Saliency Unified: A Deep Architecture for simultaneous Eye Fixation Prediction and Salient Object Segmentation

Srinivas S S Kruthiventi, Vennela Gudisa, Jaley H Dholakiya and R. Venkatesh Babu
Video Analytics Lab, Department of Computational and Data Sciences,
Indian Institute of Science, Bangalore, India

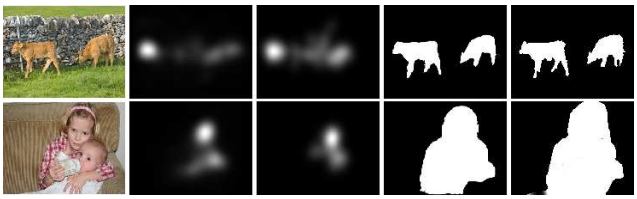
Abstract

Human eye fixations often correlate with locations of salient objects in the scene. However, only a handful of approaches have attempted to simultaneously address the related aspects of eye fixations and object saliency. In this work, we propose a deep convolutional neural network (CNN) capable of predicting eye fixations and segmenting salient objects in a unified framework. We design the initial network layers, shared between both the tasks, such that they capture the object level semantics and the global contextual aspects of saliency, while the deeper layers of the network address task specific aspects. In addition, our network captures saliency at multiple scales via inception-style convolution blocks. Our network shows a significant improvement over the current state-of-the-art for both eye fixation prediction and salient object segmentation across a number of challenging datasets.

1. Introduction

Among the various striking features of the human visual system, the ability to discriminate and selectively pay attention to a few regions in the scene over others, distinctly sets it apart. This phenomenon of selective visual attention has been a topic of interest for researchers in the fields of both neuroscience and computer vision over the past few decades [1, 2]. Modelling this *focus of attention*, also termed as visual saliency, not only gives an insight into human vision, but also has various applications such as image retargetting [3], object recognition [4], visual tracking [5], foveated video compression [6] etc.

Computational models for visual saliency often aim to solve one of the two problems - Predict the locations where an observer will fixate while free-viewing an image; Detect and segment the objects in a scene which grab our immediate attention. Eye fixation locations are considered to be indicative of the bottom-up visual attentional mechanism in humans. Models to predict fixation locations output a saliency map - 'a topographical map representing the conspicuity of each pixel in the image' [7]. The second task of salient object segmentation requires the generation of a pixel-accurate binary map indicating the presence of striking objects in the image. An example image with eye fixation and salient object segmentation maps generated by our model along with the ground-truth maps are shown in Fig. 1.



(a) (b) (c) (d) (e)

Figure 1. Illustrative images (a) with their corresponding eye fixation predictions (b), groundtruth (c) and salient object segmentation predictions (d), groundtruth (e).

Recent studies have shown that the two tasks of eye fixation prediction and salient object segmentation are correlated [8, 9]. Human eye fixations are often found to be guided by the locations of salient objects in the scene. This hypothesis of task correlation is bolstered by the work of Li *et al.* [9] who used a simple eye fixation based model for segmenting salient objects in an image and achieved state-of-the-art results. Nevertheless, only a handful of the works in visual saliency attempt to solve these two problems together [9]. In this work, we propose a deep network model which performs both these tasks simultaneously.

Early approaches for modeling saliency were driven by manually crafting features [10, 11] for over-segmented image regions and estimating their saliency using machine learning or optimization methods. However, recent advances in deep learning and the availability of large datasets have enabled models to perform end-to-end learning. Specifically, the success of Convolutional Neural Networks (CNNs) for various computer vision tasks, has led to a shift in focus from a paradigm of devising innovative features and techniques of combining them, to that of learning complex representations from data directly.

In this work, we propose a deep convolutional architecture for simultaneously predicting the human eye fixations and segmenting salient objects in an image. Our network has a branched architecture, where the shared layers, common to both the tasks, are designed to extract the crucial factors for saliency such as object level semantics and global context. The layers specialized for each of the tasks tap features from these shared layers using multi-scale convolution modules and process them further to obtain the final predictions. Our deep network has been evaluated on multiple datasets for both the tasks and is shown to achieve state-of-the-art performance across multiple metrics.

2. Related Work

In this section, we discuss a few important works in the areas of visual saliency and deep convolutional networks.

2.1. Visual Saliency

The classic work of Itti *et al.* [2] considers low-level features such as color and edge orientation at multiple scales which are combined using a neural network to predict saliency maps. Bruce *et al.* [12] attributed saliency to image patches using the criterion of maximizing the self-information derived from color based features. Another landmark work in saliency, by Harel *et al.* [13], obtained pixel saliency values by calculating equilibrium distribution of Markov chains constructed over image maps generated from low level features.

In addition to low-level features, Judd *et al.* [14] proposed a learning based approach which also uses high-level features from person and face detectors. Borji *et al.* [15] examined the role of additional high-level descriptors such as the presence of text and cars in predicting saliency.

While early saliency works were primarily aimed towards generating saliency maps for predicting eye fixation locations, the works by Liu *et al.* [16] and Achanta *et al.* [17] introduced the notion of object level saliency. These works defined immediate attention-grabbing objects in a scene as salient objects and formulated the problem of salient object segmentation to predict a pixel-accurate binary mask of the salient objects in a given test image. Achanta *et al.* [17] proposed a frequency domain approach for segmenting salient objects using low-level features of color and luminance. Perazzi *et al.* [10] obtained an abstract image representation having homogeneous regions by removing unnecessary details and assigned saliency scores based on the factors of a region's uniqueness and spatial distribution. Segmentation by assigning saliency scores to over-segmented image regions (super-pixels) using various priors (background [18], objectness [19]) has been another popular approach.

