DeepFix: A Fully Convolutional Neural Network for Predicting Human Eye Fixations

***Abstract*— Understanding and predicting the human visual attention mechanism is an active area of research in the fields of neuroscience and computer vision. In this paper, we propose DeepFix, a fully convolutional neural network, which models the bottom–up mechanism of visual attention via saliency prediction. Unlike classical works, which characterize the saliency map using various hand-crafted features, our model automatically learns features in a hierarchical fashion and predicts the saliency map in an end-to-end manner. DeepFix is designed to capture semantics at multiple scales while taking global context into account, by using network layers with very large receptive fields. Generally, fully convolutional nets are spatially invariant—this prevents them from modeling location-dependent patterns (e.g., centre-bias). Our network handles this by incorporating a novel location-biased convolutional layer. We evaluate our model on multiple challenging saliency data sets and show that it achieves the state-of-the-art results.**

***Index Terms*— Saliency prediction, eye fixations, convolutional neural network, deep learning.**

I. Introduction

IDENTIFYING conspicuous stimuli in the visual field is a key attentional mechanism in humans. While free viewing, our eyes tend to fixate on regions of the scene which have distinctive variations in visual stimuli such as a bright colour, unique texture or more complex semantic aspects such as presence of a familiar face or any sudden movements. This mechanism guides our eye gaze to the salient and informative regions in the scene.

The human visual system is dictated by two kinds of attentional mechanisms: bottom-up and top-down [1]. Bottom-up factors, which are derived entirely from the visual scene, are responsible for the automatic deployment of attention towards discriminative regions in the scene. The involuntary detection of a red coloured *STOP* sign on the road, while driving, is an example of this attentional mechanism. This kind of attention is automatic, reflexive and stimulus-driven. On the contrary, the top-down attention mechanism is driven by internal factors such as subject's prior knowledge, expectations and the task at hand, making it situational and highly subjective [2]. It uses information available in the working memory, thereby biasing attention towards areas of the scene important to the current behavioral goals [3]. The selective attention exhibited by a hungry animal while searching for its camouflaged prey is an example of the top-down mechanism.

In this work, we propose an approach for modeling the bottom-up visual attentional mechanism by predicting human eye fixations on images. This modeling, commonly referred to as visual saliency prediction, is a classic research area in the field of computer vision and neuroscience [4], [5]. This modeling has applications in vision-related tasks such as video compression [6], object and action recognition [7], [8], image retargetting [9] and surveillance systems [10]. In the past, many computational models have been developed to predict human eye fixations in the form of a saliency map- "a topographically arranged two dimensional map that represents the visual saliency of a scene" [11].

Saliency in a visual scene can arise from a spectrum of stimuli, both low-level (color/intensity, orientation, size etc.) and high-level (faces, text etc.). Most of the classic saliency models [5], [12] are biologically inspired and use multi-scale low-level visual features such as color and texture. However, these methods do not adequately capture the high level semantic aspects of a scene that can contribute to visual saliency. The wide variety of possible causes, both low-level and high-level, make it difficult to hand-craft good features for predicting saliency. This makes deep networks, which are capable of learning features from data in a task dependent manner, a natural choice for this problem.

Recently, deep networks have shown impressive results on a diverse set of perceptual tasks such as speech recognition [13], natural language processing [14] and object recognition [15]. The ability of deep neural networks to automatically learn complex patterns from data in a hierarchical fashion makes them applicable to a wide range of problems with different modalities of data. Though neural networks are being used in the field of artificial intelligence since many decades, their recent wide applicability can be attributed to the increased computational power of GPUs, efficient techniques for training [16] and the availability of very large datasets [17], [18].

In this work, we propose a fully convolutional neural network - DeepFix, for predicting human eye fixations on images in the form of a saliency map. Our model, inspired from VGG net [19], is a very deep network with 20 convolutional layers, each of a small kernel size, operating in succession on an image. The network is designed to capture object-level semantics, which can occur at multiple scales, efficiently through inception style [20] convolution blocks. Each inception module consists of a set of convolution layers

Manuscript received August 6, 2016; revised April 9, 2017; accepted May 12, 2017. Date of publication June 1, 2017; date of current version July 6, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Tolga Tasdizen. *(Corresponding author: Srinivas S. S. Kruthiventi.)*

S. S. S. Kruthiventi and R. V. Babu are with the Video Analytics Lab, Department of Computational and Data Sciences, Indian Institute of Science, Bengaluru 560012, India (e-mail: kssaisrinivas@gmail.com; venky@cds.iisc.ac.in).

K. Ayush is with the Indian Institute of Technology Kharagpur, Kharagpur 721302, India.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2710620

with different kernel sizes operating in parallel. The global context of the scene, which is crucial for saliency prediction, is captured using convolutional layers with very large receptive fields. These layers are placed towards the end of the network and replace the densely connected inner product layers commonly present in convolutional nets.

Fully Convolutional Nets (FCNs) [21], in general, are location invariant i.e., a spatial shift in the input results only in a corresponding spatial shift of the output without affecting its values. This property prevents FCNs from learning location specific patterns such as the centre-bias. The proposed DeepFix model is designed to handle this through a novel Location Biased Convolutional (LBC) layer.

Overall, our model predicts the saliency map from the image in an end-to-end manner, without relying on any hand-crafted features. Here, we summarize the key aspects of our DeepFix network:

- Large depth - to enable the extraction of complex semantic features
- Kernels of different sizes operating in parallel - to characterize the object semantics simultaneously at multiple scales
- Kernels with large receptive fields - for capturing the global context
- Location biased convolutional layers - for learning location dependent patterns such as the centre-bias present in eye fixations

We evaluate the proposed network on multiple challenging saliency datasets MIT300 [22], CAT2000 [23], PASCAL-S [24], OSIE [25], FIGRIM [26], SALICON Test Set [27], iSUN Test Set and show that it achieves state-of-the-art results on these datasets.

II. Related Work

Treisman *et al.*, in their seminal work of Feature Integration Theory (FIT) [4], proposed that preliminary features from visual stimuli are simultaneously processed in different areas of the brain resulting in multiple feature maps. These feature maps are later aggregated to aid in object recognition. Using these principles of human vision, Koch and Ullman [28] first proposed a biologically plausible computational architecture for modelling these early feature representations to artificially simulate the selective attention mechanism in humans. Later, Itti *et al.* [5], building upon the work of [28], implemented a novel system for visual saliency prediction in images. Their model extracts bottom-up visual features of color, intensity and orientation using various blocks. These features are integrated using a dynamic neural network, to construct a two-dimensional representation, called a saliency map which indicates the conspicuity of every pixel in the image. Experimentally, this model was shown to be reasonably successful in detecting centre-surround saliency in images. However, generalization of this model for saliency prediction in complex scenes is difficult because of the primitive nature of its features and the multiple parameters used for constructing them at various scales. Recently, several other works have employed more complex features maps such as isocentric curvedness [29], regional histograms [30], depth cues [31] etc. for estimating saliency. Also, Erdem and Erdem [32] have proposed a non-linear feature integration approach for saliency prediction, using regional covariance features [33].

