Automatic Image Cropping and Selection using Saliency: an Application to Historical Manuscripts

M. Cornia, S. Pini, L. Baraldi, R. Cucchiara

name.surname@unimore.it

University of Modena e Reggio Emilia

Image cropping

- Extraction of rectangular sub-regions from a given image
  - To preserve (most of) the visual content
  - And enhance the visual quality of the cropped image
  - It requires to solve the problem of “visual interestingness”
- Several applications:
  - Helping professional editors in advertisement and publishing
  - Increase presentation quality in search engines and social networks
  - Representations of image collections with a single image
- Naturally useful for *multimedia digital libraries*

Our contribution

- A saliency-based solution for image cropping, applicable to the digital humanities domain

Outline

- Introduction to saliency prediction
- Saliency Attentive Model (SAM)
- Saliency for automatic Image Cropping
- Experimental results
- Application to historical documents

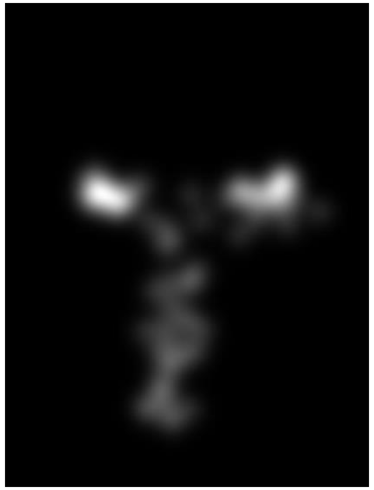
What is Saliency?

- 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 neighbours.
- 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

