VISUAL SALIENCY ANALYSIS ON FASHION IMAGES USING IMAGE PROCESSING AND DEEP LEARNING APPROACHES

by

Aashish Neupane

B.E., Tribhuvan University, Pulchowk Campus, 2017

A Thesis

Submitted in Partial Fulfillment of the Requirements for the

Master of Science Degree

Department of Biomedical Engineering

In the Graduate School

Southern Illinois University Carbondale

July 2020

THESIS APPROVAL

VISUAL SALIENCY ANALYSIS ON FASHION IMAGES USING IMAGE PROCESSING AND DEEP LEARNING APPROACHES

by

Aashish Neupane

A Thesis Submitted in Partial

Fulfillment of the Requirements for the Degree of Master of Science

In Biomedical Engineering

Approved by:

Dr. Jun Qin (Advisor)

Dr. Ada Chen

Dr. Seung Hee-Lee

Graduate School

Southern Illinois University Carbondale

July 33, 2020

# ABSTRACT

AASHISH NEUPANE, for the Master of Science degree in BIOMEDICAL ENGINEERING, presented on July 35, 2020, at Southern Illinois University Carbondale.

TITLE: VISUAL SALIENCY ANALYSIS ON FASHION IMAGES USING IMAGE PROCESSING AND DEEP LEARNING APPROACHES.

MAJOR PROFESSOR: Dr. Jun Qin

State-of-art computer vision technologies have been applied in fashion in multiple ways, and saliency modeling is one of those applications. In computer vision, a saliency map is a 2D topological map which indicates the probabilistic distribution of visual attention priorities. This study is focusing on analysis of the visual saliency on fashion images using multiple saliency models, evaluated by several evaluation metrics. A human subject study has been conducted to collect people’s visual attention on 75 fashion images. Binary ground-truth fixation maps for these images have been created based on the experimentally collected visual attention data using Gaussian blurring function. Saliency maps for these 75 fashion images were generated using multiple conventional saliency models as well as deep feature-based state-of-art models. DeepFeat has been studied extensively, with 44 sets of saliency maps, exploiting the features extracted from GoogLeNet and ResNet50. Seven other saliency models have also been utilized to predict saliency maps on these images. The results were compared over 5 evaluation metrics – AUC, CC, KL Divergence, NSS and SIM. The performance of all 8 saliency models on prediction of visual attention on fashion images over all five metrics were comparable to the benchmarked scores. Furthermore, the models perform well consistently over multiple evaluation metrics, thus indicating that saliency models could in fact be applied to effectively predict salient regions in random fashion advertisement images.

# ACKNOWLEDGEMENTS

I am very fortunate enough to complete my thesis. This thesis has been completed with the complete and regular guidance of Dr. Jun Qin, my academic and thesis advisor. It is my honor to acknowledge my sincere and deep thanks to my professor who guided and motivated me at all times. My acknowledgement is to my committee members Dr. Seung Hee-Lee and Dr. Ada Chen for giving their important feedback and sharing some knowledge to complete this thesis. Thanks to my friend Ali for helping me learn so much. At the end, this thesis wouldn’t have been completed without the support of my family and friends.

# DEDICATION

To my late Grandmother, who was a symbol of courage and love for as long as she lived.

TABLE OF CONTENTS

CHAPTER PAGE

[ABSTRACT i](#_Toc47024658)

[ACKNOWLEDGEMENTS ii](#_Toc47024659)

[DEDICATION iii](#_Toc47024660)

[LIST OF FIGURES vii](#_Toc47024661)

[CHAPTER 1 1](#_Toc47024662)

[INTRODUCTION 1](#_Toc47024663)

[1.1 Saliency Modeling on Visual Attention 1](#_Toc47024664)

[1.2 Applications of State-of-Art Computer Vision in Fashion Design 6](#_Toc47024665)

[1.3 Saliency Models for Fashion Images 6](#_Toc47024666)

[1.4 Motivation and Contribution of Our Study 7](#_Toc47024667)

[1.5 Outline of this Document 8](#_Toc47024668)

[CHAPTER 2 9](#_Toc47024669)

[BACKGROUND AND TECHNNOLOGY REVIEW 9](#_Toc47024670)

[2.1 Literature on Saliency Models 9](#_Toc47024672)

[2.2 Concepts of Softmax Activation and CAM (Class Activation Map) 11](#_Toc47024673)

[2.3 Saliency Models Used in Our Experiment 12](#_Toc47024674)

[2.3.1 DeepFeat 13](#_Toc47024675)

[2.3.2 LSTM-Based Saliency Attentive Model (SAM) 15](#_Toc47024676)

[2.3.3 Ensembles of Deep Networks (eDN) 16](#_Toc47024677)

[2.3.4 Itti Model 16](#_Toc47024678)

[2.3.5 Graph based visual saliency model (GBVS) 16](#_Toc47024679)

[2.3.6 CovSal 16](#_Toc47024680)

[2.3.7 Fixation Prediction using Unsupervised Hierarchical Models (UHF) 17](#_Toc47024681)

[2.3.8 Weighted Maximum Phase Alignment (WMAP) Model 17](#_Toc47024682)

[2.4 Evaluation Metrics for Saliency Maps 18](#_Toc47024683)

[2.4.1 ROC 18](#_Toc47024684)

[2.4.2 AUC 18](#_Toc47024685)

[2.4.3 Pearson Correlation Coefficient (CC) 19](#_Toc47024686)

[2.4.4 Normalized Scanpath Saliency (NSS) 19](#_Toc47024687)

[2.4.5 Kullback-Leibler Divergence (KL) 20](#_Toc47024688)

[2.4.6 Similarity (SIM) 20](#_Toc47024689)

[CHAPTER 3 22](#_Toc47024690)

[EXPERIMENT SETUP, METHODS AND FRAMEWORK 22](#_Toc47024691)

[3.1 Psychophysics Toolbox Version 3 (PTB-3) 22](#_Toc47024693)

[3.2 Human Subject Data Collection 22](#_Toc47024694)

[3.3 Creation of Ground-Truth Fixation points and Ground-Truth Fixation Maps 23](#_Toc47024695)

[3.4 Running Saliency Models 23](#_Toc47024696)

[3.4.1 Running DeepFeat 24](#_Toc47024697)

[3.5 Running Other Saliency Models 27](#_Toc47024698)

[3.6 Calculation of Evaluation Metrics 27](#_Toc47024699)

[3.6.1 Specifics of ROC and AUC Calculations 28](#_Toc47024700)

[CHAPTER 4 29](#_Toc47024701)

[RESULT AND DISCUSSION 29](#_Toc47024702)

[4.1 Performance of DeepFeat with GoogLeNet and ResNet50 29](#_Toc47024704)

[4.2 Performance of DeepFeat and Seven Other Saliency Models 33](#_Toc47024705)

[4.3 Visualization of Evaluation Metrics 37](#_Toc47024706)

[CHAPTER 5 40](#_Toc47024707)

[CONCLUSION, LIMITATIONS AND FUTURE DIRECTIONS 40](#_Toc47024708)

[5.1 Conclusion 40](#_Toc47024710)

[5.2 Limitations and Future Directions 40](#_Toc47024711)

[REFERENCES 42](#_Toc47024712)

[VITA 48](#_Toc47024713)

# LIST OF FIGURES

FIGURE PAGE

[Figure 1: Architecture of Itti Model [1] used in obtaining saliency map in our study 3](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014111)

[Figure 2: Example of saliency maps created by Itti saliency model. 5](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014112)

[Figure 3: (Taken from Zhou et al. [2]) : Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class. 12](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014113)

[Figure 4 Architecture of DeepFeat presented as a Block Diagram 13](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014114)

[Figure 5: Architecture of a single Inception module 26](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014115)

[Figure 6: AUC Borji and AUC Judd scores of saliency maps with DeepFeat when bottom-up maps are created using response of different layers of GoogLeNet and ResNet50. 29](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014116)

[Figure 7: CC and KL Divergence scores of saliency maps produced by DeepFeat when bottom-up maps are created using response of different layers of GoogLeNet and ResNet50. 30](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014117)

[Figure 8: NSS and SIM scores of saliency maps produced by DeepFeat when bottom-up maps are created using response of different layers of GoogLeNet and ResNet50 31](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014118)

[Figure 9: Averaged scores of four implementations of the DeepFeat model using six evaluation metrics: AUC Judd, AUC Borji, CC, KL, NSS and SIM. 32](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014119)

[Figure 10: Averaged scores of eight saliency models using six evaluation metrics: AUC Judd, AUC Borji, CC, KL, NSS and SIM. 34](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014120)

[Figure 11: Averaged ROC curves of all models for AUC Borji 36](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014121)

[Figure 12: Averaged ROC curves of all models for AUC Judd 37](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014122)

[Figure 13: Visualization of AUC Judd, CC, KL Divergence, NSS and SIM using two different saliency models 38](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014123)

[Figure 14: Visualization of saliency maps obtained using different saliency models 39](file:///C:\\Users\\prati\\Downloads\\Thesis Draft Three.docx" \l "_Toc47014124)

LIST OF TABLES

TABLE PAGE

[Table 1: Comparison of Used Saliency Models 17](#_Toc47014574)

[Table 2: A description of evaluation metrics 21](#_Toc47014575)

[Table 3: Comparison of scores of DeepFeat with GoogLeNet Inception and ResNet50 Convolutional layers for six evaluation metrics. 33](#_Toc47014576)

[Table 4: Averaged scores of eight saliency models using six evaluation metrics: AUC Judd, AUC Borji, CC, KL, NSS and SIM 35](#_Toc47014577)

# CHAPTER 1

# INTRODUCTION

## Saliency Modeling on Visual Attention

In recent years, immense strides have been made in the field of computer vision research, and saliency modeling is one of the significant achievements. In computer vision, a saliency map is a 2D topological map which indicates the probabilistic distribution of visual attention priorities. A large value of saliency at a region corresponds with a large probability for eye fixations to fall on that region. Computational saliency models provide details of the visual saliency process. These models take in a scene (an image, video, etc.) as input and the output of these models is a saliency map.

Feature Integration Theory [3] elaborated that human visual perception could be a multi-stage process. It consists of a two stages: 1) a pre-attentive stage where object is analyzed in terms of its different features (e.g., color, orientation, intensity, etc.), which are processed in different areas of the brain; 2) a focused-attention stage, which integrates different features in order to perceive the object as a whole.

