning of novel objects. J Vis, 3(6), 413-422

Which level of categorization is the right one?

Car is an object composed of:

a few doors, four wheels (not all visible at all times), a roof, front lights, windshield

If you are thinking in buying a car, you might want to be a bit more specific about your categorization.

Entry-level categories

(Jolicoeur, Gluck, Kosslyn 1984)

- Typical member of a basic-level category are categorized at the expected level
- Atypical members tend to be classified at a subordinate level.

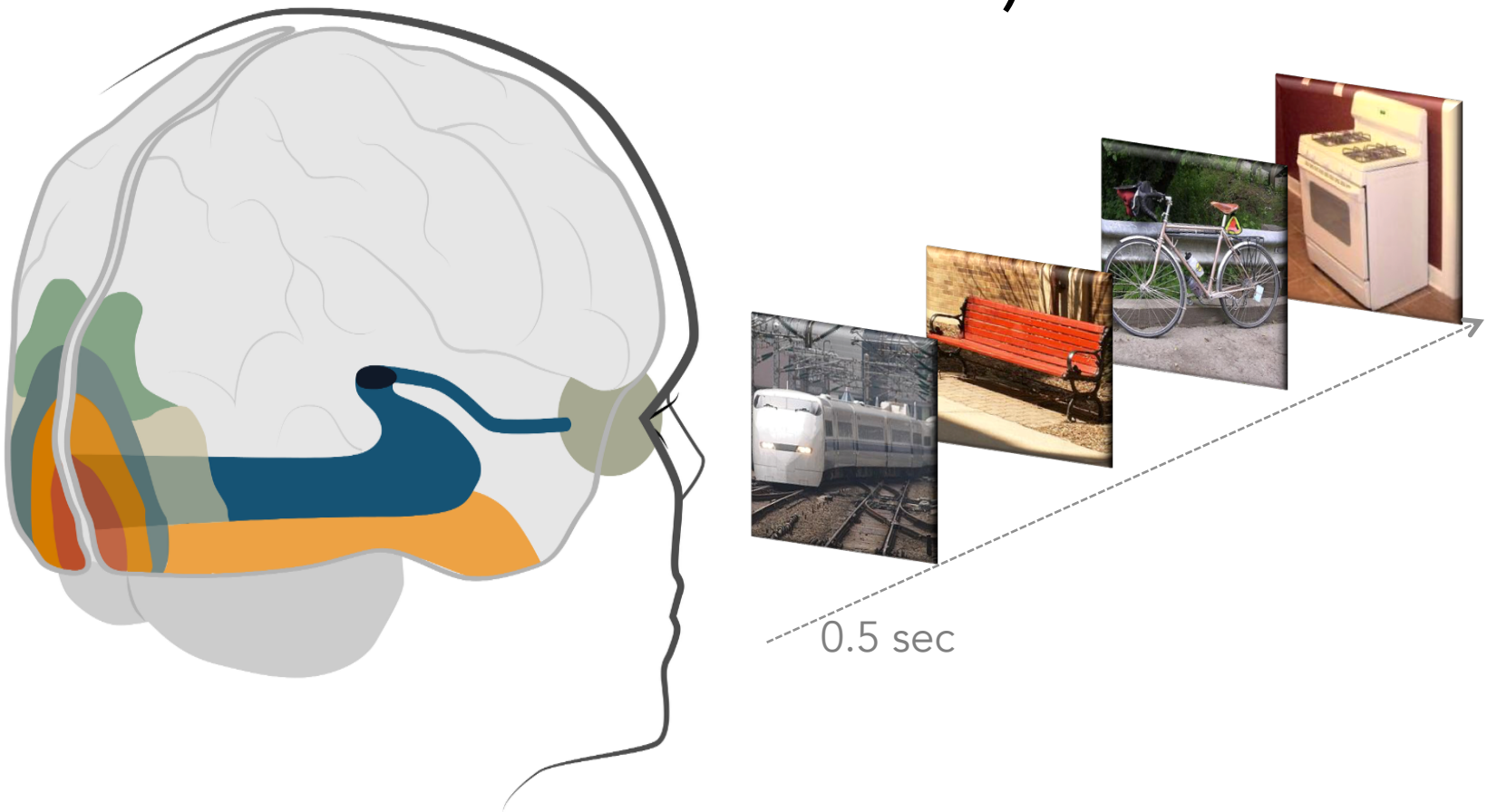
A bird

An ostrich

Creation of new categories

A new class can borrow information from similar categories

High-Powered Machine:
finding objects in context (visual search)



Even when objects are not there…

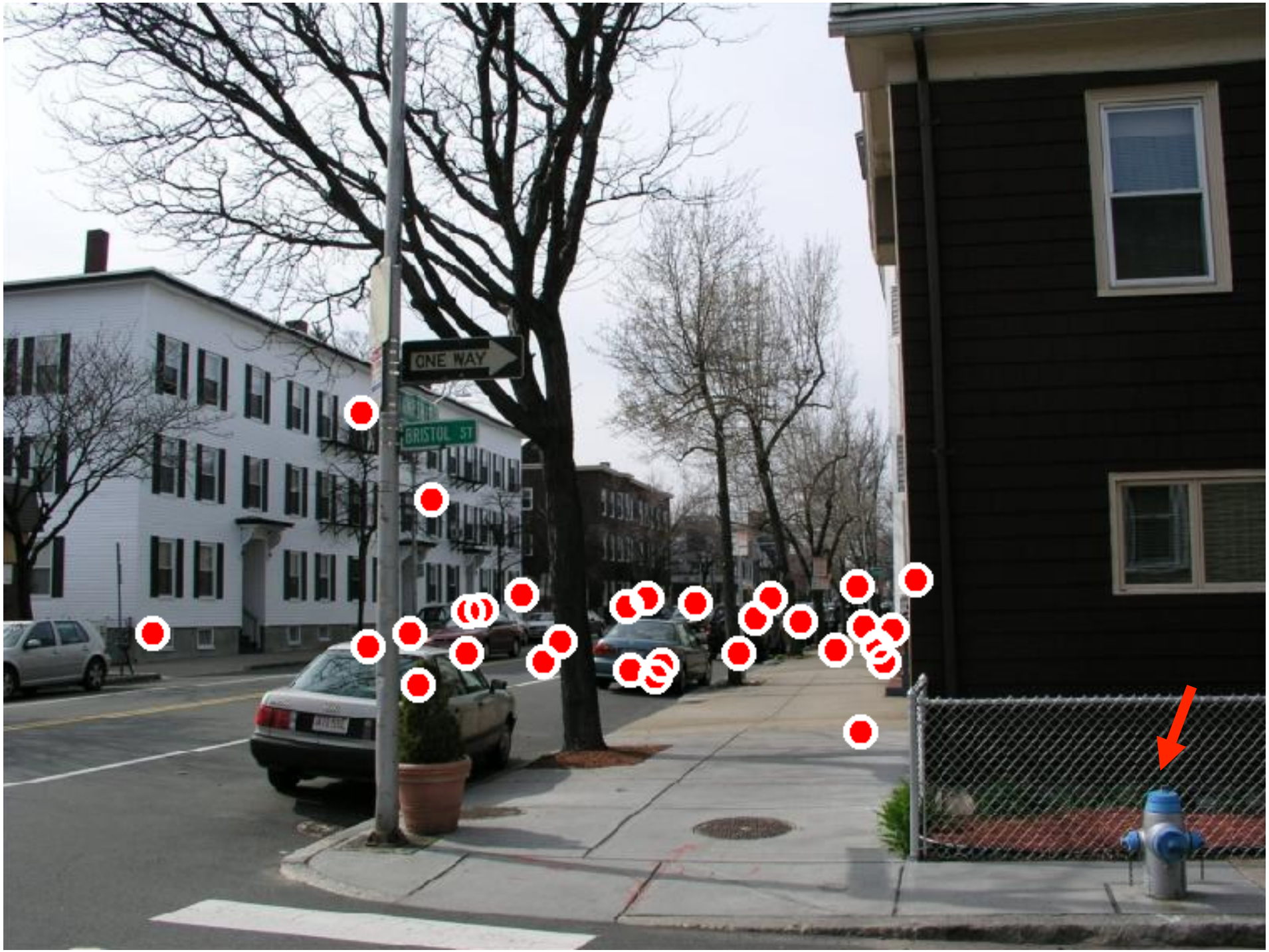
Look for a person in the next scene

+


ONE WAY
BRISTOL ST

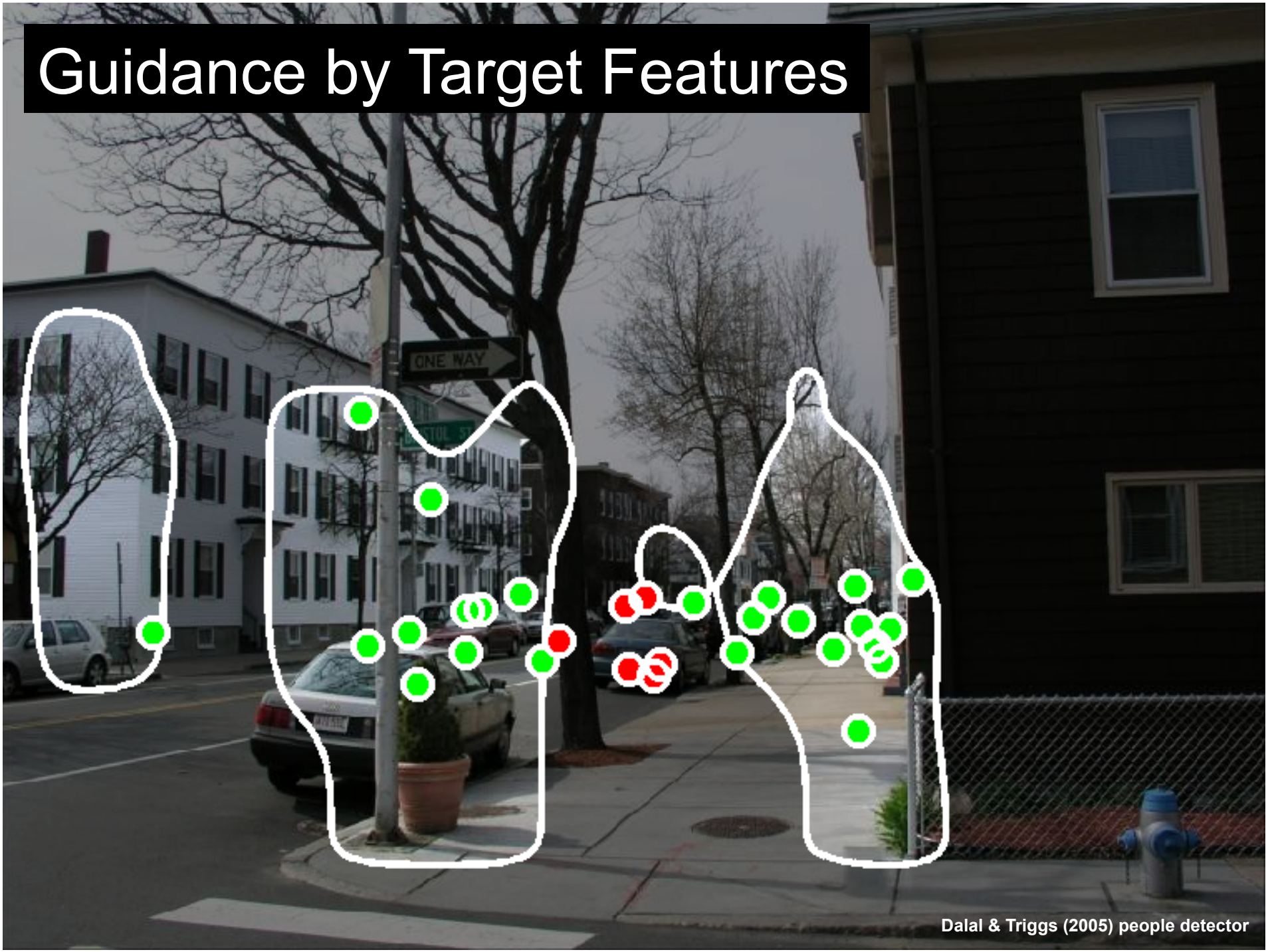
I know where you looked


ONE WAY
BRISTOL ST

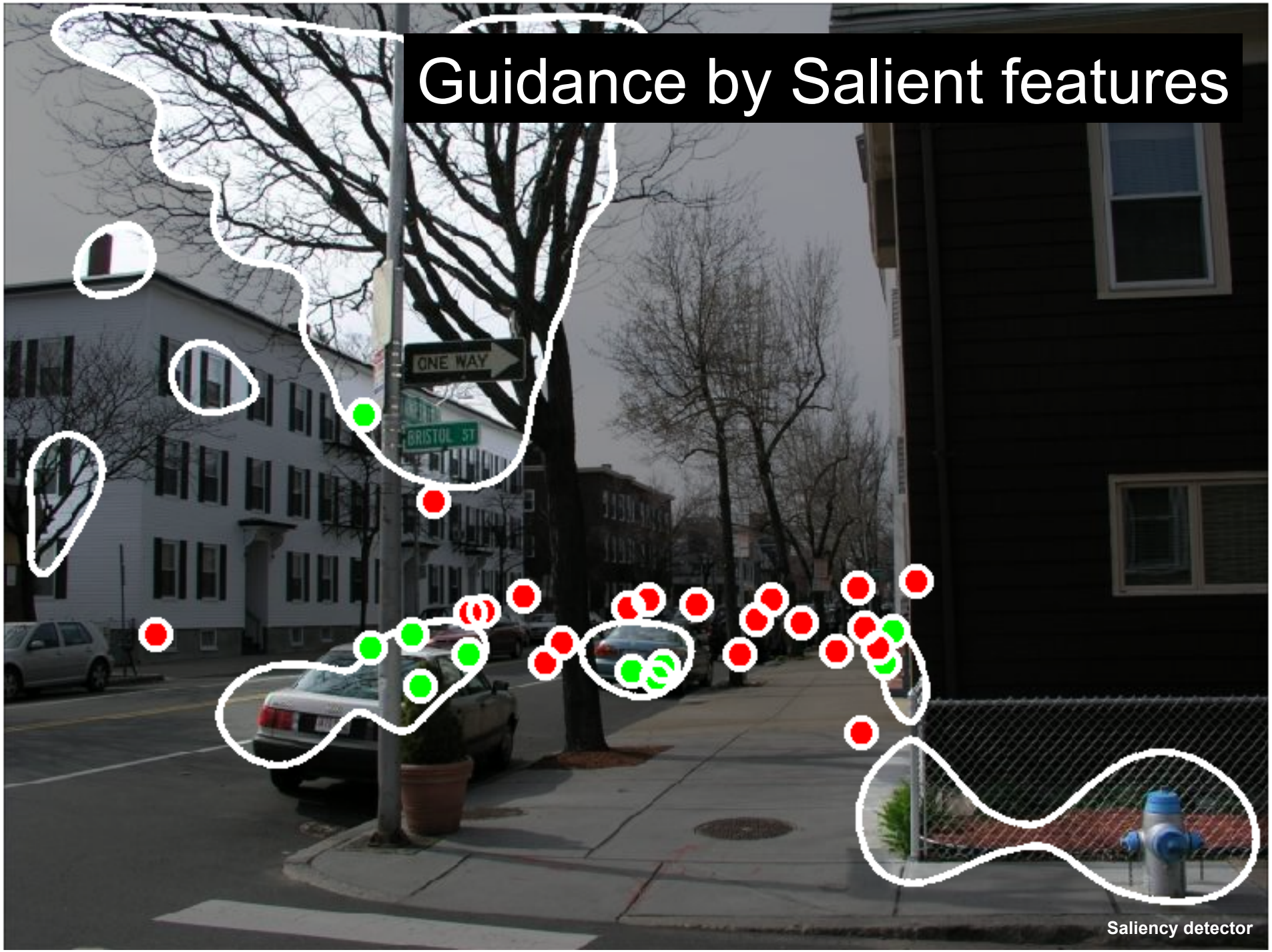

ONE WAY
BRISTOL ST


ONE WAY
BRISTOL ST

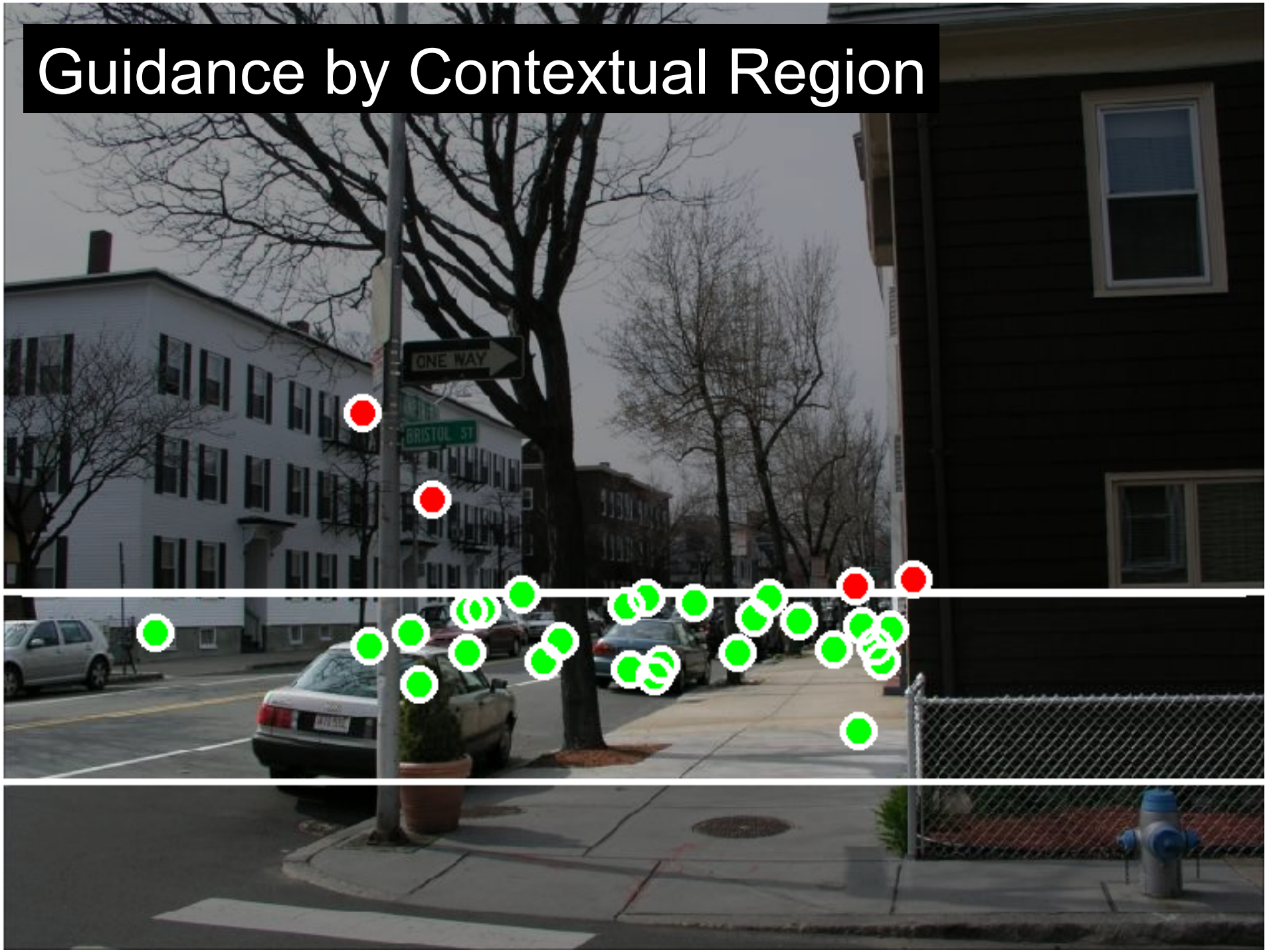
Guidance by Target Features

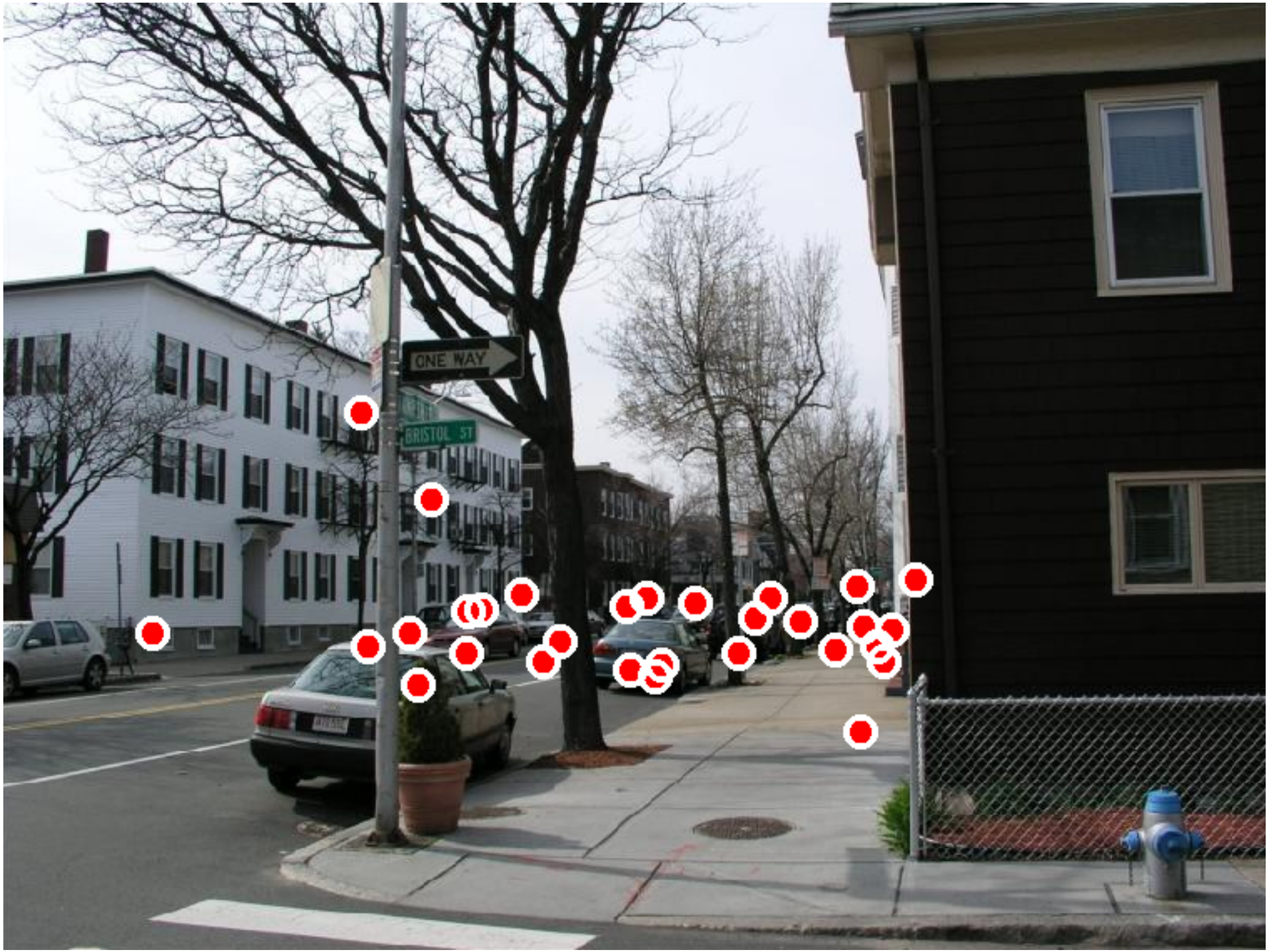


Dalal & Triggs (2005) people detector


Guidance by Salient features
ONE WAY
BRISTOL ST
Saliency detector

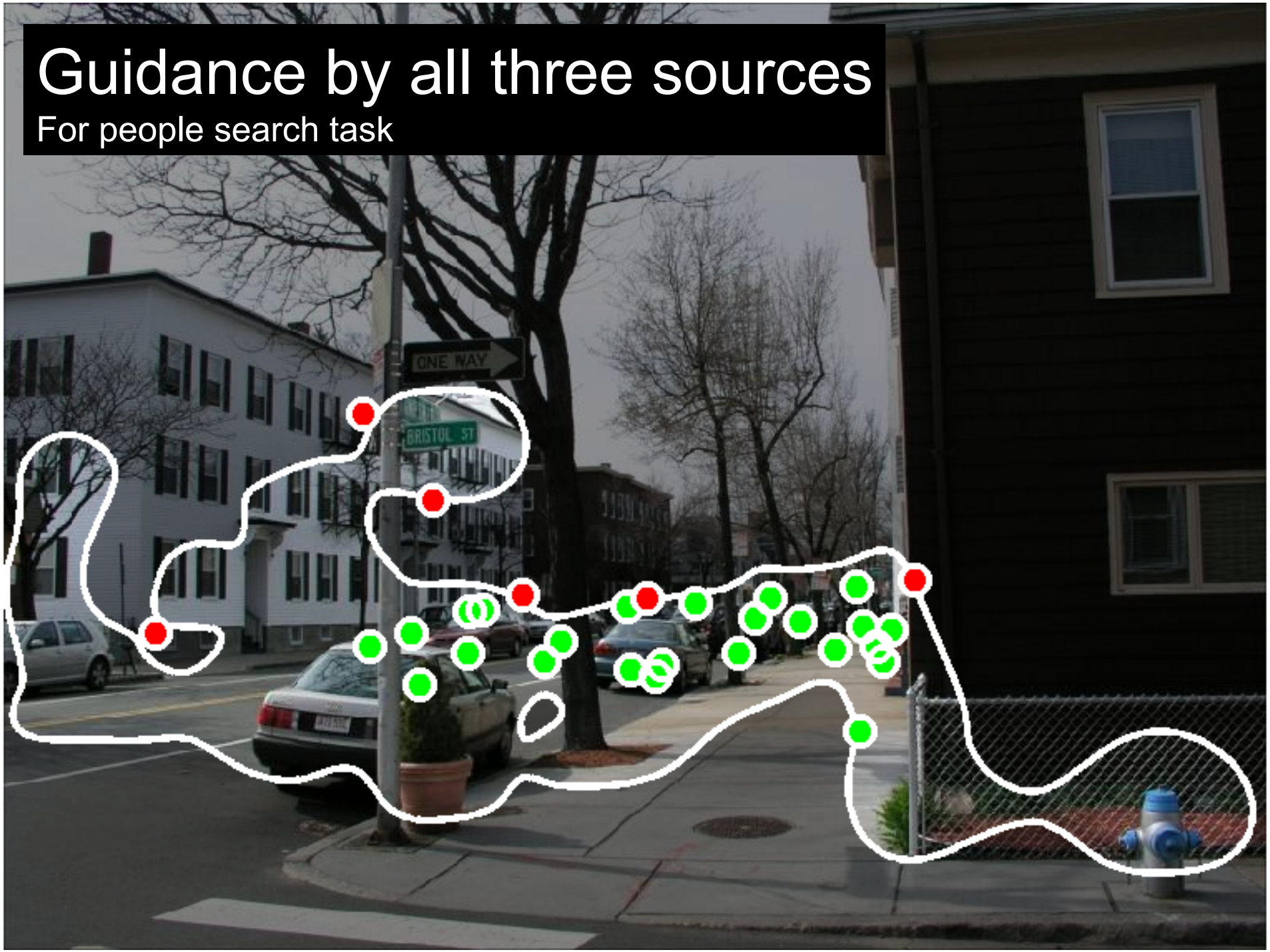
Guidance by Contextual Region

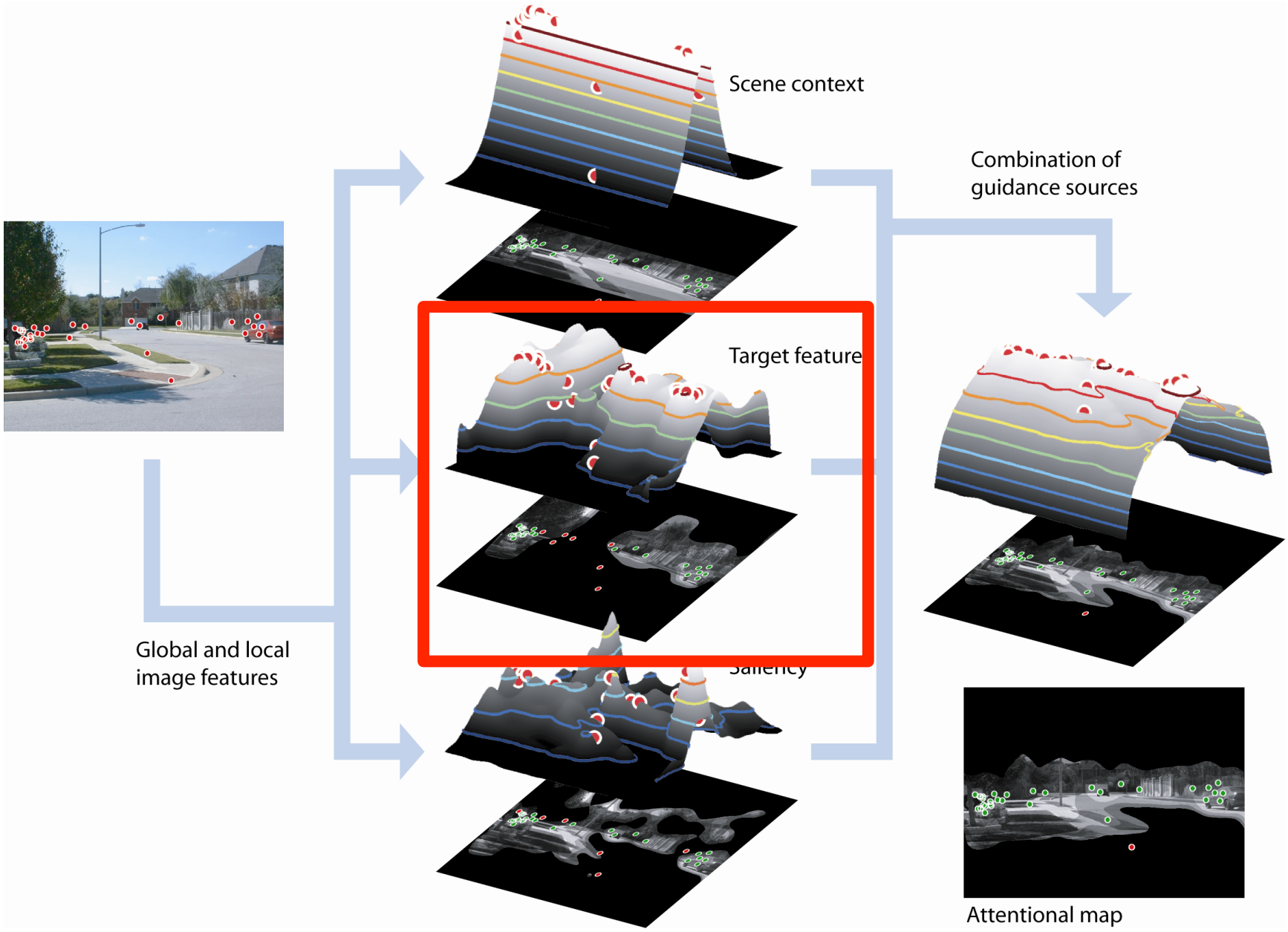



ONE WAY
BRISTOL ST

Guidance by all three sources

For people search task

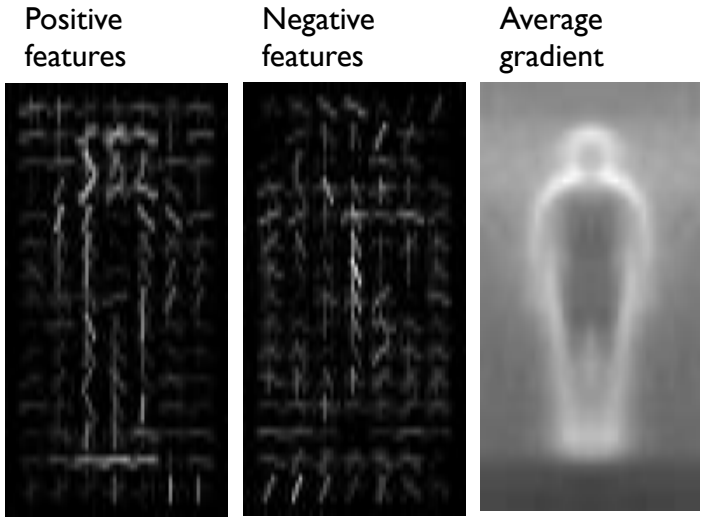




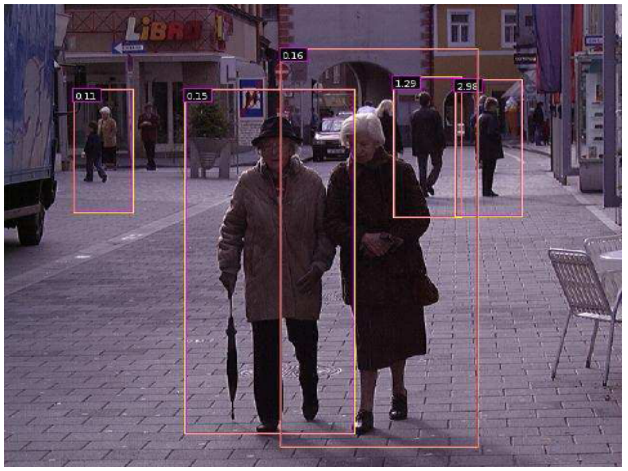
Ehinger et al (2009)

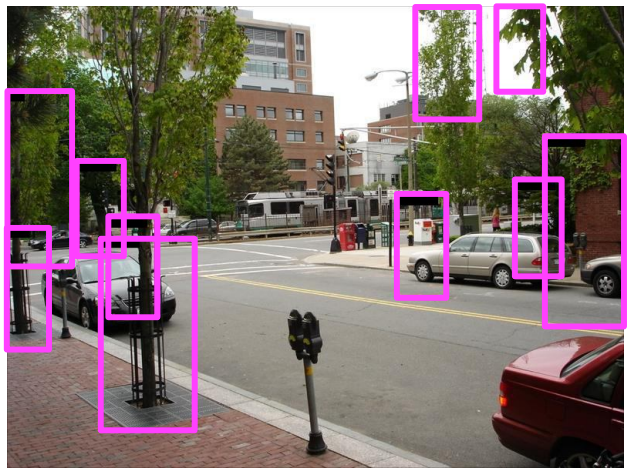
Object (target) features Model

- Dalal & Triggs (2005) detector uses histograms of oriented gradients (HOG)



Dalal & Triggs (2005)





IMAGENET **Large Scale Visual Recognition Challenge**

Steel drum

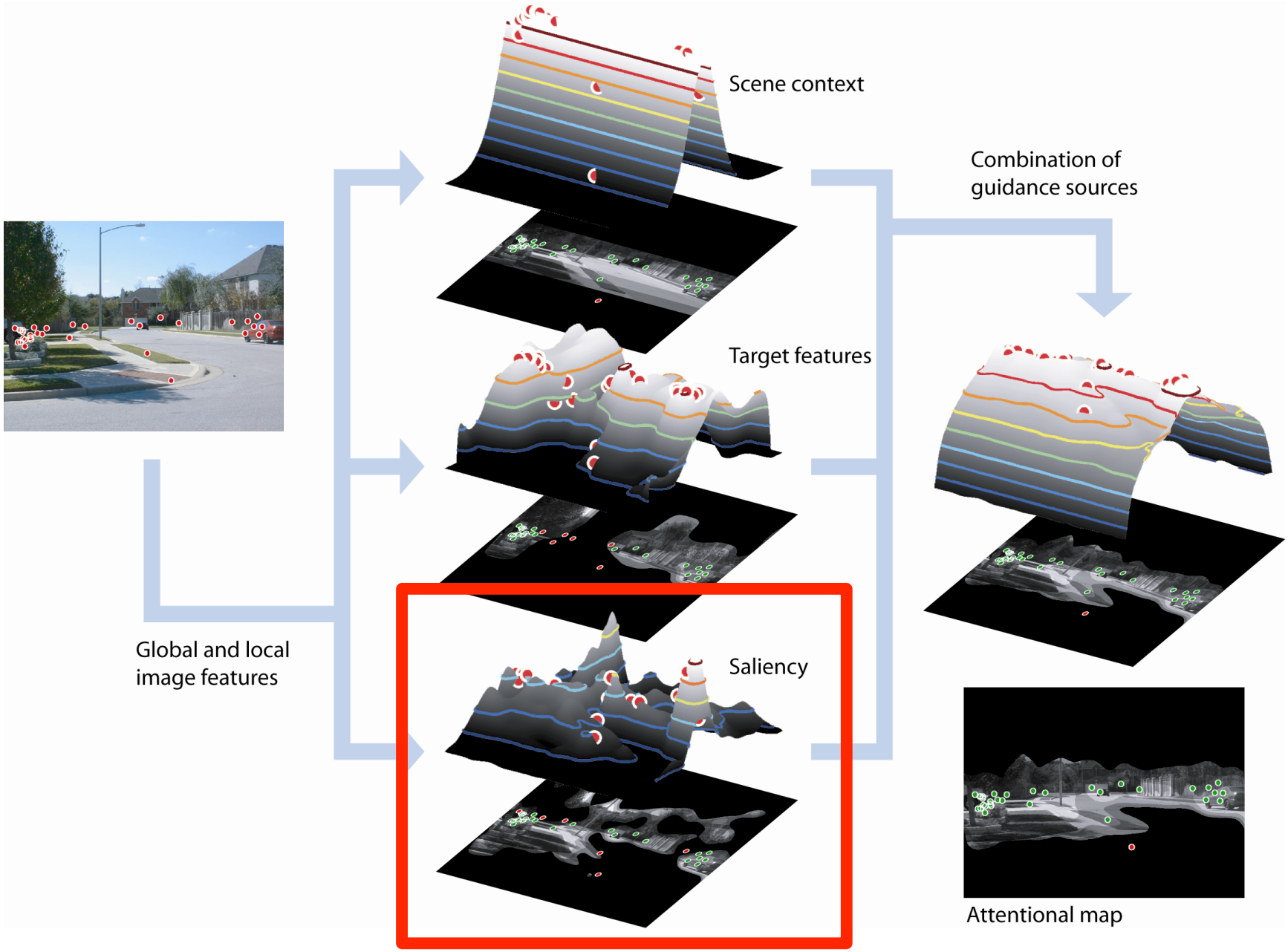
The Image Classification Challenge:
1,000 object classes
1,431,167 images

Top 5 categories

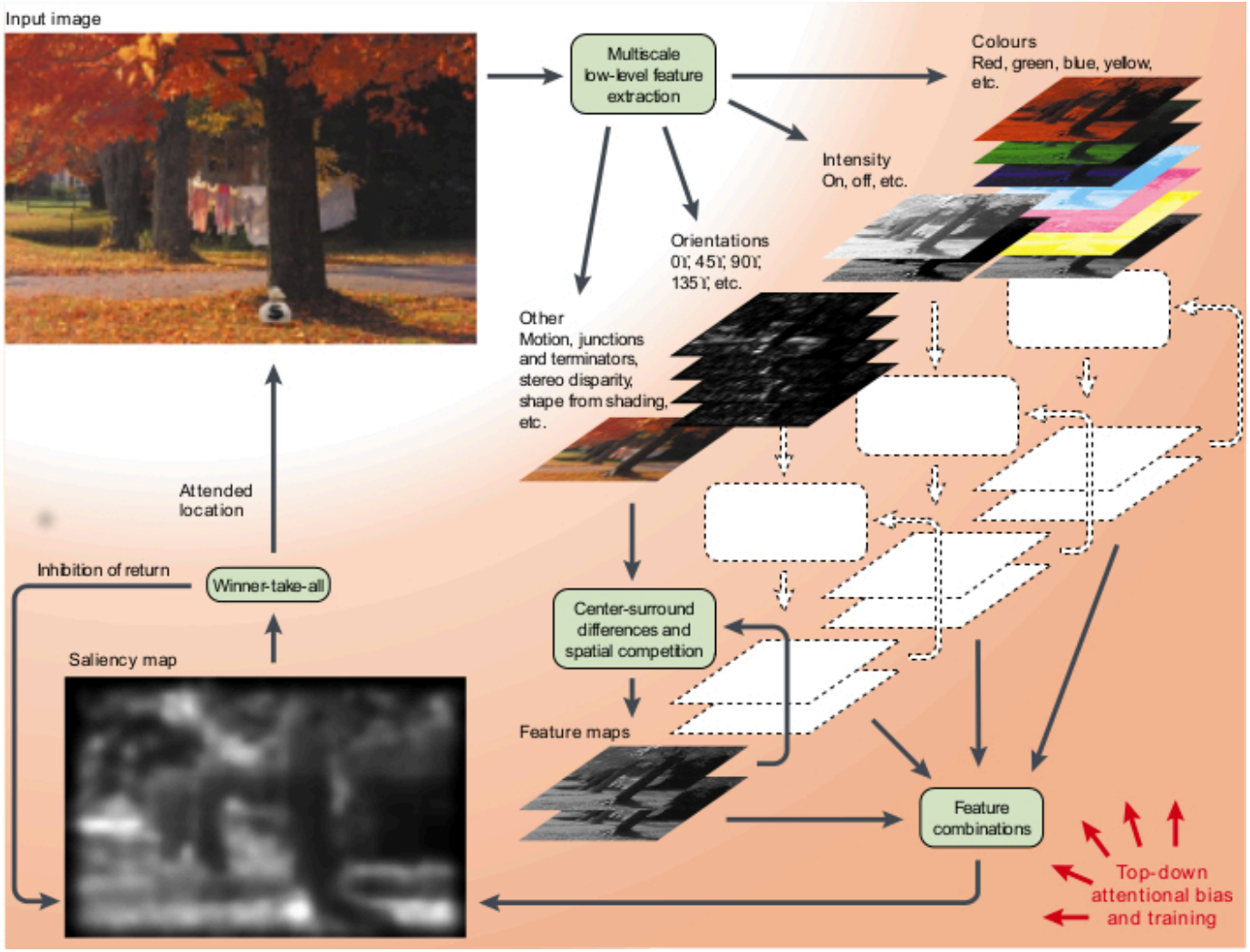
Output:		Output:	
Scale		Scale	
T-shirt		T-shirt	
Steel drum	✔	Giant panda	✗
Drumstick		Drumstick	
Mud turtle		Mud turtle	

