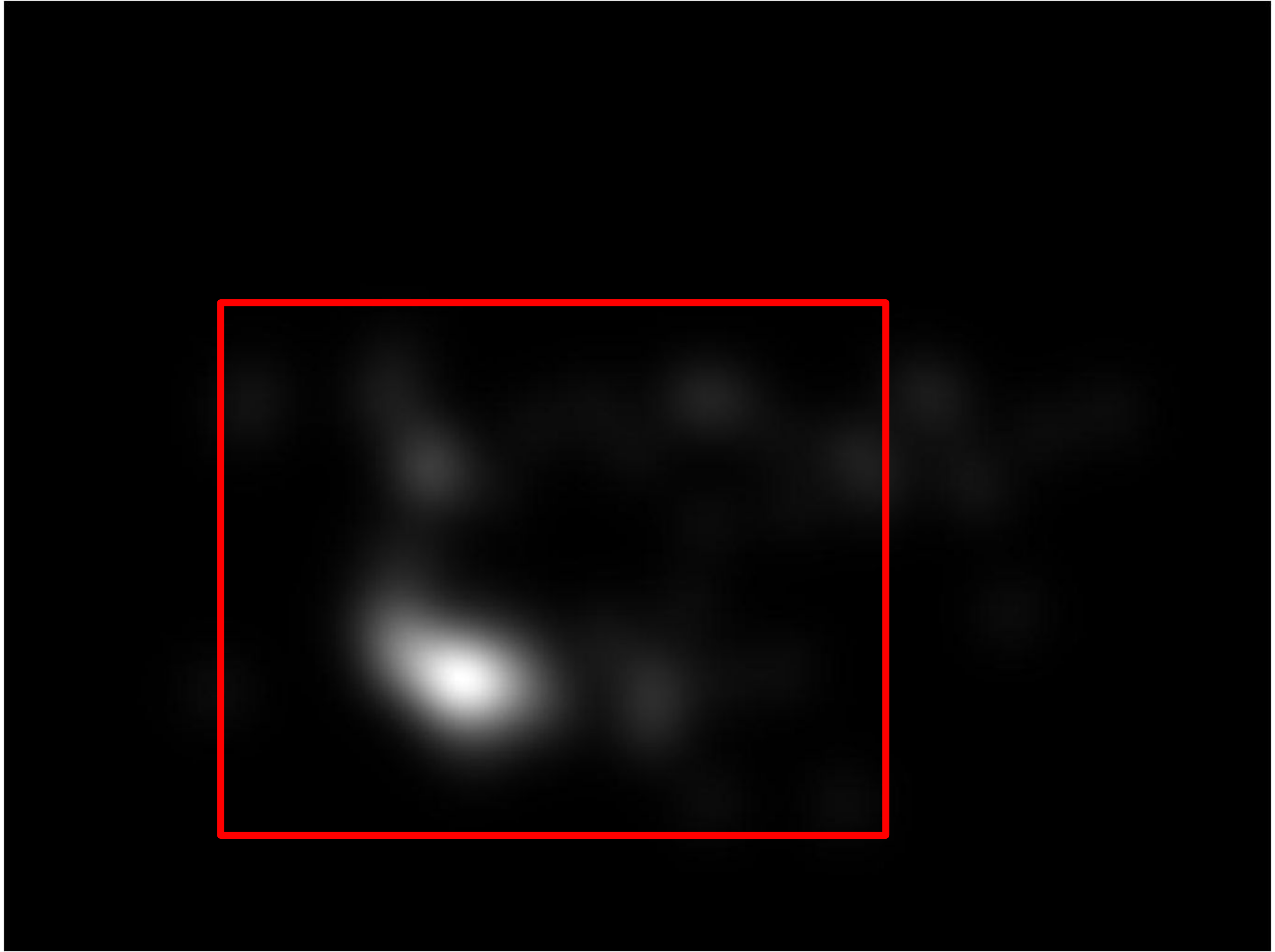
DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations

Kumar Ayush
Bachelor of Technology (Hons.)
Computer Science and Engineering
IIT Kharagpur

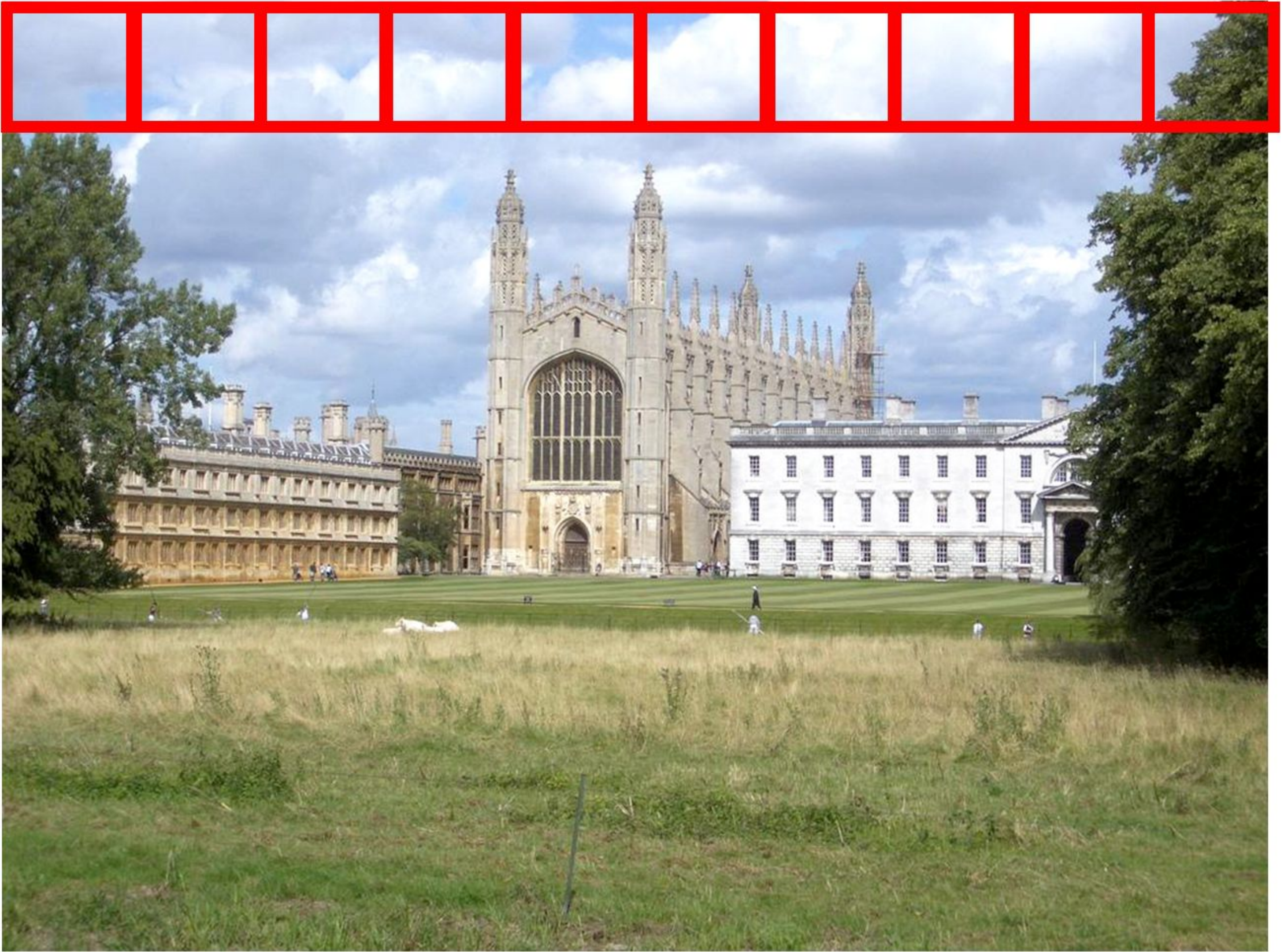
Where do you look on these images?



