ORIGINAL ARTICLE
Journal Section

Predicting Goal-directed Attention Control Using Inverse-Reinforcement Learning

Gregory J. Zelinsky[1,2] | Yupei Chen[1] | Seoyoung Ahn[1] | Hossein Adeli[1] | Zhibo Yang[2] | Lihan Huang[2] | Dimitrios Samaras[2] | Minh Hoai[2]

[1]Department of Psychology, Stony Brook University, Stony Brook, NY, 11794, USA

[2]Department of Computer Science, Stony Brook University, Stony Brook, NY, 11794, USA

Correspondence
Gregory J. Zelinsky, Department of Psychology, Stony Brook University, Stony Brook, NY 11794, USA
Email: gregory.zelinsky@stonybrook.edu

Funding information
National Science Foundation, Award Number: IIS-1763981

Understanding how goals control behavior is a question ripe for interrogation by new methods from machine learning. These methods require large and labeled datasets to train models. To annotate a large-scale image dataset with observed search fixations, we collected 16,184 fixations from people searching for either microwaves or clocks in a dataset of 4,366 images (MS-COCO). We then used this behaviorally-annotated dataset and the machine learning method of inverse-reinforcement learning (IRL) to learn target-specific reward functions and policies for these two target goals. Finally, we used these learned policies to predict the fixations of 60 new behavioral searchers (clock = 30, microwave = 30) in a disjoint test dataset of kitchen scenes depicting both a microwave and a clock (thus controlling for differences in low-level image contrast). We found that the IRL model predicted behavioral search efficiency and fixation-density maps using multiple metrics. Moreover, reward maps from the IRL model revealed target-specific patterns that suggest, not just attention guidance by target features, but also guidance by scene context (e.g., fixations along walls in the

search of clocks). Using machine learning and the psychologically meaningful principle of reward, it is possible to learn the visual features used in goal-directed attention control.

KEYWORDS
top-down attention, attention models, visual search, fixation prediction, eye behavior, reinforcement learning

1 | INTRODUCTION

Ever since Yarbus' seminal demonstration of how a goal can control attention [24], understanding goal-directed attention control has been a core aim of psychological science. This focus is justified. Goal-directed attention underlies everything that we *try* to do, making it key to understanding cognitively-meaningful behavior. Like Yarbus, we too demonstrate goal-directed control of eye-movement behavior, but here these overt attention movements are made by a deep-network model that has learned different goals.

Three factors distinguish our approach from previous work. First, it is image-computable and uses learned, rather than handcrafted, features. Our model therefore inputs an image but is not told anything about its features ("vertical", "clock", etc.), which all must be learned. This factor distinguishes the current model from most others in the behavioral literature on attention control [22, 25, 3], and makes our approach more aligned with recent computational work. [28, 27] Second, the goal-directed behavior that we study is categorical search, the visual search for any exemplar of a target-object category. [20, 8, 26] We adopt this paradigm because categorical search is the simplest (and therefore, best) goal-directed behavior to computationally model—there is a target-object goal and the task is to find it. A third and unique contribution of our approach is that we predict categorical-search fixations using a policy that was learned, through many observations of search-fixation behavior during training, to maximize the goal-specific receipt of reward. Using inverse-reinforcement learning (IRL), we obtain these reward functions and use them to prioritize spatial locations to predict the fixations made by new people searching for the learned target categories in new images. Doing this required the creation of a search-fixation-annotated image dataset sufficiently large to train deep-network models (see Methods). We show that this model successfully captured several patterns observed in goal-directed search behavior, not the least being the guidance of overt attention to the target-category goals.

1.1 | Inverse-Reinforcement Learning

Reinforcement learning, like supervised and unsupervised learning, is a basic machine learning method where agents make actions in a context for the purpose of maximizing the accumulation of reward. In reinforcement learning, knowledge of the reward is assumed and the goal is to predict the action, which in the current context is the search saccade. In inverse-reinforcement learning, knowledge of the action is assumed and the goal is to learn the reward function. In the current context, this knowledge of the action corresponds to training on the search fixation behavior.

IRL is an imitation-learning method from the machine-learning literature that learns, through observations of an expert, a reward function and policy for mimicking expert performance. We extend this framework to goal-directed behavior by assuming that the image locations fixated by searchers constitute the expert performance that the model learns to mimic. The specific IRL algorithm that we use is Generative Adversarial Imitation Learning (GAIL [11]), which

makes reward proportional to the model's ability to generate state-action pairings that imitate observed state-action pairings. Here, the action is a shift of fixation location in a search image (the model's saccade), and the state is the search context (all the information available for use in the search task). The state includes, but is not limited to, the visual features extracted from an image and the learned visual representation of the target category. Over training, and through the greedy maximization of total-expected reward, the model learns a policy for mapping states to actions that can be used to predict new actions (saccades) given new states (search images).

