ScanpathNet: A Recurrent Mixture Density Network for Scanpath Prediction

Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya
The University of New South Wales, Sydney, Australia
r.debelen@unsw.edu.au, t.bednarz@unsw.edu.au, a.sowmya@unsw.edu.au

Abstract

Understanding the mechanisms underlying human visual attention is an important research problem in cognitive neuroscience and computer vision. While existing models predict salient regions (i.e., saliency maps) and temporal sequences of eye fixations (i.e., scanpaths) in images, their designs often partially follow theoretical frameworks. Here, we introduce ScanpathNet, a deep learning model inspired by the latest theoretical model in neuroscience. It is 'guided' by a dynamic priority map influenced by semantic content and fixation history. The model leverages convolutional neural networks to extract rich semantic features, convolutional long short-term memory networks to model the inhibition of return mechanism and sequential dependencies of fixations, and mixture density networks to predict probability distributions of fixations for each pixel. Simulated human scanpaths can then be generated by sequentially sampling the output of the proposed model. Despite its simplicity, ScanpathNet showed promising qualitative and quantitative scanpath prediction performance in extensive experiments on numerous eye-tracking benchmark datasets.

1. Introduction

The human visual system acts as an information processing bottleneck. It ensures that humans' limited processing resources are allocated to the most informative region of the environment. When looking at a scene, we sequentially shift our attention to the most relevant regions. Afterwards, we perform further analysis in our high visual acuity fovea. What attracts and drives human visual attention are important research questions in the field of cognitive science and computer vision. They are relevant to understanding the cognitive processes underlying knowledge acquisition and mental health [18, 19]. In addition, they provide technical contributions to many multimedia applications, such as image and video retargeting, multimedia compression, perceptual quality assessment and medical imaging [42].

Over the past decade, researchers have developed computational models of visual attention that predict the probability distribution of fixations (i.e., saliency map). Saliency prediction can be considered a mature field with various benchmark datasets and new models published annually. However, saliency models lose temporal information by aggregating fixations into a single saliency map, neglecting the fact that visual attention is a dynamic process. This work models the rapid change of human visual attention recorded as a temporal sequence of fixations (i.e., scanpath).

Over the last few years, there has been a substantial increase in the number of scanpath models inspired by the well-known Feature Integration Theory [57]. It suggests that a set of topographic feature maps containing basic features (*e.g.* colour and orientation) can be extracted from a scene. Within each map, different spatial locations compete for attention, allowing for the location with the highest value to be attended. A seminal paper [34] where a static saliency map is computed and fed into a "winner-takes-all" network to sequentially select locations in order of decreasing saliency advocated this approach. A simple inhibition of return mechanism was used to select the next fixation location [32]. There have been numerous improvements to this approach to achieve state-of-the-art performance.

Despite the success of computational models that utilise static priority maps, recent evidence suggests that during scene exploration, visual attention is "guided" by a dynamic priority map influenced by different factors, such as semantic information and fixation history. Attention is then allocated to the peak in the current priority map. The idea that various "guidance" factors could be used to deploy visual attention was the reason for the term "Guided Search" (GS). Since its introduction in 1989, the GS model has undergone numerous revisions [62–67]. In light of new research, GS6 proposes that attention is guided by both classical top-down and bottom-up features as well as other new types of guidance, including the history of fixations, value and scene guidance. Since this work focusses on free-viewing tasks where no top-down features are used for deployment of attention, a modified GS6 model is presented in Section 3.

Our contributions are three-fold. **First**, we present the first deep learning model inspired by GS6, making our model simpler, more interpretable and intuitive than existing scanpath models. **Second**, our proposed model implicitly learns the inhibition of return mechanism observed in humans instead of direct computation from the dataset as required by previous models. **Third**, qualitative and quantitative results show the promising human-level performance of our model on scanpath generation during free-viewing.

2. Related Work

Despite the significant increase in the number of saliency map models (see [10,33] for a review), there are still considerably fewer scanpath models. In this section, related work on scanpath prediction is provided (see [38] for a review). Existing scanpath models can be broadly categorised into traditional and deep learning-based approaches.

2.1. Traditional Approach

The earliest computational scanpath model is inspired by the early primate visual system [34]. It relies on multiscale image features extracted using dyadic Gaussian pyramids. The fixation locations are simulated using a winner-takes-all strategy with an inhibition-of-return (IOR) mechanism.

Human scanpaths were modelled previously as a stochastic process with non-local transition probabilities similar to a phenomenon of random walks known as Levy flights [9, 12]. This has also been combined with low-level feature saliency and semantic content using Hidden Markov Model (HMM) with a Bag-of-Visual-Words descriptor [46].

Numerous computational models extract different information (*e.g*. Incremental Coding Length [28], Residual Perceptual Information [59] and Super Gaussian Component response [54, 55]) from each visual feature in an image. Afterwards, human scanpaths are generated by sequentially selecting features with maximum information response.

Another computational model utilises kernel density estimation to infer saliency and applies stochastic filtering to simulate human scanpaths [56]. Attention mechanism has also been simulated using handcrafted parameters [20] and optimised parameters [50]. Bottom-up saliency maps, oculomotor biases and IOR inferred from eye-tracking data have also been used to model scanpaths [43, 45] and even simulate age effects [44]. A least-squares policy iteration method [36] and a probabilistic saccadic flow model [16] have been used to model the visual exploration of humans.

A biologically inspired model takes into account the effects of foveation, saccadic bias and IOR mechanism to predict human scanpaths [60]. Inspired by the neurophysiology of the superior colliculus, another model [1] simulates the blur around the fovea, projects information into the superior colliculus space and chooses the next fixation location using averaging operations similar to brain computations.

