Behavior Research Methods (2023) 55:583–599
https://doi.org/10.3758/s13428-022-01816-5

The Oddity Detection in Diverse Scenes (ODDS) database: Validated real-world scenes for studying anomaly detection

Michael C. Hout[1,2] · Megan H. Papesh[1] · Saleem Masadeh[1] · Hailey Sandin[1] · Stephen C. Walenchok[3] · Phillip Post[1] · Jessica Madrid[1] · Bryan White[1] · Juan D. Guevara Pinto[4] · Julian Welsh[1] · Dre Goode[1] · Rebecca Skulsky[1] · Mariana Cazares Rodriguez[1]

Accepted: 18 February 2022 / Published online: 30 March 2022

Abstract
Many applied screening tasks (e.g., medical image or baggage screening) involve challenging searches for which standard laboratory search is rarely equivalent. For example, whereas laboratory search frequently requires observers to look for precisely defined targets among isolated, non-overlapping images randomly arrayed on clean backgrounds, medical images present unspecified targets in noisy, yet spatially regular scenes. Those unspecified targets are typically oddities, elements that do not belong. To develop a closer laboratory analogue to this, we created a database of scenes containing subtle, ill-specified "oddity" targets. These scenes have similar perceptual densities and spatial regularities to those found in expert search tasks, and each includes 16 variants of the unedited scene wherein an oddity (a subtle deformation of the scene) is hidden. In Experiment 1, eight volunteers searched thousands of scene variants for an oddity. Regardless of their search accuracy, they were then shown the highlighted anomaly and rated its subtlety. Subtlety ratings reliably predicted search performance (accuracy and response times) and did so better than image statistics. In Experiment 2, we conducted a conceptual replication in which a larger group of naïve searchers scanned subsets of the scene variants. Prior subtlety ratings reliably predicted search outcomes. Whereas medical image targets are difficult for naïve searchers to detect, our database contains thousands of interior and exterior scenes that vary in difficulty, but are nevertheless searchable by novices. In this way, the stimuli will be useful for studying visual search as it typically occurs in expert domains: Ill-specified search for anomalies in noisy displays.

Keywords Database · Visual search · Medical image perception · Anomaly detection

Although much has been learned about visual search from laboratory paradigms, laboratory search tasks rarely capture the challenge and complexity of consequential real-life searches (e.g., medical image screening, security checkpoints, search and rescue). For example, medical image screening involves search for potentially small and subtle anomalies in multidimensional CT. These searches are complicated by variations in grayscale, non-uniform distribution of harmless anomalies (e.g., fatty deposits), and unpredictable numbers and locations of potential "targets." These searches do, however, often present spatially regular scenes in which the placement of major landmarks (e.g., organs) is fairly typical from one image to the next. This combination of image complexity, target imprecision, and structural regularity rarely exist in laboratory search paradigms, making them a poor analogue for this type of search. In this manuscript, we describe the rationale for, and validation of, an image database meant to better capture the visual elements that characterize search environments in consequential real-life domains. Although this stimulus set cannot bridge every gap between laboratory and real-life search, it is a step toward developing laboratory paradigms capable of revealing insights into the

✉ Michael C. Hout
mhout@nmsu.edu

[1] Department of Psychology, New Mexico State University, P.O. Box 30001 / MSC 3452, Las Cruces, NM 88003, USA

[2] National Science Foundation, Alexandria, VA, USA

[3] Exponent, Tempe, AZ, USA

[4] Rollins College, Winter Park, FL, USA

dynamics of ill-specified anomaly search, such as medical image or baggage screening.

Typical laboratory visual search tasks differ from applied screening contexts in several important ways, including the extent of observers' training and expertise. For example, whereas radiologists spend a decade or more on specialized medical and diagnostic image training, laboratory participants are typically drawn from undergraduate populations. Similarly, the estimated caseload for a standard, 8-h workday for a radiologist interpreting CT and MRI images requires them to interpret one image every 3–4 s during that shift (McDonald et al., 2015) and the Transportation Security Agency screens over 1.4 million checked bags daily (and likely far more carry-on bags; tsa.gov). Laboratory participants, by contrast, rarely spend more than 1 h engaging in visual search tasks. Our focus is not on participant-level differences, but on task-relevant differences that can potentially be addressed, allowing researchers to discover insights into the basic mechanisms of anomaly search. In what follows, we compare standard laboratory visual search paradigms to applied screening contexts. We emphasize comparisons to medical image screening because it typically involves search for evidence of anomalies, rather than specific target categories (as is done in baggage screening). Note that our descriptions of both tasks are necessarily over-generalized.

Although the visual search literature is replete with diverse and theoretically meaningful techniques, stimuli, and manipulations, we consider a standard visual search paradigm in which observers scan through displays of objects searching for one or more specific targets. Such paradigms will typically cue observers with target images, names, or categories, and then present displays which may or may not contain a target. Frequently, the images in the display are randomly spatially dispersed and non-overlapping (although see Godwin et al., 2017), presented against white or neutral backgrounds (but see Wolfe et al., 2021). Observers' task may be to report target presence or absence, or to localize any targets in the display (e.g., via mouse click). Although some paradigms provide trial-by-trial or block-level feedback, it is typically not necessary to provide external feedback telling observers that they successfully spotted a target, as most paradigms adopt unambiguous targets (this feedback, of course, is more meaningful when observers fail to spot targets). Standard paradigms such as this have yielded important insights into functions like attentional guidance (Wolfe & Horowitz, 2017; Wolfe & Utochkin, 2019), resolving target–distractor relationships (e.g., Becker, 2010), and shifting quitting thresholds (e.g., Wolfe & Van Wert, 2010), among others. They do not, however, necessarily scale up to medical image screening.

Radiological screening presents a notoriously difficult cognitive and perceptual challenge with greater consequences than laboratory search tasks. Radiologists often examine two-dimensional representations of three-dimensional tissue, wherein semi-transparent anatomical structures overlap and occlude one another (Krupinski, 2010). Unless examining follow-up scans for known diseases or conditions, screeners must identify ill-defined and/or subtle abnormalities that often depend on the anatomical structure under scrutiny (e.g., breast, lung, abdomen) and underlying pathology. Determining that a scan is clear requires an exhaustive search of the entire image, but merely looking at an abnormality is not sufficient for recognizing its importance. For instance, attempts to systematize eye movements have proven effective at training searchers to sweep their eyes systematically along anatomical structures (Auffermann et al., 2015, 2018; Kok et al., 2015), but they have resulted in only modest benefits to detection accuracy and there is some concern that experts eventually disregard such strategies (Waite et al., 2019). Further, when targets rarely appear, which is the case for most abnormalities in radiological screening (Bruno et al., 2015), observers become more likely to miss them (e.g., Evans et al., 2013), often despite directly viewing them (e.g., Godwin et al., 2015a, b; Hout, et al., 2015). Even when a radiologist detects an anomaly, there may be some degree of indecision or ambiguity, as reflected by imperfect reader agreement (70–80%; Kundel & Polansky, 2003).

The differences between laboratory search and applied screening tasks (e.g., medical image analysis) are not entirely insurmountable, but may require modifications to the standard laboratory approach if researchers' goals include scaling up to applied contexts. Of the many potential modifications, we propose starting with a relatively simple one: Stimuli. In many laboratory search tasks, observers search through well-separated, albeit structurally random, arrays of objects against white or neutral backgrounds, seeking one or more specific targets. In medical image screening, by contrast, observers search through structurally regular but visually noisy displays, seeking one or more unspecified targets (e.g., nodules, duct thickening, lesions, anatomical deformations). In laboratory visual search, difficulty is often manipulated via changes to target–distractor similarity or the precision of search cues (Hout & Goldinger, 2015; Hout et al., 2017). In medical image screening, difficulty is typically defined by how subtle (i.e., more or less perceptible) the anomaly is. Finally, in laboratory search, observers experience little ambiguity when they identify targets. By contrast, in medical image screening, two doctors may disagree about whether evidence represents underlying pathology.

We suggest that creating a stimulus set that captures structural regularity, visual noise, and target ambiguity may be a first step toward developing laboratory paradigms that can reveal insights into the mechanisms of anomaly detection that characterize radiological scanning. To that end, we modified scenes of libraries/books, interior rooms, and forests to include subtle "ripple" deformations, similar to those used in research on camouflaged targets (Hess et al., 2016).

Although ripple deformations embedded within scenes are not a perfect approximation for anomalies located in medical images, we suggest that they are more appropriate than standard template-defined targets in many standard search tasks. Across two experiments, volunteers searched for and rated the subtlety of these targets, and objective salience measures were computed to compare against observers' ratings (Experiment 1); subtlety ratings were also used to predict search performance in a group of naïve observers (Experiment 2). Salience metrics were comparatively poor predictors of searcher performance, but human subtlety ratings predicted performance well in both experiments, confirming that the database contains anomalies ranging from subtle to obvious, which can be used in future research to test hypotheses about ill-specified anomaly search.

The Oddity Detection in Diverse Scenes (ODDS) Database

The ODDS Database was developed to assist researchers in creating laboratory tasks that can more closely approximate medical image perception using novice observers. Although studying radiologists would be the ideal approach, and indeed much has been learned from such studies (e.g., Aizenman et al., 2017; Drew et al., 2013; Williams & Drew, 2019; Williams et al., 2021), many researchers lack access to such a special population, but nevertheless may wish to test hypotheses about complex anomaly detection. Existing efforts to study laboratory analogues of radiological screening have employed a creative mix of standard visual search (e.g., rotated Ts and Ls) embedded among radiological images. For instance, Adamo et al. (2018) used tomosynthesis, a technique that creates three-dimensional representations of body parts using multiple two-dimensional scans of the area. Although their results nicely captured performance differences between 2 and 3D displays across professionals and novices, the targets were nevertheless well-specified (a T among Ls). Like Adamo et al. (2018), our goal was to create stimuli that retain key characteristics of medical image perception but are simple enough that novice searchers can be tested.

