UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Imagelab

Joined Directions in Video Understanding: Saliency & Captioning

Rita Cucchiara

Imagelab, Dipartimento di Ingegneria Enzo Ferrari

University of Modena e Reggio Emilia, Italy

Talk @Stanford July 18, 2017

Agenda

Imagelab

Saliency

Attention

Captioning

Question Time

Dr(eye)Ve
[IEEE IV2017]

Ml-Net
***S**aliency **A**ttentive **M**odel*
[CVPRW 2017]

Captioning &
Attention&Saliency

BA-Encoding
for Video captioning
[CVPR2017]

Neuralstory
[IEEE Trans. MM2016]

UNIMORE
UNIVERSITÀ DEGLI STUDI DI
MODENA E REGGIO EMILIA

Università di Modena e Reggio Emilia , Italy

Since 1175, about 30.000 students

15 Departments, 9 Research Centers

Dipartimento di Ingegneria “Enzo Ferrari”

Rank #1 in Italy for Engineering Courses (Censis 2017)

Imagelab @UNIMORE 2017




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

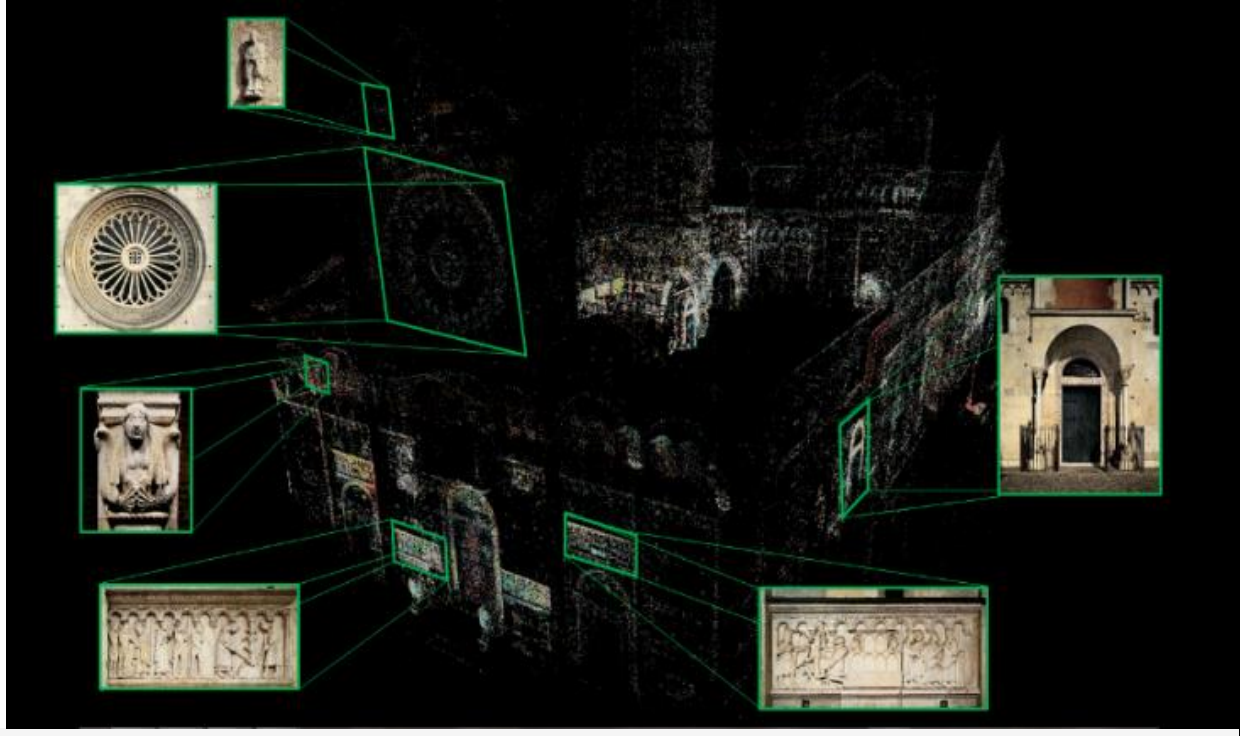
- ✓ Together with
- Panasonic (USA)
- Ferrari (I) , Maserati(I)
- Facebook FAIR (F), Eurecom (F)
- RAI, Almaviva, ..
- ISCRA Italian SuperComputing Resource Allocation – CINECA
- Many smes, EU and Italian public bodies

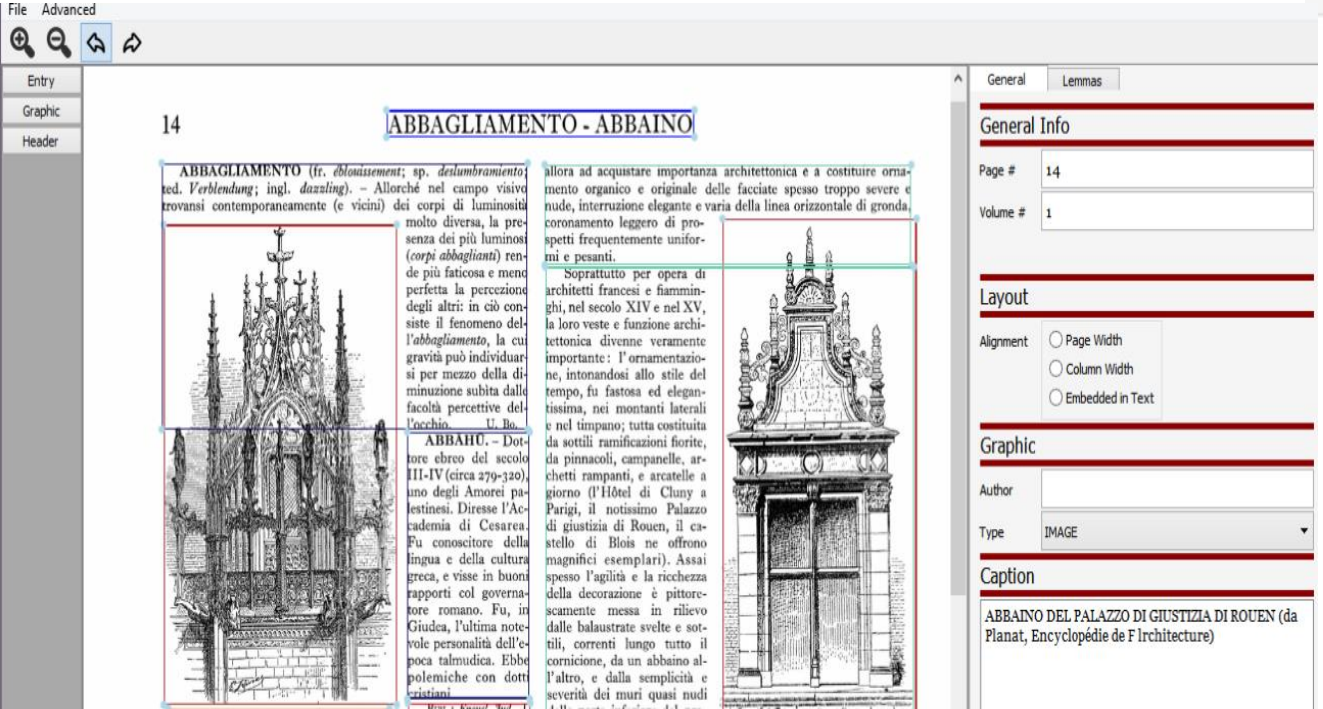
www.imagelab.unimore.it

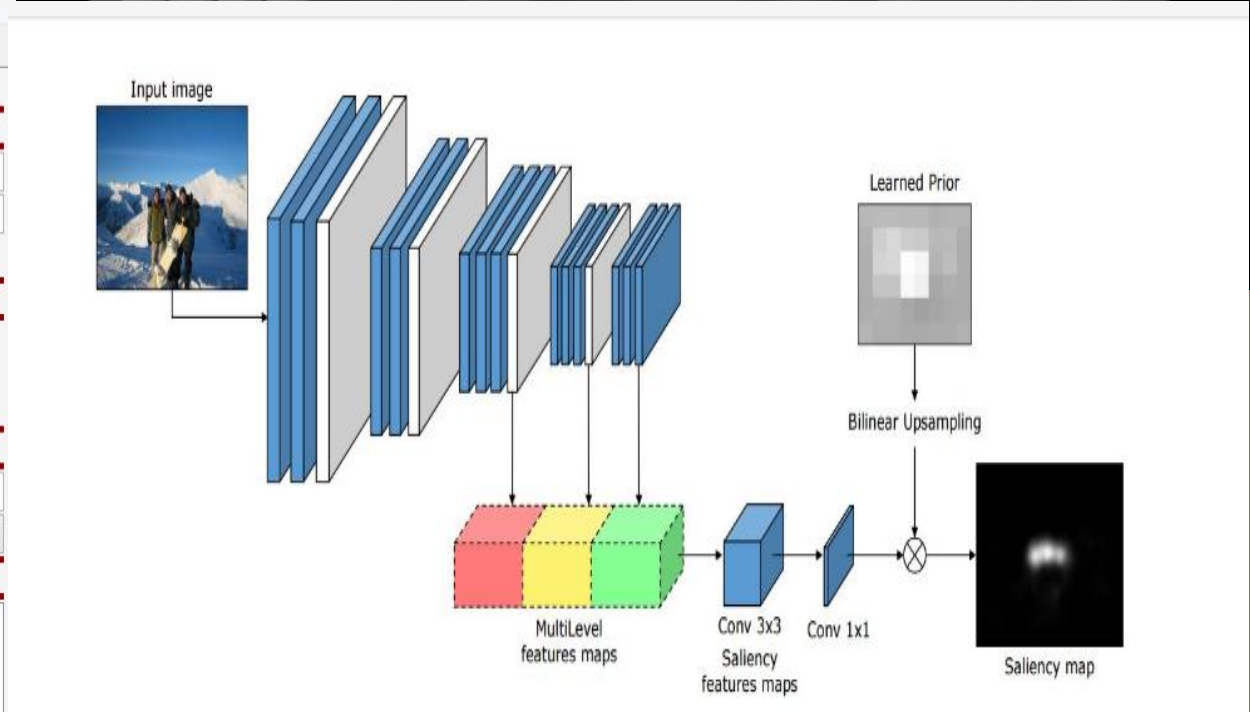
RESEARCH ACTIVITY @IMAGELAB

COMPUTER VISION, PATTERN RECOGNITION AND MULTIMEDIA

- Saliency Analysis
- Image processing , Labeling for Document Analysis
- 3D reconstruction in Egocentric Vision
- Video temporal segmentation, summarization
- Image and Video captioning



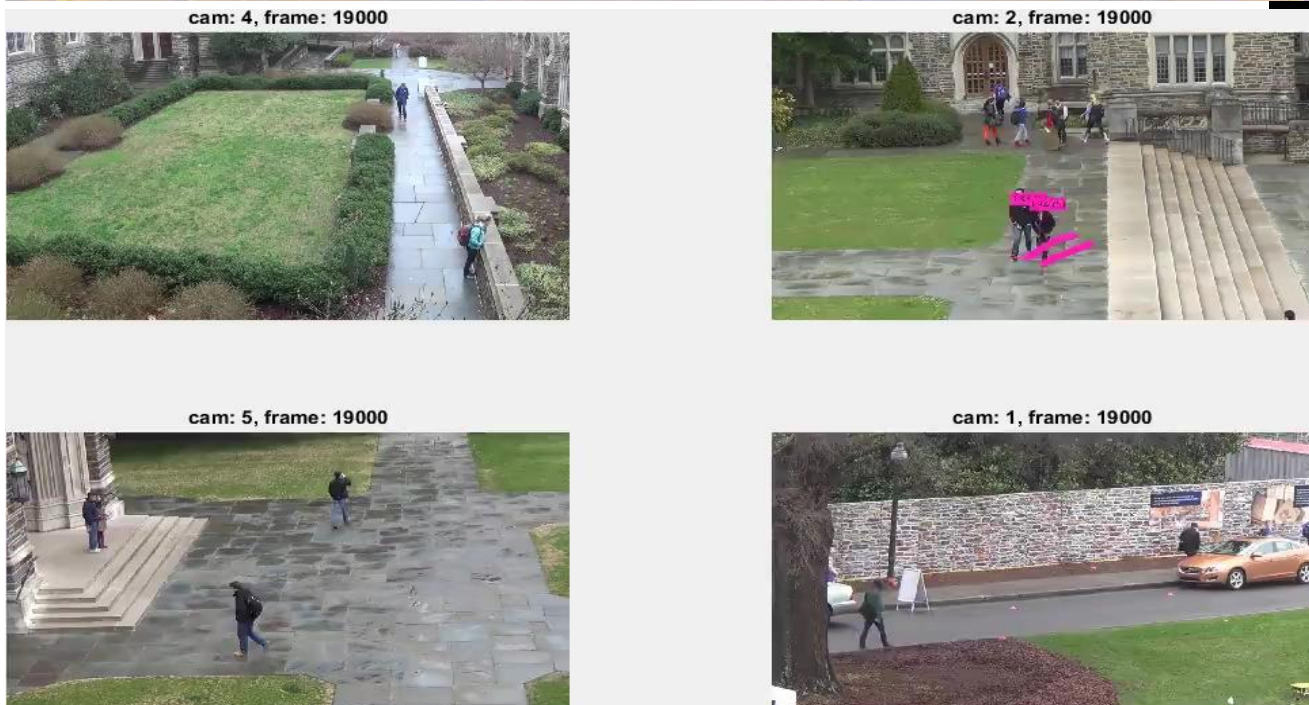




RESEARCH ACTIVITY @IMAGELAB

SURVEILLANCE AND HBU

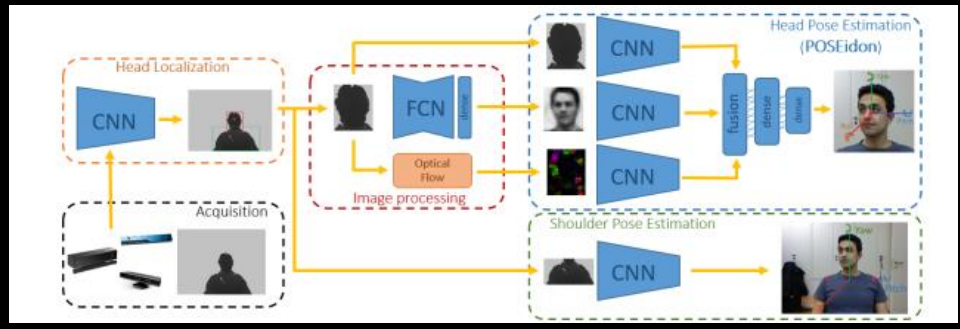
- Tracking, Tracking, Tracking
- Crowd analysis
- Surveillance in working area
- HBU: Gesture and Egocentric Vision
- Tracking body motion in sports (with saliency)
- People super-resolution with GAN



RESEARCH ACTIVITY @IMAGELAB

AUTOMOTIVE AND INDUSTRY

- Human Attention in driving
- Human Car Interaction
- Autonomous driving in smart cities
- Collaborative robotics



Other Activities at Imagelab

- ✓ Industrial research with companies in the Interdip. Center of research in ICT Softech-ICT

MuMeT 2017
visual computing and multimedia technologies

- ✓ International Phd School in ICT
- ✓ Master Mumet Ed 2017. « Visual Computing and Multimedia Technology in the Deep Learning Era»
- ✓ Laurea and Laurea Magistrale in Computer Engineering

- ✓ Scientific Organization: ICCV2017 Venezia, ICPR2020 Milano

LET'S START: SALIENCY AND ATTENTION

(thanks to **Marcella Cornia,** and Giuseppe Serra, Lorenzo Baraldi

and for Dr(eye)ve **Andrea Palazzi**, Stefano Alletto, and Simone Calderara)

From Egocentric Vision

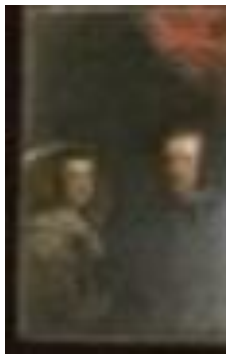
to Attention

to Saliency

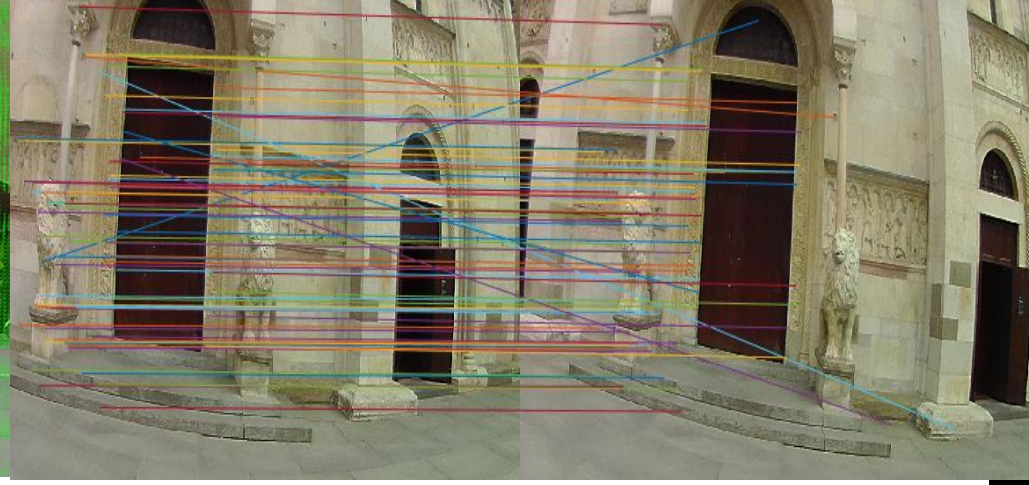
Understanding what a person sees

exploiting similar learning, perception and vision paradigms

for first-person view vision



Attentive behavior?



Attention, No blur, good matching

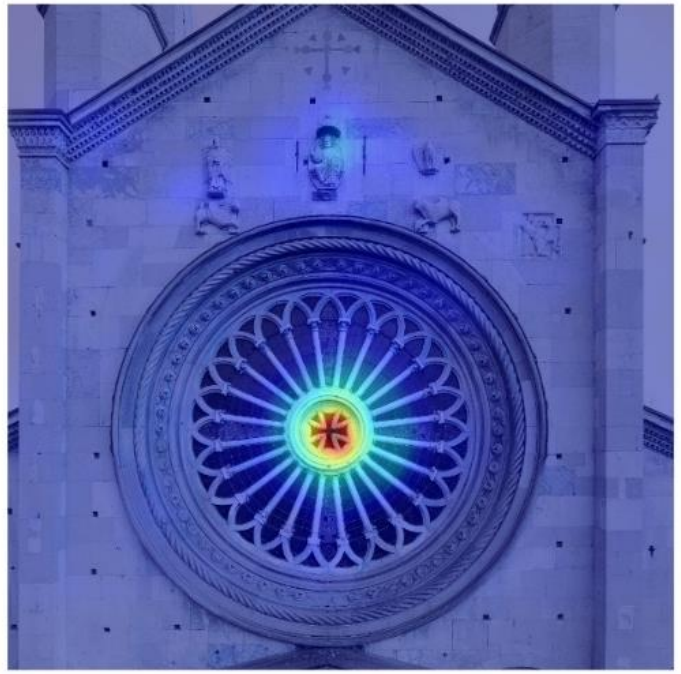
Head-body motion, Blur, no matching

Understanding Attentive behavior by head motion
Video Summarization for CH at Imagelab: [P.Varini, et al *ACM MM*, 2015],







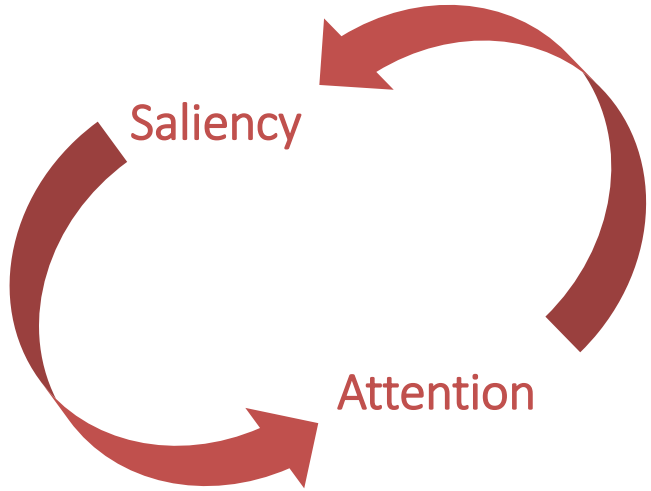
Before Attention , Saliency

Saliency: what can pop out in an image or in a view.

A bottom-up pre-attentional mechanism of human visual behavior

Saliency and Attention in Neuroscience

Saliency detection in humans is a key attentional mechanism that facilitates learning and survival by enabling organisms to focus their limited perceptual and cognitive resources on the most pertinent subset of the available sensory data.

