Insight
Centre for Data Analytics

DEEP LEARNING WORKSHOP

Dublin City University
21-22 May 2018

DCU

Learning where (and when) to look

Focus and attention in deep vision

Kevin McGuinness
kevin.mcguinness@dcu.ie

Assistant Professor
School of Electronic Engineering
Dublin City University

A salient team

Kevin McGuinness

Junting Pan

Marc Assens

Marta Coll

Xavier Giro-i-Nieto

Noel E. O'Connor

Cristian Canton

Elisa Sayrol

Jordi Torres

Eva Mohedano

Insight
Centre for Data Analytics

DCU Dublin City University
Ollscoil Chathair Bhaile Átha Cliath

facebook

Microsoft

BSC Barcelona Supercomputing Center
Centro Nacional de Supercomputación

UPC UNIVERSITAT POLITÈCNICA DE CATALUNYA
BARCELONATECH

Overview

Visual attention **models**

Applications of visual attention models

Modelling the **temporal** dimension of visual attention

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

Can we predict where humans will look?

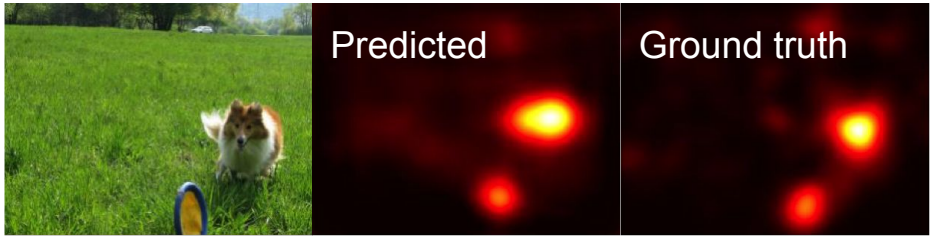
Yes! Computational models of visual saliency

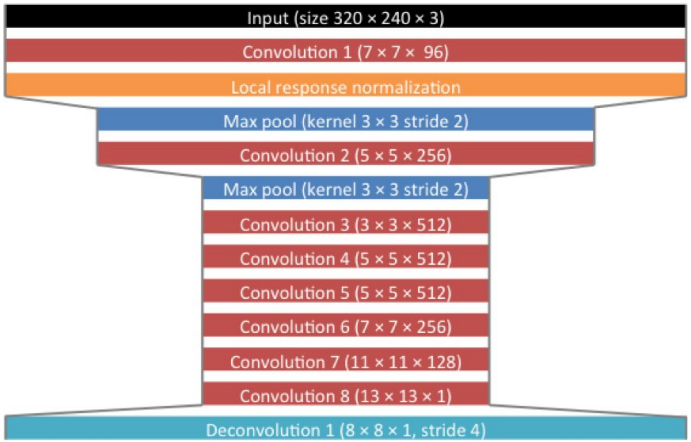
Why might this be useful?

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 (simulated crowdsourced attention dataset using mouse and artificial foveation)
- Top-5 in MIT 300 saliency benchmark http://saliency.mit.edu/results_mit300.html