Russakovsky et al. arXiv, 2014

Slide from Fei-Fei Li & Andrej Karpathy

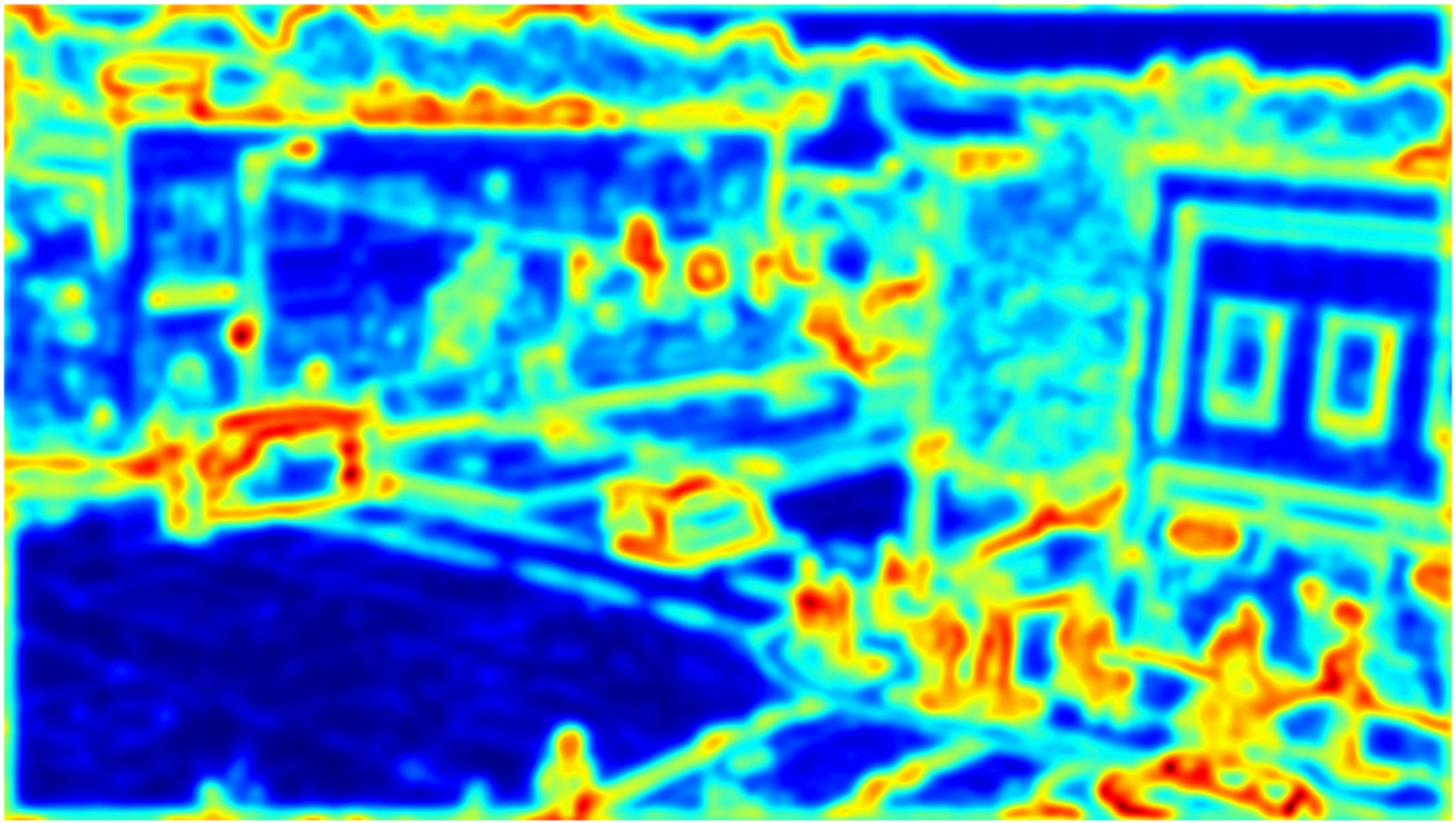

Scene context
Combination of
guidance sources
Target features
Global and local
image features
Saliency
Attentional map

First Saliency Model: Itti-Koch 2001

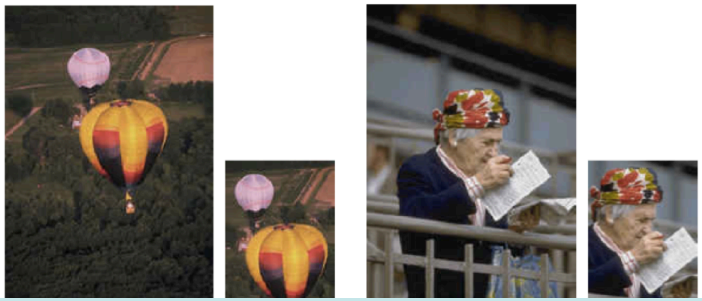


Application: image processing

Application: image processing



Application: image processing



Automatic cropping

Steniford, ICVS 2007

Automatic thumbnailing

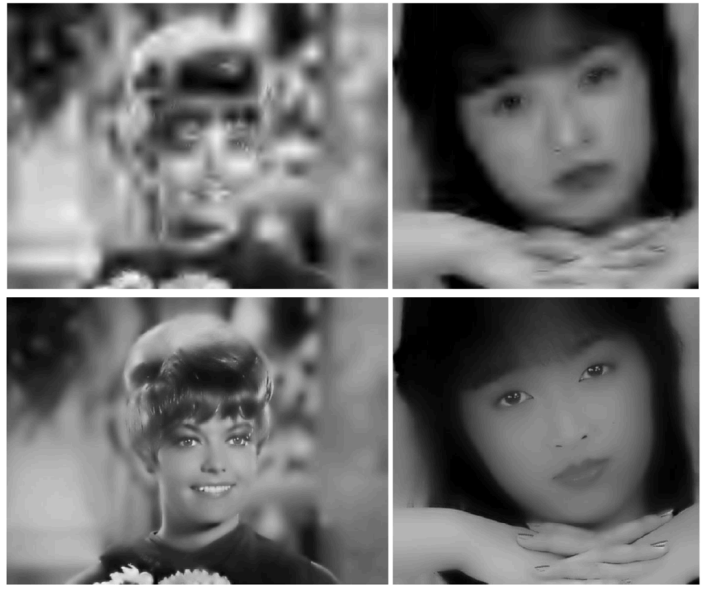
Marchesotti et al., ICCV 2009

Image re-targeting

Goferman, Zelnik-Manor, Tal, 2010, 2012

Photo collaging

Artistic effects



Smarter compression

mit saliency benchmark

home | **results** | datasets | submission | downloads

mit300 | cat2000

mit saliency benchmark results: mit300

The following are results of models evaluated on their ability to predict ground truth human fixations on our **benchmark data set containing 300 natural images with eye tracking data from 39 observers.** We post the results here and provide a way for people to submit new models for evaluation.

citations

If you use any of the results or data on this page, please cite the following:

```
@misc{mit-saliency-benchmark,
  author    = {Zoya Bylinskii and Tilke Judd and Ali Borji and Laurent Itti and Fr{\'e}do Durand and Aude Oliva and Antonio Torralba},
  title     = {MIT Saliency Benchmark},
}
```

This dataset is released in conjunction to the paper **"A Benchmark of Computational Models of Saliency to Predict Human Fixations"** by Tilke Judd, Fredo Durand and Antonio Torralba, available as a Jan 2012 MIT tech report.

```
@InProceedings{Judd_2012,
  author    = {Tilke Judd and Fr{\'e}do Durand and Antonio Torralba},
  title     = {A Benchmark of Computational Models of Saliency to Predict Human Fixations},
  booktitle = {MIT Technical Report},
  year      = {2012}
}
```

images

300 benchmark images (the fixations from 39 viewers per image are not public such that no model can be trained using this data set).

model performances

47 models, **5** baselines, **7** metrics, and counting...

Performance numbers prior to September 25, 2014.

Matlab code for the metrics we use.

Sorted by: AUC-Judd metric

Model Name	Published	Code	AUC-Judd [?]	SIM [?]	EMD [?]	AUC-Borji [?]	sAUC [?]	CC [?]	NSS [?]	Date tested [key]	Sample [img]
Baseline: infinite humans [?]			0.91	1	0	0.87	0.80	1	3.18		
SALICON	Xun Huang, Chengyao Shen, Xavier Boix, Qi Zhao		0.87	0.60	2.62	0.85	0.74	0.74	2.12	first tested: 19/11/2014 last tested: 20/03/2015 maps from authors	
DeepFix	Srinivas S S Kruthiventi, Kumar Ayush, R. Venkatesh Babu **DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations [arXiv 2015]**		0.87	0.67	2.04	0.80	0.71	0.78	2.26	first tested: 10/02/2015 last tested: 10/02/2015 maps from authors	
Deep Gaze 1	Matthias Kümmerer, Lucas Theis, Matthias Bethge. **Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet [arxiv 2014]**		0.84	0.39	4.97	0.83	0.66	0.48	1.22	first tested: 02/10/2014 last tested: 22/10/2014 maps from authors	
Boolean Map based Saliency (BMS)	Jianming Zhang, Stan Sclaroff. **Saliency detection: a boolean map approach [ICCV 2013]**	matlab, executable	0.83	0.51	3.35	0.82	0.65	0.55	1.41	first tested: 14/05/2014 last tested: 23/09/2014 maps from authors	
SalNet	Kevin McGuinness. Unpublished work.		0.83	0.52	3.31	0.82	0.69	0.58	1.51	first tested: 17/06/2015 last tested: 17/06/2015 maps from authors	
Mixture of Saliency Models	Xuehua Han, Shunji Satoh. "Unifying computational models for visual attention" [AINI 2014, Sep. (accepted)]		0.82	0.44	4.22	0.81	0.62	0.52	1.34	first tested: 08/08/2014 last tested: 23/09/2014 maps from authors	
Ensembles of Deep Networks (eDN)	Eleonora Vig, Michael Dorr, David Cox. **Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images [CVPR 2014]**	python	0.82	0.41	4.56	0.81	0.62	0.45	1.14	first tested: 16/08/2014 last tested: 23/09/2014 maps from authors	
Outlier Saliency (OS)	Chuanbo Chen, He Tang, Zehua Lyu, Hu Liang, Jun Shang, Mudar Serem. Saliency Modeling via Outlier Detection [Journal of Electronic Imaging]. Accepted, 2014.		0.82	0.50	3.33	0.81	0.62	0.54	1.38	first tested: 16/09/2014 last tested: 23/09/2014 maps from authors	
Judd Model	Tilke Judd, Krista Ehinger, Fredo Durand, Antonio Torralba. **Learning to predict where humans look [ICCV 2009]**	matlab	0.81	0.42	4.45	0.80	0.60	0.47	1.18	last tested: 23/09/2014 maps from code (DL:17/12/2013) with default params	
CovSal	Erkut Erdem, Aykut Erdem. **Visual saliency estimation by nonlinearly integrating features using region covariances [JoV 2013]**	matlab	0.81	0.47	3.39	0.67	0.57	0.45	1.22	first tested: 05/02/2012 last tested: 23/09/2014 maps from authors	
Graph-Based Visual Saliency (GBVS)	Jonathan Harel, Christof Koch, Pietro Perona. **Graph-Based Visual Saliency [NIPS 2006]**	matlab	0.81	0.48	3.51	0.80	0.63	0.48	1.24	last tested: 23/09/2014 maps from code (DL:20/08/2013) with default params	
Spatially Weighted Dissimilarity	Lijuan Duan, Chunpeng Wu, Jun Miao, Laiyun Qing, Yu Fu. **Visual Saliency Detection by Spatially Weighted**	matlab	0.81	0.46	3.89	0.80	0.59	0.49	1.27	first tested: 09/22/2014 last tested: 29/09/2014	

MIT Saliency Benchmark: keeps track of the current state-of-the-art models of saliency

mit saliency benchmark

home | **results** | datasets | submission | downloads

mit300 | cat2000

mit saliency benchmark results: mit300

The following are results of models evaluated on their ability to predict ground truth human fixations on our **benchmark data set containing 300 natural images with eye tracking data from 39 observers.** We post the results here and provide a way for people to submit new models for evaluation.

citations

If you use any of the results or data on this page, please cite the following:

```
@misc{mit-saliency-benchmark,
  author    = {Zoya Bylinskii and Tilke Judd and Ali Borji and Laurent Itti and Fr{\'e}do Durand and Aude Oliva and Antonio Torralba},
  title     = {MIT Saliency Benchmark},
}
```

This dataset is released in conjunction to the paper **"A Benchmark of Computational Models of Saliency to Predict Human Fixations"** by Tilke Judd, Fredo Durand and Antonio Torralba, available as a Jan 2012 MIT tech report.

```
@InProceedings{Judd_2012,
  author    = {Tilke Judd and Fr{\'e}do Durand and Antonio Torralba},
  title     = {A Benchmark of Computational Models of Saliency to Predict Human Fixations},
  booktitle = {MIT Technical Report},
  year      = {2012}
}
```

images

300 benchmark images (the fixations from 39 viewers per image are not public such that no model can be trained using this data set).

model performances

47 models, **5** baselines, **7** metrics, and counting...

Performance numbers prior to September 25, 2014.

Matlab code for the metrics we use.

Sorted by: AUC-Judd metric

Model Name	Published	Code	AUC-Judd [?]	SIM [?]	EMD [?]	AUC-Borji [?]	sAUC [?]	CC [?]	NSS [?]	Date tested [key]	Sample [img]
Baseline: infinite humans [?]			0.91	1	0	0.87	0.80	1	3.18		
SALICON	Xun Huang, Chengyao Shen, Xavier Boix, Qi Zhao		0.87	0.60	2.62	0.85	0.74	0.74	2.12	first tested: 19/11/2014 last tested: 20/03/2015 maps from authors	
DeepFix	Srinivas S S Kruthiventi, Kumar Ayush, R. Venkatesh Babu **DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations [arXiv 2015]**		0.87	0.67	2.04	0.80	0.71	0.78	2.26	first tested: 10/02/2015 last tested: 10/02/2015 maps from authors	
Deep Gaze 1	Matthias Kümmerer, Lucas Theis, Matthias Bethge. **Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet [arxiv 2014]**		0.84	0.39	4.97	0.83	0.66	0.48	1.22	first tested: 02/10/2014 last tested: 22/10/2014 maps from authors	
Boolean Map based Saliency (BMS)	Jianming Zhang, Stan Sclaroff. **Saliency detection: a boolean map approach [ICCV 2013]**	matlab, executable	0.83	0.51	3.35	0.82	0.65	0.55	1.41	first tested: 14/05/2014 last tested: 23/09/2014 maps from authors	
SalNet	Kevin McGuinness. Unpublished work.		0.83	0.52	3.31	0.82	0.69	0.58	1.51	first tested: 17/06/2015 last tested: 17/06/2015 maps from authors	
Mixture of Saliency Models	Xuehua Han, Shunji Satoh. "Unifying computational models for visual attention" [AINI 2014, Sep. (accepted)]		0.82	0.44	4.22	0.81	0.62	0.52	1.34	first tested: 08/08/2014 last tested: 23/09/2014 maps from authors	
Ensembles of Deep Networks (eDN)	Eleonora Vig, Michael Dorr, David Cox. **Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images [CVPR 2014]**	python	0.82	0.41	4.56	0.81	0.62	0.45	1.14	first tested: 16/08/2014 last tested: 23/09/2014 maps from authors	
Outlier Saliency (OS)	Chuanbo Chen, He Tang, Zehua Lyu, Hu Liang, Jun Shang, Mudar Serem. Saliency Modeling via Outlier Detection [Journal of Electronic Imaging]. Accepted, 2014.		0.82	0.50	3.33	0.81	0.62	0.54	1.38	first tested: 16/09/2014 last tested: 23/09/2014 maps from authors	
Judd Model	Tilke Judd, Krista Ehinger, Fredo Durand, Antonio Torralba. **Learning to predict where humans look [ICCV 2009]**	matlab	0.81	0.42	4.45	0.80	0.60	0.47	1.18	last tested: 23/09/2014 maps from code (DL:17/12/2013) with default params	
CovSal	Erkut Erdem, Aykut Erdem. **Visual saliency estimation by nonlinearly integrating features using region covariances [JoV 2013]**	matlab	0.81	0.47	3.39	0.67	0.57	0.45	1.22	first tested: 05/02/2012 last tested: 23/09/2014 maps from authors	
Graph-Based Visual Saliency (GBVS)	Jonathan Harel, Christof Koch, Pietro Perona. **Graph-Based Visual Saliency [NIPS 2006]**	matlab	0.81	0.48	3.51	0.80	0.63	0.48	1.24	last tested: 23/09/2014 maps from code (DL:20/08/2013) with default params	
Spatially Weighted Dissimilarity	Lijuan Duan, Chunpeng Wu, Jun Miao, Laiyun Qing, Yu Fu. **Visual Saliency Detection by Spatially Weighted**	matlab	0.81	0.46	3.89	0.80	0.59	0.49	1.27	first tested: 09/22/2014 last tested: 29/09/2014	

MIT Saliency Benchmark: keeps track of the current state-of-the-art models of saliency

current top players: neural network (NN) based models submitted in the last year

mit saliency benchmark

home | **results** | datasets | submission | downloads

mit300 | cat2000

mit saliency benchmark results: mit300

The following are results of models evaluated on their ability to predict ground truth human fixations on our **benchmark data set containing 300 natural images with eye tracking data from 39 observers.** We post the results here and provide a way for people to submit new models for evaluation.

citations

If you use any of the results or data on this page, please cite the following:

```
@misc{mit-saliency-benchmark,
  author    = {Zoya Bylinskii and Tilke Judd and Ali Borji and Laurent Itti and Fr{\'e}do Durand and Aude Oliva and Antonio Torralba},
  title     = {MIT Saliency Benchmark},
}
```

This dataset is released in conjunction to the paper **"A Benchmark of Computational Models of Saliency to Predict Human Fixations"** by Tilke Judd, Fredo Durand and Antonio Torralba, available as a Jan 2012 MIT tech report.

```
@InProceedings{Judd_2012,
  author    = {Tilke Judd and Fr{\'e}do Durand and Antonio Torralba},
  title     = {A Benchmark of Computational Models of Saliency to Predict Human Fixations},
  booktitle = {MIT Technical Report},
  year      = {2012}
}
```

images

300 benchmark images (the fixations from 39 viewers per image are not public such that no model can be trained using this data set).

model performances

47 models, **5** baselines, **7** metrics, and counting...

Performance numbers prior to September 25, 2014.

Matlab code for the metrics we use.

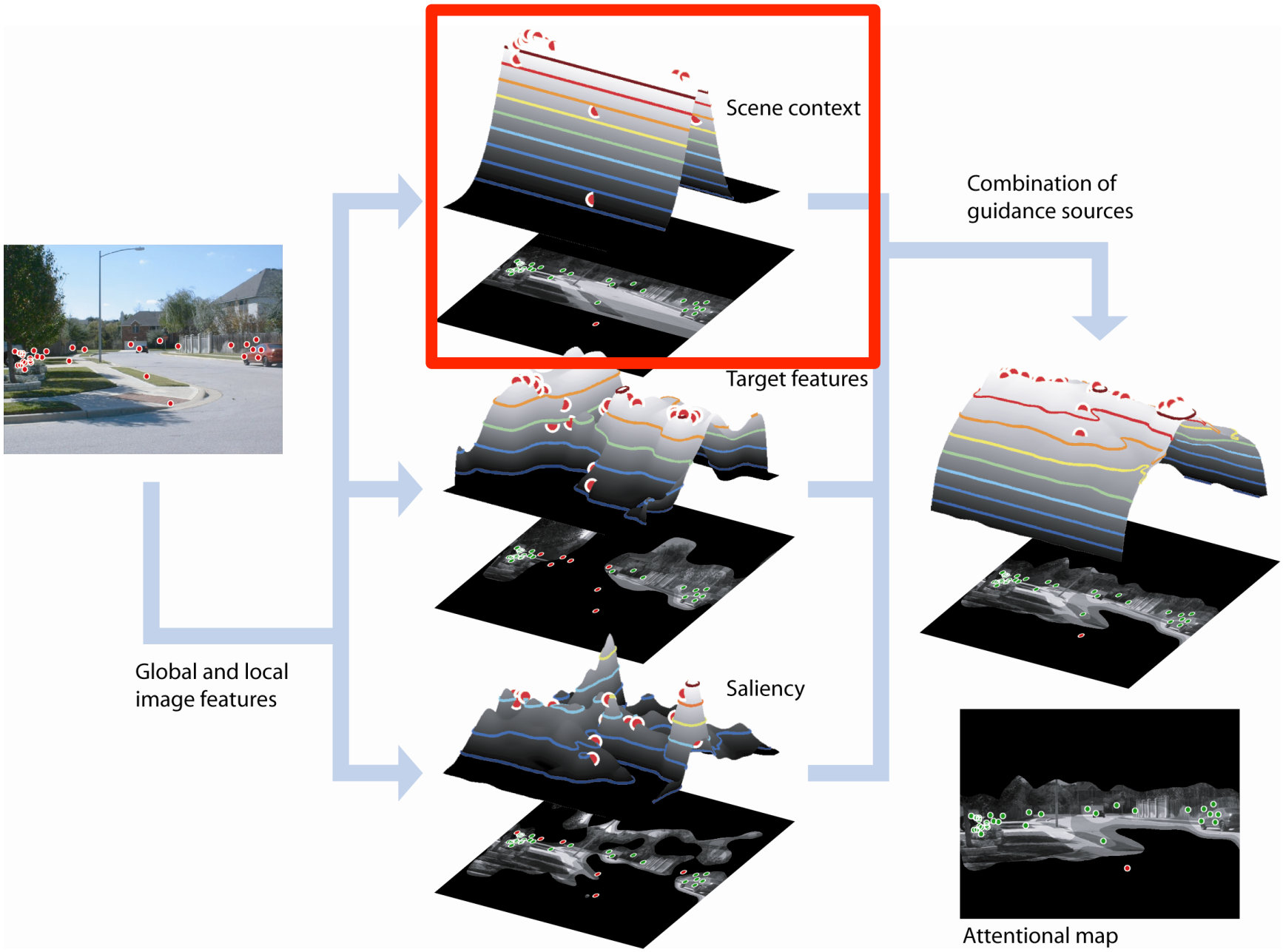
Sorted by: AUC-Judd metric

Model Name	Published	Code	AUC-Judd [?]	SIM [?]	EMD [?]	AUC-Borji [?]	sAUC [?]	CC [?]	NSS [?]	Date tested [key]	Sample [img]
Baseline: infinite humans [?]			0.91	1	0	0.87	0.80	1	3.18		
SALICON	Xun Huang, Chengyao Shen, Xavier Boix, Qi Zhao		0.87	0.60	2.62	0.85	0.74	0.74	2.12	first tested: 19/11/2014 last tested: 20/03/2015 maps from authors	
DeepFix	Srinivas S S Kruthiventi, Kumar Ayush, R. Venkatesh Babu **DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations [arXiv 2015]**		0.87	0.67	2.04	0.80	0.71	0.78	2.26	first tested: 10/02/2015 last tested: 10/02/2015 maps from authors	
Deep Gaze 1	Matthias Kümmerer, Lucas Theis, Matthias Bethge. **Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet [arxiv 2014]**		0.84	0.39	4.97	0.83	0.66	0.48	1.22	first tested: 02/10/2014 last tested: 22/10/2014 maps from authors	
Boolean Map based Saliency (BMS)	Jianming Zhang, Stan Sclaroff. **Saliency detection: a boolean map approach [ICCV 2013]**	matlab, executable	0.83	0.51	3.35	0.82	0.65	0.55	1.41	first tested: 14/05/2014 last tested: 23/09/2014 maps from authors	
SalNet	Kevin McGuinness. Unpublished work.		0.83	0.52	3.31	0.82	0.69	0.58	1.51	first tested: 17/06/2015 last tested: 17/06/2015 maps from authors	
Mixture of Saliency Models	Xuehua Han, Shunji Satoh. "Unifying computational models for visual attention" [AINI 2014, Sep. (accepted)]		0.82	0.44	4.22	0.81	0.62	0.52	1.34	first tested: 08/08/2014 last tested: 23/09/2014 maps from authors	
Ensembles of Deep Networks (eDN)	Eleonora Vig, Michael Dorr, David Cox. **Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images [CVPR 2014]**	python	0.82	0.41	4.56	0.81	0.62	0.45	1.14	first tested: 16/08/2014 last tested: 23/09/2014 maps from authors	
Outlier Saliency (OS)	Chuanbo Chen, He Tang, Zehua Lyu, Hu Liang, Jun Shang, Mudar Serem. Saliency Modeling via Outlier Detection [Journal of Electronic Imaging]. Accepted, 2014.		0.82	0.50	3.33	0.81	0.62	0.54	1.38	first tested: 16/09/2014 last tested: 23/09/2014 maps from authors	
Judd Model	Tilke Judd, Krista Ehinger, Fredo Durand, Antonio Torralba. **Learning to predict where humans look [ICCV 2009]**	matlab	0.81	0.42	4.45	0.80	0.60	0.47	1.18	last tested: 23/09/2014 maps from code (DL:17/12/2013) with default params	
CovSal	Erkut Erdem, Aykut Erdem. **Visual saliency estimation by nonlinearly integrating features using region covariances [JoV 2013]**	matlab	0.81	0.47	3.39	0.67	0.57	0.45	1.22	first tested: 05/02/2012 last tested: 23/09/2014 maps from authors	
Graph-Based Visual Saliency (GBVS)	Jonathan Harel, Christof Koch, Pietro Perona. **Graph-Based Visual Saliency [NIPS 2006]**	matlab	0.81	0.48	3.51	0.80	0.63	0.48	1.24	last tested: 23/09/2014 maps from code (DL:20/08/2013) with default params	
Spatially Weighted Dissimilarity	Lijuan Duan, Chunpeng Wu, Jun Miao, Laiyun Qing, Yu Fu. **Visual Saliency Detection by Spatially Weighted**	matlab	0.81	0.46	3.89	0.80	0.59	0.49	1.27	first tested: 09/22/2014 last tested: 29/09/2014	

Approaching human performance!

Making a lot of new applications possible.

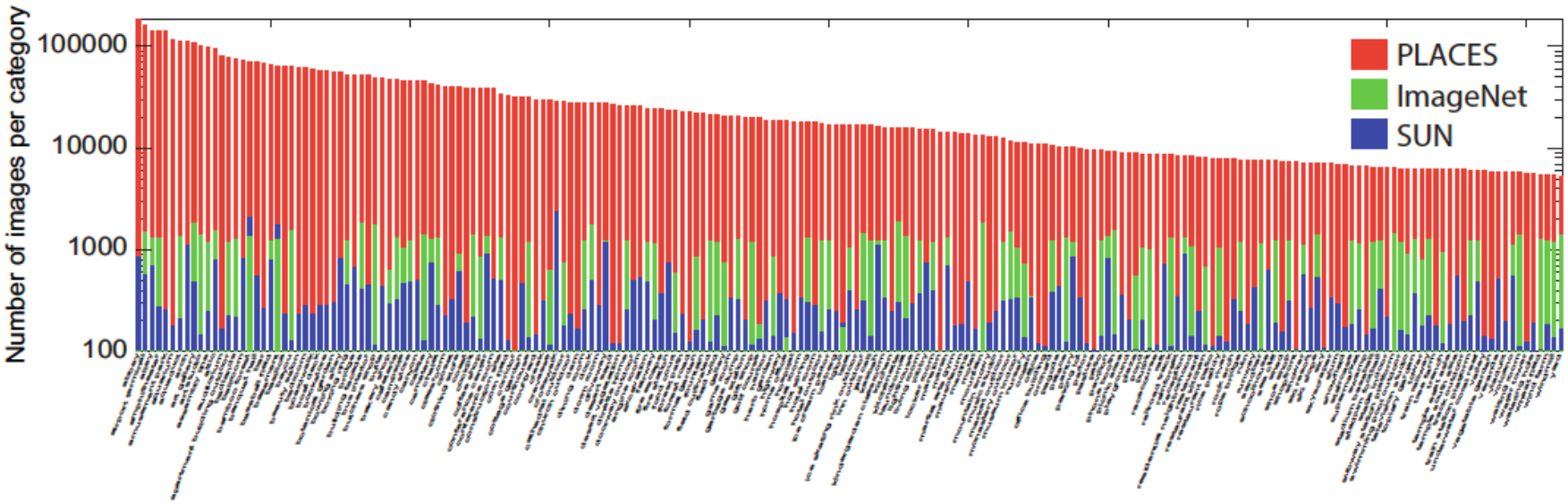
current top players: neural network (NN) based models submitted in the last year


Scene context
Target features
Saliency
Global and local
image features
Combination of
guidance sources
Attentional map

places

400 Categories, 10 M images

places.csail.mit.edu



Predictions:

- **type:** indoor
- **semantic categories:** coffee_shop:0.47, restaurant:0.17, cafeteria:0.08, food_court:0.06,

Predictions:

- **type:** indoor
- **semantic categories:** supermarket:0.96,

Predictions:

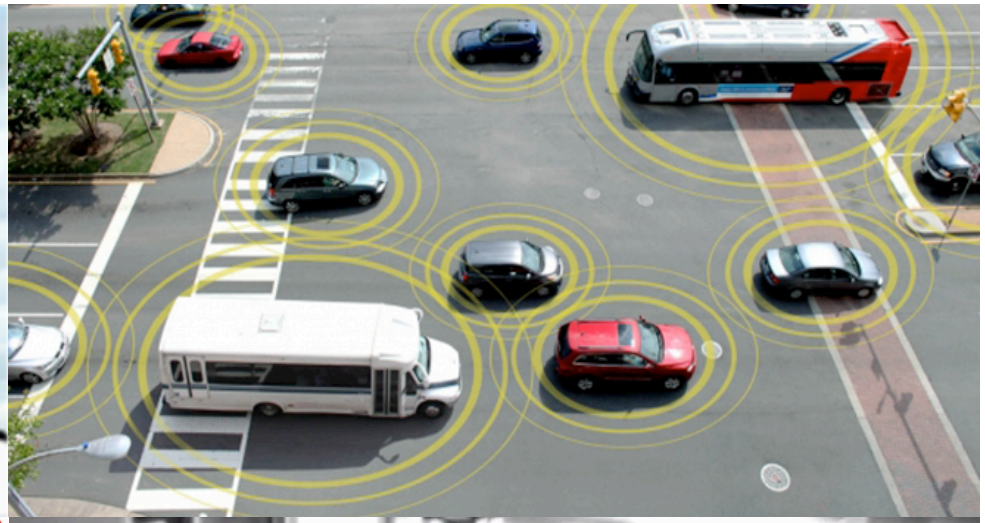
- **type:** indoor
- **semantic categories:** conference_center:0.51, auditorium:0.12, office:0.08,

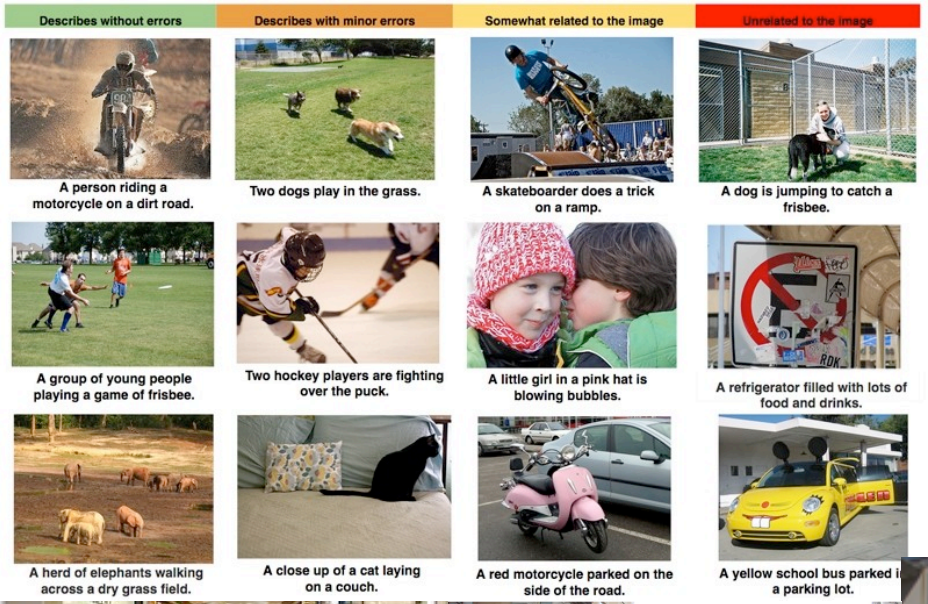
Predictions:

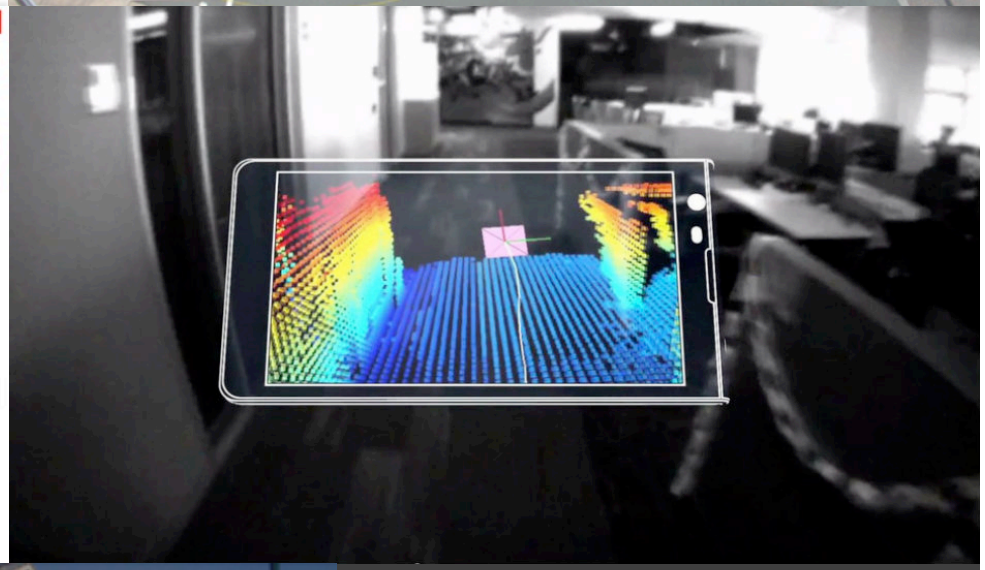
- **type:** indoor
- **semantic categories:** bus_interior:0.91,

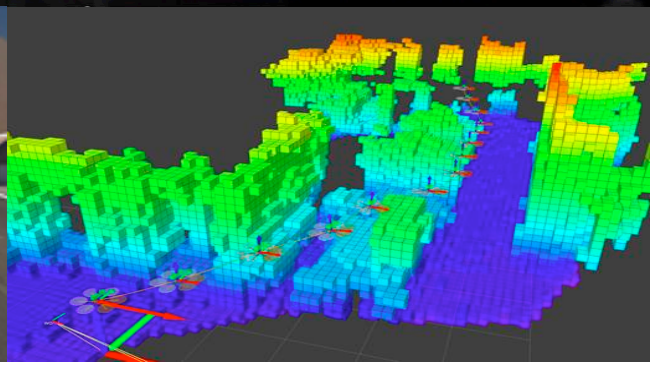
NIPS 2014 release: 2.5 million images, 205 scene categories

Zhou, Lapedriza, Xiao, Torralba & Oliva (2014), NIPS

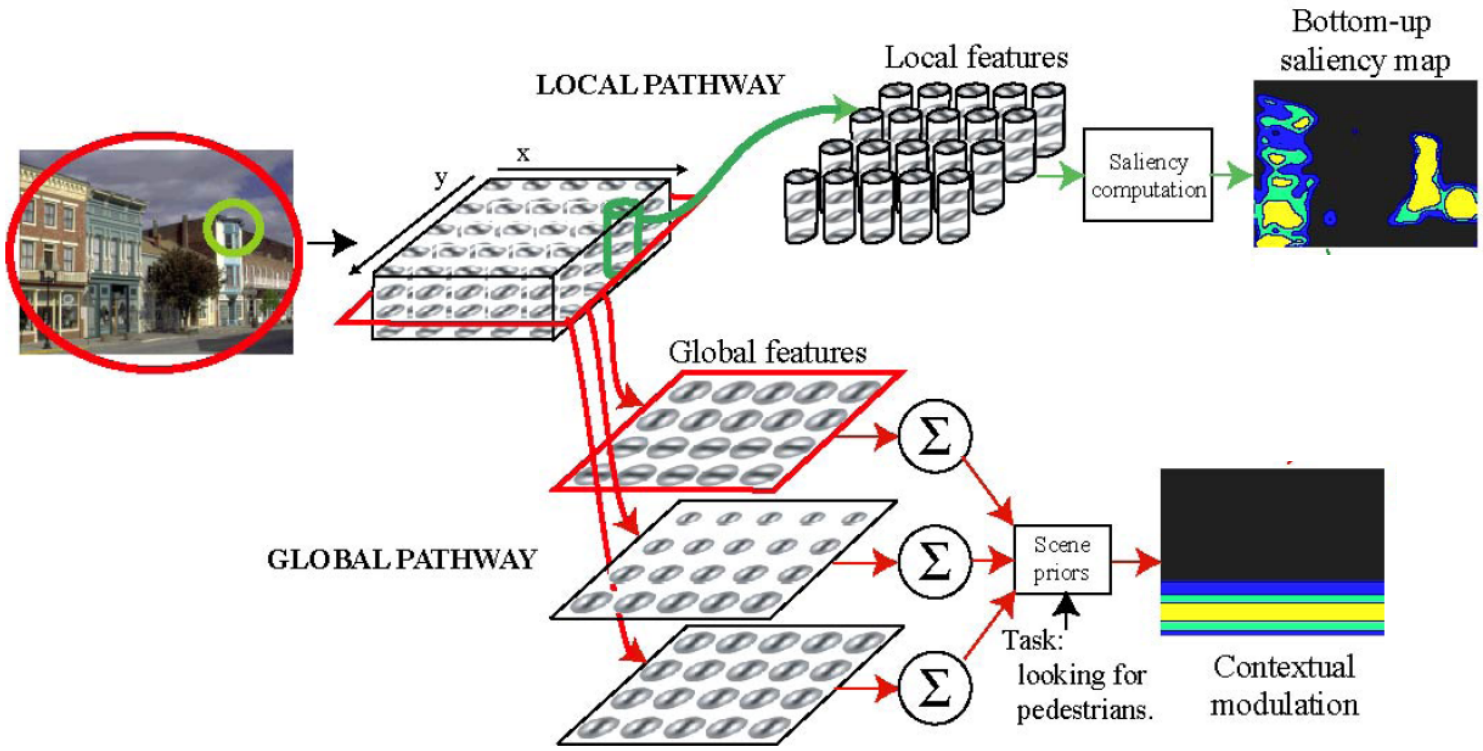



Describes without errors
Describes with minor errors
Somewhat related to the image
Unrelated to the image
A person riding a motorcycle on a dirt road.
Two dogs play in the grass.
A skateboarder does a trick on a ramp.
A dog is jumping to catch a frisbee.
A group of young people playing a game of frisbee.
Two hockey players are fighting over the puck.
A little girl in a pink hat is blowing bubbles.
A refrigerator filled with lots of food and drinks.
A herd of elephants walking across a dry grass field.
A close up of a cat laying on a couch.
A red motorcycle parked on the side of the road.
A yellow school bus parked i
a parking lot.



