Hindawi
Mobile Information Systems
Volume 2022, Article ID 3279325, 15 pages
https://doi.org/10.1155/2022/3279325

Research Article

Deep Global-Local Gazing: Including Global Scene Properties in Local Saliency Computation

Samad Zabihi, Mehran Yazdi, and Eghbal Mansoori

School of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran

Correspondence should be addressed to Mehran Yazdi; yazdi@shirazu.ac.ir

Received 23 April 2022; Revised 10 June 2022; Accepted 25 June 2022; Published 25 August 2022

Academic Editor: Sebastian Podda

Visual saliency models imitate the attentive mechanism of the human visual system (HVS) to detect the objects that stand out from their neighbors in the scene. Some biological phenomena in HVS, such as contextual cueing effects, suggest that the contextual information of the whole scene does guide the attentive mechanism. The saliency value of each image patch is influenced by its visual (local) features as well as the contextual information of the whole scene. Modern saliency models are based on deep convolutional neural networks. Because the convolutional operators operate locally and use weight sharing, such networks inherently have difficulty capturing global and location-dependent features. In addition, these models calculate the saliency value pixel-wise using local features. Therefore, it is necessary to provide global features along with local features. In this regard, we propose two approaches for capturing the contextual information from the scene. In our first method, we introduce a shift-variant fully connected component to capture global and location-dependent information. Instead of using the native CNN of our base model, in our second method, we use a VGGNet to capture the global and context information of the scene. To show the effectiveness of our methods, we use them to extend the SAM-ResNet saliency model. To evaluate our proposed approaches, four challenging saliency benchmark datasets were used. The experimental results showed that our methods could outperform the existing state-of-the-art saliency prediction models.

1. Introduction

In looking at a scene, HVS tends to gaze at the salient regions of the scene and ignores less salient parts [1]. The perceptual and cognitive resources of humans are limited. This attentive mechanism leads humans to rapidly process visual information and assign these limited resources to the salient subsets or objects of scenes [2]. This ability of HVS has been studied by neuroscientists and computer vision researchers to develop models that emulate this attention mechanism. Saliency prediction models are helpful to figure out human attention mechanisms, and also predict where people focus when they look at images or watch videos [2, 3]. Visual saliency models are useful across domains such as advertising, robotics, auto-driving [4], defense, game, assistive systems, and human–computer interaction.

In general terms, early saliency prediction models employed biologically motivated low-level features [5–8] driven from low-level stimuli, for example, color, intensity, orientation, and texture. Subsequent models incorporated semantic concepts such as face [9], text [10], and gaze direction [11]. However, these techniques are not able to generally incorporate high-level features (e.g., contextual information, center prior, and complex objects) and inherent correlation of various visual subsets in a scene (e.g., correlation of eyes, nose, ears, and mouth).

Since 2014, new sort of visual saliency models based on deep neural networks (DNNs) has emerged. They achieved strong improvements over classic saliency models. The hierarchical deep structure of convolutional neural networks (CNNs) enables the salience models to capture some complex cues whereas pioneering saliency models were not able to learn from data. However, studies revealed that they continue to fall short in capturing some high-level features of the scene such as center prior and global context [2]. Some studies have tried to compensate for such deficiencies in the

CNN structure and capture the global properties by incorporating center prior [12–15] or by the means of convolutional long short-term memory (LSTM) [14, 16].

Here, we propose two methods that incorporate contextual cues, global properties, and location-dependent features into pixel-wise saliency prediction to compensate for some deficiencies in CNN-based saliency models. Our first approach employs the VGGNet structure to capture the global and contextual information of the scene and then incorporates this information into the pixel-wise saliency prediction. This model predicts the saliency value of each image patch, by taking into account not only the locally extracted features of that patch but also the global scene properties. In our second approach, we introduce a shift-variant fully connected component to combine locally extracted features and learn the location-dependent information that simple convolution layers are incapable of capturing.

The remainder of the study is organized as follows: the next section discusses related saliency prediction models. Section 3 discusses the facts and arguments that motivated us and supported our saliency modeling. Section 4 presents our proposed saliency models. Section 5 describes some popular evaluation metrics, evaluation baselines, and saliency datasets, and provides our implementation details. Section 6 reports the evaluation results of our proposed models over several saliency benchmark datasets. Finally, Section 7 presents evaluation results and in Section 8 we conclude.

2. Related Work

The saliency prediction spreads in several research areas: free-viewing gaze prediction [4, 9–12] which tries to model the human eye fixations under free-viewing conditions, egocentric gaze prediction [13–15] which aims to predict the human eye fixations during a specific task, salient object detection [17–20] that detects and segments the salient objects in scenes and co-saliency detection [21, 22] that detects the common salient objects from a group of images. These visual saliency models are useful in vision tasks such as object recognition [17], image/video retargeting [18], segmentation [19–22], visual tracking [23], and image compression [24]. As an instance, Wang et al. [21] proposed a method that uses object-level cues for unsupervised video object segmentation.

2.1. Classic Saliency Prediction Models. Pioneer saliency prediction methods were mostly inspired by psychological and psychophysical models of attention as studied in HVS and they mainly focused on extracting better-handcrafted features and using better learning methods. Many of these bottom-up saliency models were based on Treisman's "the feature integration theory" [25], which proposed strategies for combining various kinds of visual features without any bias to find the salient subsets of the scene. In 1985, Koch and Ullman [26] were one of the first to use the feature integration theory to propose a feed-forward model for combing a set of maps of elementary cues like contrast, color, and motion to produce a map of saliency.

In 1998, Itti et al. [5] proposed an approach based on the Koch and Ullman feed-forward model [26]. In their model, they computed multi-scale center-surround contrast maps of preattentive features and then integrated these contrast maps to predict the saliency map. Their work triggered a lot of interest in the visual salience community. Many saliency models such as adapted this center-surround structure in the spatial domain [27, 28]. Itti and Baldi [29] proposed a model based on Bayesian approaches. Some methods adopted an information-theoretic justification for attentive selection [30–32]. Harel et al. [8] proposed a saliency model based on graph theory. Hou and Zhang [33] calculated saliency from frequency analysis. Some traditional saliency models used machine learning algorithms [34–36]. Some of these models incorporate high-level features such as face or text to steer the top-down process, thus they may not be purely bottom-up [37]. While many models fall into the bottom-up saliency model category, these models fail to capture the factors that contribute to attentional selection.

2.2. Deep Saliency Prediction Models. Employing deep convolutional neural networks (CNNs) in the saliency prediction model has made some drastic improvements over well-established saliency benchmark datasets [2]. Since 2014, using DNNs for saliency prediction gained much attention. To compensate for the lack of sufficiently large fixation data, most of these DNN-based models use transfer learning by employing a pretrained model that was trained for similar/different visual tasks on large image datasets.

One of the first saliency models that used DNN was proposed by Vig et al. [38]. Their model, ensemble of deep networks (eDN), generates a large number of richly-parameterized neuromorphic networks for the feature extraction phase. Then, extracted features are applied to a linear support vector machine (SVM) to predict the saliency value. In [13, 39], Kümmerer et al. introduced a deeper structure for the encoder. DeepGaze I [39] uses pretrained AlexNet and DeepGaze II [13] uses pretrained VGG-19 for extracting features from the input image. Huang et al. [40] proposed a deep CNN structure that integrates information at different image scales. They showed that adding multi-scale information improves the saliency prediction results. Kruthiventi et al. [12] introduced a fully CNN model with a new "location biased convolutional (LBC) layer" to learn "location specific patterns" such as the center bias. Jetley et al. [41] formulated saliency map prediction as a probability distribution prediction task and trained a model to learn this distribution. Liu and Han [16] introduced a saliency model with a convolutional long short-term model (LSTM) to learn the global context. Cornia et al. [14] proposed a new saliency prediction architecture that incorporates a convolutional LSTM network and a spatial attentive mechanism. In [42], a saliency model based on image segmentation was introduced that exploits the object information for the saliency prediction task. Wang et al. [3] proposed a video saliency model, called ACLNet that uses

the CNN-LSTM network to predict visual attention over dynamic scenes. Wang et al. [43] proposed a model that incorporates multi-level saliency predictions within a single network to decrease redundancy. Some researches focus on decreasing the model complexity and inference time for real-time application [44].

