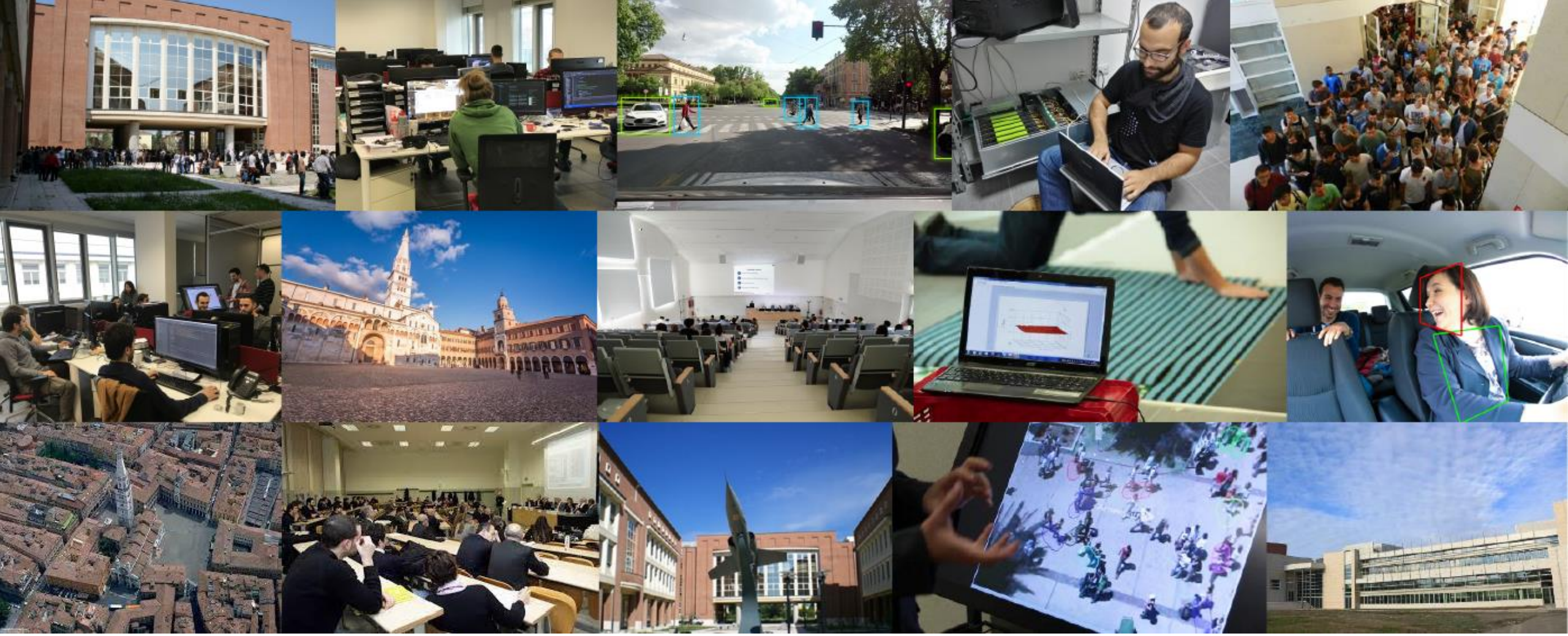


Predicting Human Eye Fixations

L. Baraldi, D. Abati, M. Cornia, R. Cucchiara

University of Modena and Reggio Emilia

AImage Lab

UNIMORE UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

WELCOME!

AImage Lab

Outline of this lab session

Introduction and theory (14:30-16:00)

- *Predicting human-eye fixations: introduction*
- *Task-agnostic saliency: a Saliency Attentive Model*
- *Task-driven saliency: the DR(eye)VE project*

Practice (16:00-18:00)

- *Hands-on introduction to PyTorch*
- *Know what you're doing: writing forward/backward passes*
- *Implementing an Attentive Convolutional LSTM*

WELCOME!

AImage Lab

What you will learn

- The basics of human fixation and saliency prediction
- State of the art approaches, in depth
- Best practices for programming in PyTorch
- Forward/backward implementations of the essential building blocks of a CNN
- Working with attention and sequences with PyTorch

About us

AImage Lab

Who

- 4 Staff people (Rita Cucchiara, Costantino Grana, Roberto Vezzani, Simone Calderara)
- 8 Phd Students
- 7 Research assistants, SW developers
- 3 (ex) spinoff companies

softech-ict
Centro Interdipartimentale di Ricerca
Softech: ICT per le Imprese

Collaborations with

- **Facebook FAIR (F),** Eurecom (F)
- **Panasonic (USA)**
- **Ferrari (I)**, Maserati (I)
- MIUR, EU and Italian public bodies
- **Italian SuperComputing Resource Allocation** – CINECA
- Many smes,
- Computer Vision Foundation, CVPL-IAPR, AIXIA

Panasonic

CINECA SCAI
SuperComputing Applications and Innovation

Facebook AI Research (FAIR)

RedVision Lab
Aimage Lab UNIMORE and Ferrari spa

www.aimagelab.unimore.it

BEFORE START CHECKLIST

AImage Lab

```
ssh -p port user@YourAzureVM
pip freeze | grep torch==0.4.0
git clone https://github.com/aimagelab/aidlda_tutorial
```

i.e.:

Verify you have a working VM on Azure

Verify that PyTorch is installed

Pull from Github slides and code

PREDICTING HUMAN EYE FIXATIONS: INTRODUCTION

What are you looking at?

What are you looking at?

What is Saliency?

- The saliency of an item (an object, a person, a pixel, etc.) is the state or quality by which it stands out relative to its neighbors.
- Classical algorithms for saliency prediction focused on identifying the **fixation points** that human viewer would focus on at first glance.

Original Image

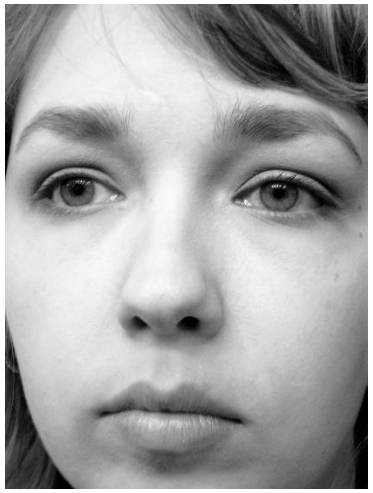
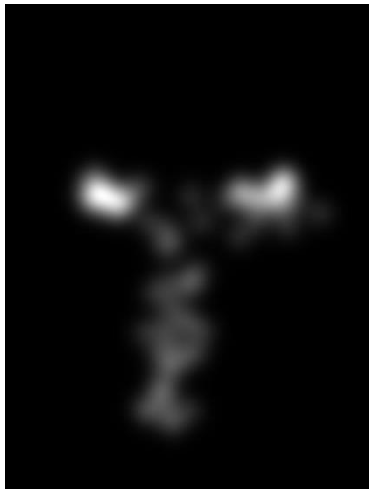