The ODDS Database was built from 284 unique scenes that can be broadly categorized as forests (144), indoor scenes (96), and libraries/books (44). Scenes were obtained from Unsplash.com, an online repository of freely usable, high-resolution images.[1] Criteria for selection included: (1) landscape orientation, (2) greater than or equal to 1920×1080 (width and height, respectively) resolution, and (3) the majority of the image must contain meaningful content (e.g., images with large open skies or blank walls were avoided because deforming flat surfaces results in targets that are nearly impossible to discriminate from the background). The original images were manipulated to contain an "oddity" target in the form of a ripple deformation in the scene (see Hess et al., 2016, for a similar image manipulation method). Each scene was manipulated multiple (independent) times, creating 16 variants of the original, each with a target located in a different spatial position (for a total of 4576 edited images). Unedited images can thus be used for "target-absent" visual search trials and edited variants can be used to vary the location of the target and the difficulty of the task (see Fig. 1 for examples from each category across a range of subtlety scores from Experiment 1). These ripple deformations may prove problematic for observers with tonic visual distortions, so the interpretations and applications of these stimuli should be considered only for observers with typical vision.[2]

Image manipulation

To create each variant of the base images, we developed a deformation process using the Python v3.6 programming language and the *OpenCV* library (an open-source library with different functions for image processing). The shape of the deformation was circular, with a 30-pixel-long radius. The formula for computing the ripple was: $ripple = radius * sigma$, where *radius* is the radius of the image section to be manipulated, and *sigma* represents how much the image content would be modified. To compute sigma, sine and cosine functions were applied to the row and column pixel indices, which we refer to as the X and Y values of the pixels, respectively. Pixels were thus "shifted" but not removed from the image or otherwise modified. To retain the pixel within a range of 30 pixels post-shifting, we kept the sigma values within the range of 0 and 1. Ripple values for each pixel were then added to the pixel's original X/Y location to create a new coordinate location for that visual information, thus "shifting" it to a new location. Collectively, this relocation process is what created the intended rippling of scene regions, as shown in Fig. 1.

Not all images were the same resolution. To standardize image manipulation, we started by creating an imaginary "box" centered on the image with a fixed dimensionality of 1920×1080; see Fig. 2, panel A. Then, inside the box we created a virtual 4×4 grid, thereby creating 16 equally

[1] The longform license from Unsplash.com states: *"Unsplash grants you an irrevocable, nonexclusive, worldwide copyright license to download, copy, modify, distribute, perform, and use photos from Unsplash for free, including for commercial purposes, without permission from or attributing the photographer or Unsplash. This license does not include the right to compile photos from Unsplash to replicate a similar or competing service."*.

[2] We are grateful to Benjamin Wolfe for identifying this important restriction.

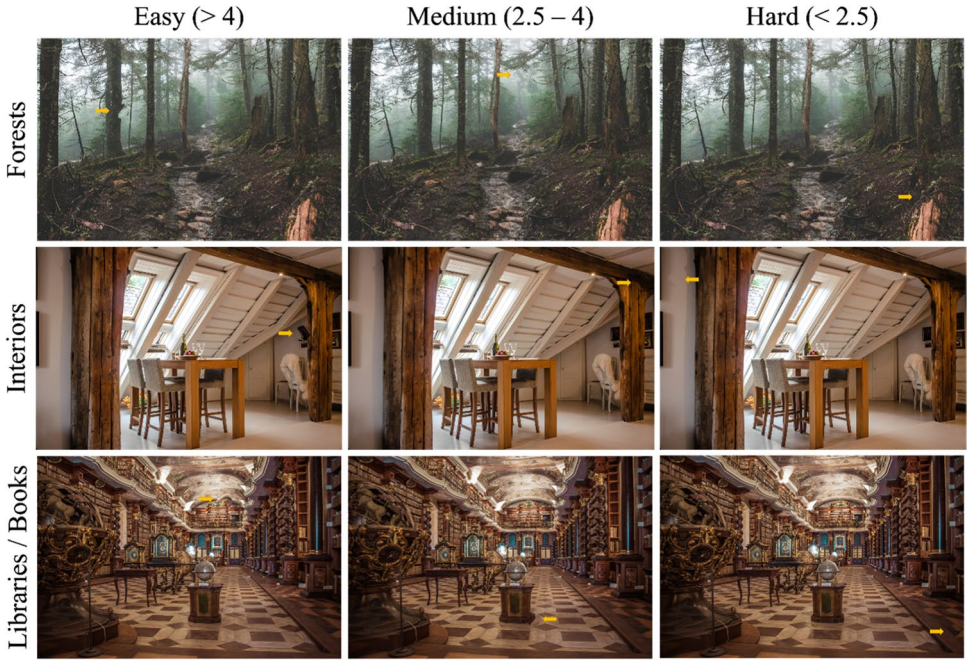

Fig. 1 Example stimuli from each category (*rows*) and across a range of subtlety scores (*columns*) obtained from Experiment 1. *Yellow arrows* point to the target. Note that easy, medium, and hard categorizations are arbitrary and the scale is biased in favor of easier to locate targets because very subtle targets are extremely difficult to appreciate when the images are a smaller size

sized "cells" (480 × 270 pixels) within which to locate the target for each of the 16 scene variants (see Fig. 2, panel B). Lastly, we created 16 different variants of the unedited image by creating a ripple within each of the cells one at a time. Target locations were randomly jittered within the cells to prevent all scene variants from forming a regular "grid" that volunteers could learn to systematically scan (see Fig. 2, panel C). The jittering process was performed independently for each scene and variant. The metadata for all scenes (shared on the OSF site: https://osf.io/b85e2/?view_only=3762e6e18d494e45bc0e387dfa8c62e2) contains the X/Y coordinates of the center point of each target for each scene variant (among other information, such as the original size of the image, subtlety ratings, search performance measures, etc.).

Experiment 1

During image manipulation, target locations were quasi-randomly determined and thus varied in their subtlety (i.e., how discriminable the target is from its background), depending on the local scene content and features present at target locations. For instance, target oddities that appeared in dense clusters of forest leaves or on large, homogeneous surfaces in indoor scenes tend to be difficult to discriminate from the background because they do not "disturb" the background content in a large enough or systematic manner to allow them to easily appear anomalous. By contrast, targets that occur near straight edges (e.g., a tree trunk, the edge of a flat surface like a countertop) are less subtle because the continuity of the background is disrupted and the target therefore more readily appears out of place.

Because of the way deformations were introduced, it was not possible to determine a priori how subtle the targets would appear in each image. The purpose of Experiment 1 was to provide subjective and objective ratings of target subtlety for each scene variant, and to validate these ratings by using them to predict visual search performance. Subjective ratings came from eight volunteers who searched for the oddity targets and, regardless of their search success, then assessed the target's subtlety in each scene variant when it was highlighted for them. Objective, image-based ratings were derived from several computed salience measures. Rather than examining an exhaustive set of salience quantifications to explore which performs optimally (for recent reviews, see Krasovskaya & MacInnes, 2019; Veale et al., 2017), we computed several common salience measures for the target region in each scene (see *Quantitative Saliency Metrics* below). Both the image-based salience measures and human ratings were then used to predict searcher performance.[3] To preface the outcome of Experiment 1, human ratings were the stronger, more reliable predictor of search outcomes.

[3] We are grateful to an anonymous reviewer for this suggestion.

Fig. 2 Demonstration of the process for selecting target locations for each scene variant (see text for details). Note that images are not drawn to scale. Panel A shows selection of the central "box," Panel B shows definition of equally sized "cells," and Panel C illustrates random jittering of target locations

Method

Volunteer participants

Eight volunteers participated in Experiment 1. Seven volunteers (HS, JM, BW, JW, DG, RS, and MCR) were members of the *Vision Sciences and Memory* laboratory at New Mexico State University (directed by the first and second authors) and one (JGP) was affiliated with Rollins College. All volunteers had normal (or corrected-to-normal) vision and self-reported typical color vision.

Design

Volunteers completed 4554 trials over the course of multiple experimental sessions. Trials were presented in random order across categories, scenes, and scene variants. The 4554 trials constituted the full set of scene variants (284 scenes * 16 variants = 4,544) plus ten additional trials. These trials were re-runs of the first ten trials volunteers completed, so that the first ten trials on the task could be treated as practice. All trial sequences were pre-randomized for each volunteer prior to the start of the first experimental session. Volunteers completed as many trials as they wished in each session and terminated the session when they felt fatigued or had to be otherwise engaged. Each new session picked up where the previous session left off, and volunteers repeated this process until all trials had been completed.

Apparatus

Because the project occurred during the COVID-19 pandemic, all data collection occurred in the volunteers' homes. This meant that hardware standardization was not possible across locations. Nevertheless, all volunteers completed the task using E-Prime vs3.0 software (Psychology Software Tools, Inc.; Pittsburgh, PA) on a PC computer whose screen resolution was set during the experiment to 1920 × 1080. Volunteers sat at a comfortable viewing distance from their screens and were instructed to minimize distractions during the experiment.

