HACETTEPE UNIVERSITY COMPUTER VISION LAB

HACETTEPE UNIVERSITY

http://vision.cs.hacettepe.edu.tr

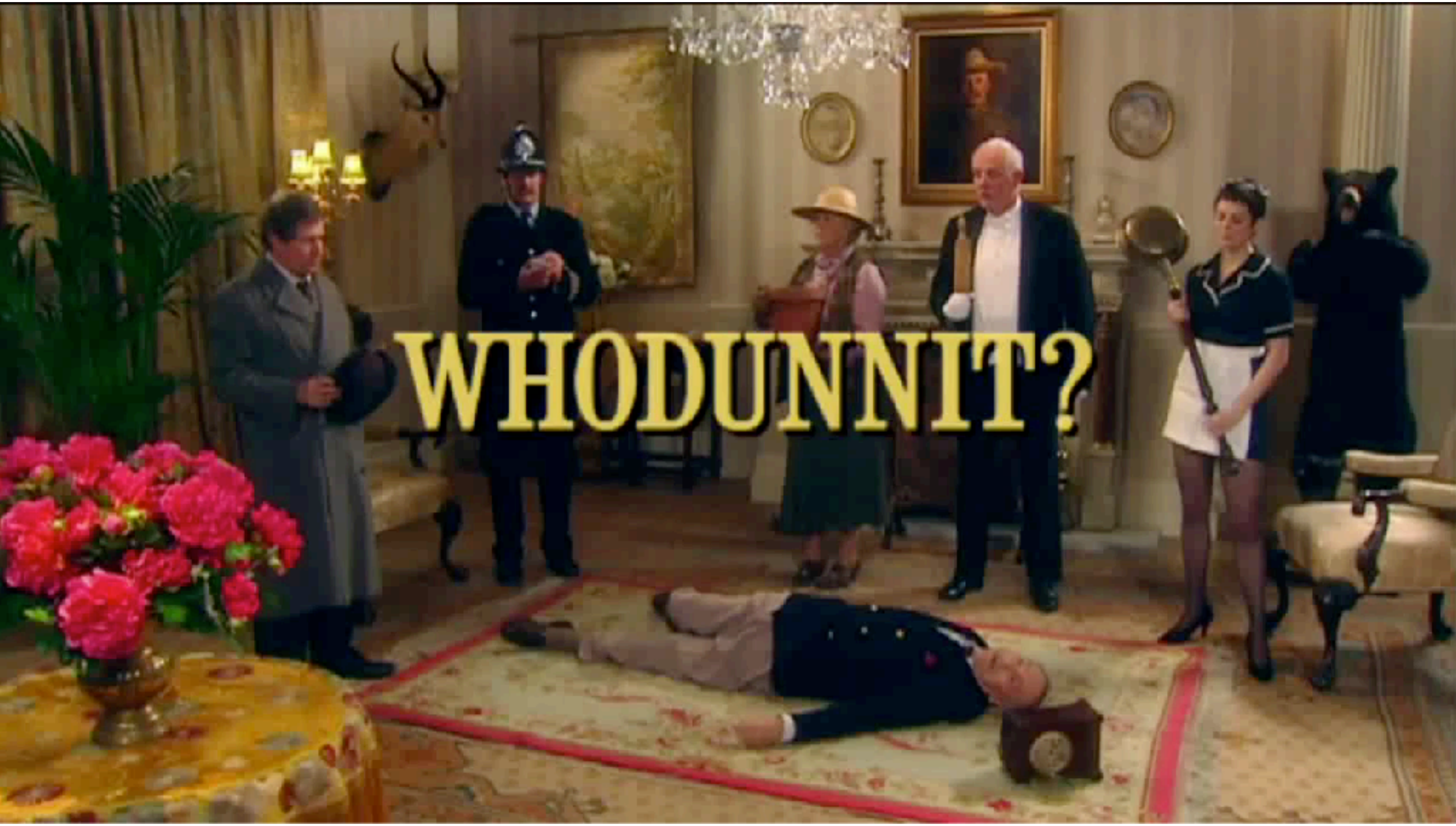
Visual saliency

Erkut Erdem
Hacettepe University
Computer Vision Laboratory

Where do we look on these

The squares shows where 15 observers looked
in eye tracking experiments

What is attention?



Why do perceptual systems need

- Limited resources
  - ➡ Our visual system processes an enormous amount of data coming from the retina. ~10^8 bits/sec [Itti, 2000]
- Warning
  - ➡ noticing predators, sudden motion, etc.
- Exploration
  - ➡ finding preys, locating objects, etc.

The amount of information coming down the optic nerve far exceeds what the brain is capable of fully processing and assimilating into conscious experience.


[Anstis, 1998]

Attentional mechanisms

- Attention is a complex set of interrelated processes:
  - ➡ selection of information (bottom-up)
  - ➡ integration of that information with existing knowledge (top-down)
- Bottom-up
  - ➡ very rapid, primitive, task-independent
- Top-down
  - ➡ slower, under cognitive control, task-dependent

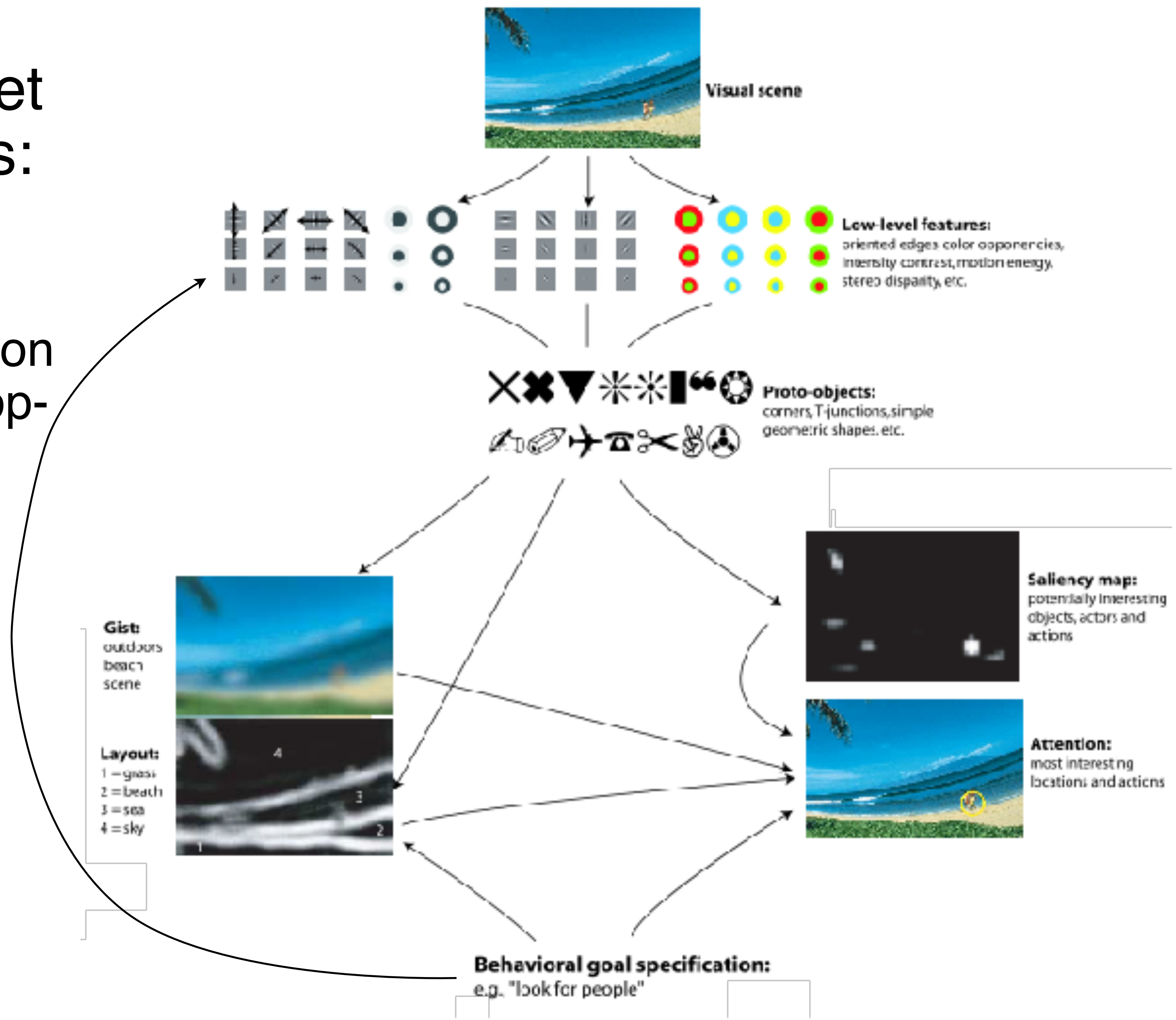
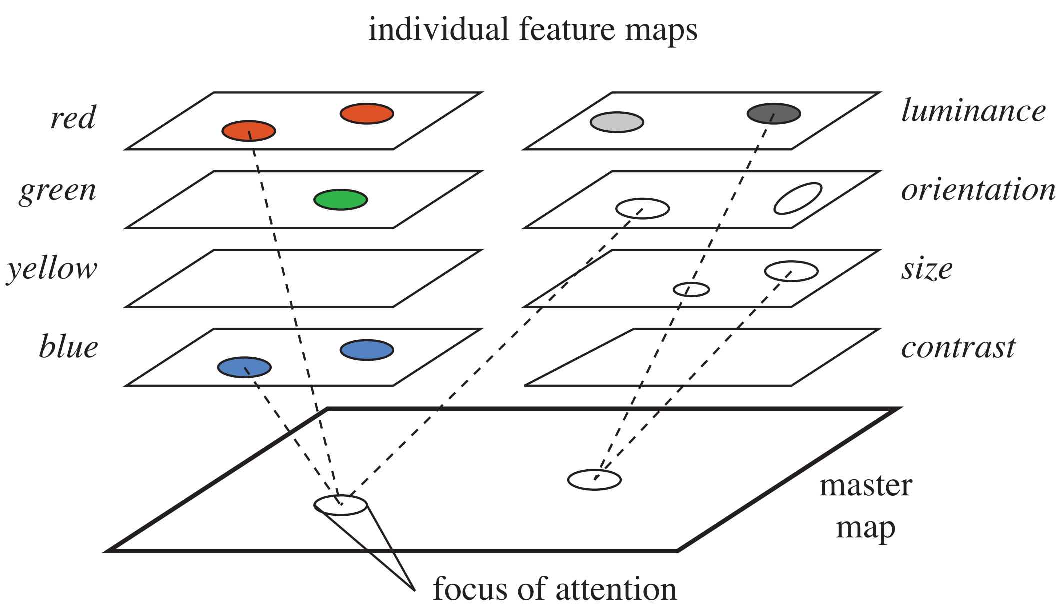
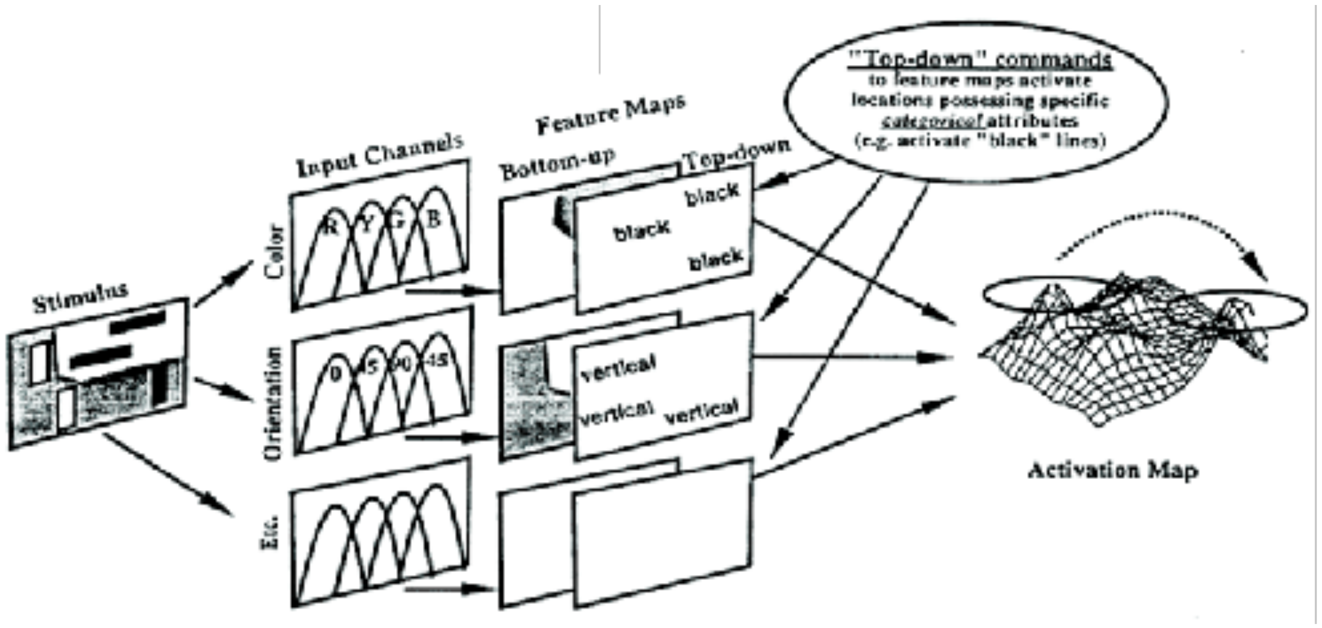


Image credit: [Itti, 2002]

Theories of visual attention



Feature-Integration Theory
[Treisman & Gelade, 1980]



Guided Search Theory
[Wolfe, 1989]

- processing occurs in parallel and focused attention occurs in serial

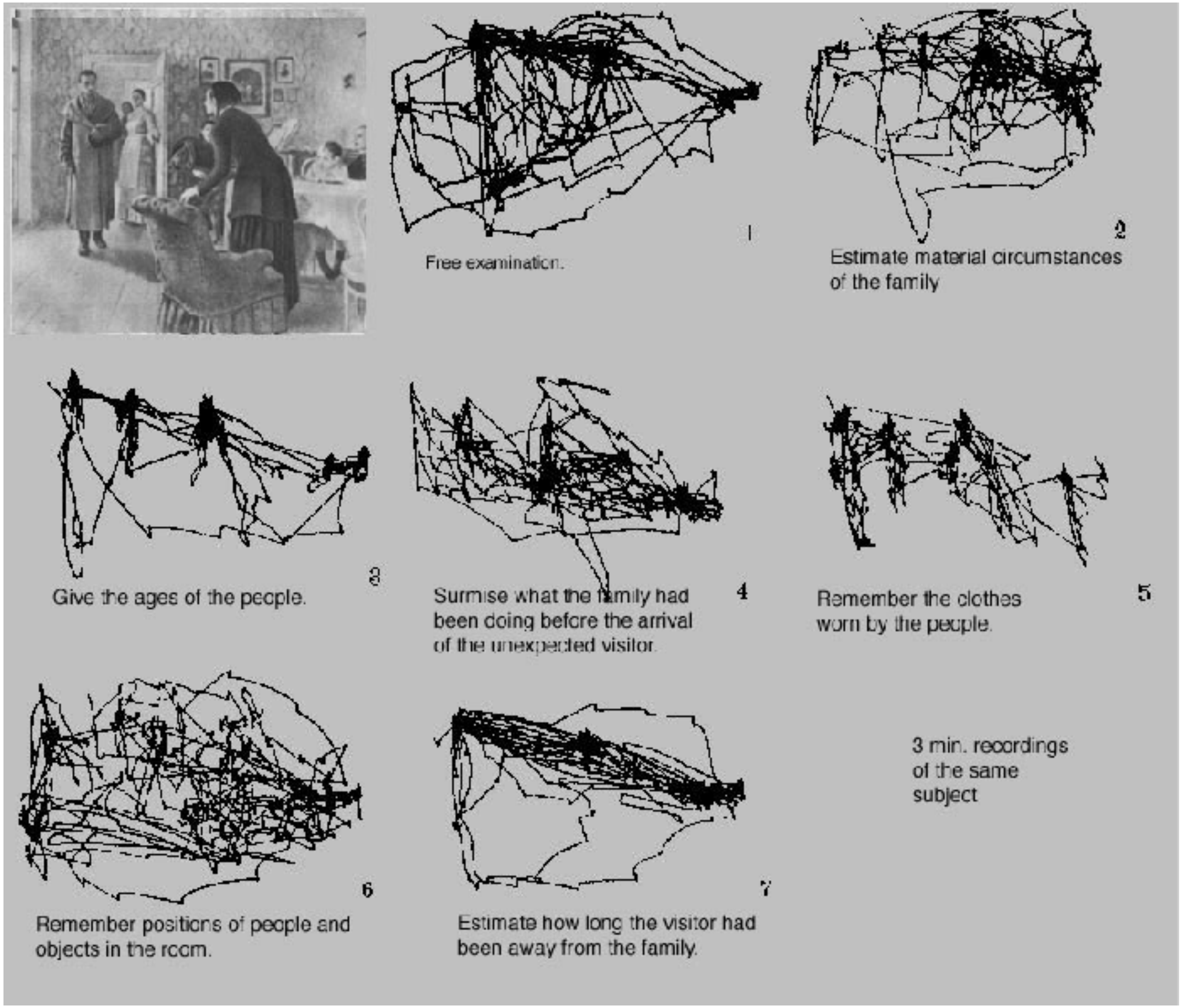
- visual search relies on a combination of bottom-up and top-down activity

Task-based visual attention

"They did not expect him" by Repin

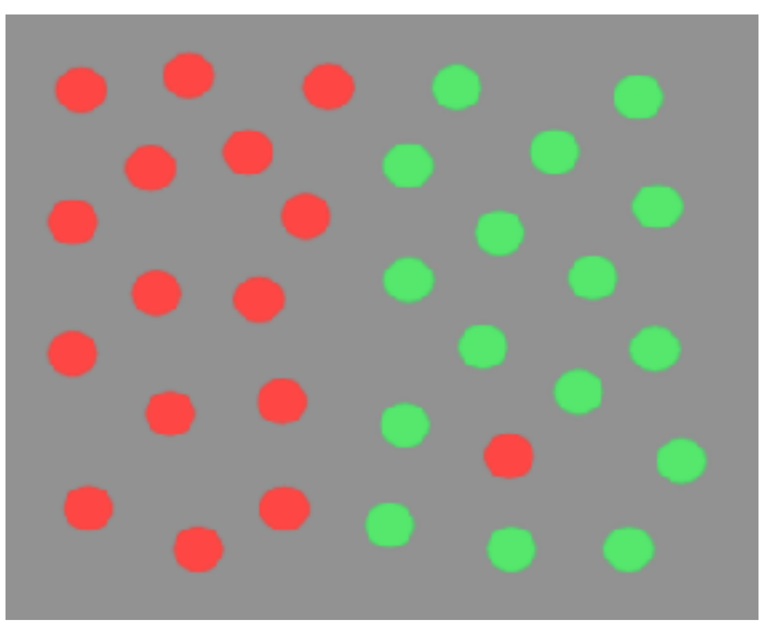
- Yarbus (1967) was the first to show that task influences eye fixation locations.

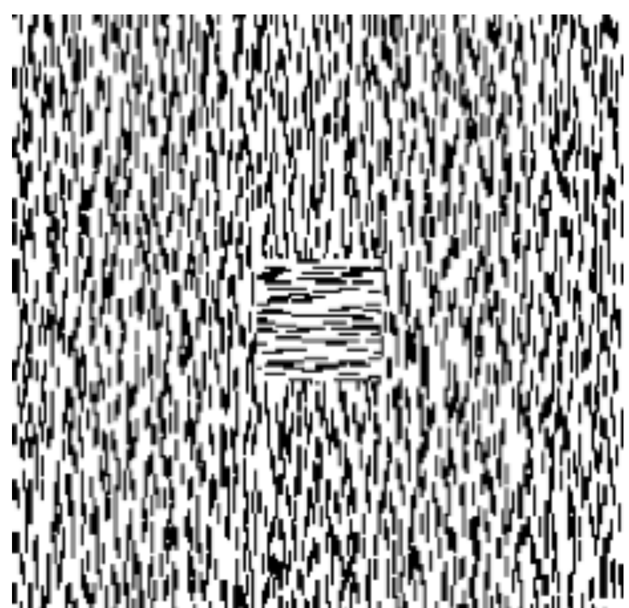
Task-based visual attention



Visual saliency

- “Saliency at a given location is determined primarily by how **different** this location is **from its surround** in color, orientation, motion, depth, etc.” [Koch & Ullman, 1985]
- “Visual salience (or visual saliency) is the **distinct** subjective perceptual quality which makes some items in the world stand out **from their neighbors** and immediately grab our attention.” [Itti, 2007]



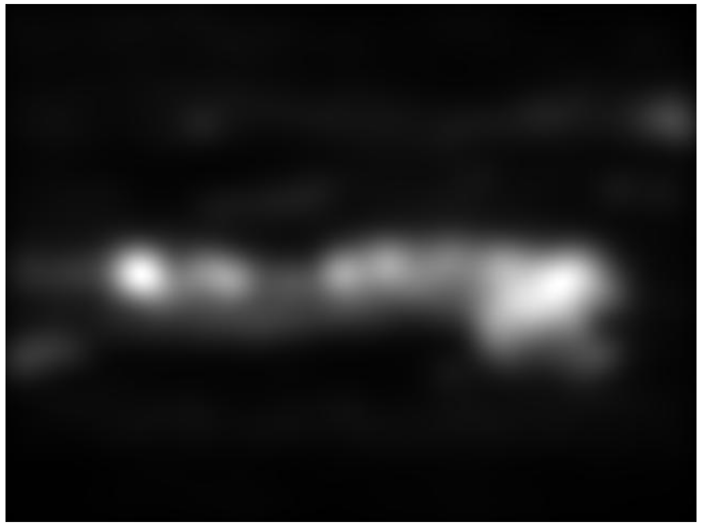


Beyond biology: Applications in Computer

- Most computer vision algorithms have relied on brute-force (e.g. sliding window) strategies.
- Attentional mechanisms provide a relatively free and fast mechanism to select a few candidates while eliminating background clutter.
- To list a few of possible applications
  - ➡ scene classification [Siagian & Itti, 2007]
  - ➡ object recognition [Gao et al., 2009; Rutishauser et al., 2004]
  - ➡ object tracking [Butko et al., 2008]
  - ➡ robotics [Frintrop et al., 2006; Siagian & Itti, 2007]
  - ➡ content-based image resizing [Achanta & Susstrunk, 2009; Avidan & Shamir, 2007]

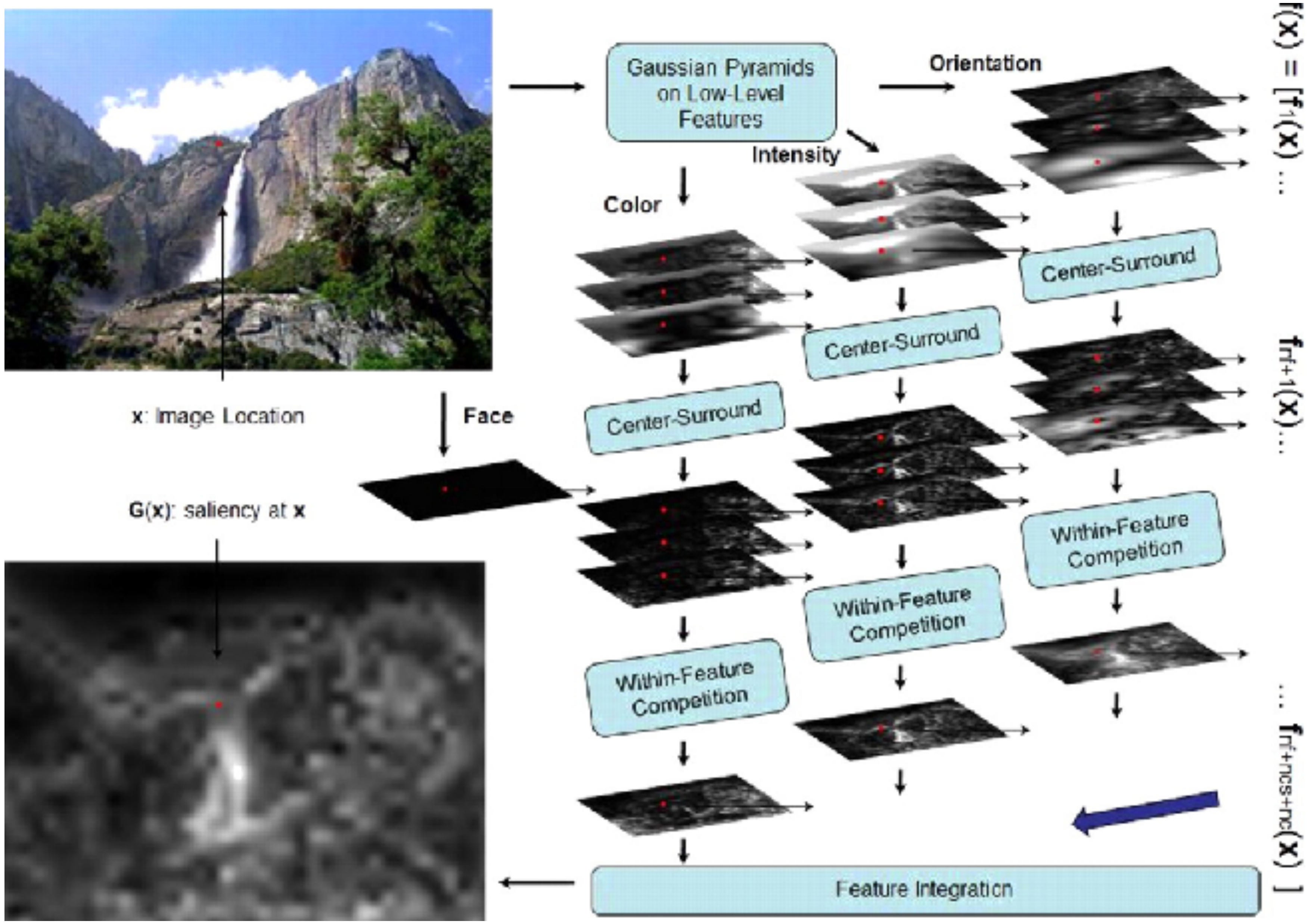
Computational models of visual

- Can machines predict where humans look at a given image?



- [Itti & Koch, 1998]
  - ➡ One of the first computational models of visual attention to predict where people look
  - ➡ A bottom-up model
  - ➡ An implementation of Koch & Ullman, 1985
  - ➡ It employs a multi-scale center-surround mechanism which imitates the workings of the retinal receptive field.

