# Supplementary Material: 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

## 1. Table of Contents

- Sec. 2 provides additional analyses comparing fixations during target-absent search (TA), target-present search (TP), and free viewing (FV).
- Sec. 3 reports the significance tests for feature map weightings.
- Sec. 4 provides additional results for classifying tasks from partial scanpaths, including an ablation study and the TP data mentioned in Fig. 3 of the main text.
- Sec. 5 reports an ablation study on the TA search termination model and additional implementation details.

## 2. Additional Fixation Comparison

### 2.1. Fixation Density Maps (FDMs)

Figure 1 shows FDMs based on the 4th, 5th, 10th and 15th fixations. Note that too few 10th and 15th fixations were available for analysis in TP search (explaining the absent FDMs), but even in the TA and FV comparison there were far more 10+ fixation trials in free-viewing than in search. We obtained FDMs by coding fixated image locations (over images and participants) with 1s and non-fixated locations with 0s and then convolving this discrete fixation point map with a Gaussian (sigma equalling one degree of visual angle, approximating the size of the fovea) to obtain the continuous FDM (see [1,2] for additional details on FDMs). Compared to the FDMs for the first three fixations (Fig. 1 in the main text), these FDMs show similar patterns. The fourth and fifth fixations in TP search again showed a bilateral distribution reflecting a target guidance signal pulling gaze away from the center. This same guidance signal explains the TA FDMs, but because this guidance is weaker in TA search it is less able to overcome the oculomotor inertia causing fixation to remain around the center. This center bias strikingly persists even after 15 new fixations during free viewing, whereas for TA search a center bias was largely dispersed by the target guidance signal after the same number of fixations.

### 2.2. Saccade Amplitude

Supplementing our analysis of fixations locations, Figure 2 shows distributions of first, second, and third saccade amplitudes in TP search, TA search and free viewing. Consistent with the FDM analyses, first saccades in both TP and TA search had overall larger amplitudes than first saccades during free viewing (which showed about a 5° reduction in range). Amplitudes tended to decrease with subsequent fixations, with their distributions becoming increasingly positively skewed. Note that second-saccade amplitudes in TA search remained comparatively larger than in free-viewing, which we again interpret as evidence for a guidance signal pulling gaze to different image locations, and away from the center. Also expected were the sharp peaks observed for second and third saccade-amplitude distributions in the case of TP search, reflecting a progressive movement of gaze toward the target and possibly small saccades around the target for confirmation.

## 3. Significance Tests for Feature Map Weightings

In Table. 1 we show the results of paired t-tests on NSS scores, Bonferroni corrected, for the saliency (Sal), target (Target), and center bias (CB) feature maps in TP search, TA search and free viewing, as discussed in Section 3.2.4 in the main paper. All claims were based on paired t-test with a $p_{bonferroni} < .016$. We see that all differences in TP Search, TA Search and FV were significant at this level of confidence.

