Visual ScanPath Transformer: Guiding Computers to See the World

Mengyu Qiu[1] Quan Rong[1] Dong Liang[1] Huawei Tu[2]
[1] College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China
[2] Department of Computer Science and Information Technology, La Trobe University, Melbourne, Australia
qmengyu@nuaa.edu.cn; rongquan0806@gmail.com; liangdong@nuaa.edu.cn; h.tu@latrobe.edu.au

Figure 1: Human scanpaths are composed of a series of fixations and saccades. Visual inputs at the fovea are processed in **high resolution** during fixations, while peripheral vision is correspondingly blurred to guide saccades. A scanpath prediction model simulates human saccadic decisions by predicting the priority probability map of the next fixation.

Abstract

We propose to exploit the scanpath prediction technology to simulate human visual system to automatically generate gaze scanpaths for VR/AR applications, to alleviate the equipment and computational cost in foveated rendering. Specifically, we propose a novel deep learning-based scanpath prediction model called Visual ScanPath Transformer (VSPT), to predict human gaze scanpaths in both free viewing and task-driven viewing situations, based on which the VR/AR systems can execute foveated rendering rapidly and cheaply. The proposed VSPT first extracts highly task-related image features from the visual scene, and then explores the global dependency relationships among all the image regions to generate each image region a global feature. Next, VSPT simulates the human visual working memory to consider all the previous fixations' influences when predicting each fixation. Experimental findings confirm that our model exhibits adherence to classical visual principles during saccadic decision-making, surpassing the current state-of-the-art performance in free-viewing and task-driven (goal-driven and question-driven) visual scenarios.

Index Terms: Human-centered computing—Human computer interaction (HCI)—Interaction techniques—Pointing; Computing methodologies—Artificial intelligence—Computer vision—Scene understanding

Dong Liang and Huawei Tu are the corresponding authors of this work. Mengyu Qiu and Rong Quan contributed equally to this work. This work is supported in part by the National Natural Science Foundation of China under grants 62206127 and 62272229, the Natural Science Foundation of Jiangsu Province under grant BK20222012, and the National Key Research and Development Program of China (No. 2021ZD0113200).

1 Introduction

Virtual Reality (VR) and Augmented Reality (AR) technologies integrate the digital world with reality through immersive experiences [39], which have been developed rapidly and brought many new ways of entertainment and industrial applications in recent years. The research on eye gaze movements, such as gaze estimation, plays a very crucial role in VR/AR applications [22]. On the one hand, gaze estimation can accurately track the users' visual behavior by analyzing their eye movements and gaze directions, enabling natural and intuitive human-computer interaction(HCI) in VR/AR [8,13,31,43]. On the other hand, gaze estimation can predict accurate gaze locations in real-time, based on which the VR/AR system can execute a more accurate foveated rendering, resulting in a more accurate rendering result with a lower computational burden. Furthermore, by analyzing gaze data, the VR/AR systems can understand and predict user attention and intent [38], subsequently providing personalized recommendations [6, 26] and optimizing interface design [29], ultimately enhancing user experiences.

Based on the descriptions above, accurate prediction of users' eye fixations and scanpaths is very important in VR and AR applications. However, traditional methods typically rely on additional cameras, electroencephalograms (EEG), and other sensors to obtain information about the human eye and face, and compute each moment's fixation based on the collected information, which is computationally heavy. In recent years, research in computer vision has made remarkable progress in human eye fixation and scanpath prediction. The human eye scanpath describes the sequence of human eye fixation when observing a visual scene. Many studies have demonstrated certain commonalities among different human visual scanpaths [20, 21, 50]. By training on the large collected scanpath datasets, the scanpath prediction models can extract these commonalities and predict the users' scanpaths accurately. Consequently, the VR/AR applications can generate a foveated rendering result in advance based on the scanpath prediction result, based on which

the users can preliminarily view the virtual environment. And If the users want to see additional regions beyond the fixations, they can activate the eye tracking system to capture their gaze location. Such an arrangement can dramatically reduce the amount of computation.

Based on our observation, people typically observe visual scenes in two different states: **free viewing** and **task-driven viewing**. Task-driven viewing can be further categorized into two types: goal-driven viewing and question-driven viewing. To investigate human visual scanpath prediction in different contexts, we have focused on three specific scenarios. Firstly, we have studied visual scanpaths under free viewing, where the observer is free to observe a visual scene without a specific task. In this situation, the observer's attention is always attracted by the salient regions in the scene. Secondly, we have examined goal-driven scanpath prediction, where the observer views the visual scene to search for a target. In this case, the observer's visual scanpath will be influenced by the target, following a particular pattern and order for searching. Lastly, we have discussed question-driven scanpath prediction, where the observer views the visual scene with a specific question in mind, expecting to find an answer. This task needs the observer to focus on question-related visual elements and thus forms a question-oriented visual scanpath.

In this paper, we propose a novel scanpath prediction model, dubbed the Visual ScanPath Transformer (VSPT), to predict human scanpaths in both free-viewing and task-driven viewing situations. VSPT formulates scanpath prediction as a sequential decision process by generating each fixation based on both the original image and the previous fixations. This simulates the human visual working mechanism, which considers the history of fixations and the visual information available at each time step. As shown in Fig. 2, VSPT consists of four main components: a saliency feature extraction module, a visual encoder module, a fixation decoder module, and a fixation generator module. The saliency feature extraction module extracts highly task-relevant visual features from the original image based on a static saliency map (in a free-viewing situation) or task guidance map (in a task-driven viewing situation). The visual encoder module then utilizes a Transformer structure to explore the global dependency relationships among all image regions, generating a global feature representation for each region. The fixation decoder module predicts each time step's fixation embedding by simulating the visual working memory. It first learns the influences of the historical fixations on current fixation, and then predicts the embedding of current fixation based on both the global feature representations and the influences. Finally, the fixation generator module predicts fixation coordinates from the fixation embeddings. Compared with previous works, our approach is heuristic-free, eliminating reliance on visual rules widely referenced in other model designs, significantly simplifying the scanpath prediction workflow and overall model architecture. Experimental results demonstrate that our model outperforms the current state-of-the-art scanpath prediction methods in free-viewing and task-driven visual scenarios and generates accurate human scanpaths for VR/AR systems.

In summary, the main contributions of this work are as follows:

- We propose to exploit the scanpath prediction technology in the computer vision field to automatically predict human-like scanpaths for each image scene in VR/AR applications, which can significantly reduce the equipment and computational cost of foveated rendering.
- We propose a heuristic-free visual scanpath predictor that can accurately predict human scanpaths under both free viewing and task-driven viewing conditions. Our approach involves extracting highly task-related features, analyzing the global long-range dependency relationship among all image regions using a Transformer structure, and simulating human visual working memory to generate fixations at each time step. By utilizing these techniques, our proposed predictor can simulate the human visual system more vividly and accurately.
- The proposed method is comprehensively evaluated on four eye-tracking datasets of free-viewing scenes, as well as a visual search dataset, and a visual question-answering dataset. It consistently achieves state-of-the-art performance, demonstrating our approach's robustness and generalization capability.

2 Related Work

The key to visual scanpath prediction is to simulate how HVS handles visual scenes. However, this is a highly challenging task due to the complexity of HVS. Research on this task primarily focuses on two scenarios: free viewing and task-driven viewing.

