Undefined 1 (2014) 1–5
IOS Press

Attentive Models in Vision: Computing Saliency Maps in the Deep Learning Era

Marcella Cornia [a,*], Davide Abati [a], Lorenzo Baraldi [a], Andrea Palazzi [a], Simone Calderara [a] and Rita Cucchiara [a]

[a] *Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia*
E-mail: {name.surname}@unimore.it

Abstract. Estimating the focus of attention of a person looking at an image or a video is a crucial step which can enhance many vision-based inference mechanisms: image segmentation and annotation, video captioning, autonomous driving are some examples. The early stages of the attentive behavior are typically bottom-up; reproducing the same mechanism means to find the saliency embodied in the images, *i.e.* which parts of an image pop out of a visual scene. This process has been studied for decades both in neuroscience and in terms of computational models for reproducing the human cortical process. In the last few years, early models have been replaced by deep learning architectures, that outperform any early approach compared against public datasets. In this paper, we discuss the effectiveness of convolutional neural networks (CNNs) models in saliency prediction. We present a set of Deep Learning architectures developed by us, which can combine both bottom-up cues and higher-level semantics, and extract spatio-temporal features by means of 3D convolutions to model task-driven attentive behaviors. We will show how these deep networks closely recall the early saliency models, although improved with the semantics learned from the human ground-truth. Eventually, we will present a use-case in which saliency prediction is used to improve the automatic description of images.

Keywords: Saliency, Human Attention, Neuroscience, Vision, Deep Learning

1. Introduction

When humans look around the world, observe an image or watch at a video sequence, attentive mechanisms drive their gazes towards salient regions. These have been studied in psychology and neuroscience for decades [22], and it is well assessed that they are mainly bottom-up in the early stages, although influenced by some contextual cues, and guided by the salient points in the scene which are scanned very quickly by the eyes (in about 25-50 ms per item). If the person is performing a task (e.g. when driving a car), top-down attentive process arise; they are slower (around 200 ms) and rely on the learned semantics of the scene. In general, the control of attention combines stimuli processed in different cortical areas to mix spatial localization and recognition, and integrates data-driven pop-outs and some learned semantics. It also has a temporal evolution, since some mechanisms such as the inhibition of return and the control of eye movements allow humans to refine attention during time.

Reproducing the same attentive process in computer vision is still an open problem. In the case of a static image, researchers have shown that salient regions can be identified by considering discontinuities in low-level visual features such as color, texture and contrast, and in high-level cues as well, like faces, text, and the horizon. When watching a video sequence, instead, static visual features have lower importance while motion gains a crucial role, motivating the need of different solutions for static images and video. In both scenarios, computational models capable of identifying salient regions can enhance many vision-based inference mechanisms, ranging from image captioning [11,12] and automatic cropping [15] to video compression [17].

*Corresponding author. E-mail: marcella.cornia@unimore.it.

Since the seminal research of Kock, Ulman and Itti [23,30], traditional prediction models have followed biological evidences using low-level features and semantic concepts [18,28]. With the advent of Deep Learning, researchers have developed data-driven architectures capable of overcoming many of the limitations of previous hand-crafted models. This success is not only due to the amount of data these architectures are trained on: deep architecture are particularly suitable for this task as they recall neural biological models. Still, it is surprising to see how much today's models share with those early works.

Motivated by these considerations, in this paper we extend the work in [9] and present an overview of different solutions that we have developed for saliency prediction on images and video. We conduct a comparison between these models and the early models for computing saliency maps, and investigate similarities and differences. We will show that today's models, based on Convolutional Neural Networks (CNNs) share many of the principles of early models, while at the same time solving many of their drawbacks. Different convolutional architectures will be presented, to deal with features extracted at multiple levels, and with the time dimension of video in the case of driver attention estimation. Those will be subsequently described in their implementation details and evaluated on standard datasets both in terms of accuracy, and space and time complexity, extending the evaluations already presented in [9]. As an additional contribution, we will also show how saliency can be beneficial in the case of automatic image description, and conclude by presenting an interesting use-case of the proposed architectures.

2. Related Work

In this section we review the literature related to saliency prediction in images and video, starting from traditional methods which combine hand-crafted features and going through recent models based on deep learning techniques.

2.1. Saliency prediction on images

Early works on saliency prediction on images were based on the Feature Integration Theory proposed by Treisman *et al.* [45] in the eighties. Itti *et al.* [23], then, proposed the first saliency computational model: this work, inspired by Koch and Ullman [30], computed a set of individual topographical maps representing low-level cues such as color, intensity and orientation and combined them into a global saliency map. The saliency map is a scalar map, as large as the image, where each point represents the visual saliency, irrespective of the feature dimension that makes the location salient. The *locus* of highest activity in the saliency map is the most probable eye fixation point or is the point where the focus of attention should be localized. After this work, a large variety of methods explored the same idea of combining complementary low-level features [5,18] and often included additional center-surround cues [53]. Other methods enriched predictions exploiting semantic classifiers for detecting higher level concepts such as faces, people, cars and horizons [28].

In the last few years, thanks to the large spread of deep learning techniques, the saliency prediction task has achieved a considerable improvement [13,14, 24,32,40]. First attempts of predicting saliency with convolutional networks mainly suffered from the absence of fine-tuning of network parameters over a saliency prediction dataset and from the lack of sufficient amount of data to train a deep saliency architecture [34,46]. The publication of the large-scale attention dataset SALICON [26] has contributed to a big progress of deep saliency prediction models and several new architectures have been proposed.

Huang *et al.* [21] introduced a deep neural network applied at two different image scales trained by using some evaluation metrics specific for the saliency prediction task as loss functions. Kruthiventi *et al.* [32] proposed a fully convolutional network called *DeepFix* that captures features at multiple scales and takes global context into account through the use of large receptive fields. Moreover, *DeepFix* takes advantage of predefined priors to predict its saliency maps, getting a large improvement from using them. Pan *et al.* [40] instead presented a shallow and a deep convnet where the first is trained from scratch while some layers of the second are initialized with the parameters of a standard convolutional network. Jetley *et al.* [24] introduced a saliency model that formulates a map as a generalized Bernoulli distribution and they used these maps to train a CNN trying different loss functions. Finally, Kruthiventi *et al.* [33] presented an unified framework that is capable of predicting eye fixations and segmenting salient objects on input images.