The Gravitational eye movement laws (G-eymol) [76] model scanpaths as a dynamic process similar to the laws of mechanics. Here, attention is subject to a gravitational field driven by the gradient of brightness, optical flow and an IOR potential. Unlike existing models, G-eymol does not compute the saliency map directly but integrates the positions of interest over time to predict scanpaths.

A Bayesian approach was used to fit a generative scanpath model based on the model's likelihood function [47]. This allows switching between two attentional states, similar to the exploration-exploitation dilemma. Due to its minimal set of assumptions, it is numerically efficient. However, it does not currently incorporate longer fixation history due to the nature of the Markov process.

2.2. Deep Learning Approach

An iterative representation learning framework to learn the saliency from an image has been proposed [69, 70]. A deep autoencoder is employed to reconstruct the input scanpath sequence and learn a reconstruction residual that estimates the saliency map. A centre bias is then applied to the computed saliency map and combined with an IOR mechanism to compute a priority map. Finally, the fixations are chosen as the locations with the highest priority values.

DeepGaze I [39] and II [41] are one of the best performing saliency models. Recently, DeepGaze III [40] has been proposed for scanpath prediction. It extracts features from a convolutional neural network (CNN) [52] and computes a spatial saliency map using a readout network. Scanpath history is then combined through a series of convolutions to yield fixation distributions.

SaltiNet is a deep neural network that uses a temporally-aware representation of saliency information for scanpath prediction on 360°images [4, 5]. PathGAN, a deep neural network trained on adversarial examples, was developed for both traditional images and omnidirectional images [3, 48].

Another model combines features (*e.g*. saccadic amplitude, CNN output) and memory bias to predict human scanpaths [51]. Instead of a recurrent neural network, a mathematical model was used to encode short-term and long-term memory. CNN [26, 52], long short-term memory (LSTM) [27] and convolutional LSTM (convLSTM) [72] were also used to generate saccade sequences [49, 58].

An extension of the Selective Tuning Attentive Reference model Fixation Controller (STAR-FC) [61] processes the input image in two streams: a peripheral stream computes low-level features while a central stream extracts features using CNNs. Both streams are combined with the fixation history to generate a priority map. The next fixation is selected from the maximum value in the priority map.

A deep convolutional saccadic model predicts fixation locations and durations [6]. Another model uses CNN, atrous spatial pyramid pooling [13] and a semantic segmen-

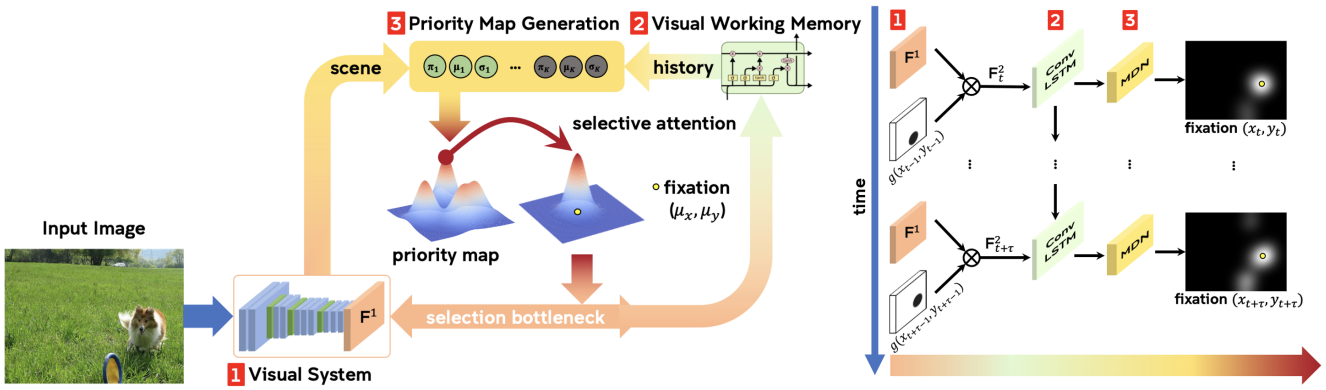

Figure 1. **(left)** A modified GS6 model where visual attention is guided by a dynamic priority map influenced by semantic content and fixation history. **(right)** ScanpathNet architecture includes three major modules: (1) Visual System (VS), (2) Visual Working Memory (VWM) and (3) Priority Map Generation (PMG). The input image is first downsampled and preprocessed with the VGG-16 network. The extracted feature representation is upsampled, multiplied by a spatial mask centred around the previous fixation location and passed to the convLSTM layer. The hidden representation from the convLSTM is flattened and passed to the MDN layer to learn the probability distribution of fixations. The MDN is sampled to identify the next fixation location.

tation module [25]. CNNs have also been used in conjunction with a variational autoencoder [22] and an SVM [71].

IOR-ROI-LSTM [15] uses CNN and LSTM to model phenomena such as IOR and gaze shift behaviour simultaneously by using a dual LSTM unit. Further improvement was achieved by including semantic segmentation masks from Mask^X-RCNN [53]. An MDN layer was also added to model fixation location and duration distributions.

GazeGenNet [78] using LSTM and Mixture Density Network (MDN) [8] was proposed to generate a large amount of data for eye-tracking classification tasks. The main difference is that it is stimuli-agnostic while our ScanpathNet generates eye-tracking data given an input image.

There are also models that generate scanpaths during task-driven search, such as categorical search [1, 74, 75, 77], target-based search [68], goal-directed search [2, 31] and visual question answering [14]. Another related work is multi-duration saliency in which saliency maps are generated for different durations [21]. While this work does not address the task-driven search problem, ScanpathNet can be extended by incorporating 'top-down' guidance in the creation of the dynamic priority map. Finally, the prediction of fixation durations is currently out of the scope of this paper.

