Characterizing Target-absent Human Attention

Yupei Chen[1], Zhibo Yang[2], Souradeep Chakraborty[2], Sounak Mondal[2], Seoyoung Ahn[2],
Dimitris Samaras[2], Minh Hoai[2], Gregory Zelinsky[2]
[1]The Smith-Kettlewell Eye Research Institute, [2]Stony Brook University

Abstract

Human efficiency in finding a target in an image has attracted the attention of machine learning researchers, but what about when no target is there? Knowing how people search in the absence of a target, and when they stop, is important for Human-computer-interaction systems attempting to predict human gaze behavior in the wild. Here we report a rigorous evaluation of target-absent search behavior using the COCO-Search18 dataset to train state-of-the-art models. We focus on two specific aims. First, we characterize the presence of a target guidance signal in target-absent search behavior by comparing it to target-present guidance and free viewing. We do this by comparing how well a model trained on one type of fixation behavior (target-present, target-absent, free viewing) can predict behavior in either the same or different task. To compare target-absent search to free viewing behavior we created COCO-FreeView, a dataset of free-viewing fixations for the same images used in COCO-Search18. These comparisons revealed the existence of a target guidance signal in target-absent search, albeit one much less dominant compared to when a target actually appeared in an image, and that the target-absent guidance signal was similar to free viewing in that saliency and center bias were both weighted more than guidance from target features. Our second aim focused on the stopping criteria, a question intrinsic to target-absent search. Here we propose to train a foveated target detector whose target detection representation is sensitive to the relationship between distance from the fovea. Then combining the predicted target detection representation with other information such as fixation history and subject ID, our model outperforms the baselines in predicting when a person stops moving his attention during target-absent search.

1. Introduction

A mechanism of attention, long known to be central to how humans prioritize and select visual information [31–34], has recently attracted computer vision researchers seeking to reproduce this selection efficiency in machines. The most often used paradigm to study this efficiency is a visual search task, where efficiency is measured with respect to how many attention shifts are needed to detect a target in an image. But what about when a target is not there? People are also extremely efficient in knowing when to end a search. Understanding this stopping behavior would not only serve applications in human computer interaction, but also has basic research significance, and no human gaze prediction model would be complete without addressing these questions.

In this study we characterize attention during target-absent (TA) search with two aims. The first aim includes two aspects (Fig. 1). First, how is TA search distinct from target-present (TP) search? TP search is known to be strongly guided by a target object representation, as evidenced by the search target being fixated far sooner than a random object [14, 40]. Although a target object is not present in a TA image, this target guidance signal may still be influencing the allocation of attention in a TA search, as evidenced by fixation preferences for target-similar non-target objects [1, 47]. Here we report a comparison of TA and TP search behavior in the context of COCO-Search18. COCO-Search18 consists of 10 people searching 6202 natural images for each of 18 target objects (microwaves, cars, bottles, etc.), and is currently the largest dataset of search fixations available [5, 41]. Critically, COCO-Search18 is evenly divided between TA and TP images, creating data subsets large enough to train separate TA and TP models. If a target guidance signal exists in TA search, then a model trained on TP search fixations should be able to predict TA search fixations. This prediction success can be compared to predictions from a comparable model that is trained on TA fixations to predict TA fixations, thereby enabling an initial estimation of the relative contribution of target guidance in TA search. In a second related characterization we ask how TA search differs from free viewing behavior. Given that a target guidance signal may be small in a TA search task, the factors remaining that are available to guide attention may be those commonly studied in the context of a free-viewing task, things like bottom-up salience, faces, text, etc. [4, 10, 17, 18]. Re-applying our experimental logic,

we will distinguish TA search from free-viewing fixations by training models on one (e.g. TA search fixations) to predict the other (e.g., free-viewing fixations), and vice versa, as well as training and testing models on the same behavior to make relative comparisons. To perform this training on free-viewing fixations we created COCO-FreeView, which is a novel contribution of our work to the computational modeling of attention (see Sec. 3.1). Although there existed other large datasets of free-viewing fixations for the purpose of model training, the combination of COCO-FreeView and COCO-Search18 creates a unique opportunity to understand the differences between these two attention behaviors, and goal-directed attention more broadly. Our first aim of characterizations of TA search differentiate it from TP search and free viewing on the basis of a target guidance signal and bias signals broadly considered by saliency models. In our second aim of characterization of TA search behavior we ask a basic question intrinsic to the TA search task—when should it stop?

2. Related Work

Although early work addressed TA search as a largely random process [16,37], this may not to be the case [6]; e.g., non-target objects visually similar to the target are more likely to be fixated in TA search [1]. Indeed, [47] was able to classify the search target category, simply by analyzing visual features of objects that were fixated in a TA search. That study, while demonstrating *some* target guidance in TA search, used only two target categories and a search task of only four non-targets (target-similar or target-dissimilar). Also unlike TP search, where there is a vigorous modeling literature on the prediction of human fixation locations [8, 45], very few models have predicted fixation-density maps (FDMs) for TA search. The only one to our knowledge is a study by [9], who combined saliency, target features, and scene context to predict search fixations in natural scenes. However, their analysis of TA search behavior was limited to a demonstration that participants searching for people tend to confine their search to sidewalks and doorways. What is clear is that the target guidance signal in TA search, although sufficient to decode some targets under some conditions, is certainly weaker than TP search guidance. Thus it is challenging for models to predict TA FDMs meaningfully better than models of TP search.

