Improving saliency models' predictions of the next fixation with humans' intrinsic cost of gaze shifts

Florian Kadner Tobias Thomas David Hoppe Constantin A. Rothkopf
Centre for Cognitive Science & Institute of Psychology, TU Darmstadt
{firstname.lastname}@tu-darmstadt.de

Abstract

The human prioritization of image regions can be modeled in a time invariant fashion with saliency maps or sequentially with scanpath models. However, while both types of models have steadily improved on several benchmarks and datasets, there is still a considerable gap in predicting human gaze. Here, we leverage two recent developments to reduce this gap: theoretical analyses establishing a principled framework for predicting the next gaze target and the empirical measurement of the human cost for gaze switches independently of image content. We introduce an algorithm in the framework of sequential decision making, which converts any static saliency map into a sequence of dynamic history-dependent value maps, which are recomputed after each gaze shift. These maps are based on 1) a saliency map provided by an arbitrary saliency model, 2) the recently measured human cost function quantifying preferences in magnitude and direction of eye movements, and 3) a sequential exploration bonus, which changes with each subsequent gaze shift. The parameters of the spatial extent and temporal decay of this exploration bonus are estimated from human gaze data. The relative contributions of these three components were optimized on the MIT1003 dataset for the NSS score and are sufficient to significantly outperform predictions of the next gaze target on NSS and AUC scores for five state of the art saliency models on three image data sets.

1. Introduction

Because of the inhomogeneous spatial acuity of the visual system, humans shift their gaze sequentially across visual scenes using saccadic gaze movements [18]. Four main factors have been shown to influence observers' eye movements: the ongoing task, image features such as contrast and intensity, semantic features such as faces and scene context, but also factors that arise from the sequential interaction of an observer with the scene including center bias, proximity preference, inhibition of return [52]. Empirically, gaze targets of multiple human observers while inspecting an image given different task instructions can be collected. Assuming that gaze prioritization is image-computable, the computational task of predicting gaze prioritization given an image is referred to as visual saliency modeling resulting in a time invariant saliency map, whereas scanpath models generate a sequence of gaze targets.

While originally developed to account for the phenomenon of pop-out [56], visual saliency modeling has been generalized to predicting the likelihood of human observers looking at image regions for arbitrary images. Initially, saliency models used handcrafted features inspired by neurophysiological properties of the visual system [26, 50], but more recently data driven approaches [25, 33, 45, 37, 13, 59, 17, 28] have commonly used features from DNNs pretrained on large image datasets, thereby improving performance on various benchmarks [7, 6]. These improvements are due to the rich image structure learned by DNNs when trained on large image datasets, e.g. the VGG19 [49] underlying Deep Gaze II [37] is trained on object recognition of one Million images before tuning to saliency problems using the SALICON dataset [30] containing 10000 images.

Scanpath models, by contrast, take an image as input and generate a full scanpath, i.e. a sequence of individual fixation locations as output [58, 5, 40, 62, 3, 2]. Progress on a variety of benchmarks and image databases has been made, but, a fundamental difficulty with scanpath models compared to saliency models is the well known variability of gaze sequences between observers, which poses particular challenges for evaluating predictions. A recent comprehensive theoretical and empirical evaluation of scanpath models [34] has revealed that common scanpath similarity metrics can score wrong models better than the true generating model. The resulting analysis in [34] establishes that a more consistent and meaningful task consists in the prediction of the next fixation target conditional on past fixations within an image, which is the task we adopt in this study.

Here, we leverage two recent developments to improve the prediction of the next fixation of human observers given an arbitrary saliency map and the sequence of preceding fixations: the theoretical analysis of scanpath models [34] and the recently measured human cost function for gaze shifts [54]. We adopt a computational account of the scanpath as a sequential decision process in the spirit of previous approaches [29, 41, 24], but differently from these approaches, our algorithm can utilize arbitrary saliency maps as input instead of estimating rewards for image features. First, we reason that saliency corresponds to the reward associated with the free-viewing task which is approximated by marginalizing over all visual tasks. The reason is, that the free-viewing task is maximally ambiguous regarding its task goal. Second, our formulation allows incorporating a map representing the human preferences for gaze shifts, which have recently been estimated for the first time independently of image content through a human psychophysical experiment [54]. This gives a computational explanation for commonly used heuristics including the proximity preference. Third, we account for past fixations through a temporally changing exploration map and present the resulting predictions of subsequent gaze targets. The relative contributions of these three components were optimized on the MIT1003 dataset for the NSS score and are sufficient to significantly outperform predictions of the next gaze target on NSS and AUC scores for five state of the art saliency models on three image data sets.

2. Related Work

The concept of saliency lies at the intersection of cognitive science, neuroscience, and computer vision [52]. Empirically, human gaze targets depend strongly on the ongoing task [20] but humans tend to look preferentially at certain areas even when free-viewing images [22]. These observations have been complemented with the discoveries of multiple retinotopic maps in the visual system [57]. While the exact relationships between attention, gaze sequences, and their neuronal underpinnings are still heavily debated, visual saliency modeling has become a canonical computer vision task.

Initially, saliency models used handcrafted lower level features like intensity, color, and orientation [26], whereas current DNN based algorithms determine salient regions in a data driven fashion by reusing learnt features e.g from CNNs [33]. Other studies have emphasized the importance of higher level information in images, such as text and faces [12, 10] or general semantic content [21, 46]. Relevance might be biased, e.g. towards text [2]. Some approaches have incorporated task goals into models of gaze selection [44, 8], albeit with a small number of tasks with respect to the broad range of human visual and visuomotor tasks. Semantic information has been incorporated by neural network approaches, for example by pretraining on object recognition [37, 25, 45, 13, 33, 59]. The work on attention in DNNs, e.g [42, 60] is somewhat complimentary [1], as it is not necessarily modeling overt shifts of attention by gaze shifts but sequential processing of internal representations. Overall, visual saliency modeling is an established field with canonical datasets and benchmarks and progress on these benchmarks has been steady [7, 6].