Procedure

Volunteers completed two tasks for every scene. First, they attempted to locate the target, and second, they were shown the location of the target (in case it was not found) and were asked to rate how subtle it was. Each trial progressed as shown in Fig. 3. At the start of each trial, volunteers were reminded what their task was and the order of events that would occur. To begin, they clicked the right mouse button, which replaced the instructions with the search scene. All images were centered on the display such that the 1920 × 1080 "box" used to establish potential target locations during image manipulation (see Fig. 2) subsumed the entire screen. Volunteers were given up to 5 s to search each scene. When the volunteer located the target, they clicked on it with the left mouse button. After the volunteer registered a response (or 5 s had elapsed), a yellow box was drawn around the target to highlight its location. The yellow box remained on screen until the volunteer clicked the right mouse button, at which point it was removed so that volunteers could view the target as it naturally appeared within the scene. During this time, volunteers evaluated how subtle the target was (i.e., how discriminable the target was from its background) in preparation for the rating task. Upon clicking the left mouse button, volunteers were shown rating instructions (see below) and provided their response via keyboard. The rating instructions were designed to approximate those used by Shiraishi et al. (2000), who aggregated a database of chest radiographs whose nodules had been rated for subtlety by medical professionals. An option for "invisible" was also included in case the image manipulation program resulted in an imperceptible target.

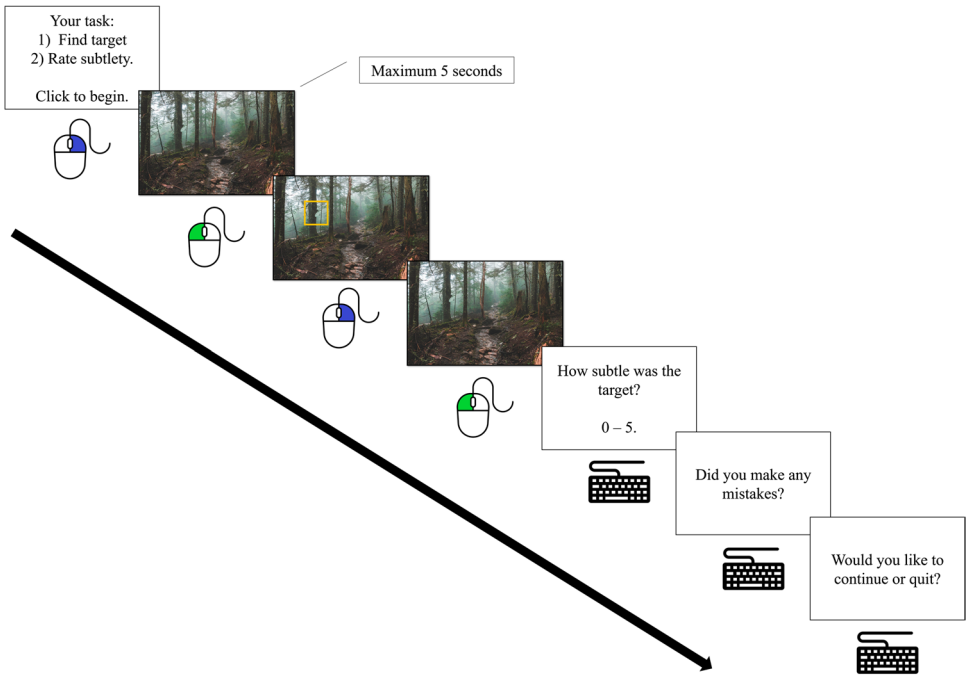

Fig. 3 Sample trial progression in Experiment 1. All displays remained on screen until the volunteer responded, with the exception of the search display which terminated after response or 5 s had elapsed. Note that instruction text is shortened for demonstrative purposes

"Please indicate how discriminable you thought the target was. Remember that this does NOT have to coincide with your subjective impression of your search performance. Sometimes you'll miss targets that in retrospect are quite obvious and sometimes you'll find ones that are quite subtle. That is why we are collecting subjective ratings and search behavior.
0 – Invisible.
1 – Extremely subtle. The target is very indistinct, and extremely difficult to detect.
2 – Very subtle. The target is very difficult to detect.
3 – Subtle. Detection is difficult.
4 – Relatively obvious. Detection is relatively easy.
5 – Obvious. Detection is very easy."

After volunteers provided their rating, they were asked if they made any mistakes that should be logged in the data file (e.g., accidentally clicking the mouse button when the target had not been found), which were registered using the keyboard. They were then asked if they would like to continue searching/rating or if they were ready to quit the session.

Quantitative saliency metrics

To quantify saliency, we utilized multiple methods built into the OpenCV computer vision software library in Python (https://www.opencv.org; see also https://www.pyimagesearch.com/2018/07/16/opencv-saliency-detection/). We first created coarse-grained and fine-grained saliency maps (Hou & Zhang, 2007; Montabone & Soto, 2010) for each image, utilizing the *StaticSaliencySpectralResidual* and *StaticSaliencyFineGrained* methods in OpenCV, each of which converts the image to grayscale, with each possible pixel value ranging from 0 to 255. We computed the average of these pixel values across each coarse-grained and fine-grained saliency map, then repeated this process for a 120×120-pixel region surrounding the target in each saliency map. This allowed us to compare the average saliency within the target region to the average saliency for the entire image. This process yielded a ratio of the mean target region salience to the mean of the entire image.[4]

We also created binary "thresholded" saliency maps (https://www.pyimagesearch.com/2018/07/16/opencv-saliency-detection/), which restricted possible pixel values to either 0 or 255, essentially creating a full-contrast saliency map for each stimulus image. Following the same steps taken for the coarse-grained and fine-grained saliency maps, we computed the average pixel value across each thresholded saliency map, then separately computed the average

[4] Target region salience, overall scene salience, and the ratio of the two are reported separately for each scene variant and each salience metric, all of which can be found in the scene metadata on our OSF site. All salience map images are available for download there as well.

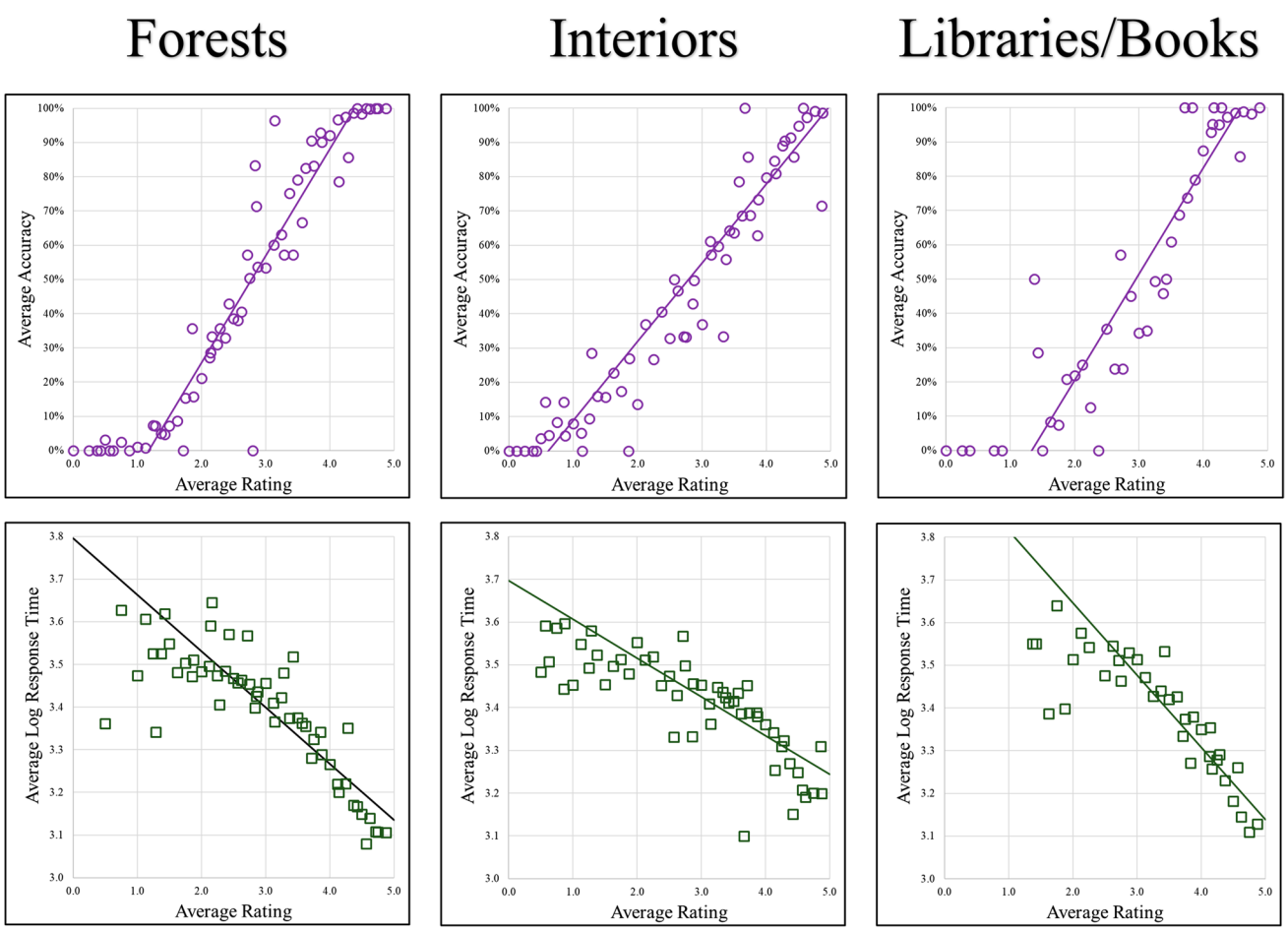

Fig. 4 Plots of linear regression analyses using average subtlety rating scores to predict search outcomes. *Top panels* show accuracy data, *bottom panels* show log RTs. The *left*, *middle*, and *right* columns present data separately for forests, interiors, and libraries/books, respectively. Individual symbols are mean performance for a given mean rating value, and solid lines plot best fitting regression equations

pixel value for each target region within these thresholded saliency maps.