When search was studied using simple arrays of objects, it was possible to exhaustively search all the objects to determine the absence of a target, although such exhaustive search patterns were not always observed [6, 38]. However, in the context of natural images the notion of an exhaustive search is ill-defined, as is the notion of an objectively countable number of objects in the image.

Broadening the question to stopping more generally, the decision-making literature interprets related TP and TA decisions as a race between an evidence accumulation process and a process based simply on the passage of time [11]. This latter process assumes the buildup of an internal signal over time since search start, eventually crossing a termination threshold.The factors affecting this time threshold are task dependent, but in the context of search they are believed to balance the expected rewards of finding a target with the expected costs of making a false negative [39], which translates roughly into how long a person is willing to search for a particular target (searching for a four-leafed clover in a clover field would likely end sooner than a search for a lost ring in the same field). The process of evidence accumulation has been engaged most often in the context of TP search, as demonstrated in the Target Acquisition Model (TAM) [44]. This model shifted a foveated retina over an image, bringing the high-resolution fovea to the target's location, extracting above-threshold evidence for the target and ending search with a TP judgement. In [48], the TAM model was extended to include TA search by adding a lower threshold to the evidence accumulation process. Target evidence can change from fixation to fixation, not only because the high-resolution fovea moves closer or farther from the target (important for TP search), but also because of inhibition of return (IOR) [23, 29] at previously fixated image locations that did not contain a target. IOR is important for TA search because its selective application to peak activity regions on the target evidence map eventually causes peak activation to drop below the TA threshold and for search to end. To our knowledge, [48] is the only evidence-based model of TA search for natural images and target categories.

3. Aim 1: Characterizing the Signal in Target-Absent Search

3.1. Approach

We significantly extend earlier TA fixation prediction work (which used methods like AdaBoost and SIFT features) [48] by leveraging state-of-the-art deep learning methods. We explore two deep network architectures, a ResNet50 (R50) [13] and a ResNet50 with a Feature Pyramid Network (R50+FPN) [27]. Given our goal of comparing TA search to both TP search and free viewing behavior, we manipulated whether models were trained on TA, TP, or free viewing fixations. Specifically, we trained ResNet50 and ResNet50+FPN versions of a model we refer to as DeepSearch (DS), using either TA or TP search fixations for training. Therefore, the DS_TA_R50+FPN model refers to DeepSearch with a ResNet50+FPN backbone trained on fixations during a TA search task. For fair comparisons to free-viewing behavior we include in our model comparison DeepFreeView (DFW), which similar to DeepSearch has both R50 and R50+FPN versions but is trained exclusively on free-viewing fixations. For both DS and DFW,

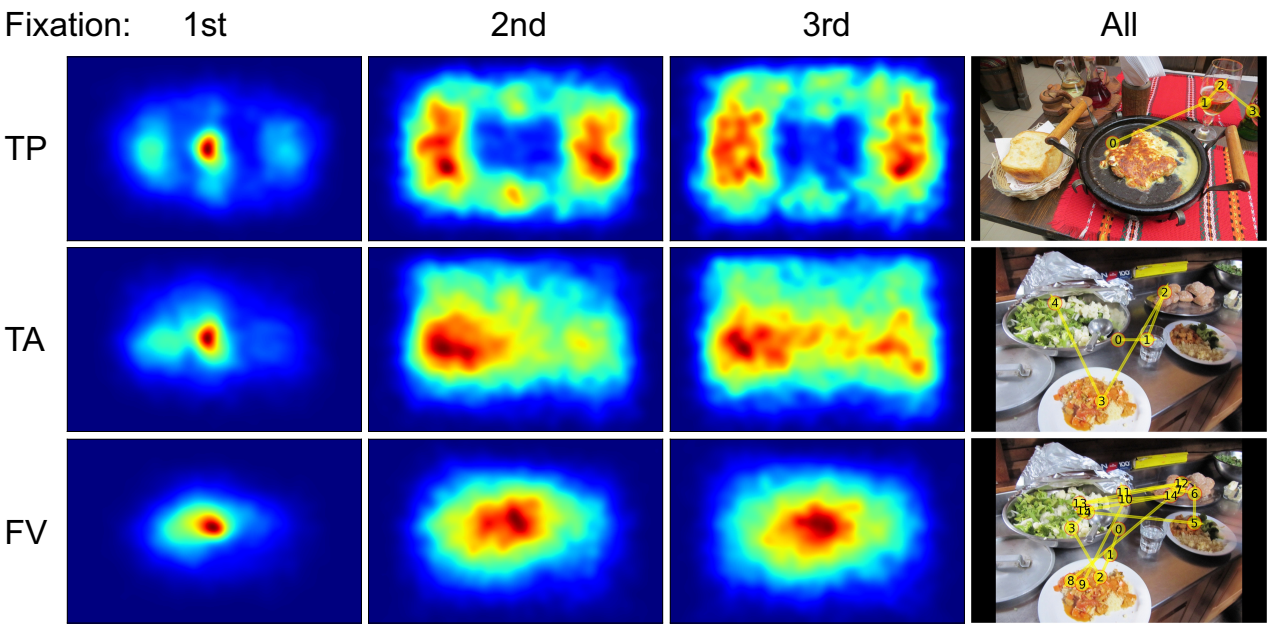

Figure 1. Fixation-density maps of the first three new fixations, with examples of complete scanpaths, for target-present search (TP), target-absent search (TA) and free viewing (FV) behaviors. Fixation patterns differ dramatically under different behavioral tasks.

