Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model

Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara

***Abstract*—Data-driven saliency has recently gained a lot of attention thanks to the use of Convolutional Neural Networks for predicting gaze fixations. In this paper we go beyond standard approaches to saliency prediction, in which gaze maps are computed with a feed-forward network, and present a novel model which can predict accurate saliency maps by incorporating neural attentive mechanisms. The core of our solution is a Convolutional LSTM that focuses on the most salient regions of the input image to iteratively refine the predicted saliency map. Additionally, to tackle the center bias typical of human eye fixations, our model can learn a set of prior maps generated with Gaussian functions. We show, through an extensive evaluation, that the proposed architecture outperforms the current state of the art on public saliency prediction datasets. We further study the contribution of each key component to demonstrate their robustness on different scenarios.**

***Index Terms*—Saliency, Human Eye Fixations, Convolutional Neural Networks, Deep Learning**

I. Introduction

VISUAL cognition science has shown that humans, when observing a scene without a specific task to perform, do not focus on each region of the image with the same intensity. Instead, attentive mechanisms guide their gazes on salient and relevant parts [1]. An intensive research effort has tried to emulate such selective visual mechanisms, as computational saliency can be applied to a wide range of applications like image retargeting [2], object recognition [3], video compression [4], tracking [5] and other data-dependent tasks such as image captioning [6].

Traditional saliency prediction methods have followed biological evidence by defining features that capture low-level cues such as color, contrast and texture or semantic concepts such as faces, people and text [7], [8], [9], [10]. However, these techniques have failed to capture the wide variety of causes that contribute to defining visual saliency maps.

With the advent of deep neural networks, saliency prediction has achieved strong improvements both thanks to specific architectures and to large annotated datasets [11], [12], [13], [14]. Although these approaches went beyond the limitations of hand-crafted models, no one has yet investigated the incorporation of machine attention models [15], [16], [17] in saliency prediction.

Machine attention [15] is a computational paradigm which sequentially attends to different parts of an input. This is usually achieved by exploiting a recurrent neural network, and by defining a compatibility measure between its internal state and regions of the input. This paradigm has been successfully applied to image captioning [15] and machine translation [18] to selectively focus on different parts of a sentence, and to action recognition [19] to focus on the relevant parts of a spatio-temporal volume. We argue that machine attention can also be effective for saliency prediction, as a powerful way to process saliency-specific features and to obtain an enhanced prediction.

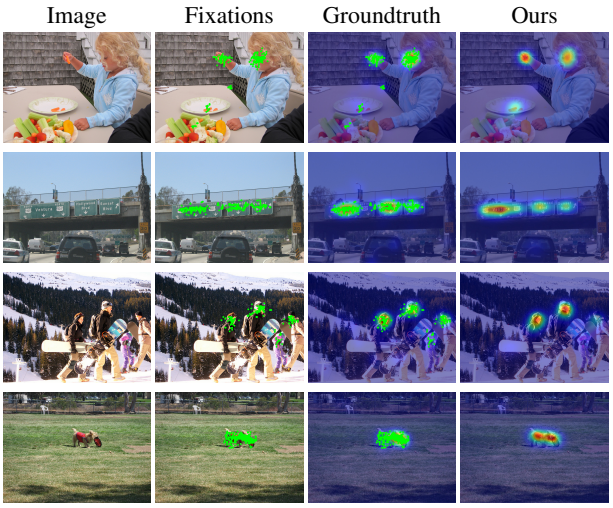


Fig. 1. Visual saliency prediction aims at predicting where humans gazes will focus on a given image. Groundtruth data is collected by means of eye-tracking glasses or mouse clicks to get eye fixation points, which are then smoothed together to obtain the groundtruth saliency map. Our model learns to predict the distribution of human fixation points by refining feature extracted from a CNN with a novel LSTM-based attentive model.

In this paper we propose a novel saliency prediction architecture that incorporates an Attentive Convolutional Long Short-Term Memory network (Attentive ConvLSTM) that iteratively focuses on relevant spatial locations to refine saliency features. The architecture is particularly original since the LSTM model is used to achieve a refinement over an image, instead of handling a temporal sequence.

Moreover, the rescaling caused by max-pooling and strides in convolutional layers deteriorates the performance of saliency prediction, we present an extension of two popular CNNs (namely, VGG-16 [20] and ResNet-50 [21]) which can reduce the downscaling effect and maintain spatial resolution. This expedient allows us to preserve detailed visual information and obtain improved feature extraction capabilities.

Finally, in order to handle the tendency of humans to fix

M. Cornia, L. Baraldi and R. Cucchiara are with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia, Modena, Italy (e-mail: {marcella.cornia, lorenzo.baraldi, rita.cucchiara}@unimore.it).

G. Serra is with the Department of Computer Science, Mathematics and Physics, University of Udine, Udine, Italy (e-mail: giuseppe.serra@uniud.it).

TABLE I
COMPARISON BETWEEN THE MAIN PROPERTIES OF OUR MODEL AND THOSE OF OTHER EXISTING SALIENCY METHODS. (*) INDICATES CITATIONS TO NON-PEER REVIEWED TEXTS.

	CNN	Attentive LSTM	Center Bias	Loss Function
SALICON [12]	AlexNet - VGG-16 - GoogleNet	✗	✗	KL-Div
DeepFix [22]	VGG-16	✗	handcrafted priors	Euclidean loss
SalNet [23]	VGG-16	✗	✗	Euclidean loss
PDP [13]	VGGNet	✗	✗	probability distances
ML-Net [14]	VGG-16	✗	single multiplicative map	normalized MSE
DSCLRCN [24] (*)	VGG-16 - ResNet-50	✗	✗	NSS
Saliency Attentive Model (SAM)	VGG-16 - ResNet-50	✓	multiple learned priors	combination of multiple saliency metrics

the center region of an image, we also introduce an explicit prior component. Unlike previous approaches that include handcrafted priors [9], [25], [11], [22], [26], our module keeps the architecture trainable end-to-end and can learn priors in an automatic way.

Figure 1 shows examples of saliency maps predicted by the proposed solution, which we call Saliency Attentive Model (SAM), compared with groundtruth saliency maps obtained from human eye fixations. We quantitatively validate our approach on three publicly available benchmark datasets: SALICON, MIT300 and CAT2000. Experimental results will show that the proposed solution significantly improves predictions. To summarize, the contributions of this paper are threefold:

- We propose a novel Attentive ConvLSTM that sequentially focuses on different spatial locations of a stack of features to enhance predictions. To the best of our knowledge, we are among the first to incorporate attentive models in a saliency prediction architecture.
- Our network is able to learn the bias present in eye fixations, without the need to integrate this information manually.
- The proposed solution overcomes by a big margin the current state of the art on the largest dataset available for saliency prediction, SALICON. Moreover, on MIT300 and CAT2000 our method achieves state of the art results showing competitive generalization properties.

We make the source code of our method and pre-trained models publicly available[1].

II. RELATED WORK

Pioneering works on saliency prediction were based on the Feature Integration Theory proposed by Treisman *et al.* [27] in the eighties. Itti *et al.* [28] defined the first computational model to predict saliency on images: this work, inspired by Koch and Ullman [29], 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. After this seminal work, a large variety of methods explored the same idea of combining complementary low-level features [30], [7], [31] and often included additional center-surround cues [32], [10]. Other methods enriched predictions exploiting semantic classifiers for detecting higher level concepts such as faces, persons, cars and horizons [33], [9], [34], [8], [35]. Related research efforts have also been done in the compressed domain, as in [36], [37].

A. Saliency and Deep Learning

Only recently, thanks to the large spread of deep learning techniques, the saliency prediction task has achieved a considerable improvement. One of the first proposals has been the *Ensemble of Deep Networks* (*e*DN) model by Vig *et al.* [25]. This model consists of three convolutional layers followed by a linear classifier that blends feature maps coming from the previous layers. After this work, Kümmerer *et al.* [11], [26] proposed two deep saliency prediction networks: the first, called *DeepGaze I*, was based on the AlexNet model [38], while the second, *DeepGaze II*, was built upon the VGG-19 network [20]. Liu *et al.* [39] presented a multi-resolution CNN (*Mr-CNN*) fine-tuned over image patches centered on fixation and non-fixation locations.

It is well known that deep learning approaches strongly depend on the availability of sufficiently large datasets. The publication of a large-scale eye-fixation dataset, SALICON [40], indeed contributed to a big progress of deep saliency prediction models. Huang *et al.* [12] introduced an architecture consisting of a deep neural network applied at two different image scales. They compared different standard CNN architectures such as AlexNet [38], VGG-16 [20] and GoogleNet [41], in particular showing the effectiveness of the VGG network.