There have been only a few works in visual saliency which have explored the relation between eye fixations and salient objects. Mishra *et al.* [20] segmented salient objects using fixation points as identification markers on objects and found an optimal contour around the fixation points. Li *et al.* [9] proposed a salient object segmentation model which used the saliency maps from existing fixation prediction algorithms to classify image regions marked by an object proposal algorithm as a salient object.

2.2. Deep Convolutional Nets

Deep Convolutional Networks, popularized by the seminal work of Krizhevsky *et al.* [21], brought a paradigm shift in vision research from hand-crafting of features to learning them from data. While initial deep networks were aimed towards image classification, they were successfully adapted to pixel-level image tasks like semantic object segmentation [22] and depth estimation [23]. Fully convolutional nets are a particular flavour of CNNs designed to make structured predictions on the image grid. They were first used by Long *et al.* [22] in their work of semantic object segmentation. Long *et al.* converted fully connected layers in existing image classification nets into convolutional layers for obtaining pixel-level predictions. Further, they introduced novel deconvolutional layers for making predictions at the original image resolution. Alternately, Chen *et al.* [24], adopted a simpler approach for retaining spatial resolution to the extent possible, by removing stride in some of the constituent layers of the network. They also introduced convolutional layers with holes which allowed the filters to have receptive fields larger than their kernel sizes. In our network, we use these layers with holes in order to capture the global context of the scene.

Recently, in the realm of salient object segmentation, Zhao *et al.* [25] proposed a multi-context approach using deep convolutional networks. They considered two different networks operating in parallel, over a subsampled and upsampled image patch around each superpixel. The network operating on subsampled image region was considered to capture global context while the network operating on upsampled region captured local context. The output features from the two networks were concatenated for determining the saliency of the corresponding superpixel.

In eye fixation prediction, Liu *et al.* [26] took an approach similar to that of Zhao *et al.* [25] by considering multiple convolutional networks, each operating at a particular scale in the image pyramid representation. This construction of multiple CNNs was termed as Multiresolution-CNN and was found to be efficient at characterizing the low-level and high-level semantics of the image.

In contrast to the works of [25, 26], our model captures the semantic context at various levels efficiently through a single network, by leveraging intermediate representations in the deep feature hierarchy for detecting saliency. The multi-scale aspects of saliency are captured through convolutional kernels of different sizes operating in parallel.

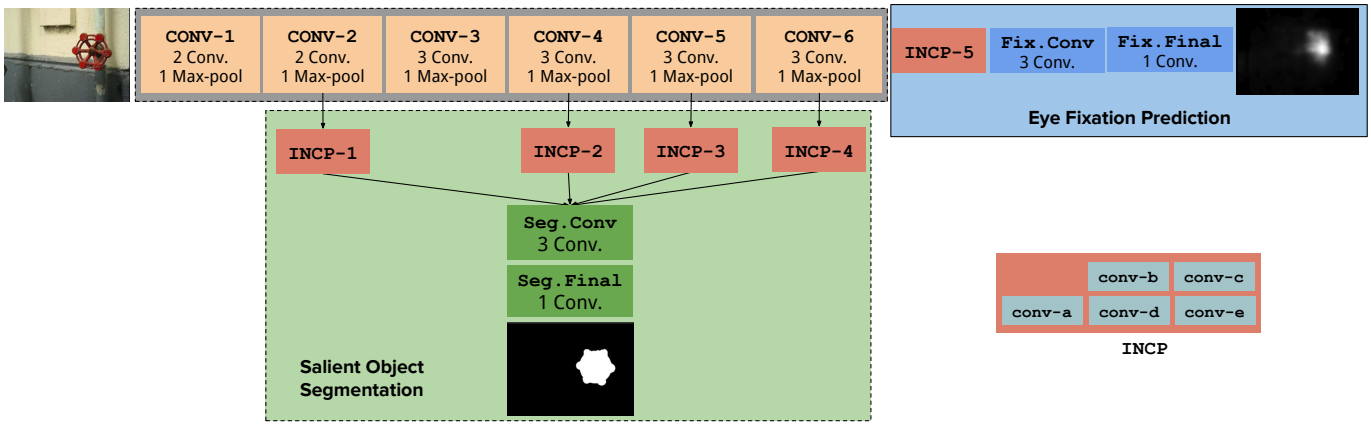


Figure 2. Architecture overview of the proposed network for simultaneously predicting human eye fixations and segmenting salient objects.

3. Network Architecture

We propose a fully convolutional deep network with a branched architecture for simultaneously predicting eye fixations and segmenting salient objects. Layers specialized to these two tasks branch out from a central shared pipeline in the network. This shared pipeline comprises of a series of 6 convolution blocks (shown in gray bounding box in Fig. 2).