Bottom-up models of visual saliency



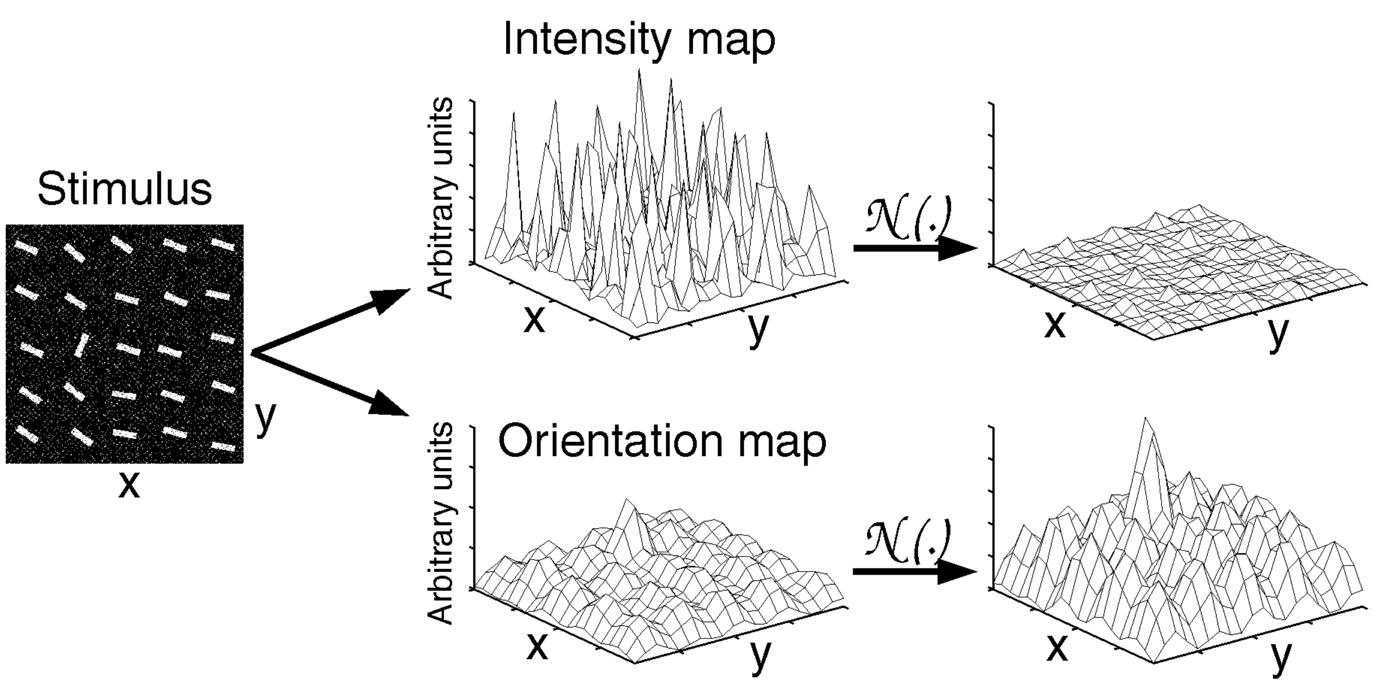
The common basic structure is:

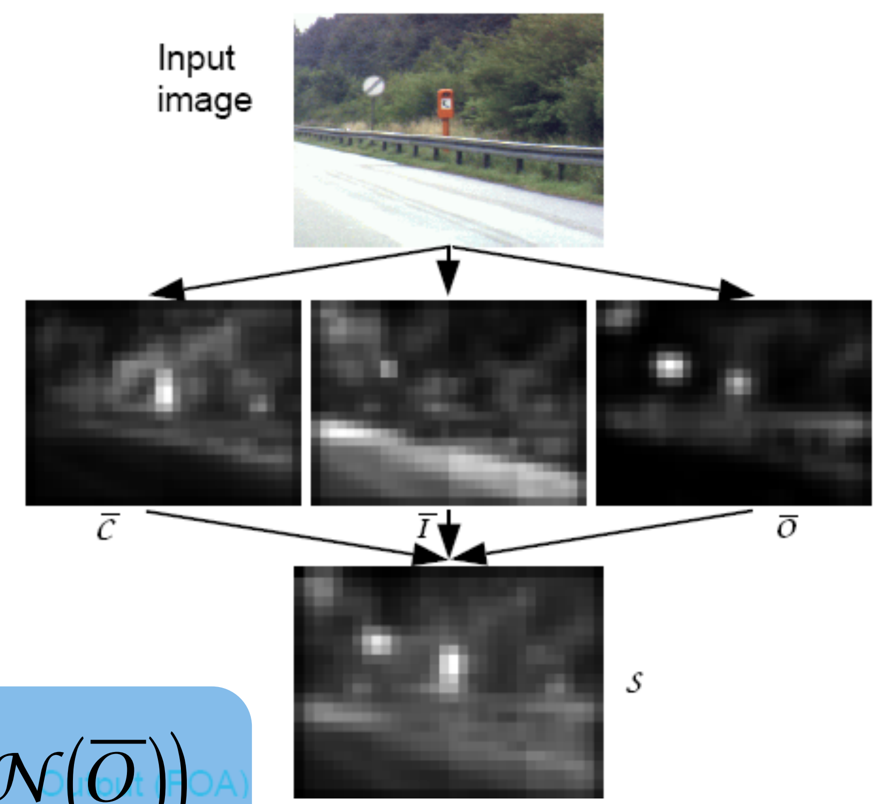
(i) Extract visual features,
(ii) Compute a saliency map for each feature channel
(iii) Compute a final saliency map by combining individual saliency maps

Image credit: [Zhao & Koch, 2012]

Feature integration step

- The most troublesome step
  - ➡ typically carried out by taking weighted average (linear summation).
  - ➡ But how different feature dimensions contribute to the overall saliency is still an open question! [Callaghan, 1989, 1990; Eckstein et al., 2000; Rosenholtz, 1999, 2001; Rosenholtz et al., 2004]



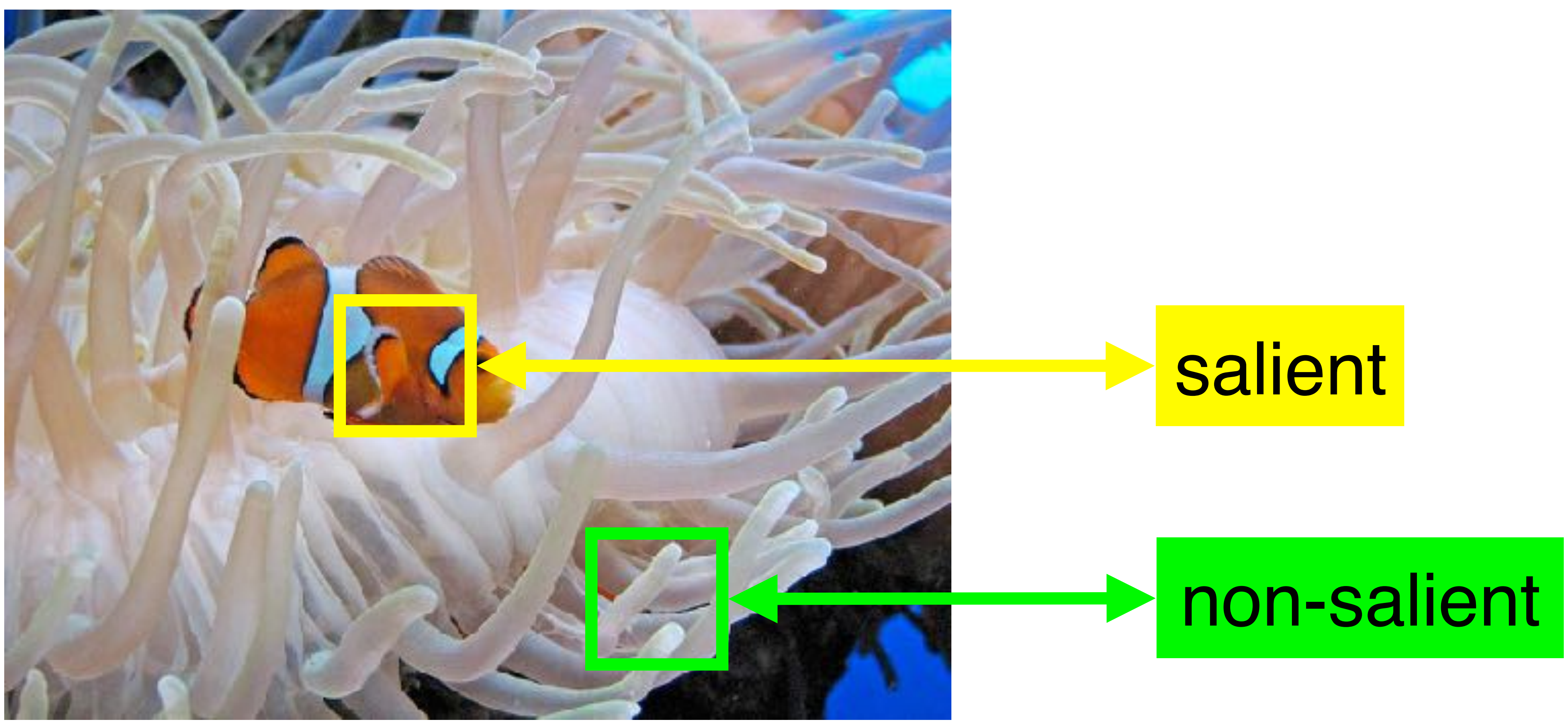


$$S = \frac{1}{3}\left(\mathcal{N}\left(\overline{I}\right) + \mathcal{N}\left(\overline{C}\right) + \mathcal{N}\left(\overline{O}\right)\right)$$

Image credit: [Itti and Koch, 2012]

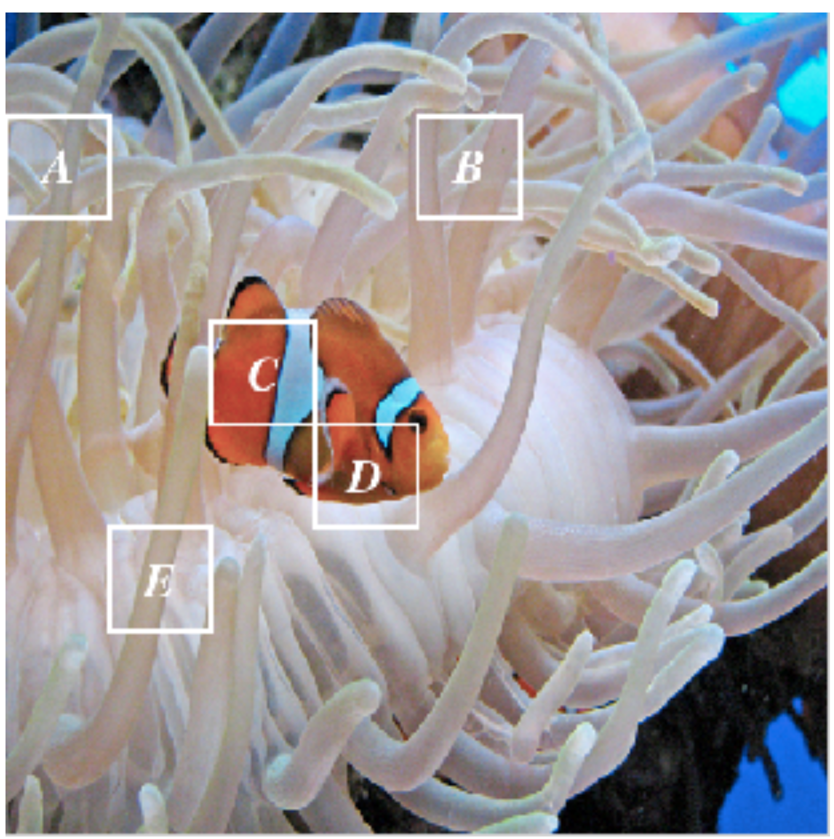
CovSal (Erdem and Erdem, 2013)

- a patch-based formulation
  - ➡ patches with rare appearance characteristics are considered as salient.



CovSal (Erdem and Erdem, 2013)

- The region covariance descriptor [Tuzel et al., 2006]
  - ➡ captures local image structures better than standard linear filters.
  - ➡ naturally provides nonlinear integration of different features by modeling their correlations.

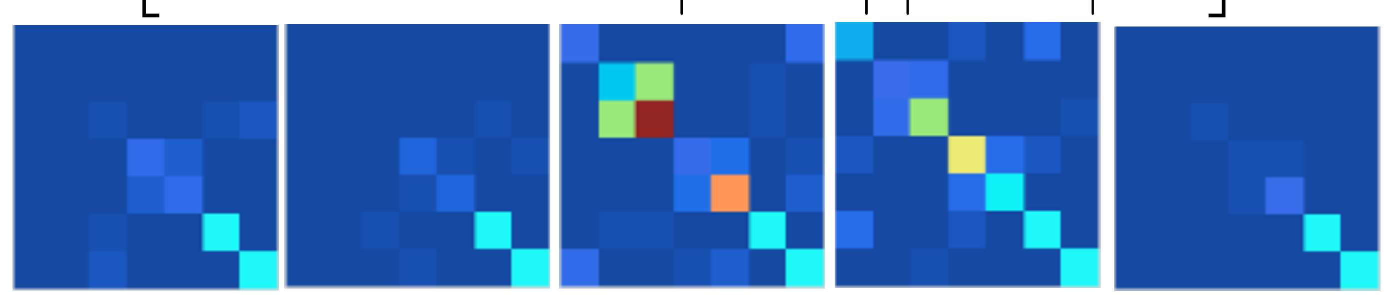


Input image

$$\mathbf{C}_R = \frac{1}{n-1}\sum_{i=1}^{n}(\mathbf{f}_i - \boldsymbol{\mu})(\mathbf{f}_i - \boldsymbol{\mu})^T$$

$\{\mathbf{f}_i\}_{i\,=\,1\ldots n}$: *d*-dimensional feature points inside R

$$\left[L(x,y)\;\; a(x,y)\;\; b(x,y)\;\; \left|\frac{\partial I(x,y)}{\partial x}\right|\;\; \left|\frac{\partial I(x,y)}{\partial y}\right|\;\; x\;\; y\right]^T$$

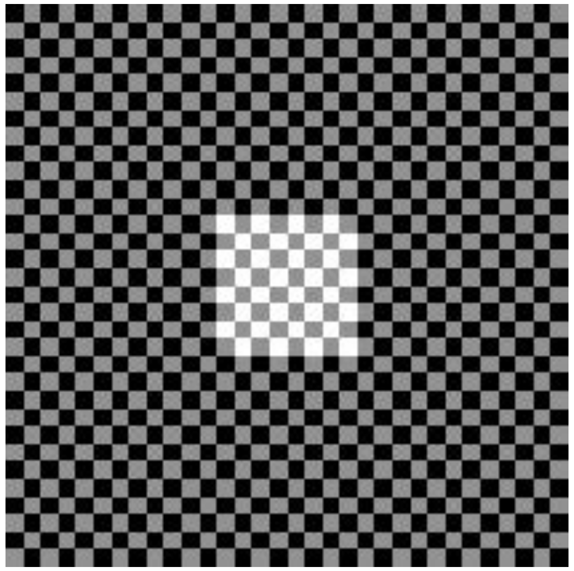


A *B* *C* *D* *E*

Extracted region covariance descriptors

CovSal (Erdem and Erdem, 2013)

- Sometimes covariances may not be enough



Covariances alone can not explain changes in the means!

- We additionally incorporate first-order statistics
  - ➡ Sigmapoints [Hong et al., 2009; Julier & Uhlmann, 1996]

$$\mathbf{s}_i = \begin{cases} \alpha\sqrt{d}\mathbf{L}_i & \text{if } 1 \leq i \leq d \\ -\alpha\sqrt{d}\mathbf{L}_i & \text{if } d+1 \leq i \leq 2d \end{cases} \qquad \mathbf{C} = \mathbf{L}\mathbf{L}^T$$

Cholesky decomposition

  - ➡ Final representation: $\Psi(\mathbf{C}) = (\mu, \mathbf{s}_1, \ldots \mathbf{s}_d, \ , \mathbf{s}_{d+1}, \ldots, \mathbf{s}_{2d})^T$

CovSal (Erdem and Erdem, 2013)

- Visual dissimilarity between two patches R_1 and R_2 can be computed by using the following metrics:

For covariance descriptor:

$$\rho(\mathbf{C}_1, \mathbf{C}_2) = \sqrt{\sum_{i=1}^{n} ln^2 \lambda_i(\mathbf{C}_1, \mathbf{C}_2)}$$

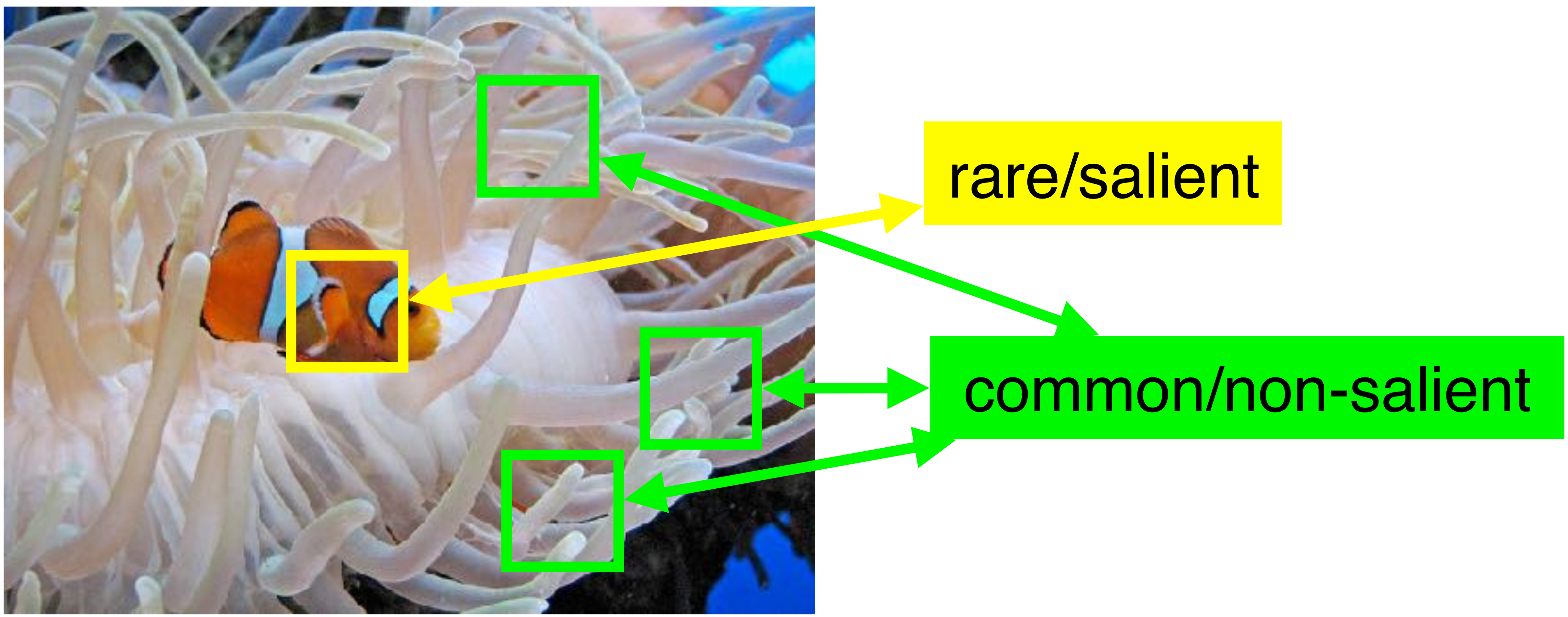
[Föerstner & Moonen, 1999]

For sigma points descriptor:

$$||\Psi(\mathbf{C}_i) - \Psi(\mathbf{C}_j)||$$

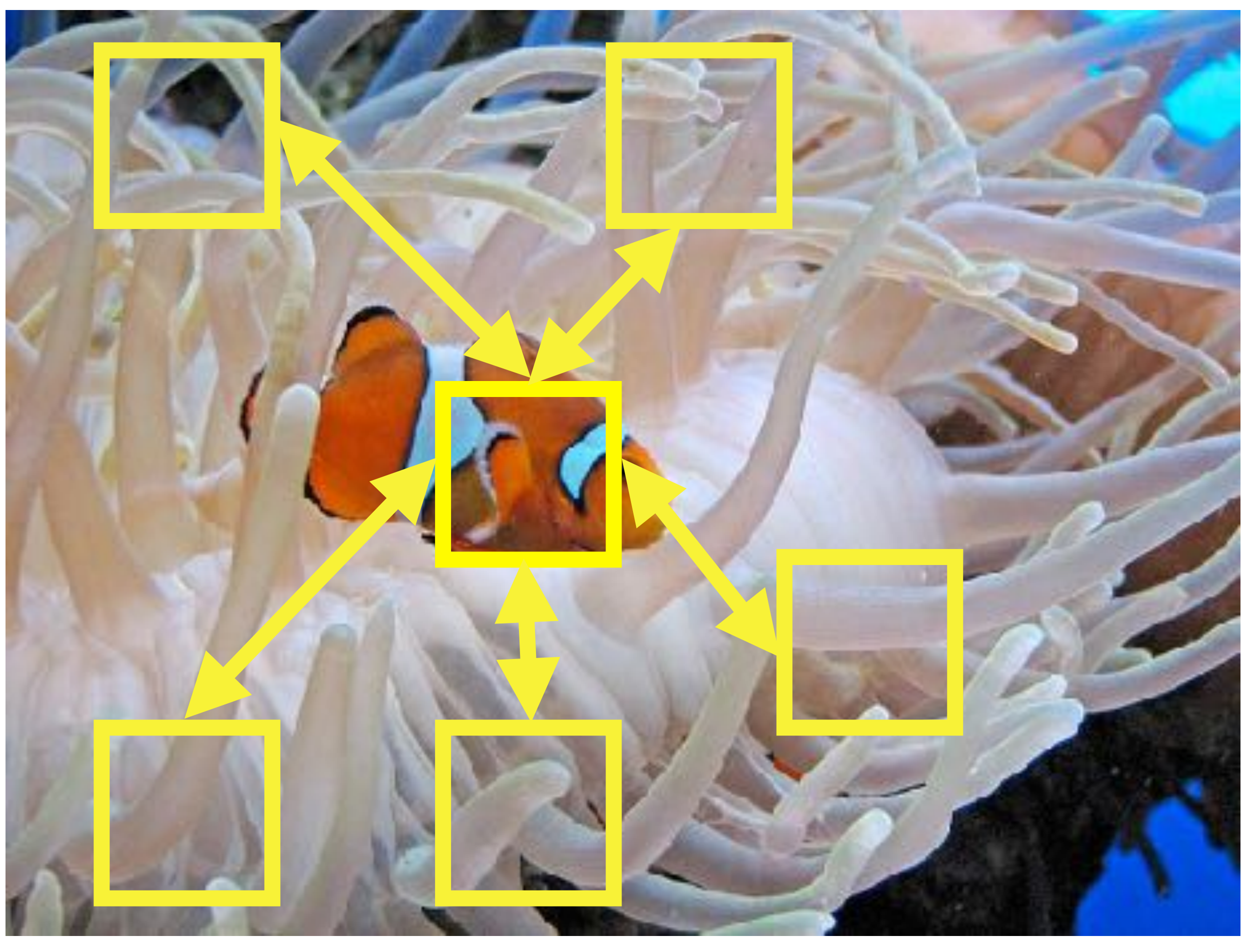
CovSal (Erdem and Erdem, 2013)

- If the patch is highly dissimilar to the patches surrounding it ⟶ rare/salient
- Otherwise ⟶ common/non-salient



CovSal (Erdem and Erdem, 2013)

- The saliency of R_i is defined as the weighted average of the dissimilarities between R_i to the $\underline{m \text{ most similar}}$ regions around it.



$$S(R_i) = \frac{1}{m}\sum_{j=1}^{m} d(R_i, R_j)$$

Model 1

$$d(R_i, R_j) = \frac{\rho(\mathbf{C}_i, \mathbf{C}_j)}{1 + \|\mathbf{x}_i - \mathbf{x}_j\|}$$

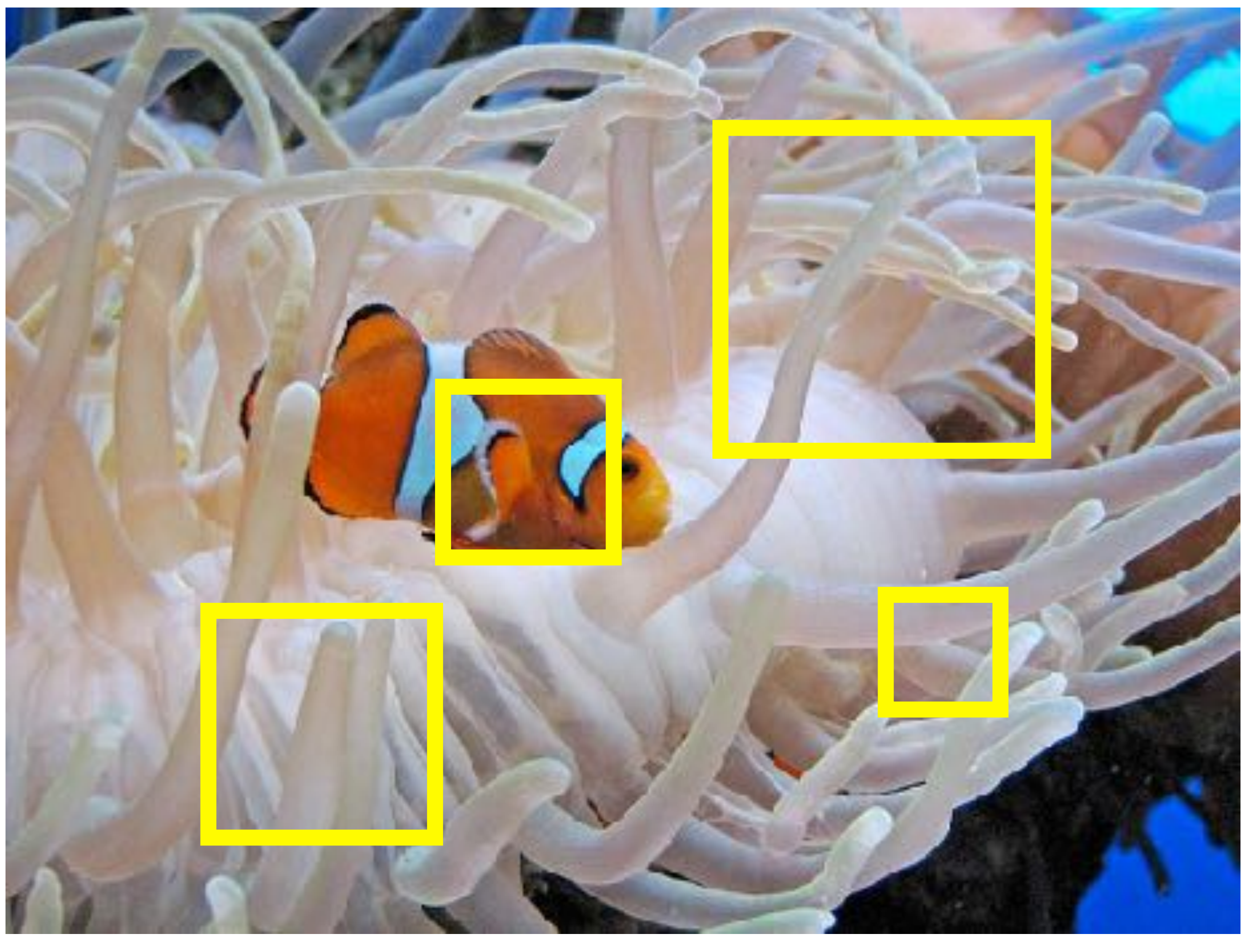
Model 2

$$d'(R_i, R_j) = \frac{\|\Psi(\mathbf{C}_i) - \Psi(\mathbf{C}_j)\|}{1 + \|\mathbf{x}_i - \mathbf{x}_j\|}$$

weighting covariance distances by inverse spatial distance decreases the influence of visually similar nearby regions

CovSal (Erdem and Erdem, 2013)

- In an image, salient parts can and do appear over a wide range of scales.
- Saliency detection should be carried out simultaneously at multiple scales.