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


Image
Ground truth
Prediction

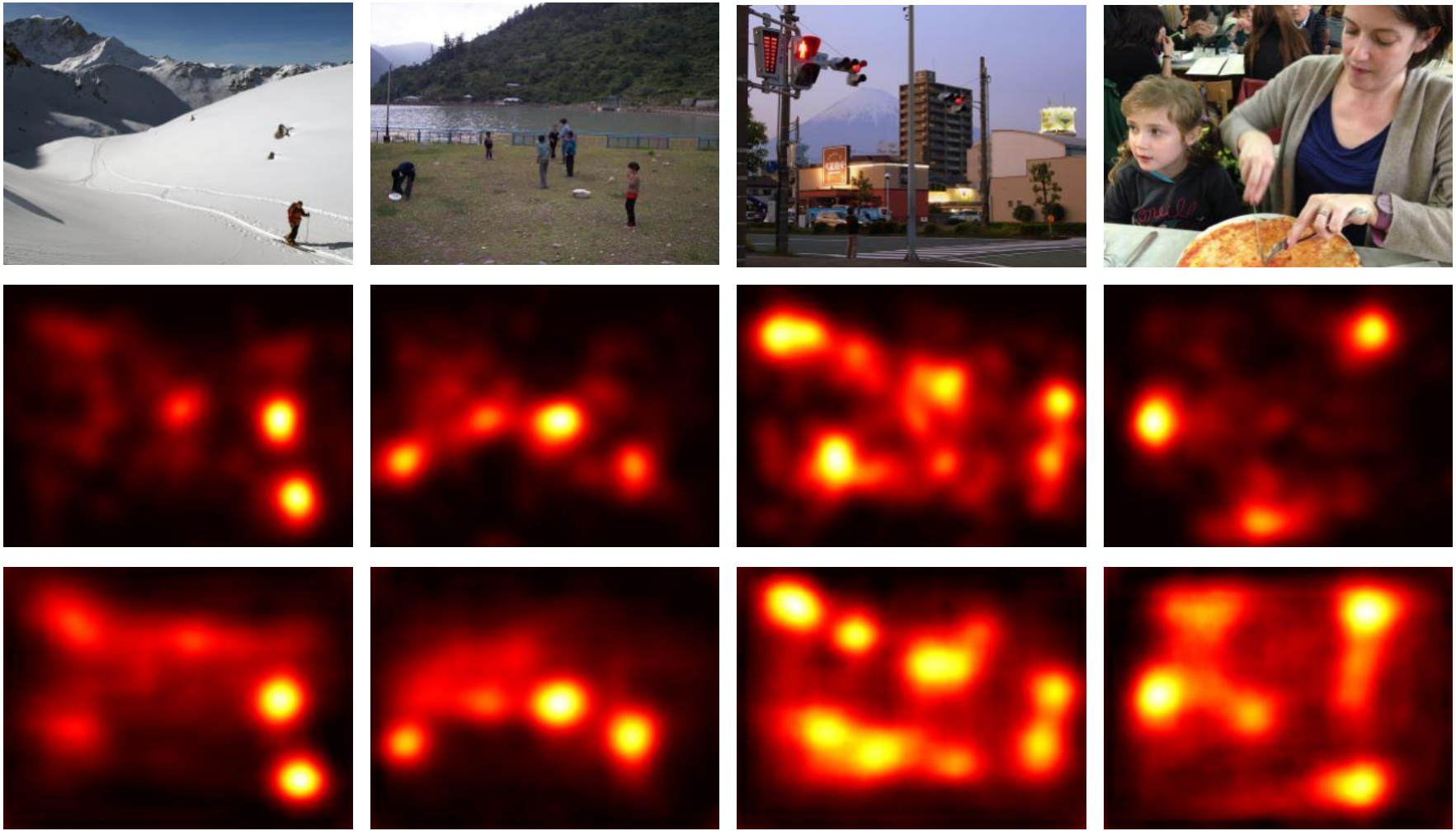


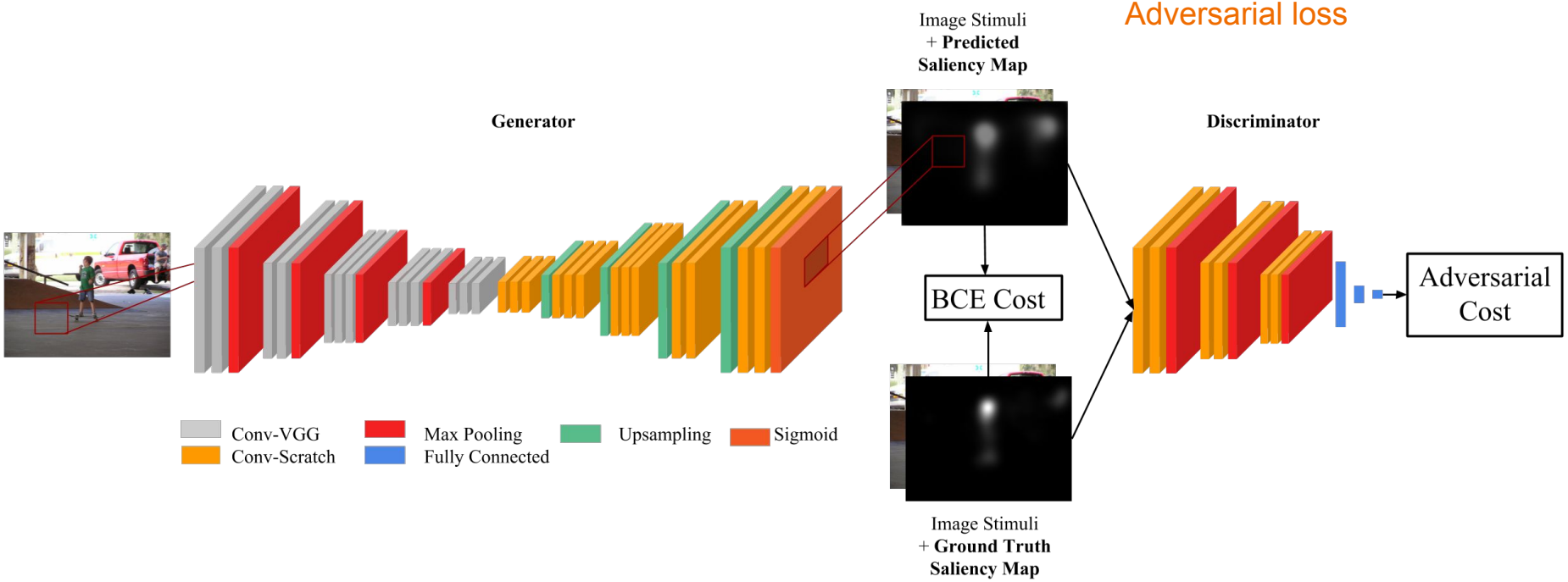
Image
Ground truth
Prediction

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

Applications of visual attention

Intelligent image cropping

Image retrieval

Improved image classification

Intelligent image cropping

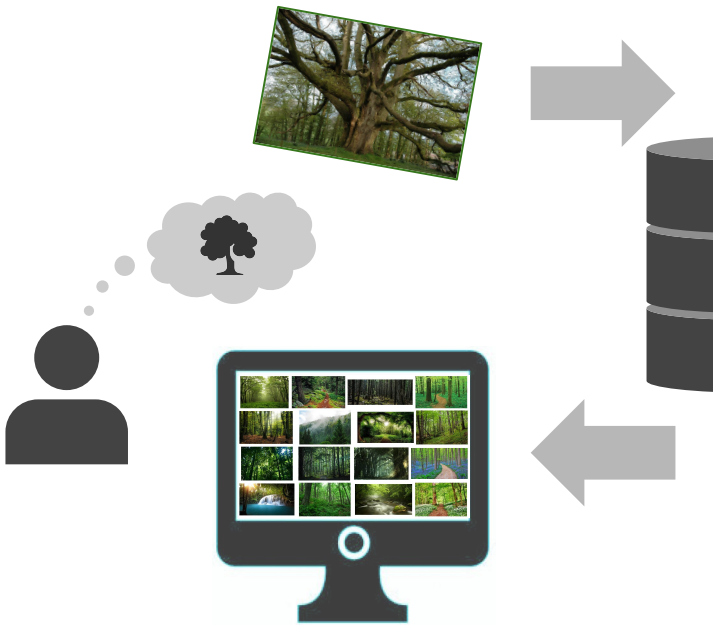
Image retrieval: query by example

Given:

- An example query image that illustrates the user's information need
- A very large dataset of images

Task:

- Rank all images in the dataset according to how likely they are to fulfil the user's information need