Inspired from VGG-16 [27], the layers in first five blocks (`CONV-1` to `CONV-5`) of the shared pipeline have small kernels of spatial size 3×3. Small kernels allow the network to have a very deep architecture with a low memory requirement while making the model more discriminative. All the five convolution blocks (`CONV-1` to `CONV-5`) end with a max-pool layer and every convolutional layer in the network is followed by a ReLU non-linear activation. The architectural details of these 5 convolution blocks are described in Table 1. In the VGG-16 network, the spatial dimensions of the data blob are halved after each block using a stride of 2 in the block's max-pool layer. This strategy of spatially subsampling the data blob is crucial in classification networks for keeping the computational demand low, as data blobs tend to have a large number of channels (usually > 1000) at fully connected layers to cater to the large number of classes. However, for fully convolutional networks, the spatial resolution of the output blob is also important as they are primarily trained for per-pixel recognition tasks. We retain the spatial resolution of the data blob at $1/8$ times that of the original image after the third convolution block (`CONV-3`). We accomplish this by reducing the stride value from 2 to 1 in the max-pool layers of fourth and fifth convolution blocks (`CONV-4`, `CONV-5`).

During training, the first five convolution blocks (`CONV-1` to `CONV-5`) of our network are initialized with weights of the VGG-16 network, which was originally trained over 1.3 million images of the ImageNet [28] dataset. In the VGG-16 net, the filters of the fifth block (`CONV-5`) were trained to operate on a data blob with a resolution of $1/16$ times the original image, unlike our network where the blob has a resolution of $1/8$ times the image. We handle this scale mismatch by introducing holes of size 2 in filters of the fifth block which doubles their receptive field [24, 29]. This allows the convolutional layers in fifth block to operate on the blob at the scale that they were originally trained for. We refer the readers to [24] for a more elaborate description of convolutional layers with holes.

Capturing Global Context: Saliency is the distinctive quality of an entity which makes it stand out from its neighbors and captures our immediate attention [30]. Efficient detection of these salient regions in an image would require the model to capture the global context of the image before assigning scores to its individual regions. To facilitate this, we employ convolutional layers with very large receptive fields in the sixth convolution block (`CONV-6`). Each of the layers in this block operates with filters of kernel size 5 and hole size 5, thus achieving an effective receptive field of 21×21. Similar to the fourth and fifth convolution blocks, the sixth block also ends with a max-pool layer of stride 1.

3.1. Salient Object Segmentation

Salient object segmentation consists of two sub-tasks: detecting the salient objects in the image and determining the spatial extent of the object by identifying its boundaries. While detection of a salient object requires the image regions to be characterized using contextually rich semantic features, the task of finding the object's spatial extent requires lower level semantics like colour, contrast, texture and part composition. These two kinds of features are referred to as global and local contexts in a recent deep saliency work [25], where two different networks are used for extracting them. However, previous studies [31] on deep architectures have shown that early layers in a convolutional network capture low-level image aspects while the later lay-

ers capture high-level semantics. Our model captures features from both the local and global contexts efficiently using this inherent feature hierarchy present in deep networks.

Recently, Hariharan *et al.* [32] have shown that information of interest for pixel-level tasks is spread across all the layers of a convolutional network. They introduced the concept of hypercolumns, defined as the concatenation of features corresponding to a spatial location across all the layers of the deep network. These features were shown to be effective for fine-grained localization tasks. We extract features from the max-pool layers of **CONV-2**, **CONV-4**, **CONV-5** and **CONV-6** blocks for the task of salient object segmentation.

The features from these blocks are tapped using multi-scale convolution kernels and are concatenated together. Li *et al.* [33] and Zhao *et al.* [25] have observed that saliency can be captured better when semantics are considered across multiple scales by upsampling and downsampling image patches. Inspired by the recent success of GoogLeNet [34], we capture this multi-scale semantic information using inception modules. Each inception module operates on its input feature maps with filters of different receptive fields i.e., 1×1, 3×3 and 5×5 capturing information from multiple scales. In order to reduce the computational costs, we replace the usual 5×5 kernel with a 3×3 kernel with 2 holes which will result in an effective receptive field of 5×5. We also reduce the number of channels in the inputs to 3×3 and 5×5 layers in the inception modules, using a 1×1 layer, similar to [34]. Apart from reducing the computational load, these 1×1 layers aid in introducing additional non-linearity.

Using inception modules to extract features from the intermediate layers, we obtain a multi-scale representation of the hierarchical deep features. Also, having multiple pathways via the inception modules (**INCP-1** to **INCP-5**) provides a plausible solution to the issue of vanishing gradients while back-propagating error through the network [34].

The concatenated output from the inception modules is fed to a block (**Seg.Conv**) with three convolutional layers each of a kernel size 3×3. The resulting output is fed to a 1×1 convolutional layer (**Seg.Final**) to predict the object saliency map at a spatial resolution of $1/8$ times the original image.

3.2. Predicting Eye Fixations

The second task of predicting eye fixation saliency maps requires the model to estimate the saliency score for every pixel in a given test image. The ground-truth saliency map for this task is generated by blurring the observer fixation locations on the image with a Gaussian kernel of a constant variance [8]. This blurring is done to take care of the noise in eye tracker equipment and the saccade landing of the observer. These saliency maps generally tend to be blurry, and do not have sharp boundaries unlike the groundtruth of salient object segmentation [8].