- Employ a fusion strategy to combine single-scale maps to come up with one final saliency map:

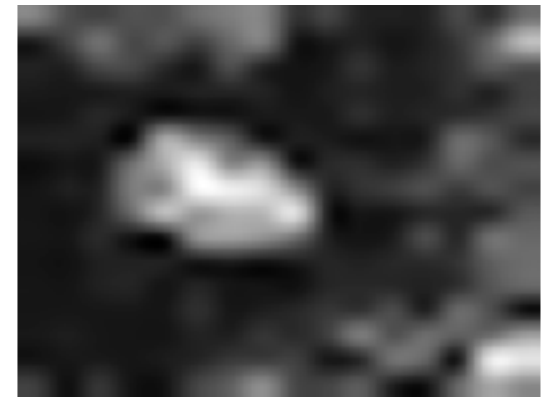
$$S(x) = G_\sigma(x) * \prod_{k \in K} \hat{S}^k(x)$$

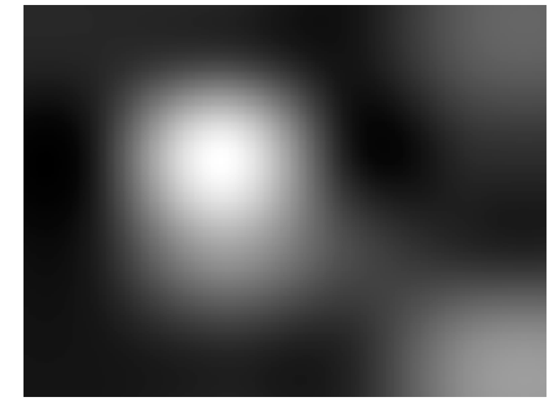
Spatial coincidence assumption:
An image part is treated as salient if it is salient at all scales.

CovSal (Erdem and Erdem, 2013)


input image

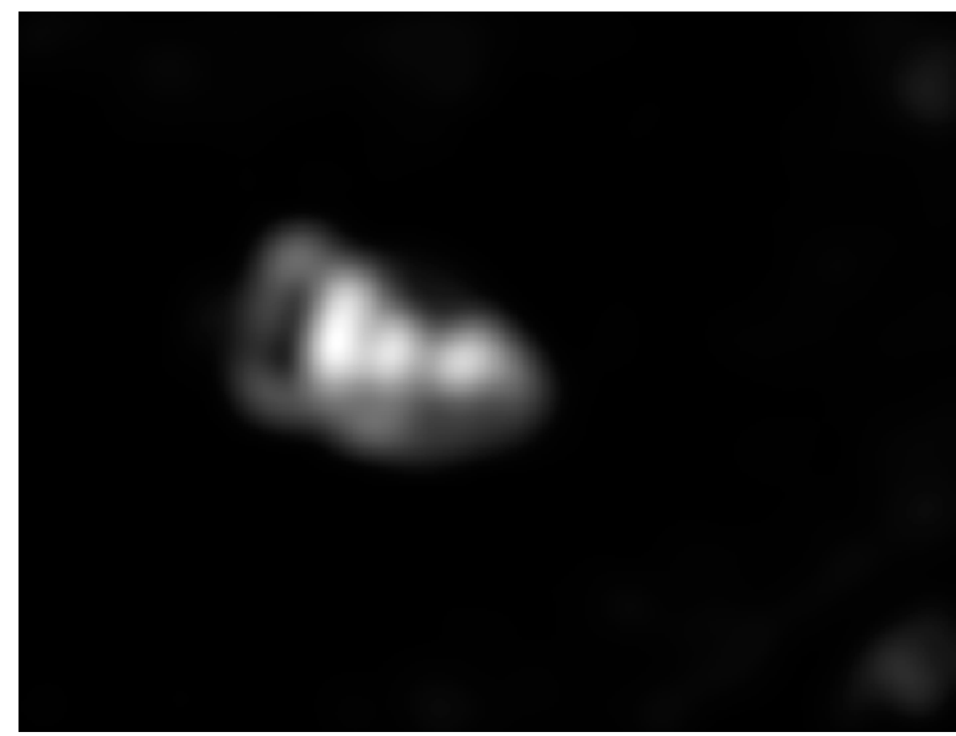

scale 1


scale 3


scale 5

- Saliency analysis at 5 different scales.

$$S(x) = G_\sigma(x) * \prod_{k \in K} \hat{S}^k(x)$$


final saliency map

Benchmark Data Sets

- Benchmark image data sets with eye fixation data (free-viewing)
  - ➡ Toronto data set [Bruce & Tsotsos, 2006]
  - ➡ MIT 1003 data set [Judd et al., 2009]
  - ➡ MIT 300 data set [Judd et al., 2012]


Fixations for one observer


Fixations from 15 observers

eye tracking experiments

[Photo Credit: Jason Dorfman CSAIL website]

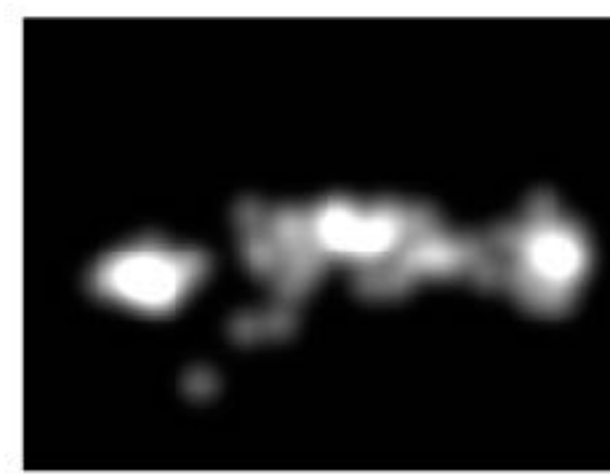
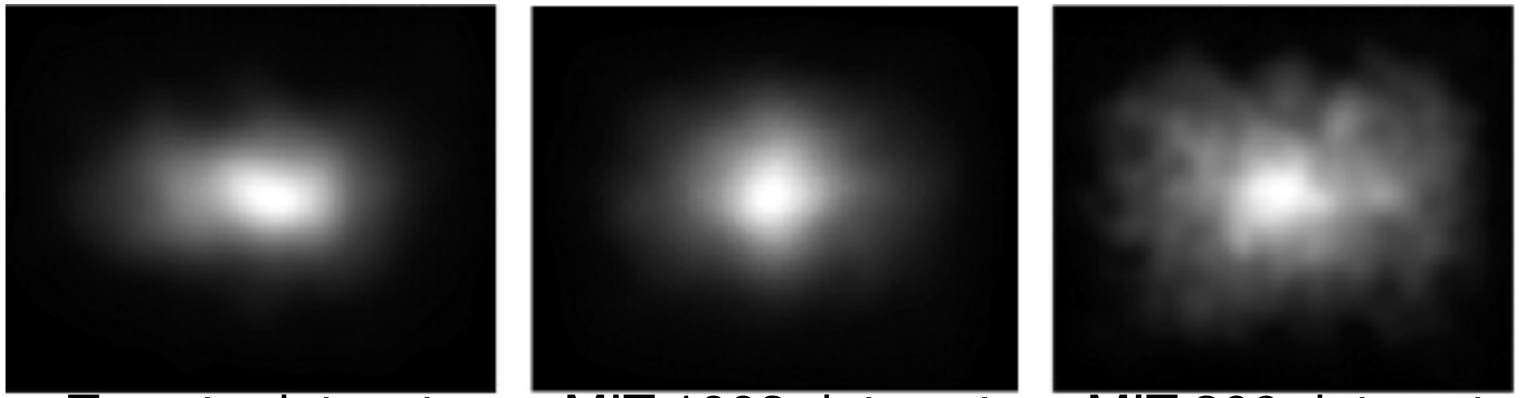

Fixation map

Image credits: T. Judd

Center bias

- Experiments show that there is a tendency in humans to look towards the image center.

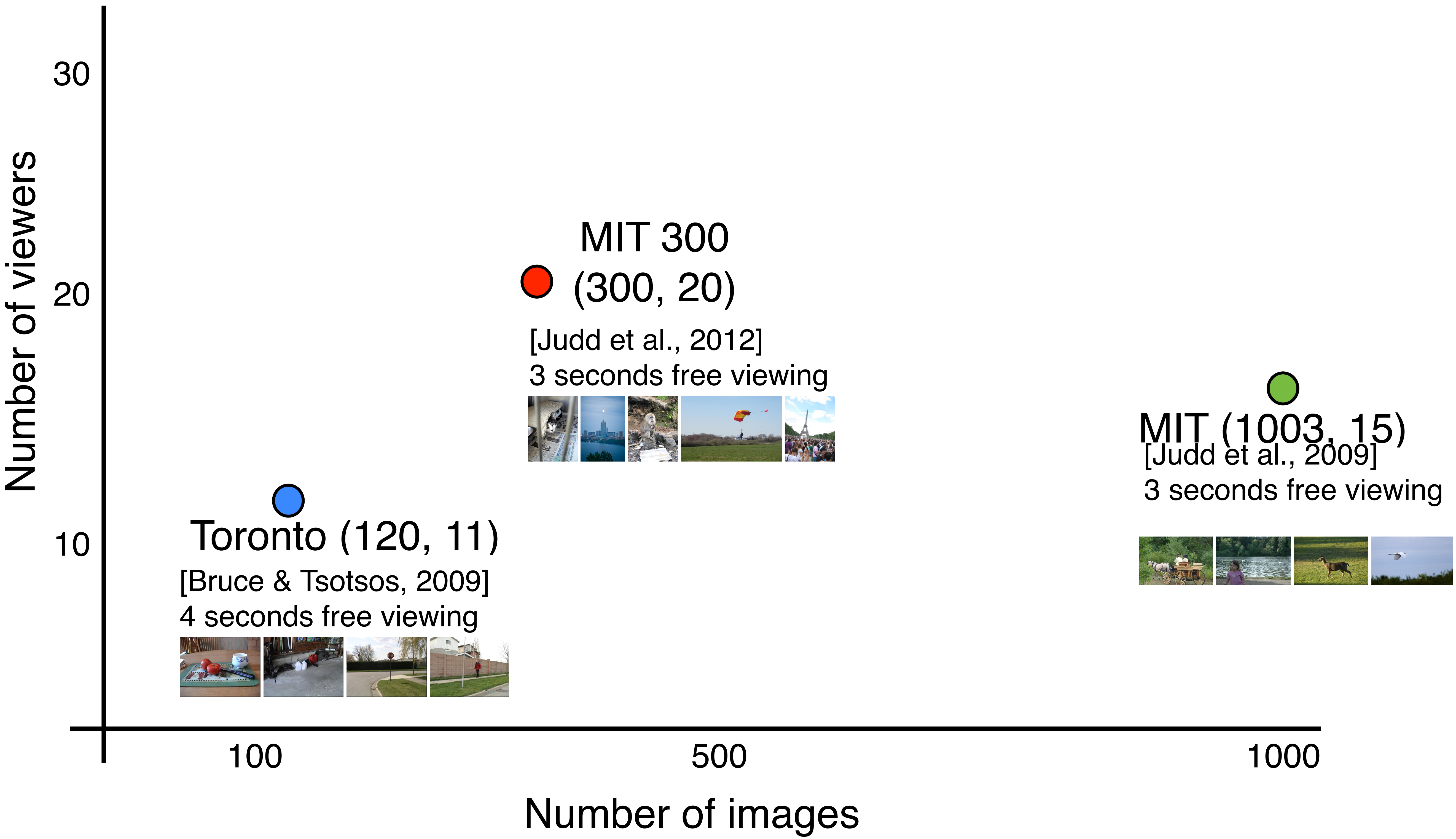
fixation maps averaged over all images



Toronto dataset MIT 1003 data set MIT 300 data set

- Why it exists?
  - ➡ photographer bias
  - ➡ viewing strategy
  - ➡ motor bias

Summary of data sets



Sample images

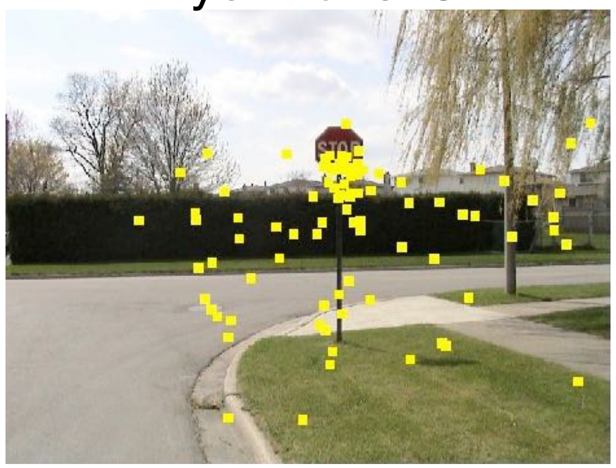
Toronto data set

MIT 1003 data set

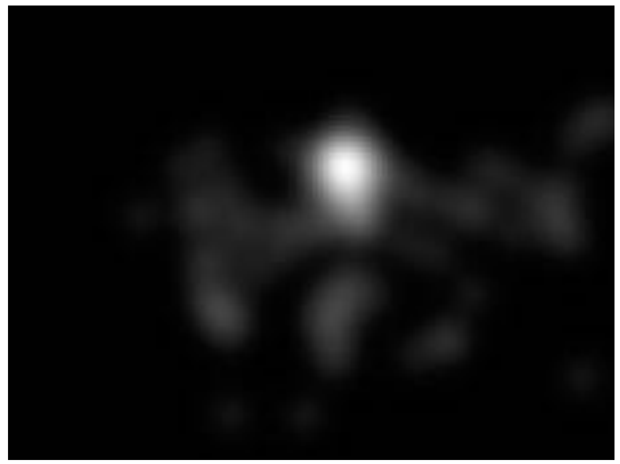
MIT 300 data set

Toronto - qualitative results

Eye fixations

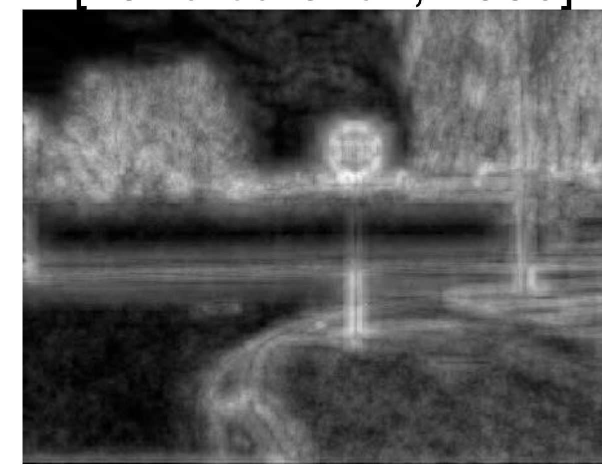


Human



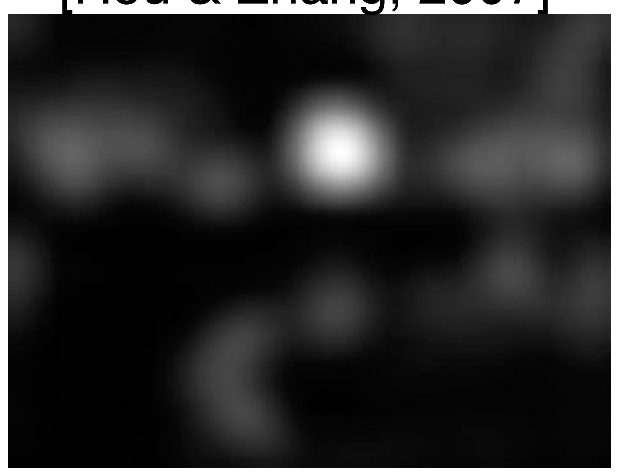
[Itti et al., 1998]

[Torralba et al., 2006]

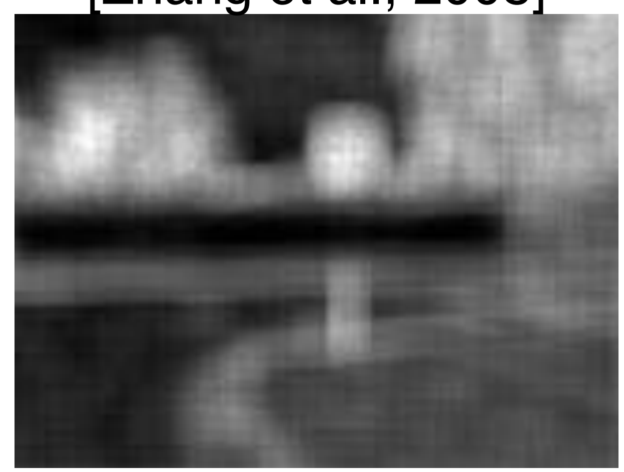


[Harel et al., 2007]

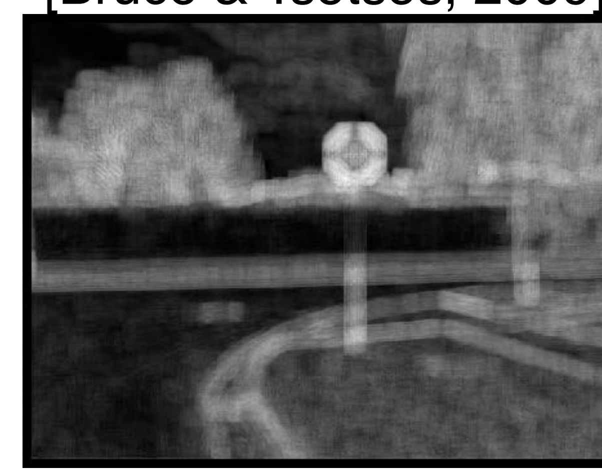
[Hou & Zhang, 2007]



[Zhang et al., 2008]



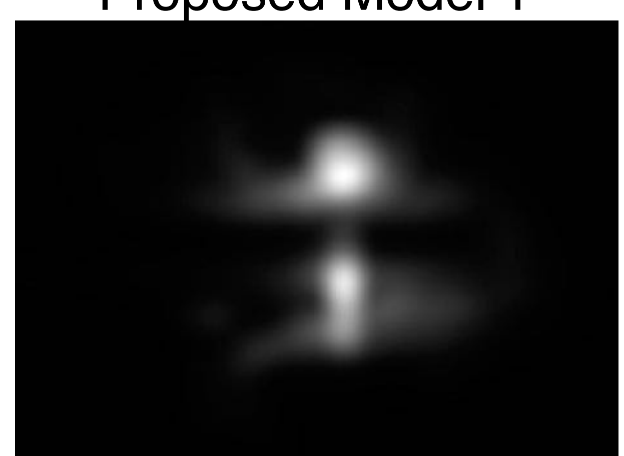
[Bruce & Tsotsos, 2009]



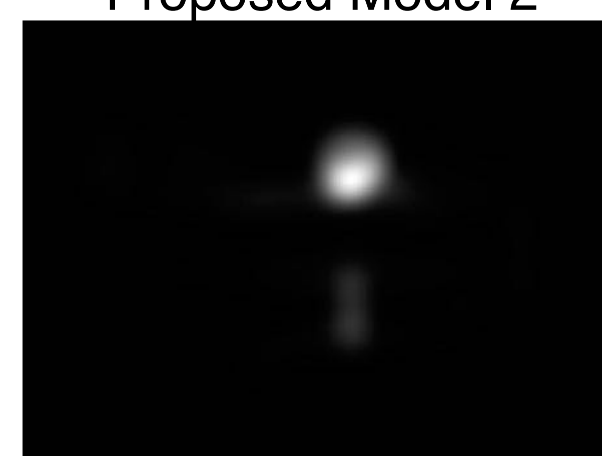
[Seo & Milanfar, 2009]

[Goferman et al., 2010]

Proposed Model 1

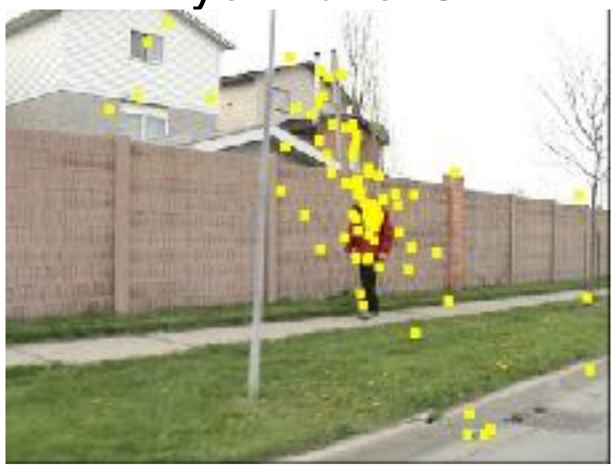


Proposed Model 2

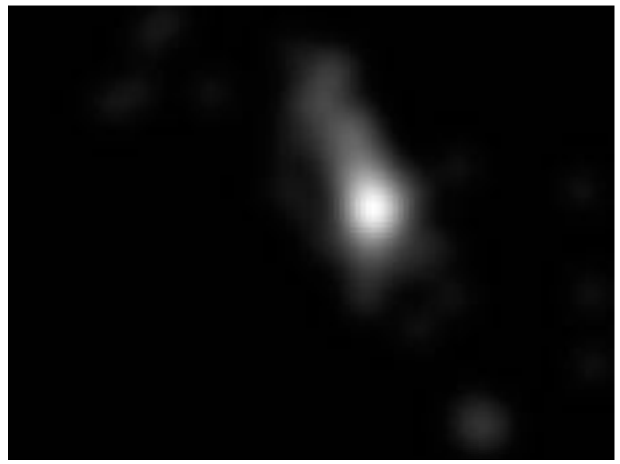


Toronto - qualitative results

Eye fixations

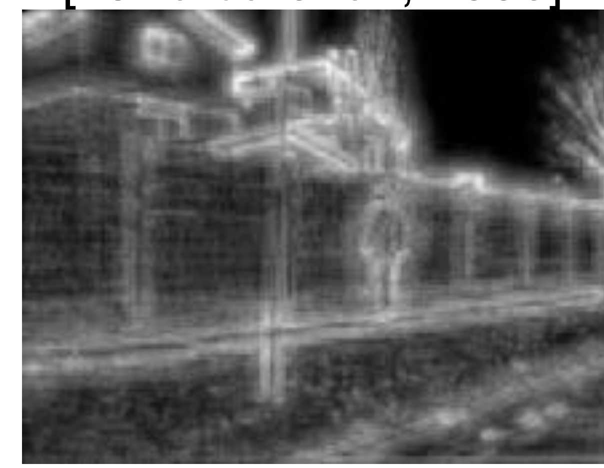


Human



[Itti et al., 1998]

[Torralba et al., 2006]

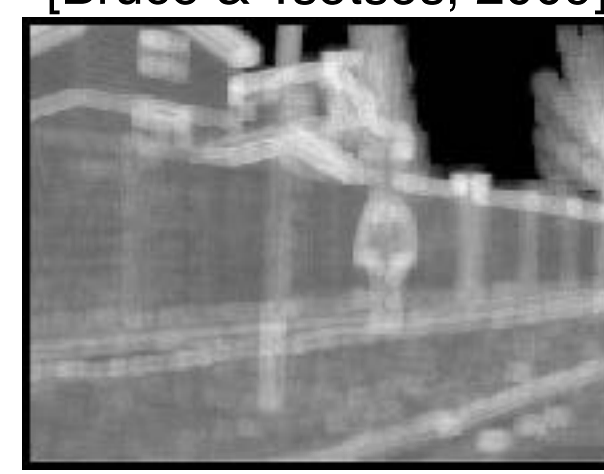


[Harel et al., 2007]

[Hou & Zhang, 2007]

[Zhang et al., 2008]

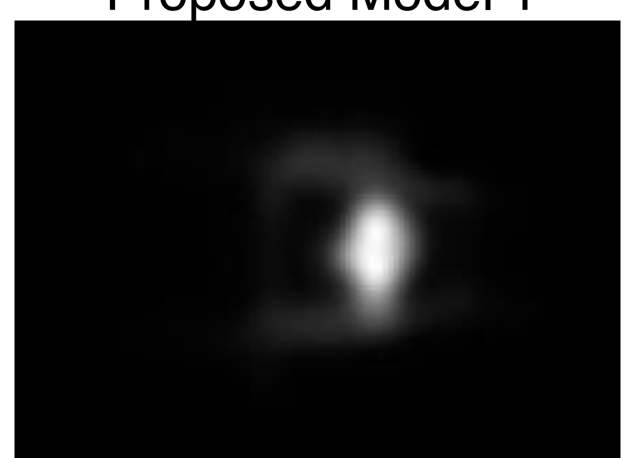
[Bruce & Tsotsos, 2009]