3. Method

Our proposed model, ScanpathNet, is a deep learning model inspired by Guided Search 6 (GS6) [64], a theoretical model of visual search in neuroscience. GS6 suggests that a dynamic priority map is computed from a weighted average of five sources of guidance, including two "classical" sources, top-down and bottom-up feature guidance, and three additional sources of guidance: history, value and scene guidance. In this work, we concentrate on the free-viewing paradigm where top-down guidance is not considered (i.e., weight set to 0 in computation). Nevertheless, our model allows extensions to allow for task-based search.

The modified GS6 is illustrated in Figure 1 left, with the numbers referring to the following processes:

1. Scene information (*e.g.* contrast, colour, orientation and brightness) is encoded into the visual system.

2. Due to the limited capacity of the brain, only one (or a very few) objects can be attended to at a time. At this point, selective attention is performed to determine the regions in the image that will pass through the bottleneck. The selected target of attention is added to the visual working memory.

3. Attention selection is rarely random. It is guided by 'bottom-up' salience based on scene information and the perceived value of the previously attended region. It is also influenced by scanpath history as represented in the visual working memory. These guidances are combined in a weighted manner to generate a dynamic priority map that represents the probability distribution of fixations. Selective attention is deployed to the 'peaks' of the distribution.

The proposed architecture is illustrated in Figure 1 right. ScanpathNet consists of three major modules: Visual System (VS), Visual Working Memory (VWM) and Priority Map Generation (PMG). The VS module is based on CNN architectures trained for object recognition. The VWM module mimics the IOR mechanism in humans using convLSTM networks. Finally, the PMG module models the stochastic nature of human scanpaths using MDN.

3.1. Visual System (VS) Module

The theory of bottom-up saliency-driven attention suggests that salient regions in an image are determined heavily by visual characteristics. It has been shown previously that CNNs trained for object recognition encode rich semantic information that can be used for saliency prediction [30]. As a result, the backbones of most state-of-the-art saliency and scanpath models incorporate CNNs (*e.g.* VGG and ResNet) to extract semantic information effectively. In this paper, the VS module uses a modified VGG-16 [52] model trained on the ImageNet dataset to produce feature representations F^0 with a downscale factor of 8, by removing the last fully connected layers and the last max-pooling layer. In order to maintain a suitable spatial resolution, one upsampling layer is added, resulting in a feature map F^1.

3.2. Visual Working Memory (VWM) Module

The IOR mechanism allows humans to inhibit processing at recently visited targets to quickly analyse the environment. The VWM module utilises convLSTM [72] to simulate IOR mechanisms. ScanpathNet is different from other convLSTM models because it utilises convLSTM to learn spatio-temporal dependencies of the fixation sequence, instead of using convLSTM to refine a saliency map. Existing scanpath models that use convLSTM networks to mimic IOR mechanisms require a computation of inhibition maps from datasets [15, 53]. The novelty of ScanpathNet is that it implicitly learns the IOR mechanism by directly inhibiting features in the previous fixation locations.

The VWM module takes image features F^1 as input and multiplies it with a spatial map g centred around the previous fixation. The input to the convLSTM is defined as:

$$F_t^2 = F^1 \otimes g(x_{t-1}, y_{t-1}) \tag{1}$$

where $\otimes$ is defined as an element-wise product over the feature map F^1 from the VS Module. From Equation 1, convLSTM receives feature representations with previously visited fixation locations inhibited. Empirical tests revealed that the spatial map controls the IOR mechanism. We compared the performance of (1) a Gaussian spatial mask and (2) a spatial mask with 0 values similar to [34, 71]. The model did not converge when setting 1 was used. Hence, the IOR setting 2 was used for the succeeding experiments. This design choice is grounded in GS6 [64] where operations seem to happen in the feature space (Figure 3 in [64]).

3.3. Priority Map Generation (PMG) Module

The PMG module uses an MDN to explicitly model the stochastic nature of human attention. MDN generates probability densities of the next fixations using a mixture of K Gaussians [8]. Recently, MDNs with RNNs have been used for handwritten text generation [23], video saliency prediction [7] and sketch generation [24].

The MDN takes flattened hidden representations of the IOR module as input and produces a parameterised Mixture of Gaussians y_t as output, consisting of a set of means μ_t^i, standard deviations σ_t^i, correlations ρ_t^i and mixture weights π_t^i, for the i-th component of K mixtures. Mathematically, this can be represented as follows:

$$y_t = \left(\{\mu_t^i, \sigma_t^i, \rho_t^i, \pi_t^i\}_{i=1}^K\right) \tag{2}$$

The parameters of MDN are constrained and normalised in order to obtain a valid probability distribution [23].

$$\begin{aligned} \mu_t^i &= \hat{\mu}_t^i \\ \sigma_t^i &= exp(\hat{\sigma}_t^i) \\ \rho_t^i &= tanh(\hat{\rho}_t^i) \\ \pi_t^i &= \frac{exp(\hat{\pi}_t^i)}{\sum_{m=1}^K exp(\hat{\pi}_t^m)} \end{aligned} \tag{3}$$

The probability distribution of the next fixation location is given by:

$$p((x_{t+1}, y_{t+1})) = \sum_{i=1}^{K} \pi_t^i \mathcal{N}(x_{t+1}, y_{t+1} | \mu_t^i, \sigma_t^i, \rho_t^i) \tag{4}$$

where $\mathcal{N}$ is a bivariate normal distribution.