Image with fixation points

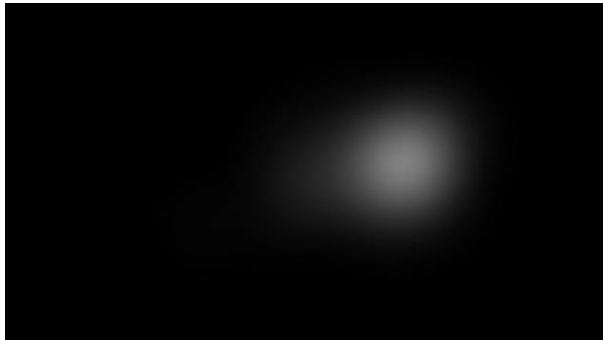
Saliency Map



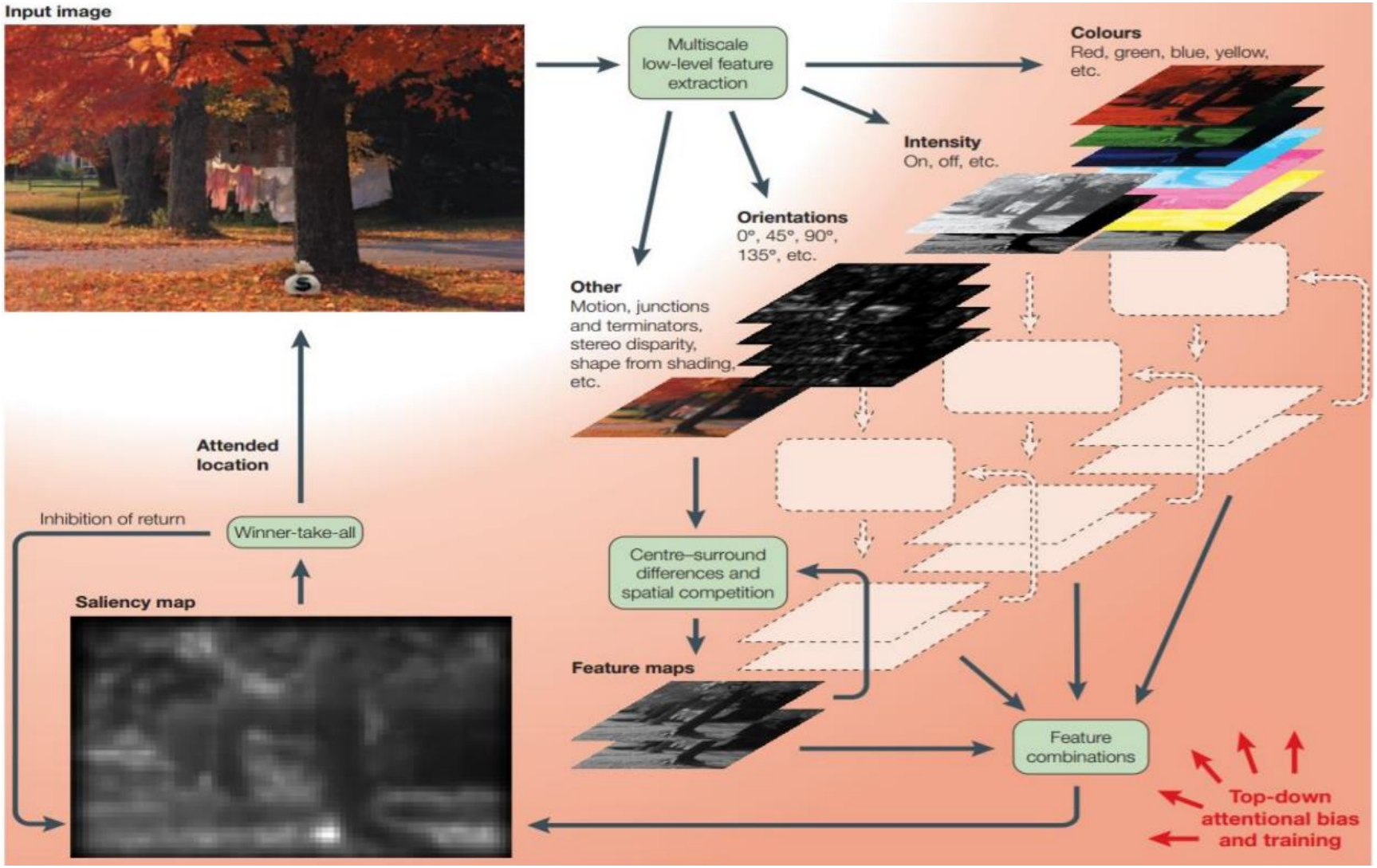
Original Video

Video with fixation points

Saliency Map



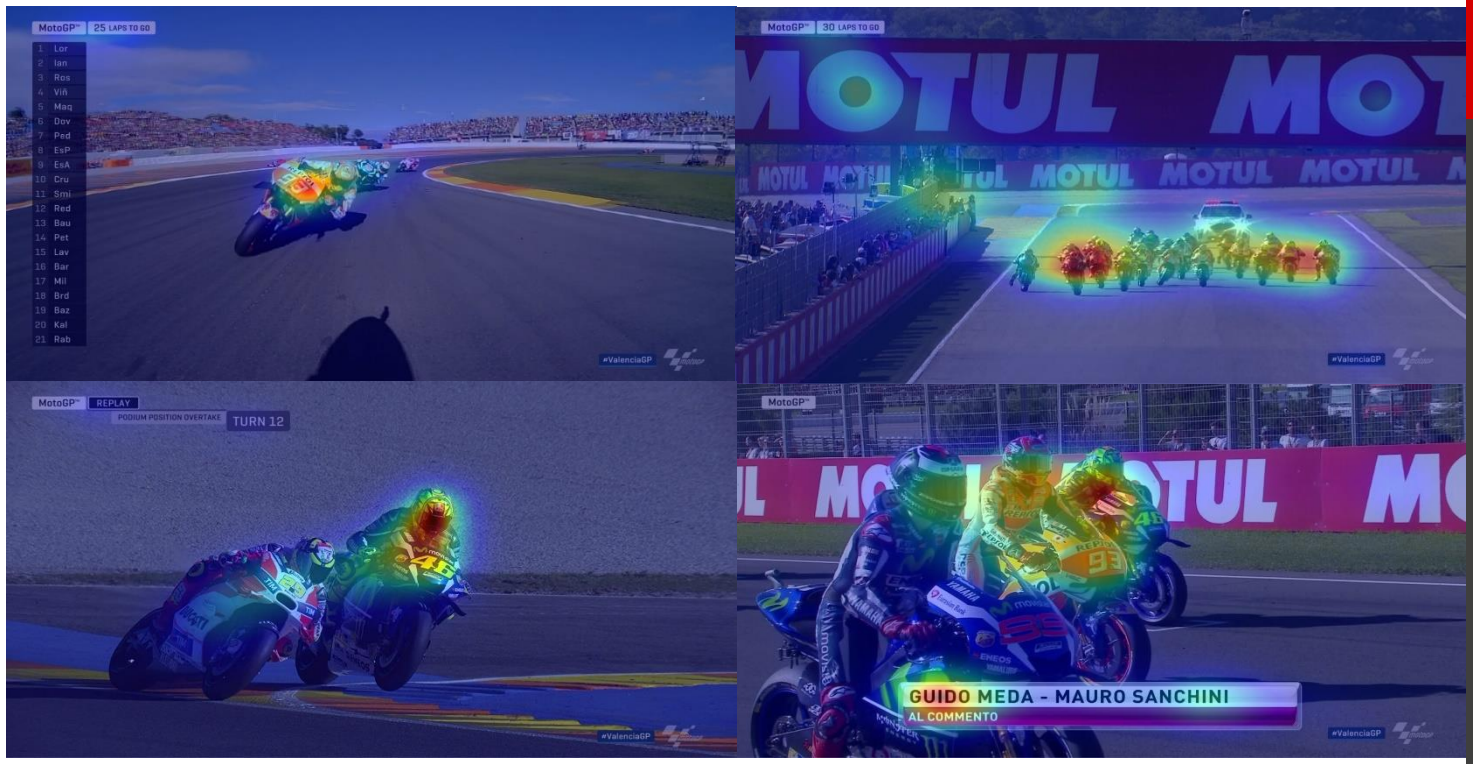
MIMIC HUMAN ATTENTIVE MECHANISMS



Itti and Koch (TPAMI 1989)

Applications: automatic annotation

Annotation of broadcast videos: saliency tells us where to see!

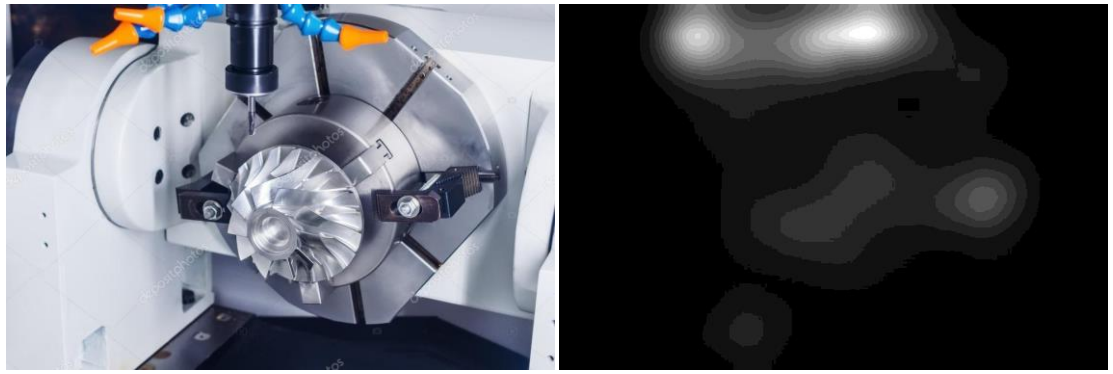


Saliency as a pre-processing step for:

- Action recognition
- Highlight detection
- Automatic image cropping

To focus on the most important details…

- Human augmented inspection with ego-vision



Applications: mimicking the driver's attention

Exploiting multi-source information

Appearance

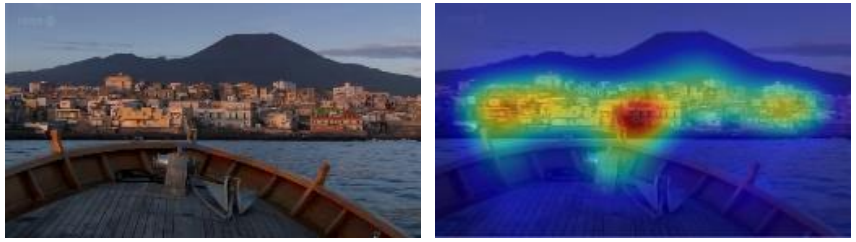
Motion

Semantics

Automatically predicted FoA

Applications: to understand visual content

Generated caption: A woman is looking at a television screen.



Generated caption: A boat is in the water near a large mountain.

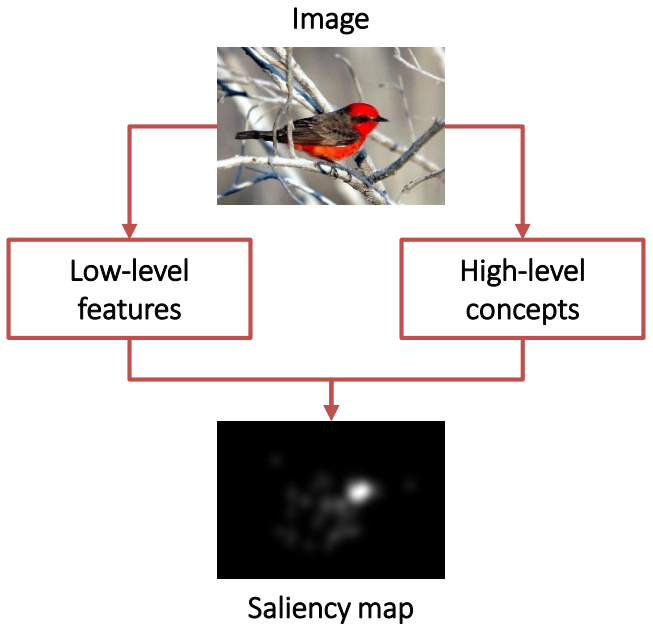
Generated caption: A city with a large boat in the water.

Generated caption: A woman in a red jacket is riding a bicycle.

Saliency Prediction

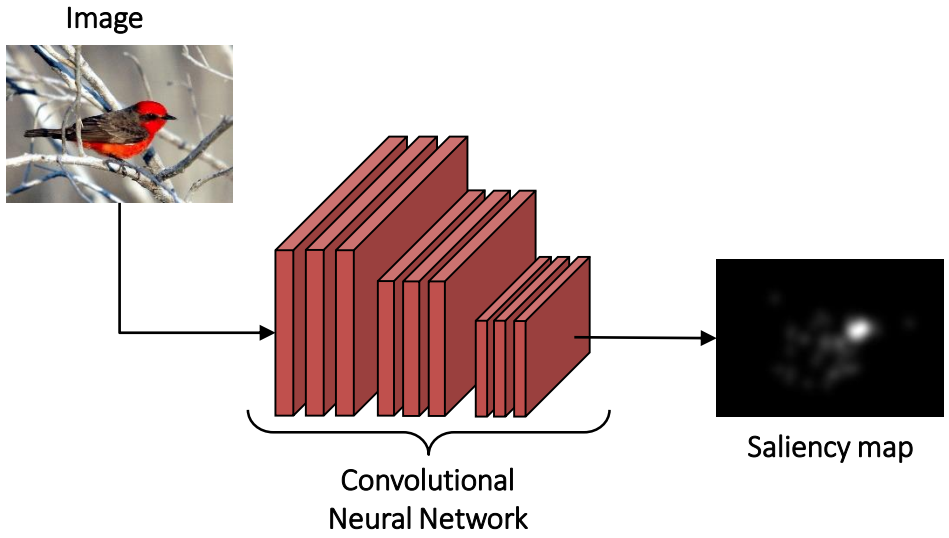
CONVENTIONAL SALIENCY

- Extraction of hand-crafted and multi-scale features:
  - Lower-level features
    - color, texture, contrast, etc.
  - Higher-level concepts
    - faces, people, text, horizon, etc.
- Difficult to combine all these factors.



DEEP SALIENCY

- Considerable progress, thanks to recent advances in deep learning.
- Fully Convolutional networks directly predict saliency maps given by a non-linear combination of high level feature maps extracted from the last convolutional layer.