After this work, several deep saliency models based on the VGG network have been published [22], [23], [13], [42], [14], [43], [44], [45]. Accordingly, we proposed a new architecture, called *ML-Net* [14], which improved previous attempts by using features coming from multiple layers of a CNN and by adding a learned prior map. In particular, we learned a matrix of weights which was applied to the output saliency map with

[1]https://github.com/marcellacornia/sam

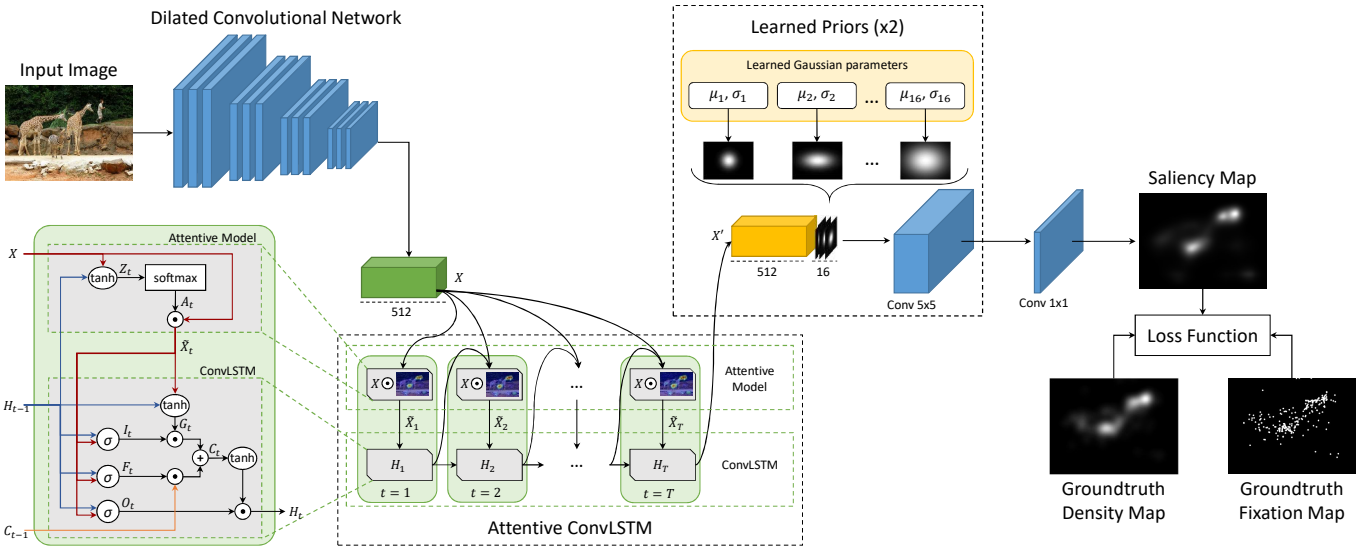


Fig. 2. Overview of our Saliency Attentive Model (SAM). After computing a set of feature maps on the input image through a new architecture called Dilated Convolutional Network, an Attentive Convolutional LSTM sequentially enhances saliency features thanks to an attentive recurrent mechanism. Predictions are then combined with multiple learned priors to model the tendency of humans to fix the center region of the image. During the training phase, we encourage the network to minimize a combination of different loss functions, thus taking into account different quality aspects that predictions should meet.

a pixel-wise multiplication. The usage of centered priors has also been investigated in [22], where multiple predefined priors were fed to a convolutional layer.

In this work, instead, we model the center bias present in human gazes using multiple learned prior maps. This is different from the approaches of [14] and [22], as we let the network learn a set of Gaussian parameters, keeping it trainable end-to-end without predefined information.

Recently, Pan *et al.* [44] introduced *SalGAN*, a deep network for saliency prediction trained with adversarial examples. As all other Generative Adversarial Networks, it is composed by two modules, a generator and a discriminator, which combine efforts to produce saliency maps.

In this work, we also employ the ResNet [21] model to extract feature maps from the input image. The only other saliency model that exploits this network is proposed by Liu *et al.* [24] and called *DSCLRCN*. This model simultaneously incorporates global and scene contexts to infer image saliency thanks to a deep spatial contextual LSTM which scans the image both horizontally and vertically.

To better highlight the differences of our model with respect to other existing saliency methods, we report in Table I a summary of the main properties of our solution and those of the most competitive methods. Note that none of the other methods incorporate an attentive mechanism or a set of prior maps directly learned by the network. In addition, differently from other previous models, we propose a loss function which is a balanced combination of different saliency metrics and that provides state of the art performances.

A related line of research is that of explaining activations of a neural model by means of techniques based on backpropagation [46]. It is worthwhile to notice that this research line is very different from that of saliency prediction, as it does not aim to replicate human fixations.

B. Salient Object Detection

Salient object detection is slightly related to the topic of this work, even though it is a significantly different task. Salient object detection consists, indeed, in identifying a binary map indicating the presence of salient objects [47], [48], [49], [50]. On the contrary, in saliency prediction the objective is to predict a density map of eye fixations.

A saliency detection approach which is in some aspects related to our work is that of Kuen *et al.* [51], in which a recurrent (non convolutional) network provides salient object detection. At each timestep, their recurrent network outputs the parameters of a spatial transformation which is used to focus on a particular location of the image, and builds the binary prediction for that location. Our recurrent network is, instead, convolutional, and is used to process saliency features by iteratively refining the prediction.

III. Model Architecture

In this section we present the architecture of our complete model, called SAM (Saliency Attentive Model).

The main novelty of our proposal is an Attentive Convolutional model, which recurrently processes saliency features at different locations, by selectively attending to different regions of a tensor. This architecture, that for the first time uses an LSTM without the concept of time, is described in Section III-A.

Predictions are then combined with multiple learned priors which are used to model the human-gaze center bias (Section III-B). To extract feature maps from input images, we employ a Convolutional Neural Network model. Instead of using a pre-defined CNN, we propose a Dilated Convolutional Network to limit the rescaling effects which can worse saliency prediction performance (Section III-C). A new combination of

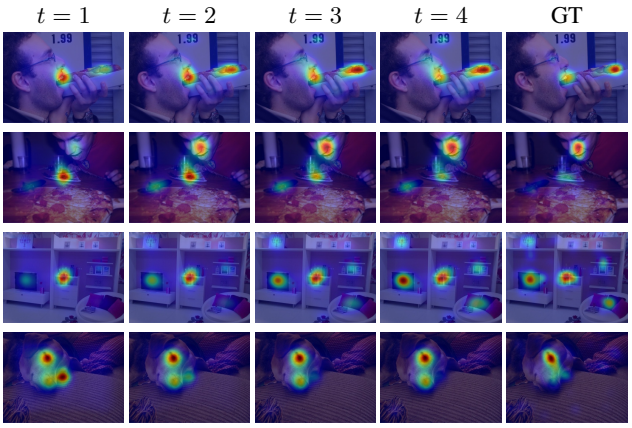


Fig. 3. Progressive refinement of predictions performed by the Attentive ConvLSTM. The first and the second row show a progressive change of focus in the saliency map, so that regions which were wrongly predicted as salient are progressively corrected, and truly salient regions are correctly identified. The third and the fourth row, instead, respectively show an increase and a reduction of saliency in regions of the image that have been (or have not been) considered as salient at the first timestep. In all cases, the result is a progressive approach of the saliency map to the groundtruth.

different loss functions is finally used to train the whole network by simultaneously taking into account different quality aspects (Section III-D). The overall architecture of our model is shown in Figure 2.

A. Attentive Convolutional LSTM

Long Short-Term Memory networks [52] achieved good performances on several tasks in which time dependencies are a key component [53], [54], [55], [56], but they can not be directly employed for saliency prediction, as they work on sequences of time varying vectors. We extend the traditional LSTM to work on spatial features: formally this is achieved by substituting dot products with convolutional operations in the LSTM equations. Moreover, we exploit the sequential nature of LSTM to process features in an iterative way, instead of using the model to deal with temporal dependencies in the input.

