http://bit.ly/dlcv2018

#DLUPC

Day 3 Lecture 5

Saliency Prediction

Kevin McGuinness
kevin.mcguinness@dcu.ie

Assistant Professor
School of Electronic Engineering
Dublin City University

DCU

Insight
Centre for Data Analytics

The importance of visual attention

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

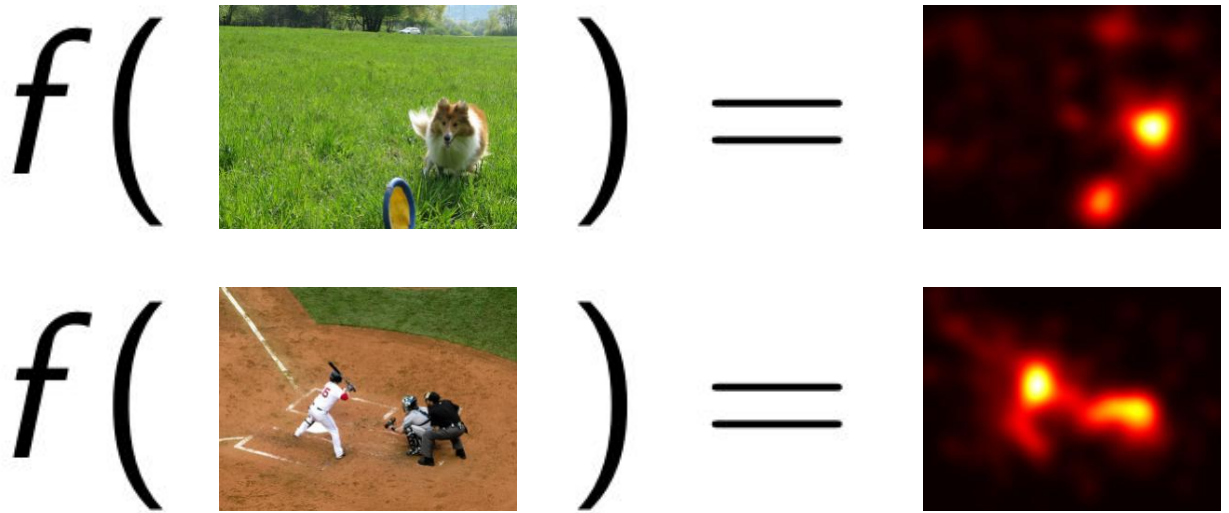
Where we look

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).



Salient object detection?

Often confused with saliency prediction, but a different task.

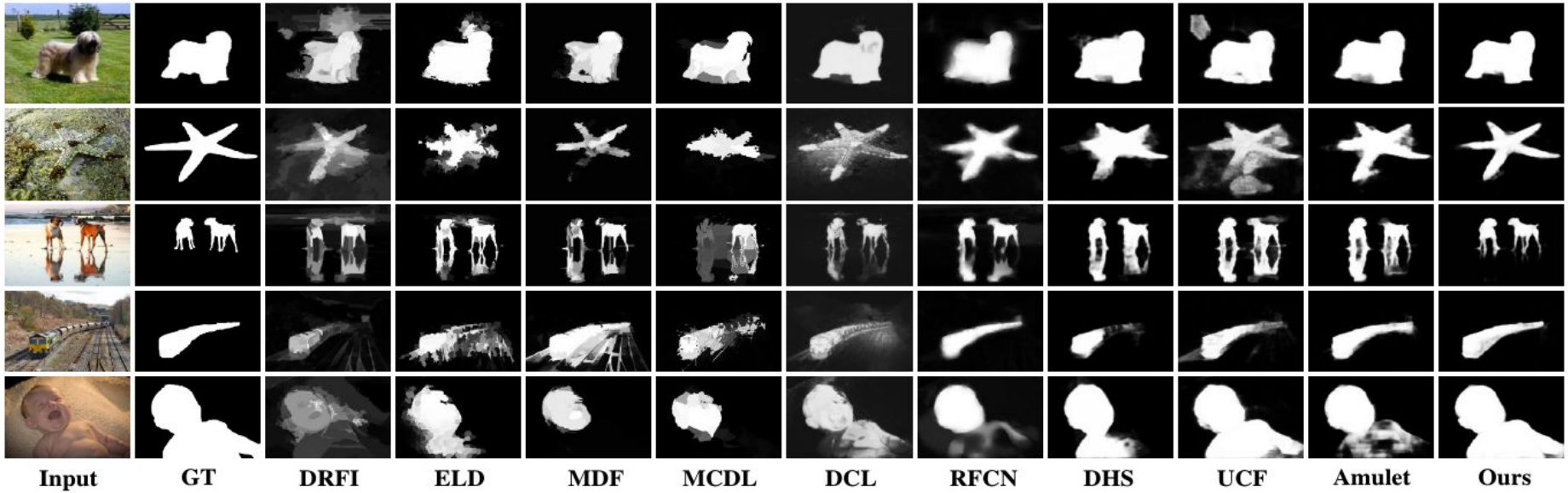


Figure from: Progressive Attention Guided Recurrent Network for Salient Object Detection, CVPR 2018

Datasets

MIT 300

300 natural indoor and outdoor scenes.

