Project Report: Saliency Prediction

[November 22, 2017]

Nittin Singh
2015csb1067@iitrpr.ac.in
Narotam Singh
2015csb1065@iitrpr.ac.in

Department of Computer Science and Engineering
Indian Institute of Technology Ropar-140001

Abstract

This report summarizes our work done in Computer Vision's Project : Visual Saliency Prediction. The approach that we have followed takes into account the local as well as the global saliency features of an image. We have trained two models one for local and other for global features. In local model approach, certain patches of pre-processed images are extracted, these patches are extracted on the basis of the saliency scores of the center pixel. Then we have trained a Deep-CNN model on this data, this trained model returns the saliency score of each pixel in the input HSV image. In the global model, the image pre-processing involves shifting the required pixel to centre after superpixel segmentation in order to get the global semantics of the pixel. Then these centered images are used as the training data for our global Deep-CNN model. The trained model predicts the saliency score of each pixel in the range of [0-1]. After generating the saliency map, we applied certain post-processing steps on the saliency map to reduce the unnecessary noise in the image and to remove the less salient portions of the map.

1 Introduction

One of the most powerful and useful abilities of human cognizant behaviour is the ability to neglect what is 'not-necessary'. This is applicable mostly in human vision system, in the images that we come across, our brain is designed to keep details of only certain objects or regions, which have some importance or tend to have some. Such features/regions of an image are called its salient features. Thus in the quest of developing a technology trying to mimic human cognitive behaviour, saliency prediction is a very fundamental problem for researchers in the field of computer vision.

Before the advent of Machine Learning and deep learning methods the main approach was that of rigorous image analysis [1], [7] . It involved the convolution of various filters and the segmentation of images based on the pixel similarities. The main idea was to identify and segregate certain objects of higher semantic meaning than others. As the technology is advancing and with better techniques of machine learning and Convolutional neural networks the 'learning' approach is widely accepted and preferred. There has been a lot of work being done in this field. The most common and intuitive approach is the bottom-up structuring of models on either local or the global semantic level. Though these approaches seem to be fine but still they miss the global perspective of an image due to their bottom up approach. Here we have employed the Deep Convolutional Neural Networks for our models. The basic idea of our model is to exploit the learning nature of Deep-CNNs both at the global as well as the local levels. The local level features contribute to the much finer details such as the edges and texture of image, while the global level features are more sensitive towards an object as a whole. Unlike local methods which are sensitive to high frequency image contents like edges and noise, global methods are less effective when the textured regions of salient objects are similar to the background (See Figure 1(d)). The combination of local and global methods has been explored by a few recent studies, where background prior, center prior, color histograms and other handcrafted features are utilized in a simple and heuristic way to compute saliency maps.

Depending on the model semantics (whether local or global) we have slightly changed the pre-processing procedures. In our local pipeline, the model is trained to focus on the local salient features and to capture finer details such as texture and edges. While in Model2 (Global), the trained model captures the details that are significant globally. Here we have tried to create a balanced model which will take care of local as well as the global saliency cues.

