COMPUTER VISION, AI & ROBOTICS FOR VISUAL INTELLIGENCE

IROS 2018 Madrid, Keynote

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

AImage Lab

Rita Cucchiara

AImageLab, Dipartimento di Ingegneria «Enzo Ferrari»

Università di Modena e Reggio Emilia, ITALIA

Artificial Intelligence and Intelligent Systems
cini National Lab

COMPUTER VISION

is a commodity
(i.e. a working software)

as a sense for ROBOTS

ROBOT

is a commodity
(i.e. a programmable, moving object)
as a moving sensor for VISION

Vision and AI are not only the ***4 Rs***
(Recognition, Reconstruction, Registration, Reorganization)

New computer vision results will provide computers, autonomous systems, robots, cars and objects with **visual intelligence,**
towards an embodied intelligence

VISUAL INTELLIGENCE

Sharpen Your Perception, Change Your Life

AMY E. HERMAN

the ability to see what's there that others don't,

to see what's not there that should be,

to see the positives and the negatives,

the opportunity, the invention, the upside, the warning signs,

the quickest way, the way out, the win

www.visualintelligence.com

the ability **to visualize the world accurately, understand quickly the saliency and the important aspect**, modify their surroundings based upon their perceptions, and recreate the aspects of their visual experiences.

People with high visual-spatial intelligence are good at remembering images, faces, and fine details. They are able to visualize objects from different angles **and predict situation**.....

Vision

+

Imagination

VISUAL INTELLIGENCE

Can I take the spritz glass with overturning the plate?

VISUAL INTELLIGENCE

Where do you put your attention?

What do you predict
while driving?

VISUAL INTELLIGENCE

What are these person doing?

How can I run out of the room?

Is that person going to move the chair?

FROM VISION TO VISUAL INTELLIGENCE

AImage Lab

1. SEE THE SALIENT (AND TELL ME BETTER WHAT YOU SEE)

   Thanks to Marcella Cornia, Lorenzo Baraldi, Andrea Palazzi [CVPR17, TMM17, TIP18, TPAMI18..]

2. SEE THE INVISIBLE (SEGMENT BY MOTION)

   Thanks to Guido Borghi and Roberto Vezzani [CVPR17, BMVC18, TPAMI19..]

3. SEE IN THE DARK (DOMAIN TRANSFER FROM RGB TO DEPTH)

   Thanks to Stefano Alletto, Davide Abati and Simone Calderara [TITS19]

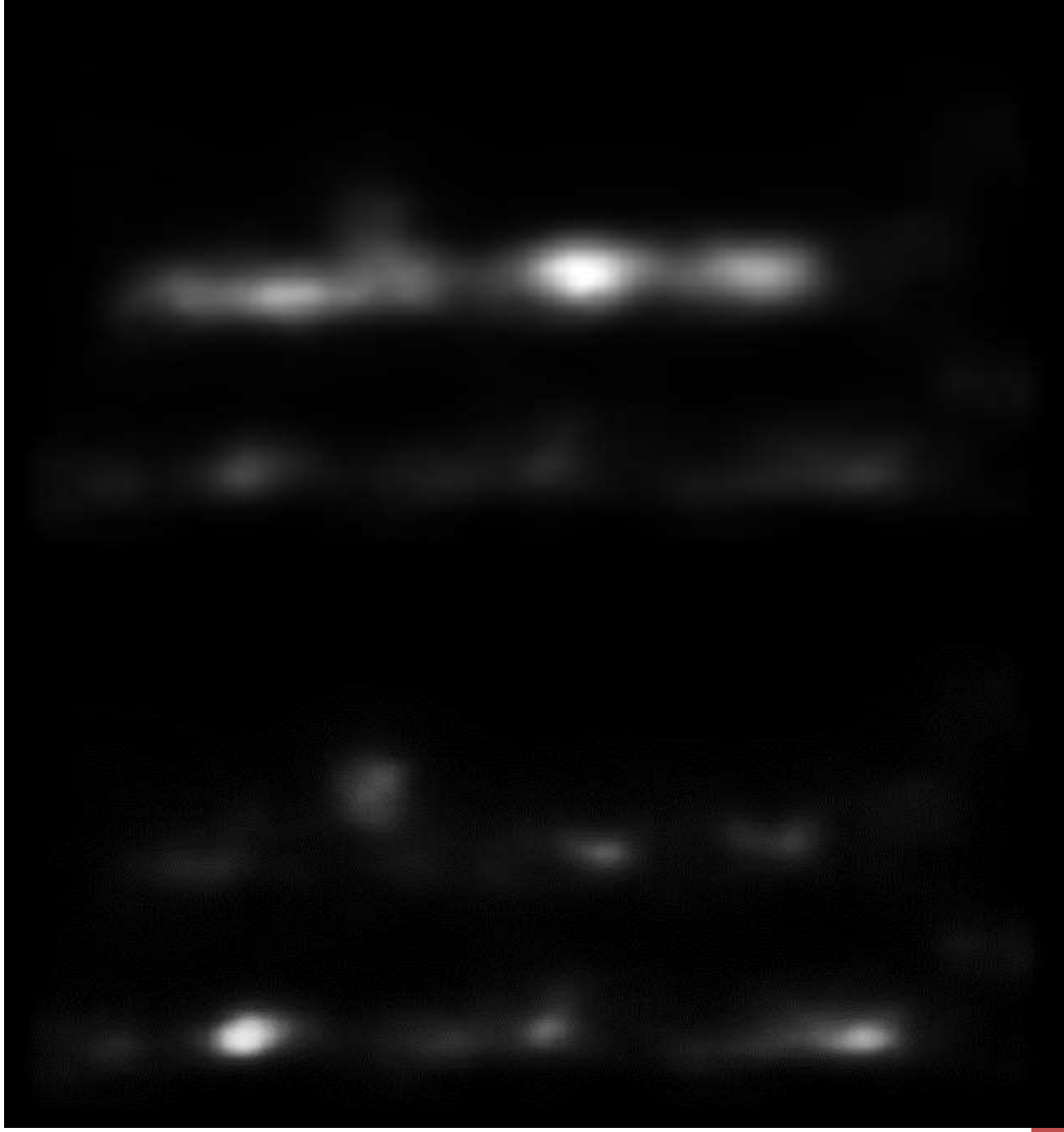
4. SEE THE HIDDEN (BY ALLUCINATING OCCLUSIONS)

   Thanks to Matteo Fabbri, Fabio Lanzi, Simone Calderara [ECCV2018]

CultMEDIA Facebook AI Research Panasonic RedVision Lab Ferrari

FROM VISION TO VISUAL INTELLIGENCE

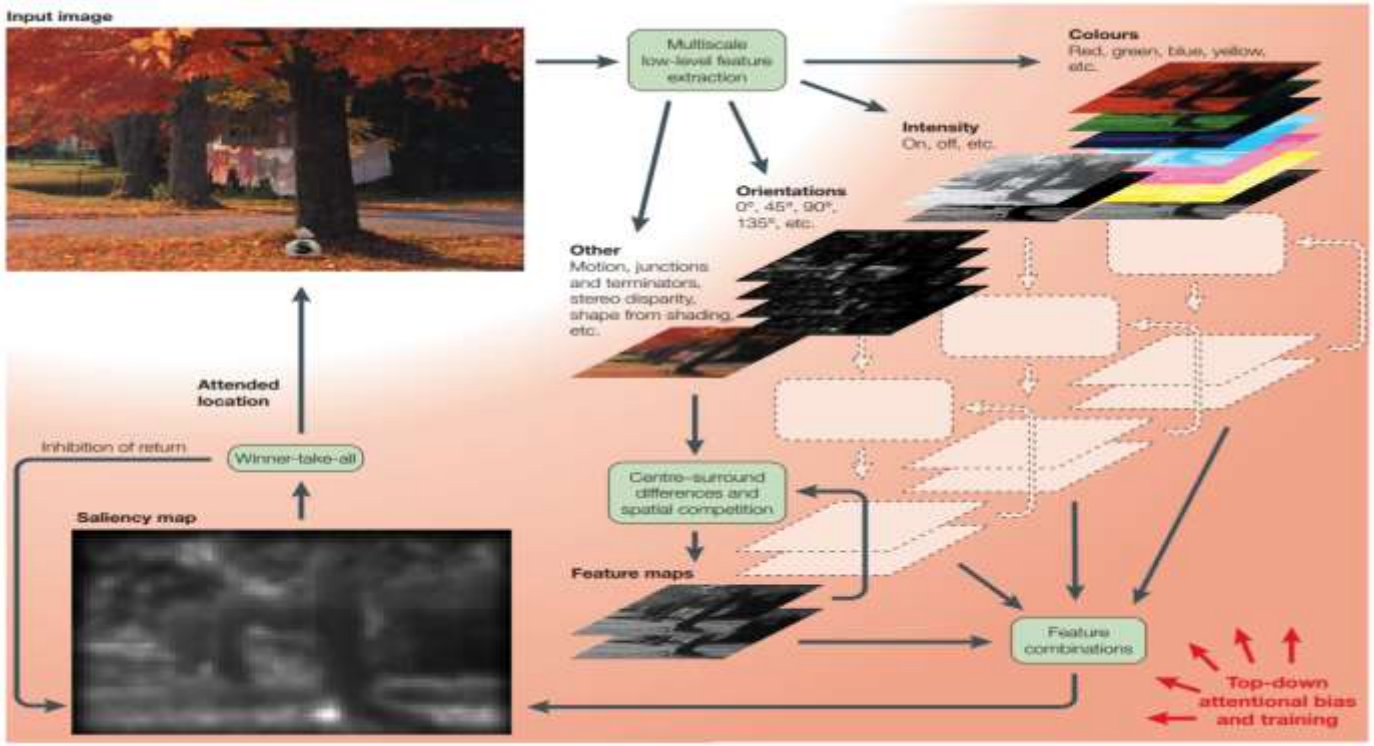
1. SEE THE SALIENT (AND TELL ME BETTER WHAT YOU SEE)



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

In Neuroscience SALIENCY:

Saliency detection is considered to be 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.



- **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)
- **Itti and Koch ,** The SALIENCY MAP (IEEE Trans on PAMI ’89)

Two forms of visual attention:

-Initial Bottom-up purely data driven

-Refined Task-driven and purposive

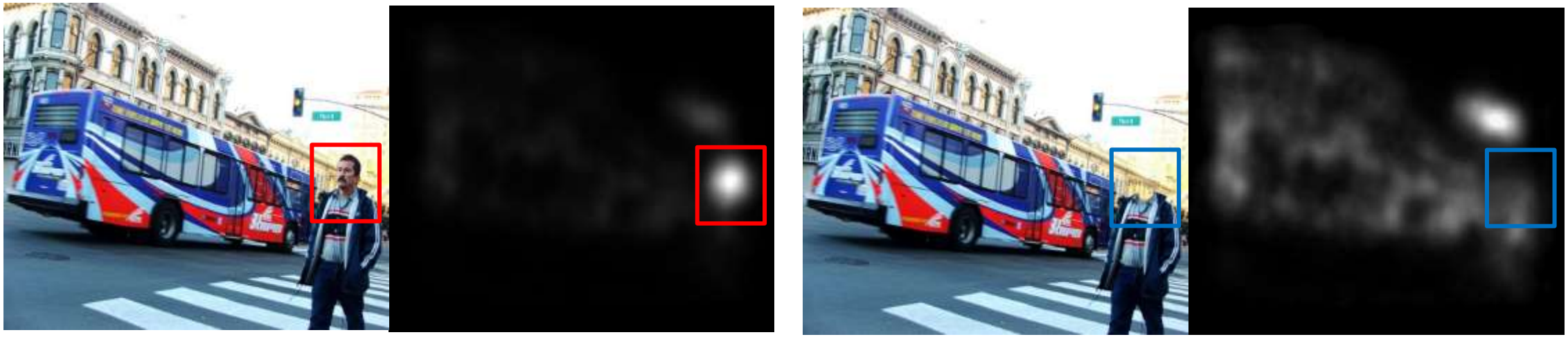
COMPUTATIONAL DETECTION OF SALIENCY MAPS BEFORE DL

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

THE DEEP LEARNING ERA

Problems of annotated Data

2014 ICCV Vig et al.: a three Convnet layer

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;

new competition **CVPR-LSUN 2017**

SALICON benchmark

TUTORIALs AND COMPETITIONS:

2016 ECCV tutorial on Saliency

2017 ICME Competition on 360° Saliency

2017 CVPR New SALICON Competition

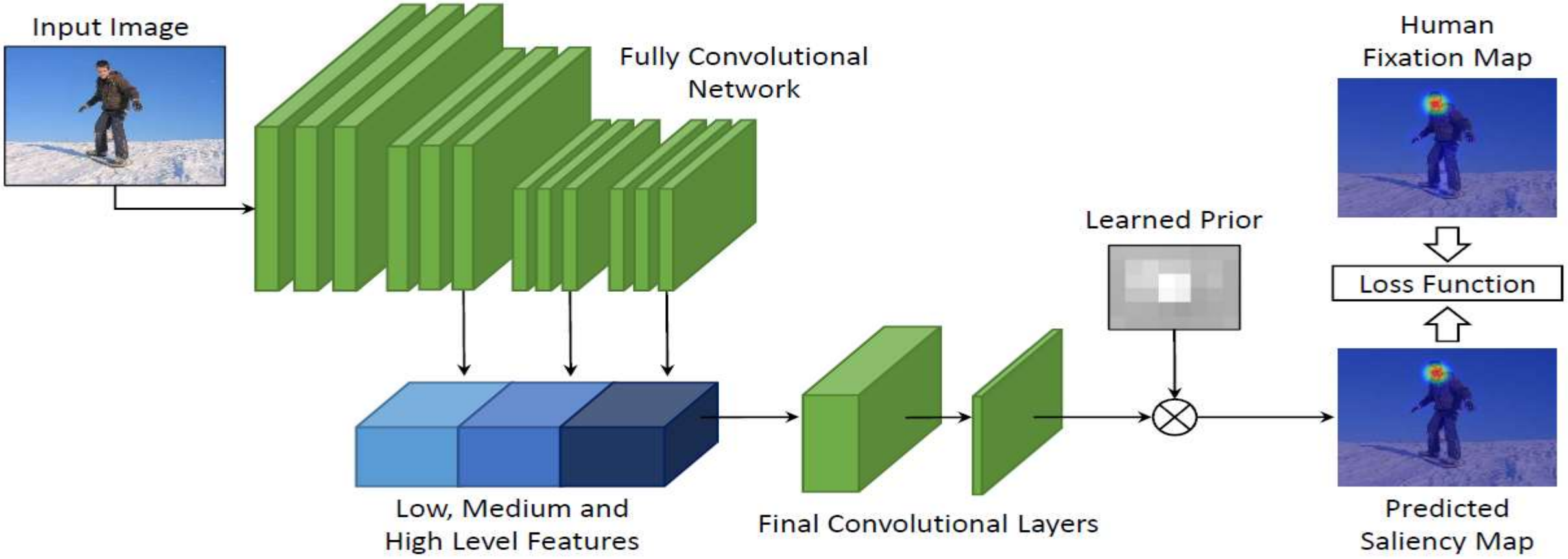
2018 CVPR competition

ARE COMPETITIONS/DATASET USEFUL?

mit saliency benchmark

Trained in SALICON and in MIT300.. Now the net is exploring the world

SALIENCY DETECTION @AIMAGELAB ML-NET*



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

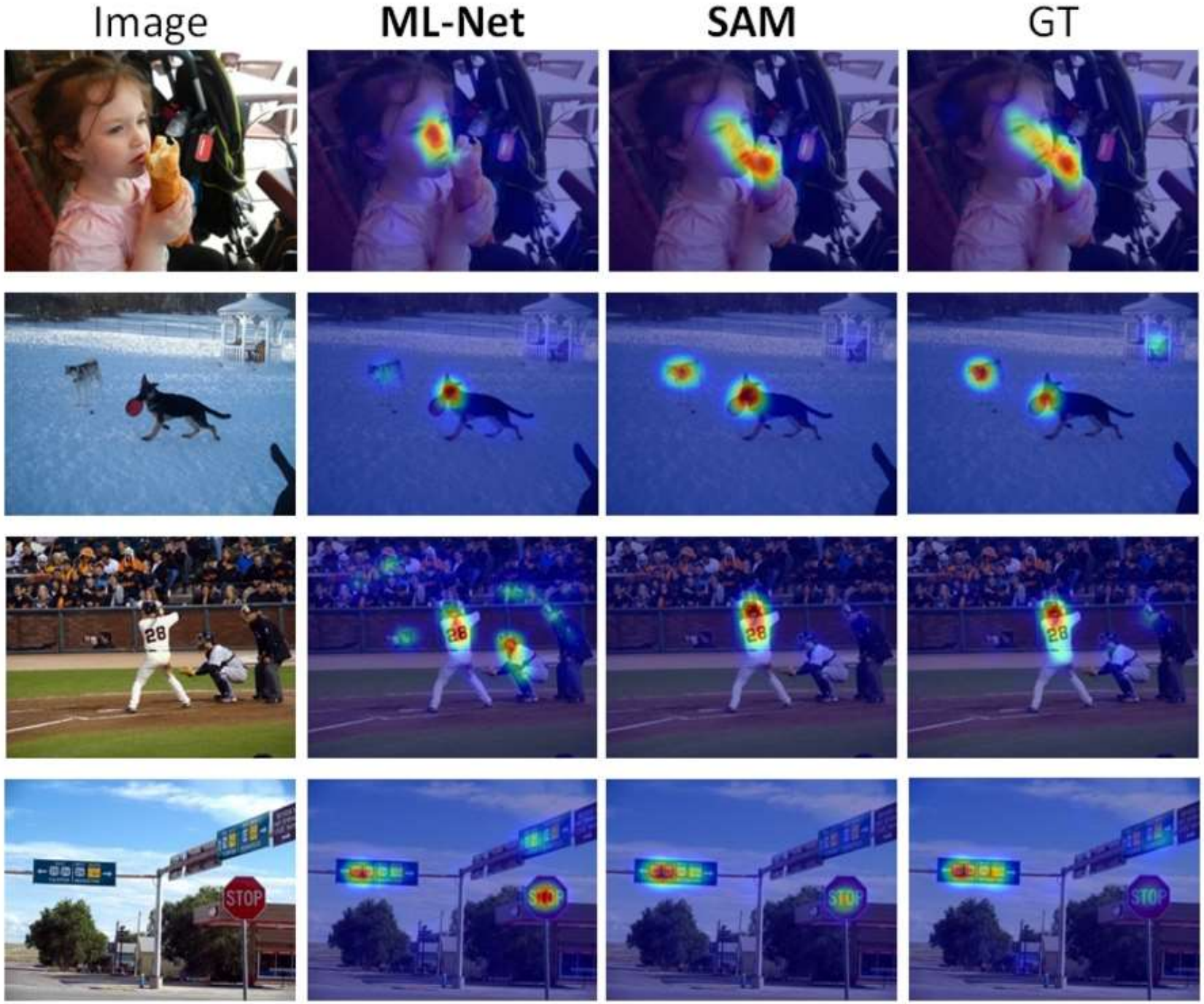
* [M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara. A *Deep Multi-Level Network for Saliency Prediction*, ICPR 2016.]