Scanpath models have received less attention compared to saliency models but recent approaches include models based on biological and cognitive facts [26, 58, 64], statistically motivated models [5, 40, 62], and models, which leverage machine learning techniques for prediction without reference to underlying mechanisms of vision [3, 2]. While some algorithms require an image as input [26, 37, 50, 17, 28, 14, 16], other models use a saliency map as input for generating a scanpath [26, 58, 64, 3, 2, 4, 5, 50, 62]. In [5, 4] the authors investigated the properties of scanpaths as function of parameters in random walks on saliency maps. [58] proposed a model incorporating an image representation map based on filter responses, foveation, and a memory module to generate sequential saliency maps. [50] used an algorithm based on projection pursuit to select image targets for simulating scanpath in order to mimic the sparsity of human gaze selection. PathGAN [2] extracts DNN features and trains recurrent layers to generate scanpaths in a training set. While PathGAN learned scanpaths end-to-end and outperformed several other models, qualitative results suggest persistent deviation to scanpaths of human observers.

Of particular relevance in this context is recent work on characterizing and evaluating the prediction accuracy of scanpath models relative to human gaze [34]. The authors' in depth analyses show that some scapath similarity metrics such as ScanMatch [15] or MultiMatch [27] can score wrong models better than the generating model given ground truth. Note also, that some of the current scanpath models employ statistics of scanpaths as a means to capture behavioral biases of gaze shifts, but these have so far never been measured independently of image content. The in depth analyses in [34] convincingly lead to the conclusion, that instead of comparing entire scanpaths it is more adequate to evaluate models regarding their prediction of the next fixation within a given scanpath.

Finally, scanpaths have also been conceptualized as sequential decision problems, which is particularly successful in situations where observers' goals are known and can therefore be formalized as rewards [42, 23, 24]. Very much related to the present approach, [41] used inverse RL to estimate implicit rewards from human gaze sequences. While this extracts reward functions in terms of image features, it is agnostic in relation to internal, behavioral costs and benefits. Other studies have used RL to predict scanpaths [29] with a state consisting of low-level features, semantic features, center bias, spatial distribution of eye-fixation shifts as well as a measure indicating previous gaze visits. However, these studies did not utilize the human cost for eye movements measured independently of image content and predicted 'fixation stages' in an experiment and not individual fixations. But empirical studies have shown, that oculomotor biases are not independent of image content, e.g. by simply rotating images [19]. Here, our goal is to leverage the recently measured human cost for making a gaze shift independently of image content [54] for arbitrary saliency models so that we do not infer the rewards of image features from scratch. This allows arbitrary saliency models to improve their predictions of the next gaze target by incorporating the intrinsic costs of a gaze shift in human observers, which interact with the prioritization of image content.

3. One-step ahead prediction model

Our general model is based on statistical decision theory and gaze sequences are viewed as reward-driven behavioral sequences, that can be described using a Markov Decision Process (MDP), similar to [41, 29, 24]. A scanpath is a sequence of gaze locations $\mathbf{x_0}, \mathbf{x_1}, \dots, \mathbf{x_t}$ visited on an Image I through movement of the visual apparatus. Each of the visited gaze locations is the result of a decision for that particular location, following a policy $\pi(s) = \arg\max_{\mathbf{x_{t+1}}} Q(s, \mathbf{x_{t+1}})$, where $Q(s, \mathbf{x_{t+1}}) = \mathbb{E}\left[G \mid s, \mathbf{x_{t+1}}\right]$ are the Q-values, i.e. the expected discounted total future rewards $G_t = \sum_{i=1}^{N} \gamma^{i-1} r_{t+i}$ when switching gaze to a location $\mathbf{x_{t+1}}$ while being in state s. The state s summarizes relevant factors that contribute to the selection of the next action $\mathbf{x_{t+1}}$, γ is the discount factor, and N is the total number of gaze shifts viewing an image. When exploring an image I, action selection is affected by past eye movements as well as the image, therefore $s = (I, \mathbf{x_0}, \dots, \mathbf{x_t})$. For some tasks, s might also include further task-relevant features or it could represent a belief state.

As has been shown repeatedly in the past, human action selection, in particular the generation of eye movements, is driven by multi-dimensional reward structures. However, the precise composition of the sources of rewards is usually unknown or not easy to measure. Here, we consider three components that have been shown to drive action selection: task-related reward, behavioral costs, and sequential effects related to the history of previous actions. In order to compute the state-action values $Q_{\text{task}}(s, \mathbf{x_{t+1}}) = \mathbb{E}\left[G \mid s, \mathbf{x_{t+1}}\right]$ in a specific task, we need to specify the rewards:

$$\begin{aligned} r(s, \mathbf{x_{t+1}}) = {} & w_0 r_{\text{task}}(s, \mathbf{x_{t+1}}) + w_1 r_{\text{internal}}(s, \mathbf{x_{t+1}}) \\ & + w_2 r_{\text{fixation history}}(s, \mathbf{x_{t+1}}) \end{aligned} \tag{1}$$

where $\mathbf{x_{t+1}}$ is a potential next eye movement location and r_{internal}, r_{task} and $r_{\text{fixation history}}$ are components contributing to the state-action value.