we fine-tuned the pre-trained object detection models on the fixation density maps (FDM) from TP, TA, or free-viewing which output priority maps to predict human fixations. We also include DeepGaze II (DG2) [26], a top performer in a fixation prediction benchmark [20], in our model comparison, although one should note that DG2 was not trained and tested on the COCO-inspired datasets used in this study. We also used other models to establish baselines and ceilings on expected prediction success. For example, we used the TA search ground-truth to predict free viewing behavior, and the free viewing ground-truth to predict the TA search behavior (the same could not be done for TP search, which used different images). To obtain a soft noise ceiling on expected model success, we additionally include a human inter-observer consistency (IOC) upper-bound obtained by using the FDMs from half of the participants to predict those from the other half. The numbers in Table 1 are the averages of 10 random selections.

Creation of COCO-FreeView. To obtain the free viewing fixations used to train DeepFreeView we created the COCO-FreeView dataset. COCO-FreeView contains the same natural images used in COCO-Search18, but labeled with 822,602 eye fixations from a free-viewing task. Specifically, 10 university students participating in exchange for course credit viewed each image for 5 seconds in anticipation of a memory test. Eye position was recorded at 1000 Hz using an EyeLink 1000 commercial eyetracker (SR research Ltd.), and fixations were extracted offline using default settings. This effort required about 12 hours per participant, distributed over 6 roughly two-hour sessions. No identifying information is included in the dataset. The experiment was approved by IRB, and informed consent was obtained from each participant at the beginning of the first session.

3.2. Results

3.2.1 Qualitative Evaluation

Fig. 1 shows fixation density maps (FDMs; see Supplemental for implementation details) visualizing the spatial distribution of fixations, organized into a grid of rows indicating the tasks (TP top, TA middle, FV bottom) and columns corresponding to first (left), second (middle left) and third (middle right) new fixations during viewing. Note that the TA and free viewing FDMs were based on identical images (the TP images were largely different), and that only fixations from correct search trials were included in our analyses. We truncated our analysis to only the first 3 fixations made during viewing for a fairer task comparison. Specifically, the mean number of fixations needed for TP search judgments in COCO-Search18 was only 2.61, whereas for TA judgements the mean number of fixations was 5.02 and even greater for the COCO-FreeView dataset (M = 14.5). By restricting our core analyses to only the first three new fixations we can therefore fairly compare the early (and most critical) guidance signal in TA search to TP search and free viewing (see Supplemental for FDMs and analyses based on all fixations, showing largely similar patterns).

Task differences clearly emerge as soon as the first new

fixation, although all three tasks still show strong center bias due to the central starting gaze position. By the second new fixations, the center bias in the TP data is replaced by a highly bilateral distribution of attention, as expected by target placement constraints in COCO-Search18. More interestingly, TA search fixations were more spatially dispersed than free-viewing fixations, suggesting a more active exploration during search compared to free viewing. TA fixations also showed a left-biased distribution of attention that is distinct from the bilateral TP fixations and the center-biased free-viewing fixations. It appears that TA search can be viewed as a combination of TP search and free viewing (see also Sec. 3.2.2 and 3.2.3).

3.2.2 Target-absent vs. Target-present Search

To evaluate model success we used three well-accepted fixation prediction evaluation metrics: Area Under the Curve (AUC), Normalized Scanpath Saliency (NSS), and Correlation Coefficient (CC) (higher values for better prediction success, see [2,3] for details). Left and middle data columns of Table 1 compare TP and TA search. The DeepSearch models (DS) all outperformed DeepGaze II (with or without fixation map priors), and many met or exceeded the Human IOC. Among different architectures of DeepSearch, the R50-FPN backbone performed the best. As expected, models trained on TP fixations also best predicted TP search, and models trained on TA search best predicted TA fixations. More interestingly, TP models also predicted TA search fixations relatively well (TA→TA = NSS of 2.389, TP→TA = NSS of 2.049), accounting for 86% of the TA-predicting-TA performance. The converse was less true, where a TA-trained model achieved only 73% of the TP-predicting-TP performance, largely due to increased false positives. This suggests that models trained on TP search fixations captured some signal guiding TA search, which we hypothesize comes from the same target representation used to guide TP search. We tested this hypothesis by additionally training TA and TP R50-FPN models on the FDMs either with or without target labels. The TP model trained without target labels was 1.2 NSS units less than the model trained with target labels. The performance of the TA model also dropped when trained without labels, although again less so. This supports our hypothesis that the signal guiding TA search is originating from a target representation.

3.2.3 Target-absent Search vs. Free Viewing

The combination of COCO-Search18 and COCO-FreeView means that a large number of images now exist that are labeled for both target-absent search fixations and free viewing fixations. Thus, we were able to train models on free-viewing behavior to predict both free viewing and TA search behavior. We refer to this model as DeepFreeView (DFV), and again explore ResNet50 and ResNet50+FPN architectures. For comparison we used the same DS model trained on TA fixations described in the previous section to predict free-viewing fixations. Once again, Table 1 shows that DFV, a model trained on free-viewing fixations, is best at predicting free viewing (NSS = 2.603). As expected, DG2 also predicted free-viewing fixations quite well (NSS = 2.237). More interestingly, DFV poorly predicted TA search fixations (NSS = 1.521), and DS was equally poor in predicting free-viewing fixations. Also notable from Table 1 is that training on target labels did not affect DFV's predictions. While predictions from both the TA models and TP models worsened when target labels were removed during training, free-viewing model performance was not impacted, which is additional evidence of target labels providing guidance during TA and TP search behaviors that is unavailable during free viewing.