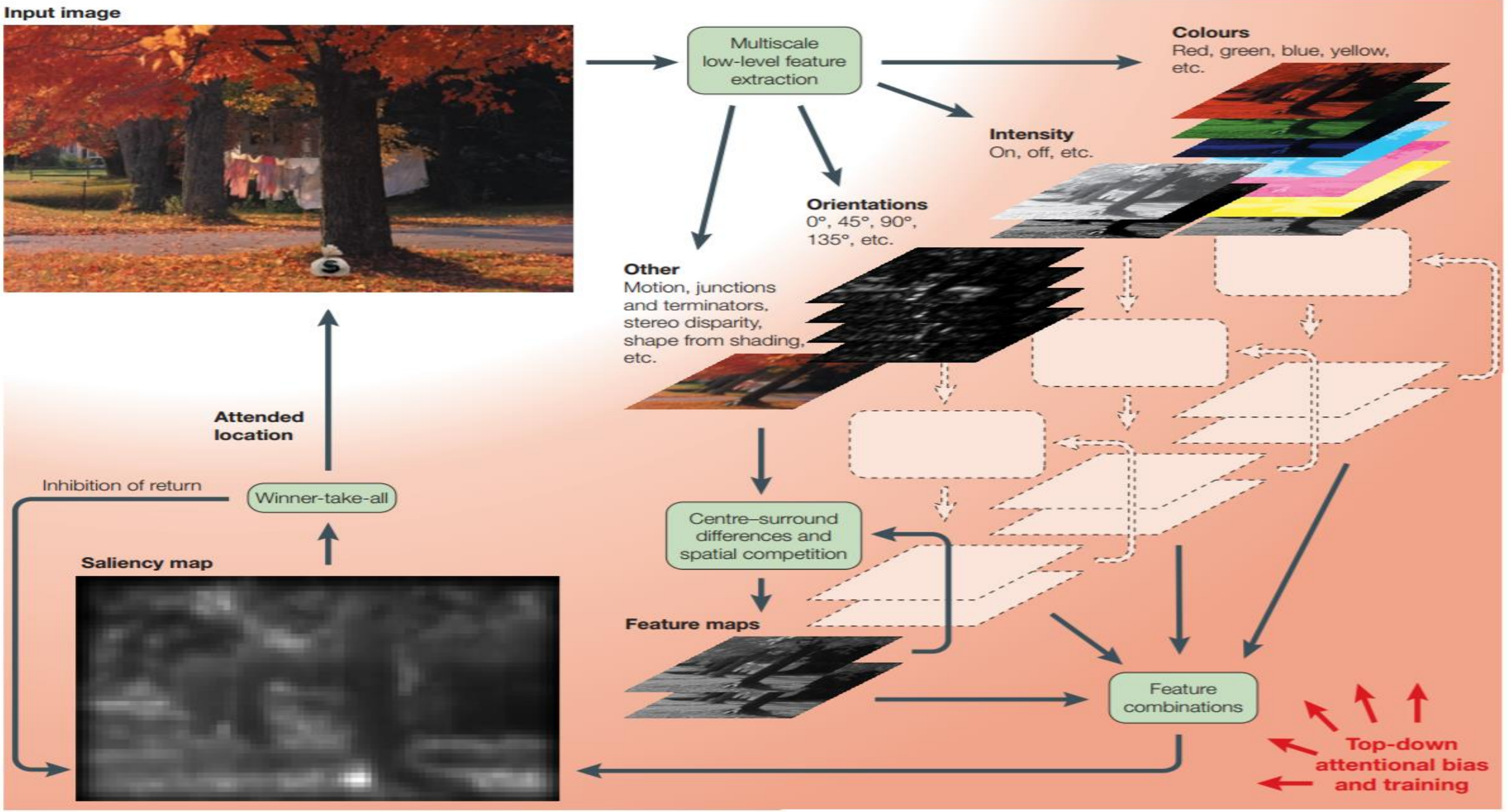


Two forms of attention:

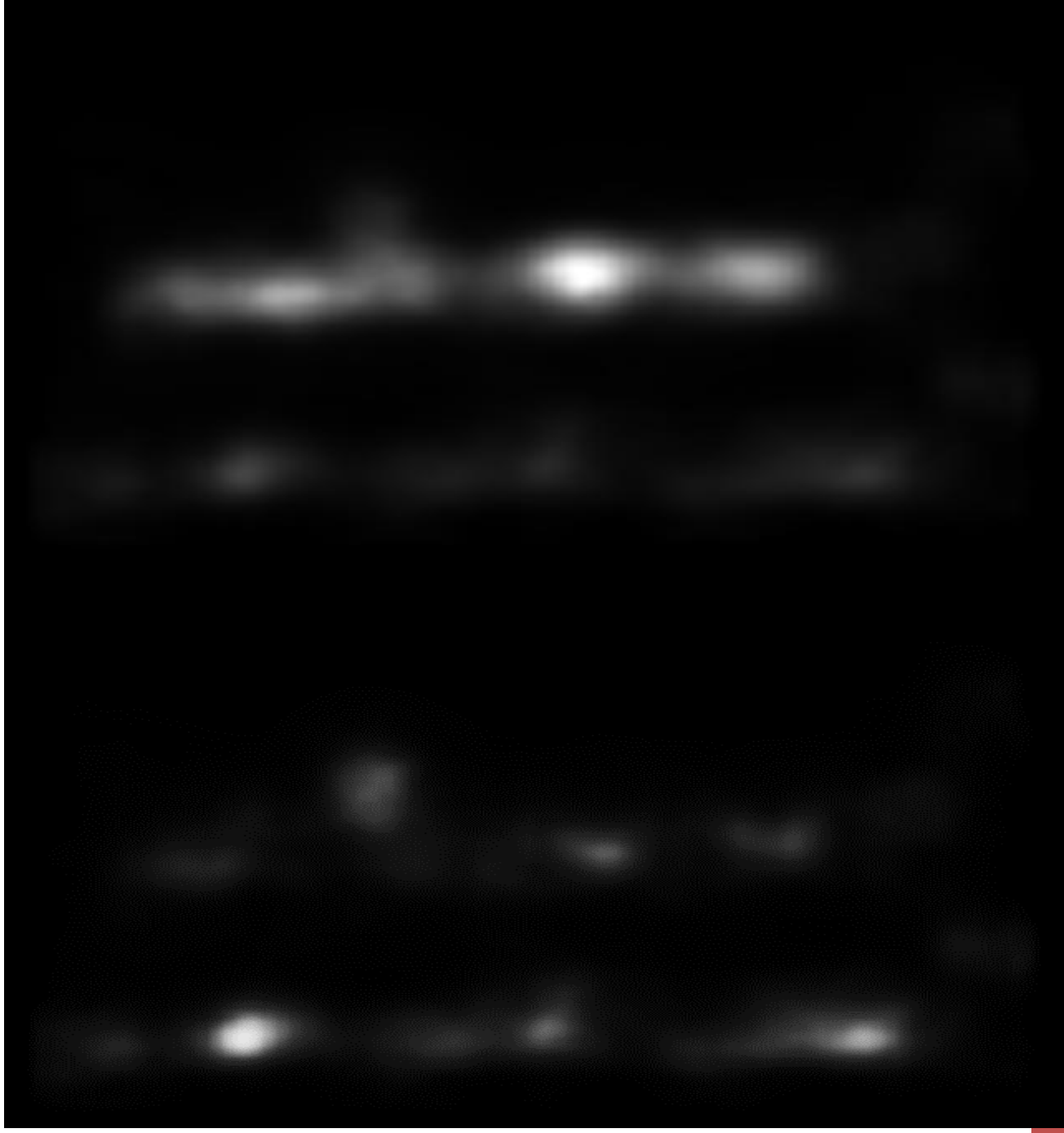
- Initial, Bottom-up purely data driven, guided by Saliency
- Refined, Task-driven and purposive

Treisman &Gelade A feature-integration theory of attention Cogn. Psych. 1980

Koch &Ulmann , Shift in selective visual attention: toward the underlying neural circuitry Hum. Neurobiol., 1985



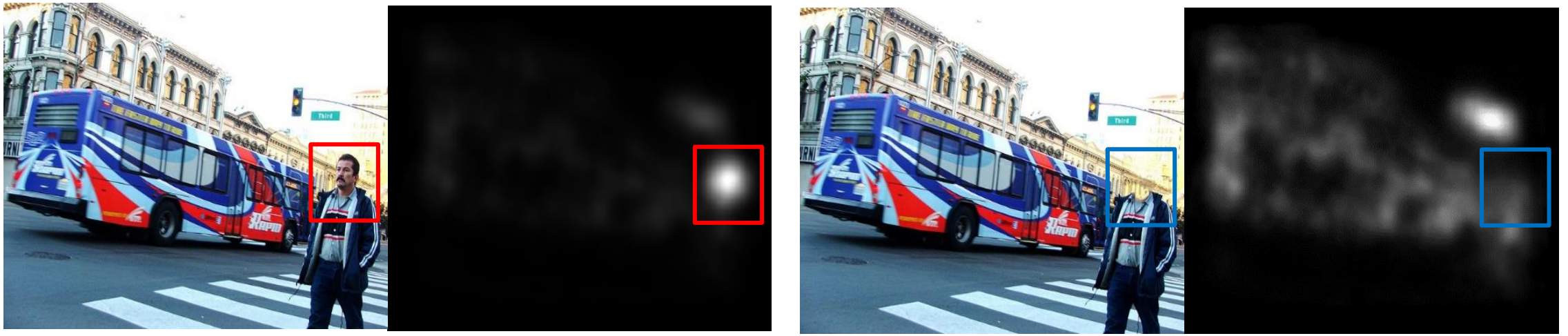
✓ "Saliency map": an image map representing areas of saliency [Itti and Koch PAMI '89, Nature Reviews 2001]



Saliency: data-driven, memory and knowledge-based or semantics driven?

Computational detection of saliency maps

- ✓ LOW LEVEL FEATURES
  - *'80—2000* ***Itti Koch***: combination color+gradient+orientation in a winner-take-all unsupervised neural network
  - *2006 NIPS* ***Perona et al.*** *Graph-based Visual* Saliency as a graph of low level features
  - ..
- ✓ ADDING MEMORY and knowledge-based higher level FEATURES (Faces, people, text..)
  - *2009 ICCV Torralba et al*
  - *2012 ICPR Biorg ICPR*
  - *2013 ICCV Sclaroff et al.*



Thanks to Marcella Cornia's thesis

The deep learning era

Problems of annotated Data

- ✓ *2014 ICCV Vig et al.:* a three Convnet layers network
- ✓ *2015 ICRLW Kummerer et al:* DeepGaze I with Alexnet (then 2016 ArXiv : Deepgaze II with VGG19)
- ✓ *2015 CVPR Lin et al* : data augmentation with image similarity

DATASETS:

MIT300 (Itti, Torralba et al) more than 70 competitors since 2014

SALICON (Jiang et al 2015), 10000 images;

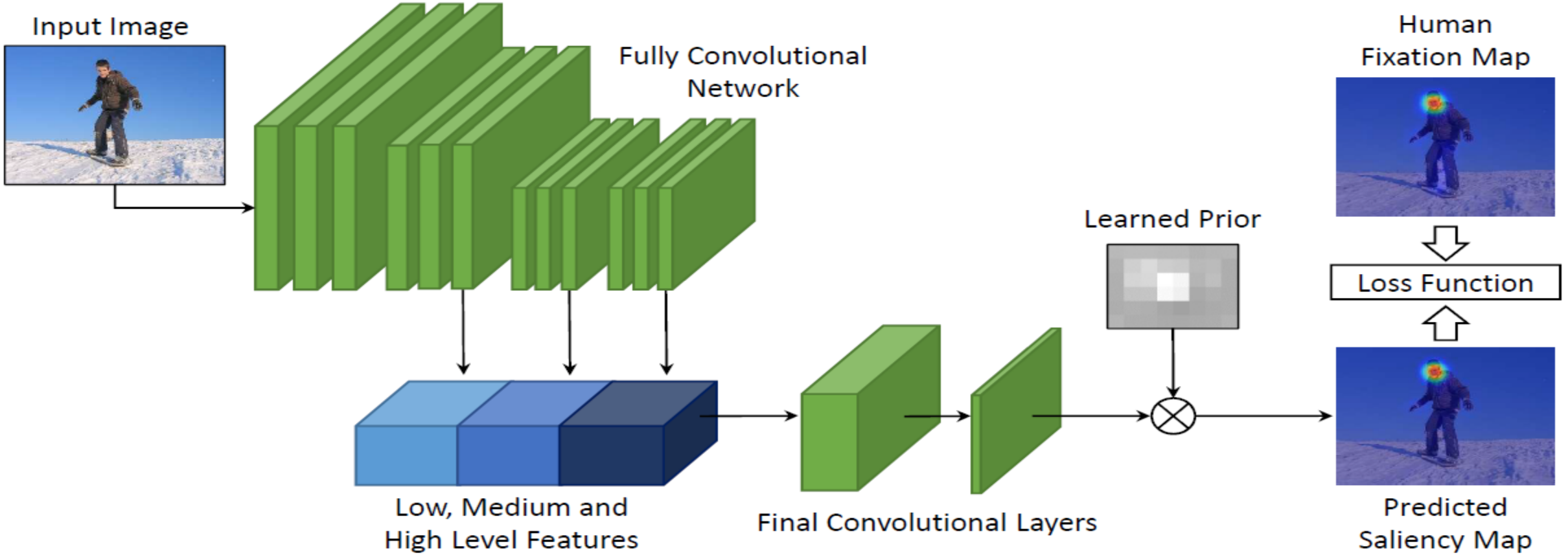
SALICON benchmark

- ✓ 2016 ECCV tutorial on Saliency
- ✓ 2017 ICME Competition on 360° Saliency
- ✓ 2017 CVPR New SALICON LSUN Competition……..

mit saliency benchmark

SALIENCY DETECTION @Imagelab : Multi-Level NET

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA



Marcella Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A *Deep Multi-Level Network for Saliency Prediction*, Proc. of ICPR 2016. (before at CVPRWoW 2016)

VGG-16; 5 blocks,13 Conv,
3 fully-connected layers + a center bias

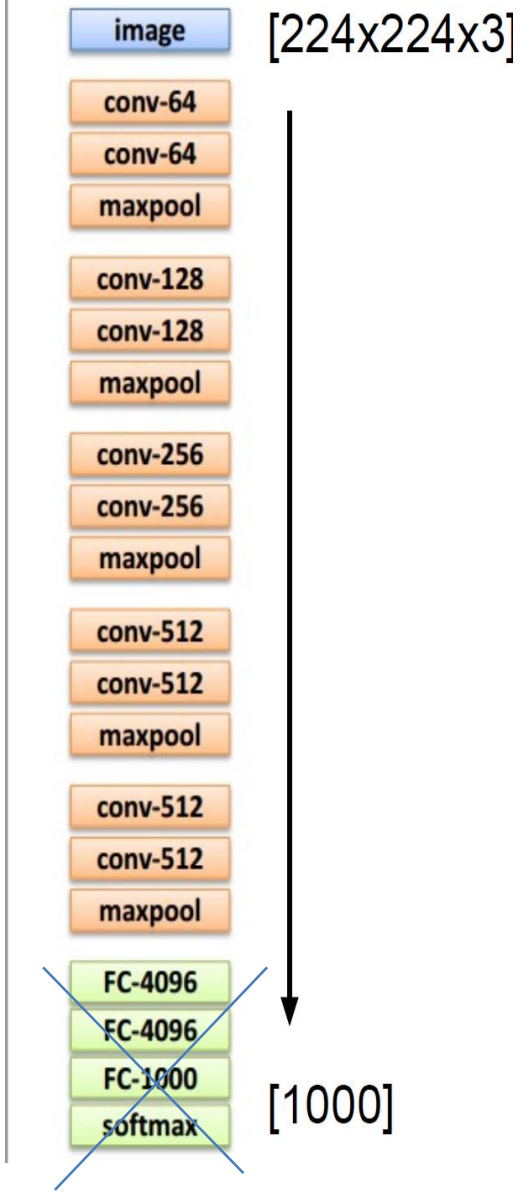
ML-NET Details

- ✓ Multi-level network: combine low-, medium- and high-level features
- ✓ A–priori learned map of central bias to emulate the center bias present in eye-fixation maps
- ✓ VGG-16; 13 layers, 5 convolutional blocks + 3 fully convolutional layers

$$L(\mathbf{w}) = \frac{1}{N}\sum_{i=1}^{N}\frac{\left\|\frac{\phi(\mathbf{x}_i)}{\max\phi(\mathbf{x}_i)} - \mathbf{y}_i\right\|^2}{\alpha - \mathbf{y}_i}$$

3 component Loss (trained with N samples with Stochastic Gradient Descent)

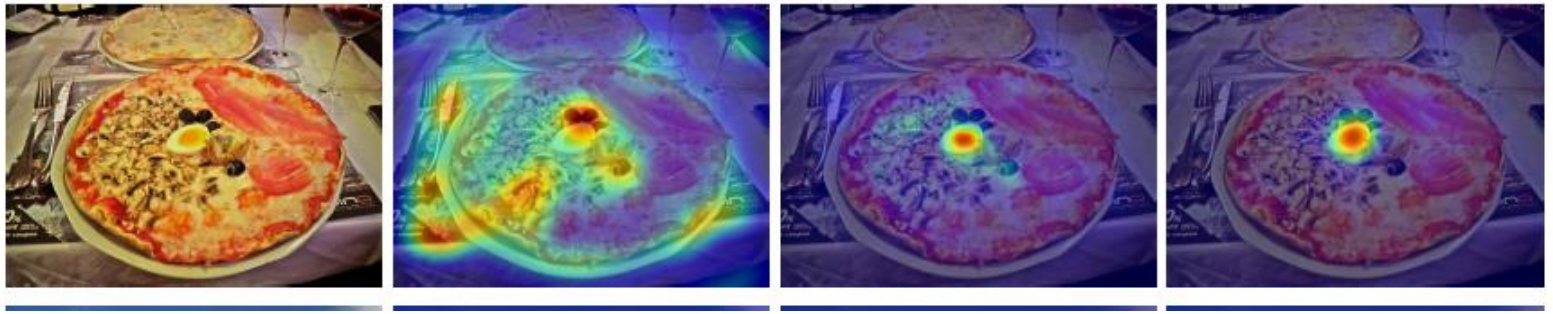
- A square error (Euclidean) loss *to make the **x** saliency similar to the **y** GT data*
- Normalized by the maximum of prediction, *not to be affected by intensity of saliency*
- Data are weighted by a linear function, *which gives more importance to pixel with high eye fixation probability*



ML-NET vs Itti and Koch: Metrics

Metrics Comparison as in [Bylinsky et al. ArXiv 2014]

Metrics	Location-based	Distribution-based
Similarity	AUC, sAUC, NSS, IG	SIM, CC
Dissimilarity		EMD, KL



Itti ML-NET humans

Table 1: Comparison results on the MIT300 dataset [21].