While the works discussed above are mainly driven by results from neuroscience and psychology, there have also been works which are motivated from an information theory perspective [34]. Oliva *et al.* [35] have taken a top-down approach wherein the statistical rarity of local features across the scene becomes a crucial factor for a region to be salient. Bruce and Tsotsos [36] have explored an information theoretic approach, where the self-information of local image content is used in predicting attention allocation. More recently, Khatoonabadi *et al.* [37] have proposed an approach to predict saliency in videos based on the number of bits needed to encode a video patch by an optimal encoder.

Recent progress in saliency prediction has mostly been driven by the advances in deep learning. Convolutional Neural Networks (CNN), whose design is motivated by the functioning of cells in visual cortex of primates, can capture semantically rich visual features in a hierarchical fashion. While some of these works extract features from pre-trained CNNs [38], a few others have trained their networks specifically for saliency prediction [39]–[41].

In contrast to the traditional usage of multi-scale hand crafted image features, Kümmerer *et al.* [38] have proposed an approach of using feature representations from AlexNet [15] trained for object classification, to perform saliency prediction. Further, the extensive experimental evaluation conducted by Kümmerer *et al.* highlights the importance of feature representations from deeper layers of a CNN in saliency prediction. Though CNN representations are usually generalizable between vision tasks (e.g., object classification to saliency prediction) [42], our work emphasises that a task-specific convolutional neural network with a larger depth, trained in an end-to-end manner, can outperform approaches using off-the-shelf CNN features by a large margin.

A recent work by Vig *et al.* [40] proposes a method for obtaining optimal features by performing a large scale search over different feature-generating model configurations. Each model is considered to be a small convolutional neural network with a maximum of 3 layers, to constrain the computational complexity of overall system. However, this model can not efficiently leverage the semantic feature extracting capabilities of CNNs because of the small depth ($<= 3$) of the individual feature extractors.

Liu *et al.* [41] construct an ensemble of CNNs, termed as Multiresolution-CNN, for predicting eye fixations. Each of these CNNs is trained to classify image patches, at a particular scale, for saliency. This approach of predicting saliency with multiple scale-specific CNNs is efficient for capturing both the low-level and high-level aspects of an image. However, since this approach presents isolated image patches to the network, it fails to capture the global context, which is often crucial for saliency. The proposed DeepFix network overcomes this by operating on the image as a whole and capturing the semantics at multiple scales through its inception style convolution blocks.

Another recent work, SALICON [43], also follows a multi-stream approach for saliency prediction with the network's objective function specifically designed for saliency. This work uses two VGG streams, each specializing for a particular scale. This increases computation since each image has to pass through two networks. In contrast, DeepFix performs a single pass and efficiently obtains multi-scale representation using inception modules.

This approach of processing the image at different scales using multiple network streams has also been utilized for the related task of salient object segmentation [44]. This task aims to identify objects which are salient in the overall scene and segment them out from the background [45], [46]. In our recent work [47], the approach of capturing semantics at multiple scales and extracting contextually rich features is shown to be effective for both predicting eye fixations and segmenting salient objects in the image as well.

III. Proposed Approach

A. DeepFix Architecture

DeepFix is a fully convolutional neural network, trained to predict pixel-wise saliency values for a given image in an end-to-end manner. The key features of the proposed CNN architecture are described below:

1) The network takes an image of size $W \times H \times 3$ (RGB image) as input. This is followed by a series of 5 convolution blocks (① to ⑤), depicted as dashed brown boxes in Fig. 1.
2) Similar to the VGG-16 net [19], the first two blocks (①, ②) contain two convolutional layers each, while the next three (③, ④, ⑤) have three convolutional layers. Each convolution filter in these five blocks is restricted to a kernel size of 3×3 and operates with a stride of 1. This allows the network to have a large depth without increasing the memory requirement. All convolutional layers in the network are followed by Rectified Linear Unit (ReLU) activation to introduce element wise non-linearity.
3) Additionally, each of the first four blocks (① to ④) have a max-pool layer (of size 3×3) following the convolutional layers. Apart from introducing local translational invariance in its output, max-pooling (with stride) reduces computation for deeper layers while preserving the important activations [48].
   Starting from the first convolutional block (①), the number of channels in the outputs of successive blocks gradually increase as 64, 128, 256, 512, depicted as numbers over dotted lines in Fig. 1. This enables the net to progressively learn richer semantic representations of the input image. However, to limit the overall blob size, the spatial dimensions of the blob are halved with every increase in the blob's depth. This is achieved by introducing a stride of 2 in the max-pool layers of the first 3 blocks. This results in an output blob of spatial dimensions $\frac{W}{8} \times \frac{H}{8}$ after the third block. These spatial dimensions are retained in all the further layers.

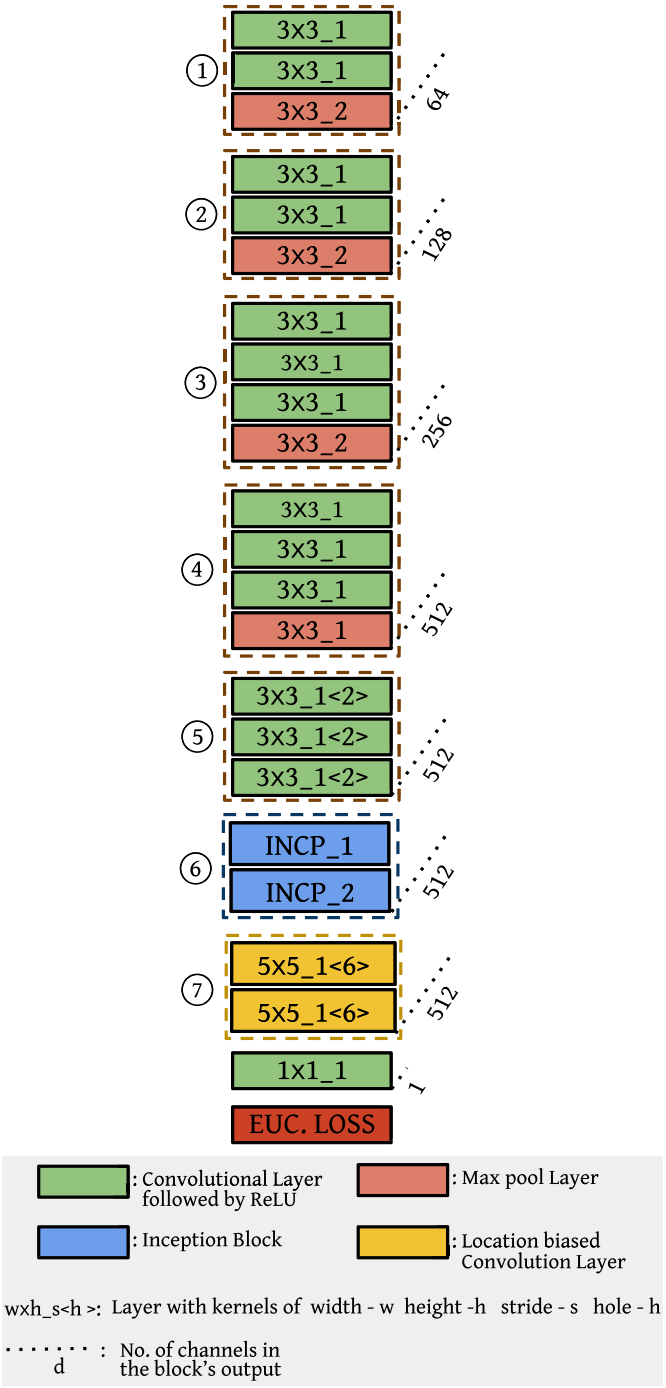


Fig. 1. Architecture of the proposed DeepFix model.

