scientific reports

Check for updates

OPEN

Integrating Bayesian and neural networks models for eye movement prediction in hybrid search

Gonzalo Ruarte[1,2]✉, Gaston Bujia[1,2], Damián Care[1], Matias Julian Ison[3,5] & Juan Esteban Kamienkowski[1,2,4,5]

Visual search is crucial in daily human interaction with the environment. Hybrid search extends this by requiring observers to find any item from a given set. Recently, a few models were proposed to simulate human eye movements in visual search tasks within natural scenes, but none were implemented for Hybrid search under similar conditions. We present an enhanced neural network Entropy Limit Minimization (nnELM) model, grounded in a Bayesian framework and signal detection theory, and the Hybrid Search Eye Movements (HSEM) Dataset, containing thousands of human eye movements during hybrid tasks. A key Hybrid search challenge is that participants have to look for different objects at the same time. To address this, we developed several strategies involving the posterior probability distributions after each fixation. Adjusting peripheral visibility improved early-stage efficiency, aligning it with human behavior. Limiting the model's memory reduced success in longer searches, mirroring human performance. We validated these improvements by comparing our model with a held-out set within the HSEM and with other models in a separate visual search benchmark. Overall, the new nnELM model not only handles Hybrid search in natural scenes but also closely replicates human behavior, advancing our understanding of search processes while maintaining interpretability.

Visual search is a critical cognitive process involved in many different daily tasks[1,2]. For instance, every morning when preparing breakfast, we have to look for a cup, coffee, a spoon and so on. At the core of this task are the various eye movements, of which the most important ones are fixations and saccades. A fixation occurs when the eye remains focused on a particular location, typically for 200 ms. When this happens, new visual information about the scene is acquired. The rapid movements that occur between fixations, known as saccades, occur too quickly to provide new visual information but are essential to move the eyes towards a different location in the scene. The ordered sequence of fixations, known as a scanpath, is known to depend on several characteristics including the task performed[3], as well as different features from the perceived scene[1,4].

Extensive research on scene perception has mostly focused on image salience, which originally proposes that image features such as luminance, contrast, and orientation are combined to generate a salient part of the image[5,6]. However, more recent approaches have highlighted the significance of context (i.e. a kitchen scenery) in guiding attention during real-world perception[7] Moreover, Henderson and colleagues[8] have shown that the guidance of attention can be better explained using a 'meaning map', where the relevance of each scene patch is defined irrespective of the scene, rather than using a 'saliency map'[9]. In visual search, it has been noted that searching for an everyday object in an indoor scene is easier than searching for the same object outside of a scene[10].

Real-life searches often involve searching for multiple objects listed in memory. Going back to the breakfast example, we typically don't perform the search in a strict sequence, nor for exact items, and we look for any cup, spoon, plate, and so on. Thus, we need to look both into the scene and into the list of items held in memory (memory search). When observers search for any of several potential targets, the task is called Hybrid search

[1]Laboratorio de Inteligencia Artificial Aplicada (LIAA), Instituto de Ciencias de la Computación (ICC), CONICET - Universidad de Buenos Aires, Buenos Aires, Argentina. [2]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina. [3]School of Psychology, The University of Nottingham, Nottingham, UK. [4]Maestría en Explotación de Datos y Descubrimiento del Conocimiento, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires, Buenos Aires, Argentina. [5]Matias Julian Ison and Juan Esteban Kamienkowski contributed equally to this work. ✉email: gruarte@dc.uba.ar

(HS)[11]. First introduced by Schneider and Shiffrin's seminal work[12], it incorporates both visual search (VS) and memory search (MS).

While in real-life searches humans are looking at scenes (mostly everyday scenes), typically experiments are done with participants looking at images on a computer screen and an eye-tracking device to record their eye movements. To date, the large majority of hybrid search findings are derived from experiments utilizing images of artificial noise or shapes against blank backgrounds[13], potentially limiting their ecological validity[14]. In visual search, an increasing body of evidence emphasizes the significance of context (i.e. a kitchen scenery) in guiding attention during real-world searches[10].

In recent years, some computational models for Visual Search (VS) in natural scenes have emerged, with some primarily designed to emulate behaviour during specific tasks, such as searching for a class of object (e.g. cups), with limited generalization ability[15]. Other approaches focused on establishing a correlation between these models and the cognitive processes associated with the task[16–19]. Bayesian models have gained increased recognition across cognitive sciences[20]. Najemnik and Geisler[21] proposed the Ideal Bayesian Searcher model (IBS), using signal detection theory where each fixation is modeled as a decision on where to look next. This seminal paper led to related approaches in more recent years. Najemnik and Geisler[22] used Entropy Limit Minimization (ELM) instead of Bayesian integration for the decision-making process. These models accumulated evidence with each fixation to update their posterior probability. Zhou and Yu[17] further extended the ELM model to take the fixation duration into account in the evidence accumulation process. They also proposed a way to limit the amount of fixations that are used to decide where to go next. These models were tested on artificial images which did not depict natural scenes. Bujia et al.[16] proposed a way to extend the IBS to natural scenes. This involves the use of Deep Neural Networks to generate saliency maps for the model's prior, and employing a similarity metric between the target and the image regions in a natural scene when estimating the model's template response. This model was further improved in[23] by using the attention map of[19] as template response and in[24] by using ELM instead of the guidance and integration from IBS. Rashidi et al.[18] developed a new model of target detectability, which was then tested in an IBS in natural scenes. To date, there remains a lack of consensus regarding the most appropriate metrics, although ongoing research endeavours are addressing this gap[23].

Remarkably, there appears to be a lack of computational models dedicated to hybrid search tasks in natural scenes. In the current study, we extend the model presented in[24] to hybrid search tasks. Briefly, we explore different ways of limiting the amount of fixations used to decide where to go next, and we improve the peripheral visibility of the model. Altogether, we achieved higher performance metrics as well as increased computational efficiency.

One of the major bottlenecks limiting comparisons between computational models and human behaviour is the availability of well-documented open datasets. A secondary objective of the current study is to generate a visual and hybrid search dataset comprising actual scanpaths from participants. Typically, participants are instructed to look for objects across a large number of images depicting natural scenes. In this work, we also present the Hybrid Search Eye Movements (HSEM) dataset, which will be available after publication.

We use a fixed subset of our hybrid search dataset to test the various modifications to the model, and its hybrid search capabilities. Furthermore, the final version is validated on a different subset of the same dataset. Moreover, we gauge its performance against state-of-the-art visual search models in the ViSioNS benchmark. The overall goal of the present work is to advance in the path of bringing these models closer to humans in a wider range of tasks while maintaining interpretability.

Methods

Hybrid search eye movements (HSEM) dataset

Participants

The participant pool consisted of two groups: 18 individuals from UBA (Universidad de Buenos Aires) and 28 from UoN (University of Nottingham). Two from the UoN were discarded due to a poor overall performance (less than 55% of accuracy). This gives a total of 44 participants (18 UBA and 26 UoN). The final sample involved 23 male and 21 female participants between 19 and 40 years old (24.5 ± 5.3 years old). All participants were naive to the experiment's objectives and possessed normal or corrected-to-normal vision. Ethical approval for the study was obtained from the respective Ethics Committees of each university (Protocol 284 from the Instituto de Investigaciones Medicas "Alfredo Lanari"—University of Buenos Aires, and Protocol F1317 from the University of Nottingham). All participants provided written and signed informed consent. All procedures were conducted in accordance with approved guidelines and regulations, in agreement with the Helsinki declaration.

Stimuli

Each stimulus (or image) was constructed by superimposing a real-world background image (depicting outdoor scenes -e.g. forest-, and indoor scenes -e.g. shelf-) (Fig. 1A) with 16 individual items (objects, animals or full-body humans) (Fig. 1A). Every item and background image was taken from COCO[25] or ImageNet[26] datasets. A total of 210 stimuli were prepared.

The items were resized so that they would fit in the background image, and placed according to scene syntax (i.e. no major violations of support, interposition, position, and size). A bounding box was defined around each item of a size of 0.33% + − 0.11% of the background image on average (Fig. 1B,C). The maximum size of a bounding box for each item was 80x80 pixels, and the background image sizes were 1280 × 1024 pixels.

The 16 individual items added to the scene comprised the visual set of the stimulus, thus, the visual set size (VSS) was fixed across the experiment (VSS = 16). Each item was presented only one time in the whole experiment, except for 53 items that appeared in two separate visual sets, and 5 items that appeared in three different visual sets. This yielded a total of 1569 items across visual sets.

