TPP-Gaze: Modelling Gaze Dynamics in Space and Time with Neural Temporal Point Processes

[1]Alessandro D'Amelio, [2]Giuseppe Cartella, [2]Vittorio Cuculo,
[1]Manuele Lucchi, [2]Marcella Cornia, [2]Rita Cucchiara, [1]Giuseppe Boccignone
[1]University of Milan, Italy [2]University of Modena and Reggio Emilia, Italy
[1]{name.surname}@unimi.it [2]{name.surname}@unimore.it

Abstract

*Attention guides our gaze to fixate the proper location of the scene and holds it in that location for the deserved amount of time given current processing demands, before shifting to the next one. As such, gaze deployment crucially is a temporal process. Existing computational models have made significant strides in predicting spatial aspects of observer's visual scanpaths (*where *to look), while often putting on the background the temporal facet of attention dynamics (*when*). In this paper we present* `TPP-Gaze`*, a novel and principled approach to model scanpath dynamics based on Neural Temporal Point Process (TPP), that jointly learns the temporal dynamics of fixations position and duration, integrating deep learning methodologies with point process theory. We conduct extensive experiments across five publicly available datasets. Our results show the overall superior performance of the proposed model compared to state-of-the-art approaches. Source code and trained models are publicly available at:* `https://github.com/phuselab/tppgaze`.

1. Introduction

Gaze, the act of directing the eyes toward a location in the visual world, is considered a good measure of overt attention and, more generally, a window to the observer's thoughts, intentions, and emotions [10, 15]. It is no surprise that research spanning decades has struggled to produce several computational models aiming at effectively predicting attention towards regions or events within the landscape of visual and multimodal stimuli. With roots in psychology and neuroscience, these approaches have gained traction in the computer vision and pattern recognition fields since the seminal Itti *et al.* [35] model; more recently, state-of-the-art approaches rely on machine learning advancements, typically employing deep neural architectures to the purpose (but see [39] or [14] for an in-depth review). As a matter of fact, the vast majority of works in the field has focused on the computational modelling of spatial saliency in the shape of saliency maps, namely, a topographic map representing the likelihood of fixating a given location of the scrutinised stimulus, a fixation being defined as the period of time during which a part of the visual stimulus (the patch) on the screen is gazed at. Nevertheless, a growing number of models (*i.e.*, scanpath models) are addressing the prediction of a sequence of fixations – namely, the scanpath – where the gaze shift from one fixation to the next represents a saccade. Beyond the salience representation, these models explicitly unfold the dynamics of overt attention allocation over a stimulus [6, 40]. It is worth remarking, though, that barely predicting the spatial sequence of fixations, does not entail proper modelling of the temporal evolution of attention. By and large, most scanpath models predict an ordered sequence of events while neglecting their continuous timestamp information. As a result, these models are able to tell *where* to look and in what order, but fail in answering *when*. In many respects, this is not an innocent flaw: human actions often rely on visual information, therefore it is important to direct attention to the right place at the right time [54]. Practically, modelling *when* to perform a saccade translates to devising scanpath models able to predict the sequence of both fixations position and corresponding

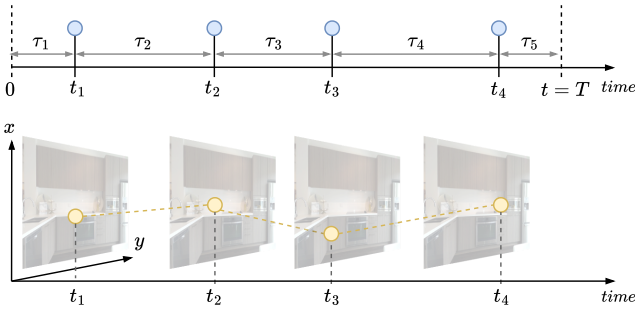

Figure 1. Scanpath dynamics as a marked TPP. Time is represented on the horizontal axis, and different scanpath fixations occurs at time t_1, t_2, t_3 and t_4.

duration. Albeit recently few approaches have successfully dealt with such problem [17, 18, 20, 47, 53] via fully engineered approaches, only a marginal subset of them has tackled it in a mathematically principled way [7, 24, 54]. This has typically resulted in a weaker generality of the methods which are tailored to specific contexts or applications. Under such circumstances, the chief concern of the present work is to introduce a fresh, general and simple view on the problem of scanpath modelling: in brief, we consider a scanpath as the realisation of a point process in space and time, precisely that of a Neural Temporal Point Process.

Temporal Point Processes (TPPs) are probabilistic generative models designed for continuous-time event sequences. Neural TPPs [46, 50, 66, 67] integrate key concepts from point process literature with deep learning methodologies, facilitating the creation of adaptable and effective models. Notably, the modelling assumptions of Neural TPPs align perfectly with the structure of scanpath data. A scanpath consists of a series of events (saccades) occurring at irregular intervals (fixation durations), which is *exactly* what Neural TPPs are designed to model. While the psychological and neuroscience literature has used traditional point processes for eye movement analysis [5, 28, 62], these tools are not well-suited for scanpath prediction due to their inability to handle stimuli. In other words, traditional TPPs are effective for studying the observer but fall short when addressing Computer Vision tasks related to attention allocation prediction. In contrast, Neural TPP-based models offer the best of both worlds: they combine the robust theoretical framework of TPPs with the flexibility and power of modern neural networks. Nevertheless, this is the first attempt to adopt them for the scanpath modelling problem.

Our key contributions can be summarised as follows: 1) We propose a novel scanpath model able to jointly learn the temporal dynamics of both fixations position and duration. 2) We extend recent Neural TPP models to deal with visual data (*i.e.*, images) and connect scanpath modelling and prediction to point process theory. To assess our proposal, which can be appreciated at a glance in Fig. 1, we conduct experiments on five publicly available datasets, showing an overall superior performance of the proposed model when compared to state-of-the-art approaches.

2. Background and Related Work

2.1. Neural Temporal Point Processes (TPPs)

Consider a sequence of generic events happening irregularly over time, TPPs model the next arrival time of an event by conditioning on the past events. Specifically, denote $\mathcal{H}_t = \{t_n \in \mathcal{T} : t_n < t\}$ (with $\mathcal{T}$ representing the sequence of strictly increasing arrival times of events) the history of arrival times of all events up to time t, the relation between the current arrival time t and the history, is typically determined by the conditional intensity function $\lambda^*(t) = \lambda(t|\mathcal{H}_t)$, whose functional form determines the properties of the TPP. Equivalently, the sequence of positive inter-event times $\tau_n = t_n - t_{n-1}$ can be considered. Knowing the conditional intensity function allows to recover the conditional probability of the inter-arrival time of an event:

$$\begin{aligned} p^*(\tau_n) &= p(\tau_n|\mathcal{H}_{t_n}) \\ &= \lambda^*(t_{n-1}+\tau_n)\exp\left(-\int_0^{\tau_n}\lambda^*(t_{n-1}+s)ds\right). \end{aligned} \tag{1}$$

For instance, under the the assumptions of no dependence on the history and constancy over time (*i.e.*, $\lambda^*(t) = k$, with $k \geq 0$), the homogeneous Poisson process is recovered, with inter-event times distributed according to the exponential distribution. Choosing more complex functional forms for $\lambda^*(t)$ allows to recover many well known TPPs such as Hawkes or self-correcting processes [30, 34]. Clearly, restricting $\lambda^*(t)$ to a specific parameterisation limits the general applicability of TPPs. For this reason, most recent solutions resorted to neural approaches (Neural TPPs) implementing learnable parametric forms of the intensity function, $\lambda^*_\theta(t)$ [27, 33]. As an example, early Neural TPPs, such as the Neural Hawkes Process [46], used RNNs to model the intensity function of the process. More recently, self-attention mechanisms have been employed to the same purpose [66, 67]. The choice of the parametric form for the intensity function has to take into account the necessity of a closed form solution of the integral in Eq. (1), thus practically restricting the expressiveness of the model. More complex parametric forms would require Monte Carlo approximation of the integral [46]. To address this, Shchur *et al.* [50] proposed to directly learn the parametric conditional distribution $p^*_\theta(\tau)$ of the inter-arrival times rather than the conditional intensity function $\lambda^*_\theta(t)$, thus recasting learning Neural TPPs as a density estimation problem.