3.1. Saliency in the context of rewards

One dimension contributing to action selection is task-related reward. Eye movements have been shown to be carried out to lead to high rewards in their respective tasks, such as visual search [43], image classification [48], and can even be planned [24]. For free viewing of natural images, however, the reward function is difficult to obtain theoretically because the task instructions are highly ambiguous: "Just look around.". Here, we conjecture that saliency can be thought of as an average reward over all possible states within all possible tasks as we will formulate in the following. One possible approach is to formulate the task-related reward structure of free viewing as the result of marginalizing over all possible tasks:

$$\begin{aligned} r_{\text{free view}}(s, \mathbf{x_{t+1}}) &= \mathbb{E}_{\text{task}}\left[\mathbb{E}_{s_{\text{task}}}\left[r_{\text{task}}(I, s_{\text{task}}, \mathbf{x_{t+1}})\right]\right] \\ &= \mathbb{E}_{\text{task}}\left[\int\limits_{s_{\text{task}}} r_{\text{task}}(I, s_{\text{task}}, \mathbf{x_{t+1}}) p(s_{\text{task}}) ds_{\text{task}}\right] \\ &= \sum_{\text{task}} \int\limits_{s_{\text{task}}} r_{\text{task}}(I, s_{\text{task}}, \mathbf{x_{t+1}}) p(s_{\text{task}})\, ds_{\text{task}}\, p(\text{task}) \\ &\approx S(I, \mathbf{x_{t+1}}) \end{aligned} \tag{2}$$

where $r_{\text{task}}(I, s_{\text{task}}, \mathbf{x_{t+1}})$ denotes the reward when performing eye movement $\mathbf{x_{t+1}}$ in image I under a specific task while being in state s_{task}. The state s_{task} summarizes all relevant information about the actions performed prior to the current decision for a specific task. The probability distribution over potential tasks $p(\text{task})$ weights the task-dependent reward according to how likely the task is. For example, information that is relevant for many visual tasks, e.g., faces, receives higher weights. Finally, $p(s_{\text{task}})$ is the probability distribution of the current state within a task, i.e. the action sequence (scanpath) prior to the current fixation and $S(I, \mathbf{x})$ is the saliency score.

In conclusion, we view the saliency of an image location during free viewing as the approximate average reward of that location across all possible tasks and all possible previous gaze shifts in that task.

3.2. Influence of past fixations

According to Equation 2 we can approximate the task related component to the reward using the predictions of a saliency model. However saliency models are time-invariant, depending only on the image. Here, we propose an extension to overcome this problem and compute saliency models taking into account past actions. Our approach is based on the fact that the next fixation depends on prior fixations. Since the exact nature of this relationship is unknown, we developed a model that quantifies the influence of a fixation within a gaze sequence on the selection of future fixation choices:

$$r_{\text{fixation history}}(s, \mathbf{x_{t+1}}) = r(\mathbf{x_0}, \ldots, \mathbf{x_{t-1}}, \mathbf{x_t}, \mathbf{x_{t+1}}) = \sum_{i=0}^{t} \phi_i \mathcal{N}(\mathbf{x_{t+1}}; \mathbf{x_i}, \mathbf{\Sigma}) \quad (3)$$

Positive values ($\phi_i > 0$) indicate that having visited location $\mathbf{x_i}$ at timestep i during the same scanpath increases the probability of targeting the next fixation to location $\mathbf{x_i}$. Negative values lead to reduced probabilities, therefore corresponding to an effect such as a spatial version of inhibition of return.

This reward can be conceptualized as the trade-off between exploration and exploitation, i.e. a reward for either parts of the state-space that have never been explored, or, if the environment can change over time, have not been explored recently [51]. Equivalently, this reward can be formulated as an exploration bonus. Therefore, this part of the reward structure encourages an agent to try long-ignored actions, i.e. visit locations that have not been visited yet or have not been visited in a long time. Since the exact nature of this relationship is yet to be understood, we estimated the parameters ϕ_i from the eye movement data. Note that we did not constrain the parameters to sum up to one, to allow both positive and negative values for already visited or not recently visited regions, in principle.

3.3. Oculomotor preference map

Saliency models commonly neglect the agent's effort expended in the actual action to gain visual information, although such internal costs influence gaze shifts [23, 24]. These costs and benefits have their origin in the effort to produce the movement, which includes cognitive costs such as deciding upon where to move next [24] and when [23]. The oculomotor preferences were recently measured independently of image content in a psychophysical experiment involving a preference elicitation paradigm [54]. Subjects repeatedly chose between two visual targets by directing gaze to the preferred target. For each choice, three properties were manipulated for both targets: the distance to the current fixation location, the absolute direction to the target (e.g., left), and the angle relative to the last saccade. Using the decisions we inferred the value of each component and integrated them in an oculo-motor preference map. This map assigns behavioral costs to each possible gaze location dependent on the last two fixations:

$$\begin{aligned} r_{\text{internal}}(s, \mathbf{x_{t+1}}) &= r_{\text{internal}}(\mathbf{x_{t-1}}, \mathbf{x_t}, \mathbf{x_{t+1}}) \\ &= \psi_0 \left(\|\mathbf{x_{t+1}} - \mathbf{x_t}\|\right) \\ &+ \psi_1 \arccos\left(\frac{(\mathbf{x_{t+1}} - \mathbf{x_t}) \cdot (\mathbf{x_t} - \mathbf{x_{t-1}})}{\|\mathbf{x_{t+1}} - \mathbf{x_t}\| \|\mathbf{x_t} - \mathbf{x_{t-1}}\|}\right) \\ &+ \psi_2 \arccos\left(\frac{(\mathbf{x_{t+1}} - \mathbf{x_t}) \cdot [1 \quad 0]}{\|\mathbf{x_{t+1}} - \mathbf{x_t}\|}\right) \end{aligned} \quad (4)$$

3.4. Approximating the value map

We proposed three factors contributing to the final reward of an image location: task-related reward (Equation 2; approximated through saliency), fixation history (Equation 3) and the oculomotor costs (Equation 4). To account for the sequential nature of visual scanpaths we extend static saliency approaches using an additional reward component, which is an exploration part based on past fixations. Note however, that only the reward component associated with the free viewing task is dependent on the image content whereas both the internal costs and the fixation history dependent part are independent of the image content.