Block	Layer	Kernel	Stride	Holes	Output Size
CONV-1	2 conv	3x3	1	1	417x417x64
	max-pool	3x3	2	1	209x209x64
CONV-2	2 conv	3x3	1	1	209x209x128
	max-pool	3x3	2	1	105x105x128
CONV-3	3 conv	3x3	1	1	105x105x256
	max-pool	3x3	2	1	53x53x256
CONV-4	3 conv	3x3	1	1	53x53x512
	max-pool	3x3	1	1	53x53x512
CONV-5	3 conv	3x3	1	2	53x53x512
	max-pool	3x3	1	1	53x53x512
CONV-6	3 conv	5x5	1	5	53x53x512
	max-pool	3x3	1	1	53x53x512
INCP-1	conv a	1x1	2	1	53x53x24
	conv b	1x1	1	1	105x105x36
	conv c	1x1	1	1	105x105x16
	conv d	3x3	2	1	53x53x72
	conv e	3x3	2	2	53x53x32
INCP-2	conv a	1x1	1	1	53x53x24
	conv b	1x1	1	1	53x53x36
	conv c	1x1	1	1	53x53x16
	conv d	3x3	1	1	53x53x72
	conv e	3x3	1	2	53x53x32
INCP-3/ INCP-4/ INCP-5	conv a	1x1	1	1	53x53x32
	conv b	1x1	1	1	53x53x80
	conv c	1x1	1	1	53x53x32
	conv d	3x3	1	1	53x53x160
	conv e	3x3	1	2	53x53x64
Fix.Conv/ Seg.Conv	3 conv	3x3	1	1	53x53x512
Fix.Final/ Seg.Final	1 conv	1x1	1	1	53x53x1

Table 1. Architectural details of the proposed deep convolutional network for segmenting salient objects and predicting eye fixations

The assignment of saliency scores requires characterizing local regions in an image with semantic features while incorporating the global context of the entire scene [29]. The layers in **CONV-6** block, owing to their large receptive fields (21×21), can provide contextually rich semantic features necessary for estimating the saliency score of local image regions. Following the multi-scale approach described earlier for salient object segmentation, we tap the max-pool layer of the **CONV-6** block using an inception module with layers of receptive fields: 1×1, 3×3 and 5×5. The output from this inception module is fed to a block with three convolutional layers (**Fix.Conv**) of spatial size 3×3, which is followed by a 1×1 convolution layer (**Fix.Final**) for estimating the fixation saliency map.

4. Refining Saliency Maps

Our network predicts the saliency maps at a sub-sampled resolution of $1/8$ times the original image resolution.

Ground-truth saliency maps for eye fixations are usually smooth. Hence, we directly interpolate the network's fixation saliency output, using bi-cubic interpolation to obtain the final saliency map.

In the case of salient object segmentation, the ground-truth maps are binary and have sharp edges at the object boundaries. Since the network's salient object predictions are coarse in resolution, we use the fully connected Conditional Random Field (CRF) formulation of Phillip *et al.* [35] to obtain the final pixel-accurate segmentation prediction.

We construct a dense graph on the image grid at its original resolution by considering each pixel as a node. The unary costs for a node to take the labels - *salient* and *background* are defined using the network's object saliency map prediction. The object saliency map is bicubic interpolated to the original image resolution and transformed using a sigmoid activation to obtain a pixel-level saliency scores. This saliency score of a pixel is assumed to be the unary cost for the corresponding graph node to take the *background* label. The additive inverse of this object saliency map is taken to be the unary cost for the *salient* label.

We use the pair-wise formulation of Phillip *et al.* [35] for defining the pair-wise cost between two nodes. This formulation connects every pixel in the image to every other pixel with an edge resulting in a densely connected graph. The pair-wise cost for two nodes taking different labels is defined as a function of the corresponding pixels' color similarity and spatial proximity. Specifically, the pair-wise cost $\psi(i,j)$ for two nodes i, j is defined as

$$\psi(i,j) = \overbrace{w_a exp\left(-\frac{|p_i-p_j|^2}{\sigma_{ap}^2}-\frac{|I_i-I_j|^2}{\sigma_{ai}^2}\right)}^{\text{appearance kernel}} + \overbrace{w_s\, exp\left(-\frac{|p_i-p_j|^2}{\sigma_{sp}^2}\right)}^{\text{smoothness kernel}}$$

Here w_a, w_s indicate the relative weights and $\sigma_{ap}, \sigma_{ai}, \sigma_{sp}$ are the standard deviation values of the Gaussian kernels in the appearance and smoothness terms. p_i, p_j are the position vectors and I_i, I_j are the RGB vectors of the pixels i, j.

The overall energy to be minimized, which is a combination of unary $\phi_u(i)$ and pair-wise $\psi_p(i,j)$ terms, can be expressed as

$$E_G = \left(\sum_{\forall i \in V_G} \phi_u(i) + \sum_{\forall (i,j) \in C_G} \psi_p(i,j)\right)$$

where V_G and C_G denote the nodes and edges in the constructed dense graph respectively.

Every pixel is binary classified into the labels of *salient* and *background* by minimizing the above energy using the approach of mean-field approximation [35].

5. Experimental Evaluation

5.1. Network Training

We train the proposed network on MSRA10K [17] and SALICON [38] datasets. MSRA10K dataset comprises of $10,000$ images picked from a variety of scenarios - natural scenes, animals, indoor, outdoor, etc. Each of these images is provided with a pixel-accurate ground truth binary mask indicating the salient object and is used for training the network for segmentation task. SALICON is a saliency dataset with $15,000$ images where eye fixation annotations are simulated through mouse movements of users on blurred images. The authors of [38] show that the mouse-contingent saliency annotations strongly correlate with actual eye-tracker annotations. We use the SALICON dataset for training the network to predict eye fixations .