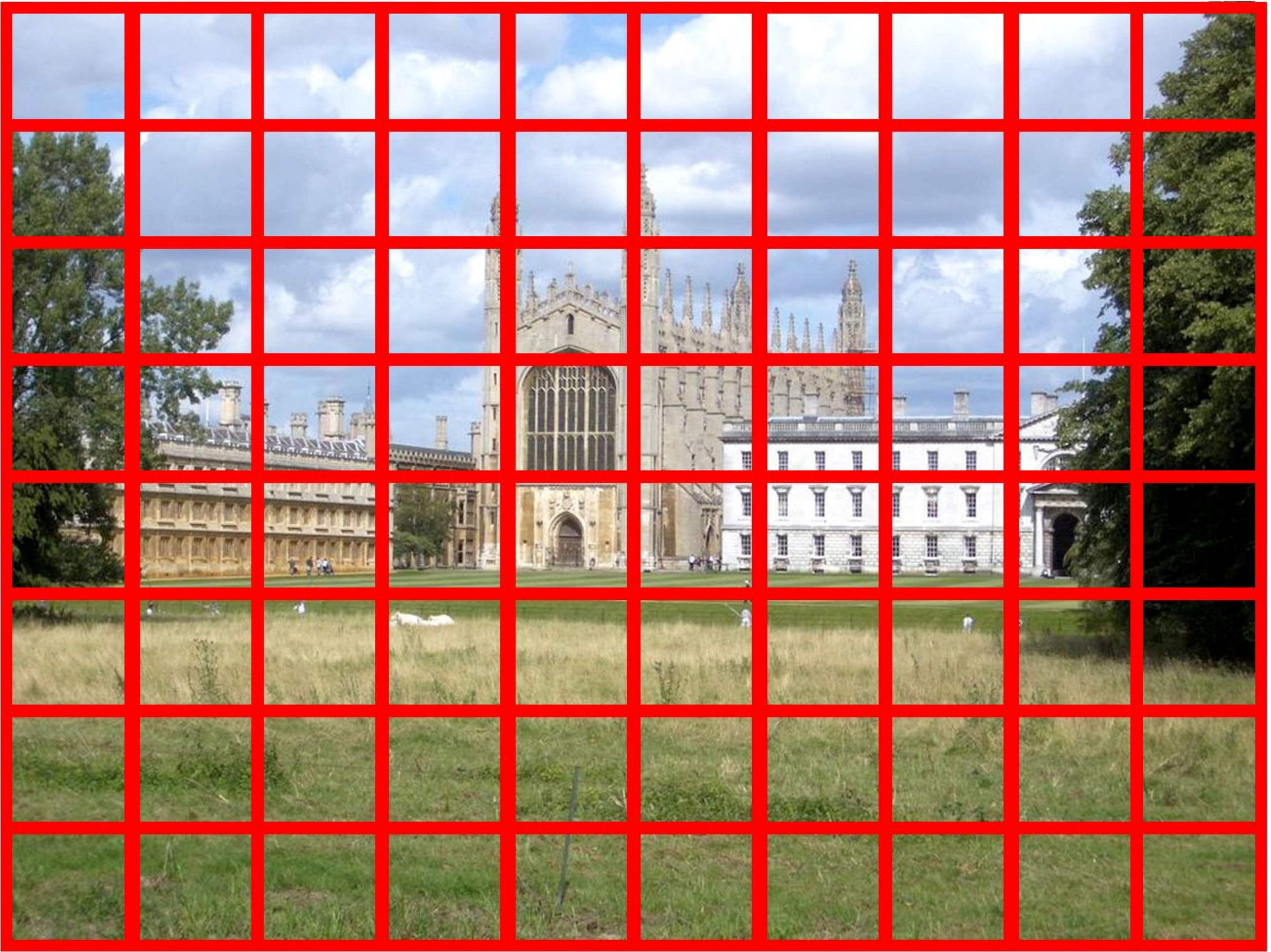
Imagine that you are a robot, and you've received this image from your camera. You need to run some expensive localization computations to decide where you are.

You could do this across the whole image....but you likely don't need to run the computation everywhere.

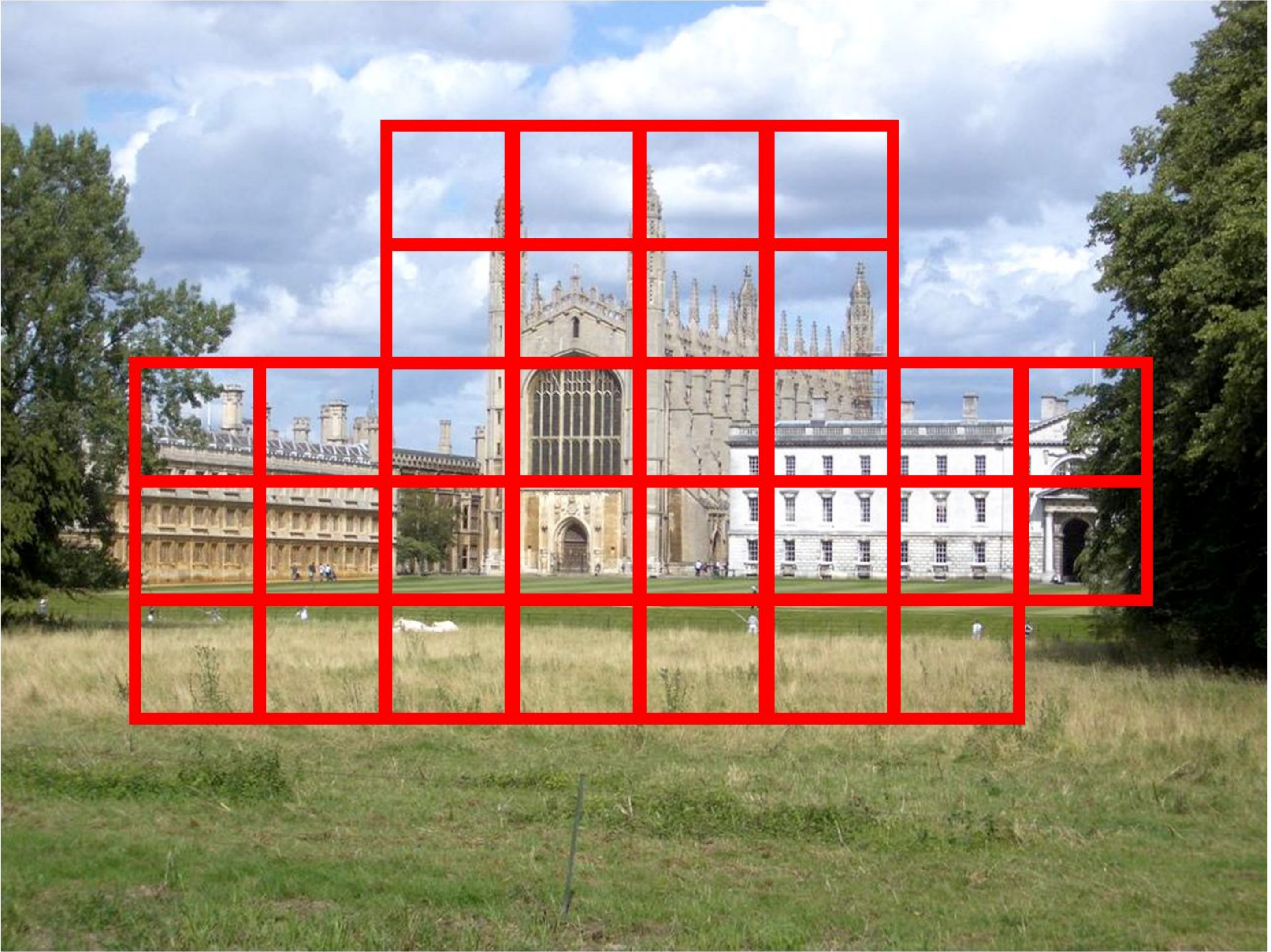
Running computations everywhere would take a long time .