3.4. Training

ScanpathNet outputs a temporal probability distribution (i.e. priority map) of fixations. To generate priority maps, the 2D probability distributions of all Gaussians are combined into a saliency map (refer to eq. 4). Instead of aggregating all fixation locations from all subjects across all times into a single fixation map, all fixations are aligned temporally and a Gaussian mask is applied to generate a temporal sequence of priority maps. The number of priority maps per image was chosen to be 6, similar to prior works [36, 46], but could be easily extended and tuned to the application.

The loss function used to train the entire model is binary cross entropy and is defined below:

$$\mathcal{L}_{\mathcal{BCE}} = -\frac{1}{N} \sum_{j=1}^{N} S_j log(\tilde{S}_j) + (1 - S_j) log(1 - \tilde{S}_j) \tag{5}$$

where S_j and $\tilde{S}_j$ are the ground truth and predicted saliency maps respectively.

3.5. Human Scanpath Generation

To generate the initial priority map, the model extracts a rich representation from the image with a Gaussian blur applied in the middle of the feature space. To generate the next fixation location, the output probability distribution (i.e., $[(x_1, y_1), (x_2, y_2), ..., (x_K, y_K)]$ is randomly sampled. For the next fixation, the model uses the feature map F^1 where the fixation location (x_{t-1}, y_{t-1}) is masked.

4. Experiments

4.1. Experiment Settings

Dataset. We performed our experiments on the OSIE [73], MIT1003 [37] and CAT2000 [11] datasets. The OSIE dataset consists of 700 natural indoor and outdoor scenes with eye-tracking data collected from 15 participants. All images have 600x800 resolution. The MIT1003 dataset consists of 1003 images with eye-tracking data collected from 15 participants. There are 779 landscape images and 228 portrait images with varying resolutions. Finally, the CAT2000 dataset consists of 2000 training images and 2000 test images from 20 different categories containing 100 images. For the CAT2000 dataset, only the train set was used since the eye-tracking data for the test set were held out. No data augmentation was performed and scanpaths with lengths less than the mean scanpath length recorded in the dataset were discarded. The datasets were randomly split into 80% training data and 20% test data.

Implementation details. All images were resized to a resolution of 300x400 pixels. The VS module generates F^0 with 512 feature maps of size 18x25. The upsampling layer scales F^0 five times, resulting in a feature map F^1 of size 90x125. The spatial map g has the same resolution as F^1 and σ value set to 5. The VWM module applies a 2x2 convolutional filter with a stride of 1 to the inhibited features F^2, resulting in a single channel representation. The number of Gaussians in the MDN layer was determined empirically and their performances are reported. For the VWM module, the pre-trained weights of the VGG-16 model on the ImageNet dataset were fixed and only the convLSTM and MDN layers were trained. The model was trained with an Adam optimizer (learning rate = 0.001) in an end-to-end supervised manner. Early stopping was performed.

State-of-the-art models. Comparison of results against state-of-the-art traditional (e.g. Itti [34], SGC [54, 55], STAR-FC) and deep learning approaches (e.g. SaltiNet [5], PathGAN [3], IOR-ROI [53], VQA [14]) was conducted. Default parameters and available pre-trained weights for all models were used. Mask^X-RCNN [29] with a threshold of 0.5 was used to compute semantic segmentation masks.

Evaluation Metrics. Different scanpath methods exist: visual inspection [9, 12] and comparisons (statistics [20], fixation density [55] and distance [17, 35]). In this paper, ScanMatch [17] and MultiMatch [35] were used for evaluation. Similar to previous works [14, 15, 53], we used the following evaluation strategy: 10 scanpaths with lengths equivalent to the mean scanpath length of the dataset were generated for each image. Each predicted scanpath was then compared to all subjects' recorded scanpaths. Human performance was also computed by measuring the similarity of every pair of ground truth scanpaths recorded in each image. The reported results are the average metric scores.

4.2. Quantitative Evaluation

MultiMatch and ScanMatch scores on three widely-used benchmark datasets are shown in Table 1. We performed sensitivity analysis on K, the number of Gaussian components. Table 1 shows that there is a slight variation in MultiMatch and ScanMatch scores of ScanpathNets with different K values. Nevertheless, ScanpathNets with K = 15 to 25 reported the highest scores in most metrics in all datasets. Since ScanpathNet does not predict fixation duration, the Duration item of the MultiMatch metric is not reported. Higher values on each item denote better performance. In addition, scores close to the human performance suggest the model's ability to generate realistic human scanpaths.

Our model achieved similar MultiMatch performance (with $\approx$ 1 to 4% difference) to VQA [14] on OSIE. Our model also scored highest in most MultiMatch metrics on MIT1003 and CAT2000 datasets.The high scores achieved by our model suggest that it generates scanpaths that have the same shape similarity as the human scanpaths. It also implies that the predictions have similar directions (i.e., small angular difference) as the human scanpaths. Furthermore, it denotes that our model generates proportional saccade amplitude (i.e., small absolute difference between the aligned predicted and ground truth). Finally, our proposed model predicted fixation locations that are positioned near the human fixation locations. This is evident when the Multimatch scores are compared to human performance.

VQA [14] achieved the best ScanMatch score on OSIE. Our model achieved similar ScanMatch performance (with only a 0.3% difference in score) to IOR-ROI [53] on MIT1003. Our model scored 3% higher on CAT2000 and also close to human performance in terms of the ScanMatch metric, suggesting the realistic prediction of our model.