SALIENCY AND INTENTIONS

AImage Lab

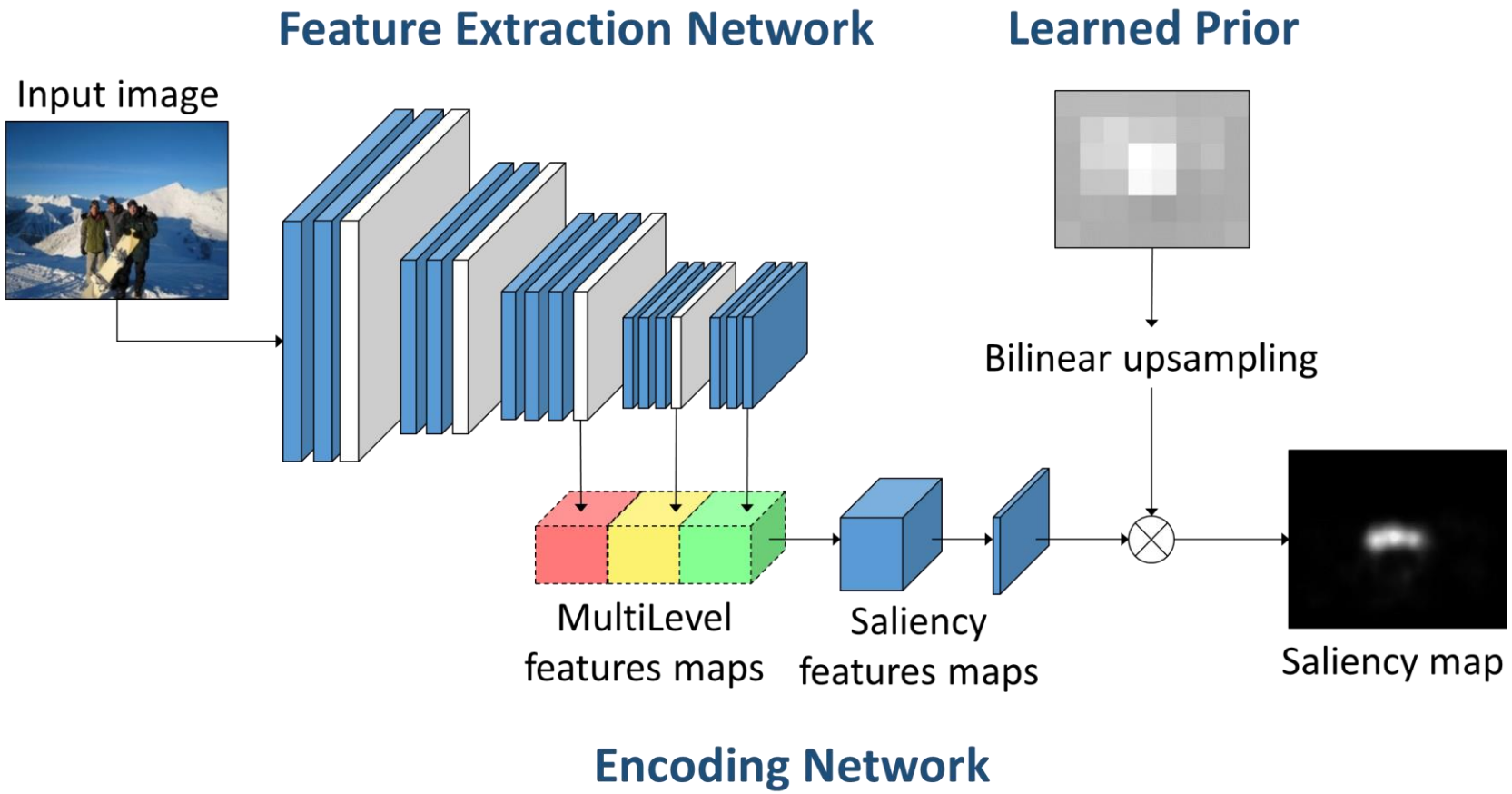
TASK-AGNOSTIC SALIENCY

what a user, without any specific goal/task, would focus on (i.e., traditional saliency)

TASK-DRIVEN SALIENCY

what a user, doing a specific task, would focus on

MULTI-LEVEL NETWORK (ML-NET) – ICPR16

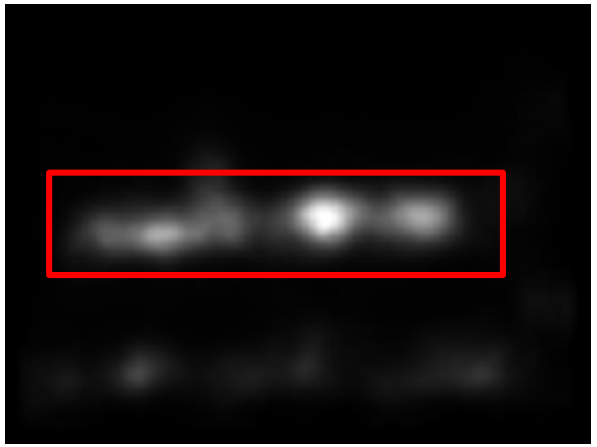


Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, Rita Cucchiara. "A Deep Multi-Level Network for Saliency Prediction." In Proceedings of the 23rd International Conference on Pattern Recognition, 2016.

Experiment

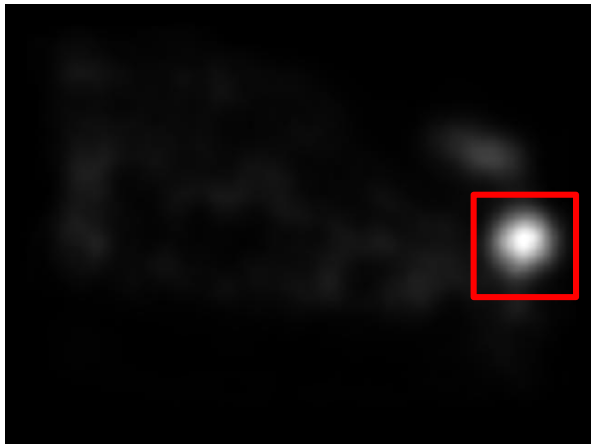
Image

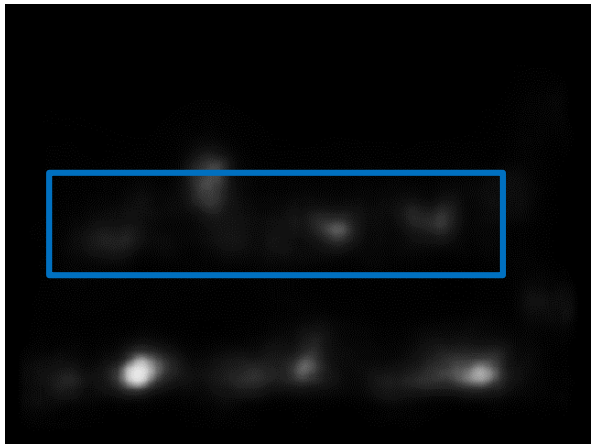
Predicted map

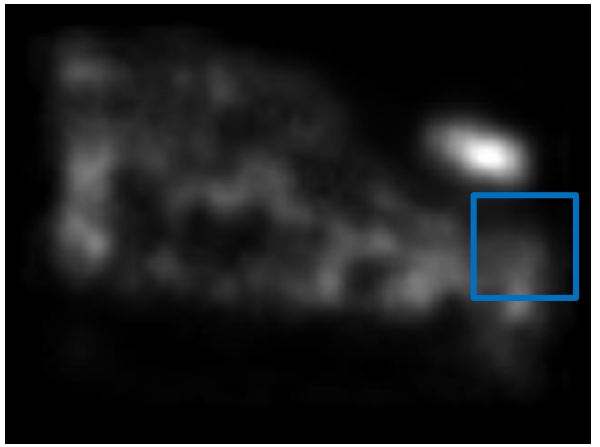


Image

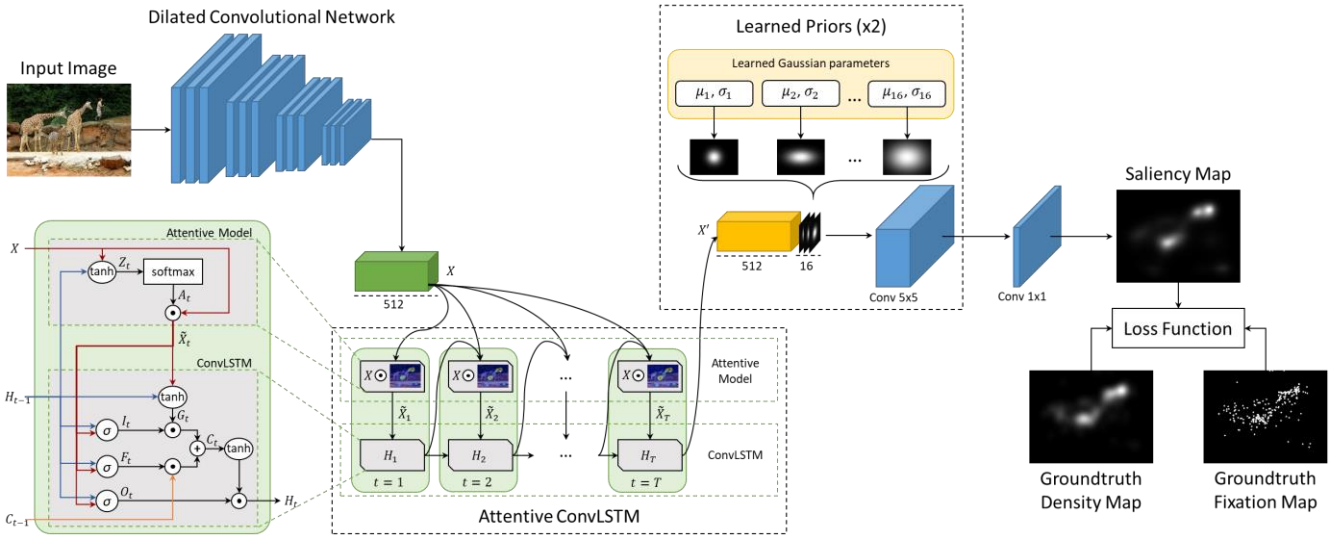
Predicted map







Saliency Attentive Model (SAM)

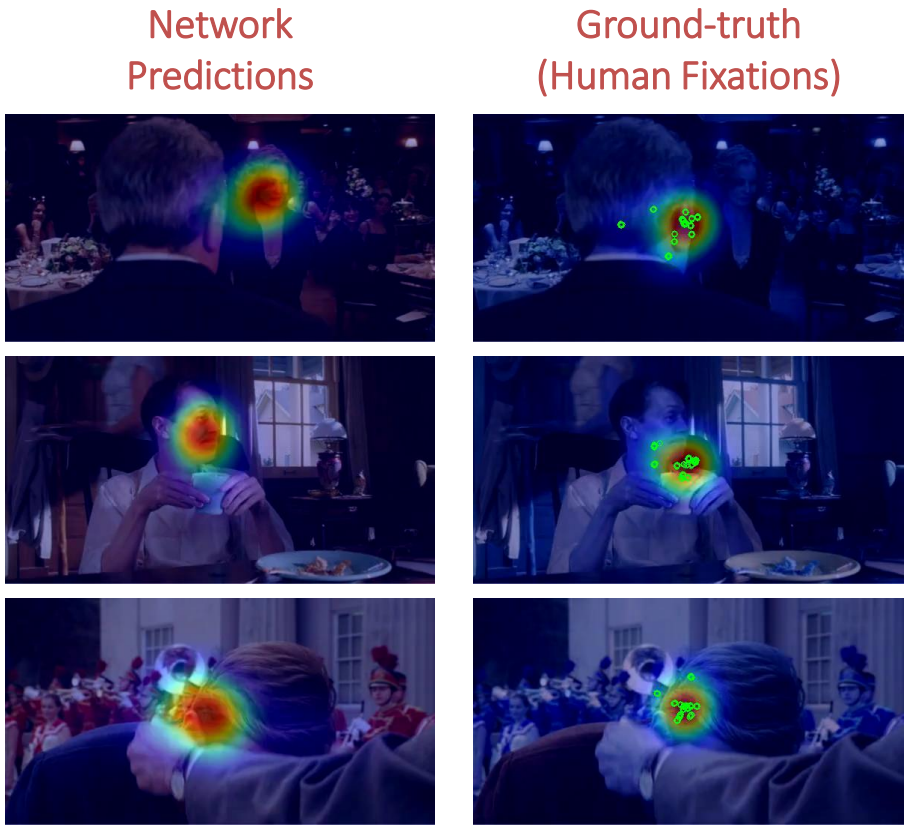


Attentive Convolutional LSTM

- Recurrent architecture that focuses on the most salient regions of the input image to iteratively refine the predicted saliency map.