39 observers. 3 sec free view.

ETL 400 ISCAN eye tracker

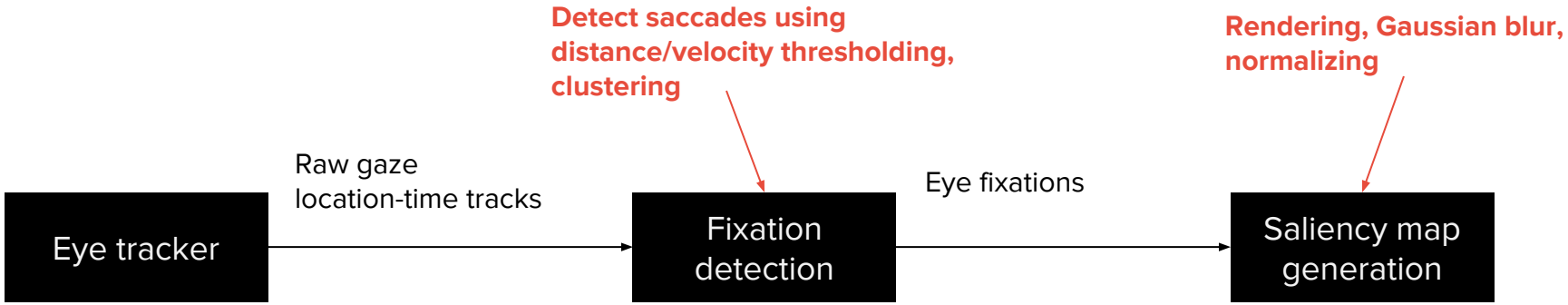
Test set only: no training data or public ground truth

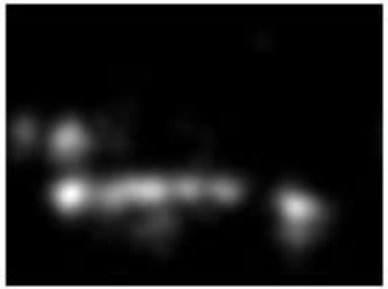
http://saliency.mit.edu/results_mit300.html

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





MIT 1003

1003 natural indoor and outdoor scenes.

15 observers. 3 sec free view.

ETL 400 ISCAN eye tracker

Training dataset for MIT 300

Learning to Predict where Humans Look [ICCV 2009]

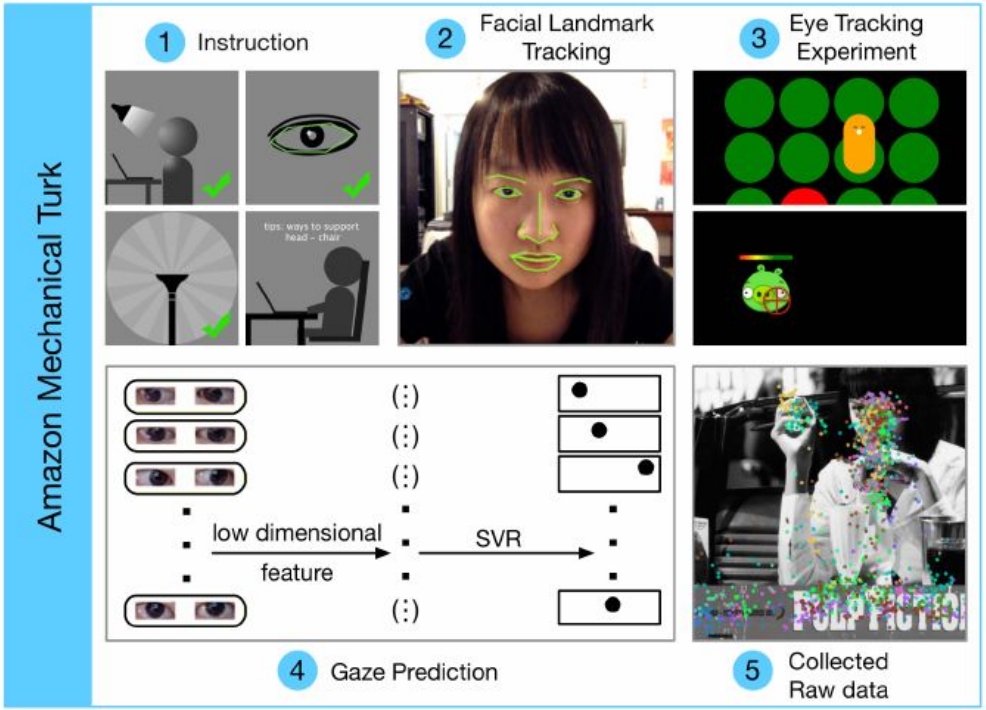
iSUN

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.

SALICON

Another large scaled dataset of images from MS COCO dataset

10K train, 5K val, 5K test

Simulated crowdsourced attention using **mouse** movements and **simulated artificial foveation**.