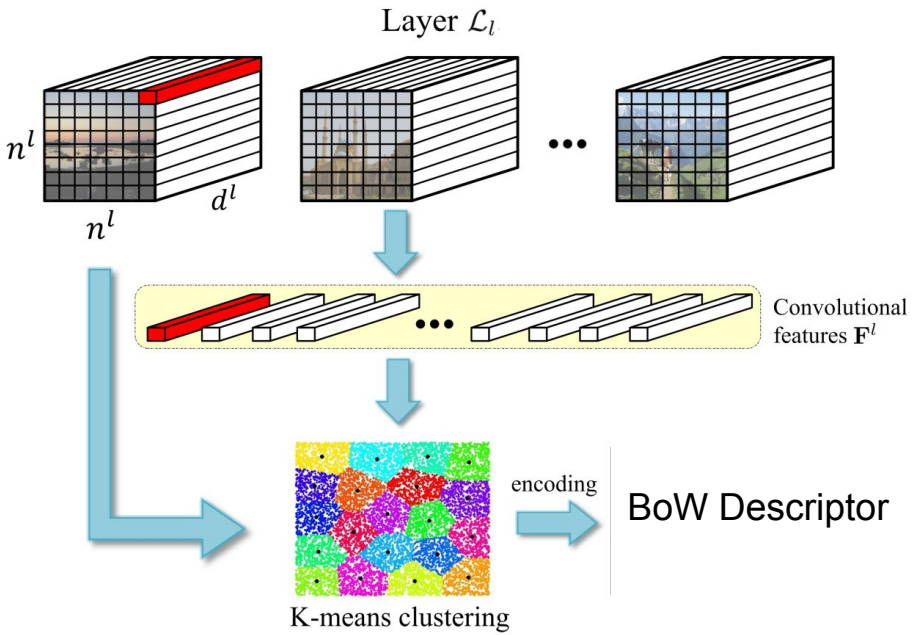
Bags of convolutional features instance search

Objective: rank images according to relevance to query image

Local CNN features and BoW

- Pretrained VGG-16 network
- Features from conv-5
- L2-norm, PCA, L2-norm
- K-means clustering -> BoW
- Cosine similarity
- Query augmentation, spatial reranking

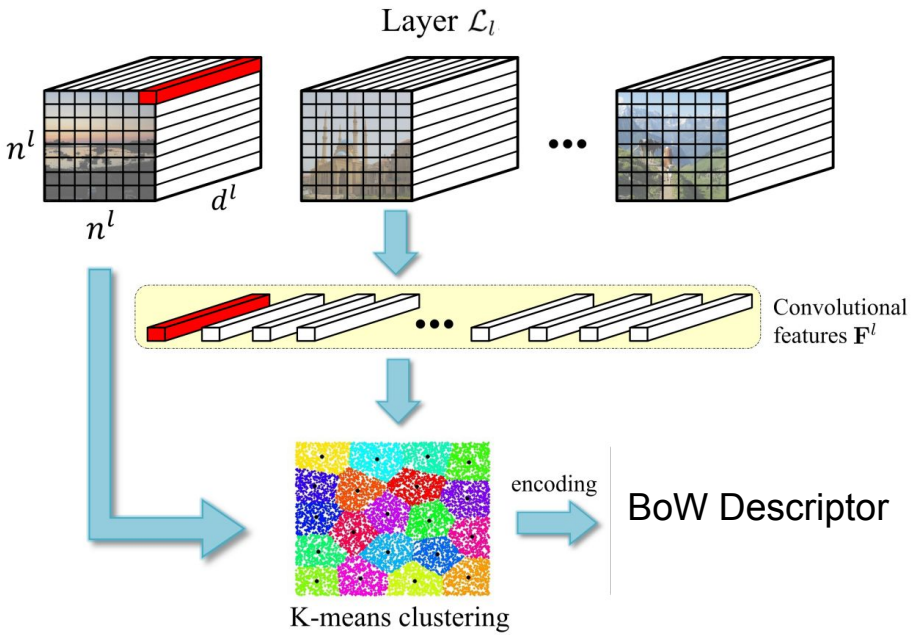
Scalable, fast, high-performance on Oxford 5K, Paris 6K and TRECVid INS



Mohedano et al. **Bags of Local Convolutional Features for Scalable Instance Search**, ICMR 2016 http://arxiv.org/abs/1604.04653

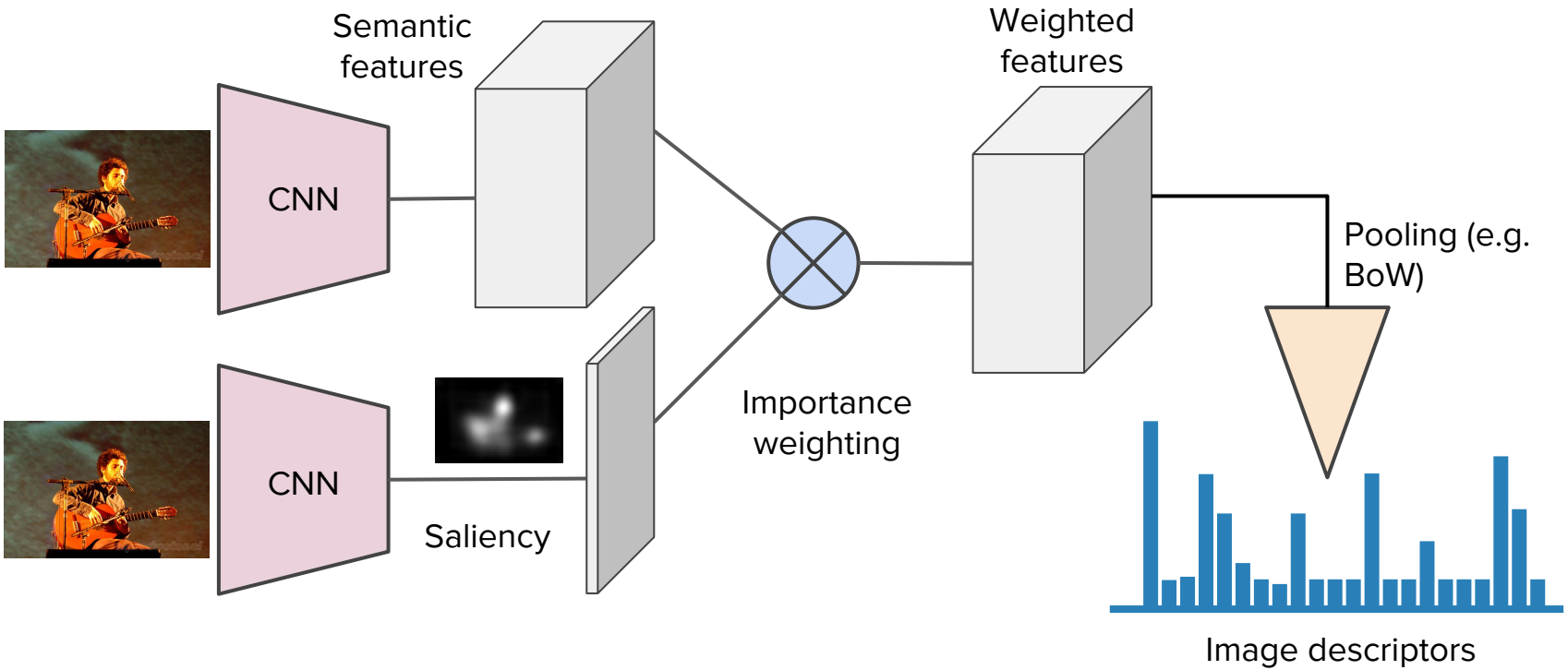
Bags of convolutional features instance search