Learned Priors

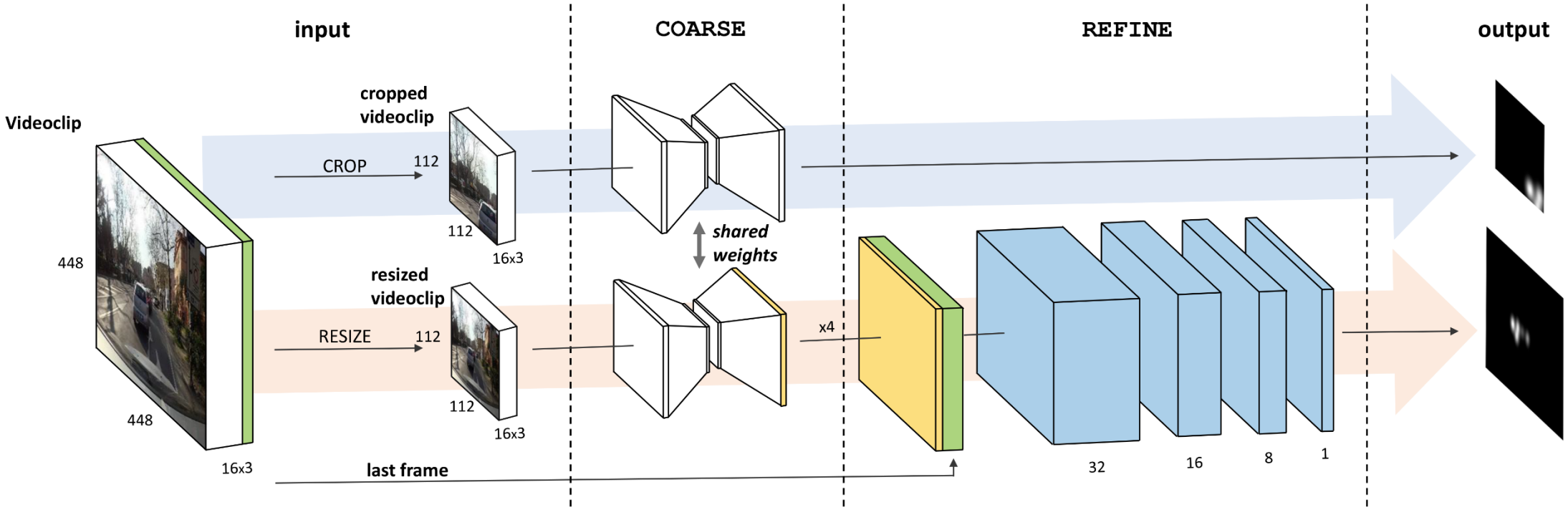
- Our network is able to learn the center bias present in eye fixations, without the need to integrate this information manually.

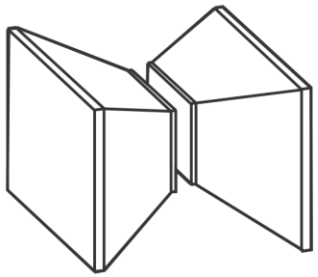


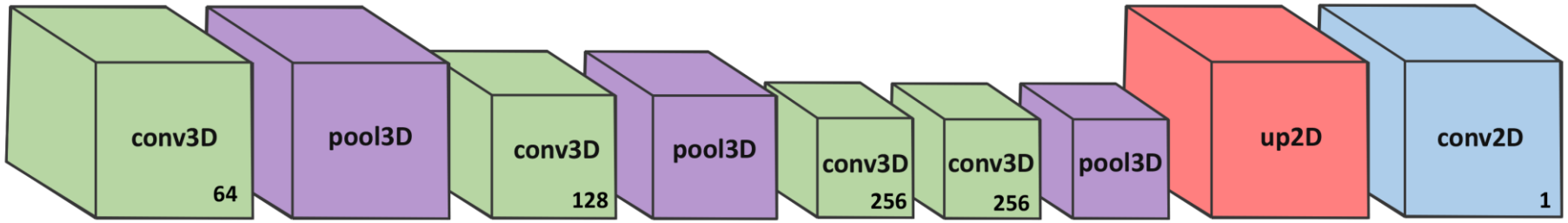
M. Cornia, L. Baraldi, G. Serra, R. Cucchiara. “Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model.” To appear in *IEEE Transactions on Image Processing*, 2018.

TASK-DRIVEN SALIENCY: DR(EYE)VE

AImage Lab

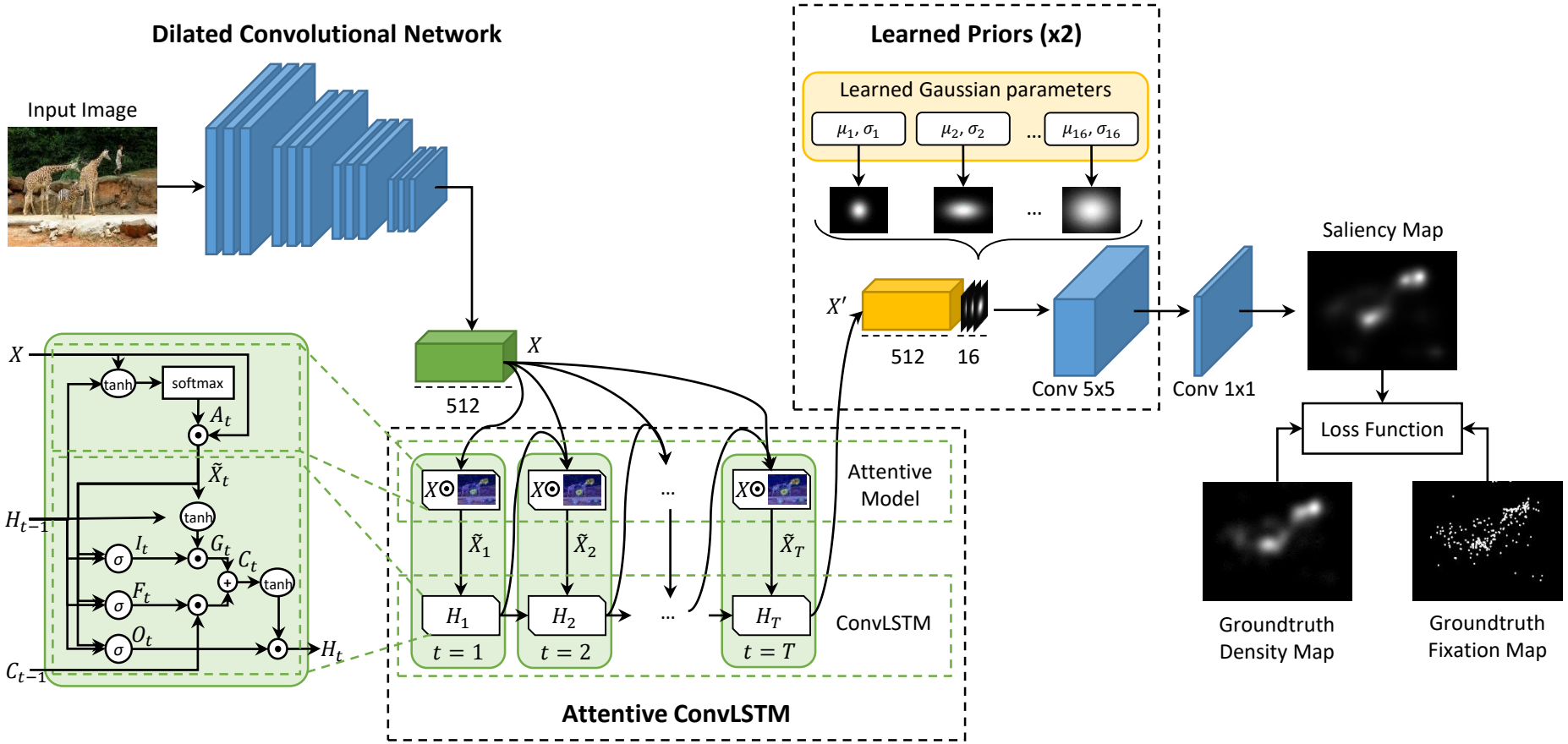






TASK AGNOSTIC SALIENCY: A SALIENCY ATTENTIVE MODEL

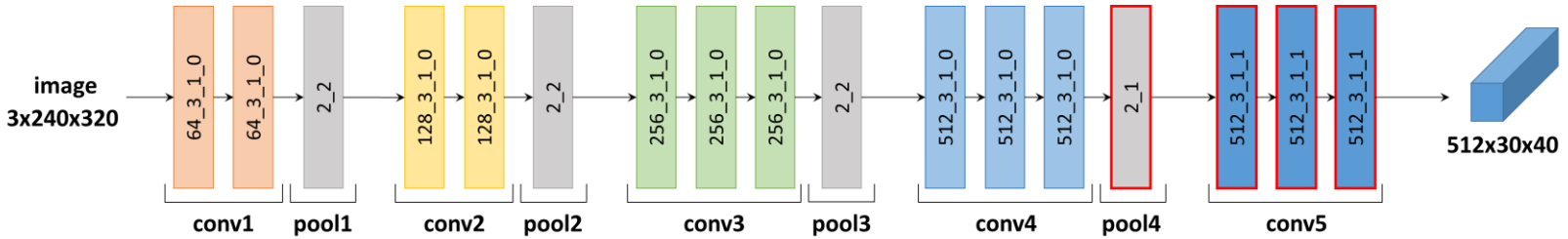
Saliency Attentive Model (SAM)



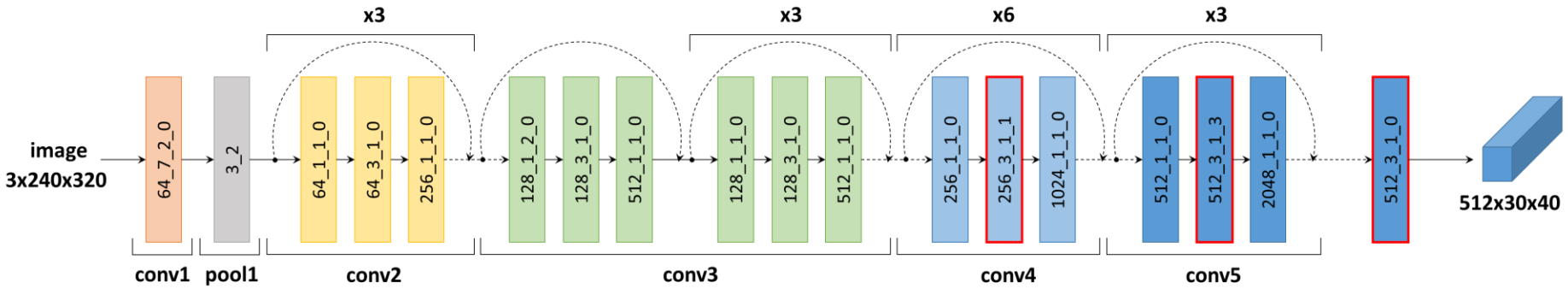
M. Cornia, L. Baraldi, G. Serra, R. Cucchiara. "Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model" arXiv preprint arXiv:1611.09571, 2017.

