Unifying Top-down and Bottom-up Scanpath Prediction Using Transformers

Zhibo Yang[1,2], Sounak Mondal[1], Seoyoung Ahn[1], Ruoyu Xue[1],
Gregory Zelinsky[1], Minh Hoai[1,3], Dimitris Samaras[1]
[1]Stony Brook University [2]Waymo LLC [3]VinAI Research

Abstract

Most models of visual attention aim at predicting either top-down or bottom-up control, as studied using different visual search and free-viewing tasks. In this paper we propose the Human Attention Transformer (HAT), a single model that predicts both forms of attention control. HAT uses a novel transformer-based architecture and a simplified foveated retina that collectively create a spatio-temporal awareness akin to the dynamic visual working memory of humans. HAT not only establishes a new state-of-the-art in predicting the scanpath of fixations made during target-present and target-absent visual search and "taskless" free viewing, but also makes human gaze behavior interpretable. Unlike previous methods that rely on a coarse grid of fixation cells and experience information loss due to fixation discretization, HAT features a sequential dense prediction architecture and outputs a dense heatmap for each fixation, thus avoiding discretizing fixations. HAT sets a new standard in computational attention, which emphasizes effectiveness, generality, and interpretability. HAT's demonstrated scope and applicability will likely inspire the development of new attention models that can better predict human behavior in various attention-demanding scenarios. Code is available at https://github.com/cvlab-stonybrook/HAT.

1. Introduction

Attention, a cognitive process that allows humans to selectively allocate their limited cognitive resources to specific regions of the visual world, plays a crucial role in human perception system. Understanding and predicting human (visual) attention will enable numerous applications such as assistive technologies that can anticipate a person's needs and intents, perceptions system that can prioritize processing regions of human interest and enhancing the accuracy and speed of various visual tasks (e.g., object detection), and image/video compression that allocates more resources to encoding and transmitting high-attention regions, optimizing the use of bandwidth.

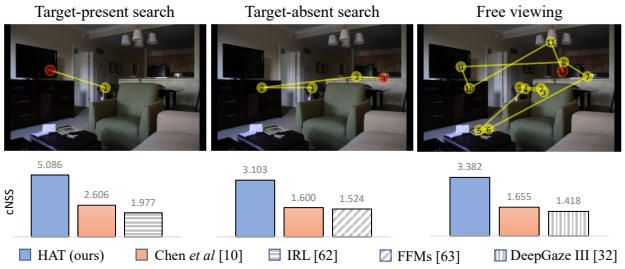

Figure 1. Given an image, the proposed **HAT** is able to predict scanpaths under three settings target-present search for TV; target-absent scanpath for sink; and free viewing. Importantly, HAT outperforms previous state-of-the-art scanpath prediction methods on multiple datasets across three settings: target-present, target-absent visual search and free viewing, that were studied separately.

Human attention control can take two broad forms. One is bottom-up, meaning that attention saliency signals are computed from the visual input and used to prioritize shifts of attention. The same visual input should therefore lead to the same shifts of bottom-up attention. The second type of attention is top-down, meaning that a task or goal is used to control attention. Given a kitchen scene, very different fixations are observed depending on whether a person is searching for a clock or a microwave oven [67]. These two types of attention control spawned two separate literatures on gaze fixation prediction (the accepted measure of attention), one where studies use a free-viewing task to study questions of bottom-up attention and the other using a goal-directed task (typically, visual search) to study top-down attention control. Consequently, most models have been designed to address *either* bottom-up *or* top-down attention, not both. *Can a single model architecture predict both bottom-up and top-down attention control?*

Our answer to this question is HAT, a Human Attention Transformer that generally predicts scanpaths of fixations, meaning that it can be applied to both top-down visual search and bottom-up free viewing tasks (Figure 1). Devising a unified model architecture capable of predicting both bottom-up and top-down attention control is nontrivial: 1) predicting human fixation scanpaths requires the model to

have a spatio-temporal understanding of the fixated image contents and their relationship to the external goals; and 2) predicting top-down and bottom-up attention requires the model to capture both low-level features and high-level semantics of the input image. HAT addresses these issues by using a novel transformer-based design and a simplified foveated retina. Together, these components forge a novel paradigm, constituting a form of *dynamically-updating visual working memory*. Traditional approaches have leaned on recurrent neural networks (RNNs) to uphold a dynamically updated hidden vector conveying information across fixations [1, 10, 51, 66]. Alternatively, simulations of a foveated retina have combined multi-resolution information at pixel [66], feature [63], or semantic levels [62]. However, these methods present drawbacks: RNNs sacrifice interpretability, while multi-resolution simulations fall short in capturing crucial temporal and spatial information integral for scanpath prediction.

In addressing these challenges, we leverage a computational attention mechanism [54] to dynamically assimilate spatial, temporal, and visual information acquired at each fixation into working memory [45, 46]. This empowers HAT to discern a set of task-specific attention weights for amalgamating information from working memory and forecasting human attention control. This innovative mechanism sheds light on the intricate relationship between human attention and working memory [17, 20], rendering HAT not only cognitively plausible but also ensuring the interpretability of its predictions. Furthermore, in contrast to prior methods [10, 62, 63], HAT treats scanpath prediction as a sequence of dense prediction tasks with per-pixel supervision, successfully avoiding the need for discretizing fixations. This enhances the method's efficacy, particularly in scenarios involving high-resolution imagery.

To demonstrate HAT's generality, we predict scanpaths under three settings, target-present (TP) and target-absent (TA) visual search, and free-viewing (FV), covering both top-down and bottom-up attention. In previous work predicting search scanpaths [10, 62, 63], separate models were trained for the TP and TA settings. HAT is a single model establishing new SOTA in both TP and TA search-scanpath prediction. When trained with FV scanpaths, HAT also achieves top performance relative to baselines. HAT advances SOTA in cNSS by 95%, 94% and 104% under the TP, TA and FV settings on the COCO-Search18 dataset [11] and the COCO-FreeView dataset [12], respectively.

Our contributions can be summarized as follows:

1. We propose HAT, a novel transformer architecture integrating visual information at two different eccentricities (approximating a foveated retina) to predict the spatial and temporal allocation of human attention.
2. We formulate scanpath prediction as a sequential dense prediction task without fixation discretization, making HAT applicable to high-resolution input.
3. The HAT architecture can be broadly applied to different attention control tasks, evidenced by the SOTA scanpath predictions in both visual search and free-viewing tasks.
4. HAT's attention predictions offer high interpretability, making it useful for studying gaze behavior.