A second version

Saliency attentive Model SAM*

AImagelab 2016-2018

IMPROVING SALIENCY DETECTION ITERATIVELY



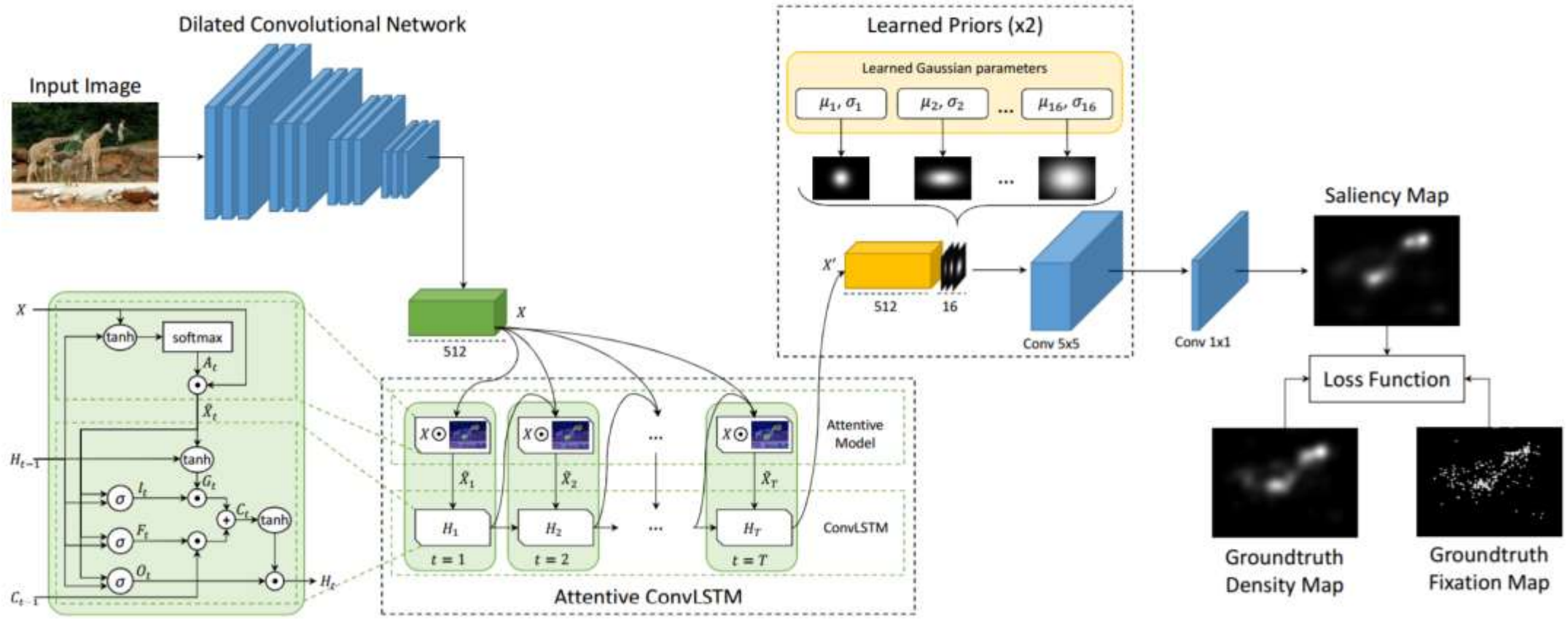
*M.Cornia, L.Baraldi, G.Serra, R.Cucchiara
"Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model"
IEEE Transactions on Image Processing, 2018

SALIENCY DETECTION @IMAGELAB SAM

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

As a sort of Pre-attentive scan-path

The IDEA: define a new **CONV-LSTM** for scan the space and not the time



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

[1] Cornia et al. “ ICPR, 2016.
[2] Kruthiventi et al.CVPR, 2016.
[3] Pan et al. “CVPR, 2016.
[4] Kümmerer et al. “DeepGaze II: arXiv 2016.

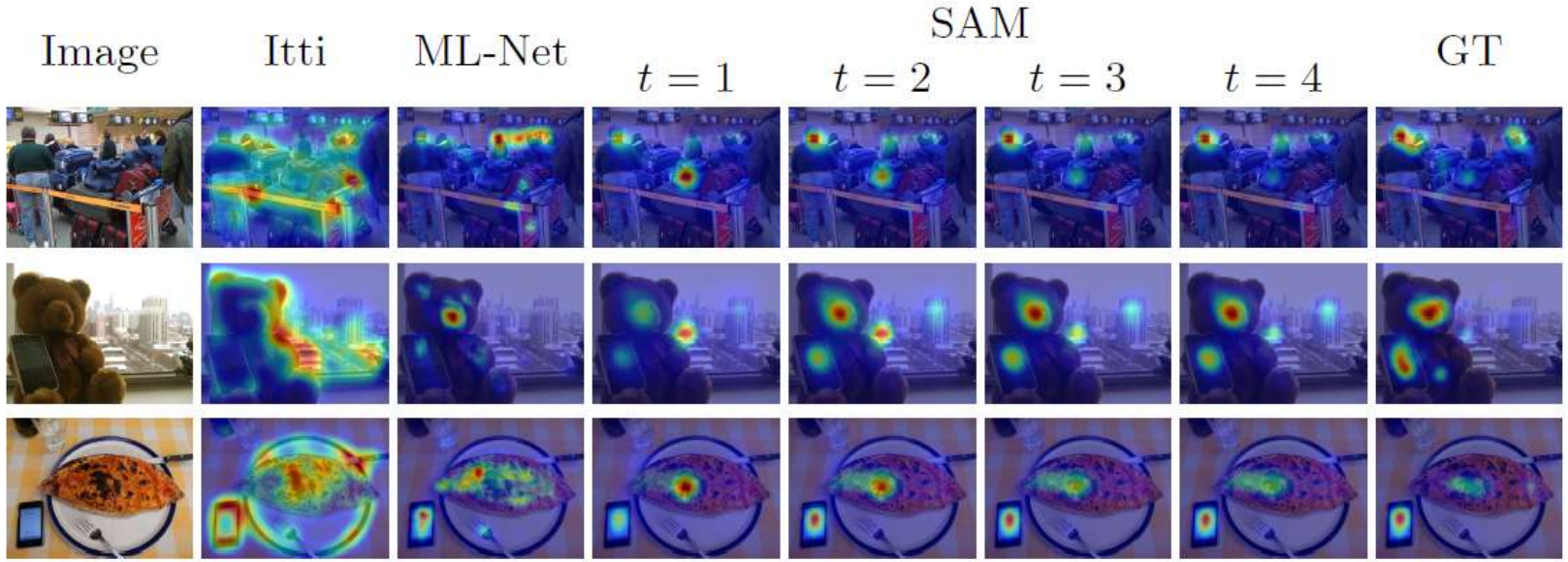
SALICON Dataset (new release)

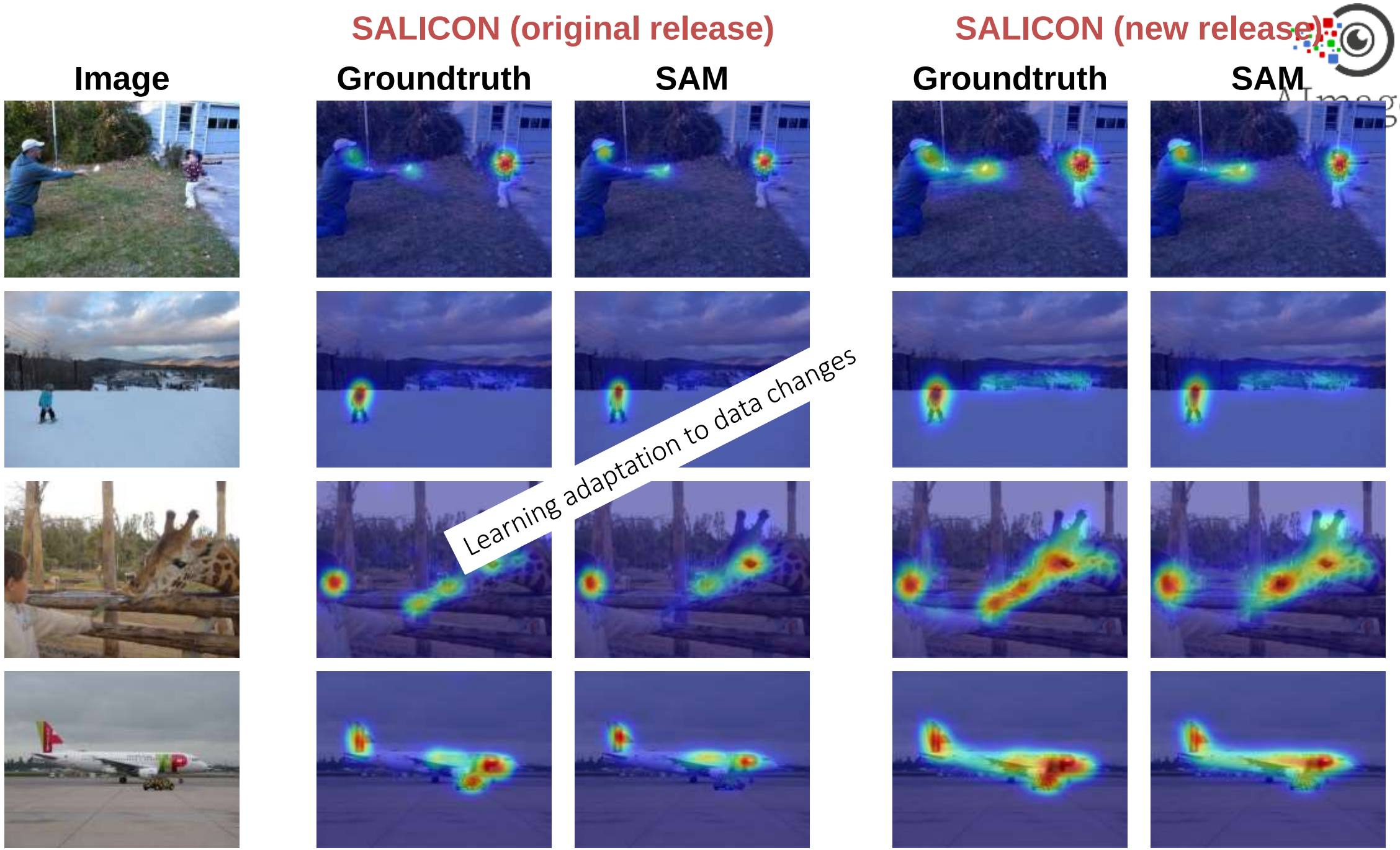
	CC	sAUC	AUC	NSS
SAM	**0.899**	**0.741**	**0.865**	**1.990**

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)

SAM : winner in the LSUN Challenge CVPR 2017



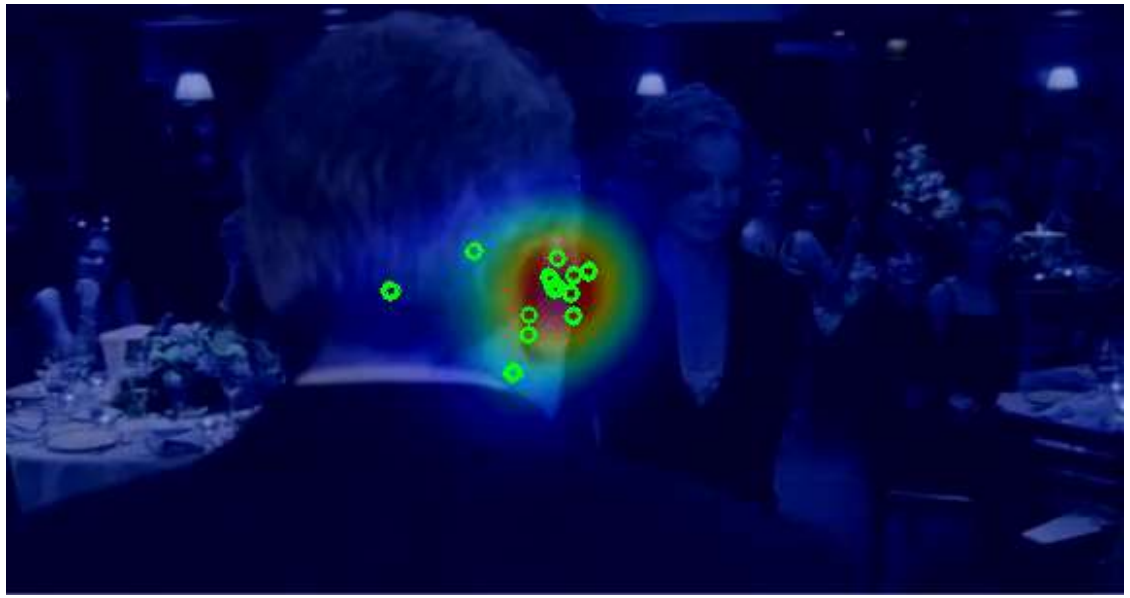

SALICON (original release)
SALICON (new release)
Image
Groundtruth
SAM
Groundtruth
SAM
Learning adaptation to data changes