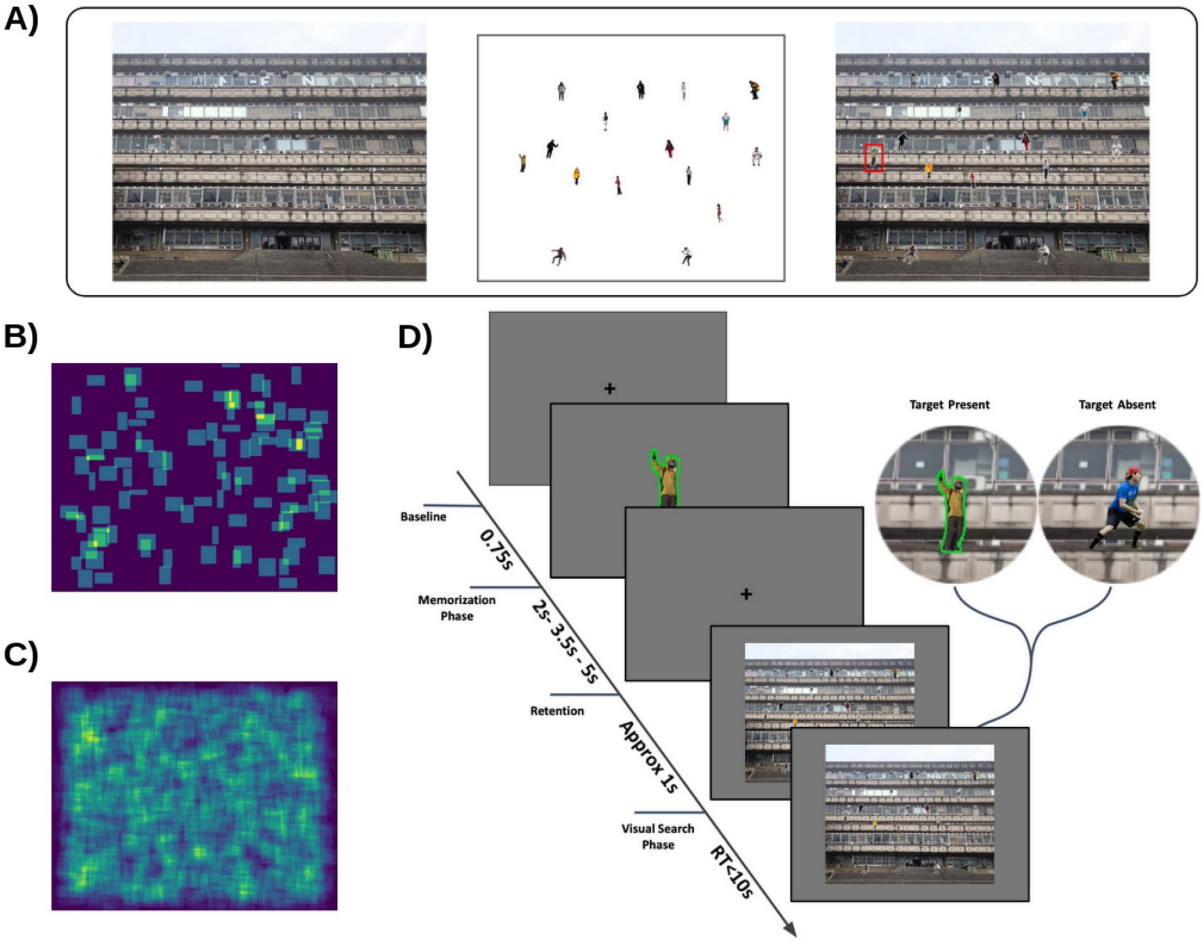

Fig. 1. Stimuli and Task. (**A**) Stimuli were constructed by superimposing real-world background image with 16 individual items. All items appearing in the image were taken from COCO and ImageNet. Due to possible copyright restrictions, the background image shown in the experiment was substituted here by a picture taken by one of the authors (GR). (**B, C**) Heatmap of the bounding boxes from the items presented at visual search phase across trials. (**B**) Targets, (**C**) Distractors. (**D**) Time progression of the trial. A detail of a highlighted version of the target at present condition and the distractor item at absent condition. After a fixation dot a variable number of potential targets were presented (Memory Set Size, MSS = {1, 2, 4}), followed by a retention period. Search screen always have 16 items (Visual Set Size, VSS), and participant's task was to respond whether one or none of the potential targets was present. In target absent trial a cat different from the targets is included in the same location. See[27].

For each stimulus, a set of 1, 2, or 4 items were selected for memorization (memory set). In this context, a trial was defined as a stimulus plus its memory set. Within each trial, both the items in the memory set and the ones in the visual set belonged to the same category (objects, animals, or people). Moreover, items in the memory set of a given trial did not appear in the memory set of other trials. There were two exceptions to this rule: two items appeared in the memory sets of two different trials. The visual and memory sets of a given trial could share at most one item: If they did share an item the trial was labelled as a target-present trial-with the shared item being the target-, and the trial was labelled as a target-absent trial otherwise. Targets were unique among target-present trials.

Only the target-present trials were considered for the current analysis. Three stimuli had a different resolution than 1280 × 1024. Trials showing any of those three stimuli were excluded from all analyses. This yielded a total of 102 trials: 35 with a memory set size (MSS) of 1, 34 with a memory set size (MSS) of 2, and 33 with a memory set size (MSS) of 4.

Task

Initially, participants were exposed to the memory set for 2 seconds if the MSS was 1, for 3.5 seconds if the MSS was 2, and for 5 seconds if the MSS was 4. To do so, four locations in the screen were previously defined. If the center of the screen was (0,0) then the four locations were: (− 300,0), (− 150,0), (150,0) and (300,0) in pixels. For each participant and trial, the items in the memory set were randomly placed in one of these locations accompanied by random vertical shifts of 0, ±50 or ±100 pixels. Subsequently, participants viewed the stimulus and were allotted 7 seconds to locate the target object (Fig. 1D). They were required to press one

key if they identified the target and a different key if they believed the target was absent. The experiment was divided into seven blocks of 30 trials and different participants were presented with a different order of trials. Every participant had to do each trial only once. The experiments were implemented and executed in Psychopy v2022.2.5 (https://www.psychopy.org/)[28].

Data collection

All the equipment used in the acquisition was the same at both universities except for the use of a 'qwerty' keyboard to introduce responses at UBA and a response pad at UoN.

The stimuli were presented on a BenQ XL2420Z monitor at UBA and in a LENOVO Y27Q-20 monitor at UoN with a screen resolution of 1920 × 1080 pixels. Participants were placed at a distance of approximately 65 cm from the monitor. All stimuli were presented using Psychopy v2022.2.5 software (https://www.psychopy.org/)[28] and synchronized with the eye tracker using ioHub (https://psychopy.org/api/iohub/index.html).

Eye-movement tracking (ET) data was recorded using an EyeLink 1000 Plus remote system in monocular mode and a sampling rate of 500 Hz. At UoN an EyeLink target sticker was used to ensure movement stabilization while a chinrest was used at UBA for head-stabilization.

There were issues in the collected data of 24 participants so some of their trials are missing. One participant had 12 trials missing but the other 23 had 1 or 2 trials missing at most. This meant a total of 4442 valid trials, 1520 for MSS 1, 1485 for MSS 2 and 1437 for MSS 4.

Eye movement pre-analysis

Fixations, saccades and blinks were detected using the built-in EyeLink algorithm. A preprocessing pipeline on the fixations followed the detection process. It included, first, checking if the Target is found, i.e. if the fixation fell less than 68 pixels away from the center of the target. Second, collapsing nearby consecutive fixations into one. Third, shifting fixations outside the image boundaries to the nearest point within the image. Finally, removing fixations after the target is found from the scanpath. See Supplementary Material and[23] for more details.

Models

Metrics

The metrics used were the same as the ones used in ViSioNS Benchmark (https://github.com/NeuroLIAA/visions)[23], but they were adapted to group results of different MSS separately.

- *Efficiency*: We measure the proportion of targets found for a given number of fixations and report the area under this curve (AUC), noted as AUCperf[29].
- *Scanpath similarity*: Multi-Match (MM)[30] is a pairwise similarity metric that represents saccades as vectors in a two dimensional space. Then, in this space, scanpaths are sequences of vectors. Two such sequences (which can differ in length) are compared on four dimensions: vector shape, vector length (saccadic amplitude), vector position, and vector direction for a multidimensional similarity evaluation. The temporal dimension was excluded as we were not considering fixations' duration. For each model, its scanpaths were compared to each participant's scanpaths (after reducing them to match the model's scanpaths dimensions) and the outcome was averaged across all dimensions and participants, resulting in a single scanpath similarity value for each image (human-model Multi- Match or hmMM). The same was done with the participants' scanpaths, comparing them within themselves, and thus obtaining a human ground truth for each image (within-human Multi-Match or whMM). These operations were performed for every scanpath with length greater than two in which the target was found. We report AvgMM as the mean value of hmMM, in the case of models, and whMM, in the case of participants, over all images. The Pearson correlation (Corr) between hmMM and whMM is computed for each model.
- *Human scanpath prediction (HSP)*: For a given participant's scanpath, we evaluated how well each model predicted the next human fixation based on the scanpath history[31]. The key idea behind this method is to force the models to follow the participant's scanpath (ignoring its own predictions). At each step, each model creates what is called a "conditional priority map" (a priority map based on the participant's previous fixations from where the next fixation is sampled) and we compared the position where the model's fixation would land (i.e. its prediction of the next fixation) against the participant's fixation[31]. By using the latter as the ground truth, this allowed for the computation of well-established metrics such as area under the curve (AUC)[32], Normalized Scanpath Saliency (NSS)[33], Information Gain relative to the center bias (IG)[34], and information gain relative to the uniform model (LL)[34]. Here, we present the average of these measures across fixations named as AUChsp, NSShsp, IGhsp and LLhsp, respectively.

Control conditions

In this project we wanted the model to behave like humans, which was why their cumulative performance and multimatch were the ground truth for our models.

The other three control conditions were the ones used in[23], which are similar to those in[31]. Their goal was to provide a lower and upper bound to the performance of the evaluated models. The lower bound was determined by the uniform and center bias models, and the upper bound by the Gold Standard model (GS). The uniform model predicts fixations to be uniformly and independently distributed over the image. The center bias model stems from[35], who show that people have a tendency to look at the center of images. As this phenomenon occurs mainly during free-viewing experiments[36], we used a Gaussian Kernel Density Estimate (GKDE) over all publicly available fixations in all images of the CAT2000 training dataset[37], excluding the first fixation as it was forced. Lastly, the Gold Standard model predicted fixations of each participant by means of a GKDE over all fixations of other participants on the same image[34]. In these last two cases, the bandwidth of the GKDE has

been selected to yield maximum log-likelihood in a 5-fold cross validation paradigm. Since the baseline models are not visual search models themselves (i.e. they don't generate scanpaths), only the fixation-by-fixation metrics (HSP) were computed on these.