2 Literature / Prior art

2.1 History

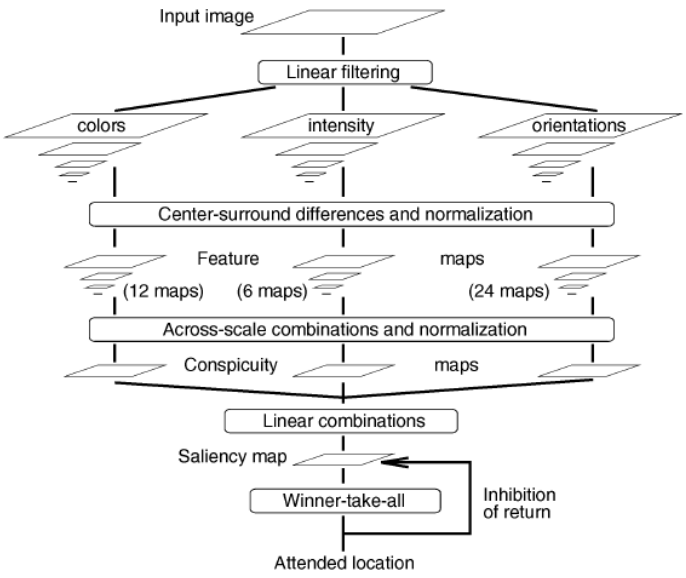

Figure 1: *Architecture of the model proposed by Itti, Koch and Niebur*

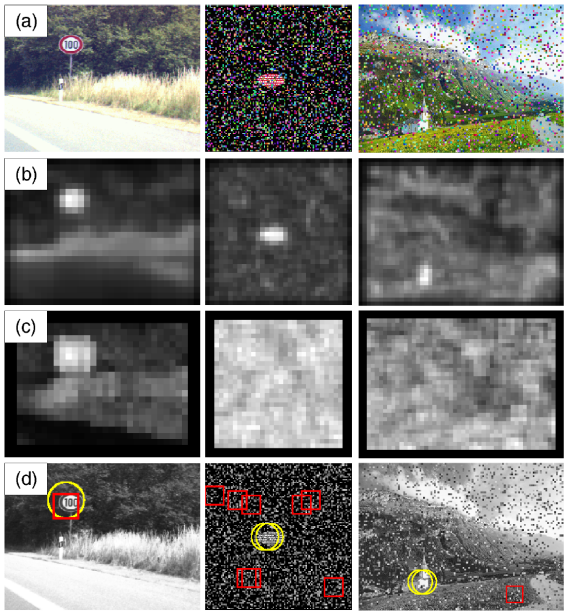

Figure 2: *Results for the above model*

Itti and Koch's [8] work is one of the earliest substantial work done in the field of saliency prediction. Their model is an embedded implementation of previous computational frameworks and psychological theories of bottom-up attention based on centre surround mechanisms. Subsequent behavioural [15] and computational [4] investigations used fixations as a means to verify the saliency hypothesis and to compare models.

The second big leap was with the works of Liu [13] and Achanta [2] who defined saliency detection as a binary segmentation problem. A huge amount of saliency models has emerged since then.

