4-24-2017

# A Comparison Study of Saliency Models for Fixation Prediction on Infants and Adults

Ali Mahdi
*Southern Illinois University Carbondale*

Mei Su

Matthew Schlesinger
*Southern Illinois University Carbondale*

Jun Qin
*Southern Illinois University Carbondale*, jqin@siu.edu

Recommended Citation

Mahdi, Ali, Su, Mei, Schlesinger, Matthew and Qin, Jun. "A Comparison Study of Saliency Models for Fixation Prediction on Infants and Adults." *IEEE Transactions on Cognitive and Developmental Systems* , No. 99 (Apr 2017): 1-14. doi:10.1109/TCDS.2017.2696439.

# A Comparison Study of Saliency Models for Fixation Prediction on Infants and Adults

Ali Mahdi, *Student Member, IEEE,* Mei Su, Matthew Schlesinger, and Jun Qin, *Member, IEEE*

***Abstract*—Various saliency models have been developed over the years. The performance of saliency models is typically evaluated based on databases of experimentally recorded adult eye fixations. Although studies on infant gaze patterns have attracted much attention recently, saliency based models have not been widely applied for prediction of infant gaze patterns. In this study, we conduct a comprehensive comparison study of eight state-of-the-art saliency models on predictions of experimentally captured fixations from infants and adults. Seven evaluation metrics are used to evaluate and compare the performance of saliency models. The results demonstrate a consistent performance of saliency models predicting adult fixations over infant 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 and baselines models performance ranking, the results show that GBVS and Itti models are among the top three contenders, infants and adults have bias toward the centers of images, and all models and the center baseline model outperformed the chance baseline model.**

***Index Terms*—Fixation, visual attention, saliency models, evaluation metrics, comparison.**

## I. Introduction

HUMAN eyes receive tremendous amounts of visual information [1]. Such information represents objects of different structures at various scales. The human visual system cannot fully process the received visual information Therefore, humans use eye, head, and body movements to direct the gaze towards the object of interest in the scene to be processed, which requires a high cognitive mechanism, visual attention. In computer vision, visual attention is represented by a topological map, known as saliency map that records levels of visual attention priority. An object of interest is salient if it is rare or novel to the surroundings. A variety of applications can be benefited from saliency modeling, e.g., object detection [2][3], image segmentation [4][5], image retargeting [6][7], image/video compression [8][9], visual tracking [10][11], gaze estimation [12], robot navigation [13], image/video quality assessment [14][15], and advertising design [16].

In the past two decades, a rich stream of saliency models have been proposed based on various hypotheses (for review see [17][18][19]). Itti et al. [20] implemented a first complete saliency model inspired by the center surround operations. Oliva et al. [21] introduced a Bayesian framework for visual search tasks using biologically inspired features and scales, in which multivariate Gaussian distributions were used to compute the joint probability of a feature vector. Similarly, Torralba et al. [22] estimated the joint probability posteriorly by multivariate generalized Gaussian distributions. Chang et al. [23] proposed salient based object detection by fusing generic objects. Zhu et al. [24] proposed a robust background measure to characterize the spatial layout of the image, and then proposed an optimized framework to integrate low-level cues. Leboran et al. [25] developed a dynamic adaptive whitening saliency, which is based on high level statistical structures. Recently, Wang et al. [26] developed a saliency model by combining 13 existing state-of-the-art saliency models.

Although saliency models are generalized models, their prediction results are often evaluated by measuring the agreement between the saliency models and datasets of human adult eye fixations [27][28][29]. This study aims to evaluate saliency models over a dataset of experimentally recorded eye fixations from infants and adults using a unified framework of several evaluation metrics. This study also aims to highlight the difference between infants and adults using saliency modeling. The reason behind this comparison is that the visual attention of infants is different than the visual attention of adults. In general, infant gaze patterns lack the physical development and infants are naive in learning, while adults have reached full development stages and are more experienced in learning. More specifically, retinal development in infants indicates that infants have poor visual acuity. In the retina, fovea diameter, cone size, and density decrease rapidly during the first 45 months [30]. Also, several studies provide an evidence for structural and functional imaging development in the brain [31][32][33][34].

Although visual attention in infants is not as consistent as adults, visual attention in infants is guided by constraints that help control the gaze during visual exploration [35][36]. Previous studies suggest that fundamental features of a saliency model influence infant gaze patterns [37][38]. During the first 2 to 4 months, infants learn to focus their eyes on tracing complex contours and following moving objects to shift their gaze toward the target of interest [39]. Such systematic patterns are developed as a result of neural growth in the structure of retina and cortical areas. Between the ages of 4 and 6 months, infants develop more complex visual attention mechanism. This mechanism exploits suppression of competing information during attention-oriented shifts [40]. Infants at 9 months suppress previously cued locations in the scene after they are visited [41][42].

Bottom up saliency models can reveal some of the above

A. Mahdi and J. Qin are with the Department of Electrical and Computer Engineering, Southern Illinois University, Carbondale, IL, 62901 USA e-mail: (ali.mahdi@siu.edu; jqin@siu.edu).

M. Su is with the College of Foreign Language, Guizhou University, Guizhou, China (email: Maysu555@163.com).

M. Schlesinger is with the Department of Psychology, Southern Illinois University, Carbondale, IL, 62901 USA email: (matthews@siu.edu).

mentioned differences between infants and adults. Such models detect visual rarity or novelty in the scene. Because visual acuity in infants is not as well developed as adults, several visual rarities are suppressed as a result of lower sampling rate of the visual scene. A few studies have used saliency models to analyze visual attention in infants. In 2007, Schlesinger et al. used Itti & Koch saliency model [43] as a multi-channel, image-filtering model to simulate the development of visual attention in infants [37]. Their model extracted four low level features (luminance, color, orientation, and motion), then performed a center surround contrast enhancement. The model examined constraints that guide visual attention in infants by varying 3 parameters: oculomotor noise, horizontal connections in the V1 area, and the recurrent processing in the parietal cortex. In 2012, Althous & Mareschal [44] proposed a combination of saliency maps and area of interest to analyze learning in 12 month old infants. Their saliency maps were obtained using a saliency filtering software [45], which consists of a one layer feed forward neural network trained by human data using a support vector machines algorithm. In 2013, Schlesinger et al. [35] used a saliency model (similar description to [37], but without the motion feature)[20][43], an entropy model, a random gaze model, 9 months infants data, and adults data to generate five sets of center of gaze (COG) samples. They trained a simple recurrent network to analyze the learnability of the sets of COG. Recently, the study was extended to investigate gaze patterns [46] and sequence [47][48] in 3 months old infants using the same saliency model setup as [37].