Marked TPPs. The basic mathematical formalism of TPPs allows to naturally handle the dynamics of arrival times of events. However, the distribution of time until the next event might depend on factors other than the history. Event data is often accompanied with some kind of covariate indicating the nature of the specific event being predicted. In the realm of TPPs, such covariate are called *marks*. More formally, a marked TPP is a random process whose realisations consists of a sequence of discrete events localised in time, $\{\mathbf{r}_{F_n}, t_n\}$, with the timing $t_n \in \mathbb{R}^+$ and the mark $\mathbf{r}_{F_n} \in \mathcal{M}$. The mark $\mathbf{r}_{F_n}$ is typically modelled as an integer representing the type of event, however other kinds of marks (*e.g.*, $\mathcal{M} = \mathbb{R}^2$) can be eventually adopted. Specifying a marked-TPP involves the definition of the joint conditional density function of the next event, with inter-event time τ_n and mark $\mathbf{r}_{F_n}$, given the history of past events: $p^*(\mathbf{r}_{F_n}, \tau_n) = p(\mathbf{r}_{F_n}, \tau_n|\mathcal{H}_{t_n})$. By assuming a conditional distribution parameterised by the

weights of a neural model, $p^*_\theta(\mathbf{r}_{F_n}, \tau_n)$, inference can be performed by maximising the joint likelihood of the N observed events in a sequence:

$$\theta^* = \arg\max_\theta \prod_{n=0}^{N} p_\theta(\mathbf{r}_{F_n}, \tau_n | \mathcal{H}_t) = \prod_{n=0}^{N} p^*_\theta(\mathbf{r}_{F_n}, \tau_n). \tag{2}$$

Applications of Neural TPPs span a variety of research fields [51], such as healthcare [29], finance [4], social network analysis [55], earthquake forecasting [11], and recommender systems [38]. In this work, we leverage the Neural TPP framework to model attention dynamics on visual data.

2.2. Scanpath Modelling

Modelling scanpaths involves defining a mapping from visual data, $\mathbf{I}$ (raw image data representing either a static picture or a stream of images), to a sequence of time-stamped gaze locations $\mathcal{S} = \{(\mathbf{r}_{F_1}, t_1), (\mathbf{r}_{F_2}, t_2), \dots (\mathbf{r}_{F_N}, t_N)\}$. Here $\mathbf{r}_{F_n} \in \mathbb{R}^2$ represents the two-dimensional vector of spatial coordinates of the $n-$th fixation on the stimulus $\mathbf{I}$, while $t_n \in \mathbb{R}^+$ represents its arrival time. Eventually, a perceptual representation of the input stimuli, $\mathcal{Z}$, is computed, with the aim of locating the relevant objects inside the scene:

$$\mathbf{I} \rightarrow \mathcal{Z} \rightarrow \{(\mathbf{r}_{F_1}, t_1), (\mathbf{r}_{F_2}, t_2), \dots (\mathbf{r}_{F_N}, t_N)\}. \tag{3}$$

Here we assume that no specific external task or goal is given to the observer (*i.e.*, free-viewing condition). Notably, the dynamics of the attentive process, which unrolls as a sequence of fixations location with corresponding duration/arrival time, is characterised by an inherent randomness which likely stems from internal stochastic fluctuations affecting sensory and information processing, movement planning, and execution [56], in both fixations location and corresponding duration. Notably, many scanpath models proposed in the recent literature [2, 3, 41, 52] get rid of fixations' timestamp information by rearranging the sequence $\{(\mathbf{r}_{F_1}, t_1), (\mathbf{r}_{F_2}, t_2), \dots\}$ as $\{\mathbf{r}_F(1), \mathbf{r}_F(2), \cdots\}$, thus assuming $(\mathbf{r}_{F_n}, t_n) = \mathbf{r}_F(n)$.

Several approaches [7, 24, 54] have dealt with this problem comprehensively, relying on specific theoretical frameworks. Tatler *et al.* [54] modeled saccade timings as an evidence accumulation process with clear neurobiological significance. Similalry, in [24] a Langevin-type SDE race model [9] was adopted to predict fixations and their duration in socially relevant contexts, while in [7] fixation duration was equated to the patch residence time of a forager searching for nourishment. Conversely, the vast majority of recent methods [17, 18, 20, 47, 53] simply model fixation duration by employing specific neural architectural choices that aim at associating each fixation to its corresponding duration.

In a different vein, this work recasts the whole visual attention allocation process in the mathematical framework of point process theory [23]. This emphasises the central role of visual attention's spatio-temporal dynamics by explicitly modelling scanpaths as sequences of discrete events happening at irregular intervals. Specifically, we conceive a scanpath as a realisation of a random process whose events happen at strictly increasing arrival times $\mathcal{T} = \{t_1, \dots, t_N\}$. Fixations duration can be recovered by resorting to inter-event times $\tau_n = t_n - t_{n-1}$, while their locations can be represented as the two-dimensional continuous mark associated to the n-th event. Under this assumption, (Neural) Temporal Point Processes (TPPs) represent the natural choice for modelling this kind of data.

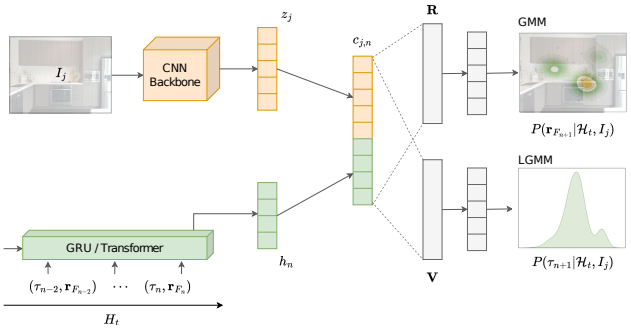

Figure 2. Overview of `TPP-Gaze` model architecture. Given a semantic representation of the image (z_j) and the history of past events (h_n), the next fixation position and duration are simulated.

3. Proposed Method

Given a stimulus (image) $\mathbf{I}_j$, an ensemble of N_{obs} observers performs a sequence of fixations and saccades (scanpath) on it, thus obtaining a set of sequences $\mathcal{C}_j = \{S^1, \dots, S^{N_{obs}}\}$. Each scanpath S^i is a sequence of pairs (events) $S^i_n = (\mathbf{r}_{F_n}, t_n)$ each composed by a fixation position (marker) $\mathbf{r}_{F_n} \in \mathbb{R}^2$, and a corresponding arrival time $t_n \in \mathbb{R}^+$. At the most general level, we are interested in modelling the stochastic generative process that given a semantic representation of the image $\mathcal{Z}_j$ and the history of past events $\mathcal{H}_t$, simulates the next fixation position and duration. More formally:

$$S^i_{n+1} \sim p_\theta(\mathbf{r}_{F_{n+1}}, t_{n+1} | \mathcal{H}_t, \mathcal{Z}_j), \tag{4}$$

where $p_\theta(\cdot)$ represents the parametric joint conditional distribution of a Neural TPP [50].

3.1. Architecture

In the following, we present the architecture of `TPP-Gaze`, implementing a scanpath model on an image as a Neural TPP.