		Oxford 5k	Paris 6k	INS 23k
BoW	GS	0.650	0.698	**0.323**
	LS	**0.739**	**0.819**	0.295
Sum pooling (as ours)	GS	0.606	0.712	0.156
	LS	0.583	0.742	0.097
Sum pooling (as in [7])	GS	0.672	0.774	0.139
	LS	0.683	0.763	0.120



Mohedano et al. **Bags of Local Convolutional Features for Scalable Instance Search**, ICMR 2016 http://arxiv.org/abs/1604.04653

Using saliency to improve retrieval



Eva Mohedano, Kevin McGuinness, Xavier Giro-i-Nieto, Noel E. O'Connor, **Saliency Weighted Convolutional Features for Instance Search** https://arxiv.org/abs/1711.10795

Saliency weighted retrieval

Table 3. Performance (mAP) of different spatial weighting schemes using the BLCF approach.

Weighting	INSTRE	Oxford	Paris
None	0.636	0.722	0.798
Gaussian	0.656	0.728	0.809
L^2-norm	0.674	0.740	**0.817**
Itti-Koch [14]	0.633	0.693	0.785
BMS [46]	0.688	0.729	0.806
SalNet [27]	0.688	0.746	0.814
SalGAN [26]	**0.698**	**0.746**	0.812
SAM-VGG [8]	0.688	0.686	0.785
SAM-ResNet [8]	0.688	0.673	0.780

Saliency weighted retrieval

Table 6. Performance comparison with the state-of-the-art with average query expansion.

Method	Off-the-shelf	dim	INSTRE	Oxford	Paris
CroW [19]	yes	512	0.613	0.741	0.855
CAM [17]*	yes	512		0.760	0.873
R-MAC [37]	yes	512	0.706	0.770	0.884
R-MAC [30]†	No	512	0.573	0.854	0.884
R-MAC-ResNet [11]†	No	2048	0.705	**0.896**	**0.953**
(ours) BLCF	yes	336	0.679	0.751	0.788
(ours) BLCF-Gaussian	yes	336	0.731	0.778	0.838
(ours) BLCF-SalGAN	yes	336	**0.757**	0.778	0.830

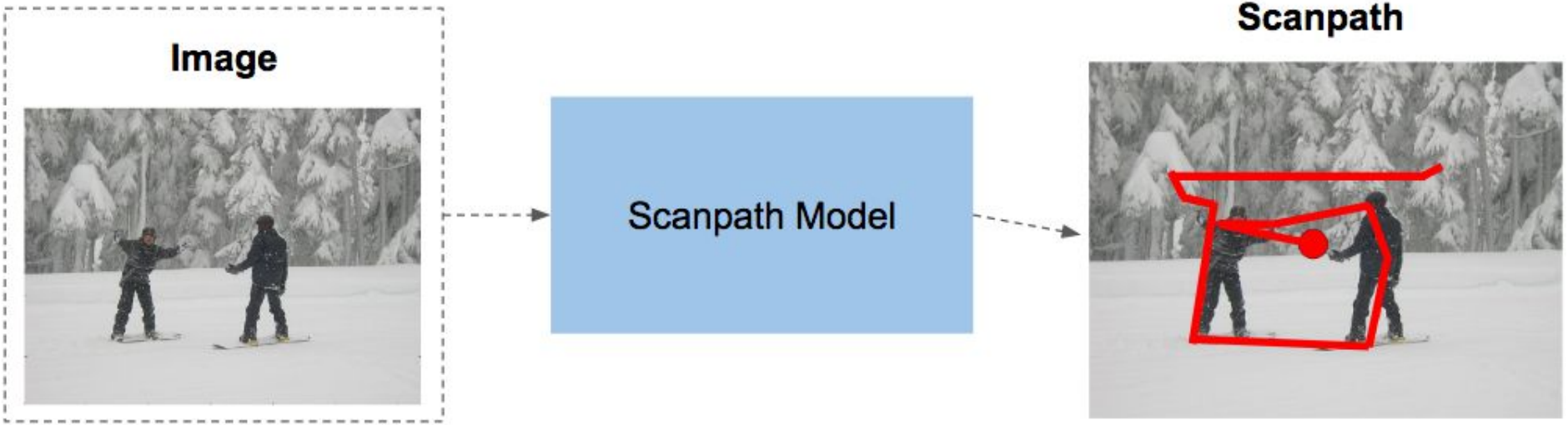
Results marked with (*) are provided by the original publications. Those marked with (†) are reported in Iscen et al. [13]. Otherwise they are based on our own implementation.

We can predict **where** people will look.

What about **when**?