Results

The primary goals of Experiment 1 were to acquire subtlety ratings for each of the scene variants, and to validate these ratings by using them to predict search performance. The secondary goal was to explore the degree to which objective measures of salience could be used to predict human ratings and searcher performance.

Only nine trials (0.02% of all trials) were discarded due to volunteers logging a user error. Additionally, occasional computer failures occurred, resulting in the volunteer repeating several trials that had already been searched and rated. When this happened, duplicate trials were discarded and only the first instance of a scene variant was analyzed. See Appendix Fig. A1 for a qualitative description of subtlety ratings and searcher performance distributions across scene categories.

Search performance predicted by subjective ratings

We conducted linear regressions using the average subtlety rating per scene variant to predict search accuracy and log-transformed RTs (for correct trials only) for that variant. Results for each scene category are presented separately. See Fig. 4 for plots of the results.

Forests Subtlety ratings reliably predicted search accuracy, $[F(1, 2302) = 5481, p < 0.001, R^2_{adj} = 0.70]$, such that scenes with higher scores (indicative of more obvious targets) resulted in higher search accuracy. Ratings also reliably predicted log-transformed search RTs, $[F(1, 1974) = 1799, p < 0.001, R^2_{adj} = 0.48]$, such that variants with more obvious targets were responded to more quickly.

Table 1 Results of linear regression analyses using salience ratios to predict observers' subjective ratings and search performance (accuracy, log-transformed RTs). F-ratios and variance accounted for are presented for each measure and for each of the three salience quantifications separately

			Coarse		Fine		Threshold	
		df	*F*	R^2 *adj*	*F*	R^2 *adj*	*F*	R^2 *adj*
Forests								
	Ratings	1, 2302	158	0.06	494	0.18	391	0.15
	Accuracy	1, 2302	118	0.05	356	0.13	293	0.11
	Log RTs	1, 1974	56	0.03	204	0.09	185	0.09
Interiors								
	Ratings	1, 1534	291	0.16	733	0.32	449	0.23
	Accuracy	1, 1534	72	0.04	222	0.13	133	0.08
	Log RTs	1, 1297	20	0.01	52	0.04	37	0.03
Libraries								
	Ratings	1, 702	119	0.14	184	0.21	116	0.14
	Accuracy	1, 702	62	0.08	76	0.10	49	0.06
	Log RTs	1, 664	60	0.08	50	0.07	39	0.05

All ps < *.001*

Interiors Subtlety ratings again positively predicted search accuracy, [$F(1, 1534) = 2658$, $p < 0.001$, $R^2_{adj} = 0.63$], and negatively predicted log-transformed RTs, [$F(1, 1297) = 563$, $p < 0.001$, $R^2_{adj} = 0.30$].

Libraries/books Again, subtlety ratings were a positive predictor of search accuracy, [$F(1, 702) = 1246$, $p < 0.001$, $R^2_{adj} = 0.64$], and a negative predictor of log-transformed RTs, [$F(1, 664) = 524$, $p < 0.001$, $R^2_{adj} = 0.44$].

Search performance predicted by objective ratings

To determine whether targets were objectively more salient than the rest of the scene (on average), we used computed target-to-scene salience ratios to predict subtlety ratings and searcher performance (accuracy, log-transformed RTs on correct trials) for each scene variant. In each case, linear regressions were statistically reliable in the predicted directions, but effect sizes varied considerably across measures and salience types. The results are summarized in Table 1.

Discussion

In Experiment 1, volunteers searched through all ODDS database scene variants and provided subtlety ratings that characterized how discriminable each target was from its background. These subjective ratings reliably predicted search accuracy and log-transformed RTs for each scene category – and accounted for more of the variance in searcher performance than algorithm-derived salience measures – suggesting that the subjective ratings are a valid way to characterize the subtlety of the targets and the difficulty of finding them.

Although we did not compute an exhaustive set of objective salience measures, those we selected were clearly capable of distinguishing salient target regions from the surrounding scene. These objective ratings also aligned with subjective human ratings to a meaningful degree, accounting for as little as 6% of the variance in human ratings in some instances but as much as 32% in others. More importantly, however, image-based salience measures were not nearly as predictive of searcher performance as human ratings; on average, objective salience measures only accounted for 7% of the variance in searcher performance, whereas subjective ratings captured an average of 53% of the variance (ranging from 30 to 70% across measures). In prior research, salience measures have often been used to predict individual oculomotor behaviors (see, for instance, Borji et al., 2013) rather than gross-level performance, as was done here.

It is perhaps unsurprising that subjective ratings better accounted for searcher performance than image-based metrics. It is difficult to determine a priori what features of the oddity targets caused human observers to perceive them as more or less subtle, and expecting a small sample of salience metrics to capture the full complexity of human vision (across such a wide variety of scenes) is probably unreasonable. Moreover, although our targets were not defined by any specific top-down target "template," it is clear that attentional guidance in scenes is driven by more than just bottom-up feature salience, and the shortcomings of purely salience-based models of attention have already been documented (e.g., Tatler & Vincent, 2009; Tatler, et al., 2011). If image databases are to be built (or expanded) while circumventing the need for human raters,

then future work will be necessary to determine ways to better quantify anomalous target salience.

The approach taken in Experiment 1 is not without limitations. For example, Experiment 1 used a small group of volunteer observers because the task of searching through and rating more than 4500 scene variants was quite onerous and time-consuming (and took place during a pandemic, which made in-lab data collection impossible). There are four potential drawbacks to this procedure: 1) Scene variants were presented in different random orders to each volunteer, but it is nevertheless possible that the experience of encountering a specific scene multiple times (irrespective of which variant was currently presented) provided some practice benefit for variants that were presented later in the experiment. 2) The sheer volume of trials may have produced burn-out (although volunteers were encouraged to quit the session if/when they felt fatigued). 3) Because the task required so many trials, we limited search time to 5 s, but it is possible that some missed targets would have been found if volunteers were given more time to look for them. And 4) the volunteers are co-authors on this project who were not naïve to the purpose of the study (although it seems unlikely that knowledge of the study's purpose would confer any benefit or bias).

Experiment 2

To address the potential shortcomings in Experiment 1 and further validate the subtlety ratings, we conceptually replicated Experiment 1 in a second validation experiment. In Experiment 2, naïve volunteers searched for up to 30 s through a small subset of the scenes (but did not provide subsequent subtlety ratings), and never encountered any scene more than once. We again found that the ratings provided by volunteers in Experiment 1 were a strong predictor of search accuracy and log-transformed search RTs. The larger sample of volunteers in Experiment 2 also allowed us to conduct inferential statistics on search outcomes to further characterize the difficulty of searching through each scene category.

Method

Participants

We conducted an a priori power analysis using G*Power vs3.1.9.7 (Erdfelder et al., 1996). To adopt a conservative approach, we used the weakest effect size observed in Experiment 1 (i.e., the regression using subtlety ratings to predict log-transformed RTs for interiors) to determine the required sample size for Experiment 2. We used the "linear multiple regression: Fixed model, R^2 deviation from zero" statistical test option with Cohen's F^2 of 0.43 (calculated from the observed squared multiple correlation value), and 1 predictor. This analysis indicated that we needed at least 33 participants to achieve the desired power of 95%.

Participants were recruited in one week increments until we met or exceeded the required sample size. Because of the ongoing pandemic, we recruited volunteers from friends, family, and members of our laboratories, but all volunteers were naïve to the purpose of the study and provided informed consent prior to participation. Fifty-three participants completed the task, but seven (13% of the total sample) were removed for data logging errors, leaving 46 participants for analyses. Nine, 12, 15, and ten participants completed the four possible experiment "packages," respectively (see *Design*, below, for more information); final samples were somewhat uneven across packages due to the unpredictable nature of the data logging problem we encountered.

Design

Experiment 2 also took place during the COVID-19 pandemic. Unlike in-lab settings wherein laboratory computers could be loaded with the entire database of stimuli – therefore allowing fully random sampling of scene stimuli across participants – we instead had to pre-select random subsamples of the database to present to participants in small enough "packages" that could be downloaded and executed on participants' home computers (see the *Apparatus* section below for more details). To accomplish this, we quasi-randomly selected one variant from each possible scene in the database with the only constraint being that across all sampled scene variants, target placement occurred equally often in each of the 16 possible "cells" (see the *Image Manipulation* section, above). This prevented participants from adopting any spatial bias from learned target locations. We then created four experiment packages that were sent to participants in counter-balanced order (determined by recruitment order). Across the four experiment packages, therefore, each of the scenes was presented only once, and target spatial locations were sampled equally; moreover, each package contained 25% of the total scenes possible from each of the stimulus categories. All participants completed 71 total trials, which consisted of 36, 24, and 11 trials from the forests, interiors, and libraries/books categories, respectively.[5]

[5] To verify that our pre-selected sample of stimuli matched the overall characteristics of the entire ODDS database, and to ensure that we did not inadvertently introduce any systematic difficulty biases, we plotted comparison histograms (see Appendix Fig. A2). These plots show the distribution of subtlety ratings for the full set of stimuli next to the pre-selected sub-samples (presented separately for each category of stimuli). As can be seen in the figure, the shape of the distributions for the sub-samples closely match the distributions in the full stimulus set.

Apparatus

Like Experiment 1, standardization of hardware was not possible because participants completed the experiment in their own homes. For this experiment, we used E-Prime Go vs1.0; this software converts E-Prime vs3.0 experiments into executable files so participants can download and execute the experiment on their own computer without having access to the full E-Prime suite. Resolution of participants' monitors was again set by the program to 1920 × 1080 at the time of experiment execution.

Procedure

Because participants in Experiment 2 were not co-authors, additional care was taken in providing instructions that maximized the quality of in-home data collection. Participants were asked to participate in a quiet and distraction-free environment, to make sure the brightness of their monitor was turned up sufficiently high, and to clean their screen in the event that they were using a touch-screen laptop that had any smudges or fingerprints on it that would potentially obscure the targets.

