The importance of visual attention

Slides credited to Kevin McGuinness@DCU

The importance of visual attention

The importance of visual attention

The importance of visual attention

Why don’t we see the changes?

We don’t really see the whole image

We only focus on small specific regions: the **salient** parts

Human beings reliably attend to the same regions of images when shown

What we perceive

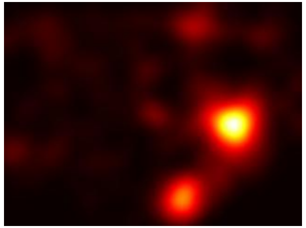
Where we look

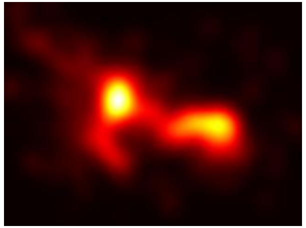
What we actually see

Saliency prediction

Produce a **computational model of visual attention**: predict where humans will look.

Often want to map an image to a **heatmap** (saliency map).

$f($  $) =$ 

$f($  $) =$ 

Deep supervised models

Datasets

MIT 300

300 natural indoor and outdoor scenes.

39 observers. 3 sec free view.

ETL 400 ISCAN eye tracker

Test set only: no training data or public ground truth

http://saliency.mit.edu/results_mit300.html

[A Benchmark of Computational Models of Saliency to Predict Human Fixations]() [MIT tech report 2012]

Fixations and saliency maps

Raw eye tracker data needs to be processed to produce saliency maps

Detect saccades using distance/velocity thresholding, clustering

Rendering, Gaussian blur, normalizing

Eye tracker → Raw gaze location-time tracks → Fixation detection → Eye fixations → Saliency map generation

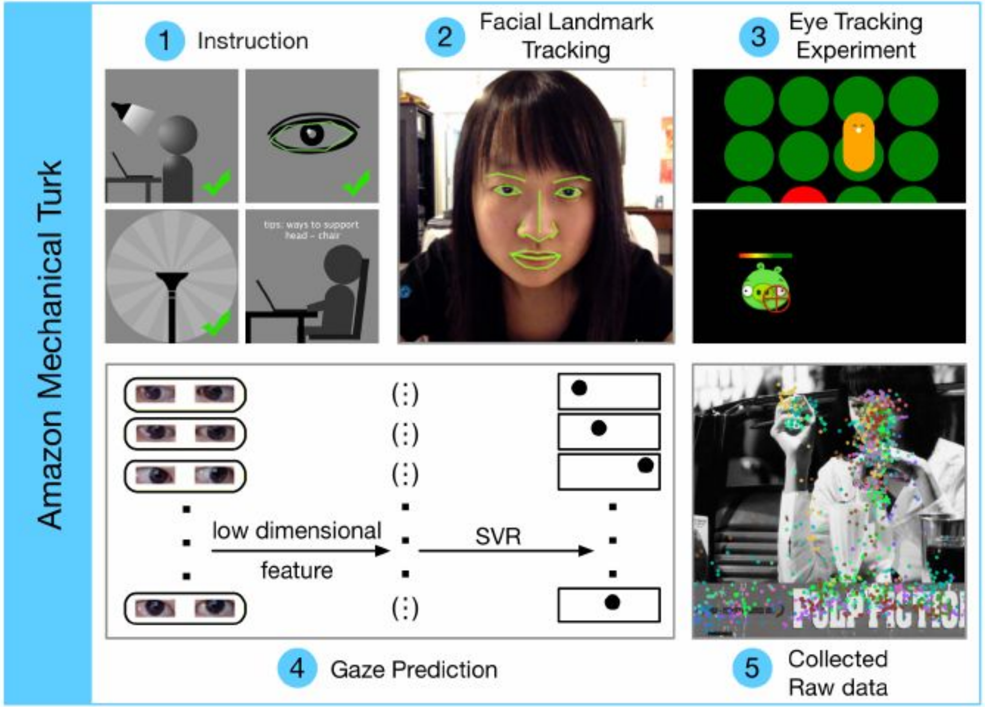
iSUN

Large scale dataset of natural scenes

20,608 images with avg. 3 observers each

Collected using webcams and Amazon Mechanical Turk

Used in [LSUN challenge](#) 2015/2016



Xu et al. TurkerGaze: Crowdsourcing Saliency with Webcam based Eye Tracking arXiv 2015.