Challenge

Create a model able to predict visual scanpaths



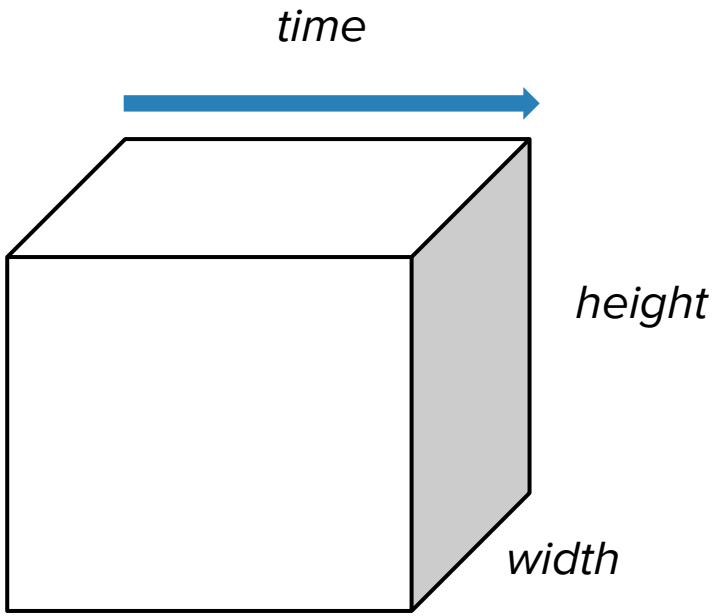
Challenge

Create a model able to predict visual scanpaths

SalTiNet solution

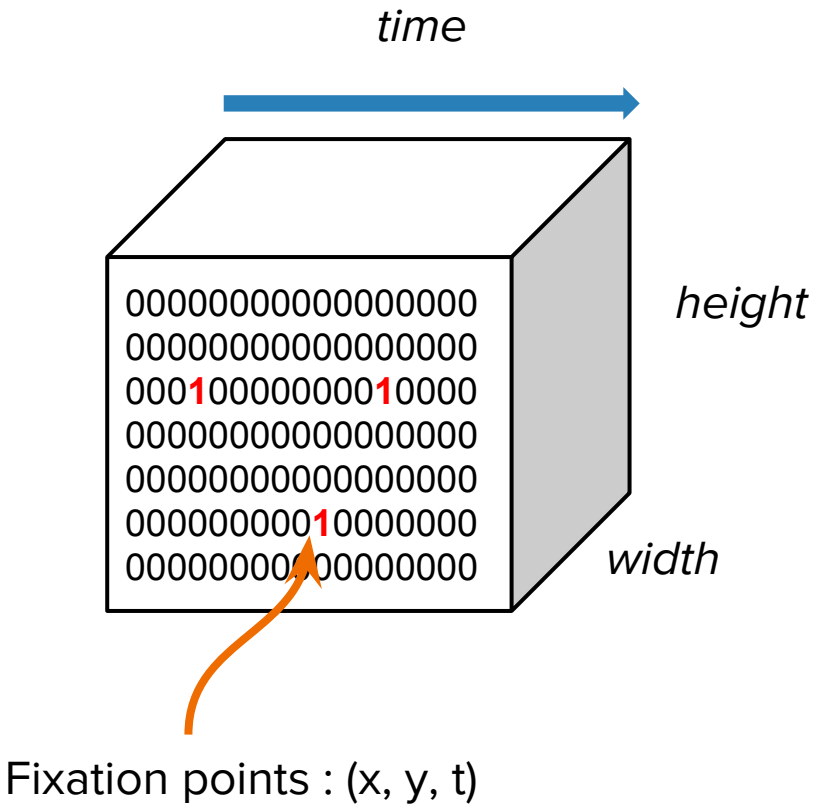
Add a **time** dimension to saliency maps with **saliency volumes** and sample from them.

Saliency volumes



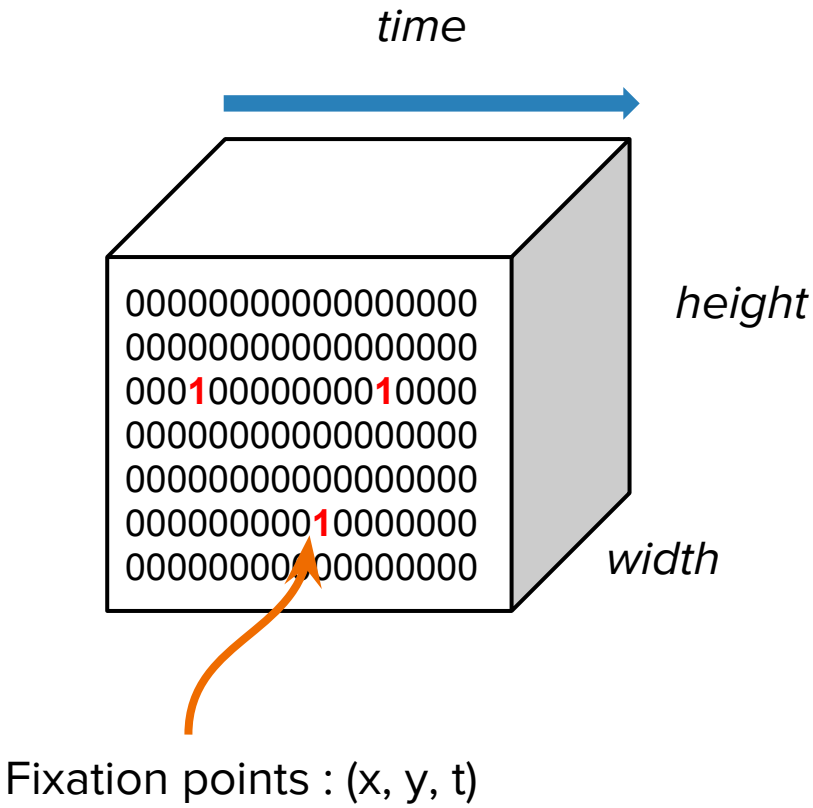
Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. "SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes." ICCVW 2017.

Saliency volumes



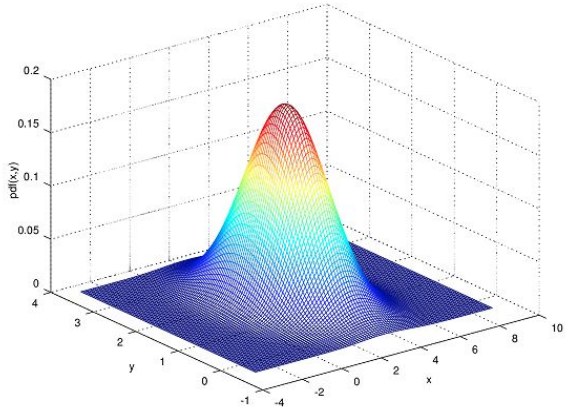
Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. "SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes." ICCVW 2017.

