Computers & Graphics 122 (2024) 103983

Contents lists available at ScienceDirect

Computers & Graphics

journal homepage: www.elsevier.com/locate/cag

Special Section on CEIG 2024

tSPM-Net: A probabilistic spatio-temporal approach for scanpath prediction

Daniel Martin *, Diego Gutierrez, Belen Masia

Universidad de Zaragoza, I3A, Spain

ARTICLE INFO

Dataset link: https://github.com/DaniMS-ZGZ/tSPM-Probabilistic-Scanpaths

Keywords:
Scanpaths
Gaze prediction
Saliency
Visual attention

ABSTRACT

Predicting the path followed by the viewer's eyes when observing an image (a scanpath) is a challenging problem, particularly due to the inter- and intra-observer variability and the spatio-temporal dependencies of the visual attention process. Most existing approaches have focused on progressively optimizing the prediction of a gaze point given the previous ones. In this work we propose instead a probabilistic approach, which we call tSPM-Net. We build our method to account for observers' variability by resorting to Bayesian deep learning and a probabilistic approach. Besides, we optimize our model to jointly consider both spatial and temporal dimensions of scanpaths using a novel spatio-temporal loss function based on a combination of Kullback–Leibler divergence and dynamic time warping. Our tSPM-Net yields results that outperform those of current state-of-the-art approaches, and are closer to the human baseline, suggesting that our model is able to generate scanpaths whose behavior closely resembles those of the real ones.

1. Introduction

Understanding and predicting human visual attention has been an active research area in computer vision for decades [1–7]. Traditionally, those works have resorted to predicting saliency as a measure of where attention is going to be directed to. Such prediction usually builds upon spatial, bottom-up analyses of the image, leading to the determination of salient areas represented as *saliency maps*, which are topographical representations by a scalar quantity of the conspicuity at every location in the visual field [1–3].

Although this representation may suffice for certain applications, saliency maps fail to capture the spatio-temporal dependencies of visual behavior [8,9]. This information is relevant in a varied number of scenarios, ranging from marketing and product placement, webpage design or scene design, to analysis of visual pathologies or realistic eye motion simulation (e.g., for avatar animation). To take into account such temporal information, a number of works have tackled the problem of *scanpath* prediction (see [10] for a review). We formally define a scanpath as a sequence of consecutive eye movements (i.e., fixations and saccades) through time and space [11]. As such, for a conventional 2D image I, a scanpath is a sequence of gaze points or image coordinates, $\{s_0, \dots, s_t\}$, $s_i \in \mathbb{R}^2$.

Previous works either resort to heuristics and hand-crafted features [12,13], or to data-driven methods [6,14,15] to perform scanpath prediction. However, most of the existing methods are designed to optimize the prediction of a *single* fixation point s_t given the previous points $\{s_0, \dots, s_{t-1}\}$ [6]; thus the scanpath is progressively built by concatenating successive single-point solutions. While this strategy can useful for several applications such as foveated rendering [16,17], it may lead to increasing deviations from actual human viewing behavior and scanpath plausibility [18], and falls short to model the large inter- and intra-observer variability [19,20].

In this paper we present tSPM-Net, a method to predict *full, plausible* scanpaths given an input image I (see Fig. 1). We leverage the fact that, despite the inter- and intra-observer variability, common patterns and behaviors do emerge when humans observe certain content [21,22]. Thus, we focus on modeling a *distribution* of scanpaths within this common behavioral space. This distribution can then be sampled to generate individual scanpaths, where every sampled scanpath represents a different, yet plausible, potential observer that maintains human behavioral patterns. To handle spatio-temporal information, and as is common in related problems, we rely on convolutional long short-term memory networks (ConvLSTM) [23]: Their recurrent architecture is well suited to capture the temporal dependency of each predicted point in a scanpath, while their convolutional nature has proven to be successful handling problems with both spatial and temporal dependencies.

Different from previous works, to obtain the distribution of plausible scanpaths we explicitly incorporate the inherent uncertainty of the problem into our model, and build our ConvLSTM module upon Bayesian deep learning [24], which has already been shown to be effective in related problems like head trajectory prediction in omnidirectional images [25]. Consistent with this explicit modeling of uncertainty, we do not learn gaze points for each temporal instant, but rather probability distributions, representing scanpaths as a sequence of what we define *time-evolving scanpath probabilistic maps* (tSPMs, see

* Corresponding author.
E-mail addresses: danims@unizar.es (D. Martin), diegog@unizar.es (D. Gutierrez), bmasia@unizar.es (B. Masia).

https://doi.org/10.1016/j.cag.2024.103983
Received 10 April 2024; Received in revised form 7 June 2024; Accepted 18 June 2024
Available online 22 June 2024

Fig. 1. Overview of this work. We present a convolutional recurrent approach to scanpath prediction. Our model relies on Bayesian deep learning, a spatialized representation of scanpaths, and a novel loss function. We propose a spatio-temporal loss function based on a combination of the Kullback–Leibler divergence and dynamic time warping and predict time-evolving scanpath probabilistic maps (tSPM), well suited to the stochastic nature of human scanpaths. Our generated scanpaths maintain the spatial and temporal characteristics of human scanpaths and outperform previous state-of-the-art approaches. An overview of the model is shown in Fig. 2.

Fig. 1). Some previous approaches have already shown the benefits of using such a spatialized representation [26]. Additionally, we introduce a novel loss function for joint spatio-temporal optimization, which combines the benefits of the Kullback–Leibler divergence and dynamic time warping (DTW). The loss is computed over the full set of tSPMs that form a scanpath.

Our resulting trained model is able to generate a distribution of plausible scanpaths for a given input image, where each scanpath mimics the visual behavior of a different human observer. We have validated our model both qualitatively and quantitatively, including an exhaustive set of existing metrics accounting for different scanpath characteristics [18]. Our model outperforms the state of the art overall, being closer to the human baseline. We also validate our design choices (Bayesian deep learning approach and loss function) through ablation studies. We will make our code and model publicly available to encourage future research.

2. Related work

2.1. Saliency prediction

First approaches towards modeling human attention were based on saliency, as a measure of how much each part of a scene attracts human attention. The seminal work by Itti et al. [1] established the basis of visual attention prediction in images, by extracting hand-crafted features to generate a saliency map. This work inspired many posterior approaches (e.g., [27,28]) which were also based on the computation of *conspicuity maps* for different visual features (such as color, intensity, orientation of edges, or faces), which were then combined into a final saliency map. Other approaches included multiple semantic segmentation and surroundness analysis [29], or known human priors such as center bias or horizon line detectors [30], to improve saliency prediction.

With the proliferation of deep learning techniques and the appearance of public datasets [20,31,32], data-driven methods emerged, yielding impressive results. These methods were mostly based on convolutional neural networks (CNN) that extract latent features from which to infer saliency [33–36]. Other approaches also leveraged the advances of generative networks [37,38] and recurrent neural networks [39,40]. None of these works, however, take into account the dynamic nature of gaze behavior, not being able to model the temporal dimension of human attention.

2.2. Scanpath prediction

Scanpath models usually aim to progressively build a scanpath by concatenating single-point predictions, which may be partially based on the previous points of the path. Ellis and Smith [21] presented a framework based on Markov stochastic processes. Later, other works proposed approaches that included known human biases, (such as the center bias, or human oculomotor constraints) [12,13,41,42]. Data-driven methods provide faster and more precise approaches, for instance using existing saliency prediction methods as a proxy to scanpath prediction, by means of winner-takes-all and inhibition-of-return strategies [6], by sampling heuristics [15,43], or simply leveraging deep features from neural networks [34].