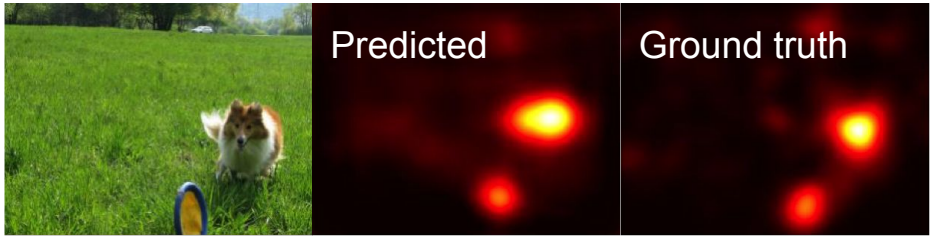
Jiang et al. SALICON: Saliency in Context, CVPR 2015

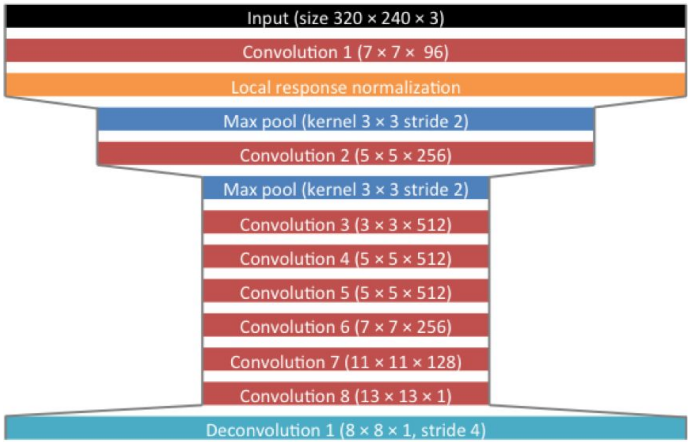
Models

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





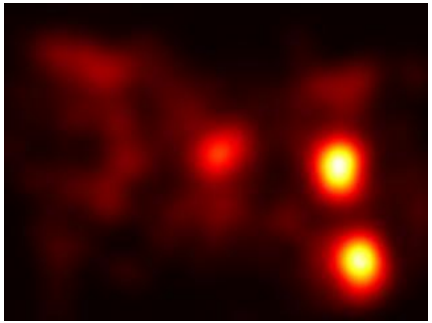
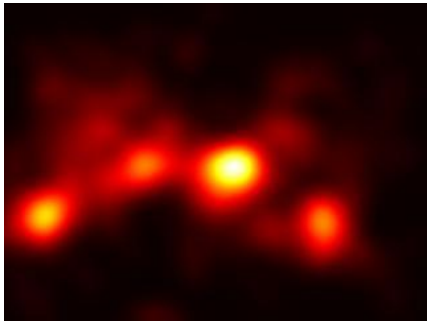
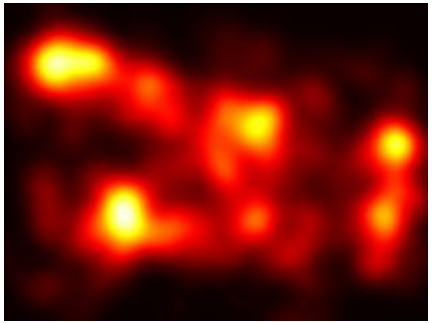
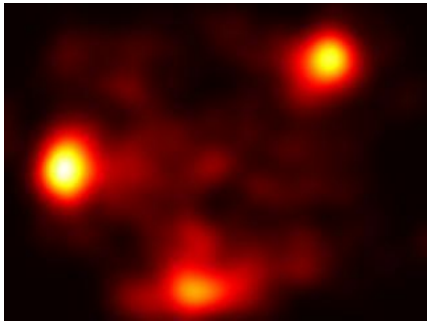
Pan, McGuinness, et al. **Shallow and Deep Convolutional Networks for Saliency Prediction**, CVPR 2016 http://arxiv.org/abs/1603.00845


Image
Ground truth
Prediction

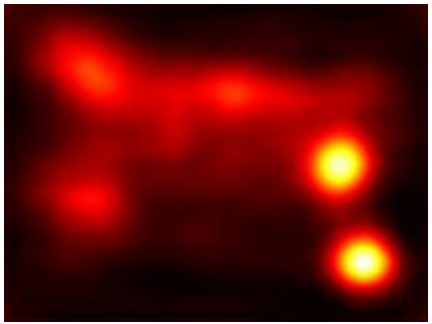
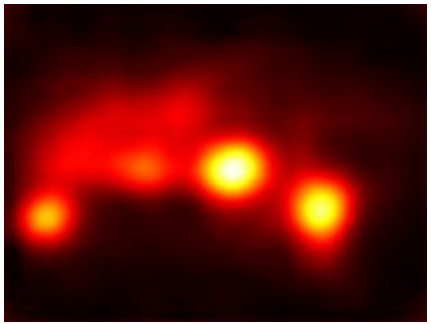
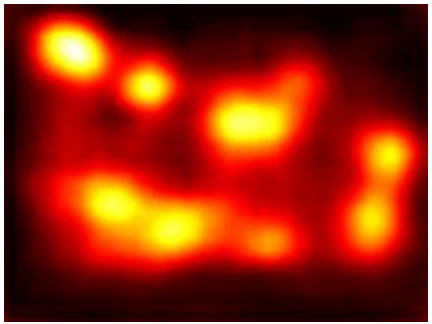
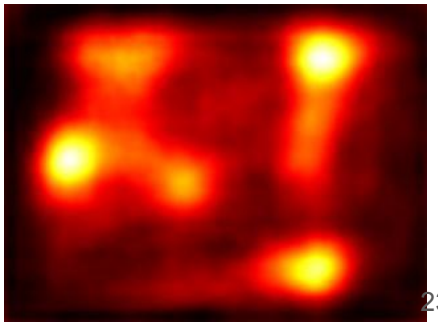
Image