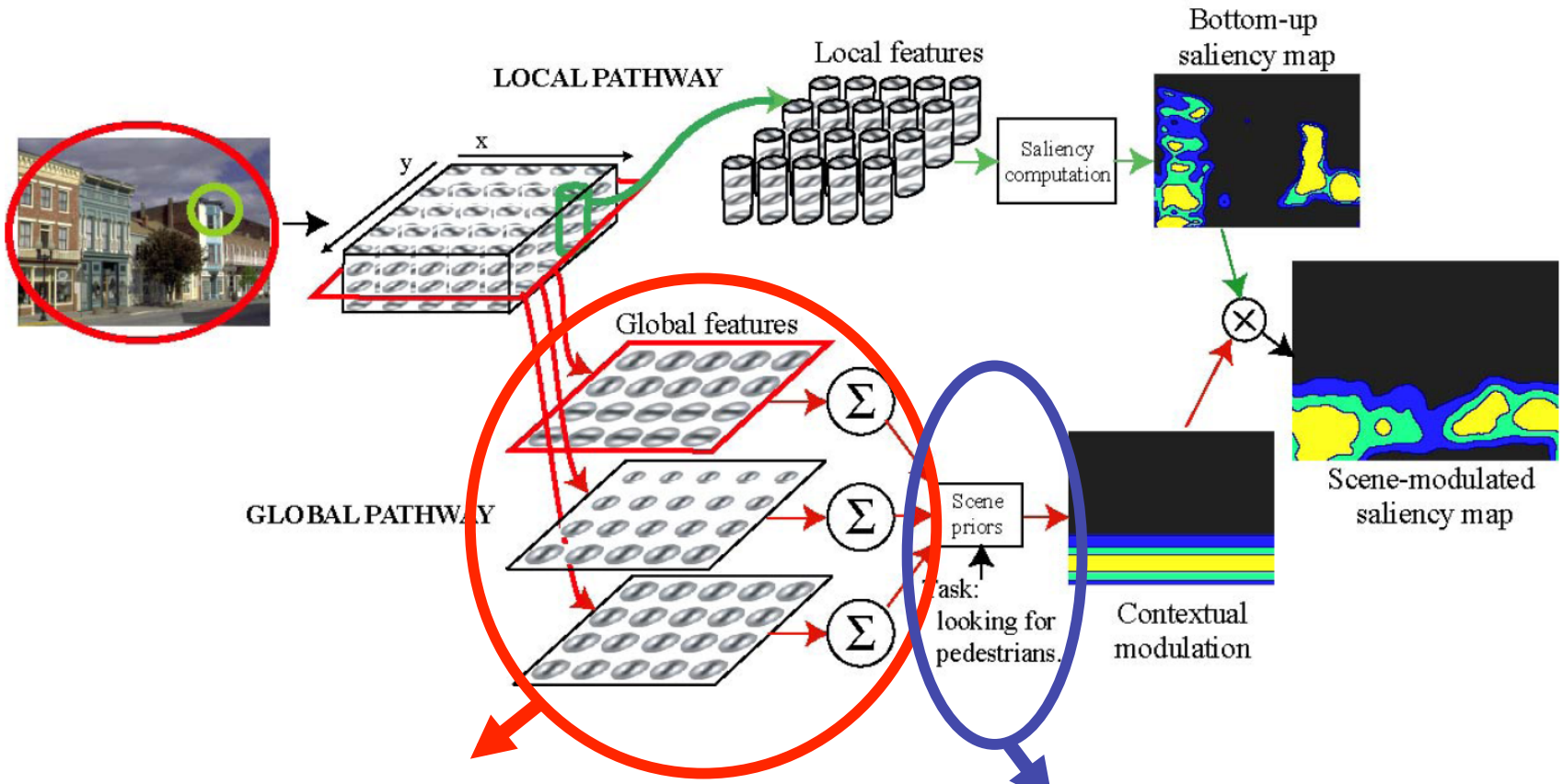


Contextual Guidance Model



Torralba, Oliva et al (2006), Torralba (2003)

Contextual Guidance Model



Torralba, Oliva et al (2006), Torralba (2003)

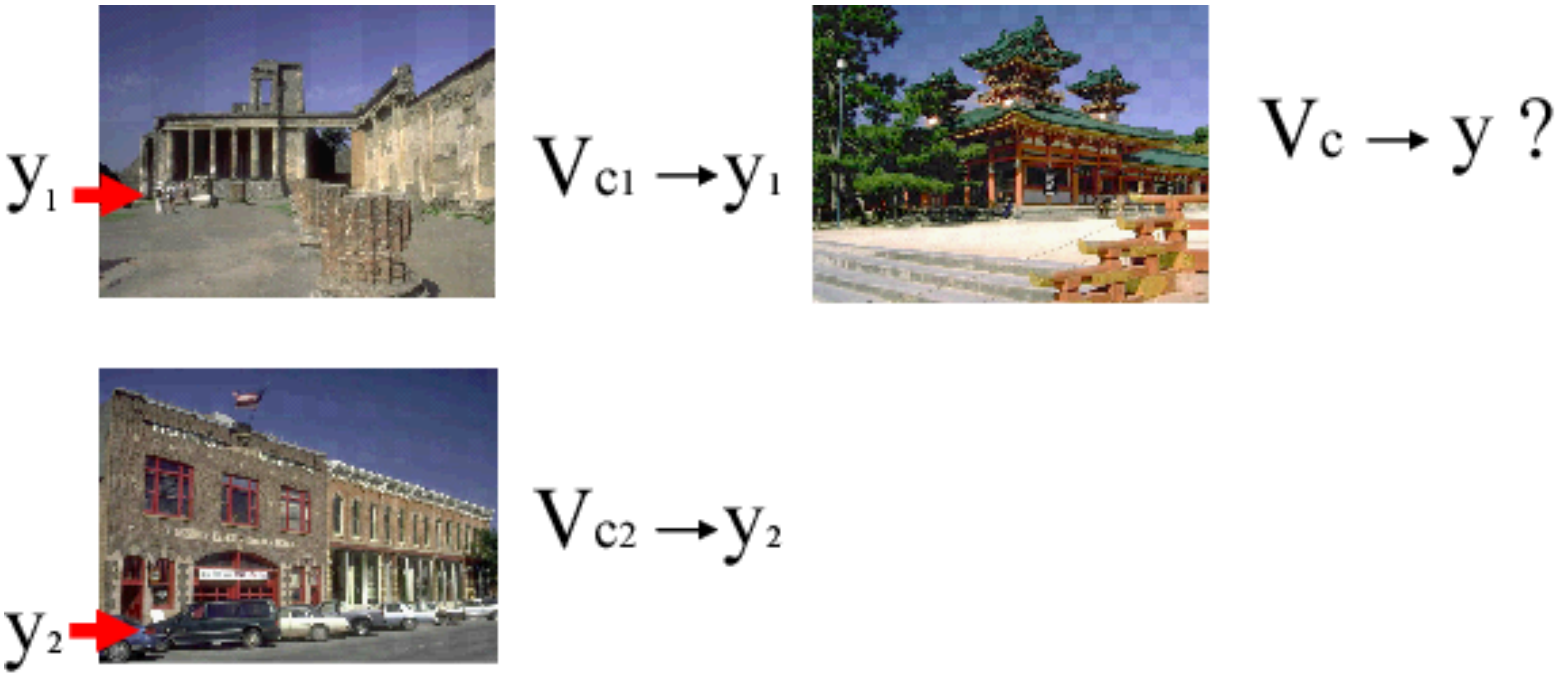
Goal

Predicting the location of the first eye movements for a given search task

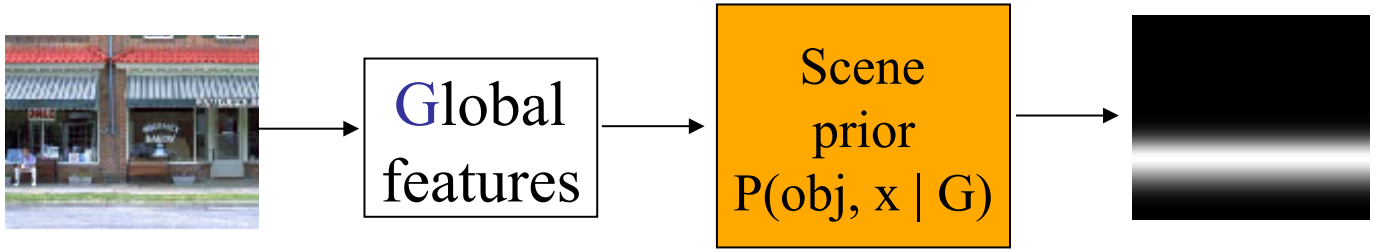
Observers fixations

Torralba, Oliva et al (2006), Oliva, Torralba, et al (2003)

Learning Scene Priors



Learning Scene Priors



• Guidance of attention by context requires a learning stage in which the system learns what are the typical locations of objects in scene.

• We trained the model to predict the location of people in the scene.

• We used a database of scenes that have been hand-labeled.

2500 images for which we know the location of people

The goal is to learn the joint distribution between global image features (Vc) and the location of the target

Torralba, Oliva et al (2006)

Categorical Priors Prototypes

Counting task

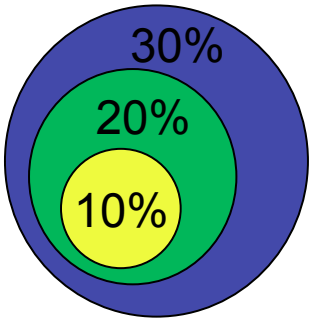
Observers search for small and camouflaged target objects

People search task

Mug and painting search task

Torralba, Oliva et al. (2006)

Comparison regions of interest

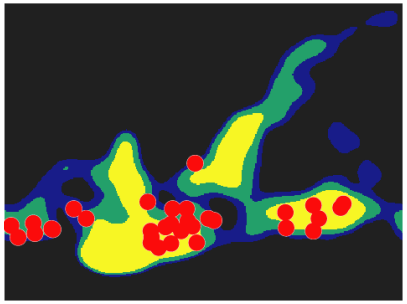


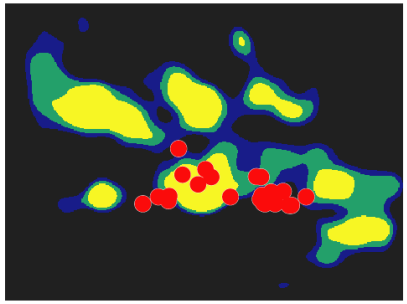
78%

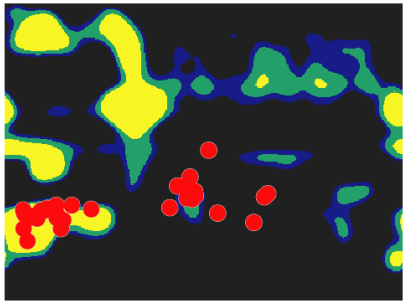
25%

63%

Saliency predictions





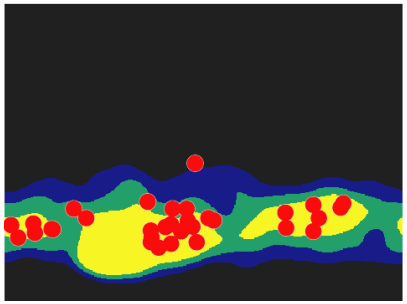


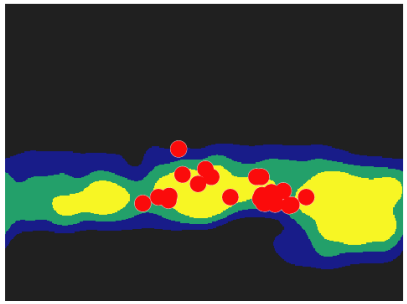
96%

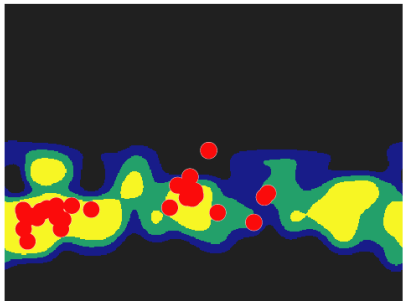
96%

90%

Saliency and Global scene priors



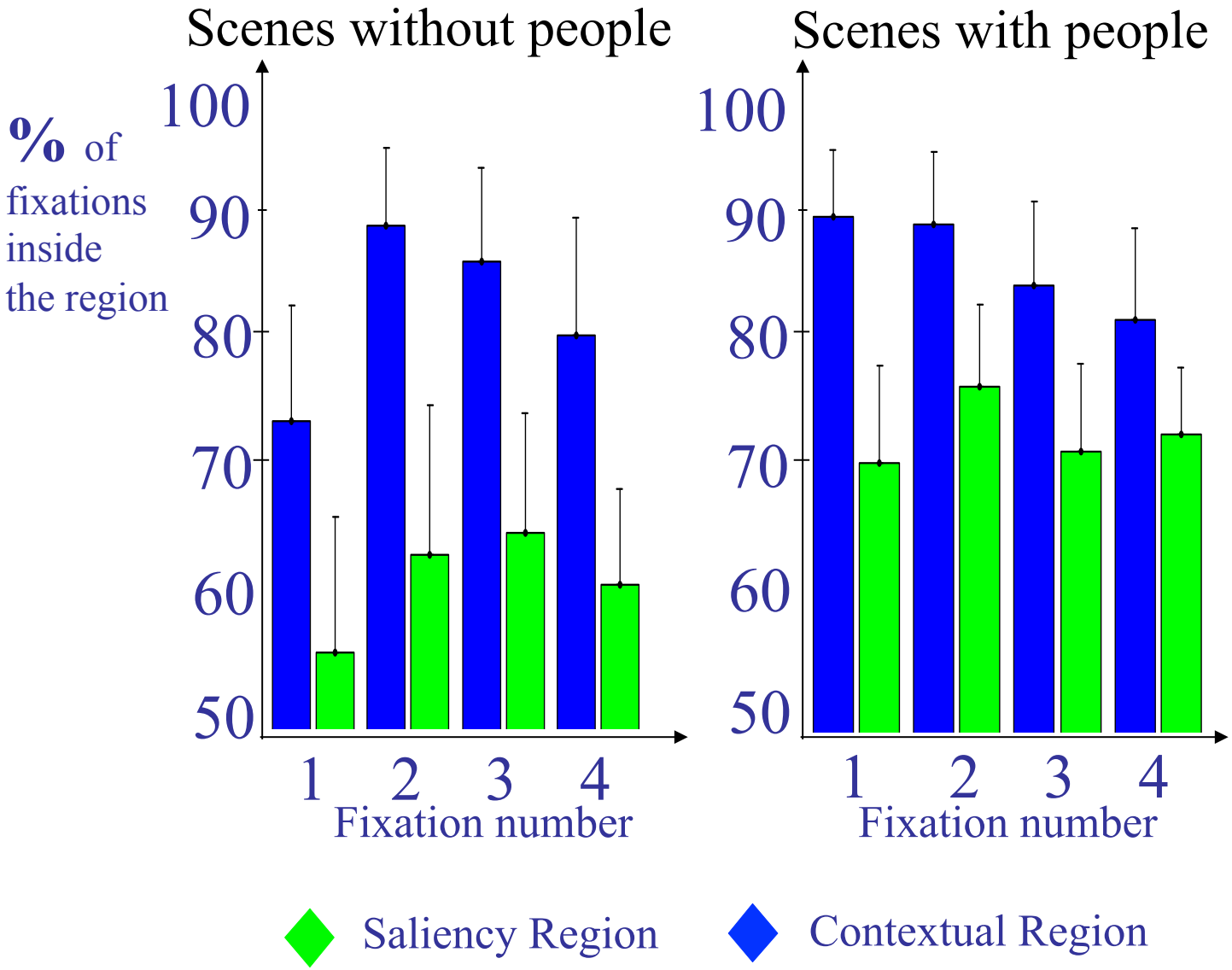




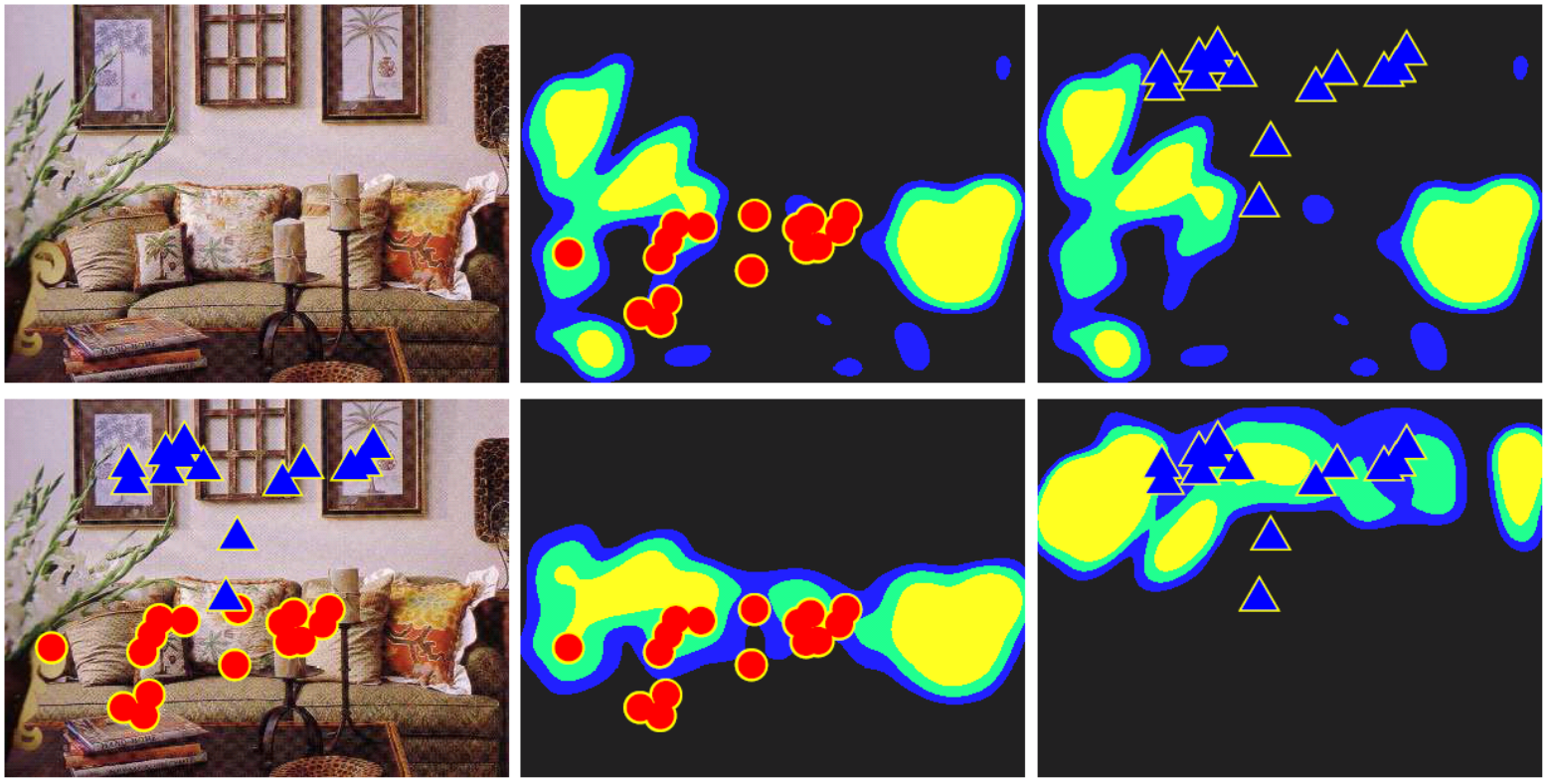
Red dots correspond to fixations 1-4

Torralba, Oliva et al. (2006)

Results: Detecting People



Task modulation



Saliency predictions

Saliency and Global scene priors

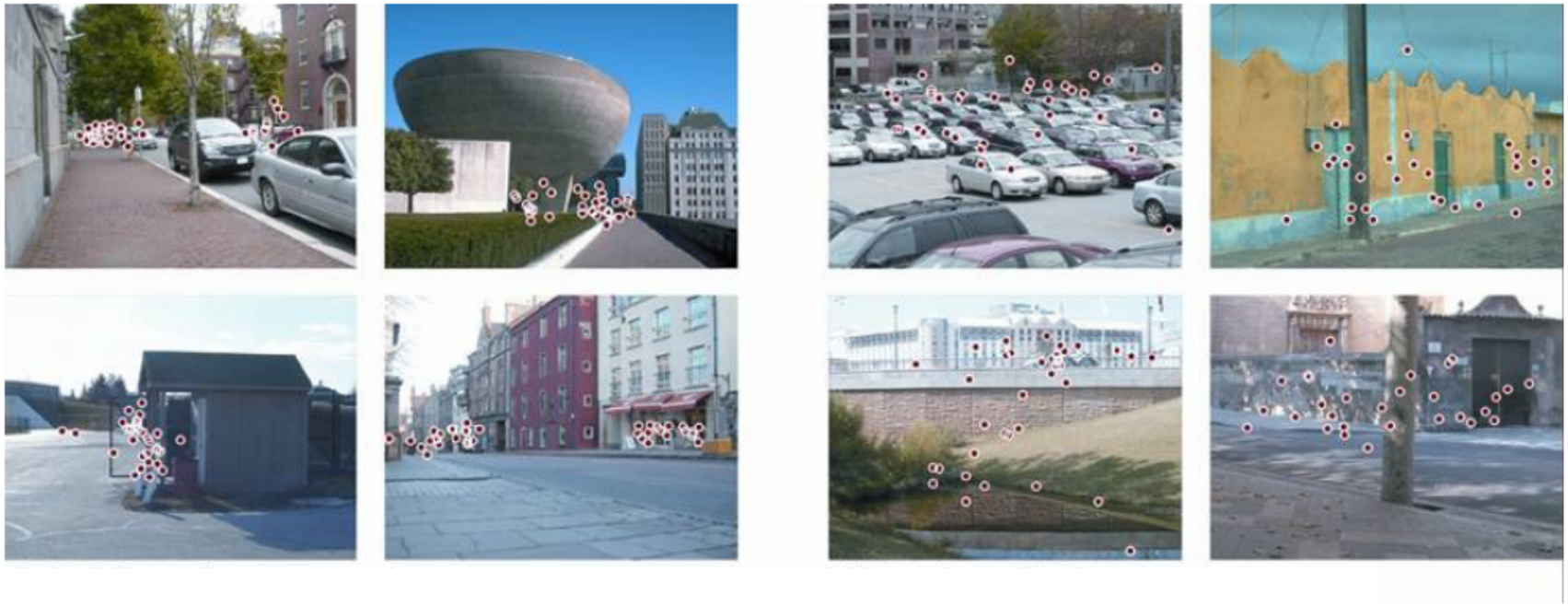
Mug search

Painting search

Torralba, Oliva et al. (2006)

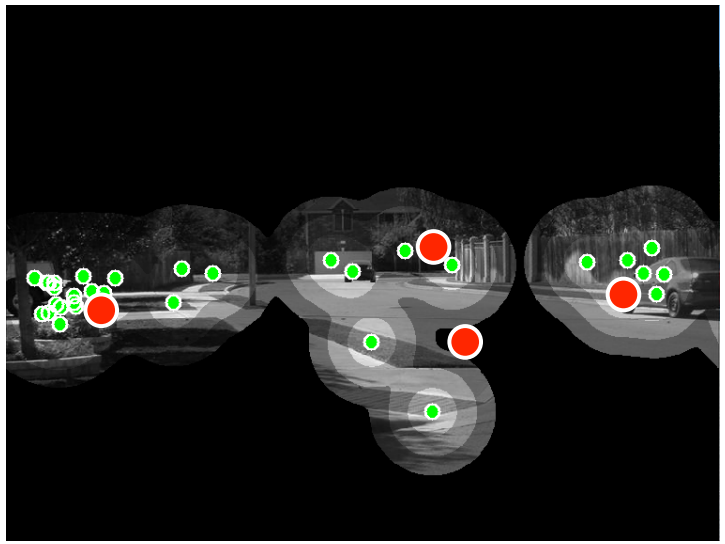

People detection in outdoors:
A thousand scenes …
& 45,144 eye fixations later …

(1) Do observers look at the same places ?
(2) Can we predict the fixated regions ?



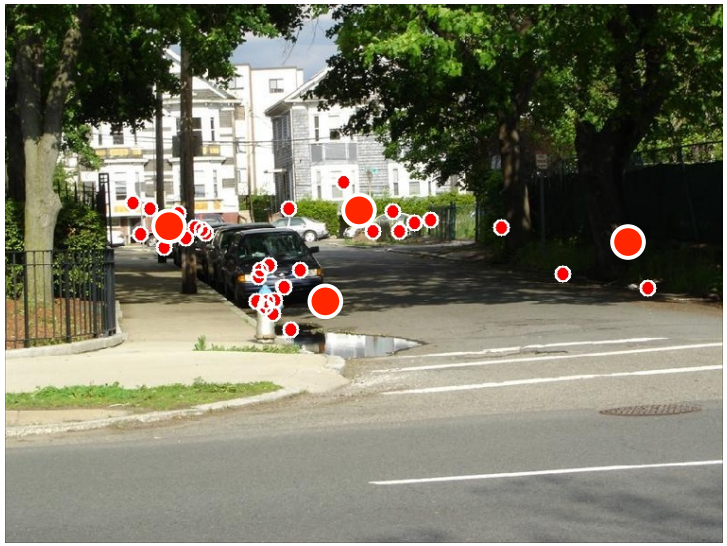
Human Agreement

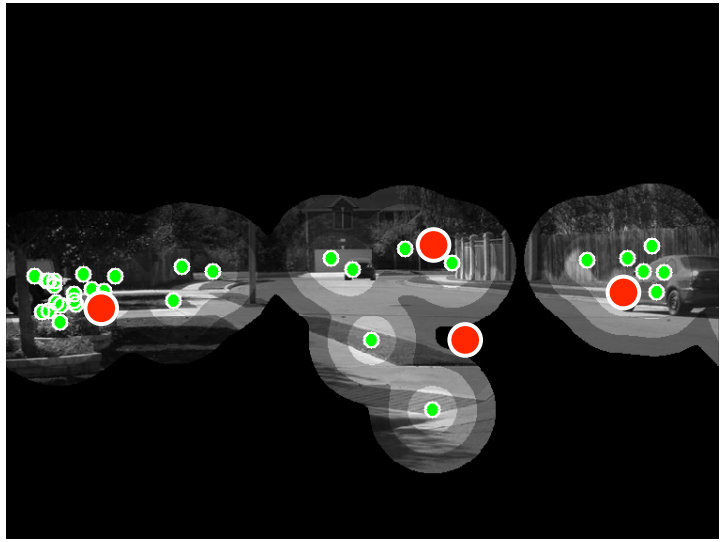
- **Inter-observer agreement** = upper bound for model performance



Human Agreement

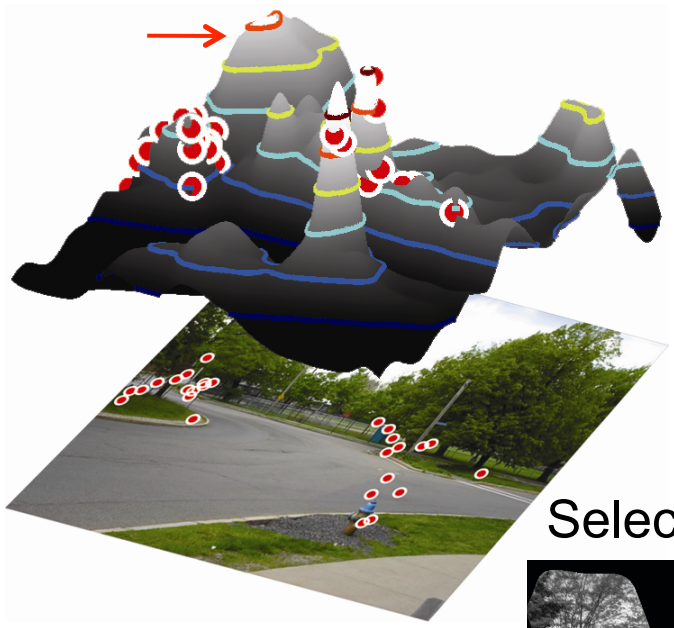
- **Inter-observer agreement** = upper bound for model performance
- **Cross-image control** = lower bound for model performance



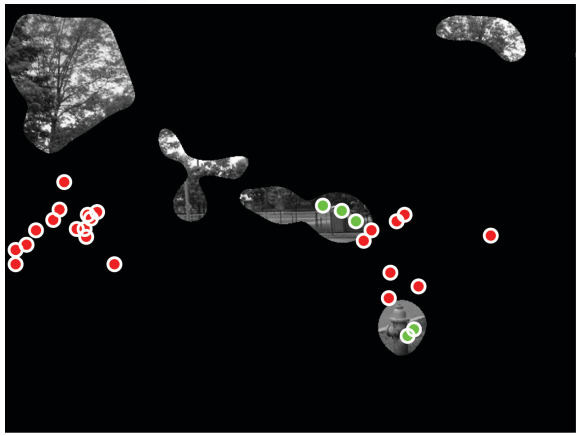


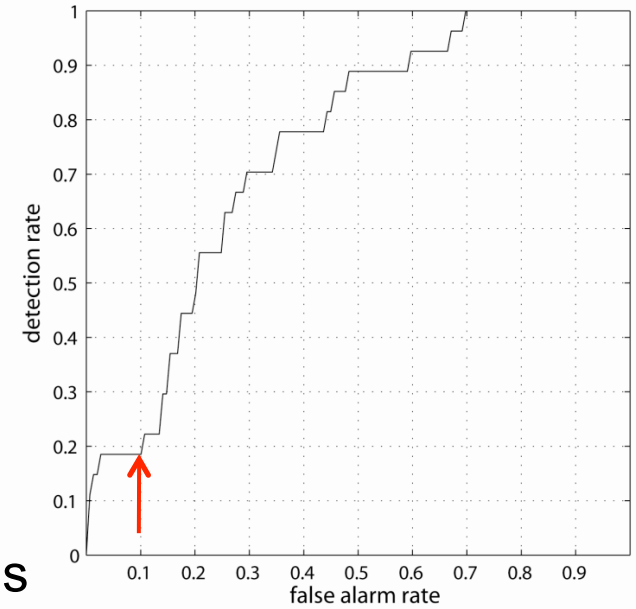
The ROC curve

Model



Selected image regions



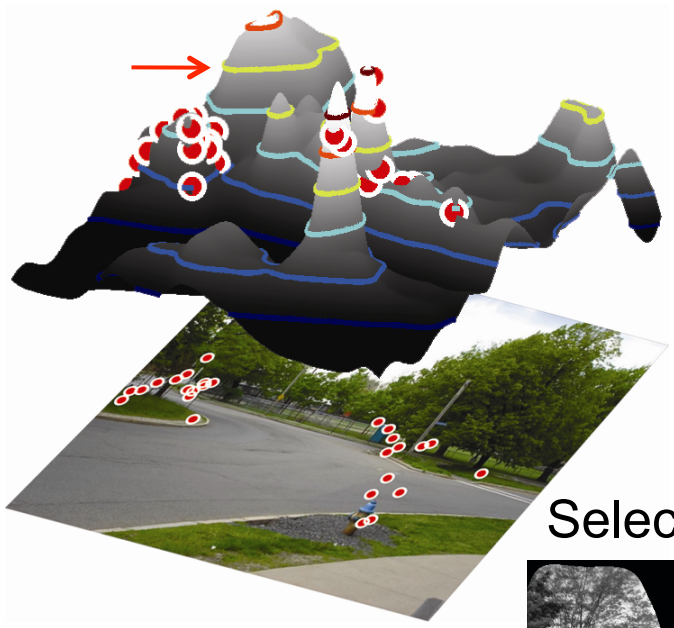


ROC curve

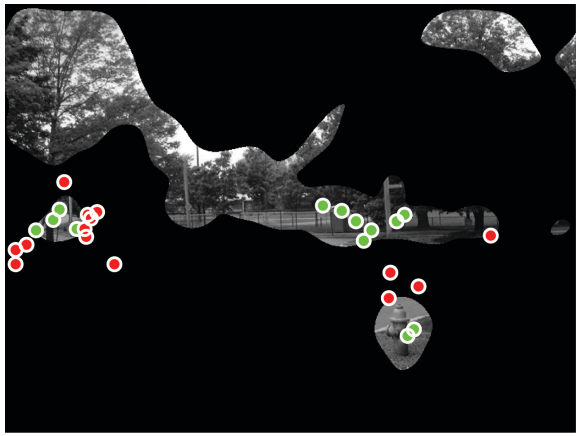
a **receiver operating characteristic** (**ROC**), or **ROC curve**, is a graphical plot that illustrates the performance of a binary classifier system as its discrimination threshold is varied. The curve is created by plotting the true positive rate or detection rate (TPR) against the false positive rate (FPR) at various threshold settings.