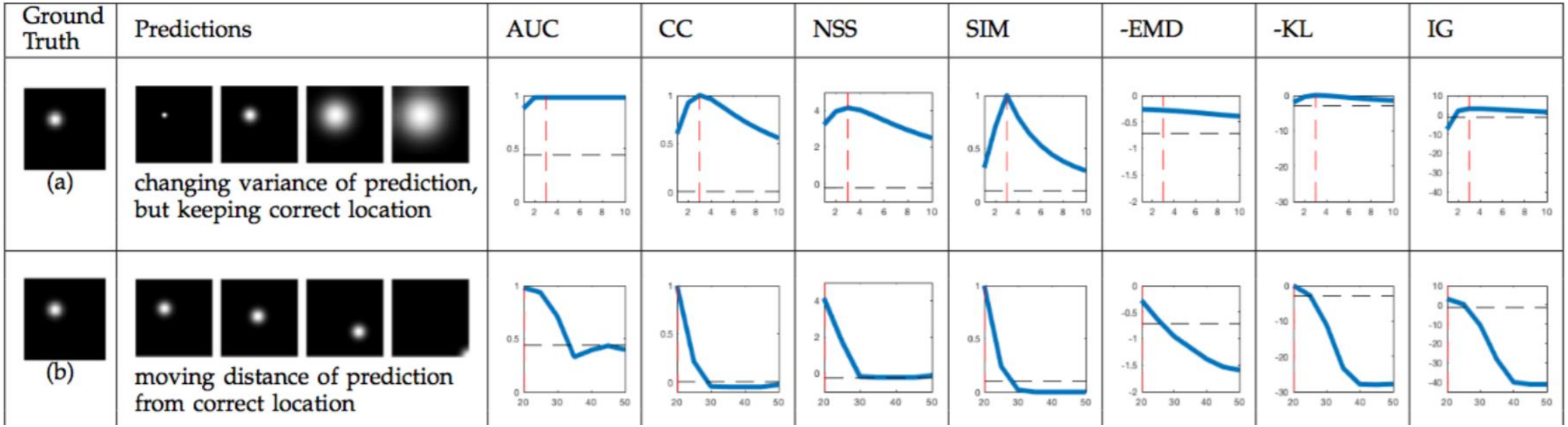
	SIM ↑	CC ↑	sAUC ↑	AUC ↑	NSS ↑	EMD ↓
Infinite humans	1.00	1.00	0.80	0.91	3.18	0.00
ML-Net	0.59	0.67	0.70	0.85	2.05	2.63
Itti	0.44	0.37	0.63	0.75	0.97	4.26

Metrics Details

- ✓ Defined for Saliency (Shuffled AUC, Information Gain, Normalized Scanpath Saliency)
- ✓ Adapted by signal detection (AUC) *[AUC not used in SALICON]*
- ✓ From image retrieval (SIMilarity, EMD)
- ✓ From information theory (KL-Divergence)
- ✓ From Statistics (Pearson Correlation Coefficient)

$$NSS(P, Q^B) = \frac{1}{N} \sum_i \overline{P_i} \times Q_i^B$$

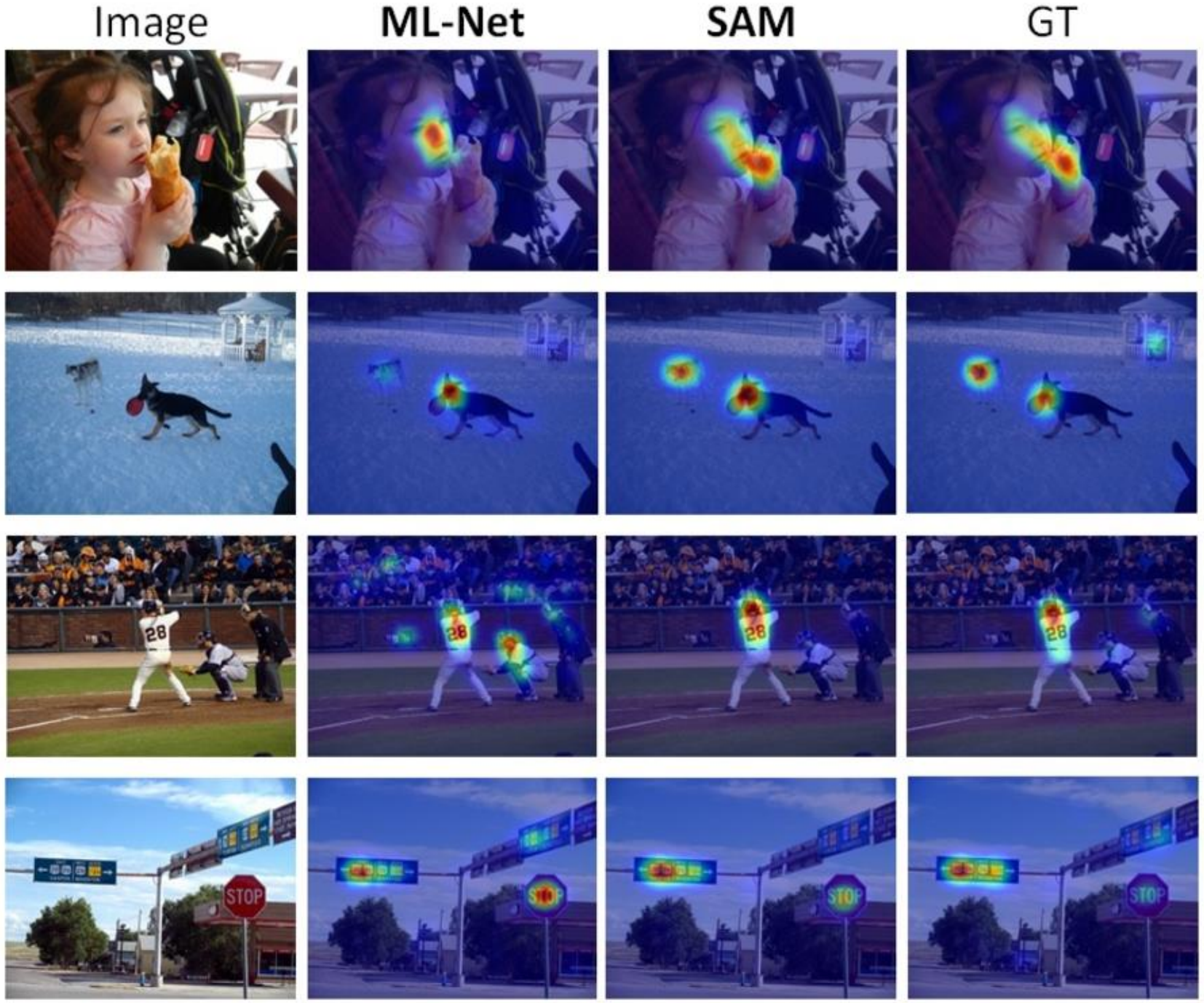
where $N = \sum_i Q_i^B$ and $\overline{P} = \frac{P - \mu(P)}{\sigma(P)}$



[Bylinsky et al. ArXiv 2014]

Saliency Attentive Model SAM

Improving saliency detection, iteratively with an attentive refinement



Marcella Cornia, L. Baraldi, G. Serra, and R. Cucchiara. *Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model*. *arXiv:1611.09571*, 2017 (submitted)

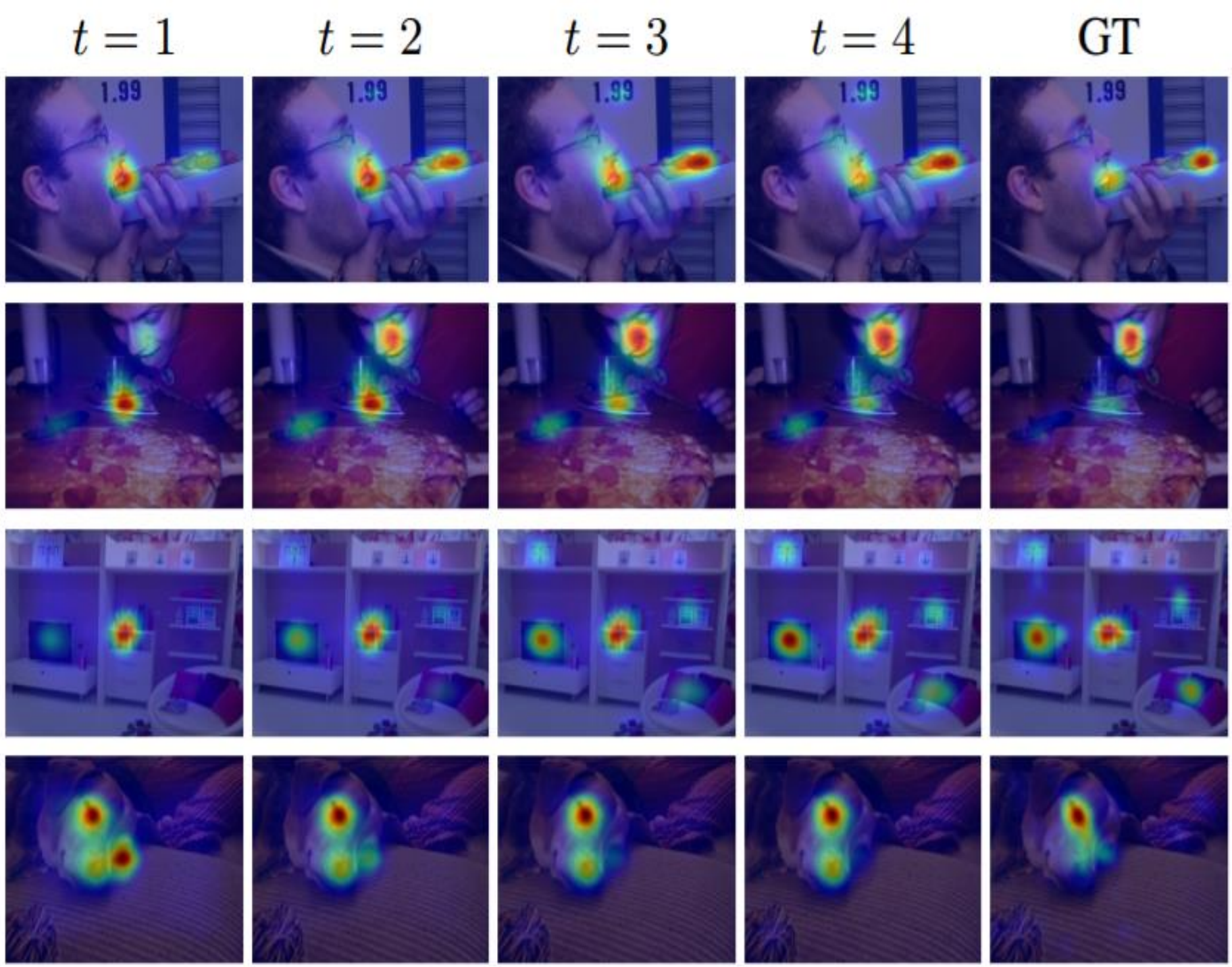
As a sort of

Pre-attentive scan-path

THE IDEA:

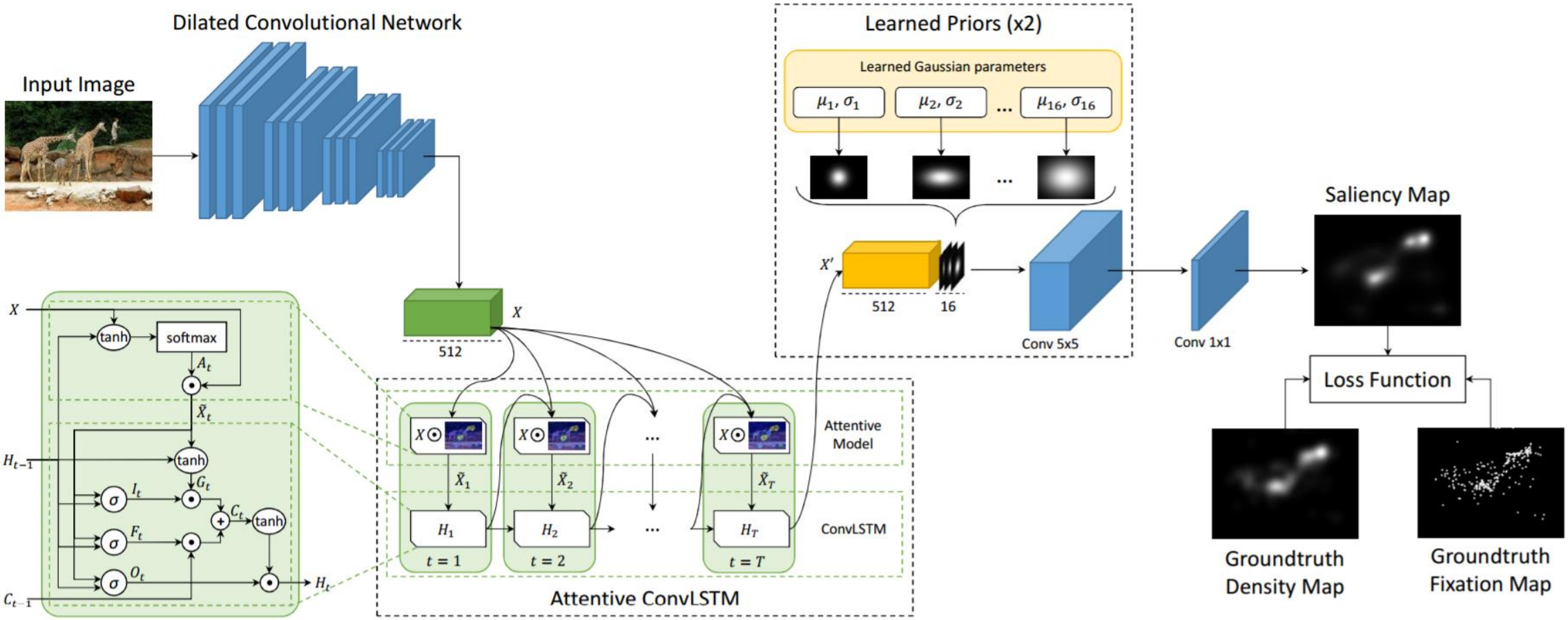
define a new **CONV-LSTM**

for scan the space and not the time



SALIENCY DETECTION @Imagelab SAM

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

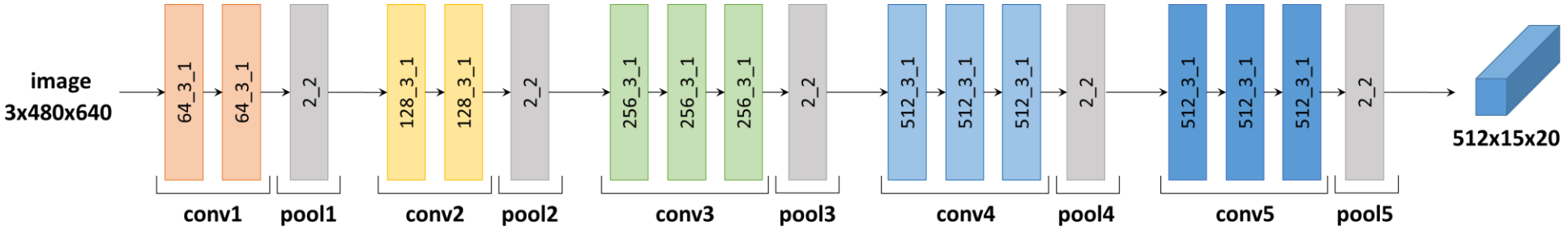


Saliency Attentive Model (SAM):
ML-NET+ LSTMs

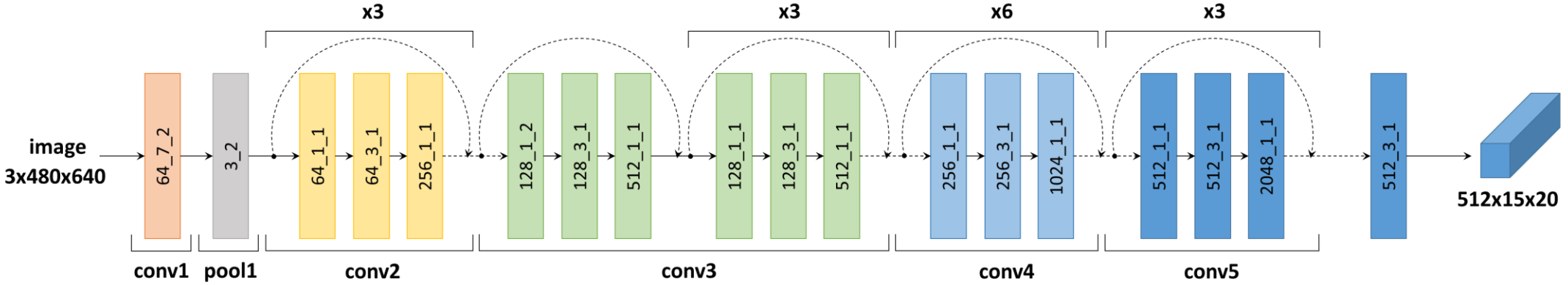
SAM details (1)

✓ 1 Features are a **STACK of channels** 512 x 30 x40 as output of

- A Dilated VGG



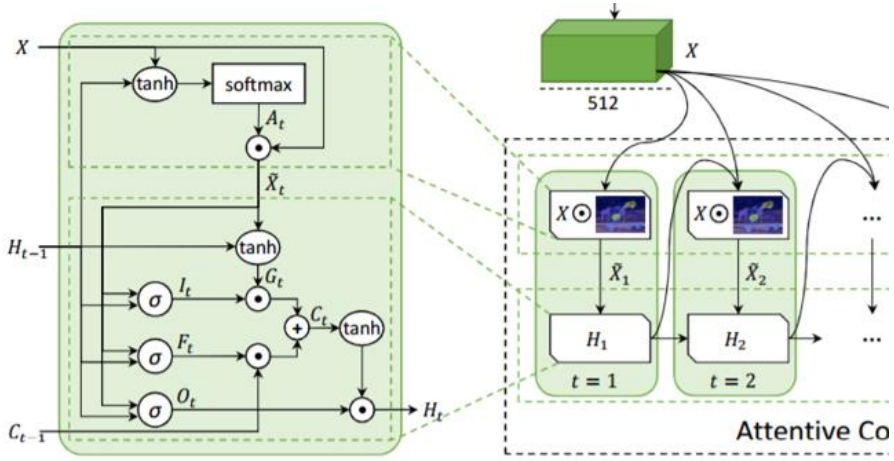
- A Dilated ResNet-50



- Removing fully-con layers and with stride 1:
- Dilated ConvNet as in [Yu , Koltun ICLR 2016]

SAM details (2)

- ✓ The attentive model is given by an **Attentive Conv LSTM**
- ✓ where input are not temporal series but spatial features and the dot product is substituted with convolution



$$I_t = \sigma(W_i * \tilde{X}_t + U_i * H_{t-1} + b_i)$$
$$F_t = \sigma(W_f * \tilde{X}_t + U_f * H_{t-1} + b_f)$$
$$O_t = \sigma(W_o * \tilde{X}_t + U_o * H_{t-1} + b_o)$$
$$G_t = \tanh(W_c * \tilde{X}_t + U_c * H_{t-1} + b_c)$$
$$C_t = F_t \odot C_{t-1} + I_t \odot G_t$$
$$H_t = O_t \odot \tanh(C_t)$$

here, the gates I_t, F_t, O_t, the candidate memory G_t, memory cell C_t, C_{t-1}, and hidden state H_t, H_{t-1} are 3-d tensors, each of them having 512 channels. $*$ represents the convolutional operator, all W and U are 2-d convolutional kernels, and all b are learned biases.

SAM details (3)

- ✓ The priori central bias is integrated in a single end-to-end pipeline
- ✓ 512 channel + 16 channel of Gaussian learned – a priori
- ✓ 528 channels than convolved in a single **Saliency map**

- ✓ Loss function
- ✓ The loss function is a combination of three measures

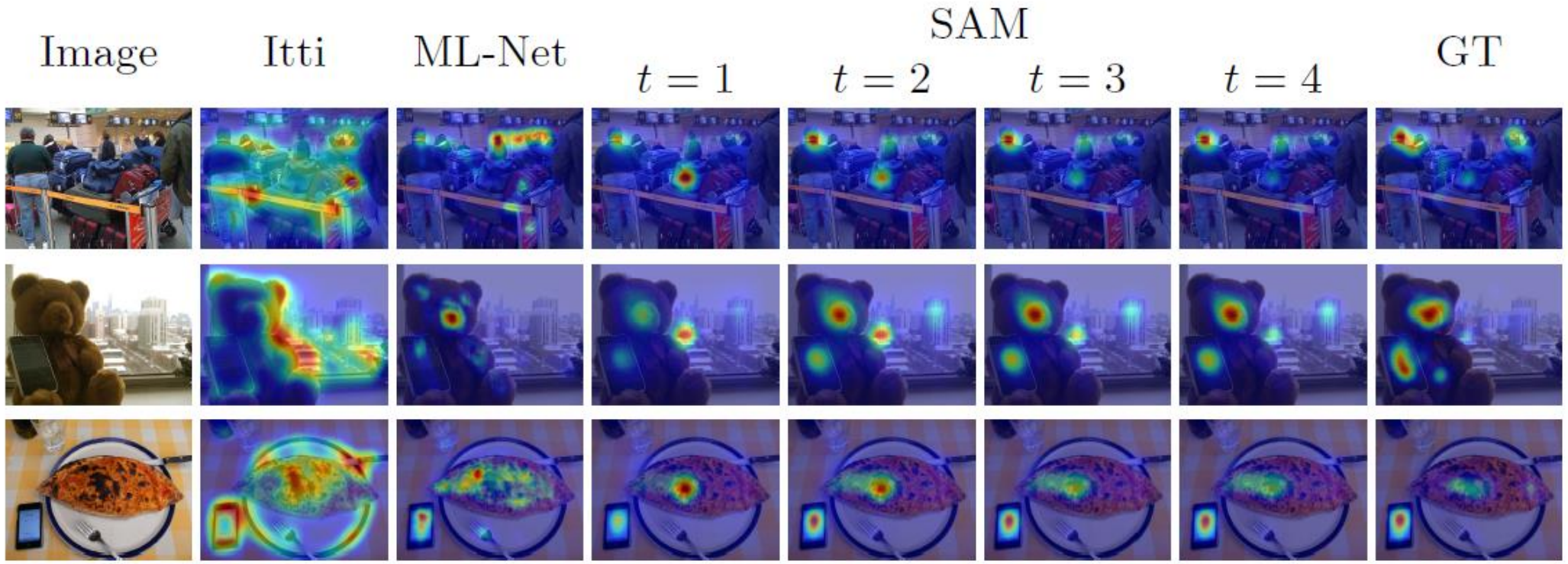
$$L(\tilde{\mathbf{y}}, \mathbf{y}^{den}, \mathbf{y}^{fix}) = \alpha L_1(\tilde{\mathbf{y}}, \mathbf{y}^{fix}) + \beta L_2(\tilde{\mathbf{y}}, \mathbf{y}^{den}) + \gamma L_3(\tilde{\mathbf{y}}, \mathbf{y}^{den})$$

- ✓ L1 Normalized Scanpath Saliency NSS
- ✓ L2 Linear Pearson's Correlation Coefficient CC
- ✓ L3 Kullback-Leiber divergence, KLD

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

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

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



Approaches share a similar model:

ITTI&Koch:

handcraft features + NN combination for winner-take-all

strongly biased by contour

DL-based and SAM:

convolutional learned features + NN combination (with an iterative LSTM)

surely **biased by collected data**

Performance analysis

SALICON Dataset (original release)

	CC	sAUC	AUC	NSS
SAM	**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

SALICON Dataset (new release)

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

Ongoing competition!