Saliency volumes

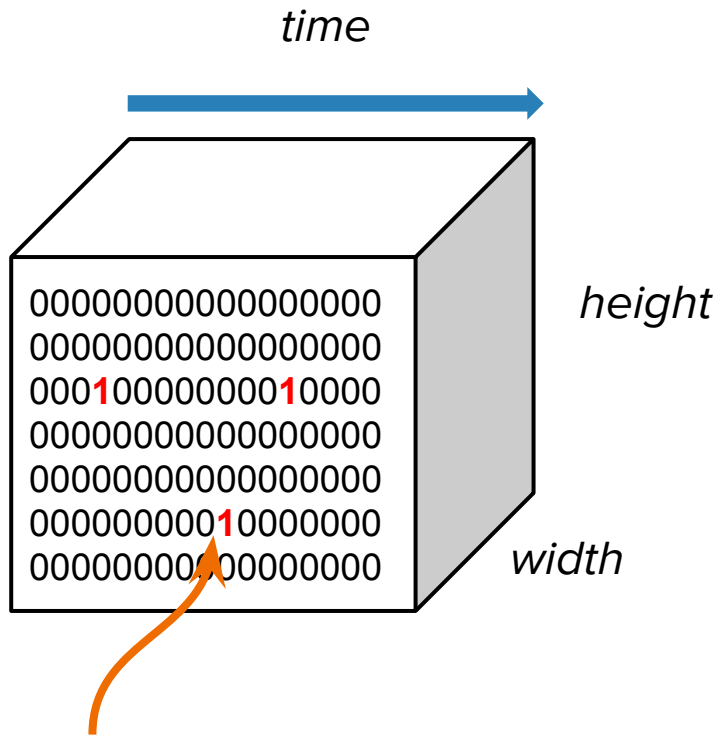


Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. "SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes." ICCVW 2017.

Saliency volumes



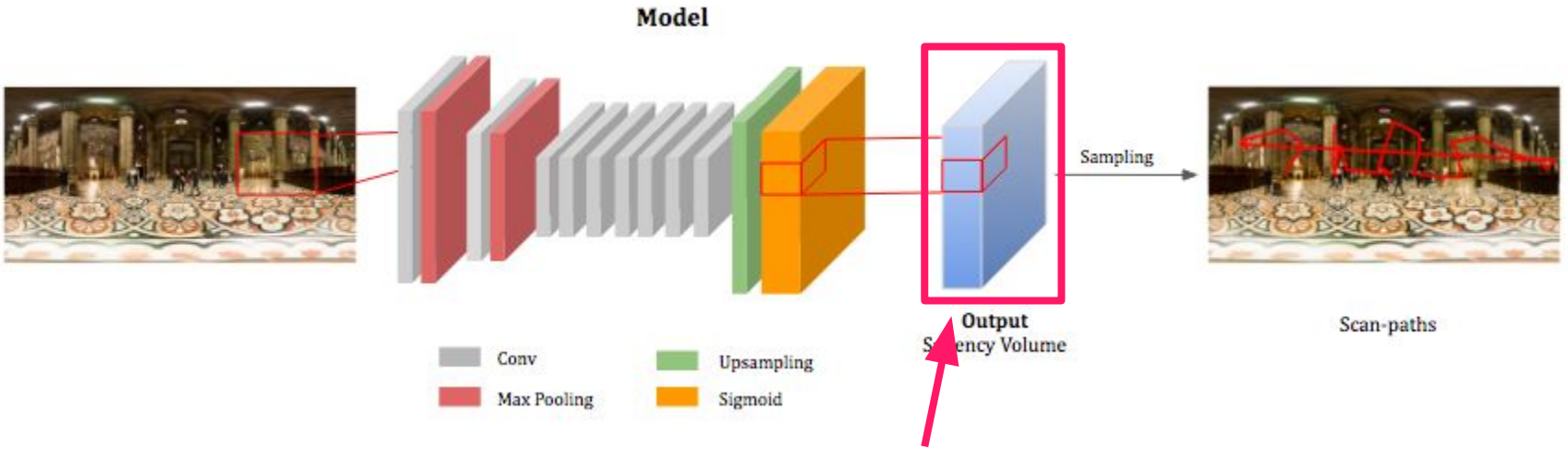
Convolution



Fixation points : (x, y, t)

Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. "SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes." ICCVW 2017.

SalTiNet

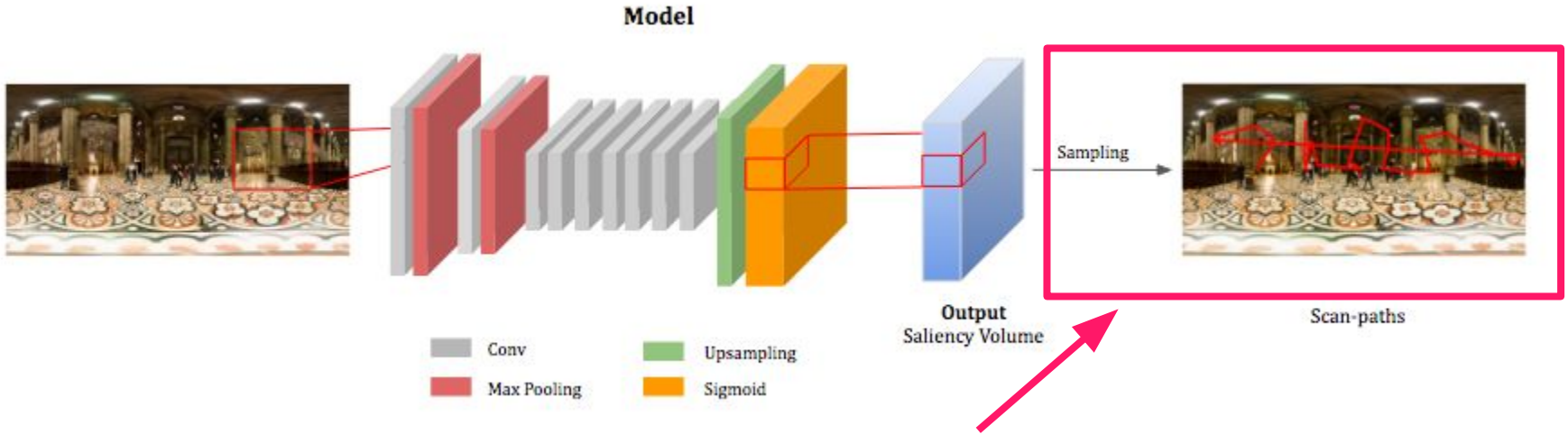


Predicted saliency volume

(a joint probability distribution over x, y, t)

Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. "SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes." ICCVW 2017.