4.3. Qualitative Evaluation

As mentioned previously, the number of Gaussian components, K, slightly affects the scanpath prediction performance. The generated scanpaths of different ScanpathNets with different K values are illustrated in Figure 2. Here, we observed that ScanpathNets with a lower value of K can only attend to a small number of objects in the image. This is evident in complex scenes where the generated fixation locations are positioned in only a small number of regions in the image. Higher K values allow the model to focus on more regions of interest, resulting in better performance on both simple and complex scenes.

What does the model see? To further examine the effect of K on performance, we visualised the predicted dynamic priority maps of different ScanpathNet as shown in Figure 3. To conduct a valid comparison, ScanpathNets with different K values are given the same image and same ground truth scanpath. The rationale is that ScanpathNets should generate priority maps similar to human priority maps.

Table 1. MultiMatch (Vector (V), Direction (D), Length (L) and Position (P)) and ScanMatch (Value) results on different benchmark datasets. Higher values denote better performance. Human performance is also reported. The best scores are in **bold**.

	OSIE Dataset					MIT1003 Dataset					CAT2000 Dataset				
Model	V ↑	D ↑	L ↑	P ↑	Value ↑	V ↑	D ↑	L ↑	P ↑	Value ↑	V ↑	D ↑	L ↑	P ↑	Value ↑
Human	0.940	0.697	0.928	0.853	0.439	0.925	0.706	0.920	0.852	0.374	0.950	0.699	0.939	0.834	0.494
Itti	0.855	0.631	0.810	0.704	0.239	0.848	0.628	0.819	0.701	0.212	0.903	0.604	0.875	0.658	0.287
SGC	0.922	0.644	0.903	0.719	0.223	0.911	0.632	0.886	0.700	0.177	0.948	0.646	0.929	0.663	0.341
Star-FC	0.931	0.662	0.910	0.777	0.334	0.917	0.642	0.902	0.781	0.299	0.949	0.655	0.936	0.817	0.406
SaltiNet	0.894	0.650	0.871	0.734	0.189	0.886	0.651	0.873	0.738	0.234	0.924	0.620	0.907	0.765	0.292
PathGAN	0.909	0.540	0.915	0.808	0.278	0.915	0.589	0.844	0.774	0.278	0.926	0.556	0.934	0.836	0.406
IOR-ROI	0.915	**0.725**	0.893	0.831	0.400	0.911	**0.719**	0.904	0.815	**0.339**	0.924	**0.689**	0.901	0.827	0.389
VQA	**0.946**	0.671	**0.928**	**0.882**	**0.445**	0.918	0.656	0.879	0.831	0.311	**0.955**	0.649	0.936	0.837	0.217
Ours (K=5)	0.924	0.655	0.909	0.793	0.312	0.912	0.680	0.825	0.784	0.330	0.944	0.651	0.929	0.833	0.408
Ours (K=10)	0.921	0.669	0.903	0.811	0.337	0.908	0.685	0.894	0.823	0.332	0.942	0.650	0.929	0.827	0.399
Ours (K=15)	0.916	0.672	0.897	0.812	0.339	0.904	0.689	0.890	0.823	0.330	0.950	0.658	**0.938**	**0.838**	0.431
Ours (K=20)	0.918	0.673	0.900	0.813	0.350	**0.920**	0.675	**0.906**	**0.831**	0.336	0.938	0.639	0.922	0.797	**0.438**
Ours (K=25)	0.931	0.663	0.916	0.837	0.322	0.906	0.684	0.893	0.822	0.329	0.945	0.659	0.933	0.833	0.408

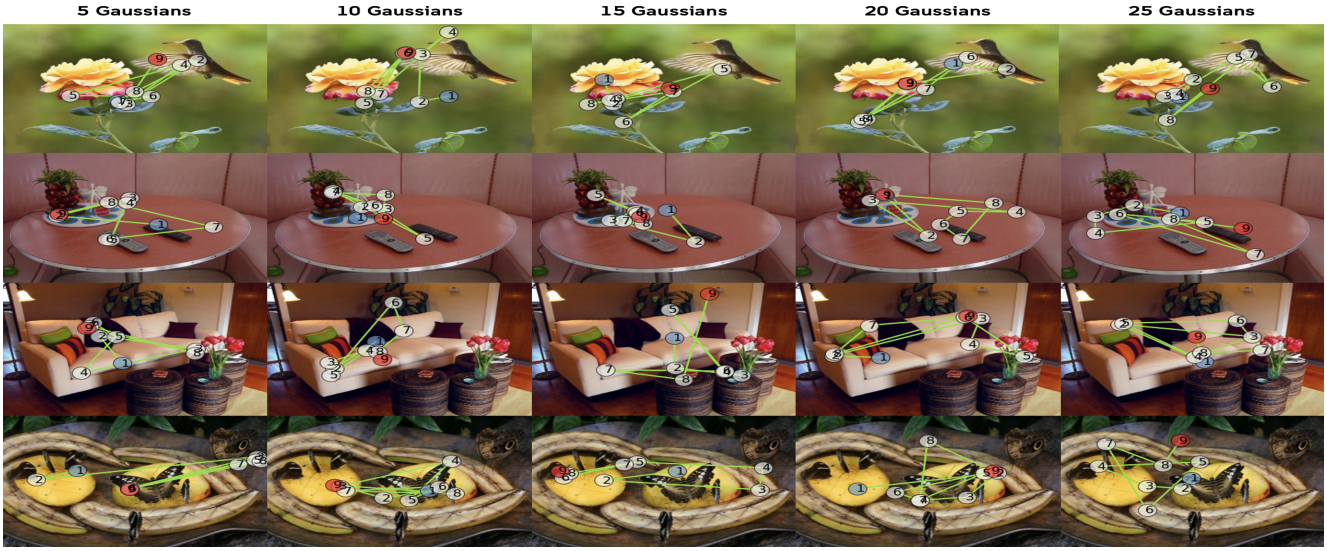

Figure 2. Each row represents an image with corresponding generated scanpaths of different ScanpathNets on each column.