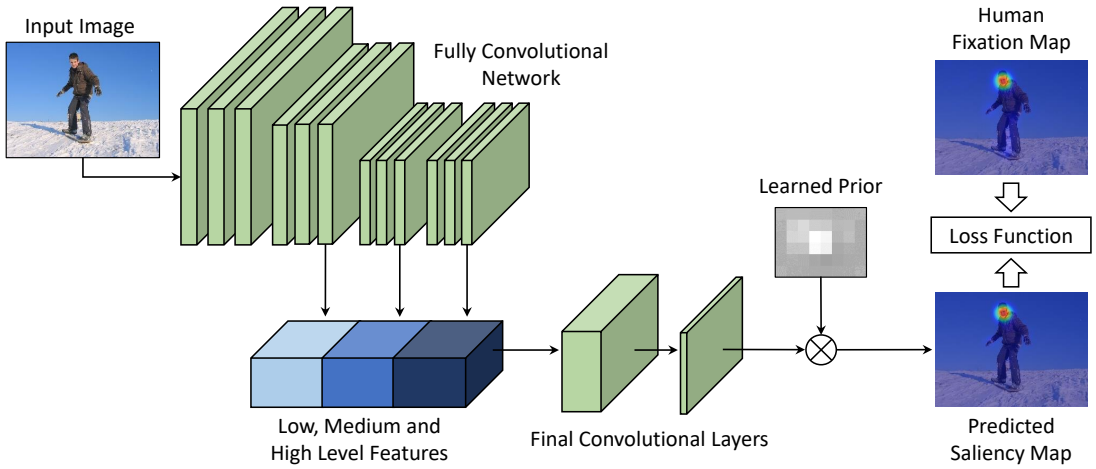

Fig. 1. Overview of our Multi-Level Network (ML-Net) [10].

2.2. *Saliency prediction in video*

When considering video inputs, saliency estimation is quite different with respect to still images. Indeed, motion is a key factor that strongly attracts human attention. Accordingly, some video saliency models pair bottom-up feature extraction with a further motion estimation step, that can be performed either by means of optical flow [54] or feature tracking [52]. Somehow differently, some models have been proposed to force the coherence of bottom-up features across time. In this setting, previous works address feature extraction both in a supervised [37] and unsupervised [47] fashion, whereas temporal smoothness of output maps can be achieved through optical flow motion cues [54] or explicitly conditioning the current map on information from previous frames [42].

As previously discussed for the image saliency setting, the representation capability of deep learning architectures, along with large labeled datasets, can yield better results. However, deep video saliency models still lack, being the work in [4] the only meaningful effort that can be found in the current literature. Such model leverages a recurrent architecture iteratively updating its hidden state over time, and emitting the saliency map at each step by means of a Gaussian Mixture Model.

3. Saliency Prediction with Deep Learning Architectures

In this section we discuss different deep learning architectures for saliency prediction on images and video. We will introduce a convolutional model for images, which incorporates low and high level visual features, and which, conceptually, extends the seminal work by Itti and Koch [23] by means of a modern neural network. A discussion on the similarities and differences between these two models will follow. We will then present an architecture for task-driven attention prediction in videos, and show how this particular domain differs from that of images in the case of driver attention prediction.

3.1. *Incorporating low-level and high-level cues in a Multi-Level Network*

In [10], we proposed a Deep Multi-Level Network (ML-Net[1]) for saliency prediction. In contrast to previous proposals, in which saliency maps were predicted from a non-linear combination of features coming from the last convolutional layer of a CNN, we effectively combined feature maps coming from three different levels of a fully convolutional network thus taking into account low, medium and high level cues. Moreover, to model the center bias present in human eye fixations, we incorporated a learned prior map by applying it to the predicted saliency map. Fig. 1 shows the overall architecture of our ML-Net model.

More in details, the first component of our architecture is a CNN based on a standard convolutional network originally designed for image classification and then employed in several other computer vision tasks.

[1]Project page is available at: `https://github.com/marcellacornia/mlnet`

This network, namely VGG-16 [43], is composed of 13 convolutional layers, divided in 5 different blocks, and 3 fully connected layers. Since we aimed at producing a 2-dimensional map (*i.e.* the predicted saliency map), we removed the fully connected layers thus obtaining a fully convolutional architecture. Several other deep saliency models [21,24,40] employ the VGG-16 as starting point for their architectures and almost all of them use feature maps coming from the last convolutional layer of the VGG-16 network. Each of them then proposes specific saliency components or different training strategies. In contrast to this dominant approach, the second component of our model takes feature maps coming from three different levels of the network: the output of the third, fourth and fifth convolutional blocks. Our model then combines these feature maps through two specific convolutional layers that merge low, medium and high level features and produce a temporary saliency map. Finally, we incorporate an important property of human gazes: when an observer looks at an image its gaze is biased towards the center of the scene. To this end, the last component of our architecture is designed to model the center bias through a learned prior map which is applied to the temporary saliency map with an element-wise multiplication, thus giving more importance to the center of the image.

Deep learning architectures are trained by minimizing a given loss function that, in the case of saliency prediction, aims at effectively approaching the predicted saliency map to the ground-truth one obtained from human fixation points. Previous deep saliency models were trained with different strategies by using a saliency evaluation metric as loss function or, more commonly, a squared error loss (such as the euclidean loss). We instead designed a specific loss function inspired by three different objectives: first, predicted saliency maps should be similar to ground-truth ones, therefore a squared error loss can be a reasonable choice. Secondly, predictions should be invariant to their maximum, and there is no point in forcing the network to produce values in a given numerical range, so we normalize predictions by their maximum. Third, the loss should give the same importance to high and low ground-truth values, even though the majority of ground-truth pixels are close to zero. For this reason, the deviation between the predicted values $\phi(\mathbf{x}_i)$ and the ground-truth values $\mathbf{y}_i$ is weighted by a linear function $\alpha - \mathbf{y}_i$ which tends to give more importance to pixels with high ground-truth fixation probability. The overall loss function is thus

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

where $\mathbf{w}$ are the network parameters and N the number of samples inside the mini-batch.

3.2. Deep Learning architectures vs. the Itti and Koch's model

The first computational model for saliency prediction, and probably the most famous, was presented in a seminal paper by Itti and Koch [23]. It proposed to extract multi-scale low-level features from the input image which were linearly combined and then processed by a dynamic neural network with a winner-takes-all strategy to select attended locations in decreasing order of saliency. As we have shown in the previous section, nowadays saliency prediction is generally tackled via CNN architectures, therefore giving more importance to learning than to hand engineering of features. However, today's models share a lot with that influential work.

The model in [23] extracted three kinds of features from input images: color (as a linear combination of raw pixels in color channels), intensity (again, computed as a linear combination of color channels), and orientation, by means of oriented Gabor pyramids [16]. It should be noted that all these features can be easily extracted by a single convolutional layer, and, indeed, visualization and inversion techniques [51] showed that filters learned in the early stages of a CNN roughly extract color and gradient features. Also, the linear combinations of color channels in [23] can be computed via a single convolutional layer with channel-wise uniform weights or with a 1×1 kernel.