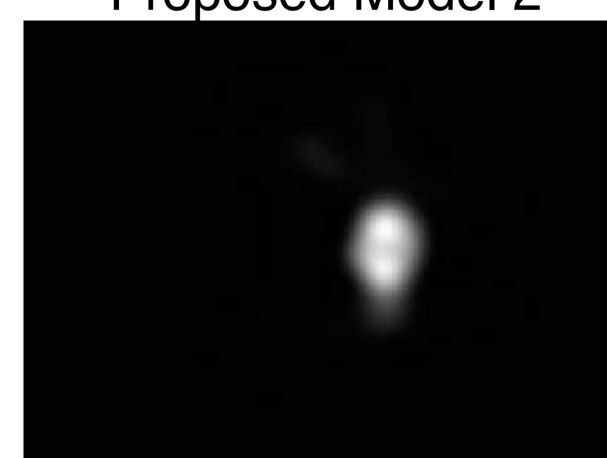
[Seo & Milanfar, 2009]

[Goferman et al., 2010]

Proposed Model 1



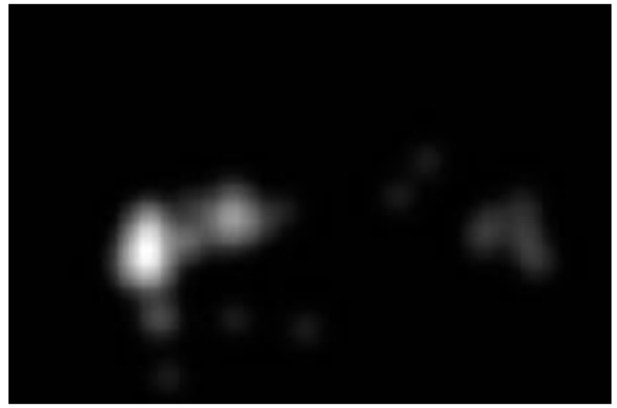
Proposed Model 2



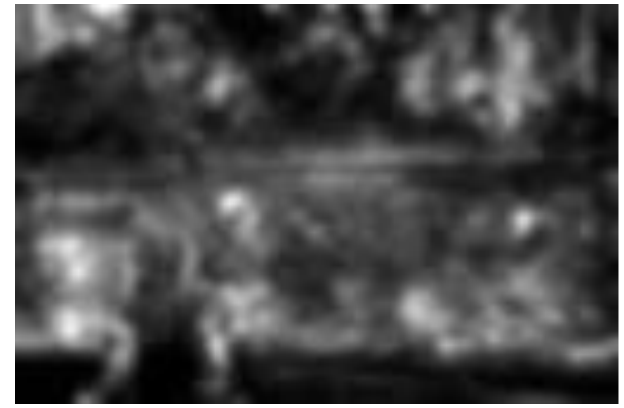
MIT 1003 - qualitative results

Eye fixations

Human



[Itti et al., 1998]

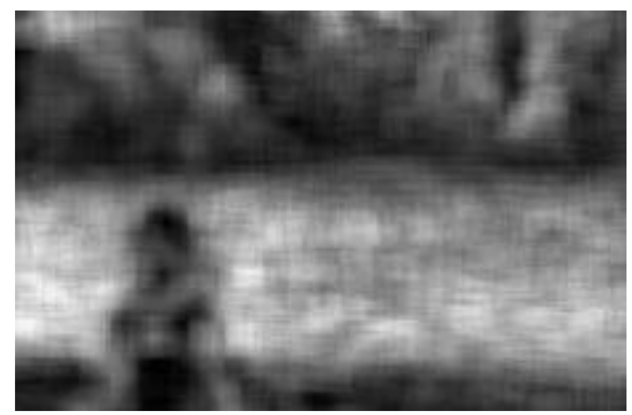


[Torralba et al., 2006]

[Harel et al., 2007]

[Hou & Zhang, 2007]

[Zhang et al., 2008]

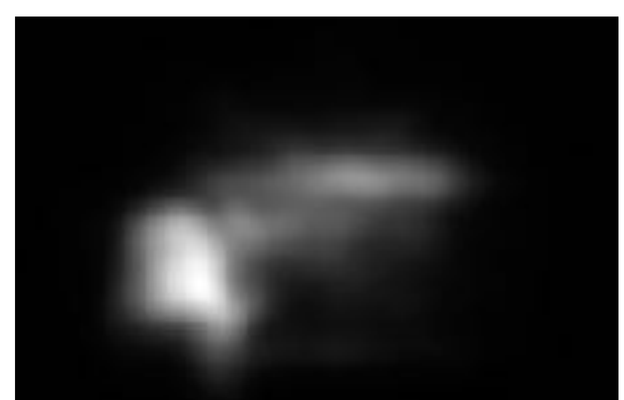


[Bruce & Tsotsos, 2009]

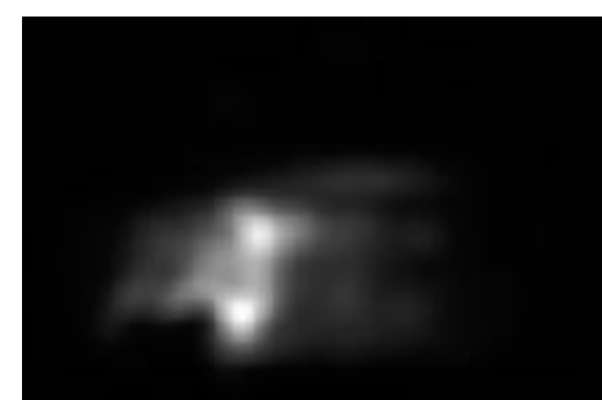
[Seo & Milanfar, 2009]

[Goferman et al., 2010]

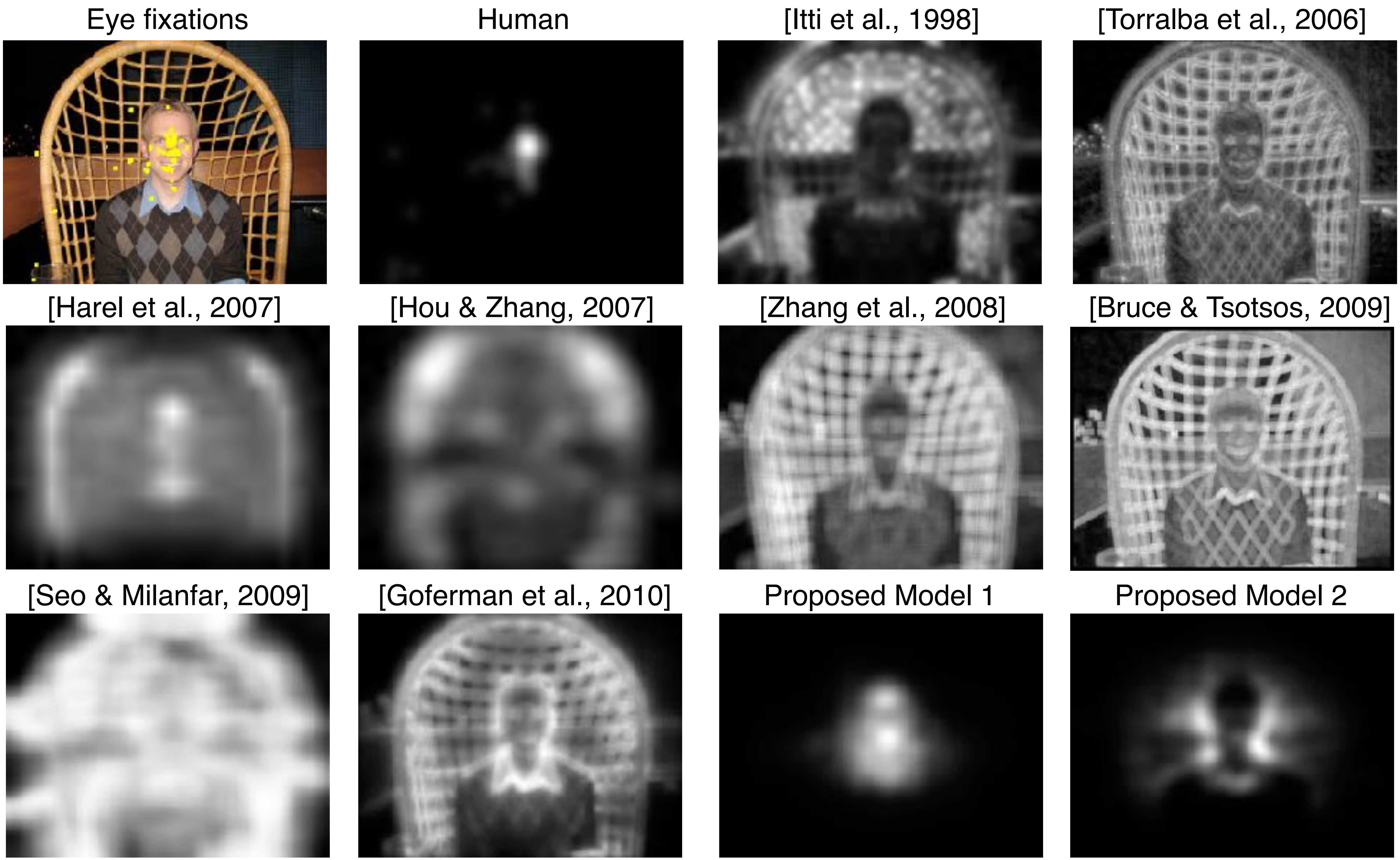
Proposed Model 1



Proposed Model 2



MIT 1003 - qualitative results



Toronto - quantitative results

	AUC		NSS		EMD		Similarity	
	Without CB	With CB	Without CB	With CB	Without CB	With CB	Without CB	With CB
Itti et al. (1998)	0.771	0.825	1.137	1.264	2.906	2.002	0.397	0.521
Harel et al. (2007)	0.829	0.835	1.533	1.533	2.014	1.886	0.519	0.556
Torralba et al. (2006)	0.710	0.832	0.805	1.185	3.467	1.868	0.330	0.528
Hou & Zhang (2007)	0.736	0.835	0.964	1.271	3.791	1.959	0.360	0.550
Zhang et al. (2008)	0.718	0.832	0.884	1.194	3.954	1.968	0.347	0.541
Bruce & Tsotsos (2009)	0.728	0.835	0.896	1.165	3.127	1.809	0.351	0.535
Seo & Milanfar (2009)	0.766	0.845	1.100	1.320	3.222	1.759	0.415	0.579
Goferman et al. (2010)	0.784	0.841	1.272	1.370	3.520	1.819	0.431	0.574
Our approach with								
Covariances only	0.767	0.834	1.184	1.342	3.142	1.931	0.408	0.546
Covariances + means	0.765	0.834	1.198	1.396	3.398	1.896	0.402	0.548
Covariances + center	0.840	0.840	1.753	1.753	1.901	1.901	0.561	0.561
Covariances + means + center	**0.851**	**0.851**	**1.891**	**1.898**	**1.728**	**1.728**	**0.581**	**0.581**
Center	–	0.803	–	0.969	–	2.401	–	0.478
Chance	0.505	0.803	−0.001	0.969	5.159	2.339	0.187	0.479

MIT 1003 - quantitative results

	AUC		NSS		Similarity	
	Without CB	With CB	Without CB	With CB	Without CB	With CB
Itti et al. (1998)	0.741	0.827	0.921	1.170	0.273	0.402
Harel et al. (2007)	0.791	0.829	1.150	1.182	0.319	0.415
Torralba et al. (2006)	0.700	0.832	0.771	1.156	0.244	0.412
Hou & Zhang (2007)	0.713	0.833	0.855	1.200	0.264	0.421
Zhang et al. (2008)	0.703	0.834	0.829	1.177	0.261	0.418
Bruce & Tsotsos (2009)	0.709	0.835	0.813	1.148	0.254	0.415
Seo & Milanfar (2009)	0.712	0.836	0.826	1.171	0.263	0.424
Goferman et al. (2010)	0.758	0.840	1.053	1.241	0.297	0.431
Our approach with						
Covariances only	0.715	0.826	0.862	1.169	0.261	0.410
Covariances + means	0.740	0.832	0.940	1.240	0.287	0.417
Covariances + center	0.833	0.833	1.468	1.486	0.417	0.418
Covariances + means + center	**0.843**	**0.843**	**1.488**	**1.543**	**0.428**	**0.432**
Center	–	0.810	–	1.004	–	0.379
Chance	0.500	0.810	−0.000	1.004	0.131	0.383

MIT 300 - quantitative results

	AUC		EMD		Similarity	
	Without CB	With CB	Without CB	With CB	Without CB	With CB
Itti et al. (1998)	0.750	0.806	4.560	3.394	0.405	0.493
Harel et al. (2007)	0.801	0.813	3.574	3.315	0.472	0.501
Torralba et al. (2006)	0.684	0.806	4.715	**3.036**	0.343	0.488
Hou & Zhang (2007)	0.682	0.804	5.368	3.200	0.319	0.487
Zhang et al. (2008)	0.672	0.799	5.088	3.296	0.340	0.473
Bruce & Tsotsos (2009)	0.751	**0.820**	4.236	3.085	0.390	0.507
Goferman et al. (2010)	0.742	0.815	4.900	3.219	0.390	**0.509**
Our approach with						
Covariances + center	0.800	0.800	3.422	3.422	0.487	0.487
Covariances + means + center	**0.806**	0.811	**3.109**	3.109	**0.502**	0.503
Center	–	0.783	–	3.719	–	0.451
Chance	0.503	0.783	6.352	3.506	0.327	0.482
Judd et al. (2009)	0.811	0.813	3.130	3.130	0.506	0.511

Salient object detection

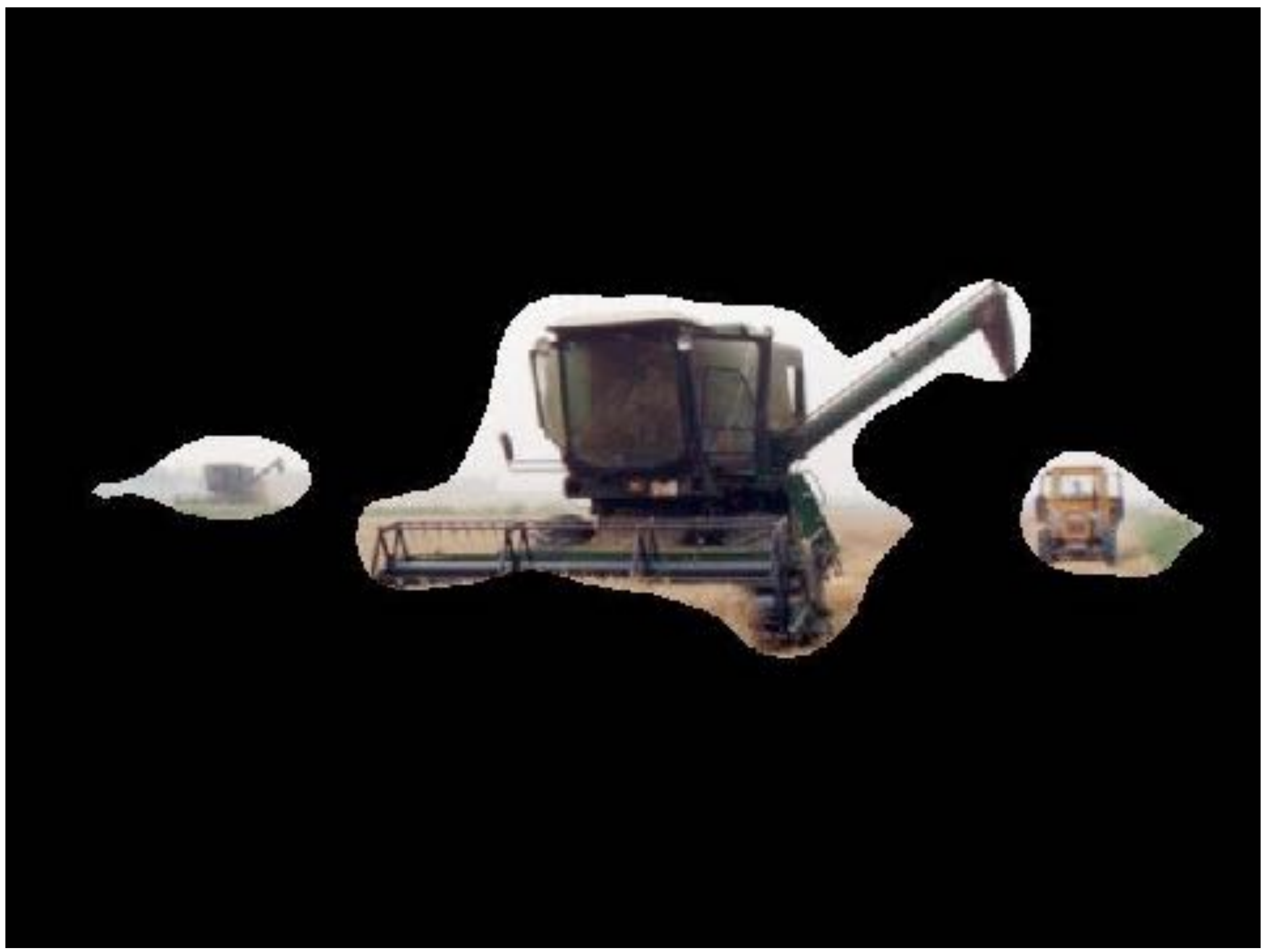
- Task of identifying foreground objects that attract more attention.
- ImgSal data set [Li et al., 2012]
  - ➡ 235 natural color images
  - ➡ Six different categories:
    - large salient regions (50 images),
    - intermediate salient regions (80 images),
    - small salient regions (60 images),
    - cluttered backgrounds (15 images),
    - repeating distractors (15 images),
    - large and small salient regions (15 images)

input image

labeling by 1 human subject

ground truth labeling agreed upon 19 subjects

ImgSal - qualitative results

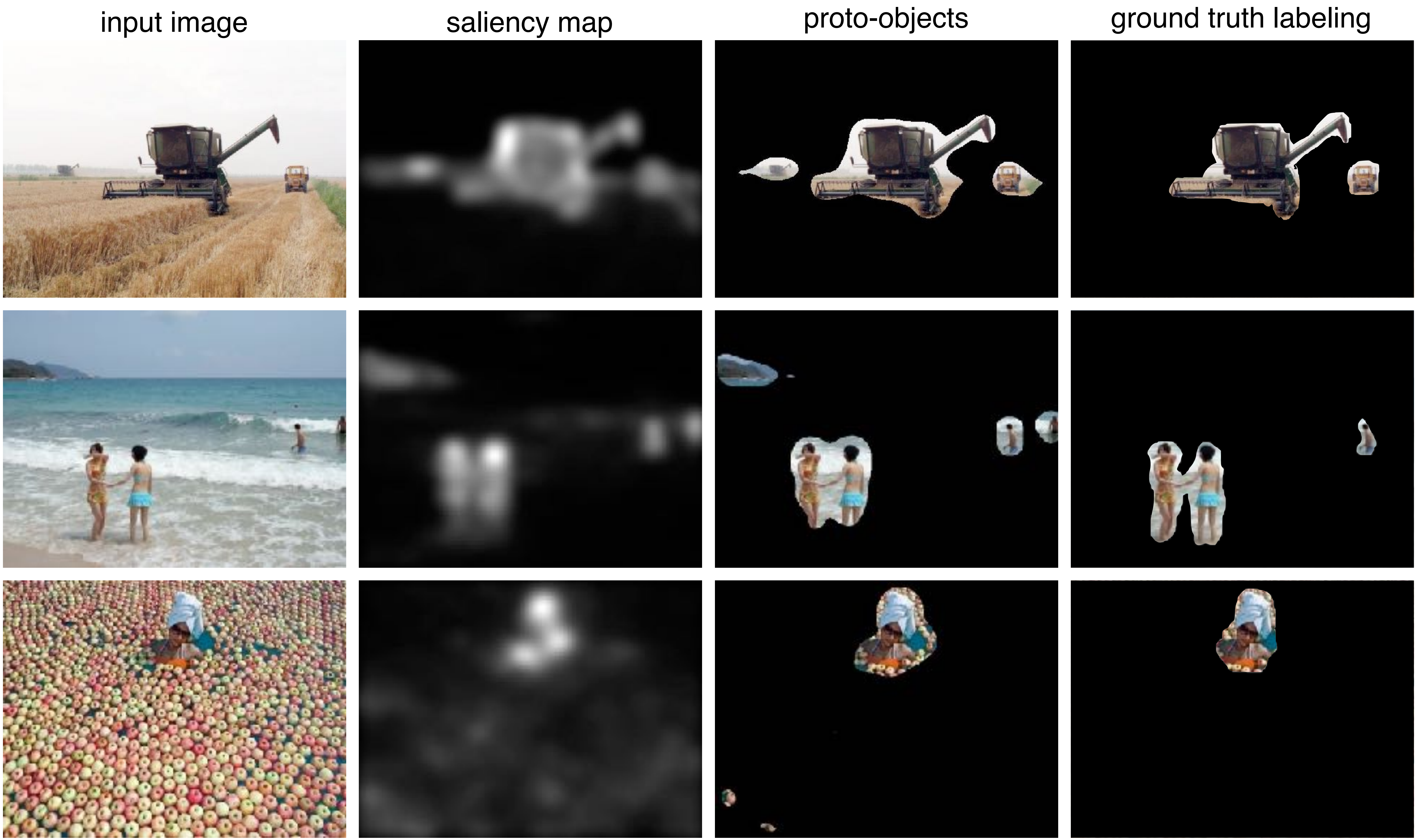


detected saliency map

extracted salient objects

Ground truth labeling

ImgSal - qualitative results



ImgSal - qualitative results

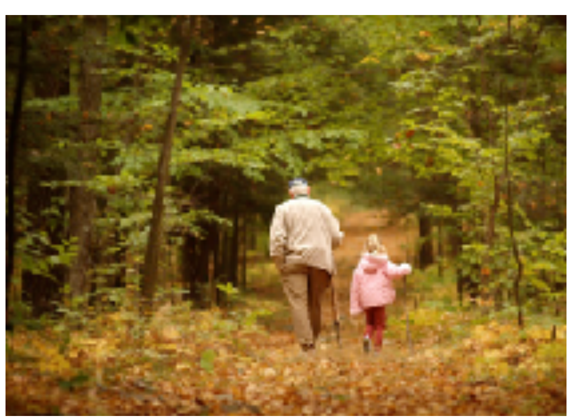
input image | saliency map | proto-objects | ground truth labeling

ImgSal - quantitative results

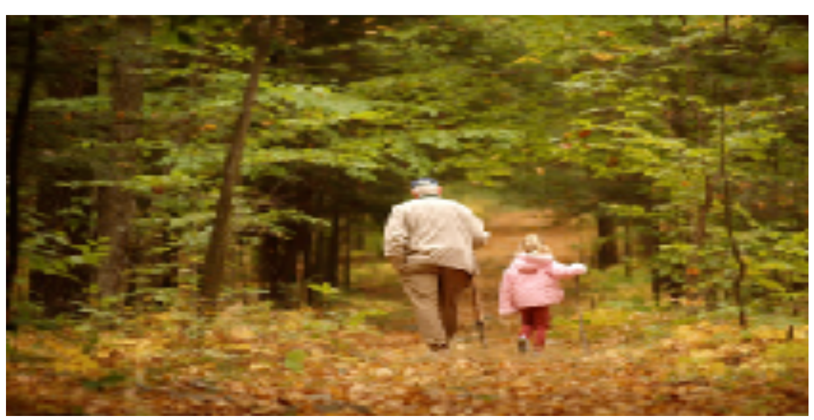
	Large salient regions		Intermediate salient regions		Small salient regions		Cluttered backgrounds		Repeating distractors		Large and small salient regions	
	AUC	DSC	AUC	DSC	AUC	DSC	AUC	DSC	AUC	DSC	AUC	DSC
Itti et al. (1998)	0.897	0.610	0.897	0.473	0.937	0.401	0.824	0.335	0.891	0.439	0.936	0.639
Harel et al. (2007)	**0.945**	**0.694**	0.925	0.529	0.951	0.463	0.916	0.499	0.934	0.557	**0.952**	0.688
Torralba et al. (2006)	0.790	0.469	0.825	0.377	0.929	0.372	0.700	0.239	0.750	0.306	0.870	0.515
Hou & Zhang (2007)	0.833	0.524	0.861	0.448	0.939	0.411	0.769	0.308	0.809	0.369	0.918	0.584
Zhang et al. (2008)	0.760	0.461	0.813	0.391	0.895	0.366	0.676	0.270	0.755	0.325	0.850	0.504
Bruce & Tsotsos (2009)	0.798	0.480	0.825	0.383	0.914	0.357	0.759	0.288	0.788	0.350	0.855	0.494
Seo & Milanfar (2009)	0.842	0.563	0.896	0.474	0.948	0.430	0.776	0.284	0.878	0.451	0.916	0.611
Goferman et al. (2010)	0.905	0.636	**0.950**	**0.610**	0.970	0.553	0.919	0.509	0.914	0.581	0.947	0.723
Our approach with												
Covariances only	0.920	0.666	0.928	0.548	0.957	0.470	**0.933**	**0.554**	0.947	0.664	0.946	0.645
Covariances + means	0.866	0.614	0.924	0.584	**0.972**	**0.586**	0.818	0.425	0.948	0.635	0.938	**0.728**
Covariances + center	0.919	0.681	0.909	0.517	0.919	0.329	0.905	0.500	**0.961**	0.654	0.893	0.574
Covariances + means + center	0.865	0.673	0.912	0.580	0.954	0.508	0.879	0.441	0.960	**0.698**	0.888	0.664