Although saliency based models have been used to investigate visual attention in infants, there are no comprehensive evaluations of performance of existing saliency models for analysis of infant visual attention. The aim of this study is to investigate the difference of bottom up visual attention in infants and adults. This study also provides readers, a comprehensive evaluation of saliency models over infants and adults. In this study, eight state-of-the-art saliency models and two baseline models are evaluated over a dataset of infants and adults eye fixations using seven evaluation metrics. The rest of this section consists of the categorization of saliency models based on their computational framework, a review of previous comparison studies, and contributions of this study.

### A. Categorization of Saliency Models

A rich stream of saliency models has been proposed [20][23][24]. These models are different in features, frameworks, applications, and the purposes for which they are designed. Although saliency models are different from one another, they share common characteristics; therefore, saliency models can be categorized based on these characteristics.

For example, saliency models are categorized as bottom up (exogenous) and top down (endogenous). Bottom up saliency models are stimulus driven, where a saliency is defined as an irregularity or visual rarity in a scene locally, regionally, or globally [49]. Such models can explain the scene partially as majority of eye fixations are driven by tasks. Top down saliency models are task driven, where such models use prior knowledge, expectation, and reward as visual cues to locate a target of interest [17]. In another categorization factor, saliency models can be classified as space based models and object based models. There is no universal agreement on whether eye fixations attend spatial locations or objects; therefore, space or object saliency can be used for fixation prediction. From another aspect, saliency models can be categorized based on the task type. Tasks are 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 [17][18]. In this section, saliency models are categorized based on the saliency computation mechanisms.

*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 proposed [50][51][52]. Itti & Baldi [53] defined a surprise as a saliency in probabilistic terms. Surprise is obtained as the Kullback Leibler divergence (KL). Jianyong et al. [54] proposed a Bayesian framework based on BING and graph models. The model uses a binarized normed gradients model to generate a coarse conspicuity map. A graph model is constructed after super pixel image abstraction. This operation is followed by a weighting to produce a prior map. After adaptive thresholding, the observation likelihood map is computed by color histogram. The two maps are then combined via Bayesian framework. Lu et al. [55] proposed a Bayesian framework to generate a saliency map based on reconstruction error. The model first obtains dense and sparse reconstructions, then measures the reconstruction error that propagates based on the contexts obtained from K-means clustering. Pixel level saliency is obtained by integration of multi-scale reconstruction errors. A Bayesian integral reconstructs a final saliency map from the pixel level saliency maps.

*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 the feature integration theory (FIT) [56] and guided search (GS) [57]. Koch & Ulman [58] proposed a concept of a saliency model. Itti et al. [20] devised the first saliency model based on that concept. Several implementations of the model have been introduced [20][59][?][60]. The model also has been modified for several applications. For example, Itti & Koch [43] modified the first saliency model to perform a visual search for overt and covert shifts of attention. The model iteratively convolved the extracted feature maps with a 2D difference of Gaussians (DoG) filter. Also, Cerf et al. [61] modified the first saliency model by adding Viola & Johns' [62] face detection as a low level feature, then performed similar feature competition and combination to emerge a saliency map. Several other cognitive models have been proposed [63][64][65][66][67]. Cognitive models are beneficial because their further development helps in better understanding the neural processing of visual information.

*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 that optimality should be driven with respect to the end task. Guo et al. [68] proposed an attention selectivity model for automatic fixation generation in a 2D space. In their model, an activation map was created by extracting early visual features and detecting meaningful objects and 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 function based on perceptual costs and rewards. The time of fixation over the regions of interest was estimated by memory learning and decaying model. In another study, Gao et al. [69] proposed a top down discriminant saliency rooted in a decision theoretic interpretation of perception. The model detected suspicious coincidences using Barlows principle, which provides two solutions for a discriminant saliency: feature selection, and saliency detection.

*4) Spectral analysis models:* Majority of existing saliency frameworks are processed to measure irregularities in the spatial domain and in the frequency domain. Several studies used the Fourier transform and its spectral analysis to compute a saliency map [70][71][72][73]. Hou & Zhang [74] analyzed the amplitude spectrum of the Fourier transform, and proposed a spectral residual saliency model, which is independent of features, parameters, and prior knowledge. Wang & Li [75] extended the residual spectral approach by adding features based on gestalt principles to detect similarity and continuity. Li et al. [76] proposed a bottom up approach for saliency detection, and they demonstrated a convolution of the image amplitude spectrum with a low pass Gaussian kernel of appropriate size that is equivalent to a saliency detector. Besides the amplitude spectrum, Guo et al. [77] pointed out that the phase spectrum is the key to saliency modeling in the frequency domain. Then, Guo & Zhang proposed a novel multi-resolution spatiotemporal saliency detection model based on the phase spectrum of the Fourier transform [9].

*5) Graphical models:* A probabilistic framework where a graph represents a conditional independent 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 markov [78][79], dynamic Bayesian networks [80], and conditional random field (CRF) [81][82]. Yang et al. [83] proposed ranking the similarity of image elements with foreground cues or background cues via graph based manifold ranking. Zhang et al. [84] proposed a novel graph based optimization for salient object detection, which employed multiple graphs to describe the complex information in the image.

*6) Information theory models:* Models in this category measure irregularity in image locations by maximizing the information sampled from one's environment [85][86][87][76]. Such models select the most informative locations and discard redundancies. Wang et al. [88] proposed a computational model inspired by information maximization for gaze shifts prediction, which computed three filter responses as a coherent representation for reference sensory responses, fovea periphery resolution discrepancy, and visual working memory. Klein et al. [89] introduced a salient object detection method, which has a similar structure to cognitive models, but acknowledges a saliency via information theoretic concept. The model extracts features, performs center surround operations, and computes feature maps. Riche et al. [90] proposed a bottom up saliency model based on the fact that locally contrasted and globally rare features are salient. Using Otsu method, the model extracted luminance and chrominance as low-level features, and image orientations as mid level features. Then, multi-scale rarity mechanisms are performed, and scaled maps are fused and normalized.

*7) Learning based models:* Learning based models are data driven functions to select, re-weight, and integrate the input visual stimuli. Such models learn a saliency map from human fixations. The majority of models in this category use a combination of bottom up and top down features to improve fixation prediction. Learning based models can be categorized into supervised and unsupervised learning models. Supervised learning models learn a function from a labeled training data. Kienzle et al. [45] 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. [91] used low, mid, and high level features to learn a saliency model using a support machine vector.