One detail, however, is missing in current convolutional architectures: authors of [23] extracted the same features at multiple scales, and then validated them by performing central differences between adjacent scales. In a CNN, instead, features are always computed at a single scale, even though the overall architecture extracts (different) features at different scales thanks to pooling stages. Of course the multi-scale validation of features was also motivated by the need of extracting robust features, something which comes almost for free in modern architectures. Moreover, many state of the art CNN models are multi-scale by construction, feeding a pyramid of images to the same

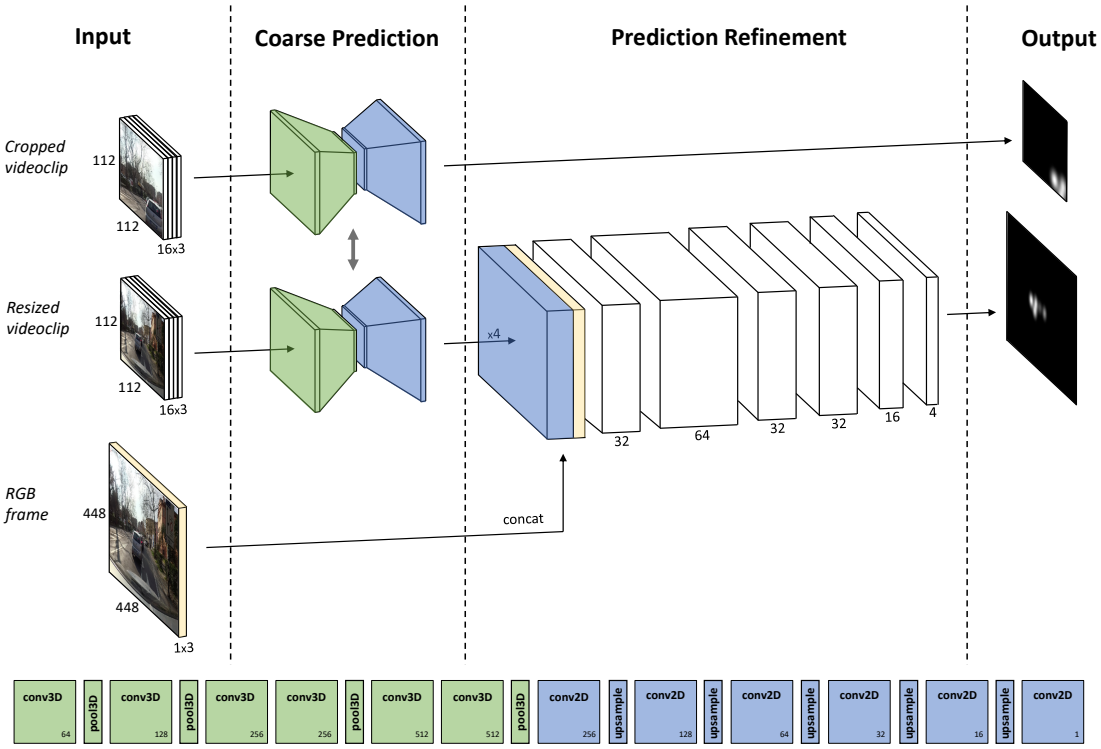

Fig. 2. Illustration of the COARSE+FINE model depicting the both streams guiding the optimization during training. Please note that in test stage the cropped stream is not used. At the bottom, the architecture of the COARSE module is illustrated.

convolutional stack. Even in our model, we combine different features extracted at different scales to form the final prediction, instead of taking only those produced by the last layer.

Conversely, the most evident characteristic that the Itti and Koch model misses with respect to today's architectures is the ability to extract higher level features, and to detect objects and part of objects. This is achieved, in today's networks, by increasing the depth of the network (e.g. 152 layers in the ResNet model [19]). This, given the big performance gap, clearly highlights the need of high-level features for saliency prediction.

3.3. Estimating task-driven attention in videos

In [39] we described a model devised for predicting driver's focus of attention on the DR(eye)VE dataset [1], capable of replicating human attentional behavior while driving. The need for a different model tailored for this specific context is twofold: first, as anticipated, objects motion in videos tends to capture human attention. Moreover, fixations recorded during the dataset acquisition in [1] are strongly related to the driving activity, and call for a task-driven model and training procedure.

Motivated by the insight that a small temporal window holds sufficient information meaningful for the task of driving, our model captures short-term correlations by means of 3D convolutions [25]. This operation resembles commonly adopted 2D convolutions, with a major difference: the input tensor explicitly encodes time in an additional axis, along which convolutional kernels stride (while still striding along spatial positions). More formally, the j-th feature map in the i-th 3D convolutional layer at position (x, y) at time t is computed as:

$$v_{i,j}^{x,y,t} = b_{i,j} + \sum_{m} \sum_{p=0}^{P_{i-1}} \sum_{q=0}^{Q_{i-1}} \sum_{r=0}^{R_{i-1}} w_{i,j,m}^{p,q,r} v_{i-1,m}^{x+p,y+q,t+r} \tag{2}$$

where m indexes different input feature maps, $w_{i,j,m}^{p,q,r}$ is the value at the position (p, q) at time r of the kernel connected to the m-th feature map, and P_i, Q_i and R_i are the dimensions of the kernel along width, height

and temporal axis respectively; $b_{i,j}$ is the bias from layer i to layer j.
Accordingly, our video architecture takes as input samples holding 16 consecutive frames (called *clips* from now on) and provides a dense probability map for the last (current) frame of the clip. The network is jointly trained with two input streams (Fig. 2), in order to tackle the central bias that usually affects saliency benchmarks in general, and is even more noticeable in the driving task.

Both streams rely on the same backbone encoder, that we name `COARSE` module as provides a first rough estimate of driver's focus of attention. Such model is based on the work by Tran *et al.* [44] and employs their C3D architecture to map pixels into a 512-dimensional encoding space. Being interested in spatially coherent feature maps, we drop the top fully connected classification module. Moreover, we discard the deepest convolutional layer, which encodings are strongly tailored to the original action recognition task, retaining only the most general features provided by previous layers. Eventually, we modify the last pooling layer to cover the whole time axis, and therefore squeeze out the temporal dimension from the output features. The resulting map, which is reduced by a 16x factor along spatial dimension and lacks the temporal axis due to pooling layers, is then processed to produce an output estimate as big as the original image and featuring a single probability channel. This is achieved by means of a series of upsampling followed by convolutions.

