8-1-2019

VISUAL SALIENCY ANALYSIS, PREDICTION, AND VISUALIZATION: A DEEP LEARNING PERSPECTIVE

Ali Majeed Mahdi
Southern Illinois University Carbondale, ali.majeed.mahdi@gmail.com

VISUAL SALIENCY ANALYSIS, PREDICTION, AND VISUALIZATION: A DEEP LEARNING PERSPECTIVE

by

Ali Majeed Mahdi

M.S., Southern Illinois University, 2013
B.S., Al-Mustansiriya University, 2007

A Dissertation
Submitted in Partial Fulfillment of the Requirements for the
Doctor of Philosophy degree

Department of Electrical & Computer Engineering
in the Graduate School
Southern Illinois University Carbondale
August 2019

DISSERTATION APPROVAL

VISUAL SALIENCY ANALYSIS, PREDICTION, AND VISUALIZATION: A DEEP LEARNING PERSPECTIVE

by

Ali Majeed Mahdi

A Dissertation Submitted in Partial

Fulfillment of the Requirements

for the Degree of

Doctor of Philosophy

in the field of Electrical & Computer Engineering

Approved by:

Jun Qin, Chair

Haibo Wang

Lalit Gupta

Mohammad Sayed

Mingqing Xiao

Graduate School
Southern Illinois University Carbondale
April 16, 2019

AN ABSTRACT OF THE DISSERTATION OF

Ali Majeed Mahdi, for the Doctor of Philosophy degree in Electrical & Computer Engineering, presented on April16, 2019, at Southern Illinois University Carbondale.

TITLE: VISUAL SALIENCY ANALYSIS, PREDICTION, AND VISUALIZATION: A DEEP LEARNING PERSPECTIVE

MAJOR PROFESSOR: Dr. Jun Qin

In the recent years, a huge success has been accomplished in prediction of human eye fixations. Several studies employed deep learning to achieve high accuracy of prediction of human eye fixations. These studies rely on pre-trained deep learning for object classification. They exploit deep learning either as a transfer-learning problem, or the weights of the pre-trained network as the initialization to learn a saliency model. The utilization of such pre-trained neural networks is due to the relatively small datasets of human fixations available to train a deep learning model. Another relatively less prioritized problem is amount of computation of such deep learning models requires expensive hardware. In this dissertation, two approaches are proposed to tackle abovementioned problems. The first approach, codenamed DeepFeat, incorporates the deep features of convolutional neural networks pre-trained for object and scene classifications. This approach is the first approach that uses deep features without further learning. Performance of the DeepFeat model is extensively evaluated over a variety of datasets using a variety of implementations. The second approach is a deep learning saliency model, codenamed ClassNet. Two main differences separate the ClassNet from other deep learning saliency models. The ClassNet model is the only deep learning saliency model that learns its weights from scratch. In addition, the ClassNet saliency model treats prediction of human fixation as a classification problem, while other deep learning saliency models treat the human fixation prediction as a regression problem or as a classification of a regression problem.

ACKNOWLEDGEMENTS

When I came to Southern Illinois University at Carbondale, I wanted to learn and grow as an engineer. I did not imagine that such experience will have a significant impact on my knowledge, experience and personality. In graduate school, I had the opportunity to become a research assistant, teaching assistant, write papers, give talks, travel to conferences, and become a researcher. Several people were a great help along the way. I would like to acknowledge some of these wonderful people:

- Jun Qin: for guiding me as a researcher and as a person. You always gave me your time when I needed help. Every advice you have given me was for my best interest. Your feedbacks on my thoughts, writing, and skills helped me become who I am today. Without your help this wouldn't be possible.
- My committees: for honoring me by accepting my invitation to serve as committees for my dissertation defense. The advices you have given me help me to be an open minded and learn to listen to other views who can be crucial for my research and my career.
- My colleagues: You directly helped me throughout my graduate study by helping me with a variety of things such as collecting eye tracking data, remote access, and for general advice.
- My professors: for giving me the required skills and knowledge to move on with my PhD study. The time you have given me to answer questions or giving me an advice made me stronger than I was.
- My friends at SIU: you are some of the most intelligent, adventurous, oriented, and driven students I have met. I have learned so much from your experiences and exchanging of thoughts.

- All my friends: for your support keeping me sane and balanced. I will always fondly remember our times together.
- My parents: you are the first reason why I have done my graduate studies. You have always told me that I will have a PhD degree. I remember those words at age of 6. You planted a seed and now it's time for reward. I will pass the same message to my kids.
- My father: who had his PhD when I was a little boy. You always explained to me why it is important to have a PhD degree. You showed me your thesis, books, and accomplishments, and taught me I can become successful too.
- My sisters: you have done a great job motivating each other and keeping the bar high. I know you are proud of me.
- The rest of my family: you have always thought of my best interest. Given me advices, supporting me, and made sure I take a rest every now and then.
- My colleagues at work: for giving me the time to finish writing my dissertation and prepare for my final defense. Also, thank you for offering to proof read my dissertation.
- Everyone else I did not mention: whether I have thanked you in person or not, I really appreciate your help no matter how small it was, it definitely had an impact on me.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

INTRODUCTION

1.1 Motivation:

During the last few decades, saliency models have been developed rapidly to leverage the understanding of human visual attention. In general, a saliency map is defined as a 2D probabilistic map that reflects a distribution of predicted fixations. A large saliency value indicates that an eye fixation has a large probability to fall on the corresponding spatial location, object, or region. Saliency modeling is beneficial to prediction of sequence or distributions of human fixations in an image [1,2]. Since human attention relies on the bottom-up and top-down influences, the developed saliency models may rely on the bottom-up influences, the top-down influences, or a combination of both. While the bottom-up attention is fast and defines saliency in terms of distinction to the surroundings [3], the top-down attention is slow and relies on prior knowledge, expectations, and rewards [4]. Visual saliency models have been applied to various applications, including object detection [5][6], image segmentation [7][8], image retargeting [9][10], image/video compression [11][12], visual tracking [13][14], gaze estimation [15], robot navigation [16], image/video quality assessment [17][18], and advertising design [19].

The first outlined description of attention is by James in 1890 [20]. In 1990, Corbetta et al. defined attention as the mental ability to select stimuli, responses, memories, and thoughts that are behaviorally relevant among several others that are behaviorally irrelevant [21].The feature integration theory suggests that the visual stimuli are processed in different regions of the brain as the bottom-up visual features in a parallel manner [22]. The resulting feature maps are assembled to advocate for object recognition. Koch & Ulman [23] then proposed a combination of such visual features to produce a saliency map. They also introduced a winner-take-all strategy to select the most salient location, and the inhibition-of-return strategy to predict the next most salient location. Based on Koch & Ulman, studies attempted to implement an attention system. Itti & Koch [24] proposed the first complete implementation of Koch & Ulman model. The model incorporates color, intensity, and orientation features at various scales using a center surround operation. Architecture of the biologically inspired model is presented in Figure 1.

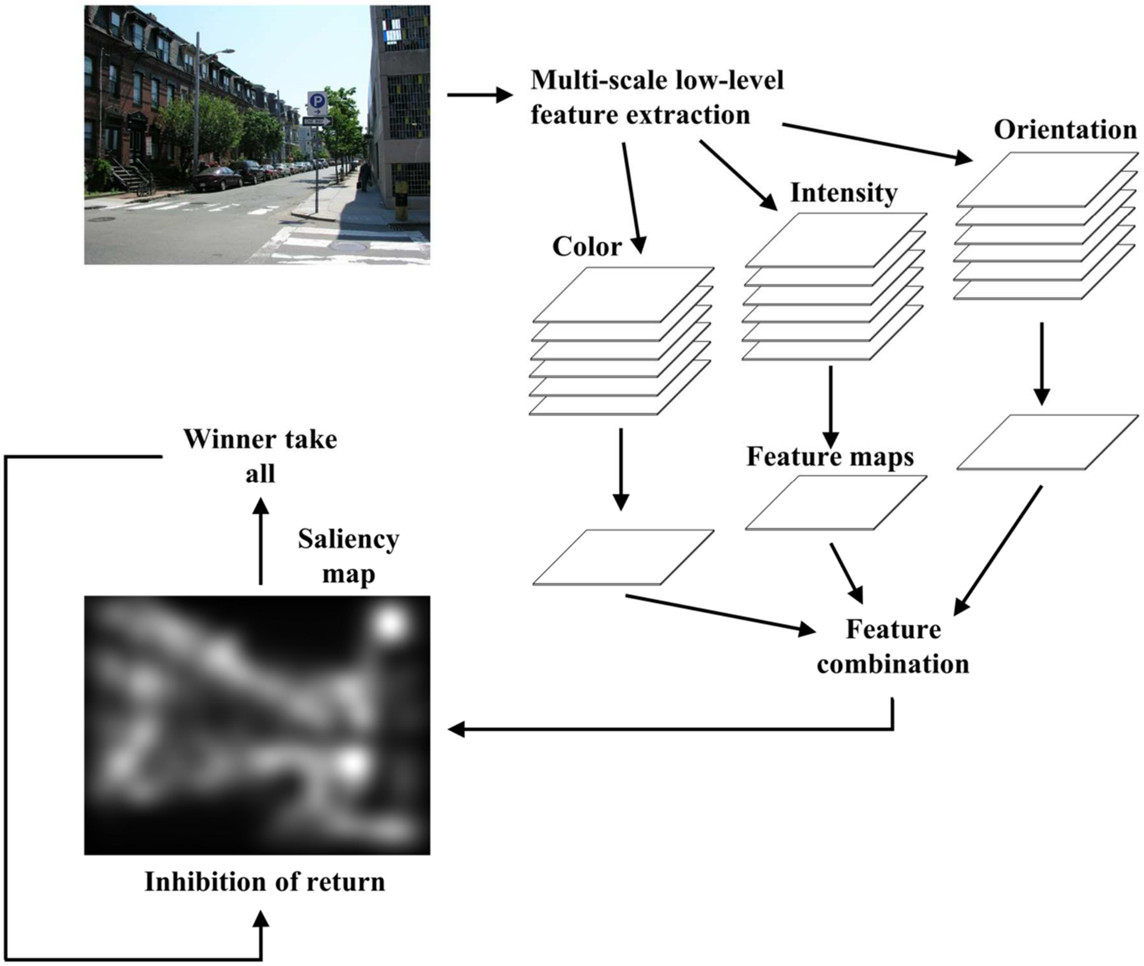


Figure 1 - Architecture of the Itti&Koch saliency model.

Moreover, several other studies exploited other handcrafted features and demonstrated exciting results. Itti & Baldi, 2006 [25] introduced surprise as a Bayesian framework to predict eye movements. Bruce & Tsotsos developed an information theoretic saliency model using the independent component analysis (ICA) as features derived from natural scenes. Navalpakkam & Itti [27] proposed a signal to noise ratio saliency model, which learned parameters of low-level features combination. Cerf et al. modified existing saliency models by incorporating face detection in a bottom-up manner [28]. Zhang et al. proposed a Bayesian framework that incorporated self-information and prior knowledge using difference of Gaussians (DoG) as visual features [29]. Judd et al. learned a saliency model via a support vector machine (SVM) [30]. The model exploited low, mid, and high-level features such as color, intensity, orientation, horizon detector, center bias, and face, car, and people detectors. Liu et al. exploited multi-scale contrast, center-surround histogram, and color spatial distribution as hand crafted features to detect salient objects [31]. Tian et al. proposed a salient region detection model using color and orientation as the bottom-up features and depth-from-focus as a top-down feature [32]. Zhang and Sclaroff proposed a Boolean saliency model using color features [33]. The model obtained Boolean maps by random thresholding of the feature maps. Zhang et al. devised a manifold ranking saliency model by segmenting the background regions of images for salient objects detection [34]. The study experimentally compared the integration of features such as locally assembled binary (LAB), local binary pattern (LBP), histograms of oriented gradients (HOG), and dis- criminative regional feature integration (DRFI). In addition, other features have also been used in saliency models, including scale invariant feature transform [35], optical flow [36], multiple superimposed orientations [37], entropy [38], gist [39], ellipses [40], flicker [41], symmetry [42], histogram of local orientations [43], isocentric curvature [44], wavelet transform [45], depth influences [46], and regional histograms [47].

Although the selections of the abovementioned handcrafted features lead to astonishing results, the predictions of such conventional saliency models are limited to the incorporated features. To overcome such bottleneck, two saliency models are developed in this study. The first saliency model, codenamed DeepFeat, which exploits the feature maps of pre-trained convolutional neural networks (CNNs) [48]. Figure2 shows two images, ground-truth maps of human fixations, and saliency maps generated by a conventional model [24], and the DeapFeat model, respectively. In both images, compared with ground-truth maps, the conventional saliency model fails to predict the animal and the baby, as such features are not incorporated in the conventional model. In contrast, the DeepFeat model can predict such missing contents: the monkey face in the first image and the baby face and drawings on the shirt in the second image. It indicates that the feature maps of pre-trained CNNs can provide more features, which a conventional saliency model may not incorporate. Such features may be benefit to saliency prediction of human gaze patterns. In this dissertation, the feature maps of pre-trained CNNs will be denoted as deep features. The second proposed saliency model, codenamed ClassNet, treats the fixation prediction as a classification problem of individual pixels. In the proposed framework, large eye fixation datasets can be derived from a relatively small dataset. Such advantage allows the proposed ClassNet model to train from scratch using random weights.

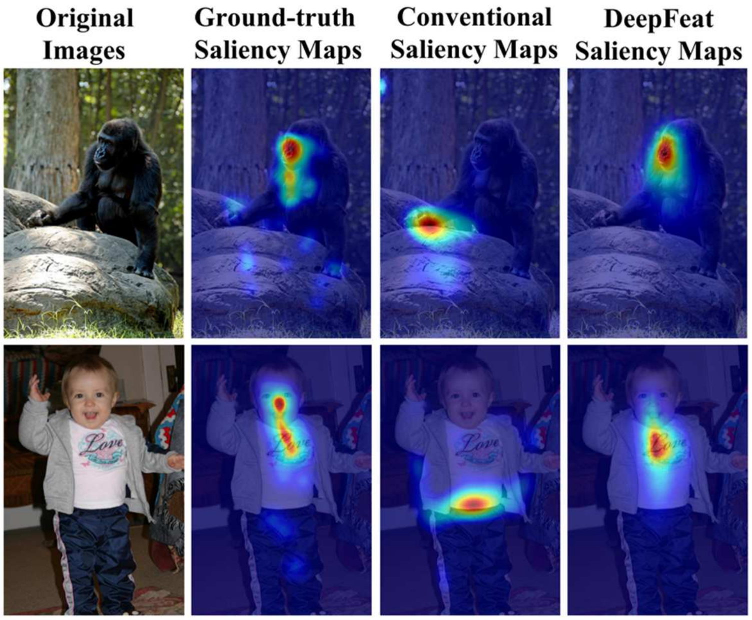


Figure 2 - Column 1 is original images, column 2 is the ground-truth maps of human fixations, column 3 is the saliency maps of a conventional saliency model [24], and column 4 is the saliency maps of the our recently developed DeepFeat model [48]. For visualization purpose, the histogram of the predicted saliency maps of both models are matched to the histogram of the dataset ground-truth.

The aim of this dissertation is to leverage the understanding of human visual attention by performing an extensive analysis, prediction, and visualization of human eye fixations. This dissertation dives deep to allow the reader to understand the previous work done in visual saliency and deep neural networks, visualize the feature maps of DCNNs, analyze infants and adults eye fixations, predict the human eye fixations, and compare deep features of DCNNs for visual saliency prediction.

1.2 Contributions:

The contributions of this dissertation can be summarized as follows:

1. **A comparison of saliency models for fixation prediction on infants and adults.**

   The gaze patterns differences between infants and adults are highlighted by using a benchmark of standard saliency models. The saliency predictions are evaluated using seven popular evaluation metrics.

2. **A proposed saliency model to predict eye fixations via deep features of DCNNs.** The first proposed model exploits deep features of DCNNs pre-trained for object recognition as optimized features to predict a saliency map. The model incorporates the deep features in a combination of bottom-up and top-down manners.
3. **An extensive analysis of deep features from various pre-trained DCNNs for saliency prediction of eye fixations.** The deep feature comparisons are conducted using four saliency implementations including bottom-up, top-down, and the combination of both with and without the incorporation of center bias. The saliency implementations are compared over seven DCNNs using both classical and CAM approaches.
4. **A proposed saliency based deep learning framework to learn from scratch.** The proposed framework consists of a data generation scheme and a modified residual network. The data generation aims to create a dataset large enough to learn a saliency model from random weights. The proposed saliency model incorporates a global contrast computation as a measure of saliency.

CHAPTER 2

BACKGROUND

2.1 Visual Saliency Computational Models:

A rich stream of saliency models has been developed [24,49,50]. These models are different in features, frameworks, applications, and the purpose which they are designed for. Although saliency models are different, they share common characteristics. Therefore, saliency models can be categorized based on these characteristics. For example, saliency models can be categorized to bottom-up (exogenous) and top-down (endogenous) models. Bottom-up saliency models are stimulus driven, where a saliency is defined as irregularity or visual rarity in a scene locally, regionally, or globally [51]. Such models can explain the scene partially as majority of eye fixations are driven by tasks. Top-down saliency models are task-driven based models, where they use prior knowledge, expectation, and reward as visual cues to locate a target of interest [52].

Saliency models also can be classified as space-based models and object-based models. There is no universal agreement whether eye fixations attend spatial locations or objects. Therefore, space or object saliency maps can be used for fixation prediction. From another aspect, saliency models can be categorized based on different task types, free viewing, visual search, and interactive tasks. In free viewing, subjects view an image freely. In visual search, subjects are asked to find a specific or odd object in an image. Interactive tasks are complex and contain subtasks like visual search, and target tracking. Other categorization factors are pointed out in previous studies [52,53]. In this section, saliency models are categorized based on the saliency computation mechanism.

2.1.1 Bayesian Models:

In visual attention, a Bayesian framework consists of a combination of sensory evidence and prior knowledge. Several Bayesian saliency models have been developed. Itti & Baldi [25] defined a surprise as a saliency in probabilistic terms, in which surprise was obtained as the Kullback-Leibler divergence (KL). Zhang et al. [29] proposed a framework that considered what the human visual system is trying to optimize. The framework was a linear combination of self-information of local image patches as bottom-up and the prior knowledge as top-down. Later, Zhang et al. [54] modified the model to predict fixations on a dynamic scene. Spatiotemporal filters were added to the model, and a general Gaussian distribution was fitted to the filter's response. Xie et al. [55] proposed a novel Bayesian framework based on low and mid-level cues. A coarse saliency region was first obtained via a convex hull. Saliency information with mid-level cues was analyzed via super pixels. A Laplacian sparse subspace clustering method grouped super pixel with local features, and then analyzed the result with respect to the coarse saliency region in order to compute the prior saliency map. Observation likelihood of the Bayesian framework was computed by the low-level cues based on the convex hull. Lu et al. [56] proposed a Bayesian framework to generate a saliency map based on reconstruction error. The model first obtained dense and sparse reconstructions, then measured the reconstruction error that propagated based on the contexts obtained from K-means clustering. Pixel level saliency was obtained by integration of multi-scale reconstruction errors. A Bayesian integral reconstructed a final saliency map from the pixel level saliency maps. Jianyong et al. [57] proposed a Bayesian framework based on BING and graph models. The model used the BING model to generate a coarse conspicuity map. A graph model was constructed after super pixel image abstraction. This operation was followed by a weighting to produce a prior map. After adaptive thresholding, the observation likelihood map was computed by color histogram. The two maps were combined via Bayesian framework.

2.1.2 Cognitive Models:

Models of saliency in early development of visual attention are biologically inspired models. Because of the biological explanations these models offer, several models were developed based on FIT. Itti et al. [24] devised the first saliency model. Several implementations of the model have been introduced including implementation of the original model [24], blur and parameters optimization [58], and an implementation for salient object detection [59,60]. The model also has been modified for several applications. For example, Itti & Koch [61] modified the first saliency model to perform a visual search for overt and covert shifts of attention. The model iteratively convolves the extracted feature maps with a two-dimensional difference of Gaussians (DoG) filter. Also, Cerf et al. [28] modified the first saliency model by adding face detection as a low-level feature, then performed similar feature competition and combination to emerge a saliency map. Other cognitive models have been proposed independently of the first saliency model. For example, Le Meur et al. [62] proposed a bottom-up model of visual attention. The model used contrast sensitivity functions, perceptual decomposition, visual masking, and center surround interactions as some of the features implemented in the model. Later, Le Meur [63] extended the model to spatiotemporal domain. The algorithm fuse saliency maps from achromatic, chromatic, and spatiotemporal channels. Kootstra et al. [42] proposed humans are sensitive to symmetry in visual patterns and developed three symmetry saliency models based on isotopic symmetry, radical symmetry, and color symmetry. Marat et al. [64] proposed a spatiotemporal saliency model for fixation prediction in video during free viewing task. The model extracted two signals that correspond to parvocellular and magnocellular. The signals were divided into elementary feature maps by cortical-like filters. The feature maps generated a static and saliency maps. Then, the two maps were fused into a spatiotemporal saliency map. Murray et al. [65] proposed a model for color appearance in human vision. The proposed model extracted color and luminance features followed by multi-scale decomposition. Multi- scale integration was performed by inverse wavelet transform. Cognitive models were beneficial, because their further development helped to better understand the neural processing of visual information.

2.1.3 Decision theoretic models:

The hypothesis of such models assumes that the perceptual system produces optimal decisions about the state of the surrounding environment. The disadvantage of decision theoretic models is optimality should be driven with respect to the end task. Gao & Vasconcelos [66] defined a top-down saliency as classification with minimal prediction error. DoG and Gabor filters were used to measure the saliency of a particular location in an image as the Kullback-Leibler divergence of the filter response of the location and the histogram of filter response of the surrounding regions. This work was extended by Mahadevan & Vasconcelos [67] to provide a spatiotemporal saliency based on biologically inspired mechanisms of motion. The model combined center surround saliency and dynamic texture. Guo & Zhang [68] proposed an attention selection model with visual memory and online learning, which consists of a sensory mapping, a novel cognitive mapping, and motor mapping. The proposed work also used Amnesic Incremental Hierarchical Discriminant Regression Tree to guide the removal of redundant information. Gu et al. [69] proposed an attention selectivity model for automatic fixation generation in a 2D space. An activation map was created by extracting early visual features and detecting meaningful objects. A retinal filter was applied on the activation map to generate regions of interest. Focus of attention was determined over the regions of interest using a belief functions based on perceptual costs and rewards. The time of fixation over the regions of interest was estimated by memory learning and decaying model. Gao et al. [70] proposed a top-down saliency rooted in a decision theoretic interpretation of perception. The model detected suspicious coincidences using Barlow's principle, which provides two solutions for a discriminant saliency, feature selection, and saliency detection.

2.1.4 Spectral analysis models:

A majority of saliency frameworks are processed to measure irregularities in the spatial domain. Irregularities can also be measured in the frequency domain. Several studies used the Fourier transform and its spectral analysis to compute a saliency map. Hou & Zhang [71] analyzed the amplitude spectrum of the Fourier transform, and proposed a spectral residual saliency model. The model was independent of features, parameters, and prior knowledge. Wang & Li [72] extended the residual spectral approach by adding feature based on gestalt principles to detect similarity and continuity. Li et al. [73] proposed a bottom-up approach for saliency detection. The authors demonstrated a convolution of the image amplitude spectrum with a low pass Gaussian kernel of appropriate size is equivalent to a saliency detector. Beside the amplitude spectrum, Guo & Zhang [74] pointed out the phase spectrum is the key to saliency modeling in the frequency domain, and then proposed a novel multiresolution spatiotemporal saliency detection model based on the phase spectrum [12]. Other saliency models have been proposed in the frequency domain. For instance, Achanta et al. [75] proposed a frequency tuned salient region detection. The model used color and luminance as low-level features. Then, saliency was obtained as the difference between the mean image feature vector, and the smoothed version of the original image. Brian & Zhang [76] proposed a biologically plausible saliency detection method based on spectral whitening. The method used a divisive normalization as estimator of spectral whitening. Li et al. [77] proposed a saliency model that combines two channels of the processed image in the frequency and spatial domains. The frequency domain channel suppressed non-distinctive patterns of the image by spectrum smoothing. The spatial domain channel enhanced those patterns by using center surround mechanism akin to mechanism in visual cortex. Xiao et al. [78] used hypercomplex discrete cosine transform for salient object detection approach based on human perception inconsistent scale. The method extracted local spectral feature, then sparse energy spectrum was calculated on local regions as visual stimulation. A visual saliency was measured on the local region and neighbor regions. A multi-scale response was performed on the saliency map.

2.1.5 Graphical Models:

A probabilistic framework where a graph represents a conditional independence structure between random variables. Graphical models treat eye fixation as time series. Several saliency models have been introduced in this category. Models in this category exploit approaches like hidden Markova, dynamic Bayesian networks, and conditional random field (CRF). Salah et al. [79] proposed an attention model based on the primate selective attention mechanism. The model was applied on face detection and handwritten digits. A bottom-up saliency map was constructed from simple features. At each region of the image, single layer perceptron was trained. Finally, the information gained were combined using an observable Markova model. Rao [80] proposed a model to modulate attention in particular image locations to neuron in V2 and V4 areas of the visual cortex. The model interpreted perception as an estimation of posterior probability of features and their location in the image using a Bayesian graphical algorithm called belief propagation. Liu et al. [81] devised a salient object detection method. They proposed multi-scale contrast, center surround histogram, and color spatial distribution as image features. Then a saliency map was emerged by learning CRF to combine the proposed features. Later, a motion feature was added to extend the model to be applied on videos [31]. A dynamic programming algorithm was devised to solve a global optimization problem. The salient object sequence detection was obtained by CRF framework. Yang et al. [82] proposed ranking the similarity of image elements with foreground cues or background cues via graph based manifold ranking. Super pixels were created and treated as nodes. Then, a k-regular graph is used to exploit the spatial relationships between the nodes. Ling et al. [83] proposed a novel saliency detection algorithm via a graph model and statistical learning. The algorithm used manifold ranking to create an initial saliency map. Then, the saliency map was optimized with absorbing Markova chain. Finally, statistical learning was performed by Bayes estimation with color statistical models to assign saliency values to pixels and refine the saliency map. Zhang et al. [84] proposed a novel graph-based optimization for salient object detection. The proposed framework employed multiple graphs to describe the complex information in the image. In the proposed work visual rarity was modeled to make the optimization framework suitable for saliency detection.

2.1.6 Information theory models:

Models in this category measure irregularity in image locations by maximizing the information sampled from surrounding environment. Such models select the most informative locations and discard the rest. Benninger et al. [85] developed a saliency model that select fixations at informative locations of the image, which reduce overall uncertainty about the visual stimulus. The model reconstructed visual information from a sequence of human fixations. After each fixation, the next fixation was selected as the fixation that would minimize the uncertainty of the stimulus. Seo & Milanfar [86] proposed a novel framework for saliency detection over static and space-time stimuli. The model computed the local regression kernels in an image to measure the likeness of a pixel to the surroundings. Then, saliency map was computed by kernel density estimation as local self-resemblance. Bruce & Tsotsos [87] built an attention model based on computational constraints derived from efficient coding and information theory. The proposed framework was an extension to previous framework based on self-information maximization. Li et al. [73] proposed a novel saliency detection method for image and video. In the method proposed, saliency was defined as minimum conditional entropy of local regions. Conditional entropy was treated as the lossy coding length of multivariate Gaussian data. The final saliency map was reconstructed by pixels and segmented to detect proto-objects. Wang et al. [88] proposed a computational model inspired by information maximization for gaze shifts prediction. The model computed three filters' responses as a coherent representation for reference sensory responses, fovea periphery resolution discrepancy, and visual working memory. Response maps from the three filters were combined into multi-band residual filter response maps, where the residual perceptual information was computed at every location. Klein et al. [89] introduced a salient object detection method, which has similar structure to cognitive models but acknowledge a saliency via information theoretic concept. The model extracted features, performed center surround operations, and computed feature maps. Riche et al. [90] proposed a bottom up saliency model based on locally contrasted and globally rare features were salient. The model extracted luminance and chrominance as low-level features. Then, image orientations were extracted as mid-level features. The extracted features were segmented using Otsu method. Then, multi-scale rarity mechanisms were performed. Finally, scaled maps were fused and normalized.

2.1.7 Learning based models:

Learning models are data driven functions to select, re-weight, and integrate the input visual stimuli. Such models learn a saliency map from human fixations. Majority of models in this category use a combination of bottom up and top down features to increase fixation prediction of the model. Learning based models can be categorized to supervised and unsupervised learning models. Supervised learning models learn a function from a labeled training data. For example, Peter & Itti [91] trained a simple regression classifier to capture the task dependent association between a given scene and the preferred gaze locations while human participants play video games. Kienzle et al. [92] introduced a non-parametric bottom up learning based saliency model. A support vector machine was trained to compute the saliency in local image patches. Similarly, Judd et al. [30] used low, mid, and high-level features to learn a saliency model using a support machine vector (SMV). Unsupervised learning models learn to predict from unlabeled training data. Several deep learning based saliency models have been developed [93-95]. Deep learning based saliency models are composed of multiple layers to learn representation of images with multiple levels of abstractions. Vig et al. [96] proposed the first deep learning based saliency model, which incorporates biologically inspired features and uses the standard learning pipeline. Kummerer et al. [97] presented a novel way to reuse existing object recognition neural networks for human fixation prediction. The model used Krizhevsky network to compute filter responses and a full convolution to learn the saliency model. Furthermore, another probabilistic model was also introduced [98]. The model used VGG-19 features and incorporated center bias. A maximum likelihood learning was used to train the model. Huang et al. [99] proposed a top down saliency model using deep convolutional neural networking (DCNN). The model used AlexNet, VGG-16, and GoogLeNet. These DCNNs contained several max-pooling layers, and a large number of convolutional and nonlinear layers between pooling layers. Kruthiventi et al. [100] proposed a fully convolutional neural network (CNN) for predicting human fixations. The model incorporated a novel location biased convolutional layer to model location dependent patterns. Liu & Han [101] proposed a deep spatial contextual long-term recurrent convolutional network to predict human fixations in natural scenes. The model learned saliency related to local features in parallel, and integrated scene context to mimic the cortical lateral inhibition mechanisms in human visual system. Jetley et al. [102] introduced a saliency model via probabilistic distribution prediction. The model was formulated as generalized Bernoulli distribution. They trained DNN using a novel loss functions that paired a SoftMax activation function with measures designed to compute distances between probability distributions. Corina et al. [95] proposed a novel DNN structure that combines features extracted at different levels of a CNN. The model consisted of three main blocks: a feature extraction CNN, a feature encoding network, and prior learning network.

2.1.8 Other models:

Categories of saliency models are interconnected. Some saliency models can fit into more than one category. On the other hand, some models do not fit to any of the aforementioned categories. In this section, models that are not fit to previous categories are briefly reviewed. Erdem & Erdem [103] addressed the integrating issue of features and proposed to exploit region covariance descriptors meta-features for saliency detection. These descriptors captured local structure information by encoding pair-wise correlations over features. Zhang & Sclaroff [33] introduced a novel saliency model based on Boolean mapping. Color feature was extracted, and Boolean maps were created from the feature map with random thresholds. Mean attention map was obtained over the randomly generated Boolean maps. The resultant attention maps were normalized then linearly combined. Liu et al. [104] proposed a novel saliency detection framework in a form of tree. The proposed framework simplified the input image to regions using adaptive color quantization and region segmentation. An initial regional saliency was formed by integrating global contrast, spatial sparsity, and object prior with regional similarities. A proposed saliency directed region merging approach with dynamic scale control scheme to create the saliency tree. A leaf node indicated primitive region, while a non-leaf node indicated non-primitive region. A regional center-surround scheme-based node selection criterion was exploited to generate a final regional saliency map.