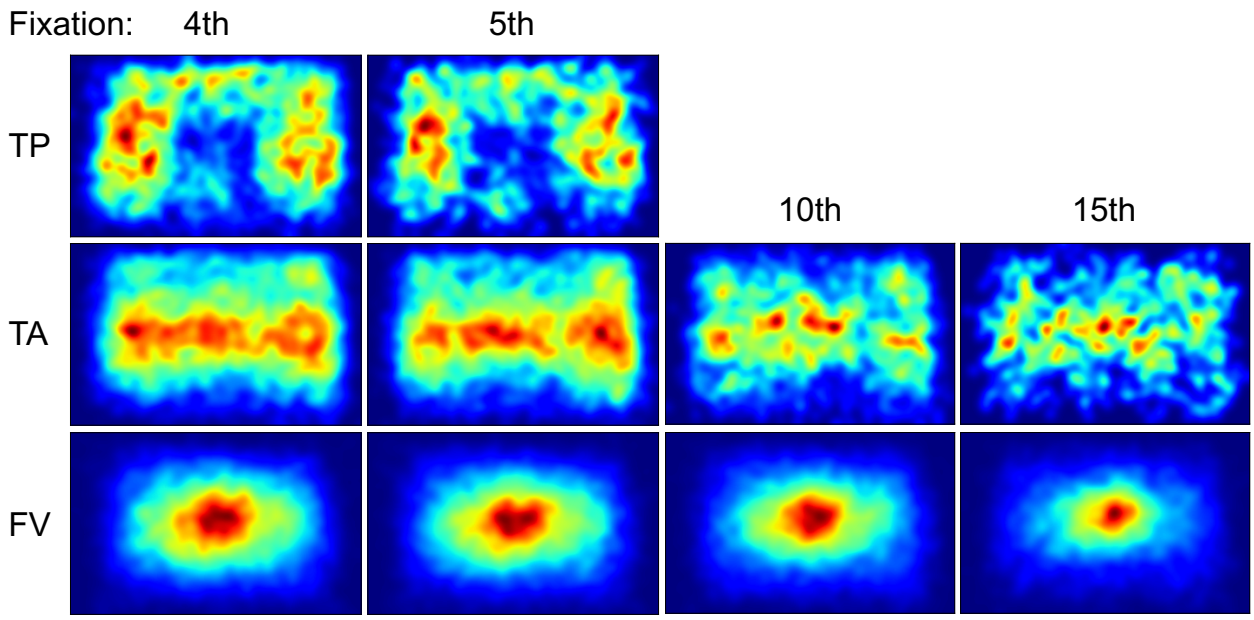



Figure 1. Fixation-density maps visualized for the 4th, 5th, 10th and 15th fixations in target-present search (TP), target-absent search (TA), and free viewing (FV) behaviors.

| **TP Search** | | | |
|---|---|---|---|
| **Fixation** | **Saliency-Target** | **Saliency-Center Bias** | **Target-Center Bias** |
| 1 | $t = -10.72, p < 0.001$ | $t = 8.94, p < 0.001$ | $t = 14.95, p < 0.001$ |
| 2 | $t = -26.00, p < 0.001$ | $t = 16.54, p < 0.001$ | $t = 34.03, p < 0.001$ |
| 3 | $t = -21.84, p < 0.001$ | $t = 14.84, p < 0.001$ | $t = 28.70, p < 0.001$ |
| **TA Search** | | | |
| **Fixation** | **Saliency-Target** | **Saliency-Center Bias** | **Target-Center Bias** |
| 1 | $t = 20.73, p < 0.001$ | $t = 8.60, p < 0.001$ | $t = -17.58, p < 0.001$ |
| 2 | $t = 11.34, p < 0.001$ | $t = 10.26, p < 0.001$ | $t = -6.02, p < 0.001$ |
| 3 | $t = 7.51, p < 0.001$ | $t = 8.77, p < 0.001$ | $t = -2.50, p = 0.012$ |
| **Free Viewing** | | | |
| **Fixation** | **Saliency-Target** | **Saliency-Center Bias** | **Target-Center Bias** |
| 1 | $t = 19.22, p < 0.001$ | $t = -3.61, p < 0.001$ | $t = -22.42, p < 0.001$ |
| 2 | $t = 10.80, p < 0.001$ | $t = 7.10, p < 0.001$ | $t = -7.26, p < 0.001$ |
| 3 | $t = 11.33, p < 0.001$ | $t = 5.90, p < 0.001$ | $t = -8.19, p < 0.001$ |

Table 1. Student's paired t-test results on NSS scores, Bonferroni corrected, for pairwise comparisons between saliency, target and center bias features in TP Search (degrees of freedom, $df_{TP} = 1223$), TA Search (degrees of freedom, $df_{TA} = 1223$) and Free Viewing (degrees of freedom, $df_{FV} = 1175$).