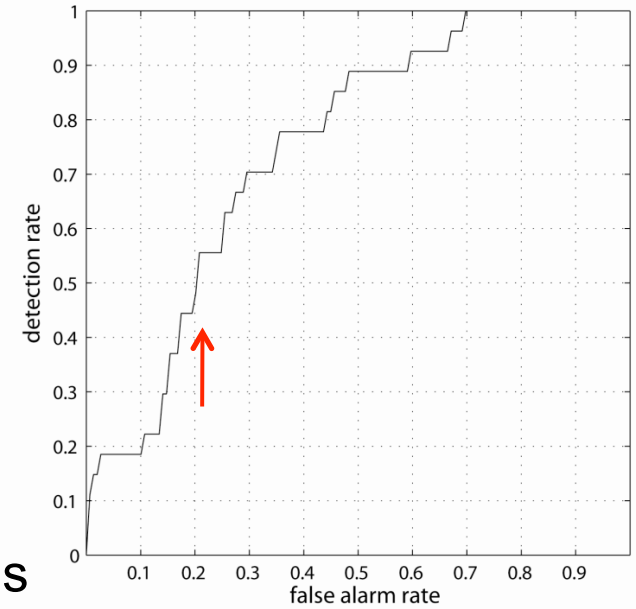
The ROC curve

Model



Selected image regions

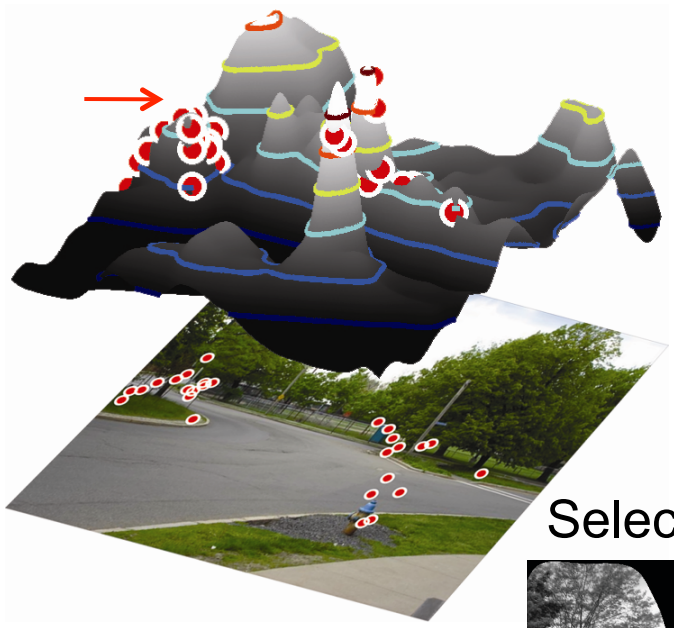




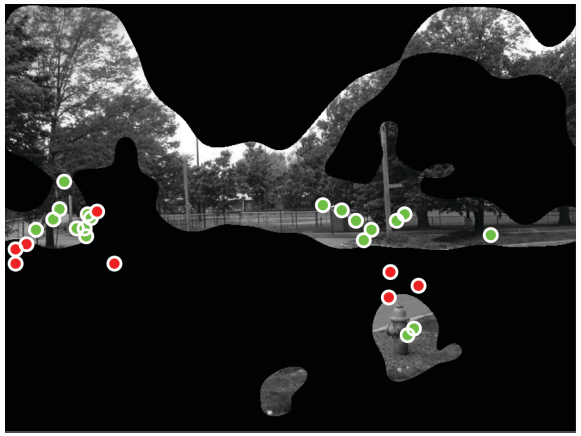
ROC curve

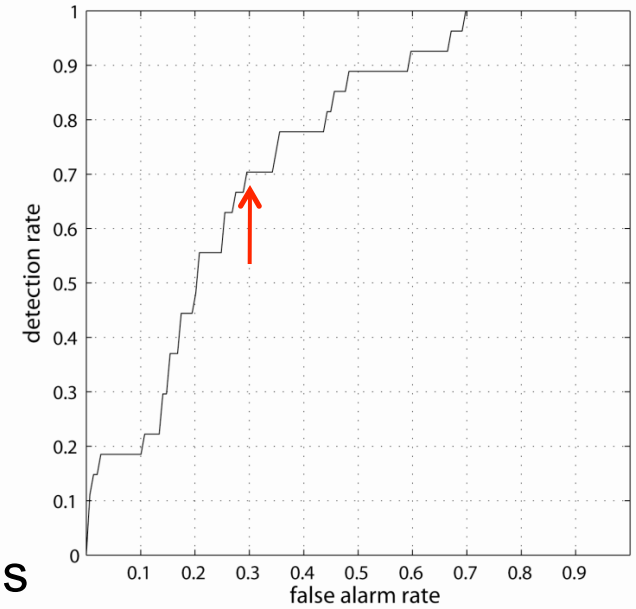
The ROC curve

Model



Selected image regions

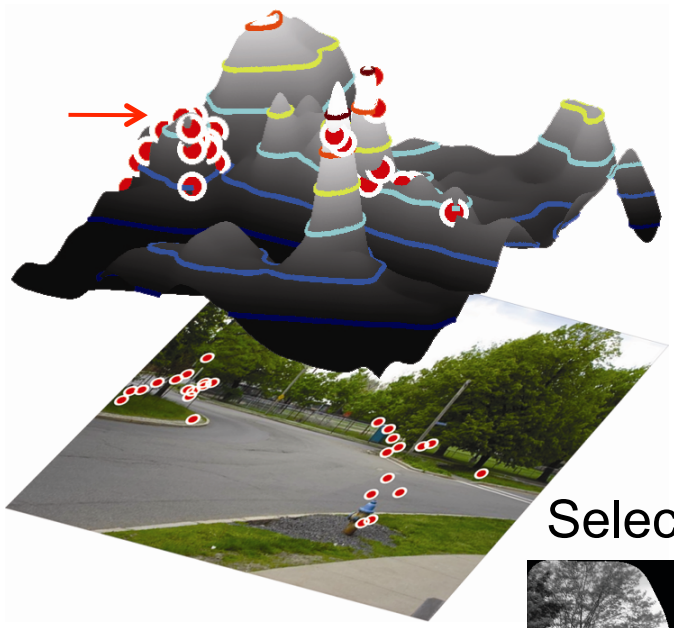




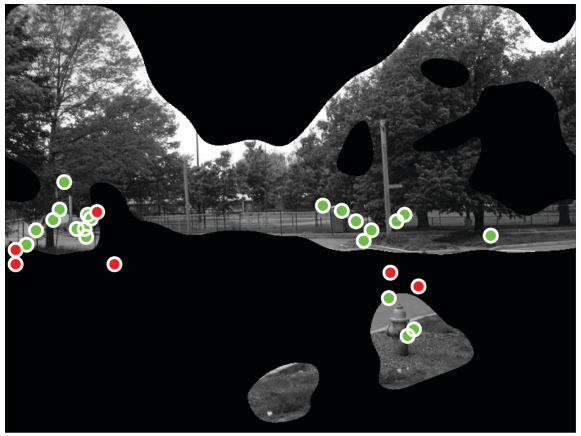
ROC curve

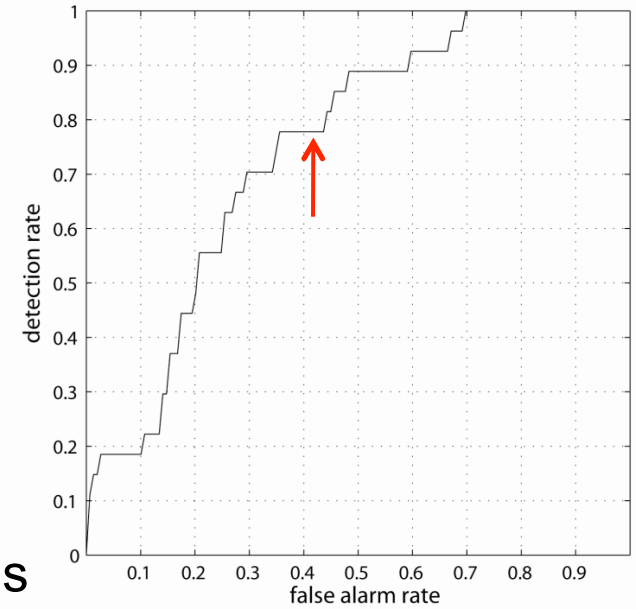
The ROC curve

Model



Selected image regions

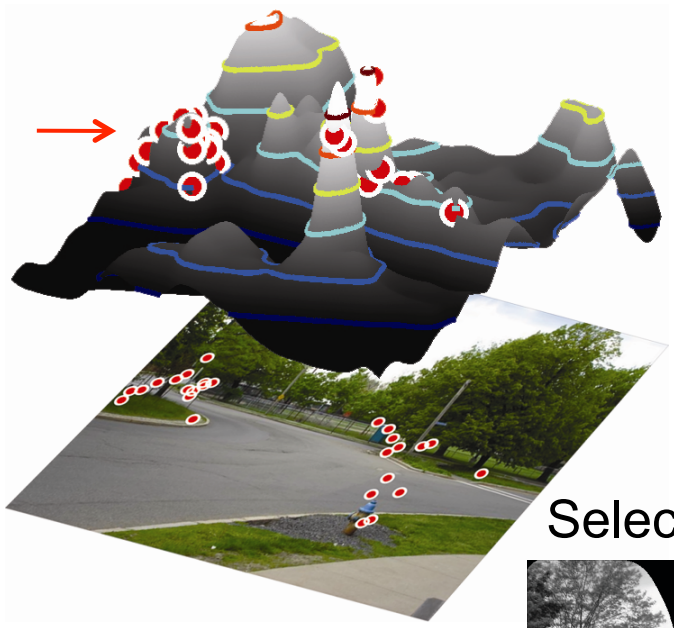




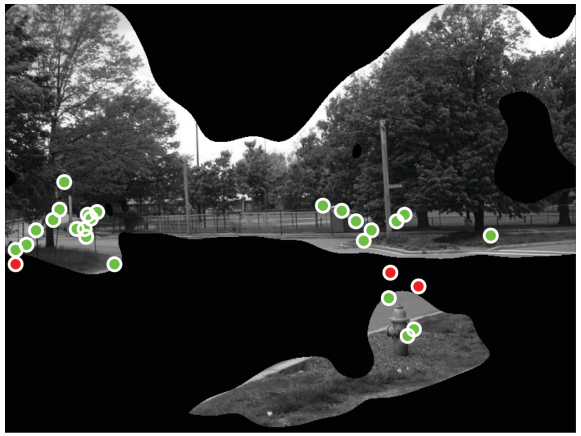
ROC curve

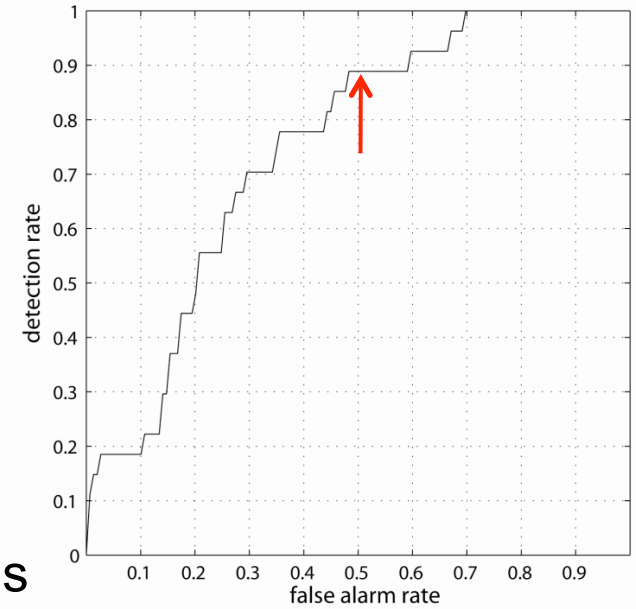
The ROC curve

Model



Selected image regions

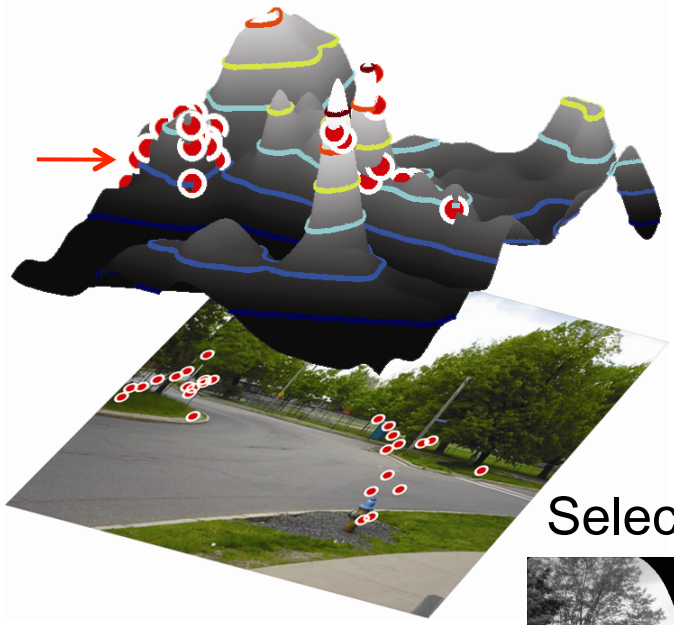




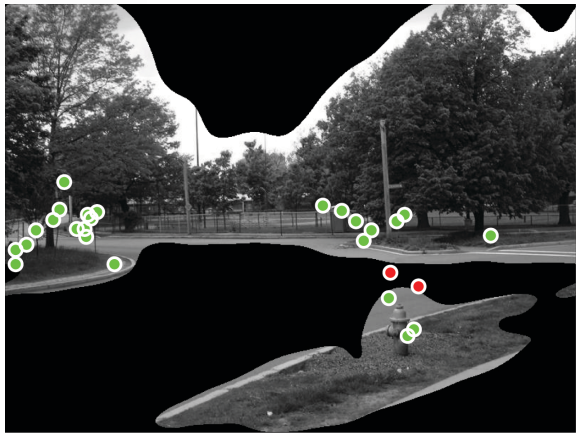
ROC curve

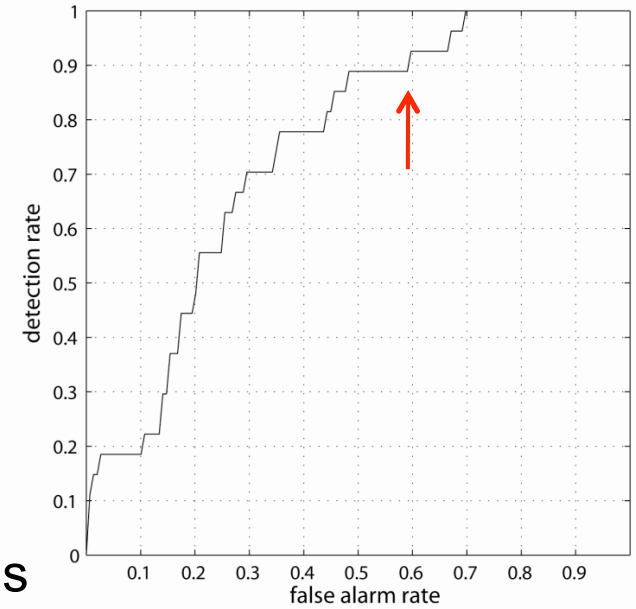
The ROC curve

Model



Selected image regions

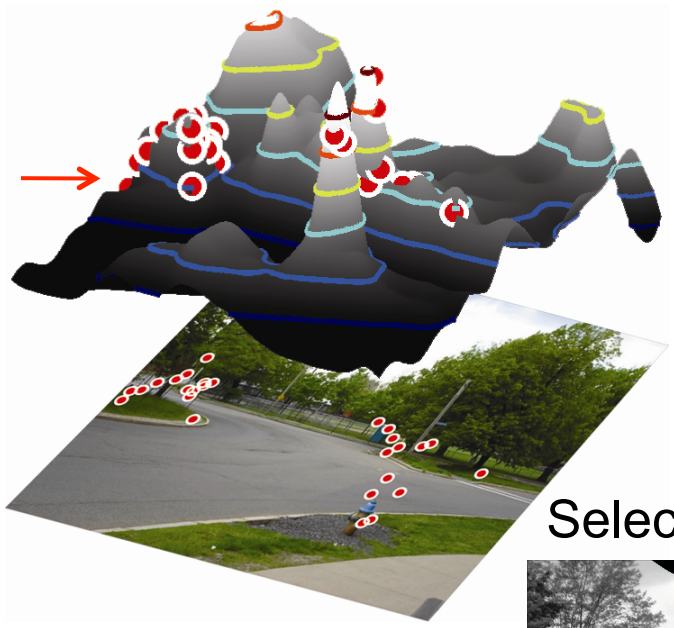




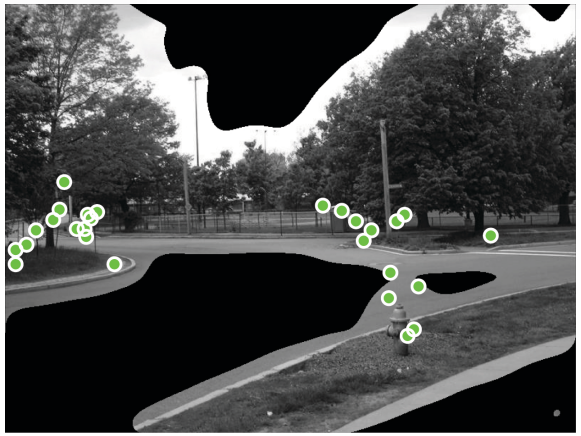
ROC curve

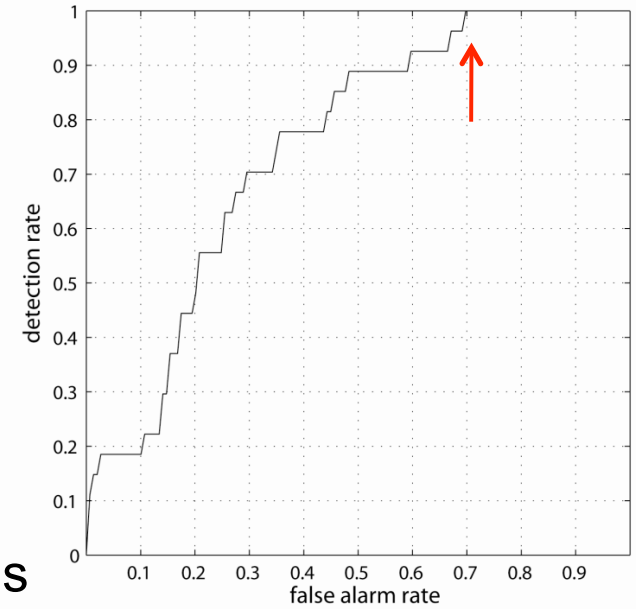
The ROC curve

Model



Selected image regions

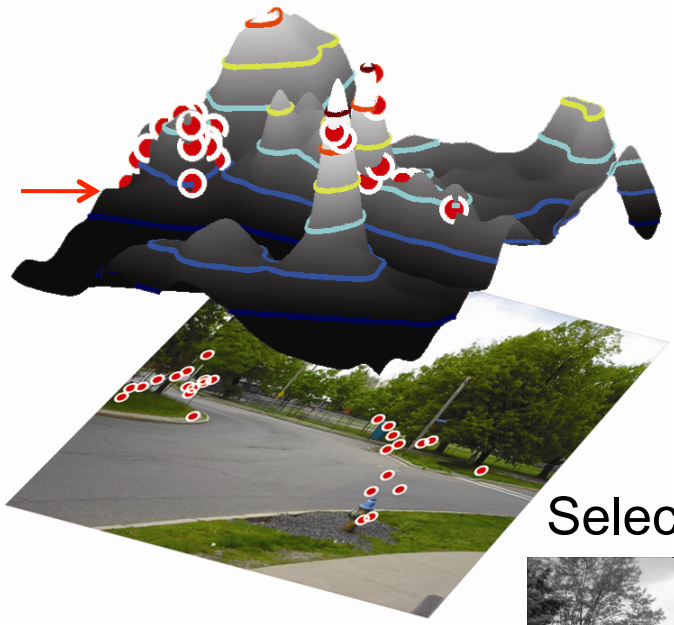




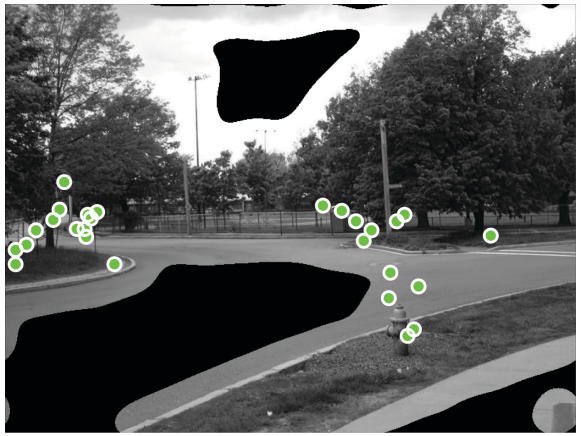
ROC curve

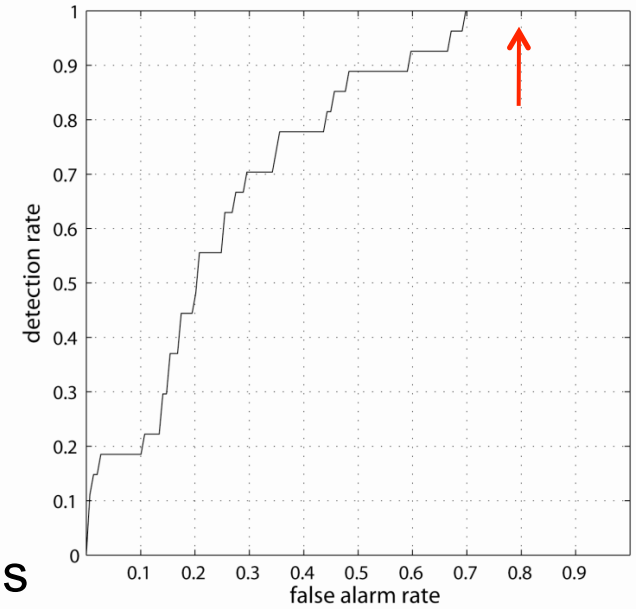
The ROC curve

Model



Selected image regions

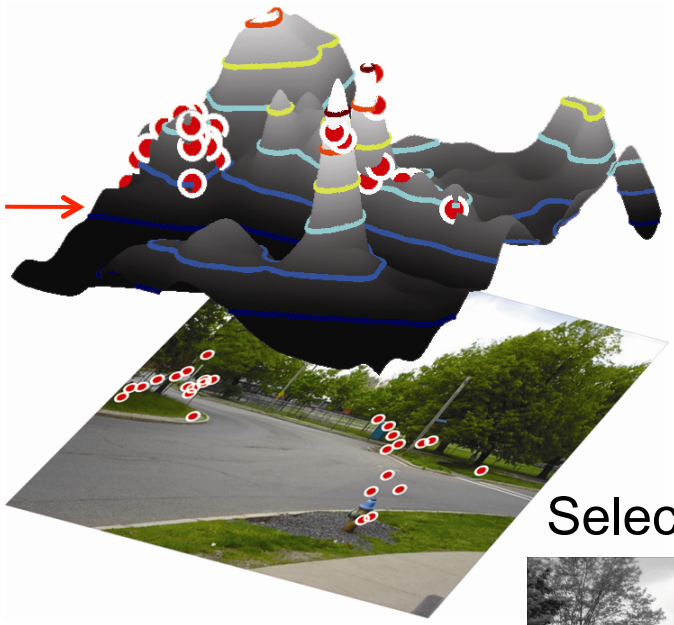




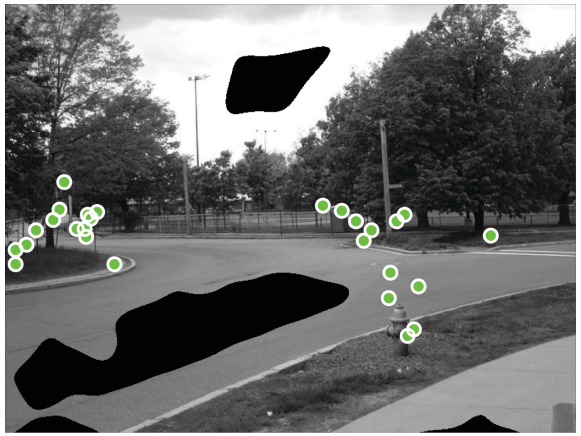
ROC curve

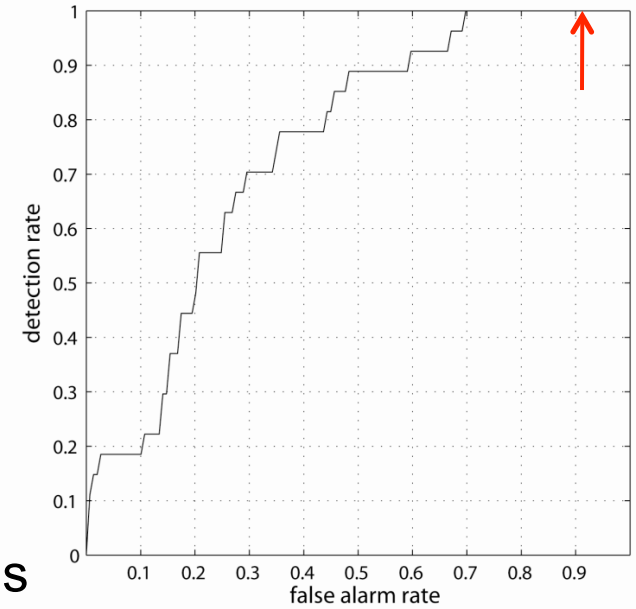
The ROC curve

Model



Selected image regions

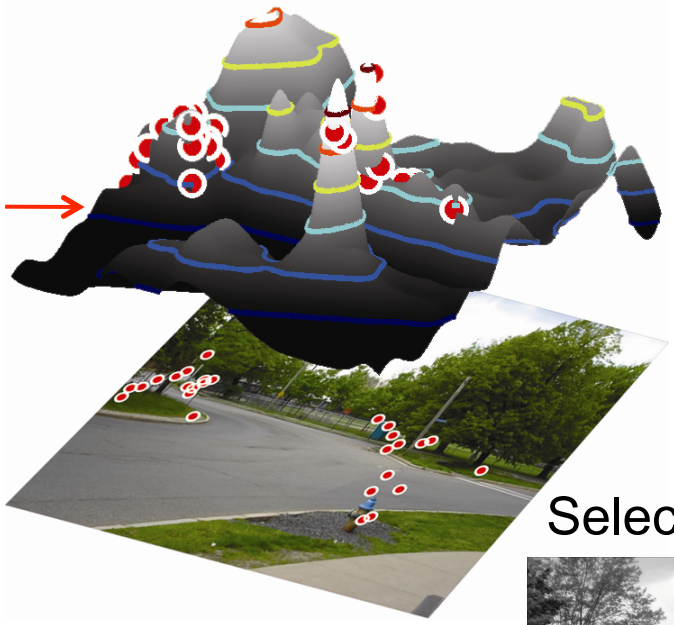




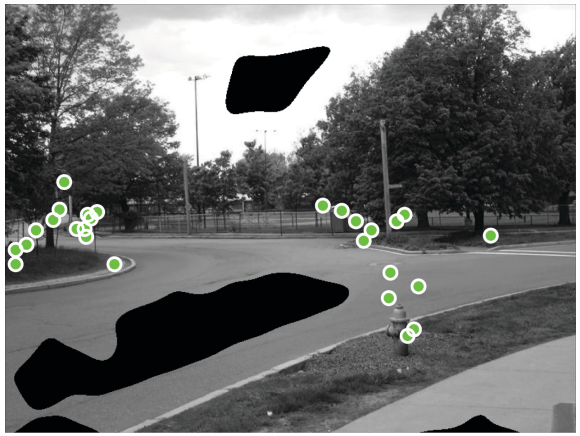
ROC curve

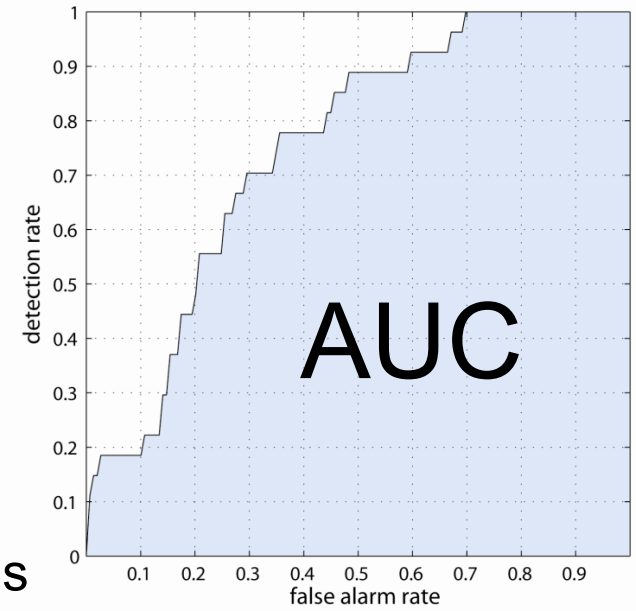
The ROC curve

Model



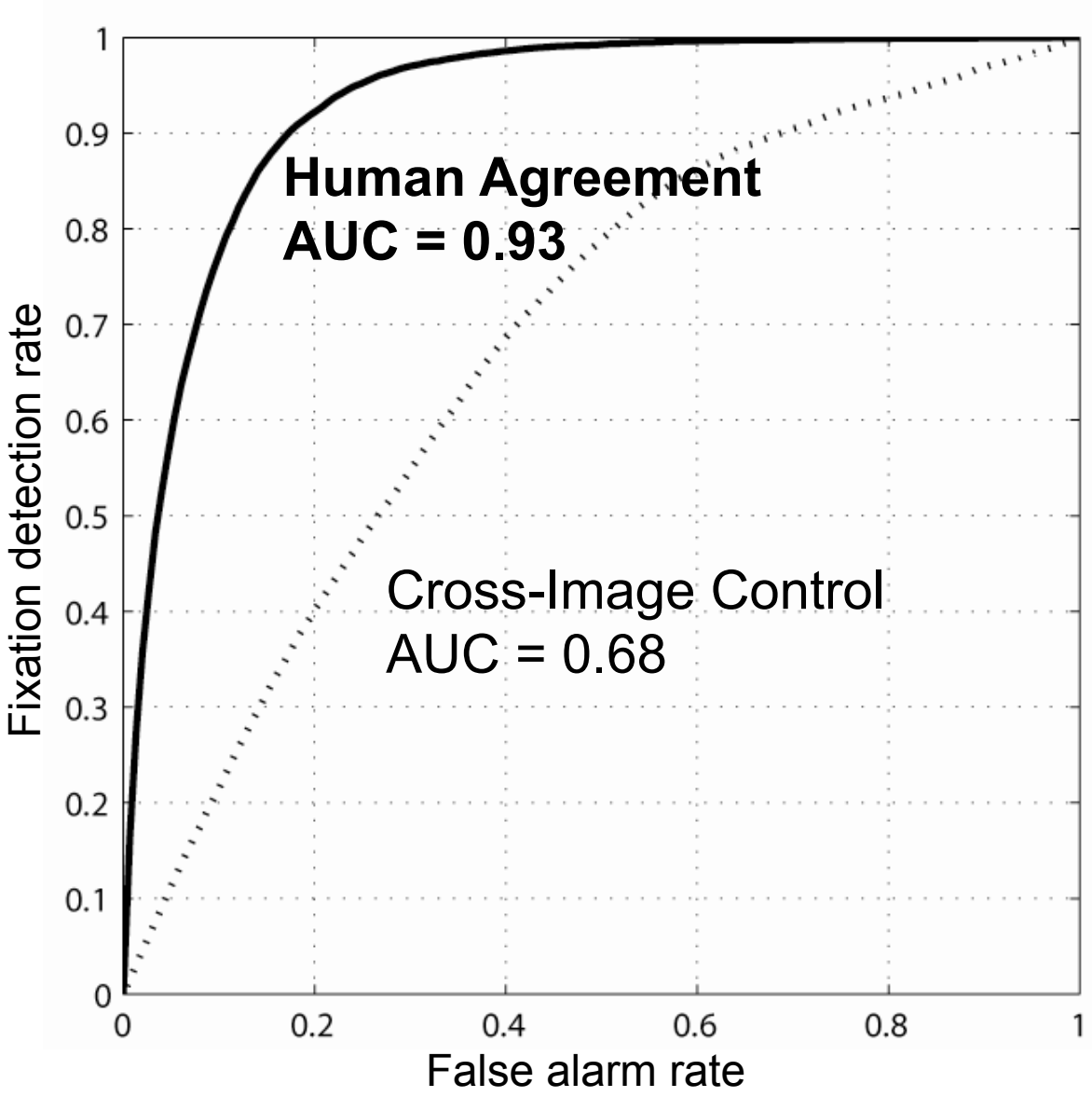
Selected image regions





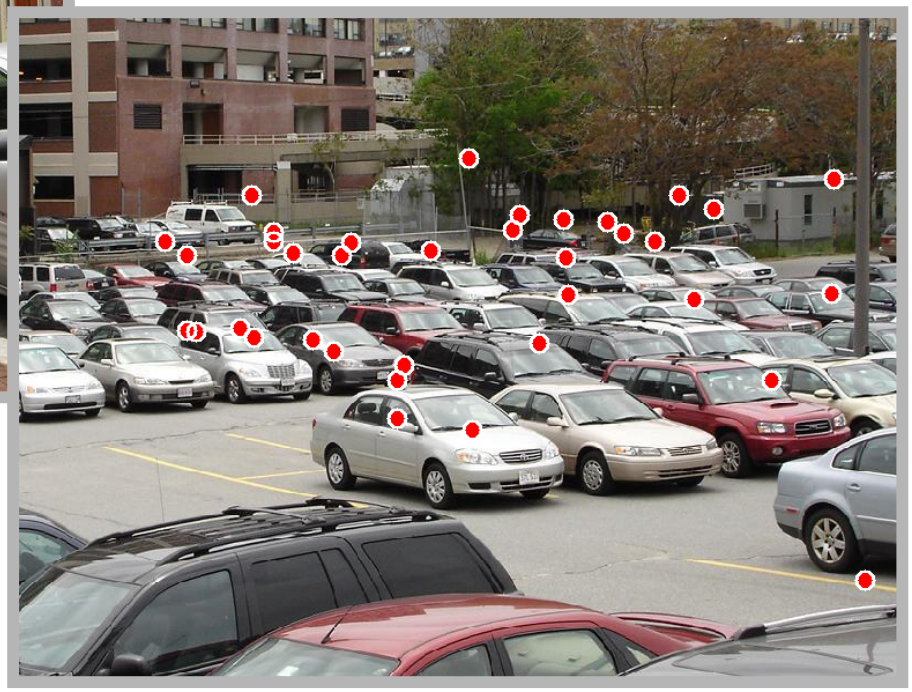
ROC curve

Human Agreement



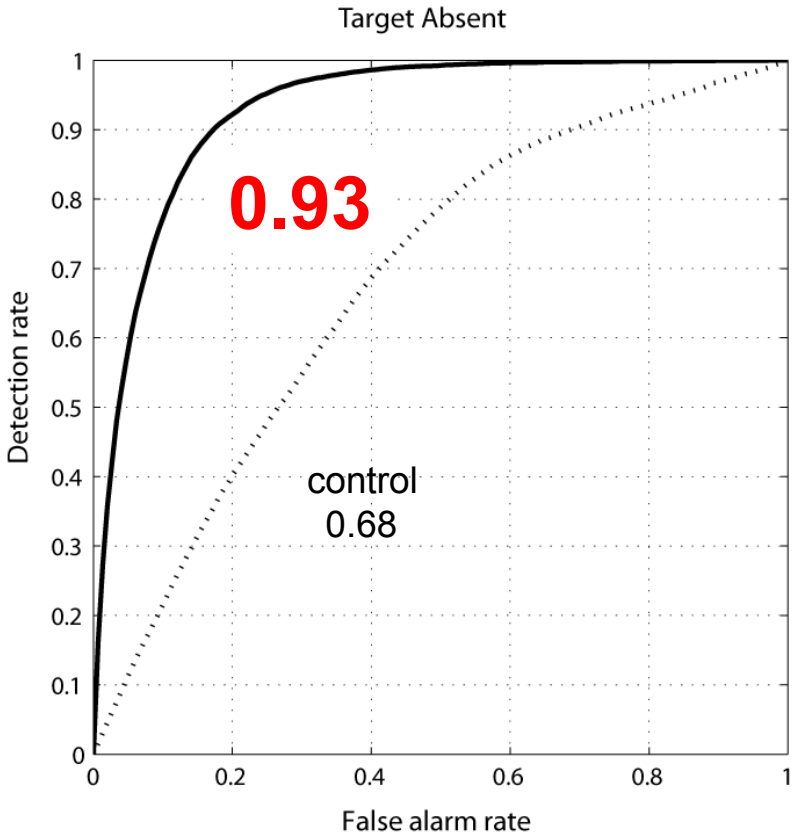
Human agreement examples

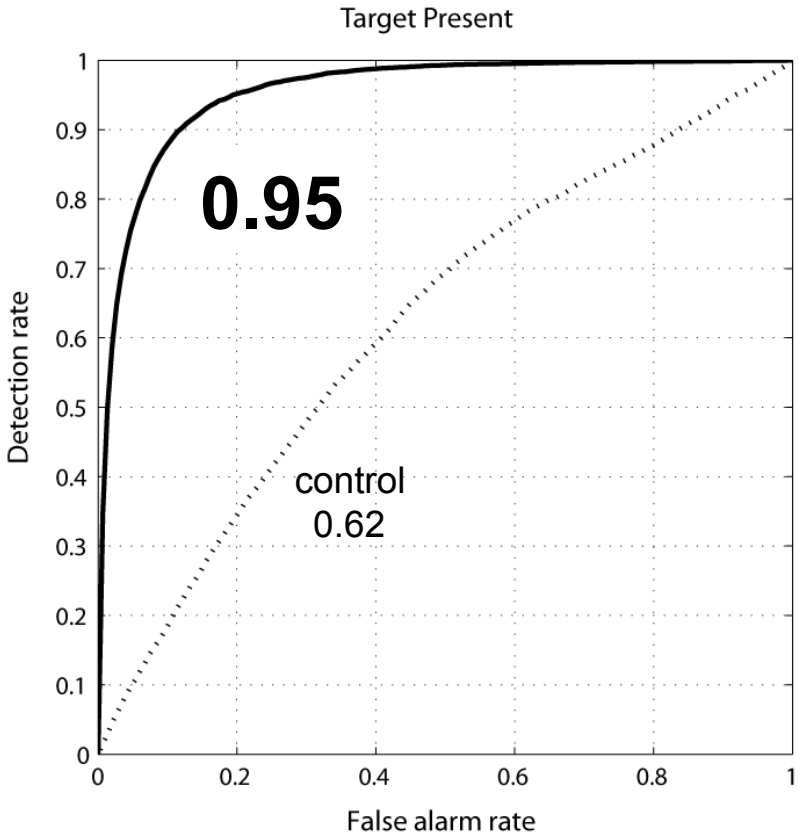
High inter-observer agreement

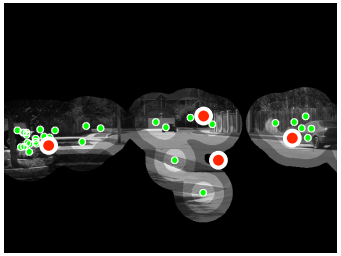


Low inter-observer agreement

Human agreement is very high !



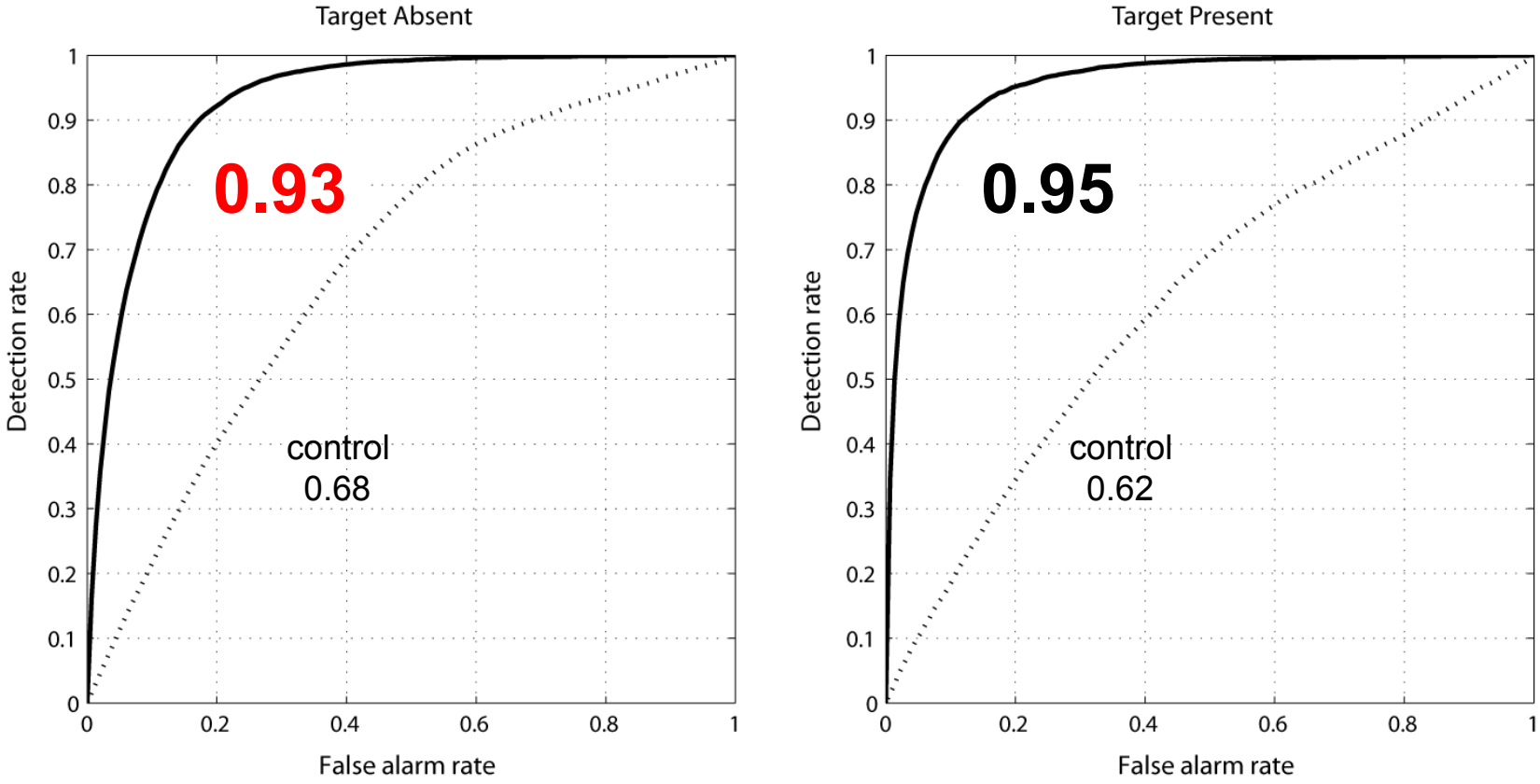




Human Agreement

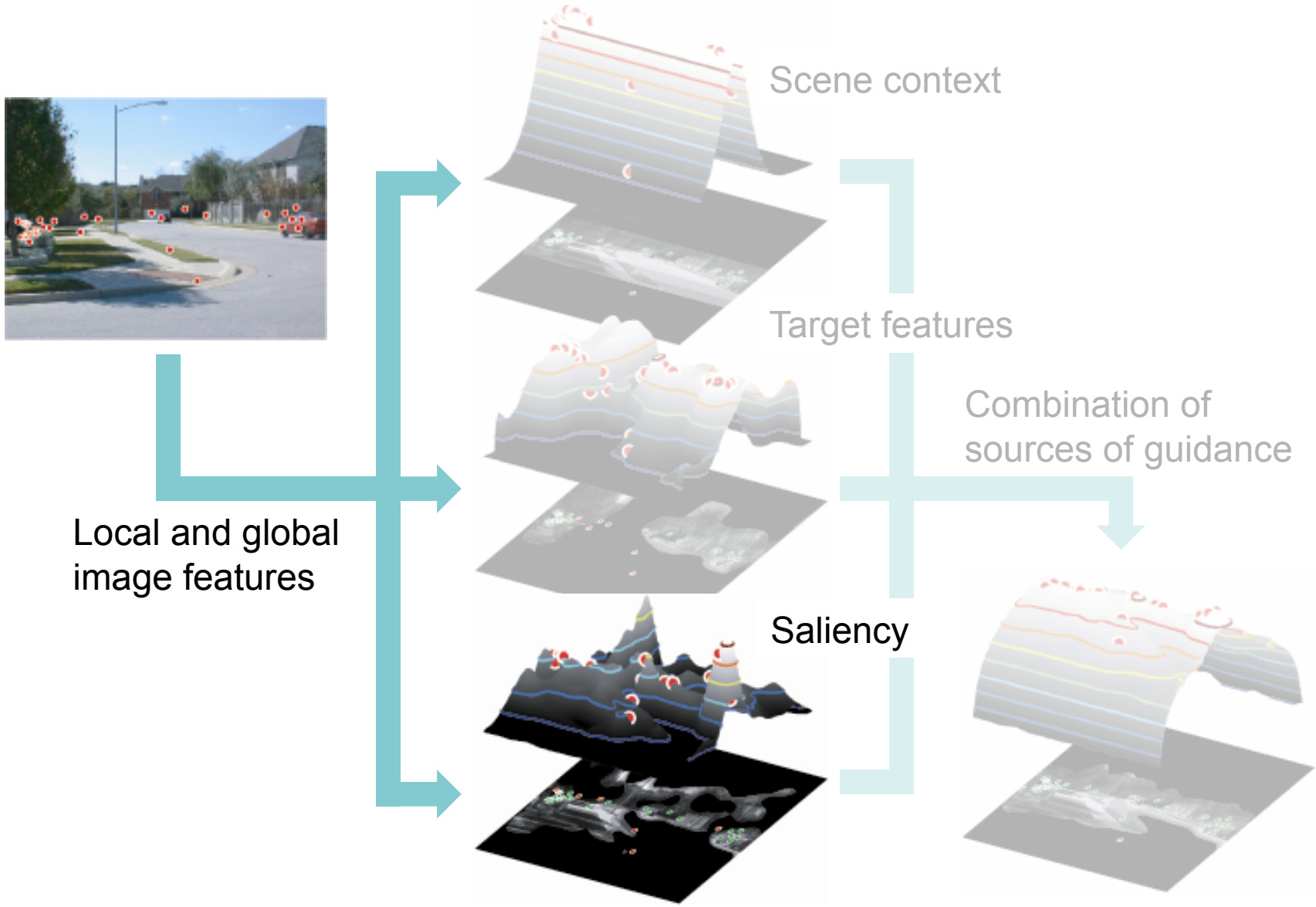
Cross-Image Control

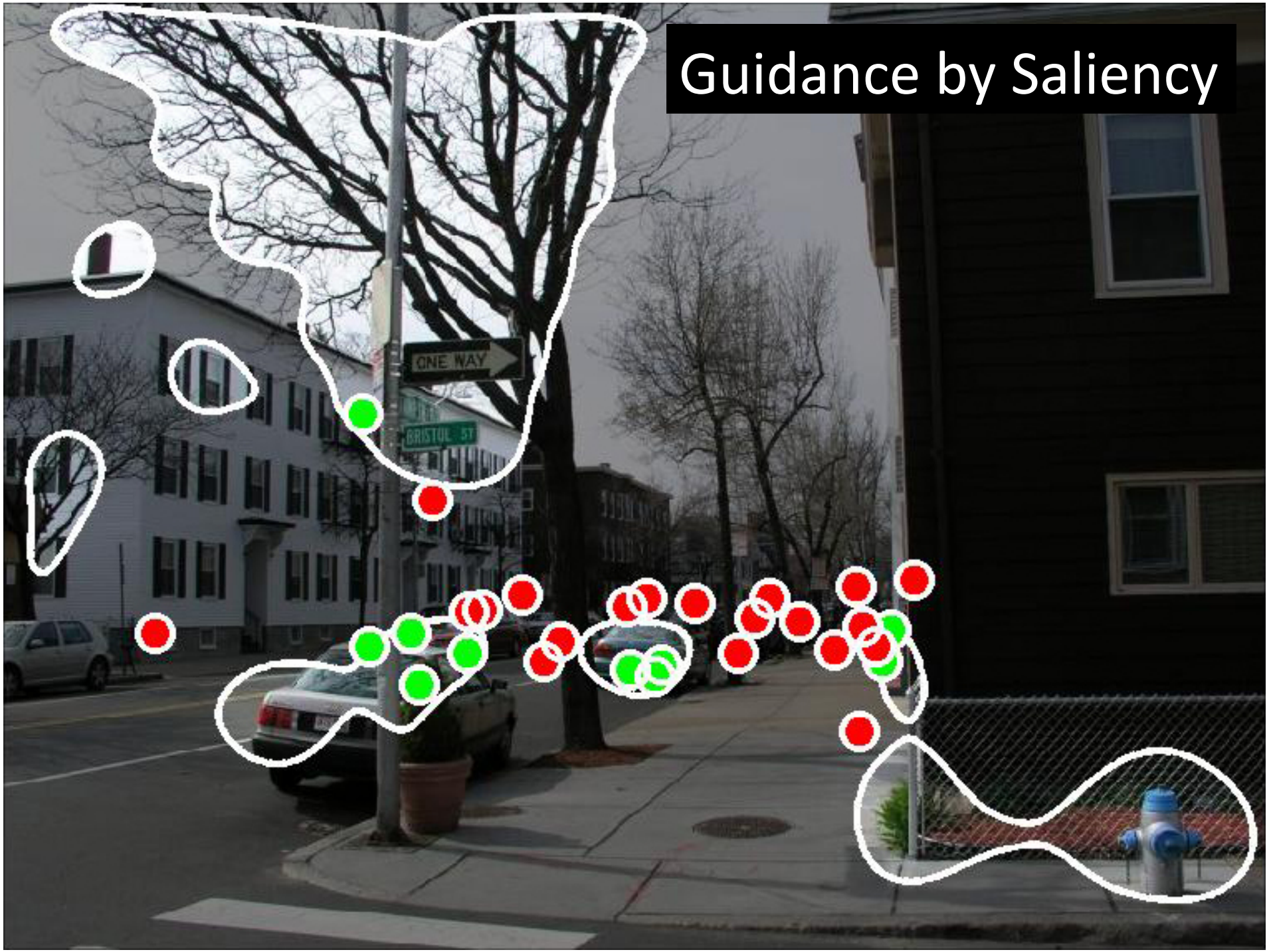
Can a model predict human fixations like another human?