2.2 Datasets of human fixations:

In order to evaluate how well a saliency model can predict the human visual attention, the existence of ground-truth maps is crucial. Therefore, a variety of human fixations datasets are collected using remote eye trackers. An eye tracker records the human eye movements (saccades). Researchers setup a delay threshold to label a set of fixations. An eye fixation is defined as a point position in the Cartesian coordinate system.

The human fixation datasets are collected for a variety of tasks, including visual search, memory, and free-viewing. The visual search task aims to detect the covert and overt shifts of attention during a visual search for a specific object. The memory task studies the attentional regions that lead to memorizing objects. The free-viewing task focuses on recording human fixations without prior knowledge about the image. In this dissertation, the fixation datasets exploited are free-viewing based datasets. In this section, the datasets used in this dissertation are described. Figure 3 presents samples from all six datasets.

1. **Infants & Adults:** Contains 16 indoor and outdoor images. These images include human objects either in the foreground or in the background. The resolution of the images is

1680 × 1050 pixels. Images were presented for 5 seconds to 20 observers, including 10 infants and 10 adults [105].

2. **MIT1003:** Consists of 1003 images. The resolution of the images is fixed on one-dimension 1024 pixels, and on the other dimension it ranges from 678 to 768. Fifteen observers (age = 18 to 35 years) freely viewed the MIT1003 images. Images were presented to 15 observers for 3 seconds [30].

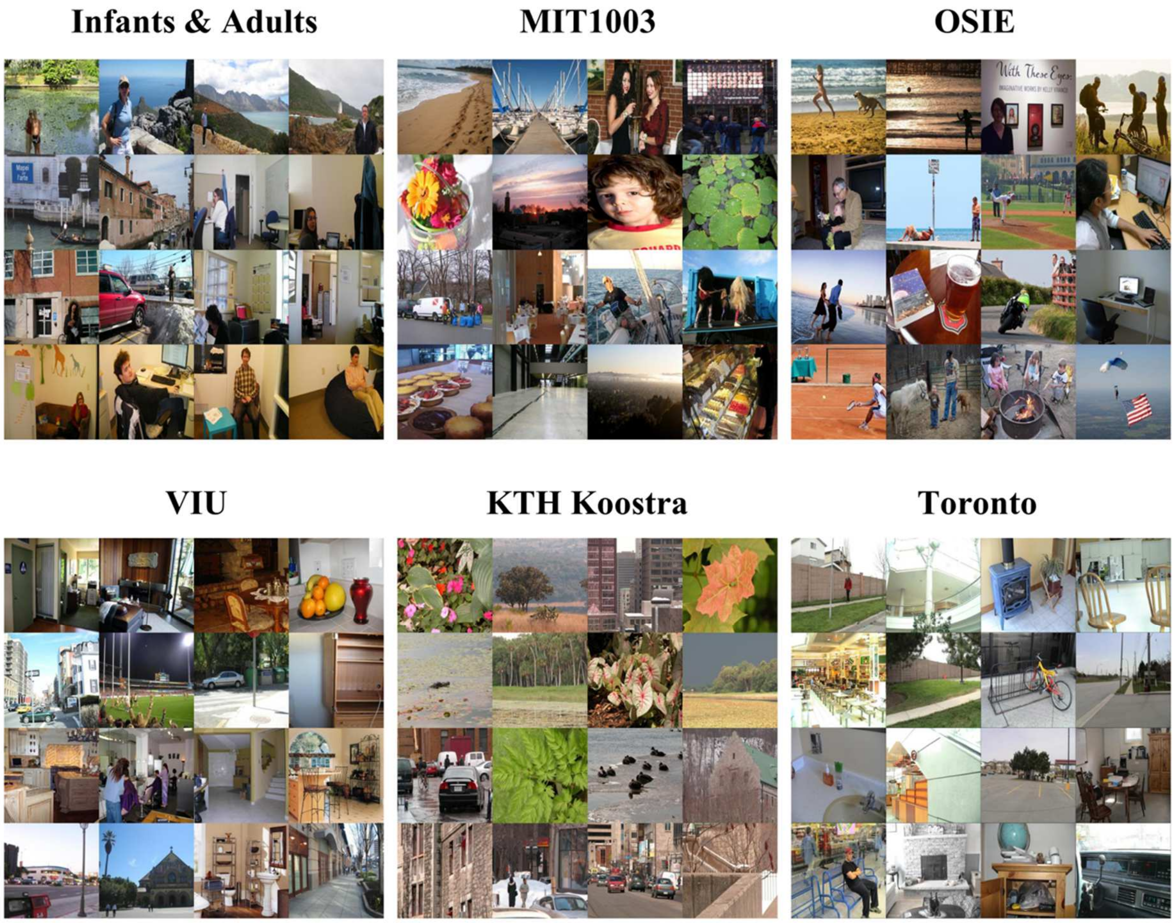


Figure 3 - Subsets of the six datasets used in this dissertation.

3. **VIU:** Consists of 800 indoor and outdoor images. The resolution across all images is 405 × 405 pixels. This dataset consists of multiple tasks (explicit saliency judgement, free-

viewing, saliency search, and cued object search). In this dissertation, human fixations under free-viewing conditions are exploited. 22 observers (age = 18 to 23 years) viewed every image for 2 seconds [106].

4. **KTH Koostra:** Contains 100 images from 5 categories. The resolution of the images is 1024 × 768 pixels. Images are free-viewed by 31 observers for 5 seconds [107].
5. **OSIE:** Includes 700 images. The resolution of the images is 800 × 600 pixels. Images are presented to 15 observers for 3 seconds [108].
6. **Toronto:** Contains 120 images. Several images in this dataset do not have regions of interest. The resolution of the images is 681 × 511 pixels. Images were presented to 20 observers for 4 seconds [109].

2.3 Evaluation Metrics:

Performance of saliency models is often compared to human fixation maps using evaluation metrics to assess the agreement between a saliency map and human fixation maps. In this dissertation, two binary classification measures and six evaluation metrics are used for evaluating saliency models. The motivation of analyzing saliency models with seven metrics is to ensure the conclusions drawn are independent of the choice of metric and consistent across all metrics. Overall, a good saliency model should perform well across all metrics.

The two binary classification measures are based on the intersection area between predicted saliency and human fixations, including receiver operating characteristics (ROC) and precision-recall (PR). From the ROC measure, the area under ROC curve (AUC) is reported as the first evaluation metric. Also, F-measure (metric) score is obtained from PR. Moreover, four metrics measuring the similarity, and two metrics measuring dissimilarity between a saliency map and a ground-truth fixation map are also used in this dissertation [110]. Four similarity-based metrics are normalized scan-path saliency (NSS) information gain (IG), similarity (SIM), and Pearson's correlation coefficient (CC). Two dissimilarity-based metrics are Kullback-Leibler divergence (KL), and earth mover's distance (EMD). Table 1 presents the evaluation metrics used in this dissertation.

Table 1 - A description of evaluation metrics.

Metric	**Denoted as**	**Theoretical range**
Area under the ROC curve	AUC	[0,1]
F measure	F-measure	[0,1]
Normalized Scan-path Saliency	NSS	[-∞,∞]
Information gain	IG	[-∞,∞]
Similarity	SIM	[0,1]
Pearson's Correlation Coefficient	CC	[-1,1]
Kullback-Leibler divergence	KL	[0, ∞]
Earth moving distance	EMD	[0, ∞]

1. **ROC:** Treats a saliency map as a binary classifier of human fixations over a set of thresholds. It plots the tradeoff between true positive and false positive rates at various thresholds of the saliency map. True positive rate (TRP) and false positive rate (FPR) are formally defined:

$$TPR = \frac{TP}{TP+FN} \quad (1)$$

$$FPR = \frac{FP}{FP+TN} \quad (2)$$

   where TP is fixated saliency map values above threshold, FP is un-fixated saliency map values above threshold, FN is the fixated saliency map values below threshold, and TN is un-fixated saliency map values below threshold.

2. **PR:** Another binary classifier, it plots the tradeoff between precision and recall for various saliency map thresholds. The precision and recall are calculated by:

$$Precision = \frac{TP}{TP+FP} \tag{3}$$

$$Recall = \frac{TP}{TP+FN} \tag{4}$$

3. **AUC:** Is the integral of the area under the ROC curve. A score higher than 0.5 indicates a prediction higher than random guessing. Throughout this dissertation, three AUC are exploited based on a unique ROC sampling and human annotation processing. Judd-AUC computes the true positive rate and false positive rate over every pixel value in the saliency map. The Borji-AUC computes the true positive rate and false positive rate over a set of thresholds sampled from the dynamic range of the saliency map. Both Judd-AUC and Borji-AUC compare saliency maps to the exact fixation points of the human fixations. A third AUC modifies the Borji-AUC by utilizing fixation maps that reflects a continuous distribution of eye fixations.
4. **F-measure:** Is a weighted harmonic mean of precision and recall. It often used because precision or recall individually cannot evaluate a saliency map. Formally:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall} \tag{5}$$

where β is a threshold suggested by previous work, β^2 = 0.3 to raise more importance to precision [111]. A β^2 is computed across the thresholds. Then, the maximum β^2 represents the maximum overlap between precision and recall along the curve. A score closer to 1, it indicates the overlap between the saliency map and ground-truth fixation map is large.

5. **NSS:** Is the normalized scan-path saliency that measures the average saliency value at the exact fixation locations:

$$NSS(S,F) = \frac{1}{N} \sum_i S'_i \times F_i \tag{6}$$

Where

$$S' = \frac{S-\mu(S)}{\sigma(S)} \quad (7)$$

where *N* denotes the number of fixation points, and *i* indexes the fixation points of the binary fixation map. A score of zero corresponds to random guessing. A positive score indicates correspondence between the two maps, and a negative score denotes anti-correspondence.

6. **IG:** Evaluates the information gain over a center bias map. It can handle center bias and it has an interpretative linear scale:

$$IG\big(S, G(x,y)\big) = \frac{1}{N}\sum_i G(x,y)_i[log_2(\epsilon + P_i) - log_2(\epsilon + B_i)] \quad (8)$$

where *S* is a saliency map, *G* is a ground-truth fixation map, *x* and *y* are the coordinates of the exact fixation location, *N* is the number of fixations, *B* is the center bias map, and *ε* is a small value for regularization. IG has to be larger than zero. A center bias map is emerged by averaging the ground-truth fixation maps of all other images in the dataset. A positive score indicates a saliency model prediction outperform the center bias map. A negative score indicates the saliency model prediction cannot compete with the center bias map.

7. **SIM:** A measure of intersection between two distributions. It measures the similarity between a saliency map and a fixation map:

$$SIM(S, G) = \sum_i min(S_i, G_i) \quad (9)$$

where

$$\sum_i S_i = \sum_i G_i = 1 \quad (10)$$

A positive score indicates an intersection between the saliency map and the fixation map, while a score of 0 indicates no intersection between the two maps.

8. **CC:** Is an evaluation of the linear relationship between a saliency map and a fixation map. It treats the saliency map and the fixation map as random variables and measures the dependence between the two variables:

$$CC(S,G) = \frac{cov(S,G)}{\sigma(S)\sigma(G)} \tag{11}$$

where $cov(S,G)$ is the covariance between the saliency map and the fixation map. A score equal to -1 or 1 indicates a perfect correlation, and a score of 0 indicates no correlation between the two maps.

9. **KL:** Is a probabilistic interpretation of the saliency and fixation maps. It measures the loss of information when a saliency map approximates a fixation map:

$$KL(S,G) = \sum_i G_i log\left(\epsilon + \frac{G_i}{\epsilon + S_i}\right) \tag{12}$$

As a dissimilarity metric, a score of 0 indicates the saliency map and the ground-truth fixation map are identical.

10. **EMD:** Is another dissimilarity metric that measures the spatial distance between two distributions. Computationally, it is the minimum cost required to move one distribution to another. Formally:

$$\widehat{EMD} = \left(min \sum_{i,j} f_{ij} d_{ij}\right) + \left|\sum_i S_i - \sum_i G_i\right| max d_{ij} \tag{13}$$

$$s.t. f_{ij} \geq 0, \sum_j f_{ij} \leq S_i, \sum_i f_{ij} \leq G_i$$

$$\sum_{i.j} f_{ij} = min\left(\sum_i S_i, \sum_j G_j\right)$$

where fij is the flow be transported from supply i to demand j, and $dij$is the ground distance (cost) between bin i and bin j in the distribution. A score of 0 indicates the distribution in the saliency map and the distribution in the fixation map are identical. As the score increases, the distance between the two distributions increases.

CHAPTER 3

DEEP FEATURES OF DEEP LEARNING NEURAL NETWORKS

3.1 Introduction:

Deep neural networks (DNNs) have achieved significant performances recently. Such neural networks can be categorized as multi-layer perceptron (MLP), recurrent neural networks (RNN), deep belief networks (DBN), generative adversarial networks (GAN), and convolutional neural networks (CNN). In image processing and computer vision, CNNs are more intuitive to be used than other DNNs due to the correlation of neighboring pixels.

CNNs are inspired by the biological process of visual information as a simulation of the visual information transmission pattern among neurons of the visual cortex [112]. In such architecture, each cortical neuron responds to a stimulus in a receptive field. The receptive fields of different neurons overlap with each other to cover the entire stimuli. In 1989, LeCun et al. learned convolutional kernels coefficients of hand-written digits using backpropagation [113]. In 2012, Krizhevski et al. achieved a breakthrough of classification performance using the concept of deep learning [114]. Later, a large number of CNNs have been proposed to achieve higher classification accuracy by increasing the depth of neural networks [115-121].

While a large number of studies achieved outstanding performances, CNNs used in this dissertation are reviewed. In addition, the learning model, and deep features mathematical computations are also reviewed in this section.

3.2 DCNN Formalization:

1. **Convolution:** Is a filter kernel that learn its weights by convolving such kernel with the input data tensor. Such operation can be formalized by:

$$F = \theta^T x + b \tag{14}$$

where θ^T denotes the weighting filter, x is the tensor of data, and b is the bias vector. Several parameters effect the output of a convolutional operation including number of filters and strides. A number of filters specifies the number of output feature maps. A stride is the distance in pixels between two pixels. For example, when stride equals 2, the convolution is computed for every other pixel causing to down sample the input tensor of data.

2. **Activation:** Is an operation that transforms the input from linear to nonlinear tensor of data. In deep learning, a variety of activation functions are popular including sigmoid, tanh, rectified linear units (ReLU), etc. All CNNs presented in this dissertation exploit ReLU activation function which can be mathematically defined as:

$$A = \max(0, F) \tag{15}$$

   Another popular activation is SoftMax, which is used to determine the probability of classified objects. The SoftMax can be formalized as:

$$P = \frac{e^{-F}}{\sum e^{-F}} \tag{16}$$

3. **Pooling:** Is a down-sampling operation that can be performed locally or globally. A local pooling function down-samples local image regions by a factor. A global pooling function returns a scalar value for every 2D feature map. A max pooling and average pooling are two pooling operations exploited by the CNNs presented in this dissertation. All presented CNNs exploit local max pooling functions. In addition, such CNNs exploit global max pooling (GMP) and global average pooling (GAP).
4. **Batch Normalization:** Is a normalization operation developed to overcome the problem of vanishing values. A batch normalization is learned by:

$$F = \gamma \hat{x}_i + \beta, i \ \in \{1, \ldots, \mathrm{m}\} \tag{17}$$

Where γ and β are learning parameters, and m is the number of mini-batches. Moreover, $\hat{x}$ is a normalization function formalized by:

$$\hat{x}_i = \frac{x_i - \mu_{\mathcal{B}}}{\sqrt{\sigma_{\mathcal{B}}^2 + \epsilon}} \tag{18}$$

The$\mathcal{B}$denotes a mini-batch that consists of m samples, and ϵ is a constant. Also, μ and σ are the mean and standard deviation of mini-batch $\mathcal{B}$.

5. **Dropout:** Is a regularization technique that aims to reduce over-fitting of neural networks. The Dropout layer is presented only in the training phase, where a constant represents the percentage of neurons to be randomly assigned to zero. In the testing phase, the Dropout layer is ignored by assigning the constant to zero.
6. **Fully Connected Layer:** Is a layer where the receptive field is an entire channel of the previous layer. A fully connected layer (FC) is usually followed by an activation layer.

3.3 Model Hyper Parameters:

In order to learn a CNN model, several hyper parameters can be adjusted manually including batch size, number of epochs, learning rate, momentum, and weight decay. A mini-batch is the number of samples required for a single iteration update. An epoch is computation of all samples over the entire dataset. Learning rate is a step size considered during training that controls the speed of the training process. Moreover, momentum is a method for accelerating the training process by moving the average of the gradient. A weight decay is a technique that prevents the learning weights from over-growing.

3.4 Learning Model:

A typical learning model consists of a forward kernel, cost function, backward kernel, and an optimization function. The forward kernel consists of a subset of convolutional layers as described previously. The cost function (also known as loss) is a function that compares the prediction of the forward kernel and the annotation label. CNNs used in this dissertation exploit the cross-entropy cost function. The backward kernel estimates the loss in prediction over the convolutional layers of the forward kernel using backpropagation. Moreover, an optimization is a technique of updating weights of the CNN. Stochastic gradient decent (SGD) is a first order optimization algorithm designed to find local minima of an objective function. The SGD reduces the prediction error rate as a function of training epochs. Adam is another iterative optimizer that updates the learning weights by using an adaptive learning rate from estimates of the first and second momentums of the gradients.

3.5 Convolutional Neural Networks:

Seven benchmark CNNs are exploited in this dissertation. Such CNNs are pre-trained for object classification and scene classification using the ImageNet and Places205 datasets. The architecture differences between these models are described below:

1. **AlexNet:** Consists of five convolutional layers followed by two fully connected (FC) layers and a probability layer [114]. The first two convolutional layers consist of 11×11 filter size and 96 feature channels, and 5×5 filter size with 256 feature channels. Each convolution layer is followed by a max pooling layer. The next three convolution layers exploit 3×3 filter size and the number of feature channels is 384, 384, and 256, respectively. Using a global maximum pooling (GMP), the two FC layers consist of 4096 neurons each. The FCs are followed by a SoftMax layer, which consists of 1000 class of objects. The architecture of the model can be illustrated in Figure 4.

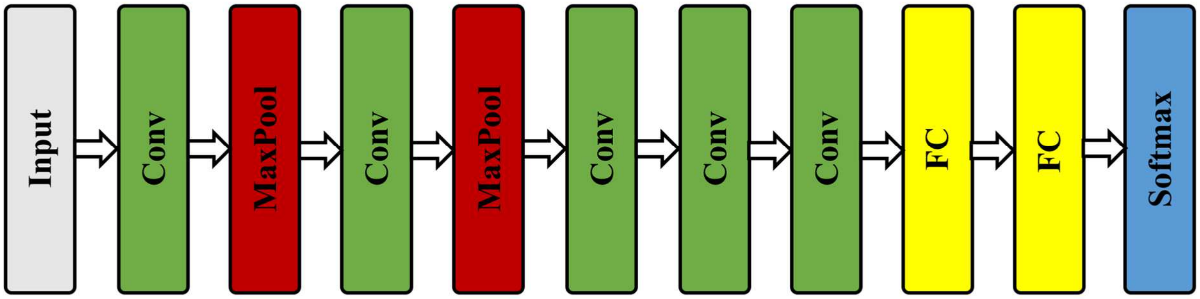


Figure 4 - Architecture of the AlexNet CNN model. Conv: convolution layer, MaxPool: max pooling layer, and FC: fully connected layer.

2. **VGG:** Demonstrated that the depth of the neural network is a critical component of object classification [122]. A general architecture of the VGG start with two blocks of two convolution layers followed by a max pooling layer. In addition, VGG employs three FC layers followed by a SoftMax layer. Figure 5 presents the general architecture of a VGG.

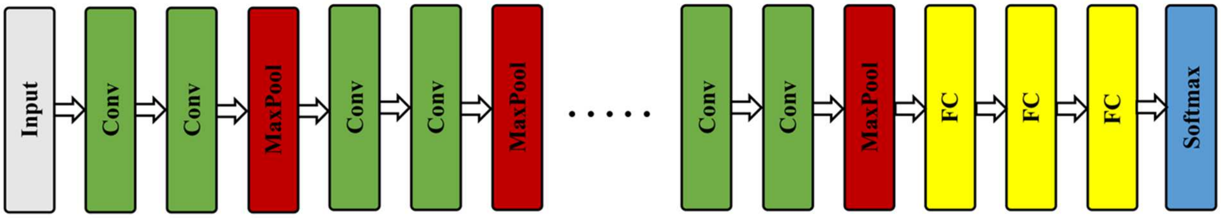


Figure 5 - General architecture of VGG. Conv: convolution layer, MaxPool: max pooling layer, and FC: fully connected layer.

Several VGG variants have been used in a variety of applications. In this dissertation, two variants of VGG are exploited: VGG16 and VGG19. As the names indicate, VGG16 consist of 16 layers. The first four convolution layers comes from the general VGG concept followed by three blocks of convolutions. Each block consists of three convolution layers followed by a max pooling layer. Similarly, VGG19 employs the first four convolutional layers of the general VGG concept followed by three blocks of four convolution layers followed by a max pooling layer. The complete structure of VGG16 and VGG19 is presented in table 2.

Table 2 - Configuration settings of VGG16 and VGG19 variants.

VGG16	**VGG19**
16 weight layers	19 weight layers
Input (64 × 64 RGB Image)	
Conv (3 × 3)-64	Conv (3 × 3)-64
Conv (3 × 3)-64	Conv (3 × 3)-64
Max Pooling	
Conv (3 × 3)-128	Conv (3 × 3)-128
Conv (3 × 3)-128	Conv (3 × 3)-128
Max Pooling	
Conv (3 × 3)-256	Conv (3 × 3)-256
Conv (3 × 3)-256	Conv (3 × 3)-256
Conv (3 × 3)-256	Conv (3 × 3)-256
	Conv (3 × 3)-256
Max Pooling	
Conv (3 × 3)-512	Conv (3 × 3)-512
Conv (3 × 3)-512	Conv (3 × 3)-512
Conv (3 × 3)-512	Conv (3 × 3)-512
	Conv (3 × 3)-512
Max Pooling	
Conv (3 × 3)-512	Conv (3 × 3)-512
Conv (3 × 3)-512	Conv (3 × 3)-512
Conv (3 × 3)-512	Conv (3 × 3)-512
	Conv (3 × 3)-512
Max Pooling	
FC-4096	
FC-4096	
FC-1000	
SoftMax	

3. **GoogLeNet**: Reduced the computation complexity in comparison to traditional CNNs [123]. The model introduces an inception module which incorporates variable receptive fields created by different kernels sizes. Figure 6 presents the architecture of the inception

module.

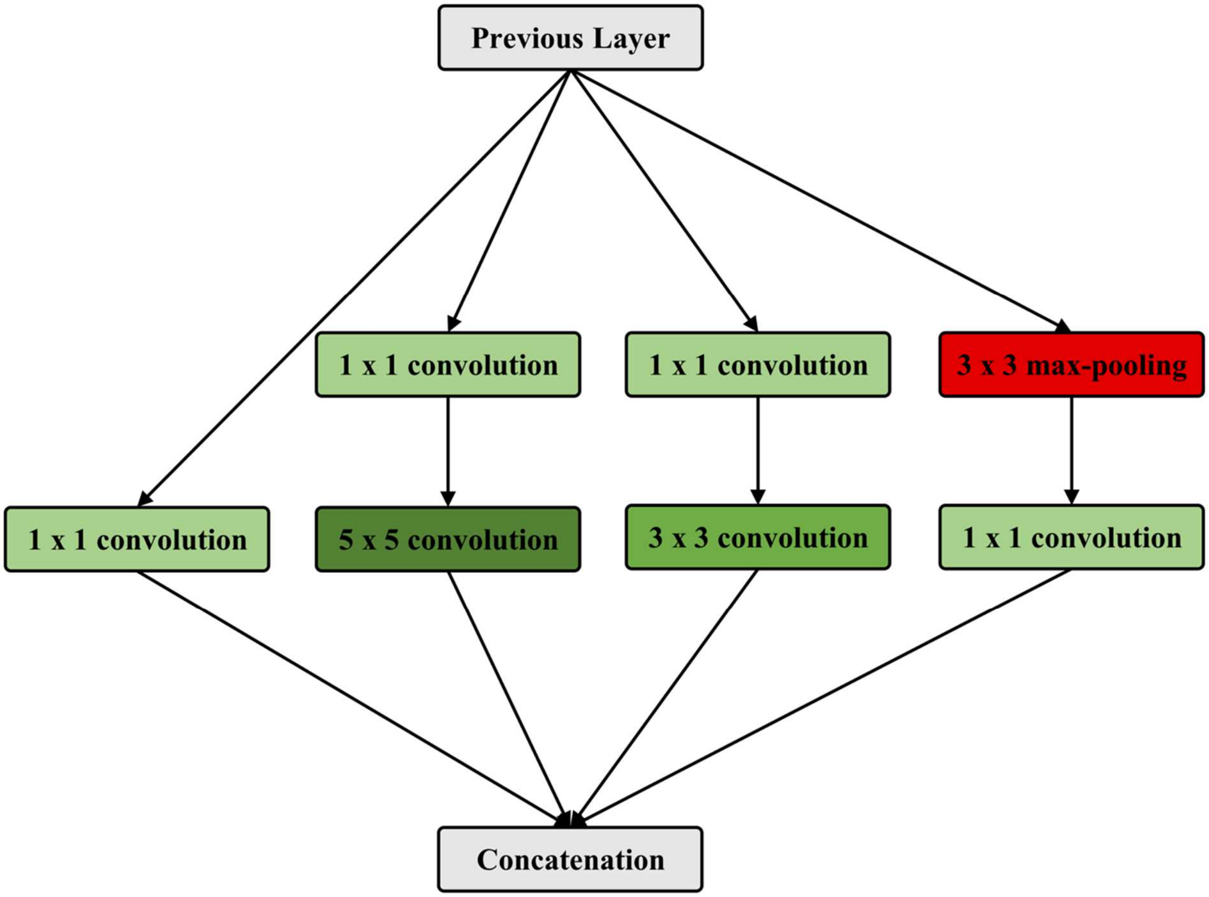


Figure 6 - Architecture of the inception module.

The GoogLeNet consists of nine inception modules following three convolution layers and two max pooling layers. One max pooling is after the first convolution layer. Another max pooling layer is after the third convolution layer. In addition, the model employs a single FC layer followed by a SoftMax layer.

4. **ResNet:** Utilizes identity learning by introducing residual paths known as skip connections [124]. A skip connection structure varies based on the ResNet variant. Figure 7 presents the general architecture of a residual block, which consists of the summation

of skip connection and previous layer in a residual network that consists of 50 layers.

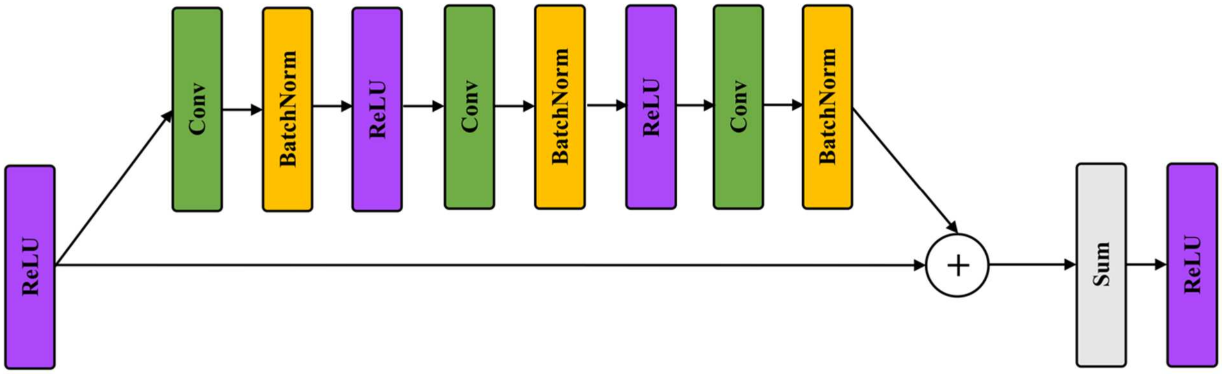


Figure7 - Architecture of a ResNet50 residual block.

A residual neural network solves the problem of vanishing gradients by using batch normalization after every convolution layer. Moreover, the skip connections allow for ultra-deep CNN development. Several variants of ResNet have been implemented including 18, 20, 34, 50, 101, 152, 1202, etc. The ResNet50 is a popular variant that consists of 49 convolution layers followed by one FC layer. Figure 8 presents deep feature maps extracted from a variety of layers of ResNet50.

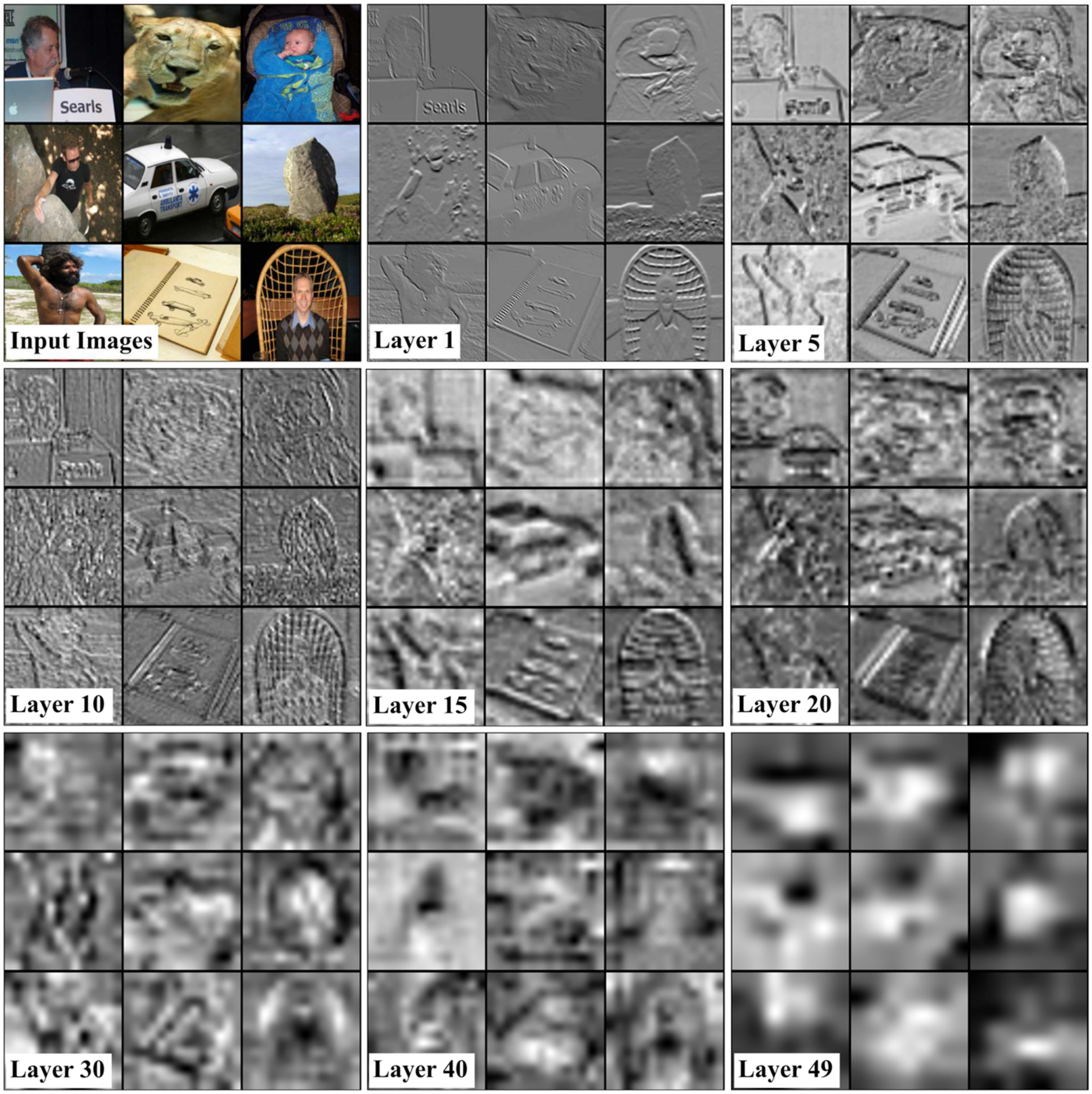


Figure 8 - Visualization of deep features of layer 1, 5, 10, 15, 20, 30, 40, and 49 of ResNet50. In each visualized layer, one convolution feature is randomly selected and presented.