## 4. Classifying Task from Scanpaths

Figure 3 shows how our LSTM-based classification model $[X; Y; D; V]$ compares to versions of the model that were ablated by removal of one or more of its input features: 2D coordinates $X \in \mathbb{R}$ and $Y \in \mathbb{R}$, durations $D \in \mathbb{R}$, and

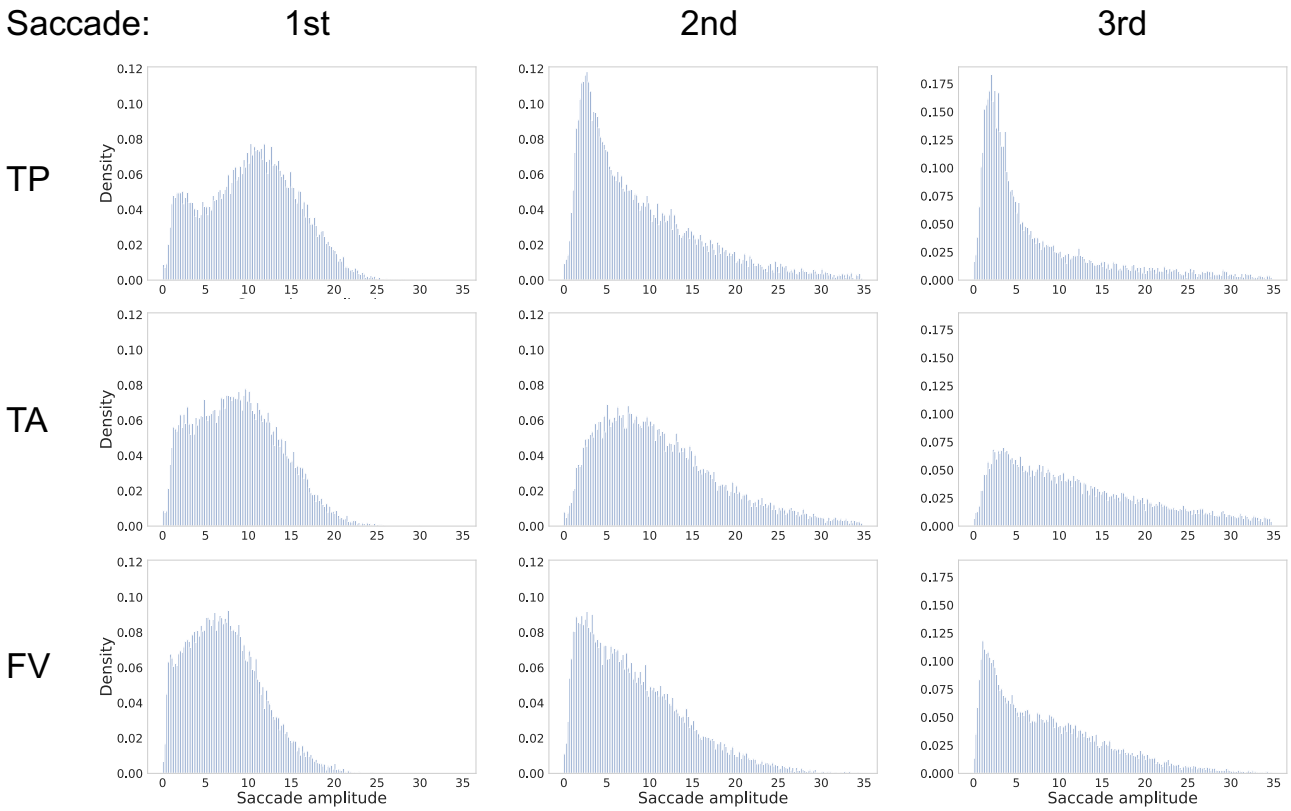



Figure 2. Distribution of saccade amplitudes computed in degrees of visual angle for the first, second, and third fixations made during TP search (TP), TA search (TA) and free viewing (FV).

visual features $V \in \mathbb{R}^C$. Unsurprisingly, the complete model, i.e. when we include all features in the form of $[X; Y; D; V]$, outperforms versions using only a subset of the features as inputs. This shows that each feature contributed to the classification of TP, TA and free-viewing tasks, and how these contributions changed with partial scanpath length. Also clear from this analysis, and reaffirming our conclusion in the main text, is the particularly important role played by visual features in predicting task classification.

