We thank the reviewers for their valuable and generally positive feedback. We are encouraged that the reviewers found our work novel in terms of "not being yet another CNN over-fitting visual search model" (R1 and R3), a significant improvement over Najemnik & Geisler (2005) (R1), supported by theory (R1) and rich in psychophysics (R1 and R3) with high amount of behavioural data (R1). We are pleased that R2 points out that considering detectability and uncertainty is an important aspect of our work that has not been addressed sufficiently in the image saliency literature. We appreciate the helpful suggestions regarding improvements to figures and wording and will incorporate these in the camera-ready version. We provide responses to the major concerns below.

@R4: Concern about the work's relevance to the NeurIPS community The goal of the paper is to model the process by which the human visual system makes saccades during search, based on a Bayesian ideal observer model; that seems like a suitable topic for the "Neuroscience and Cognitive Science" area. This focus is similar to previous NeurIPS papers (such as Deza & Eckstein, 2016; Yu, Hua, Samaras & Zelinsky, 2013; Smeulders & Lamme, 2009, among many, and we note that the other reviewers found the paper relevant.

@R4, R2: Comparison to existing saliency models We did not compare this model to existing saliency models such as DeepGaze because our goal with this approach is significantly different from those models. Deep-learning saliency models are fit to large training datasets of fixations using machine learning techniques; generally, they aren't built on models of visual processing and they focus more on fixation location than fixation order. Here, we wish to predict fixation sequences using a perceptual model of target detectability across the visual field.

@R2,R3: modeling uncertainty @R2: To approximate the uncertainty in a visual search task, we only need the d' which is a measure of discriminability between a target-present and target-absent patch. Thus, an estimate of the distributions of the target and background is not required. @R3: The area under the curve of likelihood distributions, $p(t|T)$ and $p(t|D)$, are associated with the probabilities of hit and false alarms, and calculated from eq1. As illustrated in Figure 2, the detectability d' can be calculated as the distance between the means of the two likelihood distributions.

@R1: Is this really an end-to-end CNN-based model of visual search? Why doesn't the model implement inhibition of return? The proposed pipeline can output the detectability map for any background and target pair. The model does not require inhibition of return because the Bayesian-update step after each fixation updates the prior via the posterior. This naturally causes the probability of target-present events to decrease for any previously-fixated locations that did not contain the target, thus reducing the likelihood of a return saccade.

@R1,R3: Number of fixations rather than scan-path as a measure for validating the model. Since the visual search is implemented on a statistically-stationary textured background, the scanpaths of different observers are expected to be quite different. The lengths (and by extension, number) of saccades should be similar because these depend on target discriminability. However, the directions of saccades may not be similar for all observers because one observer might make an initial saccade to the left, another to the right, etc. This makes it difficult to compare scanpaths.

@R2: Regarding the use of textured images rather than natural scenes, the single target and the scale of the target on the backgrounds. As mentioned in line 198 of the paper and noted by R3, the model is not limited to homogeneous textures and can easily be extended to natural scenes. However, the goal in training was to model target detectability on the widest possible range of backgrounds, not just the types of backgrounds on which a pedestrian target is most likely to appear in real-world scenes. Similarly, when choosing backgrounds for the human experiment, we chose backgrounds which exhibited a range of detectabilities in the model. We consider the scale of the background relative to the target to be irrelevant, because the goal is to model the perceptual detectability at different eccentricities – for this, it is most important to have a variety of target-background feature contrasts.

The model can be extended to natural scenes by considering detectability in small patches (size dependent on eccentricity) and computing a heterogeneous detectability map over the entire image. The training set of the current model includes fairly heterogenous large-scale textures, and many cases where the target fell on the boundary between two different-looking regions, so these cases should not be an issue for extending the model. To extend the model to different targets, it is necessary to recompute the detectability for that target, but this doesn't require retraining the CNNs, only the decision boundary between the target and the background. We believe this would be necessary for any human-like model of target detectability: detectability does not seem to be explained by low-level feature contrasts, so there is no simple function that could be computed in pixel space to predict discriminability of any target at any eccentricity on any background. The model is intended as a simulation; more testing on a broader range of stimuli and participants would be required before it could be deployed in a real-world application

@R1, R2: addition of related work The inspiration taken from Fridman et al. (2016) and the related Rosenholtz and Freeman work is the use of spatially larger feature-pooling regions to represent the feature compression in the visual periphery. Unlike Deza & Eckstein (2016), we compute detectability for specific targets based on the feature contrast with the background. Our signal-detection-based model is similar to Navalpakkam, V., & Itti, L. (2006) but considers more complex features to compute target detectability.