Deep networks for saliency map prediction

Naila Murray

Xerox Research Centre Europe

ECCV 2016 Tutorial: New directions in saliency research: developments in architecture, datasets and evaluation

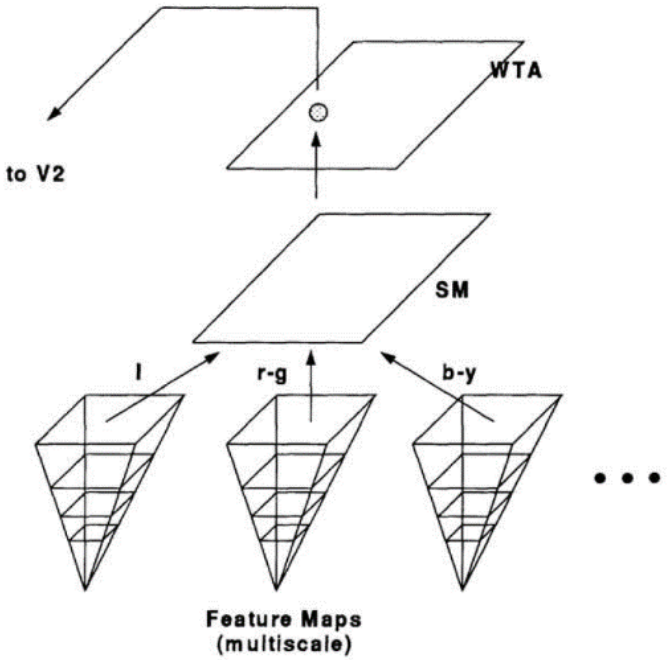
8th October, 2016

Selective attention & the saliency map

"In order to assess the global, overall conspicuity of a location, we will assume the existence of another topographical map, termed the *saliency map*, which combine the information of the individual maps into one global measure of conspicuity."[1]

Hypothesis: saliency map is sequentially scanned by attention

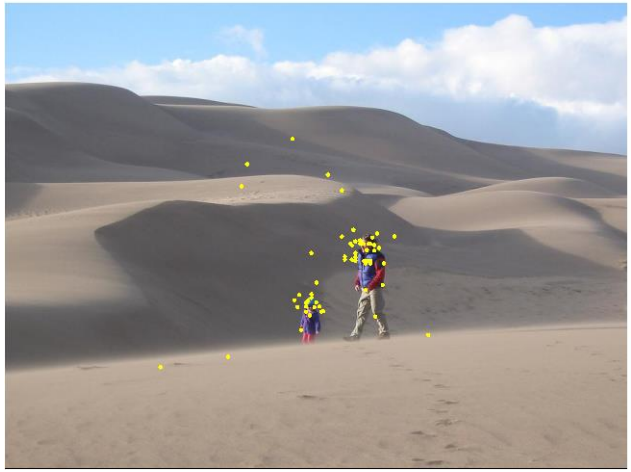
"Where" but not "what" [2]



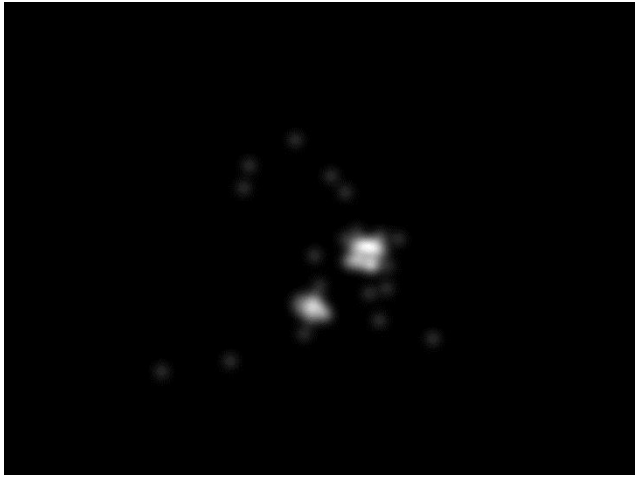
[1] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*,1985.
[2] E. Niebur & C. Koch. Control of selective visual attention: Modeling the" where" pathway. NIPS, 1995.

Saliency map prediction

Computational models aim to produce an output saliency map given an input image:



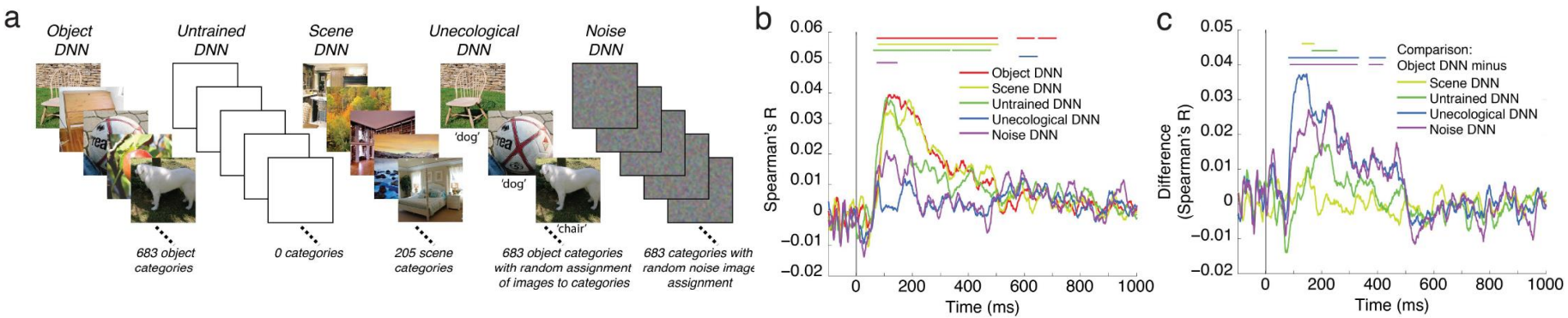
input image



saliency map

Why go deep?

- Hierarchical processing is ubiquitous in low-level human vision [1,2]
- Excellent performance on saliency map prediction task



[1] D. H. Hubel., and T. N. Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *The Journal of physiology*,1962.

[2] RM Cichy, A Khosla, D Pantazis, A Torralba, and A. Oliva. Comparison of deep neural networks to spatio-temporal cortical dynamics of human visual object recognition reveals hierarchical correspondence. Scientific reports, 2016.

Outline

- Deep unsupervised models
- Deep supervised models
- Conclusions and future directions

Deep unsupervised models

Deep unsupervised models

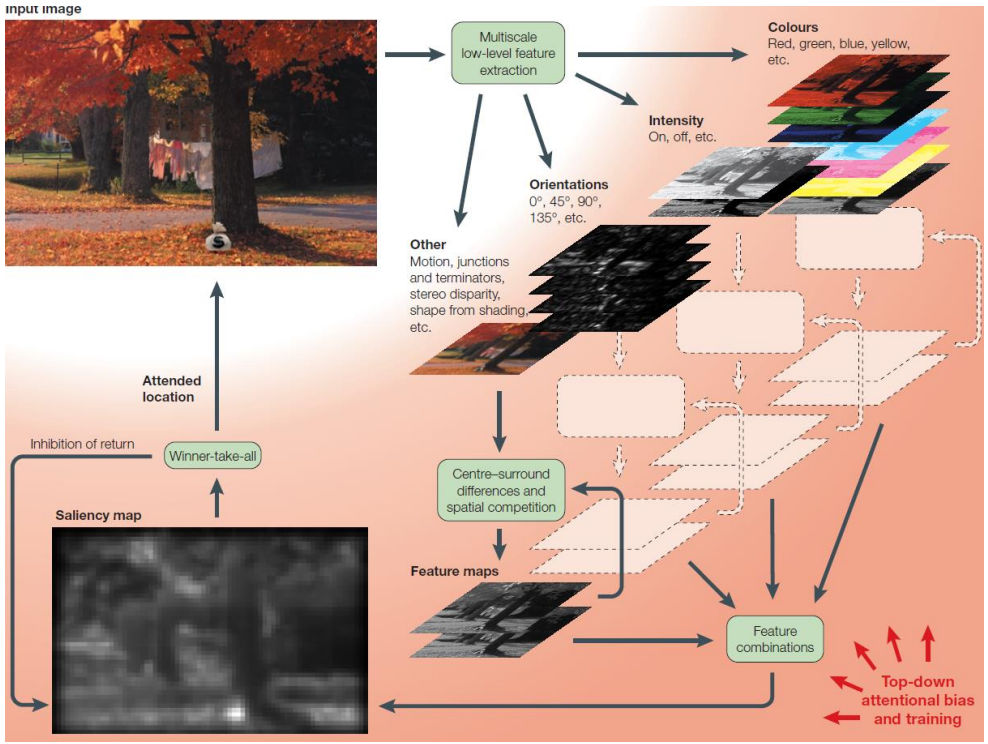
Key considerations:

- Network architecture
- Incorporation of prior cues
- Loss function

Deep unsupervised models

Classical model[1] :

- Inspired by feature-integration theory[2]
- Filters for feature map generation are engineered by hand