SAM FOR VIDEO

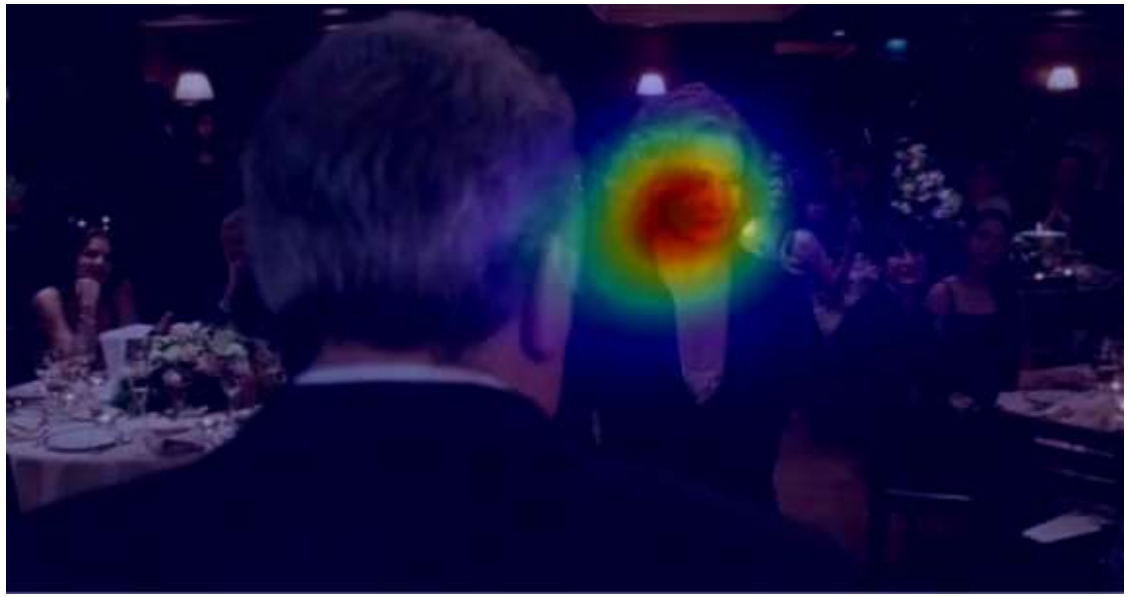
Few approaches to video

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

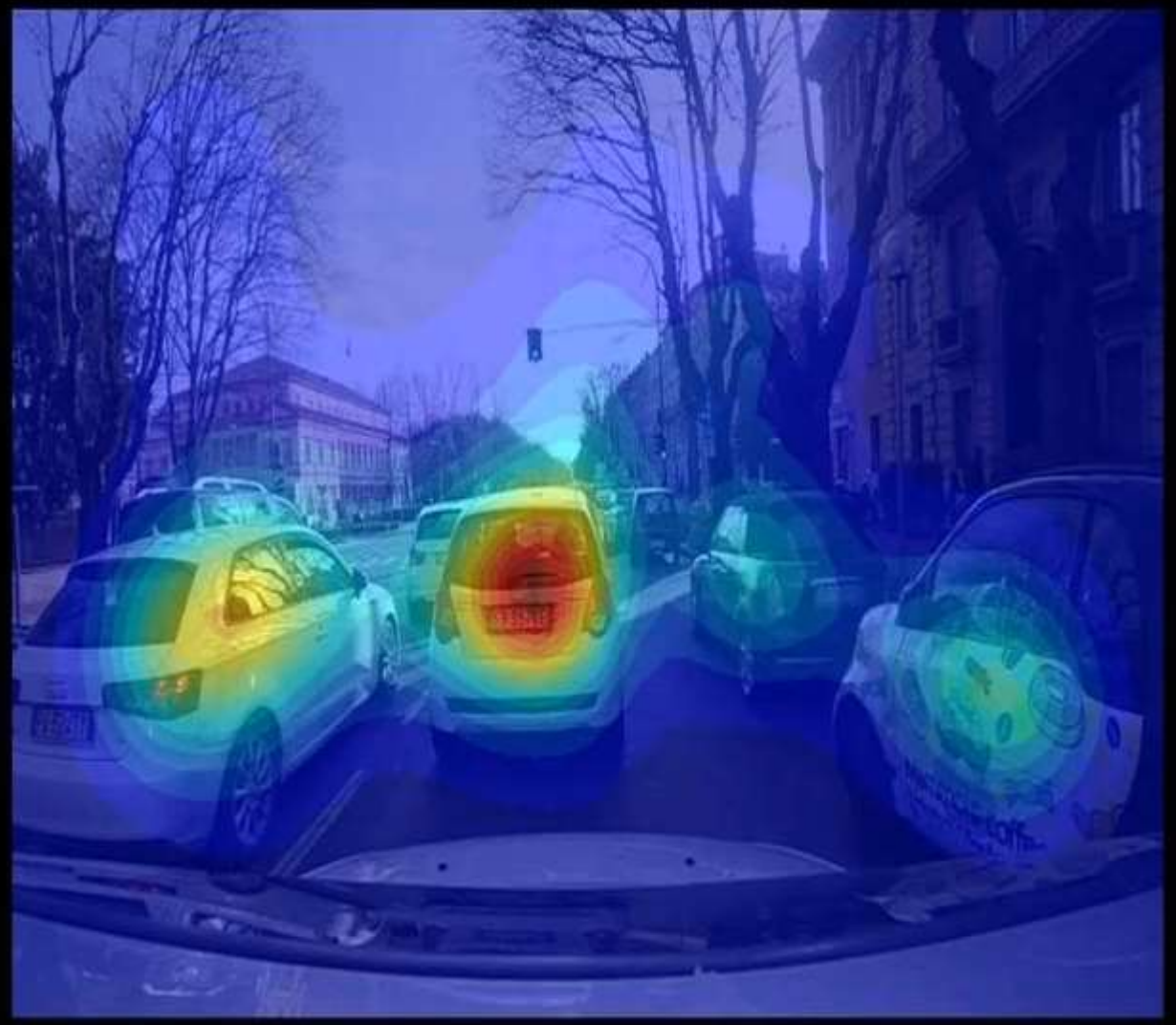


Groundtruth



SAM

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



Bottom-up saliency

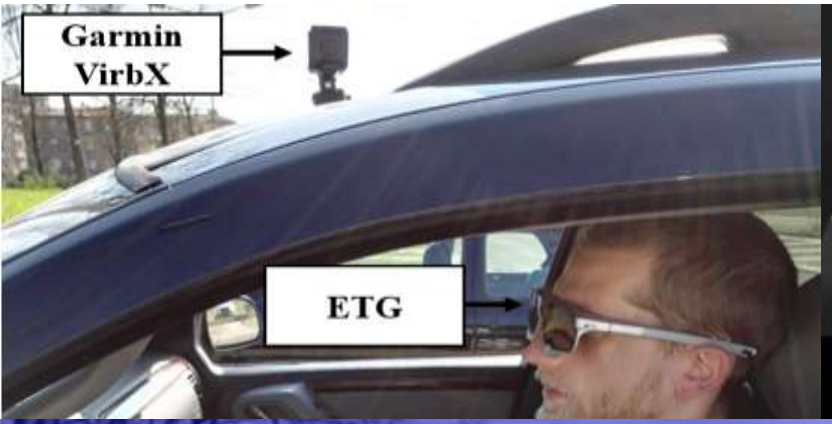
Task-driven saliency

THE DR(EYE)VE PROJECT

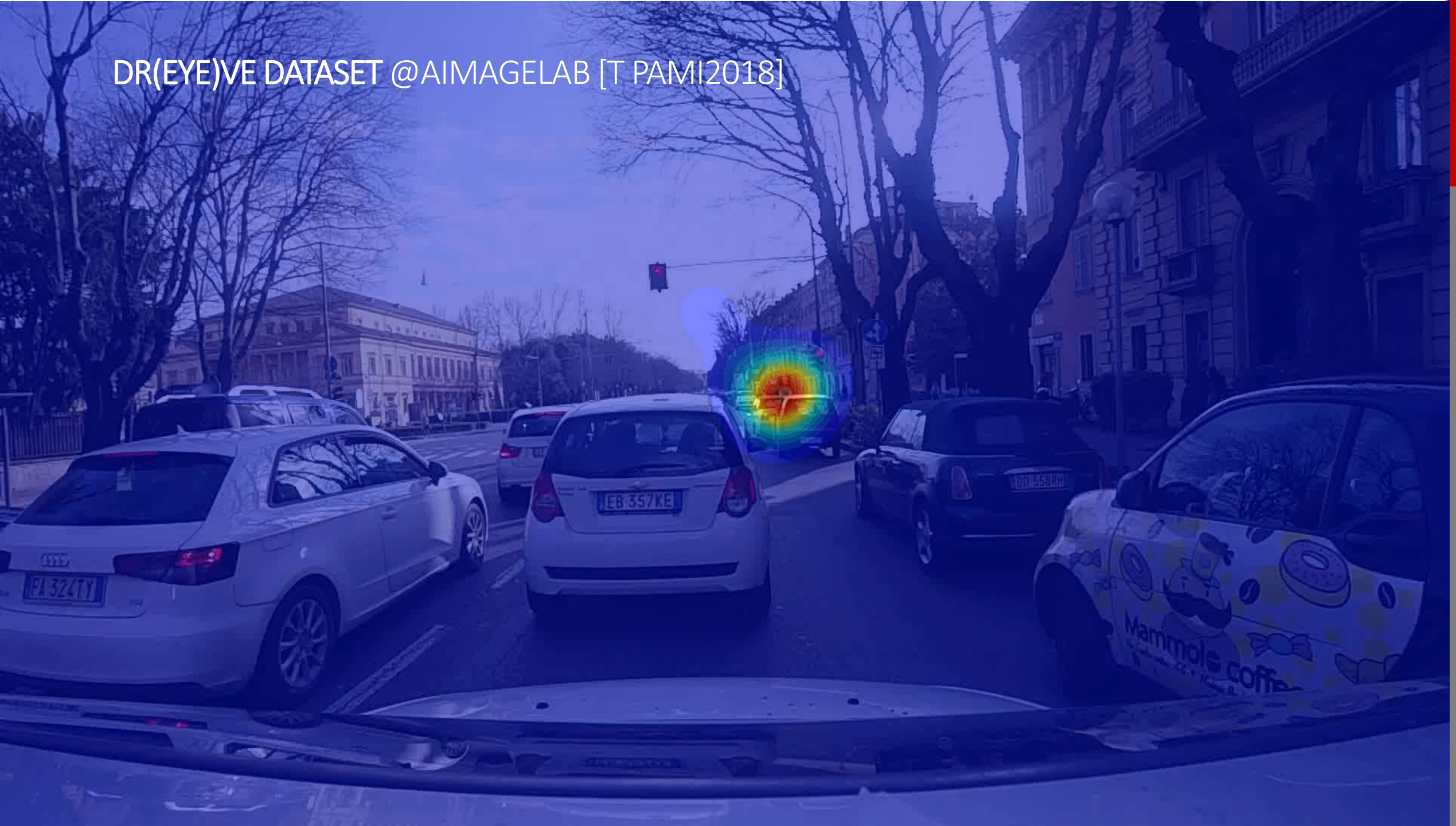
Aimagelab.unimore.it/Dreyeve

- 3 months for acquisition
- Eytrackers and camera-car
- SIFT-based image registration frame by frame
- Automatic annotation of gaze, speed, GPS position

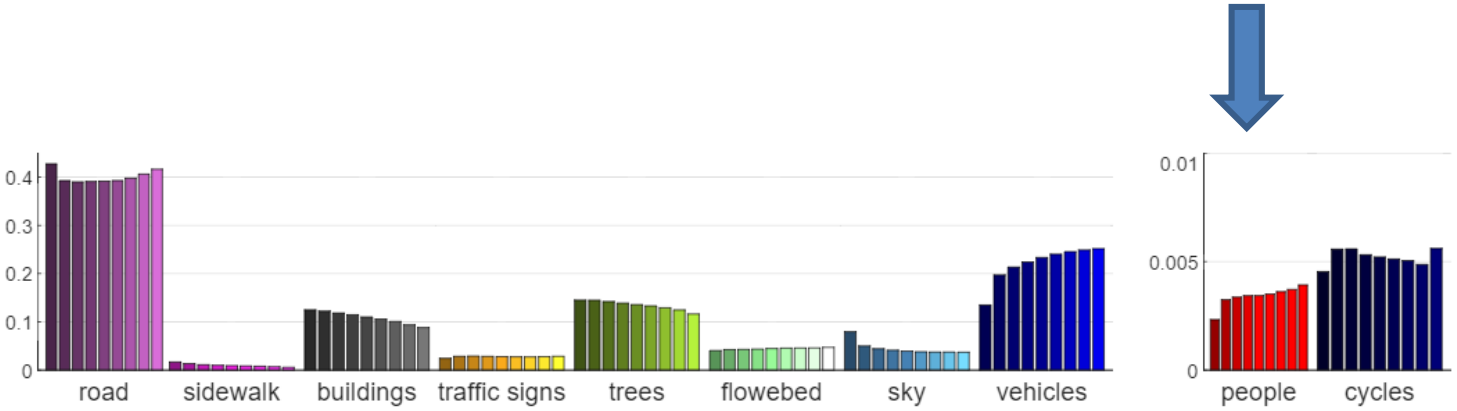
Are the drivers looking at people when driving?



DR(EYE)VE DATASET @AIMAGELAB [T PAMI2018]



(HUMAN) ATTENTIVE BEHAVIOR, MEASURED



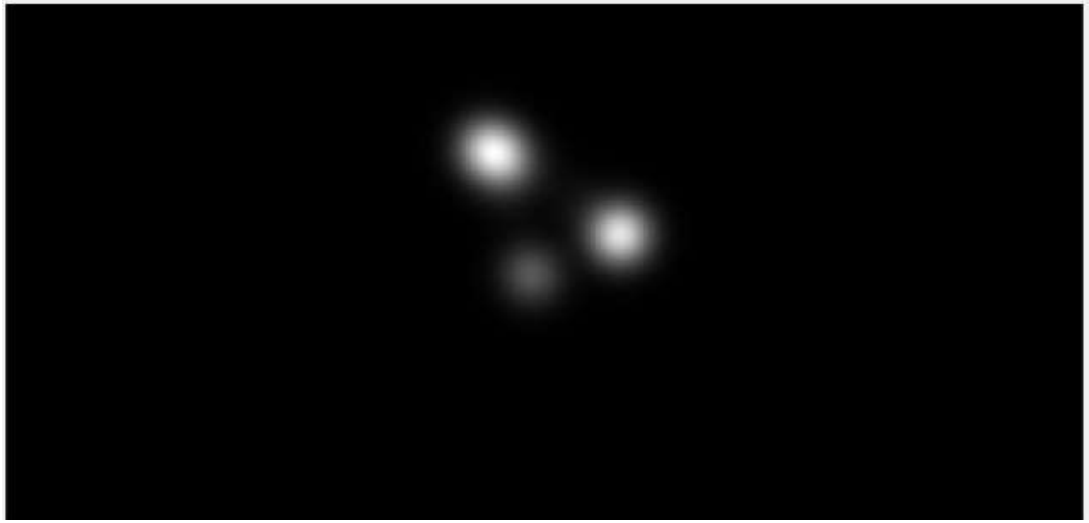
Drivers do not care pedestrians ..

A.Palazzi, D-Abati, S.Calderara, F.Solera, R.Cucchiara. «Predicting the Driver's Focus of Attention: the DR(EYE)veProject IEEE Transactions on PAMI 2018

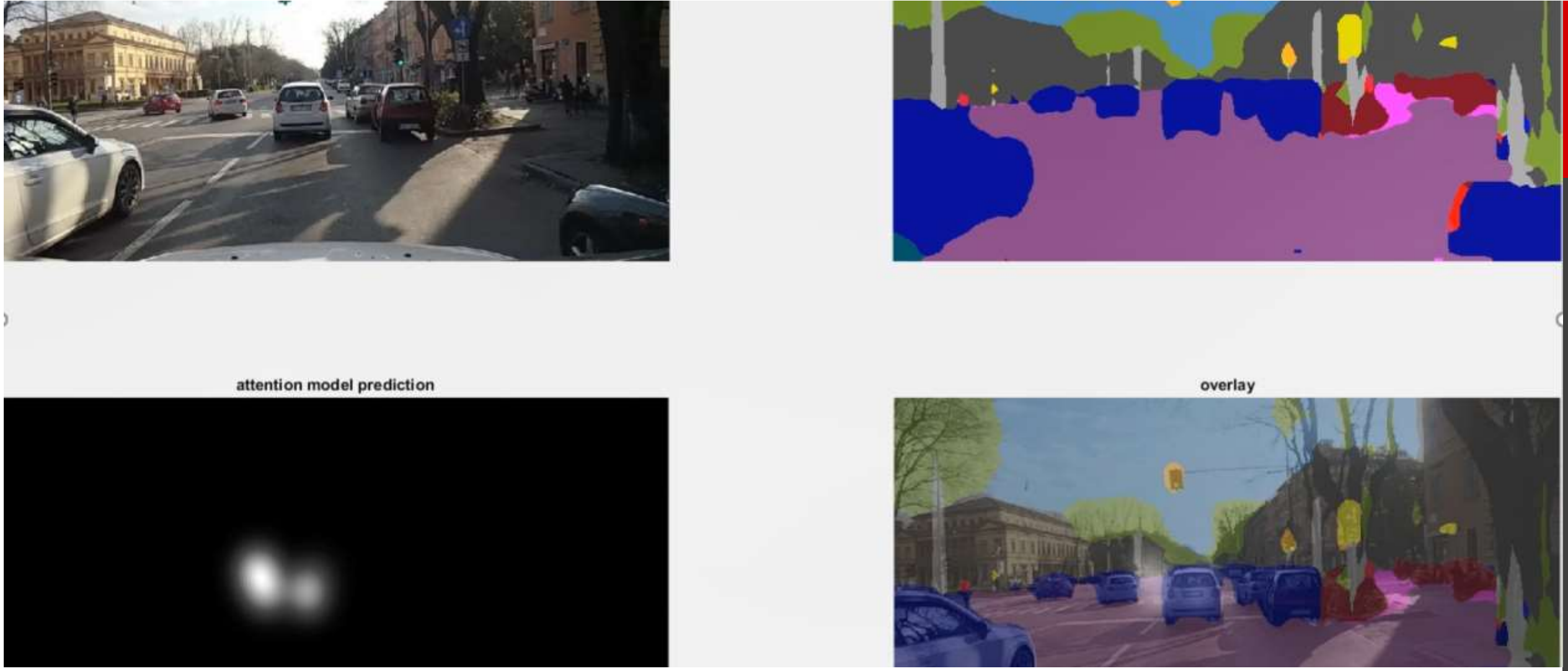
roof mounted camera

semantic segmentation

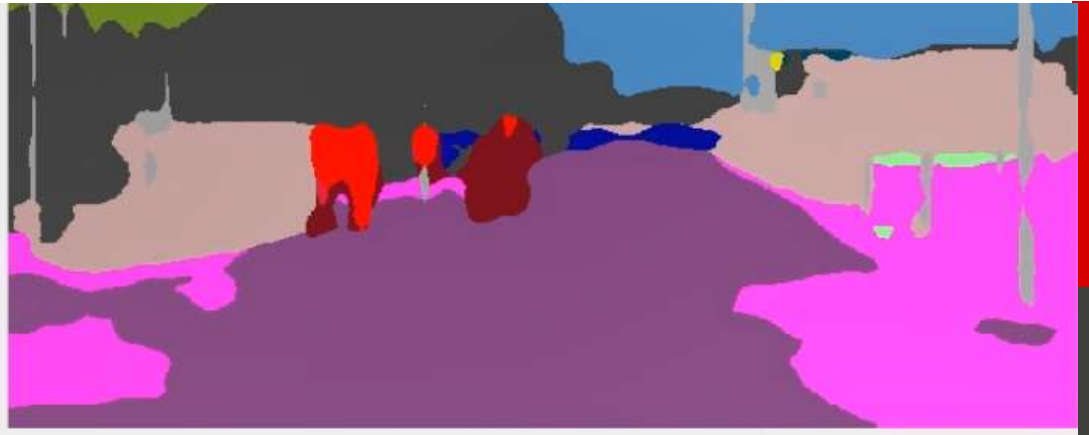
attention model prediction



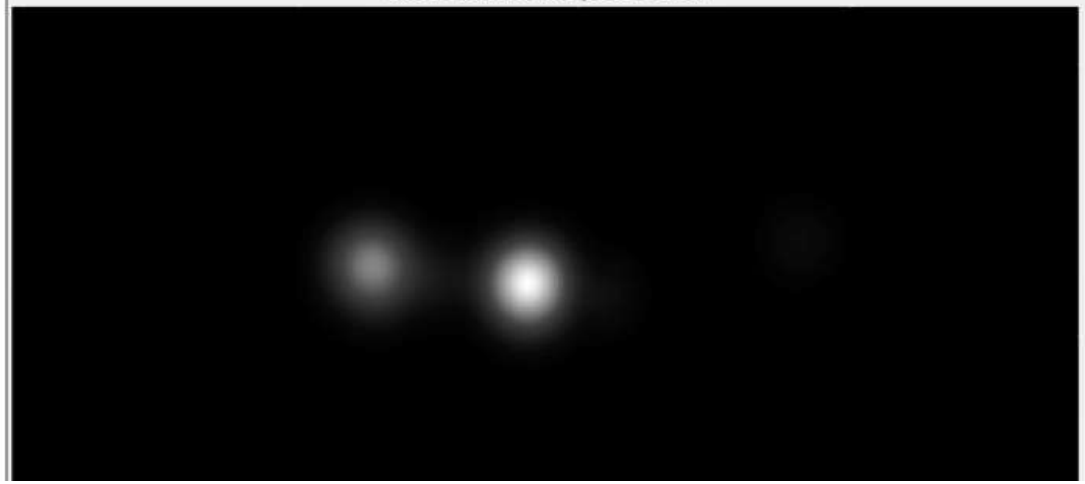
overlay



There are pedestrians around.. How many? Is the DL-based detection correct?



attention model prediction



overlay

There are pedestrians around.. How many? Where they are going?