Unsupervised learning models learn to predict from unlabeled training data, such as deep learning. Recently, several deep learning saliency models have been proposed [92][93][94]. Deep learning models are composed of multiple layers to learn representations of images with multiple levels of abstractions. Such models dramatically improved the visual attention. Vig et al. [95] proposed the first deep learning saliency model, which incorporated biologically inspired features, and used the standard learning pipeline. Kummerer et al. [96] presented a novel way to reuse existing object recognition neural networks for fixation prediction. The model used Krizhevsky network to compute filter responses and a full convolution to learn the saliency model. Further more, a probabilistic model is introduced [97], which used VGG-19 features, incorporated center bias, and used a maximum likelihood learning to train the model.

*8) Other models:* Several other saliency models do not fit to the previously mentioned categories. Syeda-Mahmood [98] proposed a saliency model based on texture feature . In their model, five attributes of regions texture were defined over four binary maps, which were linearly combined to form a saliency map. Ardizzone et al. [99] proposed a saliency model by using scale invariant feature transforms (SIFT) as local texture features. SIFT density maps were formed by measuring the density of keypoints in local image patches. Saliency was defined as the difference between the SIFT density map and the most frequent value in the map. Gao et al. [100] developed an attention model based on SIFT features and utilized bag of words to index these SIFT features. Zhang and Scarloff [101] proposed a boolean maps based saliency model. In their model, an image was decomposed to a set of binary images based on random thresholds, and then a saliency map was formed by discovering surrounded regions via topological analysis of boolean maps.

### B. Previous Comparisons

Various computational models have been proposed to predict human fixations. To measure the agreement between a computational saliency model and human fixations, several evaluation metrics and datasets have been introduced to validate the performance of developed saliency models. Since saliency models have different evaluation scheme, a few studies proposed a unified approach to comprehensively compare saliency models [102][103]. Judd et al. [104] compared 10 saliency models and 3 baseline models over a dataset of 1003 images and annotations recorded from 39 observers using 3 evaluation metrics. The center bias and blur for all models were optimized in the study. Borji et al. [105] compared 32 saliency models for prediction of fixation locations and scanpath sequence. A shuffled area under the ROC curve (sAUC) was used to analyze the models and challenges such as center bias and blurness were explored. Borji et al. [106] evaluated 35 saliency models over 54 synthetic patterns and three natural image datasets over 3 evaluation metrics. They tackled challenges of comparison, including center bias, borders effect, scores, and parameters. In another study, Borji et al. [107] compared 40 saliency models including 28 salient object detection models, 10 fixation prediction models, one objectness proposal model, and one baseline model. The comparison was conducted over six datasets using 3 evaluation metrics. Nowadays, a comparison of saliency models was conducted over two image datasets [108]. 68 saliency models and 5 baseline models were compared over the first dataset, while 22 saliency models and 5 baseline models were compared over the second dataset. All saliency models were compared using eight evaluation metrics. All existing comparison studies provide an evaluation of saliency models over datasets of adult eye fixations. There is no comparison study of saliency models for analysis of infant visual attention. Therefore, in this study, we conduct a comprehensive comparison study to evaluate the performance of selected saliency models for prediction of fixations on both infants and adults.

### C. Contributions of this study

Three contributions are presented in this study. First, this study calculates scores over seven standard and widely used evaluation metrics. Second, it demonstrates the difference between infant and adult eye fixations using eight state-of-the-art saliency models and two baseline models. Third, it presents the performance ranking of saliency models for infants and adults.

## II. Methods and Materials

### A. Computational saliency models

In this study, 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 [20] 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 study includes a slight blur as a final step [59].

2) Graph based visual saliency model (GBVS) [59] 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 [109] 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 [74] 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 [110] extracts three features: super-pixel color, closed shapes, and center bias. Then detects salient regions using contour energy computation.

6) SUN [50] model 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 [111] model 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 [112] 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 1 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 [113].

### B. 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 images human is presented in the foreground, while

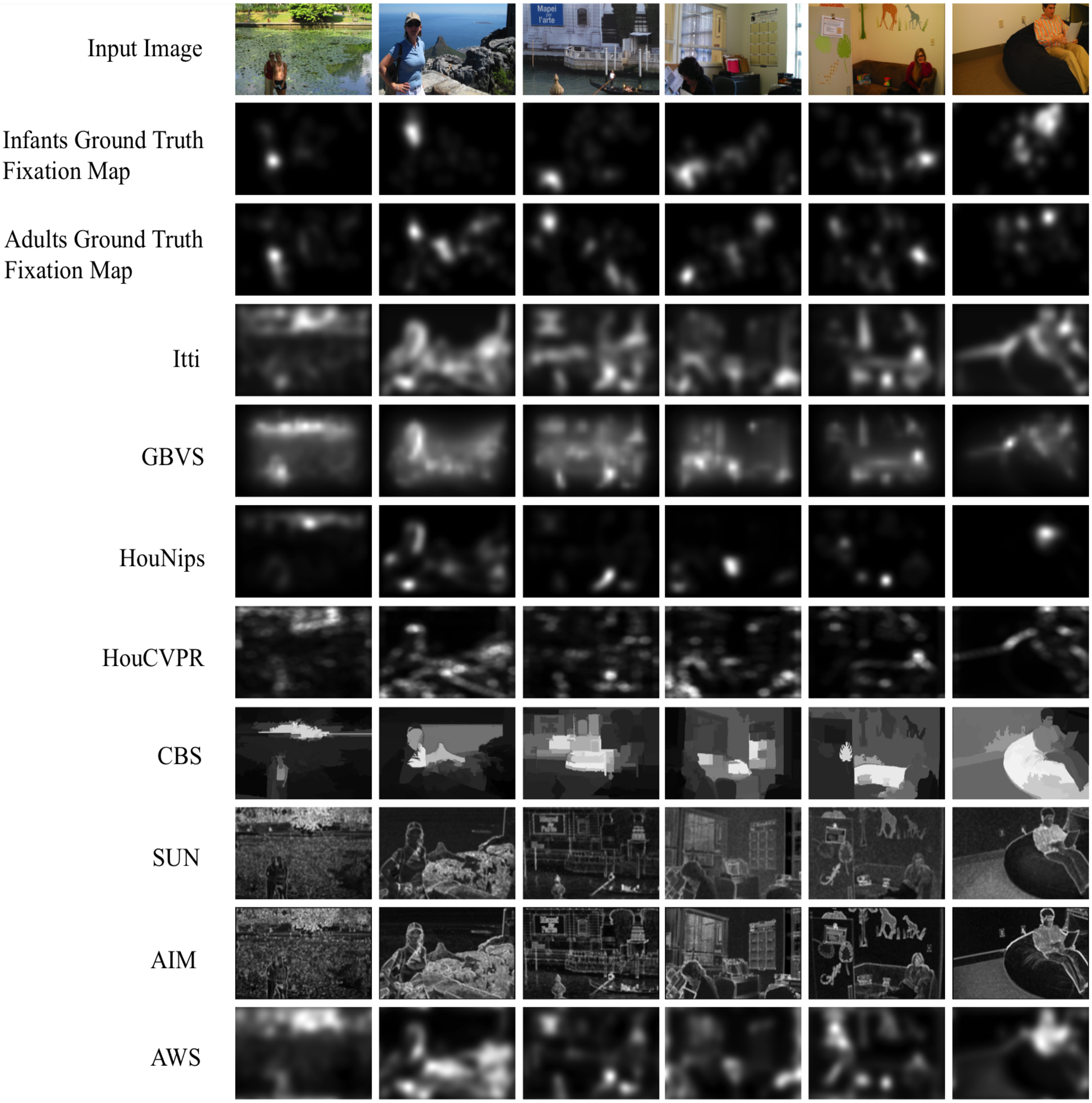
Input Image
Infants Ground Truth Fixation Map
Adults Ground Truth Fixation Map
Itti
GBVS
HouNips
HouCVPR
CBS
SUN
AIM
AWS
Mapei l'arte