4) While training, the weights of the filters in the five convolution blocks (① to ⑤) of DeepFix are initialized from the VGG-16 net. The weights of VGG-16 net have been learnt by training on 1.3 million images of the ImageNet [17] database. Initializing the weights from network trained on such a large corpus of images is observed to be important for stable and effective learning.
   However, in the VGG-16 net, the spatial dimensions of the output blob are halved at the end of each convolution block, including the fourth and fifth blocks. In our network, to allow for the filters in the fifth block (⑤) to

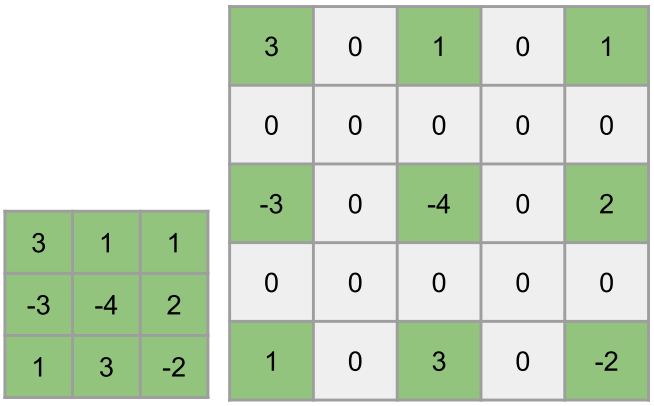


Fig. 2. Convolution kernel with holes. These kernels enable the layer to have a greater receptive field without increasing the memory footprint.

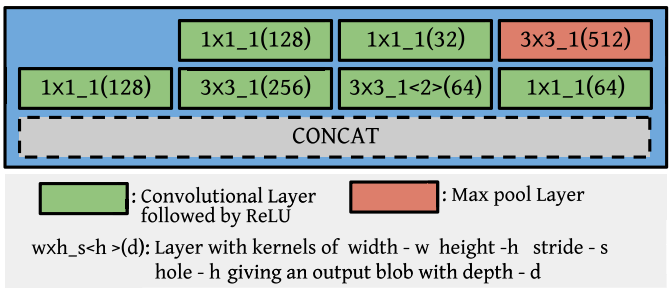


Fig. 3. Constituent layers of the Inception module used in DeepFix network.

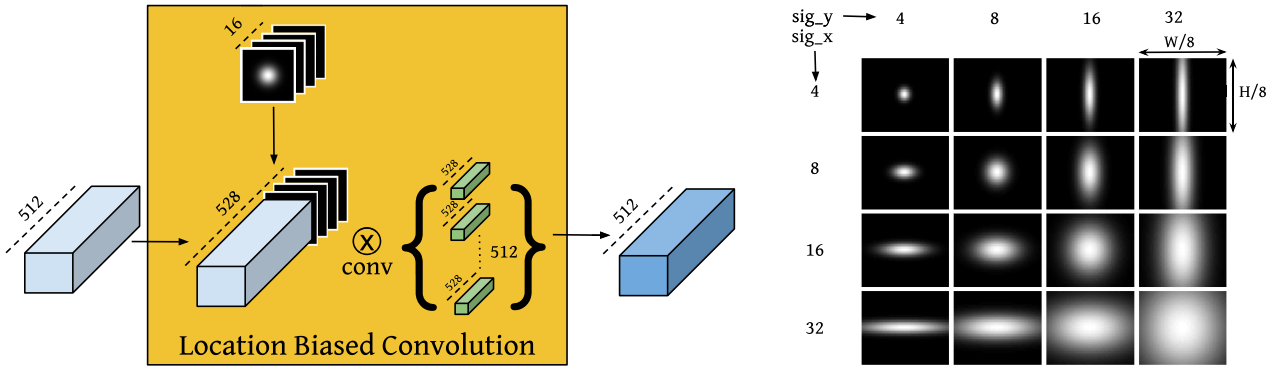
operate on the same receptive field they are originally trained for, we introduce holes of size 2 in their kernels. Convolution filters with holes are illustrated in Fig. 2. The convolution filters of the fifth block, with kernel size 3×3 and hole size 2 have a receptive field of 5×5. Chen *et al.* [49] follow a similar procedure of introducing holes in filters to facilitate weight initialization in their work of semantic image segmentation.

5) While the initial convolution blocks extract low-level image cues such as colour, contrast, texture, etc., the feature maps obtained from the fifth block are shown to characterize high-level semantic information [50]. Previous works have shown that saliency is best captured when semantics are considered from multiple scales [5], [51]. Inspired by the recent success of GoogLeNet [20], we capture the multi-scale semantic structure using two inception style convolution modules (in ⑥), illustrated in Fig. 3.
   Each inception module operates on its input feature map with filters of multiple sizes, thereby capturing the multi-scale information. In the proposed inception module, we use convolutional layers of two sizes : 1×1, 3×3. Unlike the inception module of GoogLeNet, the 5×5 convolutional layer is simulated through a 3×3 layer with holes of size 2. The number of channels in previous layer's output are brought down using 1×1 layers before feeding it to these layers. Additionally, a parallel max-pool layer is added in this inception module to bring in local invariance and it is followed by a 1×1 convolutional layer.
6) Saliency, in neuroscience literature, is often characterized as the unique quality of an entity by which it stands out with respect to its neighbours [52]. This property can be best captured when the local semantic features of an image region are examined in the context of its neighborhood. To facilitate this, the convolutional layers (in ⑦) following the inception modules are designed to have very large receptive fields by introducing holes of size 6 in their convolution kernels. As shown in Fig. 2, these filters with holes can operate on input regions larger than their actual kernel size without increasing the memory footprint. These two layers have a receptive field of 25×25 on their respective inputs.
   In addition, the convolution operation in these two layers is made location dependent to model the centre-bias observed in eye fixations. We term these layers as *Location Biased Convolutional* (LBC) layers and are explained in detail in Sec. III-B. We observe that these two layers are effective at capturing the global context of the image and have resulted in a significant performance improvement.
   To avoid over-fitting and to make the model more general, we have introduced drop-out in the output of the second LBC layer. We have chosen a dropout rate of 0.5 [53].
7) Finally, the output from the second LBC layer (in (⑦) is fed to a 1×1 convolutional layer whose output is taken to be the predicted saliency map. The obtained map has a spatial resolution of $\frac{W}{8} \times \frac{H}{8}$ due to the greater-than-unit stride present in the max-pool layers of the first three blocks. We upsample this map to the original image resolution using bicubic interpolation.

B. Centre-Bias in Eye Fixations

Statistically, it has been observed that a significant number of the human eye fixations fall in the central region of an image. This tendency of humans to gaze at the centre while free-viewing is termed as centre-bias in eye fixations and has been studied extensively in the fields of neuroscience and psychology [54], [55]. This phenomenon is often explained through photographer's bias - the innate tendency of photographers to capture the object of interest by positioning it at the centre of the view. This is found to result in a secondary effect where viewers, after repeatedly viewing such images with the photographer's bias, expect that the most informative content of an image is likely to be present at the centre [55]. This guides their attention involuntarily towards central region of an image even in the absence of the photographer's bias. This secondary effect is described in the literature as bias due to viewing strategy. Also, Borji *et al.*, in their work on CAT2000 [23], experimentally observed that uninterestingness of images could result in eye fixations towards the image centre. These findings suggest that eye fixation patterns are an outcome of both the underlying stimulus and its location.

