SAM: Pushing the Limits of Saliency Prediction Models

Marcella Cornia[1], Lorenzo Baraldi[1], Giuseppe Serra[2] and Rita Cucchiara[1]

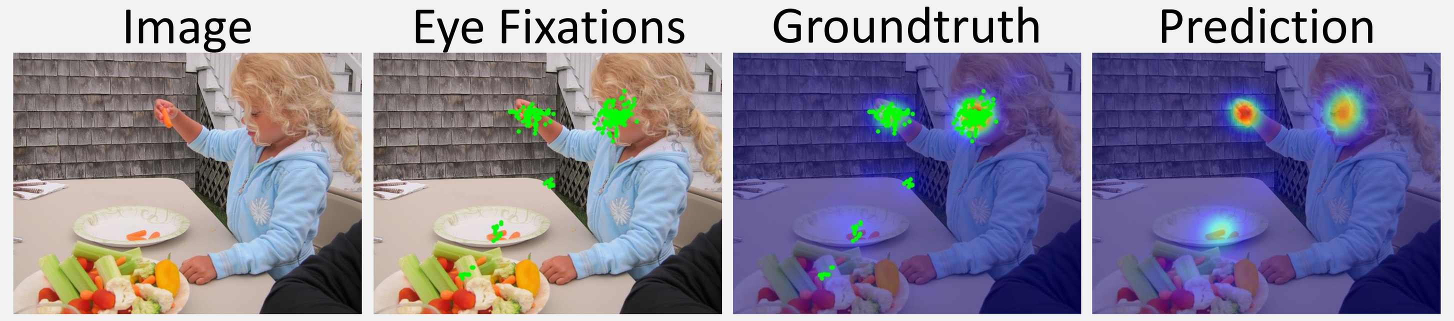
[1]University of Modena and Reggio Emilia, Italy [2]University of Udine, Italy