3.6 Neural Network Parameters:

One way to measure the computation cost of DCNNs, is to count the number of learning weight values known as parameters. For a convolution layer, the calculation of number of parameters can be formalized as:

$$Param_{\ell} = w_x \times w_y \times f_{\ell-1} \times f_{\ell} + b_{\ell} \qquad (19)$$

Theℓ is the layer number,$w_x \times w_y$ is the weight window dimensions in the x and y direction, f is the number of feature channels, and b is the bias. The total number of parameters in every network in this chapter is presented in table 3.

Table 3 - Presents the number of parameters of CNNs described in this chapter.

CNN	**Number of Layers**	**Number of Parameters**
AlexNet	5	62.4M
VGG16	16	138.4M
VGG19	19	143.7M
GoogLeNet	22	6.8M
ResNet50	50	25.4M
ResNet101	101	44.5M
ResNet152	152	60.2M

CHAPTER 4

ANALYSIS OF INFANTS AND ADULTS EYE FIXATIONS

4.1 Introduction:

A large body of computational models are proposed to predict human fixations. To measure the agreement between a computational model, several evaluation measures are introduced. Authors compared their proposed saliency model to a benchmark of popular models [125,126]. For ease of conducting comparisons, several authors provided a publicly available datasets of images and human annotations recorded from eye tracking experiments. Several authors also provide the source code for their computational models, evaluation metrics, and fixation map generation. Because of the availability of the datasets and source codes, researchers conducted fundamental comparisons. Nowadays, a comparison of saliency models is conducted over two image datasets [127]. 68 saliency models and 5 baselines are compared over the first dataset. 22 saliency models and 5 baselines are compared over the second dataset. Saliency models are compared over eight evaluation metrics including AUC Judd, AUC Borji, sAUC, NSS, CC, SIM, KL, and EMD. Borji et al. [128] compared 32 models for prediction of fixation location and scan-path sequence. A shuffled area under the ROC curve (sAUC) is used to analyze the models. Then, explored challenges such as center bias and blurring. Borji et al. [129] evaluated 35 models over 54 synthetic patterns, and three natural image datasets. Then, used three metrics to evaluate the performance of the computational models. These metrics are: Area under the ROC curve (AUC), normalized saliency scan-path (NSS), and Pearson’s correlation coefficient (CC). Finally, tackled challenges of the comparison, including center bias, borders effect, scores, and parameters. Judd et al. [130] compared 10 computational models and 3 baselines over a dataset of 1003 images and annotations recorded from 39 observers. Models were compared using 3 metrics AUC Judd, similarity (SIM), and earth movers' distance (EMD). Then, optimized the center bias and blur for all models in the dissertation. Borji et al. [111] compared 40 models including 28 salient object detection models, 10 fixation prediction models, one object detection model, and one baseline model. The comparison is conducted over six datasets. Models were compared using 3 metrics AUC, F-measure, and mean absolute error (MAE).

All previously proposed saliency models were compared to human recorded fixations using a several eye tracking datasets. The recorded eye tracking datasets were collected from human adults, because adults gaze patterns are consistent. Although infant's visual acuity is poor, their gaze patterns are not random [105].

During the first 4 months, infants learn trace complex con- tours and follow moving objects to shift their gaze toward the target of interest [131]. Such systematic patterns are developed as a result of neural growth in the structure of retina and cortical areas. Between age of 4 and 6 months, infants develop more complex visual attention mechanism. This mechanism exploits suppression of competing information during attention-oriented shifts [132]. Infants at 9 months suppress previously cued locations in the scene after they are visited [133,134]. Previous studies demonstrate that infants gaze patterns are more easily learned than adults gaze patterns [105,135].

In this chapter, a dataset of eye tracking recorded infants and adult's fixation is exploited. The ground truth fixations from infants and adults are compared to eight benchmark saliency models and two baselines using eight standard evaluation metrics. Throughout this chapter, first, a brief review of saliency comparisons is presented. Then, dataset, models, and metrics used in this chapter are explained. And then, extensive comparisons of infants and adults are demonstrated. Finally, findings are summarized and pointed out.

4.1.1 Contributions:

Three contributions are presented in this chapter. First, demonstrate how well saliency models can predict infants and adults eye fixations. Second, present the ranking order of saliency models over infants and adults. Third, highlight the differences between infants and adults gaze patterns.

4.2 Methods and Materials:

4.2.1 Computational Saliency Models:

In this chapter, eight selected bottom up saliency models and two baseline models are compared using experimental fixations dataset of infants and adults. All selected saliency models have been widely used and frequently cited in the literature. The eight selected saliency models are briefly described as follows:

1. **Itti model** [24] first extracts three visual features: color, intensity, and orientation. It then applies spatial competition via center surround operation to create conspicuous maps corresponding to the feature dimensions. The conspicuous maps are then linearly combined with equal weights into a single saliency map. The implementation of this model used in this chapter includes a slight blur as a final step [59].
2. **Graph based visual saliency model (GBVS)** [58] is a graph implementation of the Itti model. The model uses a Markov chain as an activation map and incorporates a center prior.

3. **HouNips model** [136] trains (8 × 8 pixels) RGB image patches and learns 192 feature functions. Then uses code length increment as a change of entropy with respect to feature activity probability increment.
4. **HouCVPR model** [137] processes the image in frequency domain where the difference between the logarithm of magnitude and the logarithm of blurred version of the magnitude is a residual spectral.
5. **CBS model** [138] extracts three features: super-pixel color, closed shapes, and center bias. Then detects salient regions using contour energy computation.
6. **SUN model** [29] uses a Bayesian framework to detect saliency as self-information in local image patches. The model uses difference of Gaussians (DoG) and independent component analysis (ICA) as visual features.
7. **AIM model** [26] learns a dictionary of image patches using ICA as visual features then uses self-information on local image patches to produce a saliency map.
8. **AWS model** [139] uses luminance and color to create local energy and color maps. Then generates multiple scales of the feature maps and uses principle component analysis (PCA) to de-correlate the multi-scale information of each feature map.

Figure 9 shows six representative input images and the corresponding ground-truth fixation maps for infants and adults and saliency maps obtained by eight selected saliency models. The Itti, GBVS, and AWS models produce similar results, because these three models use same features (intensity, color, and orientation). Similarly, SUN and AIM models produce similar results because both models use ICA as image features and self-information as a saliency construction operation.

In addition to infants and adults' comparisons using the saliency models, comparisons with two baseline models including chance and center are also conducted. A chance baseline model selects pixels randomly as salient locations. A center baseline model is a 2D Gaussian shape in which the center is counted as the most salient, and the salient values decrease as the distance increases from the image center [110].

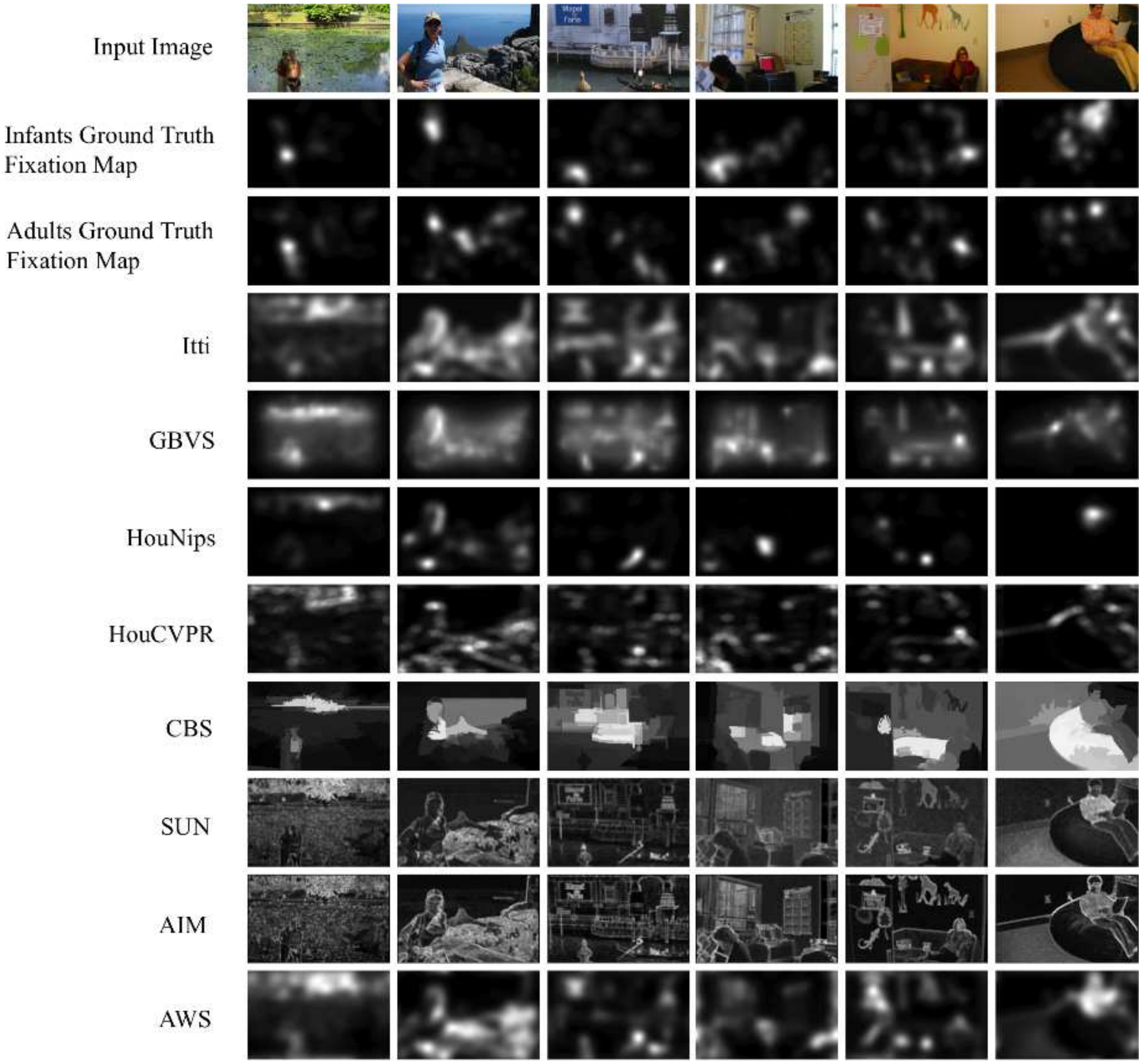

Input Image
Infants Ground Truth Fixation Map
Adults Ground Truth Fixation Map
Itti
GBVS
HouNips
HouCVPR
CBS
SUN
AIM
AWS

Figure 9 - Row 1 presents the photographs of six representative input images. The corresponding ground-truth fixation maps of infants and adults are shown in row 2 and 3, respectively. Saliency maps obtained by 8 saliency models are shown in row 4 through 11.

4.2.2 Stimuli:

Sixteen color images were used as the stimuli for collecting infants and adults eye movements. The images are 8 indoor scenes and 8 outdoor scenes. Human is presented in all images. In some image's human is presented in the foreground, while in some other image's human is presented in the background. The size of each image is 1050×1680 pixels.

4.2.3 Protocol Experiments:

In this chapter, a dataset of 16 images and recorded eye tracking data from 20 participants (10 infants and 10 adults) are used. All human data is provided by a research group at Brown University and the experimental protocol was approved by the Brown University Institutional Review Board. The detailed description of the experiments can be found in previous work [140,141].

The participants were 10 infants (mean age = 9.5 months) and 10 adults (mean age = 19 years). All participants sat at a distance of approximately 70 cm from a 22-inch (55.9 cm) computer. Infants sat at parents' lap. A remote eye tracker (SMI SensoMotoric Instruments RED system) was used to record participants' gaze path as they freely viewed each image. A digital video camera (Canon ZR960) was placed above the computer screen to record head movements. All calibrations and task stimuli in this chapter were presented using the experimental center software provided from SMI. Before starting the task, an attractive looming stimulus was presented in the upper left and lower right corners of the screen to calibrate the point of gaze (POG). The same calibration stimulus was then presented in all four corners of the screen to validate the accuracy of calibration. Images span the entire screen in a random order for 5 seconds. A central fixation target was used to return participants' POG to the center of the screen between images.

Figure 10 shows representative indoor and outdoor images with fixations distributions for infants (red circles) and adults (blue circles). In general, both infants and adults demonstrate high fixation density on human objective presented in images. Also, adult fixations show a larger distribution spread than infant fixations.

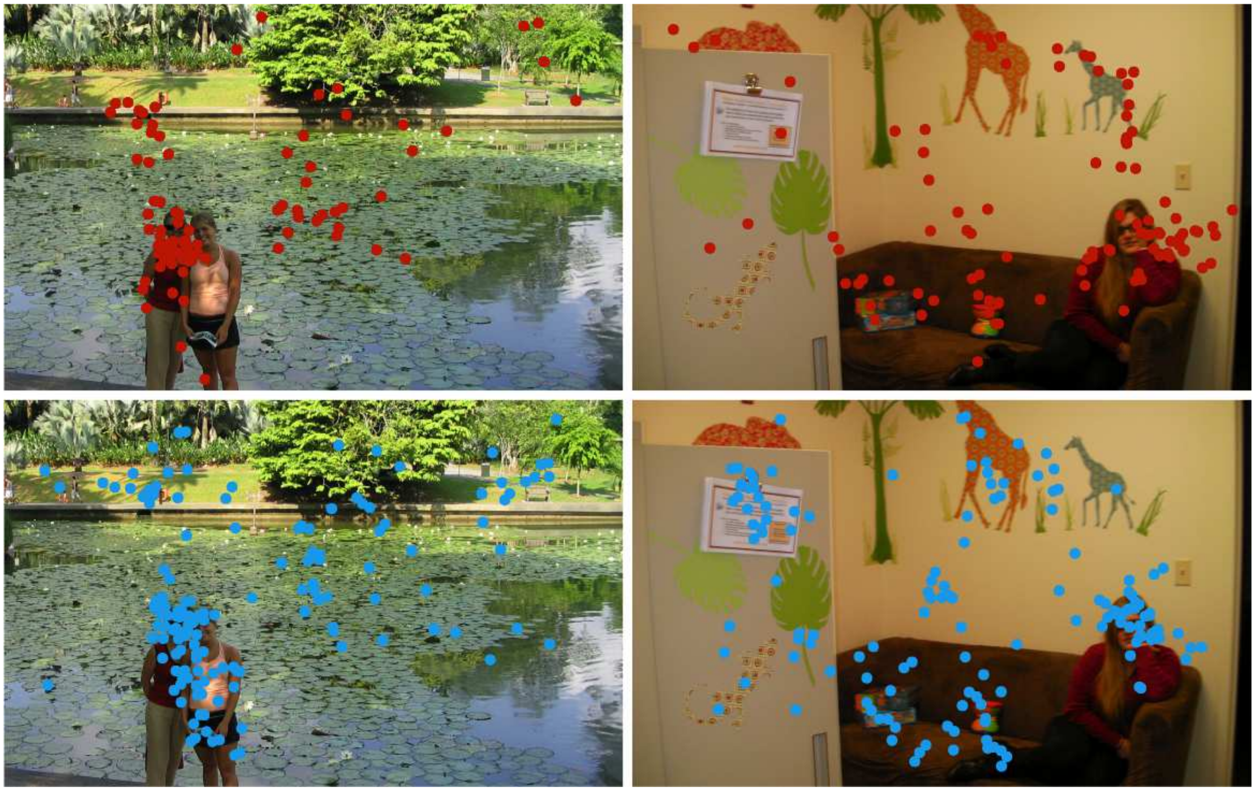


Figure 10 - Two representative images of gaze patterns of infants (top images) and adults (bottom images) over an indoor and outdoor scene. Red and blue circles highlight the fixation locations for infants (red) and adults (blue).

In order to evaluate a saliency map, the recorded eye fixations are post-processed and formatted to be ready to use. A ground-truth fixation map is obtained by convolving the binary map (one for fixation exact location and zero elsewhere) with a Gaussian function. The standard deviation of the Gaussian function is equivalent to 1° of visual angle. One degree of visual angle represents an estimation of the fovea [140].

4.2.4 Evaluation Metrics:

Performance of a saliency model is often compared to human fixations map using evaluation metrics to describe the agreement between a saliency map and human fixations map. In this chapter, seven metrics are used for evaluating the performance of selected saliency models. The motivation for analyzing saliency models with seven metrics is to ensure that the summarized conclusions are independent of the choice of metric and consistent across all metrics. Generally, a good saliency model should perform well across all metrics.

The two binary classification measures are based on the intersection of the area between predicted saliency and human fixations, including receiver operating characteristics (ROC) and precision-recall (PR). From the ROC measure, the area under ROC curve (AUC) is reported as the first evaluation metric. Also, F-measure (metric) score is obtained from PR. Moreover, three metrics measure the similarity and two metrics measure the dissimilarity between a saliency map and a ground-truth fixation map are also used in this chapter [113]. The similarity-based metrics are: information gain (IG), similarity (SIM), and Pearson's correlation coefficient (CC). The dissimilarity-based metrics are: Kullback-Leibler divergence (KL), and earth movers' distance (EMD).

4.3 Results and Discussion:

In this section, a comparison of eight saliency and two baseline models for prediction of fixations between infants and adults is presented. Then, saliency models are compared over infants and adults, separately.

4.3.1 Analysis over infants and adults:

Figure 11 presents the average receiver operating characteristics (ROC) curve, and precision recall (PR) curve of eight saliency and two baseline models over the dataset used in this chapter, for infants and adults, respectively. The ROC curves of the saliency models over infant and adult fixations are comparable. On the other hand, the PR curves of saliency models over adult fixations outperform the PR curves of the saliency models over infant fixations.

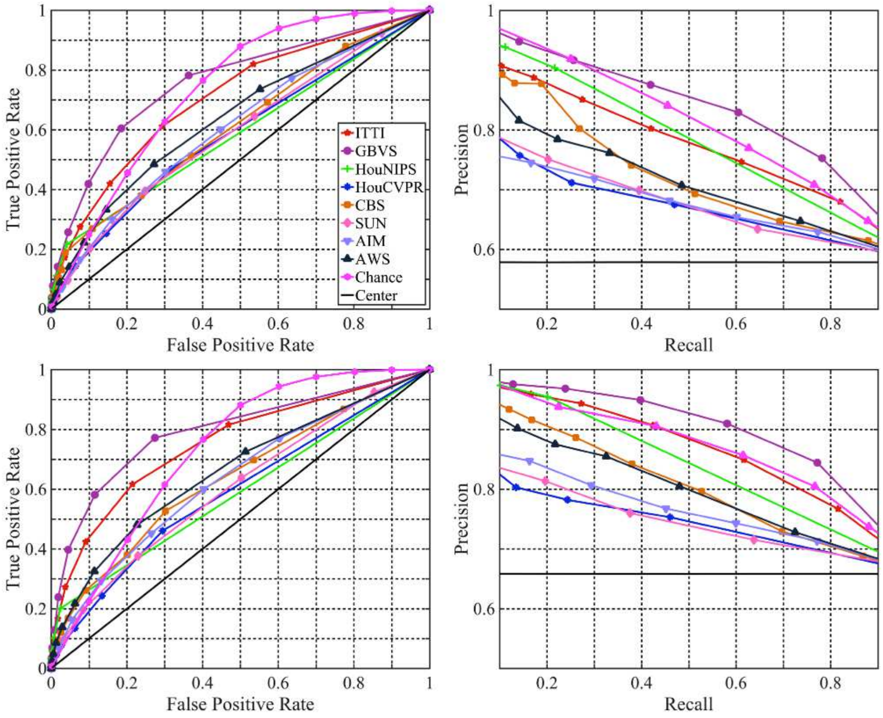


Figure 11 - Averaged ROC and PR curves of eight saliency models and two baseline models over infants (top charts) and adults (bottom charts).

To summarize the performance of saliency model's fixation prediction over the infant and adult fixations, figure 12 presents the AUC score and F-measure over the infants and adults' data. In figure 12, a comparison is conducted between infant and adult ground-truth fixation maps over all eight saliency models and two baselines. The AUC score indicates that there is no significant difference between infants and adults for all eight saliency and two baseline models. Comparatively, the F-measure (figure 12 right) over adult fixations is significantly larger than the F-measure over the infant fixations for all eight models except the HouNips model. This indicates that the overlap between the predicted and retrieved fixations for adults is larger than infants. In addition, for both baseline models, the F-measure for adults is significantly larger than that for infants.

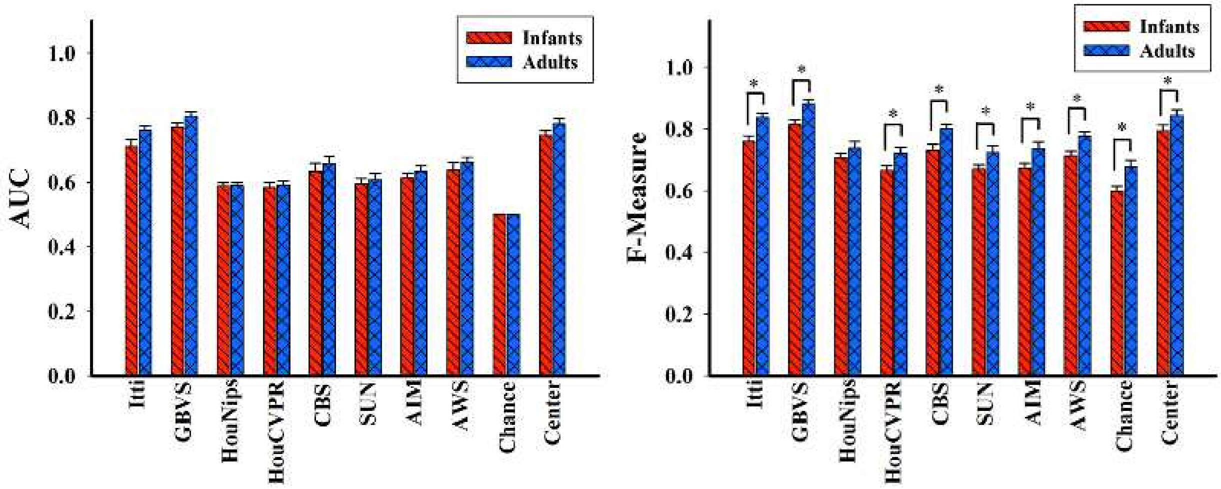


Figure 12 - Averaged AUC score and F-measure for infants and adults. A * indicates statistical significance using t-test (95%, p ≤ 0.05). Error bars indicate standard error of the mean (SEM).

Figure 13 presents the average score of information gain (IG), similarity (SIM), and correlation coefficient (CC) for infants and adults over all saliency and baseline models. As shown in figure 13 left, adults have significantly larger IG scores than infants over all saliency models except the HouNips model. Although a center bias map outperforms all saliency and baseline models for infants and adults, adults are significantly fit to their center bias maps better than infants. This is because that, the distributions predicted by saliency models are more comparable with the distribution of fixations in adults than infants.

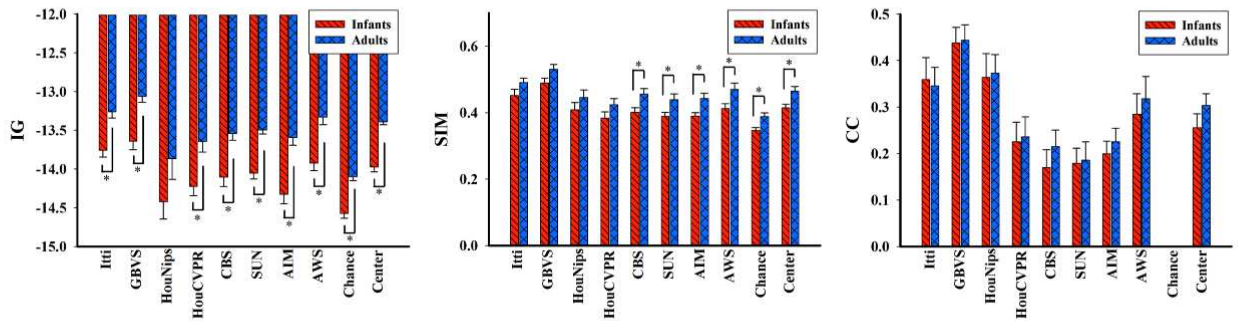


Figure 13 - Averaged IG, SIM, and CC scores for infants and adults. A * indicates statistical significance using t-test (95%, p ≤ 0.05). Error bars indicate SEM.

Furthermore, the SIM score (figure 13 middle) over adult ground-truth fixation maps is significantly larger than the SIM score over infant ground-truth fixation maps for CBS, SUN, AIM, AWS, models and both baseline models. It indicates that saliency maps are intersected with adult ground-truth fixation maps more than infants. This occurs because the difference between saliency map and fixation map at each pixel are smaller in adults than in infants.

As shown in figure 13 right, infants and adults are not significantly different in terms of CC score. Both infants and adults have positive correlation with all eight and center baseline models. Although the maps obtained from the saliency and center baseline models are not identical to the infant or adult fixation maps, the pattern of salient values in the saliency and center baseline maps change in the same direction for the corresponding values in infant or adult ground-truth fixation maps. Interestingly, both infants and adults have a score close to zero in the chance baseline model. This occurs because values of the chance baseline model change randomly, while values of the fixation maps for infants and adults change in a specific pattern. Therefore, the chance baseline model does not follow the direction of values changing in the fixation maps for infants and adults.

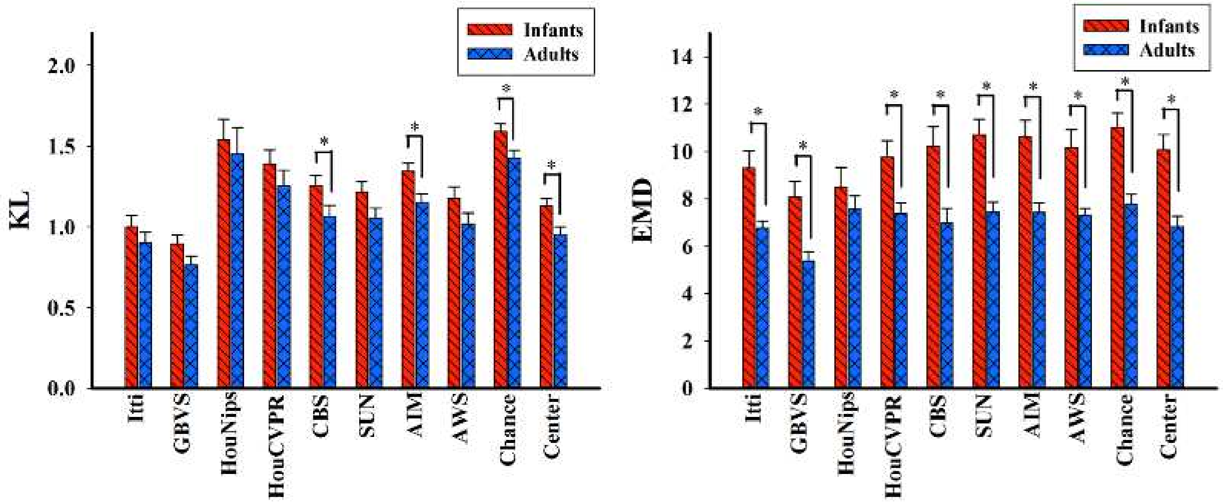


Figure 14 - Averaged KL and EMD scores for infants and adults. A * indicates statistical significance using t-test (95%, p ≤ 0.05). Error bars indicate SEM.

Two dissimilarity measures are presented in figure 14. In the left chart of figure 14, the KL scores of adults are significantly lower than that of infants in CBS, AIM, and two baseline models. This observation indicates that saliency models lose significantly less information in approximating adults than infants. In the right chart of figure 14, the EMD scores of adults is significantly lower than the corresponding values of infants for all saliency and baseline models except the HouNips model. It proves that the spatial locations in the saliency maps are significantly closer to adults' fixation locations than infant' fixation locations.

Overall, the performance of infants and adults is consistent across all seven evaluation metrics regardless of the significant difference. Adults' scores are larger than infants' scores over all similarity-based metrics. Consistently, adults' scores are smaller than infants' scores over all dissimilarity-based metrics. Such consistency of larger scores for adults than for infants indicate that adults eye falls on more salient locations than infants. It also indicates that adult distribution of fixations is more spread than infant's distribution of fixations.

4.3.2 Analysis over infants:

Table 4 presents the ranking of saliency models for infant fixation prediction over the image dataset. Although the ranking of models differs based on different metrics, some general patterns can be observed. Using the AUC score, the GBVS model has the highest score, and the center baseline and Itti models are among the top three ranking. High AUC score for the center bias indicates a high density of infant fixations near an image center. This is due to observer viewing strategies and photographic bias [142-144]. Observers tend to look near the center of the image. One explanation could be that photographers center the object of interest while capturing image. Similarly, using F-measure, GBVS scores the highest and center baseline and Itti rank second and third, respectively. High performance of the center baseline model indicates high center preference over the dataset. For the IG score, GBVS, Itti, and AWS ranked first, second, and third, respectively. This indicates that the three models are more fit to the center bias emerged from infant fixations than the center baseline model. For the SIM score, GBVS ranked first, Itti ranks second, and center baseline model ranks third. The top three models have a larger overlap with the infant ground- truth fixation map. The center baseline performs closely with AWS and HouNips models. For the CC score, GBVS scores the highest, HouNips scores second, and Itti scores third. This proves that the saliency maps obtained by these three models have a stronger positive correlation with infant ground-truth fixation maps. Also, using KL score, GBVS, Itti, and center baseline are ranked first, second, and third, respectively. It indicates a more adequate approximation of the ground-truth fixation map by the top three ranking models. Finally, for the EMD score, GBVS, HouNips, and Itti are ranked top three. The three top ranking models are less different spatially with the infant ground-truth fixation maps than the center baseline model.

Table 4 - Ranking of eight saliency and two baseline models over infants using seven evaluation metrics. Top three models are highlighted red, green, and blue, respectively.