LSUN Challenge CVPR 2017

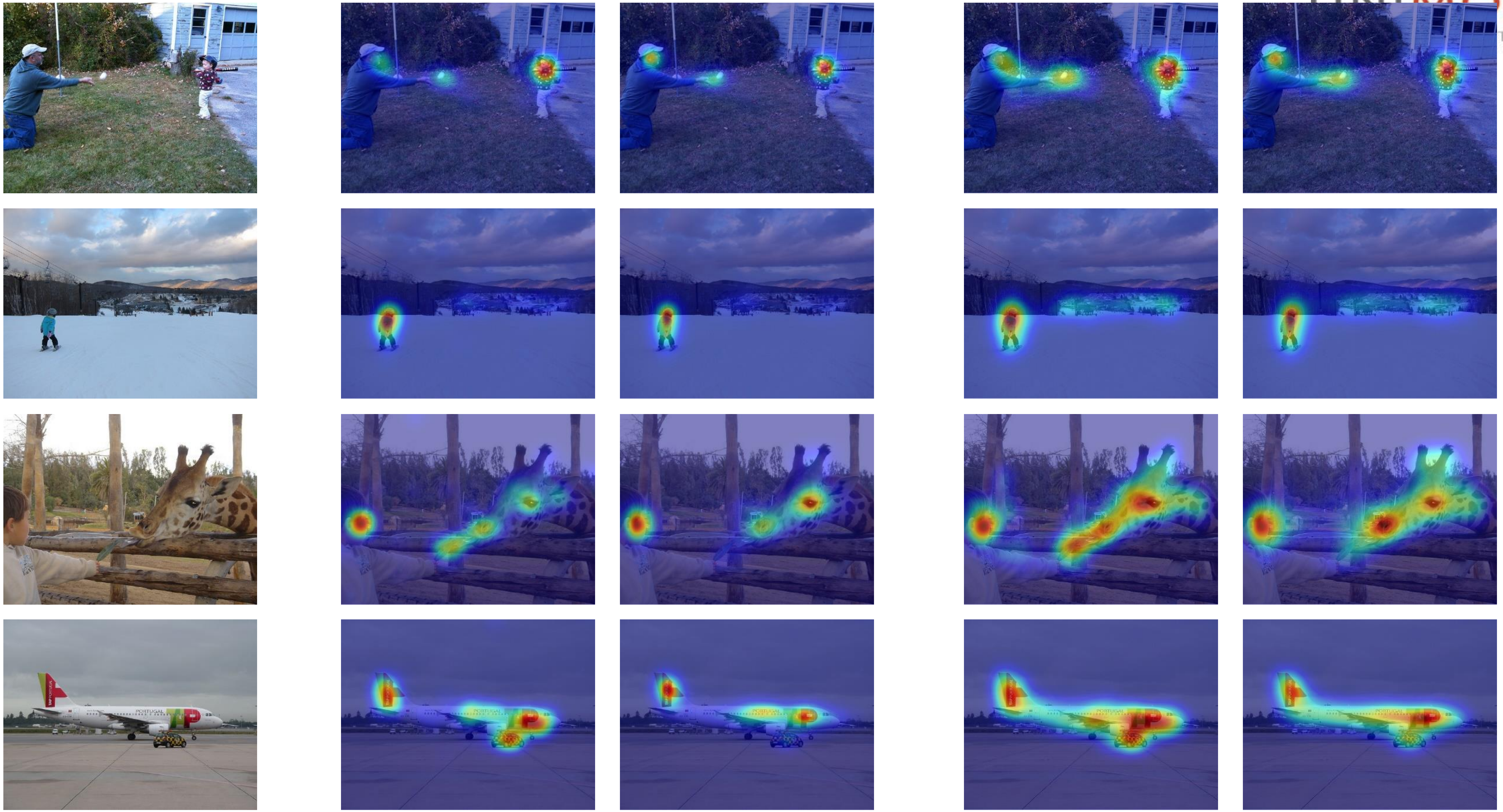
Results

#	User	SAUC ▲	IG ▲	NSS ▲	CC ▲	AUC ▲	SIM ▲	KL ▲
1	zhewuucas	0.726 (1)	0.738 (1)	1.841 (1)	0.860 (1)	0.859 (1)	0.756 (1)	0.318 (1)
2	sfdodge	0.710 (2)	0.315 (2)	1.698 (2)	0.726 (2)	0.836 (2)	0.646 (2)	0.767 (2)

Other results at 16/07

[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.

	SALICON (original release)		SALICON (new release)	
Image	Groundtruth	SAM	Groundtruth	SAM

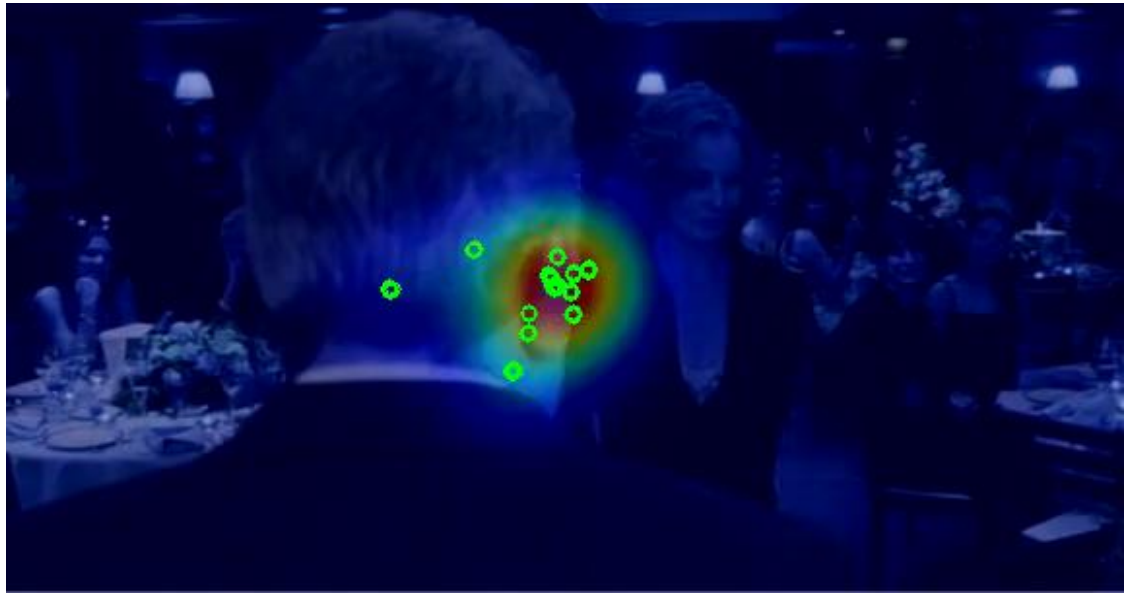


SAM for video

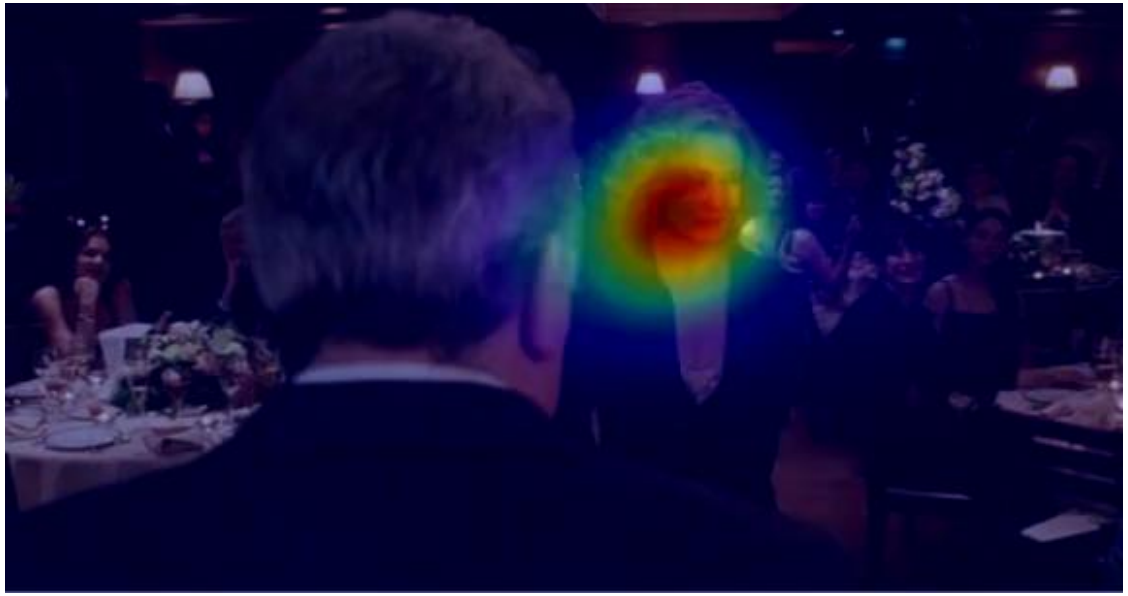
- ✓ Few approaches to video

Actions in the Eye (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

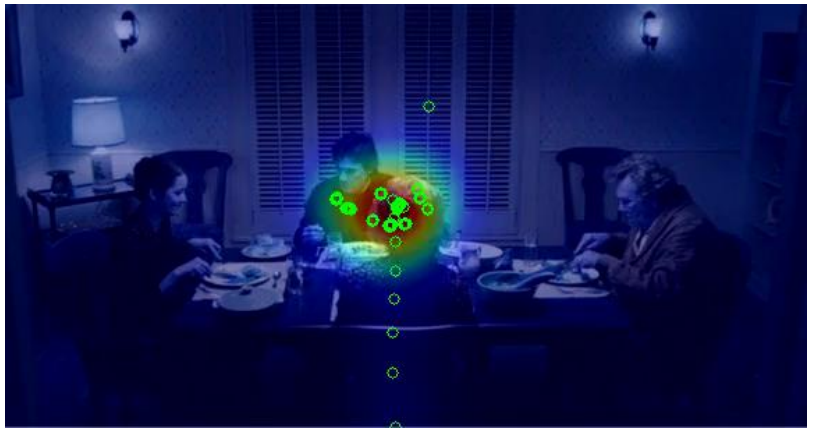


SAM

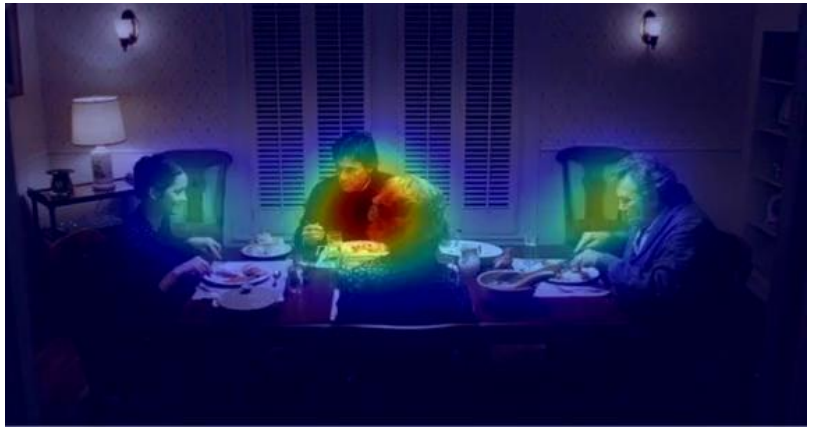
[1] Bazzani et al. “Recurrent Mixture Density Network for Spatio-temporal Visual Attention ." ICLR, 2017.

UNIMORE
UNIVERSITÀ DEGLI STUDI DI

Groundtruth



Actions in the Eye (Hollywood2) dataset



SAM

Saliency in task-driven video

Bottom-up saliency, detected by ML-NET, trained on SALICON on DR(EYE)VE dataset

http://imagelab.ing.unimore.it/dreyeve

Saliency not driven by a task..
as a passenger sees

Task-driven «Saliency» i.e. attentive behavior

Task Driven “saliency”, learned by C3D m odified by Dr(eye)ve project, trained on 15 frame windows

A.Palazzi; F.Solera; S.Calderara; S.Alletto R.Cucchiara "Learning Where to Attend Like a Human Driver" *Proc. of IEEE Intelligent Vehicles Symposium*, June 2017

The DR(eye)ve dataset

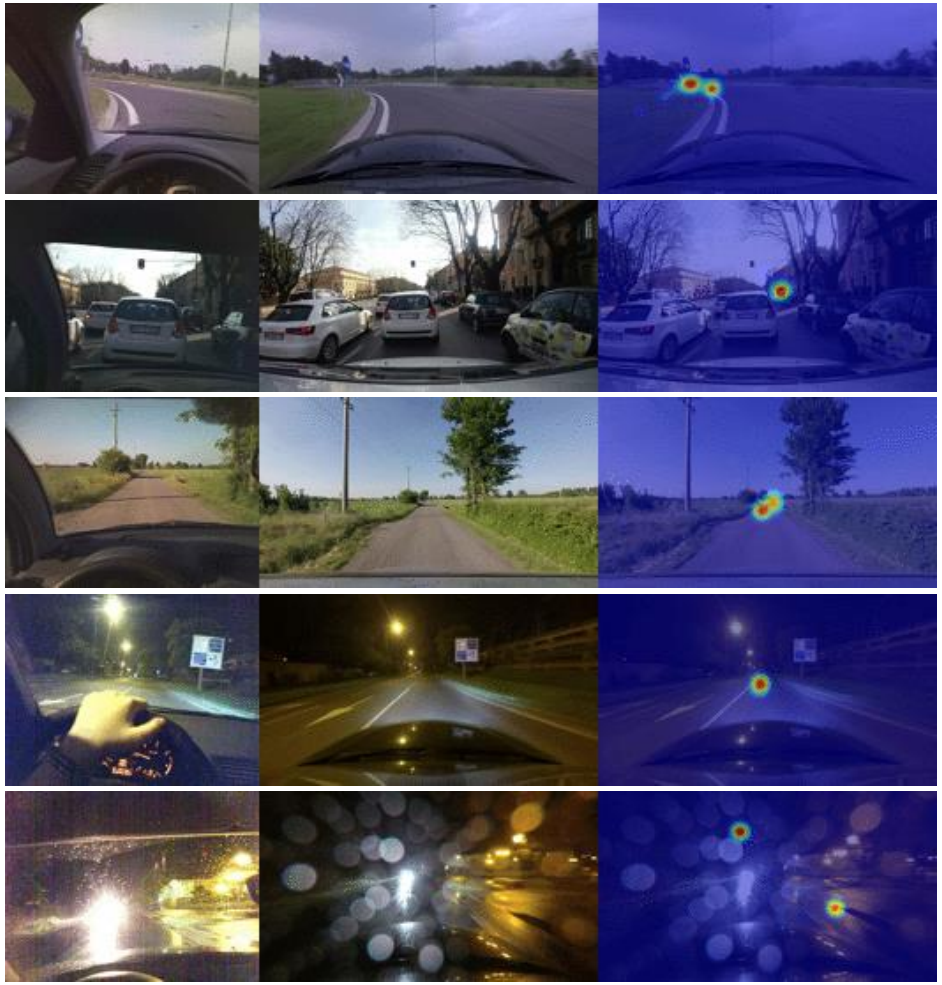


Table 2. Table summarizing the different characteristics of the dataset.

# Videos	# Frames	Drivers	Weather conditions	Lighting	Gaze Info	Metadata	Camera POVs
74	555,000	8	sunny	day	raw fixations	GPS	driver (720p)
			cloudy	evening	gaze map	car speed	car (1080p)
			rainy	night	pupil dilation	car course	

- 8 different drivers
- 3 different landscapes
  {Highway, Countryside, Downtown}
- 3 different weather's conditions:
  {Sunny, Cloudy, Rainy}
- 3 different light's conditions:
  {Morning, Evening, Night}

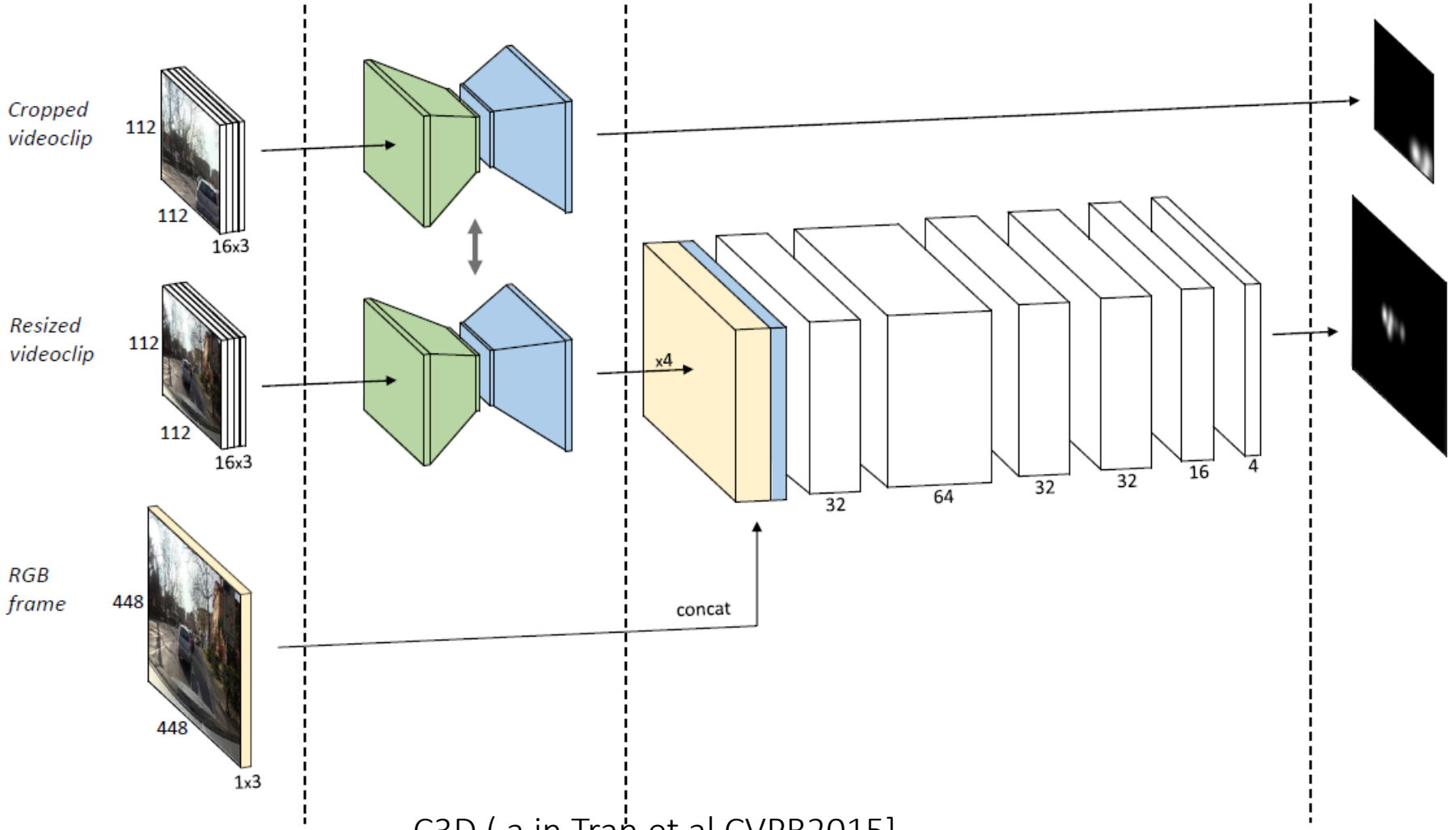
74 videos of 5 minutes each!

Stefano Alletto , Andrea Palazzi , Francesco Solera , Simone Calderara and Rita Cucchiara **DR(eye)VE** a Dataset for Attention-Based Tasks with Applications to Autonomous and Assisted Driving CVPRW2016

SIFT-BASED REGISTRATION FRAME BY FRAME

Collected with SMI ETG 2w, Frontal camera 720p/30fps + Eye pupils cameras at 60fps
GARMIN VirvX , 1080 p/25fps +GPS.