Deep supervised models

Typically, superior performance to unsupervised models

Large-scale proxy datasets have enabled effective supervised learning

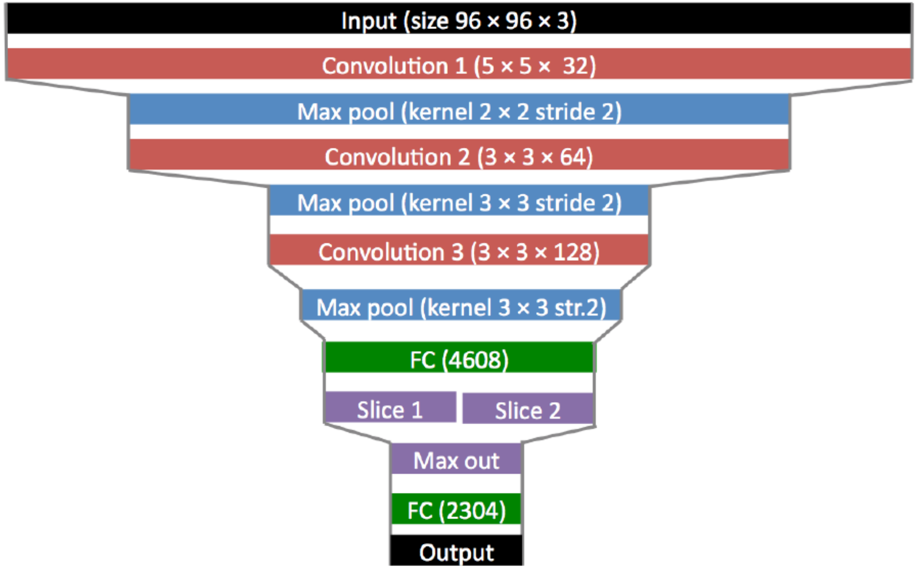
Key considerations:

- Network architecture
- Incorporation of prior cues
- Supervision mechanism
- Loss function

Deep supervised models

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

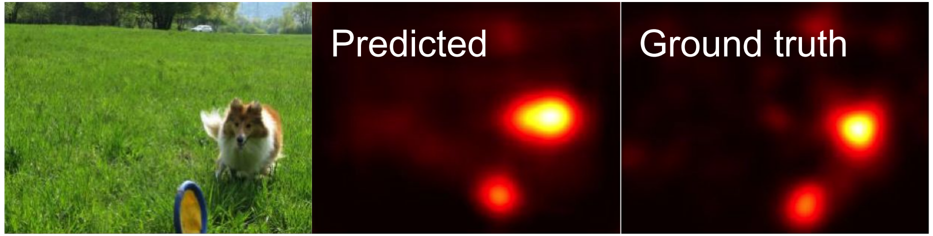


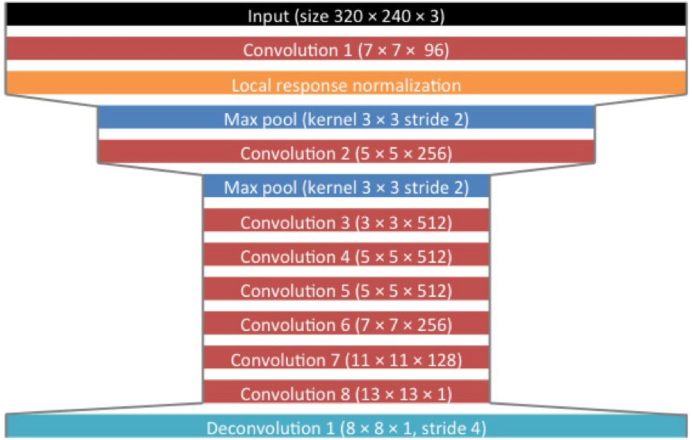
J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, N. E. O'Connor. Shallow and Deep Convolutional Networks for Saliency Prediction. CVPR, 2016.

SalNet: deep visual saliency model

Predict map of visual attention from image pixels (find the parts of the image that stand out)

- Feedforward 8 layer "fully convolutional" architecture
- Transfer learning in bottom 3 layers from pretrained VGG-M model on ImageNet
- Trained on SALICON dataset