Scanpath prediction methods can be roughly categorized into (i) biologically inspired, (ii) statistically inspired, (iii) cognitively inspired, and (iv) engineered models [10]. Biologically inspired models take into account the importance of low-level features [1,44,45], visual working memory [46], attention and inhibition-of-return [47], or neuropsychology [48]. Statistically inspired models try to mimic certain statistical properties of human scanpaths [19,38,49–51]. Cognitively inspired models assume that other cognitive processes besides low-level features can drive observers' attention, and therefore implement different human mechanisms such as low-level saliency, semantic and spatial effects [41] or region-of-interest and inhibition-of-return [6]. Finally, engineered models exploit the ability of data-driven techniques to fit to given data [7,14,43,52–54].

With this surge of data-driven approaches, and motivated by the temporal dependencies that human viewing behavior presents, some works have resorted to recurrent neural networks (RNN), which are capable of encoding previous information, and leveraging it to formulate a prediction [17]. However, scanpath prediction requires handling temporal and spatial information. To account for both, some recent approaches have built their models following ConvLSTM strategies [6, 55–57], where convolutional operators handle spatial features while LSTM architectures enable learning temporal information.

However, the aforementioned works either resort to saliency (which is not usually available as ground truth) [1,6,15,19,54], or are trained to optimize *single-point* predictions, by means of direct losses such as MSE [52] or BCE [6], and thus do not concern themselves with the plausibility of the scanpath as a whole. Closer to our work, Kummerer et al. [26] recently presented DeepGaze III, a DL-based approach for scanpath prediction. However, they optimize each predicted point based on a conditional probability on the previous history, rather than optimizing the whole scanpath with respect to the whole distribution of plausible scanpaths. Recently, the work of Martin et al. [9] presented a scanpath generation method for 360° content, where the model was optimized by means of a dynamic time warping loss function on the whole distribution of ground-truth scanpaths, rather than on a single-point solution, and was hence able to learn and mimic latent behaviors in its predictions.

In this work, and endorsed by previous literature, we resort to convolutional recurrent networks, but overcome the limitations of single-point prediction approaches by combining a novel loss function that combines dynamic time warping and Kullback–Leibler divergence, and a probabilistic approach. The loss function enables focusing on both the temporal and spatial aspects of the whole scanpaths, and optimizes our model over the whole distribution of real scanpaths, while our probabilistic approach accounts for the inherent human variability.

3. Our model

Our model, coined as tSPM-Net, performs probabilistic scanpath prediction given a single 2D image as input. The model, based on

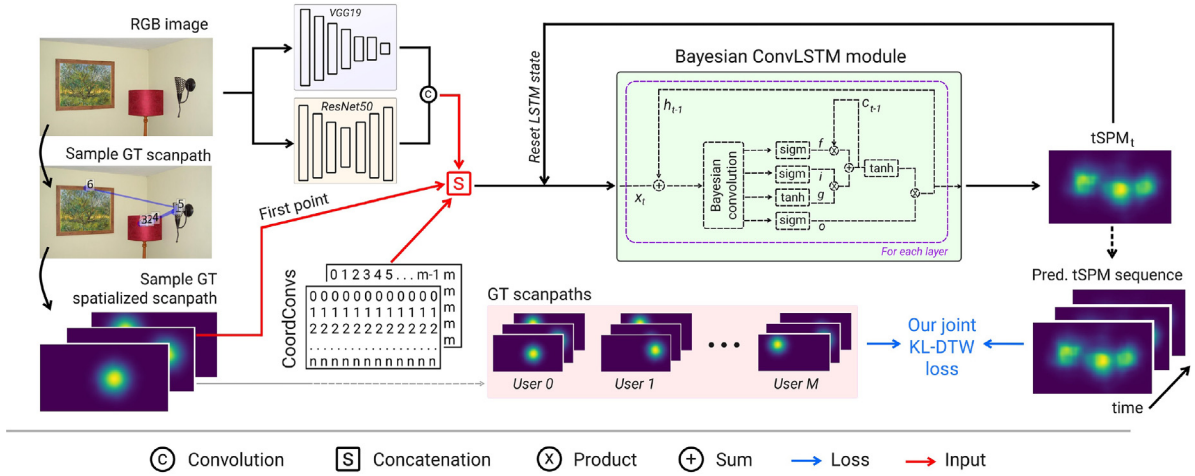

Fig. 2. Overview of our model. We represent our scanpaths as Gaussian maps to enhance spatial learning (see Section 3.1). We leverage a pretrained VGG19 and a pretrained semantic segmentator based on ResNet50 to extract meaningful visual features from the image. We then combine those features and the spatialized scanpath with a CoordConv layer, to facilitate the learning of spatial features. This input is fed to a multi-layer convolutional LSTM module that iterates over the scanpath and predicts the next fixation, until a scanpath of the target length is generated. We compute a loss based on dynamic time warping and Kullback–Leibler divergence (see Section 3.3) to optimize our model in a joint spatio-temporal fashion. The outcome of our model is a tSPM sequence (see also Fig. 4).

Fig. 3. An example of our spatialized scanpath representation. The top left image shows a RGB image with a sample ground-truth scanpath overlaid. For that sample ground-truth scanpath, we transform each fixation point into a Gaussian map centered at that particular point (see Section 3.1).

recurrent neural networks, is described in detail in this section: we introduce the representation we employ for the scanpaths (Section 3.1), a novel loss function that is able to optimize our scanpaths in a joint spatio-temporal fashion (Section 3.3), and our model architecture in depth (Section 3.4).

3.1. Scanpath representation

Scanpaths have been traditionally defined as a sequence $s = \{s_0, s_1, ..., s_{N-1}\}$ of gaze points,[1] where $s_i = (x_i, y_i)$, and (x_i, y_i) are the image coordinates of that particular gaze point.

However, there is no single ground-truth scanpath for an image [1]; scanpaths exhibit both inter- and intra-observer variability. Despite sharing some behavioral patterns [12], not all observers will explore the image in exactly the same way, thus resulting in many different, yet plausible scanpaths for the same image. Indeed, an observer watching the same image twice may follow slightly different scanpaths, and, even if asked to follow a certain path, there is a ballistic or noisy component in ocular movements (e.g., saccades or post-saccadic oscillations [58]), that results in different gaze points. As a result, scanpaths are non-deterministic.

Additionally, the aforementioned representation usually falls short for problems where it is necessary to establish a relationship between those coordinates and the position of features within an image. Indeed, convolutional networks are trained to be shift-invariant [59], and forcing them to explicitly learn the relation between a sequence of gaze point coordinates and the actual positions of image features is challenging and hinders the training process.

Given this variability, and in order to facilitate the spatial learning of the network, instead of representing each gaze point with its coordinates, $s_i = (x_i, y_i)$, a more adequate representation for gaze at a time instant is a Gaussian distribution g_i^s centered in (x_i, y_i), and defined over the whole image. In each distribution g_i, there is thus a value $g_i(x, y)$ per pixel (x, y), which represents the probability of a gaze point falling at pixel (x, y) at time step i. A scanpath s is therefore represented as a sequence $g^s = \{g_0^s, g_1^s, ..., g_{N-1}^s\}$ of Gaussian maps g_i^s; we term a scanpath represented in this way a *spatialized* scanpath. This representation facilitates spatial learning by providing a direct correlation between a scanpath and its corresponding image. We depict this representation in Fig. 3.