The Dr(eye)ve DL architecture for task-driven saliency

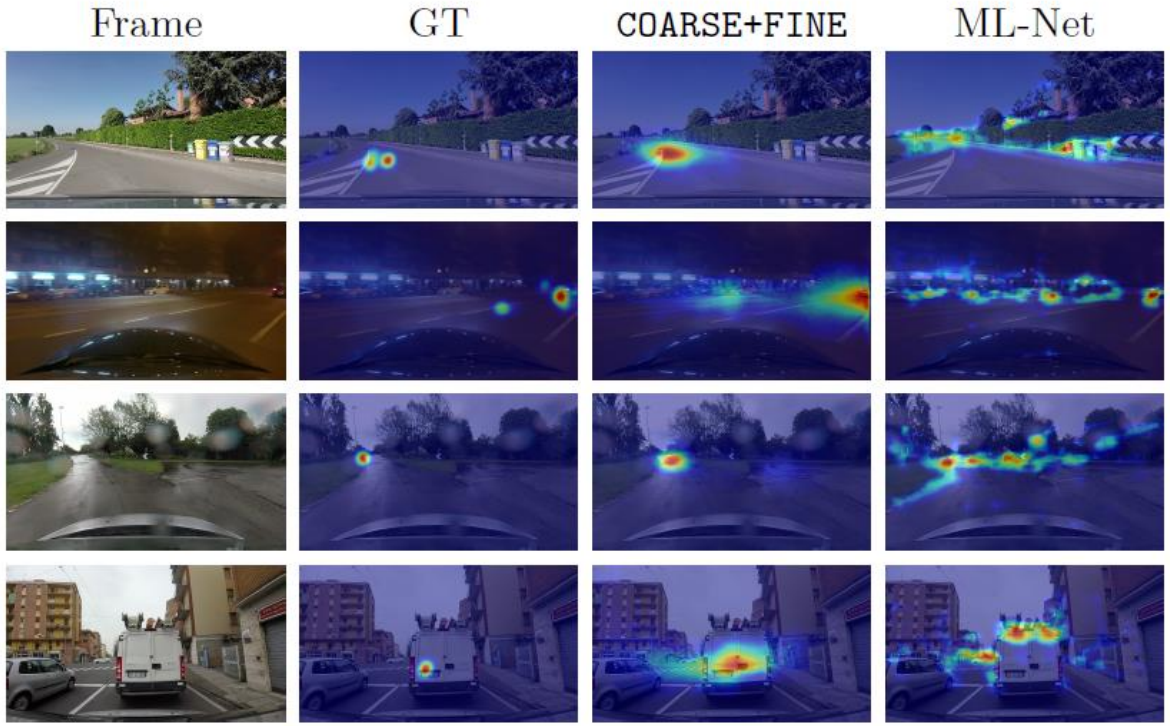


C3D (a in Tran et al CVPR2015]

Evaluation

- ✓ Loss with the Kullback-Leiber divergence between pixels P and Q groundthruth on I pixels
- ✓ Pearson's correlation coefficient (CC)

$$D_{KL}(P,Q) = \sum_i Q_i \log\left(\epsilon + \frac{Q_i}{\epsilon + P_i}\right)$$



	Test seq		Att. seq	
	$CC \uparrow$	$D_{KL} \downarrow$	$CC \uparrow$	$D_{KL} \downarrow$
Baseline (gaussian)	0.33	2.50	0.22	2.70
Baseline (mean train GT)	0.48	1.65	0.17	2.85
Wang *et al.* [41]	0.08	3.77	–	–
Wang *et al.* [40]	0.03	4.24	–	–
ML-Net	0.41	2.05	0.29	2.49
COARSE	0.44	1.73	0.19	2.74
COARSE+FINE	0.55	1.42	0.30	2.24

[40]Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. CVPR 2015
[41]Wang, W., Shen, J., Shao, L.: Consistent video saliency using local gradient ow optimization and global refinement. IEEE PAMI 2015

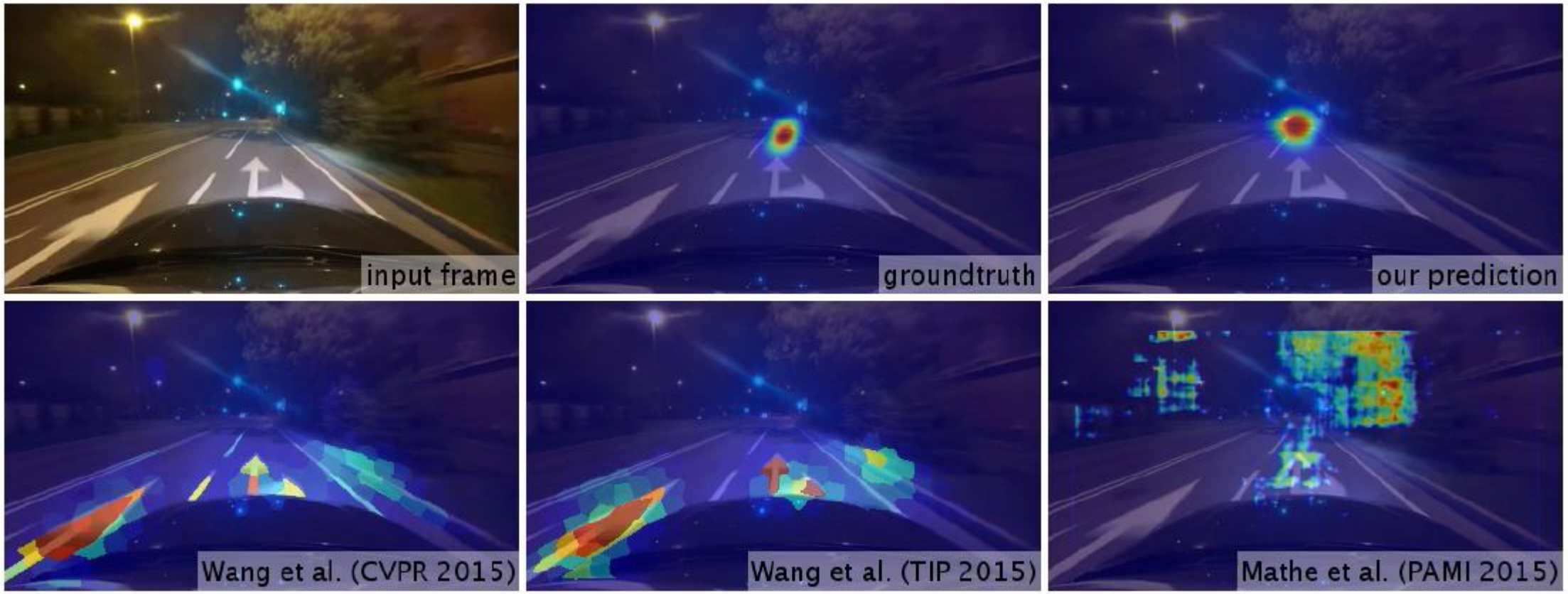


Figure 17. Visual comparison of our approach against state-of-the-art supervised (Mathe *et al.* [36]) and unsupervised methods (Wang *et al.* [63], Wang *et al.* [62]).

S. Mathe and C. Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. IEEE Transactions on PAMI 2015

In summary:

✓ Saliency wit Deep Learning

✓ Bottom-up Data driven + knowledge based (high level features)

- ML-NET: VGG convolutive structure for images with multy-layer and bias features
- SAM:Improvement with LSTM , emulating scan-path refinement
- ML-NET and SAM useful for video too if motion is not too strong

✓ Top-down Task-driven for driving

- DR(EYE)VE C3D autoencoder with cropping areas to observe every-where
- Refinements convolutive to fit the learned models driving

IMAGE AND VIDEO CAPTIONING

(thanks to **Lorenzo Baraldi** and Costantino Grana
and for Saliency and captioning also Marcella Cornia, and Giuseppe Serra)

MOTIVATION : VIDEO CONTENT ANALYSIS

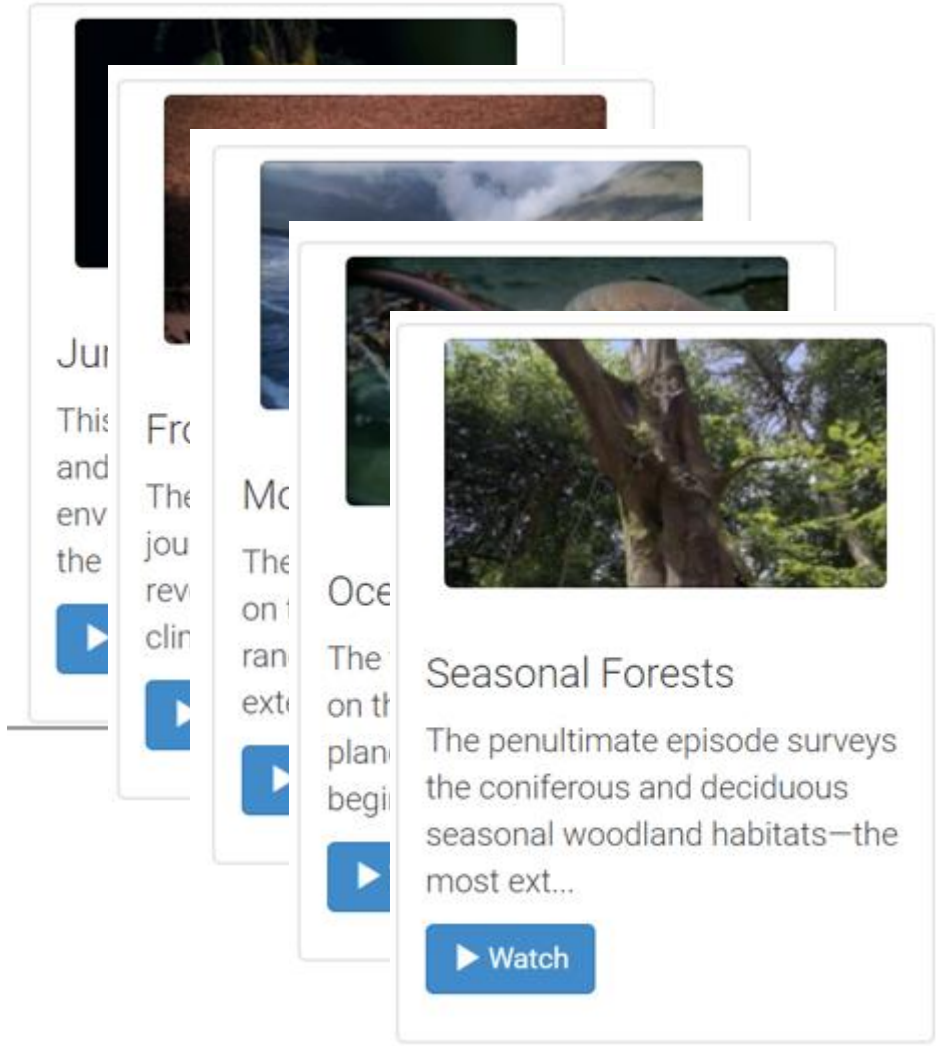
BIG MULTIMEDIA DATA FOR NEW MULTIMEDIA SERVICES

- Education
- Edutainment, Digital Humanities
- Fun in social networks
- Sport analysis
- Product Video in Industry 4.0
- Security
- ……

AUTOMATIC ANNOTATION:

FROM VISUAL DATA TO TEXTUAL INFORMATION

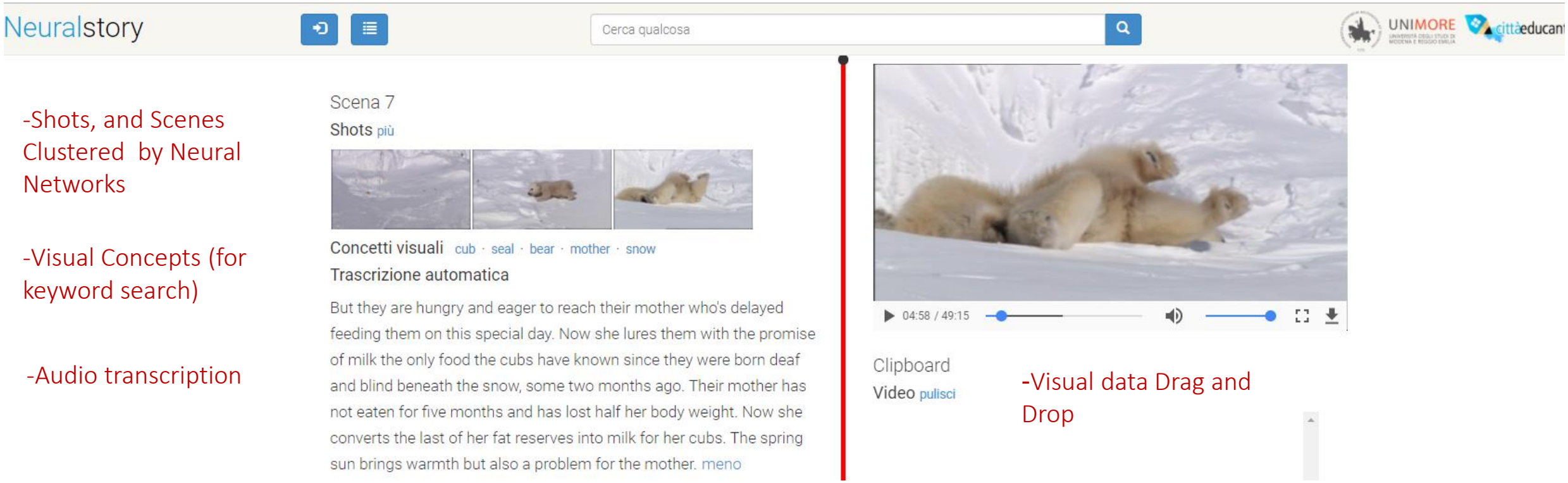
- Indexing, searching and retrieval
- Summarization
- Contextual Copy detection
- Query answering
- Natural human-computer interaction
- ..



Neuralstory

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

New web services of video annotation, summarization and re-use in education



-Shots, and Scenes Clustered by Neural Networks

-Visual Concepts (for keyword search)

-Audio transcription

-Visual data Drag and Drop

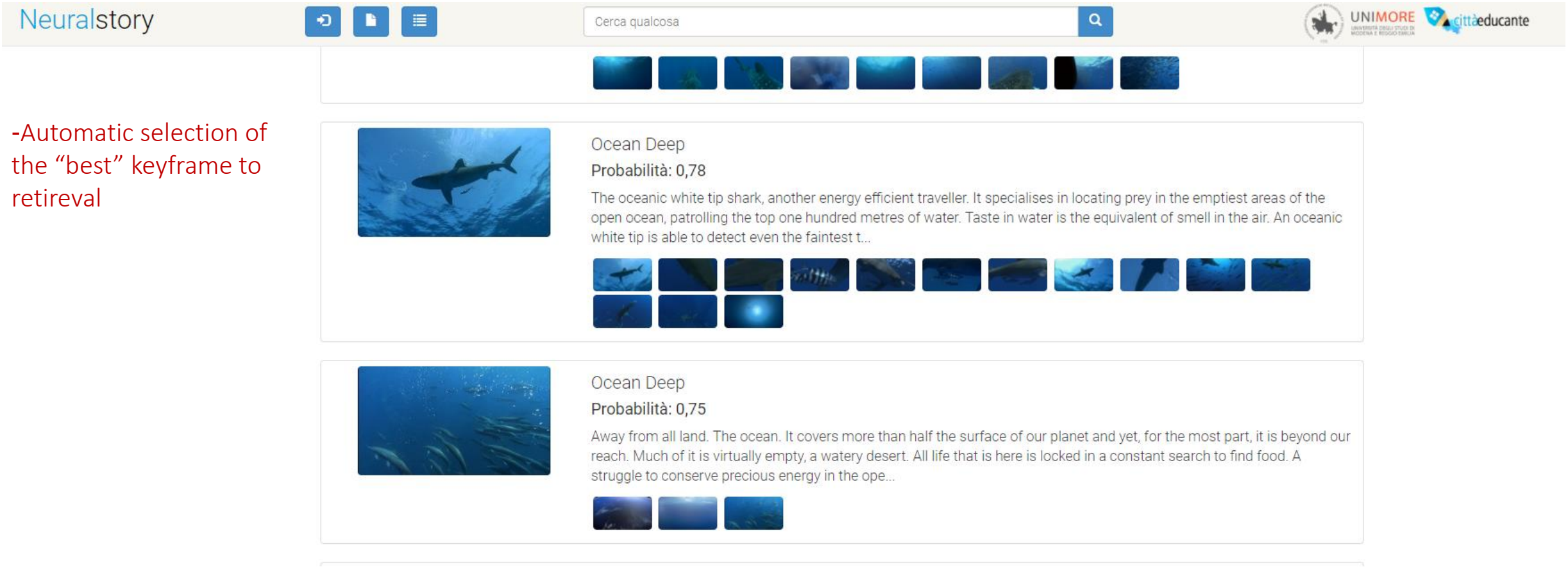
cittàeducante

Italian MIUR Project
In National Technology Cluster initiative

- ✓ L. Baraldi, C. Grana, and R. Cucchiara. *Recognizing and presenting the storytelling video structure with deep multimodal networks* IEEE Transactions on Multimedia, 2016
- ✓ L. Baraldi, C. Grana, and R. Cucchiara. *Deep siamese network for scene detection in broadcast videos* Proc of 23rd ACM MM, 2015

Neuralstory : Searching by keyword

-Automatic selection of the “best” keyframe to retireval



L. Baraldi, C. Grana, and R. Cucchiara *Scene-driven retrieval in edited videos using aesthetic and semantic deep features*; Proc of 2016 ACM ICMR 2016

Neuralstory for Edutainment Re-use

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Gli squali balena

Gruppo di studio: Arianna, Jennifer e Adrienne

LO SQUALO BALENA IN ROTTA DI COLLISIONE

Gli squali balena si lasciano trasportare lentamente dall'acqua,introducendo plancton,piccoli pesci e calamari. Vivono in acque tropicali calde, dove c'è sufficiente disponibilità di cibo.

Lo squalo balena è certamente il pesce più grande del mondo. Malgrado l'apparenza che incute timore, è però relativamente innocuo per l'uomo e si nutre quasi esclusivamente di plancton, filtranto il cibo con le branchie. La bocca si trova alla fine del muso, una posizione inusuale per uno squalo, e anche se le mascelle misurano oltre un metro, sono armate di denti molto piccoli.

Gli squali balena,che possono alimentarsi anche in posizione verticale usando la bocca come un gigantesco secchio dove convogliare il cibo, normalmente, per nutrirsi, nuotano lentamente in superficie, provocando allarme quando vengono avvistati.

8/91 secondi

Concetti visuali

fish · shark · whale · surface · bait

Trascritto

The biggest of all fish 30 tonnes in weight, 12 metres long a whale shark. Its huge bulk is sustained by near microscopic creatures of the sea, the plankton. Whale sharks cruise on regular, habitual routes between the best feeding grounds. In February, that takes them to the surface waters far from the coast of Venezuela. Others are already here. Bait fish have come for the same reason to feed on the plankton. The whale shark has timed its arrival exactly right. Oddly the tiny fish swarm around it. They're using it as a shield. Other

Finding more compact and effective description

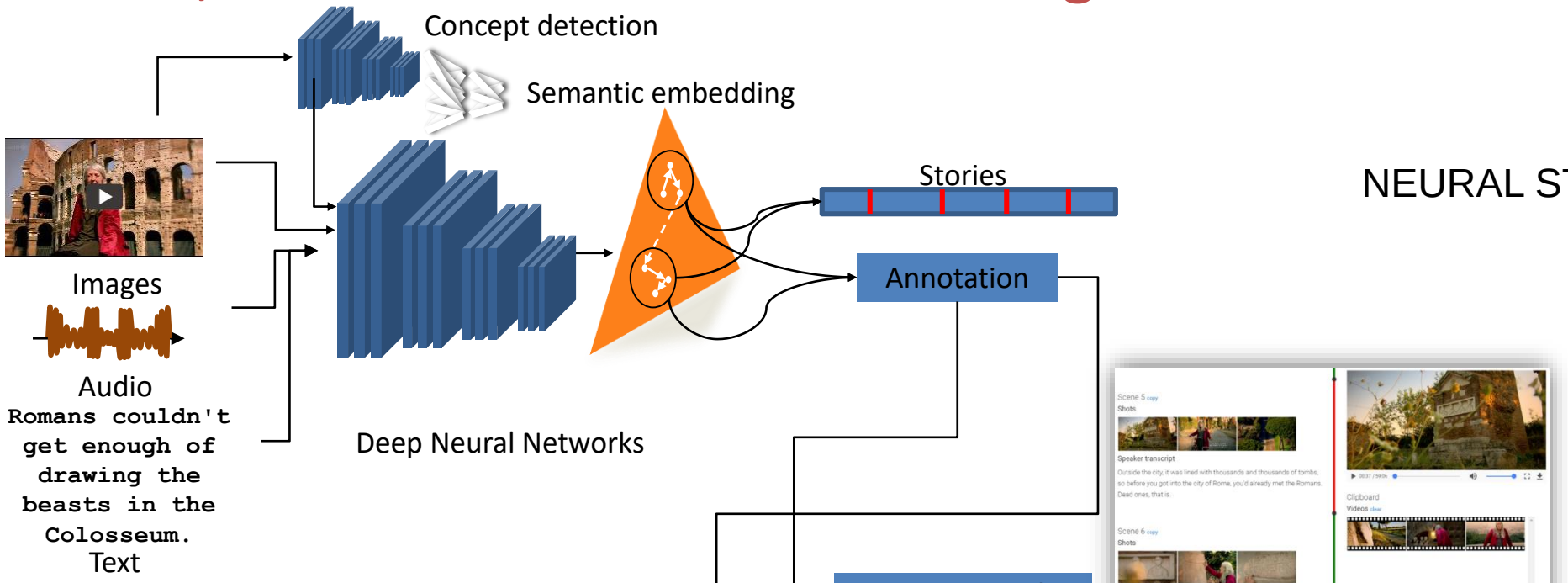
→Image and Video Captioning: *the capability of automatically describing the visual content in natural language.*

- More than keywords or tags
- More compact than long language transcription
- Extracted directly from visual data
- Learned by human-based text description

From VISUAL DATA

To TEXT SEQUENCE ..a white shark swims in the ocean water..