2 | METHODS

2.1 | Model Methods

The IRL model framework is illustrated in Fig. A1. Model training can be conceptualized as a policy generator (G) and a discriminator (D) locked in an adversarial process [11]. The generator generates fake eye movements (actions) with the goal of fooling the discriminator into believing that these actions were made by a person, while the discriminator's goal is to discriminate the real eye movements from the fake. More specifically, the generator consists of an actor-critic model [14] that learns a policy for maximizing total expected reward over all possible sequences of fixations, with greater reward given to the generator when it produces person-like actions that the discriminator miss-classifies as real (the logarithm of the discriminator output). This reward-driven adversarial process plays out during training using proximal policy optimization (PPO) [21], with the result being a generator that becomes highly adept at imitating the behavioral fixations made during categorical search. At testing, this learned Policy for mimicking people's categorical search fixations is used to predict the fixation behavior of new people searching for the same target categories in new images. These fixation predictions are quantified by what we call a *saccade map*, which is a priority map reflecting the total reward expected if saccades were to land at all the different locations in an image input. Note that this reward-based prioritization, by imitating the behavioral search fixations, captures the pursuit of reward that we assume to be driving search behavior, along with any other biases that systematically affect gaze control during search. So, although the IRL model and people are rewarded for different things, the end result is that the model recovers the reward functions that people use to guide their search behavior.

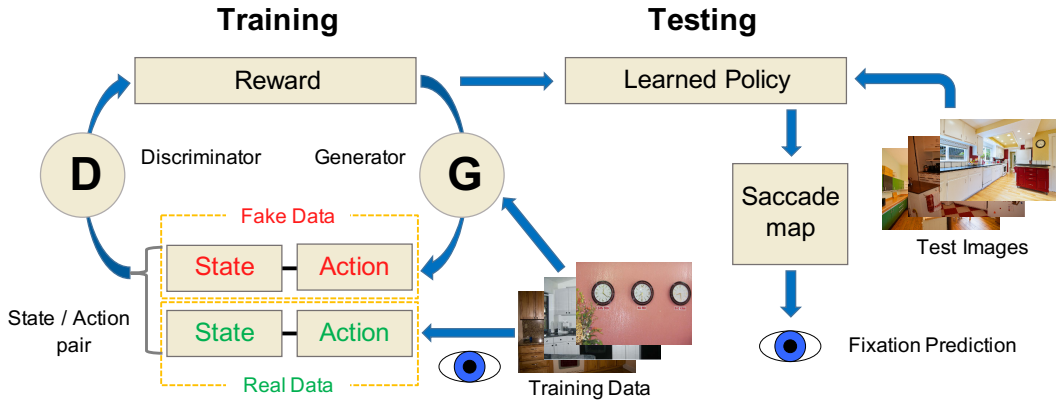

FIGURE 1 The model's adversarial imitation learning algorithm. During training it learns from fixation-annotated images a reward function and policy for predicting new search fixations in unseen test images.

2.2 | States and Actions: Cumulative Movements of a Foveated Retina

Broadly speaking, the state is the internal visual representation that is used for search, and a big part of this are the features extracted from the image input. To obtain a robust core state representation we pass each image through a pre-trained ResNet-50 [10] to get a reasonably-sized feature map output (2048x10x16). However, human search behavior is characterized by movements of a foveated retina, and each of these search fixations dramatically changes the state by re-positioning the high-resolution fovea in the visual input. We captured this fixation-dependent change in state in two steps. First, we gave the IRL model a simplified foveated retina. We did this using the method from Geisler and Perry [19] to compute a retina-transformed version of the image input (*ReT-image*), which in our implementation is an image having high resolution within a central 3° "fovea" (32x32 in the resized 512x320 pixel image) but is blurred outside of this fovea to approximate the loss of resolution that occurs with increasingly eccentric viewing in the visual periphery. Second, we accumulate these high-resolution foveal views, each a different ReT-image, over 6 new fixations in a process that we refer to as *cumulative foveation*. With each new "eye movement", the fovea is re-positioned in the image, thereby progressively de-blurring what was an initially fairly blurred visual input. Note that by adopting this cumulative-foveation state encoder we are not suggesting that people have a similar capacity to maintain high-resolution visual information once the fovea moves on, and indeed this is known not to be the case [12]. Rather, we used this fixation-by-fixation state encoder simply as a tool to integrate a dynamically changing state into the IRL method. Figure 2 shows cumulative ReT-images obtained at three successive fixation locations (0,1,2) for a sample scene, with the 7-fixation sequence of these images comprising a dynamic state representation that is input to the IRL model. The pre-trained ResNet-50 was dilated and fine-tuned on ReT-images prior to this state encoding. See the appendix on supplemental model methods for additional details.

FIGURE 2 The formation of a cumulative retina-transformed image over the first three fixations (0,1,2).

The IRL model learns to associate states with actions, but these actions must be defined in some space. We obtain an action space by first resizing a ReT-image input to 512x320 pixels, which we then discretize into a 10x16 grid of 32x32 pixel cells. The center of each cell becomes a potential fixation location, a computational necessity imposing a resolution limit on the model's oculomotor behavior. For each of the 6 new fixations generated by the model, the cumulative ReT-image input is prioritized by the saccade map and one of the 160 possible grid locations is selected for an eye movement.

2.3 | The Microwave-Clock Search Dataset

The currently most predictive models of complex fixation behavior are in the context of a free-viewing task, where the best of these models (e.g., DeepGaze II [15]) are pre-trained on SALICON [13]. SALICON is a crowd-sourced dataset consisting of images that were annotated with human mouse clicks indicating salient image locations. Without SALI-

CON, DeepGaze II and models like it would not have been possible, and our understanding of free-viewing behavior, widely believed to reflect bottom-up attention control (i.e., control solely by features extracted from the visual input), would be diminished. To date, however, there has been no comparable dataset for categorical search, and this has hindered the computational modeling of goal-directed attention control. Those suitably-sized and fixation-annotated image datasets that do exist either did not use a standard search task [17, 18], used a search task but had people search for multiple targets simultaneously [9], or used only one target category (people) [7]. Here we introduce the Microwave-Clock Search (MCS) dataset, which is now among the largest datasets of images that have been annotated with goal-directed fixations. The MCS dataset makes it possible to train deep-network models on human search fixations to predict how people will move their attention in the pursuit of target-object goals.