3.2. Overview of the model

Our model (see Fig. 2) is based on ConvLSTMs [23], a type of recurrent cell. ConvLSTMs maintain the recurrent nature of traditional LSTMs, processing data in a sequential manner, learning the temporal features of the data. Additionally, they are provided with convolutional operators that handle visual information and facilitate learning spatial features in the input sequence. Further, we resort to a Bayesian approach when modeling the ConvLSTM module, in order to better incorporate the uncertainty driven by inter- and intra-observer variability.

As explained in Section 3.1, we model gaze behavior using probability maps. Thus, the output of our ConvLSTM module is not a point, but rather a probability map (see Fig. 2). Our whole model predicts, given an input image, a sequence of time-evolving scanpath probabilistic maps (tSPM) (see Fig. 4). Each tSPM represents the probabilities of the next gaze fixation point falling on each pixel of the image at a certain time instant. We build our model leveraging pretrained neural networks on image classification tasks to facilitate feature extraction, and CoordConv layers to improve learning of spatial features.

[1] In our case, and following common practice [6,14,34], the points in a scanpath correspond to fixation points.

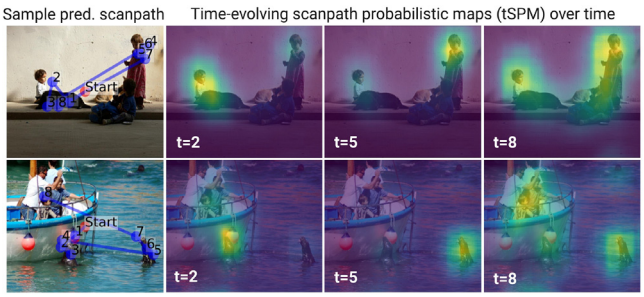

Fig. 4. We show, for two given test images, the evolution of our predicted time-evolving scanpath probabilistic maps (tSPM, see Section 3.2) over time. Our Bayesian approach enables a probabilistic weight selection for the next fixation, and allows our model to be stochastic, while the ConvLSTM module allows taking into account the previous fixations (see Section 3.4).

3.3. Loss function

As stated in Section 2, previous deep learning-based approaches have designed their models either to predict a saliency map (which neglects the temporal dimension of gaze) or to optimize the prediction of a gaze point at each time step, with element-wise loss functions, such as mean squared error (MSE) or binary cross-entropy (BCE), that penalize the prediction for *each point* in isolation (which neglects the inter- and intra-variability of viewing behaviors).

In contrast, we propose a novel loss function based on the Kullback–Leibler divergence and dynamic time warping, computed over the whole scanpath. The former allows our model to account for the spatial relations between gaze points, while the latter ensures a realistic and plausible temporal behavior of the predicted scanpaths.

Kullback–Leibler Divergence (KL-Div) KL-Div is a well-known measure of how different a probability distribution is from another one, and is one of the most commonly used losses in saliency prediction problems [55,60–62]. The Kullback–Leibler divergence (D_{KL}) is defined as:

$$D_{KL}(P \parallel Q) = \sum_j P(j) ln \frac{P(j)}{Q(j)}, \tag{1}$$

where P and Q are the probability distributions to be compared, and j refers to each point of the distribution. In our particular case, each gaze point is represented in a spatialized manner, hence KL-Div is able to give a quantitative measurement on how different two gaze points are based on their probability maps.

Dynamic Time Warping (DTW) DTW is a measure of similarity between two time series that may differ in length or speed [63]. The DTW algorithm attempts to find the optimal match between the points of two temporal sequences, r and s, by matching each point in one of them with at least one point in the other, without forcing a one-to-one correspondence between both sequences. The optimal match is found by minimizing a cost function: a distance matrix Δ stores the cost (Euclidean distance) for each possible pair of points, and the optimization searches for the matching (alignment) between r and s such that the total cost is minimized. This can be written as:

$$DTW(r, s) = \min_A \langle A, \Delta(r, s) \rangle, \tag{2}$$

where A is a binary alignment matrix between two time series r and s, $\Delta(r, s) = [\delta(r_i, s_j)]_{i,j}$ is a matrix containing the distances $\delta(\cdot, \cdot)$ between each pair of points in r and s, and $\langle \cdot, \cdot \rangle$ denotes the inner product between both matrices. Since the minimum function is not differentiable, a soft version has been proposed [64]:

$$DTW^\gamma(r, s) = \min_A{}^\gamma \langle A, \Delta(r, s) \rangle, \quad \gamma > 0 \tag{3}$$

The soft-min function $\min^\gamma$ is defined as:

$$\min{}^\gamma(a_1, \dots, a_N) = -\gamma \, log \sum_{i=1}^{N} exp\left(-\frac{a_i}{\gamma}\right), \tag{4}$$

with the γ parameter adjusting the similarity between the soft version and the original DTW algorithm, both being the same when $\gamma = 0$. Eq. (3) has been used successfully as a loss term in related contexts, such as scanpath generation for virtual reality [9] or weakly supervised action alignment and segmentation in videos [65].

Our Joint KL-DTW Loss While KL-Div accounts for the spatial similarity of two distributions, and DTW focuses on the temporal dimension, none of them suffices on its own in our particular case. We therefore propose a novel loss function, based on a combination of both KL-Div and DTW, and defined as follows:

$$\mathcal{L}_{KL-DTW}(r') = \frac{\sum_{s=1}^{S} DTW^\gamma(r', s)}{|S|}, \tag{5}$$

where r' is a predicted sequence of tSPMs (see Section 3.2), and s is a ground-truth scanpath from the set of ground-truth ones S for a given image I. DTW^γ is computed as given by Eq. (3). However, we modify the computation of the distance matrix $\Delta(r', s) = [\delta(r'_i, s_j)]_{i,j}$ such that, instead of δ being an Euclidean distance, we have:

$$\delta(r'_i, s_j) = D_{KL}(r'_i \parallel g^s_j), \tag{6}$$

where r'_i is the ith predicted tSPM, g^s_j is the spatialized representation of point s_j as described in Section 3.1, and D_{KL} is the Kullback–Leibler divergence (Eq. (1)).

This formulation allows our model to be optimized to find an alignment that minimizes both the *spatial* and the *temporal* differences between each predicted scanpath and the ground-truth ones, therefore predicting scanpaths that follow a similar distribution as the ground truth. To our knowledge, we are the first to propose such a combination of metrics.