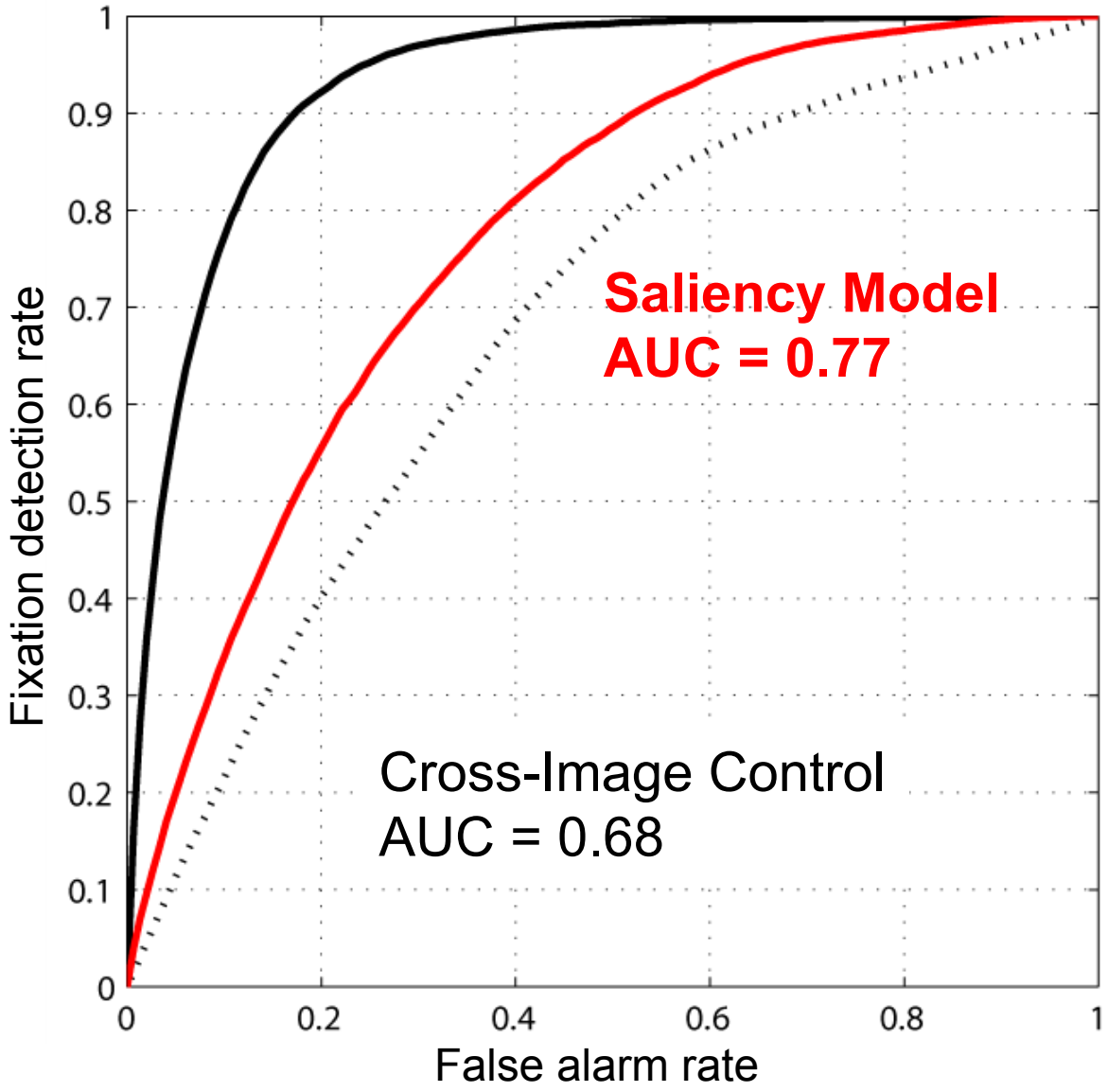
The prediction given by a model would be indistinguishable from the prediction by another human

Overview of Model



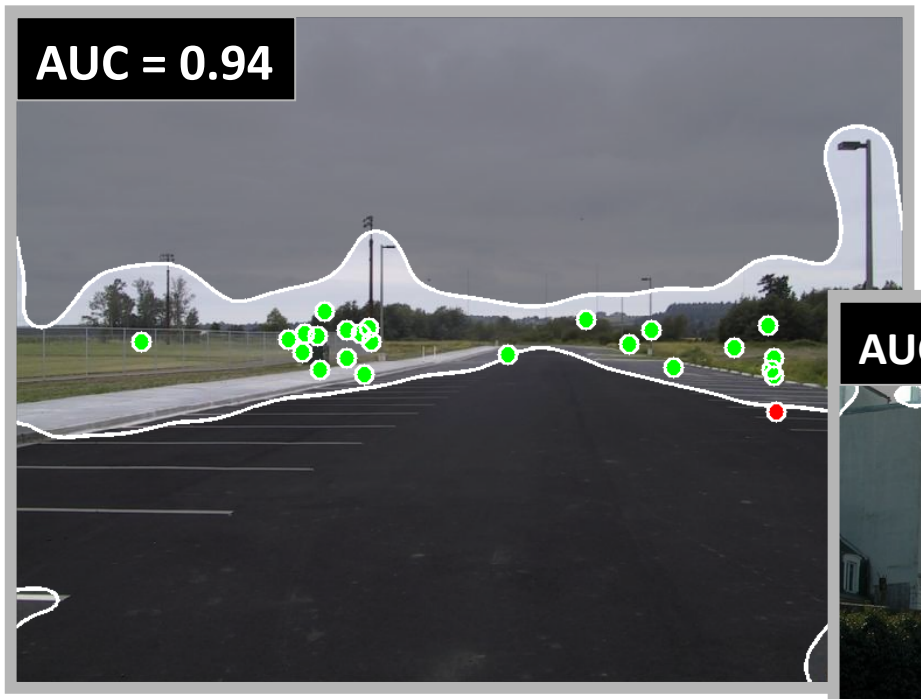

Guidance by Saliency
ONE WAY
BRISTOL ST

Saliency Model



Human Agreement
AUC = 0.93

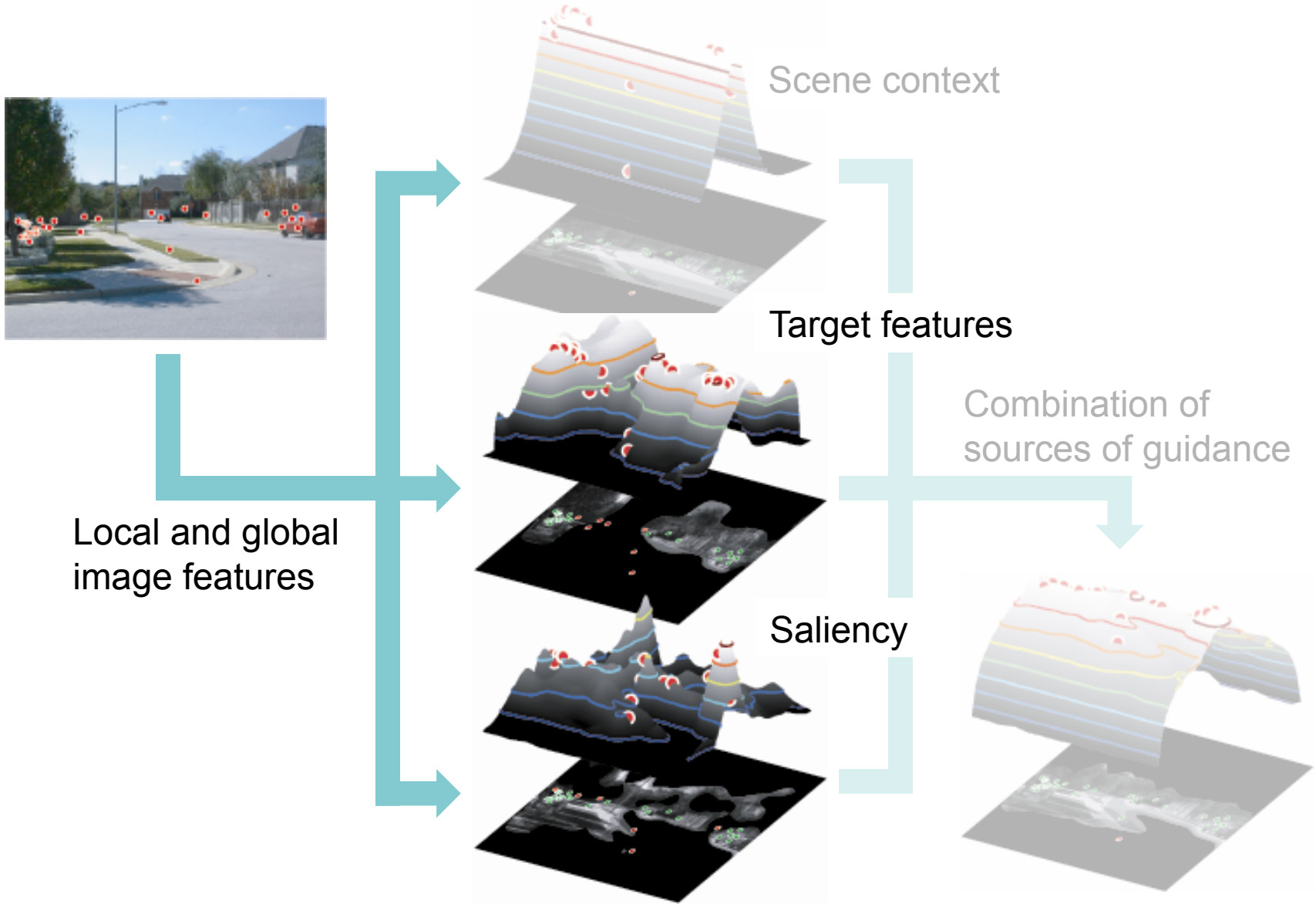
Saliency Model: Examples



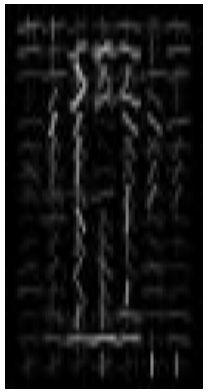
Best performance

Worst performance

Overview of Model



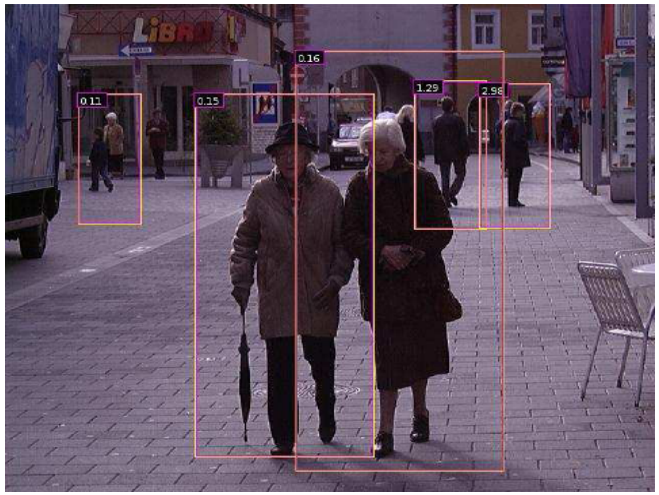
Pedestrian Detector



Positive features

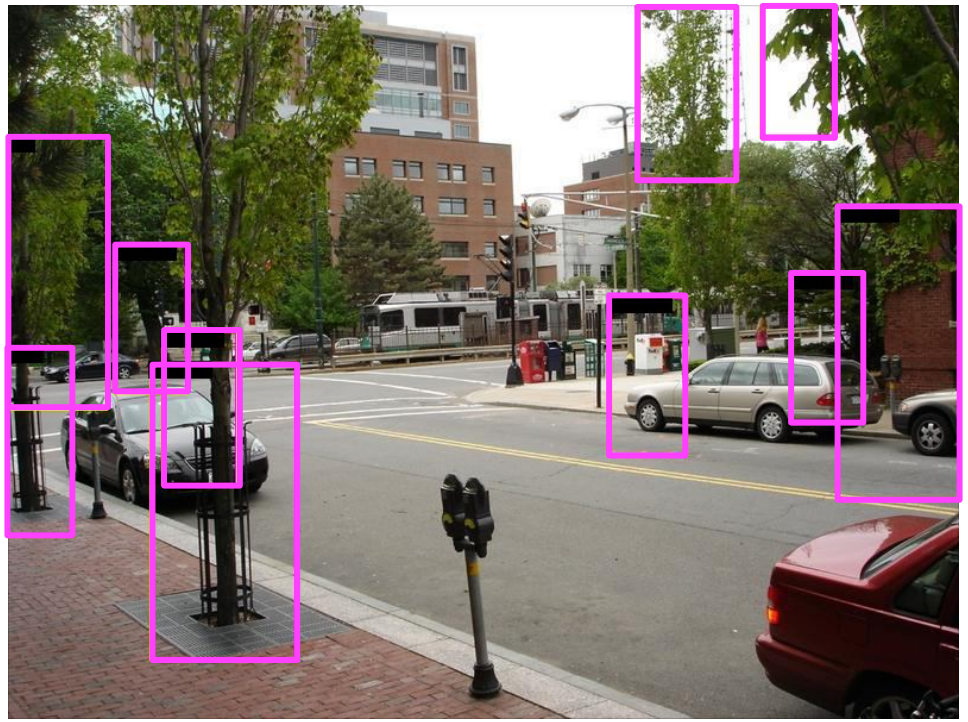
Negative features

Average gradient

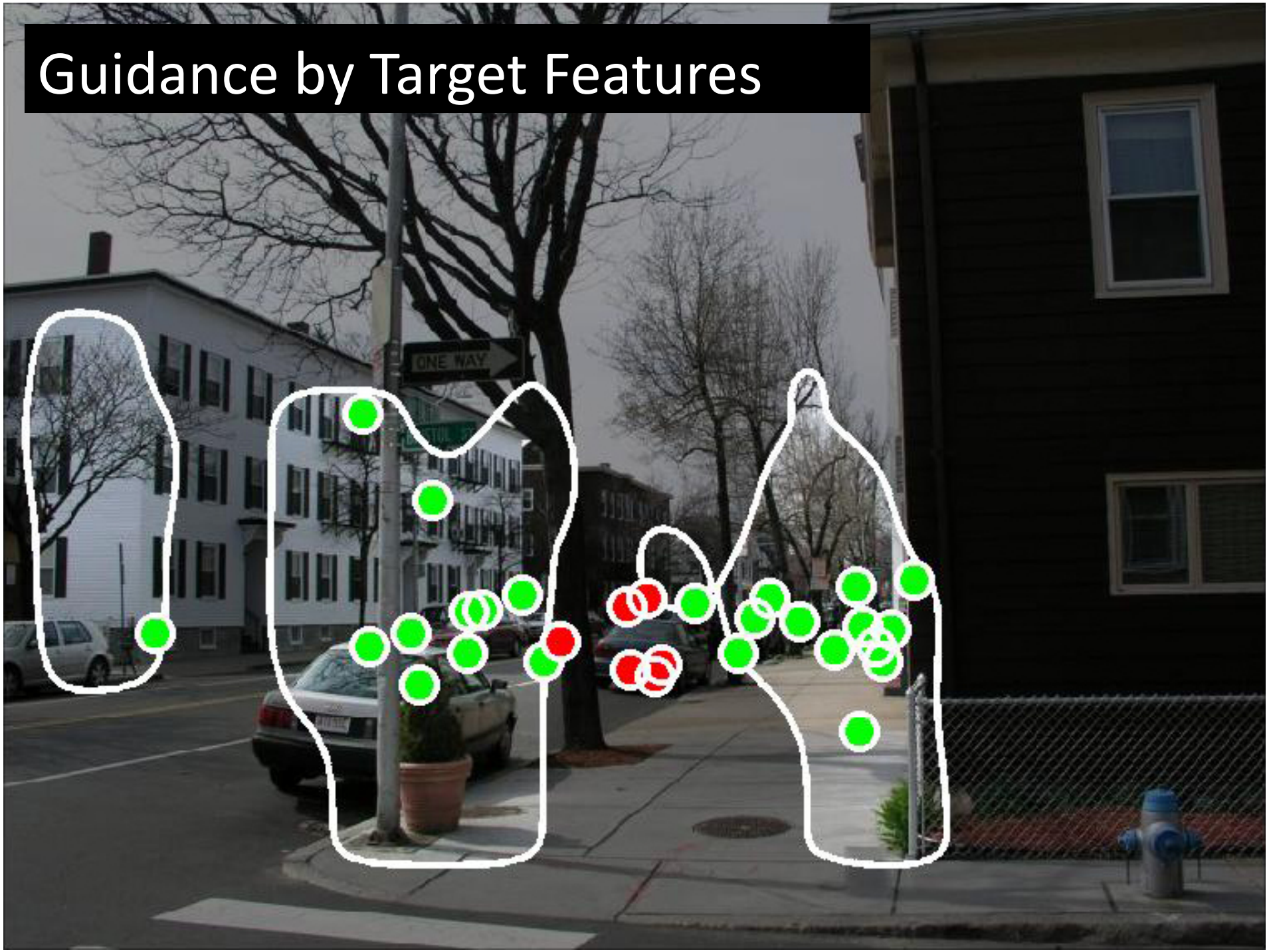


Dalal & Triggs, 2005 CVPR

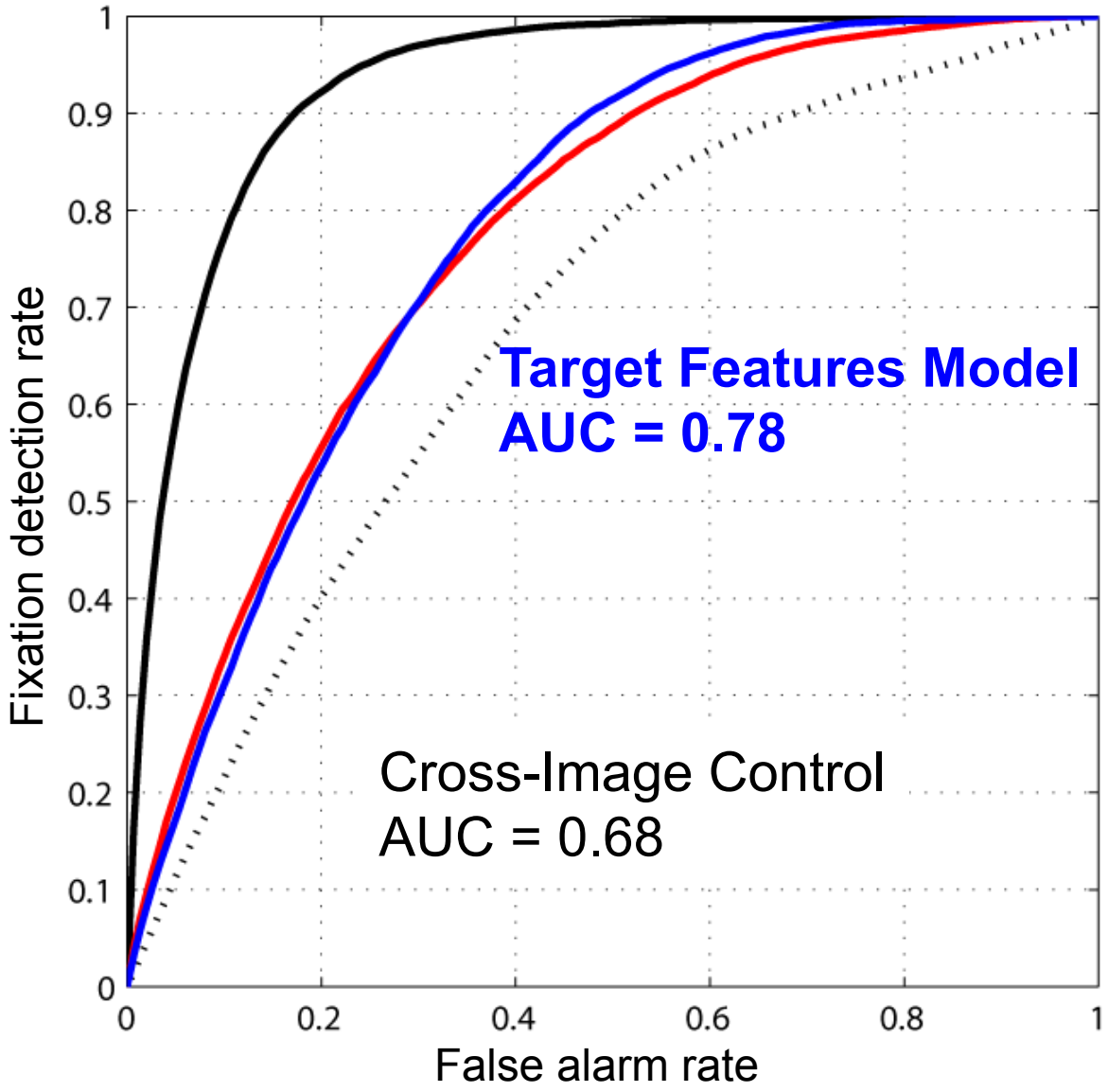
- Histograms of Oriented Gradients (HOG) detector by Dalal & Triggs



Guidance by Target Features



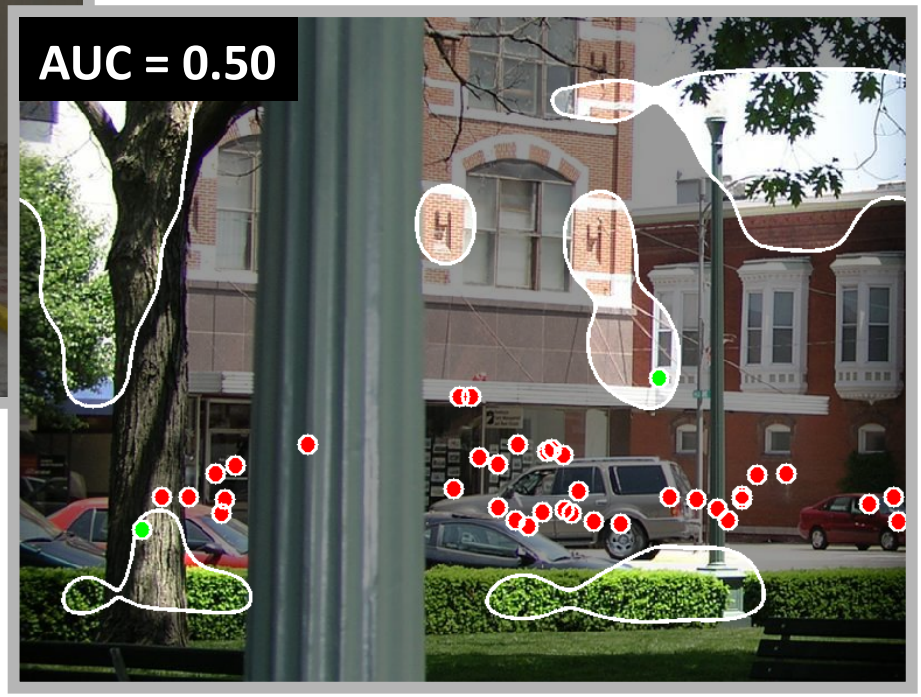
Target Features Model



Human Agreement
AUC = 0.93

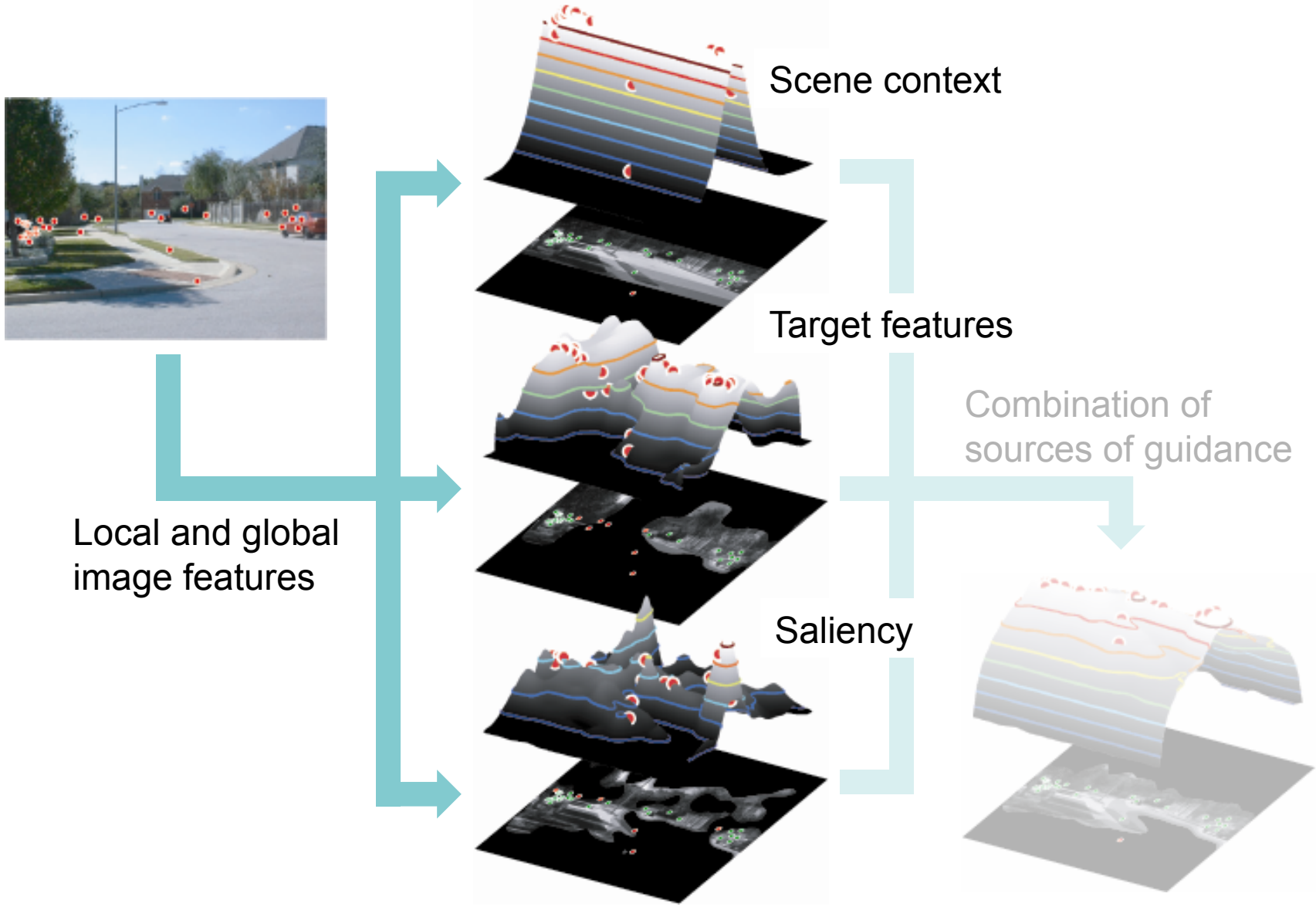
Target Features Model: Examples

Best performance



Worst performance

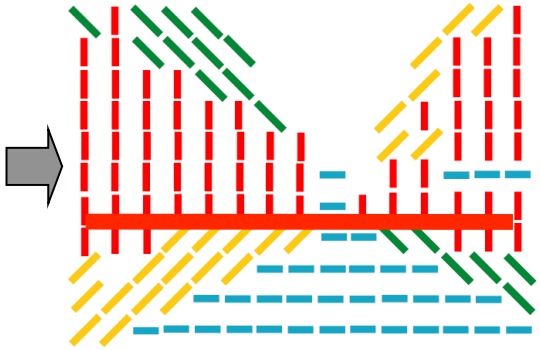
Overview of Model



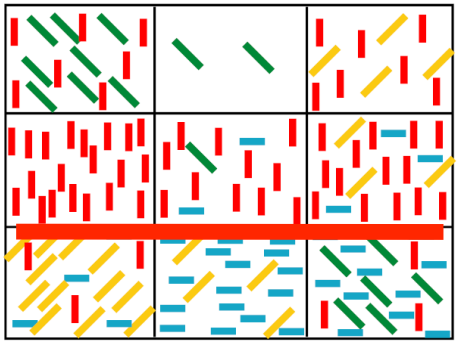
What is the context region for pedestrians?

Scene Context Model (Gist features)

Training image
(contains a pedestrian)