From Visual data, audio and text to understanding

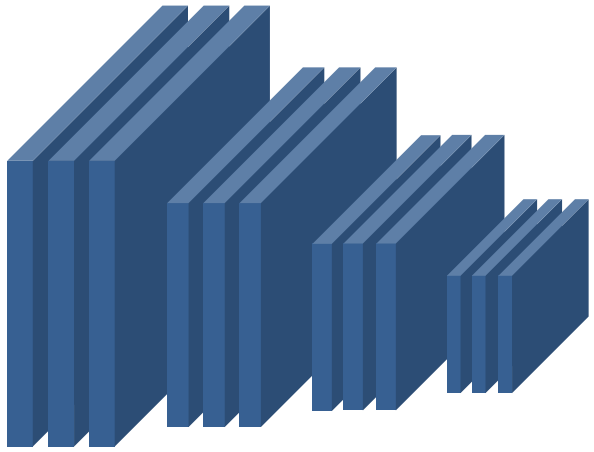


NEURAL STORY

- ✓ www.neuralstory.unimore.it

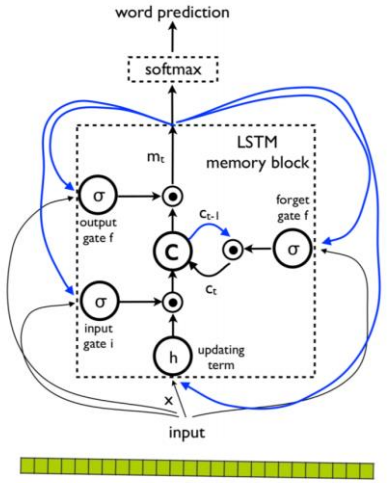
From temporal Segmentation to Captioning

..a white shark swims in the ocean water..



CONV-NET

+



Recurrent NET (LSTM)

Previous works in Conv + Recurrent

- ✓ [1] Karpathy, Andrej, and Li Fei-Fei. "Deep visual-semantic alignments for generating image descriptions." *CVPR 2015*
- ✓ [2] Vinyals, O., Toshev, A., Bengio, S. and Erhan, D., Show and tell: A neural image caption generator. *CVPR 2015*

GT1: the woman is riding a horse
GT2: the horse and rider trotted down the field
GT3: a person is riding a horse
GT4: a woman is riding a horse
GT5: a girl is riding a horse

Pr: a woman is riding a horse

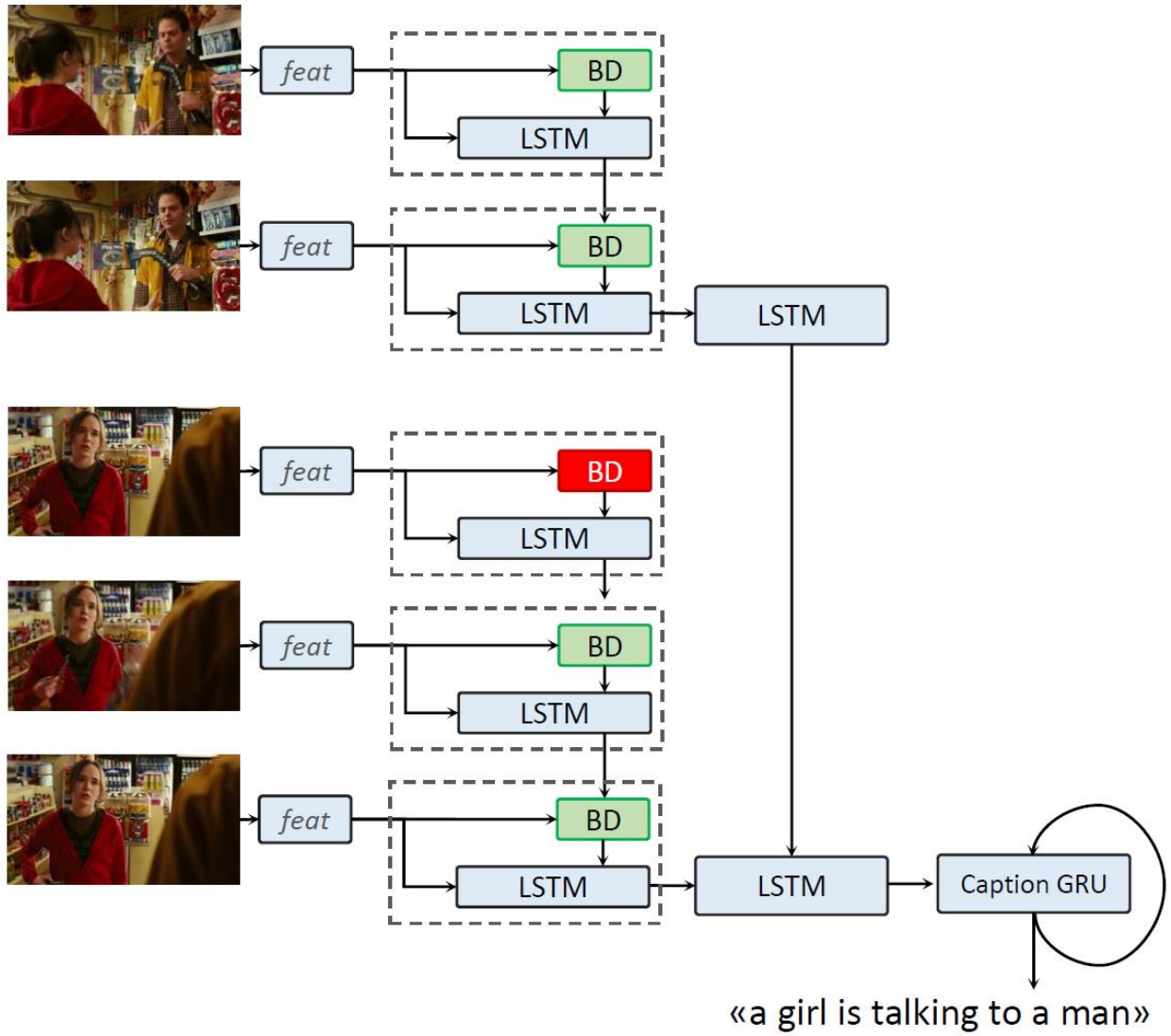
GT1: a woman is slicing potatoes
GT2: a woman is cutting a potato into small pieces
GT3: a person is slicing a potato into pieces
GT4: a woman is slicing potatoes
GT5: a woman is cutting a potato

Pr: a person is cutting a potato

Improving sentence generation in long video [CVPR17]

- ✓ A new architecture for video captioning

  With the capacity of forgetting



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

Rationale

- ✓ LSTM are spectacular but suitable for short sequence 30-80 frames [Ng et al CVPR 2015]
- ✓ Video (e.g. movies) are structured : shots and scenes

- ✓ Shots: frame sequence taken by same camera (perceptually similar)
- ✓ Scenes: shot sequences of the same story (semantically similar)

RATIONALE: Video captioning must be aware of the structure, not to mix words of consecutive shots, thanks to forget/reset mechanism

Keep in mind the *consecutio temporum*

Many attempts in NLP and hierarchical NN

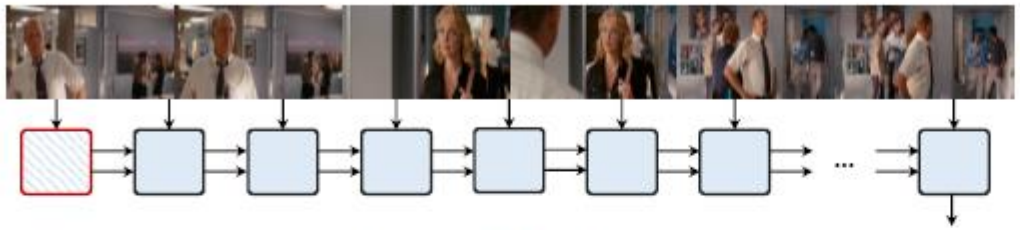
- ✓ Is a **Sequence-to-sequence translation** style: Video encoding with an LSTM and world generation with a second LSTM
- ✓ Improvement of sentence generation by hierarchy with GRU (Gated Recurrent Unit) for sentence concatenated into paragraphs: sentence of consecutive sentences
- ✓ Improvement in video encoding by feeding LSTMs with sliding windowed video chunks , passed to a second LSTM to take into account video structure

- ✓ → we propose a hierarchical LSTM to understand the video structure and add a forgetting mechanism to create consecutive sentences in a caption

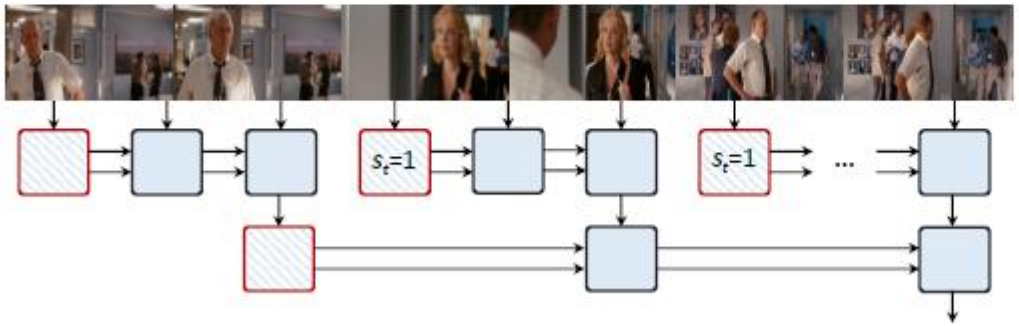
1. S. Venugopalan, M. Rohrbach, J. Donahue, R. Mooney, T. Darrell, and K. Saenko. Sequence to sequence-video to text. CVPR 2015
2. H. Yu, J. Wang, Z. Huang, Y. Yang, and W. Xu. Video paragraph captioning using hierarchical recurrent neural networks.CVPR 2016.
3. P. Pan, Z. Xu, Y. Yang, F. Wu, and Y. Zhuang. Hierarchical recurrent neural encoder for video representation with application to captioning CVPR 2016

A suitable modification of LSTM (again) with boundary detection

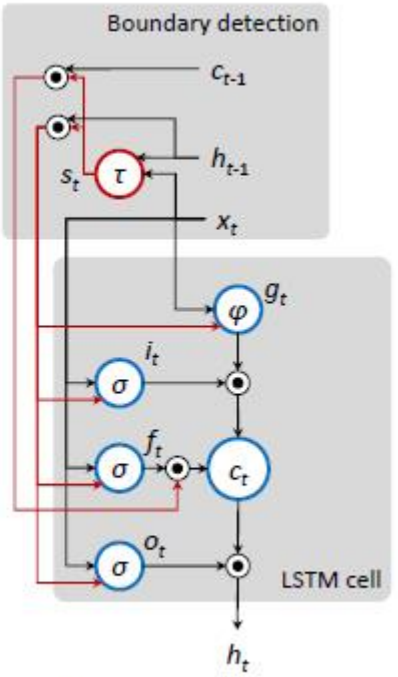
- ✓ In a single end-to-end pipeline
- ✓ if a boundary is detected, thus the memory is reset



(a) Traditional LSTM network



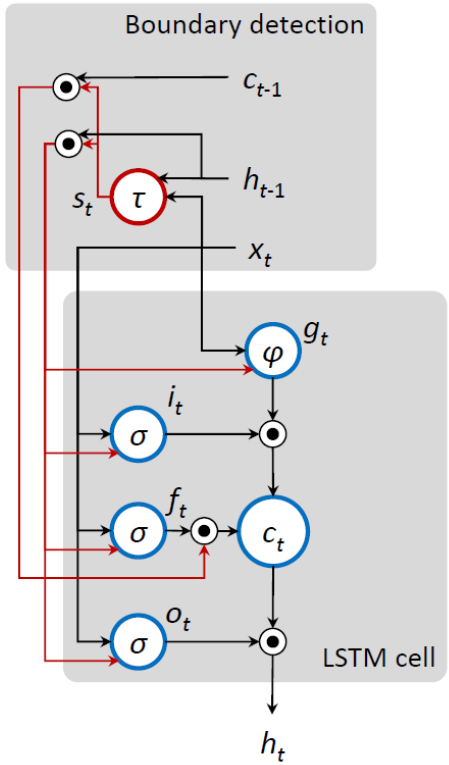
(b) Time Boundary-aware LSTM network



(c) Time Boundary-aware cell

LSTM with boundary detection

- ✓ Ct: maintains the history of the observed inputs; its update is given by modulating i_t, f_t and o_t as functions of x_t and h_{t-1}, followed by a sigmoid activation
- ✓ I_t gate controls how to add the input x_t to the cell
- ✓ F_t gate controls what the cells forget from previous acquired
- ✓ O_t gate controls if the memory should be passed as output
- ✓ A time boundary detection unit decides if the hidden state should be transferred to the cell or a reset command is given
- ✓ Given a v_t learnable row vector, Wsi, Wsh and b are learnable weights and biases
- ✓ Given s_t and h_t , c_t can be reset or continue
- ✓ Special **hints for training in order to have t(x) differentiable**: st is as a stochastic neuron [Raiko et al ICLR 2015]



$$s_t \in \{0, 1\}$$

$$\mathbf{h}_{t-1} \leftarrow \mathbf{h}_{t-1} \cdot (1 - s_t)$$
$$\mathbf{c}_{t-1} \leftarrow \mathbf{c}_{t-1} \cdot (1 - s_t).$$

$$s_t = \tau(\mathbf{v}_s^T \cdot (W_{si}\mathbf{x}_t + W_{sh}\mathbf{h}_{t-1} + \mathbf{b}_s))$$

$$\tau(x) = \begin{cases} 1, & \text{if } \sigma(x) > 0.5 \\ 0, & \text{otherwise} \end{cases}$$

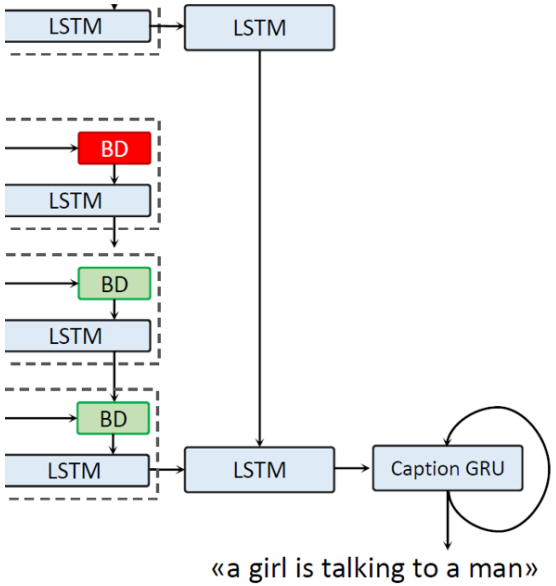
Sentence generation

- ✓ Given a v video vector, creating a sentence of words $(\mathbf{y}_0, \mathbf{y}_1, \ldots, \mathbf{y}_T)$ encoded as a 1-of-N vector
- ✓ the decoder, to find the word yt , is conditioned over the previous with a log-likelihood being w the parameters

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

- ✓ Probability is modeled as a softmax layer at the output of decoder
- ✓ to reduce dimensionality a linear encoding is used to transform 1-of-N vectors to input space and viceversa with a Wp matrix, pt is the output of decoder generated by the GRU layer

$$\Pr(\mathbf{y}_t | \mathbf{y}_{t-1}, \mathbf{y}_{t-2}, \ldots, \mathbf{y}_0, \mathbf{v}) \propto \exp(\mathbf{y}_t^T W_p \mathbf{p}_t)$$



Details on L.Baraldi et al CVPR 2017

Video caption datasets

- ✓ **M-VAD** : 84,6 hours of 92 **Hollywood** movies, 46K video clips with single description based on (descriptive Video services DVS for impairs
- ✓ **MPII** with manually corrected sentences 94HD movies with 68K sentences
- ✓ **MSVD** Microsoft 2K Youtube clips with 85K sentence from MecTurk

Performance metrics

- ✓ BLEU using four-grams as a precision (used for machine translation in NLP).
- ✓ ROUGE N computes an F-measure with a recall bias using a longest common subsequence technique. Compares N grams with co-occurrence statistics.
- ✓ METEOR, scores captions by aligning them to one or more ground truths. Alignments with synononimous, stems etc more semantically correct. Uses the F-measure with a recall bias
- ✓ CIDEr use cosine-similarities between word: **Consensus-based Image Description Evaluation** (from CVPR2016)

Details of the BA Boundary Aware encoder

- ✓ Features with ResNet 50 on Imagenet + C3D Network for Sport-1M for motion
- ✓ Ground truth versions are tokenized, and maintained only if appear at least 5 times
- ✓ Collected a Vocaboulary of
  - M-VAD 6090 words
  - MPII-MD 7198 words
  - MSVD 4125 words
  - + <BOS> and >EOS>

Our Vocabulary and results are still FAR from Natural Language!

[The average educated native English speaker knows around **35,000 words**, which is roughly 20% of 171,476 words in current usage.
-Google collected more than 1.000.000 English words
-A C2 Advanced English students know among 4000-10000 words]

Model	METEOR
SA-GoogleNet+3D-CNN [49]	4.1
HRNE [22]	5.8
S2VT-RGB(VGG) [43]	6.7
HRNE with attention [22]	6.8
Venugopalan *et al.* [42]	6.8
LSTM encoder (C3D+ResNet)	6.7
Double-layer LSTM encoder (C3D+ResNet)	6.7
Boundary encoder on shots	7.1
Boundary-aware encoder (C3D+ResNet)	**7.3**

Table 1. Experiment results on the M-VAD dataset.

Model	CIDEr	B@4	R_L	M
SMT (best variant) [30]	8.1	0.5	13.2	5.6
SA-GoogleNet+3D-CNN [49]	-	-	-	5.7
Venugopalan *et al.* [42]	-	-	-	6.8
Rohrbach *et al.* [29]	10.0	**0.8**	16.0	**7.0**
LSTM encoder (C3D+ResNet)	10.5	0.7	16.1	6.4
Double-layer LSTM encoder (C3D+ResNet)	10.6	0.6	16.5	6.7
Boundary encoder on shots	10.3	0.7	16.3	6.6
Boundary-aware encoder (C3D+ResNet)	**10.8**	**0.8**	**16.7**	**7.0**

Table 2. Experiment results on the MPII-MD dataset.

Model	B@4	M	C
SA-GoogleNet+3D-CNN [49]	41.9	29.6	-
LSTM-YT [44]	33.3	29.1	-
S2VT [43]	-	29.8	-
LSTM-E [23]	45.3	31.0	-
HRNE [22]	46.7	33.9	-
Boundary-aware encoder	42.5	32.4	63.5

Table 3. Experiment results on the MSVD dataset.

GT: A woman dips a shrimp in batter.
HRNE [22]: A woman is cooking.
BA encoder (ours): A woman is adding ingredients to a bowl of food.

GT: A boy is playing a guitar.
HRNE [22]: A man is playing a guitar.
BA encoder (ours): A boy is playing guitar.

GT: A dog is swimming in a pool.
HRNE [22]: A dog is swimming.
BA encoder (ours): A dog is swimming in the pool.

Figure 4. Example results on the MSVD dataset.

From MVAD and MPII-MD datset

GT: She gets out.
LSTM encoder: Someone stops.
BA encoder (ours): Someone gets out of the car.

GT: Shakes his head.
LSTM encoder: Someone gives her gaze.
BA encoder (ours): Someone looks at someone who shakes his head.

GT: He slows down in front of one house with a garage and box tree on the front.
LSTM encoder: Someone gets out of the car and walks out of the house.
BA encoder (ours): Someone drives up to the house.

PUTTING ALL TOGETHER

Video captioning..
and image captioning
A still un-completely solved problem

GT: he stands and offers her the small bouquet

Pr: someone looks up at someone

- ✓ Attention in persons and not in the salient objects!