Bias Regularization Term Human gaze data in 2D images is known to be strongly biased towards the center of the images [12]. Although inherent to human nature, such bias hinders the learning process of the network, which can easily overfit to that behavior. Based on this, we include a regularization loss term that penalizes scanpaths whose points tend to stay in the center of the image for a long time, hence eliciting a more exploratory behavior that better reflects ground truth data. Our regularization term is included in the pairwise cost computations $\delta(r'_i, s_j)$ for the distance matrix Δ, modifying Eq. (6) as follows:

$$\delta(r'_i, s_j) = D_{KL}(r'_i \parallel g^s_j) + \lambda_{CB} * \mathcal{L}_{Reg}(r'_i) \tag{7}$$

$$\mathcal{L}_{Reg}(r'_i) = \frac{1}{D_{KL}(r'_i \parallel g_c)}, \tag{8}$$

where g_c is a Gaussian map representing the aforementioned center bias, computed following the representation introduced in Section 3.1 for a point c in the center of the image, and λ_{CB} is a regularization weight to mimic human center bias behavior, which penalizes scanpaths that tend to stay in the center of the image for a long time. In order to set the relative weight of λ_{CB}, we analyzed the datasets used throughout this work (see Section 3.5), and found that this center bias behavior diminishes over time, with fixations being more widely spread over the image in later time instants. We measured the standard deviation of fixation positions in the ground-truth data, and found them to increase logarithmically over time ($R^2 = 0.855$); we hence increase λ_{CB} in the same way (see Fig. 5).

3.4. Model architecture

Our model features a recurrent neural network (RNN), which is able to extract and maintain temporal latent information from the scanpaths it is trained with. As introduced in Section 3.2, we choose a convolutional long short-term memory (ConvLSTM) network [23], which is an adaptation of classic LSTMs to work with 2D data, such as images. This type of network has proven to be effective in many different problems, such as weather forecasting [23], video saliency detection [66], or medical image segmentation [67]. ConvLSTMs behave

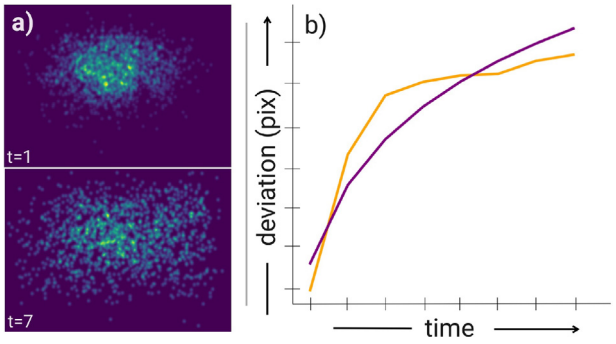

Fig. 5. We have analyzed that bias in the datasets used in this work [12] and found that this behavior diminishes over time. (a) Distribution of scanpath fixations in two different time instants for the whole training dataset. (b) We have computed the standard deviations of fixations (y-axis) over time (x-axis) for all the ground-truth scanpaths (in orange) and found they increase in a logarithmic fashion (the fitted curve is shown in purple).

in a similar way to traditional LSTMs, working over four different gates; however, since they handle spatial data, they conduct convolutional operations rather than lineal ones. The ConvLSTM used in this work is inspired by the work of Xingjian et al. [23], and follows a similar formulation as the one proposed by Blundell et al. [68], so we thus refer the reader to those works for an in-depth explanation of the used ConvLSTM architecture.

In our particular case, there is a degree of stochasticity driven by the inter- and intra-observer variability. As a result, instead of predicting the next point in a deterministic manner, we predict a probability map (i.e., the previously introduced tSPM). Further, we combine the aforementioned ConvLSTM with Bayesian deep learning [24,69]. Unlike traditional deep learning (DL), where the weights of the network are deterministic, in Bayesian DL the weights of a particular layer are sampled from a probability distribution, and thus the network itself can account for the inherent uncertainty of the data [24]. During network training, those weight distributions are also optimized. In our case, we substitute each convolutional operation of the ConvLSTM with a Bayesian 2D convolution.

Instead of feeding our Bayesian ConvLSTM with the raw images, we preprocess them to facilitate the learning process and enhance our model's performance. For this, we first extract the main image features with a pretrained VGG19 [70,71], and a semantic segmentation mask with a pretrained ResNet50 [72], which has already been shown to improve the overall network ability to focus on the relevant features [6,73]. We show some examples of this intermediate step in the supplementary material. We then convolve both of them together to obtain a final, comprehensive, single-channel image feature representation. At each time step, this feature representation is fed to the ConvLSTM alongside with (i) the corresponding Gaussian map representing a fixation, and (ii) a CoordConv layer [59] (see Fig. 2). CoordConv layers have proven to ease spatial learning and facilitate network convergence. With this input, our network is able to predict a tSPM for each time step. Then, we probabilistically sample such a map to obtain the next fixation. We repeat this procedure until the whole scanpath is predicted. Our model is composed of approximately 185M parameters and takes less than a second to predict a new scanpath.

3.5. Datasets and training details

As stated in this section, our model predicts a tSPM (time-evolving scanpath probabilistic map) at each time step. Then, to choose the actual pixel where the gaze point will fall, we follow a probabilistic weighted sampling strategy, that again accounts for the stochastic nature of human visual exploration. We first discard all pixels with a probability lower than a threshold $th = 0.7$ (see Section 4 for additional details and evaluation on this), and then sample the next point based on the predicted map's probabilities. Once a point $s_i = (x_i, y_i)$ has been sampled, a Gaussian map centered in (x_i, y_i) is again computed, and fed to the network for its posterior predictions, until the whole scanpath is predicted.

Following previous work [6], we train our model over the OSIE dataset [74], which contains 700 different images with their corresponding gaze information for a total of fifteen observers, yielding a total of approximately 10,500 scanpaths. We again follow previous work [6] and discard all the scanpaths with $N < 4$, and generate scanpaths of length $N = 8$, which is the mean length of our ground-truth data.

For the rest of the scanpaths, in order to train our model, we preprocess each to follow the representation introduced in Section 3.1. To validate our model, and again inspired by previous approaches [6], we use the MIT low-resolution dataset [75], which contains 168 natural images that were shown to 64 observers. Please refer to Section 4 for further details on our validation.

We trained our model using the Hydra [76] and Pytorch Lightning [77] frameworks for PyTorch, logging and checkpointing all the necessary parameters to restore the training process at any point. We use the Adam optimization algorithm [78]. The learning rate has a value of 10^{-4}, and we set batch size to 1. We trained our model on a Nvidia RTX 2080 Ti with 11 GB of VRAM until convergence, for a total of 22 h.

4. Evaluation

We evaluate the quality of our scanpaths against measured, ground-truth scanpaths, as well as comparing to other existing scanpath prediction methods, using a variety of scanpath similarity metrics (Section 4.1). We also analyze the variability of our generated scanpaths (Section 4.3), whether our predicted scanpaths converge to ground-truth saliency, and the performance of our model for iterative fixation prediction (Section 4.2). Finally, we include ablation studies that justify our model choices (Section 4.4).

4.1. Comparison to other approaches

We rely on the comprehensive set of metrics proposed by Fahimi and Bruce [18], which includes similarity measures from string alignment (Levenstein distance and ScanMatch), curve similarity (Hausdorff distance and discrete Frechet distance), time-series analysis (fast dynamic time warping and time delay embedding), and recurrence analysis. Following previous literature in scanpath prediction [6], we validate our model with the MIT low resolution dataset [75], and we generate ten scanpaths per image for such test set with our method and with each of the methods we compare to: Itti et al. [1], LeMeur et al. [12], IOR-ROI [6], and the recent work from Chen et al. [7]. Scanpath length is determined by the mean length of ground-truth data [6].