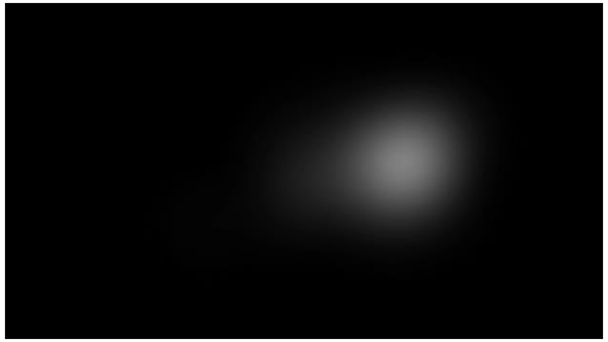
Saliency Map



Original Video

Video with fixation points

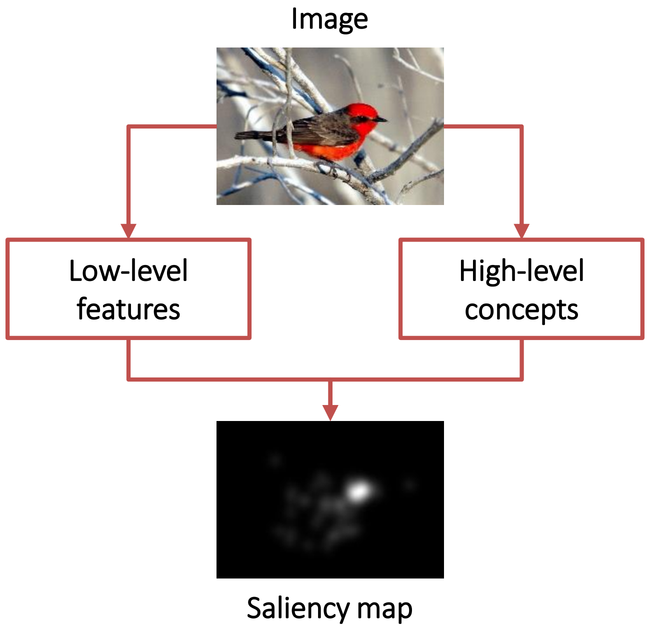
Saliency Map



Saliency Prediction

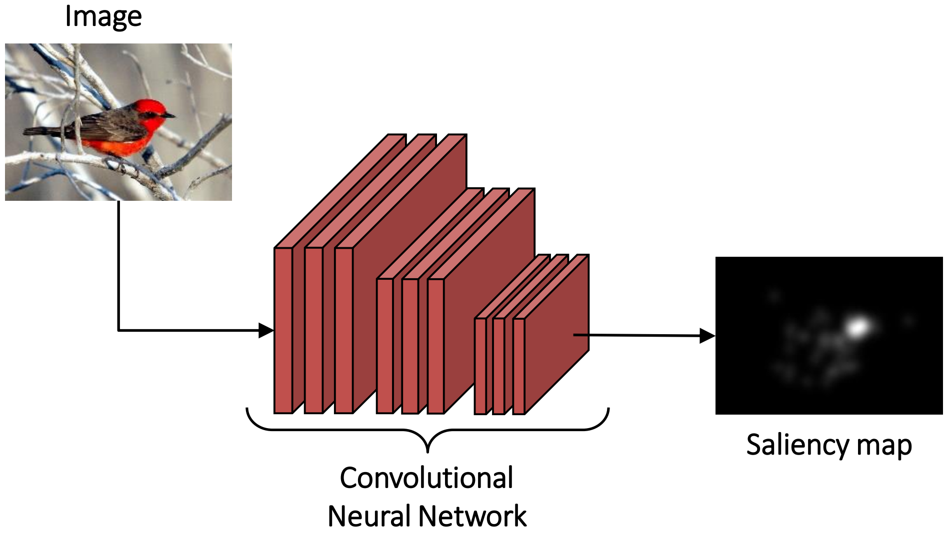
CONVENTIONAL SALIENCY

- 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.