ScanpathNet with K=5 can only attend to a few regions of interest and have Gaussian components with high standard deviations. As the value of K increases, the 'attended' area increases and the standard deviations of each Gaussian component decreases to 'fit' more components into the image. As a result, ScanpathNets with higher K values generate priority maps that are closer to the ground truth, supporting the quantitative results. While there are Gaussian components that seem to be misplaced, their mixing coefficients (π) have smaller values. This explains why ScanpathNets with higher K values perform better quantitatively and qualitatively on both simple and complex scenes.

Figure 3 shows that the IOR mechanism in humans and ScanpathNet output appears to be weak (i.e., there is saliency around previously attended objects), suggesting that there is an alignment between our model and GS6 and also with human behaviour. In fact, GS6 abandons the idea of sampling without replacement [64], allowing previously attended locations to be revisited. In contrast, most existing scanpath models seem to prioritise exploring the scene first than revisiting fixation locations, as shown in Figure 4. While more experiments are necessary to confirm this, our results suggest that GS6 provides a good theoretical framework to advance the development of scanpath models.

Figure 3. Row 1: Input image with human scanpath. Row 2: Input image masked with aggregate human priority maps at different time points. Rows 3-5: Input image masked with predicted priority maps of different ScanpathNets at different timepoints.

Comparisons against the state-of-the-art. Different images with human and generated scanpaths are shown in Figure 4 (extensive comparisons are in the supplementary material). The images are arranged by increasing scene complexity and the number of dominant objects in the image. The human scanpaths were randomly chosen. The generated scanpaths that achieved the highest ScanMatch score were selected. The ScanpathNet with the best K value for each dataset was chosen. Fixation locations are numbered in temporal order starting from 1 until the last fixation.

Our model closely resembles the ground truth scanpath in terms of fixation locations and order, as shown in Figure 4. VQA [14] performs similarly but sometimes tends to fixate on a few regions of the image. Since the IOR-ROI [53] uses semantic segmentation masks as an input, it can capture objects that may attract human attention. This is generally effective but fails in cases where there are a few dominant objects in a complex background (images 1 and 3), or where there may be no objects available (image 6). This may explain its lower performance on the CAT2000 dataset, where some images fall under categories, such as Art, Fractal and Pattern, that may not have any detected semantic segmentation masks. Similar to previous published qualitative results [53], PathGAN [3] generated scanpaths subjectively different from the human scanpaths. Although better than PathGAN, SaltiNet [5] still generates fixation locations that do not subjectively align with the most dominant object in the image. Star-FC [61] and SGC [54, 55] generate plausible scanpaths on less complex images but tend to fixate on only one or two parts on complex images. Finally, Itti's model [34] generates scanpaths that visit regions of interest without revisiting previous fixations. This results in scanpaths that span the whole image.

Limitations. Quantitative and qualitative results show that the performance of ScanpathNet is sensitive to the value of K, the number of Gaussian components. This sensitivity seems to depend on scene complexities in a given dataset. Nevertheless, experimental results revealed that ScanpathNet with K values from 15 to 25 still performed better or close to state-of-the-art scanpath models. While our model did not consistently score the highest in all metrics against the other existing scanpath models on all datasets, our model is simpler, more interpretable and intuitive to understand compared to most existing scanpath models.

Figure 4. Visualisation of the generated scanpaths from each scanpath model on images with increasing complexity.

5. Conclusion

In this work, we presented ScanpathNet, a deep learning model inspired by Guided Search 6 (GS6). This latest framework in neuroscience suggests that visual search is 'guided' by different factors: the classical top-down and bottom-up guidance, fixation history, value and scene guidance. Since this paper focusses on free-viewing tasks, a modified GS6 model with no top-down guidance was presented. ScanpathNet is composed of three major modules: (1) a Visual System modelled by a CNN extracts rich information from a given image (2) a Visual Working Memory is implemented using convLSTM that takes in the output of the Visual System and inhibits previous fixation locations and (3) a Priority Map Generation Module is modelled using MDNs that combine the semantic information and fixation history to generate a dynamic priority map. Human scanpaths are generated by sampling this dynamic priority map. Experimental results show that ScanpathNet generates scanpaths that are similar to human scanpaths. Despite its simplicity, ScanpathNet has shown promising qualitative and quantitative results in scanpath prediction.

References

[1] Hossein Adeli, Françoise Vitu, and Gregory J Zelinsky. A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *Journal of Neuroscience*, 37(6):1453–1467, 2017. 2, 3

[2] Hossein Adeli and Gregory Zelinsky. Deep-bcn: Deep networks meet biased competition to create a brain-inspired model of attention control. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 1932–1942, 2018. 3

[3] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018. 2, 5, 7

[4] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Scanpath and saliency prediction on 360 degree images. *Signal Processing: Image Communication*, 69:8–14, 2018. 2

[5] Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 2331–2338, 2017. 2, 5, 7

[6] Wentao Bao and Zhenzhong Chen. Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing*, 404:154–164, 2020. 2

[7] Loris Bazzani, Hugo Larochelle, and Lorenzo Torresani. Recurrent mixture density network for spatiotemporal visual attention. *arXiv preprint arXiv:1603.08199*, 2016. 4

[8] Christopher M Bishop. Mixture density networks. 1994. 3, 4

[9] Giuseppe Boccignone and Mario Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1-2):207–218, 2004. 2, 5