Free viewing: Early methods mainly generate scanpaths by executing some well-acknowledged human visual rules on static saliency maps [18, 19, 45]. Itti *et al.* [19] is the most representative work. They first utilized a dyadic Gaussian pyramid to generate saliency maps and then executed the Winner-Take-All (WTA) and Inhibit-Of-Return (IOR) strategies on the obtained saliency maps to generate scanpaths. However, using a static saliency map throughout the entire scanpath prediction process neglects the dynamic temporal relationship among fixations, leading to significant discrepancies between predicted and actual human visual behaviors.

Subsequent methods attempt to model the dynamical temporal relationships among fixations [25,28,40,42,46,47]. Wang *et al.* [46] integrated three human attention-driven factors named reference sensory responses, fovea-periphery resolution discrepancy, and visual working memory into a residual perceptual information map, from which new fixation is selected according to the information maximization principle. Sun *et al.* [42] exploited projection pursuit of conducting Super Gaussian Component (SGC) analysis sequentially and selected new fixation as the location with maximum SGC response. Wang *et al.* [47] used a foveated image to simulate retinal imaging and generated fixations by jointly considering foveated saliency map, a saccadic bias of gaze shift, and IOR mechanism in short-term memory. Le Meur *et al.* [25] first conducted spatial statistics of the gaze patterns from the collected human scanpaths, then developed a dynamical scanpath prediction model by integrating the obtained statistical conclusion, bottom-up saliency, and IOR mechanism. Xia *et al.* [49] conducted a deep autoencoder to form the representation from surrounding patches to central ones, based on which the perceptual residual is to guide the fixation generation. Bao *et al.* [4] integrated foveal vision and inhibition of return with deep convolutional neural networks into a recurrent model to predict human scanpaths. These methods formulate human scanpath prediction as an iterative process, constantly predicting the location of the subsequent fixation based solely on the information from the adjacent one, resulting in incomplete modeling of the temporal relationships between fixations.

Some researchers use Recurrent Neural Networks (RNNs) to model the sequential mechanism of HVS. Ngo and Manjunath [33] proposed the first RNN-based scanpath prediction model. Sun *et al.* proposed an Inhibition of Return - Region of Interest (IOR-ROI) framework to predict scanpaths, where a dual LSTM unit containing an IOR-LSTM and ROI-LSTM is constructed. Chen *et al.* [10] proposed a deep reinforcement learning-based scanpath prediction framework in visual question answering. They also exploited Conv-LSTMs as the skeleton module to model HVS.

Task-driven viewing: Compared to free viewing scanpath prediction, research on task-driven scanpaths is not yet sufficient. Common task-driven scanpaths currently include visual search and visual question-answering modes, which require searching for specific information in visual scenes to complete tasks. The task of visual search [48] is to find a target in a scene. This task requires purposeful scene scanning to identify features or patterns that match the target object. The Microwave-Clock-Search (MCS) dataset [54] was introduced as an early attempt to study goal-driven attention control in visual search tasks. Recently, a study [52] applied inverse

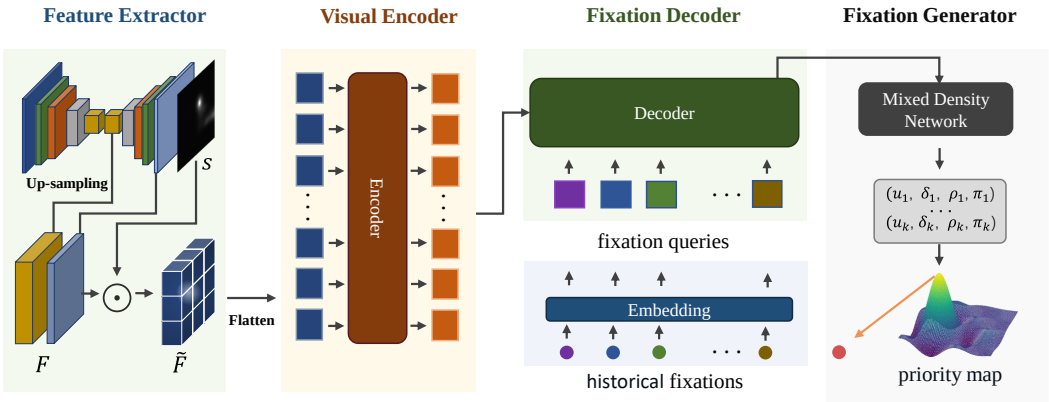

Figure 2: Architecture of Visual Scanpath Transformer. The proposed model consists of four main components: a Feature Extractor module, a Visual Encoder module, a Fiaxtion Decoder module, and a Fixation Generator module.

reinforcement learning to the scenario of visual search in scanpath and proposed a large-scale dataset of search fixations containing 18 target categories (COCO-Search18 [11]). Chen and Yang *et al.* [12, 53] compared search behaviors under target present and target-absent conditions, revealing weaker target guidance signals in target-absent searches, and proposed a visual stopping criterion prediction model based on gaze history and subject features. Hu *et al.* [17] proposed FixationNet, a novel learning-based model for forecasting human eye fixations in task-oriented virtual environments. Mondal *et al.* [30] proposed a Gazeformer model also based on a transformer encoder and decoder architecture. The differences between our model and Gazeformer are distinct. Firstly, we abstract task-guidance information into attention maps, and use the Transformer encoder to model their global relationships with the visual features, while Gazeformer encodes the search target as a feature embedding and directly concatenates it with visual features. Secondly, we use a mixture density network to generate a multimodal probability map for each fixation and an autoregressive mode to decode the fixation sequence, while Gazeformer uses a randomly initialized fixation query and outputs the entire scanpath at once.

Visual Question Answering (VQA) requires finding answers to scene-related questions while viewing the scene. Chen *et al.* [10] presents a framework for predicting scanpath in the context of visual question answering. The framework is based on deep reinforcement learning and addresses exposure bias in scanpath prediction through self-critical sequence training. It also introduces a consistency-divergence loss to generate a distinguishable scanpath between correct and incorrect answers. The framework performs well in both free-viewing and visual search scenarios.

3 Method

We propose a VSPT model for predicting scanpaths in free-viewing and task-driven visual exploration scenarios. Fig. 2 illustrates the overall framework of our proposed method. It consists of four components: a feature extractor, a visual encoder, a fixation decoder, and a fixation generator.

3.1 Feature Extractor

The input visual scene image is first passed through a feature extractor for encoding to obtain a saliency feature representation highly relevant to the scanpath prediction task. Specifically, we use the saliency prediction network SalGAN [34], which has been pre-trained on an eye-tracking dataset and internally adopts a convolutional encoder-decoder architecture, where the encoder is the same as VGG-16 [37] without the final pooling and fully connected layers. The decoder is similar to the encoder but with the layer order reversed, and upsampling layers replace pooling layers.

We first resize the input $\mathbf{I}$ to 192×256 and feed it into SalGAN, obtaining layers of feature maps from its convolutional decoder. To fuse visual features from different levels, we extract the feature maps generated by the various convolutional layers of the decoder, upsample all feature maps to the original size of the input image, and concatenate them to obtain the final feature map $\mathbf{F} \in \mathbb{R}^{C \times H_0 \times W_0}$, with typical values $C = 576$, $H_0 = 192$, and $W_0 = 256$.