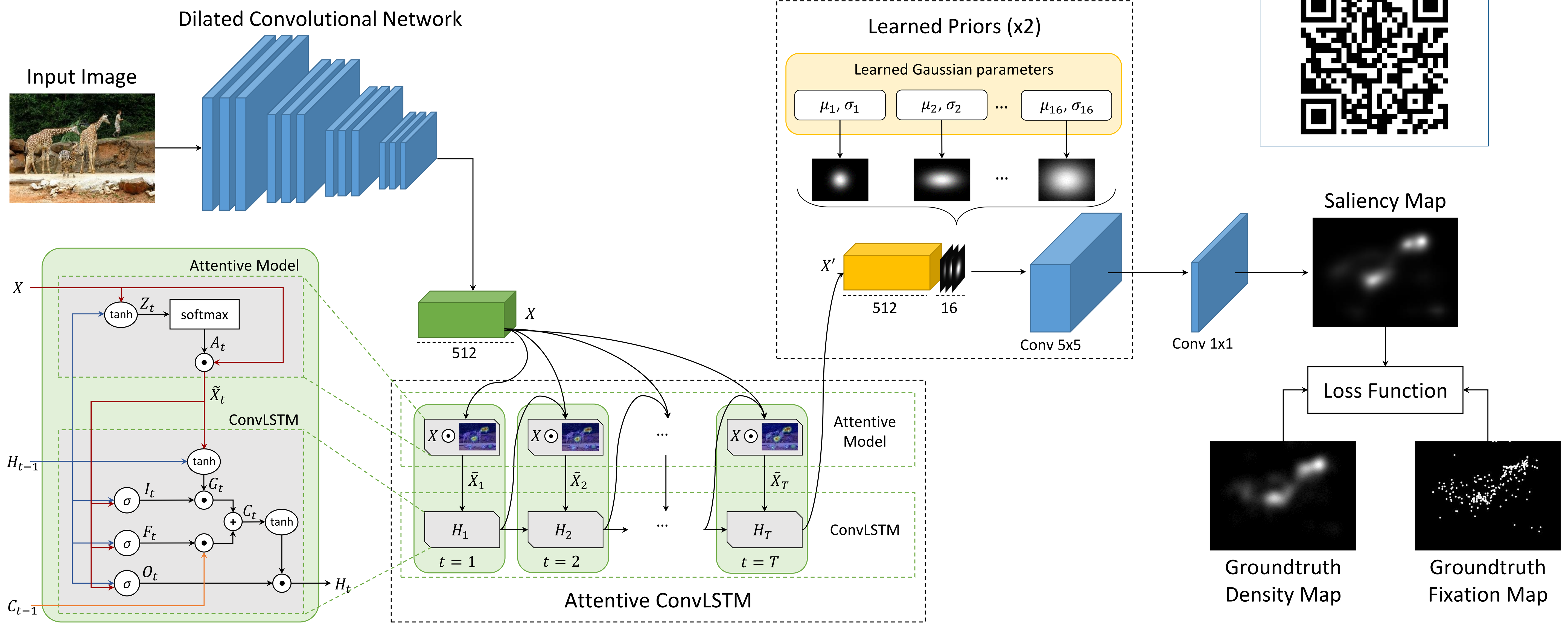
e-mail: marcella.cornia@unimore.it

Source code available at:

Overview

- **Visual saliency prediction** aims to predict where humans gazes will focus on a given image.
- Groundtruth data is collected by means of eye-tracking glasses or mouse clicks to get **eye fixation points**, which are then smoothed together to obtain the saliency map.





Saliency Attentive Model (SAM)

Attentive ConvLSTM

- Extension of the traditional LSTM to work on spatial features by substituting dot products with convolutional operations.
- Exploitation of the sequential nature of the LSTM to process features in an iterative way, without the concept of time.

The input of the LSTM layer $\tilde{X}_t$ is computed through an **attentive mechanism** which produces an attention map from the previous hidden state H_{t-1} of the LSTM and the input X

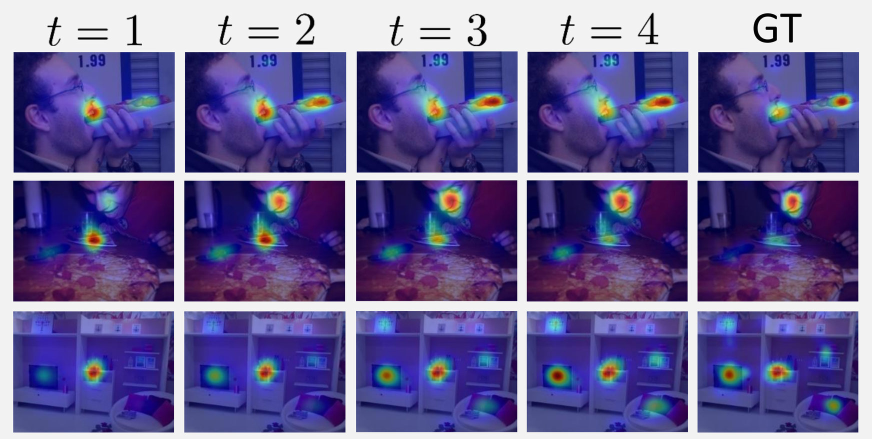
$$Z_t = V_a * \tanh(W_a * X + U_a * H_{t-1} + b_a)$$

The output of this operation is a 2-d map from which we compute a normalized spatial attention map A_t through the *softmax* operator.

The attention map is applied to the input with an element-wise product between each channel of the feature maps and the attention map

$$\tilde{X}_t = A_t \odot X$$

Progressive refinement of saliency maps

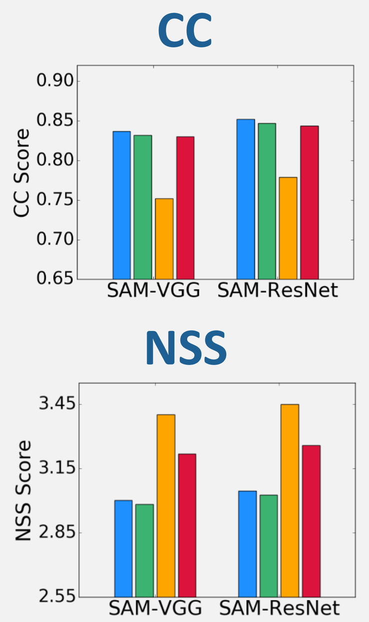


Learned Priors

- Our network is able to learn the **center bias** present in eye fixations, without integrating this information manually.
- The model learns means and variances of a set of Gaussian functions with diagonal covariance matrix and produces a prior map for each function.