A qualitative comparison can be seen in Fig. 6, while Table 1 shows the quantitative comparisons. For reference, we also include in the table a human baseline (Human BL) [38] by computing the same metrics for all the ground-truth scanpaths, plus a random baseline (Random BL) generated from random scanpaths. Our model is closest to the Human BL for seven of the ten metrics, and second in the remaining three. The qualitative comparisons in Fig. 6 illustrate some of the problems of other methods: models based on inhibition of return [6] or inspired by biological mechanisms [1] can lead to an unnatural back-and-forth behavior in certain cases. LeMeur et al.'s method [12] is closer in performance to ours, but ours still yields better results, potentially because their framework relies on a bottom-up saliency estimation from which scanpaths are then generated, while ours directly learns the scanpaths from the data. Finally, Chen et al.'s model [7] does a good job at identifying the main salient areas, but then has a tendency to stick to those and does not offer the rich variability present in the ground truth. Additional qualitative results can be found in the supplementary material.

Fig. 6. Comparison to other approaches for scanpath prediction. Each row shows representative scanpaths for a certain image; from left to right: ground truth, our method, the IOR-ROI model from Sun et al. [6], Itti et al.'s model [1], LeMeur et al.'s method [12], and Chen et al.'s method [7]. Our model generates scanpaths that present realistic variability and plausible exploration trajectories. Note that this figure only contains one sample scanpath per method for illustrative purposes. A more thorough analysis can be found in the main text. Please refer to Table 1 for quantitative metrics, and to the supplementary material for additional qualitative results.

Table 1
Results of our quantitative comparisons. We first include upper (human baseline, *Human BL*) and lower (random scanpaths, *Random BL*) baselines for reference. Then, we compare to Itti et al. [1], LeMeur et al. [19], Sun et al.'s IOR-ROI [6], and Chen et al. [7]. Boldface and underline highlight the best and second best results, respectively. Overall our model ($th = 0.7$) yields the best performance across metrics, closest to the human baseline.

Model	String alignment		Curve similarities		Time-series analysis		Recurrence analysis			
	Lev. Dist. ↓	ScanMatch ↑	Hausdorff Dist. ↓	Frechet Dist. ↓	fast DTW↓	T. D. Emb. ↓	Recurrence ↑	Determinism ↑	Laminarity ↑	CORM ↑
Human BL	10.77 (1.61)	0.38 (0.06)	95.97 (18.40)	140.02 (26.16)	550.84 (133.71)	42.40 (8.45)	6.69 (3.74)	1.72 (1.51)	6.09 (6.01)	22.11 (7.41)
Random BL	12.31 (0.88)	0.20 (0.02)	148.01 (13.76)	199.30 (13.63)	877.15 (71.66)	69.87 (4.34)	0.73 (0.34)	0.02 (0.09)	0.19 (0.25)	3.79 (1.74)
Itti et al.	14.04 (0.80)	0.23 (0.05)	160.09 (29.31)	207.97 (27.21)	1041.16 (153.97)	63.88 (9.54)	1.02 (1.98)	0.04 (0.22)	0.62 (2.03)	5.84 (6.00)
LeMeur et al.	12.58 (0.78)	**0.35** (0.04)	104.84 (12.79)	163.59 (20.52)	669.67 (108.49)	**39.75** (6.53)	2.39 (1.18)	0.40 (0.48)	2.09 (2.26)	12.54 (4.45)
IOR-ROI	13.26 (0.71)	0.30 (0.05)	115.50 (20.22)	166.07 (21.69)	777.75 (119.46)	46.98 (7.18)	1.80 (0.98)	0.18 (0.31)	0.81 (1.35)	10.28 (4.43)
Chen et al.	13.04 (1.14)	0.31 (0.07)	109.18 (27.38)	149.32 (33.03)	682.80 (183.11)	46.90 (12.55)	3.29 (3.58)	0.62 (1.10)	**6.10** (7.46)	11.70 (8.05)
Ours	**11.47** (1.13)	0.34 (0.06)	**103.44** (27.13)	**144.77** (32.77)	**610.02** (155.96)	43.74 (10.25)	**3.52** (2.86)	**0.64** (0.84)	5.05 (4.96)	**13.95** (7.92)

4.2. Additional evaluation

Spatial Convergence and Saliency We have analyzed whether our scanpaths are able to converge to saliency, i.e., whether a robust saliency map can be computed from our generated scanpaths. We have computed such maps by aggregating multiple scanpaths into a heatmap; we have then compared them against the ground-truth saliency maps computed from real observers' data. Fig. 8 depicts this comparison for four different images. We have also computed three well-known metrics for saliency evaluation (CC $= 0.49 \pm 0.21$, KLDiv $= 1.88 \pm 1.18$, SIM $= 0.49 \pm 0.12$), showing results that closely resemble the ground truth and are on par with the ability to converge to saliency of previous works in scanpath prediction that explicitly include a module trained with ground truth saliency maps [6].

Step-wise Fixation Prediction As mentioned in Section 2, most existing works take an incomplete scanpath as input, and predict the *next* fixation point. They thus build each scanpath progressively, usually by optimizing only the prediction of that last point (e.g., by means of MAE [53] or MSE [17] losses). Although this approach neglects the plausibility of the full scanpath as a whole, it may be useful in some cases. Our proposed spatio-temporal loss and probabilistic framework also offer a precise alternative in these situations. Table 3 shows quantitative results for paths of varying lengths: ○ represents points from the ground-truth scanpath fed to our network, while × represents points predicted with our model. Our method produces plausible results from a single ground-truth point, and very quickly approximates the human baseline with only four.

4.3. Scanpath variability

The output of our model is a sequence of probability maps (tSPMs), which is then sampled to produce the individual scanpaths. This allows us to adjust the variability of the generated scanpaths when doing the sampling by using a threshold th over the probability map: probabilities below th will not be sampled; thus, higher values of th lead to more similar scanpaths, while lower ones allow our scanpaths to simulate more exploratory visual behaviors. Fig. 7 illustrates this. In addition, we have conducted a quantitative analysis (see Table 2) showing how, even when eliciting a more exploratory behavior by decreasing th, our scanpaths still outperform most previous works and remain close to the human baseline.

Fig. 7. We evaluate the ability of our model to generate scanpaths with higher variability. Our model has a stochastic nature, and generates scanpaths by probabilistically sampling the predicted tSPM (see Section 3.2). Any point with a probability below a specified threshold th will be discarded before sampling. Here we show that varying th leads to different, yet realistic behaviors (see also Table 2).

Table 2
Analysis and ablation studies of our model. We first include again upper (human baseline, *Human BL*) and lower (random scanpaths, *Random BL*) baselines for reference. The next three rows show results varying our threshold th. Arrows indicate whether higher or lower is better; boldface highlights the best result for each metric. We show how decreasing the value of our parameter th (thus leading to more variability in our scanpaths) still leads to good results (see Section 4.3 for details). Finally, the last two rows show the effect of our two main design decisions, our loss function and the Bayesian 2D convolutions (see Section 4.4).