Dilated Convolutional Network

- We build two different versions of our model:
  - SAM-VGG based on the VGG-16 network;



  - SAM-ResNet based on the ResNet-50 network.

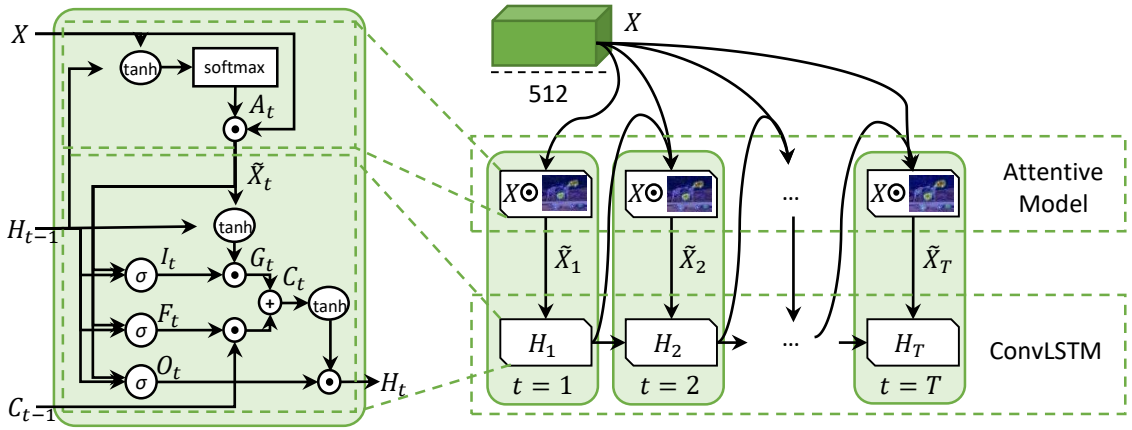


- The **rescaling of the input image** is one of the main drawbacks of using CNNs for saliency.
- We modify network structures to limit this phenomena.
- Our saliency maps are rescaled by a factor of 8 instead of 32 as in the original VGG-16 and ResNet-50 models.

Attentive Convolutional LSTM

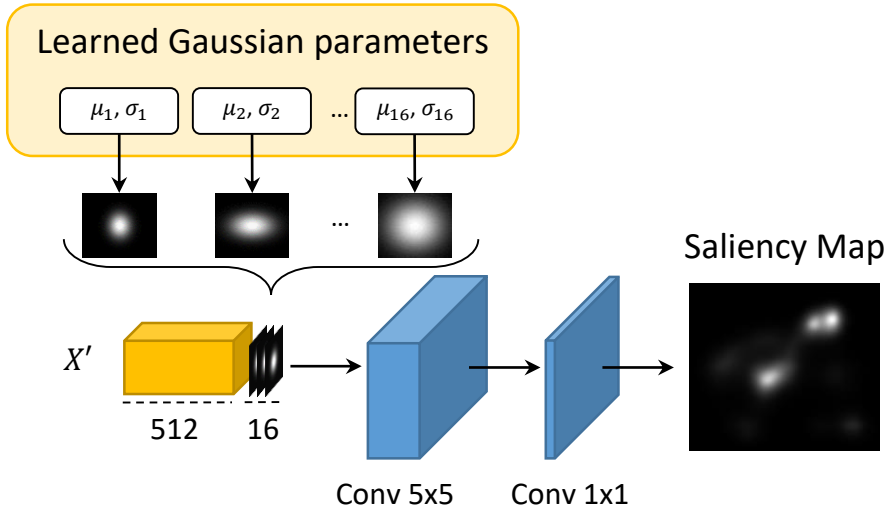
- Extension of the traditional LSTM to work on spatial features by substituting dot products with **convolutional operations**.
- Exploitation of the sequential nature of the LSTM to process features in an iterative way, **without the concept of time**.

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



Learned Priors

- Human gazes are biased toward the center of the image.
- Instead of using pre-defined priors, we integrate a module which can learn **multiple prior maps** from data.



- We model the center bias by means of a set of Gaussian functions with diagonal covariance matrix.
- Means and variances are learned for each prior map, according to:

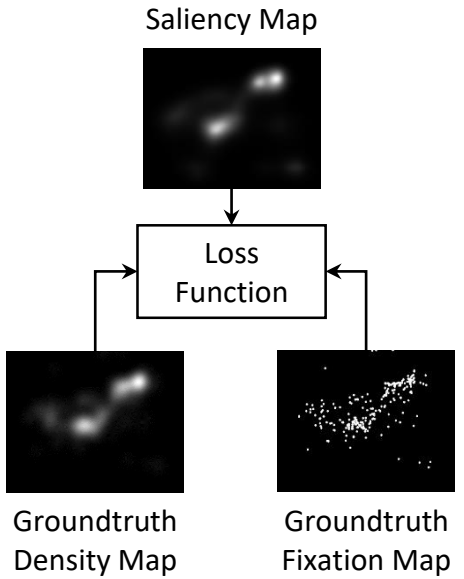
$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y}\exp\left(-\left(\frac{(x-\mu_x)^2}{2\sigma_x^2}+\frac{(y-\mu_y)^2}{2\sigma_y^2}\right)\right)$$

- These maps are concatenated with the output tensor of the Attentive ConvLSTM and fed through a convolutional layer with 512 filters.
- The last layer of our model is a convolutional operation with one filter that extracts the final saliency map.

Loss Function

- Saliency predictions are evaluated through different metrics, in order to capture several quality factors.
- We introduce a new loss function given by a linear combination of three different saliency evaluation metrics:

$$L(\tilde{\mathbf{y}}, \mathbf{y}^{den}, \mathbf{y}^{fix}) = \alpha L_1(\tilde{\mathbf{y}}, \mathbf{y}^{fix}) + \beta L_2(\tilde{\mathbf{y}}, \mathbf{y}^{den}) + \gamma L_3(\tilde{\mathbf{y}}, \mathbf{y}^{den})$$



Where $\tilde{\mathbf{y}}$, $\mathbf{y}^{den}$ and $\mathbf{y}^{fix}$ are respectively the predicted saliency map, the groundtruth density distribution and the groundtruth binary fixation map.

$$L_1\left(\tilde{\mathbf{y}}, \mathbf{y}^{fix}\right) = \frac{1}{N} \sum_i \frac{\tilde{\mathbf{y}}_i - \mu(\tilde{\mathbf{y}})}{\sigma(\tilde{\mathbf{y}})} \cdot \mathbf{y}_i^{fix}$$

$$L_2\left(\tilde{\mathbf{y}}, \mathbf{y}^{den}\right) = \frac{\sigma\left(\tilde{\mathbf{y}}, \mathbf{y}^{den}\right)}{\sigma\left(\tilde{\mathbf{y}}\right) \cdot \sigma\left(\mathbf{y}^{den}\right)}$$

$$L_3\left(\tilde{\mathbf{y}}, \mathbf{y}^{den}\right) = \sum_i \mathbf{y}_i^{den} \log\left(\frac{\mathbf{y}_i^{den}}{\tilde{\mathbf{y}}_i + \epsilon} + \epsilon\right)$$

Normalized Scanpath Saliency
that quantifies the saliency map values at the eye fixation locations and normalizes it with the saliency map variance.

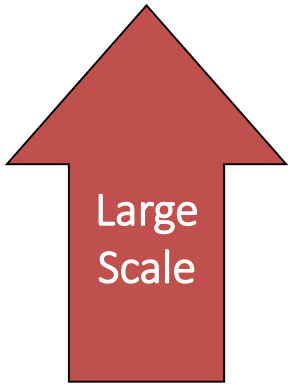
Linear Correlation Coefficient
that treats the saliency and groundtruth density maps as random variables measuring the linear relationship between them.

KL-Divergence
that evaluates the loss of information when the predicted saliency map is used to approximate the groundtruth density map.

Experimental Results

Datasets and Evaluation Metrics

	TRAIN	VALIDATION	TEST	
SALICON [1]	10,000	5,000	5,000	
CAT2000 [2]	1,800	200	2,000	
MIT300 [3]	-	-	300	

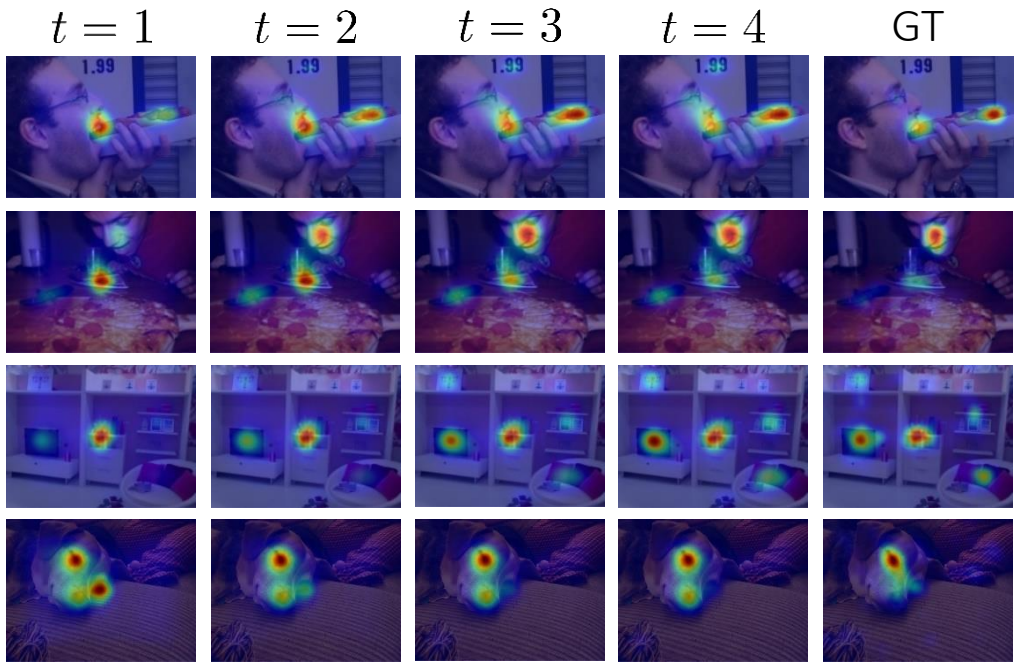


- Saliency prediction results are usually evaluated with a large variety of metrics:
  - Linear Correlation Coefficient (CC)
  - Area under the ROC Curve (AUC)
  - AUC shuffled (sAUC)
  - Normalized Scanpath Saliency (NSS)

[1] Jiang et al. "SALICON: Saliency in Context." *CVPR*, 2015.
[2] Borji et al. "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research." *CVPR Workshops*, 2015.
[3] Judd et al. "A benchmark of computational models of saliency to predict human fixations." *MIT Technical Report*, 2012.