	AUC	F-measure	IG	SIM	CC	KL	EMD
Itti	0.71 ± 0.02	0.76 ± 0.02	-13.76 ± 0.09	0.45 ± 0.02	0.36 ± 0.05	1 ± 0.07	9.31 ± 0.73
GBVS	0.77 ± 0.01	0.81 ± 0.02	-13.64 ± 0.11	0.49 ± 0.01	0.44 ± 0.03	0.9 ± 0.05	8.10 ± 0.66
HouNips	0.59 ± 0.01	0.71 ± 0.02	-14.42 ± 0.22	0.41 ± 0.02	0.36 ± 0.05	1.54 ± 0.13	8.48 ± 0.86
HouCVPR	0.58 ± 0.02	0.67 ± 0.02	-14.23 ± 0.12	0.38 ± 0.02	0.23 ± 0.04	1.39 ± 0.09	9.76 ± 0.69
CBS	0.63 ± 0.02	0.73 ± 0.02	-14.11 ± 0.12	0.40 ± 0.01	0.17 ± 0.039	1.3 ± 0.06	10.24 ± 0.81
SUN	0.59 ± 0.02	0.67 ± 0.01	-14.05 ± 0.08	0.39 ± 0.01	0.18 ± 0.03	1.22 ± 0.06	10.71 ± 0.64
AIM	0.61 ± 0.02	0.67 ± 0.01	-14.33 ± 0.12	0.39 ± 0.01	0.20 ± 0.03	1.35 ± 0.05	10.62 ± 0.67
AWS	0.64 ± 0.02	0.71 ± 0.01	-13.92 ± 0.10	0.41 ± 0.02	0.29 ± 0.04	1.18 ± 0.07	10.18 ± 0.75
Chance	0.50 ± 0	0.60 ± 0.02	-14.58 ± 0.06	0.35 ± 0.01	0±0	1.59 ± 0.05	11.03± 0.60
Center	0.75 ± 0.01	0.79 ± 0.02	-13.98 ± 0.06	0.41 ± 0.01	0.26 ± 0.03	1.13 ± 0.04	10.08± 0.64

In general, GBVS model ranks first across all evaluation metrics. It indicates that the GBVS model is more suitable to predict infants' fixations than any other models used in this chapter. The Itti model is among the top three ranking models across all metrics. This occurs because the Itti model is enhanced by slightly blurring the saliency map. Therefore, the Itti model increases the size of the predicted distribution. The center baseline model outperforms most models in AUC and F-measure. The reason is that, true positives fall near the center of the image as a result of infant's fixations bias. Therefore, the center baseline model achieves higher score than many other models. Another important observation is that, all models outperform the chance baseline model over all metrics. It indicates that infant gaze patterns are not random and follow a specific visual mechanism.

4.3.3 Analysis over adults:

Table 5 presents the ranking of saliency models over the image dataset for adults. For both AUC score and F-measure, GBVS, center baseline, and Itti models rank as top three. This shows that adult fixations are dense near the image center. The adult fixations are not only allocated near the center of the image, but also have higher overlap between the saliency map and fixation map. Using the IG score, the GBVS, Itti, and AWS rank as the top three models. It indicates that a center bias emerged from adult fixations is more fit to GVBS, Itti, and AWS models. For the SIM score, the top three models are GBVS, Itti, and AWS models, respectively. This means that's the saliency maps obtained by these three models are more correlated with the adult ground-truth fixation maps than the other models. Using the CC score, GBVS scores the highest, and the HouNips and Itti models are among the top three. The adult ground-truth fixation maps are more correlated with GBVS, Itti, and AWS models than the center baseline model. Using the KL score, GBVS ranks first, Itti model ranks second, and the center baseline model ranks third. It indicates that GBVS and Itti models have a higher approximation of the adult ground-truth fixation maps than the center baseline model. For the EMD score, the GBVS, center baseline, and Itti models rank as the top three. Also, as shown in table 3, the GBVS model has a lower EMD score than all other models. It indicates that distribution allocation of an adult ground-truth fixation map is more predictable by the GBVS model than other models in this chapter.

Generally, the GBVS model ranks as the first over all metrics. The GBVS model is more suitable for predicting the adult fixations than the other models in this chapter. Also, Itti model demonstrates its consistency ranking among the top three models. The good performance of the center baseline model over all metrics indicates a strong bias of adult fixations toward the center of the image. Finally, all models outperformed the chance baseline model for the prediction of adult fixations.

Table 5 - Ranking of eight saliency and two baseline models over adults using seven evaluation metrics. Top three models are highlighted red, green, and blue, respectively.

	AUC	F-measure	IG	SIM	CC	KL	EMD
Itti	0.76 ± 0.01	0.84 ± 0.01	-13.26 ± 0.08	0.49 ± 0.01	0.35 ± 0.04	0.90 ± 0.06	6.76 ± 0.31
GBVS	0.81 ± 0.01	0.88 ± 0.01	-13.063 ± 0.08	0.53 ± 0.02	0.44 ± 0.03	0.76 ± 0.05	5.40 ± 0.37
HouNips	0.59 ± 0.01	0.74 ± 0.02	-13.87 ± 0.27	0.45 ± 0.02	0.37 ± 0.04	1.45 ± 0.17	7.58 ± 0.56
HouCVPR	0.59 ± 0.01	0.72 ± 0.02	-13.64 ± 0.14	0.42 ± 0.02	0.24 ± 0.04	1.25 ± 0.10	7.38 ± 0.47
CBS	0.66 ± 0.02	0.80 ± 0.01	-13.54 ± 0.09	0.46 ± 0.02	0.21 ± 0.04	1.07 ± 0.07	7.0 ± 0.60
SUN	0.61 ± 0.02	0.72 ± 0.02	-13.50 ± 0.05	0.44 ± 0.02	0.19 ± 0.04	1.06 ± 0.06	7.46 ± 0.42
AIM	0.63 ± 0.02	0.74 ± 0.02	-13.60 ± 0.10	0.44 ± 0.02	0.23 ± 0.03	1.15 ± 0.05	7.42 ± 0.41
AWS	0.66 ± 0.02	0.78 ± 0.01	-13.33 ± 0.09	0.47 ± 0.02	0.32 ± 0.05	1.02 ± 0.07	7.31 ± 0.28
Chance	0.50 ± 0	0.68 ± 0.02	-14.10 ± 0.05	0.39 ± 0.01	0±0	1.43 ± 0.05	7.78 ± 0.43
Center	0.78 ± 0.02	0.84 ± 0.02	-13.39 ± 0.04	0.47 ± 0.01	0.3 ± 0.03	0.95 ± 0.05	6.84 ± 0.43

4.3.4 Discussions of different datasets:

The results over infants and adults demonstrate several differences between infants and adult’s visual attention. Such results were concluded with 16 images only. To justify the conclusions of the experimental results, MIT1003 dataset [30] was used to compare to the dataset of infants and adults. Because the MIT1003 images contain diverse scene context, a subset of 85 images were carefully selected to match the context of the images in the infants and adult’s dataset. The images are selected based on the following criteria: color, human presence, maximum size of human face is one fourth of the total image size, animals, and motion blur. Images that contained animals, motion blur, or human faces larger than one fourth the image were excluded to avoid a strong bias in the image. Saliency maps of the eight saliency models and two baseline models were computed on the subset of MIT1003 dataset. Then, scores of the seven evaluation metrics were obtained. Figure 15 shows the ranking of saliency models and baseline models over the infants and adults’ dataset and the subset of MIT1003 dataset. In the ranking scheme, statistical significance between consecutive models was measured using t-test at the significance level of $p \leq 0.05$. Although statistics of the two image datasets vary, some general patterns can be observed. The infants and adult's dataset and the MIT1003 dataset have similar trends. GBVS ranked first and all saliency models and the center baseline model outperformed the chance baseline model using all seven evaluation metrics over both datasets. Also, the scores of the two datasets are comparable for all evaluation metrics except the IG score. This occurs because the center bias map calculated for the MIT1003 dataset is an average map of larger number of images than the infants and adult's dataset.

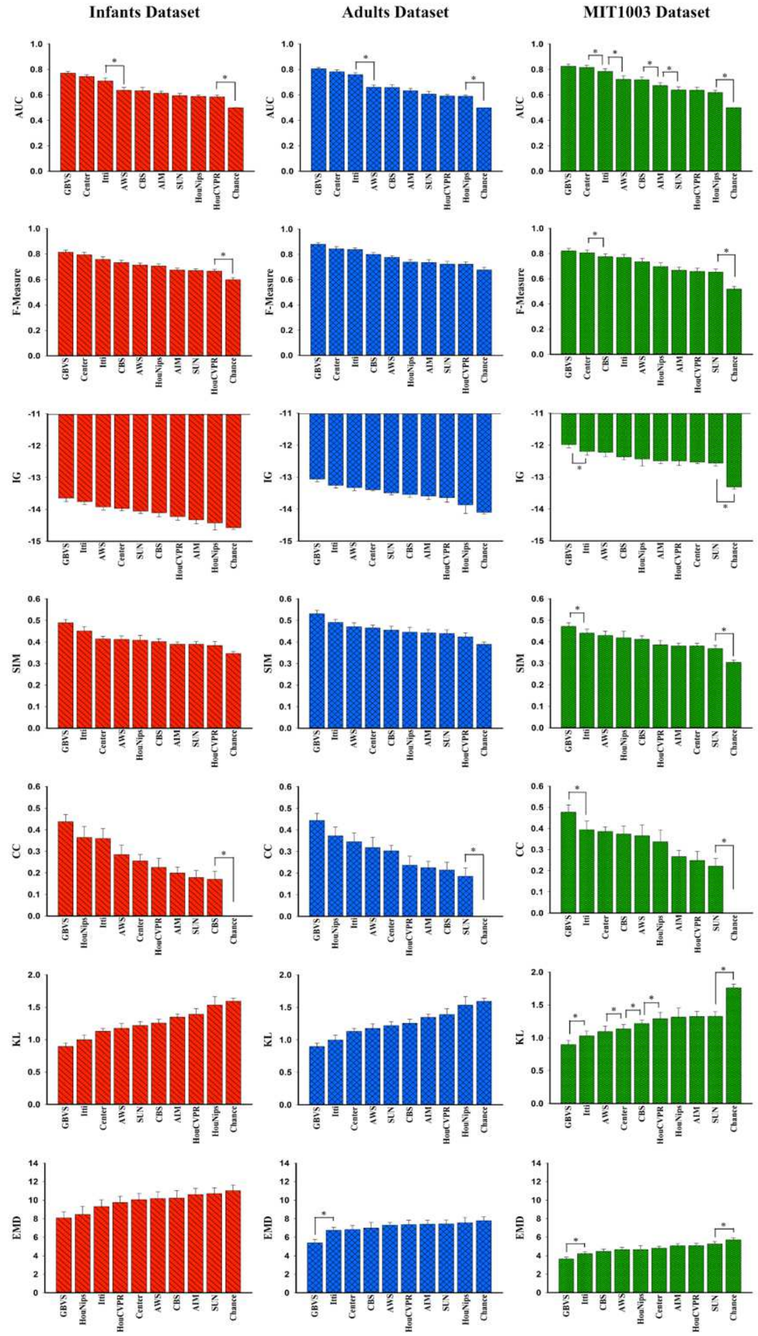


Figure 15 - Ranking visual saliency models over infants (red bar charts) and adults (blue chart bars) dataset, and a subset of 85 images (green blue charts) from the MIT1003 dataset using seven evaluation metrics: AUC, F-measure, IG, SIM, CC, KL, and EMD. A * indicates statistical significance using t-test (95%, p ≤ 0.05) between consecutive models. If no * between two models that are not consecutive, it does not indicate that they are not significantly different. In fact, models that are not consecutive have higher probability to be significantly different than consecutive models. Error bars indicate SEM.

4.4 Conclusion:

In this chapter, a dataset of images and recorded eye fixations from infants and adults is used to quantitatively analyze the difference between their gaze patterns. Eight state-of- the-art saliency and two baseline models are compared be- tween infants and adults. The ranking of eight saliency and two baseline models over both infants and adults are also provided in this dissertation. Seven standard evaluation metrics are used to evaluate the performances of all eight saliency and baseline models on prediction of fixations. The main conclusions of this comparison are: 1) Saliency models are significantly more overlapped, fit, and intersected with adult fixations than infant fixations, in terms of F-measure, IG, and SIM. 2) Saliency models have much less information loss in approximation, and spatial distance of distributions to adults than infants, in terms of KL and EMD. 3) GBVS and Itti models are among the top 3 contenders over infants and adults consistently. In other words, GBVS and Itti models are suitable for prediction of fixations for both infants and adults. 4) For the dataset used in this chapter, infant and adult fixations have bias toward the center of the image. Also, all models outperformed the chance baseline model. This demonstrates that not only adult gaze patterns are consistent, but also infant gaze patterns follow a systematic mechanism. This chapter provides a comparison of various saliency models on fixations prediction on both infants and adults. It may help the readers to understand the difference between infant and adult gaze patterns. These findings may also provide useful information on selection of saliency models for prediction of infant fixations.

CHAPTER 5

DEEPFEAT FOR VISUAL SALIENCY PREDICTION

5.1 Introduction:

The human visual system has an exceptional ability of sampling the surrounding world to pay attention to objects of interest. Such ability is the visual attention that guides the visual exploration. Visual attention requires a complex cognitive mechanism to allocate the human gaze toward the objects of interest. In computer vision, a saliency map is defined to model the human visual attention. A saliency map is a 2D topological map that indicates visual attention priorities in a numerical scale. A higher visual attention priority indicates the object of interest is irregular or rare to its surroundings. The modeling of saliency is beneficial for several applications including image segmentation [8], object detection [43], image re-targeting [9], image/video compression [12], advertising design [19], and analysis of gaze patterns [2], etc.

The research on saliency modeling is influenced by bottom- up and top-down factors. The bottom-up visual attention is triggered by stimulus, where a saliency is captured as the distinction of image locations, regions, or objects in terms of bottom-up features such as color, intensity, orientation, shape, T-conjunctions, X-conjunctions, etc. [3]. One of the bottlenecks that bottom-up saliency models suffer, is that they explain the scene partially as the majority of the human eye fixations are task driven. Following the feature integration theory (FIT) [22], the first saliency model was proposed [24]. The model exploits the biologically inspired center-surround scheme of color, intensity, and orientation at various scales to identify distinctive image locations. Bruce & Tsotsos proposed an attentional information maximization model to predict eye fixations [26]. The model uses self-information to detect saliency in local image regions. Zhang et al. derived a Bayesian framework that incorporates self-information of local image regions with prior knowledge about the image [29]. Liu t al. developed a saliency model as a decision tree of regional saliency measurements including global contrast, spatial sparsity, and object prior [104]. Zhang & Sclaroff developed a saliency map based on a Boolean approach. The model combines binary maps and attention maps [33]. The binary maps are obtained via random thresholding of the color feature of the image. Attention maps are computed using the gestalt principle of the figure-ground segregation. Leboran et al. proposed a dynamic whitening saliency model to predict fixations in videos [145]. The model uses whitening to access the relevant information by removing the second order information.

The top-down visual attention is driven by task. Top-down saliency models use prior knowledge, expectations, or rewards as high-level visual cues to identify the target of interest [52]. The recognition of an object of interest such as faces, people, and cars is an example of top-down features. Several top-down saliency models have been proposed. Such as, Oliva et al. introduced a top-down visual search model based on Bayesian framework. The model exploits cognitive features and scales [146]. Contextual features are represented by reducing dimensionality of local features. The joint probability of a feature vector is computed using multivariate Gaussian distributions. Rao proposed an attention representation as a cortical mechanism for reducing perceptual uncertainty. The model exploits belief propagation in a probabilistic framework to combine bottom-up and top-down visual factors [80]. Judd et al. developed a saliency model to predict where human look by using low, mid, and high-level cues as support vector machines [30]. Borji et al. proposed a saliency model based on top-down factors to learn task driven object based visual attention control in interacting environment [147]. Wang et al. combined 13 bottom-up and top-down saliency models using several combination strategies [148]. Then the model has been trained as a support vector machine.

Recently, deep features of the deep neural networks (DNN), have been used in several applications, including imaging and video processing, medical signal processing, large data analysis, and saliency modeling as well [94]. Although the intuition of the DNN deep features remain unclear [149], several saliency models has been trained to detect bottom-up and top-down factors. Deep features are the response images of convolution, batch normalization, activation, and pooling operations in a series of layers in a deep convolutional neural network (DCNN) [150]. Such layers encode the conspicuous information about the image. In the first layer, the network learns low level cues such as simple edges. At higher layers, the network learns higher level cues. Later layers provide higher level of abstracts such as a class of objects.

Deep learning saliency models demonstrated outstanding ability providing high accuracy prediction of human fixations. However, such models require large training times, and high cost system requirements. Several applications such as robotics requires a fast and low memory consuming saliency models. Today, robots are utilized to assist in several applications such as home service, rehabilitation, and assistant living [151], [152]. To overcome such issue, we introduce a fixed framework that uses data-driven features of DCNNs pre-trained for object classification to computes a bottom-up and top-down attention maps and combine them in a saliency map.

5.1.1 Contributions:

In this chapter, the contributions are threefold. First, a computational saliency model is proposed to predict human fixations using pre-trained deep features, codenamed DeepFeat. To our knowledge this is the only saliency model that combines deep features of pre-trained DCNNs without learning any parameters. Second, three implementations of the DeepFeat are computed and compared to investigate the role of the pre- trained deep features of three DCNNs in saliency prediction. Third, through extensive evaluation over four evaluation metrics and 9 saliency models, it is demonstrated that the DeepFeat model achieves a satisfactory performance.

5.2 Proposed Approach:

5.2.1 Visualization of deep features:

In this chapter, three popular deep convolutional neural networks are exploited to obtain the deep features. The three networks are: VGG [122], GoogLeNet [123], and ResNet [124]. All three DCNNs are pre-trained for object classification using the ImageNet dataset that consist of 1.28 million images of 1000 classes of objects to classify [153].

The VGG consist of 16 sequentially stacked convolution layers followed by rectified linear units (ReLU) nonlinearities. A max pooling is computed after every two layers in the first four layers, and after every three layers in the rest of the network.

The GoogLeNet consist of 22 convolution layers if the mid- layers in the inception module are ignored. The main novelty of GoogLeNet is the inception module which combines multiple scales of the convolution layers.

The ResNet used in this chapter consist of 50 convolution layers. The main feature of the network is that it combines the stack of convolution layers with their residual after every 3 convolutions.

Visualization of the architecture of the three DCNNs can be found online(http://www.vlfeat.org/matconvnet/models/). In this chapter, convolution response images are exploited as deep features for the bottom-up computation. In addition to that, as the fully connected layers and the last convolution layer of the VGG have a dimension mismatch, another VGG variant is used to implement the top-down saliency map [154]. All computations were done in MatConvNet [155].

5.2.2 DeepFeat Architecture:

In this chapter, the DeepFeat saliency model is formalized as a fusion of a bottom and top down saliency map using a simple combination strategy. Figure 16 shows the architecture of the saliency model presented in this chapter.

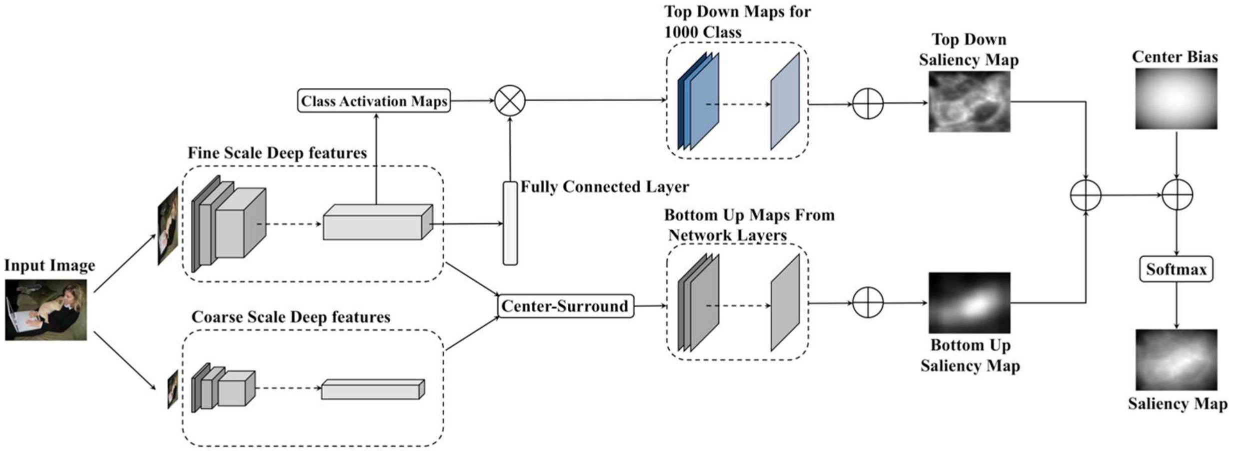


Figure 16 - Architecture of the saliency model used in this chapter.

To compute the bottom up saliency map, the fully connected layers are neglected. Let $F = \{F1, F2,\cdots, FL\}$ denotes the remaining layers of deep features treated as bottom up visual cues. The use of two scales of deep features reveals semantic cues about the image. Given $Fi = \{f1, f2,\cdot, fk\} \in R^{m\times n\times k}$, two scales of the deep features are exploited. Let $X: Fi \rightarrow R^{k}$ denotes a fine scale of the deep features, where Xjcorresponds to a response image fj. Let $Y: Fi \rightarrow R^{k}$denotes a coarse scale deep features, where Yjcorresponds to a down sampled response image fjfollowed by up-sampling. The two scales are computed using a dyadic gaussian pyramid. The pyramid consists of two operations: reduce and expand. The reduce operation is a down-sampling performed by suppressing every other row and column followed by a smoothing operation. The smoothing filter is formalized by:

$$w(r, c) = w(r)w(c) \tag{20}$$

Where

$$w(r) = w(c) = \left\{\frac{1}{4} - \frac{a}{2}, \frac{1}{4}, a, \frac{1}{4}, \frac{1}{4} - \frac{a}{2}\right\} \tag{21}$$

Where a is a constant usually between 0.3 and 0.6. In this chapter, a = 0.3. Let I be an input image, then the first down-sampling scale is denoted by:

$$G_0(x, y) = I \tag{22}$$

The following down-sampling levels are formalized by:

$$G_{i+1} = REDUCE(G_i(x, y)) \tag{23}$$

The expand operation up-samples a given feature by:

$$G_{i+1}(x, y) = 4 \sum_{m=-2}^{2} \sum_{n=-2}^{2} w(m, n) G_i(\frac{x-m}{2}, \frac{y-n}{2}) \tag{24}$$

The relationship between the two scales is formalized as a center-surround operation:

$$R^{\ell} = \sum_j \left| X^{\ell} - Y^{\ell} \right| \tag{25}$$

The total response R of layer ℓ is normalized from 0 to 1 and linearly combined with the total response images from other layers of the network to contribute equally to the computation of the bottom up saliency map:

$$M_{BU} = \sum_{\ell 1}^{L} \mathcal{N}\left(R^{\ell}\right) \tag{26}$$

The ℓ denotes the number of layers contributing to the computation of the bottom up saliency map, and $\mathcal{N}(\cdot)$ is a normalization operator.

To compute the top down saliency map, the fully connected layers are utilized to emphasize the image classification as a top down component. Let $A = \{A_1, A_2, \cdot \cdot \cdot, A_k\} \in R^{m \times n \times k}$denotes a tensor of deep features from the last activation layer (convolution, rectified linear unit, pooling, etc.). Let $W = \{W_1, W_2, \cdot \cdot \cdot, W_k\} \in R^{1 \times 1 \times k \times C}$denotes weights of the classification classes, where $W_i = [w_1, w_2, \cdots, w_c]^T$ denotes a vector of weights for unit $i \in k$. The class activation map (CAM) [155] for a class c is formalized as:

$$M_c = \sum_i W_c^T A \tag{27}$$

The CAM of a class reflects an object localization of a class or classes with the largest probability score in the fully connected layer. All object classes of an image identified by the network are localized and presented as a top down saliency map. Let $p = [p_1, p_2, \cdots, p_C]^T$ denotes the softmax of the fully connected layer. The top down saliency map is formed by:

$$M_{TD} = \sum_c P_c^T M_c \tag{28}$$

Top down factors explain majority of a scene, while bottom up factors explain the scene partially [156]. Therefore, a parameterized linear combination is defined between the top down and bottom up saliency maps:

$$\widehat{M} = (1 - \alpha) M_{BU} + \alpha M_{TD} \tag{29}$$

where α denotes a constant equal to 0.5 in this chapter. A Gaussian map is computed to reflect the bias of human eyes toward the center of the image [157], [143]. The cut off frequency of the Gaussian kernel is the maximum dimension of the image. The incorporation of the Gaussian map is formalized by:

$$\widehat{M}_{center} = \beta \widehat{M} + (1 - \beta) g \tag{30}$$

Where β is a constant equal to 0.5. A gaussian center bias map g is formalized by:

$$g(x, y) = \gamma \times exp\left(-\left(\frac{(x-x_0)^2 + (y-y_0)^2}{2\sigma^2}\right)\right) \tag{31}$$

Theγdenotes a constant equal to 1, x_0 and y_0 corresponds the center of the image. Moreover, σ is the cut-off frequency equivalent to the maximum dimension of the image. The final saliency probabilistic distribution map is formalized by:

$$S = \frac{e^{\widehat{M}_{center}}}{\sum e^{\widehat{M}_{center}}} \tag{32}$$

5.2.3 Experimental Setup:

1. **Dataset:** In this chapter, two popular datasets are explored to validate the performance of the DeepFeat saliency model. The datasets are: MIT1003 and VIU. Both datasets fall under free-viewing conditions.
2. **Evaluation Metrics:** Saliency models are usually evaluated by comparing their predictions to human fixation maps using evaluation metrics. In this work, predictions of the pro- posed framework are evaluated using four evaluation metrics including AUC, NSS, CC, and KL. In general, the AUC is a standard evaluation metric. However, it suffers multiple flaws which requires the AUC judgement to be supplemented by other evaluation metrics [163]. The AUC and NSS scores evaluate the saliency predictions over the exact fixation points (binary fixation maps). Regardless of the fixation point location, the AUC score evaluates the ranking of the saliency values at the fixation points, while NSS evaluates the saliency value at the fixation points. In addition, the CC and KL are fixation distribution-based metrics where the empirical saliency map is computed by convolving a Gaussian kernel over the map of fixation points. The cut-off frequency of the Gaussian kernel is equivalent to one degree of visual angle.
3. **Saliency Models:** To evaluate the performance of the DeepFeat model, three variants of the model are compared to nine saliency models including deep learning and conventional saliency models. The performance of the models is evaluated over the MIT1003 dataset only as the authors provides pre- computed saliency maps over the MIT1003 dataset. However, due to difficulty of compiling some of these saliency

models, the comparison of the saliency models performance over the VIU dataset is not provided. Table 6 provides a description of the saliency models used in this chapter.

Table 6 - Compared saliency models.

Model Name	Features	Category	Year	Pub.	Ref.
BMS	BU	CS	2013	ICCV	[33]
COV	BU	CS	2013	JOV	[103]
DVA	DF	DL	2018	TIP	[159]
eDN	DF	DL	2014	CVPR	[96]
iSEEL	BU	DL	2017	Neurocomp.	[160]
MLnet	DF	DL	2016	ICCV	[105]
RARE	BU	CS	2013	SPIC	[161]
SAM	DF	DL	2018	TIP	[95]
UHF	BU	ML	2016	ACCV	[162]

5.3 Results and Discussion:

5.3.1 Analysis of the Architecture:

Figure 17 presents the predicted saliency maps of three implementations of the DeepFeat model including DeepFeat bottom- up (BU) saliency map, DeepFeat top-down (TD) saliency map, and the combined bottom-up and top-down (BT) saliency map.

The three saliency implementations are computed using deep features of VGG, GoogLeNet, and ResNet implementations. For visualization of the saliency maps, the histogram of the predicted saliency maps is matched to the average histogram of the empirical saliency maps of the corresponding dataset. This technique is applied to the other visualization figures in this paper.

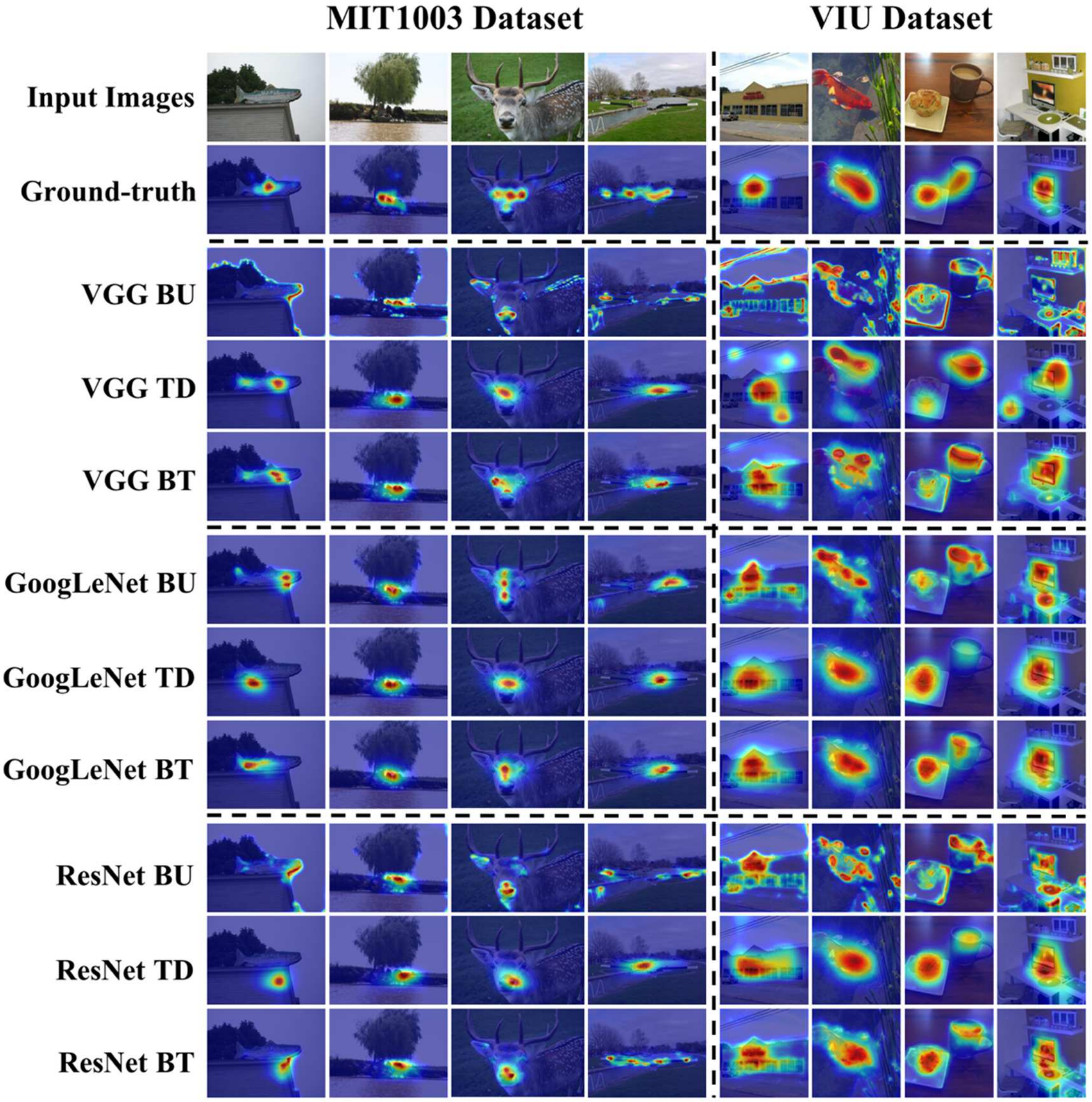


Figure 17 - Row 1 show photographs of input images from the MIT1003 and VIU datasets. Row 2 show the corresponding empirical saliency maps. Row 3 to 11 show three predicted saliency maps GoogLeNet, and ResNet.