To account for the spatial biases present in the human eye fixations, some works in the past have employed an explicit centre-bias term in their saliency prediction



(a) (b)

Fig. 4. (a). Location Biased Convolution Filter for learning location dependent patterns in data (e.g., centre-bias present in the eye-fixations). The usual bias term associated with convolutional layers and the ReLU activation are assumed to be present and are not shown explicitly in the above diagram. (b). Gaussian blobs with different horizontal and vertical variances concatenated to the input blob of LBC layers to make the layer's response location specific.

models [38], [40], [56]. In contrast, we design the DeepFix architecture to *learn* location-dependent patterns implicitly through a novel Location Biased Convolutional (LBC) layer. This layer learns to add centre-bias to saliency maps in an image-dependent manner within the CNN framework instead of adding a constant centre-bias to saliency predictions of all images. Next, we describe the construction of LBC layers.

Location Biased Convolutional (LBC) layer: The constituent layers of a fully convolutional net (Convolutional, ReLU and Max-pool layers) are location invariant i.e., a spatial shift of input image will only result in an equivalent scaled spatial shift in the output while keeping the values nearly intact. This property of fully convolutional nets prevents them from learning any location-dependent patterns such as the centre-bias. We tackle this problem by introducing location dependency in the two convolutional layers following the inception modules of the proposed DeepFix architecture.

Let $\boldsymbol{W_c}$ represent weights from c^{th} filter in a convolutional layer and b_c represent its bias. Let the feature vector at spatial location (x, y) in the input blob to this layer be $\mathbf{I}(x, y)$ and the c^{th} filter's response be $R_c(x, y)$. This convolution operation can be represented as

$$R_c(x,y) = \mathcal{R}\Big(\sum_{i,j}\Big(\mathbf{I}(x+i, y+j) \bullet \boldsymbol{W_c}(i,j) + b_c\Big)\Big) \quad (1)$$

where $\bullet$ represents dot product and $\mathcal{R}$ represents ReLU non-linearity.

Here, the weights $\boldsymbol{W_c}$ and the bias b_c are completely independent of the location (x, y) at which they operate, making the convolutional operation location invariant. Introducing spatial dependency directly by making the filter weights a function of the spatial coordinates will increase the number of layer parameters dramatically (proportional to the product of the spatial dimensions of input blob). Location specific convolution where kernels are a function of location can work well for scenarios where the appearance at each location is consistent. For e.g., in face recognition, the location of eyes, nose etc. stay consistent across registered images and filters looking for a specific pattern at a given location might be useful. However, no such consistency of appearance holds for saliency. This also goes against the principle of weight sharing in CNNs which is considered to be an important reason for their effectiveness in visual recognition. We address this problem by concatenating a data independent and location specific feature $\mathbf{L}(x, y)$ to the existing input feature $\mathbf{I}(x, y)$. This results only in a minimal increase in the number of layer parameters (i.e., linear with the dimension of the $\mathbf{L}(x, y)$) and is independent of the input blob's spatial size.

$$R_c(x,y) = \mathcal{R}\Big(\sum_{i,j}\Big(\mathbf{I}(x+i, y+j) \bullet \boldsymbol{W_c}(i,j) + \mathbf{L}(x+i, y+j) \bullet \boldsymbol{W}'_c(i,j) + b_c\Big)\Big) \quad (2)$$

While the location specific features, $\mathbf{L}(x, y)$, remain constant through the entire training procedure, the weights of a filter operating on it, $\boldsymbol{W}'_c$, are learnt over training. This enables the network to optimally combine input stimuli with its location information for predicting the final saliency map. We choose the location-specific feature $\mathbf{L}(x, y)$ to be 16 dimensional with each component giving different weights to the central region, each obtained from a Gaussian of a specific horizontal and vertical variance. The 16 constant maps from which these location specific features are obtained are shown in Fig. 4 (b).

IV. Experimental Evaluation

A. Training DeepFix

The parameters of the first five convolution blocks are initialized from the VGG-16 net [19]. The weights in the convolutional layers of the inception modules and the two LBC layers following them are drawn from a Gaussian distribution with zero mean and standard deviation of 0.01, and the biases are set to 0. The weights of the last convolutional layer are also drawn from a zero mean Gaussian distribution but with a standard deviation of 10 and the bias is set to 0.

During training, all the layers in the five convolution blocks, whose weights are initialized from VGG-16, are learnt with

an initial learning rate of 2×10^{-4}. The remaining layers, whose weights are randomly initialized from Gaussian, are assigned a higher learning rate of 2×10^{-3}. We scale down the learning rates of all the layers by a factor of 5 whenever the performance saturates on the validation set. The network parameters are learned by back-propagating the Euclidean loss of the predicted saliency map with respect to the ground truth saliency using Stochastic Gradient Descent (SGD) with momentum. The network is trained with a momentum of 0.9 and a weight decay of 0.0005.

We train our network in two stages. In the first stage, we use a mouse contingency based saliency dataset - SALICON [27] for training. Though these saliency maps do not correspond to actual eye fixations, this dataset contains 15000 images, from various indoor and outdoor scenarios, providing a rich description of the problem to the CNN. In the second stage, we train the network using smaller datasets, with ground-truth saliency maps generated from actual eye fixation data. We evaluate our network by testing on the datasets of CAT2000 [23], MIT300 [22], PASCAL-S [24], OSIE [25] and FIGRIM [26] datasets. The network used for testing on CAT2000 is trained in the second stage with CAT2000 train set while the network used for testing on MIT300, PASCAL-S, OSIE, FIGRIM is trained with images from MIT1003 [57] dataset. The entire training procedure takes about 1 day on a K40 GPU with the deeplab version [49] of caffe deep learning framework [58].

The ground-truth saliency maps for these datasets are generated by marking the eye-fixation locations for an image on a blank canvas as points and convolving this binary canvas using a Gaussian filter. This convolution operation smears the fixation points to its surroundings and results in a continuous saliency map. This smoothing operation is done to account for the uncertainty in measurements of eye-tracker equipment and eye-fixation landing [59]. The sigma for this Gaussian smoothing is generally chosen to be 1 degree of visual angle (dva) [60]. The dva values for MIT Benchmark datasets - MIT300 [22] and CAT2000 [23] are 35 and 38 pixels respectively. Now, we will briefly describe the 7 saliency datasets used during the training and testing phases of the DeepFix network.

SALICON: SALIency in CONtext (SALICON) dataset [27] contains $10,000$ training images, $5,000$ validation images and $5,000$ test images for saliency prediction. The authors of SALICON [27] propose mouse-contingent-tracking on multi-resolution images as an effective replacement for eye-contingent-tracking in saliency map generation. Further, they show, both qualitatively and quantitatively, that a high degree of similarity exists between the two. With their method, a large mouse-tracking based saliency dataset of 20,000 images has been created from MS COCO [61]. By far, this is the largest selective attention dataset with images from varying context [27]. In our work, we have used $15,000$ images ($10,000$ training images+$5,000$ validation images) during the first stage training of the DeepFix.