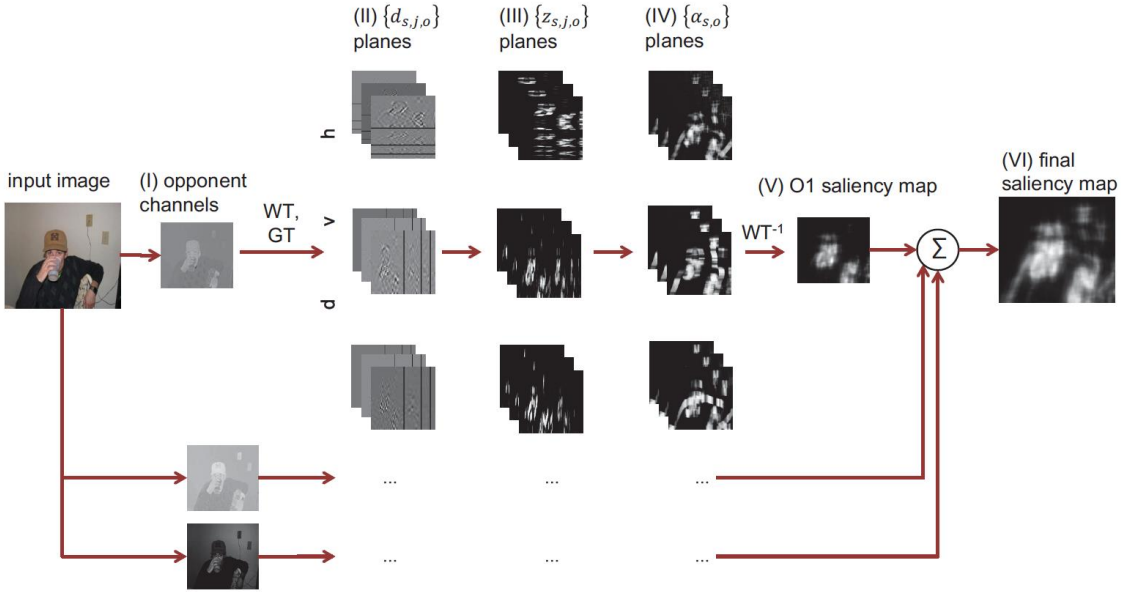
[1] Itti, Laurent, and Christof Koch. "Computational modelling of visual attention." *Nature reviews neuroscience* 2.3 (2001): 194-203.

[2] A. M. Treisman & G. Gelade. A feature-integration theory of attention.*Cognitive psychology, 1980.*

Deep unsupervised models

Wavelet-based model:

- Convolutional filters
- Normalization and non-linearities between each layer
- Feature maps are combined using inverse wavelet transform



N. Murray, M. Vanrell, X. Otazu, C. A. Parraga. Low-level spatio-chromatic grouping for saliency estimation. TPAMI, 2013.

Deep supervised models

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

Deep supervised models

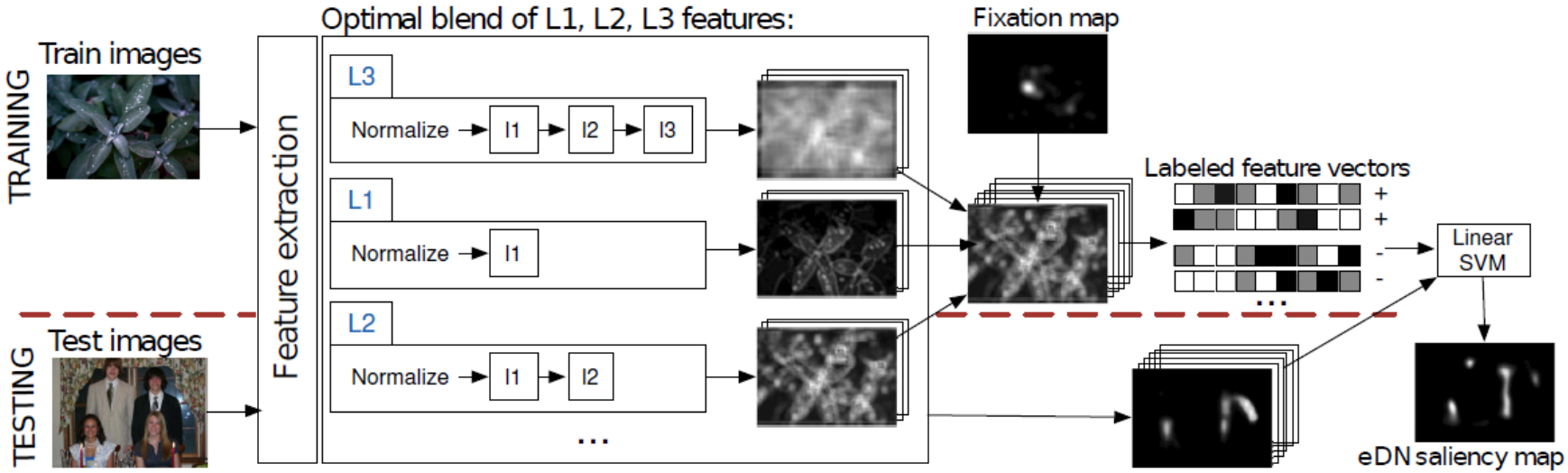
Ensemble of Deep Networks (eDN):

- 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*

E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. CVPR, 2014.
T. Judd, K. Ehinger, F. Durand, & A. Torralba. Learning to predict where humans look. ICCV, 2009.