Running computations everywhere would take a long time .



Running computations everywhere would take a long time .



Instead, doing it just here, or doing it here first could save you a lot of time.

Therefore,
need to prioritize the
visual information
and decide what is most important

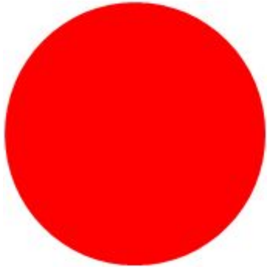
Understanding attention enables applications in computer graphics & vision, design

- Image Cropping/Thumbnailing
- Image and Video Compression
- Non-Photorealistic rendering
- Scene Understanding
- Advertising and Package Design
- Web Usability

- Localization/Recognition
- Object Detection
- Navigational Assistance
- Robot Action Vision
- Surveillance Systems
- Assistive Technology for blind or low-vision people

Where we move our eyes is dictated by two mechanisms

- Bottom-Up Mechanisms
- Top-Down Mechanisms



Visual Attention Mechanisms

Bottom-Up

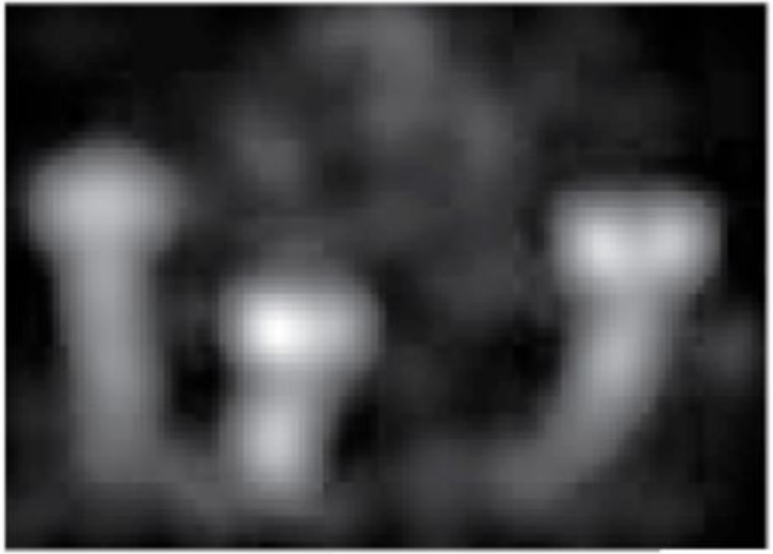
- Automatic
- Reflexive
- Stimulus-driven

Top-Down

- Subject's Prior Knowledge
- Expectations
- Task Oriented
- Memory
- Behavioural Goals

Researchers create computational models of visual attention to predict where people look

Image



Saliency Map

Proposed Solution

DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations

Achieves
state-of-the-art results on multiple challenging saliency data sets

The Ingredients

Very Deep Network

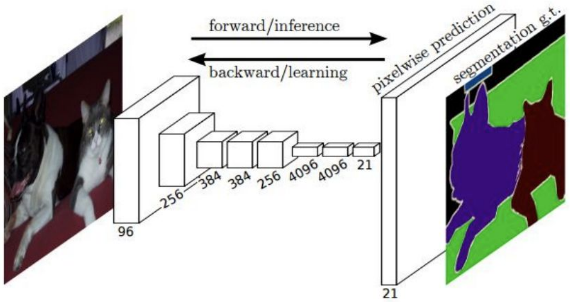
- Inspired by VGGnet (19 layers)
- 20 layers
- Small kernel sizes

K. Simonyan and A. Zisserman. (2014). Very deep convolutional networks for large-scale image recognition. https://arxiv.org/abs/1409.1556

The Ingredients

Fully Convolutional Network

- Fully connected layers at the end are replaced by convolutional layers with very large receptive fields.
- They capture the global context of the scene.
- End-to-end training

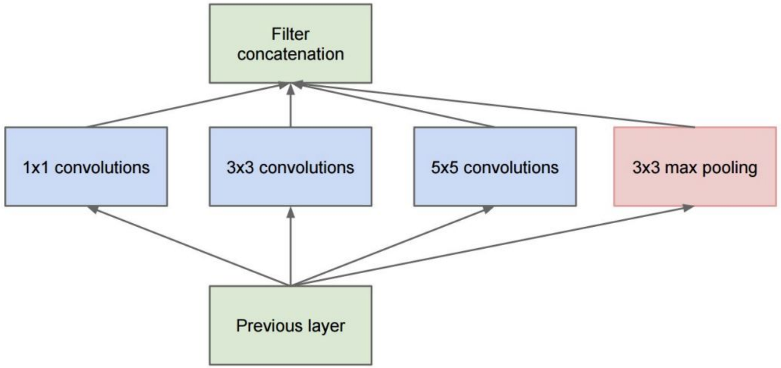


Long, J., Shelhamer, E., & Darrell, T. (2015). Fully Convolutional Networks for Semantic Segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 3431-3440)

The Ingredients

Inception Layers

- GoogLeNet
- Different kernel sizes operating in parallel.