To explain our proposal of the attentive model, let's consider the LSTM scheme on the left part of Fig. 2. Here the LSTM takes as input a stack of features extracted from the input image (X in Fig. 2) and produces a refined stack of feature maps (X' in Fig. 2) entering in the learned prior module. The LSTM works by sequentially updating an internal state, according to the values of three sigmoid gates. Specifically, the update is driven by the following equations:

$$I_t = \sigma(W_i * \tilde{X}_t + U_i * H_{t-1} + b_i) \quad (1)$$

$$F_t = \sigma(W_f * \tilde{X}_t + U_f * H_{t-1} + b_f) \quad (2)$$

$$O_t = \sigma(W_o * \tilde{X}_t + U_o * H_{t-1} + b_o) \quad (3)$$

$$G_t = \tanh(W_c * \tilde{X}_t + U_c * H_{t-1} + b_c) \quad (4)$$

$$C_t = F_t \odot C_{t-1} + I_t \odot G_t \quad (5)$$

$$H_t = O_t \odot \tanh(C_t). \quad (6)$$

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

The input of the LSTM layer $\tilde{X}_t$ is computed, at each timestep (*i.e.* at each iteration), through an attentive mechanism. In particular, an attention map is generated by convolving the previous hidden state H_{t-1} and the input X, feeding the result to a $\tanh$ activation function and finally convolving with a one channel convolutional kernel:

$$Z_t = V_a * \tanh(W_a * X + U_a * H_{t-1} + b_a). \quad (7)$$

The output of this operations is a 2-d map from which we can compute a normalized spatial attention map through the *softmax* operator:

$$A_t^{ij} = p(att_{ij}|X, H_{t-1}) = \frac{\exp(Z_t^{ij})}{\sum_i \sum_j \exp(Z_t^{ij})} \quad (8)$$

where A_t^{ij} is the element of the attention map in position (i, j). The attention map is applied to the input X with an element-wise product between each channel of the feature maps and the attention map:

$$\tilde{X}_t = A_t \odot X. \quad (9)$$

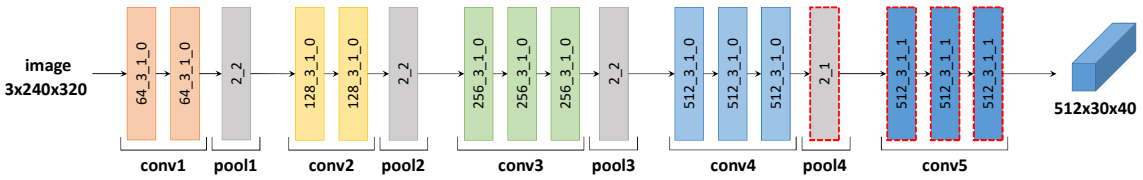
Fig. 3 shows saliency predictions on four sample images, using the output of the ConvLSTM module at different timesteps as input of the rest of the model. As can be noticed, predictions are progressively refined by modifying the initial map given by the CNN. This refinement results in an significant enhancement of the predictions.

B. Learned Priors

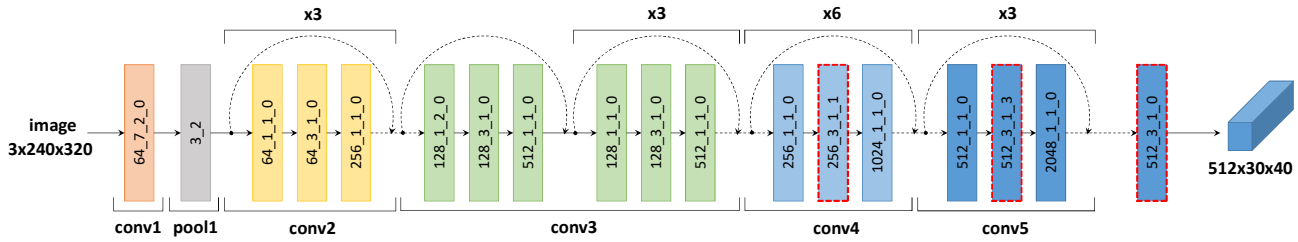
Psychological studies have shown that when observers look at images, their gazes are biased toward the center [57], [58]. This phenomenon is mainly due to the tendency of photographers to position objects of interest at the center of the image. Also, when people repeatedly watch images with salient information placed in the center, they naturally expect to find the most informative content of the image around its center [58]. Another important reason that encourages this behavior is the interestingness of the scene [59]. Indeed, when there are no highly salient regions, humans are inclined to look at the center of the image.

Based on this evidence, the inclusion of center priors is a key component of several recent works of saliency prediction [9], [25], [11], [22], [26], [14]. Differently from existing works, which included pre-defined priors, we let the network learn its own priors. To reduce the number of parameters and facilitate the learning, we constraint each prior to be a 2d Gaussian function, whose mean and covariance matrix are instead freely learnable. This lets the network learn its own priors purely from data, without relying on assumptions from biological studies.

We model the center bias by means of a set of Gaussian functions with diagonal covariance matrix. Means and vari-



(a) Dilated VGG Convolutional Network



(b) Dilated Residual Convolutional Network

Fig. 4. Overall architectures of Dilated Convolutional Networks based on the VGG-16 and ResNet-50 models. Convolutional and pooling blocks are respectively expressed in terms of `channels_kernel_stride_holes` and `kernel_stride`. On top of the ResNet model, we report the number of repetitions for each block. Red dashed edges indicate modified layers with respect to the original networks.

ances are learned for each prior map according to the following equation:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp\left(-\left(\frac{(x-\mu_x)^2}{2\sigma_x^2} + \frac{(y-\mu_y)^2}{2\sigma_y^2}\right)\right). \tag{10}$$

Our network learns the parameters of N Gaussian functions (in our experiments $N = 16$) and generates the relative prior maps. Since the X' tensor has 512 channels, after the concatenation with learned prior maps, we obtain a tensor with 528 channels. The resulting tensor is fed through a convolutional layer with 512 filters. This operation adds more non-linearity to the model and proves to be effective with respect to other previous works, as reported in Section V-C. The entire prior learning module is replicated two times.

C. Dilated Convolutional Network

One of the main drawbacks of using CNNs to extract features for saliency prediction is that they considerably rescale the input image during the feature extraction phase, thus worsening the prediction accuracy. In the following, we devise a strategy which increases the output resolution of a CNN while preserving the scale at which convolutional filters operate and the number of parameters. This makes it possible to use pre-trained weights, and thus to reduce the need for fine-tuning convolutional filters after the network structure has been modified.

The intuition of the approach is that given a CNN of choice and one of its layers having stride $s > 1$, we can increase the output resolution by reducing the stride of the layer, and adding dilation [60] to all the layers which follow the chosen layer. In this way, all convolutional filters still operate on the same scale they have been trained for. We apply this technique on two recent feature extraction networks: the VGG-16 [20] and the ResNet-50 [21].

The VGG-16 network is composed by 13 convolutional layers and 3 fully connected layers. The convolutional layers are divided in five convolutional blocks where, each of them is followed by a max-pooling layer with a stride of 2.

The ResNet-50, instead of having a series of stacked layers that process the input image as in common CNNs, performs a series of residual mappings between blocks composed by a few stacked layers. This is obtained using shortcut connections that realize an identity mapping, *i.e.* the input of the block is added to its output. Residual connections help to avoid the accuracy degradation problem [61] that occurs with the increase of the network depth, and are beneficial also in the saliency prediction case, since they improve the feature extraction capabilities of the network.

In particular, the ResNet-50 network consists of five convolutional blocks and a fully connected layer. The first block is composed by one convolutional layer followed by a max-pooling layer, both of them having a stride of 2, while the remaining four blocks are fully convolutional. All of these blocks, except the second one (`conv2`), reduce the dimension of feature maps with strides of 2.

Since the purpose of our network is to extract feature maps, we only consider convolutional layers and ignore fully connected layers which are present at the end of both networks. Moreover, it can be noticed that the downscaling factor of both of these architectures is particularly critical. For example, with an input image having a size of 240×320, the output dimension is 8×10, which is relatively small for the saliency prediction task. For this reason, we modify network structures to limit the rescaling phenomenon.

For the VGG-16 model, we also remove the last max-pooling layer and apply the aforementioned technique to the last but one pooling layer (see Figure 4a). On the contrary, for the ResNet-50 model we remove the stride and we introduce dilated convolutions in the last two blocks (see Figure 4b). In this case, since the technique is applied two times, we introduce holes of size 1 in the kernels of the block `conv4` and holes of size $2^2 - 1 = 3$ in the kernels of the block `conv5`. The output of the residual network is a tensor with

2048 channels. To limit the number of feature maps, we feed this tensor into another convolutional layer with 512 filters. Thanks to these expedients, our saliency maps are rescaled by a factor of 8 instead of 32 as in the original VGG-16 and ResNet-50 models.

We include dilated convolutions also in prior layers, thus obtaining two convolutional layers with large receptive fields that allow us to capture the saliency of an object with respect to its neighborhood. We set the kernel size of these layers to 5 and the holes size to 3 achieving therefore a receptive field of 17×17. Strides of these layers are set to 1 and both of them are followed by a ReLU activation function.

The last layer of our model is a convolutional operation with one filter and a kernel size of 1 that extracts the final saliency map. Since the predicted map has lower dimensions than the original image, it is brought to its original size via bilinear upsampling.