For training the network, we use a mini-batch of 8 images in each iteration, 4 of which have segmentation ground truth and the rest have fixation ground truth. `CONV-1` to `CONV-6` blocks, as shown in Fig. 2, are shared between both the tasks of salient object segmentation and eye fixation prediction and are trained for both the tasks simultaneously using all the images in the batch. The layers of the network specialized to each of the tasks are trained using only those images in the batch having the corresponding ground-truth.

The first five convolution blocks (`CONV-1` to `CONV-5`) in the network are initialized from the weights of VGG-16 [27]. The weights in all the other convolutional layers and inception blocks are initialized from zero mean Gaussian with a standard deviation of 0.01 and the biases are set to 0. The layers in (`CONV-1` to `CONV-5`), whose weights are initialized from VGG-16, are trained with a learning rate of 5×10^{-8} while the rest of the layers in the network are trained with a higher learning rate of 5×10^{-7}.

Before feeding the input images and ground-truth maps to the network, we scale the images such that the larger dimension is 417 and zero pad along the smaller dimension to bring the image to a fixed size of 417×417 pixels. The network is trained using stochastic gradient descent with a momentum of 0.9. The entire training procedure takes about 1 day for completion on Nvidia TITAN X GPU with deeplab version [24] of caffe deep learning framework [39].

5.2. Datasets and Evaluation

We evaluate the proposed approach on multiple datasets with images from a wide-variety of scenarios and varying resolutions, number of objects and level of background clutter. A good model should perform well consistently, on most of the datasets. PASCAL-S [9], DUT-OMRON [40], iCoSeg [41] and ECSSD [42] datasets consisting of 850, 5168, 643 and 1000 images respectively are used for evaluating the model on the task of salient object segmentation. PASCAL-S [9], DUT-OMRON [40], MIT1003 [14]

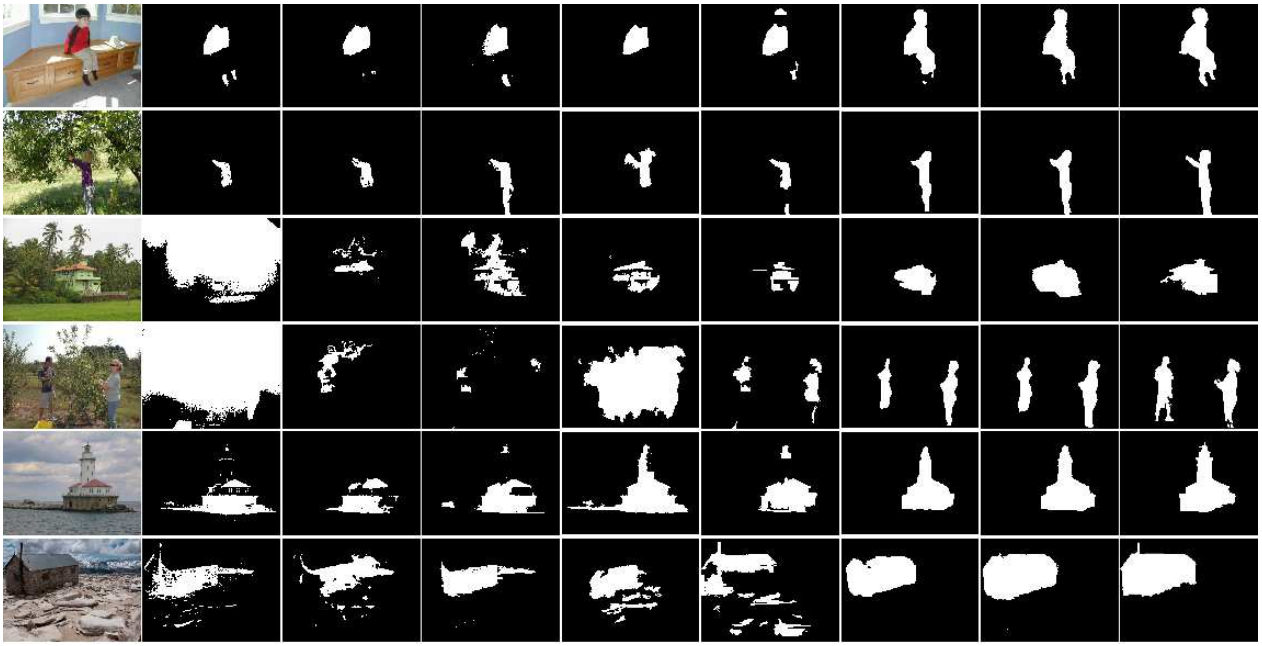


Figure 3. Qualitative results of our approach along with other state-of-the-art methods for salient object segmentation. Proposed (CNN) refers to the results from the CNN alone, whereas the Proposed (CNN+CRF) refers to the final binary segmentation results obtained after refining the maps obtained from CNN using CRF. The object saliency maps of other state-of-the-art methods and Proposed (CNN) are thresholded such that their F_β^w values are maximized.

and IS [43] datasets consisting of 850, 5168, 1003 and 235 images respectively are used for evaluating fixation prediction.

We used Mean Absolute Error (MAE) and Weighted F_β-Measure to evaluate the performance of our network for salient object segmentation. Earth Mover's Distance (EMD), Normalized Scanpath Saliency (NSS) and the shuffled-Area Under Curve (s-AUC) were used for evaluating the performance on eye-fixation prediction. We briefly describe each of these metrics in the following section.