We need Detection and Tracking together for Visual Intelligence

USE SALIENCY FOR A BETTER TEXTUAL DESCRIPTION

AUTOMATIC IMAGE/VIDEO CAPTIONING

A fantastic compress form of description

From VISUAL DATA

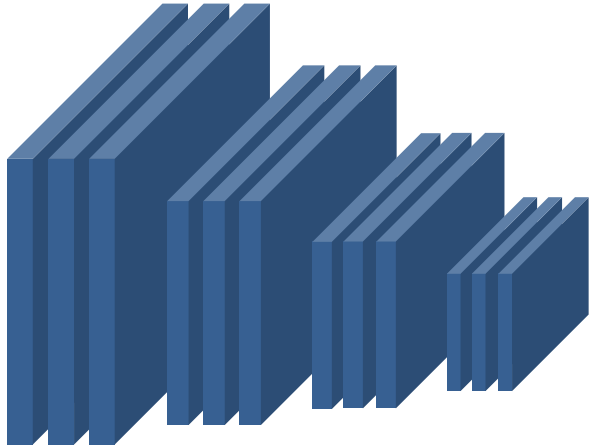
To TEXT SEQUENCE

Pr: a person is cutting a potato

From MS COCO
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

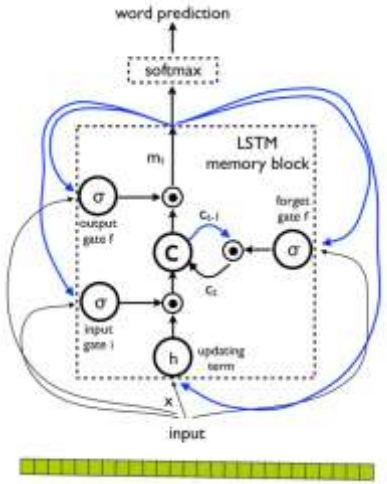
MANY ATTEMPTS TO CAPTIONING

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



CONV-NET

+



Recurrent NET (LSTM)

IMPORTANT REFERENCES

[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

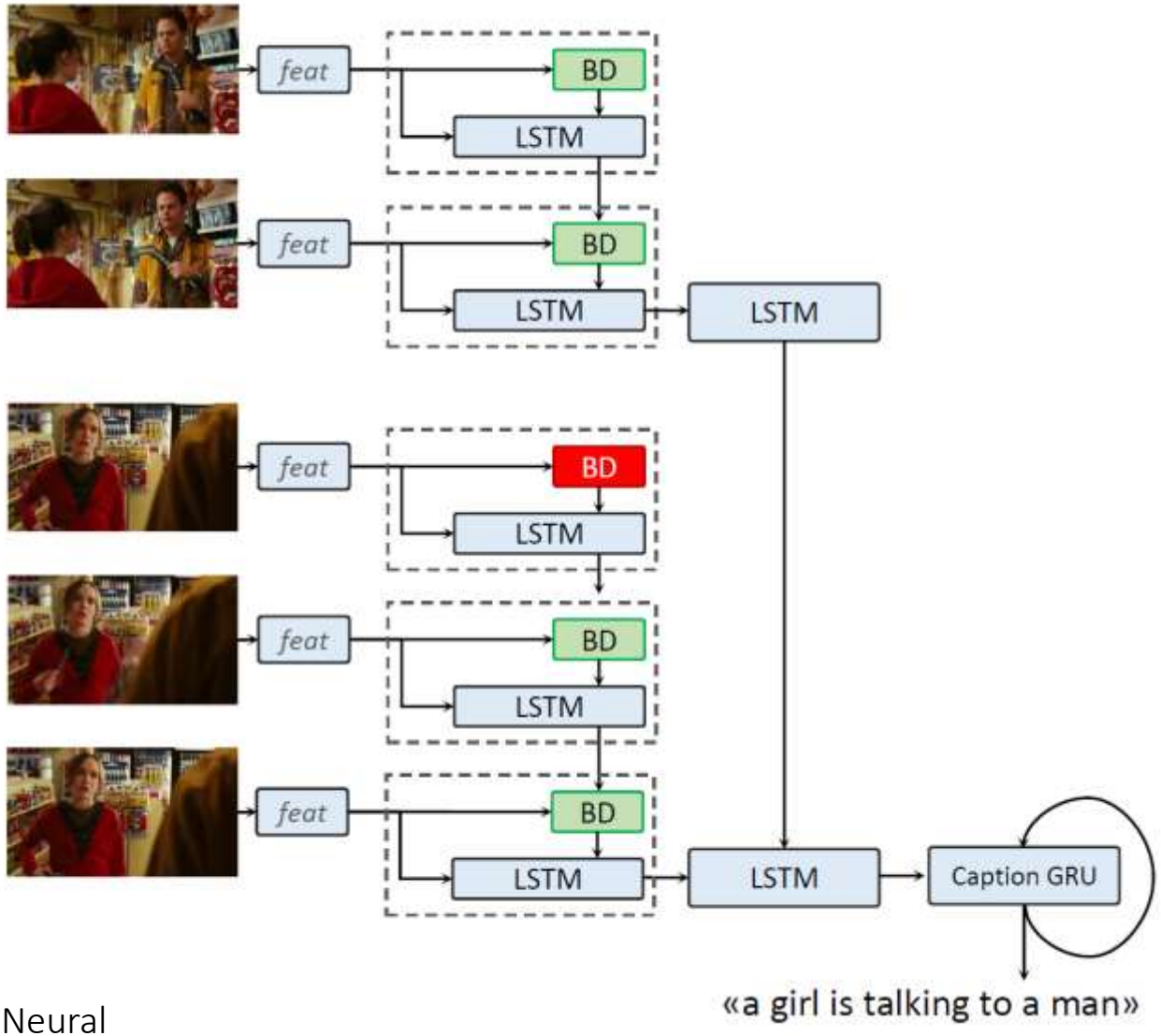
IMPROVING SENTENCE GENERATION IN LONG VIDEO

A new architecture for video captioning with the capacity of forgetting

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

A long training but now in real time

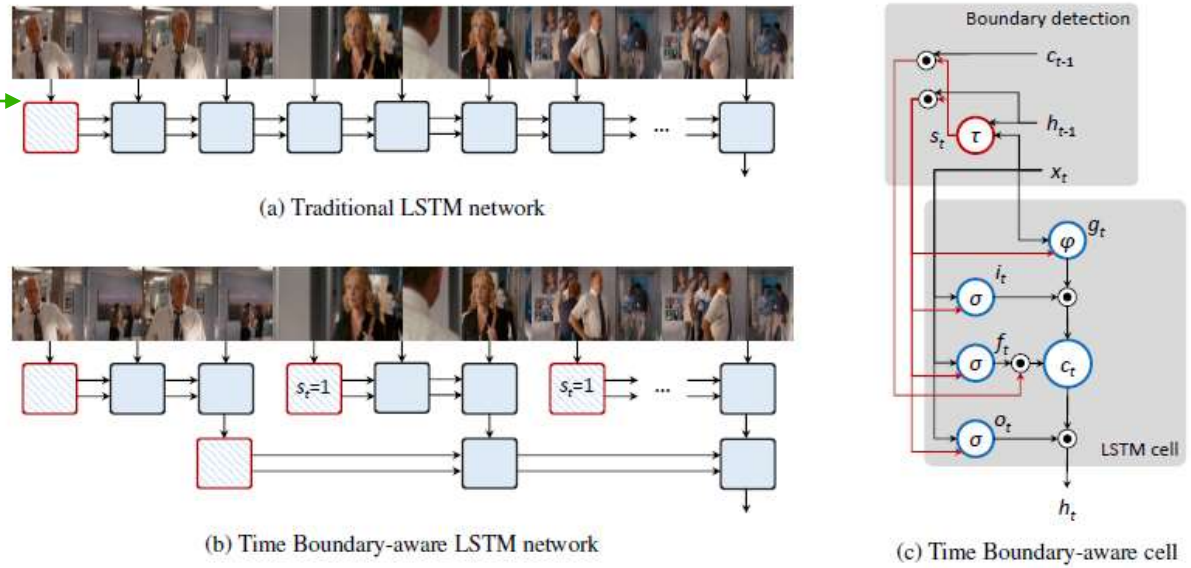


[L. Baraldi, C. Grana, and R. Cucchiara , Hierarchical Boundary-Aware Neural Encoder for Video Captioning CVPR 2017]

A SUITABLE MODIFICATION OF LSTM WITH BOUNDARY DETECTION

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

Features from
ResNet 50 (Imagenet)
+
C3D (Sport-1M for motion)



(a) Traditional LSTM network

(b) Time Boundary-aware LSTM network

(c) Time Boundary-aware cell

Vocabulary of

- M-VAD 6090 words (84,6 hours of 92 Hollywood movies, 46K video clips for impairs)
- MPII-MD 7198 words (94HD movies with 68K sentences)
- MSVD 4125 words (Microsoft 2K Youtube clips with 85K sentences)
- + <BOS> and <EOS>

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.

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 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: SALIENCY AND CAPTIONING

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

GT: he stands and offers her the small bouquet

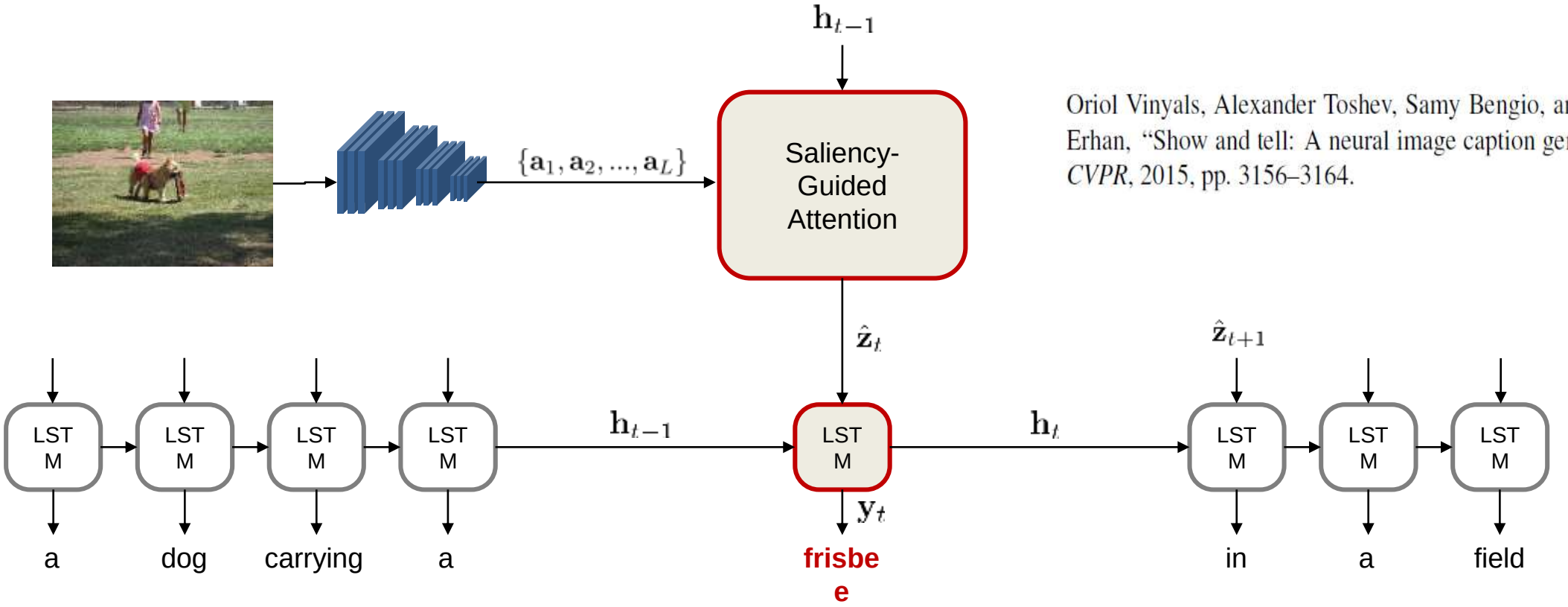
Pr: someone looks up at someone

Attention in persons and not in the salient objects!

[M. Cornia, L.Baraldi, G. Serra, R.Cucxhiara Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention ACM Transactions TOMM 2018]

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.

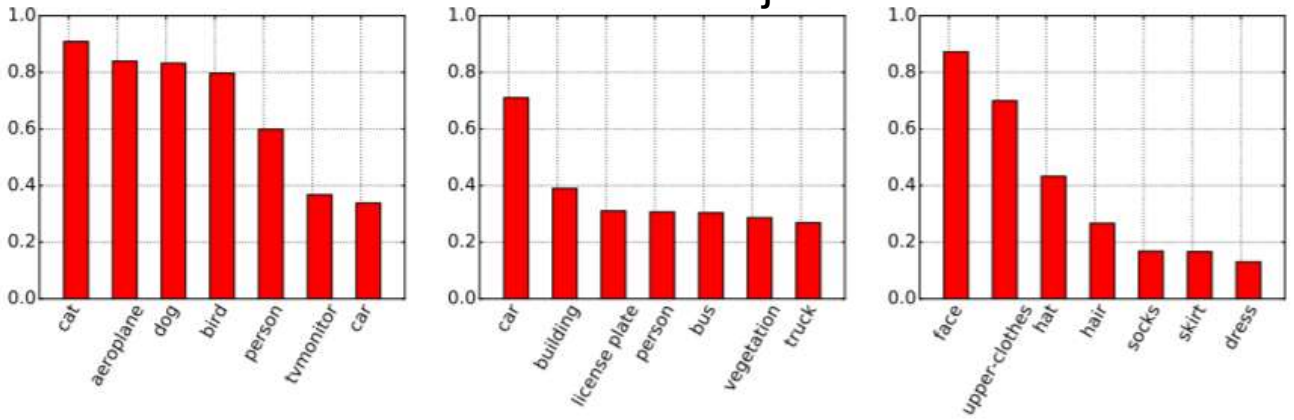


Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan, "Show and tell: A neural image caption generator," in *CVPR*, 2015, pp. 3156–3164.

PRE-ATTENTIVE SALIENCY AND CONTEXT

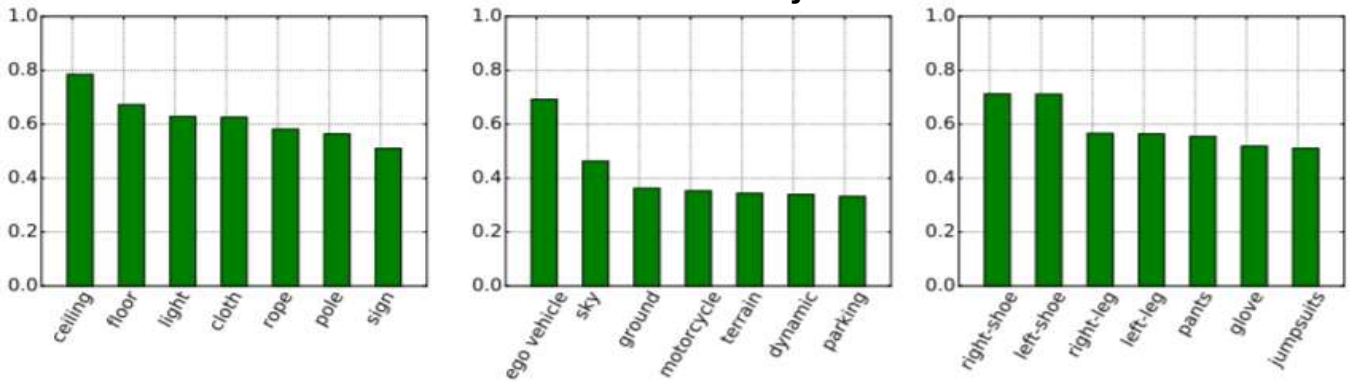
Adding **Saliency and Context as a bottom-up pre-attentive behavior**

Most salient objects



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

Less salient objects



Some Experiments:

WHAT IS HIT BY SALIENCY (SAM)