By consistently formulating the components as rewards we can combine them to yield the desired reward function:

$$\begin{aligned} r(s, \mathbf{x_{t+1}}) &= w_0 r_{\text{free view}}(s, \mathbf{x_{t+1}}) + w_1 r_{\text{internal}}(\mathbf{x_{t-1}}, \mathbf{x_t}, \mathbf{x_{t+1}}) \\ &+ w_2 r_{\text{fixation history}}(\mathbf{x_0}, \ldots, \mathbf{x_{t-1}}, \mathbf{x_t}, \mathbf{x_{t+1}}) \\ &\approx w_0 S(I, \mathbf{x_{t+1}}) + w_1 \sum_{i \in \{0,1,2\}} \psi_i(\mathbf{x_{t-1}}, \mathbf{x_t}, \mathbf{x_{t+1}}) \\ &+ w_2 \sum_{i=0}^{t} \phi_i \mathcal{N}(\mathbf{x_{t+1}}; \mathbf{x_i}, \mathbf{\Sigma}) \end{aligned} \quad (5)$$

The parameters w_0, w_1, w_2 are linear weights and control the trade-off between task-related rewards, fixation history dependent rewards and internal costs and were estimated from the data. We set w_0 equal to 1, since the scale of our final value map does not matter and for the purpose of interpretability of the other parameters.

Computing the optimal policy in the MDP framework according to the reward function specified in Equation 5 would now require knowledge of the transition function, i.e. the state dependent gaze dynamics and their associated stochasticity. Similarly, knowledge of sensory uncertainties would be needed across all possible tasks in order to find the optimal gaze shift policy within the POMDP framework. Unfortunately, both these approaches are unfeasible. Instead, we use the common approximation of selecting the optimal one-step look-ahead action, i.e. greedy approximation by selecting the action that maximizes the reward for a single subsequent gaze shift.

The approximate value map depends on the image (through S), on the location of the last fixation (through the internal costs) and on the entire sequence of past fixations (through the history dependent part). Crucially, as a consequence, the value map changes with every new fixation. The procedure of the computation of Q is illustrated in Figure 1 and examples of the respective maps for a succession of fixations is shown in Figure 2. Based on the approximate value map we can predict the future fixation locations from the policy π based on the value map $Q(s, \mathbf{x_{t+1}})$, see Algorithm 1.

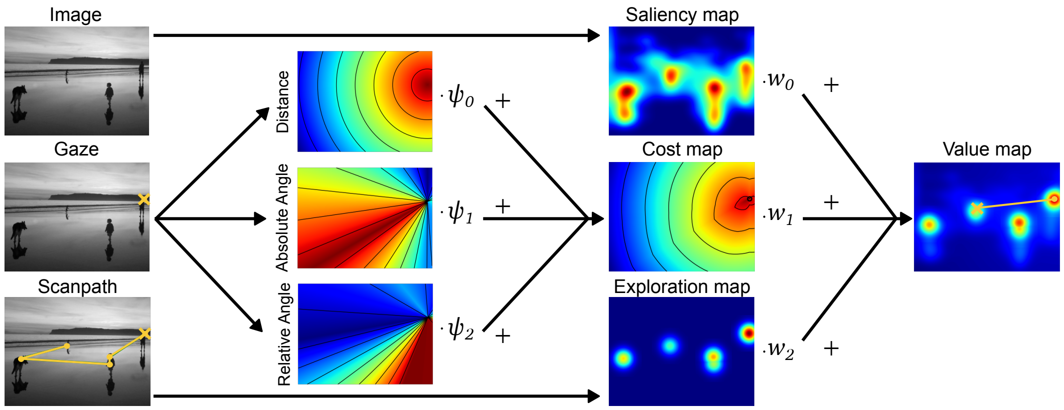

Figure 1: Schematic of the algorithm. An arbitrary saliency map and the scanpath with the current gaze position are the input. Output is a value map, which integrates the saliency map, the recomputed map for the cost of gaze shifts, and the sequential history dependent map. Note that the original image is not an input to the algorithm.

4. Experiments

First, to demonstrate the utility of our algorithm in improving the prediction of the next fixation of human observers for arbitrary saliency models, our model was implemented with four different underlying saliency models, which are currently among the ten best on the MIT/Tuebingen saliency benchmark [36] with respect to several evaluation metrics: DeepGaze II [37], SAM-ResNet [13], EML-NET [28] and CASNet II [17].

The parameters describing the three components of the behavioral costs for gaze switches corresponding to internal motor and cognitive costs were recently estimated in a psychophysical experiment from eye movement data collected in a preference elicitation paradigm [54] [1]. We collected a total of 70643 gaze shifts across 14 subjects following the experimental paradigm described in [54]. Values for the cost dimensions saccade amplitude, relative angle, and absolute angle were estimated using a random utility model [55]. The utility function was computed as the weighted sum of the individual dimensions.

[1] The estimated cost structure is available from [54]

Algorithm 1 Compute history dependent value map V at timestep t

```
Input: Arbitrary saliency map $S$ from Image $I$, human scanpath $\mathbf{X} = \{\mathbf{x_0}, \mathbf{x_1}, ..., \mathbf{x_t}\}$
for all possible fixations $\mathbf{x}$ do
    $C[\mathbf{x}] = \sum_{i\in\{0,1,2\}} \psi_i(\mathbf{x_{t-1}}, \mathbf{x_t}, \mathbf{x})$
    $E[\mathbf{x}] = \sum_{i=0}^{t} \phi_i \mathcal{N}(\mathbf{x}; \mathbf{x_i}, \mathbf{\Sigma})$
    $V[\mathbf{x}] = w_0 S[\mathbf{x}] + w_1 C[\mathbf{x}] + w_2 E[\mathbf{x}]$
end for
return $V$
```

To include the exploration map and calculate the resulting value map, the corresponding free parameters had to be estimated. Since we want to evaluate the prediction of the next n fixations, we need to find a metric suitable for comparing individual fixations. We chose the Normalized Scanpath Saliency metric [47], which is defined as NSS$(S, \mathbf{x_0}, \dots, \mathbf{x_T}) = 1/T \sum_{i=0}^{T} S_Z(\mathbf{x_i})$. where T is the total amount of fixations for the current image. Here S_Z is the saliency map standardized by its mean μ_S and its standard deviation σ_S, i.e. $S_Z = (S - \mu_S)/\sigma_S$. Thus, the metric can be viewed as an average of the standardized saliency scores at the corresponding fixation locations. For more details on the metric score see e.g. [35, 10, 31]. Since this method does not compare two continuous maps, but also considers the actual set of fixations in addition to the saliency map [38], the metric is also suitable for our case of one-step or n-step ahead prediction. In this case, we do not average over the entire gaze sequence, but optimize the value map of our model so that there is as much mass as possible at the location of the next fixation.