Image retargeting

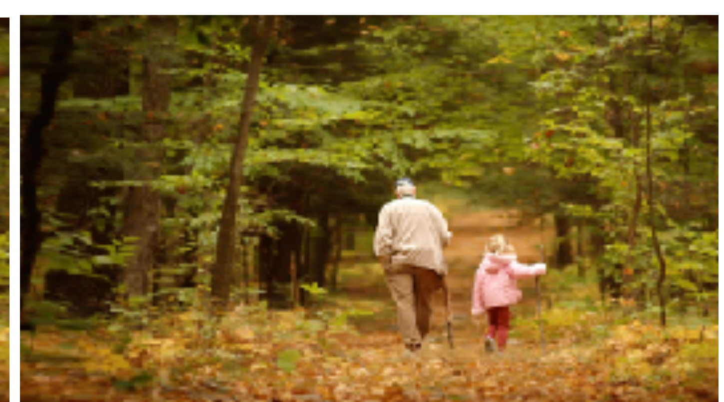
- aka content aware image resizing
- automatically resizing an image to arbitrary aspect ratios while trying to preserve important content
- ReTargetMe data set [Li et al., 2012]
  - ➡ 80 images with 92 different resizing scenarios
  - ➡ categorized into nine groups:
    - lines/clear edges,
    - faces/ people,
    - recurring texture,
    - symmetry,
    - textual elements,
    - outdoor/nature
    - indoor
    - geometric structures,
    - evident foreground objects,



input image



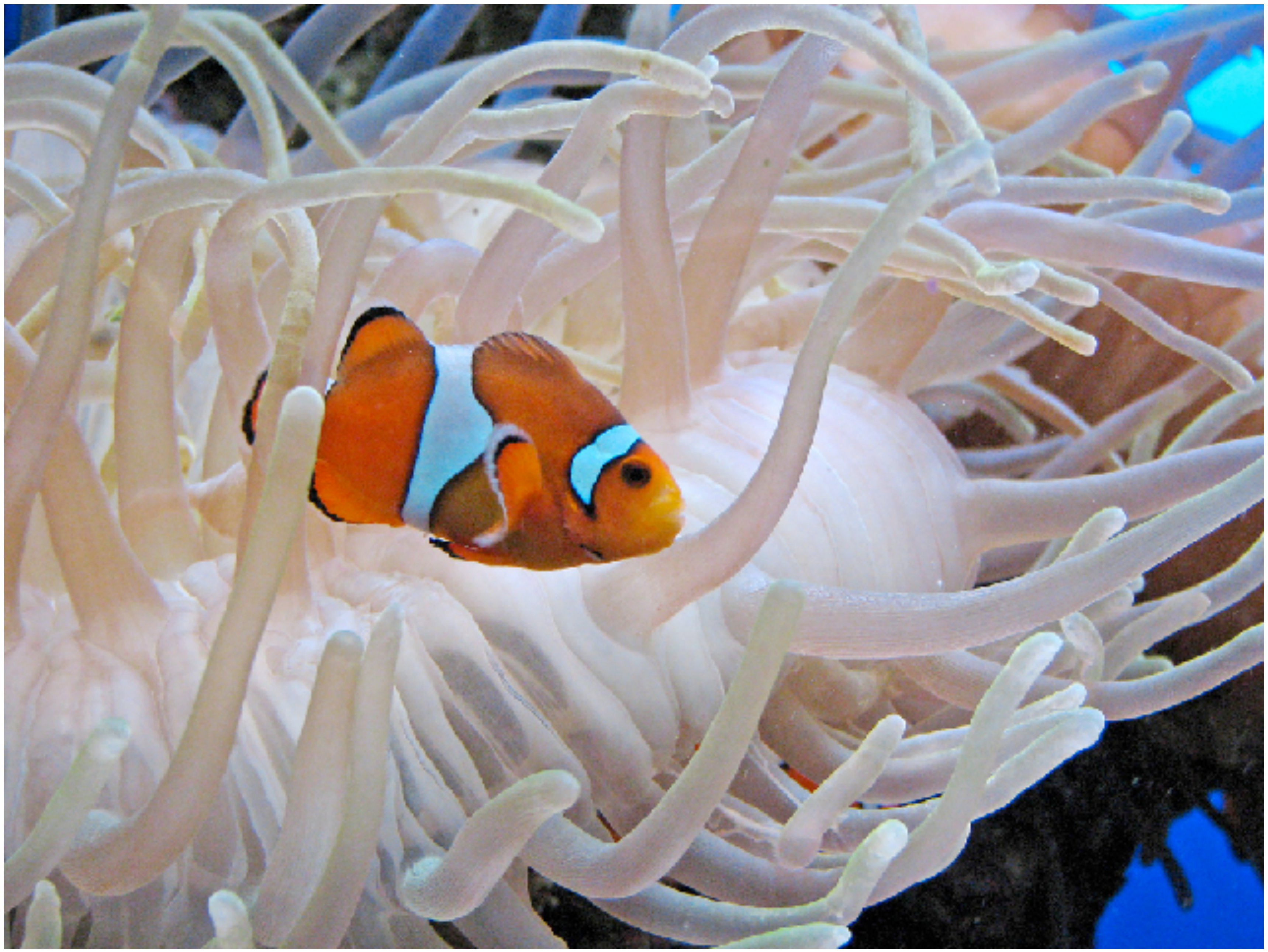
resized image



Seam Carving
[Avidan & Shamir, 2007]

replicated seams

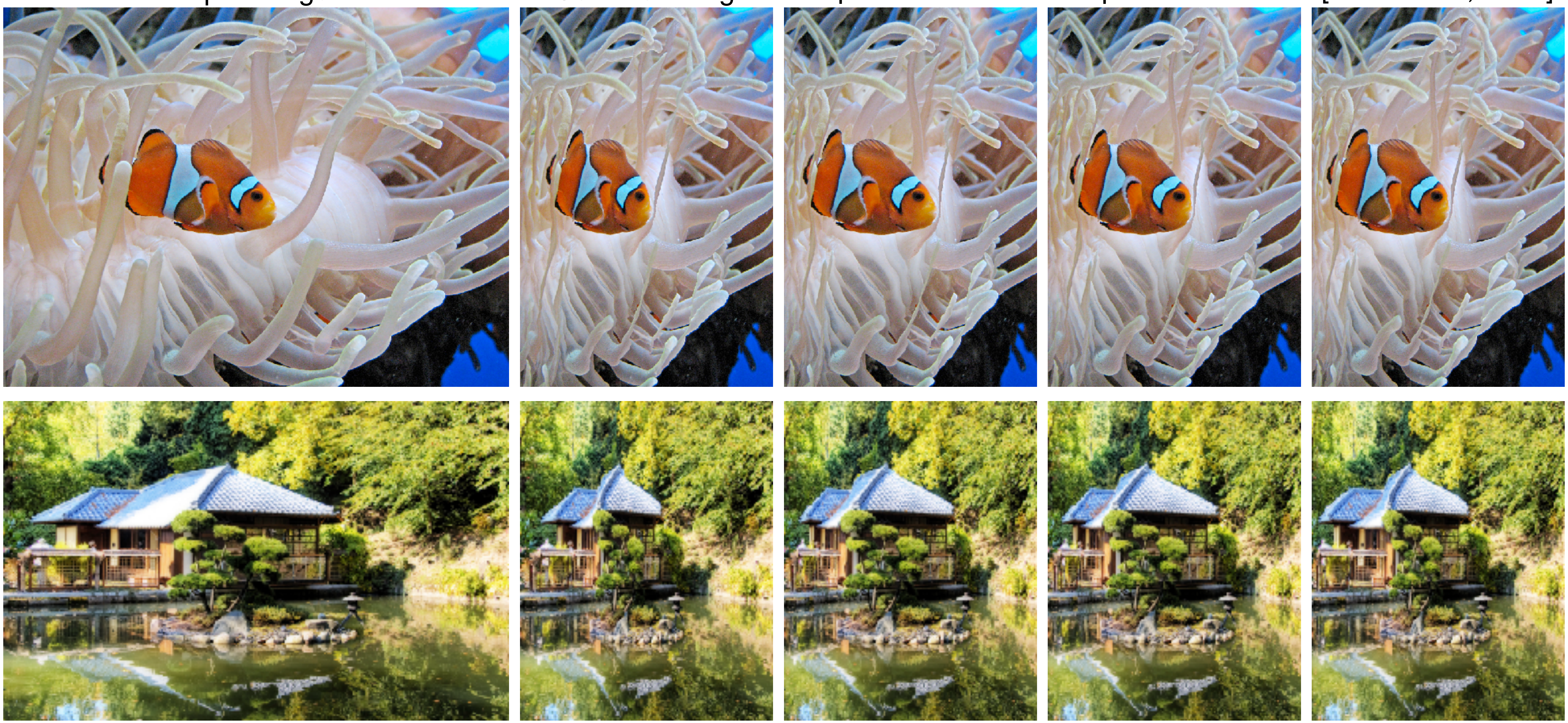
ReTargetMe - qualitative results


Input image


[Avidan & Shamir, 2007]

ReTargetMe - qualitative results

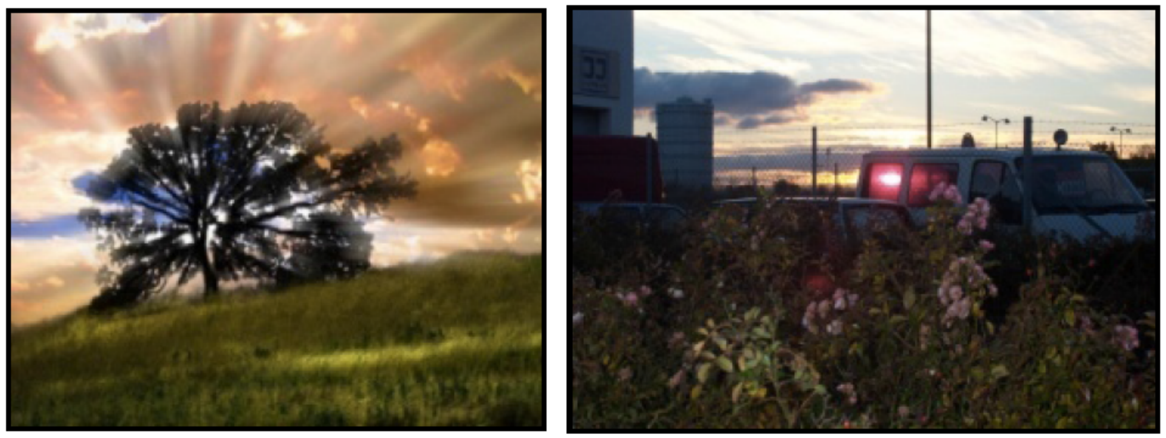
Input Image | Seam Carving | Proposed Model 1 | Proposed Model 2 | [Harel et al., 2007]



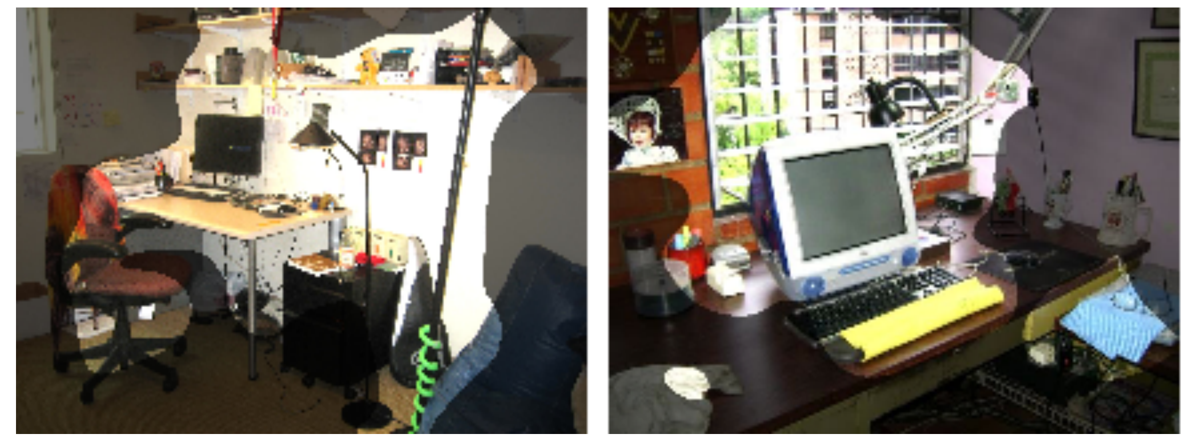
ReTargetMe - quantitative results

	Lines/clear edges		Faces/ people		Recurring texture		Evident fore-ground objects		Geometric structures		Symmetry		Textual elements		Outdoor / Nature		Indoor	
	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow	EMD	SIFTflow
Itti et al. (1998)	7.62	8.85	6.48	8.45	6.80	8.50	6.13	8.41	8.08	9.29	8.06	8.06	6.50	4.67	6.62	7.41	8.10	9.70
Harel et al. (2007)	3.95	5.71	4.10	**5.03**	4.20	8.40	4.37	5.17	4.39	6.03	3.59	**4.59**	**2.83**	8.33	**3.41**	5.78	4.50	6.00
Torralba et al. (2006)	7.71	8.60	7.86	8.17	7.20	9.00	8.48	8.57	7.63	8.84	9.65	8.94	9.83	6.67	7.97	7.68	6.40	8.40
Hou & Zhang (2007)	9.71	8.04	8.59	9.17	9.80	5.80	8.61	8.52	9.34	7.84	10.24	6.82	10.00	9.33	9.59	8.32	8.60	9.80
Zhang et al. (2008)	9.05	7.87	10.38	8.07	9.10	7.60	9.59	8.87	9.26	7.95	9.47	8.41	11.00	9.50	9.14	8.14	8.10	7.90
Bruce & Tsotsos (2009)	5.76	10.45	5.86	10.10	5.90	10.80	6.85	9.91	5.61	10.42	6.00	9.94	6.33	8.67	6.32	9.59	6.00	9.40
Seo & Milanfar (2009)	8.42	8.25	8.17	9.00	7.60	7.90	7.91	8.61	7.95	8.21	8.71	9.65	7.00	11.17	7.59	9.41	7.00	8.20
Goferman et al. (2010)	8.78	6.80	9.07	5.69	9.10	6.70	8.61	5.76	8.87	7.03	7.12	7.88	4.67	7.33	7.86	5.97	7.60	8.10
Our approach with																		
Covariances only	5.85	5.56	5.86	5.69	6.50	**4.40**	6.63	6.50	5.29	5.05	5.24	5.82	7.00	6.50	6.59	**5.68**	5.60	4.90
Covariances + means	9.04	**5.09**	8.31	6.00	9.90	5.10	7.89	5.39	9.68	**4.61**	8.24	4.82	8.67	7.00	8.59	5.95	10.50	**4.30**
Covariances + center	**2.84**	5.67	**3.10**	5.52	**2.90**	4.70	**3.26**	5.54	**2.68**	5.45	**2.59**	5.53	**2.83**	5.50	3.73	6.03	**2.70**	4.60
Covariances + means + center	4.75	5.75	5.28	5.14	4.70	5.80	4.87	**4.93**	4.76	5.97	4.18	5.53	5.17	**4.33**	5.30	6.32	7.20	5.90
Seam carving (Avidan & Shamir, 2007)	7.53	4.35	7.93	4.97	7.30	6.30	7.80	4.80	7.45	4.32	7.94	5.00	9.17	2.00	8.27	4.73	8.70	3.80

Beyond saliency - feature selection



Aesthetic class prediction,
Wong and Low, ICIP 2009



Scene recognition,
Fornoni and Caputo, BMVC 2012

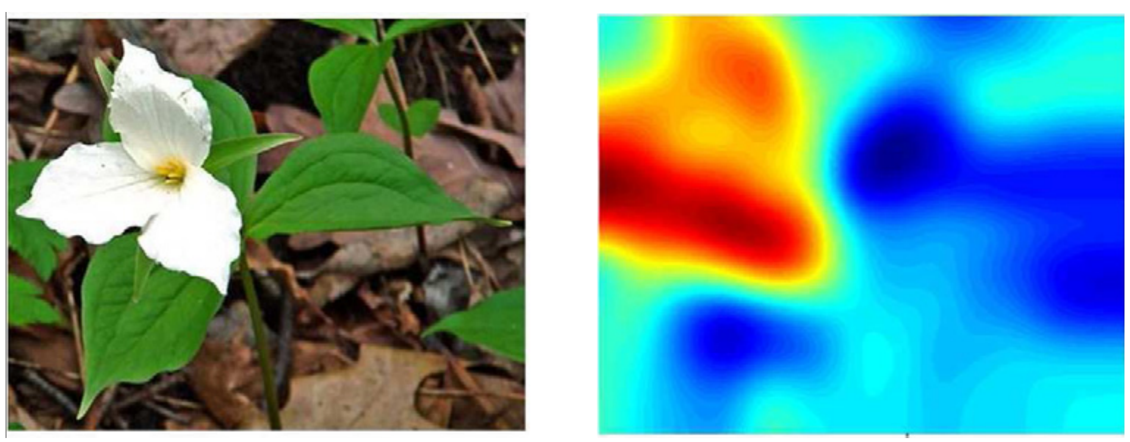
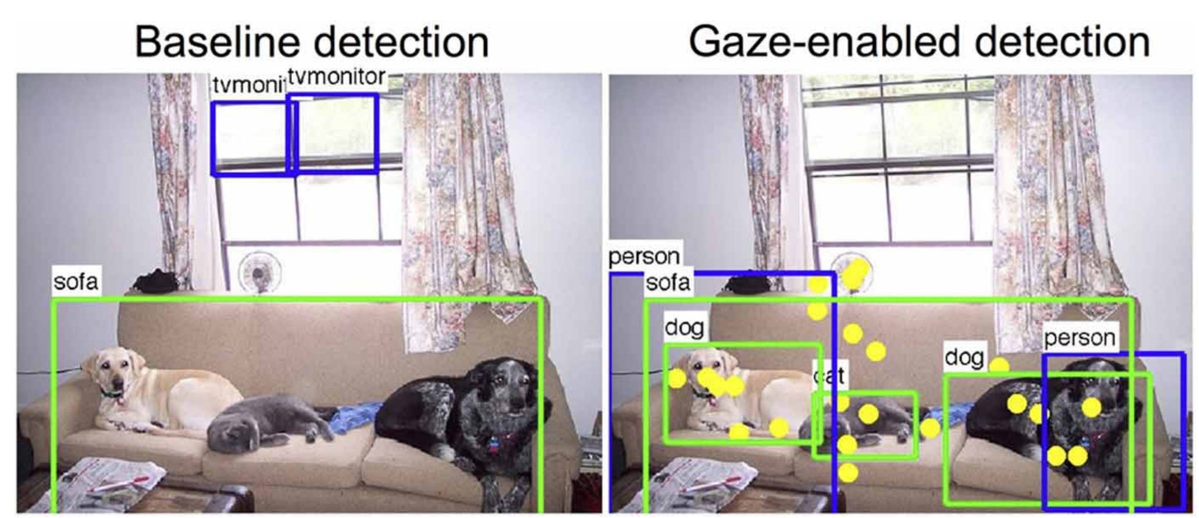


Image classification,
de Campos et al., CVIU 2012

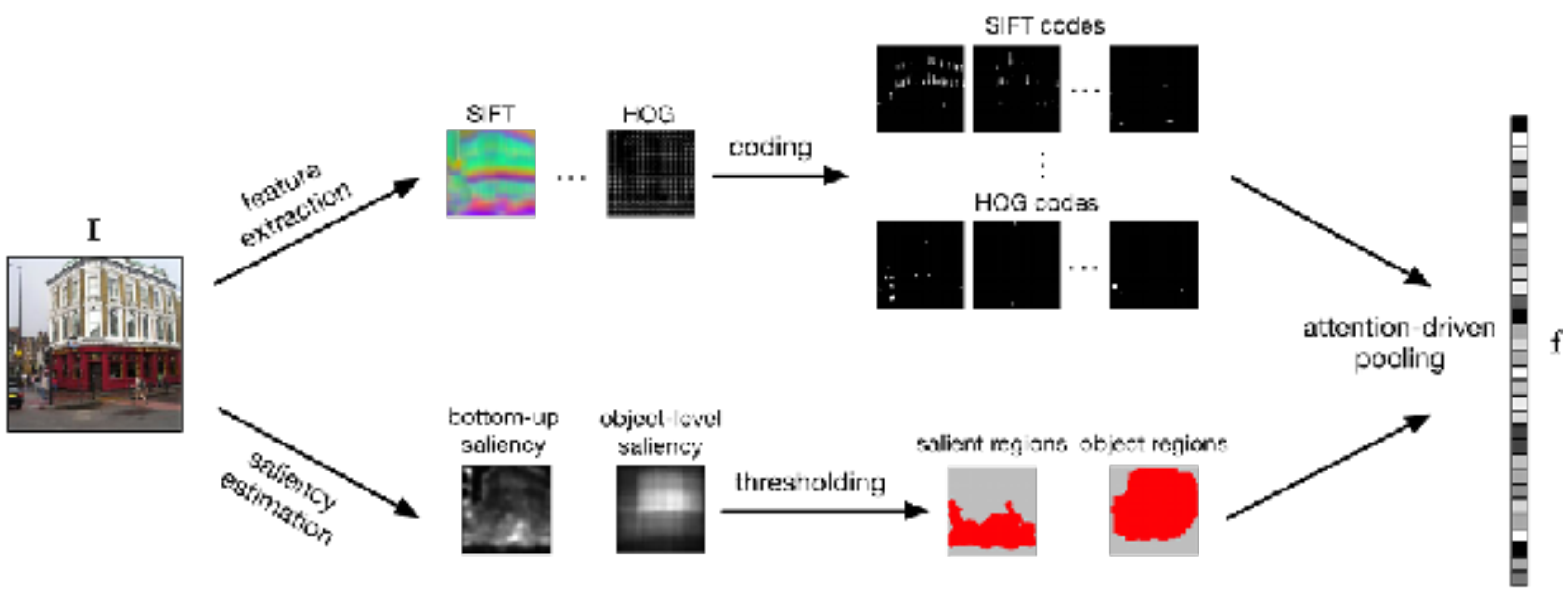


Object detection,
Yun et al., CVPR 2013

Beyond saliency - feature selection

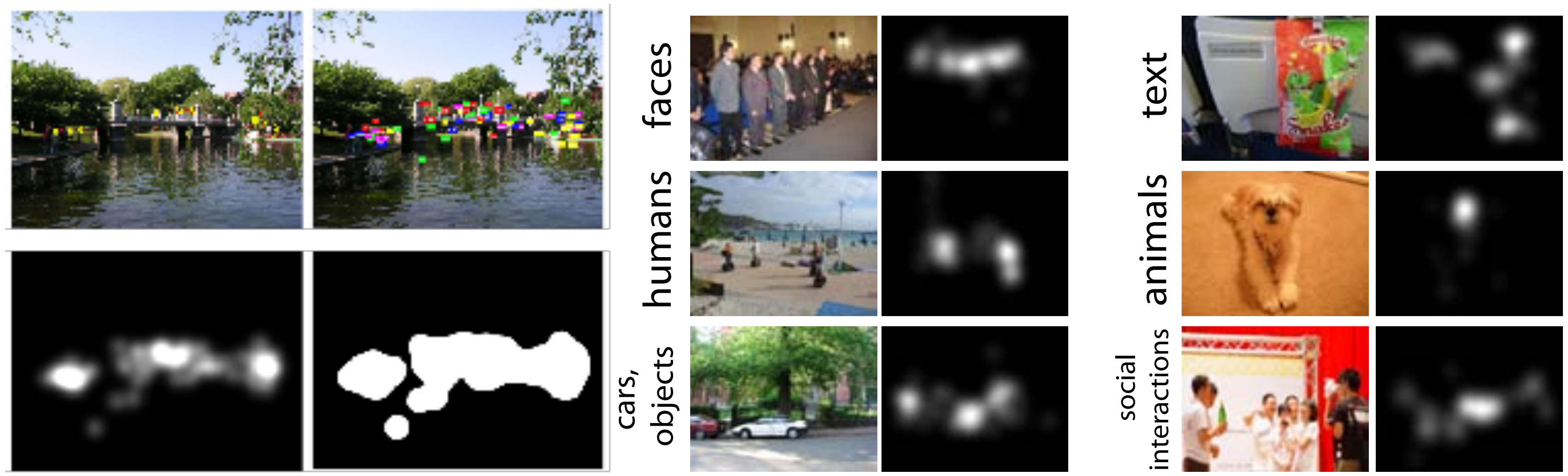
- Relationship between image memorability and attention

  B. Celikkale, A. Erdem and E. Erdem, *Predicting Memorability of Images Using Attention-driven Spatial Pooling and Image Semantics*, Image and Vision Computing, 42, pp. 35-46, October 2015 **(Editor's choice article)**

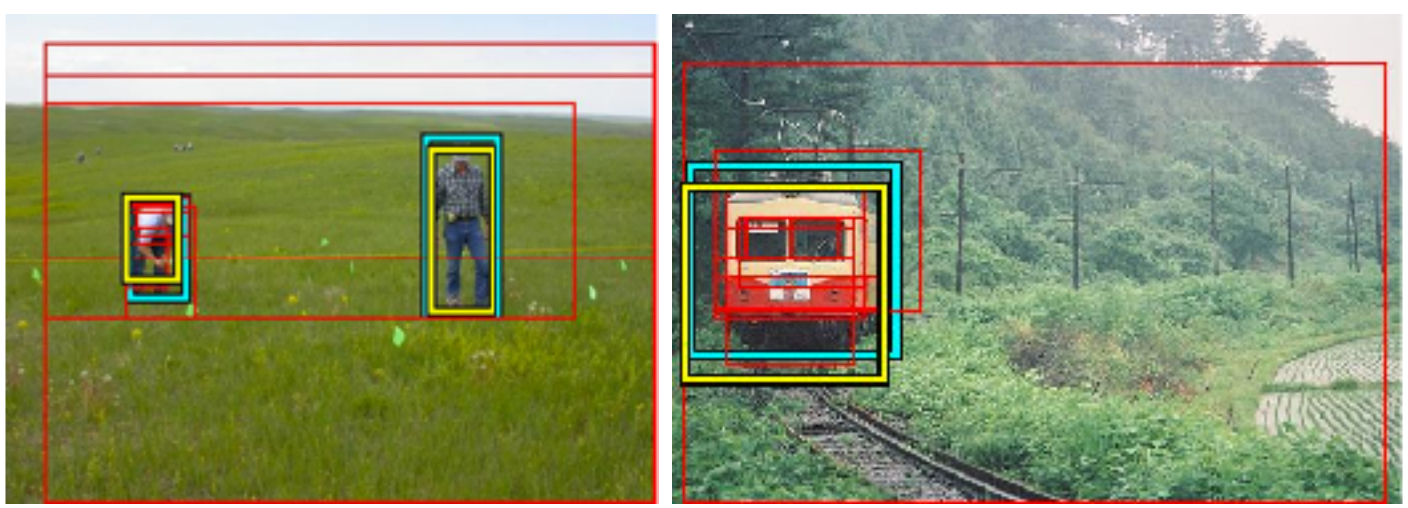


Predicted as highly memorable (89%) Predicted as typically memorable (67%) Predicted as least memorable (48%)