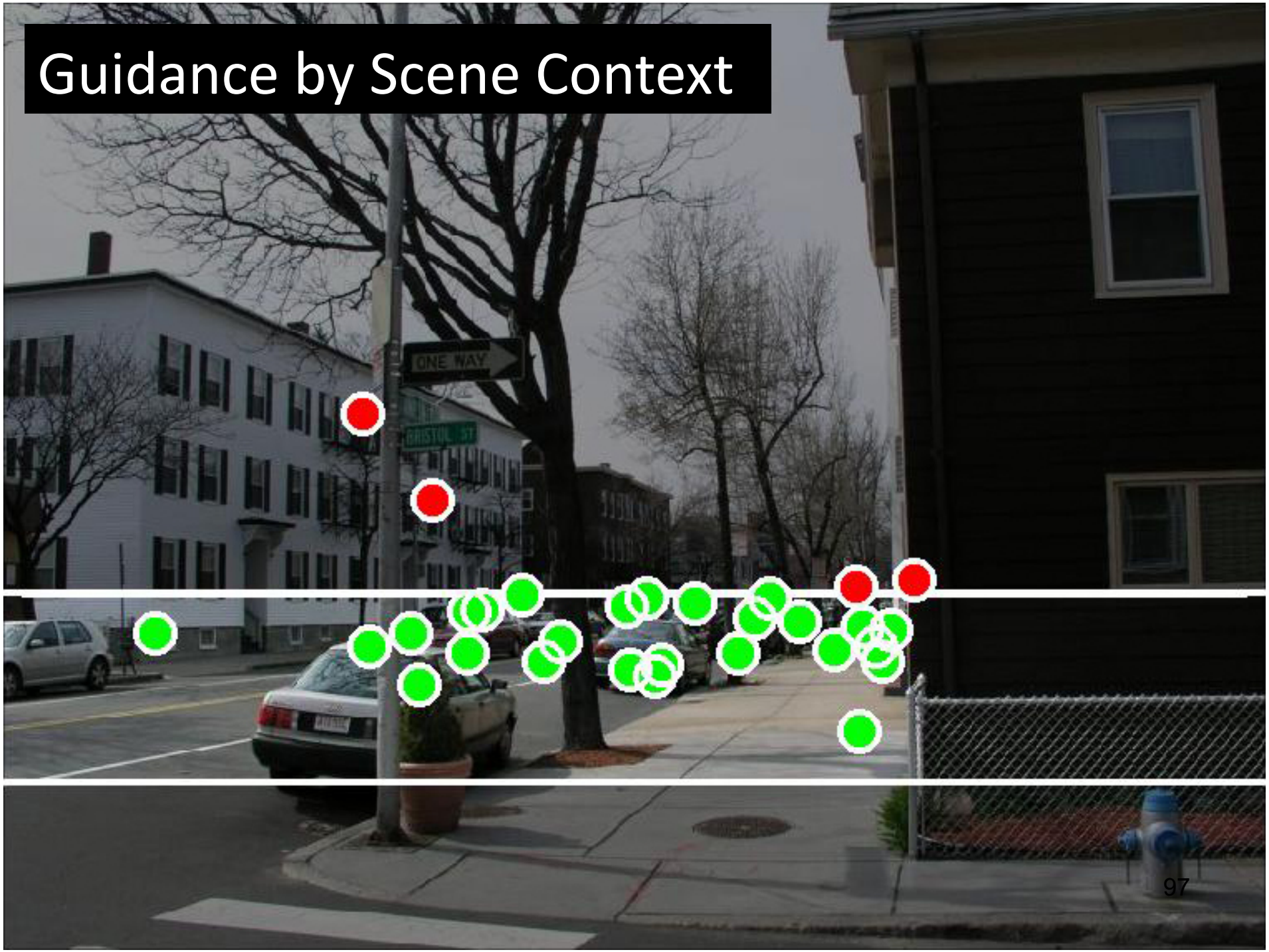
Orientations at various
spatial scales



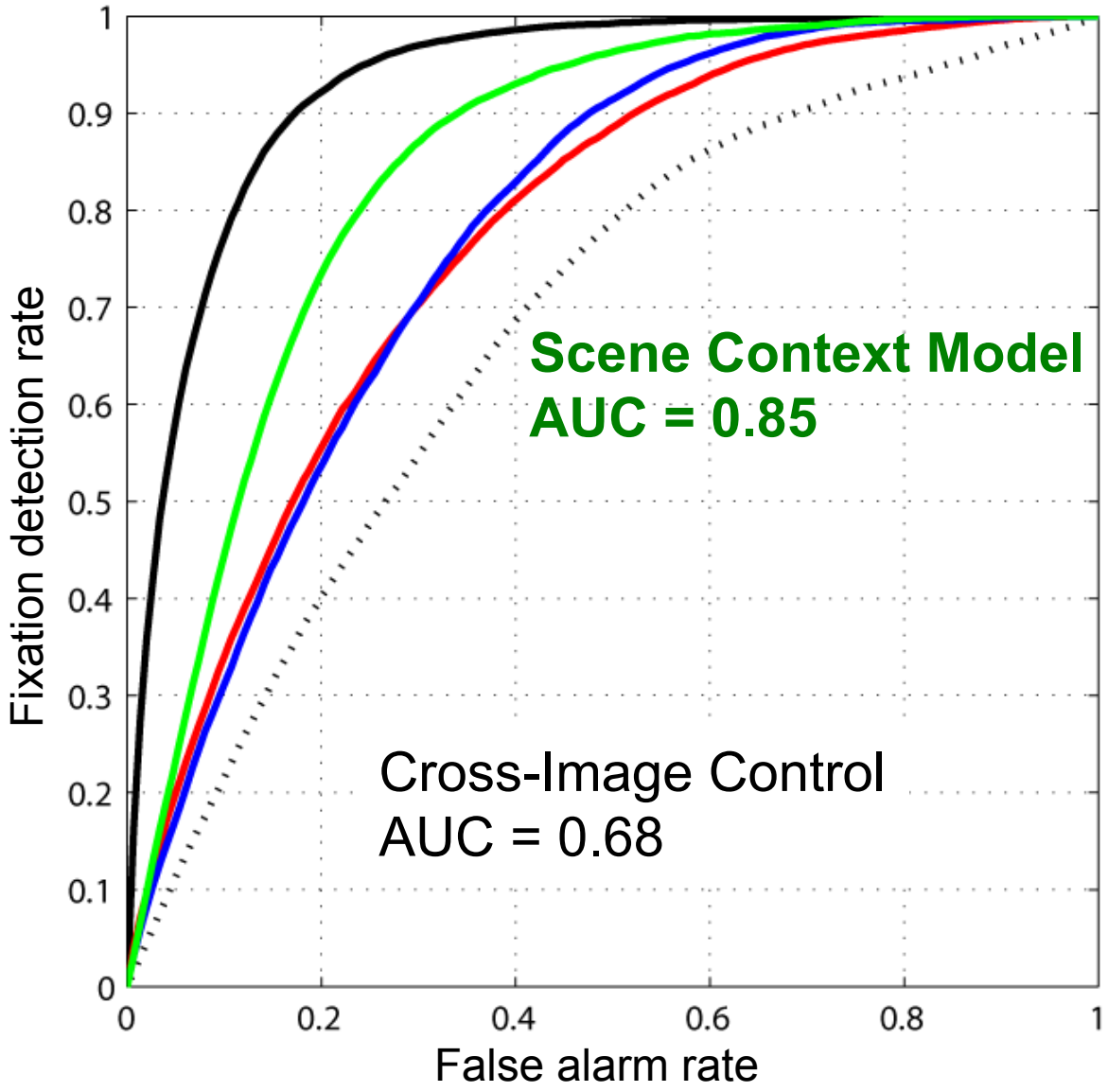
Scene “gist” + position
of pedestrian

Oliva & Torralba, 2001

Guidance by Scene Context



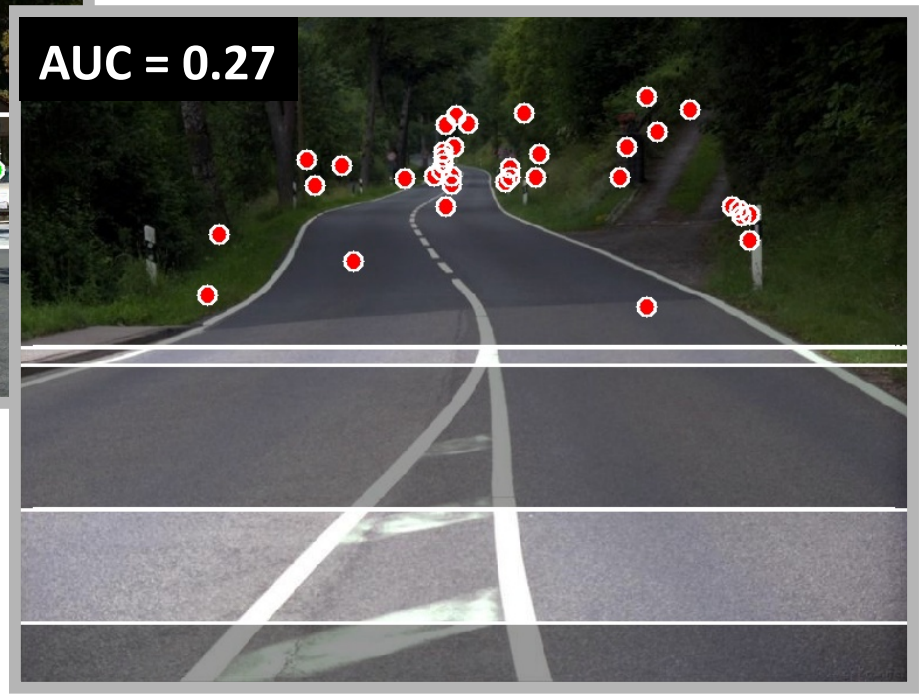
Scene Context Model



Human Agreement
AUC = 0.93

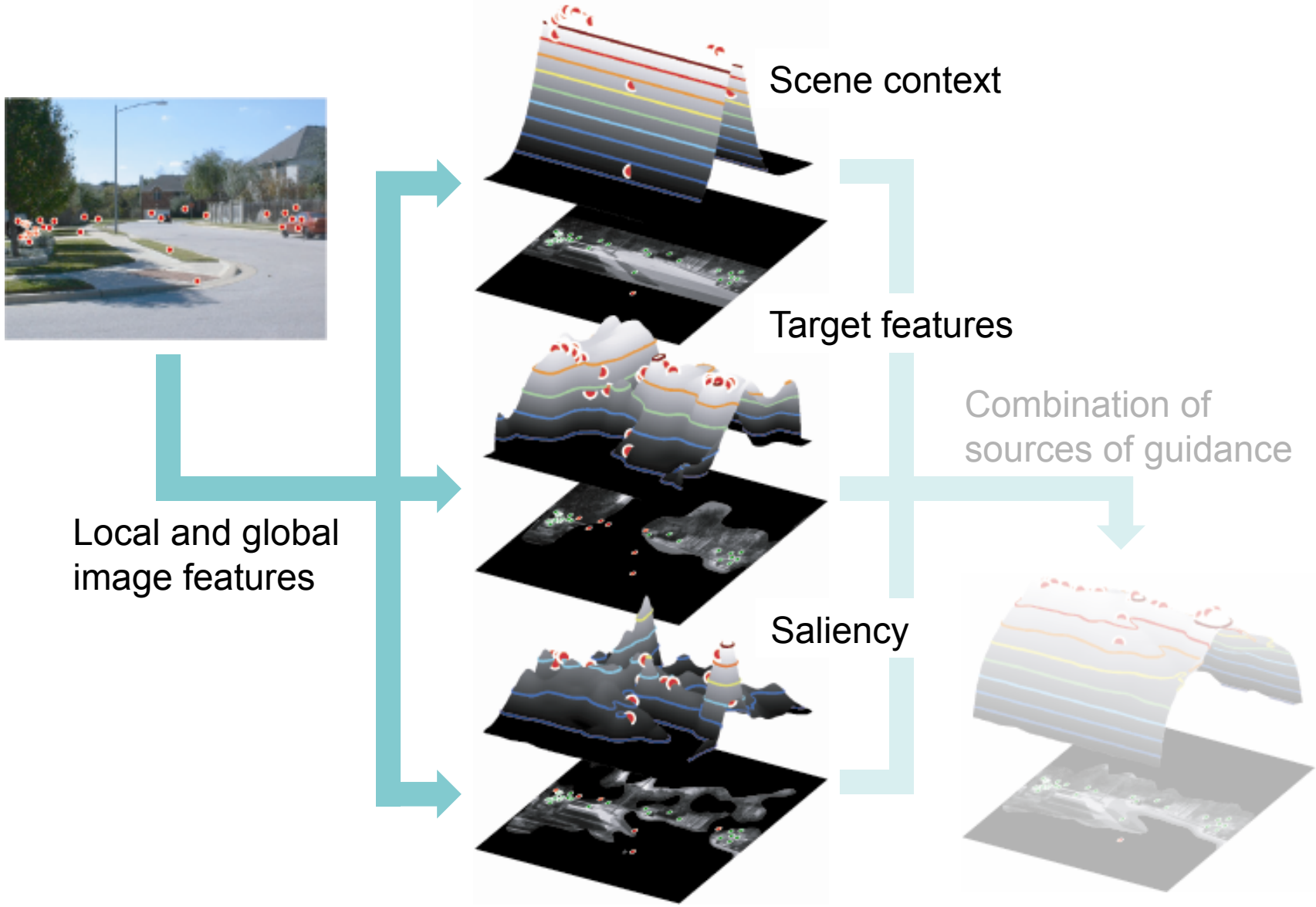
Scene Context Model: Examples

Best performance

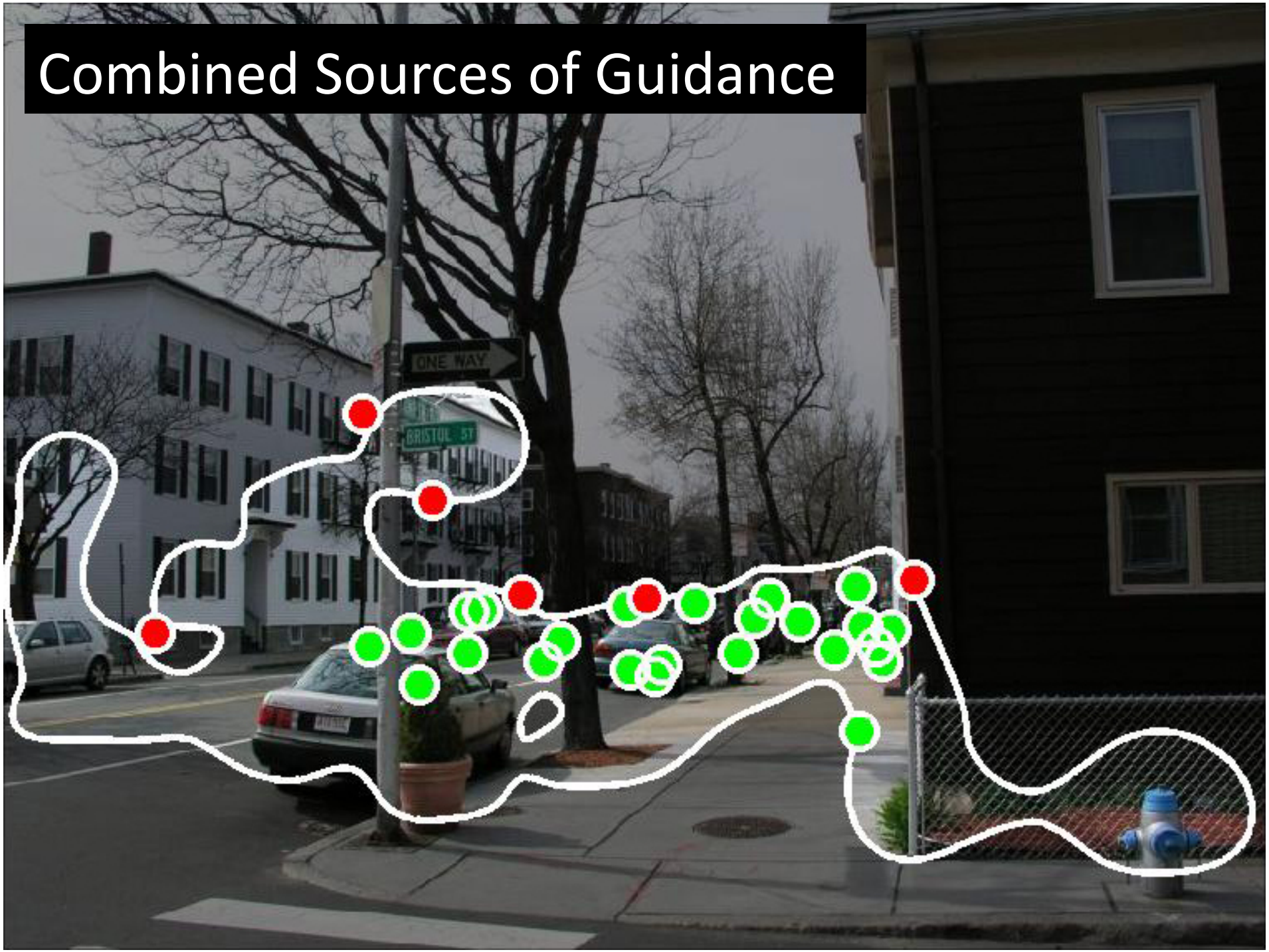


Worst performance

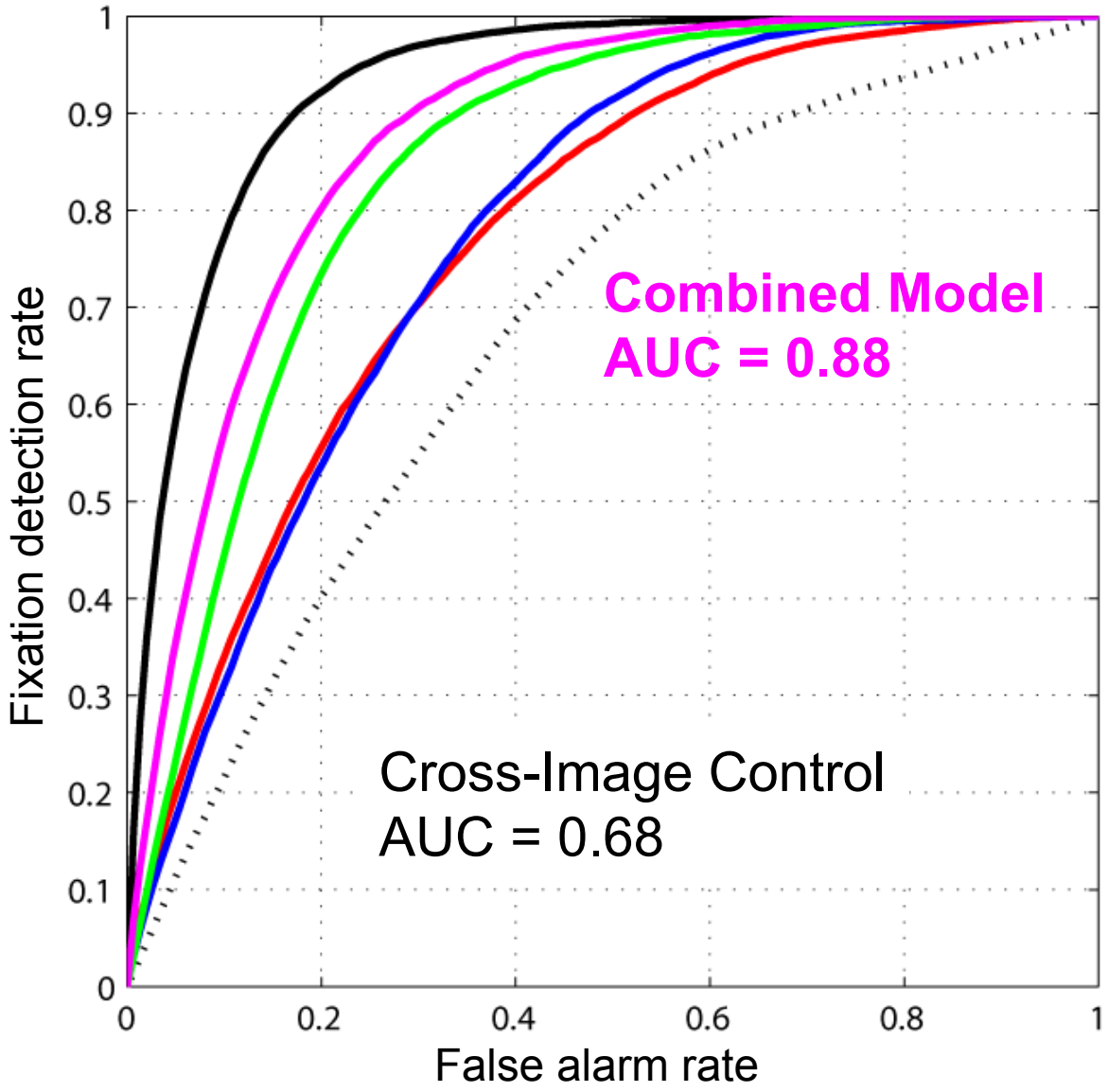
Overview of Model



Combined Sources of Guidance

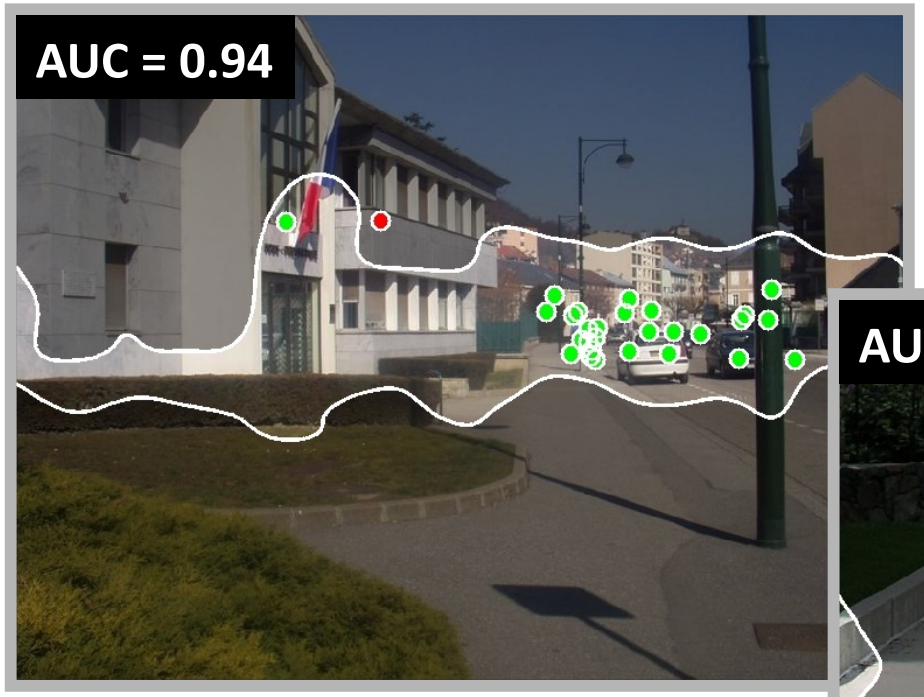


Combined Model

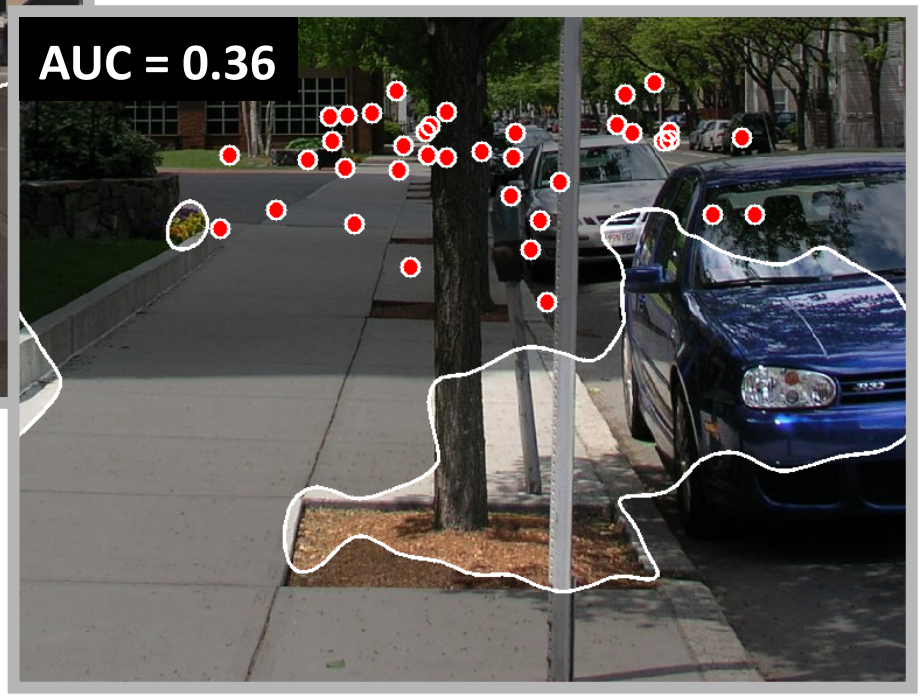


Human Agreement
AUC = 0.93

Combined Model: Examples

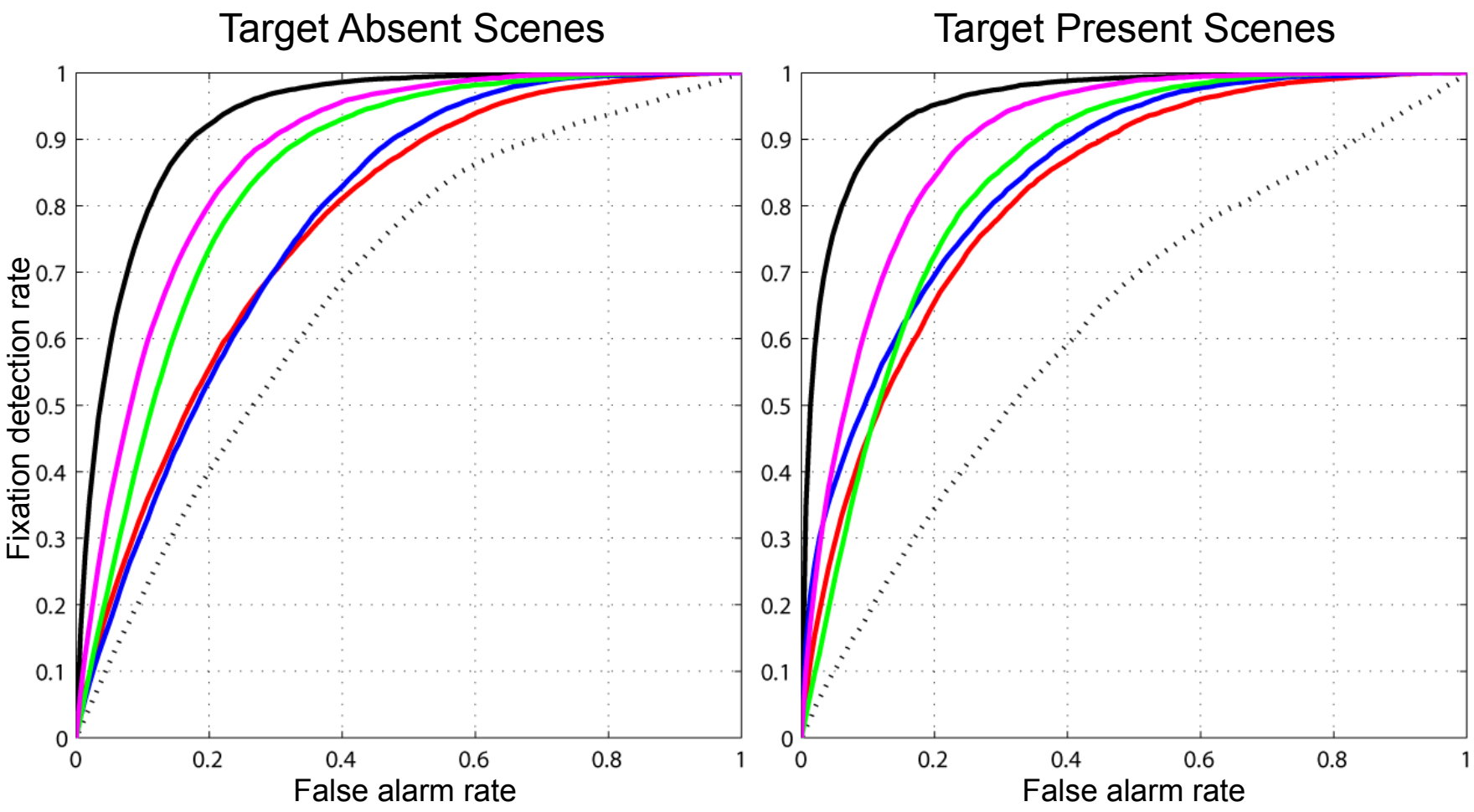


Best performance

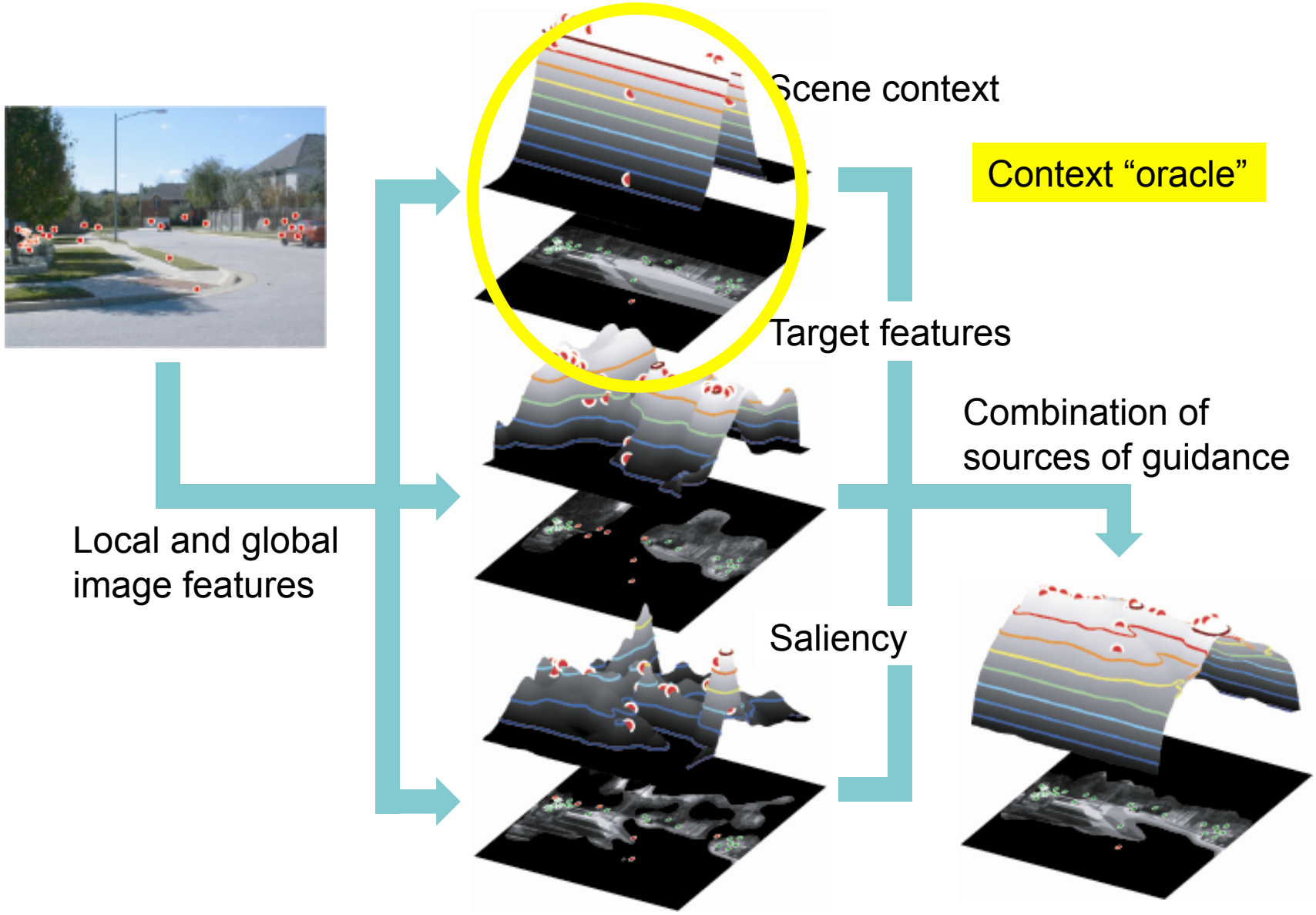


Worst performance

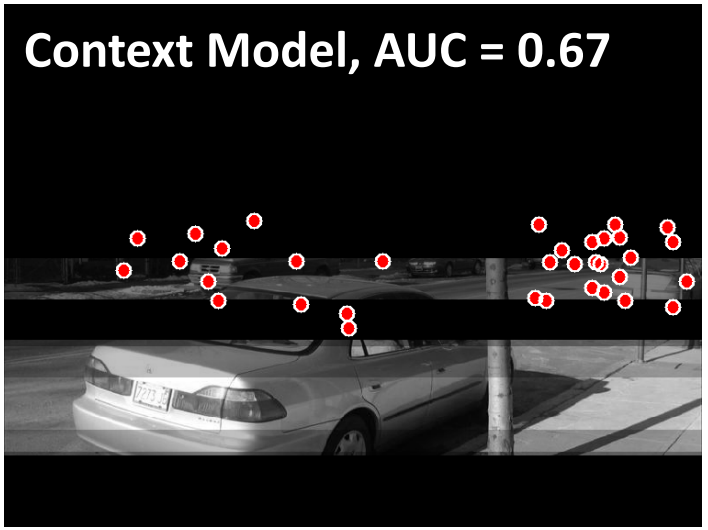
Target Absent vs. Target Present

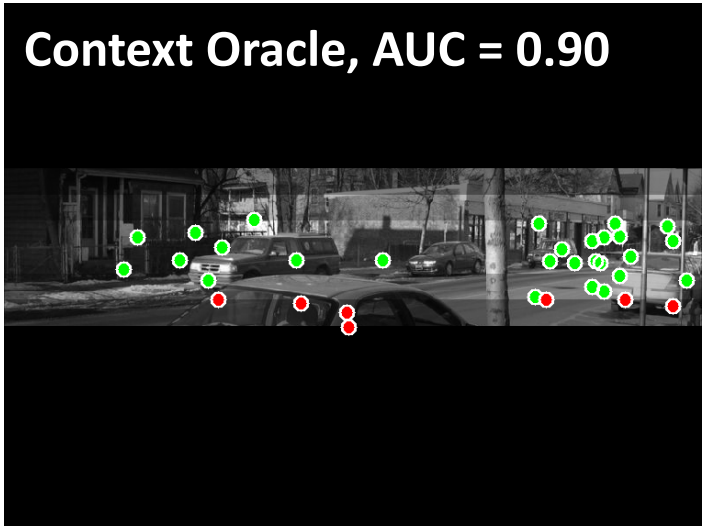


Overview of Model

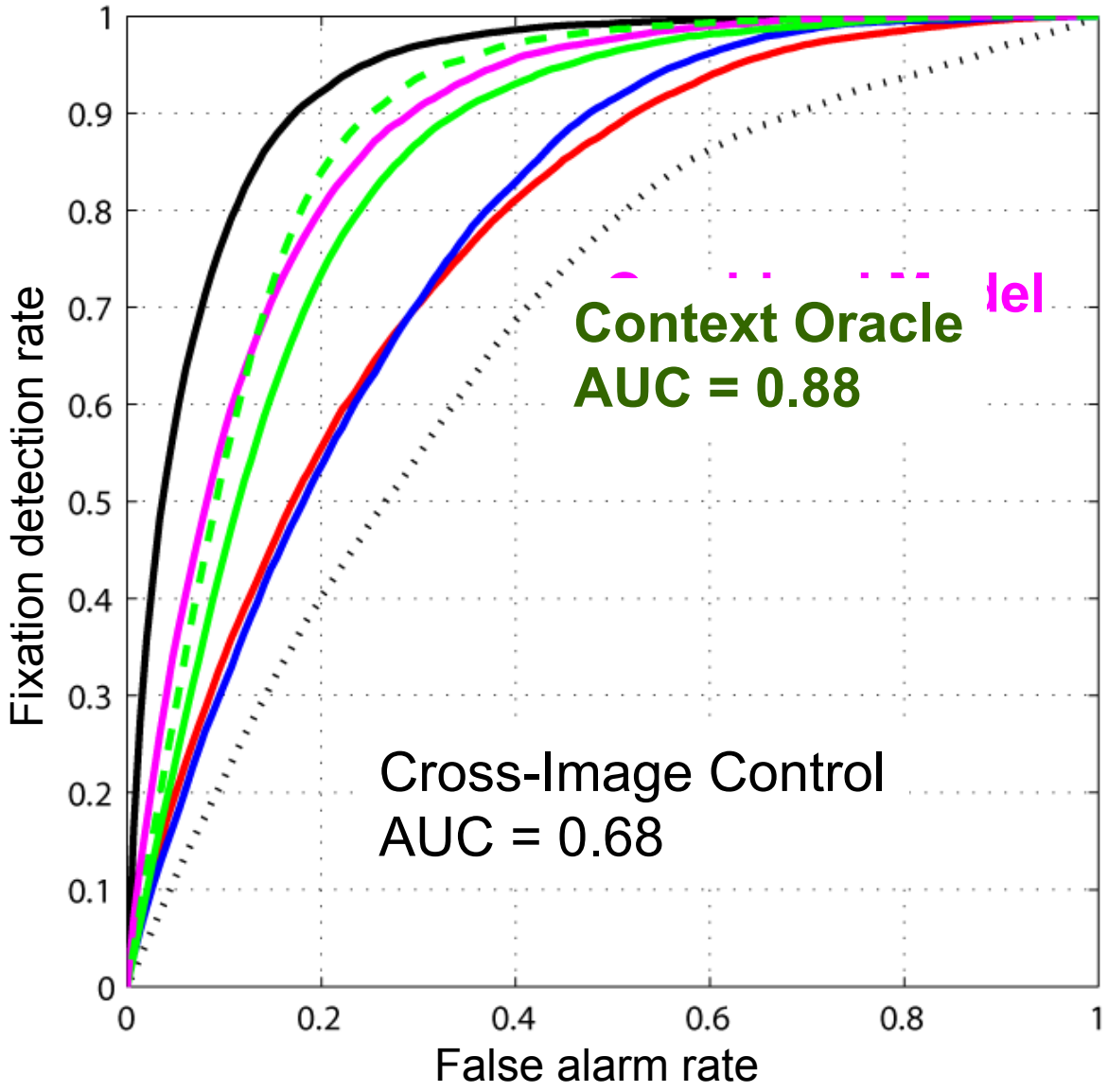


“Context Oracle” Implementation



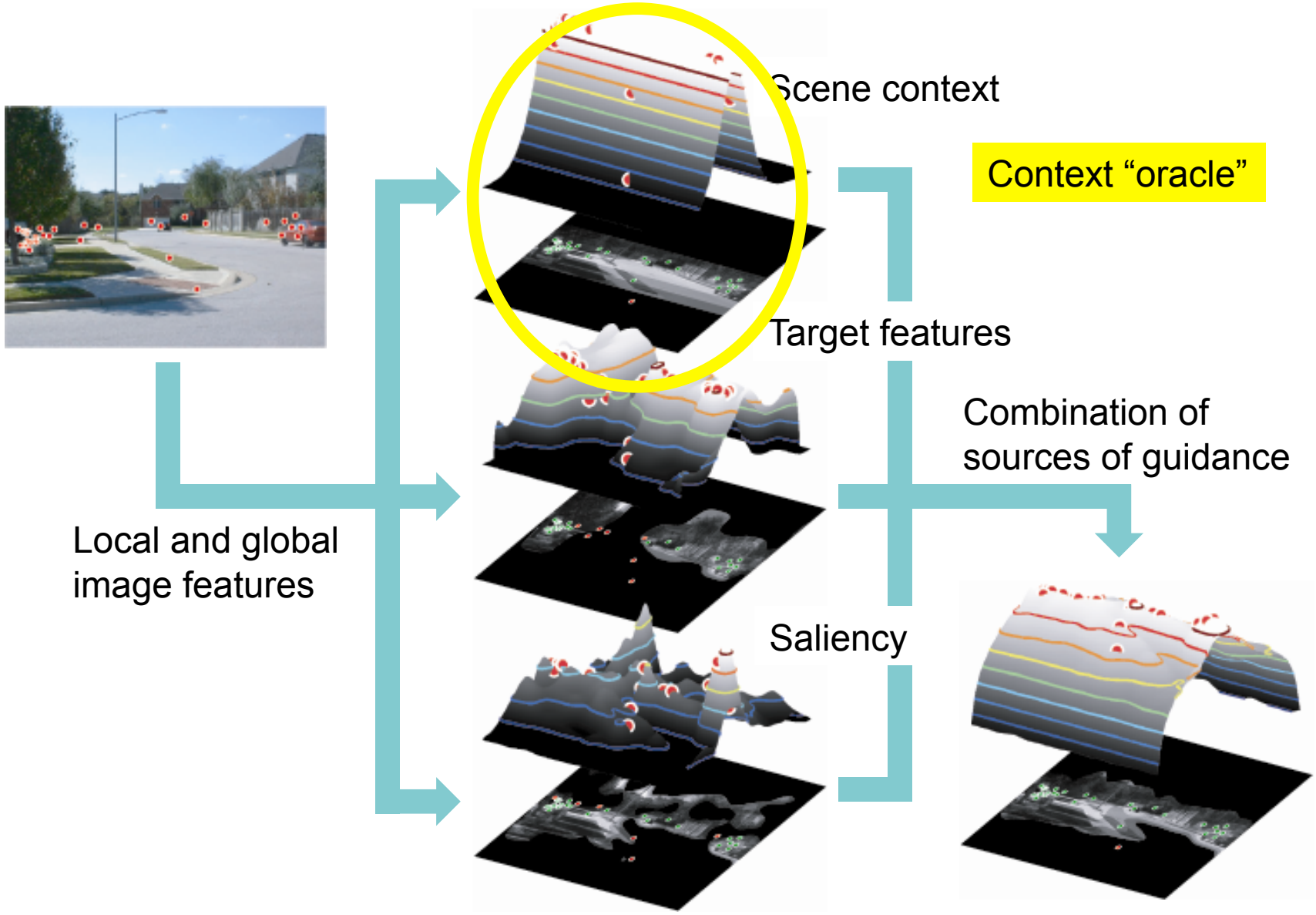


Context Oracle

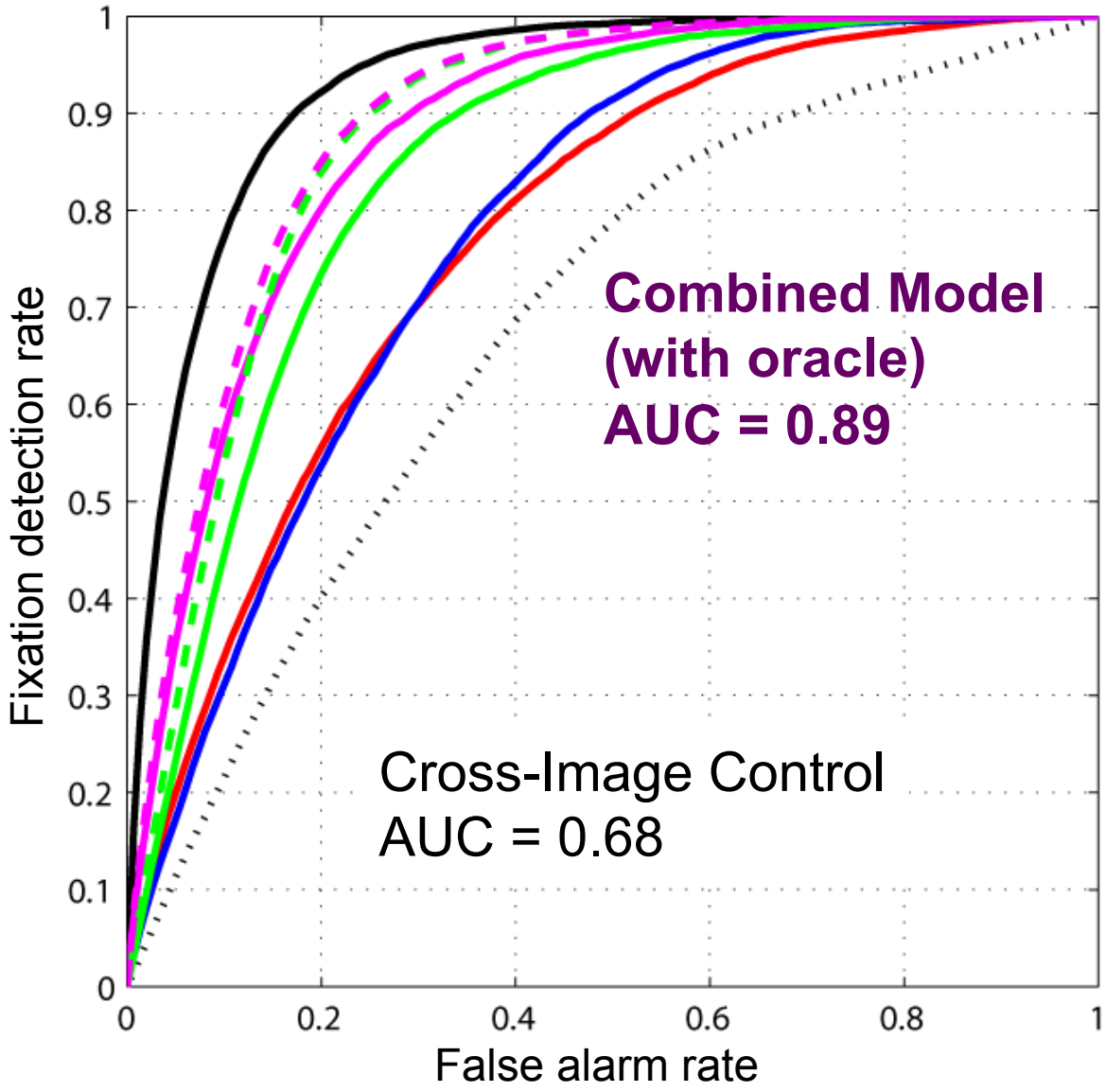


Human Agreement
AUC = 0.93

Overview of Model



Combined Model with Oracle



Human Agreement
AUC = 0.93

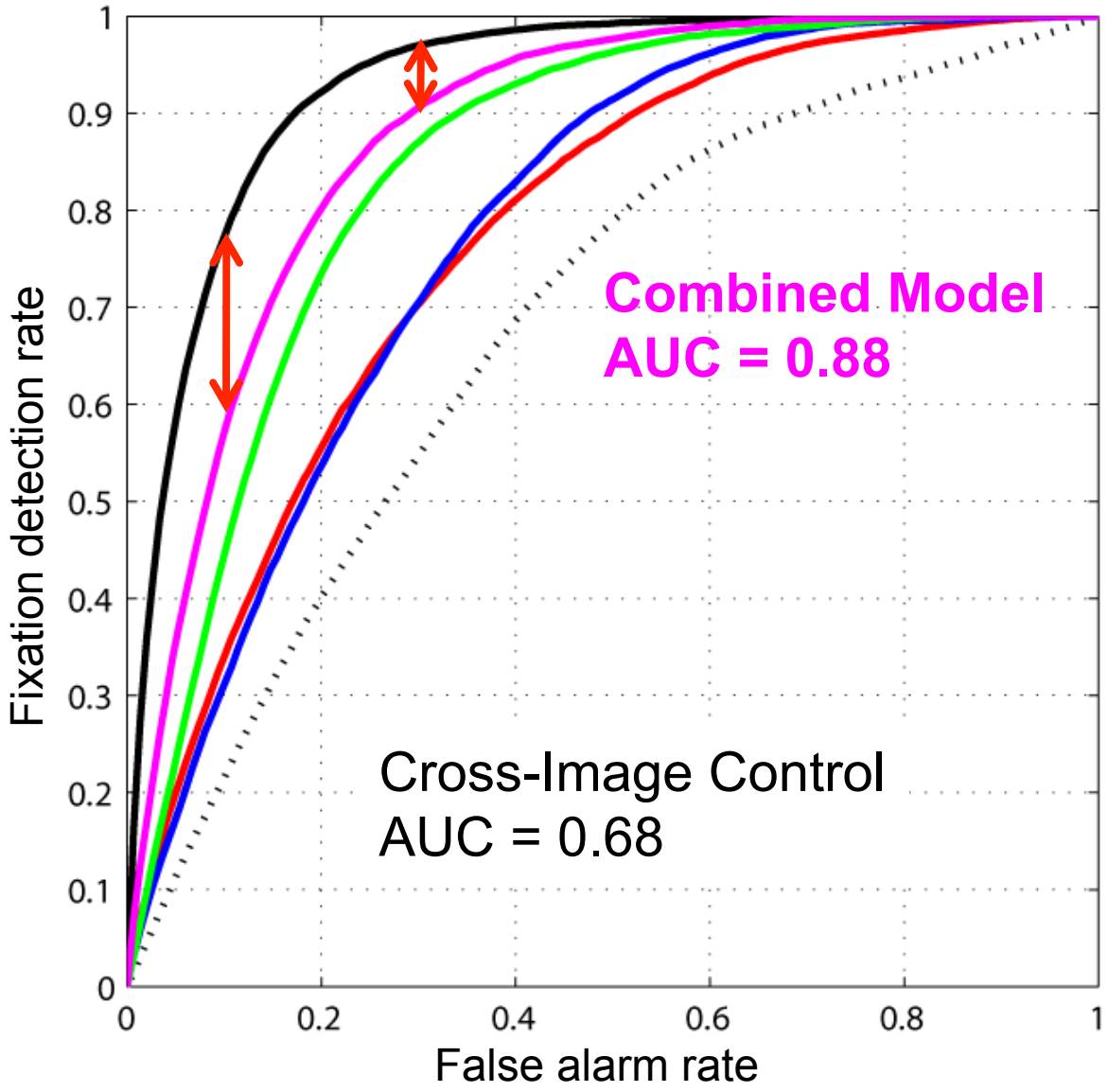
Combined Model
(computational)
AUC = 0.88

Summary of results

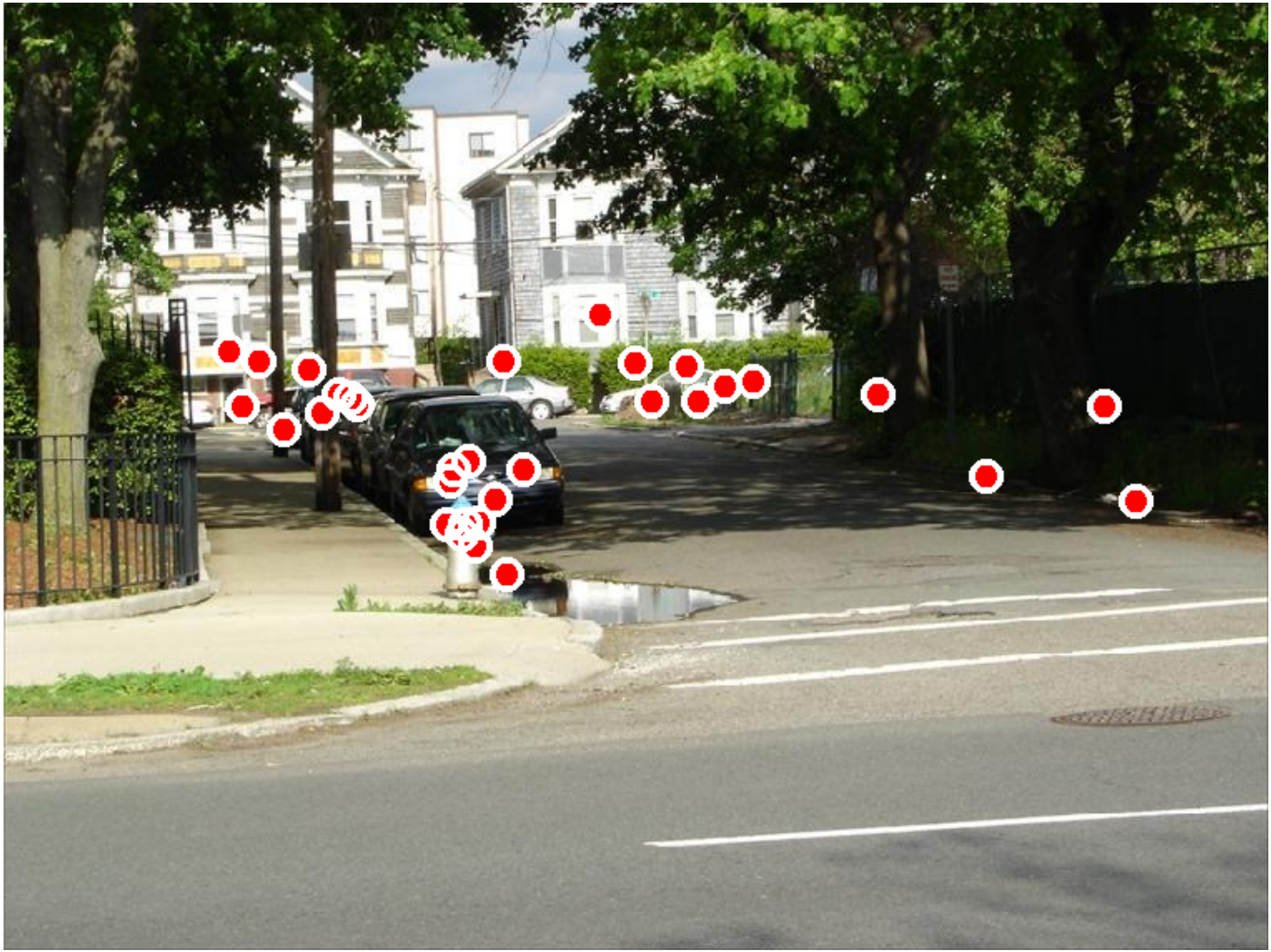
- Combined model accounts for 94% of human agreement in search fixations
- Context predicts human fixations better than saliency or target features in this search task

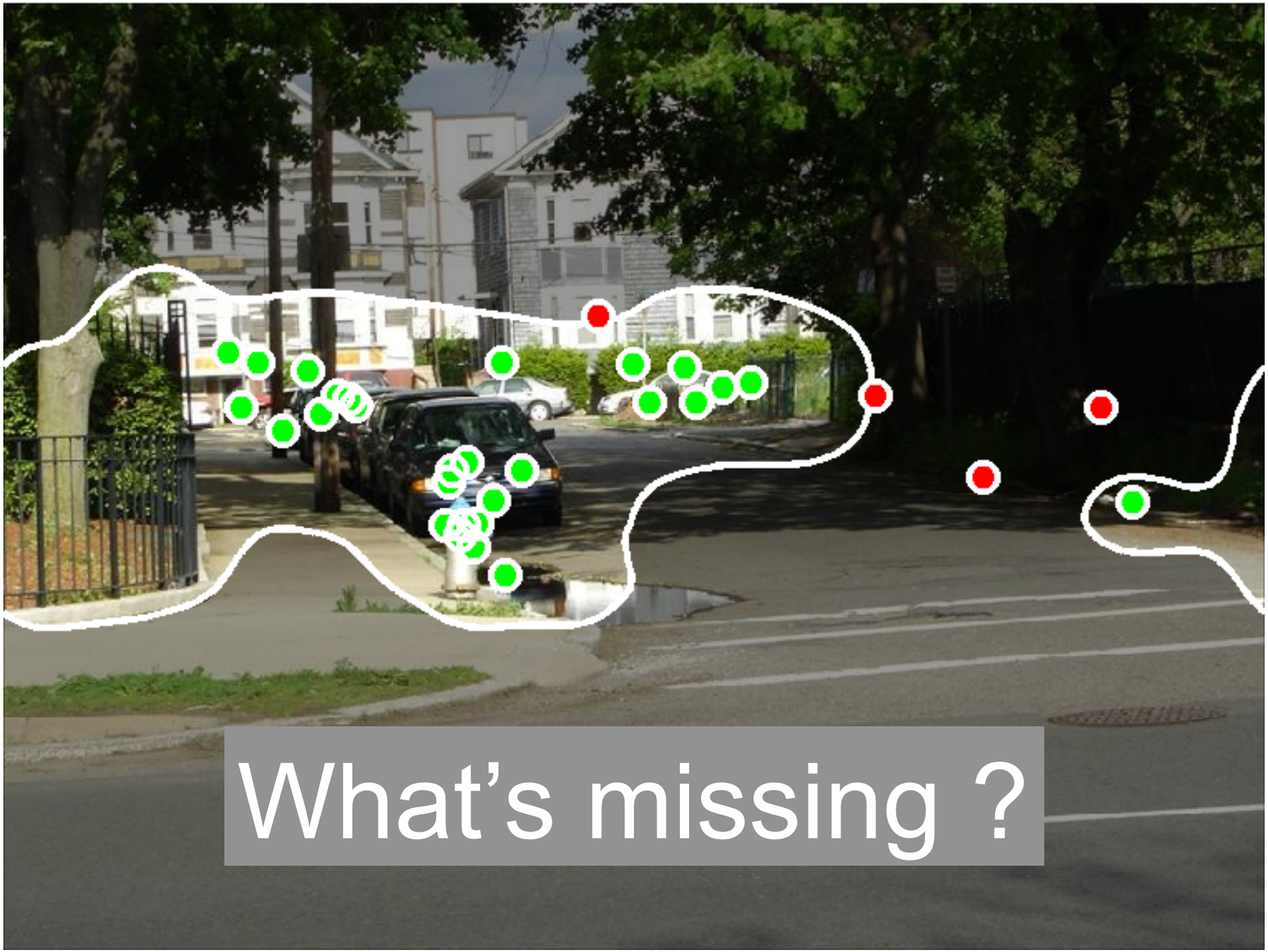
- How to get that last 6%?
  - Context “oracle”?
    - Improves performance to 95% of human agreement
  - Something else?