Fig. 1: 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.

in some other images human is presented in the background. The size of each image is 1050 × 1680 pixels.

### C. Protocol of Experiments

In this study, 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 [114][115][116].

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 study 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.

Fig. 2 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.

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 [114].

### D. Evaluation measures

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 study, 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 drawn conclusions are independent of the choice of metric and consistent across all metrics. Generally, a good saliency model should perform well across all metrics.

TABLE I: A description of seven evaluation metrics.

| Metric | Denoted as | Theoretical range |
|---|---|---|
| Area under the ROC curve | AUC | [0,1] |
| F measure | F-measure | [0,1] |
| Information gain | IG | $[-\infty,\infty]$ |
| Similarity | SIM | [0,1] |
| Pearson's correlation coefficient | CC | [-1,1] |
| Kullback leibler divergence | KL | $[0,\infty]$ |
| Earth mover's distance | EMD | $[0,\infty]$ |

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 study [113]. The similarity based metrics are: information gain (IG), similarity (SIM), and Pearsons correlation coefficient (CC). The dissimilarity based metrics are: Kullback Leibler divergence (KL), and earth movers distance (EMD). Table 1 summarizes the seven evaluation metrics used in this study.

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

$$TPR = \frac{TP}{TP + FN} \tag{1}$$

$$FPR = \frac{FP}{FP + TN} \tag{2}$$

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

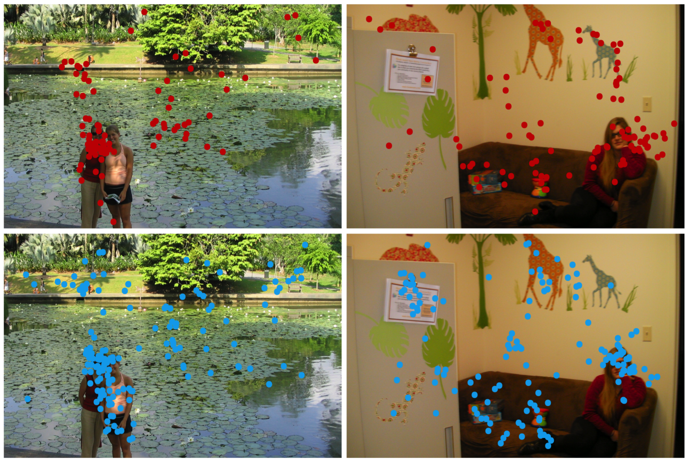

Fig. 2: Two representative images of gaze patterns of infants (top images) and adults (bottom images) over an indoor and outdoor scenes. Red and blue circles highlight the fixation locations for infants (red) and adults (blue).

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

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

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

*AUC:* The integral of the area under the ROC curve. A score higher than 0.5 indicates a prediction higher than random guessing.

*F-Measure:* A weighted harmonic mean of precision and recall. It is 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 $\beta$ is a threshold. $\beta^2 = 0.3$ to raise more importance to precision [107]. A $\beta^2$ is computed across the thresholds. Then, the maximum $\beta^2$ represent the maximum overlap between $precision$ and $recall$ along the curve. A score closer to 1 indicates that the overlap between the predicted saliency map and the ground-truth fixation map is large.

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

$$IG(S, G(x,y)) = \frac{1}{N}\sum_i G(x,y)_i [log_2(\epsilon + P_i) - log_2(\epsilon + B_i)] \tag{6}$$

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 $\epsilon$ is a small value for regularization. A center bias map emerges by averaging the ground-truth fixation maps of all other images to create a center bias of the dataset. A positive IG score indicates that saliency model prediction outperforms the center bias map. A negative IG score indicates the saliency model prediction cannot compete with the center bias map.

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

$$SIM(S, G) = \sum_i min(S_i, G_i) \tag{7}$$

where

$$\sum_i S_i = \sum_i G_i = 1$$

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

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

$$CC(S, G) = \frac{cov(S, G)}{\sigma(S)\sigma(G)} \tag{8}$$

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

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

$$KL(S, G) = \sum_i G_i log(\epsilon + \frac{G_i}{\epsilon + S_i}) \tag{9}$$

where $\epsilon$ is a regularization constant. As dissimilarity metric, a KL score of 0 indicates that the saliency map and the ground-truth fixation map are identical.

*EMD:* 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:

$$\widehat{EMD} = (min \sum_{i,j} f_{ij} d_{ij}) + |\sum_i S_i - \sum_j G_j| max d_{ij} \tag{10}$$

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

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

where $f_{ij}$ is the flow transported from supply $i$ to demand $j$, and $d_{ij}$ is the ground distance (cost) between bin $i$ and bin $j$ in the distribution. A EMD score of 0 indicates that 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 is increased.

## III. Results and Discussions

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.

### A. Analysis over infants and adults

Fig. 3 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 study, 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.

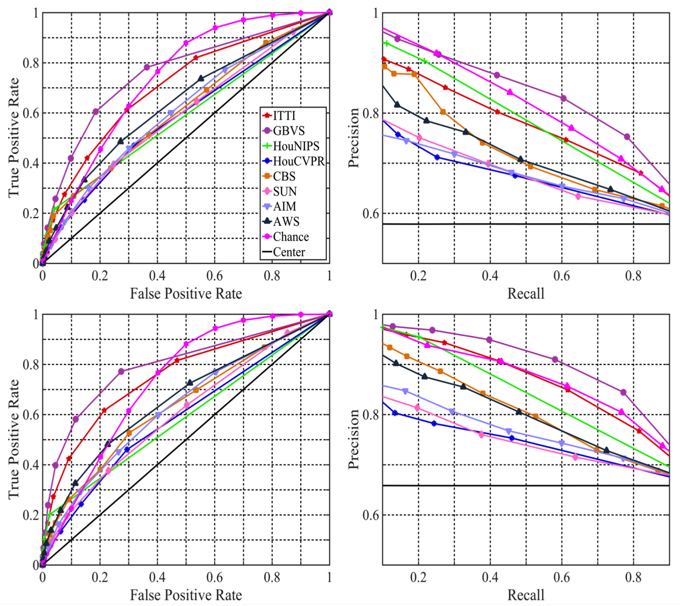

Fig. 3: Averaged ROC and PR curves of eight saliency models and two baseline models over infants (top charts) and adults (bottom charts).