Deep supervised models

eDN model:

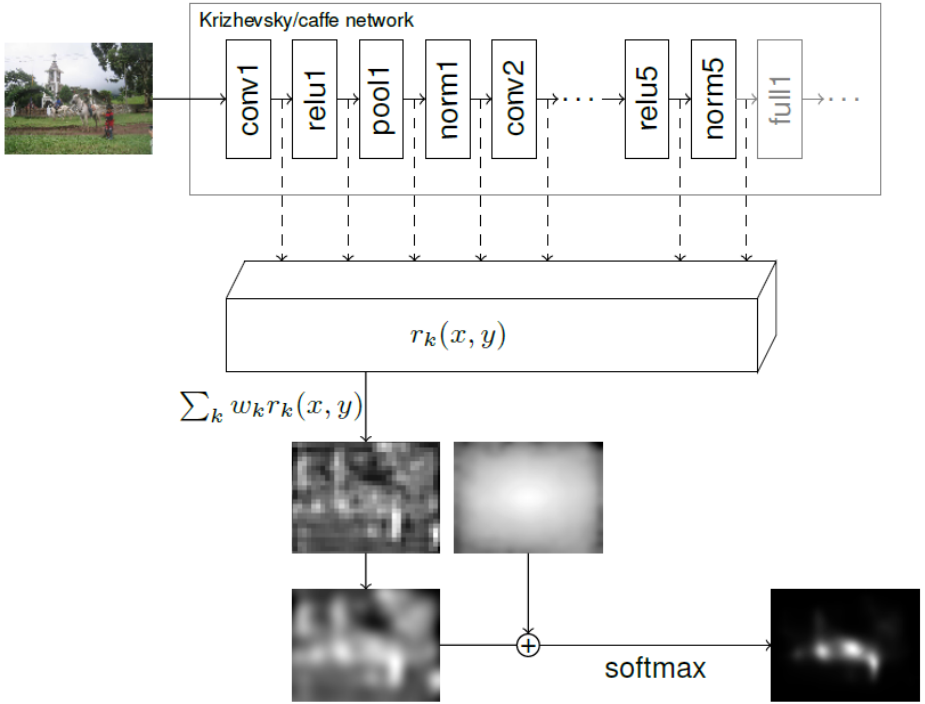


E. Vig, M. Dorr, and D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. CVPR, 2014.

Deep supervised models

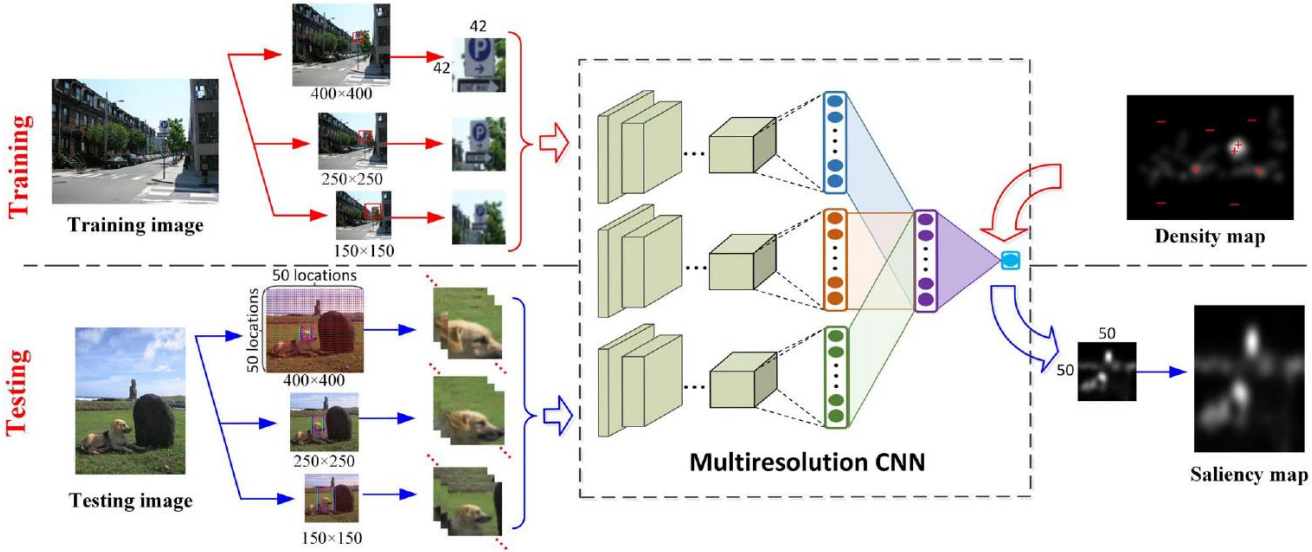
Convolutional network model:

- pre-trained for visual recognition task
- Incorporation of centre-bias prior



[M. Kümmerer, L. Theis, and M. Bethge. Deep Gaze I: Boosting Saliency Prediction with Feature Maps Trained on ImageNet. ICLR Workshop, 2015]