1) CAT2000: This dataset [23] contains 4000 images selected from a wide variety of image categories such as *Cartoons, Art, Satellite, Low resolution images, Indoor, Outdoor, Jumbled, Random, and Line drawings* etc.. Overall, this dataset contains 20 different categories with 200 images from each category. The saliency maps for 2000 images (100 from each category) are released as a part of the train set while the saliency maps for the rest of the 2000 images are held-out and they form the test set. From the set of 2000 train images, we have used 1800 images (90 images from each category) during the second stage training of the CNN while the remaining 200 images (10 images from each category) are used for validation. After the second stage training on 1800 images, the proposed method is evaluated on the CAT2000 test set using the MIT saliency benchmark [62].

2) MIT1003: MIT1003 [57] dataset is a collection of 1003 random images from Flickr and LabelMe whose saliency maps are generated using eye tracking data from fifteen users. We use 900 of these images for the second stage of training (for evaluating on MIT300) and the remaining 103 images for validation.

3) MIT300: This dataset [22] contains 300 natural images from both indoor and outdoor scenarios. The ground-truth for this entire dataset is held-out and we use this for evaluating the proposed DeepFix model using the MIT saliency benchmark [62].

4) PASCAL-S: PASCAL-S [24] consists of 850 natural images picked from the validation set of PASCAL VOC 2010. We evaluate our DeepFix model (trained with MIT1003 images in the second stage) on this dataset to test generalizability.

5) OSIE: OSIE [25] consists of 700 natural images with eye tracking annotations from 15 viewers.

6) FIGRIM: FIGRIM [26] consists of 630 images with eye tracking annotations of 16 viewers per image on average.

B. Evaluation

We have evaluated the performance of our network on the MIT Saliency Benchmark [62] with held-out test sets of MIT300, CAT2000 and also on SALICON-Test, iSUN-Test, PASCAL-S, OSIE and FIGRIM datasets. The MIT benchmark evaluates models on a variety of metrics, namely Earth Mover's Distance (EMD), Normalized Scanpath Saliency (NSS), Similarity, Linear Correlation Coefficient (CC), AUC-Judd, AUC-Borji, shuffled-AUC. Previous studies on saliency metrics by Zhang and Sclaroff [63] and Riche *et al.* [64] show that evaluating saliency models by any one of these metrics does not result in a fair comparison which can reflect the qualitative results. Here, we briefly describe these evaluation metrics:

Let G denote the ground-truth saliency map of an image and S be the map estimated using a saliency prediction model.

1) Earth Mover's Distance (EMD) : EMD is a measure of the distance between the two 2D maps, G and S. It is the minimal cost of transforming the probability distribution of the estimated saliency map S to that of the ground truth map G. Therefore, the lesser the EMD score, the better is the estimated saliency map.

2) Normalized Scanpath Saliency (NSS): Normalized Scanpath Saliency is a metric specifically introduced for saliency map evaluaiton by Peters and Itti [65]. This metric is calculated by taking the mean of scores assigned by the

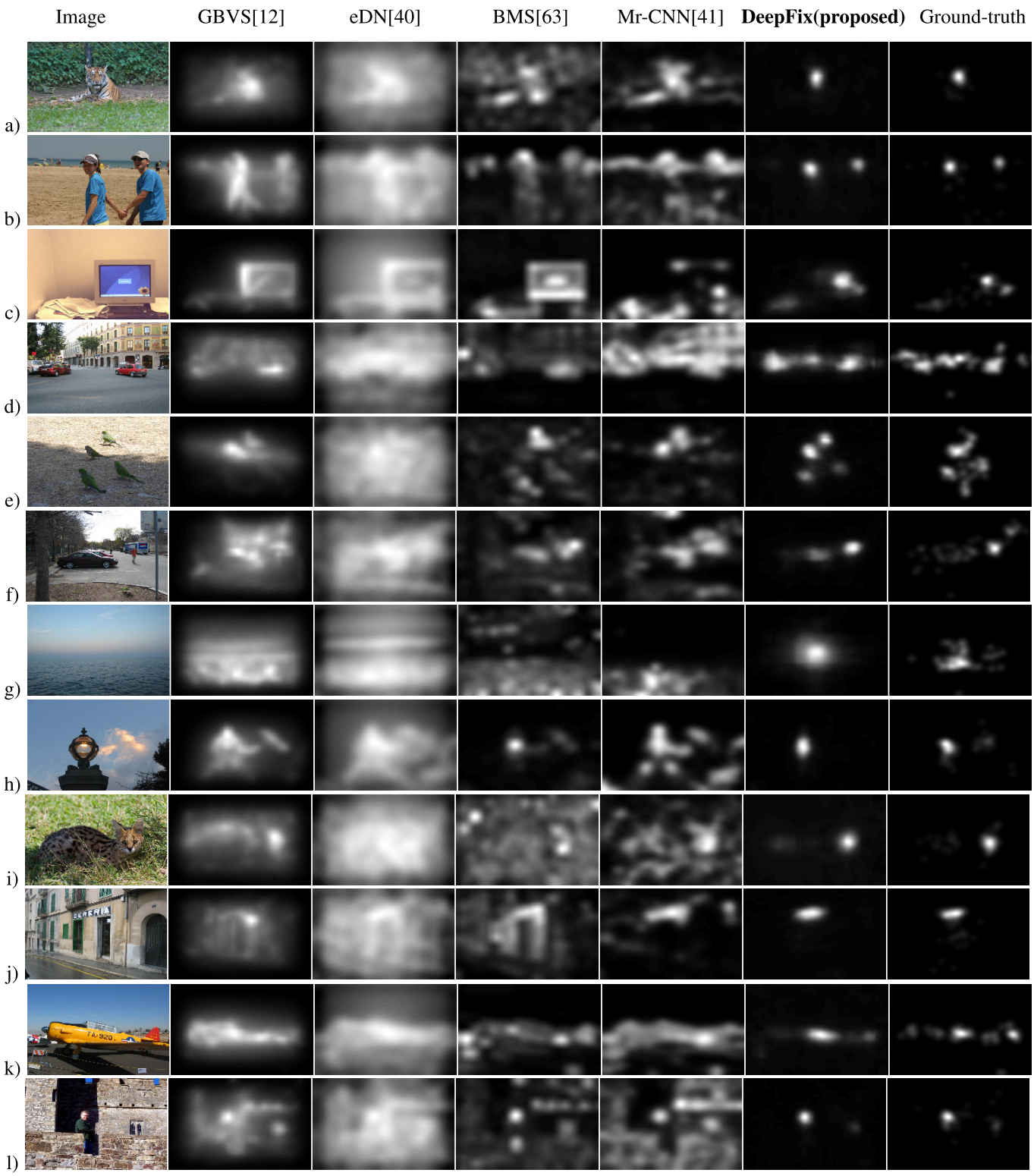


Fig. 5. Qualitative results of the proposed method on validation images from MIT1003

unit normalized saliency map $S_{\mathcal{N}}$ (with zero mean and unit standard deviation) at human eye fixations.

$$NSS = \frac{1}{N}\sum_{i=1}^{N} S_{\mathcal{N}}(i) \tag{3}$$

Here, N denotes the number of human eye positions.