Figure 4 shows the probabilities of predicting the three tasks from partial scanpath data. The left and middle panels re-plot these probabilities from the main text (Fig. 3c and 3d) for the TA and FV data, respectively. The rightmost plot shows the TP data for comparision, which was not included in the main text. In contrast to the very poor classification of a FV task from the others when partial scanpath lengths were small (about .2 in Figure 4c), our model is capable of predicting TP search even on small partial scanpaths (about .75 in Figure 4a). This suggests that TP scanpaths can be easily separated from TA search and free viewing scanpaths, with the feature responsible for this distinctiveness corresponding to what we being to be the target guidance signal. In the TA data (Figure 4b), where this signal is weaker, an intermediate data pattern is found.

We did minimal hyperparameter tuning on different combinations of input features for the models tested. For those models containing visual features $V$, the LSTM network contained 3 layers with the hidden layer size of the LSTM cells set to 20. For configurations not containing visual features $V$, the LSTM network contained 2 layers with hidden layer size of the LSTM cells set to 6. For all models, batch size was set to 128, learning rate of Adam optimizer [4] was set to 3e-4, dropout probability in the LSTM cells was set to 0.2 and dropout probability of the last MLP layer was set to 0.3. To ensure that we are able to batch process scanpath samples while ignoring padding, we use a length mask that enables us to recover the final time step for each sample in a batch. The architecture is trained using categorical cross-entropy loss averaged over each time step within the scanpath.

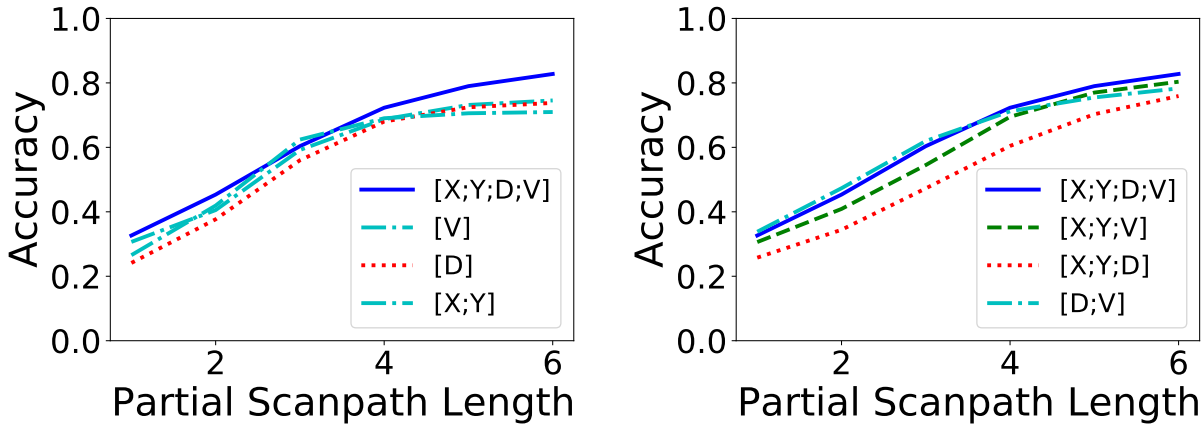
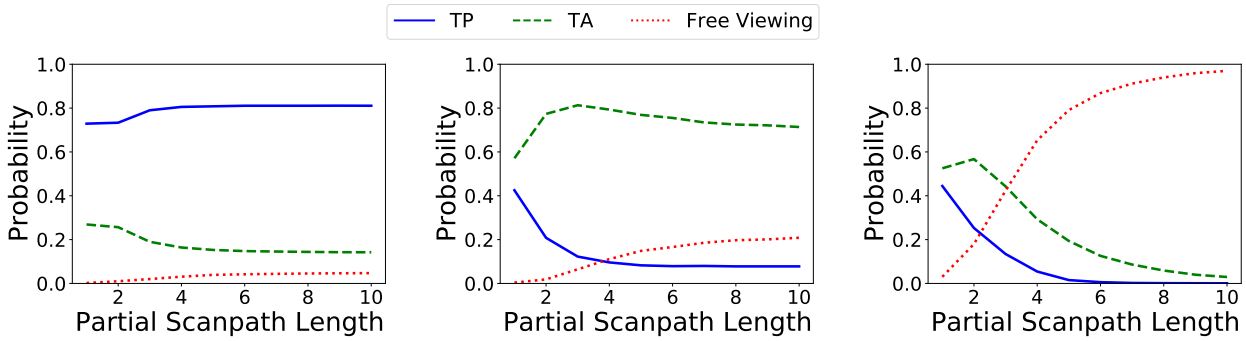