To emulate the hierarchical attention mechanisms inherent within the human visual system more accurately, we impose constraints on the extracted visual features by employing low-level saliency maps, which capture visually prominent stimuli, in conjunction with high-level context-guided maps that account for semantic context. This can be manifested in two distinct forms:

Free-viewing mode: In the context of free-viewing scenes, the human visual exploration process is stimulus-driven, characterized by a bottom-up attention mechanism. We employ saliency maps representing static visual stimulus representations to perform spatial attention operations on the acquired visual features. This guides the model to focus on important regions during the decoding phase for fixation. The specific definition of this operation is:

$$\tilde{\mathbf{F}} = \mathbf{F} \odot \mathbf{S} \tag{1}$$

Here, $\mathbf{S}$ denotes the saliency map $\mathbf{S} \in \mathbb{R}^{1 \times H_0 \times W_0}$ output by the SalGAN model, and $\odot$ represents the hadamard product. Through element-wise multiplication, we obtain the spatial-wise attention re-weighted image feature maps $\tilde{\mathbf{F}}$.

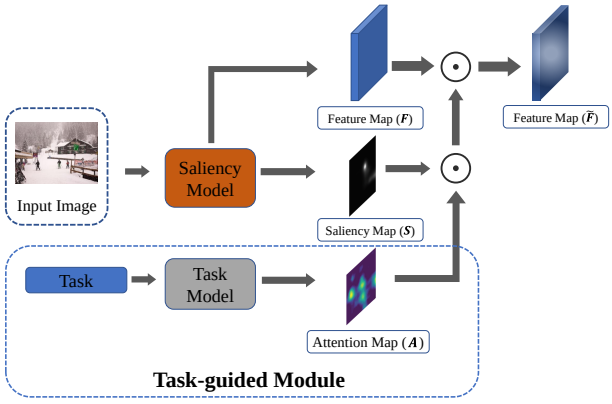

Figure 3: The structure of the task-guided module, where the task description in visual and language forms is abstracted as the task input. The $\odot$ represents the Hadamard product.

Task-driven mode: In the simulation of task-driven visual exploration processes, we employ a task-guided module for integrating semantic and visual content, expressing the influence of tasks through task guidance maps, thereby highlighting image regions associated with the task. In this mode, the control process of human visual attention is co-influenced by low-level visual stimuli and high-level contextual constraints. As shown in Fig. 3, we employ saliency maps and task guidance maps to perform a combined spatial attention operation, which is defined as follows:

$$\tilde{\mathbf{F}} = \mathbf{F} \odot (\mathbf{S} \odot \mathbf{A}) \tag{2}$$

Here, $\mathbf{S}$ represents the saliency map, $\mathbf{A}$ denotes the task guidance map, and $\odot$ signifies the Hadamard product.

An externally pre-trained model introduces task-related guidance information. Specifically, in visual exploration tasks, we use an object detector [55] to detect search targets in the visual search scene, considering regions with detected similar targets as highly task-related. In VQA scenarios, models trained on large VQA datasets can effectively represent the spatial semantics of input questions. Therefore, we adopt the machine attention maps from the VQA model [1] as spatial position guidance to generate task guidance maps. In practice, task guidance maps are processed to have a two-dimensional shape consistent with the input image size and are normalized within the range of [0, 1].

3.2 Visual Encoder

The visual encoder module further models the global dependencies between visual feature regions. First, average pooling is used to downscale the activation map $\tilde{\mathbf{F}}$ to a smaller resolution of 30x40. Subsequently, its spatial dimensions are collapsed to fit the encoder's input, resulting in a set of feature vectors $\{p_1, p_2, ..., p_N\}$, where $N = 1200$. A linear layer is then utilized to map the dimensionality c of the feature vectors to the internal dimension d of the encoder. The encoder follows the standard architecture of the Transformer [44], consisting of a stack of identical layers. Each encoder layer comprises two sub-layers: a multi-head self-attention module and a feed-forward network (FFN), both surrounded by residual connections followed by layer normalization. Before being fed into each encoder layer, the feature vectors are supplemented with fixed positional encodings [35] to allow the self-attention module to exploit positional information.

3.3 Fixation Decoder

We employ an additional fixation decoder module in conjunction with the fixation generator module to decode fixation sequences from the visual feature maps. The scanpath prediction can be considered a standard sequence generation task, and we adopt an autoregressive [16] scheme to predict fixations iteratively. Initially, for the current time step t, we utilize the fixation coordinates z_{t-1} from the previous time step t - 1 to initialize a fixation query q_t, which is responsible for encoding the current region of interest information and is ultimately transformed into the corresponding fixation embedding $\tilde{q}_t$. Specifically, the first fixation query is initialized using the image center as the previous coordinates. In particular, we use an embedding layer to initialize the fixation query, which is defined as:

$$q_t = Embedding\,(z_{t-1}), \quad t = 1, 2, 3 \ldots T \tag{3}$$

where the fixation coordinates are normalized to relative values in the range of [0, 1] according to the image size. The embedding layer uses linear mapping, and T represents the length of the predicted sequence. Next, the fixation query is fed into the decoder, which also comprises a set of identical decoder layers. In each decoder layer, it first performs self-attention with historical fixation queries to integrate the influence of historical fixation, then interacts with the visual features output by the visual encoder through cross-attention to obtain scene information, and finally transforms into fixation embeddings after passing through the feed-forward network. Compared to the visual encoder, the fixation decoder has an additional cross-attention module between each layer's self-attention and feed-forward network modules. Each decoder layer supplements the fixation query with learned positional encoding.

3.4 Fixation Generator

The distribution of fixation in visual scanpaths is often multi-modal, meaning that there are multiple possible fixations. Therefore, we adopt a Mixture Density Network (MDN) to predict the probability distribution of the current fixation. The MDN takes a fixation embedding produced by the fixation Decoder as input and predicts K sets of Gaussian distribution parameters, including the means μ, standard deviations σ, correlations ρ, and mixture weights π. The MDN is built as a 2-layer perceptron containing a hidden layer and a ReLU activation layer, utilizing K Gaussian to model the probability distribution, which can be represented as follows:

$$\left\{\tilde{\mu_t^i}, \tilde{\sigma_t^i}, \tilde{\rho_t^i}, \tilde{\pi_t^i}\right\}_{i=1}^{K} = f_{mdn}\left(\tilde{q_t}; \theta_{mdn}\right) \tag{4}$$

$$\sigma_t^i = \exp\left(\tilde{\sigma}_t^i\right), \rho_t^i = \tanh\left(\tilde{\rho}_t^i\right), \pi_t^i = \frac{\exp\left(\tilde{\pi}_t^i\right)}{\sum_{i=1}^{K} \exp\left(\tilde{\pi}_t^i\right)} \tag{5}$$

Here, $\tilde{q_t}$ is the fixation embedding produced by the fixation decoder at time step t, and θ_{mdn} denotes the weights and biases of the linear layers. Eq. 5 constrains the mixture of Gaussian parameters within a reasonable range.