[10] Ali Borji. Saliency prediction in the deep learning era: Successes and limitations. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 2

[11] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *arXiv preprint arXiv:1505.03581*, 2015. 5

[12] Dirk Brockmann and Theo Geisel. The ecology of gaze shifts. *Neurocomputing*, 32:643–650, 2000. 2, 5

[13] Liang-Chieh Chen, George Papandreou, Iasonas Kokkinos, Kevin Murphy, and Alan L Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2017. 2

[14] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10876–10885, 2021. 3, 5, 7

[15] Zhenzhong Chen and Wanjie Sun. Scanpath prediction for visual attention using ior-roi lstm. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 642–648, 2018. 3, 4, 5

[16] Alasdair DF Clarke, Matthew J Stainer, Benjamin W Tatler, and Amelia R Hunt. The saccadic flow baseline: Accounting for image-independent biases in fixation behavior. *Journal of vision*, 17(11):12–12, 2017. 2

[17] Filipe Cristino, Sebastiaan Mathôt, Jan Theeuwes, and Iain D Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42(3):692–700, 2010. 5

[18] Ryan Anthony Jalova de Belen, Tomasz Bednarz, and Arcot Sowmya. Eyexplain autism: Interactive system for eye tracking data analysis and deep neural network interpretation for autism spectrum disorder diagnosis. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7, 2021. 1

[19] Ryan Anthony J de Belen, Tomasz Bednarz, Arcot Sowmya, and Dennis Del Favero. Computer vision in autism spectrum disorder research: a systematic review of published studies from 2009 to 2019. *Translational psychiatry*, 10(1):1–20, 2020. 1

[20] Ralf Engbert, Hans A Trukenbrod, Simon Barthelmé, and Felix A Wichmann. Spatial statistics and attentional dynamics in scene viewing. *Journal of vision*, 15(1):14–14, 2015. 2, 5

[21] Camilo Fosco, Anelise Newman, Pat Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, and Zoya Bylinskii. How much time do you have? modeling multi-duration saliency. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4473–4482, 2020. 3

[22] Wolfgang Fuhl, Yao Rong, and Enkelejda Kasneci. Fully convolutional neural networks for raw eye tracking data segmentation, generation, and reconstruction. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 142–149. IEEE, 2021. 3

[23] Alex Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013. 4

[24] David Ha and Douglas Eck. A neural representation of sketch drawings. *arXiv preprint arXiv:1704.03477*, 2017. 4

[25] Yiyuan Han, Bing Han, and Xinbo Gao. Human scanpath estimation based on semantic segmentation guided by common eye fixation behaviors. *Neurocomputing*, 453:705–717, 2021. 3

[26] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016. 2

[27] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 2

[28] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: Searching for coding length increments. 2009. 2

[29] Ronghang Hu, Piotr Dollár, Kaiming He, Trevor Darrell, and Ross Girshick. Learning to segment every thing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4233–4241, 2018. 5

[30] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by

adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015. 4

[31] Mobarakol Islam, VS Vibashan, Chwee Ming Lim, and Hongliang Ren. St-mtl: Spatio-temporal multitask learning model to predict scanpath while tracking instruments in robotic surgery. *Medical Image Analysis*, 67:101837, 2021. 3

[32] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000. 1

[33] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001. 2

[34] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 1, 2, 4, 5, 7

[35] Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. A vector-based, multidimensional scanpath similarity measure. In *Proceedings of the 2010 symposium on eye-tracking research & applications*, pages 211–218, 2010. 5

[36] Ming Jiang, Xavier Boix, Gemma Roig, Juan Xu, Luc Van Gool, and Qi Zhao. Learning to predict sequences of human visual fixations. *IEEE transactions on neural networks and learning systems*, 27(6):1241–1252, 2016. 2, 4

[37] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th international conference on computer vision*, pages 2106–2113. IEEE, 2009. 5

[38] Matthias Kümmerer and Matthias Bethge. State-of-the-art in human scanpath prediction. *arXiv preprint arXiv:2102.12239*, 2021. 2

[39] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet. *arXiv preprint arXiv:1411.1045*, 2014. 2

[40] M Kümmerer, TS Wallis, and M Bethge III. Deepgaze iii: Using deep learning to probe interactions between scene content and scanpath history in fixation selection. In *2019 Conference on Cognitive Computational Neuroscience*, volume 13, page 16, 2019. 2

[41] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *arXiv preprint arXiv:1610.01563*, 2016. 2

[42] Patrick Le Callet and Ernst Niebur. Visual attention and applications in multimedia technologies. *Proceedings of the IEEE*, 101(9):2058–2067, 2013. 1

[43] Olivier Le Meur and Antoine Coutrot. Introducing context-dependent and spatially-variant viewing biases in saccadic models. *Vision research*, 121:72–84, 2016. 2

[44] Olivier Le Meur, Antoine Coutrot, Zhi Liu, Pia Rämä, Adrien Le Roch, and Andrea Helo. Visual attention saccadic models learn to emulate gaze patterns from childhood to adulthood. *IEEE Transactions on Image Processing*, 26(10):4777–4789, 2017. 2

[45] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision research*, 116:152–164, 2015. 2

[46] Huiying Liu, Dong Xu, Qingming Huang, Wen Li, Min Xu, and Stephen Lin. Semantically-based human scanpath estimation with hmms. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3232–3239, 2013. 2, 4

[47] Noa Malem-Shinitski, Manfred Opper, Sebastian Reich, Lisa Schwetlick, Stefan A Seelig, and Ralf Engbert. A mathematical model of local and global attention in natural scene viewing. *PLOS Computational Biology*, 16(12):e1007880, 2020. 2