Loss Function

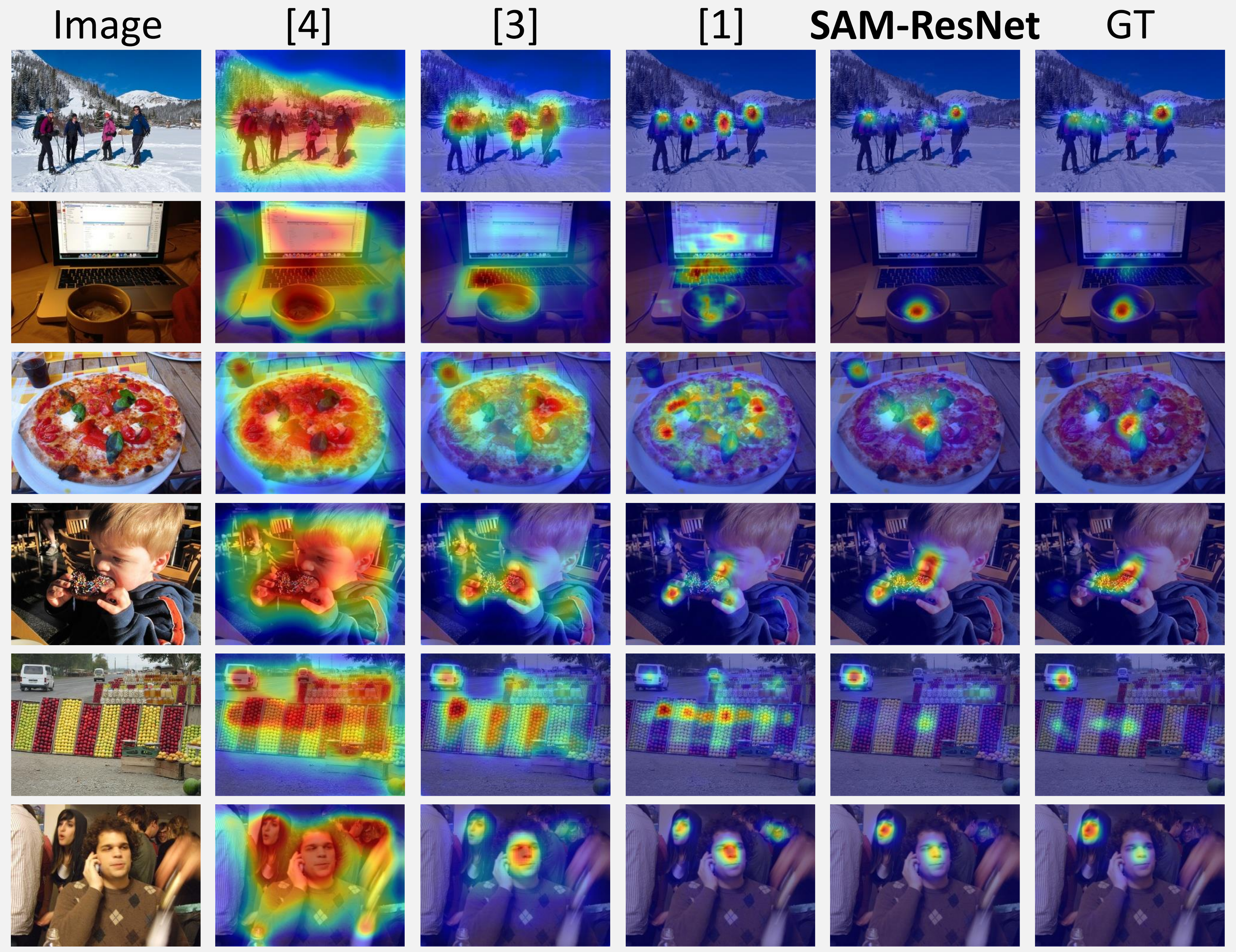
To take different quality aspects into account, we define a new loss function given by a linear combination of three saliency evaluation metrics: the **NSS**, the **CC** and the **KL-Div**.