Unlike Experiment 1, participants performed visual search but were not asked to also rate the subtlety of the targets. They completed one block of search for each of the stimulus categories; blocks were 36, 24, and 11 trials long for the forests, interiors, and libraries/books categories, respectively. Before each block, participants were instructed which category they would be searching through for the next set of trials. Scene order was randomized within blocks and block order was counterbalanced across participants. Participants were also told that the first five trials were practice, allowing them to orient themselves to the task. When the 6th trial began, they were told that practice was over and that they should try their best moving forward. All subsequent trials also reminded participants that they were no longer in practice mode.

The trials proceeded in nearly identical fashion to Experiment 1. At the start of each trial, participants were reminded of the instructions and began the trial with a right mouse click. The instructions screen was then replaced by the search scene, which remained on screen until a left mouse click was registered or 30 s had elapsed. Then, a yellow box was drawn around the target location. We retained this target highlighting process so that, in the event that the target was not found or was misidentified, participants could learn about target appearance over time (which was especially important during practice trials). Participants dismissed the yellow box with a right mouse click, after which they moved on to the next trial.

On a small subset of trials (five of the 71, randomly distributed over the course of the experiment), a "catch" stimulus was presented instead of the usual target. This stimulus was a bright yellow laughing face "emoji" placed where the target deformation would have been located. Participants were instructed that every so often these catch trials would occur to make sure they were paying attention.

Results

The primary goal of Experiment 2 was to further validate the subtlety ratings from Experiment 1. We used them to predict search outcomes from naïve participants who had more time to search, who would be unaffected by practice effects accrued via searching the same scene multiple times, and who were not accustomed to evaluating search difficulty for each scene. No participants were removed from analyses for failing to respond within 30 s on catch trials. Practice trials were not analyzed. For log-transformed RTs, only trials with correctly identified targets were analyzed. Greenhouse–Geisser corrections were applied for any sphericity violations. Figure 5 contains plots of analyses on accuracy (top row) and log-transformed RTs (bottom row).

Forests As in Experiment 1, subtlety ratings reliably predicted search accuracy, [$F(1, 142) = 278$, $p < 0.001$, $R^2_{adj} = 0.66$], such that less subtle variants resulted in higher search accuracy. Search RTs were also reliably predicted from subtlety ratings, [$F(1, 128) = 210$, $p < 0.001$, $R^2_{adj} = 0.62$], with less subtle variants producing faster RTs.

Interiors Again, subtlety ratings positively predicted search accuracy, [$F(1, 94) = 77.3$, $p < 0.001$, $R^2_{adj} = 0.45$], and negatively predicted log-transformed RTs, [$F(1, 90) = 32.4$, $p < 0.001$, $R^2_{adj} = 0.26$].

Libraries/books Subtlety ratings positively predicted search accuracy, [$F(1, 42) = 4.82$, $p = 0.03$, $R^2_{adj} = 0.08$], and negatively predicted search RTs, [$F(1, 42) = 27.3$, $p < 0.001$, $R^2_{adj} = 0.38$].

Because we had a larger participants sample than in Experiment 1, we conducted inferential statistics to examine differences in performance across scene categories. These should be interpreted with some caution, as Experiment 2 presented only a subset of the total scene variants. These results, however, may be informative for researchers interested differences between categories. See Fig. 6 for plots of the results.

We examined search errors in a 3 (Category) × 2 (Error Type: misses, misidentifications) ANOVA. We observed a main effect of Category, [$F(1.95, 87.91) = 66.96$, $p < 0.001$, $\eta^2_p = 0.60$], such that library/book trials produced the fewest errors (5%), followed by interiors (13%),

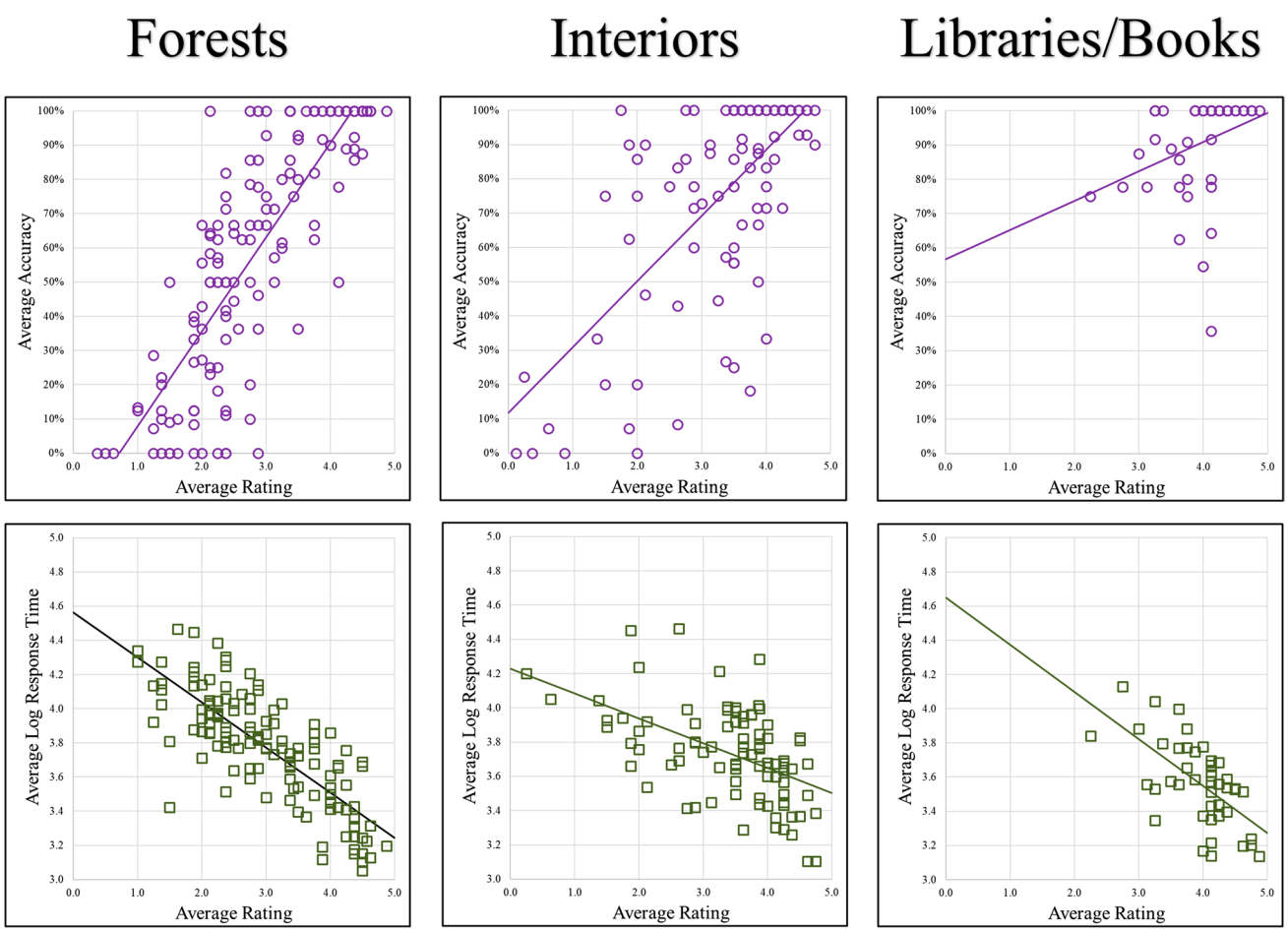

Fig. 5 Plots of linear regression analyses using average subtlety rating scores obtained in Experiment 1 to predict search outcomes from Experiment 2. *Top panels* show accuracy data, *bottom panels* show log RTs. The *left*, *middle*, and *right* columns present data separately for forests, interiors, and libraries/books, respectively. Individual *symbols* show performance for every scene variant presented

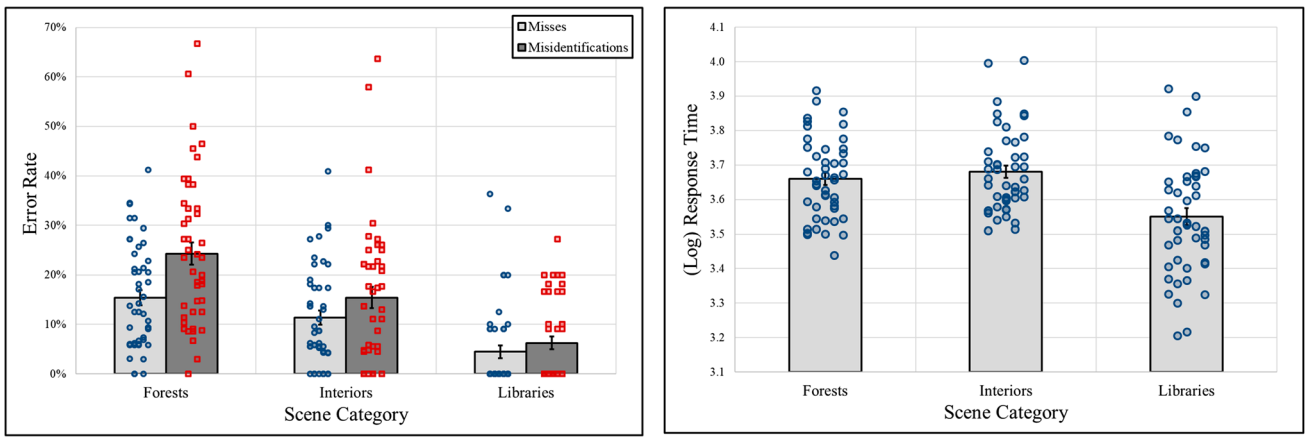

Fig. 6 Error rates (*left panel*) and log-transformed RTs (*right panel*) by scene categories from Experiment 2. In the left panel, miss rates (*light gray bars*) and misidentification rates (*dark gray bars*). *Error bars* represent one standard error of the mean. Symbols represent mean performance by individual participants

and forests (19%). We also observed a main effect of Error Type, [$F(1, 45) = 4.86$, $p = 0.03$, $\eta^2_p = 0.10$], such that misidentifications were more common (15%) than misses (10%). The interaction of both factors was marginally significant, [$F(1.6, 71.96) = 3.37$, $p = 0.05$, $\eta^2_p = 0.07$], revealing a larger difference in miss and misidentification rates during forest trials compared to interior and library/books trials.