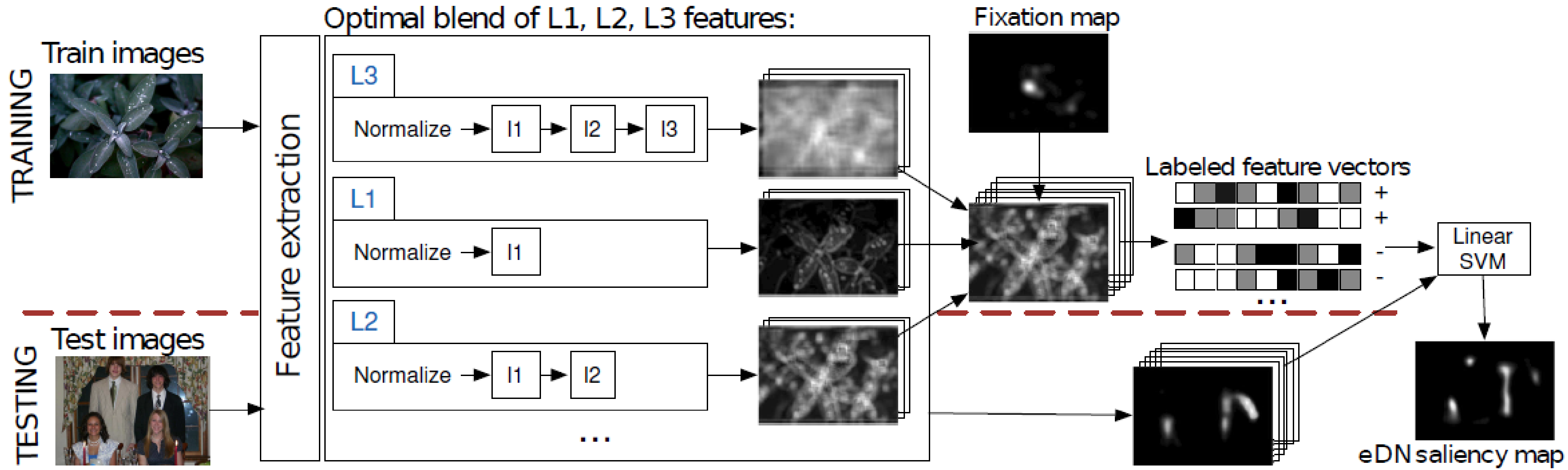




Pan, McGuinness, et al. **Shallow and Deep Convolutional Networks for Saliency Prediction**, CVPR 2016 http://arxiv.org/abs/1603.00845

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.]()