2. Related Work

Saliency prediction. Predicting and understanding human gaze control has been a topic of interest for decades in psychology [19, 59, 64, 65], but it has only recently attracted the researcher's attention in computer vision. In particular, Itti's seminal work [23] on the saliency model has triggered a lot of interest on human attention modeling in computer vision community and facilitated many other studies identifying and modeling the salient visual features of an image (i.e., saliency prediction) [3, 5, 7, 16, 22, 24, 25, 28, 30, 31, 41, 55, 56]. However, the scope of these work is often narrowly focused on predicting human natural eye-movements without a specific visual task (i.e., free-viewing), ignoring another important form of attention control, goal-directed attention. Moreover, saliency models only model the spatial distribution of fixations and do not predict the temporal order between fixations. Scanpath prediction is more challenging problem because it requires predicting not only *where* a fixation will be, but also *when* it will be there.

Scanpath prediction. Many existing scanpath prediction deep neural networks (DNN) focus on predicting the free-viewing scanpaths [1, 2, 32, 51], primarily due to their close connection to saliency modeling. However, these models are inherently constrained in their ability to capture the full spectrum of human attention control, particularly goal-directed attention—a fundamental cognitive process that underlies various everyday visual tasks such as navigation and motor control. Although goal-directed human attention has been studied for decades [33, 58, 64] in cognitive science (mainly in the context of visual search [43, 52, 65]), the development of DNNs for goal-directed scanpath prediction lags behind those designed for free-viewing tasks, partly due to the lack of data. To tackle this problem, Chen et al. [11] created the first large-scale goal-directed gaze dataset with 18 search targets, COCO-Search18. In [62], an inverse reinforcement learning model showed superior performance on COCO-Search18 in predicting TP scanpaths. Later, Chen et al. [10] showed that a reinforcement learning model directly optimized on the scanpath similarity metric can predict VQA scanpaths, as well as on TP search scanpaths. Rashidi et al. [50], Yang et al. [63] also proposed a more generalized scanpath prediction model that can be applied to both target-present and target-absent visual search scanpaths. Most recently, a transformer-based scanpath prediction model, Gazeformer [42], further advanced the TP search scanpath prediction performance on COCO-

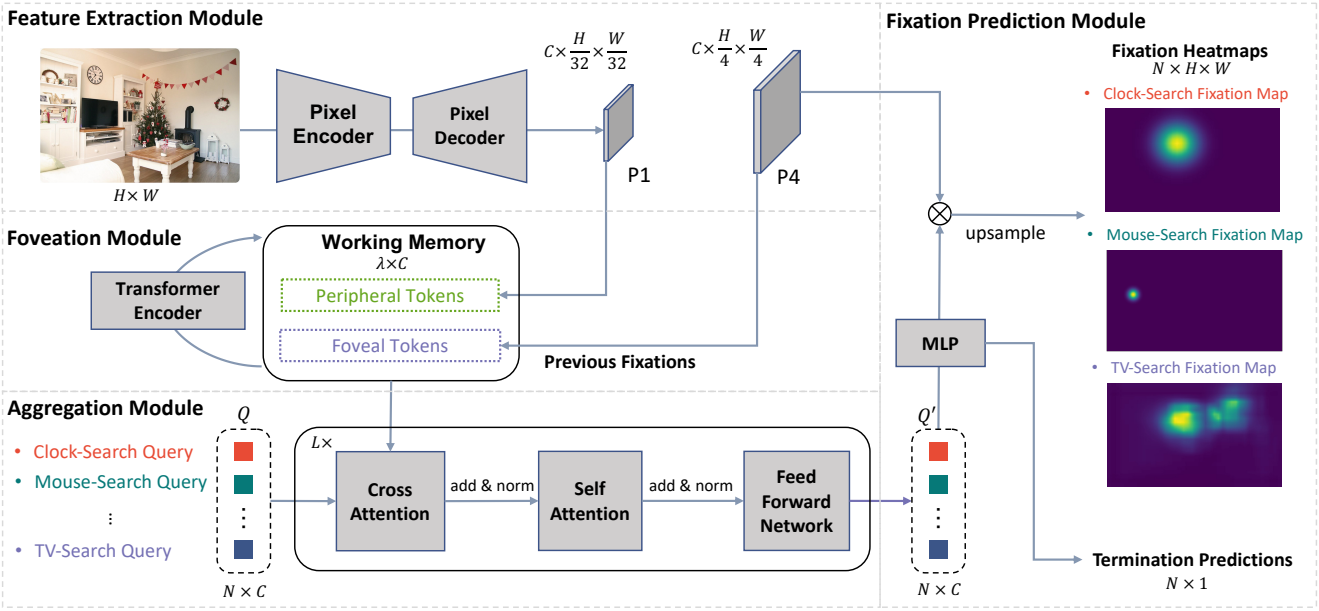

Figure 2. **HAT overview.** We use encoder-decoder CNNs to extract two sets of feature maps P_1 and P_4 of different spatial resolutions. A working memory with a capacity of λ tokens is constructed by combining all feature vectors from P_1 with the feature vectors of P_4 at previously fixated locations, representing information extracted from the periphery and central fovea. A transformer encoder is used to dynamically update the working memory at every new fixation. Then, HAT produces N per-task queries of dimension C (e.g., clock search and mouse search), with each learning to aggregates task-specific information from the shared working memory for predicting the fixations for its own task. Finally, the updated queries are convolved with P_4 to yield the fixation heatmaps after a MLP layer, and projected to the termination probabilities in parallel. Note, although this figure depicts visual search, the framework also applies for free viewing.

Search18. However, none of these work have demonstrated the generalizability to all three settings (i.e., TP, TA and FV). In this work, we design a generic scanpath model that generalizes to both free-viewing and visual search tasks.