To summarize the performance of saliency models fixation prediction over the infant and adult fixations, Fig. 4 presents the AUC score and F-measure over the infants and adults data. In Fig. 4, 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 (Fig. 4 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.

Fig. 5 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 Fig. 5 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.

Furthermore, the SIM score (Fig. 5 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 Fig. 5 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.

Two dissimilarity measures are presented in Fig. 6. In the left chart of Fig. 6, 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 Fig. 6, 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 infants distribution of fixations.

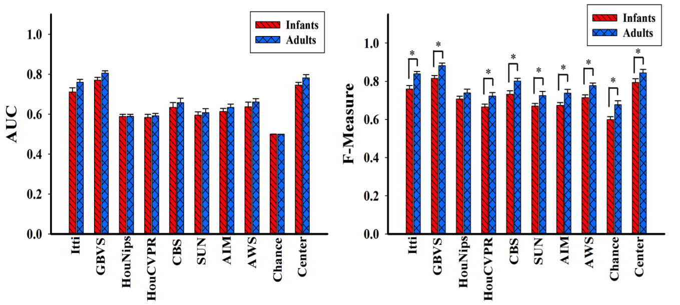

Fig. 4: Averaged AUC score and F-measure for infants and adults. A * indicates statistical significance using t-test (95%, $p \leq 0.05$). Error bars indicate standard error of the mean (SEM).

### B. Analysis over infants

Table 2 presents the ranking of saliency models for infant fixation prediction over the image dataset. Although the

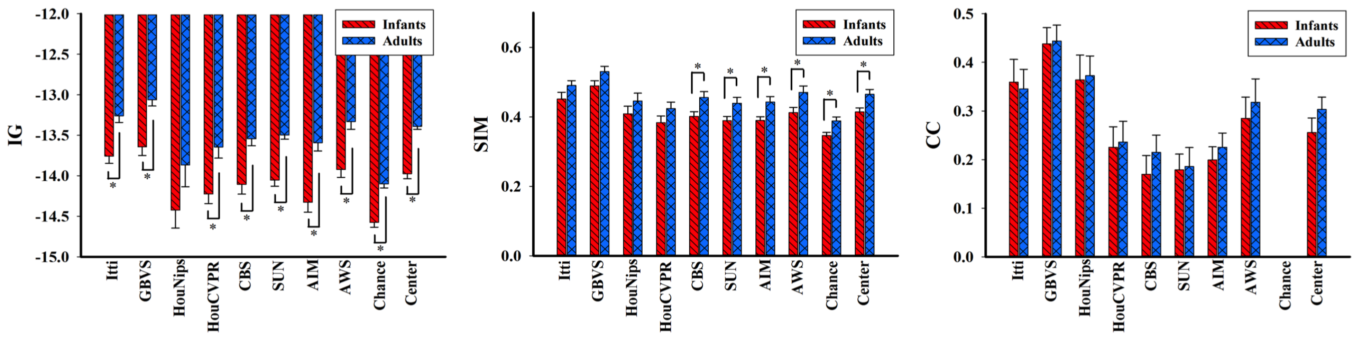

Fig. 5: Averaged IG, SIM, and CC scores for infants and adults. A * indicates statistical significance using t-test (95%, $p \leq 0.05$). Error bars indicate SEM.

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 [117][118][119]. 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.

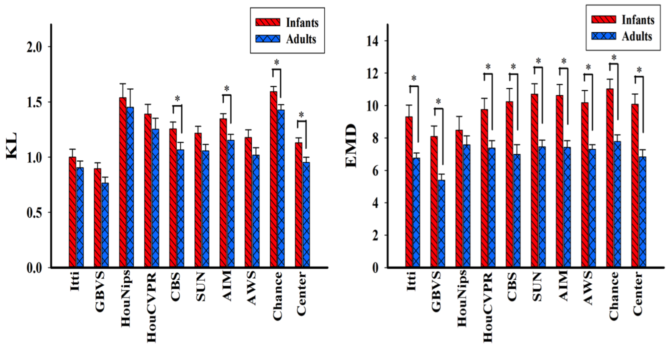

Fig. 6: Averaged KL and EMD scores for infants and adults. A * indicates statistical significance using t-test (95%, $p \leq 0.05$). Error bars indicate SEM.

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 study. 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 infants 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.

### C. Analysis over adults

Table 3 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 thats 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

TABLE II: 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 |

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 study.

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 study. 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.

### D. *Discussions of different datasets*

The results over infants and adults demonstrate several differences between infants and adults visual attention. Such results were concluded with 16 images only. To justify the conclusions of the experimental results, MIT1003 dataset [104] 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 adults 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. Fig. 7 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 adults 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 beacuse the center bias map calculated for the MIT1003 dataset is an average map of larger number of images than the infants and adults dataset.

## IV. Conclusion

In this study, 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 between infants and adults. The ranking of eight saliency and two baseline models over both infants and adults are also provided in this study. 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 study 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 study, 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 study 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.

## Acknowledgment

The authors would like to thank Dr. Dima Amso at Brown University for providing the human experimental data used in this study.