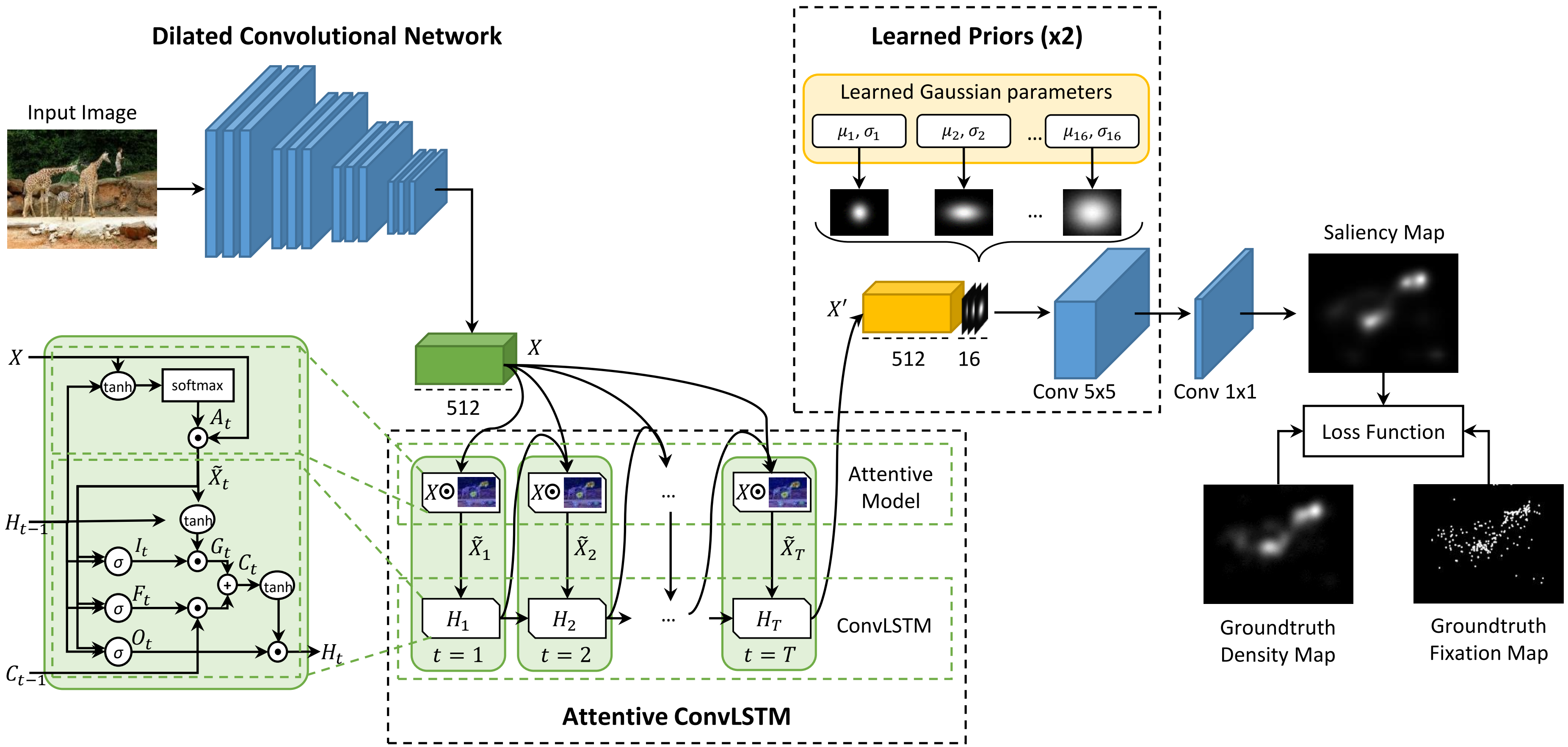


DEEP SALIENCY

- 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 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.

Results on SALICON dataset

Original Release

	CC	sAUC	AUC	NSS
SAM-ResNet	**0.842**	**0.779**	0.883	**3.204**
ML-Net [1]	0.743	0.768	0.866	2.789
SU [2]	0.780	0.760	0.880	2.610
SalNet [3]	0.622	0.724	0.858	1.859
DeepGazeII [4]	0.509	0.761	**0.885**	1.336

New Release

	CC	sAUC	AUC	NSS
SAM-ResNet	**0.899**	**0.741**	**0.865**	**1.990**

1st at LSUN Challenge

CVPR 2017

[1] Cornia et al. "A Deep Multi-Level Network for Saliency Prediction." ICPR, 2016.

[2] Kruthiventi et al. "Saliency Unified: A deep architecture for eye fixation prediction and salient object segmentation." CVPR, 2016.

[3] Pan et al. "Shallow and Deep Convolutional Networks for Saliency Prediction." CVPR, 2016.

[4] Kümmerer et al. "DeepGaze II: Reading fixations from deep features trained on object recognition."arXiv:1610.01563, 2016.

Results on MIT Saliency Benchmark

Results on MIT300 Dataset

	CC	sAUC	AUC	NSS
SAM-ResNet	**0.78**	0.70	0.87	**2.34**
SAM-VGG	0.77	0.71	0.87	2.30
DeepFix [6]	**0.78**	0.71	0.87	2.26
SALICON [7]	0.74	**0.74**	0.87	2.12
ML-Net [1]	0.67	0.70	0.85	2.05
SalGAN [3]	0.73	0.72	0.86	2.04
iSEEL [8]	0.65	0.68	0.84	1.78
SalNet [4]	0.58	0.69	0.83	1.51
DeepGazeII [5]	0.52	0.72	**0.88**	1.29