In summary, Table 1 compares the main properties of some prominent saliency prediction models and our proposed models. In Table 1, NSS, KL-D, CC, MSE, and SIM stand for normalized scanpath saliency, Kullback–Leibler divergence, linear correlation coefficient, mean square error, and similarity respectively.

2.3. Salient Object Detection. The goal of salient object detection is to detect the most salient objects of a scene. Zhang et al. [46] used the multistage refinement mechanism to propose augmenting feedforward neural networks for addressing feature resolution reduction in CNNs. Zhao et al. [47] proposed a CNNs-based architecture that uses contrast prior to enhance the depth of information for salient object detection. Zhang et al. [48] proposed a probabilistic RGB-D saliency detection model based on conditional variational autoencoders. Li et al. [49] proposed a model that uses a pixel-level fully convolutional stream and a segment-wise spatial pooling stream to overcome the problem of blurry saliency maps, especially near the boundary of salient objects. To better segment salient and preserve the salient edges, Wang et al. [50] also proposed a model with a salient edge detection module. In most of these methods, incorporation of global and local information is missing and the need to use an appropriate model of jointly considering this information is still a challenge.

3. Motivation

Psychological and neurobiological experiments have discovered the role of contextual information in guiding the attentive mechanism of HSV. To understand the influence of contextual information on local saliency prediction, assume a red apple among green apples. In this scene, this red apple is certainly a salient object because of its distinct color, but among some apples with similar shape and color, it may not be a salient object. Hence, an ideal saliency model that aims to mutate this attentive mechanism is supposed to incorporate the contextual information of the scene in saliency prediction. Despite the state-of-the-art performance of deep saliency models, some experimental results have approved that CNN-based saliency models fail to capture global information and location-dependent features of the scene. In this regard, we proposed two approaches to incorporate the global scene properties and remedy some deficiencies in CNN-based saliency models. In this section, we peruse the importance of the global scene properties in saliency prediction and deficiencies of CNN structures.

3.1. Contextual Cues and HVS. In order to understand the importance of global properties in saliency prediction, we explain how the HVS computes the visual saliency of a scene. To capture the global features of the scene which describe the context of the scene, the brain encodes the consistent properties of the scene [51]. Experimental results reported by [39] show that the neurons belonging to the visual part of the brain demonstrate tuning characteristics that can be optimized to respond to recurring features in the scenes with comparable contents [52], thus the scenes with similar global characteristics will get similar processes in the human brain.

The global visual context guides the attentive mechanism of the HVS, i.e., what to expect in the scene and where is the most salient region in familiar scenes. Indeed, HVS uses an unsupervised learning mechanism to determine the optimal features from input scenes and localize the salient regions in these scenes. When it confronts with unfamiliar scenes having comparable global properties, it uses its past experiences to efficiently process these scenes and optimally allocate the perceptual and cognitive resources [51]. In neurobiology, it is called a contextual cueing effect [53].

3.2. Deficiencies in CNN Structures. In a single convolutional layer, every neuron observes the input through an aperture called the convolution window. Prevalently, the size of this window is much smaller than the spatial size of the input, hence a convolutional layer is capable of extracting local features from images but it fails to efficiently extract the high-grade contextual features instead. A CNN typically consists of a series of convolutional layers. Every hidden layer in this structure uses the output of its previous layer as input. In the applications where contextual features are needed (e.g., image classification tasks), they employ some fully connected layers as the later stage to combine these mostly local features and to generate more effective global features.

Statistically, it has been observed that the human eye fixations are strongly biased toward the center of an image [54] which is often explained through the photographer's bias [12] or through an uninterested observer's viewing strategy [55]. This phenomenon can be observed in many saliency benchmark datasets. For instance, Figure 1 shows the average of all ground truth saliency maps in the SALICON 2017 train set. This property can be considered as a global feature of the fixations of any saliency benchmark dataset. One of the most important drawbacks of using CNN structures for saliency prediction is that fully convolutional networks (FCNs) are unable to extract the center bias of the eye fixations because of the global nature of this bias. In addition, convolutional layers use weight sharing, and hence they are location-invariant (or shift-invariant). Hence, they are incapable of learning the location-dependent patterns too [12].

To compensate for some of these aforementioned deficiencies, several methods have been proposed since 2014. It has been shown that cues like center bias may improve model performance [41]. To account for the center bias, some approaches linearly combined the saliency prediction with a fixed Gaussian blob (an estimate of the prior distribution) [13, 39]. Kruthiventi et al. [12] introduced an LBC filter for capturing location-dependent patterns. Instead of

Table 1: A comparison between the main properties of some prominent saliency models.

Model name	Center bias	Transfer learning	CNN	Loss function
LSTM-based SAM [14]	Multiple-learned priors	✓	VGG, ResNet	NSS, KL-D, CC
DSCLRCN [16]	—	✓	VGG, ResNet	NSS
DeepGaze II [13]	Gaussian prior	✓	VGG	Log-likelihood
ML-Net [15]	Single multiplicative map	✓	VGG	Normalized MSE
PDP [41]	—	✓	VGG	Probability distances
SalNet [45]	—	✓	VGG	Euclidean loss
SALICON [40]	—	✓	AlexNet, VGG, GoogLeNet	KL-D, NSS, CC, SIM
DeepFix [12]	Handcrafted priors	✓	VGG	Euclidean loss
eDN [38]	—	-	1 to 3 layer networks	Euclidean loss
GLG-I (proposed model)	Multiple-learned priors, fully connected component	✓	ResNet	NSS, KL-D, CC
GLG-II (proposed model)	Multiple-learned priors	✓	ResNet, VGG	NSS, KL-D, CC

using predefined priors, Jia et al. [1] used a prior image to capture center bias and then pixel-wise multiplied this prior image by the predicted saliency map.

4. Proposed Methods

When predicting the saliency map of an input image, the saliency value of each image patch is influenced not only by the visual features of that patch (local features) but also by the global properties of the whole scene, contextual information, and the location of the patch in that image. In this section, we propose two approaches that incorporate both locally extracted features and global scene properties into local saliency prediction. In pixel-wise saliency prediction, these methods enable the saliency model to take into account not only the locally extracted features of each pixel location but also the global scene properties. Accordingly, we call these methods the global-local gazing (GLG) based method. To evaluate the effects of employing the global-local gazing concept in saliency prediction, we use SAM-ResNet [18] as the base model and extend this model using our proposed methods.