Ground truth

Prediction

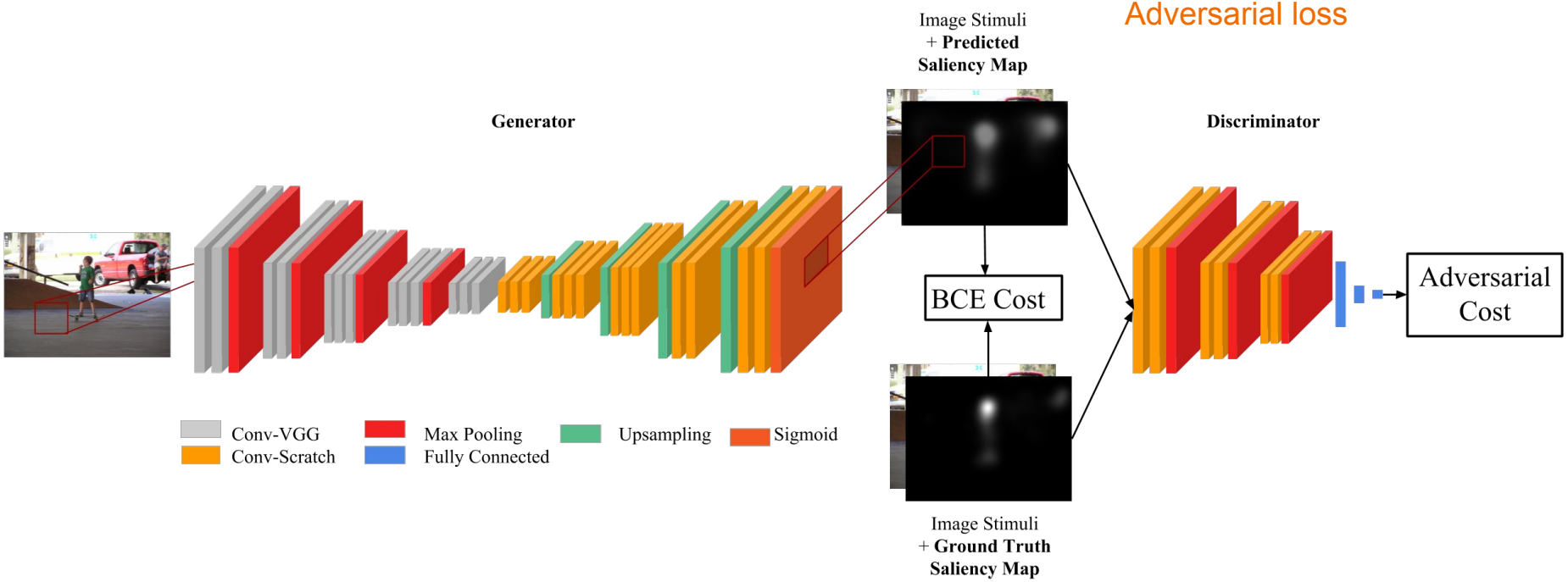





SalGAN

Data loss

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

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

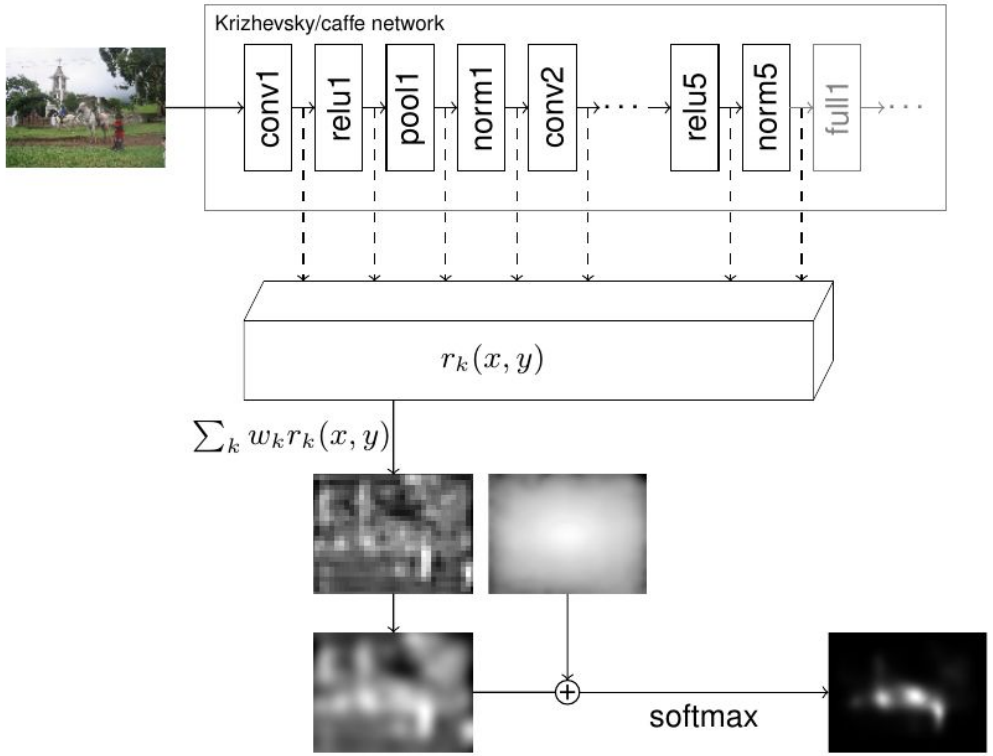
SALICON (test)	AUC-J ↑	Sim ↑	EMD ↓	AUC-B ↑	sAUC ↑	CC ↑	NSS ↑	KL ↓
DSCLRCN [24](*)	-	-	-	0.884	0.776	0.831	3.157	-