The K sets of Gaussian distributions jointly generate the final probability map, from which we select the pixel location with the highest probability as the next fixation.

$$\hat{z}_t = \arg\max_{z \in \Omega} \left(\sum_{i=1}^{K} \pi_t^i \mathcal{N}\left(z_t \mid \mu_t^i, \sigma_t^i, \rho_t^i\right) \right) \tag{6}$$

where $\mathcal{N}$ denotes the bivariate normal distribution, $\hat{z}_t$ represents the predicted fixation coordinates.

3.5 Training Objective

The model predicts a fixation sequence of length T for each input image. The probability priority map for each fixation is constructed using the K Gaussian kernel parameters output by the model. We use the actual human fixation location at the corresponding time step for supervised learning, guiding the predicted probability priority map to generate accurate fixation. The loss function employed for training is the negative log-likelihood, defined as follows:

$$\mathcal{L} = -\frac{1}{T} \sum_{t=1}^{T} \log \left(\sum_{i=1}^{K} \pi_t^i \mathcal{N}\left(z_t^* \mid \mu_t^i, \sigma_t^i, \rho_t^i\right) \right) \tag{7}$$

Here, T denotes the length of the predicted fixation sequence, and z_t^* represents the t-th actual human fixation.

Our fixation decoder employs the Transformer architecture, which differs from the RNN-inspired scanpath prediction models. This allows for parallel training of the entire fixation point generation process instead of iterating strictly in temporal order. The saliency feature extractor module of the model is pre-trained for the saliency prediction task, and its parameters are frozen during the training process. We only update the parameters of the visual encoder module, fixation decoder module, and fixation generator module.

Table 1: Comparison between our method and other scanpath prediction models on the SALICON, iSUN, OSIE, and MIT1003 datasets regarding ScanMatch, SS, and MultiMatch. 'Human' refers to human performance. The best prediction results are highlighted in bold, and the second-best results are highlighted in underlined.

Method	SALICON Dataset							iSUN Dataset						
	ScanMatch↑	SS↑	DTW-2D↓	MultiMatch↑ Vector	Direction	Length	Position	ScanMatch↑	SS↑	DTW-2D↓	MultiMatch↑ Vector	Direction	Length	Position
Human	0.2710	0.3303	468.01	0.8778	0.6072	0.8660	0.7707	0.3699	0.4471	413.89	0.9382	0.7482	0.9171	0.8607
Itti *et al.*	0.1946	0.2967	715.65	0.8857	0.6470	0.8555	0.7296	0.1321	0.2508	958.09	0.8852	0.5930	0.8421	0.7006
SGC	0.2084	0.2772	613.15	0.9034	0.6347	0.8987	0.7689	0.1588	0.2531	783.61	0.9159	0.5863	0.8840	0.7589
wang *et al.*	0.2293	0.3209	528.33	0.9308	0.6395	**0.9250**	0.8177	0.2213	0.2975	710.72	0.9014	0.5828	0.8770	0.7975
SaltiNet	0.1540	0.2850	765.17	0.9127	0.6567	0.8999	0.7239	0.1382	0.2757	948.86	0.9089	0.5853	0.8896	0.7272
PathGAN	0.0474	0.1982	1102.00	**0.9415**	0.5697	0.9219	0.5761	0.0325	0.1596	1332.78	**0.9573**	**0.6011**	**0.9411**	0.5458
DeepGazeIII	0.1778	0.3042	655.77	0.9351	0.6635	0.9245	0.7561	0.1526	0.2482	757.09	0.9349	0.5643	0.9289	0.7768
IOR-ROI	0.2732	0.3391	491.75	0.9115	**0.6934**	0.8987	0.8081	0.2357	0.3387	633.88	0.8949	0.5717	0.8791	0.7953
VQA	0.2938	0.3451	475.04	0.9354	0.6288	0.9108	0.8271	0.2541	0.3235	563.40	0.9291	0.5999	0.9088	0.8169
Ours	**0.3131**	**0.3663**	**468.15**	0.9406	0.6458	0.9248	**0.8344**	**0.2861**	**0.3590**	**501.74**	0.9410	0.5768	0.9301	**0.8235**

Method	OSIE Dataset							MIT1003 Dataset						
	ScanMatch↑	SS↑	DTW-2D↓	MultiMatch↑ Vector	Direction	Length	Position	ScanMatch↑	SS↑	DTW-2D↓	MultiMatch↑ Vector	Direction	Length	Position
Human	0.4154	0.4691	576.75	0.9401	0.6961	0.9284	0.8553	0.4016	0.4384	439.34	0.9112	0.7243	0.9051	0.8552
Itti *et al.*	0.2565	0.3318	826.86	0.8907	0.6629	0.8576	0.7526	0.2081	0.3152	826.00	0.8837	0.6874	0.8511	0.7459
SGC	0.2656	0.3181	758.95	0.9263	0.6598	0.9035	0.7754	0.2255	0.3058	702.27	0.9206	0.6572	0.8954	0.7834
wang *et al.*	0.2962	0.3587	674.01	0.9316	0.6792	0.9182	0.8000	0.2982	0.3858	586.41	0.9369	0.6966	**0.9339**	0.8271
SaltiNet	0.1949	0.3042	892.95	0.9155	0.6744	0.8968	0.7262	0.1623	0.2919	859.12	0.9141	0.6986	0.9044	0.7308
PathGAN	0.0600	0.1970	1376.16	0.9425	0.5782	0.9280	0.5846	0.0446	0.1521	1286.01	**0.9425**	0.5802	0.9258	0.5923
DeepGazeIII	0.1887	0.3272	832.38	0.9374	0.6740	0.9237	0.7581	0.1823	0.2728	711.69	0.9288	0.6958	0.9236	0.7858
IOR-ROI	0.3477	0.3922	**573.87**	0.9145	**0.7197**	0.8918	0.8282	0.3178	0.4185	549.23	0.9118	**0.7388**	0.8980	0.8285
VQA	0.3901	0.4279	589.91	0.9434	0.6397	0.9223	0.8437	0.3500	0.4255	589.91	0.9275	0.6443	0.9028	0.8459
Ours	**0.3985**	**0.4359**	581.56	**0.9482**	0.6619	**0.9342**	**0.8478**	**0.3628**	**0.4452**	**505.28**	0.9386	0.6824	0.9289	**0.8615**

4 EXPERIMENTS

4.1 Experimental Settings

Datasets. We conducted scanpath prediction experiments under free-viewing conditions on the SALICON [20], iSUN [51], OSIE [50], and MIT1003 [21] datasets. The training images were taken from the SALICON training set, and iSUN, OSIE, and MIT1003 were used to evaluate the model's performance. SALICON is currently the largest eye fixation dataset, consisting of 10,000 training images, 5,000 validation images, and 5,000 test images. All the eye fixation data was collected through mouse tracking on crowdsourcing platforms, with an average of 60 scanpaths per image. iSUN includes 6,000 training images, 926 validation images, and 2,000 test images. The OSIE dataset contains 700 natural images, each viewed by 15 participants. The MIT1003 dataset consists of 1,003 images, including 779 landscape images and 228 portrait images of varying resolutions, with eye-tracking data collected from 15 participants for all images. We conducted visual search task experiments on the COCO-Search18 dataset [52], which contains 6,202 images, half depicting instances of the specified target objects, while the other half do not, corresponding to standard Target-Present (TP) or Target-Absent (TA) search tasks. For visual question-answering, we conducted experiments on the AiR [9] dataset, with its eye-tracking data collected from 20 participants while answering the questions and associations with the correctness of their answers.