3.2.4 Weighting Features across Tasks

Large-scale datasets of search and free viewing behavior also allow us to analyze the role of different features in controlling attention in these tasks. To determine this task-based weighting we considered three incontrovertible sources of attention bias: (1) guidance from bottom-up feature contrast (saliency), (2) a center bias, and (3) guidance from target features. For a model of bottom-up saliency we used Graph-Based Visual Saliency (GBVS) [10, 24], which builds on [18] by using a graph-based approach to predict attention.To implement a center bias model (CB) we computed a 2D Gaussian map centered on image $I_c(x_0, y_0)$ and with a size determined by the image dimensions, following previous studies [28, 36]. More specifically: $CB_P = \frac{1}{\sigma_c\sqrt{2\pi}} \exp(-(P - I_c)^2 \Big/ 2\sigma_c^2)$, where CB_p denotes the Gaussian map value at image pixel P and σ_c is the standard deviation of the 2D Gaussian function. To obtain target features we used the object proposal component from MaskR-CNN [12] to compute a target map containing evidence for the target object. For all the three tasks, this was the MaskR-CNN object proposal bounding box in the image having a > 0.01 confidence that the object belongs to the target category. Although there is no "target" in free-viewing, this provides a useful contrast to the TA data to see how feature guidance changes with target designation. Finally, we applied a 2D Gaussian (σ = one-fourth of the box height, h_b) to the center of the confidence-selected bounding boxes and multiplied each by its object recognition confidence score to derive the target map. The Intersection over Union (IoU) of bounding boxes with ground truth target object labels from COCO-Search18 was 0.826 for TP search, validating our use of the MaskRCNN method.

Fig. 2 evaluates the role of saliency, target guidance, and center bias in predicting ground-truth FDMs. Reported

Table 1. FDM predictions for target-present (TP) search, target-absent (TA) search, and free viewing (FV), evaluated using AUC, NSS, and CC. Rows from top to bottom represent DeepGaze II (DG2), DeepSearch (DS), DeepFreeView (DFV), and baseline models.

	TP Search			TA Search			FreeView		
	AUC	NSS	CC	AUC	NSS	CC	AUC	NSS	CC
Fixation prior	0.772	0.926	0.156	0.792	0.983	0.246	0.839	0.858	0.227
DG2 w/o prior	0.846	1.665	0.226	0.834	1.303	0.280	0.869	2.136	0.433
DG2 with prior	0.855	1.785	0.242	0.845	1.406	0.303	0.878	2.237	0.456
DS_TP_R50	0.943	4.338	0.635	0.877	1.890	0.429	-	-	-
DS_TP_R50+FPN	**0.950**	**4.621**	**0.675**	0.883	2.049	0.462	-	-	-
DS_TP_R50+FPN w/o	0.923	3.412	0.507	0.873	1.728	0.399	-	-	-
DS_TA_R50	0.918	3.135	0.468	0.897	2.250	0.506	0.847	1.426	0.338
DS_TA_R50+FPN	0.931	3.362	0.510	**0.903**	**2.389**	**0.540**	0.845	1.504	0.359
DS_TA_R50+FPN w/o	0.895	2.308	0.360	0.893	2.141	0.489	0.856	1.624	0.389
DFV_R50	-	-	-	0.843	1.463	0.336	0.909	2.440	0.559
DFV_R50+FPN	-	-	-	0.846	1.487	0.336	**0.914**	**2.603**	**0.588**
DFV_R50+FPN w/o	-	-	-	0.848	1.521	0.345	**0.914**	2.580	0.586
TA/FW on FW/TA	-	-	-	0.770	1.312	0.281	0.770	1.300	0.281
Human IOC	0.921	5.306	0.661	0.863	2.437	0.433	0.859	2.578	0.456

values are NSS scores, normalized by row (values add to 1) to illustrate their relative importance in the task. We again limit this analysis to the first 3 new fixations, for fair comparison. Most conspicuous, and least surprising, is the dominant role played by target features in guiding attention during TP search. This was already true by the first new fixation, and progressively grew over the next two. This dominance came largely at the expense of center bias. Saliency, although consistently higher, was still weighted far less than target features, as in previous work (e.g., [43]). In contrast, for TA search and free viewing these target features played less of a role in predicting attention. Between TA search and free viewing, differences become more subtle but do exist. Most relevant to the current question is that target features were weighted significantly more in TA search compared to free viewing, consistent with our other evidence for a weak target guidance signal in TA search. All claims were based on paired t-test with a $p_{bonferroni} < .016$ (details in Supplemental).

3.2.5 Classifying Task from Scanpaths

Our previous analyses compared attention behavior across TP search, TA search, and free viewing, and current state-of-the-art in predicting fixation behavior in these tasks. We found that models trained on one task (e.g., TA) often did a poor job in predicting another (e.g., TP), based on FDMs, i.e. aggregations of distributed fixation locations. Here we ask whether these differences suffice to classify one task from another based only on the fixations in a scanpath.