4.1. Base Model. The saliency attentive model (SAM) is among the best saliency models and was proposed by Cornia et al. [14]. Figure 2 presents the architecture of this deep saliency model. It is consisting of a dilated convolutional network and a ConvLSTM network. The dilated convolutional network is an extended version of a deep convolutional neural network that has higher resolution feature maps. Cornia et al. introduced two versions of the saliency model. One of them uses VGG-16 [56] and the other version uses ResNet-50 [57] as the backbone. This dilated neural network extracts some local feature maps from the input image. The role of attentive ConvLSTM is to focus iteratively on related spatial locations to enhance extracted features. The number of timesteps for this Attentive ConvLSTM has been set to 4. An explicit prior component has been introduced in order to learn the center prior. At the final stage,

Figure 1: The average of all ground truth saliency maps in the SALICON 2017 train set.

a convolutional layer predicts the saliency map of the input image. To train and evaluate the model, a loss function has been defined as [14]:

$$L\left(\tilde{y}, y^{\text{den}}, y^{\text{fix}}\right) = \alpha NSS\left(\tilde{y}, y^{\text{fix}}\right) + \beta CC\left(\tilde{y}, y^{\text{den}}\right) + \gamma KL\left(\tilde{y}, y^{\text{den}}\right), \tag{1}$$

where $\tilde{y}$, y^{den}, and y^{fix} are the predicted saliency map, the ground truth density distribution, and the ground truth binary fixation map respectively. *NSS( )*, *CC( )*, and *KL( )* are the normalized scanpath saliency, the linear correlation coefficient, and the Kullback–Leibler divergence respectively which are among the most popular saliency measures. Loss parameters [14]: in this work, we use SAM-ResNet as the base model to evaluate the effectiveness of our proposed approaches. We extended the SAM-ResNet [14] using our GLG methods, to inject the global scene properties into local saliency prediction.

4.2. The GLG-I Saliency Model. As aforementioned, the convolutional layers use weight sharing to reduce the number of model parameters. Namely, all the neurons in a

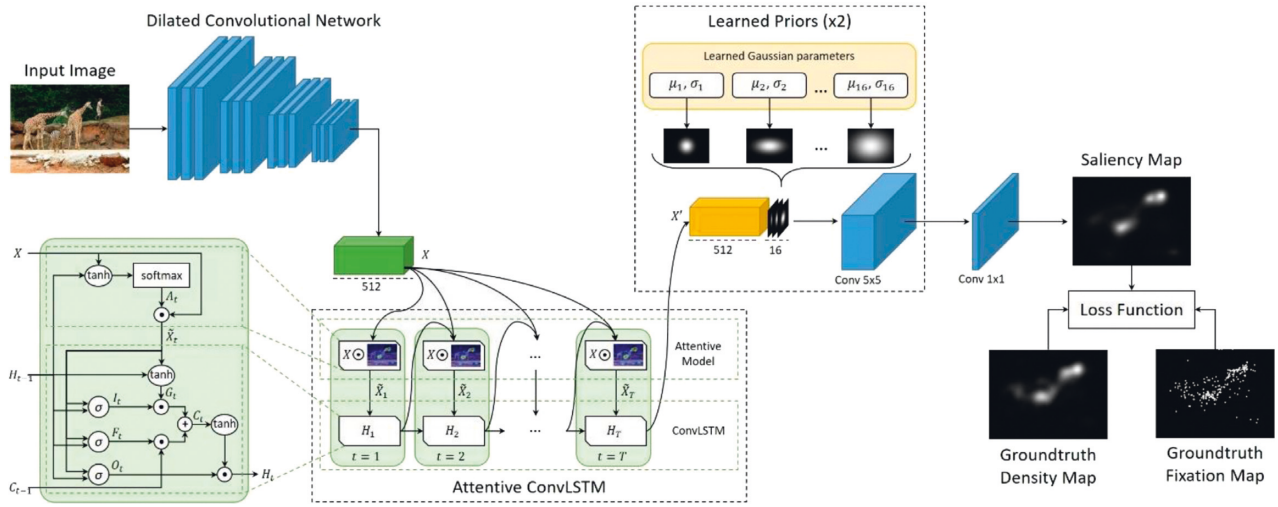

Figure 2: The saliency attentive model [14].

convolutional layer use the same weights. These weights do not depend on the location of neurons and are being used for all the spatial locations of the input. This property makes the convolutional layers location-invariant. Hence, the convolutional layers are unable to use different weights for different locations and to extract location-specific features. For CNN-based saliency models that predict the output saliency map pixel-wise, it is necessary to employ a component that compensates for such shortcomings, because the saliency value of a pixel or an image patch is very dependent on the context information of the whole scene and other global properties such as center prior.

In this subsection, we introduce a component called the fully connected component. We use this component to extend and modify the base model to create our GLG-I saliency model. This extension is able to extract location-dependent and global properties of the scene to reinforce the global information for pixel-wise saliency prediction. Figure 3 presents the architecture of our GLG-I saliency model.

To compute the location-dependent features and global scene properties, the locally extracted feature maps that are extracted by the dilated ResNet are applied to the fully connected component. The architecture of this proposed component is presented in Figure 4. This component is composed of three convolutional layers and a fully connected layer. Two convolutional layers with a core size of 3×3 are employed at the primary stage to reduce the number of input channels. These layers help the component to reduce the number of parameters. Afterward, a 2D array of 1200 fully connected neurons with a size of 30×40, called the fully connected layer, is employed to compute the location-dependent features and global scene properties. The fully connected neurons of this layer are connected to all neurons of the second convolutional layer. Unlike the convolutional layers, the fully connected layer is location-variant because every fully connected neuron in this layer has its own weights and is able to capture location-dependent patterns/features. Finally, a convolutional layer is used to smooth the output of the fully connected layer.

Through the use of fully connected neurons, compared to the base model, these neurons increase the number of parameters only by 1.6 percent. The fully connected component has a limited number of parameters too and compared to the base model, it increases the number of the parameters only by 15 percent. However, this number of parameters can still be reduced by selecting the appropriate number of cores for the first convolutional layer in the fully connected component. For example, if we set the number of cores for the first convolutional layer to 16, compared to the base model, the number of the model parameters increases only by 2 percent without any noticeable performance reduction. Table 2 presents the architectural details of our fully connected component.

The resulting feature map is concatenated with the output of the learning prior module, and then these feature maps are applied to a convolutional layer for predicting the saliency map of the input image. In the training phase, this predicted saliency map is evaluated using the ground truth. Table 2 compares the number of parameters in our GLG-I model with the base model.

4.3. The GLG-II Saliency Model. As aforementioned, GLG-I uses a fully connected component to compute the location-dependent features and global scene properties. Instead of using a fully connected component, here we introduce another approach called GLG-II that uses the output fully connected layer at the final stages of a deep neural network for extracting the contextual features of the scene. Most deep models predict the saliency value pixel-wise, and hence we use a new approach to make the contextual information available pixel-wise. To do so, we repeat this global feature vector to make it available at any spatial location of the image. Here, we use the VGG neural network (VGGNet) [56] to extend and modify our base model and to create our GLG-II saliency model, but in general, the fully connected layers of the backend neural network can be used instead to avoid using an additional deep model. Figure 5 presents the architecture of our GLG-II saliency