The factors that influence attention can be either scene-driven bottom-up factors or task-driven or expectation-driven top-down factors. Whether a saliency model relies on bottom-up cues, top-down cues, or some combination of both, is one of the most significant ways to distinguish how different models compute the saliency maps. Conventional saliency models were largely bottom-up (example [1]) . Several state-of-art models these days compute top-down maps (CAM [2] has made this computation easier)The stimulus-driven bottom up cues [4] are primarily based on the visual scene characteristics like color, intensity, shape, orientation, etc. The bottom-up models thus compute visual saliency by integrating pre-attentive features without any bias. These models, thus, do not consider prior knowledge or visual cues from predefined tasks into feature integration Human vision, however, is extensively task-driven. Although the bottom-up models are fast, failing to incorporate goal-driven cues into the bottom-up models results in partial explanation of the scene by the model. The top-down models, on the other hand, are based on high level cues such as goal, reward, expectations, and prior knowledge [5] [6]. The top-down allow the feature extraction and integration to be manipulated by the demand of a task. There have been researches on combining the information from top-down and bottom-up cues for saliency detection. [7] [8]

Most saliency models either use locally computed salient features, or similarity and comparisons all over the image in a global processing. For example, Koch and Ullman constructed one of the first saliency maps exploiting feature maps from basic visual elements like color, orientation, disparity, etc.[9] There are a variety of algorithms which have been proposed in order to computationally define salient regions and consequently to create a saliency map. These computational models consider different degrees of local and global features, or a combination of those.

Itti model [1] , one of the first saliency models to be proposed, adapted from Koch and Ullman [9], is based on extraction of three visual features (i.e., color, intensity, and orientation). There is a competition between different spatial locations for saliency within each feature map. The only locations that can persist are the ones which can locally stand out from their surround. This architecture is sensitive to local spatial discontinuities and is effective in detecting locations which stand out from their surroundings. This is a general computational principle in the human visual system. This model appears to take into consideration, that the border between adjacent luminance-defined areas appears to have higher contrast than it actually has. [10]



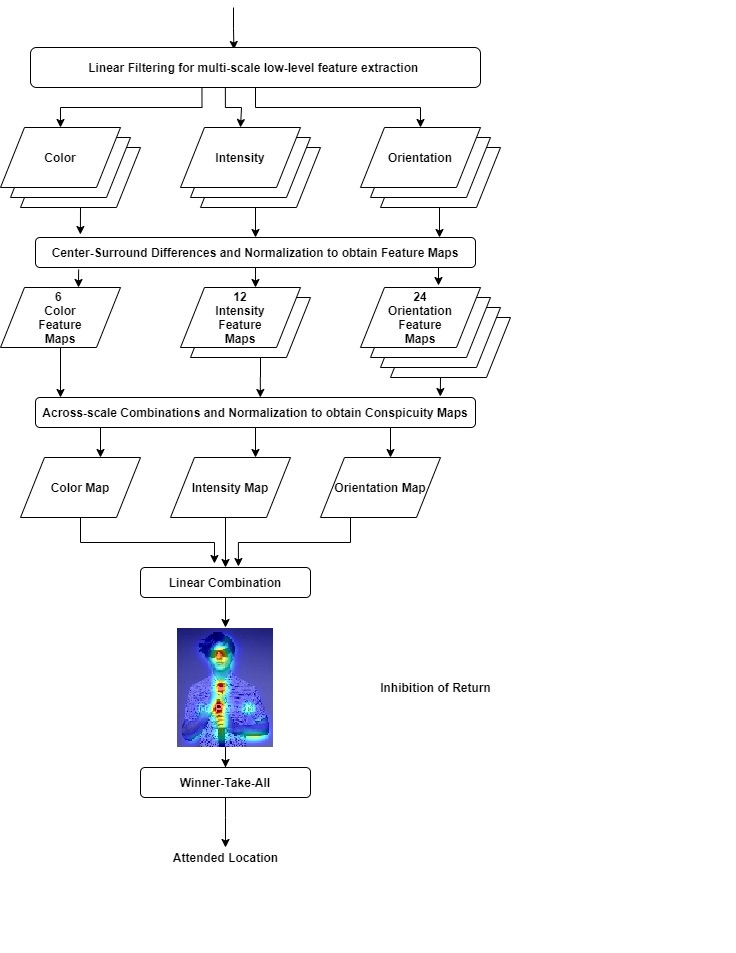


Figure 1: Architecture of Itti Model [1] used in obtaining saliency map in our study

In Itti model, the input image is first linearly filtered to obtain Gaussian pyramids of color, intensity, and orientation features. This is followed by application of spatial competition via center-surround, which is implemented as the difference between fine scale image and interpolated (to the dimension of fine scale) coarse scale image. This process is repeated for all features, and for different scales. Forty-two feature maps are obtained, which include 6 color feature maps, 12 intensity feature maps and 24 orientation feature maps. All individual feature maps are combined across scales and normalized to create individual conspicuity maps of color, intensity, and orientation features. These three maps are linearly combined with equal weights into a single master saliency map. Thus, this map topographically encodes local conspicuities, or saliencies, over the entire visual scene. In this map, the location for maximum is searched, and winner-take-all method is applied to select the most salient location. Usually, Gaussian blur is applied around the maximum and in next iteration, inhibition of return is applied to find next salient location. Figure 1 shows the architecture of Itti model, which is a general principle behind most bottom-up saliency models.

Predicting human eye fixations had relied on conventional image processing techniques, with a visible trend of continuous and modest improvements in the performance of such models over time. There have been rapid improvements across all benchmark scores, and the novel deep neural network (DNN) based saliency models dominate the chart of top scores on saliency benchmarks recently, following the advent of DNNs [11]. Some new models, for example, DeepFeat [8], exploit deep features of DNNs for object recognition as optimized features to predict a saliency map.

Visual attention modeling has been employed in an array of applications across various branches of computer science. Some examples of such applications can be seen in computer vision, image and video processing, Human-Computer Interactions, robotics, and human behavior. The manner in which the saliency information is utilized may be application specific. Some applications are in image segmentation [12] [13], image/video compression [14] [15], object detection [16] [17], robot navigation [18], detection of defects in objects [19], and advertisement design [20].

Figure 2 shows four representative fashion images and corresponding the ground truth maps (presenting experimental human visual attentions data) and prediction saliency maps using Itti mode. Row 1 and 2 consist of the input images to the model and the ground-truth maps derived from the collected mouse-click fixation points from human subject study. Row 3 consists of output saliency maps of Itti model, highlighting the predicted salient regions by the model in the input images.



Input Image

Ground Truth

Map

Itti Model

Figure 2: Example of saliency maps created by Itti saliency model.

## Applications of State-of-Art Computer Vision in Fashion Design

State-of-art computer vision technologies have been extensively applied in fashion design in multiple ways. Recently, Cheng et al. [21] conducted a comprehensive survey of the current state-of-art research progress in the fashion field and categorized the studies into four broad categories: detection, analysis, synthesis, and recommendation. From 2016 to 2019, several studies were published along with datasets [22] [23] [24], which worked towards annotating the fashion landmarks, e.g., left and right collar, sleeves, waistline and hem, on an image. Cheng et al. [21] argued that landmark detection could be considered as a critical step in fashion design application as other applications such as parsing, makeover, suggestion, etc. build on landmark detection.

Yamaguchi et al., [25] in 2012, conducted early works on fashion parsing, which is essentially semantic classification of clothing classes as per body parts. There have been multiple neural network and deep learning-based methods developed for parsing and item retrieval from fashion images [26] [27] [28] [29] . Neural networks have also been applied in recent years in fashion popularity predictions [30] [31] [32] [33] and fashion recommendations like outfit matching [34]. Virtual try-ons [35], pose transformation with clothing [36], and physical simulations [37] have been other interesting applications of computer vision in fashion design.

## Saliency Models for Fashion Images

From among various applications, saliency modeling of fashion images is one of the applications which has not been studied extensively. Wedel and Pieters [38] suggested that the abilities to result in eye fixation and registration in the viewers’ memory are both critical qualities of a good visual advertisement. They found that the ability of an advertisement to register in the viewers’ memory was in direct relation to the frequency of eye fixations in the advertisement. Researchers have studied and analyzed eye movements and eye fixations of consumers to study consumer behavior. Russo and Leclerc [39] analyzed choice processes for consumer nondurables using eye-fixation. Rayner et al. [40] studied the eye movements of the viewers to study the pattern in which they look at text and pictorial information in the print advertisements.

While the works mentioned above contributed to studying the consumer behavior, they did not provide any specific models to predict the viewers visual attention. Xiao et al. [41] incorporated Itti-Koch Model to predict saliency maps and compared it with eye tracking data. However, in their study, maps are not evaluated under any specific evaluation metric. Ma et al. [42] have evaluated advertisement using visual saliency based on foveated image, which closely mimics the behavior of human visual system [43].

## Motivation and Contribution of Our Study

Prior to our study, there has been no extensive analysis of the visual saliency on fashion images using various models and several evaluation metrics. The contribution of this study can be summarized as follows:

* We analyze visual saliency on newly collected human attention data using both conventional saliency models as well as state-of-art neural network-based saliency models.
* The values obtained from experiments are then compared with the benchmarked scores obtained by application of these models on MIT300 Dataset [44].
* This study demonstrates that saliency models can be applied to effectively predict salient regions in random fashion advertisement images.
* The results of this work support the idea that these saliency models can be used to evaluate the ability of fashion advertisements images to attract visual attention fixations.

## Outline of this Document

Chapter 2 of this thesis presents literature review and theoretical background on saliency modeling. This section also elaborates on key concepts used in creating saliency map, like Class Activation Maps (CAM) and Softmax technology, along with introducing the models used in this study. Furthermore, chapter 2 discusses the Evaluation Metrics, which are used to quantitatively evaluate the performance of saliency models.

Chapter 3 elaborates the experimental setup, methods, and framework of this study. This section specifically includes the details on how DeepFeat is implemented on the collected data. A short outline of how other saliency models are implemented is also included. Furthermore, this section describes how several evaluation metrics are calculated to evaluate the performance of each saliency model.

Chapters 4 includes the results and discussions of this study. It compares different implementations of DeepFeat with model, leading to the choice of GoogLeNet Inception saliency maps in order to compare DeepFeat with all other saliency models. Chapter 4 covers the visualization of evaluation metrics, which assists in a better understanding of the obtained results.

In chapter 5, we conclude this study and discuss the implications, limitations and future works.

# CHAPTER 2

# BACKGROUND AND TECHNNOLOGY REVIEW



## Literature on Saliency Models

Triesman and Gelade’s [3] ‘Feature Integration Theory’ marks one of the first studies about the role of features on visual attention. It proposed that visual stimuli are processed as bottom-up features in parallel and that the brain decides which features were important as per the task. The concept of producing a saliency map by integration of these features was introduced by Koch and Ullman [9]. The first complete saliency model was proposed by Itti, Koch et al. in 1998 [1], which verified and implemented the model proposed by Koch and Ullman. Parkhurst and Neibur were the ﬁrst to measure saliency models against human eye ﬁxations in free-viewing tasks [45]. The conventional saliency models primarily relied on extracting simple feature maps from the images. Such features primarily entailed intensity, color, and orientation. These features were combined after applying center-surround and normalization operations. Eventually, in order to predict the visual attention better, the newer models started to incorporate mid- and higher-level features (e.g. face and text [46] , gaze direction [47]).