FIGURE 3 Representative training images (top) and testing images (bottom) in the MCS dataset.

Half of the MCS dataset consists of COCO2014 images [16] depicting either a microwave or a clock (based on COCO labels), from which we created disjoint training and testing datasets. The training dataset was selected entirely from COCO's training images; the testing dataset was selected from COCO's training and validation images (in order to maximize the number of images selected). In selecting the training images we excluded scenes depicting people and animals (to avoid attention biases to these categories), and digital clocks in the case of the clock target category. This latter constraint was introduced because the features of analog and digital clocks are very different, and we were concerned that this would introduce unwanted variability in the search behavior. No additional exclusion criteria were used to select the training images, with our goal being to include as many images for training as possible. These criteria left 1,494 analog clock images and 689 microwave images, which we should note varied greatly in terms of their search difficulty (see Fig. 3, top). It was necessary to tolerate this variability in order to obtain a sufficient number of images for model training. Selection of the test images was more tightly controlled, resulting in the test dataset being far smaller (n=40). In addition to the exclusion criteria used in the selection of the training images, test images were further constrained to have: (1) depictions of *both a microwave and a clock* (enabling different targets to be designated in the identical images, the perfect control for differences in bottom-up saliency), (2) only a single instance of the target, (3) a target area less than 10% of the image area, and (4) targets that do not appear at the image's center (no overlap between the target and the center cell of a 5x5 grid). The latter two criteria were aimed at excluding really large targets or targets appearing too close to the center starting-gaze position, with the goal of both being to achieve a moderate level of search difficulty (see Fig. 3, bottom).

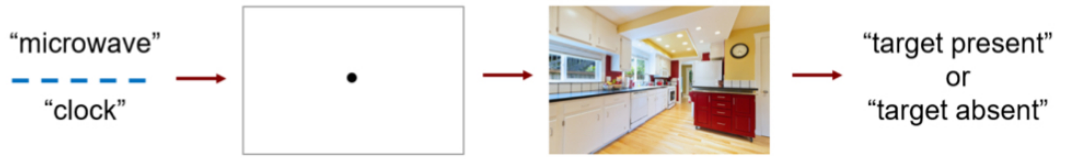

FIGURE 4 The categorical search paradigm used for behavioral data collection.

The above-described selection criteria were specific to target-present (TP) images, but an equal number of target-absent (TA) images (n=2183) were selected as well so as to create a standard TP versus TA search context. These images were selected randomly from COCO, with the constraints that: (1) none depicted the target, and (2) all depicted at least two instances of the target category's siblings. COCO defines the siblings of a microwave to be: ovens, toasters, refrigerators, and sinks, all under the parent category of "appliances". Clock siblings are defined as: books, vases, scissors, hairdryers, toothbrushes, and teddy bears, under the parent category of "indoor". Sibling membership was used as a selection criterion so as to discourage TA responses from being based on scene type (e.g., a street scene is unlikely to contain a microwave), and this criterion seemed to work well; the overwhelming majority of selected TA scenes were kitchens that did not depict a target.

The large size of the training dataset (4366 images) required data collection to be distributed over groups of searchers. Each microwave training image was searched by 2-3 people (n=27); each clock training image was searched by 1-2 people (n=26). After removing incorrect trials and TP trials in which the target was not fixated (it is not desirable to train on these), 16,184 search fixations remained for model training. Test images were each searched by a new group of 60 participants, 30 searching for a microwave target and the other 30 searching the same images for a clock target in a between-subjects design. To achieve a power and effect size of .8, based on a t-test comparing target guidance to chance (as defined in Fig. 5), we determined that a sample of 25 participants per target condition would be adequate. However, we chose to test 30 participants per condition in case of loss due to attrition or unusable eye-tracking data.

2.4 | Behavioral Search Procedure

A standard categorical search paradigm was used for both training and testing (Fig. 4). TP and TA trials were randomly interleaved within target type, and searchers made a speeded TP or TA manual response terminating each trial. Search display visual angles were $54^\circ \times 35^\circ$ for testing; for training angles ranged between $12 - 28.3^\circ$ in width and $8 - 28.3^\circ$ in height. Eye position was sampled at 1000 Hz using an EyeLink 1000 (SR Research) in tower-mount configuration (spatial resolution 0.01° rms). All participants provided informed consent in accordance with policies set by the institutional review board at Stony Brook University responsible for overseeing research conducted on human subjects.

3 | RESULTS

3.1 | Search Behavior

Table 1 provides the mean percent button-press errors and the average number of fixations made before the button-press response (which includes the starting fixation) on correct search trials. Note that the roughly doubled error rates

in the training data should be interpreted with caution, as many of these errors were due to incorrectly labelled target-object regions in COCO that create errors given correct search judgments. Rather than correcting these mislabelled objects (which would be changing COCO), we instead decided to tolerate an inflated error rate and to exclude these error trials from all analyses and interpretation.