Deep supervised models



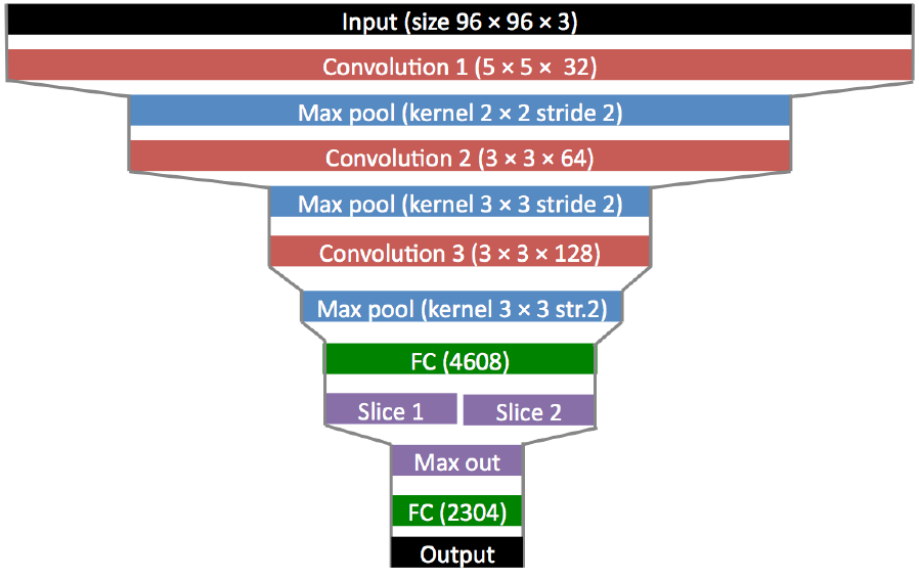
- Sample fixated and non-fixated patches
- Train end-to-end binary classifier
- At testing time, composite maps from local regions to construct global map

N. Liu, J. Han, D. Zhang, S. Wen and T. Liu. Predicting Eye Fixations using Convolutional Neural Networks. CVPR, 2015.

Deep supervised models

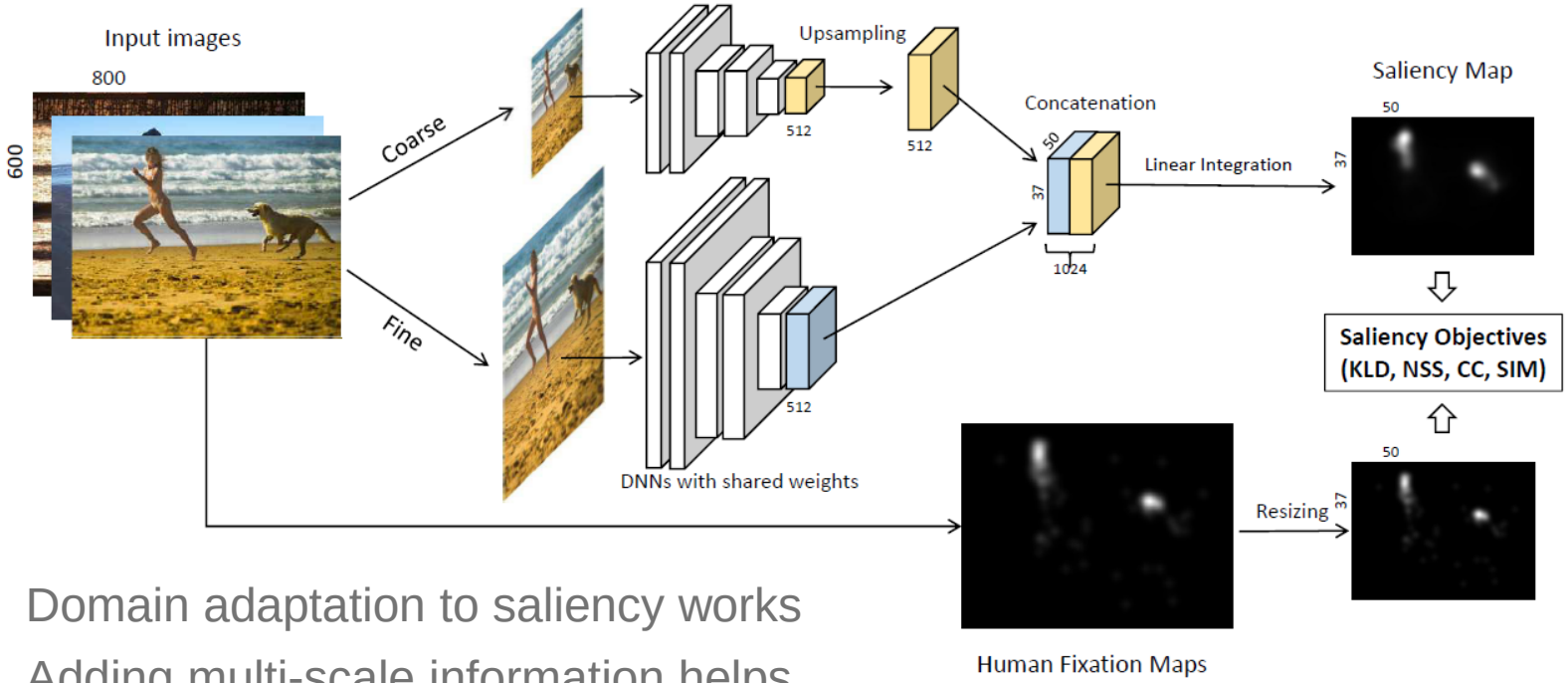
New large-scale datasets with proxy eye-fixation data
→ Training all features of larger networks