SalGAN	-	-	-	**0.884**	**0.772**	**0.781**	**2.459**	-
ML-NET [5]	-	-	-	(0.866)	(0.768)	(0.743)	2.789	-
SalNet [25]	-	-	-	(0.858)	(0.724)	(0.609)	(1.859)	-
MIT300	AUC-J ↑	Sim ↑	EMD ↓	AUC-B ↑	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

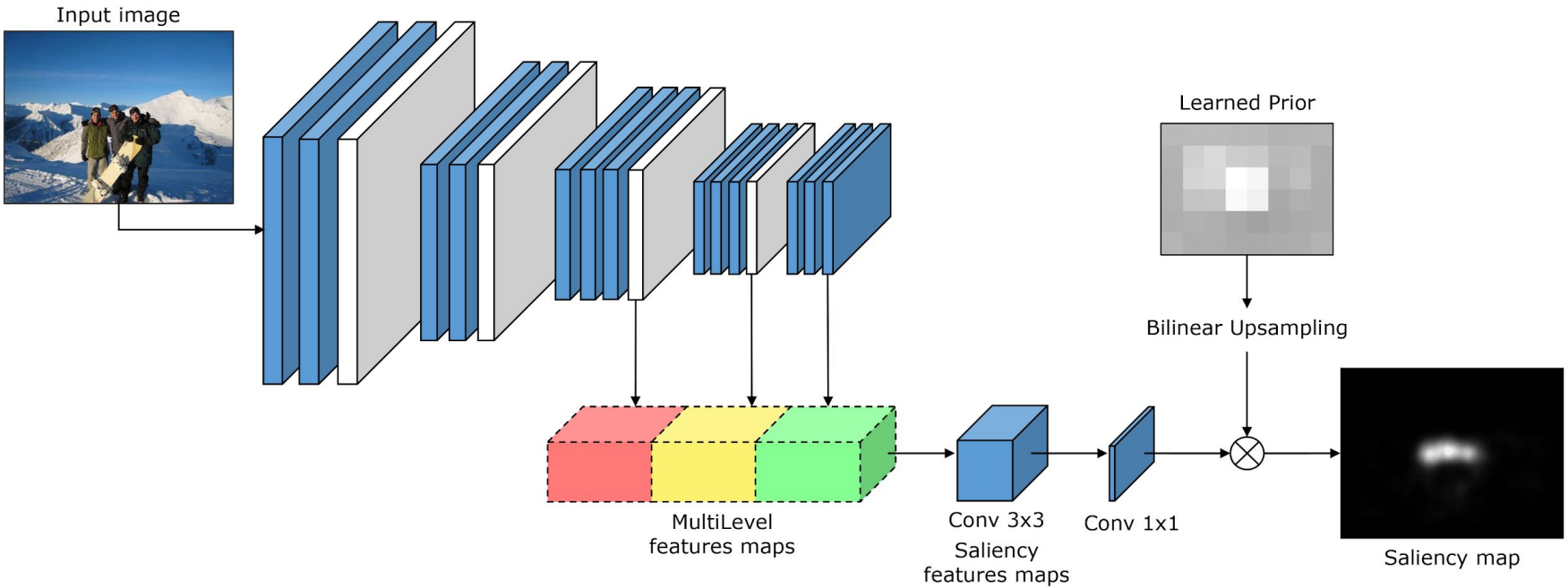
Softmax output over full image, categorical cross entropy.

L_1 regularization used to encourage sparsity.