More specific, a random subset of 10000 real human fixations from the MIT1003 dataset, was selected and the parameters of our model were optimized so that the value map could predict the single subsequent fixation as well as possible and thus maximizes the NSS score. All optimizations were done on the MIT1003 dataset [32]. All selected fixations were between the third and eleventh gaze target in their respective sequence. This fixation interval was chosen so that at least two fixations had already been carried out by the human observers to be able to compute the cost map and because only about one percent of all fixation sequences contain more than ten fixations.

We estimated the exploration values ϕ_i, the covariance matrix $\mathbf{\Sigma}$ for the Exploration Map (Equation 3) and the

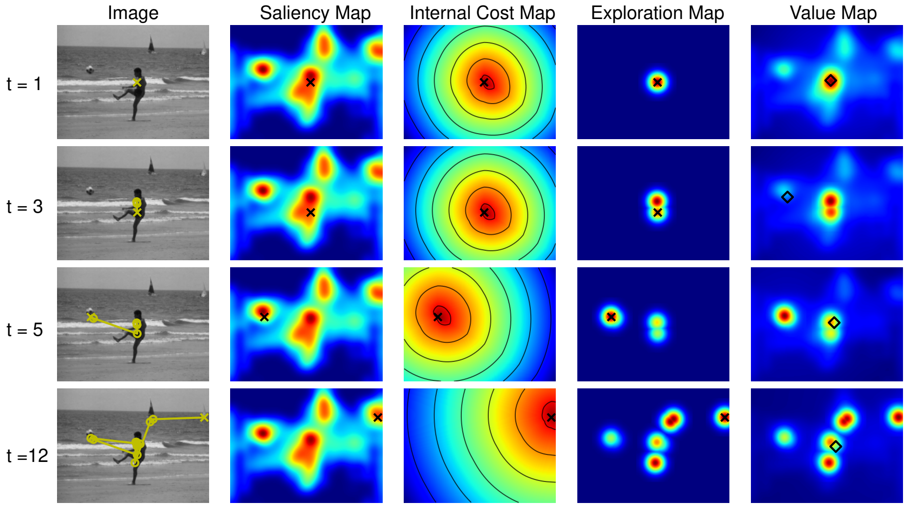

Figure 2: Example predictions of the next fixation. Each row shows the original image together with the respective preceding scanpath together with the current i-th fixation marked with a cross. The corresponding saliency, cost, and exploration maps as well as the final value map are shown from left to right. The predicted fixation is shown together with the ground truth next fixation of the human observer marked with a diamond.

weight parameters w_1,w_2 (Equation 5) through gradient based optimization. Note that the covariance matrix $\mathbf{\Sigma}$ was constrained to be a multiple of the identity matrix $\sigma^2\mathbf{I}$. We fixed the weight for the saliency map to one, so that the estimated parameters w_1,w_2 can be interpreted as quantifying the relative contributions of the costs for gaze switches and the history dependent reward relative to the saliency value. We used the Limited-memory BFGS-B algorithm [11] given the NSS score as an objective function to be maximized. The hyperparameters can be found in Section S1 in the Supplementary Material. In addition, to meet the computational cost of the multidimensional problem, the images were reduced by a factor of ten in both dimensions using bilinear interpolation.

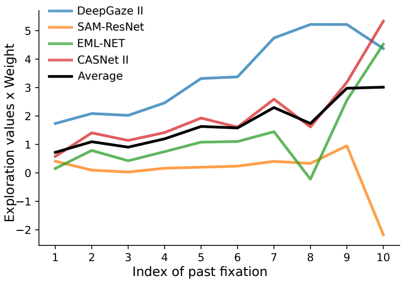

Figure 3: Estimated exploration values for four different saliency models and the averaged value (black). Note that the estimated ϕ_i are multiplied here by their associated weight w_2 to show the influence of the exploration map.

Computations were performed on a high performance computer cluster. All simulations were run on nodes with an Intel Xeon Processor E5-2680 v3 processor (2.5 GHz processor rate and 2.4 GB RAM). The results of the optimization for all parameters can be found in the Supplementary Material in Table S1. Additionally, the estimated exploration values, and thus the weighting of past fixations are shown for all four models over time in Figure 3. These intermediate results were used to derive general weights of past fixations independent of the particular model saliency. To this end, the estimated exploration values of the four models were averaged to be flexibly applied to arbitrary, new models. The resulting distribution is shown by the bold black curve in Figure 3.

The same experiment was repeated for all saliency models and image databases, except that the weight parameters were no longer co-estimated, i.e. the previously determined values were used. The results of this experiment can be found in Table S2 in the Supplementary Material. In addition, a new model was evaluated, which also belongs to the top evaluated saliency models on the MIT/Tuebingen benchmark - UNISAL [16]. This was to investigate the degree to which the optimized model parameters would generalize from the four baseline models to a new model.

5. Results

We evaluated our method on three frequently used benchmarks, the MIT1003 [32], the OSIE [63] and the Toronto dataset [9]. The MIT1003 and the OSIE dataset contain eye movements of 15 subjects during a three-second

Table 1: Evaluation results. AUC and NSS scores for the one-step and two-step ahead prediction of gaze targets based on sequential value maps compared to the respective saliency model's baseline.