During training, the model is fed with two streams. The first stream encourages the model to learn attention estimation given visual cues rather than prior spatial bias, and feeds the `COARSE` model with random crops. Cropping is also employed in the original C3D training process. Indeed, in [44] authors perform a tensor resize to 128×128 and then a random 112×112 crop. In our experience, this cropping policy is too polite, and yields models strongly biased towards the image center since ground-truth maps still suffer a poor variety. The policy we employ is immoderate, and features a 256×256 resize before the crop. This way, samples cover a small portion of the input tensor and allow variety in prediction targets, at the cost of a wider attentional area. Intuitively, the smaller crops are, the larger the attentional map will appear. Thus, the trained model was able to escape the bias when required, but unfortunately provided over-rough estimates. To address this issue, we feed the `COARSE` model with a second stream providing images resized to match the crop size. The prediction, after being resized and concatenated with the last frame of the clip, than undergoes a further block of convolutional layers (`FINE` module) that refine the map.

Estimates from both streams are modeled as a probability density P over pixels, and optimized jointly against a ground-truth map Q by means of the Kullback-Leibler divergence:

$$D_{KL}(P, Q) = \sum_i Q_i \log \left(\epsilon + \frac{Q_i}{\epsilon + P_i} \right) \qquad (3)$$

where the summation index i spans across image pixels and ϵ is a regularization constant.

Note that cropping policies can be of paramount importance for network training in presence of strong biases. This is the case as, being trained with crops the network estimates attention disregarding its spatial location. On the contrary, being trained with resized clips the model can also learn a spatial prior. Balancing the effect of the two streams clearly benefits the prediction task in its whole.

4. Experimental Evaluation

In this section, we provide experimental results of our deep learning architectures on different saliency datasets both in image and video domains.

4.1. Datasets

To validate the effectiveness of our deep learning architectures, we perform experiments on three different datasets: two datasets commonly used for image saliency prediction and one for estimating the driver's attention in videos.
- SALICON [26]: It is the largest dataset available for image saliency prediction with 20,000 images divided in training, validation and test sets and taken from the Microsoft COCO dataset [35]. Human fixation points were collected by simulating eye movements with a mouse-based strategy, instead of using an eye-tracking system. Nevertheless, authors demonstrated an high degree of similarity between their saliency maps and those collected using eye-tracking devices. Two different versions of this dataset are currently available: in the latest version, authors replaced the original velocity-based fixation detection algorithm with a new algorithm based on the Cluster Fix [31] that resulted in more eye-like fixations. In the following section, we

report experiments on both versions of this dataset.
- MIT300 [27]: Despite its limited size, it is one of the most commonly used datasets for saliency prediction. It is composed of 300 natural images, in which saliency maps have been created from eye-tracking data of 39 observers. Ground-truth saliency maps of this dataset are not publicly available and predictions must be submitted to the MIT Saliency Benchmark [6] for evaluation. To test our image saliency model on this dataset, we first fine-tuned the network, trained on the original SALICON dataset, on the MIT1003 dataset [28], which contains $1,003$ images available with human eye fixations.
- `DR(eye)VE` [1]: It was recently introduced for studying attentional behaviors during the driving task. It is composed of 74 video sequences acquired from the perspective of a car, each of which is 5 minutes long, enriched with driver's focus of attention within the surrounding urban scene and other sensors' measurements. Fixations were recorded from the driver's point of view by means of eye tracking glasses, and then projected to the car's perspective with standard image registration techniques. Eight different experienced drivers (both male and female), were asked to take part to the acquisition process. In order to measure the potential shift in attention due to different environmental contexts, sequences were acquired in different landscapes (downtown, countryside, highway), weather (sunny, cloudy, rainy) and time of day (morning, evening, night). It is split into train, validation and test set as follows: sequences 1-38 are used for training, sequences 39-74 for testing. The 500 frames in the middle of each training sequence constitute the validation set.

4.2. Implementation details

The ML-Net model was trained with mini-batches of $N = 10$ samples by using the Stochastic Gradient Descent with Nesterov momentum 0.9, weight decay 0.0005 and learning rate 10^{-3}. The α parameter was set to 1.1 in all our experiments.

Concerning the training of the `COARSE+FINE` model, training clips were augmented with horizontal mirroring. As for the optimizer we choose Adam [29], and set all its hyperparameters as suggested in the original paper, with the exception of the learning rate that we set to 10^{-4}. We trained the network until convergence without early stopping or learning rate decay policies using mini-batches of 8 examples.

Table 1
Complexity of the presented models in terms of number of trainable parameters, memory footprint during inference, amount of time required for training and inference.

	ML-Net	`COARSE`	`COARSE+FINE`
Nb. Parameters	~15.45M	~13.49M	~13.51M
Memory Occupation	~274MB	~595MB	~596MB
Training Time	~12h	~20h	~20h
Inference Time	~19.8ms	~24.5ms	~30.4ms

Both models were implemented in the Keras framework using Theano as backend. Training and experiments were performed using a NVIDIA TitanX GPU. We report in Table 1 some measurements about the complexity the proposed networks. As it can be observed, the models for video saliency feature a reduction in the number of parameters, with respect to the ML-Net model, thanks to careful architectural choices. This reduction in the number of parameters, however, does not result in lower memory occupation, given the cost of allocating video snippets and activations over the temporal axis. In terms of inference time, the proposed architecture for image saliency is faster than those presented for video saliency.

4.3. Evaluation metrics

Several evaluation metrics have been proposed for saliency prediction and the main difference between them concerns the ground-truth representation. In fact, some of these metrics consider saliency at discrete fixation locations, while others treat both predicted saliency maps and ground-truth maps, generated from fixation points, as distributions [7]. In this work, we evaluate our saliency architectures on the following evaluation metrics.

The Normalized Scanpath Saliency (NSS) metric was introduced specifically for the evaluation of saliency models [41]. The idea is to measure saliency values at eye fixation locations and to scale them according to the deviation of the whole map

$$NSS(p) = \frac{SM(p) - \mu_{SM}}{\sigma_{SM}} \tag{4}$$

where p is the location of one fixation, SM is the saliency map and μ_{SM} and σ_{SM} indicate its mean and standard deviation respectively. The final NSS score is the average of $NSS(p)$ for all fixations

$$NSS = \frac{1}{N} \sum_{p=1}^{N} NSS(p) \tag{5}$$

Table 2
Comparison results on both versions of the SALICON dataset [26].

		$CC \uparrow$	$sAUC \uparrow$	$AUC \uparrow$	$NSS \uparrow$
v2015 (val. set)	Infinite humans	1.00	0.87	0.94	3.81
	ML-Net	0.74	0.78	0.87	2.83
	Itti-Koch	0.39	0.63	0.77	1.10
v2015 (test set)	ML-Net	0.74	0.77	0.87	2.79
	Itti-Koch	0.42	0.64	0.78	1.21
v2017 (test set)	ML-Net	0.71	0.72	0.82	1.61
	Itti-Koch	0.55	0.63	0.77	1.10

where N is the total number of eye fixations.