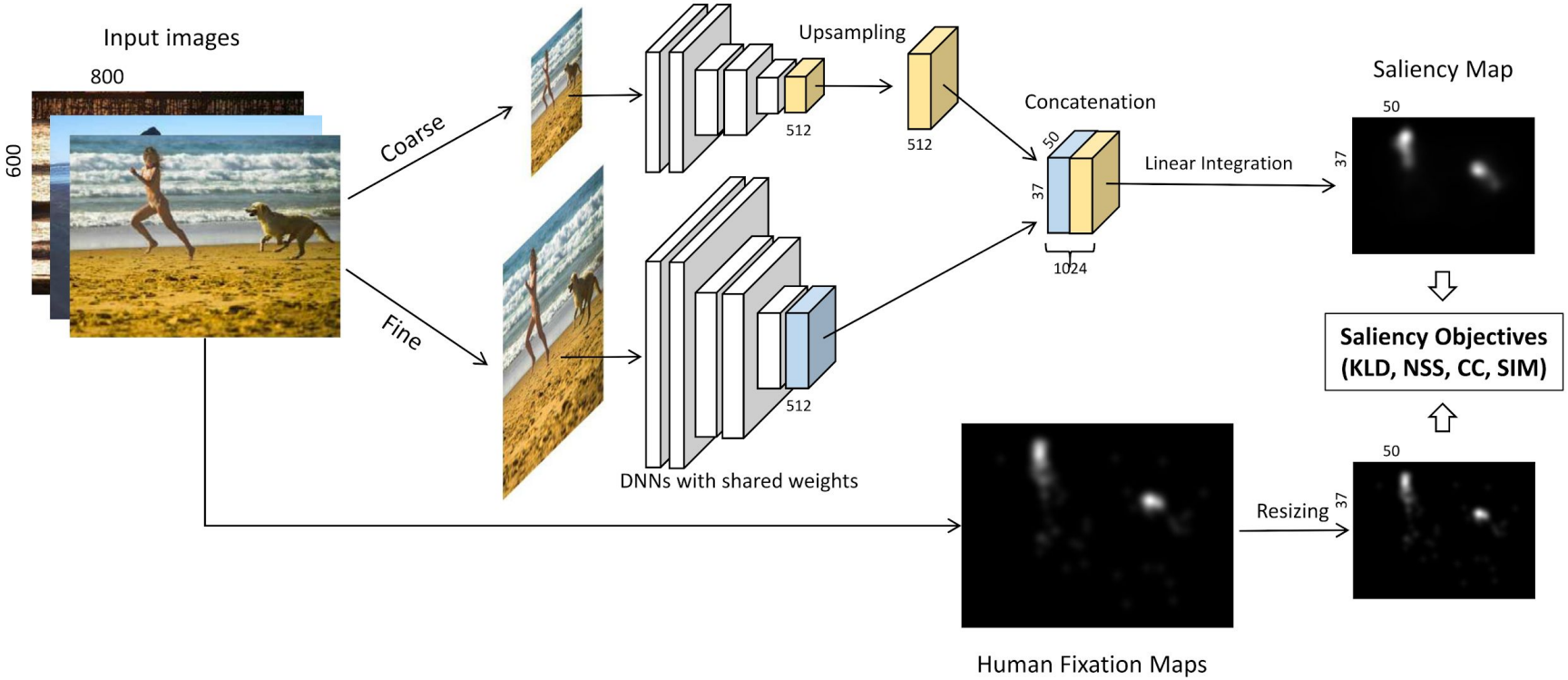
Kümmerer et al. Deep gaze 1: Boosting saliency prediction with feature maps trained on imagenet. ICLR workshops 2015

MLNet



Cornia et al. A Deep Multi-Level Network for Saliency Prediction. ICPR 2016.

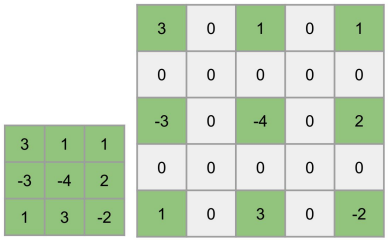
SALICON



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

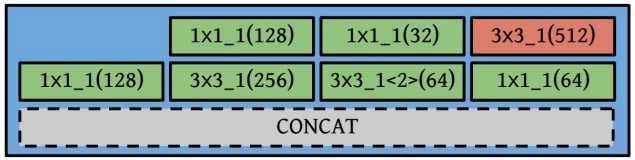
DeepFix

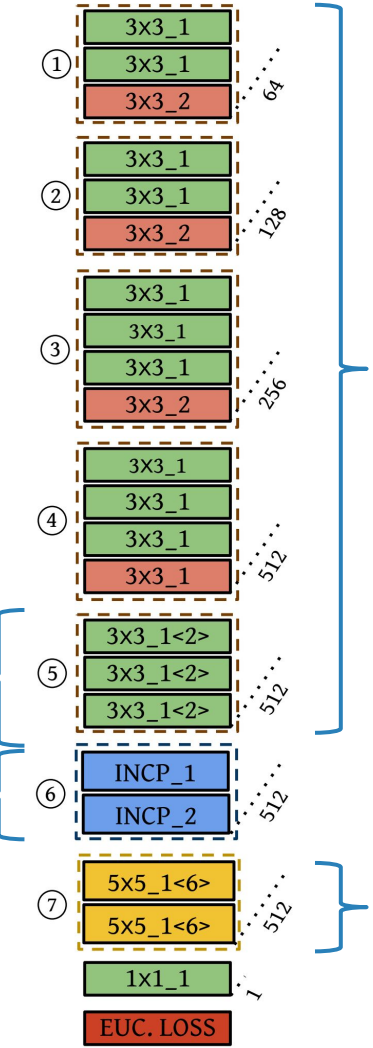
Kruthiventi et al. **DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations**
https://arxiv.org/abs/1510.02927