5.2.1 Salient Object Segmentation

Mean Absolute Error (MAE) : MAE is computed as the mean of pixel-wise absolute difference between the continuous object saliency map and the binary ground-truth .

Weighted F_β-Measure ($\mathbf{F_\beta^w}$) : Weighted F_β-measure [44] evaluates a binarized map with respect to ground truth based on weighted precision and recall values. It combines these two values into a single number by taking their weighted harmonic mean. Similar to other works, we consider $\beta^2 = 0.3$, thereby giving more importance to precision.

For binarizing the object saliency map to obtain the salient object segmentation, we follow the procedure described in [45]. Initially, the saliency map is binarized by thresholding at various intermediate values in the range of [0 255] and the F_β^w is computed for each of them. We use the mean and maximal F_β^w values to evaluate the salient object segmentation capabilities of a model.

5.2.2 Predicting Eye Fixations

Earth Mover's Distance (EMD) : EMD considers the ground-truth and predicted saliency maps to be two probability distributions and measures the cost of transforming one distribution to the other.

Normalized Scanpath Saliency (NSS) : Normalized Scanpath Saliency is the average of the response values at eye fixation locations in a model's saliency map, normalized to have zero mean and unit standard deviation.

shuffled - Area Under Curve (s-AUC) : sAUC is the area under the ROC curve of true positives vs. false positives during the binary classification of fixation and non-fixation points using saliency map at various thresholds. The non-fixation points are taken from fixations on other images in the dataset to tackle the issue of centre-bias in eye fixations.

5.3. Results

5.3.1 Salient Object Segmentation

The quantitative results obtained by the proposed method on PASCAL-S, DUT-OMRON, iCoSeg and ECSSD datasets for salient object segmentation are shown in Table 2. We compare our model against the methods – SF [10], PCA [36], DRFI [11], SMD [37] and MDF [33]with respect to the metrics discussed in Sec. 5.2.1.

As evident from the table, our model achieves state of the art results on all the datasets across these metrics. The

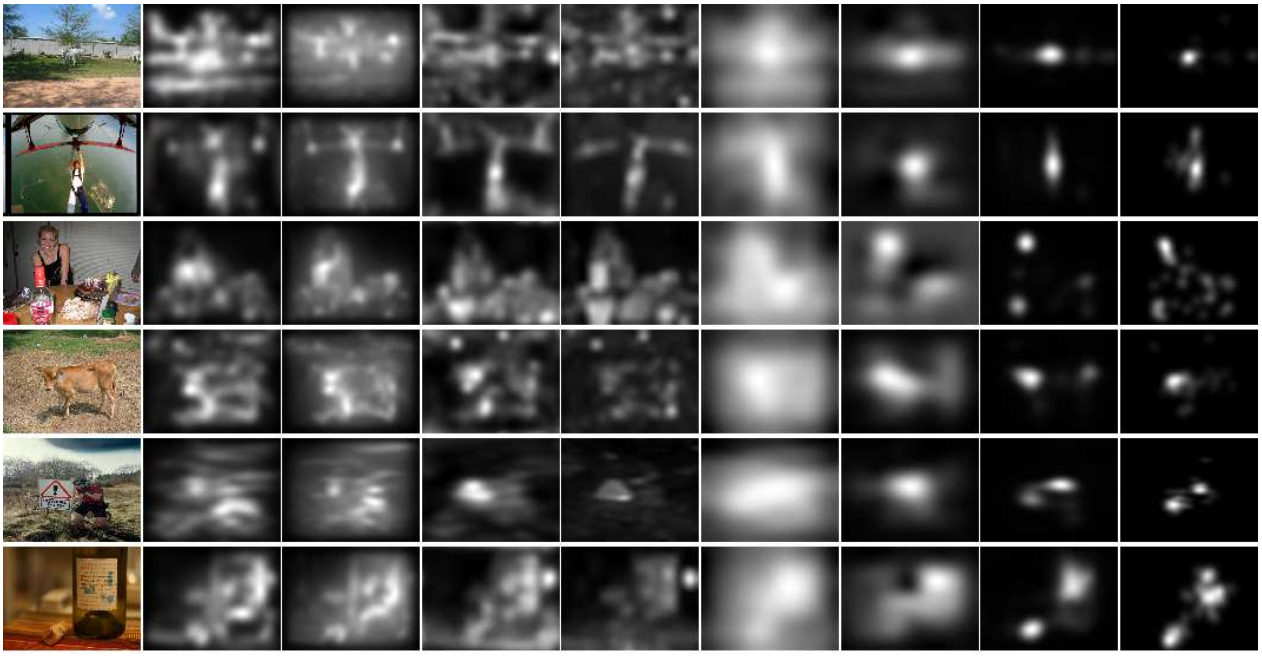


Image ITTI[2] GBVS[13] AWS[46] BMS[47] eDN[48] Junting[49] **Proposed** GT

Figure 4. Qualitative results of our approach along with other state-of-the art methods for eye fixation prediction.

qualitative results for salient object segmentation are shown in Fig. 3. As shown in the figure, our method performs well in a variety of challenging cases, e.g., multiple disconnected objects (fourth row), low contrast between object and background (second row), cluttered background (sixth row). We can also see that our method captures local features like edges and boundaries quite well (fifth row) compared to other methods. In the first image, while most of the existing methods fail to predict the person's legs as salient, our model correctly identifies the entire person as salient.