The Similarity metric [27] (SIM) is computed as the sum of pixel-wise minimums between the predicted saliency map SM and the human eye fixation map FM, after normalizing the two maps

$$SIM = \sum_{x=1}^{X} min(SM(x), FM(x)) \tag{6}$$

where SM and FM are both normalized to be probability distributions (*e.g.* sum up to one). A similarity score of one indicates that the predicted map is identical to the ground-truth one.

The linear Correlation Coefficient (CC), instead is the Pearson's linear coefficient between SM and FM and is computed as

$$CC = \frac{cov(SM, FM)}{\sigma_{SM} * \sigma_{FM}} \tag{7}$$

where cov is the covariance, σ_{SM} and σ_{FM} are the standard deviations of SM and FM. It ranges between -1 and 1, and a score close to -1 or 1 indicates a perfect linear relationship between the two maps.

The Area Under the ROC curve is one of the most widely used metrics for evaluating saliency models even though does not penalize false positives. The saliency map is treated as a binary classifier of fixations at various threshold values, and a ROC curve can be drawn by measuring the true and false positive rates under each binary classifier. There are several different implementations of this metric which differ in how true and false positives are calculated. In this work, we consider the Judd (AUC) and shuffled versions ($sAUC$).

The Earth Mover's Distance (EMD) represents the minimal cost to transform the probability distribution of the saliency map SM into the one of the human eye fixations FM. Therefore, a larger EMD indicates a larger difference between the two maps.

Finally, the Kullback-Liebler Divergence (D_{KL} in Eq. 3) evaluates the loss of information when the distribution SM is used to approximate the distribution FM, therefore taking a probabilistic interpretation of saliency and ground-truth maps.

4.4. Image saliency results

As a proof of concept, in Tables 2 and 3 we compare the results of the model in [23][2] with those of our method on both versions of the SALICON and MIT300 datasets, respectively. It can clearly be seen that CNNs overcame that early model by a big margin, with respect to all metrics, and this experimentally confirms the need of high-level features for saliency prediction, rather than just employing low-level cues such as in [23].

To give a better insight of the performance gain, we also report some qualitative results on images randomly chosen from the original SALICON dataset. We show them in Fig. 3, along with the ground-truth saliency map computed from human eye fixations. While the model of [23] tends to concentrate on color and gradient discontinuities, which often do not match with the human fixation map, our model can clearly guess most of the saliency maps in a way which is almost indistinguishable from the ground-truth. The middle image, showing a pizza, is also a good example to show the role of the center prior: when there is no a clear object which stands out in the scene, human eyes tend to fix the center of the image, as our model has learned to do. Also, predictions from our ML-Net are particularly focused on small areas, similarly to the SALICON ground-truth. This is due to the fact that, in absence of a task-driven attentive mechanism, the focus tends to be directed on what is *a-priori* known,

[2]Numerical and qualitative results of the Itti-Koch model have been generated using the re-implementation of [18], which is also the one reported in the MIT Saliency Benchmark [6].

Table 3
Comparison results on the MIT300 dataset [27].

	$SIM \uparrow$	$CC \uparrow$	$sAUC \uparrow$	$AUC \uparrow$	$NSS \uparrow$	$EMD \downarrow$
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-Koch	0.44	0.37	0.63	0.75	0.97	4.26

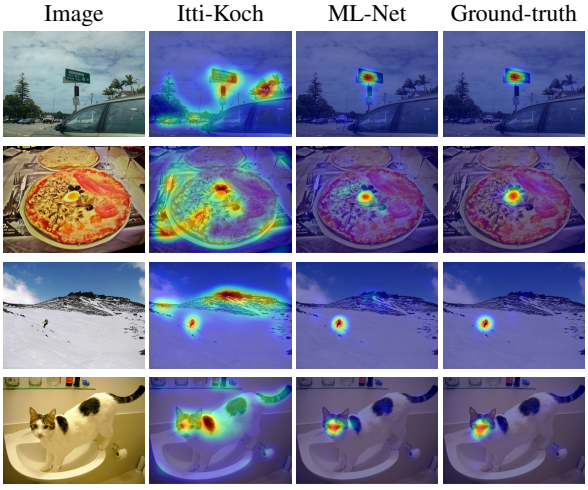

Fig. 3. Qualitative comparisons between the Itti [23] and ML-Net [10] models. Images are from the first release of the SALICON dataset (v2015) [26].

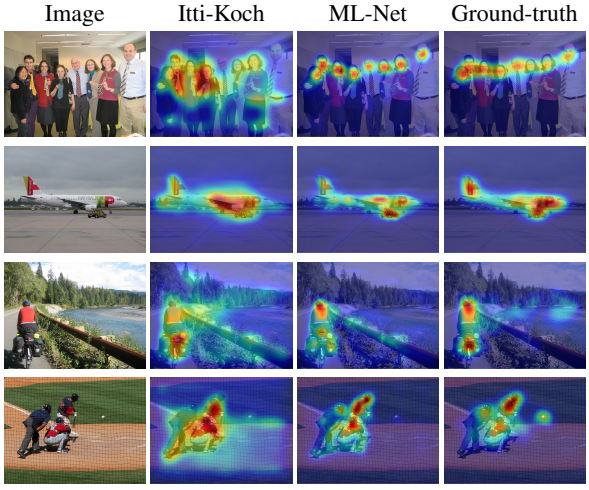

Fig. 4. Qualitative comparisons between the Itti [23] and ML-Net [10] models. Images are from the second release of the SALICON dataset (v2017) [26].

such as a person, a face, a traffic sign. The architecture, trained on similar data, does not overfit specific points, but tends to replicate the same semantic-based attentive behaviour.

Finally, we also report quantitative and qualitative results on the last release of the SALICON dataset in Table 2 and Fig. 4. As it can be seen, ground-truth saliency maps of this version are more blurred than the previous ones thus bringing the overall performance of the Itti and Koch model closer to deep learning architectures, especially on the CC metric. However, the gap between traditional methods and convolutional neural networks is still very important both quantitatively and qualitatively confirming the need of high level features for saliency prediction.