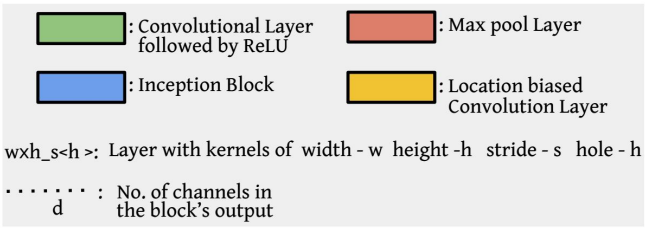
Dilated convolutions

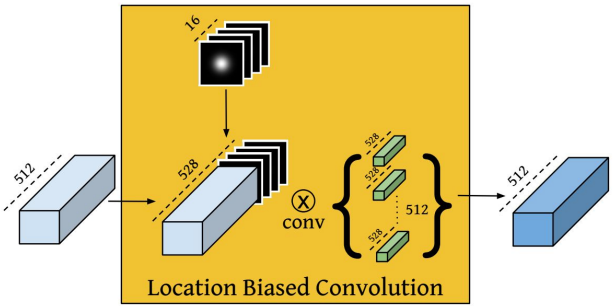
Inception layers





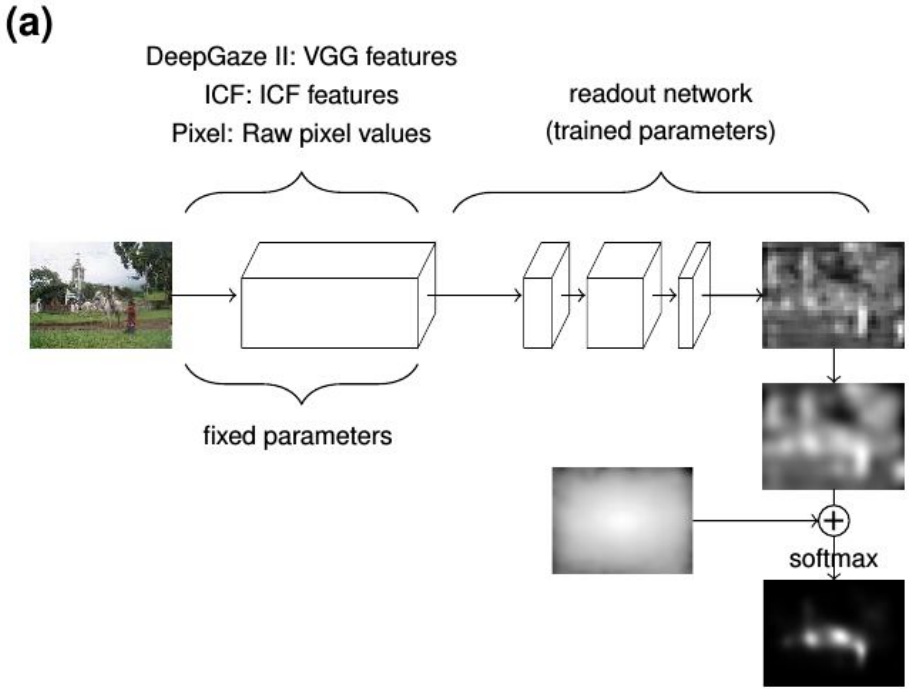
Weights initialized from VGG16 trained on ImageNet