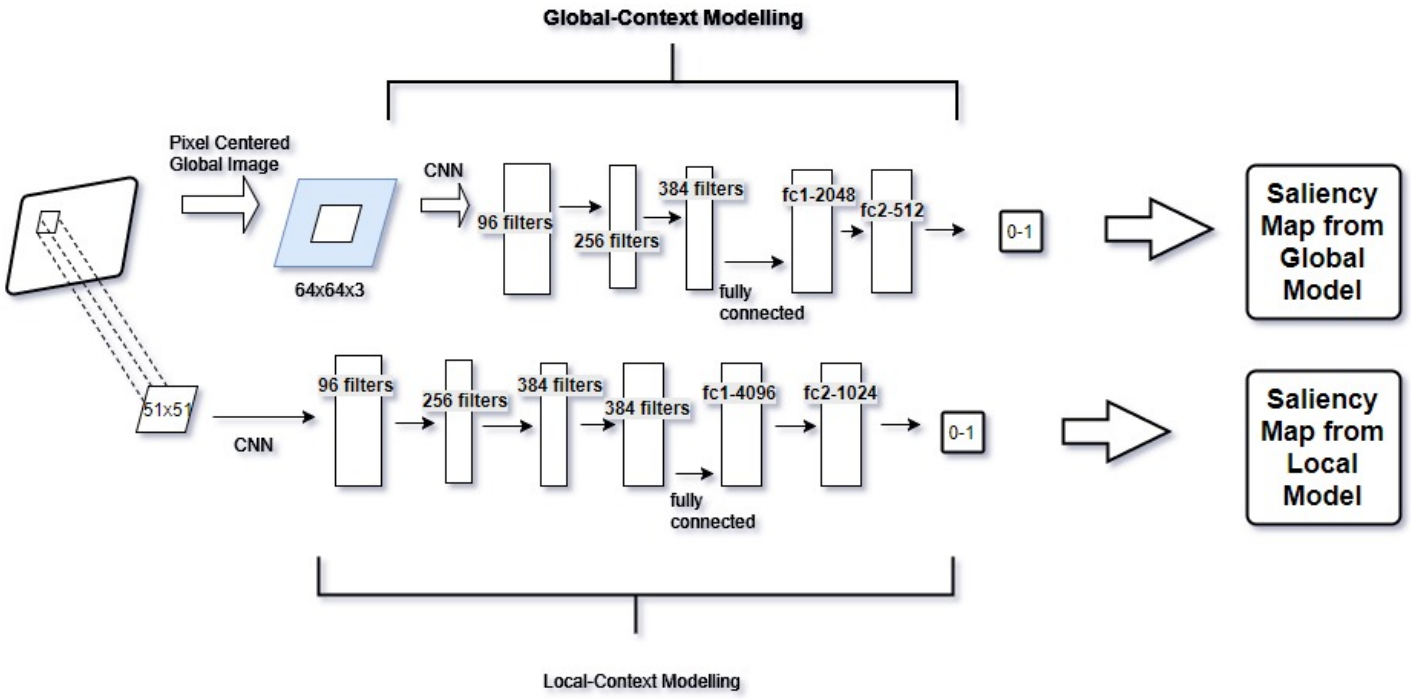

Figure 3: Proposed Model-structure

Next big development was with the advancement in technology and enhanced computing ability along with the emergence of the convolutional neural networks (CNNs), in particular with the introduction of the fully convolutional neural networks. Unlike the classical models which very mostly handcrafted, CNN-based methods eliminate the need for previously designed masks, as well as reduce the centre bias, and hence have been adopted by many researchers. A fully connected ConvNet Model depends on a huge number of variables. Depending on the receptive field range of individual neurons one can decide the local or the global features to be selected. This huge flexibility of the CNNs and their appreciable outputs have diverted many researchers to this approach.

3 Related Works

Visual Saliency models can be split into biologically inspired and computational approaches. These approaches can also be divided into bottom-up and top-down approaches.

Most of the visual saliency models available are based on the low level features and the global features [16]. They use the low level information like color, contrast orientation or texture to create the saliency maps. L. Itti [8] used these low level features which is based on bottom-up approach, is a biologically inspired model. Koch and Ullman's [12] architecture was a major source of inspiration for bottom up approaches including the works of L.Itti [8] and Harel's Graph based approach [6].

Computational models include [1], [7] which is purely computational, [5] and [6]. Centre-surround hypothesis is a part of a number of saliency models mostly bottom up models, even the top-down approaches use centre-surround in combination with other approaches to get higher accuracy. R. Achanta used center-surround distance in [2] (a purely computational model) and center-surround contrast method in [1] . In the latter, they constructed saliency maps using the frequency domain analysis of the images and it was able to outperform many state of the art models at that time.

Apart from low and midlevel features, high level features like objects, human faces representations are also reliable. Judd et al. [10], observed that humans, faces, cars, animals and other objects attract human attention and prove to be very salient regions of the image. As many of the saliency detection techniques use bottom up approaches, that do not consider top-down image semantics and often does not match actual eye movements. To address this problem, they have collected eye tracking data of 15 viewers on 1003 images and used this database as training and testing examples to learn a model of saliency based on low, middle and high-level image features.

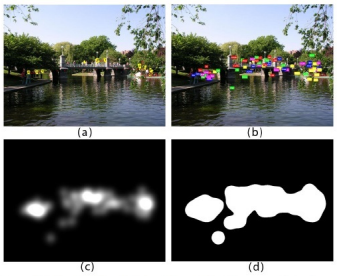

Figure 4: *(Judd) Based on the eye-tracking data on 1003 images from 15 viewers to use as the ground truth (b). A continuous saliency, map is formed (c). This map is then thresholded to some value to form the (d) image.*

Our project is based on the training on human eye gaze and fixations dataset and similar datasets are used by [11], [10] and Ali Borji's [3]. Here, Borji has combined low-level features such as orientation, color, intensity, saliency maps of previous best bottom-up models with top-down cognitive visual features (e.g., faces, humans, cars, etc.) and learnt a mapping from those features to eye fixations using Regression, SVM, and AdaBoost classifiers. Further achievement of this model was that it successfully detected the most salient object in a scene without any sophisticated image processing such as region segmentation.