D. Loss function

In order to capture several quality factors, saliency predictions are usually evaluated through different metrics. Inspired by this evaluation protocol, we introduce a new loss function given by a linear combination of three different saliency evaluation metrics. We define the overall loss function as follows:

$$L(\tilde{\mathbf{y}}, \mathbf{y}^{den}, \mathbf{y}^{fix}) = \alpha L_1(\tilde{\mathbf{y}}, \mathbf{y}^{fix}) + \beta L_2(\tilde{\mathbf{y}}, \mathbf{y}^{den}) + \gamma L_3(\tilde{\mathbf{y}}, \mathbf{y}^{den}) \tag{11}$$

where $\tilde{\mathbf{y}}$, $\mathbf{y}^{den}$ and $\mathbf{y}^{fix}$ are respectively the predicted saliency map, the groundtruth density distribution and the groundtruth binary fixation map, while α, β and γ are three scalars which balance the three loss functions. L_1, L_2 and L_3 are respectively the Normalized Scanpath Saliency (NSS), the Linear Correlation Coefficient (CC) and the Kullback-Leibler Divergence (KL-Div) which are commonly used to evaluate saliency prediction models.

The NSS metric was defined specifically for the evaluation of saliency models [62]. The idea is to quantify the saliency map values at the eye fixation locations and to normalize it with the saliency map variance:

$$L_1\left(\tilde{\mathbf{y}}, \mathbf{y}^{fix}\right) = \frac{1}{N} \sum_i \frac{\tilde{\mathbf{y}}_i - \mu(\tilde{\mathbf{y}})}{\sigma(\tilde{\mathbf{y}})} \cdot \mathbf{y}_i^{fix} \tag{12}$$

where i indexes the i^{th} pixel, $N = \sum_i \mathbf{y}_i^{fix}$ is the total number of fixated pixels and $\tilde{\mathbf{y}}$ is normalized to have a zero mean and unit standard deviation.

The CC, instead, is the Pearson's correlation coefficient and treats the saliency and groundtruth density maps, $\tilde{\mathbf{y}}$ and $\mathbf{y}_{den}$, as random variables measuring the linear relationship between them. It is computed as:

$$L_2\left(\tilde{\mathbf{y}}, \mathbf{y}^{den}\right) = \frac{\sigma\left(\tilde{\mathbf{y}}, \mathbf{y}^{den}\right)}{\sigma\left(\tilde{\mathbf{y}}\right) \cdot \sigma\left(\mathbf{y}^{den}\right)} \tag{13}$$

where $\sigma\left(\tilde{\mathbf{y}}, \mathbf{y}^{den}\right)$ is the covariance of $\tilde{\mathbf{y}}$ and $\mathbf{y}^{den}$.

The KL-Div evaluates the loss of information when the distribution $\tilde{\mathbf{y}}$ is used to approximate the distribution $\mathbf{y}^{den}$, therefore taking a probabilistic interpretation of saliency and groundtruth density maps. Formally:

$$L_3\left(\tilde{\mathbf{y}}, \mathbf{y}^{den}\right) = \sum_i \mathbf{y}_i^{den} \log\left(\frac{\mathbf{y}_i^{den}}{\tilde{\mathbf{y}}_i + \epsilon} + \epsilon\right) \tag{14}$$

where i indexes the i^{th} pixel and ϵ is a regularization constant. The KL-Div is a dissimilarity metric and a lower value indicates a better approximation of the groundtruth by the predicted saliency map.

In Section V-A, we quantitatively justify the choice of our loss combination comparing our results with those obtained using single evaluation metrics as loss function. Moreover, we compare the proposed training strategy with several other probability distances used by previous saliency methods demonstrating that our solution is able to achieve a better balance among all evaluation metrics.

IV. Experimental Setup

In this section we describe datasets and metrics used to evaluate the proposed model, and provide implementation details.

A. Datasets

For training and testing our model, we use four of the most popular saliency datasets which differ in terms of both image content and experimental settings.

- SALICON [40]: This is the largest available dataset for saliency prediction. It contains 10,000 training images, 5,000 validation images and 5,000 testing images, taken from the Microsoft COCO dataset [63]. Eye fixations are simulated with mouse movements: as shown in [40], there is a high degree of similarity between mouse-contingent saliency annotations and fixations recorded with eye-tracking systems. Groundtruth maps of the test set are not publicly available and predictions must be submitted to the SALICON challenge website[2] for evaluation.
- MIT1003 [9]: The MIT1003 dataset contains 1003 images coming from Flickr and LabelMe. Saliency maps have been created from eye-tracking data of 15 observers.
- MIT300 [64]: The MIT300 dataset is a collection of 300 natural images with saliency maps generated from eye-tracking data of 39 users. Saliency maps of this entire dataset are held out and we used the MIT Saliency benchmark [65] for evaluating our predictions. To test our network on this dataset, we fine-tune it on images of the MIT1003 randomly split in training and validation sets.
- CAT2000 [59]: This dataset contains 4,000 images coming from a large variety of categories such as *Cartoons*, *Art*, *Satellite*, *Low resolution images*, *Indoor*, *Outdoor*, *Line drawings*, ect. It is composed of 20 different categories with 200 images for each of them. Saliency maps of the testing set, composed by 2,000 images, are not available and we submitted our saliency maps to the MIT Saliency benchmark [65] for evaluation.

[2]https://competitions.codalab.org/competitions/3791

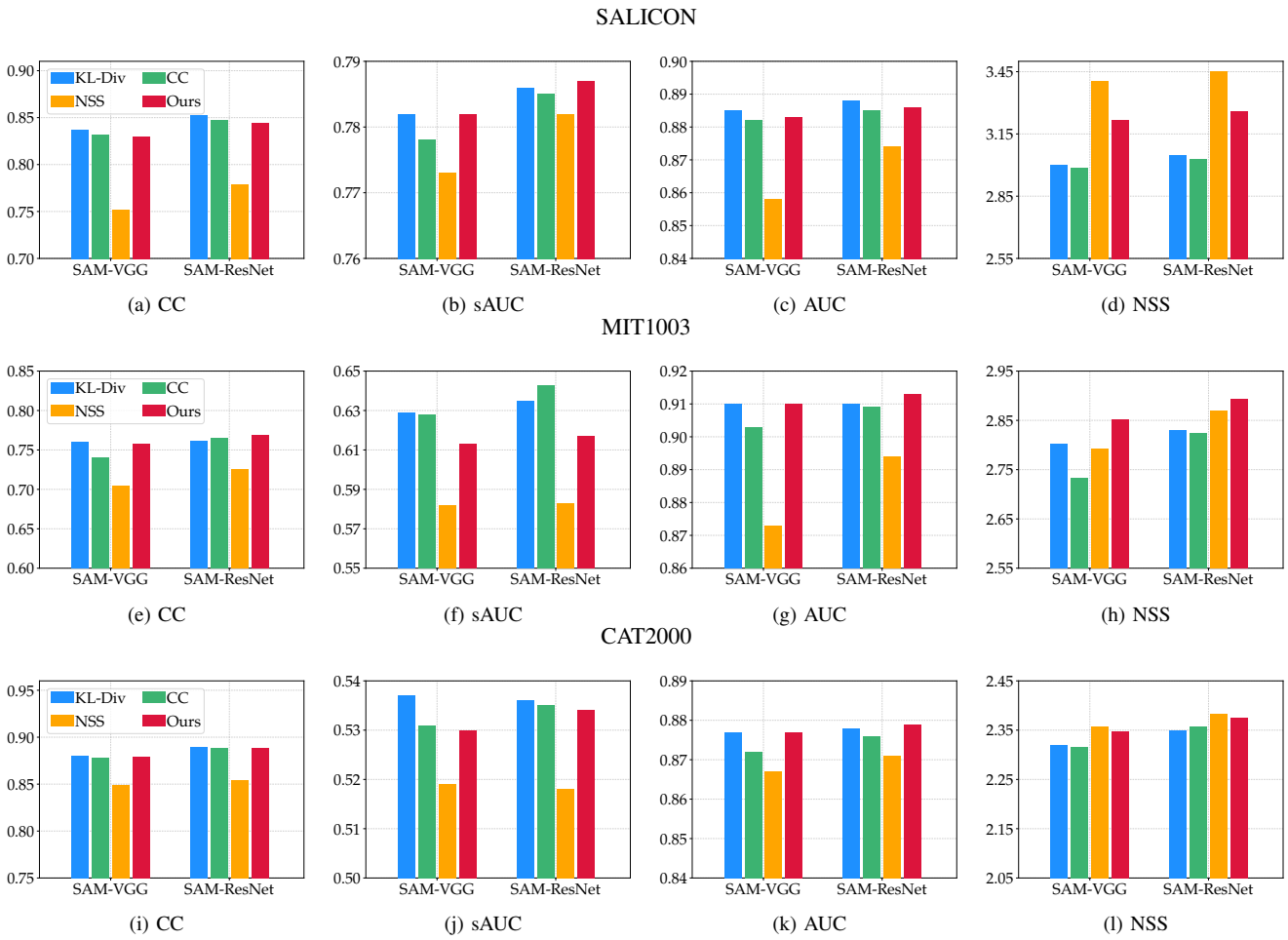


Fig. 5. Comparison between the proposed loss function and its components used individually as loss functions. We report results for both SAM-VGG and SAM-ResNet on SALICON [40] (plots a-d), MIT1003 [9] (plots e-h) and CAT2000 [59] (plots i-l) validation sets. Plots on the same row correspond to a different evaluation metric (CC, sAUC, AUC and NSS). The four color bars represent the loss functions used. As it can be observed, our loss function achieves the best balance between metrics.