Evaluation Metrics. We used five evaluation criteria to evaluate the predicted scanpath's performance, including ScanMatch [14], Sequence Score(SS) [7], MultiMatch [15], Time-Delay Embedding (TDE) [46] and Dynamic Time Warping(DTW) [23]. ScanMatch uses characters to encode the fixations and represents each scanpath as a string. Then, it used a Needleman-Wunsch [32] algorithm to match two strings and compute their similarity. SS is an improvement over ScanMatch, which clusters all actual human fixations into multiple clusters and uses each cluster center as a character. MultiMatch assesses the similarity between two scanpaths from five aspects, including the saccade's shape, direction, and length and the fixation's position and duration. Since we only predicted the fixations' temporal orders and spatial locations, we evaluated the predicted scanpath only from the aspects of shape, direction, length, and position. TDE first divides each scanpath into pieces of length k. Then, for each piece in the predicted scanpath, TDE computes its minimum distance to the split pieces of the real human scanpaths. TDE uses a Hausdorff Distance (HD) and Mean Minimum Distance (MMD) to represent the maximum and average value of all the above minimum distances. DTW is a distance-based method that first calculates the distance between each pair of elements in the two sequences, and then searches for the best matching with the minimum cumulative distance. "Human" represents the average similarity among all the real human scanpaths of each image. "Human" references the upper bound of the scanpath prediction performance.

Implementation details. We trained our model with the AdamW optimizer [27], setting the initial learning rate to 10^{-4} and using learning rate warm-up and periodic adjustment strategies. The number of epochs for the warm-up was set to 20, and the learning rate decreased by half every 50 epochs. The feature map after SalGAN was average pooled to a size of 30×40 and flattened into $N = 1200$ feature vectors. The visual encoder and fixation decoder both employed 4 stacked layers, where each multi-head attention layer used 8 attention heads of width 128 and $d_k = d_v = 64$ for the attention operation. The hidden layer size was 64 for FFN and 16 for MDN, and the Gaussian kernel $K = 5$ was used for MDN. Our method was developed on pytorch, and conducted on a single RTX 3090 GPU. We calculated the model's FLOPs, Parameters, and inference time, which are 41.37G, 35.29M, and 24 ms/fixation, respectively.

4.2 Comparison in Free-viewing

In the mode of free viewing, we compared our model with eight baseline scanpath prediction models, including Itti *et al.* [19], SGC [42], wang *et al.* [46], SaltiNet [3], PathGAN [2], DeepGazeIII [24], IOR-ROI [41], and VQA [10]. We obtained the predicted scanpaths of those methods by running their public codes. Since each image had multiple human scanpaths collected from different subjects, we evaluated each predicted scanpath's performance by first measuring its similarities with all the human scanpaths, and then averaging all the similarities to obtain a final evaluation.

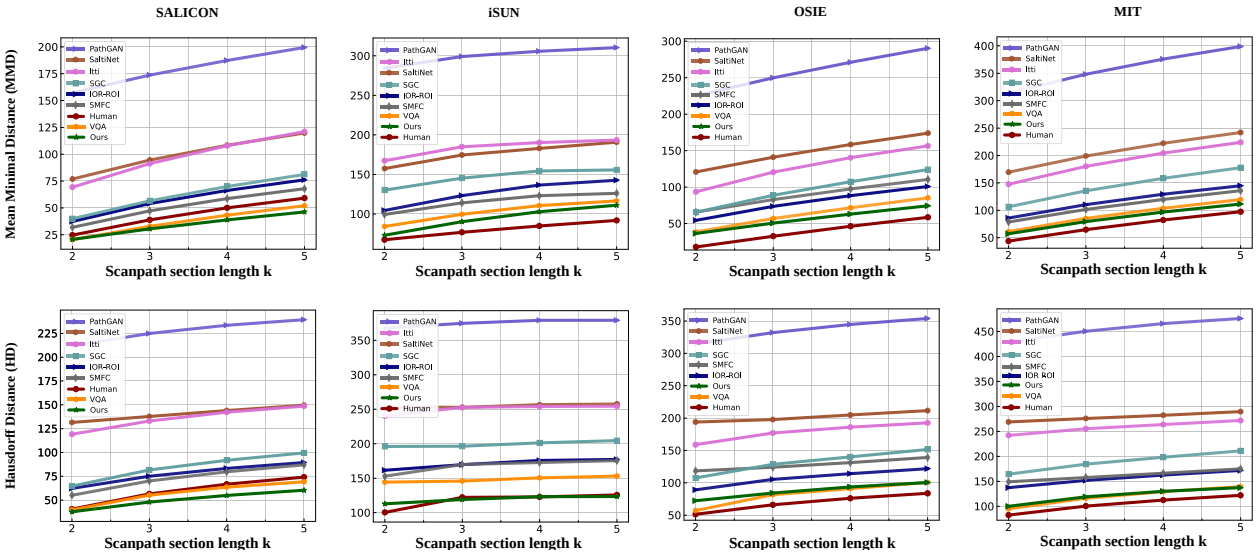

Figure 4: Comparison between our method and other scanpath prediction models on the SALICON, iSUN, OSIE, and MIT1003 datasets regarding TDE. Various methods are distinguished by different color fold lines, 'Human' refers to the inter-observer performance.

Quantitative Evaluation. Table. 1 shows the comparison results in terms of ScanMatch, SS, DTW, and MultiMatch, and Fig. 4 shows the comparison results in terms of TDE. As can be seen, our method achieved better results than the state-of-the-art methods on all four datasets. With the evaluation metrics of ScanMatch and SS, our model outperformed the other models in all four datasets. It also achieves the best results in three datasets and the second-best result in one dataset based on the DTW metric. Our model showed a great advantage in all datasets under the evaluation criteria of TDE. MultiMatch is commonly used to assess the performance of scanpath models in many works. However, we observed that the results of comparing models based on Vector, Direction, and Length of MultiMatch were not consistent with other evaluation criteria. In Fig. 4 and Table 1, the performance of PathGAN was significantly lower than other models in terms of ScanMatch, SS, MultiMatch, and TDE. Qualitative visualizations also revealed that despite high Vector, Direction, and Length of MultiMatch scores, PathGAN exhibited significant differences from the actual human scanpaths. Nonetheless, our method performed very well on most of the MultiMatch metrics.

Qualitative Evaluation. We visualized our scanpath prediction results and qualitatively compared them with the scanpaths generated from other baseline models. The comparison results were shown in Fig. 6, and the number in the lower right corner of each image represents the ScanMatch value of the corresponding scanpaths. As shown in Fig. 6, our model predicted scanpaths with shorter and more evenly distributed saccade amplitudes across different ScanMatch scores, which was consistent with the saccade amplitude biases that Le Meur *et al.* [25] summarized from the existing real human scanpaths. In addition, although some models could obtain relatively high ScanMatch scores, their predicted scanpaths had obvious limitations in visualization. For example, the predicted fixations of Wang *et al.* [46] were often gathered in a certain image region, which was obviously not a good visual exploration route.