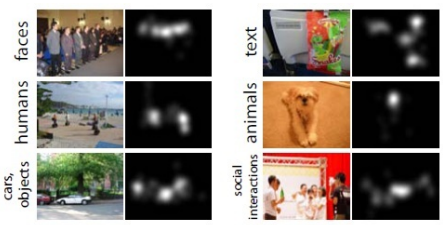

Figure 5: *(A. Borji) Human fixation map. Top-down concepts including people, social interactions, animals, cars, signs etc.*

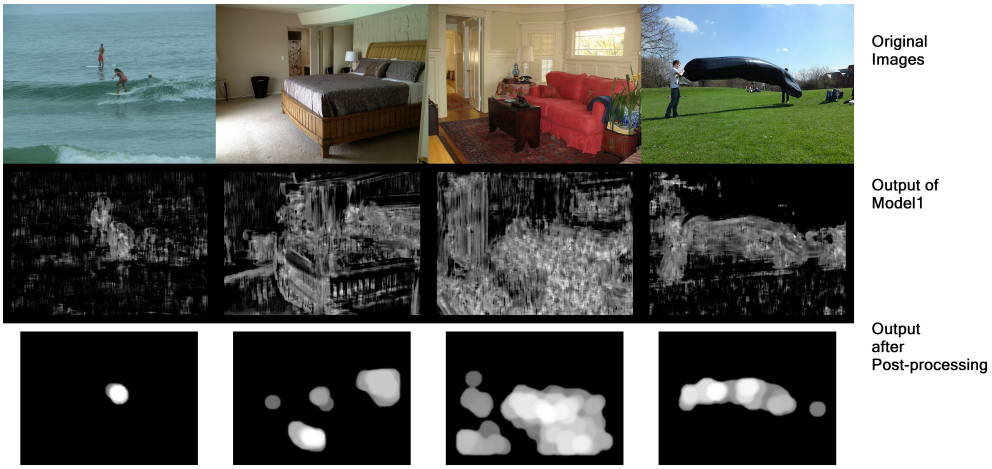

Figure 6: Results of model1 along with their post-processed Test-images

4 Local and Global Model

As shown in the Fig. 3, we have considered two models. The lower model learns the local features of the salient region while the upper model learns the global features with respect to that region. The lower model takes the 51x51 HSV image patch as input and outputs the saliency score of the centre pixel of the input patch, in the range [0-1]. The upper model takes the 64x64 RGB superpixel image as the input and outputs the saliency score of the centre pixel in the range [0-1].
Denote by#conv a convolution layer, #pool a max-pooling layer, #drop a dropout, #relu a relu activation block, #fc a fully connected layer and #sig a sigmoid activation block. The pre-processing and post-processing steps are described in the next section for both the pipelines. The structure of the local model is :
Layer1 #conv1 - #relu - #pool1{size : (3,3) stride : (2,2)}
Layer2 #drop1 - #conv2 - #relu - #pool2{size : (2,2) stride : (2,2)}
Layer3 #drop2 - #conv3 - #relu - #pool3{size : (3,3) stride (3,3)}
Fully connected layer1 #fc1(2048) - #relu
Fully connected layer2 #drop3 #fc2(512) - #relu
Fully connected layer3 #fc3(1) #sig - #output.
The filter sizes for first, second and third layer are (11,11), (5,5) and (2,2) respectively.
The structure of the global model is :
Layer1 #conv1 - #relu - #pool1{size : (2,2) stride : (2,2)}
Layer2 #drop1 - #conv2 - #relu - #pool2{size : (2,2) stride : (2,2)}
Layer3 #drop2 - #conv3 - #relu - #pool3{size : (2,2) stride (2,2)}
Layer4 #conv4 - #relu - #pool3{size : (2,2) stride (2,2)}
Fully connected layer1 #fc1(4096) - #relu
Fully connected layer2 #drop3 - #fc2(1024) - #relu
Fully connected layer3 #fc3(1) #sig - #output.
The filter sizes for first, second, third and fourth layer are (5,5), (5,5), (3,3) and (3,3) respectively.
There is partial top-down character in local model because some of the patches that we have used for training actually include human faces, human bodies, and other objects which has turned out to be useful.