Finally, in order to sort the models by a compound variable, we defined their score as the average of the individual normalized metrics, relative to Humans in the case of efficiency and scanpath similarity metrics (AUCperf, AvgMM) or by the Gold Standard in the case of the scanpath prediction metrics (AUChsp, NSShsp, IGhsp, LLhsp). For AUCperf, this was done by performing $-|AUCperf_{Model} - AUCperf_{Humans}|$, as we intended to maximize similarity to human performance. The rest of the components of the score were computed as $(Value_{Model} - Value_{Reference})/Value_{Reference}$, where Reference stands for Humans in AvgMM and for GS in all other cases. Corr did not have a reference value, so $Corr - 1$ was the component used in the score. Notice that a value of 0 for each and every component of the score means that the model was behaving like humans according to the corresponding metric. A score of 0 means that a model behaved like humans according to every metric.

nnELM model structure

The[24] model consists of a Bayesian Framework that incorporated information iteratively with each saccade (see Fig. 2 for an overview). The posterior was computed as follows:

$$p_i(T) = \frac{prior(i)\prod_{t=1}^{T} exp\left(d_{ik(t))}^2 W_{ik(t))}\right)}{\sum_{j=1}^{n} prior(j)\prod_{t=1}^{T} exp\left(d_{jk(t))}^2 W_{jk(t))}\right)} \tag{1}$$

W is what we called "Visual Evidence" (template response in[16]) whose values were sampled from a 2D normal distribution with mean and variance as follows:

$$\widetilde{W}_{ik(t)} \sim \mathcal{N}\left(\widetilde{\mu}_{ik(t)}, \widetilde{\sigma}^2_{ik(t)}\right) \tag{2}$$

$$\widetilde{\mu}_{ik(t)} = \mu_{ik(t)} \cdot \left(d_{ik(t)} + \frac{1}{2}\right) + \phi_i \cdot \left(\frac{3}{2} - d_{ik(t)}\right) \tag{3}$$

$$\widetilde{\sigma}_{ik(t)} = \frac{1}{a \cdot d_{ik(t)} + b} \tag{4}$$

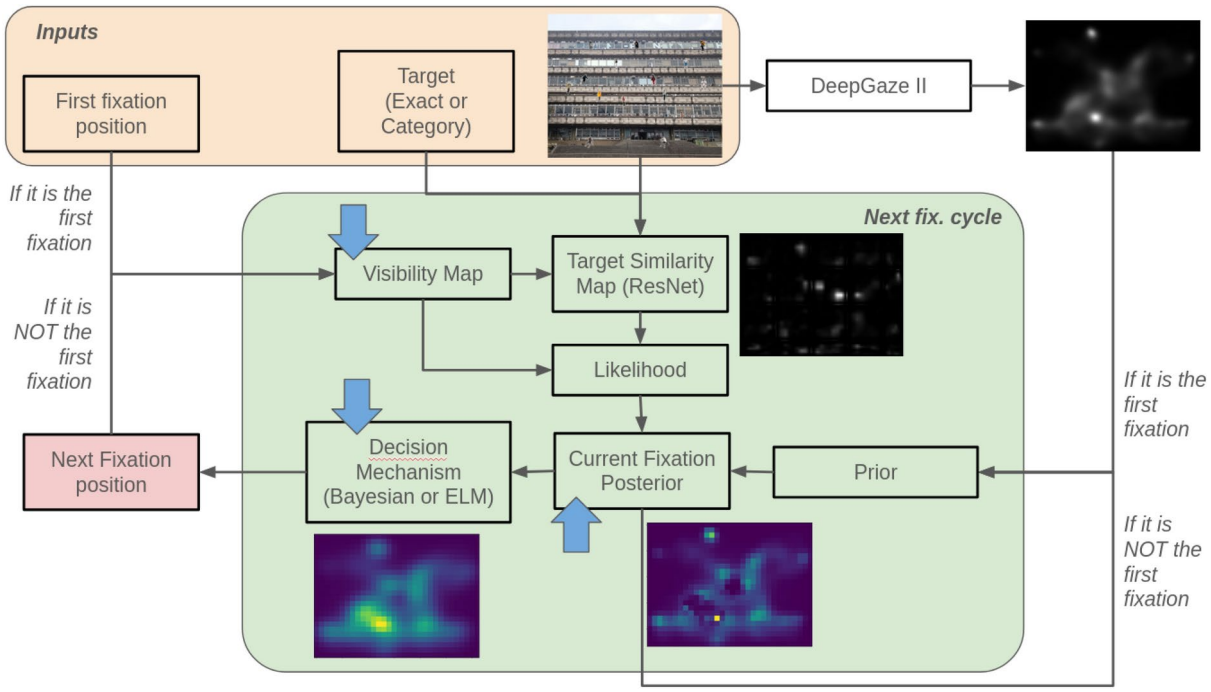

Fig. 2. The model receives the image, the target and the position of the first fixation. Before it starts moving it calculates the first prior using the DeepGaze II model which gives the first gist of the image. The model estimates the next fixation position which updates the input values. Also, the posterior becomes the prior of the next cycle. Blue arrows indicate where the model was updated in the current version: A new version of the visibility map is presented, a limited working memory is introduced in the Posterior estimation, and a decision criterion is used to adapt the model to Hybrid Search.

where in both equations k(t) is the fixation at step t; i and j are locations in the image; ϕ_i is a number between $-$ 0.5 and 0.5; $\mu_{ik(t)}$ is either -0.5 or 0.5; and d is what we called visibility map (also called detectability or eccentricity).

The prior in the first step of the iteration is a saliency map generated with Deepgaze II[38]. It is akin to the visual information a human would acquire at first glance.

The visibility map (d) is a 2D Gaussian $N(\mu, \Sigma)$ where μ is the 2D coordinate in pixels and the covariance matrix $\Sigma = \left(\begin{smallmatrix} 2600 & 0 \\ 0 & 4000 \end{smallmatrix}\right) pxs^2$, whose values were defined in[16] according to the subjects' eccentricity measured in[21].

μ is a mask derived from the specific target location:

$$\mu_{ik(t)} = 1_{(i=targetlocation)} - 0.5 \tag{5}$$

The Target Similarity Map (ϕ) captures how similar the target is to each region of the image. It is computed by initially dividing the image into blocks of size 224x224, which are then forwarded through a pre-trained convolutional neural network. Afterwards, the target item is rescaled to 128x128 and forwarded through the same network. Then, the results of the last convolutional layer (feature map) for each image block are convolved with the last feature map of the target. The results of these convolutions are rescaled back to 224 × 224 blocks. Finally, these results are put back together, yielding the Target Similarity Map. The procedure is similar to the one used in[23], but here a ResNeXt-101 32 × 8 d[39] was used instead of a VGG-16[40].

With the posterior and d, the next fixation is computed as follows:

$$E[\Delta H(T+1)|k(T+1)] = \frac{1}{2}\sum_{i=1}^{n} p_i(T) d^2_{ik(T+1)} \tag{6}$$

Equation (6) shows the Entropy Limit Minimixation (ELM)[22]. The task for which the ELM was originally derived was one in which the stimulus consisted of spatially correlated noise with an embedded target (Gabor patch). The derivation assumed that all the targets were non-overlapping (formally, they were orthogonal to each other). Moreover, the Visual Evidence (W in Eq. 1) was considered independent across possible target locations. Neither of these assumptions applies to real-life stimuli. This means that, in theory, the ELM is not necessarily a good approximation of the Ideal Bayesian Searcher in our setup. Nevertheless, empirically, we evaluated the IBS and ELM models and found that the latter significantly reduced processing times without compromising any of the metrics[24]. One potential explanation for this stems from the fact that humans are suboptimal searchers, as highlighted by[17].

Both the model and the metrics were implemented in Python 3.10.

Motivation for changes and explanation

Visibility map

The model's visibility map has an impact in two mechanisms: It modulates the information input from the image, but it also has an impact on the inhibition of return (IOR), as the model naturally reduces the probability of exploring the areas that were already visited, without adding additional parameters. The proposed model seamlessly achieves this compromise, extending previous works[16,21].

In Eq. (1), the term $W_{ik(t)} d^2_{ik(t)}$ was used to restrict the visual evidence to the foveal region while also applying an inhibition of return. $W_{ik(t)}$ had values ranging from $-$ 0.5 to 0.5 while $d^2_{k(t)}$ had positive values if i was near the center of $d^2_{k(t)}$ and it had values closer to 0 if i was far from the center of $d^2_{k(t)}$. This also implied that $W_{ik(t)} d^2_{ik(t)}$ would have been close to 0 if i was far from the center of the visibility map, whereas its values would have been either positive—when the target was there—or negative—when the target was not there—if i was near the center. These negative values were also the correlates of the inhibition of return in future steps.

Based on this, we proposed using the same principle for the periphery by defining $d'_{ik(t)}$ as follows:

$$d'_{ik(t)} = \frac{d^p_{ik(t)} \cdot max\left(d^2_{ik(t)}\right)}{max\left(d^p_{ik(t)}\right)} \tag{7}$$

This definition forced the range of values of d' to be the same as $d^2_{k(t)}$ while also having strictly positive values in a broader radius from the center. $W_{ik(t)} d'_{ik(t)}$ worked the same way as $W_{ik(t)} d^2_{ik(t)}$ but i could be further from the center and still have a non-zero value.

Finally, we needed the likelihood to incorporate information from both the periphery and the fovea, while only applying an inhibition of return to the fovea. This was achieved with a *max()* operation:

$$max\left(W_{ik(t)} \cdot d^2_{ik(t)}, W_{ik(t)} \cdot d'_{ik(t)}\right) \tag{8}$$

The zero values from $W_{ik(t)} d^2_{ik(t)}$ that were far from the center would be overridden by the positive values from $W_{ik(t)} d'_{ik(t)}$, effectively adding information from the periphery to the model. Meanwhile, the negative values from $W_{ik(t)} d'_{ik(t)}$ that were far from the center would be overridden by the zero values from $W_{ik(t)} d^2_{ik(t)}$ while the values in the foveal region remained the same, which meant that the inhibition of return also remained unchanged.