A cascade of saliency models, which used variety of computational mechanisms, were developed. Itti et al.’s first basic model was one of biologically inspired cognitive models, many of which were based on Feature Integration Theory (for example, [48]) . Several Bayesian models were proposed [49] [50] [51] , which consisted of a combination of sensory evidence and prior knowledge. Some models used Fourier transform and its spectral analysis to compute saliency maps [52] [53] . Similarly, multiple Graph- Based models [54] [55] and Information Theory models [56] [57] [58] were developed.

With neural networks and deep learning getting increasingly popular, new learning-based and pattern classification models have been developed. The supervised learning models, for example [59], [60] etc., learn functions from labelled training data. Several unsupervised deep learning models are developed, which learn to predict from non-labelled data. The models are comprised of multiple layers which learn representation of images with multiple level of abstractions.

Large-scale fixation datasets are not widely available. In order to overcome this limitation, a version of transfer learning is exploited in several deep saliency models. Multiple CNNs are trained for scene recognition which researchers apply for saliency prediction. Such saliency models exploit the weights obtained by pre-existing networks for large image datasets. This is done by fine-tuning such weights onto small scale datasets (mouse-click data, eye-tracking data, etc.). Thus, these models transfer the already learnt visual knowledge (i.e. features) from CNNs into saliency prediction task.(example, DeepGaze II [61], SALICON [62], DeepFeat [8], etc.)

There is a big performance gap between conventional saliency models and modern deep learning saliency models [44]. This can be attributed to the fact that traditional models are not nearly as capable of extracting high level features as compared to deep networks, which can combine multiple layer features linearly in the fully connected layer. Thus, intuition dictates that as compared to a traditional saliency model, a deep learning model can capture significantly higher number of individual cues (from individual layers) as well as composite cues (from addition and fully connected layers). Several deep learning models will be discussed later in this document (DeepFeat [8], eDN [63], and SAM [64]).

## Concepts of Softmax Activation and CAM (Class Activation Map)

The concept of Softmax activation is exploited in saliency modeling to create a probabilistic map. In mathematics, a Softmax function is such that it normalizes the input vector of *K* real numbers into a probability distribution consisting of *K* probabilities proportional to the exponentials of the input numbers. This output vector consists of positive numbers between (0,1) and the sum of elements always adds up to 1. This function is often used in the final layer of a

neural network-based classifier to assign decimal probabilities to each class and thus to create a normalized probabilistic classification.

The Softmax formula is given as follows:

Where values are the elements of input vector and can take any real value. The standard exponential function is applied to each element of the input vector, which gives a positive value > 0. The denominator term normalizes the output ensuring the all output values are positive values between 0 and 1, and that they add up to 1. *K*  is the total number of classes in the multi-class classifier.

A CAM for any class is used to visualize the image regions used by CNN to identify that particular class [2]. However, it can only be used to visualize the final layer as a heatmap and not to visualize previous layers in the network. (This limitation is addressed in Grad-CAM [65]). It is generated using GAP (Global Average Pool) layer in CNNs just before the Softmax layer.

The global average pooling results in an array with average scores for each feature class. When this output is weighted over each feature weight for the next dense layer, it points back to the region which caused the final feature value to be positive, thus pointing the object features for the detected class in the image. As seen in Figure 2, feature maps, i.e. p feature maps of dimension is flattened to using Global Average Pooling. Each of these averaged feature maps is then multiplied with the weights of the dense layer to obtain a heat map that visualizes the classification. The output CAMs can thus be intuitively used to create top-down saliency map.

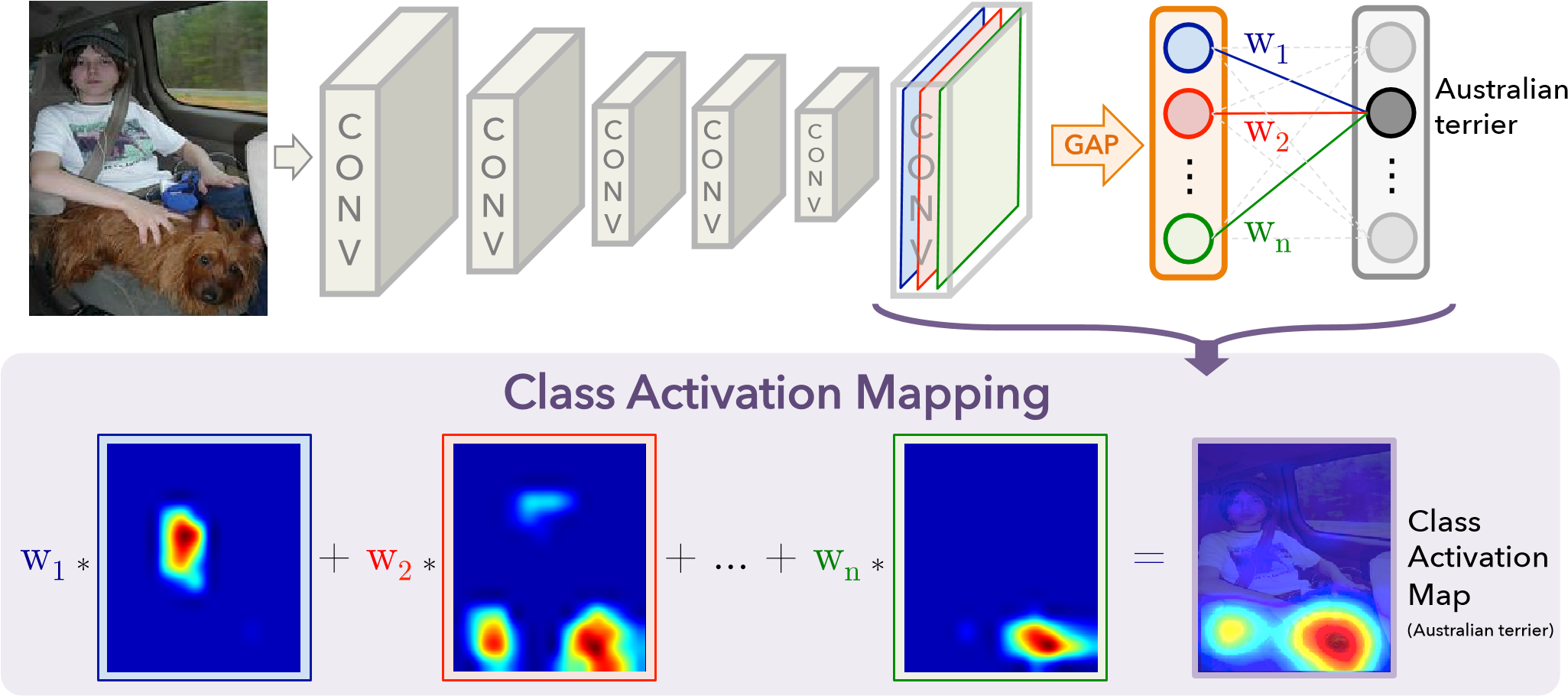


Figure 3: (Taken from Zhou et al. [2]) : Class Activation Mapping: the predicted class score is mapped back to the previous convolutional layer to generate the class.

## Saliency Models Used in Our Experiment

This study has used several traditional as well as deep learning models. These models are chosen such that we would have a variety of models to test our data with. DeepFeat, eDN and SAM are deep learning models while the others are conventional saliency models. Also, some models we used have utilized center bias, and others have not. We have not delved into details of comparison of these models based on the use of center bias. However, our final results suggest the already known superiority of deep learning models to conventional models.

### DeepFeat

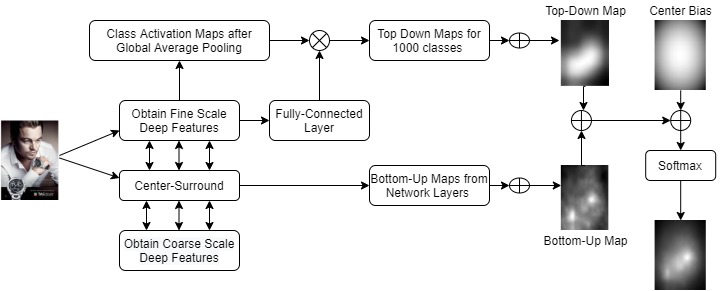
DeepFeat [8] is a novel deep feature-based saliency model that combines bottom-up and top-down features using simple linear combination strategy to obtain a saliency map. This model has been published using the pretrained feature extraction from VGG, GoogLeNet and ResNet DCNNs. The study discussed in this literature was initiated with the analysis of visual saliency of fashion images using DeepFeat saliency model, which was later extended to evaluate saliency maps using several other models. Thus, DeepFeat is elaborated in this document. Figure 3 describes the working of overall architecture of DeepFeat Model as a block diagram.

Figure 4 Architecture of DeepFeat presented as a Block Diagram

DeepFeat computes bottom-up features using fine and coarse scales of DCNN. The fully connected layer is removed to facilitate bottom-up computations and the features obtained from remaining layers are treated as bottom-up visual cues. Center-surround of fine and coarse scales are computed using a gaussian pyramid. Convolution response images are formed in this model using equation 1.

Where denotes the response of each layer to fine scale image as fine scale features, and , denotes the response of each layer to upsampled coarse scale image as coarse scale features. is the total response at layer . The resulting total response image of layer is normalized from 0 to 1 . All normalized features from all utilized layers are then linearly combined to create a bottom-up saliency map, thus, contributing equally to the computation of overall bottom-up saliency map as follows:

Where denotes the denotes the total number of layers in the network, and is the normalization operator.

As shown in Figure 3 and 4, the top-down map is computed using the CAM of the full scale DCNN, utilizing the fully connected layer. All object classes of an image identified by the network are localized and presented as a top down saliency map using CAM. The top-down saliency map is formed by:

Where, *P* denotes the Softmax of the fully connected layer and is the CAM for class *c* formalized as:

Where *A* denotes the tensor of deep features from last activation layer and *W* denotes the vector of weights for each unit.

Mahdi et al. have proposed the linear combination of top-down and bottom-up maps as follows:

(6)

Where denotes a constant equal to 0.5. Next, center bias of human eye is incorporated by computing a gaussian map. The cut off frequency of the Gaussian kernel is the maximum dimension of the image. This incorporation is formalized as:

Where is a constant equal to 0.5, and is the gaussian center bias map formalized as:

Where denotes a constant equal to 1, and correspond to the center of the image, and is the cut-off frequency equivalent to the maximum dimension of the image. The final probability distribution of saliency is formalized by DeepFeat as follows:

Thus, the final output of the model is a Softmax based probabilistic vector, which represents the probabilities of object classes. This study implemented DeepFeat with GoogLeNet [66] and ResNet50 [67] feature extraction networks in order to create saliency maps for the collected fashion images.

### LSTM-Based Saliency Attentive Model (SAM)

SAM [64] is a novel Attentive ConvLSTM (Convolutional Long Short-Term Memory) that sequentially focuses on different spatial locations of a stack of features to enhance predictions. This was the first model to incorporate attentive models in a saliency prediction architecture. This is one of the current state of the art models for saliency prediction. The source code of this method and pre-trained models are publicly available and were utilized in our study. For this study, we have utilized SAM on ResNet50 feature extraction network.