Image
Ground truth
Prediction

Image

Ground truth

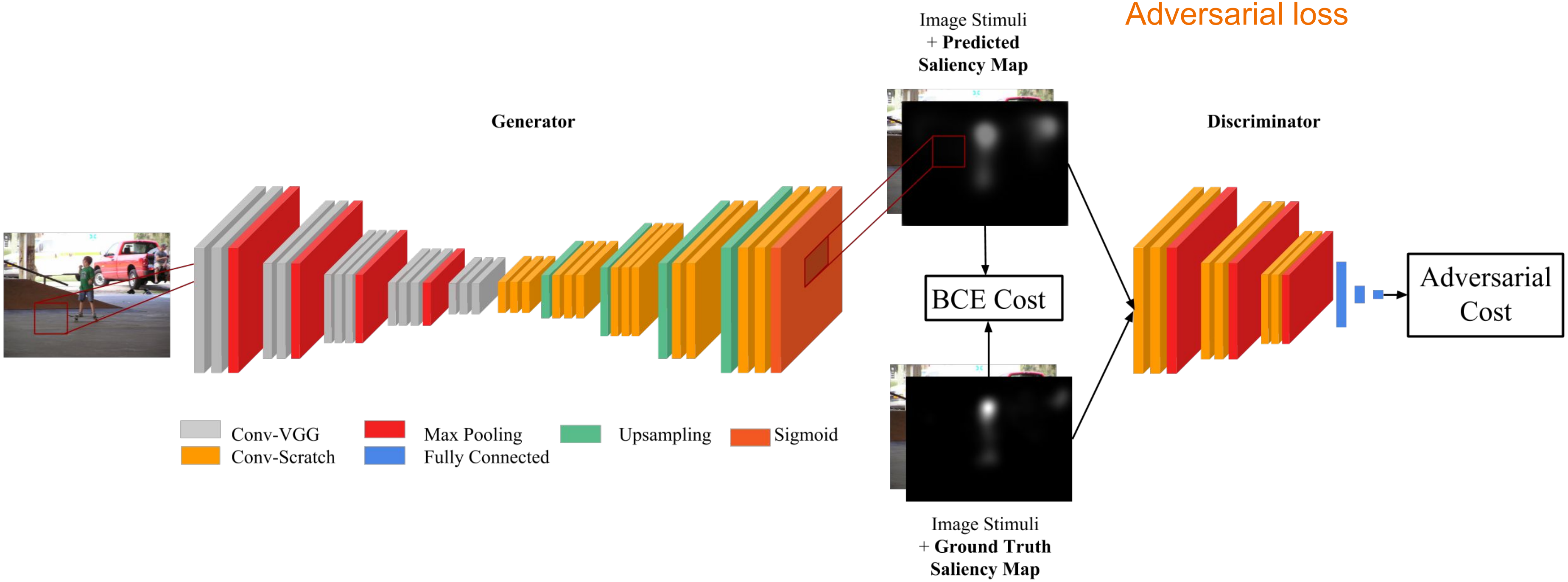
Prediction

SalGAN

Data loss

$$\alpha \cdot \mathcal{L}_{BCE} - \boxed{\log D(I, \hat{S})},$$

Adversarial loss



Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol and Xavier Giro-i-Nieto. "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks." arXiv. 2017.

SalNet and SalGAN benchmarks

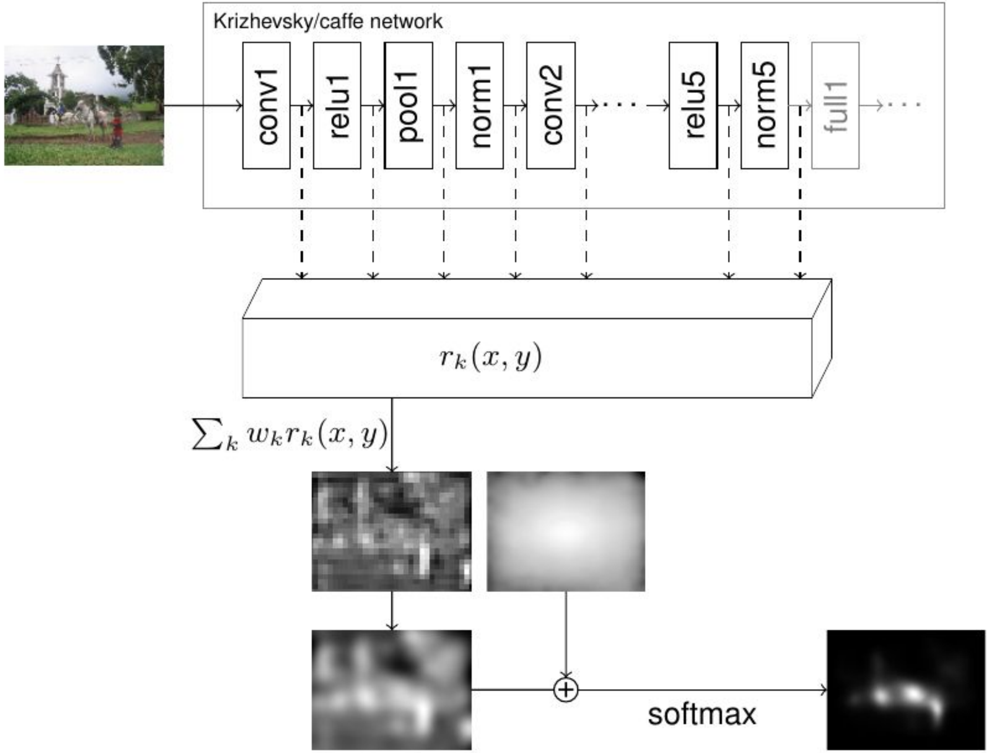
SALICON (test)	AUC-J ↑	Sim ↑	EMD ↓	AUC-B ↑	sAUC ↑	CC ↑	NSS ↑	KL ↓
DSCLRCN [24](*)	-	-	-	0.884	0.776	0.831	3.157	-
SalGAN	-	-	-	**0.884**	**0.772**	**0.781**	**2.459**	-
ML-NET [5]	-	-	-	(0.866)	(0.768)	(0.743)	2.789	-
SalNet [25]	-	-	-	(0.858)	(0.724)	(0.609)	(1.859)	-
MIT300	AUC-J ↑	Sim ↑	EMD ↓	AUC-B ↑	sAUC ↑	CC ↑	NSS ↑	KL ↓
Humans	0.92	1.00	0.00	0.88	0.81	1.0	3.29	0.00
Deep Gaze II [21](*)	0.88	(0.46)	(3.98)	0.86	0.72	(0.52)	(1.29)	(0.96)
DSCLRCN [24](*)	0.87	0.68	2.17	(0.79)	0.72	0.80	2.35	0.95
DeepFix [17](*)	0.87	0.67	2.04	(0.80)	(0.71)	0.78	2.26	0.63
SALICON [9]	0.87	(0.60)	(2.62)	0.85	0.74	0.74	2.12	0.54
SalGAN	**0.86**	**0.63**	**2.29**	**0.81**	**0.72**	**0.73**	**2.04**	**1.07**
PDP [11]	(0.85)	(0.60)	(2.58)	(0.80)	0.73	(0.70)	2.05	0.92
ML-NET [5]	(0.85)	(0.59)	(2.63)	(0.75)	(0.70)	(0.67)	2.05	(1.10)
Deep Gaze I [19]	(0.84)	(0.39)	(4.97)	0.83	(0.66)	(0.48)	(1.22)	(1.23)
iSEEL [29](*)	(0.84)	(0.57)	(2.72)	0.81	(0.68)	(0.65)	(1.78)	0.65
SalNet [25]	(0.83)	(0.52)	(3.31)	0.82	(0.69)	(0.58)	(1.51)	0.81
BMS [31]	(0.83)	(0.51)	(3.35)	0.82	(0.65)	(0.55)	(1.41)	0.81