4.3 Comparison in Task-driven Viewing

Our method can be easily extended to task-guided scanpath prediction tasks. We validated our model on the AiR dataset for visual question answering and the COCOSearch18 dataset for visual search. Due to the lack of previous work in these scenes, we followed Xianyu *et al.*'s [10] setup and customized some deep learning-based scanpath prediction models (i.e., SaltiNet, PathGAN, and IOR-ROI) as the comparison methods for supplementary scenes. In the visual question answering scene, each question was predicted as a set of correct and incorrect scanpaths, and we calculated similarity scores between the predicted results for correct and incorrect groups and human observation results, respectively. Similar to what was observed in free-viewing scenes, our method also demonstrated excellent performance in visual question answering and visual search tasks. As illustrated in Tab. 2, Tab. 3 and Fig. 5, our approach outperforms the state-of-the-art methods in two task-driven visual scanpath prediction scenarios, with superior results in ScanMatch, SS, and TDE metrics. Moreover, our method achieves the most consistent performance in the MultiMatch metric.

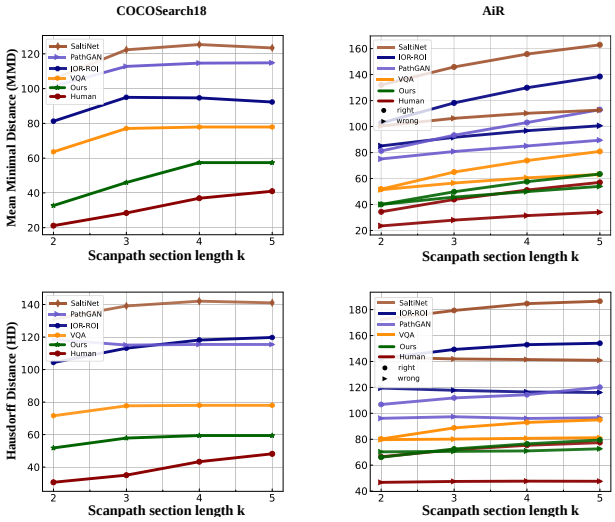

Figure 5: Comparison between our method and other scanpath prediction models on the COCOSearch18, AiR datasets regarding TDE. Various methods are distinguished by different color fold lines, 'Human' refers to the inter-observer performance.

Figure 6: Qualitative evaluation of our model, VQA, IOR-ROI, SaltiNet, PathGAN, Wang *et al.*, SGC, and Itti *et al.* The scores in the lower right represent the ScanMatch scores of the corresponding scanpaths. The ScanMatch scores of our model gradually decrease from top to bottom in the images. The comparison shows that our model exhibits good qualitative performance despite low quantitative scores.

Table 2: Comparison between our method and other scanpath prediction models on the AiR datasets regarding ScanMatch, SS, and MultiMatch. In each panel, the first row indicates the correct scanpaths and the second row indicates the incorrect scanpaths. 'Human' refers to human performance.

Method	ScanMatch↑	SS↑	MultiMatch↑			
			Vector	Direction	Length	Position
Human	0.4165	0.4807	0.9407	0.7411	0.9332	0.8735
	0.3994	0.4611	0.9366	0.7478	0.9280	0.8591
SaltiNet	0.1244	0.2219	**0.9502**	0.6673	**0.9491**	0.6990
	0.1297	0.2211	**0.9503**	0.6825	**0.9388**	0.7002
PathGAN	0.1824	0.2524	0.9442	0.6377	0.9283	0.7693
	0.1839	0.2375	0.9443	0.6323	0.9242	0.7483
IOR-ROI	0.1711	0.2667	0.9396	**0.7475**	0.9326	0.7417
	0.1803	0.2717	0.9389	**0.7590**	0.9318	0.7418
VQA	0.3726	0.4648	0.9324	0.6900	0.9302	0.7647
	0.3532	0.4290	0.9322	0.7030	0.9313	0.7663
Ours	**0.3859**	**0.4759**	0.9490	0.7052	0.9325	**0.8602**
	0.3608	**0.4313**	0.9493	0.7000	0.9342	**0.8593**

Table 3: Comparison between our method and other scanpath prediction models on the COCOSearch18 datasets regarding ScanMatch, SS, and MultiMatch. 'Human' refers to human performance.

Method	ScanMatch↑	SS↑	MultiMatch↑			
			Vector	Direction	Length	Position
Human	0.6781	0.7240	0.9744	0.7428	0.9707	0.9646
SaltiNet	0.4771	0.5358	0.9755	0.5287	**0.9768**	0.9195
PathGAN	0.5216	0.6236	0.9774	0.5594	0.9744	0.9229
IOR-ROI	0.4434	0.4716	0.7638	0.5105	0.7632	0.7220
VQA	0.6294	0.6583	0.9109	**0.6570**	0.9063	0.8738
Ours	**0.6341**	**0.6596**	**0.9803**	0.6476	0.9763	**0.9625**

4.4 Model Mechanisms that Resemble Human Behavior

Learned Inhibition of Return. In contrast to previous work, our model incorporated the influence of all historical fixations to predict the current fixation. To investigate our model's mechanism for processing historical fixation information, we set the model's output of previous fixations to specific regions of the image and observed the resulting effects on subsequent fixations. Fig. 7 illustrates the average changes in the subsequent fixation probability map by placing the sequence of previous fixations in specific regions of the image, calculated across all images in the OSIE dataset. We set up six sets of such experiments, placing the first five fixations at the image's upper-left, upper-center, upper-right, lower-left, lower-center, and lower-right. Each square in the probability map represents the sampling area for the first five fixations. Each probability map shows the change in the subsequent fixation probability at each pixel for the corresponding setting. It can be observed that when we placed the first five fixations in a certain region, the probability of subsequent fixations falling in that region was significantly reduced. This indicates that, during the prediction process, the fixation probability of the previously attended regions is significantly inhibited, which is consistent with the "inhibition of return" mechanism demonstrated in human visual psychophysics [36].

Self-Attention of fixation queries. To further investigate the internal mechanisms of the model, we visualized the attention weights in the self-attention module of the fixation decoder during scanpath prediction. This was done to observe the specific influence patterns between fixation queries when making saccadic decisions. Since the fixation queries are arranged in the temporal dimension, the current fixation query is only influenced by historical fixations. Therefore, the overall attention weight map forms a lower triangular shape. As shown in Fig. 8, in the temporal dimension, attention weights are higher for the nearest fixation to the current fixation query, indicating that the most recently observed areas have the most significant impact on current fixations. The attentional pattern of the model is highly consistent with human visual working memory (VWM). Historical information naturally decays over time, with the most recently observed areas having higher activation strength in the visual system. Additionally, in the attention map, the initial fixation

Table 4: Ablation results for different models in the SALICON and OSIE datasets. The best prediction results are highlighted in bold, and the second-best results are highlighted in underlined.