Scanpath Transformers. The transformative game-changing impact of Transformers [54] has been widely recognized in natural language processing and beyond. In computer vision, Transformers have demonstrated outstanding capabilities across a wide range of computer vision tasks, such as image recognition [18, 38, 53], object detection [9, 69] and image segmentation [14, 49, 60]. Mondal et al. [42] introduced Gazeformer, a Transformer-based model specifically designed for zero-shot visual search scanpath prediction. In contrast, our proposed model is generic, capable of predicting both visual search and free-viewing scanpaths. Additionally, our model diverges from other Transformer-based architectures by drawing inspiration from the human vision system. It incorporates a novel foveation module simulating a simplified foveated retina, thereby establishing a dynamic visual working memory for enhanced scanpath prediction.

3. Human Attention Transformer

In this section, we first formulate scanpath prediction as a sequence of dense prediction tasks using behavior cloning. We then introduce our proposed transformer-based model, HAT, for scanpath prediction. Finally, we describe how we train HAT and use it for fast inference.

3.1. Preliminaries

To avoid the precision loss caused by grid discretization present in prior fixation prediction methods [10, 62, 63, 66], we formulate scanpath prediction as a sequential prediction of pixel coordinates. Given a $H \times W$ image and an optional initial fixation f_0 (often set as the center of an image), a scanpath prediction model predicts a sequence of human-like fixation locations $f_1, \cdots, f_n$, with each fixation f_i being a pixel location in the image. Note that n is variable that may be different for each scanpath due to the different termination criteria of different human subjects. To model the uncertainty in human attention allocation, existing methods [10, 62, 63, 66] often predict a probability distribution over a coarse grid of fixation locations at each step. HAT follows the same spirit but outputs a dense fixation heatmap. Specifically, HAT outputs a heatmap $Y_i \in [0, 1]^{H \times W}$ with each pixel value indicating the chance of the pixel being fixated in the next fixation. In addition, HAT also outputs a termination probability $\tau_i \in [0, 1]$ indicating how likely the model is to terminate the scanpath at the current step i. To sample a fixation, we apply L_1-normalization on Y_i. In the

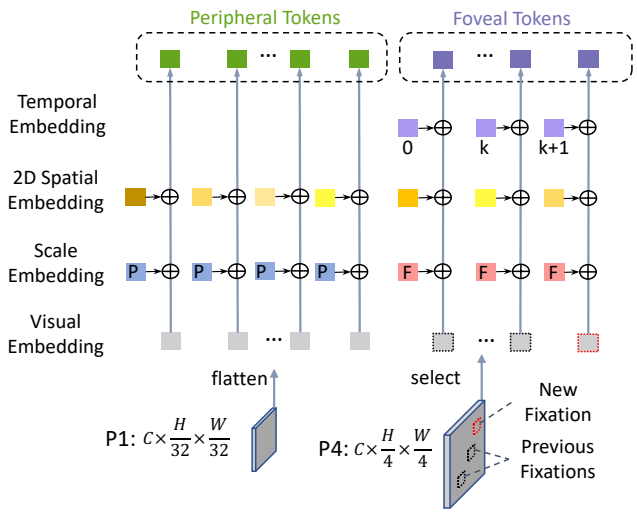

Figure 3. **Working memory construction.** We construct the working memory by starting with the visual embeddings ("what") flattened from P_1 over the spatial axes and selected from P_4 at previous fixation locations. A scale embedding is introduced to capture scale information. Spatial embeddings and temporal embeddings are further added to the tokens to enhance the "where" and "when" signals. At every new fixation (marked in red), we simply add a new foveal token while keeping other tokens unchanged.

following, we omit the subscript i for brevity.

3.2. Network Architecture

HAT is a novel transformer-based model for scanpath prediction. At each fixation, HAT outputs a set of prediction pairs $\{(Y_t, \tau_t)\}_{t=1}^{T}$ where t indicates a task, which could be a visual search task (e.g., clock search and mouse search) or a free-viewing task. Figure 2 shows an overview of the proposed model. HAT consists of four modules: 1) a feature extraction module that extracts a feature pyramid with multi-resolutional feature maps corresponding to information extracted at different eccentricities [50, 63]; 2) a foveation module which maintains a dynamical working memory representing the information acquired through fixations; 3) an aggregation module that selectively aggregates the information in the working memory using attention mechanism for each task; 4) a fixation prediction module that predicts the fixation heatmap Y_t and termination probability τ_t for each task t.

The feature extraction module consists of a pixel encoder (e.g., ResNet [21], a Swin transformer [38]), and a pixel decoder (e.g., FPN [36] and deformable attention [69]). Taking a $H\times W$ image as input, the pixel encoder encodes the input image into a high-semantic but low-resolution feature map. The pixel decoder up-samples the feature map several times, each time by a scale factor of two, to construct a pyramid of four multi-scale feature maps denoted as $P = \{P_1, \cdots, P_4\}$, where $P_1 \in \mathbb{R}^{C\times\frac{H}{32}\times\frac{W}{32}}$, $P_4 \in \mathbb{R}^{C\times\frac{H}{4}\times\frac{W}{4}}$, and C is the channel dimension.

The foveation module constructs a *dynamic* working memory using the feature maps P_1 and P_4 to represent the information a person acquires from the peripheral and foveal vision, respectively. We discard medium-grained feature maps P_2 and P_3 in computing the peripheral representation for computational efficiency. Finally, we apply a Transformer encoder [54] to dynamically update the working memory with the information acquired at a new fixation. Figure 3 illustrates the construction of the working memory. The working memory consists of two parts: peripheral tokens and foveal tokens. We first flatten the low-resolution feature map P_1 over the spatial axes to obtain the peripheral visual embeddings $V^p \in \mathbb{R}^{(\frac{H}{32}\cdot\frac{W}{32})\times C}$. Feature vectors in P_4 at each fixation location are selected as the foveal visual embeddings $V^f \in \mathbb{R}^{k\times C}$, where k is number of previous fixations. For simplicity, we round the fixation to its nearest position in P_4. Then we add a learnable **scale** embedding to each token to discern the scale/resolution of the visual embeddings. As the spatial information is shown to be important in predicting human scanpath (e.g., center bias and inhibition of return [57]), we enrich the peripheral and foveal tokens with their 2D **spatial** information in the image. Specifically, we create a lookup table of 2D sinusoidal position embeddings [35] $G \in \mathbb{R}^{H\times W\times C}$ by concatenating the 1D sinusoidal positional encoding of the horizontal and vertical coordinates of each pixel location. For a visual embedding at position (i, j) of a given feature map of stride S ($S = 32$ for P_1 and $S = 4$ for P_4), its position encoding is defined by the element at position (t_i, t_j) in G where $t_i = \lfloor i \cdot S \rfloor$ and $t_j = \lfloor j \cdot S \rfloor$. Furthermore, we add to each foveal token the **temporal** embedding, a learnable vector, according to its fixation index to capture the temporal order among previous fixations.