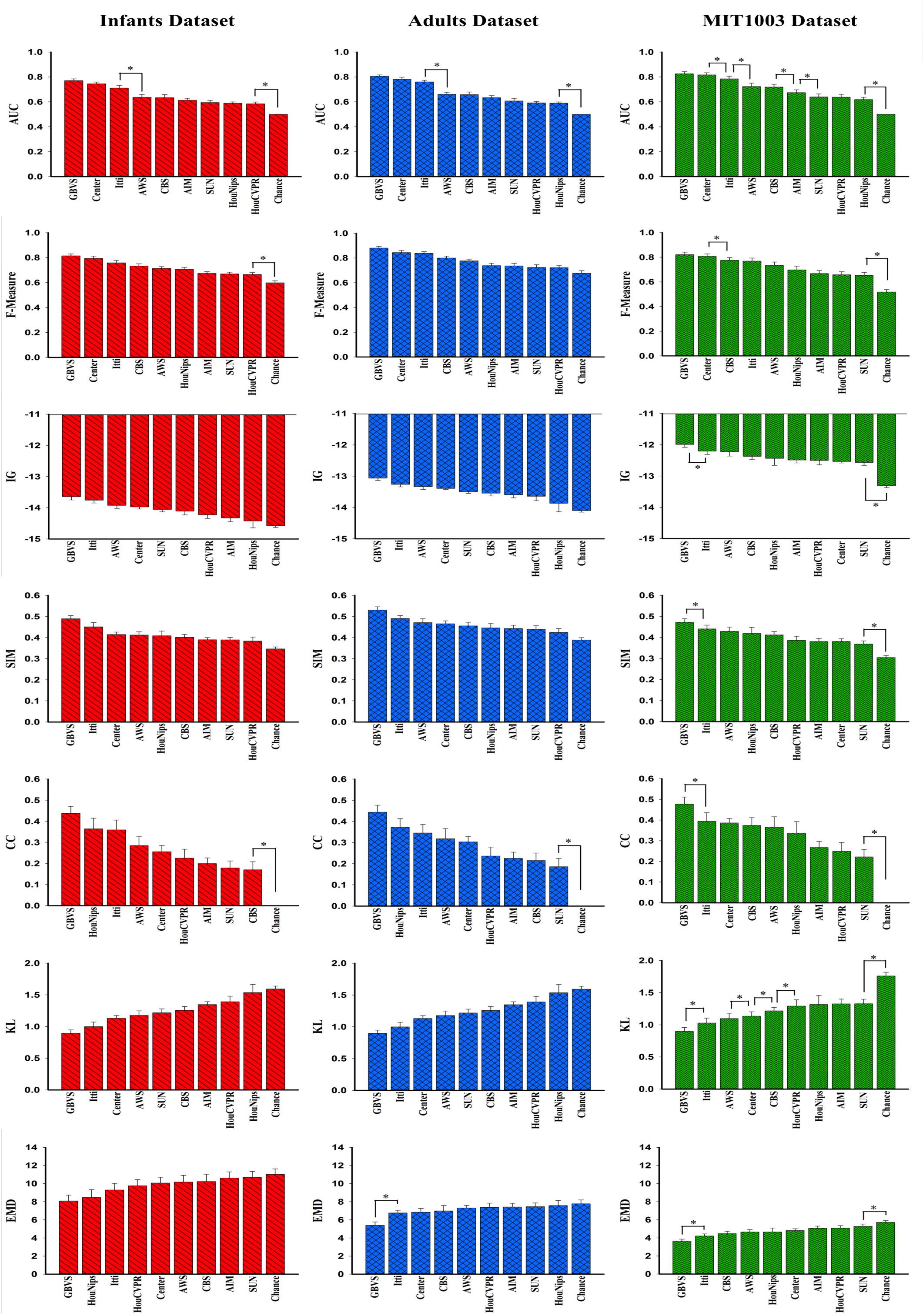

Fig. 7: Ranking visual saliency models over infants (red bar charts) and adults (blue chart bars) dataset, and a subset of 85 images (green blue charts) from the MIT1003 dataset using seven evaluation metrics: AUC, F-measure, IG, SIM, CC, KL, and EMD. A * indicates statistical significance using t-test (95%, $p \leq 0.05$) 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.

TABLE III: 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 |

## REFERENCES

[1] K. Koch, J. McLean, R. Segev, M. A. Freed, M. J. Berry, V. Balasubramanian, and P. Sterling, "How much the eye tells the brain," *Current Biology*, vol. 16, no. 14, pp. 1428–1434, 2006.

[2] N. J. Butko and J. R. Movellan, "Optimal scanning for faster object detection," in *Computer vision and pattern recognition*, 2009, pp. 2751–2758.

[3] 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.

[4] A. K. Mishra and Y. Aloimonos, "Active segmentation," *International Journal of Humanoid Robotics*, vol. 6, no. 03, pp. 361–386, 2009.

[5] 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.

[6] L. Marchesotti, C. Cifarelli, and G. Csurka, "A framework for visual saliency detection with applications to image thumbnailing," in *International Conference on Computer Vision*, 2009, pp. 2232–2239.

[7] B. Suh, H. Ling, B. B. Bederson, and D. W. Jacobs, "Automatic thumbnail cropping and its effectiveness," in *Proceedings of the 16th annual ACM symposium on User interface software and technology*, 2003, pp. 95–104.

[8] 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.

[9] 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.

[10] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Computer Vision and Pattern Recognition*, 2009, pp. 1007–1013.

[11] S. Frintrop, E. Rome, and H. I. Christensen, "Computational visual attention systems and their cognitive foundations: A survey," *ACM Transactions on Applied Perception (TAP)*, vol. 7, no. 1, p. 6, 2010.

[12] Y. Sugano, Y. Matsushita, and Y. Sato, "Appearance-based gaze estimation using visual saliency," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 2, pp. 329–341, 2013.

[13] 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.

[14] Q. Ma, L. Zhang, and B. Wang, "New strategy for image and video quality assessment," *Journal of Electronic Imaging*, vol. 19, no. 1, pp. 011 019–011 019, 2010.

[15] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric," in *International Conference on Image Processing*, vol. 2, 2007, pp. II–169.

[16] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," in *ACM transactions on graphics (TOG)*, vol. 27, no. 3, 2008, p. 16.

[17] 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.

[18] Q. Zhao and C. Koch, "Learning saliency-based visual attention: A review," *Signal Processing*, vol. 93, no. 6, pp. 1401–1407, 2013.

[19] S. Filipe and L. A. Alexandre, "From the human visual system to the computational models of visual attention: a survey," *Artificial Intelligence Review*, vol. 39, no. 1, pp. 1–47, 2013.

[20] L. Itti, C. Koch, E. Niebur *et al.*, "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.

[21] A. Oliva, A. Torralba, M. S. Castelhano, and J. M. Henderson, "Top-down control of visual attention in object detection," in *international conference on image processing*, vol. 1, 2003, pp. I–253.

[22] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search." *Psychological review*, vol. 113, no. 4, p. 766, 2006.

[23] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *International Conference on Computer Vision*, 2011, pp. 914–921.

[24] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Computer vision and pattern recognition*, 2014, pp. 2814–2821.

[25] V. Leboran, A. Garcia-Diaz, X. Fdez-Vidal, and X. Pardo, "Dynamic whitening saliency," *IEEE Transactions on pattern analysis and machine intelligence*, 2016.

[26] J. Wang, A. Borji, C.-C. J. Kuo, and L. Itti, "Learning a combined model of visual saliency for fixation prediction," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1566–1579, 2016.