A very new proposal

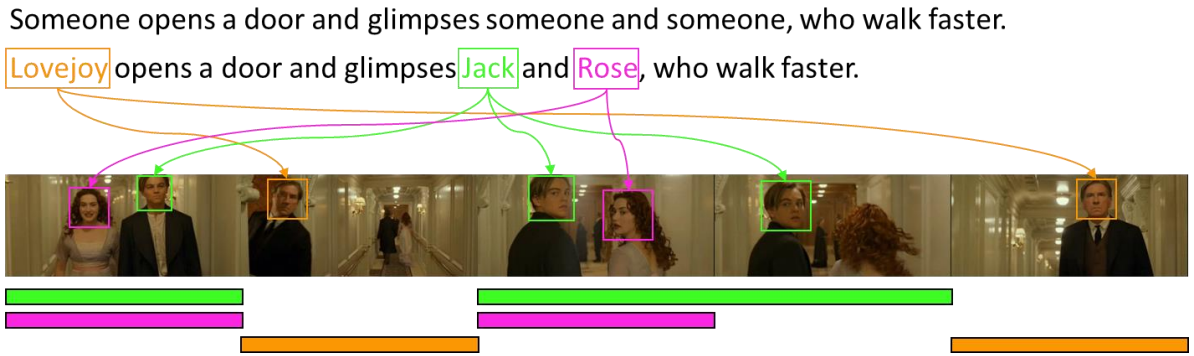
Goal:

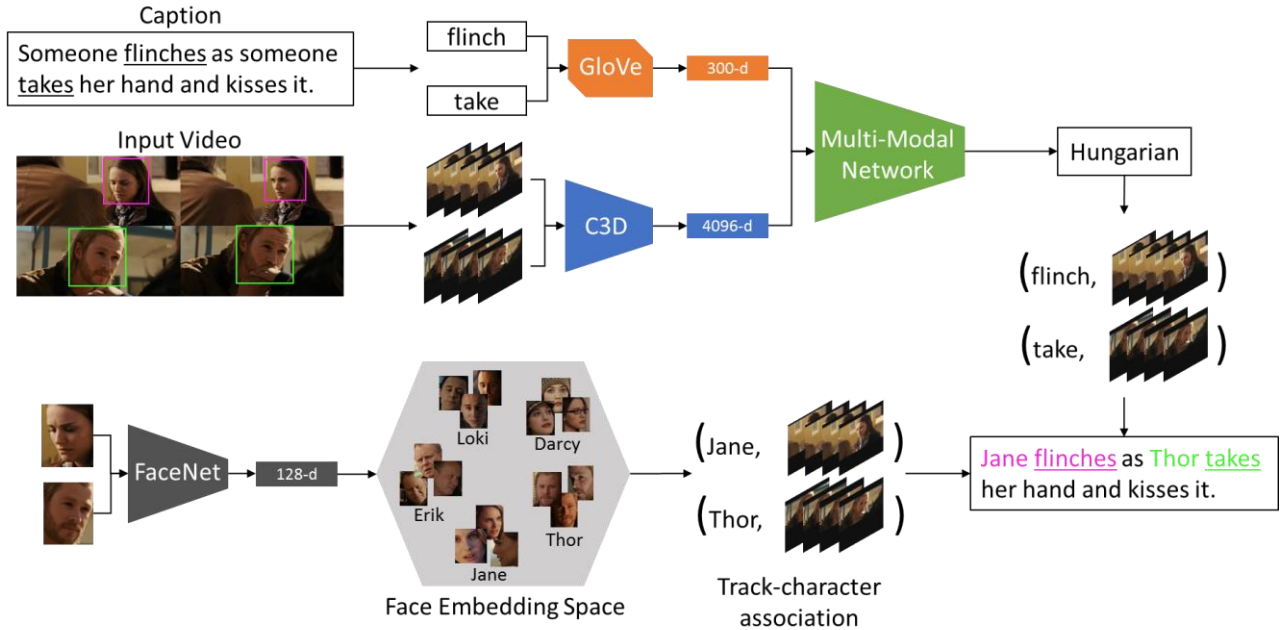
She [Mary] smiles while he [Fred] offer her a small bouquet in a garden

1. Image Captioning by LSTMs
+
2. "Machine Attention" by LSTM [Xu..Bengio ICML 2015]
+
3. Saliency & Context computation by SAM
+
4 Naming with face detectors

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, Rita Cucchiara Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention ArXiv 2017 submitted

Rationale 4. Captioning with Naming





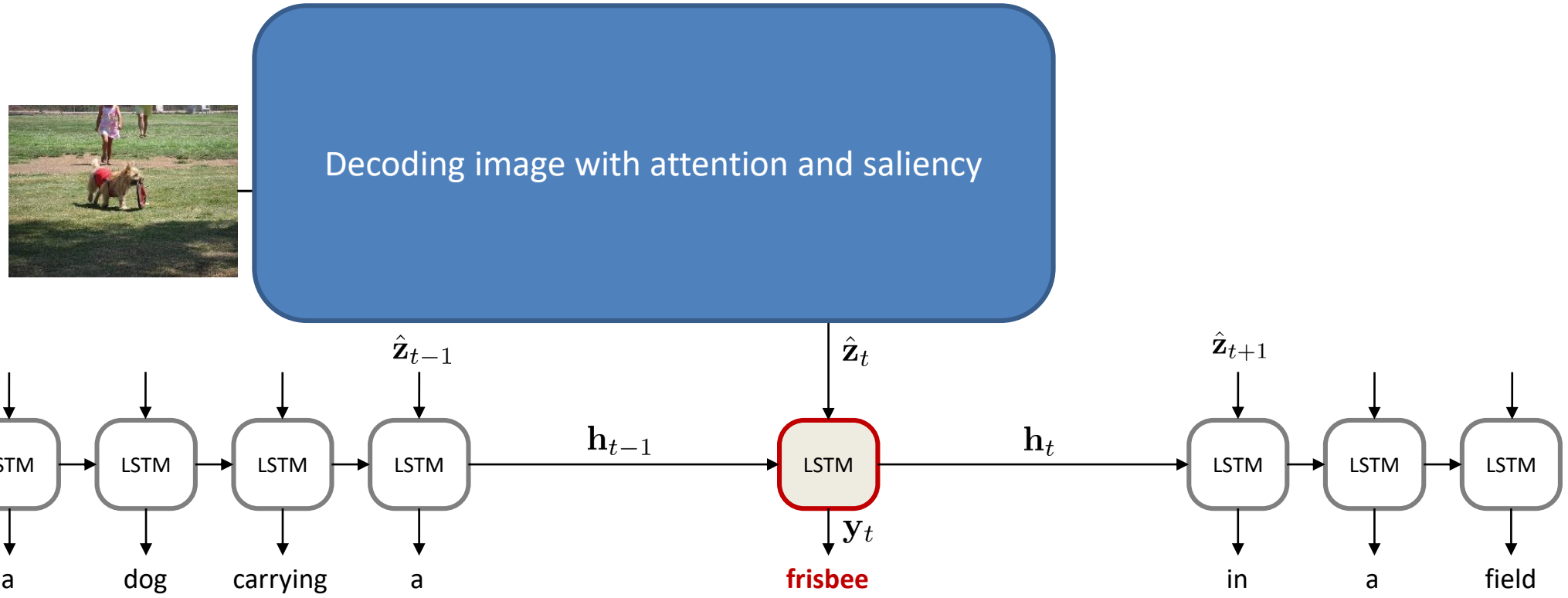
S. Pini, M. Cornia, L. Baraldi, R. Cucchiara.
Towards Video Captioning with Naming: a Novel Dataset and a Multi-Modal Approach.
Proc. of ICIAP 2017

Rationale 1: Image Caption Generation with LSTM

FROM THE END OF THE PIPELINE

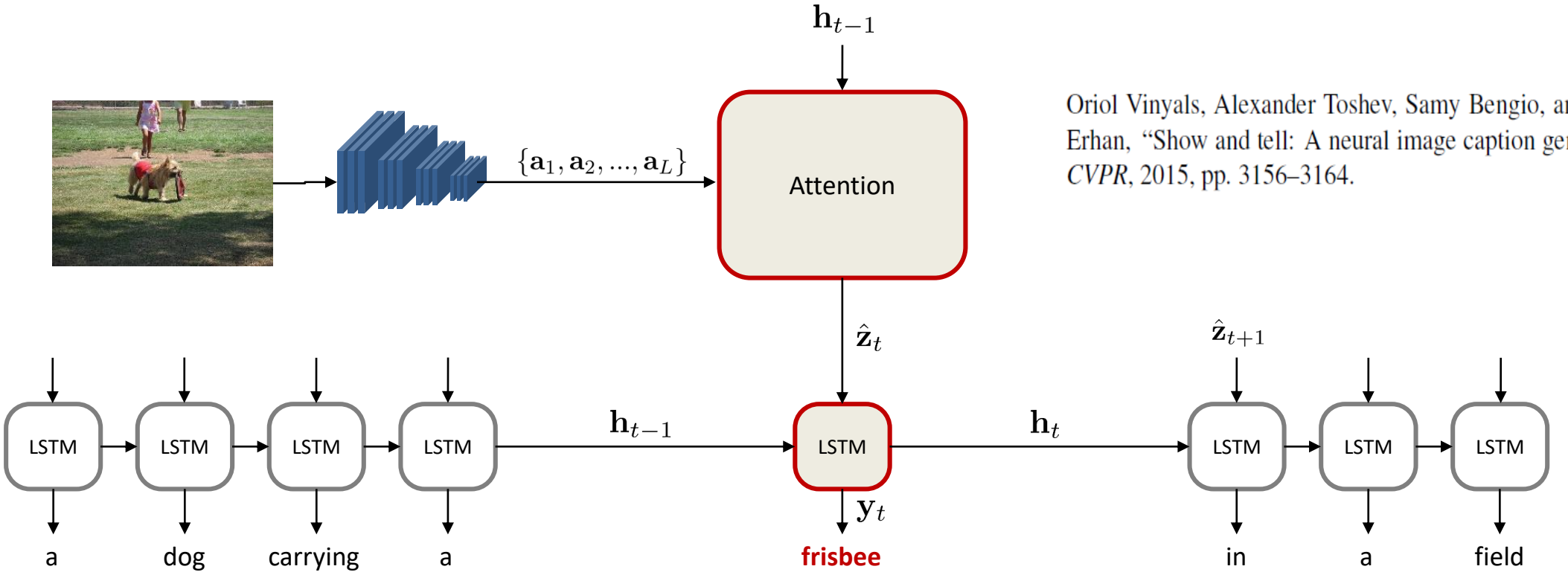
we adopt the state of the art of image and video captioning

✓ LSTM + with a starting and ending token and a one-of-N word vector



Rationale 2: Exploiting " Machine Attention"

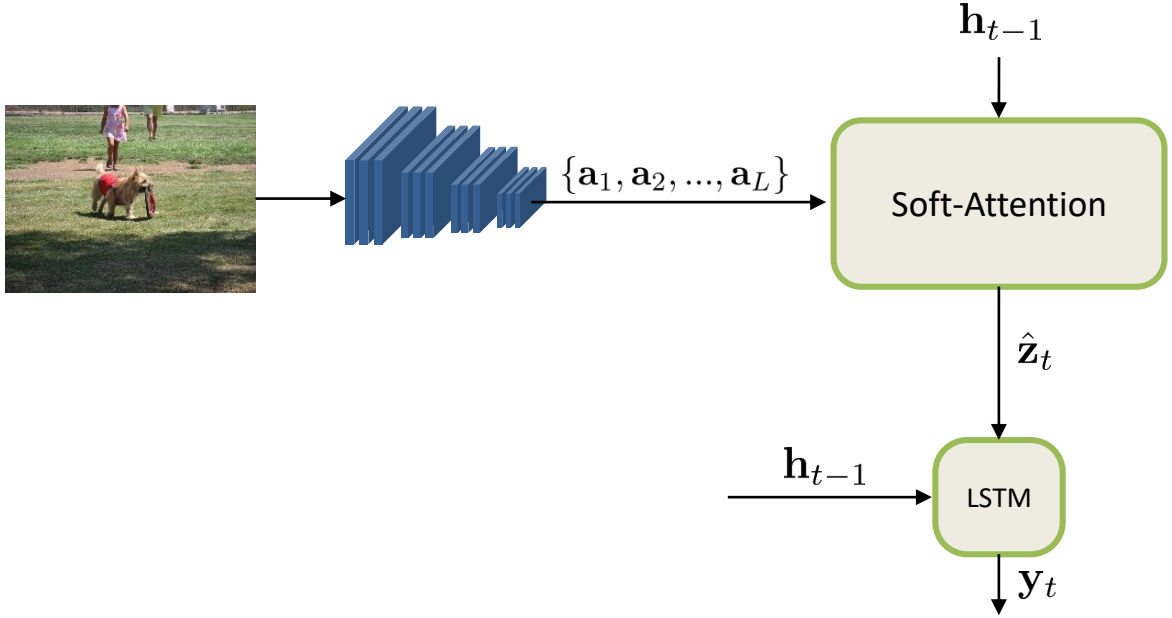
- Machine attention mechanism: a way of obtaining time-varying inputs for recurrent architectures.
- At each timestep the attention mechanism selects a region of the image, based on the previous LSTM state, and feeds it to the LSTM.
- The generation of a word is conditioned on that specific region, instead of being driven by the entire image.



Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

Machine Attention with a «Soft-Attention» method

- In the "soft-attention", the input image is encoded as a grid of feature vectors obtained from a CNN $\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_L\}$
- At each timestep, the soft-attention mechanism computes a context feature vector $$\hat{\mathbf{z}}_t = \sum_{i=1}^{L} \alpha_{ti} \mathbf{a}_i$$



where α_{ti} are weights representing the current state of the machine attention.

- These weights are driven by the original image feature vectors and by the previous hidden state of the LSTM

$$e_{ti} = v_e^T \cdot \phi(W_{ae} \cdot \mathbf{a}_i + W_{he} \cdot \mathbf{h}_{t-1}))$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}$$

where ϕ is the hyperbolic tangent $tanh$, W_{ae}, W_{he} are learned matrix weights and v_e^T is a learned row vector.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, Attend and Tell: Neural Image Caption Generation with Visual Attention. In *International Conference on Machine Learning*.

Machine attention and “soft attention”

- ✓ The idea under the table is that **“ we should look around”**….. **“attention is useful”**
- ✓ Thus some features are extracted (e.g. with VGG 16) and used as input vector for LSTM as a feed-forward network
- ✓ Very good results

  [Xu et al 2015]:

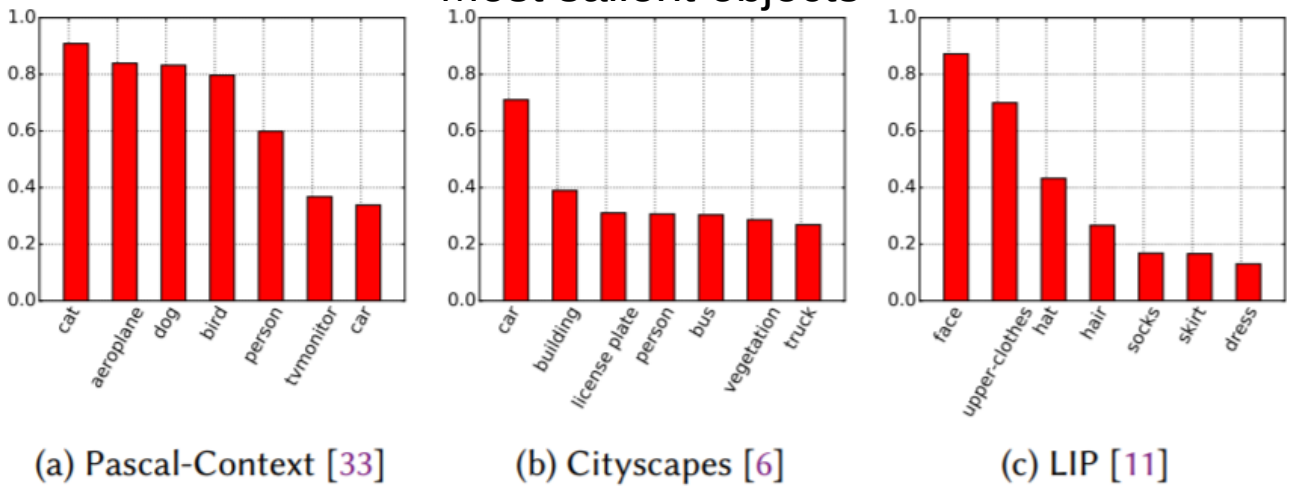
A group of people sitting on a boat in the water.

Rationale 3: Pre-attentive Saliency and Context

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

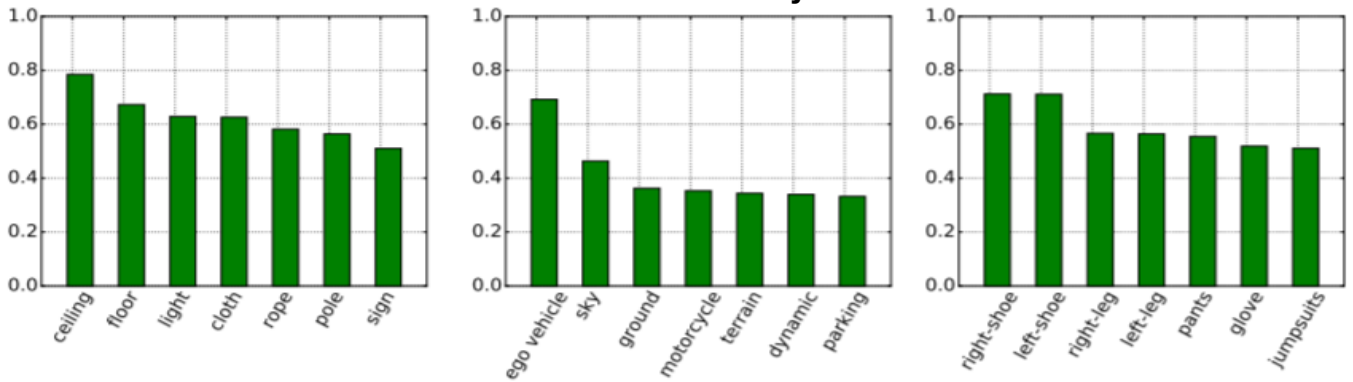
- ✓ What is salient and what is not

Most salient objects



(a) Pascal-Context [33] (b) Cityscapes [6] (c) LIP [11]

Less salient objects



Some Experiments:
WHAT IS HIT BY SALIENCY (SAM)

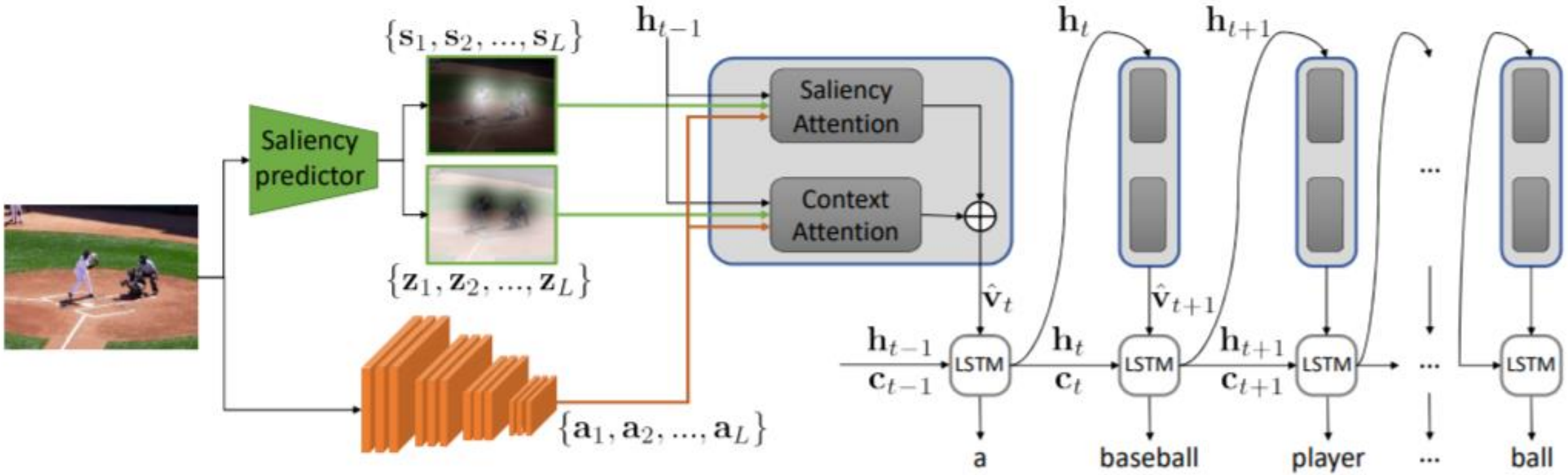
Pascal c.a. 20K images, 400 labels
Cityscapes c.a. 5K images, 30 classes
LIP c.a. 50K images , 19 human body parts