The aggregation module is a transformer decoder [54] that selectively aggregates information from the working memory using a set of learnable, task-specific queries $Q \in \mathbb{R}^{N\times C}$, where N is the number of tasks (e.g., $N = 18$ for COCO-Search18 [11] and $N = 1$ for free-viewing datasets). The transformer decoder has L layers, with each layer consisting of a cross-attention layer, a self-attention layer and a feed-forward network (FFN). Different from the standard transformer decoder [54], we follow [14] and switch the order of cross-attention and self-attention module. Firstly, each task query selectively gathers the information in working memory acquired through previous fixations using cross-attention. Then, the self-attention layer followed by a FFN is applied to exchange information in different queries which could boost the contextual cues [15] in each query. When generating a scanpath, HAT maintains its state across fixations *only* in the working

memory, the input Q is the same at each fixation prediction. **The fixation prediction module** yields the final prediction—a fixation heatmap $\hat{Y}_t$ and a termination probability $\hat{\tau}_t$ for each task t. For the termination prediction, a linear layer followed by a sigmoid activation is applied on top of each updated query $q_t \in Q$:

$$\hat{\tau}_t = \text{sigmoid}(W q_t^T + b), \tag{1}$$

where W and b are the parameters of the linear layer. For the fixation heatmap prediction, a Multi-Layer Perceptron (MLP) with two hidden layers first transforms q_t into a task embedding, which is then convolved with the high-resolution feature map P_4 to get the fixation heatmap $\hat{Y}_t$ after a sigmoid layer:

$$\hat{Y}_t = \text{sigmoid}(P_4 \odot \text{MLP}(q_t)), \tag{2}$$

where $\odot$ denotes the pixel-wise dot product operation. Finally, we upsample $\hat{Y}_t$ to the image resolution. Note that the predictions for all tasks, i.e., $\hat{Y} \in \mathbb{R}^{N \times H \times W}$ and $\hat{\tau} \in \mathbb{R}^{N \times 1}$, are yielded in parallel.

3.3. Training and Inference

Training loss. We follow [66] and use behavior cloning to train HAT. The problem of scanpath prediction is broken down into learning a mapping from the input triplet of an image, a sequence of previous fixations, and a task to the output pair of a fixation heatmap and a termination probability. Given the predicted fixation heatmaps $\hat{Y} \in \mathbb{R}^{N \times H \times W}$ and termination probabilities $\hat{\tau} \in \mathbb{R}^{N \times 1}$, the training loss is only calculated for its ground-truth task t:

$$\mathcal{L} = \mathcal{L}_{\text{fix}}(\hat{Y}_t, Y) + \mathcal{L}_{\text{term}}(\hat{\tau}_t, \tau), \tag{3}$$

where $Y \in [0,1]^{H \times W}$ and $\tau \in \{0,1\}$ are the ground-truth fixation heatmap and termination label for task t, respectively. We compute Y by smoothing the ground-truth fixation map with a Gaussian kernel with the kernel size being one degree of visual angle. $\mathcal{L}_{\text{fix}}$ denotes the fixation loss and is computed using pixel-wise focal loss [34, 37]:

$$\mathcal{L}_{\text{fix}} = \frac{-1}{HW} \sum_{i,j} \begin{cases} (1-\hat{Y}_{ij})^{\alpha} \log(\hat{Y}_{ij}) & \text{if } Y_{ij} = 1, \\ (1-Y_{ij})^{\beta} (\hat{Y}_{ij})^{\alpha} & \\ \log(1-\hat{Y}_{ij}) & \text{otherwise}, \end{cases} \tag{4}$$

where Y_{ij} represents the value of Y at location (i,j) and we set $\alpha = 2$ and $\beta = 4$ following [34, 63]. $\mathcal{L}_{\text{term}}$ is the termination loss and is computed by applying a binary cross entropy (negative log-likelihood) loss, i.e.,

$$\mathcal{L}_{\text{term}} = -\omega \cdot \tau \log(\hat{\tau}_t) - (1-\tau)\log(1-\hat{\tau}_t), \tag{5}$$

where ω is a weight to balance the loss of positive and negative training examples since there are many more negative labels than positive labels for training a termination prediction, especially for target-absent visual search and free-viewing tasks where scanpath are long. We set ω to be the ratio of the number of negative training instances to the number of positive ones.

Inference. Similarly to [10, 62, 63], HAT also generates scanpaths autoregressively, but in an efficient way. Given an image, HAT only computes the image pyramid P and peripheral tokens once. For a new fixation, a foveal token is constructed and appended to the working memory after which the aggregation module and fixation prediction module yield the fixation heatmaps and termination predictions for all tasks in parallel.

4. Experiments

Datasets. We train and evaluate HAT using four datasets: COCO-Search18 [11], COCO-FreeView [12], MIT1003 [26] and OSIE [61]. COCO-Search18 is a large-scale visual search dataset containing human scanpaths in searching for 18 different object target and it has two parts: target-present and target-absent. In total, there are 3101 target-present images and 3101 target-absent images in COCO-Search18, each viewed by 10 subjects. Following [63], we treat the target-present part and target-absent part of COCO-Search18 as two separate datasets and train models on them independently. COCO-FreeView is a "sibling" dataset of COCO-Search18 but with free-viewing scanpaths. COCO-FreeView contains the same images with COCO-Search18, each viewed by 10 subjects in a free-viewing setting. MIT1003 is a widely-used free-viewing dataset containing 1003 natural images. OSIE is a free-viewing gaze dataset with rich semantic-level annotations, containing 700 natural indoor and outdoor images. Each image in MIT1003 and OSIE is viewed by 15 subjects.