SalTiNet



Draw samples from P(x, y, t) to generate scan paths

Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. "SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes." ICCVW 2017.

Salient 360 challenge

technicolor

Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. 2017. "A Dataset of Head and Eye Movements for 360 Degree Images". MMSys 2017

technicolor

Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. 2017. "A Dataset of Head and Eye Movements for 360 Degree Images". MMSys 2017

Training

Winners of IEEE ICME 2017 Challenge

technicolor

Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. "SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes." ICCVW 2017.

With SalTiNet we **indirectly** predict scan paths by modelling the **joint probability distribution** over spatiotemporal fixation locations and sampling from it.

What about trying to predict scanpaths **directly** from the image?

PathGAN

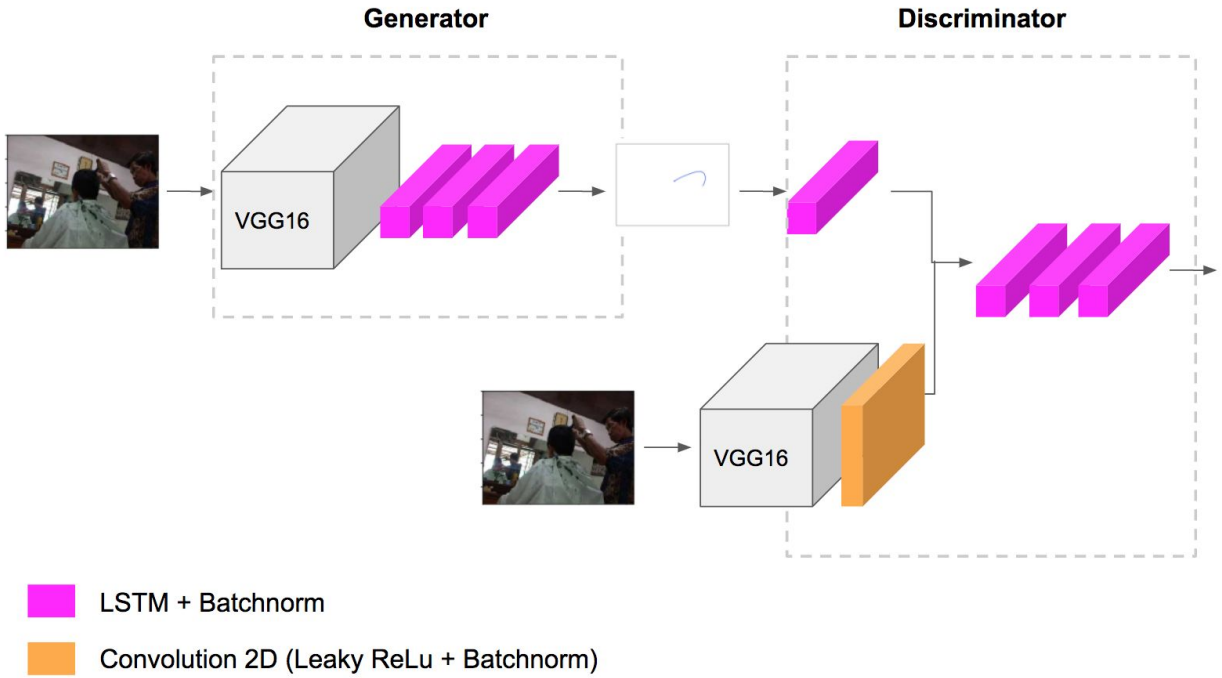
Generate a sequence of fixation points using a **recurrent neural network**.

Challenges:

- **Conditionality**: scanpath depends on the image
- **Stochasticity**: different realistic paths for the same image

Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. PathGan: Visual Scan-path Prediction with Generative Adversarial Networks arXiv 2018.

PathGAN

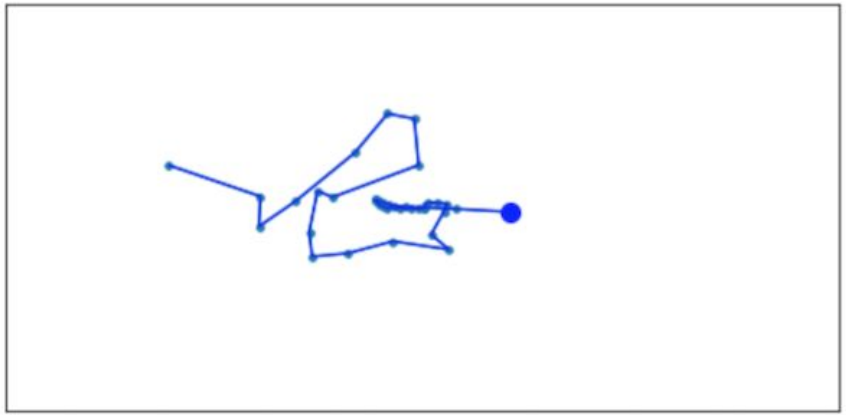


Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. PathGan: Visual Scan-path Prediction with Generative Adversarial Networks arXiv 2018.

PathGAN

Stimuli

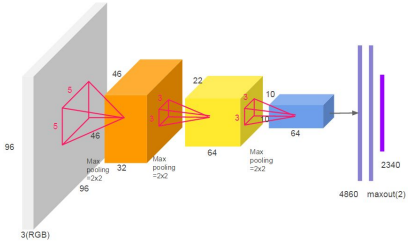
Prediction



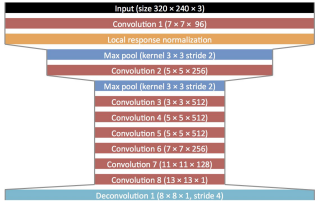
Marc Assens, Kevin McGuinness, Xavier Giro-i-Nieto and Noel E. O'Connor. PathGan: Visual Scan-path Prediction with Generative Adversarial Networks arXiv 2018.