4.5. Task-driven attention prediction results

Here we discuss the experiments performed in order to assess the design choices of our architecture for attention prediction in video. As common in public benchmarks, we first compare our model against two central baselines. The first one represents the central bias as a Gaussian $\mathcal{N}(\mu, \Sigma)$, being μ the image center and Σ a diagonal covariance matrix whose variances are coherent with the image aspect ratio. A more precise, task-driven baseline is obtained by averaging all training ground-truth maps. Furthermore, two unsupervised state-of-the-art video saliency models [47,48] are also included in the comparison. The evaluation has been carried out comparing the shift between predicted and ground-truth maps both in terms of Pearson's correlation coefficient (CC) and Kullback-Leibler divergence (D_{KL}). We report such measures evaluated both in the whole test set and in the acting subsequences only[3] in Tab. 4. Moreover, we report results of the ML-Net model, that was originally proposed for image saliency and has been properly trained from scratch on task-driven human fixations from the `DR(eye)VE` dataset.

Several conclusions can be drawn from this evaluation. Firstly, from the poor performances of unsupervised models emerges the peculiar nature of the driving context, that demands for task-driven super-

[3] *acting subsequences* in `DR(eye)VE` are clips in which the driver is looking far from the image center due to a peculiar maneuver he is performing. We refer the reader to [38] for details.

Table 4
Evaluation of the proposed models against central baselines, both on the test and acting sequences of the DR(eye)VE dataset [1].

	Test seq.		Acting 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.* [48]	0.08	3.77	–	–
Wang *et al.* [47]	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

vision. Moreover, it can be noticed that the attentive subset of samples is crucial for the evaluation, as simple input-agnostic baselines perform positively overall. Finally, an important remark is revealed by the superior performance of the proposed model w.r.t ML-Net. The gap in performance is due to the temporal nature of video data: indeed, COARSE+FINE profitably learned to extract temporal features that are meaningful for video saliency prediction, whereas the design of ML-Net cannot capture such precious dependencies. A qualitative illustration of the difference in predictions is illustrated in Fig. 5.

5. Applying saliency to image captioning: NeuralStory

As a complementary contribution, we also discuss how image saliency can be applied to boost automatic image description architectures. This work is part of a large project called *NeuralStory*, which aims at providing new services for annotation, retrieval and re-use of video material in education. The goal of the project is to re-organize video material by extracting its storytelling structure and presenting it with new forms of summarization for quick browsing. Videos are divided into shots and scenes with a deep learning-based approach [3], using images, audio and semantic concepts extracted with a suitable CNN. The resulting annotation is also enriched by means of an image captioning architecture boosted with saliency.

The goal of image captioning is to provide a natural language description of an input image. This is usually carried out by means of recurrent neural network architectures (such as LSTMs [20]), which can be conditioned on a feature vector extracted from the input image, and generate the corresponding description step by step.

5.1. Saliency-boosted image captioning model

Machine attention mechanisms [49] are a popular way of obtaining time-varying inputs for recurrent architectures. In image captioning, it is well-known that performances can be improved by providing the generative LSTM with the specific region of the image it needs to generate a word: at each timestep the attention mechanism selects a region of the image, based on the previous LSTM state, and feeds it to the LSTM, so that the generation of a word is conditioned on that specific region, instead of being driven by the entire image.

The most popular attentive mechanism is the so-called "soft-attention" [49]. The input image is encoded as a grid of feature vectors $\{\mathbf{a}_1, \mathbf{a}_2, ..., \mathbf{a}_L\}$, each corresponding to a spatial location of the image. These are usually obtained from the activations of a convolutional or pooling layer of a CNN. At each timestep, the soft-attention mechanism computes a context feature vector $\hat{\mathbf{z}}_t$ representing a specific part of the input image, by combining feature vectors $\{\mathbf{a}_i\}_i$ with weights obtained from a *softmax* operator. Formally, the context vector $\hat{\mathbf{z}}_t$ is obtained as

$$\hat{\mathbf{z}}_t = \sum_{i=1}^{L} \alpha_{ti} \mathbf{a}_i, \tag{8}$$

where α_{ti} are weights representing the current state of the machine attention. These are driven by the original image feature vectors and by the previous hidden state $\mathbf{h}_{t-1}$ of the LSTM:

$$e_{ti} = v_e^T \cdot \phi(W_{ae} \cdot \mathbf{a}_i + W_{he} \cdot \mathbf{h}_{t-1})) \tag{9}$$

$$\alpha_{ti} = \frac{\exp(e_{ti})}{\sum_{k=1}^{L} \exp(e_{tk})}, \tag{10}$$

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

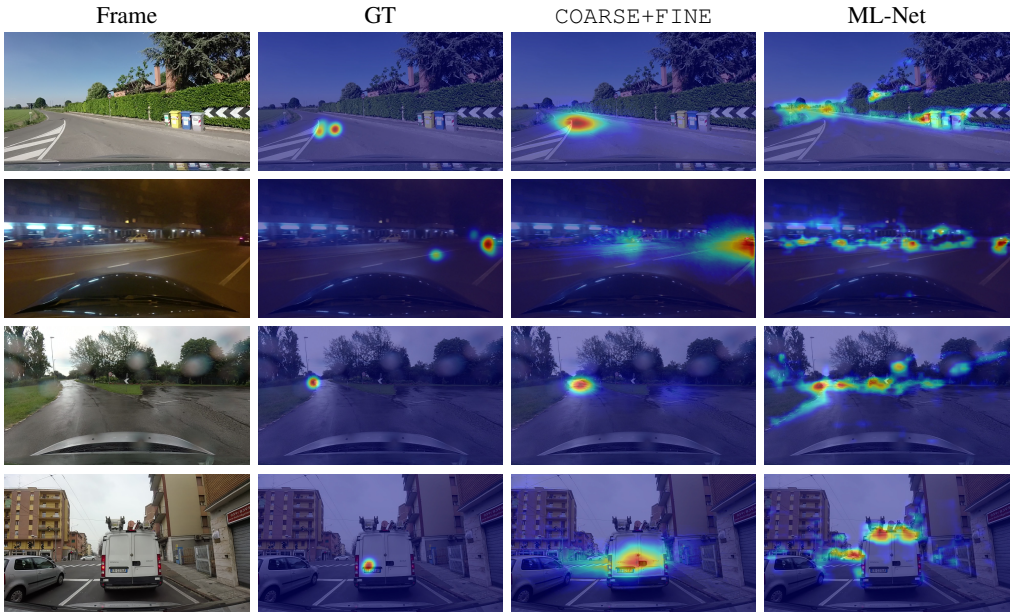

Fig. 5. Representation of differences in the video attention prediction estimation. This qualitative assessment indicates the suitability of the COARSE+FINE model in encoding temporal information. On the other hand, the ML-Net model processes still images and is more influenced by low-level non temporal features.

Table 5
Image captioning results on Microsoft COCO dataset [35] in terms of BLEU@1-4, METEOR, ROUGE$_L$ and CIDEr.