B. Evaluation Metrics

A large variety of metrics to evaluate saliency prediction models exist and the main difference between them concerns the ground-truth representation. In fact, saliency evaluation metrics can be categorized in location-based and distribution-based metrics [66], [67], [68]. The first category considers saliency maps at discrete fixation locations, while the second treats both ground-truth fixation maps and predicted saliency maps as continuous distributions.

The most widely used location-based metrics are the Area under the ROC curve, in its different variants of Judd (AUC) and shuffled (sAUC), and the Normalized Scanpath Saliency (NSS). The AUC metrics do not penalize low-valued false positives giving a high score for high-valued predictions placed at fixated locations and ignoring the others. Besides, the sAUC is designed to penalize models that take into account the center bias present in eye fixations. The NSS, instead, is sensitive in an equivalent manner to both false positives and false negatives.

For the distribution-based category, the most used evaluation metrics are the Linear Correlation Coefficient (CC), the Similarity (SIM) and the Earth Mover Distance (EMD). The CC treats both false positives and false negatives symmetrically, differently from the SIM that instead measures the intersection between two distributions and for this reason it is very sensitive to missing values. The EMD is a dissimilarity metric that penalizes false positives proportionally to the spatial distance from the groundtruth.

C. Implementation Details

We evaluate our model on SALICON, MIT300 and CAT2000 datasets. For the first dataset, we train the network on its training set and we use the $5,000$ validation images to validate the model. For the second and the third dataset, we pre-train the network on SALICON and then fine-tune on MIT1003 dataset and CAT2000 training set respectively, as suggested by the MIT Saliency Benchmark organizers. In particular, to test our model on the MIT300 dataset, we use 903 randomly selected images of the MIT1003 to fine-tune the network and the remaining 100 as validation set. For the CAT2000 dataset, instead, we randomly choose $1,800$ images of training set for the fine-tuning and we use the remaining 200 (10 for each category) as validation set.

For the SALICON, MIT1003 and MIT300 datasets, we resize input images to 240×320. Since images from MIT1003

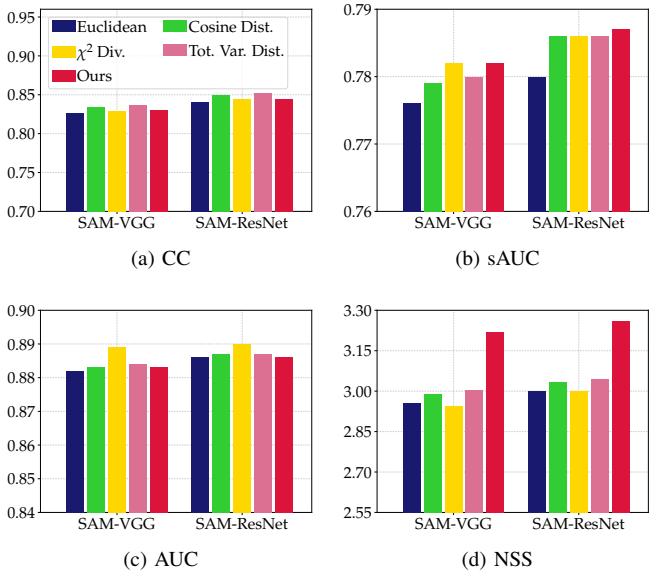


Fig. 6. Comparison between the proposed combination of saliency metrics and more traditional loss functions such as Euclidean Loss, χ^2 Divergence, Cosine Distance and Total Variation Distance. Each plot corresponds to a different evaluation metric (CC, sAUC, AUC and NSS). The five color bars represent the performance of our model trained with the considered loss functions. We report results of both SAM-VGG and SAM-ResNet models on SALICON validation set [40].

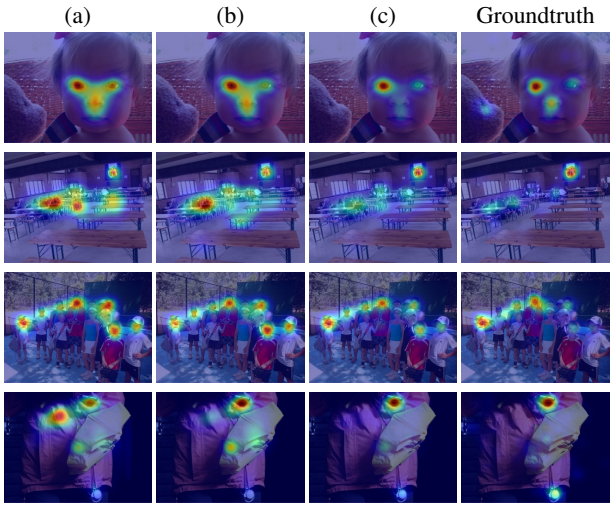


Fig. 7. Examples of saliency maps predicted by the DCN (a), the DCN with the Attentive ConvLSTM (b), and the DCN with the Attentive ConvLSTM and learned priors (c) compared with the groundtruth (d). Images are from SALICON validation set [40].

and MIT300 have different sizes, we apply zero padding bringing images to have an aspect ratio of 4:3 and then resize them to have the selected input size. Instead, images from CAT2000 dataset have all the same input size of 1080×1920. For this reason, we resize all images of this dataset to 180×320.

Predictions of all datasets are slightly blurred with a Gaussian filter. After a validation process, we set the standard deviation of the Gaussian kernel to 7.

Weights of the Dilated Convolutional Networks are initialized with those of the VGG-16 and ResNet-50 models trained on ImageNet [69]. For the Attentive ConvLSTM, following the initialization proposed in [70], we initialize the recurrent weights matrices U_i, U_f, U_o and U_c as random orthogonal matrices. All W matrices and U_a are initialized by sampling each element from the Gaussian distribution of mean 0 and variance 0.05^2. The matrix V_a and all bias vectors are initialized to zero. Weights of all other convolutional layers of our model are initialized according to [71].

At training time, we randomly sample a minibatch containing K training saliency maps, and encourage the network to minimize the proposed loss function through the RMSprop optimizer [72]. We found that a batch size of 10 is sufficient to learn the model seamlessly. Batch normalization is preserved in the ResNet-50 part of the model, and we do not add batch normalization layers elsewhere.

Loss parameters α, β and γ are respectively set to -1, -2 and 10 balancing the contribution of each loss function. Differently from the KL-Div that is a dissimilarity metric and its value should be minimized, the CC and the NSS are to be maximized to predict better saliency maps. To this end, we set α and β as negative weights. The choice of these balance weights is driven by the goal of having good results on all evaluation metrics and by taking into account the numerical range that the single metrics have at convergence.

During the training phase, we set the initial learning rate to 10^{-5} and we decrease it by a factor of 10 every two epochs for the model based on the ResNet, and every three epochs for that based on the VGG network.

V. Experimental Evaluation

In this section we perform analyses and experiments to validate the contribution of each component of the network. We also show quantitative and qualitative comparisons with other state of the art models.

A. Comparison between different loss functions

In Fig. 5 we compare results obtained by using single loss functions (KL-Div, CC, NSS) and our combination proposed in Section III-D. Results are reported for both versions of our model on SALICON, MIT1003 and CAT2000 validations sets. We call SAM-VGG the model based on the VGG network and SAM-ResNet that based on the ResNet network.

As it can be seen, our combined loss achieves on average better results on all metrics. For example on the SALICON dataset, when the model is trained using the KL-Div or the CC metrics as loss function, the performances are good especially on the CC, while the model fails on the NSS. When the model is trained using the NSS metric, instead, it achieves better results only on the NSS and fails on all other metrics. A similar behaviour is also present on the MIT1003 and CAT2000 datasets where the gain in performance obtained by our loss function is particularly evident on the CC, AUC and NSS metrics, even reaching in some cases the best results.

To further validate the effectiveness of the proposed loss function, we compare it with traditional loss functions and probability distances used by other previous saliency models [42], [23], [13]. Fig. 6 shows the comparison between our

TABLE II
ABLATION ANALYSIS OF SAM-VGG AND SAM-RESNET MODELS ON SALICON [40], MIT1003 [9] AND CAT2000 [59] VALIDATION SETS.