	Target Category	Training Dataset		Testing Dataset	
		Error (%)	Mean (SD) Fixations	Error (%)	Mean (SD) Fixations
target-present	microwave	18	5.46 (±2.6)	9	6.76 (±2.1)
	clock	15	4.52 (±3.5)	6	5.33 (±1.8)
target-absent	microwave	8	7.95 (±4.1)	4	14.36 (±2.5)
	clock	10	11.14 (±6.8)	5	15.85 (±2.3)

TABLE 1 Summary statistics showing mean errors and number of search fixations in the Microwave-Clock Search dataset.

Focusing first on the TP test data, Figure 5 plots the cumulative probability of fixating the target with each saccade made during search. The central behavioral data pattern (the topmost red and blue lines) is that attention, as measured by overt gaze fixation, is strongly guided to both the microwave and clock targets. This guidance is evidenced by the fact that 24% of the initial saccades landed on targets (averaged over microwaves and clocks). This probability of target fixation is well above chance, which we quantified using two object-based chance baselines consisting of: (1) the probability of fixating the clock when searching for a microwave (Behavior baseline - Clock), and (2) the probability of fixating the microwave when searching for a clock (Behavior baseline - Microwave). We confirmed above-chance target guidance by comparing the slopes of regression lines fit to the target and baseline data (microwave: target slope = 0.15, baseline slope = 0.03, $t(58) = 26.31$, $p = 6.20e\text{-}34 < .001$; clock: target slope = 0.17, baseline slope = 0.004, $t(58) = 52.65$, $p = 1.14e\text{-}50 < .001$). Also evident from this analysis is the importance of the first six saccades made during the search tasks. If the target was going to be fixated, it is highly likely that this would happen by the sixth eye movement. Collectively, these results indicate that there are strong microwave and clock guidance signals in the behavioral test data to predict.

3.2 | Model Performance

To determine whether the IRL model's behavior is reasonable, we conducted two initial qualitative analyses. The top row in Figure 6 shows cumulative ReT-images for the starting fixation (0 in the yellow scanpath) and the fixations following the first two saccades (1, 2). Note that the left ReT-image, because it was computed based on a center initial fixation position, is blurred on both the left and right sides. The middle and right ReT-images were computed based on the landing positions of the first and second saccades, respectively. The microwave target is indicated in each panel by the red box. The bottom row shows the saccade maps corresponding to these ReT-images, where a bluer color indicates greater total reward expected by moving fixation to different image locations. The model initially expected the greatest total reward by fixating the stove (left saccade map), but after that saccade, and the resulting change in state (top middle), the model then selected the microwave target as the location offering the greatest expected reward (bottom middle), which was fixated next (right panels). Note that the model, because it was forced to make six saccades (discussed below), continued to prioritize space even after fixating the target. This qualitative analysis

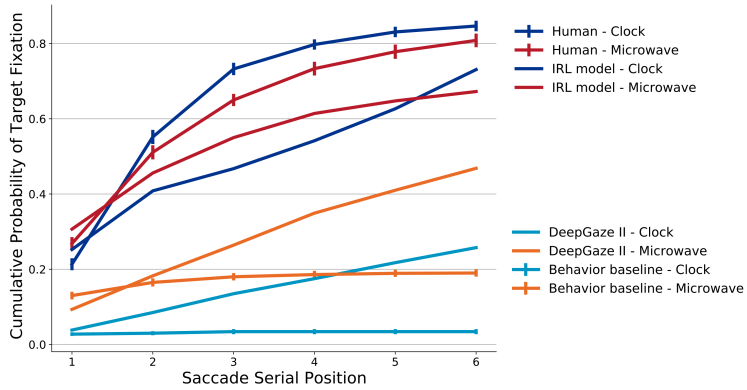

FIGURE 5 Cumulative probability of fixating the microwave (red) or clock (blue) targets on target-present trials. Behavioral participants are indicated by lines with error bars (standard error of the mean), the IRL model is indicated by lines without error bars. Microwave (orange) and clock (cyan) search is indicated for Deep Gaze II (lines without error bars) and Behavior baseline models (lines with error bars; see text for details).

shows that the model learned an association between a state (which includes the features of a microwave) and an action, and this enabled it to guide its fixations during the search of a new image for this target-category goal.

FIGURE 6 Cumulative ReT-images (top row) and corresponding saccade maps (bottom row) for the initial and first two new fixations (left to right) made by the IRL model in a microwave search task.

Figure 7 shows another qualitative evaluation, this time comparing fixation-density maps (FDMs) from people searching for a microwave (n=30) or a clock (n=30) to FDMs generated by the IRL model (sampling from probabilistic policy) as it searched for the same targets in the same two test images. In both examples, the model and behavioral searchers efficiently found the target (bright red). More interesting, however, is that they both searched the scenes differently depending on the target category. When searching for a microwave (leftmost four panels) the model and

behavioral searchers tended to look at counter-tops, but when searching for a clock (rightmost four panels) they tended to look higher up on the walls. Future work will more fully explore the potential to learn and predict these effects of scene context on search.