### Ensembles of Deep Networks (eDN)

eDN [63] follows an entirely automatic data-driven approach, performing a large-scale search for optimal features. It identifies those instances of a richly-parameterized bio-inspired model family (hierarchical neuromorphic networks) that successfully predict image saliency.

### Itti Model

This model was the first model to fully define and computationally obtain bottom-up saliency maps. The model first linearly filters the input image to obtain gaussian pyramids of color, intensity, and orientation. It then applies spatial competition via center surround operation to create conspicuous maps corresponding to each of the feature dimensions. These conspicuity maps are then linearly combined with equal weights into a single saliency map. This model was the state-of-art model for well over a decade until the advent of deep learning models along concurrent significant improvement in computational capabilities.

### Graph based visual saliency model (GBVS)

GBVS [68] is a graph implementation of the Itti model. The model uses a Markov chain as an activation map and incorporates a center prior.

### CovSal

[69] Instead of computing a separate feature map in parallel for each feature channel such as color and orientation and combining them to form a master saliency map, this model proposes using covariance matrices of simple image features as meta-features for saliency estimation. The reason they state that they propose this method is because region covariances, being low-dimensional representations of image patches, capture local image structures better than standard linear filters. Furthermore, they naturally provide nonlinear integration of different features by modeling their correlations.

### Fixation Prediction using Unsupervised Hierarchical Models (UHF)

[70] It is an unsupervised multi-scale hierarchical saliency model, which utilizes both local and global saliency pipelines.

### Weighted Maximum Phase Alignment (WMAP) Model

WMAP [71] model significantly accelerates a scene recognition application. It achieves this by combining image features to spatial saliency while simultaneously preserving recognition performance. It uses a mobile robotlike application where scene recognition is carried out through the use of image features to characterize the different scenarios, and the Nearest Neighbor rule to carry out the classification.

Table 1 provides a summary of the saliency models used in this study.

Table 1: Comparison of Used Saliency Models

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Model Name** | **Features** | **Category** | **Year** | **Pub.** | **Ref.** |
| DeepFeat | BT | DL | 2017 | TCDS | [8] |
| SAM | DF | DL | 2018 | TIP | [64] |
| eDN | DF | DL | 2014 | CVPR | [63] |
| Itti | BU | CS | 2006 | N. Net | [1] |
| GBVS | BU | CS | 2006 | NIPS | [68] |
| CovSal | BU | CS | 2013 | JOV | [69] |
| UHF | BU | ML | 2016 | ACCV | [70] |
| WMAP | BU | CS | 2011 | IntechOpen | [71] |

\*BU: Bottom-Up \* BT: Bottom-Up and Top-Down \*DF: Deep Features

\*DL: Deep Learning \*CS: Center-Surround \*ML: Machine Learning

## Evaluation Metrics for Saliency Maps

A saliency model outputs a saliency map, and the performance of this model needs to be quantitatively evaluated by comparing the output map of the model with a ground truth map that is constructed using human fixation data collected through eye-tracker or mouse-clicks. This evaluation can be done under several broadly accepted metrics. In this study, two binary classification measures have been used: Receiver Operating Characteristics (ROC) and Area Under the ROC Curve (AUC). Further, three similarity metrics and a dissimilarity metric have been measured, which are discussed in the next section. The details on currently used evaluation metrics for saliency models and what each metric means can be found in [11] and [5].

### ROC

ROC is the most popular and most widely used method for evaluating the extent of similarity between the saliency map and the ground-truth map. This metric measures and plots the tradeoff between true and false positive rates at various thresholds of the saliency map. True Positive Rate (TPR) and False Positive Rate (FPR) are defined as follows:

Where TP is the fixated saliency map values above threshold, FP is the 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.

### AUC

It is a location-based similarity metric between the output saliency map and ground-truth fixation map (points), which is given by the integral of the area under the ROC curve. A score higher than 0.5 indicates a prediction higher than chance. In this study , two variations of AUC are computed and presented. Judd-AUC [60] computes the true positive rate and false positive rate over every pixel value in the saliency map. The Borji-AUC [72] computes the true positive rate and false positive rate over a set of thresholds sampled from the dynamic range of the saliency map. This results in a complication in the creation of averaged ROC curve, which shall be discussed in the methodology section of this document. Both Judd-AUC and Borji-AUC compare saliency maps to the exact fixation points of the human fixations. A third AUC variant, called as the Shuffled-AUC, modifies the Borji-AUC by utilizing fixation maps that reflects a continuous distribution of eye fixations. The Shuffled-AUC variant is not used in this study.

### Pearson Correlation Coefficient (CC)

It is a distribution-based similarity metric that evaluates linear relationship between the saliency map produced by the model and the ground-truth fixation map. This metric measures the dependence between the saliency map and fixation map. In order to do this, it treats these two maps as random variables.

Here, is the covariance between the saliency map and the fixation map. A perfect correlation is indicated by a score of 1 or -1. A score of 0 indicates no correlation between the two maps.

### Normalized Scanpath Saliency (NSS)

It is a location-based similarity metric between saliency maps and ground-truth fixation maps (points), which measures the saliency values at fixation locations along a subject’s scanpath. This metric standardizes saliency values in order to have a zero mean and unit standard deviation, which is given as follows:

Where N denotes the number of fixation points, and indexes the fixation points of the binary fixation map. NSS score of zero indicates chance. NSS > 0 indicates an above chance correspondence between maps, while NSS < 0 points towards anti-correspondence. NSS = 1 corresponds to fixations falling on portions of the saliency map with a saliency value one standard deviation above average.

### Kullback-Leibler Divergence (KL)

The KL divergence is a distribution-based dissimilarity metric between the saliency map and the fixation map, which is used to measure the overall dissimilarity between two probability distributions. In the context of saliency, it is used to measures the loss of information when a saliency map approximates a fixation map:

Where G is the ground-truth fixation map. As a dissimilarity metric, a score of zero indicates that the saliency map approximates a fixation map.

### Similarity (SIM)

SIM is a distribution-based similarity metric between the saliency map and ground-truth fixation map. It is a measure of intersection between two distributions, given as:

Where G is the ground-truth fixation map, and:

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.

Table 2 presents an outline and comparison of the evaluation metrics used in this study.

Table 2: A description of evaluation metrics

|  |  |  |
| --- | --- | --- |
| **Metric** | **Type of metric** | **Theoretical range** |
| Area under the ROC curve (AUC) | Location-based, Similarity | [0,1] |
| Pearson’s Correlation Coefficient (CC) | Distribution-based, Similarity | [-1,1] |
| Kullback-Leibler divergence (KL) | Distribution-based, Dissimilarity | [0, ∞] |
| Normalized Scan-path Saliency (NSS) | Location-based, Similarity | [-∞,∞] |
| Similarity (SIM) | Distribution-based, Similarity | [0,1] |

## 

# CHAPTER 3

# EXPERIMENT SETUP, METHODS AND FRAMEWORK



## Psychophysics Toolbox Version 3 (PTB-3)

Psychtoolbox-3 is a set of MATLAB and GNU Octave functions for vision and neuroscience research. This toolbox interfaces MATLAB or Octave with hardware. This toolbox provides ways to easily control and present visual stimuli for our data collection process. Further details about this toolbox can be obtained at (*psychtoolbox.org*).

## Human Subject Data Collection

The ground-truth data collection required human subject study, thus, requiring us to fulfil necessary protocols. The data collection was taken under the auspices of Southern Illinois University, College of Engineering and was approved by Human Subjects Committee. Every individual participating in the data collection had provided informed consent to participate in the study, to get their responses recorded and for the results to be published.

Seventy-five fashion images were chosen from popular fashion magazines and editorials by Dr. Seung Hee-Lee at SIUC School of Architecture. The resolutions of these images varied widely. These images were randomly divided into three groups of 25 to 35 images each. 67 participant responses were taken, from 26 males and 41 females, between ages of 19 and 46. The responses were taken in three groups, each group viewing 25 to 35 different images in random order without repetition. Each image was presented on full screen span (Monitor Size: 21.5-inch Diagonal, Resolution: 1920\*1080, Refresh Rate: 60 Hz) to the viewer for 3 seconds, with a time interval of 1 second between consecutive images. The monitor was placed approximately 33.3 inches (85 cm) away from the viewer. All calibrations and task stimuli in this study were presented using Psychtoolbox-3. The participants were instructed to click on the region of the image on screen that attracted their attention.

## Creation of Ground-Truth Fixation points and Ground-Truth Fixation Maps

The co-ordinates of clicked points were recorded in a .MAT file using Psychtoolbox-3 with MATLAB. This map of binary fixation points needs to be smoothened into a continuous map so as to incorporate the fact that the viewer can click on any location on the multipixel object that catches viewer’s attention. Smoothing provides additional robustness for evaluation. In the case of few observers, smoothing the fixation locations also helps to extrapolate the existing data.

The binary fixation maps have values one for exact fixation location and zero elsewhere. A ground-truth fixation map is obtained by convolution of these fixation maps 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 [73]. Gaussian low pass filter is applied on the fixation points to create a blur and consequently, a ground-truth saliency map.

## Running Saliency Models

For much of the duration of this study, primary focus was given to analyze saliency of collected fashion data with DeepFeat, one of the higher performing models, which was developed at SIUC lab. This was followed by implementation of seven other models, which incorporated conventional as well as state-of-art saliency models.

Meanwhile, DeepFeat was implemented with slight variations using Keras and TensorFlow as well. The python implementation was verified through calculation of evaluation metrics, that yielded similar results as that of MATLAB implementation. A detailed analysis of python implementation is yet to be done.

### Running DeepFeat

All computations in DeepFeat implementation was done in MatConvNet [74] (<https://www.vlfeat.org/matconvnet/>). MatConvNet is a toolbox that implements Convolutional Neural Networks (CNNs) for MATLAB. It facilitates the access of CNN building blocks as MATLAB functions, The toolbox exposes the building blocks of CNNs as easy-to-use MATLAB functions, consists of routines that can compute linear convolutions with filter banks, provides routines for feature pooling calculations, and many more features. This toolbox allows fast prototyping of new CNN architectures and supports efficient computation on CPU as well as GPU. Hence, MatConvNet makes working with large dataset much efficient.

DeepFeat saliency model exploits deep features of CNNs pre-trained for object classification as visual cues to predict human gaze patterns without any further training [8] [75]. In this study, two implementations of DeepFeat are computed and compared, those being with features of GoogLeNet and ResNet50. GoogLeNet has 22 layers and 6.8M parameters while ResNet50 has 50 layers and 25.4M parameters. GoogLeNet stands out on a particular front that it consists of novel ‘inception’ layer that combines multiple scales of convolution layers. The main feature of ResNet50 is that it combines the stack of convolution layers with their residual after every 3 convolutions. The residual network achieves better results by gaining accuracy from increasing depth unlike plain networks which show higher training and test error when depth increases. Both these DCNNs are pretrained for object classification using the ImageNet dataset [76] that consists of 1.28 million images of 1000 classes of objects to classify. Visualization of the architecture of the three DCNNs can be found online at ‘vlfeat.org’.