We adopt a sequence modeling approach. Our specific task is to classify scanpaths into TP, TA and free viewing categories based on the fixation locations (2D coordinates $X \in \mathbb{R}$ and $Y \in \mathbb{R}$), durations $D \in \mathbb{R}$, and visual features corresponding to high-resolution panoptic FPN maps (as in [22, 41]) to construct representations referred to as Dynamic Contextual Belief (DCB) maps [41]. B denotes these maps, $B \in \mathbb{R}^{H \times W \times C}$ where H, W are spatial dimensions and $C = 134$, the number of COCO thing+stuff categories. For a given fixation, we index B along the spatial dimensions to get the visual features $V \in \mathbb{R}^C$ for this spatial location. The final feature vector for fixation at time step i, denoted by $F_i \in \mathbb{R}^{C+3}$ at location X_i, Y_i with duration D_i, is thus:$F_i = [X_i; Y_i; D_i; V_i]$, where $V_i = B[Y_i, X_i]$. For sequence modeling, we use a Long Short-Term Memory (LSTM) [15] architecture with three LSTM layers stacked on top of each other. A fully connected layer followed by a softmax layer are applied on *each* time step representation from the final LSTM layer to get classification probabilities.

Prediction of gaze behavior for every partial scanpath is necessary to investigate similarities and differences between TA, TP, and FV behaviors after every intermediate fixation. We observe that the model can discern different behaviors with high accuracy even for a 3-fixation partial scanpath. Fig. 3a shows the confusion matrix for making classification decisions based on the first three fixations of a scanpath. The classification accuracy (average of the diagonal values of the confusion matrix) is 67.53%, which is much higher than 33.33% chance. Fig. 3b shows that the classification accuracy increases as more fixations are ob-

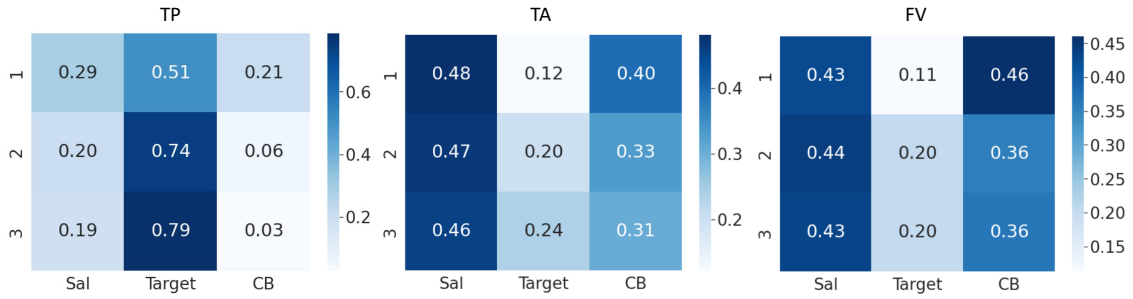

Figure 2. Normalized NSS weights for saliency (Sal), target (Target), and center bias (CB) features best predicting the ground-truth FDMs for the first three new fixations (rows) in target-present (left) and target-absent (middle) search and free viewing (right).

served and used for classification. Fig. 3b also compares the performance for the classification model with different types of input features. As can be seen, visual features are more crucial for better performance than fixation duration. Fig. 3c & d further analyze the distributions of the prediction outputs for TA and FV data. Clear from this analysis is that the model does not solely rely on the scanpath length for making the classification decision.

4. Aim 2: Predicting Stopping in Target-Absent Search

As discussed, existing models of TA stopping assume knowledge of some internal termination function, and do not offer computational solutions that can learn this function and apply it to natural image search tasks. To address this, we develop a method to predict when a TA search will stop based on previous behavioral states. Our approach adopts the cognitively meaningful evidence-accumulation stopping heuristic suggested in [44, 48], which is based on dynamically accumulating (with each shift of a simulated fovea over an image) evidence for a target on a target map.

4.1. Approach

Due to the neuroanatomy of the primate foveated retina, visual acuity lessens with increasing distance from the high-resolution central fovea. This eccentricity-dependent blur can be formulated as the probability that an actual target pixel belonging to the target, decreasing as the distance between this pixel and the current fixation location increases. Conversely, the probability of a non-target pixel belonging to a target tends to increase when moving away from the fixation point. In preliminary work we found that existing pretrained object detectors are largely insensitive to eccentricity-based blurring, and thus do not capture the human distinction between foveal and peripheral viewing. [35] addressed this problem by using a model of target detectability based on the feature distributions of TP and TA images with five predefined eccentricities. However, manual creation of a dataset for each target is impractical for larger numbers of target categories (as in COCO-Search18 with 18 target classes). Instead, our approach is to train a foveated target detector whose target detection representation is sensitive to the relationship between distance from the fovea (retinal eccentricity) and the degradation in resolution for visual targets. Notably, we train a single object detector for all 18 target categories in COCO-Search18, and based on this detection map, which dynamically changes from one fixation to the next, we predict search termination. We do this by training a simple neural network to predict stopping by treating it as a binary classification problem.

Foveated Target Detector. Object detectors are typically trained with full-resolution images as input and binary object masks as annotations. In our case, the input is a cumulative foveated image [42] where progressively greater blur was applied to pixels having larger eccentricity (i.e., appearing farther in peripheral vision). To make our target detector sensitive to fovea-induced blur (i.e., detection likelihood for true target pixels decreasing with eccentricity but false target pixels increasing with eccentricity), we apply these relationships on the binary masks (labels), rendering the masks continuous with values from 0 to 1, with 1 indicating the target pixels with the highest level of confidence. Thus, our model has different detection confidence levels with respect to eccentricity.