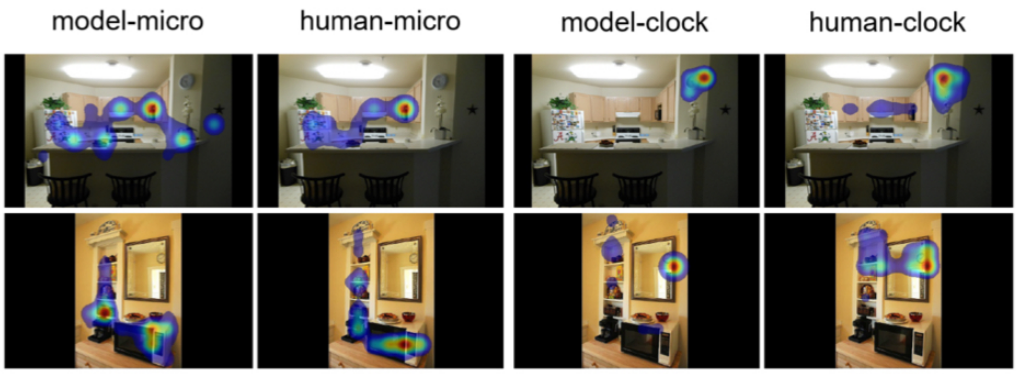

FIGURE 7 Model and behavioral fixation density maps computed for microwave (left four) and clock (right four) searches in two trials (top, bottom).

A more quantitative comparison of the IRL model to behavior appears in Figure 5. This comparison occurred on an image-by-image and fixation-by-fixation basis, but was limited to the first six movements of gaze. We introduced this limitation on the number of search saccades to reduce model computation time, but believe that it is justified given the clear adequacy of the first six saccades in revealing the goal-directed behavior of interest. The cumulative probability of target fixation over saccades is perhaps the clearest measure of search efficiency, and by this measure the model was able to guide its high-resolution foveal window to the microwave and clock targets much like how our participants controlled their search behavior. Both made between 20% and 30% of their initial saccades directly to the target, regardless of target type. The probability of fixating the target by the second saccade also increased sharply, but less so for the IRL model, whose performance was lower overall. The IRL model tended to either fixate the targets very efficiently in its first two new fixations, or not at all. To obtain a performance baseline that does not reflect goal-directed control, we also ran DeepGaze II [15] on the test images and compared its performance to the IRL model and search behavior. The behavior of this model was similarly limited to only six saccades, sampled probabilistically, although its input was full-resolution images instead of ReT images so as to maximize its prediction success. In contrast to the highly efficient initial saccades observed in the behavior and for the IRL model, DeepGaze II selected the target in its initial movement 10% or less of the time, which was no greater than chance. Note that the model was more likely to fixate microwaves than clocks because microwave targets tended to be larger, the same reason why the chance baseline is higher for this target category. The probability of DeepGaze II fixating the target increased with each saccade, but this was due to inhibitory spatial tags inserted at previously selected locations forcing samples to be taken from different locations in the image. Collectively, we interpret these results as suggesting that the IRL model was successful in learning how to control its early search saccades to different target-object goals, far more so than baselines but slightly less efficiently than human searchers.

Table 2 compares model performance to participant behavior using several metrics applied to the test data. The first column of values indicates search accuracy (ACC), where “accuracy” refers here to the proportion of trials in

Microwave	ACC↑	AvgSaccades↓	AUC↑	NSS↑	MM↑
Behavior	0.808	2.283	0.794	2.175	0.846
IRL	0.673	2.170	0.676	0.928	0.798
DG2	0.468	3.224	0.648	0.828	0.790
Clock	ACC↑	AvgSaccades↓	AUC↑	NSS↑	MM↑
Behavior	0.847	2.277	0.774	1.981	0.852
IRL	0.731	2.858	0.576	0.848	0.826
DG2	0.258	3.519	0.605	0.602	0.792

TABLE 2 Evaluation of the IRL and Deep Gaze II (DG2) models against behavior using multiple metrics. Arrow direction indicates better performance. ACC: Proportion of trials in which the target was fixated in the first six saccades (fixated-in-6 accuracy). Avg Saccades: Average number of saccades to the target on the fixated-in-6 trials. The values of Behavior for these two metrics indicate actual observed behavior. AUC: Success in predicting FDMs using the Judd Area-Under-the-Curve metric. NSS: Success in predicting fixation locations using the Normalized Scanpath Saliency metric. The values of Behavior for these two metrics indicate a Subject model computed by having a fixation-density map from half of the participants predict the FDM from the other half of the participants. MM: Predictions of search scanpaths using average MultiMatch similarity. For this metric the value for Behavior indicates a Subject model computed by averaging pairwise scanpath comparisons.

which the target was fixated in the first six eye movements (fixated-in-6 accuracy). The IRL model was about 10-13% less successful than participants in locating targets within six saccades, which was also clear from the lower performance ceiling in Figure 5. Chance fixated-in-6 accuracy is less than .25, based on a shuffling of eye data and images within each participant, and this is far lower than fixated-in-6 accuracy for the IRL model (microwave: t(58) = -31.74, p = 2.34e-38 <.001, Cohen's d = 8.20; clock: t(58) = -75.87, p = 9.73e-60 <.001, Cohen's d = 19.59). The IRL model performance can be contrasted with DeepGaze II, which managed to select the microwave target on less than half of the trials and the clock target on only about a quarter of the trials, no better than chance. Because the other performance metrics use data only from trials in which the target was fixated in the first six saccades, the poor fixated-in-6 accuracy of DeepGaze II compromises its comparison to the IRL model or behavior; the reported results will be from those few trials where DeepGaze II managed to fixate the target, which will be biased to include trials having salient target objects. For this reason we do not interpret the results of DeepGaze II for the other Table 2 metrics, although we include them for the interested reader.