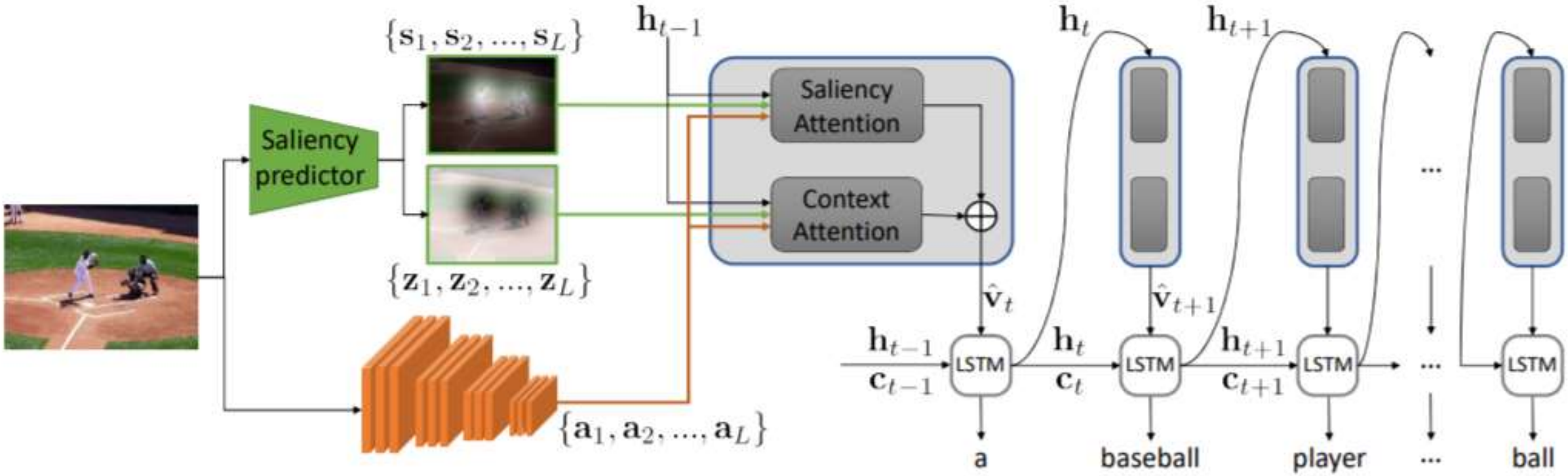
Pascal c.a. 20K images, 400 labels
Cityscapes c.a. 5K images, 30 classes
LIP c.a. 50K images , 19 human body parts

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

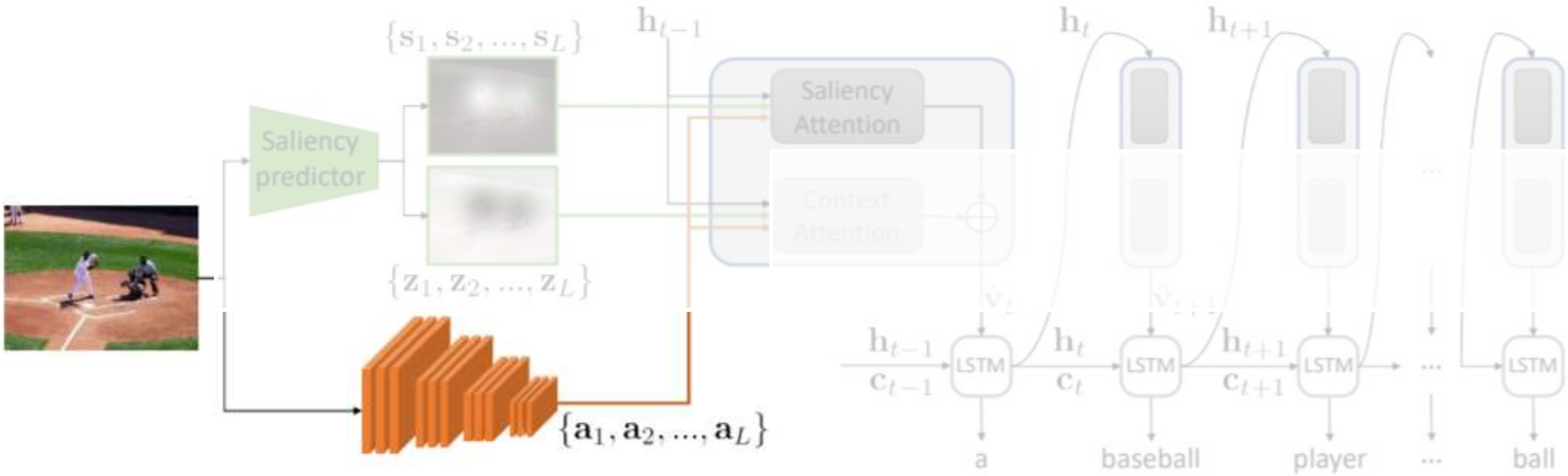
2. Saliency is independent on the object size (thus importance is not related to size)

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

RATIONALE 3 FINAL PROPOSAL

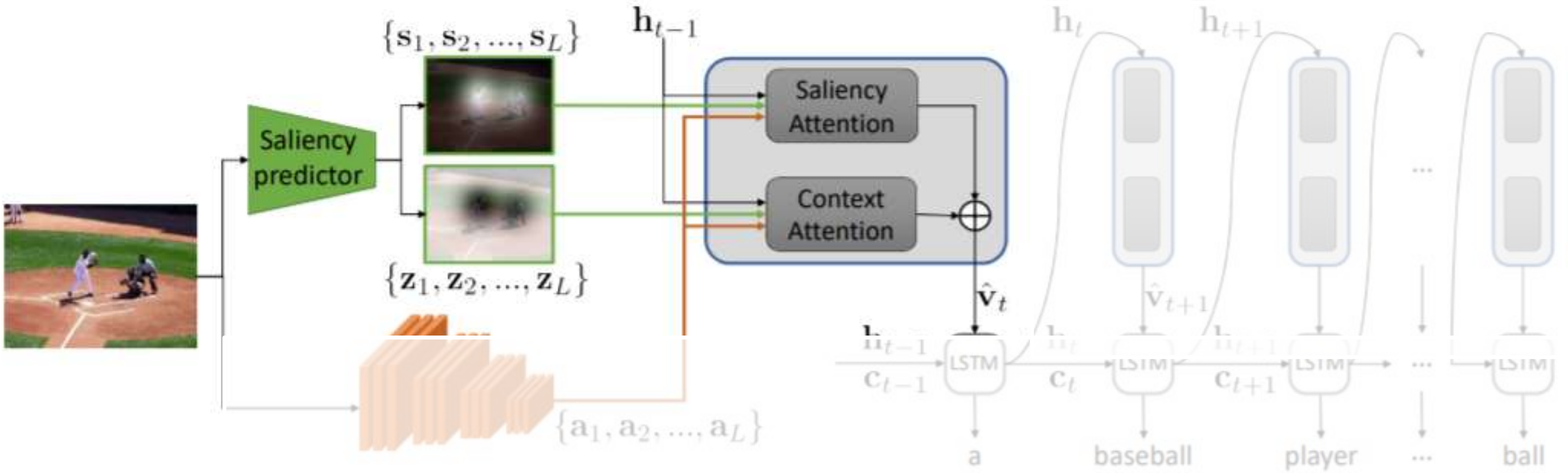


RATIONALE 3 FINAL PROPOSAL



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

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.

EXPERIMENTAL RESULTS

SALICON Dataset

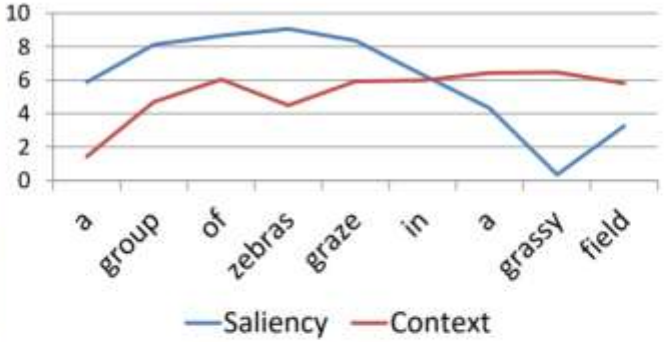
	CNN	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE$_L$	CIDEr
Soft Attention	VGG-16	0.680	0.501	0.358	0.256	0.222	**0.497**	0.691
Saliency-Guided Attention	VGG-16	**0.682**	**0.505**	**0.361**	**0.258**	**0.223**	**0.497**	**0.694**
Saliency-Guided Att. (with GT saliency maps)	VGG-16	*0.684*	*0.503*	*0.360*	*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*

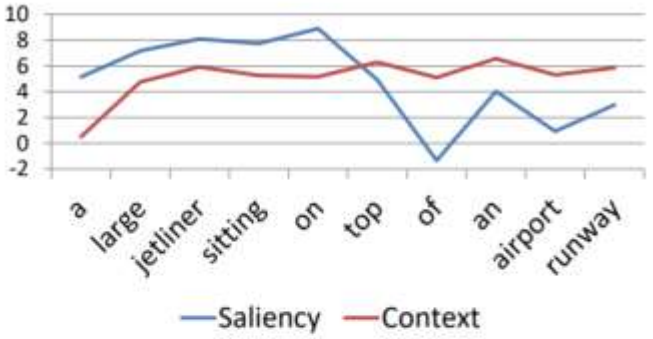
SALICON Dataset (subset of the Microsoft COCO dataset, composed by 20,000 images, largest available dataset for saliency prediction)

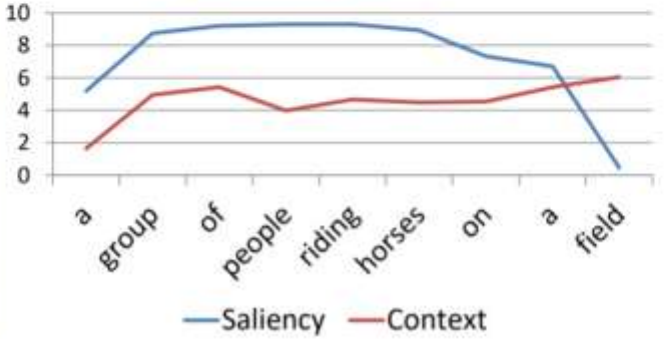
Microsoft COCO Dataset

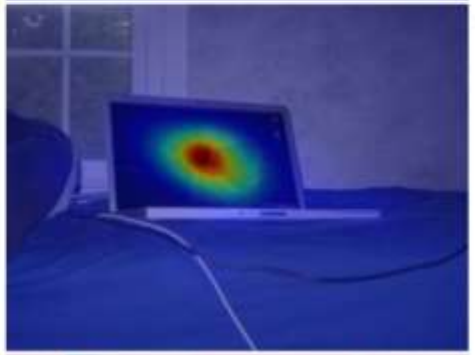
	CNN	BLEU@1	BLEU@2	BLEU@3	BLEU@4	METEOR	ROUGE$_L$	CIDEr
Soft Attention	ResNet-50	0.717	0.546	0.402	0.294	0.253	0.529	0.939
Saliency-Guided Attention	ResNet-50	**0.718**	**0.547**	**0.404**	**0.296**	**0.254**	**0.530**	**0.944**

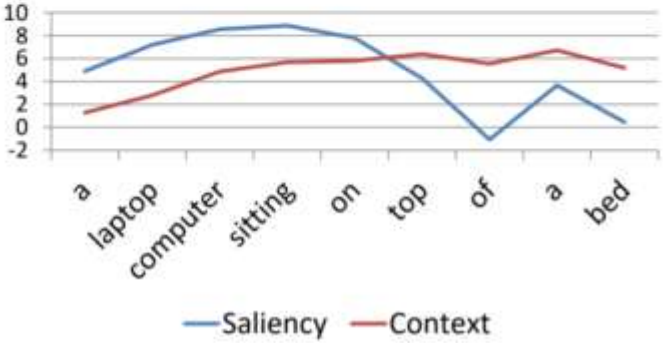
Microsoft COCO Dataset (composed by more than 120,000 images divided in training and validation sets each image is annotated with five sentences)

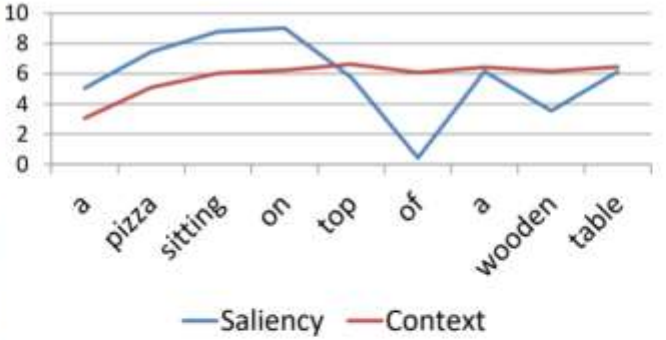

10
8
6
4
2
0
a group of zebras graze in a grassy field
Saliency
Context

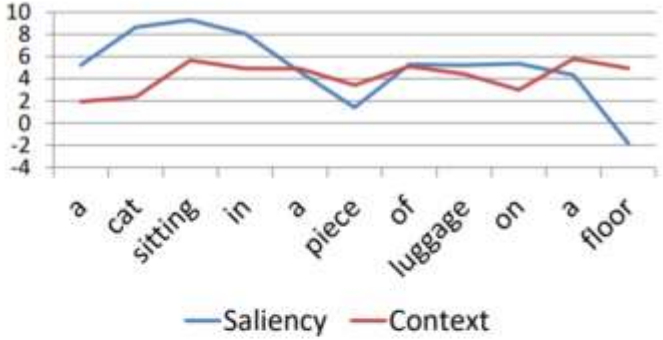

10
8
6
4
2
0
-2
a large jetliner sitting on top of an airport runway
Saliency
Context


10
8
6
4
2
0
a group of people riding horses on a field
Saliency
Context

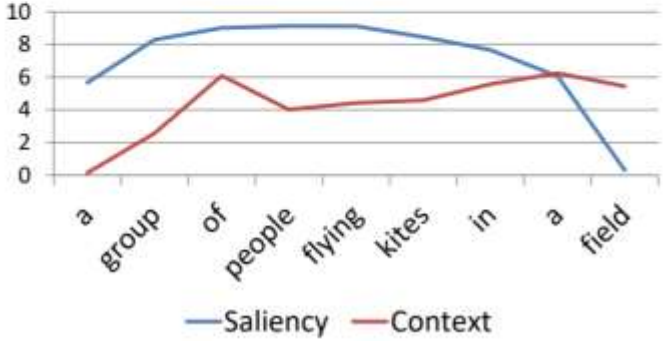



10
8
6
4
2
0
-2
a laptop computer sitting on top of a bed
Saliency
Context

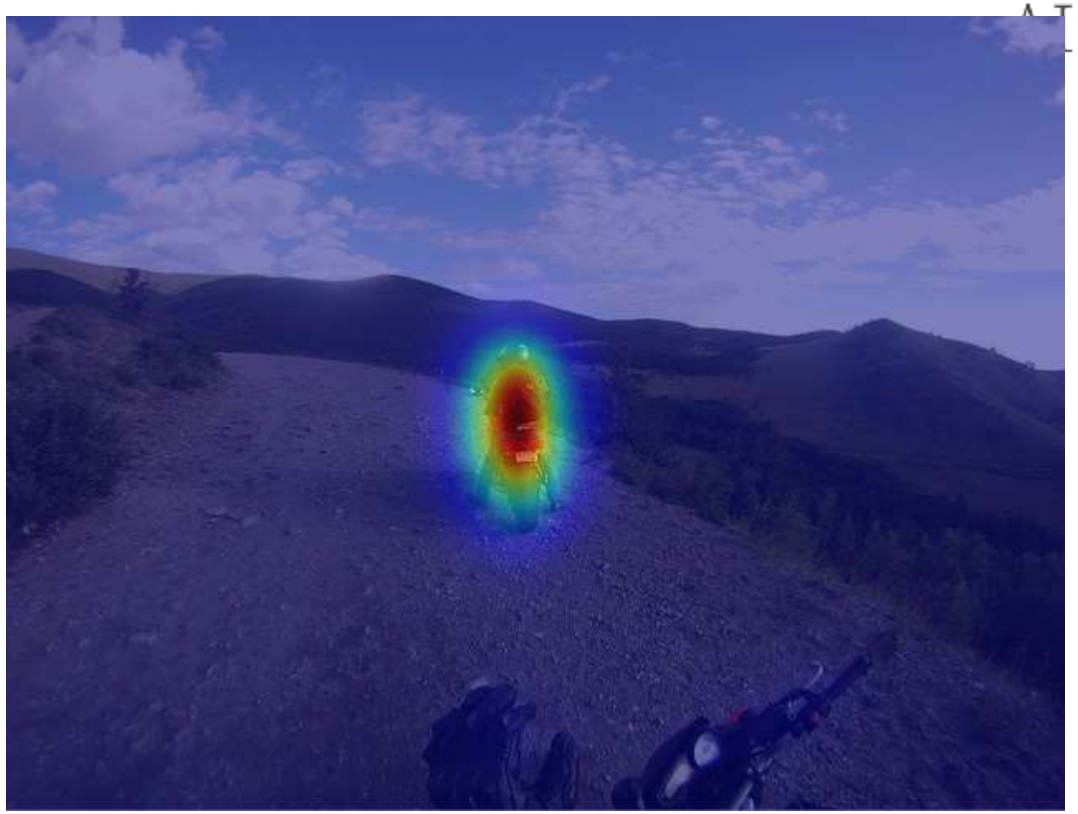

10
8
6
4
2
0
a pizza sitting on top of a wooden table
Saliency
Context


10
8
6
4
2
0
-2
-4
a cat sitting in a piece of luggage on a floor
Saliency
Context


10
8
6
4
2
0
-2
a double decker bus is parked in front of a building
Saliency
Context