1. Saliency looks at objects and less than 10% at background

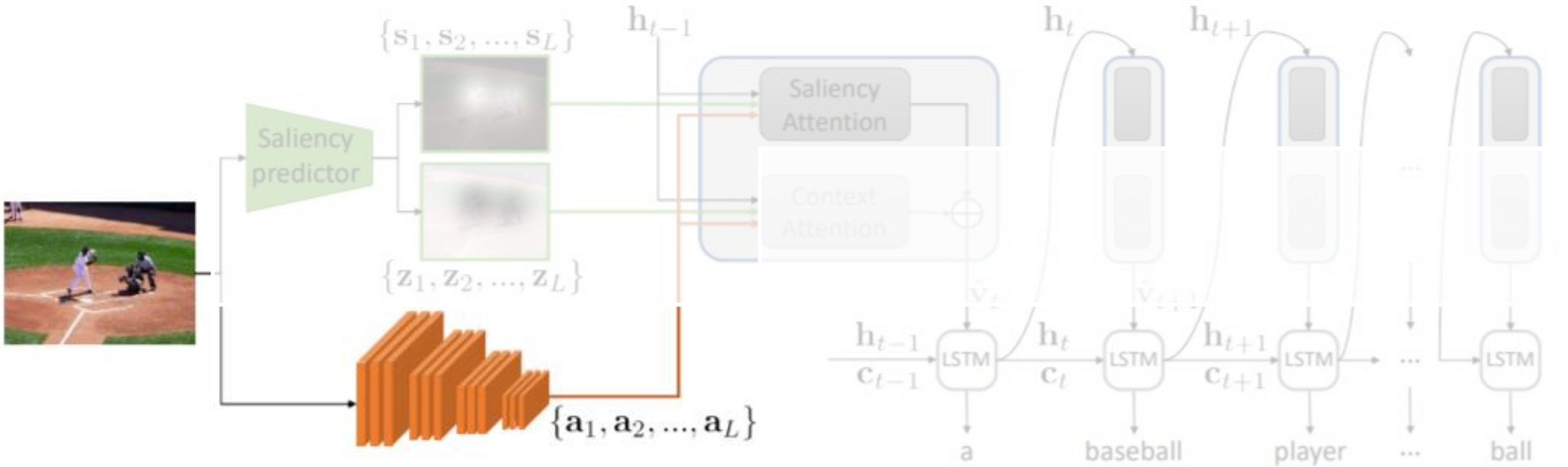
2. Saliency is independent on the object size (thus importance is not related to size)

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

Rationale 3 Final proposal

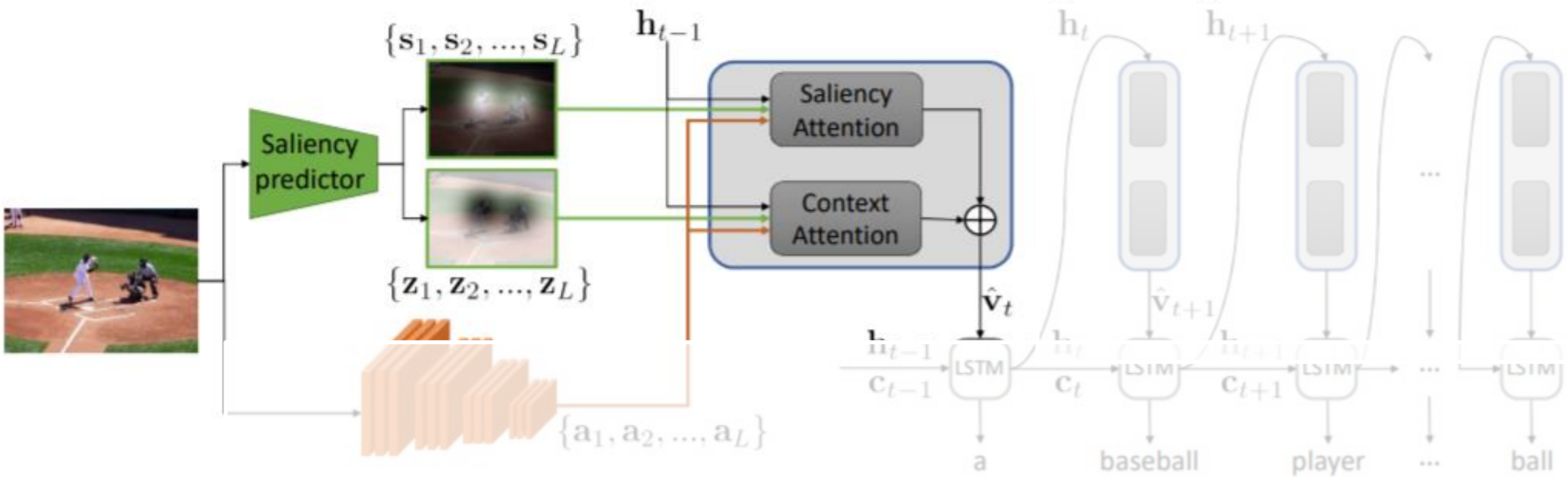


Rationale 3 Final proposal



ResNet -50, trained with Imagenet;
49 layers , output 2048 channel, +1 conv layer refined in the dataset → 512 filters

Rationale 3 Final proposal



SAM (Saliency Attentive Model) defines saliency map and the contextual map
They are input for two LSTM for “Soft attention”, trained with the same weigth

Experimental Results

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

✓ SALICON Dataset

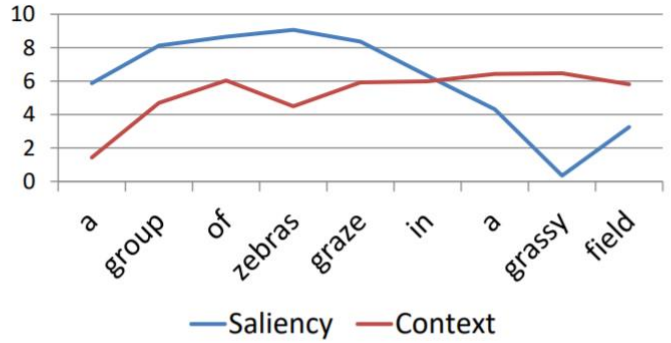
	CNN	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE$_L$	CIDEr
Soft Attention	VGG-16	0.680	0.501	0.358	0.256	0.222	**0.497**	0.691
Saliency-Guided Attention	VGG-16	**0.682**	**0.505**	**0.361**	**0.258**	**0.223**	**0.497**	**0.694**
Saliency-Guided Att. (with GT saliency maps)	VGG-16	*0.684*	*0.503*	*0.360*	*0.257*	*0.224*	*0.501*	*0.696*
Soft Attention	ResNet-50	0.700	0.523	0.379	0.274	**0.235**	0.510	0.771
Saliency-Guided Attention	ResNet-50	**0.709**	**0.534**	**0.388**	**0.280**	0.233	**0.513**	**0.774**
Saliency-Guided Att. (with GT saliency maps)	ResNet-50	*0.702*	*0.527*	*0.383*	*0.277*	*0.236*	*0.513*	*0.779*

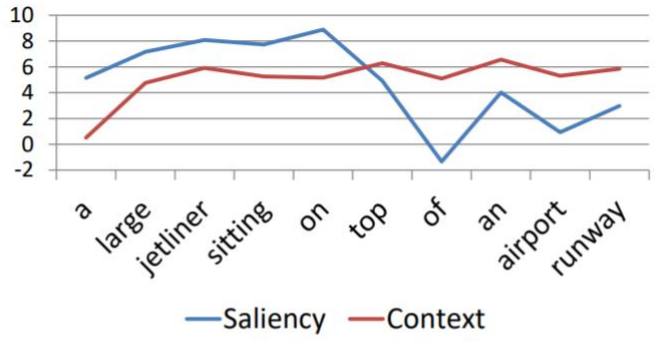
SALICON Dataset (subset of the Microsoft COCO dataset, composed by 20,000 images, largest available dataset for saliency prediction)

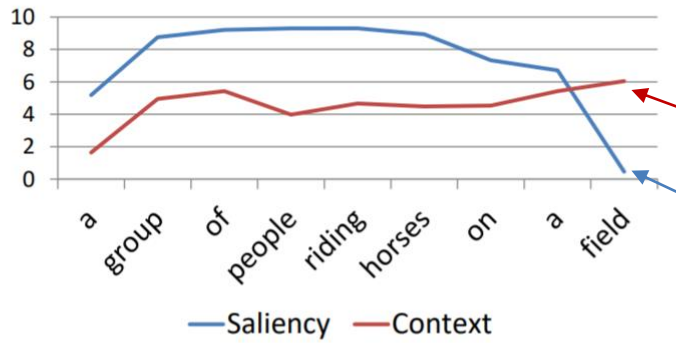
Microsoft COCO Dataset

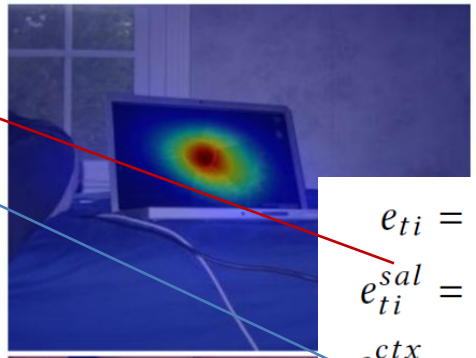
	CNN	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE$_L$	CIDEr
Soft Attention	ResNet-50	0.717	0.546	0.402	0.294	0.253	0.529	0.939
Saliency-Guided Attention	ResNet-50	**0.718**	**0.547**	**0.404**	**0.296**	**0.254**	**0.530**	**0.944**

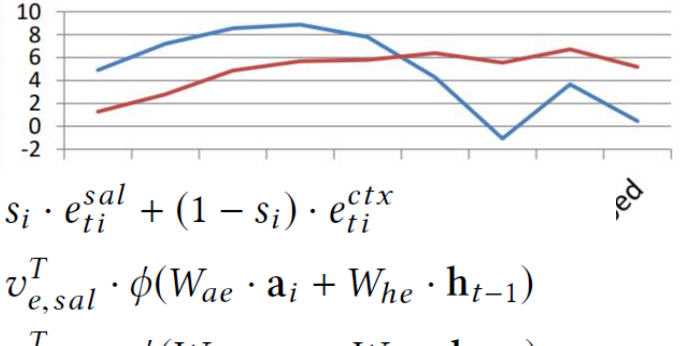
Microsoft COCO Dataset (composed by more than 120,000 images divided in training and validation sets each image is annotated with five sentences)







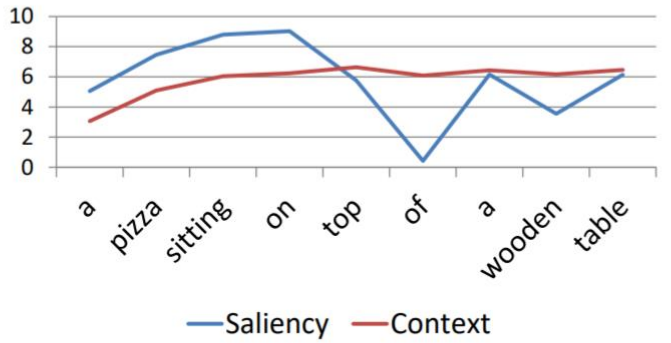


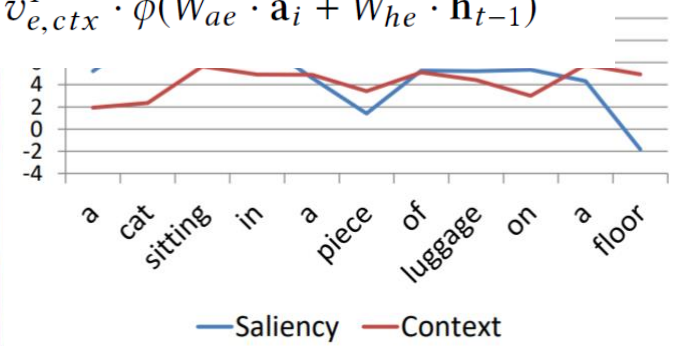


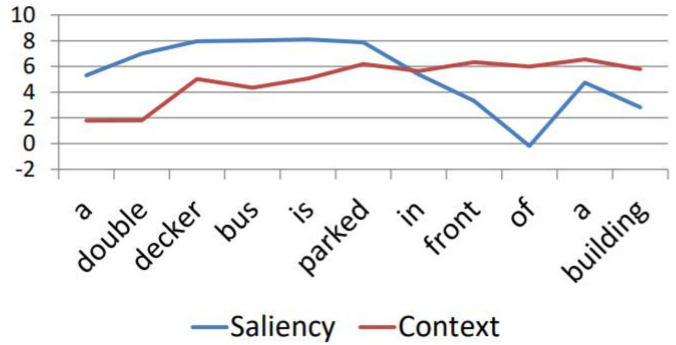
$$e_{ti} = s_i \cdot e_{ti}^{sal} + (1 - s_i) \cdot e_{ti}^{ctx}$$

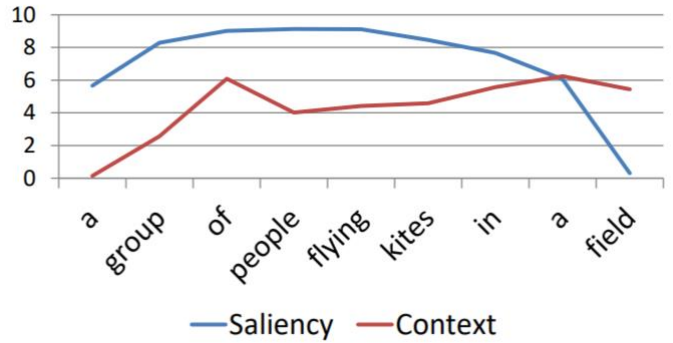
$$e_{ti}^{sal} = v_{e,sal}^T \cdot \phi(W_{ae} \cdot \mathbf{a}_i + W_{he} \cdot \mathbf{h}_{t-1})$$

$$e_{ti}^{ctx} = v_{e,ctx}^T \cdot \phi(W_{ae} \cdot \mathbf{a}_i + W_{he} \cdot \mathbf{h}_{t-1})$$

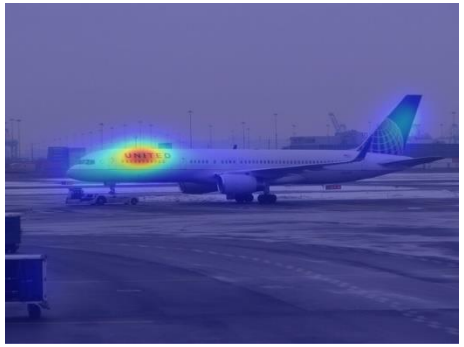








Qualitative Results

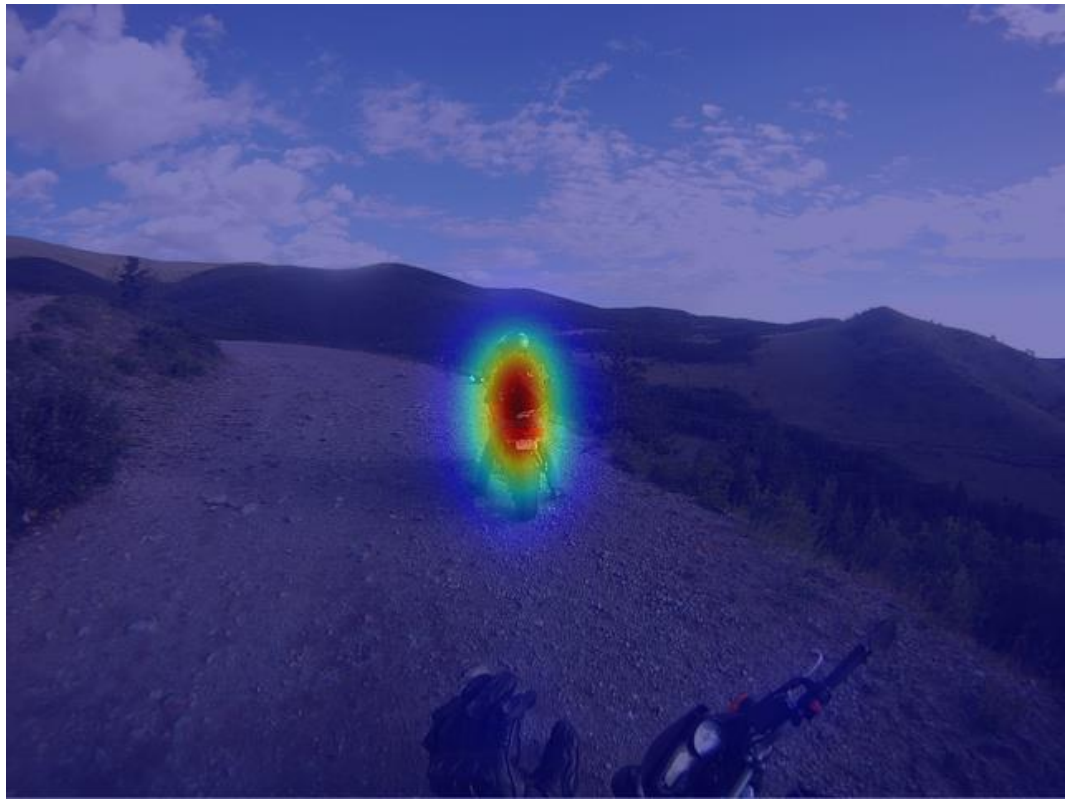



GT: A large passenger jet sitting on top of an airport runway.
With saliency&context: A large jetliner sitting on top of an airport runway.
Without: A large air plane on a runway.


GT: Family of five people in a green canoe on a lake.
With saliency&context : A group of people sitting on a boat in a lake.
Without : A group of people sitting on top of a boat.


GT: Two people in Swarthmore College sweatshirts are playing frisbee.
With saliency&context : A man and a woman are playing frisbee on a field.
Without : A man standing next to a man holding a frisbee.

Qualitative Results



With saliency&context: A person riding a motorcycle on a road.

Without : A man on a bike with a bike in the background.

Qualitative Results

With saliency&context: A person taking a picture of himself in a bathroom.

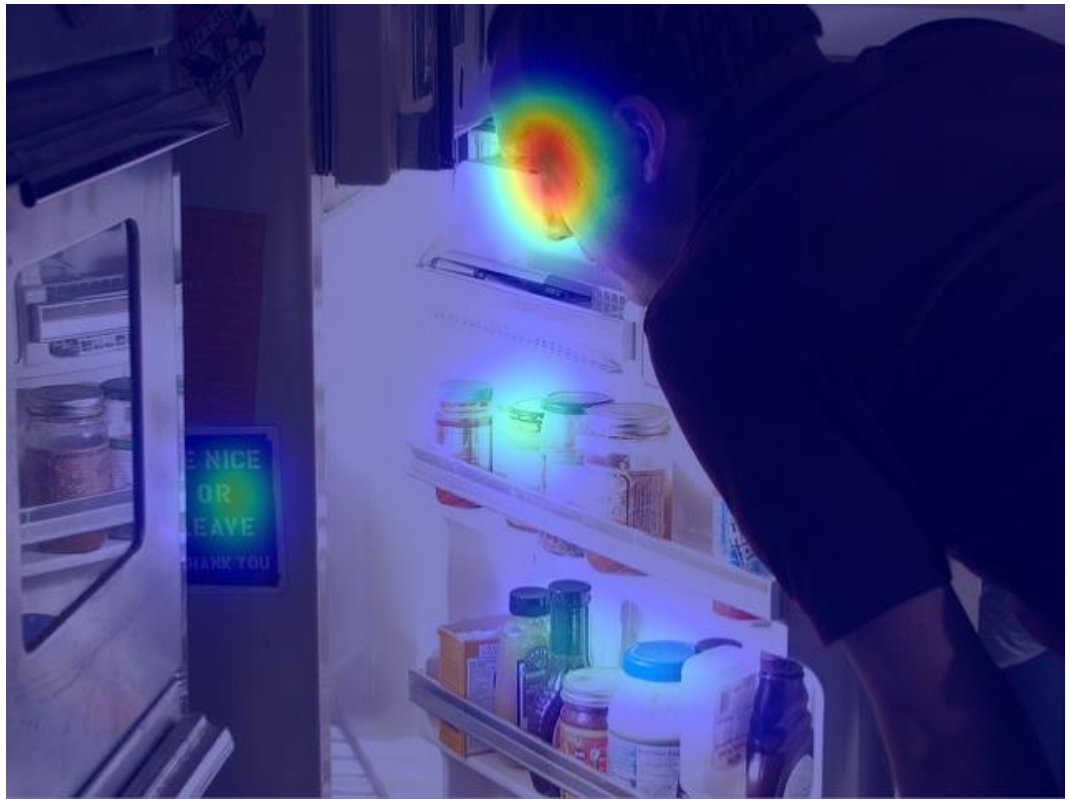
Without : A bathroom with a sink and a sink.

Qualitative Results

With saliency&context: A double decker bus driving down a street.

Without : A bus is parked on the side of the road.

Qualitative Results



With saliency&context: A man is looking inside of a refrigerator.

Without : A man is making a refrigerator in a kitchen.

In conclusion:

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

✓ Existing video captioning algorithm are not endowed with naming capabilities, neither with saliency:

"As he goes SOMEONE sips a cup of coffee. SOMEONE arrives."

"As Scott goes, Elizabeth sips a cup of coffee. Lucy arrives"

"A woman in a red dress is talking inside the Colosseum"

Next step: Video Captioning with Saliency and Naming

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

Thank you!

rita.cucchiara@unimore.it
http://imagelab.ing.unimore.it

Acknowledgements

CINECA SCAI SuperComputing Applications and Innovation
SuperComputing Applications and Innovation
Facebook AI Research
Imagelab