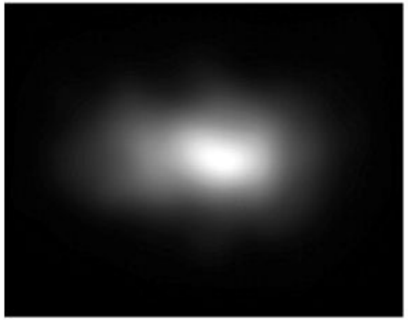
Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., ... & Rabinovich, A. (2015). Going Deeper With Convolutions. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 1-9)

The Ingredients

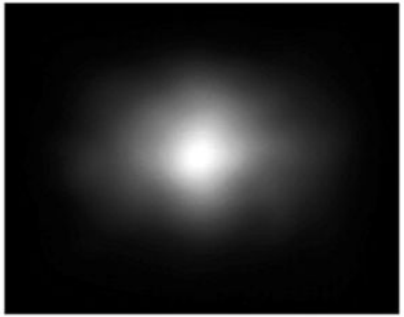
Location Biased Convolutional (LBC) layer

- Human Eye Fixations are Center Biased
  - Photographer Bias
  - Viewing Strategy

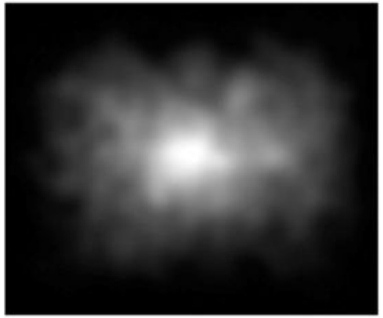
- Introducing LBC layer to model Center Bias



Bruce & Tsotsos (2005)



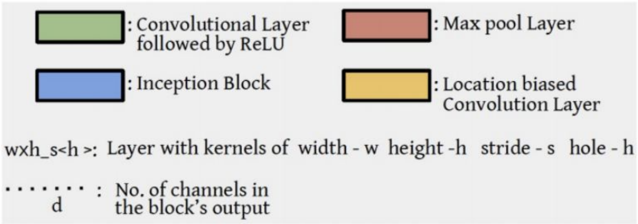
Judd (2009)

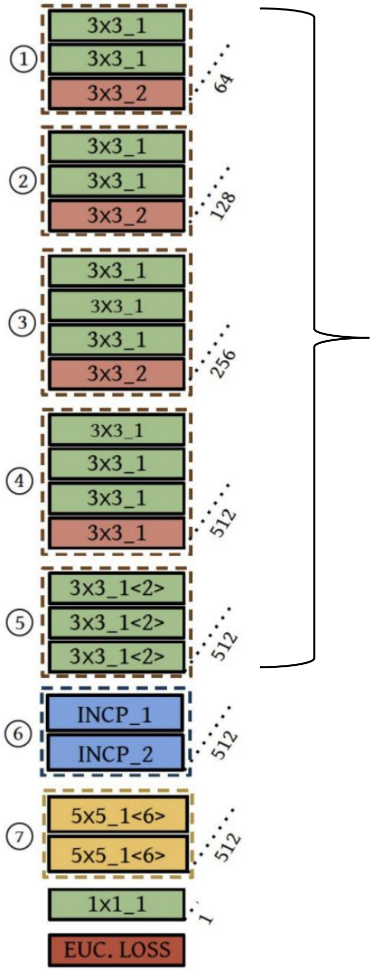


Judd (2011)

The Network

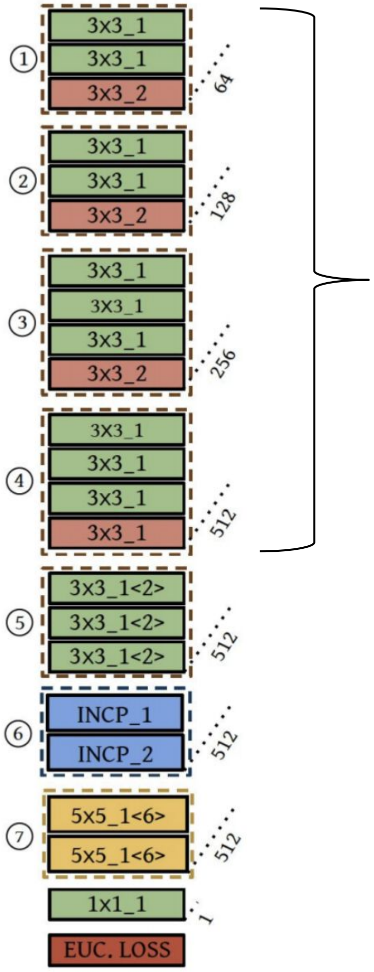
Architecture



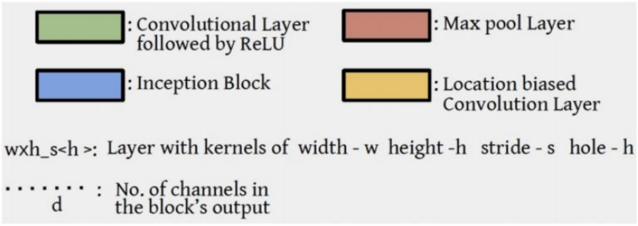


Small convolutional filters of 3x3 with stride of 1 to allow a large depth without increasing the memory requirement

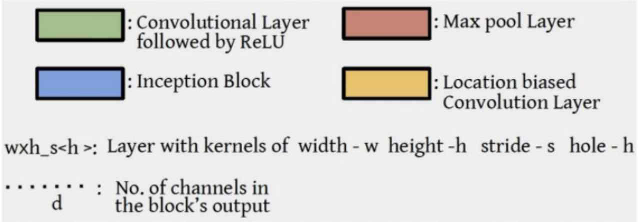
Architecture

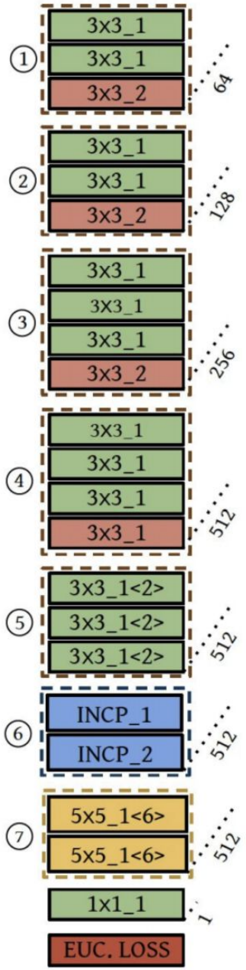


Max pooling layers (in red) reduce computation.