Results on CAT2000 Dataset

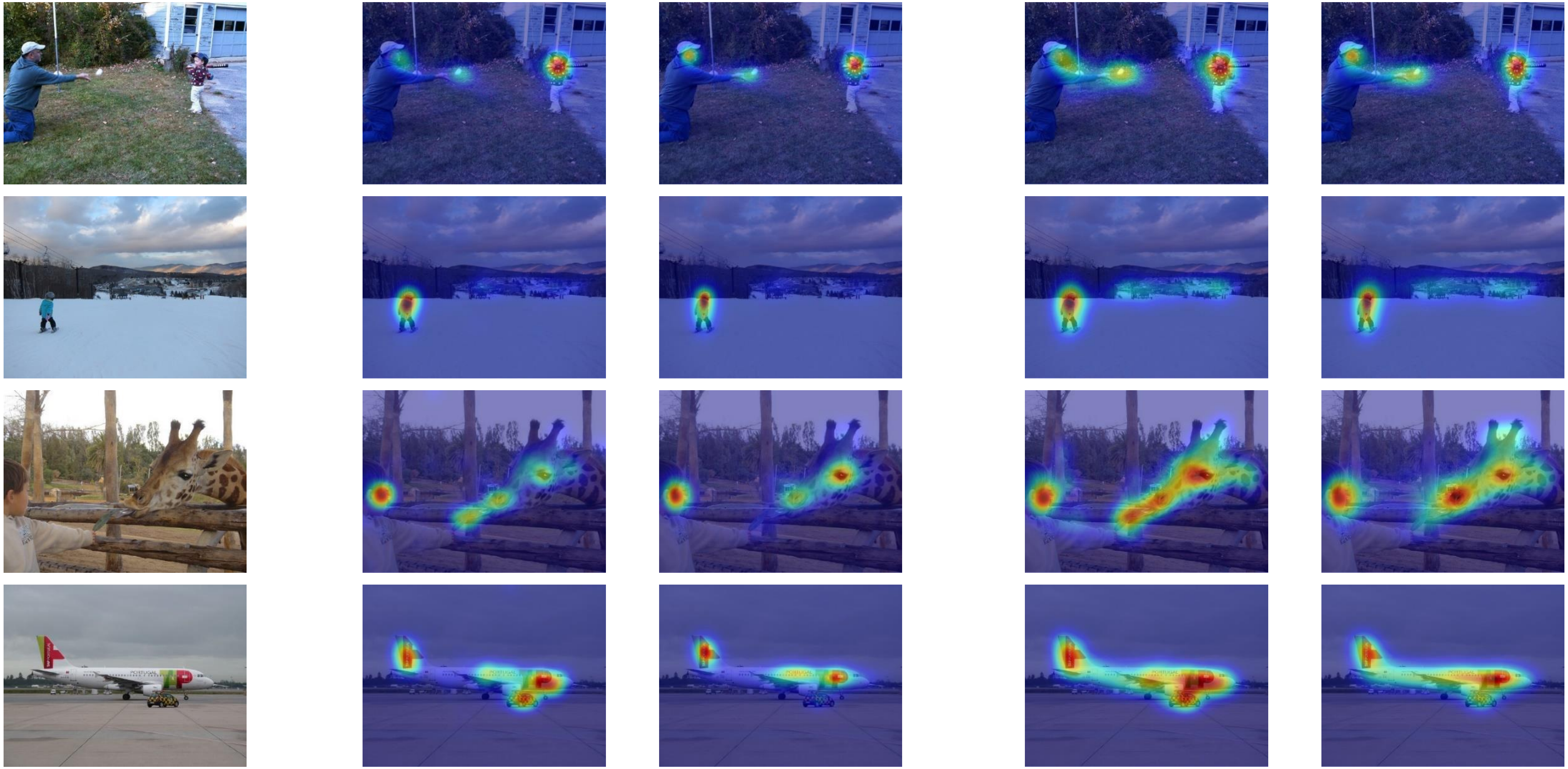
	CC	sAUC	AUC	NSS
SAM-ResNet	**0.89**	0.58	**0.88**	**2.38**
SAM-VGG	**0.89**	0.58	**0.88**	**2.38**
DeepFix [6]	0.87	0.58	0.87	2.28
MixNet [2]	0.76	0.58	0.86	1.92
iSEEL [8]	0.66	**0.59**	0.84	1.67

[1] Cornia et al. "A Deep Multi-Level Network for Saliency Prediction." ICPR, 2016.
[2] Dodge et al. "Visual Saliency Prediction Using a Mixture of Deep Neural Networks." arXiv:1702.00372, 2017.
[3] Pan et al. "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks.", arXiv:1701.01081 2017.
[4] Pan et al. "Shallow and Deep Convolutional Networks for Saliency Prediction." CVPR, 2016.
[5] Kümmerer et al. "DeepGaze II: Reading fixations from deep features trained on object recognition." arXiv:1610.01563, 2016.
[6] Kruthiventi et al. "DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations." arXiv:16rXiv:1510.02927, 2015.
[7] Huang et al. "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks." ICCV, 2015.
[8] Tavakoli et al. "Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features." Neurocomputing, 2016.