Evaluation metrics. To measure the performance, we mainly analyze the scanpath prediction models from two aspects: 1) how similar the predicted scanpaths are to the human scanpaths; and 2) how accurate a model predicts the next fixation *given all previous fixations*. To measure the scanpath similarity, we use a commonly adopted metric, sequence score (SS) [6] and its variant semantic sequence score (SemSS) [63]. SS transforms the scanpaths into sequences of fixation cluster IDs and then compares them using a string matching algorithm [44]. Different from SS, SemSS transforms a scanpath into a string of semantic labels of the fixated pixels. For next fixation prediction, we follow [29, 32, 63] and report the conditional saliency metrics, cIG, cNSS and cAUC, which measure how well a predicted fixation probability map of a model predicts the ground-truth (next) fixation when the model is provided with the fixation history of the scanpath in consideration, using the widely used saliency metrics, IG, NSS and AUC [8]. For fair comparison, we follow [63] and predict one

scanpath for each testing image, step by step selecting the most probable fixation location as the next fixation.

Baselines. We first compare our model against several heuristic baselines. Following prior works [10, 32, 62, 63, 66], the human consistency, an oracle where we use one viewer's scanpath to predict the scanpath of another, is reported as a gold-standard model. Second, we compare to a fixation heuristic method—a ConvNet trained to predict human fixation density maps, from which we select fixations sequentially with inhibition of return. For visual search scanpaths, we further include a detector baseline, which is similar to the fixation heuristic, but trained on target-present images of COCO-Search18 to predict a target detection probability map. For both fixation heuristic and detector baselines, we use the winner-take-all strategy to generate scanpaths. Furthermore, we compare HAT to the previous state-of-the-art models of scanpath prediction: IVSN [68], PathGAN [1], IRL [62], Chen *et al*. [10], DeepGaze III [32], FFMs [63] and GazeFormer [42]. Note that IVSN only applies for visual search tasks, and unlike other methods, IVSN is designed for zero-shot search scanpath prediction, hence is not trained with any gaze data. DeepGaze III only applies for free-viewing scanpaths and is trained with the SALICON dataset [25] and MIT1003 [26].

Implementation details. We use ResNet-50 [21] as the pixel encoder and MSDeformAttn [69] as the pixel decoder. For the foveation module, the transformer encoder has three layers. The transformer decoder in the aggregation module has six layers (i.e,. $L = 6$). All transformer encoder and decoder layers in HAT have 4 attention heads. The MLP in the fixation prediction module has two linear layers with 512 hidden dimensions and a ReLU activation function. We use the AdamW [39] with the learning rate of 0.0001 and train HAT for 30 epochs with a batch size of 128. All images are resized to 320×512 for computational efficiency during training and inference. Following [62], we set the maximum length of each predicted scanpath to 6 and 10 (excluding the initial fixation) for target-present and target-absent search scanpath prediction, respectively. For free viewing, the maximum scanpath length is set to 20. For more implementation details, please refer to the supplement.

4.1. Main Results

Target-present search. We compare HAT with previous scanpath prediction models under the target-present (TP) setting using the target-present part of the COCO-Search18 dataset in Tab. 1. HAT consistently outperforms all other predictive methods in predicting TP human scanpaths in nearly all metrics. The simple heuristic baselines (i.e., detector and fixation heuristic) perform quite well on TP scanpath prediction by predicting the location of the target or fixation density map as in 60% of the TP trials of COCO-Search18 humans can locate the target within 2 fixations.

	SemSS	SS	cIG	cNSS	cAUC
Human consistency	0.500	0.500	-	-	-
Detector	0.523	0.449	0.182	2.346	0.905
Fixation heuristic	0.506	0.437	1.107	2.186	0.917
IVSN [68]	0.368	0.326	-0.192	1.318	0.901
PathGAN [1]	0.280	0.239	-	-	-
IRL [62]	0.486	0.422	-9.709	1.977	0.913
Chen *et al*. [10]	0.518	0.445	-1.273	2.606	0.956
FFMs [63]	0.500	0.451	1.548	2.376	0.932
Gazeformer [42]	0.499	**0.489**	-	-	-
HAT (ours)	**0.543**	0.470	**2.399**	**5.086**	**0.977**

Table 1. **Target-present search scanpath prediction comparison** on the target-present test set of COCO-Search18. We highlight the best results in bold.

	SemSS	SS	cIG	cNSS	cAUC
Human consistency	0.372	0.381	-	-	-
Detector	0.332	0.321	-0.516	0.446	0.783
Fixation heuristic	0.309	0.298	-0.599	0.405	0.798
IVSN [68]	0.279	0.260	-0.219	0.884	0.867
PathGAN [1]	0.315	0.250	-	-	-
IRL [62]	0.329	0.319	0.032	1.202	0.893
Chen *et al*. [10]	0.340	0.331	-3.278	1.600	0.925
FFMs [63]	0.376	0.372	0.729	1.524	0.916
Gazeformer [42]	0.374	0.357	-	-	-
HAT (ours)	**0.382**	**0.402**	**1.686**	**3.103**	**0.961**

Table 2. **Target-absent search scanpath prediction comparison** on the target-absent test set of COCO-Search18. We highlight the best results in bold.

	SS	cIG	cNSS	cAUC
Human consistency	0.349	-	-	-
Fixation heuristic	0.329	0.319	1.621	0.930
PathGAN [1]	0.181	-	-	-
IRL [62]	0.300	-0.213	1.018	0.888
Chen *et al*. [10]	0.365	-1.263	1.655	0.922
DeepGaze III [32]	0.339	0.140	1.418	0.910
FFMs [63]	0.329	0.329	1.432	0.918
Gazeformer [42]	0.280	-	-	-
HAT	**0.369**	**1.485**	**3.382**	**0.965**

Table 3. **Comparing free-viewing scanpath prediction algorithms** (rows) using multiple metrics (columns) on the test set of COCO-FreeView. The best results are highlighted in bold.

However, they have low scores on saliency metrics (i.e., cIG, cNSS and cAUC) as they ignore the inter-dependencies between fixations. Compared to FFMs [63] and Chen *et*

al. [10] which have high saliency scores, HAT further improves the performance significantly for all metrics. Particularly, HAT is better than Chen *et al*. [10] (the second best) in cNSS by 95%. HAT slightly lags behind the most recent GazeFormer [42] in SS but is significantly better in semSS. We also demonstrate in the supplement that HAT learns the entire scanpath distribution from multiple subjects whereas GazeFormer overfits to the "average person" and fails to predict scanpaths from different subjects. Moreover, HAT surpasses the human consistency in semSS, suggesting that HAT well captures the semantics behind fixations.