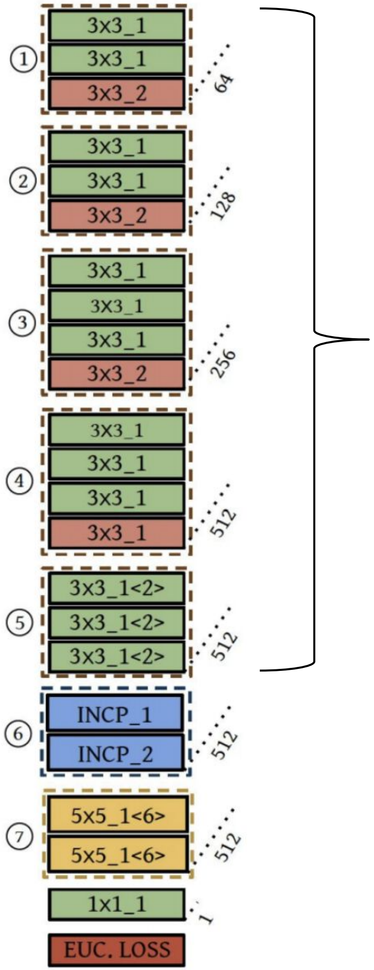
Architecture



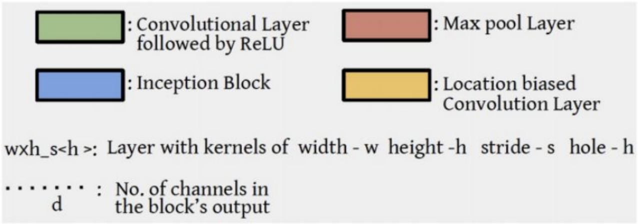


Gradual increase in the amount of channels to progressively learn richer semantic representations: 64, 128, 256, 512...