5.3.2 Predicting Eye-Fixations

The quantitative results obtained by the proposed method on PASCAL-S, DUT-OMRON, MIT1003 and IS datasets for eye-fixation prediction are shown in Table 3. We quantify our results in terms of the previously discussed EMD, NSS and s-AUC metrics. Results illustrate that our method outperforms existing methods with respect to NSS and EMD by a huge margin and achieves state of art results with respect to s-AUC as well.

The qualitative results for eye fixation prediction are shown in Fig. 4. The proposed network is able to detect saliency arising from faces of both humans (third row) and animals (fourth row) efficiently. From the images in fifth and sixth rows, we also note that our model can correctly detect the text and sign boards as salient. Our network also captures multiple salient objects (third row) and weighs their relative importance in the scene appropriately.

In addition to the above four datasets, we further evaluate our model on another large-scale test set – SALICON which comprises of 5000 test images. We obtain the metric scores on this dataset by submitting our fixation predictions to the SALICON challenge website. Results, shown in Table. 4 illustrate that our method outperforms the winner of LSUN Saliency Challenge 2015, JuntingNet [49, 50], a CNN based model, by a significant margin.

For both the tasks of eye fixation prediction and salient object segmentation, our model has been evaluated in a cross-dataset manner i.e., the train and test sets have been taken from different datasets. Inspite of this, our model performs consistently well across various metrics on both the tasks highlighting its generalizability.

5.3.3 Simultaneous versus Independent Training

In order to understand the effect of simultaneously training the network, we train relevant parts of the network independently for the tasks of salient object segmentation and eye fixation prediction. We compare the results from these independently trained models with the proposed simultaneously trained model on DUT-OMRON dataset. Results, shown in Table. 5, illustrate that simultaneously training the network retains the performance of independently trained models across MAE, sAUC metrics while giving a small improvement across F_β^w, EMD, NSS metrics. Also, a model which can simultaneously predict eye fixations and segment salient objects is computationally efficient compared to independent models as the former shares the low-level feature computations for both tasks.

Dataset	Metric	SF [10]	PCA [36]	DRFI [11]	SMD [37]	MDF [33]	**Proposed (CNN)**	**Proposed (CNN+CRF)**
PASCAL-S [9]	MAE ↓	0.26	0.25	–	0.21	0.15	0.12	0.10
	mean F_β^w ↑	0.33	0.35	–	0.49	0.64	0.75	0.77
	max F_β^w ↑	0.43	0.49	–	0.56	0.68	0.78	0.77
DUT-OMRON [40]	MAE ↓	0.27	0.21	0.14	0.17	0.09	0.09	0.07
	mean F_β^w ↑	0.30	0.34	0.48	0.45	0.59	0.65	0.68
	max F_β^w ↑	0.42	0.46	0.58	0.53	0.61	0.69	0.68
iCoSeg [41]	MAE ↓	0.25	0.20	0.14	0.14	0.10	0.10	0.08
	mean F_β^w ↑	0.36	0.42	0.58	0.61	0.67	0.76	0.79
	max F_β^w ↑	0.54	0.58	0.67	0.68	0.74	0.77	0.79
ECSSD [42]	MAE ↓	0.29	0.25	0.16	0.17	0.11	0.08	0.06
	mean F_β^w ↑	0.33	0.39	0.59	0.58	0.74	0.85	0.88
	max F_β^w ↑	0.46	0.54	0.71	0.67	0.77	0.87	0.88

Table 2. Quantitative results of our approach on salient object segmentation compared against other state-of-the art methods on PASCAL-S, DUT-OMRON, iCoSeg and ECSSD datasets. Proposed (CNN) refers to the results using the CNN alone, whereas the Proposed (CNN+CRF) refers to the final results obtained after refining the CNN's output using CRF. The best results are shown in red and the second best in blue. ↑ indicates higher scores on the metric are better and ↓ indicates lower scores on the metric are better.

Dataset	Metric	ITTI [2]	GBVS [13]	AWS [46]	BMS [47]	eDN [48]	MrCNN [26]	JuntingNet [49]	**Proposed**
PASCAL-S [9]	s-AUC ↑	0.64	0.65	0.67	0.67	0.65	–	0.69	0.72
	EMD ↓	1.21	1.16	1.38	1.32	1.29	–	1.03	0.73
	NSS ↑	1.30	1.36	1.12	1.28	1.42	–	1.90	2.22
DUT-OMRON [40]	s-AUC ↑	0.78	0.81	0.78	0.79	0.80	–	0.83	0.83
	EMD ↓	1.47	1.32	1.62	1.58	1.56	–	1.37	1.03
	NSS ↑	1.54	1.71	1.51	1.66	1.33	–	2.03	3.02
MIT1003 [14]	s-AUC ↑	0.66	0.66	0.69	0.69	0.66	0.71	0.68	0.73
	EMD ↓	2.33	2.19	2.54	2.40	2.39	2.30	1.91	1.49
	NSS ↑	1.06	1.17	1.07	1.19	1.24	1.28	1.60	2.08
IS [43]	s-AUC ↑	0.66	0.67	0.72	0.71	0.61	–	0.65	0.70
	EMD ↓	1.30	1.22	1.49	1.43	1.49	–	1.11	0.77
	NSS ↑	1.50	1.58	1.58	1.74	1.27	–	1.72	2.30

Table 3. Quantitative results of our approach on eye fixation prediction compared against other state-of-the-art methods on PASCAL-S, DUT-OMRON, MIT1003 and IS datasets. The best results are shown in red and the second best in blue.