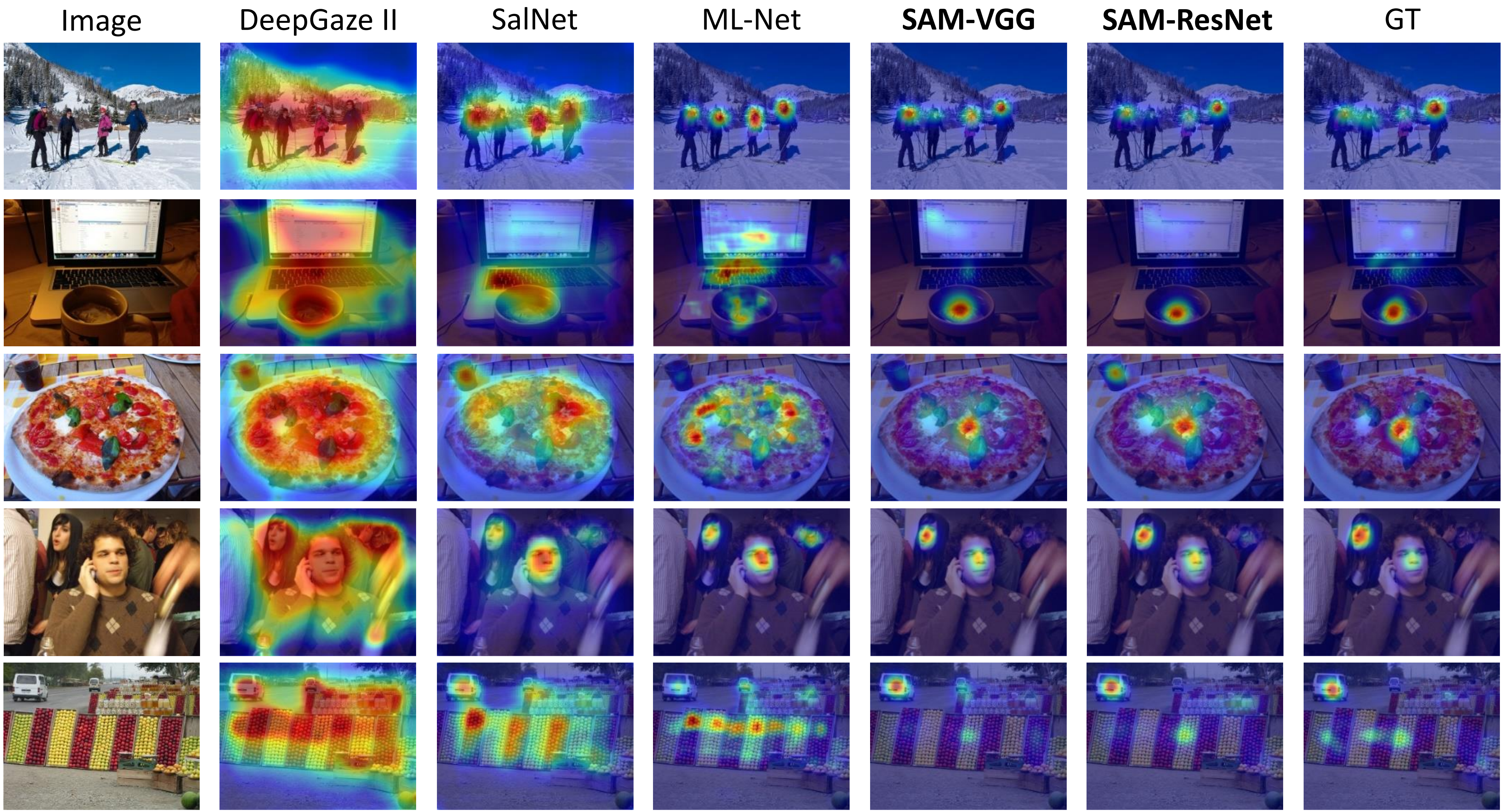
Qualitative results

SALICON (original release) | SALICON (new release)

Image | Groundtruth | SAM-ResNet | Groundtruth | SAM-ResNet

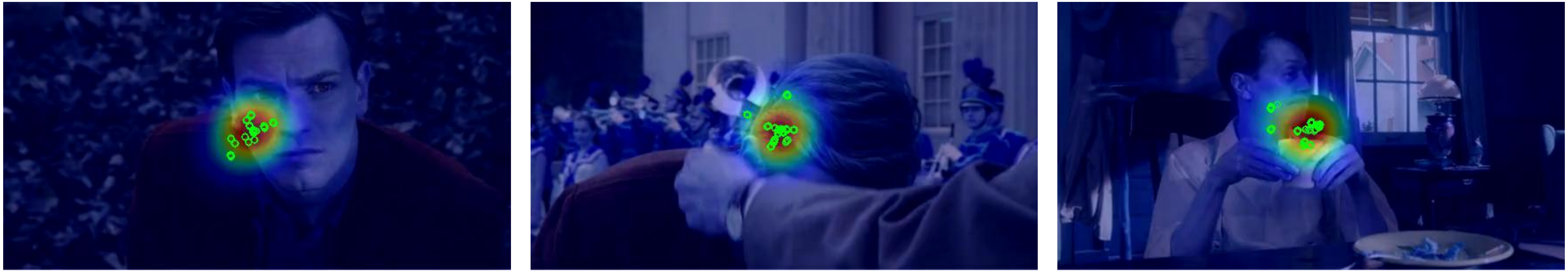


Qualitative results



Qualitative results (Hollywood2 dataset)