3) Linear Correlation Coefficient(CC): The correlation coefficient metric between G and S is given by:

$$CC = \frac{cov(G,S)}{\sigma_G * \sigma_S} \tag{4}$$

It gives a measure of the linear relationship between the two maps. A score close to $+1$ indicates almost a perfect linear relationship between the maps.

4) Similarity Metric: Similarity metric computes the sum of the minimum values at each pixel location for the two distributions ($S_{\mathcal{N}}$ and $G_{\mathcal{N}}$). Here, $S_{\mathcal{N}}$ and $G_{\mathcal{N}}$ are normalized to be probability distributions.

$$Sim = \sum_{i=1}^{N} min(S_{\mathcal{N}}(i), G_{\mathcal{N}}(i)) \quad (5)$$

Here, $S_{\mathcal{N}}$ and $G_{\mathcal{N}}$ are the normalized distributions and N denotes all the pixel locations in the 2D maps. As the name suggests, a score of 1 denotes that the two maps are the same.

5) Area Under Curve (AUC): Area Under the ROC curve (AUC) is one of the widely used metrics for the evalaution of the maps estimated by saliency models. In AUC, two image locations are used: the actual human fixations as the positive set (fixation distribution) and some points randomly sampled from the image as the negative set (non-fixation distribution). Depending upon the choice of the non-fixation distribution, there are mainly two versions of AUC: AUC with uniform distribution of non-fixation points (AUC-Judd and AUC-Borji) and the shuffled-AUC. The shuffled-AUC uses human fixations of other images in the dataset (to take into account the center-bias) as the non-fixation distribution. Thus shuffled-AUC tends to give a lower score to those models which explicitly incorporate center-bias [66].

C. Results

The qualitative results obtained by the DeepFix network, along with that of other recent methods on a few example images from validation set of MIT1003 are shown in Fig. 5. As shown in the figure, the proposed network is able to efficiently capture saliency arising from both low-level features such as colour variations (rows c, e), shape (row h) etc. as well as the more high-level aspects such as text (rows j, k), faces of animals (rows a, i), and humans (row l). In the case of images without any strikingly salient regions (row g), the network relies on the learnt location pattern of humans gazing towards the centre to predict saliency accurately. The network is also able to analyze the relative importance of these factors and weigh them appropriately in the final saliency map. For instance, in row 2 of Fig. 5, DeepFix attributes saliency not to the bright colored T-shirts worn by the people (low-level), but to their faces (high-level). We attribute the performance of the proposed CNN architecture to its large depth allowing it to learn richer semantic representations, filters capturing multi-scale semantics and global context, as well as the implicit learning of location dependent patterns in eye fixations.

The quantitative results obtained on the datasets of CAT2000, MIT300, PASCAL-S, OSIE, FIGRIM are presented in Tables. I, II, III, IV, V respectively. The results obtained show that the proposed method achieves state-of-the-art results on all the datasets. As evident from these tables, our approach outperforms other methods by a huge margin with respect to a majority of the metrics–NSS, EMD, CC and Similarity, on both the datasets. The proposed method does not show a similar gain in performance on the AUC metrics. The AUC metrics reward methods primarily based on true positives, while the false positives generated do not incur heavy penalties. This can often result in blurred/hazy saliency maps receiving good scores as shown in Fig. 6. This drawback of AUC metrics has been previously observed by Borji *et al.* [71] and Zhao and Koch [72].

TABLE I
EXPERIMENTAL EVALUATION ON CAT2000 TEST SET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.87**	**0.75**	**1.11**	0.81	0.57	**0.88**	**2.29**
CAS [68]	0.77	0.50	3.09	0.76	**0.60**	0.42	1.07
Judd [57]	0.84	0.46	3.61	**0.84**	0.56	0.54	1.30
GBVS [12]	0.80	0.51	2.99	0.79	0.58	0.50	1.23

TABLE II
EXPERIMENTAL EVALUATION ON MIT300 TEST SET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.87**	**0.67**	**2.04**	0.80	0.71	**0.78**	**2.26**
Salicon [43]	**0.87**	0.60	2.62	**0.85**	**0.74**	0.74	2.12
Mr-CNN [41]	0.77	0.45	4.33	0.76	0.69	0.41	1.13
DG-I [38]	0.84	0.39	4.97	0.83	0.66	0.48	1.22
BMS[63]	0.83	0.51	3.35	0.82	0.65	0.55	1.41
eDN [40]	0.82	0.41	4.56	0.81	0.62	0.45	1.14
CAS [68]	0.74	0.43	4.46	0.73	0.65	0.36	0.95
Judd [57]	0.81	0.42	4.45	0.80	0.60	0.47	1.18
GBVS [12]	0.81	0.48	3.51	0.80	0.63	0.48	1.24

TABLE III
EXPERIMENTAL EVALUATION ON PASCAL-S DATASET

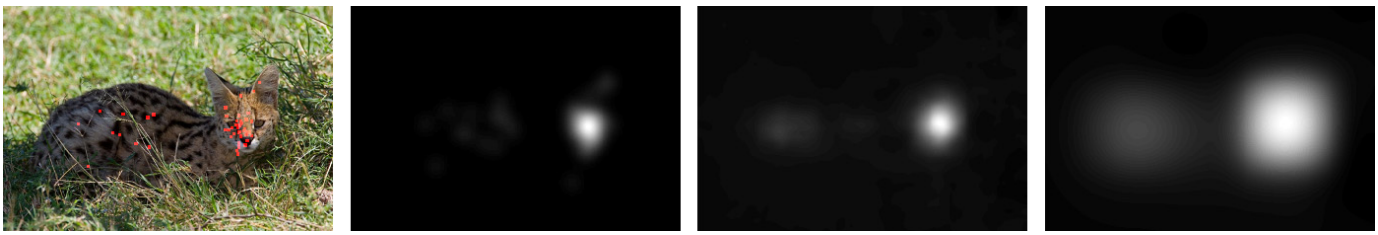
Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.91**	**0.65**	**0.54**	0.82	**0.73**	**0.78**	**2.60**
Salicon [43]	-	-	-	-	0.72	-	-
SU [47]	0.89	0.59	0.73	0.81	0.72	0.69	2.22
JN [69]	0.88	0.50	1.04	0.86	0.69	0.68	1.90
eDN [40]	0.89	0.39	1.29	**0.87**	0.65	0.55	1.42
BMS[63]	0.80	0.41	1.32	0.78	0.67	0.44	1.28
GBVS [12]	0.84	0.43	1.16	0.82	0.65	0.51	1.36

TABLE IV
EXPERIMENTAL EVALUATION ON OSIE DATASET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.91**	**0.66**	**1.04**	**0.83**	**0.79**	**0.80**	**3.04**
eDN [40]	0.82	0.36	2.02	0.82	0.68	0.40	1.16
BMS[63]	0.83	0.43	1.89	0.82	0.76	0.46	1.47
GBVS [12]	0.82	0.42	1.67	0.80	0.68	0.44	1.35
AWS [70]	0.82	0.42	1.93	0.81	0.76	0.45	1.45

TABLE V
EXPERIMENTAL EVALUATION ON FIGRIM DATASET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DeepFix	**0.90**	**0.66**	**1.10**	0.84	**0.67**	**0.80**	**2.51**
eDN [40]	0.87	0.37	2.88	**0.86**	0.62	0.50	1.38
BMS[63]	0.76	0.38	3.00	0.73	0.64	0.34	1.05
GBVS [12]	0.82	0.43	2.29	0.81	0.62	0.45	1.26
AWS [70]	0.72	0.36	3.20	0.74	0.64	0.29	0.89



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

Fig. 6. Illustration for the limitation of AUC metrics in penalizing false positives. For the image (a), with ground-truth saliency map (b), we have calculated the metric scores for two predictions (c) and (d). While (c) matches closely to the ground-truth, (d) can be observed to be highly blurred. The scores obtained for (c) are EMD = 1.04, NSS = 4.95, shuffled AUC = 0.88, AUC-Borji = 0.91. On the other hand, the scores obtained for (d) are EMD = 1.48, NSS = 3.23, shuffled AUC = 0.88, AUC-Borji = 0.94. It can be observed that the false positives in (d) are penalized by EMD and NSS metrics significantly. However, the shuffled AUC assigns the same score for both (c), (d) and contrary to expectations, AUC-Borji assigns a higher score for (d) than for (c). This limitation of AUC measures for saliency prediction has been observed before by [71] and [72].