Deep Gaze

Simple linear model trained on activations of all conv layers (upsampled) from AlexNet

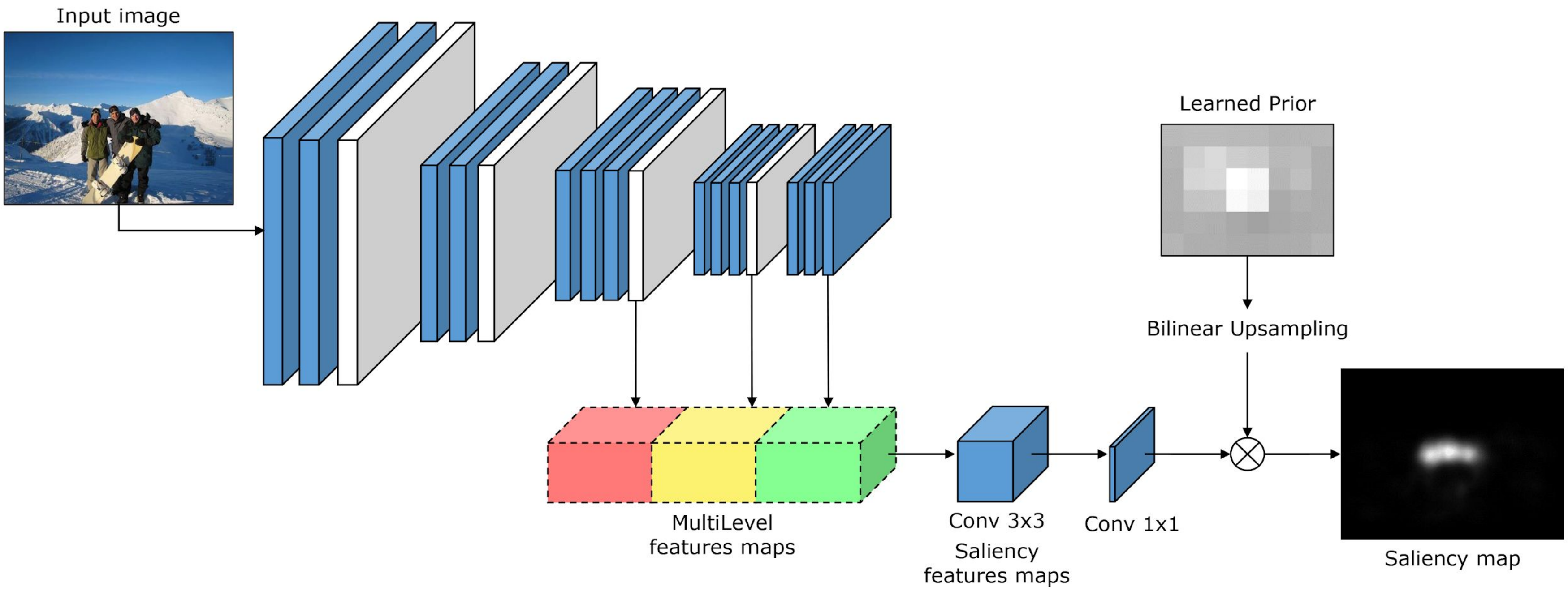
Softmax output over full image, categorical cross entropy.

L_1 regularization used to encourage sparsity.