Still small-scale compared to networks designed for semantics prediction



J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, N. E. O'Connor. Shallow and Deep Convolutional Networks for Saliency Prediction. CVPR, 2016.

Deep supervised models

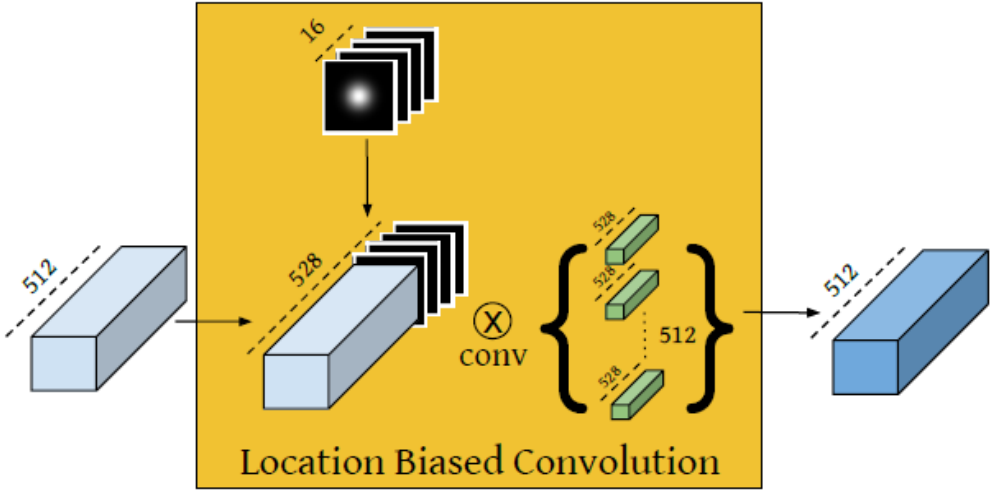


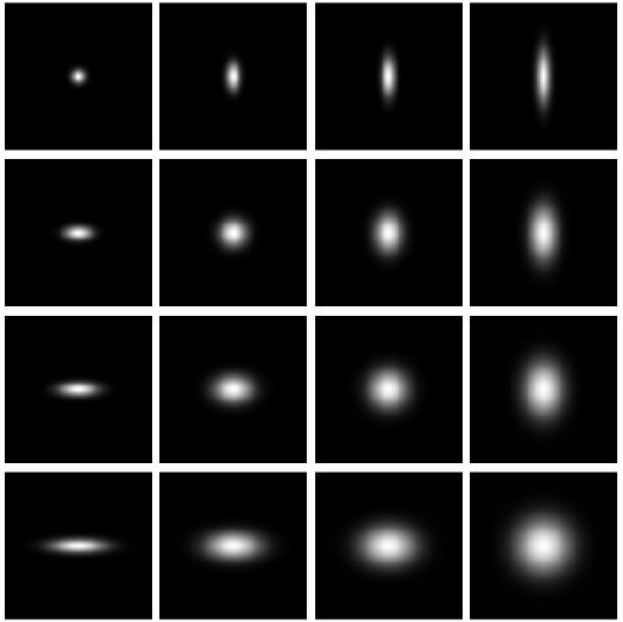
- Domain adaptation to saliency works
- Adding multi-scale information helps

[X. Huang, C. Shen, X. Boix, Q. Zhao. SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. ICCV, 2015.]