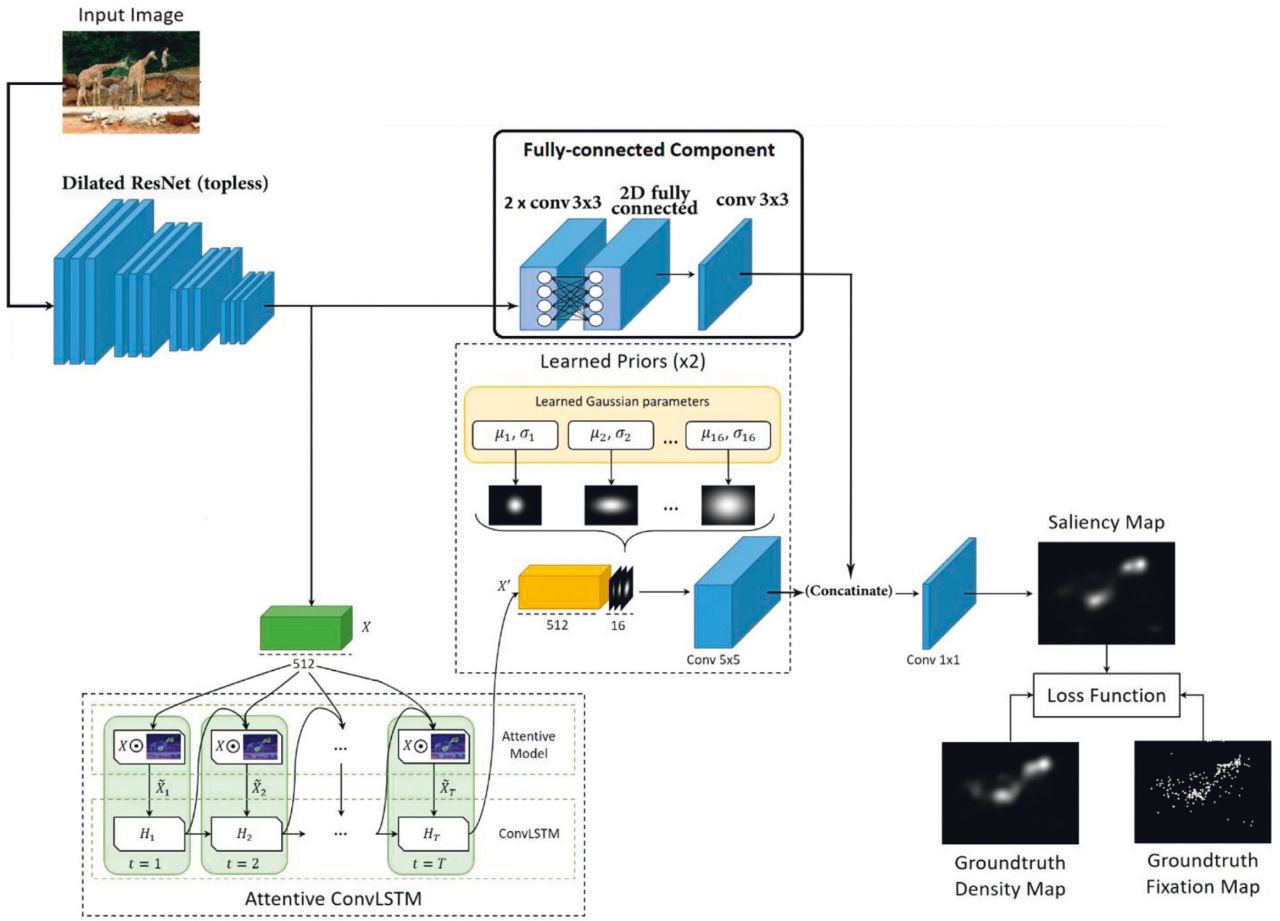

Figure 3: The GLG-I saliency model.

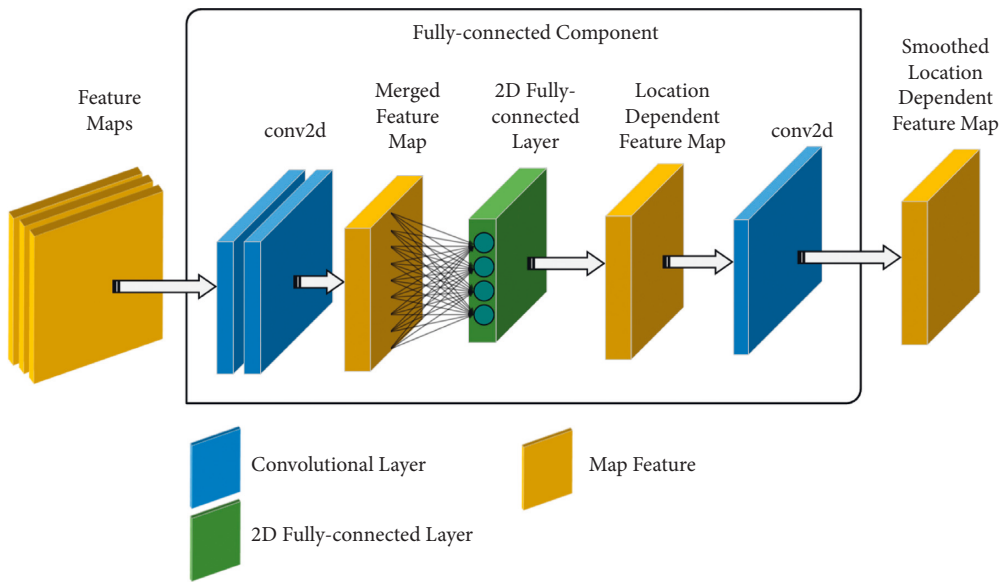

Figure 4: The fully connected component.

Table 2: The architecture of our fully connected component.

	Core size	Number of cores	Output size	Number of parameters
Conv2d	3×3	512	$30 \times 40 \times 512$	9,437,696
Conv2d	3×3	1	$30 \times 40 \times 1$	4,608
2D fully connected layer	—	—	$30 \times 40 \times 1$	1,124,400
Conv2d	3×3	1	$30 \times 40 \times 1$	10
	Total number of parameters:			10,566,715

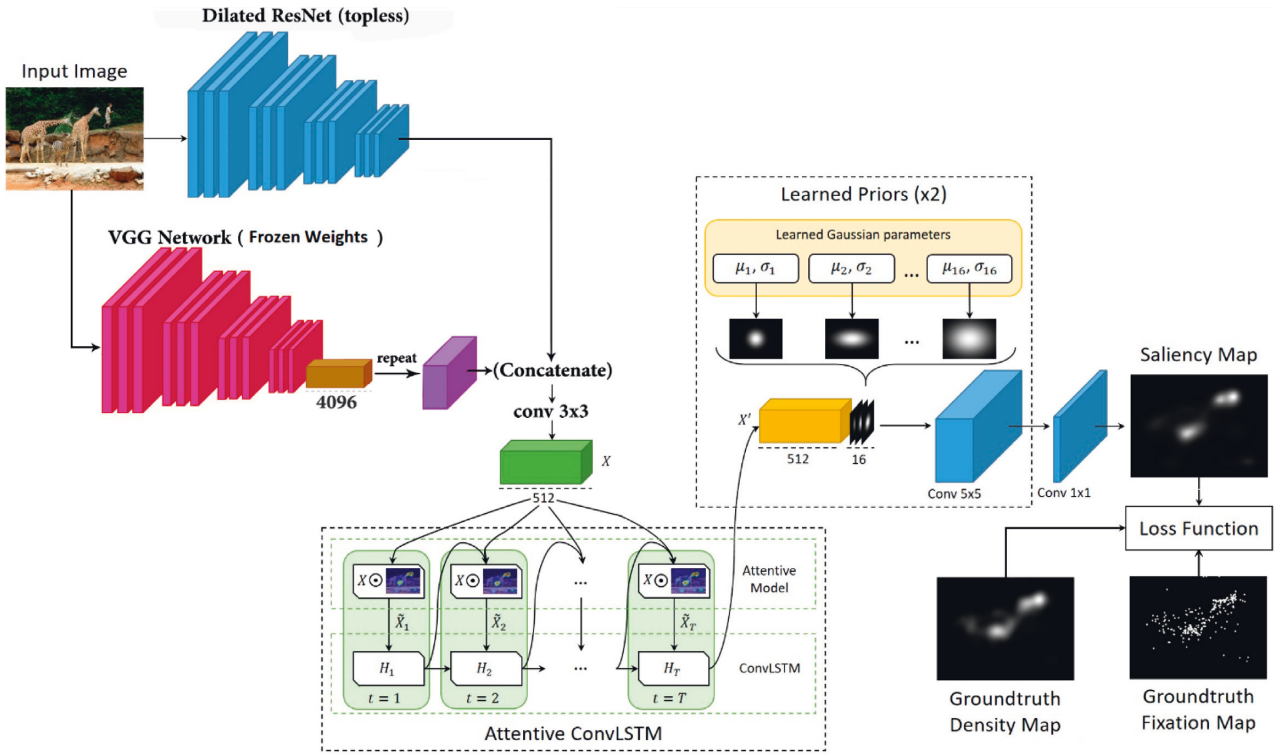

Figure 5: The GLG-II saliency model.

model. The weights of this VGGNet are initialized with that of the VGG-16 trained on ImageNet [58]. The output of the second fully connected layer of this VGG structure was considered as contextual features because this neural network has been trained to classify input images based on their context, and thus the features that are extracted at later layers are expected to describe the contextual information of the input image.