Saliency maps

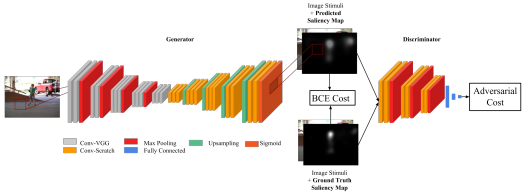
JuntingNet



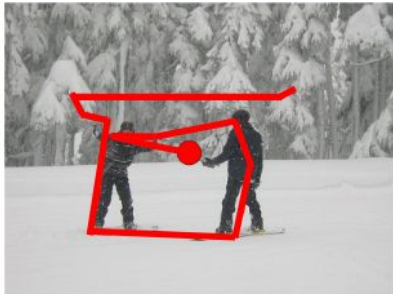
SalNet



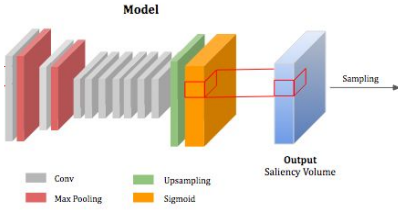
SalGAN



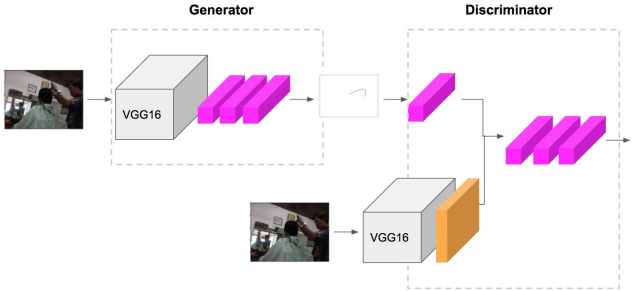
Scanpaths

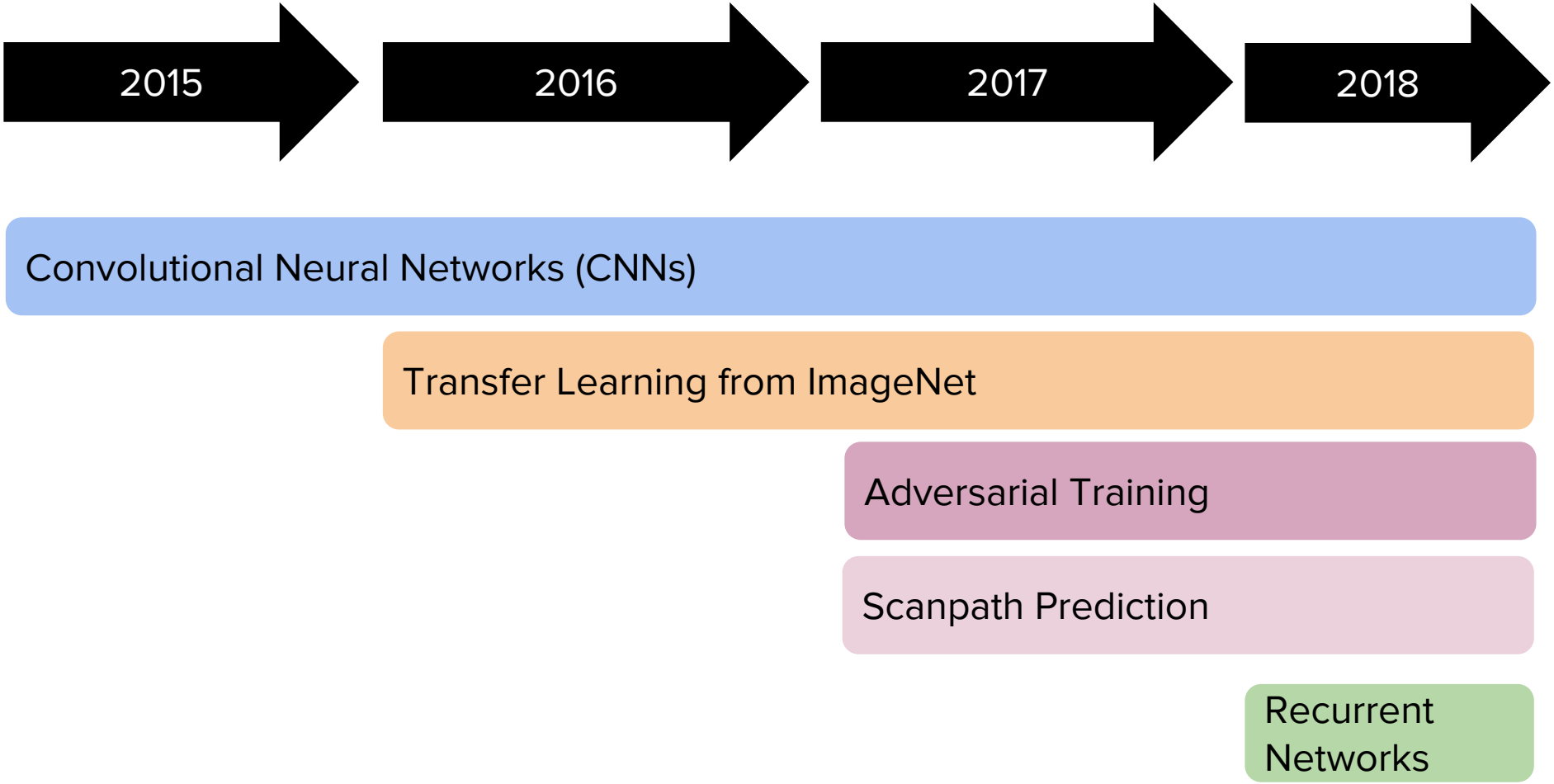


SaltiNet



PathGAN




2015
2016
2017
2018
Convolutional Neural Networks (CNNs)
Transfer Learning from ImageNet
Adversarial Training
Scanpath Prediction
Recurrent Networks

Questions?