Method	SALICON Dataset						OSIE Dataset					
	ScanMatch↑	SS↑	MultiMatch↑				ScanMatch↑	SS↑	MultiMatch↑			
			Vector	Direction	Length	Position			Vector	Direction	Length	Position
Ours	**0.3131**	**0.3663**	0.9406	0.6458	0.9248	**0.8344**	**0.3985**	**0.4369**	0.9463	0.6619	**0.9342**	**0.8478**
w/o pre-trained in saliency	0.2855	0.3320	**0.9453**	0.6080	**0.9267**	0.8236	0.3259	0.3754	**0.9482**	0.6080	0.9325	0.8157
w/o restrain	0.2863	0.3417	0.9399	0.6387	0.9247	0.8230	0.3619	0.4044	0.9445	0.6462	0.9318	0.8340
w/o encoder	0.3041	0.3578	0.9368	**0.6507**	0.9236	0.8277	0.3873	0.4232	0.9413	0.6610	0.9282	0.8389
w/o autoregression	0.2956	0.3471	0.9391	0.6416	0.9202	0.8283	0.3666	0.4016	0.9432	0.6552	0.9273	0.8345
w/o self-attention in decoder	0.2880	0.3379	0.9384	0.6501	0.9232	0.8192	0.3556	0.3864	0.9425	**0.6652**	0.9278	0.8216

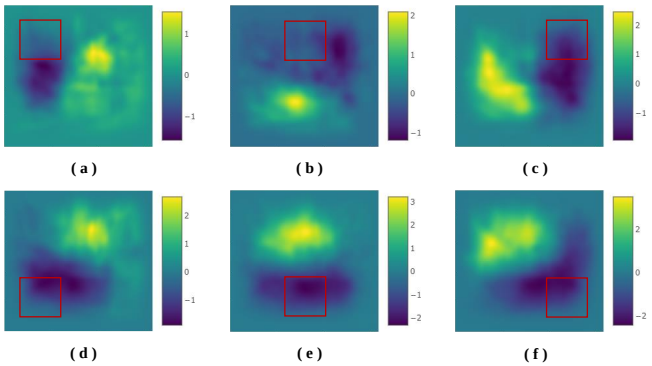

Figure 7: Probability discrepancy map of subsequent fixations after fixing the first five fixation locations. The red squares represent the region where the first five fixations are sampled.

receives more attention, consistent with human observation patterns. When we observe a scene, the brain rapidly integrates preliminary visual information to form an overall perception of the environment. This perception influences people's subsequent observation and interpretation of the finer details [5].

4.5 Ablation Studies

We conducted ablation experiments to demonstrate the effectiveness of different components and configurations. The experiments were conducted on the SALICON validation set and OSIE dataset.

Visual Feature-related components. For the feature extractor module, we employed a saliency pre-trained network for feature extraction and further optimized the feature maps with saliency spatial attention operation. We compare the results of using a VGG-19 feature extraction network pre-trained on an image classification task instead of a saliency feature extractor and a model without saliency spatial attention operation, as shown in Table. 4 with "w/o pre-trained in saliency" and "w/o restrain". The benefit of saliency feature extraction for predicting more reasonable scanpaths is significant, and saliency spatial attention operation is also more beneficial for improving the performance of the model. In addition, the experimental performance of "w/o encoder" suggests that it is advantageous to use the visual encoder after the saliency feature extractor module to learn the global correlation between regions. Still, this module has a relatively small impact on the overall performance.

Fixation-related components. For the fixation decoder module, we predicted fixations one by one in an autoregressive manner, where the location of the previous fixation was used to initialize the fixation query for the next fixation. However, this is not the sole method, and the experimental results of initializing all fixation queries randomly were labeled as "w/o autoregression" which showed that predicting fixations one by one in an autoregressive manner is more effective. In addition, we also verified the effectiveness of modeling the correlation between all historical fixations and the current fixation in the fixation decoder, and the experimental results were labeled as "w/o self-attention in the decoder". The experimental results showed that removing self-attention from the decoder led to a significant decrease in performance, indicating that it is crucial to consider the effect of historical fixations.

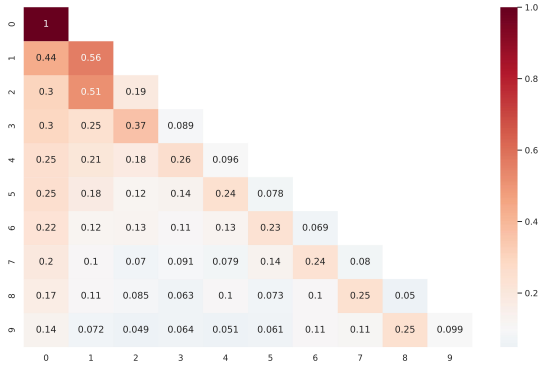

Figure 8: Visualization of the attention between fixations. Each row represents the attention of the current fixation with others. The values represent the attention weights.

5 Limitations and Future Work

In task-driven visual exploration scenarios, we abstract the task's influence in the form of guidance maps, which can be further investigated to incorporate inputs from different modalities of the task. We proposed a novel Visual ScanPath Transformer to predict people's visual scanpaths in both free-viewing and task-driven scenarios. When predicting the scanpaths, we only focus on the fixation locations and orders, ignoring the duration of each fixation. In the future, we will predict the accurate duration for each fixation, and analyze the influence of fixation duration on the whole fixation sequence.

6 Conclusion

We propose VSPT, a novel deep-learning-based visual scanpath prediction model that is applicable to both free-viewing and task-driven visual exploration. We integrate the saliency of low-level visual stimuli with contextual semantic constraints and learn the influence of historical fixations on saccade decisions by modeling the dependencies between fixation, significantly simplifying the scanpath workflow and the overall model architecture. Experiments show that VSPT can simulate the decision-making process during human exploration of visual scenes and outperforms the current state-of-the-art in both free-viewing and task-driven (goal-driven and question-driven) visual scenarios. Advances in visual scanpath prediction performance will contribute to the application of eye-tracking technology in virtual reality/augmented reality, enhancing human-computer interaction (HCI) and rendering quality.

REFERENCES

[1] P. Anderson, X. He, C. Buehler, D. Teney, M. Johnson, S. Gould, and L. Zhang. Bottom-up and top-down attention for image captioning and visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6077–6086, 2018.

[2] M. Assens, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. In *ECCVW*, September 2018.

[3] M. Assens Reina, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *ICCV*, pp. 2331–2338, 2017.

[4] W. Bao and Z. Chen. Human scanpath prediction based on deep convolutional saccadic model. *Neurocomputing*, 404:154–164, 2020.

[5] M. Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004.

[6] M. Bielikova. 5.5 utilizing eye tracking data for user modeling in personalized recommendation. *Ubiquitous Gaze Sensing and Interaction*, p. 98.

[7] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *ICCV*, pp. 921–928, 2013.

[8] A. Bulling and H. Gellersen. Toward mobile eye-based human-computer interaction. *IEEE Pervasive Computing*, 9(4):8–12, 2010.

[9] S. Chen, M. Jiang, J. Yang, and Q. Zhao. Air: Attention with reasoning capability. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part I 16*, pp. 91–107. Springer, 2020.

[10] X. Chen, M. Jiang, and Q. Zhao. Predicting human scanpaths in visual question answering. In *CVPR*, pp. 10876–10885, 2021.