In figure 17, two consistent trends over all three model variations can be observed. The bottom-up saliency maps predict salient contours, while the top-down saliency maps predict localized objects. Moreover, the DeepFeat implementations demonstrate that GoogLeNet computes smoother saliency maps than VGG and ResNet, and ResNet provides smoother saliency maps than VGG. This occurs because GoogLeNet merges deep features of different levels of blur. Similarly, the ResNet merges the residual deep features and the feed-forward blocks of deep features, while VGG combines feed-forward deep features only.

To quantitatively analyze the saliency implementations over the deep features of the three DCNNs, four metrics, AUC, NSS, CC, and KL, were used for evaluations over MIT1003 and VIU datasets. Figure 18 shows scores of four metrics for three implementations of the proposed DeepFeat model using deep features of VGG, GoogLeNet, and ResNet with and without center bias. To measure the statistical significance, a t-test is used at the significance rate of $p \leq 0.05$. In figure 18 without center bias, the combination of bottom- up and top-down (BT), top-down (TD), and bottom-up (BU) implementations are ranked first, second, and third, respectively. Such results are consistent over all three DCNNs and four metrics in both datasets. It indicates that the prediction of human fixation is more accurate when both bottom-up and top-down factors are assembled into the DeepFeat model. Moreover, it also can be found that the center bias significantly boosts the performance of the BU implementations more than the TD and BT implementations. This occurs because the BU implementation of the DeepFeat model computes the global contrast in terms of the deep features without any preference toward the center of the image. By adding a center bias to the bottom-up saliency map, salient regions toward the center receives more credit than those at the edges. The top-down implementation of the DeepFeat model detects objects of interest, which usually falls around the center of the image due to photography strategies [158].

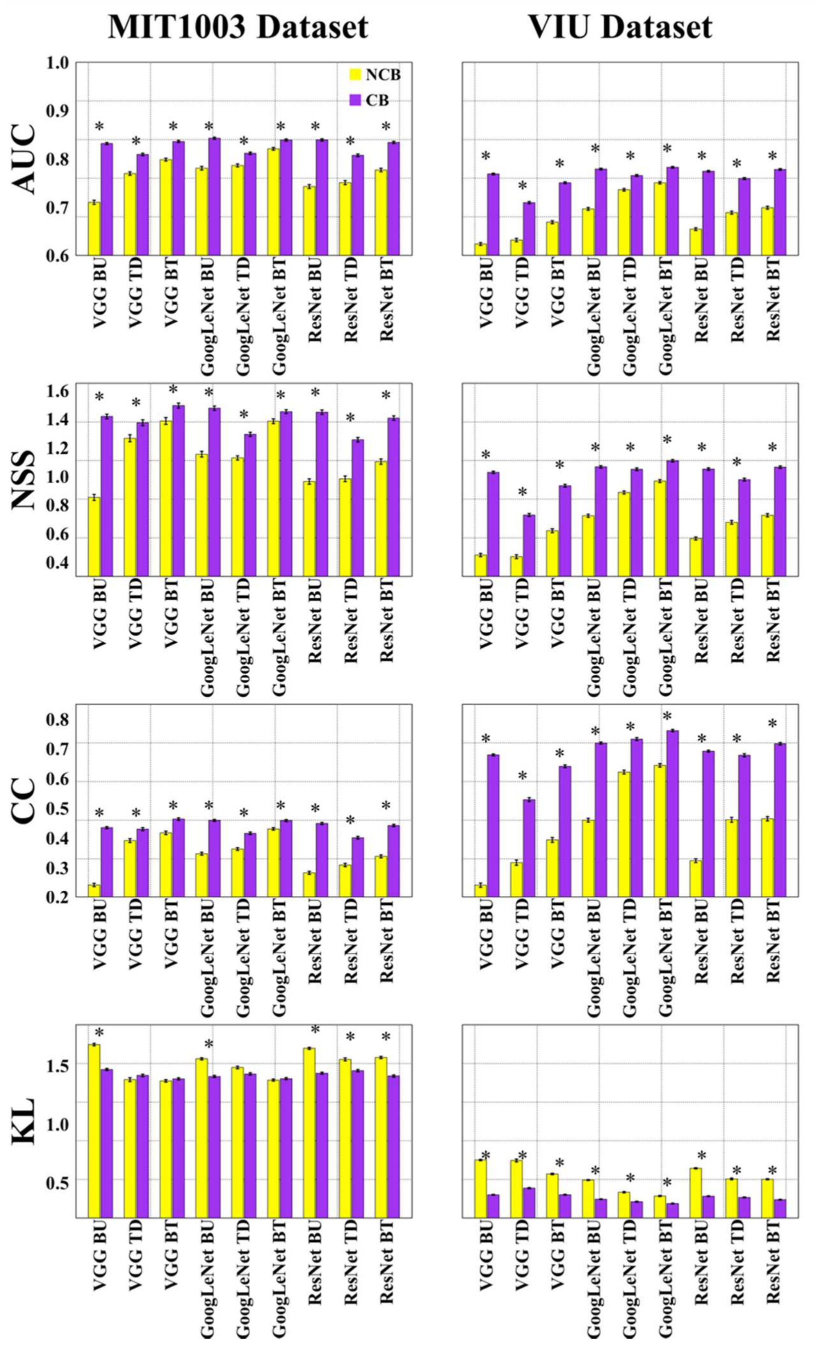


Figure 18 - Averaged scores of three implementations (BU, TD and BT) of the proposed DeepFeat model using deep features of VGG, GoogLeNet, and ResNet with and without center bias. The analysis of score are presented using four evaluation metrics: AUC, NSS, CC, and KL over the MIT1003 and VIU datasets. A * indicates the two comparing models are significantly different using t-test at confidence level of $p \leq 0.05$. Standard error of the mean (SEM) is indicated by the error bars.

In general, all three implementations with and without center bias achieve a certain agreement with the human annotations over all four metrics in both datasets. It indicates that the deep features of all three DCNNs are rich with semantic information that can be useful to predict human fixation.

5.3.2 Comparison with Other State-of-the-Art Saliency Models:

In this section, the performance of the BT implementation of DeepFeat model to a variety of saliency models is evaluated. In this chapter, the VGG, GoogLeNet, and ResNet variants of the DeepFeat model are denoted as VGG, GoogLeNet, and ResNet, respectively. These three variants are compared to nine saliency models including BMS, COV, DVA, eDN, iSEEL, MLnet, RARE, SAM, and UHF. The description of the models can be found in table 4.

Figure 19 shows 10 representative images from the MIT1003 dataset along with the corresponding empirical saliency maps and predicted saliency maps from the DeepFeat models and other nine saliency models. Fig. 20 shows the AUC, NSS, CC, and KL scores of twelve saliency models (three DeepFeat models and nine other models) over the MIT1003 dataset. Although the models ranking order is not identical over the four scores, some general patterns can be observed. Over the AUC score, all three DeepFeat models (GoogLeNet, VGG, and ResNet) are ranked in the top group together with the eDN, SAM, and iSEEL models. They are significantly higher than the other six models. In the NSS and CC scores, four deep learning-based models (SAM, DVA, MLnet, iSEEL) outperformed all other eight saliency models. The VGG, GoogLeNet, and ResNet are ranked fifth, sixth, and seventh ranking including three DeepFeat models. For the KL score, SAM, DVA, and MLnet are the top three ranking models. The VGG, GoogLeNet, and ResNet are ranked fifth, sixth, and eighth.

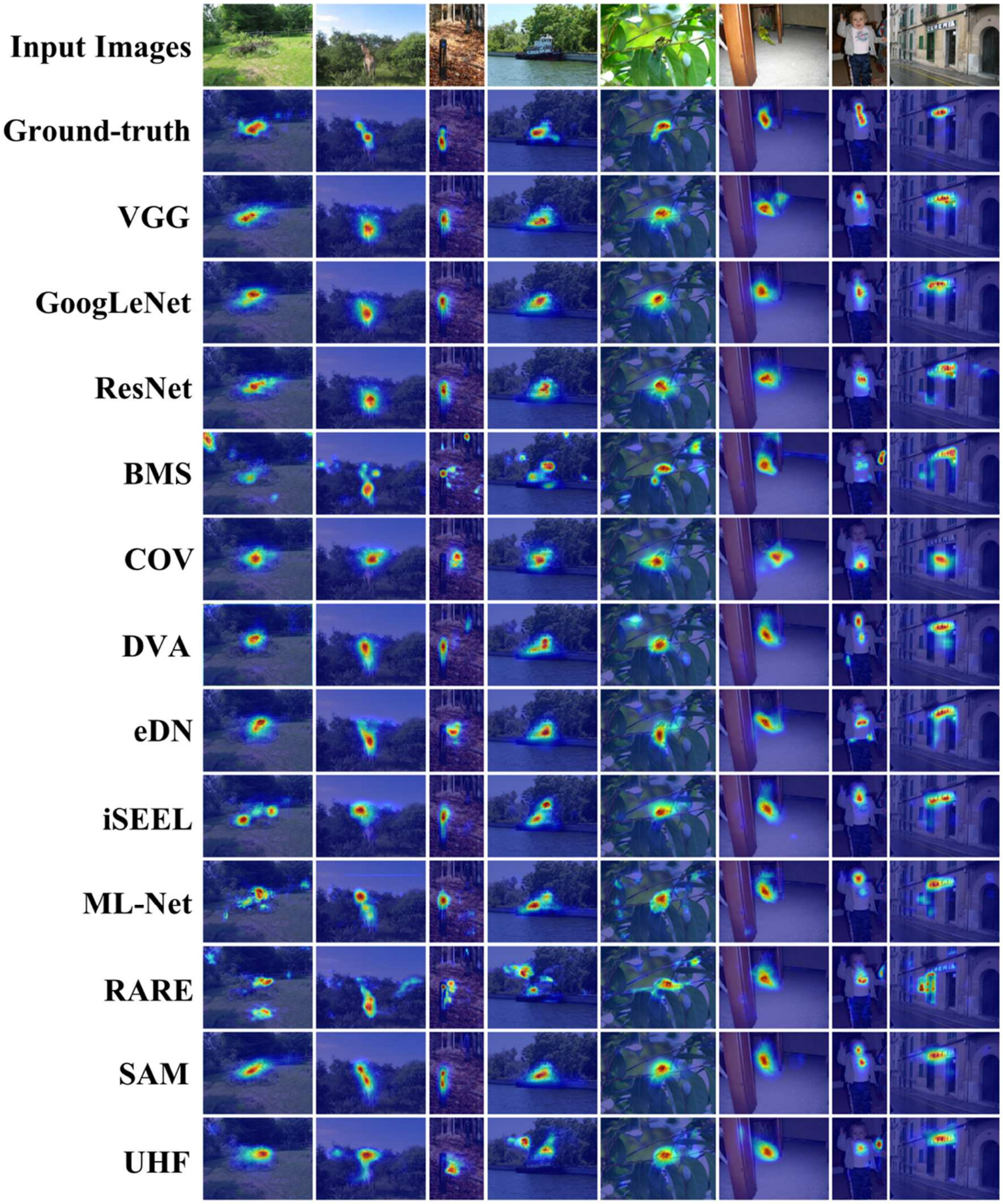


Figure 19 - Row 1 show the photographs of ten input images in MIT1003 dataset. Row 2 show the corresponding empirical saliency maps. The predicted saliency maps computed by three variants of the proposed DeepFeat model (VGG, GoogLeNet, and ResNet) are shown in row 3 to 5. Rows 6 to 15 present saliency maps computed by 9 other saliency models.

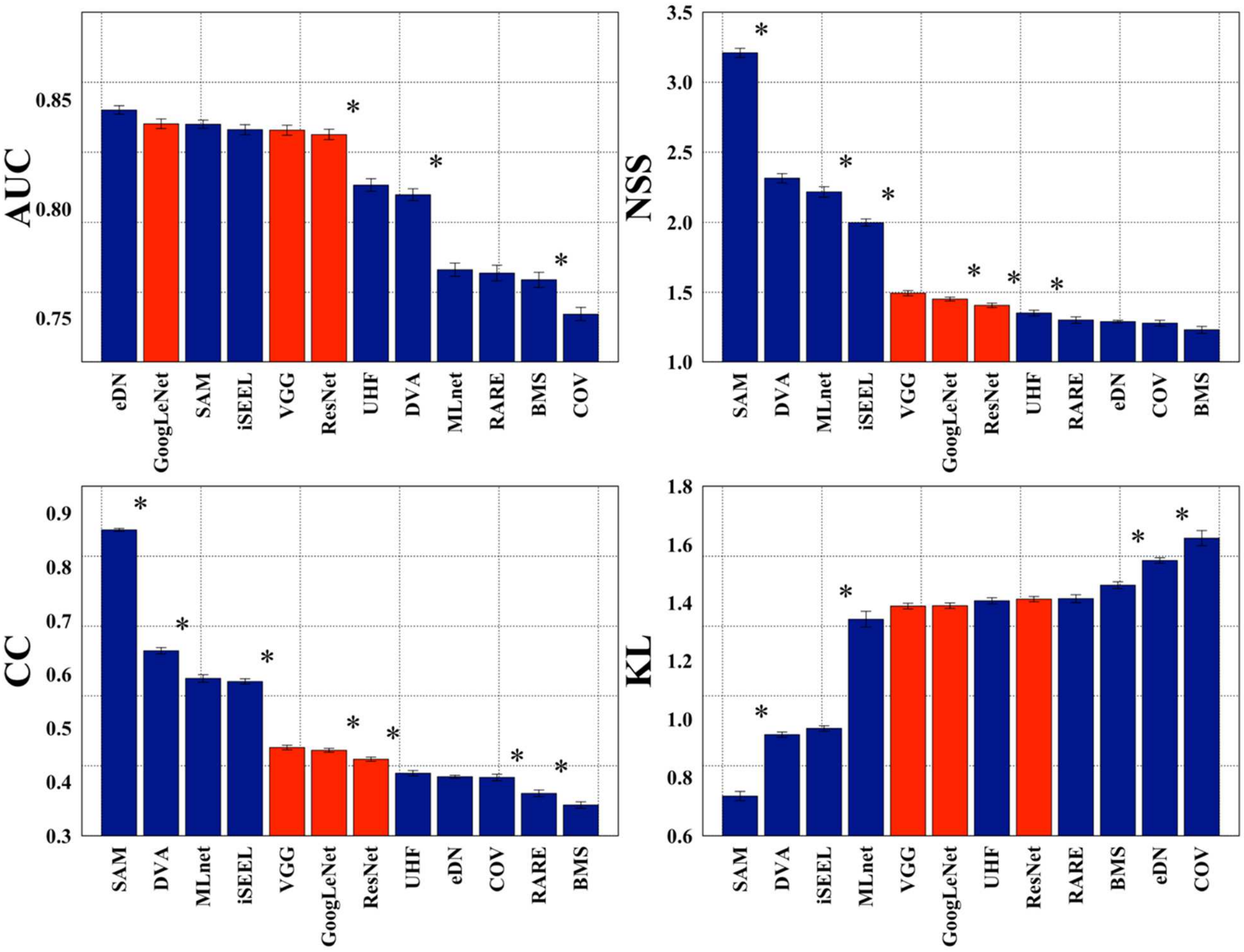


Figure 20 - Averaged AUC, NSS, CC, and KL scores of twelve saliency models including three variants of the DeepFeat model (VGG, GoogLeNet, and ResNet) and 9 other saliency models over the MIT1003 dataset. A * indicates the two consecutive models are significantly different using t-test at confidence level of $p \leq 0.05$. Models that are not consecutive have a larger probability to achieve statistical significance.

Generally speaking, the proposed DeepFeat models outperform the conventional saliency models and baseline learning models. It also can be found that the DeepFeat models can achieve comparable performance with top deep learning base saliency models in AUC score. The DeepFeat models cannot reach the performance of the top deep learning-based saliency models in NSS, CC, KL scores. However, the DeepFeat model does not require learning, which requires large training dataset and computational time. The DeepFeat model can be potentially applied to predict human gaze pattern in the cases of lacks training dataset, such as infant gaze pattern prediction.

5.3.3 Discussions:

The proposed DeepFeat model exploits deep features of a pre-trained DCNN. One advantage of the DeepFeat model is its fusion of a bottom-up and top-down saliency maps. In Eq. 5, the constant α is used as a weight for combining bottom-up and top-down maps. When α is 0, the saliency map is top- down. At α equal 1, the saliency map is bottom-up. The α may affect the performance of the DeepFeat model. To evaluate the effect of α, the saliency maps are computed by changing α ranging from 0 to 1 with 0.1 step. Figure 21 presents the mean scores of four metrics for the DeepFeat model with various α using the MIT1003 dataset. The results indicate that the combination of bottom-up and top-down saliency maps improves the prediction of human fixations. There is no consistent pattern on what the optimized value of is α. For GoogleNet and ResNet, the best performance is achieved when the α is 0.5-0.6. For VGG, the optimal α is about 0.25- 0.4. It indicates that the α value could be varied when using deep features pre-trained by different DCNNs. In our future work, more experiments will be conducted to optimize α by using more datasets and evaluation metrics.

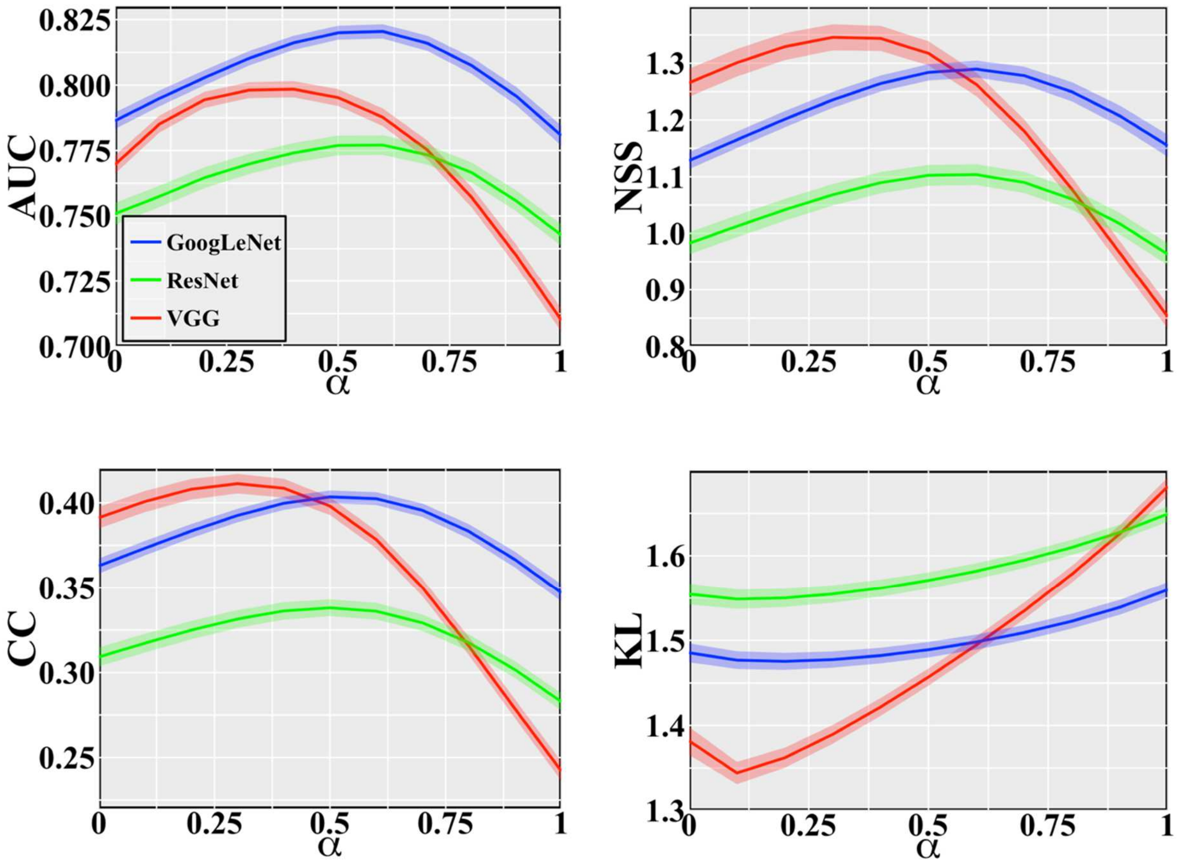


Figure 21 - Averaged curves of the combination of bottom-up and top-down over AUC, NSS, CC, and KL metrics using MIT1003 dataset. The smooth region surrounding the curves indicates SEM.

As shown in figure 21, the result indicates the top-down saliency maps outperform the bottom-up saliency maps with- out center bias. However, in few cases the bottom-up saliency maps outperformed the top-down saliency maps. Figure 22 presents three cases where the bottom-up saliency maps outperform the top-down saliency map. In figure 22, the top-down saliency maps fail to detect human or text which are not labels of the ImageNet dataset, while the detected objects are dominant in the images and belong to the ImageNet labels. While the top down fails to detect the human and text in figure 22, the bottom up predicts the missed salient regions. Such result indicates the combination of bottom up and top down improves the prediction of saliency. Moreover, the computed top-down saliency maps have no bias toward a class of objects in the fully connected layer. This occur because the MIT1003 and VIU datasets content include various scenarios of objects represented in images of the datasets.

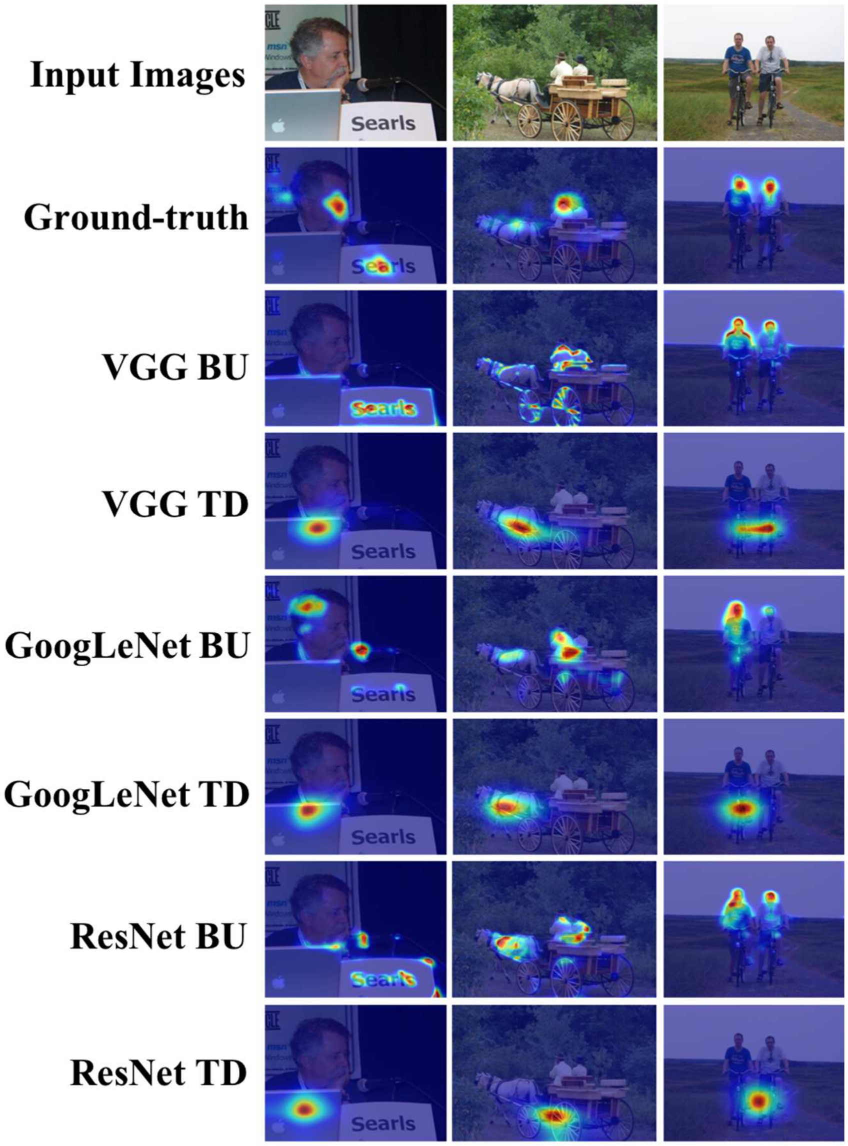


Figure 22 - Examples of bottom-up saliency maps outperforming top-down saliency maps. Row 1 show the photographs of three input images in MIT1003 dataset. Row 2 show the corresponding empirical saliency maps. Bottom-up and top- down saliency maps computed using three variants of the proposed DeepFeat model (VGG, GoogLeNet, and ResNet) are shown in row 3 to 8.

5.4 Conclusion:

In this chapter, a deep feature-based saliency model is proposed, which combines bottom-up and top-down visual factors obtained from pre-trained deep features of VGG, GoogLeNet, and ResNet DCNNs. To validate the performance of the DeepFeat model, different implementations of the DeepFeat model are investigated using four evaluation metrics over the MIT1003 and VIU datasets. The results demonstrate that the implementation of the DeepFeat model with incorporation of bottom-up and top-down saliency maps outperform the bottom-up and top-down saliency maps individually. Moreover, performance of the proposed DeepFeat model is evaluated in a comparison to nine state-of-the-art and conventional saliency models using four evaluation metrics over the MIT1003 dataset. The experimental results show that the pro- posed DeepFeat model outperforms the conventional saliency models. In future work, the performance of the DeepFeat model will be investigated on datasets other than natural image datasets such as webpages or text datasets. In addition, a parameterized version of the model will be learned, where Eq. 2 will be modified to a weighted sum of the response images of a layer.

CHAPTER 6

FEATURE BASED COMPARISON OF DEEP LEARNING NEURAL NETS

6.1 Introduction:

The recent advances of deep learning (DL) lead to a swarm of studies exploring ways to learn saliency models, including data representation and learning model architecture. In several studies, the DL based saliency models fine-tune pre-trained CNNs by adapting the weights of such CNNs as the initial weights for the saliency models. Pan and Giro proposed a CNN based model for saliency prediction [162]. The output of the CNN is obtained using max-out operation, and then convolved with a Gaussian filter to slightly smooth the saliency map. Kruthiventi et al. proposed a CNN based saliency model, which is inspired by the VGG network to predict pixel-wise saliency values [164]. Jetley et al. developed a saliency model based on a deep learning architecture, which formalizes a generalized Bernoulli distribution, and then trains a CNN with an architecture of convolutional layers identical to a VGG network [102]. Wang and Shen proposed an encoder-decoder based DL approach for saliency prediction [159]. The encoder consists of the first 13 layers of VGG-16 network. The multi-scale features are processed by the decoder which up-samples the multi-scale features, performs a deconvolution operation, and reduces the dimensionality of the feature maps. Pan et al. proposed a generative adversarial network for saliency prediction, and their proposed network consisting of a generator and a discriminator [165]. The generator is a CNN with identical structure as VGG-16 network, which consists of an encoder and a decoder. The discriminator consisted of convolutions and sigmoid activations followed by a fully connected layer. Cornia et al. modified the VGG/ResNet network by reducing the strides of the convolutional filters and adding a dilate after each layer, and then used the modified deep features as inputs to a long short-term memory (LSTM) network [166].

One bottleneck of the DL based saliency models is that the size of the avail- able training dataset is relatively small. As an alternative, transfer learning technology can be a viable solution, in which CNNs are pre-trained for one task, and used for another. Several DL based saliency models imply transfer learning by using deep features of pre-trained CNNs as inputs to new network. Vig et al. blended deep features from the first three layers of a biologically inspired CNN, and then combined the features using SVM classifier [96]. Huang et al. developed a deep learning-based saliency model that used two scales of pre-trained CNN [99]. They explored the feature maps of three pre-trained networks (i.e., AlexNet, VGG-16, and GoogLeNet), and then learned saliency weights using backpropagation. Kummerer et al. used deep feature of a pre-trained AlexNet CNN as inputs to SVM [97]. In their model, deep features are extracted and linearly combined, and then are processed by slightly blurring the result and adding a center bias. Later, they used deep features of a VGG-19 network as inputs to readout network, which consists of four convolution layers followed by ReLU nonlinearities [98]. Tavakoli et al. formalized a saliency model based on ensemble of extreme learning machines and inter-image similarities [160]. The model exploits deep features of a VGG-16 network to detect low-level visual cues, contextual information, and memorable events. Liu and Han exploited the deep features of two CNNs. One CNN solves a regression problem, and another CNN solves a classification problem. The two CNNs are combined and input to a LSTM network [101].

All above-mentioned studies demonstrate that the deep feature maps of pretrained CNNs for object classification can be fine-tune or optimized for prediction of human gaze patterns. In addition, the DeepFeat saliency model is developed, which exploits deep features of CNNs pre-trained for object classification as visual cues to predict human gaze patterns without any further training [48,167]. In this chapter, the framework of the DeepFeat model is used to investigate the role of deep features in saliency prediction, and extensively analyze and evaluate the deep features of different CNNs in a bottom-up manner, a top-down manner, and a combination of both bottom-up and top-down with and without the incorporation of the center bias.

6.1.1 Contributions:

The contributions of this chapter are threefold:

1. A comparison of 35 implementations of the bottom-up saliency maps is conducted using groups of deep features extracted from seven CNNs.
2. The influence of top-down visual attention is analyzed by comparing seven CNNs pre-trained for object classifications.
3. The role of the center bias in weighting the combined deep features from seven CNNs is evaluated.

6.2 Methods and Materials:

6.2.1 Deep Features:

In this chapter, 10 popular CNNs approaches are explored to evaluate how the deep features impact the saliency prediction. These 10 networks include seven classical CNN approaches (i.e., AlexNet, VGG-16, VGG-19, GoogLeNet, ResNet-50, ResNet-101, and ResNet-152) [114,122-124] and three CNN approaches based on CAM (AlexNet, VGG-16, and GoogLeNet) [154]. An extensive comparison is conducted to evaluate subsets of the network variants for the bottom-up implementation, the top-down implementation, and the combination of bottom- up and top-down implementations with and without the center bias.

To evaluate the bottom-up saliency implementation, 35 selections of activation layers from the seven classical CNN approaches are extensively compared. The complete description of 35 deep features selected for the bottom-up implementation is presented in Table 7.

Table 7 - Description of activation layers used as deep features for bottom up saliency implementation.

Activation	**Description**
	AlexNet
Conv	Convolution activations.
ReLU	Rectified linear unit activations.
Pool	Max pooling activations.
All	All activations of the network.
	VGG (16 & 19 layers)
Conv	Convolution activations.
ReLU	Rectified linear unit activations.
Pool	Max pooling activations.
All	All activations of the network.
	GoogLeNet
Conv	Convolution activations.
ReLU	Rectified linear unit activations.
Pool	Max pooling activations.
Incep	Inception module outputs.
All	All activations of the network.
	ResNet (50, 101, and 152)
Conv	Convolution activations.
Batch	Batch normalization.
ReLU	Rectified linear unit activations.
Concat	Concatenation between the network blocks and residuals.
Blocks	All network blocks except the residual short cuts.
All	All activations of the network including the residuals.

For top-down saliency implementation, the object localization is evaluated using four classical CNN approaches (GoogLeNet, ResNet-50, ResNet-101, and ResNet-152) and three CAM based CNN approaches. The CAM not only matches the size of the last activation and the score weights, but also modifies the object localization using global average pooling (GAP) instead of global maximum pooling (GMP). In this chapter, the top-down saliency implementation using GAP network variants are computed.

The combination of a bottom-up and top-down saliency implementations is computed by matching every top-down network to a selection of layers from a bottom-up network. The selected layers outperform other layers of the corresponding network.

6.2.2 Implementation details:

All four CNNs and their classical approaches are pre-trained on 1.28 million images of ILSVRC for 1000 classes of objects for object classification. The pre- trained classical CNN approaches are publicly available. The CAM based CNNs are pre-trained over the Places dataset which consists of 2.5 million images and 205 classes. The source code and the pre-trained CAM based CNN approaches are available online. All computations were done in MatConvNet and Caffe [155,168].