Dataset	Model	SAM-VGG				SAM-ResNet			
		CC	sAUC	AUC	NSS	CC	sAUC	AUC	NSS
SALICON	Plain CNN	0.743	0.765	0.870	2.333	0.771	0.762	0.876	2.404
	Dilated Convolutional Network	0.801	**0.786**	0.876	3.122	0.823	0.774	0.879	3.187
	DCN + Attentive ConvLSTM	0.809	0.784	0.878	3.142	0.841	0.786	0.885	3.256
	DCN + Learned Priors	0.824	0.782	0.882	3.209	0.840	0.784	0.885	3.235
	DCN + Attentive ConvLSTM + Learned Priors	**0.830**	0.782	**0.883**	**3.219**	**0.844**	**0.787**	**0.886**	**3.260**
MIT1003	Plain CNN	0.638	**0.625**	0.889	2.147	0.667	**0.631**	0.895	2.255
	Dilated Convolutional Network	0.718	0.596	0.906	2.704	0.748	0.609	0.902	2.845
	DCN + Attentive ConvLSTM	0.749	0.601	0.908	2.812	0.756	0.613	0.912	2.860
	DCN + Learned Priors	0.750	0.621	0.908	2.805	0.746	0.613	0.908	2.816
	DCN + Attentive ConvLSTM + Learned Priors	**0.757**	0.613	**0.910**	**2.852**	**0.768**	0.617	**0.913**	**2.893**
CAT2000	Plain CNN	0.751	0.546	0.862	1.886	0.819	**0.538**	0.870	2.052
	Dilated Convolutional Network	0.791	**0.548**	0.870	2.067	0.881	0.527	0.877	2.368
	DCN + Attentive ConvLSTM	0.851	0.537	0.874	2.253	0.882	0.528	0.878	2.367
	DCN + Learned Priors	0.877	0.532	0.876	2.328	0.885	0.528	0.878	**2.377**
	DCN + Attentive ConvLSTM + Learned Priors	**0.879**	0.530	**0.877**	**2.347**	**0.888**	0.534	**0.879**	2.375

combination of saliency metrics and four other loss functions: the Euclidean loss, the Cosine Distance, the χ^2 Divergence and the Total Variation Distance. Also in this case, our loss function achieves a better balance among all metrics. The gap with respect to all other traditional losses is particularly evident on the NSS metric, while, on all other metrics, the proposed combined loss, if it does not reach the best results, it is very close to them.

Overall, our combined loss reaches competitive results on all metrics differently from the other loss functions. For this reason, results of all following experiments are obtained by training the network with our combination of loss.

B. Model Ablation Analysis

We evaluate the contribution of each component of the architecture, on SALICON, MIT1003 and CAT2000 validation sets. To this end, we construct five different variations: the plain CNN architecture without the last fully convolutional layer (as a baseline), the Dilated Convolutional Network (DCN), the DCN with the proposed ConvLSTM model, the DCN with the proposed learned priors module and the final version of our model with all its components.

Table II shows the results of the ablation analysis using both versions of our model on three different datasets. The results emphasize that the overall architecture is able to predict better saliency maps in both SAM-VGG and SAM-ResNet variants and each proposed component gives an important contribution to the final performance on all considered datasets. In particular, on the SALICON dataset, it can be seen that there is a constant improvement on all metrics. For example, the VGG baseline achieves a result of 0.743 in terms of CC, while the DCN achieves a relative improvement of $\frac{0.801-0.743}{0.743} = 7.8\%$. This result is further improved by 1% when adding the Attentive ConvLSTM or by 2.9% when adding the learned priors. The overall architecture adds an important improvement of 2.6% to the DCN with the Attentive ConvLSTM and 0.7% to the DCN with learned priors. The ResNet baseline, instead, achieves a CC result of 0.771 that is improved by a 6.7% when adding the dilated convolutions. The Attentive ConvLSTM and the learned priors respectively add an improvement of 2.2% and 2.1%. These results are further improved using the overall architecture with all proposed components by 0.4% and 0.5%.

It is also noteworthy that, with our pipeline, a VGG-based network and a ResNet-based network achieve almost the same performance, so one of the two models can be equally chosen according to speed and memory allocation needs, without considerably affecting prediction performance.

Figure 7 shows some qualitative examples of saliency maps predicted by our SAM-ResNet model and by only some of its main components with respect to the groundtruth. As it can be seen, there is a constant improvement of predictions which, by adding our key components, are more qualitatively similar to the groundtruth.

C. Contribution of the attentive model and learned priors

Table IV reports the performance of our model when using the output of the Attentive ConvLSTM module at different timesteps as input for the rest of the model. Results clearly show that the refinement carried out by the Attentive model results in better performance. No further significant improvements were observed for $t > 4$: while CC, sAUC and AUC almost saturated, NSS slightly decreased after four iterations.

To assess the effectiveness of our prior learning strategy, we compare it with the approach in [14], in which a low resolution prior map is learned and applied element-wise to the predicted saliency map, after performing bilinear upsampling. We chose to compare our solution to that in [14] because it is the only other attempt to incorporate the center bias in a deep learning

TABLE III
COMPARISON RESULTS BETWEEN OUR LEARNED PRIORS AND THAT PROPOSED IN [14] ON SALICON [40], MIT1003 [9] AND CAT2000 [59] VALIDATION SETS.

	SALICON				**MIT1003**				**CAT2000**			
	CC	sAUC	AUC	NSS	CC	sAUC	AUC	NSS	CC	sAUC	AUC	NSS
SAM-VGG (prior of [14])	0.811	**0.783**	0.878	3.150	0.738	0.610	0.908	2.754	0.845	**0.539**	0.874	2.233
SAM-VGG (learned priors)	**0.830**	0.782	**0.883**	**3.219**	**0.757**	**0.613**	**0.910**	**2.852**	**0.879**	0.530	**0.877**	**2.347**
SAM-ResNet (prior of [14])	0.840	0.785	0.884	3.249	0.766	0.609	0.912	**2.899**	0.886	0.528	0.878	**2.386**
SAM-ResNet (learned priors)	**0.844**	**0.787**	**0.886**	**3.260**	**0.768**	**0.617**	**0.913**	2.893	**0.888**	**0.534**	**0.879**	2.375

TABLE IV
RESULTS ON SALICON VALIDATION SET [40] WHEN USING THE OUTPUT OF THE ATTENTIVE CONVLSTM MODULE AT DIFFERENT TIMESTEPS AS INPUT OF THE REST OF THE MODEL.

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

model without the use of hand-crafted prior maps. Results on SALICON, MIT1003 and CAT2000 validation sets are reported in Table III. Using multiple Gaussian learned priors, instead of learning an entire prior map, with no pre-defined structure, shows to be beneficial according to all metrics.

D. Comparison with state of the art

We quantitatively compare our method with state of the art models on SALICON, MIT300 and CAT2000 test sets. Not all saliency methods report experimental results on all considered datasets. For this reason, comparison methods are different depending on each dataset. We decide to sort model performances by the NSS metric as suggested by the MIT Saliency Benchmark [65], [67], [68].

Table V shows the results on the SALICON dataset in terms of CC, sAUC, AUC and NSS. As it can be observed, our SAM-ResNet solution outperforms all competitors by a big margin especially on CC and NSS metrics and obtains the best result also on the sAUC. In particular, our method overcomes the other ResNet-based model [24] with an improvement of 1.5% according to NSS metric, 1.3% and 0.4% according to CC and sAUC. For a fair comparison with other methods, we also include the results achieved by our SAM-VGG model. The improvement with respect all other VGG-based methods is even more significant than that obtained by the SAM-ResNet model. In detail, our SAM-VGG overcomes all other VGG-based methods with an improvement of 12.7% and 5.6% according to NSS and CC metrics.

TABLE V
COMPARISON RESULTS ON SALICON TEST SET [40]. THE RESULTS IN BOLD INDICATE THE BEST PERFORMING METHOD ON EACH EVALUATION METRIC. (*) INDICATES CITATIONS TO NON-PEER REVIEWED TEXTS. METHODS ARE SORTED BY THE NSS METRIC.

	CC	sAUC	AUC	NSS
SAM-ResNet	**0.842**	**0.779**	0.883	**3.204**
DSCLRCN [24] (*)	0.831	0.776	0.884	3.157
SAM-VGG	0.825	0.774	0.881	3.143
ML-Net [14]	0.743	0.768	0.866	2.789
MixNet [45] (*)	0.730	0.771	0.861	2.767
SU [42]	0.780	0.760	0.880	2.610
SalGAN [44] (*)	0.781	0.772	0.781	2.459
SalNet [23]	0.622	0.724	0.858	1.859
DeepGazeII [26]	0.509	0.761	**0.885**	1.336

With the proposed model, we have also participated to the LSUN Challenge 2017, where we reached the first place on the saliency prediction task[3].