The other data columns in Table 2 show the results of analyses of the accurate fixated-in-6 trials. The second column is the mean number of saccades needed to find the target (Avg Saccades). By this measure, the IRL model tended to find the target as efficiently as the participants, needing only about half a fixation more in the case of clocks. The third and fourth columns of the table show metrics for comparing the spatial fixation predictions of the models to behavior on the fixated-in-6 trials. The AUC metric has a scale between 0 and 1, where higher values indicate better success in predicting the behavioral FDMs [2, 4]. Higher values for the NSS metric also indicate better predictive success [2, 4]. The "Behavior" values for the AUC and NSS metrics refer to Subject models, which are useful to obtain a practical noise limit on a model's ability to predict group behavior [4]. For both metrics, the Subject model was created by randomly splitting the 30 subjects into two groups of 15 (exploration of different random splits made little difference), then having FDMs from one group predict the FDMs from the other group. These analyses show good model predictions of behavioral fixation locations in the test images, with the greater room for improvement being in the clock search task. Finally, AUC and NSS aggregate the fixations made during the search of an image, and are

therefore purely spatial metrics, but search fixations happen over time, ultimately producing a scanpath. Because the models also make sequences of fixations, we were able to compare their 6-saccade scanpaths to the 6-saccade scanpaths from the behavioral searchers. We conducted this scanpath comparison using MultiMatch [6], excluding the metric's fixation duration component. The rightmost table column is average MultiMatch similarity of the fixated-in-6 scanpaths, where the IRL model did a very good job in predicting the spatio-temporal sequences of fixations made by the behavioral searchers in their first six saccades, nearly as well as could be expected from the behavioral data. We determined this using a Subject model computed by averaging the pairwise scanpath comparisons for a given image. This improved prediction, relative to the Subject model for the spatial metrics, underscores the importance of spatio-temporal information in the prediction of search behavior.

4 | DISCUSSION

Models of search behavior have traditionally aimed at describing relatively coarse patterns (e.g., set-size effects) in highly simplified contexts [22, 25], limitations that were imposed by a reliance on handcrafted features to create a guidance signal. In this study we adopted the radically different approach of training a model simply on many observations of search behavior. We found that the policy learned by this model predicted multiple measures of overt goal-directed attention control. The success of these predictions is significant in that it requires a re-setting of the goal posts with respect to model evaluation. While once computational methods limited attention models to fitting patterns of search data in simple contexts, with deep networks it is possible to predict individual fixations made in the search for categories of objects in realistic scenes.

Training this model required creating the Microwave-Clock Search dataset, which is among the only datasets of goal-directed attention (search fixations) large enough to train deep-network models. We encourage people to download this dataset from `https://you.stonybrook.edu/zelinsky/datasetscode/` and use it in their own predictive-modeling work, citing this publication. Our hope is that the availability of this dataset will promote greater model development and comparison, which is needed to meaningfully advance the understanding of goal-directed attention control [1].

The visual search for an object category is a goal-directed behavior of unique importance, shared by pigeons and people and most species in between. Because of its fundamental role in survival, search is likely to use the most basic of control processes—reward. [1] Using the MCS dataset and inverse-reinforcement learning, we showed that the target-specific reward functions learned by our model predicted the goal-directed fixations made by new people searching new images for the learned target categories. With this machine learning method it is now possible to learn the reward functions underlying goal-directed attention control. In future work we plan to manipulate the different types of reward used in model training, explore the potential of learning scene context effects, and apply IRL to

[1] During the course of bringing the currently-described work to publication, related work from our group was presented at the *Computer Vision and Pattern Recognition* annual meeting [23]. Similar to the present study, that work aimed to apply IRL to the prediction of fixations made during object-category search. However, both chronologically and conceptually the work described in the present study came first. Our focus in this initial effort was to demonstrate the plausibility of using IRL to study goal-directed attention, which we did by showing that the different microwave and clock representations learned by the IRL model could predict the different search behaviors observed to these two target classes. We also definitively ruled out bottom-up saliency in explaining these behavioral differences by using test images depicting instances of both a microwave and a clock. With these validations, the goal of the CVPR work was to explore new state representations in the context of a larger dataset of search behavior, needed to train more powerful IRL models. This dataset is COCO-Search18 [23, 5], which consists of approximately 300,000 fixations from 10 people searching for 18 different categories of target objects in 6202 images of natural scenes. COCO-Search18 is larger than the MCS dataset, and contains more target categories, but these datasets also differ in that each MCS test image depicted an instance of both a microwave and a clock whereas no such requirement was used to select the COCO-Search18 test images. The training images from the two datasets also only partially overlap with respect to their microwave and clock target categories. Because these are different datasets aimed at different model evaluations, and given the largely simultaneous progression of these efforts, we do not include in the present study comparisons to models trained using COCO-Search18.

questions in individual-difference learning.

data availability

The dataset described in this paper is available in the *Microwave-Clock Search (MCS) dataset* repository: `https://you.stonybrook.edu/zelinsky/datasetscode/`.

acknowledgements

We would like to thank the National Science Foundation for their generous support through award IIS-1763981, and members of the EyeCog Lab for their help with data collection and invaluable feedback.

author contributions

M.H., D.S., and G.J.Z. conceptualized the research; H.A., Y.C., and G.J.Z. collected the dataset, L.H., Z.Y., D.S., and M.H. implemented the model. All authors analyzed and interpreted data, but especially Z.Y., Y.C., L.H., and S.A. G.J.Z., Y.C., S.A., and H.A. wrote the paper.

conflict of interest

The authors declare no competing interests.

references

[1] Brian A Anderson. A value-driven mechanism of attentional selection. *Journal of vision*, 13(3):7–7, 2013.