Contribution of the Attentive Model

- Overall performance when using the output of the Attentive ConvLSTM at different timestep as input for the rest of the model.
- The refinement performed by the attentive model results in better performance.



	T	CC	sAUC	AUC	NSS
SAM-VGG	1	0.821	0.777	**0.884**	3.168
SAM-VGG	2	0.827	0.777	0.883	3.224
SAM-VGG	3	0.828	0.781	0.883	**3.226**
SAM-VGG	4	**0.830**	**0.782**	0.883	3.219
SAM-ResNet	1	0.785	0.737	0.879	3.050
SAM-ResNet	2	0.829	0.764	**0.886**	3.214
SAM-ResNet	3	0.842	0.779	**0.886**	3.256
SAM-ResNet	4	**0.844**	**0.787**	**0.886**	**3.260**

Comparison with the state of the art

SALICON Dataset (2015 release)

	CC	sAUC	AUC	NSS
SAM-ResNet	**0.842**	**0.779**	0.883	**3.204**
SAM-VGG	0.825	0.774	0.881	3.143
ML-Net [1]	0.743	0.768	0.866	2.789
SalGAN [2]	0.781	0.772	0.781	2.459
SalNet [3]	0.622	0.724	0.858	1.859
DeepGazeII [4]	0.509	0.761	**0.885**	1.336

SALICON Dataset (2017 release)

	CC	Sim	AUC	NSS
SAM-ResNet	**0.899**	**0.793**	**0.865**	**1.990**
SAM-VGG	0.891	0.786	0.864	1.971
EAD [5]	0.871	0.760	0.852	1.896
SalGAN [2]	0.844	0.728	0.857	1.816
SalNet [3]	0.763	0.639	0.840	1.555

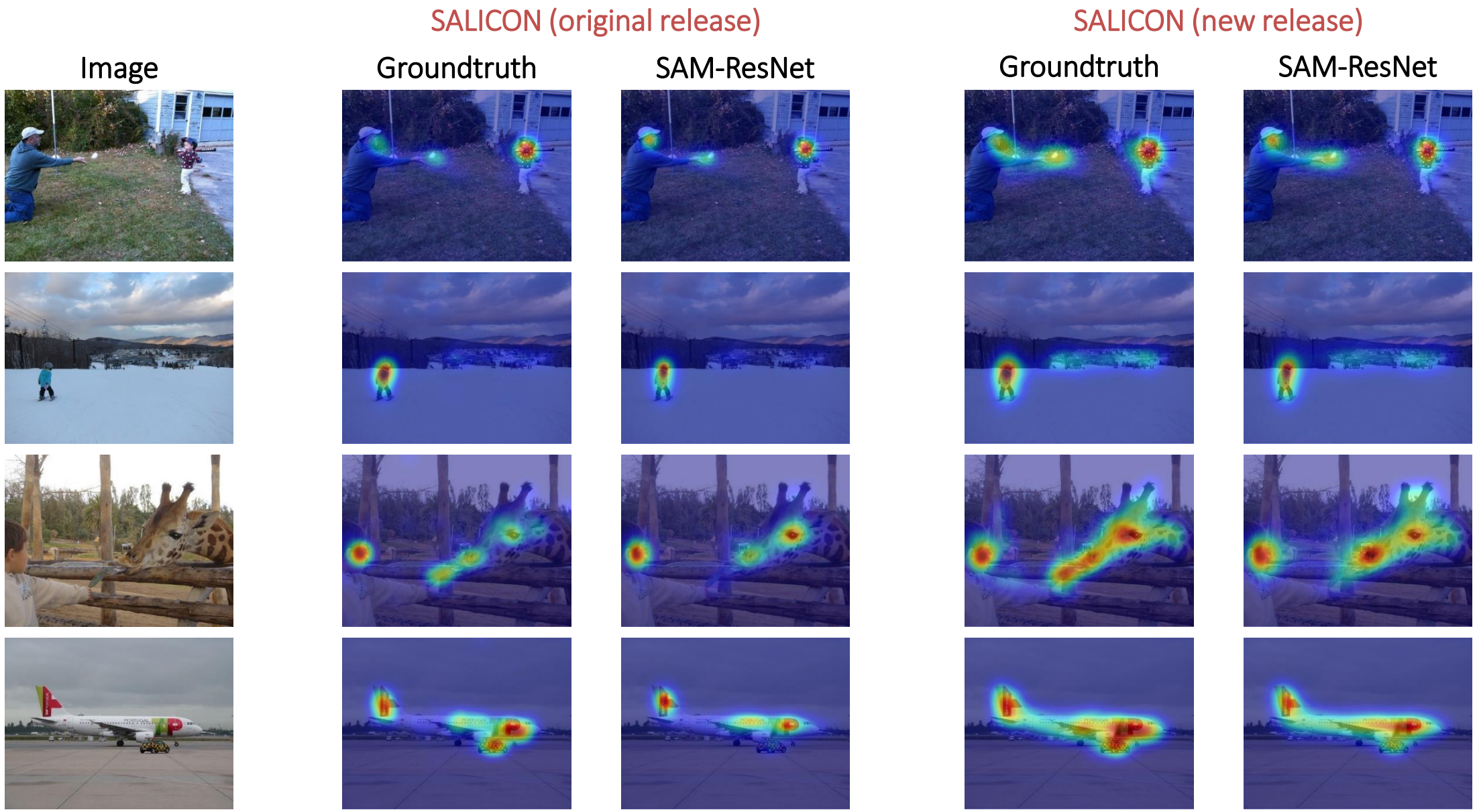
[1] Cornia et al. “A Deep Multi-Level Network for Saliency Prediction." ICPR, 2016.
[2] Pan et al. “Shallow and Deep Convolutional Networks for Saliency Prediction.” CVPR, 2016.
[3] Pan et al. “SalGAN: Visual Saliency Prediction with Generative Adversarial Networks.” CVPR Workshops, 2017.
[4] Kümmerer 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 preprint arXiv:1803.05753, 2018.

Comparison with the state of the art

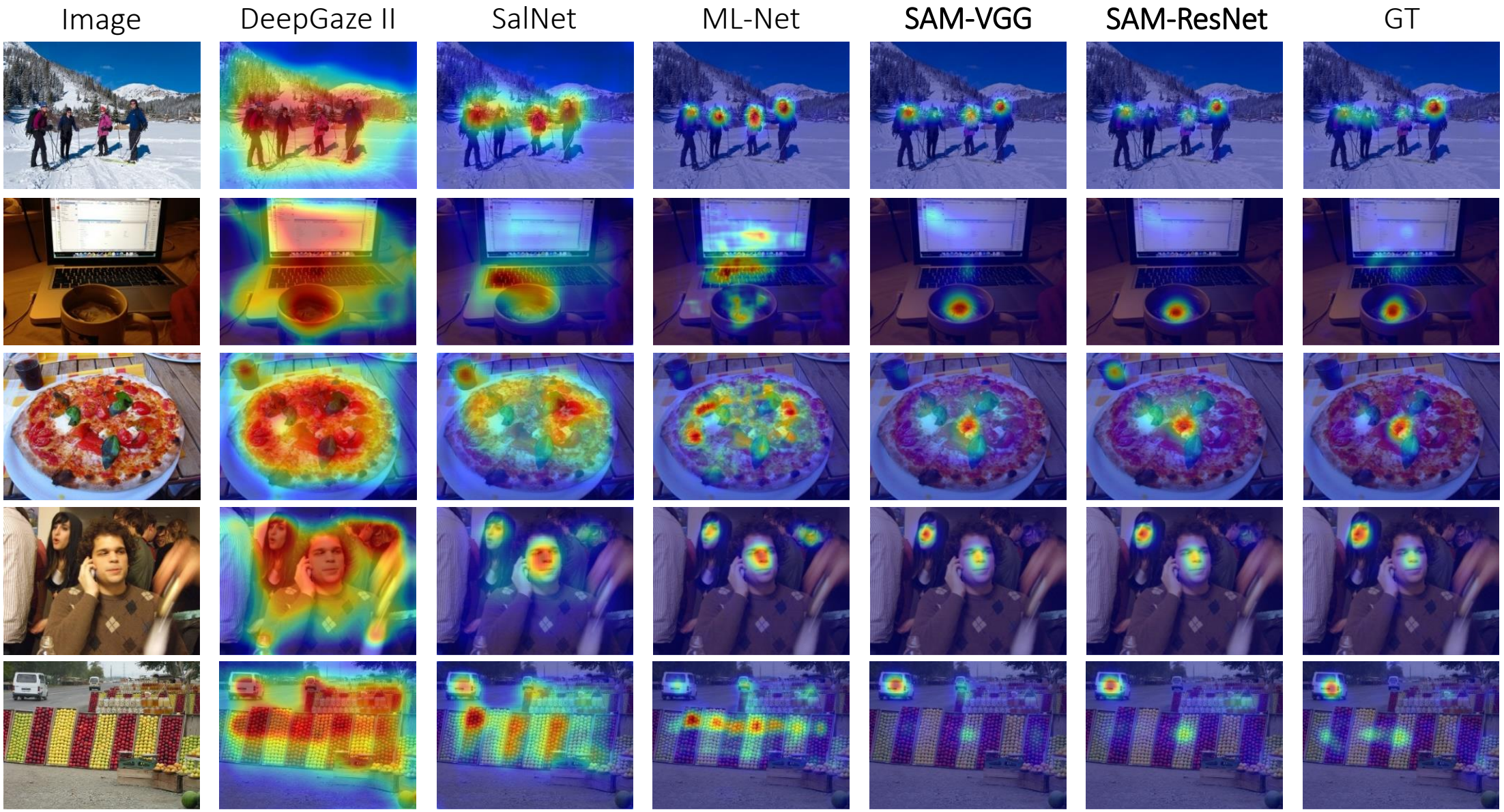
	MIT1003			DUT-OMRON			TORONTO			PASCAL-S		
	CC	AUC	NSS	CC	AUC	NSS	CC	AUC	NSS	CC	AUC	NSS
Itti [1]	0.33	0.77	1.10	0.46	0.83	1.54	0.48	0.80	1.30	0.42	0.82	1.30
GBVS [2]	0.42	0.83	1.38	0.53	0.87	1.71	0.57	0.83	1.52	0.45	0.84	1.36
eDN [3]	0.41	0.85	1.29	-	-	1.33	0.50	0.85	1.25	-	-	1.42
Mr-CNN [4]	0.38	0.80	1.36	-	-	-	0.49	0.80	1.41	-	-	-
DVA [5]	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

[1] Itti et al. “A model of saliency-based visual attention for rapid scene analysis." IEEE TPAMI, 1998.
[2] Harel et al. “Graph-based visual saliency.” NIPS, 2006.
[3] Vig et al. “Large-scale optimization of hierarchical features for saliency prediction in natural images.” CVPR, 2014.
[4] Liu et al. “Predicting eye fixations using convolutional neural networks.” CVPR, 2015.
[5] Wang et al. “Deep Visual Attention Prediction.” IEEE Transactions on Image Processing, 2018.