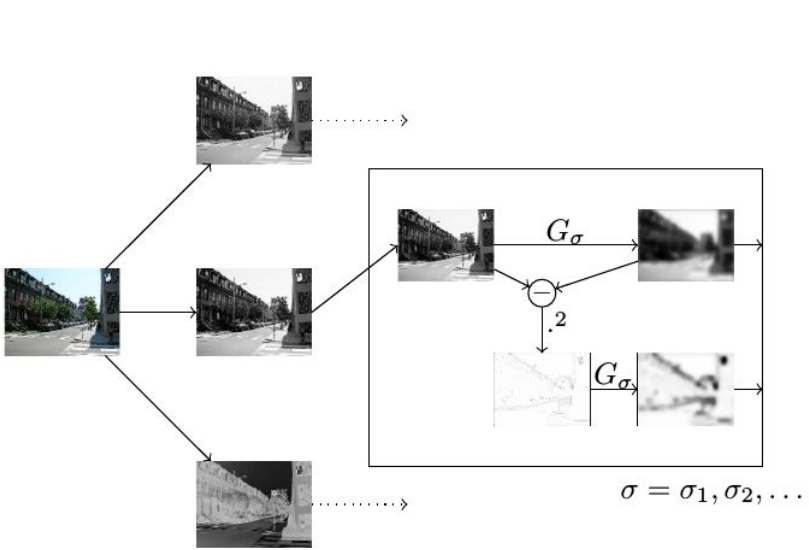




Location biased convolutions

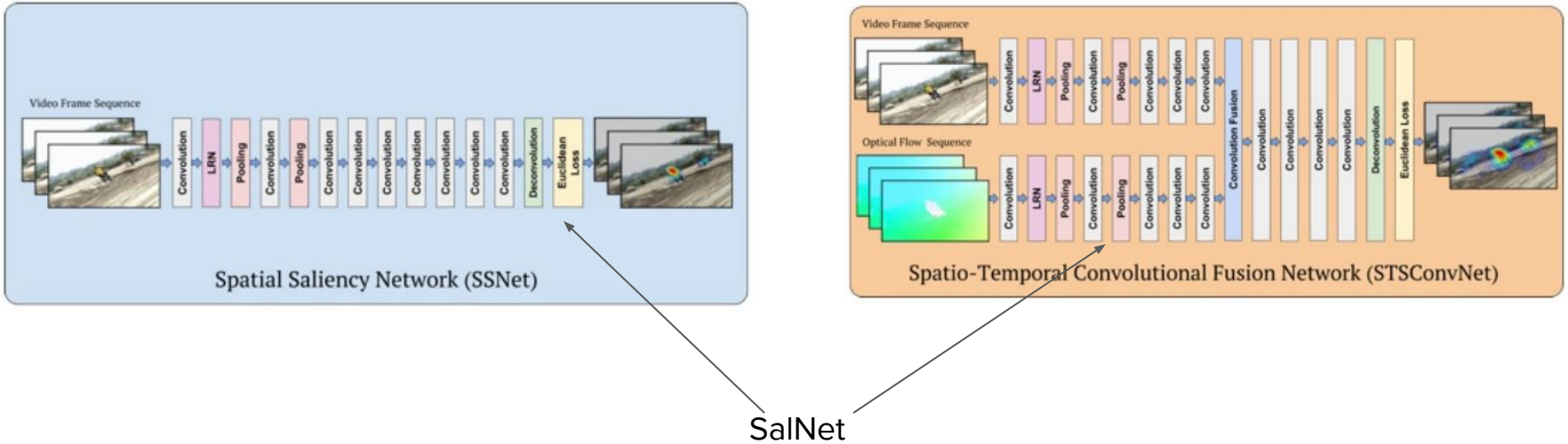
Deep Gaze II





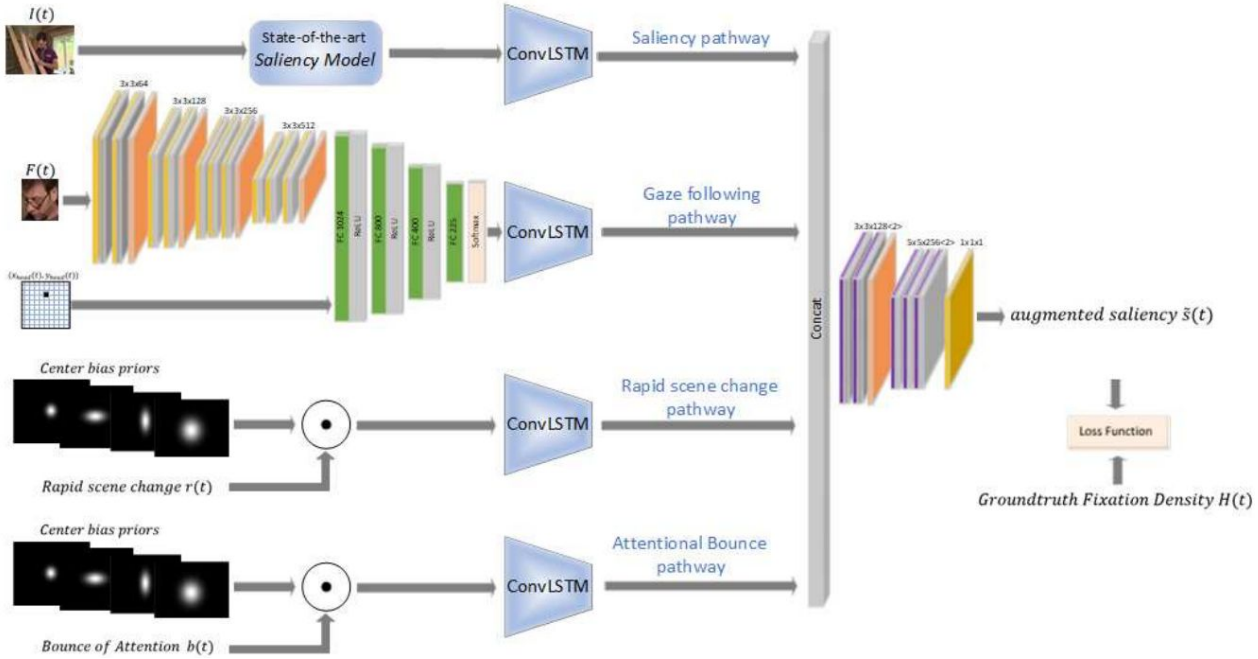
Kümmerer, et al. Understanding low-and high-level contributions to fixation prediction. ICCV 2017

From image to video saliency?



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

From image to video saliency



Gorji and Clark, Going From Image to Video Saliency: Augmenting Image Salience With Dynamic Attentional Push, CVPR 2018

Questions?