Figure 3. Classification accuracy as a function of partial scanpath length for models using either fixation location $(X, Y)$, fixation duration $(D)$, or visual features at the fixated location $(V)$ as inputs (left plot) or a pairwise combination of inputs (right plot). For comparision, tThe complete model, $[X; Y; D; V]$, is shown by the solid blue line in both plots. Note that two plots are used instead of one simply to avoid clutter.




(a) On TP Data (b) On TA Data (c) On Free Viewing Data

Figure 4. Model predictions for classifying TP scanpaths, of variable length, as coming from either a TP search, TA search, or free-viewing task.

## 5. Termination Prediction

### 5.1. Label Foveation

Figure 5 gives an overview of our label foveation process. Our aim is to generate a continuous "retina" mask where the target likelihood of the target pixels and non-target pixels decreases and increases as a function of retinal eccentricity, respectively. To this end, we first create a six-level pyramid of the label maps from $P_0$ to $P_5$, where $P_0$ is the ground-truth binary target mask (1 for the target pixels and 0 for the non-target pixels). From $P_1$ to $P_5$, the values for the target and non-target pixels linearly decrease or increase, respectively. Formally,

$$P_i^j = \begin{cases} 1 - \alpha i & \text{if } j \text{ is a target pixel} \\ \alpha i & \text{if } j \text{ is a non-target pixel,} \end{cases}$$

where $i \in \{0, \cdots, 5\}$ and $j$ indicates the pixel location. We use a linear slope $\alpha = 0.1$ in our experiments. Hence, $P_5$ is 0.5 everywhere, meaning that the target is utterly indistinguishable from the background if viewed from the furthest peripheral vision. To generate the final continuous label map, we follow the foveation algorithm in [3, 5, 6] and create a gaze-contingent resolution map to combine all of the label maps in the pyramid. For multiple fixations, we follow the cumulative foveation