For every input image, the locally extracted features and contextual information are computed using dilated ResNet and VGGNet respectively. The VGG neural network generates a vector of 4096 features at its second fully connected layer. We use this feature vector as the contextual information of the scene. CNN-based saliency models predict the saliency value pixel-wise. To incorporate the contextual information in the saliency prediction of every pixel, we embed the contextual information in every spatial location of each pixel. To do this, we repeat this feature vector along two spatial dimensions (width and height) to generate a 3D global feature array. The globally and locally extracted feature maps are concatenated to enable the model to predict the saliency value of each pixel by using both of this information. A 3×3 convolutional layer is employed to reduce the number of channels in concatenated feature maps and as a result, the number of model parameters reduces. However, the increase in the number of model parameters compared to the base model is due to this layer. Then, a convolutional LSTM fine-tunes the resulting features. After the prior module, at the final stage, a convolutional layer predicts the saliency map of the input image. For the training phase, this predicted saliency map is evaluated using the ground truths.

We initialize the weights of the VGGNet with that of the VGG-16 trained on ImageNet [58]. As we want to use VGGNet to extract contextual information, the trained weights on ImageNet would be enough, and no training phase is required for our VGGNet. That is, the weights of VGGNet would stay frozen and the number of the model parameter would not increase by employing VGGNet. Table 3 compares the number of parameters in our GLG-II model with the base model.

Table 3: The number of parameters in GLG models compared to the base model.

Model name	Number of parameters
SAM-ResNet	70,093,441
GLG-I	80,660,157
GLG-II	88,967,809

5. Experimental Setup

In this section, some popular evaluation metrics, evaluation baselines, and saliency datasets are described, and then implementation details are provided.

5.1. Evaluation Metrics. For measuring the saliency model performance, several measures are being used. Some of these evaluation measures are distribution based and they compare predicted saliency maps and fixation maps. Other metrics are location based and compute some statistics at fixated locations. In this section, these metrics are concisely described.

5.1.1. Pearson's Correlation Coefficient. The correlation coefficient measure (CC), calculates the correlation between the ground truth map G and the predicted map P. It can be measured as [59]:

$$CC(P,G) = \frac{\operatorname{cov}(P,G)}{\operatorname{std}(P) \times \operatorname{std}(G)}, \tag{2}$$

where std( ) and cov( ) compute the standard deviation and covariance, respectively. The CC ranges between −1 and 1. A value of 1 shows a complete positive correlation between P and G. A value of 0 shows no relationship between these two maps.

5.1.2. Kullback–Leibler Divergence. Kullback–Leibler divergence (KL-D) can be used to calculate the difference between two probability distributions. If we interpret the predicted map P and ground truth map G, it can be computed as [59]:

$$\mathrm{KL}(P,G) = \sum_i G_i \log\left(\varepsilon + \frac{G_i}{\varepsilon + P_i}\right), \tag{3}$$

where ε is a constant is used for regularization and i indexes the ith pixel. As can be seen, the KL score is asymmetric. A larger KL value shows a larger difference between the predicted saliency map and fixation map while a KL score of zero indicates that the model is predicting the saliency values perfectly.

5.1.3. Earth Mover's Distance. The Earth mover's distance (EMD) measures the spatial distance between the predicted map and the ground truth map over a region. EMD computes how much transformation the predicted saliency map would need to match the fixation map [59].

A larger difference between the predicted map and fixation maps results in a larger EMD value while a zero value shows that the predicted and fixation maps are the same.

5.1.4. Similarity or Histogram Intersection. The similarity metric (SIM) measures the similarity between the predicted saliency map P and ground truth fixation map G. SIM is computed as the sum of the minimum values of the normalized P and G at each pixel. It can be computed as [59]:

$$\begin{aligned} \mathrm{SIM}(P,G) &= \sum_i \min(P_i, G_i), \\ \text{where} \sum_i P_i &= \sum_i G_i = 1. \end{aligned} \tag{4}$$

The SIM ranges between zero and one. A value of 1 shows P and G are the same. A value of 0 shows no overlap between P and G.

5.1.5. Normalized Scanpath Saliency. The normalized scanpath saliency (NSS) calculates the correspondence of predicted saliency maps P and the binary fixation map of G^B. It measures the average of the predicted saliency values in fixated points after normalization and can be computed as [59]:

$$\begin{aligned} \mathrm{NSS}\left(P, G^B\right) &= \frac{1}{N} \sum_i \overline{P}_i \times G_i^B, \\ \text{where} N &= \sum_i G_i^B \text{and} \overline{P} = \frac{P - \operatorname{mean}(P)}{\operatorname{std}(P)}, \end{aligned} \tag{5}$$

where std( ) and mean( ) compute the standard deviation and average, respectively, i indicates the ith pixel, and N is the number of fixation points.

A larger NSS indicates higher saliency values in fixated points and better performance of the model. An NSS of zero shows that the saliency model does not work better than a random number generator and a negative NSS shows that the saliency model performs worse than a random number generator.

5.1.6. Area under ROC Curve. By interpreting a saliency model as a binary classifier that classifies each pixel into fixated and nonfixated. The ROC curve can be created by plotting the true positive rate (TPR) versus the false positive rate (FPR) at various discrimination thresholds. The area under the ROC curve (AUC) is used to evaluate the model's performance. Various kinds of AUC have been proposed which differ in how TPR and FPR are calculated. The AUC values range between 0.0 and 1.0. A larger AUC indicates that the model performs better.

The AUC-Judd [36] (or AUC in this study) uses the samples from saliency map values as the thresholds. For a given threshold, the percentage of saliency values smaller

than the threshold at fixation locations is TPR and the percentage of saliency values higher than the threshold at unfixated pixels is FPR.

To bypass the effects of the center bias on FPR calculation, The AUC-Borji [60] calculates the FPR at random pixels that are sampled uniformly from all image pixels and the shuffled AUC (sAUC) [61, 62] calculates the FPR at random pixels that are sampled uniformly from fixations on other images. Despite the difference in the definition of TPR, the AUC-Judd, the AUC-Borji, and the shuffled AUC calculate the TPR similarly.

5.1.7. Information Gain. Information gain (IG) [63] is an information-theoretic metric that computes the average information gain of the saliency map P for the center-prior baseline B at fixated locations G^B [59]. Information gain is computed as:

$$\mathrm{IG}\left(P, G^B\right) = \frac{1}{N}\sum_i G_i^B\left[\log_2\left(\varepsilon + P_i\right) - \log_2\left(\varepsilon + B_i\right)\right], \quad (6)$$

where ε is a constant for regularization, i indicates the ith pixel, and N is the number of fixation points. An IG score above zero indicates the model outperforms the center prior to baseline in the prediction of ground truth fixations.

5.2. Evaluation Baselines

(i) Infinite: this baseline uses the fixation points of an infinite number of observers to predict the fixation points of another infinite number of observers.

(ii) One human: this baseline uses the fixation points of an observer to predict the fixation points of the other observers.

(iii) Center: this baseline uses a symmetric 2D Gaussian map as the predicted fixation map of the input image.

(iv) Permutation: this baseline uses fixation points of a randomly selected image as the predicted fixation points of the input image.

(v) Chance: this baseline uses a randomly generated saliency map as the predicted fixation map of the input image.