#### DeepFeat with GoogLeNet

First, four variants of saliency map were computed using pretrained features from GoogLeNet for all 75 fashion images: 1. Top-down map 2. Bottom-up map 3. Top-down and bottom-up maps combined with no center bias and 4. Top-down and bottom-up maps combined, with center bias. The top down maps were obtained using Class Activation Maps as elaborated in section 2.2. The bottom-up maps were computed using center-surround of two scales of deep features. Center-surround is elaborated in section 1.1.

Five variants of bottom-up saliency maps were computed using the responses from various layers of GoogLeNet: Convolutional response images, ReLU activation response images, maximum pooling response images, response of the inception layers and finally response of all activations of the network. Thus, in total, 20 batches of saliency maps were computed for all 75 images with GoogLeNet alone.

In a typical cascade layer, one always needs to choose what type of filters, and what type of layers to use. Inception layer of GoogLeNet is a cascade of inception blocks [66] (inception module), which is such that each module is a composed by 4 filters, and not just one choice: a convolution filter, a convolution filter followed by a convolution, a convolution followed by a convolution, and an average pooling followed by a convolution. At the top, the outputs of each of the above four processes are concatenated. Using convolution before a and a convolutions enables reducing the computational cost (number of multiplications) by as much as ten folds. GoogLeNet Inception consists of 8 inception blocks. Figure 8 shows the architecture of a single inception block.

Additionally, two of the GoogLeNet inception modules in hidden layers consist of side branch with Average pooling followed by fully connected layers and Softmax, which tries to make label predictions based on the hidden layers. The features computed at these intermediate levels have a regularizing effect on the network and add more information to prevent overfitting. In section 4, we consider the saliency maps created using GoogLeNet Inception layers responses in order to compare the performance of DeepFeat along with seven other saliency models.



Figure 5: Architecture of a single Inception module

#### DeepFeat with ResNet50

Similar process as with GoogLeNet wax repeated with ResNet50 as well. The top down maps were obtained using Class Activation Maps. Six variants of bottom-up saliency maps were computed using the responses from various layers of ResNet50: Convolutional response images, ReLU activation response images, batch normalization response images, responses of all network blocks except the residual short cuts, responses of concatenation between the network blocks and residuals, and finally response of all activations of the network including the residuals. In total, 24 batches of saliency maps were computes for all 75 images with ResNet50.

In section 4, we consider the saliency maps created using convolutional responses of ResNet50 in order to compare the performances of DeepFeat between GoogLeNet and ResNet.

## Running Other Saliency Models

Seven other saliency models, as listed in section 2.3, were used to create saliency maps for the collected fashion data. The links to the implementations were available in MIT Saliency Benchmarking website (<http://saliency.mit.edu/>) in MATLAB or python codes. This benchmarking [77] was done as per the performance of the models on MIT300 Dataset. Some models utilized MATLAB with MatConvNet while a few were implemented with TensorFlow and Theano. Computation was done using GPU resources with CUDA and CUDNN whenever applicable.

## Calculation of Evaluation Metrics

The results of all above-mentioned saliency models were analyzed and evaluated using five evaluation metrics. The evaluation metrics were calculated using [44]. Five evaluation metrics are calculated in this study to ensure that any conclusions thus drawn are independent of the choice of metric and consistent across as many metrics as possible. A decent performance across 5 metrics would point towards an overall good saliency model. CC, KL Divergence, NSS and SIM scores were evaluated for each individual image using [44], followed by calculation of the final average score and standard error of mean for each saliency model used in the study. Visualization of these metrics is discussed in section 4 of this document.

### Specifics of ROC and AUC Calculations

Average AUC score and Standard Error of Mean were calculated from the individual AUC scores of 75 saliency maps. Obtaining an average ROC curve, however, was implemented slightly different than by directly averaging the True Positive and False Positive points obtained using the MIT saliency benchmarking code.

The Borji-AUC [72] computes true positive rates and false positive rates over a set of thresholds sampled from the dynamic range of the saliency map. Thus, if we wanted to take the average TP and FP rates at given thresholds, we could take those averages for each of the 75 images and create average ROC curve.

However, Judd-AUC [60] computes the true positive rate and false positive rate over every pixel value in the saliency map. Thus, the total number of TPR and FPR points varies across all 75 images according to the image dimension. Averaging the scores like in Borji-AUC would not be feasible in this case. In order to get around this limitation, we set up a new x-axis, i.e. FPR values, as a linearly spaced vector with 20 points. The TPR and FPR points of each individual image were then interpolated across the new x-axis to create a set of values which could now be averaged in order to create an average ROC curve. The visualization is discussed in section 4.

# CHAPTER 4

# RESULT AND DISCUSSION

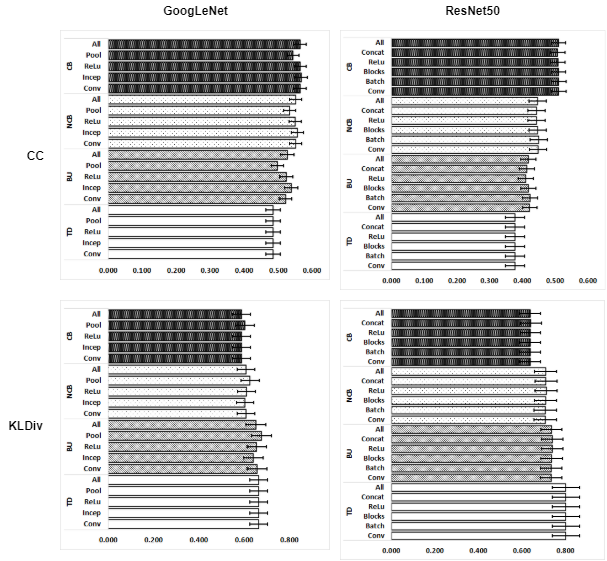


## Performance of DeepFeat with GoogLeNet and ResNet50

\* \*

\* \*

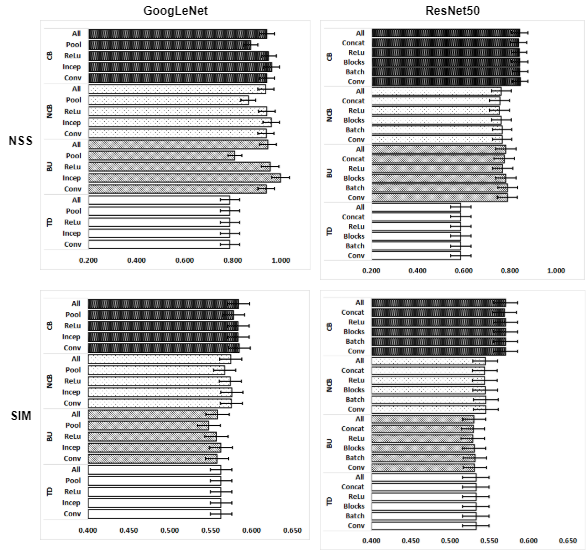
Figure 6: AUC Borji and AUC Judd scores of saliency maps with DeepFeat when bottom-up maps are created using response of different layers of GoogLeNet and ResNet50.

In Figures 6, 7 and 8, a \* indicates that the scores of two consecutive map types (TD: Top-Down, BU: Bottom-Up, NCB: Top-down and Bottom-Up with no Center Bias, or CB: Top-down and Bottom-Up with no Center Bias) are significantly different using t-test at confidence level of p ≤ 0.05. Map types that are not consecutive have a larger probability to achieve statistical significance. Standard error of the mean (SEM) is indicated by the error bars.

\*

Figure 7: CC and KL Divergence scores of saliency maps produced by DeepFeat when bottom-up maps are created using response of different layers of GoogLeNet and ResNet50.

Figure 6 shows that while there is a clear improvement in AUC scores between top-down maps and the other (BU, NCB and CB) maps, the above results do not show a significant difference between the performance of DeepFeat in terms of AUC scores in the latter three maps. Also, near to no difference is noticed between the performances of the model using different set of network response images to create bottom-up maps, which consequently resulted in absence of significant differences in NCB and CB maps. Figures 7 and 8 point towards a similar trend in terms of CC, KL divergence, NSS, and SIM scores.



\* \*

Figure 8: NSS and SIM scores of saliency maps produced by DeepFeat when bottom-up maps are created using response of different layers of GoogLeNet and ResNet50

The explanation behind this result could be, that identical to similar features are extracted when we take the response images from different combinations of layers in these networks. We compared the results of maps created using convolutional layers of ResNet 50 and GoogLeNet Inception since they performed the best among different layer responses of their own network.



Figure 9: Averaged scores of four implementations of the DeepFeat model using six evaluation metrics: AUC Judd, AUC Borji, CC, KL, NSS and SIM.

In Figure 9, a \* indicates the two consecutive models are significantly different using t-test at confidence level of p ≤ 0.05. Models that are not consecutive have a larger probability to achieve statistical significance. Standard error of the mean (SEM) is indicated by the error bars. As we can see from Figure 9 and Table 3, GoogLeNet Inception layers result in saliency Maps that outperform ResNet50 maps over all evaluation metrics.

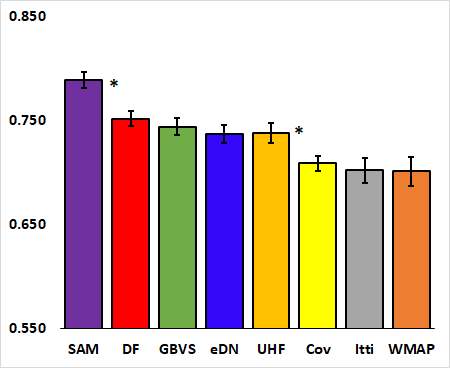
Table 3: Comparison of scores of DeepFeat with GoogLeNet Inception and ResNet50 Convolutional layers for six evaluation metrics.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **GoogLeNet Inception Layer** | | | | **ResNet50 Convolutional Layers** | | | |
| **Metric** | TD | BU | NCB | CB | TD | BU | NCB | CB |
| AUC B | 0.709 ± 0.010 | **0.753 ± 0.008** | 0.749 ± 0.008 | **0.748 ± 0.007** | 0.659 ± 0.011 | 0.710 ± 0.010 | 0.708 ± 0.010 | **0.728 ± 0.008** |
| AUC J | 0.712 ± 0.010 | **0.759 ± 0.008** | 0.753 ± 0.008 | **0.752 ± 0.007** | 0.666 ± 0.011 | 0.716 ± 0.011 | 0.713 ± 0.011 | **0.732 ± 0.008** |
| CC | 0.484 ± 0.021 | 0.539 ± 0.019 | 0.556 ± 0.018 | **0.568 ± 0.018** | 0.377 ± 0.029 | 0.422 ± 0.023 | 0.448 ± 0.026 | **0.510 ± 0.023** |
| KL | 0.664 ± 0.038 | 0.640 ± 0.043 | 0.604 ± 0.039 | **0.588 ± 0.040** | 0.800 ± 0.063 | 0.733 ± 0.049 | 0.706 ± 0.051 | **0.636 ± 0.048** |
| NSS | 0.789 ± 0.041 | **0.998 ± 0.037** | 0.960 ± 0.034 | **0.962 ± 0.031** | 0.586 ± 0.044 | 0.789 ± 0.043 | 0.765 ± 0.042 | **0.844 ± 0.034** |
| SIM | 0.563 ± 0.013 | 0.562 ± 0.014 | 0.576 ± 0.014 | **0.584 ± 0.014** | 0.533 ± 0.017 | 0.531 ± 0.015 | 0.545 ± 0.016 | **0.571 ± 0.015** |