Representing Scene Semantics. As outlined in Eq. (3), the sequence of events composing a scanpath depends not only on the history of past events, but on a perceptual representation of the input stimulus $\mathbf{I}_j$, encoding scene semantics and

relevant objects location. We extract the perceptual representation of the input image via a CNN architecture inspired by [41]. Specifically, the input image is first processed by a pre-trained DenseNet201 CNN [32]. Activation maps from various convolutional layers (as reported in [41]) are extracted, thus obtaining a $2,048$ channels volume, each representing the location of semantic features inside the scene. It is worth noticing that learning to predict fixations location (*i.e.*, marks) involves a mapping between coordinates in Cartesian space, a task in which standard convolutions have been reported to fail [43]. In the vein of [45, 52], we adopt a CoordConv layer [43] to give convolutions access to their own input coordinates. This results in a $2,051$ channels volume which is fed as input to 3 layers of 1×1 convolutions, followed by a linear layer mapping to $\mathbf{z}_j$ acting as our semantic representation.

Representing History. Neural TPPs employ either Recurrent Neural Networks (RNNs) and their variants (*e.g.*, LSTM, GRU) [27, 50, 57] or Transformer encoders [66, 67] to model the nonlinear dependency over both the markers and the timings from past events [51]. As shown in Fig. 2, the pair $(\mathbf{r}_{F_n}, \tau_n)$ representing the event occurring at the time t_n with fixation position $\mathbf{r}_{F_n}$ and duration $\tau_n = t_n - t_{n-1}$, is fed as the input into either a GRU or a Transformer encoder as described in [67]. The Transformer/GRU state embedding $\mathbf{h}_n$ represents the influence of the history up to the $n-$th fixation. Hence, can be employed as a vector space representation of $\mathcal{H}_{t_n}$. Taking into account the semantic representation $\mathbf{z}_j$ and the history embedding $\mathbf{h}_n$, Eq. (4) can be rewritten as:

$$S^i_{n+1} \sim p_\theta(\mathbf{r}_{F_{n+1}}, t_{n+1} | \mathbf{h}_n, \mathbf{z}_j). \tag{5}$$

Fixation Duration Generation. We model the conditional dependence of the distribution $p_\theta(\tau_{n+1}|\mathbf{h}_n, \mathbf{z}_j)$ on both past events and stimulus by concatenating the history embedding and semantic vectors into a context vector $\mathbf{c}_{j,n} = [\mathbf{h}_n || \mathbf{z}_j]$. In the vein of [50], the latter is employed to learn the parameters of a Log-Gaussian Mixture Model (LGMM) via an affine transform:

$$\begin{gathered} \mathbf{w} = softmax(\mathbf{V}_\mathbf{w}\mathbf{c}_{j,n}) \quad \mathbf{s} = \exp(\mathbf{V}_\mathbf{s}\mathbf{c}_{j,n}) \\ \boldsymbol{m} = \mathbf{V}_\mathbf{m}\mathbf{c}_{j,n} \end{gathered} \tag{6}$$

where $\boldsymbol{w} \in \mathbb{R}^K_+$ are the mixture weights, $\boldsymbol{m} \in \mathbb{R}^K$ are the mixture means, and $\boldsymbol{s} \in \mathbb{R}^K_+$ are the standard deviations. K represents the number of mixture components. The fixation duration for the $n-$th event can be generated by sampling from the LGMM defined by:

$$\begin{aligned} p^*_\theta(\tau_n|\mathbf{c}_{j,n}) &= p(\tau_n|\mathbf{w}, \boldsymbol{m}, \mathbf{s}) \\ &= \sum_{k=1}^{K} w_k \frac{1}{\tau_n s_k \sqrt{2\pi}} \exp\left(-\frac{(\log \tau_n - m_k)^2}{2s_k^2}\right). \end{aligned} \tag{7}$$

Fixation Position (Mark) Generation. Similarly, given the context vector $\mathbf{c}_{j,n}$, we define the conditional probability of the next mark (fixation position), $p_\theta(\mathbf{r}_{F_{n+1}}|\mathbf{h}_n, \mathbf{z}_j)$, as a 2D Gaussian Mixture Model (GMM) whose parameters are obtained via another affine projection:

$$\begin{gathered} \boldsymbol{\omega}_g = softmax(\mathbf{R}^g_{\boldsymbol{\omega}}\mathbf{c}_{j,n}) \quad \boldsymbol{\Sigma}_g = diag(\exp(\mathbf{R}^g_{\boldsymbol{\Sigma}}\mathbf{c}_{j,n})) \\ \boldsymbol{\mu}_g = \mathbf{R}^g_{\boldsymbol{\mu}}\mathbf{c}_{j,n} \end{gathered} \tag{8}$$

where $\boldsymbol{\omega}_g \in \mathbb{R}^2_+$ are the mixture weights, $\boldsymbol{\mu}_g \in \mathbb{R}^2$ are the mixture means, and $\boldsymbol{\Sigma}_g \in \mathbb{R}^{2\times 2}$ are the diagonal covariance matrices of G bi-variate Gaussian distributions. The x and y coordinates of the $n-$th fixation can be generated by sampling from the GMM defined by:

$$\begin{aligned} p^*_\theta(\mathbf{r}_{F_n}|\boldsymbol{c}_{j,n}) &= p(\mathbf{r}_{F_n}|\boldsymbol{\omega}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) \\ &= \sum_{g=1}^{G} \boldsymbol{\omega}_g \frac{\exp\left(-\frac{1}{2}(\mathbf{r}_{F_n} - \boldsymbol{\mu}_g)^\mathrm{T} \boldsymbol{\Sigma}^{-1} (\mathbf{r}_{F_n} - \boldsymbol{\mu}_g)\right)}{\sqrt{(2\pi)^2 |\boldsymbol{\Sigma}_g|}}. \end{aligned} \tag{9}$$

3.2. Model Inference

Consider a set of stimuli $\mathcal{I} = \{\mathbf{I}_1, \ldots, \mathbf{I}_j, \ldots, \mathbf{I}_J\}$ each gazed by N_{obs} human observers. Each observer produces an ensemble of scanpaths $\mathcal{C}_j = \{S^1, \ldots, S^{N_{obs}}\}$ with $S^i_n = (\mathbf{r}^i_{F_n}, \tau^i_n)$ representing an event (*i.e.*, fixation position and duration). Model inference is performed by minimising a negative log-likelihood loss with respect to the parameters of the semantic network, the GRU/Transformer encoding history of events, and the affine transforms of the LGMM and GMM. Formally, the loss function is defined as follows:

$$\mathcal{L}(\boldsymbol{\theta}) = -\sum_j \sum_i \sum_n \left[\log p^*_\theta(\tau^i_n|\mathbf{c}_{j,n}) + \log p^*_\theta(\mathbf{r}^i_{F_n}|\mathbf{c}_{j,n})\right]. \tag{10}$$

4. Experiments

4.1. Experimental Setup

Datasets. Regarding the stimuli and eye tracking data, we select five publicly available datasets of human recorded scanpaths comprising both fixation positions and durations: COCO-FreeView, MIT1003, OSIE, NUSEF, and FiFa.

COCO-FreeView [61] is a high-quality dataset capturing free viewing behaviour, featuring the same natural images adopted in COCO-Search18, annotated with $822,602$ eye fixations from a free-viewing task. Only train and validation splits are publicly released. Each image was presented for 5 seconds. The MIT1003 dataset [37] comprises $1,003$ images primarily featuring natural scenes. It provides eye movement data from 15 subjects, observing stimuli for 3 seconds. The OSIE dataset [58] comprises 700 images with eye-tracking data of 15 viewers. The dataset was explicitly

devised to incorporate high-level semantic attributes. The NUSEF (NUS Eye Fixation) dataset [48] features a diverse collection of images, representing a range of semantic concepts and capturing objects with varying scale, illumination, and orientation. Each free-view experiment lasted 5 seconds. The Fixations In Faces (FiFa) database [16] shares data related to observers' viewing of faces in natural settings.Each image was presented for 2 seconds.