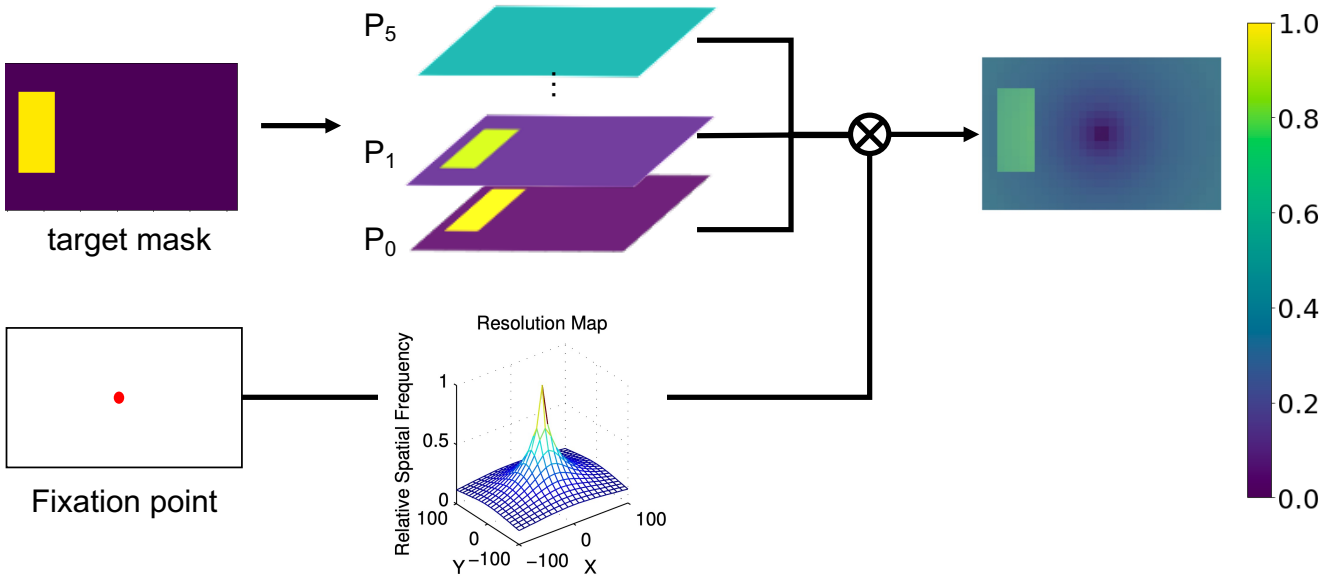



Figure 5. Overview of the label foveation process.

algorithm at [6] and take the element-wise maximum at the resolution maps to produce the final resolution map used to fuse the label pyramid.

### 5.2. Further Implementation Details

We considered two different training paradigms: one in which we jointly trained a foveated target detector and a termination predictor and another in which we treated these as two separate stages. We adopted the two-stage training approach after finding that joint training led to a performance drop (the average precision fell from 0.462 to 0.386), which is likely caused by difficulty in finding a trade-off to balance the two factors. In the first stage of training, we train the foveated target detector on the TP image training set from COCO-Search18 with randomly generated scanapths having a maximum scanpath length of 5. We do this to learn a target signal, which is best done on TP images. In the second stage, we train the termination predictor with the training TA trials from COCO-Search18. We use the trained foveated target detector to compute the detection map that is used as input to the termination predictor. Note that the parameters of the foveated target detector are fixed during the second stage. To compute the precision, recall, and f1-score for our model and the DCB-based model, we use the decision threshold having the maximum f1-score in the validation set and report the results on the testing set. All models were trained using Adam optimizer [4] with a learning rate of $10^{-4}$, a decay rate of $10^{-8}$, and a batch size of 128, until the validation loss plateaued.

### 5.3. Ablation Study

In an ablation study we systematically removed each of the components in our model, which included the history fixation map, subject ID, and task ID. The precision, recall, and f1-score of the different ablated models in provided in Table 2. Removing any one of the components incurs a non-negligible performance drop relative to the intact model. However, among the three considered components removing subject ID impacted the model the most, reducing the F1-score from 0.462 to 0.326. This result suggests that different searchers employ different stopping criteria, because knowing the searcher's identity helped most to predict the termination of a search. Interestingly, removing target ID incurred the least cost, suggesting that knowing the category of search target was relatively unimportant to the stopping behavior compared to the other components.

## References

[1] Ali Borji, Hamed R. Tavakoli, Dicky N. Sihite, and Laurent Itti. Analysis of scores, datasets, and models in visual saliency prediction. 2013. 1

[2] 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. 1

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

| | Precision | Recall | F1-score |
|---|---|---|---|
| Ours | **0.402** | **0.543** | **0.462** |
| Ours w/o history map | 0.395 | 0.432 | 0.413 |
| Ours w/o subject ID | 0.285 | 0.380 | 0.326 |
| Ours w/o target ID | 0.365 | 0.499 | 0.422 |


[3] 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. 4

[4] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3, 5

[5] 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. 4

[6] 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. 4, 5