To create these continuous "retina" masks, we follow the foveation algorithm in [19, 30, 42] by creating a six-level pyramid of the label maps, where the lowest level is a binary mask (1 for the target pixels and 0 for the non-target pixels) and for each higher level the values for the target and non-target pixels linearly decrease or increase, respectively. A final continuous label map is obtained through weighted combination of all label maps in the pyramid. The same gaze-contingent weight map was used for image foveation (see Fig. 4 for an example and Supplemental for further illustration). So our label foveation builds on the log-linear model of eccentricity in [35], but extends it significantly.

Fig. 4 shows an overview of the training pipeline for our target detector. In particular, we follow [41, 42, 46] and dis-

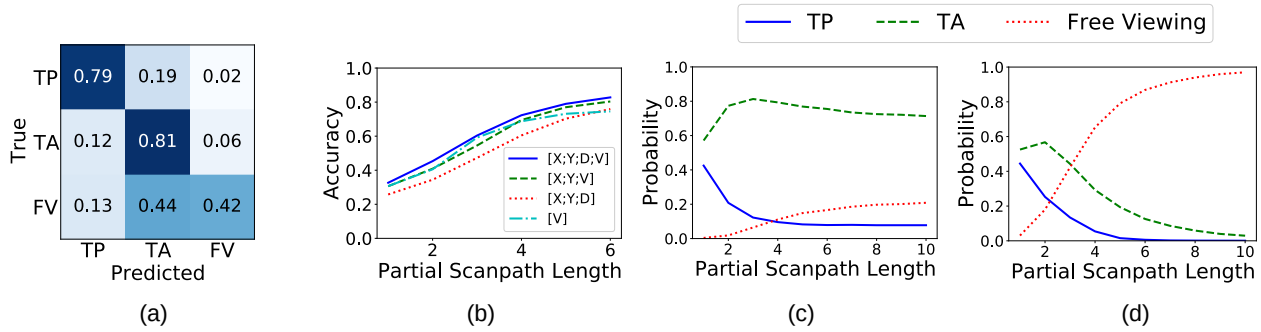

Figure 3. (a) Confusion matrix of LSTM-model's predictions for partial scanpaths comprised of only the first three fixations. The average of the diagonal elements of the confusion matrix is 67.53%. (b) Classification accuracy as a function of the input features and the length of the partial scanpath used. (X,Y), D, V are the fixation location, fixation duration, and visual feature at the fixated location respectively. (c)&(d): distribution of model predictions for classifying scanpath data from either TA (c) or FV (d) behaviors. See the supplementary material for the plot for TP data.

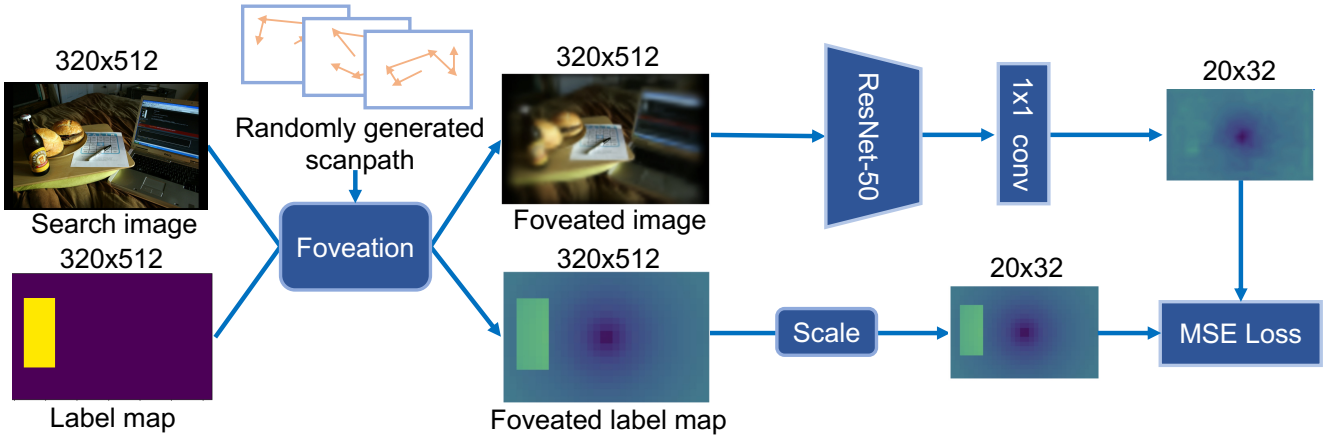

Figure 4. Training pipeline of the foveated target detector. Depicted shows an example of bottle search; similar for other target classes. We train the target detector with (cumulative) foveated images and label maps using randomly generated scanpaths. We use mean square error (MSE) loss to train the network. Yellow and blue in the heatmaps (label maps) denote 1 and 0, respectively, with yellower pixels indicating larger values.

cretize the fixations into a 20×32 grid. We input a 320×512 foveated image to the network and output a 20×32 detection map. We use the first 4 convolution blocks of the ImageNet [7] pre-trained ResNet-50 [13] followed by a 1×1 convolution layer to map the feature maps to a 20×32 target detection map. We then train the network using the mean square error between the detection map and the down-scaled foveated label map.