What's missing?

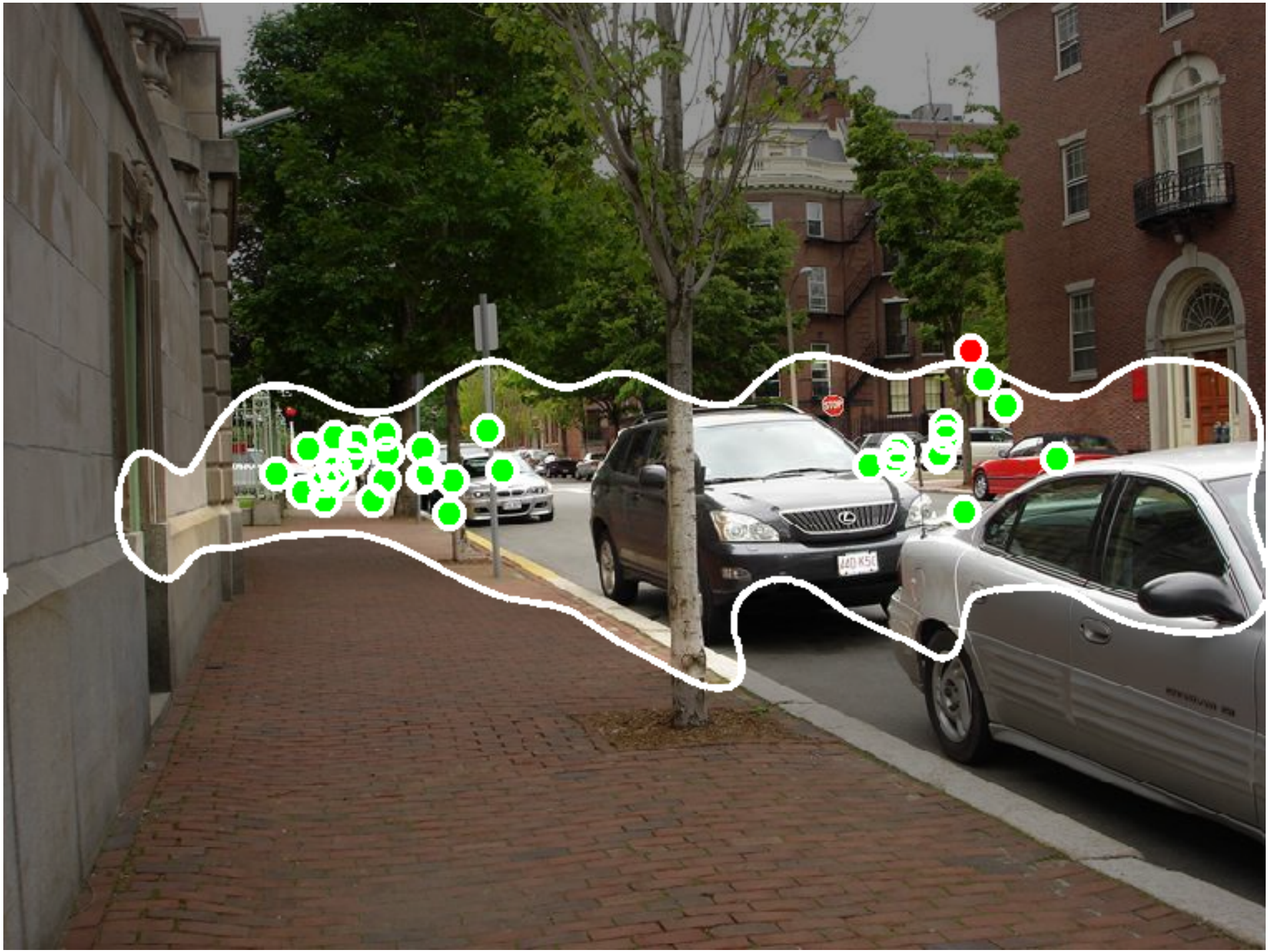


Human Agreement
AUC = 0.93

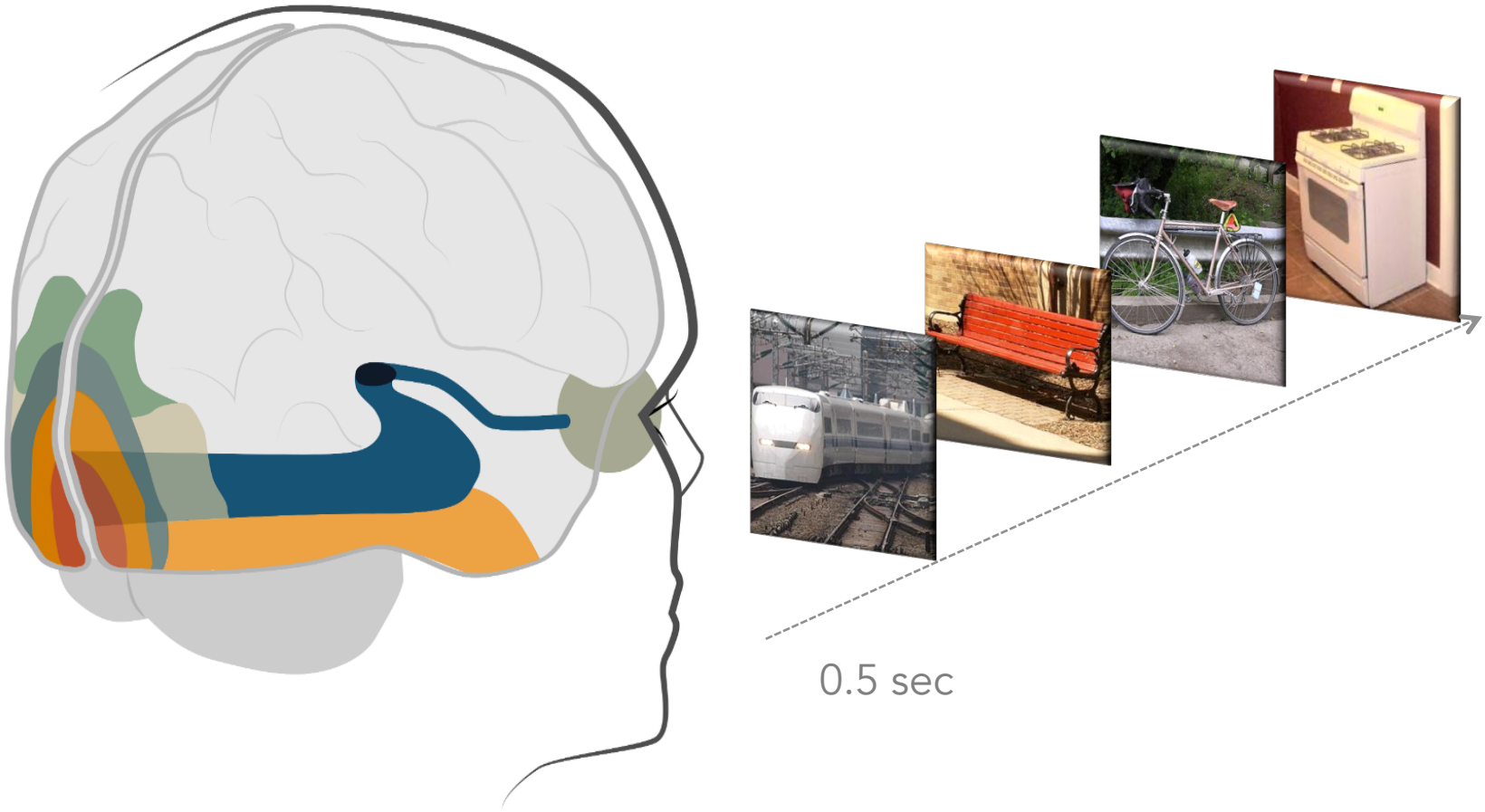



What’s missing ?

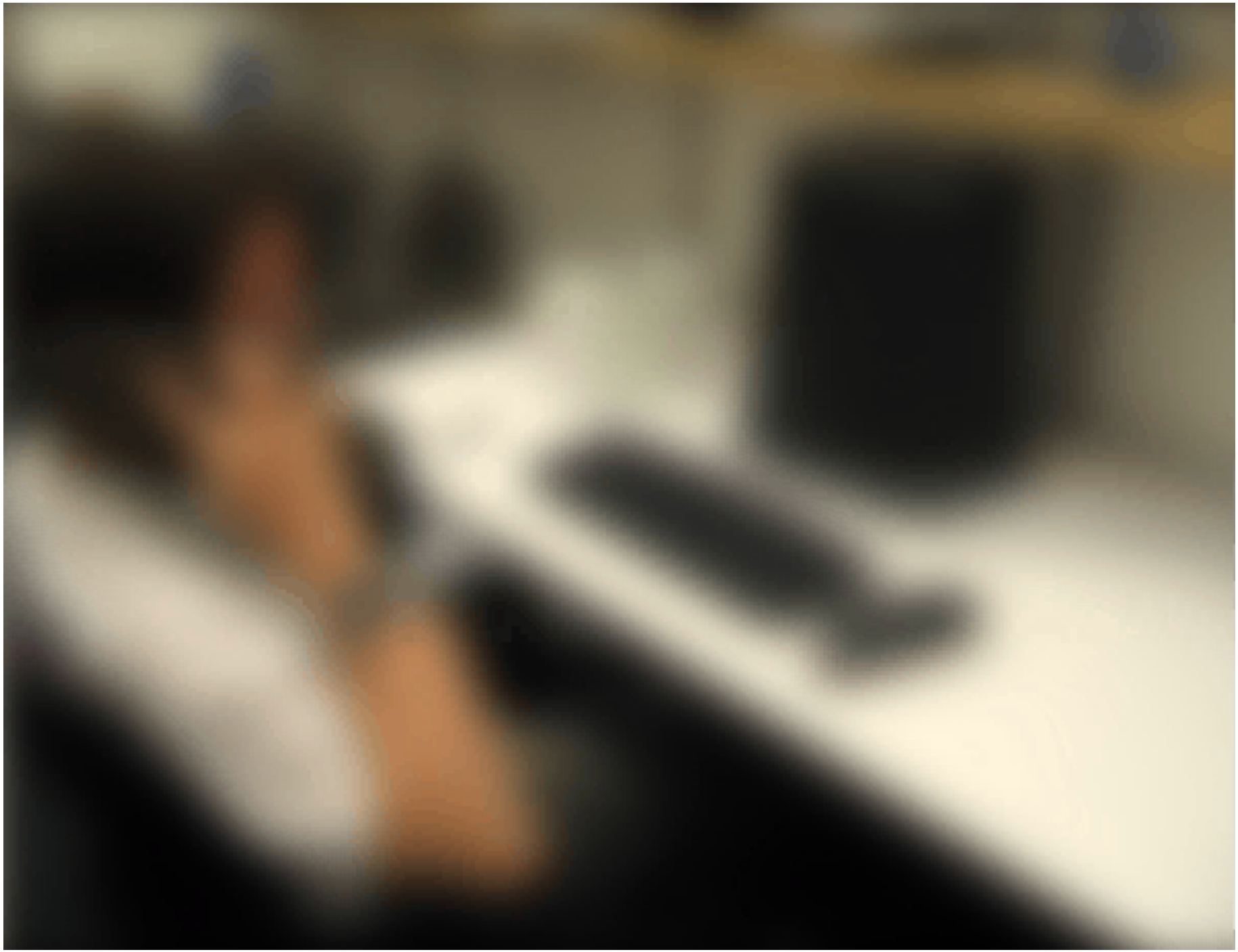
When context fails, we learn



High-Powered Machine: Visual context/scene



Look at the following video and try to make sense of it



Let’s see the real video ...

Standard approach to scene analysis

1) Object representation based on intrinsic features:

2) Detection strategy:

3) The scene representation

Is local information enough?

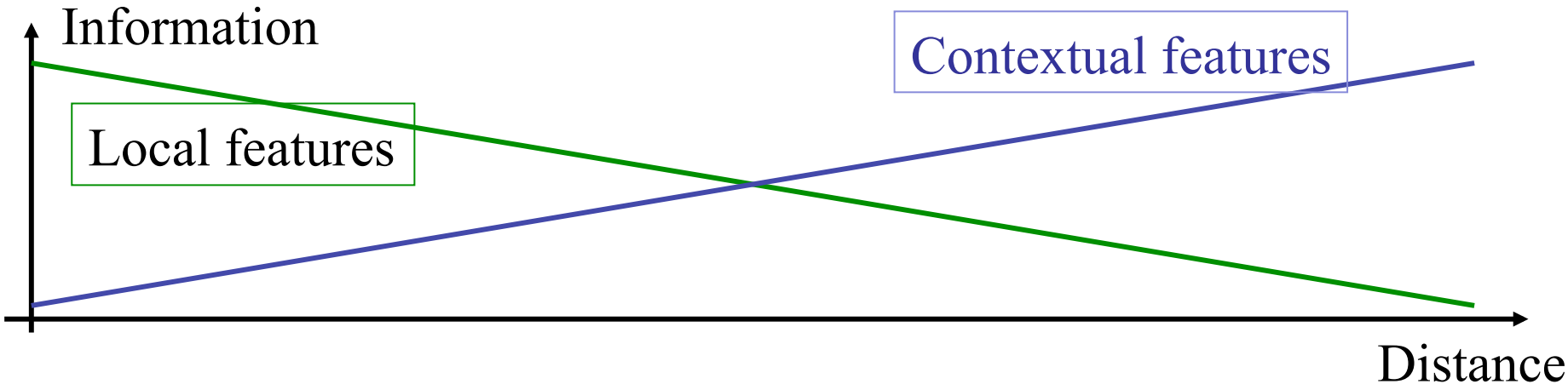
With hundreds of categories

If we have 1000 categories (detectors), and each detector produces 1 fa every 10 images, we will have 100 false alarms per image…

Is local information even enough?

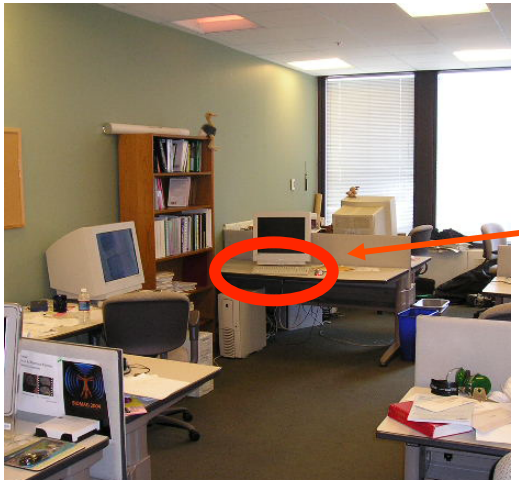




The system does not care about the scene, but we do…

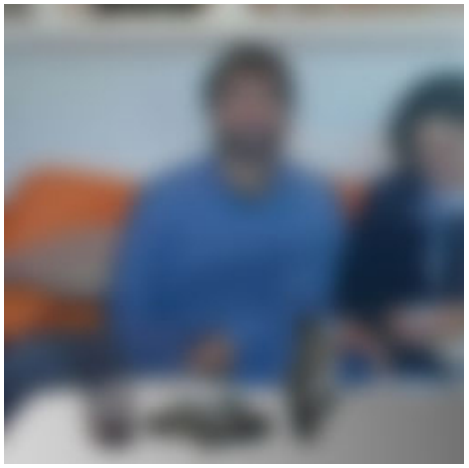
We know there is a keyboard present in this scene even if we cannot see it clearly.

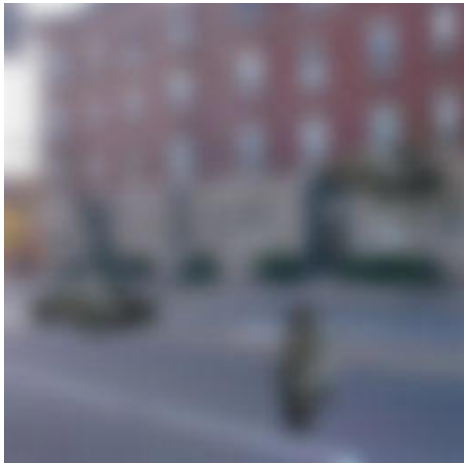


We know there is no keyboard present in this scene

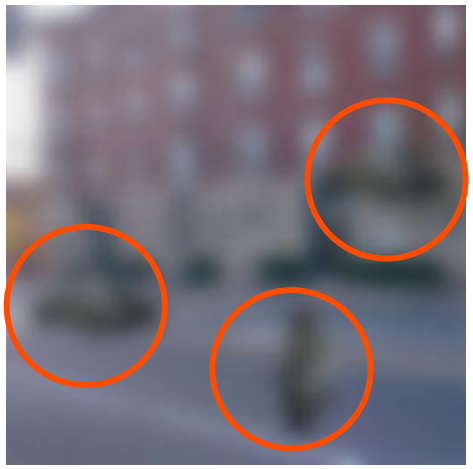
… even if there is one indeed.

The multiple personalities of a blob





The multiple personalities of a blob



A B C

12
13
14

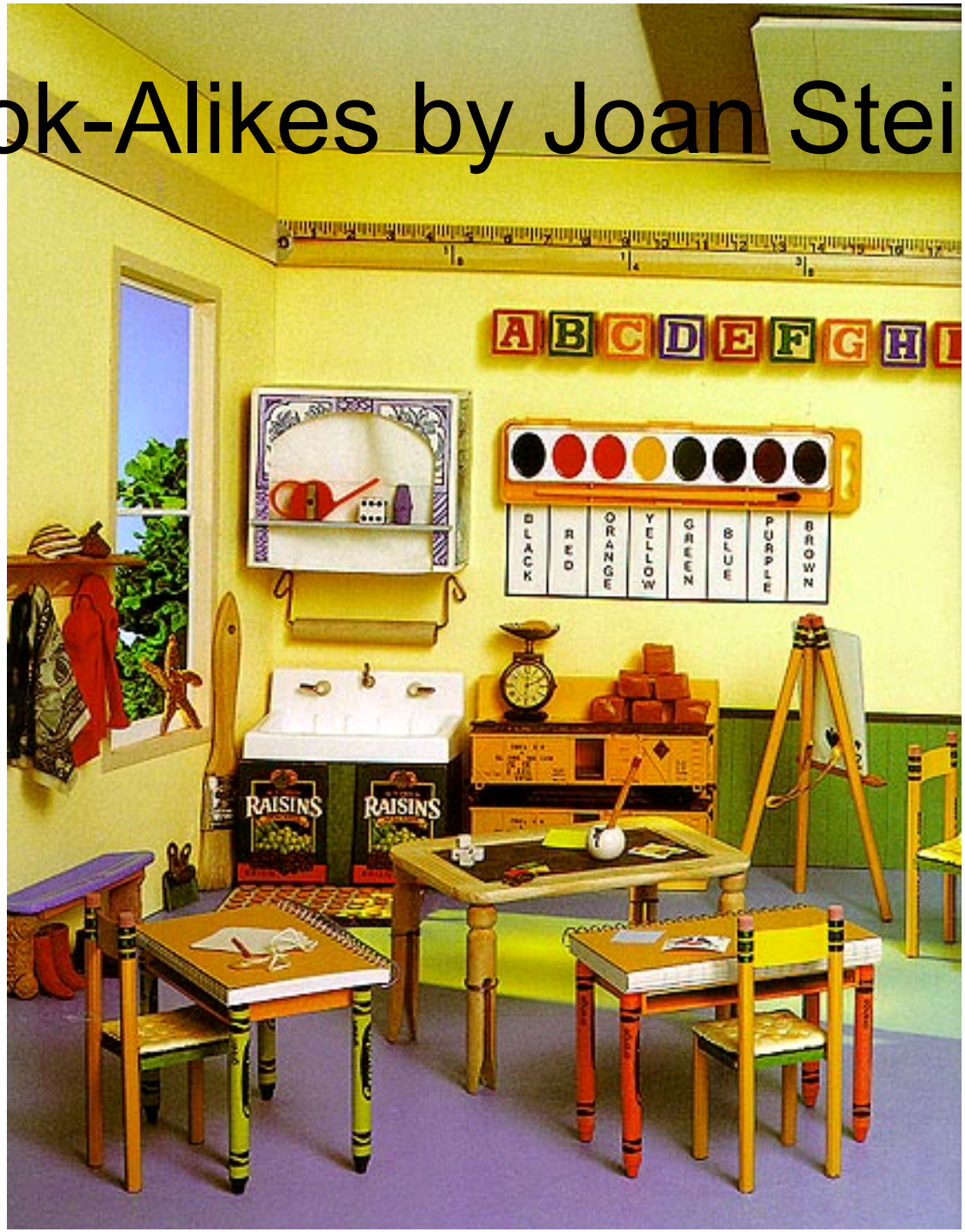
A B C

12
13
14

12
A B C
14

Look-Alikes by Joan Steiner

Look-Alikes by Joan Steiner



Look-Alikes by Joan Steiner

Why is context important?

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

- Context defines what an **unexpected event** is

The influence of an object extends beyond its physical boundaries

The context challenge

How far can you go without using an object detector?

What are the hidden objects?

What are the hidden objects?

The importance of context

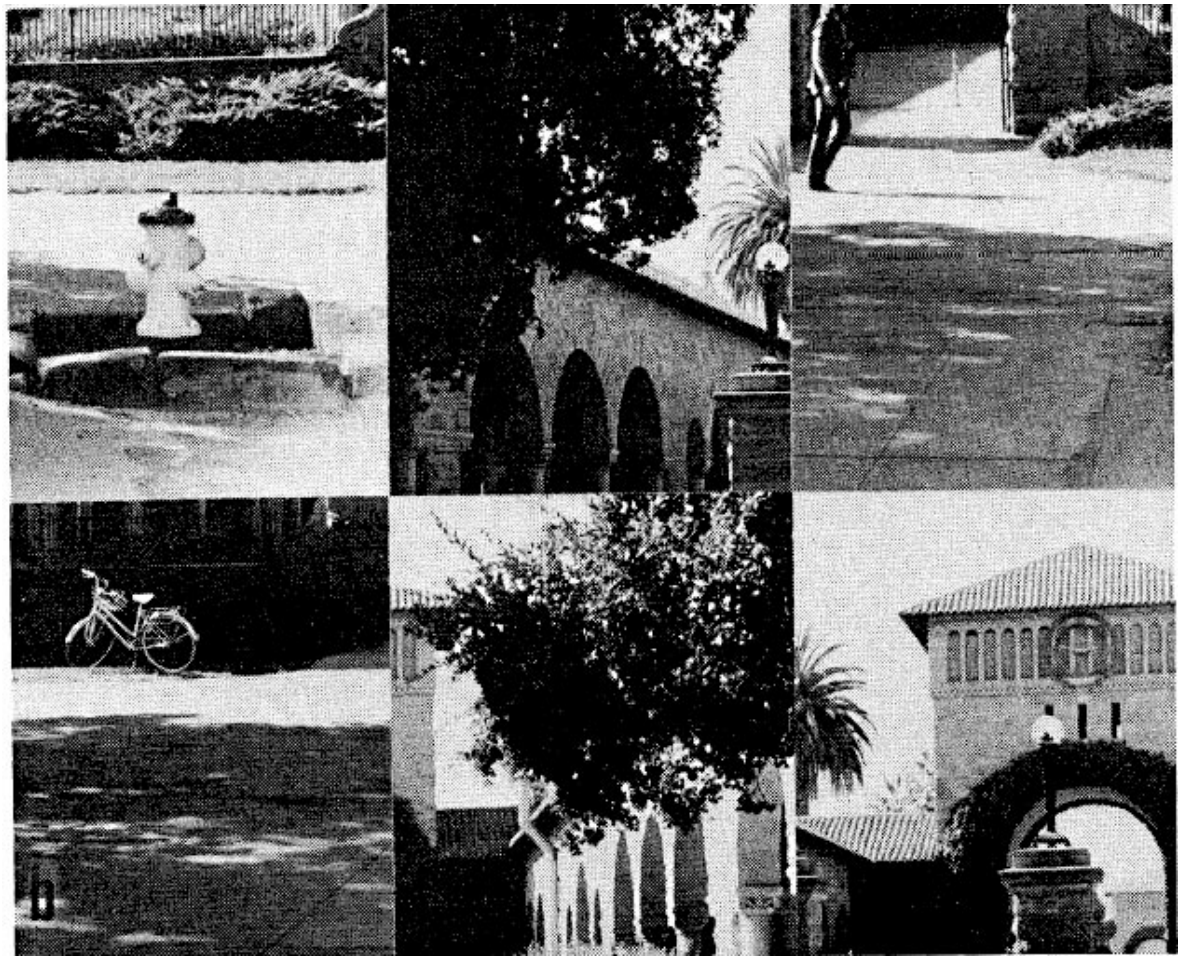
- Cognitive psychology
  - Palmer 1975
  - Biederman 1981
  - …

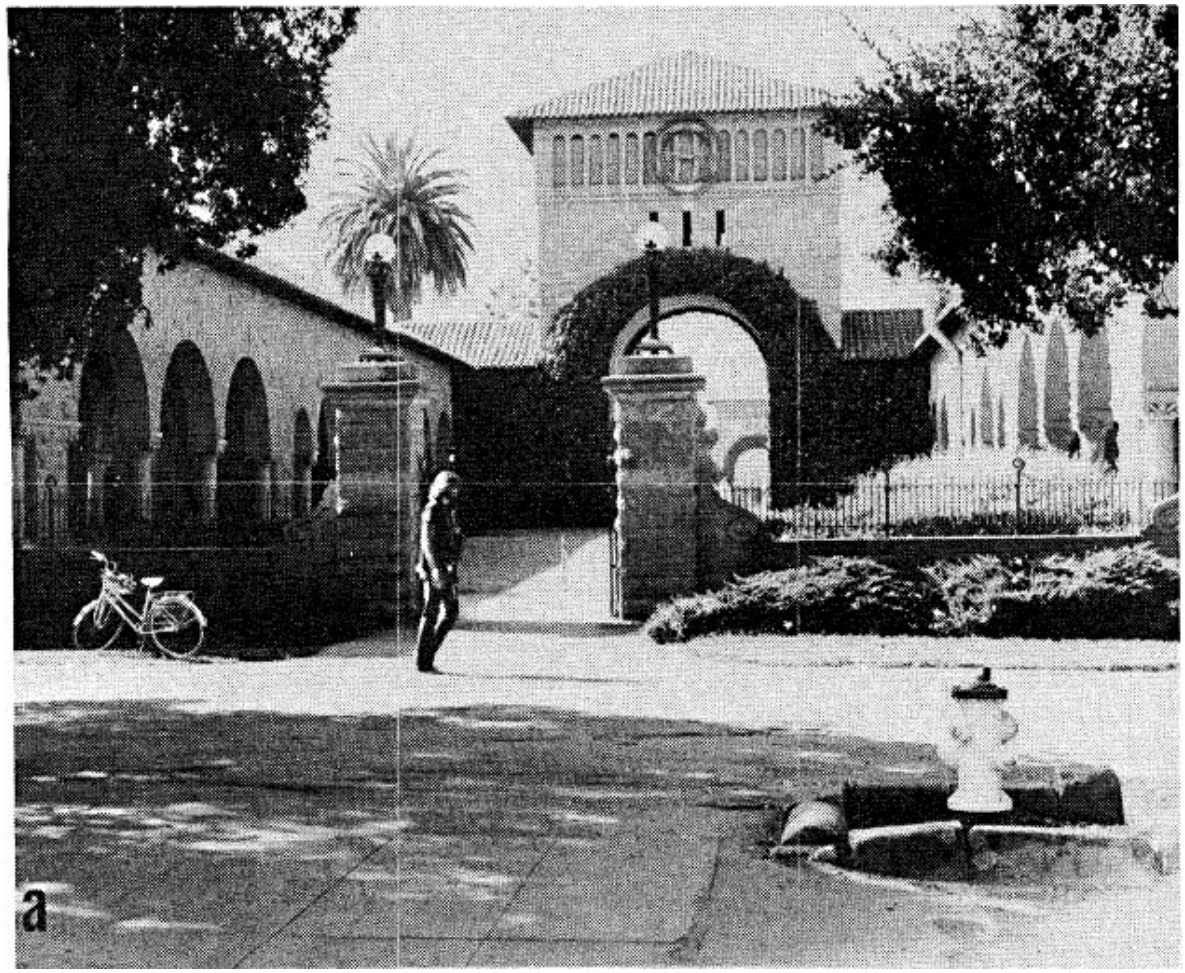
- Computer vision
  - Noton and Stark (1971)
  - Hanson and Riseman (1978)
  - Barrow & Tenenbaum (1978)
  - Ohta, kanade, Skai (1978)
  - Haralick (1983)
  - Strat and Fischler (1991)
  - Bobick and Pinhanez (1995)
  - Campbell et al (1997)

Class	Context elements	Operator
SKY	ALWAYS	ABOVE-HORIZON
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-CLEAR ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	BLUE
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	BRIGHT
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY	UNTEXTURED
SKY	SKY-IS-OVERCAST ∧ TIME-IS-DAY ∧ RGB-IS-AVAILABLE	WHITE
SKY	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGE-IS-UNDEFINED
SKY	CAMERA-IS-HORIZONTAL	NEAR-TOP
SKY	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-sky)	ABOVE-SKYLINE
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-INTENSITY
SKY	CLIQUE-CONTAINS(sky)	SIMILAR-TEXTURE
SKY	RGB-IS-AVAILABLE ∧ CLIQUE-CONTAINS(sky)	SIMILAR-COLOR
GROUND	CAMERA-IS-HORIZONTAL	HORIZONTALLY-STRIATED
GROUND	CAMERA-IS-HORIZONTAL	NEAR-BOTTOM
GROUND	SPARSE-RANGE-IS-AVAILABLE	SPARSE-RANGES-FORM-HORIZONTA
GROUND	DENSE-RANGE-IS-AVAILABLE	DENSE-RANGES-FORM-HORIZONTA
GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(complete-ground)	BELOW-SKYLINE
GROUND	CAMERA-IS-HORIZONTAL ∧ CLIQUE-CONTAINS(geometric-horizon) ∧ ¬ CLIQUE-CONTAINS(skyline)	BELOW-GEOMETRIC-HORIZON
GROUND	TIME-IS-DAY	DARK

Biederman 1972

- Arrow appeared before or after picture.
- Selected object from 4 pictures.



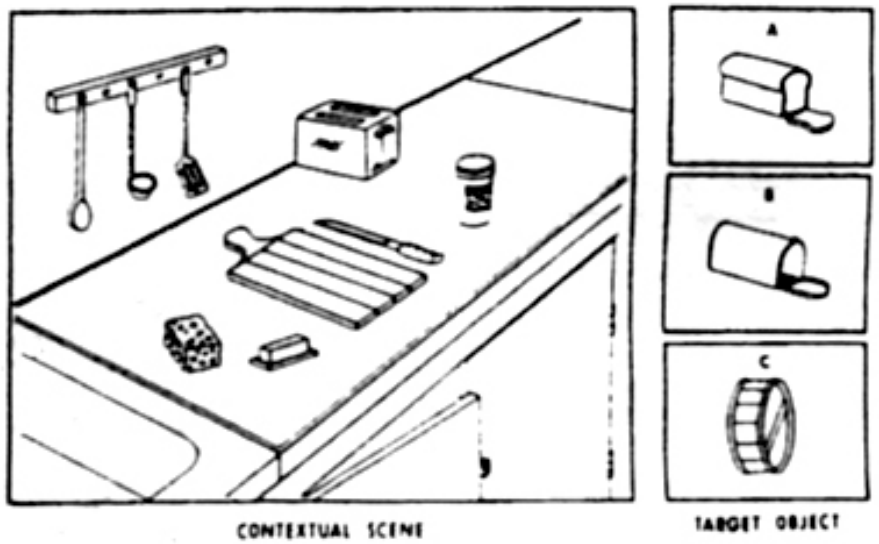

a

Biederman 1972

- Better accuracy with normal scene and with pre-cue.
- Coherence of surroundings affected object perception.
- But, jumbled pictures had unnatural edge artifacts.

Palmer 1975

- Scene preceded object to identify.
- Better identification when preceded by a semantically consistent scene.



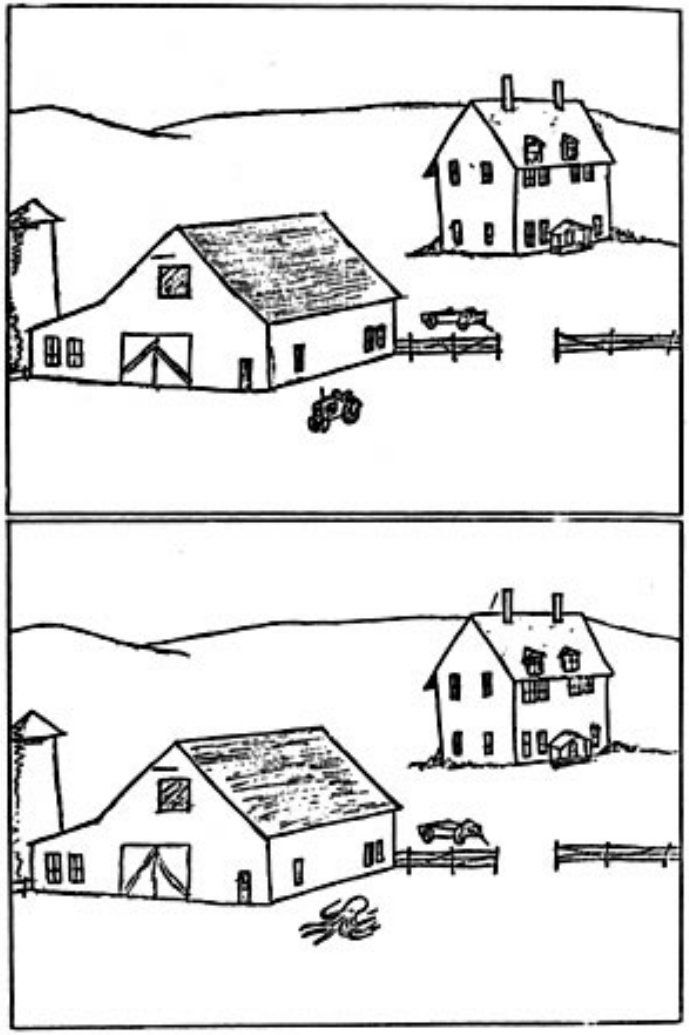
Objects seen for 20, 40, 60 or 120 ms.

Palmer

- Scenes shown ahead of time for 2 s.
- More accurate recognition of consistent objects than inconsistent objects.
- Similar looking objects were misnamed, showing a bias effect.

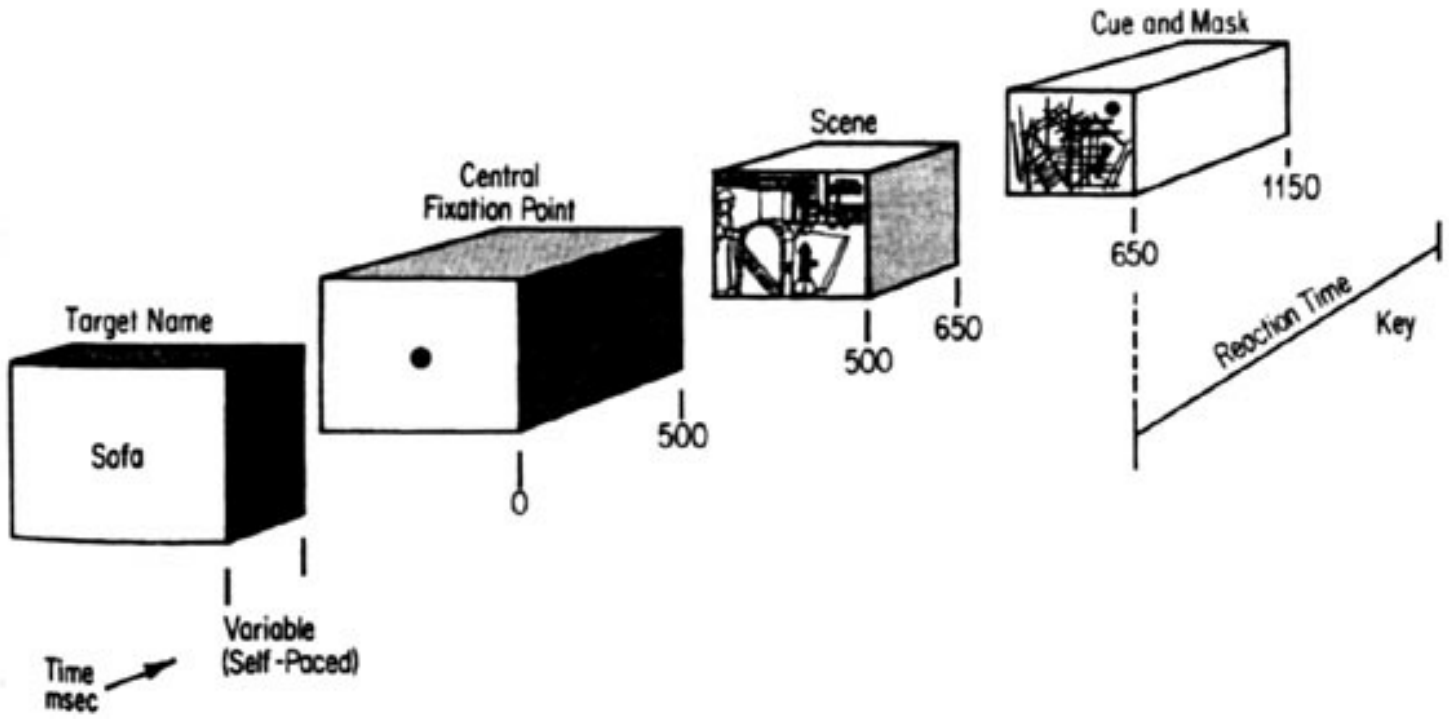
Loftus & Mackworth

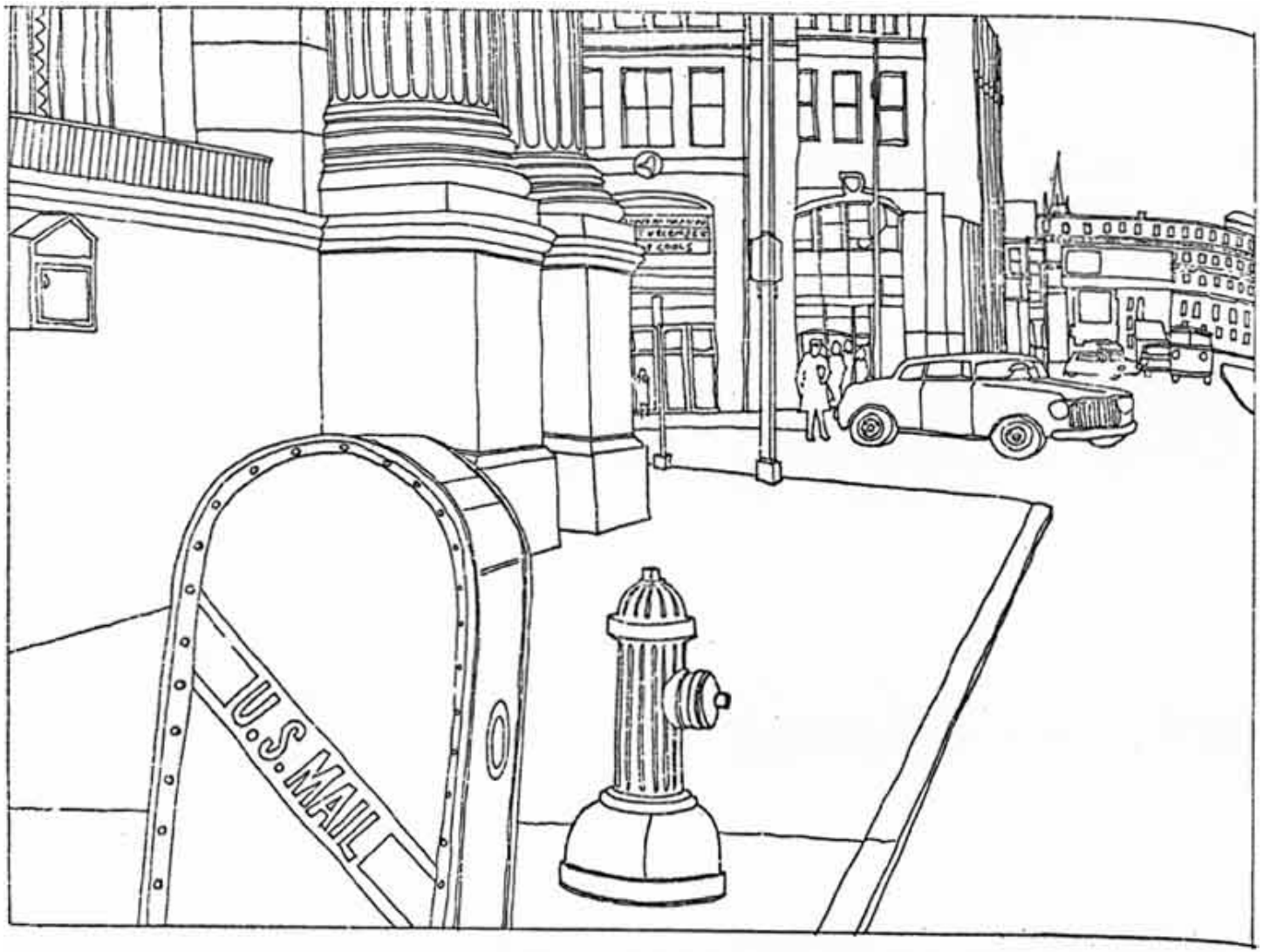
- Inconsistent objects fixated earlier and longer.
- Suggested additional processing of objects out of context.
- Similar results found by Friedman (1979).