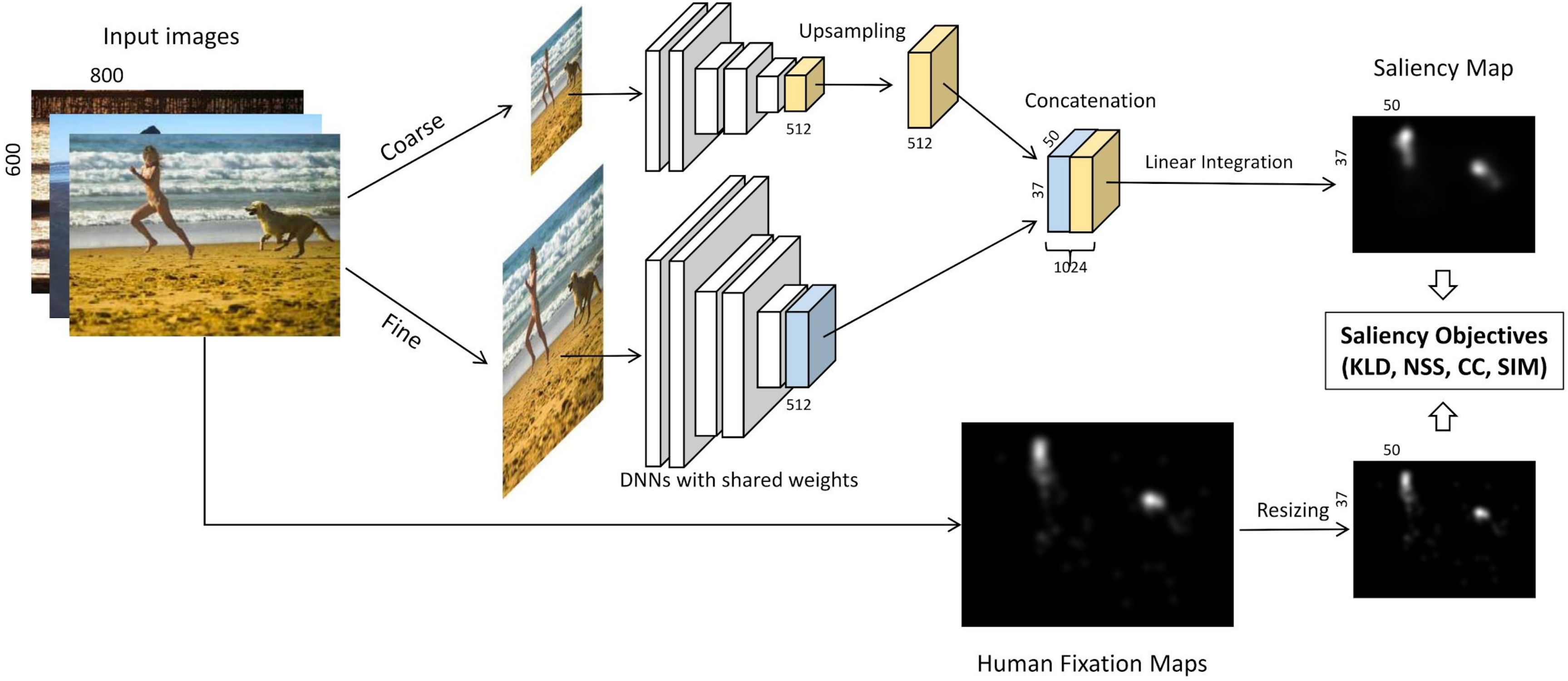
Kümmerer et al. Deep gaze 1: Boosting saliency prediction with feature maps trained on imagenet. ICLR workshops 2015

MLNet



Cornia et al. [A Deep Multi-Level Network for Saliency Prediction](). ICPR 2016.

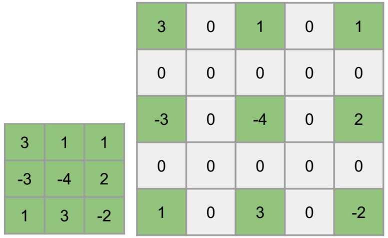
SALICON



Huang et al., SALICON: [Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks](), ICCV 2015