The results on MIT300 and CAT2000 datasets are respectively reported in Tables VI and VII. Our method achieves state of the art results on all metrics, except for the sAUC, on the CAT2000 dataset surpassing other methods by an important margin especially on SIM, CC, NSS and EMD metrics. On the MIT300 dataset, instead, we obtain results very close to the best ones.

Our model does not obtain a big gain in performance on AUC metrics. This can be explained considering that the AUC metrics are primarily based on true positives without significantly penalizing false positives. For this reason, hazy or blurred saliency maps like the ones predicted by [26] achieve high AUC values [73], [34], despite being visually very different from the groundtruth annotations, as we will show in the following.

Qualitative results obtained by our models on SALICON and MIT1003 validations sets, together with those of other state of the art models, are shown in Figure 8. As it can be noticed, our network is able to predict high saliency values on people, faces, objects and other predominant cues. It also produces good saliency maps when images do not contain strong saliency regions, such as when saliency is concentrated in the center of the scene or when images portray a landscape.

[3]https://competitions.codalab.org/competitions/17136

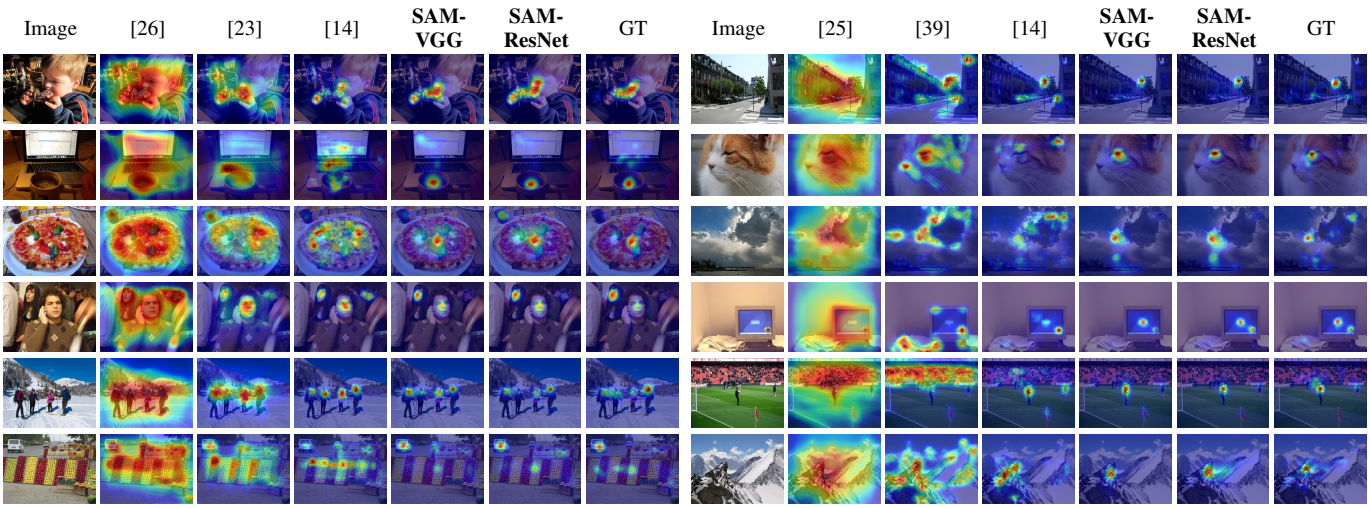


Fig. 8. Qualitative results and comparison with other state of the art models. Left images are from SALICON validation set [40], while right images are from MIT1003 validation set [9].

TABLE VI
COMPARISON RESULTS ON MIT300 DATASET [64]. THE RESULTS IN BOLD INDICATE THE BEST PERFORMING METHOD ON EACH EVALUATION METRIC. (*) INDICATES CITATIONS TO NON-PEER REVIEWED TEXTS. METHODS ARE SORTED BY THE NSS METRIC.

	SIM	CC	sAUC	AUC	NSS	EMD
DSCLRCN [24] (*)	**0.68**	**0.80**	0.72	0.87	**2.35**	2.17
SAM-ResNet	**0.68**	0.78	0.70	0.87	2.34	2.15
SAM-VGG	0.67	0.77	0.71	0.87	2.30	2.14
DeepFix [22]	0.67	0.78	0.71	0.87	2.26	**2.04**
SALICON [12]	0.60	0.74	**0.74**	0.87	2.12	2.62
PDP [13]	0.60	0.70	0.73	0.85	2.05	2.58
ML-Net [14]	0.59	0.67	0.70	0.85	2.05	2.63
SalGAN [44] (*)	0.63	0.73	0.72	0.86	2.04	2.29
iSEEL [43]	0.57	0.65	0.68	0.84	1.78	2.72
SalNet [23]	0.52	0.58	0.69	0.83	1.51	3.31
BMS [10]	0.51	0.55	0.65	0.83	1.41	3.35
Mr-CNN [39]	0.48	0.48	0.69	0.79	1.37	3.71
DeepGazeII [26]	0.46	0.52	0.72	**0.88**	1.29	3.98
GBVS [7]	0.48	0.48	0.63	0.81	1.24	3.51
eDN [25]	0.41	0.45	0.62	0.82	1.14	4.56

TABLE VII
COMPARISON RESULTS ON CAT2000 TEST SET [59]. THE RESULTS IN BOLD INDICATE THE BEST PERFORMING METHOD ON EACH EVALUATION METRIC. (*) INDICATES CITATIONS TO NON-PEER REVIEWED TEXTS. METHODS ARE SORTED BY THE NSS METRIC.

	SIM	CC	sAUC	AUC	NSS	EMD
SAM-ResNet	**0.77**	**0.89**	0.58	**0.88**	**2.38**	**1.04**
SAM-VGG	0.76	**0.89**	0.58	**0.88**	**2.38**	1.07
DeepFix [22]	0.74	0.87	0.58	0.87	2.28	1.15
MixNet [45] (*)	0.66	0.76	0.58	0.86	1.92	1.63
iSEEL [43]	0.62	0.66	**0.59**	0.84	1.67	1.78
BMS [10]	0.61	0.67	**0.59**	0.85	1.67	1.95
eDN [25]	0.52	0.54	0.55	0.85	1.30	2.64
GBVS [7]	0.51	0.50	0.58	0.80	1.23	2.99

We notice, from a qualitative point of view, that the model can sometimes infer the relative importance of different people in the same scene, a human behaviour which saliency models still struggle to replicate, as discussed in [74].

VI. CONCLUSION

We described a novel Saliency Attentive Model which can predict human eye fixations on natural images. The main novelty of the proposal is an Attentive Convolutional LSTM specifically designed to sequentially enhance saliency predictions. The same idea could potentially be employed in other tasks in which an image refinement is profitable. Furthermore, we captured an important property of human gazes by optimally combining multiple learned priors, and effectively addressed the downscaling effect of CNNs. The effectiveness of each component has been validated through extensive evaluation, and we showed that our model achieves state of the art results on two of the most important datasets for saliency prediction. Finally, we contribute to further research efforts by releasing the source code and pre-trained models of our architecture.

ACKNOWLEDGMENT

We thank the organizers of the LSUN Challenge and of the MIT Saliency Benchmark for enabling us to compare with other published approaches.

This work was partially supported by JUMP project, funded by the Emilia-Romagna region within the POR-FESR 2014-2020 program. We acknowledge the CINECA award under the ISCRA initiative, for the availability of high performance computing resources and support. We also gratefully acknowledge the support of Facebook AI Research and NVIDIA Corporation with the donation of the GPUs used for this research.

REFERENCES

[1] R. A. Rensink, "The Dynamic Representation of Scenes," *Visual Cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[2] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch, "Automatic Image Retargeting," in *International Conference on Mobile and Ubiquitous Multimedia*, 2005.

[3] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition - a gentle way," in *International Workshop on Biologically Motivated Computer Vision*, 2002.

[4] H. Hadizadeh and I. V. Bajic, "Saliency-Aware Video Compression," *IEEE Transactions on Image Processing*, vol. 23, no. 1, pp. 19–33, 2014.

[5] V. Mahadevan and N. Vasconcelos, "Biologically Inspired Object Tracking Using Center-Surround Saliency Mechanisms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 3, pp. 541–554, 2013.

[6] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Paying More Attention to Saliency: Image Captioning with Saliency and Context Attention," *ACM Transactions on Multimedia Computing, Communications, and Applications*, vol. 14, no. 2, p. 48, 2018.

[7] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2006.

[8] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[9] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *IEEE International Conference on Computer Vision*, 2009.

[10] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *IEEE International Conference on Computer Vision*, 2013.

[11] M. Kümmerer, L. Theis, and M. Bethge, "DeepGaze I: Boosting saliency prediction with feature maps trained on ImageNet," in *International Conference on Learning Representations Workshops*, 2015.

[12] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks," in *IEEE International Conference on Computer Vision*, 2015.