Implementation and Training Details. COCO-FreeView and MIT1003 datasets are used for model training. To this end, 70% of the images from both datasets are used for training, while the remaining 30% is equally partitioned between validation and test sets. We use AdamW as optimizer, with weight decay set to 10^{-1}, and the learning rate set to 10^{-3}. Batch size is equal to 128. We employ early stopping after 20 epochs with no improvement on the validation set. Following previous literature [20, 41], during training and evaluation, we discard the first fixation and removed all scanpaths containing less than four fixations.

Scanpath Evaluation Metrics. A variety of scanpath evaluation metrics have been proposed to quantitatively assess the similarity between real and simulated eye-movements [1, 39]. Here we employ the MultiMatch, ScanMatch, and Sequence Score evaluation metrics since they explicitly consider fixation duration in the evaluation process. Moreover the String Edit Distance is adopted to further evaluate predicted scanpaths.

MultiMatch (MM) [25, 36] assesses scanpaths based on five features: shape (Sh), length (Len), direction (Dir), position (Pos), and duration (Dur). Scanpaths are temporally aligned and compared using the Dijkstra algorithm. Similarity is determined by applying vector arithmetic to the aligned saccade pairs. ScanMatch (SM) [22] encodes scanpaths as letter sequences by segmenting them into spatial and temporal bins. In our experiments, the longest dimension of the stimuli is divided into 14 bins, while the shortest is split into 8 bins. The temporal bin size is set to 50 ms for scanpath models delivering fixation duration estimates. The encoded scanpaths are then aligned and compared, with higher scores reflecting greater spatial, temporal, and sequential similarity. Sequence Score (SS) [59] transforms the human and predicted scanpaths into sequences of fixation cluster IDs and compares them using a string-matching algorithm. String Edit Distance (SED) [10], first partitions the input stimulus into an $n \times n$ grid. Scanpaths are then transformed into strings and the string-edit algorithm calculates the distance between them.

Evaluation Protocol. We compare the scanpaths synthesised from various models with those recorded from human observers. The objective is to evaluate whether the simulated behaviours exhibited statistical properties closely resembling those exhibited by human subjects who are eye-tracked while viewing a given stimulus. The evaluation protocol unfolds as follows. Suppose there are N_{obs} human observers. For each stimulus, we first compute the evaluation scores for every possible pair of the N_{obs} observers (Real vs. Real). Then, for each model, (i) we generate gaze trajectories from artificial observers and (ii) calculate the evaluation scores for every possible pair of real and artificial scanpaths (Real vs. Simulated).

For a given metric this procedure yields a target distribution P of similarity scores between observers (Real vs. Real) and a distribution Q of similarity scores for the given model w.r.t. humans (Real vs. Simulated). As reported in [40], MM, SM, and SS average values may deliver inconsistent results: models exhibiting less variability w.r.t. humans, can score systematically better than the ground truth model. This issue can be tackled by considering a good model as the one that minimises the discrepancy between the target and model-derived score distributions. We quantified such discrepancy using the Kullback-Leibler Divergence (KL-Div): $D_{\mathrm{KL}}(P \parallel Q) = \sum_{x \in \mathcal{X}} P(x) \log(P(x)/Q(x))$. Conversely, as SED is an evaluation metric not requiring alignment, it is not susceptible to the inconsistency issues associated with MM, SM, and SS. Consequently, its values are directly reported without any further processing.

	Dim		GMM		MM (KL-Div) ↓		SM (KL-Div) ↓		SED ↓
CNN	Img	TPP	K	G	Dur	Avg	w/ Dur	w/o Dur	Avg
Image Backbone									
RN	256	256	4	16	**0.011**	0.037	0.113	0.101	17.575
DN	256	256	4	16	0.012	**0.028**	**0.078**	**0.060**	**17.032**
Image and TPP Dimensionalities									
DN	128	128	4	16	0.010	0.031	0.094	0.069	16.959
DN	128	256	4	16	0.012	0.030	0.084	0.063	**16.887**
DN	256	128	4	16	0.009	0.037	0.105	0.082	17.413
DN	256	256	4	16	0.012	0.028	**0.078**	**0.060**	17.032
DN	256	512	4	16	0.010	0.031	0.101	0.095	17.462
DN	512	256	4	16	**0.008**	0.032	0.110	0.093	17.497
DN	512	512	4	16	0.009	**0.027**	0.104	0.077	17.154
Mixture Components									
DN	256	256	2	16	0.014	**0.027**	0.092	0.071	**16.944**
DN	256	256	4	16	0.012	0.028	**0.078**	**0.060**	17.032
DN	256	256	2	32	**0.009**	0.030	0.109	0.098	17.216
DN	256	256	4	32	**0.009**	0.031	0.103	0.076	17.252

Table 1. Ablation study results comparing different model configurations and hyperparameters. We report the results for ResNet50 (RN) and DenseNet201 (DN) visual backbones, various embedding vector dimensions for the image representation and the TPP history, and different numbers of Gaussian mixture components.

4.2. Scanpath Prediction

Ablation Studies. The `TPP-Gaze` architecture consists of three main blocks: image encoding (CNN backbone), history encoding (RNN/Transformer), and fixation/inter-time prediction (GMM/LGMM). To break down these components and make the adopted design choices explicit, we perform extensive ablation studies. Specifically, we evaluate

	COCO-FreeView											MIT1003										
	MM (KL-Div) ↓						SM (KL-Div) ↓		SS (KL-Div) ↓		SED ↓	MM (KL-Div)↓						SM (KL-Div) ↓		SS (KL-Div) ↓		SED ↓
	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur	Avg	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	w/ Dur	w/o Dur	Avg
Itti-Koch [35]	0.42	0.40	0.21	1.02	-	0.51	-	2.54	-	1.01	14.00	0.91	0.64	0.71	1.53	-	0.95	-	2.27	-	6.30	8.86
CLE (Itti) [8, 35]	0.07	0.30	0.35	1.43	-	0.54	-	2.50	-	1.27	14.37	0.10	0.10	0.32	1.04	-	0.39	-	2.16	-	6.25	9.09
CLE (DG) [8, 42]	0.06	0.18	0.23	1.31	-	0.44	-	2.37	-	1.22	14.31	-	-	-	-	-	-	-	-	-	-	-
G-Eymol [63]	0.37	0.73	0.93	1.22	1.99	1.05	9.00	6.67	8.75	6.30	14.20	0.68	0.68	0.46	1.54	1.03	0.88	15.90	4.89	3.32	6.96	6.96
IOR-ROI-LSTM [20]	1.15	0.47	0.03	0.19	0.05	0.38	1.54	0.76	0.56	0.64	13.55	0.59	0.27	0.07	0.57	0.05	0.31	0.69	0.45	5.08	8.61	8.61
DeepGazeIII [41]	0.04	0.02	0.03	**0.03**	-	**0.03**	-	0.33	-	0.33	13.15	-	-	-	-	-	-	-	-	-	-	-
Scanpath-VQA [17]	0.05	0.16	0.10	0.06	0.25	0.12	1.07	0.34	0.43	0.28	12.76	0.04	0.05	0.08	0.05	0.14	0.07	0.06	0.05	0.05	0.11	7.26
DeepGazeIII [41]	**0.01**	0.03	0.05	0.05	-	0.04	-	0.34	-	0.36	**13.15**	0.05	**0.01**	0.20	0.05	-	0.08	-	0.19	-	5.06	8.28
Scanpath-VQA [17]	0.62	0.41	**0.02**	0.05	0.03	0.23	**0.08**	**0.03**	**0.03**	0.31	14.34	0.20	0.14	0.15	0.08	0.02	0.12	0.23	0.19	0.14	0.25	9.27
`TPP-Gaze` (GRU)	0.06	0.02	**0.02**	**0.03**	**0.01**	**0.03**	**0.08**	0.06	0.05	**0.11**	17.03	**0.01**	0.03	**0.09**	**0.04**	**0.01**	**0.04**	**0.15**	**0.11**	**0.12**	**0.11**	**7.21**
`TPP-Gaze` (Trans.)	0.05	**0.01**	**0.02**	**0.03**	**0.01**	**0.03**	0.10	0.07	0.06	0.12	16.93	**0.01**	0.02	**0.09**	0.07	0.02	**0.04**	0.22	0.16	0.14	0.14	7.33