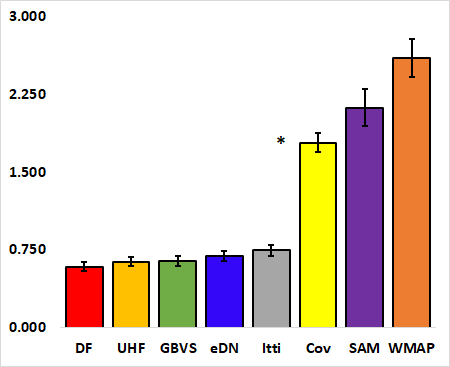
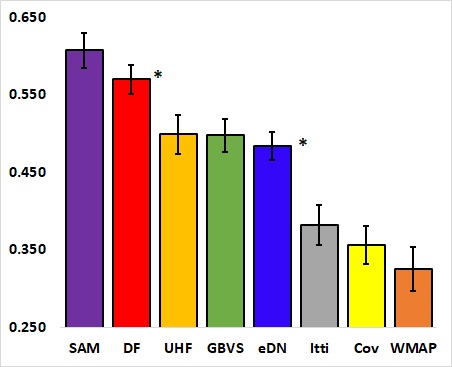
Thus, for the remaining of this chapter, we consider the saliency maps created using GoogLeNet Inception layers responses in order to compare the performance of DeepFeat along with seven other saliency models on our fashion data.

## Performance of DeepFeat and Seven Other Saliency Models

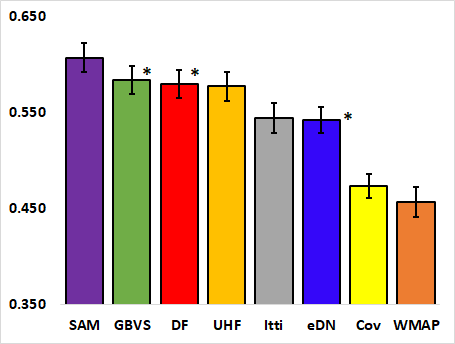
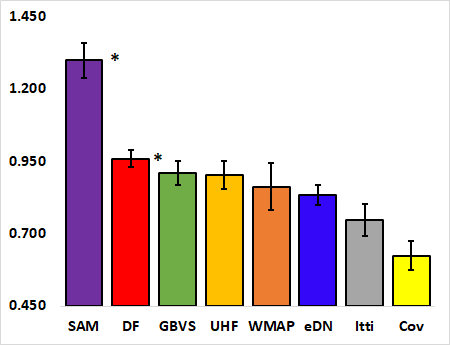
Saliency maps obtained by passing the collected fashion data onto several saliency models were compared with the ground truth map over six evaluation metrics. As shown in Figure 10 and Table 4, the traditional image-processing models as well as state-of-art neural network and deep feature-based models produced decent results, with the neural network and deep feature-based models clearly seen to be outperforming the conventional methods over a majority of metrics.



**AUC Borji** **AUC Judd**



**CC KL Div**



**NSS**  **SIM**

Figure 10: Averaged scores of eight saliency models using six evaluation metrics: AUC Judd, AUC Borji, CC, KL, NSS and SIM.

In Figure 10, a \* indicates that the two consecutive models are significantly different using t-test at confidence level of p ≤ 0.05. Models that are not consecutive have a larger probability to achieve statistical significance. Standard error of the mean (SEM) is indicated by the error bars.

Table 4: Averaged scores of eight saliency models using six evaluation metrics: AUC Judd, AUC Borji, CC, KL, NSS and SIM

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Evaluation Metrics | | | | | |
| Saliency Models | AUC Borji | AUC Judd | CC | KL Div | NSS | SIM |
|  | Neural network and Deep Feature-based models | | | | | |
| DeepFeat | **0.748 ± 0.007** | **0.752 ± 0.007** | **0.568 ± 0.018** | **0.588 ± 0.040** | **0.962 ± 0.031** | **0.584 ± 0.013** |
| SAM ResNet | 0.720 ± 0.009 | **0.790 ± 0.008** | **0.607 ± 0.023** | 2.126 ± 0.176 | **1.299 ± 0.061** | **0.607 ± 0.015** |
| eDN | 0.732 ± 0.009 | 0.738 ± 0.009 | 0.484 ± 0.018 | 0.692 ± 0.045 | 0.834 ± 0.033 | 0.542 ± 0.014 |
|  | Conventional image processing-based Models | | | | | |
| GBVS | **0.732 ± 0.008** | 0.745 ± 0.008 | 0.498 ± 0.021 | 0.644 ± 0.051 | 0.910 ± 0.042 | 0.584 ± 0.015 |
| UHF | 0.732 ± 0.010 | 0.738 ± 0.010 | 0.499 ± 0.026 | **0.639 ± 0.047** | 0.903 ± 0.049 | 0.578 ± 0.015 |
| Itti | 0.693 ± 0.012 | 0.703 ± 0.012 | 0.383 ± 0.026 | 0.747 ± 0.053 | 0.748 ± 0.056 | 0.544 ± 0.015 |
| CovSal | 0.624 ± 0.009 | 0.709 ± 0.008 | 0.356 ± 0.025 | 1.785 ± 0.092 | 0.624 ± 0.050 | 0.474 ± 0.012 |
| WMAP | 0.637 ± 0.013 | 0.702 ± 0.014 | 0.325 ± 0.028 | 2.604 ± 0.187 | 0.863 ± 0.082 | 0.457 ± 0.015 |

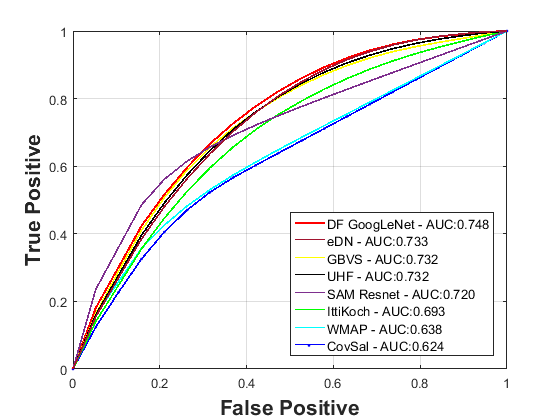
The averaged ROC Curves for AUC Borji and AUC Judd variants are presented in Figures 11 and 12, respectively. The graph is averaged for 75 images as discussed in section 3.6.1. We can see from the plot that area under the curve for SAM Resnet is significantly higher than the other models for Judd variant, the visualization having been confirmed by the model leading in AUC Judd score as well. However, the model performs slightly inferior in terms of AUC Borji. The area under the ROC curve can be observed to be proportional to the respective AUC scores of the models. A straight line represents chance model, which implies that all applied models are significantly better than chance in predicting the salient objects in the images.

Figure 11: Averaged ROC curves of all models for AUC Borji

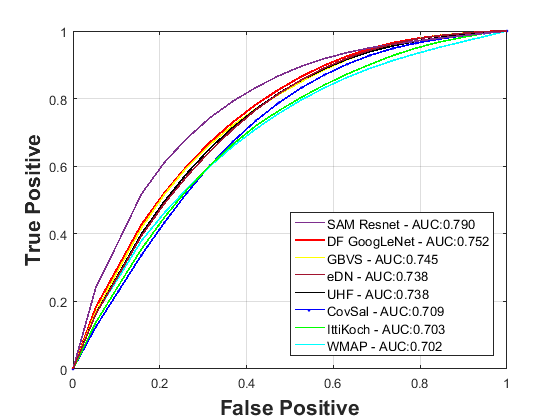


Figure 12: Averaged ROC curves of all models for AUC Judd

## Visualization of Evaluation Metrics

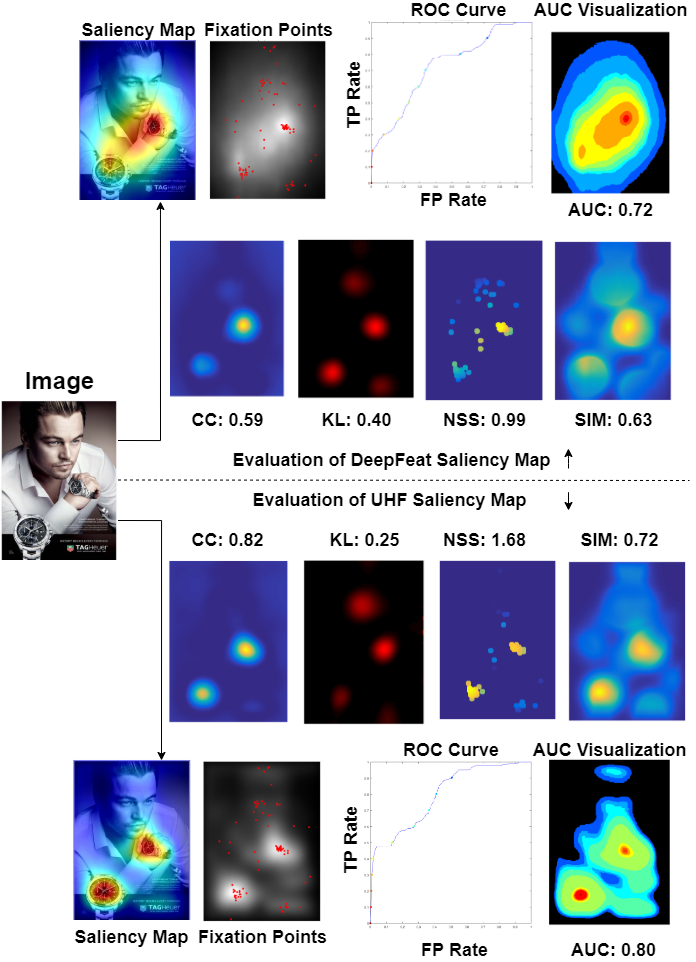


Figure 13: Visualization of AUC Judd, CC, KL Divergence, NSS and SIM using two different saliency models