	BLEU@1↑	BLEU@2↑	BLEU@3↑	BLEU@4↑	METEOR↑	ROUGE$_L$ ↑	CIDEr↑
Soft Attention	71.7	54.6	40.2	29.4	25.3	52.9	93.9
Saliency-Guided Attention	71.8	54.7	40.4	29.6	25.4	53.0	94.4
ATT [50]	70.9	53.7	40.2	30.4	24.3	-	-
SCA-CNN [8]	71.9	54.8	41.1	31.1	25.0	-	-
PG-SPIDEr [36]	74.3	57.8	43.3	32.2	25.1	54.4	100.0

In [11], to investigate the role of visual saliency in the context of attentive captioning models, we extended this schema by splitting the machine attention into saliency and non saliency regions, and learning different weights for both of them. Given a visual saliency predictor which predicts a saliency map $\{s_1, s_2, ..., s_L\}$, having the same resolution of the feature vector grid $\{\mathbf{a}_i\}_i$, and with $s_i \in [0, 1]$, we proposed to modify Eq. 9 as follows:

$$e_{ti}^{sal} = v_{e,sal}^T \cdot \phi(W_{ae} \cdot \mathbf{a}_i + W_{he} \cdot \mathbf{h}_{t-1})) \tag{11}$$

$$e_{ti}^{nosal} = v_{e,nosal}^T \cdot \phi(W_{ae} \cdot \mathbf{a}_i + W_{he} \cdot \mathbf{h}_{t-1})) \tag{12}$$

$$e_{ti} = s_i \cdot e_{ti}^{sal} + (1 - s_i) \cdot e_{ti}^{nosal}. \tag{13}$$

Notice that our model learns different weights for saliency and non-saliency regions ($v_{e,sal}^T$ and $v_{e,nosal}^T$ respectively), and combines them into a final attentive map in which the contributions of salient and non-salient regions are merged together. Similarly to the classical soft-attention approach, the proposed generative LSTM can focus on every region of the image, but the focus on salient region is driven by the output of the saliency predictor.

5.2. Results

Table 5 compares the performances of our approach against the unsupervised machine attention approach in [49]. Due to different implementation details, the numerical results that we report are not directly comparable with those in the original Soft Attention paper

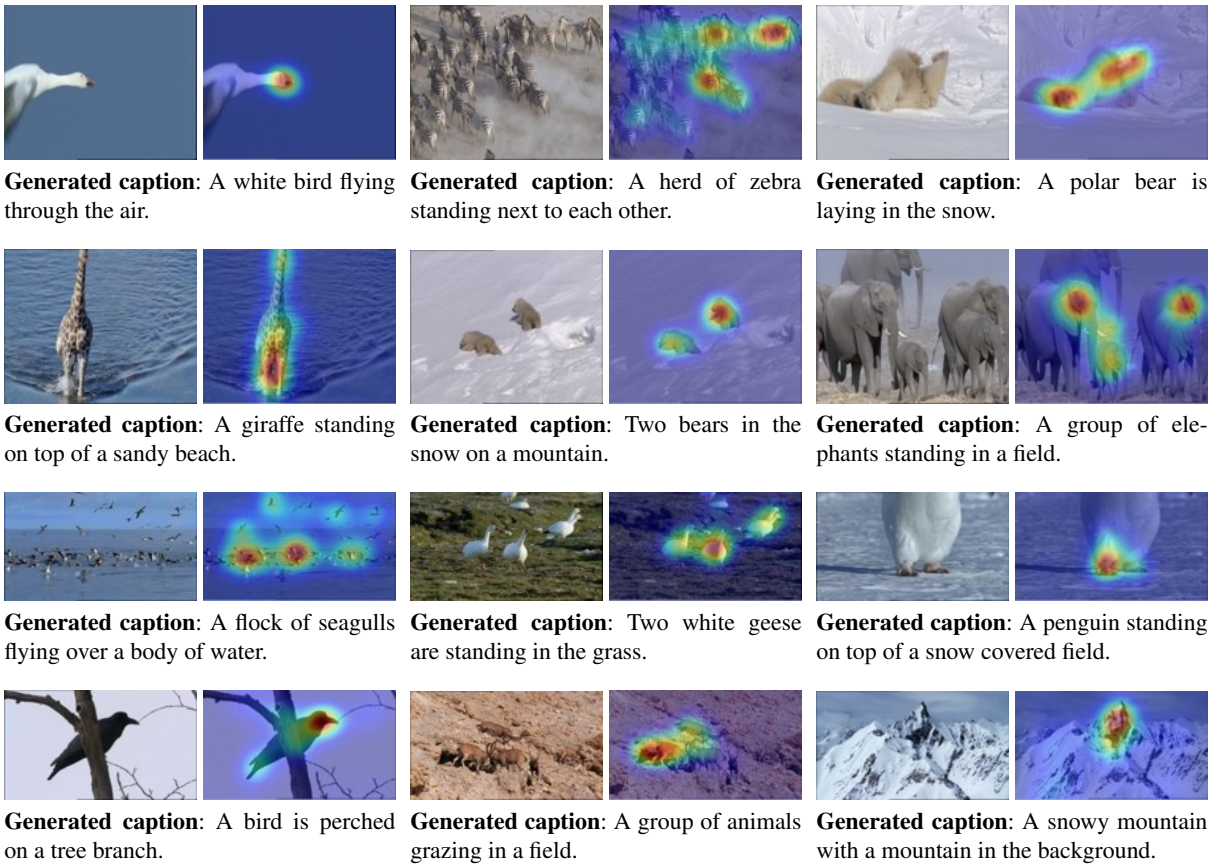

Fig. 6. Saliency maps and captions generated on sample images taken from the BBC Planet Earth dataset [2].

(ours are in general higher than the original ones). As it can be seen, our attention model, which incorporates visual saliency, is able to achieve better results on all metrics with respect to the Soft Attention approach.

For a complete evaluation, we also report the results of some recent state of the art image captioning models. It is easy to see that, even though in some cases our model obtains slightly worse results, the overall performance results satisfactory and confirms the role of saliency in image captioning.

Figures 6 and 7 show some captions automatically generated by our architecture on images respectively taken from the BBC educational TV series *Planet Earth* and an art documentary which are part of *NeuralStory*. As it can be seen, even though the model has been trained on a different domain, it is still able to generalize and provide appropriate captions. With this work we intend to enrich the annotation and key-frame description on the web interface. Automatically generated captions will be useful for human search, for automatic search by query, and possibly for future query-answering services.

6. Conclusions

In this work we presented different deep learning architectures for saliency prediction on images and video, showing the importance of multi-level features and the ability of convolutional architectures to deal with video sequences. The comparison between today's models and the early model by Itti and Koch [23] has shown several similarities in the way feature are extracted, and motivated the gap in performances with current models, which is not merely due to the their brute-force nature, but also to their ability to recall very closely early saliency and biological models, although improved with the semantics learned on the ground-truth. Finally, we showcased how saliency prediction can be incorporated in a image description ar-

Generated caption: A statue of a man standing next to a building.