Model	String alignment		Curve similarities		Time-series analysis		Recurrence analysis			
	Lev. Dist. ↓	ScanMatch ↑	Hausdorff Dist. ↓	Frechet Dist. ↓	fast DTW↓	T. D. Emb. ↓	Recurrence ↑	Determinism ↑	Laminarity ↑	CORM ↑
Human BL	10.77 (1.61)	0.38 (0.06)	95.97 (18.40)	140.02 (26.16)	550.84 (133.71)	42.40 (8.45)	6.69 (3.74)	1.72 (1.51)	6.09 (6.01)	22.11 (7.41)
Random BL	12.31 (0.88)	0.20 (0.02)	148.01 (13.76)	199.30 (13.63)	877.15 (71.66)	69.87 (4.34)	0.73 (0.34)	0.02 (0.09)	0.19 (0.25)	3.79 (1.74)
Ours (th = 0.7)[a]	**11.47** (1.13)	**0.34** (0.06)	**103.44** (27.13)	**144.77** (32.77)	**610.02** (155.96)	**43.74** (10.25)	**3.52** (2.86)	**0.64** (0.84)	**5.05** (4.96)	**13.95** (7.92)
Ours (th = 0.5)	11.60 (0.98)	0.33 (0.06)	103.97 (23.23)	149.37 (30.09)	636.08 (146.44)	45.46 (10.30)	3.01 (2.28)	0.50 (0.58)	3.14 (2.85)	12.96 (6.73)
Ours (th = 0.35)	13.26 (0.71)	0.30 (0.05)	102.17 (19.77)	149.99 (28.42)	639.07 (138.27)	45.77 (9.63)	2.82 (2.09)	0.44 (0.54)	2.43 (2.59)	12.88 (6.21)
DTW w/o KLDiv	13.48 (0.59)	0.13 (0.03)	169.21 (15.87)	224.89 (13.08)	1134.56 (63.00)	79.74 (4.07)	0.36 (0.25)	0.01 (0.08)	0.13 (0.30)	2.45 (1.61)
w/o DTW	11.54 (1.17)	0.32 (0.08)	115.18 (33.23)	149.12 (36.60)	639.75 (177.56)	49.66 (13.48)	3.37 (3.17)	0.58 (0.90)	4.48 (4.77)	12.88 (8.05)
w/o Bayesian	11.57 (1.08)	**0.34** (0.06)	104.77 (23.76)	149.16 (29.56)	628.47 (148.55)	44.12 (10.22)	3.29 (2.60)	**0.64** (0.81)	4.87 (4.92)	13.05 (7.84)

[a] $th = 0.7$ for the last two rows.

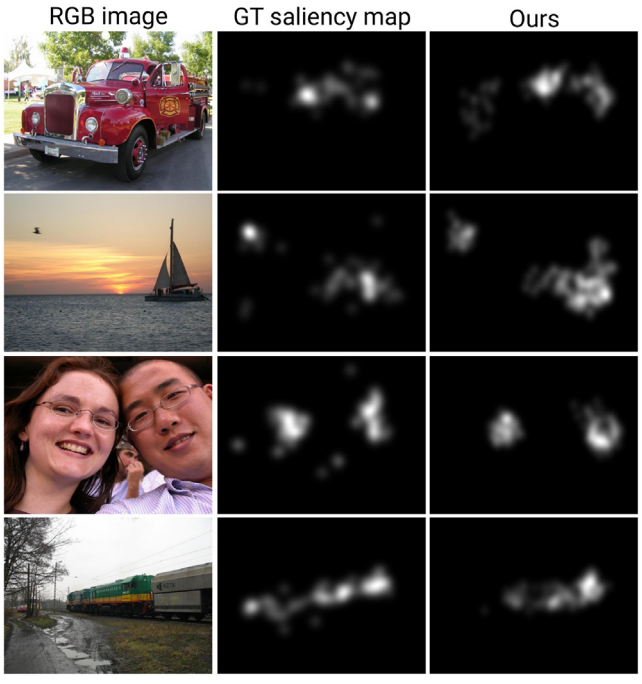

Fig. 8. We evaluate the spatial convergence of our predicted scanpaths: Our model generates scanpaths that focus on salient regions, and whose aggregation closely resembles ground-truth saliency maps.

Table 3
We have evaluated the ability of our model to complete sequences of scanpaths whose first n points are known: ○ represents such n points from the ground-truth scanpath (i.e., points that are already known), and × the points predicted with our model. Each row is computed by completing every scanpath on our test set 10 times (see Section 4.2). The first and last rows show a random baseline and the human baseline, respectively.

Scanpath	SCAM↑	HAU↓	fDTW↓	REC↑
Random BL	0.21	192.96	703.19	0.72
○××××	0.42	131.02	421.72	5.16
○○×××	0.44	129.52	407.57	6.32
○○○××	0.46	128.81	393.83	6.99
○○○○×	0.47	129.03	381.93	7.24
Human BL	0.49	126.73	387.75	7.17

4.4. Ablation studies

We have analyzed the effectiveness of the different elements that compose our model, and in particular of our main contributions: the loss function and the Bayesian convolutions. We (1) train our model on a loss function based only on KL-Div (and thus not using DTW, as indicated in Section 3.3), and (2) on a version with DTW using a basic MSE. Besides, we (3) train another variant using traditional convolutions instead of Bayesian ones in our ConvLSTM module (Section 3.4). Quantitative results of the three aforementioned studies can be found in the last three rows of Table 2: the combination of DTW and KL-Div and the Bayesian approach are key to the performance of our model.

5. Discussion and future work

Our work is not free from limitations and offers interesting avenues for future work. Since we rely on an intermediate segmentation process,

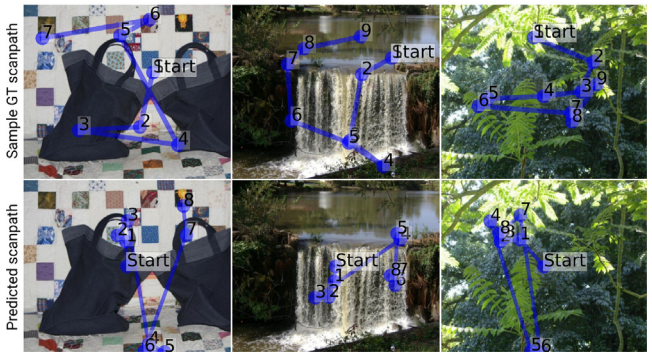

Fig. 9. When the visual features of an image are complex or too abstract, our model's performance diminishes: It sometimes fails to localize the areas of interest, or remains within the same areas for several fixations. Feeding our model with larger datasets with more varied images would probably make it more robust to these cases.

performance decreases when the input image is too abstract or complex (see Fig. 9). The inclusion of additional cues, such as depth information, could potentially improve the results in such situations.

Using a larger dataset and more ground-truth data would also likely ameliorate this. Adding more behavioral priors may also be helpful, although the difficulty resides in finding out which priors would apply in the presence of complex visual content. A more comprehensive understanding and modeling of viewing behavior uncertainty still remains an open problem. Further investigating how Bayesian deep learning could alleviate this remains an interesting avenue. Finally, our model has been shown to yield scanpaths with better spatial behavior than previous approaches; adding explicit modeling of the *duration* of the fixations and finding suitable loss functions could further enhance its performance.

6. Conclusion

We have presented a novel method for scanpath prediction in 2D images, generating a *distribution* of scanpaths whose behavior resembles that of real observers, while maintaining the inter- and intra-variability that exist amongst different viewers. To achieve this, we resort to Bayesian deep learning, and follow a probabilistic approach. Besides, we introduce a novel loss function tailored to the spatio-temporal particularities of scanpaths, based on a combination of dynamic time warping and Kullback–Leibler divergence. We have evaluated our model and compared it to state-of-the-art methods on a large set of metrics that analyze different aspects of scanpaths, showing that our model outperforms previous approaches.