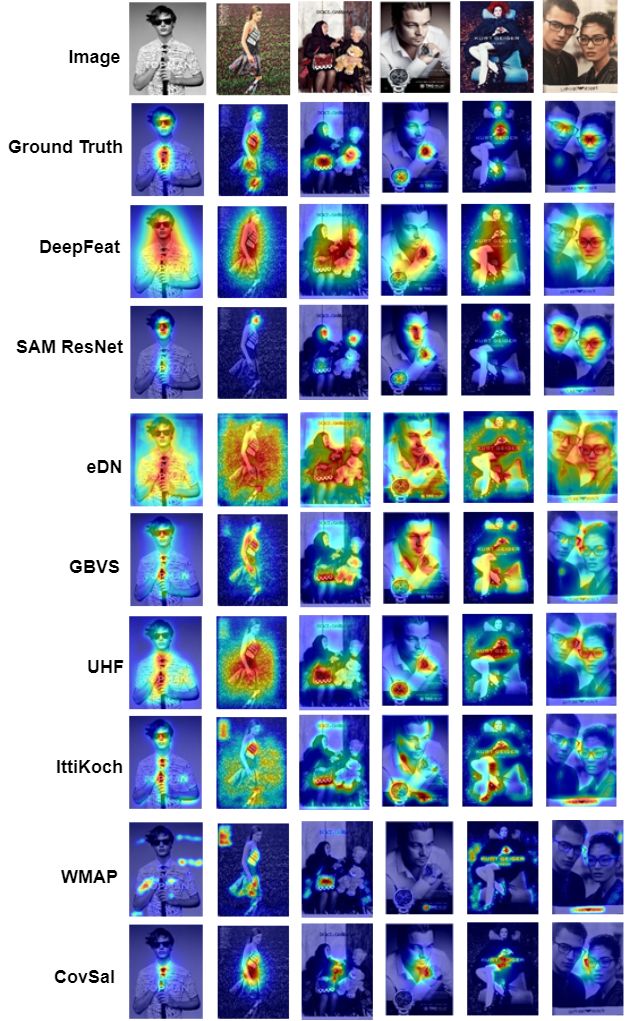


Figure 14: Visualization of saliency maps obtained using different saliency models on 6 out of the 75 collected images

The results in Figure 10 and Table 4 show that not all models perform well in all evaluation metrics. The results of figures 13 further support this observation. Results portrayed in figure 13 are obtained using [44]. These images visualize all evaluation metrics incorporated in the previous sections, and visually demonstrate how a good saliency map performs consistently over multiple metrics and also how different models may perform differently across same metrics. Figure 14 presents the saliency maps created by various models tested on the fashion dataset.

# CHAPTER 5

# CONCLUSION, LIMITATIONS AND FUTURE DIRECTIONS



## Conclusion

This study incorporated collection of attention fixation data from human subjects on random fashion images. Ground-truth saliency maps were constructed by incorporating gaussian blur on the fixation points. Saliency maps were then predicted using multiple saliency models and the results compared with the ground-truth map over six evaluation metrics. The results of this study indicate that saliency models could in fact be applied to effectively predict salient regions in random fashion advertisement images. The averaged AUC scores of each of the tested models, as well as the scores for individual images, were well beyond chance (chance score being 0.5 and visualized by a straight line with slope 1 on ROC curve). The performance of saliency models on our collected data over the computed evaluation metrics was comparable to the benchmarked scores. Most importantly, the models perform well consistently over multiple evaluation metrics, hence, indicating overall efficacy.

## Limitations and Future Directions

This study was conducted with mouse click data which on one hand, could be more accurate about collected attention points, but on the other hand, only very few fixation points could be taken in. The next step in this study will be moving towards analysis of eye tracking data using various saliency models and comparing the results with mouse click data.

Furthermore, initially, this study expected a better result with fashion data using top-down models than bottom up, which would make more sense given that the fashion images focus more on objects. However, we realized that CAM, which is used to obtain top-down maps, can take in features from only the final layer before SoftMax, and not from previous hidden layers. Another direction in which this study can be augmented, is to use Grad-Cam instead of CAM and see if any improvements can be observed by enabling taking top-down features from hidden layers as well.

Building a learning algorithm from scratch for prediction of saliency in fashion specific images and incorporating transfer learning to predict saliency in other image task specific image sets can be other ambitious directions to move in.

# REFERENCES

[1] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 20, no. 11, pp. 1254-1259, 1998, doi: 10.1109/34.730558.

[2] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning Deep Features for Discriminative Localization," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 2921-2929, doi: 10.1109/CVPR.2016.319.

[3] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology,* vol. 12, no. 1, pp. 97-136, 1980/01/01/ 1980, doi: <https://doi.org/10.1016/0010-0285(80)90005-5>.

[4] H.-C. Nothdurft, "CHAPTER 38 - Salience of Feature Contrast," in *Neurobiology of Attention*, L. Itti, G. Rees, and J. K. Tsotsos Eds. Burlington: Academic Press, 2005, pp. 233-239.

[5] A. Borji and L. Itti, "State-of-the-Art in Visual Attention Modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 35, no. 1, pp. 185-207, 2013, doi: 10.1109/TPAMI.2012.89.

[6] J. M. Henderson and A. Hollingworth, "HIGH-LEVEL SCENE PERCEPTION," *Annual Review of Psychology,* vol. 50, no. 1, pp. 243-271, 1999/02/01 1999, doi: 10.1146/annurev.psych.50.1.243.

[7] G. Zhu, Q. Wang, and Y. Yuan, "Tag-Saliency: Combining bottom-up and top-down information for saliency detection," *Computer Vision and Image Understanding,* vol. 118, pp. 40-49, 2014/01/01/ 2014, doi: <https://doi.org/10.1016/j.cviu.2013.07.011>.

[8] A. Mahdi, J. Qin, and G. Crosby, "DeepFeat: A Bottom-Up and Top-Down Saliency Model Based on Deep Features of Convolutional Neural Networks," *IEEE Transactions on Cognitive and Developmental Systems,* vol. 12, no. 1, pp. 54-63, 2020, doi: 10.1109/TCDS.2019.2894561.

[9] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," (in eng), *Hum Neurobiol,* vol. 4, no. 4, pp. 219-27, 1985.

[10] S. Yantis, *Sensation and Perception*. New York: Worth Publishers, 2014.

[11] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What Do Different Evaluation Metrics Tell Us About Saliency Models?," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 41, no. 3, pp. 740-757, 2019, doi: 10.1109/TPAMI.2018.2815601.

[12] A. K. Mishra, and Yiannis Aloimonos, "Active segmentation," *International Journal of Humanoid Robotics,* vol. 06, no. 03, pp. 361-386, 2009.

[13] A. Maki, P. Nordlund, and J.-O. Eklundh, "Attentional Scene Segmentation: Integrating Depth and Motion," *Computer Vision and Image Understanding,* vol. 78, no. 3, pp. 351-373, 2000/06/01/ 2000, doi: <https://doi.org/10.1006/cviu.2000.0840>.

[14] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Transactions on Image Processing,* vol. 13, no. 10, pp. 1304-1318, 2004, doi: 10.1109/TIP.2004.834657.

[15] C. Guo and L. Zhang, "A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression," *IEEE Transactions on Image Processing,* vol. 19, no. 1, pp. 185-198, 2010, doi: 10.1109/TIP.2009.2030969.

[16] N. J. Butko and J. R. Movellan, "Optimal scanning for faster object detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 20-25 June 2009 2009, pp. 2751-2758, doi: 10.1109/CVPR.2009.5206540.

[17] K. A. Ehinger, B. Hidalgo-Sotelo, A. Torralba, and A. Oliva, "Modelling search for people in 900 scenes: A combined source model of eye guidance," *Visual Cognition,* vol. 17, no. 6-7, pp. 945-978, 2009/08/01 2009, doi: 10.1080/13506280902834720.

[18] S. Baluja and D. A. Pomerleau, "Expectation-based selective attention for visual monitoring and control of a robot vehicle," *Robotics and Autonomous Systems,* vol. 22, no. 3, pp. 329-344, 1997/12/01/ 1997, doi: <https://doi.org/10.1016/S0921-8890(97)00046-8>.

[19] M. Mancas, Unay, B., Goselin, B., and Macq, D., "Computational attention for defect localisation," in *Proceedings of ICVS Workshop on Computational Attention & Applications.*, 2007.

[20] S. Jain and S. Kamath S, *Saliency Prediction for Visual Regions of Interest with Applications in Advertising*. 2017, pp. 48-60.

[21] W.-H. Cheng, S. Song, C.-Y. Chen, S. C. Hidayati, and J. Liu, "Fashion Meets Computer Vision: A Survey," *ArXiv,* vol. abs/2003.13988, 2020.

[22] Y. Ge, R. Zhang, L. Wu, X. Wang, X. Tang, and P. Luo, "DeepFashion2: A Versatile Benchmark for Detection, Pose Estimation, Segmentation and Re-Identification of Clothing Images," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 5332-5340, 2019.

[23] Z. Liu, P. Luo, S. Qiu, X. Wang, and X. Tang, "DeepFashion: Powering Robust Clothes Recognition and Retrieval with Rich Annotations," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 1096-1104, doi: 10.1109/CVPR.2016.124.

[24] Z. Liu, S. Yan, P. Luo, X. Wang, and X. Tang, "Fashion Landmark Detection in the Wild," *ArXiv,* vol. abs/1608.03049, 2016.

[25] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Parsing clothing in fashion photographs," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, 16-21 June 2012 2012, pp. 3570-3577, doi: 10.1109/CVPR.2012.6248101.

[26] X. Liang *et al.*, "Deep Human Parsing with Active Template Regression," (in eng), *IEEE Trans Pattern Anal Mach Intell,* vol. 37, no. 12, pp. 2402-14, Dec 2015, doi: 10.1109/tpami.2015.2408360.

[27] L. Xiaodan *et al.*, "Human Parsing with Contextualized Convolutional Neural Network," (in eng), *IEEE Trans Pattern Anal Mach Intell,* vol. 39, no. 1, pp. 115-127, Jan 2017, doi: 10.1109/tpami.2016.2537339.

[28] W. Wang, Z. Zhang, S. Qi, J. Shen, Y. Pang, and L. Shao, "Learning Compositional Neural Information Fusion for Human Parsing," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 27 Oct.-2 Nov. 2019 2019, pp. 5702-5712, doi: 10.1109/ICCV.2019.00580.

[29] K. Yamaguchi, M. H. Kiapour, L. E. Ortiz, and T. L. Berg, "Retrieving Similar Styles to Parse Clothing," (in eng), *IEEE Trans Pattern Anal Mach Intell,* vol. 37, no. 5, pp. 1028-40, May 2015, doi: 10.1109/tpami.2014.2353624.

[30] L. Lo, C. Liu, R. Lin, B. Wu, H. Shuai, and W. Cheng, "Dressing for Attention: Outfit Based Fashion Popularity Prediction," in *2019 IEEE International Conference on Image Processing (ICIP)*, 22-25 Sept. 2019 2019, pp. 3222-3226, doi: 10.1109/ICIP.2019.8803461.

[31] B. Wu, W.-H. Cheng, P. Liu, Z. Zeng, and J. Luo, "SMP Challenge: An Overview of Social Media Prediction Challenge 2019," *Proceedings of the 27th ACM International Conference on Multimedia,* 2019.

[32] B. Wu, W.-H. Cheng, Y. Zhang, Q. Huang, J. Li, and T. Mei, "Sequential Prediction of Social Media Popularity with Deep Temporal Context Networks," *ArXiv,* vol. abs/1712.04443, 2017.