Log-transformed RTs were examined in a one-way ANOVA using a single within-subjects factor of Category. A main effect of Category, [$F(1.8, 81.02) = 13.2$, $p < 0.001$, $\eta^2_p = 0.23$], revealed fastest search times for libraries/books (3.55), followed by forests (3.66) and interiors (3.68).

Discussion

Experiment 2 replicated the main findings from Experiment 1, further validating the subtlety ratings obtained in Experiment 1. Subtlety ratings predicted search outcomes in naïve participants who were given longer durations to search and never saw any scene more than once. That the subtlety scores reliably predicted search outcomes (accuracy and log-transformed RTs) for all categories despite variability in participants' hardware and local environments suggests that these ratings will be generally predictive in other search paradigms and contexts. Because of the larger sample, Experiment 2 also revealed differences across scenes, such that forests produced the highest error rates and libraries produced the fastest search times.

General discussion

The ODDS Database contains over 4000 scene images that have been edited to contain an "oddity" in the form of a subtle, ill-specified malformation of the scene content. Each scene variant was rated by eight volunteer observers for target subtlety. In our validation experiments, we found that these ratings were reliable predictors of the raters' prior search behavior (Experiment 1) and the search behavior of naïve participants (Experiment 2). Moreover, these ratings were better predictors of search performance than computer-derived salience measures. The ratings can thus be used by visual search researchers to manipulate search difficulty in new empirical work that employs these scenes.

The goal in developing the ODDS Database was to provide a means for researchers to better approximate applied anomaly search (e.g., medical image screening) using laboratory paradigms and novice (non-expert) participants. Although studying search behaviors in experts would be the most incisive way to investigate such an important search domain, many researchers lack the access or resources to secure such an expert population. Without access to experts, researchers can still discover interesting insights into the basic mechanisms of anomaly search that may reveal fundamental differences between standard laboratory search and ill-specified anomaly search. Establishing these basic mechanisms may ultimately "scale up" to medical image screening or provide researchers with proof-of-concept data to support grant applications to recruit experts.

Efforts to develop laboratory analogues of medical image screening are challenging, given the many differences between medical image search and laboratory visual search. For example, in addition to differences in observer ages, qualifications, and backgrounds, medical image screening requires observers to scan through visually noisy, but structurally regular, displays, searching for evidence of damage or disease. That evidence varies based on the anatomy under scrutiny and the potential underlying pathology. Laboratory search, by contrast, typically requires observers to search for known targets (or target categories) among randomly-arrayed, but non-overlapping objects against plain backgrounds.[6] Our approach was to address stimulus differences across medical image screening and laboratory paradigms. To that end, we created a database of complex scene images in which targets are anomalies that vary in subtlety and location. Target subtlety introduces some degree of ambiguity in target detection, which is meant to capture the fact that radiologists do not always agree about whether anomalies reflect underlying pathology (Kundel & Polansky, 2003).

Although the ODDS Database is only an approximation of medical images, and does not perfectly match the task that medical image readers undertake, its images can be successfully searched by novice participants. This is desirable when one wants to study novice searchers who require lengthy training to develop enough perceptual expertise to recognize anomalous tissue in medical images. For example, Sha et al. (2020) explored perceptual learning in novice searchers learning to recognize and localize lung cancer in chest radiographs. In the first session, participants' ability to choose which of two radiographs contained cancer was only slightly better than chance, and their localization ability was even worse (between 30 and 35%). Over the course of four training sessions, these abilities improved, but it required participants to return to the lab on consecutive days (see

[6] Note that many exceptions to this exist. For example, Adamo et al. (2018) developed a paradigm in which participants searched for Ts among Ls embedded into medical images.

Sowden et al., 2000, for a similar approach). By contrast, the naïve volunteers in our second experiment accurately localized between 76 and 94% of the targets, depending on stimulus category. Given that many experiments in university settings consist of only a single session, researchers may wish to use noisy, occasionally ambiguous stimuli from the ODDS Database to answer questions about poorly specified anomaly search.

Limitations and future directions

Although the images presented in the ODDS Database are well-structured scenes, the search processes they encourage are better described as anomaly search than scene search. For example, scene search is often heavily guided by scene semantics (e.g., objects presented in typical or atypical locations, given surrounding context), whereas the anomalies in the ODDS Database are random with respect to surrounding context. This more closely aligns the ODDS Database with paradigms in which difficult perceptual deformations (e.g., Gabor patches) or small embedded letters serve as targets. Kosovicheva and Bex (2021) recently examined the influence of local image statistics (e.g., luminance, edges/boundaries, salience) on search performance for briefly presented Gaussian patches embedded in natural visual scenes. They found that observers' attention, as indexed by both eye movements and target localization responses, was quickly captured by salient information (e.g., dark regions, edges) or landmarks within the scene. Although our paradigm differed in several ways from that used by Kosovicheva and Bex, their findings may nevertheless explain observers' subtlety ratings in Experiment 1. For example, subjective impression of the examples in Fig. 1 suggests that subtlety ratings were affected by whether targets appeared in areas with dense local features (see also Bex et al., 2009; Wallis & Bex, 2012) or at edges/boundaries. Although raters were instructed to ignore their own search performance when evaluating target subtlety, the drop in subtlety–performance correlations across Experiments 1 and 2 suggests that they could not entirely discount their own metacognitive evaluations of search. It is possible that their metacognitive assessments were affected by the speed with which they located targets, which is partially driven by local image statistics.

The images in this database are meant to be broadly analogous to medical images, and should be useful to researchers interested in studying visual search domains that are more challenging than what is captured by typical laboratory paradigms. Although many forms of medical image screening are guided by patient history and symptom presentation, incidental anomaly detection (e.g., in emergency department imaging) nevertheless remains a common cancer detection method for certain cancers (e.g., colorectal; Esteva et al., 2018). The ODDS stimuli could prove useful for examining incidental detections or the development of expertise, and how it changes both perception and the oculomotor patterns engaged by observers (e.g., Papesh et al., 2021). This could lead to investigations of gist processing and localization, as in studies where participants quickly categorize medical images as anomalous or not, followed by localization decisions (e.g., Brunye et al., 2021). By tracking eye movements during search through the ODDS Database, researchers could explore the different types of errors that searchers make (search errors, recognition errors, decision errors; Kundel et al., 1978) or the extent to which these errors depend on the definition of a useful field of view (UFOV; Wolfe et al., 2021).

The image manipulations used for the ODDS scenes were modeled after those used by Hess et al. (2016), who studied search for camouflaged targets. "Camouflage-breaking" can be considered a special instance of visual search wherein observers must identify objects of interest that blend into the background (Branch et al., 2021). The ODDS stimuli might be useful for understanding search behaviors in tasks that routinely require camouflage breaking. For instance, search and rescue responders might be asked to locate missing hikers or hunters whose clothing blends into their surroundings (cf., Koester, 2008), or police and/or military personnel may need to localize threats (e.g., adversaries, weapons) that have been intentionally hidden from view (Riggs et al., 2018).

The stimuli in this database may also prove useful in basic laboratory paradigms. For example, Lancry-Dayan et al. (2021) have suggested that active template maintenance is not necessary for successful visual search. This is an exciting possibility, but thus far has only been tested with faces as stimuli. The nature of the distortions in the ODDS Database stimuli make specific template maintenance unlikely, making them a reasonable choice to investigate search without templates.

Conclusions

In conclusion, the ODDS database presents a new tool for studying the types of searches routinely performed by medical professionals. These stimuli circumvent many of the problems with using standard visual search paradigms, because the scenes present ill-specified and unnamable targets embedded, with varying degrees of subtlety, within noisy scene contexts. The images in this database should allow researchers to explore new questions about anomaly search using non-expert participants.