[11] Y. Chen, Z. Yang, S. Ahn, D. Samaras, M. Hoai, and G. Zelinsky. Coco-search18 fixation dataset for predicting goal-directed attention control. *Scientific reports*, 11(1):8776, 2021.

[12] Y. Chen, Z. Yang, S. Chakraborty, S. Mondal, S. Ahn, D. Samaras, M. Hoai, and G. Zelinsky. Characterizing target-absent human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5031–5040, 2022.

[13] H. Chennamma and X. Yuan. A survey on eye-gaze tracking techniques. *arXiv preprint arXiv:1312.6410*, 2013.

[14] F. Cristino, S. Mathôt, J. Theeuwes, and I. D. Gilchrist. Scanmatch: A novel method for comparing fixation sequences. *Behavior research methods*, 42(3):692–700, 2010.

[15] R. Dewhurst, M. Nyström, H. Jarodzka, T. Foulsham, R. Johansson, and K. 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.

[16] A. Graves. Generating sequences with recurrent neural networks. *arXiv preprint arXiv:1308.0850*, 2013.

[17] Z. Hu, A. Bulling, S. Li, and G. Wang. Fixationnet: Forecasting eye fixations in task-oriented virtual environments. *IEEE Transactions on Visualization and Computer Graphics*, 27(5):2681–2690, 2021. doi: 10.1109/TVCG.2021.3067779

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

[19] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *TPAMI*, 20(11):1254–1259, 1998.

[20] M. Jiang, S. Huang, J. Duan, and Q. Zhao. Salicon: Saliency in context. In *CVPR*, 2015.

[21] T. Judd, F. Durand, and A. Torralba. Fixations on low-resolution images. *JOV*, 11(4):14–14, 2011.

[22] A. Kar and P. Corcoran. A review and analysis of eye-gaze estimation systems, algorithms and performance evaluation methods in consumer platforms. *IEEE Access*, 5:16495–16519, 2017.

[23] E. Keogh and C. A. Ratanamahatana. Exact indexing of dynamic time warping. *Knowledge and information systems*, 7:358–386, 2005.

[24] M. Kümmerer, M. Bethge, and T. S. Wallis. Deepgaze iii: Modeling free-viewing human scanpaths with deep learning. *Journal of Vision*, 22(5):7–7, 2022.

[25] O. Le Meur and Z. Liu. Saccadic model of eye movements for free-viewing condition. *VR*, 116:152–164, 2015.

[26] Y. Li, P. Xu, D. Lagun, and V. Navalpakkam. Towards measuring and inferring user interest from gaze. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pp. 525–533, 2017.

[27] I. Loshchilov and F. Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[28] D. Martin, A. Serrano, A. W. Bergman, G. Wetzstein, and B. Masia. Scangan360: A generative model of realistic scanpaths for 360° images. *IEEE Transactions on Visualization and Computer Graphics*, 28(5):2003–2013, 2022. doi: 10.1109/TVCG.2022.3150502

[29] Y. K. Meena, H. Cecotti, K. Wong-Lin, A. Dutta, and G. Prasad. Toward optimization of gaze-controlled human–computer interaction: Application to hindi virtual keyboard for stroke patients. *IEEE transactions on neural systems and rehabilitation engineering*, 26(4):911–922, 2018.

[30] S. Mondal, Z. Yang, S. Ahn, D. Samaras, G. Zelinsky, and M. Hoai. Gazeformer: Scalable, effective and fast prediction of goal-directed human attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 1441–1450, 2023.

[31] C. H. Morimoto and M. R. Mimica. Eye gaze tracking techniques for interactive applications. *Computer vision and image understanding*, 98(1):4–24, 2005.

[32] S. B. Needleman and C. D. Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J MOL BIOL*, 48(3):443–453, 1970.

[33] T. Ngo and B. Manjunath. Saccade gaze prediction using a recurrent neural network. In *ICIP*, pp. 3435–3439. IEEE, 2017.

[34] J. Pan, C. C. Ferrer, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. *arXiv preprint arXiv:1701.01081*, 2017.

[35] N. Parmar, A. Vaswani, J. Uszkoreit, L. Kaiser, N. Shazeer, A. Ku, and D. Tran. Image transformer. In *ICML*, pp. 4055–4064. PMLR, 2018.

[36] M. I. Posner, Y. Cohen, et al. Components of visual orienting. *Attention and performance X: Control of language processes*, 32:531–556, 1984.

[37] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[38] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. Wetzstein. Saliency in vr: How do people explore virtual environments? *IEEE transactions on visualization and computer graphics*, 24(4):1633–1642, 2018.

[39] M. Slater. Immersion and the illusion of presence in virtual reality. *British journal of psychology*, 109(3):431–433, 2018.

[40] X. Sui, Y. Fang, H. Zhu, S. Wang, and Z. Wang. Scandmm: A deep markov model of scanpath prediction for 360○ images. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2023.

[41] W. Sun, Z. Chen, and F. Wu. Visual scanpath prediction using ior-roi recurrent mixture density network. *TPAMI*, 43(6):2101–2118, 2019.

[42] X. Sun, H. Yao, and R. Ji. What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency. In *CVPR*, pp. 1552–1559, 2012.

[43] V. Tanriverdi and R. J. Jacob. Interacting with eye movements in virtual environments. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*, pp. 265–272, 2000.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *NIPS*, 30, 2017.

[45] D. Walther and C. Koch. Modeling attention to salient proto-objects. *NN*, 19(9):1395–1407, 2006.

[46] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao. Simulating human saccadic scanpaths on natural images. In *CVPR*, pp. 441–448. IEEE, 2011.

[47] Y. Wang, B. Wang, X. Wu, and L. Zhang. Scanpath estimation based on foveated image saliency. *Cognitive processing*, 18(1):87–95, 2017.

[48] J. M. Wolfe. Visual search. *Current biology*, 20, 2010.

[49] C. Xia, F. Qi, and G. Shi. An iterative representation learning framework to predict the sequence of eye fixations. In *ICME*, pp. 1530–1535, 2017.

[50] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao. Predicting human gaze beyond pixels. *JOV*, 14(1):28–28, 2014.

[51] P. Xu, K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and

J. Xiao. Turkergaze: Crowdsourcing saliency with webcam based eye tracking. *arXiv preprint arXiv:1504.06755*, 2015.

[52] Z. Yang, L. Huang, Y. Chen, Z. Wei, S. Ahn, G. Zelinsky, D. Samaras, and M. Hoai. Predicting goal-directed human attention using inverse reinforcement learning. In *CVPR*, pp. 193–202, 2020.

[53] Z. Yang, S. Mondal, S. Ahn, G. Zelinsky, M. Hoai, and D. Samaras. Target-absent human attention. In *Computer Vision–ECCV 2022: 17th European Conference, Tel Aviv, Israel, October 23–27, 2022, Proceedings, Part IV*, pp. 52–68. Springer, 2022.

[54] G. Zelinsky, Z. Yang, L. Huang, Y. Chen, S. Ahn, Z. Wei, H. Adeli, D. Samaras, and M. Hoai. Benchmarking gaze prediction for categorical visual search. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 0–0, 2019.

[55] X. Zhou, D. Wang, and P. Krähenbühl. Objects as points. *arXiv preprint arXiv:1904.07850*, 2019.