Qualitative Results

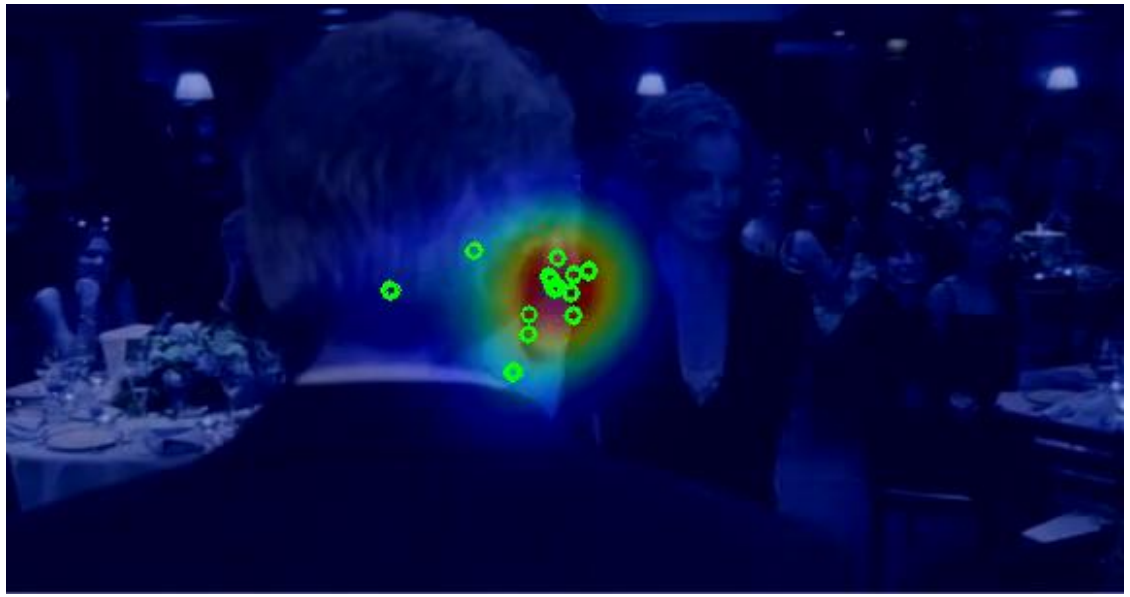


Qualitative Results

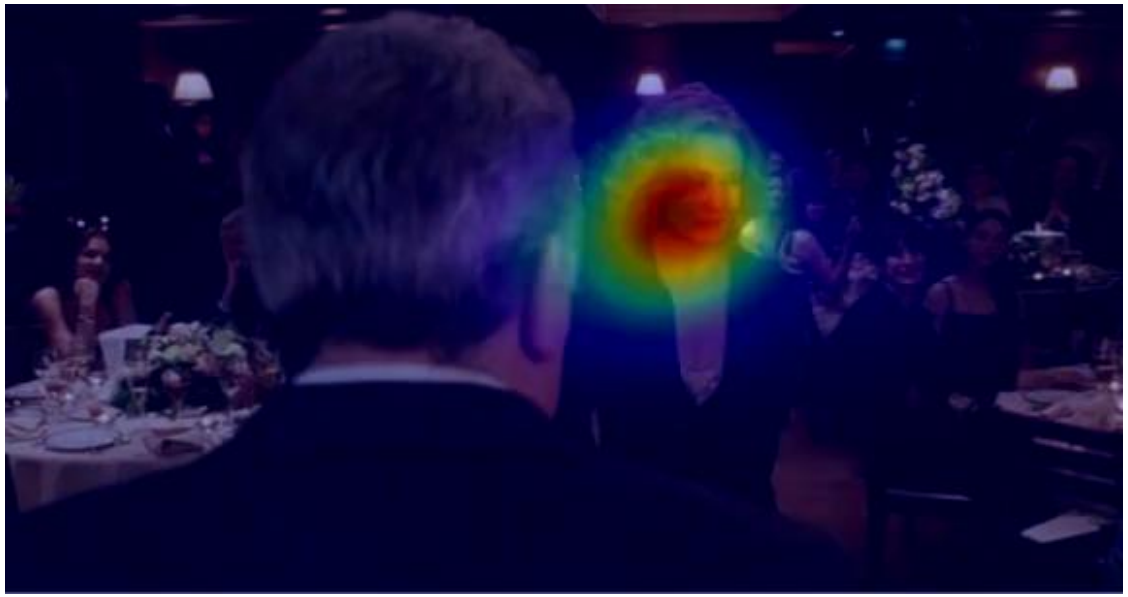


Results on Hollywood2 dataset

	CC	Similarity	AUC	NSS
SAM	**0.694**	**0.574**	**0.922**	**3.202**
RMDN [1]	0.613	0.535	0.904	2.646



Groundtruth



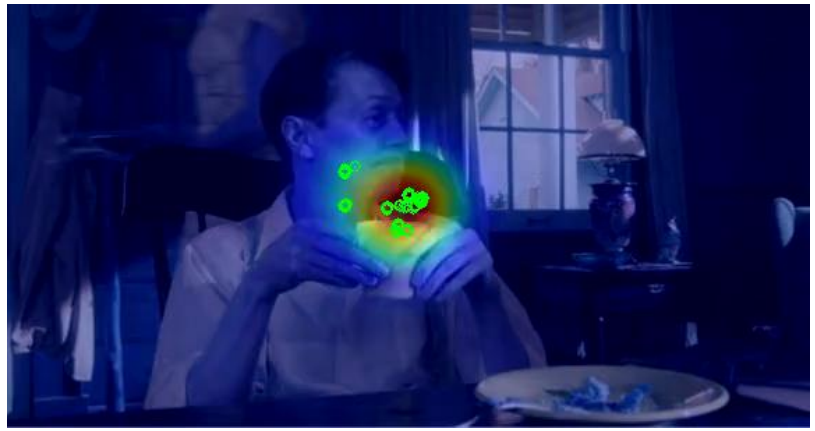
SAM

[1] Bazzani et al. “Recurrent Mixture Density Network for Spatiotemporal Visual Attention ." ICLR, 2017.