6.2.3 Datasets:

Four public datasets are exploited in this chapter under free-viewing conditions to ensure a comprehensive evaluation of the deep features using a variety of image contents and experimental settings. Four exploited datasets including KTH Koostra, MIT1003, OSIE, and Toronto. The complete description of the datasets can be found in chapter 2.

6.2.3 Evaluation Metrics:

In this chapter, three popular evaluation metrics are exploited to measure the agreement between the saliency predictions and the human annotations over four datasets. The three metrics are area under the receiver operator characteristic (ROC) curve (AUC), Pearson's correlation coefficient (CC), and similarity (SIM).

6.3 Experimental Results:

In this chapter, the bottom-up saliency maps, the top-down saliency maps, and combination of bottom-up and top-down saliency maps are evaluated over four datasets using three evaluation metrics. To measure the statistical significance, t-test for mean of scores is used at significance level $p \leq 0.05$. In addition, a comparison of the highest performance deep feature-based saliency implementation to six other popular saliency models over the MIT300 dataset is conducted.

6.3.1 Analysis of the bottom up saliency maps:

Figure 23 shows the ranking of the 35 bottom-up implementations over four datasets using three evaluation metrics. Although the ranks of the bottom- up implementations are varied over four datasets and three metrics, a general pattern can be observed. The GoogLeNet Incep is ranked first and the AlexNet ReLU ranked last. This is because the inception module in the GoogLeNet Incep network incorporates multiple levels of blur of deep features, which allows the object of interest to stand out from its surroundings. This conclusion is confirmed by the ResNet implementations, where the ResNet Concat for 50, 101, and 152 layers outperform the other ResNet implementations. In addition, the VGG16 Pool and the VGG19 Pool outperform the other VGG implementations. The only anomaly is that the AlexNet Conv ranks the highest among the AlexNet category and outperforms the AlexNet Pool. This may be caused by the depth of the AlexNet network. While the AlexNet is not as deep as the other exploited CNNs in this chapter, the other CNNs include a larger number of layers to average, and therefore, they tend to provide more suppression of non-salient regions.

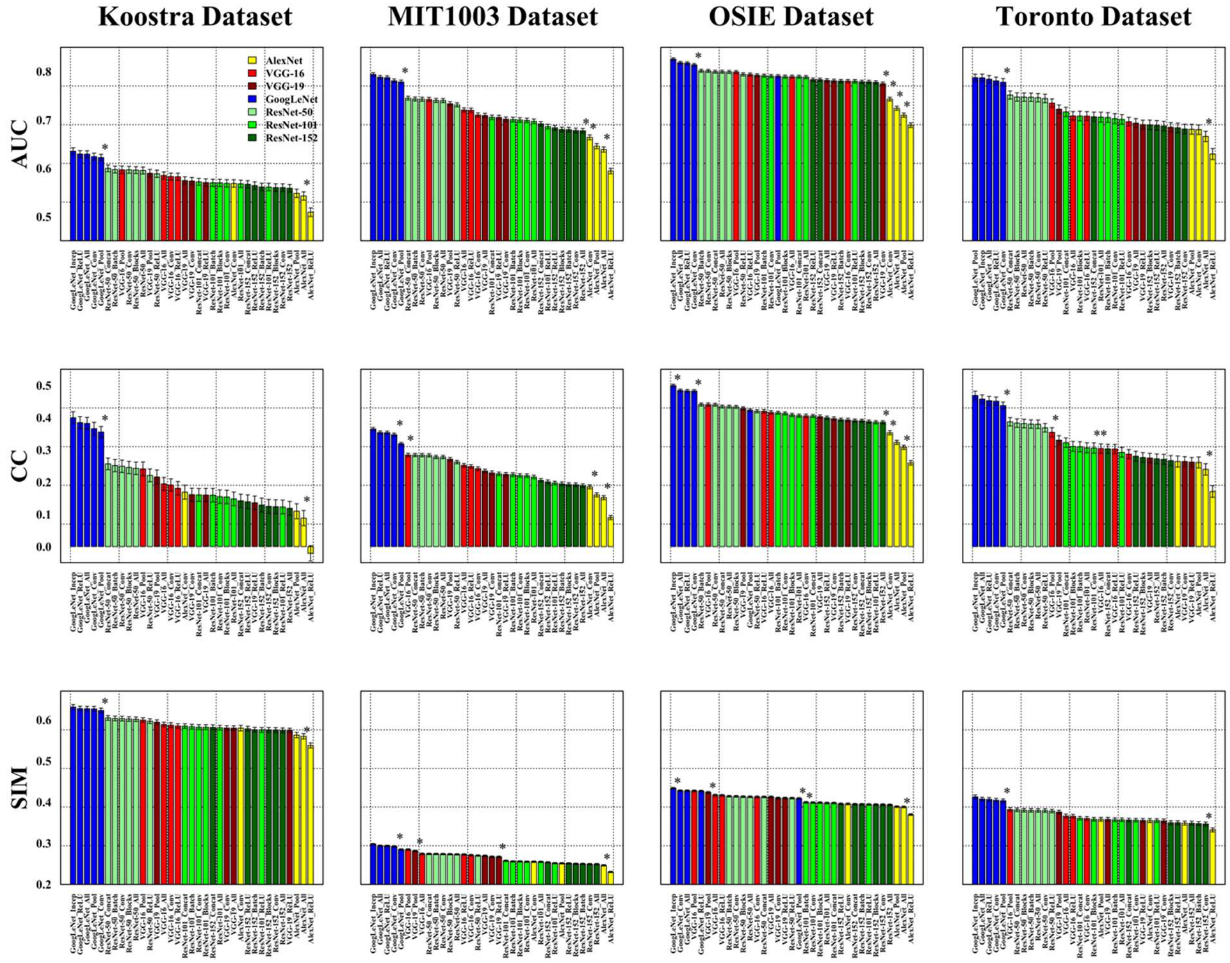


Figure 23 - Ranking of 35 bottom-up saliency implementations over four datasets using AUC, CC, and SIM evaluation metrics. A * indicates a significance at $p \leq 0.05$ between two consec- utive models using t-test. Non-consecutive models have a high probability to be significantly different. The error bars indicate standard error of the mean (SEMs).

Overall, the implementations of the GoogLeNet significantly outperform all other implementations. It indicates that the deep features of the GoogLeNet are highly correlated with the human visual system. Moreover, the implementations of the ResNet-50 outperform the implementations of the ResNet-101 and the ResNet-152. Also, the implementations of the VGG-16 outperform the implementations of the VGG-19. It indicates that the accuracy of bottom-up saliency map is not proportional to the number of network layers. This is because the effect of each layer is averaged and the increase of the number of layers may suppress the distinctive areas that appear in a smaller number of layers.

6.3.2 Analysis of the top down saliency maps:

To evaluate the top-down saliency maps, seven implementations using four classical CNN approaches and three CAM based CNN approaches are presented. Figure 24 presents the ranking of the seven top-down implementations over four datasets using three metrics.

In figure 24, consistent patterns are observed regard- less of the variation in ranking of the implementations. The GoogLeNetCAM implementation outperforms all other implementations over four datasets and three metrics. It indicates that the GoogLeNetCAM provides a better localization than the other implementations. In general, the CAM based implementations are among the top three rankings and outperform the other four top-down implementations. This is because the CAM based CNN approaches are pre-trained on Places dataset, which is larger than the ImageNet dataset. The result indicates that the deep features of the CAM based CNN approaches are more optimized than the deep features of the classical CNN approaches.

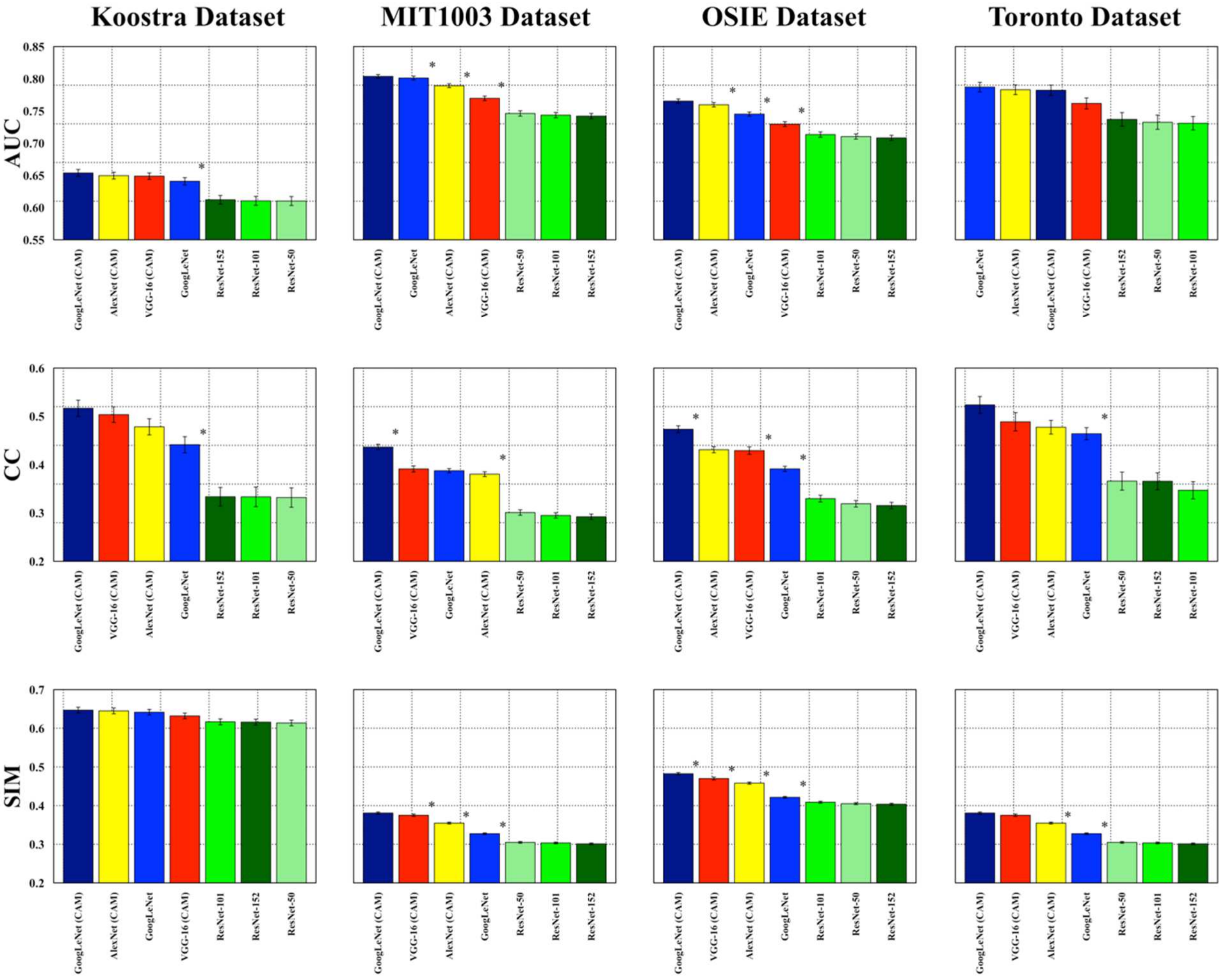


Figure 24 - Ranking of 7 top-down saliency implementations over four datasets using AUC, CC, and SIM evaluation metrics. A * indicates a significance at $p \leq 0.05$ between two consecutive models using t-test. Non-consecutive models have a high probability to be significantly different. The error bars indicate SEMs.

6.3.3 Analysis of the top down saliency maps:

Figure 25 presents eight representative images from four datasets, the corresponding ground-truth fixation maps, and four GoogLeNet based saliency maps, including the bottom-up GoogLeNet Incep, the top-down GoogLeNetCAM, and the combination of GoogLeNetCAM with and without the center bias. Overall, all saliency maps achieve a certain accuracy compared with the ground-truth fixation maps in all eight images.

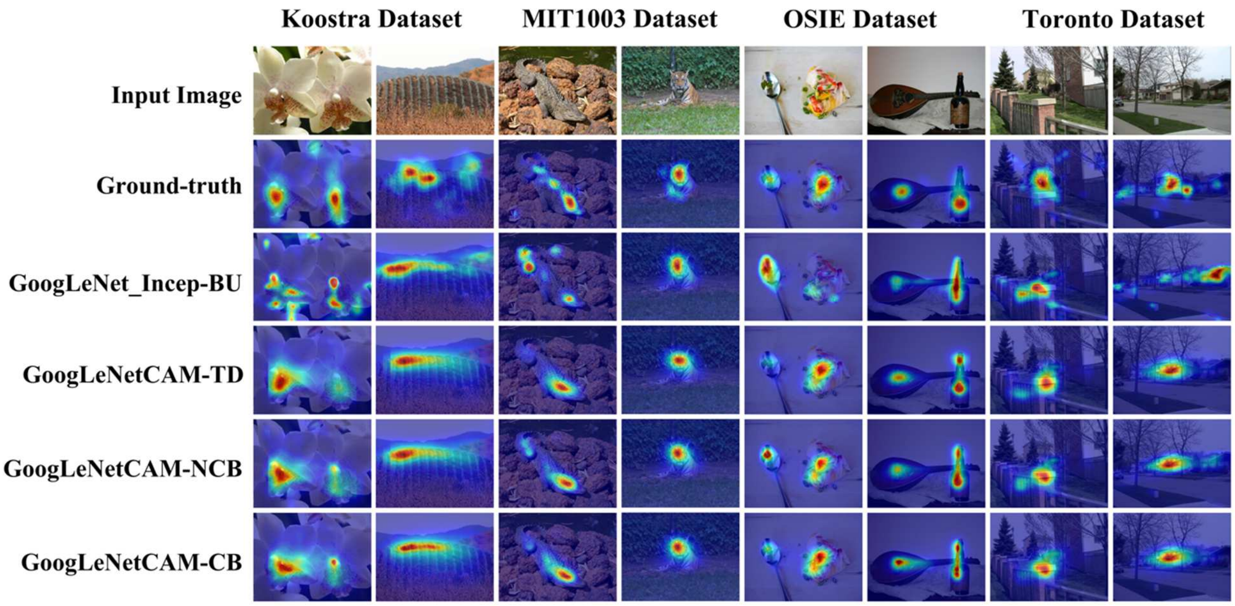


Figure 25 - Row 1 presents eight representative images from four datasets. Row 2 is the ground-truth maps of the corresponding images. Row 3 to row 6 are the four saliency maps of the GoogLeNet implementations, including the bottom-up GoogLeNet Incep, the top-down GoogLeNetCAM, and the combination of GoogLeNetCAM with and without the center bias, respectively. For visualization purpose, the histogram of the predicted saliency maps of both models are matched to the histogram of the dataset ground-truth.

Figure 26 presents the combined implementations with and without the center bias over four datasets using three evaluation metrics. Over the AUC scores, implementations that incorporate the center bias outperform all the implementations without the center bias over the Koostra, MIT1003, and Toronto datasets. Such result may be caused by the property of AUC, where it gives more credit to predictions near the center of the image. The incorporation of the center bias boosts the AUC scores of the saliency models with the center bias. In addition, saliency implementations demonstrate inconsistent performance over the OSIE dataset. While the AlexNet, and ResNets (50, 101, and 152) implementations with the center bias outperform their corresponding implementations without the center bias, the VGG-16, the GoogLeNet, and the GoogLeNet- CAM implementation without the center bias outperform their corresponding implementations with the center bias. It indicates the predictions of these three implementations without the center bias are more localized toward the center than these implementations without the center bias.

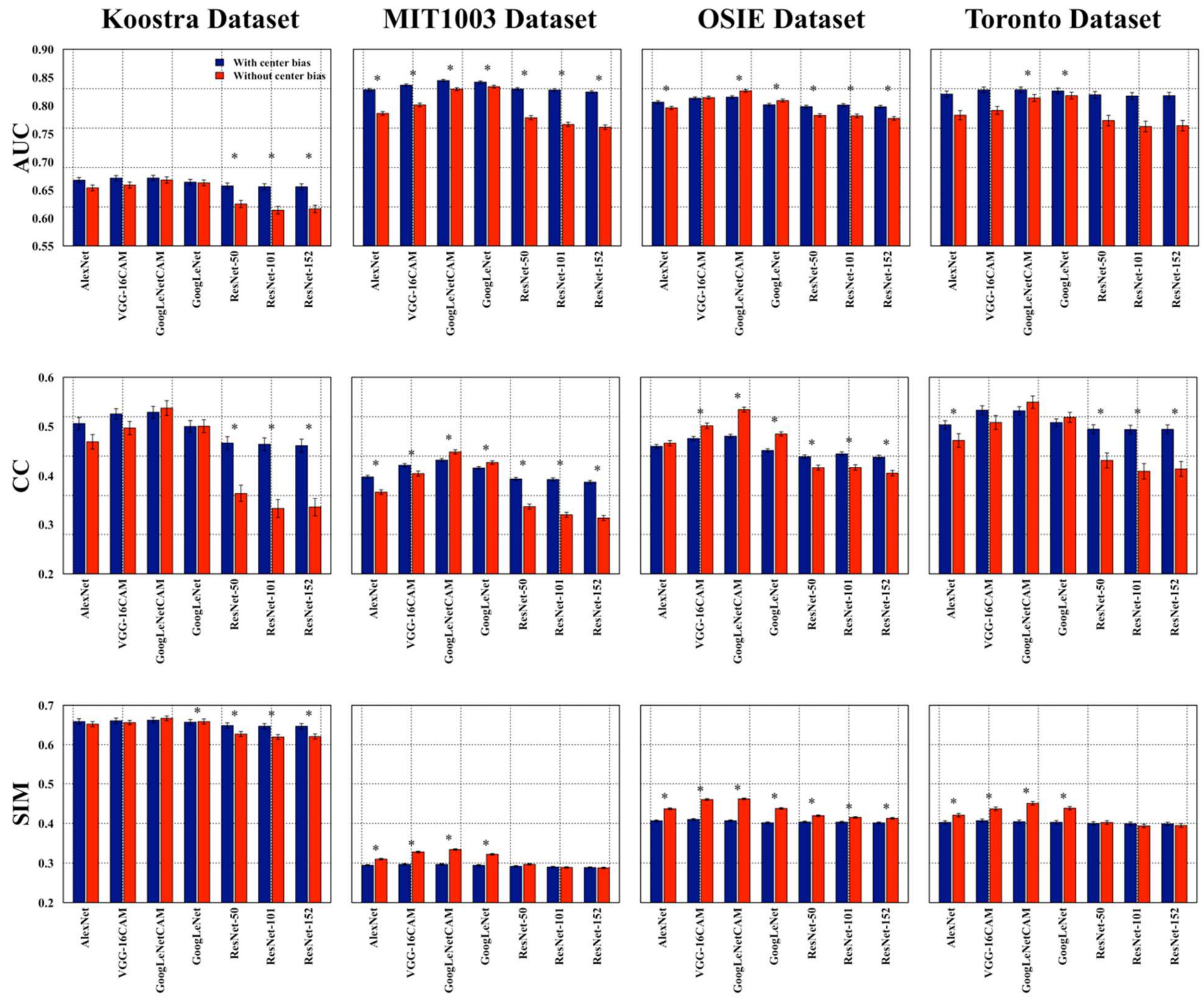


Figure 26 - Average AUC, CC, and SIM scores of various saliency maps, which are combinations of bottom-up and top-down implementations with and without the center bias over four datasets. A * indicates a significance at $p \leq 0.05$ between two consecutive models using t-test. The error bars indicate SEMs.

For the CC scores, the GoogLeNet and the GoogLeNetCAM implementations without the center bias outperform their corresponding implementations with the center bias over all four datasets. All other implementations with the center bias outperform their corresponding implementations without the center bias. Using the SIM metric, the VGG-16CAM, the GoogLeNetCAM, and the GoogLeNet implementations without the center bias outperform the corresponding implementations with the center bias over all four datasets. The AlexNet implementation without the center bias outperforms the AlexNet implementation with the center bias over all datasets except the Koostra dataset. The performances of the ResNet implementations fluctuate over four datasets. Overall, the center bias boosts the performance of the saliency implementations except the GoogLeNet and the GoogLeNetCAM.

In addition, for different datasets, the Koostra dataset has the lower AUC scores and the higher SIM scores compared with the other three datasets. It may be caused by different complexity of scenes of the datasets. The complete comparisons are described in table 8.

Table 8 - The combination of bottom up and top down results with and without center bias over four datasets using three evaluation metrics. Red, green, and blue color scores indicate the top three rankings models over individual scores, respectively.

Koostra Dataset						
	With center bias			**Without center bias**		
Implementation Name	**AUC**	**CC**	**SIM**	**AUC**	**CC**	**SIM**
AlexNetCAM	0.668 ± 0.005	0.506 ± 0.012	0.659 ± 0.007	0.654 ± 0.005	0.469 ± 0.015	0.652 ± 0.07
VGG-16CAM	0.671 ± 0.005	0.526 ± 0.011	0.661 ± 0.007	0.659 ± 0.005	0.497 ± 0.014	0.656 ± 0.005
GoogLeNetCAM	0.671 ± 0.005	0.529 ± 0.012	0.662 ± 0.007	0.668 ± 0.006	0.538 ± 0.015	0.667 ± 0.006
GoogLeNet	0.664 ± 0.005	0.500 ± 0.012	0.657 ± 0.007	0.663 ± 0.005	0.501 ± 0.013	0.657 ± 0.007
ResNet-50	0.657 ± 0.005	0.467 ± 0.013	0.649 ± 0.007	0.625 ± 0.005	0.364 ± 0.017	0.627 ± 0.006
ResNet-101	0.656 ± 0.005	0.464 ± 0.013	0.647 ± 0.007	0.614 ± 0.007	0.333 ± 0.019	0.620 ± 0.006
ResNet-152	0.656 ± 0.005	0.461 ± 0.013	0.647 ± 0.007	0.617 ± 0.007	0.336 ± 0.018	0.621 ± 0.006
MIT1003 Dataset						
AlexNetCAM	0.828 ± 0.002	0.398 ± 0.003	0.294 ± 0.002	0.786 ± 0.003	0.367 ± 0.005	0.310 ± 0.002
VGG-16CAM	0.836 ± 0.002	0.422 ± 0.003	0.297 ± 0.002	0.801 ± 0.003	0.405 ± 0.005	0.328 ± 0.002
GoogLeNetCAM	0.845 ± 0.002	0.432 ± 0.003	0.297 ± 0.002	0.829 ± 0.003	0.448 ± 0.005	0.334 ± 0.002
GoogLeNet	0.842 ± 0.002	0.416 ± 0.003	0.294 ± 0.002	0.834 ± 0.003	0.427 ± 0.004	0.322 ± 0.002

ResNet-50	0.830 ± 0.003	0.394 ± 0.003	0.292 ± 0.002	0.779 ± 0.004	0.337 ± 0.005	0.297 ± 0.002
ResNet-101	0.828 ± 0.002	0.393 ± 0.003	0.290 ± 0.002	0.767 ± 0.004	0.320 ± 0.005	0.290 ± 0.002
ResNet-152	0.824 ± 0.003	0.387 ± 0.003	0.289 ± 0.002	0.762 ± 0.004	0.313 ± 0.005	0.288 ± 0.002
OSIE Dataset						
AlexNetCAM	0.806 ± 0.003	0.460 ± 0.004	0.407 ± 0.002	0.796 ± 0.003	0.466 ± 0.005	0.437 ± 0.002
VGG-16CAM	0.813 ± 0.002	0.476 ± 0.004	0.410 ± 0.002	0.814 ± 0.003	0.502 ± 0.006	0.460 ± 0.002
GoogLeNetCAM	0.815 ± 0.002	0.481 ± 0.004	0.407 ± 0.002	0.826 ± 0.003	0.534 ± 0.005	0.462 ± 0.002
GoogLeNet	0.802 ± 0.003	0.452 ± 0.004	0.402 ± 0.002	0.809 ± 0.003	0.485 ± 0.004	0.438 ± 0.002
ResNet-50	0.798 ± 0.003	0.439 ± 0.004	0.404 ± 0.002	0.783 ± 0.003	0.417 ± 0.005	0.419 ± 0.002
ResNet-101	0.801 ± 0.003	0.445 ± 0.004	0.403 ± 0.002	0.782 ± 0.003	0.417 ± 0.006	0.415 ± 0.002
ResNet-152	0.798 ± 0.003	0.438 ± 0.004	0.401 ± 0.002	0.777 ± 0.003	0.401 ± 0.006	0.413 ± 0.002
Toronto Dataset						
AlexNetCAM	0.821 ± 0.006	0.504 ± 0.008	0.403 ± 0.004	0.783 ± 0.008	0.472 ± 0.014	0.421 ± 0.004
VGG-16CAM	0.828 ± 0.005	0.533 ± 0.009	0.407 ± 0.004	0.792 ± 0.007	0.508 ± 0.014	0.437 ± 0.005
GoogLeNetCAM	0.828 ± 0.005	0.532 ± 0.008	0.404 ± 0.004	0.814 ± 0.006	0.550 ± 0.013	0.451 ± 0.005
GoogLeNet	0.826 ± 0.005	0.508 ± 0.007	0.403 ± 0.004	0.818 ± 0.006	0.519 ± 0.010	0.438 ± 0.004
ResNet-50	0.819 ± 0.006	0.495 ± 0.009	0.400 ± 0.004	0.774 ± 0.009	0.431 ± 0.015	0.402 ± 0.005
ResNet-101	0.817 ± 0.006	0.494 ± 0.009	0.399 ± 0.004	0.763 ± 0.010	0.409 ± 0.016	0.394 ± 0.005
ResNet-152	0.818 ± 0.006	0.495 ± 0.009	0.399 ± 0.004	0.764 ± 0.009	0.414 ± 0.015	0.395 ± 0.005

6.3.4 Comparison with other saliency models:

To evaluate the ability of deep features to predict human fixations, two GoogLeNetCAM based implementations with and without the center bias (GoogLeNetCAM-CB and GoogLeNetCAM-NCB) to six other popular saliency models are compared. For a fair comparison, predictions of our two GoogLeNetCAM implementations are computed and evaluated over the MIT300 dataset, which consists of 300 indoor and outdoor images observed by 39 observers for 3 seconds [169]. Six other saliency models include two deep learning-based saliency models (DeepGaze1 [97], and eDN [96]), a shallow learning-based saliency model (Judd [30]), and three conventional saliency models (GBVS [58], LGS [170], and RC [171]).

The complete results over 8 evaluation metrics are available online [110,127]. In this chapter, a comparison results over three evaluation metrics is presented.

Table 9 summarizes the comparison results of the two GoogLeNetCAM implementations and 6 other saliency models over the MIT300 dataset using AUC, CC, and SIM evaluation metrics. For the AUC scores, DeepGaze1 is ranked first, the GoogLeNetCAM-CB is ranked second, and eDN is ranked third.

Over the CC scores, GoogLeNetCAM-NCB is ranked first, GoogLeNetCAM-CB, DeepGaze1, and GBVS are ranked second, and Judd and RC are ranked third. Using the SIM metric, GBVS and RC are ranked first, GoogLeNetCAM- NCB is ranked second, and GoogLeNetCAM-CB, Judd, and LGS are ranked third. In general, both GoogLeNetCAM implementations perform among the top three ranking models in the comparison. It indicates that the deep features of CNNs highlights the image semantics that can be used to model a saliency map comparable to popular saliency models.

Table 9 - The comparison of two deep features of CNNs based saliency implementations and 6 state-of-the-art saliency models over the MIT300 dataset. The top three ranking models are marked red, green, and blue, respectively.

Saliency Model	**AUC**	**CC**	**SIM**
GoogLeNetCAM-CB	0.82	0.48	0.42
GoogLeNetCAM-NCB	0.80	0.49	0.45
DeepGaze1	0.83	0.48	0.39
eDN	0.81	0.45	0.41
GBVS	0.80	0.48	0.48
Judd	0.80	0.47	0.42
LGS	0.76	0.39	0.42
RC	0.78	0.47	0.48

6.4 Conclusion:

In this chapter, deep features are explored via different saliency implementations to evaluate the effects of deep features on saliency prediction of human gaze patterns. Such deep features are obtained from seven popular CNNs. The networks are pre-trained using the original proposed approaches and the modified class activation maps approaches. A series of comparisons are conducted to evaluate the performances of various implementations, including the bottom- up, top-down, and the combination of both with and without the center bias, over four datasets using three evaluation metrics. In addition, the performances of the deep features-based saliency models are evaluated by comparing to six other popular saliency models. The experimental results indicate that the deep features from all pre-trained CNNs are useful for saliency modeling. The increase in number of layers may not be helpful for detecting low level factors. Instead, the incorporation of multiple levels of blurred features boosts the detection of low-level cues. CAM based CNN approaches provide more localized objects that are useful for top-down saliency modeling. Moreover, the incorporation of the center bias boosts the performance of saliency predictions over several implementations.

CHAPTER 7

CLASSNET: A CLASSIFIER FOR VISUAL ATTENTION PREDICTION

7.1 Introduction:

A visual stimulus triggers cells and photoreceptors of the human eye. The resulting signals travel through the optic nerve and stimulate neurons of the brain to give visual representations of the surrounding world [172-174]. In visual perception, the human visual system tends to minimize its neural resources by sampling the most informative areas of the visual stimuli [175]. Such mechanism is known as visual attention. Both biological (exogeneous) and psychological (endogenous) influences guide the human visual attention [176,1].

In computer vision, a saliency map is defined as a prediction of the human attention probability [24]. It is modeled as a 2D map that assigns levels of attention priority to spatial locations of the map. Saliency modeling is viable for a variety of applications [2,7,9,12]. A previous chapter demonstrates that the human visual system processes the visual stimuli using preliminary features in multiple scales [22]. These features are processed in parallel and fused to aid for recognition of the input visual stimuli. Such work lead to the development of a tremendous number of saliency models relying on hand crafted features such as color, intensity, and orientation [52]. The limitation of such saliency models is that their predictions correspond to the incorporated features only. Later, the development of deep learning-based saliency models overcome the limitation of the hand-crafted features. Several studies learned a saliency model in an end-to-end manner [176] or using transfer learning [97-98]. The end-to-end saliency models exploit weights of pre-trained CNN as initializers for their proposed models. Moreover, transfer learning-based saliency models learn a combination of deep features of pre-trained CNNs. Such directions in visual saliency prediction due to the relatively small datasets available to train a deep learning model. To this extend, there is no deep learning saliency model that learns its weights from scratch.

In this chapter, a deep learning classification framework is proposed. The framework focuses on preprocessing the image dataset to be useful for training a saliency model from scratch. In addition, a modified version of ResNet is proposed for visual saliency prediction. The proposed approach is codenamed ClassNet. While previous deep learning saliency models treat the saliency prediction as a regression problem, the proposed framework treats the saliency prediction as a classification problem.

7.1.1 Contributions:

The contributions of this chapter are threefold:

1. A data generation protocol is proposed to increase the number if samples for training a deep learning saliency model from scratch.
2. A modification of ResNet20 [124] is proposed in order to train a saliency model from scratch.
3. An evaluation of the proposed framework over five datasets using four evaluation metrics.