Table 2. Comparison of various models on COCO-FreeView and MIT1003. **Gray color** indicates models trained under the same settings and datasets. Within this group, **bold** values represent the best performance for each metric. Underline values indicate the overall best performance across all models and metrics.

	OSIE								NUSEF								FiFa							
	MM (KL-Div) ↓						SM (KL-Div) ↓		MM (KL-Div) ↓						SM (KL-Div) ↓		MM (KL-Div) ↓						SM (KL-Div) ↓	
	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur	Sh	Len	Dir	Pos	Dur	Avg	w/ Dur	w/o Dur
Itti-Koch [35]	1.62	0.89	0.45	3.69	-	1.66	-	2.22	0.63	0.44	0.17	0.56	-	0.45	-	0.61	1.51	0.51	1.08	3.46	-	1.64	-	6.08
CLE (Itti) [8, 35]	0.13	0.03	0.20	0.75	-	0.28	-	1.98	0.26	0.03	0.09	0.42	-	0.20	-	0.79	0.38	0.10	0.29	1.14	-	0.48	-	3.97
CLE (DG) [8, 42]	0.17	0.03	0.15	0.60	-	0.24	-	1.43	0.28	0.06	0.06	0.18	-	0.15	-	0.50	0.40	0.14	0.36	0.97	-	0.46	-	3.10
G-Eymol [63]	1.18	1.08	0.25	2.12	1.18	1.16	16.17	7.29	0.38	0.30	0.05	0.29	3.02	0.81	1.76	0.55	0.34	0.57	0.59	2.48	2.40	1.28	17.36	11.71
IOR-ROI-LSTM [20]	1.72	0.73	0.03	0.96	0.03	0.69	0.75	0.76	0.90	0.36	0.12	0.23	0.17	0.36	0.11	0.13	1.24	0.51	0.10	1.71	0.05	0.72	1.25	1.56
DeepGazeIII [41]	0.14	0.08	0.06	0.15	-	0.11	-	0.12	0.10	0.06	0.08	0.05	-	0.07	-	0.07	0.28	0.12	0.21	0.34	-	0.24	-	0.60
Scanpath-VQA [17]	0.07	0.07	0.04	0.04	0.16	0.08	0.03	0.03	0.11	0.04	0.02	0.05	0.08	0.06	0.02	0.03	0.14	0.04	0.13	0.07	0.12	0.10	0.03	0.13
DeepGazeIII [41]	0.04	**0.03**	0.09	0.14	-	0.08	-	**0.22**	0.11	0.07	0.09	0.04	-	0.08	-	0.06	0.25	0.13	0.40	**0.18**	-	0.24	-	0.69
Scanpath-VQA [17]	0.49	0.35	0.09	0.20	**0.02**	0.23	0.40	0.28	0.11	0.07	0.06	0.03	0.16	0.09	0.06	0.06	0.44	0.26	0.33	0.31	0.08	0.28	0.47	0.79
`TPP-Gaze` (GRU)	0.03	0.04	**0.05**	**0.12**	0.03	**0.05**	**0.20**	0.30	**0.03**	0.02	**0.01**	0.02	**0.10**	**0.04**	**0.04**	0.04	**0.05**	**0.05**	**0.12**	0.25	**0.05**	**0.10**	**0.23**	**0.47**
`TPP-Gaze` (Trans.)	**0.02**	0.04	0.06	0.14	0.05	0.06	0.25	0.44	**0.03**	**0.01**	0.02	**0.01**	0.13	**0.04**	**0.04**	**0.01**	0.06	**0.05**	**0.12**	0.30	**0.05**	0.12	0.32	0.52

Table 3. Comparison of various models on OSIE, NUSEF, and FiFa datasets. **Gray color** indicates models trained under the same settings and datasets. Within this group, **bold** values represent the best performance for each metric. Underline values indicate the overall best performance across all models and metrics.

two different CNN backbones for image encoding (*i.e.*, a ResNet50 [31] and a DenseNet201 [32]) as well as three embedding vector dimensions for the image semantic representation ($\mathbf{z}_j$) and the history embedding ($\mathbf{h}_n$, TPP dimensionality). Moreover, different numbers of components for the GMM/LGMM are considered. Table 1 reports the results of the ablation studies conducted on the COCO-FreeView dataset. In our experiments, we select the hyperparameters yielding the best trade-off according to the considered evaluation metrics, resulting in a DesNet201 backbone and a dimensionality equal to 256 for all embedding vectors. Moreover, the parameters K and G representing mixture components are respectively set to 4 and 16.

Comparison with the State of the Art. To compare the proposed approach with others, we include state-of-the-art approaches that either reach high performance in recent scanpath benchmarks [39], offer source code availability, and are representative of different approaches and architectures. As to the latter criteria, following the taxonomy proposed in [39], scanpath models can be aggregated into the following categories: biologically inspired (*e.g.* Itti-Koch model [35] and G-Eymol [63]); statistically inspired (*e.g.* CLE model [8]); cognitively inspired (*e.g.* IOR-ROI-LSTM [20]); engineered models (*e.g.* DeepGazeIII [41] and Scanpath-VQA [17]); but see [39–41] for an in-depth review. Under such circumstances, we assess the performance of `TPP-Gaze` against the aforementioned models.

Table 2 reports quantitative results on the COCO-FreeView and MIT1003 datasets in terms of all considered metrics, while model performance on OSIE, NUSEF, and FiFa are shown in Table 3 in terms of MM and SM[1]. In all experiments, we compare the aforementioned approaches using the pre-trained model weights released by the authors. As DeepGaze models were trained on the entire MIT1003 dataset, the results from DeepGazeIII and CLE (DG) have not been included in this comparison. Additionally, to explicitly measure the effect of the proposed architecture and mathematical framework, we retrain and test the two most recent models (DeepGazeIII and Scanpath-VQA) under the same conditions adopted for `TPP-Gaze` (see Sec. 4.1). Specifically, beyond training on the same data, the large-scale pre-training of DeepGazeIII as well as

[1]We refer to the supplementary material for the results in terms of SS and SED on OSIE, NUSEF, and FiFa datasets.