CRediT authorship contribution statement

Daniel Martin: Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Methodology, Formal analysis, Data curation, Conceptualization. **Diego Gutierrez:** Writing – review & editing, Writing – original draft, Validation, Supervision, Investigation, Conceptualization. **Belen Masia:** Writing – review & editing, Writing – original draft, Validation, Supervision, Project administration, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Code will be available at https://github.com/DaniMS-ZGZ/tSPM-Probabilistic-Scanpaths.

Acknowledgments

This work has been supported by grant PID2022-141539NB-I00, funded by MICIU/AEI/10.13039/501100011033 and by ERDF, EU. This project has also received funding from the PRIME project (MSCA-ITN, grant agreement No. 956585), and from the Government of Aragon's Departamento de Ciencia, Universidad y Sociedad del Conocimiento through the Reference Research Group "Graphics and Imaging Lab" (ref T34-20R). Daniel Martin was additionally supported by a Gobierno de Aragon predoctoral grant (2020-2024).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.cag.2024.103983.

References

[1] Itti L, Koch C, Niebur E. A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 1998;20(11):1254–9.
[2] Duan L, Wu C, Miao J, Qing L, Fu Y. Visual saliency detection by spatially weighted dissimilarity. In: CVPR 2011. IEEE; 2011, p. 473–80.
[3] Yan Q, Xu L, Shi J, Jia J. Hierarchical saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2013, p. 1155–62.
[4] Soo Park H, Shi J. Social saliency prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2015, p. 4777–85.
[5] Pan J, Sayrol E, Giro-i Nieto X, McGuinness K, O'Connor NE. Shallow and deep convolutional networks for saliency prediction. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2016, p. 598–606.
[6] Sun W, Chen Z, Wu F. Visual scanpath prediction using IOR-ROI recurrent mixture density network. IEEE Trans Pattern Anal Mach Intell 2019;43(6):2101–18.
[7] Chen X, Jiang M, Zhao Q. Predicting human scanpaths in visual question answering. In: Proc. computer vision and pattern recognition. CVPR, 2021, p. 10876–85.
[8] Kapoula Z. The influence of peripheral preprocessing on oculomotor programming in a scanning task. In: Eye movements and psychological functions. Routledge; 2021, p. 101–14.
[9] Martin D, Serrano A, Bergman AW, Wetzstein G, Masia B. ScanGAN360: A generative model of realistic scanpaths for 360° images. IEEE Trans Vis Comput Graph 2022;28(5):2003–13.
[10] Kümmerer M, Bethge M. State-of-the-art in human scanpath prediction. 2021, arXiv preprint arXiv:2102.12239.
[11] Goldberg JH, Helfman JI. Visual scanpath representation. In: Proc. symposium on eye-tracking research & applications. 2010, p. 203–10.
[12] Le Meur O, Liu Z. Saccadic model of eye movements for free-viewing condition. Vis Res 2015;116:152–64.
[13] Tatler BW, Vincent BT. The prominence of behavioural biases in eye guidance. Vis Cogn 2009;17(6–7).
[14] Bao W, Chen Z. Human scanpath prediction based on deep convolutional saccadic model. Neurocomputing 2020;404.
[15] de Belen RAJ, Bednarz T, Sowmya A. ScanpathNet: A recurrent mixture density network for scanpath prediction. In: Proc. computer vision and pattern recognition (CVPR) workshops. 2022, p. 5006–16. http://dx.doi.org/10.1109/CVPRW56347.2022.00549.
[16] Arabadzhiyska E, Tursun OT, Myszkowski K, Seidel H-P, Didyk P. Saccade landing position prediction for gaze-contingent rendering. ACM Trans Graph 2017;36(4):1–12.
[17] Nguyen A, Yan Z, Nahrstedt K. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In: Proc. ACM international conference on multimedia. 2018, p. 1190–8.
[18] Fahimi R, Bruce ND. On metrics for measuring scanpath similarity. Behav Res Methods 2020;1–20.
[19] Le Meur O, Coutrot A. Introducing context-dependent and spatially-variant viewing biases in saccadic models. Vis Res 2016;121:72–84.
[20] Judd T, Ehinger K, Durand F, Torralba A. Learning to predict where humans look. In: Proc. international conference on computer vision. ICCV, 2009, p. 2106–13.
[21] Ellis SR, Smith JD. Patterns of statistical dependency in visual scanning. Eye Mov Hum Inf Process 1985;221–38.
[22] Sitzmann V, Serrano A, Pavel A, Agrawala M, Gutierrez D, Masia B, Wetzstein G. Saliency in VR: How do people explore virtual environments? IEEE Trans Vis Comput Graph 2018;36(4).
[23] Xingjian S, Chen Z, Wang H, Yeung D-Y, Wong W-K, Woo W-c. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Advances in neural information processing systems. 2015, p. 802–10.
[24] Kendall A, Gal Y. What uncertainties do we need in bayesian deep learning for computer vision? Adv Neural Inf Process Syst 2017;30.