The AUC-Shuffled metric is specifically designed to penalize models which account for the centre-bias present in eye fixations. Since our network learns this centre-bias as a location dependent fixation pattern, we obtain lower scores on shuffled AUC, despite the qualitative similarity of the predicted saliency maps with the ground-truth. For example, for the image in row 7 of Fig. 5, the saliency map predicted by DeepFix receives a shuffled AUC score of 0.68 where as the map predicted by eDN [40] receives a higher score of 0.70.

D. Large-Scale Scene Understanding (LSUN) Saliency Challenge

We further evaluate our model by submitting it to LSUN saliency challenge 2016 [73]. This challenge evaluates models on the test sets of SALICON and iSUN datasets with respect to multiple metrics. For evaluation on iSUN datset, our model is fine-tuned in a single stage with its 9000 image training set and is validated upon the remaining 926 images. Our model, used for predicting saliency on SALICON test set, is fine-tuned in a single stage using 12,500 images of SALICON dataset while the rest of its 2,500 images are used for validation.

The quantitative results of DeepFix along with that of the other competitors in the challenge[2] are shown in Table. VI and VII. Our model outperforms other competitors on iSUN, a dataset annotated for saliency using eye fixations captured from web camera. On SALICON dataset, where the saliency annotation for images is obtained using psycho-physical paradigm of tracking user's mouse gestures, we achieve reasonably good performance. Our model is the winner of LSUN saliency challenge 2016.

E. Analysis

In this subsection, we analyze the effect of Inception modules and LBC layers in the saliency map prediction. For this analysis, we construct three variations of our model.

1) DeepFix with no Inception modules (DF-No-Incp.): The inception modules in the network (in ⑥), illustrated in Fig. 3 are aimed at capturing semantics at multiple scales for saliency prediction. We remove these two inception blocks from the network in this variation.
2) DeepFix with no LBC layers (DF-No-LBC): The data-independent feature concatenated to the input of LBC layers, discussed in Sec. III-B, is aimed at introducing location dependence in the convolution operation. We remove this from the proposed DeepFix architecture converting the LBC layers to the usual convolutional layers.
3) DeepFix (DF): This is the proposed architecture described in Sec. III and shown in Fig. 1.

The above discussed models are trained as described in Sec. IV-A and are evaluated on the validation set of MIT1003.

[2]The results of our approach along with that of the other competitors are obtained from LSUN leaderboard-http://lsun.cs.princeton.edu/leaderboard/index_2016.html

TABLE VI

QUANTITATIVE RESULTS ON LSUN CHALLENGE [73] - ISUN (EYE FIXATION BASED SALIENCY DATASET)

Team	AUC-Judd	CC	IG	sAUC
VAL (Ours)	**0.862**	**0.815**	**0.179**	**0.550**
DEEPATTENT	**0.862**	0.814	0.174	**0.550**
NPU_HanLab	0.861	**0.815**	0.156	**0.550**
XRCE	0.855	0.787	0.102	0.538
UPC-Microsoft-BSC	0.860	0.798	0.136	0.541

TABLE VII

QUANTITATIVE RESULTS ON LSUN CHALLENGE [73] - SALICON (MOUSE-GESTURE BASED SALIENCY DATASET)

Team	AUC-Judd	CC	IG	sAUC
VAL (Ours)	0.761	0.804	0.315	0.630
DEEPATTENT	**0.767**	**0.890**	**0.326**	0.631
NPU_HanLab	0.756	0.775	0.318	**0.637**
XRCE	0.756	0.821	0.304	0.632
UPC-Microsoft-BSC	0.754	0.797	0.292	0.636

TABLE VIII

ANALYSIS OF THE PROPOSED METHOD ON MIT1003 VALIDATION SET

Method	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
DF-No-Incp.	0.89	0.51	1.72	**0.86**	0.74	0.69	2.37
DF-No-LBC	0.90	0.52	1.45	0.85	**0.75**	0.70	2.54
DF	0.90	**0.54**	**1.28**	0.85	0.74	**0.72**	**2.58**

TABLE IX

STANDARD ERROR FOR MEAN SCORES OF OUR MODEL ON OSIE DATASET

	AUC Judd	SIM	EMD	AUC-Borji	shuff. AUC	CC	NSS
Mean Score	0.91	0.66	1.04	0.83	0.79	0.80	3.04
Std. Error	0.001	0.003	0.015	0.003	0.003	0.004	0.034

The quantitative results obtained for the above three scenarios are presented in Table. VIII. The results emphasize that learning multi-scale semantics through inception modules and implicitly learning location dependent patterns through LBC layers results in better saliency prediction.

We have also computed the standard error for the mean scores obtained by our model for the test set - OSIE. These are presented in Table. IX and indicate a good statistical accuracy for the mean scores obtained.

V. CONCLUSION

In this work, we have proposed a fully convolutional neural net - DeepFix for predicting human eye fixations on images. The proposed deep net utilizes the potential of the 'inception module' to extract complex semantic features at multiple scales and also exploits the ability of 'filters with holes' to capture global context via large receptive fields. We introduce *LBC - Location Biased Convolutional filters*, a novel technique which enables the deep network to learn location dependent patterns. We show the advantage of LBC over traditional techniques of explicit bias addition, by means of an ablation analysis. Lastly, we show that an effective combination of the above mentioned concepts is able to outperform other state-of-the-art methods by a considerable margin.

REFERENCES

[1] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biol.*, vol. 14, no. 19, pp. R850–R852, 2004.

[2] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Trans. Appl. Perception*, vol. 7, no. 1, p. 6, 2010.

[3] E. Awh, E. K. Vogel, and S.-H. Oh, "Interactions between attention and working memory," *Neuroscience*, vol. 139, no. 1, pp. 201–208, 2006.

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

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[6] H. Hadizadeh and I. V. Bajić, "Saliency-aware video compression," *IEEE Trans. Image Process.*, vol. 23, no. 1, pp. 19–33, Jan. 2014.

[7] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition—A gentle way," in *Biologically Motivated Computer Vision*. Berlin, Germany: Springer, 2002, pp. 472–479.

[8] K. Rapantzikos, Y. Avrithis, and S. Kollias, "Dense saliency-based spatiotemporal feature points for action recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1454–1461.