Object Detection

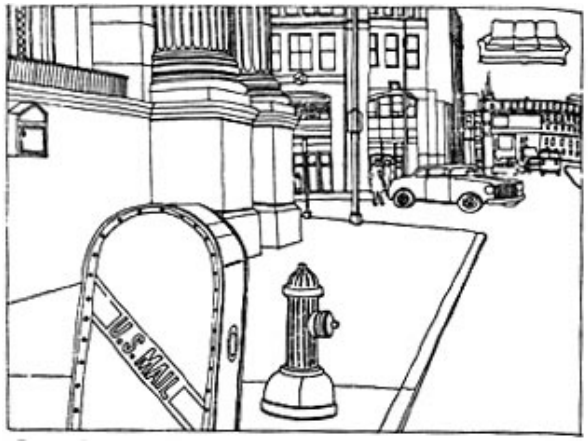
- Biederman et al. 1982, relational violations




U.S. MAIL

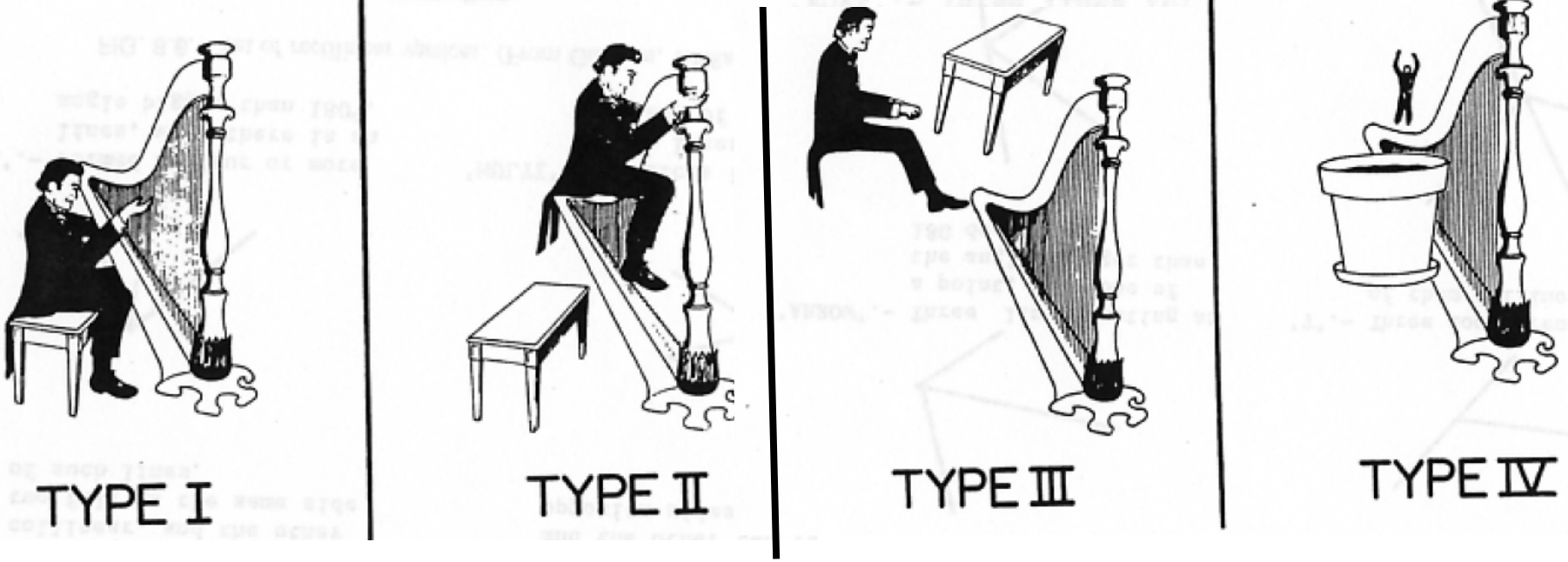
Biederman 1982

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



Objects and Scenes

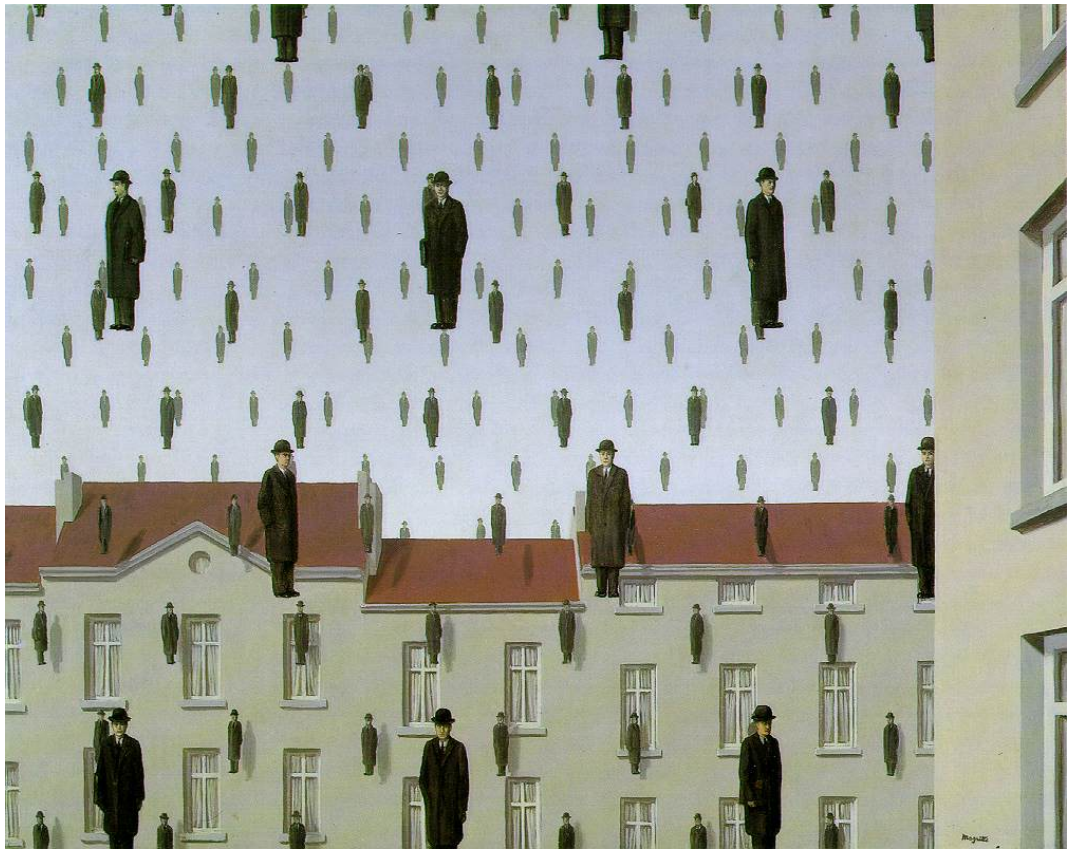
Stimuli from Hock, Romanski, Galie, and Williams (1978).



Biederman's violations (1981):

1. *Support* (e.g., a floating fire hydrant). The object does not appear to be resting on a surface.
2. *Interposition* (e.g., the background appearing through the hydrant). The objects undergoing this violation appear to be transparent or passing through another object.
3. *Probability* (e.g., the hydrant in a kitchen). The object is unlikely to appear in the scene.
4. *Position* (e.g., the fire hydrant on top of a mailbox in a street scene). The object is likely to occur in that scene, but it is unlikely to be in that particular position.
5. *Size* (e.g., the fire hydrant appearing larger than a building). The object appears to be too large or too small relative to the other objects in the scene.

Support



[*Golconde* Rene Magritte]

Interposition

[*Blank Check* Rene Magritte]

Size

[*The Listening Room* Rene Magritte]

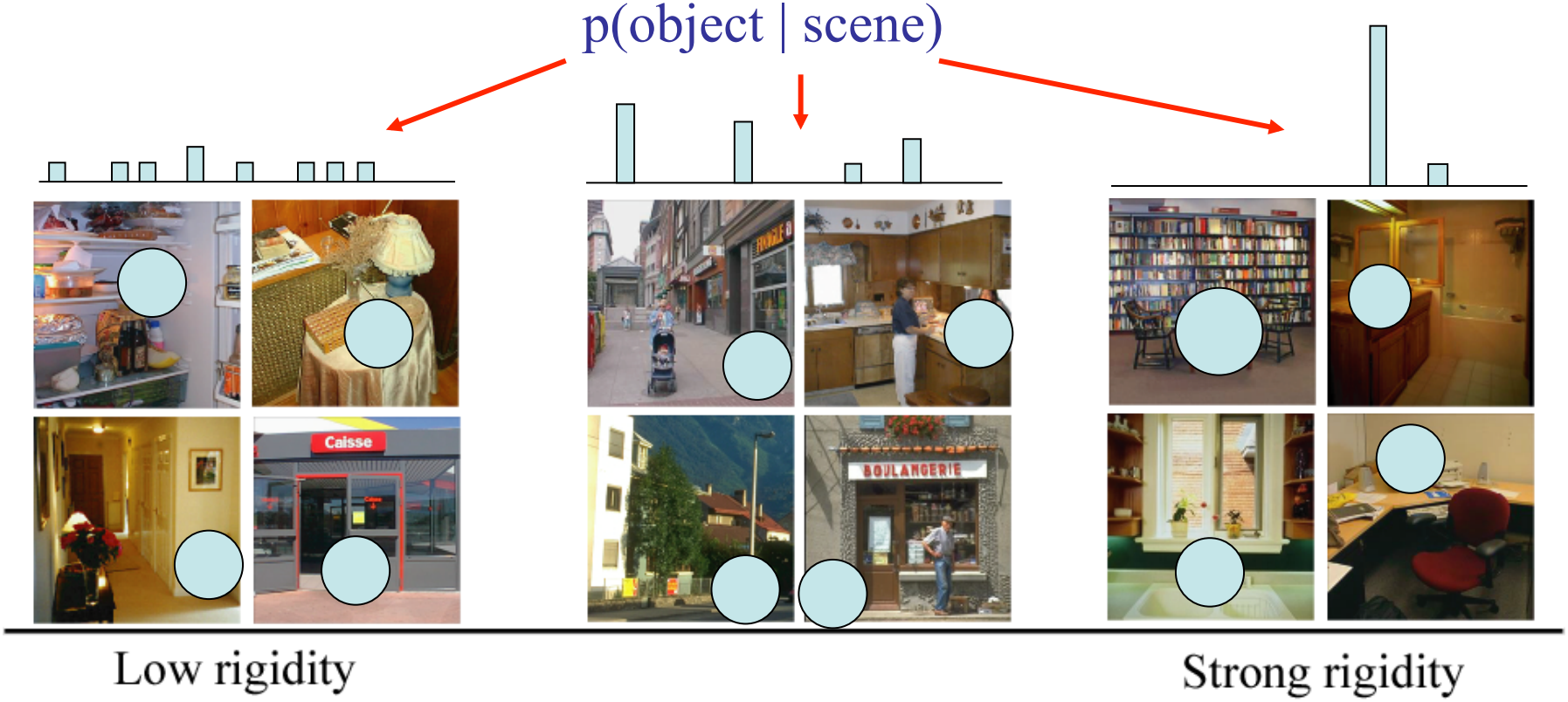
Position, Probability

[*Personal Values* Rene Magritte]

Object priming

→ Car, pedestrian, mailbox, …

p(object | scene)

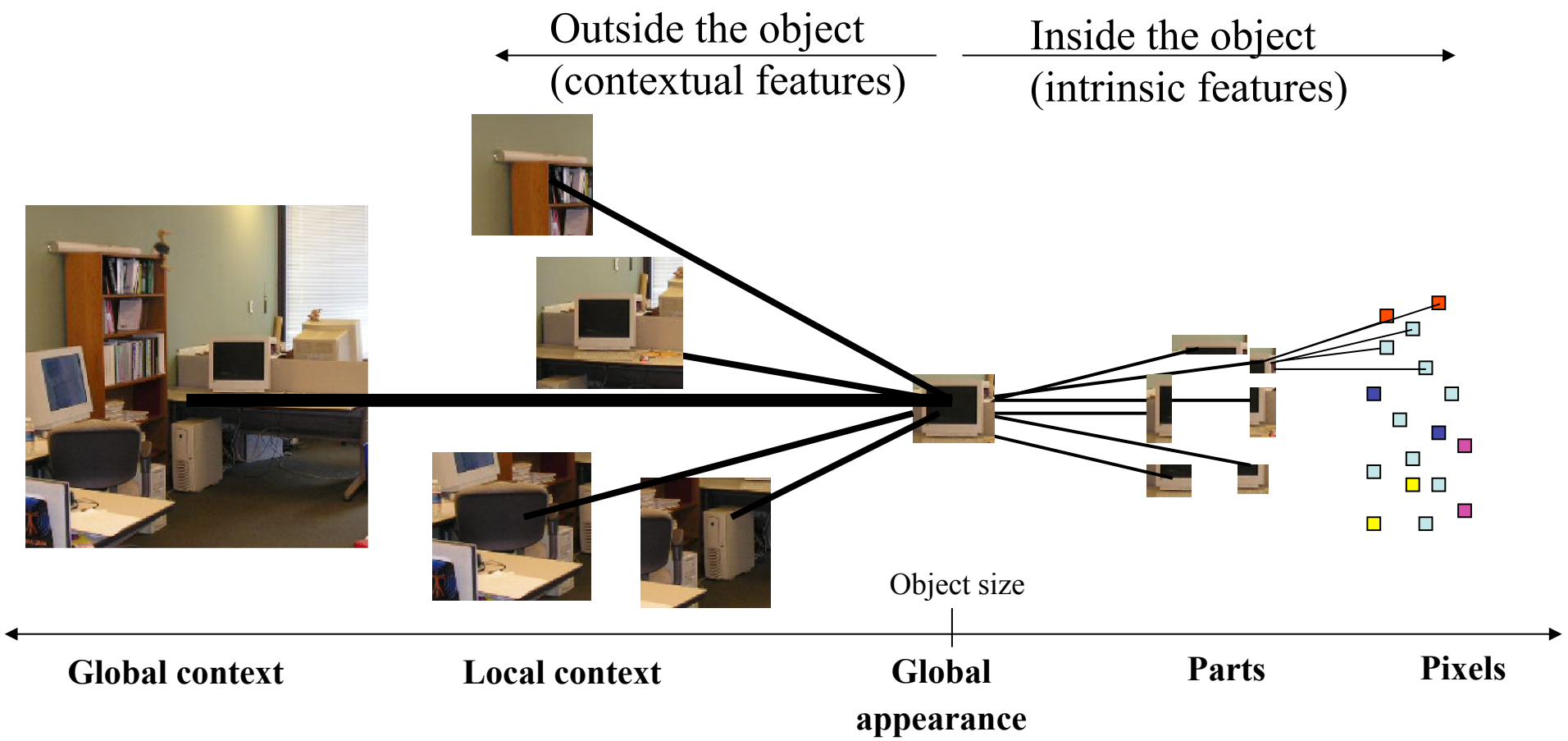


Low rigidity

Strong rigidity

Torralba, Sinha, Oliva, VSS 2001

Looking outside the bounding box



Kruppa & Shiele, (03), Fink & Perona (03)

Carbonetto, Freitas, Barnard (03), Kumar, Hebert, (03)

He, Zemel, Carreira-Perpinan (04), Moore, Essa, Monson, Hayes (99)

Strat & Fischler (91), Torralba (03), Murphy, Torralba & Freeman (03)

Agarwal & Roth, (02), Moghaddam, Pentland (97), Turk, Pentland (91),Vidal-Naquet, Ullman, (03)

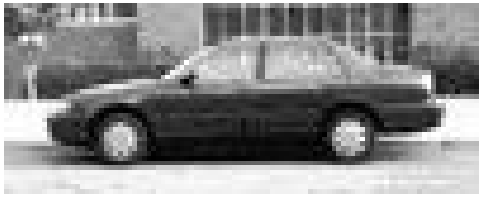
Heisele, et al, (01), Agarwal & Roth, (02), Kremp, Geman, Amit (02), Dorko, Schmid, (03)
Fergus, Perona, Zisserman (03), Fei Fei, Fergus, Perona, (03), Schneiderman, Kanade (00), Lowe (99)
Etc.

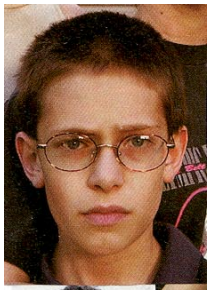
Current approaches

1) **Scene to object dependencies**

2) Object to object dependencies

Many object types co-occur…





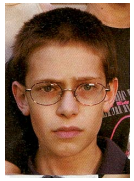
… but this co-occurrence has a hidden common “cause”: the scene

offices









streets

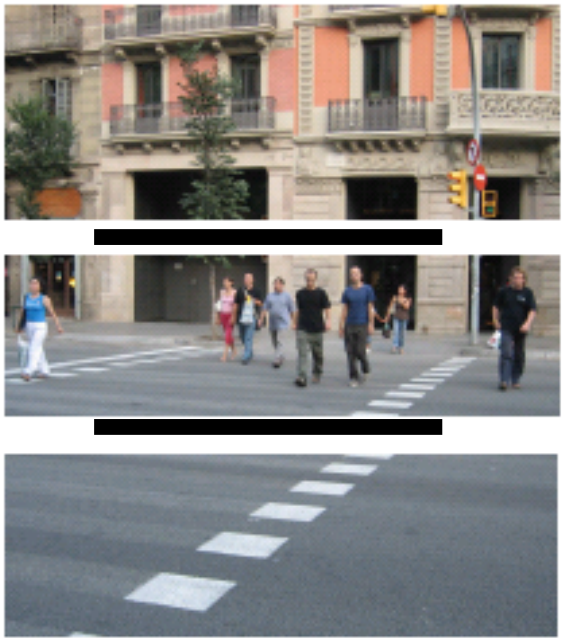





It is easier to first recognize the scene, then predict object presence, than running local object classifiers

The layered structure of scenes

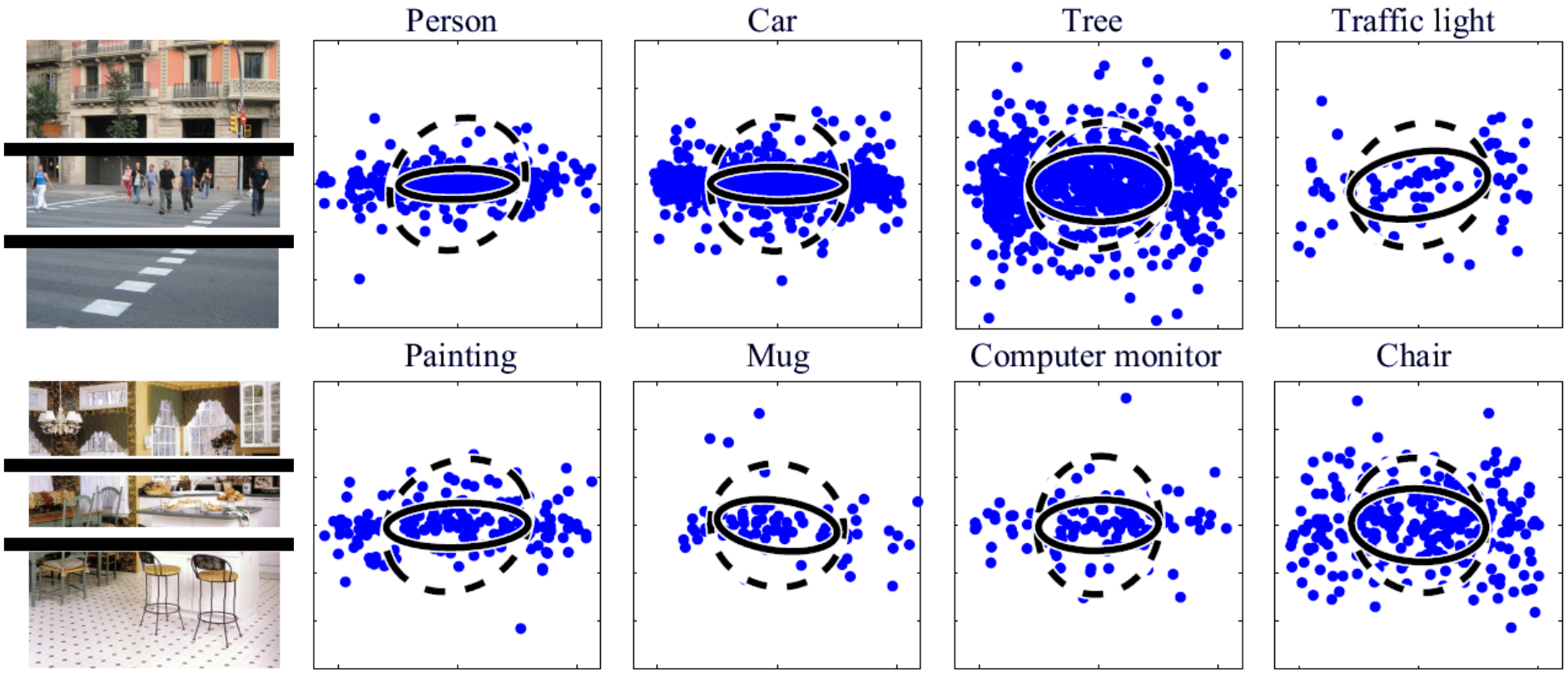
Assuming a human observer standing on the ground



In a display with multiple targets present, the location of one target constraints the ‘y’ coordinate of the remaining targets, but not the ‘x’ coordinate.

The layered structure of scenes

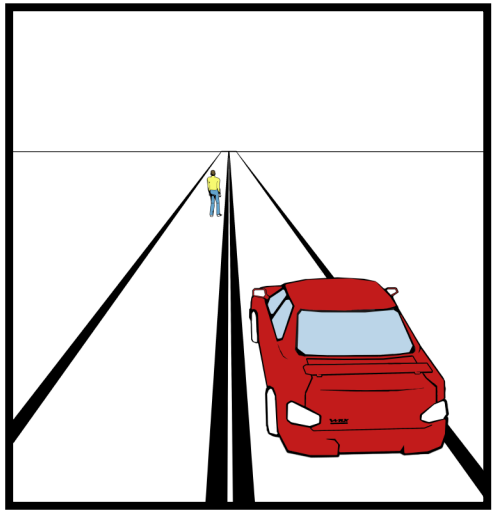
Assuming a human observer standing on the ground



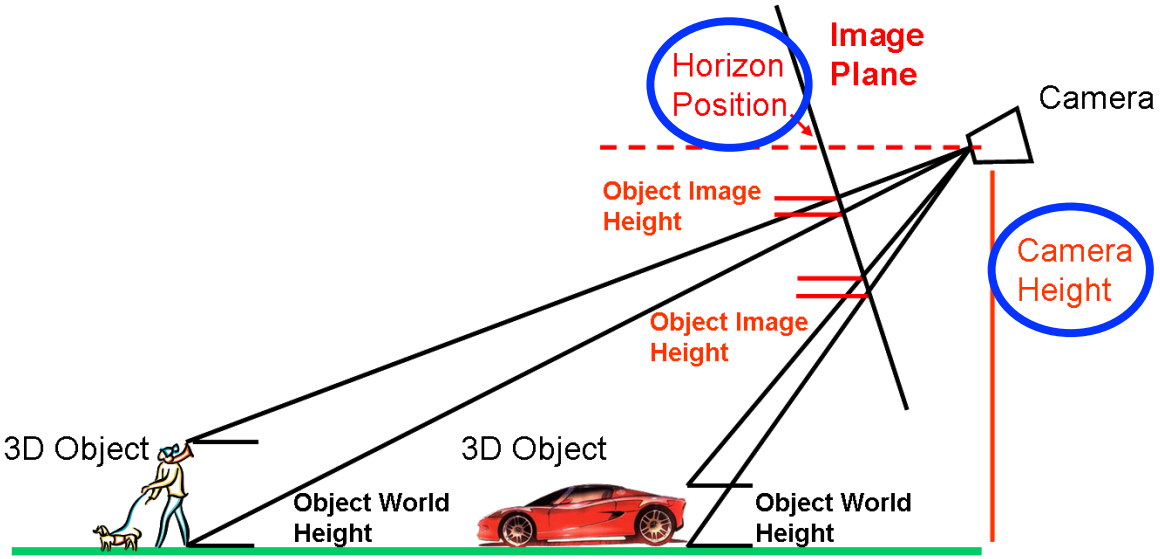
In a display with multiple targets present, the location of one target constraints the ‘y’ coordinate of the remaining targets, but not the ‘x’ coordinate.

Torralba, Oliva, Castelhano, Henderson. (2006).

3d Scene Context



Image



World

Hoiem, Efros, Hebert ICCV 2005

Current approaches

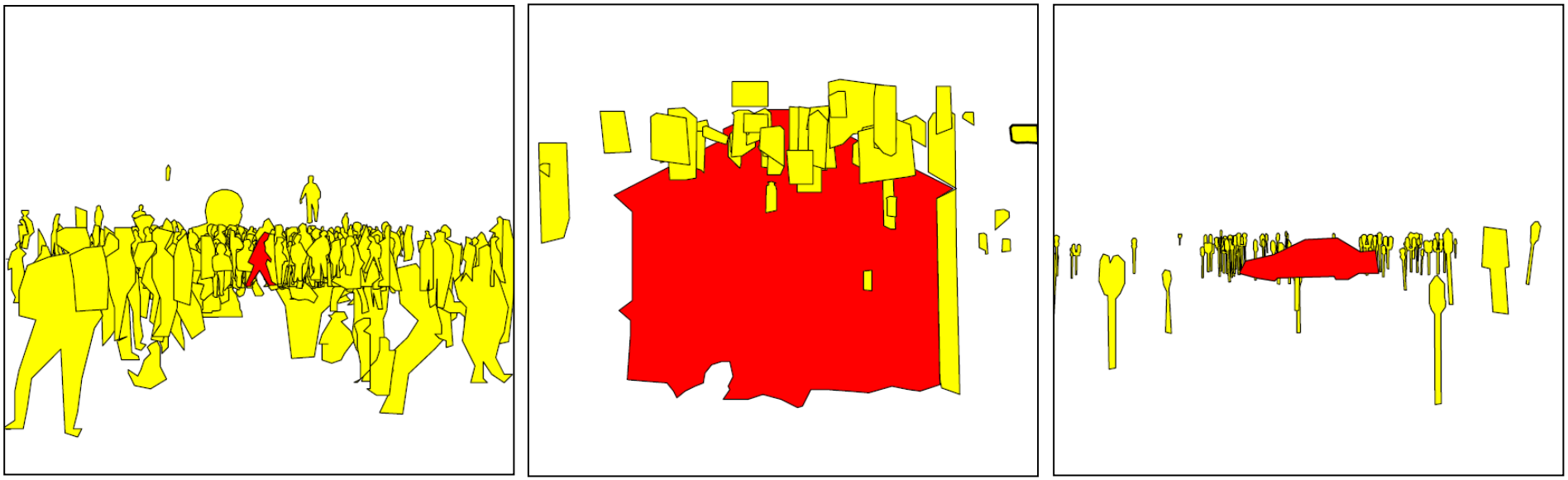
1) Scene to object dependencies

2) Object to object dependencies

Where should I put the silverware?

Sampling from the labels

Sampling from the labels



Cf. Hoiem et al; Hays, Efros. Siggraph 2007

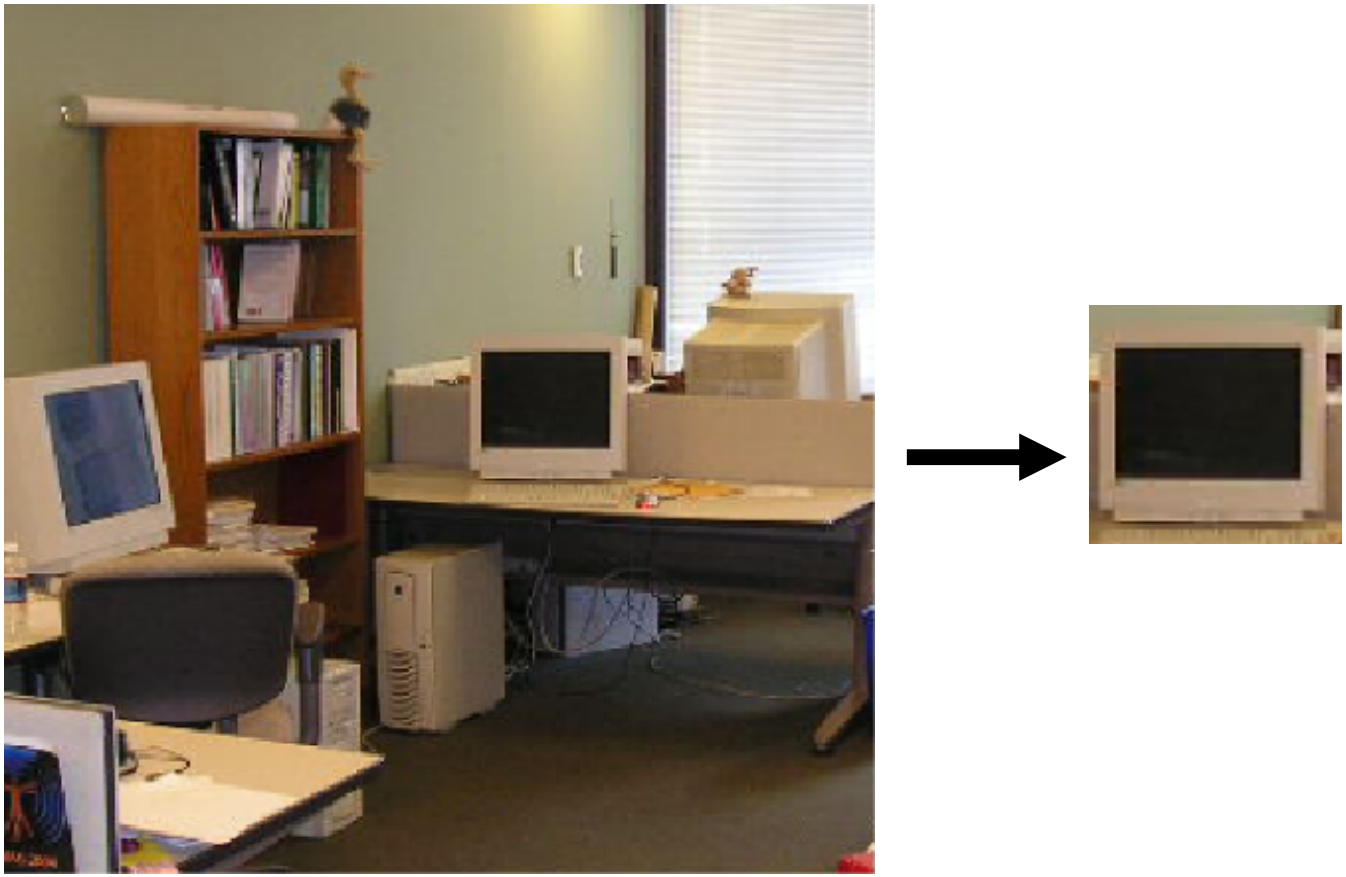
Detecting difficult objects

→ Office → Maybe there is a mouse

Start recognizing the scene

Torralba, Murphy, Freeman. NIPS 2004.

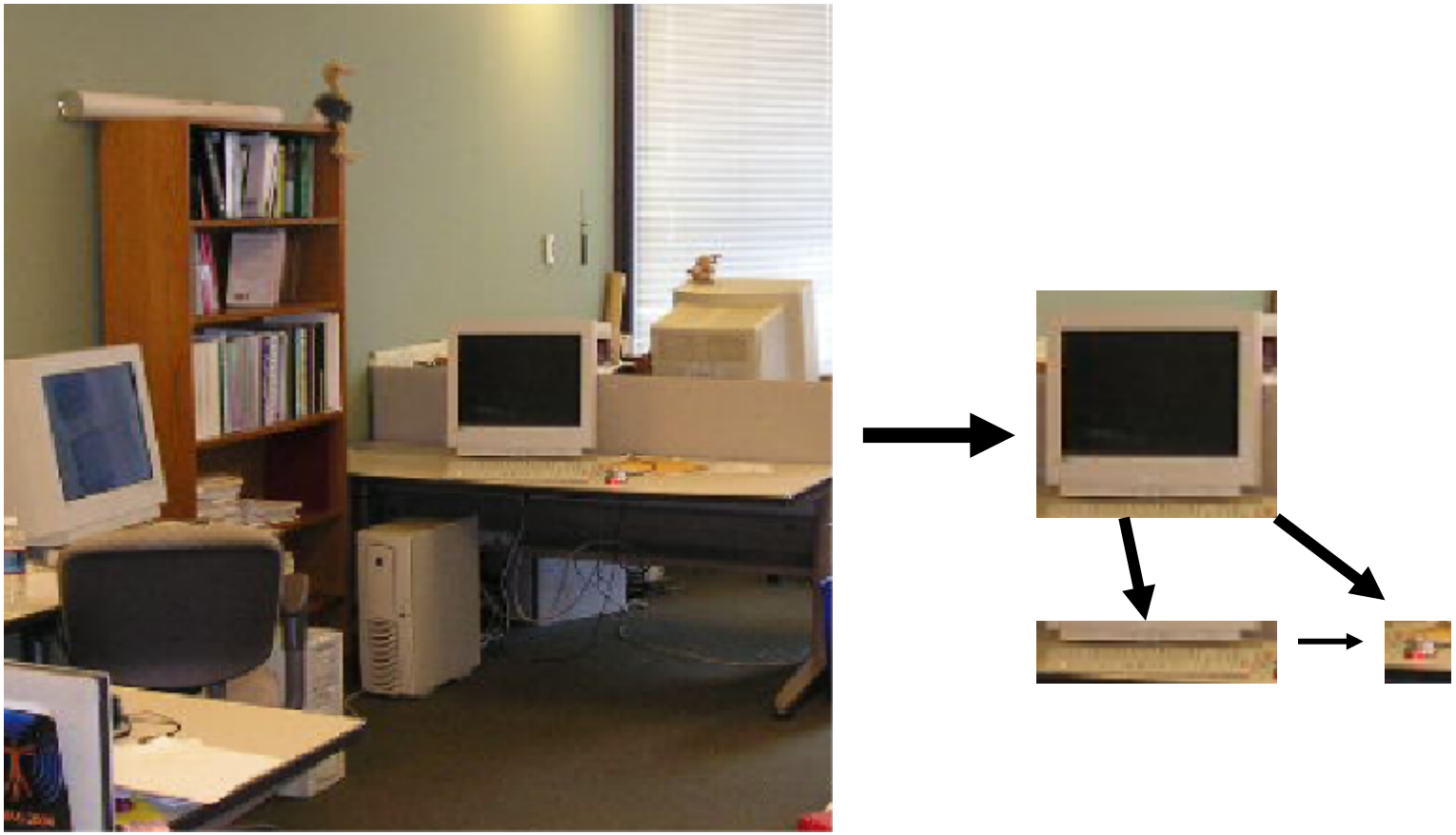
Detecting difficult objects



Detect first simple objects (reliable detectors) that provide **strong contextual constraints to the target** (screen -> keyboard -> mouse)

Torralba, Murphy, Freeman. NIPS 2004.

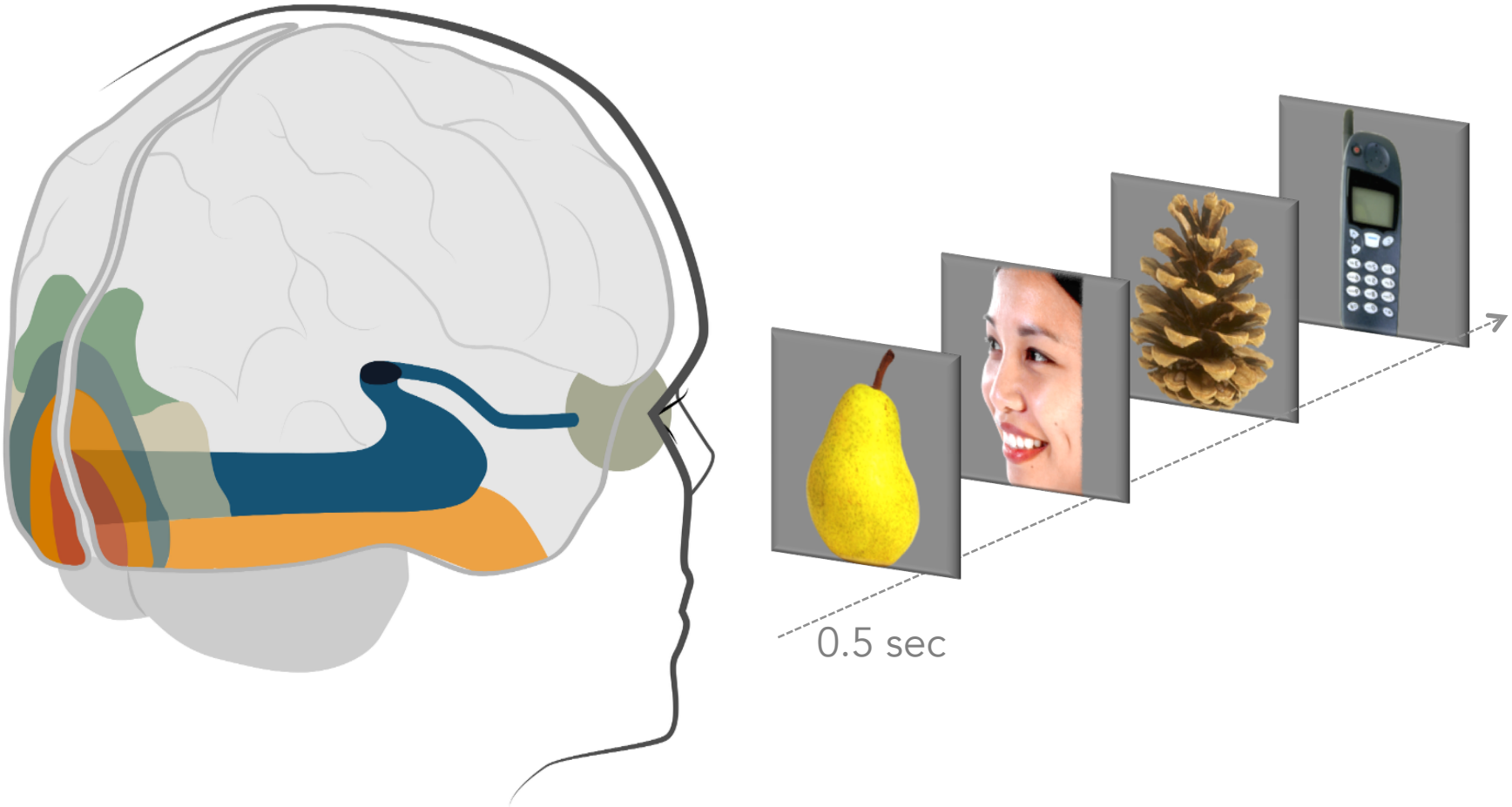
Detecting difficult objects



Detect first simple objects (reliable detectors) that provide strong contextual constraints to the target (screen -> keyboard -> mouse)

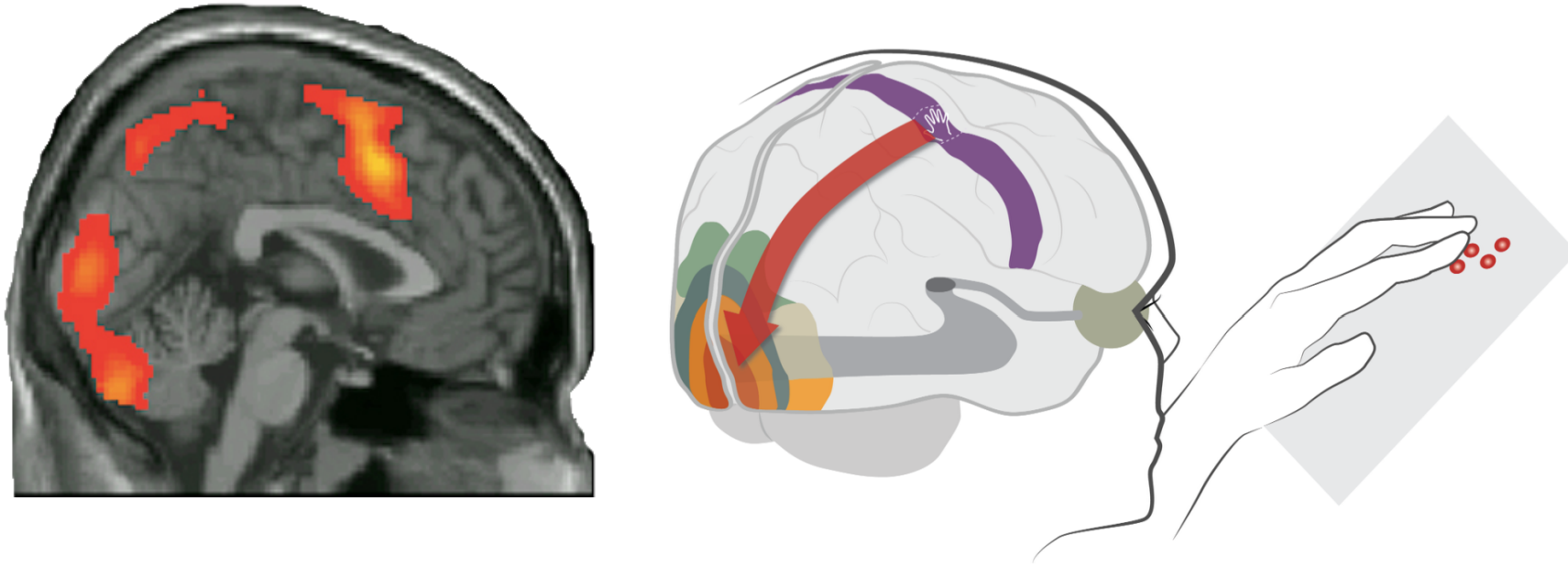
Torralba, Murphy, Freeman. NIPS 2004.

High-Powered Machine: Principles



I. Plasticity

Nothing is lost, everything is transformed



Feeling touch with the “visual” brain

Teng, Cichy, Pantazis, Oliva

II. Growth

Immediate

Short-term

Long-term