Dilated Convolutional Network

- We build two different versions of our model based on **VGG-16** and **ResNet-50**.
- To limit rescaling, we use dilated convolutions thus obtaining saliency maps rescaled by a factor of 8 instead of 32.

Qualitative Results



Experimental Results

Results on SALICON 2015

	CC	AUC	NSS
SAM-Resnet	**0.84**	0.88	**3.20**
SAM-VGG	0.83	0.88	3.14
ML-Net [1]	0.74	0.87	2.79
SalGAN [2]	0.78	0.78	2.46
SalNet [3]	0.62	0.86	1.86
DeepGazeII [4]	0.51	**0.89**	1.34

Results on SALICON 2017

	CC	AUC	NSS
SAM-ResNet	**0.90**	**0.87**	**1.99**
SAM-VGG	0.89	0.86	1.97
EAD [5]	0.87	0.85	1.89
SalGAN [2]	0.84	0.86	1.82
SalNet [3]	0.76	0.84	1.56
SALICON [6]	0.66	0.81	1.56

- **First place in the LSUN Saliency Prediction Challenge (CVPR 2017)!**
- State of the art on both versions of SALICON, the largest dataset available for saliency.
- Very good results on several other datasets such as MIT300, MIT1003, CAT2000, DUT-OMRON, TORONTO and PASCAL-S.

Results on MIT1003, DUT-OMRON, TORONTO and PASCAL-S

	MIT1003			DUT-OMRON			TORONTO			PASCAL-S		
	CC	AUC	NSS	CC	AUC	NSS	CC	AUC	NSS	CC	AUC	NSS
GBVS [7]	0.42	0.83	1.38	0.53	0.87	1.71	0.57	0.83	1.52	0.45	0.84	1.36
Mr-CNN [8]	0.38	0.80	1.36	-	-	-	0.49	0.80	1.41	-	-	-
DVA [9]	0.64	0.87	2.38	0.67	0.91	3.09	0.72	**0.86**	2.12	0.66	0.89	2.26
SAM-VGG$_{2015}$	0.61	0.88	2.25	0.65	0.91	2.91	0.69	**0.86**	2.14	0.72	**0.90**	**2.48**
SAM-VGG$_{2017}$	0.65	**0.89**	2.33	0.69	0.91	2.95	**0.74**	**0.86**	**2.15**	0.73	0.89	2.31
SAM-ResNet$_{2015}$	0.65	0.88	**2.48**	0.69	0.91	**3.21**	0.69	**0.86**	2.12	0.69	0.89	2.34
SAM-ResNet$_{2017}$	**0.66**	**0.89**	2.35	**0.70**	**0.92**	2.97	**0.74**	**0.86**	2.14	**0.74**	**0.90**	2.34

References:

[1] Cornia, et al. "A Deep Multi-Level Network for Saliency Prediction." *ICPR*, 2016.
[2] Pan, et al. "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks." *CVPR Workshops*, 2017.
[3] Pan, et al. "Shallow and Deep Convolutional Networks for Saliency Prediction." *CVPR*, 2016.
[4] Kümmerer, et al. "Understanding Low- and High-Level Contributions to Fixation Prediction." *ICCV*, 2017.
[5] He, et al. "What Catches the Eye? Visualizing and Understanding Deep Saliency Models." *arXiv*, 2018.
[6] Huang, et al. "SALICON: Reducing the Semantic Gap in Saliency Prediction." *ICCV*, 2015.
[7] Harel, et al. "Graph-based Visual Saliency." *NIPS*, 2006.
[8] Liu, et al. "Predicting Eye Fixations using Convolutional Neural Networks." *CVPR*, 2015.
[9] Wang, et al. "Deep Visual Attention Prediction." *IEEE Trans. on Image Processing*, 2018.