7.2 Proposed Approach:

7.2.1 Data Preparation Protocol

In order to learn a saliency model from scratch, a human fixation dataset is preprocessed to increase the number of image samples. As a classification problem, the number of samples is increased by cropping an image and labeling the cropped sample to fixation or non-fixation. An OSIE dataset is utilized to create a large enough dataset for training a deep learning saliency model. For each image, all fixation points that falls in the top 20% salient locations of the distribution-based fixation map are labeled as fixation. The spatial points selected from the top 20% and bottom 30% ensures a pattern difference between the two labels. Spatial locations that falls in the bottom 30% salient locations of the distribution-based fixation map are labeled as non-fixation. The number of non-fixation points randomly selected is equal to the number of fixation points in the same image. Figure 27 presents fixation and non-fixation labels of patches for an image. Moreover, for every selected label the image is cropped to 64×64 pixels where the center of the cropped patch corresponds to the location of the label location. The resulting dataset consist of $61,\ 080$ samples which is large enough to classify two classes only.

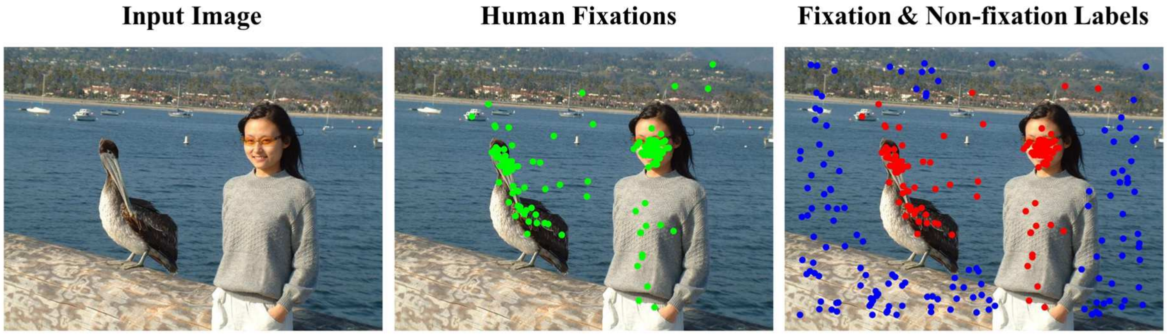


Figure 27 - Training and testing labels of fixations overlay an input image. Green points are actual fixation points, red points are a subset of the actual fixation points labeled as fixation, and the blue points are non-fixated points labeled as non-fixation.

7.2.2 Residual Learning:

In this dissertation, the degradation problem in deep neural networks is exploited. Such that, a stack of nonlinear layers fit a residual mapping instead of fitting stacked layers directly. Formally, let $\mathcal{H}(x)$ denote the desired underlying mapping. Then, the stacked nonlinear layers fit another mapping:

$$\mathcal{F}(x) \coloneqq \mathcal{H}(x) - x \tag{33}$$

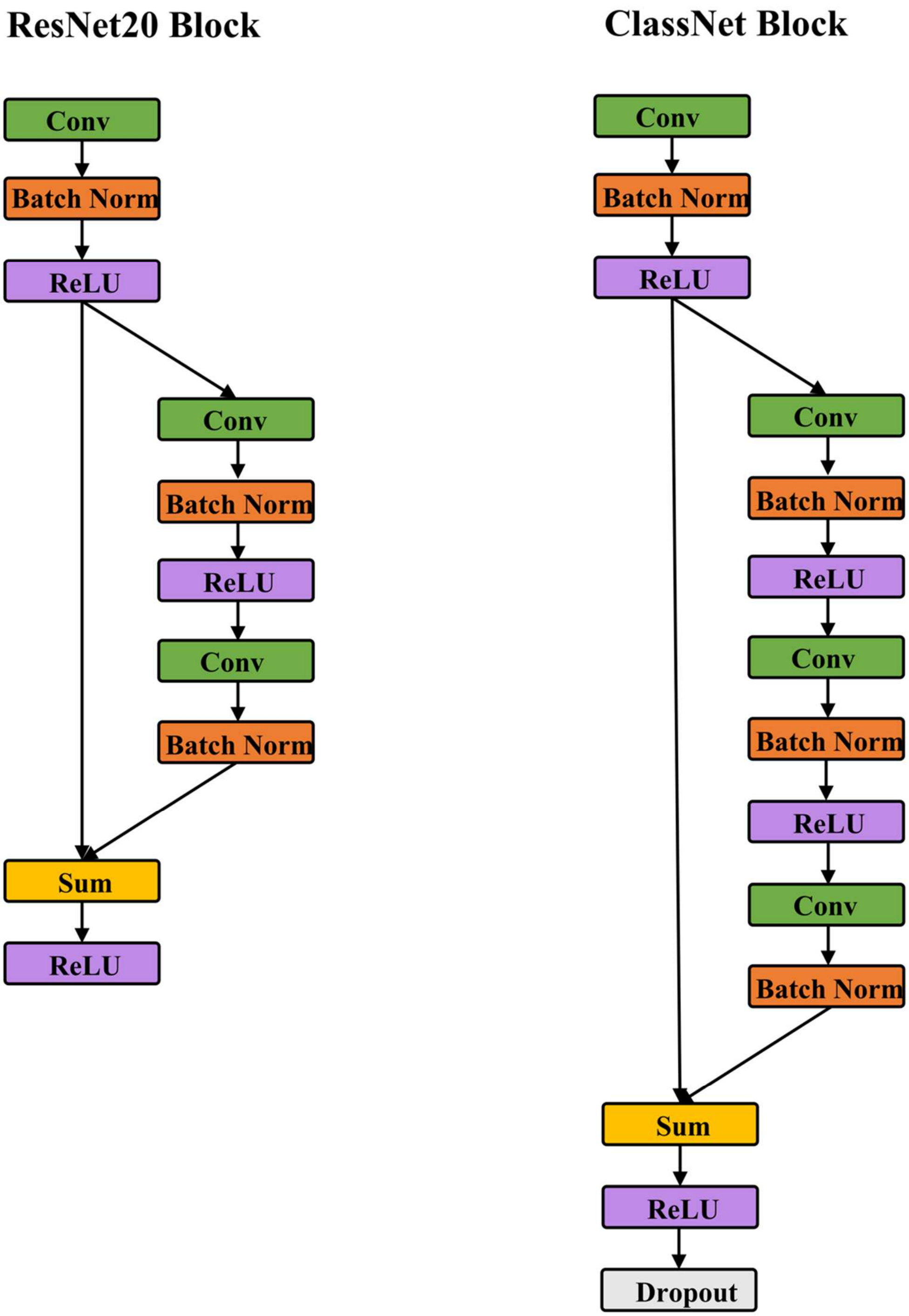
Where x denote the input to these layers. The residual learning hypothesis indicates it is easier to optimize the residual mapping, than to optimize the original mapping. Therefore, the residual learning can be performed by:

$$\mathcal{H}(x) \coloneqq \mathcal{F}(x) + x \tag{34}$$

The reformulation suggests a deeper model should have training error no larger than its shallow counterpart. The degradation problem suggests that the approximation of identity mapping using multiple nonlinear layers is a difficult task. Therefore, the use of residual mapping reformulation may drive weights of multiple nonlinear layers toward zero to achieve identity mapping.

7.2.3 ClassNet Saliency Model:

Residual learning neural networks achieved outstanding performances on several computer vision applications. However, they can’t be employed directly for saliency prediction without any fine-tuning. A modified version of ResNet20 is proposed. The main difference between the ResNet20 and ClassNet is the architecture of the residual block. The residual block of ClassNet consist of three convolution layers in the skip connection. The first two convolutions in the residual block are followed by a batch normalization layer followed by a ReLU activation. The third convolution layer of the residual block is followed by a batch normalization layer. Furthermore, the summation of previous layer and skip connection is followed by a ReLU activation and a dropout layer. Figure 28 presents the differences between the residual block of ResNet20 and ClassNet.



ResNet20 Block

ClassNet Block

Figure 28 - A comparison of the ResNet 20 residual block architecture, and ClassNet residual block architecture.

After every six residual blocks, the number of features is doubles while the feature maps are down-sampled to preserve the computation complexity per layer. The down-sampling is performed by convolutions with a stride of 2. After the last convolution layer, the global average pooling (GAP) is replaced with global contrast measure weighted by:

$$w_i = 0.5\left\{cos\left[\frac{\pi}{r}\sqrt{(x_i - x_c)^2 + (y_i - y_c)^2}\right] + 1\right\} \tag{35}$$

where r denotes the patch radius, $x_i$and $y_i$denotes the spatial location of the ith pixel in the image, and $x_c$and $y_c$denotes the spatial location of the center of the patch. Note that, an image spatial location is expressed in pixels in the vertical and horizontal directions. On every spatial location, the global contrast for every feature map is measured by:

$$C = \sqrt{\frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \frac{(I_i - I)^2}{I^2}} \tag{36}$$

The n denotes the number of pixels in the patch, $I_i$denotes the luminance value of the ith pixel in the patch, and I denote the mean luminance of the patch calculated by:

$$I = \frac{1}{\sum_{i=1}^{n} w_i}\sum_{i=1}^{n} w_i \, I_i \tag{37}$$

Finally, the model contains a single FC that has two neurons.

7.2.4 Data Augmentation

To increase the variance and number of patch samples, a data augmentation is performed. In this work, the patch width is randomly shifted horizontally by a fraction of 10% of the patch width. Similarly, the patch height is randomly shifted vertically by 10% of the patch height. In addition, images are randomly flipped vertically and horizontally.

7.2.5 Implementation details

In order to train a deep learning model, the OSIE dataset is divided into 500 training images and 200 testing images. The training and testing images are normalized individually. Each image is normalized by:

$$I_{\mathcal{N}} = \frac{I-\min(I)}{\max(I)-\min(I)} \tag{38}$$

The I denotes an image to be normalized. Using the normalized images, a larger dataset of image patch samples is created. The training set consist of 42,992 image patch samples, and the testing set consist 18,088 image patch samples. The image patches are randomly shuffled and divided into mini-batches of 32 patch samples each. Figure 29 presents examples of fixation and non-fixation samples.

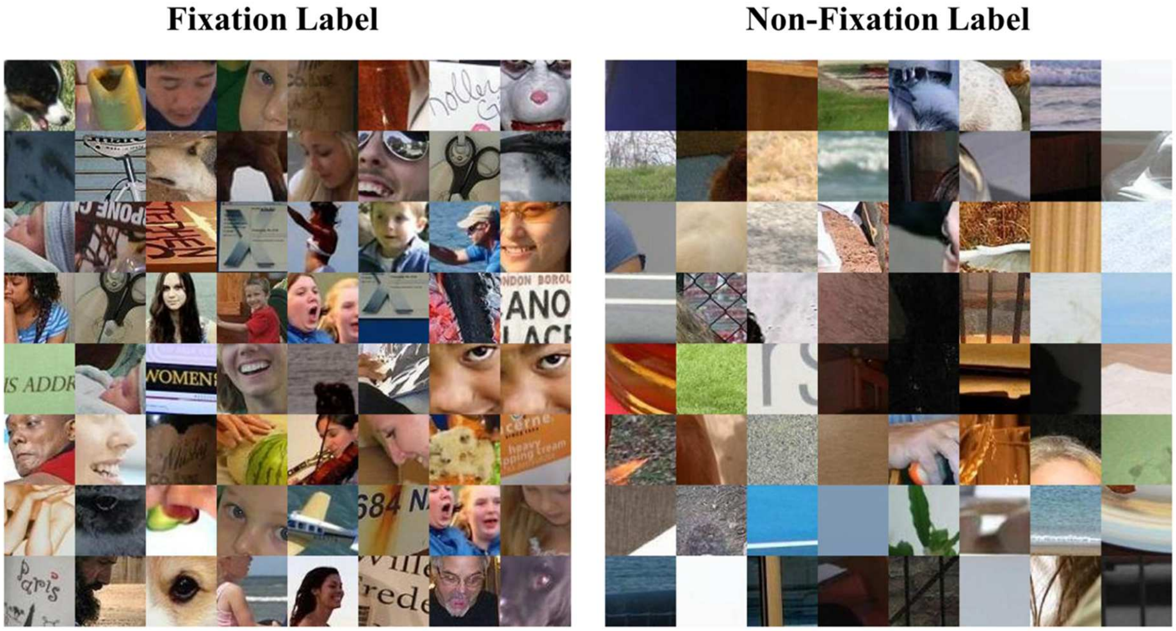


Figure 29 - Example of patch datasets labeled as fixation in the left panel and non-fixation on the right panel.

The loss function between the predicted label and the actual label is defined as the cross-validation. The loss is updated using ADAM optimizer with a learning rate equal to 0.001. The training process continues for 200 epochs. For validation, the testing set is exploited without data augmentation. The training process took about 90 minutes of training using two Titan X GPU's. The testing set serves as a validation technique to evaluate the training performance. For accuracy performance measure, the accuracy is defined as the ration of correctly predicted classes to the total number of predictions.

For image prediction, an image is normalized, and the learned weights are applied to the normalized image pixels. To reduce the computation complexity, pixels are selected for prediction in a square lattice with 32 pixels distance between every two pixels horizontally and vertically. The size of 32 pixels to ensure the selected patches do have an overlap with each other. Furthermore, the prediction array of selected pixels is resized to the size of the original input image.

7.3 Experimental Setup:

7.3.1 Datasets

Five datasets are exploited in this chapter including MIT1003, VIU, OSIE, Toronto, and Koostra. Although the OSIE dataset is used for training the model, the model performance over all five datasets is compared to evaluate the amount of overfitting occurred during the training.

7.3.2 Evaluation Metrics

To measure the agreement between the model predictions and the human annotations, four evaluation metrics are used. All four metrics are similarity-based scores including two fixation-points based scores and two fixation-map based scores. The four scores are: AUC (Judd), NSS, SIM, and CC.

7.4 Experimental Results

Performance of the proposed framework is evaluated over five datasets including the dataset used for training. Figure 30 presents two sample images from every dataset used in this chapter and their corresponding predictions. In figure 30, the ClassNet predictions are more square regions. This is mainly due to the fact the interpolation is between spatial values of ones.

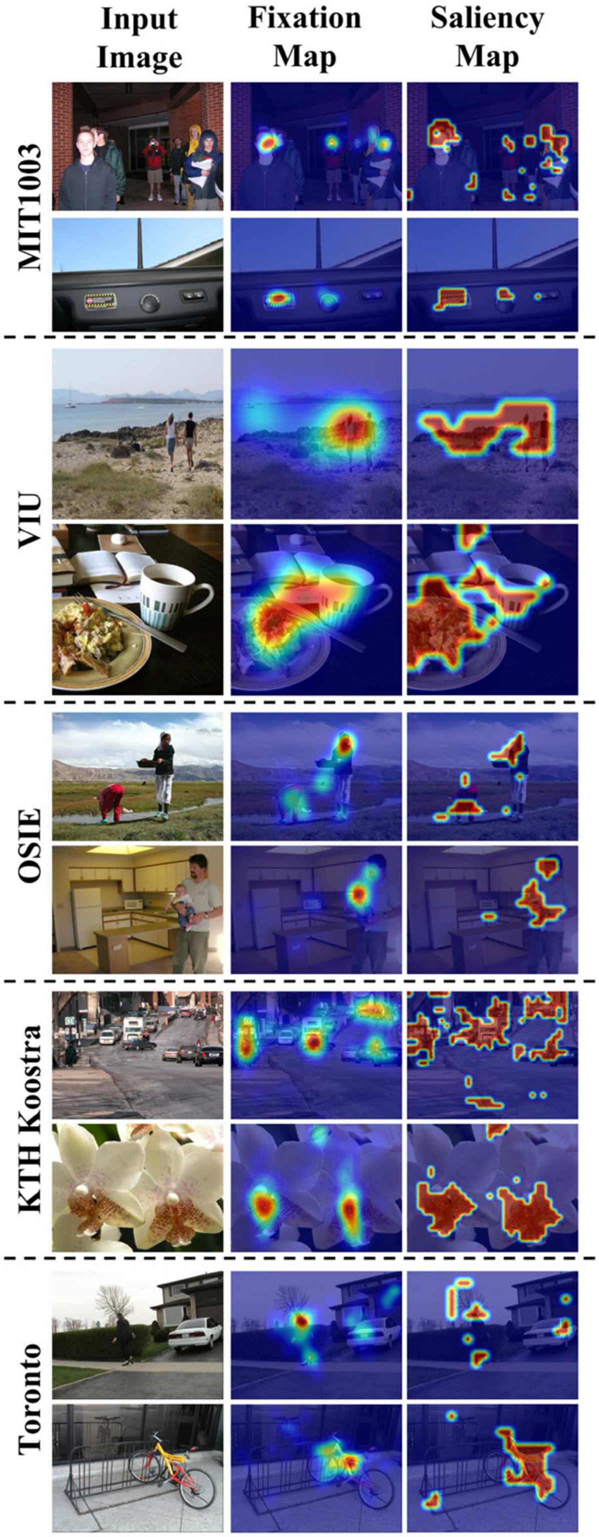


Figure 30 - Column 1 presents ten representative images from five datasets. Column 2 is the ground- truth maps of the corresponding images. Column is saliency maps of ClassNet.

Moreover, to analyze performance of the proposed framework, four scores are drawn as shown in figure 31. Using the AUC score, predictions of the model over all datasets is over 0.5 which demonstrate the proposed framework perform higher than random guessing. Over the NSS, SIM, and CC, performance of the proposed model over all five datasets achieves a certain agreement with the ground-truth. Such result indicates that the proposed framework is a valid approach to train a fixation prediction model from scratch.

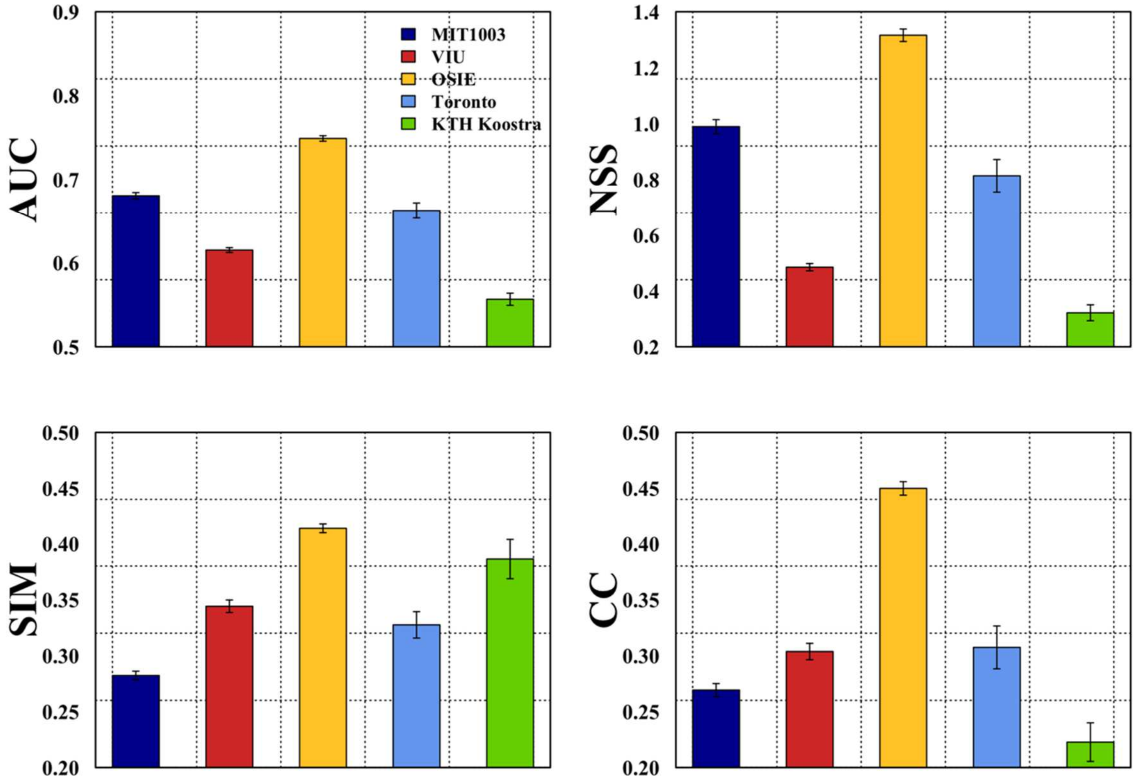


Figure 31 - Averaged AUC, NSS, SIM, and CC scores of the proposed framework over five datasets. The error bars indicate SEM.

In general, the performance of the ClassNet over the OSIE dataset outperform the performance of the model over other datasets. This occurs because the model overfits over the

OSIE which the training dataset is. However, this result is preliminary, and the purpose was to demonstrate the scalability of small datasets to train a saliency model from scratch. The complete analysis results are presented in table 10.

Table 10 - Average scores of ClassNet over five datasets.

Dataset	**AUC**	**NSS**	**SIM**	**CC**
MIT1003	0.681±0.004	0.990±0.025	0.282±0.004	0.269±0.006
VIU	0.615±0.003	0.485±0.013	0.334±0.006	0.304±0.007
OSIE	0.749±0.003	1.316 ±0.022	0.414±0.004	0.450±0.006
Toronto	0.663±0.009	0.813±0.058	0.327±0.012	0.307±0.019
KTH Koostra	0.557±0.007	0.322±0.028	0.386±0.018	0.223±0.017

7.5 Conclusion

Although the recent trends demonstrate high prediction accuracy of human visual attention, all deep learning saliency models exploit pre-trained CNNs before they start training their models. Such structure is useful for prediction. However, to chapter the relationship between the learned weights and the response filters in the human eye it is essential to train a saliency model from random weights. This dissertation presents a deep learning framework to generate a large datasets of patch samples from a small dataset of images. Also, a modification of ResNet20 is presented and codenamed ClassNet. The validity of such framework is evaluated over five datasets of images. While the experimental results are preliminary, the results demonstrate the proposed framework is valid and achieves higher than random guessing.

CHAPTER 8

CONCLUSIONS

8.1 Summary:

A tremendous number of saliency models have been developed over the years. The performance of saliency models is usually evaluated on datasets that carry out eye fixations recorded from adults. Despite the consistency in adults gaze patterns, infants gaze patterns are not random. To explore infants and adults gaze patterns, Infants and adults extensive comparisons using 8 state of the art saliency models and two baselines are conducted. Seven standard evaluation metrics are exploited to measure the agreement between the models and eye fixations from infants and adults. The results demonstrate a consistent performance of saliency models predicting adults' fixations over infants' fixations in terms of overlap, center fitting, intersection, information loss of approximation, and spatial distance between the distributions of saliency map and fixation map. In saliency models and baselines ranking, the GBVS and Itti are among the top 3 contenders, infants and adults have bias toward the center, and all models and the center baseline outperformed the chance baseline.

A deep feature based saliency model (DeepFeat) is developed to leverage the understanding of the prediction of human fixations. Conventional saliency models often predict the human visual attention relying on few image cues. Although such models predict fixations on a variety of image complexities, their approaches are limited to the incorporated features. This chapter aims to utilize the deep features of convolutional neural networks by combining bottom-up and top-down saliency maps. The proposed framework is applied on deep features of three popular deep convolutional neural networks. Four evaluation metrics are exploited to evaluate the correspondence between the proposed framework and the ground-truth fixations over two datasets. The key findings of the results demonstrate that the deep features of pre-trained deep convolutional neural networks over the ImageNet dataset are strong predictors of the human fixation. The incorporation of bottom-up and top-down saliency maps outperforms the individual bottom-up and top-down implementations. Moreover, in comparison to nine saliency models including four state-of-the-art and five conventional saliency models, our proposed DeepFeat model outperforms the conventional saliency models over all four evaluation metrics.

Based on transfer learning, feature maps of deep convolutional neural networks (DCNNs) trained for object classification have been used to predict human gaze patterns. Such studies either fine-tune the DCNNs or use a transfer learning framework to learn the combination of such feature maps. Since the DeepFeat saliency model is a transfer-learning approach, extensive comparisons are conducted to investigate effects of feature maps on the predictions of the human gaze patterns using the DeepFeat saliency model framework. Four different implementations of the model have been used to create saliency maps, including a bottom-up implementation, a top-down implementation, and a combination of bottom-up and top-down implementations with and without the center bias. Feature maps of four pre-trained DCNNs are exploited using classical and class activation maps approaches. The performances of various saliency implementations are evaluated over four public datasets using three evaluation metrics. The results demonstrate that feature maps of the pre-trained DCNNs can be used to predict human gaze patterns. The incorporation of multiple levels of blurred and multi-scale feature maps improves the extraction of salient regions. Moreover, DCNNs pre-trained using the Places dataset provide more localized objects that can be beneficial to the top-down saliency maps. In addition, the incorporation of the center bias may boost the performance of some saliency implementations. Keywords:

Convolutional neural networks, feature maps, human fixation prediction, saliency map, transfer learning.

In another problem direction, a deep learning saliency framework is proposed to learn its weights from scratch. The model investigates a data generation protocol to create a large enough dataset to train a saliency model using random weights. The OSIE dataset is exploited to label a large number of patch samples as fixation or non-fixation. The generated data is used to train a saliency model using a residual learning framework. The proposed deep learning model is a modification of ResNet20 to fit for human visual attention prediction. Validity of the proposed framework is evaluated over five datasets including the original training dataset using four evaluation metrics. While the trained model slightly overfits, the preliminary results demonstrate the proposed hypothesis of data generation.

8.2 Future Work:

There is exciting works that have been conducted to compare human gaze patterns at different ages. However, majority of studies focus on adults' gaze patterns using foveated image processing to simulate the stimulus in the adult retina. Due to the rapid development of the biological structure of the eye in infants', foveal vision in infants haven't been simulated yet. In order to highlight the differences between infants and adults, formalization of foveal vision in infants is necessary to analyze infants saccades (velocity and magnitude) in addition to eye fixation.

Another area of future work would be validation of the DeepFeat saliency model over other datasets of different objective. Current saliency models are trained on relatively small datasets which may prune to overfitting over different eye fixation datasets such as fashion design, video games, etc. The ability of the DeepFeat model to highlight a human object is interesting to chapter, although the model exploit CNNs pre-trained over the ImageNet dataset which does not include the human as an object class to classify.

While transfer learning models of saliency exploit deep features from a variety of CNN layers, the feature selection task remains unclear. This dissertation demonstrates how a variety of feature selections behave over the DeepFeat model. In future work, one significant contribution would be the investigation of Deep features of CNNs pre-trained for other tasks than object classification. For example, face detection, image segmentation, and scene classification can provide deep feature that may be able to provide a better highlight to some attentive objects.

One area that requires an extensive amount of attention is data gathering. Although collection of a very large dataset of human eye fixation is extensively exhausting, the requirement of such dataset is necessary not only to provide more accurate predictions of human fixations, but also allows researchers to develop deep learning models that are trained from scratch, and therefore, the learned weights maybe a useful tool to leverage the understanding of features that are ensemble to guide the human attention. Moreover, the proposed model performance can be improved by adjusting the model hyper parameters, loss function, and reduce the distance between every two patches which leads to reduce the interpolation for resizing the resulting image to the original image size.

REFERENCES

[1] Mahdi, Ali, Matthew Schlesinger, Dima Amso, and Jun Qin. "Infants gaze pattern analyzing using contrast entropy minimization." In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 106-111. IEEE, 2015.

[2] Mahdi, Ali, Mei Su, Matthew Schlesinger, and Jun Qin. "A comparison study of saliency models for fixation prediction on infants and adults." *IEEE Transactions on Cognitive and Developmental Systems* 10, no. 3 (2018): 485-498

[3] Nothdurft, Hans-Christoph. "Salience of feature contrast." In *Neurobiology of attention*, pp. 233-239. Academic Press, 2005.

[4] Itti, Laurent, and Christof Koch. "Computational modelling of visual attention." *Nature reviews neuroscience* 2, no. 3 (2001): 194.

[5] Butko, Nicholas J., and Javier R. Movellan. "Optimal scanning for faster object detection." In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2751-2758. IEEE, 2009.

[6] Ehinger, Krista A., Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. "Modelling search for people in 900 scenes: A combined source model of eye guidance." *Visual cognition* 17, no. 6-7 (2009): 945-978.

[7] Mishra, Ajay K., and Yiannis Aloimonos. "Active segmentation." *International Journal of Humanoid Robotics* 6, no. 03 (2009): 361-386.

[8] Maki, Atsuto, Peter Nordlund, and Jan-Olof Eklundh. "Attentional scene segmentation: integrating depth and motion." *Computer Vision and Image Understanding* 78, no. 3 (2000): 351-373.

[9] Marchesotti, Luca, Claudio Cifarelli, and Gabriela Csurka. "A framework for visual saliency detection with applications to image thumbnailing." In *2009 IEEE 12th International Conference on Computer Vision*, pp. 2232-2239. IEEE, 2009.

[10] Suh, Bongwon, Haibin Ling, Benjamin B. Bederson, and David W. Jacobs. "Automatic thumbnail cropping and its effectiveness." In *Proceedings of the 16th annual ACM symposium on User interface software and technology*, pp. 95-104. ACM, 2003.

[11] Itti, Laurent. "Automatic foveation for video compression using a neurobiological model of visual attention." *IEEE Transactions on Image Processing* 13, no. 10 (2004): 1304-1318.

[12] Guo, Chenlei, and Liming Zhang. "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression." *IEEE transactions on image processing* 19, no. 1 (2010): 185-198.

[13] Mahadevan, Vijay, and Nuno Vasconcelos. "Saliency-based discriminant tracking." In *2009 IEEE conference on computer vision and pattern recognition*, pp. 1007-1013. IEEE, 2009.

[14] Frintrop, Simone, Erich Rome, and Henrik I. Christensen. "Computational visual attention systems and their cognitive foundations: A survey." *ACM Transactions on Applied Perception (TAP)* 7, no. 1 (2010): 6.

[15] Sugano, Yusuke, Yasuyuki Matsushita, and Yoichi Sato. "Appearance-based gaze estimation using visual saliency." *IEEE transactions on pattern analysis and machine intelligence* 35, no. 2 (2013): 329-341.

[16] Baluja, Shumeet, and Dean A. Pomerleau. "Expectation-based selective attention for visual monitoring and control of a robot vehicle." *Robotics and autonomous systems* 22, no. 3-4 (1997): 329-344.

[17] Ma, Qi, Liming Zhang, and Bin Wang. "New strategy for image and video quality assessment." *Journal of Electronic Imaging*19, no. 1 (2010): 011019.

[18] Ninassi, Alexandre, Olivier Le Meur, Patrick Le Callet, and Dominique Barba. "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric." In *2007 IEEE International Conference on Image Processing*, vol. 2, pp. II-169. IEEE, 2007.

[19] Rubinstein, Michael, Ariel Shamir, and Shai Avidan. "Improved seam carving for video retargeting." In *ACM transactions on graphics (TOG)*, vol. 27, no. 3, p. 16. ACM, 2008.

[20] James, W. "The principles of psychology, Vol. 2. NY, US: Henry Holt and Company." (1890).

[21] Corbetta, Maurizio, Francis M. Miezin, Susan Dobmeyer, Gordon L. Shulman, and Steven E. Petersen. "Attentional modulation of neural processing of shape, color, and velocity in humans." *Science* 248, no. 4962 (1990): 1556-1559.

[22] Treisman, Anne M., and Garry Gelade. "A feature-integration theory of attention." *Cognitive psychology* 12, no. 1 (1980): 97-136.

[23] Koch, Christof, and Shimon Ullman. "Shifts in selective visual attention: towards the underlying neural circuitry." In *Matters of intelligence*, pp. 115-141. Springer, Dordrecht, 1987.

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

[25] Itti, Laurent, and Pierre F. Baldi. "Bayesian surprise attracts human attention." In *Advances in neural information processing systems*, pp. 547-554. 2006.

[26] Bruce, Neil, and John Tsotsos. "Saliency based on information maximization." In *Advances in neural information processing systems*, pp. 155-162. 2006.

[27] Navalpakkam, Vidhya, and Laurent Itti. "Search goal tunes visual features optimally." *Neuron* 53, no. 4 (2007): 605-617.