Termination Predictor. To predict TA search termination, we not only rely on the detection map obtained from the foveated target detector but also information about: 1) history of fixation locations (encoding coverage of the search space); 2) subject ID (as different subjects likely have different termination criteria), and; 3) target ID (as some targets generate stronger guidance signals than others, which could be used in predicting termination). Hence, we use a two-layer MLP to embed the 20×32 history fixation map (1 at the fixated locations and 0 elsewhere) into a vector, and use a trainable encoding vector for each subject and each target. Finally, we concatenate these embedding vectors and input them into a two-layer MLP for the termination prediction and use binary cross-entropy loss to train the termination predictor.

4.2. Results

We evaluate our model on the TA trials of COCO-Search18, based on a random split of the dataset into 70% training, 10% validation, and 20% testing sets, within each target category. We report precision, recall, F1-score and average precision (when applicable).

Implementation Details. The linear slope used to create the label pyramid for our foveated target detector is set to

Table 2. Results of target-absent search termination prediction.

	Precision	Recall	F1-score	Average Precision
Avg. scanpath length	0.233	0.361	0.283	-
Subject-specific avg. scanpath length	0.297	0.347	0.320	-
DCB-based model	0.393	0.431	0.411	0.387
Our model	**0.402**	**0.543**	**0.462**	**0.424**

0.1, so the label map at the highest level of the label pyramid is 0.5 everywhere corresponding to the largest eccentricity. We first train the foveated target detector on the TP image training set in COCO-Search18 with randomly generated scanpaths (up to 5 fixations, approximating the average scanpath length). The randomly generated scanpath can be viewed as a data augmentation scheme to prevent overfitting. Then we train the termination predictor with the training TA trials of COCO-Search18, while keeping the foveated target detector fixed. A dropout layer ($p = 0.5$) is attached to every linear layer (except the last layer of the termination predictor) to prevent overfitting. Hidden size of MLP layers and embedding size of the history, subject and target in the termination predictor are all set to 32. We train the networks using an Adam optimizer [21] with learning rate of 10^{-4} and a decay rate of 10^{-8} (more details in Supplemental).

Baselines. We compare our method with: (1) **Avg. scanpath length**: search stops when number of fixations is larger than average scanpath length (i.e., simple time-based stopping). (2) **Subject-specific avg. scanpath length**: similar to Avg. scanpath length, but the stopping criteria is specific for each subject. (3) **DCB-based model**: to evaluate the foveated target detector, we replace it with a 1×1 convolution layer that inputs the dynamic contextual beliefs (DCB) proposed in [41], and then train the termination predictor following our method.

Experimental Results. Tab. 2 gives termination prediction results for models and metrics. Our model outperforms all baselines at all metrics, demonstrating that it is not trivial. Our method also outperformed the DCB-based baseline in average precision, indicating that our foveated target detector better characterizes the change of target evidence over the course of search. In addition, the fact that subject-specific avg. scanpath length performs better than the avg. scanpath length suggests that different subjects could use different termination criteria in TA search. Please see Supplemental for further ablation study.

5. Discussion, Limitations, and Broader impacts

Our first aim in this study was to characterize similarities and differences between TA search fixations and fixations made during TP search and free viewing. To make possible the direct comparison of search fixations (available from COCO-Search18) to free-viewing fixations we created COCO-FreeView. We show the existence of a target guidance signal even when the target is not there. However, we also show that this target guidance is weak compared to TP search, and indeed factors such as saliency and center bias play even larger roles in predicting TA FDMs. From this we conclude that TA search behavior is a blend of TP search and free viewing, a claim that we supported in computational experiments in which we trained and compared models of FDM prediction and viewing behavior classification.

Our second aim was to predict when people stop a TA search. Grounded in a theory of stopping based on target evidence accumulation, we proposed a termination predictor based on a foveated target detector that outputs a dynamically evolving detection map, one that is sensitive to the degradation in visual resolution that occurs with increasing retinal eccentricity. We showed that our method outperformed the baselines by a large margin, highlighting the importance of accurate stopping prediction in any complete understanding of attention during search. Our work paves a path for future work on this neglected question that has applications ranging from efficient HCI to gaze-based annotation creation.

A limitation of our work is that it did not disentangle internal and external factors potentially contributing to the TA guidance signal. We demonstrated the importance of image features in creating a target guidance, but the internal cognitive state of a person, such as short-term and long-term memory, can also influence guidance and this is not captured by our model. In future work we hope to extend our approach to include guidance from scene semantics in TA search, which we believe to be another question ripe for engagement by state-of-the-art attention prediction methods. We also plan to extend our search stopping model to scanpath prediction, making our model more aligned with the latest fixation prediction models in TP search and free viewing [25, 41].

References

[1] Robert G Alexander and Gregory J Zelinsky. Visual similarity effects in categorical search. *Journal of vision*, 11(8):9–9, 2011. 1, 2

[2] Ali Borji, Hamed R. Tavakoli, Dicky N. Sihite, and Laurent Itti. Analysis of scores, datasets, and models in visual saliency prediction. 2013. 4

[3] 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. 4

[4] Moran Cerf, E Paxon Frady, and Christof Koch. Faces and text attract gaze independent of the task: Experimental data and computer model. *Journal of vision*, 9(12):10–10, 2009. 1

[5] 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. 1

[6] Marvin M Chun and Jeremy M Wolfe. Just say no: How are visual searches terminated when there is no target present? *Cognitive psychology*, 30(1):39–78, 1996. 2

[7] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. Ieee, 2009. 7

[8] Miguel P Eckstein. Visual search: A retrospective. *Journal of vision*, 11(5):14–14, 2011. 2