Appendix

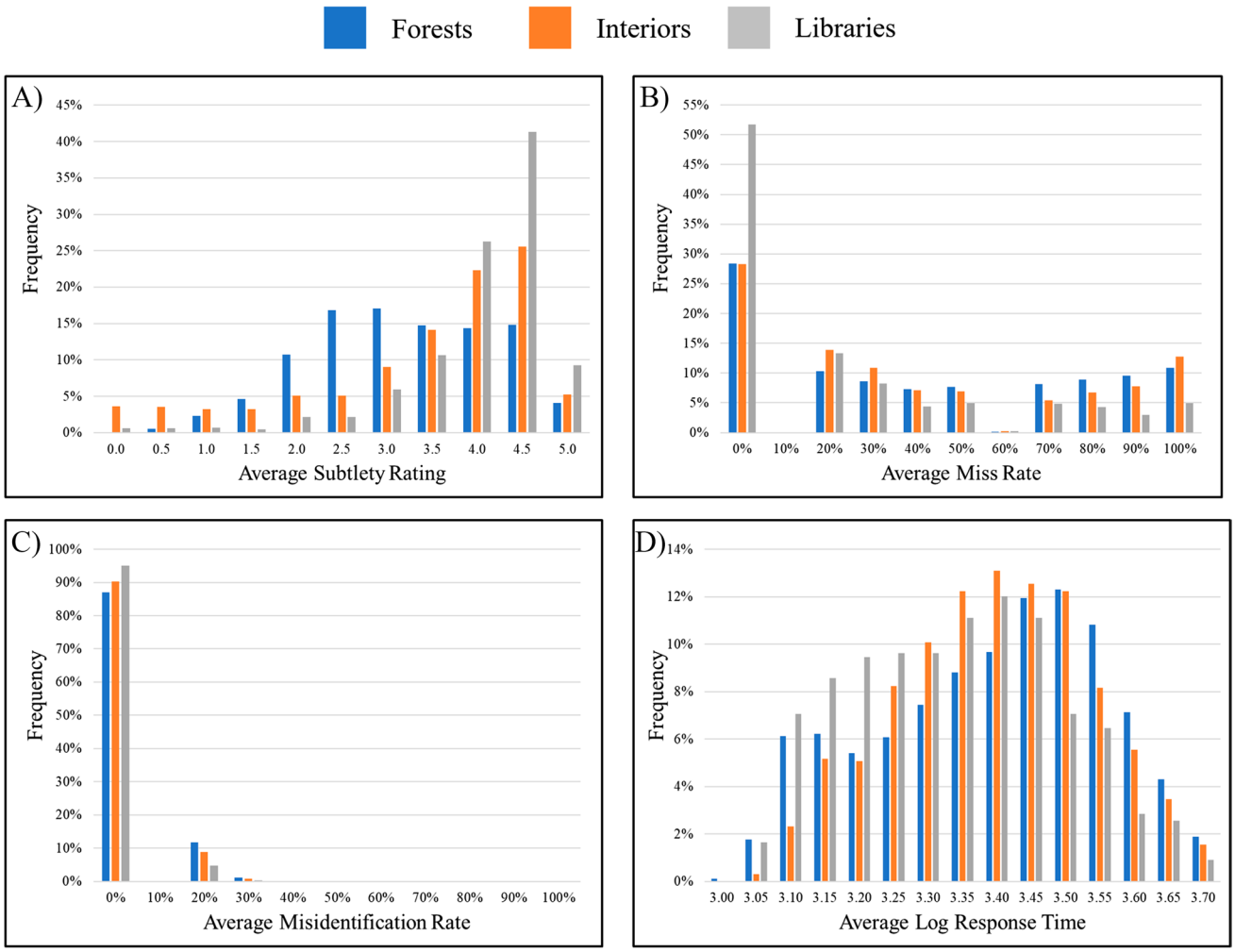

Fig. 7 Histograms of subtlety ratings and search performance presented separately for forests (*blue bars*), interiors (*orange bars*), and libraries/books (*gray bars*). Panel **A** presents subtlety ratings (with lower values indicating more subtle targets), Panels **B** and **C** present search errors (misses and misidentifications, respectively), and Panel **D** presents log-transformed search RTs

Histograms of aggregate subtlety ratings and search performance across categories in Experiment 1 are shown in Fig. 7. We did not conduct inferential statistics on these measures because of our small sample size and, more importantly, because cross-category differences are not of key theoretical interest. Rather, these descriptive statistics show broad similarities and differences between categories and may prove useful for selecting stimuli in future experiments.

Panel A depicts average subtlety ratings for each category. The distribution of ratings for forests approximates a normal curve, but the interiors and libraries/books are skewed negatively. This is most likely due to the fact that interiors and libraries contain more straight edges than the natural environments and therefore targets tended to be more noticeable in these scenes. Panels B and C present search errors; Panel B shows average miss rates (i.e., trials in which the volunteer did not click before 5 s had elapsed) and Panel C shows average misidentification rates (i.e., trials in which the volunteer clicked on the wrong area of the display, defined as being more than 60 pixels from the center point of the target). Misses are distributed across the entire spectrum, with many scenes having targets that were never missed, few scenes in which the target was always missed, and no obvious differences across categories. Misidentifications, by contrast, were quite rare overall, with most scenes exhibiting no misidentifications at all, and again no obvious differences between categories. Lastly, Panel D presents log-transformed search response times (RTs; for correctly identified targets only). Log-transformed RTs appear approximately normally distributed with no apparent differences across categories.

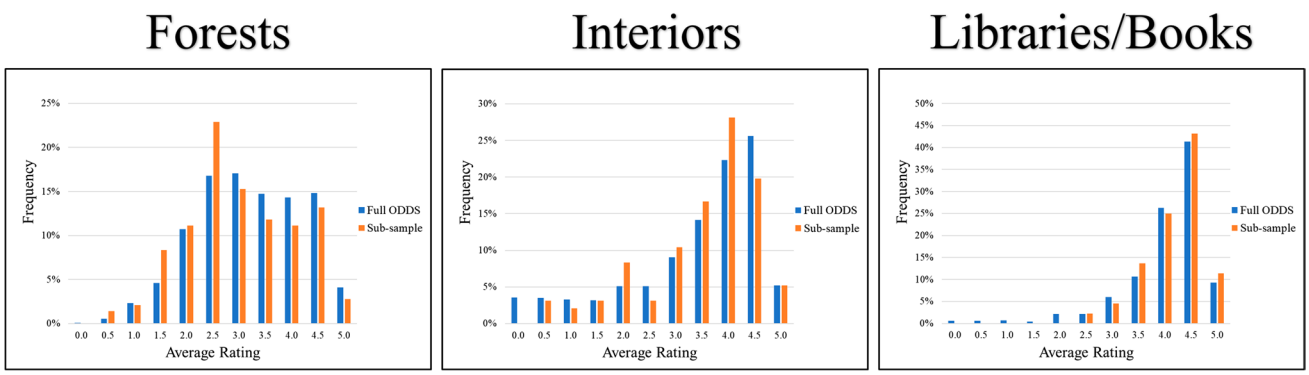

Fig. 8 Histograms showing the distribution of subtlety ratings for the entire ODDS database (*in blue bars*) and the pre-selected sub-samples (*in orange bars*) from Experiment 2. Results are presented separately for each stimulus category (forests, interiors, libraries/books in the *left*, *middle*, and *right* panels, respectively). Note that plots have different scales on the *Y*-axis

Acknowledgments Research reported in this publication was supported by an *Institutional Development Award (IDeA)* from the *National Institute of General Medical Sciences* of the *National Institutes of Health* under grant number P20GM103451. It was also supported by a gift donated by the *Electronic Caregiver, Inc.*

Author Note Any opinions, findings, and conclusions expressed in this material are those of the authors and do not necessarily reflect the views of the *National Science Foundation*. Research reported in this publication was supported by an *Institutional Development Award* (IDeA) from the *National Institute of General Medical Sciences* of the *National Institutes of Health* under grant number P20GM103451. Our work was also supported by a gift from the *Electronic Caregiver, Inc.* We thank Alice Godwin for insightful comments on a draft of this work.

Data availability All de-identified processed data, metadata, and stimuli are available at the Open Science Foundation: https://osf.io/b85e2/?view_only=3762e6e18d494e45bc0e387dfa8c62e2.

Code availability Not applicable.

Declarations

Ethics approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards. The study was approved by the Institutional Review Board (IRB), granted by the New Mexico State University Office of the Vice President for Research, application #21,430.

Consent to participate All volunteers indicated acknowledgment of informed consent to participate (via keyboard press) in the experiments prior to the start of the tasks.

Conflicts of Interest / Competing interests The authors declare no conflicts of interest nor competing interests.

References

Adamo, S. H., Ericson, J. M., Nah, J. C., Brem, R., & Mitroff, S. R. (2018). Mammography to tomosynthesis: Examining the differences between two-dimensional and three-dimensional visual search. *Cognitive Research: Principles and Implications, 3*, 17. https://doi.org/10.1186/s41235-018-0103-x

Aizenman, A., Drew, T., Ehinger, K. A., Georgian-Smith, D., & Wolfe, J. M. (2017). Comparing search patterns in digital breast tomosynthesis and full-field digital mammography: An eye tracking study. *Journal of Medical Imaging, 4*, 045501.

Auffermann, W. F., Krupinski, E. A., & Tridandapani, S. (2018). Search pattern training for evaluation of central venous catheter positioning on chest radiographs. *Journal of Medical Imaging, 5*(3), 031407. https://doi.org/10.1117/1.JMI.5.3.031407

Auffermann, W. F., Little, B. P., & Tridandapani, S. (2015). Teaching search patterns to medical trainees in an educational laboratory to improve perception of pulmonary nodules. *Journal of Medical Imaging, 3*(1), 011006. https://doi.org/10.1117/1.jmi.3.1.011006

Becker, S. I. (2010). The role of target–distractor relationships in guiding attention and the eyes in visual search. *Journal of Experimental Psychology: General, 139*(2), 247.

Bex, P. J., Solomon, S. G., & Dakin, S. C. (2009). Contrast sensitivity in natural scenes depends on edge as well as spatial frequency structure. *Journal of Vision, 9*(10), 1–19. https://doi.org/10.1167/9.10.1

Borji, A., Sihite, D. N., & Itti, L. (2013). What stands out in a scene? A study of human explicit saliency judgment. *Vision Research, 91*, 62–77.