Results on Hollywood2 dataset

Groundtruth



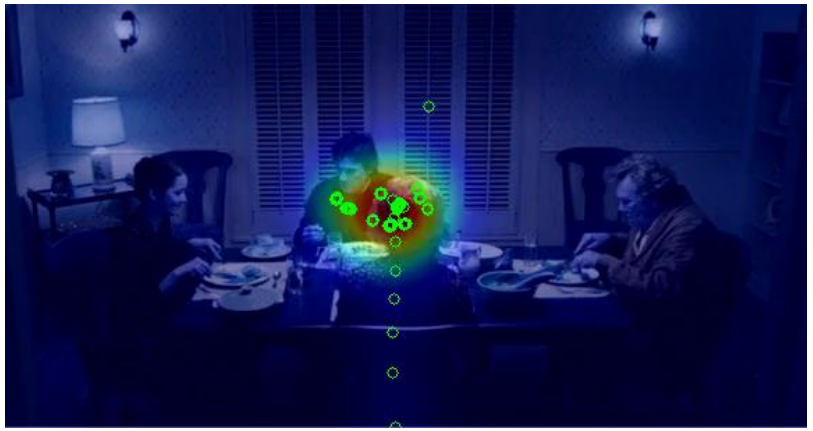


SAM

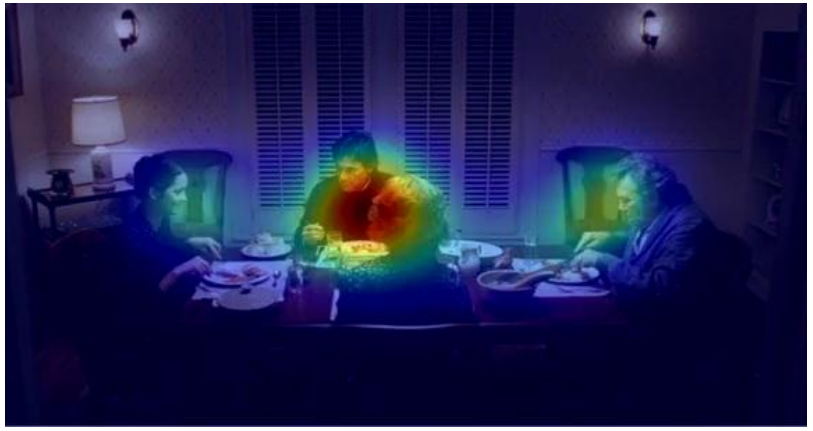


AImage Lab

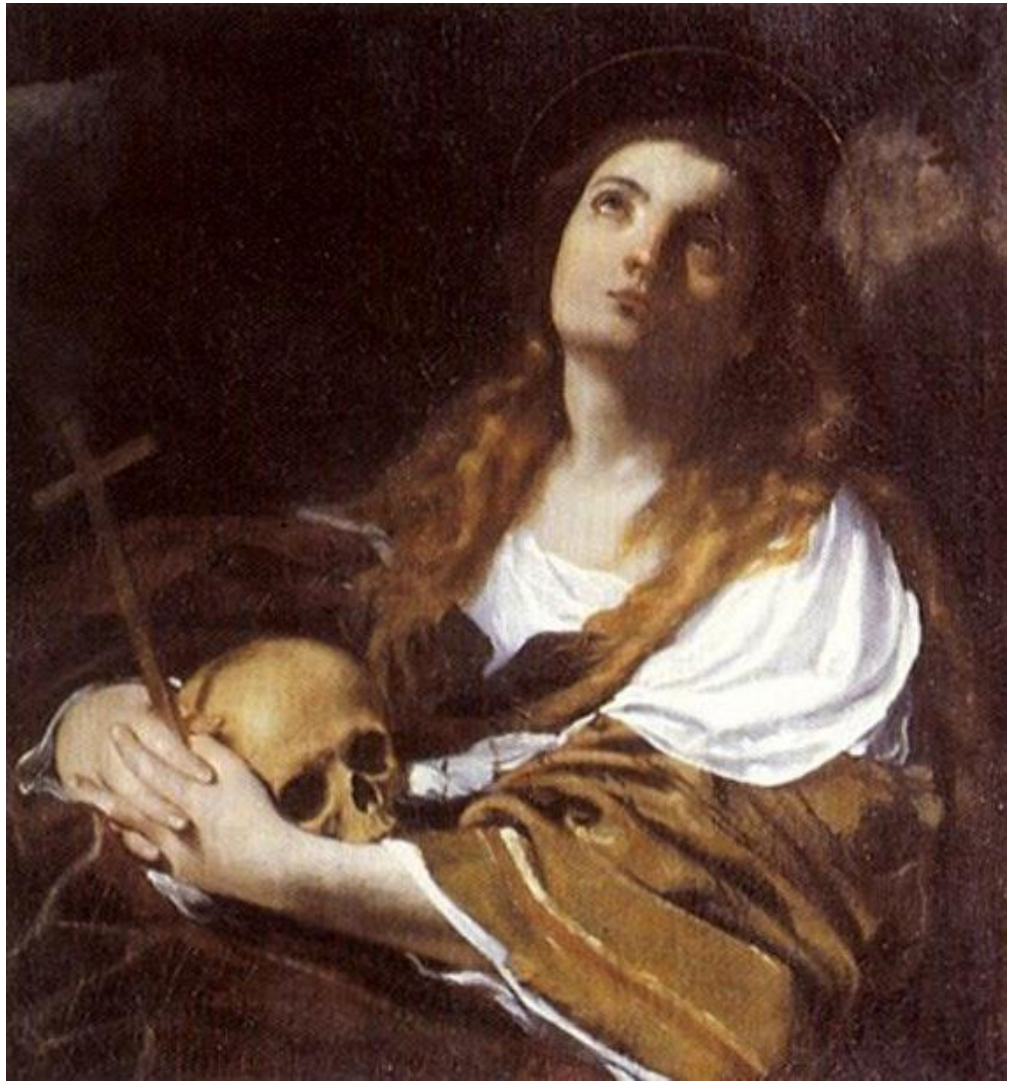
Groundtruth



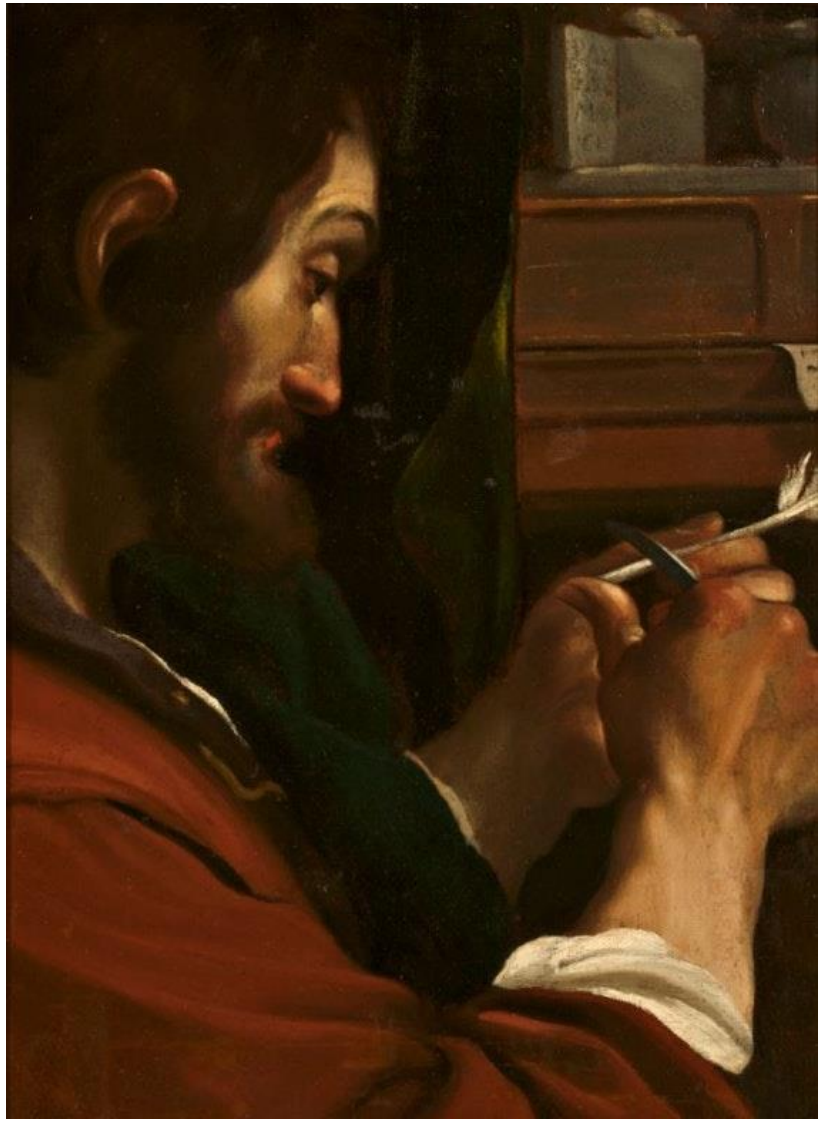
SAM

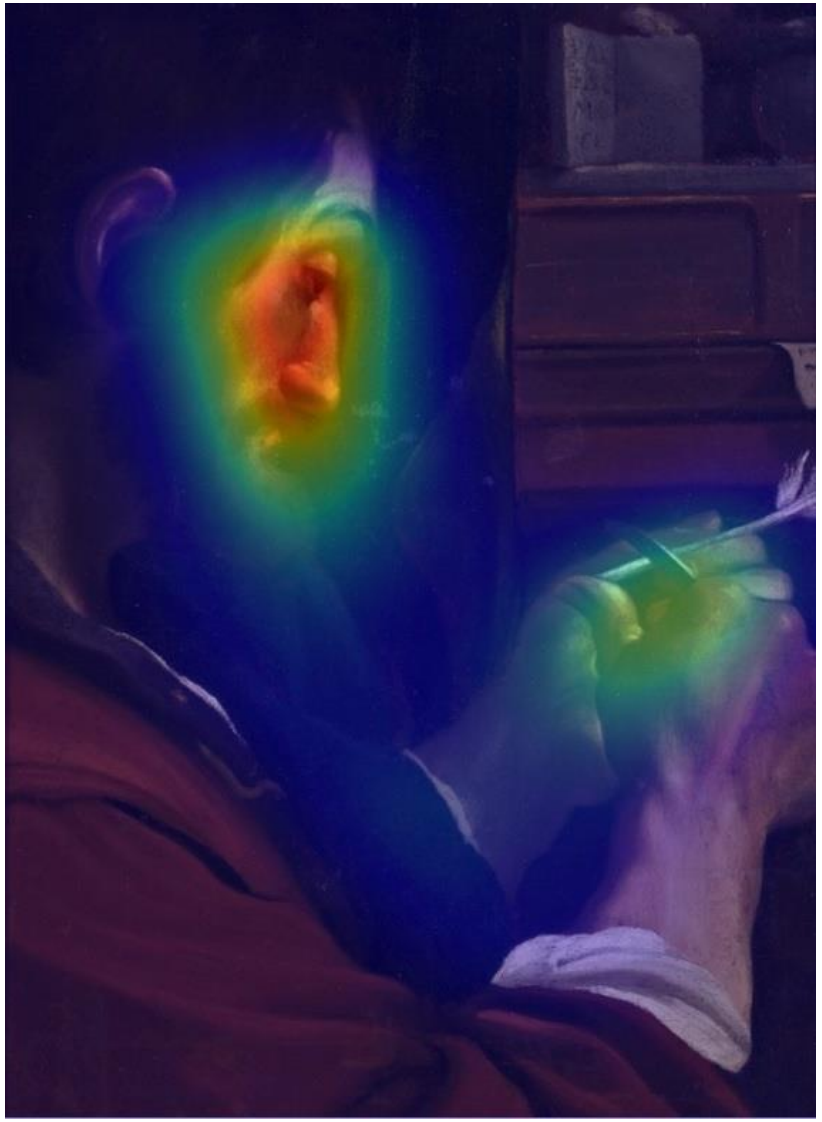


Saliency on art



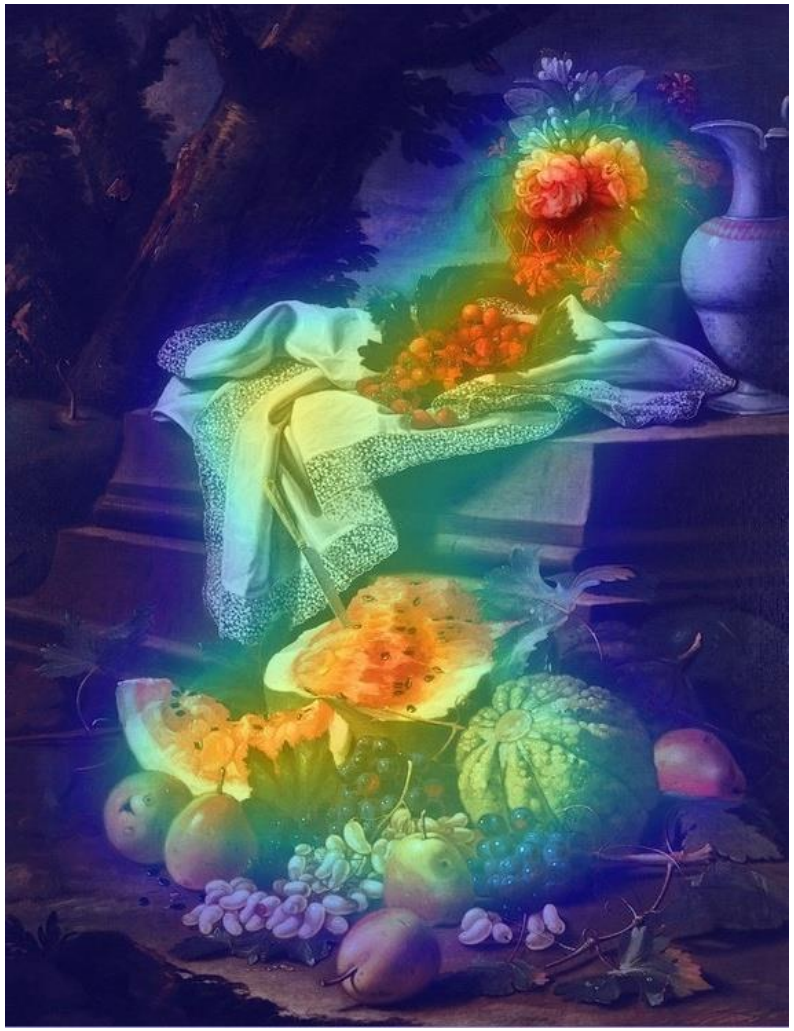
Saliency on art





Saliency on art

Saliency on art



From saliency to captioning

LSTM networks as language models

At training time: condition on the image and train to predict the next word given the previous (GT) words

- only at the first timestep
- at every timestep

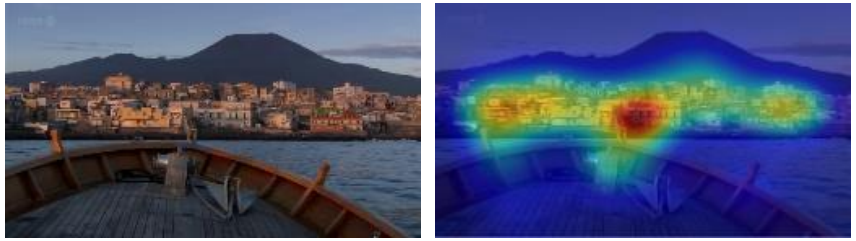
$$\max_{\mathbf{w}} \sum_{t=1}^{T} \log \Pr(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, ..., \mathbf{y}_0)$$

Using a vocabulary of more than 10.000 words

Automatic annotation

Automatically generated captions will be useful for human search, for automatic search by query, and for future query-answering services.

Generated caption: A woman is looking at a television screen.



Generated caption: A boat is in the water near a large mountain.

Generated caption: A city with a large boat in the water.

Generated caption: A woman in a red jacket is riding a bicycle.

L. Baraldi, C. Grana, R. Cucchiara, "Hierarchical Boundary-Aware Neural Encoder for Video Captioning" CVPR, 2017

M. Cornia, L. Baraldi, G. Serra, R. Cucchiara, "Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention" ACM TOMM, 2017