[9] L.-Q. Chen, X. Xie, X. Fan, W.-Y. Ma, H.-J. Zhang, and H.-Q. Zhou, "A visual attention model for adapting images on small displays," *Multimedia Syst.*, vol. 9, no. 4, pp. 353–364, Oct. 2003.

[10] T. Yubing, F. A. Cheikh, F. F. E. Guraya, H. Konik, and A. Trémeau, "A spatiotemporal saliency model for video surveillance," *Cognit. Comput.*, vol. 3, no. 1, pp. 241–263, 2011.

[11] E. Niebur, "Saliency map," *Scholarpedia*, vol. 2, no. 8, p. 2675, 2007.

[12] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, 2006, pp. 545–552.

[13] G. Hinton *et al.*, "Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups," *IEEE Signal Process. Mag.*, vol. 29, no. 6, pp. 82–97, Nov. 2012.

[14] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1097–1105.

[16] L. Bottou, "Large-scale machine learning with stochastic gradient descent," in *Proc. COMPSTAT*, 2010, pp. 177–186.

[17] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Apr. 2015.

[18] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva, "Learning deep features for scene recognition using places database," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 487–495.

[19] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556

[20] C. Szegedy *et al.* (2014). "Going deeper with convolutions." [Online]. Available: https://arxiv.org/abs/1409.4842

[21] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3431–3440.

[22] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," Dept. Comput. Sci. Artif. Intell. Lab, MIT, Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, 2012. [Online]. Available: http://hdl.handle.net/1721.1/68590

[23] A. Borji and L. Itti. (2015). "Cat2000: A large scale fixation dataset for boosting saliency research." [Online]. Available: https://arxiv.org/abs/1505.03581

[24] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[25] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *J. Vis.*, vol. 14, no. 1, pp. 1–20, 2014.

[26] Z. Bylinskii, P. Isola, C. Bainbridge, A. Torralba, and A. Oliva, "Intrinsic and extrinsic effects on image memorability," *Vis. Res.*, vol. 116, pp. 165–178, Nov. 2015.

[27] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1072–1080.

[28] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.

[29] R. Valenti, N. Sebe, and T. Gevers, "Image saliency by isocentric curvedness and color," in *Proc. Int. Conf. Comput. Vis.*, 2009, pp. 2185–2192.

[30] Z. Liu, O. Le Meur, S. Luo, and L. Shen, "Saliency detection using regional histograms," *Opt. Lett.*, vol. 38, no. 5, pp. 700–702, 2013.

[31] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *Proc. Eur. Conf. Comput. Vis.*, 2012, pp. 101–115.

[32] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.*, vol. 13, no. 4, p. 11, Mar. 2013.

[33] F. Porikli, O. Tuzel, and P. Meer, "Covariance tracking using model update based on lie algebra," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2006, pp. 728–735.

[34] R. Kountchev and K. Nakamatsu, *Advances in Reasoning-Based Image Processing Intelligent Systems: Conventional and Intelligent Paradigms*, vol. 29. Springer, 2012. [Online]. Available: http://www.springer.com/in/book/9783642246920

[35] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *Proc. Int. Conf. Image Process.*, 2003, pp. I-253–I-256.

[36] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Proc. Adv. Neural Inf. Process. Syst.*, 2005, pp. 1–8.

[37] S. H. Khatoonabadi, N. Vasconcelos, I. V. Bajić, and Y. Shan, "How many bits does it take for a stimulus to be salient?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 5501–5510.

[38] M. Kümmerer, L. Theis, and M. Bethge. (2014). "Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet." [Online]. Available: https://arxiv.org/abs/1411.1045

[39] C. Shen and Q. Zhao, "Learning to predict eye fixations for semantic contents using multi-layer sparse network," *Neurocomputing*, vol. 138, pp. 61–68, Aug. 2014.

[40] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2798–2805.

[41] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 362–370.

[42] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "CNN features off-the-shelf: An astounding baseline for recognition," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2014, pp. 512–519.

[43] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Jun. 2015, pp. 262–270.

[44] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 1265–1274.

[45] S. R. Srivatsa and R. V. Babu, "Salient object detection via objectness measure," in *Proc. IEEE Int. Conf. Image Process.*, Sep. 2015, pp. 4481–4485.

[46] C. Sheth and R. V. Babu, "Object saliency using a background prior," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, Mar. 2016, pp. 1931–1935.

[47] S. S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. V. Babu, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 5781–5790.

[48] *Deeplearning.net*, accessed on Aug. 5, 2016. [Online]. Available: http://deeplearning.net/tutorial/lenet.html#maxpooling

[49] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. (2014). "Semantic image segmentation with deep convolutional nets and fully connected CRFs." [Online]. Available: https://arxiv.org/abs/1412.7062

[50] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. (2014). "Return of the devil in the details: Delving deep into convolutional nets." [Online]. Available: https://arxiv.org/abs/1405.3531

[51] G. Li and Y. Yu. (2015). "Visual saliency based on multiscale deep features." [Online]. Available: https://arxiv.org/abs/1503.08663

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

[53] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 1929–1958, 2014.

[54] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, p. 4, 2007.

[55] P.-H. Tseng, R. Carmi, I. G. M. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *J. Vis.*, vol. 9, no. 7, p. 4, 2009.

[56] Y. Yang, M. Song, N. Li, J. Bu, and C. Chen, "What is the chance of happening: A new way to predict where people look," in *Proc. Eur. Conf. Comput. Vis.*, 2010, pp. 631–643.

[57] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE Int. Conf. Comput. Vis.*, Sep./Oct. 2009, pp. 2106–2113.

[58] Y. Jia *et al.* (2014). "Caffe: Convolutional architecture for fast feature embedding." [Online]. Available: https://arxiv.org/abs/1408.5093

[59] A. Borji, "What is a salient object? A dataset and a baseline model for salient object detection," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 742–756, Feb. 2015.

[60] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: Strengths and weaknesses," *Behavior Res. Methods*, vol. 45, no. 1, pp. 251–266, Mar. 2013.

[61] T.-Y. Lin *et al.*, "Microsoft COCO: Common objects in context," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 740–755.

[62] Z. Bylinskii *et al. MIT Saliency Benchmark*, accessed on Aug. 5, 2016. [Online]. Available: http://saliency.mit.edu/

[63] 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 *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 1153–1160.

[64] R. J. Peters and L. Itti, "Applying computational tools to predict gaze direction in interactive visual environments," *ACM Trans. Appl. Perception*, vol. 5, no. 2, p. 9, 2008.

[65] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 8, pp. 2397–2416, 2005.

[66] S. Barthelmé, H. Trukenbrod, R. Engbert, and F. Wichmann, "Modeling fixation locations using spatial point processes," *J. Vis.*, vol. 13, no. 12, p. 1, 2013.

[67] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[68] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 153–160.

[69] J. Pan, K. McGuinness, E. Sayrol, N. O'Connor, and X. Giro-i-Nieto. (2016). "Shallow and deep convolutional networks for saliency prediction." [Online]. Available: https://arxiv.org/abs/1603.00845

[70] 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," *J. Vis.*, vol. 12, no. 6, p. 17, 2012.

[71] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 921–928.

[72] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, p. 9, 2011.

[73] (2016). *Large-Scale Scene Understanding (LSUN): Challenge Leaderboard*. [Online]. Available: http://lsun.cs.princeton.edu/leaderboard/index_2016.html