[9] Krista A Ehinger, Barbara Hidalgo-Sotelo, Antonio Torralba, and Aude Oliva. Modelling search for people in 900 scenes: A combined source model of eye guidance. *Visual cognition*, 17(6-7):945–978, 2009. 2

[10] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. 2007. 1, 4

[11] Guy Hawkins and Andrew Heathcote. Racing against the clock: Evidence-based vs. time-based decisions. *Psychol Rev.*, (2):222–263, 2019. 2

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask r-cnn. 2017. 4

[13] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. 2016. 2, 7

[14] John M Henderson, Phillip A Weeks Jr, and Andrew Hollingworth. The effects of semantic consistency on eye movements during complex scene viewing. *Journal of experimental psychology: Human perception and performance*, 25(1):210, 1999. 1

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997. 5

[16] 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

[17] Laurent Itti and Christof Koch. Computational modelling of visual attention. *Nature reviews neuroscience*, 2(3):194–203, 2001. 1

[18] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on pattern analysis and machine intelligence*, 20(11):1254–1259, 1998. 1, 4

[19] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015. 6

[20] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012. 3

[21] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 8

[22] Alexander Kirillov, Ross Girshick, Kaiming He, and Piotr Dollár. Panoptic feature pyramid networks. 2019. 5

[23] Raymond Klein. Inhibitory tagging system facilitates visual search. *Nature*, 334(6181):430–431, 1988. 2

[24] Chrishtoph Koch. Saliency map algorithm : Matlab source code. `http://people.vision.caltech.edu/~harel/share/gbvs.php`. Accessed: 2010-09-30. 4

[25] M Kümmerer, TS Wallis, and M Bethge III. Deepgaze iii: Using deep learning to probe interactions between scene content and scanpath history in fixation selection. In *Conference on Cognitive Computational Neuroscience*, 2019. 8

[26] Matthias Kümmerer, Thomas S.A. Wallis, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. 2017. 3

[27] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2017. 2

[28] Sophie Marat, Anis Rahman, Denis Pellerin, Nathalie Guyader, and Dominique Houzet. Improving visual saliency by adding 'face feature map'and 'center bias'. *Cognitive Computation*, 5(1):63–75, 2013. 4

[29] Koorosh Mirpour, Fabrice Arcizet, Wei Song Ong, and James W Bisley. Been there, seen that: a neural mechanism for performing efficient visual search. *Journal of neurophysiology*, 102(6):3481–3491, 2009. 2

[30] Jeffrey S Perry and Wilson S Geisler. Gaze-contingent real-time simulation of arbitrary visual fields. In *Human vision and electronic imaging VII*, volume 4662, pages 57–69. International Society for Optics and Photonics, 2002. 6

[31] Steven E Petersen and Michael I Posner. The attention system of the human brain: 20 years after. *Annual review of neuroscience*, 35:73–89, 2012. 1

[32] Michael I Posner. Attention: the mechanisms of consciousness. *Proceedings of the National Academy of Sciences*, 91(16):7398–7403, 1994. 1

[33] Michael I Posner. *Cognitive neuroscience of attention*. Guilford Press, 2011. 1

[34] Michael I Posner and Steven E Petersen. The attention system of the human brain. *Annual review of neuroscience*, 13(1):25–42, 1990. 1

[35] Shima Rashidi, Krista Ehinger, Andrew Turpin, and Lars Kulik. Optimal visual search based on a model of target detectability in natural images. *Advances in Neural Information Processing Systems*, 33, 2020. 6

[36] Na Tong, Huchuan Lu, Lihe Zhang, and Xiang Ruan. Saliency detection with multi-scale superpixels. *IEEE Signal Processing Letters*, 21(9):1035–1039, 2014. 4

[37] Anne M Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive psychology*, 12(1):97–136, 1980. 2

[38] Jeremy M Wolfe. What can 1 million trials tell us about visual search? *Psychological Science*, 9(1):33–39, 1998. 2

[39] Jeremy M Wolfe. Guided search 6.0: An updated model of visual search. *Psychonomic Bulletin & Review*, pages 1–33, 2021. 2

[40] Hyejin Yang and Gregory J Zelinsky. Visual search is guided to categorically-defined targets. *Vision research*, 49(16):2095–2103, 2009. 1

[41] 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. 2020. 1, 5, 6, 8

[42] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019. 6

[43] Gregory Zelinsky, Wei Zhang, Bing Yu, Xin Chen, and Dimitris Samaras. The role of top-down and bottom-up processes in guiding eye movements during visual search. In *Advances in neural information processing systems*, pages 1569–1576. Citeseer, 2006. 5

[44] Gregory J Zelinsky. A theory of eye movements during target acquisition. *Psychological review*, 115(4):787, 2008. 2, 6

[45] Gregory J Zelinsky, Yupei Chen, Seoyoung Ahn, and Hossein Adeli. Changing perspectives on goal-directed attention control: The past, present, and future of modeling fixations during visual search. *Gazing Toward the Future: Advances in Eye Movement Theory and Applications*, 73:231, 2020. 2

[46] 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*, 2021, 2021. 6

[47] Gregory J Zelinsky, Yifan Peng, and Dimitris Samaras. Eye can read your mind: Decoding gaze fixations to reveal categorical search targets. *Journal of vision*, 13(14):10–10, 2013. 1, 2

[48] Wei Zhang, Hyejin Yang, Dimitris Samaras, and Gregory Zelinsky. A computational model of eye movements during object class detection. 2005. 2, 6