(a) One-step ahead predictions

	MIT 1003		OSIE		Toronto	
	AUC	NSS	AUC	NSS	AUC	NSS
DeepGaze II	0.844	1.506	0.906	1.867	0.497	-0.031
Our extension	**0.874**	**1.856**	**0.908**	**2.569**	**0.632**	**0.823**
SAM-ResNet	0.864	2.222	0.905	3.088	0.477	-0.105
Our extension	**0.881**	**2.323**	**0.917**	**3.315**	**0.639**	**0.706**
EML-NET	0.864	2.255	0.902	3.050	0.490	-0.073
Our extension	**0.882**	**2.329**	**0.919**	**3.330**	**0.638**	**0.656**
CASNet II	0.860	1.993	0.898	2.587	0.515	-0.059
Our extension	**0.879**	**2.155**	**0.915**	**3.003**	**0.684**	**1.033**
UNISAL	0.889	2.612	0.890	2.755	0.542	0.020
Our extension	**0.898**	**2.653**	**0.909**	**3.159**	**0.626**	**0.451**

(b) Two-step ahead predictions

MIT 1003		OSIE		Toronto	
AUC	NSS	AUC	NSS	AUC	NSS
0.844	1.506	0.906	1.867	0.497	-0.031
0.8554	**1.725**	0.888	**1.899**	**0.602**	**0.624**
0.864	2.222	0.905	3.088	0.477	-0.105
0.862	**2.301**	0.894	2.862	**0.598**	**0.535**
0.864	2.255	0.902	3.050	0.490	-0.073
0.869	**2.332**	**0.903**	2.897	**0.601**	**0.508**
0.860	1.993	0.898	2.587	0.515	-0.059
0.865	**2.098**	0.894	2.475	**0.608**	**0.598**
0.889	2.612	0.890	2.755	0.542	0.020
0.888	**2.667**	**0.893**	**2.841**	**0.604**	**0.371**

free viewing task on 1003 and 700 natural indoor and outdoor scenes, respectively. The Toronto dataset consits of 20 subjects during a four-second free viewing task on 120 color images of outdoor and indoor scenes.

5.1. One-step ahead predictions

We evaluated the one-step ahead predictions of our model with the NSS metric on the three datasets. Additionally we used a second metric, the Area under Curve (AUC) (see [61, 35, 10, 31] for details) for a second evaluation measurement, which was not considered during optimization. AUC is also a well known hybrid measure for evaluating fixation prediction and saliency models [38], which can be understood as a binary classifier for whether pixels are fixated or not.

These two metrics can be used in evaluating the prediction of the next fixation, see [34]. Other saliency metrics, like KL-divergence, Correlation Coefficient or Information gain are distribution based, so they assume the ground truth map to be a density and not a single fixation. Therefore, they cannot be used to evaluate models predicting the next gaze target or any other per fixation evaluation. Example images with best and worst NSS scores are provided in Figure S2 and S3 of the Supplementary Material.

Regarding scanpath prediction metrics (like ScanMatch or MultiMatch), we follow the evidence and argumentation from [34], arguing that it makes more sense to evaluate the capability of a model to predict the next fixation, which is exactly what saliency metrics do. For evaluation, both metrics were calculated on all fixations of the three datasets above. For the first fixation, the model selects a target exclusively based on the saliency map as neither the internal cost nor an fixation history can contribute. To predict the second fixation, we assumed that the fixation prior to image onset was at the image's center. This is true for most experiments and this only influences the relative angle of the cost map.

The baseline saliency models were evaluated equivalently, but instead of using our dynamic value maps, the static history-independent maps were used. The results on the three different datasets with five different baseline models are shown in Table 1(a). We reached higher scores on all three datasets compared to all baseline models, even for the UNISAL model, which was not used in the estimation of the parameters of the exploration map. These results transferred in all cases to the AUC score, which had not been used in the optimization. Thus, subsequent fixations on the datasets are better predicted by our dynamic one-step ahead prediction maps compared to the static baseline saliency models. This provides evidence, that including the independently measured human cost function for carrying out eye movements improves predictions by saliency maps.

5.2. n-step ahead predictions

Although the free parameters of the model were optimized to maximize predictions of the single next fixation on the NSS score for the MIT1003 data set, we can test the performance of the n-step predictions. Table 1(b) and Table S3 in the Supplementary Material report the results of the two-step and three-step predictions respectively. These results show, that the present model performs better consistently on the NSS score for both the MIT1003 and Toronto datasets across the second and third fixation predictions for all tested saliency models. Performance on the AUC score starts deteriorating for the prediction of the third fixation on the MIT1003 dataset but not the Toronto dataset. By comparison, both AUC and NSS scores are weaker already for the predictions of the second fixations on the OSIE dataset for all tested saliency models.

5.3. Influence of past fixations

In addition to predicting the next fixation in a gaze sequence, our model allows quantifying and explaining the relative influence of past fixations. Since the exploration

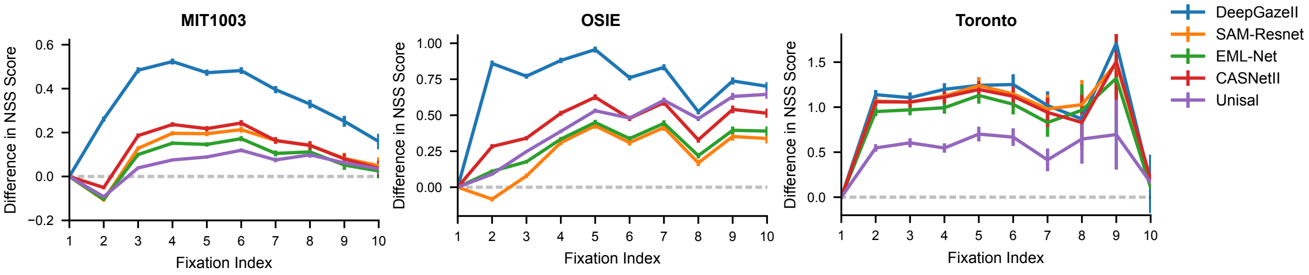

Figure 4: Differences in the NSS scores between our dynamic value maps and the underlying static saliency maps. Positive values indicate that our dynamic model predicted the subsequent fixation better than the baseline model. The errorbars indicate ± standard error of the mean.

values ϕ_i were not constrained, we are able to interpret them directly. Figure 3 shows the relative value of past fixations over time. Overall, the value of refixating an image location increases approximately linearly over time. This indicates that having visited location $\mathbf{x_i}$ i fixations ago during the same gaze sequence increases the probability of targeting the next fixation at location $\mathbf{x_i}$. This effect increases with increasing i, which means that fixation locations visited longer ago become more attractive for the observer.