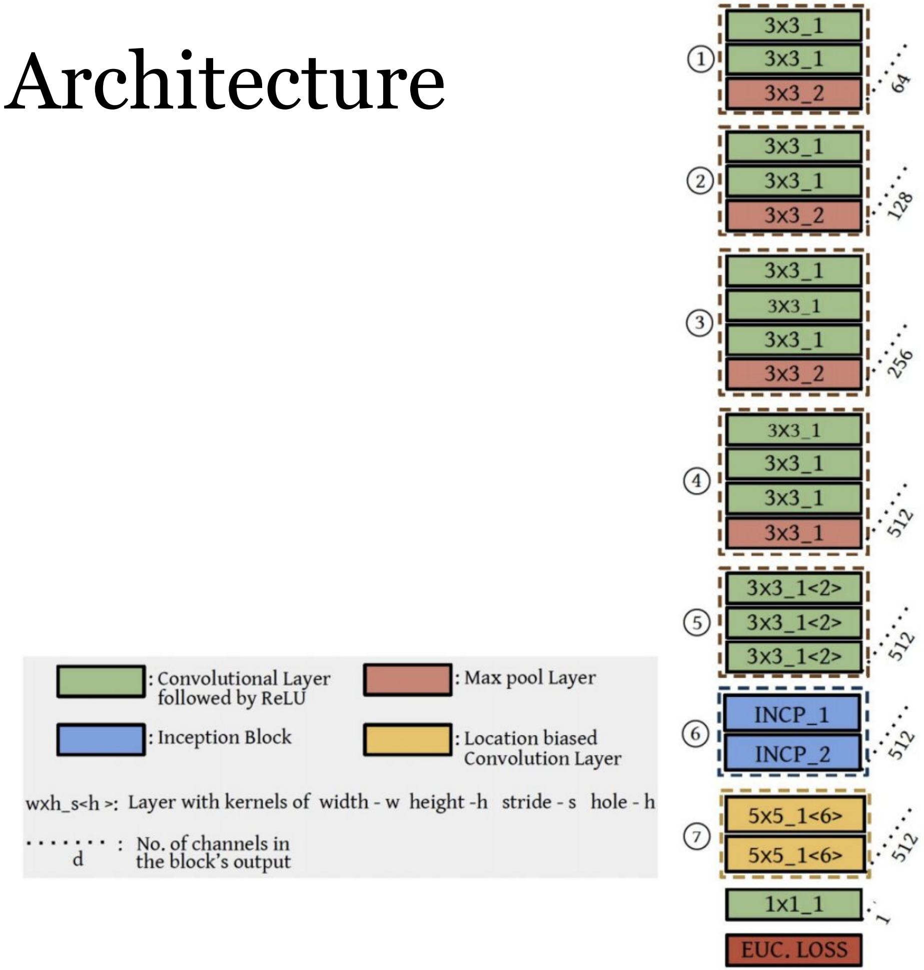
Architecture

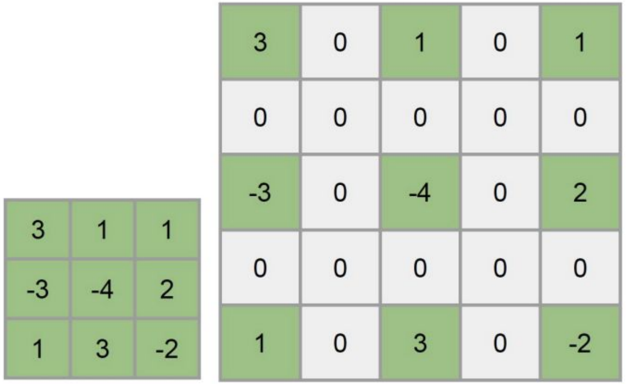


Weights initialized from VGG-16 net for stable and effective learning



Architecture



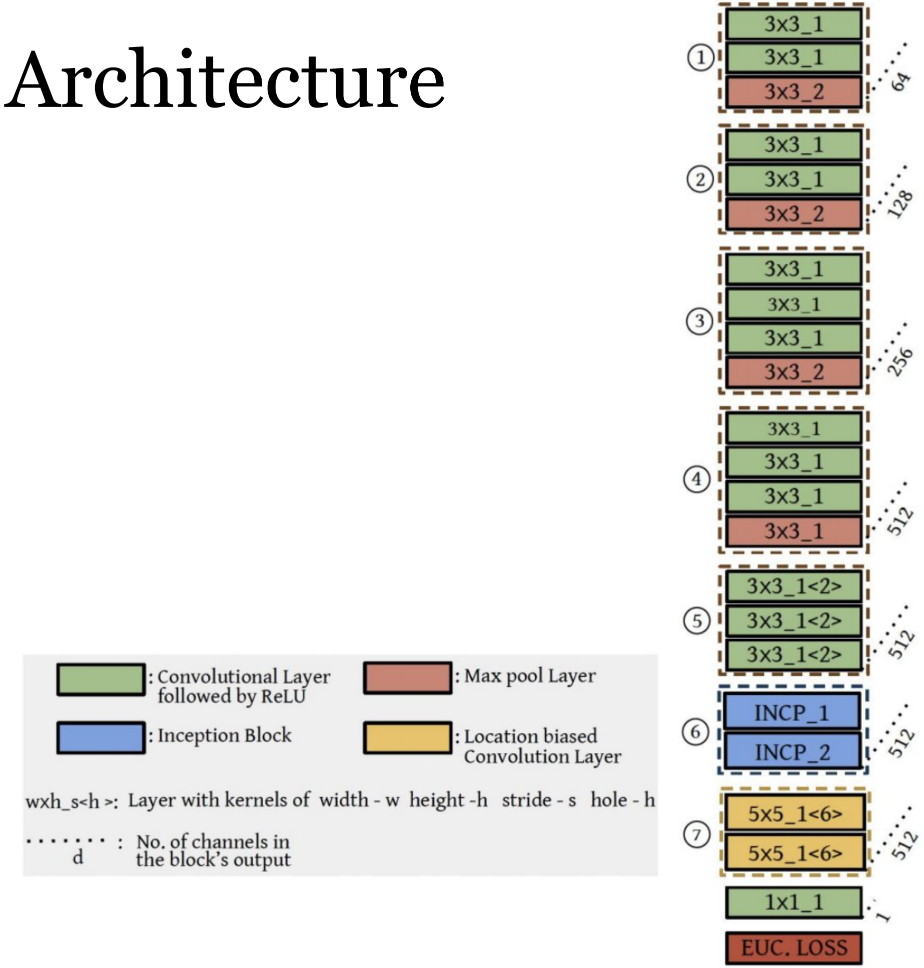


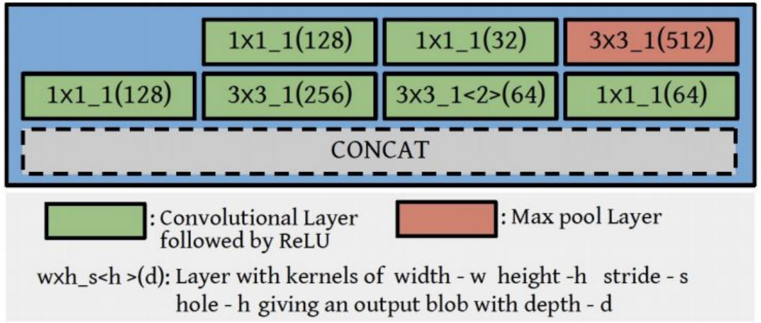
(a) Conv. kernel of size 3x3

(b) Conv. kernel of size 3x3 with hole - 2

Convolution kernel 3x3 with hole size 2 have a receptive field of 5x5

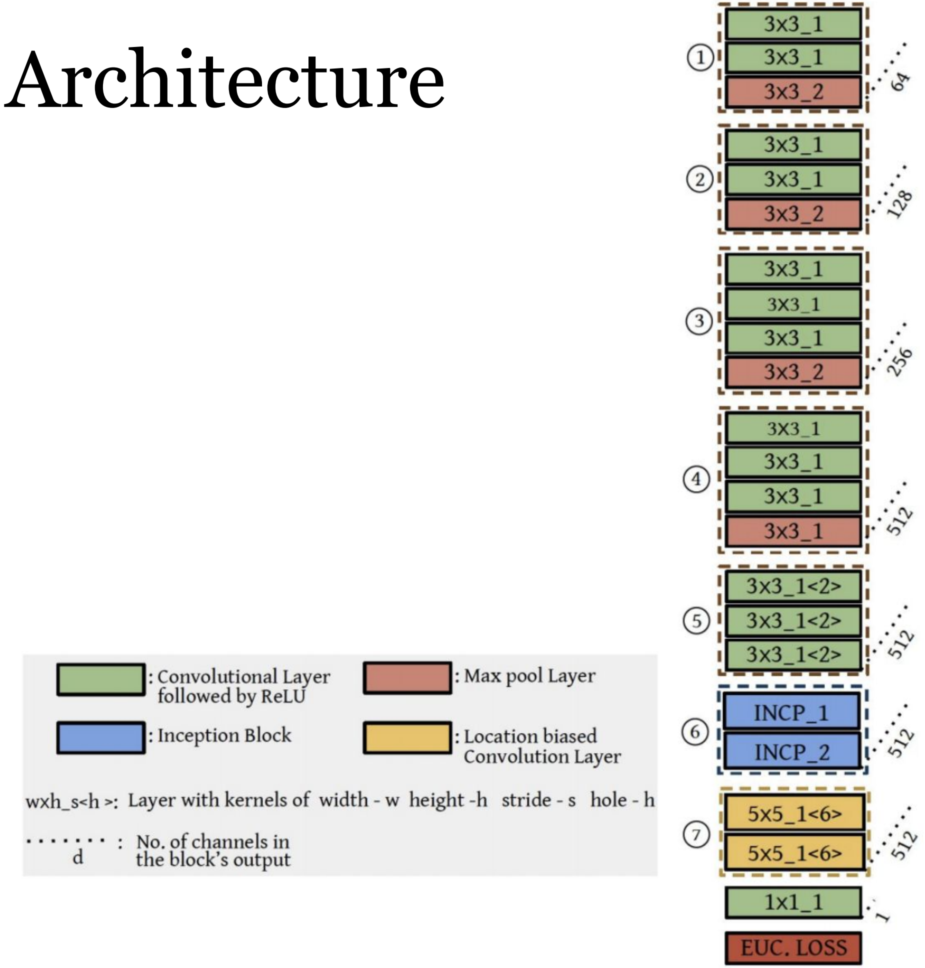
Architecture





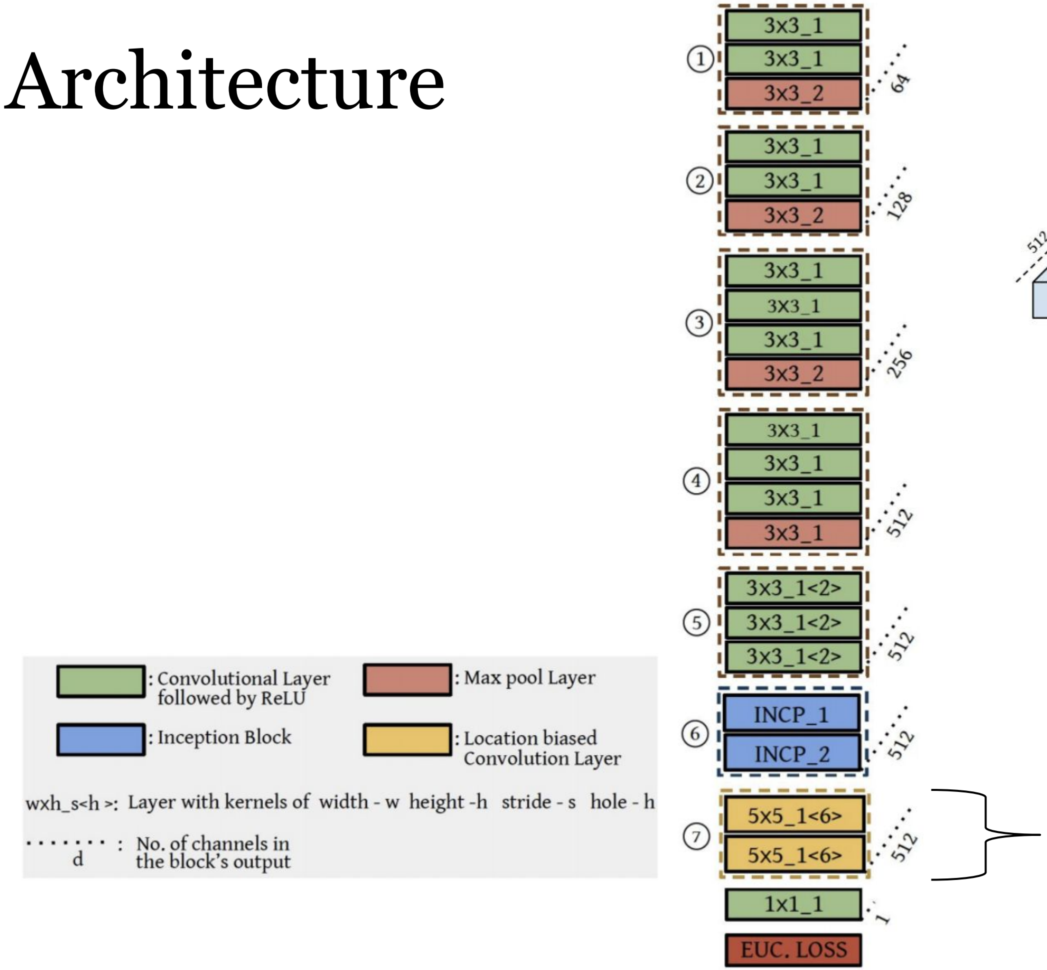
Capture multi-scale semantic structure using two inception style convolutional modules