[27] A. Borji and L. Itti, "Cat2000: A large scale fixation dataset for boosting saliency research," *arXiv preprint arXiv:1505.03581*, 2015.

[28] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "Salicon: Saliency in context," in *Computer vision and pattern recognition*. IEEE, 2015, pp. 1072–1080.

[29] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *European Conference on Computer Vision*. Springer, 2010, pp. 30–43.

[30] C. Yuodelis and A. Hendrickson, "A qualitative and quantitative analysis of the human fovea during development," *Vision research*, vol. 26, no. 6, pp. 847–855, 1986.

[31] N. Gogtay, J. N. Giedd, L. Lusk, K. M. Hayashi, D. Greenstein, A. C. Vaituzis, T. F. Nugent, D. H. Herman, L. S. Clasen, A. W. Toga *et al.*, "Dynamic mapping of human cortical development during childhood through early adulthood," *Proceedings of the National academy of Sciences of the United States of America*, vol. 101, no. 21, pp. 8174–8179, 2004.

[32] E. R. Sowell, P. M. Thompson, C. M. Leonard, S. E. Welcome, E. Kan, and A. W. Toga, "Longitudinal mapping of cortical thickness and brain growth in normal children," *The Journal of neuroscience*, vol. 24, no. 38, pp. 8223–8231, 2004.

[33] A. Raznahan, P. W. Shaw, J. P. Lerch, L. S. Clasen, D. Greenstein, R. Berman, J. Pipitone, M. M. Chakravarty, and J. N. Giedd, "Longitudinal four-dimensional mapping of subcortical anatomy in human development," *Proceedings of the National Academy of Sciences*, vol. 111, no. 4, pp. 1592–1597, 2014.

[34] S. N. Vandekar, R. T. Shinohara, A. Raznahan, D. R. Roalf, M. Ross, N. DeLeo, K. Ruparel, R. Verma, D. H. Wolf, R. C. Gur *et al.*, "Topologically dissociable patterns of development of the human cerebral cortex," *The Journal of Neuroscience*, vol. 35, no. 2, pp. 599–609, 2015.

[35] M. Schlesinger and D. Amso, "Image free-viewing as intrinsically-motivated exploration: estimating the learnability of center-of-gaze image samples in infants and adults," vol. 4, pp. 1–12, 2013.

[36] M. Schlesinger, D. Amso, and S. P. Johnson, "Increasing spatial competition enhances visual prediction learning," in *Development and Learning (ICDL)*, vol. 2. IEEE, 2011, pp. 1–6.

[37] M. Schlesinger, D. Amso, and S. P. Johnson, "The neural basis for visual selective attention in young infants: A computational account," *Adaptive Behavior*, vol. 15, no. 2, pp. 135–148, 2007.

[38] M. Schlesinger, D. Amso, and S. P. Johnson, "Simulating the role of visual selective attention during the development of perceptual completion," *Developmental science*, vol. 15, no. 6, pp. 739–752, 2012.

[39] M. Schlesinger, "Investigating the origins of intrinsic motivation in human infants," in *Intrinsically motivated learning in natural and artificial systems*. Springer, 2013, pp. 367–392.

[40] D. Amso and S. P. Johnson, "Development of visual selection in 3-to 9-month-olds: Evidence from saccades to previously ignored locations," *Infancy*, vol. 13, no. 6, pp. 675–686, 2008.

[41] M. L. Dixon, P. D. Zelazo, and E. De Rosa, "Evidence for intact memory-guided attention in school-aged children," *Developmental Science*, vol. 13, no. 1, pp. 161–169, 2010.

[42] D. Amso and G. Scerif, "The attentive brain: insights from developmental cognitive neuroscience," *Nature Reviews Neuroscience*, vol. 16, no. 10, pp. 606–619, 2015.

[43] L. Itti and C. Koch, "A saliency-based search mechanism for overt and covert shifts of visual attention," *Vision research*, vol. 40, no. 10, pp. 1489–1506, 2000.

[44] N. Althaus and D. Mareschal, "Using saliency maps to separate competing processes in infant visual cognition," *Child development*, vol. 83, no. 4, pp. 1122–1128, 2012.

[45] W. Kienzle, M. O. Franz, B. Schölkopf, and F. A. Wichmann, "Center-surround patterns emerge as optimal predictors for human saccade targets," *Journal of vision*, vol. 9, no. 5, pp. 7–7, 2009.

[46] M. Schlesinger, S. P. Johnson, and D. Amso, "Prediction-learning in infants as a mechanism for gaze control during object exploration," *Frontiers in psychology*, vol. 5, 2014.

[47] M. Schlesinger, S. P. Johnson, and D. Amso, "Learnability of infants' center-of-gaze sequences predicts their habituation and posthabituation looking time," in *Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2014, pp. 275–280.

[48] M. Schlesinger, S. P. Johnson, and D. Amso, "Do infants' gaze sequences predict their looking time? testing the sequential-learnability model," in *Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*, 2015, pp. 162–167.

[49] H. Nothdurft, "Salience of feature contrast," 2005.

[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.

[51] L. Zhang, M. H. Tong, and G. W. Cottrell, "Sunday: Saliency using natural statistics for dynamic analysis of scenes," in *Proceedings of the 31st annual cognitive science conference*. AAAI Press Cambridge, MA, 2009, pp. 2944–2949.

[52] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Transactions on Image Processing*, vol. 22, no. 5, pp. 1689–1698, 2013.

[53] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[54] L. Jianyong, T. Zhenmin, and X. Wei, "Improved bayesian saliency detection based on bing and graph model," *Open Cybernetics & Systemics Journal*, vol. 9, pp. 648–656, 2015.

[55] H. Lu, X. Li, L. Zhang, X. Ruan, and M.-H. Yang, "Dense and sparse reconstruction error based saliency descriptor," *IEEE Transactions on Image Processing*, vol. 25, no. 4, pp. 1592–1603, 2016.

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

[57] J. M. Wolfe, K. R. Cave, and S. L. Franzel, "Guided search: an alternative to the feature integration model for visual search." *Journal of Experimental Psychology: Human perception and performance*, vol. 15, no. 3, p. 419, 1989.

[58] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of intelligence*. Springer, 1987, pp. 115–141.

[59] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, 2006, pp. 545–552.

[60] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[61] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," in *Advances in neural information processing systems*, 2008, pp. 241–248.

[62] P. Viola and M. J. Jones, "Robust real-time face detection," *International journal of computer vision*, vol. 57, no. 2, pp. 137–154, 2004.

[63] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model the bottom-up visual attention." *IEEE transactions on pattern analysis and machine intelligence*, vol. 28, pp. 802–817, 2006.

[64] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision research*, vol. 47, no. 19, pp. 2483–2498, 2007.