This modification allowed the model to incorporate information from the whole field of vision while keeping the inhibition of return restricted to the fovea. $max(d^2_{ik(t)})/max(d^p_{ik(t)})$ was the term that allowed all of these behaviours at the same time.

The term d^2 was also used in Eqs. (3)–(5). As for Eqs. (3) and (4), $W_{ik(t)}$ was sampled from a normal distribution and both μ' and σ' depend on d. The formula in Eq. (3) gives a lot of weight to the real target within the fovea and less weight elsewhere (first addend). At the same time it gives more weight to the real target and distractors in the peripheral vision and less weight in the fovea (second addend). That is to say, it will only pay attention to an object within the fovea if it is indeed the real target, but it can also pay attention to other objects that are not within the fovea.

$\widetilde{\sigma}_{ik(t)}$ indicates how much noise there is: the idea is that we want little noise within the fovea and more noise as we move further away.

Previously these effects in mu and sigma did not matter much in the sense that everything in the peripheral vision was discarded afterwards, so only the desired effects within the fovea were what mattered. After the change in Eq. (1) the information from the peripheral vision started to be considered as well, so both of these formulas started mattering much more. Still, there was a distinction: in $\widetilde{\mu}_{ik(t)}$ we wanted to distinguish distractors from the real target, but by definition, they are called distractors because we are not sure whether they are the real target or not until we look at them, so we only cared about making a clear distinction between target and distractor within the fovea. Then it made sense to use $d^2_{ik(t)}$ in Eq. (3) instead of $d'_{ik(t)}$:

$$\widetilde{\mu}_{ik(t)} = \mu_{ik(t)} \cdot \left(d^2_{ik(t)} + \frac{1}{2}\right) + \phi_i \cdot \left(\frac{3}{2} - d^2_{ik(t)}\right) \tag{9}$$

On the other hand, we cared about $\widetilde{\sigma}_{ik(t)}$ all over the visual field, so it made sense to use $d'_{ik(t)}$ in Eq. (4):

$$\widetilde{\sigma}_{ik(t)} = \frac{1}{a \cdot d'_{ik(t)} + b} \tag{10}$$

In Eq. (5) we did not want a bigger foveal region so no changes were made.

Visual working memory

We did not include the modifications to the peripheral vision in the following equations to maintain simplicity.

In visual search and hybrid search, the visual working memory is the amount of fixations being remembered while performing the search. The first strategy to limit the model's visual working memory was to take the last N fixations and completely forget the rest, like what was done in[17]. Equation (1) changed as follows:

$$p_i(T) = \frac{prior(i)\prod_{t=min(1,T-x)}^{T} exp\left(d^2_{ik(t))}W_{ik(t))}\right)}{\sum_{j=1}^{n} prior(j)\prod_{t=min(1,T-x)}^{T} exp\left(d^2_{jk(t))}W_{jk(t))}\right)} \tag{11}$$

where x is the amount of fixations that the model remembers.

This brought the following variant into the fold: if the model forgets fixations it made, it might made sense to forget the initial prior as well. The equation remained the same when $T \leq x$, but when $T > x$ the equation changed as follows:

$$p_i(T) = \frac{\prod_{t=min(1,T-x)}^{T} exp\left(d^2_{ik(t))}W_{ik(t))}\right)}{\sum_{j=1}^{n}\prod_{t=min(1,T-x)}^{T} exp\left(d^2_{jk(t))}W_{jk(t))}\right)} \tag{12}$$

We also developed another approach in which instead of completely forgetting fixations we applied an exponential decay to the working memory. The idea was to progressively remember fixations less and less as the search went on. The parameters of the exponential were selected so that after approximately 8 fixations, the model will almost completely forget about the current one:

$$p_i(T) = \frac{prior(i)\prod_{t=1}^{T} exp\left(d^2_{ik(t))}W_{ik(t))}exp((T-t)/b)\right)}{\sum_{j=1}^{n} prior(j)\prod_{t=1}^{T} exp\left(d^2_{jk(t))}W_{jk(t))}exp((T-t)/b)\right)} \tag{13}$$

this version could also forget the prior like this:

$$p_i(T) = \frac{exp(log(prior(i))exp((T-t-1)/b))\prod_{t=1}^{T} exp\left(d^2_{ik(t))}W_{ik(t))}exp((T-t)/b)\right)}{\sum_{j=1}^{n} exp(log(prior(j))exp((T-t-1)/b))\prod_{t=1}^{T} exp\left(d^2_{jk(t))}W_{jk(t))}exp((T-t)/b)\right)} \tag{14}$$

where in both cases $b > 0$ was a parameter to explore.

Hybrid task adaptation

In order to make IBS work in hybrid tasks, we had to allow it to process more than one object. We did that by computing the likelihood and posterior corresponding to each object in the memory set at each step. Those posteriors would have to be somehow combined or selected. To do so, we implemented four strategies (Fig. 3):

- Random: Selecting the posterior of a random object.
- MinEntropy: Selecting the posterior with less entropy (it can be a 2D or regular entropy).
- CorrectTarget: Selecting the posterior of the real target. This one served as a baseline.
- LikelihoodMean: Taking the mean of the foveated visual evidences ($W_{ik(t)} \cdot d^2_{ik(t)}$) over the memory set and computing a single likelihood and posterior with that.

Validation set and benchmark

Participants from the Hybrid Search dataset were split up into train and validation sets using the train_test_split function from Scikit learn Python package (https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.train_test_split.html) with 21% validation set size and stratified by subject accuracy. This procedure resulted in 34 and 10 participants in train and validation sets respectively.

Also, an external validation was performed using datasets and measures available in the ViSioNS benchmark of visual search in natural scenes (https://github.com/NeuroLIAA/visions)[23]. Briefly, the Interiors dataset from our previous work comprises 134 images and 57 participants who had to look for an exact target in photos of interiors[16]. The MCS and COCOSearch18 datasets used by Zelinsky and collaborators comprises 1687 and 612 images and 23 and 10 participants respectively[29] that had to look for any target belonging to a given category (for instance, a "cup") in a varied set of images. The Unrestricted dataset collected by Zhang, Kreiman, and collaborators comprises 234 images and 15 participants who also had to look for a target within a category as MCS and COCOSearch18[19].

Results

Human behavior

The response time of the participants performing the Hybrid Search followed a logarithmic dependence with the memory set size (MSS), consistent with previous findings[27] (Fig. 4A). Target detection accuracy decayed with MSS = 4 (Fig. 4B).

In order to complete the task, participants performed a total of 9.03 ± 0.15, 10.45 ± 0.17, and 12.79 ± 0.19 fixations across different images, for MSS equal to 1, 2, and 4 respectively (Fig. 4C,D). Most fixations were on new items, as the mean number of re-fixations were 0.59 ± 0.02, 0.72 ± 0.03, and 1.02 ± 0.03 respectively. The saccade direction distribution presented a bias towards horizontal as is typically seen in natural scene exploration and search (Fig. 4E). The main sequence of saccades (Fig. 4F) was also similar to previous reports in almost every task[41].

Improving visual search model's performance: effects of visibility and working memory capacity

Before extending the model from Visual Search (MSS=1) to Hybrid Search (for trials with MSS = 2 and 4), we focused on improving the Visual Search model. Recently[24], proposed using the Entropy-Limit Minimization (ELM) as a model of decision and guidance (see also[17,18]). We suggested two modifications to achieve a greater similarity between the model and humans: Weighting the tails of the visibility map distribution to model the peripheral vision and limiting the visual working memory capacity. The previously defined train set (79%) was used to tune the parameters of the modifications and both the validation set (21%) and the ViSioNS benchmark[23] were used to validate those parameters.

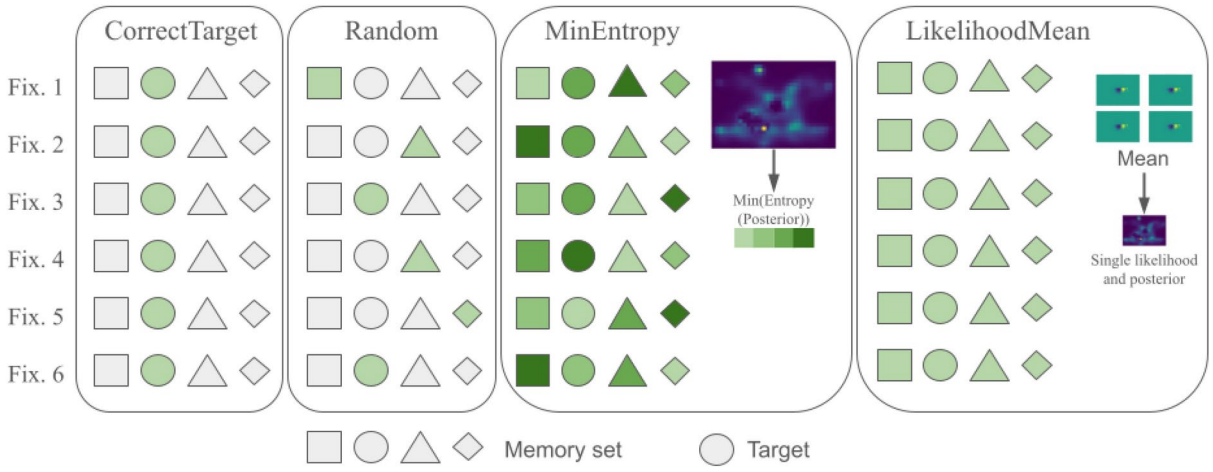

Fig. 3. Overview of the selection criteria. CorrectTarget, Random and MinEntropy perform a selection while LikelihoodMean does not..