Branch, F., Lewis, A. J., Santana, I. N., & Hedge, J. (2021). Expert camouflage-breakers can accurately localize search targets. *Cognitive Research: Principles and Implications, 6*, 27. https://doi.org/10.1186/s41235-021-00290-5

Bruno, M. A., Walker, E. A., & Abujudeh, H. H. (2015). Understanding and confronting our mistakes: The epidemiology of error in radiology and strategies for error reduction. *Radiographics, 35*, 1668–1676. https://doi.org/10.1148/rg.2015150023

Brunye, T. T., Drew, T., Saikia, M. J., Kerr, K. F., Eguchi, M. M., Lee, A. C., May, C., Elder, D. E., & Elmore, J. G. (2021). Melanoma

in the blink of an eye: Pathologists' rapid detection, classification, and localization of skin abnormalities. *Visual Cognition, 29*, 386–400. https://doi.org/10.1080/13506285.2021.1943093

Drew, T., Vo, M.L.-H., Olwal, A., Jacobson, F., Seltzer, S. E., & Wolfe, J. M. (2013). Scanners and drillers: Characterizing expert visual search through volumetric images. *Journal of Vision, 13*, 1–13.

Erdfelder, E., Faul, F., & Buchner, A. (1996). GPOWER: A general power analysis program. *Behavior Research Methods, Instruments & Computers, 28*, 1–11. https://doi.org/10.3758/BF03203630

Esteva, M., Ruiz-Díaz, M., Sánchez, M. A., Pértega, S., Pita-Fernández, S., Macià, F., et al. (2018). Emergency presentation of colorectal patients in Spain. *PLoS ONE, 13*(10), e0203556. https://doi.org/10.1371/journal.pone.0203556

Evans, K. K., Birdwell, R. L., & Wolfe, J. M. (2013). If you don't find it often, you often don't find it: why some cancers are missed in breast cancer screening. *PLoS ONE, 8*(5), e64366.

Godwin, H. J., Menneer, T., Liversedge, S. P., Cave, K. R., Holliman, N. S., & Donnelly, N. (2017). Adding depth to overlapping displays can improve visual search performance. *Journal of Experimental Psychology: Human Perception and Performance, 43*, 1532–1549.

Godwin, H. J., Menneer, T., Riggs, C. A., Cave, K. R., & Donnelly, N. (2015a). Perceptual failures in the selection and identification of low-prevalence targets in relative prevalence visual search. *Attention, Perception, & Psychophysics, 77*, 150–159.

Godwin, H. J., Menneer, T., Riggs, C. A., Taunton, D., Cave, K. R., & Donnelly, N. (2015b). Understanding the contribution of target repetition and target expectation to the emergence of the prevalence effect in visual search. *Psychonomic Bulletin & Review, 23*, 809–816.

Hess, A. S., Wismer, A. J., Bohil, C. J., & Neider, M. B. (2016). On the hunt: Search for poorly defined camouflaged targets. *PLoS ONE, 11*, 1–18.

Hou, X. & Zhang, L. (2007). Saliency detection: A spectral residual approach. *2007 IEEE Conference on Computer Vision and Pattern Recognition.* https://doi.org/10.1109/CVPR.2007.383267

Hout, M. C., & Goldinger, S. D. (2015). Target templates: The precision of mental representations affects attentional guidance and decision-making in visual search. *Attention, Perception & Psychophysics, 77*, 128–149. https://doi.org/10.3758/s13414-014-0764-6

Hout, M. C., Robbins, A., Godwin, H. J., Fitzsimmons, G., & Scarince, C. (2017). Categorical templates are more useful when features are consistent: Evidence from eye-movements during search for societally important vehicles. *Attention, Perception, & Psychophysics, 79*, 1578–1592. https://doi.org/10.3758/s13414-017-1354-1

Hout, M. C., Walenchok, S. C., Goldinger, S. D., & Wolfe, J. M. (2015). Failures of perception in the low-prevalence effect: Evidence from active and passive visual search. *Journal of Experimental Psychology: Human Perception & Performance, 41*, 977–994. https://doi.org/10.1037/xhp0000053

Koester, R. J. (2008). Lost person behavior: A search and rescue guide on where to look for land, air, and water. *Dbs Production, LLC (Charlottesville, VA).*

Kok, E. M., Jarodzka, H., de Bruin, A. B. H., BinAmir, H. A. N., Robben, S. G. F., & van Merriënboer, J. J. G. (2015). Systematic viewing in radiology: Seeing more, missing less? *Advances in Health Sciences Education, 21*(1), 189–205. https://doi.org/10.1007/s10459-015-9624-y

Kosovicheva, A., & Bex, P. J. (2021). Gravitational effects of scene information in object localization. *Scientific Reports, 11*, 11520. https://doi.org/10.1038/s41598-021-91006-8

Krasovskaya, S., & MacInnes, W. J. (2019). Salience models: A computational cognitive neuroscience review. *Vision, 3*(4), 56. https://doi.org/10.3390/vision3040056

Krupinski, E. A. (2010). Current perspectives in medical image perception. *Attention, Perception, & Psychophysics, 72*(5), 1205–1217. https://doi.org/10.3758/APP.72.5.1205

Kundel, H. L., Nodine, C. F., & Carmody, D. (1978). Visual scanning, pattern recognition, and decision-making in pulmonary nodule detection. *Investigative Radiology, 13*, 175–181.

Kundel, H. L., & Polansky, M. (2003). Measurement of observer agreement. *Radiology, 228*(2), 303–308.

Lancry-Dayan, O. C., Gamer, M., & Pertzov, Y. (2021). Search for the unknown: Guidance of visual search in the absence of an active template. *Psychological Science*. https://doi.org/10.1177/09567976219966 60

Montabone, S., & Soto, A. (2010). Human detection using a mobile platform and novel features derived from a visual saliency mechanism. *Image and Vision Computing, 28*, 391–402.

McDonald, R. J., Schwartz, K. M., Eckel, L. J., Diehn, F. E., Hunt, C. H., Bartholmai, B. J., Erickson, B. J., & Kallmes, D. F. (2015). The effects of changes in utilization and technological advancements of cross-sectional imaging on radiologist workload. *Academic Radiology, 22*, 1191–1198.

Papesh, M. H., Hout, M. C., Guevara Pinto, J. D., Robbins, A., Lopez, A. (2021). Eye movements reflect the development of expertise in hybrid search. *Cognitive Research: Principles and Implications, 6(7).* https://doi.org/10.1186/s41235-020-00269-8

Psychology Software Tools, Inc. [E-Prime v3.0]. (2012). http://www.pstnet.com

Riggs, C. A., Godwin, H. J., Mann, C. M., Smith, S. J., Boardman, M., Liversedge, S. P., & Donnelly, N. (2018). Rummage search by expert dyads, novice dyads and novice individuals for objects hidden in houses. *Visual Cognition, 26*, 334–350. https://doi.org/10.1080/13506285.2018.1445678

Sha, L. Z., Toh, Y. N., Remington, R., & Jiang, Y. V. (2020). Perceptual learning in the identification of lung cancer in chest radiographs. *Cognitive Research: Principles and Implications, 5*, 1–13. https://doi.org/10.1186/s41235-020-0208-x

Shiraishi, J., Katsuragawa, S., Ikezoe, J., Matsumoto, T., Kobayashi, T., Komatsu, K.-I., Matsui, M., Fujita, H., Kodera, Y., & Doi, K. (2000). Development of a digital image database for chest radiographs with and without a lung nodule: Receiver operating characteristic analysis of radiologists' detection of pulmonary nodules. *American Journal of Roentgenology, 174*, 71–74.

Tatler, B. W., Hayhoe, M. M., Land, M. F., & Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of Vision, 11*(5), 1–23.

Tatler, B. W., & Vincent, B. T. (2009). The prominence of behavioural biases in eye guidance. *Visual Cognition, 17*, 1029–1054.

Veale, R., Hafed, Z. M., & Yoshida, M. (2017). How is visual salience computed in the brain? Insights from behavior, neurobiology and modelling. *Philosophical Transactions of the Royal Society B., 372*, 20160113. https://doi.org/10.1098/rstb.2016.0113

Waite, S., Grigorian, A., Alexander, R. G., Macknik, S. L., Carrasco, M., Heeger, D. J., & Martinez-Conde, S. (2019). Analysis of Perceptual Expertise in Radiology – Current Knowledge and a New Perspective. *Frontiers in Human Neuroscience, 13*, 213. https://doi.org/10.3389/fnhum.2019.00213

Wallis, T. S., & Bex, P. J. (2012). Image correlates of crowding in natural scenes. *Journal of Vision, 12*(7), 1–19. https://doi.org/10.1167/12.7.6

Williams, L. H., Carrigan, A. J., Mills, M., Auffermann, W. F., Rich, A. N., & Drew, T. (2021). Characteristics of expert search behavior in volumetric medical image interpretation. *Journal of Medical Imaging, 8*, 1–24. https://doi.org/10.1117/1.JMI.8.4.041208

Williams, L. H., & Drew, T. (2019). What do we know about volumetric medical image interpretation?: A review of the basic science and medical image perception literatures. *Cognitive Research: Principles and Implications, 4*, 21.

Wolfe, J. M., & Horowitz, T. S. (2017). Five factors that guide attention in visual search. *Nature Human Behaviour, 1*, 0058. https://doi.org/10.1038/s41562-017-0058

Wolfe, J. M., & Utochkin, I. S. (2019). What is a preattentive feature? *Current Opinion in Psychology, 29*, 19–26. https://doi.org/10.1016/j.copsyc.2018.11.005

Wolfe, J. M., & van Wert, M. J. (2010). Varying target prevalence reveals two dissociable decision criteria in visual search. *Current Biology, 20*, 121–124.

Wolfe, J. M., Wu, C.-C., Li, J., & Suresh, S. B. (2021). What do experts look at and what do experts find when reading mammograms? *Journal of Medical Imaging, 8*, 1–22. https://doi.org/10.1117/1.JMI.8.4.045501

Open Practices Statement The data and materials for all experiments are available at the Open Science Foundation (https://osf.io/b85e2/?view_only=3762e6e18d494e45bc0e387dfa8c62e2). None of the experiments were pre-registered.