Target-absent search. For target-absent (TA) search scanpath prediction, we compare HAT to different approaches on the TA test set of COCO-Search18 in Tab. 2. Different from TP search results shown in Tab. 1, we see in Tab. 2 that the gap between heuristic methods to human consistency is much larger for TA search, demonstrating that TA search scanpath prediction is a more challenging task than TP scanpath prediction. Indeed, the predominant influence on human attention in TP search (i.e., the target) is now absent [12], making other factors such as the spatial cues provided by the anchor objects [4], the contextual cues from global scene understanding [52] and object co-occurrence [40] stand out. The discernment of these factors necessitates a robust semantic understanding of the input image. Tab. 2 shows that HAT sets a new state-of-the-art at **all** metrics, outperforming the previous state-of-the-art (Chen *et al*. [10]) by 94% in cNSS. More importantly, HAT achieves a sequence score surpassing human consistency for the first time. These results suggest that comparing to other methods HAT better captures the semantics of the image and learns the relation between other objects and targets.

Free-viewing. In addition to visual search, HAT can predict free-viewing scanpaths by treating free-viewing as a standalone task. In Tab. 3, we compare HAT with the baselines using COCO-FreeView. Note that Detector and IVSN are excluded here as the free-viewing fixations are not tasked to searching for a target like visual search. HAT outperforms all other methods in cIG, cNSS and cAUC, especially HAT is 351% and 104% better than the second best (FFMs and Chen *et al*. [10]) in cIG, cNSS, respectively. This reaffirms the effectiveness of HAT as a generic framework for scanpath prediction. We further validated the effectiveness of our proposed HAT using OSIE [61] and MIT1003 [26], and the generalizability of HAT to new scenes, please refer to the supplement for detailed results.

4.2. Qualitative Analysis

Scanpath visualization. In this section, we qualitatively compare the predicted scanpaths of different methods to each other and to the ground-truth human scanpaths in the TP, TA and FV settings. As shown in Fig. 4, when searching for bottles in the TP setting, HAT not only correctly predicted the terminal fixation on the heavily-occluded target, but also predicted fixations on all the distractor objects that look similar to the target, like humans do. Other methods either missed the distractor objects or failed to find the target. Similarly, for the TA stop sign search, HAT was the only one that looked at both sides of the road in searching for a stop sign like the human subject would, showing a use of semantic and context cues to control attention. In the FV setting, HAT also predicted the most human-alike scanpaths among all methods in (1) the fixation locations (where), (2) the semantics (what), and (3) the order (when) of the fixations. More scanpath visualizations can be found in the supplement.

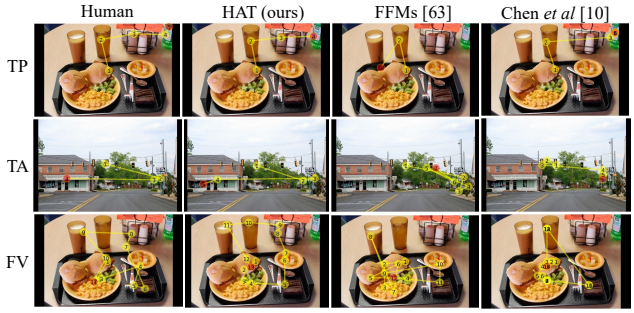

Figure 4. **Visualization of the ground-truth human scanpaths and predicted scanpaths of different methods (columns).** Three different settings (rows) including target-present bottle search, target-absent stop sign search and free viewing are shown from the top to bottom. The final fixation of each scanpath is highlighted in red circle. For methods without termination prediction, i.e., IRL, detector and fixation heuristic, we visualize the first 6 fixations for visual search and 15 for free viewing. The rightmost column shows the predicted scanpaths of the heuristic methods (detector 630 for visual search and fixation heuristic for free-viewing)

Model interpretability. A distinctive attribute of HAT lies in its interpretability, facilitated by the computational attention mechanism and the foveation module design. HAT enables quantitatively measuring the contribution of both peripheral and foveal tokens to fixation allocation. The contribution of a token is computed as the attention weight from the last cross-attention layer of the aggregation module in HAT. By computing the normalized contribution of each peripheral token, we create a *peripheral contribution map*, which offers insights for the human gaze behavior. We further analyze how the peripheral contribution map evolves across a sequence of fixations. Fig. 5 shows the predicted scanpath, peripheral contribution maps and predicted fixation heatmaps of HAT in a TP laptop search task. We observe that the encoded periphery features not only align with the location of the next fixation (e.g., when the occluded laptop is encoded in the left-bottom periphery, the model makes a fixation to the target and terminates

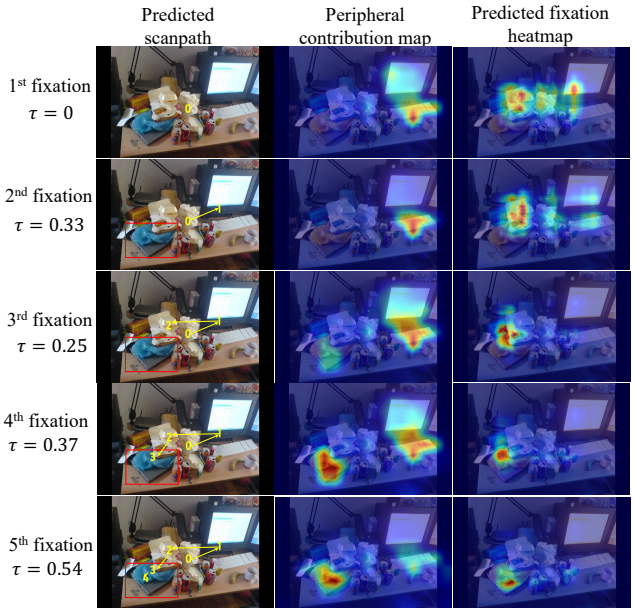

Figure 5. Visualization of the **predicted scanpath, peripheral contribution map and fixation heatmap** (columns) of HAT for target-present laptop visual search examples at every fixation (rows). We also include the predicted termination probability τ for each step on the left. The model terminates searching if $\tau > 0.5$.

the search), but also provides the *contextual cues* where a target might be located (e.g., near the keyboard and the monitor where a laptop is usually found). We also observe a similar pattern for the TA setting (see illustration in the supplement). In the supplement, we also collectively analyze the contribution of peripheral and foveal tokens in predicting human attention control, which has shown that the peripheral vision plays different roles under different settings. These all have demonstrated that HAT can make highly *interpretable* predictions.