10
8
6
4
2
0
a group of people flying kites in a field
Saliency
Context

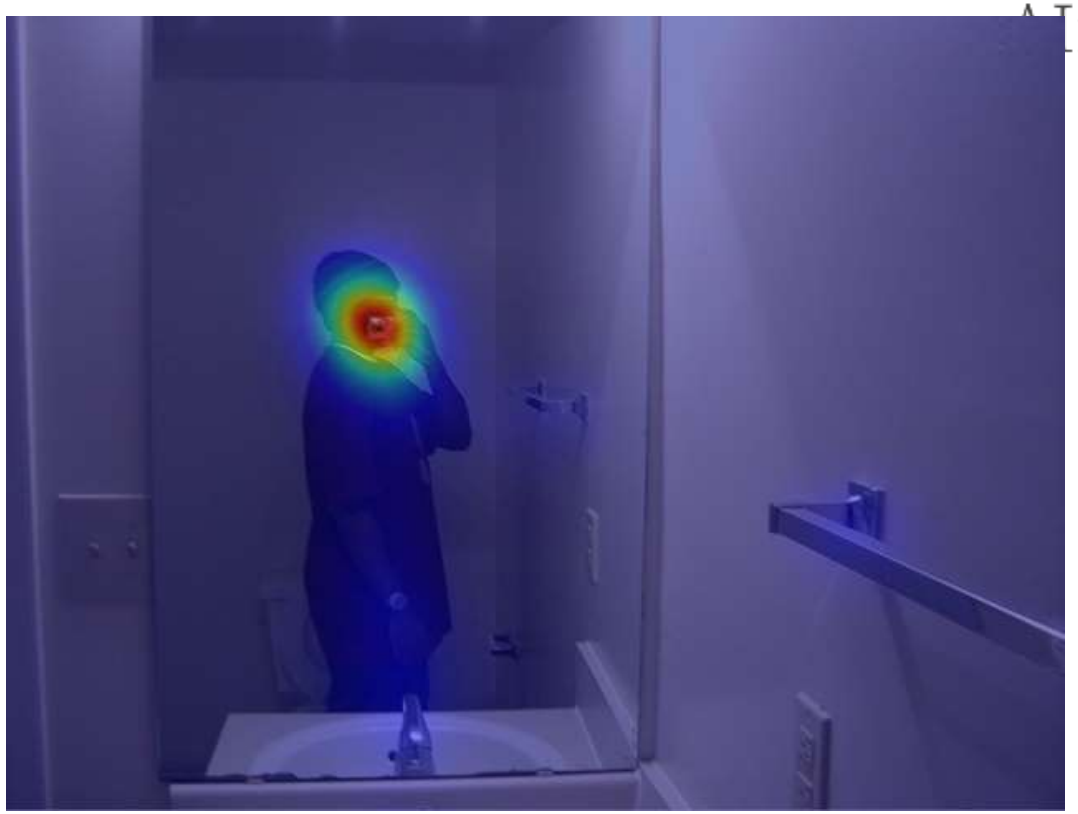
QUALITATIVE RESULTS



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

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

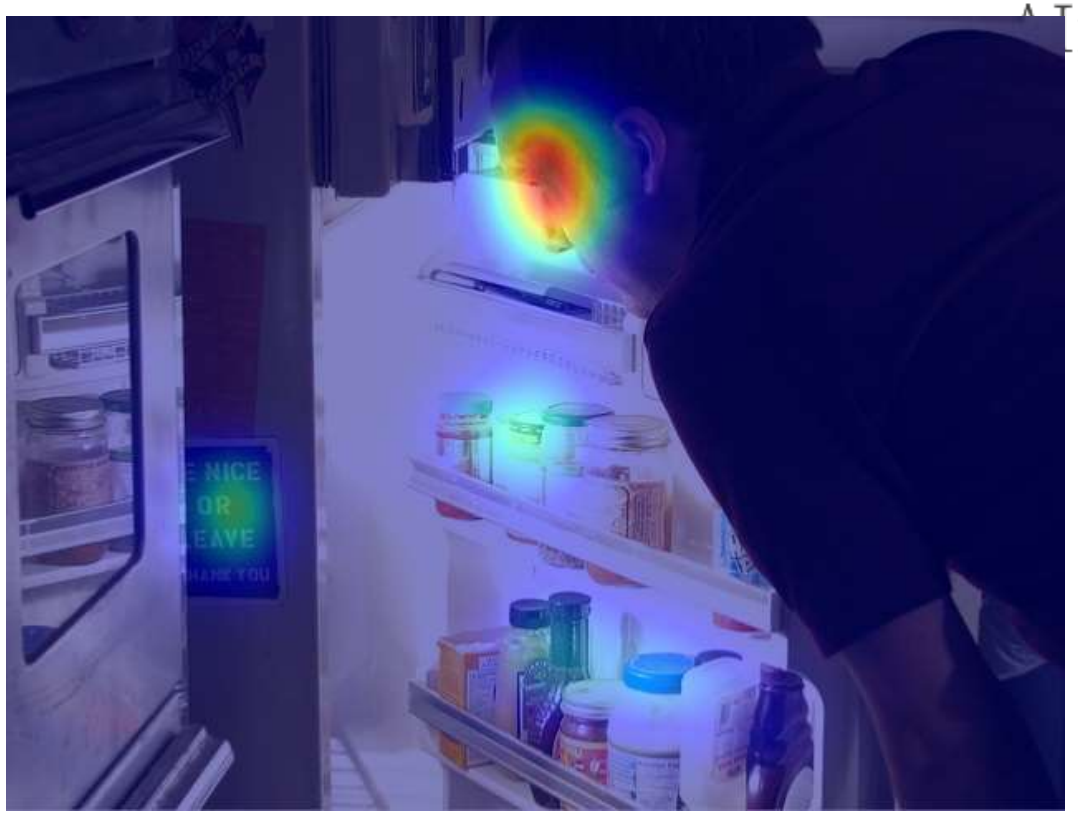
QUALITATIVE RESULTS



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

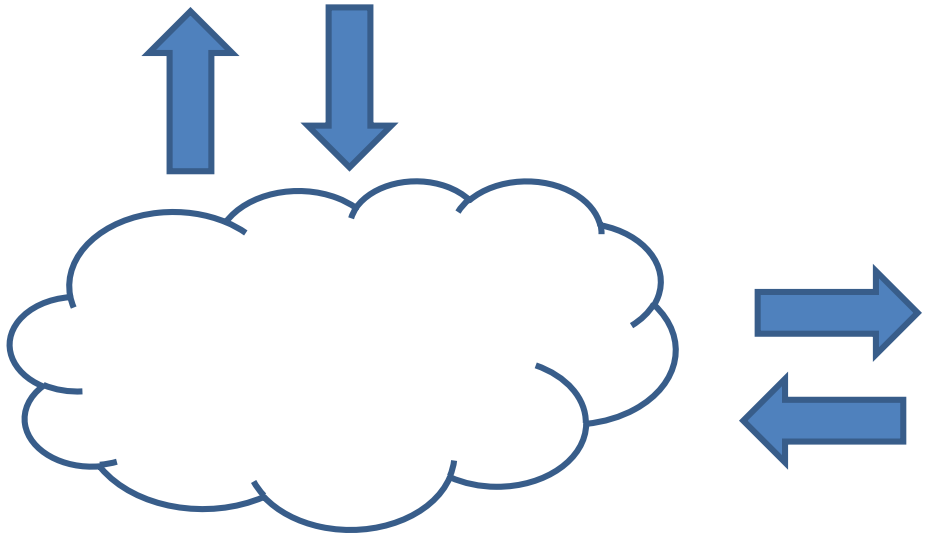
Without : A bathroom with a sink and a sink.

QUALITATIVE RESULTS



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

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



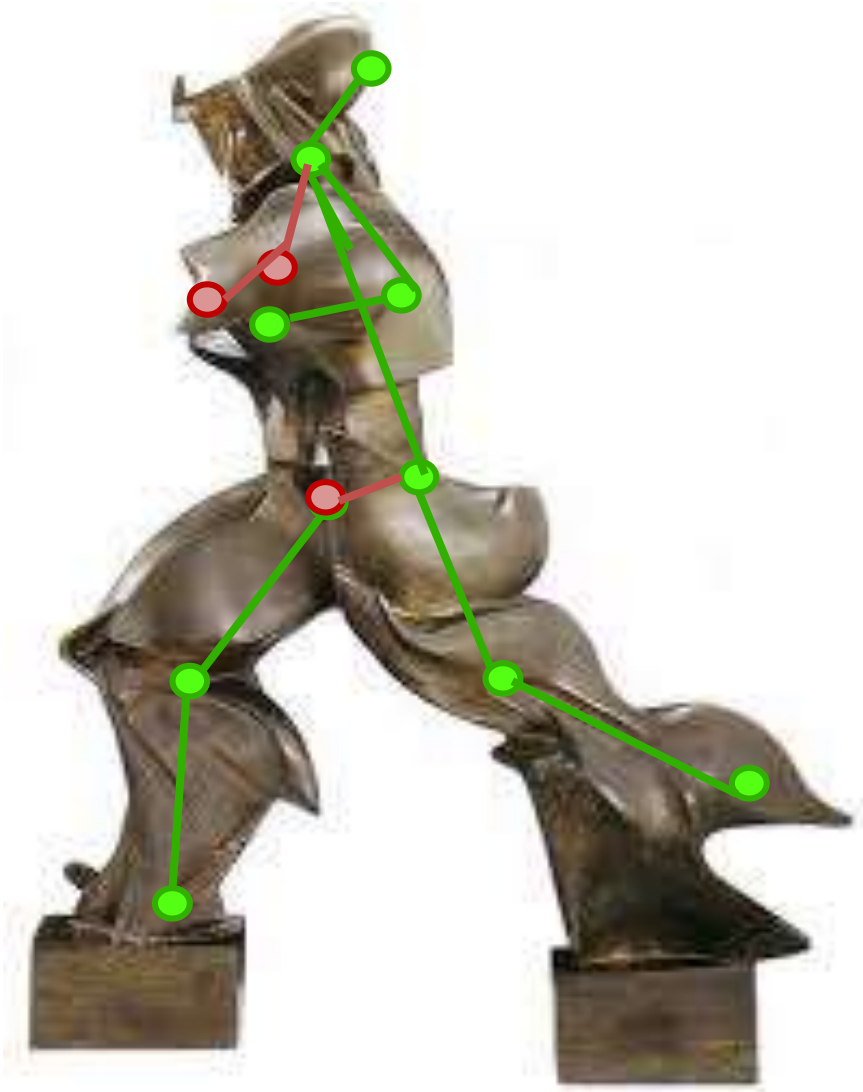
From the Loomo Robot to Facebook GPU Server and vice-versa (about 15 frame per seconds)


"two standing women have a phone and a cup "

3. SEE THE HIDDEN (OCCLUSIONS)

THE NEW TREND

DETECTION BY POSE



MANY RELATED WORKS – POSE DETECTION 2017-2019

images

- **Associative Embedding:** End-to-End Learning for Joint Detection and Grouping [Newell et. Al, NIPS2017]
- **RMPE:** Regional Multi-person Pose Estimation [Fang et. Al, ICCV2018]
- ***Realtime Multi-Person 2D Pose Estimation*** *using Part Affinity Fields* [Cao et. Al, CVPR2017]
- **Cascaded Pyramid Network** for Multi-Person Pose Estimation [Chen et. Al, CVPR2018]

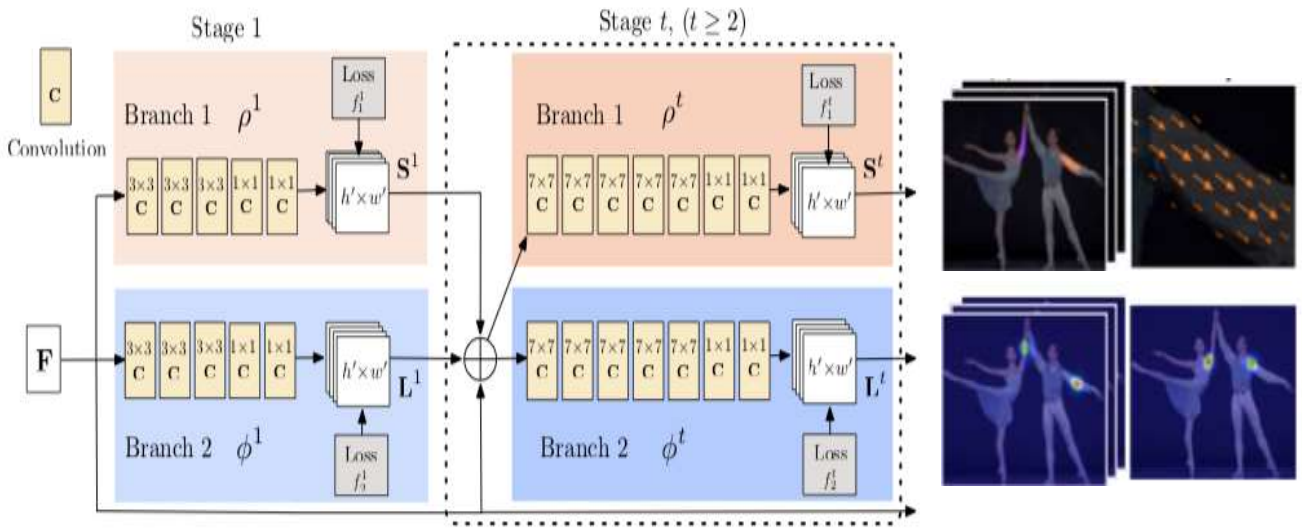
video

- **ArtTrack:** Articulated Multi-person Tracking in the Wild [Insafutdinov et. Al. CVPR2017]
- **PoseTrack:** Joint Multi-Person Pose Estimation and Tracking [Iqbal et al. CVPR2017, ECCV2018]
- Simple, efficient and effective **keypoint tracking** [Girdhar et. Al, ICCVws2017]
- **Towards Multi-Person Pose Tracking:** Bottom-up and Top-down Methods [Jin et. Al ICCVws2017]
- **OpenPose:** Real-time multi-person keypoint detection library for body, face, and hands estimation [S. Wei et al CVPR 2016, → CVPR 2017, CVPR 2018],

RELATED WORKS: OPENPOSE CMU

Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields[CVPR2017]

- Two-branch multi-stage CNN inspired by CMP for 2D multi person-pose estimation that jointly learns and parts association
- Introduces **a novel bottom-up approach for joints association via Part Affinity Fields (PAFs)**, a set of 2D vector fields that encode the location and orientation of limbs in the image domain.



[S. Wei V. Ramakrishna T.Kanade and Y. Sheikh Convolutional pose machines CVPR 2017]

WHERE ARE THE MAIN CHALLENGES?

1. Reliable **DL architectures** for pose detection
2. Coping with **occlusions** (and body self-occlusions), p
3. A global spatio-temporal association for

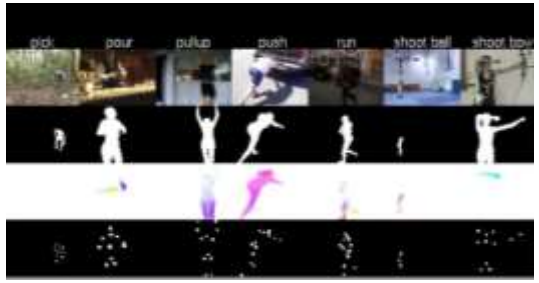
   detection, tracking and then prediction

1. **A Large and General Annotated dataset**

RECENT ANNOTATED DATASETS FOR POSE DETECTION AND/OR TRACKING

Dataset	#Clips	#Frames	#PpF	3D	Occlusion	Tracking	Pose	Type
Penn Action	2,326	159,633	1				X	Sports
JHMDB	5,100	31838	1				X	Diverse
YouTube Pose	50	5,000	1				X	Diverse
Video Pose 2.0	44	1,286	1				X	Diverse
MOT-16	14	11,235	6-51		X	X		Urban