Groundtruth



SAM

Saliency for automatic image cropping

- Being saliency a proxy of visual interestingness, we apply it to automatic image cropping
- The problem can be casted as that of finding a rectangular region *R* with maximum saliency.
  - Which boils down to finding the **minimum bounding box** of all salient pixels above a threshold

Datasets

- Flickr-Cropping dataset
  - 1,743 images, associated with *crowd-sourced* annotations
  - 1,395 for training, 348 for test
- CUHK Image Cropping dataset
  - 950 images cropped by *experienced photographers*
  - 3 annotations for each image

Metrics

- Intersection-over-union (area)
- Boundary Displacement Error (distance between sides)

$$\mathrm{IoU} = \frac{1}{N}\sum_{i}^{N}\frac{GT_i \cap P_i}{GT_i \cup P_i}$$

$$\mathrm{BDE} = \frac{1}{4}\frac{1}{N}\sum_{i}^{N}\left(\frac{|x_1^{GT_i} - x_1^{P_i}|}{w_i} + \frac{|y_1^{GT_i} - y_1^{P_i}|}{h_i} + \frac{|x_2^{GT_i} - x_2^{P_i}|}{w_i} + \frac{|y_2^{GT_i} - y_2^{P_i}|}{h_i}\right)$$

Results on Flickr-Cropping dataset

Two baselines:

- *Saliency density*: maximizes the difference of averaged saliency between the selected BB and the outer region
- *VGG activations*: saliency maps are replaced with activations from the last convolutional layer of the VGG-16

Method	Avg IoU	Avg BDE
eDN [1]	0.4857	0.1372
RankSVM+DeCAF$_7$ [1]	0.6019	0.1060
VFN [2]	**0.6744**	**0.0872**
A2-RL [3]	**0.6564**	**0.0914**
Saliency Density	0.6193	0.0997
VGG Activations	0.6004	0.1088
Ours	**0.6589**	**0.0892**

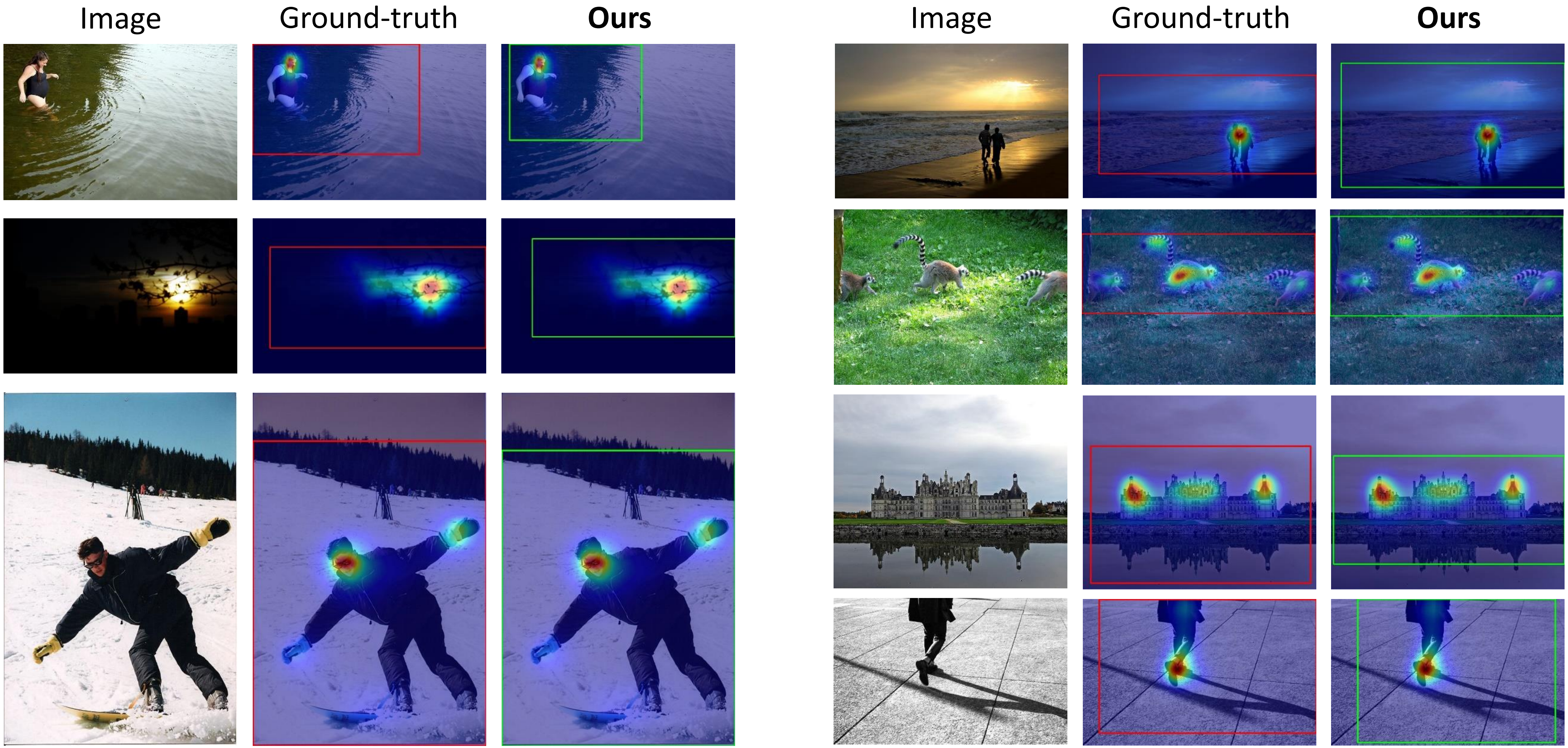
[1] Chen et al. "Quantitative analysis of automatic image cropping algorithms: A dataset and comparative study." *WACV*, 2017.
[2] Chen et al. "Learning to compose with professional photographs on the web." *arXiv preprint arXiv:1702.00503*, 2017.
[3] Li et al. "A2-RL: Aesthetics Aware Reinforcement Learning for Automatic Image Cropping." *arXiv preprint arXiv:1709.04595*, 2017.