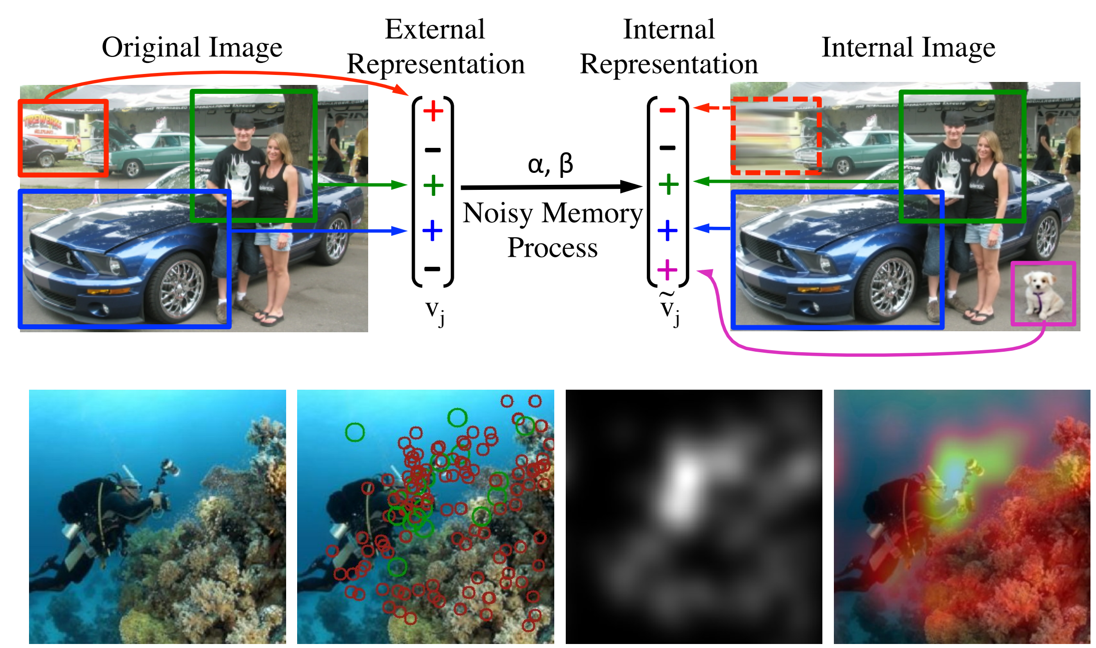
Beyond saliency - as a feature



Learning saliency,
Judd et al., ICCV 2009, Borji, CVPR 2012



Generic objectness,
Alexe et al., CVPR 2010



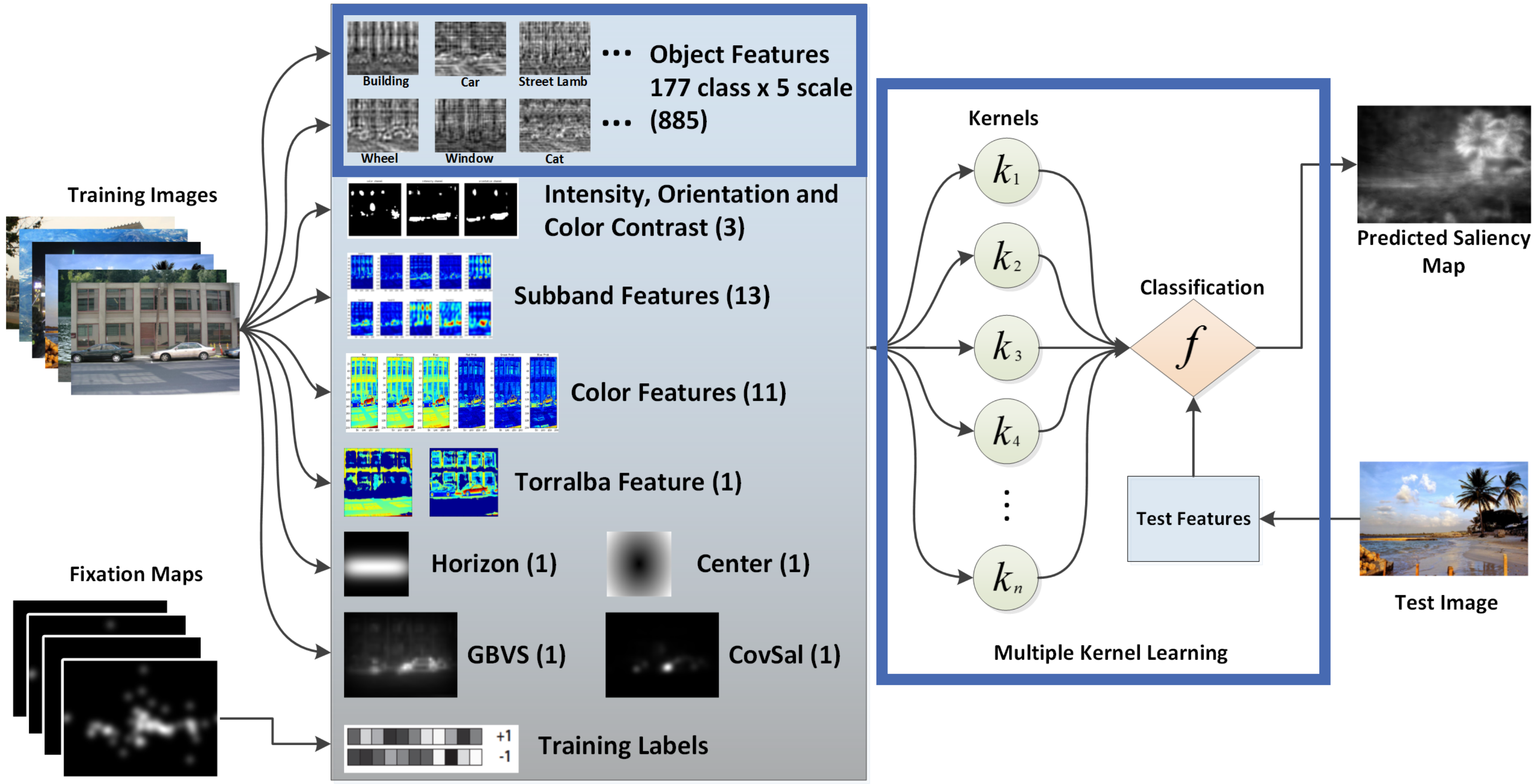
Memorability prediction,
Khosla et al., NIPS 2012
Mancas and le Meur, ICIP 2013

Beyond saliency - as a feature

- Learning visual saliency

  Y. Kavak, E. Erdem and A. Erdem, *Visual saliency estimation by integrating features using multiple kernel learning,* 6th International Symposium on Attention in Cognitive Systems (ISACS 2013), Beijing, China, August 2013.

- Automatically choose features relevant to visual saliency by learning specific feature weights and normalization schemes in the integration step.



Problems with saliency models?

- Important information may not be visually salient (e.g., stop sign in a cluttered scene)
- Salient information may not be important
- Can not account for many fixations when there is a task

Original image

Bottom-up saliency

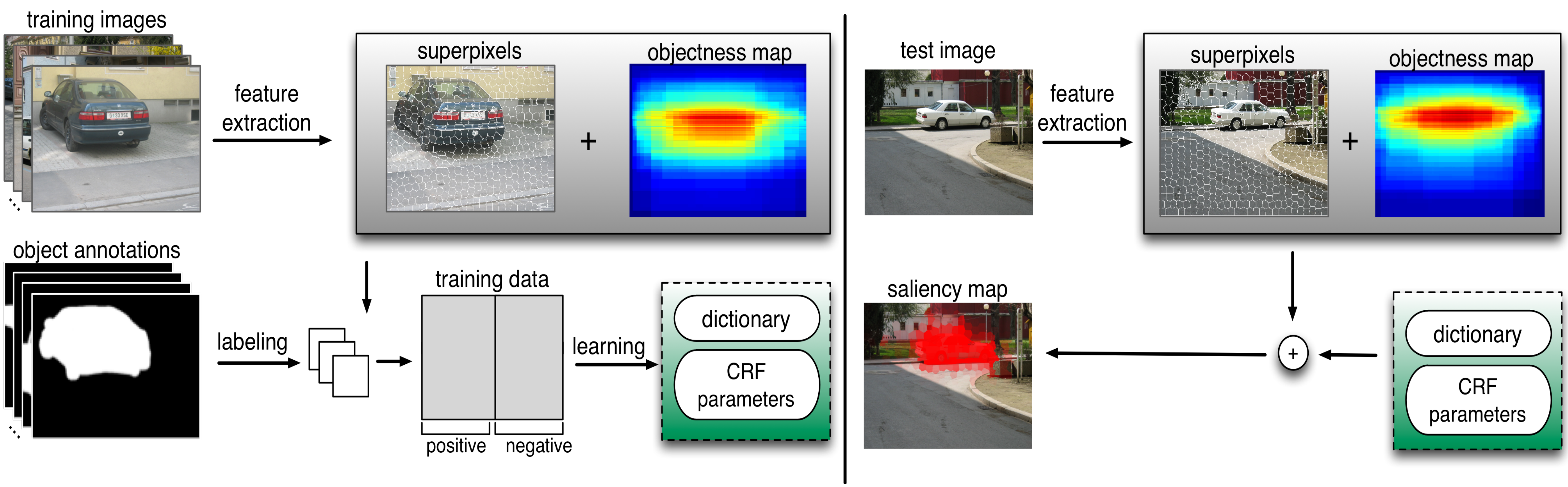
Task-driven fixations

Hayhoe and Ballard, 2009

Top-down saliency estimation

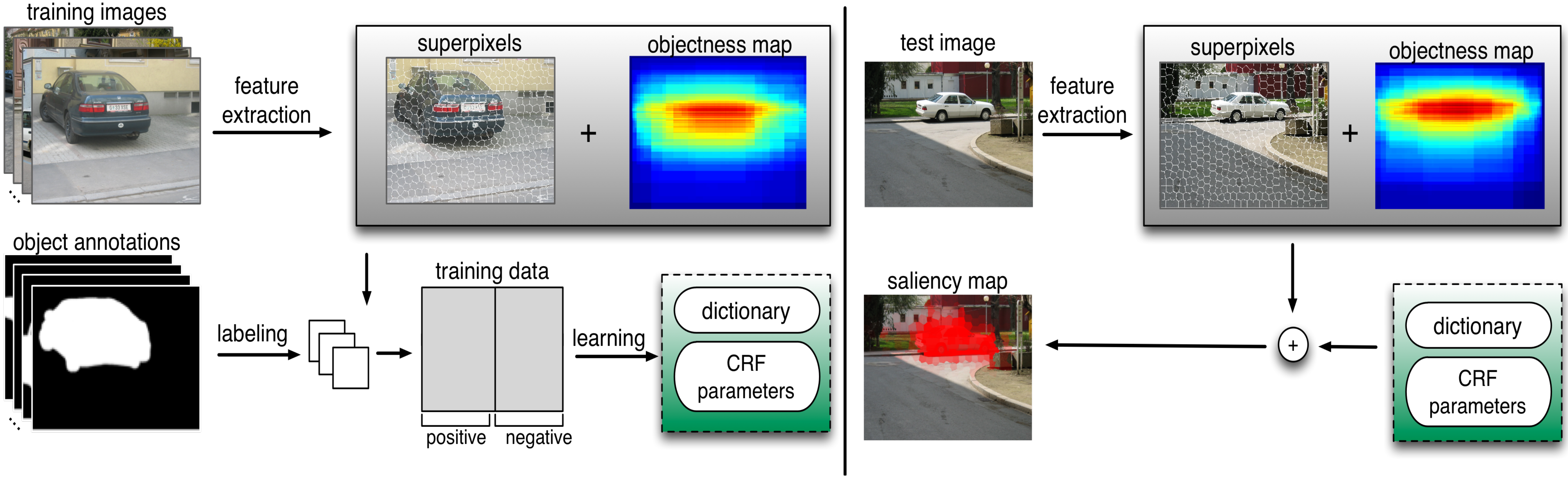
- A. Kocak, K. Cizmeciler, A. Erdem and E. Erdem, Top down saliency estimation via superpixel based discriminative dictionaries, BMVC 2014
- A superpixel-based top-down saliency model via joint discriminative dictionary and CRF learning
- **Task:** Task-driven such as detecting an object instance from a certain category

Top-down saliency estimation



Training:
(1) Segment the images into superpixels and represent them with the sigma points descriptor.
(2) Extract the objectness maps.
(3) Jointly learn the dictionary and the CRF parameters for each object category.

Top-down saliency estimation



Testing:
(1) Segment the images into superpixels and represent them with the sigma points descriptor.
(2) Compute the sparse codes of superpixels with dictionaries learned from data.
(3) Estimate the objectness map.
(4) Use the CRF model to infer the saliency scores.

CRF and dictionary learning

- Construct a CRF model with nodes representing the superpixels and edges describing the connections among them.

$$\log P(\mathbf{Y}|\mathbf{X},\mathbf{D},\theta) = \underbrace{\textstyle\sum_{i\in\mathcal{V}} \psi_i(y_i,\mathbf{x}_i;\mathbf{D},\theta)}_{\text{dictionary potential}} + \underbrace{\textstyle\sum_{i\in\mathcal{V}} \gamma_i(y_i,\mathbf{x}_i;\theta)}_{\text{objectness potential}}$$

$$+ \underbrace{\textstyle\sum_{(i,j)\in\mathcal{E}} \phi_{i,j}(y_i,y_j,\mathbf{x}_i,\mathbf{x}_j;\theta)}_{\text{edge potential}} - \log Z(\theta,\mathbf{D})$$

CRF and dictionary learning

$$\log P(\mathbf{Y}|\mathbf{X},\mathbf{D},\theta) = \underbrace{\textstyle\sum_{i\in\mathcal{V}} \psi_i(y_i,\mathbf{x}_i;\mathbf{D},\theta)}_{\text{dictionary potential}} + \underbrace{\textstyle\sum_{i\in\mathcal{V}} \gamma_i(y_i,\mathbf{x}_i;\theta)}_{\text{objectness potential}}$$
$$+ \underbrace{\textstyle\sum_{(i,j)\in\mathcal{E}} \phi_{i,j}(y_i,y_j,\mathbf{x}_i,\mathbf{x}_j;\theta)}_{\text{edge potential}} - \log Z(\theta,\mathbf{D})$$

- **Dictionary potential:** Use a sparse codes-based linear classifier as a unary potential.

$$\psi_i(y_i,\mathbf{x}_i;\mathbf{D},\theta) = -y_i\mathbf{w}^T\boldsymbol{\alpha}_i$$

$$\boldsymbol{\alpha}_i(\mathbf{x}_i,\mathbf{D}) = \arg\min_{\boldsymbol{\alpha}} \frac{1}{2}\|\mathbf{x}_i - \mathbf{D}\boldsymbol{\alpha}\|^2 + \lambda\|\boldsymbol{\alpha}\|_1$$

CRF and dictionary learning

$$\log P(\mathbf{Y}|\mathbf{X},\mathbf{D},\theta) = \underbrace{\textstyle\sum_{i\in\mathcal{V}} \psi_i(y_i,\mathbf{x}_i;\mathbf{D},\theta)}_{\text{dictionary potential}} + \underbrace{\textstyle\sum_{i\in\mathcal{V}} \gamma_i(y_i,\mathbf{x}_i;\theta)}_{\text{objectness potential}}$$

$$+ \underbrace{\textstyle\sum_{(i,j)\in\mathcal{E}} \phi_{i,j}(y_i,y_j,\mathbf{x}_i,\mathbf{x}_j;\theta)}_{\text{edge potential}} - \log Z(\theta,\mathbf{D})$$

- **Objectness potential:** a class-independent unary potential

$$\gamma_i(y_i,\mathbf{x}_i;\theta) = -\beta y_i\left(2P(obj|\mathbf{x}_i) - 1\right)$$

CRF and dictionary learning

$$\log P(\mathbf{Y}|\mathbf{X},\mathbf{D},\theta) = \underbrace{\sum\nolimits_{i\in\mathcal{V}} \psi_i(y_i,\mathbf{x}_i;\mathbf{D},\theta)}_{\text{dictionary potential}} + \underbrace{\sum\nolimits_{i\in\mathcal{V}} \gamma_i(y_i,\mathbf{x}_i;\theta)}_{\text{objectness potential}}$$
$$+ \underbrace{\sum\nolimits_{(i,j)\in\mathcal{E}} \phi_{i,j}(y_i,y_j,\mathbf{x}_i,\mathbf{x}_j;\theta)}_{\text{edge potential}} - \log Z(\theta,\mathbf{D})$$

- **Edge potential:**

$$\phi_{i,j}(y_i,y_j,\mathbf{x}_i,\mathbf{x}_j;\theta) = \rho\left(1-\delta(y_i-y_j)\right)$$

CRF and dictionary learning

$$\log P(\mathbf{Y}|\mathbf{X},\mathbf{D},\theta) = \underbrace{\textstyle\sum_{i\in\mathcal{V}} \psi_i(y_i,\mathbf{x}_i;\mathbf{D},\theta)}_{\text{dictionary potential}} + \underbrace{\textstyle\sum_{i\in\mathcal{V}} \gamma_i(y_i,\mathbf{x}_i;\theta)}_{\text{objectness potential}}$$
$$+ \underbrace{\textstyle\sum_{(i,j)\in\mathcal{E}} \phi_{i,j}(y_i,y_j,\mathbf{x}_i,\mathbf{x}_j;\theta)}_{\text{edge potential}} - \log Z(\theta,\mathbf{D})$$

- **Learning:** Simultaneously learn the CRF parameters θ and the dictionary **D** by optimizing:

$$(\mathbf{D}^*,\theta^*) = \arg\max_{\mathbf{D},\theta} \prod_{m=1}^{M} P(\mathbf{Y}^{(m)}|\mathbf{X}^{(m)},\mathbf{D},\theta)$$

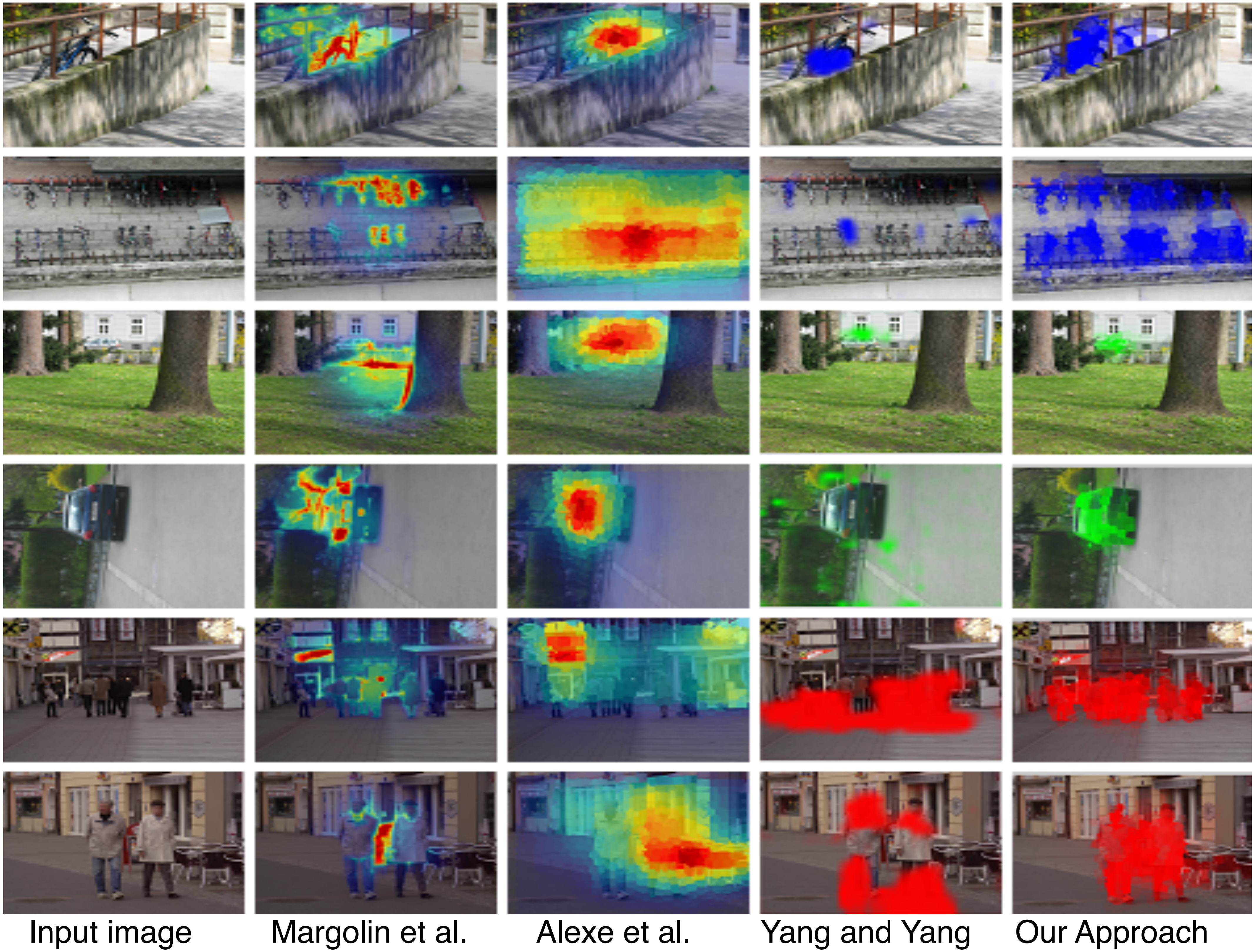
Quantitative analysis

EER results on the Graz-02 dataset

	Bike	Car	People
Margolin *et al*. (2013)	25.6	16.9	17.4
Perazzi *et al*. (2012)	11.4	13.8	14.3
Yang and Zhang (2013)	14.8	13.7	14.9
Objectness (Alexe et al., 2010)	53.5	48.3	43.5
Aldavert *et al*. (2010)	71.9	64.9	58.6
Khan and Tappen (2013)	72.1	-	-
Marszalek and Schmid (2012)	61.8	53.8	44.1
Yang and Yang (2012)	62.4	60.0	62.0
Our approach (setting 1)	71.9	61.9	65.5
Our approach (setting 2)	71.7	62.0	64.9
Our approach (setting 3)	**73.9**	**68.4**	**68.2**

Qualitative analysis

Saliency maps on the Graz-02 dataset

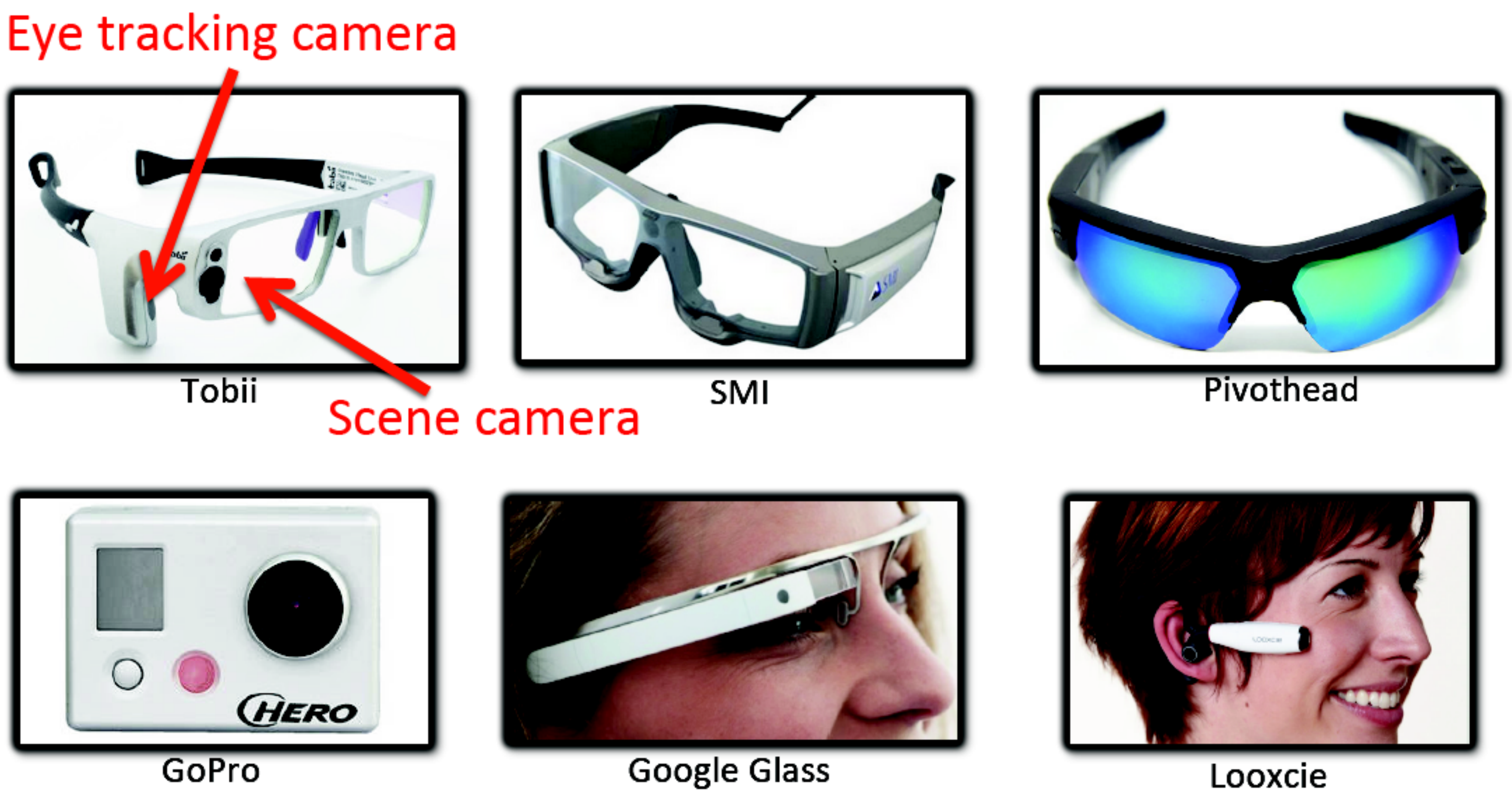


Main insights from natural tasks

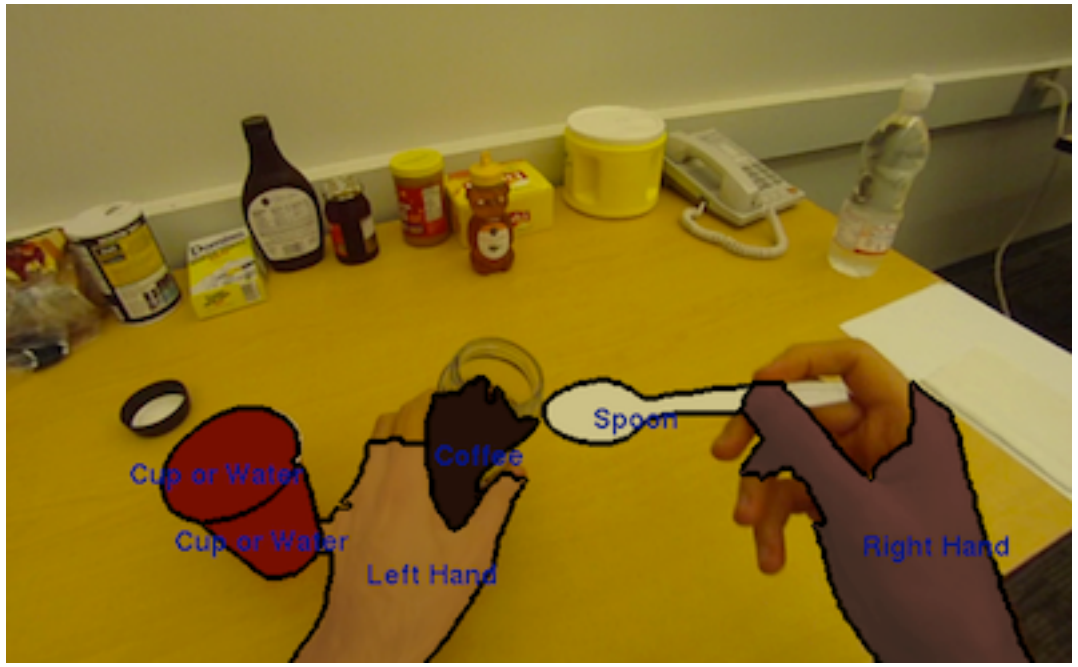
- Vision is **active** not passive.
  - Specific information is usually acquired at the fixation point.
  - Information is acquired “just-in-time”.
- Fixations patterns reflect learning at several levels:
  - what objects are relevant
  - where information is located
  - order of sub-tasks/properties of world.
- Fixations tightly linked to actions.