5.3. Saliency Datasets. In this work, we train and evaluate our models over four datasets: the dataset of SALICON Challenge 2015, the dataset of SALICON Saliency Prediction Challenge (LSUN 2017), MIT300, and MIT1003 that are among the most popular image-based saliency datasets.

5.3.1. SALICON 2015 and SALICON 2017. The dataset of SALICON Challenge 2015 [64] and the dataset of the SALICON Saliency Prediction Challenge (LSUN 2017) are among the richest saliency datasets based on the MS COCO image dataset [65]. They consist of 10,000 images for training, 5,000 images for validation data, and 5,000 images for the test. We call these datasets SALICON 2015 and SALICON 2017 respectively. Presently, the model evaluation over SALICON 2015 test set is not available because it has been closed by the provider.

Deep neural networks need abundant data for the training phase. Currently, many studies train their deep saliency models on the SALICON dataset and then fine-tune on other saliency datasets for predicting fixations of small datasets. Considering the evaluation result of state-of-the-art saliency models over the SALICON 2015 test set that is available in [2], our base model, SAM-ResNet [14], is among the best models over SALICON 2015 test set.

5.3.2. MIT300. The MIT300 [66] consists of 300 color images of natural indoor and outdoor scenes in JPG format that is used as a benchmark test set. The ground truth (fixation points and saliency map) of this dataset is not provided and the MIT/Tuebingen Saliency Benchmark [67, 68] uses it for evaluation of the saliency models according to multiple metrics.

5.3.3. MIT1003. The MIT1003 [36] consists of 1003 color images of natural indoor and outdoor scenes in JPG format. The ground truth (fixation points and saliency map) of this dataset is provided and it is available as the training data for MIT/Tuebingen Saliency Benchmark [67, 68].

5.4. Implementation Details. As mentioned before, our models are evaluated on SALICON 2015, SALICON 2017, and MIT300. For SALICON 2015 and SALICON 2017, we train our model on the training data and are validated on the validation set of these datasets using the loss function in (1). For SALICON datasets, a batch size of 10 samples is chosen for the training and validation phase. As instructed by the MIT Saliency Benchmark [67], for MIT300, we pretrain our models on the SALICON and then fine-tune them on MIT1003. To find the appropriate version of the SALICON dataset that leads to better performance on MIT300, we tested both SALICON 2015 and SALICON 2017 for the pretraining phase separately. To fine-tune the models on MIT1003, this dataset is split randomly into 904 images of the training set and 99 images of the validation set. In the pretraining phase on SALICON and fine-tuning phase on MIT1003, batch sizes of 10 and 9 samples are chosen respectively.

For the pretraining and finetuning stages, the learning rate is initialized to 10^{-4} and after every two epochs it is decreased by a factor of 10. Finally, the models with the best validation loss are chosen for evaluation on the test set.

We use a computer with 16 GB RAM and NVIDIA Tesla K80 GPU. The number of rows and columns of the input images is 240 and 320 pixels, respectively. The inference time of the base model using the aforementioned GPU is about 200 ms. The inference times of our models are about 250 ms which shows only a 25 percent increase. This indicates that our methods do not increase the overall inference time of the model. The reason for this is that the base model uses a recursive component that requires a lot of time to calculate its output.

Table 4: Performance of GLG models, compared to state-of-the-art saliency models over the SALICON 2015 validation set, compiled from SALICON challenge 2015 website.

Model name	AUC	CC	SAUC	NSS
EOF-MODEL [42]	0.886	**0.851**	**0.791**	3.026
GLG-I	**0.89**	0.846	0.788	3.243
GLG-II	0.887	0.843	**0.791**	**3.262**
DSCLSTM [16]	0.887	0.835	0.788	3.221
DSCLRCN [16]	0.887	0.835	0.785	3.221
SAM-ResNet [14]	0.886	0.844	0.787	3.26
DeepGaze II [13]	0.886	0.505	0.767	1.34
FSM [44]	0.884	0.803	0.775	2.756
SAM-VGG [14]	0.883	0.83	0.782	3.219
ML-Net [15]	0.869	0.744	0.776	2.829
SalNet: deep convnet [45]	0.858	0.609	0.727	1.822
SalNet: shallow convnet [45]	0.817	0.548	0.658	1.625

6. Experimental Results

The evaluation results of our GLG models over the SALICON 2015 validation set, the SALICON 2017 test set, and MIT300 are reported and compared with the state-of-the-art saliency models. Currently, evaluations on the CAT2000 test set and the SALICON 2015 test set are closed and are not available anymore.

Considering Table 4, over the SALICON 2015 validation set, the GLG-I model outperforms the base model according to AUC, CC, and sAUC and outperforms all other existing state-of-the-art saliency models according to AUC. The GLG-II model outperforms the base model (SAM-ResNet) according to AUC, sAUC, and NSS and outperforms almost other existing state-of-the-art saliency models according to sAUC and NSS.

Considering Table 5, our models outperform the base model according to AUC, CC, KL, IG, and SIM. Our models also outperform other state-of-the-art models according to CC, AUC, and SIM, and over the SALICON 2017 test set.

Considering Table 6, over the MIT300, the GLG-I model outperforms the base model according to EMD, AUC-B, sAUC, CC, and KL, and the GLG-II model outperforms the base model according to EMD, AUC-B, sAUC, CC, NSS, and KL. In Table 6, the evaluation results were sorted based on SIM, CC, and AUC-B. Overall, our proposed models also outperform as well as the best state-of-the-art models.

It also shows that pretraining on SALICON 2017 and SALICON 2015 does not affect noticeably on model performance over MIT300.

It can be concluded from the evaluation results over SALICON 2015, SALICON 2017, and MIT300 that our methods improved the performance of the base model. These extensions on the base model enable the saliency model to capture global information better and improve the accuracy of the saliency prediction task. In Figure 6, we compare the output of our model with EML-NET and SAM-ResNet. Figure 6 demonstrates that by using our proposed methods for including the contextual information and location-dependent patterns, the focus of attention gets corrected in most cases and the model performance improves according to several evaluation metrics.

7. Discussion

As aforementioned, convolutional layers use weight sharing and as a result, they are location-invariant. Hence, the fully convolutional neural networks [44] make them incapable of learning the location-dependent patterns [12], and global scene properties. In our GLG-I model, we propose a novel fully connected component to incorporate these properties into the local saliency prediction. Unlike the convolutional layers, the fully connected layer is location-variant because every fully connected neuron in this layer has its own weights and is able to capture location-dependent patterns/features. Considering the performance of the GLG-I model on different datasets, it can be concluded that by employing some location-variant structures in the model, the performance of saliency prediction improves considerably.

Experimental results demonstrate that the neurons of the visual part of the brain show tuning properties that can be optimized to better react to recurring features in the scenes with comparable contents [52]. HVS provides a good platform to learn the best features and locations of the salient region of a scene and extend this for similar scenes [51]. Our GLG-II model imitates this mechanism in the human brain and employs an additional VGGNet to extract and incorporate the contextual information of the scene. Considering the performance of the GLG-II model on different datasets, it can be concluded that as expected from the contextual cueing effect [53], by incorporating the contextual features of the scene into the local saliency prediction, the performance of saliency prediction improves.

Despite the fact that the deep state-of-the-art saliency models have shown tremendous improvements over the classic saliency models, these models mainly suffer from a high number of parameters. Although deep saliency models are suitable for applications that require high accuracy, they are not recommended for real-time applications due to their high number of parameters. The models with high complexity require more calculation and powerful and expensive hardware for training and test phases. The new studies need to focus not only on higher performance but on the lower model complexity. Some domains with the real-time application demand light models with mediocre performance.

Table 5: Performance of GLG models compared to SAM-ResNet over the SALICON 2017 test set.