Results on CUHK dataset

Annotation	Method	Avg IoU	Avg BDE
1	LearnChange [30]	0.7487	0.0667
	VFN [7]	**0.7847**	**0.0581**
	A2-RL [17]	**0.7934**	**0.0545**
	Saliency Density	0.6345	0.0971
	VGG Activations	0.7788	0.0574
	Ours	**0.8017**	**0.0500**
2	LearnChange [30]	0.7288	0.0720
	VFN [7]	**0.7763**	**0.0614**
	A2-RL [17]	**0.7911**	**0.0554**
	Saliency Density	0.6053	0.1075
	VGG Activations	0.7648	0.0624
	Ours	**0.7711**	**0.0594**
3	LearnChange [30]	0.7322	0.0719
	VFN [7]	0.7602	0.0653
	A2-RL [17]	**0.7826**	**0.0551**
	Saliency Density	0.6153	0.1040
	VGG Activations	**0.7612**	**0.0618**
	Ours	**0.7675**	**0.0599**

[1] Yan et al. "Learning the change for automatic image cropping." *CVPR*, 2013.
[2] Chen et al. "Learning to compose with professional photographs on the web." *arXiv preprint arXiv:1702.00503*, 2017.
[3] Li et al. "A2-RL: Aesthetics Aware Reinforcement Learning for Automatic Image Cropping." *arXiv preprint arXiv:1709.04595*, 2017.

Qualitative Results



Application to Historical Manuscripts

- We apply our image cropping approach to select the best pages to represent historical manuscripts.
- Application: improvement of the navigation of historical digital libraries: users can visually identify the content of a book watching its most representative images, without the need of opening it.
- Visually representative pages:
  - Those with a big contrast between salient and non salient regions
  - i.e., those that contain valuable details

Dataset

- A set of digitized manuscripts from the Estense Library Collection (Modena)

Qualitative Results

Qualitative Results

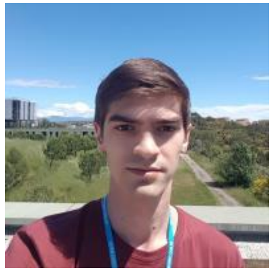
Qualitative Results

Thank you!

Any question?

lorenzo.baraldi@unimore.it
http://aimagelab.ing.unimore.it

Marcella Cornia



Stefano Pini

Lorenzo Baraldi

Rita Cucchiara