4.3. Ablation studies

We ablate HAT under the TA setting as TA search fixations exhibit characteristics of both target-present search fixations and free-viewing fixations [12].

Peripheral and foveal tokens. We verify the effectiveness of peripheral tokens and foveal tokens by ablating them one at a time. It is shown in Tab. 4 that ablating any one of them incur a performance drop over all metrics. This suggests that all of these components contribute to the superior performance of HAT. In comparison, removing foveal tokens incurs a larger performance drop (cIG decreases by 30%). This decline is expected as foveal tokens embody the knowledge accumulated from prior fixations. Without them, HAT can be regarded as a static fixation density map predictor, akin to the fixation heuristic baseline. Conversely, the removal of peripheral tokens has a relatively minor effect, possibly attributed to the adaptive capacity of foveal tokens (P_2) compensating for information loss in peripheral tokens (P_1) during training.

	SemSS	SS	cIG	cNSS	cAUC
baseline (80×128)	**0.382**	**0.402**	**1.686**	**3.103**	**0.961**
– peripheral tokens	0.375	0.396	1.600	3.003	0.960
– foveal tokens	0.358	0.385	1.179	2.380	0.948
low-res (20×32)	0.374	0.389	1.534	2.760	0.955

Table 4. **Ablation study** of HAT. These experiments are done on the TA set of COCO-Search18. The best results are in bold.

Output resolution. HAT has a default output resolution of 80×128 due to the convolution with the high-resolution feature map P_4 (see Fig. 2). In Tab. 4 (last row), we change the convolution operant from P_4 to P_2 to yield an output resolution of 20×32, same as FFMs [63] and IRL [62] but smaller than Chen *et al*. [10] (30×40). Despite that a reduced resolution incurs a noticeable performance drop in HAT, HAT still outperforms prior state-of-the-art FFMs with the same output resolution and Chen et al. [10] using a higher output resolution. This underscores HAT's effectiveness and design flexibility. Additional ablations can be found in the supplement.

5. Conclusions

With the rapid development of Augmented Reality (AR) and Virtual Reality (VR) technologies, there is an increasing demand for predicting and understanding human gaze behavior [27, 47, 48], with scanpath prediction being a challenging task. For those AR/VR applications requiring a high input resolution (360°), discretizing fixations into a coarse grid incurs a non-negligible loss in accuracy. In this work we presented HAT, a generic attention scanpath prediction model. Built from a simple dense prediction framework [13], HAT circumvents the drawbacks of discretizing fixations as in prior state of the arts [10, 62, 63]. Inspired by the human vision system, HAT uses a novel foveated working memory which dynamically updates its knowledge about the scene as it changes its fixation. We show that HAT achieves new SOTA performance, not only in predicting free-viewing fixation scanpaths, but also scanpaths in target-present and target-absent search. In demonstrating this broad scope, our HAT model sets a new bar in the computational attention of attention control.

Acknowledgement. The authors would like to thank Xianyu Chen for providing the source code for PathGAN. This project was supported by US National Science Foundation Awards IIS-1763981, IIS-2123920, NSDF DUE-2055406, and the SUNY2020 Infrastructure Transportation Security Center, and a gift from Adobe.

References

[1] 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. 2, 6

[2] 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. 2

[3] David J Berg, Susan E Boehnke, Robert A Marino, Douglas P Munoz, and Laurent Itti. Free viewing of dynamic stimuli by humans and monkeys. *Journal of vision*, 9(5): 19–19, 2009. 2

[4] Sage EP Boettcher, Dejan Draschkow, Eric Dienhart, and Melissa L-H Võ. Anchoring visual search in scenes: Assessing the role of anchor objects on eye movements during visual search. *Journal of vision*, 18(13):11–11, 2018. 7

[5] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):185–207, 2013. 2

[6] Ali Borji, Hamed R Tavakoli, Dicky N Sihite, and Laurent Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *ICCV*, 2013. 5

[7] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE transactions on image processing*, 24(12):5706–5722, 2015. 2

[8] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell us about saliency models? *IEEE transactions on pattern analysis and machine intelligence*, 41(3):740–757, 2018. 5

[9] Nicolas Carion, Francisco Massa, Gabriel Synnaeve, Nicolas Usunier, Alexander Kirillov, and Sergey Zagoruyko. End-to-end object detection with transformers. In *ECCV*, 2020. 3

[10] Xianyu Chen, Ming Jiang, and Qi Zhao. Predicting human scanpaths in visual question answering. In *CVPR*, 2021. 2, 3, 5, 6, 7, 8

[11] 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. 2, 4, 5

[12] Yupei Chen, Zhibo Yang, Souradeep Chakraborty, Sounak Mondal, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Characterizing target-absent human attention. In *CVPR Workshops*, 2022. 2, 5, 7, 8

[13] Bowen Cheng, Alex Schwing, and Alexander Kirillov. Per-pixel classification is not all you need for semantic segmentation. In *NeurIPS*, 2021. 8

[14] Bowen Cheng, Ishan Misra, Alexander G Schwing, Alexander Kirillov, and Rohit Girdhar. Masked-attention mask transformer for universal image segmentation. In *CVPR*, 2022. 3, 4

[15] Marvin M Chun and Yuhong Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive psychology*, 36(1):28–71, 1998. 4

[16] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018. 2

[17] Robert Desimone, John Duncan, et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995. 2

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *ICLR*, 2020. 3

[19] John M Findlay and Iain D Gilchrist. Visual attention: The active vision perspective. In *Vision and attention*, pages 83–103. Springer, 2001. 2

[20] Adam Gazzaley and Anna C Nobre. Top-down modulation: bridging selective attention and working memory. *Trends in cognitive sciences*, 16(2):129–135, 2012. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 4, 6

[22] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *ICCV*, 2015. 2

[23] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision research*, 40(10-12):1489–1506, 2000. 2