For further analysis, we can quantify how well our predictions work for individual ordinal positions in the gaze sequence. For this, we selected all predictions by their ordinal position and averaged the NSS scores grouped by their fixation index. The progression of the goodness of the predictions can be seen in Figure S1 for all five models on all three datasets in comparison to the underlying baseline saliency models. The differences in NSS scores can be seen in Figure 4. These results demonstrate, that the prediction accuracy is higher throughout the entire sequence up to the tenth gaze target, which was the last considered for almost all combinations of saliency models and image data sets. This further supports the usefulness and validity of the current approach.

6. Discussion

In this paper, we introduced a computational model utilizing arbitrary saliency maps for computing sequential value maps to predict the next gaze target in human fixation sequences [34]. We conceptualized gaze sequences as sequential decision making within the framework of statistical decision theory, similar to previous approaches [41, 29, 24]. Given a saliency map of arbitrary origin and a sequence of previous gaze targets on an image, the model generates predictions of the next most likely fixation. The intrinsic preferences for gaze shifts used in the algorithm were recently estimated through a preference elicitation experiment independently of image content [54] and the spatial and temporal parameters of the influence of fixation history were inferred based on the MIT1003 data set. Finally, the relative contributions of the three value maps were optimized on the same data set to maximize prediction of the next fixation. The algorithm can be applied to arbitrary saliency models and is available upon request from the authors.

The results demonstrate that the three components of the intrinsic costs for human gaze shifts [54] are sufficient to improve predictions of subsequent gaze targets obtained from a saliency model. These results are evidence that the common simplifying assumption that human scan paths are independent of behavioral preferences in gaze selection does not hold. Instead, the analysis of the distribution of preferred angles demonstrates, that image content and preferences in gaze shifts interact in non-trivial ways, a fact that has previously been demonstrated empirically [19]. Although some previous approaches in scanpath modeling have acknowledged or implemented statistics of human gaze shifts [5, 58, 53, 39, 64], these were not measured independently of image content. The problem this gives rise to, is that the empirical statistics e.g. of saccade lengths measured in free viewing is the result of the preferences for gaze shifts and the distribution of image features. Thus, predictions of the next fixation need to be generated by taking the actual human costs of gaze shifts into account instead of the empirical distributions of gaze obtained from the databases, because the latter are the result of the interaction between image features and the costs for gaze shifts.

Acknowledgements

Calculations for this research were conducted on the Lichtenberg high performance computer of the TU Darmstadt. FK and CR acknowledge the support by the German Research Foundation (DFG, grant: RO 4337/3-1; 409170253). This research was supported by the Hessian research priority program LOEWE within the project WhiteBox.

References

[1] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps. In *Advances in Neural Information Processing Systems*, pages 9505–9515, 2018.

[2] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E. O'Connor. PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks. In Laura Leal-Taixé and Stefan Roth, editors, *Computer Vision – ECCV 2018 Workshops*, volume 11133, pages 406–422. Springer International Publishing, Cham, 2019.

[3] Marc Assens, Kevin McGuinness, Xavier Giro-i Nieto, and Noel E. O'Connor. SaltiNet: Scan-path Prediction on 360 Degree Images using Saliency Volumes. *arXiv:1707.03123 [cs]*, Aug. 2017. arXiv: 1707.03123.

[4] Giuseppe Boccignone, Vittorio Cuculo, and Alessandro D'Amelio. *How to Look Next? A Data-Driven Approach for Scanpath Prediction*, pages 131–145. Springer, 08 2020.

[5] 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, Jan. 2004.

[6] Ali Borji. Saliency Prediction in the Deep Learning Era: Successes, Limitations, and Future Challenges. *arXiv:1810.03716 [cs]*, May 2019. arXiv: 1810.03716.

[7] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2013.

[8] Ali Borji, Dicky N Sihite, and Laurent Itti. Probabilistic learning of task-specific visual attention. In *2012 IEEE Conference on computer vision and pattern recognition*, pages 470–477. IEEE, 2012.

[9] Neil Bruce and John Tsotsos. Attention based on information maximization. *Journal of Vision*, 7(9):950–950, Jun 2007.

[10] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Fredo Durand. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(3):740–757, Mar. 2019.

[11] Richard H. Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

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

[13] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, Oct. 2018. arXiv: 1611.09571.

[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*, 27(10):5142–5154, 2018.

[15] 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.

[16] Richard Droste, Jianbo Jiao, and J. Alison Noble. Unified Image and Video Saliency Modeling. In *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020.

[17] Shaojing Fan, Zhiqi Shen, Ming Jiang, Bryan L. Koenig, Juan Xu, Mohan S. Kankanhalli, and Qi Zhao. Emotional attention: A study of image sentiment and visual attention. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2018.

[18] John M Findlay, John M Findlay, Iain D Gilchrist, et al. *Active vision: The psychology of looking and seeing*. Oxford University Press, 2003.

[19] Tom Foulsham, Alan Kingstone, and Geoffrey Underwood. Turning the world around: Patterns in saccade direction vary with picture orientation. *Vision research*, 48(17):1777–1790, 2008.

[20] Mary Hayhoe and Dana Ballard. Eye movements in natural behavior. *Trends in cognitive sciences*, 9(4):188–194, 2005.

[21] John M. Henderson, Taylor R. Hayes, Candace E. Peacock, and Gwendolyn Rehrig. Meaning and attentional guidance in scenes: A review of the meaning map approach. *Vision (Basel, Switzerland)*, 3(2):19, May 2019. 31735820[pmid].