LSUN Saliency Challenge 2015 - SALICON				
Method	s-AUC↑	CC↑	AUC-Borji↑	NSS↑
Proposed	0.76	0.78	0.88	2.61
JuntingNet [49]	0.67	0.60	0.83	–

Table 4. Quantitative results of our approach on SALICON Test set compared against JuntingNet - the winner of LSUN 2015 Saliency Challenge. The best results are shown in red.

Simultaneous vs. Independent Training					
	Segmentation		Fixation		
Method	MAE ↓	F_β^w ↑	sAUC ↑	EMD ↓	NSS ↑
Simul.	0.07	0.68	0.83	1.03	3.02
Indp.	0.07	0.67	0.83	1.07	2.80

Table 5. Quantitative Results on DUT-OMRON dataset when the networks are trained simultaneously versus independently for the tasks of eye fixation prediction and salient object segmentation. The best results are shown in red.

6. Conclusion

In this work, we have proposed a novel deep convolutional architecture capable of simultaneously predicting human eye fixations and segmenting the salient objects in an image. Our network captures the global context, which is crucial for saliency, through layers with large receptive fields and handles multi-scale aspects of saliency using inception modules. Also, our network has a branched architecture to efficiently capture both the low-level and high-level semantics necessary for salient object segmentation.

We evaluate our method on four datasets of eye fixation prediction and salient object segmentation and show that it outperforms the existing state-of-the-art approaches.

7. Acknowledgements

This work was supported by Defence Research and Development Organization (DRDO), Government of India. We thank Nvidia for their hardware grant and Google for the travel grant.

References

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

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

[3] V. Setlur, S. Takagi, R. Raskar, M. Gleicher, and B. Gooch, "Automatic image retargeting," in *ACM MUM*, 2005.

[4] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *NIPS*, 2004.

[5] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *CVPR*, 2009.

[6] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *Image Processing, IEEE Transactions on*, vol. 13, no. 10, pp. 1304–1318, 2004.

[7] E. Niebur, "Saliency map," *Scholarpedia*, vol. 2, no. 8, 2007.

[8] A. Borji, "What is a salient object? a dataset and a baseline model for salient object detection," *Image Processing, IEEE Transactions on*, vol. 24, no. 2, pp. 742–756, 2015.

[9] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014.

[10] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012.

[11] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013.

[12] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *NIPS*, 2005.

[13] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NIPS*, 2006.

[14] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *ICCV*, 2009.

[15] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *CVPR*, 2012.

[16] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 2, pp. 353–367, 2011.

[17] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009.

[18] C. Sheth and R. V. Babu, "Object saliency using a background prior," in *ICASSP*, 2016.

[19] S. S. R and R. V. Babu, "Salient object detection via objectness measure," in *ICIP*, 2015.

[20] A. K. Mishra, Y. Aloimonos, L.-F. Cheong, A. Kassim, *et al.*, "Active visual segmentation," *Pattern Analysis and Machine Intelligence, IEEE Transactions On*, vol. 34, no. 4, pp. 639–653, 2012.

[21] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *NIPS*, 2012.

[22] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[23] D. Eigen, C. Puhrsch, and R. Fergus, "Depth map prediction from a single image using a multi-scale deep network," in *NIPS*, 2014.

[24] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected CRFs," *arXiv preprint arXiv:1412.7062*, 2014.

[25] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015.

[26] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *CVPR*, 2015.

[27] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[28] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, *et al.*, "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, pp. 1–42.

[29] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," *arXiv preprint arXiv:1510.02927*, 2015.

[30] L. Itti, "Visual salience," *Scholarpedia*, vol. 2, no. 9, p. 3327, 2007.

[31] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman, "Return of the devil in the details: Delving deep into convolutional nets," *arXiv preprint arXiv:1405.3531*, 2014.

[32] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *CVPR*, 2015.

[33] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," *arXiv preprint arXiv:1503.08663*, 2015.

[34] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *arXiv preprint arXiv:1409.4842*, 2014.

[35] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected CRFs with gaussian edge potentials," *arXiv preprint arXiv:1210.5644*, 2012.

[36] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?," in *CVPR*, 2013.

[37] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. Maybank, "Salient object detection via structured matrix decomposition," *Technical Report*, pp. 1–14, 2015.

[38] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *CVPR*, 2015.

[39] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," *arXiv preprint arXiv:1408.5093*, 2014.

[40] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013.

[41] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, "iCoSeg: Interactive co-segmentation with intelligent scribble guidance," in *CVPR*, 2010.

[42] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *CVPR*, 2013.

[43] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 35, no. 4, pp. 996–1010, 2013.

[44] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps," in *CVPR*, 2014.

[45] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *ECCV*, 2012.

[46] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *Journal of vision*, vol. 12, no. 6, p. 17, 2012.

[47] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *ICCV*, 2013.

[48] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *CVPR*, 2014.

[49] J. Pan and X. Giró-i Nieto, "End-to-end convolutional network for saliency prediction," *arXiv:1507.01422*, 2015.

[50] J. Pan, K. McGuinness, E. Sayrol, N. O'Connor, and X. Giro-i Nieto, "Shallow and deep convolutional networks for saliency prediction," in *CVPR*, 2016.