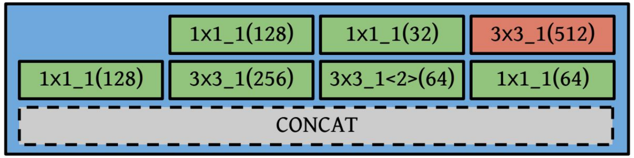
DeepFix

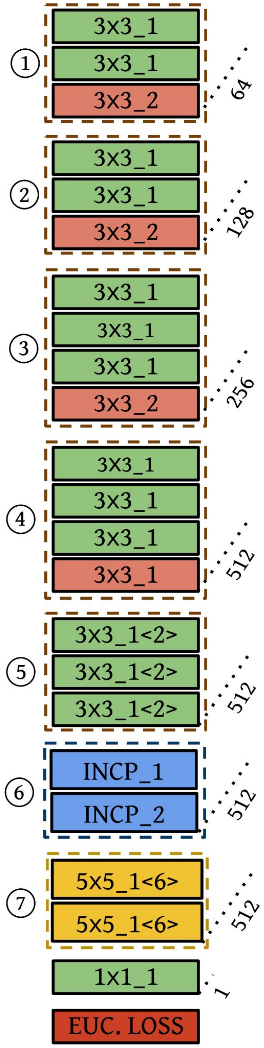
Kruthiventi et al. **DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations**
https://arxiv.org/abs/1510.02927

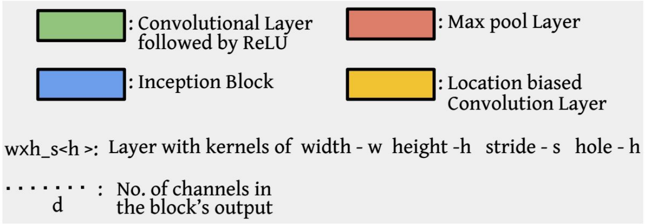


Dilated convolutions

Inception layers







Weights initialized from VGG16 trained on ImageNet

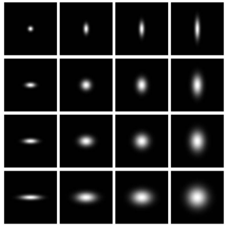
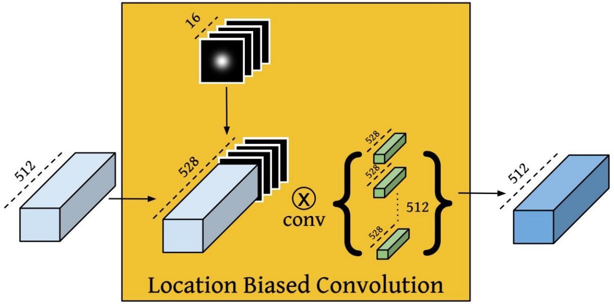
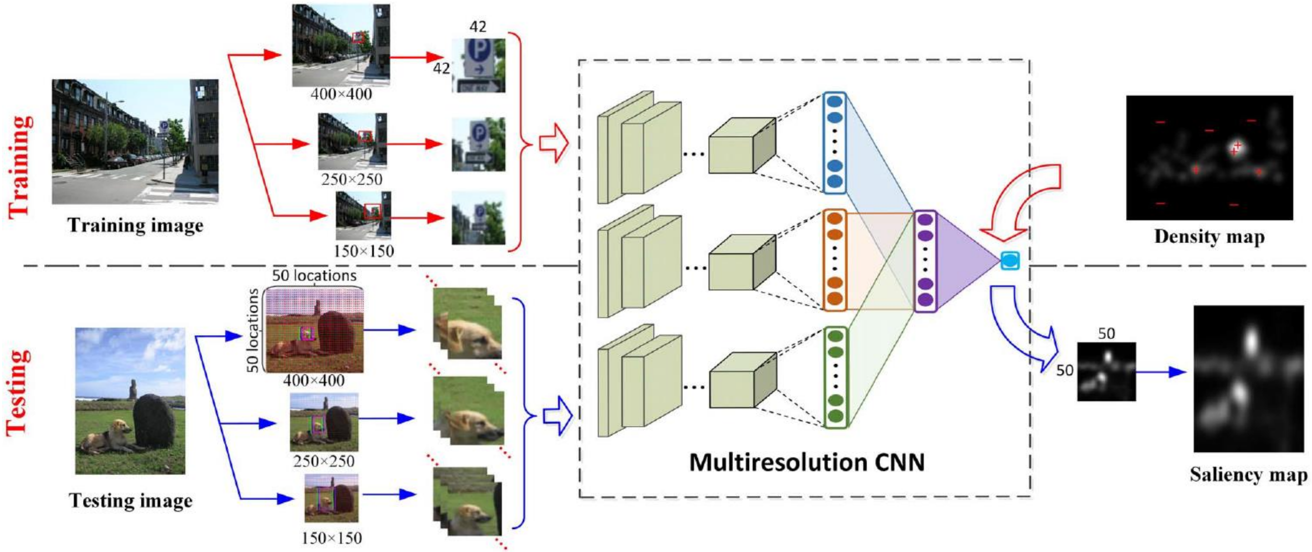


Fig. 6: Gaussian blobs with different horizontal and vertical variances concatenated to the input blob of LBC layers to make the layer's response location specific.