[22] John M Henderson and Andrew Hollingworth. Eye movements during scene viewing: An overview. In *Eye guidance in reading and scene perception*, pages 269–293. Elsevier, 1998.

[23] David Hoppe and Constantin A Rothkopf. Learning rational temporal eye movement strategies. *Proceedings of the National Academy of Sciences*, 113(29):8332–8337, 2016.

[24] David Hoppe and Constantin A Rothkopf. Multi-step planning of eye movements in visual search. *Scientific reports*, 9(1):1–12, 2019.

[25] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the Semantic Gap in Saliency Prediction by Adapting Deep Neural Networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*, pages 262–270, Santiago, Chile, Dec. 2015. IEEE.

[26] 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*, 20(11):1254–1259, Nov. 1998.

[27] 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.

[28] Sen Jia and Neil D.B. Bruce. Eml-net: An expandable multi-layer network for saliency prediction. *Image and Vision Computing*, 95:103887, 2020.

[29] 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, June 2016.

[30] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in Context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1072–1080, Boston, MA, USA, June 2015. IEEE.

[31] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[32] T. Judd, K. Ehinger, F. Durand, and A. Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*, pages 2106–2113, 2009.

[33] Srinivas S. S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu. DeepFix: A Fully Convolutional Neural Network for predicting Human Eye Fixations. *arXiv:1510.02927 [cs]*, Oct. 2015. arXiv: 1510.02927.

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

[35] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In Vittorio Ferrari, Martial Hebert, Cristian Sminchisescu, and Yair Weiss, editors, *Computer Vision – ECCV 2018*, Lecture Notes in Computer Science, pages 798–814. Springer International Publishing, 2018.

[36] Matthias Kümmerer, Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT/Tübingen Saliency Benchmark.

[37] M. Kümmerer, T. S. A. Wallis, L. A. Gatys, and M. Bethge. Understanding Low- and High-Level Contributions to Fixation Prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4799–4808, 2017.

[38] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, Mar. 2013.

[39] Olivier Le Meur, Antoine Coutrot, Adrien Le Roch, Andrea Helo, Pia Rämä, and Zhi Liu. Age-dependent saccadic models for predicting eye movements. In *2017 IEEE International Conference on Image Processing (ICIP)*, pages 3740–3744. IEEE, 2017.

[40] Huiying Liu, Dong Xu, Qingming Huang, Wen Li, Min Xu, and Stephen Lin. Semantically-Based Human Scanpath Estimation with HMMs. In *2013 IEEE International Conference on Computer Vision*, pages 3232–3239, Sydney, Australia, Dec. 2013. IEEE.

[41] Stefan Mathe and Cristian Sminchisescu. Action from still image dataset and inverse optimal control to learn task specific visual scanpaths. In *Advances in neural information processing systems*, pages 1923–1931, 2013.

[42] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 2204–2212, Cambridge, MA, USA, 2014. MIT Press.

[43] Jiri Najemnik and Wilson S Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005.

[44] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision research*, 45(2):205–231, 2005.

[45] Junting Pan, Elisa Sayrol, Xavier Giro-I-Nieto, Kevin McGuinness, and Noel E. OConnor. Shallow and Deep Convolutional Networks for Saliency Prediction. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 598–606, Las Vegas, NV, USA, June 2016. IEEE.

[46] Marek A Pedziwiatr, Thomas SA Wallis, Matthias Kümmerer, and Christoph Teufel. Meaning maps and deep neural networks are insensitive to meaning when predicting human fixations. *Journal of Vision*, 19(10):253c–253c, 2019. Publisher: The Association for Research in Vision and Ophthalmology.

[47] Robert J. Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision Research*, 45(18):2397–2416, 2005.

[48] Matthew F Peterson and Miguel P Eckstein. Looking just below the eyes is optimal across face recognition tasks. *Proceedings of the National Academy of Sciences*, 109(48):E3314–E3323, 2012.

[49] Karen Simonyan and Andrew Zisserman. Very Deep Convolutional Networks for Large-Scale Image Recognition. *arXiv:1409.1556 [cs]*, Apr. 2015. arXiv: 1409.1556.

[50] Sun, Xiaoshuai, Yao, Hongxun, and Ji, Rongrong. 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, Providence, RI, June 2012. IEEE.

[51] Richard S Sutton. Integrated architectures for learning, planning, and reacting based on approximating dynamic programming. In *Machine learning proceedings 1990*, pages 216–224. Elsevier, 1990.

[52] Benjamin W Tatler, Mary M Hayhoe, Michael F Land, and Dana H Ballard. Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5–5, 2011.

[53] 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.

[54] Tobias Thomas, David Hoppe, and Constantin A. Rothkopf. The neuroeconomics of individual differences in saccadic decisions. *bioRxiv*, 2022.

[55] Kenneth E Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

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

[57] Stefan Treue. Visual attention: the where, what, how and why of saliency. *Current opinion in neurobiology*, 13(4):428–432, 2003.

[58] 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.

[59] Wenguan Wang, Jianbing Shen, Fang Guo, Ming-Ming Cheng, and Ali Borji. Revisiting Video Saliency: A Large-Scale Benchmark and a New Model. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4894–4903, Salt Lake City, UT, June 2018. IEEE.

[60] Sean Welleck, Jialin Mao, Kyunghyun Cho, and Zheng Zhang. Saliency-based sequential image attention with multiset prediction. *Advances in Neural Information Processing Systems*, 2017-December:5174–5184, Jan. 2017. 31st Annual Conference on Neural Information Processing Systems, NIPS 2017 ; Conference date: 04-12-2017 Through 09-12-2017.

[61] Niklas Wilming, Torsten Betz, Tim C. Kietzmann, and Peter König. Measures and limits of models of fixation selection. *PLOS ONE*, 6(9):1–19, 09 2011.

[62] 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.

[63] Juan Xu, Ming Jiang, Shuo Wang, Mohan S. Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):28–28, 01 2014.

[64] 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.