[65] G. Kootstra, A. Nederveen, and B. De Boer, "Paying attention to symmetry," in *British Machine Vision Conference*, 2008, pp. 1115–1125.

[66] S. Marat, T. H. Phuoc, L. Granjon, N. Guyader, D. Pellerin, and A. Guérin-Dugué, "Modelling spatio-temporal saliency to predict gaze direction for short videos," *International journal of computer vision*, vol. 82, no. 3, pp. 231–243, 2009.

[67] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Computer vision and pattern recognition*, 2011, pp. 433–440.

[68] C. Guo and L. Zhang, "An attention selection model with visual memory and online learning," in *International Joint Conference on Neural Networks*, 2007, pp. 1295–1301.

[69] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Advances in neural information processing systems*, 2004, pp. 481–488.

[70] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Computer vision and pattern recognition*, 2009, pp. 1597–1604.

[71] P. Bian and L. Zhang, "Biological plausibility of spectral domain approach for spatiotemporal visual saliency," in *International Conference on Neural Information Processing*. Springer, 2008, pp. 251–258.

[72] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 4, pp. 996–1010, 2013.

[73] L. Xiao, C. Li, Z. Hu, and Z. Pan, "Multi-scale spectrum visual saliency perception via hypercomplex dct," in *International Conference on Intelligent Computing*. Springer, 2016, pp. 645–655.

[74] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Computer vision and pattern recognition*, 2007, pp. 1–8.

[75] Z. Wang and B. Li, "A two-stage approach to saliency detection in images," in *International Conference on Acoustics, Speech and Signal Processing*, 2008, pp. 965–968.

[76] Y. Li, Y. Zhou, J. Yan, Z. Niu, and J. Yang, "Visual saliency based on conditional entropy," in *Asian Conference on Computer Vision*. Springer, 2009, pp. 246–257.

[77] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform," in *Computer vision and pattern recognition*, 2008, pp. 1–8.

[78] A. A. Salah, E. Alpaydin, and L. 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*, vol. 24, no. 3, pp. 420–425, 2002.

[79] L. Huang, S. Tang, J. Hu, and W. Deng, "Saliency region detection via graph model and statistical learning," in *Chinese Conference on Pattern Recognition*. Springer, 2016, pp. 3–13.

[80] R. P. Rao, "Bayesian inference and attentional modulation in the visual cortex," *Neuroreport*, vol. 16, no. 16, pp. 1843–1848, 2005.

[81] J. Sun, T. Liu, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *Computer cision and pattern recognition*, 2007, pp. 1–8.

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

[83] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Computer vision and pattern recognition*, 2013, pp. 3166–3173.

[84] 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.

[85] L. W. Renninger, J. M. Coughlan, P. Verghese, and J. Malik, "An information maximization model of eye movements," in *Advances in neural information processing systems*, 2004, pp. 1121–1128.

[86] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *Journal of vision*, vol. 9, no. 12, pp. 15–15, 2009.

[87] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of vision*, vol. 9, no. 3, pp. 5–5, 2009.

[88] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating human saccadic scanpaths on natural images," in *Computer vision and pattern recognition*, 2011, pp. 441–448.

[89] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *International Conference on Computer Vision*, 2011, pp. 2214–2219.

[90] N. Riche, M. Mancas, B. Gosselin, and T. Dutoit, "Rare: A new bottom-up saliency model," in *International Conference on Image Processing*, 2012, pp. 641–644.

[91] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *International Conference on Computer Vision*, 2009, pp. 2106–2113.

[92] H. Kato and T. Harada, "Visual language modeling on cnn image representations," *arXiv preprint arXiv:1511.02872*, 2015.

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

[94] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Computer vision and pattern recognition*, 2015, pp. 362–370.

[95] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Computer vision and pattern recognition*, 2014, pp. 2798–2805.

[96] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," *arXiv preprint arXiv:1411.1045*, 2014.

[97] M. Kümmerer, T. S. Wallis, and M. Bethge, "Deepgaze ii: Reading fixations from deep features trained on object recognition," *arXiv preprint arXiv:1610.01563*, 2016.

[98] T. F. Syeda-Mahmood, "Detecting perceptually salient texture regions in images," *Computer Vision and Image Understanding*, vol. 76, no. 1, pp. 93–108, 1999.

[99] E. Ardizzone, A. Bruno, and G. Mazzola, "Visual saliency by keypoints distribution analysis," in *International Conference on Image Analysis and Processing*. Springer, 2011, pp. 691–699.

[100] K. Gao, S. Lin, Y. Zhang, S. Tang, and H. Ren, "Attention model based sift keypoints filtration for image retrieval," in *Computer and Information Science, 2008. ICIS 08. Seventh IEEE/ACIS International Conference on*. IEEE, 2008, pp. 191–196.

[101] J. Zhang and S. Sclaroff, "Exploiting surroundedness for saliency detection: a boolean map approach," *IEEE transactions on pattern analysis and machine intelligence*, vol. 38, no. 5, pp. 889–902, 2016.

[102] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of vision*, vol. 8, no. 7, pp. 13–13, 2008.

[103] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of vision*, vol. 11, no. 3, pp. 9–9, 2011.

[104] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *MIT Technical Report*, 2012.

[105] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *International conference on computer vision*, 2013, pp. 921–928.

[106] 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.

[107] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[108] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark."

[109] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in neural information processing systems*, 2009, pp. 681–688.

[110] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior." in *British Machine Vision Conference*, vol. 6, no. 7, 2011, p. 9.

[111] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in neural information processing systems*, 2005, pp. 155–162.

[112] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Decorrelation and distinctiveness provide with human-like saliency," in *International Conference on Advanced Concepts for Intelligent Vision Systems*. Springer, 2009, pp. 343–354.

[113] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *arXiv preprint arXiv:1604.03605*, 2016.

[114] 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.

[115] D. Amso, S. Haas, and J. Markant, "An eye tracking investigation of developmental change in bottom-up attention orienting to faces in cluttered natural scenes," *PloS one*, vol. 9, no. 1, p. e85701, 2014.

[116] A. Mahdi, M. Schlesinger, D. Amso, and J. Qin, "Infants gaze pattern analyzing using contrast entropy minimization," in *Development and Learning and Epigenetic Robotics (ICDL-EpiRob)*. IEEE, 2015, pp. 106–111.

[117] D. J. Parkhurst and E. Niebur, "Scene content selected by active vision," *Spatial vision*, vol. 16, no. 2, pp. 125–154, 2003.

[118] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: effects of scale and time," *Vision research*, vol. 45, no. 5, pp. 643–659, 2005.

[119] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *Journal of vision*, vol. 7, no. 14, pp. 4–4, 2007.