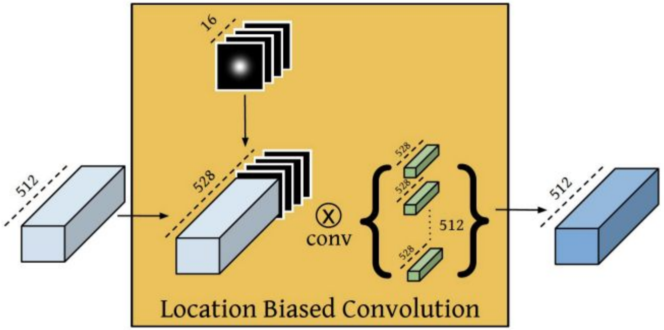
Architecture



Very large receptive fields of 25x25 by introducing holes of size 6 in kernels

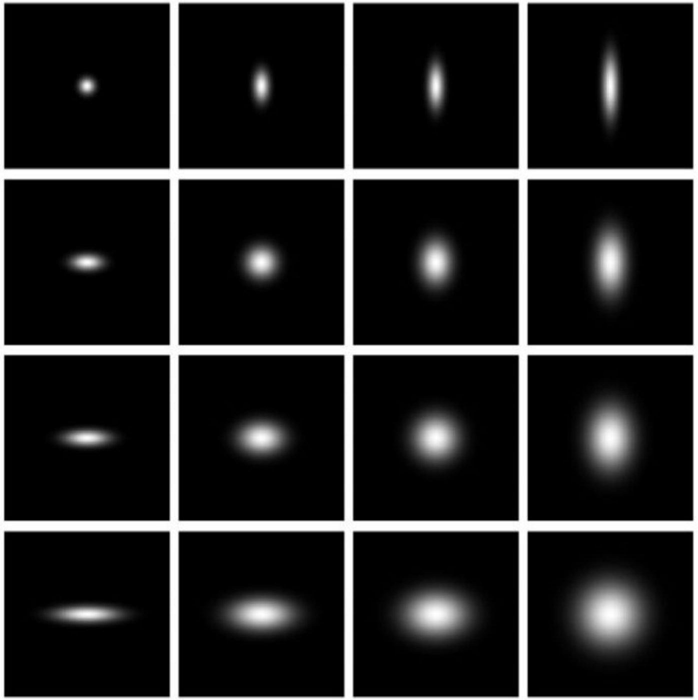
Architecture

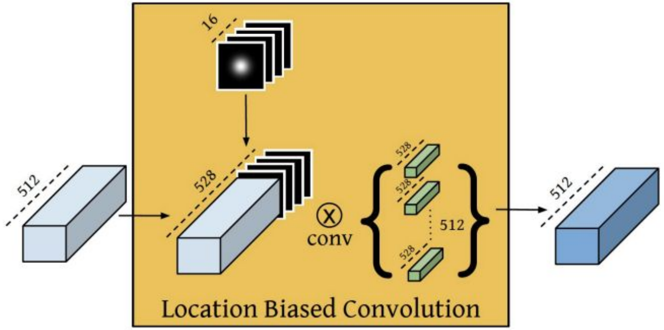




Location Biased Convolutional (LBC) layers

Architecture

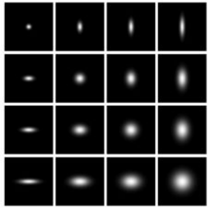




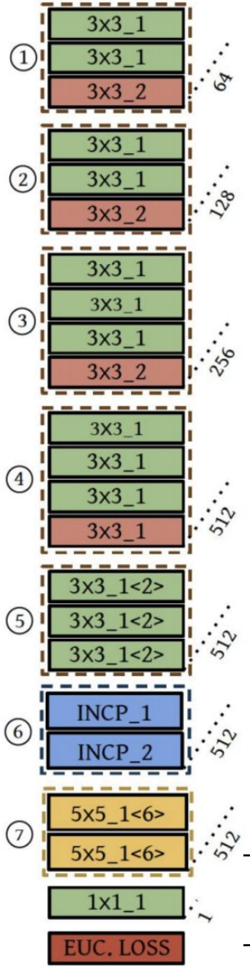
Location Biased Convolutional (LBC) layers

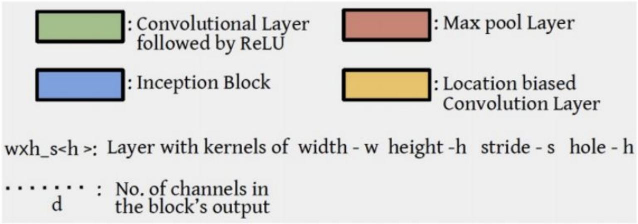
Architecture

$$R_c(x,y) = \mathcal{R}\left(\sum_{i,j}\left(\overbrace{\mathbf{I}(x+i,y+j)}^{\text{input blob}} * \overbrace{\mathbf{W_c}(i,j)}^{\text{weights from c'th filter in a convolutional layer}} + \underbrace{\mathbf{L}(x+i,y+j)}_{\text{constant during training}} * \underbrace{\mathbf{W}'_c(i,j) + b_c}_{\text{learnt during training}}\right)\right)$$



Architecture





Final output W/8xH/8 is upsampled.