5 Experiments

5.1 Dataset: SALICON dataset [9]

The current release comprises 10,000 training images and 5,000 validation images with saliency annotations. For training and validation sets, the color images in JPG format are present and ground truth (including gaze trajectory, fixation points, and saliency map) are also provided. The test set with 5,000 images is released without ground-truth. This dataset was used by [14].

5.2 Procedure

In the dataset, we are given images and their corresponding saliency maps, and the eye gaze-fixations of different subjects. Saliency score are the pixel intensity values of the saliency maps provided in the dataset in the range of [0-1]. For salient and non salient regions, the eye gaze fixations of all the subjects are considered. Further other regions with pixel intensity less than a threshold are also considered to increase the training data for non salient regions. For local-context model, before extracting any local patches for training, all the original training images are converted from RGB to HSV space. In the same model, if the saliency score of a pixel in the saliency map is greater than the threshold (0.1) then we will consider a patch of size 51x51 with that pixel at center from the HSV source train image as the potential patch. In this patch, if 70 % of pixels have saliency score more than the threshold then we consider this patch for training, otherwise its discarded. Similarly for non-salient features, we considered the pixel with score less than the threshold and again the patch around that pixel is considered if and only more than 70% of the pixels in that patch are having scores less than the threshold value. All such 51x51 patches are the input to the first model.
In the pre-processing stage of model2 i.e. global-context model, we took a slightly different path. In this model, initially all the training images are converted into superpixel segmented images and the RGB space is retained unlike the previous model. Here again we select those eyegaze-fixation points which have scores greater than the threshold and are surrounded with more than 70% of such pixels. We take the good pixel and bring that pixel to the center shifting the remaining image along with it. To prevent the image to go out of bounds, the image is padded with the mean pixel value of the training images. Image is then resized to 64x64 to meet the model requirement. Similarly the non-salient superpixel centered global images are also obtained, and the model is trained as shown in the Figure 3.

5.3 Post Processing

As we can see in Fig. 6, the output of the model1 is not much decent so we will perform a further post-processing to the final output. The post processing involves a series of image 'eroding' steps followed with a series of image 'dilating' steps which ensures the maintenance of overall saliency structure and minimizes the noise. Afterwards we ran a quadratic kernel over the entire pixels in the saliency map which are above certain threshold (30% of max) and then gaussian kernel is applied to obtain a smooth final saliency structure.

6 Results

6.1 Evaluation Metrics

- Precision-recall : Precision is defined as :

$$P = TP/(TP+FP) \tag{1}$$

Recall is defined as :

$$R = TP/(TP+FN) \tag{2}$$

TP is true positive, FP is false positive and FN is false negative. A (P,R) curve can be plotted for different thresholds in order to compare the saliency map to the ground truth.

- F-measure or F1 score : Used as a measure of test's accuracy and defined as the weighted harmonic average of precision and recall. Defined as

$$F = \frac{(1-\gamma^2)Precision \times Recall}{\gamma^2 Presicion + Recall}$$

$$where \quad \gamma = 0.3$$

To find out precision,recall and f-score of the predicted saliency maps, we binarized the continuous saliency map with certain threshold.