Model name	CC	AUC	SIM	KL	IG	NSS
GLG-I	**0.903**	**0.867**	0.798	0.37	0.764	1.99
GLG-II	**0.903**	**0.867**	**0.799**	0.43	0.708	1.987
EOF-MODEL [42]	0.900	0.866	0.794	0.392	0.723	1.954
SAM-ResNet [14]	0.899	0.865	0.793	0.61	0.538	1.99
MSI-Net [69]	0.889	0.865	0.784	**0.307**	**0.793**	1.931
EML-NET [1]	0.886	0.866	0.78	0.52	0.736	2.05
GazeGAN [70]	0.879	0.864	0.773	0.376	0.72	1.899
FSM [44]	0.875	0.862	0.772	0.365	0.716	1.863
MD-SEM [71]	0.868	—	—	0.568	—	**2.058**
SalNet [45]	0.622	—	—	—	—	1.859

Table 6: Performance of GLG models compared to state-of-the-art saliency models over the MIT300 dataset, compiled from [2].

Model name	SIM	CC	AUC-B	KL	EMD	sAUC	NSS	AUC-J
Baseline: infinite	**1**	**1**	**0.88**	**0**	**0**	**0.81**	**3.29**	**0.92**
EOF-MODEL [42]	*0.68*	0.79	0.80	1.05	2.10	0.72	2.31	0.87
GLG-I_salicon 2017	*0.68*	0.79	0.80	1.13	2.00	0.71	2.34	0.87
GLG-II_salicon 2017	*0.68*	0.79	0.80	1.10	1.99	0.71	2.34	0.87
GLG-I_salicon 2015	*0.68*	0.79	0.80	0.99	2.03	0.71	2.33	0.87
GLG-II_salicon 2015	*0.68*	0.79	0.79	1.24	2.05	0.71	2.35	0.87
EML-NET [1]	*0.68*	0.79	0.77	0.84	1.84	0.70	2.47	0.88
SAM-ResNet [12]	0.68	0.78	0.78	1.27	2.15	0.70	2.34	0.87
SAM-VGG [12]	0.67	0.77	0.78	1.13	2.14	0.71	2.30	0.87
FSM [44]	0.65	0.74	0.80	0.80	2.32	0.71	2.10	0.86
SalGAN [60]	0.63	0.73	0.81	1.07	2.29	0.72	2.04	0.86
PDP [37]	0.6	0.70	0.80	0.92	2.58	0.73	2.05	0.85
ML-Net [13]	0.59	0.67	0.75	1.10	2.63	0.70	2.05	0.85
DVI [43]	0.58	0.68	0.78	-	3.05	0.71	1.98	0.85
SalNet [38]	0.52	0.58	0.82	0.81	3.31	0.69	1.51	0.83
GBVS [6]	0.48	0.48	0.80	0.87	3.51	0.63	1.24	0.81
Deep gaze 2 [11]	0.46	0.52	0.86	0.96	3.98	0.72	1.29	0.88
Baseline: center	0.45	0.38	0.77	1.24	3.72	0.51	0.92	0.78
IttiKoch2 [3]	0.44	0.37	0.74	1.03	4.26	0.63	0.97	0.75
eDN [34]	0.41	0.45	0.81	1.14	4.56	0.62	1.14	0.82
Deep Gaze 1 [35]	0.39	0.48	0.83	1.23	4.97	0.66	1.22	0.84
Baseline: 1 human	0.38	0.52	0.66	6.19	3.48	0.63	1.65	0.80
SUN saliency [45]	0.38	0.25	0.66	1.27	5.10	0.61	0.68	0.67
Baseline: Perm.	0.34	0.20	0.59	6.12	4.59	0.50	0.49	0.68
Baseline: Chance	0.33	0	0.50	2.09	6.35	0.50	0	0.50
IttiKoch [61]	0.2	0.14	0.54	2.30	5.17	0.53	0.43	0.60

For instance, in [44] a compact and light saliency prediction model with acceptable performance has been proposed for real-time applications on CPU.

As can be seen in the second row of Figure 6, based on the given ground truth image, an observer finds the man's face and the plastic bag as the salient objects of the input scene, but saliency models including our GLG models were not able to detect the bag as a salient object. It is mainly due to the partial occlusion of the plastic bag. None of the saliency models in Figure 6 perceived the connection between the man and the bag in his hand. As a result, we can conclude that complex backgrounds and partially occluded objects are two big challenges for saliency models. Another example of the partially occluded salient object is the third cow in the first input image in Figure 6. The head of the cow is occluded and as a result, none of the saliency models in Figure 6 (including our GLG models) could find it as a salient object. On the other hand, the human brain can easily identify the brown spot behind the second cow as the third cow by semantically completing missing parts in partially occluded objects.

Figure 6: Qualitative results and comparison to the state of the art.

8. Conclusion

In this study, we proposed two novel saliency models to predict human attention during scene free-viewing of natural scenes. To investigate the effectiveness of our methods, we used the SAM-ResNet [14] as the base model. We extended the base model using our proposed methods to inject contextual cues and capture location-dependent patterns/features in order to overcome the deficiencies of CNN structures in the base model. In our first approach, a novel fully connected component is used to incorporate the location-dependent and global scene properties. In the second approach, a VGGNet is employed to extract the contextual information of the scene.

Experimental results showed that our GLG models outperform not only the base model but also most previous saliency models over SALICON 2015, SALICON 2017, and MIT300 datasets. Our effort to incorporate the contextual information and global scene properties may supply new inspirations for future works on saliency models to apply such an amendment to the computational saliency models.

Data Availability

Experiments have been done based on a database available at the MIT saliency benchmark (http://saliency.mit.edu/) and MIT/Tuebingen Saliency Benchmark (https://saliency.tuebingen.ai/ and http://salicon.net/).

Conflicts of Interest

The authors have no conflicts of interest to disclose.

References

[1] S. Jia and N. D. Bruce, "Eml-net: an expandable multi-layer network for saliency prediction," *Image and Vision Computing*, vol. 95, Article ID 103887, 2020.

[2] A. Borji, "Saliency prediction in the deep learning era: successes and limitations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 1, no. 1, pp. 1–5, 2019.

[3] W. Wang, J. Shen, J. Xie et al., "Revisiting Video Saliency Prediction in the Deep Learning Era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, 2019.

[4] T. Deng, H. Yan, L. Qin, T. Ngo, and M. BJIToITS, "How do drivers allocate their potential attention?" *Driving fixation prediction via convolutional neural networks*, vol. 21, no. 5, pp. 2146–2154, 2019.

[5] 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.

[6] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Reviews Neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[7] Y.-F. Ma, "Zhang H-J Contrast-based image attention analysis by using fuzzy growing," in *Proceedings of the Eleventh ACM International Conference on Multimedia*, pp. 374–381, ACM, Berkeley, CA, USA, November 2003.

[8] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in Neural Information Processing Systems*, vol. 19, pp. 545–552, 2007.

[9] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *Advances in Neural Information Processing Systems*, vol. 20, pp. 241–248, 2008.

[10] 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, p. 10, 2009.

[11] 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.

[12] S. S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: a fully convolutional neural network for predicting human eye fixations," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, 2017.

[13] M. Kümmerer, T. S. Wallis, and M. Bethge, "DeepGaze II: reading fixations from deep features trained on object recognition," 2016, https://arxiv.org/abs/1610.01563.

[14] 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.

[15] M. Cornia, L. Baraldi, and G. Serra, "Cucchiara R Multi-level net: a visual saliency prediction model," in *European Conference on Computer Vision*, pp. 302–315, Springer, Berlin Germany, 2016.

[16] N. Liu and J. Han, "A deep spatial contextual long-term recurrent convolutional network for saliency detection," *IEEE Transactions on Image Processing*, vol. 27, no. 7, pp. 3264–3274, 2018.