[2] 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.

[3] Claus Bundesen. A theory of visual attention. *Psychological review*, 97(4):523, 1990.

[4] 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] Yupei Chen, Zhibo Yang, Seoyoung Ahn, Dimitris Samaras, Minh Hoai, and Gregory Zelinsky. Coco-search18: A dataset for predicting goal-directed attention control. *bioRxiv*, 2020.

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

[7] 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.

[8] Martin Eimer. The neural basis of attentional control in visual search. *Trends in cognitive sciences*, 18(10):526–535, 2014.

[9] Syed Omer Gilani, Ramanathan Subramanian, Yan Yan, David Melcher, Nicu Sebe, and Stefan Winkler. Pet: An eye-tracking dataset for animal-centric pascal object classes. In *2015 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE, 2015.

[10] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. 2015.

[11] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in Neural Information Processing Systems*, pages 4565–4573, 2016.

[12] David E Irwin. Integrating information across saccadic eye movements. *Current Directions in Psychological Science*, 5(3):94–100, 1996.

[13] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[14] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

[15] Matthias Kummerer, Thomas SA Wallis, Leon A Gatys, and Matthias Bethge. Understanding low-and high-level contributions to fixation prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4789–4798, 2017.

[16] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer, 2014.

[17] Stefan Mathe and Cristian Sminchisescu. Actions in the eye: Dynamic gaze datasets and learnt saliency models for visual recognition. *IEEE transactions on pattern analysis and machine intelligence*, 37(7):1408–1424, 2014.

[18] Dim P Papadopoulos, Alasdair DF Clarke, Frank Keller, and Vittorio Ferrari. Training object class detectors from eye tracking data. In *European conference on computer vision*, pages 361–376. Springer, 2014.

[19] 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 57–70. International Society for Optics and Photonics, 2002.

[20] Joseph Schmidt and Gregory Zelinsky. Search guidance is proportional to the categorical specificity of a target cue. *The Quarterly Journal of Experimental Psychology*, 62(10):1904–1914, 2009.

[21] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[22] Jeremy M Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic bulletin & review*, 1(2):202–238, 1994.

[23] 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 193–202, 2020.

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

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

[26] Gregory J Zelinsky, Yifan Peng, Alexander C Berg, and Dimitris Samaras. Modeling guidance and recognition in categorical search: Bridging human and computer object detection. *Journal of Vision*, 13(3):30–30, 2013.

[27] Gregory J 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 Conference on Computer Vision and Pattern Recognition Workshops*, 2019.

[28] 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):3730, 2018.

Appendix: Supplemental Model Methods

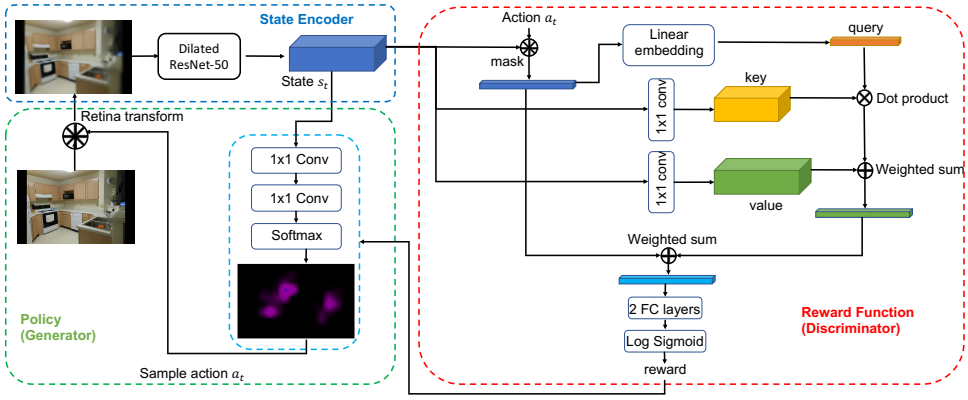

FIGURE A1 Detailed architecture of the IRL model.

| Network Architecture

The method we used is generative adversarial imitation learning (GAIL [1]). Figure A1 shows the pipeline and detailed architecture of our model. It has three primary components: 1) the state encoder that extracts the state representation from a cumulative foveated image (blue dotted box); 2) the generator (policy) that tries to imitate human gaze behavior by generating scanpaths similar to those of the behavioral participants (green dotted box); and 3) the discriminator (reward function) that gives high reward to generated scanpaths (represented by state-action pairs) that are similar to human behavior and low value to those that are not (red dotted box).