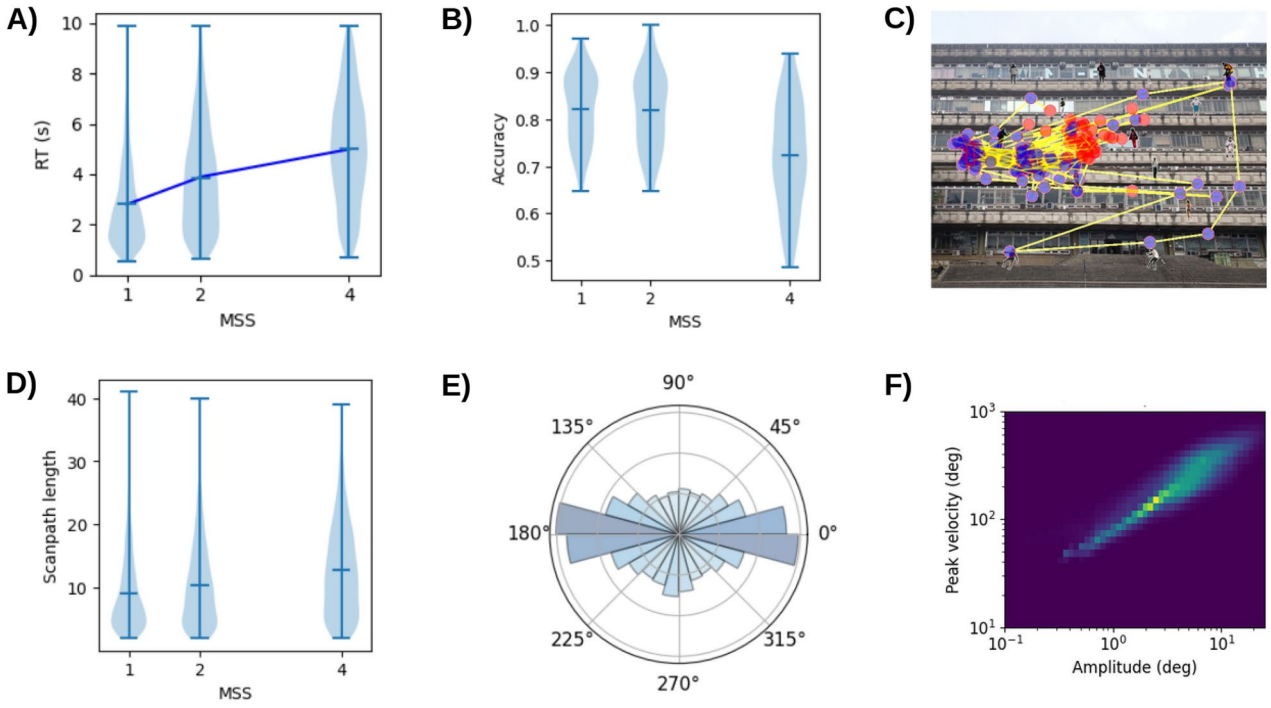

Fig. 4. Participants' Eye Movements. (**A, B**) Performance measured as Response Times (on Target Present trials) and Accuracy (on Target Present trials) as function of Memory Set Size (MSS). (**C**) Scanpaths of all participants in a sample image. (**D**) Scanpath length as function of MSS. (**E**) Saccade direction. (**F**) Main sequence of saccades.

Until[24] the model acquired new information from the fovea only (see Fig. 5A). Since the visibility map followed a Gaussian distribution, the likelihood value in the peripheral region remained constant at 1 (given that $exp(0) = 1$), which was non-informative. In contrast, the foveal region could display positive values. Values larger than 1 in a region within the fovea suggested the presence of a relevant feature, and most likely the model would have fixated on that location next, especially if it had been close to the target already. When values were close to 0, it indicated that there was no relevant information, which elicited an inhibition of return across the entire fovea. After this happened, the model would rely on the prior to guide the next fixation. Recursively, if an inhibition of return had been applied to every fixation, then the saliency map (initial prior) would have dictated the location of subsequent fixations. To improve this approach, we made the likelihood more informative by allowing it to capture information from the peripheral vision. We modified Eq. (1) by dividing the visual evidence into foveal and peripheral terms that compete with each other according to a max operation, which also kept the inhibition of return restricted to the foveal region (see Eq. 6). This led to the likelihood displaying values larger than 1 (see Fig. 5A). We also replaced the term d by d^2 in Eq. (3), and d by d' in Eq. (4) (see Eqs. 9, 10). These changes made the model more consistent with what is currently known about the foveal and peripheral areas (see Fig. 5B,C).

This model outperformed the nnELM model[24] in every metric (Fig. 5D). In order to choose a value for the exponent (p in Eq. 7) we found a compromise between more general metrics like the AUCperf and MM and the fixation-based metrics, in particular IG and LL. Thus, the best model had $p = 0.2$ (Fig. 5E). It is important to note that small changes in the exponent did not significantly affect the results.

In a previous version of the model, we observed that after some time the saccades tended to be longer, presumably because the saliency map was what primarily dictated where the model went. By improving the visibility map the saccade amplitude also got closer to human saccades along the whole scanpath (Fig. 5B,C,F), without any ad-hoc constraint on the saccade amplitude[17].

Human participants have memory limitations that makes them less likely to remember which regions have already been visited in long searches. Meanwhile, previous IBS and ELM models[16,23,24] had unlimited memory capacity and would eventually find the target if the scanpath was large enough. To address this discrepancy, we implemented memory constraints within the model, thereby bringing its behavior into closer conformity with that of human participants during extended searches. Because of that, we decided to take into account only the last N fixations and completely forget the rest, like what was done in[17] (see Eq. 10).

We had a non-trivial prior, which is why it was important to decide whether the model could forget the prior or not as if it was another fixation. Firstly, we took the best model from the previous experiment, made it forget the prior and explored the number of fixations that it could maintain in visual working memory (see Eq. 11). In this case all limited memory models outperformed the unlimited memory model, with 4 being the number of fixations that worked best (Fig. 6A,B). It is worth mentioning that the metrics' values were very close to those of

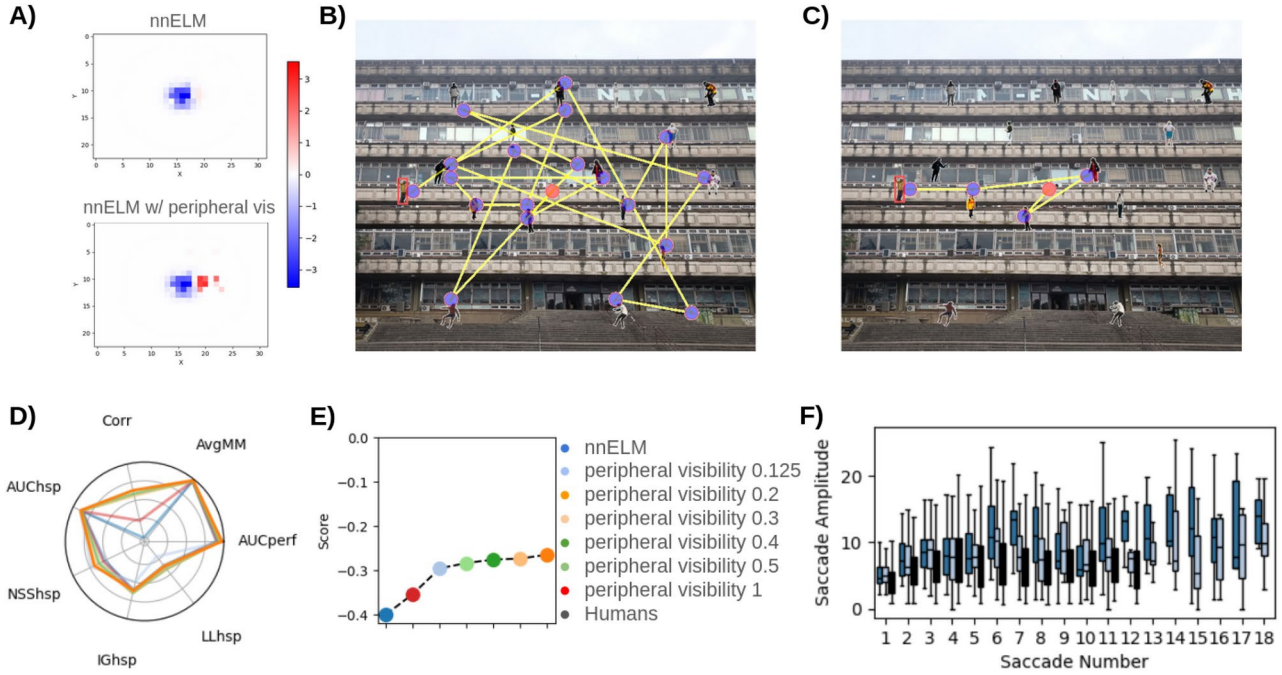

Fig. 5. Peripheral Visibility. (**A**) Change in likelihood after adding peripheral visibility to the model. (**B**) Model scanpath before adding peripheral visibility to the model. (**C**) Model scanpath after adding peripheral visibility to the model. (**B, C**) The red dot is the initial fixation, the blue dots are the other fixations and the yellow lines are saccades. (**D**) Model performance according to each metric. With $p = 0.2$ the model has a higher value for almost every metric. (**E**) Model score. With $p = 0.2$ the model has a greater score than the rest. (**F**) Saccade amplitude boxplots for the model with and without peripheral visibility, and how they compare against humans, only for trials in which the target was found. 90% quantile of the saccade number was used to filter longer scanpaths.