[24] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *CVPR*, 2016. 2

[25] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *CVPR*, 2015. 2, 6

[26] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *ICCV*, 2009. 5, 6, 7

[27] Anton S Kaplanyan, Anton Sochenov, Thomas Leimkühler, Mikhail Okunev, Todd Goodall, and Gizem Rufo. Deepfovea: Neural reconstruction for foveated rendering and video compression using learned statistics of natural videos. *ACM Transactions on Graphics (TOG)*, 38(6):1–13, 2019. 8

[28] Srinivas SS Kruthiventi, Kumar Ayush, and R Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9):4446–4456, 2017. 2

[29] Matthias Kümmerer and Matthias Bethge. State-of-the-art in human scanpath prediction. *arXiv preprint arXiv:2102.12239*, 2021. 5

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

[31] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *ICCV*, 2017. 2

[32] 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. 2, 5, 6

[33] Michael Land and Benjamin Tatler. *Looking and acting: vision and eye movements in natural behaviour*. Oxford University Press, 2009. 2

[34] Hei Law and Jia Deng. Cornernet: Detecting objects as paired keypoints. In *ECCV*, 2018. 5

[35] Yang Li, Si Si, Gang Li, Cho-Jui Hsieh, and Samy Bengio. Learnable fourier features for multi-dimensional spatial positional encoding. In *NeurIPS*, 2021. 4

[36] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 4

[37] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *ICCV*, 2017. 5

[38] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin transformer: Hierarchical vision transformer using shifted windows. In *ICCV*, 2021. 3, 4

[39] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *ICLR*, 2019. 6

[40] Stephen C Mack and Miguel P Eckstein. Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of vision*, 11(9):9–9, 2011. 7

[41] Christopher Michael Masciocchi, Stefan Mihalas, Derrick Parkhurst, and Ernst Niebur. Everyone knows what is interesting: Salient locations which should be fixated. *Journal of vision*, 9(11):25–25, 2009. 2

[42] 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, 6, 7

[43] Jiri Najemnik and Wilson S Geisler. Optimal eye movement strategies in visual search. *Nature*, 434(7031):387–391, 2005. 2

[44] Saul B Needleman and Christian D Wunsch. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of molecular biology*, 48(3):443–453, 1970. 5

[45] Klaus Oberauer. Working memory and attention—a conceptual analysis and review. *Journal of cognition*, 2019. 2

[46] Christian NL Olivers and Pieter R Roelfsema. Attention for action in visual working memory. *Cortex*, 131:179–194, 2020. 2

[47] Sohee Park, Arani Bhattacharya, Zhibo Yang, Mallesham Dasari, Samir R Das, and Dimitris Samaras. Advancing user quality of experience in 360-degree video streaming. In *IFIP Networking*, 2019. 8

[48] Sohee Park, Arani Bhattacharya, Zhibo Yang, Samir R Das, and Dimitris Samaras. Mosaic: Advancing user quality of experience in 360-degree video streaming with machine learning. *IEEE Transactions on Network and Service Management*, 18(1):1000–1015, 2021. 8

[49] René Ranftl, Alexey Bochkovskiy, and Vladlen Koltun. Vision transformers for dense prediction. In *ICCV*, 2021. 3

[50] Shima Rashidi, Krista Ehinger, Andrew Turpin, and Lars Kulik. Optimal visual search based on a model of target detectability in natural images. In *NeurIPS*, 2020. 2, 4

[51] Wanjie Sun, Zhenzhong Chen, and Feng Wu. Visual scanpath prediction using ior-roi recurrent mixture density network. *IEEE transactions on pattern analysis and machine intelligence*, 43(6):2101–2118, 2019. 2

[52] Antonio Torralba, Aude Oliva, Monica S Castelhano, and John M Henderson. Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. *Psychological review*, 113(4):766, 2006. 2, 7

[53] Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *ICML*, 2021. 3

[54] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NeurIPS*, 2017. 2, 3, 4

[55] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5): 2368–2378, 2017. 2

[56] Wenguan Wang, Jianbing Shen, Jianwen Xie, Ming-Ming Cheng, Haibin Ling, and Ali Borji. Revisiting video saliency prediction in the deep learning era. *IEEE transactions on pattern analysis and machine intelligence*, 43(1):220–237, 2019. 2

[57] Zhiguo Wang and Raymond M Klein. Searching for inhibition of return in visual search: A review. *Vision research*, 50 (2):220–228, 2010. 4

[58] JM Wolfe. Visual search. pashler, h.(ed.), attention, 1998. 2

[59] Jeremy M Wolfe and Todd S Horowitz. Five factors that guide attention in visual search. *Nature Human Behaviour*, 1(3):1–8, 2017. 2

[60] Enze Xie, Wenhai Wang, Zhiding Yu, Anima Anandkumar, Jose M Alvarez, and Ping Luo. Segformer: Simple and efficient design for semantic segmentation with transformers. In *NeurIPS*, 2021. 3

[61] Juan Xu, Ming Jiang, Shuo Wang, Mohan S Kankanhalli, and Qi Zhao. Predicting human gaze beyond pixels. *Journal of vision*, 14(1):28–28, 2014. 5, 7

[62] 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*, 2020. 2, 3, 5, 6, 8

[63] Zhibo Yang, Sounak Mondal, Seoyoung Ahn, Gregory Zelinsky, Minh Hoai, and Dimitris Samaras. Target-absent human attention. In *ECCV*, 2022. 2, 3, 4, 5, 6, 8

[64] AL Yarbus. Eye movements and vision plenum. *New York*, 1967. 2

[65] Gregory Zelinsky. A theory of eye movements during target acquisition. *Psychological review*, 115(4):787, 2008. 2

[66] Gregory Zelinsky, Zhibo Yang, Lihan Huang, Yupei Chen, Seoyoung Ahn, Zijun Wei, Hossein Adeli, Dimitris Samaras, and Minh Hoai. Benchmarking gaze prediction for categorical visual search. In *CVPR Workshops*, 2019. 2, 3, 5, 6

[67] Gregory J. Zelinsky, Yupei Chen, Seoyoung Ahn, Hossein Adeli, Zhibo Yang, Lihan Huang, Dimitrios Samaras, and

Minh Hoai. Predicting goal-directed attention control using inverse-reinforcement learning. *Neurons, Behavior, Data analysis, and Theory*, 5(2):1–9, 2021. 1

[68] 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):1–15, 2018. 6

[69] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. In *ICLR*, 2021. 3, 4, 6