Deep supervised models

- Saliency map priors
- Increased resolution: dilation layers

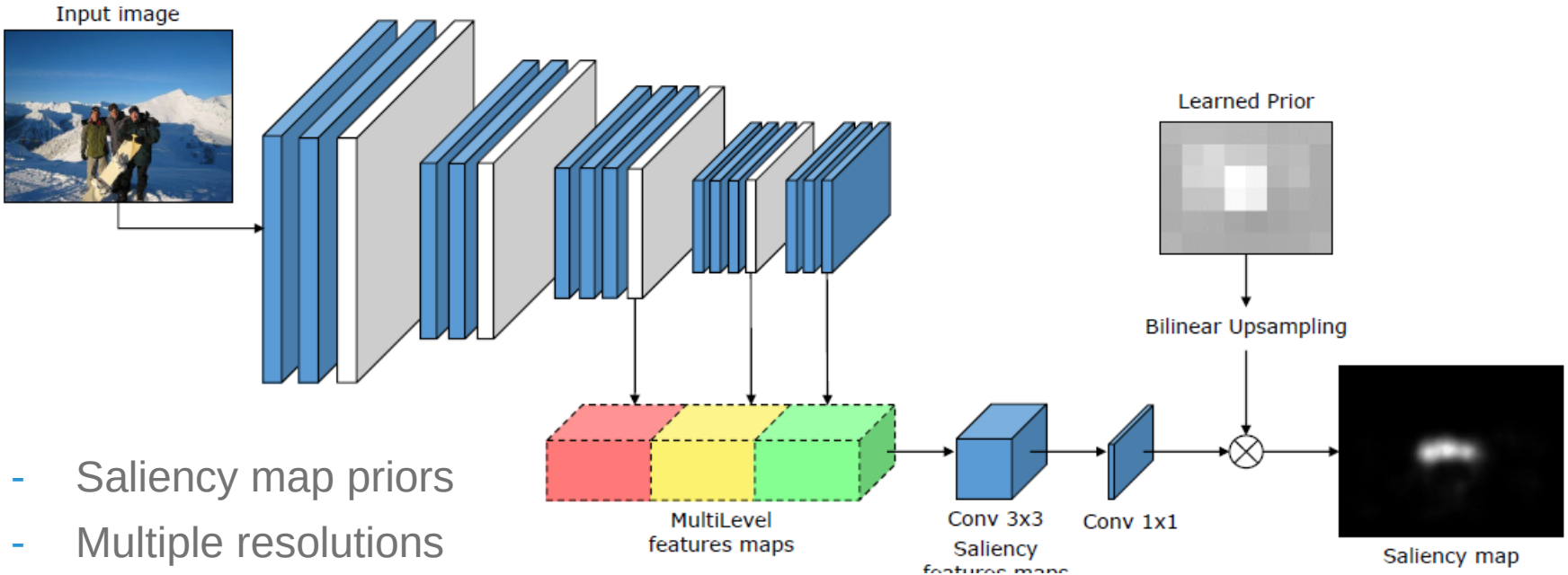




Gaussian priors

[S. Kruthiventi, K. Ayush, and R. Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. arXiv, 2015.]

Deep supervised models



- Saliency map priors
- Multiple resolutions

M. Cornia, L. Baraldi, G. Serra, R. Cucchiara. A Deep Multi-Level Network for Saliency Prediction. ICPR, 2016.

Deep supervised models

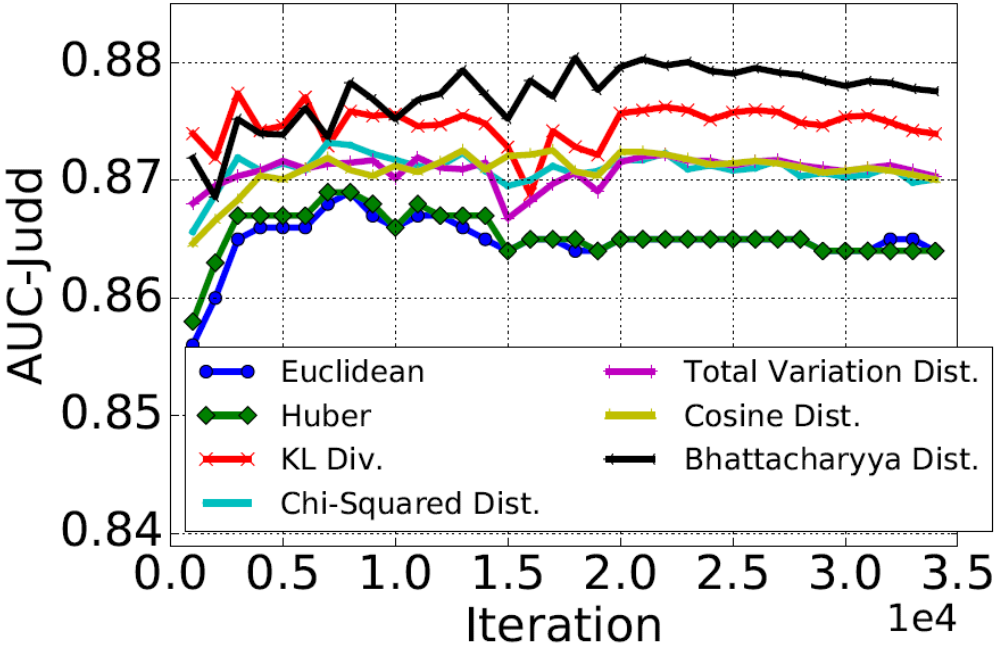
Dense prediction problem - which loss functions to use?