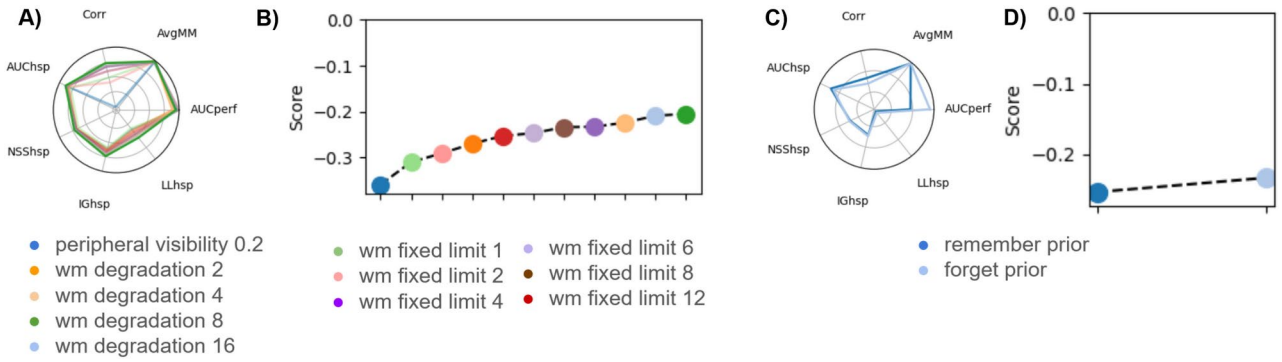

Fig. 6. [Limited Memory. (**A**) Model performance according to each metric. Degrading the memory with an exponential (exponent = 0.125) works best for almost every metric. (**B**) Model score. An exponent of 0.125 yields the highest score. (**C**) Forgetting the prior doesn't yield better results with the exception of corr and AUChsp. (**D**) The score remains higher when forgetting the prior].

the models with a visual working memory of 6, 8 and 12 fixations (Fig. 6A,B), which is consistent with[17], who find good values to be between 4 and 12 fixations.

Secondly, models of other memory domains suggest a decay in the retention of previous items instead of a fixed all-none limit[42]. Thus, we replaced the stepwise function by an exponential decay and explore the time constant. We found that the best model has a decay of 0.125, which corresponds to a decrease of 0.4 for 4 fixations and 0.63 after 8 fixations (Fig. 6A,B). This worked even better than the fixed-limit models.

Finally, we took the best model and made it remember the prior, and found that it was worse overall (Fig. 6C,D).

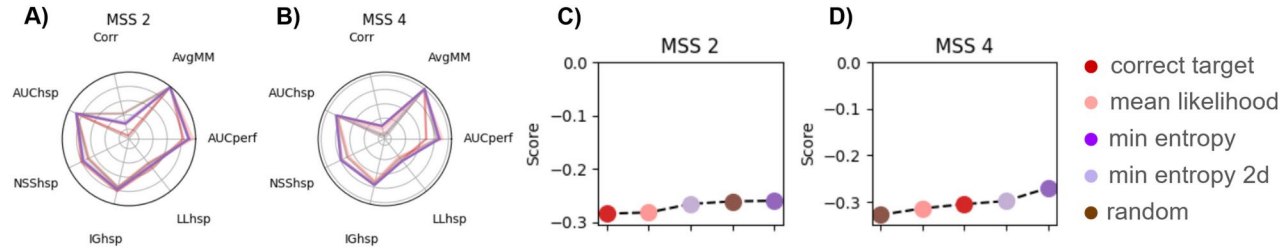

Fig. 7. Hybrid Search (**A, B**) Model performance according to each metric for MSS 2 trials and MSS 4 trials respectively. MinEntropy works better for almost every metric. (**C, D**) Model score for MSS 2 trials and MSS 4 trials respectively. MinEntropy yields the highest score in both cases.

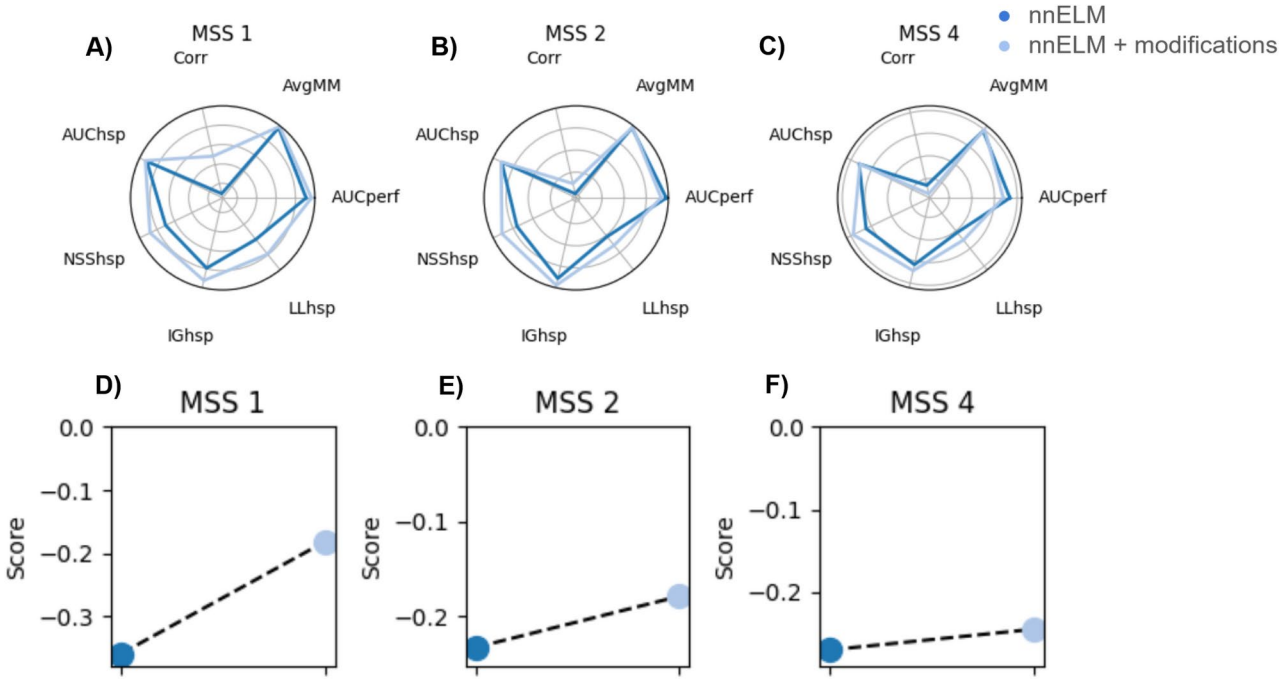

Fig. 8. Validation. (**A**–**C**) Model performance according to each metric for MSS 1, 2 and 4 trials respectively. The modifications improve every metric for MSS 1, and almost every metric for MSS 2 and 4. (**D**–**F**) Model score for MSS 1, 2 and 4 trials respectively. The modifications improve the overall score for every MSS.

Adapting the model to hybrid search: target selection criteria

For the rest of the experimentation we proceeded with the best model identified so far, which included an improved visibility map (with an exponent of 0.2) and an exponential decay of previous fixations (with an exponent of 0.125 and considering the prior as fixation zero). For a single target, the model determined where to look next based on a single probability map. However, when dealing with many potential targets, the observer has to decide at each step which map to use or how to combine them. Here we propose that the posterior probability map for every target is estimated in parallel, and the model guides the next fixation towards the map that contains the largest amount of information (i.e., minimum entropy). We compared this criterion with different alternatives, such as selecting maps at random in each fixation or knowing the correct target in advance and only focusing on that map. The MinEntropy approach worked better for memory set sizes of 2 (see Fig. 7A,C) and 4 (see Fig. 7B,D).

Model validation

A validation set stratified by accuracy was separated at the beginning of the experiment. The best model also showed better performance on this validation set for the three MSS values (Fig. 8). In the case of nnELM we used the CorrectTarget strategy for MSS 2 and 4 because originally it could not perform hybrid search tasks.

Moreover, our best model showed better performance in every metric than previously published models such as nnIBS and other state-of-the-art models. Table 1 corresponds to the average across the four datasets of the ViSioNS Benchmark (to date these are: Unrestricted, Interiors, COCOSearch18 and People)[23]. This result lowered the possibility of overfitting to the HSEM dataset.

Model	AUCperf	AvgMM	Corr	AUCshp	NSShsp	IGhsp	LLhsp	Score
Human performance	0.617	0.871	–	–	–	–	–	0.0
gold standard baseline	–	–	–	0.908	2.922	2.746	2.248	0.0
nnELM + mods	0.673	0.852	0.256	0.807	1.801	1.872	1.115	− 0.303
nnELM	0.52	0.843	0.219	0.796	1.445	1.623	0.902	− 0.364
nnIBS	0.507	0.843	0.228	0.799	1.444	1.22	0.499	− 0.411
Gupta et al.[48]	0.465	0.828	0.169	0.681	1.006	− 0.866	− 1.312	− 0.691
uniform baseline	–	–	–	0.5	0.0	0.499	0.0	− 0.817
center bias baseline	–	–	–	0.66	0.543	0.0	− 0.499	− 0.827
Zhang et al.[19]	0.569	0.79	0.122	0.619	0.85	− 2.054	− 2.528	− 0.846
Yang et al.[15]	0.3	0.599	0.058	0.49	0.897	− 3.053	− 3.366	− 1.048

Table 1. Benchmark evaluation. Average across all datasets.

Discussion

This work introduces a novel computational model for Hybrid Search tasks in natural scenes, extending the capabilities of existing models focused solely on Visual Search. By incorporating mechanisms like limited memory capacity and an adapted visibility function, our model effectively captures aspects of human behaviour observed in hybrid search experiments. These adaptations allowed the model to perform better than previous approaches within the visual search benchmark[23]. This work conveys the release of a new Hybrid Search Eye Movements (HSEM) dataset, with 4442 scanpaths from 44 participants performing a hybrid visual and memory search where targets and distractors were superimposed to natural scenes.