Developments in eye tracking

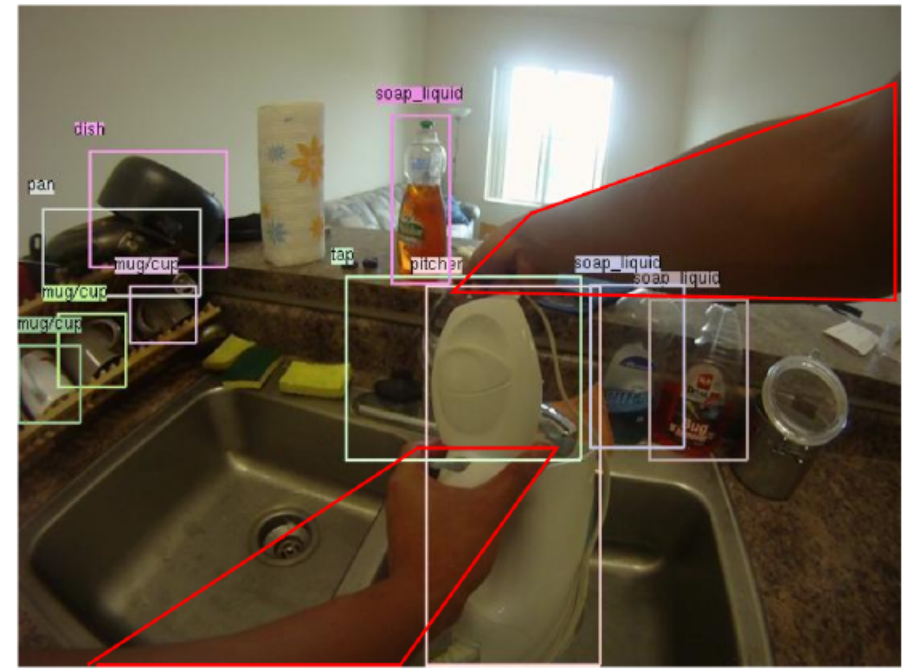
- <u>Head free:</u>
  - Head mounted IR video-based systems
  - Remote systems with head tracking!
  - Scene camera



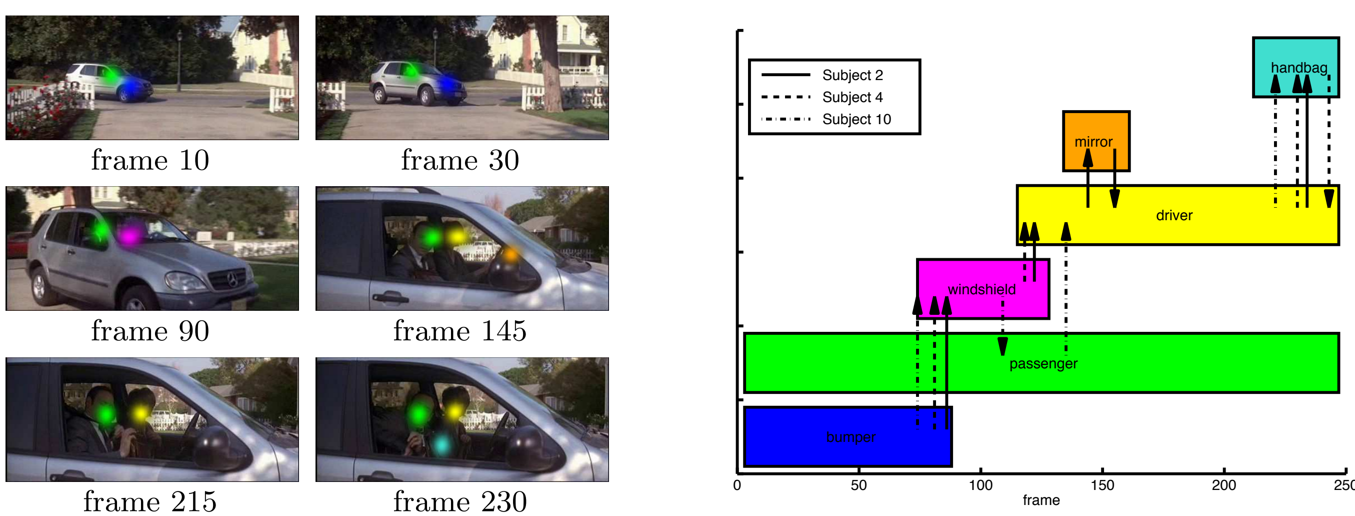
Ego Centric Vision a.k.a First person vision
(Lucas Kanade)



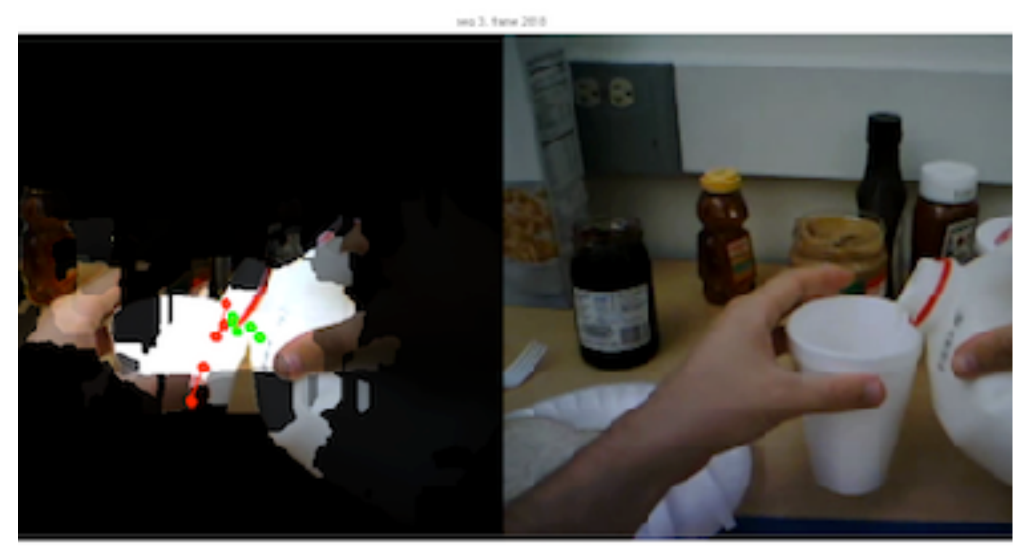
Fathi et al., CVPR 2011



Pirsiavash and Ramanan, CVPR 2012



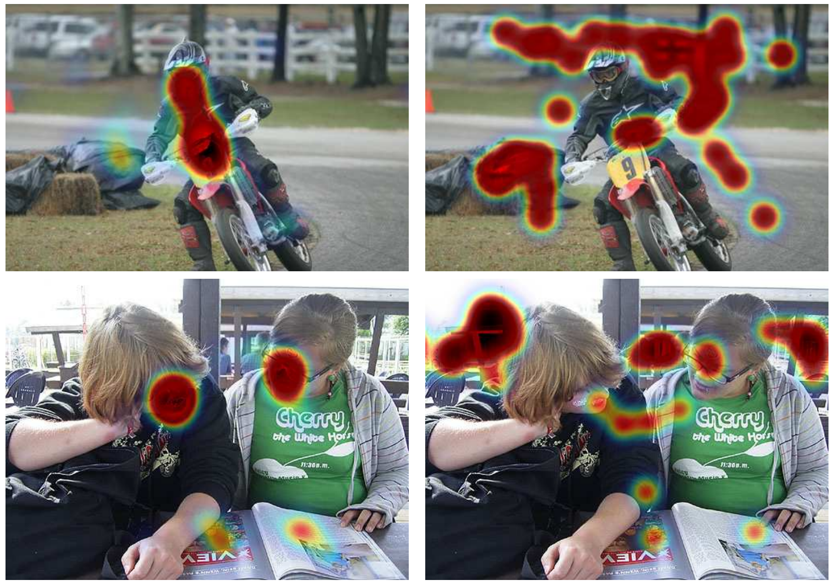
Mathe and Sminchisescu, ECCV 2012



Fathi et al., ECCV 2012

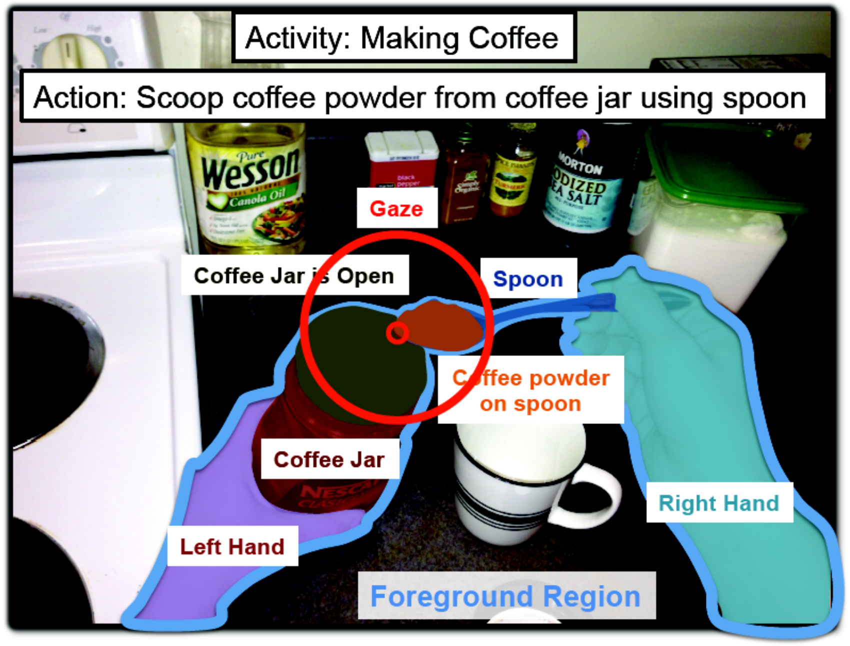
Ego Centric Vision a.k.a First person vision

(Lucas Kanade)

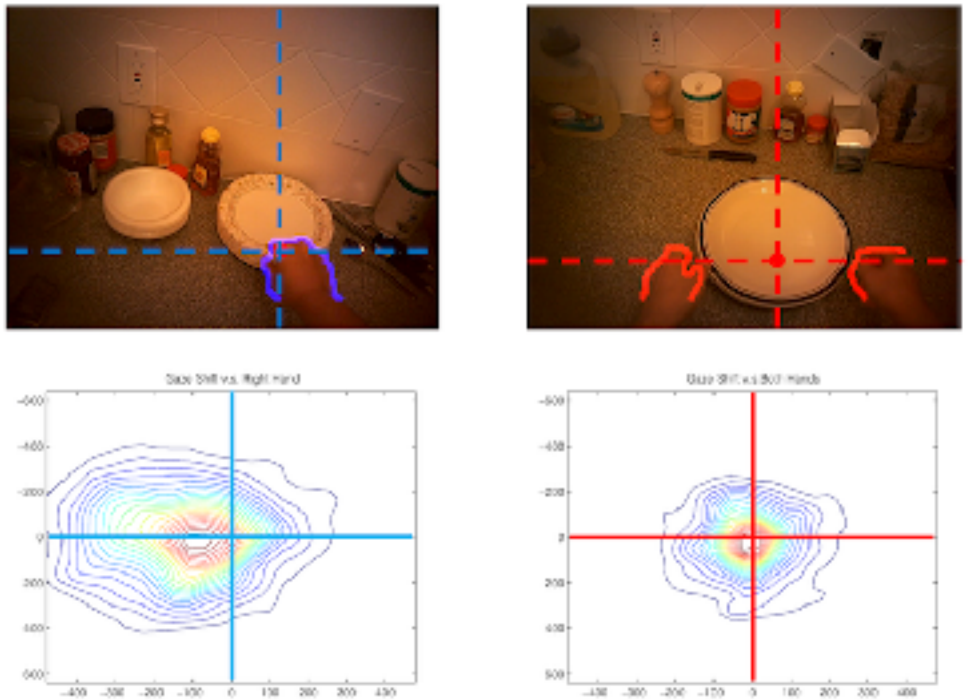


Mathe and Sminchisescu, NIPS 2013

Shapovalova et al., NIPS 2013



Fathi and Rehg, CVPR 2013



Li et al., ICCV 2013

What is the current state-of-the-art?

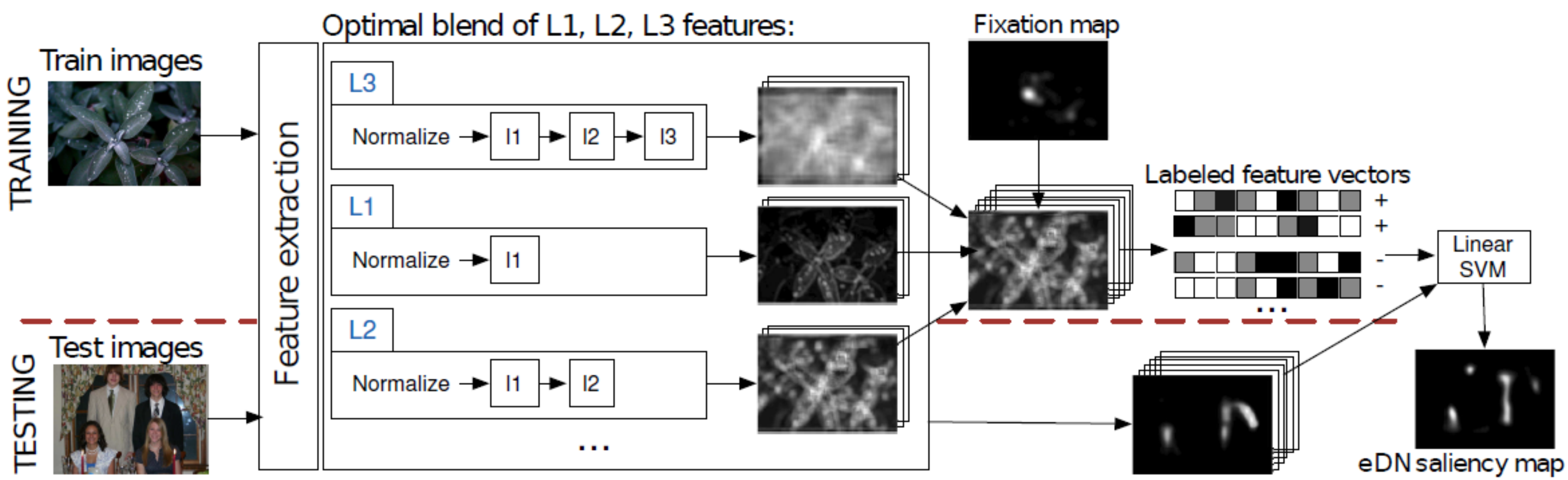
- Hierarchical processing is ubiquitous in low-level human vision.
- Deep unsupervised models have been present for over a decade.
- Nowadays, go deep and use supervision!
- Mimic human visual system and learn a saliency model in an end-to-end manner.

Deep supervised models

- Typically, superior performance to unsupervised models
- Large-scale proxy datasets have enabled effective supervised learning
- Key considerations:
  - Network architecture
  - Incorporation of prior cues
  - Supervision mechanism
  - Loss function

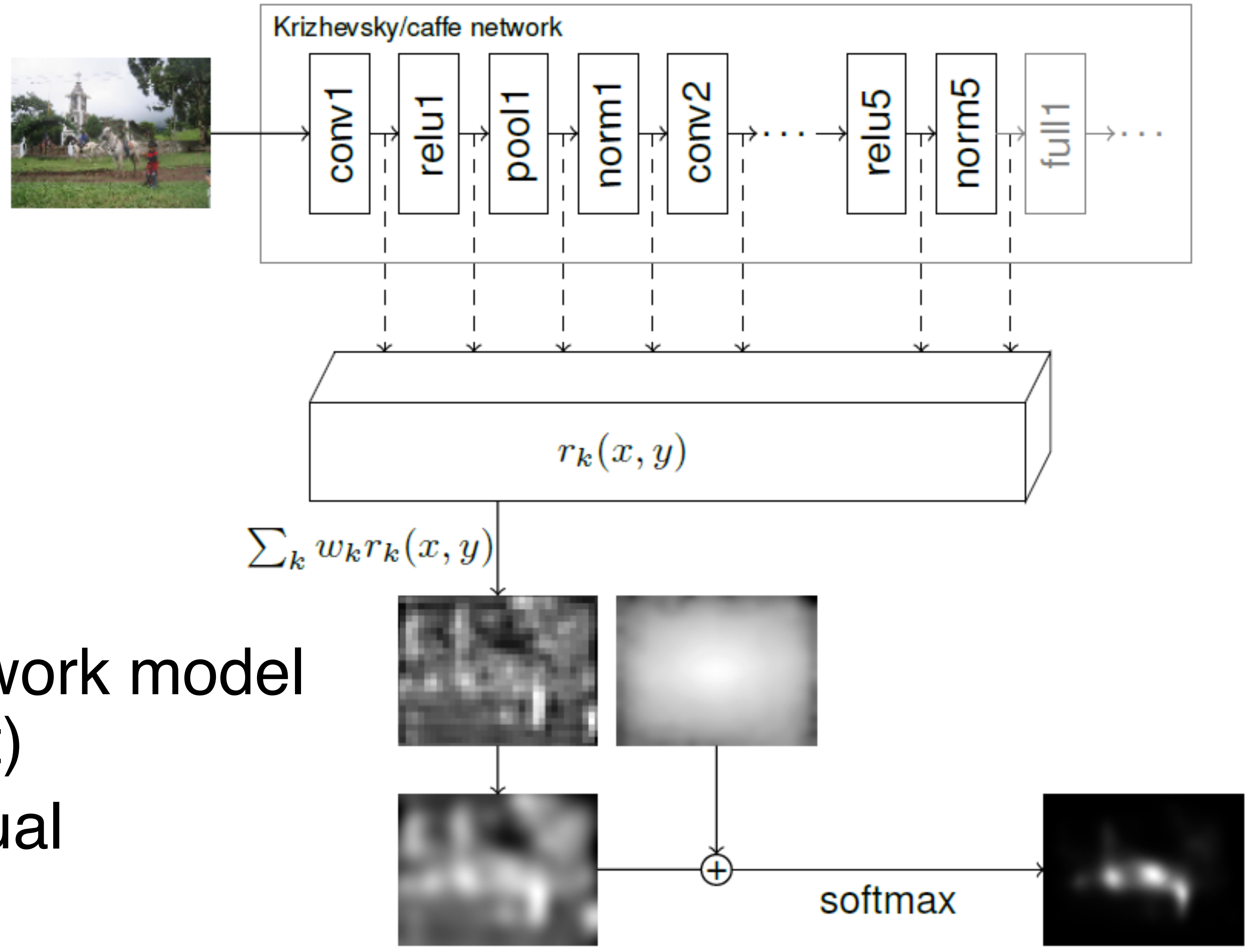
adopted from N. Murray

eDN Model (Vig et al., 2014)



- 1-3 layer networks
- Up to 43 hyper-parameters
- Linear patch classifier is learned
- fixated and non-fixated regions used to supervise training
- Small-scale dataset used for training
- Filters are drawn randomly

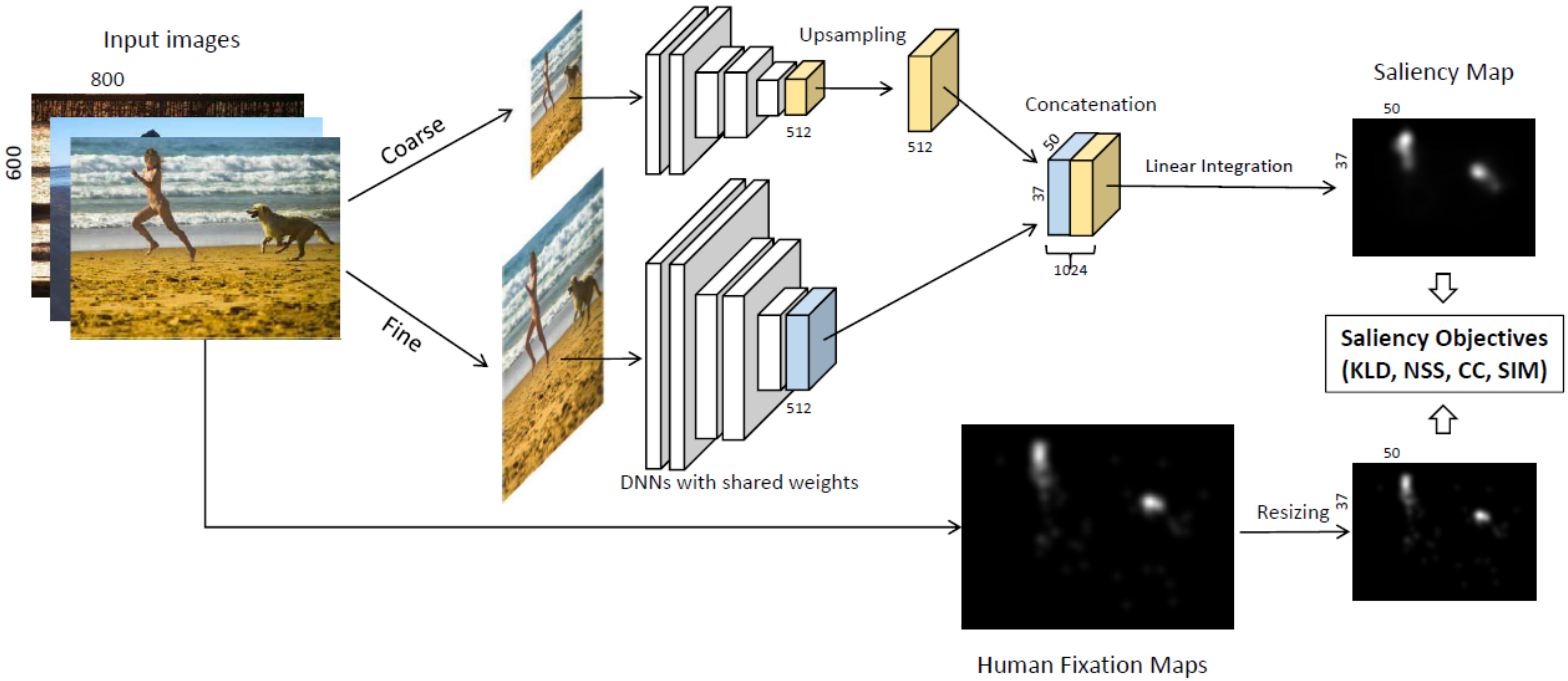
Deep Gaze (Kummerer et al., 2015)



- Convolutional network model (based on AlexNet)
- pre-trained for visual recognition task
- Incorporation of centre-bias prior

adopted from N. Murray

SALICON Model (Huang et al., 2015)