[33] K. Yamaguchi, T. L. Berg, and L. E. Ortiz, "Chic or Social: Visual Popularity Analysis in Online Fashion Networks," presented at the Proceedings of the 22nd ACM international conference on Multimedia, Orlando, Florida, USA, 2014. [Online]. Available: <https://doi.org/10.1145/2647868.2654958>.

[34] X. Zhang *et al.*, "Trip Outfits Advisor: Location-Oriented Clothing Recommendation," *IEEE Transactions on Multimedia,* vol. 19, no. 11, pp. 2533-2544, 2017, doi: 10.1109/TMM.2017.2696825.

[35] H. Dong, X. Liang, X. Shen, B. Wu, B. Chen, and J. Yin, "FW-GAN: Flow-Navigated Warping GAN for Video Virtual Try-On," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 27 Oct.-2 Nov. 2019 2019, pp. 1161-1170, doi: 10.1109/ICCV.2019.00125.

[36] G. Balakrishnan, A. Zhao, A. V. Dalca, F. Durand, and J. V. Guttag, "Synthesizing Images of Humans in Unseen Poses," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 8340-8348, 2018.

[37] I. Santesteban, M. A. Otaduy, and D. Casas, "Learning-Based Animation of Clothing for Virtual Try-On," *Comput. Graph. Forum,* vol. 38, pp. 355-366, 2019.

[38] M. Wedel and R. Pieters, "Eye Fixations on Advertisements and Memory for Brands: A Model and Findings," *Marketing Science,* vol. 19, no. 4, pp. 297-312, 2000/11/01 2000, doi: 10.1287/mksc.19.4.297.11794.

[39] J. E. Russo and F. Leclerc, "An Eye-Fixation Analysis of Choice Processes for Consumer Nondurables," *Journal of Consumer Research,* vol. 21, no. 2, pp. 274-290, 1994, doi: 10.1086/209397.

[40] K. Rayner, C. M. Rotello, A. J. Stewart, J. Keir, and S. A. Duffy, "Integrating text and pictorial information: eye movements when looking at print advertisements," (in eng), *J Exp Psychol Appl,* vol. 7, no. 3, pp. 219-26, Sep 2001, doi: 10.1037//1076-898x.7.3.219.

[41] Y. Xiao, G. Peng, and S. Yuling, "Advertisement Evaluation Based On Visual Attention Mechanism," 2017/01 2017: Atlantis Press, pp. 765-772, doi: <https://doi.org/10.2991/icemeet-16.2017.159>. [Online]. Available: <https://doi.org/10.2991/icemeet-16.2017.159>

[42] Z. Ma, L. Qing, J. Miao, and X. Chen, "Advertisement evaluation using visual saliency based on foveated image," in *2009 IEEE International Conference on Multimedia and Expo*, 28 June-3 July 2009 2009, pp. 914-917, doi: 10.1109/ICME.2009.5202644.

[43] A. P. Sripati and C. R. Olson, "Representing the Forest before the Trees: A Global Advantage Effect in Monkey Inferotemporal Cortex," *The Journal of Neuroscience,* vol. 29, no. 24, pp. 7788-7796, 2009, doi: 10.1523/jneurosci.5766-08.2009.

[44] T. J. Z. Bylinski, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba. "MIT Saliency benchmark." <http://saliency.mit.edu> (accessed.

[45] D. Parkhurst, K. Law, and E. Niebur, "Modeling the role of salience in the allocation of overt visual attention," (in eng), *Vision Res,* vol. 42, no. 1, pp. 107-23, Jan 2002, doi: 10.1016/s0042-6989(01)00250-4.

[46] M. Cerf, E. P. Frady, and C. Koch, "Faces and text attract gaze independent of the task: Experimental data and computer model," *Journal of Vision,* vol. 9, no. 12, 2009, doi: 10.1167/9.12.10.

[47] D. Parks, A. Borji, and L. Itti, "Augmented saliency model using automatic 3D head pose detection and learned gaze following in natural scenes," *Vision Research,* vol. 116, pp. 113-126, 2015/11/01/ 2015, doi: <https://doi.org/10.1016/j.visres.2014.10.027>.

[48] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural Networks,* vol. 19, no. 9, pp. 1395-1407, 2006/11/01/ 2006, doi: <https://doi.org/10.1016/j.neunet.2006.10.001>.

[49] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," (in eng), *Vision research,* vol. 49, no. 10, pp. 1295-1306, 2009/06// 2009, doi: 10.1016/j.visres.2008.09.007.

[50] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *Journal of Vision,* vol. 8, no. 7, pp. 32-32, 2008, doi: 10.1167/8.7.32.

[51] Y. Xie, H. Lu, and M. Yang, "Bayesian Saliency via Low and Mid Level Cues," *IEEE Transactions on Image Processing,* vol. 22, no. 5, pp. 1689-1698, 2013, doi: 10.1109/TIP.2012.2216276.

[52] X. Hou and L. Zhang, "Saliency Detection: A Spectral Residual Approach," in *2007 IEEE Conference on Computer Vision and Pattern Recognition*, 17-22 June 2007 2007, pp. 1-8, doi: 10.1109/CVPR.2007.383267.

[53] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*, 20-25 June 2009 2009, pp. 1597-1604, doi: 10.1109/CVPR.2009.5206596.

[54] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency Detection via Graph-Based Manifold Ranking," presented at the Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, 2013. [Online]. Available: <https://doi.org/10.1109/CVPR.2013.407>.

[55] J. Zhang, K. A. Ehinger, H. Wei, K. Zhang, and J. Yang, "A novel graph-based optimization framework for salient object detection," *Pattern Recognition,* vol. 64, pp. 39-50, 2017/04/01/ 2017, doi: <https://doi.org/10.1016/j.patcog.2016.10.025>.

[56] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, "Visual Saliency Based on Conditional Entropy," in *Computer Vision – ACCV 2009*, Berlin, Heidelberg, H. Zha, R.-i. Taniguchi, and S. Maybank, Eds., 2010// 2010: Springer Berlin Heidelberg, pp. 246-257.

[57] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: an information theoretic approach," (in eng), *J Vis,* vol. 9, no. 3, pp. 5.1-24, Mar 13 2009, doi: 10.1167/9.3.5.

[58] N. Riche, M. Mancas, B. Gosselin, and T. Dutoit, "Rare: A new bottom-up saliency model," in *2012 19th IEEE International Conference on Image Processing*, 30 Sept.-3 Oct. 2012 2012, pp. 641-644, doi: 10.1109/ICIP.2012.6466941.

[59] R. J. Peters and L. 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*, 17-22 June 2007 2007, pp. 1-8, doi: 10.1109/CVPR.2007.383337.

[60] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *2009 IEEE 12th International Conference on Computer Vision*, 29 Sept.-2 Oct. 2009 2009, pp. 2106-2113, doi: 10.1109/ICCV.2009.5459462.

[61] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "DeepGaze II: Reading fixations from deep features trained on object recognition," *ArXiv,* vol. abs/1610.01563, 2016.

[62] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in Context," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7-12 June 2015 2015, pp. 1072-1080, doi: 10.1109/CVPR.2015.7298710.

[63] E. Vig, M. Dorr, and D. Cox, "Large-Scale Optimization of Hierarchical Features for Saliency Prediction in Natural Images," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, 23-28 June 2014 2014, pp. 2798-2805, doi: 10.1109/CVPR.2014.358.

[64] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting Human Eye Fixations via an LSTM-Based Saliency Attentive Model," *IEEE Transactions on Image Processing,* vol. 27, no. 10, pp. 5142-5154, 2018, doi: 10.1109/TIP.2018.2851672.

[65] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," *International Journal of Computer Vision,* vol. 128, no. 2, pp. 336-359, 2020/02/01 2020, doi: 10.1007/s11263-019-01228-7.

[66] C. Szegedy *et al.*, "Going deeper with convolutions," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 7-12 June 2015 2015, pp. 1-9, doi: 10.1109/CVPR.2015.7298594.

[67] K. He, X. Zhang, S. Ren, and J. Sun, "Deep Residual Learning for Image Recognition," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 27-30 June 2016 2016, pp. 770-778, doi: 10.1109/CVPR.2016.90.

[68] S. Bernhard, P. John, and H. Thomas, "Graph-Based Visual Saliency," in *Advances in Neural Information Processing Systems 19: Proceedings of the 2006 Conference*: MITP, 2007, pp. 545-552.

[69] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of Vision,* vol. 13, no. 4, pp. 11-11, 2013, doi: 10.1167/13.4.11.

[70] H. R. Tavakoli and J. Laaksonen, "Bottom-Up Fixation Prediction Using Unsupervised Hierarchical Models," in *Computer Vision – ACCV 2016 Workshops*, Cham, C.-S. Chen, J. Lu, and K.-K. Ma, Eds., 2017// 2017: Springer International Publishing, pp. 287-302.

[71] F. L. p. García, X. R. n. Fdez-Vidal, X. Manuel Pardo, and R. Dosil, "Scene Recognition through Visual Attention and Image Features: A Comparison between SIFT and SURF Approaches," in *Object Recognition*: IntechOpen, 2011, ch. 12.

[72] A. Borji, D. N. Sihite, and L. Itti, "Quantitative Analysis of Human-Model Agreement in Visual Saliency Modeling: A Comparative Study," *IEEE Transactions on Image Processing,* vol. 22, no. 1, pp. 55-69, 2013, doi: 10.1109/TIP.2012.2210727.

[73] O. Le Meur and T. Baccino, "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behavior Research Methods,* vol. 45, no. 1, pp. 251-266, 2013/03/01 2013, doi: 10.3758/s13428-012-0226-9.

[74] A. Vedaldi and K. Lenc, "MatConvNet: Convolutional Neural Networks for MATLAB," presented at the Proceedings of the 23rd ACM international conference on Multimedia, Brisbane, Australia, 2015. [Online]. Available: <https://doi.org/10.1145/2733373.2807412>.

[75] A. Mahdi and J. 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)*, 4-7 March 2018 2018, pp. 247-250, doi: 10.1109/BHI.2018.8333415.

[76] O. Russakovsky *et al.*, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision,* vol. 115, no. 3, pp. 211-252, 2015/12/01 2015, doi: 10.1007/s11263-015-0816-y.

[77] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Saliency Benchmarking: Separating Models, Maps and Metrics," *ArXiv,* vol. abs/1704.08615, 2017.

# VITA

Graduate School

Southern Illinois University, Carbondale

Aashish Neupane

[aashish.neupane@siu.edu](mailto:aashish.neupane@siu.edu)

[neo.aashish@gmail.com](mailto:neo.aashish@gmail.com)

Tribhuvan University, Pulchowk Campus, Nepal

Bachelor of Engineering in Electronics & Communication Engineering

September 2017.

Thesis Title: Visual saliency analysis on fashion images using conventional image processing and deep learning approaches.

Major Professor: Dr. Jun Qin.