The **state encoder**, here a ResNet-50 pre-trained on ImageNet, inputs a foveated image and outputs a $2048 \times 10 \times 16$ feature map. This is the state representation used for model training and testing. However, before this feature extraction could occur the model had to first learn to "see" through a foveated retina. The problem is that ImageNet consists of full-resolution images, making models trained on ImageNet completely unfamiliar with images transformed to approximate a foveated retina. We therefore needed to fine-tune the model on foveated images, which we created by first randomly selecting a location in each training image and then applying the retina-transformation method from Perry & Geisler [3] at this location. We did this for each epoch during the fine-tuning of the dilated ResNet-50 [6] (pre-trained on ImageNet). The task used for fine-tuning was to predict from these foveated images the target locations (from ground-truth labels) and behavioral FDMs. This pre-training was done on both tasks simultaneously (i.e. predicting FDMs and target locations for an image) so that the encoded features contained information for both tasks. The features extracted from foveated images by this fine-tuned ResNet-50 make up the state representation passed on to the generator and discriminator.

The **generator** consists of two parts: the Actor network and the Critic network. The Actor network maps a $2048 \times 10 \times 16$ state to a 10×16 action map (saccade map), from which we sample an action (make a saccade) and generate the next state. The Critic network maps the state to a value indicating the expected return (accumulated long-term reward) of being in that state. The Actor network consists of two 1×1 convolutional layers and a softmax layer. The Critic network is an 1×1 convolutional layer followed by a fully-connected layer. The policy is trained using

a Proximal policy optimization algorithm (PPO) [4] and an Actor-Critic algorithm [2], with rewards provided when the generator generates fixations that fool the discriminator.

The **discriminator** outputs the log probability that a given state-action pair (S, a) corresponds to human behavior (true data), as opposed to the IRL agent (fake data). Changes in fixation location are encoded as the spatio-temporal sequence of changing foveated images. Each fixation is evaluated separately by the discriminator, but is conditioned on all previous fixations in the emerging scanpath by the accumulation of the non-blurred (foveal) pixels from all previously fixated locations (cumulative retina-transformed image; Figure 2 in the main text).

We also used a **self-attention** module [5] with our reward function to augment the local image features extracted by the model with features capturing a more global context. We did this to exploit any non-local contextual dependencies that may exist during visual search, although we did not explore the contribution of this global context here. Specifically, the image feature vector x_a at location a of state S is first embedded into a 32-dimensional query vector q_a. The image features (state) are mapped into a key feature space f and value feature space g using 11 convolution, respectively, where $f(S) \in \mathbb{R}^{32\times10\times16}$ and $g(S) \in \mathbb{R}^{2048\times10\times16}$. A contextual feature vector c_a is computed as the weighted sum of the value $c_a = \sum_j \alpha_j g(s_j)$, where s_j is a 32-dimensional feature vector at a spatial location j of the 1016 image grid. The attention weights are calculated as

$$\alpha_j = \frac{\exp(q_a^\top f(S_j))}{\sum_j \exp(q^\top f(S_j))}. \tag{1}$$

Inspired by human attention control, known to be a combination of top-down contextual biases and bottom-up biases from image features, we combine the contextual feature vector c_a with the image feature vector using a learnable weight γ: $z_a = \gamma c_a + x_a$. The feature vector z_a is then mapped into a probability using two fully-connected layers and a sigmoid function. The reward is computed as:

$$r(S, a) = \log(\mathrm{sigmoid}(\mathrm{linear}(\mathrm{ReLU}(\mathrm{linear}(z_a))).$$

| Model Training

Let D and G denote the discriminator and the generator, respectively. The discriminator aims to differentiate human state-action pairs from fake state-action pairs generated by the policy. Hence, it is trained by maximizing the following objective function:

$$\begin{aligned} {}_D =& \mathbb{E}_r[\log(D(S, a))] + \mathbb{E}_f[\log(1 - D(S, a))] \\ & - \lambda \mathbb{E}_r[\|\nabla D(S, a))\|^2]. \end{aligned} \tag{2}$$

The generator aims to fool the discriminator, and its objective is to maximize the log likelihood of the generated state-action pairs, i.e., to maximize:

$${}_G = \mathbb{E}_f[\log(D(S, a))] = \mathbb{E}_f[r(S, a)] \tag{3}$$

The generator is an RL policy, meaning its objective can be equivalently reformulated as an RL objective and optimized by Proximal Policy Optimization (PPO) [4].

$$_{\pi} = \mathbb{E}_{\pi}[\log(\pi(a|S))A(S,a)] + H(\pi), \tag{4}$$

$$\text{where} \quad A(S,a) = Q(S,a) - V(S), \tag{5}$$

$$H(\pi) = -\mathbb{E}_{\pi}[\log(\pi(a|S))]. \tag{6}$$

In the above, A is the advantage function – the difference between the state-action value function Q and the state value function V in reinforcement learning. H is the entropy in max-entropy inverse reinforcement learning [7].

references

[1] Jonathan Ho and Stefano Ermon. Generative adversarial imitation learning. In *Advances in neural information processing systems*, pages 4565–4573, 2016.

[2] Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

[3] 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 57–70. International Society for Optics and Photonics, 2002.

[4] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008, 2017.

[6] Fisher Yu, Vladlen Koltun, and Thomas Funkhouser. Dilated residual networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 472–480, 2017.

[7] Brian D. Ziebart, Andrew Maas, J. Andrew Bagnell, and Anind K. Dey. Maximum entropy inverse reinforcement learning. In *AAAI Conference on Artificial Intelligence*, 2008.