[28] Cerf, Moran, Jonathan Harel, Wolfgang Einhäuser, and Christof Koch. "Predicting human gaze using low-level saliency combined with face detection." In *Advances in neural information processing systems*, pp. 241-248. 2008.

[29] Zhang, Lingyun, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. "SUN: A Bayesian framework for saliency using natural statistics." *Journal of vision* 8, no. 7 (2008): 32-32.

[30] Judd, Tilke, Krista Ehinger, Frédo Durand, and Antonio Torralba. "Learning to predict where humans look." In *2009 IEEE 12th international conference on computer vision*, pp. 2106-2113. IEEE, 2009.

[31] Liu, Tie, 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, no. 2 (2011): 353-367.

[32] Tian, Huawei, Yuming Fang, Yao Zhao, Weisi Lin, Rongrong Ni, and Zhenfeng Zhu. "Salient region detection by fusing bottom-up and top-down features extracted from a single image." *IEEE Transactions on Image processing* 23, no. 10 (2014): 4389-4398.

[33] Zhang, Jianming, and Stan Sclaroff. "Saliency detection: A Boolean map approach." In *Proceedings of the IEEE international conference on computer vision*, pp. 153-160. 2013.

[34] Zhang, Lihe, Chuan Yang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. "Ranking saliency." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 9 (2017): 1892-1904.

[35] Gao, Ke, Shouxun Lin, Yongdong Zhang, Sheng Tang, and Huamin Ren. "Attention model based sift keypoints filtration for image retrieval." In *Seventh IEEE/ACIS International Conference on Computer and Information Science (icis 2008)*, pp. 191-196. IEEE, 2008.

[36] Tsotsos, John K., Scan M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. "Modeling visual attention via selective tuning." *Artificial intelligence* 78, no. 1-2 (1995): 507-545.

[37] Privitera, Claudio M., and Lawrence W. Stark. "Algorithms for defining visual regions-of-interest: Comparison with eye fixations." *IEEE Transactions on Pattern Analysis & Machine Intelligence* 9 (2000): 970-982.

[38] Kadir, Timor, and Michael Brady. "Saliency, scale and image description." *International Journal of Computer Vision* 45, no. 2 (2001): 83-105.

[39] Oliva, Aude, and Antonio Torralba. "Modeling the shape of the scene: A holistic representation of the spatial envelope." *International journal of computer vision* 42, no. 3 (2001): 145-175.

[40] Lee, K., Hilary Buxton, and J. Feng. "Selective attention for cue-guided search using a spiking neural network." In *International Workshop on Attention and Performance in Computer Vision*, pp. 55-62. 2003.

[41] Itti, Laurent, Nitin Dhavale, and Frederic Pighin. "Realistic avatar eye and head animation using a neurobiological model of visual attention." In *Applications and Science of Neural Networks, Fuzzy Systems, and Evolutionary Computation VI*, vol. 5200, pp. 64-79. International Society for Optics and Photonics, 2003.

[42] Kootstra, Gert, Arco Nederveen, and Bart De Boer. "Paying attention to symmetry." In *British Machine Vision Conference (BMVC2008)*, pp. 1115-1125. The British Machine Vision Association and Society for Pattern Recognition, 2008.

[43] Ehinger, Krista A., Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. "Modelling search for people in 900 scenes: A combined source model of eye guidance." *Visual cognition* 17, no. 6-7 (2009): 945-978.

[44] Valenti, Roberto, Nicu Sebe, and Theo Gevers. "Image saliency by isocentric curvedness and color." In *2009 IEEE 12th International Conference on Computer Vision (ICCV)*, pp. 2185-2192. IEEE, 2009.

[45] Li, Jia, Yonghong Tian, Tiejun Huang, and Wen Gao. "Probabilistic multi-task learning for visual saliency estimation in video." *International journal of computer vision* 90, no. 2 (2010): 150-165.

[46] Lang, Congyan, Tam V. Nguyen, Harish Katti, Karthik Yadati, Mohan Kankanhalli, and Shuicheng Yan. "Depth matters: Influence of depth cues on visual saliency." In *Computer vision–ECCV 2012*, pp. 101-115. Springer, Berlin, Heidelberg, 2012.

[47] Liu, Zhi, Olivier Le Meur, Shuhua Luo, and Liquan Shen. "Saliency detection using regional histograms." *Optics letters*38, no. 5 (2013): 700-702.

[48] Mahdi, Ali, Jun Qin, and Garth Crosby. "DeepFeat: A Bottom-Up and Top-Down Saliency Model Based on Deep Features of Convolutional Neural Nets." *IEEE Transactions on Cognitive and Developmental Systems* (2019).

[49] Chang, Kai-Yueh, Tyng-Luh Liu, Hwann-Tzong Chen, and Shang-Hong Lai. "Fusing generic objectness and visual saliency for salient object detection." In *2011 International Conference on Computer Vision*, pp. 914-921. IEEE, 2011.

[50] Zhu, Wangjiang, Shuang Liang, Yichen Wei, and Jian Sun. "Saliency optimization from robust background detection." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2814-2821. 2014.

[51] Itti, Laurent, Geraint Rees, and John K. Tsotsos, eds. *Neurobiology of attention*. Elsevier, 2005.

[52] Borji, Ali, and Laurent Itti. "State-of-the-art in visual attention modeling." *IEEE transactions on pattern analysis and machine intelligence* 35, no. 1 (2013): 185-207.

[53] Filipe, Sílvio, and Luís A. Alexandre. "From the human visual system to the computational models of visual attention: a survey." *Artificial Intelligence Review* 39, no. 1 (2013): 1-47.

[54] Zhang, Lingyun, Matthew H. Tong, and Garrison W. Cottrell. "SUNDAy: Saliency using natural statistics for dynamic analysis of scenes." In *Proceedings of the 31st annual cognitive science conference*, pp. 2944-2949. Cambridge, MA: AAAI Press, 2009.

[55] Xie, Yulin, Huchuan Lu, and Ming-Hsuan Yang. "Bayesian saliency via low and mid level cues." *IEEE Transactions on Image Processing* 22, no. 5 (2013): 1689-1698.

[56] Lu, Huchuan, Xiaohui Li, Lihe Zhang, Xiang Ruan, and Ming-Hsuan Yang. "Dense and sparse reconstruction error based saliency descriptor." *IEEE Transactions on Image Processing*25, no. 4 (2016): 1592-1603.

[57] Jianyong, Lv, Tang Zhenmin, and Xu Wei. "Improved Bayesian saliency detection based on bing and graph model." *Open Cybernetics & Systemics Journal* 9 (2015): 648-656.

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

[59] Frintrop, Simone. *VOCUS: A visual attention system for object detection and goal-directed search*. Vol. 3899. Springer, 2006.

[60] Walther, Dirk, and Christof Koch. "Modeling attention to salient proto-objects." *Neural networks* 19, no. 9 (2006): 1395-1407

[61] Itti, Laurent, and Christof Koch. "A saliency-based search mechanism for overt and covert shifts of visual attention." *Vision research* 40, no. 10-12 (2000): 1489-1506.

[62] Le Meur, Olivier, Patrick Le Callet, Dominique Barba, and Dominique Thoreau. "A coherent computational approach to model bottom-up visual attention." *IEEE transactions on pattern analysis and machine intelligence* 28, no. 5 (2006): 802-817.

[63] Le Meur, Olivier, Patrick Le Callet, and Dominique Barba. "Predicting visual fixations on video based on low-level visual features." *Vision research* 47, no. 19 (2007): 2483-2498.

[64] Marat, Sophie, Tien Ho Phuoc, Lionel Granjon, Nathalie Guyader, Denis Pellerin, and Anne Guérin-Dugué. "Modelling spatio-temporal saliency to predict gaze direction for short videos." *International journal of computer vision* 82, no. 3 (2009): 231.

[65] Murray, Naila, Maria Vanrell, Xavier Otazu, and C. Alejandro Parraga. "Saliency estimation using a non-parametric low-level vision model." In *CVPR 2011*, pp. 433-440. IEEE, 2011.

[66] Gao, Dashan, Sunhyoung Han, and Nuno Vasconcelos. "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31, no. 6 (2009): 989-1005.

[67] Mahadevan, Vijay, and Nuno Vasconcelos. "Spatiotemporal saliency in dynamic scenes." *IEEE transactions on pattern analysis and machine intelligence* 32, no. 1 (2010): 171-177

[68] Guo, Chenlei, and Liming Zhang. "An attention selection model with visual memory and online learning." In *2007 International Joint Conference on Neural Networks*, pp. 1295-1301. IEEE, 2007.

[69] Gu, Erdan, Jingbin Wang, and Norman I. Badler. "Generating sequence of eye fixations using decision-theoretic attention model." In *International Workshop on Attention in Cognitive Systems*, pp. 277-292. Springer, Berlin, Heidelberg, 2007.

[70] Gao, Dashan, and Nuno Vasconcelos. "Discriminant saliency for visual recognition from cluttered scenes." In *Advances in neural information processing systems*, pp. 481-488. 2005.

[71] Hou, Xiaodi, and Liqing Zhang. "Saliency detection: A spectral residual approach." In *2007 IEEE Conference on computer vision and pattern recognition*, pp. 1-8. IEEE, 2007.

[72] Wang, Zheshen, and Baoxin Li. "A two-stage approach to saliency detection in images." In *2008 IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 965-968. IEEE, 2008.

[73] Li, Yin, Yue Zhou, Junchi Yan, Zhibin Niu, and Jie Yang. "Visual saliency based on conditional entropy." In *Asian Conference on Computer Vision*, pp. 246-257. Springer, Berlin, Heidelberg, 2009.

[74] Guo, Chenlei, Qi Ma, and Liming Zhang. "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform." In *2008 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8. IEEE, 2008.

[75] Achanta, Radhakrishna, Sheila Hemami, Francisco Estrada, and Sabine Süsstrunk. "Frequency-tuned salient region detection." In *IEEE international conference on computer vision and pattern recognition (CVPR 2009)*, no. CONF, pp. 1597-1604. 2009.

[76] Bian, Peng, and Liming Zhang. "Biological plausibility of spectral domain approach for spatiotemporal visual saliency." In *International conference on neural information processing*, pp. 251-258. Springer, Berlin, Heidelberg, 2008.

[77] Li, Jian, Martin D. Levine, Xiangjing An, Xin Xu, and Hangen He. "Visual saliency based on scale-space analysis in the frequency domain." *IEEE transactions on pattern analysis and machine intelligence* 35, no. 4 (2013): 996-1010.

[78] Xiao, Limei, Ce Li, Zhijia Hu, and Zhengrong Pan. "Multi-scale spectrum visual saliency perception via hypercomplex DCT." In *International Conference on Intelligent Computing*, pp. 645-655. Springer, Cham, 2016.

[79] Salah, Albert Ali, Ethem Alpaydin, and Lale Akarun. "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24, no. 3 (2002): 420-425.

[80] Rao, Rajesh PN. "Bayesian inference and attentional modulation in the visual cortex." *Neuroreport* 16, no. 16 (2005): 1843-1848.

[81] Liu, Tie, Jian Sun, Nan-Ning Zheng, Xiaoou Tang, and Heung-Yeung Shum. "Learning to Detect A Salient Object." In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8. IEEE, 2007.

[82] Yang, Chuan, Lihe Zhang, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. "Saliency detection via graph-based manifold ranking." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3166-3173. 2013.

[83] Huang, Ling, Songguang Tang, Jiani Hu, and Weihong Deng. "Saliency region detection via graph model and statistical learning." In *Chinese Conference on Pattern Recognition*, pp. 3-13. Springer, Singapore, 2016.

[84] Zhang, Jinxia, Krista A. Ehinger, Haikun Wei, Kanjian Zhang, and Jingyu Yang. "A novel graph-based optimization framework for salient object detection." *Pattern Recognition* 64 (2017): 39-50.

[85] Renninger, Laura W., James M. Coughlan, Preeti Verghese, and Jitendra Malik. "An information maximization model of eye movements." In *Advances in neural information processing systems*, pp. 1121-1128. 2005.

[86] Seo, Hae Jong, and Peyman Milanfar. "Static and space-time visual saliency detection by self-resemblance." *Journal of vision* 9, no. 12 (2009): 15-15.

[87] Bruce, Neil DB, and John K. Tsotsos. "Saliency, attention, and visual search: An information theoretic approach." *Journal of vision* 9, no. 3 (2009): 5-5.

[88] Wang, Wei, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. "Simulating human saccadic scanpaths on natural images." In *CVPR 2011*, pp. 441-448. IEEE, 2011.

[89] Klein, Dominik A., and Simone Frintrop. "Center-surround divergence of feature statistics for salient object detection." In *2011 International Conference on Computer Vision*, pp. 2214-2219. IEEE, 2011.

[90] Riche, Nicolas, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. "Rare: A new bottom-up saliency model." In *2012 19th IEEE International Conference on Image Processing*, pp. 641-644. IEEE, 2012.

[91] Peters, Robert J., and Laurent Itti. "Beyond bottom-up: Incorporating task-dependent influences into a computational model of spatial attention." In *2007 IEEE conference on computer vision and pattern recognition*, pp. 1-8. IEEE, 2007.

[92] Kienzle, Wolf, Matthias O. Franz, Bernhard Schölkopf, and Felix A. Wichmann. "Center-surround patterns emerge as optimal predictors for human saccade targets." *Journal of vision* 9, no. 5 (2009): 7-7.

[93] Liu, Nian, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. "Predicting eye fixations using convolutional neural networks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 362-370. 2015.

[94] Zhao, Rui, Wanli Ouyang, Hongsheng Li, and Xiaogang Wang. "Saliency detection by multi-context deep learning." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1265-1274. 2015.

[95] Cornia, Marcella, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. "A deep multi-level network for saliency prediction." In *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp. 3488-3493. IEEE, 2016.

[96] Vig, Eleonora, Michael Dorr, and David Cox. "Large-scale optimization of hierarchical features for saliency prediction in natural images." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2798-2805. 2014.

[97] Kümmerer, Matthias, Lucas Theis, and Matthias Bethge. "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet." *arXiv preprint arXiv:1411.1045* (2014).

[98] Kummerer, Matthias, Thomas SA Wallis, Leon A. Gatys, and Matthias Bethge. "Understanding low-and high-level contributions to fixation prediction." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 4789-4798. 2017.

[99] Huang, Xun, Chengyao Shen, Xavier Boix, and Qi Zhao. "Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks." In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 262-270. 2015.

[100] Kruthiventi, Srinivas SS, Kumar Ayush, and R. Venkatesh Babu. "Deepfix: A fully convolutional neural network for predicting human eye fixations." *IEEE Transactions on Image Processing* 26, no. 9 (2017): 4446-4456.

[101] Liu, Nian, and Junwei Han. "A deep spatial contextual long-term recurrent convolutional network for saliency detection." *IEEE Transactions on Image Processing* 27, no. 7 (2018): 3264-3274.

[102] Jetley, Saumya, Naila Murray, and Eleonora Vig. "End-to-end saliency mapping via probability distribution prediction." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5753-5761. 2016.

[103] Erdem, Erkut, and Aykut Erdem. "Visual saliency estimation by nonlinearly integrating features using region covariances." *Journal of vision* 13, no. 4 (2013): 11-11.

[104] Liu, Zhi, Wenbin Zou, and Olivier Le Meur. "Saliency tree: A novel saliency detection framework." *IEEE Transactions on Image Processing* 23, no. 5 (2014): 1937-1952.

[105] Schlesinger, Matthew, and Dima Amso. "Image free-viewing as intrinsically-motivated exploration: estimating the learnability of center-of-gaze image samples in infants and adults." *Frontiers in psychology* 4 (2013): 802.

[106] Koehler, Kathryn, Fei Guo, Sheng Zhang, and Miguel P. Eckstein. "What do saliency models predict?." *Journal of vision* 14, no. 3 (2014): 14-14.

[107] Kootstra, Gert, Bart de Boer, and Lambert RB Schomaker. "Predicting eye fixations on complex visual stimuli using local symmetry." *Cognitive computation* 3, no. 1 (2011): 223-240.

[108] Xu, Juan, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. "Predicting human gaze beyond pixels." *Journal of vision* 14, no. 1 (2014): 28-28.

[109] Bruce, Neil, and John Tsotsos. "Attention based on information maximization." *Journal of Vision* 7, no. 9 (2007): 950-950.

[110] Bylinskii, Zoya, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. "What do different evaluation metrics tell us about saliency models?." *IEEE transactions on pattern analysis and machine intelligence* 41, no. 3 (2019): 740-757.

[111] Borji, Ali, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. "Salient object detection: A benchmark." *IEEE transactions on image processing* 24, no. 12 (2015): 5706-5722.

[112] Fukushima, Kunihiko. "Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position." *Biological cybernetics* 36, no. 4 (1980): 193-202.

[113] LeCun, Yann, Bernhard Boser, John S. Denker, Donnie Henderson, Richard E. Howard, Wayne Hubbard, and Lawrence D. Jackel. "Backpropagation applied to handwritten zip code recognition." *Neural computation* 1, no. 4 (1989): 541-551.

[114] Krizhevsky, Alex, Ilya Sutskever, and Geoffrey E. Hinton. "Imagenet classification with deep convolutional neural networks." In *Advances in neural information processing systems*, pp. 1097-1105. 2012.

[115] Chollet, François. "Xception: Deep learning with depthwise separable convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1251-1258. 2017.

[116] Szegedy, Christian, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. "Rethinking the inception architecture for computer vision." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2818-2826. 2016.

[117] Szegedy, Christian, Sergey Ioffe, Vincent Vanhoucke, and Alexander A. Alemi. "Inception-v4, inception-resnet and the impact of residual connections on learning." In *Thirty-First AAAI Conference on Artificial Intelligence*. 2017.

[118] Howard, Andrew G., Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. "Mobilenets: Efficient convolutional neural networks for mobile vision applications." *arXiv preprint arXiv:1704.04861* (2017).

[119] Huang, Gao, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q. Weinberger. "Densely connected convolutional networks." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4700-4708. 2017.

[120] Zoph, Barret, Vijay Vasudevan, Jonathon Shlens, and Quoc V. Le. "Learning transferable architectures for scalable image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 8697-8710. 2018.

[121] Sandler, Mark, Andrew Howard, Menglong Zhu, Andrey Zhmoginov, and Liang-Chieh Chen. "Mobilenetv2: Inverted residuals and linear bottlenecks." In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4510-4520. 2018.

[122] Simonyan, Karen, and Andrew Zisserman. "Very deep convolutional networks for large-scale image recognition." *arXiv preprint arXiv:1409.1556* (2014).

[123] Szegedy, Christian, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. "Going deeper with convolutions." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1-9. 2015.

[124] He, Kaiming, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. "Deep residual learning for image recognition." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770-778. 2016.

[125] Gao, Dashan, Vijay Mahadevan, and Nuno Vasconcelos. "On the plausibility of the discriminant center-surround hypothesis for visual saliency." *Journal of vision* 8, no. 7 (2008): 13-13.

[126] Zhao, Qi, and Christof Koch. "Learning a saliency map using fixated locations in natural scenes." *Journal of vision* 11, no. 3 (2011): 9-9.

[127] Bylinskii, Zoya, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. "Mit saliency benchmark." (2015): 402-409.

[128] Borji, Ali, Hamed R. Tavakoli, Dicky N. Sihite, and Laurent Itti. "Analysis of scores, datasets, and models in visual saliency prediction." In *Proceedings of the IEEE international conference on computer vision*, pp. 921-928. 2013.

[129] Borji, Ali, Dicky N. Sihite, and Laurent Itti. "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study." *IEEE Transactions on Image Processing* 22, no. 1 (2013): 55-69.

[130] Judd, Tilke, Frédo Durand, and Antonio Torralba. "A benchmark of computational models of saliency to predict human fixations." (2012).

[131] Schlesinger, Matthew. "Investigating the origins of intrinsic motivation in human infants." In *Intrinsically motivated learning in natural and artificial systems*, pp. 367-392. Springer, Berlin, Heidelberg, 2013.

[132] Amso, Dima, and Scott P. Johnson. "Development of visual selection in 3-to 9-month-olds: Evidence from saccades to previously ignored locations." *Infancy* 13, no. 6 (2008): 675-686.

[133] Dixon, Matthew L., Philip David Zelazo, and Eve De Rosa. "Evidence for intact memory□guided attention in school□aged children." *Developmental Science* 13, no. 1 (2010): 161-169.

[134] Amso, Dima, and Gaia Scerif. "The attentive brain: insights from developmental cognitive neuroscience." *Nature Reviews Neuroscience* 16, no. 10 (2015): 606.

[135] Schlesinger, Matthew, Scott P. Johnson, and Dima Amso. "Prediction-learning in infants as a mechanism for gaze control during object exploration." *Frontiers in psychology* 5 (2014): 441.

[136] Hou, Xiaodi, and Liqing Zhang. "Dynamic visual attention: Searching for coding length increments." In *Advances in neural information processing systems*, pp. 681-688. 2009.

[137] Hou, Xiaodi, and Liqing Zhang. "Saliency detection: A spectral residual approach." In *2007 IEEE Conference on computer vision and pattern recognition*, pp. 1-8. IEEE, 2007.

[138] Jiang, Huaizu, Jingdong Wang, Zejian Yuan, Tie Liu, Nanning Zheng, and Shipeng Li. "Automatic salient object segmentation based on context and shape prior." In *BMVC*, vol. 6, no. 7, p. 9. 2011.

[139] Garcia-Diaz, Antón, Xosé R. Fdez-Vidal, Xosé M. Pardo, and Raquel Dosil. "Decorrelation and distinctiveness provide with human-like saliency." In *International Conference on Advanced Concepts for Intelligent Vision Systems*, pp. 343-354. Springer, Berlin, Heidelberg, 2009.

[140] Le Meur, Olivier, and Thierry Baccino. "Methods for comparing scanpaths and saliency maps: strengths and weaknesses." *Behavior research methods* 45, no. 1 (2013): 251-266.

[141] Amso, Dima, Sara Haas, and Julie Markant. "An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes." *PLoS One* 9, no. 1 (2014): e85701.

[142] Parkhurst, Derrick J., and Ernst Niebur. "Scene content selected by active vision." *Spatial vision* 16, no. 2 (2003): 125-154.

[143] Tatler, Benjamin W., Roland J. Baddeley, and Iain D. Gilchrist. "Visual correlates of fixation selection: Effects of scale and time." *Vision research* 45, no. 5 (2005): 643-659.

[144] Tatler, Benjamin W. "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions." *Journal of vision* 7, no. 14 (2007): 4-4.

[145] Leboran, Victor, Anton Garcia-Diaz, Xosé R. Fdez-Vidal, and Xosé M. Pardo. "Dynamic whitening saliency." *IEEE transactions on pattern analysis and machine intelligence* 39, no. 5 (2017): 893-907.

[146] Oliva, Aude, Antonio Torralba, Monica S. Castelhano, and John M. Henderson. "Top-down control of visual attention in object detection." In *Proceedings 2003 International Conference on Image Processing (Cat. No. 03CH37429)*, vol. 1, pp. I-253. IEEE, 2003.

[147] Borji, Ali, Majid Nili Ahmadabadi, Babak Nadjar Araabi, and Mandana Hamidi. "Online learning of task-driven object-based visual attention control." *Image and Vision Computing* 28, no. 7 (2010): 1130-1145.

[148] Wang, Jingwei, Ali Borji, C-C. Jay Kuo, and Laurent Itti. "Learning a combined model of visual saliency for fixation prediction." *IEEE Transactions on Image Processing* 25, no. 4 (2016): 1566-1579.

[149] Zeiler, Matthew D., and Rob Fergus. "Visualizing and understanding convolutional networks." In *European conference on computer vision*, pp. 818-833. Springer, Cham, 2014.

[150] Schmidhuber, Jürgen. "Deep learning in neural networks: An overview." *Neural networks* 61 (2015): 85-117.

[151] Saxena, Ashutosh, Justin Driemeyer, and Andrew Y. Ng. "Robotic grasping of novel objects using vision." *The International Journal of Robotics Research* 27, no. 2 (2008): 157-173.

[152] Yamazaki, Kimitoshi, Ryohei Ueda, Shunichi Nozawa, Mitsuharu Kojima, Kei Okada, Kiyoshi Matsumoto, Masaru Ishikawa, Isao Shimoyama, and Masayuki Inaba. "Home-assistant robot for an aging society." *Proceedings of the IEEE*100, no. 8 (2012): 2429-2441.

[153] Russakovsky, Olga, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang et al. "Imagenet large scale visual recognition challenge." *International journal of computer vision* 115, no. 3 (2015): 211-252.

[154] Zhou, Bolei, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. "Learning deep features for discriminative localization." In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2921-2929. 2016.

[155] Vedaldi, Andrea, and Karel Lenc. "Matconvnet: Convolutional neural networks for matlab." In *Proceedings of the 23rd ACM international conference on Multimedia*, pp. 689-692. ACM, 2015.

[156] Henderson, John M., and Andrew Hollingworth. "High-level scene perception." *Annual review of psychology* 50, no. 1 (1999): 243-271.

[157] Parkhurst, Derrick J., and Ernst Niebur. "Scene content selected by active vision." *Spatial vision* 16, no. 2 (2003): 125-154.

[158] Tseng, Po-He, Ran Carmi, Ian GM Cameron, Douglas P. Munoz, and Laurent Itti. "Quantifying center bias of observers in free viewing of dynamic natural scenes." *Journal of vision* 9, no. 7 (2009): 4-4.

[159] Wang, Wenguan, and Jianbing Shen. "Deep visual attention prediction." *IEEE Transactions on Image Processing* 27, no. 5 (2018): 2368-2378.

[160] Tavakoli, Hamed R., Ali Borji, Jorma Laaksonen, and Esa Rahtu. "Exploiting inter-image similarity and ensemble of extreme learners for fixation prediction using deep features." *Neurocomputing* 244 (2017): 10-18.

[161] Riche, Nicolas, Matei Mancas, Matthieu Duvinage, Makiese Mibulumukini, Bernard Gosselin, and Thierry Dutoit. "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis." *Signal Processing: Image Communication* 28, no. 6 (2013): 642-658.

[162] Pan, Junting, and Xavier Giró-i-Nieto. "End-to-end convolutional network for saliency prediction." *arXiv preprint arXiv:1507.01422* (2015).

[163] Gide, Milind S., and Lina J. Karam. "A locally weighted fixation density-based metric for assessing the quality of visual saliency predictions." *IEEE Transactions on Image Processing*25, no. 8 (2016): 3852-3861.

[164] Kruthiventi, Srinivas SS, Kumar Ayush, and R. Venkatesh Babu. "Deepfix: A fully convolutional neural network for predicting human eye fixations." *IEEE Transactions on Image Processing* 26, no. 9 (2017): 4446-4456.

[165] Pan, Junting, Cristian Canton Ferrer, 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).

[166] Cornia, Marcella, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. "Predicting human eye fixations via an lstm-based saliency attentive model." *IEEE Transactions on Image Processing* 27, no. 10 (2018): 5142-5154

[167] Mahdi, Ali, and Jun Qin. "Bottom up saliency evaluation via deep features of state-of-the-art convolutional neural networks." In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 247-250. IEEE, 2018.

[168] Jia, Yangqing, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. "Caffe: Convolutional architecture for fast feature embedding." In *Proceedings of the 22nd ACM international conference on Multimedia*, pp. 675-678. ACM, 2014.

[169] Judd, Tilke, Frédo Durand, and Antonio Torralba. "A benchmark of computational models of saliency to predict human fixations." (2012).

[170] Borji, Ali, and Laurent Itti. "Exploiting local and global patch rarities for saliency detection." In *2012 IEEE conference on computer vision and pattern recognition*, pp. 478-485. IEEE, 2012.

[171] Cheng, Ming-Ming, Niloy J. Mitra, Xiaolei Huang, Philip HS Torr, and Shi-Min Hu. "Global contrast based salient region detection." *IEEE Transactions on Pattern Analysis and Machine Intelligence* 37, no. 3 (2015): 569-582.

[172] Geisler, Wilson S. "Visual perception and the statistical properties of natural scenes." *Annu. Rev. Psychol.* 59 (2008): 167-192.

[173] Odermatt, Benjamin, Anton Nikolaev, and Leon Lagnado. "Encoding of luminance and contrast by linear and nonlinear synapses in the retina." *Neuron* 73, no. 4 (2012): 758-773.

[174] Oesch, Nicholas W., and Jeffrey S. Diamond. "Ribbon synapses compute temporal contrast and encode luminance in retinal rod bipolar cells." *Nature neuroscience* 14, no. 12 (2011): 1555.

[175] Koch, Kristin, Judith McLean, Ronen Segev, Michael A. Freed, Michael J. Berry II, Vijay Balasubramanian, and Peter Sterling. "How much the eye tells the brain." *Current Biology* 16, no. 14 (2006): 1428-1434.

[176] Egeth, Howard E., and Steven Yantis. "Visual attention: Control, representation, and time course." *Annual review of psychology* 48, no. 1 (1997): 269-297.

[177] Oyama, Taiki, and Takao Yamanaka. "Influence of image classification accuracy on saliency map estimation." *CAAI Transactions on Intelligence Technology* 3, no. 3 (2018): 140-152.

VITA

Graduate School
Southern Illinois University

Ali M. Mahdi

amahdi@dcscorp.com
ali.majeed.mahdi@gmail.com

Al-Mustansiriya University, IRAQ
Bachelor of Science, Computer Engineering, July 2007

Southern Illinois University Carbondale
Master of Science, Electrical & Computer Engineering, May 2013

Special Honors and Awards:

2018 SIU Graduate Professional Student Council Research Award.
2018 IEEE BHI Conference Student Travel Award.
2018 SIU Graduate Professional Student Council Career Development Award.
2006 Al-Mustansiriya University Computer Maintenance Competition Award.
2006 Al-Mustansiriya University Computer Architecture Competition Award.

Dissertation Paper Title:

Visual Saliency Analysis, Prediction, and Visualization: A Deep Learning Perspective

Major Professor: Jun Qin

Publications:

Mahdi, Ali, and Jun Qin. "Evaluation of Bottom Up Saliency Models Using Deep Features Pre-trained by Convolutional Neural Networks." *Journal of Electronic Imaging* (2019).

Mahdi, Ali, Jun Qin, and Garth Crosby. "DeepFeat: A Bottom-Up and Top-Down Saliency Model Based on Deep Features of Convolutional Neural Nets." *IEEE Transactions on Cognitive and Developmental Systems* (2019).

Mahdi, Ali, Mei Su, Matthew Schlesinger, and Jun Qin. "A comparison study of saliency models for fixation prediction on infants and adults." *IEEE Transactions on Cognitive and Developmental Systems* 10, no. 3 (2018): 485-498.

Mahdi, Ali, and Jun Qin. "Bottom up saliency evaluation via deep features of state-of-the-art convolutional neural networks." In *2018 IEEE EMBS International Conference on Biomedical & Health Informatics (BHI)*, pp. 247-250. IEEE, 2018.

Sun, Pengfei, Ali Mahdi, Jianhong Xu, and Jun Qin. "Speech enhancement in spectral envelop and details subspaces." *Speech Communication* 101 (2018): 57-69.

Mahdi, Ali, Matthew Schlesinger, Dima Amso, and Jun Qin. "Infants gaze pattern analyzing using contrast entropy minimization." In *2015 Joint IEEE International Conference on Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, pp. 106-111. IEEE, 2015.

Qin, J., Y. Jiang, and A. Mahdi. "Recent developments on noise induced hearing loss for military and industrial applications." *Biosensors Journal* 3 (2014): e101.

Mahdi, Ali Majeed. "Validation of the Touching Corn Separation Using Improved Convex Segmentation." *Thesis,Southern Illinois University Carbondale* (2013).