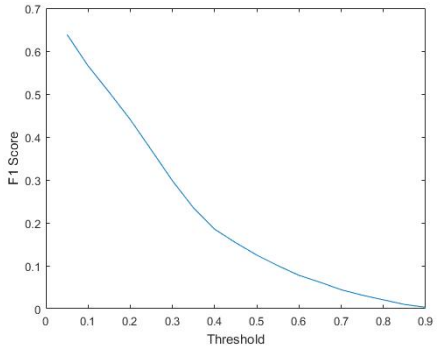

Figure 7: graph between 'f-score' and 'threshold'

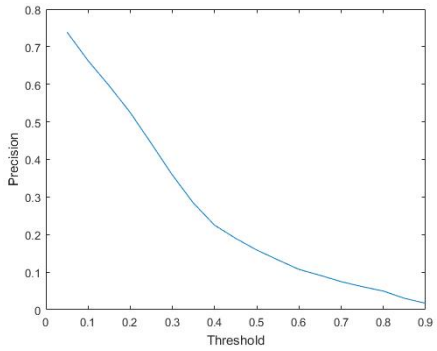

Figure 8: graph between 'precision' and 'threshold'

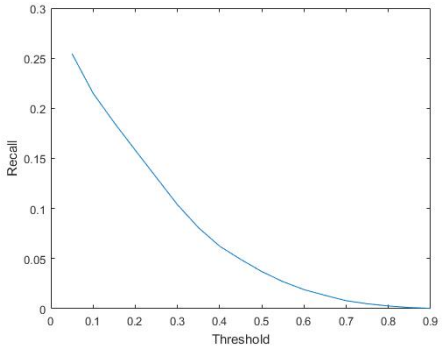

Figure 9: graph between 'recall' and 'threshold'

7 Conclusion

The global model has not performed as expected. From the Fig.10, its visible that the model is considerably underfit and there is considerable amount of noise. So we have discarded the global model.

The local model has performed reasonably well, one of the major challenges of the saliency prediction is the 'center-bias' which is due to the natural human tendency of looking at the center of the images. But here we have tried to tackle the problem and we are somewhat able to reduce

Figure 10: Result is not center biased

Figure 11: Saliency Map predictions of Model1

the center bias as shown in the example fig.11: We can say that with the state of art machines and technology this task of saliency prediction can be completed with promising accuracy.

8 In future

To improve the global model, it need to be trained for a larger number of epochs. In the local model we performed various pre-processing and post-processing tasks which helped us in improving the accuracy of the local model to a significant level. However this task involved a lot of parameters which we have chosen heuristically. But these parameters can be learned for optimum results, which is one possible direction this project can go into. Other major problem is that the results obtained are very slow these are not real time results, one reason is huge prediction model, however the task of saliency prediction is a very real time application thus further major improvement over this project would be a real time application.

9 References

[1] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1597–1604, June 2009. doi: 10.1109/CVPR.2009.5206596.

[2] Radhakrishna Achanta, Francisco Estrada, Patricia Wils, and Sabine Süsstrunk. Salient region detection and segmentation. *Computer Vision Systems*, pages 66–75, 2008.

[3] Ali Borji. Boosting bottom-up and top-down visual features for saliency estimation. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 438–445. IEEE, 2012.

[4] Neil Bruce and John Tsotsos. Saliency based on information maximization. In *Advances in neural information processing systems*, pages 155–162, 2006.

[5] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *Computer vision and pattern recognition, 2008. cvpr 2008. ieee conference on*, pages 1–8. IEEE, 2008.

[6] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552, 2007.

[7] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral resid-

ual approach. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[8] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998.

[9] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[10] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *Computer Vision, 2009 IEEE 12th international conference on*, pages 2106–2113. IEEE, 2009.

[11] Wolf Kienzle, Felix A Wichmann, Matthias O Franz, and Bernhard Schölkopf. A nonparametric approach to bottom-up visual saliency. In *Advances in neural information processing systems*, pages 689–696, 2007.

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

[13] Tie Liu, Zejian Yuan, Jian Sun, Jingdong Wang, Nanning Zheng, Xiaoou Tang, and Heung-Yeung Shum. Learning to detect a salient object. *IEEE Transactions on Pattern analysis and machine intelligence*, 33(2):353–367, 2011.

[14] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[15] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision research*, 42(1):107–123, 2002.

[16] Keyang Shi, Keze Wang, Jiangbo Lu, and Liang Lin. Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2115–2122, 2013.