Penn Action



JHMDB

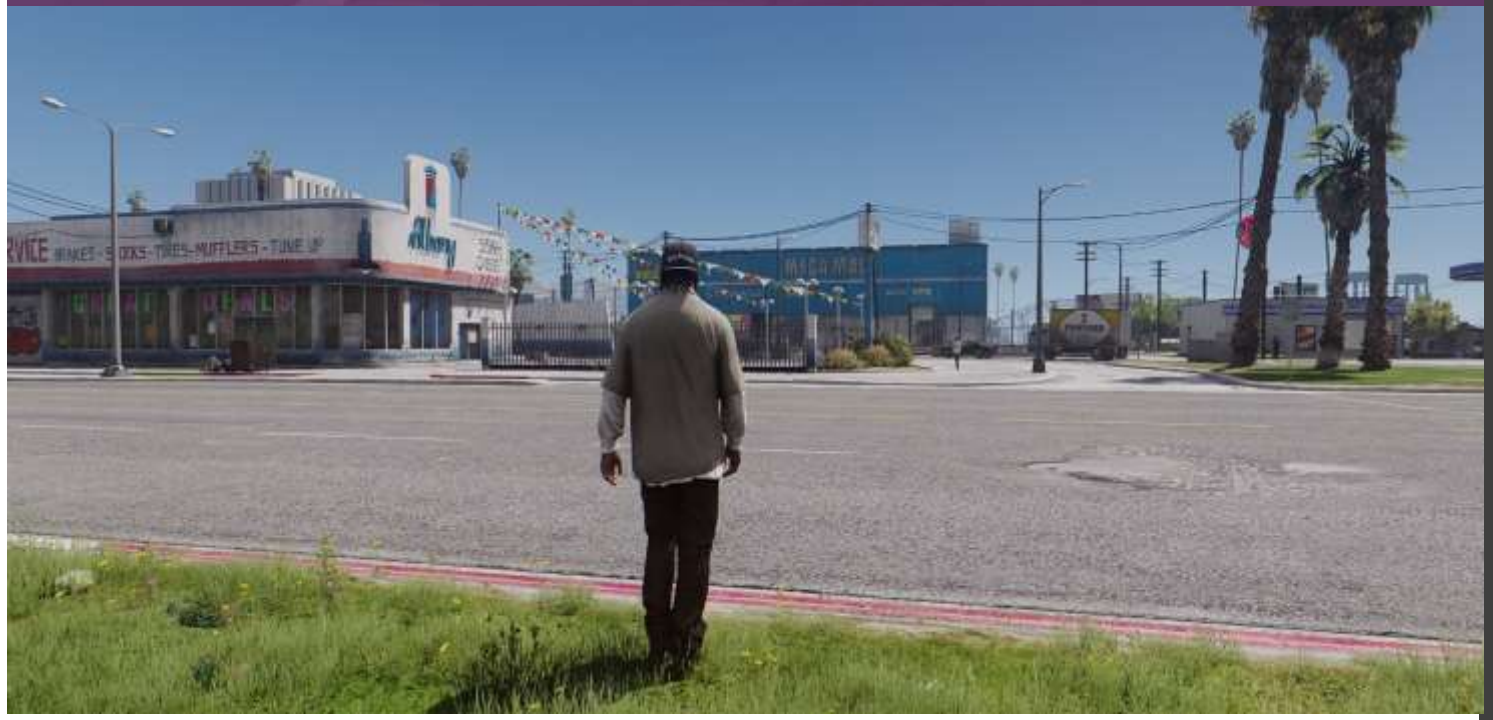
YouTube Pose

YouTube Pose 2.0

Whenever data annotation is extremely expensive…

Synthetic data become

THE SOLUTION




TED
NIKE

GTA V SCRIPTHOOK

AImage Lab

grand theft auto FIVE

+

ScriptHook library

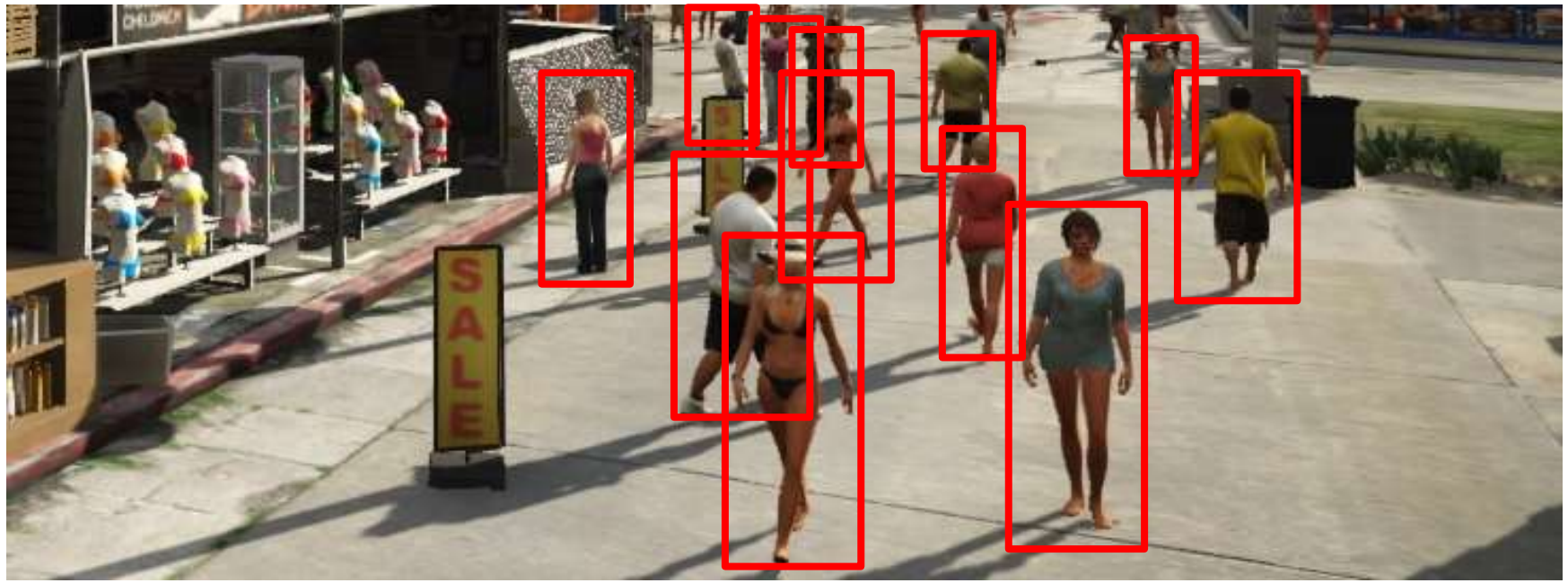
- Photorealistic
- Plausible dynamics
- Lifelike entity AI

- Access to native GTA functions
- Customizable
- Extract all the information available to the game engine

+ Phd male students

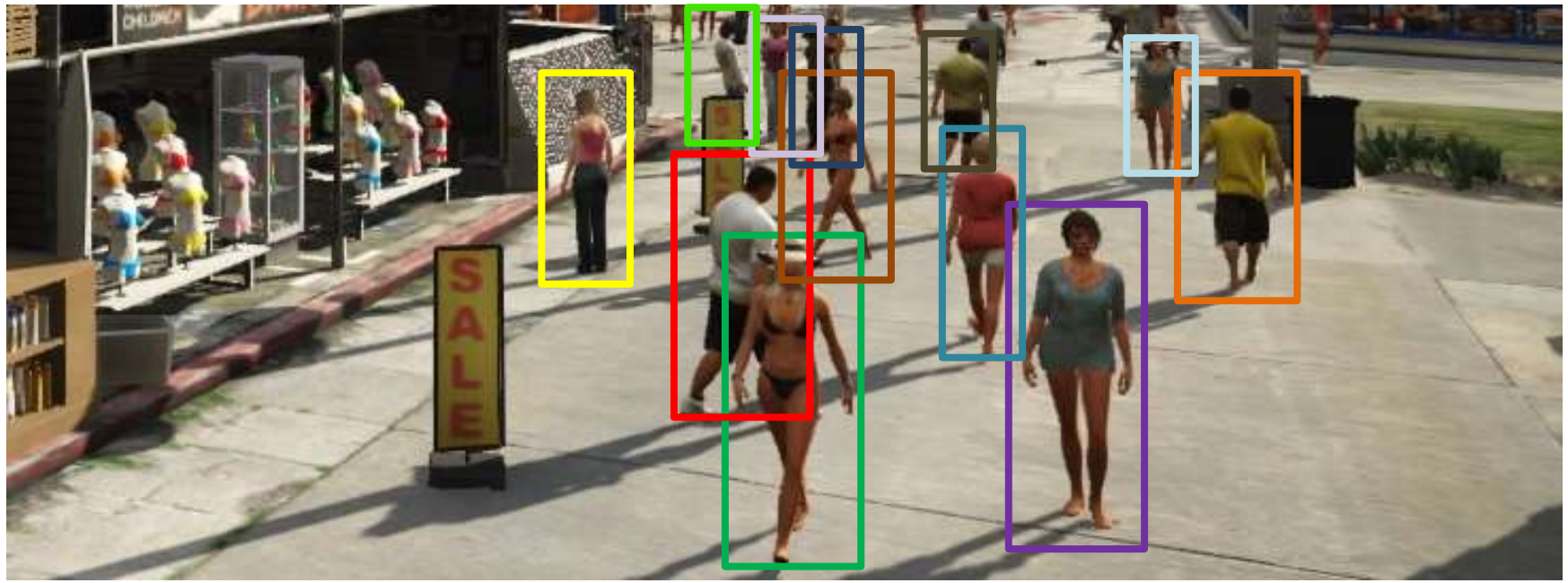
EXTRACTING DATA FROM GAME ENGINE

EXTRACTING DATA FROM GAME ENGINE



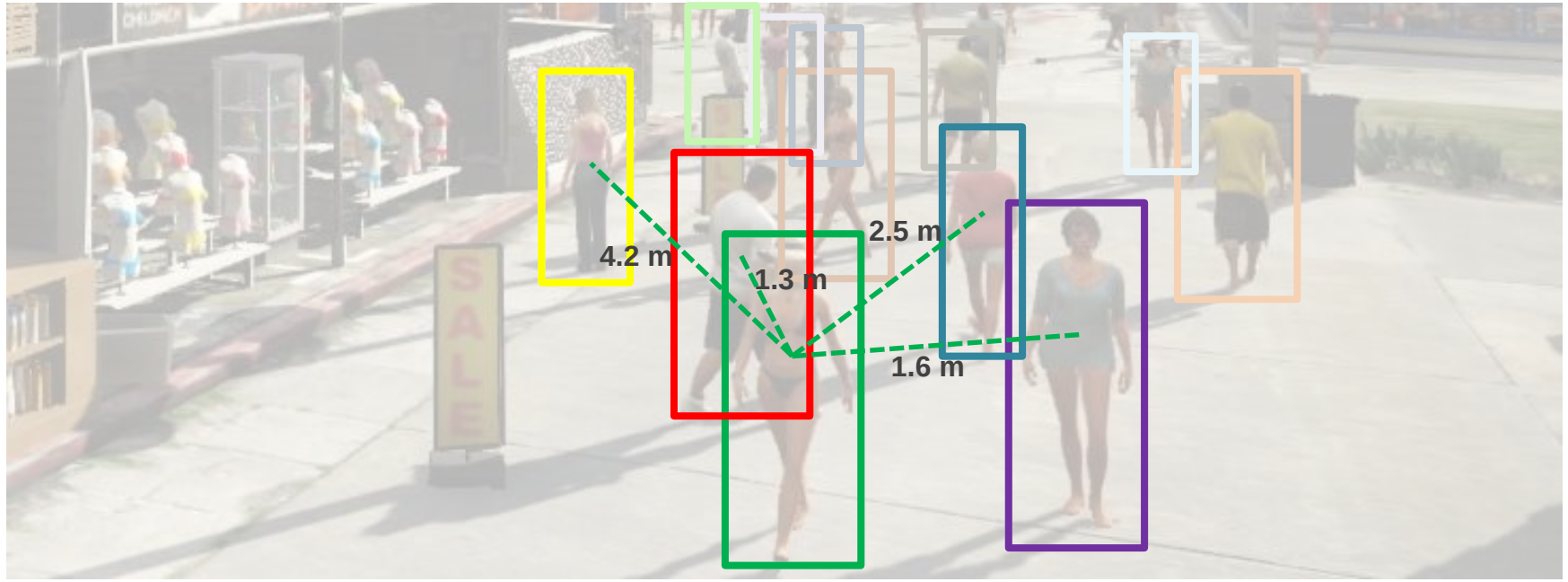
- Bounding boxes

EXTRACTING DATA FROM GAME ENGINE



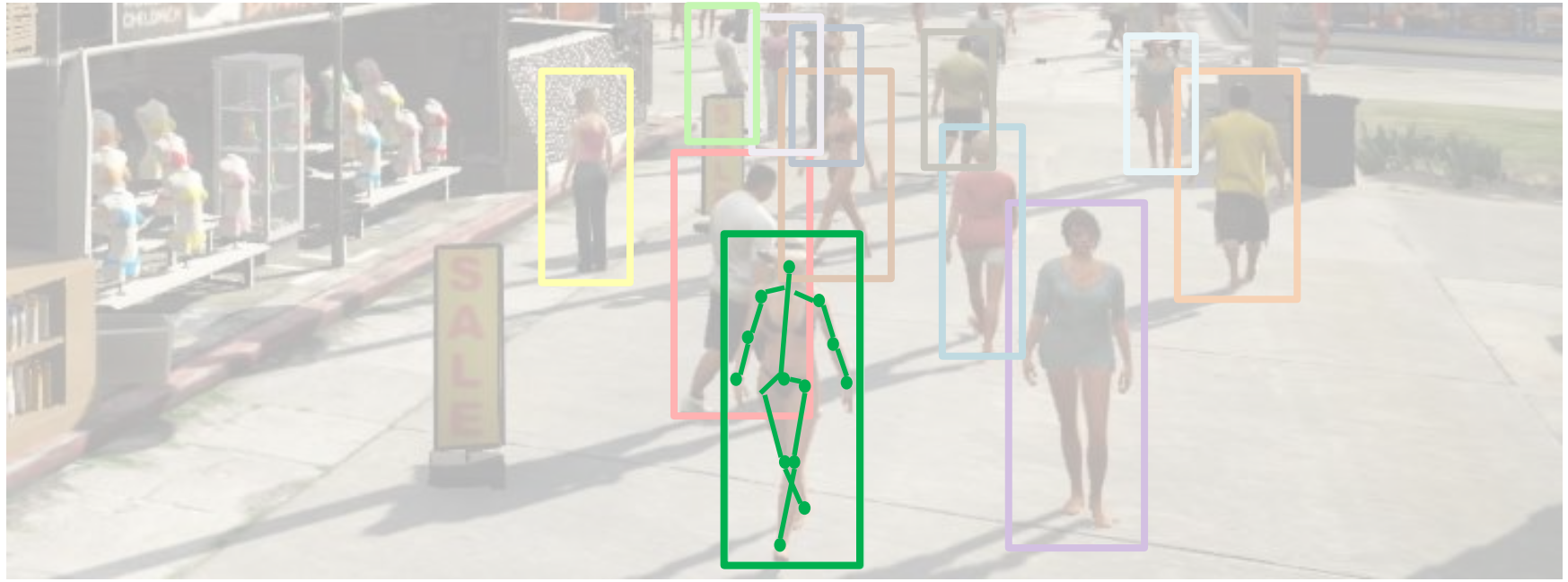
- Bounding boxes
- **Instances**

EXTRACTING DATA FROM GAME ENGINE



- Bounding boxes
- Instances
- **Reciprocal distances**

EXTRACTING DATA FROM GAME ENGINE



- Bounding boxes
- Instances
- Reciprocal distances
- **Joints coordinates**

THE NEW JTA DATASET BY AIMAGELAB UNIMORE 2018 [ECCV18]

About the Dataset

JTA (Joint Track Auto) is a huge dataset for pedestrian pose estimation and tracking in urban scenarios created by exploiting the highly photorealistic video game *Grand Theft Auto V* developed by *Rockstar North*. We collected a set of 512 full-HD videos (256 for training and 256 for testing), 30 seconds long, recorded at 30 fps.

THE NEW JTA DATASET

	JTA	Posetrack
Data type	Synthetic	Real
#Clips	512	514
#Frames	>460,000	>22,000
#Poses	9,836,194	153,615
#PpF	0-60	1-13
3D	X	
Occlusion	X	
Tracking	X	X
Pose	X	X
Type	Urban	Diverse
Pose variation	Low	High

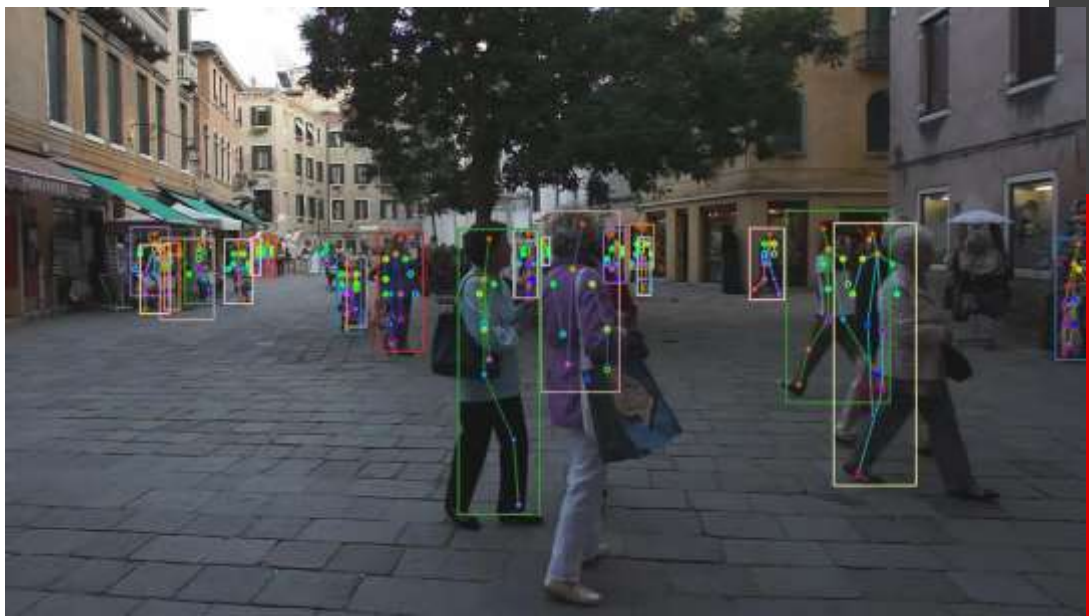
The J-GTA Dataset is available at
www.aimagelab.unimore.it/dataset
Contact Matteo.Fabbri@unimore.it


iPod 9X
Escalone

NOW A DL-BASED PROPOSAL

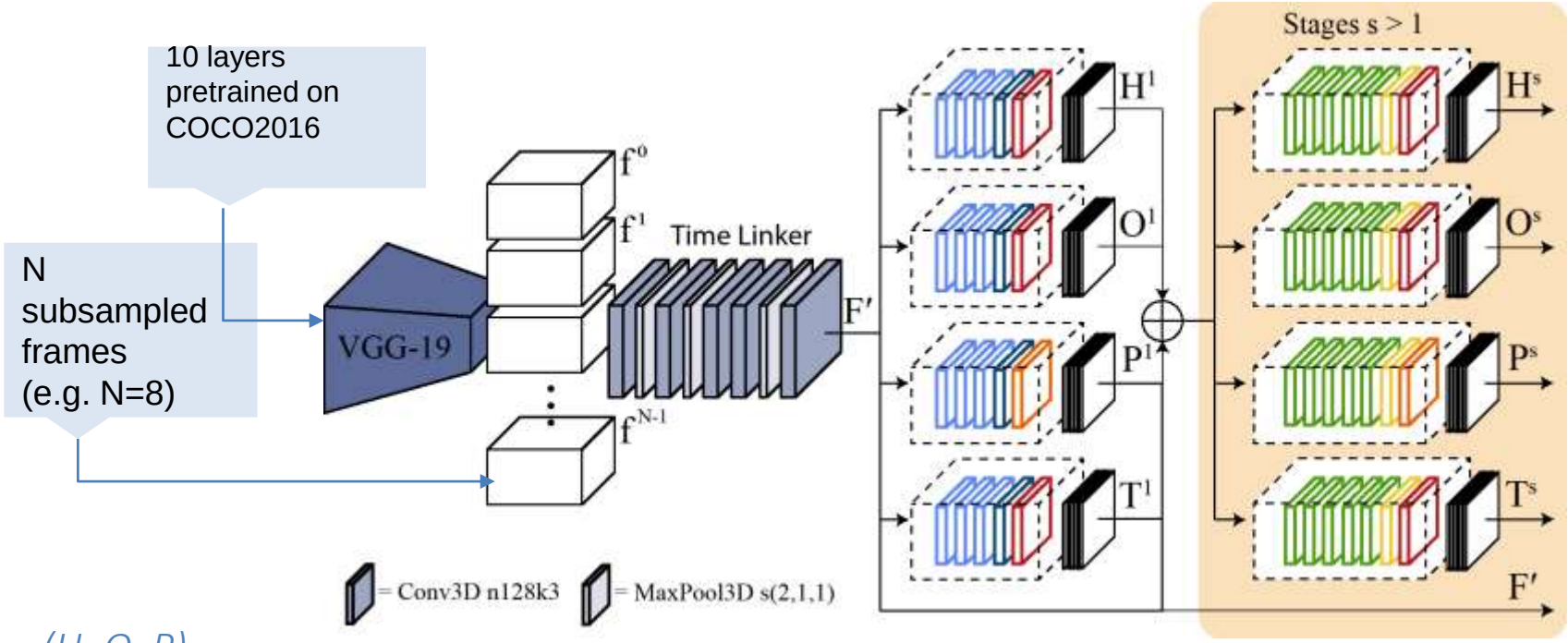
1. **Reliable DL architecture for pose detection**
2. **Coping with occlusions (and body self-occlusions)**
3. **Aglobal spatio-temporal association for tracking**
4. Large and General Annotated dataset

Starting from CPM by CMU



THOPA-NET FOR POSE ESTIMATION AND TRACKING

Detection by joints **H**eatmaps, **O**cclusions, **P**art Affinity Fields (PAFs) and **Temporal** Affinity Fields (TAF)
Tracking is then a NN- association by minimizing TAF scores in short tem.