Generated caption: A large body of water surrounded by mountains.

Generated caption: A person holding a book in their hand.

Generated caption: A woman and a man are standing together.

Generated caption: A woman walking down a street with a bag.

Generated caption: A woman in a red dress riding a bicycle.

Generated caption: A woman is sitting at a table with a plate of food.

Generated caption: A person is standing on a boat in the water.

Generated caption: A large body of water with a sky background.

Fig. 7. Saliency maps and captions generated on sample images taken from the *Meet the Romans with Mary Beard* TV series.

chitecture, and evaluated this last scenario both quantitatively and qualitatively.

References

[1] 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. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2016.

[2] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. A deep siamese network for scene detection in broadcast videos. In *ACM International Conference on Multimedia*, 2015.

[3] Lorenzo Baraldi, Costantino Grana, and Rita Cucchiara. Recognizing and presenting the storytelling video structure with deep multimodal networks. *IEEE Transactions on Multimedia*, 19(5):955–968, 2017.

[4] Loris Bazzani, Hugo Larochelle, and Lorenzo Torresani. Recurrent mixture density network for spatiotemporal visual attention. In *International Conference on Learning Representations*, 2017.

[5] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in Neural Information Processing Systems*, pages 155–162, 2005.

[6] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. http://saliency.mit.edu/.

[7] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *arXiv preprint arXiv:1604.03605*, 2016.

[8] Long Chen, Hanwang Zhang, Jun Xiao, Liqiang Nie, Jian Shao, Wei Liu, and Tat-Seng Chua. SCA-CNN: Spatial and Channel-Wise Attention in Convolutional Networks for Image Captioning. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

[9] Marcella Cornia, Davide Abati, Lorenzo Baraldi, Andrea Palazzi, Simone Calderara, and Rita Cucchiara. Attentive Models in Vision: Computing Saliency Maps in the Deep Learning Era. In *Conference of the Italian Association for Artificial Intelligence*, pages 387–399. Springer, 2017.

[10] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A Deep Multi-Level Network for Saliency Prediction. In *International Conference on Pattern Recognition*, 2016.

[11] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Visual Saliency for Image Captioning in New Multimedia Services. In *IEEE International Conference on Multimedia and Expo Workshops*, 2017.

[12] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 14(2):48, 2018.

[13] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 2018.

[14] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. SAM: Pushing the Limits of Saliency Prediction Models. In *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*, 2018.

[15] Marcella Cornia, Stefano Pini, Lorenzo Baraldi, and Rita Cucchiara. Automatic image cropping and selection using saliency: An application to historical manuscripts. In *Digital Libraries and Multimedia Archives*, volume 806, 2018.

[16] Hayit Greenspan, Serge Belongie, Rodney Goodman, Pietro Perona, Subrata Rakshit, and Charles H Anderson. Over-

complete steerable pyramid filters and rotation invariance. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 1994.

[17] Hadi Hadizadeh and Ivan V Bajic. Saliency-aware video compression. *IEEE Transactions on Image Processing*, 23(1):19–33, 2014.

[18] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pages 545–552, 2006.

[19] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[20] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[21] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In *IEEE International Conference on Computer Vision*, 2015.

[22] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3):194–203, 2001.

[23] Laurent Itti, Christof Koch, Ernst Niebur, et al. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(11):1254–1259, 1998.

[24] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[25] Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231, 2013.

[26] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[27] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[28] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *IEEE International Conference on Computer Vision*, 2009.

[29] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[30] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of intelligence*, pages 115–141. Springer, 1987.

[31] Seth D König and Elizabeth A Buffalo. A nonparametric method for detecting fixations and saccades using cluster analysis: Removing the need for arbitrary thresholds. *Journal of neuroscience methods*, 227:121–131, 2014.

[32] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017.

[33] Srinivas SS Kruthiventi, Vennela Gudisa, Jaley H Dholakiya, and R Venkatesh Babu. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[34] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. DeepGaze I: Boosting saliency prediction with feature maps trained on ImageNet. In *ICLR Workshops*, 2015.

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*, 2014.

[36] Siqi Liu, Zhenhai Zhu, Ning Ye, Sergio Guadarrama, and Kevin Murphy. Improved Image Captioning via Policy Gradient Optimization of SPIDEr. In *IEEE International Conference on Computer Vision*, 2017.

[37] Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37, 2015.

[38] Andrea Palazzi, Davide Abati, Simone Calderara, Francesco Solera, and Rita Cucchiara. Predicting the Driver's Focus of Attention: the DR (eye) VE Project. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[39] Andrea Palazzi, Francesco Solera, Simone Calderara, Stefano Alletto, and Rita Cucchiara. Learning to attend like a human driver. In *Intelligent Vehicles Symposium*, 2017.

[40] Junting Pan, Kevin McGuinness, Sayrol E., N. O'Connor, and X. Giró-i Nieto. Shallow and Deep Convolutional Networks for Saliency Prediction. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[41] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of Bottom-Up Gaze Allocation in Natural Images. *Vision research*, 45(18):2397–2416, 2005.

[42] Dmitry Rudoy, Dan B Goldman, Eli Shechtman, and Lihi Zelnik-Manor. Learning video saliency from human gaze using candidate selection. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2013.

[43] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[44] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *IEEE International Conference on Computer Vision*, 2015.

[45] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980.

[46] Eleonora Vig, Michael Dorr, and David Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[47] Wenguan Wang, Jianbing Shen, and Fatih Porikli. Saliency-aware geodesic video object segmentation. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[48] Wenguan Wang, Jianbing Shen, and Ling Shao. Consistent video saliency using local gradient flow optimization and global refinement. *IEEE Transactions on Image Processing*, 24(11):4185–4196, 2015.

[49] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard S Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International Conference on Machine Learning*, 2015.

[50] Quanzeng You, Hailin Jin, Zhaowen Wang, Chen Fang, and Jiebo Luo. Image captioning with semantic attention. In *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[51] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *European Conference on Computer Vision*. 2014.

[52] Yun Zhai and Mubarak Shah. Visual attention detection in video sequences using spatiotemporal cues. In *ACM International Conference on Multimedia*, 2006.

[53] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *IEEE International Conference on Computer Vision*, 2013.

[54] Sheng-hua Zhong, Yan Liu, Feifei Ren, Jinghuan Zhang, and Tongwei Ren. Video saliency detection via dynamic consistent spatio-temporal attention modelling. In *AAAI*, 2013.