[25] Yang L, Xu M, Guo Y, Deng X, Gao F, Guan Z. Hierarchical Bayesian LSTM for head trajectory prediction on omnidirectional images. IEEE Trans Pattern Anal Mach Intell 2021;44(11):7563–80.
[26] Kümmerer M, Bethge M, Wallis TS. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. J Vis 2022;22(5):7.
[27] Walther D, Koch C. Modeling attention to salient proto-objects. Neural Netw 2006;19:1395–407.
[28] Zhao Q, Koch C. Learning a saliency map using fixated locations in natural scenes. J Vis 2011;11:9.
[29] Lu Y, Zhang W, Jin C, Xue X. Learning attention map from images. In: Proc. computer vision and pattern recognition. CVPR, 2012.
[30] Borji A. Boosting bottom-up and top-down visual features for saliency estimation. In: Proc. computer vision and pattern recognition. CVPR, 2012.
[31] Bylinskii Z, Judd T, Borji A, Itti L, Durand F, Oliva A, Torralba A. MIT Saliency Benchmark. 2019, http://saliency.mit.edu/.
[32] Yang C, Zhang L, Lu H, Xiang R, Yang M-H. Saliency detection via graph-based manifold ranking. In: Proc. computer vision and pattern recognition. CVPR, 2013, p. 3166–73.
[33] Vig E, Dorr M, Cox D. Large-scale optimization of hierarchical features for saliency prediction in natural images. In: Proc. computer vision and pattern recognition. CVPR, 2014.
[34] Kümmerer M, Wallis TSA, Bethge M. DeepGaze II: Reading fixations from deep features trained on object recognition. 2016, arXiv preprint arXiv:1610.01563.
[35] Pan J, Sayrol E, Giro-i Nieto X, McGuinness K, O'Connor NE. Shallow and deep convolutional networks for saliency prediction. In: Proc. computer vision and pattern recognition. CVPR, 2016.
[36] Martin D, Serrano A, Masia B. Panoramic convolutions for 360° single-image saliency prediction. In: CVPR workshop on computer vision for augmented and virtual reality. 2020.
[37] Pan J, Canton C, McGuinness K, O'Connor NE, Torres J, Sayrol E, Giro-i Nieto Xa. SalGAN: Visual saliency prediction with generative adversarial networks. 2018, arXiv preprint arXiv:1701.01081.
[38] Xia C, Han J, Qi F, Shi G. Predicting human saccadic scanpaths based on iterative representation learning. IEEE Trans Image Process 2019;28(7):3502–15.
[39] Cornia M, Baraldi L, Serra G, Cucchiara R. Predicting human eye fixations via an lstm-based saliency attentive model. IEEE Trans Image Process 2018;27(10).
[40] Wang W, Shen J, Dong X, Borji A. Salient object detection driven by fixation prediction. In: Proc. computer vision and pattern recognition. CVPR, 2018.
[41] Liu H, Xu D, Huang Q, Li W, Xu M, Lin S. Semantically-based human scanpath estimation with hmms. In: Proc. international conference on computer vision. ICCV, 2013, p. 3232–9.
[42] Tavakoli HR, Rahtu E, Heikkilä J. Stochastic bottom–up fixation prediction and saccade generation. Image Vis Comput 2013;31(9):686–93.
[43] Assens Reina M, Giro-i Nieto X, McGuinness K, O'Connor NE. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In: Proc. international conference on computer vision (ICCV) workshops. 2017, p. 2331–8.
[44] Tatler BW, Brockmole JR, Carpenter RH. LATEST: A model of saccadic decisions in space and time. Psychol Rev 2017;124(3):267.
[45] Zanca D, Melacci S, Gori M. Gravitational laws of focus of attention. IEEE Trans Pattern Anal Mach Intell 2019;42(12):2983–95.
[46] Wang W, Chen C, Wang Y, Jiang T, Fang F, Yao Y. Simulating human saccadic scanpaths on natural images. In: Proc. computer vision and pattern recognition. CVPR, 2011, p. 441–8.
[47] Engbert R, Trukenbrod HA, Barthelmé S, Wichmann FA. Spatial statistics and attentional dynamics in scene viewing. J Vis 2015;15(1):14.
[48] Adeli H, Vitu F, Zelinsky GJ. A model of the superior colliculus predicts fixation locations during scene viewing and visual search. J Neurosci 2017;37(6):1453–67.
[49] Boccignone G, Ferraro M. Modelling gaze shift as a constrained random walk. Phys A 2004;331(1–2):207–18.
[50] Sun X, Yao H, Ji R, Liu X-M. Toward statistical modeling of saccadic eye-movement and visual saliency. IEEE Trans Image Process 2014;23(11):4649–62.
[51] Clarke AD, Stainer MJ, Tatler BW, Hunt AR. The saccadic flow baseline: Accounting for image-independent biases in fixation behavior. J Vis 2017;17(11):12.
[52] Assens M, Giro-i Nieto X, McGuinness K, O'Connor NE. PathGAN: Visual scanpath prediction with generative adversarial networks. In: Proc. European conference on computer vision (ECCV) workshops. 2018.
[53] Hu Z, Li S, Zhang C, Yi K, Wang G, Manocha D. DGaze: CNN-based gaze prediction in dynamic scenes. IEEE Trans Vis Comput Graph 2020;26(5):1902–11.
[54] Wloka C, Kotseruba I, Tsotsos JK. Active fixation control to predict saccade sequences. In: Proceedings of the IEEE conference on computer vision and pattern recognition. 2018, p. 3184–93.
[55] Qiao M, Xu M, Wang Z, Borji A. Viewport-dependent saliency prediction in 360° video. IEEE Trans Multimedia 2020;23:748–60.
[56] Li C, Zhang W, Liu Y, Wang Y. Very long term field of view prediction for 360-degree video streaming. In: IEEE conference on multimedia information processing and retrieval. MIPR, 2019, p. 297–302.
[57] Xu Y, Zhang Z, Gao S. Spherical DNNs and their applications in 360° images and videos. IEEE Trans Pattern Anal Mach Intell 2021.
[58] Larsson L, Nyström M, Stridh M. Detection of saccades and postsaccadic oscillations in the presence of smooth pursuit. IEEE Trans Biomed Eng 2013;60(9).
[59] Liu R, Lehman J, Molino P, Such FP, Frank E, Sergeev A, Yosinski J. An intriguing failing of convolutional neural networks and the coordconv solution. In: Neural information processing systems. 2018, p. 9605–16.
[60] Zhang K, Chen Z, Liu S. A spatial-temporal recurrent neural network for video saliency prediction. IEEE Trans Image Process 2020;30.
[61] Palazzi A, Abati D, Solera F, Cucchiara R, et al. Predicting the driver's focus of attention: the DR (eye) VE project. IEEE Trans Pattern Anal Mach Intell 2018;41(7):1720–33.
[62] Wu X, Wu Z, Zhang J, Ju L, Wang S. SalSAC: A video saliency prediction model with shuffled attentions and correlation-based ConvLSTM. In: Proc. AAAI conference on artificial intelligence. Vol. 34, 2020, p. 12410–7.
[63] Müller M. Dynamic time warping. Inf Retr Music Mot 2007;69–84.
[64] Cuturi M, Blondel M. Soft-DTW: a differentiable loss function for time-series. 2017, arXiv preprint arXiv:1703.01541.
[65] Chang C-Y, Huang D-A, Sui Y, Fei-Fei L, Niebles JC. D3TW: Discriminative differentiable dynamic time warping for weakly supervised action alignment and segmentation. In: Proc. computer vision and pattern recognition. CVPR, 2019.
[66] Song H, Wang W, Zhao S, Shen J, Lam K-M. Pyramid dilated deeper convlstm for video salient object detection. In: Proc. European conference on computer vision. ECCV, 2018, p. 715–31.
[67] Azad R, Asadi-Aghbolaghi M, Fathy M, Escalera S. Bi-directional ConvLSTM U-Net with densley connected convolutions. In: Proc. international conference on computer vision (ICCV) workshops. 2019.
[68] Blundell C, Cornebise J, Kavukcuoglu K, Wierstra D. Weight uncertainty in neural networks. 2015, arXiv:1505.05424.
[69] Wang H, Yeung D-Y. Towards Bayesian deep learning: A framework and some existing methods. IEEE Trans Knowl Data Eng 2016;28(12).
[70] Russakovsky O, Deng J, Su H, Krause J, Satheesh S, Ma S, Huang Z, Karpathy A, Khosla A, Bernstein M, et al. Imagenet large scale visual recognition challenge. Int J Comput Vis 2015;115(3):211–52.
[71] Simonyan K, Zisserman A. Very deep convolutional networks for large-scale image recognition. 2014, arXiv preprint arXiv:1409.1556.
[72] He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: Proc. computer vision and pattern recognition. CVPR, 2016, p. 770–8.
[73] Lasheras-Hernandez B, Masia B, Martin D. DriveRNN: Predicting drivers' attention with deep recurrent networks. In: Spanish computer graphics conference. CEIG, 2022.
[74] Xu J, Jiang M, Wang S, Kankanhalli MS, Zhao Q. Predicting human gaze beyond pixels. J Vis 2014;14(1):1–20.
[75] Judd T, Durand F, Torralba A. Fixations on low-resolution images. J Vis 2011;11(4):14.
[76] Yadan O. Hydra - A framework for elegantly configuring complex applications. 2019, URL https://github.com/facebookresearch/hydra. [Last Accesed June 2023].
[77] Wa F, et al. PyTorch lighning. Github; 2019, URL https://github.com/PyTorchLightning/pytorch-lightning. [Last Accesed June 2023].
[78] Kingma DP, Ba J. Adam: A method for stochastic optimization. In: ICLR. 2014, Last updated in arXiv in 2017.