(H, O, P)

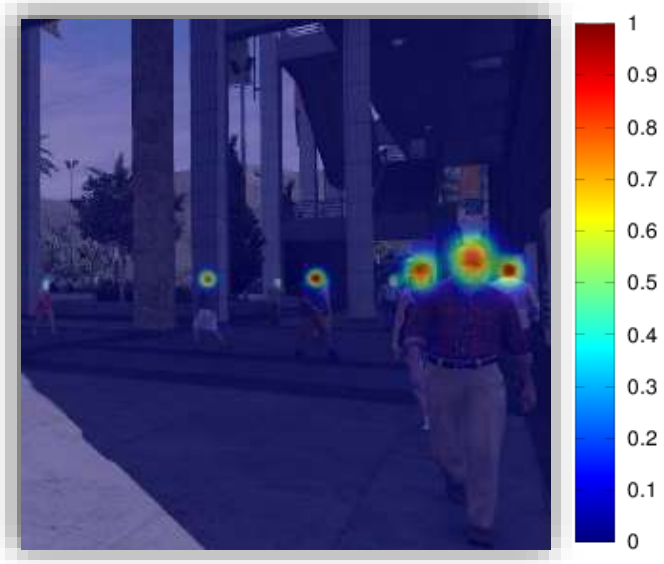
$$l_X^s = \sum_i \sum_{x=1}^{w'} \sum_{y=1}^{h'} M(x,y) \odot (X_i^s(x,y) - X_i^*(x,y))^2 \qquad L = \sum_{s=1}^{S} (l_H^s + l_O^s + l_P^s + l_T^s)$$

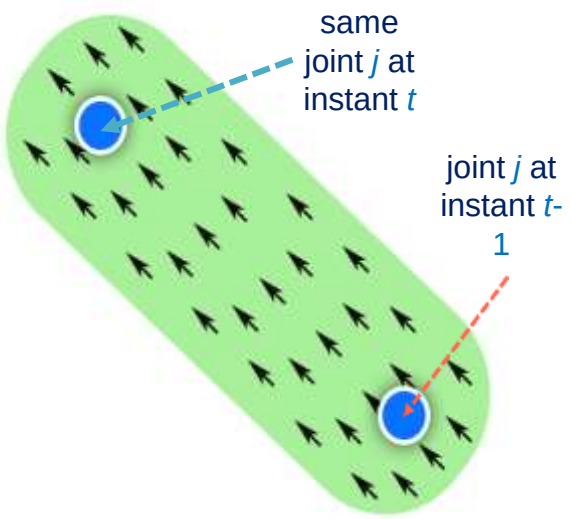
[M.Fabbri, F.Lanzi, S.Calderara, A.Palazzi, R.Vezzani, R.Cucchiara Learning to Detect and Track Visible and Occluded Body Joints in a Virtual World ECCV 2018]

HEATMAPS, OCCLUSIONS, PAFS AND TAFS

Heatmaps model the part locations and occlusions as Gaussian peaks in the map;

One heatmap for each type of joint (eg. “nose”, “neck”, “left sholder”, …) with Scale awareness without the need of multi-scale branches or pyramid inputs



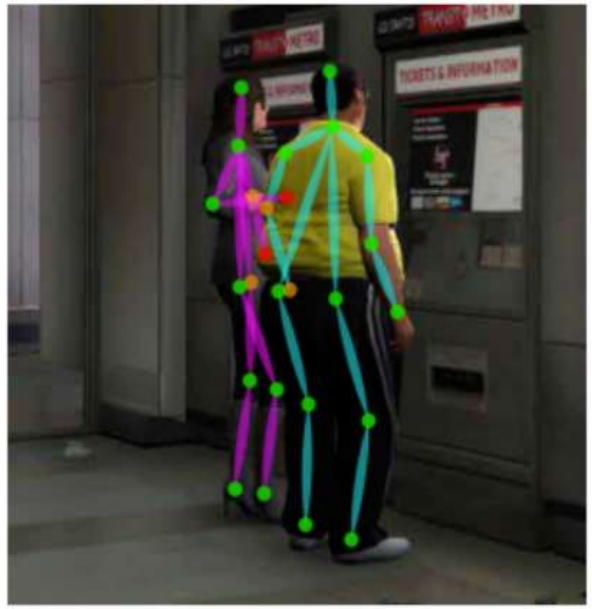


The PAFs (Part Affinity Fields) are 2D versor encoding he direction that points from one joint to the other.

The TAF (Temporal Affinity Field) links corresponding joints of the same person in consecutive frames (for an unknown number of people).

Spatio-temporal association of joints to person is given by an optimization takin into account PAFs, TAFS and human anatomy constraints

Automatically annotated joints, occluded joints, connections (as PAF) and temporal connections (TAF)



Detected joints, occluded joints, connections (as PAF) and temporal connections (TAF) (color means direction)

OCCLUSION HALLUCINATION (FROM ECCV2018)

The network is able to **hallucinate** the position of not visible joints.

Table 2. Detection results on JTA Dataset

	Joints	Detection		
	Mean Average Prec.	Precision	Recall	F1 Score
Single Image no occ	50.9	81.5	64.1	71.6
Single Image + occ	56.3	87.9	71.8	78.4
Complete	**59.3**	**92.1**	**77.4**	**83.9**
[7]	50.1	86.3	55.8	69.5

[7] [Cao et al. CVPR2017]

RESULTS – DETECTION AND TRACKING BY POSE AT ECCV2018

AImage Lab

Table 3. Results on JTA dataset

	MOTA	IDF1	MT	ML	FP	FN	IDs	FRAG
[3] + our det	57.4	57.3	45.3	21.7	40096	103831	15236	15569
[3] + DPM det	31.5	27.6	25.3	41.7	80096	170662	10575	19069
THOPA-net	**59.3**	**63.2**	**48.1**	**19.4**	**40096**	**103662**	**10214**	**15211**

Experiments of JTA THOPAnet in short term tacking (1 sec)
w.r.t the detection and a standard nn tracking,
and w.r.t a classical people detector+ nntracking,

Experiments of real data
With THOPAnet and final refinement

Table 4. Results on MOT-16 benchmark ranked by MOTA score

	MOTA	IDF1	MT	ML	FP	FN	IDs	FRAG
[25]	**66.1**	**65.1**	**34.0**	20.8	5061	**55914**	805	3093
[26]	61.4	62.2	32.8	**18.2**	12852	56668	781	2008
THOPA-net	56.0	29.2	25.2	27.9	9182	67059	4064	5557
[27]	47.2	46.3	14.0	41.6	**2681**	92856	774	1675
[28]	46.0	50.0	14.6	43.6	6895	91117	**473**	**1422**
[29]	43.9	45.1	10.7	44.4	6450	95175	676	1795
[30]	38.8	42.4	7.9	49.1	8114	102452	965	1657

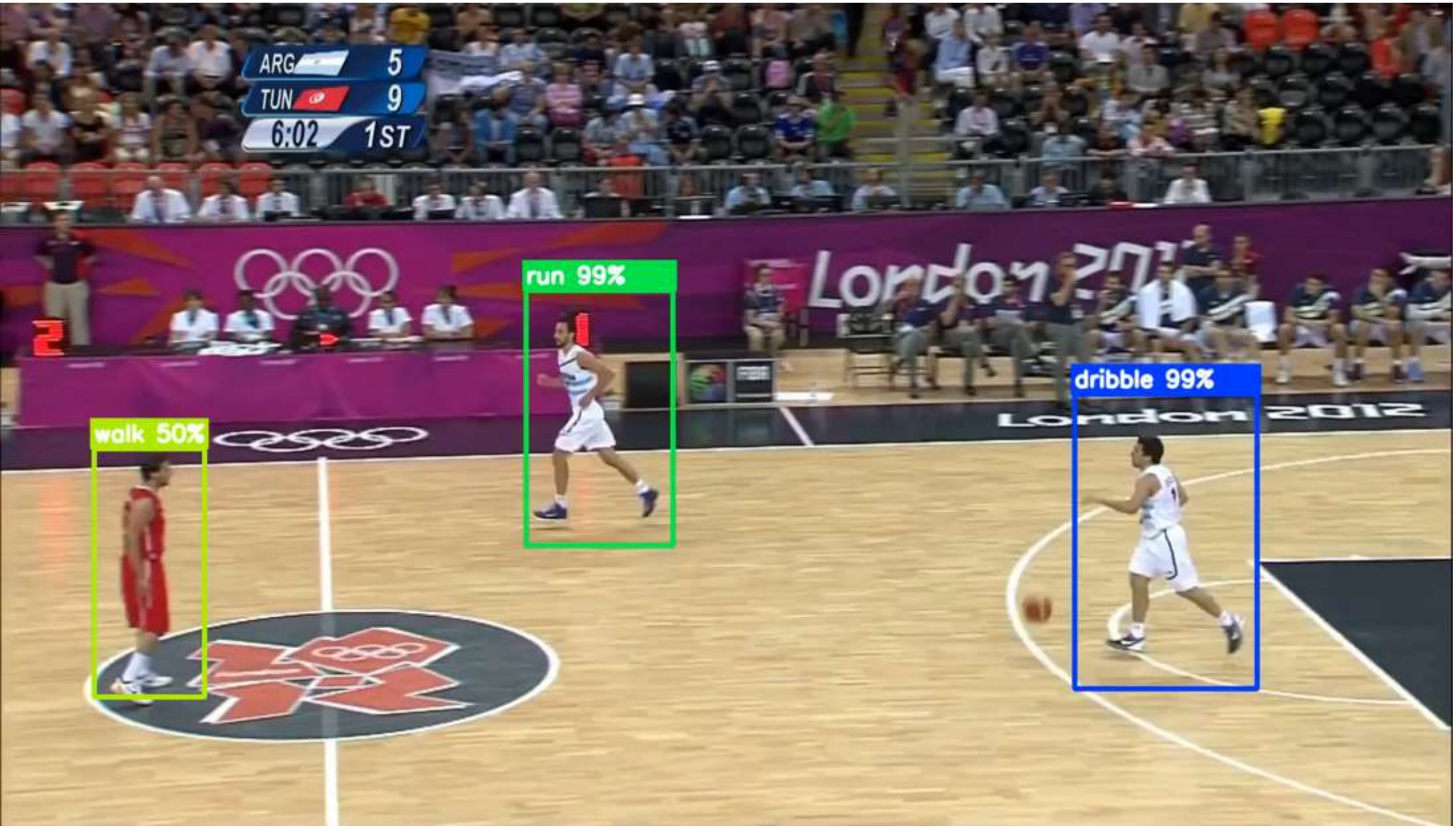
[25] Yu et al ECCV Ws2016; [26] Woike et al ICIAP2017;
[27] Sadeghian et al ICCV2017Tracking the untrackable
[28]Chu et al ICCV 2017 [29]Bae et al T-PAMI 2018
[30] Sanchez Mattilla et al ECCV WS 2016

[3] Solera et. al AVSSS BPA 2015 Towards the evaluation of r
eproducible robustness in tracking-by-detection.

RESULTS ON TRACKING PEOPLE (JTA)

RESULTS ON
TRACKING PEOPLE (DR(EYE)VE)

JUMP PROJECT – ACTION RECOGNITION



AImage Lab


AImage Lab

AImage Lab

THEN TOWARD PREDICTION (DRAFT)

Actual, predicted joints , reconstruction (with a specific U-NET)

IN CONCLUSION…WHAT WE LEARNED, RECENTLY.

AImage Lab

Computer vision, after about 30 years, with DL is still FASHINATING and COMPLEX

Each result requires TIME, PEOPLE, CPU/GPU, and STUDY

Data are necessary. Big Data are not

God gives bread to those who have no teeth and viceversa… invent solutions..

Vision (also starting from dataset) can be used around…
Robotics and automotive are some examples

Enjoy with the others. Meet researchers, work together and organize each event as the best scientific party you can…

SEE YOU AT **ICPR 2020 MILANO, ITALY**

(AND AT ECCV 2022 TELAVIV, ISRAEL)

AImage Lab

UNIMORE
UNIVERSITÀ DEGLI STUDI DI MODENA E REGGIO EMILIA

btw..OPEN POSITIONS at Aimagelab at University of Modena and Reggio Emilia, ITALY .
Send my your CV!

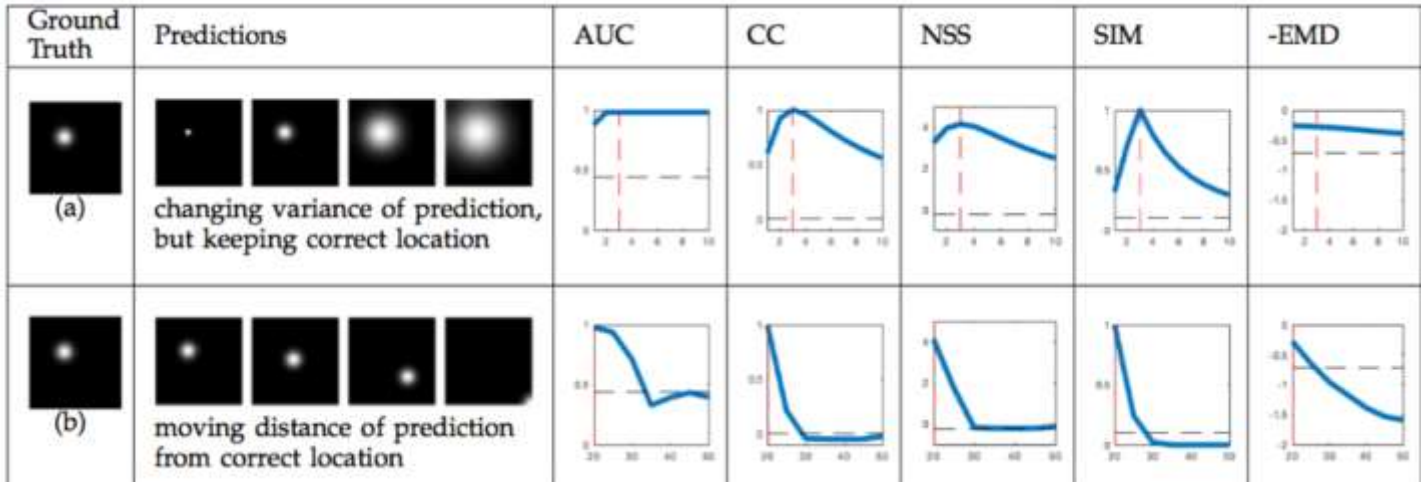
Thanks

ADDENDUM

MEASURE IN SALIENCY

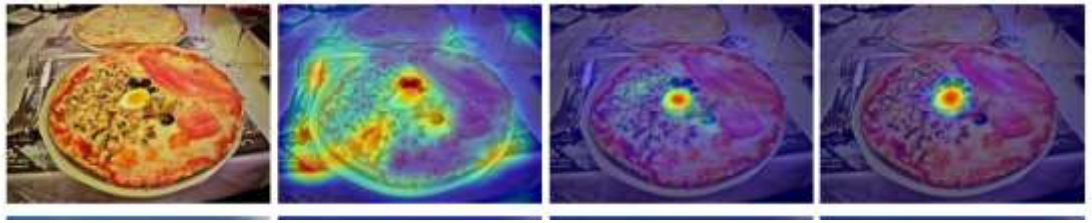
Comparison as in [Bylinsky et al. ArXiv 2014]

- Similarity (SIM)
- Correlation Coefficient (CC)
- Area Under the ROC Curve (AUC)
- shuffled version of AUC (sAUC)
- NSS, Normalized Scanpath Saliency (NSS)
- -Earth-Mover Distance (EMD).



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

Image, Itti, ML-NET, GT



LEGENDA: NO COMPLETE CONSENSUS IN METRICS

BLEU (**bilingual evaluation understudy**) is an algorithm for evaluating the quality of text which has been machine-translated from one natural language to another. BLEU uses a modified form of precision to compare a candidate translation against multiple reference translations.

METEOR (**Metric for Evaluation of Translation with Explicit ORdering**) is a metric for the evaluation of machine translation output. The metric is based on the harmonic mean of unigram precision and recall, with recall weighted higher than precision.

ROUGE, or **Recall-Oriented Understudy for Gisting Evaluation**,[1] is a set of metrics and a software package used for evaluating automatic summarization and machine translationsoftware in natural language processing. ROUGE-N: N-gram[3] based co-occurrence statistics

CIDEr: Consensus-based Image Description Evaluation (from CVPR2016)