[17] H. Li, X. Su, J. Wang et al., "Image processing strategies based on saliency segmentation for object recognition under simulated prosthetic vision," *Artificial Intelligence in Medicine*, vol. 84, pp. 64–78, 2018.

[18] W. Wang, J. Shen, H. Ling, and m intelligence, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 7, pp. 1531–1544, 2019.

[19] Y. Zhao, J. Zhao, J. Yang et al., "Saliency driven vasculature segmentation with infinite perimeter active contour model," *Neurocomputing*, vol. 259, pp. 201–209, 2017.

[20] E. Ahn, J. Kim, L. Bi et al., "Saliency-based lesion segmentation via background detection in dermoscopic images," *IEEE Journal of Biomedical and Health Informatics*, vol. 21, no. 6, pp. 1685–1693, 2017.

[21] W. Wang, J. Shen, R. Yang, F. Porikli, and m. intelligence, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20–33, 2018.

[22] W. Wang, J. Shen, X. Lu, S. C. H. Hoi, H. Ling, and M. Intelligence, "Paying attention to video object pattern understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2413–2428, 2021.

[23] F. Tu, S. S. Ge, Y. Tang, and C. C. Hang, "Saliency guided hierarchical robust visual tracking," 2018, https://arxiv.org/abs/1812.08973.

[24] S. Li, M. Xu, Y. Ren, and Z. Wang, "Closed-form optimization on saliency-guided image compression for HEVC-MSP," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 155–170, 2018.

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

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

[27] Y. Hu, D. Rajan, and L.-T. Chia, "Adaptive local context suppression of multiple cues for salient visual attention detection," in *Proceedings of the 2005 IEEE International Conference on Multimedia and Expo*, p. 4, July 2005.

[28] D. Gao and V. Mahadevan, "Vasconcelos N the discriminant center-surround hypothesis for bottom-up saliency," *Advances in Neural Information Processing Systems*, vol. 20, pp. 497–504, 2008.

[29] L. Itti, "Baldi PF Bayesian surprise attracts human attention," *Advances in Neural Information Processing Systems*, vol. 49, pp. 547–554, 2006.

[30] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection

speed," in *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, pp. 2049–2056, New York, NY, USA, June 2006.

[31] N. Bruce, "Tsotsos J Saliency based on information maximization," *Advances in Neural Information Processing Systems*, vol. 18, pp. 155–162, 2006.

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

[33] X. Hou, "Zhang L Saliency detection: a spectral residual approach," in *Proceedings of the Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pp. 1–8, IEEE, Minneapolis, MN, USA, June 2007.

[34] M. Z. Aziz and B. Mertsching, "Fast and robust generation of feature maps for region-based visual attention," *IEEE Transactions on Image Processing*, vol. 17, no. 5, pp. 633–644, 2008.

[35] 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, p. 7, 2009.

[36] T. Judd, K. Ehinger, and F. Durand, "Torralba A Learning to predict where humans look," in *Proceedings of the Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2106–2113, IEEE, Kyoto, Japan, September 2009.

[37] 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.

[38] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2798–2805, Columbus, Ohio, USA, June 2014.

[39] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: boosting saliency prediction with feature maps trained on imagenet," 2014, https://arxiv.org/abs/1411.1045.

[40] X. Huang, C. Shen, X. Boix, and Q. S. Zhao, "Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 262–270, Santiago, Chile, December 2015.

[41] S. Jetley and N. Murray, "Vig E End-to-end saliency mapping via probability distribution prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 5753–5761, Las Vegas, NV, USA, 2016.

[42] S. Zabihi, E. Mansoori, and M. Yazdi, "Exploiting object features in deep gaze prediction models," *Journal of Visual Communication and Image Representation*, vol. 73, Article ID 102931, 2020.

[43] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2018.

[44] S. Zabihi, H. R. Tavakoli, A. Borji, and E. Mansoori, "A compact deep architecture for real-time saliency prediction," *Signal Processing: Image Communication*, vol. 104, Article ID 116671, 2022.

[45] J. Pan, E. Sayrol, X. Giro-i-Nieto, K. McGuinness, and O. Connor, "NE Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 598–606, Las Vegas, NV, USA, June 2016.

[46] L. Zhang, J. Wu, T. Wang, A. Borji, G. Wei, and H. J. I. ToI. P. Lu, "A multistage refinement network for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 3534–3545, 2020.

[47] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang, "Contrast prior and fluid pyramid integration for RGBD salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3927–3936, Long Beach, CA, USA, June 2019.

[48] J. Zhang, D.-P. Fan, Y. Dai et al., "Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, vol. 44, pp. 8582–8591, 2020.

[49] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 478–487, Las Vegas, Nevada, USA, June 2016.

[50] W. Wang, S. Zhao, J. Shen, and S. C. Hoi, "Borji A Salient object detection with pyramid attention and salient edges," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1448–1457, Long Beach, CA, USA, June 2016.

[51] J. Li and W. Gao, "Visual saliency computation," *A machine learning perspective*, Vol. 8408, Springer, Berlin Germany, 2014.

[52] M. M. Chun, "Contextual guidance of visual attention," in *Neurobiology of Attention*, pp. 246–250, Elsevier, Amsterdam, Netherlands, 2005.

[53] L. Itti, G. Rees, and J. K. Tsotsos, *Neurobiology of Attention*, Elsevier, Amsterdam, Netherlands, 2005.

[54] 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, p. 4, 2007.

[55] A. Borji and L. Itti, "Cat2000: a large scale fixation dataset for boosting saliency research," 2015, https://arxiv.org/abs/1505.03581.

[56] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," 2014, https://arxiv.org/abs/1409.1556.

[57] K. He, X. Zhang, and S. Ren, "Sun J Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778, Las Vegas, Nevada, USA, June 2016.

[58] O. Russakovsky, J. Deng, H. Su et al., "Imagenet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[59] 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.

[60] 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.

[61] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti, "Analysis of scores, datasets, and models in visual saliency prediction," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 921–928, Sydney, NSW, Australia, December 2013.

[62] 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, p. 32, 2008.

[63] M. Kümmerer, T. S. A. Wallis, and M. Bethge, "Information-theoretic model comparison unifies saliency metrics," *Proceedings of the National Academy of Sciences*, vol. 112, no. 52, pp. 16054–16059, 2015.

[64] M. Jiang, S. Huang, J. Duan, and Q. S. Zhao, "Saliency in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1072–1080, Boston, MA, USA, June 2015.

[65] T.-Y. Lin, M. Maire, S. Belongie et al., "Microsoft coco: common objects in context," in *European Conference on Computer Vision*, pp. 740–755, Springer, Berlin Germany, 2014.

[66] T. Judd, F. Durand, and A. Torralba, *A benchmark of computational models of saliency to predict human fixations*, MIT, MIT tech report, 2012.

[67] Z. Bylinskii, T. Judd, A. Borji et al., "Mit saliency benchmark," 2015, http://saliency.mit.edu/.

[68] M. Kümmerer, Z. Bylinskii, T. Judd et al., "Torralba A MIT/Tuebingen Saliency Benchmark," 2020, https://saliency.tuebingen.ai/.

[69] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual Encoder-Decoder Network for Visual Saliency Prediction," *Neural Networks*, vol. 129, 2019.

[70] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, "How is gaze influenced by image transformations? dataset and model," *IEEE Transactions on Image Processing*, vol. 29, pp. 2287–2300, 2020.

[71] C. Fosco, A. Newman, P. Sukhum, Y. B. Zhang, A. Oliva, and Z. Bylinskii, "How many glances? Modeling multi-duration saliency," *SVRHM Workshop at NeurIPS*, 2019.