[13] S. Jetley, N. Murray, and E. Vig, "End-to-End Saliency Mapping via Probability Distribution Prediction," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[14] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "A Deep Multi-Level Network for Saliency Prediction," in *International Conference on Pattern Recognition*, 2016.

[15] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio, "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention," in *International Conference on Machine Learning*, 2015.

[16] K. Gregor, I. Danihelka, A. Graves, D. J. Rezende, and D. Wierstra, "DRAW: A recurrent neural network for image generation," *arXiv preprint arXiv:1502.04623*, 2015.

[17] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention Is All You Need," in *Advances in Neural Information Processing Systems*, 2017.

[18] J. K. Chorowski, D. Bahdanau, D. Serdyuk, K. Cho, and Y. Bengio, "Attention-based models for speech recognition," in *Advances in Neural Information Processing Systems*, 2015.

[19] Z. Li, K. Gavrilyuk, E. Gavves, M. Jain, and C. G. Snoek, "VideoLSTM Convolves, Attends and Flows for Action Recognition," *Computer Vision and Image Understanding*, vol. 166, pp. 41–50, 2018.

[20] K. Simonyan and A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[22] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, 2017.

[23] J. Pan, K. McGuinness, E. Sayrol, 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.

[24] N. Liu and J. Han, "A Deep Spatial Contextual Long-term Recurrent Convolutional Network for Saliency Detection," *arXiv preprint arXiv:1610.01708*, 2016.

[25] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2014.

[26] M. Kümmerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low-and high-level contributions to fixation prediction," in *IEEE International Conference on Computer Vision*, 2017.

[27] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[28] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[29] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of Intelligence*, 1987, pp. 115–141.

[30] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*, 2005.

[31] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision*, vol. 13, no. 4, pp. 11–11, 2013.

[32] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2011.

[33] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in Neural Information Processing Systems*, 2008.

[34] Q. Zhao and C. Koch, "Learning a Saliency Map using Fixated Locations in Natural Scenes," *Journal of Vision*, vol. 11, no. 3, pp. 9–9, 2011.

[35] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2012.

[36] Y. Fang, Z. Chen, W. Lin, and C.-W. Lin, "Saliency detection in the compressed domain for adaptive image retargeting," *IEEE Transactions on Image Processing*, vol. 21, no. 9, pp. 3888–3901, 2012.

[37] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, and C.-W. Lin, "A video saliency detection model in compressed domain," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 24, no. 1, pp. 27–38, 2014.

[38] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," in *Advances in Neural Information Processing Systems*, 2012.

[39] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[40] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[41] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going Deeper with Convolutions," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[42] S. S. Kruthiventi, V. Gudisa, J. 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.

[43] H. R. Tavakoli, A. Borji, J. Laaksonen, and E. Rahtu, "Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features," *Neurocomputing*, vol. 244, pp. 10–18, 2017.

[44] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, "SalGAN: Visual Saliency Prediction with Generative Adversarial Networks," *arXiv preprint arXiv:1701.01081*, 2017.

[45] S. Dodge and L. Karam, "Visual Saliency Prediction Using a Mixture of Deep Neural Networks," *arXiv preprint arXiv:1702.00372*, 2017.

[46] J. Zhang, Z. Lin, J. Brandt, X. Shen, and S. Sclaroff, "Top-down neural attention by excitation backprop," in *European Conference on Computer Vision*, 2016.

[47] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[48] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[49] D. Zhang, J. Han, C. Li, and J. Wang, "Co-saliency detection via looking deep and wide," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[50] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[51] J. Kuen, Z. Wang, and G. Wang, "Recurrent Attentional Networks for Saliency Detection," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[52] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[53] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-Term Recurrent Convolutional Networks for Visual Recognition and Description," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[54] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[55] Q. Wu, P. Wang, C. Shen, A. v. d. Hengel, and A. Dick, "Ask Me Anything: Free-form Visual Question Answering Based on Knowledge from External Sources," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2016.

[56] L. Baraldi, C. Grana, and R. Cucchiara, "Hierarchical Boundary-Aware Neural Encoder for Video Captioning," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2017.

[57] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of Vision*, vol. 7, no. 14, pp. 4–4, 2007.

[58] P.-H. Tseng, R. Carmi, I. G. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of Vision*, vol. 9, no. 7, pp. 4–4, 2009.

[59] A. Borji and L. Itti, "CAT2000: A Large Scale Fixation Dataset for Boosting Saliency Research," in *CVPR Workshops*, 2015.

[60] F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations*, 2016.

[61] K. He and J. Sun, "Convolutional Neural Networks at Constrained Time Cost," in *IEEE International Conference on Computer Vision and Pattern Recognition*, 2015.

[62] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of Bottom-Up Gaze Allocation in Natural Images," *Vision research*, vol. 45, no. 18, pp. 2397–2416, 2005.

[63] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common Objects in Context," in *European Conference on Computer Vision*, 2014.

[64] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *MIT Technical Report*, 2012.

[65] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "MIT Saliency Benchmark," http://saliency.mit.edu/.

[66] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and Human Fixations: State-of-the-Art and Study of Comparison Metrics," in *IEEE International Conference on Computer Vision*, 2013.

[67] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016.

[68] M. Kümmerer, T. S. Wallis, and M. Bethge, "Information-Theoretic Model Comparison Unifies Saliency Metrics," *National Academy of Sciences*, vol. 112, no. 52, pp. 16 054–16 059, 2015.

[69] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[70] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," in *International Conference on Learning Representations*, 2015.

[71] X. Glorot and Y. Bengio, "Understanding the difficulty of training deep feedforward neural networks," in *International Conference on Artificial Intelligence and Statistics*, 2010.

[72] T. Tieleman and G. Hinton, "RMSProp: Divide the gradient by a running average of its recent magnitude," *Coursera Course: Neural Networks for Machine Learning*, 2012.

[73] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of Scores, Datasets, and Models in Visual Saliency Prediction," in *IEEE International Conference on Computer Vision*, 2013.

[74] Z. Bylinskii, A. Recasens, A. Borji, A. Oliva, A. Torralba, and F. Durand, "Where should saliency models look next?" in *European Conference on Computer Vision*, 2016.

Marcella Cornia received the B.Sc. degree in Computer Science and the M.Sc. degree in Computer Engineering from the University of Modena and Reggio Emilia, Modena, Italy. She is currently pursuing the Ph.D. degree at the AImageLab Laboratory at the Department of Engineering "Enzo Ferrari" of the University of Modena and Reggio Emilia. Her research interests include visual saliency prediction, image captioning, and cross-modal retrieval.

Lorenzo Baraldi received the M.Sc. degree in Computer Engineering and the Ph.D. degree cum laude in Information and Communication Technologies from the University of Modena and Reggio Emilia, Modena, Italy, in 2014 and 2018, respectively. He is currently a Research Fellow with the Department of Engineering "Enzo Ferrari", University of Modena and Reggio Emilia. He was a Research Intern at Facebook AI Research (FAIR) in 2017. He has authored or coauthored more than 30 publications in scientific journals and international conference proceedings. His research interests include video understanding, deep learning and multimedia. He regularly serves as a Reviewer for international conferences and journals.

Giuseppe Serra is currently an Assistant Professor with the University of Udine, Udine, Italy. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, and at Telecom ParisTech/ENST, Paris, France, in 2006 and 2010, respectively. He has authored or coauthored more than 80 publications in scientific journals and international conference proceedings. His research interests include egocentric vision, and image and video analysis. Prof. Serra has been an Associate Editor for the IEEE Transactions on Human-Machine Systems since 2017. He was a Technical Program Committee member of several workshops and conferences. He regularly serves as a Reviewer for international conferences and journals such as CVPR and ACM Multimedia.

Rita Cucchiara received the M.Sc. degree in Electronics Engineering and the Ph.D. degree in Computer Engineering from the University of Bologna, Bologna, Italy, in 1989 and 1992, respectively. She is currently a Full Professor of computer engineering with the University of Modena and Reggio Emilia, Modena, Italy. She is the Director of the Research Center Softech-ICT, and heads the AImageLab Laboratory at the University of Modena and Reggio Emilia. She has authored or coauthored more than 350 papers in journals and international proceedings, and is a reviewer for several international journals. She has been a coordinator of several projects in computer vision and pattern recognition, and in particular on video surveillance, human behavior analysis, and video understanding. In the field of multimedia, she works on annotation, retrieval, and human-centered searching in images and video big data for cultural heritage. Prof. Cucchiara is a Member of the ACM, a Member of the IEEE Computer Society, and a Fellow of the IAPR. She is the President of the "Associazione Italiana per la ricerca in Computer Vision, Pattern recognition e machine Learning" (affiliated with IAPR), and a Member of the Advisory Board of the Computer Vision Foundation.