Location biased convolutions

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

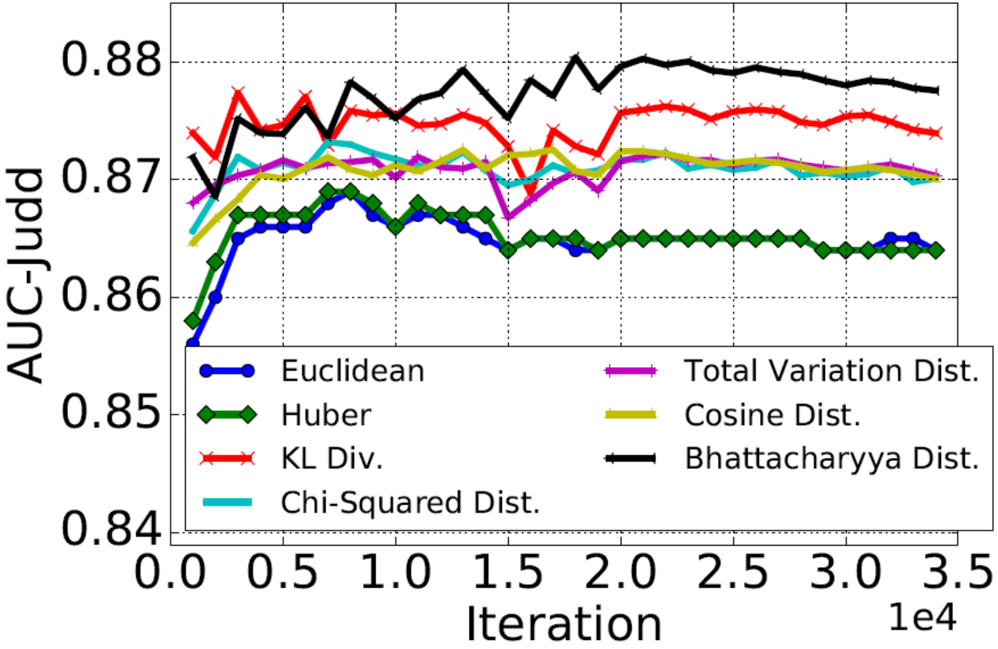
Dense prediction problem - which loss functions to use?

- 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.]

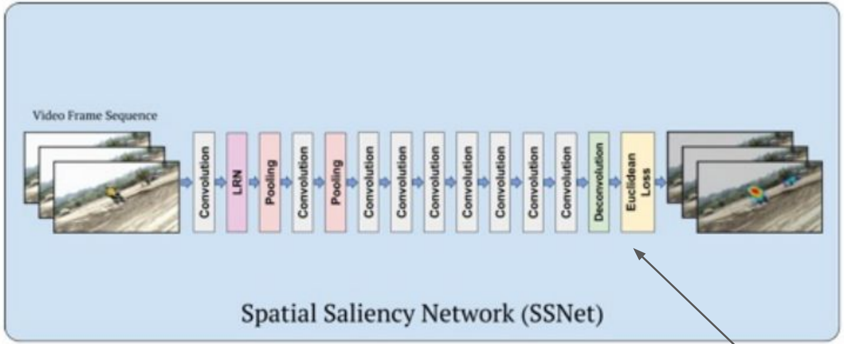
Deep supervised models

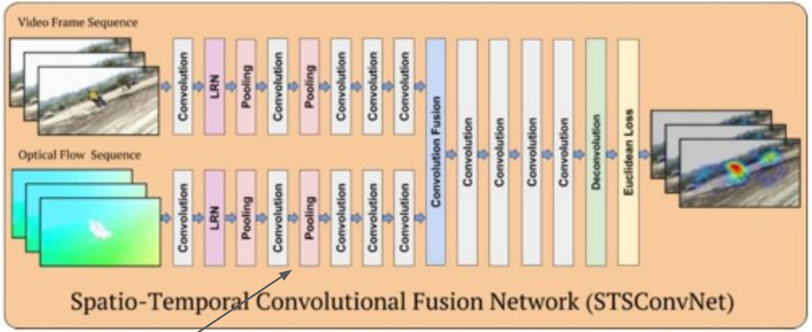


Convergence of AUC using different loss functions

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

From image to video saliency?





SalNet

Bak et al. Spatio-temporal saliency networks for dynamic saliency prediction. IEEE Trans. Multimedia 2017

Thanks!