Our model's performance demonstrates the importance of peripheral visibility and visual working memory (VWM) in hybrid search tasks. Previous studies have noted that human searchers utilize peripheral information even as the foveal region remains the primary source of high-resolution detail[43]. In particular, the periphery has an important role in the guidance of eye movements by incorporating information of the context[6,44,45]. For instance,[44] showed that scene context acts as a framework to guide our vision, particularly in peripheral vision. Nuthmann and Malcom[6] highlighted the importance of colour in the periphery to help localize targets. By enhancing our visibility map to reflect peripheral information, we achieved significant improvements in alignment with human-like saccadic behaviour and fixation patterns, particularly in scenarios with higher memory set sizes. Geisler and collaborators also stressed the importance of peripheral vision in search guidance[21,46,47]. In their paradigms, they made a strong effort to estimate the visibility map for each participant. They used the first 16 sessions only to measure the dependence of the visibility with the eccentricity in each direction. Those measurements were then used to estimate the parameters of the IBS's visibility map for each participant[21]. This approach was also used in the recent paper by[17] for the same kind of artificial stimuli. While this approach could potentially incorporate individual differences among participants, their generalization capability might be more limited. In contrast, our approach avoids a potential leak of information about the viewing patterns from the participants to the model because its parameters are not fitted to each participant[16,48].

Visual Working Memory (VWM) is a core mechanism in our visual system. We evaluated two approaches to limit the VWM capacity: adding a fixed amount of fixations or an exponential decay to the model's memory. Zhou and Yu[17] implemented a model for visual search with limited VWM in artificial displays. Interestingly, they found the optimal value to be around 8 fixations, coinciding precisely with our modeling results. Moreover, this goes in line with behavioural results of Kaunitz and collaborators[49], who found that participants remembered up to seven fixated non-target faces with more than 70% accuracy in an incidental memories paradigm. Our second approach, incorporating an exponential decay to the model's memory, improved the model's performance even further, suggesting that the memory of recently explored locations exponentially decays over fixations. This resembles other memory systems, such as iconic memory, in which the exponential decay seems to be the best model[42]. Importantly, comparisons with other systems are relevant as the hybrid search task we studied differs from a standard working memory paradigm, in which the stimulus disappears after a short presentation. In hybrid search, the stimuli and their context remain on the screen and it is the active vision process that generates a sequence of foveated images.

In this work, we generated the similarity map with a ResNeXt-101 32x8d network[39] instead of the VGG-16 network[40] used in previous works[19,23]. This novel neural network shows better alignment with visual cortex activity (BrainScore (https://www.brain-score.org/vision/)), and it can potentially extract more informative features from the images, which might have been missed by earlier architectures. While this network yielded an increase in some of the metrics, their effect sizes were low.

A critical aspect of hybrid search relies on the decision of how to prioritize possible targets in each fixation. This prioritization could be influenced by both the current state of the search and biases generated during the memorization phase. Previous studies have modeled this process as competing drift-diffusion models, in which it is possible to introduce modulations in both the evidence accumulating rate and the initial bias[50]. However, these models did not take into account the sequential nature of the search, where each fixation represents a decision point in which accumulated evidence may be re-evaluated. Here, we modeled how this decision is taken at each step. The logarithmic dependence of reaction time on memory set size suggests that memory search operates in parallel[50]. Building on this, we assumed that the posterior distribution for each target is processed in parallel, followed by a selection process that determines which map will guide the following saccade. We

found that a selection process based on the potential information gained after the next saccade outperformed strategies guided either randomly or by a model that is only aware of the real target. This result is in line with evidence-based integration processes proposed by drift-diffusion models of Hybrid Search[27,50]. Importantly, the model incorporates the discrete nature of cognitive processes[51,52], where saccades are considered the natural boundaries of attentional episodes[53,54].

Although the present model outperforms previous models in the visual search benchmark and also extends the scope to complex search tasks like hybrid search, it still has a number of limitations. Firstly, the prior does not explicitly incorporate contextual information. Bujia et al.[16] recently explored different prior distributions, including a simple center bias, a saliency measure taken from the Itti & Koch model[5], based only on low-level features, and those taken from deep learning models such as DeepGaze II, which is a convolutional neural network that uses both low-level and high-level features. Since DeepGaze II was trained on an exploration task, it relies on the stimulus (and not the target object) to generate a saliency map. In real-life search, the scanpath differs if you are looking for a bird (you would search in the sky) or for a rabbit (you would not search in the sky). This contextual information is important both in visual search and hybrid search[11,27,55], and it could be approached using novel generative AI models[56,57]. These novel models are proficient at understanding semantic/syntactic relationships between objects in images and that information could be exploited to inform the prior. Importantly, this process will require overcoming technical challenges, such as segmenting objects of diverse sizes as well as generating prompts that could be informative of the relative positions between objects across different scenarios.

Secondly, the current visibility map, based on 2D Gaussian, is a rough approximation to the eccentricity in humans[16] who participated in a visual search experiment[21]. However, there are other studies specifically focused on how the visibility decays with the periphery[58]. The modification presented in this work allows decoupling the inhibition of return from the visibility map, easily allowing the integration of a more suitable model for both the fovea and periphery in the future.

Thirdly, we used a ResNeXt-101 32x8d model to build the target similarity map. This approach was embedding-based –it did not rely on the actual pixels of the image–. In[23], we also evaluated another embedding-based approach[19] and some pixel-based approaches to build the target similarity map. As these approaches relied on the pixels themselves to compute similarity, they could only work when the target was an accurate representation of the image, but they are not robust against physical manipulations, such as changing the object size, rotations or lighting. In contrast, embedding-based approaches are robust against such manipulations. Going forward, we intend to explore new ways of generating those embeddings using neural networks.

Fourthly, the IBS model is an optimal searcher model whereas humans are known to be suboptimal searchers[17], and the ELM could not be the best approximation of the Ideal Bayesian Searcher in natural scenes because the targets are not necessarily non-overlapping and the visual evidences are not necessarily independent across possible target locations[22]. A framework that makes more realistic assumptions is still needed.

Finally, regarding the biases generated during the memorization phase, one limitation of our approach is that our models assumed all targets were remembered equally. This simplification prevented us from incorporating the increased difficulty that humans experience as memory set size grows, where some targets may be encoded with different strengths during the memorization phase. For instance, primacy and recency effects, typically observed during the encoding of item sequences[59], were not evident in the current paradigm, likely due to the low memory set sizes used. Increasing set value sizes could show some of these effects[11,27] which could be readily incorporated into future models. Moreover, in this study all potential targets were similar and belonged to the same category, preventing an evaluation of biases arising from differences in target attributes, such as saliency or target-context congruence. New experiments could specifically evaluate how potential targets' attributes interact during memory encoding, which could provide the basis for extending our model to a more realistic memory encoding framework.

This study unveiled the first computational model for eye movements during hybrid search in natural scenes, providing a robust framework with key components relevant to explaining human behaviour in these tasks. Additionally, we presented the Hybrid Search Eye Movements (HSEM) dataset, which includes a large variety of images under different conditions, and the corresponding human scanpaths. Using the ViSioNS Benchmark, we validated the enhancements done to our model, demonstrating that our new version outperformed existing models[15,19,23,48]. Overall, this work illustrates how cognitively inspired model adjustments can lead to a more accurate representation of human search behaviour across complex tasks.

Data availability

The HSEM dataset used for the main analysis is available in https://osf.io/e97ws/. The external datasets used for the Model Validation section are available through the ViSioNS Benchmark (https://github.com/NeuroLIAA/visions repository).

Code availability

The code for running the model and the analysis is publicly available at https://github.com/NeuroLIAA/ruarte-hs-2024.