[48] Daniel Martin, Ana Serrano, Alexander W Bergman, Gordon Wetzstein, and Belen Masia. Scangan360: A generative model of realistic scanpaths for 360. *arXiv preprint arXiv:2103.13922*, 2021. 2

[49] Thuyen Ngo and BS Manjunath. Saccade gaze prediction using a recurrent neural network. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3435–3439. IEEE, 2017. 2

[50] Heiko H Schütt, Lars OM Rothkegel, Hans A Trukenbrod, Sebastian Reich, Felix A Wichmann, and Ralf Engbert. Likelihood-based parameter estimation and comparison of dynamical cognitive models. *Psychological review*, 124(4):505, 2017. 2

[51] Xuan Shao, Ye Luo, Dandan Zhu, Shuqin Li, Laurent Itti, and Jianwei Lu. Scanpath prediction based on high-level features and memory bias. In *International Conference on Neural Information Processing*, pages 3–13. Springer, 2017. 2

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014. 2, 4

[53] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE transactions on pattern analysis and machine intelligence*, 2019. 3, 4, 5, 7

[54] Xiaoshuai Sun, Hongxun Yao, and Rongrong Ji. What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In *2012 IEEE conference on computer vision and pattern recognition*, pages 1552–1559. IEEE, 2012. 2, 5, 7

[55] Xiaoshuai Sun, Hongxun Yao, Rongrong Ji, and Xian-Ming Liu. Toward statistical modeling of saccadic eye-movement and visual saliency. *IEEE Transactions on Image Processing*, 23(11):4649–4662, 2014. 2, 5, 7

[56] Hamed Rezazadegan Tavakoli, Esa Rahtu, and Janne Heikkilä. Stochastic bottom–up fixation prediction and saccade generation. *Image and Vision Computing*, 31(9):686–693, 2013. 2

[57] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 1

[58] Ashish Verma and Debashis Sen. Hmm-based convolutional lstm for visual scanpath prediction. In *2019 27th European Signal Processing Conference (EUSIPCO)*, pages 1–5. IEEE, 2019. 2

[59] Wei Wang, Cheng Chen, Yizhou Wang, Tingting Jiang, Fang Fang, and Yuan Yao. Simulating human saccadic scanpaths on natural images. In *CVPR 2011*, pages 441–448. IEEE, 2011. 2

[60] Yixiu Wang, Bin Wang, Xiaofeng Wu, and Liming Zhang. Scanpath estimation based on foveated image saliency. *Cognitive processing*, 18(1):87–95, 2017. 2

[61] Calden Wloka, Iuliia Kotseruba, and John K Tsotsos. Active fixation control to predict saccade sequences. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3184–3193, 2018. 2, 7

[62] Jeremy Wolfe, Matthew Cain, Krista Ehinger, and Trafton Drew. Guided search 5.0: Meeting the challenge of hybrid search and multiple-target foraging. *Journal of vision*, 15(12):1106–1106, 2015. 1

[63] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994. 1

[64] Jeremy M Wolfe. Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, pages 1–33, 2021. 1, 3, 4, 6

[65] Jeremy M Wolfe, Kyle R Cave, and Susan L Franzel. Guided search: an alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human perception and performance*, 15(3):419, 1989. 1

[66] Jeremy M Wolfe and Gregory Gancarz. Guided search 3.0. In *Basic and clinical applications of vision science*, pages 189–192. Springer, 1997. 1

[67] Jeremy M Wolfe and W Gray. Guided search 4.0. *Integrated models of cognitive systems*, pages 99–119, 2007. 1

[68] Jinghan Wu, Meiqi Lu, Yuping Lin, and Xuetao Zhang. Scanpaths generation for target search based on deep learning. In *2020 Chinese Automation Congress (CAC)*, pages 1443–1448. IEEE, 2020. 3

[69] Chen Xia, Junwei Han, Fei Qi, and Guangming Shi. Predicting human saccadic scanpaths based on iterative representation learning. *IEEE Transactions on Image Processing*, 28(7):3502–3515, 2019. 2

[70] Chen Xia, Fei Qi, and Guangming Shi. An iterative representation learning framework to predict the sequence of eye fixations. In *2017 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1530–1535. IEEE, 2017. 2

[71] Chen Xia and Rong Quan. Predicting saccadic eye movements in free viewing of webpages. *IEEE Access*, 8:15598–15610, 2020. 3, 4

[72] SHI Xingjian, Zhourong Chen, Hao Wang, Dit-Yan Yeung, Wai-Kin Wong, and Wang-chun Woo. Convolutional lstm network: A machine learning approach for precipitation nowcasting. In *Advances in neural information processing systems*, pages 802–810, 2015. 2, 4

[73] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 5

[74] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 193–202, 2020. 3

[75] Chen-Ping Yu, Huidong Liu, Dimitrios Samaras, and Gregory J Zelinsky. Modelling attention control using a convolutional neural network designed after the ventral visual pathway. *Visual Cognition*, 27(5-8):416–434, 2019. 3

[76] Dario Zanca, Stefano Melacci, and Marco Gori. Gravitational laws of focus of attention. *IEEE transactions on pattern analysis and machine intelligence*, 42(12):2983–2995, 2019. 2

[77] Gregory Zelinsky, Zhibo Yang, Lihan Huang, Yupei Chen, Seoyoung Ahn, Zijun Wei, Hossein Adeli, Dimitris Samaras, and Minh Hoai. Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 0–0, 2019. 3

[78] Raimondas Zemblys, Diederick C Niehorster, and Kenneth Holmqvist. gazenet: End-to-end eye-movement event detection with deep neural networks. *Behavior research methods*, 51(2):840–864, 2019. 3