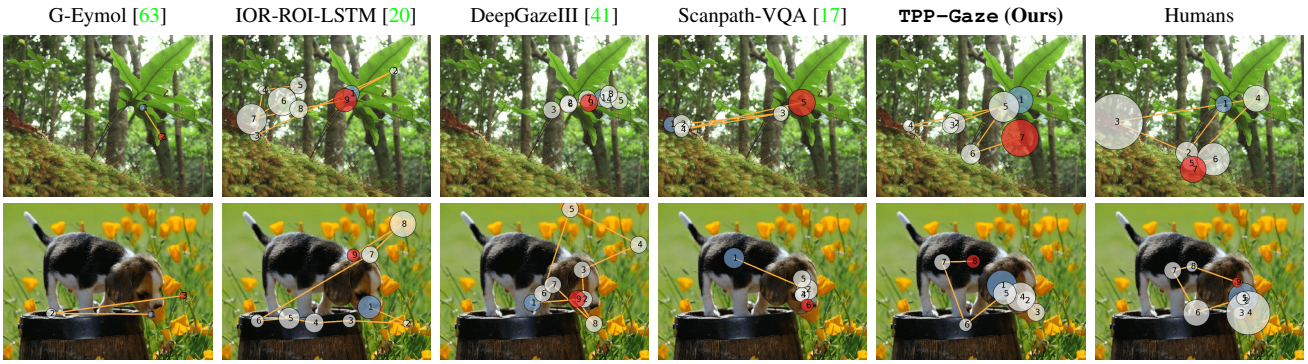

Figure 3. Comparison of simulated and human scanpaths. Each circle represents a fixation point, with its diameter proportional to the fixation duration. For methods that do not model fixation duration, circles are shown with a uniform size.

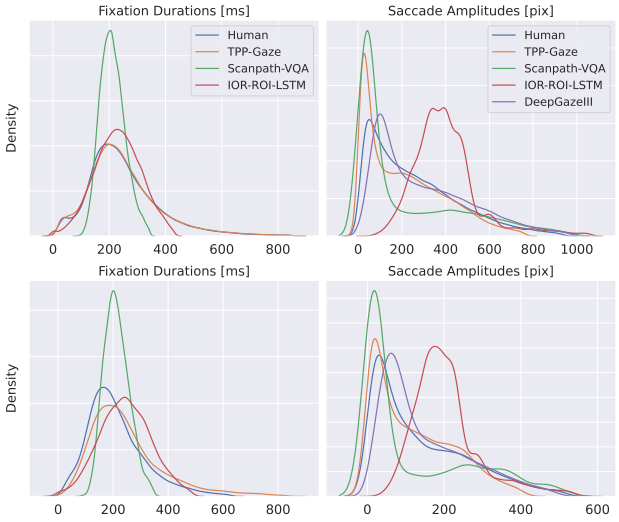

Figure 4. Statistical properties exhibited by TPP-Gaze and other methods relative to those of human observers, in terms of empirical fixation durations and saccade amplitudes on the COCO-FreeView (top row) and OSIE (bottom row) datasets.

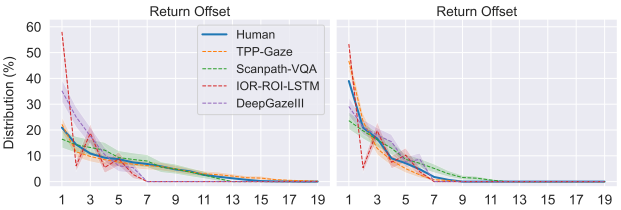

Figure 5. Return fixations analysis comparing TPP-Gaze with other methods and human observers. Results are shown on COCO-FreeView (left plot) and OSIE (right plot) datasets.

the fine-tuning stage of Scanpath-VQA based on reinforcement learning [49] have been inhibited. These results are reported in gray color at the bottom of the tables.

As can be observed, when trained under the same settings and datasets, TPP-Gaze (with either GRU or Transformer-based history encoding) outperforms all the other approaches on most of the adopted metrics. Interestingly, in many cases the proposed approach offers the best overall performance, even when considering the pre-trained models released by the authors, except for ScanMatch where Scanpath-VQA, which is directly optimized via reinforcement learning on this metric, understandably proves to be the best. Some qualitative results are shown in Fig. 3, where we report sampled scanpaths from five models alongside those from humans. Notably, TPP-Gaze can predict fixations that better align with those recorded from human subjects, confirming the advantages of the proposed approach for predicting scanpaths during free-viewing.

Additional analyses are reported in Fig. 4 that shows empirical distributions summarizing TPP-Gaze's scanpath statistics compared to those yielded by human observers and other methods. Beyond common scanpath statistics, we further evaluate the proposed approach using a return fixations (RF) analysis [64]. RF analysis describes the tendency of observers (either human or simulated) to revisit previously foveated locations. The frequency of RFs and the temporal offset (*i.e.*, the number of intervening fixations before returning to a location) at which they occur, provide a more nuanced description of the cognitive processes underlying attention allocation [64]. Fig. 5 reports the results of this analysis in comparison with existing methods across two datasets. Notably, although TPP-Gaze was not explicitly trained for this objective, it produces the most accurate RF patterns with respect to human behavior when compared to state-of-the-art approaches[2].

4.3. Applications

Saliency Prediction. The performance of TPP-Gaze are further evaluated by comparing the saliency maps "backward" generated from fixations with those of human observers across all evaluated scanpath models. The results, presented in Table 4, are measured using three commonly adopted saliency metrics [12, 13, 21]: Kullback-Leibler Divergence (KL-Div), Judd's Area Under the Curve (AUC), and Normalised Scanpath Saliency (NSS). DeepGazeIII and CLE (DG) are reported here only as references for the per-

[2]Results of the RF analysis on OSIE, NUSEF and FiFa are reported in the supplementary material.

	COCO-FreeView			MIT1003			OSIE			NUSEF			FiFa		
	KL-Div ↓	AUC ↑	NSS ↑	KL-Div ↓	AUC ↑	NSS ↑	KL-Div ↓	AUC ↑	NSS ↑	KL-Div ↓	AUC ↑	NSS ↑	KL-Div ↓	AUC ↑	NSS ↑
Saliency-based															
CLE (DG) [8, 42]	8.65	0.55	0.09	-	-	-	5.08	0.59	0.28	4.99	0.63	0.38	6.39	0.59	0.25
DeepGazeIII [41]	0.85	0.84	1.75	-	-	-	0.32	0.87	2.01	0.49	0.85	1.89	0.62	0.88	2.52
Saliency-free															
Itti-Koch [35]	8.94	0.56	0.24	5.01	0.64	0.47	3.35	0.65	0.51	4.84	0.63	0.40	5.47	0.64	0.42
CLE (Itti) [8, 35]	7.45	0.54	0.07	4.15	0.61	0.23	3.45	0.61	0.23	3.36	0.63	0.31	4.84	0.60	0.23
G-Eymol [63]	10.98	0.56	0.26	7.64	0.62	0.35	4.58	0.67	0.60	5.09	0.66	0.55	9.04	0.62	0.47
IOR-ROI-LSTM [20]	1.30	0.77	0.99	**0.78**	0.81	1.40	**0.50**	0.83	1.46	**0.74**	0.80	1.32	**0.83**	0.85	1.72
Scanpath-VQA [17]	3.53	0.77	1.56	2.12	0.82	2.01	1.26	**0.84**	**2.12**	2.45	0.80	**1.76**	1.88	**0.86**	**2.89**
TPP-Gaze (GRU)	**1.01**	**0.84**	**1.65**	**0.78**	**0.86**	**2.06**	0.67	**0.84**	1.72	0.84	**0.84**	1.71	1.07	**0.86**	2.06
TPP-Gaze (Transformer)	1.11	0.83	1.54	0.83	0.85	1.93	0.68	**0.84**	1.68	0.79	**0.84**	1.70	1.11	0.85	1.91

Table 4. Saliency prediction results on COCO-FreeView, MIT1003, OSIE, NUSEF, and FiFa datasets. Models are grouped into saliency-based and saliency-free methods, where the former (*i.e.*, CLE (DG) and DeepGazeIII) incorporate components trained to predict saliency maps. **Bold** values represent the best performance within each metric, while underline values indicate the second-best results.