- Euclidean / Huber loss
- Losses based on probability distance measures:

Probability distances	$L(\boldsymbol{p}, \boldsymbol{g})$	$\frac{\partial L(\boldsymbol{p},\boldsymbol{g})}{\partial x_i^p}$
χ^2 divergence	$\sum_j \frac{(g_j)^2}{p_j} - 1$	$p_i \sum_{j \neq i} \frac{g_j^2}{p_j} - \frac{g_i^2}{p_i}(1 - p_i)$
Total Variation distance	$\frac{1}{2} \sum_j \lvert g_j - p_j \rvert$	$\frac{1}{2} \left[p_i \sum_{j \neq i} \frac{g_j - p_j}{\lvert g_j - p_j \rvert} p_j - p_i \frac{g_i - p_i}{\lvert g_i - p_i \rvert}(1 - p_i) \right]$
Cosine distance	$1 - \frac{\sum_j p_j g_j}{\sqrt{\sum_j p_j^2} \sqrt{\sum_j g_j^2}}$	$\frac{1}{C} \left[p_i \sum_{j \neq i} p_j (g_j - p_i \frac{\sqrt{\sum_i g_i^2}}{\sqrt{\sum_i p_i^2}} R) - p_i (g_i - p_i R)(1 - p_i) \right]$; where $R = \frac{\sum_i p_i g_i}{C}$ and $C = \sqrt{\sum_i p_i^2} \sqrt{\sum_i g_i^2}$.
Bhattacharyya distance	$-\ln \sum_j (p_j g_j)^{0.5}$	$\frac{-1}{2 \sum_j (p_j g_j)^{0.5}} \left[p_i \sum_{j \neq i} (p_j g_j)^{0.5} - (p_i g_i)^{0.5} (1 - p_i) \right]$
KL divergence	$\sum_j g_j \log \frac{g_j}{p_j}$	$p_i \sum_{j \neq i} g_j - g_i (1 - p_i)$

S. Jetley, N. Murray, and E. Vig. End-to-End Saliency Mapping via Probability Distribution Prediction. CVPR, 2016.

Deep supervised models



Convergence of AUC using different loss functions

S. Jetley, N. Murray, and E. Vig. End-to-End Saliency Mapping via Probability Distribution Prediction. CVPR, 2016.

Conclusions & Future Directions

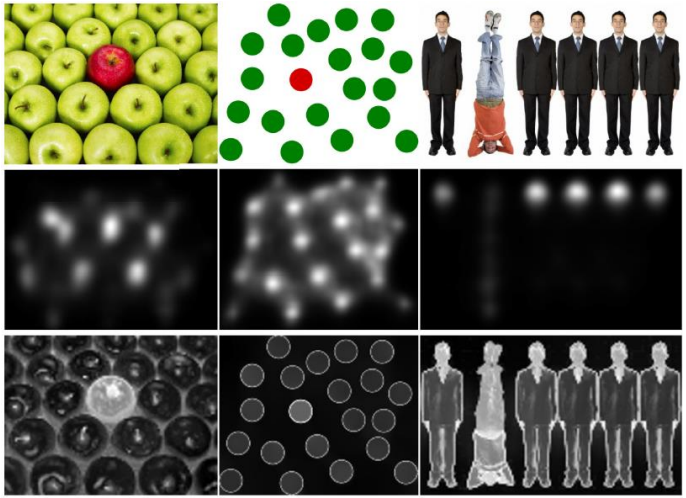
Conclusion

Using deep networks is a classical approach to modelling saliency maps

Supervised networks achieve state-of-the-art performance on standard benchmarks

But:

- Datasets are biased towards semantic objects
- What about psychophysical stimuli?
- Saliency v.s. eye-fixation prediction



S. Rahman & N. Bruce. Saliency, Scale and Information: Towards a Unifying Theory. NIPS, 2015.

What’s next for saliency map networks

- Spatio-temporal saliency networks
- From saliency maps to selective attention
- Incorporating saliency map networks into larger pipelines

V. Mnih, N. Heess, and A. Graves. "Recurrent models of visual attention." *NIPS* 2014.

Additional references

S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, R. V. Babu. Saliency Unified: A Deep Architecture for Simultaneous Eye Fixation Prediction and Salient Object Segmentation. CVPR, 2016.

N. Murray, M. Vanrell, X. Otazu and C. A. Párraga. Saliency Estimation Using a Non-Parametric Low-Level Vision Model. CVPR, 2011.

M. Riesenhuber& T. Poggio. "Hierarchical models of object recognition in cortex. Nature neuroscience, 1999.

L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. TPAMI, 1998.

M.C. Mozner & M. Sitton. Computational modelling of spatial attention. Attention. Eds. H. Pashler. Psychology Press, 1998.

J.K. Tsotsos, S.M. Culhane, W. Yan, K. Wai, Y. Lai, N. Davis, F. Nuflo. Modeling visual attention via selective tuning. Artificial intelligence, 1995.

P. Sandon, P. Simulating visual attention. J. Cognitive Neuroscience, 1990.

S. Grossberg, E. Mingolla, and D. Todorovic. A neural network architecture for preattentive vision." IEEE Transactions on Biomedical Engineering,1989.

K. Fukushima. A neural network model for selective attention in visual pattern recognition. Biological Cybernetics, 1986.

Thanks!