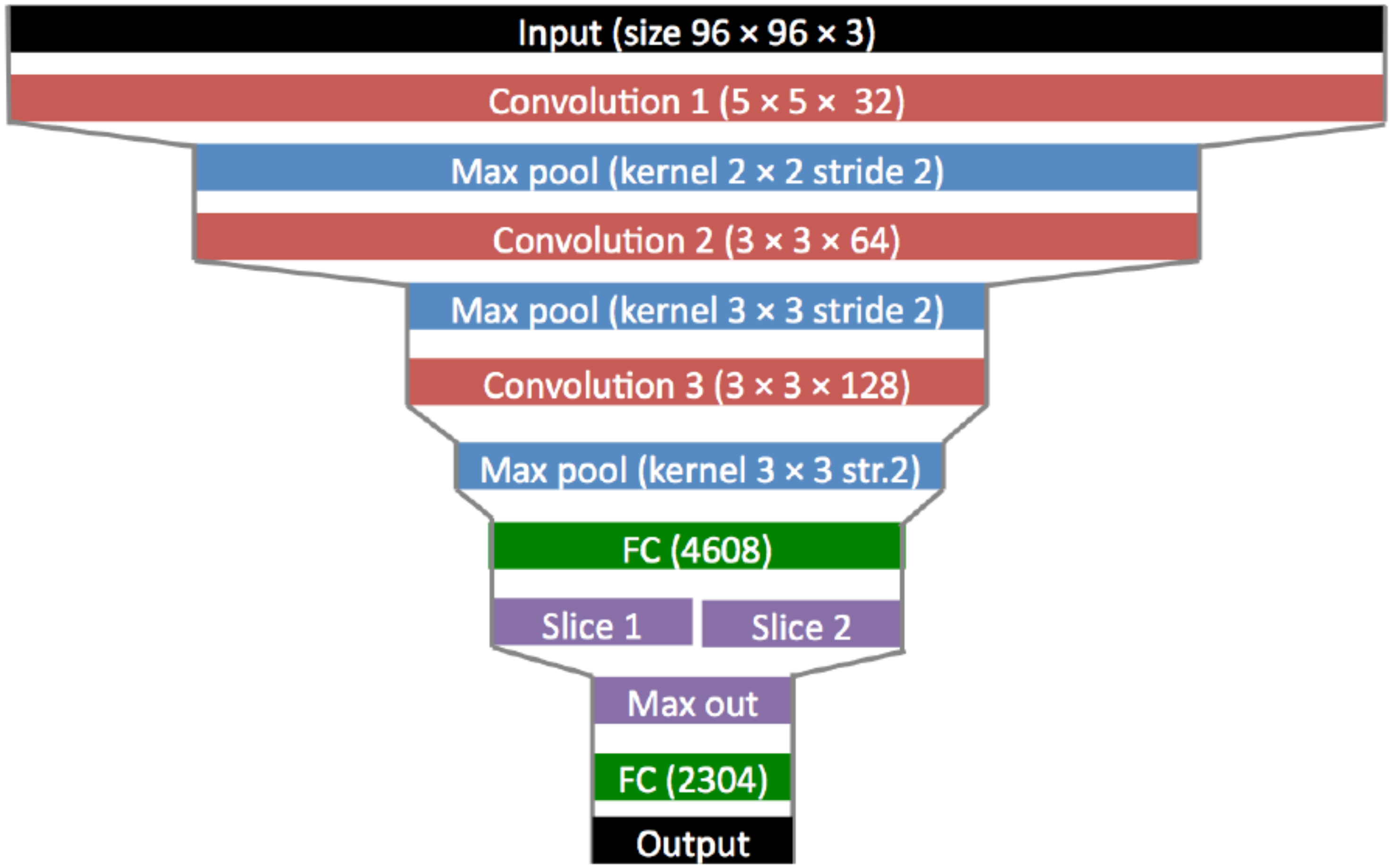


- Domain adaptation to saliency works
- Adding multi-scale information helps

adopted from N. Murray

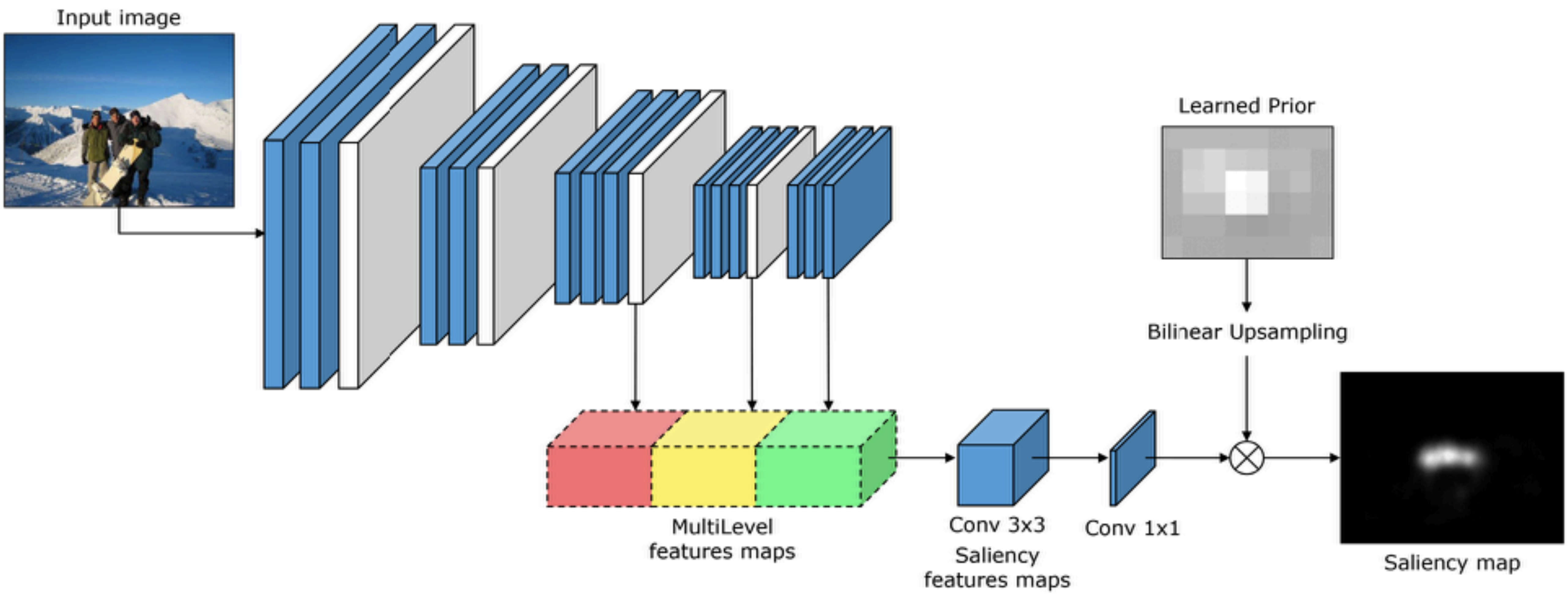
DeepSal Model (Pan et al., 2016)

- New large-scale datasets with proxy eye-fixation data
- Training all features of larger networks
- Still small-scale compared to networks designed for semantics prediction



adopted from N. Murray

ML-Net Model (Cornia et al., 2016)



- Saliency map priors
- Multiple resolutions

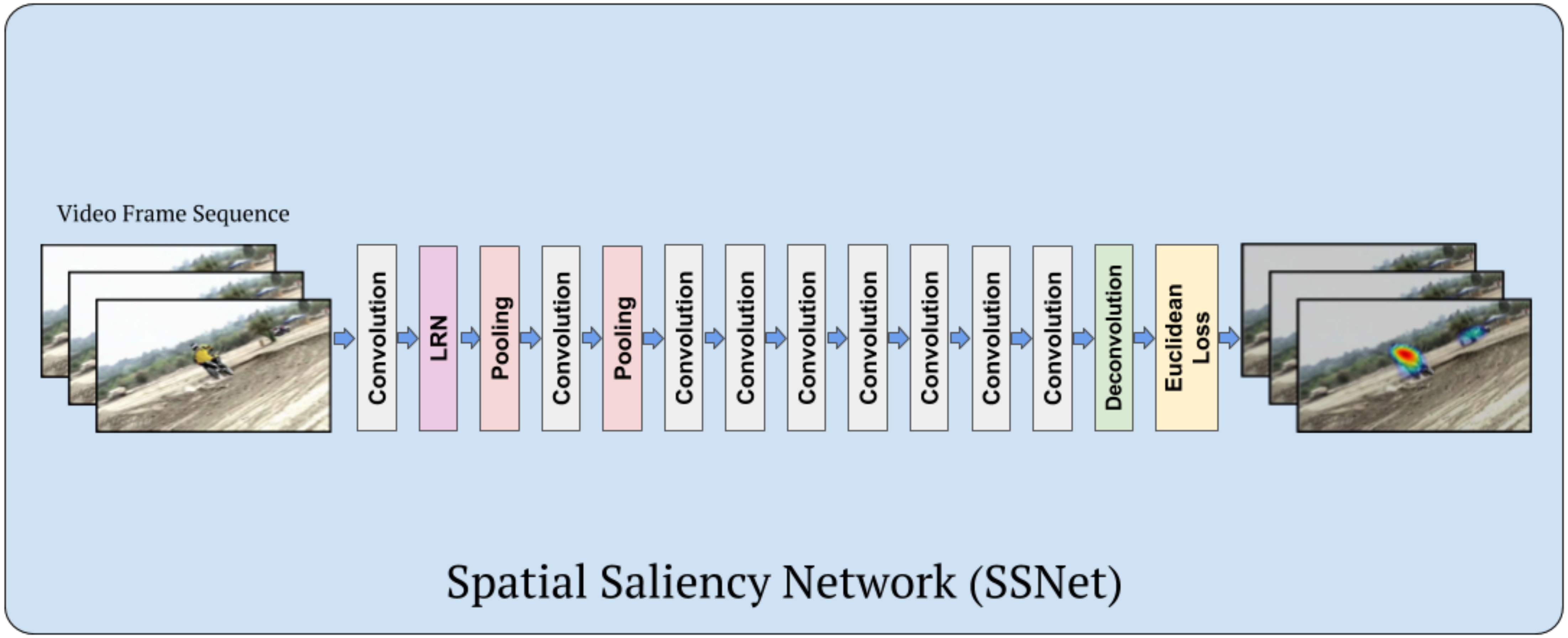
adopted from N. Murray

Spatio-Temporal Saliency Networks (Bak et al., 2018)

- Two-stream CNNs for saliency prediction from videos
  - One of the first deep models for dynamic saliency prediction
- Element-wise and convolutional fusion strategies to integrate spatial and temporal information.
- Experiments on
  - DIEM (Mital et al., 2011) : 84 videos, fixations from 50 subjects
  - UCF-Sports (Mathe and Sminchisescu, 2015) : 150 videos, fixations from 16 subjects

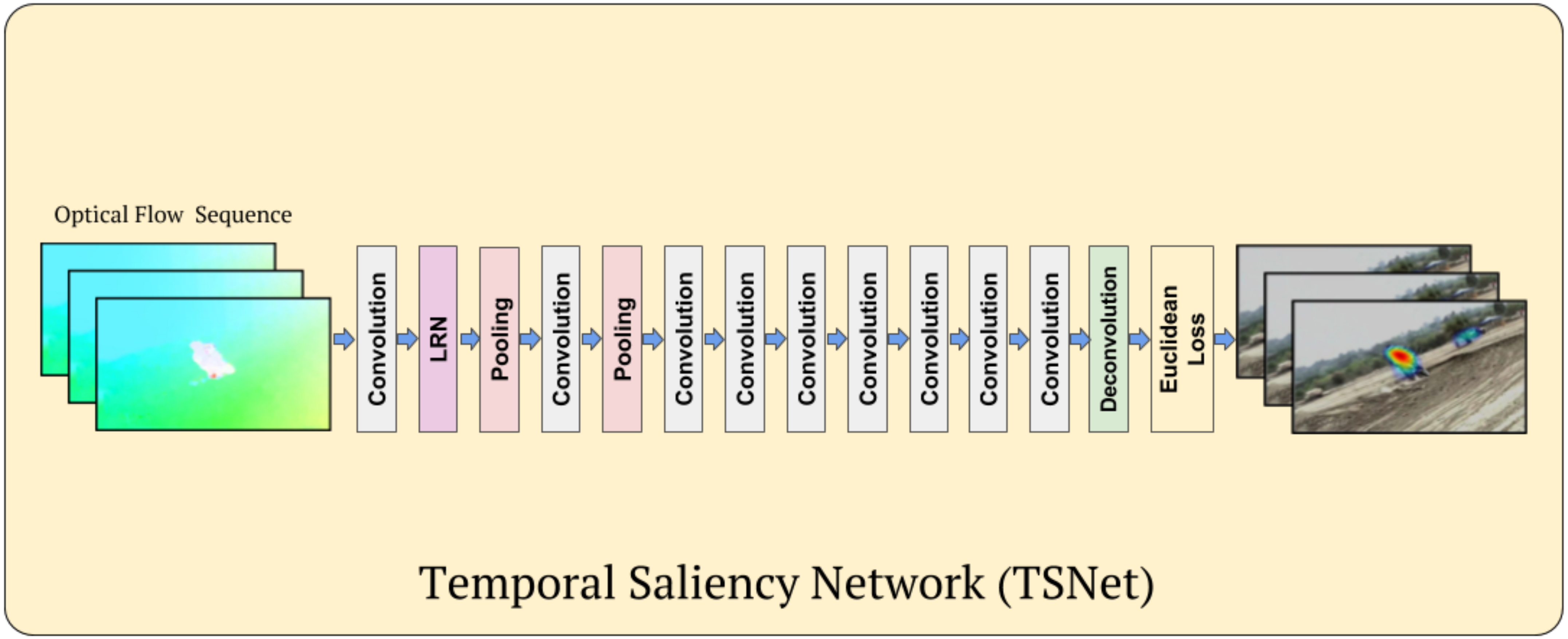
Spatio-Temporal Saliency Networks (Bak et al., 2018)

- Two base network models
- 9 convolution + 1 deconvolution layers
- 25.8M parameters
- Appearance stream



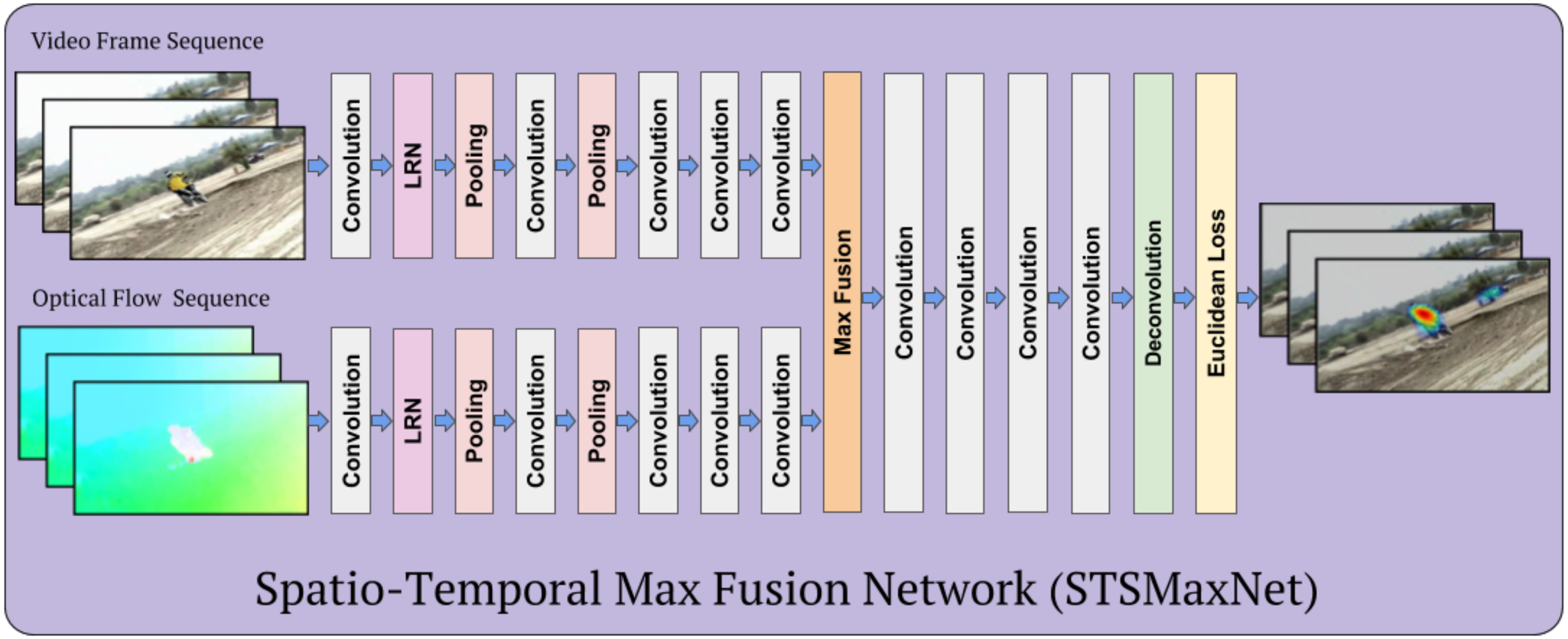
Spatio-Temporal Saliency Networks (Bak et al., 2018)

- Two base network models
- 9 convolution + 1 deconvolution layers
- 25.8M parameters
- Motion stream



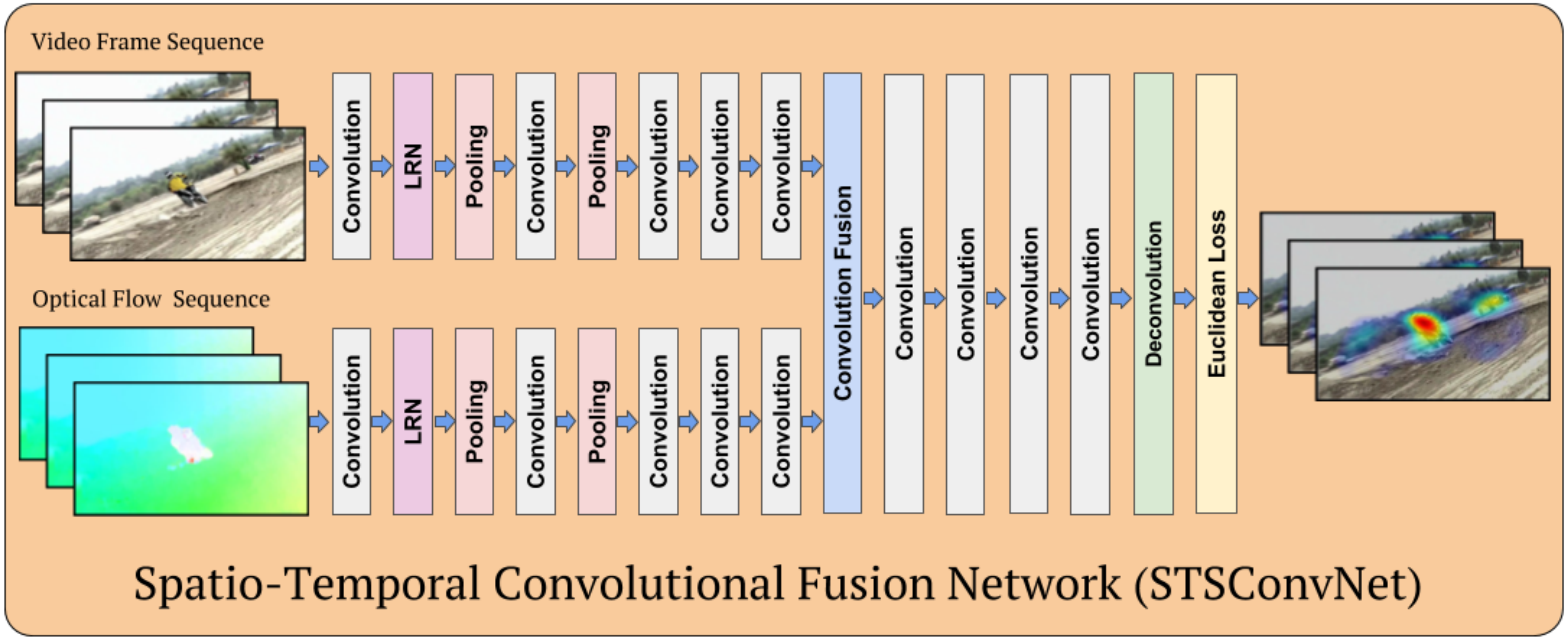
Spatio-Temporal Saliency Networks (Bak et al., 2018)

- STSMaxNet
- Element-wise max fusion
- 34.7M parameters

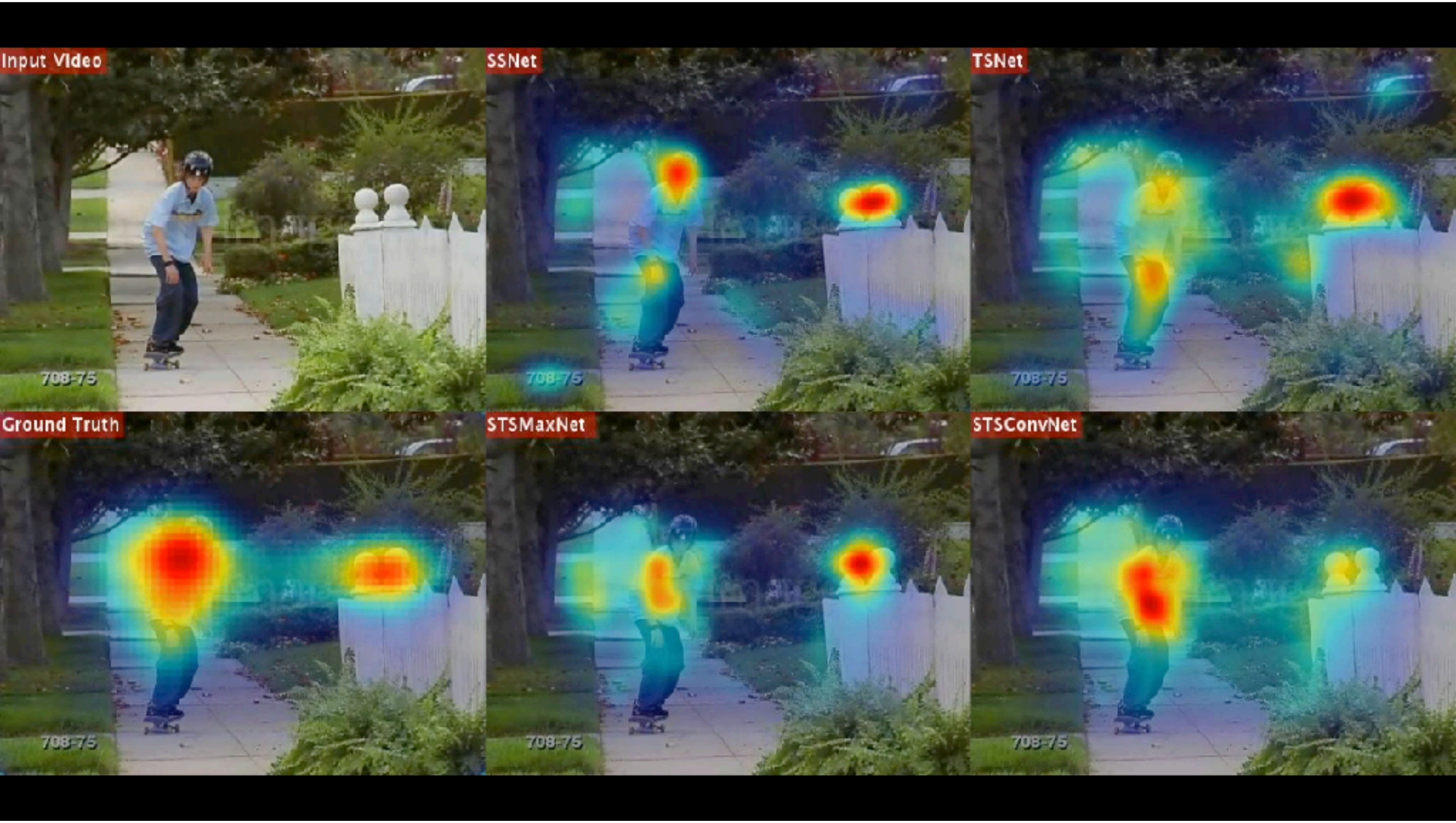


Spatio-Temporal Saliency Networks (Bak et al., 2018)

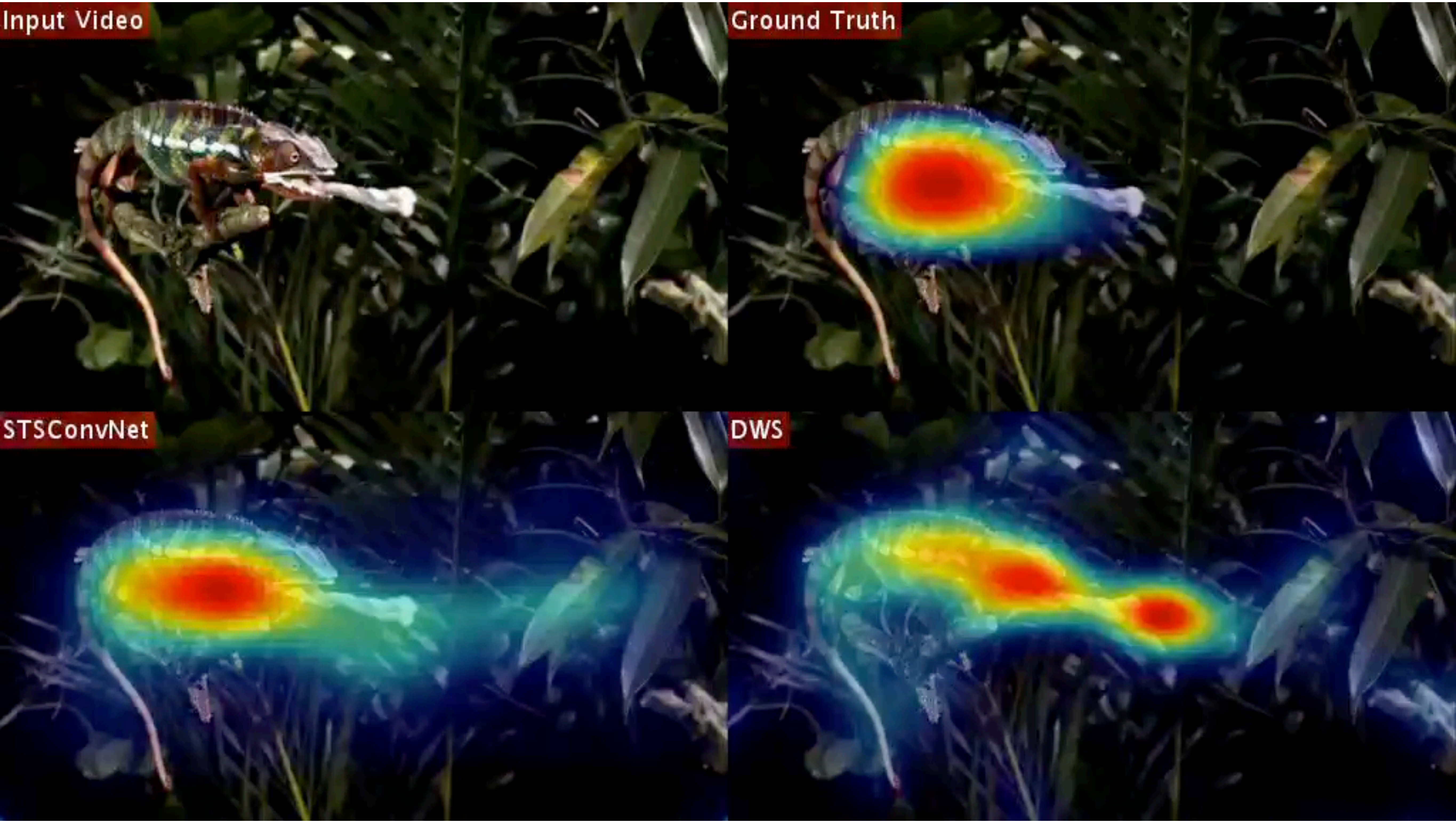
- STSConvNet
- Convolution fusion
- 51.6M parameters



Spatio-Temporal Saliency Networks (Bak et al., 2018)



Spatio-Temporal Saliency Networks (Bak et al., 2018)



Thanks for your **attention!**