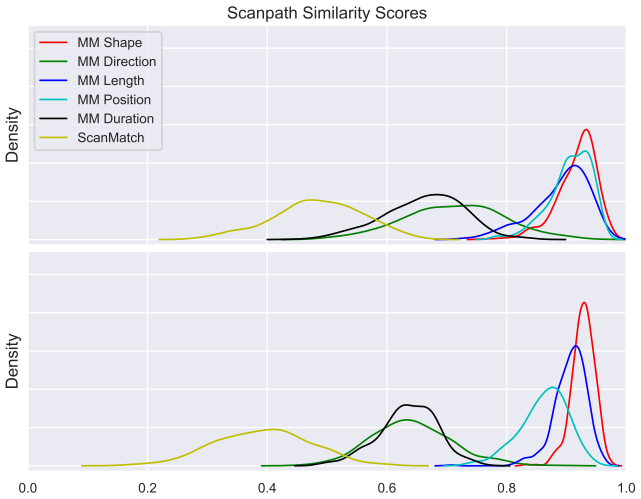

Figure 6. Empirical distributions of the adopted metrics quantifying inter-humans (top) and human *vs.* TPP-Gaze (bottom) scanpath similarity for the visual search task on COCO-Search18.

formance of a saliency prediction model, given their adoption of an extensive pre-training phase designed expressly for saliency generation (DeepGazeIII), or explicit adoption of a saliency model (CLE (DG)). Overall, TPP-Gaze obtains the best or second-best performance across all metrics and datasets. It yields results that are comparable to or surpass those of IOR-ROI-LSTM [20] and Scanpath-VQA [17], which are significantly better than all other approaches. This further demonstrates the effectiveness of our approach in predicting fixation points that better resemble human scanpaths than those predicted by existing methods.

Extending the Model to Visual Search Tasks. Recently, several works [17, 26, 47, 60, 65] have focused on predicting attention allocation on specific targets (visual search tasks). Although TPP-Gaze was originally devised and evaluated for the free-viewing scenario, it can be extended to tackle the visual search problem in various ways. Here, we propose a proof-of-concept model featuring a simple architectural variation that enables goal-directed attention prediction with TPP-Gaze. In a nutshell, we use RoBERTa [44] to perform a linguistic embedding of the search target and learn a target-oriented image semantic representation[3]. (*cf.* Sec. 3.1). Preliminary results show encouraging trends on the COCO-Search18 dataset [19, 59], as illustrated in Fig. 6, where visual search patterns produced by TPP-Gaze are compared to human patterns using the MM and SM metrics. A qualitative example is depicted in Fig. 7.

Figure 7. Human (left) and simulated (right) scanpaths for the visual search task. Search objective is "Sink".

5. Conclusion

We presented TPP-Gaze, a novel approach that explicitly models the evolution of visual attention through scanpaths using Neural Temporal Point Processes. TPP-Gaze enables principled modelling of both fixation positions and their durations. Extensive experiments conducted on five publicly available datasets demonstrate the effectiveness of the proposed approach in capturing gaze spatio-temporal dynamics, as reflected in state-of-the-art performance in scanpath similarity and fixation duration prediction. Additionally, it demonstrates human-like return fixation patterns and achieves competitive results in saliency prediction and task-driven attention allocation.

Acknowledgments

This work was supported by a grant from Università degli Studi di Milano (Bando Linea 3 My First SEED – DM 737/2021 MUR) and by the PNRR project "Italian Strengthening of Esfri RI Resilience (ITSERR)" funded by the European Union - NextGenerationEU (CUP B53C22001770006).

[3] More details and simulations are shown in the supplementary material

References

[1] Nicola C Anderson, Fraser Anderson, Alan Kingstone, and Walter F Bischof. A comparison of scanpath comparison methods. *Behavior Research Methods*, 47(4):1377–1392, 2015. 5

[2] Marc Assens, Xavier Giro i Nieto, Kevin McGuinness, and Noel E. O'Connor. PathGAN: Visual Scanpath Prediction with Generative Adversarial Networks. In *ECCV Workshops*, 2018. 3

[3] Marc Assens Reina, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. SaltiNet: Scan-Path Prediction on 360 Degree Images Using Saliency Volumes. In *ICCV Workshops*, 2017. 3

[4] Emmanuel Bacry, Iacopo Mastromatteo, and Jean-François Muzy. Hawkes processes in finance. *Market Microstructure and Liquidity*, 1(01):1550005, 2015. 3

[5] Simon Barthelmé, Hans Trukenbrod, Ralf Engbert, and Felix Wichmann. Modeling fixation locations using spatial point processes. *Journal of Vision*, 13(12):1–1, 2013. 2

[6] Giuseppe Boccignone, Vittorio Cuculo, and Alessandro D'Amelio. Problems with Saliency Maps. In *ICIAP*. Springer International Publishing, 2019. 1

[7] Giuseppe Boccignone, Vittorio Cuculo, Alessandro D'Amelio, Giuliano Grossi, and Raffaella Lanzarotti. On gaze deployment to audio-visual cues of social interactions. *IEEE Access*, 8:161630–161654, 2020. 2, 3

[8] Giuseppe Boccignone and Mario Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1-2):207–218, 2004. 6, 8

[9] Rafal Bogacz, Eric Brown, Jeff Moehlis, Philip Holmes, and Jonathan D Cohen. The physics of optimal decision making: a formal analysis of models of performance in two-alternative forced-choice tasks. *Psychological Review*, 113(4):700, 2006. 3

[10] Stephan A Brandt and Lawrence W Stark. Spontaneous eye movements during visual imagery reflect the content of the visual scene. *Journal of Cognitive Neuroscience*, 9(1):27–38, 1997. 1, 5

[11] Andrew Bray and Frederic Paik Schoenberg. Assessment of point process models for earthquake forecasting. *Statistical Science*, 28(4):510–520, 2013. 3

[12] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. MIT Saliency Benchmark. http://saliency.mit.edu/. 7

[13] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What Do Different Evaluation Metrics Tell Us About Saliency Models? *IEEE Trans. PAMI*, 41(3):740–757, 2019. 7

[14] Giuseppe Cartella, Marcella Cornia, Vittorio Cuculo, Alessandro D'Amelio, Dario Zanca, Giuseppe Boccignone, and Rita Cucchiara. Trends, Applications, and Challenges in Human Attention Modelling. In *IJCAI*, 2024. 1

[15] Giuseppe Cartella, Vittorio Cuculo, Marcella Cornia, and Rita Cucchiara. Unveiling the Truth: Exploring Human Gaze Patterns in Fake Images. *IEEE Signal Processing Letters*, 31:820–824, 2024. 1

[16] M. Cerf, J. Harel, W. Einhäuser, and C. Koch. Predicting human gaze using low-level saliency combined with face detection. In *NeurIPS*, 2008. 5

[17] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting Human Scanpaths in Visual Question Answering. In *CVPR*, 2021. 2, 3, 6, 7, 8

[18] Xianyu Chen, Ming Jiang, and Qi Zhao. Beyond average: Individualized visual scanpath prediction. In *CVPR*, 2024. 2, 3

[19] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Scientific Reports*, 11(1):1–11, 2021. 8

[20] Zhenzhong Chen and Wanjie Sun. Scanpath Prediction for Visual Attention using IOR-ROI LSTM. In *IJCAI*, 2018. 2, 3, 5, 6, 7, 8

[21] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. SAM: Pushing the Limits of Saliency Prediction Models. In *CVPR Workshops*, 2018. 7

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

[23] Daryl J Daley and David Vere-Jones. *An introduction to the theory of point processes: volume II: general theory and structure*. Springer Science & Business Media, 2007. 3