Received: 20 January 2025; Accepted: 28 April 2025
Published online: 12 May 2025

References

1. Treisman, A. M. & Gelade, G. A feature-integration theory of attention. *Cogn. Psychol.* **12**, 97–136 (1980).

2. Eckstein, M. P. Visual search: A retrospective. *J. Vis.* **11**, 14–14 (2011).
3. Yarbus, A. L. *Eye Movements and Vision* 171–211 (Springer, 1967).
4. Schütt, H. H., Rothkegel, L. O. M., Trukenbrod, H. A., Engbert, R. & Wichmann, F. A. Disentangling bottom-up versus top-down and low-level versus high-level influences on eye movements over time. *J. Vis.* **19**, 1. https://doi.org/10.1167/19.3.1 (2019).
5. Itti, L. & Koch, C. Computational modelling of visual attention. *Nat. Rev. Neurosci.* **2**, 194–203 (2001).
6. Nuthmann, A. & Malcolm, G. L. Eye guidance during real-world scene search: The role color plays in central and peripheral vision. *J. Vis.* **16**, 3. https://doi.org/10.1167/16.2.3 (2016).
7. Peacock, C. E., Singh, P., Hayes, T. R., Rehrig, G. & Henderson, J. M. Searching for meaning: Local scene semantics guide attention during natural visual search in scenes. *Q. J. Exp. Psychol.* **76**, 632–648 (2022).
8. Henderson, J. M. & Hayes, T. R. Meaning-based guidance of attention in scenes as revealed by meaning maps. *Nat. Hum. Behav.* **1**, 743–747 (2017).
9. Henderson, J. M. & Hayes, T. R. Meaning guides attention in real-world scene images: Evidence from eye movements and meaning maps. *J. Vis.* **18**(10), 1–18 (2018).
10. Wolfe, J. M., Alvarez, G. A., Rosenholtz, R., Kuzmova, Y. I. & Sherman, A. M. Visual search for arbitrary objects in real scenes. *Attent. Percept. Psychophys.* **73**, 1650–1671. https://doi.org/10.3758/s13414-011-0153-3 (2011).
11. Wolfe, J. M. Saved by a log: How do humans perform hybrid visual and memory search?. *Psychol. Sci.* **23**, 698–703. https://doi.org/10.1177/0956797612443968 (2012).
12. Schneider, W. & Shiffrin, R. M. Controlled and automatic human information processing: I. Detection, search, and attention. *Psychol. Rev.* **84**, 1–66 (1977).
13. Wolfe, J. M., Võ, M. L., Evans, K. K. & Greene, M. R. Visual search in scenes involves selective and nonselective pathways. *Trends Cogn. Sci.* **15**, 77–84 (2011).
14. Nastase, S. A., Goldstein, A. & Hasson, U. Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience. *NeuroImage* **222**, 117254 (2020).
15. Yang, Z. *et al.* Predicting goal-directed human attention using inverse reinforcement learning, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* 193–202 (2020).
16. Bujia, G., Sclar, M., Vita, S., Solovey, G. & Kamienkowski, J. E. Modeling human visual search in natural scenes: A combined Bayesian searcher and saliency map approach. *Front. Syst. Neurosci.* **16**, 882315 (2022).
17. Zhou, Y. & Yu, Y. Human visual search follows a suboptimal Bayesian strategy revealed by a spatiotemporal computational model and experiment. *Commun. Biol.* **4**, 1–16 (2021).
18. Rashidi, S. et al. An active foveated gaze prediction algorithm based on a Bayesian ideal observer. *Pattern Recogn.* **143**, 109694 (2023).
19. Zhang, M. et al. Finding any Waldo with zero-shot invariant and efficient visual search. *Nat. Commun.* **9**, 1–15 (2018).
20. Griffiths, T. L., Chater, N. & Tenenbaum, J. B. *Bayesian Models of Cognition: Reverse Engineering the Mind* (MIT Press, 2024).
21. Najemnik, J. & Geisler, W. S. Optimal eye movement strategies in visual search. *Nature* **434**, 387–391 (2005).
22. Najemnik, J. & Geisler, W. S. Simple summation rule for optimal fixation selection in visual search. *Vis. Res.* **49**, 1286–1294 (2009).
23. Travi, F., Ruarte, G., Bujia, G. & Kamienkowski, J. E. ViSioNS: Visual search in natural scenes benchmark. In *Advances in Neural Information Processing Systems* Vol. 35 (eds Koyejo, S. et al.) 11987–12000 (Curran Associates, Inc., 2022).
24. Bujia, G., Ruarte, G., Sclar, M., Solovey, G. & Kamienkowski, J. E. Uncertainty during visual search: Insights from a computational model and behavioral experiment in natural stimuli. *bioRxiv* 2025–01 (2025).
25. Lin, T.-Y. et al. Microsoft COCO: Common objects in context. In *Computer Vision - ECCV 2014* (eds Fleet, D. et al.) 740–755 (Springer, 2014).
26. Deng, J. *et al.* ImageNet: A large-scale hierarchical image database, in *2009 IEEE Conference on Computer Vision and Pattern Recognition* 248–255 (2009).
27. Barbosa, A., Ruarte, G., Ries, A. J., Kamienkowski, J. E. & Ison, M. J. Investigating the effects of context, visual working memory, and inhibitory control in hybrid visual search. *Front. Hum. Neurosci.* **18**, 1436564 (2024).
28. Peirce, J. et al. PsychoPy2: Experiments in behavior made easy. *Behav. Res. Methods* **51**, 195–203. https://doi.org/10.3758/s13428-018-01193-y (2019).
29. Chen, Y. et al. COCO-Search18 fixation dataset for predicting goal-directed attention control. *Sci. Rep.* **11**, 8776 (2021).
30. Dewhurst, R. et al. It depends on how you look at it: Scanpath comparison in multiple dimensions with MultiMatch, a vector-based approach. *Behav. Res. Methods* **44**, 1079–1100 (2012).
31. Kümmerer, M. & Bethge, M. *State-of-the-Art in Human Scanpath Prediction* arXiv:abs/2102.12239 (2021).
32. Bylinskii, Z., Judd, T., Oliva, A., Torralba, A. & Durand, F. What do different evaluation metrics tell us about saliency models?. *IEEE Trans. Pattern Anal. Mach. Intell.* **41**, 740–757 (2019).
33. Peters, R. J., Iyer, A., Itti, L. & Koch, C. Components of bottom-up gaze allocation in natural images. *Vis. Res.* **45**, 2397–2416 (2005).
34. Kümmerer, M., Wallis, T. S. A. & Bethge, M. Information-theoretic model comparison unifies saliency metrics. *Proc. Natl. Acad. Sci.* **112**, 16054–16059. https://doi.org/10.1073/pnas.1510393112 (2015).
35. Tatler, B. W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor Biases and image feature distributions. *J. Vis.* **7**, 4–4 (2007).
36. Chen, Y. *et al.* Characterizing target-absent human attention, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops* 5031–5040 (2022)
37. Borji, A. & Itti, L. CAT2000: A large scale fixation dataset for boosting saliency research, in *CVPR 2015 Workshop on "Future of Datasets"*. arXiv:1505.03581 (2015).
38. Kümmerer, M. & Bethge, M. DeepGaze II: A big step towards explaining all information in image-based saliency. *J. Vis.* **16**, 330–330 (2016).
39. Xie, S., Girshick, R., Dollar, P., Tu, Z. & He, K. Aggregated residual transformations for deep neural networks, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (2017).
40. Simonyan, K. & Zisserman, A. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. arXiv:1409.1556 [cs.CV] (2015).
41. Otero-Millan, J., Troncoso, X. G., Macknik, S. L., Serrano-Pedraza, I. & Martinez-Conde, S. Saccades and microsaccades during visual fixation, exploration, and search: Foundations for a common saccadic generator. *J. Vis.* **8**, 21–21. https://doi.org/10.1167/8.14.21 (2008).
42. Teeuwen, R. R. M., Wacongne, C., Schnabel, U. H., Self, M. W. & Roelfsema, P. R. A neuronal basis of iconic memory in macaque primary visual cortex. *Curr. Biol.* **31**, 5401–5414. https://doi.org/10.1016/j.cub.2021.09.052 (2021).
43. Benson, N. C., Kupers, E. R., Barbot, A., Carrasco, M. & Winawer, J. Cortical magnification in human visual cortex parallels task performance around the visual field. *Elife* **10**, e67685. https://doi.org/10.7554/eLife.67685 (2021).
44. Pereira, E. J. & Castelhano, M. S. Peripheral guidance in scenes: The interaction of scene context and object content. *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 2056–2072. https://doi.org/10.1037/a0037524 (2014).
45. Deza, A. & Eckstein, M. Can peripheral representations improve clutter metrics on complex scenes? *Adv. Neural Inf. Process. Syst.* **29** (2016).
46. Geisler, W. S., Perry, J. S. & Najemnik, J. Visual search: The role of peripheral information measured using gaze-contingent displays. *J. Vis.* **6**, 1. https://doi.org/10.1167/6.9.1 (2006).

47. Bradley, C., Abrams, J. & Geisler, W. S. Retina-V1 model of detectability across the visual field. *J. Vis.* **14**, 22–22. https://doi.org/10.1167/14.12.22 (2014).
48. Gupta, S. K., Zhang, M., Wu, C. C., Wolfe, J. & Kreiman, G. Visual search asymmetry: Deep nets and humans share similar inherent biases. *Adv. Neural Inf. Process. Syst.* **34**, 6946–6959 (2021).
49. Kaunitz, L. N., Rowe, E. G. & Tsuchiya, N. Large capacity of conscious access for incidental memories in natural scenes. *Psychol. Sci.* **27**, 1266–1277. https://doi.org/10.1177/0956797616658869 (2016).
50. Drew, T. & Wolfe, J. M. Hybrid search in the temporal domain: Evidence for rapid, serial logarithmic search through memory. *Attent. Percept. Psychophys.* **76**, 296–303. https://doi.org/10.3758/s13414-013-0606-y (2014).
51. Wyble, B., Potter, M. C., Bowman, H. & Nieuwenstein, M. Attentional episodes in visual perception. *J. Exp. Psychol. Gen.* **140**, 488 (2011).
52. VanRullen, R. Perceptual cycles. *Trends Cogn. Sci.* **20**, 723–735 (2016).
53. Kamienkowski, J. E., Navajas, J. & Sigman, M. Eye movements blink the attentional blink. *J. Exp. Psychol. Hum. Percept. Perform.* **38**, 555 (2012).
54. Navajas, J., Sigman, M. & Kamienkowski, J. E. Dynamics of visibility, confidence, and choice during eye movements. *J. Exp. Psychol. Hum. Percept. Perform.* **40**, 1213 (2014).
55. Rosman, B. & Ramamoorthy, S. Learning spatial relationships between objects. *I. J. Robot. Res.* **30**, 1328–1342 (2011).
56. DeepSeek-AI *et al. DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning* (2025). arXiv:2501.12948 [cs.CL].
57. OpenAI *et al. GPT-4 Technical Report*. arXiv:2303.08774 [cs.CL] (2024).
58. Anstis, S. Picturing peripheral acuity. *Perception* **27**, 817–825 (1998).
59. Glanzer, M. & Cunitz, A. R. Two storage mechanisms in free recall. *J. Verb. Learn. Verb. Behav.* **5**, 351–360 (1966).

Acknowledgements
J.E.K received research grants from CONICET (PIP 11220150100787CO) and ANPCyT (PICT 2018-2699). J.E.K. and M.J.I. received an award from ARL (Cooperative Agreement Numbers W911NF1920240 and W911NF2120237). We thank Alessandra Barbosa for their collaboration with the data acquisition, and Joaquin Gonzalez and Anthony Ries for insightful discussions.

Author contributions
GR implemented the computational models. GB developed the ELM version of the IBS model and contributed to the development of the experimental stimuli. DC preprocessed the eye-movement data and contributed to the development of the experimental stimuli. MJI and JEK conceptualized the study, designed the experiment, and supervised the project. GR, MJI and JEK analyzed the data and wrote the manuscript.

Declarations

Competing interests
The authors declare no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-025-00272-3.

Correspondence and requests for materials should be addressed to G.R.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.