Experiments

Results

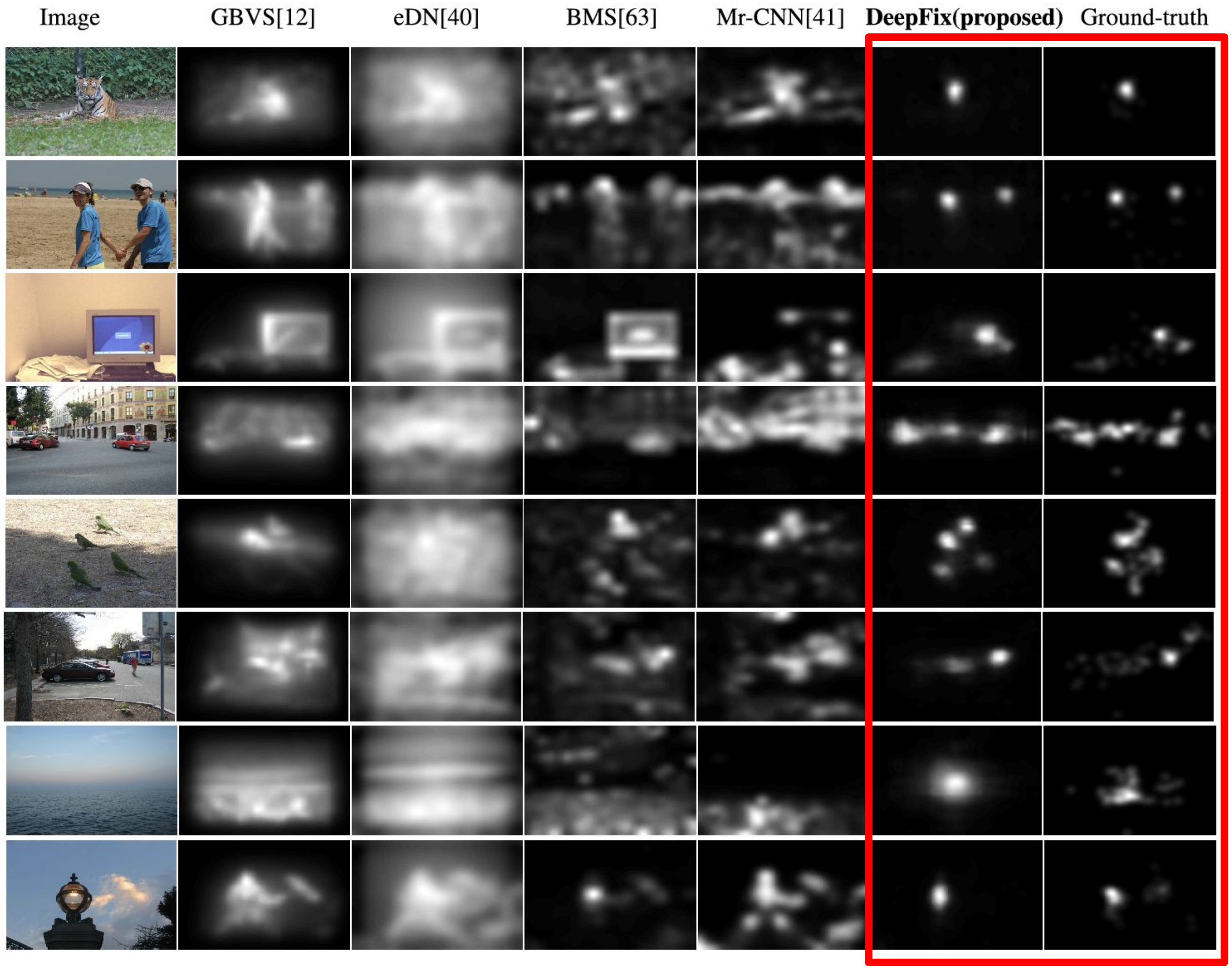


Image
GBVS[12]
eDN[40]
BMS[63]
Mr-CNN[41]
DeepFix(proposed)
Ground-truth

Comparison of Ground Truth and Predicted Saliency Map

Various metrics are used to evaluate the performance of a given Saliency Model

- AUC - Judd
- AUC - Borji
- Shuffled - AUC
- Earth Mover's Distance
- Similarity
- Correlation Coefficient
- Normalized Saliency Scanpath
- Kullback-Leibler divergence

Results

TABLE I

EXPERIMENTAL EVALUATION ON CAT2000 TEST SET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.87**	**0.75**	**1.11**	0.81	0.57	**0.88**	**2.29**
CAS [68]	0.77	0.50	3.09	0.76	**0.60**	0.42	1.07
Judd [57]	0.84	0.46	3.61	**0.84**	0.56	0.54	1.30
GBVS [12]	0.80	0.51	2.99	0.79	0.58	0.50	1.23

TABLE II

EXPERIMENTAL EVALUATION ON MIT300 TEST SET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.87**	**0.67**	**2.04**	0.80	0.71	**0.78**	**2.26**
Salicon [43]	**0.87**	0.60	2.62	**0.85**	**0.74**	0.74	2.12
Mr-CNN [41]	0.77	0.45	4.33	0.76	0.69	0.41	1.13
DG-I [38]	0.84	0.39	4.97	0.83	0.66	0.48	1.22
BMS[63]	0.83	0.51	3.35	0.82	0.65	0.55	1.41
eDN [40]	0.82	0.41	4.56	0.81	0.62	0.45	1.14
CAS [68]	0.74	0.43	4.46	0.73	0.65	0.36	0.95
Judd [57]	0.81	0.42	4.45	0.80	0.60	0.47	1.18
GBVS [12]	0.81	0.48	3.51	0.80	0.63	0.48	1.24

Results

TABLE III

EXPERIMENTAL EVALUATION ON PASCAL-S DATASET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.91**	**0.65**	**0.54**	0.82	**0.73**	**0.78**	**2.60**
Salicon [43]	-	-	-	-	0.72	-	-
SU [47]	0.89	0.59	0.73	0.81	0.72	0.69	2.22
JN [69]	0.88	0.50	1.04	0.86	0.69	0.68	1.90
eDN [40]	0.89	0.39	1.29	**0.87**	0.65	0.55	1.42
BMS[63]	0.80	0.41	1.32	0.78	0.67	0.44	1.28
GBVS [12]	0.84	0.43	1.16	0.82	0.65	0.51	1.36

Results

TABLE IV

EXPERIMENTAL EVALUATION ON OSIE DATASET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.91**	**0.66**	**1.04**	**0.83**	**0.79**	**0.80**	**3.04**
eDN [40]	0.82	0.36	2.02	0.82	0.68	0.40	1.16
BMS[63]	0.83	0.43	1.89	0.82	0.76	0.46	1.47
GBVS [12]	0.82	0.42	1.67	0.80	0.68	0.44	1.35
AWS [70]	0.82	0.42	1.93	0.81	0.76	0.45	1.45

TABLE V

EXPERIMENTAL EVALUATION ON FIGRIM DATASET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.90**	**0.66**	**1.10**	0.84	**0.67**	**0.80**	**2.51**
eDN [40]	0.87	0.37	2.88	**0.86**	0.62	0.50	1.38
BMS[63]	0.76	0.38	3.00	0.73	0.64	0.34	1.05
GBVS [12]	0.82	0.43	2.29	0.81	0.62	0.45	1.26
AWS [70]	0.72	0.36	3.20	0.74	0.64	0.29	0.89

Thanks!