[24] Alessandro D'Amelio and Giuseppe Boccignone. Gazing at social interactions between foraging and decision theory. *Frontiers in Neurorobotics*, 15:639999, 2021. 2, 3

[25] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behavior Research Methods*, 44(4):1079–1100, 2012. 5

[26] Zhiwei Ding, Xuezhe Ren, Erwan David, Melissa Vo, Gabriel Kreiman, and Mengmi Zhang. Efficient Zero-shot Visual Search via Target and Context-aware Transformer. *arXiv preprint arXiv:2211.13470*, 2022. 8

[27] Nan Du, Hanjun Dai, Rakshit Trivedi, Utkarsh Upadhyay, Manuel Gomez-Rodriguez, and Le Song. Recurrent marked temporal point processes: Embedding event history to vector. In *ACM SIGKDD*, 2016. 2, 4

[28] Ralf Engbert, Hans A Trukenbrod, Simon Barthelmé, and Felix A Wichmann. Spatial Statistics and Attentional Dynamics in Scene Viewing. *Journal of Vision*, 15(1):14–14, 2015. 2

[29] Joseph Enguehard, Dan Busbridge, Adam Bozson, Claire Woodcock, and Nils Hammerla. Neural temporal point processes for modelling electronic health records. In *Machine Learning for Health*, 2020. 3

[30] Alan G Hawkes. Spectra of some self-exciting and mutually exciting point processes. *Biometrika*, 58(1):83–90, 1971. 2

[31] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[32] Gao Huang, Zhuang Liu, Laurens Van Der Maaten, and Kilian Q Weinberger. Densely connected convolutional networks. In *CVPR*, 2017. 4, 6

[33] Hengguan Huang, Hao Wang, and Brian Mak. Recurrent poisson process unit for speech recognition. In *AAAI*, 2019. 2

[34] Valerie Isham and Mark Westcott. A self-correcting point process. *Stochastic Processes and Their Applications*, 8(3):335–347, 1979. 2

[35] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. PAMI*, 20:1254–1259, 1998. 1, 6, 8

[36] Halszka Jarodzka, Kenneth Holmqvist, and Marcus Nyström. A Vector-based, Multidimensional Scanpath Similarity Measure. In *ETRA*, 2010. 5

[37] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009. 4

[38] Srijan Kumar, Xikun Zhang, and Jure Leskovec. Predicting dynamic embedding trajectory in temporal interaction networks. In *ACM SIGKDD*, 2019. 3

[39] Matthias Kümmerer and Matthias Bethge. State-of-the-Art in Human Scanpath Prediction. *arXiv preprint arXiv:2102.12239*, 2021. 1, 5, 6

[40] Matthias Kümmerer and Matthias Bethge. Predicting visual fixations. *Annual Review of Vision Science*, 9, 2023. 1, 5, 6

[41] Matthias Kümmerer, Matthias Bethge, and Thomas SA Wallis. DeepGaze III: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022. 3, 4, 5, 6, 7, 8

[42] Matthias Kümmerer, Lucas Theis, and Matthias Bethge. Deep Gaze I: Boosting saliency prediction with feature maps trained on ImageNet. *arXiv preprint arXiv:1411.1045*, 2014. 6, 8

[43] Rosanne Liu, Joel Lehman, Piero Molino, Felipe Petroski Such, Eric Frank, Alex Sergeev, and Jason Yosinski. An intriguing failing of convolutional neural networks and the coordconv solution. In *NeurIPS*, 2018. 4

[44] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*, 2019. 8

[45] Daniel Martin, Ana Serrano, Alexander W Bergman, Gordon Wetzstein, and Belen Masia. ScanGAN360: A Generative Model of Realistic Scanpaths for 360° Images. *IEEE Trans. VCG*, 28(5):2003–2013, 2022. 4

[46] Hongyuan Mei and Jason M Eisner. The neural hawkes process: A neurally self-modulating multivariate point process. In *NeurIPS*, 2017. 2

[47] Sounak Mondal, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Gregory Zelinsky, and Minh Hoai. GazeFormer: Scalable, Effective and Fast Prediction of Goal-Directed Human Attention. In *CVPR*, 2023. 2, 3, 8

[48] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. In *ECCV*, 2010. 5

[49] Steven J Rennie, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. Self-Critical Sequence Training for Image Captioning. In *CVPR*, 2017. 7

[50] Oleksandr Shchur, Marin Biloš, and Stephan Günnemann. Intensity-Free Learning of Temporal Point Processes. In *ICLR*, 2020. 2, 3, 4

[51] Oleksandr Shchur, Ali Caner Türkmen, Tim Januschowski, and Stephan Günnemann. Neural Temporal Point Processes: A Review. In *IJCAI*, 2021. 3, 4

[52] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. ScanDMM: A Deep Markov Model of Scanpath Prediction for 360○ Images. In *CVPR*, 2023. 3, 4

[53] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual Scanpath Prediction Using IOR-ROI Recurrent Mixture Density Network. *IEEE Trans. PAMI*, 43(6):2101–2118, 2019. 2, 3

[54] Benjamin W Tatler, James R Brockmole, and Roger HS Carpenter. Latest: A model of saccadic decisions in space and time. *Psychological Review*, 124(3):267, 2017. 1, 2, 3

[55] Rakshit Trivedi, Mehrdad Farajtabar, Prasenjeet Biswal, and Hongyuan Zha. Dyrep: Learning representations over dynamic graphs. In *ICLR*, 2019. 3

[56] R.J. van Beers. The sources of variability in saccadic eye movements. *The Journal of Neuroscience*, 27(33):8757–8770, 2007. 3

[57] Shuai Xiao, Junchi Yan, Xiaokang Yang, Hongyuan Zha, and Stephen Chu. Modeling the intensity function of point process via recurrent neural networks. In *AAAI*, 2017. 4

[58] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of Vision*, 14(1):28–28, 2014. 4

[59] Zhibo Yang, Lihan Huang, Yupei Chen, Zijun Wei, Seoyoung Ahn, Gregory Zelinsky, Dimitris Samaras, and Minh Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *CVPR*, June 2020. 5, 8

[60] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent Human Attention. In *ECCV*, 2022. 8

[61] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Predicting Human Attention using Computational Attention. *arXiv preprint arXiv:2303.09383*, 2023. 4

[62] Anna-Kaisa Ylitalo. *Statistical inference for eye movement sequences using spatial and spatio-temporal point processes*. PhD thesis, University of Jyväskylä, 2017. 2

[63] Dario Zanca, Stefano Melacci, and Marco Gori. Gravitational laws of focus of attention. *IEEE Trans. PAMI*, 42(12):2983–2995, 2020. 6, 7, 8

[64] Mengmi Zhang, Marcelo Armendariz, Will Xiao, Olivia Rose, Katarina Bendtz, Margaret Livingstone, Carlos Ponce, and Gabriel Kreiman. Look twice: A generalist computational model predicts return fixations across tasks and species. *PLoS Computational Biology*, 18(11):1–38, 2022. 7

[65] Mengmi Zhang, Jiashi Feng, Keng Teck Ma, Joo Hwee Lim, Qi Zhao, and Gabriel Kreiman. Finding any Waldo with zero-shot invariant and efficient visual search. *Nature Communications*, 9(1), 2018. 8

[66] Qiang Zhang, Aldo Lipani, Omer Kirnap, and Emine Yilmaz. Self-attentive hawkes process. In *ICML*, 2020. 2, 4

[67] Simiao Zuo, Haoming Jiang, Zichong Li, Tuo Zhao, and Hongyuan Zha. Transformer hawkes process. In *ICML*, 2020. 2, 4